Testing nature to the limit: the Large Hadron Collider

The Large Hadron Collider is intended to probe to the very heart of nature. As it restarts after its teething problems, **Glen Cowan** explains how this huge experiment will depend crucially on statistics to interpret what it finds.

Great fleas have little fleas upon their backs to bite 'em,
And little fleas have lesser fleas, and so *ad infinitum*.
(Augustus de Morgan, 1806–1871)

Over the last couple of centuries the concept of an elementary particle has evolved. Atoms are no longer indivisible: they are made up of nuclei and electrons. Nuclei can be split into protons and neutrons; protons and neutrons in turn are made up of quarks. The sizes of the supposed ultimately-small particles have become smaller and the mathematical theories governing their behaviour more intriguing. The quest to learn more about what particles exist and how they interact has taken place in a variety of settings. Many of the most dramatic contributions have come from particle accelerators. These huge devices produce beams of elementary particles moving close to the speed of light. They bring them into collision and allow scientists to study what takes place.

The newest of these accelerators is actually the largest scientific instrument ever built. The Large Hadron Collider (LHC) is a circular ring 27 km in circumference. It is buried 100 m underground at the European Organisation for Nuclear Research (CERN) near Geneva, Switzerland. Its function is summed up by its name. It is large; it takes beams of protons (an elementary particle that comes under a general classification known as "hadrons") and makes them collide. The debris may contain particles that are as yet undiscovered or unconfirmed.

Higgs bosons, dark matter and a billion events a second: the biggest scientific instrument ever needs the help of statistics

Shortly after its switch-on in September 2008 the LHC suffered an electrical fault that has set the project back by more than a year. Repairs of the machine have been underway since then and everything appears to be on track to restart operation towards the end of 2009.

The basic setup at the LHC can be seen in Figure 1. Beams of protons travel around the huge circumference ring in opposite directions, guided along the circular path by superconducting magnets. The energy of an individual proton will reach up to 7 TeV (tera or10^7 electron-volts), seven times higher than the current record holder, an accelerator called the Tevatron near Chicago.

At four places along the LHC ring the protons are brought into head-on collisions and around these points experimenters have constructed large particle detectors. Two of these are so-called "general purpose detectors", which go by the names ATLAS and CMS. The other two, called LHCb and ALICE, have more specialised goals related to studying particles called B-hadrons and heavy ions respectively. The detectors each represent collaborative efforts by several thousand physicists from hundreds of universities and research centres around the world.

In a proton–proton collision, referred to as an "event", hundreds of other particles can be produced, which emerge in different directions from the collision point. This is a classic example of the conversion of energy into mass, as described by Einstein's famous equation: $E = mc^2$. Many of the phenomena that nature may have in store could involve particles with high masses—

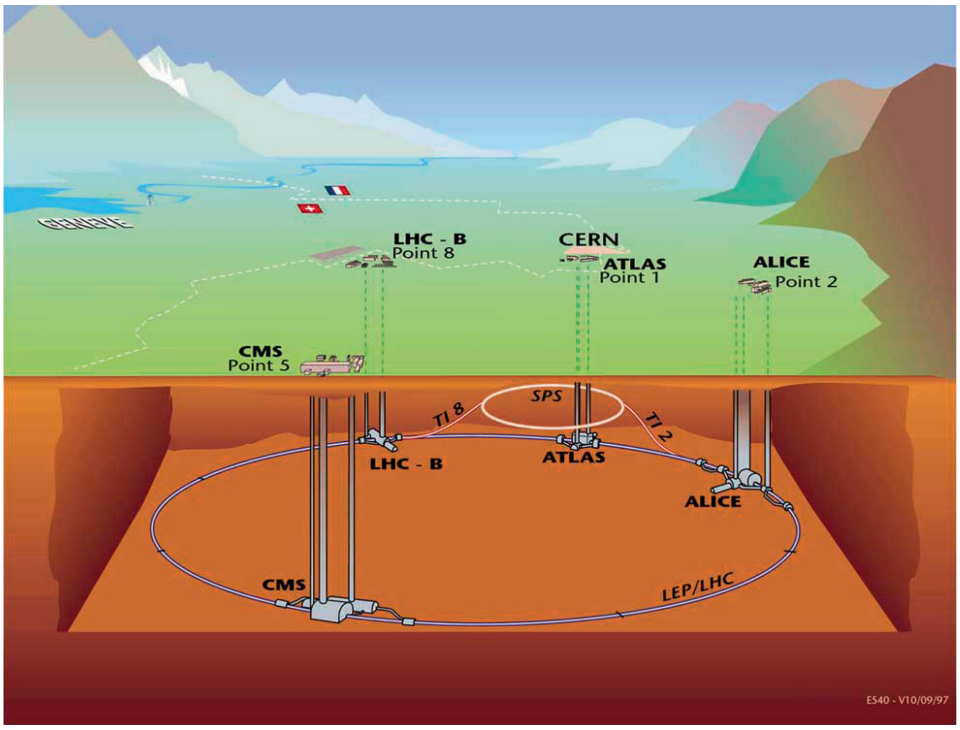


Figure 1. Overall view of the LHC experiments

hence the need for the high energies, and vast size, of the LHC. Nothing less would be able to produce them and study their properties.

A computer simulation of a collision is shown in Figure 2, which represents a cut-away view of the ATLAS detector, 46 m long and 25 m in diameter. The innermost components are designed to track the trajectories of electrically charged particles, which leave deposits of ionisation as they fly through the detector. Outside these are devices called calorimeters, which can detect both electrically charged and also neutral particles such as photons (particles of light) by absorbing their energy. Particles called muons, similar to but heavier than electrons, can be distinguished by their almost unique ability to penetrate the calorimeters and leave signals in a very outer layer of particle detectors. From a single event, one can measure the energy and direction of many hundreds of particles together with some information on their type.

The Standard Model and beyond

Our best theory of elementary particles describes all matter as consisting of quarks and leptons with six types, or "flavours", of each, as shown in Figure 3. For each of these particles there exists a so-called antiparticle. For example, the proton (a hydrogen nucleus) is composed of two "up" quarks and one "down" quark. Leptons and quarks interact by exchanging other particles called gauge bosons, of which there are four types, the photon being an important example. The behaviour of these basic ingredients is described by a mathematical theory that took shape in the 1970s and is now known as "The Standard Model". Like all quantum mechanical theories, it cannot predict what will happen on an event-by-event basis, but gives the probabilities for different outcomes.

The Standard Model contains, in its current form, 25 adjustable parameters, which include the masses of the quarks, leptons and gauge bosons, and other constants that describe the strengths of particle interactions. Most of these have been measured accurately so that the model can make predictions for a wide variety of observable phenomena.

An unconfirmed piece of the theory is the Higgs boson, a particle for which there is, as yet, no direct experimental evidence. Without the Higgs, however, the mathematical consistency of the theory runs into serious trouble as soon as one considers non-zero masses for the other particles. Most of these masses have been accurately estimated (and are non-zero), so most physicists believe that the Higgs boson, or something like it, must exist. The mass of the Higgs itself is indirectly constrained to lie in a range roughly between 100 and 200 GeV, and it is by virtue of such a relatively high mass that it would be produced only very rarely and thus could elude discovery. (The "GeV" or giga-electron-volt is a unit of energy, strictly

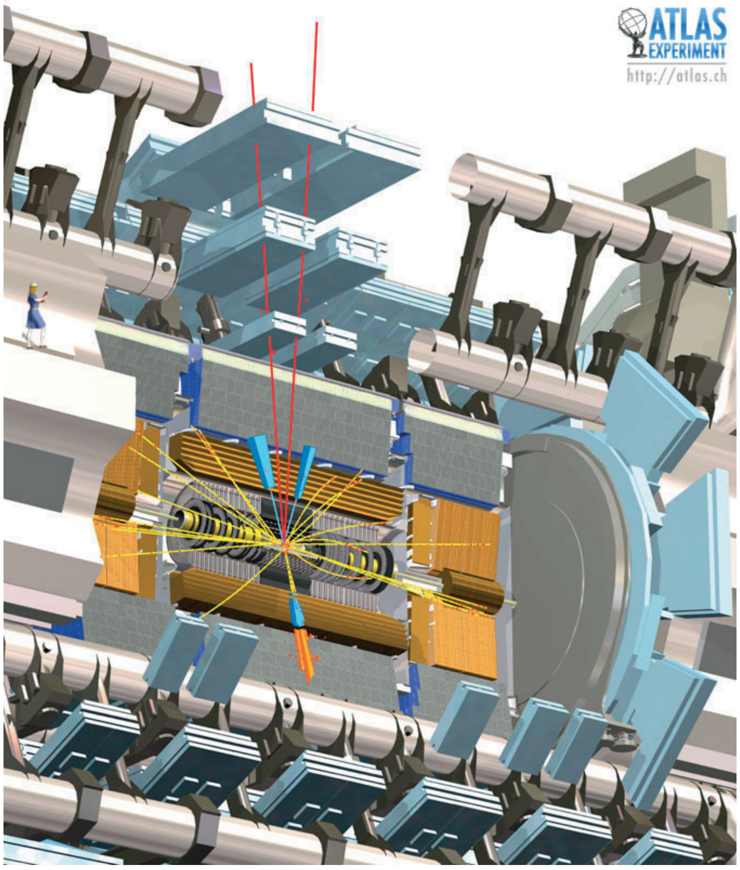


Figure 2. A computer simulation of a Higgs boson collision in the ATLAS detector of the Large Hadron Collider. Photograph: Joao Pequenao, courtesy of CERN

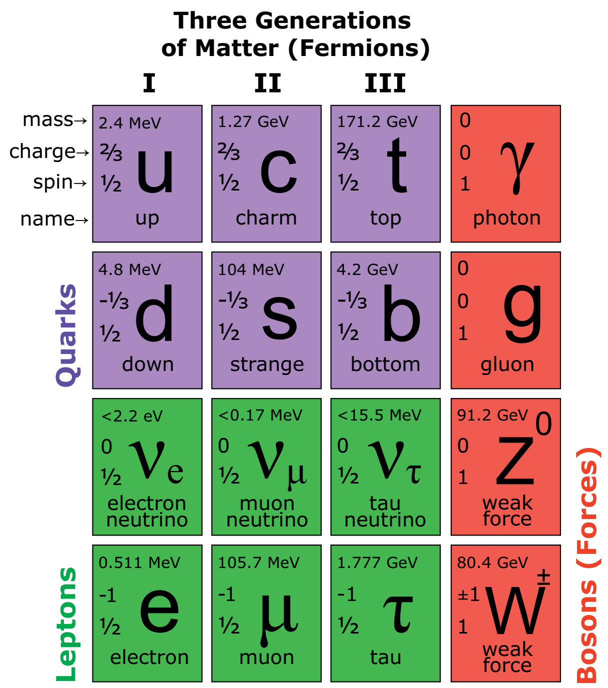


Figure 3. The Standard Model of particle physics

speaking, but is often used to quantify a particle's mass by exploiting the relation $E = mc^2$. On this scale, the mass of a hydrogen atom is somewhat less than 1 GeV.)

A key goal of the LHC is to establish whether the Higgs boson actually exists and, if so, to measure its properties.

For many years the predictions of the Standard Model have agreed extremely well with essentially all measurements. The few cases where one sees marginally significant discrepancies may be hints of new phenomena, but may also reflect fluctuations or systematic uncertainties that are not fully understood. Nevertheless, we have good reasons to believe that the Standard Model, even including its elusive Higgs boson, cannot be a complete description of particle interactions. A number of hints indicate that nature is described by some deeper theory and that this should reduce to the Standard Model when considering processes at low enough energy. By studying higher energy particle collisions at a machine such as the LHC we hope to find direct evidence for whatever more fundamental theory lies beyond.

Many extensions to the Standard Model have been proposed, including those with additional particle types or where space has more than the usual number of three dimensions. An important type of alternative hypothesis comes under the general name of supersymmetry or "SUSY". This is a class of theory where for every known type of particle there exists a new partner particle, which should have a different angular momentum or "spin". The super-partners also apparently have high masses, which would explain why none have been seen in lower-energy accelerators.

Supersymmetric theories can involve more than a hundred adjustable parameters beyond those of the Standard Model, although specific SUSY models may have less than a half a dozen. Now already this seems to introduce a lot of additional complexity and perhaps Occam's razor should warn us away from such speculation, but in fact there are a number of important reasons to believe that supersymmetry, or something like it, could represent a true description of nature.

Supersymmetry can help solve a theoretical mystery as to why the relevant energy scales for elementary particles cover such an enormously broad range, from the mass of the Higgs at around 10^2 GeV up to the scale where we believe gravitational interactions should be involved—more than 10^{19} GeV. One type of SUSY model also predicts the existence of a particle called the "neutralino", which could be as massive as a heavy nucleus, but which would have an almost negligibly weak interaction with other particles.

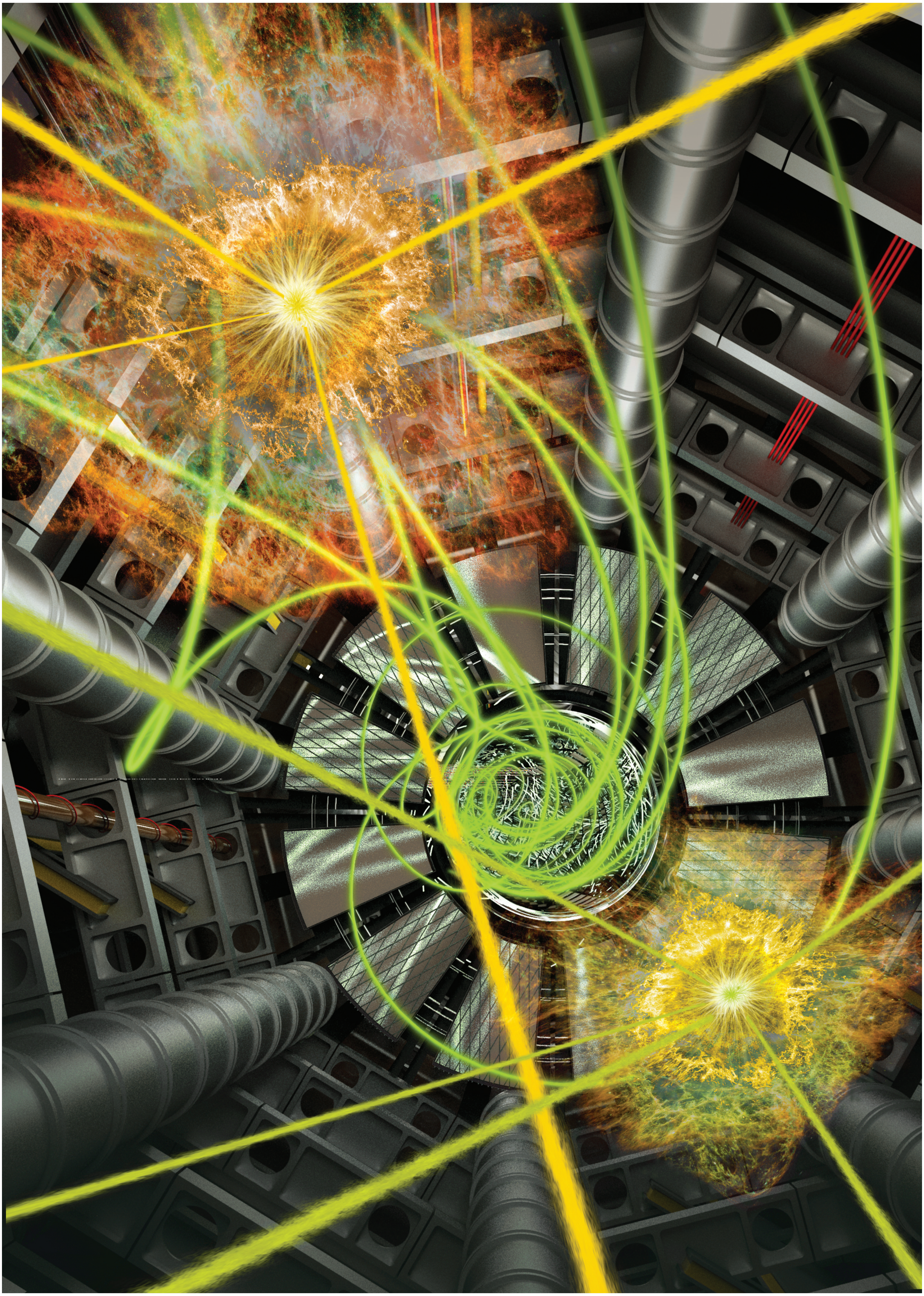
If such a SUSY model is correct, then large numbers of neutralinos should have been produced in the ultra-hot universe just after the "Big Bang". As the universe expanded and cooled, these neutralinos would form a sort of background gas, attracted only by gravity. In this way the neutralino could provide an explanation for "dark matter", matter whose existence is seen only through its influence on other gravitating bodies such as galaxies. Astronomical evidence for dark matter has been gathering for many years and, in the form of neutralinos or otherwise, it has become a key ingredient in cosmological models. So the discovery of a neutralino at the LHC would have a major impact on cosmology and provide a spectacular link between the science of nature's smallest particles and that of the largest structures in the universe.

Statistics for particles

Physics is a fundamentally mathematical science and its most important theoretical ingredient, quantum mechanics, is based on probability. One might then suppose that the analysis of data in particle physics would involve rigorous statistical methods. In fact, many of the important discoveries have been so clear cut that a simple "error analysis" has been sufficient to make the case. In recent years, however, with the rising costs of experiments—the LHC programme will cost around 10 billion US dollars—it has become important to extract the maximum possible information from the data and so advanced statistical methods have become increasingly relevant.

When running at full intensity, the LHC should produce close to a billion events per second. After a quick sifting, the data from around 200 per second are recorded for further study, resulting in more than a billion events per year. But only a tiny fraction of these are of potential interest. If one of the speculative theories such as SUSY turns out to be realised in nature, then this will result in a subset of events having characteristic features and these "SUSY events" will simply be mixed in randomly with a much larger number of Standard Model events.

An event's relevant distinguishing features depend on what new physics nature chooses to reveal, but one might see particles of a certain type emitted at large angles relative to the beamline, or "jets" containing many particles moving in roughly the same direction. For example, a SUSY process might lead to an event containing two neutralinos as well as a host of other particles. The neutralinos would simply sail through the detector without a trace. The total momentum of all of the particles must still balance, however, and so if one were to see jets carrying a large amount of energy on one side of the detector not compensated by a corresponding amount on the other, then this would be characteristic of escaped neutralinos, or "missing energy".



Unfortunately, Standard Model processes can often mimic these features and one will not be able to say with certainty that a given event shows clear evidence for something new such as SUSY. For example, even Standard Model events can contain very light particles called neutrinos, which also escape undetected. However, the typical amount and pattern of missing energy in these events differs, on average, from what a SUSY event would give and so a statistical analysis can be applied to test whether something besides Standard Model events are present.

In a typical analysis there is a class of events we are interested in finding—the signal—and these, if they exist at all, are mixed in with the rest of the events—the background. The data for each event is some collection of numbers, $x = (x_1, ..., x_n)$, representing particle energies, momenta, etc. And the probabilities are joint densities for x given the signal (s) or background (b) hypotheses: $f(x|s)$ and $f(x|b)$.

The use of a statistical test to distinguish between two classes of events (signal and background) comes up in different ways. Sometimes both event classes are known to exist, and the goal is to select one class (signal) for further study. For example, proton–proton collisions leading to the production of top quarks are a well-established process. By selecting these events one can carry out precise measurements of the top quark's properties such as its mass. In other cases, the signal process could represent an extension to the Standard Model, say, SUSY, whose existence is not yet established, and the goal of the analysis is to see if one can do this. Rejecting the Standard Model with a sufficiently high significance level amounts to discovering something new and, of course, one hopes that the newly revealed phenomena will provide important insights into how nature behaves.

What the physicist would of course like to have is a test with maximal power with respect to a broad class of alternative hypotheses. For a given signal model, for example, one would like to choose the acceptance and rejection regions based on the likelihood ratio $f(x|s)/f(x|b)$. In principle the signal and background theories should allow us to work out the required functions $f(x|s)$ and $f(x|b)$ but in practice the calculations are too difficult and we do not have explicit formulas for these.

What we have instead of $f(x|s)$ and $f(x|b)$ are complicated Monte Carlo programs: that is, we can sample x to produce simulated signal and background events. Because of the multivariate nature of the data, where x may contain at least several or perhaps even hundreds of components, it is a non-trivial problem to construct a test with a power approaching that of the likelihood ratio.

Often physicists begin by making simple "cuts", that is, signal and background are separated using a set of rectangular decision boundaries in the space of the input variables x. Here one can at least exercise some physical intuition as to where the cuts should be placed, but the resulting statistical test cannot possibly exploit all of the information available in the data.

Another possible boundary for separating signal and background events could be a hyperplane in the n-dimensional space of measured variables. Such linear classifiers have been used for many years and since the 1990s methods from machine learning and neural computing that allow for non-linear boundaries have become increasingly popular. The artificial neural network has long been a standard tool but, more recently, classifiers such as boosted decision trees and support vector machines have emerged on the scene.

Rejecting the Standard Model with a sufficiently high significance level amounts to a new discovery about nature

A search for a particular signal could identify a region in the space of measured variables where one expects to see as many signal and as few background events as possible. If the data reveal a number of events in excess of the expected background, this leads one to believe that something new has been discovered. The significance of the observed signal is often quantified by a p-value taken as the probability, assuming only background events are present, of finding as many events as actually found or more.

Traditionally the significance is translated into the equivalent number of standard deviations that would lead to the same p-value for a one-sided fluctuation of a Gaussian random variable. Common practice has been to regard a 5-standard-deviation effect as sufficient to announce a discovery but, of course, the actual degree of belief that a new phenomenon has been found will depend on many other factors, especially the plausibility of the signal and one's trust in the modelling of the background.

The models for signal and background processes involve not only predicting what comes out of the proton–proton collisions but also the complex, and by no means perfect, response of the detector. The models can be improved by including more free parameters whose values are estimated from the data but this then results in a degradation of the significance of a potential discovery. It is precisely in this area of accurate model building that most of the effort in a statistical analysis is invested.

In the case where the number of events found is compatible with the expected background from Standard Model processes, one can try to see which alternatives can be excluded. Usually this amounts to placing limits on the parameters of the proposed models, which often take the form of lower limits on the masses of the particles involved. For example, if the neutralino exists but has a mass greater than 47 GeV, then it would be so heavy that we would not be sensitive to it. For smaller masses, theory and data would be deemed incompatible (here based on a p-value below a threshold of 0.05).

The use of Bayesian methods in particle physics appears to be on the increase, but perhaps not as quickly as in many other fields. The hesitation can usually be traced to technical, philosophical or sociological difficulties in the assignment of prior probabilities to models about which there is no clear consensus. Some important applications of Bayesian methods can be found in attempts to constrain model parameters using non-informative or reference prior probabilities. In other cases one may have prior information based not on measurements but on purely theoretical considerations, and one can attempt to bring this into the analysis in the Bayesian framework. Computational issues have also been a major hurdle but advances in computing, especially Markov chain Monte Carlo methods, are now allowing Bayesian methods to enter the standard toolbox of a particle physics analysis.

We have many reasons to be optimistic that new and exciting discoveries will emerge from the LHC. Of course this depends on many factors, most crucially on whether nature chooses to place new phenomena within our reach. In any case it will take a tremendous performance from the accelerator, detectors and analysts to understand the enormous data sample that will soon emerge. The discovery phase of the project is, we hope, about to begin.

Glen Cowan is a physicist at Royal Holloway, University of London, and is developing statistical methods for particle physics, including applications of Bayesian statistics to measurements that will be made at the LHC. He is a member of its ATLAS Experiment team.