2 Multiple Patterns, Multiple Explanations

Steve Petersen

Introduction

At the heart of the unificationist account of scientific explanation is the idea that we explain events by subsuming them into wider *patterns* (Kitcher 1989). We can supplement this key idea with a formal theory of patterns, according to which a pattern is a regularity in the explananda that allows for data compression. This notion is lifted from algorithmic information theory (AIT), which also goes by the name "Kolmogorov complexity theory." (AIT studies theoretical limits of data compressibility and identifies the information content of a particular data string with the length of its best compression (Li and Vitányi 2008).) This formal pattern-based approach results in a robust version of explanation unificationism that is both immune to its usual criticisms and able to incorporate the best insights of rival accounts. A detailed defense of this "patternist" account of explanation is in the works. For this volume, though, I would like to highlight an independent feature: the patternist account of explanation can provide both a rigorous sense of how data can admit multiple explanations and a rigorous sense of how some of those explanations can conjoin, while others compete.

I frame this as a response to James McAllister (2007), who argues that three AIT-based model selection techniques—such as the patternist one I propose—are not adequate, exactly because they *cannot* accommodate the multiple overlapping patterns that data sets frequently exhibit.[1] He gives three helpful examples of data sets with overlapping patterns, and we will focus on the simplest: a time series of temperatures at a particular spot on Earth. McAllister points out such a data set will have cyclical patterns such as daily and yearly variation, as well as longer-term cycles from sunspots and the Earth's precession. There will also be non-cyclical patterns, such as the "hockey stick" of global climate change. McAllister says

> each of these models [diurnal variation, sunspots, *etc.*] must be regarded, in the light of our current knowledge, as very close to the

Acknowledgements: Thanks to James Evershed, James McAllister, Jonah Schupbach, and especially to Tomoji Shogenji for helpful thoughts and comments.

DOI: 10.4324/9781003184324-4

> truth: there are strong grounds for considering each pattern to be a genuine component of the data, and for regarding the hypothesized cause of the pattern to be a real physical phenomenon.
>
> (p. 888)

He then argues that

> standard quantitative techniques for choosing among data models [such as from AIT] . . . lack the conceptual resources to allow for the possibility that a data set can be correctly analyzed in several different ways.
>
> (p. 890)

On the full account of my view, such compressing models are patterns, and those patterns can themselves be explanatory. This immediately seems wrong for the toy data set before us: the mere regularity of daily temperature variation is clearly not itself *explanatory* of the data. Rather, the explanation of that variation is (roughly) the rotation of the Earth. But this is just an artifact of the toy example because the explanation adverting to the Earth's rotation is relative to a different data set that includes such astronomical facts. Patternist explanation, as a form of unificationism, is a *global* affair. When we consider all data actually available to us, the laws of physics come out as the fundamental explanatory regularities. (Patternism also allows for higher-level explanations at different levels of abstraction, but that is a long story.) In this toy example, I am pretending that our only evidence is this data series. Thus we are pretending the daily variation is a brute regularity that is minimally explanatory, but not itself explained (in the same way fundamental laws of physics could be unexplained regularities that explain).

So although the chosen example does not make much sense of why I take models to be explanatory, that is beside the point here; the example is sufficient to make McAllister's concern about accounts like mine clear. If we take such models to be explanatory, then McAllister's examples illustrate how a data set can have multiple, noncompeting explanations. We would like a way to say that any such pattern *partially* explains the data and consider how multiple partial explanations can combine or compete. McAllister holds that AIT-based accounts cannot accommodate this desideratum; here I aim to show that mine can.

Patternist Explanation

First I present the core of my patternist view, focusing on the relevant portions for this issue. Start with the "data set" at hand, such as the time series of temperatures from McAllister's examples. Consider those data as encoded in one binary string x. (One simple example of binary data encoding: a spreadsheet file containing the data, as represented in bits on your computer.) Next, fix a friendly universal Turing machine (UTM), U.[2] By definition the universal U can emulate any other Turing machine, as run on any input; we simply encode the Turing machine to emulate, and the input to that emulated Turing machine (TM), as an ordered pair (p,n). (We can think of the emulated TM p as the "Program" and n as the "iNput" to that program.) The result is written $U(p,n)$.

The Kolmogorov complexity of data x, written $K_U(x)$, is the length of its best compression—that is, its complexity is the length of the shortest (p,n) required for U to output x. The standard example of how regularities allow for compression is a very long string of m 1s for some large enough m. Code like "for i from 1 to m: print 1" will be much briefer than the original string, showing the string to be quite simple. In the tradition of Daniel Dennett's "Real Patterns" (1991), any such *compressing* regularity is basically all that I mean by a *pattern.*

Pattern

> p is a pattern in data x iff it is the program portion of a compression of x, that is there is an n such that $U(p,n) = x$ and $\text{len}(p,n) < \text{len}(x)$.

Since being a pattern is a necessary condition for explanation on my account, we could call any such pattern a *potential* explanation of the data. For our purposes we can think of the input n to p as the *noise* term, although "noise" isn't quite right, since it can contain details of realization in addition to error terms and may carry patterns itself. Calling n the "noise" is only appropriate insofar as it is *intended* to carry the non-patterned information. The simplicity of the pattern and of the noise are measured by their lengths, that is, the number of bits they each require to be fed into the UTM in order to recreate the original data set exactly. For example, we could model our time series of temperatures by trying to curve-fit it to some polynomial. If we pick a very simple polynomial, such as a straight line, then the p portion of the compression will be quite short—but we will also need a lot of error terms, encoded into a much longer n, to reproduce the original data losslessly. On the other hand we can pick a polynomial with *no* error terms if it has as many degrees as there are data points. But then of course the polynomial will be extremely complicated, resulting in a very long p portion. Seen this way, the game of curve-fitting is to find the right trade-off between model simplicity and model fit. AIT provides a common currency in which to make that trade: the length in bits of p (model) and n (noise).

It is important to emphasize, especially in response to McAllister, that program p is only a pattern in the data if its length *together with the length of the noise term n* are shorter than the original data x. This is crucial in the AIT tradition of Minimum Description Length for finding a good trade-off between model simplicity and data fit: any additional model complexity must pay for itself with smaller error terms, and larger error terms must pay for themselves in model simplicity (Grünwald 2007). Only data sets with some real regularity can actually be *compressed* by a trade-off between these two considerations.

A data set like our time series of temperatures x will exhibit multiple patterns in this sense: it is very plausible that each of the regularities McAllister mentions will, on their own, be sufficient to compress x. But this does not yet show that data sets can have multiple *explanations*, because not all patterns are genuinely explanatory on my account. For example, a mere preponderance of 1s over 0s will be enough to compress a long-enough binary string (by Shannon-Fano encoding), and so will be a genuine pattern in the string by my definition—but this compressing regularity tells us nothing about what we would intuitively consider the *reason* for such a preponderance.[3]

To account for this, patternist explanation requires a fundamental notion that I call a "proper" explanation. A proper explanation is basically an "ideal compression" of the data, in a specific sense: not only are the (p,n) together minimal in length, but the p is the shortest possible of all such pairs in that minimal length.[4] So a proper explanation is the simplest program portion of the best compression of the data.

Proper Explanation

> Pattern p^* properly explains data x iff p^* is a shortest pattern portion of a maximal compression, that is, $K_U(x) < \text{len}(x)$ and for some n, $U(p^*,n) = x$, and for any q and m, if $U(q,m) = x$ then $\text{len}(p^*,n) \leq \text{len}(q,m)$ and $\text{len}(p^*) \leq \text{len}(q)$.[5]

Preference for such a model seems to be roughly the concern McAllister had in mind: this "one model to rule them all" looks like it would crowd out all the particular, individual explanatory patterns in x that might interest a scientist. Worse, these proper explanations are extremely demanding; it is very unlikely we have identified *all* patterns in temperature variation, for example, and in general extracting all the explanatory regularities from a data set is an uncomputable ideal. Thus we possess very few if any *proper* explanations. Yet it seems that there is at least some important sense in which science does, now, possess *good* explanations—in particular, as McAllister's example illustrates, it seems that even though we are unlikely to have found the best possible explanation of x in terms of all its regularities, we already possess several *partial* explanations of it, none of which is the whole story.

So to address concerns like McAllister's and accommodate the possibility of multiple explanations, patternist explanation must make sense of such partial explanations. The key move is to define partial explanations as any pattern that, in a precise algorithmic sense, provides some information about the proper explanation. String *a provides information* about another string *b* just in case $K_U(b \mid a) < K_U(b)$, where $K_U(b \mid a)$ is the *conditional Kolmogorov complexity*: the length of the shortest (p, n) required to produce *b* given string *a* as input "for free." In sum, *a* provides information in this sense about *b* when *b* is easier to compress if *a* is already known. The measure of *how much* easier, in bits, is called the *algorithmic mutual information* between *a* and *b*:[6]

$$I(a : b) = K_U(b) - K_U(b \mid a)$$

Note this sense of "provides information" contrasts with a more standard reading, where to provide information is to eliminate some possibilities. To say string *x* starts with a 1 provides information about *x* only in the latter sense.

Algorithmic mutual information allows us to define a partial explanation as one that gives some information about the proper explanation:

> **Partial Explanation**
>
> Pattern *p* partially explains data *x* if and only if *p* provides information about *x*'s proper explanation p^*, that is, for some n, $U(p, n) = x$, and $\text{len}(p, n) < \text{len}(x)$, and $K_U(p^* \mid p) < K_U(p^*)$.

Thus patternism formalizes a strategy for partial explanation that is perhaps familiar from Peter Railton's (1981) proposal: we start with an "ideal explanatory text" (what I'm calling the "proper" explanation) and count the right information *about* that ideal text as partially explanatory.

An ideal compression of *x* will exploit *all* patterns McAllister mentions and then some—but simply noting the variation from a 24-hour cycle will surely be enough to compress the data to some extent, and this pattern seems very likely to be part of the *best* compression. Roughly put, a programmer trying to compress the data as far as possible would happily incorporate a subroutine that can adjust for daily variation and then layer other factors (such as the yearly cycle) on top of it. This is why, on my view, it is right to say the daily cycle helps explain the temperature variation at that spot. It may also of course be the most *relevant* partial explanation in some particular context. McAllister worries the best compression "disregards all the other patterns" (p. 890), but on this account it *incorporates* all the patterns that are partially explanatory.

Note that because we won't typically have the proper explanation in hand, we typically won't be able to *know* whether some pattern provides information about the proper explanation, so we won't know whether the pattern is partially explanatory. Strictly speaking any account according to which explanation is factive will run this risk—we can always think we have an explanation and be wrong. But my account may seem more worrisome on this score because it is harder to see how we could be *justified* in thinking some pattern is part of the ideal explanation. Compressions are rare, though; there can certainly be patterns that tell us nothing about the proper explanation, but I think just finding one is some evidence we are on the right track. In practice the patternist about explanation will simply seek the *best* patterns for the purpose. When we are lucky enough to find two or more patterns in the data, we can consider the degree to which they conjoin or compete (in the sense cited later), slowly triangulating on the proper explanation.

McAllister closes his paper by suggesting that an algorithmic approach to model choice must, at the least, be able to take a pre-specified tolerance for noise into account since plausibly this is what practicing scientists do when they work at different levels of abstraction, examining different patterns. He claims that approaches like mine cannot accommodate this. Here I have tried to show how they can: by appealing to genuine patterns in the data, partial explanations allow for approaching a data set at different levels of noise tolerance. But this does not mean "anything goes" either; to be good objects of scientific inquiry, the patterns in play must still compress, even with the noise term. And to count as explanatory, they must tell us at least *something* about the full, "proper" explanation.

McAllister's Anticipatory Response

McAllister anticipates a response like mine, namely

> to claim that there is indeed a unique best model of any such data set—the one corresponding to the sum of several or all the patterns that can be identified in the data—and that the quantitative techniques can be expected to pick out this pattern as the closest to the truth.
>
> (p. 891)

He gives two reasons this response will not work; I will respond to them in turn.

> First, the sum of all patterns that can be identified in the data would probably coincide with the complete data set itself, since any discrepancy between a data set and a pattern identified in it can be endlessly analyzed as a sum of further patterns.
>
> (p. 891)

Perhaps in the grip of my own view, I confess it is not easy for me to make sense of this passage; I suspect McAllister means something quite different by "pattern," or perhaps "sum." In *my* defined sense of "pattern," at least, it is clear that the sum of all patterns does not "coincide" with the data set itself. For a simple example, think again of the program "for i from 1 to m: print 1," which I suppose is the one best compression of a long string of 1s. That program is thus the "sum of all patterns" for a long string of 1s, but it is not the same as that long string.

Also—again, in my defined sense of "pattern"—it is not true that the discrepancy between the pattern and the data set (which I take to be the *n* term) can be "endlessly analyzed as a sum of further patterns." Though *partial* explanations may leave some patterns in the *n* term, the ideal compressing program behind the "proper explanation" must squeeze out any such regularities, leaving its noise term incompressible.[7] At any rate, however McAllister understands "discrepancies," they cannot be *endlessly* analyzed as compressing patterns (I'm not sure how literally he meant this); in general a lossless compression cannot itself be compressed.[8]

So let us put this objection down to a miscommunication about "patterns" and turn to McAllister's second response:

> Second, scientists adduce individual patterns in data as evidence for claims about the contributions of individual causal factors. The evidence for a claim about the existence and effect of a causal factor consists of the component pattern that is determined by that causal factor alone: it does not consist of the resultant pattern determined by the combination of several or all causal factors operating in a physical system . . . . For these reasons, the notion of a sum of several or all patterns does not nullify the reality or the significance of each component pattern.
>
> (p. 891–892)

McAllister rightly points out that scientists will want to isolate different such patterns; in the case of *x*, for example, climate change scientists will likely focus on the long-term patterns, while meteorologists will focus on more daily ones. I hope it's clear that patternism can account for this. We often focus on one aspect of the "ideal explanatory text" or the other for pragmatic reasons. The climate change scientist and the meteorologist are both studying legitimate *partial* explanations of the variation in *x*.

McAllister summarizes his position this way:

> In this paper, I argue that the assumption that an empirical data set provides evidence for just one phenomenon is mistaken. It frequently occurs that data sets provide evidence for multiple phenomena, in the form of multiple patterns that are exhibited in the data with differing noise levels. This means that, in these cases, several different models of a data set must be regarded as equally close to the truth. In the light of this fact, none of the standard techniques for selecting among

> models of data sets can be considered adequate, since none allows for the possibility that a data set may admit multiple models.
> (p. 886–887)

I think the best thing to say, in cases like the temperature time series, is not that the diurnal, annual, *etc*. models are all "equally close to the truth"—rather, they are all *part* of the whole truth, and some may be bigger parts than others.

Measuring Competition and Conjunction

I would like to close with a related advantage. Not only can patternist explanation accommodate multiple explanations, but it also can provide a precise measure of the extent to which different partial explanations of data can conjoin or compete. Recall the information that partial explanation p provides about the proper explanation p^* is measured in bits by $I(p : p^*) = K_U(p^*) - K_U(p^* \mid p)$. It seems to me that if partial explanations p and q each capture different aspects of the proper explanation, so that they are perfectly complementary, then this means that together they would provide as much information about proper explanation p^* as each individually. Where pq is the concatenation of the two programs, then, we should have[9]

$$I(p : p^*) + I(q : p^*) = I(pq : p^*)$$

I would consider such a pair of partial explanations to be perfectly *conjunctive*. On the other hand, p and q might be totally redundant—that is, once you have the information from one, the other does not help to compress p^* further at all. In this case the savings of both together will be no better than the most informative alone. If p is the more informative pattern, so that $I(p : p^*) > I(q : p^*)$, then complete redundancy would mean

$$I(p : p^*) = I(pq : p^*)$$

(Note it can never be the case that p gives *more* information than p and q do together.) When p and q are redundant like this, they are perfect *competitors*; there is no reason to take both on board. We might prefer p, since it contains all the information in q and more—it "screens off" q. Or we might prefer q for its more narrow focus given a specific interest, especially if it is shorter. But in no situation would we want to use both.

There are many possibilities between these, where there is some competing overlap of information but also some coordination between the two partial explanations. Since the worst that can happen in conjoining the two is no improvement over the best of the two (perfect competition), and the best that can happen is for each to maintain its full explanatory force, so that the two together are as powerful as each separately (perfect conjunction), we can measure the *degree* of complementarity by comparing how each does separately vs. both together. That is, take the sum of bits saved by each individually, and subtract off the bits saved by the two together. The result will range between zero for perfect conjunction, and the size of the wasted number of bits of the worse explanation for perfect competition. We can thus normalize by this worst possible case, to get a measure in $[0,1]$:

$$0 \leq \frac{\mathrm{I}(p:p^*)+\mathrm{I}(q:p^*)-\mathrm{I}(pq:p^*)}{\min\left(\mathrm{I}(p:p^*),\mathrm{I}(q:p^*)\right)} \leq 1$$

Here 0 is perfect conjunction, and 1 is perfect competition.[10]

Readers of this collection especially may be familiar with Jonah Schupbach and David Glass's two desiderata for hypothesis competition (2017):

1. "Hypothesis competition is a matter of degree."
2. "There are two pathways to hypothesis competition: a direct pathway and an indirect pathway via the evidence."

We have just seen how patternism captures the first of these. The second is not so straightforward in this AIT framework. In the tradition of inference to the best explanation (Harman 1965), all the hypotheses are intended as explanations, and explanations always have their *explananda* as their evidence. So it is not clear how hypotheses can compete "directly" as explanations, independently of what they purportedly explain.[11]

We would further like to be able to compare two hypotheses in practice, where we usually don't know the proper explanation. When we have two *potential* explanations p and q of data x—that is, two compressing regularities that may or may not provide part of the proper explanation—we can ask the extent to which they overlap in compressing x using similar mechanisms as mentioned previously. Since the Kolmogorov complexities of our strings will generally be unknown, we can instead ask whether pattern p can help compress the noise term for q, or *vice versa*. There is no straightforward, tractable algorithm here; it is a matter of understanding the patterns well enough to see whether and how they might interact.[12] As a simplified example, suppose p divides the temperature time series x into 24-hour chunks and exploits the predictable curve for each such chunk well enough to compress them—but it treats the average temperature for each such chunk as unexplained noise. Suppose q, meanwhile, exploits the yearly pattern in the average of each 24-hour chunk but treats the variation within each 24-hour chunk as unexplained noise. Then they can each compress each other's noise terms, and we have (apparently) conjunctive explanations; p and q together will compress x by about as much as the sum of each individual compression. On the other hand, if pattern r takes into account yearly variation *and* the temperature trend from climate change, it is a clear competitor with yearly pattern q—we might choose the simpler q for some purposes, or the more accurate r for others, but never both together.

As cases like this illustrate, I wholeheartedly agree that data sets can exhibit multiple explanatory patterns, some pairs of which compete and some pairs of which conjoin. In my book, this is just one more reason to approach explanation as a patternist.

Notes

1. Specifically he argues that AIT, Minimum Description Length (MDL) (Grünwald 2007), and the related Akaike Information Criterion (Forster and Sober 1994) for model selection all fail to account for multiple patterns in data. My patternism is closely allied with MDL, which I think is more accurately taken as a branch of AIT.
2. Which data encoding we choose does not matter much, assuming it is computable, since it comes out in the wash when choosing the universal Turing machine. I use "friendly" basically to mean that U should be both *prefix-free* and *additively optimal*; see Li and Vitányi (2008). Normally the subscript for the reference UTM is suppressed, since as a function all friendly UTMs differ only by a constant. But since the Turing-machine-relativity may be of philosophical significance, we will conscientiously preserve it.
3. If on the other hand there is no further fundamental regularity responsible for that preponderance—as for example a universe consisting solely of one pure Bernoulli process—then I would say the mere statistical preponderance is the best (because only) explanation available.
4. Note that there will typically be a number of program-input pairs that can reproduce x in the minimal length, since we could hard-code an argument into the program, or load some of the program portion as data input.
5. The clause "$K_U(x) < \text{len}(x)$" guarantees that x is compressible and so guarantees that p is a pattern as defined.
6. This is intended, of course, to be analogous with conditional probabilities and the more traditional mutual information from Shannonian information theory. The "mutual" is justified in both cases because this relation is symmetric—or more carefully, in the algorithmic case, it is symmetric up to a constant, once defined a bit more carefully. See Grünwald and Vitányi (2003) Section 5.2.
7. See Vereshchagin and Vitanyi (2004) for the proof, which strictly speaking holds up to an additive $O(\log \text{len}(x))$ for overhead.
8. Otherwise we could then compress *that* compression losslessly, and so forth. But lossless decompressions are unique: no matter the technique, at most two strings can be compressed down to one bit, at most four more can be compressed down to two bits, and so on. So clearly not just any string can be "endlessly" compressed.
9. I am neglecting small constant fudge factors for concatenation and such throughout.
10. We could generalize this to any finite set $\{p_i\}_1^n$ of partial explanations:

$$\frac{\sum_i \mathrm{I}(p_i : p^*) - \mathrm{I}(p_1 p_2 \dots p_n : p^*)}{\sum_i \mathrm{I}(p_i : p^*) - \max \mathrm{I}(p_i : p^*)}$$

11. I did find some potential ways to characterize something like "direct" hypothesis competition in my framework, but they are probably not worth the space here.
12. This is not to say there's no algorithm for doing such inference—only no algorithm that is both straightforward *and* tractable.

References

Dennett, Daniel C. (1991). "Real Patterns." *The Journal of Philosophy* 88 (1): 27–51.

Forster, M. R., & Elliott Sober. (1994). "How to Tell When Simpler, More Unified, or Less Ad Hoc Theories Will Provide More Accurate Predictions." *British Journal for the Philosophy of Science* 45: 1–36.

Grünwald, Peter D. (2007). *The Minimum Description Length Principle*. Cambridge: MIT Press.

Grünwald, Peter D., & Paul M. B. Vitányi. (2003). "Kolmogorov Complexity and Information Theory." *Journal of Logic, Language, and Information* 12: 497–529.

Harman, Gilbert. (1965). "The Inference to the Best Explanation." *The Philosophical Review* 74 (1): 88–95.

Kitcher, Philip. (1989). "Explanatory Unification and the Causal Structure of the World." In *Scientific Explanation*, edited by Philip Kitcher and Wesley C. Salmon, XIII:410–505. Minnesota Studies in the Philosophy of Science. Minneapolis: University of Minnesota Press.

Li, Ming, & Paul M. B. Vitányi. (2008). *An Introduction to Kolmogorov Complexity and Its Applications*. Third edition. New York: Springer.

McAllister, James W. (2007). "Model Selection and the Multiplicity of Patterns in Empirical Data." *Philosophy of Science* 74 (5): 884–94.

Railton, Peter. (1981). "Probability, Explanation, and Information." *Synthese* 48: 233–56.

Schupbach, Jonah N., & David H. Glass. (2017). "Hypothesis Competition Beyond Mutual Exclusivity." *Philosophy of Science* 84 (5): 810–24.

Vereshchagin, N. K., & Paul M. B. Vitanyi. (2004). "Kolmogorov's Structure Functions and Model Selection." *IEEE Transactions on Information Theory* 50 (12): 3265–90.