FUNDAMENTAL AS FEWER BITS

Terry Bollinger
January 22, 2018

Abstract

In this essay I propose that a subtle but universal principle of binary conciseness *lies behind the most powerful and insightful equations in physics and other sciences. The principle is that if two or more precise descriptive models (theories) address the same experimental data, the theory that is more concise in terms of Kolmogorov complexity will also be more fundamental in the sense of having the deepest insights.*

The World's Most Famous Equation

It is arguably the best known and most often quoted equation in the world:

$$E = mc^2$$

Its dual association with the genius of Albert Einstein [1] and the power of nuclear energy makes it unforgettable. It is the only physics equation that can be heard regularly in non-technical conversations, where it is used as shorthand to refer to levels of intellectual insights far beyond everyday norms. Its message is easy to understand and deeply unexpected: Stodgy, static mass and dynamic, moving energy are in some mysterious way two aspects of a single quantity.

Another compelling feature of $E = mc^2$ is its conciseness. It needs only three letters, a number, and two high-school level algebraic operations. It is so concise that it was easier for me to embed it in the previous paragraph than to reference it as a separate figure. It is also concise at the concept level, altering the perspective of its more avid recipients until they reflexively interpret mass and energy as a single pool of resources rather than as two separately tracked issues.

Fundamental Means Mathematically Concise

The claim of this essay is that the conciseness of $E = mc^2$ is not an accident, but an example of a subtle but universal *principle of binary conciseness*. The principle is this: If two or more precise descriptive models (theories) address the same data, then the theory that is more concise in terms of *Kolmogorov complexity* [2] is more fundamental. That is, the theory whose form is closest to the *Kolmogorov minimum* will provide the deepest, most insightful understanding of that topic.

Kolmogorov Complexity as Data Compression

So what then is Kolmogorov complexity? While it is not usually presented as such, Kolmogorov complexity is simply a mathematically precise model for how to do lossless data compression.[3] Its precision is made possible by requiring that the compressed file be written and interpreted as a computer program, one which when executed exactly reproduces the original data set. The *Kolmogorov minimum* (or *descriptive minimum*) is simply the maximally compressed form of the original file, that is, the shortest computer program capable of reproducing the original data set.

The idea of compressing data by converting it into a computer program sounds exotic, but in fact is so easy and commonplace that people do it without realizing it. For example, if someone asked you to write a program to replace a gigabyte file containing one billion copies of the bit string 01001101, what would you suggest? You just read one possible answer: "one billion copies of the bit string 01001101". Even in English this phrase is an unambiguous executable program capable of reproducing the original file. Since it uses only 45 characters or bytes, the result is a compression ratio of over 22 million to 1. A very modest effort to shorten this program even further gives "10^9*01001101" with a length of 13 bytes, which achieves a compression ratio of almost 77 million to one.

Two observations are worth making about such Kolmogorov programs. The first is that *factoring* is one of the most common and fundamental ways to compress data, though not the only one. Factoring looks for "something" that occurs multiple times in a data sequence. This allows it to be represented just once, with pointers to the other locations. The "something" is not necessarily as simple or obvious as a string-level repetition of bits, however. The second observation is that when the data file becomes more complex, the process of factoring it into smaller pieces also becomes more complex. This is why commercial data compression programs require more time and computer resources to achieve higher levels of data file compression.

Compression as an Exponentially Difficult Asymptotic Limit

Given how easy it was to make the earlier gigabyte file program more concise, it seem likely that it could be compressed even more. Alas, this is where the "shortest possible program" aspect of Kolmogorov reduction starts getting tricky. Andrey Kolmogorov in fact proved that it is *undecidable* whether any given program is "the" shortest compression possible.

The problem is that there is no way to prove that some unexplored domain of math or logic might not provide even greater compression. For example, the 20 digit

sequence 66276378959982317513 looks fully random and thus uncompressible. However, a broader search of mathematics shows this sequence is a substring of π occurring at position 39,025,353. Thus the 20 digit sequence could in principle be replaced by a short binary program that generates and indexes π.

Such examples suggest a strategic rule of thumb for extreme compression:

> *The closer you get to the Kolmogorov minimum, the more you will need to explore novel data transformations and mathematical options.*

There was nothing in the 20 digit sequence that hinted that it might be part of π, yet the definition of π turns out to be its best option for compression. The only way to find such cryptic opportunities is to broaden the search. Similarly, as the search for repetition becomes harder, it will become necessary to transform the data in new ways that bring out hidden connections and redundancies.

Pragmatically, all of this translates into a simple but unfortunate rule of thumb:

> *The closer you get to the Kolmogorov minimum, the higher the cost will be in terms of resources needs, time used, creativity required, and failure rates.*

If theories are messages, then these heuristics provide a different way of looking at why it is hard to derive new, truly fundamental theories. Fundamental theories are well-factored theories, and so are likely to be near their Kolmogorov limits. This in turn means that any further reductions of their size and complexity usually require novel interpretations, more resources, and greater tolerance for failure. Given these counterincentives, it will always be easier to stop searching at a point where the theory is compact, but still relatively far from its Kolmogorov minimum.

Physics as Information Theory

The universe indisputably possesses a wide range of well-defined structures and behaviors that exist independently of human knowledge and actions. In a nutshell, the role of science is first to identify such pre-existing structures and behaviors, and then to document them in sufficient detail to understand and predict how they work. The resulting data sets can be massive and complex, but ultimately they are just data and thus qualify as messages. I will at times refer to them as *foundation messages* to emphasize that their content must reflect only content from the as-is universe, despite the extensive work that humans must perform to obtain them.

It is worth pointing out that it is in their handling of foundation messages that mathematics and physics are most sharply distinguished from each other. In mathematics the sole criterion for whether a theorem is correct is whether it can be derived from a small set of previously agreed-to axioms. In physics the sole criterion for whether a theory is correct is whether it accurately reproduces the data in foundation messages. Unlike an axiom set, that data can contain disturbing and unexpected surprises that are contrary to the axioms of otherwise closely related mathematical methods. One particularly notable example of axiomatic instability in

physics occurred when Einstein found it necessary to violate Euclid's fifth "parallel lines" postulate to create the curved spacetime of General Relativity.[4]

While the need for physics to predict and replicate data in foundation messages distinguishes it from axiomatic mathematics, it simultaneously makes physics more compatible with the perspectives of information theory. This is particularly true for the Kolmogorov version of information theory, which focuses specifically on the issue of ensuring that the compressed, program-style forms versions of messages reproduce the original data exactly and completely.

The implication is that a better way to think of physics is not as some form of axiomatic mathematics, but as a type of information theory. The universe in this interpretation behaves like the message-sending equivalent of a reluctant witness: It has all the data needed to build valid scientific theories, but it releases that data only very grudgingly and only through the active efforts of those interested in receiving such foundation messages. The highly compressed Kolmogorov program encodings of foundation messages become the theories of physics, with the level of insight of the theory corresponding in a surprisingly direct fashion to how well it has converted raw data into compact, program-like code and equations.

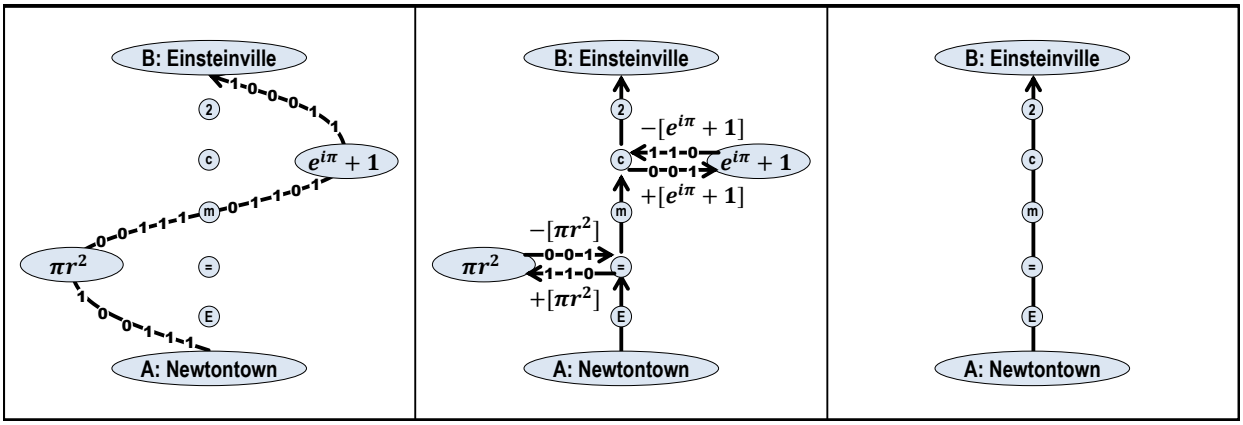
More subtly, data compression and theory development share more features than is typically realized. Both strive to create compact, singular formal representations of structures and behaviors that are otherwise spread across the uncompressed data. Both strive to find and eliminate irrelevant factors that obfuscate patterns. The searches in data compression for factorable identical units can also be described as searches for symmetries, that is, for situations in in which some mathematical transformation of the data results in an identical pattern. Modern particle physics relies intensely on the mathematics of symmetry groups,[5] but the resulting symmetries are also akin to data compression in the sense that they often point to an underlying shared pattern (or even particle) that allows the symmetry to exist.

So, if physics really is a specialized form of information theory, and its experimental data sets are its uncompressed messages from the universe at large, how are these data sets actually transformed into meaningful theories? How can manipulating bit strings be reinterpreted as theory building?

Messages as Maps

An answer to that question requires a better understanding of what message are. The defining feature of a message is that it *changes the state* of the recipient. That is, a message must alter the way in which the receiver behaves or reacts in some current or future situation. Conversely, if such a change in the state of the recipient does *not* occur, it means that message was either never received or discarded after receipt. Since states are often represented as locations in some abstract, multi-dimensional space, a simple map analogy turns out to be a good way to visualize and explain states, state changes, and the various properties of the messages.

Kolmogorov Reduction as Map Simplification



Imagine your task is to create a map (message) that tells the receiver how to move from point A to point B. The simplest and shortest path will be a straight line from A to B, but you could also draw curved paths that traverse other locations on their way from A to B. The first box in the map figure is an example. The curves in such paths amount to "side trips" to locations that are irrelevant to reaching the final destination. Note that while there is only one straight line path, infinitely many curved paths are possible. These longer paths are valid in the sense that they still lead the receiver from A to B. However, they are also slower and require more bits to encode side trips. In terms of Kolmogorov messages, the straight line path is the Kolmogorov minimum, and the infinitely large family of curved paths represents all the longer and less efficient messages that convey the same semantic content.

The second box shows a procedure for moving messages closer to the Kolmogorov minimum. The goal is to seek out and eliminate the curved "side trip" components that eventually end up cancelling themselves out. This is different from redundancy elimination, since the central goal in side trip elimination is to find and factor out the curve-inducing oppositely signed components that create the wasteful side trips. Uncovering and trimming away (third box) these computationally and conceptually wasteful side trips not only improves the clarity of theories, but can dramatically improve the efficiency of software that relies on such theories.[6]

The Trampoline Effect

Note that the closer one moves to the straight path Kolmogorov minimum in a state change map, the harder it becomes to find further simplifications. This can happen even when factors such as too many arbitrary constants, too much reliance on raw data, or unexplained features suggest that further simplifications are needed. The unsettling truth in most such cases is that theorists are making wrong assumptions about what is "fundamental" or "atomic" and what need to be broken down farther.

Unfortunately, this resistance to further compression when approaching Kolmogorov minima can easily lead to what I call the trampoline effect: Bouncing *off* of the near-minimum region by adding new ideas that seem relevant, yet in the end just add more complexity and more curves. The result is to create theories that

in effect bounce off the taut Kolmogorov minimum path and instead send theorists out on far and sometimes fascinating side trips, but ones that ultimately have very little to do with the original simplification problem. One of the signs that this has happened is when the literature for a topic that once requires only few concentrated pages of math to describe suddenly explodes into a huge spectrum of papers and ideas that no longer converge to any obvious resolution of the original problem.

The trampoline effect helps explains the baffling sequence of events that ensued after completion in the 1970s Standard Model of particle physics. Judging by its remarkable predictive success and relatively small size, the 1970s model likely was already relatively close to its Kolmogorov minimum.[7] Attempts to shrink the Standard Model further instead resulted in a spectacular explosion of almost entirely untestable and heavily mathematical papers, with *string theory* [8][9] in particular dominating theoretical physics research for decades. One indicator that this was a trampoline bounce is that the Standard Model was left largely unaffected.

The Spekkens Principle

Robert Spekkens contemplated something very close to the trampoline effect in his 2012 FQXi essay [10] when he addressed the curiously complementary relationship between using bits to describe where a particle "is" at a given moment — its "kinematic state" — and how that particle and its state changes as it moves into the future — its "dynamics." The principle that Spekkens recognized was that the kinematic and dynamic descriptions in quantum theories can take on dramatically different forms as long as the two sides remain complementary in some deeper fashion. From this Spekkens speculated that there must exist a more fundamental fulcrum point from which these many various pairings of kinematics and dynamics emerge, much as in the mutually cancelling side paths I describe for the trampoline effect. He even proposed a specific approach, causal structure, as a starting point for uncovering this theoretical fulcrum. In Kolmogorov terminology, the fulcrum that Spekkens postulated would be the Kolmogorov minimum for quantum theories, and the various interpretation pairs would be examples of "side trips" into areas that theorists such as John Bell [11] (a pilot wave advocate) and David Deutsch [12] (a many-worlds advocate) felt needed to be addressed.

Three Challenges

I would like to end this essay with three challenges, two of which originated with Nobel Laureate Richard Feynman, and one of which originates broadly with the particle physics community.

Challenge #1: What is the full physics meaning of Euler's identity, $e^{i\pi} + 1 = 0$?

One of Richard Feynman's distinguishing traits was his exceptionally good nose for the profound, and he found Euler's identity enthralling.[13] Why? Because it compactly connects four (or five) of the most fundamental and profound constants in all of mathematics: e, i, π, 1, and implicitly -1 by subtracting 1 from both sides.

Euler's identity is already arguably the basis for much of the mathematics used in quantum mechanics, since it is the starting point for expressing wave mechanics in an exceptionally elegant and compact form. However, my challenge (not Feynman's *per se*) is a bit different: I am asserting that due to its extreme brevity, Euler's identity is most likely an overlooked example of a Kolmogorov minimum relevant to the physics of our universe. My postulate is that we don't think of Euler's identity as physics only because we do not yet understand how it maps into experimental reality. Identifying such connections might lead to some new factoring of physics in general and of quantum mechanics in particular, one in which Euler's identity pops out and brings together concepts that previously were thought to be unrelated.

Challenge #2: What is the simple explanation for fermion-boson spin statistics?

For over 20 years, Richard Feynman thought about what seems at face value to be an amazing coincidence.[14] All known fundamental particles in physics fall into one of two categories: fermions that refuse to share the same state, and bosons that love to share the same state. Fundamental fermions include electrons, quarks, and neutrinos, and also composite protons such as neutrons. These fermions form what we call matter. Fundamental bosons include photons and gluons, and are the basis both for energy (e.g. a beam of light) and, in virtual form, fields (e.g. electromagnetic fields).

Every fundamental particle also has a quantized form of angular momentum called spin, and its spin has a fascinating relationship to these two families. Particles that include a very strange and originally unexpected form of angular momentum called ½ spin are always fermions, while particles that use only the much more understandable whole integer spins (eg. 0, 1, or 2) are always bosons.

The question that troubled Feynman for decades, and which he never was able to answer to his own full satisfaction, was this: What is the *simple* explanation for this connection between spin and the two families of particles?

I should hasten to note that the necessity of this correlation was proven decades ago, so in that sense it is not a mystery! The problem that troubled Feynman was that for so simple a rule, there should also be a similarly simple explanation. The current proofs of the connection are anything but that, requiring pages of complicated arguments that leave the reader thinking no better off in terms of understanding *why* such a thing should be so.

Given that the very concept of spin ½ is nonsensical when applied to ordinary three-dimensional space, the lack of simplicity in this case likely stems from our inability understand what spin ½ really means at a deeper level. The great early quantum physicist Wolfgang Pauli unfortunately became so frustrated with his own inability to resolve the spin ½ issue that he finally (and angrily, as was his tendency when frustrated) declared it a "property" of quantum systems that had no need for further analysis by him or anyone else. Pauli thus set up a pattern that persists strongly to this day of simply ignoring one of the most fascinating clues in all of

physics, which is the existence in all fermions of a type of angular momentum that *makes no sense* from any classical perspective.

The second challenge thus is to stop treating this astonishing half-spin mystery as "irrelevant" and instead seek out a deeper, more *fundamental* understanding of how half spin can even exist in our universe. After all, if you have a mysterious behavior (in this case "why do fermions refuse to share the same state?") that is firmly and profoundly attached to an even more mysterious and opaque box ("what exactly is half spin?"), the odds are quite good that figuring the mystery of the box works will also provide insights into the unique behavior associated with that box.

Challenge #3: Refactor the Standard Model *without* adding gravity or complexity.

While often lauded as the most successful predictive model in all of physics, the Standard Model of particle physics is also frustratingly incomplete. The evidence for its incompleteness shows up vividly in its large number of arbitrary constants and baffling "givens," such as why there are three generations of fermions.[7]

The three generations of fermions are a particularly pointed example of our lack of a deeper understanding of fermions. The first generation of fermions contains the stable particles such as electrons and proton-forming quarks that constitute nearly all of the visible matter in the universe. But for some reason, the universe also allows two sets of nearly identical fermions that differ from the first set only in mass. As with spin ½, this experimental finding was so totally unexpected that when theorist Isidor I. Rabi first heard about out about the muon, the second generation heavy version of the electron, he exclaimed "Who ordered *that*?" [15]

Despite its reliance on several key points on unfactored experimental data, the Standard Model qualifies overall as a remarkably compact and thus fundamental framework for describing most of the universe. It was the desire to factor it further that lead to work in areas such a string theory, in which the mind-bogglingly large vibration modes of tiny strings and loops in higher dimensional spaces are assumed to explain not just the particles and fields of the Standard Model, but also gravity. Curiously, although decades of effort in string theory have produced an enormous number of often very arcane, hard-to-understand papers, what it has not produced are any simple or convincing insights into the most blatant unexplained features of the Standard Model, such as why the three generations of fermions even exist.

One factor in why string theory and related efforts to explain the Standard Model became so complex is their insistence on including gravity. Because gravity is so weak, principles of quantum mechanics drove the scale of such models into both extremely small length scales and extraordinarily high energies. This in turn helped unleash so many new options for "exploration" that the original Standard Model simply got lost in an almost unimaginably large sea of possibilities.[9]

Thus my suggestion for anyone interested in bit-reduction refactoring the Standard Model is simple: *Stop trying to include gravity in the refactoring*. Instead, take what was already in the 1970s original version and look for novel ways to factor it that

reduce its size instead of expanding it. Furthermore, take issues such as the half-spin conundrum and the existence of three fermion generations as first-order clues that need to be integral parts of the final explanation.

Another strategy is to look for unexpected symmetries, but this time *without* insisting on using group theory first. While powerful, group theory is like software: It only takes what you put into it. If what you feed into the powerful machinery of group theory ignores or skims over issues such as why ½ spin exists, or why there are three fermion generations, it guaranteed that whatever sausage comes out the other end of your symmetry grinder will be just as oblivious to these issues.

Regarding gravity, here's a thought: If someone can succeed in uncovering a smaller, simpler, more factored version of the Standard Model, who is to say that the resulting model might not enable new insights into the nature of gravity? A more fundamental quantum model of the fermion and bosons could for example point to emergent effects relevant to gravity. There are after powerful theoretical reasons for arguing that gravity is *not* identical in nature to the other forces of the Standard Model. That reason is the very existence of Einstein's General Theory of relativity, which explains gravity using geometric concepts that bear no significant resemblance to the quantum field models used for other forces. Focusing on clarifying the relationships of the clearly quantum forces thus might open up opportunities to clarify why gravity looks so different, in ways that embrace and complement the geometric power of General Relativity instead of ignoring it.

Final Thoughts

If you have gained anything by reading this essay, my hope is that it is the belief that simplicity is just as important now as it was in the early 1900s heydays of relativity and quantum theory. Simplicity is more akin to a rare gemstone than it is to a massive building, and it is more likely to be found by someone who likes to pull on unexplained dangling threads. This includes in particular threads that have been dangling for so long that no one bothers to look at them closely anymore.

If you see such a thread and find it intriguing, your first step should be to find and immerse yourself in the details of any high-quality experimental data relevant to that thread. Some obscure detail from that data could become the unexpected clue that helps you break a major conceptual barrier. With hard work and insight, you might just become the person who finds a hidden gemstone of simplicity by unravelling the threads of misunderstanding that for decades have kept it hidden.

References

[1] Albert Einstein. Does the Inertia of a Body Depend on its Energy Content? *Ann Phys*, 18:639–641, 1905.

[2] Andrey N. Kolmogorov. Three Approaches to the Quantitative Definition of Information. *Problemy Peredachi Informatsii*, 1(1):1–7, 1965.

[3] Khalid Sayood. *Introduction to Data Compression (Fifth Edition)*. Morgan Kaufmann, 2017.

[4] A. Einstein. The Foundation of the Generalised Theory of Relativity. *Annalen der Physik*, 354(7):769–822, 1916.

[5] Stephen Haywood. *Symmetries and Conservation Laws in Particle Physics: An Introduction to Group Theory for Particle Physicists*. World Scientific, 2011.

[6] Jean Michel Sellier. A Signed Particle Formulation of Non-Relativistic Quantum Mechanics. *Journal of Computational Physics*, 297:254–265, 2015.

[7] Necia Grant Cooper and Geoffrey B West. *Particle Physics: a Los Alamos Primer*. Cambridge University Press, 1988.

[8] Lee Smolin. *The Trouble with Physics: The Rise of String Theory, the Fall of a Science, and What Comes Next*. Houghton Mifflin, 2006.

[9] Brian Greene. The Elegant Universe: *Superstrings, Hidden Dimensions, and the Quest for the Ultimate Theory*. Vintage, 1999.

[10] Robert W Spekkens. The paradigm of kinematics and dynamics must yield to causal structure. In *Questioning the Foundations of Physics*, pages 5–16. Springer, 2015.

[11] J.S. Bell. *Speakable and Unspeakable in Quantum Mechanics: Collected Papers on Quantum Philosophy*. Cambridge University Press, 1987.

[12] David Deutsch. *The Fabric of Reality*. Penguin Books, 1997.

[13] Richard P Feynman, Robert B Leighton, and Matthew Sands. *The Feynman Lectures on Physics, Vol. I: The New Millennium Edition: Mainly Mechanics, Radiation, and Heat*, volume 1. Basic Books, 2011.

[14] Richard Phillips Feynman, Steven Weinberg, Richard MacKenzie, Paul Doust, and Steven Weinberg. *Elementary Particles and the Laws of Physics: The 1986 Dirac Memorial Lectures*. Cambridge University Press, 1987.

[15] Marcia Bartusiak. Science & Technology: Who Ordered the Muon? *New York Times*, Sept 27, 1987.