Occam's razor:
From Ockham's *via moderna* to modern data science

HUGO A. VAN DEN BERG

Mathematics Institute, University of Warwick, Coventry CV4 7AL, UK

The principle of parsimony, also known as "Occam's razor," is a heuristic dictum that is thoroughly familiar to virtually all practitioners of science: Aristotle, Newton, and many others have enunciated it in some form or other. Even though the principle is not difficult to comprehend as a general heuristic guideline, it has proved surprisingly resistant to being put on a rigorous footing – a difficulty that has become more pressing and topical with the "big data" explosion. We review the significance of Occam's razor in the philosophical and theological writings of William of Ockham, and survey modern developments of parsimony in data science.

Keywords: *Occam's razor, William of Ockham, principle of parsimony, large deviations, Kolmogorov complexity, Bayesian inference*

Hugo van den Berg teaches mathematical biology at the University of Warwick. In addition to textbooks on several topics, including evolutionary dynamics and the general theory and practice of mathematical modelling, he has written over fourscore papers on energetics & homeostasis, T-cell immunity, evolutionary biology, bioinformatics, and the physiology of uterine contractility. *email*: hugo@maths.warwick.ac.uk

Introduction

Parsimony is a virtue, but not a guarantor of truth. Yet this is what popular renditions of Occam's "razor" maxim would have us believe: the simplest explanation is said to be the best. An explanation is a theory or hypothesis consistent with the known facts, and we would be forgiven to infer that by the best is meant: *correct*, or else perhaps: *the most likely*.

A counterexample readily establishes that parsimony is not necessarily the path to truth, nor even the most likely approximant to truth. For suppose someone makes the following observations:

$$?, ?, ?, 0, ?, ?, ?, 0, ?, ?, ?, \ldots$$

where the question mark indicates a missing data point. Then a simple pattern consistent with these data is as follows:

$$0, 0, 0, \underline{0}, 0, 0, 0, \underline{0}, \ldots$$

where the actual data have been underlined. This, most would agree, is the most parsimonious "explanation" even though we should perhaps not be all that confident given the paucity of the data set. Now let another observer gather the following data:

?, ?, 16, ?, 36, ?, 64, ?, ?, ?, …

and surmise that these observations reflect the following underlying pattern of squares:

4, 9, 16, 25, 36, 49, 64, 81, …

– again a parsimonious account of the data (arguably "the simplest"). It turns out that both observers have partially observed the *very same* phenomenon:

4, 9, 16, 0, 36, 32, 64, 0, …

and both have come up with incorrect inter- and extrapolations of the fragmentary data. The above is the start of a well-defined mathematical sequence with its own underlying regularity as well as a physical meaning, viz. the number of coincidence site lattices of index n in the Z^4 lattice[1]. The first observer would expect the next term to be 0, whereas the second observer would guess 100 (the square of 10), but both would be wrong… the actual value being 168.

What makes this example work is that (i) both observers have to work with pretty sparse data sets, and (ii) the complexity underpinning the sequence is far greater than the naïve observer might have thought (the reader will be spared the actual formula). But these circumstances are entirely typical of the scientific endeavour: even in the present age of "big data" the cumulative corpus of observational data collected by humans or instruments acting on their behalf is infinitesimal compared to the data that *might* have been gathered by some omnipotent being – indeed, it is a small miracle that science seems to be possible at all. As for underlying complexity, this is a more contentious point, since there is no generally agreed-upon (and universally applicable) concept of complexity, and even then: the best scientific theories we have at present are quite simple, but only when looked at in the right light. The lesson that the history of science teaches is that, as our insights (notably our mathematics) evolve, we continually change our opinions about what constitutes complexity, elegance and the like.

In sum, even if the above example seems somewhat contrived, it is in fact perfectly representative of the existential dilemmas routinely faced by scientists. What might seem to be an elegant and appealing theory to one particular scientist, working with a partial view of the world, will not necessarily correspond to what will appear to be the simplest explanation to another scientist working with a different corpus on information (as well as with different cultural and personal prejudices, intellectual skills, et cetera). If we admit to this much by way of scientific relativism, are we then bound to conclude that "anything goes"[2]? It is here that parsimony becomes a virtue, as it reigns in the wildest flights of fancy and, we hope, enables us to extricate ourselves more easily from yesterday's misconceptions. For one thing, theories that are parsimonious are more likely to satisfy the falsifiability criterion.

Occam's razor in Ockham's time

William of Ockham (*c.* 1285-1349) is of course is not to blame for our bungling equating of simplicity and truth. In fact, the principle of parsimony does not even originate with him. As is so often the case, we can find the beginnings with Aristotle, who wrote in his *Posterior Analytics*: "Let that demonstration be better which, other things being equal, depends on fewer postulates or suppositions or propositions."[3] Commenting on this passage, Robert Grosseteste (*c.* 1168-1253) expanded this to a more sweeping “that is better and more valuable which requires fewer, other circumstances being equal.”[4] For Grosseteste, the logical conclusion of Aristotle’s maxim was that best of all is that which requires no assumptions whatsoever. Here we may reflect that for medieval philosophers, who were first and foremost men of faith, those truths that were immediately perceptible as such – that could be unequivocally apprehended by the human mind – took center stage in their metaphysics. By contrast, modern scientists and mathematicians have learned to be especially weary, at least outside the arena of theology, of “axiomatic truths” that would appear to be self-evident. For instance, Frege was quite convinced of the obvious consistency of his formalization of set theory, and yet Russell famously found a fatal flaw, which a hastily penned addendum could not repair[7]. Furthermore, whenever a proposition is undecidable within the framework of a formal system, that system can be extended by adding either the undecidable proposition or its negation as an axiom, in either case without loss of consistency, which makes clear that the technical meaning of “axiom” has strayed from its original meaning of “self-evident truth” (from *ἄξιος* meaning “worthy” as in: worthy of consideration).

Grosseteste did not just produce a translation of the complete works of Aristotle (complete with extended commentary), but he also performed original research on light and colour, which can be appreciated as an early adumbration of the scientific method.[4] From a modern perspective, we are struck particularly by Grosseteste’s promulgation of mathematics as a unifying framework for the description of nature. Indeed, it could be argued that the adoption of mathematics by natural philosophers is itself a far-reaching expression of the principle of economy.

For Ockham, however, parsimony was primarily a matter of ontological minimalism, which he may have derived from his near-contemporary Duns Scotus (*c.* 1265-1308)[5]. Ockham viewed abstractions and generalizations as mental concepts derived from the perceiving of particulars, and discerning among these similarities, affinities, derivations and the like. Conception amounts to the act of understanding individual objects in certain ways, each such way corresponding to one particular concept. But abstractions (in the widest sense of the word) were *not*, for Ockham, to be regarded as entities in their own right. Such entities, called universals, left classical logic hamstrung – and universals were but one problem among many that made the ancient syllogistic logic unwieldy, cumbersome, and consequently far less powerful than the much more streamlined logical systems we use today[6]. Ockham also observed that NOT (A AND B) ≡ NOT A .OR NOT B and NOT (A OR B) ≡ NOT A .AND NOT B, which today we call the fundamental Boolean equivalences.

In a similar vein, Ockham rejected the distinction between essence and existence, thought by some to be required to distinguish the creator from his creatures[5]. For Ockham it sufficed to note that God’s essence and existence were co-extensive in virtue of the necessity of God, whereas any other creature’s existence is caused by divine will. Ockham also applied his razor to the “faculties” of the human mind[5]. For instance, willing (volition) as opposed to understanding or perceiving (intellect) were at the time regarded as being as distinct as, say, the sensory modalities of hearing or smelling (which we know to be based on distinct anatomical structures). Ockham viewed willing and understanding simply as two different acts of the rational soul, and similarly for other mental faculties[5]. Generalizing, we might say that his universe consisted of fewer things, each of which could do more, thus breaking with the traditions of medieval thought and giving a distinctly modern feel to his arguments.

Ockham’s ontological minimalism constituted a new way (*via moderna*, as opposed to the old way, *via antiqua*) that came to be known as *nominalism* or *termism*[5]. Even if his philosophical system was not fully satisfactory, we can recognise with the benefit of hindsight that his thinking displayed considerable affinity with modern analytical philosophy. Although his efforts were almost exclusively bent toward religious matters, as we might expect of a 14th-century Franciscan, his striving to achieve the maximum possible with the smallest possible set of conceptual tools feels congenial to our own intellectual predilections.

It seems hardly likely that Ockham deliberately set out to revolutionize (or even just streamline) Western thought. His primary motivation seems to have been religious, in particular to safeguard Christian doctrine from Greco-Islamic necessitarianism, which posits that divine will is itself still subject to moral and ethical norms[5]. On this principle, there are moral imperatives related to good versus evil, right versus wrong, that one has to accept as being in some sense greater than God insofar as they impose boundaries on his will.

The proper Christian perspective, for Ockham, was that God is restricted *only* by logic: God cannot perform acts that are logically impossible, but this is truly the only constraint on his powers[5]. Although this move should technically count as a liberation of God, the mystics that reacted against Ockhamism appear to have felt that Ockham had made logic itself greater than God; these reactionaries disapproved of what they saw as Ockham's arid and ultra-refined scholasticism[5].

Ockham's insistence on the primacy of logic and ontological economy, in which we now perceive the seeds of later developments, did land him in hot water. Freed from necessitarianism, God *could* compel a person to do what would normally be regarded as evil, and the question whether that act would then constitute a sin was an issue that exercised his contemporaries[5].

For Ockham, God *could* also confound a person's senses and cause false perceptions, since it followed from divine omnipotence that whatever God does by means of secondary causes he can do without them (in modern parlance we might say that God can bypass the laws of physics, the latter merely being his default way of making the world happen) [5]. This was another vexed matter since truths immediately apprehended by the senses (and the intellect) were, for many, metaphysically essential and a prerequisite for faith. The idea that God *could* cause a false experience, be it a sensory hallucination or the bliss of revelation itself, was deemed highly upsetting if not blasphemous.

To make matters worse, Ockham questioned the possibility that the immortality of the soul could be proven, or even that the existence of God could be demonstrated, if by God we understand an entity that is absolutely supreme AND perfect AND unique AND infinite[5]. On the other hand, if by God we merely mean an entity unsurpassed in perfection and nobility, existence is trivially provable but uniqueness is not; alternatively, if by God we mean the "first efficient cause" then he must exist by logical necessity – in this Ockham followed the consensus of his time[5].

Such doggedly logical points of view were not well received by the church hierarchy. We may well imagine a situation in which, on the one hand, we have the sincerely devout Ockham who in the light of his new way of thinking (*via moderna*), imagines that he has only underlined the primacy of faith, which for him has become the only way in which a mortal creature can apprehend that, in point of fact, there *is* a God who is absolutely supreme AND perfect AND unique AND infinite (and who does not confound our senses because, *as it happens*, he is benevolent), and on the other hand, we have Ockham's superiors who as a matter of practical intelligence perceive all too well how Ockham's teachings were bound to go down with the masses. It will, then, not come as a huge surprise that Ockham never received the licence to teach; he was to remain a lowly "inceptor" and effectively denied full professorship because he was too far ahead of his time[5].

Ockham's razor primarily served to sever faith from rational enquiry. This instigated the separation of theology from philosophy (and hence what we call natural science), a split that soon became more prominent and gained support from thinkers belonging to various religious orders (often mendicants)[5]. From our modern point of view, we could readily view this is as a first step towards, and perhaps even a precondition for, the later developments of the renaissance and the Enlightenment.

Modern elaborations of Occam's razor

We saw in the foregoing section that Occam's razor could with equal justice be called Aristotle's maxim or perhaps Grosseteste's parsimony. If it has become associated with Ockham's name, this is because of the wide-ranging impact that his ontological minimalism had on 14th-century thought. Many of Ockham's preoccupations are looked upon quite differently today. For instance, Islamic necessitarianism is hardly viewed now as the most pressing threat to the propriety of Christendom, modern neural science is comfortable with the idea of mental faculties as different activities performed by the brain (although the issue of anatomical or histological correlates remains topical), and the paradoxes of omnipotence no longer appear to incite ecclesiastical schisms.

Thus, for us, Occam's razor boils down to merely the general guiding principle that had already been enunciated by Aristotle, and whose prominence in modern scientific thinking was perhaps sealed by Newton's admonishment that "we are to admit no more causes of natural things than such as are both true and sufficient to explain their appearances […] Nature is pleased with simplicity, and affects not the pomp of superfluous causes […] *hypotheses non fingo*."[8] Ockham was no pantheist[5], but if we allow ourselves some sort of identification of the divine with nature or the laws of nature, which are here being personified by Newton (if only for rhetorical reasons), we may begin to perceive a continuity of thought. A pantheistic rendition of Occam's razor might read: "do not presume too much on the mind of God."

Although the idea is clear enough, a worry remains: we are not to multiply hypotheses unnecessarily, but how does one actually count hypotheses? As every working researcher knows, hypotheses come in complex networks of similar ideas with all manner of variations and interdependencies. Counting them is by no means straightforward: it is rather like counting clouds. Moreover, if we are faced with the choice between two plausible hypotheses of modest scope and a single outrageous one, which option is the more economical one? Is there a way to put the principle of parsimony on a more rigorous footing? We shall briefly discuss three promising avenues.

The Large Deviation principle

When we consider the "typical case" say corresponding to the average (or perhaps the mode) of the statistical distribution of some quantity of interest, and we scrutinize the relationship between the concentration of probability mass and the underlying combinatorics, we often find that this concentration is due to the fact that values near to the mean (or the mode) can be brought about in many different ways, e.g. in terms of configurations of the underlying system. The numerical imbalance between the "typical case" and any other cases (which are variously called "deviations" or "fluctuations" or simply "the tail") can become truly overwhelming in statistical physics and gives rise to the remarkable robustness of thermodynamical principles[9].

Once a deviation (that is to say an *a priori* very unlikely event) is observed, the situation is radically altered: although, as with the mean, the deviation could have come about in many different ways, there tends to be one particular way that is overwhelmingly more likely (*given* that the deviation has occurred) then any of the other ones[10]. The other ways dwarf this one special case in their number, but are in turn suppressed by their being exponentially less likely, which renders them irrelevant even in the aggregate[10]. In other words: if an exceedingly unlikely fluctuation does happen, we can be exceedingly sure of the way in which it happens. The word "exceedingly" in the previous sentence gains in force as we move out into "the tail" – around the mean, where fluctuations are rare but as yet not all that unlikely to happen, the singling-out effect is weak. (The mathematically inclined reader will have intuited that these broad, sweeping statements are the verbal upshot of a series of precise technical statements taking the form of limit theorems[10].)

The universality of this principle, which has numerous applications in such disparate fields as statistical physics[9], critical transitions[11], and mathematical immunology[12], has led to its being given a special name: The Large Deviation Principle (LDP)[10]. It is perhaps not as famous as the Central Limit Theorem, but it deserves to be.

How is the LDP relevant to Occam's razor? Well, that one particular way in which the unlikely event is virtually bound to have happened given that, against the odds, it did – that one particular way also corresponds to the *least complicated* way in which it could have happened. This is admittedly an imprecise and contentious statement: first, it begs the question of defining complexity, which has not yet been accomplished, and second, if a suitable definition of complexity were at hand, we might find that the state of affairs pointed to by the LDP is not always the "simplest". All the same, the LDP configuration does tend to correspond to what one would intuitively call the simplest way.

We could turn this on its head and *define* simplicity as whatever the LDP identifies as the most likely. This is by no means as outrageous as it might seem. For one thing, we would end up with a concept of simplicity that ties in well with our intuition, and for another thing, this new concept would be as unambiguous and rigorous as the LDP principle itself. We would be able to forge a strong link between overwhelming likelihood and simplicity, i.e. parsimony.

It is tempting to put the LDP forward, however tentatively, as a modern version of the principle of parsimony. However, there is one obvious fly in the ointment: the LDP is about large deviations, that is to say, large fluctuations away from the norm. How can this be relevant to the discovery of natural law, which after all is primarily concerned with characterizing "the norm"? There is no satisfactory answer to this question, but let us venture a provocative answer: our entire universe is *a priori* tremendously unlikely – according to Roger Penrose, for every way the universe could have existed at the instant of the Big Bang, there are exp{exp{284}} *other* ways for the universe (at any instant) to be[13], and this vast number is probably a gross underestimate in the context of the present discussion. Consequently, the way the world is put together is as the LDP tells us it should be. We find ourselves straying back into theological territory, cosmogony in particular, as we appear to be presupposing a statistical distribution over an ensemble of universes. The idea of a multiverse seems to throw all ontological economy to the wind, and one may well question its cogency.

Kolmogorov complexity

As traditionally formulated, the principle of parsimony seems to require that we rank and compare rival theories by simply counting their hypotheses, suppositions, assumptions, and the like. If counting, as such, does not appear to be well-defined, it makes sense to cast about for a more suitable numerical measure by means of which we might quantify how "involved" a theory is.

One promising candidate is Kolmogorov complexity[14]. Given any data object, we may consider the computer programs that produce that object as output. The length of the shortest program among these is the *Kolmogorov complexity* of that object. The basic idea is that the effort it takes to specify a thing tells us how complex that thing really is. Kurt Gödel was already thinking among similar lines in 1936 when he considered the length of proofs as a measure of complexity[15]. The Kolmogorov complexity is not to be confused with the size in raw bits of the data object: if the latter is large, but very regular in its structure, the shortest program needed to produce it can be much shorter. One may think that the choice of programming language matters, but this turns out to be unimportant: if two programming languages are equally powerful, one can always add a conversion module (of bounded length) to the code, and the extra burden drops out as soon as we start doing comparisons between data objects[14].

Now if the Kolmogorov complexity of a binary data object x is $K(x)$, the *algorithmic probability* of x is defined as $2^{-K(x)}$ (here $K(x)$ is the variant of Kolmogorov complexity known more specifically as algorithmic prefix complexity[14]). This probability is of the same order ("big-O," for the mathematically inclined reader) as the universal *a priori* probability of the data object[14]. More complex thus becomes tantamount to less likely, again providing the "Occam connection" we have been looking for.

Whereas the universal *a priori* probability accumulates the probabilities over all programs that produce x, the algorithmic probability is based on the shortest one. The probabilities nonetheless turn out to be pretty much the same. This can only happen because the universal *a priori* probability is dominated by the contribution from the shortest ("simplest") one[14], an effect strongly reminiscent of the LDP.

If theories can be encoded as data objects, their Kolmogorov complexity gives us a means to compare their parsimoniousness. Again one might quibble about the encoding language that is used, but once more the language chosen only adds a constant to the complexity that cancels when comparisons are being made. So that answers our initial question on how to define complexity.

However, things can be taken further. Consider the data that are available regarding one particular phenomenon: we can regard this as a corpus-to-date, a data object that grows in size every time observations are made, experiments and measurements being done, and so on. The Kolmogorov complexity of this corpus will change as well every time information is being added to it. In fact, inasmuch as a "lawlike" theory of the phenomenon is possible at all, we should expect that the Kolmogorov complexity tops out and levels off. The reason for this is that the shortest possible program reproducing the corpus is equivalent to a formal theory accounting for the phenomenon at hand. Thus, as soon as sufficient data have been added to the corpus to determine such a theory (bearing in mind our introductory example, we reflect that this may require a substantial mass of data), the shortest program will "stabilise" and the corresponding Kolmogorov complexity should no longer increase. This stabilized shortest program is known as the *Occam* for the data corpus[14].

In much the same way that Turing machines render the concept of effective computation unambiguous, but are poor blueprints for actual calculation devices, the Kolmogorov complexity of a growing data set is an excellent conceptual device to elucidate the principle of parsimony in the process of theory discovery, but not necessarily a suitable recipe for actual "automated" science. In fact, we should not expect computers to take over the scientific discovery process any time soon. Humans (as opposed to an imaginary "Kolmogorov computer") have been able to construct superb theories on the basis of comparatively scant data[13], possibly because they bring the imagination and intuition to the game that allows them to make the correct conceptual leaps, and think of the experiment that will decide the crucial question. This psychological aspect of creative genius is obviously missing from the Kolmogorov account, in fact deliberately so, since the objective was just to isolate the algorithmic-informational side of the problem.

Incidentally, the foregoing argument does not rule out the possibility of building machines capable of scientific discovery (or any other aspect of human thought). Indeed, insofar as the mind is the expression of a physical process, we ought to expect that this process could equally well be set up in an artificial contrivance (whether or not we would still be comfortable calling such a device a "computer"). One may go one step further and, citing the celebrated Church-Turing thesis, maintain that mental processes are emulatable/simulatable by a classic computing device[16].

Returning to the concept of the "Occam program," we may consider two roughly equivalent ways of looking at the asymptotic behaviour of the Kolmogorov complexity as the data corpus grows, which are interesting since they come each with their own flavour and connotations. One is that of *extrapolation*, the other that of *data compression*. As regards extrapolation, let us consider again the corpus as it grows whenever observational data are added. At each stage we have the best candidate-Occam thus far, i.e. the shortest program associated with the Kolmogorov complexity of the data corpus. Letting this candidate produce the outcome of the next experiment (i.e. extrapolation or prediction), and then actually performing the experiment, we find ourselves able to compare prediction and outcome and hence obtain an idea of how close we are to the "leveling off" point. The world being noisy, things are complicated to no small extent by the fact that we have to allow for some margin of error[14].

As regards data compression, we observe that the data object encoding the shortest program (along with some suitable inputs etc.) contains all that is worth knowing of the original object. This allows for substantial savings when the corpus data are to be transmitted over some channel of communication. We transmit the (much shorter) program object instead of the full data object, and the receiver "unzips" it. The saturation point of the data corpus is achieved when the compressed counterpart of the growing data object stops increasing in size. Incidentally, data objects that resist compression – whose Kolmogorov complexity continues to increase in proportion to their raw size – can for that very reason be defined as "devoid of pattern," and a rigorous notion of randomness can be developed, taking this as a starting point[14].

Perhaps the receiver actually finds the short object more useful anyway, in much the same way that we prefer the formulas $F = m{\cdot}a$ or $E = m{\cdot}c^2$ to the respective (and sizable) corpora of data that warrant these laws. We may even come to think of such "laws" as extremely succinct summaries of experimental data sets: a valid point of view, if not to everyone's taste.

Bayesian inference

LDP and Kolmogorov complexity both lend support to the idea that "simplest is likeliest" and lead us to suspect that likelihood should take logical precedence, with simplicity being secondarily defined in terms of wherever likelihood takes us. An approach that is entirely grounded in the idea of likelihood departs from a fundamental equality in mathematical statistics[17]:

$$P[A|B]\times P[B] = P[A \text{ AND } B] = P[B|A]\times P[A]$$

where A and B stand for any two events and P[…] means probability of … and P[A|B] denotes the probability of A, given that B occurs. This can be rearranged as P[A|B] = P[B|A]×P[A]/P[B] which is a simple form of "Bayes' Law"[17].

To apply this to the problem at hand, we replace A by some hypothesis H_i and B by the corpus-to-date of data D, and we have an expression for the likelihood of the hypothesis given the data:

$$P[H_i|D] = P[D|H_i]\times P[H_i]/P[D]$$

where we may compute P[D] as $\sum_i P[D|H_i]\times P[H_i]$ using the principle of total probability[17]. The idea is now that we have data D in hand and keep this as a constant; we seek to maximize $P[H_i|D]$ over all hypotheses H_1, H_2, …, H_i,… and the argmax of this procedure is the hypothesis to be preferred.

Conceptually, the term $P[D|H_i]$ is not problematic. Even though it might well be the case that capturing H_i in the form of a suitable mathematical model requires substantial skill and intellectual resources, or that the actual calculation of $P[D|H_i]$ imposes a huge strain on the available computational resources, the matter is nonetheless pretty straightforward from a philosophical/metaphysical perspective.

Things stand dramatically different with the term $P[H_i]$: how can we interpret the probability of a hypothesis *an sich* in a meaningful way? Bayesians like to point out that $P[H_i]$ is simply where we left things the last time we updated $P[H_i|D]$ when the corpus D has data added to its mass, and it is in this idea of incremental updating and improvement that the present argument connects to the phenomenon of Kolmogorov complexity levelling off as data keep being added. Another reasonable point made by the Bayesians is that as D grows, $P[H_i|D]$ settles on a value that is independent of the starting point. Still, there remains that niggling problem of choosing an appropriate starting distribution in the first place…

As far as ontological economy goes, the notion of a prior statistical distribution over a space of hypotheses looks utterly wasteful indeed. It rather makes our earlier speculations about multiverses look like child's play. Bayes himself seems to have preferred the notion of assuming a uniform distribution[14]. Philosophical misgivings[18] aside, the pragmatic advantage becomes immediately obvious when we take logarithms:

$$\ln P[H_i|D] = \ln P[D|H_i] + \ln P[H_i] - \ln \sum_i P[D|H_i]\times P[H_i]$$

where the latter two terms cancel whenever we consider the differences between any two candidate hypotheses. Also, if we restrict ourselves to a space of hypotheses that are structurally the same and only differ with respect to the values of their parameters, we find that maximizing just the term $\ln P[D|H_i]$ becomes substantially more tractable. Moreover, if the prior probability of the data D (as obtained) is low, we should expect the quantity $\ln P[H_i|D]$ to be governed by the LDP, which is perhaps another indication that we should focus our attention on this term.

If we reject the idea of a universal uniform prior, we may look to Kolmogorov complexity to derive *a priori* probabilities. We saw that in the context of this theory, both H_i and D are treated as instances of data objects where we implicitly think of them as essentially defined by the programs that produce them as output. In turn, these programs are themselves nothing more than data objects that, as such, can be systematically enumerated. This systematic enumeration creates an *a priori* objective universe of "machines" that allows us to work out at which frequencies finite data strings are produced as initial segments of the output of the machines, when fed certain suitable inputs. These various claims are all rather loosely put here, as the non-trivial technical details would take us too far afield[14]. The upshot is that it is possible to define a "universal" statistical distribution over the space of hypotheses per se. This distribution is instrumental in the evaluation of the predictions produced by a candidate-Occam program, as discussed above[14]. In fact, since the enumeration has to weighed in one way or another, the construction can be said to let the assumption of a uniform prior distribution back in through the backdoor, and as such it remains susceptible to the same philosophical objections, [18] mitigated by the fact that the uniformity assumption is now being made at an extremely deep level, which ought to bestow some measure of objectivity.

Prospects

Nominalism is nowadays considered to be untenable, at least in any of its canonical formulations[19]. Nevertheless, the move from medieval Scholasticism to modern empiricism constituted a turning away from intimate communion with the true inner nature of things, and towards the primacy of observations and confining theories to "mere" description. The saving grace of the latter is the astounding succinctness that can be achieved. The formal apparatus of modern physics achieves an awful lot with, well, next to nothing, daunting as that nothing may seem to the uninitiated[13]. This sea change in Western thought can, with hindsight, be seen to hinge on Ockham's penchant for ontological minimalism (along with similar arguments found with Robert Grossteste, Roger Bacon, and others[5]).

Modern data scientists are searching for a more thoroughgoing formalization and instrumentation of Occam's razor. As we have seen, promising developments are found in the LDP, Kolmogorov complexity, and Bayesian inference. The numerous points of commonality that we have encountered among these approaches seem to indicate that we are only catching glimpses of a more mature theory in which all these ideas find a satisfactory and elegant unification.

References

[1] Baake, M. (1997) Solution of coincidence problem in dimensions *d*≤4. pp. 9–44 in R. V. Moody, ed., The Mathematics of Long-Range Aperiodic Order, Dordrecht: Kluwer.

[2] Feyerabend, P. K. (1975) Against Method: Outline of an Anarchist Theory of Knowledge. New York: New Left Books.

[3] Barnes, J. (1984) The Complete Works of Aristotle (The revised Oxford translation, Vol. 2). Princeton: Princeton University Press.

[4] Crombie, A. C. (1953) Robert Grosseteste and the Origins of Experimental Science 1100–1700. Oxford: Oxford University Press.

[5] Copleston, F. C. (1972) A History of Medieval Philosophy. Notre Dame: University of Notre Dame Press.

[6] Russell, B. (1945) A History of Western Philosophy. London: George Allen & Unwin.

[7] Quine, W. V. O. (1955) On Frege's Way Out. Mind **64**: 145–159.

[8] Newton, I. (1713) Philosophiæ Naturalis Principia Mathematica, 2nd Ed. *Phil. Trans. Roy. Soc.*

[9] Dorlas, T. C. (1999) Statistical Mechanics: Fundamentals and Model Solutions. Boca Raton: CRC Press.

[10] Hollander, F. den (2000) Large Deviations. *Fields Institute Monographs* **14**.S.

[11] Boettinger, C. and Hastings, A. (2013) No early warning signals for stochastic transitions: insights from large deviation theory. *Proc. Biol. C.* **280(1766)**: art no 20131372.

[12] Zint, N., Baake, E. and Hollander, F. den (2008) How T-cells use large deviations to recognize foreign antigens. *J. Math. Biol.* **57**: 841–861.

[13] Penrose, R. (2004) The Road to Reality. London: Jonathan Cape.

[14] Li, M. and Vitányi, P. (2008) An Introduction to Kolmogorov Complexity and its Applications. Berlin: Springer.

[15] Gödel, K. (1936) Über die Länge von Beweisen. *Ergebnisse eines mathematischen Kolloquiums* **7**: 23–24.

[16] Hofstadter, D. R. and Bennet, D. C. (1992) The Mind's I: Fantasies and Reflections on Self and Soul. New York: Bantam Books.

[17] Bain, L. J. and Engelhardt, M. (2000). Introduction to Probability and Mathematical Statistics. Boston: Cengage Learning

[18] Gilles, D. (2000) Philosophical Theories of Probability. London: Routledge.

[19] Peñaranda, J. J. L. (2013) Ontology: minimalism and truth-conditions. *Philos. Stud.* **162**: 683–696.