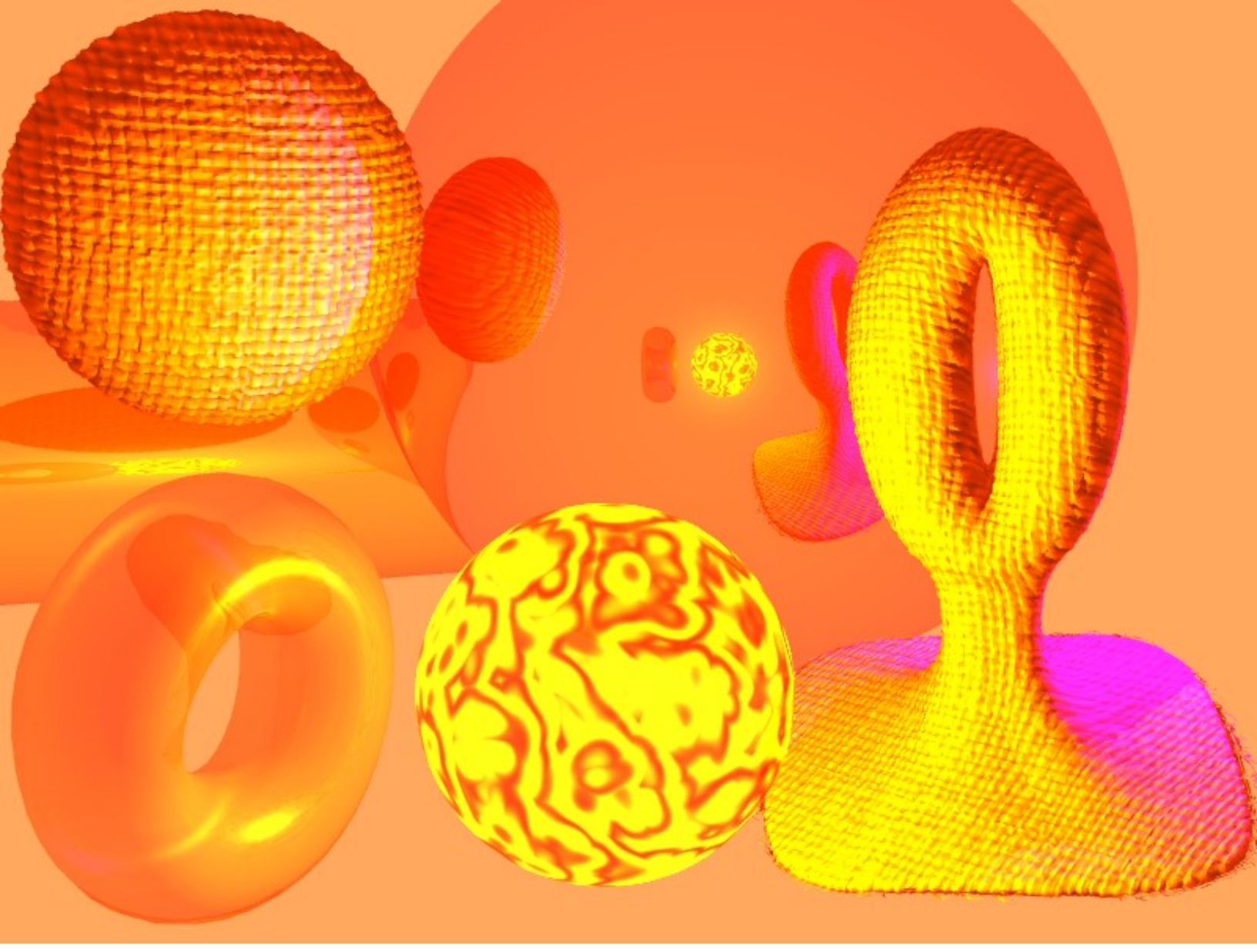

WARREN SIEGEL

Particles, Strings, & Other Things

Warren Siegel

Particles, Strings, & Other Things

Di Renzo Editore

Table of Contents

Biography	1
The life and mind of a physicist	11
Modern physics	20
Particle physics	36
Infinities	48
Superstring theory	53
The future of string theory?	63
Glossary	68

Biography

Early years

I was born in the USA in 1952 to emigrants from a town that would be destroyed by the Nazis in what is now Ukraine; here they owned and ran a small grocery store. As a child, I was immediately interested in learning. I started reading at age 3. At first, I was most interested in chemistry, because I thought atoms were the most fundamental objects. But by age 8 I learned about nuclei and particles, so I became interested in physics. (Earlier I thought physics was about just levers and pulleys, and I had no interest in becoming a carpenter.) Although I had been exposed to a layman's view of science, it was only at that age that I also learned how to do long division in school. The (non-math) teacher was not skilled enough to explain the procedure to anyone else in my class, but it was enough to spark my interest in mathematics. I studied more arithmetic on my own, and in a couple of years my math teacher had me explaining to her and her other classes tricks I had learned on how to add long lists of numbers more efficiently. Soon I learned how to take square roots longhand, and how to make slide rules using logarithms. Somehow my teachers were always most impressed with my math skills, but I was more interested in physics. Unfortunately, the amount of science taught in grade school was far less than the amount of math (including "New Math"), so physics had to wait.

The 1960's were a period of great cultural change, particularly in the US, in many ways --- music, fashion, civil rights, the peace movement, etc.

The space program showed what a dedicated, large-scale effort could achieve toward a specific goal in science. It also inspired the public's interest in (and thus funding for) science. Overall, the 60's produced an aura of optimism and progress, in contrast to the postwar depression of the 50's, in spite of the fact that the Cold War continued. This accelerated progress in society probably reinforced my interest in the future of society and science. Unfortunately, many things kind of stagnated (or even regressed) in the following decades; shortly after man landed on the Moon, the manned space program was canceled, as a scapegoat for those afraid to cut the military budget. But by then my course was pretty much set.

This attitude toward science was reflected in the science fiction of the period, unlike the predominant nuclear apocalyptic fiction of the 50's. Undoubtedly some of my optimism was a matter of interpretation: I remember an Italian science fiction movie dubbed into English ("*Il Pianeta degli Uomini Spenti*", or "*Battle of the Worlds*" in the English version), where at the end the protagonists pitied a "mad" scientist who stayed behind on a doomed planet just to learn the secrets of the universe, while I thought they must be fools to miss such an opportunity in order to go back to their mundane lives. On the other hand, the TV show *Star Trek* portrayed a positive future and a positive view of science, and separated science from war.

In high school we didn't get physics until our senior year. Before then all the physics I got I picked up on my own, from laymen's books, my older sisters' college textbooks (for non-scientists), and public libraries. I read a lot of the magazines *Scientific American* and *Science Digest*. I also subscribed to the journal *Science*, which automatically made me a member of the American Association for the Advancement of Science. At the public library, and once at the physics library of the local campus of the University of Michigan (where no one questioned my access), I also read journals like *Physics Today* and *The Astrophysical Journal*, but they were a bit above my head. I didn't

find my senior physics class as interesting as I'd hoped, as it had no "modern" physics, but at least I finally learned vectors.

I was born and grew up in Flint, Michigan. When I left there to go to college in 1970, it was the second largest city in Michigan, based entirely on the automotive industry. For the same reason, it's now dropped to half that size, and has the cheapest real estate of any city in the country (except maybe those recently hit by hurricanes). Besides, after finishing drivers' education, I realized I didn't like cars: More people are killed in car crashes than in wars. Around the same time, I became a vegetarian, which also didn't seem to fit well into the Midwestern life style. So I went to Berkeley (the only college to which I applied), as much for the cultural differences as for the education (and the weather was nicer, too).

The cultural change from Flint to Berkeley was quite refreshing: For example, at the time Flint was rather racist. Although my high school was well integrated (it was mostly black), and a lot of people from different ethnic groups were friendly, there was always a bit of tension. When Detroit had its big riot, Flint had its own smaller version (although I was actually visiting Detroit at the time). On a later Martin Luther King Day, kids were allowed to leave school for a ceremony downtown. Some from another school came to ours for general hooliganism; I happened to be in a classroom directly above the main entrance, which was unfortunately just where the police chose to release tear gas. In contrast, one of the first things I saw coming onto the Berkeley campus was a black male and a white female student walking hand-in-hand, which would probably not have been considered so romantic in Flint. There were many other political differences, e.g., I was only one of two male students at my high school who wore long hair, but in college that was common. Berkeley also went a bit further: It had two local political parties --- liberal, and radical. The San Francisco Bay Area was also very cosmopolitan; e.g., there was a wide variety of vegetarian restaurants --- Chi-

nese, Nigerian, Buddhist, etc. (In contrast, once when I gave a talk in Texas and was invited to dinner by a physics professor, I told him I was a vegetarian. After some consultation with fellow faculty, he suggested a steak house.) The cultural breadth included science, with museums such as the Exploratorium (founded the year before I arrived) and the Lawrence Hall of Science (established the year before that); in Flint the best one could do was the planetarium. And cars seemed less important; the subway began operation a couple of years after I arrived, with the main Berkeley station a block from my apartment.

The University never responded to my first letter asking for application forms, and by the time they answered my second, the College of Letters and Science (which included the Physics Department) had already filled its quota, so I got accepted into the College of Chemistry, and later transferred. I attended the University of California for both my undergraduate and graduate education. I left home the day after my high-school graduation ceremony, and started summer sessions a few days later, taking freshman courses in physics and math. This allowed me to start sophomore physics and math in the fall. Taking a little more than the usual course load, plus more summer sessions, I managed to complete a double major in physics and math (even though I had no intention of making math my profession) in 2 1/2 years. By the end of that period I had already taken first-quarter graduate courses in physics and math.

I was originally also interested in philosophy, but one course I took included logical positivism, which made it clear that religion and philosophy describe nothing more than their proponents' personalities (often entertaining in themselves), although the philosophy teachers seemed not to appreciate that fact. Ironically, this gave me a greater appreciation of those they rejected by vacuous arguments, like Zeno of Citium, Thoreau, and Ayer. Philosophy is a relic of early attempts to understand the world: "Metaphysics" has been

replaced by true physics, "epistemology" by scientific method, and "ethics" by psychology and sociology. ("Science" was known as "natural philosophy" as recently as the 19th century.)

Research years

The time I started graduate school was exactly the same time that theorists' interest returned to "quantum field theory". Unfortunately, I was located in the place that was considered the center for the opposite point of view. So, even though I wrote the first paper on what was much later to be called "D-branes" (a problem suggested by my advisor, Marty Halpern), nobody else really cared about strings anymore: They had gone back to either field theory, or the more phenomenological side of "S-matrix theory". So, due to the bad timing of history, what was perhaps the most prestigious school for theoretical high-energy physics when I entered was probably not so influential by the time I left. I received no offers of postdoctoral positions, so I went to Harvard as a "freebie", an act which was unusual and almost always unsuccessful. The Cambridge/Boston area was the East Coast analog of the Berkeley/San Francisco Bay Area, both being cosmopolitan, and more European in style. (In "Humbead's Revised Map of the World", they are represented as contiguous, and form most of the world.)

By the time I arrived in Cambridge in 1977, I had decided to start research on supersymmetry and supergravity. (I had briefly considered "instantons" as an alternative, but I could see that area was rapidly dying, the last resort of a program to look at extended solutions of classical field equations in increasing dimensions.) Unfortunately, it seemed that all the people working on supersymmetry in that locale (and the rest of the USA) had decided to spend the year at CERN in Europe. One exception was Jim Gates, but although we had similar interests, it was many months before we worked to-

gether because, although his position was at Harvard, his soul was at his alma mater MIT, and it took a while to pin him down. Before that happened, I wrote a few papers at Harvard. So the people from the Boston area at CERN wondered who I was, because they hadn't seen me there, while the people still at Harvard wondered who I was, because they didn't work in supersymmetry. Furthermore, because of the pecking order at Harvard, it took months for my papers to get typed (with frequent and long interruptions); but they also shipped (as in "boat") preprints abroad, which took further months. (On at least one occasion I purposely sent out a preprint with typographical errors because the delay in having them corrected would have been too great, and a previous attempt at correction had only resulted in new errors; the publishers actually did a better job of corrections than the local secretaries.) So, by the time other people in supersymmetry saw my results, some had already been independently rederived.

There were also delays in publication because the journals rejected all my earliest papers: The method I had used to derive my results, "superspace", was much more efficient than earlier methods usually used for supergravity. But the referees demanded I translate my results into the older language, even though I had already proven equivalence. The translation was almost as difficult as working in the older formalism in the first place, taking months to translate in collaboration with Jim Gates what took weeks to derive by myself, so most of those results did not appear till over a year later. (Some appeared much earlier, because I translated them into another, published superspace formalism that had obtained incomplete results.) Nowadays dissemination of information is much quicker, using electronic methods for both writing and communicating results, long before any journal can publish or even evaluate them.

Refereeing is a random process: Referees are often not only incapable of evaluating a paper, but also incapable of realizing it. (Dishonesty may also

be involved.) Furthermore, editors are sometimes unqualified to make a judgment on the matter. Almost every scientist has had papers rejected for absurd reasons: Once a referee said he knew my result was wrong, even though he was unable to find an error, or provide any evidence to that effect, by his own argument or from published sources. On another occasion a referee provided a list of "reasons" why my paper was wrong, each of which I refuted, in response to which he provided a new list, which I again proved incorrect, in response to which he provided yet another list; the editor then refused my paper on the grounds that it had been rejected too many times. So, I submitted the paper to another journal, who gave it to the same referee, who sent his first list again. I sent back to the editor all that referee's lists, along with all my refutations, and told him I was not interested in publishing in a journal that used such dishonest referees. (Eventually my paper was published in yet another journal, which managed to find another referee.) On the other hand, I once rejected a paper I refereed, on the grounds that results obtained in one section of the paper directly contradicted results obtained in another section; the editor responded that he would accept the paper anyway because it was "controversial".

Jim and I wrote more papers together, and the people who left for the year returned, including Marc Grisaru and Martin Roček, and we wrote many papers together. After a total of 1 1/2 years at Harvard, I got an unusual one-semester postdoc at nearby Brandeis, followed by a job at the Institute for Advanced Study in Princeton. During the first year of work at IAS, I was offered a postdoc at Caltech. Ordinarily I wouldn't have been interested, but there was no interest in supersymmetry in the Princeton area, and I was told that Jim and Martin had also been made postdoctoral offers, and Marc a visiting professorship offer. I guessed they would go if I did, so I accepted. (I was aware that Julius Wess, one of the discoverers of supersymmetry, was coming to visit Princeton, but I didn't think much would come of that. How-

ever, when Ed Witten came at the same time, he became interested in supersymmetry.)

The weather and politics of southern California make it seem like a different country from northern California: The smog in Pasadena was so bad that it was part of the daily weather report through the summer; during a second-stage smog alert it literally hurt to take a deep breath outdoors during the middle of the day. When I first arrived there via the Los Angeles airport, the plane descended through an orange cloud; I lived in Pasadena for several days before I realized there were mountains only a few miles away, as they were totally camouflaged by bluish-white smog. The environment was decidedly unfriendly to pedestrians (probably from unfriendliness toward ethnic groups most likely to lack cars): They actually ticketed jaywalkers there, and once when Jim and I went for a walk a few blocks south of our offices, into the adjacent suburb of San Marino, we were stopped and questioned by two squad cars and a helicopter.

In contrast, Caltech itself was a nice place to work. Every office had a computer terminal, a condition ahead of its time. This allowed me to learn more about computers, like how to transmit a character code directly to a terminal in another office that would completely lock it up, thus spreading frustration and confusion. On the more serious side, it also allowed us to typeset our own papers, and in particular provided the perfect environment for Jim, Marc, Martin, and myself to start the book *Superspace*. We also wrote many papers on supersymmetry, along with other collaborators. After two years there, I was a total of five years out of my Ph.D., so I applied for assistant professorships. After getting no offers, I took up the IAS on their offer, made when I left there, to complete the second year of my original postdoc there, but curiously was told that I would have to re-apply for it. So, I accepted a three-year postdoc position at Berkeley. During my first few months at Berkeley I continued working on the book *Superspace*. I briefly

rejoined my collaborators in Caltech to finish the book. It took a little longer than expected, so they had to return to their respective institutions, leaving me to tie up loose ends and send it off to the publisher. We went to dinner at a Chinese restaurant the night before they left; my fortune cookie said I was working too hard. Soon after returning to Berkeley, I visited Japantown in San Francisco. In a bookstore there I found a book titled *Superspace*; it was a book of paintings of spaceships and the like. Taking that as a good time to switch topics, I returned to string theory. (In parallel with the book *Superspace*, I wrote, with some help from my colleagues, the joke paper *Stuperspace*, which was later published as part of the proceedings of a conference, and is also available at my website. It was the first of a long series.)

By the end of my stay at Berkeley, it was 8 years since I got my Ph.D., by which time almost anyone else would have long-since finished being a postdoc one way or another. In the meantime, Jim had taken a tenure-track position at Maryland, and Martin at Stony Brook. Jim made me an offer to join him as an assistant professor, so off I went. At Maryland I worked mostly on string theory, and wrote the book *Introduction to String Field Theory*. In two years I was made full professor. Shortly after that, for the first time in my career, I received offers from more than one place at the same time. I chose to come to Stony Brook in 1988, mostly because of their larger theory group, where I have been ever since. Here I wrote *Fields*, a textbook on quantum field theory for second-year graduates that I use for teaching courses on field theory, relativity, and string theory; it's been available for free on the Web since I first wrote it, as now are the other two books.

You can find more information at my web site:

http://insti.physics.sunysb.edu/~siegel/plan.html

It also has some physics summaries, and some parodies on physics and other things.

The life and mind of a physicist

Research

You might think you can understand physics better if you can understand physicists. Nothing could be further from the truth. Many people think scientists are some strange lot, totally devoted to science, which permeates their way of life. Some even come to the conclusion that this very fact limits their view of the world. I, on the other hand, thought this would make scientists a very interesting group of people to meet and with whom to associate. So I was very disappointed when I learned that scientists are almost indistinguishable from anyone else, except for their line of work. The most one could say is that scientists must be well educated, so the statistical distribution of their opinions on politics, philosophy, food, sports, weather, etc., will closely resemble that of almost any group of Ph.D.'s; unfortunately, I don't find enough substance in any of those topics to make them interesting for more than 5 minutes. This means that your typical physicist will be neither much more nor less interesting than an average person when discussing any topic other than physics. Of course, you might find exceptional physicists who have something unique to say about something other than physics, but they are about as common as such exceptional people outside of physics.

For many physicists, physics is mostly just a job. So they work hard for success, as measured by pay, position, or awards. Generally it's a job they like, or else they wouldn't have worked so hard to get it. But for some it's just what was expected of them, so they do it until they fail and are forced

into another line of work, or have some mid-life crisis that leads them into another field, like religion or business. Conversely, some physicists sort of stumble into physics, after majoring in other areas as undergraduates.

For many other physicists, physics is an art. They are willing to suffer failure, if it means they can discover what they perceive to be the beauty of nature. Many of these are diverted into other areas if they find that nature does not agree with their tastes. In science it is always important to keep an open mind, and not fool oneself into beliefs, which are merely assumptions one wants to be true. At each step one needs to think about why one is doing what one is doing.

Much of the success of scientists can be attributed to extensive training, hard work, and good luck. A lot of it is also politics: For example, a physicist needs to travel around the world giving speeches ("talks", "seminars", "colloquia") about their papers to other physicists, who can't be bothered to read them. There is also a tendency toward conformism; this is generally due as much to insufficient individuality as to outside pressure. But all these things apply to most professions.

Physicists are also expected to be very smart. Intelligence is an important factor in science: Science cannot develop without new ideas. But while the average physicist might be a bit smarter than the average person, certainly very few physicists should be considered geniuses (assuming one can define what a "genius" is).

More importantly, it is far from clear how well intelligence is applied. For example, many physicists smoke, in spite of the fact that it has long been known that smoking ruins one's health and shortens one's life-span. (In the US alone, more than half a million deaths per year can be attributed to smoking, dwarfing terrorism by several orders of magnitude.) One might then ar-

gue that other factors (childhood "training", orders from your DNA, etc.) are dominating intelligence, forcing people to do something that doesn't seem very smart. But if intelligence can take a back seat even in matters of life and death, what guarantees that it is in full control in one's career? Yet surprisingly, many physicists are perfectly rational and analytic in their work, while in other matters they seem to shut off that side of their brains, as if they lived in a dualistic world where the same laws of reason did not apply. Often they are incapable of even giving a rational explanation of why they are physicists, or why they chose the area of physics in which they work, and may actually resist providing such a reason when prompted for one. Fortunately, science itself is nothing like scientists in that respect.

Science has a built-in system of checks and balances. Unlike human behavior, which is regulated by laws that are ultimately set by popular consent (whether by vote under a democracy, or coup under a dictatorship), the laws of science are set by nature. So eventually one can be confident that science is certain, and not merely a point of view; true knowledge is by definition objective. However, every scientific law develops from a hypothesis, or "educated guess". Generally, there are a lot of guesses, until enough evidence accumulates to sort them out. Finding a new law of physics is a lot like solving a murder mystery, only there are no motives to consider, and the case involves a serial killer, who continues his *modus operandi* even after he is apprehended.

Often there is some question as to the status of science, or of scientific development. This is generally confused with the psychological status of scientists themselves. As in other human endeavors, there are conservatives and liberals (those who are more positive about the past or the future), and pessimists and optimists (those who are more negative or positive about the present). Anything unproven is necessarily controversial; anything proven is treated as obvious. As in court, in science it is necessary for both sides of a

question to be adequately represented, but often the opposing views are presented with more emotion than reason. Scientists as a whole act as the jury, but there is often no judge to preside over the inquiry.

Teaching

Being a scientist generally involves not only research, but also teaching. (However, there is some truth in the George Bernard Shaw saying, "He who can, does. He who cannot, teaches.") Unfortunately, most of the style of teaching, as reflected in textbooks, is to simply tag new material onto the end of the old. While this is usually OK for topics like history, it is often contradictory to the principles of science. Even the scientific notation of some courses is outdated: Freshman physics is usually taught in about the same way as it was in the 19th century, so it might as well be written in Latin. By the time one reaches graduate physics, the way one learned introductory physics looks quite quaint. More importantly, many principles are understood much more simply in modern terms than when they were originally discovered, but the baggage of concepts that proved less general or efficient has not been discarded. Sometimes the excuse is used that such an approach gives insight into the way ancient physicists thought, but hypocritically, ancient wrong theories are not discussed in detail, and time spent discussing outdated modes of thought is time lost for learning modern concepts. For example, special relativity is a simple idea mathematically, and requires no calculus, so can be taught to college freshman, or even to high school seniors. Furthermore, it is generally considered by non-physicists to be one of the most interesting subjects in physics. But often it is lost somewhere, and might not even be taught in physics courses for non-scientists.

There are several reasons for this situation, due not to the nature of science but to the nature of scientists, who are not so different from average

people. Most scientists are still busy doing research, and want to spend the minimum required time in teaching, and so simply follow the available textbooks, whose authors spent the minimum effort in updating them (if at all) from previously available ones. A closely related reason is that using a more modern textbook requires taking the time to learn a new way of teaching, rather than teaching the way the teacher learned it, even though it may have been a generation earlier. Ironically, "You can't teach an old dog new tricks," often applies to the teachers themselves. To a great extent, this attitude is reflected in research: The research areas and methods of most physicists differ little from those of their student days.

Because of a dissatisfaction with many physics textbooks in this regard, I have written one of my own on topics required as prerequisites to my kind of research, *Fields*. (I also authored and coauthored a couple of more-advanced books at the research level.) Unlike most textbooks, I released this one for free over the World Wide Web. It's available at my web site, and also at arXiv.org, the "e-print" archive for physics, mathematics, computer science, and quantitative biology.

ArXiv.org was first introduced in 1991 (under another name) as a way to distribute research papers before publication ("preprints"), originally for high-energy physics, but now beyond physics. Its popularity gradually but steadily increased to becoming the nearly universal way to distribute new research results. Before then, preprints had already become more important than "official" publications, but their distribution depended on "snail" mail, which was slower, more expensive, and less efficient, requiring people to sort through piles of paper to find something of interest. And one also had to do the same with journals anyway, in case some article had not been distributed as a preprint. Since then, the amount of paper (or at least its rate of increase) taking up space in physics research libraries has dramatically decreased, not

only because preprints in paper form have essentially disappeared, but because physics journals have become an almost unnecessary expense.

However, the idea has not yet quite caught on that physics can not only be *distributed* electronically, but *read* the same way. So many physicists still download preprints from arXiv.org, only to print them out on paper and throw them onto a pile in the corner (or sometimes, every part) of their offices, perhaps to never read them, or to download them again when they realize that's easier than trying to find them in a pile of paper. For those people, the printer has replaced the photocopier. They have not yet learned how to use their computers to magnify text to make it easier to read, to instantly search for text rather than slowly flipping through pages, to use bookmarks in PDF files, etc. Ironically, most of these people already use computers to *write* papers. On a related note, many of them still insist on having letters of recommendation for jobs printed for mailing, rather than being transmitted by email (even though secretaries do the work).

So it is not totally surprising that few teachers are yet ready to teach electronically. The burden is born mostly by the students, who are not offered the option to replace shelves of textbooks with a laptop computer.

What it takes to become a physicist

On the other hand, it isn't too difficult to find people who have learned how to live in an electronic world, but don't understand physics: It doesn't take a degree in physics to learn how to send email or use the web. Most of these people fall into 3 categories:

(1) people who don't know physics, but care about it, and want to learn,

(2) people who don't know and don't care, and

(3) people who don't know, and would like to get rid of it. (Note that the borders between these categories are not sharp, and many people show all these characteristics to varying degrees.)

The last category is sometimes known as "quacks". A quack has seen enough physics to know that he doesn't like it, and would like to replace it with his own beliefs, which may or may not form part of an organized religion. He denies established facts as lies. He will never admit he has made a mistake; if any fault is found with him, he will accuse his accuser of the same fault, citing falsified or irrelevant evidence. He pretends to know how to use "real" physics, with arguments so transparently wrong they could not possibly fool anyone who bothered to check them. Such a person can easily be distinguished from those who are merely ignorant, by a brief conversation: Any evidence presented that his arguments are wrong is immediately countered with denial, insults, self-contradiction, and new arguments that are even worse. If given enough rope to hang himself, the quack will willfully reduce himself to the object of *reductio ad absurdum*. Finally, when faced with rejection by the entire physics community (both theorists and experimentalists), being unable to publish except through his own expenses or in an unknown journal, he will compare himself to Galileo, totally unaware of the fact that he much more closely resembles those who rejected Galileo.

The second category is the "couch potatoes". Such a person is willing to learn as much physics as possible without applying any effort, i.e., none. He may accumulate trivia through various media, such as television, the web, or non-technical books, in the same way that a parrot collects phrases. But he is unable to distinguish a quack from a scientist, since even the elementary logic required to recognize a self-contradiction would require the effort of actual reasoning. Often such people are most interested in scientific topics such as black holes, the Big Bang, or dinosaurs, because they are big, scary things that might eat you.

The first category is "students", in the broadest sense of the term. This book is for them. Ultimately any student will find this book somewhat unsatisfying, since it has few technical details, but maybe it can serve as a starting point toward enlightenment.

Sometimes scientists are asked, "Why are you interested in science?" This question answers itself, since "why" is a request for knowledge, and would not have been asked if the questioner himself did not have such an interest. Once a friend asked me if I "liked to think". I didn't appreciate the question, because the only alternative was to be a dumb animal. Another friend once asked me if I "meditated". I asked him to define "meditation", but he wasn't able to distinguish it from thinking. I guess he had never meditated on that question. Unfortunately, many people never ask even such elementary questions, because they have never really thought about it.

The first thing one needs to understand to learn science is what science is; this is basically understanding the meaning of "knowledge". True knowledge is not merely the accumulation of data, but organizing it by finding relationships. This method is "inductive reasoning", which can be defined by 3 principles: consistency, generality, and simplicity. Knowledge must be both self-consistent (logical) and consistent with nature (experiment); the application of this principle is "deductive reasoning". To be more than just a recording of data, facts need to be generalized in such a way that a single fact can describe many observations, and be useful for predicting results that have not yet been observed; this process requires imagination. The simpler that such a fact can be stated, the easier it is to use and generalize; this step requires some pragmatism. The well-rounded scientist is good at all 3 of these things; in practice most scientists tend to be much better at one than the rest.

Since science is always an ongoing process, the result is never perfect. There will always be some new experiment that will disagree with established science. This does not mean that science as we know it will be thrown out, only that it will be improved to be more accurate and more general. Such improvements take time, during which the quality of the improvements will increase. The stages have been given names, "hypothesis", "theory", and "law", but the distinctions are only relative (like, e.g., "class", "order", and "family" in biological taxonomy); the actual development is more continuous, and its degree is difficult to measure.

The fact that the development of science always builds on top of established science (with some repairs along the way) means that any scientist must first learn established science (at least in his area of research) before attempting any revisions. This was not necessarily true during the Scientific Revolution, since there was little science at the time, and what little there was was infused with philosophy and religion. But today's technology is based on today's science, providing everyday proof of its veracity.

Modern physics

Before getting into some of the details of my research area, I would like to briefly review some of the many more-elementary ideas on which it is built.

Orders of magnitude

Things come in many different sizes and weights. That's not very new, but the range of sizes and weights in modern science varies from those of the (observed) Universe down to subatomic particles (and perhaps smaller). These are conveniently arranged in powers of 10 ("orders of magnitude"), to avoid long strings of zeros (before or after the decimal point) that take too long to count. Some idea of the vastness of this range is given by the following diagram:

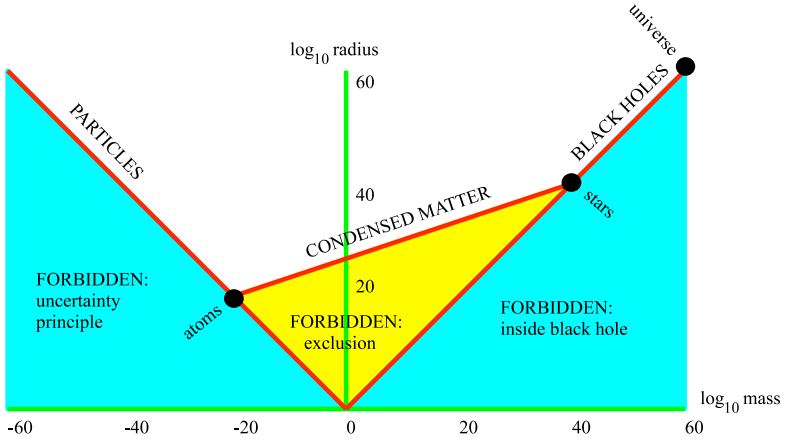

Fig. 1: *Mass-radius diagram*

The axes for the mass and radius (for spherical objects) are "logarithmic", in that only the power of 10 is given. Since the powers of 10 range in the dozens, we have rounded that power itself to multiples of 20 ($10^0 = 1$, 10^{20}, 10^{40}, ...), neglecting mere factors of 10 or 100. Besides the axes, the straight lines in the diagram correspond to certain physical criteria:

(1) The line on the left labeled "particles" indicates that an object of a given mass cannot have a radius below a certain size, because the "Uncertainty Principle" of quantum mechanics (see below) prevents the localization of a mass below a certain limit.

(2) The line on the right labeled "black holes" indicates that an object of a given mass cannot have a radius below a certain other size or else, according to General Relativity, it will collapse to form a black hole, in which case you won't be able to see into smaller than that radius anyway. (Actually, the concept of black hole can be somewhat generalized to Newton's law of gravity, where below a similar radius the pull of gravity would be so strong that light could not escape, because escape velocity would be greater than the speed of light.) Astronomical objects lie near the right-hand line, from planets and stars to the Universe.

(3) The line in the middle labeled "condensed matter" is a line of (roughly) constant density, corresponding to packing atoms together. (The mass goes as the number of atoms, which goes as the volume, which goes as the cube of the radius.) Atoms can't be packed more closely than their size allows, unless they are converted into something else, which generally causes them to shrink to sizes close to that of black holes. Almost all the things we think of as "objects" lie near these 3 lines. For example, solid (or liquid) matter lies near this middle line, packing together atoms numbering from 1 to

about 10^{60}, which makes a star. (You yourself lie roughly in the middle of that line.)

So we can get qualitative ideas about things by ignoring factors of 2 or π that might show up in detailed calculations. This often allows us to estimate numbers without knowing too much about the physics. For example, the equations for these lines follow from just knowing what physical constants appear in the topic of discussion, as well as the appropriate units for mass and length appearing there, which we have not yet defined:

(1) The line for particles comes from quantum mechanics and Special Relativity. From the former we have $\hbar$, "Planck's constant" (over 2π), the fundamental constant that defines quantum mechanics. It has dimensions of $(\text{mass})(\text{length})^2/(\text{time})$. From the latter we have c, the speed of light (in a vacuum), which has dimensions of (length)/(time). A free particle is defined by its mass m. (There is also "spin", see below; but it is measured in units of $\hbar$, and so introduces nothing new to our analysis. There are also other constants that describe the interactions of particles.) If we put these things together, the only radius we can get is $\hbar/mc$. This identification gives the "particles" line in the figure. There is a better reason for this line (see below), but we have derived it by just "dimensional analysis". In particular, we get the slope -1 on the logarithmic plot from the fact that this radius goes as the inverse of the mass.

(2) The line for black holes is obtained by replacing quantum mechanics by gravity: Newton's gravitational constant G has dimensions $(\text{length})^3/(\text{mass})(\text{time})^2$. Then the only radius coming from c, G, and m is Gm/c^2. So we get a line of slope $+1$. The origin is the location of the intersection of these 2 lines: This defines our units as "Planck units". We can then write expressions for the units of mass and length in terms of c, G, and $\hbar$, but it's more convenient to do the reverse, and simply define those 3 constants as be-

ing equal to 1 in Planck units; this also gives a unit of time. (In practice, the meter is already defined in terms of the second, by measuring the distance traveled by light in a certain time; so c is already fixed to a certain number by definition even in the metric system. Soon something similar will be done for the kilogram, so ħ's value will be a definition in the metric system.)

(3) Finally, we can get the third line by plugging in the proton mass (or mass of some atom) into the equation for the particles line. It so happens that such a mass is about 10^{-20} Planck masses. The slope is 1/3 because of constant density. The observed part of the Universe (which might be infinite) has a mass of the order of 10^{60} Planck masses, setting the borders of the graph.

In conventional units, the Planck mass is of the order of 10^{-8} kilograms, the Planck length, 10^{-35} meters, and the Planck time, 10^{-43} seconds.

Symmetry

"Symmetry" is one of the most basic concepts in science. It is a general relation between things that might have been different. Symmetry can be important even when it isn't exact, like the fact that people's bodies are almost, but not quite, the same on the left and right sides ("mirror symmetry").

One important example of an exact symmetry in nature is "rotational symmetry", the simple fact that the laws of physics describing motion in one direction are the same as those describing motion in another direction. A related symmetry is "translational symmetry", that these laws are the same in one part of the Universe as another. Both these symmetries are "continuous symmetries", in that we can rotate any direction continuously (cumulatively by arbitrarily small angles at a time) into any any other, or translate any position continuously (by arbitrarily small distances) into any other. (This con-

trasts with "discrete symmetries" like mirror symmetry, which involves reflections about some axis.)

In practice, we use Cartesian coordinates, (x,y,z), to label the position in space of any object. Translational symmetry says that any force between two objects will depend only on the difference between the positions of the two objects,

$$(\Delta x, \Delta y, \Delta z) = (x_1 - x_2, y_1 - y_2, z_1 - z_2)$$

while rotational symmetry then further says that the magnitude of any such force will depend only on the distance r between the two objects (measured by a ruler or other straight object connecting them), as given by the Pythagorean theorem and Euclidean geometry,

$$r^2 = (\Delta x)^2 + (\Delta y)^2 + (\Delta z)^2$$

The distance r is "invariant" under both translations and rotations: It is unchanged if we translate, e.g., just x as

$$x \rightarrow x + a$$

for some constant a, meaning we change the x coordinate of everything by a:

$$x_1 \rightarrow x_1 + a, \qquad x_2 \rightarrow x_2 + a$$

$$\Delta x = x_1 - x_2 \rightarrow \Delta x$$

and r is also unchanged if we rotate, e.g., x into y by some angle θ ("rotation about the z axis") as

$$x \rightarrow x \cos\theta + y \sin\theta, \qquad y \rightarrow y \cos\theta - x \sin\theta$$

$$(\Delta x)^2 + (\Delta y)^2 \rightarrow (\Delta x)^2 + (\Delta y)^2$$

Thus the way we choose our Cartesian coordinates is merely a convention; by translations and rotations we can change to other choices of Cartesian coordinates in which the laws of physics will take the same form.

Symmetries are related to conservation laws: Newton's laws of motion are more conveniently replaced with the law of conservation of momentum. But this conservation is a consequence of translational symmetry, as momentum is associated with translation. Similarly, there is a conserved "angular momentum" associated with rotational symmetry.

The Special Theory of Relativity

Another exact symmetry principle that applies to all nature is Einstein's Special Theory of Relativity. Although some areas of physics do not require it explicitly, because they deal only with speeds much smaller than the speed of light in the vacuum, even such nonrelativistic concepts are simpler in the "light" of relativity. The basic idea of relativity is most simply embodied in the concept of "Minkowski space", that space and time are part of the same structure, with distance measured in Cartesian coordinates for both space and time as

$$-s^2 = r^2 - (\Delta t)^2 = (\Delta x)^2 + (\Delta y)^2 + (\Delta z)^2 - (\Delta t)^2$$

where now we include time differences Δt. Except for the funny relative sign, this is a direct generalization of the Pythagorean theorem above.

(x,y,z,t) now form a relativistic "4-vector", a generalization of the "3-vector" (x,y,z). Here we use units where c = 1, e.g., by measuring distances in light years and times in years, or the Planck units defined above. In the relativistic case, the "proper time s" can be measured by a clock that travels between the two events in spacetime along a "straight" (constant-velocity) path. Just as nonrelativistically we have "rotational symmetry", and so are free to choose orthogonal axes for our coordinates x, y, z, but the distance r is independent of that choice, relativistically we are free to choose also the time axis, and s (but not r) remains the same. An extreme example is light, along whose path s = 0 between any two points, since its speed is $r/\Delta t = 1$: No matter how we change our time axis, by moving at a constant velocity with respect to our previous axes, s will still be measured as 0 --- the speed of light is the same in all reference frames.

Thus relativity adds two new symmetries to our list by extending space to include time: Time translations were obvious before relativity,

$$t \to t + a$$

but now we also have "Lorentz transformations", the generalizations of rotations that allow mixing of space and time, e.g.,

$$x \to x \cosh\theta + t \sinh\theta, \qquad t \to t \cosh\theta + x \sinh\theta$$

$$(\Delta x)^2 - (\Delta t)^2 \to (\Delta x)^2 - (\Delta t)^2$$

where the trigonometric functions sine and cosine have been replaced by their hyperbolic analogs, satisfying $\cosh^2 - \sinh^2 = 1$ instead of $\cos^2 + \sin^2 = 1$.

The fact that we can mix space and time in the same way we can mix length, height, and width has interesting physical consequences. For exam-

ple, we know that lying down you seem "shorter" (measured vertically) than standing up; you haven't changed, only rotated at an angle from the point of view of someone still standing. Similar things can happen with respect to the "angle" between space and time: "Tilting" your direction in that case means moving at some velocity; instead of changing your "slope" $\Delta y/\Delta x$, you change your speed $\Delta x/\Delta t$. This will have strange effects, as seen by an observer who has not speeded up: You will look shorter (in your direction of motion), and your watch will seem to run slower. Since the natural measure of speed is that of light, these effects will become large only if your speed is significant when compared to that of light.

An interesting way to think about the extra signs that relate Minkowski space to Euclidean space is with complex numbers: If we think of time as imaginary space, then we can write

$$(is)^2 = (\Delta x)^2 + (\Delta y)^2 + (\Delta z)^2 + (i\Delta t)^2$$

since $i^2 = -1$. This also explains the use of hyperbolic trigonometric functions, in terms of their definitions using exponentials:

$$e^{\pm i\theta} = \cos\theta \pm i \sin\theta, \qquad e^{\pm\theta} = \cosh\theta \pm \sinh\theta$$

There is a similar relation between the energy "E", momentum "p", and (rest) mass "m" of a particle, namely

$$-m^2 = p^2 - E^2$$

(p is a 3-vector, like position, but we have used just its magnitude for brevity, writing the analog of $-s^2 = r^2 - (\Delta t)^2$.) Energy and momentum form a 4-vector. Thus, energy conservation is associated with time translation symmetry.

Fields and waves

The concepts we have discussed so far are most conveniently ascribed to solid objects. In particular, they are most simply used to describe point-like objects, since we can attribute to them a position at a particular point in space at a particular "point" in time.

But some features of nature seem to prefer a completely different, more continuous description: For example, magnetism is generally described by a "magnetic field", which is not localized at a point, although we can ask how "strong" it is at any particular point. The same is true of gravity. Classically, there is this dichotomy between "matter" and "energy", where matter is localized, and can be divided into smaller pieces, while energy is exchanged between matter to influence its motion.

This energy is described by a "field", and the force attributed to it is proportional to its magnitude. The field is thus a function of space and time, as is the force it produces. The "energy density" due to this field (energy per unit volume, since the energy is distributed continuously throughout space) is proportional to the square of the field.

Since fields already fill all of space, any motion associated with them must be described by a change in their spatial dependence as a function of time. The simplest such dependence is called a "wave": For example, consider

$$\phi(x,t) = A\, e^{i(kx - \omega t)}$$

(which we can express in terms of trigonometric functions according to the relation above to make the oscillatory behavior of the wave more obvious). This describes a field ϕ at point x (for simplicity, we assume no dependence on y or z --- a "plane wave") and time t, where the constant A is its "amplitude", the constant k related to its position dependence is its "wave number", and the constant ω related to its time dependence is its "angular frequency". Since (pulling out a factor of k) ϕ depends on x and t only through the combination $x - (\omega/k)t$, we see that ϕ remains constant along a path $x = (\omega/k)t +$ constant, and thus ω/k is the velocity of the wave (in the x direction). Also, since $e^{i2\pi}=1$, the wave goes through a complete cycle (i.e., the "phase" $kx-\omega t$ goes through an "angle" of 2π radians) whenever x changes by $2\pi/k$ (for fixed t) or t changes by $2\pi/\omega$ (for fixed x). Thus, $2\pi/k$ is the "wavelength" and $2\pi/\omega$ is the "period". (Its inverse $\omega/2\pi$ is the "frequency".)

Waves produced by fields associated with electromagnetism or gravity travel at the speed of light, so $k = \omega$. For fields not of the simple wave form above, this relation can be generalized to a certain differential equation (involving derivatives with respect to space and time), called a "wave equation".

Quantum mechanics

There is often more than one way to describe the same physics. The most familiar example is wave-particle duality: Long before the days of quantum mechanics, there was disagreement among physicists on how to describe light, as waves or particles. In fact, at the macroscopic level, the distinction is rather moot, and either description can be used satisfactorily (once one knows about both "phase" and "group" velocities). But it seemed that the difference still might be determined at the microscopic level. When individ-

ual photons were first observed, through the photoelectric effect, rather than settling the matter in favor of particles, it was realized that the two descriptions were in fact equivalent at the quantum level: Mathematically, a point particle is described by its position (and perhaps other variables), while a wave is described by a field whose strength is a function of position. Quantum mechanically, a particle is described by a "wave function" which yields the probability of observing it at any position. This function is closely related to the field; e.g., both satisfy wave equations, and thus exhibit wave behavior. The same remarks apply to all forms of matter and energy, which are no longer distinguished (at least in this respect): Protons, electrons, photons, gravitons, etc., can all be described as either waves or particles.

You need a wavelength to identify a size for a wave. A more convenient quantity is the inverse of the wavelength (times 2π), or wave number k, as seen above. The quantity k is actually a (3-)vector, whose direction is that of the propagation of the wave. For the "size" in the time direction we have the angular frequency ω (frequency times 2π), an inverse time: You need the period of an oscillation to identify "when" a wave is.

Quantum mechanically, there is a relation between these geometric quantities and energy/momentum:

$$E = \hbar\omega, \qquad p = \hbar k$$

in terms of Planck's constant (divided by 2π) $\hbar$. We can choose clever units (e.g., Planck units) where $\hbar = 1$, just as we did for c. This means that energy should be considered as an inverse time, and momentum as an inverse length. We can then write the plane wave given above as

$$\phi(x,t) = A\, e^{i(px - Et)}$$

The fact that p is associated with x, and E with t, in this particular way is directly related to the fact that conservation of momentum p is equivalent to spatial translation symmetry, and conservation of energy E to time translation symmetry.

Since one has to measure a wave over a distance of a wavelength to determine that quantity, or wait for a period of oscillation to determine its frequency, the above relations yield the Uncertainty Principle, which states that momentum p and position can't be measured simultaneously with arbitrary accuracy, and similarly for time and energy E:

$$\Delta E\, \Delta t \geq \hbar/2, \qquad \Delta p\, \Delta x \geq \hbar/2$$

where "ΔE" is the uncertainty (inherent measurement inaccuracy) in E, etc. (in some appropriate definition, which gives the factors of 1/2). The main point is that to measure small distances, you need large momenta. Combining this with relativity, we see that it also requires large energies.

Particles can orbit each other, if bound by some force. The corresponding physical quantity is "orbital angular momentum", which is related to the angle of revolution in the same way that ordinary momentum is related to position. We thus have

$$\Delta L\, \Delta\theta \geq \hbar/2$$

in terms of orbital angular momentum L and angle θ, measured in radians. However, the fact that an angle can never exceed 2π leads to a minimum uncertainty in L: L must always occur in integer multiples of ħ --- it is "quantized".

In the same way, a particle can rotate on its axis. This property is "spin", or "internal angular momentum", the magnitude of which is "quantized" in *half* units of ħ.

In two dimensions, there is only one way to measure an angle about any point; but in three dimensions, one can measure an angle about any axis. (The angle is really a 3-vector.) Therefore, although spin is a vector, the component of that vector measured along any axis will be quantized. However, there is an uncertainty in the simultaneous measurements of different components of that vector. This is related to the fact that even classically a rotation about one axis followed by a rotation about a second axis will generally not produce the same result as performing those rotations in the opposite order. (This statement is not true when applied to translations, motion in straight lines.) Thus, one can only know one component of a spin (or any type of angular momentum) vector at a time, and that will be quantized. For example, an electron has spin 1/2 (times ħ), so any component of its spin can have values +1/2 or −1/2; a Z boson has spin 1, so any component of its spin can have values 1, 0, or −1. In general, a particle of spin s can have values s, $s-1$, ..., $-s$. Thus, a particle of spin s has $2s+1$ different "spin states": The direction of its spin is quantized with respect to whatever axis its direction is measured.

Locality

There is a general principle in physics that is often stated in different ways in different contexts, and sometimes not stated explicitly at all. The basic idea is "locality", that one object cannot influence another unless they come in direct contact. There is a weaker version of this principle, called "causality", that events at any one time are determined by the events immediately preceding it, and an even weaker version, called "macro-causality", that these events are determined by all the events occurring at some point in time

arbitrarily far back in the past. But special relativity combined with causality implies locality: Two events at different positions in space but perceived to be at the same time by one observer will appear to be at different times (with either one being earlier than the other) to an observer who is moving at some velocity with respect to the first observer, because he has a different choice of space and time "axes". Thus, to be at a slightly earlier time as seen by all observers means to also be only slightly separated in space. Within the framework of quantum field theory, macro-causality also implies locality.

For example, electromagnetism and gravity can cause one object to influence another only through particles (or waves) that carry the force from one object to the other. Although these particles are often invisible to the human eye, the fact that they are carriers of forces is no more strange than the fact the wind is composed of air molecules. In fact, sight itself is an example of such an influence, carried by particles of light. According to special relativity, particles can travel no faster than the speed of light; but light is much faster than the wind, so the fact that forces require mediators, and are not instantaneous, was not apparent to all the earliest physicists.

The resulting picture of nature is very simple: Everything in nature is particles, which interact only through collision. This point of view was foreseen by Democritus, but challenged by some later physicists, who advocated waves as the carriers of forces. With the advent of quantum mechanics, it was revealed that particles and waves are aspects of the same phenomenon. However, unlike Democritus' eternal atoms, particles can be transformed through such collisions: Besides two particles bouncing off each other, one particle can absorb the other, or they can combine to form a new particle (not simply the original 2 particles stuck together), or a particle can decay into 2 or more different particles, or 2 particles can scatter off each other creating a new, third particle, etc.

A natural way to examine and describe this local behavior of everything is through scattering experiments: Take a beam of something and use it to hit a target of something else, measuring the probability of collisions of individual particles in each, and the particles produced, as a function of their energies and angles of scattering. The picture of the particles involved is that two free particles scatter directly or by exchange of other particles, something like a game of pool, except that the only balls that are actually observed are the cue ball and the balls that wind up in the pockets (and the balls don't bounce off the edges of the table). The actual number measured is called a "cross section", and represents the effective area of the target particle. Since the target is a point particle, this area is really a measure of the range of the interaction.

This description of nature is exhibited through a "Feynman diagram", a picture showing the types of particles involved, their paths through space-time, and how they interact through collision.

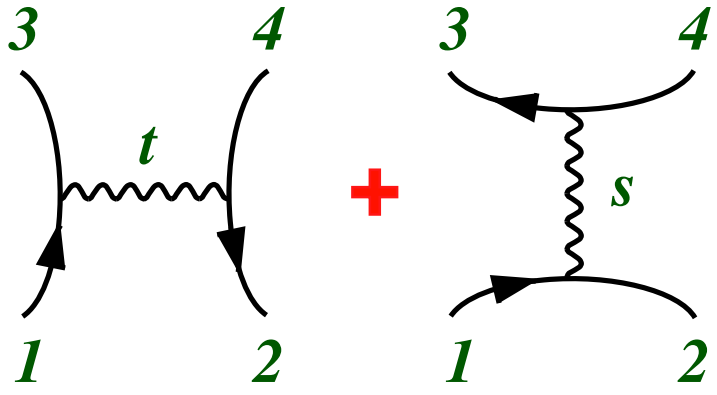

Fig. 2: *Feynman diagram*

This diagram is more than just an illustration, but is associated with an equation that allows calculation of a cross section. For example, each line

has associated with it a particular value of energy and momentum; the total energy and momentum is conserved at any interaction point (vertex). Other particle properties are associated with the lines, such as electric charge, which is also conserved, and spin, which restricts the type of interaction allowed; some of these properties may be indicated by labels or line styles. Different energy-momentum-dependent factors are then associated with each line and vertex and multiplied together. But more than one diagram can be associated with the same process: They only need to have the same particles going in and coming out; what happens in the middle may vary. Then the functions obtained from each diagram are added together, and the result can be used to obtain a numerical value for the cross section.

Particle physics

I work in "theoretical high-energy physics". This is the most fundamental area of physics; it is the study of nature at the smallest scale, smaller than the atom or even the nucleus. "Particle physics" is the treatment of this subject in terms of point particles; it is generally taken to mean all of high-energy physics except string theory.

Forces

Free particles can be described by their energy-momentum (and thus mass) and spin. These are their only "kinematic" properties, those that define how they relate to spacetime. But they can also be labeled by other properties that determine how they interact with each other. There are four known interactions:

(1) The force discovered first was gravity. Ironically, it is the weakest interaction, but dominates physics at large distances because its effect is always cumulative for multiple objects: It always attracts, while the other forces can also repel, and thus tend to cancel. Gravitational force is proportional to ("couples to") the energies of the interacting objects, and energy is always positive. Also, gravity propagates at the speed of light, and thus has infinite range, satisfying the usual inverse-square law.

(2) The next force to be understood was electromagnetism. It describes both electricity and magnetism, which are related to each other by relativity (in the same way as space and time, or energy and momentum). It also propagates at the speed of light (in fact, light itself is an effect of its propagation), but repels likes and attracts opposites: It couples to "electric charge", which comes with either positive or negative sign. Since charge, like energy, is additive, opposites continue to attract until they produce combinations whose net charge cancels. This can always happen because electric charge always appears quantized, in units of the electron's charge. The result is atoms, whose total charge cancels between protons and an equal number of electrons. However, since the protons are all located in the nucleus and the electrons a short distance away, their net force will not exactly cancel on objects close enough to notice this difference in range. This residual "Van der Waals" force falls off much faster than the inverse-square law, and is responsible for such short-range effects as friction.

(3) The "weak interaction" was first seen as responsible for the decay of nuclei, and in particular the decay of a neutron into a proton plus other stuff. It is also seen as a force between particles, but less frequently than the other interactions because it has the shortest range, and thus becomes important only at high energies, except in those special cases where the other forces (except gravity) do not contribute (as in most nuclear decays).

(4) The "strong interaction" is associated with binding neutrons and protons (or collectively, "nucleons") in the nucleus. It needs to be stronger than electromagnetism to overcome the latter's repulsion among protons. It appears to be a short-range force between nucleons; however, it is better understood as a Van der Waals type of force resulting form the true strong force that binds together the parts of a nucleon: Thus a nucleon itself is a type of "atom".

Fundamental vs. composite

An important concept that is often not appreciated even by some physicists is the distinction between "fundamental" and "composite" particles. An amazing duality in 2 spacetime dimensions relates particles of different spin: A free, massless particle of spin 1/2 can be described in terms of free, massless particles of spin 0, and vice versa. Getting spin 0 from spin 1/2 isn't so mysterious, since 2 spin 1/2 particles can be arranged so their spins cancel. But getting spin 1/2 from spin 0 is strange: It requires "superpositions" of arbitrarily large numbers of spin 0 particles. Thus, either the spin 0 or the spin 1/2 particle can be considered fundamental, as either can be constructed as a bound state of the other. (Alternatively, but less conveniently, one can use a redundant description where both are treated as fundamental.) The "binding" is trivial in this case, since massless particles in 1 space dimension going in the same direction all travel together at the speed of light. But in the massive case the same results can be obtained when interactions are included: The "sine-Gordon model" (spin 0) is equivalent ("dual") to the "massive Thirring model" (spin 1/2).

The lesson is that the distinction between "fundamental" and "composite" is formal; there is no physical difference. However, it is often clear that one description of a theory is far more useful (and clear) than another. For example, any quantum description of the atom treating it as fundamental would be far more awkward than the usual one, which treats electrons, protons, neutrons, and photons as fundamental. This would not necessarily be the case if electromagnetism were much stronger than it is: The strength of electromagnetism is described by the "fine-structure constant", $\alpha \equiv e^2/\hbar c \approx 1/137$, in terms of the electron charge e. The quantum mechanical size of an electron is given by its "Compton wavelength" (over 2π), $\hbar/mc$, in terms of

its mass m. (The proton is thus "smaller", since it's heavier.) The quantum mechanical size of the hydrogen atom is given by the Bohr radius, $\hbar^2/me^2$, which is larger by a factor of $1/\alpha$. However, if α were much larger than 1 instead of much smaller, the atom would be much smaller than the electron of which it is made, so its composite nature would be obscure. In general, weak coupling allows a simpler interpretation, since weak coupling is "close" to none at all. Such behavior is seen explicitly in the sine-Gordon/Thirring model: The coupling constant of one is large when the other is small.

The Standard Model

spin	**weak**	**strong**
2	graviton	
3/2	gravitino?	
1	W, Z, photon	gluons
1/2	e, μ, τ; neutrinos	quarks
0	Higgs?	

Table 1: *Fundamental particles*

The Standard Model embodies all the theoretical high-energy physics that has been verified by experiment (with the exception of the "Higgs boson", the most eagerly sought particle). The theory is defined by its spectrum of particles and how they interact. These particles are the "gauge bosons", of spin 1, which mediate the forces, the "fermions", of spin 1/2, which can be considered the basic constituents of matter, and the Higgs bosons, of spin 0, which are responsible for giving particles mass. The

"graviton" has spin 2, but is usually omitted because the energy scale at which gravity's strength becomes comparable to the other interactions is 20 or so orders of magnitude higher, so it's only seen in situations where the other forces cancel, like macroscopic bodies.

The gauge bosons are the mediators of the other 3 interactions: strong, electromagnetic, and weak. The one for electromagnetism is the "photon"; it is massless (like the graviton), so it travels at the speed of light (hence the name), and has infinite range, satisfying the usual inverse-square law. A force mediated by a massive boson, like the weak interactions, has a range equal to the inverse of its mass (again in units of ħ and c, the Compton wavelength), yielding an exponential decrease with distance.

The Standard Model is formulated in a way where all the "fundamental" particles are massless. It is only in this formulation that infinities can be removed in an obvious way. (There are also other significant technical advantages.) But few observed particles are massless: the photon, the graviton, and (maybe) the neutrinos. (Either some of the neutrinos of the Standard Model are massive, or some new ones are.) The other fundamental particles appear only in massive bound states.

There are 3 ways that bound states form, depending on the nature of the interaction:

(1) For electromagnetism, there is the Coulomb type of bound state, like atoms. States of the hydrogen atom have arbitrarily large spin, but their mass has an upper limit, beyond which ionization occurs. (When an atom is thought of as composite, its spin is called "total angular momentum", comprised of the spins of its constituents and the relative "orbital angular momentum". Likewise, its mass is the sum of the masses of its constituents plus the relative "kinetic energy" and the "binding energy".)

(2) For the strong interaction, there is confinement. "Gluons" (spin 1) bind themselves and "quarks" (spin 1/2) to form "hadrons" (such as the nucleons) of arbitrarily large spin and mass. These hadrons then interact with each other through residual strong interactions, in the same way that atoms interact through Van der Waals forces.

(3) For the weak interactions, the Higgs mechanism binds the Higgs particles to each other, to most of the fermions, and to the gauge bosons other than the gluons and photon. The resulting spectrum of bound states is limited in both spin and mass, and might include only "ground states", with no "excited states". 3 fundamental gauge bosons bind to the Higgs to form 3 massive bound states, called the W^+, W^-, and Z^0 (where the superscripts indicate their charges with respect to electromagnetism). The Higgs have only a single bound state, out of the original 4 fundamental fields; in picturesque language, 3 massless gauge bosons are said to have "eaten" the other 3 of the Higgs to become heavy. (Often these bound states are not distinguished from the fundamental ones, since binding with the Higgs doesn't change their spin, and excited states might not be seen. This is also why the Higgs mechanism is so convenient.) Although the Higgs mechanism might be replaced with confinement, it has the technical advantage of working already at the classical level, by a simple redefinition of fields (change of variables), whereas confinement requires a complicated (and perhaps intractable) summation of all quantum corrections.

These different types of bound states can be described by plotting their spin J as a function of their $(\text{mass})^2$, t: $J = \alpha(t)$. This is known as a "Regge trajectory". Cross sections can also be conveniently described in terms of this function, with t then representing the $(\text{energy})^2$ of the scattering. The above example describes bosons, so particles occur only when J takes positive integer or zero values.

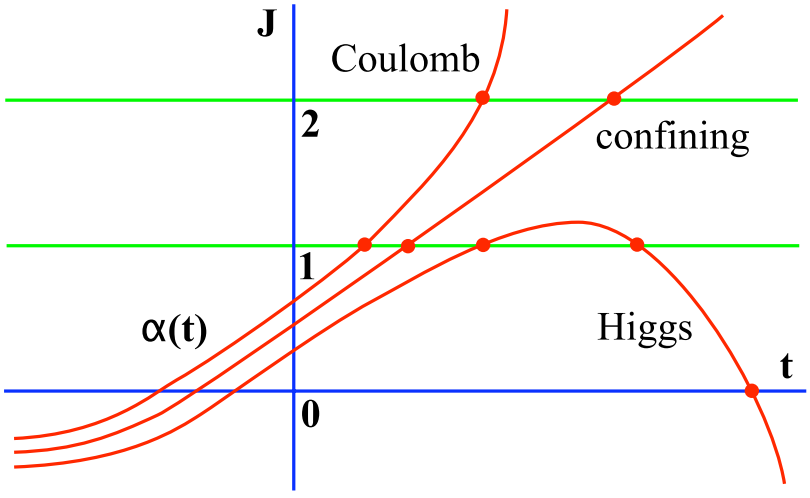

Fig. 3: *Regge trajectory*

force	**couples to**	**finite range from**
gravitational	energy	nothing
electromagnetic	electric charge	cancelation in atoms
weak	"flavor"	Higgs
strong	"color"	confinement

Table 2. *Fundamental forces*

Internal symmetry

We have already seen how relativity is an important symmetry of spacetime, restricting the form of theories, relating not only space and time but also energy and momentum, electricity and magnetism, etc. In particle physics there are other symmetries that are just as important, but unrelated to spacetime. (However, there have been proposals that such symmetries come from hidden extra spatial dimensions.) The most well-known example is the symmetry that relates the neutron and proton, called "isospin". The neutron and proton are identical in how they couple to the strong interaction. However, they differ in how they couple to the weak interaction, and to electromagnetism: The proton is electrically charged, the neutron is neutral. This accounts for the fact that they differ in mass, but only by a fraction of a percent. Thus isospin is only an approximate symmetry, yet a very accurate one as regards masses and the strong interactions.

Although isospin differs from spin in that it's unrelated to spacetime and is only an approximate symmetry, it's similar to spin in its quantization. Thus the proton and neutron can be regarded as two states of the same particle, the "nucleon" (with isospin 1/2), "rotated" in different directions in "isospin space".

Isospin generalizes to larger, but more approximate, symmetries. An easy way to see this is to note that quarks come in different "flavors", with only two of those flavors appearing in the nucleon: "up" and "down", referring to their "direction" in isospin space. But there are 6 known flavors of quark, corresponding to different directions in not just isospin but in a much larger internal symmetry space. However, while the up and down quarks are very light, the other quarks are heavier, varying from 10 times lighter than the nucleon to hundreds of times heavier. So it's more useful to think of hadrons as made of quarks than as different states of some flavor symmetry.

"isospin" ΔQ	quark (Q = ΔQ+1/6)	lepton (Q = ΔQ−1/2)
−1/2	down (.006)	electron (.0005)
+1/2	up (.003)	electron neutrino (0)
−1/2	strange (.10)	muon (.11)
+1/2	charm (1.2)	muon neutrino (0)
−1/2	bottom (4)	tauon (2)
+1/2	top (200)	tauon neutrino (0)

Table 3. *Fundamental fermions*

Masses given in parentheses in GeV. (Proton mass is .9, for comparison.) Q is electric charge (−1/3 or +2/3 for quarks, −1 or 0 for leptons). Quarks, besides their flavors, also come in 3 colors.

But there is another kind of internal symmetry that is exact and very useful, called local, or "gauge" symmetry. This is a symmetry that holds independently at every point in spacetime. So, instead of rotating everything in the Universe through the same angle in the internal space, things at different points in *spacetime* can be rotated at different angles in the *internal space*. (The internal-space angle is itself a function of the spacetime coordinates.) The result is that the internal angles measured at different spacetime points can't be compared, so they have little obvious physical meaning. But gauge symmetries are an important physical principle when combined with locality: The only way for such a symmetry to have any physical consequences is if there exists a field that instead of (just) rotating, is changed by the *derivative* (with respect to spacetime) of the angle of rotation. Since derivatives form a vector, this "gauge field" must itself be a vector, i.e., have spin 1. Since the

derivative of the angle is proportional to the difference in its value at 2 nearby points, the gauge field can be used to compare the value of this angle at those points, or more generally at 2 arbitrary points in spacetime. All spin-1 particles can be considered as having their own gauge symmetries; the charges of other particles under these symmetries then describe how the gauge field couples to them: Each gauge symmetry corresponds to a force.

In particular, the weak interactions couple to isospin. We then have 3 "families" of quarks/leptons (pairs out of the 6 flavors): Within each family, there are different couplings to the weak, electromagnetic, and strong interactions, but the 3 families are almost duplicates of each other, differing only by how they couple to the Higgs, and therefore in their masses. All these gauge symmetries are symmetries of the fundamental fields: For example, the observed, massive electron is a composite of the fundamental, massless electron and the Higgs; that's why the observed "isospin" symmetry between the electron and its neutrino is not exact (and things get worse for the more massive families), while the unobserved isospin *gauge* symmetry is exact (as required for coupling to the fundamental, but unobserved, massless W and Z bosons, which become massive after eating the Higgs).

The General Theory of Relativity

As a generalization, one can consider what happens if the gauge symmetry is a spacetime symmetry, rather than an internal symmetry. This generalization of the Special Theory of Relativity leads to the General Theory of Relativity. Since the symmetry transformations themselves are now described by a vector in spacetime (rather than in some internal space), the gauge boson now has spin 1+1 = 2: It's the graviton.

Special relativity introduced the concept of spacetime; separate measurements of distances in space and in time are physically meaningless. Just

as one must specify a time zone when stating the time (or worse yet, perhaps also whether one is using Daylight Savings Time), one must also specify the motion of the clock, to determine its relative velocity. General relativity generalized this concept by recognizing that spacetime is "curved": Just as measurements of distances on the surface of the Earth violate Euclid's axioms of geometry for a flat two-dimensional space ("non-Euclidean geometry"), measurements of distance in spacetime in the presence of gravity are not the same as in Minkowski space.

"Straight" lines can still be defined, essentially as giving the shortest distances between two points (although in spacetime they are really the longest distances, because of the funny extra signs in the definition of relativistic distance). However, there are no longer "parallel" lines in the usual sense: On the surface of the Earth, straight lines are really great circles, which always meet somewhere. Straight lines in curved spacetime describe objects in "free fall", allowing gravity to act on them without resistance. Such objects, initially not moving with respect to each other, may eventually meet at the source of gravitational attraction.

As in Newton's Laws (but not the one for gravity), the fact that curvature affects all objects means that all objects also create curvature. The curvature is due to their energy and momentum; curvature in the temporal and spatial directions is related in the same way as energy is related to momentum.

Gravity is generally not considered part of the Standard Model for several reasons:

(1) It is so weak that it plays little part in scattering experiments. Certainly the gravitational force between scattering particles is many orders of magnitude smaller than what can be measured presently, 40 or so orders of

magnitude smaller than the other forces. And the gravitational attraction of the Earth is irrelevant because it affects all particles in the same way, and doesn't measurably alter the path of the particles over the duration of the short-ranged weak and strong interactions.

(2) For similar reasons, it is not yet possible to measure quantum effects of gravity. Not only is it too difficult to observe the particles of gravity, gravitons, but even the waves of gravity have not yet been detected (though experiments are now looking).

(3) The fact that gravity is so weak at low energies means that it is very strong at high energies: Gravity couples to energy (and momentum) itself, rather than the various fixed charges to which the other forces couple. Although the energy scale at which gravity becomes comparable to the other interactions is 20 orders of magnitude greater than that of modern scattering experiments, one can consider the ramifications of its existence. The result is that it is difficult to find a satisfactory description of quantum gravity, for reasons we now consider.

Infinities

The appearance of infinities in quantum field theory has been a problem since its earliest days. The original solution, "renormalization", removed the infinities in a way that seemed consistent for any finite number of quantum corrections. But when all the corrections were added up, the problems returned. The only solution is to choose theories that are finite from the start, which requires supersymmetry.

Renormalization

Particle physics is based on renormalizable, relativistic quantum field theory: "Relativistic" means consistent with the Special Theory of Relativity; "field theory" means local (in time, but by relativity also space) when expressed in terms of waves; "quantum" means consistent with quantum theory; "renormalizable" means that the process of obtaining the quantum theory as corrections to the classical theory does not introduce new arbitrary parameters (masses, coupling constants) into the theory.

Calculations in particle theory are performed "perturbatively": First the classical calculation is performed, then quantum corrections are found. There is a whole series of such corrections, which is conventionally expressed as an expansion in powers of ħ, although since ħ is effectively 1, a more proper

way to formulate the expansion is in powers of the "coupling constants", a measure of the strength of the interaction, like the electron charge.

The first problem in such an approach is that some quantum corrections appear to be infinite: Classically, two particles scatter by exchanging a particle, thereby trading some amount of energy and momentum. That amount is determined by conservation of energy and momentum: By examining these quantities before and after the interaction, the difference gives the amount attributed to the particle mediating the force. Quantum corrections involve exchanging more than one particle. (This is "quantum" in the language of scattering classical waves, not classical particles: It is higher order in coupling constants, since it involves multiple interactions.) However, specifying the energy and momentum before and after the total process does not determine that of each of the exchanged particles, but only their sum. There is thus an infinite number of ways these quantities could be distributed, and performing this sum (actually an "integral" in the sense of calculus) can lead to an infinite result, depending on the energy- and momentum-dependence of each individual process.

Fortunately, it turns out that in any case not all the pieces of such a calculation can be infinite: The infinite piece itself looks like yet another classical process. The procedure of "renormalization" involves defining the original classical process to be infinite in such a way that it cancels the infinities of the quantum corrections. This is not an arbitrary procedure, because the classical theory is required to be local and relativistic.

The second problem in such an approach is that this procedure might require an infinite number of different types of interaction, each with its own coupling constant. Although the calculations themselves would still give finite answers, the predictions would be useless, because there would be an infinite number of arbitrary coupling constants to fix, by an infinite number of different measurements. (There might still be limited value, in that generally

each type of interaction has a different energy dependence, and so only a finite number of the interaction types, and thus a finite number of coupling constants, would be important at low energy, to some order of approximation.) Fortunately, it turns out that a certain finite number of types of interaction do not require new types of interaction to perform renormalization. Such interactions are called "renormalizable". This is *the most important restriction* in quantum field theory, because it severely limits with what types of classical theories one can start: In particular,

(1) it limits all spins to be only 0, 1/2, or 1 (and thus not the spin 2 of the graviton);

(2) spin 1 can become massive only by eating spin 0; and

(3) the number of dimensions of spacetime must be no more than 4 (i.e., the real world has the maximum number of dimensions).

Resummation

For this expansion to be useful, the coupling constant should be much less than 1; then each correction will be much smaller than the preceding one (and all less than the classical term), so the classical expression is a good approximation, and each correction makes the estimate better and better. Unfortunately, this does not happen; the corrections do not "converge" to the answer. There is a way to (almost) fix this, by reorganizing the sum. Thus, while the procedure of "renormalization" removed infinities in each quantum correction, the "resummation" procedure removes infinities from adding up the sum of an infinite number of terms. However, resummation, like renormalization, can replace infinities with ambiguities; the method of removing infinities isn't always unique.

Consider the following simpler analogy, an expression we can easily write as an infinite sum:

$$1/(1 - x) = 1 + x + x^2 + x^3 + ...$$

This can easily be proven by multiplying both sides by 1 – x; on the right side all terms will cancel except the 1. We can also check some simple examples: x = 0 works trivially. For x = 1/2 we get 2 on the left; on the right, as we add one term at a time in the sum, we get 1, 1 1/2, 1 3/4, ..., as each successive term takes us half of the way closer to 2. But if we try x = – 1, on the right we get an oscillation between 1 and 0, while on the left we get the average value, 1/2. Thus, the expression on the left gives a well defined result for the sum, even for the cases where the sum does not converge term by term. A more extreme example is the case x = 2: On the left we get – 1, while on the right we get a divergent series, each term positive and bigger than the term before. Clearly the expansion itself is useless, although we might still hope that this method of performing the sum could provide a useful result.

Unfortunately, even worse types of sums can occur in field theory. Consider

$$\sqrt{1 + x} = 1 + (1/2)x - (1/8)x^2 + (1/16)x^3 + ...$$

(The values of the coefficients can be checked by squaring both sides.) Take the case x = – 2 (or any negative number < – 1): While again the right side doesn't converge, the left side has the worse problem that it gives $\sqrt{-1}$, which has two values, + i and – i. (In fact, the square root of any number has two values, differing only in sign, since $(-1)^2 = + 1$, but for x > – 1 we might argue that it is more natural to choose the positive real number solution, while for imaginary numbers there is no clear preference.)

This type of ambiguity appears in any resummation of a perturbation expansion for a theory that required renormalization. Worse yet, such a theory has an infinite number of such ambiguities, appearing at higher and higher energies. Thus, any theory that appears renormalizable at finite orders of quantum corrections will be found nonrenormalizable once the sum has been performed, even after methods of resummation are used to make divergent sums converge. The moral is that removing infinities does not solve the problem, but merely transforms it into a problem of an infinite amount of ambiguity.

Superstring theory

What's next?

Unfortunately, little tangible progress has been made in theoretical high-energy physics since the 1970's (again, coincidentally, the time I was a student), as a result of which theory and experiment in this area have diverged to the point where a third approach, "phenomenology", was invented to try to bridge the gap between the two. Certainly that makes this a very challenging time for the subject. This lack of progress has led to a bit of polarization in the high-energy physics community: On the pessimists' side, it ranges from a rejection of the more popular string theory approach, to a move from theory to phenomenology, to an abandonment of high-energy physics altogether for more lucrative areas, such as condensed matter physics. On the optimists' side, string theory has amassed such a huge amount of new mathematical results and new ways of interpreting old physics results, compared to any would-be competitors, that it is often accepted as true without solid evidence.

Although the Standard Model does a very good job (especially for electromagnetism, where it has been found accurate to a few parts per trillion), it has a few shortcomings:

(1) Confinement of quarks and gluons to produce hadrons has not been proven and, more importantly, no effective method of calculation is known

which would allow one to calculate scattering cross sections for hadrons (although some pieces can be).

(2) The Higgs hasn't been seen yet, but so far this hasn't presented a contradiction, and its discovery is expected at the Large Hadron Collider. Of course, some alternatives have been suggested, just in case; the most notable is "technicolor", which proposes to produce the Higgs boson, or perhaps just its eaten part, by a new kind of confinement. So, either the Higgs will be found, or we will just extend the confinement problem.

(3) The Standard Model isn't as simple as it could be, being comprised of three unrelated forces acting on three families of assorted particles. One approach to this is called "Grand Unified Theories". Many such theories have been proposed, and await experiments to sort them out. They simplify everything except the Higgs bosons (but maybe technicolor can help there, too).

(4) It is renormalizable only to any finite order of quantum corrections. However, when all the (infinite number of) corrections are summed up, renormalizability is lost (even after resummation). This is known as the "renormalon" problem. Essentially, the problem of infinities in quantum corrections, which was thought removed by the renormalization procedure, was only postponed. The only solution is to work with a finite theory, where no renormalization is needed. Such finite theories require "supersymmetry", a symmetry that relates particles of different spin. Many such theories are known; they would (at least) double the number of particles, because no supersymmetry "partners" have yet been seen.

(5) Gravity is not included, at least not at the quantum level. Quantum gravity is nonrenormalizable.

A solution?

Superstring theory has been proposed to solve all these problems: (a) It seems to be finite, partly as a result of supersymmetry. (b) It includes gravity (without destroying finiteness). (c) It is relatively unique, apparently solving the unification problem. (d) It can be used to describe hadrons, thus offering a solution to the confinement problem.

However, it is not clear how useful some of these "solutions" are. The main problem is that superstring theory is naturally defined in 10 dimensions (or perhaps 11, if it is really a theory of supermembranes). There is a great deal of ambiguity in how to reduce this to the usual 4 dimensions of space and time, making any unification rather hollow. In particular, in some choices of such "compactifications", the string states correspond to hadrons, while in other cases they correspond to the quarks and gluons that make up the hadrons, as well as to gravitons. So it seems that superstring theory can have many different interpretations, solving all the problems of high-energy physics, but not all at the same time. This ambiguity in its interpretation is a major symptom of the 10D problem.

But there is a worse symptom of this problem, a difficulty that sometimes is not mentioned because it is a drawback with *all* methods that have tried to extend the Standard Model by addressing the topics of confinement or quantum gravity: It is too difficult to calculate with it. Historically there have been many approaches to theoretical high-energy physics based on clever ideas, grounded in solid physical principles, and able to derive new results in a concrete way. But their results were limited to constants, such as masses or couplings, and other features of low-energy (by the standards of high-energy) physics. Although such results are significant in areas where no better progress has been made, they pale in comparison to the Standard Mod-

el in areas where it has predicted functions, such as scattering cross sections that depend on energies and angles. Predicting constants is nice, but physicists are willing to accept constants as inputs if they can get out predictions of functions.

So, if the Standard Model works so well, and no confirmed solutions have yet been found to its few remaining problems, why is so much effort, perhaps even more than is now devoted to the Standard Model, spent on string theory, or any other theory for that matter? There are several reasons:

(1) It's always good to have alternatives, even if they aren't as good. That makes it more clear which results of the Standard Model are independent, and not merely consequences of simpler assumptions.

(2) Almost everything that can be done with the Standard Model has already been done. To get really new physics, one needs either to add something new to it, or to find a new way to calculate with it.

(3) New experimental results will eventually be obtained. There are already some indications from astronomy and cosmology, such as masses for neutrinos (a small modification to the Standard Model), and dark matter: missing matter that contributes to the mass of the universe, not accounted for by the Standard Model, that is unseen because of its lack of electromagnetic interaction. Much more is expected from the Large Hadron Collider (although there are pessimists). Many physicists propose models in advance, in the lack of any experimental evidence for their modifications of the Standard Model. Although this is like shooting in the dark, it might save time once the data becomes available.

(4) Many things can be learned without new experimental input. When Einstein discovered the General Theory of Relativity, there was no great need

for a new theory of gravity: Newton's theory seemed to work quite well. But it had a problem of consistency with the Special Theory of Relativity. Einstein's explanation of gravity seemed much more fundamental: It explained gravity as a feature of the geometry of spacetime, making it clear why it should have such a universal relation to all forms of matter and energy.

(5) Even if there is no inconsistency with experiment or theory, there may exist simpler explanations of known results. The simpler explanation might not even make new predictions but, because of its simplicity, allow or suggest new generalizations or applications. For example, Minkowski's description of special relativity in terms of measuring lengths in four-dimensional spacetime led to Einstein's general relativity. Also, Schrödinger's wave equation was a reformulation of Heisenberg's matrix mechanics that made calculations in quantum mechanics much easier, and Feynman's diagrams did the same for quantum field theory.

(6) Models that are wrong are not necessarily useless. Classical physics is "wrong", but it is a good approximation to quantum physics under many situations, especially at macroscopic distances. There are also "toy models", which are known to be wrong, but are easier with which to work than realistic models, and have enough realistic characteristics to be used as learning tools. For example, "asymptotic freedom" was a feature of large-angle scattering discovered in the late 60's that indicated that the constituents of hadrons behaved as if they were free from the strong interactions at high energies. Unfortunately, all field theories known at the time predicted that the opposite should be true. But field theory continued as a model for the strong interactions, even though it seemed to require *ad hoc*, and unsatisfactory, modifications to avoid (or ignore) this problem. By the early 70's, calculations indicated that Yang-Mills theory actually solved the problem, and in the form of Quantum ChromoDynamics became the Standard Model of the strong interactions.

Supersymmetry and supergravity

The most common generalization of the Standard Model is supersymmetry. (Grand Unified Theories generally make predictions only for energies so high they will not be accessible in the near future. Technicolor requires confinement, with which it is difficult to calculate.) The simplest generalization to general relativity is "supergravity", which is the supersymmetrization of Einstein's gravity. Supersymmetry and supergravity are also parts of string theory.

Supersymmetry is a symmetry that relates particles of different spin. There are many reasons people became interested in supersymmetry, but the main reason is that it improves the behavior of field theories with respect to renormalizability. Although there is little if any experimental evidence in favor of supersymmetry at this time, it is seldom realized even by physicists that supersymmetry is as crucial to particle physics as the Higgs boson. Higgs bosons are the only known calculable way that masses can be given to self-interacting particles of spin 1 (namely, the W and Z bosons) while maintaining renormalizability. However, renormalizability is not enough: The method of removing infinities isn't always unique. The condition that it should be unique restricts the set of allowed theories; e.g., there can be no spins higher than 1, and Higgs bosons are needed for certain properties. In the same way, resummation makes further requirements for uniqueness, including supersymmetry. While the Higgs should probably be seen eventually at the LHC, to avoid coupling constants that are too large, the predictions for seeing supersymmetry are not as particular, though eagerly sought: In any (renormalizable) supersymmetric theory, for any spin-0 or spin-1 particle type there must be a similar particle type of spin 1/2, and vice versa. However, supersymmetry may be "broken" in a way that does not prevent the exis-

tence of such particles, but makes them more massive, making them harder to produce.

Supergravity was originally hoped to do for quantum gravity what supersymmetry does for quantum field theories of lower spins. Pure gravity is not renormalizable, because the gravitational coupling is proportional to the energy. This means that it grows at high energies more quickly than the forces of the Standard Model, producing more severe infinities in quantum corrections. Supergravity removes some of these infinities, because supersymmetry helps cancel infinities. But while supersymmetry can sometimes cancel all infinities in theories of spins 0, 1/2, and 1, it doesn't appear to be strong enough to cancel infinities from spins 2 and 3/2 (the graviton and "gravitino").

Just as the concept of Minkowski space clarified special relativity and paved the way to general relativity, "superspace" simplified the concept of supersymmetry, and made it easier to apply. In some of my early research I found how to reformulate supergravity in superspace and thus find new theories (with Gates), and how to apply superspace to quantum calculations (with Grisaru and Roček); these superspace techniques are now standard, and some have found applications even for nonsupersymmetric theories. I also found new ways to formulate and quantize superstrings in superspace; this investigation is still ongoing.

Strings

A harmonic oscillator is physically described by a spring connecting two weights; a string is an infinite number of infinitesimal springs all linked in a chain. For the relativistic string all the vibrational modes are perpendicular to the length of the string, so effectively all the little springs are linked

side-to-side instead of end-to-end. You can see this behavior in an ordinary string, like on a violin: The string will vibrate in waves, with wavelengths that can be (twice) any fraction of the length of the string; there are an infinite number of such modes (harmonics), each with its own amplitude of oscillation. But a single spring can vibrate only in one way (although also with arbitrary amplitude): A tuning fork is a spring, and has only one note, of arbitrary volume; a single violin string can play an infinite number (ideally) of different notes.

Classically the frequencies of the notes of a string are quantized, but quantum mechanically the volumes are also. Just as the energy levels of an atom can take only fixed values, the same is true for a single spring. A hydrogen atom has a "ground state" (state of lowest energy), with "excited states" whose energies get higher and higher, but closer and closer, converging to the energy where the electron and proton break apart. But a (ideal) spring has equally spaced energy levels (or energy squared in the relativistic case), stretching to infinity. For the string, this feature is confinement: The quarks and gluons that make up the string never escape each other. A string may break, but only into more strings.

One interesting feature of the hadronic spectrum is that one can identify some particles as ground states of which other particles can be seen as excited states with similar properties except for mass (rest energy) and spin. Because the coupling dependence of a particle goes as its energy to the power of the spin of the particle conveying the force, the highest spin at any particular energy level is the most evident in high-energy scattering. On a graph of spin vs. energy squared, these states plot as a straight line. This already indicates springy behavior, but closer analysis indicates strings.

When string theory was invented the problem with dimensions was not evident. In fact, string theory originated from a scattering amplitude for 2

hadrons into 2 hadrons, in 4 dimensions. This amplitude was cooked up to satisfy properties expected from treating hadrons as bound states at various excitation levels (although originally "bound state of what" was not specified, nor even sought). This single function was soon generalized to amplitudes for more particles. The mathematics was recognized as that of a set of harmonic oscillators, which were then identified with the vibrational modes of a string. Favorable comparisons with experiment were made, with respect to not only spectrum but also high-energy scattering at small angles.

However, high-energy scattering at *large* angles showed a discrepancy with experiment, even qualitatively. More serious problems appeared when quantum corrections were considered, which could be solved only by recognizing that the theory was naturally defined in 10 dimensions (or worse yet, 26 dimensions for the original model, which lacked fermions). These two problems in fact had a common origin: The "partons" (would-be quarks and gluons) that made up these strings did not behave as ordinary particles, or show any signs of physical degrees of freedom; they did not have any classical analog. This feature allowed them to avoid the usual nonrenormalizability problems; but renormalizability is the feature of relativistic quantum theory that determines the dimension of spacetime to be 4. (There is also confinement, but these partons have no degrees of freedom to confine.) As a result, most string people gave up on using them to describe hadrons, in spite of the fact that they were the only particles for which strings had shown any experimental success, and studied them instead as a theory of unification.

Another problem with string theory, that shows up together with the dimension problem, is the appearance of a massless spin-2 particle, unlike hadrons, which are all massive, and there is no known way to give mass to that particle. This particle can be identified with the graviton, so the string can be interpreted as describing quantum gravity. Along with the graviton come other massless particles, especially because of supersymmetry, suggest-

ing the identification of some of them with gluons. Continuing in this direction, one can try to interpret string theory as a unified theory of all forces and particles. However, the distance/energy scale of a unified theory of gravity is much different from that of a theory of hadrons: While the stringiness of hadrons was observed in the 1960's, any stringiness of gravity will probably not be observed in the foreseeable future. (However, there are models, with or without strings, where the effects of extra dimensions might be seen earlier.) That makes it much more difficult to find advantages of such a string theory, or to distinguish its predictions from similar theories that don't require strings.

Although string theory has yet to succeed as a useful description of the real world, many ideas already have sprung from it that have applications to ordinary field theory, including the invention of supersymmetry and supergravity. Furthermore, the fact that the known string theories provide scattering amplitudes that are relatively easy to calculate, and incorporate many observed properties of hadrons that no other theory does, although in the wrong dimension, indicates that at least some of the principles on which it is based will eventually be incorporated into a successful theory of nature.

The future of string theory?

New strings?

One problem on which I am working is to find a new string theory that is defined in four dimensions. Confinement is a long-distance property (quarks and gluons not escaping to infinity), renormalizability is a short-distance property. The greater the number of dimensions, the weaker forces become at long distances, and the stronger they become at short distances. Four is the maximum number of dimensions for both confinement and renormalizability, at least for field theories of particles.

But confinement works in a funny way in known string theories. Formulating these strings as bound states of particles, we find that these "particles" are trivial: If we heat up ordinary matter enough it eventually turns into a plasma, making the atoms fall apart ("ionize") into its constituent electrons and nuclei. If we do the same for the nuclei themselves (though much hotter), they ionize into quarks and gluons. But if we would do the same for higher-dimensional strings, we would find almost nothing --- neither strings nor particles, nor anything with which we might associate a position. This is in strong contrast with the nucleons, which behave like strings, yet ionize to particles.

In string theory there is a symmetry between position and momentum, called "T-duality". This symmetry formally holds in all dimensions (although strings themselves are consistent quantum mechanically only in certain num-

bers of dimensions), as a consequence of the trivial nature of the strings' constituents. On the other hand, if we try to formulate a string theory as bound states of true particles, we find that T-duality can exist only in four dimensions.

These things suggests that strings that are constructed from particles will require 4 dimensions of spacetime, and that the usual string theories require higher dimensions because they are lacking some important ingredient. Part of what is necessary is a formalism where the strings and the particles of which they are composed can be seen at the same time.

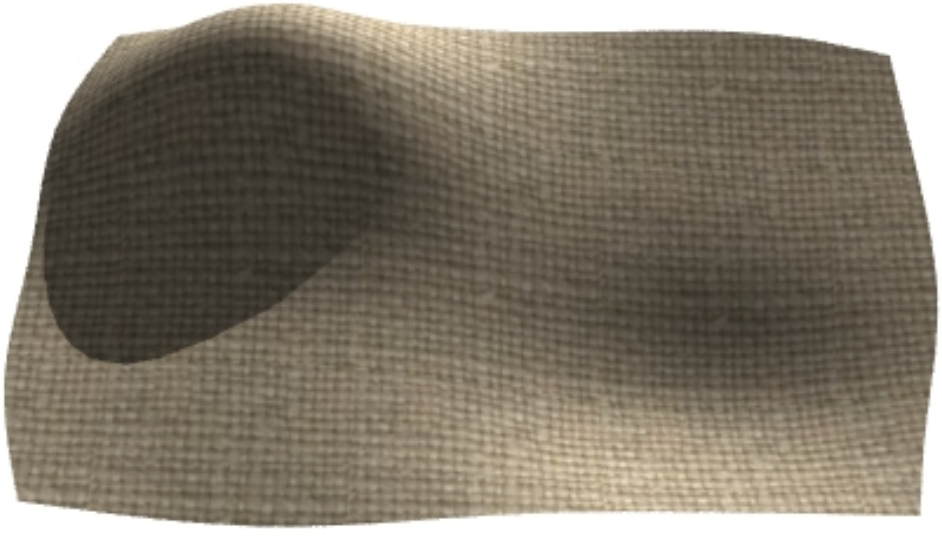

Fig. 4: *The string sheet is woven from particle lines*

Just as a point particle produces a line as its path through spacetime, a string sweeps out a sheet. (Actually an open string will sweep out a sheet, but a closed string, which has no ends, will sweep out something more like a cylinder.) But a real sheet is woven from threads. If we think of these

threads as the lines of particles, we can see how to interpret a string as a bound state of many particles, whose individuality is lost unless we examine the string in close detail. Since by quantum mechanics short distances means high energies, we expect that the same cross section that describes scattering of strings (hadrons) will exhibit their partons (quarks and gluons) when the energies become sufficiently large.

Relations between strings and particles

String theory has been useful as a toy model, telling us many things about particles that we had missed. For example, supersymmetry was discovered from string theory, as were certain simplifications in quantum corrections to particle theory. In fact, the first paper on strings (although not string theory as we know it today) was written in 1747 by Jean le Rond d'Alembert, and was the first appearance of the wave equation, which laid the foundation for field theory (especially electromagnetism), special relativity, and quantum mechanics.

There are methods of calculation in string theory that are necessary because of its extreme complexity. Part of my research involves finding ways to apply these methods to particle theory. In principle these should also simplify particle calculations, but have not yet done so because some of the relations between particles and strings are still obscure. Conversely, there are some standard methods in particle physics that have not yet been fully applied to string theory: For example, some of my well-known, early work (partly with Barton Zwiebach) involved formulating the field theory of strings; one consequence was the discovery of how to write the free field theory of arbitrary types of particles. More importantly, understanding the relationships between particles and strings would give an explanation of confinement.

Modifications at short distances

In quantum mechanics, various things are quantized, such as angular momentum, energy levels of atoms, etc. General relativity is fundamentally about how distances are measured, which varies from point to point, since space is curved, and the curvature varies from point to point. This suggests that in quantum gravity distances themselves must be quantized in some sense. I am currently investigating several methods of incorporating such a quantization of spacetime. One of these involves treating the space of momentum as a sphere: Since large momenta correspond to short distances, thus limiting the maximum value of momentum also limits the smallest values of distances, in a way whose symmetry reflects that of the sphere. Such approaches necessarily have a strange effect on supersymmetry, since it can be considered as a type of square-root of momentum.

A related short-distance modification of spacetime leaves the usual interpretation of momentum but weakens the short-distance behavior of interactions to produce a similar effect. In this approach spacetime still exists at short distances, but the gravity associated with the curvature of spacetime does not. In such a theory of gravity black holes would not exist, because the singular behavior of classical gravity in such situations would be removed. (Large aggregates of mass characteristic of black holes have been observed in the universe through their strong gravitational fields, but the resulting black holes predicted by classical General Relativity have not.) Similarly, although there would still be a Big Bang at the beginning of the Universe, there would be no corresponding infinities, so it really would have been more of a Big Bounce.

This is closely related to the treatment of the graviton as a bound state: At high energies the graviton breaks apart into its constituents, so gravity disappears. Such behavior is seen in string theory: The graviton appears as a closed-string state, but the closed string appears as the bound state of two open strings (by connecting their ends).

Glossary

Angular momentum: Position multiplied by momentum orthogonal to the position; analogous to momentum, but defined with respect to angular (instead of linear) motion. It can be either "orbital" (with respect to revolution), or "spin" (with respect to rotation).

Boson: Particle of integer spin.

Chromodynamics: Theory of (strongly) interacting quarks and gluons.

Closed string: String that occurs as a closed loop, with no ends.

Confinement: Particles being bound together so strongly that they cannot be separated. This usually refers to quarks and gluons.

Cross section: Area of a particle as seen in a scattering experiment. This depends on the nature of the scattering: angle, energy, type of particle scattered, etc.

D-branes: Extended objects to which ends of open strings may be attached.

Dimensional analysis: Checking that units agree on both sides of an equation after doing algebra.

Duality: Equivalence of two different descriptions.

Family: Set of fermions (quarks and leptons) whose members couple differently from one another to strong, weak, and electromagnetic interactions. There are 3 known families, each with members identical to those of another family, except for couplings to Higgs (and thus masses).

Fermion: Particle of half-integer spin.

Field theory: Description of phenomena as waves.

Gauge boson: Particle of spin 1. These are the mediators of all forces except gravity.

Gauge symmetry: Internal symmetry, with independent internal angles for each point in spacetime. Its gauge field changes under this symmetry by the derivative of the angle.

Gluon: Massless particle that mediates strong interactions.

Gravitino: Spin-3/2 particle related to graviton by supersymmetry.

Graviton: Particle that mediates gravity. It is massless and has spin 2.

Ground state: Lowest-energy state of a system.

Hadron: Strongly interacting particle, formed as bound state of quarks and gluons. They come in all spins, and all are massive.

Harmonic oscillator: Idealized spring. It is a simple approximation to many dynamical systems. Its energy levels are equally spaced.

Higgs (boson): As-yet-unseen particle of spin 0, responsible for the simplest description of how the particles of the Standard Model aquire mass in a way consistent with renormalization.

High-energy physics: Physics at the highest energies, and thus, according to quantum mechanics, at the shortest distances. The most fundamental area of science.

Internal symmetry: Symmetry unrelated to spacetime (as far as we know).

Isospin: Approximate, rotation-like symmetry relating, e.g., protons and neutrons. It is quantized like spin.

Neutrino: Massless (or nearly so) particle of spin 1/2. It interacts only through gravity and weak interactions.

Nucleon: A particle that makes up the nucleus of an atom, with the neutron and proton considered as its two forms, or isospin states.

Open string: String with two ends.

Order of magnitude: Power of 10; the number of digits in a number before the decimal point (minus one), or if the number is less than one, minus the number of zeros immediately after (minus one). Generally expressed by writing that number as a superscript on "10".

Phenomenology: Area of high-energy physics that tries to tie together theory and experiment. It keeps closer ties with experiment than theory, while avoiding hypotheses that deviate too much from established theory, except perhaps in simplified or narrow form.

Photon: Massless particle that mediates electromagnetism (as light, e.g.).

Planck units: Units in which the speed of light in a vacuum (c), Newton's gravitational constant (G), and Planck's constant divided by 2π ($\hbar$) all take the value 1. Thus these units are the most natural ones (as opposed to artificial, human ones like the metric system) for special and general relativity and quantum mechanics. These 3 constants fix the units of mass, length, and time.

Quark: Particle of spin 1/2 that is confined in hadrons. It is the only type of fundamental particle that couples to all the interactions.

Renormalization: Procedure by which quantum corrections to classical field theory are evaluated in such a way as to eliminate infinities. Although it eliminates ambiguities at any finite number of steps, ambiguities return when the infinite number of steps are summed. It thus should be considered only as a stop-gap measure, that can be fixed only by supersymmetry.

Resummation: Getting a result out of a divergent sum.

Scattering amplitude: In quantum mechanics, a complex number (or function) representing the scattering of particles, whose absolute value (squared) yields a probability of scattering.

S(cattering)-matrix theory: Theory of elementary particles based on scattering amplitudes as fundamental quantities.

Spin: Instrinsic angular momentum of a particle; the property that identifies a particle as a "top" of quantized rotation. Its quantum mechanical value is an (non-negative) integer or half-integer multiple of ħ.

Standard Model: Theory of everything except gravity. It is well verified, even at the quantum mechanical level, except for the existence of the Higgs boson.

String: Simple generalization of the particle to an extended object, in a way consistent with quantum mechanics and special relativity.

Supergravity: Supersymmetric extension of gravity.

Superspace: Supersymmetric extension of spacetime.

Supersymmetry: Symmetry that relates particles of different spin. It is better behaved under renormalization; some such theories eliminate the procedure of renormalization altogether, along with its infinities and resultant ambiguities.

Symmetry: Relation between various quantities, so they can be described in the same way. For example, the laws of physics are the same in London as in Beijing, and the same at noon as at midnight.

T-duality: Symmetry between position and momentum.

Technicolor: Confinement as a way to avoid the Higgs.

Uncertainty principle: The position and momentum of a particle can't be measured simultaneously with arbitrary accuracy.

Unification: Simplifying the particles and forces by increasing the symmetry.

Van der Waals force: A type of force that is not fundamental. It is the remains of a more fundamental force whose charges cancel, but are not located at the same point.

Yang-Mills theory: Description of self-interacting spin-1 particles, such as gluons.