proceedings

MDPI

Proceedings

Towards Demystifying Shannon Entropy, Lossless Compression and Approaches to Statistical Machine Learning †

Hector Zenil [1,2,3]

1 Information Dynamics Lab, Karolinska Institute, 171 76 Stockholm, Sweden; hector.zenil@cs.ox.ac.uk
2 Oxford Immune Algorithmics, Reading RG3 1EU, UK
3 Algorithmic Nature Group, LABORES, 75005 Paris, France

† Conference Morphological, Natural, Analog and Other Unconventional Forms of Computing for Cognition and Intelligence (MORCOM), Berkeley, CA, USA, 2–6 June 2019.

Published: 19 June 2020

Abstract: Current approaches in science, including most machine and deep learning methods, rely heavily at their core on traditional statistics and information theory, but these theories are known to fail to capture certain fundamental properties of data and the world related to recursive and computable phenomena, and they are ill-equipped to deal with high-level functions such as inference, abstraction, modelling and causation, being fragile and easily deceived. How is it that some of these approaches have (apparently) been successfully applied? We explore recent attempts to adopt more powerful, albeit more difficult methods, methods based on the theories of computability and algorithmic probability, which may eventually display and grasp these higher level elements of human intelligence. We propose that a fundamental question in science regarding how to find shortcuts for faster adoption of proven mathematical tools can be answered by shortening the adoption cycle and leaving behind old practices in favour of new ones. This is the case for randomness, where science continues to cling to purely statistical tools in disentangling randomness from meaning, and is stuck in a self-deluding pattern of still privileging regression and correlation despite the fact that mathematics has made important advances to better characterise randomness that have yet to be incorporated into scientific theory and practice.

Keywords: Shannon entropy; machine learning; algorithmic complexity; Kolmogorov complexity; feasibility; LZW; causality vs. correlation; Algorithmic Information Dynamics

1. Introduction

Scientific evidence, even in the case of the greatest scientific advances, relies for the most part on theories developed decades, if not centuries ago, and it does not necessarily build upon these theories, but often uses them in misleading ways. We are barely catching up with a theory such as Bayes', which dates back 250 years, or with Shannon's classical information theory, which was developed 80 years ago. While Shannon's theory has seen a lot of developments, from Granger causality to transfer entropy and Partial Information Decomposition, and Bayes' too, in the form of Bayesian networks and causal diagrams, they are still fundamentally reliant on classical probability and traditional statistical approaches. Indeed, Shannon entropy continues to be widely used to quantify randomness despite the mathematical consensus that established that randomness can only be characterised by algorithmic randomness (or algorithmic complexity) because recursively generated data will be misidentified as random by a purely statistical (entropic) approach. A fundamental question in science should concern how to find shortcuts to the adoption of new mathematical tools to be incorporated more quickly into practical methods.

Researchers in the field of Algorithmic Information Theory (AIT) can roughly be divided between, on the one hand, those who study the algorithmic randomness of infinite sequences and degrees of randomness, and, on the other hand, those interested in the applications of algorithmic complexity to various areas of science, when it comes to supplying causal explanations.

It is my strong belief, however, that AIT has the potential to make an impact on the way we approach science, on our currently predominant data-reliant—as opposed to model-centred—scientific method. Unfortunately, AIT has had little impact to date, and has often taken the wrong direction when applied to different areas of science, having focused for far too long on weak methods, and sometimes misleading assumptions, some of which we explore here.

2. Fooling Ourselves

The following quotation attributed to Arnold Sommerfeld can easily be applied to what has happened in the case of statistical lossless compression algorithms such as LZW [1], to their misuse and abuse in applications of algorithmic complexity, and is all the more relevant given the connection of LZW to entropy, the latter being what Sommerfeld is talking about (in responding to a question about why he never wrote a book [2]):

> Thermodynamics is a funny subject. The first time you go through it, you do not understand it at all. The second time you go through it, you think you understand it, except for one or two small points. The third time you go through it, you know you do not understand it, but by that time you are so used to it, it does not bother you any more.

This opinion can be used also to characterise the way in which popular lossless compression algorithms such as LZW are exploited, used and abused to claim connections to algorithmic complexity quite fooling ourselves.

Let me illustrate a common misconception concerning Shannon entropy. At a conference on computability where I was an invited speaker, I gave a talk on algorithmic complexity related to causality, and I used as an example of a random variable a process in which an observer is looking at a stream of digits emerging in their natural order from the digits of the mathematical constant π in a given base (say base 10), with a view to determining what kind of object it is. The experiment assumes that the observer is not already in possession of the information that the object in question is in fact π, so one can model it with a random variable. Immediately, an expert in both Shannon entropy and algorithmic complexity intervened aggressively, claiming that one could not take π as a random variable because π was a deterministic constant. Not only did this expert miss the point of the thought experiment, but despite his mastery of the subject, he did not realise something elementary that speaks to the weaknesses of the measure, something that is germane to an elementary understanding of both entropy and algorithmic complexity. In all models in which a random variable is involved, these variables are variable only in an epistemological sense, as in my thought experiment. Indeed, the quintessential example of a random variable given in the context of an introduction to Shannon entropy in any textbook is the process of throwing a die or tossing a coin. Nothing in either of these processes is random according to classical mechanics, which is believed to govern every aspect of the evolution of both coins and dice. Indeed, the precise initial conditions, basic force and gravitation laws fully determine the dice and coin outcomes, making them no different from my experiment with π. It is merely that we do not have access to the knowledge that π is in fact π, just as we have no access to the full initial conditions of the coin or die, forcing us to model these processes with random variables. In [3], we explored a specific example of a network to demonstrate how fragile is Shannon entropy in the face of language description and object representation and how advisable is to move away from using exclusively statistical measures (e.g., Shannon entropy or LZW) and put so little to no effort in methodological innovation. Indeed, when characterising an object by Shannon entropy (or LZW), one has to make arbitrary decisions, such as the choice of feature of interest (e.g., for a network, the degree sequence vs. say its adjacency matrix) that can describe the same object but in different

ways and for each retrieve different and divergent values of, e.g., Shannon entropy. What we did [3] was to construct a network that had a degree distribution near maximal entropy that determined a single growing network that when described by its sparse adjacency matrix would suggest it actually had an entropy limit equal to zero. Shannon entropy is subjective and epistemological; it measures our ignorance and difficulty estimating the underlying probability distribution, but even experts in the field often fail to appreciate this and end up fooling themselves.

For example, in what is considered a landmark paper [4], certain conceptual and foundational mistakes were made and overlooked. The paper claims to instantiate a form of integrated information, it claims to have found an application, but it is bedevilled by the same fallacy, offering what is claimed to be a measure of causality (and even consciousness), but one which is arrived at using tools of a statistical nature that are no more powerful than regression and correlation. In a similar fashion, landmark papers that made early contributions to the area using information distances can be reproduced resorting to measures related to Shannon entropy, and on some occasions simpler methods, such as GC content for genome sequences. The damage done by the use, misuse and abuse of LZW to "estimate algorithmic complexity" still haunts us to this day.

Machine learning is no stranger to these weaknesses. At the core of most approaches in machine learning and deep learning there is a function known as cross-entropy, one of the more sophisticated functions (beyond those based on simple distances). The cross entropy between two probability distributions is a function that measures the average number of bits needed to identify an event drawn from each distribution, one being assumed to be ground truth and the other the one to be approximated by reducing its distance according to the statistical function. It has also widely believed that the power of deep learning resides in the continuous nature of the search space on which differentiable manifolds and methods from continuous calculus are needed. We have shown that this is not the case [5] and that neither continuity nor differentiable search methods are needed to perform a successful exploration. Moreover, the methods in [5] give a glance of how symbolic computation can reach similar success to traditional deep learning but requires exponentially more resources at it does implement a model-driven approach that requires them to build causal chains and state-to-state representations.

The lossless compression approach to algorithmic complexity worked for reasons not related to the theory of algorithmic complexity (or as a result of it), but because of the already established connection with Shannon entropy. Indeed, statistical compression algorithms cannot capture all (non-)algorithmic random properties in data, not only because they are not universal in the Turing sense, but also because they are designed only to spot repetitions. However, accounting for non-statistical regularities should be crucial, since these regularities represent the main advantage of using algorithmic complexity over, for example, Shannon entropy with a uniform probability distribution. There are many papers that use Shannon entropy (and even popular lossless compression algorithms) to analyse data and do not make any false connection to algorithmic complexity.

A thorough review of these challenges, limitations and new directions of research to make progress in the application of algorithmic complexity can be found in [6].

3. How Misconceptions Can Drive a Scientific Agenda and Vice Versa

We used to think that science was entirely objective, but history has taught us that it is also driven by community choices. Whether these choices lead to radical effects is another matter, but it is undeniable that this happens at all levels in the practice of science.

One common dictum, for example, would have it that complex systems are those in which interactions cannot be accounted for in a reductionist fashion, that is, those in which there are too many interactions among too many elements to be accounted for individually when attempting to explain their overall behaviour and the fact that they produce a result greater that the sum of their component elements.

This is all true, but some researchers have indicated that this was already the case before the advent of complex systems, as distinct from areas such as dynamical systems. Indeed, in the theory

of dynamical systems, there are objects and processes for which very small changes produce very different behaviour, without their being any need for interacting elements or too many elements, just trivial equations and mappings operating on some continuous space or time.

Discrete chaotic systems, such as the logistic map, can exhibit strange attractors whatever their dimensionality and continuous dynamical systems; the Poincaré–Bendixson theorem establishes that a strange attractor can only arise in three or more dimensions. Indeed, finite-dimensional linear systems cannot show complex (chaotic) behaviour. For a dynamical system to display chaotic behaviour, it must be either nonlinear or infinite-dimensional. However, in all cases, continuous space and/or time is a requirement by definition, as chaotic behaviour in dynamical systems can only be defined for a ϵ of arbitrarily small length, quantifying the distance among initial conditions that make the system diverge. Millennia-old algorithms such as Euclid's division algorithms can also be said to be of this type; the algorithm itself is computable and outputs can diverge for very small changes of the quotient or dividend (e.g., the sequence of remainders), but the algorithm's domain is the continuous field of real numbers.

In contrast, in the theory of computation, everything is discrete at any given point, in both space and time, and at all times, so the fact that chaotic behaviour research and results may predate what appeared to be complex behaviour in the simplest of the computer programs over fully discrete regimes (under an adapted definition of behavioural divergence based on single bit distance enumerations of initial conditions [7]) came as a surprise when first disclosed by researchers such as Wolfram, as summarised in [8]. Indeed, this is what I believe set the area of complexity apart and led to its emergence as a field in its own right, namely the realisation that there are many layers of irreducibility beyond only uncomputability and unreachability, that of divergent qualitative behaviour intrinsic to even the most simple computer programs running in full discrete space and time and for the simplest of the initial conditions [8]. The promise of the field was that it would make a difference in science and that methods pointed out by Wolfram would deliver a new type of science. One major challenge, however, has been how to conduct a guided rather than blind and fruitless search for computable models other than toy models. What we have done is precisely to offer a first set of tools to perform such an exploration of the space of computable candidate scientific models, in the form of what we call Algorithmic Information Dynamics.

4. Algorithmic Information Dynamics

Algorithmic Probability and the Universal Distribution approach the challenge of algorithmic inference [9–11] from the standpoint of the theory of computation. Both algorithmic complexity and algorithmic probability are not computable but semi-computable, meaning that approximations from above and below are possible and are deeply related to each other, a formal connection being represented by the so-called (algorithmic) Coding theorem that establishes that a short computer program is also algorithmically highly probably and vice versa. Their uncomputability has meant that, for decades after their discovery, very few attempts were made to apply them to other areas of science, with most researchers' reaction to the field, especially as regards applications, being sceptical, their misgivings arising from the issue of uncomputability.

The notion behind AP is very intuitive. If one wished to produce the digits of π randomly, one would have to try time after time until one managed to hit upon the first numbers corresponding to an initial segment of the decimal expansion of π. The probability of success is extremely small: $1/10$ digits multiplied by the desired quantity of digits. For example, it would be $1/10^{2400}$ for a segment of 2400 digits of π. However, if instead of shooting out random numbers one were to shoot out computer programs to be run on a digital computer, the result would be very different.

A program that produces the digits of π would have a higher probability of being produced by a computer. Concise and known formulas for π could be implemented as short computer programs that would generate any arbitrary number of digits of π. We demonstrated in [12] how AP can account for up to 60% of the simplicity bias in the output distribution of subuniversal models of computation

when running on a set of random initial conditions, thus strongly suggesting that AP is of fundamental value in explaining processes in science, even when relaxing the criterion of Turing universality under which AP was conceived.

Indeed, algorithmic probability is of fundamental interest to science because it can address some of its most pressing challenges, such as inference, model generation and causation, which are the topics of interest in our research programme (beyond simple estimations of algorithmic complexity represented by a single real-number value). This relevance to science was pointed out by Vitányi and colleagues in a very engaging article [13], and, more recently, in research in areas such as AI and machine learning, by people such as Marvin Minsky, who claimed that Algorithmic Probability to be one of the most, if not the most important theory for applications [14]. Approaches to algorithmic complexity and algorithmic probability (AP) that take into consideration finite resources and applications have been proposed before, such as resource-bounded Kolmogorov complexity and Universal Search to approach algorithmic complexity in practice based on, for example, dovetailing all possible programs and cutting short runtimes.

Unlike most complexity measures that are designed for static objects, except those related to dynamical systems, the measure of algorithmic complexity I have led the introduction of is adapted for dynamical systems and designed to characterise the change of algorithmic complexity of an object evolving over time. The measure is universal in the sense that it can deal with any computable feature that a system may display over time, either spontaneously or as a result of an external perturbation/intervention/interaction.

At the core of Algorithmic Information Dynamics [15–17] (AID), the algorithmic causal calculus that we have introduced, is the quantification of the change of complexity of a system under natural or induced perturbations, particularly the direction (sign) and magnitude of the difference of algorithmic information approximations denoted by C between an object G, such as a cellular automaton or a graph and its mutated version G', e.g. the flip of a cell bit (or a set of bits) or the removal of an edge e from G (denoted by $G\backslash e = G'$). The framework is agnostic to underlying method but it has been relying on the Block Decomposition Method [18] and the Coding Theorem Method [21] motivated by Algorithmic Probability. The difference $|C(G) - C(G\backslash e)|$ is an estimation of the shared algorithmic mutual information of G and $G\backslash e$. If e does not contribute to the description of G, then $|C(G) - C(G\backslash e)| \leq \log_2 |G|$, where $|G|$ is the uncompressed size of G, i.e. the difference will be very small and at most a function of the graph size, and thus $C(G)$ and $C(G\backslash e)$ have almost the same complexity. If, however, $|C(G) - C(G\backslash e)| \leq \log_2 |G|$ bits, then G and $G\backslash e$ share at least n bits of algorithmic information in element e, and the removal of e results in a loss of information. In contrast, if $C(G) - C(G\backslash e) > n$, then e cannot be explained by G alone, nor is it algorithmically not contained/derived from G, and it is therefore a fundamental part of the description of G, with e as a generative causal mechanism in G. Or else it is not part of G but has to be explained independently, e.g. as noise. Whether it is noise or part of the generating mechanism of G depends on the relative magnitude of n with respect to $C(G)$ and to the original causal content of G itself. If G is random, then the effect of e will be small in either case, but, if G is richly causal and has a very small generating program, then e as noise will have a greater impact on G than would removing e from an already short description of G. However, if $|C(G) - C(G\backslash e)| \leq \log_2 |G|$, where $|G|$ is, e.g., the vertex count of a graph, or the runtime of a cellular automaton, G, then e is contained in the algorithmic description of G and can be recovered from G itself (e.g., by running the program from a previous step until it produces G with e from $G\backslash e$).

Algorithmic Information Dynamics is an algorithmic probabilistic framework for causal discovery and causal analysis. It guides a search for computable models and is an alternative or complement to other approaches and methods of experimental inference, such as statistical machine learning and classical information theory. AID is related to areas such as computational mechanics and program synthesis. However, unlike other methods such as Bayesian networks, AID does not rely on graphical models or calculations of mass probability distributions.

AID is the result of combining Algorithmic Information Theory, Causation and Perturbation Analysis where interventions are induced or simulated in an open system where external influences may be detected and identified by AID (if the signal to noise ratio allows it). AID has laid the foundations and devised methods to make Algorithmic Information Theory more applicable in scientific discovery and analysis in a wide range of areas ranging from dynamical systems to cognition to molecular biology and genetics, and it constitutes an alternative to estimating algorithmic complexity by popular lossless compression algorithms such as LZW, which are based on entropy rate estimations removed from some of the key features of algorithmic information.

We have shown how we can infer and reconstruct space-time evolutions by quantification of the disruptiveness of a perturbation [15,19]. We could then extract a set of generating mechanisms from the ordered time indices, from least to most disruptive, and produce candidate generating models. Simpler rules have simpler hypotheses, with an almost perfect correspondence in row order. Some systems may look more disordered than others, but locally the relationship between single rows is for the most part preserved (indicating local reversibility).

We have also shown, for example, how evolution seen as a search in the space of computable instantiations of 'living' organisms (simulated by computable processes) produces similar effects to unexplained phenomena such as diversity explosions and sudden extinctions. Even modularity can be explained in a natural algorithmic fashion [20].

5. Conclusions

Studying randomness from the computing perspective affords us a framework for studying the nature of the world that discerns the way in which patterns in the universe are distributed, and can help make a real difference and advance the practice of science and the validation of scientific evidence. These tools can significantly contribute to all areas of science, and we should find ways to speed up the translation and adoption of mature and validated methods from mathematics to science. Our algorithms have filled a void left by other heuristics and have contributed to enriching the discussion of approaches rooted in or better motivated by generative mechanisms, first principles and mature measures of algorithmic information.

References

1. Ziv, J.; Lempel, A. Compression of individual sequences via variable-rate coding, *IEEE Trans. Inf. Theory* **1978**, *24*, 530.
2. Angrist, S.W.; Helper, L.G. *Order and Chaos–Laws of Energy and Entropy*; Basic Books: New York, NY, USA, 1967; p. 215.
3. Zenil, H.; Kiani, N.A.; Tegnér, J. Low Algorithmic Complexity Entropy-deceiving Graphs. *Phys. Rev. E* **2017**, *96*, 012308.
4. Casali, A.G.; Gosseries, O.; Rosanova, M.; Boly, M.; Sarasso, S.; Casali, K.R.; Casarotto, S.; Bruno, M.-A.; Laureys, S.; Tononi, G.; et al. A Theoretically Based Index of Consciousness Independent of Sensory Processing and Behavior. *Sci. Transl. Med.* **2013**, *5*, 198ra105.
5. Hernández-Orozco, S.; Zenil, H.; Riedel, J.; Uccello, A.; Kiani, N.A.; Tegnér, J. Algorithmic Probability-guided Machine Learning On Non-differentiable Spaces. *arXiv* **2019**, arXiv:1910.02758.
6. Zenil, H. A Review of Methods for Estimating Algorithmic Complexity: Options, Challenges, and New Directions. *Entropy* **2020**, *22*, 612.
7. Zenil, H. Asymptotic Behaviour and Ratios of Complexity in Cellular Automata Rule Spaces. *Int. J. Bifur. Chaos* **2013**, *23*, 1350159.
8. Wolfram, S. *A New Kind of Science*; Wolfram Media: Champaign, IL, USA, 2002.
9. Solomonoff, R.J. Complexity–Based Induction Systems: Comparisons and Convergence Theorems. *IEEE Trans. Inf. Theory* **1978**, *24*, 422–432.
10. Solomonoff, R.J. The Application of Algorithmic Probability to Problems in Artificial Intelligence. In *Uncertainty in Artificial Intelligence*; Kanal, L.N., Lemmer, J.F., Eds.; Elsevier: Amsterdam, The Netherlands, 1986; pp. 473–491.

11. Solomonoff, R.J. A System for Incremental Learning Based on Algorithmic Probability. In Proceedings of the Sixth Israeli Conference on Artificial Intelligence, Computer Vision and Pattern Recognition, Tel Aviv, Israel, 26–27 December 1989; pp. 515–527.
12. Zenil, H.; Badillo, L.; Hernández-Orozco, S.; Hernandez-Quiroz, F. Coding-theorem Like Behaviour and Emergence of the Universal Distribution from Resource-bounded Algorithmic Probability. *Int. J. Parallel Emerg. Distrib. Syst.* **2018**, *34*, 161–180.
13. Kirchherr, W.; Li, M.; Vitányi, P. The miraculous universal distribution. *Math. Intell.* **1997**, *19*, 7–15.
14. Minsky, M. Panel discussion on The Limits of Understanding. In Proceedings of the World Science Festival, New York, NY, USA, 14 December 2014.
15. Zenil, H.; Kiani, N.A.; Marabita, F.; Deng, Y.; Elias, S.; Schmidt, A.; Ball, G.; Tegnér, J. An Algorithmic Information Calculus for Causal Discovery and Reprogramming Systems. *iScience* **2019**, *19*, 1160–1172.
16. Zenil, H.; Kiani, N.A.; Tegnér, J. Algorithmic Information Dynamics of Emergent, Persistent, and Colliding Particles in the Game of Life. In *From Parallel to Emergent Computing*; Adamatzky, A., Ed.; Taylor & Franci/CRC Press: Boca Raton, FL, USA, 2019; pp. 367–383.
17. Zenil, H. Algorithmic Information Dynamics, Scholarpedia. Available online: http://www.scholarpedia.org/article/Algorithmic_Information_Dynamics (accessed on 10 March 2020).
18. H. Zenil, S. Hernández-Orozco, N.A. Kiani, F. Soler-Toscano, A. Rueda-Toicen, A Decomposition Method for Global Evaluation of Shannon Entropy and Local Estimations of Algorithmic Complexity. *Entropy* **2018**, *20*, 605.
19. Zenil, H.; Kiani, N.A.; Zea, A.; Tegnér, J. Causal Deconvolution by Algorithmic Generative Models. *Nat. Mach. Intell.* **2019**, *1*, 58–66.
20. Hernández-Orozco, S.; Kiani, N.A.; Zenil, H. Algorithmically Probable Mutations Reproduce Aspects of Evolution, such as Convergence Rate, Genetic Memory, and Modularity. *R. Soc. Open Sci.* **2018**, *5*, 180399.
21. H. Zenil, F. Soler-Toscano, J.-P. Delahaye, N. Gauvrit, Two-dimensional Kolmogorov complexity and an empirical validation of the Coding theorem method by compressibility. *PeerJ Comput. Sci.* **2015**, *1*, e23.