BEYOND THE BITTER LESSON

VISHAL VASAN

Abstract. Sutton's "bitter lesson" holds that general methods which exploit computation have repeatedly outperformed methods that encode human domain knowledge, and that this pattern is likely to persist because computation scales exponentially while human insight does not. The essay is now frequently invoked to dismiss architectural alternatives to the dominant scaling paradigm, including neuro-symbolic hybrids, world models, and self-supervised joint-embedding architectures, on the grounds that any explicit structural commitment is a regression to the losing, hand-crafted side of the essay's dichotomy. We argue that this use of the essay rests on an equivocation between two senses of "general" that the essay itself does not clearly separate: not tailored to a single problem instance, and approaching context-free applicability. The no-free-lunch theorems show that the second sense is unattainable, and that every successful method, whether or not it is compute-scaled, encodes a substantive assumption about the structure of the problems it will meet. We propose that the distinction actually supported by the historical record is not general versus domain-specific but compressive versus enumerative: whether a method converges on a low-complexity, extrapolation-licensing structure, often expressible as an invariant of its domain, or merely accumulates an open-ended list of corrections to fit observed cases. We develop this distinction through the transition from Ptolemaic epicycles to Keplerian and Newtonian mechanics, and we show it is measurable inside trained neural networks, in the phenomenon of grokking, in structural probing of language models, and in the asymptotic account of such transitions given by singular learning theory. We conclude that the bitter lesson, properly restricted, is a claim about which agent is likely to find a domain's compressive structure first, and not a claim about what has been found, still less grounds for treating any non-scaling-first proposal as domain-specific hand-crafting.

Keywords. Bitter lesson, no-free-lunch theorems, minimum description length, singular learning theory, geometric deep learning, inductive bias.

1. Introduction

Sutton's short essay [24] makes an empirical claim about seventy years of research in artificial intelligence: general-purpose methods that leverage computation, principally search and learning, have repeatedly overtaken methods in which researchers encode their own understanding of a problem, and the margin has not been close. Chess fell to brute-force search rather than to hand-built positional heuristics. Go fell to self-play and search rather than to encoded strategic principles. Speech recognition and computer vision fell to statistical, data-driven pipelines rather than to hand-built phonological rules or edge detectors. Sutton calls the lesson bitter because it must be relearned by each generation of researchers, who are drawn, understandably, to the intellectual satisfaction of building their own insight into a system.

The essay is brief, cites four examples, and has since acquired a role in the field's discourse that its length does not obviously warrant. It is now invoked, rarely by Sutton himself, as a general objection to any architectural proposal that departs from "add more compute and data to the current recipe". Brooks raised this concern within days of the essay's publication, arguing that the convolutional architecture credited with defeating hand-built computer-vision pipelines is itself a piece of encoded domain knowledge, namely translational invariance, built in because natural images possess that symmetry [5]. More recently, Rieck has argued that the essay now functions in practice as a mechanism by which proposals are rejected for a perceived lack of scalability, often without examining whether that perception is well founded [17]. Our purpose here is narrower than adjudicating whether Sutton's four examples support his conclusion, which we take them to do; it is to identify what licenses the further step from those examples to a general argument against structural proposals, and to show that the licence usually assumed is not available. We do not offer an alternative to Sutton's historical thesis; we challenge a methodological conclusion often attached to it, namely that the demonstrated value of scalable search creates a presumption against structural innovation at the architectural level.

We proceed as follows. In Section 2 we separate the claim the essay demonstrates from the stronger claim it is used to support. In Section 3 we use the no-free-lunch theorems to show that the stronger claim is not merely unproven but false as stated, and we relocate the distinction the historical examples actually support. In Section 4 we develop that distinction through the transition from Ptolemaic to Newtonian astronomy, which we take to be the cleanest available case study for it. In Section 5 we give the distinction a precise vocabulary in terms of description length, and we review evidence that it is directly measurable inside trained neural networks. In Section 6 we distinguish several distinct ways geometric structure can come to be present in a method, since these are frequently conflated in this debate. In Section 7 we draw out the consequence for current debates concerning alternatives to the dominant scaling recipe, before concluding in Section 8.

2. The claim and its extension

The "*Bitter Lesson*" supports a specific and, we think, correct claim: within the domains Sutton names, methods whose quality scales with available computation have outperformed methods whose quality was bounded by how much structure a single researcher could identify and encode in a lifetime. This is a claim about the comparative economics of two processes of discovery. It says nothing, in itself, about the objects discovered by either process.

The claim actually invoked in contemporary argument is considerably broader: that any proposal carrying an explicit structural or architectural commitment, as opposed to a further scaling of the current recipe, repeats the error the essay documents and may be dismissed on that basis. This extension does not survive examination of the systems it is invoked on behalf of. The convolutional network that displaced hand-built computer-vision pipelines encodes, by construction, the assumption that natural images are translationally invariant [5]. The transformer architecture that displaced earlier sequence models encodes its own explicit commitments: a positional encoding, an attention mechanism that presumes a useful notion of pairwise relevance among tokens, a tokenizer that presumes a useful discrete unit of language. Bronstein et al. make this point systematically, arguing that essentially every successful modern architecture can be read as a choice of which symmetries of a domain to preserve by construction [4]. Rieck adds a further complication: the convolutional architecture is not a product of the scaling era. LeCun et al. described it in 1989 [13], more than two decades before the hardware existed to make it competitive [17]. By the standard now applied in retrospect, this would have made the architecture insufficiently general at the time; in fact it was simply ahead of the hardware available to it, which suggests that scalability, assessed in the present, is not the stable property of an idea it is often taken to be.

None of this contradicts the narrow claim documented in the preceding paragraph. It shows only that the narrow claim does not entail the broad one: "general methods outperforming a hand-tailored encoding of a single problem's specifics" is compatible with "every winning method, general or not, encoding some structural commitment of its own". The distinction the slogan fails to mark, and which we develop in the remainder of this paper, is between a commitment that is narrow and independently re-derived for each new instance and one that is broad and shared across instances.

3. No-free-lunch and the myth of the context-free method

In the discourse surrounding "general-purpose methods", the word general has two distinct senses. In the first sense, a method is general if it is not tailored to a single problem instance: gradient descent on a suitably parameterised function approximator is general in this sense relative to a hand-built chess evaluation function, since the same procedure applies to Go, to protein structure prediction, or to language, with no re-derivation required. In the second sense, a method is general if its performance does not depend on any assumption about the structure of the problems it will meet, that is, if it would perform as well on an arbitrary, unstructured task as on the tasks we actually care about.

The second sense is not merely unattained; it is unattainable. Wolpert and Macready proved that, averaged uniformly over the space of all possible objective functions, every optimisation algorithm, including one that samples points uniformly at random, achieves identical expected performance [28]. Wolpert had earlier established the corresponding result for supervised learning, namely that no learning algorithm can be shown to outperform another when both are evaluated over all logically possible data-generating distributions without begging the question [27], and Schaffer independently proved a restricted version of the same

fact [19]. The result is not confined to abstract search. Shalev-Shwartz and Ben-David give the corresponding statement for empirical risk minimisation: for any fixed hypothesis class there exist data-generating distributions on which that class fails badly while another succeeds, and no data-only algorithm can be justified as adequate across all distributions without an a priori restriction on which distributions are admissible [20]. Sterkenburg and Grünwald give a careful account of how this sits alongside the practical success of Occam's-razor-style learning guarantees, concluding that every workable guarantee is bought by restricting the space of situations under consideration rather than by escaping the theorem [23].

It is sometimes observed that the force of these theorems disappears once the uniform prior over all possible problems is replaced by the actual, structured subset of problems that physical, biological, and cultural processes produce. This observation is correct, and it is the point we wish to draw out. No real task is drawn uniformly from the space of all conceivable tasks. Gradient descent on a differentiable, compositional function approximator is not general in the no-free-lunch sense: it would perform at chance on a uniformly random reward landscape or a uniformly random string-completion task, exactly as the theorem requires. What allows it to perform well on chess, on images, or on language is a specific wager, namely that the loss surface is smooth, that the target function is compositional, and that natural data exhibit statistical regularity, and this wager happens to hold broadly across problems that humans care about. A hand-built chess evaluation function makes a narrower wager, that a particular feature matters for a particular game. Both sides of Sutton's dichotomy, on this view, place a wager; what distinguishes the historically successful side is not the absence of a wager but its breadth, namely a structural commitment, discovered once, that turns out to be shared across an enormous number of instances, set against a commitment that must be independently re-derived, by a single bounded researcher, for every new instance it meets.

A defender of scalable methods might resist even this weaker formulation. It might be said that nobody serious believes a successful method is literally assumption-free. The claim is only that a structural commitment should be treated as presumptively narrowing a method's applicability relative to a less explicitly structured alternative. Moreover this presumption, not context-free performance, is what has historically favoured compute-scaled search. This is a more defensible position than the one the no-free-lunch theorems dispatch, and it deserves a direct answer rather than being absorbed into the argument above by default. The answer is that the presumption, stated this way, is itself an empirical claim about which structural commitments tend to be broad, not a conclusion scale can settle in advance; investigating it directly, rather than assuming it, is precisely the task the remainder of this paper undertakes.

4. Cataloguing versus structure: epicycles, Kepler, and Noether

The clearest test of this distinction is not one of Sutton's own examples but an older one, the transition from Ptolemaic to Newtonian astronomy. It represents a harder case for the *general-beats-specific* reading than first appears, since both sides of the transition are, in a precise sense, curve-fitting exercises.

A Ptolemaic epicycle model represents a planet's angular position as a sum of rotating circles, $\sum_n c_n e^{in\theta}$, a truncated complex Fourier series in representations of the rotation group. Each additional epicycle is an additional harmonic term added to reduce residual error against observation, with no accompanying claim about why the planet moves as it does. This is not the absence of a data-driven method. Indeed it is data-driven curve-fitting of a remarkably literal kind, accurate enough to serve astronomers for over a millennium, and for part of that period more accurate than the early heliocentric models proposed against it, which retained circular orbits and required their own corrective epicycles to match observation.

Kepler's break with this tradition did not consist in refusing to fit data. He fit Tycho Brahe's positional record for Mars over several years, testing circles and ovals before arriving at the ellipse. The break consisted in terminating the fit in a low-parameter, invariant geometric object, an ellipse fully specified by an eccentricity and a semi-major axis, in place of an open series that must be extended indefinitely to improve its accuracy. Newton subsequently supplied the reason the ellipse is the correct shape: for an inverse-square central force, the rotational symmetry of the potential implies, by an argument later generalised by Noether's theorem, conservation of angular momentum, from which Kepler's second law, equal areas swept in equal times, follows as a consequence rather than as an independently fitted fact re-derived for each planet. A single symmetry and a single law account for every orbit, whether of a planet, a comet, or a binary star, with no free parameters beyond initial conditions and mass. The power of the theory was demonstrated when

perturbations in the observed orbit of Uranus were used to predict the position of an unobserved eighth planet, subsequently located close to the predicted position.

The distinction between the two theories is therefore not that one involves geometry and the other does not; epicycles are already a geometric object, namely rotations composed with rotations. The distinction is that the epicyclic model uses its group as an independently tuned basis for each planet separately, with no constraint transferring from one planet to the next, while the Newtonian model uses the same symmetry as a single generative law shared across every instance of the system. This is the sense in which a hand-built chess evaluation function differs from a trained value network. Both are attempts to approximate the same underlying object. The hand-built version is smaller because human designers, however expert, can only encode features they already have a concept for and can only test a small fraction of the ways those features might interact. A compute-scaled search process can explore both at a scale no team can match.

An objection suggests itself at this point. Newtonian celestial mechanics does not, in practice, generate every orbit from the two-body solution alone: once more than two bodies are involved, exact closed-form solutions are unavailable, and predicting the Moon's orbit or the motion of Jupiter's satellites to useful precision requires a perturbation series, an expansion in correction terms of increasing order. If an added correction term is what makes a theory enumerative, does Newtonian mechanics not become enumerative too, given sufficient precision? We think not, and we make the reason precise here, since it separates two practices that can look superficially alike. A perturbation series is generated by the same law that generates the unperturbed solution: each successive term is a derived consequence of applying Newton's equations to the small parameter distinguishing the true configuration from the idealised two-body case, and the series can in principle be extended to any order by a fixed procedure with no further appeal to observation. An epicycle, by contrast, is an independently chosen free parameter fitted to the residual error of the previous fit, with no procedure generating its frequency or amplitude from the law itself; the only input available for choosing it is the very data it is added to explain. In the vocabulary of Section 5 below, the distinction is between a correction whose description length is paid for once, in the law that generates the entire series, and a correction whose description length must be paid for again with each new term. This is why perturbation theory does not compromise Newtonian mechanics' claim to compactness in the relevant sense, while an open-ended stack of epicycles compromises Ptolemy's.

This particular historical transition also falls within the scope of an existing line of argument in the philosophy of science. Forster and Sober formalise the trade-off between fit and simplicity using Akaike's information criterion: a theory's expected predictive accuracy on data not yet observed is estimated by penalising its fit to the data already in hand by the number of adjustable parameters used to achieve that fit [8]. On this account, an ever-extended epicycle stack improves its fit to observed positions only at the cost of a penalty that erodes its expected accuracy on future observations, while a low-parameter law such as Kepler's or Newton's incurs a much smaller such penalty and is, in expectation, the more predictively accurate theory even before any future observation is made. The argument of this section is a restatement, in the vocabulary of dynamical symmetry, of the point Forster and Sober establish in the vocabulary of model selection; the following section restates it once more in the vocabulary of description length.

We state the chain of reasoning connecting this section to the last plainly before proceeding. Section 3 established that no learner is free of assumptions about the structure of the problems it will meet. The historical case developed here establishes a second, independent fact: some such assumptions, once found, generalise far beyond the instance that produced them, while others do not generalise past the instance they were fitted to. Together these two facts fix the productive question as which broad assumptions are available for a given class of problems, not whether an architecture carries any assumptions at all. Additionally these facts fix what the bitter lesson can and cannot settle in advance, namely the comparative speed at which a human or a compute-scaled search process arrives at such an assumption, and not which assumption a newly encountered domain will in fact reward.

The Ptolemy-to-Newton transition is not offered here as a claim that every successful learning system discovers a law of comparable cleanliness. It isolates, in an unusually uncluttered historical case, a distinction that is considerably harder to observe directly in Sutton's own examples, where a trained value network's or a policy's transferable structure, if it exists, is unlikely to be expressible as anything so tidy as an ellipse.

5. Compression as a common currency

A formal vocabulary for this distinction, one that does not depend on a case presenting itself as tidily as Kepler's, is available in the minimum description length principle. Rissanen's formulation asks a learner to minimise a two-part code, the length of the model together with the length of the data as encoded under that model, $L(\mathcal{H}) + L(D \mid \mathcal{H})$ [18]. A lookup table, whether a memorised training set or an epicycle stack extended with a fresh correction term for each anomaly, drives the second term toward zero at the cost of a first term that grows with the data itself, so that no compression has actually taken place. Kepler's ellipse, and Newton's law, keep the first term small at the cost of an honestly reported residual.

Compression, in this sense, is not identical to breadth. The rule that all primes greater than two are odd is extraordinarily compact and useful only within an extremely narrow domain, while a moderately complex description can still be broadly reusable. The claim of this paper is narrower than a general identification of the two. When a structural commitment does generalise across many instances, amortising its description length in the manner described above is the mechanism that allows it to do so cheaply; but whether any given compact description is in fact broad remains an empirical question that compactness alone does not settle.

Solomonoff and Kolmogorov generalised this idea from stochastic sources to individual objects: the Kolmogorov complexity $K(x)$ of a string is the length of the shortest program that outputs it on a fixed universal machine, and Solomonoff's theory of induction weights hypotheses by $2^{-K(h)}$ [12,21]. $K(x)$ is, however, provably uncomputable, by a diagonalisation argument in the style of the halting problem, so that any practical compressor, whether a trained model or a physicist proposing a law, can only exhibit an upper bound on true compressibility and never certify a lower one. The search for compact generative structure therefore has no guaranteed terminus; it produces successively better compressions rather than a final one. One consequence follows directly: no comparison of two live architectural proposals can be settled in advance of trial by a procedure that certifies which comes closer to a compact structure supporting accurate extrapolation on that domain. Indeed such a procedure would be a decision procedure for a quantity this section has just shown admits none.

This is a narrower claim than it may first appear, and we should be precise about its scope. It does not rule out the practical, approximate tools already available once some trial has occurred, such as a compression-based generalisation bound computed from an already-trained model, or a comparison of predictive performance under cross-validation; nor does it rule out extrapolation from scaling behaviour observed at smaller scale. What it rules out is a domain-independent oracle, applicable prior to any empirical exposure to the domain in question, that certifies which of two architectures is nearer to a compact structure supporting accurate extrapolation on that domain; the tools available in practice are themselves approximate, and are applied after some degree of trial rather than as a substitute for it.

Compressive and enumerative, in this vocabulary, name the two poles of a spectrum rather than a strict dichotomy into which any given method must fall. A lookup table sits at one pole, with $L(\mathcal{H})$ doing essentially all the work, since the model just is a copy of the data, and $L(D \mid \mathcal{H})$ negligible; Kepler's ellipse sits at the other, with $L(\mathcal{H})$ reduced to a handful of parameters and $L(D \mid \mathcal{H})$ carrying a genuine, non-trivial residual. Most successful systems, including the neural networks discussed below, sit somewhere between the two, encoding a mixture of transferable regularity and instance-specific detail within the same parameters, and a large language model in particular should be expected to store both at once rather than to fall cleanly on one side. What varies between methods, and over the course of training within a single method, is the ratio of the two components rather than the presence or absence of either. The comparative question this paper is concerned with is accordingly not whether a method contains any enumerative content, since essentially all real methods do, but which component a method's additional capacity is spent on as scale increases.

Carlini et al. found that as GPT-style language models trained by next-token prediction are scaled up, verbatim memorisation of their training data grows log-linearly with parameter count, with no change to architecture or objective required to produce the effect [6]: scale applied to a fixed choice of objective does not by itself favour the compressive pole, and whether it does depends on the objective and architecture doing the scaling rather than on the amount of computation supplied to them.

This distinction is not confined to the historical or philosophical register; it is measurable inside trained neural networks. Networks trained on small algorithmic tasks such as modular addition initially memorise

their training set, an interpolating solution comparable to an unbounded epicycle stack, before undergoing, well past the point of zero training error and under continued weight decay, a transition to a compact geometric representation in which the inputs are arranged on a circle and the outputs are computed via trigonometric identities. Power et al. first documented this transition and named it grokking [16]; Nanda et al. subsequently reverse-engineered the mechanism, exhibiting the circular embedding and the trigonometric computation directly in the trained weights [15]. This is the same transition in model-selection character, in miniature, that distinguishes an extended epicycle fit from a compact generative law, driven here by continued compression pressure under weight decay rather than by an explicit instruction to search for a symmetry.

Watanabe's singular learning theory provides an asymptotic framework relevant to such transitions. Because the map from a neural network's parameters to the function it computes is highly singular, in that many parameter settings compute the same function, the classical model-selection quantities developed for regular statistical models, such as a raw parameter count, do not describe a network's effective complexity. Watanabe shows instead that the real log canonical threshold of the singularity governing this map controls the asymptotics of the Bayesian free energy, providing the correct complexity term for this class of model [26], and the resulting theory has since been used to track phase transitions of the kind grokking exhibits. Separately, Arora et al. showed that trained deep networks, despite carrying far more parameters than training examples, can typically be compressed by low-rank factorisation and quantisation to a substantially smaller effective description without loss of performance, and that it is this compressed size, rather than the raw parameter count, that predicts generalisation through a PAC-Bayes bound [2]. The double descent phenomenon documented by Belkin et al. complicates this picture in a useful way: test error, which classical bias-variance analysis predicts should increase once a model is large enough to interpolate its training data exactly, in fact decreases again past that threshold in modern over-parameterised networks [3]. This indicates that gradient descent does not invariably select the memorising solution among the many that fit the data equally well, though why it tends to prefer a compressible one remains an open question rather than a settled one.

This paper uses minimum description length, Kolmogorov complexity, the PAC-Bayes compression bound, and the real log canonical threshold of singular learning theory as a shared conceptual vocabulary for compression, not as a claim that these formally distinct quantities coincide. Each captures a related but different notion of complexity, developed under different assumptions and for different purposes; the argument here rests on the qualitative pattern they jointly illustrate rather than on any identification between them.

Two different senses of compression have been at work in this section, and they are worth separating before drawing further conclusions. The first is a formal, information-theoretic property: a description is compressive if it is short relative to the data it accounts for, in the sense of the two-part code above, whether or not that description takes a form a human would recognise. Arora et al.'s compression bounds and the double descent phenomenon are evidence of compression in this first sense, and this sense remains representation-dependent in exactly the way already conceded above: a network can be short relative to its training corpus while remaining long, or simply opaque, under any other coding scheme one might choose to apply to it. The second, stronger sense is the one Kepler's ellipse and grokking's circular embedding both satisfy: not merely a short description under some encoding, but a low-dimensional, recognisable invariant of the kind discussed in Section 6 below. Grokking is compelling evidence for this paper's argument precisely because it satisfies the stronger sense; nothing shown in this section licenses assuming that formal compression inside a large trained network generally satisfies it as well.

This is not a coincidence of the particular task chosen. Modular addition is, in the sense made precise in Section 6 below, already a geometric object, namely the finite cyclic group under addition, and the network's circular embedding is a discovery of exactly the structure the task already possesses. Nothing in this guarantees that an equally compact structure exists for next-token prediction over natural text, and the memorisation result discussed above cannot, on its own, be read as evidence either way on that question.

A trained network's function is, moreover, jointly determined by three factors: the architecture's hypothesis space, the training procedure that searches it, and the data it is given. A compact description becomes reachable only when a genuinely compact structure exists in the interaction of all three, not in the data alone; an architecture with the wrong inductive bias, or an optimiser pursuing the wrong objective, can fail to expose a structure that is nonetheless present, and increasing scale under a fixed architecture and

objective addresses only one of the three factors that must be aligned. Carlini et al.'s finding is accordingly compatible with at least two very different explanations, between which it does not adjudicate: that some substantial fraction of a large text corpus is genuinely idiosyncratic, in the way that the initial conditions of a Newtonian orbit are idiosyncratic even though the orbit itself is law-governed, and no architecture could compress it further without loss; or that the particular combination of architecture, tokenisation, and objective in current use is simply not the combination best suited to exposing whatever compact structure natural language does possess. Nothing in the framework developed in this paper predicts which of these is correct, any more than it predicts in advance which architectural wager is the broad one; that, as Section 7 argues, is a question only experimentation can settle.

6. Where geometric structure comes from

The two threads developed above meet at a single point. Section 4 identified the invariant that survives the passage from Ptolemy to Newton as a shared geometric object, a symmetry acting the same way across every instance of the system rather than a catalogue re-fit for each one; Section 5 identified the same passage, inside a trained network, as a reduction in description length that persists on data not yet seen. We take these to be two descriptions of a single phenomenon: what a successful compression buys, beyond a short code for the data already in hand, is very often a geometric structure of exactly this kind.

We use the term geometric throughout this section, as we have implicitly used it above, in the sense established by Klein's Erlangen program, namely as whatever remains invariant under a specified group of transformations, rather than in the narrower sense of a shape embedded in physical space [11]. This is a substantive rather than a merely convenient usage: it is precisely what unifies the rotational symmetry underlying Kepler's ellipse, the translational invariance built into a convolution, and the invariance of the Fisher metric under sufficient statistics discussed below, as instances of a single notion rather than as a family of loose analogies. We do not claim, however, that every compact structure a learning method might discover is geometric in this sense. Algebraic, logical, or combinatorial invariants that resist description as invariants of a transformation group would fall outside the scope of the claim made here, and nothing in what follows depends on ruling out that possibility; the claim is only that, in each case examined below, the structure a method is shown to have discovered has in fact taken this form.

There are, however, four distinct ways in which a method can come to possess such a structure. The first is geometry imposed on a representation in advance: a convolution encodes translational invariance, message passing in a graph neural network encodes a given adjacency structure, an equivariant layer is constructed to respect a chosen symmetry group. The geometric deep learning programme of Bronstein et al. is explicitly of this kind [4], and it is, by the logic of Section 3, exactly as exposed as any other hand-supplied prior to being overtaken by a less constrained model given sufficient data and computation; whether sufficiently large unconstrained architectures learn the relevant invariances from data without such constraints remains an active question. Vapnik's support vector machine belongs to the same category, in a more classical register: margin maximisation is by construction a distance computation, and the kernel trick deliberately builds an embedding in which a geometric fact, linear separability, is made available to the algorithm [25]. Here too the geometry is supplied by the designer rather than found by the method.

The second use is different in kind: geometry discovered by a learning process, emerging in what a generically trained system's weights come to represent, in a manner agnostic to the representation used to find it. Hewitt and Manning's structural probe is a clear instance. They show that syntactic parse-tree distances between words in a sentence are recoverable, to a high degree of accuracy, as approximately squared Euclidean distances under a single learned linear transformation of a contextual language model's embeddings [10]. No parse tree was built into the architecture, and nothing in the training objective called for one; self-supervised next-word prediction left syntactic structure recoverable from the resulting representations in geometric form, without the network being directed to produce anything of the kind. This is what makes it evidence for the claim at stake here: that the object a learning process converges on, when it succeeds, is frequently a compact geometric invariant rather than an unbounded catalogue of cases.

A third and distinct case is geometry fixed at the level of the learning objective itself, prior to any weights being trained. Diffusion and flow-matching generative models are constructed as a transport, a continuous path or vector field, connecting a simple reference object, an isotropic Gaussian, to the data manifold, formalised as a reverse-time stochastic differential equation by Song et al. [22] and as a direct construction

of the connecting vector field by Lipman et al. [14]. Here the geometric object, the transport between two distributions, is guaranteed by the choice of objective before training begins; what training discovers is only which particular vector field implements it, and the network parameterising that field is incidental, whether convolutional, attention-based, or a plain multilayer perceptron. This is not spontaneous discovery in the sense of the structural-probe example above, but neither is it geometry imposed on the architecture in the sense of convolutional or equivariant design choices: the constraint lives in the objective rather than in the network's connectivity, and the system remains free to discover any function that satisfies it.

A further case is geometric structure that is not merely discovered by a training process, but is, in a stronger sense, a mathematical necessity. Chentsov's characterisation theorem is the clearest example: it shows that the Fisher information metric is, up to scale, essentially determined by the requirement that a metric on probability distributions be invariant under sufficient statistics [1, 7]. Statistical inference is not merely analogous to geometry on this account; it is geometry, as a consequence of an invariance requirement rather than a choice of description, and this is as true of a Bayesian calculating by hand as it is of a trained network. What distinguishes this case from the three preceding ones is that it does not depend on an architect's choice, an empirical accident of training, or a choice of objective: the structure is forced on any correct solution to the inference problem, not stipulated as part of the problem's definition.

7. Consequences for current architectural debate

This bears directly on how the essay is presently used. A joint-embedding predictive architecture, which asks a learned system to predict in a latent space rather than at the level of raw pixels or tokens, is not hand-crafted domain trivia in the sense the essay disparages. It is a rival, equally general-purpose wager about which broad, shared structural commitment a compute-scaled learner should be directed toward, argued at the same level of generality as next-token prediction rather than as a retreat from it. The same holds for world models, for architectures that separate a learned dynamics model from a policy, and for neuro-symbolic proposals that route computation through an explicit, reusable structure rather than encode it implicitly in undifferentiated weights. Invoking the bitter lesson to foreclose on any of these proposals in advance of the relevant empirical evidence is not an application of Sutton's historical observation; it lends the essay's accumulated authority to one contestable wager, namely that token-level autoregression over the present architecture is the correct shared structure, as though that wager had already been settled by the mere existence of scale.

This is not an argument for indifference between architectures, still less for a return to encoding domain intuitions by hand in place of experiment. The historical pattern Sutton documents is real, and the underlying economics, namely that a bounded researcher will generally be outpaced by a compute-scaled search process feeling its way toward a better analogy, is likely to continue holding wherever it has held before. What does not follow is that every question concerning which analogy a learning process should be directed toward, or which shared structure a new architecture wagers on, has already been closed by the existence of scale.

It is sometimes objected that this account gives no way to say, in advance of trial, which of two rival architectural wagers carries the broader structural commitment, and that the criterion of Section 3 is therefore only ever applied in hindsight. The objection has real force, but it is not one this account can be faulted for failing to meet. Section 5 showed that the uncomputability of Kolmogorov complexity rules out a universal, domain-independent procedure that certifies, for an arbitrary domain and representation, which of two candidate descriptions comes closer to optimal. This does not imply that any particular architectural comparison is impossible; it implies only that no domain-independent criterion can settle such comparisons in general, prior to any empirical exposure to the domain in question. This does not deny that partial tools, such as scaling-law extrapolation from smaller trained models, cross-validation, or a compression-based generalisation bound computed after training, provide genuine comparative evidence once some trial has occurred; it denies only that such a tool can be run in advance of all trial, or extended into a universal, domain-independent oracle.

Short of a trial, there is one further partial criterion worth naming: whether a proposed architecture can be shown to recover a previously validated compact structure as a special or limiting case, in the manner of a correspondence principle. This is best treated as a desideratum rather than a strict requirement, and it is neither necessary nor sufficient. It is not necessary, since a genuinely new structure need not be derivable from what it succeeds, and such a derivation is often unavailable even for well-established physical theories: no

one has derived Euler's theory of beam buckling from the Schrödinger equation, and the electronic structure of anything beyond the smallest molecules can be recovered from quantum mechanics only approximately, if at all. It is not sufficient, since reproducing known cases does not establish that a model generalises beyond them.

What does not follow from any of this is that the presumption of continued success belongs to one side of this debate and not the other. Repeated success under trial is genuine evidence, in the ordinary sense that it legitimately raises confidence in a hypothesis; the point is only that this form of evidence is available in principle to every candidate wager and not solely to the one already in wide use. A method tried and found wanting on some class of problems is itself evidence, weighted accordingly, that a rival wager may fare better there, just as a method's repeated success is evidence in its own favour; a method not yet tried at the relevant scale contributes no evidence either way until it is. The conclusion drawn here is accordingly not a case for indifference between architectures, still less a licence to prefer an untested proposal over a validated one without reason, but a case for continued diversity of experiment: if evidence is what settles which structural wager is the broader one, only trial across a range of candidate architectures can supply it, resource and practical constraints permitting.

8. Conclusion

Three distinct claims have been in play throughout this paper, and they do not stand or fall together. The first is about how a particular text is currently used: Sutton's essay functions, in practice, as a mechanism for foreclosing architectural proposals, a pattern documented independently by Brooks and by Rieck [5, 17]. This pattern is what makes this paper timely, and it is also the least durable of the three claims, since it concerns a discourse, and discourses change.

The second is a claim of logic, independent of anyone's behaviour. The inference from general methods having historically outperformed hand-tailored ones to the conclusion that any structural commitment is illegitimate does not follow, and the no-free-lunch theorems show why, regardless of whether anyone has actually drawn that inference aloud.

The third is this paper's substantive contribution: a positive claim about a central feature distinguishing methods that generalise well from methods that do not, though not the only such feature. The relevant question is not whether a method carries structural assumptions, since every successful method does, but whether those assumptions compress, in the sense made precise in Sections 5 and 4, or merely enumerate.

This third claim does not depend on the first. It would remain true, and would remain the correct lesson to draw from the historical record, even in a world where Sutton's essay had never been written: the passage from Ptolemy to Newton, the uncomputability of Kolmogorov complexity, and the phase transition documented in grokking would stand as argued here regardless, and would still bear on which architectural wagers are worth pursuing. Sutton's essay, and its subsequent misreading, supply the occasion for setting this argument out, not its content.

We close by naming directly what this account takes learning itself to be, since the word analogy has been used throughout without definition: not a domain-free procedure and not a catalogue of facts, but the construction of a structure-preserving map between a domain not yet understood and one already partly understood, in the sense established in cognitive science by Gentner's structure-mapping theory [9], a map that is necessarily contentful on both sides and hence the opposite of context-free.

When properly restricted, the bitter lesson asserts something narrower than the slogan it has become: a bounded researcher, encoding one analogy by hand, will generally be outpaced by a compute-scaled search process feeling its way toward better ones, provided the domain in question possesses a compressible structure worth finding and provided the search is directed, by some prior human decision about objective and architecture, toward a region in which that structure is reachable. Each of these provisos is substantive and contestable, and each is precisely where current disagreement about architecture actually resides. The bitter lesson, on the reading offered here, is a lesson about which agent finds the analogy first. It has comparatively little to say about which kind of analogy is there to be found, and it lends no support to the practice of using the essay to close that second question in advance.

Acknowledgements

This essay grew out of an extended dialogue with Claude (Anthropic) examining what Sutton's essay claims and what it has come to be used for in current debate over model architecture. It is offered as a companion to Rieck's *The Better Lesson?* [17], approached from a different direction.

References

1. Shun-ichi Amari and Hiroshi Nagaoka, *Methods of information geometry*, Translations of Mathematical Monographs, vol. 191, American Mathematical Society, Providence, RI, 2000.
2. Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang, *Stronger generalization bounds for deep nets via a compression approach*, Proceedings of the 35th International Conference on Machine Learning (ICML), 2018, pp. 254–263.
3. Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal, *Reconciling modern machine-learning practice and the classical bias–variance trade-off*, Proceedings of the National Academy of Sciences **116** (2019), no. 32, 15849–15854.
4. Michael M. Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst, *Geometric deep learning: Going beyond Euclidean data*, IEEE Signal Processing Magazine **34** (2017), no. 4, 18–42.
5. Rodney Brooks, *A better lesson*, Blog post, March 2019.
6. Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr, and Chiyuan Zhang, *Quantifying memorization across neural language models*, Proceedings of the 11th International Conference on Learning Representations (ICLR), 2023.
7. Nikolai N. Chentsov, *Statistical decision rules and optimal inference*, Translations of Mathematical Monographs, vol. 53, American Mathematical Society, Providence, RI, 1982.
8. Malcolm Forster and Elliott Sober, *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** (1994), no. 1, 1–35.
9. Dedre Gentner, *Structure-mapping: A theoretical framework for analogy*, Cognitive Science **7** (1983), no. 2, 155–170.
10. John Hewitt and Christopher D. Manning, *A structural probe for finding syntax in word representations*, Proceedings of NAACL-HLT 2019, 2019, pp. 4129–4138.
11. Felix Klein, *Vergleichende betrachtungen über neuere geometrische forschungen*, Mathematische Annalen **43** (1893), 63–100.
12. Andrei N. Kolmogorov, *Three approaches to the quantitative definition of information*, Problems of Information Transmission **1** (1965), no. 1, 1–7.
13. Yann LeCun, Bernhard Boser, John S. Denker, Donnie Henderson, Richard E. Howard, Wayne Hubbard, and Lawrence D. Jackel, *Backpropagation applied to handwritten zip code recognition*, Neural Computation **1** (1989), no. 4, 541–551.
14. Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le, *Flow matching for generative modeling*, Proceedings of the 11th International Conference on Learning Representations (ICLR), 2023.
15. Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt, *Progress measures for grokking via mechanistic interpretability*, Proceedings of the 11th International Conference on Learning Representations (ICLR), 2023.
16. Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Ilya Sutskever, *Grokking: Generalization beyond overfitting on small algorithmic datasets*, arXiv:2201.02177, 2022.
17. Bastian Rieck, *The better lesson? Geometry and topology in the era of deep learning*, Blog post, February 2026.
18. Jorma Rissanen, *Modeling by shortest data description*, Automatica **14** (1978), no. 5, 465–471.
19. Cullen Schaffer, *A conservation law for generalization performance*, Proceedings of the 11th International Conference on Machine Learning (ICML), 1994, pp. 259–265.
20. Shai Shalev-Shwartz and Shai Ben-David, *Understanding machine learning: From theory to algorithms*, Cambridge University Press, Cambridge, 2014.
21. Ray J. Solomonoff, *A formal theory of inductive inference, Part I and Part II*, Information and Control **7** (1964), no. 1–2, 1–22, 224–254.
22. Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole, *Score-based generative modeling through stochastic differential equations*, Proceedings of the 9th International Conference on Learning Representations (ICLR), 2021.
23. Tom F. Sterkenburg and Peter D. Grünwald, *The no-free-lunch theorems of supervised learning*, Synthese **199** (2021), 9979–10015.
24. Richard S. Sutton, *The bitter lesson*, Blog post, March 2019.
25. Vladimir N. Vapnik, *Statistical learning theory*, Wiley, New York, 1998.
26. Sumio Watanabe, *Algebraic geometry and statistical learning theory*, Cambridge Monographs on Applied and Computational Mathematics, vol. 25, Cambridge University Press, Cambridge, 2009.
27. David H. Wolpert, *The lack of a priori distinctions between learning algorithms*, Neural Computation **8** (1996), no. 7, 1341–1390.
28. David H. Wolpert and William G. Macready, *No free lunch theorems for optimization*, IEEE Transactions on Evolutionary Computation **1** (1997), no. 1, 67–82.

ICTS-TIFR Bangalore
Email address: vishal.vasan@icts.res.in