Why the Renormalization Group Is a Good Thing

Steven Weinberg

My text for today is a paper by Francis Low and Murray Gell-Mann. It is "Quantum Electrodynamics at Small Distances," published in the *Physical Review* in 1954.

This paper is one of the most important ever published in quantum field theory. To give you objective evidence of how much this paper has been read, I may mention that I went to the library to look at it again the other day to check whether something was in it, and the pages fell out of the journal. Also it is one of the very few papers for which I know the literature citation by heart. (And all the others are by me.) This paper has a strange quality. It gives conclusions which are enormously powerful; it's really quite surprising when you read it that anyone could reach such conclusions: The input seems incommensurate with the output. The paper seems to violate what one might call the First Law of Progress in Theoretical Physics, the Conservation of Information. (Another way of expressing this law is: *You will get nowhere by churning equations.* I'll come to two other laws of theoretical physics later.)

I want here to remind you first what is in this paper, and try to explain why for so long its message was not absorbed by theoretical physicists. Then I will describe how the approach used in this paper, which became known as the method of the renormalization group, finally began to move into the center of the stage of particle physics. Eventually I will come back to the question in the title of my talk—why the renormalization group is a good thing. Why does it yield such powerful conclusions? And then at the very end, very briefly, I'll indicate some speculative possibilities for new applications of the ideas of Gell-Mann and Low.

Let's first take a look at what Gell-Mann and Low actually did. They started by considering an ancient problem, the Coulomb force between two charges, and they asked how this force behaves at very short distances. There's a naive argument that, when you go to a very high momentum

transfer, much larger than the mass of the electron, the mass of the electron should become irrelevant, and therefore, since the potential has the dimensions (with $\hbar = c = 1$) of an inverse length, and since there is no other parameter in the problem with units of mass or length but the distance itself, the potential should just go like the reciprocal of the distance r. That is, you should have what is called naive scaling at very large momentum transfers or, in other words, at very short distances. Now, this doesn't happen, and this observation is the starting point of the paper by Gell-Mann and Low. The leading term in the potential, due to a one-photon exchange, is indeed just α/r. However, if you calculate the first radiative correction to the potential by inserting an electron loop in the exchanged photon line, you find a correction which has a logarithm in it:

$$V(r) = \frac{\alpha}{r}\left[1 + \frac{2\alpha}{3\pi}\left\{\ln\left(\frac{1}{\gamma m_e r}\right) - \frac{5}{6}\right\}\right] \quad (\gamma = 1.781\ldots). \tag{1}$$

This does not behave like $1/r$ as r goes to zero.

The questions addressed in the paper by Gell-Mann and Low are, first, why does the naive expectation of simple dimensional analysis break down? And, second, can we characterize the way this will happen in higher-order perturbation theory? And, third, what does the potential look like at really short distances, that is, when the logarithm is so large as to compensate for the smallness of $2\alpha/3\pi$? Those distances are incredibly short, of course, because α is small and the logarithm doesn't get big very fast. In this particular case the distance at which the logarithmic term becomes large is 10^{-291} cm. Nevertheless, the question of the behavior of the potential at short distances is an important matter of principle, one that had been earlier discussed by Landau and Källén and others, and that becomes also a matter of practical importance for forces that are stronger than electromagnetism.

Gell-Mann and Low immediately realized that the only reason that there can be any departure from a $1/r$ form for the potential is because the naive expectation that at large momentum transfer the electron mass should drop out of the problem is simply wrong. The potential does not have a smooth limit when r is very small compared to the Compton wavelength of the electron or, in other words, when the electron mass goes to zero. You can see from (1) that when the electron mass goes to zero the logarithm blows up. The failure of the naive expectation for the Coulomb force at short distances is entirely due to the fact that there is a singularity at zero electron mass. But where did that singularity come from? It's a

little surprising that there should be a singularity here. In fact, if you look at the Feynman diagram in which an electron loop is inserted in the exchanged photon line, you can see that the momentum transfer provides an infrared cutoff and, in fact, there's no way that this diagram can have a singularity for zero electron mass. What is going on here?

Gell-Mann and Low recognized that the singularity at the electron mass is entirely due to the necessity of renormalization, in particular of what is called charge renormalization. If you calculate the one-loop diagram using an ultraviolet cutoff at momentum Λ to make the integral finite then the formula you get before you go to any limit is something like this (simplified a little bit):

$$V(r) = \frac{\alpha}{r}\left[1 + \frac{2\alpha}{3\pi}\ln\left(\sqrt{\frac{1 + r^2\Lambda^2}{1 + r^2 m_e{}^2}}\right) + \cdots\right]. \tag{2}$$

This is, as expected, not singular as the electron mass goes to zero. Consequently the naive expectation that the potential should go like $1/r$ at short distances is indeed correct for (2): the potential approaches α/r. The potential also behaves like $1/r$ at very large distances, but here with a different coefficient:

$$V(r) \to \frac{\alpha}{r}\left[1 + \frac{2\alpha}{3\pi}\ln\left(\frac{\Lambda}{m_e}\right) + \cdots\right] \quad \left(\text{for } r \gg \frac{1}{m_e} \gg \frac{1}{\Lambda}\right). \tag{3}$$

But the electric charge is *defined* in terms of this coefficient, because we measure charge by observing forces at large distances. That is, if we want to interpret α as the observed value of the fine-structure constant, then in (2) we should make the replacement

$$\alpha \to \alpha\left[1 - \frac{2\alpha}{3\pi}\ln\left(\frac{\Lambda}{m_e}\right) + \cdots\right], \tag{4}$$

so that (2) becomes (to second order in α)

$$V(r) = \frac{\alpha}{r}\left[1 - \frac{2\alpha}{3\pi}\ln\left(\frac{\Lambda}{m_e}\right) + \frac{2\alpha}{3\pi}\ln\left(\sqrt{\frac{1 + r^2\Lambda^2}{1 + r^2 m_e{}^2}}\right)\right]. \tag{5}$$

Now we can let the cutoff Λ go to infinity, and we get (1) (aside from nonlogarithmic terms, which are not correctly given by the simplified formula (2)). The singularity at zero electron mass arises solely from the renormalization (4) of the electric charge.

That is the diagnosis—now what is the cure? This too was provided by Gell-Mann and Low. They advised that since the logarithm of the electron mass was introduced by a renormalization prescription which defines the

electric charge in terms of Coulomb's law at very large distances, we shouldn't do that; we should instead define an electric charge in terms of Coulomb's law at some arbitrary distance, let's say R; that is, we should define a renormalization-scale-dependent electric charge as simply the coefficient of $1/R$ in the Coulomb potential:

$$\alpha_R \equiv RV(R). \tag{6}$$

You might think that this wouldn't get you very far, but it does. Let's for a moment just use dimensional analysis, and not try to calculate any specific Feynman diagrams. If I set out to calculate the Coulomb potential at some arbitrary distance r, and I use as an input parameter the value of the fine structure constant α_R at some other distance R, then on dimensional grounds the answer must be a factor $1/r$ times a function of the dimensionless quantities α_R, r/R, and $m_e R$:

$$V(r) = \frac{1}{r} F\left(\alpha_R, \frac{r}{R}, m_e R\right). \tag{7}$$

Since we are expressing the answer in terms of α_R rather than $\alpha \equiv \alpha_\infty$, there should be no singularity at $m_e = 0$, and hence for r and R much less than $1/m_e$, the dependence on m_e should drop out here. Multiplying with r then gives our development equation for α:

$$\alpha_r = F\left(\alpha_R, \frac{r}{R}\right). \tag{8}$$

This is usually written as a differential equation $r d\alpha_r/dr = -\beta(\alpha_r)$, with $\beta(\alpha) \equiv -[\partial F(\alpha, x)/\partial x]_{x=1}$. However, it makes no difference in which form it is written; the important thing is that we have an equation for α_r in which $\alpha = 1/137$ and m_e do not enter, except through the initial condition that for $r = 1/m_e$, α_r is essentially equal to α.

This has remarkable consequences. First of all, one consequence which is not of stunning importance, but is useful: since 1/137 and the mass of the electron only enter together, through the initial condition, you can relate the number of logarithms to the number of powers of 1/137. For instance, we have seen that in the Coulomb potential to first order in 1/137 there is only one logarithm, and it can be shown that in second order there's still only one logarithm, in third order there are two logarithms, and so on. That's interesting. It is surprising that one can obtain such detailed information about higher orders with so little work, but what is really remarkable is what (8) says about the very short distance limit. In the limit of very short distances, there are only two possibilities.

First α_R may not have a limit as R goes to 0, in which case the conclusion would be that the bare charge is infinite and probably (although I can't say this with any certainty) the theory makes no sense. Such a theory probably develops singularities at very short distances, like the so-called ghosts or tachyons, which make the theory violate the fundamental principles of relativistic quantum mechanics.

The second possibility is that α_R does have a limit as R goes to 0, and the limit is nonzero, but since 1/137 enters in this whole business just as the initial condition on (8), this limit is, of course, independent of 1/137. By letting r and R both go to zero in (8) with arbitrary ratio x, you can see that the limit α_0 of α_r as $r \to 0$ is defined as the solution of the equation

$$\alpha_0 = F(\alpha_0, x) \quad \text{(for all } x\text{)}. \tag{9}$$

This limit is called a fixed point of the development equation. (Another way of expressing this is that α_0 is a place where the Gell-Mann–Low function $\beta(\alpha)$ vanishes.)

The one thing which isn't possible in quantum electrodynamics is that the limit of α_r as $r \to 0$ should be 0. Although we can't calculate the development function in general, we can calculate it when α_r is small, so we can look and see whether or not, if α_r is small, it will continue to decreases as r goes to 0. The answer is no, it doesn't. When α_r is small, it's given by (1) as

$$\alpha_r = \alpha_R \left[1 + \frac{2\alpha_R}{3\pi} \ln\left(\frac{R}{r}\right) + \cdots\right]. \tag{10}$$

You see that when r gets very small α_r does not decrease, it increases. Eventually it increases to the point where you can't use the power series any more; this happens at a distance of 10^{-291} cm. About what happens at such short distances, this equation tells you essentially nothing, but the one thing it does tell you for sure is that when r goes to 0, α_r does not go to 0, because if it did go to 0 then you could use perturbation theory, and then you would see it doesn't go to 0; so it doesn't.

This analysis gives information about much more than the short-distance behavior of the Coulomb potential. Consider any other amplitude, let's say, for the scattering of light by light. This will have a certain dimensionality, let's say length to the dth power. So write this amplitude as the renormalization scale to the dth power times some dimensionless function of the momenta $k_1, k_2, \ldots$ of the various photons, the electron mass, the renormalization scale R, and the fine-structure constant α_R at that renormalization scale. (Remember the idea. We're

defining the electric charge in terms of the Coulomb potential not at infinity but at some distance R.) That is, the amplitude A takes the form

$$A = R^d \mathcal{f}(k_1 R, k_2 R, \ldots, m_e R, \alpha_R). \tag{11}$$

In order to study the limit in which $k_1 = kx_1$, $k_2 = kx_2$, ... with x_1, x_2, ... fixed and the overall scale k going to infinity, it is very convenient to choose $R = 1/k$. No one can stop you from doing that. You can renormalize anywhere you want; the physics has to be independent of where you renormalize. Now there is no singularity here at zero electron mass, because we renormalizing not at large distances but at short distances; hence we can replace $m_e R$ by 0 in the limit $R \to 0$. With $R = 1/k$ and $k \to \infty$, the amplitude (11) has the behavior

$$A \to k^{-d} \mathcal{f}(x_1, x_2, \ldots, 0, \alpha_{1/k}). \tag{12}$$

The factor k^{-d} is what we would expect from naive dimensional analysis, ignoring problems of mass singularities or renormalization. Aside from this, the asymptotic behavior depends entirely on the behavior of the function α_r for $r \to 0$. In particular if α_r approaches a finite limit as $r \to 0$, then the amplitude does exhibit naive scaling for $k \to \infty$, but with a coefficient of k^{-d} that is not easy to calculate. (There are complications here that I have left out, having to do with matters like wavefunction renormalization. The above discussion is strictly valid only for suitably averaged cross sections. However the result of naive scaling for α_0 finite is valid for purely photonic amplitudes. For other amplitudes, there are corrections to the exponent d.)

Now this is really amazing—that one can get such conclusions without doing a lot of difficult mathematics, without really ever trying to look at the high orders of perturbation theory in detail. Nevertheless, the paper by Gell-Mann and Low suffered a long period of neglect—from 1954, when it was written, until about the early 1970s. There are a number of reasons for this; let me just run through what I think were the important ones.

First of all, there was a general lack of understanding of what it was that was important in the Gell-Mann–Low paper. There had been a paper written the year before Gell-Mann and Low, by Stueckelberg and Petermann, which made the same remark Gell-Mann and Low had made, that you could change the renormalization point freely in a quantum field theory, and the physics wouldn't be affected. Unfortunately, when the book on quantum field theory by Bogoliubov and Shirkov was published in the late 1950s, which I believe contained the first mention in a book of

these matters, Bogoliubov and Shirkov seized on the point about the invariance with respect to where you renormalize the charge, and they introduced the term "renormalization group" to express this invariance. But what they were emphasizing, it seems to me, was the least important thing in the whole business.

It's a truism, after all, that physics doesn't depend on how you define the parameters. I think readers of Bogoliubov and Shirkov may have come into the grip of a misunderstanding that if you somehow identify a group that then you're going to learn something physical from it. Of course, this is not always so. For instance when you do bookkeeping you can count the credits in black and the debits in red, or you can perform a group transformation and interchange black and red, and the rules of bookkeeping will have an invariance under that interchange. But this does not help you to make any money.

The important thing about the Gell-Mann–Low paper was the fact that they realized that quantum field theory has a scale invariance, that the scale invariance is broken by particle masses but these are negligible at very high energy or very short distances if you renormalize in an appropriate way, and that then the only thing that's breaking scale invariance is the renormalization procedure, and that one can take that into account by keeping track of the running coupling constant α_R. This didn't appear in the paper by Stueckelberg and Petermann, and it was pretty well submerged in the book by Bogoliubov and Shirkov. I say this with some bitterness because I remember around 1960 when that book came out thinking that the renormalization group was pretty hot stuff, and trying to understand it and finding it just incomprehensible and putting it away. I made the mistake of not going back and reading carefully the paper by Gell-Mann and Low, which is quite clear and explains it all very well. (Incidentally, the later textbook by Bjorken and Drell gave a good clear explanation of all this, following the spirit of the Gell-Mann–Low paper.)

The second reason, I think, for these decades of neglect of the Gell-Mann–Low paper was the general distrust of quantum field theory that set in soon after the brilliant successes of quantum electrodynamics in the late 1940s. It was realized that the strong interactions were too strong to allow the use of perturbation theory and the weak interactions did not seem to have the property that the electromagnetic interactions did, of being renormalizable. (Renormalizability means that you can have a Lagrangian or a set of field equations with a finite number of constants, and all the infinities can always be absorbed into a redefinition of the

constants, as I've already shown here that you can do with the cutoff dependence of the Coulomb potential.) Since people were not all that enthusiastic about quantum field theory, it was not a matter of high priority to study its properties at very short distances. Finally, we have seen, in quantum electrodynamics the Gell-Mann–Low analysis itself tells you that perturbation theory fails at very short distances, and then you just have to give up. There didn't seem to be much more that one could do.

The great revival of interest in the renormalization group came in the early 1970s, in part from a study of what are called anomalies. Anomalies are things that happen in higher orders of quantum field theory that you don't expect and that don't appear when you use the field equations in a formal way. I guess you could say the anomalies represent an instance of the Second Law of Progress in Theoretical Physics, which can be stated: *Do not trust arguments based on the lowest order of perturbation theory.* Some of these anomalies were studied here at MIT by Jackiw and Bell and Johnson and Low, and at Princeton by Steve Adler. In 1971 Callan, Coleman, and Jackiw were studying the scaling behavior of higher-order contributions to scattering amplitudes, and found as Gell-Mann and Low had found earlier in a different context that these amplitudes did not have the sort of "soft" nonsingular dependence on particle masses as the lowest-order contribution. A little later, Coleman and Jackiw traced this failure of naive scaling to an anomaly in the trace of the energy-momentum tensor. In the limit of zero masses one would expect the trace of the energy-momentum tensor to vanish. (For hydrodynamics, for instance, the trace of the energy-momentum tensor is 3 times the pressure minus the density. And everyone knows that for massless particles like light, the pressure is 1/3 the energy density. So you should get zero.) And, in fact, in quantum electrodynamics you do get zero if you just use lowest-order perturbation theory, in the limit where the electron mass is zero—but even with the electron mass equal to zero, if you calculate matrix elements of the energy-momentum tensor beyond the lowest-order perturbation theory you find that its trace is not zero. At about the same time, Callan and Symanzik set up a formalism for studying the failure of naive scaling. Their results turned out to look very much like the Gell-Mann–Low formalism. With the benefit of hindsight, this should not be surprising at all because, as I have emphasized here, the essential point of Gell-Mann and Low was that naive dimensional analysis breaks down precisely because of renormalization. The fact that the Coulomb potential is not just proportional to $1/r$ at short distances is one symptom of this breakdown of scale invariance, and the nonvanishing of the trace of the energy-momentum tensor is another symptom. The formalism used for one is related to the formalism used for the other.

Another theoretical influence: in the early 1970s non-Abelian gauge theories began to be widely studied, both with regard to the electroweak interactions and soon also with regard to the strong interactions. Politzer and Gross and Wilczek realized that the plus sign in the logarithmic term of (10), which prevented the use of perturbation theory in quantum electrodynamics at short distances, for non-Abelian gauge theories is a minus sign. The important thing about non-Abelian gauge theories for these purposes is that instead of one photon you have a family of "photons," and each member of this family of "photons" carries the "charge" that other members interact with. The prototypical non-Abelian gauge theory is that of Yang and Mills, in which there are three "photons." Because the "photons" interact with "photons," in addition to the usual diagrams for the "Coulomb" potential where you have loops of fermions like electrons inserted into exchanged "photon" lines, here you also have "photon" loops, and these have opposite sign. In fact, not only do they have opposite sign but they're bigger. In place of the characteristic factor of 2/3 in (10), each "photon" loop carries a factor of $-11/3$. In the theory of strong interactions the fermions are quarks and there are 8 "photons" known as gluons. So unless you have an awful lot of quarks, the gluons are likely to overpower the quarks and give the logarithm in (10) a large negative coefficient, while in quantum electrodynamics you find a positive one. This makes all the difference because it means that as you go to shorter distances the forces get weak rather than getting strong and you can then use perturbation theory at very short distances. This is called asymptotic freedom. Politzer, Gross, and Wilczek instantly realized that this explains an experimental fact which had been observed in a famous experiment on deep inelastic electron proton scattering done by an MIT-SLAC collaboration in 1968. This was that at very high momentum transfer, in other words, at very short distances, the strong interactions seem to turn off and the formulas for the form factors in deep inelastic electron scattering seem to obey a kind of naive scaling, "Bjorken scaling." This had been a mystery because it would require that somehow or other the strong interactions must disappear at short distances. It had been this result that in part had stimulated all this theoretical work on scaling. Now suddenly this was understood.

Also, if the force gets small as one goes to short distances, there's a least a good chance that it will get big as you go to large distances. At

first it was generally supposed that this did not happen. It was assumed that the "photons" here are heavy, getting their mass (like the intermediate vector bosons of the weak interactions) from the vacuum expectation values of scalar fields. But scalar fields would have raised all sorts of problems for the theory. Then Gross and Wilczek and I guessed that there are no strongly interacting scalars; that the gluons, the strongly interacting "photons," are therefore massless; that consequently the force does continue to increase with distance; and that this might explain why we don't observe the gluons, and also why we don't observe the quarks. Putting together all the pieces, at last we had a plausible theory of the strong interactions. It was christened (by Murray Gell-Mann, who with Fritsch and Minkowski had developed some of these ideas before the discovery of asymptotic freedom) quantum chromodynamics, that is, the same as quantum electrodynamics except that the quantity called color replaces electric charge.

There's an interesting side to the history of all this. Ken Wilson, perhaps alone of all theoretical physicists, was well aware of the importance of using the renormalization group ideas of Gell-Mann and Low through the late 1960s and early 1970s. He used these ideas to consider all kinds of interesting things that might happen at high energy. He considered, for example, the possibility that the coupling constant would go to a nonzero fixed point, which is exactly what Gell-Mann and Low thought might happen in quantum electrodynamics, or that we might find a limit cycle where the coupling constant goes round and round and just keeps oscillating in a periodic way. He wrote papers about how this would appear from various points of view experimentally, whereas the experimentalists at the same time were showing that, in fact, everything is very simple—that at high energies the strong interactions go away altogether. To the best of my knowledge, Ken Wilson missed only one thing—the possibility that the coupling constant might go to zero at short distances. He just didn't consider that possibility because he knew it didn't happen in quantum electrodynamics. On the other hand, Tony Zee was very much aware of that possibility, and wrote a paper saying, wouldn't it be simply grand if the coupling constant did go to zero at high energy, then we could understand the MIT-SLAC experiment. He sat down and calculated the logarithmic terms in the vacuum polarization effect in various theories and he found he got the plus sign, the one that you get in quantum electrodynamics, in all his calculations, and gave up in disgust. The one case he did not consider was the case of a non-Abelian gauge theory like the Yang-Mills theory. The reason that he didn't consider it

was because at that time the rules for calculating those theories, with Fadeev-Popov ghosts and all the rest of the boojums, were not very widely known and he didn't feel confident in doing the calculation. So he gave up the idea. On the other hand, Gerard 't Hooft, who knows everything about how to calculate in a non-Abelian gauge theory, did this calculation and, in fact, found that the sign factor in the Gell-Mann–Low function was opposite to what it is in quantum electrodynamics. He announced the result of this calculation at a conference on gauge theory at Marseille in June 1972, but he waited to publish it while he was doing other things, so his result did not attract much attention.

Finally, however, it did all come together. From 1973 on, I would say, most theorists have felt that we now understand the theory of the strong interactions. It is, of course, very important to test this understanding, and I certainly wouldn't claim that quantum chromodynamics is indisputably verified. My own feeling is that quantum chromodynamics *will* be indisputably verified in machines like LEP, in which electron-positron annihilation produces jets of quarks and antiquarks and gluons, and that this verification will be very much like the verification of quantum electrodynamics, not in the 1940s when the problem was the loop graphs, but in the 1930s when quantum electrodynamics was verified for processes like Bhabha scattering and Møller scattering and Compton scattering, using only tree diagrams. I say this in part because of a theorem, that if you calculate the cross section not for producing a certain definite number of quarks or gluons but instead for producing a certain definite number of quark or gluon jets (a jet being defined as a cone within which there can be any number of particles) then these cross sections satisfy the assumptions of the Gell-Mann–Low paper, that in the limit of very short distance or very high energy they can simply be calculated by perturbation theory. The other case in which one would like to verify quantum chromodynamics is, of course, at large distances or low energy, where the question of quark trapping arises. We'd like to be able to calculate the mass of the proton, the pion-nucleon scattering at 310 MeV, and all sorts of other quantities. Many people are working on this very difficult problem. I will come back at the end of my talk to one idea about how this kind of calculation might be done.

The wonderful discovery by Politzer, Gross, and Wilczek of the decrease of the strong interactions at high energy also had an immediate impact on our understanding of the possibilities for further unification. Ideas about unifying the strong and electroweak interactions with each other have been presented in papers by Pati and Salam, Georgi and Glashow, and

many others. However, there was from the start an obvious problem with any such idea: strong interactions are strong and the others aren't. How can you unify interactions that have such different coupling constants? Once quantum chromodynamics was discovered, the possibility opened up that because the strong interactions, although strong at ordinary energies, get weak as you go to high energy or short distances, at some very high energy they fuse together with the electroweak interactions into one family of "grand unified" interactions. This idea was proposed in 1974 by Georgi, Quinn, and me, and we used it to calculate the energy at which the strong and electroweak couplings come together. After my earlier remarks, it should come as no surprise to you that the energy that we found is an exponential of an inverse square coupling constant, like the energy that Gell-Mann and Low found where electromagnetism would become a strong interaction. (They expressed this in terms of distances, but it's the same thing, except for taking a reciprocal.) Instead of the Gell-Mann–Low energy of $\exp(3\pi/2\alpha)$ electron masses, we found that (in a large class of theories) the strong and electroweak forces come together at an energy which is larger than the characteristic energy of quantum chromodynamics by a factor roughly $\exp(\pi/11\alpha)$. (The 11 is that magic number I mentioned earlier that is always contributed by a loop of gauge bosons.) This factor, in other words, is something like the 2/33 power of the enormous factor that corresponds to the incredibly short distance at which Gell-Mann and Low found that perturbation theory begins to break down in quantum electrodynamics. The energy here turns out to be something still very high but not so inconceivably high, only about 10^{15} GeV. This suggests that there's a whole new world of physics at very high energies of which we in studying physics at 100 GeV or thereabouts are only seeing the debris.

There may be all sorts of new physical effects that come into play at 10^{15} GeV. For example, there's no real reason to believe that baryon number would be conserved at such energies. The fact that it is conserved at ordinary energy can be understood without making any assumption about baryon conservation as an exact symmetry of nature. We might expect a proton lifetime of the order of magnitude of $(10^{15}\ \text{GeV})^4/(\alpha^2 m_p{}^5)$, essentially as estimated in the paper by Georgi, Quinn, and me. This comes out to be about 10^{32} years, which is nice because it's a little bit beyond the lifetimes that have been looked for so far experimentally, but not hopelessly beyond them. Of course, we are all anxious to find out whether or not the proton does decay with some such lifetime.

After the strong and electroweak interactions have hooked up with

each other, what happens then? Does the grand unified interaction, which then would have only one independent coupling constant, satisfy the idea of asymptotic freedom, that the coupling constant goes on decreasing? Or does the coupling start to rise, presenting us back again with the same problem that Gell-Mann and Low faced, of a coupling constant which increases as you go to short distances or high energies and, therefore, ultimately makes it impossible to use perturbation theory. And, of course, at 10^{15} GeV you're not very far below the energy at which gravity becomes important. Perhaps that cancels all bets.

In addition to the applications of the renormalization group to the strong interactions and thence to grand unified theories, there had even a little earlier been an entirely different development due to Ken Wilson and Michael Fisher and Leo Kadanoff and others—the application of renormalization-group methods to critical phenomena. It is interesting that in this volume there are two papers that deal with fixed points and the renormalization group and so on. The first of these, by Ken Johnson, is entirely about quantum field theory. The second one, by Mitchell Feigenbaum, is entirely about statistical phenomena. In fact, there seems to be no overlap between these papers except for the language of the renormalization group.

I think it is really surprising that the same ideas can be applied to such apparently diverse realms. When you're dealing with critical phenomena, you're not concerned about short-distance (or high-energy) behavior; you're concerned about long-distance behavior. You're asking about matters like critical opalescence, about the behavior of the correlation function when two points go to very large separation, not very short separation. Well, that alone is perhaps not such an enormous difference. After all even in quantum electrodynamics you might be interested in such questions, not in the real world where the electron mass provides an infrared cutoff which makes all such questions irrelevant, but say in a fictitious world where the electron mass is zero. If the electron mass really were zero, it would be very interesting to say what happens to quantum electrodynamics at very long distances. The Gell-Mann–Low formalism answers that question. At very long distance, massless quantum electrodynamics becomes a free field theory. In quantum chromodynamics all we know for sure is that it does *not* become a free field theory, just the reverse of what we know about the short-distance behavior of these theories. Now, when you're talking about critical phenomena there is something analogous to the mass of the electron—there's the difference between the actual temperature and the critical temperature. The critical

temperature at which a second-order phase transition occurs is defined in such a way that at that temperature there's nothing that's providing an infrared cutoff, and, therefore, for example, correlation functions don't exponentially damp as you go to very large separations. So, in other words, setting the temperature equal to the critical temperature in a statistical mechanics problem is analogous to studying what happens in quantum electrodynamics when you actually set the electron mass equal to zero and then consider what happens as you go to very large distances. Of course, we can't dial the value of the electron mass. We can, however, set thermostats, so there are things that are of interest in statistical mechanics that aren't of that much interest in quantum field theory, because the value of the temperature really is at our disposal. When you look at it from this point of view you can see the similarity between what people who work in critical phenomena are doing and what Gell-Mann and Low did. They're all exploiting the scale invariance of the theory, scale invariance, that is, except for the effects of renormalization, and corrected by the Gell-Mann–Low formalism for the effects of renormalization.

There is another difference between high energy particle physics and statistical physics. After all, ordinary matter is, in fact, not scale invariant. Where does scale invariance come from when you're talking about critical opalescence in a fluid going through a phase transition? In what sense is there any scale invariance with or without renormalization corrections? If you construct a kind of field theory to describe what's happening in a fluid, in which the field ϕ might be a pressure or density fluctuation of some kind, the Hamiltonian would include a huge number of terms, ϕ^2, ϕ^4, ϕ^6, ... because there's no simple principle of renormalizability here that limits the complexity of the theory. It doesn't look like a scale-invariant theory at all. Well, in fact, you can show that if you're interested in the long-distance limit then all the higher terms such as ϕ^6, ϕ^8, etc., become irrelevant. The ϕ^2 term also would break scale invariance, but this is precisely the effect we eliminated by going to the critical temperature. Finally, the ϕ^4 term also breaks scale invariance (in classical statistical mechanics its coupling constant has the dimensions of a mass); but this is taken care of by the same renormalization group manipulations that are needed anyway to deal with renormalization effects. If C is a function with dimensionality d that describes correlations at separation r, then at the critical temperature dimensional analysis gives

$$C = R^d F(r/R, R\lambda(R)), \tag{13}$$

where $\lambda(R)$ is the ϕ^4 coupling constant, defined by some renormalization prescription at a scale R. Once again, set R equal to r; (13) then becomes

$$C = r^d F(1, r\lambda(r)). \tag{14}$$

Furthermore, the dimensionless quantity $r\lambda(r)$ satisfies a Gell-Mann–Low equation like (8). If this quantity approaches a fixed point for $r \to 0$, then (14) indicates that we have naive scaling ($C \propto r^d$) for $r \to 0$. (Once again, I am ignoring complications having to do with the renormalization of the field ϕ, or equivalently of the operator $\partial_\mu \phi \partial^\mu \phi$. These change the value of the power of r as $r \to 0$.)

There is still something mysterious about all this, which takes me back to my starting question: Why is the renormalization group a good thing? What in the world does renormalization have to do with critical phenomena? Renormalization was invented in the 1940s to deal with the ultraviolet divergences in quantum field theory. Theories of condensed matter are not renormalizable field theories. They don't look like quantum electrodynamics at all. If you throw away the higher terms in the Hamiltonian (ϕ^6, ϕ^8, etc.) on the grounds that you're only interested in long-distance behavior (these terms are what in statistical mechanics are called irrelevant operators), then you're left with a theory that doesn't have any need for renormalization to eliminate ultraviolet divergences. (This is because when you deal with critical phenomena you're working with 3 and not 4 dimensions.) But then why does the use of the renormalization group help at all in understanding critical phenomena?

I think the answer to the last question gets us to essence of what really is going on in the use of the renormalization-group method. The method in its most general form can I think be understood as a way to arrange in various theories that the degrees of freedom that you're talking about are the relevant degrees of freedom for the problem at hand. If you renormalize in the conventional way in quantum electrodynamics in terms of the behavior of the Coulomb potential at large distances, then for any process like scattering of light by light you will have momenta running around in the Feynman diagram which go down to small values, small meaning of the order of the electron mass. Even if what you're really interested in is the scattering of light by light at 100 GeV, when you calculate the Feynman diagram you'll find that the integrals get important contributions from momenta which go all the way down to one-half MeV, the electron mass. In other words, the conventional renormalization scheme in quantum electrodynamics, although it does not actually introduce any mistakes, emphasizes degrees of freedom which, when

you're working at very high energy, are simply not the relevant degrees of freedom. The Gell-Mann–Low trick of introducing a sliding renormalization scale effectively suppresses those low-energy degrees of freedom in the Feynman integrals. If you define a renormalization scheme, so that when you calculate scattering of light by light at 100 GeV you use a definition of the electric charge which is renormalized at 100 GeV, then you will in fact find that all of the Feynman integrals you have to do get their important contributions from energies roughly of order 100 GeV. In other words, the Gell-Mann–Low procedure gets the degrees of freedom straight. The same is true in the renormalization-group approach to critical phenomena, whether you implement it as Wilson did by simply integrating out the very short wave numbers, or if you do what Brezin, LeGuillou, and Zinn-Justin do and use the renormalization scheme itself to provide an ultraviolet cutoff in close analogy to the Gell-Mann–Low approach to field theory. Either way, you are arranging the theory in such a way that only the right degrees of freedom, the ones that are really relevant to you, are appearing in your equations. I think that this in the end is what the renormalization group is all about. It's a way of satisfying the Third Law of Progress in Theoretical Physics, which is that *you may use any degrees of freedom you like to describe a physical system, but if you use the wrong ones, you'll be sorry*.

Now let me briefly come to some possibilities for future developments. We still have with us the problem of quantum chromodynamics at very large distances. This is a somewhat paradoxical problem because in fact for a long time we have had a perfectly good quantum field theory for strong interactions at very large distances. For simplicity, I will adopt here the fiction that the bare quark masses are zero, which for many purposes is a good approximation. In that case, the pion is massless because it's a Goldstone boson. The Lagrangian that describes strong interactions at a very low energy like 1 eV, where the only degree of freedom is the massless pion, is the nonlinear Lagrangian which was originally written down by Gell-Mann and Levy in 1960, and which as I showed in 1967 actually reproduces all the theorems of current algebra. The Lagrangian is

$$\mathscr{L} = -\partial_\mu \boldsymbol{\pi} \cdot \partial^\mu \boldsymbol{\pi} / (1 + \boldsymbol{\pi}^2 / F_\pi^2)^2, \tag{15}$$

where $\boldsymbol{\pi}$ is the pion field, and F_π is an empirically determined constant, about 190 MeV. This then is the field theory of the strong interactions at very low energy, always with the proviso that the bare quark masses are zero. (It's not much more complicated otherwise.) So we have a perfectly

good field theory for strong interactions at low energies, and we also have a perfectly good field theory for strong interactions at very high energies, the quantum chromodynamics in which we all believe. The question is not so much how we can solve the strong interactions at low energy, or at large distances, as how we can prove that there's any connection between these two theories. How can we prove that if you start with quantum chromodynamics which we think is, in some sense, an underlying theory, that then if you then treat it in the limit of very long distances or low energies you go over to the theory described by (15)?

I wonder if the answer is not that we should expand once again our idea of what the renormalization group means. To me the essence of the renormalization-group idea is that you concentrate on the degrees of freedom that are relevant to the problem at hand. As you go to longer and longer wave lengths you integrate out the high-momentum degrees of freedom because they're not of interest to you and then you learn about correlation functions at long distances; or, vice versa, you do what Gell-Mann and Low did, and as you go to shorter and shorter wave lengths you suppress the long wavelengths. But sometimes the choice of appropriate degrees of freedom is not just a question of large or small wavelength, but a question of what kind of excitation we ought to consider. At high energy the relevant particles are quarks and gluons. At low energy they're massless pions. What we need is a version of the renormalization group in which as you go from very high energy down to low energy you gradually turn on the pion as a collective degree of freedom, and turn off the high-energy quarks. Now I don't really know how to do that. I do have some ideas about it. There are ways of introducing fields for particles like the pion which are not elementary, and then making believe that they are elementary. The question is whether the dynamics generate a kinematic term for $\boldsymbol{\pi}$ in the Lagrangian. I'm working on this and certainly have no progress to report. I have asked my friends in statistical mechanics whether or not when they use renormalization-group ideas they find that they have to not only continually change the wavelength cutoff but actually introduce new degrees of freedom as they go along. Apparently this has not been done in statistical mechanics. Collective degrees of freedom, like say the Cooper pair field in superconductivity, are just introduced at the beginning of the calculation and are not turned on in a smooth way as you go to long wavelengths. But perhaps this readjustment of degrees of freedom might be useful also in statistical mechanics.

Finally, I want to come to what is perhaps the most fundamental question of all: What is the behavior of nonrenormalizable theories at

short distances? This is an important problem above all because so far no one has succeeded in embedding gravity into the formalism of a renormalizable quantum field theory. As far as we know, the Lagrangian for gravity, in order to cancel all infinities, has to be taken to have an infinite number of terms, in fact all conceivable terms which are allowed by general covariance and other symmetries. For instance, for pure gravity the Lagrangian must be taken as

$$\mathscr{L} = \frac{1}{16\pi G}R + fR^2 + f'R^{\mu\nu}R_{\mu\nu} + hR^3 + \cdots . \tag{16}$$

(I've written here only terms involving the metric but in reality there are an infinite number of terms involving matter as well.) This is not at all in contradiction with experiment; the success of Einstein's theory does not contradict this. The leading term, the R term, has a coefficient of about 10^{38} GeV^2; that is the square of the Planck mass. If we believe that this is the only unit of mass in the problem then the coefficients f and f' in the next two terms are of order 1; the coefficients h, etc., in the next few terms are of order 10^{-38} GeV^{-2}; and so on. Any experiment which is carried out at distances large compared to 10^{-19} GeV^{-1} (which, of course, all experiments are) would only see the R term. So we don't know anything experimentally about the higher terms in (16). There's no evidence for or against them except that if gravity isn't renormalizable they would all have to be there.

What would be the short-distance or the high-energy behavior of such a theory? Well, suppose we make a graph in coupling-constant space showing the trajectory of the coupling constants G, f, f', h, etc., as we vary the renormalization scale. The renormalization group applies here; a theory doesn't have to be renormalizable for us to apply the renormalization-group method to it. These trajectories simply describe how all the couplings change as you go from one renormalization scale to another. Now many of those trajectories—in fact, perhaps most of them—go off to infinity as you go to short-distance renormalization scales. However, it may be that there's a fixed point somewhere in coupling-constant space. A fixed point, remember, is defined by the condition that if you put the coupling constant at that point it stays there as you vary the renormalization scale. Now, it is a fairly general phenomenon that for each fixed point there are usually some trajectories that hit the point, but these trajectories do not fill up much of coupling-constant space. That is, there may be some trajectories that you can draw that run into a given fixed point, but the surface that these trajectories map out is usually finite-

dimensional, whether the theory has an infinite number of couplings or not.

There's even experimental evidence for this property of fixed points. In fact, the whole lore of second-order phase transitions in a sense can be quoted as experimental evidence for this statement. In second-order phase transitions, where you're considering not the behavior at short distances but at large distances, this statement translates into the statement that the *normals* to the surfaces of trajectories (now going the other way!) that hit a given fixed point form a finite-dimensional set. That is why in statistical mechanics, for example, if you want to produce a second-order phase transition, you only have to adjust one or a few parameters, so that the coupling constants have no components along these normals. Water is an extremely complicated substance, with a huge number of parameters describing all its molecules, but if you want to produce a second-order phase transition in water, all you have to do is adjust the temperature and the pressure; you don't also have to adjust the mass of the water molecules or the various force constants. This means that the surface formed by the trajectories which are attracted by the fixed point as you go to very long distances has only two independent normals. If you go to short distances instead, then that statement translates into the statement that the space of trajectories that are attracted to the fixed point is only 2-dimensional.

If the parameters of a theory lie on a trajectory that hits a fixed point at short-distance renormalization scales, then the physical amplitudes of the theory may be expected to behave smoothly at short distances or high energies—often just a power-law behavior, perhaps with anomalous exponents. The behavior of such a theory is just like that found by Gell-Mann and Low for quantum electrodynamics with an ultraviolet fixed point. On the other hand, one may suspect that a theory which is on a trajectory which does *not* hit any fixed point is doomed to encounter a Landau ghost or a tachyon or some other terrible thing. Then you have a reason for believing that nature has to arrange the infinite number of parameters in a nonrenormalizable field theory like the theory of gravity so that the trajectories do hit the fixed point. This would leave only a finite number of free parameters. Indeed, conceivably this finite-dimensional surface is only 1-dimensional—conceivably it's just a line running into the fixed point. In this case we would have a physical theory in which the demands of consistency, the demands of unitarity and analyticity and so on which rule out ghosts and tachyons, dictate all the parameters of the theory, except for one scale parameter which just specifies the unit of length. What could be better?