Beauty in Machine Learning: Fluency and Leaps

Benjamin Bloem-Reddy*
Department of Statistics
University of British Columbia
benbr@stat.ubc.ca

Abstract

The extrapolative leaps from training to deployment are precisely where ML's best methods for system development succeed, and where most—and the most consequential—failures occur. I argue that the leap is where beauty can and does play a role. Beauty also serves to highlight some important aspects of the leap that are not specific to beauty. Based on this, and drawing on research on beauty from psychology, cognitive science, and philosophy of science, I articulate some fundamental problems and suggest directions for potential solutions.

Should beauty play a role in machine learning (ML) research? If so, how? In many scientific fields, particularly physics, beauty is valued. The mathematician and physicist Hermann Weyl famously said, "My work always tried to unite the true with the beautiful; but when I had to choose one or the other, I usually chose the beautiful" [1]. The physicist Steven Weinberg wrote, "The physicist's sense of beauty is also supposed to serve a purpose—it is supposed to help the physicist select ideas that help us to explain nature" [2]. These ideas are intuitively appealing and have helped in the development of successful theories, but putting them into practice is challenging. Beauty has been used to describe different things by different scientists. There is also the risk of conflating it with truth and using it as part of a non-empirical "logic of justification," or reason to believe a theory without empirical evidence [3, 4]. For example, the physicist Sabine Hossenfelder has argued at length that an over-emphasis on beauty (particularly the elegance of string theory) was largely responsible for a stagnation in the foundations of physics, making much of it "post-empirical" [5, 6]. So beauty might be a good guide to individual scientists for developing novel theories, but a theory's beauty should not be its primary justification, particularly for the continued efforts of large groups of researchers. Ultimately, it must be reconciled with empiricism. Is the situation the same in ML?

For concreteness, I take the following textbook example of supervised learning with empirical risk minimized by gradient descent as my template for the development of an ML system, though it is straightforward to adapt to less conventional examples. To start, an ML researcher has a training set of labeled examples, say images of animals. The researcher also has an objective in mind, say to obtain a function that predicts the label of the animal in an image with high accuracy, on more images of animals. In our template, an ML researcher specifies a hypothesis class of functions or distributions, $\mathcal{H}$, and a loss function quantifying the accuracy of a prediction function, possibly with a regularization term to encourage a solution with certain properties. A single hypothesis h^* is selected as follows. Starting with some $h_0 \in \mathcal{H}$, the compatibility of h_0 with the current data is assessed by the empirical risk, i.e., the average loss on the training data. A response in the form of a gradient-based update yields h_1, and the process is iterated until the search is terminated with some h^*. There may be an additional process for selecting h^*, such as cross-validation. The selected hypothesis h^* is now treated as a *theory*, and deployed as an explanation for the data generating process or, increasingly, as part of a technology that relies on the predictions of h^* on new inputs.

*This really is a draft intended to be workshopped. It has evolved substantially (including a title change) since the first draft was submitted to the workshop. All comments and feedback welcome.

I (Still) Can't Believe It's Not Better Workshop at NeurIPS 2021.

There is an important difference between this process and the process of developing a theory in most sciences. Predictions of a scientific theory are tested, and empirical evidence allows scientists to reason *inductively* [7]; as evidence accumulates the predictions of a theory *are likely* to hold (or not hold), based on the evidence alone. In contrast, an ML system is typically developed using one dataset, subject to a small battery of tests or selection criteria, and then deployed—potentially in settings in which the system's predictions have direct and major consequences for one or more humans. In the act of deployment, there is an implicit, untested hypothesis that the system will perform similarly on data encountered in deployment, and fulfill objectives in deployment that may not have been considered during development. Here, except in special cases, a different mode of reasoning is involved. Reasoning that the predictions of a theory *may be true*, before enough evidence has accumulated to deem them probable or improbable, is *abductive reasoning* [8, 9]. In many sciences, such as physics, abductive reasoning is often used as a "logic of pursuit," i.e., using various non-empirical arguments for pursuing the development of a theory before it has been sufficiently tested [4]. Ultimately, the theory must be tested before it is deemed likely to hold, and usually it can be tested experimentally without major, direct impact on the lives of humans who are not involved in the research.[1] In ML, however, similar testing may never be performed and if it is, testing occurs after the deployed system has already had consequences.

The extrapolative leaps from training to deployment are precisely where ML's best methods for system development succeed, and where most—and the most consequential—failures occur. As I argue below, the leap is where beauty can and does play a role. Beauty also serves to highlight some important aspects of the leap that are not specific to beauty. Based on this, and drawing on research on beauty from psychology, cognitive science, and philosophy of science, I articulate some fundamental problems and suggest directions for potential solutions.

Theories of beauty Before making any attempts to address the role of beauty in ML research, we need a working definition of beauty, and how it might possibly manifest in ML. Philosophers, artists, and scientists have debated the meaning of beauty for millennia—the Western historical record of such debates dates at least 2,500 years [10]. Historically, much of the debate focused on the distinction between objective beauty (it is a feature of an object, independent of any perceiver) and subjective beauty (completely "in the eye of the beholder" [10]), eventually leading to Kant, Hume, and others to postulate an intersubjective theory (having a social aspect with which we might reason). Modern theories have combined elements of those two traditions with empirical research from psychology and neuroscience to understand beauty as interactionist: "emerg[ing] from patterns in the way people and objects interact" [11].

The interactionist perspective is adopted here. Within ML research, there are two main categories of interaction to which beauty might be relevant. The first category consists of the interactions between an ML system and data. In these types of interactions, the system processing data has much in common with a human brain processing sensory input. Viewed this way, research on beauty from cognitive science is informative. I explore this in detail in Section 1. The second category involves ML researchers interacting with ML systems in the course of research. For these types of interactions, philosophical investigations of the role of beauty in scientific research are relevant. This is the subject of Section 2.

1 Processing fluency in ML systems

From an interactionist perspective, the role of beauty in ML systems, if there is one, is determined by the interactions between those systems and objects represented by the data they process. A particularly potent cognitive theory of beauty is the theory of *processing fluency*, summarized by the psychologists Reber, Schwarz, and Winkielman [11] as follows: "[A]esthetic experience is a function of the perceiver's processing dynamics: The more fluently the perceiver can process an object, the more positive is his or her aesthetic response" [11, p. 365]. Crucially, specific assumptions of processing fluency theories [12] adapted to ML systems entail three aspects:

(A1) a *compatibility* between the system and certain features of data that *might* be processed;

[1] An important exception to this is, of course, an experiment on human subjects, though ethical standards and protocols are in place to mitigate negative impacts. In general, there also may be indirect impact in the form of taxes funding research, etc.

(A2) an internal signal indicating the compatibility of the system with the current data being processed; and

(A3) a response that is more positive with higher compatibility.

Our textbook example of supervised learning attempts to find an ML system that satisfies (A1) by iteratively assessing compatibility (A2) using empirical risk and responding (A3) with a gradient-based update. The standard assessment of (A1)-style compatibility is out-of-training-sample generalization error, under the assumption that all data that might be processed will be drawn from the same distribution as the training data. This has long been and remains the dominant framework in ML. Adopting processing fluency as our framework for beauty, we find that certain aspects of processing fluency already play significant roles in ML systems, particularly those to which the template is well suited. We might also understand some of the shortcomings of current ML systems as failures to be properly compatible with the data they might process (A1) or with the task on which they are deployed, and the lack of a reliable, integrated, real-time "compatibility response" (A2-A3).

1.1 "Objective" beauty in ML: quantifiable usefulness

Some of ML's best methods rely on notions of beauty that are conceptually similar to aspects of *perceptual fluency*, or "the ease of identifying the physical identity of the stimulus," [11, p. 366] which humans rely on to perceive physical stimuli. On these types of problems, ML has made, and continues to make, exceptional progress. Reber, Schwarz, and Winkielman [11] listed three features that, based on a large body of empirical evidence, might be considered "objectively" beautiful features: simplicity/amount of information, symmetry, and contrast/clarity. Moreover, those features contribute primarily to perceptual fluency. For ML problems that correspond to this type of fluency, the purposeful beauty of these features can be found throughout ML; we review some of them here.

Simplicity: generalization and theory selection As summarized in [11], experimental research in psychology indicates that visual stimuli with less information (as measured by high redundancy and/or symmetry; see below) are identified faster and are found more pleasing, consistent with the theory that simpler objects are processed more fluently by the human brain. In textbook ML system development, out-of-training-sample generalization error has a notion of beauty-in-simplicity built in. Statistical learning theory provides bounds on generalization error (e.g., PAC-Bayes [13, 14] or bounds based on the VC dimension or Rademacher complexity [15]) with the general form:

$$\text{Generalization error of } h^* \leq \text{Empirical error of } h^* + \text{Complexity of } h^* \,. \tag{1}$$

The practical guidance provided by this bound is clear: in order to select the most compatible theory from those that achieve the same empirical error, we should select the simplest. There is a fluency in (1): one can obtain general compatibility between system and possible data (A1) by obtaining specific compatibility between system and observed data (A2) and imposing the requirement that the system be no more complex than necessary. This is the conceptual basis of Occam's Razor, information criteria for statistical model selection (e.g., AIC and BIC) [16], the minimum description-length principle [17], and the maximum entropy principle [18], among others.

Symmetry: decomposition and simplification Symmetry is widely held to be a beautiful feature [11]. From a perceptual fluency perspective, an object with high degree of symmetry is easier to process; in fact, it has been used in psychology experiments as a proxy for information content [11]. Symmetry also has a long history in statistics and ML as a reliable way to obtain appropriately simple and accurate models. For example, we might specify a model family $\mathcal{H}_{\text{symm}} \subset \mathcal{H}$ that only contains symmetric models (e.g., CNNs or families of invariant distributions). Alternatively, we might ask for an algorithm that yields a model from the full $\mathcal{H}$ that is "almost" in $\mathcal{H}_{\text{symm}}$, or approximately invariant (e.g., data augmentation). If the symmetry assumption is accurate—that is, if the underlying process really does obey those symmetries—then something useful happens: compared to a non-symmetric method, a symmetric method is simpler (lower complexity) and the empirical risk is at least as good. This fluency can be quantified in a number of ways: statistical benefits are quantified by generalization bounds specializing (1) [19, 20, 21]; in information theoretic terms, invariance is compressible [22]; and CNNs have a smaller number of parameters than their fully connected counterparts [23, 24].

Contrast and clarity: signal-to-noise In the same way that high-contrast images are processed more fluently [11], data with high signal-to-noise ratio is learned from more efficiently, and results in lower generalization error.

1.2 Objectifying beauty

Perhaps a better name for these aspects of "objective" beauty is "objective function" beauty. In each case, the beautiful feature can be described well in information theoretic terms, leading to a quantifiable statistical and computational benefit. More generally, to the extent that an aesthetic feature can be quantified, it can be incorporated into ML methods as a term in an objective function. Indeed, regularization towards lower complexity has been standard practice for some time [e.g., 25, 26], and regularization towards invariance through data augmentation has become standard practice in deep learning in recent years. So—is the inclusion of beauty necessary for a good ML method? The statistical and computational usefulness of "objective function" beauty suggest an answer in the affirmative, not as an end unto itself, but as an indicator of a system well suited to its environment.

But beauty is not sufficient, and observed compatibility can be misleading. The compatibility assessment in (1) implies (A1)—compatibility between system and data that *will* be processed—only under a certain condition: new data is *drawn from the same distribution as the training data*. Relatively newer[2] lines of ML research pertaining to robustness, causality, fairness, accountability, and transparency have highlighted the risks posed by an over-reliance on that particular version of compatibility which, in many settings, turns out to be restrictively myopic. As an example, consider the well-known problem of distribution shift: after an ML system is optimized for performance with (1) on data from distribution P, it is deployed and encounters data from distribution $P' \neq P$. The generalization guarantee of (1) is no longer a guarantee; it can be arbitrarily, catastrophically bad. Without anticipating how the distribution might change in deployment, no notion of objective function beauty can salvage the situation. Within the context of processing fluency, the system is not compatible with the data it processes. Particularly problematic for ML system deployment is the lack of reliable real-time signal of and response to the lack of compatibility. Although many types of incompatibility arise from a distribution shift, others can arise from various types of bias [27], or incompatibility between a system and its intended use (e.g., a lack of interpretability [28, 29]).

1.3 Towards increased processing fluency in ML systems

How can we address the risks presented by incompatibility between an ML system and the data it processes? One general purpose approach is to imitate human processing fluency with a separate, meta-cognitive system for assessing compatibility. Crucially, the loss function should be only one part of such a system: If the only signal for compatibility in deployment is high prediction error, then by the time that signal is given the system has erred, potentially with unacceptable consequences.

Computational models of meta-cognitive processes already exist. For example, recent work in [30] models the so-called "Aha! moment" using methods similar to those for predicting algorithm runtime [31, 32]. Alternatively, uncertainty quantification has played the role of incompatibility signal in statistics and other fields, but a reliable, general purpose, efficiently computable method has remained elusive for most ML systems, particularly those using deep neural networks. Ultimately, the design of a reliable compatibility assessment system should fulfill various criteria—but which criteria? Again, perceptual fluency can provide a framework.

Experiential and meta-perceptive beauty Clearly, human experiences of beauty are more complicated than the types of beauty we can currently encode in an objective function. Research indicates that numerous other temporal and learning factors including repeated exposure, prototypicality of a stimulus, implicit learning of structure, and the development of expertise all play important roles, as features influencing processing fluency [11]. For lack of a better term, I will categorize these as "experiential." Moreover, high-level features and meta-cognitive processes appear to moderate our perception of beauty. These include assessing how processing fluency deviates from expectations, the attribution of processing fluency to relevant sources, and internal competition between perceptual and conceptual fluency [11]. I will categorize these as "meta-perceptive" features.

Whereas the usefulness of objective function beauty can be formalized by statistical and computational quantities, experiential and meta-perceptive beauty appear more complicated; to my knowledge, no broadly applicable formalism exists. If one does, it does not appear to be widespread in ML research.[3]

[2] Robustness, causality, interpretability, etc., have long histories in statistics, related applied fields, and small sub-areas of ML; they are somewhat newer to mainstream ML.

[3] The only work I am aware of that does anything like this combines reinforcement learning with processing fluency as a computational model for certain cognitive processes [33, 12].

However, recent research in structured representation learning aims to learn implicit structure; causal learning attempts to learn causal attribution; and continual learning naturally adapts the system to changes in data. Quantities derived from those lines of research could provide ingredients for analogues to the compatibility determined by experiential and meta-perceptive processes. Suitable variations of statistical learning theory and information theory, respectively, seem like sensible starting points, though much work remains.

2 Stronger reasoning in the leap

The success of methods based on objective function beauty is tied to our ability to reason about the leap from training to deployment in terms of quantifiable hypotheticals. For example, (1) makes a strong argument for a balance between minimal training error and simplicity when the question is, "How should we minimize expected loss if more data from the same distribution, P, is input into the trained system?" The bound in (1) allows us to pass from abductive reasoning [8] (suggestive that a theory *may* hold but has not been tested) to induction (a theory *does* hold with certain probability); from inferring that h^* may perform well on other data drawn from P, to concluding that it will perform well with quantifiably high probability.

The leap from abduction to induction depends on the relationship between the tests performed and methods used to select h^*, on the one hand, and the settings and purposes for which we imagine h^* will be applied, on the other. The relationship between a sample from P (on which h^* is trained) and other samples from P is very different from the relationship between a sample from P and a sample from a mixture of P and another distribution P', for example. Minimizing expected risk is very different from minimizing worst-case risk.

The general hypothetical in ML is something like, "How should we achieve criteria $\mathcal{Q}$ if data from a set $\mathcal{P}$ of distributions is input?" How can we reason about the likelihood of an ML system achieving $(\mathcal{P}, \mathcal{Q})$ when we cannot test it directly before deployment? Passage from abductive to inductive reasoning requires a specific $(\mathcal{P}, \mathcal{Q})$, its relationship to the training setting, and observable constraints on the properties of h^*, such as complexity. In practice, at least one of these ingredients is missing. Especially difficult for ML researchers developing a new method is the challenge of anticipating the widest possible range of $(\mathcal{P}, \mathcal{Q})$-settings in which their method might be used. This is precisely where beauty and other values can (and do) enter: in the absence of empirically testable properties and standards or known deployment conditions. In their absence, most research problems are underdetermined. Beauty can impose useful constraints, as can other values.

Pursuing beauty Of the possible values that might impose useful constraints, suppose that ML researchers as a whole value beauty. How, then, should it be pursued, by individual researchers and the field as a whole? How should ML researchers judge a system to have beautiful properties? The mathematician Gian-Carlo Rota wrote, "Mathematicians may say that a theorem is beautiful when they really mean to say that the theorem is enlightening," used as a device "to avoid facing up to the messy phenomenon of enlightenment," or understanding [34, pp. 181-2]. In ML, as in mathematics, we may use "beautiful" to describe an experiment or a theoretical result that "gives away the secret ... [that] leads us to perceive the actual" [34, p. 182]. In other words, there is a certain conceptual fluency present (and it's messy). Within ML, there have been calls for this type of enlightenment. A well-known recent example was the 2017 NeurIPS Test of Time Award acceptance speech by Rahimi and Recht [35, 36], which articulated a need for better understanding of the methods researchers develop and use. Better understanding of our methods enables stronger reasoning about them for a wider range of $(\mathcal{P}, \mathcal{Q})$. In this sense, beauty as something that gives enlightenment or encodes understanding can play a key part. But as Rota wrote, enlightenment is messy, at least partially subjective—how can we achieve it?

One possible bridge is *conceptual fluency*, described in [11, p. 366] as "the ease of mental operations concerned with stimulus meaning and its relation to semantic knowledge structures." For ML researchers, building such a bridge requires work from both end-points: the development of ML methods that are able to infer meaning from data and relate it to (human) semantic knowledge structures; and the development of the appropriate knowledge structures with which to understand ML methods. In many ways, researchers regularly cross such a bridge individually and informally. Using simple models that are widely interpretable and easy to reason about is one path, but it seems limiting. Alternatively, researchers can do the hard work of illuminating the inner workings of a new

method, of its strengths, limitations, and potential performance under a wide range of specific $(\mathcal{P}, \mathcal{Q})$. Broadly, this seems to require new empirical and theoretical tools, such as those being pursued by researchers working in interpretability and explainability [28, 29]. Work causality-inspired ML and structured representation learning is also directly relevant. Given the growing role of ML systems in our everyday lives, the development of such tools should be considered at least as important as the development of the next benchmark dataset leader or new large language model.

Crossing the bridge Perhaps conceptual fluency provides a bridge. However, even a sturdy bridge does not provide a single complete answer. Researchers may use the bridge differently. For example, the discussion of enlightenment ignores the fact that the degree of understanding obtained by one human may be completely different from the degree of understanding obtained by another. Semantic knowledge structures vary from human to human. As with aesthetic preferences, they depend on one's cultural environment, one's lived experience, one's education [11]. To illustrate, Rota [34, p. 173] gives an example of mathematical beauty in "the Galois theory of equations," which in all likelihood I would find incomprehensible (though perhaps with some study, I might come to appreciate the theory's beauty). In the context of ML, simplicity and interpretability don't have universally agreed-upon definitions [37]. There may not be a one-size-fits-all solution to the problem.

This is true more generally, and the diversity can be a good thing, especially when researchers carefully articulate their reasoning. Diversity here is broadly construed to encompass all of the factors that influence how a researcher arrives at the varied and complex decisions involved in the course of conducting research. These largely overlap with the factors that influence one's perception of beauty, whether inherited, learned, or experienced. As the philosopher of science T. S. Kuhn wrote, research decisions, or "application of values," are "sometimes considerably affected by the features of individual personality and personal biography that differentiate" researchers, and that "individual variability in the application of shared values may serve essential functions in science" [38]. These essential functions are "risk-spreading" (i.e., efficient exploration of the space of possible solutions) and the paradigm-shifting discoveries that often require high-risk research [39]. Combining Kuhn with perceptual fluency, in the pursuit of the shared value of beauty, ML researchers with diverse aesthetic preferences will, collectively, develop ML systems that are more compatible with their inputs and tasks.

Beauty is one of an infinitude of values. The philosopher G. J. Morgan extrapolated Kuhn's arguments from the diversity of applying shared values, to the diversity of values themselves, which may or may not include aesthetic preferences [3]. Morgan concluded that ultimately, "[t]here must also be what might be called the spontaneity of individual values" [3, p. 14]. Scientists might vary in their aesthetic preferences and the relative value they assign to them, and a scientist might not value beauty at all. Although Morgan justifies such spontaneity with the freedom of inquiry, the benefits of diverse values within a research community are largely the same as those articulated by Kuhn. This is particularly important for ML research into systems deployed a large scales, and whose predictions can have meaningful consequences. Greater diversity translates into consideration of a wider set of $\mathcal{P}$'s and $\mathcal{Q}$'s, better exploration of the solution space and the formulation of important new solution spaces.

Acknowledgments and Disclosure of Funding

I am grateful to the workshop organizers for creating a space for investigating and sharing these ideas, and to the five anonymous reviewers for thoughtful feedback on an earlier draft. My research is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC): Discovery Grant (RGPIN-2020-04995), Discovery Accelerator Supplement (RGPAS-2020-00095), and Discovery Launch Supplement (DGECR-2020-00343).

References

[1] F. J. Dyson. "Prof. Hermann Weyl, For.Mem.R.S." *Nature* 177.4506 (1956), pp. 457–458. URL: https://doi.org/10.1038/177457a0.

[2] S. Weinberg. *Dreams of a Final Theory*. New York: Pantheon Books, 1992.

[3] G. J. Morgan. "The Value of Beauty in Theory Pursuit: Kuhn, Duhem, and Decision Theory". *Open Journal of Philosophy* 3.1 (2013), pp. 9–14.

[4] P. Achinstein. "How to Defend a Theory Without Testing It: Niels Bohr and the "Logic of Pursuit"". *Midwest Studies In Philosophy* 18.1 (1993), pp. 90–120.

[5] S. Hossenfelder. *The present phase of stagnation in the foundations of physics is not normal*. Nov. 2018. URL: http://backreaction.blogspot.com/2018/11/the-present-phase-of-stagnation-in.html.

[6] S. Hossenfelder. *Lost in Math: How Beauty Leads Physics Astray*. Basic Books, 2018.

[7] J. Hawthorne. "Inductive Logic". *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta. Spring 2021. Metaphysics Research Lab, Stanford University, 2021.

[8] I. Douven. "Abduction". *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta. Summer 2021. Metaphysics Research Lab, Stanford University, 2021.

[9] J. Schickore. "Scientific Discovery". *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta. Summer 2018. Metaphysics Research Lab, Stanford University, 2018.

[10] C. Sartwell. "Beauty". *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta. Winter 2017. Metaphysics Research Lab, Stanford University, 2017.

[11] R. Reber, N. Schwarz, and P. Winkielman. "Processing Fluency and Aesthetic Pleasure: Is Beauty in the Perceiver's Processing Experience?" *Personality and Social Psychology Review* 8.4 (2004), pp. 364–382.

[12] A. A. Brielmann and P. Dayan. "A computational model of aesthetic value". *43rd European Conference on Visual Perception (ECVP 2021)*. 2021. URL: psyarxiv.com/eaqkc.

[13] D. A. McAllester. "Some PAC-Bayesian Theorems". *Machine Learning* 37.3 (1999), pp. 355–363.

[14] B. Guedj. "A Primer on PAC-Bayesian Learning" (2019). arXiv: 1901.05353 [stat.ML].

[15] S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[16] K. P. Burnham and D. R. Anderson. *Model Selection and Multimodel Inference*. Springer-Verlag, 2002.

[17] P. D. Grünwald. *The Minimum Description Length Principle*. MIT Press, 2007.

[18] E. T. Jaynes. "Information Theory and Statistical Mechanics". *Phys. Rev.* 106 (4 May 1957), pp. 620–630. DOI: 10.1103/PhysRev.106.620. URL: https://link.aps.org/doi/10.1103/PhysRev.106.620.

[19] J. Sokolic et al. "Generalization Error of Invariant Classifiers". *AISTATS*. Vol. 54. 2017, pp. 1094–1103.

[20] S. Chen, E. Dobriban, and J. H. Lee. "A Group-Theoretic Framework for Data Augmentation". *Journal of Machine Learning Research* 21.245 (2020), pp. 1–71.

[21] C. Lyle et al. "On the Benefits of Invariance in Neural Networks" (2020). arXiv: 2005.00178 [cs.LG].

[22] Y. Dubois et al. "Lossy Compression for Lossless Prediction". *Advances in Neural Information Processing Systems*. 2021. arXiv: 2106.10800 [cs.LG].

[23] J. Shawe-Taylor. "Building symmetries into feedforward networks". *1989 First IEE International Conference on Artificial Neural Networks, (Conf. Publ. No. 313)*. Oct. 1989, pp. 158–162.

[24] S. Ravanbakhsh, J. Schneider, and B. Póczos. "Equivariance Through Parameter-Sharing". *International Conference on Machine Learning (ICML)*. Ed. by D. Precup and Y. W. Teh. 2017, pp. 2892–2901.

[25] H. Zou and T. Hastie. "Regularization and variable selection via the elastic net". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.2 (2005), pp. 301–320.

[26] A. Krogh and J. Hertz. "A Simple Weight Decay Can Improve Generalization". *Advances in Neural Information Processing Systems*. 1992.

[27] N. Mehrabi et al. "A Survey on Bias and Fairness in Machine Learning". *ACM Comput. Surv.* 54.6 (2021).

[28] C. Molnar. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. 2019.

[29] C. Rudin. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead". *Nature Machine Intelligence* 1.5 (2019), pp. 206–215.

[30] R. Dubey et al. *Aha! moments correspond to meta-cognitive prediction errors*. June 2021. DOI: 10.31234/osf.io/c5v42. URL: psyarxiv.com/c5v42.

[31] F. Hutter et al. "Algorithm runtime prediction: Methods & evaluation". *Artificial Intelligence* 206 (2014), pp. 79–111.

[32] K. Eggensperger, M. Lindauer, and F. Hutter. "Neural Networks for Predicting Algorithm Runtime Distributions". *IJCAI*. 2018.

[33] J. Schmidhuber. "Formal Theory of Creativity, Fun, and Intrinsic Motivation (1990–2010)". *IEEE Transactions on Autonomous Mental Development* 2.3 (2010), pp. 230–247.

[34] G.-C. Rota. "THE PHENOMENOLOGY OF MATHEMATICAL BEAUTY". *Synthese* 111.2 (1997), pp. 171–182.

[35] A. Rahimi and B. Recht. *Reflections on Random Kitchen Sinks*. URL: http://www.argmin.net/2017/12/05/kitchen-sinks/.

[36] A. Rahimi and B. Recht. *An Addendum to Alchemy*. URL: http://www.argmin.net/2017/12/11/alchemy-addendum/.

[37] Z. C. Lipton. "The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability is Both Important and Slippery." *Queue* 16.3 (2018), pp. 31–57.

[38] T. S. Kuhn. *The Structure of Scientific Revolutions*. 2nd. Chicago: University of Chicago Press, 1970.

[39] F. D'Agostino. "Kuhn's Risk-Spreading Argument and the Organization of Scientific Communities". *Episteme* 1.3 (2005), pp. 201–209.