Event-Symmetric Space-Time

By Philip Gibbs

Event-Symmetric Space-Time

A free electronic book

How much can physics explain? "Event-Symmetric Space-Time" presents a startlingly integrated world view from the forefront of physics. So often we read about the new quantum paradigm which has replaced the old mechanistic philosophy of physics, but seldom do we find *"what the paradigm is"* spelt out so succinctly. "The universe is made of stories, not of atoms." (Muriel Rukeyser) This is the storyteller's point of view. Through a literal interpretation of those words we transcend causality and determinism to see the quantum multiverse as a whole.

Throughout this book, the author returns to the principle of event symmetry -- in particle physics, in cosmology, in superstring theory, in epistemology. Coupled to the storyteller's paradigm this new idea of philosophy and physics dares to free us from the constraints of our intuition, to reveal natures truths. We are in the midst of a revolution in our understanding of physics and the universe. This new interpretation of superstring theory is slowly helping to bring physicists' long search for the holy grail of knowledge to fruition.

At the debut of the twentieth century Einstein revealed how the laws of nature are independent of any co-ordinate system. According to general relativity, no matter how a reference frame of space-time is turned, pulled and stretched, the laws of physics remain the same because gravity keeps track of the changes. Einstein's only restriction was that he did not allow space-time to tear. You cannot cut out two pieces of space-time and swap them over expecting the forces of nature to compensate, or can you? Research attempting to form a theory of quantum gravity suggests that space-time *can* tear and reconnect in ways which change its topology. This book suggests that Einstein's symmetry must be extended to allow space-time to be atomised into space-time events which can be pulled apart and recombined in any permutation. The unified forces of nature must permit this "event symmetry" just as gravity already permits the more restricted co-ordinate transformations.

Recently theorists have discovered matrix models of superstring theories which vindicate these ideas. In this new picture pioneered by Leonard Susskind at

Stanford, the co-ordinates of space-time have been replaced by anti-commuting matrices in the same way that quantum theory modified the commuting observables of position and momentum seventy years before. Now it is space-time itself which is being remoulded. In the limit where the co-ordinate matrices commute they can be diagonalised simultaneously so that their eigenvalues represent the co-ordinates of classical space-time events. The order of these events can be permuted under the symmetry of the model and thus the principle of event-symmetry is realised. In the true non-commuting geometry this is generalised to a matrix group symmetry which unifies gauge symmetry and particle statistics. These features had been previously predicted as a natural outcome of event symmetry. Other predictions to be found in this book, such as the relationship between multiple-quantisation and dimension, may further help string theorists to understand the nature of space and time.

In a style which mixes technical notes with clear expositions, "Event-Symmetric Space-Time" will be of interest to researchers in physics and interested non-professionals alike.

"I admired the way in which he so lucidly expressed such complex ideas."

- Caroline Pretty, Assistant Editor, Penguin Books

"I must say that I like these idea quite a lot. I experimented with strings on discrete lattices and found some very interesting behavior. The strings were formed from sequences of links on an ordinary lattice but I am convinced that one really must link any pair of points so that the theory should have this huge permutation symmetry. I think that physics at and beyond the Hagedorn temp will require such concepts. In particular I do not believe that the black hole mess can be sorted out without understanding these things."

- Leonard Susskind, Professor of Physics, Stanford

Table of Contents

V.
What About Causality?

VI.
The Superstring Supermystery

VII.
The Principle of Event Symmetry

VIII.
Event-Symmetric String Theory

IX.
Is String Theory in Knots?

X.
The Theory of Theories

The Storyteller

Between a story and the world

The storyteller, surrounded by his enthralled audience, softly ended his tale. After a few moments of silence a young voice from the front asked a question. "What is the difference between a story and the world?"

The storyteller replied "There is no big difference. The world is just a story told with too much irrelevant detail."

"That's nonsense!" The words came from a teacher listening from the back. "The world is real, tangible, concrete. A story is just made up fiction."

"A child knows that a story can be as real as anything." said the storyteller. "As people grow older they learn to separate a part they see as the real world from the rest, but they are mistaken. Some continue to regard certain stories as real which others come to regard as fiction. A story is not made up. It is discovered!"

The storyteller and the teacher might argue for many hours about what is real. For centuries physical science has been based on a paradigm which considers the universe as real and material. Other things are held apart and regarded as part of the imagination. In the real world, events are governed by the laws of physics and causality. In our imagination anything goes.

As the second millennium draws to an end, science is searching for a new paradigm. Many surprising discoveries have been made over the past century and causality has been cast into doubt. Above all our own place in the universe is a great mystery. Often physicists have remarked that the laws of physics seem to be designed so that life could evolve. But if the universe was designed just for us why was it necessary that we evolve? Why not just put us there? In quantum physics it seems to be impossible to separate the laws of physics from our role as observers. Does the universe depend on us to work? And what about consciousness? What, if anything, does it mean to be aware of our own existence? In the past these questions were regarded as unscientific but now many scientists are trying to tackle them and the old paradigm is totally inadequate.

Our storyteller sees the world differently. To him all stories already exist and are real. We do not create them. We find them. The universe is no different. It might be helpful to see it as a coherent collection of stories which unfold. He may not be able to persuade you to accept this immediately, so in the best storyteller's tradition, he asks you to suspend your disbelief. If you can take his advice it will help you to come to terms with some of the unusual things in physics which I am going to describe in this book. I want to tell you about how space can evaporate and how time might change direction. Some people find such things hard to accept as a possible part of real experience, yet somewhere, somewhen they may happen.

Try to imagine that there is a very large number of real or hypothetical storytellers all telling their favourite stories. They may be in this universe - past, present or future - or perhaps they are somewhere else, they may be very different from storytellers as we know them. It does not really matter. Some storytellers will be telling the same stories as others, perhaps with different details, or they may be telling stories which start the same but end differently. There are so many possible storytellers in our imagination that this is not really a coincidence. Some will tell stories which are sequels or prequels of others. Sometimes one story will seem to be the story of what is going on next door to the location of another. Many of the stories will be very imaginative when compared to our limited experience. They may even make little sense to us, but somewhere in the whole collection any possible story is being told.

Stories can be broken down into components such as chapters, sentences and words. Those elements might fit together in other ways. So the stories fit together to create whole universes like random jigsaws. Just for your entertainment here is a story broken down into phrases and jumbled up. It is a well-known anecdote told by a famous physicist who himself has an important role to play in this chapter. Can the phrases be put together uniquely?

```
"Hee-heh-heh-heh-heh. Surely You're joking, Mr. Feynman."
and there are some ladies,
"I'll have both thank you," I say,
I go through the door,
and some girls, too.
when I hear a voice behind me.
still looking for where I'm going to sit,
"Would you like cream or lemon in your tea, Mr. Feynman?
and I'm thinking about where to sit down
It's Mrs Eisenhart, pouring tea.
and should I sit next to this girl, or not,
It's all very formal
when suddenly I hear
and how should I behave,
```

You might solve this puzzle, either exactly or with a slight variation which does not change the meaning. If there were many more phrases, or if they were broken down into words you might end up with a story different from the original. If I gave you just a jumble of letters and punctuation marks, you could produce just about anything. Putting together the vast number of stories which can be told would be the same. There would be no unique solution but you could make some order out of the chaos.

To understand the physics of event-symmetric space-time which I am going to explain, you must imagine that the universe is built this way. There are many possible stories and where stories fit together in a self-consistent way they combine to form many different universes. Each of us has a life which is a story somewhere in these universes. We should not expect our future to be completely determined since what we have experienced up to now could fit into many stories with different endings. Even our pasts, and events happening elsewhere in our present, may not be fully determined, yet we are guaranteed a consistent story in the end. The storyteller's arena of universes is called the *multiverse* and this is the storyteller's paradigm.

If you are not very impressed, remember that a paradigm is not a theory. It is just an empty vessel within which you can place a theory. The storyteller's paradigm is much more flexible

than other paradigms such as mechanism, materialism and causality. It needs to be if new physics is to be comprehensible.

Dreams of Rationalism

On the night of November 10th, 1619, René Descartes was serving in the army of the Duke of Bavaria. They were in the midst of the thirty years war which burned across the continent. Outside it was bitterly cold and Descartes, 23 years old, had fallen into an uneasy sleep in the stove-heated room.

During that night he had three dreams, showing him his past, present and future. The first dream terrified him. A ghostly presence showed him a melon which he interpreted as a sign of solitude and human preoccupations. He was in pain; a punishment. In the second dream he heard thunder which brought home his present uncomfortable predicament, but the thunder was the Spirit of Truth coming for him. He lay awake reflecting on these signs before having his third and most revealing dream. In front of him on a table he saw two books, a dictionary and a book of poems. A stranger appeared and showed him a poem, "Est et Non" by Pythagoras.

This was the turning point in his life. He changed his ways. From that time on, Descartes would pursue a reconstruction of knowledge based on physics and mathematics. He came to believe that a unified system of truth was attainable. The realisation of that vision has been sought by generations of scientists throughout the centuries which followed. Today we have not yet reached it but we seem closer than ever before.

On that night in 1619 the time was certainly right for a new science. Just ten years before, Galileo had looked to the sky with his telescope. He had seen mountains on the moon, the phases of Venus, moons of Jupiter, sunspots and millions of new stars not known before. Never since in the history of our world, has one person announced a catalogue of so many unexpected discoveries all at once. With these observations Galileo had crushed the old worldview and physics of Aristotle. Now it was clear that the Earth was just like another planet circling the Sun as Copernicus and Kepler had surmised. Galileo also judged that the same laws of physics which act on Earth must also rule the heavens. Just imagine the excitement of those times. Plainly it was the beginning of something big. Much more could be seen and known than previously thought possible. A new physics would have to be worked out to fit the new facts and a new philosophy to go with it.

Descartes had heard of Galileo's discoveries as a 15 year old student at La Flèche. In response, Descartes drew up a picture of the world as the workings of a complicated machine whose motion is governed by simple physical laws. He said that everything which happened must have a prior cause. He hoped that the right laws could be found by looking to mathematics and logic. By knowing the equations and solving them, humankind would understand the mechanism of the universe.

This Cartesian rationalism can be understood as two elements of causality. There is *temporal causality* which means that if we know the positions and velocities of all particles at a given time, and the laws which govern the forces between them, then we can understand their motions

at all future times. To Descartes, rationalism also meant that all things had a deeper explanation in terms of simpler causes. This is *ontological causality*. Nothing comes from nothing. The Cartesian philosophy was a reaction to the scientific method which had been described by Francis Bacon just a few years before. What mattered to Bacon was experiment and observation, but Descartes put more weight on the use of rational logic and deduction to work out how things should be.

People often criticise scientific theories, saying that they do not explain anything. They say that Maxwell's electromagnetism does not explain what charge or magnetic fields are, or that general relativity does not explain what space-time or inertia is. Physicists will argue that explanation in this sense is not what counts. The important thing is that the theory provides a successful means of predicting the result of experiments. The scientific method requires that physical theories must be drawn up in response to observations and tested empirically. Anything more is just metaphysical.

Yet physicists are themselves always searching for deeper explanations and often express their wishes for an underlying theory from which all phenomena can, in principle, be derived. What scientists do is often different from what they report. To Descartes, experimental results are just hints that we need because we are not clever enough to work things out from first principles. He admitted the shortcomings of his method and resorted to experiments himself, but he hoped to rectify the matter later. The last in order of discovery would be the first in order of knowledge. This dichotomy between the scientific method and Cartesian rationalism has survived intact since the time of Descartes and Bacon and has become an ironic feature of scientific progress. Descartes himself predicted that the journey on the road to that ultimate discovery was to be a long one taking centuries to follow.

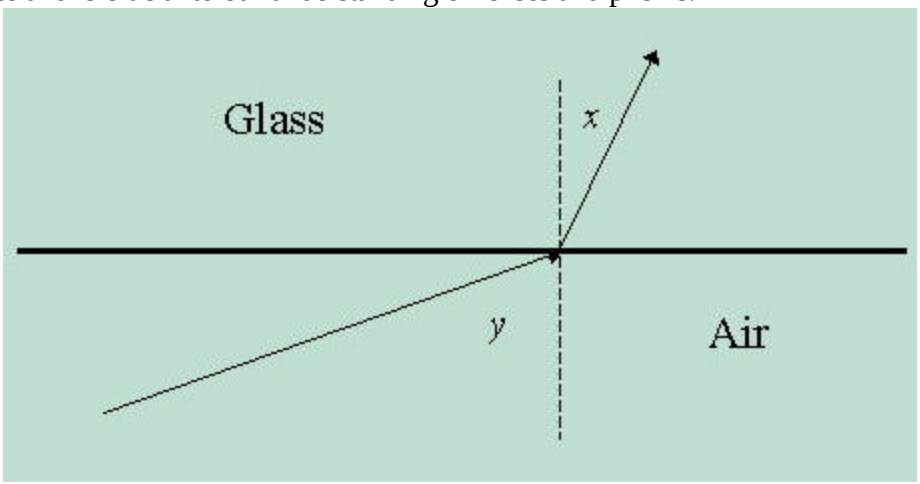
Descartes became a great mathematician. He became the founder of analytic geometry as well as modern western philosophy. When Newton spoke of "standing on the shoulders of giants" he meant Descartes as well as Galileo, Kepler and Copernicus who had set in motion the scientific revolution during the previous century. Together those individuals, and many others who joined them, established a new order which would last until the twentieth century. Newton used his prodigious mathematical skills to bring Descartes's dream to life. Applying Cartesian geometry, he defined absolute space and time as the arena for deterministic mechanical law.

The pillars of absolute space, time and determinism were the supporting structures of physics until the end of the nineteenth century. Then they crumbled, but the notion that all cause comes from the past and from deeper laws has remained as the foundation stone of all science. Causality is now firmly embedded in our thought but it was not always so. Before the mechanistic paradigm, philosophers viewed change as part of becoming towards a purpose. To Aristotle an acorn has a destiny to become a tree, it has *telos* and that is why it grows. At least some of the cause was seen to lie in the future. A child will become an adult, always developing towards perfection. Lead will become gold in the fullness of time. Descartes had expelled Aristotle's final cause, but Newton had reservations and believed that final cause may yet play its part. What can be said of temporal causality could also be said of ontological causality. The reasons for existence may not all lie in the past or in the underlying laws of nature. We have come too far to return to teleology and mysticism, but we need to prepare for a wider view of

causality. There may be no first cause, no deepest cause, no final cause or highest cause; just a sea of interdependent possibilities; a synthesis of consistent stories.

Light on Light

Among the many scientific discoveries made by Descartes is a contribution to optics which is commonly known as Snell's sine law of refraction. It was named after the Dutch mathematician, Willebrod Von Roijen Snell who discovered it just prior to Descartes in 1625. Snell died just a year after his discovery and did not publish, so the law was not widely known until Descartes published it in 1637. The law tells us how light bends when passing between two mediums such as air and glass and is crucial to our understanding of lenses and prisms.



The product of the refractive index and the sine of the angle of incidence of a ray in one medium is equal to the product of the refractive index and the sine of the angle of refraction in a successive medium.

$$r_{glass} \sin x = r_{air} \sin y$$

Descartes provided a derivation of Snell's law which we now know to be incorrect, even though it gave the right answer. He envisaged light as the motion of small spherical particles. He could see that it is easy to explain light reflected from a mirror as a stream of particles which bounce off the smooth surface, as balls bounce from a wall. The component of velocity of the particles tangent to the surface does not change while the normal component is reversed. In accordance with his general methods, Descartes wanted a similar mechanical description of refraction. When light passes from air into a denser medium such as glass, it turns towards the normal of the surface. If the tangential component of velocity is to remain unchanged for refraction as it is for reflection, light must go faster in the denser medium.

Newton later perfected Descartes derivation and agreed with his conclusion. He claimed that particles of light are attracted to denser mediums when they enter, and so gain momentum perpendicular to the surface. We can compare the situation with balls which roll across a flat surface until they descend a short downward slope onto another flat surface. They will gain

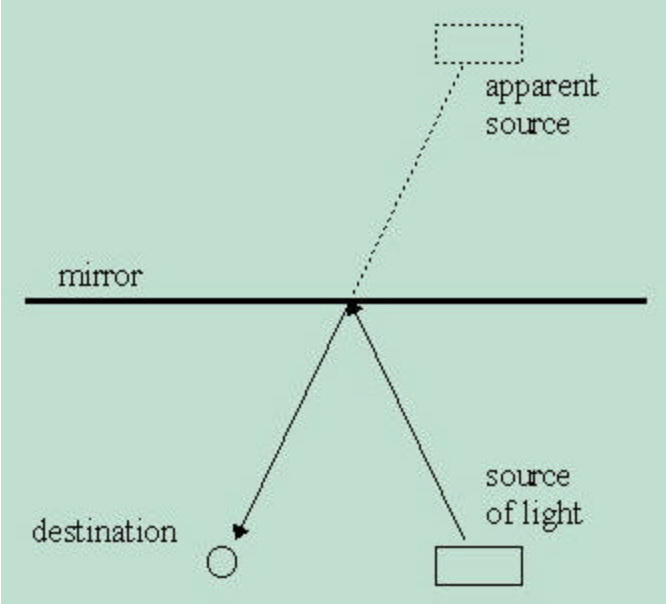
energy and speed up, but only the normal component of velocity changes. The result is that they change direction, and if the initial velocity is fixed then the angles of deflection will mimic Snell's sine law. This is the essence of the Cartesian-Newtonian mechanistic explanation of refraction.

At that time, the French mathematician Marin Mersenne was acting as a clearing house for scientific information in Europe. It is no accident that knowledge began to expand rapidly after Johann Gutenberg introduced the printing press to Europe in 1450. Communication has always been of vital importance in the development of science. Mersenne's role was the 17th century equivalent of today's electronic e-print archives on the internet. When he received Descartes's manuscript on optics in 1637 he circulated copies to other scientists including Fermat.

Pierre de Fermat was by profession a councillor of the French parliament, but his passion was mathematics and his theorems in number theory are legendary. When he read Descartes's derivation of the sine law of refraction he was not impressed. For one thing, he felt that some unjustified assumptions had been made. He also felt that, if anything, light should slow down in a denser medium, not speed up. The ensuing argument between Descartes and Fermat petered out quickly without resolution.

Some twenty years later Fermat decided to try and conclude the matter by finding a better explanation for refraction. His philosophy was very different from that of Descartes. Instead of seeking a mechanical analogy he fell back on the old idea of Aristotle that nature always takes the most economical way. In 125 AD Heron of Alexandria had shown that the law of reflection from a mirror could be explained if rays of light were taking the shortest path from the source to destination via the surface of the mirror. This can be easily seen by looking through the mirror at the path of light before reflection. The ray traces a straight line from the apparent position of the object in the mirror to the destination.

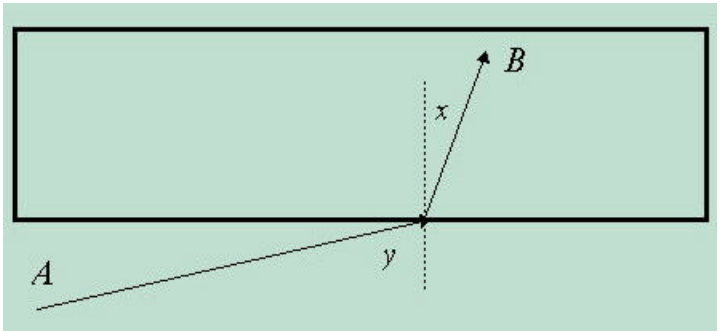
If the angle of incidence were not the same as the angle of reflection it would not be a straight line and would therefore be a longer path.



Fermat was interested in problems of finding maxima and minima before Newton and Leibniz developed the general methods of differential calculus. He considered the hypothesis that the path of the ray of light might give a minimum in the time taken for light to go from A to B. This would work equally well as minimum distance for reflection and could also explain refraction.

Imagine that instead of a light ray passing into a block of glass, it is a life guard at the swimming pool. While standing at position A she sees a swimmer in distress at position B. She needs to get to him as quickly as possible but can run twice as fast as she can swim. To get from A to B in the shortest time she would have to follow the path shown.

It is not the path of shortest distance.



She must first get to a point at the side of the pool nearer to the swimmer. The optimum route is given by the equivalent of Snell's law,

$$2 \sin x = 1 \sin y$$

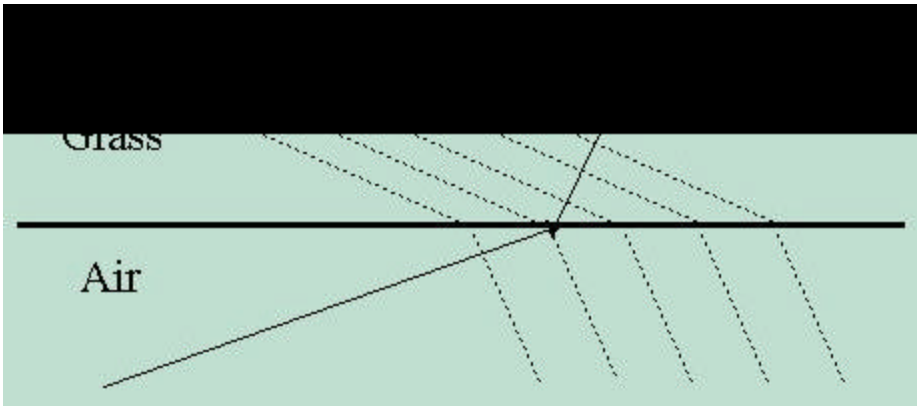
A ray of light going from a point A to a point B in a rectangular block of glass with a refractive index of two would take the same route. Thus, in 1657, Fermat showed that if light was being

slowed down in a medium by a factor equal to its refractive index, then he could derive Snell's sine law from a principle of least time. He was astonished that he got the same refraction law as Descartes even though his alternative theory predicted a slowing down of light in dense media instead of a speeding up. It was not until 1850, almost 200 years later, that Jean Foucault was able to measure directly the speed of light in different media. He confirmed that light slowed down in water. Fermat was right and Descartes was wrong.

The beauty of Fermat's principle of least time is its generality. The implication is that a ray of light passing through any complex set-up of mirrors and lenses takes a route which gives at least a local minimum of time to go from start to finish. According to Descartes's notion of causality, Fermat's principle is a bizarre way to formulate a law of physics. What we expect are laws which allow us to begin with a starting point and direction for a ray of light, and then work out the route it takes and where it will end up. Of course, Fermat's principle can be used in this way via a derivation of Snell's law, but it seems to work as if the light was given a starting and end position and then worked out the optimum route between them. This is quite absurd in terms of temporal causality.

By the mid 17th century the nature of light was a subject of hot debate. Important experiments by the Italian Francesco Grimaldi in 1648 were then becoming known. Grimaldi had observed diffraction of light and proposed that light had a wavelike nature.

At this time a wave theory of sound was already well established. Galileo had studied a vibrating string and clarified the relationship between frequency and pitch in 1600. In 1636 Mersenne had made the first measurements of the speed of sound by timing the return of an echo and in 1660 Robert Boyle demonstrated that sound could not travel through a vacuum by placing a bell in a jar and pumping out the air. The conclusion was inescapable. Sound must be due to compression waves travelling through the air. Using this theory, Isaac Newton was able to calculate the speed of sound from first principles and obtain a result in agreement with Mersenne's measurement.



Newton's rival, Robert Hooke, was one of those who wanted an analogous theory of light but he failed to see that light must slow down in dense media rather than speed up. In 1673 Ignace Pardies corrected Hooke's oversight and provided a new explanation for Snell's law. If light propagated in a direction perpendicular to wave fronts and slowed down as it passed through a dense medium, then waves become closer together and would be deflected in accordance with

the sine law. Christian Huygens agreed but wanted a deeper understanding. Why should the wave theory be in agreement with Fermat's principle? Huygens was from Amsterdam so it is easy to imagine how he might have seen the effects of water waves on the many canals of the city as he walked home across the bridges. He developed an intuition for the behaviour of waves which enabled him to grasp a deep relation between the wave theory of light and the principle of least time. Newton and Huygens were both followers of Descartes's mechanistic philosophy, but they had very different views of the road ahead. Newton liked Descartes's theory of light and incorporated it into his corpuscular theory. Huygens started from a different observation made by Descartes, that crossed beams of light pass through each other without interacting. He must have noticed that water waves and sound waves pass through each other in a similar way. He could not see how this would be possible for light if it was composed of streams of particles.

Huygens explained instead that light propagated from each point of a luminous source in spherical waves. These are analogous to the circular waves propagating from a disturbance on the surface of water, but with immense speed and short wavelength. The speed of light was deduced by Olaus Roemer in 1676 to account for a discrepancy in the timing of eclipses of Jupiter's moons. The short wavelength could be confirmed by an experiment which Newton performed, now known as Newton's rings. Huygens noticed that if water waves pass through a tiny hole smaller than their wavelength they again spread out from that point in spherical waves. He said that spherical secondary waves propagated from any point but are only seen clearly when a barrier shields the contributions from other points. At that time the mathematics needed to express the propagation of waves in the form of differential equations was not available, but by combining Huygens's principle of secondary waves with the effects of interference, it is possible to explain refraction and diffraction. It is even possible to see why Fermat's principle of least time applies: Constructive interference appears at points where light wave fronts passing by different routes from the source arrive after the same time of travel so that they are in phase. This corresponds to the paths of least time. This conveniently reduced Fermat's principle to a deeper wave principle which, to Huygens, had the greater merit of being explicitly causal and Cartesian.

Newton saw things very differently. In his theory, light was composed of particles or corpuscles. These corpuscles undulated with a frequency depending on their colour. This was his explanation for the experiment in which he was able to measure the wavelengths of light of different colours by observing the rings of light between two glass surfaces.

There the matter rested without further progress during the whole of the eighteenth century. Newton's corpuscular theory and Huygens wave principle were seen as opposing theories. Because of the huge success of Newton's mechanics and theory of gravitation, he was the greater authority and his ideas were favoured. Newton objected to the wave hypothesis because light casts a sharp shadow whereas sound and water waves can bend round an obstruction. In the nineteenth century, opinion swung the other way. Thomas Young and Augustin Fresnel were first to revive the wave theory of light with new theory and experiments to study interference and diffraction. With the superior mathematical methods of Fourier and Laplace and the experimental basis of Ampere, Faraday, Henry, Oersted and others, rapid progress was made. James Clerk Maxwell presented the unified theory of electromagnetism in 1864. Nine years later he had derived the speed of light by supposing it to be a form of electromagnetic wave. With

this, all aspects of light known at the time including colour and polarisation could be explained. Newton's corpuscular theory was no longer needed, it seemed.

Light-Quanta

Occasionally an important breakthrough in physics comes about because of someone asking an important question which others had not thought of. History will give the greater glory to the one who finds the answer but often it is the person who posed the question who made the greater contribution to science. This was the case in 1860 when Gustav Kirchhoff asked: "What is the electromagnetic spectrum from a black-body?" He realised that the radiation inside a uniformly heated box must not depend on the characteristics of the walls, otherwise the second law of thermodynamics could be violated by letting radiation pass from box to another at a slightly higher temperature. In that case the energy in the radiation from an ideal black body must be a function of wavelength and temperature which should be explainable solely in terms of fundamental physics. However there was no theory at that time which could be used to derive the answer and experiment could give only a rough guide. In the decades that followed Maxwell's theory was to be found wanting when applied to Kirchhoff's simple question. As the nineteenth century drew to a close Lord Rayleigh showed that Maxwell's equations and the laws of thermodynamics predicted a spectrum which worked well at low infra-red frequencies but which would give a nonsensical increasing intensity of emission at higher ultra-violet frequencies. In fact there would be an infinite radiation of heat. Something was badly wrong with the theory. In Berlin at the world's best equipped physics laboratory of the time, two teams were painstakingly measuring black-body radiation at temperatures from well below freezing up to as high as 1500 °C. Most theorists could do little better than guess equations which might fit the empirical curves. Finally it was Max Planck who wrote down the correct law which fitted the data. Then Planck went a step further than guesswork. He concluded, reluctantly, that the spectrum at high frequencies diminished because the radiation was emitted in discrete quanta. Thus in 1900, the quantum era began.

It was not easy for physicists to accept the new idea. At first it was thought that the quantisation may apply only to emission and perhaps absorption of light, and not as a property of light propagation. For the first two decades of the twentieth century, Albert Einstein alone believed that light quanta were real. He applied the same idea to explain the photoelectric effect and successfully predicted the correct law, $E = hf - P$, of photoelectric emission. In 1915 after 10 years of experiment a sceptical Robert Millikan conceded that the formula was correct. It was Einstein who in 1909 saw the need for a theory of particle-wave duality. It was he too, who in 1917 saw the first signs that determinism was threatened. He understood that in the phenomenon of stimulated light emission, the exact moment at which each light quantum would be emitted, could not be determined from the initial state. To Einstein this was an unacceptable breakdown of causality which he hoped to fix later in a deeper theory. To other physicists who followed it became an experimentally verified fact of life. The breakdown of causality was, however, postponed by a semantic adjustment. We now say that quantum mechanics is indeterministic rather than acausal. We mean that although we cannot determine the outcome of an experiment, the result is still influenced only by the past state and not the future. Cartesian temporal causality could live to see another century.

In 1913 Niels Bohr used the theory of light quanta to explain the Balmer series of emission lines in the spectrum of hydrogen, but what did it mean? In 1923, Arthur Compton derived the relativistic expression for hard scattering of a quantum of light from an electron. The term "light quanta" was replaced by the word "photon" as if to celebrate its wider recognition as a particle. No longer would the reality of photons be questioned. It was impossible to deny the particular side to their nature when the Compton effect was photographed in cloud chambers and energy and momentum conservation was verified.

The almost fantastic story of those discoveries and the years that followed have filled many volumes on the history of science. In that golden age of physics many great scientists rose to the challenge. Heisenberg, Pauli, Fermi, Schrödinger, Dirac, ... the roll-call is endless. Now is a good moment to turn the clock back to the time of Newton and his theory of undulatory corpuscles. One can only marvel at the profound insight implied by this theory. To be sure, Newton was wrong to think that light is faster in dense media. Huygens and Fermat were correct that it slows down. It must also be admitted that everything Newton had observed was later consistent with the wave theory when it found its final form in Maxwell's equations. Yet Newton's anticipation of the quantum theory was no fluke. It grew out of a belief that the laws of physics were unified. Following the chemist and philosopher Robert Boyle, he guessed that everything was built from elementary units. It was Boyle who had christened them corpuscles. History recounts that this was inspired by alchemist sympathies. They wanted to believe that any form of matter could be transformed into another because they dreamt of becoming rich by transforming lead into gold. But their guess that such transformations might come about by rearrangements of the constituent corpuscles was founded on many observations of other physical processes. It was natural for Newton to suppose that light was produced by another transformation of this sort. We know now that he was right, and we should not scoff just because the theory was not based purely on empirical induction from solid observations.

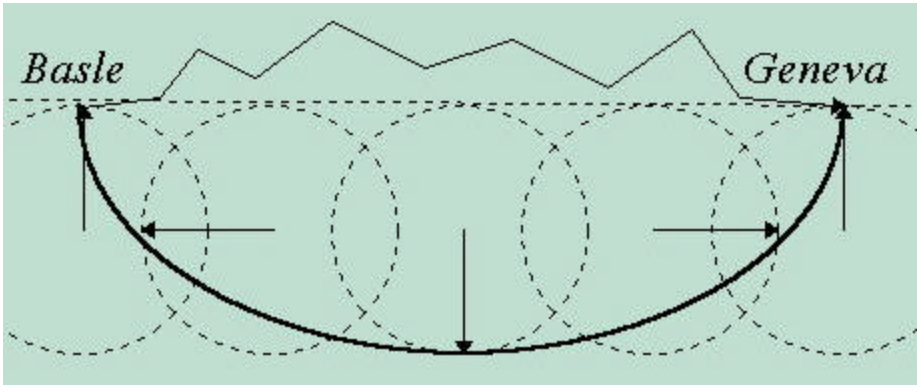
With hindsight we can see the modern theory of light as a synthesis of the principles of Newton, Fermat and Huygens. Explaining how, will lead up to my thesis of the storyteller's paradigm, but first we must go back and trace the development of another principle.

The Principle of Least Action

At the end of the seventeenth century, European mathematicians liked to show off their prowess by posing and solving puzzles. The Bernoulli brothers particularly enjoyed this game and Jean Bernoulli, the 10th child of Nicolaus Bernoulli, set an especially tricky problem for his rival and older brother Jacques. In 1690 he asked him to identify the curve of the brachistochrone, the curve down which a particle will slide in the shortest time from one given point to another. An interesting application of this problem would be to build an underground train between two towns powered only by gravity. Suppose the line was to go from the Bernoulli's home town of Basle to Geneva, 259km to the south-west. By descending down a steep slope from Basle, it could pick up momentum to cover the distance on frictionless tracks. Then, using its kinetic energy, it would finish by climbing back up to Geneva where it would come perfectly to rest. What would be the optimum shape of the track to minimise the travel time? Jean failed to trip up his brother with this problem and other mathematicians solved it too. Newton is reputed to have cracked the problem overnight when it was given to him. The solution is a cycloid; the curve

traced out by a point on the rim of a rolling wheel. To get from Basle to Geneva the train would follow the sweep of a point on a circle as it did a full revolution.

It would descend to a maximum depth of 82.4km where it would reach a speed of 4580 km per hour and it would complete its journey in only 6 minutes 47 seconds.



The brachistrochrone puzzle influenced other mathematicians to look for general methods of solving other similar optimisation problems which involved curves, and so the calculus of variations was invented. Since it grew out of a physical problem, physicists wondered how the new maths might be applied to Newton's laws of mechanics more widely. Remember that according to Fermat's principle, a ray of light follows the line of shortest time through any system of mirrors and prisms. Could there be a more general principle to be found? Gottfried Leibniz was especially keen on the idea. He did not like the Cartesian exclusion of final cause and saw Fermat's principle as an example that demonstrated his point.

But applying Fermat's principle directly to mechanics does not work. Particles do not seem to be trying to get from A to B in the least time possible, otherwise they would accelerate towards their destinations. A free particle goes in a straight line so its path has the minimum length, but it would be better to have a principle which explains why it goes at constant speed too. Leibniz proposed that mechanics optimises the use of another quantity which he called *action*. Later, in 1744, Pierre de Maupertuis discovered how to make this idea work. For the single particle subjected to no forces the action is energy multiplied by time which is also half momentum times distance integrated along the path. When a particle travels from A to B in a fixed time interval, it does so with the least possible action. Maupertuis attached great philosophical significance to this principle and was ridiculed by Voltaire for doing so. Yet it is hard for a student learning mechanics not to be struck by the beauty and generality of the principle of least action when he first encounters it. Richard Feynman was one such student who heard about it from his high school physics teacher. The consequences for Feynman and for physics were profound, as we shall see.

The calculus of variations and the principle of least action were further developed in the eighteenth century by mathematicians such as Leonhard Euler and Joseph Lagrange. For any mechanical system moving in an energy potential, the action is defined as the kinetic energy minus the potential energy integrated with respect to time.

When the system evolves from an initial state to a final state at given times, it does so in a way which minimises the action. Euler and Lagrange showed how to derive the equations of motion of any system of particles from this principle. This energy difference in the integral is now called the Lagrangian and finding its form for more general situations is the key to any problem of theoretical physics. The principle of least action is a curious discovery from the point of view of causality in the same fashion as for Fermat's optical principle of least time. Recall that in classical mechanics (meaning deterministic motion without the quantum theory), given the initial positions and velocities of particles and the equations of force acting on them, you can in principle predict their subsequent motion. This is the principle of temporal causality. However, the principle of least action tells us how a system evolves given the initial and final positions of the particles and the equation for the action. It is as if the evolution of the system is determined equally by the past and future. Causality is only found indirectly through the derivation of the equations of motion and, apparently, our own psychological bias for prior cause.

The next in line to work on the action principle were William Hamilton and Carl Jacobi. They developed techniques now known as the Hamilton-Jacobi formalism which took them to the brink of discovering quantum mechanics in 1834, eighty years before its time. Recall that Huygens had used his theory of secondary waves to provide an explanation for Fermat's principle which reconciled it with causality. If Hamilton or Jacobi had considered a similar explanation of the principle of least action they could easily have found quantum wave mechanics. As it turned out, we only see this with the hindsight which came from eighty more years of experimentation. It is amusing to consider that we could write a fictional but almost plausible sounding history in which mathematicians discovered all the fundamental principles of physics without ever doing an experiment! In practice, Descartes has to concede that we need those empirical signposts to keep us from straying onto false paths. Does it have to be that way or is it just a human weakness?

In the real story it was 1923 that became the breakthrough year for quantum mechanics. Einstein had already suggested particle-wave duality for light quanta in 1909, but only when Louis de Broglie suggested that the same must apply to electrons did all become clear. He was only a student at the time but he realised immediately that the Hamilton-Jacobi theory pointed in that direction. Duality was, and still is, a hard lesson to learn. It had to be accepted because it made sense, at last, of Bohr's model of the atom. Many who would otherwise have doubted were swayed by convincing experiments. Electron diffraction from metals was seen as the perfect confirmation of deBroglie's matter wave theory. It was the time of the greatest revelations in physics. Within three short years the full theory of quantum mechanics was established and ten Nobel Laureates had earned their physics prizes in the process.

Feynman Meets Dirac

It is difficult to think of two twentieth century physicists less alike in character than Paul Dirac and Richard Feynman. Born in Bristol, West of England, Dirac was a quiet genius, a man of few words, over-typically the reserved Englishman. He was a master of imaginative speculation; exploiting mathematical beauty to invent new physics. He discovered the relativistic equation of the electron and founded quantum field theory. Later in life, he anticipated string theory, membrane theory and magnetic monopoles thirty years in advance of their time. His masterpiece

was the systematic construction of the quantisation process described in his book, "The Principles of Quantum Mechanics". It showed how to derive a quantum theory from any classical Hamiltonian mechanics by introducing a quantum state vector and replacing classical commuting quantities with non-commuting quantum operators.

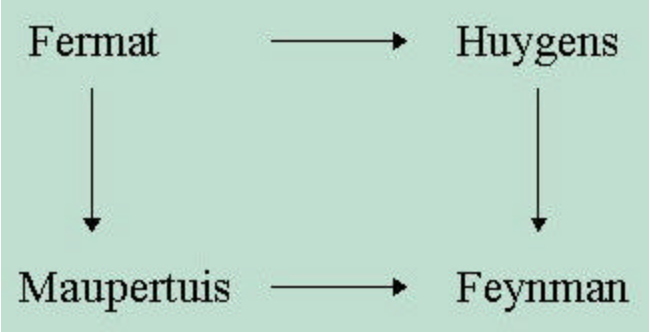
Feynman was born in New York City, 16 years younger than Dirac. He was a popular genius, an outspoken character, over-typically American. His approach to physics was practical and down to Earth. He was brilliant at finding new ways to look at things more clearly and solving physical problems. He found the modern approach to quantum field theory and renormalisation. He explained superfluids and tackled quantum gravity directly. He wrote a series of lecture notes on theoretical physics which will remain standard texts for decades to come. His masterpiece was an alternative formulation of the process of quantisation using path integrals.

Despite these different styles, Feynman was a great admirer of Dirac's work. In 1946 they met for the first time at a series of lectures which had been organised to celebrate the bicentennial of Princeton University. After giving a talk, Feynman found Dirac resting on the lawn outside by himself, and went out to talk to him. He wanted to ask about an expression which Dirac had written in a paper in 1933, about the relation between quantum mechanics and the principle of least action. Dirac had found what he thought was an approximate relationship but Feynman saw that it was exact. This was his opportunity to ask Dirac if he actually knew that. In fact Dirac had not known but said it was a very interesting observation. As a result, Feynman thought some more about it and had a marvellous flash of insight. Suddenly he could see a very direct and intuitive relation between the classical action and quantum theory.

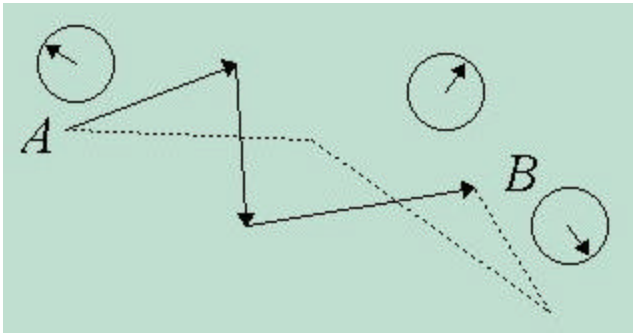
Feynman's Sum Over Stories

To understand what Feynman came up with let us first look at the simple case of a single particle. In 1923 Louis De Broglie suggested that if light waves behave as particles, then other particles must also be considered to have wave properties. Almost immediately Davisson and Kunsman were able to verify De Broglie's conjecture by observing electron diffraction effects. In 1926 Erwin Schrödinger came up with a more detailed wave theory in which the state of the particle at any time is actually described by a complex valued number assigned to each point in space. Soon after that, Max Born interpreted Schrödinger's wave function as a description of the probability of finding a particle at any point in space. The probability density is given by the square of the wave amplitude.

The wave evolves according to a wave equation which Schrödinger gave us and which was later generalised by many others. Now Feynman, inspired by Dirac, realised that the evolution of the wave could also be described in terms of what he called "path integrals". The relationship between Feynman's path integral and Maupertuis's principle of least action is the same as that between Huygen's principle of secondary waves and Fermat's principle of least time. The square was completed.



According to Feynman, in order to find the evolution of the wave function for a single particle between given starting and end times, we must consider all possible starting points A, all possible finishing points B and all paths P which the particle could take in going from A to B. The value of the wave function at the start time is a complex number which can be pictured as the position of the hand of a clock. Suppose that initially the particle has a definite position at A so the wave function takes the value 1 there and zero everywhere else. We now want to know what the wave function will look like at some later finishing time. As a path P from A to B is traced out, the action can be calculated using the classical equations of Lagrange. Imagine that the hand of the clock turns as if clocking up action along the path until it gets to B so that it ends up at some other position on the clock face. For each path from A to B there is a different position value. To get the final amplitude of the wave function at B you have to sum up, or integrate, the values for all the paths. This path integral has a built in normalisation so that the final answer has a sensible value.



The evolution is wavelike since the turning hands of the clock are like the phase of a wave. When the dials read the same values they add together like constructive interference. When they point in opposite directions they cancel like destructive interference. Constructive interference is most pronounced when paths near to the minimum of the action are added together. This explains why the principle of least action describes the motion of the particle in the classical limit.

The path integral makes sense, at last, of the theories of light of both Huygens and Newton. Previously seen as rivals, they are now seen as complementary. The path integral incorporates Huygen's secondary waves and generalises his explanation of Fermat's principle, but it also describes light as particles with an undulatory nature as Newton wanted.

But quantum mechanics deals with much more than just light. Any system which has a classical principle of least action can be quantised using the methods of Dirac or Feynman. A system of many particles interacting through forces which conserve energy can be dealt with in this way. An example is an atom consisting of a nucleus with its entourage of electrons. Classically we would describe such a multi-particle system by giving the positions of each particle in space. The quantum wave function of one particle is a complex valued function on the 3 co-ordinates of space, so it might have been expected that the quantum wave function of n particles would involve n such functions. In fact it is more complicated than that. The wave function is a much bigger complex valued function of the 3n co-ordinates of the positions of all the particles.

It is non-local in the sense that it does not just give independent wave functions for each particle. It also describes correlations between them. If a group of n friends goes out to town for the evening you could give a probability for each bar, club and cinema, that each friend will be there at 11 o'clock. If there are h such haunts that they like to go to, there would be nh such probabilities. However, these probabilities alone would be a very poor description of the total behaviour because some friends like to stick together and are more likely to be found together. There are actually h^n possible situations at 11 o'clock and to account for all possible circumstances you must give the probability for each one. The situation for particles is similar except for a few important details. Firstly, as already said, the wave function gives a complex number rather than a real number for each possibility. Also, there are an infinite number of places the particle can be at any given instant, but it may be useful to suppose that space is discrete and finite with only a fixed number h of points. Another crucial distinction between particles and our group of friends is that particles do not have names. There is no way to tell photons apart. They are absolutely identical. This means that we cannot distinguish the difference in circumstance if any two photons are swapped over. We only need to give a probability for the number of photons which can be found at each place. This is less than h^n but it is still a large number.

Electrons are a little different again. They are also indistinguishable like photons, but they never appear together in the same place. Electrons are like a group of anti-social friends who detest each other so much that each one avoids being found in the same haunt as any other. Particles actually have just these two kinds of social behaviour. Either they are like photons and do not mind being together, or they are like electrons which stay apart. Particles which are like electrons are called *bosons* and those like electrons are called *fermions*.

In the path integral of the system we cannot deal with the path of each point separately because they interact through electromagnetic forces. We must consider all ways in which the system of many particles can evolve from a given classical starting state to a final one. The action for each such possible history contributes to the evolution of the wave function. I hope that the reason for calling it a sum over stories is now emerging. We are looking at stories of particles, like a story of a group of friends who go out on the town. The story has a given beginning and a given ending and we must consider all possible stories which fit; where they could be at each moment of time. In the macroscopic world where physics appears classical, we see only one story but we know that in the microscopic world there are many stories. We are just seeing the one which dominates through constructive interference.

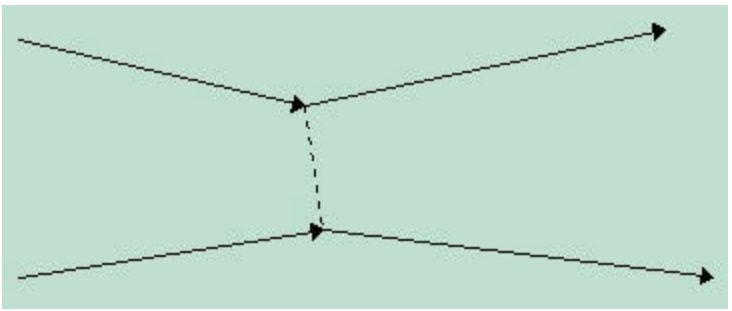
It is worth taking a moment to contemplate the complexity of the system being described. If you were an engineer charged with the task of programming a computer to simulate a galaxy at a level of detail where each particle is described individually you would balk at the task. Even doing it classically, you would require a high precision variable for each co-ordinate of some $n = 10^{70}$ particles, plus a field strength for the electromagnetic forces at each point of a closely spaced lattice over the entire galaxy. That might need $h = 10^{80}$ points. If you are required to solve the problem with quantum mechanics you need to cover the full wave function. If each particle was behaving independently you could get away with about $hn = 10^{150}$ variables, but the full wave function requires more like $(h/n)n = 10^{\wedge 10^{71}}$. Even with today's powerful computers some further approximations will be necessary.

Sometimes people talk about the "many worlds" interpretation of quantum mechanics and the multiverse of possible universes. Sceptics cannot accept it because it is hard to believe that so many things are going on in parallel. Yet quantum mechanics *is* a theory of many things happening at once and the huge size of the wavefunction for all the particles of the universe is what makes quantum mechanics work. Today physicists are looking at ways to harness the power which lies hidden in these functions. It may be possible to tame them in quantum computers which will do many simultaneous computations as if they are each happening as a separate story.

The Feynman sum over stories is a realisation of the storyteller's paradigm. It is the most fundamental principle known in physics. The quantum theory is more general and more fundamental than any other theory because it must apply to all physics if it applies to any. If we wish to understand why we exist we should not look to the big bang where we think the universe began because the temporal causality of Descartes is not what this paradigm is about. Our real origins lie in the quantum principles which are held in the physics of all times and all places.

Second Quantisation

There is a twist in the tale of quantisation which was introduced by Pascual Jordan in 1925. A single particle which is quantised becomes a field, i.e. values assigned to each point in space like the classical electromagnetic fields. A field theory can also be derived from a principle of least action and can therefore also be quantised. The field theory of the single particle Schrödinger equation can be quantised in this way as if it were a classical field. The result of this second quantisation works out to be the same as the quantum theory of a many-particle system. The Schrödinger equation is linear but quantisation can be applied to field theories with non-linear terms. The interaction between the electromagnetic field and Dirac's equation for an electron is a non-linear relativistic generalisation of the Schrödinger equation. This is still called second quantisation but not everyone likes the term used in this way. Many physicists prefer to think that the first quantisation was a mistake and quantum field theory alone is correct.



The quantum field theories always describe the quantum interactions of many particle systems. Feynman was able to use his path integrals to understand the process better. He found that the equations of quantum field theory could be written out as a sum over diagrams, now known as Feynman diagrams, which show the paths and interactions of particles

The diagrams look just like the paths of particles which described the first quantisation of many particles except now there are nodes where particles can interact. There is a subtle duality between the fields and particles. Quantising particles gives fields, and quantising fields gives particles. Like the cliché of a novel about a writer, second quantisation is confusing and perhaps there is more to be understood about what the double process means.

The Storyteller's Paradigm

A story is a cultural thing. Different peoples of the world have different traditional stories. If we found that a tribe in the Amazon knew a story which was identical to one told by the Eskimos, we would think that it was either a fantastic coincidence or that there had been some communication between them. Science is different. We expect different countries to have similar theories about biology for example, but written in different languages. This would be true even if they had not shared their discoveries because their citizens are all the same form of life and must have the same biology. If we ever make contact with intelligent life on another planet we will be interested to hear about their biology because it is likely to be rather different from terrestrial biology. However, their laws of physics will surely be the same even though they express them differently. They will know about conservation of energy and will have a list of particles which matches ours once we have sorted out how to convert terms and units. What if there are different universes where the laws of physics are different? What would life in those universes have in common with us? We would expect them to know the same mathematics because mathematical logic is more abstract than physics. They may choose different axioms as fundamental and will certainly have a different notation, but there should be a correspondence between what they judge as true and what we do.

Pure mathematicians do not usually use ideas from physics to decide what is worth studying. Yet often mathematicians working independently discover the same theorems. Perhaps one day computers will be so powerful that we will be able to simulate creative thought in a computer. Then we will verify that the same mathematical concepts can develop without any influence from physics.

According to Plato's theory of forms, the world of mathematics exists in its own right and knowledge is attainable through the study of logic. There is a hierarchy which puts maths at the foundation, physics above, natural history over that and cultural knowledge at the top. This is the scene of reductionism through Descartes' ontological causality. All knowledge is dependent on what is below, but in our lives we have more direct experience of our culture and natural history. Ultimately we want to explain our own perceptions. There is a positivist philosophy which takes the opposite extreme to Platonism saying that only the things we perceive directly are real. Perhaps the truth is a mixture of both. Is there a larger realm beyond mathematics where different rules of logic can be tried out? Perhaps there is, but it seems like it must contain itself.

The role which mathematics plays in physics is certainly a curious one. It is true that mathematics is the language of the universe. No physicist can work without it. A theory which is expressed in words may have some meaning but it is impossible to verify its correctness unless it is backed up with a mathematical model which makes testable predictions. It is hard to resist believing in an even greater significance of mathematics because we find that the most abstract concepts are applicable to the real world. It is this that Plato recognised so long ago.

If our experiences are like stories then the laws of physics are the grammar of the language in which it is written. But the same story can be told in many languages so how important is the language of physics? We still could not tell a story without words or something similar. The laws of physics can also be written in many different equivalent ways and it is not clear that any one way is more fundamental. This is a special characteristic of the laws of physics. Feynman remarked that if you modify the laws much you find that you can only write them in fewer ways.

In one language of physics, the Feynman diagrams are the words and sentences. We could collect together many diagrams and connect them together in different ways just as we can put together sentences to make paragraphs and chapters. The stories of our experience are told in that way. There are symmetries and dualities which translate from one language to another. In the Platonic sense those diagrams are the forms which exist in the world of mathematics. They join together in every possible way which the rules of logic, the grammar, permit. There is no need for temporal causality in this language. We do not need to look to some creation event where the universe was set in motion. The illusion of temporal causality itself may emerge from such an event but it does not have to be fundamental. It is a part of our story but stories with less linear structure are also possible.

What about the storyteller? Remember that in his mind he did not invent the story. He discovered it. He himself is part of another story. Perhaps this is reflected in the rule of second quantisation. Why do the Feynman diagrams obey the particular rules they do? Those rules determine which particles exist and how they interact. Do they represent some especially rich language? If the storyteller's paradigm were taken to its logical conclusion there would be no fixed Feynman rules. Feynman's sum over stories should be just part of a much larger sum over all possibilities. All of these things remain mysterious and we do not yet know the full grammar and vocabulary of physics.

The Beauty of the Tiger

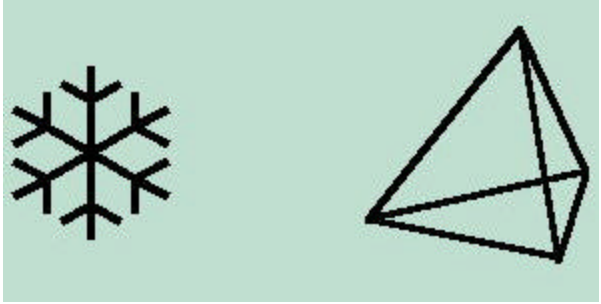
Natural Beauty

We do not have to examine nature very closely to admire its beauty. A bird, a forest or a galaxy has a form of beauty which is typical of complex organised systems in our universe. A tiger has another element to its beauty which is also very common in nature but which is often only evident on close inspection. We call it *symmetry*. The common meaning of symmetry is a well-balanced shape or design but it also has a more specific mathematical meaning. The tiger's shape and pattern are certainly well balanced but he has this mathematical symmetry too. More specifically, this symmetry of a tiger is *bilateral*: Divide his face and body by a vertical line, and the left hand side is a mirror image of the right hand side. Many animals including us have bilateral symmetry but it is especially engaging on the tiger because it is seen in his striped patterns.

A few animals and many flowers have more than bilateral symmetry. A daisy or a starfish has *radial* symmetry from its centre. Crystals also form symmetrical shapes such as octahedra and cubes. A snowflake is a crystal of ice with 6-fold radial symmetry and it is particularly elegant. How does it acquire its shape?

The snowflake begins its life as a minute hexagonal crystal forming in a cloud. The origins of this structure lie in a lattice arrangement of the water molecules which form the ice. During its passage from the clouds to the ground, it experiences a sequence of changes in temperature and humidity which cause it to grow at varying rates. Its history is recorded in the variations of thickness in its six petals as it grows. This process ensures that each petal is almost identical to any other and accounts for the snowflake's symmetry.

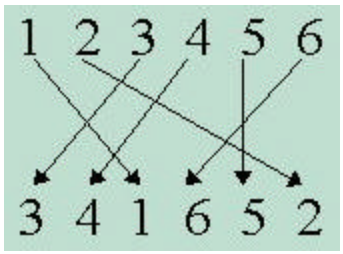
When a snowflake is rotated through an angle of 60 degrees about its centre, it returns to a position where it looks the same as before. Its shape is said to be *invariant* (meaning unchanged) under such a transformation. It is invariance which characterises symmetry in mathematics. The shape of the snowflake is also invariant if it is rotated through 120 degrees. It is invariant again if it is turned over. By combining rotations and turning over it is possible to find 12 different transformations which leave its shape invariant (including the identity transformation which does nothing). We say that the *order* of the snowflake's symmetry is 12.



Consider now the symmetry of a regular tetrahedron. That is a solid shape in the form of a pyramid with a triangular base for which all four faces are equilateral triangles. The shape of a regular tetrahedron is invariant when it is rotated 120 degrees about an axis passing through a vertex and the centre of the opposite face. It is also invariant when rotated 180 degrees about an axis passing through the midpoints of opposite edges. If you make a tetrahedron and experiment with it, you will find that it also has a symmetry of order 12. But the symmetry of the tetrahedron is not quite the same as that of a snowflake. The snowflake has a transformation which must be repeated six times to restore it to its original position and the tetrahedron does not.

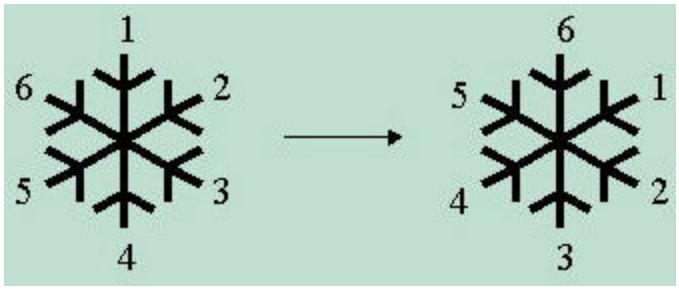
Mathematicians have provided precise definitions of what I meant by "not quite the same". The invariance transformations or *isometries* of any shape form an algebraic structure called a *group*. You can consider *composition* of transformations as a kind of multiplication. For example, two isometries of the snowflake are a rotation of 60 degrees clockwise (call it a) and a reflection about the vertical axis (call it b). The transformations are composed by doing one and then the other, a followed by b. The result is a reflection about a different axis set at 30 degrees to vertical which is also an isometry (call it c). This composition is expressed algebraically as ab = c, as if it were a multiplication. The algebraic structure defined by these elements of symmetry is the group. The order of the symmetry is the number of elements in the group. Two groups are *isomorphic* if there is a one-to-one mapping between them which respects the multiplication. Two groups which are isomorphic are often regarded as essentially the same thing. The symmetry group of a snowflake is not isomorphic to that of a tetrahedron but it is isomorphic to that of a hexagon. Groups can be considered to be a mathematical abstraction of symmetry. Many of them have symbolic names. The symmetry group of the snowflake and hexagon is called D6 while that of the tetrahedron is called A4.

The historical origins of group theory can be traced back to tragic events of May 30th 1832. That morning a young Frenchman named Évariste Galois died in a dual. At 21 years old his life was already a tale of rejection and failure as a mathematician, yet the night before he met his death he wrote a letter which brought about a revolution in abstract thought. Galois developed a theory about which polynomial equations could be solved exactly using simple arithmetic operations such as addition, multiplication and square roots. Polynomials up to degree four could be solved in this way but quintic equations had been proven insoluble by the Norwegian mathematician Niels Abel in 1823. Galois found that the answer lay in the group of permutations of the solutions of the equations. A *permutation* is a way of rearranging or shuffling an ordered set of objects. Suppose, for example, that there are six numbered objects in numerical order 1, 2, 3, 4, 5, 6. A possible permutation would be 3, 4, 1, 6, 5, 2. It can be shown as a diagram,

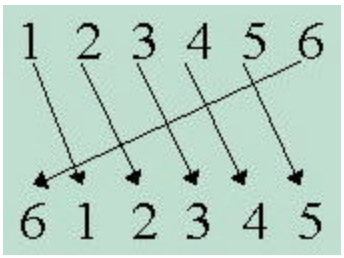


It is not really the numbers which are important. It is the arrows which permute them. There are 720, (6! = 1 x 2 x 3 x 4 x 5 x 6) different possible permutations of six objects.

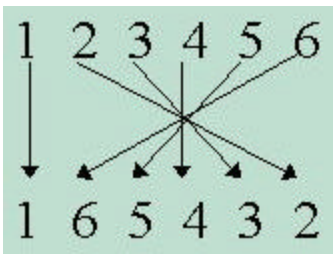
A rotation of a snowflake can be regarded as a permutation of its arms. Number them clockwise and look at the 60 degree rotation.



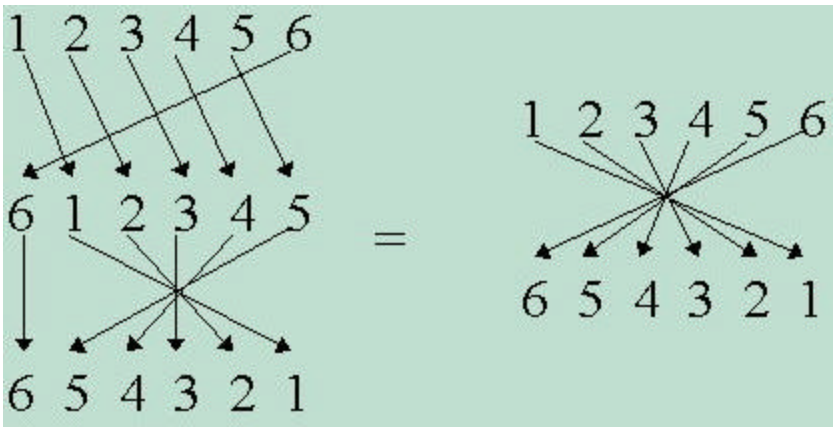
This is a permutation of the arms



Likewise a reflection about the vertical axis is another permutation



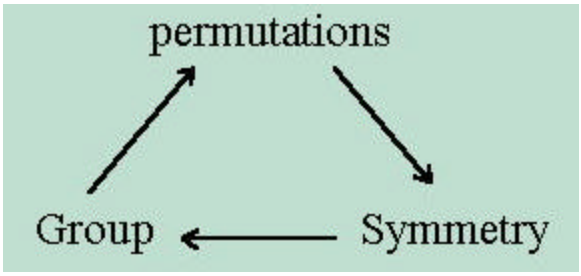
Any of the twelve transformations which leave the shape of the snowflake invariant can be shown as a permutation. To appreciate the algebraic structure of the group formed by the transformations we need to see how they can be combined. This is how it works for the rotation followed by the reflection



So combining permutations by joining the arrows is equivalent to performing one isometry followed by another. This is the same as multiplication in the group of isometries.

Like ordinary multiplication of numbers this kind of multiplication is associative, i.e. a(bc) = (ab)c for any three transformations a, b and c, but unlike ordinary multiplication it is not always commutative (ab) != (ba). There is always an identity which has the property, a1 = 1a = a. Each element has an inverse, $aa^{-1} = a^{-1}a = 1$. These algebraic rules are taken as the definition of a group.

Permutations, symmetry and groups all go together. A permutation is just a one-to-one mapping from some set to itself. A symmetry is a subset of permutations which leaves something (like shape) invariant. The algebraic structure of symmetries and the ways they combine is a group. To complete the triangle any group can also be seen as a collection of permutations of its own elements because multiplication by any element of the group is a one-to-one mapping onto itself.



The geometric symmetries of the snowflake and tiger are just one type of symmetry which leaves the shape of an object invariant. The permutations on a set of n objects also form a group which is called the *symmetric group* of the set or S_n for short. All these things are very important in physics but the theory of groups and symmetries also has its own intrinsic power and beauty which makes it interesting to mathematicians.

Permutations are not only applied to finite sets. There are also infinite order symmetries described by infinite groups and permutations of infinite numbers of objects. The simplest example is the group of rotations in a plane about the origin. It describes part of the symmetry of a circle and is known as U(1).

Symmetry in Physics

Symmetry is important in physics because there are all kinds of transformations which leave the laws of physics invariant. For example, we know that the laws of physics are the same everywhere. We can detect no difference in the results of any self contained experiment which depends on where we do it. Galileo realised just how universal this principle is when he looked at the planets through his telescope in 1609. He saw moons going round Jupiter in the same way as our moon goes round the Earth. He proposed that the laws of physics which describe the motions of the planets should be the same as those which govern the motion of objects here. This was very different from the way people had thought before. Another way to say the same thing is that the laws of physics are invariant under a translation transformation which would displace all objects by the same distance in the same direction. This is a kind of symmetry of physics which is just like the symmetry of shapes. The infinite order group of translations is a symmetry of the laws of physics.

The next important example is rotation symmetry. The laws of physics are invariant under rotations in space about any axis through some origin. An important difference between the translation symmetry and the rotation symmetry is that the former is *abelian* while the latter is *non-abelian*. An Abelian group is one in which the order of multiplication does not matter, they *commute* (ab = ba). This is true of translations but it is not true of rotations about different axis.

If the laws of physics are invariant under both rotations and translations then they must also be invariant under any combination of a rotation and a translation. In this way we can always combine any two symmetries to form a larger one. The smaller symmetries are contained within the larger one. Note that the symmetry of a snowflake is already contained within rotation symmetry. Mathematicians say that the invariance group of the snowflake is a *subgroup* of the rotation group. They are both subgroups of the full group of permutations of points of space which leave the distance between any two points invariant.

Such symmetry is important because we can use it to test new theories of physics. Once we have accepted that certain symmetries are exactly observed in nature we can check that any set of equations looks the same after applying the transformations under which physics is supposed to be invariant. If they are not then they cannot form any part of the laws of physics. Mathematicians often go much further than this and work out all possible forms the laws of physics might take to respect the symmetry. Given translational and rotational symmetry we know that the equations can be expressed using scalars, vectors, tensors and spinors; quantities which can be combined in certain ways such as using vector and scalar multiplications. Nature has been kind to physicists. With these rules they waste much less time dreaming up useless theories of physics than they would if there was no symmetry. The more symmetry they know about, the better physicists can do. This is one of the secret of their success.

Hidden Symmetry

Symmetry in physics is not always evident at first sight. When we are comfortably seated on the ground we notice a distinct difference between up and down, and between the horizontal and the

vertical. If we describe the motion of falling objects in terms of physical laws which have the concept of vertical and horizontal built in then we do not find the full rotational symmetry in those laws. As another example, many ancient philosophers thought that the Earth marked a special place at the centre of the universe. In such a case we could not say that the laws of physics were invariant under translations. In medieval times the symmetry of rotational and translational invariance in the laws of physics remained hidden to philosophers despite many centuries of observation and thought.

It was the Copernican revolution that changed all that. Nicolaus Copernicus described a cosmology in which the Earth had no special place and initiated a new freedom of thought taken up by Galileo. Newton, in response to Galileo, discovered his law of gravity which could at the same time account for falling objects on Earth and the motion of the planets in the Solar system. If the moon was subject to Earthly forces why did it not fall down like objects do on Earth.

Newton's answer was that the moon does fall, but it moves horizontally fast enough to keep it from coming down. From that point on it could be seen that the laws of physics are invariant under rotations and translations. It was a profound revelation. Whenever new symmetries of physics are discovered the laws of physics become more unified. Newton's discovery meant that it was no longer necessary to have different theories about what was happening on Earth and what was happening in the heavens.

Once the unifying power of symmetry is realised and combined with the observation that symmetry is hidden and not always recognised at first sight, the unique importance of symmetry is clear. Physicists have discovered that as well as the symmetries of space transformations, there are also more subtle *internal* symmetries which exist as part of the forces of nature. These symmetries are important in particle physics. In recent times it has been discovered that symmetry can be hidden through mechanisms such as spontaneous symmetry breaking. Such mechanisms are thought to account for the apparent differences between the known forces of nature. This increases the hope that other symmetries remain to be found.

Conservation Laws

During the centuries which followed Galileo and Newton, physicists and mathematicians came to realise that there is a deep relationship between symmetry and conservation laws in physics. The law of conservation of momentum is related to translation invariance, while angular momentum is related to rotation invariance. Conservation of energy is due to the invariance of the laws of physics with time.

The relationship was finally established in a very general mathematical form known as Noether's theorem. Mathematicians had discovered that classical laws of physics could be derived from the philosophically pleasing *principle of least action*. In 1918 Emmy Noether showed that any laws of this type which have a *continuous* symmetry, like translations and rotations, would have a conserved quantity which could be derived from the action principle.

Although Noether's work was based on classical Newtonian notions of physics, the principle has survived the quantum revolution of the twentieth century. In quantum mechanics we find that the

relationship between symmetry and conservation is even stronger. There are even conservation principles related to *discrete* symmetries.

An important example of this is *parity*. Parity is a quantum number which is related to symmetry of the laws of physics when reflected in a mirror. Mirror symmetry is the simplest symmetry of all since it has order two. If the laws of physics were indistinguishable from their mirror inverse then according to the rules of quantum mechanics parity would be conserved. This is the case for electromagnetism, gravity and the strong nuclear force. It was quite a surprise to physicists when they discovered that parity is not conserved in the rare weak nuclear interactions. Because these interactions are not significant in our ordinary day-to-day life, we do not normally notice this asymmetry of space.

Simple laws of mechanics involving the forces of gravity and electricity are invariant under time reversal as well as mirror reflection. If you could freeze every particle in the universe and then send them on their way with exactly reversed velocity, they would retrace their history in reverse. This is a little surprising because our everyday world does not appear to be symmetric in this way. There is a clear distinction between future and past. In the primary laws of physics time reversal is also only broken by the weak interaction but not enough to account for the perceived difference. There is an important combined operation of mirror inversion, time reversal and a third operation which exchanges a particle with its antiparticle image. This is known as CPT. Again the universe does not appear to realise particle-antiparticle symmetry macroscopically because there seems to be more matter than anti-matter in the universe. However, CPT is an exact symmetry of all interactions, as far as we know.

Relativity

There is another symmetry which is found in ordinary mechanics. If you are travelling in a modern high speed train like the French TGV, moving at constant speed on a long straight segment of track, it is difficult to tell that you are moving without looking out of the window. If you could play a game of billiards on the train, you would not notice any effects due to the speed of the train until it turned a corner or slowed down.

This can be accounted for in terms of an invariance of the laws of mechanics under a *Galilean transformation* which maps a stationary frame of reference onto one which is moving at constant speed. Galileo used this symmetry to explain how the Earth could be moving without us noticing it but he used a ship at sea rather than a train to demonstrate the principle.

When you examine the laws of electrodynamics discovered by Maxwell you find that they are not invariant under a Galilean transformation. Light is an electrodynamic wave which moves at a fixed speed c. Because c is so fast compared with the speed of the TGV, you could not notice this on the train. However, towards the end of the nineteenth century, a famous experiment was performed by Michelson and Morley. They hoped to detect changes in the speed of light due to the changing direction of the motion of the Earth. To everyone's surprise they could not detect the difference.

Maxwell believed that light must propagate through some medium which he called ether. The Michelson-Morley experiment failed to detect the ether. The discrepancy was finally resolved by Einstein and Poincaré when they independently discovered special relativity in 1905. The Galilean transformation, they realised, is just an approximation to a Lorentz transformation which is a perfect symmetry of electrodynamics. The correct symmetry was there in Maxwell's equations all along but symmetry is not always easy to see. In this case the symmetry involved an unexpected mixing of space and time co-ordinates. Minkowski later explained that relativity had unified space and time into one geometric structure which was thereafter known as space-time. Symmetry was again a unifying principle.

It seems that Einstein was more strongly influenced by symmetry than he was by the Michelson-Morley experiment. According to the scientific principle as spelt out by Francis Bacon, theoretical physicists should spend their time fitting mathematical equations to empirical data. Then the results can be extrapolated to regions not yet tested by experiment in order to make predictions. In reality physicists have had more success constructing theories from principles of mathematical beauty and consistency. Symmetry is an important part of this method of attack. Of course these principles are still based on observations and empiricism serves as a check on the correctness of the theory afterwards, yet by using symmetry it is possible to leap ahead of where you would get to using just simple induction.

Einstein demonstrated the power of symmetry again with his dramatic discovery of general relativity. This time there was no experimental result which could help him. Actually there was an observed discrepancy in the orbit of Mercury, but this might just as easily have been corrected by some small modification to Newtonian gravity or even by some more mundane effect due to the shape of the sun. Einstein knew that Newton's description of gravity was inconsistent with special relativity. Even if there were no observation which showed it up, there had to be a more complete theory of gravity which complied with the principle of relativity.

Since Galileo's experiments with weights dropped from the leaning tower of Pisa, it was known that inertial mass is equal to gravitational mass. Otherwise objects of different mass would fall at different rates even in the absence of air resistance. Einstein realised that this would imply that an experiment performed in an accelerating frame of reference could not separate the apparent forces due to acceleration from those due to gravity. This suggested to him that a larger symmetry which included acceleration might be present in the laws of physics.

It took several years and many thought experiments before Einstein completed the work. He knew that the equivalence principle implied that space-time must be curved, and the force of gravity is a direct consequence of this curvature. In modern terms the symmetry he discovered is known as diffeomorphism invariance. It means that the laws of physics take the same form when written in any 4d co-ordinate system on space-time. The form of the equations which express the laws of physics must be the same when transformed from one space-time co-ordinate system to another no matter how curvilinear the transformation equations are.

The symmetry of general relativity is a much larger one than any which had been observed in physics before Einstein. We can combine rotation invariance, translation invariance and Lorentz invariance to form the complete symmetry group of special relativity which is known as the

Poincaré group. The Poincaré group can be parameterised by ten real numbers. We say it has dimension 10.

Diffeomorphism invariance, on the other hand, cannot be parameterised by a finite number of parameters. It is an infinite-dimensional symmetry. Already we have passed from finite order symmetries like that of the snowflake, to symmetries which are of infinite order but finite dimensional like translation symmetry. Now we have moved on to infinite-dimensional symmetries and we still have a long way to go.

Diffeomorphism invariance is another hidden symmetry. If the laws of physics were invariant under any change of co-ordinates in a way which could be clearly observed, then we would expect the world around us to behave as if everything could be deformed like rubber. Diffeomorphisms leave the physics invariant under any amount of stretching and bending of space-time. The symmetry is hidden by the local form of gravity just as the constant vertical gravity seems to hide rotational symmetry in the laws of physics. On cosmological scales the laws of physics do show a more versatile form allowing space-time to deform, but on smaller scales only the Poincaré invariance is readily observed.

Einstein's field equations of general relativity which describe the evolution of gravitational fields, can be derived from a principle of least action. It follows from Noether's theorems that there are conservation laws which correspond to energy, momentum and angular momentum but it is not possible to distinguish between them. A special property of conservation equations derived from the field equations is that the total value of a conserved quantity integrated over the volume of the whole universe is zero, provided the universe is closed. This fact is useful when sceptics ask you where all the energy in the universe came from if there was nothing before the big bang! However, the universe might not be finite.

A final remark about relativity is that the big bang breaks diffeomorphism invariance in quite a dramatic way. It singles out one moment of the universe as different from all the others. It is even possible to define absolute time as the proper time of the longest curve stretching back to the big bang. According to relativity there should be no absolute standard of time but we can define cosmological time since the big bang. This fact does not destroy relativity provided the big bang can be regarded as part of the solution rather than being built into the laws of physics. In fact we cannot be sure that the big bang is a unique event in our universe. Although the entire observable universe seems to have emerged from this event it is likely that the universe is much larger than what is observable. In that case we can say little about its structure on bigger scales than those which are observable.

Gauge Symmetry and Economics

What about electric charge? It is a conserved quantity so is there a symmetry which corresponds to charge according to Noether's theorem? The answer comes from a simple observation about electric voltage. It is possible to define an electrostatic potential at any point in space. The voltage of a battery is the difference in this potential between its terminals. In fact there is no way to measure the absolute value of the electrostatic potential. It is only possible to measure its difference between two different points. Voltage is relative. In the language of symmetry we

would say that the laws of electrostatics are invariant under the addition of a value to the potential which is the same everywhere. This describes an internal symmetry which through Noether's theorem can be related to conservation of electric charge.

The electric potential is just one component of the electromagnetic vector potential which can be taken as the dynamical variables of Maxwell's theory allowing it to be derived from an action principle. In this form the symmetry is much larger than the simple one parameter invariance I just described. It corresponds to a change in a scalar field of values defined at each event throughout space-time. Like the diffeomorphism invariance of general relativity this symmetry is infinite-dimensional. Symmetries of this type are known as gauge symmetries. The principles of gauge theories were first recognised by Herman Weyl in 1918. He hoped that the similarities between the gravitational and electromagnetic forces would herald a unification of the two. It was many years before the full power of his ideas was appreciated.

There is an analogy of gauge symmetry in the world of finance. Consider the money which circulates in an economy. If one day the government wants to announce a currency devaluation, it has to be implemented in such a way that nobody loses out. Every price can be adjusted to be one tenth of its previous value, but everybody's wage must be changed in the same way, as must their savings. If done correctly the effect would be cosmetic. The economy is invariant under a global change in the scale of currency. It is a symmetry of the system.

What about the combined system of economies of the different countries of the world? Any one currency can revalue its currency but to avoid any economic effect the exchange rates with other currencies must also reflect the change. In this larger system there is a degree of symmetry for each currency of the world.

This is analogous to a local gauge symmetry which allows a gauge transformation to take place independently anywhere in space. Prices and wages are analogous to the wave functions of matter. Exchange rates are like the gauge fields of gravity and electromagnetism. The purpose of these fields which propagate the forces of nature is to allow the gauge symmetry to change locally, just as varying exchange rates allow economies to adjust and interact. In both cases the variables change dynamically, evolving in response to market forces in the case of economy and evolving in response to natural forces in the case of physics.

Both diffeomorphism invariance and the electromagnetic symmetry are local gauge symmetries because they correspond to transformation which can be parameterised as fields throughout space-time. In fact there are marked similarities between the forms of the equations which describe gravity and those which describe electrodynamics, but there is an essential difference too. Diffeomorphism invariance describes a symmetry of space-time while the symmetry of electromagnetism acts on some abstract internal space of the components of the field.

The gauge transformation of electrodynamics acts on the matter fields of charged particles as well as on the electromagnetic fields. In 1927 Fritz London noted that to implement the gauge transformation the phase of the wave function of matter fields is multiplied by a phase factor, which is a complex number of modulus one. Such factors have no physical effects since only the modulus of the wave function is observable. Through this action the transformation is related to

the group of complex numbers of modulus one which is isomorphic to the rotation symmetry group of the circle, U(1).

In the 1960s physicists were looking for quantum field theories which could explain the weak and strong nuclear interactions as they had already done for the electromagnetic force. They realised that the U(1) gauge symmetry could be generalised to gauge symmetries based on other continuous groups. As I have already said, an important class of such symmetries has been classified by mathematicians. In the 1920s Elie Cartan proved that a subclass known as *semi-simple Lie groups* can be described as matrix groups which fall into three families parameterised by an integer N and five other *exceptional* groups:

- The special orthogonal groups SO(N)
- The special unitary groups SU(N)
- The special symplectic groups Sp(N)
- Exceptional Groups G2 F4 E6 E7 E8

The internal gauge symmetry should be made up of combinations of these groups. They can be combined using a direct product denoted *X* in which both groups are independent subgroups.

The best thing about gauge symmetry is that once you have selected the right group the possible forms for the action of the field theory are extremely limited. Einstein found that for general relativity there is an almost unique most simple form with a curvature term and an optional cosmological term. For internal gauge symmetries the corresponding result is Yang-Mills field theory developed by Chen Ning Yang and Robert Mills in 1954. Maxwell's equations for electromagnetism are a special case of Yang-Mills theory corresponding to the gauge group U(1) but there is a generalisation for any other gauge group. From tables of particles, physicists were able to conjecture that the strong nuclear interactions used the gauge group SU(3) which is metaphorically referred to as colour. This symmetry is hidden by the mechanism of confinement which prevents quarks escaping from the proton and neutron to reveal the colour charge. For the weak interaction it turned out that the symmetry was SU(2) *X* U(1) but that it was broken by a *Higgs mechanism*. There is a Higgs boson whose vacuum state breaks the symmetry at low energies. By these uses of symmetry theoretical physicists were able to construct the complete *standard model* of particle physics by 1972.

The rapid acceptance of gauge theories at that time was due to the discovery by 't Hooft and Veltman that Yang-Mills theories are *renormalisable,* even when the symmetry is broken. Other theories of the nuclear interactions were plagued with divergences when calculations were attempted. The infinite answers rendered the theory useless. These divergences are also present in Yang-Mills theory but a process of renormalisation can be used to cancel out the infinities leaving sensible consistent results. In the years that followed this discovery, experiments at the world's great particle accelerator laboratories have rigidly confirmed the correctness of the standard model. Of the four forces only gravity remains in a form which stubbornly refuses to be renormalised.

Supersymmetry

Symmetry is proving to be a powerful unifying tool in particle physics because through symmetry and symmetry breaking, particles which appear to be different in mass, charge, etc. can be understood as different states of a single unified field theory. Ideally we would like to have a completely unified theory in which all particles and forces of nature are related through a hierarchy of broken symmetries.

A possible catch to this hope is that fermions and bosons cannot be related by the action of a classical symmetry based on a group. One way out of this problem would be if all bosons were revealed to be bound states of fermions so that at some fundamental level only elementary fermions would be necessary. This is an unlikely solution because gauge bosons such as photons appear to be fundamental.

A more favourable possibility is that fermions and bosons are related by supersymmetry. Supersymmetry is an algebraic construction which is a generalisation of the Lie group symmetries already observed in particle physics. It is a new type of symmetry which cannot be described by a classical group. It is defined as a different but related algebraic structure which still has all the essential properties which make symmetry work.

If supersymmetry existed in nature we would expect to find that fermions and bosons came in pairs of equal mass. In other words there would be bosonic *squarks* and *selectrons* with the same masses as the quarks and electrons, as well as fermionic *photinos* and *higgsinos* with the same masses as photons and Higgs. The fact that no such partners have been observed implies that supersymmetry should be broken if it exists.

It is probably worth adding that there may be other ways in which supersymmetry is hidden. For example, If quarks are composite then the quark constituents could be supersymmetric partners of gauge particles. Also, superstring theorist Ed Witten has found a mechanism which allows particles to have different masses even though they are supersymmetric partners and the symmetry is not spontaneously broken.

Supersymmetry unifies more than just fermions and bosons. It also goes a long way towards unifying internal gauge symmetry with space-time gauge symmetry. If gravity is to be unified with the electromagnetic and nuclear forces there should be a larger symmetry which contains diffeomorphism invariance and internal gauge invariance. In 1967 Coleman and Mandula proved a theorem which says that any group which contained both of these must separate in to a direct product of two parts each containing one of them. In other words, they simply could not be properly unified, or at least, not with classical groups. The algebraic structure of supersymmetry is a supergroup which is a generalisation and a classical group and is not covered by the Coleman-Mandula theorem, so supersymmetry provides a way out of the problem. There are still a limited number of ways of unifying gravity with internal gauge symmetry using supersymmetry and each one gives a theory of *supergravity*.

There is now some indirect experimental evidence in favour of supersymmetry, but the main reasons for believing in its existence are purely theoretical. During the 1970s it was discovered that supergravity provides a perturbative quantum field theory which has better renormalisation behaviour than gravity on its own. This was one of the first breakthroughs of quantum gravity.

The big catch with supergravity theories is that they work best in ten or eleven-dimensional space-time. To explain this discrepancy with nature, theorists revived an old idea called *Kaluza-Klein theory* which was originally proposed as a way to unify electromagnetism with gravity geometrically. According to this idea space-time has more dimensions than are apparent. All but four of them are compacted into a ball so small that we do not notice it. Particles are then supposed to be modes of vibration in the geometry of these extra dimensions. Yang-Mills theory emerges from space-time curvature in the compacted dimensions so Kaluza-Klein theory is an elegant way to unify internal gauge symmetry with the diffeomorphism invariance of general relativity. If we believe in supergravity then even fermions fall into this scheme.

Supergravity theories were popular around 1980 but it was found to be just not quite possible to have a version with the right structure to account for the particle physics we know about. The sovereign theory of supergravity lives in 11 dimensions and nearly manages to generate enough particles and forces when compactified down to 4 dimensions, but unfortunately it was not possible to get the left-right asymmetry in that way. It was also realised eventually that these field theories could not be perfectly renormalisable. Supergravity was quickly superseded by superstring theory. String theories had earlier been considered as a model for strong nuclear forces but, with the addition of supersymmetry it became possible to consider them as a unified theory including gravity. In fact, supergravity is present in superstring theories.

Enthusiasm for superstring theories became widespread after John Schwarz and Michael Green discovered that a particular form of string theory was not only renormalisable, it was even finite to all orders in perturbation theory. That event started many research projects which are a story for another chapter. All I will say now is that string theory is believed to have much more symmetry than is understood, but its nature and full form are still a mystery.

Universal Symmetry

We have seen how symmetry in nature has helped physicists uncover the laws of physics. Symmetry is a unifying concept. It has helped combine the forces of nature as well as joining space and time. There are other symmetries in nature which I have not yet mentioned. These include the symmetry between identical particles and the symmetry between electric and magnetic fields in Maxwell's equations of electrodynamics known as *electromagnetic duality.* Symmetry is often broken or hidden so it is quite possible that there is more of it than we know about, perhaps a lot more.

Let us look again at the symmetry we have seen so far. There is the SU(3)*X*SU(2)*X*U(1) internal gauge symmetry of the of the strong, weak and electromagnetic forces. Since these groups are gauged there is actually one copy of the group acting at each event of space-time so the group structure is symbolically raised to the power of the number of points in the space-time manifold $\mathcal{M}$. The symmetry of the gravitational force is the group of diffeomorphisms on the manifold which is indicated by *diff*($\mathcal{M}$). However, the combination of the diffeomorphism group with the internal gauge groups is not a direct product because diffeomorphisms do not commute with internal gauge transformations. They combine with what is known as a semi-direct product indicated by |*X*. The known symmetry of the forces of nature is therefore:

$$G(\mathcal{M}) = diff(\mathcal{M}) \times (SU(3)^{\mathcal{M}} \times SU(2)^{\mathcal{M}} \times U(1)^{\mathcal{M}})$$

There is plenty of good reason to think that this is not the full story. This group will be the residual subgroup of some larger one which is only manifest in circumstances where very high energies are involved, such as the big bang. Both general relativity and quantum mechanics are full of symmetry so it would be natural to imagine that a unified theory of quantum gravity would combine those symmetries into a larger one. String theory certainly seems to have many forms of symmetry which have been explored mathematically. There is evidence within string theory that it contains a huge symmetry which has not yet been revealed. Whether or not string theory is the final answer, it seems that there is some *universal symmetry* in nature that has yet to be found. It will be a symmetry which includes the gauge symmetries and perhaps also others such as the symmetry of identical particles and electromagnetic duality. The existence of this symmetry is a big clue to the nature of the laws of physics and may provide the best hope of discovering them if experiments are not capable of supplying much more empirical data.

What will the universal symmetry look like? The mathematical classification of groups is incomplete. Finite simple groups have been classified and so have semi-simple Lie groups, but infinite-dimensional groups appear in string theory and these are so far beyond classification. Furthermore, there are new types of symmetry such as supersymmetry and quantum groups which are generalisations of classical symmetries. These symmetries are algebraic constructions which preserve an abstract form of invariance. They turn up in several different approaches to quantum gravity including string theory so they are undoubtedly important. This may be because of their importance in understanding topology. At the moment we do not even know what should be regarded as the most general definition of symmetry let alone having a classification scheme.

Particle Permutations

The importance of the symmetry in a system of identical particles is often overlooked. The symmetry group is the permutation group acting to exchange particles of the same type. The reason why this symmetry is not considered to be as important as gauge symmetry lies in the relationship between classical and quantum physics. There is an automatic scheme which allows a classical system of field equations derived from a principle of least action to be quantised. This can be done either through Dirac's canonical quantisation or Feynman's path integral. The two are formally equivalent. In modern quantum field theory a classical field theory is quantised. Particles appear as a consequence of this process. Gauge symmetry is a symmetry of the classical field which is preserved in the process of quantisation. The symmetry between identical particles, however, does not exist in the classical theory. It appears along with the particles during the process of quantisation. Hence it is a different sort of symmetry.

But the matter cannot simply be left there. In a non-relativistic approximation of atomic physics it is possible to understand the quantum mechanics of atoms by treating them first of all as a system of classical particles. The system is quantised in the usual way and the result is the Schrödinger wave equation for the atom. This is known as first quantisation because it was discovered before the second quantisation of the Schrödinger wave equation which became a

part of quantum field theory. In the first quantisation we have gone from a classical particle picture to a field theory and the symmetry between particles existed as a classical symmetry.

This observation suggests that the relationship between classical and quantum systems is not so clear as it is often portrayed and that the permutation group could also be a part of the same universal symmetry as gauge invariance. This claim is now supported by string theory which appears to have a mysterious duality between classical and quantum aspects. A further clue may be that the algebra of fermionic creation and annihilation operators generate a supersymmetry which includes the permutation of identical particles. This opens the door to a unification of particle permutation symmetry and gauge symmetry.

Event symmetry

Even now we can make some guesses. The universal symmetry must be fundamental to the laws of physics. When the right symmetry is known the laws of physics might be fully determined by the constraints imposed by invariance under the action of the symmetry. Surely it should be some unique fundamental mathematical structure, but G(M), the symmetry group we have so far, is dependent on the topology of the space-time manifold M. Should we expect the topology of space-time to be fixed by the laws of physics? There are many different topologies which space-time could have and it would seem too arbitrary to make the choice at so fundamental a level. This poses quite a puzzle.

There are two possible solutions that I know of. The first is the *principle of event symmetry* which is the central theme of this book. It says that we must simply forget the topology of space-time at the most fundamental level and regard the space-time manifold as just a set of discrete space-time events. The diffeomorphism group of any manifold is a subgroup of the symmetric group of permutations on the set of points in the manifold. The internal gauge symmetries also fall into this pattern. This solution to the puzzle generates many new puzzles and in later chapters I will describe them and start to resolve them.

The second solution to this puzzle is to generalise symmetry using the mathematical theory of categories. A category can describe mappings between different topologies and a group is a special case of a category. If the concept of symmetry is extended further to include more general categories it should be possible to incorporate different topologies in the same categorical structure. How should we interpret these two solutions to a difficult problem when at first one solution seemed difficult to find? Is only one right, or are they both different aspects of the same thing?

There seems little doubt that there is much to be learnt in both mathematics and physics from the hunt for better symmetry. The intriguing idea is that there is some special algebraic structure which will unify a whole host of subjects through symmetry, as well as being at the root of the fundamental laws of physics.

In a Grain of Sand

Discrete Matter

At a seaport in the Aegean around the year 500BC the philosopher Democritus pondered the idea that matter was made of indivisible units separated by void. He had been handed the idea by his mentor Leucippus who had in turn heard about it from the Ionian philosopher Anaxagoras. Was it a remarkable piece of insight or just a lucky guess? At the time there was certainly no compelling evidence for such a hypothesis. Perhaps they were inspired by the coarseness of natural materials like sand and stone. The insight of Anaxagoras went far beyond such observations and his theories of cosmological origins were just as uncanny. There is no accounting for the similarity of these ideas to the modern view. With such bold claims Anaxagoras had become one of the first heretics. He was punished for his impiety and his books were burnt.

Democritus extended the atomic concept as far as it could go, claiming that not just matter, but everything else from colour to the human soul must also consist of atoms. These atoms were indivisible but had different shapes and could combine in a variety of ways to form the substances of the world. He saw creation as the natural consequence of the ceaseless whirling motion of atoms in space. Atoms would collide and spin, forming larger aggregations of matter.

These ideas were soon rivalled by the very different philosophies of Aristotle from the school of Plato, who believed that matter was infinitely divisible and that nature was constructed from perfect symmetry and geometry. According to Empedocles substance was composed of four continuous elements; Earth, Air, Fire and Water. Only with the Islamic Caliphates who studied the earlier Greek philosophers, did the atomistic theory hold out during the middle ages. Al-Razi of Persia is credited with an atomistic revival in the ninth century but Aristotle's physics remained the dominant doctrine in European philosophy until the seventeenth century.

In the 1660s Robert Boyle, a careful chemist and philosopher proposed a corpuscular theory of matter to explain behaviour of gases such as diffusion. According to Boyle there was only one fundamental element, all corpuscles would be identical. Different substances would be constructed by combining the corpuscles in different ways. The theory was based as much on the alchemist's belief in the existence of a philosopher's stone which could turn lead into gold, as it is was on empirical evidence. Newton built on the corpuscular theory. He saw the corpuscles as units of mass and introduced the laws of mechanics to explain their motion.

In 1808 the atomic theory was again resurrected by a school teacher and amateur scientist by the name of John Dalton. He discovered a law of partial pressures of gases which revealed how gases of equal volume contribute pressures in nearly integer ratios. He concluded that these were ratios of atomic weights which were a characteristic of indivisible atoms. This would also explain chemical composition and the nature of the chemical elements. Amedeo Avogadro developed the molecular theory and his law that all gases at the same temperature, pressure and volume contain the same number of molecules even though their weights are different. By the mid nineteenth century the number of molecules in a volume of gas could be measured. Maxwell

and Boltzmann went on to explain the laws of thermodynamics through the statistical physics of molecular motion. The atomic theory was having unprecedented success in explaining a wide variety of physical phenomena.

Despite this indirect evidence, positivists led by Ernst Mach remained sceptical about the kinetic theory. They argued that since atoms could not be directly observed they are no more than metaphysical constructs with no basis in reality. The pressure of such disputes was too much for Boltzmann who took his own life in 1906. Ironically, Einstein had provided what would transpire to be the clinching evidence for atoms just the previous year. In the early eighteenth century, a biologist Robert Brown had observed random motion of particles suspended in gases. Einstein explained that this Brownian motion could be seen as direct experimental evidence of molecules which were jostling the particles with their own movements. In 1956 the field ion microscope made it possible to form images of individual atoms for the first time.

How far has modern physics gone towards the ideal of Democritus that everything should be composed of discrete units?

The story of light parallels that of matter. The Greeks saw an atomistic theory of light as the explanation of light rays. In the Arabic world of the middle ages Al-hazen used a ballistic theory of light to explain reflection. Newton extended Boyle's corpuscular theory to light even though such a supposition had no empirical foundation at that time. Everything he had observed and much more was later explained by Maxwell's theory of Electromagnetism in terms of waves in continuous fields. It was Planck's Law and the photoelectric effect which later upset the continuous theory. These phenomena could only be explained in terms of light quanta. Today we can detect the impact of individual photons on CCD cameras even after they have travelled across most of the observable universe from the earliest moments of galaxy formation.

Those who resisted the particle concepts had, nevertheless, some good sense. Light and matter, it turns out, are both particle and wave at the same time. This paradox is explained mathematically as a consequence of quantum field theory but the interpretation remains unintuitive and mysterious.

As it turned out, the atomic theory of Dalton was a long way short of the end of the road for divisibility. The atom was split and broken down into its constituent particles, and they were in turn further divided. The way we now describe the composition of matter is no longer so simple. When a neutron is observed to decay spontaneously into a proton, neutron, electron and neutrino we do not suppose that those four particles were parts of the neutron which broke apart. Particles can transform and interact in a way which is not simply division and recombination of immutable parts. Physicists continue their journey into the heart of matter, and the final picture has not yet been seen.

Unification

Since Newton set the foundations of mechanics, the major leaps forward have come mostly in the form of unification of two or more previously unrelated concepts. Newton took the first leap himself when he achieved the unification of celestial and terrestrial mechanics demanded by

Galileo. The Newtonian theory of gravity and dynamics could explain both the fall of an apple to Earth and the motion of moons around Jupiter which Galileo had seen in 1609.

Two hundred years after Newton, James Clerk Maxwell unified electricity, magnetism and light into one theory of electromagnetism. This unification was the result of a series of experiments starting in 1820 when Hans Christian Oersted observed that an electric current deflected a compass and Andre Ampere measured the corresponding reaction force on a current in a magnetic field. Above all it was Michael Faraday who appreciated the significance of these results and devised the experiments which would unveil the unity of nature. He showed that a moving magnet could induce a current in a wire and also noticed that a magnetic field could change the polarisation of light passing through a medium. Faraday is regarded as possibly the greatest experimental physicist who ever lived and he proposed the idea of force lines but he never used equations to describe his theories. It was only when Maxwell applied mathematics to the problem that the full power of electromagnetic unification was realised.

The atomic theory was the other important unification step of the nineteenth century. Prior to 1808 chemistry was little more than a catalogue of chemicals and their reactions, although the distinction between elements and compounds had been recognised by Antoine Lavoisier in 1786. The molecular theory was also already part of the kinetic theory of gases when John Dalton proposed that molecules were composed of immutable atoms. By 1869 Dmitri Mendeleyev had laid out the periodic table of the elements in order of atomic weights. By the end of the nineteenth century most everyday observations could be accounted for in terms of well-known physics, and some scientists thought that little remained to be understood. They failed to see the lack of unity which remained in their theories. Mass, energy, space, time, charge, the ether and atoms were the basic constituents whose behaviour followed the laws of mechanics, electromagnetics, gravity, chemistry, electricity and thermodynamics. Other sciences such as biology and astronomy could have been regarded as reducible to these terms but the case for vitality in biology still held sway and astronomy was still a realm apart.

Even then there were other new phenomena, and unexplained enigmas were appearing: By 1900 the electron, X-rays and nuclear radiation had been discovered. Experiments had failed to detect the ether and electromagnetism and thermodynamics could not explain black body radiation. The spectral lines in light already seen by Fraunhofer in 1814, the anomalous perihelion shift of Mercury discovered by Le Verrier in 1859 and the photoelectric effect of Hertz in 1887, were all indications of future revolutions. That is easy for us to see now, but at the turn of the century these things might just have easily been accounted for by making small adjustments to known physics. Many physicists were unprepared for what was to come, but not all. At the dawn of the new century Henri Poincaré wrote that there was a whole new world of which none had expected the existence but that further progress would show how these complete the general unity.

Our greatest lesson of the twentieth century is what Poincaré foresaw, that the universe is governed by a profound unity of physical law. The revelation began with the special relativity of Poincaré and Einstein which Minkowski recognised as a unification of space and time into a single space-time geometry. Mass and energy were then also seen as equivalent, or at least interchangeable. In the same decade the Planck-Einstein theory of light quanta brought together electromagnetics and thermodynamics. Then Einstein unified space-time and gravity into one

theory of general relativity and the atomic theory was reduced to quantum mechanics by Bohr, Heisenberg, Schrödinger and others. The quantum theory also produced an unexpected unification of particles and waves. Later, when Dirac brought together special relativity and quantum mechanics he predicted anti-matter particles which were found shortly after. At the same time as all this unification, new things like the nuclear forces, new particles, superfluidity, and quantum spin were being found but they were all part of the new physics. The total number of fundamental concepts needed to account for nature had diminished drastically.

By the end of the first half of the century the theory of quantum electrodynamics was complete. The world was then recovering from the second world war. Physicists had served their part, for better or for worse, by developing radar and the atomic bomb. No doubt it was by way of repayment, or the hope of further military spin-offs, that they were granted funds to build the large accelerators which were to dominate the discoveries in physics of the following decades. Suddenly there was a new wealth of particles and properties to explain. In 1960 physics was a messy catalogue of particle properties, but the lesson had already been learnt and the search for unity prevailed again. Yang-Mills gauge theories were the key to understanding the forces. By the mid-seventies the quark theory, quantum chromodynamics and the electro-weak force were part of a standard model of particle physics.

At the end of the twentieth century physics is able to explain much more than everyday observations. It can explain just about every fundamental observation that we have been capable of making up to now, from the laboratory to the cosmos. The last quarter of the century has been a tough time for experimenters. They were impotent in their search for new phenomena and could do no more than verify the standard model in ever greater detail. That is not to say that experiments made no contribution to knowledge since the mid-seventies. While the standard model has been verified, many new theories which were advanced have been ruled out through negative results, allowing the theorists to concentrate their efforts on those which remain.

But the main impetus which has been pushing forward the front of physics over the last twenty years has come from a belief in complete unity. According to conventional wisdom among physicists, the process of unification will continue until all physics is unified into one neat and tidy theory. There is no a priori reason to be so sure that this must happen. It is quite possible that physicists will always be discovering new forces, or finding new layers of structure in particles, without ever arriving at a final theory. It is quite simply the unified nature of the laws of physics as we currently know them, the lesson of the twentieth century, that inspires the belief that we are getting closer to that end.

After physicists discovered the atom, they went on to discover that it was composed of electrons and a nucleus, then that the nucleus was composed of protons and neutrons, then that the protons and neutrons were composed of quarks. Should we expect to discover that quarks and electrons are made of smaller particles? This is possible but there are reasons to suppose not. Firstly there are far fewer particles in the standard model than there ever were at higher levels. Secondly, their interactions are described by a clean set of gauge bosons through renormalisable field theories. Composite interactions, such as pion exchange, do not take such a tidy form. These reasons in themselves are not quite enough to rule out the possibility that quarks, electrons and gauge bosons are composite but they reduce the number of ways such a theory could be constructed. In

fact all viable theories of this type which have been proposed are now all but ruled out by experiment. There may be a further layer of structure but it is likely to be different. It is more common now for theorists to look for ways that different elementary particles can be seen as different states of the same type of object. The most popular candidate for the ultimate theory of this type is *superstring theory*, in which all particles are just different vibration modes of very small loops of string.

Physicists construct particle accelerators which are like giant microscopes. The higher the energy they can produce, the smaller the wavelength of the colliding particles and the smaller the distance scale they probe. In this way, physicists can see the quarks inside protons, not through direct pictures but through scattering data. They have already examined quarks at a scale of 10^{-19} metres and they still look pointlike. Such resolution is impressive given that atoms have a typical size of 10^{-10} metres and nucleons have structure on the scale of 10^{-15} metres. Suppose you have a cannon ball about 10 centimetres in diameter in your hand. Imagine you scale it up until it is as big as the Earth (a factor of 10^8). The bumps and scratches on the surface would have become mountain ranges and great ravines. As you walked over the surface you could look down at the ground and would see that it is made of atoms scaled up to the size of marbles 1 or 2 centimetres across. Each atom would be a hazy cloud of electrons around the tiny nucleus which appears as just a point in the centre.

Now scale one of those atoms again by the same factor. It would now be about the size of Pluto. The nucleus will have expanded to a huge jumble of nucleons, each the size of a house but appearing as a fuzz of quarks. If you could now stop one of the electrons or quarks in the atom and look at it closely with the naked eye, you would be seeing it on the scale which today's biggest accelerators have probed, so we know that it would still look like a point. Despite this impressive achievement we have only gone half way towards the smallest scale. If the superstring theory is right and electrons and quarks have no structure until you see them on the string scale, it will be necessary to scale them up twice again by the same factor before they become visible as little loops of string. The atom, now scaled up by a factor of 10^{32}, would then be about a million light years across. The scale of inner space is as impressive as the scale of outer space.

In the first decade of the 21st century new accelerator experiments at CERN will probe beyond the electro-weak scale. There is some optimism that new physics will be found but nothing is certain. After that, experimental particle physics may become more difficult. There is a limit to how much funding for larger accelerators can be found, even with many nations clubbing together. Perhaps other observational clues will come from cosmic rays and big bang cosmology. Perhaps experimenters will get lucky and find a better way to accelerate particles. If they could have a wish granted it might be the discovery of a stable charged elementary particle with a 1000 times the mass of the proton. It could then be produced in quantity and accelerated to much higher energy. Alternatively they might ask for a new form of stable matter which can be built into superdense substances. Even with such luck there is a long way to go before reaching the scale of grand unification, but ingenuity and the unexpected should never by underestimated in experimental physics.

In any case, that empirical route is just the low road, and there is an alternative high road which the theorists can take while the lower remains blocked. Progress may come from the mathematical search for greater unity. The electromagnetic and nuclear forces are now only partially unified. They still have separate coupling strengths in the standard model. There are also three generations of quark-lepton matter quadruplets and that need to be explained. Perhaps there should be unification of the gauge bosons of the force fields and the fermionic matter fields. Above all gravity must be brought together with the other forces. That will require a unification of general relativity and quantum mechanics. By searching through the mathematical possibilities for new forms of unity, physicists may be able to bypass the huge gap in energy between current day experiments and the higher unification scales. Ironically, as a result of such endeavours, we may already know more about physics at distance scales of 10^{-36} metres than we do at scales of 10^{-24} metres.

Quantum Gravity

The search for a theory of quantum gravity is reputed to be one of the most difficult puzzles of science. In practical terms it is probably of no direct relevance in our lives and may even be impossible to verify by experiment. But to physicists it is their holy grail. It may enable them to complete the unification of all fundamental laws of physics.

The problem which they face is to put together general relativity and quantum mechanics into one self consistent theory. The difficulty is that the two parts seem to be incompatible, both in concept and in practice. A direct approach, attempting to combine general relativity and quantum mechanics, while ignoring conceptual differences, leads to a meaningless quantum field theory with unmanageable divergences. Conceptually, it is the nature of space and time, seen differently from each side, which present the fundamental differences. There have, in fact, been many attempts to create a theory of quantum gravity. From some of these it appears that the combination of general relativity and the quantum theory will also be a unification of much more. It will probably require all four forces and the matter fields to be brought together. It may also require a deeper unification of space-time and matter. If this is true, a complete theory of quantum gravity will then be the realisation of Descartes's visionary dream. It will be the final step on the long road of unification which he foresaw.

Einstein's Geometrodynamics

General relativity is Einstein's monumental theory of gravity and it is rightly seen as the most elegant physical theory we know. It was partially anticipated by the mathematician Bernhard Riemann who developed a large part of the mathematics of curved surfaces. In 1854 he gave a lecture "on the hypothesis which underlie geometry" and speculated that physical objects may be a consequence of non-Euclidean structures in space on both large and small length scales.

Einstein's special relativity was the culmination in 1905 of the work of many physicists such as Lorentz and Poincaré. Mechanics and electrodynamics were placed in a new kinematic framework in which space and time were no longer absolute. When Minkowski described a geometric formulation of special relativity in which space and time were combined into a single

space-time continuum, at first Einstein did not like it. Soon he changed his mind as he recognised that this geometric way to understand relativity was more easy to generalise than his original mechanical approach. He wanted to extend relativity to include gravity. His genius is demonstrated by the way in which he was able to perceive the correct principles which were needed and follow their consequences to the right conclusion.

General relativity is based on two fundamental principles: *The principle of relativity* which states that all basic laws of physics should take a form which is independent of any reference frame, and *The principle of equivalence* which states that it is impossible to distinguish (locally) the effects of gravity from the effects of being in an accelerated frame of reference.

Einstein struggled with the consequences of these principles for several years, constructing many thought experiments to try to understand what they meant. He had already recognised the value of the equivalence principle in 1907. Finally he learnt about Riemann's mathematics of curved geometry and in 1912 realised that a new theory could be constructed in which the force of gravity was a consequence of the curvature of space-time.

In constructing that theory, Einstein was not significantly influenced by any experimental result which was at odds with the Newtonian theory of gravity. He knew of the anomalous precession of the perihelion of Mercury and hoped that a new theory might explain it but there is no route to develop general relativity directly from such an observation. He also knew, however, that Newtonian gravity was inconsistent with his theory of special relativity and he knew there must be a more complete self consistent theory. A similar inconsistency now exists between quantum mechanics and general relativity and, even though no experimental result is known to violate either theory, physicists now seek a more complete theory in the same spirit.

By 1915 Einstein's work was complete. The force of gravity was now a consequence of geometrodynamics; the dynamic geometry of space-time. The equations for the gravitational field are complicated but are an almost unique consequence of the relativity principles which require that they must be independent of any co-ordinate system. Einstein calculated the motion of Mercury in his theory and found that the relativistic corrections to the Newtonian prediction correctly accounted for its anomalous motion. He then predicted that star-light passing the sun would be deflected by twice the Newtonian amount. Arthur Eddington measured this deflection on a South American expedition to observe a solar eclipse in 1919. When he announced to the world that the result agreed with the prediction of general relativity, Einstein became a household name synonymous with "genius".

In the decades that have followed Einstein's discovery, a number of other experimental confirmations of general relativity have been found, and geometrodynamics has become the cornerstone of cosmology. There still remains a possibility that it may not be accurate on very large scales, or under very strong gravitational forces. There are, however, no alternative theories with the force of elegance found in general relativity. The fortuitous discovery by Hulse and Taylor of a binary pulsar in 1974, made it possible to test and verify general relativistic effects to very high precision. Still, the theory is sure to break down finally under the conditions which are believed to have existed at the big bang where quantum gravity effects were important.

One of the most spectacular predictions of general relativity is that a dying star of sufficient mass will collapse under its gravitational weight into an object so compressed that not even light can escape its pull. Such collapsed objects were designated "black holes" by John Wheeler in 1967 and the picturesque term has stuck. Astronomers now have a growing list of celestial objects which they believe are black holes because of their apparent high density and because of evidence of matter apparently falling silently through the event-horizon. The accuracy of Einstein's theory may be stringently tested again in the near future when gravitational wave observatories such as LIGO come on-line to observe such catastrophic events as the collisions between black holes.

The Planck Scale

The Quantum theory was founded before Einstein began his theory of relativity but it took much longer to be completed and understood. Max Planck's observations of quanta in the spectrum of black body radiation first produced signs that the classical theories of mechanics were due for major revisions.

Unlike general relativity which was essentially the work of one man, the quantum theory required major contributions from Bohr, Einstein, Heisenberg, Schrödinger, Dirac and many others, before a complete theory of quantum electrodynamics was formulated. In practical terms, the consequences of the theory are more far reaching than those of general relativity. Applications such as transistors and lasers are now an integral part of our lives and, in addition, the quantum theory allowed us to understand chemical reactions and many other phenomena.

Despite such spectacular success, confirmed in ever more detail in high energy accelerator experiments, the quantum theory is still criticised by some physicists who feel that its indeterministic nature and its dependency on the role of observer suggest an incompleteness. For others the major task is to combine general relativity and quantum mechanics. Opinions differ as to how much revision of quantum mechanics is required to achieve it. Perhaps quantum mechanics is more fundamental than general relativity or perhaps it is the other way round. The answers lie in the realms of ultra-high energy physics, well beyond what can be attained experimentally with known techniques. This leaves us with theory as the only means of moving ahead for the time being at least.

At first thought it might seem ridiculous to suppose that we can invent valid theories about physics at high energies before doing experiments. However, theorists have already demonstrated a remarkable facility for doing just that. The standard model of particle physics was devised in the 1960s by theoretical physicists. It described the physics of energies several orders of magnitude beyond what had been observed before. Experimentalists have spent the last three decades verifying it. The reason for this success is that physicists recognised the importance of certain types of symmetry and self-consistency conditions in quantum field theory which led to an almost unique model for physics up to the electro-weak unification energy scale, with only a few parameters such as particle masses to be determined.

The situation now is a little different. Experimentalists are about to enter a new scale of energies and theorists do not have a single unique theory about what can be expected there. They do have

some ideas, in particular it is hoped that supersymmetry may be observed, but we will have to wait and see.

Despite these unknowns there are other more general arguments which tell us things about what to expect at higher energies. When Planck initiated the quantum theory he recognised the significance of fundamental constants in physics, especially the speed of light (known as c), Boltzmann's constant (known as k) and Planck constant (known as h). Scientists and engineers have invented a number of systems of units for measuring lengths, masses, temperature and time, but they are entirely arbitrary and must be agreed by international convention. Planck realised that there should be a natural set of units in which the laws of physics take a simpler form. The most fundamental constants, such as c, k and h would simply be equal to one unit in that system.

If one other suitable fundamental constant could be selected, then the units for measuring mass, length and time would be determined. Planck decided that Newton's gravitational constant (known as G) would be a good choice. Actually there were not many other constants, such as particle masses known at that time, otherwise his choice might have been more difficult. By combining c, h, k and G, Planck defined a system of units now known as the Planck scale. In 1899 he wrote that it is possible to give units for length, mass, time and temperature which retain their meaning for all time and all cultures, even extraterrestrial ones. He calculated that the Planck unit of length is very small, about 10^{-35} metres. To build an accelerator which could see down to such lengths would require energies about 10^{16} times larger than those currently available. The Planck scale is not very good for practical engineering, partly because the units are mostly either too small or too big compared with everyday quantities. More importantly, it is not possible to make accurate enough measurements using Planck units because it would be necessary to measure the mass of an object by measuring its gravitational pull on other objects. However, Planck units are very convenient for physicists studying quantum gravity because the values of the constants c, h, k and G are equal to one and can be left out of the equations.

Physicists have since sought to understand what the Planck scale of units signifies. One possibility is that at the Planck scale all the four forces of nature, including gravity, are unified. Physicists who specialise in general relativity have a different idea. In 1955 John Wheeler argued that when you combine general relativity and quantum mechanics you will have a theory in which the geometry of space-time is subject to quantum fluctuations. He computed that these fluctuations would become significant if you could look at space-time on length scales as small as the Planck length. Sometimes physicists talk about a space-time foam at this scale but we do not yet know what it really means. For that we will need the theory of quantum gravity.

Without really knowing too much for certain, physicists guess that at the Planck scale all forces of nature are unified *and* quantum gravity is significant. It is at the Planck scale that they expect to find the final and completely unified theory of the fundamental laws of physics.

It seems clear that to understand quantum gravity we must understand the structure of space-time at the Planck length scale. In the theory of general relativity space-time is described as a smooth continuous manifold but we cannot be sure that this is correct for very small lengths and times. We could compare general relativity with the equations of fluid dynamics for water. They describe a continuous fluid with smooth flows in a way which agrees very well with experiment.

Yet we know that at atomic scales, water is something very different and must be understood in terms of forces between molecules whose nature is completely hidden in the ordinary world. If space-time also has a complicated structure at the tiny Planck length, way beyond the reach of any conceivable accelerator, can we possibly hope to discover what it is?

If you asked a group of mathematicians to look for theories which could explain the fluid dynamics of water, without them knowing anything about atoms and chemistry, then they would probably succeed in devising a whole host of mathematical models which work. All those models would probably be very different, limited only by the imagination of the mathematicians. None of them would correspond to the correct description of water molecules and their interactions. The same might be true of quantum gravity in which case there would be little hope of finding out how it worked without further empirical information. Nevertheless, the task of putting together general relativity and quantum mechanics together into one self consistent theory has not produced a whole host of different and incompatible theories. The clever ideas which have been developed have things in common. It is quite possible that all the ideas are partially correct and are aspects of one underlying theory which is within our grasp. It is time now to look at some of those ideas.

The Best Attempts

The physics of the electromagnetic and nuclear forces is successfully described by quantum field theories which are constructed by applying a quantisation process to the classical field equations. This is not a straight forward matter. Troublesome infinite quantities appear in the calculation of physical quantities. A messy renormalisation must be applied to make the answers finite. Although it cannot be said for sure that this defines a mathematically rigorous theory, it does at least provide an apparently consistent means of calculation and prediction. It is rather fortuitous that this works. Only a small class of field theories can be renormalised in this way and the ones which describe the known particles are the right sort.

In this scheme, particles are a consequence of the field quantisation and are seen as less fundamental than the field waves out of which they appear. The particles carry spin in integer or half-integer multiples of Planck's constant. They may be spin zero, spin a half or spin one according to the type of field which is quantised. All the known fundamental fermions such as quarks and electrons are spin half. The gauge bosons which mediate the electromagnetic and nuclear forces are spin one. There are also thought to be Higgs particles which have spin zero but they have not yet been found in experiments. The interactions between these particles can be most easily worked out using a perturbation theory. The clearest form of this is a diagrammatic system which was worked out by Richard Feynman.

In principle it should be possible to apply the same quantisation methods to the gravitational field. It is necessary to first construct a system of non-interacting *graviton* particles which represent a zero order approximation to quantised gravitational waves in flat space-time. These hypothetical gravitons must be massless particles carrying spin two, because of the form of the gravitational field in general relativity. The next step is to describe the interactions of these gravitons using the perturbation theory. Feynman himself spent a significant amount of time trying to get it to work, but for gravity this simply cannot be done in the way that works for the

Yang-Mills gauge fields. The calculations are plagued by infinite quantities which cannot be renormalised. The resulting quantum field theory is incapable of giving any useful result.

Because quantum gravity is an attempt to combine two different fields of physics, there are two distinct groups of physicists involved. These two groups form a different interpretation of the failure of the direct attack. The relativists say that it is because gravity cannot be treated perturbatively. To try to do so destroys the basic principles on which relativity was founded. It is, for them, no surprise that this should not work. Perturbation theory requires that you define a fixed approximate background and treat the full physics as if it was a perturbative deviation from there. The fixed background breaks the relativistic symmetry of general covariance. On the other hand, particle physicists say that if a field theory is non-renormalisable then it is because it is incomplete. The theory must be modified and new fields might have to be added to cancel divergences, or it may be that the observed fields are approximate composite structure of more fundamental constituent fields.

Supergravity

The first significant progress in the problem of quantum gravity was made by particle physicists. They discovered that a new kind of symmetry called supersymmetry was very important. particles can be classed into two types; *fermions* such as quarks and electrons, and *bosons* such as photons and Higgs particles. Supersymmetry allows the two types to intermix. With supersymmetry we have some hope to unify the matter fields with radiation fields.

Particle physicists discovered that if the symmetry of space-time is extended to include supersymmetry, then it is necessary to supplement the metric field of gravity with other matter fields. Miraculously these fields led to cancellations of many of the divergences in perturbative quantum gravity. This has to be more than coincidence. At first it was thought that such theories of supergravity might be completely renormalisable. After many long calculations this hope faded. A strange thing about supergravity was that it works best in ten or eleven-dimensional space-time. This inspired the revival of an old theory from the 1920s called Kaluza-Klein theory, which suggests that space-time has more dimensions than the four obvious ones. The extra dimensions are not apparent because they are curled up into a small sphere with a circumference as small as the Planck length. This theory provides a means to unify the gauge symmetry of general relativity with the internal gauge symmetries of particle physics.

The next big step taken by particle physicists came along shortly after. Two physicists Michael Green and John Schwarz were looking at a theory which had originally been studied as a theory of the strong nuclear force but which was actually more interesting as a theory of gravity because it included spin-two particles. This was the new beginning of string theory. Combining string theory and supergravity to form superstring theory quickly led to some remarkable discoveries. A few string theories in ten dimensions were perfectly renormalisable and finite. This was exactly what they were looking for.

It seemed once again that the solution was near at hand, but nature does not give up its secrets so easily. The problem now was that there is a huge number of ways to apply Kaluza-Klein theory to the superstring theories. Hence there seem to be a huge number of possible unified theories of

physics. The perturbative formulation of string theory makes it impossible to determine the correct way.

Canonical Quantum Gravity

While particle physicists were making much noise about superstring theory, relativists have been quietly trying to do things differently. Many of them take the view that to do quantum gravity properly you must respect its diffeomorphism symmetry or general covariance. Starting from the old quantisation methods of Dirac it is possible to formally derive the *Wheeler-DeWitt equation* together with a *Hamiltonian constraint equation*, which describe the way in which the quantum state vector should evolve according to this *canonical approach*.

For a long time there seemed little hope of finding any solutions to the Wheeler-DeWitt equation. Then in 1986 Abhay Ashtekar found a way to reformulate Einstein's equations of gravity in terms of new variables. Soon afterwards a way was discovered to find solutions to the equations. This is now known as the loop representation of quantum gravity. Mathematicians were surprised to learn that knot theory was an important part of the concept.

The results from the canonical approach seem very different from those of string theory. There is no need for higher dimensions or extra fields to cancel divergences. Relativists point to the fact that a number of field theories which appear to be unrenormalisable have now been quantised exactly. There is no need to insist on a renormalisable theory of quantum gravity. On the other hand, the canonical approach still has some technical problems to resolve. It could yet turn out that the theory can only be made fully consistent by including supersymmetry.

As well as their differences, the two approaches have some striking similarities. In both cases they are trying to be understood in terms of symmetries based on loop like structures. It seems quite plausible that they are both aspects of one underlying theory. Other mathematical topics are common features of both, such as knot theory and topology. Indeed there is now a successful formulation of quantum gravity in three-dimensional space-time which can be regarded as either a loop representation or a string theory. A small number of physicists such as Lee Smolin are looking for a more general common theory uniting the two approaches.

Non-Commutative Geometry

A technique which introduces such a minimum length into physics by quantising space-time was attempted by Hartland Snyder in 1947. In analogy to the non-commuting operators of position and momentum in quantum mechanics, Snyder introduced non-commutative operators for space-time co-ordinates. These operators have a discrete spectrum and so lead to a discrete interpretation of space-time. The model was Lorentz invariant but failed to preserve translation invariance so no sensible physical theory came of it. Similar methods have been tried by others since and although no complete theory has come of these ideas there has been a recent upsurge of renewed interest in quantised space-time, now re-examined in the light of quantum groups and non-commutative geometry. The traditional definition of a field in physics is a function from the co-ordinates of space-time events to field variables which may be real, complex or whatever.

Fields can be multiplied together event by event. Differential operators which act on the fields are defined using the continuous nature of the space-time co-ordinates. The equations of evolution for the fields are specified using these operations which ensure their causal and local nature. In the new approach fields are defined by their algebraic properties and space-time co-ordinates are ignored. Fields are any kind of mathematical structure which can be multiplied together and which can be operated on by some operators which obey rules analogous to those of differentiation, such as Leibniz rule for products.

If enough algebraic rules are applied the new type of fields will be equivalent to the old traditional definition for a space-time with some kind of topology. If the rules are allowed to differ then a more general structure than space-time is defined. The rule which is the most likely candidate for change is that fields should multiply together commutatively. This is analogous to the step taken in going from classical to quantum physics where observables are replaced by non-commuting observables. Now the same idea is used to define non-commutative geometry.

The technique can also be applied successfully to groups by generalising the algebraic properties of a function from the group to the real numbers. The result in this case is the discovery of quantum groups which have all the important algebraic properties of functions on a group except commutivity. Space-time structure can be derived from its group of symmetries in a way which can be generalised to quantum groups. The result is various forms of quantum space-time. The hope of this program is that general relativity and quantum field theories can also be generalised and that the results will not suffer from the infinite divergences which are the primary obstacle to a theory of quantum gravity.

Black Hole Thermodynamics

Although there is no direct empirical input into quantum gravity, physicists hope to accomplish unification by working on the requirement that there must exist a mathematically self consistent theory which accounts for both general relativity and quantum mechanics as they are separately confirmed experimentally. It is important to stress the point that no complete theory satisfying this requirement has yet been found. If just one theory could be constructed then it would have a good chance of being correct.

Because of the stringent constraints that self consistency enforces, it is possible to construct thought experiments which provide strong hints about the properties a theory of quantum gravity has to have. There are two physical regimes in which quantum gravity is likely to have significant effects. In the conditions which existed during the first Planck unit of time in our universe, matter was so dense and hot that unification of gravity and other forces would have been reached. Likewise, a small black hole whose mass corresponds to the Planck unit of mass also provides a thought laboratory for quantum gravity.

Black holes have the classical property that the surface area of their event horizons must always increase. This is suggestively similar to the law that entropy must increase, and in 1972 it led Jacob Bekenstein to conjecture that the area of the event horizon of a black hole is in fact proportional to its entropy. If this is the case then a black hole would have to have a temperature and obey the laws of thermodynamics. Stephen Hawking investigated the effects of quantum

mechanics near a black hole using semi-classical approximations to quantum gravity. Against his own expectations he discovered that black holes must emit thermal radiation in a way consistent with the black hole entropy law of Bekenstein.

This forces us to conclude that black holes can emit particles and eventually evaporate. For astronomical sized black holes the temperature of the radiation is minuscule and certainly beyond detection, but for small black holes the temperature increases until they explode in one final blast. Hawking realised that this creates a difficult paradox which would surely tell us a great deal about the nature of quantum gravity if we could understand it.

The entropy of a system can be related to the amount of information required to describe it. When objects are thrown into a black hole the information they contain is hidden from outside view because no message can return from inside. Now if the black hole evaporates, this information will be *lost* in contradiction to the laws of thermodynamics. This is known as the *black hole information loss paradox*.

A number of ways on which this paradox could be resolved have been proposed. The main ones are:

- The lost information escapes to another universe
- The final stage of black hole evaporation halts leaving a remnant particle which holds the information.
- There are strict limits on the amount of information held within any region of space to ensure that the information which enters a black hole cannot exceed the amount represented by its entropy.
- Something else happens which is so strange we cannot bring ourselves to think of it.

The first solution would imply a breakdown of quantum coherence. We would have to completely change the laws of quantum mechanics to cope with this situation. The second case is not quite so bad but it does seem to imply that small black holes must have an infinite number of quantum numbers which would mean their rate of production during the big bang would have been divergent.

Assuming that something has not been missed out, which is a big assumption, we must conclude that the amount of entropy which can be held within a region of space is limited by the area of a surface surrounding it. This is certainly counterintuitive because you would imagine that you could write information on bits of paper and the amount you could cram in would be limited by the volume only. This is false because any attempt to do that would eventually cause a black hole to form. Note that this rule does not force us to conclude that the universe must be finite because there is a hidden assumption that the region of space is static.

If the amount of information is limited then the number of physical degrees of freedom in a field theory of quantum gravity must also be limited. Inspired by this observation, Gerard 't Hooft, Leonard Susskind and others have proposed that the laws of physics should be described in terms of a discrete field theory defined on a space-time surface rather than throughout space-time.

They liken the way this might work to that of a hologram which holds a three-dimensional image within its two-dimensional surface.

Rather than being rejected as a crazy idea, this theory has been recognised by many other physicists as being consistent with other ideas in quantum gravity, especially string theory.

If Susskind is right, this solution to the information loss problem may have even stranger consequences. What happens in the case of an observer, Mr. X, who falls into a black hole. From his point of view he will pass through the event horizon without incident and continue to his gruesome fate at the black-hole singularity. Any knowledge and information he carries will stay with him till the end. To an outside observer, Miss Y, the situation must be different. Gravitational time dilation ensures that she will watch Mr. X slow down so much as he approaches the event horizon that he will never cross it. Eventually he will fade from her view but the information he carries must still be accessible. If Miss Y waits long enough the black hole will evaporate and the information will be returned in the radiation. At least it should be in principle even if it is too jumbled to be read in practice.

There is a conceptual difficulty which accompanies this situation. The course of events as witnessed by Mr. X is very different from that seen by Miss Y. If they are ever brought together in a court of law and asked to account for what happened to the information their stories will not be consistent. Mr X will claim he carried it to his cosmic grave where time ended for him but Miss Y will say that it never got past the event horizon and was brought back into the outside universe as the hole evaporated. The judge and jury will be forced to conclude that one of them was lying. This paradox is resolved by the simple fact that the two witnesses never can be brought back together. Presumably this must even be true if the black hole harboured a wormhole through to another universe through which Mr. X could escape his fate.

Susskind has called this the *black hole complementarity principle* in deference to Niels Bohr's complementarity principle of quantum mechanics. Just as there is no conflict between the dual properties of matter as both particle and wave because no observation brings them into contradiction, so too there is no conflict between the contrary observations of Mr. X and Miss Y. The implications of Susskind's principle may be even harder to contemplate than Bohr's. In ordinary quantum mechanics observers who can communicate freely should be able to agree what the probability of future events is. However, if one plans to take a swan dive into a black hole he may not agree on the most likely future events with his partner who plans to rest outside. This removes physics further from the conventional causal paradigms. The full implications may only be understood when we have a complete consistent theory which embraces the new complementarity.

Although there has been considerable progress on the problem of quantising gravity, it seems likely that it will not be possible to complete the solution without some fundamental change in the way we think about space-time. To face the quantum gravity challenge we need new insights and more new principles like those which guided Einstein to the correct theory of gravity.

Is There a Theory of Everything?

This is a good moment to take a pause and look at where we are. If the physics lesson of the twentieth century is that progress comes through unification, then how far can that unification go? It seems likely that it will continue until all fundamental physical laws are unified. There is more than unification of the four fundamental forces. We have also seen how space and time, mass and energy, thermodynamics and gravity and much more have become unified. The final step may lead to a unification of matter and space-time. Will that be the end of physics?

At one point supergravity looked very promising as a theory which might unify all physics. At the time I was a student at Cambridge University where Stephen Hawking was taking up his position as the new Lucasian professor of mathematics. There was great anticipation of his inaugural lecture to take place on 29th April 1980. Even though I made a point of turning up early I found only standing room at the back of the auditorium. It was an exciting talk at which Hawking made some of his most quotable comments. He cautiously predicted that the end of theoretical physics was in sight. The goal might be achieved in the not-too-distant future, perhaps by the end of the century.

But early hopes faded as the perturbative calculations in supergravity became difficult and it seemed less likely that it defined a renormalisable field theory. There were other difficulties such as the problem of fitting in the distinction between left and right which we find in the weak force. Hawking pointed out himself that he was joining a list of physicists who had thought they were near the end. Faraday thought a unification of gravity and electromagnetism would lead to a complete theory but he could not detect any effect linking the two as he had with electricity and magnetism. After the rapid progress in the foundations of quantum mechanics in the 1920s Max Born told a meeting of scientists that physics would be over in six months. Einstein, in his later years, also thought that a unified theory was within reach. Those hopes were premature.

In 1985 The phrase "Theory of Everything" entered the minds of theoretical physicists. It came up in articles written for science magazines such as *New Scientist* and *Science* and later appeared in the title of a number of books. The discovery that set things going was that the heterotic superstring theory is finite in all orders of perturbation theory and has the potential to encompass all the known theories of particle physics and gravity too. In other words it provided potentially a unified theory of all the known underlying laws of physics.

It was not long before scientists from other disciplines and physicists too, started to question the validity of the claim that superstring theory was a theory of everything. For one thing it did not really make any testable predictions, leading some to retort that it was more like a theory of nothing. More to the point, they questioned whether any theory of physics could rightly be called a theory of everything. They were quite right.

The term *Theory of Everything* is a desperately misleading one. Physicists usually try to avoid it but the media apparently cannot help themselves. "Physicists on the verge of finding theory of everything." It makes too good a headline. If physicists find a complete unified set of equations for the laws of physics, then that would be a fantastic discovery. The implications would be enormous, but to call it a theory of everything would be nonsense.

For one thing, it would be necessary to solve the equations to understand anything. No doubt many problems in particle physics could be solved from first principles, perhaps it would be possible to derive the complete spectrum of elementary particles including their relative masses and the coupling constants of the forces which bind them. However, there would certainly be limits to the solvability of the equations. We already find that it is almost impossible to derive the spectrum of hadrons composed of quarks, even though we believe we have an accurate theory of strong interactions. *In principle* any set of well-defined equations can be solved numerically given enough computer power. The whole of nuclear physics and chemistry ought to be possible to calculate from the laws we now have. *In practice* computers are limited and experiments will never be obsolete.

Furthermore, it is not even possible to derive everything *in principle* from the basic laws of physics. Many things in science are determined by historical accident. The foundations of biology fall into this category. The final theory of physics will not tell us how life on Earth originated. The most ardent reductionist would retort that, in principle, it would be possible to derive a list of all possible forms of life from the basic laws of physics. Such justification is weak. No theory of physics is likely to answer all the unsolved problems of mathematics, chemistry, biology, astronomy or medicine.

Finally it must be said that even given a convincing unified theory of physics, it is likely that it would still have the indeterminacy of quantum mechanics. This would mean that no argument could finally lay to rest questions about paranormal, religion, destiny or other such things, and beyond that there are many matters of philosophy and metaphysics which might not be resolved, not to mention an infinite number of mathematical problems.

But string theorists never claimed that their work was applicable to any of these things. Steven Weinberg tried to clarify what it was all about in his 1988 book "Dreams of a Final Theory". Physicists, he argued, are seeking to take the last step of unification on a climb which started as least as far back as Newton. Those steps could lead us towards one "Final Theory" in which all the underlying laws of physics are unified. Weinberg's term "Final Theory" is actually not much better than "Theory of Everything". It suggests, to some, that the theory will mark the end of science and there will be no new theories after. Again, this is not what is meant. Finding the final laws of physics will be like arriving at the summit of the highest mountain. It is a special place from where you can see far, but getting there does not mean you have been everywhere.

In my youth I found time to explore the mountains of Scotland where I lived. Often as you climb one of those rounded peaks, you see ahead what appears to be the top. As you get closer you realise that it is a false summit with a further climb beyond. Sometimes there are several of them before you reach the true summit and at last take in the panoramic view, if the mist and rain have cleared. Approaching the final theory of physics seems to be a very similar experience. There have already been many false summits and again we see another ahead. A mountaineer always knows that there is a final summit and it can be reached if he has the courage to continue. Can physicists know that their summit is there too? Hawking feels that it is. After Cambridge the next time I had the opportunity to hear Hawking lecture was 17 years on at a conference for string theorists.

Hawking had never moved on from supergravity to string theory as other physicists had, until then. His liking for strings appeared to have improved when it was discovered in 1995 that string theories can be unified under a mysterious form of supergravity in 11 dimensions. Hawking must have felt that he had been vindicated in his prediction that supergravity was near the end. With a characteristic touch of humour he told us, "twenty years ago, I said there was a 50/50 chance that we would have a complete picture of the universe in the next twenty years. That is still my estimate today but the 20 years start now."

There are a few who are not so certain. John Taylor in his book "When the clock struck zero" argued that there could be an infinite structure of levels of physical law to find. No-one thinks that there will be a final theory of mathematics and if mathematics is so strongly reflected in physics why should there be a limit to its application? For what my opinion is worth, I too think we really are near the summit.

Is Space-Time Discrete?

Seeking the ultimate indivisible

We have seen how atomic physics and quantum mechanics have reduced matter and light to discrete components. Today history is repeating itself for a third time and now it is space-time which is threatened to be reduced to discrete events. The idea that space or time could be discrete has been a recurring one in the scientific literature of the twentieth century and its origins go back much further. A survey of just a few examples reveals that discrete space-time can actually mean many things and is motivated by a variety of philosophical or theoretical influences. As we shall see, it is only recently that theories of quantum gravity have suggested the true scale at which the structure of space-time breaks up.

It has been apparent since early times that there is something different about the mathematical properties of the real numbers and the quantities of measurement in physics at small scales. Riemann himself remarked on this disparity even as he constructed the formalism which would be used to describe the space-time continuum for the next century of physics in 1876.

In mathematics numbers have unphysical properties like being an exact ratio of two integers. When you measure a distance or time interval you cannot declare the result to be a rational or irrational number no matter how accurate you manage to be. Furthermore it appears that there is a limit to the amount of detail contained in a volume of space. If we look under a powerful microscope at a grain of dust we do not expect to see minuscule universes supporting the complexity of life seen at larger scales. Structure becomes simpler at smaller distances. Surely

there must be some minimum length at which the simplest elements of natural structure are found and surely this must mean space-time is discrete rather than continuous?

This style of argument tends to be persuasive only to those who already believe the hypothesis. It will not make many conversions. After all, the modern formalism of axiomatic mathematics leaves no room for Zeno's paradox. In the fifth century BC the philosopher Parmedies and his disciple Zeno of Elea tried to discredit the senses by posing paradoxes about the divisibility of space-time. In a race between the Archiles and the tortoise, the tortoise was given a head start. To catch him up Archiles must first half the distance between them, then half the remaining distance again. No matter how many times he halves the distance he will not have caught the tortoise. If space and time are infinitely divisible Archiles cannot pass the tortoise according to Zeno. Such thoughts influenced the atomists of ancient Greece, and a more complete philosophy of atomic space and time was developed by the Kalam of Bahgdad from the 9th century.

But axiomatic mathematics has dispelled Zeno's paradox. It is possible to talk about limits and infinity without reaching any mathematical contradiction and it can be proven that the sum of an infinite number of halving intervals is finite. Although some philosophers such as Bertrand Russell persisted with such arguments and developed a detailed and general philosophy of atomism, there are few physicists who would agree that logic and philosophy alone can tell us whether or not space and time are discrete.

However, experimental facts are a different matter and the discovery of quantum theory with its discrete energy levels and the Heisenberg uncertainty principle led physicists to speculate that space-time itself may be discrete as early as the 1930s. In 1936 Einstein expressed the general feeling that the success of the quantum theory points to a purely algebraic method of description of nature and the elimination of continuous function and space-time continuum from physics.

Heisenberg himself noted that the laws of physics must have a fundamental length in addition to Planck's constant and the speed of light, to set the scale of particle masses. At the time it was thought that this length scale would be around 10^{-15}m corresponding to the masses of the heaviest elementary particles known at the time. Searches for non-local effects in high energy particle collisions have now given negative results for scales down to about 10^{-19}m and today the consensus is that it must correspond to the much smaller Planck length at 10^{-35}m.

The belief in some new space-time structure at small length scales was reinforced with the discovery of ultraviolet divergences in Quantum Field Theory. From 1929 it was found that infinite answers appear when you sum up contributions to a physical quantity from waves of ever smaller wavelength. In 1930 Viktor Ambarzunmian and Mitrij Dmitrevich Iwanenko were the first of many physicists to propose that space should be treated as discrete to resolve the problems. Even after it was found possible to perform accurate calculations by a process of renormalisation in 1948 many physicists felt that the method was incomplete and would break down at smaller length scales unless a natural cut-off was introduced.

Another aspect of the quantum theory which caused disquiet was its inherent indeterminacy and the essential role of the observer in measurements. The Copenhagen interpretation seemed inadequate and alternative hidden variable theories were sought. It was felt that quantum

mechanics would be a statistical consequence of a more profound discrete deterministic theory in the same sense that thermodynamics is a consequence of the kinetic gas theory.

Lattice Theories

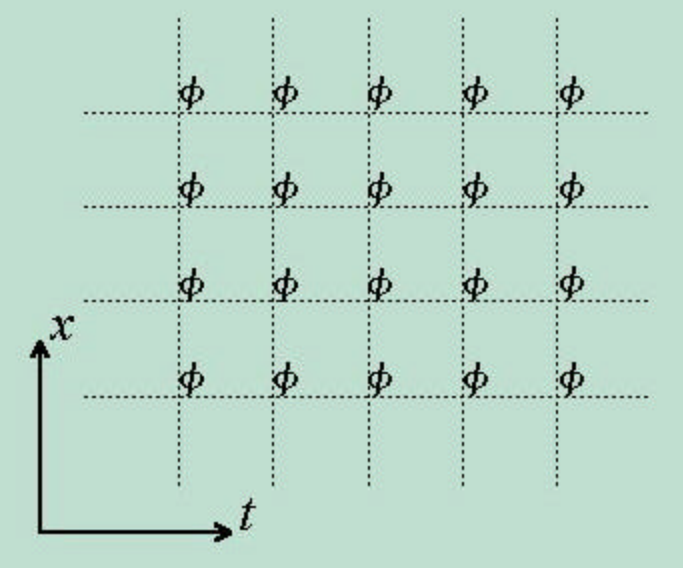
One way to provide a small distance cut-off in field theory is to formulate it on a discrete lattice with space-time events placed in a regular array like the molecules of a crystal. The numerical method for solving differential equations is to replace continuous space or time by discrete intervals as an approximation. This has been used since at least the eighteenth century and the possibility of applying such techniques to a discrete geometry of space was investigated by Oswald Veblen and William Bussey as early as 1906 but only later was it studied in any depth.

Classical field theories are described in terms of quantities which vary continuously over space and time according to certain wave equations. For example, electromagnetism has an electric field and a magnetic field each of which is described by three real numbers for each event of space-time. The equations which determine how they evolve are Maxwell's equations. The equations have derivatives in them which only make sense on continuous space and time, so if space-time is really a discrete lattice the equations will have to be replaced by some alternative which avoids the derivatives and approximates the original equations at large scales.

To make things simpler we will look at how this could be done for a simpler wave equation. The massless Klein-Gordon equation in two dimensions has just one field value at each event. The value will be a complex number since the Klein-Gordon equation was first proposed as a relativistic generalisation of the single particle Schrödinger equation. Usually it is denoted by ϕ (x,t). The equation is as follows:

$$\frac{\partial^2 \phi}{\partial t^2} - \frac{\partial^2 \phi}{\partial x^2} + m^2 \phi = 0$$

This has solutions which describe localised wave packets of energy like particles of mass m moving at less than a speed of 1 unit which is the speed of light. In discrete space-time the values of ϕ are only defined on the sites of a lattice which are spaced regularly at a distance d apart in the space dimensions and also in the time dimension.



The derivatives which appeared in the wave equation can no longer be defined exactly but they can be approximated using finite differences. E.g.

$$\frac{\partial^2 \phi(x,t)}{\partial x^2} \cong \frac{\phi(x+d,t) - 2\phi(x) + \phi(x-d,t)}{d^2}$$

If this and a similar approximation for the time derivative is substituted into the Klein-Gordon equation we get an equation which is well defined on the lattice.

$$\frac{\phi(x,t+d) + \phi(x,t-d) - \phi(x+d,t) - \phi(x-d,t)}{,2} + m^2 \phi(x,t) = 0$$

to real life but on a much smaller scale. A sceptic might ask about what happens between the discrete time steps or what lies in the space between the sites of the lattice. The answer is simply that there is nothing between. The sites are the only events of space-time which exist and the fields interact directly with their neighbours. Particles are formed as wave packets which are spread over many sites of the lattice so we never need to think of them as travelling between sites.

Lattice Quantum Field Theory

Part of the beauty of lattice theories is their simplicity. Continuum field theories are expressed in terms of differential equations while lattice theories are written with simple arithmetic operations such as subtraction. This economy of concepts is even more striking when we move on from the classical theory to the quantum. Quantum field theory is notoriously difficult to learn because it requires many mathematical concepts to describe. Even with these things understood quantum field theory is not as complete and rigorously defined as a mathematician would want. In contrast, lattice quantum field theories are quite simple, and so long as we do not concern ourselves with the continuum limit, they are usually well defined.

Quantum field theory as expressed by Richard Feynman starts from the Lagrangian formalism. In the case of the Klein-Gordon equation a Lagrangian density is defined as follows:

$$L = \left|\frac{\partial\phi}{\partial x}\right|^2 - \left|\frac{\partial\phi}{\partial t}\right|^2 + m^2|\phi|^2$$

The modulus squared of the complex numbers is used so that the Lagrangian is always real. The action is given by

$$S(\phi) = \int L(\phi) d^3x dt$$

By the principle of least action for the classical filed theory, this must be minimised subject to boundary conditions which fix the value of ϕ at any given start and end times. By an application of the calculus of variations the Klein-Gordon field equations can be derived from this principle. According to Feynman the quantum theory replaces the principle of least action with a path integral which defines a transition amplitude for going from each initial field configuration to a final one.

$$P = \int e^{\frac{i}{2\pi\hbar}S(\phi)} D\phi$$

The path integral must be taken over all possible evolutions of the field between the start and end. Not only does this sound complicated, it is not even possible to define rigorously except when the field equations are linear. Ordinary integration has been around since Newton and

Leibniz and was rigorously defined by Riemann in the eighteenth century. Path integrals only appeared in the latter half of the twentieth century and are still not well defined accept in restricted cases. Informally the path integral is a sum over all possible ways the field can vary over space and time but defining exactly what such an infinite-integral means is less simple to do.

By comparison the lattice version of the same thing is much easier to grasp. The lattice Lagrangian is just a discretised version of the continuum Lagrangian.

$$L = \left|\frac{\phi(x+a,t)-\phi(x,t)}{a}\right|^2 - \left|\frac{\phi(x,t+a)-\phi(x,t)}{a}\right|^2 + m^2|\phi(x,t)|^2$$

The action is a sum over the lattice sites.

$$S = \sum_{x,t} a^2 L(x,t)$$

The classical lattice field equations already given above can be derived from the action relatively easily by just requiring that the action is minimised with respect to variation of each field variable ϕ(x,t).

The lattice quantum field theory is then specified in a similar way as for the continuum field except that now the integral is a multi-variable integral over each field variable. This may still sound complicated but at least multi-variable integrals are well defined (when they converge) which is a big improvement over path integrals. If we believed that space-time was a lattice we would never have to worry about problems like renormalisation because the lattice spacing sets a cut-off scale which turns the divergences of field theory into well-defined finite answers. Such convenience does not make them right, of course, but it might count for something.

Lattice Gauge Theories

It is instructive to see how lattice theories work in more complicated cases. We know that the standard model of particle physics is built around gauge theories so it would certainly be worth while to look at gauge theories on the lattice. The obvious thing to do would be to take the continuum Lagrangian for Yang-Mills theory and replace all the derivatives with finite differences as we did for the Klein-Gordon equation. I have not described the Yang-Mills equations here so instead we shall see how lattice gauge theories can be formulated directly from the symmetry principles of gauge theory applied to the lattice Klein-Gordon Lagrangian.

The action for two-dimensional Klein-Gordon theory can be written differently by expanding the squares and collecting together the square terms in the sum over lattice sites. Actually the square terms from the difference terms cancel in the sum and we are left with a sum over an alternative Lagrangian.

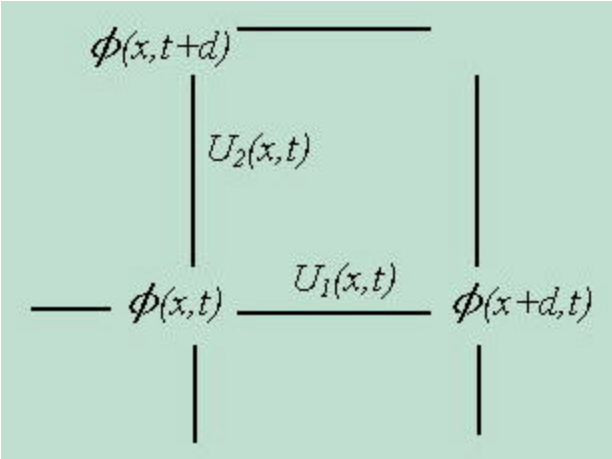
$$L = \frac{2\,\mathrm{Re}[\phi^*(x,t+d)\,\phi(x,t)]}{d} - \frac{2\,\mathrm{Re}[\phi^*(x+d,t)\,\phi(x,t)]}{d} + m^2\left|\phi(x,t)\right|^2$$

Recall the gauge symmetry for the electromagnetic field is invariance of the wave equation when the wave function is multiplied by a complex phase.

$$\phi(x,t) \to e^{i\theta(x,t)}\phi(x,t)$$

The Lagrangian for the lattice Klein Gordon equation is already invariant under this transformation when the phase θ(x,t) is a global constant, independent of x and t. The principles of gauge theory require us to introduce a gauge field in such a way that the Lagrangian is an invariant even when the phase is not a constant. As it stands the Lagrangian is not invariant because the field values at (x,t) are directly multiplied by field values at (x+d,t) and (x,t+d). Notice that the mass term does not suffer this problem and is already invariant.

Remember the analogy between gauge fields and economics. Multiplying field values together at different places is like trying to exchange money between different countries with different currencies. An exchange rate must be used. In the gauge theory the exchange rate is a phase factor U which is a unit complex number. Since the Lagrangian has products extending between any site and its nearest neighbours we must introduce such a factor on each link between sites of the lattice in both space and time directions. We will use U_i(x,t) for the variables linking site (x,t) to (x+d,t) and (x,t+d).



These phases are the field values of the gauge field. They represent the electromagnetic force on the lattice. When a local gauge transformation changes the matter field variables by a phase which can vary from one site to another, the gauge field must also be adjusted, just as exchange rates must be modified by a factor if the values of currencies change.

The gauge transformation is as follows.

$$\phi(x,t) \to e^{i\theta(x,t)}\phi(x,t)$$

$$U_1(x,t) \to e^{i\theta(x,t)}U_1(x,t)e^{-i\theta(x+d,t)}$$

$$U_1(x,t) \to e^{i\theta(x,t)}U_1(x,t)e^{-i\theta(x+d,t)}$$

gerorally with co-ordinate transformations. Only in

the time direction. The greater difficulty lies with r
displaced by any multiple of the lattice spacing alon
theories on a regular lattice have discrete translatio
version of gauge invariance are points in its favour.
of physics at very small length scales. The simple f
Here we are more interested in the possibility that
the lattice spacing tends to zero if the fields and co
Lattice gauge theory is an approximation to Yang-

dramatic chromodynamics.

structure of particles composed of quarks and pro
important part of a method for performing numeric
groups in any number of dimensions. Using Wilson
form which every given a discrete lattice approxima
two dimensions. In 1974 Ken Wilson discovered t
force. When this term is added to the matter field it
β - is just a coupling constant parameter, which co

$$L^{gauge} = \beta K\sigma[U_1(x,t)U_2(x$$

product of four gauge fields round a square of link
of course, it must be gauge invariant and real. It tur
term and all others in the Lagrangian are then
transformation. However, the Lagrangian is still inc
term some dynamics. The Lagrangian should includ

$$S K\sigma[\phi_*(x,t+\sigma$$

becomes

$$S K\sigma[\phi_*(x,$$

the appropriate gauge field in between the product
With these fields the Lagrangian can be modified to

$$U_j(x,t) \to e_{\ldots}$$

theory. If space-time was such a lattice there would be a preferred set of space axis and a preferred reference frame but such things contradict relativity and have never been observed.

If the continuum limit is not to be restored by taking the limit where the lattice spacing goes to zero then the issue of the loss of rotational invariance must be addressed. A space-time constructed as a discrete lattice is analogous to a crystal whose atoms are arranged on a regular array. At first sight the internal structure of a crystal solid appears isotropic but there its mechanical properties can be carefully measured to determine the directions in which the atoms are aligned. If space-time was a regular lattice its loss of rotational invariance would also be present even though it might not be detectable with present technology. Lorentz invariance would also be lost so relativity would be violated in a way which is hard for theorists to accept.

The fact is that lattice theories of space-time cannot easily be ruled out but they are just too plain ugly to be right! The laws of physics seem to be based on elegant principles such as symmetry which help determine the correct form the laws of physics must take. If we abandon those principles we have little hope of making progress. Lattice theories are arbitrary in their form. There is an infinity of ways to approximate any field theory on a lattice. How would we know which is right if experiment can never probe at sufficiently small length scales? This arbitrariness is the price you pay whenever you abandon a principle of symmetry.

Nevertheless, the fact that we can accommodate gauge invariance on the lattice may be telling us something. If we could represent diffeomorphism invariance in such a clean discrete form too, there would be some hope. The discrete version of diffeomorphism invariance is permutation invariance. Diffeomorphisms are one-to-one mappings of the set of space-time events to itself which preserve its continuum properties. Permutations are one-to-one mappings of a discrete set of events to itself. We call this event symmetry. The event-symmetric analogue of a lattice gauge theory is a gauge glass with events each linked to each other using gauge fields. The lattice structure is discarded. This gives a complete model of symmetries but how could such a structureless model be anything to do with physics?

Fading Motivations

Over the years many of the problems which surrounded the development of the quantum theory have diminished. Renormalisation itself has become acceptable and is proven to be a consistent procedure in perturbation theory of Yang-Mills gauge field physics. The perturbation series itself may not be convergent but Yang-Mills theories can be regularised non-perturbatively on a discrete lattice using the prescription introduced by Ken Wilson. There is good reason to believe that consistent quantum field theory can be defined on continuous space-time at least for non-abelian gauge theories which are asymptotically free. In lattice QCD the lattice spacing can be taken to zero while the coupling constant is systematically rescaled. In the continuum limit there are an infinite number of degrees of freedom in any volume no matter how small. This would be a counterexample to any claim that physical theories must be discrete.

Quantum indeterminacy, which was another motivation for looking to discrete space-time, has also become an acceptable aspect of continuum physics. In 1964 John Bell showed that most ideas for local hidden variable theories would violate an important inequality of quantum

mechanics. This inequality was directly verified in a careful experiment by Alain Aspect in 1982. There are still those who try to get round this with new forms of quantum mechanics such as that of David Bohm, but now they are a minority pushed to the fringe of established physics. Hugh Everett's thesis which leads us to interpret quantum mechanics as the dynamics of a multiverse has been seen as a resolution of the measurement problem for much of the physics community. Others are simply content with the fact that quantum mechanics provides the same way of doing calculations no matter what interpretation is used.

Without the physical motivation discrete space-time has been disfavoured by many physicists but others have found reason to persist with the idea.

It from Bit

In the late 1970s the increasing power of computers seems to have been the inspiration behind some new discrete thinking in physics. Monte Carlo simulations of lattice field theories were found to give useful numerical results with surprisingly few degrees of freedom where analytic methods had made only limited progress. Their newly found close contact with computers seems to have led some physicists to wonder if the universe is itself some sort of giant computer.

In 1947 Claude Shannon laid the foundations of information theory. The smallest unit of information used in computers is the binary digit or bit. Each bit can just have a value 0 or 1 but many bits can record vast amounts of information in the form of numbers or binary coded characters. Shannon's information theory turned out to be important in physics as well as computers. It seems that the entropy of a system may be a measure of the amount of information it contains but it is difficult to make sense of such an idea unless the amount of information in a physical system is finite. If the positions and orientations of molecules can be specified to any degree of precision then there is no limit to the number of bits needed to describe the state of a gas in a box so entropy from information may only make sense if there is some minimum distance which can be measured.

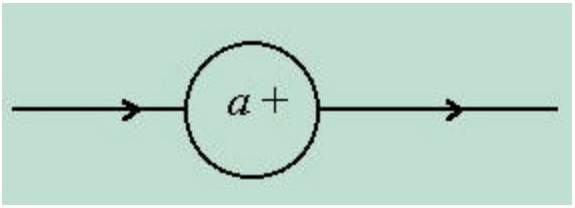
Such reasoning has created a school of thought about the role of information processing in the fundamental laws of physics. John Wheeler has sought to extend this idea so that every physical quantity derives its ultimate significance from bits. He calls this idea "It from Bit." For Wheeler and his followers the continuum is a myth, but he goes further than just making space-time discrete.

Space-time itself, he argues, must be understood in terms of a more fundamental pregeometry. In the pregeometry there would be no direct concepts of dimension or causality. Such things would only appear as emergent properties in the space-time idealisation. All would be the consequence of complex interactions based on very simple basic elements, just as a complex piece of computer software is built from a simple set of instructions.

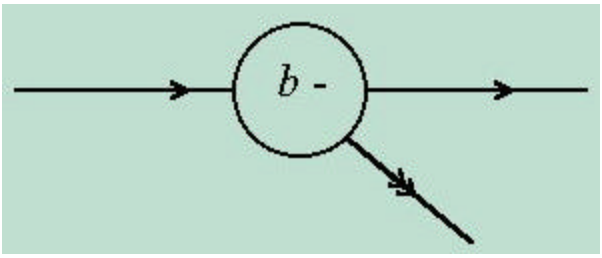
There are many different instruction sets which have been used to control computers. In RISC processors the number of different instructions is kept to a minimum. In the theory of computers, without the practical constraints of efficiency, it is possible to reduce the instruction set to very few elements indeed and still be able to use it to do any computation which is theoretically

possible. Such a machine is called a *universal computer*. In 1979 while I was a student I attended an extra-circular course on logic given by the mathematics professor John Conway. He introduced the class to a hypothetical computer called a Minsky machine which had been devised by computer science theorist Marvin Minsky. The computer can store an unlimited number of non-negative integer values which are given variable names a, b, c, … etc. The computer obeys two fundamental instructions:

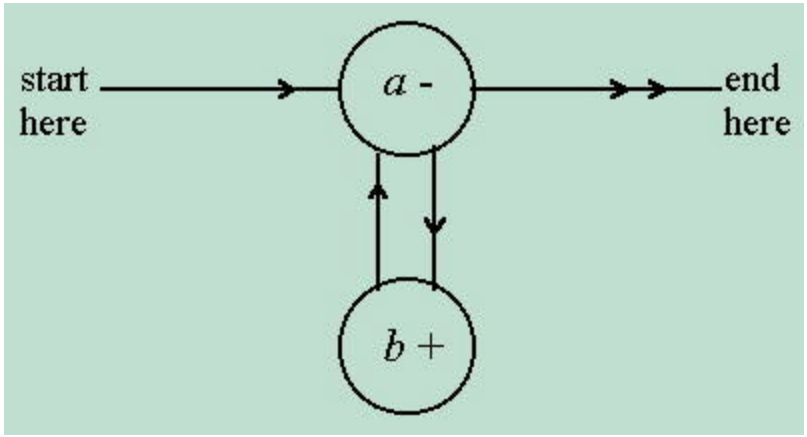
- (1) increment a variable by adding one. E.g. the instruction to increment variable a can be written schematically like this



- (2) decrement a variable by subtracting one, unless it is zero in which case branch E.g. the instruction to decrement variable b or branch is shown as follows

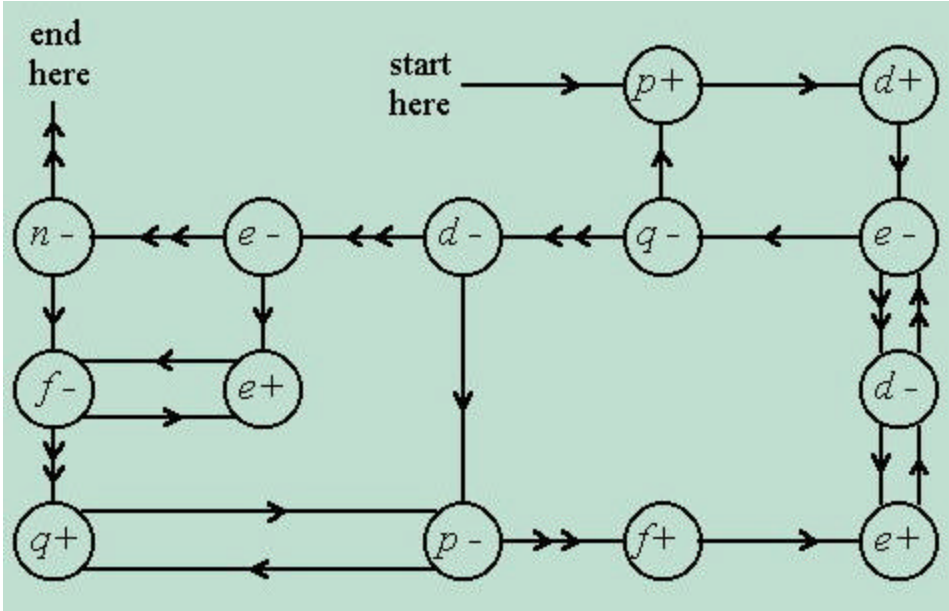


The branch with the double arrow is taken if b is zero on entering the circle. A program for a Minsky machine is a diagram made up of these two instructions. Here is an example of a simple program to add a to b.



If you want an interesting puzzle to solve try and work out what is the largest number which a Minsky machine can generate in a variable when it stops if it is only allowed to have k instructions where k is some small number of your choice.

In one lecture of the course, Conway showed us a program he had written for a Minsky machine which could calculate the nth prime number. It had only 16 instructions and he challenged us to do better. The next week I showed him how to do it with only 14 instructions. Can you do better still? Here is the program. Start with all variables set to zero except n. When you arrive at the end p will be the nth prime number. This Minsky machine program illustrates how the simplest of rules can be used to generate non-trivial systems. Perhaps some equally simple set of rules will account for physics.



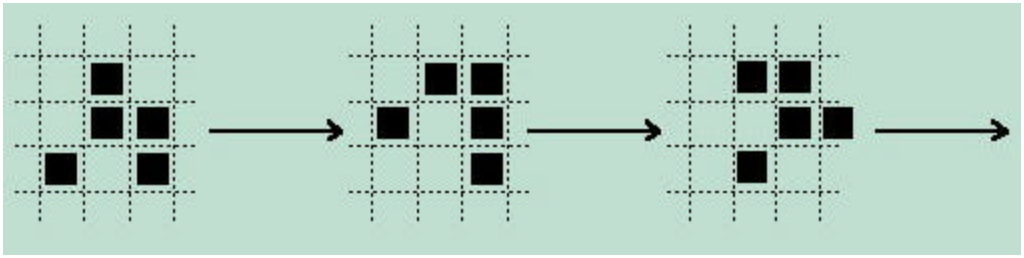
Cellular Automata

A similar idea which seems closer to the real world is the cellular automata. Cellular automata became popular in the 1970s with Conway's invention of the Game of Life. Its simple rules made it popular with people who liked recreational mathematics and was partly responsible for Conway's popularity as a lecturer.

The game of life is played on a two-dimensional array of square cells. Each cell at any given time step is either alive or dead. The state of the game at the next time step is determined by rules which are meant to mimic the life and death of living cells. If a living cell at one moment is isolated or it is accompanied by no more than one other living cell in the nearest neighbouring 8 cells, it will be dead the next moment through lack of support. If it is surrounded by two or three living cells in its neighbourhood it will continue to live but if there are more it will die from over competition. On the other hand, a cell which is dead will be revived if it is surrounded by exactly three living cells. Otherwise it remains dead. When these rules are applied iteratively to an initial

picture of living and dead cells, the system evolves and patterns emerge. A computer can readily be made to simulate the game and display the progress.

Typically regions of cells will die out or stabilise into patterns which do not change such as an isolated square of four cells, or which repeat such as a line of three cells. From time to time a group of living cells will appear to separate from the activity and move away on its own. These are called gliders. The most common variety reflects about a diagonal axis after each second step and moves diagonally.



Despite its simple rules defined on a discrete lattice of cells the game has some features in common with the laws of physics. There is a maximum speed for causal propagation which plays a role similar to the speed of light in special relativity. Even more intriguing is the comparison of gliders with elementary particles. Cellular automata go a step further than lattice field theories. Even the continuous values of the field variables have been replaced with discrete quantities.

A great deal of research has been done to find out how cellular automata like this one behave on very large arrays. Numerical simulations suggest that stable regions develop but some activity can continue for a long time. It seems that *self organised criticality* is established. This means that the system stops evolving leaving steady or cyclic configurations of cells, but a small perturbation such as a glider wandering in from outside can set the thing off again like a spark lighting a fire. Little is known about how cellular automata might behave on very large arrays and over very large numbers of time steps. Recall that the smallest scales in physics seem to be around 10^{-35} m. To correspond in size to our universe, a cellular automaton would have to have an array of something like 10^{240} cells.

Despite its simple rules the game of life has sufficient complexity that we cannot imagine how an array that big would behave. On large scales some kind of physical laws may emerge from the statistical behaviour of the system. It is quite possible that complex organised structures would evolve. It is plausible that some cellular automata specified in 3-dimensions may be sufficiently interesting places for life to develop inside them. At present we have no idea if such things are likely. For those seeking to reduce physics to simple deterministic laws this was an inspiration to look for cellular automata as toy models of particle physics despite the obvious flaw that they broke space-time symmetries. Edward Fredkin is one of those people who suggests that the universe really does operate like a gigantic computer. Fredkin is a computer specialist with an interest in physics who has managed to influence a number of respected physicists to take the idea seriously.

In 1981 Fredkin was one of the organisers of a conference at MIT which he wanted to be called something like "On computational models of physics." Fredkin managed to persuade Richard

Feynman to be the keynote speaker at the meeting, but when Feynman heard the title he said "Well if you have that as a name, and it implies that there are computational models of physics, then I am not coming." The title was changed to "Physics and computing" and so Feynman went. However, by time Feynman arrived to give his talk he had changed his mind and gave a talk about computational models of physics. He and many other speakers spoke about cellular automata which were very topical by then. Other speakers at the conference included Wheeler, Minsky and Fredkin himself. This conference and especially the presence of Feynman was very influential on the subject.

There has been some progress towards using cellular automata to study hydrodynamics and turbulence but there seems to be an impassable hurdle when we attempt to apply the automata to quantum physics. The evolution of automata is always based on what happens locally to any cell in the array, but Bell's inequality and the experiments of Aspect and others strongly suggest that quantum reality is not local in such a strong sense.

Another notable physicist who has been influenced by Fredkin is Gerard 't Hooft. He is not put off by locality arguments and suggests that the states of a cellular automaton could be seen as the basis of a Hilbert space on which quantum mechanics is formed. Although the idea is not popular, some interesting things may yet be learnt from such research.

Discreteness in Quantum Gravity

We have seen how some of the early motivations behind theories of discrete space-time have faded with time, but recently new evidence has come in to take their place. It is only when we try to include gravity in quantum theory that we find solid reason to believe in discrete space-time. With quantisation of gravity all the old renormalisation issues return and many new problems arise.

Whichever approach to quantum gravity is taken the conclusion seems to be that the Planck length is a minimum size beyond which the Heisenberg Uncertainty Principle prevents measurement if applied to the metric field of Einstein Gravity. In ordinary quantum field theory the ability to measure small distances is limited only by the energy of the particles available and according to relativity there should be no theoretical limit to energy. When gravity is included, however, the metric itself becomes uncertain. At smaller distances the quantum fluctuations of the metric become more significant until, at the scale of the Planck length, it is impossible to do any reliable measurements.

Does this mean that space-time is discrete at such scales with only a finite number of degrees of freedom per unit volume? Recent theoretical results from string theories and the loop-representation of gravity do suggest that space-time has some discrete aspects at the Planck scale. These are akin to the discrete quantum numbers of the quantum mechanics of an atom which still also has a continuum description so the answer may be that space and time have a dual discrete and continuous nature.

The far reaching work of Bekenstein and Hawking on black hole thermodynamics has led to some of the most compelling evidence for discreteness at the Planck scale. The *black hole*

information loss paradox which arises from semi-classical treatments of quantum gravity is the nearest thing physicists have to an experimental result in quantum gravity. Its resolution is likely to say something useful about a more complete quantum gravity theory. There are several proposed ways in which the paradox may be resolved most of which imply some problematical breakdown of quantum mechanics while others lead to seemingly bizarre conclusions.

One approach is to suppose that no more information goes in than can be displayed on the event horizon and that it comes back out as the black hole evaporates by Hawking radiation. Bekenstein has shown that if this is the case then very strict and counter-intuitive limits must be placed on the maximum amount of information held in a region of space. It has been argued by 't Hooft that this finiteness of entropy and information in a black-hole is also evidence for the discreteness of space-time. In fact the number of degrees of freedom must be given by the area in Planck units of a surface surrounding the region of space. This has led to some speculative ideas about how quantum gravity theories might work through a holographic mechanism, i.e. it is suggested that physics must be formulated with degrees of freedom distributed on a two-dimensional surface with the third spatial dimension being dynamically generated.

At this point it may be appropriate to discuss the prospects for experimental results in quantum gravity and small scale space-time structure. Over the past twenty years or more, experimental high energy physics has mostly served to verify the correctness of the standard model of particle physics as established theoretically between 1967 and 1973. We now have theories extending to energies way beyond current accelerator technology but it should not be forgotten that limits set by experiment have helped to narrow down the possibilities and will presumably continue to do so.

It may seem that there is very little hope of any experimental input into quantum gravity research because the Planck energy is so far beyond reach. However, a theory of quantum gravity would almost certainly have low energy consequences which may be in reach of future experiments. The discovery of supersymmetry, for example, would have significant consequences for theoretical research on space-time structure.

Lattice Quantum Gravity

If discrete space-time is a feature of quantum gravity then the early ideas of lattices and cellular automata were just not inventive enough. A lattice is surely too rigid a structure to model curved space-time. General relativity is about invariance of the form of laws of physics under co-ordinate transformations but the space-time co-ordinates are really artificial constructs without any direct physical basis. In 1961 Tulio Regge came up with a way of doing relativity without any co-ordinates. He imagined space-time as a network of points joined together by links, triangles, tetrahedrons and pentahedroids. These are simplexes of dimension 1 to 4. The structure is analogous to the faceted surface of a geodesic dome. Just as the curving vaults of a modern building can be approximated by a surface of flat triangles, so too can curved space-time by approximated to any desired accuracy with the simplicial structure of the Regge skeleton.

The concept is very much like a lattice except that it is not rigid. Instead of varying field values on sites the length of the links between the sites is allowed to be variable. It is sufficient to

specify how the sites are connected and the lengths of all the links. Then the size and shape of all the simplexes can be determined. The curvature of the space-time surface can be derived from the angles of the simplexes around any site. It is possible to work out the equations which express the dynamics of the structure and which reduce to Einstein's field equations of general relativity in the limit where the size of the simplices becomes very small. The Regge calculus is therefore a discrete version of general relativity. Useful numerical simulations of either the classical or quantum dynamics can be done on a fast computer.

To Regge this discrete space-time was just an approximation scheme which would give ordinary general relativity in the fine limit. To us it could also be a pregeometric model of space-time, valid even while discrete. If space-time was a Regge skeleton we would have to find some rules about how it should be split into simplexes. Loss of space-time symmetry is also a problem just as it was with a regular lattice.

An alternative scheme which has proved to work better in numerical studies of quantum gravity is *random triangulation*. Instead of varying the lengths of the links joining sites the links are all the same length and the way space-time is divided into simplexes is varied. Space-time curvature varies with the number of simplexes which meet at each site. The path integral of quantum gravity is then effectively a sum over all the ways of triangulating a four-dimensional surface. The action can be given in terms of just the numbers of simplexes in the lattice. Discrete effects are averaged out so that rotational symmetry is exact in the quantum version. This is an interesting pregeometric model though it would be surprising if it was anything like reality.

Pregeometry

For John Wheeler simplicial space-time was not radical enough. He demanded a pregeometry much more basic than the space-time manifold or any discrete approximation to it. A true description of the structure of space-time at the smallest scales may require us to discard some other properties which it appears to have at larger scales. For example, ***dimension*** may not be a fundamental quantity. We know that space-time is four-dimensional on scales at least as large 10^{-19} m which have been probed with particle accelerators, but at the Planck scale the number of dimensions may change. It may even become a vague concept with no definite meaning. Other features which space-time physics may lose along with ***continuity*** include its ***metric***, ***topology***, ***symmetry***, ***locality*** or ***causality***. We cannot be sure that space-time ***events*** have a precise meaning or that ***quantum mechanics*** works the same way. In short it is difficult to imagine what space-time may be like at all.

Any pregeometric model can be characterised according to which of the highlighted properties in the previous paragraph it throws out and which it keeps. For example, lattice models discard continuity and symmetry but keep dimension, metric, events, etc. Cellular automata also discard quantum mechanics. Some physicists have played the game of building toy models which throw out all but a few of these concepts, the ones which they feel might be the most fundamental. They might try to keep causality, locality and quantum mechanics for example, because they think these things are of primary significance and must be part of the laws of physics at the most fundamental level. Another feature like topology, a metric or even information might be thrown in just to see what it led to.

Before about 1980 only a rare few physicists had made any serious attempts at this sort of thing. The best examples were Hartland Snyder with quantum space-time, David Finkelstein with his quantum net dynamics, Carl von Weizsäcker with Ur-theory and Roger Penrose with spin networks and twisters. Then in the 1980s and early 1990s there was a flurry of new speculative ideas. The time seemed right for bold ideas. Chris Isham and others looked at the quantum mechanics of spaces with just a distance metric between scattered points, or topologies of sets or even just random networks of links between space-time events.

Is there really any hope that such methods can tell us something about the real world? Physicists have succeeded before with theories they devised with little more than mathematics and insight. Dirac was a strong advocate of the power of mathematical beauty as an indicator of truth and successfully predicted the positron on such a basis. If you examine the pregeometries which have been studied up till now it is easy to dismiss them because none is complete.

However, rather than discarding each one because of some feature which does not correspond to reality, you can also look for features which seem promising. Better theories can then be produced by combining things from different models which might work well together. It seems improbable that someone is going to have complete success by such methods alone, but if clues from superstring theory and canonical quantum gravity are also considered there may be some hope.

Sadly, there is little encouragement or funding for such speculative research. Happily there are still a hand full of physicists and one or two journals which keep it alive.

The Metaphysics of Space-Time

Space and time have been favourite subjects of debate for philosophers since at least the ancient Greeks. The paradoxes of the infinite and the infinitesimal are reinvented each day by children with inquisitive minds. How can space be infinite? If it is not infinite what would lie beyond the end? Can the universe have a beginning and an end? What is the smallest thing and what can it be made of? What is time? Do time and space really exist?

How have modern physicists learnt to deal with these questions? The simplest answer is that they use mathematics to construct models of the universe from basic axioms. Mathematicians can define the system of real numbers from set theory and prove all the necessary theorems of calculus that physicists need. With the system of real numbers they can go on to define many different types of geometry. In this way it was possible to discover non-Euclidean geometries in the nineteenth century which were used to build the theory of general relativity in the twentieth.

The self consistency of general relativity can be proven mathematically from the fundamental axioms within known limitations. This does not make it correct, but it does make it a viable model whose accuracy can be tested against observation. In this way there are no paradoxes of the infinite or infinitesimal. The universe could be infinite or finite, with or without a boundary. There is no need to answer questions about what happened before the beginning of the universe because we can construct a self-consistent mathematical model of space-time in which time has a beginning with no before.

So long as we have a consistent mathematical model we know there is no paradox, but nobody yet has an exact model of the whole universe. Newton used a very simple model of space and time described by Euclidean geometry. In that model space and time are separate, continuous, infinite and absolute.

This is consistent with what we observe in ordinary experience. Clocks measure time and normally they can be made to keep the same time within the accuracy of their working mechanisms. It as if there were some universal absolute standard of time which flows constantly. It can be measured approximately with clocks but never directly.

So long as there is no complete theory of physics we know that any model of space-time is likely to be only an approximation to reality which applies in a certain restricted domain. A more accurate model may be found later and although the difference in predicted measurement may be small, the new and old model may be very different in nature. This means that our current models of space and time may be very unrealistic descriptions of what they really are even though they give very accurate predictions in any experiment we can perform.

Philosophers sometimes try to go beyond what physicists can do. Using reason alone they consider what space and time might be beyond what can be observed. Even at the time of Newton there was opposition to the notion of absolute space and time from his German rival Leibniz. He, and many other philosophers who came after, have argued that space and time do not exist in an absolute form as described by Newton.

If we start from the point of view of our experiences, we must recognise that our intuitive notions of space, time and motion are just models in our minds which correspond to what our senses find. This is a model which exists like a computer program in our head. It is one which has been created by evolution because it works. In that case there is no assurance that space and time really exist in any absolute sense.

The philosophical point of view developed by Gottfried Leibniz, the Bishop Berkeley and Ernst Mach is that space and time should be seen as formed from the relationships between objects. We experience objects through their relationships with our senses and infer space and time more indirectly. The mathematical models used by physicists turn this inside-out. They start with space and time, then they place objects in it, then they predict our experiences as a result of how the objects interact.

Mach believed that space and time do not exist in the absence of matter. The inertia of objects should be seen as being a result of their relation with other objects rather than their relation with space and time. Einstein was greatly influenced by Mach's principle and hoped that it would follow from his own postulates of relativity.

In the theory of special relativity he found that space and time do not exist as independent absolute entities but Minkowski showed that space-time exists as a combination of the two. In General Relativity Einstein found, ironically, that the correct description of his theory must use the mathematics of Riemannian geometry. Instead of confirming Mach's principle he found that space-time can have a dynamic structure in its own right. Not only could space-time exist

independent of matter but it even had interesting behaviour on its own. One of the most startling predictions of general relativity; that there should exist gravitational waves, ripples in the fabric of space-time itself, may soon be directly confirmed by detection in gravitational wave observatories. In short, relativity succeeded in showing that all motion is relative but it failed to construct a complete relational model of physics.

Einstein's use of geometry was so elegant and compelling that physicists thereafter have always sought to extend the theory to a unified description of matter through geometry. Examples include the Kaluza-Klein models in which space-time is supposed to have more than four dimensions with all but four compacted into an undetectably small geometry. This is the opposite of what the philosophers prescribed. Thus physicists and philosophers have become alienated over the subject of space and time during the twentieth century.

Recent theories of particle physics have been so successful that it is now very difficult to find an experimental result which can help physicists go beyond their present theories. As a result they have themselves started to sound more philosophical and are slowly reviewing old ideas. The fundamental problem which faces them is the combination of general relativity and quantum theory into a consistent model.

According to quantum theory a vacuum is not empty. It is a sea of virtual particles. This is very different from the way that space and time were envisioned in the days of Mach. In a theory of quantum gravity there would be gravitons; particles of pure geometry. Surely such an idea would have been a complete anathema to Mach. But suppose gravitons could be placed on a par with other matter. Perhaps then Mach would be happy with gravitons after all. The theory could be turned on its head with space-time being a result of the interactions between gravitons.

Leibniz might also have been satisfied with such an answer. In his philosophy everything is constructed from monads. These could be packets of energy or more abstract entities. A discrete space-time would fit in well with the idea. Discrete elements of space-time can be put on a par with particles of matter suggesting the final unification of space-time and matter.

In string theory, the most promising hope for a complete unified theory of physics, we find that gravitons are indeed on an equal footing with other particles. All particles are believed to be different modes of vibration in loops of string. Even black holes, one of the ultimate manifestations of the geometry of space-time are thought to be examples of single loops of string in a very highly energised mode. There is no qualitative distinction between black holes and particles, or between matter and space-time.

The problem is that there is as yet no mathematical model which makes this identity evident. The equations we do have for strings are somewhat conventional. They describe strings moving in a background space-time. And yet, the mathematics holds strange symmetries which suggest that string theories in different background space-times and even different dimensions are really equivalent. To complete our understanding of string theory we must formulate it independently of space-time. The situation seems to be analogous to the status of electrodynamics at the end of the 19th century. Maxwell's equations were described as vibrations in some ether pervading

space. The Michelson-Morley experiments failed to detect the hypothetical ether and signalled the start of a scientific revolution.

Just as Einstein banished the ether as a medium for electromagnetism we must now complete his work by banishing space-time as a medium for string theory. The result will be a model in which space-time is recovered as a result of the relationship between interacting strings. It will be the first step towards a reconciliation of physics and philosophy. Perhaps it will be quickly followed by a change of view, to a point from where all of our universe can be seen as a consequence of our possible experiences just as the old philosophers wanted us to see it. What other ways will we have to modify our understanding to accommodate such a theory? Not all can be foreseen.

So is it or isn't it?

There do seem to be good reasons to suppose that space-time is discrete in some sense at the Planck scale. Theories of quantum gravity suggest that there is a minimum length beyond which measurement cannot go, and also a finite number of significant degrees of freedom. In canonical quantisation of gravity, volume and area operators are found to have discrete spectra, while topological quantum field theories in 2+1 dimensions have exact lattice formulations.

At the same time, the mathematics of continuous manifolds seems to be increasingly important. Topological structures such as instantons and magnetic monopoles appear to play their part in field theory and string theory. Can such things be formulated on a discrete space?

Hawking says that he sees no reason to abandon the continuum theories that have been so successful. It is a valid point but it may be possible to satisfy everyone by invoking a discrete structure of space-time without abandoning the continuum theories if the discrete-continuum duality can be resolved as it was for light and matter.

The philosopher Immanuel Kant may have had some insight into this question. The human mind can pose questions about nature which have contradictory but perfectly logical answers. One such question is whether the world is made of elementary parts. The answer can be both yes and no. The riddle may be resolved through a dual theory of space-time which has both discrete and continuous aspects.

What About Causality?

Causality in the news

I read a news article recently which reported that family conflict can stunt the growth of young children. A survey had shown that parents who divorce or separate tend to have smaller children. According to the team who conducted the study this is scientific evidence of how conditions in childhood can have lifelong consequences.

But how right were they? To conduct the survey someone visited schools and measured the height of many children with the same age. The results were then compared statistically with the circumstances of their parents. Presumably they found a statistically significant negative correlation between height and indicators of family conflict such as divorce, thus proving the link. Fine so far, but can we conclude that the conflict caused children to be smaller? Would it not have been equally valid to conclude that having small children leads to divorce? The scientist in charge speculated that stress may reduce the amount of growth hormone that young children produce.

In fact he applied his prejudices and drew a conclusion which sounds reasonable without realising that the converse was also a possible explanation of the survey results. It is not difficult to believe his theory but there was nothing from the survey which proved it. In fact the real reason behind the correlation may have been one or more third factors such as wealth. Children of poorer families may have worse standards of nutrition resulting in slower growth, and lack of money might also lead to higher divorce rates. Another cause may have been a genetic trait which shows up in both the growth and temperament of family individuals. Such effects are equally likely to show up as a correlation in the survey but the news article said nothing about such possibilities.

The difference between the possible conclusions from the survey is not just one of semantics. People reading the article could blame their frequent family rows for having a small child. Such feelings of guilt are unlikely to help the situation. They may have been right but I suspect they would have been wrong. Surveys such as this are common and are often reported in the media by people who do not appreciate the traps that statistics can lead us into.

When responsible scientists wish to establish causal links between different effects they are more careful. For example, when a new drug is tested it is necessary to know how effective it is and what side effects it may produce. To do this a group of volunteers is selected for trials. The group is divided in two at random and one half is given the drug. The other half is given a placebo pill which is known to have no effect. Nobody taking part knows which group they are in. Both groups are then monitored for possible effects. The effect is known to be real if it is significantly more noticeable amongst those who took the drug than those who took the placebo. It is then certain that taking the drug really *caused* the effect. The difference between this example and the survey is that the choice of who got the drug and who did not was controlled. In the survey

which claimed to link height and family strife there was no control over whose parents were divorced which were not so it was impossible to distinguish cause from effect or rule out other factors with certainty.

Causality in Physics

Suppose your child bumped into a table and an expensive vase fell off, smashing into pieces on the floor. Would you conclude that her carelessness caused the vase to be broken? Probably you would, but why would you not conclude that the vase falling off the table caused her to bump, quite innocently, into the table? Your response might be that, for one thing, the vase was broken after her collision with the table so the direction of the causal link is incontestable. This reflects the modern concept of causality: Cause precedes effect. Yet the logical relation between the two events; her bumping into the table and the vase falling off, are symmetrically related. If one has happened the other probably has too. Is it just our prejudices which have made us favour a causal link or is it justified by physics?

Philosophers such as David Hume have been sceptical about these notions of causality. In 1740 Hume questioned the basic idea of causation. It is sometimes thought that his rejection of causation implies a rejection of scientific laws but it does not. What it really implies is a rejection of *free will*. Compare the case of the broken vase with the survey and the medical trials. Which does it more closely resemble? You might argue that it is more like the medical trials because a person has control over whether or not they do something like bump into a table. They have free will. The vase breaking is a response to an action of free will, even if it was an accident. If an action is controlled then it must be the cause rather than the effect. If we accept the contention of Hume we deny any distinction between cause and effect so we must also deny our free will.

Causality was not always characterised so simply as it is today. In ancient Greece, at the Lyceum in Athens, Aristotle taught that there were four types of cause: the *material*, the *formal*, the *efficient* and the *final*. If you build a boat he would have said that the causes were the materials you used, the plans you drew up, the labour you put into it and what you wanted to do with it. If any of these four things were not there, the boat would not be made.

In terms of modern physics we would regard the efficient and final cause as the two extremes of *temporal causality*, that is, causality related to time. The efficient cause is the initial set of conditions and the final cause is the final set of conditions. Likewise we can regard the material and formal causes as two opposite views of *ontological causality*, that is causality related to the way in which something is formed.

Let us imagine another example. You are very proud because you have successfully grown a good crop of potatoes in your back garden. You bring a handful in to show your daughter saying "Look, I grew some potatoes!" "Why did they grow?" she inquires, as children do. How would you respond? Just suppose that you are rather philosophical in your ways and you respond according to which types of causality you believe in. The conversation might continue as follows:

"They grew because of biological processes such as photosynthesis."

"Why are there biological processes like photosynthesis?"

"because of atoms and the laws of chemistry which make biological processes work"

"Why are there atoms and laws of chemistry?"

"because of nuclear physics and electromagnetic forces which make atoms out of protons, and electrons?"

"why...?"

"because of more elementary particles and laws of physics which we don't know everything about yet!"

These answers characterise a reductionist or atomist who believes that all explanation can be reduced to underlying laws of physics which may one day be explained through some deep principle of mathematics. Aristotle would say that you had invoked the material cause.

In another mood you might answer differently:

"They grew because I planted them"

"Why did you plant them?"

"because I knew they would be good to eat when they were ready"

"Why did you know?"

"because I learnt such things in school"

"Why did you go to school?"

"because a long time ago people realised that having an education was useful"

... and so on.

This time the conversation might continue through the history of humanity, life on Earth and cosmology until you explain that everything is a result of what happened at the big bang. Of course we are stuck again because we cannot say what caused the big bang. This may be a strange way to explain why potatoes grow but it is exactly how conventional wisdom describes causality in physics. Aristotle would have called it the efficient cause.

Since the 17th century scientists have replaced Aristotle's four causes with just those two: The efficient or prior cause and the material cause, or physical laws. The final and formal cause is gone. Descartes's mechanistic causality is the most widely accepted today. We would say that a cause of an event is any preceding event without which it would not have happened. In addition

to this temporal causality many physicists believe that there are fundamental laws of physics to which are other phenomena can be reduced. This reductionism is the material cause and it is what is left of the ontological causality.

If your mind is opened a little by my story of the survey in the news article, then you may also be ready to reconsider your notions of causality in physics. How would you explain the growth of your potatoes if you believed in a final cause?

"They grew to become potatoes"

"Why did they become potatoes?"

"So that we could eat them and grow ourselves"

"Why do we grow?"

"So that we can become strong enough to do our jobs"

Eventually it seems that this will lead towards some ultimate unknown destiny of humanity. These days most scientists do not believe in destiny but Aristotle would defend the final cause. A seed grows because it is destined to become a plant and produce more seeds. His error is easily exposed if we tear up the plant before it matures. It grew just the same to begin with even though the final cause was taken away. The same would not be true if we intervened before the seeds grew. Prior cause *seems* to be more right than final cause but notice that we have invoked our free will again to prove it.

It could be harder to explain growth in terms of the formal cause. We would have to suppose that the potatoes grew because it had a design purpose. You might say:

"They grew because if they didn't we would have nothing to eat. Then we would not be here to ask such questions!"

This may sound like an invalid explanation at first. Yet it is an explanation which might be given by someone who advocates the *anthropic principle*. Such people claim that the laws of physics and other aspects of our world are the way they are because they must be that way for us to be here. Reductionism and the anthropic principle are opposing philosophies of ontological causality. They correspond to Aristotle's material and formal cause respectively. Aristotle accepted both types of explanation but most people prefer one or the other.

Let us put ontological causality aside for now and consider temporal causality in more detail. Do the laws of physics justify Descartes who threw out final cause in favour of prior cause?

To keep things simple, let us start by considering just classical Newtonian mechanics. The form which the laws of physics take is crucial to our understanding of causality. Newton's laws take the form of a set of differential equations describing the motion of particles under forces that act between them. If we know the initial positions and velocities of all the particles at an initial time

then their positions are determined at any future time. So does this form for the laws of physics allow us to justify our concept of temporal causality, that cause comes always from the past and precedes its effect? It would seem so because the initial conditions seem to be causing all that happens in the future.

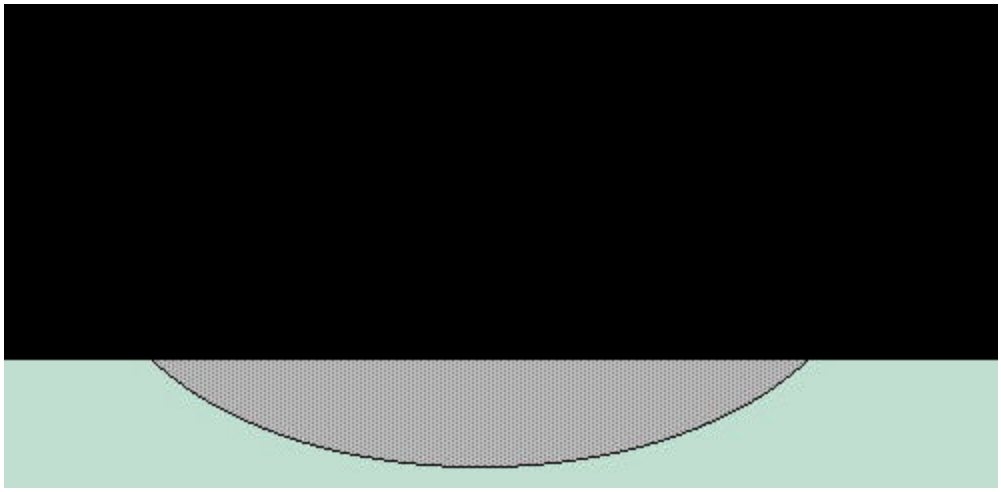
There is a catch. The laws of physics in this form can be made to work identically in reverse. If we know the final state of a system we can just as easily determine its past. Furthermore, the classical laws of mechanics do not allow any room for free will. All actions are predetermined by any complete past state. They are also postdetermined by any future state. Newton's laws do not explain why past events are the cause of future events.

A Block Universe

It is difficult to think clearly and rationally about causality because it is bound up with our experience. It is sometimes difficult to separate logical deduction from intuition. We are so used to the flow of time that it is almost impossible to detach ourselves from it and appreciate time as part of physics. Time flows past while space remains, yet since the debut of the theory of relativity we have known that the distinction between space and time is not so profound.

To appreciate the physics without being misled by intuition we must imagine ourselves separated from space and time. We need to imagine space-time as a single entity which does not evolve. Like a block of existence, the universe just is. Our lives are worldlines through the block stretching between birth and death. We might equally well say that they stretch between death and birth. On close examination we can tell which way our lives went from past to future because we recognise the symptoms of ageing but there are no time stamps built in to space-time. The block universe has no past, present or future. It is just a collection of events.

If the universe is finite and closed with a beginning at the big bang and an end at the big crunch you can think of it as a kind of rugby ball shaped surface which narrows at either end. Space-time is four-dimensional and has nothing outside or inside but we have to visualise it as a two-dimensional surface sitting in space. This limitation of our minds does not matter. We do not have to visualise something to understand it.



People often discuss what came before the big bang. Some think that there must have been something. Others say there was nothing. When we think about the block universe we see that there was no "before". The surface of the sausage is all that there is to the universe and time is part of it. We should not think of an empty space around it since that space does not exist. Of course we do not know that the universe is really rugby ball shaped and there could have been something before the big bang, but it is not necessarily so. We should not let our experience influence our reasoning since our experience is limited to a small part of the universe and prejudices our judgement. It is not easy to imagine a universe which is curved but which has no outside, no before and no after, but we can describe the shape of space-time mathematically without referring to anything outside, so an outside is not necessary. Asking about what came before the big bang is like asking what comes before the letter A in the alphabet. Asking about what is outside the universe or where it is, is like asking what is outside the alphabet or where it is. It is nowhere or everywhere. It just is.

Nevertheless, we can *imagine* that we are examining the universe from outside as a psychological crutch to support out thoughts. We look closely to see if there are signs of causality but if we are outside we have no control over events. We are in the place of someone who does a survey and tries to establish causal relationships between things we observe. Without control any judgement about causality is subjective. We may be able to measure a correlation between certain sets of events but we have no definitive way of knowing which is cause and which is effect unless we could draw from our experience of how we think past influences future.

Does such a view of a block universe from outside make sense? It is a classical view which ignores the quantum nature of the world. In quantum mechanics it is impossible to separate observer from observed. It is difficult to know what is the significant of quantum theory to causality. There are many different interpretations of quantum mechanics and some would suggest a different answer to others. Time is an infamous problem when applied to quantum mechanics and general relativity. Without a theory of quantum gravity we cannot be sure of any response to the question.

I will adopt a position on quantum mechanics which extends the block universe metaphor. Our space-time can be cut like a sliced sausage. Each thin slice represents the universe at one moment in time and records the state of everything classically at that instant. According to physicist Julian Barbour, the quantum multiverse is a heap of slices. The heap contains all possible slices from all possible universes and is not ordered. Time and change have no absolute meaning and just represent the different ways that the slices can be put back together to make histories of the universes. Our passage through the quantum world is just one of many possible sequences which can follow from each instant. A different analogy of the same notion has been described by David Deutsch. Each slice is a snapshot of the universe. They can be put together as frames in a sequence of film which tells the story of a universe. Indeed, this is the film version of the storyteller's paradigm. Our experience of the universe is like a showing of the film, but even when the film lies in the can the universe still exists without any frame singled out as the present moment. The unordered heap of all possible frames is the multiverse.

Einstein and Minkowski taught us that space and time cannot be separated. A universe can be sliced up in different ways just as a sausage could be sliced at different angles. A natural development of the time slice analogies is to break each slice down further into small morsels. If space-time is minced up finely enough the multiverse is reduced to a heap of events. The rules which tell us how they can be put back together are the Feynman rules of quantum gravity, whatever they may be. Just like a story broken down into sentences and then words and then letters, there are fewer components each time. The finer the universe is chopped, the smaller is the heap, but each bit can be used many times and combined in an infinity of different permutations. Such a view of the universe seems to demand event symmetry. The heap is unordered and shuffling its contents has no consequence to the multiverse. It should follow that event symmetry, the symmetric group acting to permute space-time events, should be part of the universal symmetry of nature.

Where does this leave the present? At some time we all ask ourselves "why now?" What distinguishes this moment from others? Given that the universe lasts many billions of years it seems a fantastic coincidence that the present even falls within our lifetime. Of course this is nonsense. It could be no other time than "now". When we view the block universe we see all moments at a glance. There before us are all the moments when we asked "why now?" It becomes a stupid question, a trick of our psychology which has a need to know something it calls consciousness. Within the universe it is a hotly debated subject. From outside the question loses its meaning and we judge it differently. It is fortunate that we do not need to apply our philosophy of physics to our everyday lives otherwise we would lose all sense of purpose.

The Second Law of Thermodynamics

It is all very well to say that temporal causality is not absolute, but then we must explain why it is such a good illusion. How about the laws of thermodynamics? If we have a system of many particles then we cannot determine all their positions and velocities exactly. When we know only some statistical information about them they obey laws which seem not to be reversible. The second law of thermodynamics says that entropy must always increase. Could this be linked to causality?

Indeed, the continual increase of entropy is intimately linked to our perception of causality. Entropy is a measure of disorder in a system and defines a thermodynamic arrow of time which can be linked to the psychological arrow of time. There is, however, a catch again. The second law of thermodynamics is inexplicable in terms of the underlying laws of physics which, as far as we know, are reversible. This is enshrined in a theorem of relativistic quantum field theory which proves the necessity of CPT conservation.

The increase of entropy can be understood in certain idealised experiments. For example, take two closed containers filled with gases which are each in thermal and chemical equilibrium, and allow them to mix by connecting the two systems without allowing any energy to escape or enter. When the system comes back into equilibrium the entropy of the final state can be shown theoretically to be higher than the combined entropies in the two original systems. This seems to be theoretical evidence for increasing entropy and it is confirmed by experiment, but we must not be misled. The assumption that prepared systems tend towards equilibrium has been justified, but

theory would tell us that they tend towards equilibrium in the past as well as the future. We are victims of our prejudices about causality again and have devised an argument with circular reasoning to support it.

Such attempts to prove the second law of dynamics originated in the 19th century with the work of physicists such as Ludwig Boltzmann. Such a feat can never be achieved because the laws of physics are time symmetric and it is impossible to derive a time asymmetric result from time symmetric assumptions. Boltzmann slipped in some time-asymmetric assumptions in order to derive the result. Physicists have devised many other arguments for why entropy always increases, trying to get round the problem of CPT symmetry.

Here are a few possibilities:

- CPT symmetry exchanges matter for antimatter so perhaps entropy would decrease for antimatter.

  Fault: Electromagnetic radiation cannot be distinguished from its antimatter image, and yet it obeys the second law of thermodynamics.

- CPT symmetry does not apply to the collapse of the wavefunction in quantum mechanics which is a time asymmetric process.

  Query: Does this mean that the third law of thermodynamics is not valid for classical statistical mechanics?

- CPT conservation is violated by quantum gravity.

  This could be true but can the laws of thermodynamics be a result of quantum gravity whose effects are normally thought to be irrelevant except in the most extreme physical regimes?

- Entropy increases as a result of the fact that it started very low at the beginning of time. Thus it is due to the initial conditions being set in a special way, and from then on it could only increase.

  But then why were initial conditions set rather than final or mixed boundary conditions?

When I was an undergraduate student I naively thought that physicists understood entropy. Some have produced arguments based on any or all of the above possibilities. In retrospect I think now that I should be no more convinced by any of those arguments than I should if I heard someone arguing that family strife stunts the growth of children based on the correlation reported in the survey. One of the difficulties is that we do not really have an ideal definition of entropy for systems which are not in equilibrium. We can understand it as a measure of disorder in a closed system. More generally we have to resort to some kind of coarse graining process in which we

imagine that a non-equilibrium system can be seen as made of small sub-systems, or grains, which are in equilibrium themselves but not in equilibrium with each other.

Entropy might be better understood in terms of information. It can be linked to the number of bits which are needed to describe a system accurately. In a hot disordered system you need to specify the individual state of each particle, while a cold lattice can be described in terms of its lattice shape, size and orientation. Far less information is needed for the low entropy system.

The claim that entropy increases because it started low in the big bang is perhaps the one which has fallen into conventional wisdom, even if it is admitted that we do not understand why it started low. Perhaps it is because of some huge unknown symmetry which was valid at the high temperatures of the big bang and broken later. This is also my opinion but I think that if the universe were closed we would have to apply the argument in reverse at the big crunch too.

In a completely deterministic system the evolution of the system is equally well determined by its final state as by its initial so we could argue that the amount of information in the system must be constant. The difficulty there is that we are assuming an exact knowledge of state which is impossible. In any case, quantum mechanics is not deterministic. If we make a perfect crystal with an unstable isotope, as time passes some of the atoms will decay. The amount of information needed to track the decayed atoms increases. Perhaps, then, it really is quantum mechanics and the collapse of the wave function which is responsible.

If physicists used to think they understood entropy then their faith was deeply shaken when Stephen Hawking and Jacob Bekenstein discovered that the laws of thermodynamics could be extended to the quantum mechanics of black holes. The entropy is given by the area of the black hole but its temperature can only be understood through quantum mechanical effects. This shows that classical understanding of thermodynamics is indeed incomplete and perhaps only a complete theory of quantum gravity can explain the laws fully.

Could the Universe be Gold?

In the mid 1960s there was a widely held belief that the universe should be closed. The simple homogeneous cosmological models can describe a space which is finite in size, curving gradually so that it eventually joins back on itself like the surface of a sphere. Time would start at the big bang from where it expands for many billions of years. Eventually, according to the equations of general relativity, gravity must arrest the expansion and it will contract again like a ball falling back to Earth towards its final crunch.

At present the universe is certainly expanding, as demonstrated by Hubble in 1929 when he started measuring the red-shifts of far away galaxies and correlating them to their distance. This defines a cosmological arrow of time which distinguishes past from future. In 1962 J. E. Hogarth suggested the possibility that this cosmological arrow could be linked to the thermodynamic arrow of time. Thomas Gold proposed that when the universe starts to contract the increase of entropy might reach a turning point. As the universe collapses history would run in reverse.

Needless to say, Gold's model of the universe is quite controversial. Intuition suggests that the arrow of time cannot change direction. It would be a complete reversal of causality with events being determined by the future instead of the past. In 1985, Stephen Hawking unexpectedly came out in support of Gold. He published a paper demonstrating that a time reversal was to be expected because the physics of the final crunch must be the same as the physics of the big bang. We might try to understand the quantum state of the entire universe by using Feynman's path integral formulation of quantum mechanics. We must form a sum over all possible space-time manifolds allowed in general relativity. Hawking has argued that we can understand entropy in this way if the universe is an entirely closed system, finite in both time and space but with no boundary. There would be no initial or final conditions to worry about, and both the end and start of the universe would be a consequence of the same laws of physics which are obeyed at all times. If the laws of physics are time reversal invariant we should then expect the end to be like a reversed playback of the beginning.

Before Hawking's paper had passed through the publishing process he was already under pressure to change his mind. His colleagues Laflamme and Page set out to convince him that he had made an error. Before the paper went to press they succeeded and he added a note to the paper admitting his mistake. He now claims that there are two possible ways a universe could start or end. One has low entropy the other high. The only consistent picture is one in which it is low at one end and high at the other hence temporal symmetry is broken.

If this argument could be made solid then it would be a powerful one. The path integral formulation avoids problems of time since it is a sum over all possible universes rather than an evolution with separate boundary conditions. However, Hawking's method uses an incomplete semi-classical description of quantum gravity. The argument could only be made complete when we understand quantum gravity better. Until then it is an open question whether or not a closed cosmological model will have a time reversal at half time or not.

There remain very few scientists who have argued in favour of a Gold universe and stuck to it. Most cosmologists have sought reasons to rule it out and have often claimed success. As the philosopher Huw Price has shown, most of those arguments are based on double standards of reasoning. Often time asymmetric conclusions are drawn from time symmetric assumptions. This is just about impossible unless there is some spontaneous symmetry breaking such as that proposed by Hawking.

Intuition suggests that the arrow of time could never reverse. If we could meet other intelligent life-forms who are evolving in reverse, many paradoxes would present themselves. Their past would be our future. What would there be to prevent them from telling us about events in our future? Suppose we decided they were a threat and decided to destroy them. If we succeeded they would cease to exist in their own past. What is to prevent us from bringing about such a paradoxical situation?

The only reasonable answer must be that the arrow of time will only reverse when we are long gone and other time-reversed life-forms are not there either. In other words, the epoch in which the universe will reverse its collapse must be lifeless. Some people already find it hard to accept that the human race must be extinguished at the big crunch. To suggest that we cannot even

survive for half as long even when there is no such catastrophe to wipe us out seems almost unthinkable. After many trillions of years the stars will have faded. The universe will be a cold place, hard to live in with so few sources of energy. Could we not at least hope to build a powerful computerised automaton which could be programmed to hibernate through the aeons, using the least power possible to steer away from black holes and other places where it would be destroyed? If so it would be able to take a message of our past into the future? In the collapsing universe it might revive and deliver a message to the anti-thermodynamic inhabitants of the other half of space-time. Sadly the answer must be no since it would create unresolvable paradoxes, but unless we can explain what would stop it we must give up the possibility of a Gold universe.

Anti-thermodynamic light from the future

Although such reasoning may be what motivates disbelief in reversal of time's arrow, most attempts to rule out the Gold universe have concentrated on arguments which may be simpler and more certain. Physicists such as Murray Gell-Mann have asked about the fate of starlight. We know that starlight can cross the universe for billions for years without being absorbed. Each photon loses energy as it is red-shifted by the expanding universe but still it can continue with only a very small chance of hitting another particle. As the universe expands the matter becomes more thinly spread. The chance of a collision grows smaller. According to a calculation by Jason Twamley and Paul Davies in 1995, a photon which heads out into space has only a small probability of being lost no matter how long the universe lasts before it arrives in the collapsing universe. If that is so then most of the light being emitted by stars now will be present in the collapse.

Conversely, the time-reversed stars of the future will absorb photons because they are time reversed. Those photons should be around now. Could we see them?

Gell-Mann believes that if they are there they could be detected. He says that they would add to the background light of the universe which could be measured. If the light is not there a Gold universe might be ruled out. Huw Price pointed out that it is not so simple. The light from future stars cannot be detected simply by looking at the night sky with a telescope. These photons would be heading for a time reversed star in the future. If you block their passage with any kind of detector such as a photographic film they will simply not be there because they would not then be around in the future. Their behaviour is distinctly acausal. According to Price they would be invisible by ordinary means.

If you hold up a piece of paper in space. Photons of future starlight would not be absorbed. Instead they would be emitted as if they were being drawn out of the paper by a future cause. You are probably thinking that all this is already just too absurd to be possible anyway, but you must suspend your disbelief until a contradiction with either logic or observation has been reached. Light drawn off a surface like this would not register in the ordinary way. It is actually quite difficult to predict what would really happen because the photons are acausal and the paper is not. Would the effect of the photons be detected before or after they are emitted? Despite such logical difficulties we know that energy must be conserved what ever happens. This means that energy will be drawn off the paper. It should be detectable in principle.

It is not absolutely clear whether or not observation can already rule this out but I think they can. The anti-thermodynamic radiation would be present at many wavelengths. Light photons may be difficult to detect in this way but radio waves would be likely to affect radio telescopes and gamma rays would also surely leave their mark. Above all an anti-thermodynamic cosmic background radiation destined for the big crunch would be similar in energy and temperature to the cosmic background radiation from the big bang. Instead of imparting heat to a detector it would take it away. The net effect of both the big bang and big crunch radiation would be no heating. Yet the heat of the cosmic background was detected by Andrew Mckellar in cosmic cyanogen as long ago as 1941 even before its significance was recognised.

A Crystal Ball

There is another reason why we should suspect that anti-thermodynamic radiation is not present in the universe today: If it was, we would be able to use it to send messages back in time.

When you hold up your hand to light it casts a shadow behind it. Even faint starlight casts such a shadow. What about our anti-thermodynamic light from the future? If you could expose your hand to anti-thermodynamic radiation you expect it to have photons drawn off it destined for some anti-thermodynamic star in the distant future as the universe collapses. Radiation would surround your hand but instead of casting a shadow behind the direction the light is travelling, there would be a kind of anti-shadow in front of it from the direction the radiation is coming. This is simply because light in front of your hand is blocked in its passage towards its destiny.

If you move your hand in front of a lamp, the shadow moves with it. Because of the finite speed of light there is always a slight delay and the movement of the shadow lags behind the movement of your hand by an imperceptible amount. The anti-shadow cast by anti-thermodynamic light behaves differently. It is not difficult to see that it must move ahead of the hand, anticipating every move by the instant of time it would take the light to travel from the shadow to the hand.

This effect could be used in principle to send messages back in time. To do it effectively the distance from the hand to where the shadow was cast would have to be made large. A mirror could be used to reflect the shadow from a long distance away back to a point near where the hand is moving. By detecting the anti-shadow you could see what your hand is about to do. You could literally use hand signals to send messages into the past. It is difficult to see how the paradoxes presented by such a phenomenon could be avoided unless anti-thermodynamic light is invisible, but as I have already argued, it should be detectable. Either anti-thermodynamic light is not available to us or we will have to face up to these paradoxes.

Mixing or Meeting

The arguments I have presented so far have made the assumption that a Gold universe would contain a mix of what we have been calling anti-thermodynamic matter (or radiation) with ordinary thermodynamic matter. For anti-thermodynamic matter the arrow of time is reversed and its behaviour is affected by future causes. This is the opposite of the more familiar

thermodynamic matter for which cause precedes effect. Thermodynamic and anti-thermodynamic matter might co-exist in the present universe.

There is an alternative way in which a Gold universe might work. It could be that thermodynamic and anti-thermodynamic matter and radiation never mix. Instead they might meet in the middle of time when thermodynamic matter may slowly transform into anti-thermodynamic matter. Thermodynamic matter would only be present in the expanding half of the universe and anti-thermodynamic matter would only be present in the collapsing half .

Think again about the electromagnetic radiation. Remember it was argued that light left over from the stars in the expanding universe and the cosmic background radiation would survive into the collapsing half. It was assumed that this radiation would be randomly dispersed so that it would strike any objects that are around during the collapse. However, this assumes that each photon is causally influenced only by its dim and distant past, never the future. On reflection this is not what would be most probable. It is more likely that the radiation would fall under the influence of its destiny if the collapse is anti-thermodynamic. In that case the photons which are radiated from stars now and pass into the collapsing phase of the universe will be the same photons which are anti-thermodynamically absorbed by the anti-thermodynamic stars in the collapse. If this were to be the case then there would likewise be no anti-thermodynamic radiation from the future around now and we would not be able to send paradoxical messages back in time. There would be no inconsistency.

You might think that a huge coincidence would be required for all the photons emitted by stars now to conspire to fall onto anti-thermodynamic stars in the future, but the whole point is that a low entropy phase of the universe already appears as a fantastic statistical fluke. This comes about because the initial and final state force it to happen and the rest of time has to cope with it. It drives evolution and other acts of the universe which would otherwise seem highly improbable. A fluke such as photons travelling through the aeons and hitting an anti-thermodynamic star must be weighed against the equally unlikely events which must happen if it hits a cold anti-thermodynamic surface.

Matter and Anti-matter

I have been saying a lot about anti-thermodynamic matter and radiation and you might have been wondering if it is related to anti-matter. They are certainly not the same thing because there is no distinction between photons and anti-photons yet we can talk about thermodynamic radiation and anti-thermodynamic radiation in terms of whether they are causally effected by the past or future.

Substance made out of protons, electrons and neutrons is a different matter. Time reversal (T) alone is not an exact symmetry of nature but if we combine it with charge conjugation (C) and parity inversion (P) we do get an exact symmetry called CPT. This operation effectively exchanges matter and anti-matter. In 1967, Andrei Sakharov found a way to account for why the universe is dominated by matter with very little anti-matter. It is due to the slight CP violating effects in the nuclear forces. In the heat of the big bang these would have been significant enough to account for the imbalance left over from the first instants. If this is correct then a similar effect must apply in reverse at the big crunch which we are assuming is anti-

thermodynamic. The alarming consequence is that the collapsing phase shall be dominated by anti-matter.

It is going to be more difficult to explain how thermodynamic matter can transform into anti-thermodynamic anti-matter somewhere around the middle of time because CP violating effects are improbable at low temperatures. If the universe lasts long enough the problem will be resolved because protons can decay to produce positrons and then the electrons can anti-decay to make anti-protons. But the half life of this process is at least 10^{32} years, so unless the universe is set to live much longer than that there is a problem. Proton decay could be forced to happen as the statistically least costly way of making the transformation but if so it would probably be happening already. Experiments which try to detect proton decay say otherwise.

A second possibility is that all the matter falls into black holes where matter is indistinguishable from anti-matter. The anti-matter would then have to emerge from white holes in the reverse fashion. This brings us to the next problem. Where are the white holes? Unless the universe is going to go on long enough for all the protons to decay we will need them. Even if it *is* going to go on long enough for the protons to decay, there are other particles such as neutrinos which may never reach an equilibrium state with an equal mix of particles and anti-particles. Only photons and other particles which are their own anti-particles can be guaranteed to carry over from the expanding phase to the collapsing phase without spoiling the time symmetry.

Black Holes, White Holes.

If black holes can solve the matter to anti-matter problem, they themselves may present a greater problem. It is a fundamental property of black holes in classical general relativity that they swallow up matter which can never escape again. They can only get bigger and bigger. This is the second law of black hole thermodynamics. How then, can the black holes which form from collapsed stars and galaxies in the expanding universe be reconciled with an anti-thermodynamic collapsing universe?

The gravitational field equations of classical general relativity are symmetric under time reversal just as for all the other forces. To complement black holes there can also be white holes which are the time reversal of black holes. Just as black holes swallow matter, always get bigger and can never be destroyed, white holes can release matter, always get smaller and can never be created. If black holes survive after the first half of the history of the universe as the classical theory says they must, then a Gold universe must likewise contain white holes which are their time reversal. Those white holes would have to be out there now and must have been already there at the big bang, even though the true cause of their creation is in the future.

The white holes would be dormant, waiting for the distant future when their destiny will be to release all the anti-thermodynamic anti-matter which makes up the anti-stars of the collapsing universe.

There seem to be some probable inconsistencies in this scenario. We should be able to detect those white holes because they will act as gravitational lenses even if they are alone in deep intergalactic space. Astronomers are increasingly finding that black holes are common and that

they range in size from a few solar masses up to billions of solar masses. The white holes would have to be at least as common and as big. We do see gravitational lenses but they appear to be due to ordinary galaxies and it already seems unlikely that we can account for so many white holes in the universe. There are other conceptual problems if white holes are around. What if they were to collide with ordinary stars, galaxies or even dust. White holes must attract ordinary matter yet it is not supposed to be able to fall in. Dormant white holes would be very paradoxical objects, especially if we could locate them. The difficulties would be even greater in the early universe where they would inevitably have had a significant influence.

It begins to look like we have finally found a likely contradiction which would rule out a Gold universe, but once again we have only considered the mixing solution for black and white holes. Could there be a better meeting solution as there seems to be for radiation and matter? The only way out would be if black holes could somehow transform gently into white holes. Then there would be no need to account for white holes in the universe now. The black holes which are being discovered all over the universe now, would transform into the complementary white holes which will have to be around in the collapsing universe.

The transformation of black holes into white holes is not easy to understand. In classical physics it simply cannot happen. In quantum mechanics the situation is a little different. According to Hawking, black holes radiate and *can* lose mass. When Hawking considered the possibility of a Gold universe he considered whether it would be possible for the transformation to happen. A black hole would become quiet when all the matter around it had been pulled in. It could gently radiate but any black hole of the size we have found them to be would be too large to radiate significantly. How could it switch to throwing out matter like a white hole?

As a matter of fact, a dormant black hole would be virtually indistinguishable from a dormant white hole from an external point of view. The gravitational field around them is the same. Only internally are they different. Hawking argued that if a black hole comes into thermal equilibrium with the radiation that surrounds it, so that it radiates the same as it takes in, then it should be in a time symmetric state. This leaves open the possibility that the transformation could take place. From outside the black hole Hawking radiation would just appear to get stronger until what was a black hole is behaving just like a white hole. What would happen internally? A black hole has an internal singularity which lies in the future of anyone who falls in, whereas a white hole must have one in the past from which any outgoing matter originates.

H- Dieter Zeh is one physicist who has continued to study this possibility. Matter which fell into the black hole would seem to be frozen on the event horizon from the point of view of someone who stays outside. Zeh has suggested that quantum effects could simply cause it to turn round and come back out again. The black hole singularity would never form. Unfortunately it is difficult to envisage how the dynamics could work. The curvature at the event horizon of a large black hole is slight and quantum effects should be small. From the point of view of what we are trying to imagine here there is an even worse problem. We were going to claim that the matter which fell into the black hole would later re-emerge as anti-matter from the white hole, but if it is the same matter which turns round and heads back out it cannot change from matter to anti-matter.

It is interesting that Stephen Hawking still believes that black holes and white holes are identical when they are very small. Such virtual quantum black/white holes must be part of the vacuum but they are very different from the macroscopic ones which form from collapsed stars. They would be more like elementary particles and may even turn out to be the same thing as particles when we understand quantum gravity. It would be extraordinary if large black holes could also be identified with white holes. They would have to have both a future and past singularity.

As it happens, the classic static model of a black hole found by Schwarzschild does have a future and past singularity, but a more realistic model of a black hole which formed from a collapsing star cannot have such time symmetry in classical general relativity. If it is possible for it to happen when quantum gravity effects are taken into account it will be very different from what we expect classically. Yet, despite the strangeness of the idea, the possibility cannot be ruled out. The closest description of what it would be like in the language of physics we understand would be that the inside of a black hole would be a quantum supposition of the wave functions of a black hole and a white hole.

The black hole complementarity principle proposed by physicists considering the information loss problem gives further hope to the possibility that a black hole can transform into a white hole. The principle asserts that there is no inconsistency between the point of view of an observer who falls past the event horizon of a black hole towards its singularity and another observer outside who sees him stop at the horizon and eventually return as thermal Hawking radiation. If this is true then we should also accept that there is no inconsistency if there is a third observer who emerges from the event horizon as if it were a white hole too. It is as if the event horizon were a cross-roads in time.

The Shape of Things to Come

I have put together a picture of a Gold universe in which a closed universe expands from a big bang then collapses towards a big crunch. The collapsing phase will be like the expanding phase only in reverse. Galaxies, stars and planets will be made of anti-matter and will absorb light and other radiation rather than emitting it and will run their history in reverse. Life would also evolve backwards driven by a decreasing entropy unlike the increasing entropy of the expanding phase.

The sources of low entropy are both the initial and the final singularity of the universe. Thus it has two origins. Entropy follows its natural statistical tendency to increase away from those origins where some unknown principle of quantum gravity must be responsible for the extraordinary low entropy. Although life evolves backwards, intelligent life in the collapsing phase will have experiences similar to ours. Their future is our past and they can find no record of it.

The light radiation from our thermodynamic stars, as well as the cosmic background radiation which fills space today, will survive into the collapsing phase. It will gradually transform from being thermodynamic to being anti-thermodynamic. All matter made of particles with mass is most likely to fall into the black holes which are the dead remnants of stars and galaxies. Even neutrinos must follow such a fate, which may only be possible for them if they have a small

mass. The black holes themselves will slowly transform to white holes from which the anti-thermodynamic matter of the collapsing phase emerges.

Perhaps the most difficult part of this vision for us is the fate of ourselves and other life. It cannot survive until the collapse or even leave any reminder of its past. Otherwise there might be a paradoxical mixing of thermodynamic and anti-thermodynamic life. The universe will see to it that this does not happen and its job will certainly be made easier if the universe grows to a very old age before the expansion stops.

Wider Perspectives

The universe can only be as Gold proposed if it is finite and closed. This used to be the preferred model of theoretical cosmology. Cosmologists favoured a universe which is finite in space and time mostly for philosophical reasons. These days they are generally more open minded. Still it is most common to read about the standard homogeneous cosmologies which were first worked out by Alexsandr Friedmann in 1922. These can be either open or closed. The closed case corresponds to the geometry of the Gold universe but the open one is asymmetric in time. There is a single big bang from which the entire universe emerges and then expands forever. Space is infinite and time is indefinite into the future. There would be no need for any time reversal in such a universe.

The question of homogeneity has always been a controversial one in cosmology. In 1933 just a few years after Hubble had shown that the universe is expanding, Arthur Milne proposed homogeneity as a *cosmological principle*. It is certainly a convenient principle because homogeneous models of the universe are much easier to analyse, but why should we believe it is true? Even in the 1930s Fritz Zwicky was arguing for the presence of galactic clusters in the cosmos, evidence for less homogeneity than others wanted. In 1953 Gérard de Vaucouleurs also produced evidence for large scale structure but still most were sceptics. In the 1980s when detailed maps of the distribution of galaxies were produced the doubters had to concede. There are huge voids and walls on scales which extend to a significant fraction of the size of the observable universe.

Our measurements of the cosmic microwave backgrounds show a high degree of isotropy and this is taken as proof that the universe is homogeneous on larger scales. Our observation is limited by a horizon defined by the age of the universe and the speed of light. Thus we cannot observe anything beyond about 15 billion light years distance. Why should we imagine that the size of the universe is a similar order of magnitude to its current age? We have been unable to measure the extent to which space is curved and cannot place limits on its size.

Martin Rees has compared our view of the universe with a seascape as seen from a ship in the middle of the ocean. As far as the eye can see it seems unchanging except for the waves which we see at close range. The view is limited to the horizon and beyond who knows what there is. It seems to be only an application of Occam's razor which justifies the assumption that space is homogenous on scales hundreds of orders of magnitude larger than the observable horizon.

Occam's Razor

Occam's (or Ockham's) razor is a principle attributed to the 14th century logician and Franciscan friar; William of Occam. Ockham was the village in the English county of Surrey where he was born. The principle states that **"Entities should not be multiplied unnecessarily."** Sometimes it is quoted in one of its original Latin forms to give it an air of authenticity.

"Pluralitas non est ponenda sine neccesitate"

"Frustra fit per plura quod potest fieri per pauciora"

"Entia non sunt multiplicanda praeter necessitatem"

In fact, only the first two of these forms appear in his surviving works and the third was written by a later scholar. Many scientists have adopted or reinvented Occam's Razor. Isaac Newton stated the rule: "We are to admit no more causes of natural things than such as are both true and sufficient to explain their appearances."

The most useful statement of the principle for scientists is, **"When you have two competing theories which make exactly the same predictions, the one that is simpler is the better."**

In physics we use the razor to cut away metaphysical concepts. The canonical example is Einstein's theory of special relativity compared with Lorentz's theory that ruler's contract and clocks slow down when in motion through the Ether. Einstein's equations for transforming space-time are the same as Lorentz's equations for transforming rulers and clocks, but Einstein and Poincaré recognised that the Ether could not be detected according to the equations of Lorentz and Maxwell. By Occam's razor it had to be eliminated.

But the non-existence of the ether cannot be deduced from Occam's Razor alone. It can separate two theories which make the same predictions but does not rule out other theories which might make a different prediction. Empirical evidence is also required and Occam himself argued for empiricism, not against it.

Ernst Mach advocated a version of Occam's razor which he called the Principle of Economy, stating that **"Scientists must use the simplest means of arriving at their results and exclude everything not perceived by the senses."** Taken to its logical conclusion this philosophy becomes positivism; the belief that what cannot be observed does not exist. Mach influenced Einstein when he argued that space and time are not absolute but he also applied positivism to molecules. Mach and his followers claimed that molecules were metaphysical because they were too small to detect directly. This was despite the success the molecular theory had in explaining chemical reactions and thermodynamics. It is ironic that while applying the principle of economy to throw out the concept of the ether and an absolute rest frame, Einstein published almost simultaneously a paper on Brownian motion which confirmed the reality of molecules and thus dealt a blow against the use of positivism. The moral of this story is that Occam's razor should not be wielded blindly. As Einstein put it in his autobiographical notes:

> "This is an interesting example of the fact that even scholars of audacious spirit and fine instinct can be obstructed in the interpretation of facts by philosophical prejudices."

Occam's razor is often cited in stronger forms than Occam intended, as in the following statements...

"If you have two theories which both explain the observed facts then you should use the simplest until more evidence comes along"

"The simplest explanation for some phenomenon is more likely to be accurate than more complicated explanations."

"If you have two equally likely solutions to a problem, pick the simplest."

"The explanation requiring the fewest assumptions is most likely to be correct."

... or in the only form which takes its own advice...

"Keep things simple!"

Notice how the principle has strengthened in these forms which should be more correctly called the law of parsimony, or the rule of simplicity. To begin with we used Occam's razor to separate theories which would predict the same result for all experiments. Now we are trying to choose between theories which make different predictions. This is not what Occam intended. Should we not test those predictions instead? Obviously we should eventually, but suppose we are at an early stage and are not yet ready to do the experiments. We are just looking for guidance in developing a theory.

This principle goes back at least as far as Aristotle who wrote **"Nature operates in the shortest way possible."** Aristotle went too far in believing that experiment and observation were unnecessary. The principle of simplicity works as a heuristic rule-of-thumb but some people quote it as if it is an axiom of physics. It is not. It can work well in philosophy or particle physics, but less often so in cosmology or psychology, where things usually turn out to be more complicated than you ever expected.

Simplicity is subjective and the universe does not always have the same ideas about simplicity as we do. Successful theorists often speak of symmetry and beauty as well as simplicity. Paul Dirac said that if requirements for simplicity and beauty clash we should strive for mathematical beauty first and simplicity second. The law of parsimony is no substitute for insight, logic and the scientific method. It should never be relied upon to make or defend a conclusion. As arbiters of correctness only logical consistency and empirical evidence are absolute. Dirac was very successful with his method. He constructed the relativistic field equation for the electron and used it to predict the positron. But he was not suggesting that physics should be based on mathematical beauty alone. He fully appreciated the need for experimental verification.

The final word falls to Einstein, himself a master of the quotable one liner. He warned,

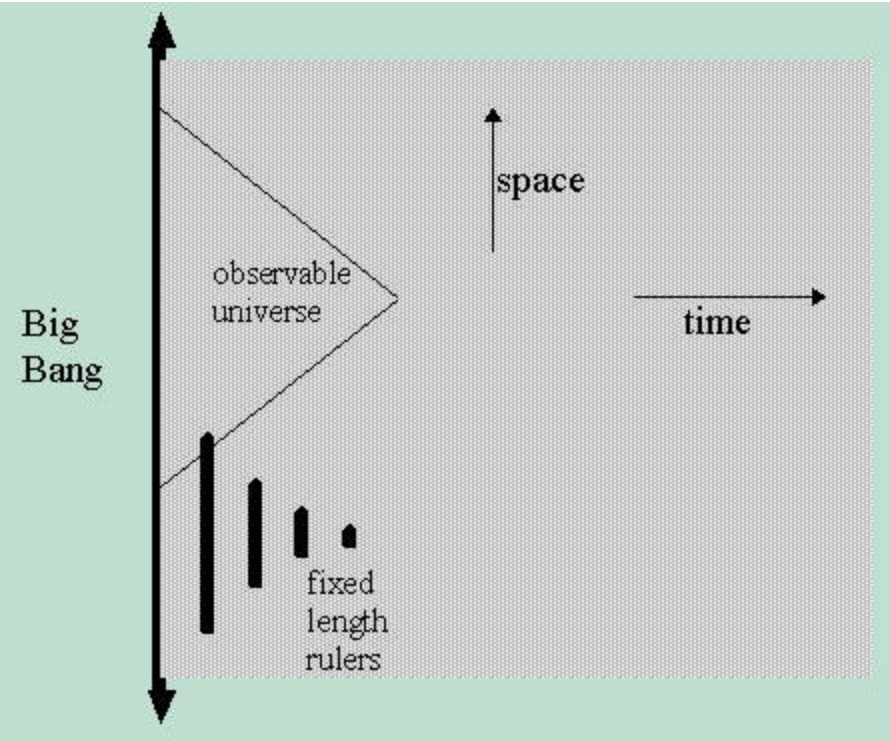
"Everything should be made as simple as possible, but not simpler."

An Inhomogeneous Universe

If there is enough matter in the expanding universe space will have a positive curvature and the expansion will be slowing down. Eventually it will stop and start to recollapse. If the best observational data we have is taken at face value there is not enough matter and the universe will continue to expand. It used to be thought that there would be sufficient unseen dark matter to place the universe near the critical point between eventual collapse and continued expansion but a series of indirect observations now seems to indicate otherwise. Unless further corrections change the situation again we must now assume that the universe is not the simple closed cosmology.

Now cosmologists are turning to the open homogeneous cosmologies as the most likely model of our universe. Time starts at a big bang singularity and space is infinite from that moment onwards. The observable universe is a small finite part of the whole universe which lies inside the light cone traced back to the big bang. In the diagram below, the size of the observable universe appears bigger near the singularity but this is not an isometric diagram. In fact the universe is expanding as illustrated by the sequence of fixed length rulers which get smaller with time just as a scale gets smaller on a flat map of the world with increasing distance from the poles.

The net result is that the size of the observable universe shrinks to zero near the horizon even though the whole universe remains infinite.



This model of the universe poses some paradoxes. The singularity appears as a region of infinite extent yet it is everywhere uniform and flat. There is nothing mathematically inconsistent about such a universe and it does not come into contradiction with any known laws of physics, but is it a reasonable model of the universe? The uniformity suggests a difficult horizon problem: How is it co-ordinated over the infinite extent of the universe just an instant after the big bang. In a finite closed universe the horizon problem can be explained away by invoking inflationary theories, but no matter how rapidly the universe may have expanded in the first instants you cannot explain correlations over unlimited distances.

One possible way to explain this homogeneity would be Penrose's *Weyl curvature hypothesis*. This suggests that there is some physical law which applies to singularities and ensures that the Weyl part of the curvature tensor must be zero there. That would be sufficient to resolve the problem and it is quite possible that it could be a consequence of the unknown theory of quantum gravity which is significant at the singularity. However, the singularities which form in black holes cannot be subject to the same law since black holes are finite in size. The only known distinction between black hole singularities and the big bang is that the former always sits in the future light cone of all observers while the latter is in the past. A law which applies to one and not the other would have to break CPT invariance. Penrose has conjectured this possibility but the favourite theories of quantum gravity like superstring theory are all CPT invariant. What is the solution to this puzzle?

In truth there are several acceptable resolutions, but which is the most reasonable? How should Occam's razor be applied here? We could postulate two physically different types of singularity for the big bang and black holes to keep the simplest homogeneous model alive, or we can break CPT, or we can discard the homogeneous universe. In my opinion the last of these is the

preferred but this is just my philosophical prejudice. I would like the universe to be symmetrical in time. It does not have to be so clearly symmetrical in shape as the Gold universe. It may have a random distribution of regions where time's arrow points in different directions and others where the absence of matter or thermal equilibrium makes the direction of flow indeterminate. All this must be happening far beyond our currently observable horizon. This description of reality fits best the storyteller's paradigm since it means the universe is more diverse. Of course the universe has no obligation to satisfy anyone's philosophical preferences but it is at least worth while exploring this possibility. A future unified theory may be able to tell us what the universe is like on very large scales, but it might equally well remain an unanswerable question.

Is The Big Bang a White Hole?

When people hear about the big bang theory they often ask "Where is its centre?" The standard answer is that it has no centre because it is expanding uniformly everywhere. In giving this answer cosmologists are forgetting about alternative models which Georges Lemaître first discovered in 1927 when he developed Friedmann's original work into the big bang theory. Lemaître found solutions to the equations of general relativity which are centred on a point in space. They are inhomogeneous spherically symmetric models of the universe which have been rediscovered many times since, but they are rarely considered as plausible cosmological models on very large scales.

The time reversal of Lemaître's models can also describe the formation of a black hole from a pressureless, spherically symmetrical, non-rotating cloud of dust. A particular case of this was studied by Oppenheimer and Snyder in 1939. A sphere of dust is uniform in density with empty space outside. The dust sphere collapses to form a black hole. The interesting thing about this solution of the equations of gravity is that the geometry inside the sphere is identical to the standard homogenous cosmology of Friedmann except that it runs in reverse. The lesson to be learnt from this is that the same model in reverse is a possible model of the big bang. It looks identical to the standard homogeneous big bang within a region which might cover the whole observable universe.

In other words, the big bang could be a white hole which is indistinguishable from the standard cosmological models for restricted observers such as us. Lemaître's solutions were more general than this. The density of the dust could vary gradually away from the centre, but so long as the variation was gradual this could describe the universe with our observable universe being one small region well within the event horizon.

The idea that the big bang may be a white hole is not popular with many serious cosmologists. One reason is that classically white holes cannot form. Since I have discounted causality I can accept the possibility of a white hole as easily as I can a black hole. Indeed, the white hole could also be a black hole in accordance with Hawking's complementarity. Once it was thought that the universe consisted of just our galaxy which had a centre and no stars outside a certain limit. Now I am suggesting that the big bang could be a similarly isolated object on a much larger scale. Just as our galaxy turned out to be one of many, so too may the big bang.

It is quite possible, as far as we can tell, that the big bang is actually just a huge white hole which formed in a larger universe. Perhaps on some huge scale there is a population of black and white holes of vastly different sizes. What does that say about the laws of thermodynamics? We can expect that inside a very large white hole time's arrow is flowing away from the singularity as we observe in our neighbourhood of the universe. The opposite can be expected in a very large black hole. The big bang is represented by a large object which is both a black hole and a white hole with time flowing outwards in both directions which we would call past and future. There might be many such objects in the universe. Within them there are smaller black holes which form from collapsing stars. These will eventually emerge from the large white hole and may subsequently fall into another large black hole. Then their arrow of time will reverse as they become white holes.

According to this model black holes always become white holes as the arrow of time reverses yet there are two distinct possibilities. For small black and white holes the arrow of time always flows in, while for large ones it always flows out. This is not inconsistent. The arrow of time must be most strongly influenced by the largest singularity in the past light cone. The full explanation will have to await a more unified theory of physics. The effects of quantum gravity near a singularity must determine the extent of its homogeneity and low entropy. Over all the universe is not governed by temporal causality. Time flows in both directions. For example, the near flatness of the universe near the big bang is due to influence from the future, not the past.

Occam's razor does not have a very good track record in cosmology. Usually space turns out to contain more complexity than we imagined before we looked. It will be billions of years before we are able to see beyond the current horizon defined by the speed of light. In the shorter term, theory is our only hope to know what the structure of space-time is like on very large scales.

Time Travel

Apart from entropy there are other aspects of causality. We know that in general relativity causal effects are limited by the light cones which are part of the geometry of space-time. But the geometry is itself dynamic. In general relativity it is possible to construct space-time models which have closed time-like paths. If such things really exist in the universe we would be able to travel back to our past.

Traditionally physicists have simply said that such universes must be ruled out because if we could travel back to our past we could change our history, which seems to raise contradictions. Recently some physicists have started to question this assumption. It seems possible that quantum mechanics may allow closed time like curves through space-time wormholes to be constructed, at least in principle. The contradictions which were thought to be a consequence of time travel do not stand up to close examination.

Perhaps it would be possible to travel back to the past and see our parents but some chance event would prevent us from being able to change their lives in ways which we know never happened. If that is a correct interpretation then it attacks our faith in our own free will.

There is perhaps little that we can conclude reliably about causality from our current understanding of physics. Only when we have found and understood the correct theory for quantum gravity will we be able to know the truth. We may be prevented from finding that theory if we hold fast to our prejudices.

The Superstring Supermystery

Everything or Nothing?

In 1984 Michael Green and John Schwarz made a discovery which might turn out to be the greatest advance in physics of all time, if it is right. They found that a particular quantum field theory of supersymmetric strings in 10 dimensions gives finite answers at all orders in perturbation theory. This was a breakthrough because the superstring theory had the potential to include all the particles and forces in nature. It could be a completely unified theory of physics. By 1985 the press had got hold of it. Articles appeared in Science and New Scientist. They called superstrings a *Theory Of Everything*.

Following the media reports about string theory there was an immediate backlash. People naturally asked what this *Theory Of Everything* had to tell us. The answer was that it could not yet tell us anything, even about physics, yet. On closer examination it was revealed that the theory is not even complete. It exists only as a perturbation series with an infinite number of terms. Although each term is well defined and finite, the sum of the series will diverge.

To understand string theory properly it is necessary to define the action principle for a non-perturbative quantum field theory. In the physics of point particles it is possible to do this at least formally, but in string theory success has evaded all attempts. To get any useful predictions out of string theory it will be necessary to find non-perturbative results. The perturbation theory simply breaks down at the Planck scale where stringy effects should be interesting.

More bad news was to come. Systematic analysis showed that there were really several different ten-dimensional superstring theories which are well defined in perturbation theory. If you count the various open and closed string theories with all possible chirality modes and gauge groups which have no anomalies, there are five in all. This is not bad when compared to the infinite number of renormalisable theories of point particles, but one of the original selling points of string theory was its uniqueness. Worse still, to produce a four-dimensional string theory it is necessary to compactify six dimensions into a small curled up space. There are estimated to be many thousands of ways to do this. Each one predicts different particle physics. With the *Heterotic string* it is possible to get tantalisingly closed to the right number of particles and

gauge groups. At the moment there are just too many possibilities and the problem is made more difficult because we do not know how the supersymmetry is broken.

All this makes string theory look less promising. Some critics called it a *theory of nothing* and advocated a more conservative approach to particle physics tied more closely to experimental results. Yet a large number of physicists have persisted. There is something about superstring theory which is very persuasive.

Why String Theory?

The most commonly asked question from the public about string theory is *Why?* To understand why physicists study string theory rather than theories of surfaces or other objects we have to go back to its origins. The first person to consider string theories was Paul Dirac in 1950. Dirac had a way of doing physics which few others managed so well. His motto was that "mathematics can lead us in a direction we would not take if we only followed up physical ideas by themselves." The whole idea of it will seem crazy to most people who have not seen this principle at work, but many theoretical physicists now practice the same technique.

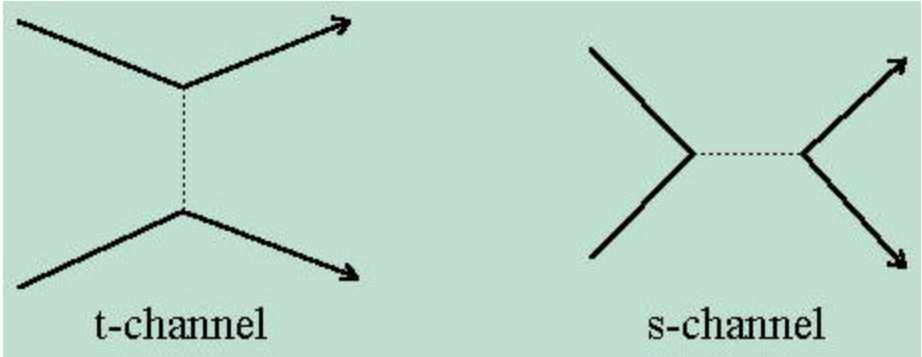
In 1950 it was known that physics holds fast to solid principles including the principle of relativity, causality and the quantum version of the principle of least action. These impose very tight mathematical constraints on the kind of theories you can build. One day those principles may be superseded but it is not easy to modify them without destroying the successes of the past. You cannot just replace linear quantum mechanics with some non-linear version and expect it to make sense, nor can you break the symmetries of relativity without invalidating the whole thing. There is more sense in thinking about how physical theories can be generalised within these principles and that is what Dirac was doing.

At the time particle physics was understood in terms of quantum field theory derived from quantised interaction of point particles. There is very limited scope for relativistic theories of this type which are renormalisable. We now know that Yang-Mills theory with spin half and spin zero particles with a few possible interaction terms is all that is permitted. Dirac considered the possibility that more general theories might start from string-like and membrane-like objects rather than point particles. It may seem like a wild idea but actually there is not much else you can do without revising our concepts of space-time or quantum mechanics. As a mathematical problem in its own right you can study the full class of possible theories of p-dimensional surfaces, known as p-branes moving in D-dimensional space. 0-branes are just particles, 1-branes are strings and 2-branes are membranes. You can work out all the ways these objects might interact which are consistent with relativity and then try to work out which of those can be consistently quantised and which are consistent with causality. The final step would be to see which of the remaining possibilities matches the real world. It is an ambitious program which is far from easy to complete.

As it turned out Dirac's ideas about strings and membranes were forgotten and history delivered string theory by a less direct route. In 1968 physicists were trying to understand the nature of the strong nuclear interactions which held the quarks together in nucleons. It was by no means clear that quantum field theory was adequate to solve the problem. Even the quark hypothesis was not

universally accepted although experiments were just beginning to see signs of their effects. One way to tackle the problem was to work directly with the matrix of scattering amplitudes, the S-matrix, which describes how hadronic particles interact. Instead of trying to derive it from some underlying field theory it could be considered fundamental. The rules of quantum mechanics and relativity restrict the S-matrix to satisfy a set of equations. It was hoped that a few more additional principles might pin it down to some unique form.

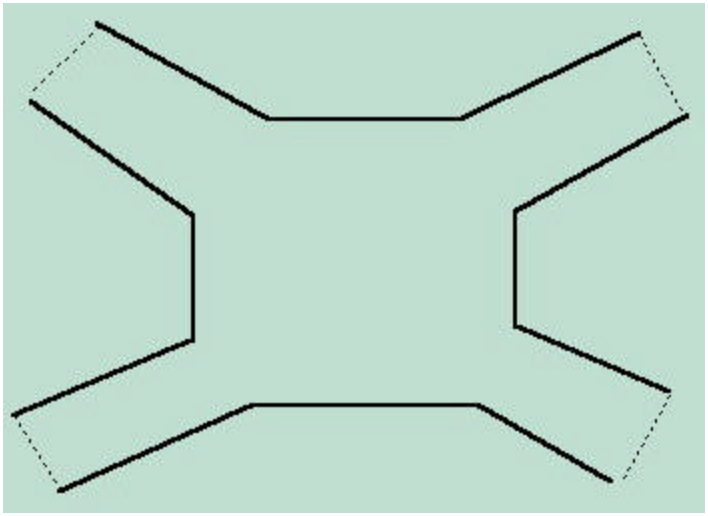
An extra principle which would help was a form of duality. When two particles come together, interact and scatter off each other they could have done one of two things. It could be that they exchanged an intermediate particle, like an electron and positron exchanging a photon. Or, it could be that they join to form a new particle which then reverts back to the original two, like an electron and positron which annihilate briefly and are then recreated from a photon. These two scattering modes are known as the t-channel and s-channel respectively. For strong interactions it was found experimentally that these two amplitudes were approximately the same. There might be a principle which meant that the two channels were somehow really the same thing. Could there be an underlying interaction which possessed such duality exactly?



No sooner had the idea been thought of when Gabriele Veneziano came up with a simple formula for the scattering amplitude which did indeed possess this duality. He gave no model of what it was going on during the scattering process, just a formula which satisfied the constraints on the S-matrix. It was not long before the answer emerged suddenly from three different people. Lenny Susskind, then at Yeshiva University published his "Dual-symmetric theory of hadrons". Holger Nielsen of the Niels Bohr institute in Copenhagen called his paper "An almost physical interpretation of the dual N point function" while Yoichiro Nambu in Chicago produced "Quark model and the factorisation of the Veneziano amplitude". It was 1970 and string theory had been reborn.

By that time the evidence in favour of quarks as constituents of the proton and neutron was becoming more convincing, but nobody could understand why they were never seen on their own. They seemed to be bound together inside the hadrons. According to string theory "bound" was just the right word. The quarks were always attached to the end of strings which resisted them being pulled apart. When stretched too far it would break but a new quark anti-quark pair formed from the energy released would take hold of the lose ends. The process could also reverse when strings join together. In space-time the strings sweep out a surface or world sheet. The scattering of two mesons would now be described by a process in which two strings joined momentarily and then broke. When the world sheet is drawn the explanation for duality suddenly

becomes clear. The same picture can be interpreted as either a t-channel or s-channel scattering mode.



String theory was considered as a theory of strong interactions for some time but it had problems. It only worked correctly in 26-dimensional space-time, not a very physical feature. Eventually this theory gave way to another theory called Quantum Chromo Dynamics which explained the strong nuclear interaction in terms of colour charge on gluons. In any case, string theory may have sounded good for mesons made of two quarks but protons have three. A string cannot have three ends. It looked like string theory was about to be lost for a second time.

String theory suffered from certain inconsistencies apart from its dependence on 26 dimensions of space-time. It also had *tachyons*, particles with imaginary mass which must travel faster than light. Tachyons could reek havoc with causality and would destabilise the vacuum, but string theory had already cast its spell on a small group of physicists who felt there must be something more to it. Pierre Ramond, Andre Neveu and John Schwarz looked for other forms of string theory and found one with fermions in place of bosons. The new theory in 10 dimensions was supersymmetric and, magically, the tachyon modes vanished.

What then was the interpretation of this new model? Schwarz teamed up with Joel Scherk and found that at low energies the strings would appear as particles. Only at very high energies would these particles be revealed as bits of string. The strings could vibrate in an infinite tower of quantised modes in an ever increasing range of mass, spin and charge. The lowest modes could correspond to all the known particles. Better still, the spin two modes would behave like gravitons. The theory was necessarily a unified theory of all interactions including quantum gravity. In 1978 the leading candidate for a super unified theory was eleven-dimensional supergravity and superstrings were largely ignored. Despite early hopes, supergravity was not quite renormalisable and it just failed to have the right properties to explain the left-right asymmetry of particle physics. Then came the historic 1984 paper of Green and Schwarz and their discovery of almost miraculous anomaly cancellations in one particular theory. Almost instantly superstrings took over as the hottest topic of research.

To come back to the original question, *why string theory?* The answer is simply that it has the right mathematical properties to be able to reduce to theories of point particles at low energies, while being a perturbatively finite theory which includes gravity. The simple fact is that there are no other known theories which accomplish so much. Of course physicists have now studied the mathematics of vibrating membranes in any number of dimensions. The fact is that there are only a certain number of possibilities to try and only the known string theories work out right in perturbation theory.

Of course it is possible that there are other completely different self-consistent theories but they would lack the important perturbative form of string theories. The fact is that string theorists are now turning to other p-brane theories. Harvey, Duff and others have found equations for certain p-branes which suggest that self-consistent field theories of this type might exist, even if they do not have a perturbative form.

All Is String

In 1985 string theory developed rapidly. It was discovered to have a rich and compelling mathematical structure which persuaded a growing band of physicists that it must be the next step forward. All particles were imagined to be tiny threads vibrating like resonating guitar strings. The strings can be open ended or they can be closed loops. The different harmonics correspond to different particles with different mass, spin, charge etc. In experiments physicists will only have seen the first few modes of vibration among the particles we know since most of them will have relatively high mass. There are modes which can have as high a mass and spin as you may demand. The strings are not made of anything in particular. It is wrong to say they are made of energy because energy is actually just one of the properties they carry. They are best thought of as strands of pure substance with length but no thickness.

One of the strengths of string theory is that it also included massless spin two bosons in its repertoire. These were identified as gravitons; quantum particles of gravity. Physicists had thought before then that they could see how to fit together the electromagnetic and nuclear forces but the gravitational force had been a big problem. Now they were replacing quantum field theory, which could not include gravity, with string theory which must include it.

By 1981 Green and Schwarz had identified two separate types of superstring theory. Type I is the theory of open strings but it must include closed strings as well to be complete. The other known as Type II has only closed strings. In the Type II theories the bosons and fermions appear as wave modes which circle round the strings in opposite directions. There is a version of either type for each gauge group, but the breakthrough of 1984 was the discovery that the quantisation of Type I is only free of infinities when the gauge group is SO(32) .

They also found that Type II theory worked with the same group and that it had two versions Type IIa and Type IIb. In 1985 the family of string theories was enlarged by the arrival of the *heterotic string*. This version discovered at Princeton by David Gross, Jeffrey Harvey, Emil Martinec and Ryan Rohm, also had two versions which were finite. One with gauge group SO(32) again, and the other with E8%E8. The total number of possibilities was therefore five, sometimes denoted I, IIa, IIb, HO and HE. No other theories with the same good behaviour can

be found. String theorists would like to have a unique theory so five is an embarrassment of choice. On the other hand it is much better than the situation regarding quantum field theory which works with any gauge group and a whole variety of possible matter fields, yet cannot unify all the forces.

All five superstring theories only work in 10 dimensions, 9 space dimensions plus 1 time dimension. If they have anything to do with real physics then six of the space dimensions must be rolled up or *compactified* just as a two-dimensional sheet of paper can be rolled into a narrow tube which becomes a one-dimensional line. If the distance around the compact dimension is very small, perhaps the Planck length, then we would not be aware of it. While there is only one way to roll up one dimension giving a tubular cross-section which is a circle, more dimensions can be rolled up in many different ways. With two dimensions there is already the choice of a sphere, torus or other surfaces with more than one hole.

These are topologically distinct and for any given choice of compactification for each string theory a different theory of the universe with different particles is found. The number of ways you can go about reducing string theory to four dimensions in this fashion is just mind boggling. It is too difficult to find the one which should correspond to our universe.

String theory is a superb example of unification. Through supersymmetry, matter is united with force. There is only one type of object; the string. If it vibrates one way it can be a quark, another way it is an electron, change its mode again and it becomes a force carrying photon or even a graviton.

But by 1988 string theory was in trouble. Past history shows that breakthroughs in physics are at first largely ignored until experiment forces the community of physicists to accept them. Such had been the case with atoms, relativity, parity violation, quark theory and electroweak unification. By contrast string theory was immediately taken up by a huge proportion of physicists and then it failed to make any experimental predictions which could be tested. Richard Feynman was one of those who spoke against his mostly younger colleagues who supported string theory. He did not like the fact that string theorists were not calculating anything which would allow them to check their ideas empirically.

Yet they carried on. String theory was still young and rather than letting its critics stop them they would rise to the challenge. The acknowledged leader in the fight to understand string theory is Ed Witten. He speaks in a very different tone, explaining that the critics do not seem to have fully grasped the scope and richness of the structure involved in string theory. They are too impatient for quick answers.

Duality

In 1986 one of the niggling problems in superstring theory was the fact that there were 5 different versions. Which one would correspond to our world and what is the point of the other four? Then there was a sequence of big discoveries which brought new hope.

A fine example of the rich and beautiful structure of string theory is T-duality, short for target space duality. The target space of a string theory is just the space-time in which it is placed. The five principal superstring theories are most at home in flat ten-dimensional space-time infinite in all directions, but they can also be placed in space-times where some of the dimensions have been compactified.

The simplest case is where one of the space dimensions is rolled up round a circle of radius R. A string theory in such a space-time appears like a nine-dimensional theory of strings. The rolled up dimension becomes invisible and the compactification radius R becomes just one of many arbitrary parameters.

Since there are five superstring theories in 10 dimensions and only one way to compactify to 9 dimensions, you would expect there to be five superstring theories in 9 dimensions too. In actual fact there are only three. The two different heterotic theories in 10 dimensions, HE and HO, reduce to the same nine-dimensional theory. The compactification radii R_E for HE and R_O for HO heterotic string appear as a parameter in this theory but they are related inversely $R_E = \alpha' /R_O$. HE is recovered as the limit of the nine-dimensional string theory as R_E is made large and HO is the limit as R_O is made large. So the two heterotic string theories are really two aspects of the same theory. They are said to be T-dual. The same magic can be applied to the two TypeII theories. IIa is T-dual to IIb. This leaves us with just three separate superstring theories Type I, Type II and Heterotic.

That is how the situation stood in 1993 but then another kind of duality was found. It concerns a relation between electric charges and magnetic monopoles.

Maxwell's equations for electromagnetic waves in free space are symmetric between electric and magnetic fields. A changing magnetic field generates an electric field and a changing magnetic field generates an electric one. The equations are the same in each case, apart from a sign change. If you take the equations and switch the electric and magnetic fields, while changing the sign of one of them, you arrive back at the same form. The free fields without charges are invariant but if electric charges are included there must also be magnetic charges to complete the symmetry. However, it is an experimental observation that there are no magnetic monopole charges in nature which mirror the electric charge of electrons and other particles. Despite some quite careful experiments only dipole magnetic fields which are generated by circulating electric charges have ever been seen.

In classical electrodynamics there is no inconsistency in a theory which places both magnetic and electric monopoles together. In quantum electrodynamics this is not so easy. To quantise Maxwell's equations it is necessary to introduce a vector potential field from which the electric and magnetic fields are derived by differentiation. This procedure cannot be done in a way which is symmetric between the electric and magnetic fields.

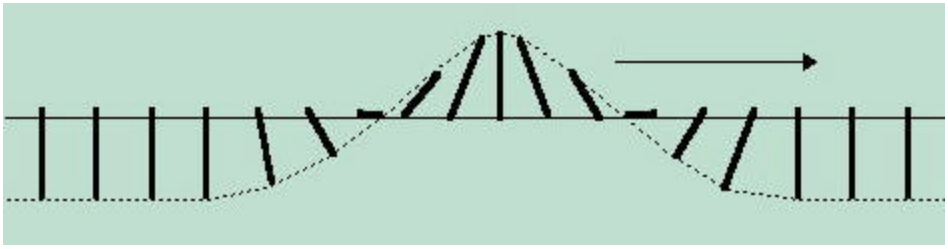
Forty years ago Paul Dirac was not convinced that this ruled out the existence of magnetic monopoles. Again motivated by mathematical beauty in physics, he tried to formulate a theory in which the gauge potential could be singular along a string joining two magnetic charges in such a way that the singularity could be displaced through gauge transformations and must therefore be

considered physically inconsequential. The theory was not quite complete but it did have one saving grace. It provided a tidy explanation for why electric charges must be quantised as multiples of a unit of electric charge.

In the 1970s it was realised by 't Hooft and Polyakov that grand unified theories which might unify the strong and electro-weak forces would get around the problem of the singular gauge potential because they had a more general gauge structure. In fact these theories would predict the existence of magnetic monopoles. Even their classical formulation could contain these particles which would form out of the matter fields as *topological solitons*.

There is a simple model which gives an intuitive idea of what a topological soliton is. Imagine first a straight wire pulled tight like a washing line with many clothes pegs strung along it. Imagine that the clothes pegs are free to rotate about the axis of the line but that each one is attached to its neighbours by elastic bands on the free ends. If you turn up one peg it will pull those nearby up with it. When it is let go it will swing back like a pendulum but the energy will be carried away by waves which travel down the line. The angles of the pegs approximate a field along the one-dimensional line.

The equation for the dynamics of this field is known as the sine-Gordon equation. It is a pun on the Klein-Gordon equation which is the correct linear equation for a scalar field and which is the first order approximation to the sine-Gordon equation for small amplitude waves. If the sine-Gordon equation is quantised it will be found to be a description of interacting scalar fields in one dimension.



The interesting behaviour of this system appears when some of the pegs are swung through a large angle of 360 degrees over the top of the line. If you grab one peg and swing it over in this way you would create two twists in the opposite sense around the line. These twists are quite stable and can be made to travel up and down the line. A twist can only be made to disappear in a collision with a twist in the opposite direction.

These twists are examples of topological solitons. They can be regarded as being like particles and antiparticles but they exist in the classical physics system and are apparently quite different from the scalar particles of the quantum theory. In fact the solitons also exist in the quantum theory but they can only be understood non-perturbatively. So the quantised sine-Gordon equation has two types of particle which are quite different.

What makes this equation so remarkable is that there is a non-local transformation of the field which turns it into another one-dimensional equation known as the Thirring model. The transformation maps the soliton particles of the sine-Gordon equation onto the ordinary quantum

excitations of the Thirring model, so the two types of particle are not so different after all. We say that there is a duality between the two models, the sine-Gordon and the Thirring. They have different equations but they are really the same because there is a transformation which takes one to the other.

The relevance of this is that the magnetic monopoles predicted in GUT's are also topological solitons, though the configuration in three-dimensional space is more difficult to visualise than for the one dimension of the clothesline. It would be nice if there was a similar duality between electric and magnetic charges as the one discovered for the sine-Gordon and Thirring equations. If there was then a duality between electric and magnetic fields would be demonstrated. It would not be quite a perfect symmetry because we know that magnetic monopoles must be very heavy if they exist.

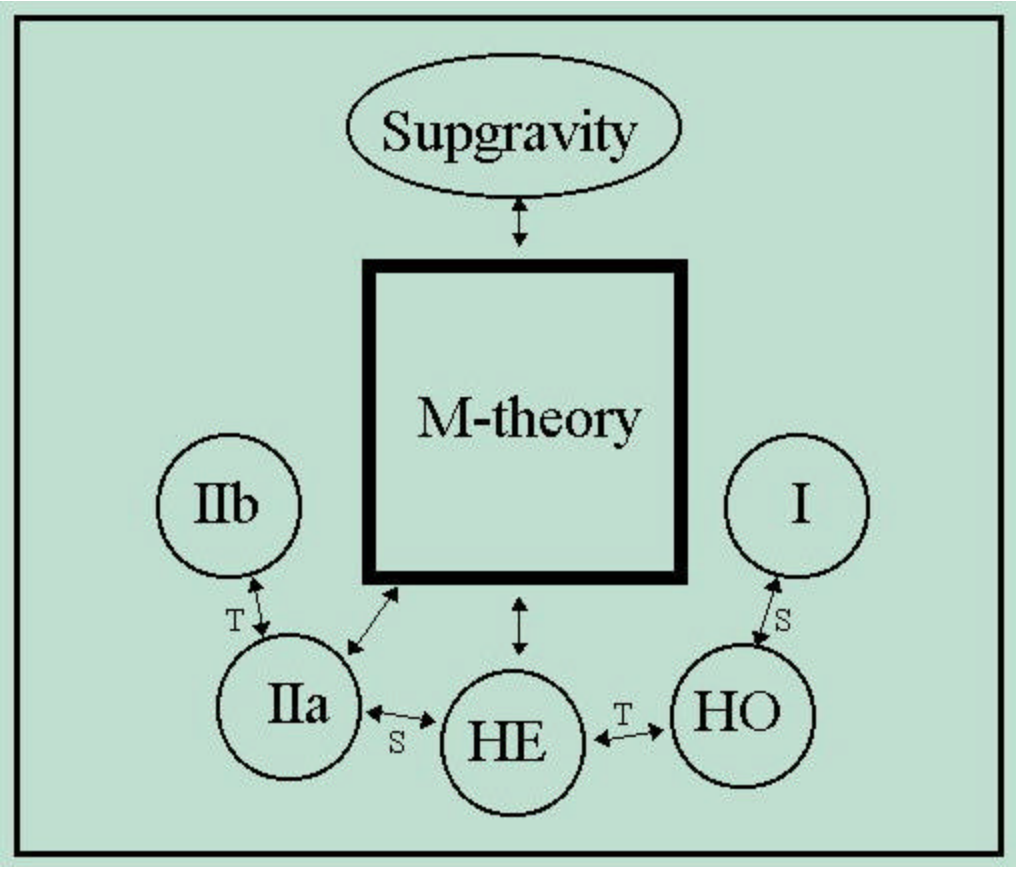
In 1977 Olive and Montenen conjectured that this kind of duality could exists, but the mathematics of field theories in 3 space dimensions is much more difficult than that of one dimension and it seems beyond hope that such a duality transformation can be constructed. But they made one step further forward. They showed that the duality could only exist in a supersymmetric version of a GUT. This is quite tantalising given the increasing interest in supersymmetric GUT's which are now considered more promising than the ordinary variety of GUT's for a whole host of reasons.

Until 1994 most physicists thought that there was no good reason to believe that there was anything to the Olive-Montenen conjecture. Then Nathan Seiberg and Ed Witten made a breakthrough which rocked the worlds of physics and mathematics. By means of a special set of equations they demonstrated that a certain supersymmetric field theory did indeed exhibit electro-magnetic duality. As a bonus their method can be used to solve many unsolved problems in topology and physics. The duality exchanges strong coupling with weak coupling. This is very significant for theories like QCD where the strong coupling limit is not understood.

This kind of duality is now known as S-duality to distinguish it from T-duality. In string theory S-duality is very natural. There is a general rule about the dimensions of dual objects. An "electric" p1-brane which is a fundamental construct of a theory in D dimensions can have a p2-brane "magnetic" soliton when p1 + p2 = D - 4. In the familiar case the electric and magnetic charges in D=4 are particles, i.e. 0-branes. In D=10 string theory the strings are 1-branes so their duals must be (10-4-1)-dimensional 5-branes. In the last year physicists have discovered how to apply tests of duality to different string and p-brane theories in various dimensions. Conjectures have been made and tested. This does not prove that the duality is correct but each time a test has had the potential to show an inconsistency it has failed to destroy the conjectures.

It now seems that any string theory with sufficient supersymmetry must have an S-dual waiting to be found. What makes this discovery so useful is that the dualities are a non-perturbative feature of string theory. Now many physicists see that p-brane theories can be as interesting as string theories in a non-perturbative setting. Using T-duality we made reduced the five superstring theories to three. Now with S-duality we can make further links which leave them all connected. Type I is S-dual to HO while HE is S-dual to IIa (but only when compactified to six dimensions). The last of the five IIb is self dual.

That was not quite the end of the story. If these five theories are all part of the same thing then what is that thing? The answer, it seems, is that they are all derived from something called M-theory in 11 dimensions. M-theory is like string theory except that it is a theory of membranes (2-branes) rather than strings (1-branes). It also has an S-duality between its 2-branes and solitonic 5-branes. All five string theories are special points in the parameter space of this one theory, but so is eleven-dimensional supergravity theory, the same theory that string theory ousted as the most popular super-unified theory in 1984.



This may be too simple a picture of M-theory which really includes open and closed strings, membranes, p-branes etc. Each of the string theories appears in some corner of M-theory where particular states become weakly coupled and can be described using perturbation theory.

It would be wrong to say that very much of this is understood yet. There is still nothing like a correct formulation of M-theory or p-brane theories in their full quantum form, but there is new hope because now it is seen that all the different theories can be seen as part of one unique theory. The best way to formulate that theory is not yet known.

Black Strings

As if one major conceptual breakthrough was not enough, string theorists had to come to terms with a whole wave of new finds which started around 1994. Just as physicists have been quietly

speculating about electro-magnetic duality for decades, a few have also speculated that somehow elementary particles could be the same things as black holes so that matter could be regarded as a feature of the geometry of space-time.

It is curious that various stellar objects under the influence of strong gravity parallel various entities from particle physics. A white dwarf star is like an atom in that it resists collapse due to the Pauli exclusion principle. A more massive star will collapse further to a neutron star which is like a stable nucleus. A stronger gravitational force can reduce it to a quark star which is like a neutron. The final stage of gravitational collapse reduces the star to a black hole. If the analogy continues to hold, the black hole should be like a quark or other elementary particle.

The theory started to look a little less ridiculous when Hawking postulated that black holes actually radiate particles. The process could be likened to a very massive particle decaying. If a black hole were to radiate long enough it would eventually lose so much energy that its mass would reduce to the Planck scale. This is still much heavier than any elementary particle we know but quantum effects would be so overwhelming on such a black hole that it would be difficult to see how it might be distinguished from an extremely unstable and massive particle in its final explosion.

To make such an idea concrete requires a full theory of quantum gravity and since string theory claims to be just that, it seems a natural step to compare string states and black holes. We know that strings can have an infinite number of states of ever increasing spin, mass and charge. Likewise a black hole, according to the *no hair conjecture* is also characterised only by its spin, mass and charge.

With magnetic duality we can add magnetic charge to the list. It is therefore quite plausible that there is a complementarity between string states and black hole states, and in fact this hypothesis is quite consistent with all mathematical tests which have been applied. It is not something which can be established with certainty simply because there is not a suitable definition of string theory to prove the identity. Nevertheless, many physicists now consider it reasonable to regard black holes as being single string states which are continually decaying to lower states through Hawking radiation.

It was discovered that if you consider Planck mass black holes in the context of string theory then it is possible for space-time to undergo a smooth transition from one topology to another. This means that many of the possible topologies of the curled up dimensions are connected and may pave a way to a solution of the selection of vacuum states in string theory.

String Symmetry

Superstring theory is full of symmetries. There are gauge symmetries, supersymmetries, covariance, dualities, conformal symmetries and many more. But superstring theory is supposed to be a unified theory which should mean that its symmetries are unified. In the perturbative formulation of string theory that we have, the symmetries are not unified.

One thing about string theory which was discovered very early on was that at high temperatures it would undergo a phase transition. The temperature at which this happens is known as the Hagedorn temperature after a paper written by Hagedorn back in 1968, but it was in the 1980s that physicists such as Witten and Gross explored the significance of this for string theory.

The Hagedorn temperature of superstring theory is very high, such temperatures would only have existed during the first 10-43 seconds of the universe existence, if indeed it is meaningful to talk about time in such situations at all. Calculations suggest that certain features of string theory simplify above this temperature. The implication seems to be that a huge symmetry is restored. This symmetry would be broken or hidden at lower temperatures, presumably leaving the known symmetries as residuals.

The problem then is to understand what this symmetry is. If it was known, then it might be possible to work out what string theory is really all about and answer all the puzzling questions it poses. This is the superstring mystery.

A favourite theory is that superstring theory is described by a *topological quantum field theory* above the Hagedorn temperature. TQFT is a special sort of quantum field theory which has the same number of degrees of gauge symmetry as it has fields, consequently it is possible to transform away all field variables except those which depend on the topology of space-time. Quantum gravity in (2+1)-dimensional space-time is a TQFT and is sufficiently simple to solve, but in the real world of (3+1)-dimensional Einstein Gravity this is not the case, or so it would seem.

But TQFT in itself is not enough to solve the superstring mystery. If space-time topology change is a reality then there must be more to it than that. Most physicists working in string theory believe that a radical change of viewpoint is needed to understand it. At the moment we seem to be faced with the same kind of strange contradictions that physicists faced exactly 100 years ago over electromagnetism. That mystery was finally resolved by Einstein and Poincaré when they dissolved the ether. To solve string theory it may be necessary to dissolve space-time altogether.

In string theory as we understand it now, space-time curls up and changes dimension. A fundamental minimum length scale is introduced, below which all measurement is possible. It will probably be necessary to revise our understanding of space-time to appreciate what this means. Even the relation between quantum mechanics and classical theory seems to need revision. String theory may explain why quantum mechanics works according to some string theorists.

All together there seem to be rather a lot of radical steps to be made and they may need to be put together into one leap in the dark. Those who work at quantum gravity coming from the side of relativity rather than particle physics see things differently. They believe that it is essential to stay faithful to the principles of diffeomorphism invariance from general relativity rather than working relative to a fixed background metric as string theorists do. They do not regard renormalisability as an essential feature of quantum gravity.

Working from this direction they have developed the canonical theory of quantum gravity which is also incomplete. It is a theory of loops, tantalisingly similar in certain ways to string theory, yet different. Relativists such as Lee Smolin hope that there is a way to bridge the gap and develop a unified method

---

The Principle of Event Symmetry

The Bucket of Dust

Many theoretical physicists, and other people besides, will ask themselves at some time "What could the most fundamental laws of physics be like?" It is next to impossible to find the answer but it is still a useful question to think about. Most people will give an answer tainted by what they are familiar with. Descartes thought the answer would be mechanical and causal because that was what was familiar at the time. Today we might think of quantum mechanics instead.

As we ascend a mountain the scenery changes. We may pass from grassy pastures to harsher slopes, through alpine forest, up rocky cliffs till beyond the snow line we find the summit. As we climb the mountain of scientific truth our experience is similar. What will remain of our familiar surroundings when we reach the top, if indeed there is a top. When we passed from the land of classical certainty to the indeterminism of quantum mechanics Einstein said it was like the ground had been pulled out from under us leaving nothing to stand on. He was left behind as others climbed on. As we rise higher space-time is fading from our grasp and we have even less to hold on to.

A philosopher would tell you that the only thing which remains at the top is the realm of our perceptions. According to the storyteller's paradigm the universe is no more than the sum total of all possible experiences which can be perceived. This is realised in the multiverse of quantum mechanics described by Feynman's path integral. Thus some remnant of quantum mechanics should be valid on at least the final slopes. All else must emerge further down the levels of thought. Indirectly we apprehend events and the relations between them. According to a dictionary an event is anything which happens, but to a physicist an event is also a point of space-time; a place and a moment where something could happen.

Events are also what the physicist sees in his experiments when particles come together and interact. Particle physics, both theoretical and experimental is the pursuit of the most basic events and the rules which join them. Space-time is made of events but events are more fundamental than their *when* and *where*. Space-time forms out of the relationships between events.

In 1925 Alfred North Whitehead, philosopher of science, asked us to regard events as primordial. Space-time is constructed by us from the prehension of events. A physics based on events is sometimes called Whiteheadian but the origins of such philosophy can be traced back through the monadology of Leibniz to the atomistic doctrine of space and time in the Kalám of tenth century Baghdad, and perhaps beyond to the ancient Greeks.

With heavy irony John Archibald Wheeler described a universe constructed out of events as "a bucket of dust". He sought a pregeometry for space-time but felt that starting from the set of events is premature. A deeper guiding principle must be found.

The Universal Lattice

After I had finished my doctorate in 1985 I also wondered what the fundamental laws of physics might be like. My thesis had been about lattice gauge theories so I was used to thinking about space-time as made up of discrete events (or lattice sites) with links joining nearest neighbours together. Fields are represented by numbers attached to events and links. It is just an approximation trick for doing calculations. The continuum is supposed to be regained from the cubic array of the lattice in the limit when the distance between lattice points goes to zero. In fact the sites can be linked in other ways, so long as they make some kind of four-dimensional lattice. The continuum limit should be the same in all cases.

I imagined what might happen if the fixed linkage structure of the lattice was discarded. It could be made dynamic allowing any site to link to any other nearby site at random. Why not even allowing linkage to any site no matter how far away? For maximum simplicity each site should have no preferences for which other sites it likes to link to. When doing lattice gauge theory calculations, the path integral of quantum mechanics becomes a sum over different configurations of the field variables weighted by a factor related to the action. Dynamic links changing at random fit into the sum quite naturally. It now includes a sum over all the ways of linking up the lattice sites as well as a sum over the values of the field variables. You can even look for interesting physics in models where there are no field variables, just random links between events.

This paints a rather strange image of the universe. Events and links between events would be fundamental objects but there would be no built in structure to space-time, no continuity, no dimension. The dynamics would be determined by the form chosen for the action as a function of the way the events were linked up. It might take into account the number of links meeting at each event, the number of triangles which form and other similar quantities which depend on the network of connections. For the right choice of action, lattices with a four-dimensional structure might be favoured and the structure of space-time could be determined dynamically. In some appropriate limit a continuum might emerge. If it could be done it would show how the laws of physics, including the nature of space-time, could be derived from much simpler equations than those normally used to specify them.

Such speculations are often naive and unlikely to work out right, which is why Wheeler likened such models to a bucket of dust. Nevertheless you have to try these things out because if you do not make a few mistakes you never learn anything. The attractive thing about the idea for me was

that you could simulate such systems on a computer and watch what happened. The results I got were not overly encouraging. There is no simple and natural way to specify the dynamics of the lattice so that it tends to form structures like space-time, unless you build in some preference for which sites want to join up. To go further it would be necessary to think more carefully about how space-time is expected to behave.

Witten's Puzzle

Back in 1958 John Wheeler suggested that when general relativity and quantum theory were put together there would be astonishing things going on at the very small length scale known as the Planck length (about 10^{-35} metres). If we could look down to such distances we would see space changing wildly. In general relativity gravity results from space-time curvature. If gravity is quantised the curvature should fluctuate. Wheelers rough calculations showed that at the Planck scale the fluctuations would be so wild that space would be likely to tare open forming microscopic wormholes and other topological variations. The structure of this space-time foam has been a mysterious area of research ever since.

Topology change is found to be an important feature of superstring theory, so again string theorists seem to be on the right track. When they try to understand together the concepts of topology change and universal symmetry they come up against a strange enigma known as Witten's Puzzle after the much cited string theorist, Ed Witten, who first described it.

The difficulty is that both diffeomorphism invariance and internal gauge symmetry are strictly dependent on the topology of the space. Different topologies lead to non-equivalent symmetries. The diffeomorphism group of smooth mappings on a sphere is not isomorphic to the diffeomorphism group on a torus. The same applies to internal gauge groups. If topology change is permitted then it follows that the universal symmetry must, in some fashion, contain the symmetry structures for all allowable topologies at the same time. Witten admitted he could think of no reasonable solution to this problem.

An old maxim of theoretical physics says that once you have ruled out reasonable solutions you must resort to unreasonable ones. As it happens there is one unreasonable but simple solution to Witten's puzzle. It can already be identified as a property of the universal lattice where any event has no preference for which other events it connects to. This implies a simple permutation symmetry on events.

Consider diffeomorphisms to begin with. A diffeomorphism is a suitably smooth one to one mapping of a space onto itself. The set of all such mappings form a group under composition which is the diffeomorphism group of the space. A group is an algebraic realisation of symmetry. One group which contains all possible diffeomorphism groups as a subgroup is the group of all one-to-one mappings irrespective of how smooth or continuous they are. This group is the symmetric group on the manifold. Unlike the diffeomorphism groups, the symmetric groups on two topologically different space-times are algebraically identical. A solution of Witten's puzzle would therefore be for the universal group to contain the symmetric group acting on space-time events.

This is called **The Principle of Event Symmetry** which states that: *The universal symmetry of the laws of physics includes the symmetric group acting on space-time events.*

The principle of event symmetry is realised by the universal lattice, but it is more general. The universal lattice is a naive model of space-time whereas event symmetry is a deep principle which solves the puzzle of combining symmetry and topology change. There are also philosophical reasons for holding to the principle of event symmetry. According to the storyteller's paradigm, the multiverse describes all ways of putting together events. The events are taken from a heap within which they are not ordered. If something is not ordered then it does not matter how its contents are mixed up. They can be permuted without consequence. The symmetric group is a symmetry of the heap.

In its simplest form, event symmetry is realised in a heap of discrete events. The universal lattice is a good example. But the symmetric group can be a subgroup of a larger group allowing the individuality of events to be blurred. There are other ways of including event symmetry within larger symmetries. You can have a mapping from a larger symmetry onto a smaller one which preserves its structure. This is called a homomorphism. You can also deform symmetries by introducing a more general symmetry structure with a deformation parameter which reduces to something containing the symmetric group for one special case of that parameter. I will describe examples of all of these. The beauty of event symmetry is revealed in the ways it can become part of the full universal symmetry.

Space-Time and Soap Films

There are a number of reasons why this principle of event symmetry may seem unreasonable. For one thing it suggests that we must treat space-time at some level as a discrete set of events. In fact, as I have already explained, there are plenty of reasons to believe in discrete space-time. Theorists working on quantum gravity in various forms agree that the Planck scale defines a minimum length beyond which the Heisenberg uncertainty principle makes measurement impossible. In addition, arguments based on black hole thermodynamics suggest that there must be a finite number of physical degrees of freedom in a region of space.

A more direct reason to doubt the principle would be that there is no visible or experimental evidence of such a symmetry. The principle suggests that the world should look the same after permutations of space-time events. It should even be possible to swap events from the past with those of the future without consequence. This does not seem to accord with experience. Event symmetry cannot be a principle of nature unless it is well hidden. Since the symmetric group acting on space-time can be regarded as a discrete extension of the diffeomorphism group in general relativity, it is worth noting that the diffeomorphism invariance is not all that evident either. If it were then we would expect to be able to distort space-time in ways reminiscent of the most bizarre hall of mirrors without consequence. Everything around us would behave like it is made of liquid rubber. Instead we find that only a small part of the symmetry which includes rigid translations and rotations is directly observed on human scales. The rubbery nature of space-time is more noticeable on cosmological scales where space-time can be distorted in quite counterintuitive ways.

If space-time is event-symmetric then we must account for space-time topology as it is observed. Topology is becoming more and more important in fundamental physics. Theories of magnetic monopoles, for example, are heavily dependent on the topological structure of space-time. To solve this problem is the greatest challenge for the event-symmetric theory.

To get a more intuitive idea of how the event symmetry of space-time can be hidden we use an analogy. Anyone who has read popular articles on the Big Bang and the expanding universe will be familiar with the analogy in which space-time is compared to the surface of an expanding balloon. The analogy is not perfect since it suggests that curved space-time is embedded in some higher-dimensional flat space, when in fact, the mathematical formulation of curvature avoids the need for such a thing. Nevertheless, the analogy is useful so long as you are aware of its limitations.

We can extend the balloon analogy by imagining that space-time events are like a discrete set of particles populating some higher-dimensional space. The particles might float around like a gas of molecules interacting through some kind of forces. In any gas model with just one type of molecule the forces between any two molecules will take the same form dependent on the distance between them and their relative orientations. Such a system is therefore invariant under permutations of molecules. In other words, it has the same symmetric group invariance as that postulated in the principle of event-symmetric space-time, except that it applies to molecules rather than events.

Given this analogy we can use what we know about the behaviour of gases and liquids to gain a heuristic understanding of event-symmetric space-time. For one thing we know that gases can condense into liquids and liquids can freeze into solids. Once frozen, the molecules stay fixed relative to their neighbours and form rigid objects. In a solid the symmetry among the forces still exists but because the molecules are held within a small place the symmetry is hidden.

Another less common form of matter gives an even better picture. If the forces between molecules are just right then a liquid can form thin films or bubbles. This is familiar to us whenever we see soap suds. A soap film takes a form very similar to the balloon which served as our analogy of space-time for the expanding universe. The permutation symmetry of the molecular forces is hidden and all that remains is a surface. The same idea works in higher dimensions so it is possible that four-dimensional space-time may condense out of something like a gas of events, just like the formation of a soap bubble. Curvature of space-time is similar to the curvature of the surface of the soap film.

Permutation City

In 1991 I had worked out the basic ideas behind the principle of event symmetry. At that time I was working as a contract software engineer and was isolated from front line research in theoretical physics. I did not take my physics very seriously and I imagined that such a simple and obvious notion as event symmetry would have been considered already by physicists. They would, I thought, have already extracted any useful consequences there might be. I was wrong.

Two years later the world went through a new revolution in information technology: the internet. Its impact on science rivals the introduction of the printing press into Europe in the fifteenth century. The internet had already existed for some time. I had used it myself as a research student in 1984 when I used to control computers in Germany from my base in the University of Glasgow. But in 1993 the internet came out of academic institutes into the wider world, where I was then working as a programmer in France. I gained access to usenet and the world wide web and I regained access to what was happening in physics. I could download the latest papers in physics which appeared as electronic pre-prints each day. I could search databases of papers compiled over the previous twenty years. Best of all, I could write my own papers and circulate them on the internet. In April 1994 my first tentative paper about event-symmetric space-time emerged and drew no response.

I decided that it would be prudent to find out who else had done similar work in the past. Using on-line databases I searched the literature for papers with titles that had anything to do with discrete space-time and then followed their hyperlinked references and citations to find other relevant papers. I discovered the work on Wheeler, Finkelstein and others which I had not heard of before. There were, in fact, just a few examples of such work which dared to speculate about the small scale structure of space-time with models not unlike my universal lattice. Some of what I found was more mathematically sophisticated, yet not one example expressing the principle of event symmetry came to light. I continued my work. A couple of years later a contact on the internet told me about a book which discussed ideas similar to mine. It was not a physics book. It was 'Permutation City', a science fiction novel by Greg Egan, but it was a science fiction novel with more interesting things to say about the philosophy of physics than many physicists or philosophers.

In 2045 the protagonist, Paul Durham, programs a simulation of himself into a computer. Applying the strong AI hypothesis, the story line continues from the point of view of the copy. It is another invocation of the storyteller's paradigm. A computer simulation can be regarded as a sophisticated way of recounting a story. As the storyteller told us, there is no need to distinguish between the story and reality. Durham performs some experiments with his copy, now referred to as Paul, in the simulation. He divides the program up and changes the order in which states are computed. The events of Paul's simulated life are permuted but he does not experience anything different from normal.

Paul tries to understand what is happening to him in terms of the theory of general relativity. Relativity declares that points of view of different observers are equally valid, but only observers whose reference frames can be related by continuous co-ordinate transformations. The mapping between the events of Paul's existence and the events of space-time outside the computer were discontinuous. In relativity influences have to be localised travelling from point to point at a finite velocity. Paul thought that if you chop up space-time and rearrange it, then causal structure would fall apart.

Finally Paul appreciates the principle of event symmetry, or as Egan calls it; the dust theory. It would be a new principle of equivalence, a new symmetry between observers. Relativity threw out absolute space and time but it did not go far enough. Absolute cause and effect must go too.

Permutation City was first published in 1994 and parts were adapted from a story called 'Dust' which was first published in Isaac Asimov's Science Fiction Magazine, July 1992.

More Symmetry

When Einstein decided to try to revise Newton's gravity he was advised not to waste his time. The problem was regarded as too difficult. Einstein persisted and succeeded against short odds in formulating a relativistic theory of gravity because he recognised the importance of the principle of equivalence. He deduced that the principle required curved space-time and reduced it to a need for generally covariant equations. This was the powerful symmetry which we now call diffeomorphism invariance. It was sufficiently stringent as a requirement that Einstein was able to deduce the essential form of the field equations for gravity leaving only Newton's gravitational constant and the possibility of a cosmological constant to be determined empirically.

The principle of event symmetry is stronger, in a sense, than diffeomorphism symmetry because it is larger, but it also allows for more general models of space-time as discrete sets. Einstein was able to assume that space-time was a continuous manifold with one temporal and three spatial dimensions. We no longer have such a restriction and consequently there are too many possible ways to devise event-symmetric theories. Event symmetry on its own is not very powerful. To go further the symmetry must be extended.

So far we have seen how the principle of event-symmetric space-time allows us to retain space-time symmetry in the face of topology change. Beyond that we would like to find a way to incorporate internal gauge symmetry into the picture too. It turns out that there is an easy way to embed the symmetric group into matrix groups. This is interesting because, as it happens, matrix models are already studied as simple models of string theory. String theorists do not normally interpret them as models on event-symmetric space-time but it would be reasonable to do so in the light of what has been said here.

To see how event-symmetry leads naturally to matrices consider how the universal random lattice may be represented. Each event could be labelled with an index i. For each pair of events (i, j) there may or may not be a link joining them in the lattice. This could be represented by a matrix of variables a_{ij} each of which is zero or one. One indicates that events i and j are linked, and zero indicates that they are not linked.

$$a_{ij} = a_{ji}$$

$$a_{ii} = 0$$

So the state of the random lattice is specified by a symmetric square matrix with zero diagonal other entries may be zero or one.

To put a model of a gauge theory on this lattice, field variables ϕ_i can be associated with each event and gauge variables U_{ij} with each link. The field variables form a column vector Φ and the gauge variables can again be collected together in a matrix A. If it is a Z_2 gauge theory, the elements of the matrix are now always zero or plus or minus one. The matrix A can be

symmetric but it may be more convenient to make it antisymmetric since the diagonal elements are then necessarily zero without imposing an extra condition. Gauge invariant quantities which could be used in an action for this model can be expressed in matrix notation e.g.

$$S = m\Phi^T\Phi + \Phi^T A\Phi + tr[A^4]$$

A gauge transformation can be effected as a similarity transformation on the matrix and vector. That is,

$$\Phi \to \Phi T$$
$$A \to T^{-1}AT$$

For the Z_2 gauge transformation T is a diagonal matrix with 1 and -1 down the diagonal. For example,

$$T = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix}$$

All of this generalises easily to other gauge groups. For an SO(N) gauge transformation T is a block diagonal matrix with blocks of N by N orthogonal matrices down the diagonal.

What about event symmetry? A permutation of events is also a symmetry of an action expressed in matrix notation as above. Columns and rows of the matrix and vector are permuted. This can also be effected by a similarity transformation T which is a permutation matrix. I.e. T has a single element equal to 1 in each row and column and all other elements equal to zero. For example,

$$T = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

Now that we have put internal gauge symmetry and event-symmetry into similar forms it is tempting to unify them. In both cases the similarity transformations are orthogonal matrices. If the elements of Φ and A are allowed to be any real numbers the matrix action has a full symmetry of orthogonal matrix transformations which includes the gauge transformations and event permutations as special cases. The same can be done with other gauge groups using orthogonal or unitary matrix models.

In these models the total symmetry of the system is a group of rotation matrices in some high-dimensional space. The number of dimensions corresponds to the total number of space-time

events in the universe, which may be infinite. Permutations of events now correspond to rotations in this space which swap over the axes.

So does this mean that the universal symmetry of physics is an infinite-dimensional orthogonal matrix? The answer is probably no since an orthogonal matrix is too simple to account for the structure of the laws of physics. For example, orthogonal groups do not include supersymmetry which is important in superstring theories. The true universal symmetry may well be some much more elaborate structure which is not yet known to mathematicians.

Before moving on it is worth taking note of how the amount of symmetry has increased in going over to matrix models. In conventional gauge theory there are a few degrees of symmetry for each event so the symmetry is of dimension N; the number of space-time events. With the matrix model there is a degree of symmetry for each independent element of the matrix so the symmetry is of dimension N^2. This is just the first step towards the much larger symmetries which may be present in the universe.

Identical Particles

Theorists often talk about unifying the gauge symmetries which are important to our understanding of the four natural forces. There are, however, other symmetries in nature which are rarely mentioned in the context of unification. These symmetries take the form of an invariance under exchange of identical particles. For example, every electron in the universe is the same, they all have the same charge, mass etc. If we swap one electron in the universe with another the universe will carry on as before.

The symmetry involved here is described by the symmetric groups, just like event-symmetric space-time. Obviously we should ask ourselves whether or not there is any connection between the two. Could the symmetric group acting to exchange identical particles be part of the symmetric group acting on space-time events? If it were, then that would suggest a deep relation between space-time and matter. It would take the process of unification beyond the forces of nature towards a more complete unification of matter and space-time.

As we shall see it is natural to combine the permutation symmetry of particles and event-symmetry and it will imply a unification of particle statistics and gauge symmetries which has now become apparent in superstring theories.

Clifford's Legacy

On its own, the principle of event-symmetric space-time is not very fruitful. What is needed is a mathematical model which incorporates the principle and which gives body to some of the speculative ideas outlined above.

It turns out that such a model can be constructed using Clifford algebras. These algebras are very simple in principle but have a remarkable number of applications in theoretical physics. They

first appeared to physicists in Dirac's relativistic equation of the electron. They also turn out to be a useful way to represent the algebra of fermionic annihilation and creation operators.

If we regard a Clifford algebra as an algebra which can create and annihilate fermions at space-time events then we find we have defined a system which is event-symmetric. It can be regarded as an algebraic description of a quantum gas of fermions.

This is too simple to provide a good model of space-time but there is more. Clifford algebras also turn out to be important in construction of supersymmetries and if we take advantage of this observation we might be able to find a more interesting supersymmetric model.

The definition of Clifford Algebras is very simple. It is an algebra generated by a set of elements γ_i such that

$$\gamma_i\gamma_j + \gamma_j\gamma_i = 2\delta_{ij}$$

A general element of the algebra can be expressed as sums of products of these elements. Since they square to one each need appear only once in any product. If there is one generator for each of N space-time events then the algebra has 2^N independent terms. To each of these we can assign a field variable. Each one is the coefficient of k different γ_i with k < N and can be interpreted as a field variable for a k-simplex with the k events as vertices. In comparison with the matrix model which had a field variable for each event and each pair of linked events, a model using Clifford algebras will have these plus a variable for each triplet of events, each quadraplet etc.

Back to Superstrings

Superstring theory was an important part of the motivation for proposing the principle of event-symmetric space-time in the first place. String theorists seem to believe that the subject they are studying is already the correct theory of physics, but they are probably missing the key to understanding its most natural formulation.

The situation seems to parallel Maxwell's theory of electromagnetism as it was seen at the end of the 19th century. Many physicists did not accept the validity of the theory at that time. This was largely because of the apparent need for a medium of propagation for light known as the ether, but experiment had failed to detect it. Einstein's theory of special relativity showed why the ether was not needed. He did not have to change the equations to correct the theory.

Instead he introduced a radical change in the way space and time were viewed. It is likely that the equations we have for string theory are also correct, although they are not as well formed as Maxwell's were. To complete the theory it is again necessary to revise our concept of space-time and remove some of its unnecessary structure just as Einstein removed the ether.

It would be natural to search for an event-symmetric string model. We might try to generalise the fermion model described by Clifford algebras to something which was like a gas of strings. A string could be just a sequence of space-time events connected in a loop. The most significant

outcome of the event-symmetric program so far is the discovery of an algebra which does just that. It is an algebraic model which can be interpreted as an algebra of strings made of closed loops of fermionic partons.

The result is not sophisticated enough to explain all the rich mathematical structures in string theory but it may be a step towards that goal. Physicists have found that new ideas about knot theory and deformed algebras are important in string theory and also in the canonical approach to quantisation of gravity. This has inspired some physicists to seek deeper connections between them. Through a turn of fate it appears that certain knot relations have a clear resemblance to the relations which define the discrete event-symmetric string algebras. This suggests that there is a generalisation of those algebras which represents strings of anyonic partons, that is to say, particles with fractional statistics.

Event-Symmetric Physics

What can this theory tell us about the universe? Since it is incomplete it is limited. The one place where a theory of quantum gravity would have most significance would be at the big bang. In the first jiffy of existence the temperature was so high that the structure of space-time would have been disrupted. It is known that in string theory there is a high temperature phase transition in which the full symmetry is realised. If the principle of event-symmetric space-time is correct then that is a much larger symmetry than people have previously imagined. At such high temperature space-time would cease to exist in the form we would know it, and only a gas of interacting strings would be left. A reasonable interpretation of this state of affairs would be to say that space-time has evaporated. The universe started from such a state, then space-time condensed and the rest is history.

Event-Symmetric String Theory

Leap Frog

In my mind, the principle of event symmetry would be a mere curiosity if it were not for string theories. Although they appear conceptually similar to quantum field theories with particles replaced by strings and higher-dimensional p-branes, it has become clear that string theories are really an altogether different and much stranger animal. For quantum field theories space-time is just a static arena within which the action is played out, but in string theory space-time is part of the show. String theory seems to understand the small scale structure of space-time better than we do. The best part of its trick is to fool us into thinking that space-time is real, flat and

continuous. We should not be fooled into taking this for anything other than the clever illusion which it must surely be.

There have been many amazing discoveries about superstring theory, but there are still some deep conceptual problems concerning the way it is formulated. The most profound of these is that string theory does not directly account for the equivalence principle. We know that superstring theory has gravitons and supergravity is therefore a component of the effective theory of strings at low energy. Supergravity is generally covariant and so incorporates ordinary general relativity with its equivalence principle. Thus string theory seems to include the equivalence principle, but the formulations we know are not generally covariant. There are versions which are Lorentz covariant but that is a long way short of the **general covariance** under all co-ordinate transformations. It is a little surprising and frustrating that this is the case and it may well be a key part of why we do not fully understand string theory.

The principle of event-symmetric space-time is the solution which I propose as a resolution of the superstring mystery. Event symmetry is a step beyond the diffeomorphism invariance of general covariance. If we can formulate string theory in a way which is event-symmetric we can leap frog over the conceptual hurdles.

Eight Reasons to Believe

Why should anyone believe that string theory is event-symmetric? I cannot prove it to you but I can give seven good reasons why I think it is right. The first is the problem of **general covariance** I just described. If string theory cannot be made covariant it seems hopeful that it may be event-symmetric instead.

Another reason which I already covered is the solution to **Witten's puzzle**. Topology change and the universal symmetry put together are difficult to reconcile without event symmetry.

The third reason is the presence of a **very large symmetry** of string theory beyond its Hagedorn temperature. It is not known what this symmetry is but it seems to reduce the effective number of degrees of freedom enormously. It is likely that there must be one dimension of symmetry to match each degree of freedom of the string. No mere gauge symmetry can achieve this but event symmetry is much larger than any gauge theory in quantum field theory.

Next I cite the important idea that strings can be considered as composites of **discrete partons**; particles bound together like beads on a necklace. Space-time too seems to have a discrete character. This picture may seem opposed to the usual formulation of strings as cords of continuous substance, yet it can explain many mysteries especially in the context of black holes. In that case it is easy to picture strings as loops connecting discrete points of space, and with such discreteness, event symmetry is easily imagined.

After that comes **matrix models**. String theory may ultimately be described by something like a model of random matrices whose rows and columns may index particles, colours of gauge symmetry or space-time events. Models on event-symmetric space-time also drive physics towards the dynamics of matrices. The matrix model which seems to contain the essence of M-

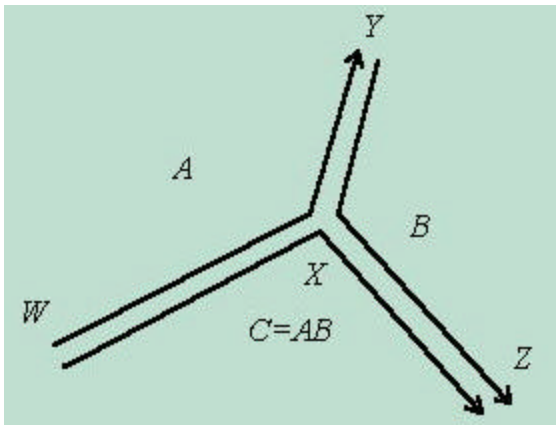
theory can be interpreted in any of these ways, bringing event symmetry a step clearer. A **unification of gauge symmetry and particle statistics** was a prediction of the principle of event symmetry which soon after appeared as a feature of this matrix model.

Then there are the new S-dualities which reverse the roles of solitons and particles, or more generally, solitonic p-branes with fundamental p-branes. But string theory also has **instantons**, sometimes called (-1)-branes because they have one less dimension than particles which are 0-branes and two fewer than strings which are 1-branes. Instantons are excitations of a field which exist for an instant. Their importance in non-abelian gauge theories such as QCD has been known for many years and now they are playing a starring role in string theories too. In passing through a duality transformation the instanton must reverse its role with a fundamental (-1)-brane and what other character can that be than a space-time event? Like particles and any other p-brane instantons have statistics; a symmetry over their permutations. This symmetry must be dual to a corresponding symmetry of space-time events; event symmetry.

I have now given seven bits of evidence that event symmetry is a feature of string theory. Some of them are more convincing than others. None of them are absolutely conclusive. The final proof would be a version of string theory which explicitly exhibited event symmetry and which was equivalent to the familiar string theories. I cannot offer that yet, but I can describe some **string inspired supersymmetries** which appear to lead the way. These supersymmetries are especially elegant and, of course, they include event symmetry.

String Inspired Symmetry

Superstrings are, of course, full of supersymmetry. They also have other symmetry which comes in various forms and includes all the types of symmetry which have been observed in nature, as well as almost all others which have ever been studied but never yet seen. String theory is meant to be a unified theory of everything so its symmetries should also be unified but apparently they are not. When a set of physical equations is found their symmetry does not always jump out at you from the start. For example, Maxwell's equations for electromagnetism at first only appeared to have rotational and translational invariance. Later they were found to be invariant under the Poincaré group of special relativity and then they were found to have an internal gauge symmetry. These symmetries can be made much more explicit by reformulating them in a different but equivalent way. It is likely that string theories also have much more symmetry than we now recognise, but it is hidden because we are forcing ourselves to write the equations in terms of concepts which we are accustomed to.



There have been many discoveries or near discoveries of new symmetry in string theory, but there is one which I found particularly inspirational. It was the string inspired symmetries of Michio Kaku. Symmetry is about groups so to discover a new symmetry all you really need is a way of defining an associative product with an inverse and a unit on whatever objects come to mind. So how might open ended strings be multiplied? Strings can interact by joining together at their ends so we could think about multiplying them in a similar way. Think of open strings as continuous paths through space starting at one point and ending at another. We will multiply them together by joining them together if the end of the first coincides with the start of the second, cancelling out the part where they join.

Take one string A starting at a point W passing through point X and ending at point Y and multiply it by another string B which starts at Y, passes back through X and ends at Z. B follows the same path in reverse as A took from X to Y. The product C=AB is then the path from W to Z passing through X and following the same path as A between W and X and the same path as B between X and Z. This product of strings is nicely associative, i.e. (AB)C = A(BC) but it fails miserably to make a group. It has no unit, no inverses and it only defines multiplication for strings which join together at their ends.

What we are looking for is the stringy generalisation of gauge symmetry. The group elements of ordinary local gauge theories are described by a field, that is an element of the base group at each event in space-time.

For example, if we are talking about the U(1) gauge symmetry of the electromagnetic field there is an element of U(1) (i.e. a complex number of modulus one) at each event. In other words the gauge transformation is specified by a function f(X) from space-time events X to the complex numbers. The charged matter fields are gauge transformed by multiplying by this phase factor at each event with the accompanying gauge transformation of the electromagnetic field. To generalise this, think of events in space-time as possible points that a particle worldline can pass through. The stringy generalisation of a gauge transformation would be specified by a function f(A) from all possible string paths A to the complex numbers. A string path is just one of the path segments through space-time which we have already thought about. So what we are really looking for is a group of objects with a complex number assigned to each string.

Gauge transformations are multiplied together by on a simple event by event basis. If f(X) is one gauge transformation and g(X) is another, then the product h(X) is just,

$$h(X) = f(X)g(X)$$

For strings we do things a little differently like this,

$$h(C) = \sum f(A)g(B) \quad : \quad C = AB$$

The sum is over all pairs of strings A and B whose product according to the previous definition is C. For a complete field there would be an infinite number of such strings and the sum becomes a difficult to define integral, but we will not worry about this detail just yet.

This definition of string gauge fields actually includes ordinary particle field gauge transformations if a particle at X is identified with a zero length string which starts and ends at the same point X. A little thought will show that string fields which are non-zero only for such strings will multiply together in the same way as particle fields. Now we can also see that this multiplication has a group-like identity. It is the string field which is equal to one for every zero length string and zero for all others. Not all string fields have inverses for this multiplication but some do, and the set of those that do forms a group. This group is then what we will consider as the general gauge group for continuous open strings. It is essentially the symmetry which Kaku defined in 1988.

Of course we would need to define some model of string dynamics which was invariant under the action of this group. That is what Kaku tried to do with some success.

These open strings, however, are less interesting than closed strings, formed from closed loops. Indeed open string theory is incomplete without closed strings along side. Kaku tried to work out a version of gauge symmetry which also works for closed strings. It is not so easy. Closed strings can interact by coming together and joining where they touch to form a single loop, but if you multiply loops together by joining them in this way you do not get an associative algebra like we did by joining open strings at their ends. Kaku solved the problem by looking at the commutators of the product and defining a supersymmetry in a clever way, or at least he almost solved it. In fact there were cases which did not quite work out. The symmetry was flawed and sadly it never proved useful as a way to understand string gauge symmetry.

Discrete String Theory

Now I will turn to another question. Are strings discrete? In string theory as we currently know it there is not much indication that string theory is discrete. Strings are described as continuous loops in space. However, there has been some interesting work by Susskind and others which does seem to suggest that string theory could be discrete. It may be possible to describe strings as objects made of small partons strung together. These partons would not exist as hard objects but can be conceptually subdivided and rejoined. They are points on the string which describe the topology of its interactions.

If the partons can be subdivided then they must be permitted to have fractional statistics. They must live on the string world sheet. The statistics of a whole loop of string would be the sum of the fractional statistics of its partons and would be an integer or half integer so that the string can live in three-dimensional space. If space-time is event-symmetric and we wish to consider event-symmetric string field theory, then a discrete string approach is essential. The partons of the string can be tied to the events through which the string passes. It will be permitted to pass through space-time events in any order it likes. In this way strings can tie together the events of space-time and provide an origin of topology in an otherwise unstructured event-symmetric universe.

If strings are formed from loops of partons with fractional statistics then it seems natural to allow them to be knotted. We should look for ways of describing this algebraically in an event-symmetric string theory.

String theorists are now also turning to higher-dimensional p-brane theories. If strings can be made of partons then surfaces, or 2-branes, can be made from strings. The process could continue ad infinitum. Space-time itself might be viewed as a membrane built in this way. There may be structures of all dimensions in physics. The two-dimensional string world sheets and three-dimensional space-time are more visible only because they stand out as a consequence of some as yet unknown quirk in the maths.

Event-Symmetric Open String Theory

In 1994 I decided that if I was to do anything serious with the principle of event symmetry I would have to apply it to string theory. String theory seemed to be crying out for a new type of symmetry and I thought that event symmetry could be a part of it. The obvious place to begin was from was Kaku's string gauge symmetry. They can be reconstructed for discrete strings with interesting results. Imagine space-time as a large number N of discrete events which are arbitrarily numbered 1, 2, ... , N. In analogy to continuous strings, an open ended string will be defined simply by the sequence of events it passes through. An example would be

$$A = 15213$$

A general string of length 4 might be written

$$B = abcd$$

a, b, c and d are variables for the events the string passes through.

The shortest permissible strings have length 2 because they must have at least start and end points, even if these coincide at the same event. Strings can be any finite length from the 2 upwards.

These strings are taken as the defining basis of a vector space. This is just a way of saying that we are going to look at fields defined over these strings as we did for continuous fields. The field is a function from the set of all strings to the complex numbers. Those fields can be added, subtracted and multiplied by complex number constants like vectors, so we call the collection of fields over strings a vector space.

I define multiplication of strings where the end of one coincides with the start of the other by joining them together and summing over all possibilities where identical events are cancelled. If they do not meet it is convenient to define the product to be zero. e.g., using a dot for the product

$$5431.12 = 5432$$
$$1234.4351 = 123351 + 1251$$
$$637.346 = 0$$

The multiplication is associative. It defines not a product for the strings, but a product for the vector space. It also has a unit. Just as the unit for continuous strings came from the shortest strings with just the same start and end point, so also the unit for this algebra is the sum,

$$I = 11 + 22 + 33 + ... + NN$$

What I have defined then, is an infinite-dimensional unital associative algebra.

From any such algebra a group can be formed simply by taking the subset of everything which has an inverse. This group could be the algebra of a symmetry of discrete open strings. Of course we would need to define some model of string dynamics which was invariant under the action of this group. This can be done in the same way as it is done for random matrix models. In fact, what I have defined is really just an extension of matrix algebra since the sub-algebra formed of strings of length two multiplies in the same way as N by N matrices.

A benefit of the discrete string version is that it is easy to go from the bosonic discrete open string to the supersymmetric version. Strings of even length are taken to be bosonic and strings of odd length are taken to be fermionic. This describes a rather simple sort of string theory which does not do very much except have super-symmetry. The interpretation is that these are open strings made of discrete fermionic or bosonic partons at space-time events. The model is event-symmetric in the sense that the order in which the events are numbered is irrelevant, but the transformations of event symmetry which would permute the numbering of events are not a part of the symmetry algebra. This is a disappointing failure which means that string gauge symmetry and general covariance are not yet unified for open strings.

Event-Symmetric Closed String Theory

Can we do the same thing with discrete closed strings? Kaku had attempted this with his formulation of string gauge theory so why not?

What is needed is a Lie superalgebra defined on a basis of closed discrete cycles. It actually took me quite a lot of investigation before I discovered the correct way to do this. I started by writing down strings of events just like for open strings, but if they are closed strings the starting point should not matter. For example a loop which went through the events numbered 2, 5, 3, 4 and 1 returning back to 2 can also start and return at any other of the five events, so long as it went round in the same cyclic order. This is signified by equations such as this,

$$25341 = 53412 = 34125 = 41253 = 12534$$

I found that if the number of events in a loop is even it is better to use,

$$7134 = -1347 = 3471 = -4713$$

You cannot do that for strings of odd length because you would go round the cycle and arrive back at the beginning and find that the string was minus itself. It is not easy to define a product directly for two closed strings and make it associative but to construct groups all you really need to define is a commutator in the algebra. i.e.

$$[A, B] = AB - BA$$

Commutators satisfy a special equation known as the Jacobi relation

$$[[A, B], C] + [[B, C], A] + [[C, A], B] = 0$$

Since closed strings are meant to interact by joining together I tried defining commutators by cancelling out bits of strings wherever they went through the same events. I experimented endlessly to work out which rules about sign factors could fit in with the Jacobi equations. I discovered that I could get it to work, but only for even length strings. The cancellation of common bits of string must only be done when there is an odd number of them in a row. In short there was only one way to make it work and it seemed lucky that it worked at all.

What about odd length strings, were they to be excluded? The answer was not difficult to guess as with open strings the odd length loops could be considered as fermionic. The commutators for fermionic variables must be replaced with anti-commutators where the minus sign is changed to a plus sign. These define a supersymmetry algebra in place of a classical symmetry. This was a very satisfying result. I had found myself forced to use supersymmetry for closed strings even before I had begun to think about any dynamics, or anomalies or any of the things which were usually used to justify supersymmetry in string theory.

There was one other satisfying result. The way the strings of length two commuted with all other strings was exactly what was required to define a rotation matrix acting on the vector space where events correspond to axis. A rotation can be used to permute axis, in other words, event symmetry must be part of the symmetry algebra I had discovered. This seemed to happen only by chance, if the signs had needed to be different, or it had been necessary to cancel out even length bits of string instead of odd length bits, this would simply not have worked. Yet I had had no choice in the matter. It was a sign that I was doing something right. It meant that if I built a model of strings with this supersymmetry algebra, it would have space-time symmetries unified with internal gauge symmetry; something that had never been achieved with string theories in continuum space-time.

I wanted to know if the supersymmetry algebra I had discovered was already known to mathematicians. The way the relations worked out was rather mysterious. Usually when you find something like this there turns out to be some deeper explanation of why it exists. Anything I could turn up might help me understand what to do next.

In 1995 a strange coincidence helped me out. I saw a paper about the role of Borcherds algebras in superstring theory. Borcherds was a name I recognised. The algebras had been discovered by an old friend of mine. I had become aquatinted with Richard Borcherds at high school when we used to participate in mathematics competitions. In fact Richard and I had been the joint winners of the 1978 British Mathematical Olympiad. We had both been in the same British team for the International Mathematical Olympiads two years running and then we knew each other at Cambridge University.

However, we had very different tastes in what kind of maths we liked. Richard was definite that he wanted to do pure maths, whereas I was becoming interested in mathematical physics. It was a bit of a surprise to discover 15 years later that Richard had made his name from a discovery about string theory, but he had approached the subject as a pure mathematician to study its symmetry. He had found a rigorous way to define an infinite-dimensional supersymmetry algebra of string theory which was of interest to mathematicians.

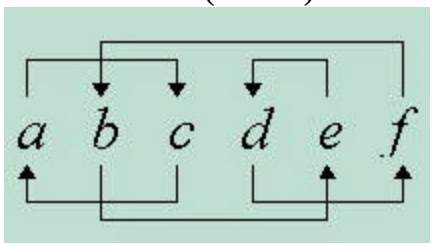
I sent an e-mail to Richard with an explanation of my super-symmetry algebra. I knew that they were not the same thing but perhaps there would be a relation between them. I was a little surprised when Richard quickly replied to tell me that my algebra did not quite work. He had found a particular case which failed to satisfy the Jacobi identity. In fact he too had already looked at Kaku's definitions of superstring gauge theory and had found that they were flawed. He easily found a similar fault in my discrete string versions.

Fortunately, as so often happens, the flaw itself gave the clue to how it should be repaired. I had to extend my algebra to include more than one loop at a time, and I had to allow them to interact by touching at more than one point of contact so that two loops which could come together and split into two others. At first it seemed like this was going to be even harder to define but I found that actually there was a conceptually simpler way to do it. This new way would give further clues about what the algebra meant.

Start with a set E of N events. Write sequences of events in the same way as for the open strings

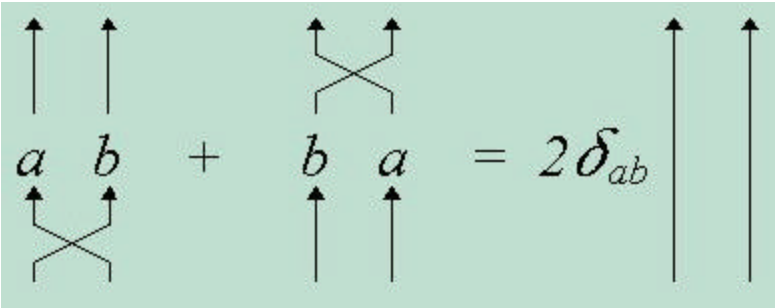
$$A = abcdef, a, b, .. \in E$$

To introduce closed loops we define permutations on these sequences. The permutation can be shown as arrows going from each event to another (or itself). An example would look like this,



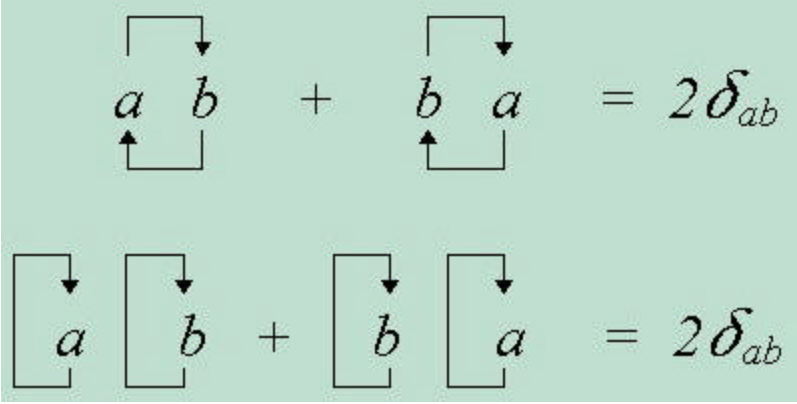
The permutation is composed of cycles. In the example there are two cycles, one of length 2 and one of length 4. But the order of the events across the page is also important.

As before these objects form the basis of a vector space. An associative algebra is defined on these objects by simply taking multiplication to be concatenation of two of these objects together. The empty sequence is a unit for this algebra. A more interesting algebra is now formed by factoring out a set of relationships among these elements. The relations are defined in the following diagram.



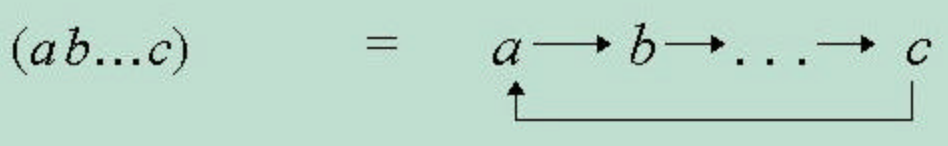
This says that the order of two events can be interchanged keeping the loop connections intact. The sign is reversed and if the two events are the same an extra reduced term must be included. To get a complete relation the ends of the string in these diagrams must be connected to something.

If they are just joined together the following two equations can be formed,



The first shows the cyclic relationships for a loop of two events. The second is the anti-commutation relation for two loops of single events.

Since the relationship can be used to order the events as we wish, it is possible to reduce every thing to a canonical basis which is a product of ordered loops. A more convenient notation without the connections shown is then introduced.



This notation allows the relations to be written in a way similar to those of the open strings, but now the cyclic relations mean that they must be interpreted as closed loops.

The algebra is associative and it is consistent to consider combinations of loops with an odd total number of events as fermionic, and with an even number of events as bosonic. So again this generates a supersymmetry using the appropriate commutator and anti-commutators. As far as I

know this infinite-dimensional supersymmetry has never been studied by mathematicians. It is possible that it can be reduced to something well known but until this is demonstrated I will assume that it is original and interesting.

Here are a few important properties of the discrete closed string algebra which did not apply to the open string algebra.

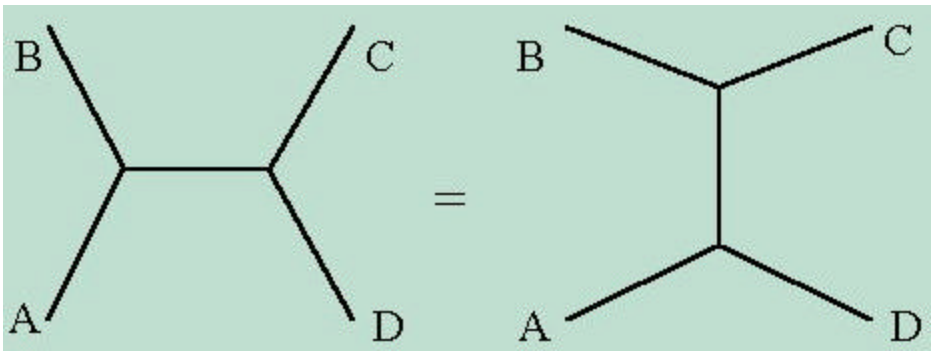
- Closed strings which do not have any events in common commute or anticommute. This is important because it can be interpreted to mean that strings only interact when they touch.
- The algebra contains a subalgebra isomorphic to a Clifford algebra. It also has a homomorphism onto a Clifford algebra which is defined by stripping out the string connections. This is important because the algebra of creation and annihilation operators for fermions is also isomorphic to a Clifford algebra. This justifies the interpretation of this algebra as a model of discrete closed strings made from fermionic partons.
- The length two strings generate an orthogonal group acting on the vector space spanning events. The symmetric group permuting events exists as a subgroup of this. It follows that the symmetry of event-symmetric space-time is included in this supersymmetry.

Algebraic String Theory

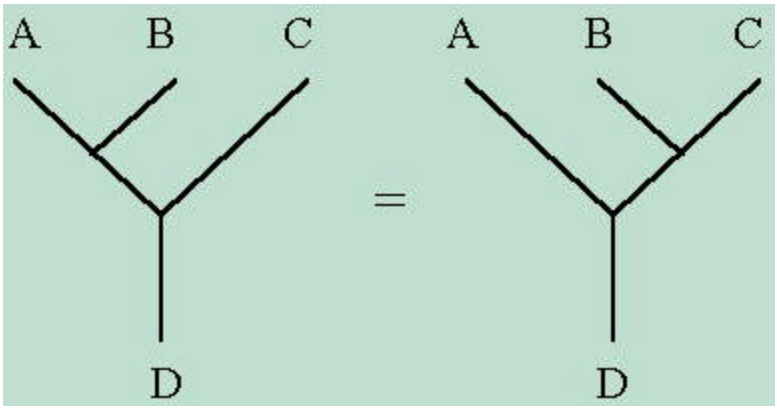
Although great strides have been taken towards an understanding of non-perturbative string theory, there is still little progress towards a formulation which shows manifest *general* covariance. In previous work I have tackled the issue by employing the *principle of event-symmetry* as a means of incorporating topology change. Space-time is regarded as a discrete set of events with the permutation group on the events being contained in the universal symmetry of physics. The symmetric group on events trivially contains the diffeomorphism group over any topology.

It may be that string theory has to be formulated in the absence of space-time which will then emerge as a derived property of the dynamics. Another interpretation of the event-symmetric approach which embodies this is that instantons are fundamental. Just as solitons may be dual to fundamental particles instantons may be dual to space-time events. Event-symmetry is then dual to instanton statistics. In that case a unification between particle statistics and gauge symmetry follows on naturally from the principle of event-symmetry. It is encouraging that this unification also appears in the matrix model of M-Theory.

The final string theory may be founded on a mixture of geometry, topology and algebra. The dual theory origins of string theory hide a clue to an underlying algebraic nature. In dual theories the s-channel and t-channel amplitudes are supposed to be equal. At tree level, in terms of Feynman diagrams this means that,



This diagram could also be distorted to look like this,



This figure is familiar to many mathematicians who recognise it as a diagrammatic representation of the associative law,

$$D = (A\ B)\ C = A\ (B\ C)$$

In developing an algebraic string theory the first step would be to define creation and annihilation operators for strings analogous to Dirac's operators for bosonic and fermionic particles. It might be possible to do this if strings are described as composites of particles like a string of beads. The creation and annihilation operators can then be strings of ordinary bosonic or fermionic operators. The algebras I have just defined are symmetry algebras for superstrings but they are also isomorphic to algebras of string creation and annihilation operators so they represent the first steps towards an algebraic theory of strings.

Is String Theory in Knots?

When I was a mathematics student at Cambridge back in 1980, I remember going to one of John Conway's popular lectures which he gave to the mathematics clubs. This one was about knot theory. Conway performed a series of tricks with bits of rope to demonstrate various properties of knots. A fundamental unsolved problem in knot theory, he told us, is to discover an algorithm which can tell when a loop of string is a knot or not.

It is possible to tie up closed loops of string into complicated tangles which can nevertheless be untied without cutting the string. But suppose I gave you a tangled loop of string. How could you determine if it could be untied?

Conway showed us a clever trick with groups which enabled him to determine that some knotted loops could not be untied, but there were others which were not classified in this way. Conway had generalised a polynomial invariant of knots first discovered by Alexandria many years ago. The Conway Polynomial was quite a powerful tool to distinguish some knots from others, but it could not separate all. I remember thinking at the time that this was a piece of pure maths which would never have any useful applications apart from providing a way of proving that your boat cannot slip its moorings, perhaps. Mathematicians delight in this kind of problems.

Ten years later a dramatic change had taken place. Knot theory now looked like it was going to have applications to solving quantum gravity and probably other problems in condensed matter theory. Louis Kauffman had even written a substantial book called *Knots and Physics* (World Scientific). Conway's Knot Polynomial had been generalised and the problem of classifying knots seemed all but solved.

To summarise, I will list just a few points of interest here:

- Knot theory is important in understanding the physics of particles with fractional statistics: anyons or parafermions. These particles, which can exist in one or two dimensions have properties between fermions and bosons. The symmetric group is the symmetry of fermions and bosons, while the braid group from knot theory plays the same role for anyons.
- Knot theory is important in canonically quantised quantum gravity. Where knotted loop states provide a basis of solutions to the quantum gravity equations. This is described in the important loop representation of quantum gravity.
- Knot theory is closely related to quantum groups. These are a generalisation (or deformation) of classical Lie groups and are important in condensed matter theory, string theory and other physics. Knot theory seems to be very closely related to symmetry.

Quantum groups are also used to construct Topological Quantum Field theories which can be used to find invariants of manifolds.

From this point on things are going to get more technical and I am going to assume that the reader knows some maths.

Strings and knots

Knotted loops have turned out to be important in the canonical approach to quantum gravity and it is natural to wonder if these loops are the same stuff as the strings of string theory, the other important approach to quantum gravity. It would be nice to think that the two are related, surely it is not a coincidence, but we must not become carried away.

By way of illustration consider the following:

When Wheeler took some of the first steps in the development of canonical gravity he used the term "superspace" to refer to the three-dimensional geometry of space which describes the states of the theory. Similarly, in the early days of string theory, they discovered that space-time symmetry must be generalised to something they also called "superspace" . Are these two types of superspace related? Surely it is not a coincidence!

But, of course, it was just a coincidence. Wheeler's superspace has nothing to do with the new superspace of superstring theories. They are very different. Likewise, most string theorists hold the opinion that there is probably no connection between the loops of the loop representation of quantum gravity and the strings of string theory. The knot which the loops make in space cannot pass through each other without changing the quantum state discontinuously. On the other hand, superstrings can pass through each other and themselves without consequence. Despite this there is a small group of people such as John Baez and Lee Smolin who have suggested that there might be a connection all the same. The strings and loops both have a common origin in gauge theories and they share some mathematics such as quantum groups in their description.

The Symmetric Group to the Braid Group

The principle of event-symmetric space-time states that the universal symmetry of physics must have a homomorphism onto the symmetric group acting on space-time events. Now the symmetric group can be defined by the following relations among the transposition generators a_1, a_2, a_3,...

$$a_i a_j a_i = a_j a_i a_j$$
$$a_i a_i = 1$$

The braid group is defined in the same way but with only the former relation. Put into words, this means that the braid group describes a symmetry where it does not matter in which order you exchange things but if you exchange two things then exchange them again you do not necessarily get back to where you were before.

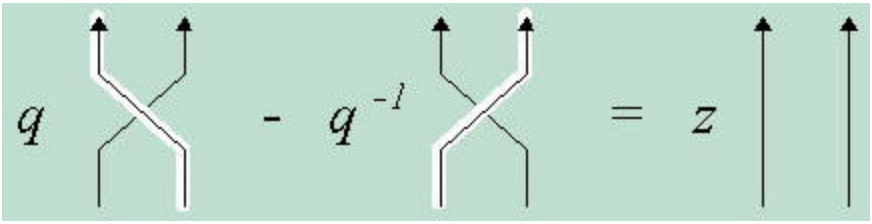
There is a homomorphism from the braid group onto the symmetric group generated by the second relation. This means that the braid group is also a candidate for part of the universal symmetry according to the principle of event-symmetric space-time. In that case space-time events would behave like particles with fractional statistics.

A String made of anyons?

It is almost certainly incorrect to model strings as loops of fermions. They must have some continuous form. To achieve this in an event-symmetric framework it will be necessary to replace the fermions with partons having fractional statistics which can be divided, i.e. anyons.

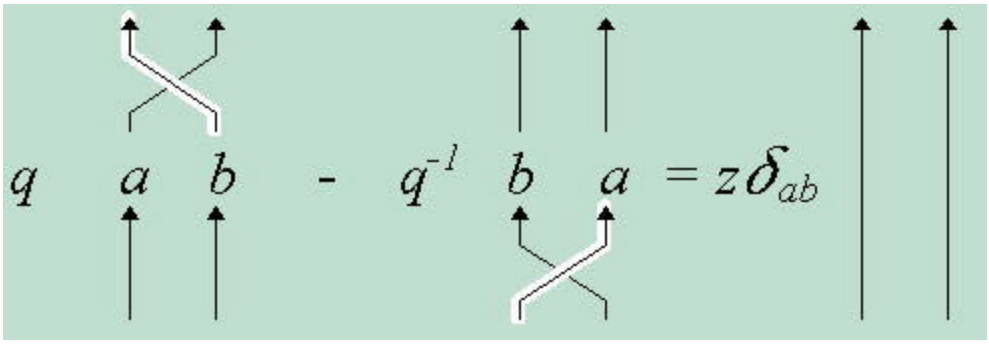
Defining creation and annihilation operators for anyons is not a simple matter. Various schemes have been proposed but none seem ideal. However, here we have the advantage that our anyons are strung together. The statistics and symmetries of anyons must be described by knot theory.

The commutation relations used to generate the closed string algebra will remind anyone who knows about knot polynomials of Skein relations. This suggests a generalisation may be possible if the string connections are replaced by knotted cords which can be tied. These could be subject to the familiar Skein relations which define the HOMFLY polynomial.



In the special case where q=1 and z=0 this relation says that string can pass through itself. This is what we have for the strings which join the fermions. The crucial question is, are there generalisations of the parton commutation relations which are consistent with the general Skein relation?

One way to do it is as follows, but does this define a consistent algebra? It is not easy to say without some interpretation of what these symbols mean. A deeper understanding could guide us towards the right solution.



Multiple Quantisation

Baron Carl Friedrich von Weizsäcker had an inauspicious beginning to his career as a physicist. In 1938 he had made an important contribution to the theory of the 'carbon cycle' of nuclear fusion in stars. Then in 1939 war broke out and Weizsäcker became a key scientist under Heisenberg in the team which failed to build the atomic bomb for Nazi Germany. After the war he became a director of a department in the Max Planck Institute of Physics in Göttingen, but the

centre of research in physics had then shifted to America and working in Germany at that time must have seemed like being cut off from the main action.

Perhaps that is why Weizsäcker came up with a fundamental idea which seemed completely out of touch with what anybody else was doing at the time. He proposed a bold theory of a way that space-time and physics might be constructed from a single bit of information by repeatedly applying the process of quantisation.

A binary digit or bit can take the value zero or one. You could think of a bit as about the simplest universe possible. Any amount of information can be coded using a sufficient number of bits. The universe is quantised, so quantise the bit. Now you have the quantum of spin-1/2, the spin of an electron which can take to values, spin-up or spin-down. The spin state is a unit length vector with two complex components which rotates under the action of SU(2) matrices.

This group is also a double covering of SO(3); the group of rotations in three-dimensional space. Weizsäcker wrote the two components as u_r where r = 1 or 2, so he called them urs and the theory was ur-theory, but ur- is also a prefix meaning 'original' or 'primitive' in German so there is a double meaning.

Just as bits can be combined to make volumes of information, urs can be combined by tensor products to define higher-dimensional state spaces. It is also possible to quantise a second time, each u_r of the quantum bit is replaced with a creation and annihilation operator, just as when a harmonic oscillator is quantised. This defines a more structured object which includes the symmetries of space-time. Just as quantisation of a field generates a multi-particle theory, the urs can be quantised again. This third quantisation generates a primitive form of field theory. Perhaps further quantisation can produce more of the structures of physics but the work remains incomplete.

Penrose Spin Networks

In 1971, Roger Penrose initiated an inspired attempt to derive the properties of space-time from combinatorics. Like Weizsäcker, he recognised the importance of spin-half and the way spins can be combined to make higher spins. Penrose was able to define discrete networks of spins which possessed geometric properties of three-dimensional space. Later a connection was found between the spin networks and Regge's discrete lattice approach to quantum gravity. It was discovered that spin networks solved quantum gravity in three dimensions. If only this could be extended to four dimensions we would have found the holy grail of physics; a theory of space-time combining general relativity and quantum mechanics. However, gravity in three dimensions is much simpler than in four dimensions. There are no gravitational waves in a universe with one less space dimension than ours.

But spin-networks turned out to be significant for four-dimensional quantum gravity too. Using the canonical quantisation methods which had led to the loop representation, relativists discovered that spin-networks should define a base of states for quantum gravity. If only they could discover the correct dynamics the breakthrough would be complete. There has already been much progress towards a four-dimensional theory of spin foams.

An interesting aspect to this story which makes it relevant here, is a remarkable parallel between the spin-network program started by Penrose and the ur-theory of Weizsäcker. Both are based in properties of SU(2) spinors. In ur-theory these spinors are regarded as the first quantisation of a bit, and are then quantised twice more. Spin networks are also derived by quantising SU(2) twice, but in rather different ways. SU(2) is first quantised to give the quantum group $SU_q(2)$, an algebraic deformation of the original group which was discovered in the 1980s. Then in 1992 Boulatov showed how you could define a quantisation of functions on quantum groups which formed spin networks. This achieved the same end as Weizsäcker but in a mathematically more powerful form.

What all this suggests is that multiple quantisation is of some fundamental importance to physics. It had been known since nearly the beginning of quantum theory that second quantisation was the way to construct quantum field theory, but this has always been regarded as a quirk rather than a fundamental feature. The first quantisation is often seen as a mistake of little significance. Some physicists even want to get rid of the term second quantisation because they dislike that interpretation so much. It is possible that they will turn out to be utterly wrong and Weizsäcker's multiple quantisation will be seen as a great insight many years ahead of its time when he first wrote about it in 1955.

What is Quantisation?

Quantisation as a formal process was introduced by Dirac as a generalisation of Heisenberg's mechanics of non-commuting matrices. Dirac showed that in principle you can take any classical system based on a principle of least action and turn it into a quantum theory. You just have to systematically find the momenta p_i corresponding to each position variable x_i in the system and then substitute operators for each position and momentum such that they satisfy a commutator relation,

$$[x_i, p_j] = \hbar \delta_{ij}$$

The operators act on a state wavefunction Ψ which evolves according to a general form of the Schrödinger equation.

If Planck's constant h were zero this would merely mean that all operators commute like real numbers, which is what happens in classical mechanics. Quantum mechanics is said to be a deformation because it reduces to classical mechanics as a special case.

It is rather curious that this process of quantisation exists. We now think of classical mechanics as just an approximation to the real quantum mechanics. The fact that it is possible to derive the quantum mechanics from the classical approximation through a process of quantisation is just a handy trick of nature to which we should attach no great significance, or should we?

The fact that we have to do a second quantisation to get field theory is also just a curiosity, after all, it only works exactly for a simple non-relativistic system of non-interacting electrons. In the real world the Schrödinger equation must be modified to make it relativistic and gauged to

introduce forces *between* the first and second quantisation. This certainly mucks up the procedure. Then again, it is *very* curious that things should work that way at all. Could multiple quantisation as we now understand it nevertheless be an echo of some deep feature of the final theory which just happens to become messed up as that theory is reduced to the approximation we know of it?

In modern times the term quantisation has been used to mean things other than what Dirac and Feynman meant. A symmetry from a classical matrix group like SU(N) can be quantised to give a quantum group $SU_q(N)$. Here quantisation is another type of deformation. q is a complex number parameter and in the special case where q = 1 the quantum group reduces to the classical one. This is not quite the same process as Dirac's quantisation but the analogy goes further than just borrowing the terminology. There is a real sense in which quantising a group with q=exp(ih) is very similar to quantising a system of mechanics. The suggestion is that there is some much more general algebraic process of quantisation of which both these things are a special case. We do not yet know what that general process is.

Since Dirac's first formulation, other equivalent ways to quantise a classical system were found. The most revealing of those was Feynman's path integral. Again you could in principle take any classical system with an action and quantise it using the path integral to define how the wave function evolves. Mathematicians have found ways in which quantum groups can arise through path integration too, but it is less direct.

Path integrals may give a clearer picture of what quantisation really is. Quantising a system which has different states seems to have something about all the different ways of going from A to B which are two different states of the system. In quantum mechanics these ways are the possible time evolutions of the system between the two states but it may be possible to generalise the concept further. In quantum field theory multiple particle systems are a derived consequence of quantising a classical field theory.

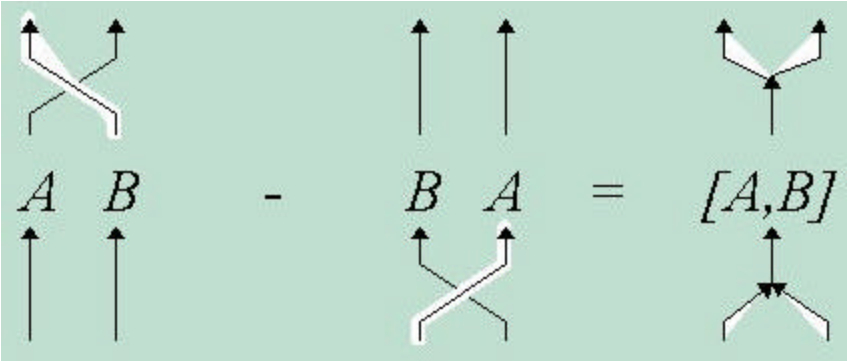
Strangely, there are other types of particle which appear as solutions of some classical systems. They are called solitary waves or just solitons. A special kind of soliton was discovered to be a solution of classical non-abelian gauge theories and they are interpreted as magnetic monopoles. What makes these especially strange is that they exist in the classical system and yet there may be a duality between monopoles and the electrically charged particles which only appear in the quantum field theory. The duality mixes up classical and quantum. There could be no clearer signal that the role of quantisation in physics is more special than it has often been given credit for.

The Supersymmetric ladder

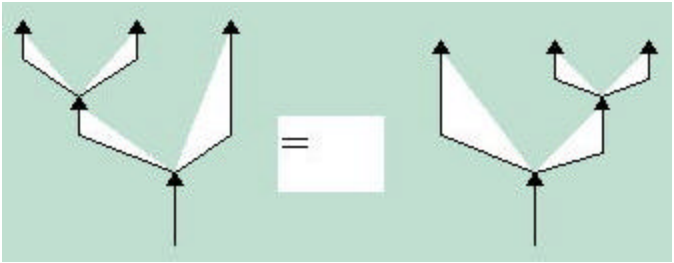
I shall now demonstrate a supersymmetric ladder construction which generalises the discrete fermion string symmetry. This construction may explain why structures of so many different dimensions are important in string theory. It may also provide some clues about what multiple quantisation is.

The fermionic operators which are strung together in the discrete string model form a Heisenberg Lie superalgebra when the strings are removed. The universal enveloping algebra of this is then a Clifford algebra. I would like to repeat the string construction starting from a general Lie superalgebra. To keep things simple I will begin with just an ordinary Lie algebra A.

As before, the elements of the Lie-algebra can be strung together on strings but this time the commutation relations will look like this,



The commutation relations can be shown to be consistent with the Jacobi relations provided the functors satisfy the following associativity relationship,



and also the similar coassociativity relationship upside down. In this way we can take out Lie algebra A and generate a new Lie algebra L(A). The process can be generalised to a Lie superalgebra. In the case where A is a Heisenberg superalgebra there is a homomorphism from L(A) onto the discrete string algebra which I defined previously. So this process can be regarded as a generalisation.

The interesting thing to do now is look at what happens if we apply the L ladder operator to the string algebra. This can be visualised by circling the discrete strings around the network so that they are replaced with tubes. The interpretation is that we generate a supersymmetry algebra as string world sheets. The ladder operator can be applied as many times as desired to generate higher-dimensional symmetry algebras. Furthermore. There is always a homomorphism from L(A) back onto A. This makes it impossible to apply the ladder operator an infinite number of times to generate a single algebra which contains all the previous ones.

This last observation raises some interesting mathematical puzzles. The algebra formed by applying the ladder operator an infinite number of times will have the property that it is isomorphic to the algebra formed by applying the ladder operator to itself. It is certainly of

interest to ask whether this situation actually arises after just a finite number of steps of the ladder. Would it be too daring to conjecture that the algebra becomes complete after only 26 steps in the ordinary Lie algebra case and 10 steps in the supersymmetric case?

To progress further it will be necessary to study more general categories like those defined by Skein relations. Mathematical physicists such as Louis Crane have looked at ways to construct n-categories by stepping up a ladder of dimensions. The symmetries I have described here could be a related to such structures. The hope is that a full theory of quantum gravity and string theory can be constructed algebraically in such a fashion.

The ladder of dimensions

In string theory there is evidence that membranes and space-times of various different dimensions play important roles. According to a principle of p-brane democracy we should not regard any particular objects as more fundamental than others. Some may be seen as composites in one manifestation but in a dual theory the roles may be reversed. What simple explanation can account for such a diversity of fundamental objects.

It is possible to go down the scale of dimensions by compactifying space-times. From M-theory in 11 dimensions or F-theory in 12 dimensions it is possible to construct the important critical string theories in 10 dimensions. The strings themselves arise by winding membranes round the compactified dimensions so embedded objects can also be reduced in dimension. To construct such theories from first principles it may be necessary to go the other way and open up hidden dimensions but what is the process which performs this operation?

The suggestion of this chapter is that it is quantisation which allows us to go back up the dimensional ladder. This is supported in string theory by the observation that second quantised string theory in 10 dimensions is first quantised M-theory in 11 dimensions. In general we should expect a k-times quantised D-dimensional theory to correspond to a (k-1)-times quantised theory in (D+1) dimensions.

The ultimate theory may have the property that it is equivalent to itself under quantisation. In other words, quantisation acts as a symmetry on the theory. This is consistent with the observation of classical/quantum dualities in compactified string theories. Invariance under quantisation may be a fundamental principle which explains p-brane democracy.

Quantisation raises the dimensions of objects as well. Quantisation of a p-brane generates a (p+1)-brane. Everything is ultimately built out of instantons and the process of composition is multiple quantisation, but instantons too can be regarded as higher-dimensional objects which have been compactified so the process has no bottom as well as no top.

This dream of a structured theory of p-branes invariant under quantisation will only be realised if a suitable definition of quantisation can be found. It must be an algebraic definition which can be applied recursively. The best candidate for a mathematical discipline in which such a definition may be possible is category theory and its generalisation to n-category theory. Category theory is a way to describe objects and morphisms between them. n-categories permit higher-dimensional

processes which map between morphisms. It is known that n-categories are related to n-dimensional topological quantum field theories but there is still much about them which is not understood.

Mathematical physicists such as John Baez have been studying their properties which relate beautifully to quantum theory and geometry. If the process of quantisation could be defined as a constructive mapping from an n-category to an (n+1)-category the link between dimension and quantisation would be established. A complete theory may be defined as the Ω-category which is equivalent to itself under quantisation.

The Theory of Theories

The Theory That Flies

As everybody knows, the job of a theoretical physicist is to invent theories of the universe. A non-professional might ask a physicist "What is charge?" or "What is time?" or "What is gravity?" He will be disappointed when the physicist replies that his theories do not even try to explain what these things are. Theories are just mathematical models which make predictions about how they will behave in experiments.

When pressed the physicist will probably admit that he does physics because he too seeks deeper explanations of what things are and why things are the way they are in the universe. One day he hopes to understand the most basic laws of physics and he hopes that they will provide an answer to the most difficult question of all, "Why do we exist?"

Physicists can be justly proud of the fact that almost everything in physics can be accounted for with just a small number of basic equations embracing general relativity and the standard model of particle physics. There remain many puzzles but those will probably be solved once a unified theory of quantum gravity and the other forces is found. Such a theory would be the final fundamental theory, although it will not be the end of physics. The equations may be cast in other forms but they would always be exactly equivalent. There is no *a priori* reason why such a theory should exist but, as Steven Weinberg argues in "Dreams of a Final Theory", the convergence of principles in modern physics seems to suggest that it does.

How many physicists have not wondered what principle of simplicity and beauty underlies that final theory? Could we not take an intellectual leap and work it out from what we already know? Surely the equations which describe the evolution of the universe at its most fundamental level must possess some magical properties to distinguish them all the other equations which merely describe hypothetical universes. What could be so unique about them that they take on a life of their own? As John Wheeler put it: What makes them fly?

Some people imagine that some reason for existence was present at the moment of creation. Some cause must have brought the universe into being in a "big bang" and the laws of physics were set there and then, they say. I have already argued against such temporal causality in all forms and I also see no reason to believe that the Big Bang is not a unique event in the cosmos. That leaves ontological causality which is what I am discussing here.

The Nature of Nature

If there is really a unique principle on which the laws of physics are founded then to understand it we should look for clues in the nature of nature, or as Feynman called it; the character of physical law. One thing is clear: Nature uses mathematics. If this were not the case, if nature was governed instead by a committee of demons who made nature follow their whims, then there would be little hope for us to understand physics and predict the outcome of experiments or invent new technology. Scientists would be replaced by sorcerers.

The relationship between physics and mathematics seems to be much deeper than we yet understand. In early history there was little distinction between a mathematician and a physicist but in modern times pure mathematicians have explored their subject independently of any potential application. Mathematics has an existence of its own. Those pure mathematicians have constructed a huge web of logical structures which have a remarkable inner beauty only apparent to those who take the time to learn and explore it. They would usually say that they discovered new mathematics rather than invented it. It is almost certain that another intelligence on another planet, or even in a different universe, would have mathematicians who discover the same theorems with just different notation.

What becomes so surprising is the extent to which mathematical structures are applicable to physics. Sometimes a physicist will discover a useful mathematical concept only to be told by mathematicians that they have been studying it for some time and can help out with a long list of useful theorems. Such was the case when Heisenberg formulated a theory of quantum mechanics which used matrix operations previously unfamiliar to physicists. Other examples abound, Einstein's application of non-Euclidean geometry to gravitation and, in particle physics, the extensive use of the classification of Lie groups.

Recently the mathematical theory of knots has found a place in theories of quantum gravity. Before that, mathematicians had considered it an area of pure mathematics without application (except to tying up boats of course). Now the role played by knots in fundamental physics seems so important that we might even guess that the reason space has three dimensions is that it is the only number of dimensions within which you can tie knots in strings. Such is the extent to which mathematics is used in physics that physicists find new theories by looking for beautiful mathematics rather than by trying to fit functions to empirical data as you might expect. Dirac explained that it was this way that he found his famous equation for the electron. The laws of physics seem to share the mathematician's taste for what is beautiful. It is a deep mystery as to why this should be the case. It is what Wigner called "the unreasonable effectiveness of mathematics in the natural sciences".

It has also been noted by Feynman that physical law seems to take on just such a form that it can be reformulated in several different ways. Quantum mechanics can be formulated in terms of Heisenberg's matrix mechanics, Schrödinger's wave mechanics or Feynman's path integrals. All three are mathematically equivalent but very different. It is impossible to say that one is more correct than the others.

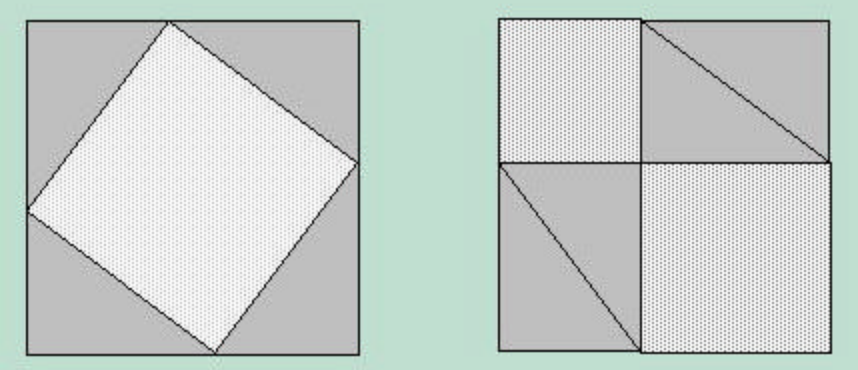
Perhaps there is a unique principle which determines the laws of physics and which explains why there is such a tight relationship between mathematics and physics. Some people imagine that the principle must be one of simplicity. The laws of physics are supposed to be the simplest possible in which intelligent life could exist. I consider this a non-starter. Simplicity is very subjective. You might attempt to define simplicity objectively by measuring the minimum length of a computer program designed to carry out a simulation of the universe but I do not accept that this is workable. The simplest complex universe might then be something like a cellular automaton and the details would depend on the syntax of the computer language we choose. A principle of simplicity would suggest that there is an optimal simplest form of the laws of physics whereas we have seen that they want to be expressed in many equally valid mathematical forms.

Furthermore, if the laws of physics were merely some isolated piece of mathematics chosen for its simple beauty then there would be no explanation why so much of mathematics is incorporated into physics. There is no reason why one set of equations should "fly". The fundamental principle of physics must be something more general. Something which embraces all of mathematics. It is the principle which explains the nature of nature. So what is it?

Can we ask why?

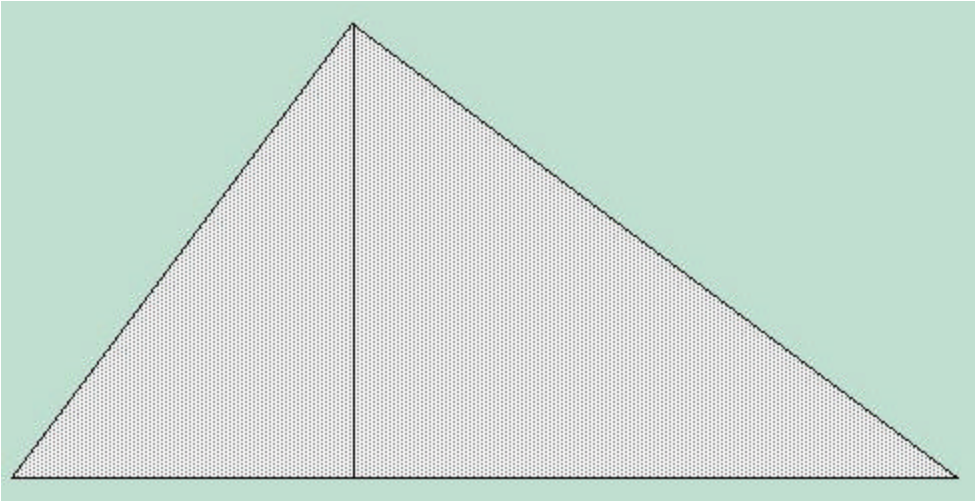
Perhaps we need to be more modest and first ask ourselves if we have the right to ask questions about why we exist. Do *why* questions make sense? Causality originally meant the principle that everything has a cause. We have come to doubt this, especially in the temporal form which says that everything has a cause in the past. A neutron left on its own for a few minutes spontaneously decays. Nothing came in from outside to make it happen and there is no clock inside a neutron which counts down to the moment at which the decay must be set off. It just happens without a cause. There are, however, reasons why neutrons decay. It can be explained in terms of the interactions to which its constituents are subject. Does everything have such an ontological cause?

First ask the question in mathematics where we think we understand the rules better. Let us take an example. Why is Pythagoras's theorem true? It is easy to prove. Look at these pictures



The two outer squares are the same size and shape and so are the areas of the four right triangles inside. Therefore the remaining areas inside must also be equal so the square on the hypotenuse is equal to the sum of the squares on the other two sides. This proof makes the theorem obviously true at a glance but is it the reason *why* it is true?

In an alternative proof a right triangle is divided in two by a line perpendicular to the hypotenuse like this



The triangle is split into two smaller right triangles and examination of the angles shows that they must both be the same shape as the original but with different size and orientation. It is known that the areas of such similar shapes are proportional to the square of the length of a side such as the hypotenuse. Once the hypotenuse of each the three triangles is identified it is then easy to see that Pythagoras's theorem follows.

Now we have two alternative proofs and hence two alternative reasons for why the theorem is true. There is no obvious relation between them so they appear to be distinct reasons. We can at least say then that there is no unique reason why something is true in mathematics. Pythagoras theorem follows by such proofs from the axioms of geometry chosen by Euclid, but modern mathematics is often founded on a different set of axioms such as those of set theory. Using sets it is possible to construct a model of the natural numbers, then the rational numbers and then the reals. Euclidean space is then defined using Cartesian co-ordinates and the distance between two

pairs of co-ordinates is defined to be the answer given by Pythagoras theorem. In this approach Pythagoras is true (for some triangles at least) by definition.

Certainly there are some theorems in mathematics which have direct proofs which can be considered to be the unique reason that they are true. In general, truth in mathematics is independent of proof and "why" questions cannot be said to have absolute answers. If this is true in mathematics then we should not expect it to be different in physics. No such absolute causality can be guaranteed. We may well find a reason "why" for many things that happen but they might not be unique and may often not exist at all. The question "why do we exist?" probably does not have a final answer but we might at least hope to understand why the laws of physics take the form that they do – as yet unknown – even if the answer is not unique.

Many Anthropic Principles

The universe is populated by an impressive menagerie of objects which exhibit organised complexity; a crystal, a flower, a planet, a star, a galaxy. They exist on all length scales from the atomic to the cosmological. Most impressive of all (that we know of) are living beings like ourselves.

Examination of the way that chemistry, nuclear physics, astrophysics, cosmology and other sciences are dependent on the details of the laws of physics suggests that the existence of so much complexity is no accident. The precise values of various constants of nature, such as the fine structure constant, seem to be just right to allow organised complexity to develop. Perhaps we might even say, just right to allow life to develop. There are many famous examples such as the nuclear resonance of carbon-12 which was predicted by Fred Hoyle in 1953. He realised that without it the higher weight elements would not have formed and we would not exist.

This observation has inspired much faith among physicists and philosophers in the *anthropic principle*. The anthropic principle supposes that the laws of physics are indeed selected so that intelligent life has a maximum chance of developing in the universe. Believers ask us to consider first why our planet Earth is so well suited to the evolution of life while other planets in the solar system seem to be more hostile. The answer is that we would not be on this planet to consider the question if it were not suitable for life to evolve here. The same principle can then be extended to the whole universe.

One way to understand the anthropic principle is to imagine that all possible universes exist with a validity which is equal to our own. When we say all possible universes we might mean any system which can be described by mathematics. Each such system has a set of physical laws which allow its structure to be determined in principle. Sometimes they will be simple and beautiful and often they will be complex and ugly. Sometimes the phenomenology of such a system will be dull or easily determined and nothing interesting will happen. Sometimes it will be so complicated that nothing can be determined, even a hypothetical computer simulation would reveal little of interest in the turmoil of those universes. Somewhere in between would exist our universe which has just the right balance of equations in its physical laws for intelligent life to exist and explore the nature of its environment.

Another interpretation of the anthropic principle, developed by Lee Smolin, is that there is one universe with a set of physical laws much as we know them. Those laws may have a number of variables which determine the physical *constants* but which can vary in certain extreme situations such as the collapse of massive stars into black holes. Universes governed by such laws might give birth to baby universes with different physical constants. Through a process of natural selection universes might evolve over many generations to have constants which are conducive to further procreation.

This might mean that they are optimised for the production of black holes and, from them, more baby universes. Within this population of worlds there will be some with laws conducive to life, indeed, the production of black holes may be linked to the existence of advanced life-forms which could have an interest in fabricating black holes as energy sources. This scenario makes a number of demands on the nature of physical laws. In particular, it is essential that some physical parameters such as the fine structure constants should be able to vary rather than being determined by some equation. Future theories of quantum gravity may tell us if this is so. Smolin's explanation of the laws of physics calls on temporal causality so it is not in line with the philosophy of this book.

Is the Anthropic Principle Enough?

The Anthropic Principle is compelling enough for us to wonder if it can determine the laws of physics on its own. I know of no convincing argument that it can. There is nothing in the anthropic principle which explains why so many of the most elegant discoveries of mathematics are so important in physics. There is nothing to explain why there is so much symmetry in physics, or why the elegant principle of least action is important or even why the laws of physics should be the same in one place as they are in another.

You might try to argue that the laws of physics have to take a certain form because otherwise they would be impossible to understand. I don't buy it! I am convinced that a suitable mathematical system, perhaps even something as simple as a cellular automaton, can include sufficient complexity that intelligent life would evolve within it. There must be a huge variety of possible forms the laws of physics could have taken and there must be many in which life evolves. In the case of cellular automata, the cellular physicists living in it would probably be able to work out the rules of the automata because its discrete nature and simple symmetry would be clear and easily uncovered. They would not need to know so much sophisticated mathematics as we do to explore the physics of our universe.

The anthropic principle may well play a role in shaping our universe. The arguments given by its proponents include lists of ways in which the laws of physics are apparently tuned to suit life. It is hard not to be swayed even taking into account that we cannot be sure that life will not develop in different unknown ways in universes with different laws. Whether or not the principle is valid as an explanation for some of the characteristics of nature and the values of its parameters I believe that there must be some other principle which explains those other aspects of physical law.

To understand the Theory of Theories we start from the same premise as we do with the anthropic principle, i.e. that all mathematically consistent models exist just as our own universe exists. We can simply take this to be our definition of existence.

We know from Feynman's Path Integral formulation of quantum mechanics that the evolution of the universe can be understood as a superposition of all possible histories that it can follow classically. The expectation values of observables are dominated by a small subset of possibilities whose contributions are reinforced by constructive interference. The same principle is at work in statistical physics where a vast state space is dominated by contributions at maximum entropy leading to thermodynamic behaviour.

We might well ask if the same can be applied to mathematical systems in general to reveal the laws of physics as a universal behaviour which dominates the space of all possible theories and which transcends details of the construction of individual theories. If this was the case then we would expect the most fundamental laws of physics to have many independent formulations with no one of them standing out as the simplest. This might be able to explain why such a large subset of mathematics is so important in physics.

Can we use the Theory of all Theories to explain why symmetry is so important in physics? There is a partial answer to this question which derives from an understanding of critical behaviour in statistical physics. Consider a lattice approximation to a Yang-Mills quantum field theory in the Euclidean sector. The Wilson discretisation preserves a discrete form of the gauge symmetry but destroys the space-time rotational symmetry. If we had more carelessly picked a discretisation scheme we would expect to break all the symmetry. We can imagine a space of discrete theories around the Yang-Mills theory for which symmetry is lost at almost all points. The symmetric continuum theory exists at a critical point in this space. As the critical point is approached correlation lengths grow and details of the discretisation are lost. Symmetries are perfectly restored in the limit, and details of all the different discretisations are washed out. If this is the case then it seems that the critical point is surrounded by a very high density of points in the space of theories.

This is exactly what we would expect if universal behaviour dominating in theory space was to exhibit high symmetry. It also suggests that a dominant theory could be reformulated in many equivalent ways without any one particular formulation being evidently more fundamentally correct than another. Perhaps ultimately there is an explanation for the unreasonable effectiveness of mathematics in physics contained in this philosophy.

If physics springs in such a fashion from all of mathematics then it seems likely that discovery of these laws will answer many old mathematical puzzles. There is no *a priori* reason to believe that mathematical theories should have some universal behaviour, but if they did it might explain why there is so much cross-reference in mathematics. Perhaps mathematicians sense intuitively when they are near the hot spots in the space of theories. They notice the heightened beauty, the multitude of unexpected connections. Eventually, left to their own devices mathematicians might be capable of finding the central source of the heat, if physicists do not get there first.

I am not alone in thinking along these lines. Physicist Holger Nielsen has made a similar conjecture and Edward Fredkin has suggested that the laws of physics may be found in a universality class of cellular automata. The general philosophy is the storyteller's paradigm. All stories are out there, told as mathematical possibilities. The rules of physics follow from a dominating universal property of the ensemble of universes.

I think therefore I am...

So, is it really possible to derive the laws of physics from pure mathematics without any reference to empirical observations as Descartes thought? If the Theory of Theories is correct then the answer should be "yes". At first it seems rather hard to make progress with the theory of theories beyond the philosophical conception, since it is necessary to define an appropriate topology and measure in the space of all mathematical theories. Mathematics is just too large for this, or is it?

Perhaps we could search for a universal behaviour in the set of all possible computer programs. The set is sufficiently diverse to cover all mathematics because, in principle, we can write a computer program to explore any mathematical problem. John Wheeler proposed this as a place to start and called it *It From Bit*. Simple computer programs can be very complex to understand, but we are not interested in understanding the details of any one. We are concerned about the universal behaviour of very big programs randomly written in some (any) computer language.

The variables of a large program would evolve in some kind of statistical manner. Perhaps the details would fade into the background and the whole could be understood using the methods of statistical physics. Suppose one system (one theory, one universe) had a number N of variables; its degrees of freedom.

$$a_1, a_2, \ldots a_N$$

In addition there must be an energy function,

$$E(a_1, a_2, \ldots a_N)$$

In the system, a possible set of values for these variables would appear with a weight given by

$$Z = \exp[- E(a_1, a_2, \ldots a_N)]$$

I have not said much about the values of these variables. They could be discrete variables or real numbers, or points on a higher-dimensional manifold. Somewhere in this complete set of systems you could find something *close* to any mathematical universe you thought of. For example, cellular automata would exist as limiting cases where the energy function forced discrete variables to follow rules.

What did I mean when I said "close"? Two different systems would be *isomorphic* if there was a one to one mapping between them which mapped the weight function of one onto the weight function of the other. We could define a distance between two systems by finding the function

mapping one to the other which minimised the correlations between them. This defines a metric space with the minimum correlation as metric.

A powerful property of metric spaces is that they can be *completed* by forming Cauchy sequences. Hence we can define a larger set of theories as the completed metric space of statistical systems. By means of this technique we include even renormalisable lattice gauge theories into the theory space. The renormalisation process can be defined as a Cauchy sequence of finite statistical systems. It remains to define a natural measure on this space and determine if it has a universal point where the total measure within any small radius of this point is larger than the measure on the rest of the space.

Needless to say, this is quite a difficult mathematical problem and I am not going to solve it. Perhaps I did not really get much further than Descartes!
