EXCERPTED FROM

STEPHEN WOLFRAM

A NEW KIND OF SCIENCE

NOTES FOR CHAPTER 9:

Fundamental Physics

Fundamental Physics

The Notion of Reversibility

■ **Page 437 · Testing for reversibility.** To show that a cellular automaton is reversible it is sufficient to check that all configurations consisting of repetitions of different blocks have different successors. This can be done for blocks up to length n in a 1D cellular automaton with k colors using

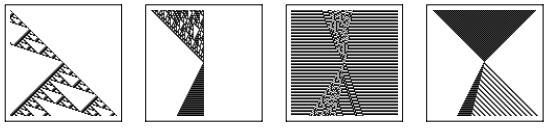
```
ReversibleQ[rule_, k_, n_] := Catch[Do[
  If[Length[Union[Table[CAStep[rule, IntegerDigits[i, k, m]],
    {i, 0, k^m - 1}]]] ≠ k^m, Throw[False]], {m, n}]; True]
```

For $k = 2$, $r = 1$ it turns out that it suffices to test only up to $n = 4$ (128 out of the 256 rules fail at $n = 1$, 64 at $n = 2$, 44 at $n = 3$ and 14 at $n = 4$); for $k = 2$, $r = 2$ it suffices to test up to $n = 15$, and for $k = 3$, $r = 1$, up to $n = 9$. But although these results suggest that in general it should suffice to test only up to $n = k^{2r}$, all that has so far been rigorously proved is that $n = k^{2r} (k^{2r} - 1) + 2r + 1$ (or $n = 15$ for $k = 2$, $r = 1$) is sufficient.

For 2D cellular automata an analogous procedure can in principle be used, though there is no upper limit on the size of blocks that need to be tested, and in fact the question of whether a particular rule is reversible is directly equivalent to the tiling problem discussed on page 213 (compare page 942), and is thus formally undecidable.

■ **Numbers of reversible rules.** For $k = 2$, $r = 1$, there are 6 reversible rules, as shown on page 436. For $k = 2$, $r = 2$ there are 62 reversible rules, in 20 families inequivalent under symmetries, out of a total of 2^{32} or about 4 billion possible rules. For $k = 3$, $r = 1$ there are 1800 reversible rules, in 172 families. For $k = 4$, $r = 1$, some of the reversible rules can be constructed from the second-order cellular automata below. Note that for any k and r, no non-trivial totalistic rule can ever be reversible.

■ **Inverse rules.** Some reversible rules are self-inverse, so that applying the same rule twice yields the identity. Other rules come in distinct pairs. Most often a rule that involves r neighbors has an inverse that also involves at most r neighbors. But for both $k = 2$, $r = 2$ and $k = 3$, $r = 1$ there turn out to be reversible rules whose inverses involve larger numbers of neighbors. For any given rule one can define the neighborhood size s to be the largest block of cells that is ever needed to determine the color of a single new cell. In general $s \le 2r + 1$, and for a simple identity or shift rule, $s = 1$. For $k = 2$, $r = 1$, it then turns out that all the reversible rules and their inverses have $s = 1$. For $k = 2$, $r = 2$, the reversible rules have values of s from 1 to 5, but their inverses have values $\bar{s}$ from 1 to 6. There are only 8 rules (the inequivalent ones being 16740555 and 3327051468) where $\bar{s} > s$, and in each case $\bar{s} = 6$ while $s = 5$. For $k = 3$, $r = 1$, there are a total of 936 rules with this property: 576, 216 and 144 with $\bar{s} = 4$, 5 and 6, and in all cases $s = 3$. Examples with $\bar{s} = 3$, 4, 5 and 6 are shown below. For arbitrary k and r, it is not clear what the maximum $\bar{s}$ can be; the only bound rigorously established so far is $\bar{s} \le r + 1/2\, k^{2r+1} (k^{2r} - 1)$.



2828556973047 3762560660157 538556225233 3066231781977

■ **Surjectivity and injectivity.** See page 959.

■ **Directional reversibility.** Even if successive time steps in the evolution of a cellular automaton do not correspond to an injective map, it is still possible to get an injective map by looking at successive lines at some angle in the spacetime evolution of the system. Examples where this works include the surjective rules 30 and 90.

■ **Page 437 · Second-order cellular automata.** Second-order elementary rules can be implemented using

```
CA2EvolveList[rule_List, {a_List, b_List}, t_Integer] :=
  Map[First, NestList[CA2Step[rule, #] &, {a, b}, t]]
CA2Step[rule_List, {a_, b_}] := {b, Mod[a + rule[[
    8 - (RotateLeft[b] + 2 (b + 2 RotateRight[b]))]], 2]}
```

where *rule* is obtained from the rule number using *IntegerDigits[n, 2, 8]*.

The combination *Drop[list, -1] + 2 Drop[list, 1]* of the result from *CA2EvolveList* corresponds to evolution according to a first-order $k = 4$, $r = 1$ rule.

■ **History.** The concept of getting reversibility in a cellular automaton by having a second-order rule was apparently first suggested by Edward Fredkin around 1970 in the context of 2D systems—on the basis of an analogy with second-order differential equations in physics. Similar ideas had appeared in numerical analysis in the 1960s in connection with so-called symmetric or self-adjoint discrete approximations to differential equations.

■ **Page 438 · Properties.** The pattern from rule 67R with simple initial conditions grows irregularly, at an average rate of about 1 cell every 5 steps. The right-hand side of the pattern from rule 173R consists three triangles that repeat progressively larger at steps of the form $2(9^s - 1)$. Rule 90R has the property that of the diamond of cells at relative positions *{{-n, 0}, {0, -n}, {n, 0}, {0, n}}* it is always true for any n that an even number are black.

■ **Page 439 · Properties.** The initial conditions used here have a single black cell on two successive initial steps. For rule 150R, however, there is no black cell on the first initial step. The pattern generated by rule 150R has fractal dimension $\text{Log}[2, 3 + \sqrt{17}] - 1$ or about 1.83. In rule 154R, each diagonal stripe is followed by at least one 0; otherwise, the positions of the stripes appear to be quite random, with a density around 0.44.

■ **Generalized additive rules.** Additive cellular automata of the kind discussed on page 952 can be generalized by allowing the new value of each cell to be obtained from combinations of cells on s previous steps. For rule 90 the combination c can be specified as *{{1, 0, 1}}*, while for rule 150R it can be specified as *{{0, 1, 0}, {1, 1, 1}}*. All generalized additive rules ultimately yield nested patterns. Starting with a list of the initial conditions for s steps, the configurations for the next s steps are given by

```
Append[Rest[list],
  Map[Mod[Apply[Plus, Flatten[c#]], 2] &, Transpose[
    Table[RotateLeft[list, {0, i}], {i, -r, r}], {3, 2, 1}]]]
```

where $r = (\text{Length}[\text{First}[c]] - 1)/2$.

Just as for ordinary additive rules on page 1091, an algebraic analysis for generalized additive rules can be given. The objects that appear are solutions to linear recurrences of order s, and in general involve s^{th} roots. For rule 150R, the configuration at step t as shown in the picture on page 439 is given by $(u^t - v^t)/\text{Sqrt}[4 + h^2]$, where $\{u, v\} = z /. \text{Solve}[z^2 == h z + 1]$ and $h = 1/x + 1 + x$. (See also page 1078.)

■ **Page 440 · Rule 37R.** Complicated structures are fairly easy to get with this rule. The initial condition *{1, 0, 1}* with all cells 0 on the previous step yields a structure that repeats but only every 666 steps. The initial condition *{{0, 1, 1}, {1, 0, 0}}* yields a pattern that grows sporadically for 3774 steps, then breaks into two repetitive structures. The typical background repeats every 3 steps.

■ **Classification of reversible rules.** In a reversible system it is possible with suitable initial conditions to get absolutely any arrangement of cells to appear at any step. Despite this, however, the overall spacetime pattern of cells is not arbitrary, but is instead determined by the underlying rules. If one starts with completely random initial conditions then class 2 and class 3 behavior are often seen. Class 1 behavior can never occur in a reversible system. Class 4 behavior can occur, as in rule 37R, but is typically obvious only if one starts say with a low density of black cells.

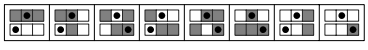
For arbitrary rules, difference patterns of the kind shown on page 250 can get both larger and smaller. In a reversible rule, such patterns can grow and shrink, but can never die out completely.

■ **Emergence of reversibility.** Once on an attractor, any system—even if it does not have reversible underlying rules—must in some sense show approximate reversibility. (Compare page 959.)

■ **Other reversible systems.** Reversible examples can be found of essentially all the types of systems discussed in this book. Reversible mobile automata can for instance be constructed using

```
Table[(IntegerDigits[i, 2, 3] → If[First[#] == 0, {#, -1},
    {Reverse[#], 1}] &)[IntegerDigits[perm[[i]], 2, 3]], {i, 8}]
```

where *perm* is an element of *Permutations[Range[8]]*. An example that exhibits complex behavior is:



Systems based on numbers are typically reversible whenever the mathematical operations they involve are invertible. Thus, for example, the system on page 121 based on successive multiplication by 3/2 is reversible by using division by 3/2. Page 905 gives another example of a reversible system based on numbers.

Multiway systems are reversible whenever both $a \to b$ and $b \to a$ are present as rules, so that the system corresponds mathematically to a semigroup. (See page 938.)

■ **Reversible computation.** Typical practical computers—and computer languages—are not even close to reversible: many inputs can lead to the same output, and there is no unique

way to undo the steps of a computation. But despite early confusion (see page 1020), it has been known since at least the 1970s that there is nothing in principle which prevents computation from being reversible. And indeed—just like with the cellular automata in this section—most of the systems in Chapter 11 that exhibit universal computation can readily be made reversible with only slight overhead.

Irreversibility and the Second Law of Thermodynamics

■ **Time reversal invariance.** The reversibility of the laws of physics implies that given the state of a physical system at a particular time, it is always possibly to work out uniquely both its future and its past. Time reversal invariance would further imply that the rules for going in each direction should be identical. To a very good approximation this appears to be true, but it turns out that in certain esoteric particle physics processes small deviations have been found. In particular, it was discovered in 1964 that the decay of the K^0 particle violated time reversal invariance at the level of about one part in a thousand. In current theories, this effect is not attributed any particularly fundamental origin, and is just assumed to be associated with the arbitrary setting of certain parameters. K^0 decay was for a long time the only example of time reversal violation that had explicitly been seen, although recently examples in B particle decays have probably also been seen. It also turns out that the only current viable theories of the apparent preponderance of matter over antimatter in the universe are based on the idea that a small amount of time reversal violation occurred in the decays of certain very massive particles in the very early universe.

The basic formalism used for particle physics assumes not only reversibility, but also so-called CPT invariance. This means that same rules should apply if one not only reverses the direction of time (T), but also simultaneously inverts all spatial coordinates (P) and conjugates all charges (C), replacing particles by antiparticles. In a certain mathematical sense, CPT invariance can be viewed as a generalization of relativistic invariance: with a speed faster than light, something close to an ordinary relativistic transformation is a CPT transformation.

Originally it was assumed that C, P and T would all separately be invariances, as they are in classical mechanics. But in 1957 it was discovered that in radioactive beta decay, C and P are in a sense each maximally violated: among other things, the correlation between spin and motion direction is exactly opposite for neutrinos and for antineutrinos that are emitted. Despite this, it was still assumed that CP and T would be true invariances. But in 1964 these too were found to be violated. Starting with a pure beam of K^0 particles, it turns out that quantum mechanical mixing processes lead after about 10^{-8} seconds to a certain mixture of $\overline{K}^0$ particles—the antiparticles of the K^0. And what effectively happens is that the amount of mixing differs by about 0.1% in the positive and negative time directions. (What is actually observed is a small probability for the long-lived component of a K^0 beam to decay into two rather than three pions. Some analysis is required to connect this with T violation.) Particle physics experiments so far support exact CPT invariance. Simple models of gravity potentially suggest CPT violation (as a consequence of deviations from pure special relativistic invariance), but such effects tend to disappear when the models are refined.

■ **History of thermodynamics.** Basic physical notions of heat and temperature were established in the 1600s, and scientists of the time appear to have thought correctly that heat is associated with the motion of microscopic constituents of matter. But in the 1700s it became widely believed that heat was instead a separate fluid-like substance. Experiments by James Joule and others in the 1840s put this in doubt, and finally in the 1850s it became accepted that heat is in fact a form of energy. The relation between heat and energy was important for the development of steam engines, and in 1824 Sadi Carnot had captured some of the ideas of thermodynamics in his discussion of the efficiency of an idealized engine. Around 1850 Rudolf Clausius and William Thomson (Kelvin) stated both the First Law—that total energy is conserved—and the Second Law of Thermodynamics. The Second Law was originally formulated in terms of the fact that heat does not spontaneously flow from a colder body to a hotter. Other formulations followed quickly, and Kelvin in particular understood some of the law's general implications. The idea that gases consist of molecules in motion had been discussed in some detail by Daniel Bernoulli in 1738, but had fallen out of favor, and was revived by Clausius in 1857. Following this, James Clerk Maxwell in 1860 derived from the mechanics of individual molecular collisions the expected distribution of molecular speeds in a gas. Over the next several years the kinetic theory of gases developed rapidly, and many macroscopic properties of gases in equilibrium were computed. In 1872 Ludwig Boltzmann constructed an equation that he thought could describe the detailed time development of a gas, whether in equilibrium or not. In the 1860s Clausius had introduced entropy as a ratio of heat to temperature, and had stated the Second Law in terms of the increase of this quantity. Boltzmann then showed that his

equation implied the so-called H Theorem, which states that a quantity equal to entropy in equilibrium must always increase with time. At first, it seemed that Boltzmann had successfully proved the Second Law. But then it was noticed that since molecular collisions were assumed reversible, his derivation could be run in reverse, and would then imply the opposite of the Second Law. Much later it was realized that Boltzmann's original equation implicitly assumed that molecules are uncorrelated before each collision, but not afterwards, thereby introducing a fundamental asymmetry in time. Early in the 1870s Maxwell and Kelvin appear to have already understood that the Second Law could not formally be derived from microscopic physics, but must somehow be a consequence of human inability to track large numbers of molecules. In responding to objections concerning reversibility Boltzmann realized around 1876 that in a gas there are many more states that seem random than seem orderly. This realization led him to argue that entropy must be proportional to the logarithm of the number of possible states of a system, and to formulate ideas about ergodicity. The statistical mechanics of systems of particles was put in a more general context by Willard Gibbs, beginning around 1900. Gibbs introduced the notion of an ensemble—a collection of many possible states of a system, each assigned a certain probability. He argued that if the time evolution of a single state were to visit all other states in the ensemble—the so-called ergodic hypothesis—then averaged over a sufficiently long time a single state would behave in a way that was typical of the ensemble. Gibbs also gave qualitative arguments that entropy would increase if it were measured in a "coarse-grained" way in which nearby states were not distinguished. In the early 1900s the development of thermodynamics was largely overshadowed by quantum theory and little fundamental work was done on it. Nevertheless, by the 1930s, the Second Law had somehow come to be generally regarded as a principle of physics whose foundations should be questioned only as a curiosity. Despite neglect in physics, however, ergodic theory became an active area of pure mathematics, and from the 1920s to the 1960s properties related to ergodicity were established for many kinds of simple systems. When electronic computers became available in the 1950s, Enrico Fermi and others began to investigate the ergodic properties of nonlinear systems of springs. But they ended up concentrating on recurrence phenomena related to solitons, and not looking at general questions related to the Second Law. Much the same happened in the 1960s, when the first simulations of hard sphere gases were led to concentrate on the specific phenomenon of long-time tails. And by the 1970s, computer experiments were mostly oriented towards ordinary differential equations and strange attractors, rather than towards systems with large numbers of components, to which the Second Law might apply. Starting in the 1950s, it was recognized that entropy is simply the negative of the information quantity introduced in the 1940s by Claude Shannon. Following statements by John von Neumann, it was thought that any computational process must necessarily increase entropy, but by the early 1970s, notably with work by Charles Bennett, it became accepted that this is not so (see page 1018), laying some early groundwork for relating computational and thermodynamic ideas.

▪ **Current thinking on the Second Law.** The vast majority of current physics textbooks imply that the Second Law is well established, though with surprising regularity they say that detailed arguments for it are beyond their scope. More specialized articles tend to admit that the origins of the Second Law remain mysterious. Most ultimately attribute its validity to unknown constraints on initial conditions or measurements, though some appeal to external perturbations, to cosmology or to unknown features of quantum mechanics.

An argument for the Second Law from around 1900, still reproduced in many textbooks, is that if a system is ergodic then it will visit all its possible states, and the vast majority of these will look random. But only very special kinds of systems are in fact ergodic, and even in such systems, the time necessary to visit a significant fraction of all possible states is astronomically long. Another argument for the Second Law, arising from work in the 1930s and 1940s, particularly on systems of hard spheres, is based on the notion of instability with respect to small changes in initial conditions. The argument suffers however from the same difficulties as the ones for chaos theory discussed in Chapter 6 and does not in the end explain in any real way the origins of randomness, or the observed validity of the Second Law.

With the Second Law accepted as a general principle, there is confusion about why systems in nature have not all dissipated into complete randomness. And often the rather absurd claim is made that all the order we see in the universe must just be a fluctuation—leaving little explanatory power for principles such as the Second Law.

▪ **My explanation of the Second Law.** What I say in this book is not incompatible with much of what has been said about the Second Law before; it is simply that I make more definite some key points that have been left vague before. In particular, I use notions of computation to specify what kinds of initial conditions can reasonably be prepared, and what kinds of measurements can reasonably be made. In a sense

what I do is just to require that the operation of coarse graining correspond to a computation that is less sophisticated than the actual evolution of the system being studied. (See also Chapters 10 and 12.)

▪ **Biological systems and Maxwell's demon.** Unlike most physical systems, biological systems typically seem capable of spontaneously organizing themselves. And as a result, even the original statements of the Second Law talked only about "inanimate systems". In the mid-1860s James Clerk Maxwell then suggested that a demon operating at a microscopic level could reduce the randomness of a system such as a gas by intelligently controlling the motion of molecules. For many years there was considerable confusion about Maxwell's demon. There were arguments that the demon must use a flashlight that generates entropy. And there were extensive demonstrations that actual biological systems reduce their internal entropy only at the cost of increases in the entropy of their environment. But in fact the main point is that if the evolution of the whole system is to be reversible, then the demon must store enough information to reverse its own actions, and this limits how much the demon can do, preventing it, for example, from unscrambling a large system of gas molecules.

▪ **Self-gravitating systems.** The observed existence of structures such as galaxies might lead one to think that any large number of objects subject to mutual gravitational attraction might not follow the Second Law and become randomized, but might instead always form orderly clumps. It is difficult to know, however, what an idealized self-gravitating system would do. For in practice, issues such as the limited size of a galaxy, its overall rotation, and the details of stellar collisions all seem to have major effects on the results obtained. (And it is presumably not feasible to do a small-scale experiment, say in Earth orbit.) There are known to be various instabilities that lead in the direction of clumping and core collapse, but how these weigh against effects such as the transfer of energy into tight binding of small groups of stars is not clear. Small galaxies such as globular clusters that contain less than a million stars seem to exhibit a certain uniformity which suggests a kind of equilibrium. Larger galaxies such as our own that contain perhaps 100 billion stars often have intricate spiral or other structure, whose origin may be associated with gravitational effects, or may be a consequence of detailed processes of star formation and explosion. (There is some evidence that older galaxies of a given size tend to develop more regularities in their structure.) Current theories of the early universe tend to assume that galaxies originally began to form as a result of density fluctuations of non-gravitational origin (and reflected in the cosmic microwave background). But there is evidence that a widespread fractal structure develops—with a correlation function of the form $r^{-1.8}$—in the distribution of stars in our galaxy, galaxies in clusters and clusters in superclusters, perhaps suggesting the existence of general overall laws for self-gravitating systems. (See also page 973.)

As mentioned on page 880, it so happens that my original interest in cellular automata around 1981 developed in part from trying to model the growth of structure in self-gravitating systems. At first I attempted to merge and generalize ideas from traditional areas of mathematical physics, such as kinetic theory, statistical mechanics and field theory. But then, particularly as I began to think about doing explicit computer simulations, I decided to take a different tack and instead to look for the most idealized possible models. And in doing this I quickly came up with cellular automata. But when I started to investigate cellular automata, I discovered some very remarkable phenomena, and I was soon led away from self-gravitating systems, and into the process of developing the much more general science in this book. Over the years, I have occasionally come back to the problem of self-gravitating systems, but I have never succeeded in developing what I consider to be a satisfactory approach to them.

▪ **Cosmology and the Second Law.** In the standard big bang model it is assumed that all matter in the universe was initially in completely random thermal equilibrium. But such equilibrium implies uniformity, and from this it follows that the initial conditions for the gravitational forces in the universe must have been highly regular, resulting in simple overall expansion, rather than random expansion in some places and contraction in others. As I discuss on page 1026 I suspect that in fact the universe as a whole probably had what were ultimately very simple initial conditions, and it is just that the effective rules for the evolution of matter led to rapid randomization, whereas those for gravity did not.

▪ **Alignment of time in the universe.** Evidence from astronomy clearly suggests that the direction of irreversible processes is the same throughout the universe. The reason for this is presumably that all parts of the universe are expanding—with the local consequence that radiation is more often emitted than absorbed, as evidenced by the fact that the night sky is dark. Olbers' paradox asks why one does not see a bright star in every direction in the night sky. The answer is that locally stars are clumped, and light from stars further away is progressively red-shifted to lower energy. Focusing a larger and larger distance away, the light one sees was emitted longer and longer ago. And eventually one sees light emitted when the universe was filled with hot opaque

gas—now red-shifted to become the 2.7K cosmic microwave background.

■ **Poincaré recurrence.** Systems of limited size that contain only discrete elements inevitably repeat their evolution after a sufficiently long time (see page 258). In 1890 Henri Poincaré established the somewhat less obvious fact that even continuous systems also always eventually get at least arbitrarily close to repeating themselves. This discovery led to some confusion in early interpretations of the Second Law, but the huge length of time involved in a Poincaré recurrence makes it completely irrelevant in practice.

■ **Page 446 · Billiards.** The discrete system I consider here is analogous to continuous so-called billiard systems consisting of circular balls in the plane. The simplest case involves one ball bouncing around in a region of a definite shape. In a rectangular region, the position is given by *Mod[a t, {w, h}]* and every point will be visited if the parameters have irrational ratios. In a region that contains fixed circular obstructions, the motion can become sensitively dependent on initial conditions. (This setup is similar to a so-called Lorentz gas.) For a system of balls in a region with cyclic boundaries, a complicated proof due to Yakov Sinai from the 1960s purports to show that every ball eventually visits every point in the region, and that certain simple statistical properties of trajectories are consistent with randomness. (See also page 971.)

■ **Page 449 · Entropy of particles in a box.** The number of possible states of a region of m cells containing q particles is *Binomial[m, q]*. In the large size limit, the logarithm of this can be approximated by $q\,$*Log[m/q]*$/m$.

■ **Page 457 · Periods in rule 37R.** With a system of size n, the maximum possible repetition period is 2^{2^n}. In actuality, however, the periods are considerably shorter. With all cells 0 on one step, and a block of nonzero cells on the next step, the periods are for example: *{1}*: *21*; *{1, 1}*: $3n-8$; *{1, 0, 1}*: *666*; *{1, 1, 1}*: $3n-8$; *{1, 0, 0, 1}*: irregular ($< 24n$; peaks at $6j+1$); *{1, 0, 0, 1, 0, 1}*: irregular ($\lesssim 2^n$; *857727* for $n = 26$; *13705406* for $n = 100$). With completely random initial conditions, there are great fluctuations, but a typical period is around $2^{n/3}$.

Conserved Quantities and Continuum Phenomena

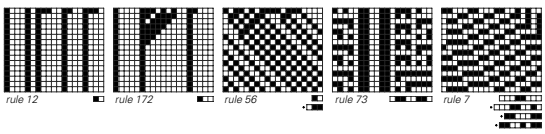
■ **Physics.** The quantities in physics that so far seem to be exactly conserved are: energy, momentum, angular momentum, electric charge, color charge, lepton number (as well as electron number, muon number and τ lepton number) and baryon number.

■ **Implementation.** Whether a k-color cellular automaton with range r conserves total cell value can be determined from

```
Catch[Do[
  (If[Apply[Plus, CAStep[rule, #] - #] ≠ 0, Throw[False]] &)[
   IntegerDigits[i, k, m]], {m, w}, {i, 0, k^m - 1}]; True]
```

where w can be taken to be k^{2r}, and perhaps smaller. Among the 256 elementary cellular automata just 5 conserve total cell value. Among the 2^{32} $k = 2$, $r = 2$ rules 428 do, and of these 2 are symmetric, and 6 are reversible, and all these are just shift and identity rules.

■ **More general conserved quantities.** Some rules conserve not total numbers of cells with given colors, but rather total numbers of blocks of cells with given forms—or combinations of these. The pictures below show the simplest quantities of these kinds that end up being conserved by various elementary rules.



Among the 256 elementary rules, the total numbers that have conserved quantities involving at most blocks of lengths 1 through 10 are *{5, 38, 66, 88, 102, 108, 108, 114, 118, 118}*.

Rules that show complicated behavior usually do not seem to have conserved quantities, and this is true for example of rules 30, 90 and 110, at least up to blocks of length 10.

One can count the number of occurrences of each of the k^b possible blocks of length b in a given state using

```
BC[list_] :=
  With[{z = Map[FromDigits[#, k] &, Partition[list, b, 1, 1]]},
    Map[Count[z, #] &, Range[0, k^b - 1]]]
```

Conserved quantities of the kind discussed here are then of the form $q \cdot$ *BC[a]* where q is some fixed list. A way to find candidates for q is to compute

```
NullSpace[Table[With[{u = Table[Random[Integer,
    {0, k - 1}], {m}]}, BC[CAStep[u]] - BC[u]], {s}]]
```

for progressively larger m and s, and to see what lists continue to appear. For block size b, k^{b-1} lists will always appear as a result of trivial conserved quantities. (With $k = 2$, for $b = 1$, *{1, 1}* represents conservation of the total number of cells, regardless of color, while for $b = 2$, *{1, 1, 1, 1}* represents the same thing, while *{0, 1, -1, 0}* represents the fact that in going along in any state the number of black-to-white transitions must equal the number of white-to-black ones.) If more than k^{b-1} lists appear, however, then some must correspond to genuine non-trivial conserved quantities. To identify any such quantity with certainty, it turns out to be enough to look at the k^{b+2r-1} states where no block of length

$b + 2r - 1$ appears more than once (and perhaps even just some fairly small subset of these).

(See also page 981.)

▪ **Other conserved quantities.** The conserved quantities discussed so far can all be thought of as taking values assigned to blocks of different kinds in a given state and then just adding them up as ordinary numbers. But one can also imagine using other operations to combine such values. Addition modulo n can be handled by inserting $Modulus \to n$ in *NullSpace* in the previous note. And doing this shows for example that rule 150 conserves the total number of black cells modulo 2. But in general not many additional conserved quantities are found in this way. One can also consider combining values of blocks by the multiplication operation in a group—and seeing whether the conjugacy class of the result is conserved.

▪ **PDEs.** In the early 1960s it was discovered that certain nonlinear PDEs support an infinite number of distinct conserved quantities, associated with so-called integrability and the presence of solitons. Systematic methods now exist to find conserved quantities that are given by integrals of polynomials of any given degree in the dependent variables and their derivatives. Most randomly chosen PDEs appear, however, to have no such conserved quantities.

▪ **Local conservation laws.** Whenever a system like a cellular automaton (or PDE) has a global conserved quantity there must always be a local conservation law which expresses the fact that every point in the system the total flux of the conserved quantity into a particular region must equal the rate of increase of the quantity inside it. (If the conserved quantity is thought of like charge, the flux is then current.) In any 1D $k = 2$, $r = 1$ cellular automaton, it follows from the basic structure of the rule that one can tell what the difference in values of a particular cell on two successive steps will be just by looking at the cell and its immediate neighbor on each side. But if the number of black cells is conserved, then one can compute this difference instead by defining a suitable flux, and subtracting its values on the left and right of the cell. What the flux should be depends on the rule. For rule 184, it can be taken to be 1 for each ■□ block, and to be 0 otherwise. For rule 170, it is 1 for both □□ and ■□. For rule 150, it is 1 for □□ and ■■, with all computations done modulo 2. In general, if the global conserved quantity involves blocks of size b, the flux can be computed by looking at blocks of size $b + 2r - 1$. What the values for these blocks should be can be found by solving a system of linear equations; that a solution must exist can be seen by looking at the de Bruijn network (see page 941), with nodes labelled by size $b + 2r - 1$ blocks, and connections by value differences between size b blocks at the center of the possible size $b + 2r$ blocks. (Note that the same basic kind of setup works in any number of dimensions.)

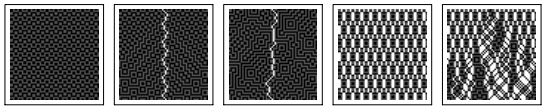
▪ **Block cellular automata.** With a rule of the form *{{1, 1}→{1, 1}, {1, 0}→{1, 0}, {0, 1}→{0, 0}, {0, 0}→{0, 1}}* the evolution of a block cellular automaton with blocks of size n can be implemented using

```
BCAEvolveList[{n_Integer, rule_}, init_, t_] :=
  FoldList[BCAStep[{n, rule}, #1, #2] &, init, Range[t]] /;
    Mod[Length[init], n] == 0
BCAStep[{n_, rule_}, a_, d_] := RotateRight[
    Flatten[Partition[RotateLeft[a, d], n] /. rule], d]
```

Starting with a single black cell, none of the $k = 2$, $n = 2$ block cellular automata generate anything beyond simple nested patterns. In general, there are $k^{n k^n}$ possible rules for block cellular automata with k colors and blocks of size n. Of these, $k^n!$ are reversible. For $k = 2$, the number of rules that conserve the total number of black cells can be computed from *q = Binomial[n, Range[0, n]]* as *Apply[Times, q^q]*. The number of these rules that are also reversible is *Apply[Times, q!]*. In general, a block cellular automaton is reversible only if its rule simply permutes the k^n possible blocks.

Compressing each block into a single cell, and n steps into one, any block cellular automaton with k colors and block size n can be translated directly into an ordinary cellular automaton with k^n colors and range $r = n/2$.

▪ **Page 461 · Block rules.** These pictures show the behavior of rule (c) starting from some special initial conditions.



The repetition period with a total of n cells can be 3^n steps. With random initial conditions, the period is typically up to about $3^{n/2}$. Starting with a block of q black cells, the period can get close to this. For $n = 20$, $q = 17$, for example, it is 31,300.

Note that even in rule (b) wraparound phenomena can lead to repetition periods that increase rapidly with n (e.g. 4820 for $n = 20$, $q = 15$), but presumably not exponentially.

In rule (d), the repetition periods can typically be larger than in rule (c): e.g. 803,780 for $n = 20$, $q = 13$.

▪ **Page 464 · Limiting procedures.** Several different limiting procedures all appear to yield the same continuum behavior for the cellular automata shown here. In the pictures on this

page a large ensemble of different initial conditions is considered, and the density of each individual cell averaged over this ensemble is computed. In a more direct analogy to actual physical systems, one would consider instead a very large number of cells, then compute the density in a single state of the system by averaging over regions that contain many cells but are nevertheless small compared to the size of the whole system.

■ **PDE approximations.** Cellular automaton (d) in the main text can be viewed as minimal discrete approximations to the diffusion equation. The evolution of densities in the ensemble average is analogous to a traditional finite difference method with a real number at each site. The cellular automaton itself uses in effect a distributed representation of the density.

■ **Diffusion equation.** In an appropriate limit the density distribution for cellular automaton (d) appears to satisfy the usual diffusion equation $\partial_t f[x, t] == c\, \partial_{xx} f[x, t]$ discussed on page 163. The solution to this equation with an impulse initial condition is $Exp[-x^2/t]$, and with a block from $-a$ to a it is $(Erf[(a-x)/\sqrt{t}\,] + Erf[(a+x)/\sqrt{t}\,])/a$.

■ **Derivation of the diffusion equation.** With some appropriate assumptions, it is fairly straightforward to derive the usual diffusion equation from a cellular automaton. Let the density of black cells at position x and time t be $f[x, t]$, where this density can conveniently be computed by averaging over many instances of the system. If we assume that the density varies slowly with position and time, then we can make series expansions such as

$$f[x + dx, t] == f[x, t] + \partial_x f[x, t]\, dx + 1/2\, \partial_{xx} f[x, t]\, dx^2 + \ldots$$

where the coordinates are scaled so that adjacent cells are at positions $x - dx$, x, $x + dx$, etc. If we then assume perfect underlying randomness, the density at a particular position must be given in terms of the densities at neighboring positions on the previous step by

$$f[x, t + dt] == p_1\, f[x - dx, t] + p_2\, f[x, t] + p_3\, f[x + dx, t]$$

Density conservation implies that $p_1 + p_2 + p_3 == 1$, while left-right symmetry implies $p_1 == p_3$. And from this it follows that

$$f[x, t + dt] == c\,(f[x - dx, t] + f[x + dx, t]) + (1 - 2\,c)\, f[x, t]$$

Performing a series expansion then yields

$$f[x, t] + dt\, \partial_t f[x, t] == f[x, t] + c\, dx^2\, \partial_{xx} f[x, t]$$

which in turn gives exactly the usual 1D diffusion equation $\partial_t f[x, t] == \xi\, \partial_{xx} f[x, t]$, where ξ is the diffusion coefficient for the system. I first gave this derivation in 1986, together with extensive generalizations.

■ **Page 464 · Non-standard diffusion.** To get ordinary diffusion behavior of the kind that occurs in gases—and is described by the diffusion equation—it is in effect necessary to have perfect uncorrelated randomness, with no structure that persists too long. But for example in the rule (a) picture on page 463 there is in effect a block of solid that persists in the middle—so that no ordinary diffusion behavior is seen. In rule (c) there is considerable apparent randomness, but it turns out that there are also fluctuations that last too long to yield ordinary diffusion. And thus for example whenever there is a structure containing s identical cells (as on page 462), this typically takes about s^2 steps to decay away. The result is that on page 464 the limiting form of the average behavior does not end up being an ordinary Gaussian.

■ **Conservation of vector quantities.** Conservation of the total number of colored cells is analogous to conservation of a scalar quantity such as energy or particle number. One can also consider conservation of a vector quantity such as momentum which has not only a magnitude but also a direction. Direction makes little sense in 1D, but is meaningful in 2D. The 2D cellular automaton used as a model of an idealized gas on page 446 provides an example of a system that can be viewed as conserving a vector quantity. In the absence of fixed scatterers, the total fluxes of particles in the horizontal and the vertical directions are conserved. But in a sense there is too much conservation in this system, and there is no interaction between horizontal and vertical motions. This can be achieved by having more complicated underlying rules. One possibility is to use a hexagonal rather than square grid, thereby allowing six particle directions rather than four. On such a grid it is possible to randomize microscopic particle motions, but nevertheless conserve overall momenta. This is essentially the model used in my discussion of fluids on page 378.

Ultimate Models for the Universe

■ **History of ultimate models.** From the earliest days of Greek science until well into the 1900s, it seems to have often been believed that an ultimate model of the universe was not far away. In antiquity there were vague ideas about everything being made of elements like fire and water. In the 1700s, following the success of Newtonian mechanics, a common assumption seems to have been that everything (with the possible exception of light) must consist of tiny corpuscles with gravity-like forces between them. In the 1800s the notion of fields—and the ether—began to develop, and in the 1880s it was suggested that atoms might be knotted vortices in the ether (see page 1044). When the electron was discovered in 1897 it was briefly thought that it might be the fundamental constituent of everything. And later it was imagined that perhaps electromagnetic fields could underlie

everything. Then after the introduction of general relativity for the gravitational field in 1915, there were efforts, especially in the 1930s, to introduce extensions that would yield unified field theories of everything (see page 1028). By the 1950s, however, an increasing number of subatomic particles were being found, and most efforts at unification became considerably more modest. In the 1960s the quark model began to explain many of the particles that were seen. Then in the 1970s work in quantum field theory encouraged the use of gauge theories and by the late 1970s the so-called Standard Model had emerged, with the Weinberg-Salam $SU(2) \otimes U(1)$ gauge theory for weak interactions and electromagnetism, and the QCD SU(3) gauge theory for strong interactions. The discoveries of the c quark, τ lepton and b quark were largely unexpected, but by the late 1970s there was widespread enthusiasm for the idea of a single "grand unified" gauge theory, based say on SU(5), that would explain all forces except gravity. By the mid-1980s failure to observe expected proton decay cast doubts on simple versions of such models, and various possibilities based on supersymmetry and groups like SO(10) were considered. Occasional attempts to construct quantum theories of gravity had been made since the 1930s, and in the late 1980s these began to be pursued more vigorously. In the mid-1980s the discovery that string theory could be given various mathematical features that were considered desirable made it emerge as the main hope for an ultimate "theory of everything". But despite all sorts of elegant mathematical work, the theory remains rather distant from observed features of our universe. In some parts of particle physics, it is still sometimes claimed that an ultimate theory is not far away, but outside it generally seems to be assumed that physics is somehow instead an endless frontier—that will continue to yield a stream of surprising and increasingly complex discoveries forever—with no ultimate theory ever being found.

▪ **Theological implications.** Some may view an ultimate model of the universe as "leaving no room for a god", while others may view it as a direct reflection of the existence of a god. In any case, knowing a complete and ultimate model does make it impossible to have miracles or divine interventions that come from outside the laws of the universe—though working out what will happen on the basis of these laws may nevertheless be irreducibly difficult.

▪ **Origins of physical models.** Considering the reputation of physics as an empirical science, it is remarkable how many significant theories were in fact first constructed on largely aesthetic grounds. Notable examples include Maxwell's equations for electromagnetism (1880s), general relativity (1915), the Dirac equation for relativistic electrons (1928), and QCD (early 1970s). This history makes it seem more plausible that one might be able to come up with an ultimate model of physics on largely aesthetic grounds, rather than mainly by working from detailed experimental observations.

▪ **Simplicity in scientific models.** To curtail absurdly complicated early scientific models Occam's razor principle that "entities should not be multiplied beyond necessity" was introduced in the 1300s. This principle has worked well in physics, where it has often proven to be the case, for example, that out of all possible terms in an equation the only ones that actually occur are the very simplest. But in a field like biology, the principle has usually been regarded as much less successful. For many complicated features are seen in biological organisms, and when there have been guesses of simple explanations for them, these have often turned out to be wrong. Much of what is seen is probably a reflection of complicated details of the history of biological evolution. But particularly after the discoveries in this book it seems likely that at least some of what appears complicated may actually be produced by very simple underlying programs—which perhaps occur because they were the first to be tried, or are the most efficient or robust. Outside of natural science, Occam's principle can sometimes be useful—typically because simplicity is a good assumption in some aspect of human behavior or motivation. In looking at well-developed technological systems or human organizations simplicity is also quite often a reasonable assumption—since over the course of time parts that are complicated or difficult to understand will tend to have been optimized away.

▪ **Numerology.** Ever since the Pythagoreans many attempts to find truly ultimate models of the universe have ended up centering on derivations of numbers that are somehow thought to be characteristic of the universe. In the past century, the emphasis has been on physical constants such as the fine structure constant $\alpha \simeq 1/137.0359896$, and usually the idea is that such constants arise directly from counting objects of some specified type using traditional discrete mathematics. A notable effort along these lines was made by Arthur Eddington in the mid-1930s, and certainly over the past twenty or so years I have received a steady stream of mail presenting such notions with varying degrees of obscurity and mysticism. But while I believe that every feature of our universe does indeed come from an ultimate discrete model, I would be very surprised if the values of constants which happen to be easy for us to measure in the end turn out to be given by simple traditional mathematical formulas.

▪ **Emergence of simple laws.** In statistical physics it is seen that universal and fairly simple overall laws often emerge

even in systems whose underlying molecular or other structure can be quite complicated. The basic origin of this phenomenon is the averaging effect of randomness discussed in Chapter 7 (technically, it is the survival only of leading operators at renormalization group fixed points). The same phenomenon is also seen in quantum field theory, where it is essentially a consequence of the averaging effect of quantum fluctuations, which have a direct mathematical analog to statistical physics.

▪ **Apparent simplicity.** Given any rules it is always possible to develop a form of description in which these rules will be considered simple. But what is interesting to ask is whether the underlying rules of the universe will seem simple—or special, say in their elegance or symmetry—with respect to forms of description that we as humans currently use.

▪ **Mechanistic models.** Until quite recently, it was generally assumed that if one were able to get at the microscopic constituents of the universe they would look essentially like small-scale analogs of objects familiar from everyday life. And so, for example, the various models of atoms from the end of the 1800s and beginning of the 1900s were all based on familiar mechanical systems. But with the rise of quantum mechanics it came to be believed throughout mainstream physics that any true fundamental model must be abstract and mathematical—and never ultimately amenable to any kind of direct mechanistic description. Occasionally there have been mechanistic descriptions used—as in the parton and bag models, and various continuum models of high-energy collisions—but they have typically been viewed only as convenient rough approximations. (Feynman diagrams may also seem superficially mechanistic, but are really just representations of quite abstract mathematical formulas.) And indeed since at least the 1960s mechanistic models have tended to carry the stigma of uninformed amateur science.

With the rise of computers there began to be occasional discussion—though largely outside of mainstream science—that the universe might have a mechanism related to computers. Since the 1950s science fiction has sometimes featured the idea that the universe or some part of it—such as the Earth—could be an intentionally created computer, or that our perception of the universe could be based on a computer simulation. Starting in the 1950s a few computer scientists considered the idea that the universe might have components like a computer. Konrad Zuse suggested that it could be a continuous cellular automaton; Edward Fredkin an ordinary cellular automaton (compare page 1027). And over the past few decades—normally in the context of amateur science—there have been a steady stream of systems like cellular automata constructed to have elements reminiscent of observed particles or forces. From the point of view of mainstream physics, such models have usually seemed quite naive. And from what I say in the main text, no such literal mechanistic model can ever in the end realistically be expected to work. For if an ultimate model is going to be simple, then in a sense it cannot have room for all sorts of elements that are immediately recognizable in terms of everyday known physics. And instead I believe that what must happen relies on the phenomena discovered in this book—and involves the emergence of complex properties without any obvious underlying mechanistic set up. (Compare page 860.)

▪ **The Anthropic Principle.** It is sometimes argued that the reason our universe has the characteristics it does is because otherwise an intelligence such as us could not have arisen to observe it. But to apply such an argument one must among other things assume that we can imagine all the ways in which intelligence could conceivably operate. Yet as we have seen in this book it is possible for highly complex behavior—ultimately not dissimilar to intelligence—to arise from simple programs in ways that we never came even close to imagining. And indeed, as we discuss in Chapter 12, it seems likely that above a fairly low threshold the vast majority of underlying rules can in fact in some way or another support arbitrarily complex computations—potentially allowing something one might call intelligence in a vast range of very different universes. (See page 822.)

▪ **Physics versus mathematics.** Theoretical physics can be viewed as taking physical input in the form of models and then using mathematics to work out the consequences. If I am correct that there is a simple underlying program for the universe, then this means that theoretical physics must at some level have only a very small amount of true physical input—and the rest must in a sense all just be mathematics.

▪ **Initial conditions.** To find the behavior of the universe one potentially needs to know not only its rule but also its initial conditions. Like the rule, I suspect that the initial conditions will turn out to be simple. And ultimately there should be traces of such simplicity in, say, the distribution of galaxies or the cosmic microwave background. But ideas like those on page 1055—as well as inflation—tend to suggest that we currently see only a tiny fraction of the whole universe, making it very difficult for example to recognize overall geometrical regularities. And it could also be that even though there might ultimately have been simple initial conditions, the current phase of our universe might be the result of some sequence of previous phases, and so effectively have much more complicated initial conditions. (Proposals discussed in quantum cosmology since the 1980s

that for example just involve requiring the universe to satisfy final but not initial boundary condition constraints do not fit well into my kinds of models.)

▪ **Consequences of an ultimate model.** Even if one knows an ultimate model for the universe, there will inevitably be irreducible difficulty in working out all its consequences. Indeed, questions like "does there exist a way to transmit information faster than light?" may boil down to issues analogous to whether it is possible to construct a configuration that has a certain property in, say, the rule 110 cellular automaton. And while some such questions may be answered by fairly straightforward computational or mathematical means, there will be no upper bound on the amount of effort that it could take to answer any particular question.

▪ **Meaning of the universe.** If the whole history of our universe can be obtained by following definite simple rules, then at some level this history has the same kind of character as a construct such as the digit sequence of π. And what this suggests is that it makes no more or less sense to talk about the meaning of phenomena in our universe as it does to talk about the meaning of phenomena in the digit sequence of π.

The Nature of Space

▪ **History of discrete space.** The idea that matter might be made up of discrete particles existed in antiquity (see page 876), and occasionally the notion was discussed that space might also be discrete—and that this might for example be a way of avoiding issues like Zeno's paradox. In 1644 René Descartes proposed that space might initially consist of an array of identical tiny discrete spheres, with motion then occurring through chains of these spheres going around in vortices—albeit with pieces being abraded off. But with the rise of calculus in the 1700s all serious fundamental models in physics began to assume continuous space. In discussing the notion of curved space, Bernhard Riemann remarked in 1854 that it would be easier to give a general mathematical definition of distance if space were discrete. But since physical theories seemed to require continuous space, the necessary new mathematics was developed and almost universally used—though for example in 1887 William Thomson (Kelvin) did consider a discrete foam-like model for the ether (compare page 988). Starting in 1930, difficulties with infinities in quantum field theory again led to a series of proposals that spacetime might be discrete. And indeed by the late 1930s this notion was fairly widely discussed as a possible inevitable feature of quantum mechanics. But there were problems with relativistic invariance, and after ideas of renormalization developed in the 1940s, discrete space seemed unnecessary, and has been out of favor ever since. Some non-standard versions of quantum field theory involving discrete space did however continue to be investigated into the 1960s, and by then a few isolated other initiatives had arisen that involved discrete space. The idea that space might be defined by some sort of causal network of discrete elementary quantum events arose in various forms in work by Carl von Weizsäcker (ur-theory), John Wheeler (pregeometry), David Finkelstein (spacetime code), David Bohm (topochronology) and Roger Penrose (spin networks; see page 1055). General arguments for discrete space were also sometimes made—notably by Edward Fredkin, Marvin Minsky and to some extent Richard Feynman—on the basis of analogies to computers and in particular the idea that a given region of space should contain only a finite amount of information. In the 1980s approximation schemes such as lattice gauge theory and later Regge calculus (see page 1054) that take space to be discrete became popular, and it was occasionally suggested that versions of these could be exact models. There have been a variety of continuing initiatives that involve discrete space, with names like combinatorial physics—but most have used essentially mechanistic models (see page 1026), and none have achieved significant mainstream acceptance. Work on quantum gravity in the late 1980s and 1990s led to renewed interest in the microscopic features of spacetime (see page 1054). Models that involve discreteness have been proposed—most often based on spin networks—but there is usually still some form of continuous averaging present, leading for example to suggestions very different from mine that perhaps this could lead to the traditional continuum description through some analog of the wave-particle duality of elementary quantum mechanics. I myself became interested in the idea of completely discrete space in the mid-1970s, but I could not find a plausible framework for it until I started thinking about networks in the mid-1980s.

▪ **Planck length.** Even in existing particle physics it is generally assumed that the traditional simple continuum description of space must break down at least below about the Planck length $Sqrt[\hbar G/c^3] \approx 2 \times 10^{-35}$ meters—since at this scale dimensional analysis suggests that quantum effects should be comparable in magnitude to gravitational ones.

▪ **Page 472 · Symmetry.** A system like a cellular automaton that consists of a large number of identical cells must in effect be arranged like a crystal, and therefore must exhibit one of the limited number of possible crystal symmetries in any particular dimension, as discussed on page 929. And even a

generalized cellular automaton constructed say on a Penrose tiling still turns out to have a discrete spatial symmetry.

■ **Page 474 · Space and its contents.** A number of somewhat different ideas about space were discussed in antiquity. Around 375 BC Plato vaguely suggested that the universe might consist of large numbers of abstract polyhedra. A little later Aristotle proposed that space is set up so as to provide a definite place for everything—and in effect to force it there. But in geometry as developed by Euclid there was at least a mathematical notion of space as a kind of uniform background. And by sometime after 300 BC the Epicureans developed the idea of atoms of matter existing in a mostly featureless void of space. In the Middle Ages there was discussion about how the non-material character of God might fit in with ideas about space. In the early 1600s the concept of inertia developed by Galileo implied that space must have a certain fundamental uniformity. And with the formulation of mechanics by Isaac Newton in 1687 space became increasingly viewed as something purely abstract, quite different in character from material objects which exist in it. Philosophers had meanwhile discussed matter—as opposed to mind—being something characterized by having spatial extent. And for example in 1643 Thomas Hobbes suggested that the whole universe might be made of the same continuous stuff, with different densities of it corresponding to different materials, and geometry being just an abstract idealization of its properties. But in the late 1600s Gottfried Leibniz suggested instead that everything might consist of discrete monads, with space emerging from the pattern of relative distances between them. Yet with the success of Newtonian mechanics such ideas had by the late 1700s been largely forgotten—leading space almost always to be viewed just in simple abstract geometrical terms. The development of non-Euclidean geometry in the mid-1800s nevertheless suggested that even at the level of geometry space could in principle have a complicated structure. But in physics it was still assumed that space itself must have a standard fixed Euclidean form—and that everything in the universe must just exist in this space. By the late 1800s, however, it was widely believed that in addition to ordinary material objects, there must throughout space be a fluid-like ether with certain mechanical and electromagnetic properties. And in the 1860s it was even suggested that perhaps atoms might just correspond to knots in this ether (see page 1044). But this idea soon fell out of favor, and when relativity theory was introduced in 1905 it emphasized relations between material objects and in effect always treated space as just some kind of abstract background, with no real structure of its own. But in 1915 general relativity introduced the idea that space could actually have a varying non-Euclidean geometry—and that this could represent gravity. Yet it was still assumed that matter was something different—that for example had to be represented separately by explicit terms in the Einstein equations. There were nevertheless immediate thoughts that perhaps at least electromagnetism could be like gravity and just arise from features of space. And in 1918 Hermann Weyl suggested that this could happen through local variations of scale or "gauge" in space, while in the 1920s Theodor Kaluza and Oskar Klein suggested that it could be associated with a fifth spacetime dimension of invisibly small extent. And from the 1920s to the 1950s Albert Einstein increasingly considered the possibility that there might be a unified field theory in which all matter would somehow be associated with the geometry of space. His main specific idea was to allow the metric of spacetime to be non-symmetric (see page 1052) and perhaps complex—with its additional components yielding electromagnetism. And he then tried to construct nonlinear field equations that would show no singularities, but would have solutions (perhaps analogous to the geons discussed on page 1054) that would exhibit various discrete features corresponding to particles—and perhaps quantum effects. But with the development of quantum field theory in the 1920s and 1930s most of physics again treated space as fixed and featureless—though now filled with various types of fields, whose excitations were set up to correspond to observed types of particles. Gravity has never fit very well into this framework. But it has always still been expected that in an ultimate quantum theory of gravity space will have to have a structure that is somehow like a quantum field. But when quantum gravity began to be investigated in earnest in the 1980s (see page 1054) most efforts concentrated on the already difficult problem of pure gravity—and did not consider how matter might enter. In the development of ordinary quantum field theories, supergravity theories studied in the 1980s did nominally support particles identified with gravitons, but were still formulated on a fixed background spacetime. And when string theory became popular in the 1980s the idea was again to have strings propagating in a background spacetime—though it turned out that for consistency this spacetime had to satisfy the Einstein equations. Consistency also typically required the basic spacetime to be 10-dimensional—with the reduction to observed 4D spacetime normally assumed to occur through restriction of the other dimensions to some kind of so-called Calabi-Yau manifold of small extent, associated excitations with various particles through an analog of the Kaluza-Klein mechanism. It has always been hoped that this kind of seemingly arbitrary setup would somehow automatically

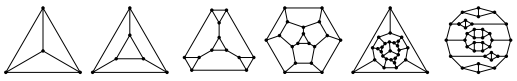
emerge from the underlying theory. And in the late 1990s there seemed to be some signs of this when dualities were discovered in various generalized string theories—notably for example between quantum particle excitations and gravitational black hole configurations. So while it remains impossible to work out all the consequences of string theories, it is conceivable that among the representations of such theories there might be ones in which matter can be viewed as just being associated with features of space.

Space as a Network

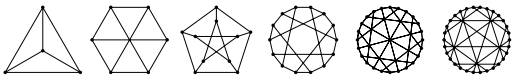
▪ **Page 476 · Trivalent networks.** With n nodes and 3 connections at each node a network must always have an even number of nodes, and a total of $3n/2$ connections. Of all possible such networks, most large ones end up being connected. The number of distinct such networks for even n from 2 to 10 is $\{2, 5, 17, 71, 388\}$. If no self connections are allowed then these numbers become $\{1, 2, 6, 20, 91\}$, while if neither self nor multiple connections are allowed (yielding what are often referred to as cubic or 3-regular graphs), the numbers become $\{0, 1, 2, 5, 19, 85, 509, 4060, 41301, 510489\}$, or asymptotically $(6n)!/((3n)!(2n)!288^n e^2)$. (For symmetric graphs see page 1032.) If one requires the networks to be planar the numbers are $\{0, 1, 1, 3, 9, 32, 133, 681, 3893, 24809, 169206\}$. If one looks at subnetworks with dangling connections, the number of these up to size 10 is $\{2, 5, 7, 22, 43, 141, 373, 1270, 4053, 14671\}$, or $\{1, 1, 2, 6, 10, 29, 64, 194, 531, 1733\}$ if no self or multiple connections are allowed (see also page 1039).

▪ **Properties of networks.** Over the past century or so a variety of global properties of networks have been studied. Typical ones include:

- Edge connectivity: the minimum number of connections that must be removed to make the network disconnected.
- Diameter: the maximum distance between any two nodes in the network. The pictures below show the largest planar trivalent networks with diameters 1, 2 and 3, and the largest known ones with diameters 4, 5 and 6.



- Circumference: the length of the longest cycle in the network. Although difficult to determine in particular cases, many networks allow so-called Hamiltonian cycles that include every node. (Up to 8 nodes, all 8 trivalent networks have this property; up to 10 nodes 25 of 27 do.)
- Girth: the length of the shortest cycle in the network. The pictures below show the smallest trivalent networks with girths 3 through 8 (so-called cages). Girth can be relevant in seeing whether a particular cluster can ever occur in network.



- Chromatic number: the minimum of colors that can be assigned to nodes so that no adjacent nodes end up the same color. It follows from the Four-Color Theorem that the maximum for planar networks is 4. It turns out that for all trivalent networks the maximum is also 4, and is almost always 3.

▪ **Regular polytopes.** In 3D, of the five regular polyhedra, only the tetrahedron, cube and dodecahedron have three edges meeting at each vertex, corresponding to a trivalent network. (Of the 13 additional Archimedean solids, 7 yield trivalent networks.) In 4D the six regular polytopes have 4, 4, 6, 8, 4 and 12 edges meeting at each vertex, and in higher dimensions the simplex ($d+1$ vertices) and hypercube (2^d vertices) have d edges meeting at each vertex, while the co-cube ($2d$ vertices) has $2(d-1)$. (See also symmetric graphs on page 1032, and page 929.)

▪ **Page 476 · Generalizations.** Almost any kind of generalized network can be emulated by a trivalent network just by introducing more nodes. As indicated in the main text, networks with more than three connections at each node can be emulated by combining nodes into groups, and looking only at the connections between groups. Networks with colored nodes can be emulated by representing each color of node by a fixed group of nodes. Going beyond ordinary networks, one can consider hypernetworks in which connections join not just pairs of nodes, but larger numbers of nodes. Such hypernetworks are specified by adjacency tensors rather than adjacency matrices. But it is possible to emulate any hypernetwork by having each generalized connection correspond to a group of connections in an ordinary trivalent network.

▪ **Maintaining simple rules.** An important reason for considering models based solely on trivalent networks is that they allow simpler evolution rules to be maintained (see page 508). If nodes can have more than three connections, then they will often be able to evolve to have any number of connections—in which case one must give what is in effect an infinite set of rules to specify what to do for each number of connections.

■ **Page 477 · 3D network.** The 3D network (c) can be laid out in space using *Array[x[8{##}] &, {n, n, n}]* where

```
x[m:{_, _, _}] := {x1[m], x1[m + 4],
    x2[m + {4, 2, 0}], x2[m + {0, 6, 4}]}
x1[m:{_, _, _}] := Line[Map[# + m &, {{1, 0, 0}, {1, 1, 1},
    {0, 2, 1}, {1, 1, 1}, {3, 1, 3}, {3, 0, 4}, {3, 1, 3}, {4, 2, 3}}]]
x2[{i_, j_, k_}] :=
  x1[{-i - 4, -j - 2, k}] /. {a_, b_, c_} → {-a, -b, c}
```

The resulting structure is a cubic array of blocks with each block containing 8 nodes. The shortest cycle that returns to a particular node turns out to involve 10 edges. The structure does not correspond to the way that chemical bonds are arranged in any common crystalline materials, probably because it would be likely to be mechanically unstable.

■ **Continuum limits.** For all everyday purposes a region in a network with enough nodes and an appropriate pattern of connections can act just like ordinary continuous space. But at a formal mathematical level this can happen rigorously only in an infinite limit. And in general, there is no reason to expect that all properties of the system (notably for example the existence of particles) will be preserved by taking such a limit. But in understanding the structure of space and comparing to ordinary continuous space it is convenient to imagine taking such a limit. Inevitably there are several scales involved, and one can only expect continuum behavior if one looks at scales intermediate between individual connections in the underlying network and the overall size of the whole network. Yet as I will discuss on pages 534 and 1050 even at such scales it is far from straightforward to see how all the various well-studied properties of ordinary continuous space (as embodied for example in the theory of manifolds) can emerge from discrete underlying networks.

■ **Page 478 · Definitions of distance.** Any measure of distance—whether in ordinary continuous space or elsewhere—takes a pair of points and yields a number. Several properties are normally assumed. First, that if the points are identical the distance is zero, and if they are different, it is a positive number. Second, that the distance between points A and B is the same as between B and A. And third, that the so-called triangle inequality holds, so that the distance AC is no greater than the sum of the distances AB and BC. With distance on a network defined as the length of shortest path between nodes one immediately gets all three of these properties. And even though all distances defined this way will be integers, they still make any network formally correspond in mathematical terms to a metric space (or strictly a path metric space). If the connections on the underlying network are one-way (as in causal networks) then one no longer necessarily gets the second property, and when a continuum limit exists it can correspond to a (perhaps discontinuous) section through a fiber bundle rather than to a manifold. Note that as discussed on page 536 physical measures of distance will always end up being based not just on single paths in a network, but on the propagation of something like a particle, which typically in effect requires the presence of many paths. (See page 1048.)

■ **Page 478 · Definitions of dimension.** The most obvious way to define the dimension of a space is somehow to ask how many parameters—or coordinates—are needed to specify a point in it. But starting in the 1870s the discovery of constructs like space-filling curves (see page 1127) led to investigation of other definitions. And indeed there is some reason to believe that around 1884 Georg Cantor may have tried developing a definition based on essentially the idea that I use here of looking at growth rates of volumes of spheres (balls). But for standard continuous spaces this definition is hard to make robust—since unlike in discrete networks where one can define volume just by counting nodes, defining volume in a continuous space requires assigning a potentially arbitrary density function. And as a result, in the late 1800s and early 1900s other definitions of dimension were developed. What emerged as most popular is topological dimension, in which one fills space with overlapping balls, and asks what the minimum number that ever have to overlap at any point will be. Also considered was so-called Hausdorff dimension, which became popular in connection with fractals in the 1980s (see page 933), and which can have non-integer values. But for discrete networks the standard definitions for both topological and Hausdorff dimension give the trivial result 0. One can get more meaningful results by thinking about continuum limits, but the definition of dimension that I give in the main text seems much more straightforward. Even here, there are however some subtleties. For example, to find a definite volume growth rate one does still need to take some kind of limit—and one needs to avoid sampling too many or too few nodes in the network. And just as with fractal dimensions discussed on page 933 there are issues about whether a definite power law for the growth rate will emerge, and how one should average over results for different parts of the network. There are some alternative approaches to defining dimension in which some of these issues at least become less explicit. For example, one can imagine not just forming a ball on the network, but instead growing something like a cellular automaton, and seeing how big a pattern it produces after some number of steps. And similarly, one can for example look at the statistics of random walks on the network. A slightly different but still related approach is to study the

density of eigenvalues of the Laplace operator—which can also be thought of as measuring the number of solutions to equations giving linear constraints on numbers assigned to connected nodes. More sophisticated versions of this involve looking at invariants studied in topological field theory. And there are potentially also definitions based for example on considering geodesics and seeing how many linearly independent directions can be defined with them. (Note that given explicit coordinates, one can check whether one is in d or more dimensions by asking for all possible points

```
Det[Table[(x[i] - x[j]) . (x[i] - x[j]), {i, d + 3}, {j, d + 3}]] == 0
```

and this should also work for sufficiently separated points on networks. Still another related approach is to consider coloring the edges of a network: if there are $d + 1$ possible colors, all of which appear at every node, then it follows that d coordinates can consistently be assigned to each node.)

■ **Page 478 · Counting of nodes.** The number of nodes reached by going out to network distance r (with $r > 1$) from any node in the networks on page 477 is (a) $4r - 4$, (b) $3r^2/2 - 3r/2 + 1$, and (c)

```
First[Select[4 r^3/9 + 2 r^2/3 +
    {2, 5/3, 5/3} r - {10/9, 1, -4/9}, IntegerQ]]
```

In any trivalent network, the quantity $f[r]$ obtained by adding up the numbers of nodes reached by going distance r from each node must satisfy $f[0] = n$ and $f[1] = 3n$, where n is the total number of nodes in the network. In addition, the limit of $f[r]$ for large r must be n^2. The values of $f[r]$ for all other r will depend on the pattern of connections in the network.

■ **Page 479 · Cycle lengths.** The lengths of the shortest cycles (girths) of the networks on page 479 are (a) 3, (b) 5, (c) 4, (d) 4, (e) 3, (f) 5, (g) 6, (h) 10, (i) ∞, (j) 3. Note that rules of the kind discussed on page 508 which involve replacing clusters of nodes can only apply when cycles in the cluster match those in the network.

■ **Page 479 · Volumes of spheres.** See page 1050.

■ **Page 480 · Implementation.** Networks are conveniently represented by assigning a number to each node, then having lists of rules which specify what nodes the connection from a particular node go to. The tetrahedron network from page 476 is for example given in this representation by

```
{1 → {2, 3, 4}, 2 → {1, 3, 4}, 3 → {1, 2, 4}, 4 → {1, 2, 3}}
```

The list of nodes reached by following up to n connections from node i are then given by

```
NodeLists[g_, i_, n_] :=
  NestList[Union[Flatten[# /. g]] &, {i}, n]
```

The network distance corresponding to the length of the shortest path between two nodes is given by

```
Distance[g_, {i_, j_}] := Length[NestWhileList[
    Union[Flatten[# /. g]] &, {i}, ! MemberQ[#, j] &]] - 1
```

■ **Finding layouts.** One way to lay out a network g so that network distances in it come as close as possible to ordinary distances in d-dimensional space, is just to search for values of the $x[i, k]$ which minimize a quantity such as

```
With[{n = Length[g]}, Apply[Plus,
  Flatten[(Table[Distance[g, {i, j}], {i, n}, {j, n}]^2 - Table[
      Sum[(x[i, k] - x[j, k])^2, {k, d}], {i, n}, {j, n}])^2]]]
```

using for example *FindMinimum* starting say with $x[1, \_] \to 0$ and all the other $x[\_, \_] \to$ *Random*[]. Rarely is there a unique minimum that can be found, but the approach nevertheless seems to work fairly well whenever a good layout exists in a particular number of dimensions. One can imagine weighting different network distances differently, but usually I have found that equal weightings work best. If one ignores all constraints beyond network distance 1, then one is in effect just trying to build the network out of identical rigid rods. It turns out that this is almost always possible even in 2D (though not in 1D); the only exception is the tetrahedron network. And in fact very few trivalent structures are rigid, in the sense the angles between rods are uniquely determined. (In 3D, for example, this is true only for the tetrahedron.)

■ **Hamming distances.** In the so-called loop switching method of routing messages in communications systems one lays out a network on an m-dimensional Boolean hypercube so that the distance on the hypercube (equal to Hamming distance) agrees with distance in the network. It is known that to achieve this exactly, m must be at the least the number of either positive or negative eigenvalues of the distance matrix for the network, and can need to be as much as $n - 1$, where n is the total number of nodes.

■ **Continuous mathematics.** Even though networks are discrete, it is conceivable that network-based models can also be formulated in terms of continuous mathematics, with a network-like structure emerging for example from the pattern of singularities or topology of continuous surfaces or functions.

The Relationship of Space and Time

■ **History.** The idea of representing time graphically like space has a long history—and was used for example by Nicholas Oresme in the mid-1300s. In the 1700s and 1800s the idea of position and time as just two coordinates was widespread in mathematical physics—and this then led to notions like "travelling in time" in H. G. Wells's 1895 *The Time Machine*. The mathematical framework developed for relativity theory in the early 1900s (see page 1042) treated space and time very

symmetrically, leading popular accounts of the theory to emphasize a kind of fundamental equivalence between them and to try to make this seem inevitable through rather confusing thought experiments on such topics as idealized trains travelling near the speed of light.

In the context of traditional mathematical equations there has never been much reason to consider the possibility that space and time might be fundamentally different. For typically space and time are both just represented by abstract symbolic variables, and the formal process of solving equations as a function of position in space and as a function of time is essentially identical. But as soon as one tries to construct more explicit models of space and time one is immediately led to consider the possibility that they may be quite different.

▪ **Page 482 · Discreteness in time.** In present-day physics, time, like space, is always assumed to be perfectly continuous. But experiments—the most direct of which are based on looking for quantization in the measured decay times of very short-lived particles—have only demonstrated continuity on scales longer than about 10^{-26} seconds, and there is nothing to say that on shorter scales time is not in fact discrete. (The possibility of a discrete quantum of time was briefly discussed in the 1920s when quantum mechanics was first being developed.)

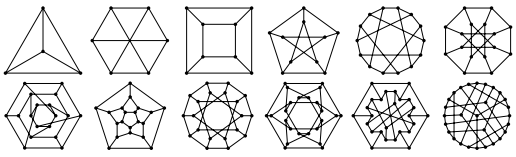
▪ **Page 483 · Network constraint systems.** Cases (a), (f) and (p) allow all networks that do not contain respectively cycles of length 1 (self-loops), cycles of length 3 or less, and cycles of length 5 or less. In cases where an infinite sequence of networks is allowed, there are typically particular subnetworks that can occur any number of times, making the sizes of allowed networks form arithmetic progressions. In cases (m), (n) and (o) respectively triangle, pentagon and square subnetworks can be repeated.

The main text excludes templates that have no dangling connections, and are thus themselves already complete networks. There are 5 such templates involving nodes out to distance one, but of these only 3 correspond to networks that satisfy the constraint that around each node the network has the same form as the template. Among templates involving nodes out to distance two there are 106 that have no dangling connections, and of these only 8 satisfy the constraints.

The main text considers only constraints based on a single template. One can also allow each node to have a neighborhood that corresponds to any of a set of templates. For templates involving nodes out to distance one, there are 13 minimal sets in the sense of page 941, of which only 6 contain just one template, 6 contain two and 1 contains three.

If one does allow dangling connections to be joined within a single template, the results are similar to those discussed so far. There are 52 possible templates involving nodes out to distance two, of which 12 allow complete networks to be formed, none forced to be larger than 12 nodes. There are 46 minimal sets, with the largest containing 4 templates, but none forcing a network larger than 16 nodes.

▪ **Symmetric graphs.** The constraints in a network constraint system require that the structure around each node agrees with a template that contains some number of nodes. A symmetric graph satisfies the same type of constraint, but with the template being the whole network. The pictures below show the smallest few symmetric graphs with 3 connections at each node (with up to 100 nodes there are still only 37 such graphs; compare page 1029).



▪ **Cayley graphs.** As discussed on page 938, the structure of a group can be represented by a Cayley graph where nodes correspond to elements in the group, and connections specify results of multiplying by generators. The transitivity of group multiplication implies that Cayley graphs always have the property of being symmetric (see above). The number of connections at each node is fixed, and given by the number of distinct generators and inverses. In cases such as the tetrahedral group A_4 there are 3 connections at each node. The relations among the generators of a group can be thought of as constraints defining the Cayley graph. As mentioned on page 938, there are finite groups that have simple relations but at least very large Cayley graphs. For infinite groups, it is known (see page 938) that in most cases Cayley graphs are locally like trees, and so do not have finite dimension. It appears that only when the group is nilpotent (so that certain combinations of elements commute much as they do on a lattice) is there polynomial growth in the Cayley graph and thus finite dimension.

▪ **Page 485 · Spacetime symmetric rules.** With $k = 2$ and the neighborhoods shown here, only the additive rules 90R, 105R, 150R and 165R are space-time symmetric. For larger k and larger neighborhoods, there presumably begin to be non-additive rules with this property.

Time and Causal Networks

▪ **Causal networks.** The idea of using networks to represent interdependencies of events seems to have developed with the systematization of manufacturing in the early 1900s—

notably in the work of Frank and Lillian Gilbreth—and has been popular since at least the 1940s. Early applications included switching circuits, logistics planning, decision analysis and general flowcharting. In the last few decades causal networks have been widely used in system specification methods such as Petri nets, as well as in schemes for medical and other diagnosis. Since at least the 1960s, causal networks have also been discussed as representations of connections between events in spacetime, particularly in quantum mechanics (see page 1027).

Causal networks like mine that are ultimately associated with some evolution or flow of activity always have certain properties. In particular, they can never contain loops, and thus correspond to directed acyclic graphs. And from this it follows for example that even the most circuitous path between two nodes must be of finite length.

Causal networks can also be viewed as Hasse diagrams of partially ordered sets, as discussed on page 1040.

■ **Implementation.** Given a list of successive positions of the active cell, as from *Map[Last, MAEvolveList[rule, init, t]]* (see page 887), the network can be generated using

```
MAToNet[list_] := Module[{u, j, k}, u[_] = ∞; Reverse[
    Table[j = list[[i]]; k = {u[j - 1], u[j], u[j + 1]}; u[j - 1] =
      u[j] = u[j + 1] = i; i → k, {i, Length[list], 1, -1}]]]
```

where nodes not yet found by explicit evolution are indicated by ∞.

■ **Page 488 · Mobile automata.** The special structure of mobile automata of the type used here leads to several special features in the causal networks derived from them. One of these is that every node always has exactly 3 incoming and 3 outgoing connections. Another feature is that there is always a path of doubled connections (associated with the active cell) that visits every node in some order. And in addition, the final network must always be planar—as it is whenever it is derived from the evolution of a local underlying 1D system.

■ **Computational compression.** In the model for time described here, it is noteworthy that in a sense an arbitrary amount of underlying computation can take place between successive moments in perceived time.

■ **Page 496 · 2D mobile automata.** As in 2D random walks, active cells in 2D mobile automata often do not return to positions they have visited before, with the result that no causal connections end up being created.

The Sequencing of Events in the Universe

■ **Implementation.** Sequential substitution systems in which only one replacement is ever done at each step can just be implemented using */.* as described on page 893. Substitution systems in which all replacements are done that are found to fit in a left-to-right scan can be implemented as follows

```
GSSEvolveList[rule_, s_, n_] :=
  NestList[GSSStep[rule, #] &, s, n]
GSSStep[rule_, s_] :=
  g[rule, s, f[StringPosition[s, Map[First, rule]]]]
f[{}] = {}; f[s_] := Fold[If[Last[Last[#1]] ≥ First[#2],
      #1, Append[#1, #2]] &, {First[s]}, Rest[s]]
g[rule_, s_, {}] := s; g[rule_, s_, pos_] := StringReplacePart[
    s, Map[StringTake[s, #] &, pos] /. rule, pos]
```

with rules given as *{"ABA" → "BAAB", "BBBB" → "AA"}*.

■ **Generating causal networks.** If every element generated in the evolution of a generalized substitution system is assigned a unique number, then events can be represented for example by *{4, 5} → {11, 12, 13}*—and from a list of such events a causal network can be built up using

```
With[{u = Map[First, list]}, MapIndexed[Function[
    {e, i}, First[i] → Map[(If[# === {}, ∞, #[[1, 1]]] &)[
      Position[u, #]] &, Last[e]]], list]]
```

■ **The sequential limit.** Even when the order of applying rules does not matter, using the scheme of a sequential substitution system will often give different results. If there is a tree of possible replacements (as in *"A" → "AA"*), then the sequential substitution system in a sense does depth-first recursion in the infinite tree, never returning from the single path it takes. Other schemes are closer to breadth-first recursion.

■ **Page 502 · Rule (b).** The maximum number of steps for which the rule can be applied occurs with initial conditions consisting of a white element followed by n black elements, and in this case the number of steps is $2^n + n$.

■ **String theory.** The sequences of symbols I call strings here have absolutely no direct connection to the continuous deformable 1D objects known as strings in string theory.

■ **String overlaps.** The total numbers of strings with length n and k colors that cannot overlap themselves are given by

```
a[0] = 1; a[n_] := k a[n - 1] - If[EvenQ[n], a[n/2], 0]
```

Up to reversal and interchange of A and B, the first few overlap-free strings with 2 colors are A, AB, AAB, $AAAB$, $AABB$.

The shortest pairs of strings of 2 elements with no self- or mutual overlaps are *{"A", "B"}*, *{"AABB", "AABAB"}*, *{"AABB", "ABABB"}*; there are a total of 13 such pairs with strings up to length 5, and 85 with strings up to length 6.

The shortest non-overlapping triple of strings is *{"AAABB", "ABABB", "ABAABB"}* and its variants. There are a total of 36 such triples with no string having length more than 6.

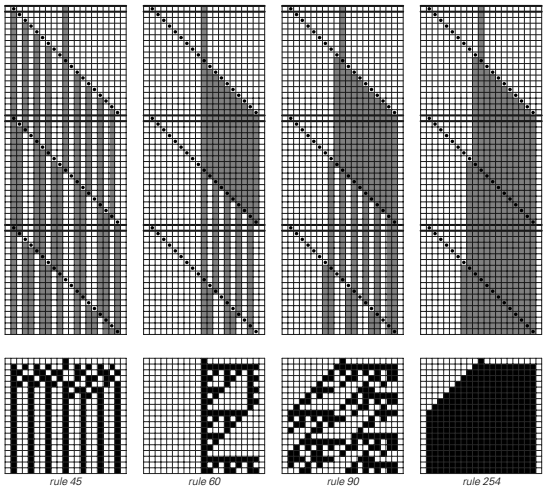
■ **Simulating mobile automata.** Given a mobile automaton like the one from page 73 with rules in the form used on page

887—and behavior of any complexity—the following will yield a causal-invariant substitution system that emulates it:

```
Map[StringJoin, Map[{"AAABB", "ABABB", "ABAABB"}[[
    # + 1]] &, Map[Insert[#[[1]], 2, 2] →
    Insert[#[[2, 1]], 2, 2 + #[[2, 2]]] &, rule], {2}], {2}]
```

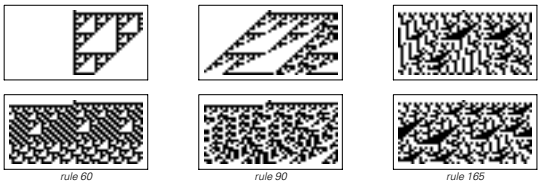
▪ **Sequential cellular automata.** Ordinary cellular automata are set up so that every cell is updated in parallel at each step, based on the colors of neighboring cells on the previous step. But in analogy with generalized substitution systems, one can also consider sequential cellular automata, in which cells are updated sequentially rather than in parallel. The behavior of such systems is usually very different from that of corresponding ordinary cellular automata, mainly because in sequential cellular automata the new color of a particular cell can depend on new rather than old colors of neighboring cells.

The pictures below show the behavior of several sequential cellular automata with $k = 2$, $r = 1$ elementary rules. In the top picture of each pair every individual update is indicated by a black dot. In the bottom picture each line represents one complete step of evolution, including one update of each cell. Note that in this representation, effects can propagate all the way across the system in a single step.

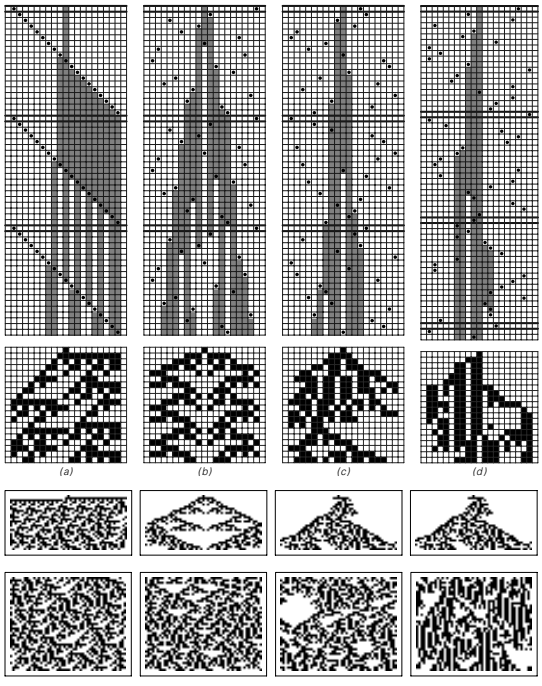


Size dependence. Because effects can propagate all the way across the system in a single step, the overall size, as well as boundary conditions, for the system can be significant after just a few steps, as illustrated in the pictures of rule 60 below.

Additive rules. Among elementary sequential cellular automata, those with additive rules turn out to yield some of the most complex behavior, as illustrated below. The top row shows evolution with the boundary forced to be white; the bottom row shows cyclic boundary conditions. Even though the basic rule is additive, there seems to be no simple traditional mathematical description of the results.



Updating orders. Somewhat different results are typically obtained if one allows different updating orders. For each complete update of a rule 90 sequential cellular automaton, the pictures below show results with (a) left-to-right scan, (b) random ordering of all cells, the same for each pass through the whole system, (c) random ordering of all cells, different for different passes, (d) completely random ordering, in which a particular cell can be updated twice before other cells have even been updated once.



History. Sequential cellular automata have a similar relationship to ordinary cellular automata as implicit updating schemes in finite difference methods have to explicit ones, or as infinite impulse response digital filters have to finite ones. There were several studies of sequential or asynchronous cellular automata done following my work on ordinary cellular automata in the early 1980s.

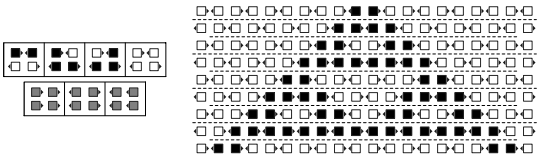
Implementation. The following will update triples of cells in the specified order by using the function *f*:

```
OrderedUpdate[f_, a_, order_] := Fold[ReplacePart[
    #1, f[Take[#1, {#2 - 1, #2 + 1}]], #2] &, a, order]
```

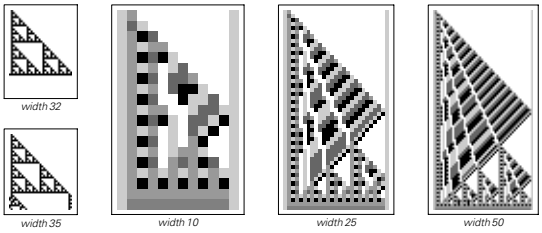
A random ordering of *n* cells corresponds to a random permutation of the form

```
Fold[Insert[#1, #2, Random[Integer, Length[#1]] + 1] &,
  {}, Range[n]]
```

▪ **Intrinsic synchronization in cellular automata.** Taking the rules for an ordinary cellular automaton and applying them sequentially will normally yield very different results. But it turns out that there are variants on cellular automata in which the rules can be applied in any order and the overall behavior obtained—or at least the causal network—is always the same. The picture below shows how this works for a simple block cellular automaton. The basic idea is that to each cell is added an arrow, and any pair of cells is updated only when their arrows point at each other. This in a sense forces cells to wait to be updated until the data they need is ready. Note that the rules can be thought of as replacements such as *"A><B" → "<AB>"* for blocks of length 4 with 4 colors.



▪ **"Firing squad" synchronization.** By choosing appropriate rules it is possible to achieve many forms of synchronization directly within cellular automata. One version posed as a problem by John Myhill in 1957 consists in setting up a rule in which all cells in a region go into a special state after exactly the same number of steps. The problem was first solved in the early 1960s; the solution using 6 colors and a minimal number of steps shown on the right below was found in 1988 by Jacques Mazoyer, who also determined that no similar 4-color solutions exist. Note that this solution in effect constructs a nested pattern of any width (it does this by optionally including or excluding one additional cell at each nesting level, using a mechanism related to the decimation systems of page 909). If one drops the requirement of cells going into a special state, then even the 2-color elementary rule 60 shown on the left can be viewed as solving the problem—but only for widths that are powers of 2.



▪ **Distributed computing.** Many of the basic issues about the progress of time in a universe consisting of many separate elements have analogs in the progress of computations that are distributed across many separate computing elements. In practice, such computations are most often done by requiring explicit synchronization of all elements at appropriate points, and implementing this using a mechanism that is outside of the computation. But more theoretical investigations of formal concurrent systems, temporal logics, dataflow systems, Petri nets and so on have led to ideas about distributed computing that are somewhat closer to the ones I discuss here for the universe. And, as it happens, in the mid-1980s I tried hard, though at the time without much success, to use updating rules for networks as the basis for a new kind of programming language intended for massively parallel computers.

Uniqueness and Branching in Time

▪ **Page 506 · String transformations.** An example of a rule that allows one to go from any string of *A*'s and *B*'s to any other is

```
{"A" → "AA", " AA" → "A", " A" → "B", " B" → "A"}
```

(Compare page 1038.)

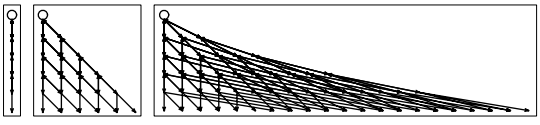
▪ **Parallel universes.** The idea of parallel universes which somehow interact with each other has been much explored in science fiction. And one might think that if the history of each universe corresponds to one path in a multiway system then the convergence of paths might represent interactions between universes. But in fact, much as in the case of time travel, such connections do not represent additional observable effects; they simply imply consistency conditions, in this case between universes whose paths converge.

▪ **Many-worlds models.** The notion of "many-figured time" has been discussed since the 1950s in the context of the many-worlds interpretation of quantum mechanics. There are some similarities to the multiway systems that I consider here. But an important difference is that while in the many-worlds

approach, branchings are associated with possible observation or measurement events, what I suggest here is that they could be an intrinsic feature of even the very lowest-level rules for the universe. (See also page 1063.)

▪ **Spacetime networks from multiway systems.** The main text considers models in which the steps of evolution in a multiway system yield a succession of events in time. An alternative kind of model, somewhat analogous to the ones based on constraints on page 483, is to take the pattern of evolution of a multiway system to define directly a complete spacetime network. Instead of looking separately at strings produced at each step, one instead maintains just a single copy of each distinct string ever produced, and makes that correspond to a node in the network. Each node is then connected to the nodes associated with the strings reached by one application of the multiway rule, as on page 209.

It is fairly straightforward to generate in this way networks of any dimension. For example, starting with n A's the rule $\{"A" \to "AB", "AB" \to "A"\}$ yields a regular n-dimensional grid, as shown below.



If each node in a network is associated with a point in spacetime, then one slightly peculiar feature is that every such point would have an associated string—something like an encoded position coordinate. And it then becomes somewhat difficult to understand why different regions of spacetime seem to behave so similarly—and do not, for example, seem to depend on the details of their coordinates.

▪ **Page 507 · Commuting operations.** If replacements on strings are viewed as mathematical operations, then when the replacements give the same result if applied in any order, the corresponding operations commute.

▪ **Conditions for convergence.** One way to guarantee that there is convergence after one step is to require as in the previous section that blocks to be replaced cannot overlap with themselves or each other. And of the 196 possible rules involving two colors and blocks of length at most three, 112 have this property. But there are also an additional 20 rules which allow some overlap but which nevertheless yield convergence after one step. Examples are $"AAA" \to "A"$ and $"AA" \to "ABA"$. In these rules some of the elements essentially just supply context, but are not affected by the replacement. These elements can then overlap while not affecting the result. Note that unless one excludes the context elements from events, paths in the multiway system will converge, but the causal networks on these paths will be locally slightly different.

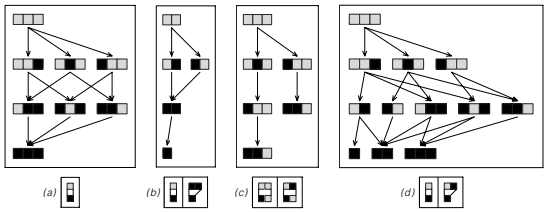
Much as in the previous section, even if paths do not converge for every possible string, it can still be true that paths converge for all strings that are actually generated from a particular initial string.

In general, one can consider convergence after any number of steps, requiring that any two strings which have a common ancestor must at some point also have a common successor. Note that a rule such as $\{"A" \to "B", "A" \to "C", "B" \to "A", "B" \to "D"\}$ exhibits convergence for all paths that have diverged for only one step, but not for all those that have diverged for longer. In general it is formally undecidable whether a particular multiway system will eventually exhibit convergence of all paths.

▪ **Confluence.** As mentioned on page 938, multiway systems have been studied in mathematical logic, typically under names such as rewrite systems, since the early 1900s. The property of path convergence discussed in the main text has been considered since the 1930s, usually under the name of confluence, or sometimes the Church-Rosser property. (Also considered is strong confluence—that paths can always converge in at most one step, and local confluence—that paths can converge after diverging for one step but not necessarily more. Early in its history confluence was most often studied for symbolic systems and lambda calculus rather than ordinary multiway systems.)

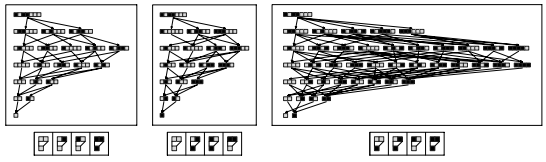
Confluence is important in defining a notion of equivalence for strings. One can say that two strings are equivalent if they can both be transformed to the same string by using the rules of the multiway system. And with such a definition, confluence is what is needed to obtain transitivity for equality, so that $p == q$ and $q == r$ implies $p == r$.

Most often confluence is studied in the context of terminating multiway systems—multiway systems in which eventually strings are produced to which no further replacements apply. If a terminating multiway system has the confluence property, then this implies that regardless of the path taken, a given string will always evolve to a unique string that can be thought of as giving a canonical or normal form for the original string. Examples (a) through (c) below have this property; (d) does not. In example (a), the canonical form is all elements black; in (b) it is a single black element, and in (c) all elements are black, except the last one, which is white if there were any initial white elements. Note that the first example on page 507 has a canonical form consisting of a sorted string.



The process of evaluation in mathematics or in a computer language such as *Mathematica* can be thought of as involving the application of a sequence of replacement rules. Only if these rules have the confluence property will the results always be unique, and independent of the order of rule application.

The evaluation of functions with attribute *Flat* in *Mathematica* provides an example of confluence. If *f* is *Flat*, then in evaluating *f[a, b, c]* one can equally well start with *f[f[a, b], c]* or *f[a, f[b, c]]*. Showing only the arguments to *f*, the pictures below illustrate how the flat functions *Xor* and *And* are confluent, while the non-flat function *Implies* is not.



■ **Completion.** If one has a multiway system that terminates but is not confluent then it turns out often to be possible to make it confluent by adding a finite set of new rules. Given a string p which gets transformed either to q or r by the original rules, one can always imagine adding a new rule $q \to r$ or $r \to q$ that makes the paths from p immediately converge. To do this explicitly for all possible p that can occur would however entail having infinitely many new rules. But as noted by Donald Knuth and Peter Bendix in 1970 it turns out often to be sufficient just iteratively to add new rules only for each so-called critical pair q, r that is obtained from strings p that represent minimal overlaps in the left-hand sides of the rules one has. To decide whether to add $q \to r$ or $r \to q$ in each case one can have some kind of ordering on strings. For the procedure to work this ordering must be such that the strings generated on successive steps in every possible evolution of the multiway system follow the ordering. A number of variations of the basic procedure—using different orderings and with different schemes for dropping redundant rules—have been proposed for systems arising in different kinds of applications. The original Knuth-Bendix procedure was for equations (of the form $a \leftrightarrow b$) had the feature that it could terminate yet not give a confluent multiway system. But in the 1980s so-called unfailing completion algorithms (see page 1158) were developed that—if they terminate—guarantee to give confluent systems. (The question of whether any procedure of this type will terminate in a particular case is nevertheless in general undecidable.)

The basic idea of so-called critical pair completion procedures has arisen several times—notably in the Gröbner basis approach of Bruno Buchberger from 1965 to finding canonical forms for systems of polynomials.

■ **Relationships between types of networks.** Each arrow on each path in a multiway system corresponds to a node in a causal network. Each element in each string in a multiway system corresponds to a connection in a causal network. Each complete string in a multiway system corresponds to a possible slice that goes through all connections across a causal network. Such a slice can be considered in traditional physics terms as a spacelike hypersurface (see page 1041).

Evolution of Networks

■ **Page 509 · Neighbor-independent rules.** Even though the same replacement is performed at each node at each step, the networks produced are not homogeneous. In the first case shown, the picture produced after t steps has $4 \times 3^{t-k-1}$ regions with 3×2^k edges. In the limit $t \to \infty$, the picture has the geometrical form of an Apollonian circle packing (see page 986). The number of nodes at distance up to r from a given node is at most *1 + Sum[c[i] + c[i - 1], {i, n}]* where *c[i_] := 2^DigitCount[i, 2]*. In practice this number fluctuates greatly with r, making pictures like those on page 479 not exhibit smooth profiles. Averaged over all nodes, however, the number of nodes at distance up to r approximates *r^Log[2, 3]*, implying an effective dimension of *Log[2, 3]*. Note that there is no upper limit on the dimension that can be obtained with appropriate neighbor-independent rules.

■ **Implementation.** For many practical purposes the best representation for networks is the one given on page 1031. But in updating networks a particularly straightforward implementation of one scheme can be obtained if one uses instead a more explicit symbolic representation such as

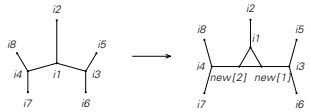
u[1 → v[2, 3, 4], 2 → v[1, 3, 4], 3 → v[1, 2, 4], 4 → v[1, 2, 3]]

This allows one to capture the basic character of networks by

Attributes[u] = {Flat, Orderless}; Attributes[v] = Orderless

Updating rules can then be written in terms of ordinary *Mathematica* patterns. A slight complication is that the patterns have to include all nodes whose connections go to

nodes whose labels are changed by the update. The rule at the top of page 509 must therefore be written out as



and this corresponds to the *Mathematica* rule

```
u[i1_ → v[i2_, i3_, i4_], i3_ → v[i1_, i5_, i6_],
    i4_ → v[i1_, i7_, i8_]] :> u[i1 → v[i2, new[1], new[2]],
    new[1] → v[i1, new[2], i3], new[2] → v[i1, new[1], i4],
    i3 → v[new[1], i5, i6], i4 → v[new[2], i7, i8]]
```

(Strictly there also need to be additional rules to cover where for example nodes 3 and 4 are actually the same.) With rules in this form the network update is simply

```
NetStep[rule_, net_] := Block[{new},
    net /. rule /. new[n_] → n + Apply[Max, Map[First, net]]]
```

Note that just as we discussed for strings on page 1033 the direct use of `/.` here corresponds to a particular scheme for applying the update rule.

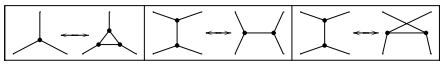
■ **Identifying subnetworks.** The problem of finding where in a network a given subnetwork can occur turns out in general to be computationally difficult. For strings the analogous problem is straightforward, since in a string of length n one can ultimately just try each of the n possible starting points for the substring and see for which of them a match occurs. But for a network with n nodes, a similar procedure would require one to check n^k possible configurations in order to find out where a subnetwork of size k occurs. In practice, however, for fixed subnetworks, one can devise fairly efficient procedures. But the general problem of so-called subgraph isomorphism is formally NP-complete.

■ **Page 509 · Number of replacements.** The total number of distinct replacements that maintain planarity, involve clusters with up to five nodes and have from 3 to 7 dangling connections is $\{16, 8, 125, 24, 246\}$. Not maintaining planarity, the numbers are $\{14, 5, 13, 2, 2\}$. (See page 1039.)

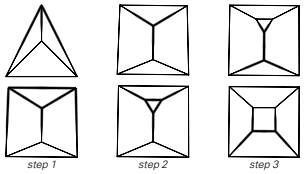
■ **Cycles in networks.** See page 1031.

■ **Planar networks.** One feature of a planar network is that it is always possible to identify definite regions or faces bounded by connections in the network. And from Euler's formula $f + n = e + 2$, it then follows that the average number of edges of each face is always $6(1 - 2/f)$, where f is the total number of faces. Note that with my definition of dimension for networks, the fact that a network is planar does not necessarily mean that it has be two-dimensional—and for example the networks on page 509 are not.

■ **Arbitrary transformations.** By applying the string transformation rules on page 1035 at appropriate locations, it is possible to transform any string of A's and B's to any other. And the analog of this for networks is that by applying the rules shown below at appropriate locations it is possible to transform any network into any other. These rules correspond to the moves invented by James Alexander in 1923 in connection with transforming one knot into another. (Note that the first two rules suffice for all planar networks, and are sometimes called respectively T2 and T1.)



As an example, the pictures below show how a tetrahedron network can be transformed into a cube.



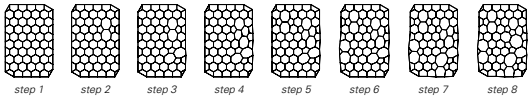
■ **Random networks.** One way to generate the connections for a "completely random" trivalent network with n nodes is just to apply a random permutation:

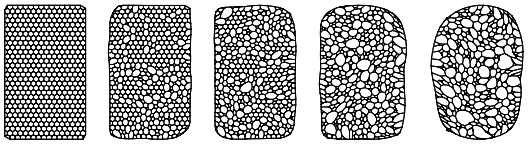
```
RandomNetwork[n_?EvenQ] := Partition[
    Fold[Insert[#1, #2, Random[Integer, Length[#1]] + 1] &,
      {}, Floor[Range[1, n + 2/3, 1/3]]], 2]
```

Networks obtained in this way are usually connected, but will almost always contain self-loops and multiple edges. Properties of random networks are discussed on page 963. A convenient way to get somewhat random planar networks is from 2D Voronoi diagrams of the kind discussed on page 987.

■ **Random replacements.** As indicated in the note above, applying the second rule (T1, shown as (b) on page 511) at an appropriate sequence of positions can transform one planar network into any other with the same number of nodes. The pictures below show what happens if this rule is repeatedly applied at random positions in a network. Each time it is applied, the rule adds two edges to one face, and removes them from another. After many steps the pictures below show that faces with large numbers of edges appear. The average number of edges must always be 6 (see note above), but in a sufficiently large network the probability for a face to have n edges eventually approaches an equilibrium value of $8(n-2)(2n-3)!!(3/8)^n/n!$. (For large n this is approximately λ^n with $\lambda = 3/4$; if 1- and 2-edged regions are allowed then $\lambda = (3 + \sqrt{3})/6 \simeq 0.79$.) There may be some easy way to derive such results, but so far it has only been done using fairly sophisticated techniques from quantum field theory developed in the late 1970s. The starting point is to look at a

ϕ^3 field theory with SU(n) internal symmetry and to note that in the limit $n \to \infty$ what dominates are Feynman diagrams that have the structure of planar trivalent networks (see page 1040). And it then turns out that in zero spacetime dimensions the complete path integral for the theory can be evaluated exactly—yielding in effect a generating function for the number of possible networks. Parametric differentiation (to yield n-point correlation functions) then gives results for n-sided regions. Another result that has been derived is that the average total number $m[n]$ of edges of all faces around a given face with n edges is $7n + 3 + 9/(n + 1)$. Note that the networks obtained always have dimension 2 according to my definitions.





▪ **Cellular structures.** There are many systems in nature that consist of assemblies of discrete regions—and the lines that define the interfaces between these regions form networks. In many cases the regions are fixed once established (compare page 988). But in other cases there is continuing evolution, as for example in soap and other foams and froths, grains in metals and perhaps some biological tissues. In 2D situations the lines between regions generically form a trivalent planar network. In a soap foam, the geometrical layout of this network is determined by surface tension forces—with connections meeting at 120° at each node, though being slightly curved and of different lengths. Pressure differences lead to diffusion of gas and on average to von Neumann's Law that the area of an n-sided region changes linearly with time, at a rate proportional to $n - 6$. Typically the network topology of a foam continually rearranges itself through cascades of seemingly random T1 processes (rule (b) from page 511), with regions that reach zero size disappearing through T2 processes (reversed rule (a)). And as noted for example by Cyril Smith in the early 1950s there is a characteristic coarsening that occurs. Something similar is already visible in the pure T1 pictures in the note above. But results such as the so-called Aboav-Weaire law that $m[n]$ from the note above is in practice about $5n + c$ suggest that T2 processes are also important. (Processes like cell division in 2D biological tissue in effect directly add connections to a network. But this can again be thought of as a combination of T1 and T2 processes, and in appropriate idealizations can lead to very similar results.)

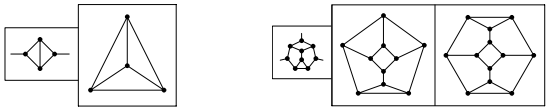
▪ **Page 514 · Cluster numbers.** The following tables give the total numbers of distinct clusters—with number of nodes going across the page, and number of dangling connections going down. (See also page 1038.)

	1	2	3	4	5	6	7	8	9	10
0	0	0	0	1	0	2	0	5	0	19
1	0	0	0	0	1	0	4	0	19	0
2	0	0	0	1	0	5	0	23	0	132
3	1	0	1	0	3	0	15	0	91	0
4	0	1	0	2	0	9	0	54	0	390
5	0	0	1	0	4	0	22	0	166	0
6	0	0	0	2	0	9	0	63	0	551

	1	2	3	4	5	6	7	8	9	10
7	0	0	0	0	2	0	17	0	157	0
8	0	0	0	0	0	4	0	38	0	424
9	0	0	0	0	0	0	6	0	80	0
10	0	0	0	0	0	0	0	11	0	180
11	0	0	0	0	0	0	0	0	18	0
12	0	0	0	0	0	0	0	0	0	37

▪ **Page 515 · Non-overlapping clusters.** The picture shows all distinct clusters with 3 dangling connections and 9 nodes that are not self-overlapping. The only smaller cluster with the same property is the trivial one with just a single node.

Most clusters that can overlap will be able to do so in an infinite number of possible networks. (One can see this by noting that they can overlap inside clusters with dangling connections, not just closed networks.) But there are some clusters that can overlap only in a few small networks. The pictures below show examples where this happens. The pictures in the main text still treat such clusters as non-overlapping.



If two clusters overlap, then this means that there is some network in which there are copies of these clusters that involve some of the same nodes. And it is possible to search for such a network by starting from a single node and then sequentially trying to take corresponding pieces from the two clusters.

▪ **1- and 2-connection clusters.** Clusters with just one or two dangling connections can always in effect be thought of just as adding extra structure to single connections in a network. But this extra structure can be important in the application of other rules—and can for example emulate something like having multiple colors of connections.

▪ **Connectedness.** It is not clear whether a network that represents the universe must remain globally connected, or whether pieces can break off. But any replacements that take connected clusters and yield connected clusters must always maintain the connectedness of any network.

■ **Reversibility.** By including both forward and backward versions of every transformation it is straightforward to set up reversible rules for network evolution. It is not clear, however, whether the basic rules for the universe are really reversible. It could well be that the apparent reversibility we see arises because the universe is effectively on an attractor, as discussed on page 1018. Note that if pieces of the universe can break off, but cannot reconnect, then there will inevitably be an irreversible loss of information.

■ **1/*n* expansion.** If there are n possible colors for each connection in a network, then for large n it turns out that the vast majority of networks will be planar. This idea was used in the 1980s as a way of simplifying the Feynman diagrams to consider in QCD and other quantum field theories. (See page 1039.)

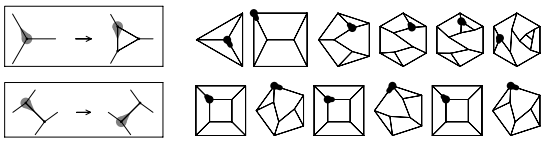
■ **Feynman diagrams.** In the standard approach to particle physics, possible interaction processes are represented by networks in which each node corresponds to an elementary interaction, and the nodes are joined by connections which correspond to the propagation of particles in spacetime. I can see no direct physical relationship between such diagrams and the networks I consider. However, at a mathematical level, the set of trivalent networks with n nodes formally corresponds to the set of n^{th} order Feynman diagrams in a ϕ^3 field theory. (Compare page 1039.)

■ **Chemical analogy.** The evolution of a network can be thought of as an idealized version of a chemical process in which molecules are networks of bonds. (See page 1193.)

■ **Symbolic representations.** Expressions in which common subexpressions are shared correspond to networks, as do collections of relations between objects representing nodes.

■ **Graph grammars.** The notion of generalizing substitutions for strings to the case of networks has been discussed in computer science since the 1960s—and a fair amount of formal work has been done on so-called graph grammars for specifying formal languages whose elements are networks. Even a good analog of regular languages has, however, not yet been found. But applications to constructing or verifying practical network-based system description schemes are quite often discussed. In mathematics rather little is usually done with anything but very trivial network substitutions. In mathematics, rather little is usually done with network substitutions, though the proof of the Four-Color Theorem in 1976 was for example based on showing that 300 or so possible replacement rules—if applied in an appropriate sequence—can transform any graph to have one of 1936 smaller subgraphs that require the same number of colors. (32 rules and 633 subgraphs are now known to be sufficient.)

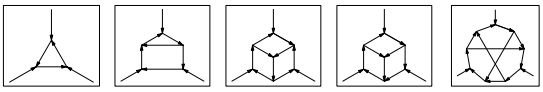
■ **Network mobile automata.** The analog of a mobile automaton can be defined for networks by setting up a single active node, then having rules which replace clusters of nodes around this active node, and move its position. The pictures below show two simple examples.



The total number of replacements that can be used in the rules of a network mobile automaton and which involve clusters with up to four nodes and have from 1 to 4 dangling connections is *{14, 10, 2727, 781}*. Despite looking at several hundred thousand cases I have not been able to find network mobile automata with especially complicated behavior.

Note that by having a cluster of nodes with a unique form it is possible to emulate a network mobile automaton using an ordinary network substitution system.

■ **Directed network systems.** If one adds directionality to the connections in a network it becomes particularly easy to set up rules for clusters of nodes that cannot overlap. For no two clusters whose dangling connections all point inwards can ever overlap, at least so long as neither of these clusters themselves contain subclusters whose dangling connections similarly all point inwards. The pictures below show a few examples of such clusters. Note that in a random network of n nodes, about $n/8$ such clusters typically occur.



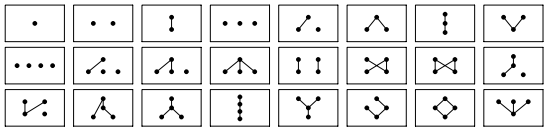
Space, Time and Relativity

■ **Page 516 · Posets.** The way I set things up, collections of events can be thought of as partially ordered sets (posets). If all events occurred in a definite sequence in time, this would define a total linear ordering for them. But with the setup I use, there is only a partial ordering of events, defined by causal connections. The causal networks I draw are so-called Hasse or order diagrams of the posets of events. If a connection goes directly from x to y in this network then x is said to cover y. And in general if there is a path from x to y then one writes $x > y$. The collection of all events that will lead to a given set of events (the union of their past light cones) is known as the filter of that set. Within a poset, there

can be sequences of elements that are totally ordered, and these are called chains. (The maximum length of any chain is sometimes called the dimension of a poset, but this is unrelated to the notions of dimension I consider.) There can also be sets of elements between which no ordering relations at all are defined, and these are called antichains.

Standard examples of posets include subsets of a set ordered by the subset relation, complex numbers ordered by magnitude, and integers ordered by divisibility. Posets first arose as general concepts in the late 1800s in connection with the development of mathematical logic, and to some extent abstract algebra. They became somewhat popular in the mid-1900s, both as formal generalizations in lattice theory, and as structures in various combinatorics applications. It was already noted in the 1920s that events in relativity theory formed posets.

The pictures below show the first few distinct possible Hasse diagrams for posets. For successive numbers of elements the total numbers of these are 1, 2, 5, 16, 63, 318, 2045, 16999, ...



▪ **Page 517 · Spacelike slices.** The definition of spacelike slices used here is directly analogous to what is used in traditional relativity theory (typically under names like spacelike hypersurfaces and Cauchy surfaces). There will normally be many different possible choices of spacelike slices, but in all cases a particular such slice is set up to represent what can consistently be thought of as all of space at a given time. One definition of a spacelike slice is then a maximal set of points in which no pair are causally related (corresponding to a maximal antichain in a poset). Another definition (equivalent for any connected causal network) is that spacelike slices are what consistently divide a causal network into a past and a future. And an intermediate definition is that a spacelike slice contains points that are not themselves causally related, but which appear in either the past or the future of every other point. Given a spacelike slice in a causal network, it is always possible to construct another such slice by finding all those points whose immediate predecessors are all included either in the original slice or its predecessors.

▪ **Page 518 · Speed of light.** In a vacuum the speed of light is 299,792,458 meters/second (and this is actually what is taken to define a meter). In materials light mostly travels slower—basically because there are delays when it is absorbed and reemitted by atoms. In a first approximation, the slowdown factor is the refractive index. But particularly in materials which can amplify light a whole sequence of peculiar effects have been observed—and it is fairly subtle to account correctly for incoming and outgoing signals, and to show that at least no energy or information is transmitted faster than c. The standard mathematical framework of relativity theory implies that any massless particle must propagate at c in a vacuum—so that not only light but also gravitational waves presumably go at this speed (and the same is at least approximately true of neutrinos). The effective mass for massive particles increases by a factor $1/\text{Sqrt}[1 - v^2/c^2]$ at speed v, making it take progressively more energy to increase v. At a formal mathematical level it is possible to imagine tachyons which always travel faster than c. But the structure of modern physics would find it difficult to accommodate interactions between these and ordinary particles.

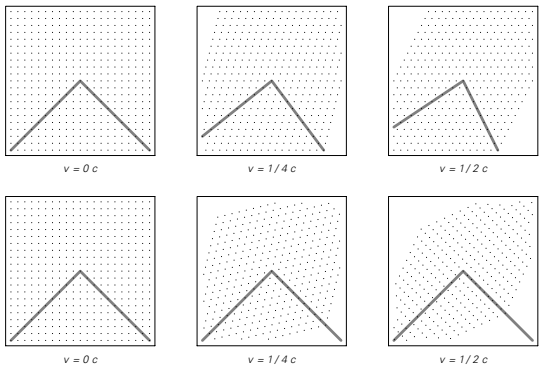
▪ **Page 522 · History of relativity.** (See also page 1028.) The idea that mechanical processes should work the same regardless of how fast one is moving was expressed by Galileo in the early 1600s, particularly in connection with the motion of the Earth—and was incorporated in the laws of mechanics formulated by Isaac Newton in 1687. But when the wave theory of light finally became popular in the mid-1800s it seemed to imply that no similar principle could be true for light. For it was generally assumed that waves of light must correspond to explicit disturbances in a medium or ether that fills space. And it was thus expected that for example the apparent speed of light would depend on how fast one was moving with respect to this ether. And indeed in particular this was what the equations for electromagnetism developed by James Maxwell in the 1860s seemed to suggest. But in 1881 an experiment by Albert Michelson (repeated more accurately in 1887 as the Michelson-Morley experiment and now done to the 10^{-20} level) showed that in fact this was not correct. Already in 1882 George FitzGerald and Hendrik Lorentz noted that if there was a contraction in length by a factor $\text{Sqrt}[1 - v^2/c^2]$ in any object moving at speed v (with c being the speed of light) then this would explain the result. And in 1904 Lorentz pointed out that Maxwell's equations are formally invariant under a so-called Lorentz transformation of space and time coordinates (see note below). Then in 1905 Albert Einstein proposed his so-called special theory of relativity—which took as its basic postulates not only that the laws of mechanics and electrodynamics are independent of how fast one is moving, but that this is also true of the speed of light.

And while at first these postulates might seem incompatible, what Einstein showed was that they are not—at least if modifications are made to the basic laws of mechanics. In the few years that followed, various formulations of this result were given, with Hermann Minkowski in 1908 showing that it could be derived if one just assumes that space and time enter all physical laws together in a certain kind of 4D vector. In the late 1800s Ernst Mach had emphasized the idea of formulating science and particularly mechanics in terms only of concepts that can actually be measured by observers. And in this framework Einstein and others gave what seemed to be almost purely deductive arguments for relativity theory—with the result that it generally came to be assumed that there was no meaningful sense in which one could ever imagine deriving relativity from anything more fundamental. Yet as I discussed earlier in the chapter, if a complete theory of physics is to be as simple as possible, then most things like relativity theory must in effect be derived from more basic features of the theory—as I start to try to do in the main text of this section.

■ **Standard treatment.** In a standard treatment of relativity theory one way to begin is to consider setting up a square grid of points in space and time—and then to ask what kind of transformed grid corresponds to this same set of points if one is moving at some velocity v. At first one might assume that the answer would just be a grid that has been sheared by the simple transformation $\{t, x\} \to \{t, x - v\,t\}$, as in the first row of pictures below. And indeed for purposes of Newtonian mechanics this so-called Galilean transformation is exactly what is needed. But as the pictures below illustrate, it implies that light cones tip as v increases, so that the apparent speed of light changes, and for example Maxwell's equations must change their form. But the key point is that with an appropriate transformation that affects both space and time, the speed of light can be left the same. The necessary transformation is the so-called Lorentz transformation

$$\{t, x\} \to \{t - v\,x/c^2,\ x - v\,t\}/\mathrm{Sqrt}[1 - v^2/c^2]$$

And from this the time dilation factor $1/\mathrm{Sqrt}[1 - v^2/c^2]$ shown on page 524 follows, as well as the length contraction factor $\mathrm{Sqrt}[1 - v^2/c^2]$. An important feature of the Lorentz transformation is that it preserves the quantity $c^2\,t^2 - x^2$—with the result that as v changes in the pictures below a given point in the grid traces out a hyperbola whose asymptotes lie on a light cone. Note that on a light cone $c^2\,t^2 - x^2$ always vanishes. Note also that the intersection of the past and future light cones for two events separated by a distance x in space and t in time always has a volume proportional exactly to $c^2\,t^2 - x^2$.



■ **Inferences from relativity.** The pictures on page 524 show that an idealized clock based on bouncing light between mirrors will exhibit relativistic time dilation. And from such derivations it is often assumed that the same result must hold for any possible clock system. But as a practical matter it does not. And indeed for example the clocks in GPS satellites are specifically set up so as to remove the effects of time dilation. And in the twin paradox one can certainly imagine that each twin could have an accelerometer whose readings they use to correct their clocks. Indeed, even when it comes to individual particles there are subtle effects associated with acceleration and radiation (see page 1062)—so that in the end not entirely clear that something like a biological system would actually in practice exhibit just standard time dilation.

One feature of relativity is that it implies that only relative motion is ultimately ever detectable. (This was also implied by Newtonian mechanics for purely mechanical systems.) And from this it is often concluded that there can be nothing like an ether that one can consider as defining an absolute state of rest in the universe. But in fact the cosmic microwave background in effect does exactly this. For in standard cosmological models it fills the universe, but is everywhere at rest relative to the global center of mass of the universe. And from the anisotropies we have observed in the microwave background it is thus possible to conclude that the Earth is moving at an absolute speed of about $c/10^3$ relative to the center of mass of the universe. In particle physics standard models also in effect introduce things that are assumed to be at rest relative to the center of mass of the universe. One example is the Higgs condensate discussed in connection with particle masses (see page 1047). Other possible examples include zero-point fluctuations in quantum fields.

Outside of science, relativity theory is sometimes given as evidence for various general ideas of cultural relativism (compare page 1131)—which have existed since well before

relativity theory in physics, and seem in the end to have no meaningful connection to it.

■ **Particle physics.** Relativity theory was originally formulated just for mechanics and electromagnetism. But its predictions like $E = m c^2$ were immediately applied for example to radioactivity, and soon it came to be assumed that the theory would work for any system at all—unless it involved gravity. So this has meant that in particle physics $c^2 t^2 - x^2 - y^2 - z^2$ is at some level the only quantity that ever appears. And to make mathematical work easier, what is very often done is to carry out the so-called Wick rotation $t \to i t$—so relativistic invariance is just independence on 4D orientation. (See page 1061.) But except in rather simple cases there is practically no evidence that results obtained after Wick rotation have anything to do with physical reality—and certainly the transformation removes some very basic phenomena such as particle propagation. One feature of it, however, is that it maps the equation for quantum mechanical time evolution into the equation for probabilities in statistical mechanics, with imaginary time corresponding to inverse temperature. And while it is conceivable that this mapping may have some deep significance, none has so far ever been identified.

■ **Time travel.** The idea that space and time are similar suggests that it might be possible to move backwards and forwards in time just like it is possible to move backwards and forwards in space. And indeed in the partial differential equations that define general relativity, it is formally possible for the motion of particles to achieve this, at least when there is sufficient negative energy density from matter or a cosmological constant. But even in this case there is no real progression in which one travels backwards in time. Instead, the possibility of motion that leads to earlier times simply implies a requirement of consistency between behavior at earlier and later times.

Elementary Particles

■ **Note for physicists.** My goal in the remainder of this chapter is not to present a specific ultimate model for physics, but rather to discuss at a fairly general level some features that I believe such a model will have, given the overall discoveries of this book, and the specific results I have described in this chapter. I am certainly aware that many physicists will want to know more details. But particularly in making contact with existing physics it is almost inevitable that all sorts of technical formalism will be needed—and to maintain balance in this book I have not included this here. (Given my own personal background in theoretical physics it will come as no surprise that I have often used such formalism in the process of working out what I describe in these sections.)

■ **Page 525 · Types of particles.** Current particle physics identifies three basic types of known elementary particles: leptons, quarks and gauge bosons. The known leptons are the electron (e), muon (μ) and tau lepton (τ), and their corresponding neutrinos (ν_e, ν_μ, ν_τ). Quarks exist inside hadrons like the proton and pion, but never seem to occur as ordinary free particles. Six types are known: u, d, c (charm), s (strange), t (top), b. Gauge bosons are associated with forces. Those currently known are the photon (γ) for electromagnetism (QED), W and Z for so-called weak interactions, and the gluon (g) for QCD interactions between quarks. Gravitons associated with gravitational forces presumably also exist. In ordinary matter, the only particles that contribute in direct ways to everyday physical, chemical and even nuclear properties are electrons, photons and effectively u and d quarks, and gluons. (These, together presumably with some type of neutrino, are the only types of particles that never seem to decay.) The first reasonably direct observations of the various types of particles were as follows (some were predicted in advance): e (1897), γ (~1905), u, d (1914/~1970), μ (1937), s (1946), ν_e (1956), ν_μ (1962), c (1974), τ, ν_τ (1975), b (1977), g (~1979), W (1983), Z (1983), t (1995).

Most particles exist in several variations. Apart from the photon (and graviton), all have distinct antiparticles. Each quark has 3 possible color configurations; the gluon has 8. Most particles also have multiple spin states. Quarks and leptons have spin 1/2, yielding 2 spin states (neutrinos could have only 1 if they were massless). Gauge bosons normally have spin 1 (the graviton would have spin 2) yielding 3 spin states for massive ones. Real massless ones such as the photon always have just 2. (See page 1046.)

In the Standard Model the idea of spontaneous symmetry breaking (see page 1047) allows particles with different masses to be viewed as manifestations of single particles, and this is effectively done for W, Z, γ, as well as for each of the 3 so-called families of quarks and leptons: u, d; c, s; t, b and e, ν_e; μ, ν_μ; τ, ν_τ. Grand unified models typically do this for all known gauge bosons (except gravitons) and for corresponding families of quarks and leptons—and inevitably imply the existence of various additional particles more massive than those known, but with properties that are somehow intermediate. Some models also unify different families, and supersymmetric models unify quarks and leptons with gauge bosons.

■ **History.** The idea that matter—and light—might be made up of discrete particles was already discussed in antiquity

(see page 876). But it was only in the mid-1800s that there started to be real evidence for the existence of some kind of discrete atoms of matter. Yet at the time, the idea of fields was popular, and it was believed that the universe must be filled with a continuous fluid-like ether responsible at least for light and other electromagnetic phenomena. So for example following ideas of William Rankine from 1849 William Thomson (Kelvin) in 1867 suggested that perhaps atoms might be like knotted stable vortex rings in the ether—with different knots corresponding to different chemical elements. But though it initiated the mathematical classification of knots, and now has certain conceptual similarities to what I discuss in this book, the details of this model did not work out—and it had been largely abandoned even before the electron was discovered in 1897. Ernest Rutherford's work in the 1910s on scattering from atoms introduced the idea of an atomic nucleus, and after the discovery of the neutron in 1932 it became clear that the main constituents of nuclei were protons and neutrons. The positron and the muon were discovered in cosmic rays in the 1930s, followed in the 1940s by a handful of other particles. By the 1960s particle accelerators were finding large numbers of new particles every year. And the hypothesis was then suggested that all these particles might actually be composed of just three more fundamental particles that became known as quarks. An alternative so-called democratic or bootstrap hypothesis was also suggested: that somehow any particle could just be viewed as a composite of all others with the same overall properties—with everything being determined by consistency in the web of interactions between particles, and no particles in a sense being more fundamental than others. But by the early 1970s experiments on so-called deep inelastic scattering had given increasingly direct evidence for point-like constituents inside particles like protons—and by the mid-1970s these were routinely identified with quarks.

As soon as the electron was discovered there were questions about its possible size. For if its charge was distributed over a sphere of radius r, this was expected to lead to electrostatic repulsion energy proportional to $1/r$. And although it was suggested around 1900 that effects associated with this might account for the mass of the electron, this ran into problems with relativity theory, and it also remained mysterious just what might hold the electron together. (A late suggestion made in 1953 by Hendrik Casimir was that it could be forces associated with zero-point fluctuations in quantum fields—but at least with the simplest setup these turned out to have wrong sign.)

The development of quantum theory in the 1920s showed that discrete particles will inevitably exhibit continuous wave-like features in their spatial distribution of probability amplitudes. But traditional quantum mechanics and quantum field theory are both normally formulated with the assumption that the basic particles they describe have zero intrinsic spatial size. Sometimes nonzero size is taken into account by inserting additional interaction parameters—as done in the 1950s with magnetic moments and form factors of protons and neutrons. But for example in quantum electrodynamics the definite assumption is made that electrons are intrinsically of zero size. Quantum fluctuations make any particle in an interacting field theory effectively be surrounded by virtual particles. Yet not unlike in classical electrodynamics having zero intrinsic size for the electron still immediately suggests that an electron should have infinite self-energy. In the 1930s ideas about avoiding this centered around modifying basic laws of electrodynamics or the structure of spacetime (see page 1027). But the development of renormalization in the 1940s showed that these infinities could in effect just be factored out. And by the 1960s a long series of successes in the predictions of QED had led to the almost universal belief that its assumption of point-like electrons must be correct. It was occasionally suggested that the muon might be some kind of composite object. But experiments seemed to indicate that it was in every way identical to the electron, except in mass. And although no reasonable explanation for its existence was found, it came to be generally assumed by the 1970s that it was just another point-like particle. And indeed—apart from few rare suggestions to the contrary—the same is now assumed throughout mainstream practical particle physics for all of the basic particles that appear in the Standard Model. (Actual experiments based on high-energy scattering and precision magnetic moment measurements have shown only that electrons and muons must have sizes smaller than about $\hbar\, c/(10\ \mathrm{TeV}) \simeq 10^{-20}\ m$—or about 10^{-5} times the size of a proton. One can make arguments that composite particles this small should have masses much larger than are observed—but it is easy to find theories that avoid these.)

In the 1980s superstring theory introduced the idea that particles might actually be tiny 1D strings—with different types of particles corresponding essentially just to strings in different modes of vibration. Since the 1960s it has been noted in many simplified quantum field theories that there can be a kind of duality in which a soliton or other extended field configuration in one representation becomes what acts like an elementary particle in another representation. And in the late 1990s there were indications that such phenomena could occur in generalized string theories—leading to suggestions of at least an abstract correspondence between

for example particles like electrons and gravitational configurations like black holes.

■ **Page 526 · Topological defects.** An idealized vortex in a 2D fluid involves velocity vectors that in effect wind around a point—and can never be unwound by making a series of small local perturbations. The result is a certain kind of stability that can be viewed as being of topological origin. One can classify forms of stability like this in terms of the mathematics of homotopy. Most common are point and line defects in vector fields, but more complicated defects can occur, notably in liquid crystals, models of condensates in the early universe, and certain nonlinear field theories. Analogs of homotopy can presumably be devised to represent certain forms of stability in systems like the networks I consider.

■ **Page 527 · Kuratowski's theorem.** Any network can be laid out in 3D space. (This is related to the Whitney embedding theorem that any d-dimensional manifold can be embedded in $(2d+1)$-dimensional space.) When one says that a network is planar what one means is that it can be laid out in ordinary 2D space without any lines crossing. Kuratowski's theorem that planarity is associated with the absence of specific subgraphs in a network is an important result in graph theory established in the late 1920s. A subgraph is formally defined to be what one gets by selecting just some subset of connections in a network—and with this definition Kuratowski's theorem must allow extensions of K_5 and $K_{3,3}$ where extra nodes have been inserted in the middle of connections. (K_5 and $K_{3,3}$ are examples of so-called complete graphs, obtained by taking sets of specified numbers of nodes and connecting them in all possible ways.) Another approach is to consider reducing whole networks to so-called minors by deleting connections or merging connected nodes, and in this case Wagner's theorem shows that any non-planar network must be exactly reducible to either K_5 or $K_{3,3}$.

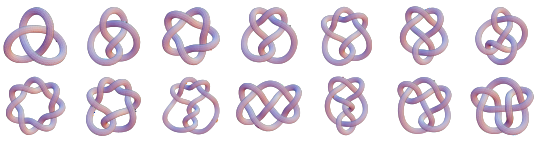
One can generalize the question of planarity to asking whether networks can be laid out on 2D surfaces with various topological structures—and in fact the genus of a graph can be defined to be the number of handles that must be added to a plane to embed the graph without crossings. But even on a torus it turns out that there is no finite set of (extended) subgraphs whose absence guarantees that a network can successfully be laid out. Nevertheless, if one considers minors a finite list does suffice—though for example on a torus it is known that at least 800 (and perhaps vastly more) are needed. (There is in fact a general theorem established since the 1980s that absolutely any list of networks—say for example ones that cannot be laid on a given surface—must actually in effect always all be reducible to some finite list of minors.) Note that finding the genus for a particular trivalent network is in general NP-complete.

■ **Page 527 · Gauge invariance.** It is often convenient to define quantities for which only differences or derivatives matter. In classical physics an example is electric potential, which can be shifted by any constant amount without affecting voltage differences or the electric field given by its gradient. In the mid-1800s the idea emerged of a vector potential whose curl gives the magnetic field, and it was soon recognized—notably by James Clerk Maxwell—that any function whose curl vanishes (and that can therefore normally be written as a gradient) could be added to the vector potential without affecting the magnetic field. By the end of the 1800s the general conditions on electromagnetic potentials for invariance of fields were known, though were not thought particularly significant. In 1918 Hermann Weyl tried to reproduce electromagnetism by adding the notion of an arbitrary scale or gauge to the metric of general relativity (see page 1028)—and noted the "gauge invariance" of his theory under simultaneous transformation of electromagnetic potentials and multiplication of the metric by a position-dependent factor. Following the introduction of the Schrödinger equation in quantum mechanics in 1926 it was almost immediately noticed that the equations for a charged particle in an electromagnetic field were invariant under gauge transformations in which the wave function was multiplied by a position-dependent phase factor. The idea then arose that perhaps some kind of gauge invariance could also be used as the basis for formulating theories of forces other than electromagnetism. And after a few earlier attempts, Yang-Mills theories were introduced in 1954 by extending the notion of a phase factor to an element of an arbitrary non-Abelian group. In the 1970s the Standard Model then emerged, based entirely on such theories. In mathematical terms, gauge theories can be viewed as describing fiber bundles in which connections between values of group elements in fibers at neighboring spacetime points are specified by gauge potentials—and curvatures correspond to gauge fields. (General relativity is in effect a special case in which the group elements are themselves related to spacetime coordinates.)

■ **Page 527 · Identifying particles.** In something like a class 4 cellular automaton it is quite straightforward to start enumerating possible persistent structures—as we saw in Chapter 6. But in a network system it can be much more difficult. Ultimately what one wants to do is to find what possible types of forms for local regions are inequivalent under the application of the underlying rules. But in general it may be undecidable even whether two such forms are actually equivalent (compare the notes below and on page

1051)—since to tell this one might need to be able to apply the rules infinitely many times. In specific cases, however, generalizations of concepts like planarity and homotopy may provide useful guides. And a first step may be to look at small closed networks and try to determine which of these can be transformed into each other by a given set of rules.

■ **Knot theory.** Somewhat analogous to the problem in the note above is the problem of classifying knots. The pictures below show some of the simplest distinct knots. But given presentations of two knots, no finite procedure is known that determines in general whether the knots are equivalent (or constructs a sequence of Reidemeister moves that transform one into the other). Quite probably this is in general undecidable, though since the 1920s a few polynomial invariants have been discovered—with recent ones being related to ideas from quantum field theory—that have allowed some progress to be made. (Even the problem of determining whether a knot specified by line segments is trivial is known to be NP-complete.)



■ **Page 528 · Charge quantization.** It is an observed fact that the electric and other charges of all particles are simple rational multiples of each other. In the context of electromagnetism alone, there would be no particular reason to expect this (unless magnetic monopoles exist). But as soon as different particles are related by a non-Abelian symmetry group, then the discreteness of the representations of such a group immediately implies that all charges must be rational multiples of each other.

■ **Spin.** Even when they appear to be of zero size, particles exhibit intrinsic angular momentum known as spin. The total spin is always a fixed multiple of the basic unit $\hbar$: 1/2 for quarks and leptons, 1 for photons and other ordinary gauge bosons, 2 for gravitons, and in theory 0 for Higgs particles. (Observed mesons have spins up to perhaps 5 and nuclei up to more than 50.) Particles of higher spin in effect require more information to specify their orientation (or polarization or its analog). And in the context of network models it could be that spin is somehow related to something as simple as the number of places at which the core of a particle is attached to the rest of the network. Spin values can be thought of as specifying which irreducible representation of the group of symmetries of spacetime is needed to describe a particle after momentum has been factored out. For ordinary massive particles in d-dimensional space the group is Spin(d), while for massless particles it is $E(d-1)$ (the Euclidean group). (For tachyons, it would be fundamentally non-compact, forcing continuous spin values.) For small transformations, Spin(d) is just the ordinary rotation group SO(d), but globally it is its universal cover, or SU(2) in 3D. And this can be thought of as what allows half-integer spins, which must be described by spinors rather than vectors or tensors. Such objects have the property that they are not left invariant by 360 ° rotations, but only by 720 ° ones—a feature potentially fairly easy to reproduce with networks, perhaps even without definite integer dimensions. In the standard formalism of quantum field theory it can be shown that (above 2D) half-integer spins must always be associated with fermions (which for example satisfy the exclusion principle), and integer spins with bosons. (This spin-statistics connection also seems to hold for various kinds of objects defined by extended field configurations.)

■ **Page 528 · Particle masses.** The measured masses of known elementary particles in units of GeV (roughly equal to the proton mass) are: photon: 0, electron: 0.000510998902; muon: 0.1056583569; τ lepton: 1.77705; W: 80.4; Z: 91.19. Recent evidence suggests a mass of about 10^{-11} GeV for at least one type of neutrino. Quarks and gluons presumably never occur as free particles, but still act in many ways as if they have definite masses. For all of them their confinement contributes perhaps 0.3 GeV of effective mass. Then there is also a direct mass: gluons 0; u: ~0.005; d ~0.01; s: ~0.2; c: 1.3; b: 4.4; t: 176 GeV. Note that among sets of particles that have the same quantum numbers—like d, s, b or γ, Z—mixing occurs that makes states of definite mass—that would propagate unchanged as free particles—differ by a unitary transformation from states that are left unchanged by interactions. When one sets up a quantum field theory one can typically in effect insert various mass parameters for particles. Self-interactions normally introduce formally infinite corrections—but if a theory is renormalizable then this means that there are only a limited number of independent such corrections, with the result that relations between masses of different particles are preserved. In quantum field theory any particle is always surrounded by a kind of cloud of virtual particles interacting with it. And following the Uncertainty Principle phenomena involving larger momentum scales will then to probe progressively smaller parts of this cloud—yielding different effective masses. (The masses tend to go up or down logarithmically with momentum scale—following so-called renormalization group equations.)

The Standard Model starts off with certain symmetries that force the masses of all ordinary particles to be zero. But then one assumes that nonzero masses are generated by spontaneous symmetry breaking. One starts by taking each particle to be coupled to a so-called Higgs field. Then one introduces self-interactions in this field so as to make its stable state be one that has constant nonzero value throughout the universe. But this means that as particles propagate, their interactions with the background give them an effective mass. And by having Higgs couplings be proportional to observed particle masses, it becomes inevitable that these will be the masses of particles. One prediction of the usual version of this mechanism for mass is that a definite Higgs particle should exist—which in the minimal Standard Model experiments should observe fairly soon. At times there have been hopes of so-called dynamical symmetry breaking giving the same effective results as the Higgs mechanism, but without an explicit Higgs field—perhaps through something similar to various phenomena in condensed matter physics. String theory, like the Standard Model, tends to start with zero mass particles—and then hopes that an appropriate Higgs-like mechanism will generate nonzero ones.

▪ **More particles.** To produce more massive particles requires higher-energy particle collisions, and today's accelerators only allow one to search up to masses of perhaps 200 GeV. (Sufficiently stable particles could have survived from the early universe, and a few cosmic ray interactions in principle give higher energies—but are normally too rare to be useful.) I am not sure whether in my approach one should expect an infinite series of progressively more massive particles. The example of nonplanarity might suggest not, but even in the class 4 cellular automata discussed in Chapter 6 it is not clear whether fundamentally different progressively larger structures will appear forever. In quantum field theory particles of any mass can always in principle exist for short times in virtual form. But normally their effects decrease like powers of their mass—making them hard to measure. In two kinds of cases, however, this does not happen: one is so-called anomalies, the other interactions with the Higgs field, in which couplings are proportional to mass. In the minimal Standard Model it turns out to be impossible to get quarks or leptons with masses much above about 200 GeV without destabilizing the vacuum (a fact pointed out by David Politzer and me in 1979). But with more complicated models one can avoid this constraint. In supersymmetric models—and string theory—there are typically also all sorts of other types of particles, assumed to have high masses since they have not been observed. There is evidence against any more than the three known generations of quarks and leptons in that the decay process $Z^0 \to \nu\,\bar{\nu}$ has a rate that rather accurately agrees with what is expected from just three types of low-mass neutrinos.

▪ **Page 530 · Expansion of the universe.** See page 1055.

The Phenomenon of Gravity

▪ **History.** With the Earth believed to be the center of the universe, gravity did not seem to require much explanation: it was just a force bringing things to a natural place. But with the advent of Copernican astronomy in the 1500s something more was needed. In the early 1600s Galileo noted that the force of gravity seems to depend only on the mass of an object, and not on any of its other features. In 1687 Isaac Newton then suggested a universal inverse square law of gravity between objects. In the 1700s and 1800s all sorts of celestial mechanics was done on the basis of this—with occasional observational anomalies being resolved for example by the discovery of new planets. Starting in the mid-1800s there were attempts to formulate gravity in the same way as electromagnetism—and in 1900 it was for example suggested that gravitational effects might propagate at the speed of light. Following his introduction of relativity theory in 1905, Albert Einstein began to seek a theory of gravity that would fit in with it. Ordinary special relativity has the feature that it assumes that systems behave the same regardless of their overall velocity—but not regardless of their acceleration. In 1907 Einstein then suggested the equivalence principle that gravity always locally has the same effect as an acceleration. (This principle requires only slightly more than Galileo's idea of the equivalence of gravitational and inertial mass, which has now been verified to the 10^{-12} level.) But by 1912 Einstein realized that if the effective laws of physics were somehow to remain the same in systems with different accelerations (or in different gravitational fields) then this would require a change in their perceived geometry. And building on ideas of differential geometry and tensor calculus from the late 1800s Einstein then began to formulate the concept that gravity is associated with curvature of space. In the late 1800s Ernst Mach had argued that phenomena like acceleration and rotation could ultimately be defined only relative to matter in the universe. And partly on this basis Einstein used the idea that curvature in space must be like a field produced by matter—leading eventually to his formulation in 1915 of the standard Einstein equations for general relativity. An immediate prediction of these was a deviation from the inverse square law, explaining an observed precession in the orbit of Mercury. After a dramatic verification in 1919 of predicted bending of light by

the Sun, general relativity began to be widely accepted. In the 1920s expansion of the universe was discovered, and this was seen to be consistent with general relativity. In the 1940s study of the evolution of stars then led to discussion of what became known as black holes. But for the most part general relativity was still viewed as being highly elegant though of little practical relevance. In the 1960s, however, more work began to be done on it. The discovery of the cosmic microwave background in 1965 led to increasing interest in cosmology. Precision tests—particularly with spacecraft—were designed. In calculations it was sometimes difficult to tell what was a genuine effect, and what was just a feature of the particular coordinates used. But a variety of increasingly abstract mathematical methods were developed, leading notably to general theorems about inevitability of singularities. Detailed calculations tended to require complicated symbolic tensor manipulation (with some associated problems being NP-complete), but with the development of computer algebra this gradually became more feasible—and by the mid-1970s approximate numerical methods were also being used. Various alternative formulations of general relativity were proposed, based for example on tetrads, spinors and twistors (and more recently on connection, loop and non-commutative geometry methods)—but none led to any great simplification. Meanwhile, there continued to be ever more accurate experimental tests of general relativity in the solar system—and at least in the weak gravitational fields available there (with metrics differing from the identity by at most one part in 10^6), all have worked out to around the 10^{-3} level. Starting in the 1960s, more and more ambitious gravitational wave detectors have been built—although none as yet have actually observed anything. Measurements done on a binary pulsar system are nevertheless consistent at a 10^{-3} level with the emission of gravitational radiation in a fairly strong gravitational field at the rate implied by general relativity. And since the 1980s there has been increasing conviction that at least indirect effects of black holes associated with very strong gravitational fields are being observed.

Over the years, some variants of general relativity have been proposed. At least when formulated in terms of tensors, none have quite the simplicity of the original theory—but some lead to rather different predictions, such as an absence of singularities like black holes. Ever since quantum theory began in the early 1900s there has been discussion of quantum gravity—and almost every major method developed for handling other quantum phenomena has been tried on gravity. Starting in the 1980s a variety of methods more specific to quantum gravity were also pursued, but none have yet had convincing success. (See page 1054.)

▪ **Differential geometry.** Standard descriptions of properties like curvature—as used for example in general relativity—are normally based on differential geometry. In its usual formulation this assumes that space is continuous, and can always effectively be treated as some kind of deformed version of ordinary Euclidean space—thus forming what is known as a manifold. The result of this is that points in space can always be specified by lists of coordinates—although historically one of the objectives of differential geometry has been to find ways to define properties like curvature so that they do not depend on the choice of such coordinates. The geometrical properties of a space are in general specified by its so-called metric—and this metric allows one to compute quantities based on lengths and angles from coordinates. The metric can be written as a matrix g, defined so that the analog for infinitesimal vectors u and v of $u \cdot v$ in ordinary Euclidean space is $u \cdot g \cdot v$. (This is essentially equivalent to saying that infinitesimal arc length is related to infinitesimal coordinate distances by $ds^2 = g_{i,j}\, dx_i\, dx_j$.) In d dimensions the metric g for a so-called Riemannian space can in general be any $d \times d$ positive-definite symmetric matrix—and can vary with position. But for ordinary flat Euclidean space it is always just *IdentityMatrix[d]* (at least with Cartesian coordinates). Within say a surface whose points $\{x_1, x_2, \ldots\}$ are obtained by evaluating an expression e as a function of parameters p (so that for example $e = \{x, y, f[x, y]\}$, $p = \{x, y\}$ for a *Plot3D* surface) the metric turns out to be given by

```
(Transpose[#] . # &)[Outer[D, e, p]]
```

In ordinary Euclidean space a defining feature of geometry is that the shortest path between two points is a straight line. But in an arbitrary space things can be more complicated, and in general such a path will be a geodesic (see note below) which can have a more complicated form. If the coordinates along a path are given by an expression s (such as $\{t, 1 + t, t^2\}$) that depends on a parameter t, and the metric at position p is $g[p]$, then the length of a path turns out to be

```
Integrate[Sqrt[∂t s . g[s] . ∂t s], {t, t1, t2}]
```

and geodesics then correspond to paths that extremize this quantity. In ordinary Euclidean space, such paths are straight lines, so that the length of a path between points with lists of coordinates a and b is just the ordinary Euclidean distance *Sqrt[(a - b) . (a - b)]*. But in general, even though geodesics are not straight lines their lengths can still be used to define a so-called geodesic distance—which turns out to have all the various properties of a distance discussed on page 1030.

If one draws a circle of radius r on a page, then the smaller r is, the more curved the circle will be—and one can define the

circle to have a constant curvature equal to $1/r$. If one draws a more general curve on a page, one can define its curvature at every point by seeing what size of circle fits it best at that point—or equivalently what the coefficients are in a quadratic approximation. (Compare page 418.) With a 2D surface in ordinary 3D space, one can imagine fitting quadrics (generalized ellipsoids). But these are now specified by two radii, yielding two principal curvatures. And in general these curvatures depend on the way the surface is laid out in 3D space. But a crucial point noted by Carl-Friedrich Gauss in the 1820s is that the product of such curvatures—the so-called Gaussian curvature—is always independent of how the surface is laid out, and can thus be viewed as intrinsic to the surface itself, and for example determined purely from the metric for the 2D space corresponding to the surface.

In a 2D space, intrinsic curvature is completely specified just by Gaussian curvature. In higher-dimensional spaces, there are more components, but in general they are all part of the so-called Riemann tensor—a rank-4 tensor introduced by Bernhard Riemann in 1854. (In *Mathematica*, the explicit form of such a tensor can be represented as a nested list for which `TensorRank[list] == 4`.) Several descriptions of the Riemann tensor can be given. One is based on looking at infinitesimal vectors u, v and w and asking how much w differs when transported two ways around the edges of a parallelogram, from x to $x+u+v$ via $x+u$ and via $x+v$. In ordinary flat space there is no difference, but in general the difference is a vector that is defined to be `Riemann . u . v . w`. (The `Riemann` that appears here is formally R_{ijk}^{l}.) Another description of the Riemann tensor is based on geodesics. In flat Euclidean space any two geodesics that start parallel always remain so. But a defining feature of general non-Euclidean spaces is that this is not in general so. And it turns out that the Riemann tensor is what determines the rate at which geodesics deviate from being parallel. Still another description of the Riemann tensor is as the coefficient of the quadratic terms in an expansion of the metric about a particular point, using so-called normal coordinates set up to make linear terms vanish. In general the Riemann tensor can always be computed from the metric, though it is somewhat complicated. If p is a list of coordinate parameters that appear in a d-dimensional metric g, then

```
Riemann = Table[∂p[[j]] Γ[[i, k]] - ∂p[[i]] Γ[[j, k]] +
    Γ[[i, k]] . Γ[[j]] - Γ[[j, k]] . Γ[[i]], {i, d}, {j, d}, {k, d}]
```

where the so-called Christoffel symbol Γ_{ij}^{k} is

```
Γ = With[{gi = Inverse[g]}, Table[Sum[
      gi[[l, k]] (∂p[[j]] g[[i, l]] + ∂p[[i]] g[[j, l]] - ∂p[[l]] g[[i, j]]),
      {l, d}], {i, d}, {j, d}, {k, d}]]/2
```

There are d^4 elements in the nested lists for `Riemann`, but symmetries and the so-called Bianchi identity reduce the number of independent components to $1/12\, d^2\, (d^2-1)$—or 20 for $d=4$. One can then compute the Ricci tensor ($R_{ik} = R_{ijk}^{\ j}$) using

```
RicciTensor = Map[Tr, Transpose[Riemann, {1, 3, 2, 4}], {2}]
```

and this has $1/2\, d\, (d+1)$ independent components in $d>2$ dimensions. (The parts of the Riemann tensor not captured by the Ricci tensor correspond to the so-called Weyl tensor; for $d=2$ the Ricci tensor has only one independent component, equal to the negative of the Gaussian curvature.) Finally, the Ricci scalar curvature is given by

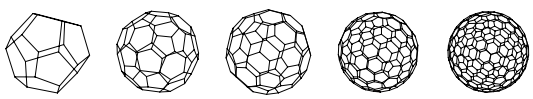
```
RicciScalar = Tr[RicciTensor . Inverse[g]]
```

▪ **Page 531 · Geodesics.** On a sphere all geodesics are arcs of great circles. On a surface of constant negative curvature (like (c)) geodesics diverge exponentially, as noted in early work on chaos theory (see page 971). The path of a geodesic can in general be found by requiring that the analog of acceleration vanishes for it. In the case of a surface defined by `z == f[x, y]` this is equivalent to solving

```
x''[t] == -(f^(1,0)[x[t], y[t]] (y'[t]^2 f^(0,2)[x[t], y[t]] +
     2 x'[t] y'[t] f^(1,1)[x[t], y[t]] + x'[t]^2 f^(2,0)[x[t], y[t]]))/
  (1 + f^(0,1)[x[t], y[t]]^2 + f^(1,0)[x[t], y[t]]^2)
```

together with the corresponding equation for y'', as already noted by Leonhard Euler in 1728 in connection with his development of the calculus of variations.

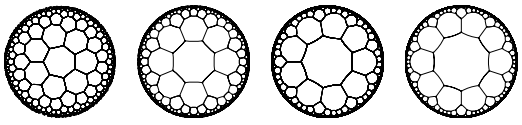
▪ **Page 532 · Spherical networks.** One can construct networks of constant positive curvature by approximating the surface of a sphere—starting with a dodecahedron and adding hexagons. (Euler's theorem implies that at any stage there must always be exactly 12 pentagonal faces.) The following are examples with 20, 60, 80, 180 and 320 nodes:



The object with 60 nodes is a truncated icosahedron—the shape of a standard soccer ball, as well the shape of the fullerene molecule C_{60}. (Note that in C_{60} one of the connections at each node is always a double chemical bond, since carbon has valence 4.) Geodesic domes are typically duals of such networks—with three edges on each face.

▪ **Hyperbolic networks.** Any surface that always has positive curvature must eventually close up to form something like a sphere. But a surface that has negative curvature (and no holes) must in some sense be infinite—more like cases (c) and (d) on page 412. Yet even in such a case one can always define coordinates that nominally allow the surface to be drawn in a finite way—and the Poincaré disk model used in the pictures below is the standard way of doing this. In ordinary flat space, regular polygons with more than 6

sides can never form a tessellation. But in a space with negative curvature this is possible for polygons with arbitrarily many sides—and the networks that result have been much studied as Cayley graphs of Fuchsian groups. One feature of these networks is that the number of nodes reached in them by following r connections always grows like 2^r. But if one intersperses hexagons in the networks (as in the main text) then one finds that for small r the number of nodes just grows like r^2—as one would expect for something like a 2D surface. But if one tries to look at growth rates on scales that are not small compared to characteristic lengths associated with curvature then one again sees exponential growth—just as in the case of a uniform tessellation without hexagons.



▪ **Page 533 · Sphere volumes.** In ordinary flat Euclidean space the area of a 2D circle is πr^2, and the volume of a 3D sphere $4 \pi r^3/3$. In general, the volume of a sphere in d-dimensional Euclidean space is $s[d]\, r^d$ where $s[d] = \pi^{d/2}/(d/2)!$ (the surface area is $d\, s[d]\, r^{d-1}$). (The function $s[d]$ has a maximum around $d = 5.26$, then decreases rapidly with d.)

If instead of flat space one considers a space defined by the surface of a 3D sphere—say with radius a—one can ask about areas of circles in this space. Such circles are no longer flat, but instead are like caps on the sphere—with a circle of radius r containing all points that are geodesic (great circle) distance less than r from its center. Such a circle has area

$$2 \pi a^2 (1 - \text{Cos}[r/a]) = \pi r^2 (1 - r^2/(12 a^2) + r^4/(360 a^4) - \ldots)$$

In the d-dimensional space corresponding to the surface of a $(d+1)$-dimensional sphere of radius a, the volume of a d-dimensional sphere of radius r is similarly given by

$$d\, s[d]\, a^d \text{Integrate}[\text{Sin}[\theta]^{d-1}, \{\theta, 0, r/a\}] = s[d]\, r^d (1 - d (d-1) r^2/((6 (d+2)) a^2) + (d (5 d^2 - 12 d + 7)) r^4/((360 (d+4)) a^4) + \ldots)$$

where

$$\text{Integrate}[\text{Sin}[x]^{d-1}, x] = -\text{Cos}[x]\ \text{Hypergeometric2F1}[1/2, (2-d)/2, 3/2, \text{Cos}[x]^2]$$

In an arbitrary d-dimensional space the volume of a sphere can depend on position, but in general it is given by

$$s[d]\, r^d (1 - \text{RicciScalar}\, r^2/(6 (d+2)) + \ldots)$$

where the Ricci scalar curvature is evaluated at the position of the sphere. (The space corresponding to a $(d+1)$-dimensional sphere has $\text{RicciScalar} = d (d-1)/a^2$.) The $d = 2$ version of this formula was derived in 1848; the general case in 1917 and 1939. Various derivations can be given. One can start from the fact that the volume density in any space is given in terms of the metric by $\text{Sqrt}[\text{Det}[g]]$. But in normal coordinates the first non-trivial term in the expansion of the metric is proportional to the Riemann tensor, yet the symmetry of a spherical volume makes it inevitable that the Ricci scalar is the only combination of components that can appear at lowest order. To next order the result is

$$s[d]\, r^d (1 - \text{RicciScalar}\, r^2/(6 (d+2)) + (5\, \text{RicciScalar}^2 - 3\, \text{RiemannNorm} + 8\, \text{RicciNorm} - 18\, \text{Laplacian}[\text{RicciScalar}])\, r^4/(360 (d+2) (d+4)) + \ldots)$$

where the new quantities involved are

```
RicciNorm = Norm[RicciTensor, {g, g}]

RiemannNorm = Norm[Riemann, {g, g, g, Inverse[g]}]

Norm[t_, gl_] := Tr[Flatten[t Dual[t, gl]]]

Dual[t_, gl_] := Fold[Transpose[#1 . Inverse[#2], RotateLeft[
                    Range[TensorRank[t]]]] &, t, Reverse[gl]]

Laplacian[f_] := Inner[D, Sqrt[Det[g]]
                   (Inverse[g] . Map[∂# f &, p]), p]/Sqrt[Det[g]]
```

In general the series in r may not converge, but it is known that at least in most cases only flat space can give a result that shows no correction to the basic r^d form. It is also known that if the Ricci tensor is non-negative, then the volume never grows faster than r^d.

▪ **Cylinder volumes.** In any d-dimensional space, the volume of a cylinder of length x and radius r whose direction is defined by a unit vector v turns out to be given by

$$s[d-1]\, r^{d-1}\, x\ (1 - (d-1) (\text{RicciScalar} - \text{RicciTensor} \,.\, v \,.\, v)\, r^2/(d+1) + \ldots)$$

Note that what determines the volume of the cylinder is curvature orthogonal to its direction—and this is what leads to the combination of Ricci scalar and tensor that appears.

▪ **Page 533 · Discrete spaces.** Most work with surfaces done on computers—whether for computer graphics, computer-aided design, solving boundary value problems or otherwise—makes use of discrete approximations. Typically surfaces are represented by collections of patches—with a simple mesh of triangles often being used. The triangles are however normally specified not so much by their network of connections as by the explicit coordinates of their vertices. And while there are various triangulation methods that for example avoid triangles with small angles, no standard method yields networks analogous to the ones I consider in which all triangle edges are effectively the same length.

In pure mathematics a basic idea in topology has been to look for finite or discrete ways to capture essential features of continuous surfaces and spaces. And as an early part of this Henri Poincaré in the 1890s introduced the concept of approximating manifolds by cell complexes consisting of collections of generalized polyhedra. By the 1920s there was

then extensive work on so-called combinatorial topology, in which spaces are thought of as being decomposed into abstract complexes consisting say of triangles, tetrahedra and higher-dimensional simplices. But while explicit coordinates and lengths are not usually discussed, it is still imagined that one knows more information than in the networks I consider: not only how vertices are connected by edges, but also how edges are arranged around faces, faces around volumes, and so on. And while in 2D and 3D it is possible to set up such an approximation to any manifold in this way, it turns out that at least in 5D and above it is not. Before the 1960s it had been hoped that in accordance with the Hauptvermutung of combinatorial topology it would be possible to tell whether a continuous mapping and thus topological equivalence exists between manifolds just by seeing whether subdivisions of simplicial complexes for them could be identical. But in the 1960s it was discovered that at least in 5D and above this will not always work. And largely as a result of this, there has tended to be less interest in ideas like simplicial complexes.

And indeed a crucial point for my discussion in the main text is that in formulating general relativity one actually does not appear to need all the structure of a simplicial complex. In fact, the only features of manifolds that ultimately seem relevant are ones that in appropriate limits are determined just from the connectivity of networks. The details of the limits are mathematically somewhat intricate (compare page 1030), but the basic approach is straightforward. One can find the volume of a sphere (geodesic ball) in a network just by counting the number of nodes out to a given network distance from a certain node. And from the limiting growth rate of this one can immediately get the Ricci scalar curvature—just as in the continuous case discussed above. To get the Ricci tensor one also needs a direction. But one can get this from a geodesic—which is in effect the analog of a straight line in the network. Note that unlike in a continuous space there is however usually no obvious way to continue a geodesic in a network. And in general, some—but not all—of the standard constructions used in continuous spaces can also immediately be used in networks. So for example it is straightforward to construct a triangle in a network: one just starts from a particular node, follows geodesics to two others, then joins these with a geodesic. But to extend the triangle into a parallelogram is not so easy—since there is no immediate notion of parallelism in the network. And this means that neither the Riemann tensor, nor a so-called Schild ladder for parallel transport, can readily be constructed.

Since the 1980s there has been increasing interest in formulating notions of continuous geometry for objects like Cayley graphs of groups—which are fundamentally discrete but have infinite limits analogous to continuous systems. (Compare page 938.)

▪ **Manifold undecidability.** Given a particular set of network substitution rules there is in general no finite way to decide whether any sequence of such rules exists that will transform particular networks into each other. (Compare undecidability in multiway systems on page 779.) And although one might not expect it on the basis of traditional mathematical intuition, there is an analog of this even for topological equivalence of ordinary continuous manifolds. For the fundamental groups that represent how basic loops can be combined must be equivalent for equivalent manifolds. Yet it turns out that in 4D and above the fundamental group can have essentially any set of generators and relations—so that the undecidability of the word problem for arbitrary groups (see page 1141) implies undecidability of equivalence of manifolds. (In 2D it is straightforward to decide equivalence, and in 3D it is known that only some fundamental groups can be obtained—roughly because not all networks can be embedded in 2D—and it is expected that it will ultimately be possible to decide equivalence.)

▪ **Non-integer dimensions.** Unlike in traditional differential geometry (and general relativity) my formulation of space as a network potentially allows concepts like curvature to be defined even outside of integers numbers of dimensions.

▪ **Page 534 · Lorentzian spaces.** In ordinary Euclidean space distance is given by $Sqrt[x^2 + y^2 + z^2]$. In setting up relativity theory it is convenient (see page 1042) to define an analog of distance (so-called proper time) in 4D spacetime by $Sqrt[c^2 t^2 - x^2 - y^2 - z^2]$. And in terms of differential geometry such Minkowski space can be specified by the metric $DiagonalMatrix[\{+1, -1, -1, -1\}]$ (now taking $c = 1$). To set up general relativity one then considers not Riemannian manifolds but instead Lorentzian ones in which the metric is not positive definite, but instead has the signature of Minkowski space.

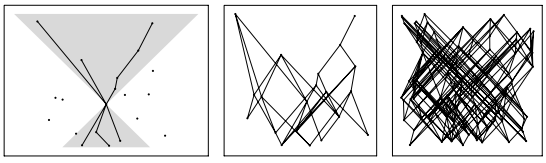
In such Lorentzian spaces, however, there is no useful immediate analog of a sphere. For given any point, even the light cone that corresponds to points at zero spacetime distance from it has an infinite volume. But with an appropriate definition one can still set up cones that have finite volume. To do this in general one starts by picking a vector e in a timelike direction, then normalizes it to be a unit vector so that $e \cdot g \cdot e == -1$. Then one defines a cone of height t whose apex is a given point to be those points whose displacement

vector v satisfies $0 > e \cdot g \cdot v > -t$ (and $0 > v \cdot g \cdot v$). And the volume of such a cone then turns out to be

$$s[d]\, t^{d+1} (1 - t^2 (d+1)(d\, \text{RicciScalar} + 2(d+1)(\text{RicciTensor} \cdot e \cdot e))/((d+2)(d+3)) + \ldots)/(d+1)$$

■ **Torsion.** In standard geometry, one assumes that the distance from one point to another is the same as the distance back, so that the metric tensor can be taken to be symmetric, and there is zero so-called torsion. But in for example a causal network, connections have definite directions, and there is in general no such symmetry. And if one looks at the volume of a cone this can then introduce a correction proportional to r. But as soon as there is enough uniformity to define a reasonable notion of static space, it seems that this effect must vanish. (Note that in pure mathematics there are several different uses of the word "torsion". Here I use it to refer to the antisymmetric parts of the metric tensor.)

■ **Random causal networks.** If one assumes that there are events at random positions in continuous spacetime, then one can construct an effective causal network for them by setting up connections between each event and all events in its future light cone—then deleting connections that are redundant in the sense that they just provide shortcuts to events that could otherwise be reached by following multiple connections. The pictures below show examples of causal networks obtained in this way. The number of connections generally increases faster than linearly with the number of events. Most links end up being at angles that are close to the edge of the light cone.



■ **Page 534 · Einstein equations.** In the absence of matter, the standard statement of the Einstein equations is that all components of the Ricci tensor—and thus also the Ricci scalar—must be zero (or formally that $R_{ij} = 0$). But since the vanishing of all components of a tensor must be independent of the coordinates used, it follows that the vacuum Einstein equations are equivalent to the statement $\text{RicciTensor} \cdot e \cdot e == 0$ for all timelike unit vectors e—a statement that can readily be applied to networks of the kind I consider in the main text. (A related statement is that the 3D Ricci scalar curvature of all spacelike hypersurfaces must vanish wherever these have vanishing extrinsic curvature.)

Another way to state the Einstein equations—already discussed by David Hilbert in 1915—is as the constraint that the integral of $\text{RicciScalar}\ \text{Sqrt}[\text{Det}[g]]$ (the so-called Einstein-Hilbert action) be an extremum. (An idealized soap film or other minimal surface extremizes the integral of the intrinsic volume element $\text{Sqrt}[\text{Det}[g]]$, without a RicciScalar factor.) In the discrete Regge calculus that I mention on page 1054 this variational principle turns out to have a rather simple form.

The Einstein-Hilbert action—and the Einstein equations—can be viewed as having the simplest forms that do not ultimately depend on the choice of coordinates. Higher-order terms—say powers of the Ricci scalar curvature—could well arise from underlying network systems, but would not contribute noticeably except in very high gravitational fields.

Various physical interpretations can be given of the vanishing of the Ricci tensor implied by the ordinary vacuum Einstein equations. Closely related to my discussion of the absence of t^2 terms in volume growth for 4D spacetime cones is the statement that if one sets up a small 3D ball of comoving test particles then the volume it defines must have zero first and second derivatives with time.

Below 4D the vanishing of the Ricci tensor immediately implies the vanishing of all components of the Riemann tensor—so that the vacuum Einstein equations force space at least locally to have its ordinary flat form. (Even in 2D there can nevertheless still be non-trivial global topology—for example with flat space having its edges identified as on a torus. In the Euclidean case there were for a long time no non-trivial solutions to the Einstein equations known in any number of dimensions, but in the 1970s examples were found, including large families of Calabi-Yau manifolds.)

In the presence of matter, the typical formal statement of the full Einstein equations is $R_{\mu\nu} - R\, g_{\mu\nu}/2 = 8\pi\, G\, T_{\mu\nu}/c^4$, where $T_{\mu\nu}$ is the energy-momentum (stress-energy) tensor for matter and G is the gravitational constant. (An additional so-called cosmological term $\lambda\, g_{\mu\nu}$ is sometimes added on the right to adjust the effective overall energy density of the universe, and thus its expansion rate. Note that the equation can also be written $R_{\mu\nu} = 8\pi\, G (T_{\mu\nu} - 1/2\, T^{\mu}_{\mu}\, g_{\mu\nu})/c^4$.) The μ, ν component of $T_{\mu\nu}$ gives the flux of the μ component of 4-momentum (whose components are energy and ordinary 3-momentum) in the ν direction. The fact that T_{00} is energy density implies that for static matter (where $E = m c^2$) the equation is in a sense a minimal extension of Poisson's equation of Newtonian gravity theory. Note that conservation of energy and momentum implies that $T_{\mu\nu}$ must have zero divergence—a result guaranteed in the Einstein equations by the structure of the left-hand side.

In the variational approach to gravity mentioned above, the *RicciScalar* plays the role of a Lagrangian density for pure gravity—and in the presence of matter the Lagrangian density for matter must be added to it. At a physical level, the full Einstein equations can be interpreted as saying that the volume v of a small ball of comoving test particles satisfies

$$\partial_{tt} v[t]/v[t] == -1/2\,(\rho + 3\,p)$$

where ρ is the total energy density and p is the pressure averaged over all space directions.

To solve the full Einstein equations in any particular physical situation requires a knowledge of $T_{\mu\nu}$—and thus of properties of matter such as the relation between pressure and energy density (equation of state). Quite a few global results about the formation of singularities and the absence of paths looping around in time can nevertheless be obtained just by assuming certain so-called energy conditions for $T_{\mu\nu}$. (A fairly stringent example is $0 \le p \le \rho/3$—and whether this is actually true for non-trivial interacting quantum fields remains unclear.)

In their usual formulation, the Einstein equations are thought of as defining constraints on the structure of 4D spacetime. But at some level they can also be viewed as defining how 3D space evolves with time. And indeed the so-called initial value formulations constructed in the 1960s allow one to start with a 3D metric and various extrinsic curvatures defined for a 3D spacelike hypersurface, and then work out how these change on successive hypersurfaces. But at least in terms of tensors, the equations involved show nothing like the simplicity of the usual 4D Einstein equations. One can potentially view the causal networks that I discuss in the main text as providing another approach to setting up an initial value formulation of the Einstein equations.

▪ **Page 536 · Pure gravity.** In the absence of matter, the Einstein equations always admit ordinary flat Minkowski space as a solution. But they also admit other solutions that in effect represent configurations of pure gravitational field. And in fact the 4D vacuum Einstein equations are already a sophisticated set of nonlinear partial differential equations that can support all sorts of complex behavior. Several tens of families of solutions to the equations have been found—some with obvious physical interpretations, others without.

Already in 1916 Karl Schwarzschild gave the solution for a spherically symmetric gravitational field. He imagined that this field itself existed in a vacuum—but that it was produced by a mass such as a star at its center. In its original form the metric becomes singular at radius $2\,G\,m/c^2$ (or $3\,m$ km with m in solar masses). At first it was assumed that this would always be inside a star, where the vacuum Einstein equations would not apply. But in the 1930s it was suggested that stars could collapse to concentrate their mass in a smaller radius. The singularity was then interpreted as an event horizon that separates the interior of a black hole from the ordinary space around it. In 1960 it was realized, however, that appropriate coordinates allowed smooth continuation across the event horizon—and that the only genuine singularity was infinite curvature at a single point at the center. Sometimes it was said that this must reflect the presence of a point mass, but soon it was typically just said to be a point at which the Einstein equations—for whatever reason—do not apply. Different choices of coordinates led to different apparent locations and forms of the singularity, and by the late 1970s the most common representation was just a smooth manifold with a topology reflecting the removal of a point—and without any specific reference to the presence of matter.

Appealing to ideas of Ernst Mach from the late 1800s it has often been assumed that to get curvature in space always eventually requires the presence of matter. But in fact even the vacuum Einstein equations for complete universes (with no points left out) have solutions that show curvature. If one assumes that space is both homogeneous and isotropic then it turns out that only ordinary flat Minkowski space is allowed. (When matter or a cosmological term is present one gets different solutions—that always expand or contract, and are much studied in cosmology.) If anisotropy is present, however, then there can be all sorts of solutions—classified for example as having different Bianchi symmetry types. And a variety of inhomogeneous solutions with no singularities are also known—an example being the 1962 Ozsváth-Schücking rotating vacuum. But in all cases the structure is too simple to capture much that seems relevant for our present universe.

One form of solution to the vacuum Einstein equations is a gravitational wave consisting of a small perturbation propagating through flat space. No solutions have yet been found that represent complete universes containing emitters and absorbers of such waves (or even for example just two massive bodies). But it is known that combinations of gravitational waves can be set up that will for example evolve to generate singularities. And I suspect that nonlinear interactions between such waves will also inevitably lead to the analog of turbulence for pure gravity. (Numerical simulations often show all sorts of complex behavior—but in the past this has normally been attributed just to the approximations used. Note that for example Bianchi type IX solutions for a complete universe show sensitive dependence on initial conditions—and no doubt this can also happen with nonlinear gravitational waves.)

As mentioned on page 1028, Albert Einstein considered the possibility that particles of matter might somehow just be localized structures in gravitational and electromagnetic fields. And in the mid-1950s John Wheeler studied explicit simple examples of such so-called geons. But in all cases they were found to be unstable—decaying into ordinary gravitational waves. The idea of having purely gravitational localized structures has also occasionally been considered—but so far no stable field configuration has been found. (And no purely repetitive solutions can exist.)

The equivalence principle (see page 1047) might suggest that anything with mass—or energy—should affect the curvature of space in the same way. But in the Einstein equations the energy-momentum tensor is not supposed to include contributions from the gravitational field. (There are alternative and seemingly inelegant theories of gravity that work differently—and notably do not yield black holes. The setup is also somewhat different in recent versions of string theory.) The very definition of energy for the gravitational field is not particularly straightforward in general relativity. But perhaps a definition could be found that would allow localized structures in the gravitational field to make effective contributions to the energy-momentum tensor that would mimic those from explicit particles of matter. Nevertheless, there are quite a few phenomena associated with particles that seem difficult to reproduce with pure gravity—at least say without extra dimensions. One example is parity violation; another is the presence of long-range forces other than gravity.

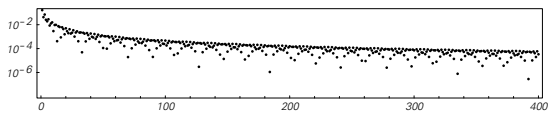
▪ **Quantum gravity.** That there should be quantum effects in gravity was already noted in the 1910s, and when quantum field theory began to develop in the 1930s, there were immediately attempts to apply it to gravity. The first idea was to represent gravity as a field that exists in flat spacetime, and by analogy with photons in quantum electrodynamics to introduce gravitons (at one point identified with neutrinos). By the mid-1950s a path integral (see page 1061) based on the Einstein-Hilbert action had been constructed, and by the early 1960s Feynman diagram rules had been derived, and it had been verified that tree diagrams involving gravitons gave results that agreed with general relativity for small gravitational fields. But as soon as loop diagrams were considered, infinities began to appear. And unlike for quantum electrodynamics there did not seem to be only a finite number of these—that could be removed by renormalization. And in fact by 1973 gravity coupled to matter had been shown for certain not to be renormalizable—and the same was finally shown for pure gravity in 1986. There was an attempt in the 1970s and early 1980s to look directly at the path integral—without doing an expansion in terms of Feynman diagrams. But despite the fact that at least in Euclidean spacetime a variety of seemingly relevant field configurations were identified, many mathematical difficulties were encountered. And in the late-1970s there began to be interest in the idea that supersymmetric field theories might make infinities associated with gravitons be cancelled by ones associated with other particles. But in the end this did not work out. And then in the mid-1980s one of the great attractions of string theory was that it seemed to support graviton excitations without the problem of infinities seen in point-particle field theories. But it had other problems, and to avoid these, supersymmetry had to be introduced, leading to the presence of many other particles that have so far not been observed. (See also page 1029.)

Starting in the 1950s a rather different approach to quantum gravity involved trying to find a representation of the structure of spacetime in which a quantum analog of the Einstein equations could be obtained by the formal procedure of canonical quantization (see page 1058). Yet despite a few signs of progress in the 1960s there was great difficulty in finding appropriately independent variables to use. In the late 1980s, however, it was suggested that variables could be used corresponding roughly to gravitational fluxes through loops in space. And in terms of these loop variables it was at least formally possible to write down a version of quantum gravity. Yet while this was found in the 1990s to have a correspondence with spin networks (see below), it has remained impossible to see just how it might yield ordinary general relativity as a limit.

Even if one assumes that spacetime is in a sense ultimately continuous one can imagine investigating quantum gravity by doing some kind of discrete approximation. And in 1961 Tullio Regge noted that for a simplicial complex (see page 1050) the Einstein-Hilbert action has a rather simple form in terms of angles between edges. Starting in the 1980s after the development of lattice gauge theories, simulations of random surfaces and higher-dimensional spaces set up in this way were done—often using so-called dynamic triangulation based on random sequences of generalized Alexander moves from page 1038. But there were difficulties with Lorentzian spaces, and when large-scale average behavior was studied, it seemed difficult to reproduce observed smooth spacetime. Analytical approaches (that happened to be like 0D string theory) were also found for 2D discrete spacetimes (compare page 1038)—but they were not successfully extended to higher dimensions.

Over the years, various attempts have been made to derive quantum gravity from fundamentally discrete models of

spacetime (compare page 1027). In recent times the most widely discussed have been spin networks—which despite their name ultimately seem to have fairly little to do with the systems I consider. Spin networks were introduced in 1964 by Roger Penrose as a way to set up an intrinsically quantum mechanical model of spacetime. A simple analog involves a 2D surface made out of triangles whose edges have integer lengths j_i. If one computes the product of $Exp[i(j_1 + j_2 - j_3)]$ for all triangles, then it turns out for example that this quantity is extremized exactly when the whole surface is flat. In 3D one imagines breaking space into tetrahedra whose edge lengths correspond to discrete quantum spin values. And in 1968 Tullio Regge and Giorgio Ponzano suggested—almost as an afterthought in technical work on $6j$ symbols—that the quantum probability amplitude for any form of space might perhaps be given by the product of $6j$ symbols for the spins on each tetrahedron. The $SixJSymbol[\{j_1, j_2, j_3\}, \{j_4, j_5, j_6\}]$ are slightly esoteric objects that correspond to recoupling coefficients for the 3D rotation group SO(3), and that arose in 1940s studies of combinations of three angular momenta in atomic physics—and were often represented graphically as networks. For large j_i they are approximated by $Cos[\theta + \pi/4]/Sqrt[12 \pi v]$, where v is the volume of the tetrahedron and θ is a deficit angle. And from this it turns out that limits of products of $6j$ symbols correspond essentially to $Exp[i s]$, where s is the discrete form of the Einstein-Hilbert action—extremized by flat 3D space. (The picture below shows for example $Abs[SixJSymbol[\{j, j, j\}, \{j, j, j\}]]$. Note that for any j the $6j$ symbols can be given in terms of *HypergeometricPFQ*.)



In the early 1990s there was again interest in spin networks when the Turaev-Viro invariant for 3D spaces was discovered from a topological field theory involving triangulations weighted with $6j$ symbols of the quantum group $SU(2)_q$—and it was seen that invariance under Alexander moves on the triangulation corresponded to the Biedenharn-Elliott identity for $6j$ symbols. In the mid-1990s it was then found that states in 3D loop quantum gravity (see above) could be represented in terms of spin networks—leading for example to quantization of all areas and volumes. In attempting extensions to 4D, spin foams have been introduced—and variously interpreted in terms of simplified Feynman diagrams, constructs in multidimensional category theory, and possible evolutions of spin networks. In all cases, however, spin networks and spin foams seem to be viewed just as calculational constructs that must be evaluated and added together to get quantum amplitudes—quite different from my idea of associating an explicit evolution history for the universe with the evolution of a network.

▪ **Cosmology.** On a large scale our universe appears to show a uniform expansion that makes progressively more distant galaxies recede from us at progressively higher speeds. In general relativity this is explained by saying that the initial conditions must have involved expansion—and that there is not enough in the way of matter or gravitational fields to produce the gravity to slow down this expansion too much. (Note that as soon as objects get gravitationally bound—like galaxies in clusters—there is no longer expansion between them.) The standard big bang model assumes that the universe starts with matter at what is in effect an arbitrarily high temperature. One issue much discussed in cosmology since the late 1970s is how the universe manages to be so uniform. Thermal equilibrium should eventually lead to uniformity—but different parts of the universe cannot come to equilibrium until there has at least been time for effects to propagate between them. Yet there seems for example to be overall uniformity in what we see if we look in opposite directions in the sky—even though extrapolating from the current rate of expansion there has not been enough time since the beginning of the universe for anything to propagate from one side to the other. But starting in the early 1980s it has been popular to think that early in its history the universe must have undergone a period of exponential expansion or so-called inflation. And what this would do is to take just a tiny region and make it large enough to correspond to everything we can now see in the universe. But the point is that a sufficiently tiny region will have had time to come to thermal equilibrium—and so will be approximately uniform, just as the cosmic microwave background is now observed to be. The actual process of inflation is usually assumed to reflect some form of phase transition associated with decreasing temperature of matter in the universe. Most often it is assumed that in the present universe a delicate balance must exist between energy density from a background Higgs field (see page 1047) and a cosmological term in the Einstein equations (see page 1052). But above a critical temperature thermal fluctuations should prevent the background from forming—leading to at least some period in which the universe is dominated by a cosmological term which yields exponential expansion. There tend to be various detailed problems with this scenario, but at least with a sufficiently complicated setup it seems possible to get results that are consistent with observations made so far.

In the context of the discoveries in this book, my expectation is that the universe started from a simple small network, then

progressively added more and more nodes as it evolved, until eventually on a large scale something corresponding to 4D spacetime emerged. And with this setup, the observed uniformity of the universe becomes much less surprising. Intrinsic randomness generation always tends to lead to a certain uniformity in networks. But the crucial point is that this will not take long to happen throughout any network if it is appropriately connected. Traditional models tend to assume that there are ultimately a fixed number of spacetime dimensions in the universe. And with this assumption it is inevitable that if the universe in a sense expands at the speed of light, then regions on opposite sides of it can essentially never share any common history. But in a network model the situation is different. The causal network always captures what happens. And in a case like page 518—with spacetime always effectively having a fixed finite dimension—points that are a distance t apart tend to have common ancestors only at least t steps back. But in a case like (a) on page 514—where spacetime has the structure of an exponentially growing tree—points a distance t apart typically have common ancestors just *Log[t]* steps back. And in fact many kinds of causal networks—say associated with early randomly connected space networks—will inevitably yield common ancestors for distant parts of the universe. (Note that such phenomena presumably occur at around the Planck scale of 10^{19} GeV rather than at the 10^{15} GeV or lower scale normally discussed in connection with inflation. They can to some extent be captured in general relativity by imagining an effective spacetime dimension that is initially infinite, then gradually decreases to 4.)

Quantum Phenomena

■ **History.** In classical physics quantities like energy were always assumed to correspond to continuous variables. But in 1900 Max Planck noticed that fits to the measured spectrum of electromagnetic radiation produced by hot objects could be explained if there were discrete quanta of electromagnetic energy. And by 1910 work by Albert Einstein, notably on the photoelectric effect and on heat capacities of solids, had given evidence for discrete quanta of energy in both light and matter. In 1913 Niels Bohr then made the suggestion that the discrete spectrum of light emitted by hydrogen atoms could be explained as being produced by electrons making transitions between orbits with discrete quantized angular momenta. By 1920 ideas from celestial mechanics had been used to develop a formalism for quantized orbits which successfully explained various features of atoms and chemical elements. But it was not clear how to extend the formalism say to a problem like propagation of light through a crystal. In 1925, however, Werner Heisenberg suggested a new and more general formalism that became known as matrix mechanics. The original idea was to imagine describing the state of an atom in terms of an array of amplitudes for virtual oscillators with each possible frequency. Particular conditions amounting to quantization were then imposed on matrices of transitions between these, and the idea was introduced that only certain kinds of amplitude combinations could ever be observed. In 1923 Louis de Broglie had suggested that just as light—which in optics was traditionally described in terms of waves—seemed in some respects to act like discrete particles, so conversely particles like electrons might in some respects act like waves. In 1926 Erwin Schrödinger then suggested a partial differential equation for the wave functions of particles like electrons. And when effectively restricted to a finite region, this equation allowed only certain modes, corresponding to discrete quantum states—whose properties turned out to be exactly the same as implied by matrix mechanics. In the late 1920s Paul Dirac developed a more abstract operator-based formalism. And by the end of the 1920s basic practical quantum mechanics was established in more or less the form it appears in textbooks today. In the period since, increasing computational capabilities have allowed coupled Schrödinger equations for progressively more particles to be solved (reasonably accurate solutions for hundreds of particles can now be found), allowing ever larger studies in atomic, molecular, nuclear and solid-state physics. A notable theoretical interest starting in the 1980s was so-called quantum chaos, in which it was found that modes (wave functions) in regions like stadiums that did not yield simple analytical solutions tended to show complicated and seemingly random forms.

Basic quantum mechanics is set up to describe how fixed numbers of particles behave—say in externally applied electromagnetic or other fields. But to describe things like fields one must allow particles to be created and destroyed. In the mid-1920s there was already discussion of how to set up a formalism for this, with an underlying idea again being to think in terms of virtual oscillators—but now one for each possible state of each possible one of any number of particles. At first this was just applied to a pure electromagnetic field of non-interacting photons, but by the end of the 1920s there was a version of quantum electrodynamics (QED) for interacting photons and electrons that is essentially the same as today. To find predictions from this theory a so-called perturbation expansion was made, with successive terms representing progressively more interactions, and each

having a higher power of the so-called coupling constant $\alpha \simeq 1/137$. It was immediately noticed, however, that self-interactions of particles would give rise to infinities, much as in classical electromagnetism. At first attempts were made to avoid this by modifying the basic theory (see page 1044). But by the mid-1940s detailed calculations were being done in which infinite parts were just being dropped—and the results were being found to agree rather precisely with experiments. In the late 1940s this procedure was then essentially justified by the idea of renormalization: that since in all possible QED processes only three different infinities can ever appear, these can in effect systematically be factored out from all predictions of the theory. Then in 1949 Feynman diagrams were introduced (see note below) to represent terms in the QED perturbation expansion—and the rules for these rapidly became what defined QED in essentially all practical applications. Evaluating Feynman diagrams involved extensive algebra, and indeed stimulated the development of computer algebra (including my own interest in the field). But by the 1970s the dozen or so standard processes discussed in QED had been calculated to order α^2—and by the mid-1980s the anomalous magnetic moment of the electron had been calculated to order α^4, and nearly one part in a trillion (see note below).

But despite the success of perturbation theory in QED it did not at first seem applicable to other issues in particle physics. The weak interactions involved in radioactive beta decay seemed too weak for anything beyond lowest order to be relevant—and in any case not renormalizable. And the strong interactions responsible for holding nuclei together (and associated for example with exchange of pions and other mesons) seemed too strong for it to make sense to do an expansion with larger numbers of individual interactions treated as less important. So this led in the 1960s to attempts to base theories just on setting up simple mathematical constraints on the overall so-called S matrix defining the mapping from incoming to outgoing quantum states. But by the end of the 1960s theoretical progress seemed blocked by basic questions about functions of several complex variables, and predictions that were made did not seem to work well.

By the early 1970s, however, there was increasing interest in so-called gauge or Yang-Mills theories formed in essence by generalizing QED to operate not just with a scalar charge, but with charges viewed as elements of non-Abelian groups. In 1972 it was shown that spontaneously broken gauge theories of the kind needed to describe weak interactions were renormalizable—allowing meaningful use of perturbation theory and Feynman diagrams. And then in 1973 it was discovered that QCD—the gauge theory for quarks and gluons with SU(3) color charges—was asymptotically free (it was known to be renormalizable), so that for processes probing sufficiently small distances, its effective coupling was small enough for perturbation theory. By the early 1980s first-order calculations of most basic QCD processes had been done—and by the 1990s second-order corrections were also known. Schemes for adding up all Feynman diagrams with certain very simple repetitive or other structures were developed. But despite a few results about large-distance analogs of renormalizability, the question of what QCD might imply for processes at larger distances could not really be addressed by such methods.

In 1941 Richard Feynman pointed out that amplitudes in quantum theory could be worked out by using path integrals that sum with appropriate weights contributions from all possible histories of a system. (The Schrödinger equation is like a diffusion equation in imaginary time, so the path integral for it can be thought of as like an enumeration of random walks. The idea of describing random walks with path integrals was discussed from the early 1900s.) At first the path integral was viewed mostly as a curiosity, but by the late 1970s it was emerging as the standard way to define a quantum field theory. Attempts were made to see if the path integral for QCD (and later for quantum gravity) could be approximated with a few exact solutions (such as instantons) to classical field equations. By the early 1980s there was then extensive work on lattice gauge theories in which the path integral (in Euclidean space) was approximated by randomly sampling discretized field configurations. But—I suspect for reasons that I discuss in the note below—such methods were never extremely successful. And the result is that beyond perturbation theory there is still no real example of a definitive success from standard relativistic quantum field theory. (In addition, even efforts in the context of so-called axiomatic field theory to set up mathematically rigorous formulations have run into many difficulties—with the only examples satisfying all proposed axioms typically in the end being field theories without any real interactions. In condensed matter physics there are nevertheless cases like the Kondo model where exact solutions have been found, and where the effective energy function for electrons happens to be roughly the same as in a relativistic theory.)

As mentioned on page 1044, ordinary quantum field theory in effect deals only with point particles. And indeed a recurring issue in it has been difficulty with constraints and redundant degrees of freedom—such as those associated with extended objects. (A typical goal is to find variables in which one can carry out what is known as canonical quantization: essentially applying the same straightforward

transformation of equations that happens to work in ordinary elementary quantum mechanics.) One feature of string theory and its generalizations is that they define presumably consistent quantum field theories for excitations of extended objects—though an analog of quantum field theory in which whole strings can be created and destroyed has not yet been developed.

When the formalism of quantum mechanics was developed in the mid-1920s there were immediately questions about its interpretation. But it was quickly suggested that given a wave function ψ from the Schrödinger equation $Abs[\psi]^2$ should represent probability—and essentially all practical applications have been based on this ever since. From a conceptual point of view it has however often seemed peculiar that a supposedly fundamental theory should talk only about probabilities. Following the introduction of the uncertainty principle and related formalism in the 1920s one idea that arose was that—in rough analogy to relativity theory—it might just be that there are only certain quantities that are observable in definite ways. But this was not enough, and by the 1930s it was being suggested that the validity of quantum mechanics might be a sign that whole new general frameworks for philosophy or logic were needed—a notion supported by the apparent need to bring consciousness into discussions about measurement in quantum mechanics (see page 1063). The peculiar character of quantum mechanics was again emphasized by the idealized experiment of Albert Einstein, Boris Podolsky and Nathan Rosen in 1935. But among most physicists the apparent lack of an ordinary mechanistic way to think about quantum mechanics ended up just being seen as another piece of evidence for the fundamental role of mathematical formalism in physics.

One way for probabilities to appear even in deterministic systems is for there to be hidden variables whose values are unknown. But following mathematical work in the early 1930s it was usually assumed that this could not be what was going on in quantum mechanics. In 1952 David Bohm did however manage to construct a somewhat elaborate model based on hidden variables that gave the same results as ordinary quantum mechanics—though involved infinitely fast propagation of information. In the early 1960s John Bell then showed that in any hidden variables theory of a certain general type there are specific inequalities that combinations of probabilities must satisfy (see page 1064). And by the early 1980s experiments had shown that such inequalities were indeed violated in practice—so that there were in fact correlations of the kind suggested by quantum mechanics. At first these just seemed like isolated esoteric effects, but by the mid-1990s they were being codified in the field of quantum information theory, and led to constructions with names like quantum cryptography and quantum teleportation.

Particularly when viewed in terms of path integrals the standard formalism of quantum theory tends to suggest that quantum systems somehow do more computation in their evolution than classical ones. And after occasional discussion as early as the 1950s, this led by the late 1980s to extensive investigation of systems that could be viewed as quantum analogs of idealized computers. In the mid-1990s efficient procedures for integer factoring and a few other problems were suggested for such systems, and by the late 1990s small experiments on these were beginning to be done in various types of physical systems. But it is becoming increasingly unclear just how the idealizations in the underlying model really work, and to what extent quantum mechanics is actually in the end even required—as opposed, say, just to classical wave phenomena. (See page 1147.)

Partly as a result of discussions about measurement there began to be questions in the 1980s about whether ordinary quantum mechanics can describe systems containing very large numbers of particles. Experiments in the 1980s and 1990s on such phenomena as macroscopic superposition and Bose-Einstein condensation nevertheless showed that standard quantum effects still occur with trillions of atoms. But inevitably the kinds of general phenomena that I discuss in this book will also occur—leading to all sorts of behavior that at least cannot readily be foreseen just from the basic rules of quantum mechanics.

▪ **Quantum effects.** Over the years, many suggested effects have been thought to be characteristic of quantum systems:

- Basic quantization (1913): mechanical properties of particles in effectively bounded systems are discrete;
- Wave-particle duality (1923): objects like electrons and photons can be described as either waves or particles;
- Spin (1925): particles can have intrinsic angular momentum even if they are of zero size;
- Non-commuting measurements (1926): one can get different results doing measurements in different orders;
- Complex amplitudes (1926): processes are described by complex probability amplitudes;
- Probabilism (1926): outcomes are random, though probabilities for them can be computed;
- Amplitude superposition (1926): there is a linear superposition principle for probability amplitudes;
- State superposition (1926): quantum systems can occur in superpositions of measurable states;

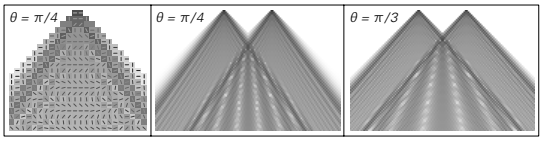
- Exclusion principle (1926): amplitudes cancel for fermions like electrons to go in the same state;
- Interference (1927): probability amplitudes for particles can interfere, potentially destructively;
- Uncertainty principle (1927): quantities like position and momenta have related measurement uncertainties;
- Hilbert space (1927): states of systems are represented by vectors of amplitudes rather than individual variables;
- Field quantization (1927): only discrete numbers of any particular kind of particle can in effect ever exist;
- Quantum tunnelling (1928): particles have amplitudes to go where no classical motion would take them;
- Virtual particles (1932): particles can occur for short times without their usual energy-momentum relation;
- Spinors (1930s): fermions show rotational invariance under SU(2) rather than SO(3);
- Entanglement (1935): separated parts of a system often inevitably behave in irreducibly correlated ways;
- Quantum logic (1936): relations between events do not follow ordinary laws of logic;
- Path integrals (1941): probabilities for behavior are obtained by summing contributions from many paths;
- Imaginary time (1947): statistical mechanics is like quantum mechanics in imaginary time;
- Vacuum fluctuations (1948): there are continual random field fluctuations even in the vacuum;
- Aharanov-Bohm effect (1959): magnetic fields can affect particles even in regions where they have zero strength;
- Bell's inequalities (1964): correlations between events can be larger than in any ordinary probabilistic system;
- Anomalies (1969): virtual particles can have effects that violate the original symmetries of a system;
- Delayed choice experiments (1978): whether particle or wave features are seen can be determined after an event;
- Quantum computing (1980s): there is the potential for fundamental parallelism in computations.

All of these effects are implied by the standard mathematical formalism of quantum theory. But it has never been entirely clear which of them are in a sense true defining features of quantum phenomena, and which are somehow just details. It does not help that most of the effects—at least individually—can be reproduced by mechanisms that seem to have little to do with the usual structure of quantum theory. So for example there will tend to be quantization whenever the underlying elements of a system are discrete. Similarly, features like the uncertainty principle and path integrals tend to be seen whenever things like waves are involved. And probabilistic effects can arise from any of the mechanisms for randomness discussed in Chapter 7. Complex amplitudes can be thought of just as vector quantities. And it is straightforward to set up rules that will for example reproduce the detailed evolution of amplitudes according say to the Schrödinger equation (see note below). It is somewhat more difficult to set up a system in which such amplitudes will somehow directly determine probabilities. And indeed in recent times consequences of this—such as violations of Bell's inequalities—are what have probably most often been quoted as the most unique features of quantum systems. It is however notable that the vast majority of traditional applications of quantum theory do not seem to have anything to do with such effects. And in fact I do not consider it at all clear just what is really essential about them, and what is in the end just a consequence of the extreme limits that seem to need to be taken to get explicit versions of them.

▪ **Reproducing quantum phenomena.** Given molecular dynamics it is much easier to see how to reproduce fluid mechanics than rigid-body mechanics—since to get rigid bodies with only a few degrees of freedom requires taking all sorts of limits of correlations between underlying molecules. And I strongly suspect that given a discrete underlying model of the type I discuss here it will similarly be much easier to reproduce quantum field theory than ordinary quantum mechanics. And indeed even with traditional formalism, it is usually difficult to see how quantum mechanics can be obtained as a limit of quantum field theory. (Classical limits are slightly easier: they tend to be associated with stationary features or caustics that occur at large quantum numbers—or coherent states that represent eigenstates of raising or particle creation operators. Note that the exclusion principle makes classical limits for fermions difficult—but crucial for the stability of bulk matter.)

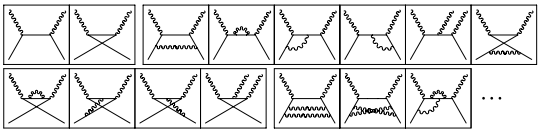
▪ **Discrete quantum mechanics.** While there are many issues in finding a complete underlying discrete model for quantum phenomena, it is quite straightforward to set up continuous cellular automata whose limiting behavior reproduces the evolution of probability amplitudes in standard quantum mechanics. One starts by assigning a continuous complex number value to each cell. Then given the list of such values the crucial constraint imposed by the standard formalism of quantum mechanics is unitarity: that the quantity $Tr[Abs[list]^2]$ representing total probability should be conserved. This is in a sense analogous to conservation of total density in diffusion processes. From

the discussion of page 1024 one can reproduce the 1D diffusion equation with a continuous block cellular automaton in which the new value of each block is given by $\{\{1-\xi, \xi\}, \{\xi, 1-\xi\}\} . \{a_1, a_2\}$. So in the case of quantum mechanics one can consider having each new block be given by $\{\{\text{Cos}[\theta], i\,\text{Sin}[\theta]\}, \{i\,\text{Sin}[\theta], \text{Cos}[\theta]\}\} . \{a_1, a_2\}$. The pictures below show examples of behavior obtained with this rule. (Gray levels represent magnitude for each cell, and arrows phase.) And it turns out that in suitable limits one generally gets essentially the behavior expected from either the Dirac or Klein-Gordon equations for relativistic particles, or the Schrödinger equation for non-relativistic particles. (Versions of this were noticed by Richard Feynman in the 1940s in connection with his development of path integrals, and were pointed out again several times in the 1980s and 1990s.)



One might hope to be able to get an ordinary cellular automaton with a limited set of possible values by choosing a suitable θ. But in fact in non-trivial cases most of the cells generated at each step end up having distinct values. One can generalize the setup to more dimensions or to allow $n \times n$ matrices that are elements of SU(n). Such matrices can be viewed in the context of ordinary quantum formalism as S matrices for elementary evolution events—and can in general represent interactions. (Note that all rules based on matrices are additive, reflecting the usual assumption of linearity at the level of amplitudes in quantum mechanics. Non-additive unitary rules can also be found. The analog of an external potential can be introduced by progressively changing values of certain cells at each step. Despite their basic setup the systems discussed here are not direct analogs of standard quantum spin systems, since these normally have local Hamiltonians and non-local evolution functions, while the systems here have local evolution functions but seem always to require non-local Hamiltonians.)

▪ **Page 540 · Feynman diagrams.** The pictures below show a typical set of Feynman diagrams used to do calculations in QED—in this case for so-called Compton scattering of a photon by an electron. The straight lines in the diagrams represent electrons; the wavy ones photons. At some level each diagram can be thought of as representing a process in which an electron and photon come in from the left, interact in some way, then go out to the right. The incoming and outgoing lines correspond to real particles that propagate to infinity. The lines inside each diagram correspond to virtual particles that in effect propagate only a limited distance, and have a distribution of energy-momentum and polarization properties that can differ from real particles. (Exchanges of virtual photons can be thought of as producing familiar electromagnetic forces; exchanges of virtual electrons as yielding an analog of covalent forces in chemistry.)



To work out the total probability for a process from Feynman diagrams, what one does is to find the expression corresponding to each diagram, then one adds these up, and squares the result. The first two blocks of pictures above show all the diagrams for Compton scattering that involve 2 or 3 photons—and contribute through order α^3. Since for QED $\alpha \simeq 1/137$, one might expect that this would give quite an accurate result—and indeed experiments suggest that it does. But the number of diagrams grows rapidly with order, and in fact the k^{th} order term can be about $(-1)^k \alpha^k (k/2)!$, yielding a series that formally diverges. In simpler examples where exact results are known, however, the first few terms typically still seem to give numerically accurate results for small α. (The high-order terms often seem to be associated with asymptotic series for things like $\text{Exp}[-1/\alpha]$.)

The most extensive calculation made so far in QED is for the magnetic moment of the electron. Ignoring parts that depend on particle masses the result (derived in successive orders from 1, 1, 7, 72, 891 diagrams) is

$$2\,(1 + \alpha/(2\,\pi) + (3\,\text{Zeta}[3]/4 - 1/2\,\pi^2\,\text{Log}[2] + \pi^2/12 + 197/144)\,(\alpha/\pi)^2 + (83/72\,\pi^2\,\text{Zeta}[3] - 215\,\text{Zeta}[5]/24 - 239\,\pi^4/2160 + 139\,\text{Zeta}[3]/18 + 25/18\,(24\,\text{PolyLog}[4, 1/2] + \text{Log}[2]^4 - \pi^2\,\text{Log}[2]^2) - 298/9\,\pi^2\,\text{Log}[2] + 17101\,\pi^2/810 + 28259/5184)\,(\alpha/\pi)^3 - 1.4\,(\alpha/\pi)^4 + \ldots)$$

or roughly

$$2. + 0.32\,\alpha - 0.067\,\alpha^2 + 0.076\,\alpha^3 - 0.029\,\alpha^4 + \ldots$$

The comparative simplicity of the symbolic forms here (which might get still simpler in terms of suitable generalized polylogarithm functions) may be a hint that methods much more efficient than explicit Feynman diagram evaluation could be used. But it seems likely that there would be limits to this, and that in the end QED will exhibit the kind of computational irreducibility that I discuss in Chapter 12.

Feynman diagrams in QCD work at the formal level very much like those in QED—except that there are usually many more of them, and their numerical results tend to be larger,

with expansion parameters often effectively being $\alpha\,\pi$ rather than α/π. For processes with large characteristic momentum transfers in which the effective α in QCD is small, remarkably accurate results are obtained with first and perhaps second-order Feynman diagrams. But as soon as the effective α becomes larger, Feynman diagrams as such rapidly seem to stop being useful.

■ **Quantum field theory.** In standard approaches to quantum field theory one tends to think of particles as some kind of small perturbations in a field. Normally for calculations these perturbations are on their own taken to be plane waves of definite frequency, and indeed in many ways they are direct analogs of waves in classical field theories like those of electromagnetism or fluid mechanics. To investigate collisions between particles, one thus looks at what happens with multiple waves. In a system described by linear equations, there is always a simple superposition principle, and waves just pass through each other unchanged. But what in effect leads to non-trivial interactions between particles is the presence of nonlinearities. If these are small enough then it makes sense to do a perturbation expansion in which one approximates field configurations in terms of a succession of arrangements of ordinary waves—as in Feynman diagrams. But just as one cannot expect to capture fully turbulent fluid flow in terms of a few simple waves, so in general as soon as there is substantial nonlinearity it will no longer be sufficient just to do perturbation expansions. And indeed for example in QCD there are presumably many cases in which it is necessary to look at something closer to actual complete field configurations—and correlations in them.

The way the path integral for a quantum field theory works, each possible configuration of the field is in effect taken to make a contribution $Exp[i\,s/\hbar]$, where s is the so-called action for the field configuration (given by the integral of the Lagrangian density—essentially a modified energy density), and $\hbar$ is a basic scale factor for quantum effects (Planck's constant divided by $2\,\pi$). In most places in the space of all possible field configurations, the value of s will vary quite quickly between nearby configurations. And assuming this variation is somehow random, the contributions of these nearby configurations will tend to cancel out. But inevitably there will be some places in the space where s is stationary (has zero variational derivative) with respect to changes in fields. And in some approximation the field configurations in these places can be expected to dominate the path integral. But it turns out that these field configurations are exactly the ones that satisfy the partial differential equations for the classical version of the field theory. (This is analogous to what happens for example in classical diffraction theory, where there is an analog of the path integral—with $\hbar$ replaced by inverse frequency—whose stationary points correspond through the so-called eikonal approximation to rays in geometrical optics.) In cases like QED and QCD the most obvious solutions to the classical equations are ones in which all fields are zero. And indeed standard perturbation theory is based on starting from these and then looking at the expansion of $Exp[i\,s/\hbar]$ in powers of the coupling constant. But while this works for QED, it is only adequate for QCD in situations where the effective coupling is small. And indeed in other situations it seems likely that there will be all sorts of other solutions to the classical equations that become important. But apart from a few special cases with high symmetry, remarkably little is known about solutions to the classical equations even for pure gluon fields. No doubt the analog of turbulence can occur, and certainly there is sensitive dependence on initial conditions (even non-Abelian plane waves involve iterated maps that show this). Presumably much like in fluids there are various coherent structures such as color flux tubes and glueballs. But I doubt that states involving organized arrangements of these are common. And in general when there is strong coupling the path integral will potentially be dominated by large numbers of configurations not close to classical solutions.

In studying quantum field theories it has been common to consider effectively replacing time coordinates t by $i\,t$ to go from ordinary Minkowski space to Euclidean space (see page 1043). But while there is no problem in doing this at a formal mathematical level—and indeed the expressions one gets from Feynman diagrams can always be analytically continued in this way—what general correspondence there is for actual physical processes is far from clear. Formally continuing to Euclidean space makes path integrals easier to define with traditional mathematics, and gives them weights of the form $Exp[-\beta\,s]$—analogous to constant temperature systems in statistical mechanics. Discretizing yields lattice gauge theories with energy functions involving for example $Cos[\theta_i - \theta_j]$ for color directions at adjacent sites. And Monte Carlo studies of such theories suggest all sorts of complex behavior, often similar in outline from what appears to occur in the corresponding classical field theories. (It seems conceivable that asymptotic freedom could lead to an analog of damping at small scales roughly like viscosity in turbulent fluids.)

One of the apparent implications of QCD is the confinement of quarks and gluons inside color-neutral hadrons. And at some level this is presumably a reflection of the fact that QCD forces get stronger rather than weaker with increasing distance. The beginnings of this are visible in perturbation

theory in the increase of the effective coupling with distance associated with asymptotic freedom. (In QED effective couplings decrease slightly with distance because fields get screened by virtual electron-positron pairs. The same happens with virtual quarks in QCD, but a larger effect is virtual gluon pairs whose color magnetic moments line up with a color field and serve to increase it.) At larger distances something like color flux tubes that act like elastic strings may form. But no detailed way to get confinement with purely classical gluon fields is known. In the quantum case, a sign of confinement would be exponential decrease with spacetime area of the average phase of color flux through so-called Wilson loops—and this is achieved if there is in a sense maximal randomness in field configurations. (Note that it is not inconceivable that the formal problem of whether quarks and gluons can ever escape to infinity starting from some given class of field configurations may in general be undecidable.)

▪ **Vacuum fluctuations.** As an analog of the uncertainty principle, one of the implications of the basic formalism of quantum theory is that an ordinary quantum field can in a sense never maintain precisely zero value, but must always show certain fluctuations—even in what one considers the vacuum. And in terms of Feynman diagrams the way this happens is by virtual particle-antiparticle pairs of all types and all energy-momenta continually forming and annihilating at all points in the vacuum. Insofar as such vacuum fluctuations are always exactly the same, however, they presumably cannot be detected. (In the formalism of quantum field theory, they are usually removed by so-called normal ordering. But without this every mode of any quantum system will show a zero-point energy $\hbar\omega/2$—positive in sign for bosons and negative for fermions, cancelling for perfect supersymmetry. Quite what gravitational effects such zero-point energy might have has never been clear.) If one somehow changes the space in which a vacuum exists, there can be directly observable effects of vacuum fluctuations. An example is the 1948 Casimir effect—in which the absence of low-energy (long wavelength) virtual particle pairs in the space between two metal plates (but not in the infinite space outside) leads to a small but measurable force of attraction between them. The different detailed patterns of modes of different fields in different spaces can lead to very different effective vacuum energies—often negative. And at least with the idealization of impermeable classical conducting boundaries one predicts (based on work of mine from 1981) the peculiar effect that closed cycles can be set up that systematically extract energy from vacuum fluctuations in a photon field.

If one has moving boundaries it turns out that vacuum fluctuations can in effect be viewed as producing real particles. And as known since the 1960s, the same is true for expanding universes. What happens in essence is that the modes of fields in different background spacetime structures differ to the point where zero-point excitations seem like actual particle excitations to a detector or observer calibrated to fields in ordinary fixed flat infinite spacetime. And in fact just uniform acceleration turns out to make detectors register real particles in a vacuum—in this case with a thermal spectrum at a temperature proportional to the acceleration. (Uniform rotation also leads to real particles, but apparently with a different spectrum.) As expected from the equivalence principle, a uniform gravitational field should produce the same effect. (Uniform electric fields lead in a formally similar way to production of charged particles.) And as pointed out by Stephen Hawking in 1974, black holes should also generate thermal radiation (at a temperature $\hbar c^3/(8\pi G k M)$). A common interpretation is that the radiated particles are somehow ones left behind when the other particle in a virtual pair goes inside the event horizon. (A similar explanation can be given for uniform acceleration—for which there is also an event horizon.) There has been much discussion of the idea that Hawking radiation somehow shows pure quantum states spontaneously turning into mixed ones, more or less as in quantum measurements. But presumably this is just a reflection of the idealization involved in looking at quantum fields in a fixed background classical spacetime. And indeed work in string theory in the mid-1990s may suggest ways in which quantum gravity configurations of black hole surfaces could maintain the information needed for the whole system to act as a pure state.

▪ **Page 542 · Quantum measurement.** The basic mathematical formalism used in standard quantum theory to describe pure quantum processes deals just with vectors of probability amplitudes. Yet our everyday experience of the physical world is that we observe definite things to happen. And the way this is normally captured is by saying that when an observation is made the vector of amplitudes is somehow replaced by its projection s into a subspace corresponding to the outcome seen—with the probability of getting the outcome being taken to be determined by $s \cdot \text{Conjugate}[s]$.

At the level of pure quantum processes, the standard rules of quantum theory say that amplitudes should be added as complex numbers—with the result that they can for example potentially cancel each other, and generally lead to wave-like interference phenomena. But after an observation is made, it is in effect assumed that a system can be described by ordinary real-number probabilities—so that for example no

interference is possible. (At a formal level, results of pure quantum processes are termed pure quantum states, and are characterized by vectors of probability amplitudes; results of all possible observations are termed mixed states, and are in effect represented as mixtures of pure states.)

Ever since the 1930s there have been questions about just what should count as an observation. To explain everyday experience, conscious perception presumably always must. But it was not clear whether the operation of inanimate measuring devices of various kinds also should. And a major apparent problem was that if everything—including the measuring device—is supposed to be treated as part of the same quantum system, then all of it must follow the rules for pure quantum processes, which do not explicitly include any reduction of the kind supposed to occur in observations.

One approach to getting around this suggested in the late 1950s is the many-worlds interpretation (see page 1035): that there is in a sense a universal pure quantum process that involves all possible outcomes for every conceivable observation, and that represents the tree of all possible threads of history—but that in a particular thread, involving a particular sequence of tree branches, and representing a particular thread of experience for us, there is in effect a reduction in the pure quantum process at each branch point. Similar schemes have been popular in quantum cosmology since the early 1990s in connection with studying wave functions for the complete universe.

A quite different—and I think much more fruitful—approach is to consider analyzing actual potential measurement processes in the context of ordinary quantum mechanics. For even if one takes these processes to be pure quantum ones, what I believe is that in almost all cases appropriate idealized limits of them will reproduce what are in effect the usual rules for observations in quantum theory. A key point is that for one to consider something a reasonable measurement it must in a sense yield a definitive result. And in the context of standard quantum theory this means that somehow all the probability amplitudes associated with the measuring device must in effect be concentrated in specific outcomes—with no significant interference between different outcomes.

If one has just a few quantum particles—governed say by an appropriate Schrödinger equation—then presumably there can be no such concentration. But with a sufficiently large number of particles—and appropriate interactions—one expects that there can be. At first this might seem impossible. For the basic rules for pure quantum processes are entirely reversible (unitary). So one might think that if the evolution of a system leads to concentration of amplitudes, then it should equally well lead to the reverse. But the crucial point is that while this may in principle be possible, it may essentially never happen in practice—just like classical reversible systems essentially never show behavior that goes against the Second Law of thermodynamics. As suggested by the main text, the details in the quantum measurement case are slightly more complicated—since to represent multiple outcomes measuring devices typically have to have the analogs of multiple equilibrium states. But the basic phenomena are ultimately very similar—and both are in effect based on the presence of microscopic randomness. (In a quantum system the randomness serves to give collections of complex numbers whose average is essentially always zero.)

This so-called decoherence approach was discussed in the 1930s, and finally began to become popular in the 1980s. But to make it work there needs to be some source of appropriate randomness. And almost without exception what has been assumed is that this must come through the first mechanism discussed in Chapter 7: that there is somehow randomness present in the environment that always gets into the system one is looking at. Various different specific mechanisms for this have been suggested, including ones based on ambient low-frequency photons, background quantum vacuum fluctuations and background spacetime metric fluctuations. (A somewhat related proposal involves quantum gravity effects in which irreversibility is assumed to be generated through analogs of the black hole processes mentioned in the previous note.) And indeed in recent practical experiments where simple pure quantum states have carefully been set up, they seem to be destroyed by randomness from the environment on timescales of at most perhaps microseconds. But this does not mean that in more complicated systems more characteristic of real measuring devices there may not be other sources of randomness that end up dominating.

One might imagine that a possibility would be the second mechanism for randomness from Chapter 7, based on ideas of chaos theory. For certainly in the standard formalism, quantum probability amplitudes are taken to be continuous quantities in which an arbitrary number of digits can be specified. But at least for a single particle, the Schrödinger equation is in all ways linear, and so it cannot support any kind of real sensitivity to initial conditions, or even to parameters. But when many particles are involved the situation can presumably be different, as it definitely can be in quantum field theory (see page 1061).

I suspect, however, that in fact the most important source of randomness in most cases will instead be the phenomenon of intrinsic randomness generation that I first discovered in systems like the rule 30 cellular automaton. Just like in so

many other areas, the emphasis on traditional mathematical methods has meant that for the most part fundamental studies have been made only on quantum systems that in the end turn out to have fairly simple behavior. Yet even within the standard formalism of quantum theory there are actually no doubt many closed systems that intrinsically manage to produce complex and seemingly random behavior even with very simple parameters and initial conditions. And in fact some clear signs of this were already present in studies of so-called quantum chaos in the 1980s—although most of the specific cases actually considered involved time-independent constraint satisfaction, not explicit time evolution. Curiously, what the Principle of Computational Equivalence suggests is that when quantum systems intrinsically produce apparent randomness they will in the end typically be capable of doing computations just as sophisticated as any other system—and in particular just as sophisticated as would be involved in conscious perception.

As a practical matter, mechanisms like intrinsic randomness generation presumably allow systems involving macroscopic numbers of particles to yield behavior in which interference becomes astronomically unlikely. But to reproduce the kind of exact reduction of probability amplitudes that is implied by the standard formalism of quantum theory inevitably requires taking the limit of an infinite system. Yet the Principle of Computational Equivalence suggests that the results of such a limit will typically be non-computable. (Using quantum field theory to represent infinite numbers of particles presumably cannot help; after appropriate analysis of the fairly sophisticated continuous mathematics involved, exactly the same computational issues should arise.)

It is often assumed that quantum systems should somehow easily be able to generate perfect randomness. But any sequence of bits one extracts must be deduced from a corresponding sequence of measurements. And certainly in practice—as mentioned on pages 303 and 970—correlations in the internal states of measuring devices between successive measurements will tend to lead to deviations from randomness. Whatever generates randomness and brings measuring devices back to equilibrium will eventually damp out such correlations. But insofar as measuring devices must formally involve infinite numbers of particles this process will formally require infinitely many steps. So this means that in effect an infinite computation is actually being done to generate each new bit. But with this amount of computation there are many ways to generate random bits. And in fact an infinite computation could even in principle produce algorithmic randomness (see page 1067) of the kind that is implicitly suggested by the traditional continuous mathematical formalism of quantum theory. So what this suggests is that there may in the end be no clear way to tell whether randomness is coming from an underlying quantum process that is being measured, or from the actual process of measurement. And indeed when it comes to more realistic finite measuring devices I would not be surprised if most of the supposed quantum randomness they measure is actually more properly attributed to intrinsic randomness generation associated with their internal mechanisms.

■ **Page 543 · Bell's inequalities.** In classical physics one can set up light waves that are linearly polarized with any given orientation. And if these hit polarizing ("anti-glare") filters whose orientation is off by an angle θ, then the waves transmitted will have intensity $Cos[\theta]^2$. In quantum theory the quantization of particle spin implies that any photon hitting a polarizing filter will always either just go through or be absorbed—so that in effect its spin measured relative to the orientation of the polarizer is either +1 or -1. A variety of atomic and other processes give pairs of photons that are forced to have total spin 0. And in what is essentially the Einstein-Podolsky-Rosen setup mentioned on page 1058 one can ask what happens if such photons are made to hit polarizers whose orientations differ by angle θ. In ordinary quantum theory, a straightforward calculation implies that the expected value of the product of the two measured spin values will be $-Cos[\theta]$. But now imagine instead that when each photon is produced it is assigned some "hidden variable" ϕ that in effect explicitly specifies the angle of its polarization. Then assume that a polarizer oriented at $0°$ will measure the spin of such a photon to have value $f[\phi]$ for some fixed function f. Now the expected value of the product of the two measured spin values is found just by averaging over ϕ as

```
Integrate[f[ϕ] f[θ - ϕ], {ϕ, 0, 2 π}]/(2 π)
```

A version of Bell's inequalities is then that this integral can decrease with θ no faster than $\theta/(2\pi) - 1$—as achieved when $f = Sign$. (In 3D ϕ must be extended to a sphere, but the same final result holds.) Yet as mentioned on page 1058, actual experiments show that in fact the decrease with θ is more rapid—and is instead consistent with the quantum theory result $-Cos[\theta]$. So what this means is that there is in a sense more correlation between measurements made on separated photons than can apparently be explained by the individual photons carrying any kind of explicit hidden property. (In the standard formalism of quantum theory this is normally explained by saying that the two photons can only meaningfully be considered as part of a single "entangled" state. Note that because of the probabilistic nature of the

correlations it turns out to be impossible to use them to do anything that would normally be considered communicating information faster than the speed of light.)

A basic assumption in deriving Bell's inequalities is that the choice of polarizer angle for measuring one photon is not affected by the choice of angle for the other. And indeed experiments have been done which try to enforce this by choosing the angles for the polarizers only just before the photons reach them—and too close in time for a light signal to get from one to the other. Such experiments again show violations of Bell's inequalities. But inevitably the actually devices that work out choices of polarizer angles must be in causal contact as part of setting up the experiment. And although it seems contrived, it is thus at least conceivable that with a realistic model for their time evolution such devices could end up operating in just such a way as to yield observed violations of Bell's inequalities.

Another way to get violations of Bell's inequalities is to allow explicit instantaneous propagation of information. But traditional models involving for example a background quantum potential again seem quite contrived, and difficult to generalize to relativistic cases. The approach I discuss in the main text is quite different, in effect using the idea that in a network model of space there can be direct connections between particles that do not in a sense ever have to go through ordinary intermediate points in space.

When set up for pairs of particles, Bell's inequalities tend just to provide numerical constraints on probabilities. But for triples of particles, it was noticed in the late 1980s that they can give constraints that force probabilities to be 0 or 1, implying that with the assumptions made, certain configurations of measurement results are simply impossible.

In quantum field theory the whole concept of measurement is much less developed than in quantum mechanics—not least because in field theory it is much more difficult to factor out subsystems, and so to avoid having to give explicit descriptions of measuring devices. But at least in axiomatic quantum field theory it is typically assumed that one can somehow measure expectation values of any suitably smeared product of field operators. (It is possible that these could be reconstructed from combinations of idealized scattering experiments). And to get a kind of analog of Bell's inequalities one can look at correlations defined by such expectation values for field operators at spacelike-separated points (too close in time for light signals to get from one to another). And it then turns out that even in the vacuum state the vacuum fluctuations that are present show nonzero such correlations—an analog of ordinary quantum mechanical entanglement. (In a non-interacting approximation these correlations turn out to be as large as is mathematically possible, but fall off exponentially outside the light cone, with exponents determined by the smallest particle mass or the measurement resolution.) In a sense, however, the presence of such correlations is just a reflection of the idealized way in which the vacuum state is set up—with each field mode determined all at once for the whole system.