Investigating the Foundations of Physical Law
3 The most primitive concepts

What are the most primitive concepts in physics?

This lecture and the next will be the most important in the whole series. Here, we will begin to justify the claim that Foundations of Physics is a subject in its own right and with its own methodology. What we are going to do will be almost entirely inductive, but also deeply mathematical. It won't yet look like conventional mathematical physics because that is a product of complexity and emergence, not of fundamental simplicities, but it will certainly lead to it, and we will show how in the lectures that follow. We have established our methodology and laid out our mathematical toolkit. Now we will apply these to physics at its most fundamental, 'embryonic' level. If our application is correct, then physics can never be looked at in the same way again. Although we will be using the methodology generated by our philosophical reasoning, this work will not itself be an exercise in philosophy, but a completely physical discussion pitched at the most primitive level at which physical concepts can be identified as such. Questions that cannot be answered in physics at a more developed, emergent level, will, we expect, find answers when pitched at this foundational level.

When we are working at this level, the inductive mode has to go into overdrive. We need entirely new techniques of organizing the information that is confronting us in every direction. We have to use our methodological principles and the pre-prepared mathematics to see past all the conflicting claims and the layers of complication that nature, as well as ourselves, has put in the way of getting directly to the central core of information. We need to develop an instinctive feel for the 'big picture' and to be able to sort out the ideas that are truly basic from the accumulation of complexity that is beginning to interfere with them. The big simple ideas, like why time flows only one way, need big simple answers – not ones based on the emergent consequences, as we may be tempted to think – and we have to develop an aptitude for recognising the recurring patterns. But this inductive thinking is very different from pure speculation. There should be no constraints on the ideas we generate, but there are, and should be, very definite and very strong constraints on the ones we *accept*, because the methodology allows us immediately to work out the consequences that a *foundational* idea will generate. Because they have such a general application, foundational ideas that are wrong will immediately generate a mass of unacceptable consequences. Ones that are broadly correct will tend to produce smaller issues, which successive iterations are likely to resolve. The special advantage of working in a purely abstract mode and avoiding model-dependent ideas is that faults in the structure can't be hidden in the details, as they can with a model. To a large extent, it either works or it doesn't, and the main problems arise with deciding which ideas are more basic than which others.

Near the end of the first lecture we outlined those concepts that we thought the most 'primitive', based on those that survived as we changed our scale of operation. These included space, time, something representing matter in its point-like state and something representing energy or the connections between the points of matter. We can now approach the last two concepts more directly for it is clear that they, in some sense, represent the sources of the four known physical interactions, which, apart from space and time, seem to be all that could be truly fundamental in physics. The concept related to the point-like nature of matter can be identified as charge. However, this is not as simple as we might think. Apart from electric charge, the weak and strong interactions require a similar concept, and the parallel is certainly assumed in the concept of 'charge conjugation'. In fact, in view of the already successful partial unification involved in the electroweak theory, and the potential for further unification at higher energy scales, it seems more meaningful to define a single parameter with three components than to imagine that there are three totally separate concepts.

However, something seems to break the symmetry between the three charge components that we might otherwise expect in such a picture, and that has been assumed to exist at some particular energy regime in Grand Unification, and that is suggested by the common $U(1)$ component that each of the forces has. A situation like this needs to be attacked using a bold conjecture (in fact, at this level we have no alternative), based perhaps on the idea that we may have seen the pattern before. In fact, we have seen in our mathematical discussion how 'packaging' where we have two 3-dimensional structures can lead to a broken symmetry. Since an 'embryo' theory must be expected to show differences from a fully-fledged emergent one, we may reasonable take it as a provisional assumption that this is what is happening here. If this is a correct assumption, and, as far as it goes, it is perfectly compatible with Standard Model physics as we now know it, then the exact mathematical structure for symmetry-breaking which we have established in the previous lecture should be reflected in the physics that emerges, and we will investigate this in the next lecture. The point is that, if the conjecture is correct, the structures that separate the interactions should begin to emerge from the mathematical pattern we have imagined might create them. If they don't, it will soon become obvious.

The other concept that might be primitive has an established and single structure. It is the source of the gravitational interaction, and has to be differentiated from the other three. It can be called energy, mass-energy or mass. Though some people like to reserve the last term for the invariant or rest mass of material particles, I prefer to use it for the last primitive quantity, rather than energy. One reason is to emphasize its role as a source for gravity; and I want to show later how the concept of 'energy' in the quantum sense and its relation to mass seen as the gravitational source are emergent properties from the packaging of the more primitive structure. In any case, there is no physical system in which the rest mass can actually be separated from some dynamic component. All mass is, in fact, in some sense dynamic, and, in the next lecture, we will see that even this so-called rest mass can be seen as a product of the subtle quantum dynamical process known as *zitterbewegung*, in the sense that the frequency of the latter. $\approx 2mc^2 / \hbar$, is totally determined by the mass and so can be seen as a measure of it.

The concepts identified here could almost have been derived by the old method of dimensional analysis and they feature at a fundamental level in such a deeply significant result as the *CPT* theorem. They should be seen as purely abstract – none, as we will show, is any more 'real' than any other. They can all be seen almost as manifestations of pure algebra. Also, we are not assuming that these concepts are the most fundamental imaginable, but the most fundamental imaginable that can be described as purely *physical*. The first chapter in my book, *Zero to Infinity* (World Scientific, 2007), in fact, provides a kind of information process by which they and their mathematical structures can be generated from zero, and which plays a more general role in all self-organizing processes in nature. However, this is physics at a level where it can no longer be distinguished from mathematics.

Measurement

The most primitive abstract concepts must be more fundamental than either the laws of physics or the structure of matter, for neither of these can be imagined without them. It must be significant that of all the four concepts we have suggested may be primitive, only one, space, is actually susceptible to direct measurement, even though measurement is seemingly the only means we have of investigating nature. We have created many ingenious devices for measuring and recording data, but however sophisticated the system, they can always be reduced to the equivalent of moving a pointer across a scale.

For some reason, we have found it necessary to channel our measurements of space through three other conceptual structures as well, as though they had an independent existence. The entertainment industry is totally dependent on the fact that attributes of space, such as shape, colour and sound vibration, can be used to simulate things that are not meant to be spatial. The make-believe of films, sound recordings and holograms, for example, depends entirely on the idea that variations in space can induce in the viewer or listener a sense of the passage of time or the presence of matter. In fact, space is not only unique in making measurement possible, it is also universal in that any object or collection of objects will automatically create a measurement standard for space, and this will happen at all times and in all places throughout the universe.

Many people will think that we can also measure time, but, in fact, though we can sense time passing through the laws of thermodynamics, and can detect whether the simulation of a sequence of events is running in the wrong direction, we can never actually measure it. When we claim to be 'measuring' time, we are really only measuring space. Of course, it's a bit different from measuring space in the ordinary way, because it has to be repetitive. We have to find some kind of device in which something repeatedly traverses the same section of space, so that we can construct a time interval from the frequency of repetition. Traditional devices, such as pendulum clocks and watches governed by a balance spring, worked on the principle that simple harmonic motion was *isochronous*, or regular in its repeating cycle. The alternative was to use an astronomical measure, like the rotation of the Earth or the orbit of the Earth round the Sun. Relatively modern devices like atomic and digital clocks work on the same principle, with internal oscillations which are counted automatically. But in all the devices the space is observed not the time.

Clearly, we need special conditions to 'measure' time, and these would be impossible without acceleration and force. Even sending a light signal over a known path requires a reflection, and so still uses the same principles of force and repetition. Acceleration and force are second order in time and so are yet further removed from anything like a direct time measurement. The same is true of 'measurements' of mass and charge, which again are only possible in the presence of force, and again require the equivalent of a pointer moving over a scale. The mass of solid objects is determined by observing the force of gravity on them at the Earth's surface, and this involves both a spatial measurement and the use of a clock and its spatial repetitions. The same is true for astronomical objects, where we observe the dynamics due to gravity on a large scale. There are various ways of measuring the mass-energy of a particle, but they all involve using a force or a heating effect, with consequent reliance on spatial observation. Charge, of course, is not even detectable without force.

Now, the special nature of space has long been of interest to philosophers and scientists, and several have tried to reduce the whole of nature to this one quantity. Descartes believed that there was nothing other than extension in nature and that matter had no separate existence. Einstein, of course, set out to build a physics in which there was nothing but space. Time became the fourth dimension of the new combined concept of 'space-time' in the special theory of relativity, while mass-energy was expressed as space-time curvature in the general theory. He never succeeded in incorporating electromagnetism, which Kaluza and Klein decided needed the addition of a fifth dimension. No one has come close to incorporating the rest of the Standard Model into a space-like structure even by incorporating yet more dimensions, though this appears to be the ultimate aim of string theory.

But even with such 'unifications' as we have already achieved, there appear to be significant problems. A unification which created a single multi-dimensional spacelike structure would not fulfil our criterion of leaving nothing arbitrary in a fundamental theory. Space itself, and dimensionality, would remain unexplained, and there would also be no explanation of why the dimensions had such distinct manifestations. If the multidimensional 'space-time' really was the foundation for physics, there would be nothing more foundational to explain its components. In any case, quantum mechanics, which is by far the most successful physical theory ever devised, seems to be telling us that space and time are fundamentally different in ways that suggest a theory relying on a *complete* union, while suggesting some interesting consequences, will fail at some significant point. One of the most significant differences in quantum mechanics is that space is an observable, while time is not, exactly in line with our more general analysis. The indications are that the union between space and time in relativity, is an emergent one, at the first stage of complexity, not a foundational one that can't be broken down further.

It is surely significant that, although space is the only observable and measurable quantity, nature seems to be telling us that we also need time, mass and charge, concepts whose relationship with space is anything but direct. There have been numerous attempts, especially in the twentieth century, to reconstruct physics as an observer-centred subject, in which the only parameters appearing in physical theories should be those that are directly observable, and quantum mechanics originated from one such attempt. However, it quickly became clear that quantum mechanics also required quantities that could not be observed, or could not be successfully observed at the same time as others. The main effect was to change *which* quantities were to be classified as observables; it was unable to specify that observables must be exclusive. Ultimately, this is what we must expect if we are true to the principles outlined in the first lecture. Measurement and observability, however desirable, are no more likely to be universal aspects of a nature that cannot be absolutely characterised than is anything else.

Nevertheless, there is clearly something special in the fact that we can get so near in *almost* reducing the properties of the other parameters to those of space, even though the attempt gets increasingly problematic as we work through the connections with time, then mass, and finally charge. There has to be a connection which shows why space becomes a 'privileged' concept to which these others so nearly relate, and we will return to this later in the series with a quite unexpected solution.

Our main task now is to apply the principles we have established to examining the structures of the parameters in relation to each other. At first, this may appear low key, but the comparisons on a fundamental basis will show a progressive tightening of the options available and a progressive tightening of the descriptions, which will eventually lead to an understanding from which there will be no going back. Once we have reached this stage, it will become apparent that this process, which you can imagine having gone through many previous iterations, has begun to explain even the most seemingly inexplicable fundamental questions within a framework that is mathematically rigorous, though we haven't yet needed to introduce a single mathematical equation. The following three sections will be the most significant in the entire series in showing that the Foundations of Physics methodology has a power to construct a basis from which more sophisticated aspects of physics can emerge at the first level of complexity.

Conservation and nonconservation

Though we have been unable to reduce all physical concepts to aspects of space, we remain convinced that there must be *relationships* between space and the other fundamental concepts. If we can't establish *identity* between concepts there is an alternative, which physics and our methodology allows us, in establishing *symmetries* between them. At this level, we could guess that the most likely symmetry to be found is the most basic one, duality, represented by the C_2 group. If we had a perfect duality between two things, we would expect to find some characteristics in which they were identical, and some in which they were absolutely opposite. If we had exactly dual concepts, then they might well look alike in many, even most, respects, and we might even believe for a long time, that they were identical. It might then take a lot more searching before we found the areas where they were different.

Perhaps this is the position we have now reached with space and time, or even space and all the other quantities. Possibly, we should consider them as dual, rather than identical. In fact, for a fundamental theory, duality would offer a better route to explanation than identity, for, if all other concepts were versions of space, then we would have no route to finding an explanation of space itself, whereas dualities might only be explicable if we could explain the concepts themselves. So, do we have any dualities at the foundations of physics? This and the next two sections will explore three possible cases, the first being that between conservation and nonconservation.

If we were asked to guess which laws of physics might be absolutely true, a good bet would be the conservation laws. Most descriptions of physical systems seem to involve a statement, direct or indirect, that some quantity is conserved – for example, mass-energy, momentum, angular momentum, charge – while others – for example, space and time – are not. Noticeably, these are largely concerned with our fundamental parameters. Now the conservation laws of mass (or mass-energy) and charge are among the most fundamental in the whole of physics. They are also very specific. Mass and charge are not just globally, but *locally* conserved, that is a point-charge or an element of mass is conserved at a point in space and time and cannot be destroyed at one point in space and time to be recreated in another. It is as though each elementary charge and each element of mass had an *identity*, or unique label, which it carried with it throughout any changes brought about by its interactions, except in so far as a charge of one sign can be destroyed by one of the opposite sign.

What about quantities that are not conserved? Is there a property that can be called ‘nonconservation’? Surprisingly, there is and it is just as definite a property as conservation, and an exact dual to it. This is the property of the quantities we call variable, in particular, space and time, quantities which have *no identity*. We can’t single out a unit of space and time, like we can those of mass and charge, and there are three major symmetries which say exactly that. According to the translation symmetry of time, one moment in time is the same as any other. There is no way of pinning down a moment. We can ‘translate’ or move linearly along the time direction without any noticeable effect. There is, similarly, a translation symmetry of space which says that one element of space is the same as any other, and that absolute position in space is arbitrary. Space, however, is also a 3-dimensional quantity, and this leads to a third, rotation, symmetry, which says that one direction in space is the same as any other. The translation and rotation symmetries of space combine to produce its affine structure, which allows us to reconstruct a vector in space along axes in any direction.

If translation and rotation symmetry are properties of nonconserved quantities, the conserved quantities should have exactly opposite properties, and they do. Mass and charge are both translation *a*symmetric and we can guess that charge, if it is truly a 3-dimensional quantity, is additionally rotation *a*symmetric. The translation asymmetry is obvious: a unit that is unique cannot be replaced by another. This is clearly true of charges. Here, we note that even though quantum mechanics says that the *wavefunctions* of identical fundamental particles are ‘indistinguishable’, this is a property of the *observed* space and time aspects, not of the charges. The same is true of the positioning: wavefunctions are extended, charges are not. For mass(-energy), though the element may undergo a continual transformation of form, it nevertheless retains its identity. (The ‘identity of energy’ was a remarkable contribution made by Oliver Lodge to the explanation of the Poynting theorem on the flow of electromagnetic energy.)

The power of the foundational method now becomes apparent when we ask if charge, as a 3-component or 3-‘dimensional’ quantity, shows any property which we might reasonable interpret as rotation *a*symmetry. Experimental evidence so far suggests very strongly that it does. The three types of ‘charge’ (electric, strong and weak) do not ‘rotate’ into each other, despite the partial unification of the electric and weak forces in the $SU(2) \times U(1)$ electroweak theory. They must be separately conserved. Now, in particle physics, the composite baryons, such as the proton and neutron, are the only particles known with net strong ‘charge’, which manifests itself as ‘baryon number’. If strong charges are conserved separately from electric and weak charges, then there is no end product for a baryonic decay except another baryon. Baryons, along with leptons, are classed as fermions. Essentially, fermions, as distinct from bosons, are particles which are sources of the weak interaction and have net weak charges. (The *W* and *Z* bosons are carrier of the weak interaction, but not sources of it – the same applies to photons with respect to the electric interaction.) Leptons cannot decay into particles which have no net weak charge, and cannot decay into baryons, which have net strong charges, so leptons can only decay into other leptons. Experimental results have repeatedly shown that both baryon and lepton number are conserved in all particle interactions. Though there have been repeated speculations that protons could decay in such a way as to violate these laws, for example into a neutral pion (with no charges) and an antielectron (with just electric and weak charges) experimental results have always contradicted them, and the limits on the proton lifetime have been extended way beyond the predictions originally made for decay on the basis of grand unification.

Another important property of nonconserved quantities, which comes very close to translation / rotation symmetry, is *gauge invariance*, which we see in both classical and quantum contexts. A quantity that is invariant remains constant when everything that can change does so. To define a conservative system, we need to define it as one in which the conservation laws apply. The universe is such a system and so are individual particles, but there are also many systems in between where they apply to a very good approximation, and physics generally operates by constructing such systems. Under the principle of gauge invariance, the field terms which determine the strength of interactions are unchanged even when there are arbitrary changes in quantities that are subject to translations (or rotations) in the space and time coordinates, such as the vector and scalar potentials, or phase changes in the quantum mechanical wavefunction. In effect, a conservative system can absorb arbitrary changes in the space and time coordinates as long as there are no changes in the values of conserved quantities, such as charge, energy, momentum and angular momentum.

From classical physics, we have the examples of the scalar electric and gravitational potentials, which are each defined, in principle, as the ratio of the source (charge or mass) to the distance from the source. The quantity will vary with the zero point from which we measure the distance. However, in real cases, it is the potential *difference* between two points which determines the energy transfer between them, and, as long as the charge or mass value is conserved, the zero position is irrelevant. In quantum terms, changing the absolute values of the nonconserved quantities, space and time, is equivalent to changing the arbitrary phase involved in the interaction, an expression of the *U*(1) symmetry involved in defining the interaction sources in terms of scalar potentials. The absolute value of the phase term in an interaction remains unknown because it has no effect on any physical outcome. We can thus divide fundamental physical quantities into those whose absolute values are significant (the conserved ones) and those where they are irrelevant (the nonconserved ones). Significantly, in the Yang-Mills theories, which govern particle interactions in the Standard Model, gauge invariance is *local*, exactly like the conservation laws. The property of local conservation of charge and mass leads to an exactly opposite dual property of local nonconservation of space and time.

Now, gauge invariance and translation and rotation symmetry are not merely passive constraints. They force us to construct physical equations in such a way that nonconserved quantities have properties exactly opposite to those of the conserved ones, and that it is *explicitly shown* that this is the case. So, we write the laws of physics in terms of *differential equations*, with the nonconserved or variable quantities expressed only in terms of the rates of change, which specify that they are not fixed. The differential equations ensure that the conserved quantities – mass and charge (and also others derived from them, such as energy, momentum and angular momentum) – remain unchanged while space and time, expressed as the differentials, dx, dt, vary absolutely. It is not enough to say that space and time have no fixed values. They must be *seen* to have none.

The intrinsic variability or nonconservation of space and time can be seen as the ultimate origin of the path-integral approach to quantum mechanics, where we must sum over all possible paths. None can be privileged. It is also responsible for many of the aspects of quantum mechanics that seem to concern the naïve realists. The fundamental meaning of nonconservation seems to require that God does ‘play dice’. So we have to accept that space and time, as nonconserved quantities, are in principle subject to absolute variation, as long as they do not violate conservation principles. It is the conservation principles alone that restrict the range of variation of space and time when particles and systems interact. On a large scale, with many such principles acting simultaneously, the variation can be reduced to the point where we can make a classical ‘measurement’. But it is not the *measurement* that makes the situation become classical. The degree of variability, in fact, becomes restricted by the application of external potentials, requiring new conservation conditions, as the isolated system interacts with its external environment (the ‘rest of the universe’). The so-called ‘collapse of the wavefunction’ is nothing more significant than the extension of an isolated quantum system to incorporate some part of its environment, so introducing a degree of decoherence.

A free electron has complete variability and can be anywhere at any time. It remains free of any conservation principle that would restrict it (except the conservation of charge). If we now bring it near to a proton, so constructing a hydrogen atom, we find that it is now subject to new principles of conservation of energy and angular momentum which apply to the system. Nevertheless, the electron can still be anywhere at any time, as long as those principles are obeyed. The electron’s position cannot be fixed, but its range of variability is no longer the whole of space, but rather that determined by the conservation principles. If we now bring our hydrogen near to another one to form a hydrogen molecule, the electron’s position is still not fixed, but its range of variability has been changed according to the new conservation of energy and angular momentum principles which apply to the hydrogen molecule. Eventually, we can extend the system to the point where the variability is below anything we can observe.

It is clear that a great deal can be learned about both the conserved and the nonconserved parameters by realising that the distinction comes from a dual pairing. There is even a well-known mathematical result which is an expression of the duality. According to Noether's theorem, every variational (i.e. variable) property in physics leads to a conserved quantity. Three classic examples of this relate to the translation-rotation properties we have already discussed.

translation symmetry of time	≡	conservation of energy
translation symmetry of space	≡	conservation of momentum
rotation symmetry of space	≡	conservation of angular momentum

According to our terminology, and the mass-energy relation $E = mc^2$, the first of these suggests that translation symmetry of time also demands the conservation of mass. This is exactly what we would expect if nonconservation and conservation were exactly dual, as applied to time and mass. If our general methodological principles are true, this would lead us to expect a corresponding link between nonconserved space and conserved charge. In fact, as these are assumed to be 3-dimensional quantities, the link would manifest itself in two different ways, referring to translation and rotation. So we should expect:

translation symmetry of time	≡	conservation of mass
translation symmetry of space	≡	conservation of value of charge
rotation symmetry of space	≡	conservation of type of charge

Is the link true? Work done as long ago as 1927 by Fritz London suggests that the translation part, at least, might be valid. In fact, it is almost obvious from the principle of gauge invariance. London showed that the conservation of electric charge was identical to 'invariance under transformations of electrostatic potential by a constant representing changes of phase', the phase changes being of the same kind as those involved in the conservation of momentum. In principle, as in gauge invariance, the electric charge is conserved while the scalar electric potential (or ratio of charge to distance from the source) is not. In fact, we can extend the result beyond the electric charge, for the strong and weak charges each have an associated potential of the same kind as the electric one (a Coulomb potential). This potential essentially determines the value of the coupling constant for the interaction and so can be said to determine the 'value of charge'. So, it looks like this result will fulfil the test.

What about the other? This is a major development. The foundational method proposes a result which is completely unprecedented, and, frankly, looks bizarre. How can we make the conservation of angular momentum relate to the conservation of *type* of charge? How can it show that there is no mutual transformation between weak, strong and electric charges, and that the laws of baryon and lepton conservation will therefore hold? It looks impossible, but there is, in fact, an extraordinarily simple explanation which relates to the broken symmetry between these forces in the Standard Model. It provides the first significant test of our whole approach. However, we have to establish a few other things before we can give the complete explanation.

Real and imaginary

The next area of investigations brings in the algebras we have considered in the previous lecture, and especially the norm 1 and norm –1 units that make up Clifford algebra. We will use 'real' to mean norm 1 quantities, those whose units square to 1, and 'imaginary' to represent the norm –1 quantities, whose units square to –1, whether or not they are vector or scalar, commutative or anticommutative. The question of whether quantities are real or imaginary is a very significant one in physics because squaring occurs in nearly all aspects of the subject, for example, in Pythagorean addition for space and time, the amplitudes in quantum mechanics, and even in the interactions of masses and charges.

Imaginary numbers have been important to quantum mechanics from the beginning, in addition to noncommutative algebras, but, even in classical physics, it was obvious from Euler's theorem that the mathematics of waveforms was greatly simplified by the use of complex numbers. The introduction of Minkowski space-time for relativity led to the development of 4-vectors, or quantities with 3 real parts and one imaginary, almost as Hamilton had imagined in the early days of quaternions. So, from the representation of space and time as a version of Pythagoras' theorem in 4-D,

$$r^2 = x^2 + y^2 + z^2 - c^2t^2 = x^2 + y^2 + z^2 + i^2c^2t^2$$

we extract the 4-vector

$$r = \mathbf{i}x + \mathbf{j}y + \mathbf{k}z + ict.$$

And from the Einstein energy-momentum relation (with c = 1)

$$m = E^2 - p_x{}^2 - p_y{}^2 - p_z{}^2 = i^2E^2 - p_x{}^2 - p_y{}^2 - p_z{}^2$$

$$p_x{}^2 + p_y{}^2 + p_z{}^2 - E^2 = p_x{}^2 + p_y{}^2 + p_z{}^2 - i^2E^2$$

we can extract the 4-vector

$$\mathbf{i}p_x + \mathbf{j}p_y + \mathbf{k}p_z + iE,$$

which, with the c terms included, becomes

$$\mathbf{i}p_xc + \mathbf{j}p_yc + \mathbf{k}p_zc + iE.$$

Why is the time or energy component imaginary compared to the space or momentum? Why do they have different norms. If we want a *physical* argument based on relativity, it is because the light signal is retarded. But we have to remember that, at the foundational level, there is no light and there is no relativity. Sometimes, people describe the 3 + 1 real-imaginary representation of space and time as a mathematical 'trick', but mathematical tricks only work if they are needed. We have seen that vectors in our representation require an imaginary fourth component (a pseudoscalar) because they derive from complexified quaternions. There seems, then, to be a possible mathematical explanation for the representation, but there is also a *physical* one.

Physics consistently tells us that quantities containing time to the first power, such as uniform velocity, have no real significance or physical meaning. This only comes when they incorporate time squared, as we find with acceleration and force. We have already seen that time 'measurement' requires these quantities, even for time-measuring devices that use light itself. This is totally consistent with what we might expect for an imaginary quantity, but there is yet another physical reason. One thing that we haven't yet discussed about imaginary quantities is that they are intrinsically dual. They have + and – solutions which can't be distinguished. It's not that we can accept one value and discard the other – we have to always include both. Any equation with a 'positive' imaginary term in its solution has to have a dual solution with 'negative' imaginary terms. If we are using imaginary numbers, then we are automatically accepting the duality that is built into them. So, we have no option if we use *it* or *ict* but to regard it as implying a duality in its sign. Now, one of the best known aspects of time is that it flows only one way, a fact that we can detect from the increased entropy or disorder that follows any physical event. Physical equations, however, seem to ignore this completely and are constructed to have two directions of *time symmetry*. Even if we can't reverse time, we can extract physical meaning from reversing the sign of the time parameter (as with *CP* violation in particle physics). This constitutes the famous reversibility paradox. However, there is no paradox at all if time really is imaginary, and the one-way flow of time comes from an entirely different aspect of the parameter (as we will see in the next section). A parallel case can be seen in relativistic quantum mechanics where two signs of the energy parameter derive from time via its representation as $\partial / \partial t$, though there is only one sign of physical energy.

If our methodology is correct, then we might expect to find a real-imaginary distinction occurring also with mass and charge. Of course, we have the problem that our picture of 'charge' is complicated by the fact that there is a broken symmetry involved, which ensures significant distinctions between all three interactions for which we suppose it is the source. On such occasions, I tend to assume that a real unbroken symmetry is there which is exact in principle, and that the breaking of the symmetry is an effect of emergence or complexity, as our analysis of the Clifford algebra would suggest. Subsequent lectures will show how this assumption can be completely justified.

It is widely believed that there is some energy regime at which the weak, strong and electric interactions would lose their distinguishing features and become alike, but they are already alike in at least one aspect. That is, that they have a 'Coulomb' or inverse square force term, representing the $U(1)$ symmetry of a scalar phase. In this, they are also like Newtonian gravity (an aspect which is imported even into general relativity as the Newtonian potential). The weak and strong interactions differ from the electric interaction in having additional terms in their force laws which give them additional properties. It is possible then that, at grand unification, these extra components could be seen to shrink, leaving all three interactions as purely Coulomb in form.

Now, the inverse square force or Coulomb force has a relatively simple explanation. It is the exact result we would expect for a charged point source in a 3-dimensional space with spherical symmetry. In all known interactions, it relates to the coupling constant. This is not strictly the charge, which is really just a pure number which indicates whether or not a particular source of one of the interactions is present or not, but one can define the *magnitude* of the charge (electric, weak or strong) in terms of the electromagnetic, weak or strong coupling constants. In quantum terms, the coupling constant squared becomes the probability of absorbing or emitting the boson that carries the interaction, and this will be zero if the particular charge is not present.

One of the big unanswered questions of physics has always been why particles with identical masses attract, whereas particles with identical charges of any kind repel. This is seen clearly if we write down the force laws for the gravitational force between masses m_1 and m_2 and the electric force between charges e_1 and e_2 over the same distance r. The force for gravity is negative, which signifies attraction, meaning that the force has the opposite direction to the space vector **r**. However, the electric force is positive, signifying repulsion, or a force in the same direction as **r**.

$$F = -constant \times \frac{m_1 m_2}{r^2}$$

$$F = constant \times \frac{e_1 e_2}{r^2}$$

The force between e_1 and e_2 will only be attractive if they have opposite signs. We can, however, find an immediate solution if we suppose that the charges are imaginary, say ie_1 and ie_2. The force laws then assume an identical form.

$$F = -constant \times \frac{m_1 m_2}{r^2}$$

$$F = -constant \times \frac{ie_1 ie_2}{r^2}$$

We are almost drawn to this solution from the knowledge that we need to accommodate three ‘charges’ with the same property, but quite distinct from each other, and that the mathematics for this is readily available in the form of a quaternion, with components such as ***i****s*, ***j****e*, ***k****w*, where *s*, *e* and *w* are the strong, electric and weak charges. In this context, we remember that a quaternion needs a real fourth term, or scalar, just as our multivariate vectors needed a pseudoscalar, and that mass is available to play this part. We can then propose that the units of charge and mass could act as the three imaginary plus one real parts of a quaternion, just as the units of space and time act as the three real and one imaginary parts of a multivariate 4-vector.

space	time	charge	mass
i*x* **j***y* **k***z*	*it*	***i****s* ***j****e* ***k****w*	1*m*

Perhaps this could be the final vindication of Hamilton, giving quaternions a direct role in nature, as well as the indirect one of being the progenitor of the 4-vectors linking space and time. In fact the quaternion representation would be logically ‘prior’, not only in the mathematical sense, but also in the physical sense, the character of space-time being predetermined by the necessity of symmetry with charge-mass, whose structure is completely determined by the quaternionic form. In addition, such a symmetry would constrain the vector character of space to the extended form required from a complexified quaternion or Clifford algebra, and not the restricted form of the Gibbs-Heaviside algebra, meaning that spin would be automatically factored in to the structure of space, and not be an unexplained additional extra brought in with quantum mechanics.

This all depends on choosing charge to be the imaginary quantity, rather than mass. So, could we have chosen mass to be imaginary instead of charge? To show that we could not, we return to the fundamental property of imaginary numbers that we discussed in relation to time: they can only exist as a dual pair, with both + and – signs. This is true of extended imaginary numbers, such as quaternions, as much of ordinary complex numbers. All the indications are that mass, as we know it, has only one sign. Whether we call it positive or negative, there is only one version. Mass is ‘unipolar’. The case is quite different with charge. There, we always have both + and – versions, whether the charges are electric, strong or weak.

This is the explanation of ‘antimatter’. For every particle with a charge structure of any kind, there has to be a particle with charges of the opposite sign. We even call the switching of particle and antiparticle by the name of *charge conjugation*, and we are fully aware that it isn’t just about particles with *electric* charge. Neutrons, which have no electric charge, have antiparticles, because the neutron still has strong and weak charges, and the antiparticle requires these to take the opposite signs. The only exception comes with particles like the photon, which have a totally zero charge structure, and are *their own* antiparticles, with only the spins reversed. Significantly, for all antiparticles, the masses are exactly the same as those of the respective particles, emphasizing once again the intrinsic unipolarity of mass by comparison with that of charge.

As with time, there is also another reason why charge must be the imaginary quantity while mass remains real. In effect, we can observe space directly by observation or measurement, or through its squared value, in Pythagoras' theorem or vector addition, whereas time can only be apprehended through its squared value in force or acceleration. Similarly, we can apprehend mass physically in two different ways, either directly or through its squared value. The direct method is through inertia or force = mass × acceleration. The apprehension via the squared quantity is through gravitation. Through inertia, we can apprehend a mass even if no other mass is present, though we need at least two masses for gravitation. In the case of charge, only apprehension through the squared quantity is available to us, via Coulomb's law; we can never apprehend an imaginary quantity like charge unless another one is present to create a real effect. A photon, for example, can only interact with a charge if it has already been radiated by another. Ultimately, in the case of both time and charge, the imaginary status is not a mathematical convention; it represents a real physical property.

Commutative and anticommutative

Already our methodology is forcing certain constraints on the way we view the fundamental parameters. We have established that there are at least two dualities which connect them, and there is, in fact, a third. We have already specified that mass and time are, respectively, scalar and pseudoscalar, which makes them commutative, while space, as a vector, must be anticommutative. If charge is correctly described by a quaternionic structure, then this must be anticommutative as well. Anticommutativity requires a quantity to be both dimensional and specifically 3-dimensional, and the reverse argument is also true. Commutative quantities, like time and mass, must be non-dimensional, or, as it is sometimes termed, one-dimensional.

In addition to dimensionality, there is another very significant consequence of anticommutativity. This is the idea of discreteness or discontinuity introduced with the fact that the three components of an anticommutative system are very much like the components of a closed discrete set. Can we now further postulate that anticommutative or dimensional quantities are necessarily also discrete or divisible, while commutative or nondimensional are correspondingly always continuous or indivisible? It seems odd to imagine that the only discrete quantities in physics are 3-

dimensional, but that seems to be the logic of what is happening, and it would be yet another remarkable consequence of Hamilton's original discovery.

Continuous quantities clearly cannot be dimensional, because a dimensional system cannot be conceived without an origin, a zero or crossover point, and this is incompatible with continuity. Also, a one-dimensional quantity cannot be measured, because scaling requires crossover points into another dimension. Though it is often claimed that a point in space has zero dimensions, a line one dimension and an area two dimensions, this is actually impossible because a single dimension cannot generate structure. A line can only be seen as a one-dimensional structure, within a two-dimensional world, which itself can only exist in a three-dimensional one, because a two-dimensional mathematical structure always necessarily generates a third.

But, does this mean, by the counter-argument, that space, as a 3-dimensional quantity, is necessarily also discrete? The answer has to be yes. If space weren't discrete we couldn't observe it. The whole of our measuring process is based on the fact that space is fundamentally discrete. In the past, this has caused confusion, due to a fundamental misunderstanding about the nature of real numbers, and the lack of a methodology which could separate the different aspects of the fundamental parameters into primitive properties. We know, for certain, that *charge*, which we think may also be 3-dimensional, is certainly discrete, because it comes in fixed point-like units or 'singularities', which are easily countable. But space is nothing like this. Mathematicians frequently represent it, or one of its dimensions, by a real number line, which gives the appearance of being continuous. However, space, unlike charge, is a nonconserved quantity, and its units cannot be fixed. Its discreteness is one that is endlessly reconstructed. The real number line is not absolutely continuous, it is *infinitely divisible*. It presents exactly the characteristics required by non-standard arithmetic and non-Archimedean geometry. It is made up of real numbers constructed by an algorithmic process, and so is necessarily countable. If it were not, then measurement, and dimensionality, would be impossible.

There is, as far as we know, a very deep distinction between space and time, beyond any consideration of their mathematical nature as real and imaginary quantities. Time cannot be split into dimensions in the same way as space, which means that it cannot be discrete and must be continuous. The physical consequences are very significant, and allow us to complete the solution of the reversibility paradox. As a 'nondimensional' and continuous quantity, time is necessarily irreversible. To reverse time, we would have to create a discontinuity or zero-point. In addition, as a nondiscrete quantity, time can never be observed, which is exactly what quantum mechanics tells us. Observability always requires discreteness. The lack of observability is the exact reason why we treat time as the *independent variable*, by comparison with space. We write dx / dt, in fundamental equations, not dt / dx, because time varies independently of our measurements, represented by *dx*, which respond in turn to the unmeasurable variation in time.

Unlike space, which is infinitely divisible, time is *absolutely continuous*, that is not divisible at all. The two conditions are mathematically opposite, about as different as any physical or mathematical properties could conceivably be, though a fundamental duality in nature allows one to be substituted for the other. Clocks, as we have seen, do not measure time, but a space with which it has an indirect relation; the divisions that we measure are those of the space which is repeatedly traversed, and they require the fact that space, as a dimensional quantity can be reversed because we can define it to have an origin. They also, very often, use the fact that space has more than one dimension.

The absolute continuity of time gives us a complete explanation of a very old paradox, one of the oldest in physics. This is the famous paradox of Achilles and the tortoise, due to Zeno of Elea. In this paradox, Achilles can run ten times faster than the Tortoise, so, over a hundred metre race, the Tortoise gets a start of ten metres. While Achilles runs ten metres to catch up, taking one second, the Tortoise runs one metre. Achilles then runs another metre while the Tortoise runs a tenth of a metre. If Achilles runs this tenth of a metre, the Tortoise will run a hundredth of a metre and so still be ahead. Achilles may be ten times faster, but he will never actually catch up. Zeno also produced several related paradoxes, including one of a flying arrow, which to go any distance, must first go half the distance, and then half of this distance, and so on, so it will never get started at all.

Everyone knows that a faster runner will always beat a slower one, given enough time to catch up, and that arrows do fly, and most discussions of these paradoxes point to their physical absurdity, but most cannot explain why they are physically absurd. The nearest that anyone has come is in seeing that it is connected with the assumption that we can divide time into the same observational units as we divide space. For example, the philosopher G. J. Whitrow writes that: 'One can, therefore, conclude that the idea of the infinite divisibility of time must be rejected, or ... one must recognize that it is ... a logical fiction.' Motion is 'impossible if time (and, correlatively, space) is divisible *ad infinitum*'. And the science writers Peter Coveney and Roger Highfield suggest that that: 'Either one can seek to deny the notion of 'becoming', in which case time assumes essentially space-like properties; or one must reject the assumption that time, like space, is infinitely divisible into ever smaller portions.' Nevertheless, there seems to be a reluctance on the part of such commentators, and others, to proceed to the logical conclusion that Zeno's paradoxes arise from the assumption that space and time are alike in their fundamental physical properties. In fact, as we have seen, it arises from the fact that, though space is indeed 'infinitely divisible into ever smaller portions', time is not divisible at all, and that the 'divisions of time' that we actually observe are only seen through the medium of space.

One of the usual strategies for tackling the problem has been to invoke calculus, and in particular to define it in terms of the 'limit' of a function as approaching a particular value. Achilles overtakes the Tortoise at the point where the limit is reached. While this leads to a solution in mathematical terms, it doesn't explain the physical reason why the limit has to be invoked. As we have seen, however, there are two valid methods of calculus, only one of which involves limits. We can now see that they are, in fact, based on differentiation with respect to two different quantities with different physical properties, namely, space and time. Calculus, in principle, has nothing to do with the distinction between continuity and discontinuity, but is concerned with whether a quantity is variable or conserved. Variable quantities can be either continuous or discontinuities, and this leads to two ways of approaching the differential. If we differentiate with a discrete quantity like space, we end up with infinitesimals and non-standard analysis (cf lecture 2). If we differentiate with respect to time, we generate standard analysis and the theory of limits. One requires an imagined line that is absolutely continuous, the other one that is infinitely divisible. It is a classic example of the 'unreasonable effectiveness of physics in mathematics' parallelling the 'unreasonable effectiveness of mathematics in physics'. Remarkably, only the method of limits can be used to 'solve' Zeno's problem, and this is because it is really concerned with differentiation with respect to time.

Exactly the same duality also applies in physics, though this time it is not quite at the most primitive level, but rather at the first level of complexity. The mathematical connection between space and time does not automatically require the kind of *physical* connection supposed by Minkowski, who renamed them as one physical concept of 'space-time'. In fact the connection is, as we will show, a result of 'packaging'. This will become clearer in the next lecture. The physical identity is denied by quantum mechanics, which proclaims that time, unlike space, is not an observable. The reason will emerge only when we make quantum mechanics relativistic. In principle, the combination of space and time in a 4-vector format, while possible mathematically, cannot be done in a physical way. We are obliged to go to the nearest physical equivalent by either making time spacelike or space timelike. We will argue that this is the origin of wave-particle duality, for the first solution makes everything discrete, or particle-like, while the second makes everything continuous, or wavelike. The mathematical connection diverges into two physical connections, neither of which is completely valid. Wave-particle duality would not exist if space-time was a truly physical quantity.

The duality, which runs through both classical and quantum physics, extends even to the existence of two forms of nonrelativistic quantum mechanics. Heisenberg gives us the particle-like solution, while Schrödinger gives us a version based on waves. Neither is more valid than the other, but the ideas of the theories cannot be mixed. In all cases, classical as well as quantum, the duality is absolute. Though the attempt has often been made to validate one at the expense of the other, such strategies have never succeeded. Nature gives us duality between discrete and continuous processes where space and time occur at the same level, and this is because of the fundamental difference between them due to their respective properties of discreteness and continuity.

There is just one issue to be resolved in relation to discreteness and continuity, but it is a very important one, with major implications for the way we interpret both quantum mechanics and gravity. Where does mass stand with regard to this question? There can only be one answer, either from symmetry or its intrinsic nondimensionality. Such a parameter can only be completely continuous in the same way as time. We have already stated that there is no mass which is fully discrete, even though we are accustomed to defining a rest mass or invariant mass for fundamental particles. No particle, in fact, has such a mass, for all are in motion with the relativistic energy which this involves, and even the 'rest' mass arises dynamically from the subtle quantum mechanical motion known as *zitterbewegung*. But mass-energy, in any case, is a continuum which is present at all points in space. This is seen in several different forms, in particular, the Higgs field or vacuum, consisting of 246 GeV of energy at every point in space, without which the rest masses could not be generated. It is possible, in fact, that the discrete and continuous options could be responsible for the respective ideas of the local and global, the transition from global to local gauge invariance being the point in the Higgs mechanism at which the continuous field leads to the generation of a discrete invariant mass. Besides the Higgs field, manifestations of continuous mass include the zero-point energy and even ordinary fields, which cannot be localised at points. It is the continuity of mass which is the reason for its 'unipolarity' or single sign, and for the absence of a zero or crossover point, which would indicate dimensionality.

The three dualities we have discussed in this lecture appear to be astonishingly exact, and in the long period I have been thinking about these ideas and of putting them to the test I have yet to find an exception. If they represented a truly primitive level in physics, this is exactly what we would expect to find. We would expect nature at this level to be neither totally conserved nor totally nonconserved, neither totally real nor totally imaginary, neither totally continuous nor totally discrete. In the case of the last distinction, it is impossible to imagine defining discreteness without also describing continuity. We can only know what something is if we also know what it isn't. The thing that seems to be completely excluded at this level is *extended* discreteness, except insofar as it applies to space. Continuity is often described as an 'illusion', but, at the fundamental level, where abstractions are dominant, 'illusions' or ideas are an intrinsic part of reality – we couldn't even have an idea unless it was somehow part of the abstractions which nature makes available. At this level, we would not expect to find the 'best fit' compromises that might appear at a more complex level.

While it has often been claimed that physics would fit ‘reality’ better if we made it totally discrete, like measurement, continuity always seems to force its way in, as in the second law of thermodynamics. The four fundamental parameters can also be interpreted in terms of system and measurement, ontology and epistemology, neither being dominant over the other. We can make the system (or theoretical superstructure) discrete, as Heisenberg did, but then continuity will appear in the measurement, as in the Heisenberg uncertainty. Alternatively, we can make the system continuous, as Schrödinger did, and find that discreteness appears in the measurement, in this case the ‘collapse of the wavefunction’. As with the divisions between conserved and nonconserved, and real and imaginary parameters, this one seems to be an exact symmetry of absolute opposites.

Peter Rowlands
Physics Department, University of Liverpool
18 March 2013