Journal for the History of Knowledge

ALEXANDER CAMPOLO

Loss

A Notion of Error in Machine Learning

▼ **RESEARCH ARTICLE**

▼ **ABSTRACT** This essay compares two statistical notions of error to draw out their distinctive epistemological and normative implications. The first sense crystallized in the nineteenth century as practical techniques for producing estimates from discrepant observations were interpreted as metaphysical laws of error. Historians have shown how these interpretations produced new forms of social knowledge and control over normal types. Although the metaphysical understanding of these laws was abandoned, error continued to be a major theme in twentieth-century statistical thought. A second sense of error associated with machine learning emerged in the second part of the century. This began when John von Neumann and Frank Rosenblatt developed statistical theories of computing, notably as machines able to learn from their environments. In the 1980s, researchers developed techniques such as backpropagation that used error measurements to improve a model's performance on learning tasks. Comparing these conceptions of error, I argue that we can perceive a larger shift from a politics of normal types revealed by the regularity of error to what I term a "politics of tasks" in which errors are used to refine desired behaviors.

▼ **KEYWORDS** error; statistics; machine learning; computing

▼ **ISSUE** Volume 6 (2025)

Alexander Campolo • Durham University, UK, alexander.campolo@durham.ac.uk

Cite this article: Alexander Campolo, 'Loss', *Journal for the History of Knowledge*, 6 (2025), 305–325
<https://dx.doi.org/10.55283/jhk.18717>

DOI: 10.55283/jhk.18717

BREPOLS

> Error is viewed, therefore, not as an extraneous and misdirected or misdirecting accident, but as an essential part of the process under consideration.
>
> —*John von Neumann*[1]

A typical paper in the field of machine learning describes a model architecture and reports performance on a suite of benchmark tasks. These measures often take the form of error rates. One well-known paper opens, "We trained a large, deep convolutional neural network to classify the 1.2 million high-resolution images in the ImageNet LSVRC-2010 content into the 1000 different classes. On the test data, we achieved top-1 and top-5 error rates of 37.5% and 17.0% which is considerably better than the previous state-of-the-art."[2]

This routinized reporting of error rates fits within a scientific context that has rendered error mundane. This was not always the case. As its obverse, error deserves special scrutiny within the history of knowledge. Errors are always to be avoided, but older ethical and theological senses have waned—the failure to follow divine command, to "wander" off the proper moral course.[3] The cliché "to err is human" gestures to these deeper anthropological stakes: error is the result of human finitude, contrasted with the infinity and perfection of God. It is the modern sciences that have taught us that errors need not indicate wayward souls or demonic presences, as difficult as these have been to fully exorcise.[4] A defective instrument can throw off the value of an observation; a momentary lapse in concentration can derail intricate calculations; various bugs break computer programs. Such errors are ordinary, unavoidable, and—in a revaluation—perhaps even desirable occurrences, the source of learning.[5] In place of a fallen human nature, psychologists today speak in deflated terms of cognitive "biases" that lead to "systematic errors" in judgment under conditions of uncertainty.[6]

For some well-informed observers, the culture of machine learning has even made error a special virtue. The statistician David Donoho argues that the public, competitive evaluation of machine learning models—more than any single set of architectural innovations—has made possible a new paradigm of computational science, one that he calls "frictionless reproducibility."[7] Even if one regards such epochal claims with skepticism, the practical and discursive centrality of error within machine learning systems is striking. Their training process is governed by loss functions, which measure the difference between

---

1 Von Neumann, "Probabilistic Logics," 43.
2 Krizhevsky et al., "ImageNet Classification," 1097.
3 Talcott, *The Problem of Error*, 91.
4 Canales, *Bedeviled*, 237.
5 Mayo, *Error*.
6 Tversky and Kahneman, "Judgment under Uncertainty," 1124.
7 Donoho, "Data Science," 4.

a model's actual—erroneous—and desired outputs. These measurements are then used as a signal to update the model's weights in order to improve its performance on a task.

Historians have used error to understand machine learning's political effects. Claudia Aradau and Tobias Blanke draw comparisons between a nineteenth-century biometric paradigm and the contemporary uses of machine learning for facial recognition, where error is "optimized" in machine learning.[8] Across the social sciences, researchers are studying how the many forms of error within machine learning are experienced and interpreted: training error, generalization error, as well as adjacent concepts such as bias, variance, and overfitting.[9] I wish to pursue this line of inquiry by contextualizing machine learning within the history of statistical thought, following the "continuity" identified by Adrian Mackenzie between the "advent of the Normal distribution" and the probabilistic ethos of machine learning researchers.[10] This entails an expansive conception of error, not limited to a quantity to be minimized, scientific or engineering mistakes, or even a way of drawing epistemological boundaries between the true and the false. In addition to all these senses, I am interested in the way that the concept of error works as an *evaluation*—creating norms, indicating ways to behave, govern, or live.[11]

This article compares two moments in the history of a statistical sense of error. The first moment culminates in the elaboration of laws of error in the nineteenth century. These laws entailed a revolutionary transformation in the value of error. Before them, as errors multiplied, so did confusion, leading to chaos. After them, errors, when properly aggregated, pointed toward true values at the mean. Eventually, so many things—bodies, deaths, or crimes—seemed to conform to statistical laws of error that they became a way of understanding order in the natural universe and human societies. This led to political interpretations that centered on average types—of a nation, class, or race. However, the moral of many excellent histories of this episode of statistical thought is that the enthusiasm that accompanied the rise of error laws precipitated a fall. Their universal status could not be sustained when scientists perceived different distributions in their data and conceptually when deviations came to be understood not as errors but in more positive terms. Theodore Porter thus concludes that the "reinterpretation [of the Gaussian distribution] as a law of genuine variation rather than of mere error, was the central achievement of nineteenth-century statistical thought."[12]

---

8 Aradau and Blanke, "Algorithmic Surveillance," 10.

9 Lin and Jackson, "Bias to Repair"; Annany, "Seeing Like an Algorithmic Error." For a perspective from within the machine learning community, see Zhang et al., "Understanding deep learning."

10 Mackenzie, *Machine Learners*, 105.

11 Galison, "Author of Error," 72; Daston, "Scientific Error," 3; Grosman and Reigeluth, "Perspectives on Algorithmic Normativities."

12 Porter, *Rise of Statistical Thinking*, 91.

The collapse of error laws did not, however, mean the end of "mere" error as a matter of statistical concern. In the main part of the article, I show how technical innovations and new intellectual contexts such as cybernetics, information theory, and neurophysiology produced discontinuities that transformed this stochastic conception of error as it migrated into the field of machine learning. No single, lawlike form of error could possibly govern the complex, multidimensional distributions that computers could now model. Instead of *revealing* the iconic, pre-existing form of a stable, normal curve, new techniques such as backpropagation made it possible for models to *learn* to estimate functions, distributing errors to *improve* performance on a given task. The earlier statistical sense of error referred to a distance between a known set of measurements and an unknown true value. The machine learning sense instead draws on known pairings of inputs and outputs to model them. Whereas the earlier statistical sense of error was interpreted politically as a source of spontaneous underlying order and stable social *types*, machine learning uses error—now understood as the difference between predicted and desired responses—as a source of feedback to drive improvement on learning *tasks*.

While the wide, stylized arc of this comparison renders some aspects of machine learning's sense of error intelligible, it necessarily obscures others. Certainly, it is not possible here to follow all the tantalizing philosophical paths that error marks out; even my narrower focus on the statistical sense of error brings a wide range of epistemological and political implications into play. Studying longer-term conceptual change in a comparative way also means less attention to microsociological detail—the practices of machine learning engineers as they tinker with loss functions, model architectures, and hardware configurations.[13] However, this methodological choice can indicate alternative entryways into histories of artificial intelligence, connecting it with probability and statistics alongside existing lineages in computing, logic, or psychology.[14]

The comparison developed in this article also presents an alternative to a view of machine learning—shared by the research community and historians—as a primarily non- or even antitheoretical engineering practice, where error is simply incorrect prediction.[15] While this conception is important, I seek to show that it exists alongside a substantive, theoretical valuation of error, which makes different normative interpretations possible. As we shall see, members of the machine learning community themselves draw on comparisons from the history of statistical thought to theorize their own practices.

---

13 Mackenzie, *Machine Learners*.

14 Joque, *Revolutionary Mathematics*; Jones, *Reckoning with Matter*; Dick, "Of Models and Machines"; Boden, *Mind as Machine*.

15 Breiman, "Statistical Modeling"; Jones, "How We Became Instrumentalists."

Error as Statistical Law

The first statistical sense of error begins with discrepant observations. Why is it that when we measure the same quantity repeatedly, we obtain different results? This practical question begs an epistemological one: given a set of discrepant observations, how should we distinguish between the true value and errors? These questions are ancient. The theoretically minded Greeks "enjoyed full power over their observations," *selecting* the ones that conformed to mathematical law and consigning the rest, presumed erroneous, to oblivion.[16] Other methods were practiced, with reasonable justifications. One might select the observation made under the best weather conditions or at a moment when one's instruments were working well.[17] In these cases, a value could be selected through careful judgment based on the observer's contextual knowledge.

A new method emerged late in the sixteenth century which proved so useful that it quickly gained adherents. This was, in the words of Churchill Eisenhart, "the practice of taking the arithmetic mean of two or more measurements or observed values of a single quantity as *the* value of the quantity indicated by these measurements of observations."[18] The combination of observations made under similar conditions—aggregation—replaced the selection of a single best value. Although the arithmetic was simple, it slowly led to a profound reconceptualization of error that was made explicit and justified by more sophisticated mathematical means over time: one that treats errors collectively, holistically, rather than as individual mistakes; as the probabilistic outcome of measurement process rather than a failure of that process; not as multiplying and leading to chaos but rather as a lawlike source of order, in the form of a distribution.

Crucial to this transformation was a change in focus, initiated in the middle of the eighteenth century by the mathematician Thomas Simpson, from the observations and unknown true values themselves to the specification of a distribution of errors from which a most likely estimate could be inferred.[19] The question of how to specify an error curve remained contentious as techniques for combining observations grew more sophisticated, reaching a high point around the turn of the nineteenth century when Adrien-Marie Legendre provided a compelling criterion for the more general problem of determining a set of coefficients for linear equations to fit a set of observational data: to minimize the sum of the squared value of the error terms.[20] Stephen Stigler has shown how Carl Friedrich Gauss was able to link the practical technique of least squares to parallel developments in probability being pursued by Pierre-Simon Laplace, by assuming that the most probable value for discrepant observations

16 Sheynin, "Density Curves," 191.
17 Eisenhart, "Laws of Error," 531.
18 Ibid., 530.
19 Simpson, "Advantage of Taking the Mean"; Stigler, *History of Statistics*, 91.
20 Ibid., 13.

is its mean and then specifying the probability distribution of errors—now, the normal or Gaussian curve—that "leads back to least squares." Although the circular logic of Gauss's derivation left much to be desired, it immediately inspired Laplace to ground it using his own work on the central limit theorem to derive the normally distributed error curve, thereby connecting probability and the combination of observations.[21]

The practical success of least squares and the sophistication of its probabilistic justifications led to metaphysical inflation: a law of error in the form of a distribution centered on the mean value that governed all measurements and observations. Porter gives a sense of this exuberance, quoting the French mathematician Joseph Fourier's 1819 ode to the universality of the error curve, which "unites the most diverse effects, and discovers in them common properties. Its object has nothing of the contingent, nothing of the fortuitous. Imprinted in all of nature, it is a preexistent element of universal order."[22] During these heady days the Belgian astronomer-turned-social scientist Adolphe Quetelet noticed that many moral and social phenomena, from the number of crimes to measurements of bodies, were distributed in ways that resembled the astronomical error curve, and today he is often blamed for drawing sweeping conclusions from this observation. In the words of Ian Hacking, Quetelet "transformed the theory of measuring unknown physical quantities, with a definite probable error, into the theory of measuring ideal or abstract properties of a population. Because these could be subjected to the same formal techniques, they became real qualities."[23]

Not only were these abstract properties or types endowed with reality, they thereby became, as many historians and philosophers have observed, *normative*. Georges Canguilhem, for instance, notes how this movement from error to normality draws on "the realist philosophical tradition" in which "a generality observable in fact takes the value of realized perfection, and a common characteristic, the value of an ideal type."[24] These social types, just like the true value of an astronomical measurement, were unobservable but powerfully real when they could be modeled using the astronomical error curve. This made new kinds of value judgments possible, based on the ambiguous normativity of the astronomical conception of error, with more precise, quantitative distinctions between a desirable normality and pathology or deviance, now associated with error. This valuation of types was also political, as Porter and Hacking

21 Ibid., 143.

22 Fourier, "Théorie analytique des assurances," 189, quoted in Porter, *Rise of Statistical Thinking*, 99.

23 Hacking, *Taming of Chance*, 108. An exception to this consensus is Stigler, who is more willing to acknowledge that contemporary judgments of Quetelet's apparent naïveté may fail to appreciate the deep epistemological difficulties involved with the use of statistics in the social sciences. See Stigler, *History of Statistics*, 161.

24 Canguilhem, *Normal and the Pathological*, 125.

stress; the submission of chaotic social causes to the ordered symmetry of the normal distribution promised a way out of post-revolutionary anarchy through an enlightened, self-governing, liberal order.[25]

Error and Desired Behaviors

We no longer speak of a single law of error, but of a plurality of possible models for analyzing data.[26] The universality of the error law proved incompatible with the observation and derivation of other distributions, beginning with Francis Galton and later Karl Pearson in heredity, William Lexis in demography, and Gustav Fechner in psychophysics.[27] But our more modest, pluralistic outlook does not mean that error ceased to be salient. Many important statistical ideas from the first part of the twentieth century concerned errors, often in terms of sampling. Ideas about error were transformed in new estimation criteria such as "efficiency" and "sufficiency" in R.A. Fisher's maximum likelihood framework.[28] Jerzy Neyman and Egon Pearson refined notions of error in experimental design, separating two different "sources" or types of error in evaluating hypotheses of a distinctive statistical form: that a given sample has been randomly drawn from a population.[29] Moving from the sciences to engineering, the practical imperatives of quality control, not least for weapons manufacturing during the Second World War, spurred the development of techniques such as the minimax and sequential analysis by Abraham Wald—where the term "loss" appeared in a statistical sense in 1939.[30] This sense was expanded in the decision sciences, where errors were quantified in terms of "costs," "risk," or "regret" when certain potential states of the world were realized—terminology that persists in machine learning.[31] Other concepts such as "noise" in information theory pointed to a wider range of related concerns, ranging from the degradation of communicative signals to the disordering of thermodynamic systems.[32]

---

25 Porter, *Rise of Statistical Thinking*, 104; Hacking, *Taming of Chance*, 168.

26 Eisenhart, "Laws of Error," 565.

27 Ibid., 531.

28 Fisher, "Mathematical Foundations," 316. See Edwards, "History of Likelihood"; Aldrich, "Making of Maximum Likelihood."

29 Neyman and Pearson, "Certain Test Criteria," 177. See also Gigerenzer et al., "Inference Experts." Interestingly, Fisher strongly objected to the use of the word "error" in reference to their acceptance procedures, based on what he saw as fundamental differences between the use of statistics for "learning" in science versus their more "mechanistic" application in industry or engineering. See Fisher, "Statistical Methods," 73.

30 "In many cases, especially in statistical questions concerning industrial production, we are able to express the importance of an error in terms of money, that is to say, we can express the loss caused by the error considered in terms of money." Wald, "Theory of Statistical Estimation," 302. See also Wallis, "Statistical Research Group"; Klein, "Economics for a Client."

31 Savage, *Foundations of Statistics*, 163.

32 Shannon, "Mathematical Theory."

In the movement from these midcentury contexts to machine learning in the present, the conception of error took on an increasingly temporalized, processual form: from running experiments, to evaluating manufacturing processes, to training computational models. This marks a continuity with the earlier astronomical conception in certain respects, which was able to envision measurement as a stochastic process that, when repeated, converged toward true values. However, this sense of error was also shaped within the context of cybernetics, where the concept of feedback as a process of error correction played a central role. In their 1943 paper, Arturo Rosenblueth, Norbert Wiener, and Julian Bigelow famously broadened the scope of the term "feedback" beyond mechanical engineering to refer to any "behavior…controlled by the margin of error at which the object stands at a given time with reference to a relatively specific goal"—although this more abstract presentation obscured its original impetus: Wiener's probabilistic techniques for anti-aircraft targeting.[33] Here, feedback named a means for moving from inputs to outputs, itself a broadening of the relationship between stimulus and response inherited from behaviorist psychology.[34]

As William Aspray has noted, John von Neumann connected this notion of feedback with a stochastic conception of error in a series of lectures on the topic of self-reproducing "automata" (logical representations of neurons that turn an input signal into an output response) beginning in the late 1940s.[35] The significance of these lectures lies in the way they connect a statistical model of error, supplemented by more recent innovations in thermodynamics and communication engineering, to the world of computing, which was then understood principally in logical terms.[36] This was notably a *temporal* conception of error. "Time," von Neumann observed, "never occurs in logic, but every network or nervous system has a definite time lag between the input signal and the output response. A definite temporal sequence is always inherent in the operation of such a real system."[37]

In von Neumann's usage, "error" refers not to discrepancies due to imperfections of observation and instrumentation in measurement, but rather to "malfunctions" that occur stochastically during a self-directed process of synthesis—the combination of automata into networks. His objective was to design a network so that it could "control" error—that is, operate with some acceptable degree of "accuracy" in the movement from input to output signal even when individual components malfunction with some defined probability.[38] Instead of aggregating measurements to produce an optimal estimate of

33 Rosenblueth et al., "Behavior, Purpose, and Teleology," 19; Galison, "Ontology of the Enemy," 262–63.
34 Dupuy, *Mechanization of the Mind*, 46.
35 Aspray, *John von Neumann*, 200.
36 Von Neumann, "Probabilistic Logics," 43. The dichotomy between logical and statistical conceptions of computation is a theme in the history of artificial intelligence. See Boden, *Mind as Machine*, 190.
37 Von Neumann, "Probabilistic Logics," 44.
38 Ibid., 64.

an unknown quantity, von Neumann drew from communication engineering to propose a system of "multiplexing" whereby messages are carried simultaneously on many rather than a single best line.[39] Rather than producing the most likely measurement, error is managed to ensure some acceptable level of the overall performance of a system over time.[40]

The problem of ensuring the functioning of the whole system in spite of the stochastic malfunctioning of its parts served as inspiration for a work associated with the birth of machine learning: Frank R. Rosenblatt's 1958 paper introducing the perceptron. Rosenblatt credited von Neumann as being one of "a relatively small number of theorists…concerned with the problems of how an imperfect neural network, containing many random connections can be made to perform reliably."[41] Rosenblatt's work on neural networks has of course featured centrally in histories of machine learning and artificial intelligence, which embed it in controversies between symbolic and connectionist approaches to artificial intelligence, and even the political implications of Rosenblatt's neurophysiological and psychological analogies.[42] Drawing on the wider comparative perspective of the history of error in statistics, I will focus on his formulation of a statistical pattern recognition problem, and the place of error within "a mathematical analysis of learning."[43]

Rosenblatt began with a venerable statistical rejoinder to the logical view of computing that predominated at the time: the complexity of computing technology, like that of a gas in physics or a population in sociology, means that "only the gross organization can be characterized, and the precise structure is unknown."[44] In a subsequent talk, Rosenblatt invoked the second law of thermodynamics to argue that an exclusively logical computing system that executes explicit rules written by humans can never "improve in its ability to organize, and to draw valid conclusions from information."[45] Only a system capable of "observing and learning from the organization of the surrounding world" might be able to adapt its structure to the world around it.[46] This imperative to "improve performance," to respond to stimulus signals in a differentiated environment—rather than estimate the true value of a physical parameter—remains at the heart of machine learning's distinctive notion of error.[47]

---

39 Ibid., 63–64.

40 Galison argues that this systematicity distinguishes the cybernetic concept of feedback. "Ontology of the Enemy," 262.

41 Rosenblatt, "The Perceptron," 387.

42 See Minsky and Papert, *Perceptrons*; Boden, *Mind as Machine*; McCorduck, *Machines Who Think*; Olazaran, "Perceptrons Controversy"; Halpern, "The Future Will Not Be Calculated."

43 Rosenblatt, "The Perceptron," 394.

44 Ibid., 387–88.

45 Rosenblatt, "Two Theorems," 424.

46 Ibid.

47 This imperative was already articulated in Rosenblueth et al., "Behavior, Purpose, and Teleology," 18.

In a subsequent report, "Principles of Neurodynamics," Rosenblatt detailed how such learning might actually occur, through the optimization of connections between three elementary units: sensory units, which process input signals from the environment and activate (emitting an output signal) if a threshold is reached; association units, which sum multiple input signals and emit an output if this quantity is greater than a threshold; and response units, which emit signals outside of the network based on the inputs they receive.[48] Connected in a network, these elements form a simple perceptron. In this system, "learning" means feedback-driven changes in the connections between units so that they come to correctly classify environmental stimuli. Rosenblatt proposed a number of feedback-based reinforcement mechanisms, notably an "error-correction system," in which negative feedback is given for an incorrect response.[49]

This led Rosenblatt to a more basic conceptual level: "we must first re-examine the concept of 'error' which has been employed so far as a criterion for reinforcement."[50] Error refers no longer, as in astronomy and geodesy, to the difference between an observed value and an *unknown* actual position, but rather to the difference between the "obtained" and a not only known but "desired" response—the latter term stemming from behaviorist contexts and indicating a subtle shift from an epistemology of uncertainty to normativity.[51] Tantalizingly, he sketched a probabilistic technique for using these errors to modify connections between units in a training process: "the back-propagating error correction procedure"—so-called because it begins with errors in the terminal response units, "propagating corrections back towards the sensory end of the network."[52] At a high level, the procedure compares the state of the units in the output layer for an exemplary, desired response to the response obtained during the training process, with errors (differences between these two states) used to probabilistically change the configuration of the network toward the desired state.

Although this statistical, connectionist approach was subjected to strong criticism during the 1960s and 1970s, by the middle of the 1980s conceptual and technical advances prompted a resurgence of interest among a group of researchers including Yann LeCun, David Rumelhart, Geoffrey Hinton, and Ronald Williams.[53] Their training method was similar to that of Rosenblatt: compare desired outputs to model outputs and use these measurements to

48 Rosenblatt, "Principles of Neurodynamics," 81–82.

49 Ibid., 91–92.

50 Ibid., 287.

51 Ibid., 288. Compare this formulation with more neutral earlier pairing of "stimulus and response" in Skinner, *The Behavior of Organisms*, 25. See also Campolo and Schwerzmann, "From Rules to Examples" for a theoretical discussion of this "exemplary" form of normativity.

52 Rosenblatt, "Principles of Neurodynamics," 292.

53 Minsky and Papert, *Perceptrons*; Olazaran, "Perceptrons Controversy."

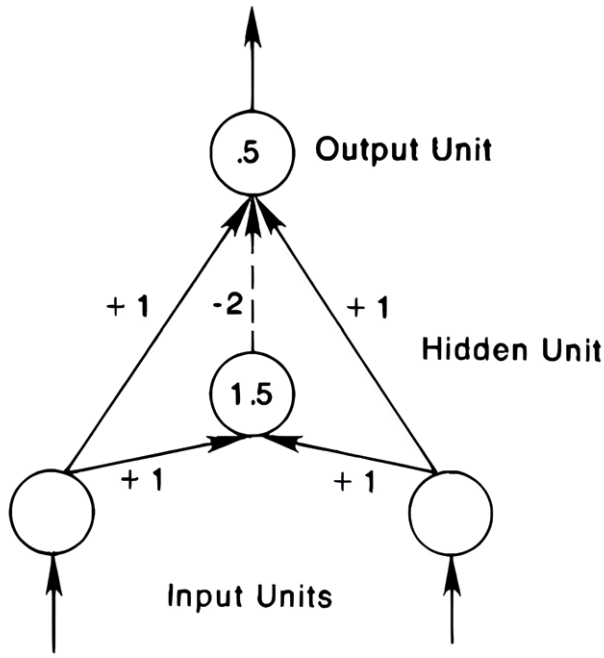

Figure 1. A simple network with an intermediate hidden unit. Source: Rumelhart et al., "Learning Internal representations," 321. © 1986 Massachusetts Institute of Technology, by permission of The MIT Press.

adjust weights so as to reduce this difference—error.[54] However, their models contained an innovation: hidden units that produce "internal representations" as input signals move through the network, allowing these neural networks to address a much wider range of problems in which outputs do not strictly resemble inputs [**Fig. 1**].

To update the weights of these intermediate layers, Rumelhart, Hinton, and Williams introduced a second, backward pass through the network, which "propagates the errors" beginning with the final layer, recursively updating "weight changes for all connections that feed into the final layer," repeating in each subsequent step, a process that Mackenzie describes as a "shuttling movement."[55] In these more complex networks, it is not clear what the overall shape of the "error surface" is, making it impossible to prove that their method of gradient descent—"using the derivatives of the error measure with respect to each weight"—will find an optimal set of weights without getting "stuck"

54 Rumelhart et al., "Learning Internal Representations"; LeCun, "Learning Process"; Anderson and Rosenfeld, *Talking Nets*.

55 Rumelhart et al., "Learning Internal representations," 321; Mackenzie, *Machine Learners*, 207.

in local minima. However, drawing on evidence from simulations, they argued that this method, called "back-propagation of errors," worked empirically on a wide range of tasks by 1986, mitigating concerns about its theoretical optimality.[56]

At this point, a concept of error in machine learning is becoming more distinctive. In the earlier error laws, aggregation was used to resolve discrepancies among observations to produce a most probable estimate of an unknown parameter. A great advantage of this method was that once a plausible distribution of these errors was specified, very little needed to be known or assumed about the different causes or sources of error, which tended to cancel each other out. Over the course of the twentieth century, this conception of error was reconfigured in the context of computing and the intellectual matrix of cybernetics. Instead of aggregating errors to converge toward a most probable estimate, errors were measured against an output and propagated backwards to refine the model. This required a collection of exemplary pairings of inputs and outputs—known, *desired* outputs—to which a model's output can be compared. This sense of error took on a recursive, backwards-facing temporality, in which weights are updated processually, beginning with the end—in the sense both of the final layer and the use of a desired or target value to compute an error measure.[57] Only once a set of adequate weights has been computed over time through a training process is the model tested on unknown examples, measuring its ability to *generalize*, a different test of inductive ability.

This account of error in machine learning is sometimes interpreted as a sign of the field's pragmatic engineering ethos.[58] Errors are merely incorrect predictions; they do not and perhaps should not tell us anything about the data-generating process. It is thought to be a great virtue that this "algorithmic modeling culture" obviates the need for the misleading theoretical interpretations characteristic of earlier statistical thought.[59] However, there is an alternative branch of machine learning research that draws on precisely these histories to theorize machine learning itself, indicating deeper connections between these conceptions of error. One researcher in this tradition, Vladimir Vapnik, began his career at the Institute of Control Sciences in Moscow, publishing important work in probability with his colleague Alexey Chervonenkis.[60] Later, he emigrated to the United States and worked at the Adaptive System Research Department of Bell Laboratories, where he further developed theories of statistical learning.[61]

---

56 Rumelhart et al., "Learning Internal Representations," 328. In addition to Rosenblatt's usage, the term "backpropagation" was also used in Paul Werbos's 1974 dissertation, "Beyond Regression."

57 For more on this "retroactive" modeling logic, see Amoore, "Machine Learning Political Orders," 9.

58 Jones, "How We Became Instrumentalists," 678.

59 Breiman, "Statistical Modeling."

60 Vapnik and Chervonenkis, "Uniform Convergence."

61 Law, "Bell Labs."

Vapnik framed his work in comparative historical terms. In his 1998 book *Statistical Learning Theory*, he delineated two statistical paradigms in historical succession. The first "particular" or "parametric" approach emerged fully in the 1920s, when Fisher began to consider statistical problems in terms of estimating functions from empirical data. This paradigm addressed well-defined, applied problems of statistical inference, assuming the researcher's familiarity with the problem, knowledge of both "the physical law that generates the stochastic properties of the data" and "the function to be found up to a finite number of these parameters."[62] A pillar of this "philosophy"—a term that Vapnik uses frequently to address these epistemological issues—is the normal law and the central limit theorem as articulated by the earlier error theorists.[63]

The second paradigm, that of "general statistical inference," emerged due to technological changes, beginning with Rosenblatt and running through Rumelhart, Hinton, and Williams; Vapnik refers to the latters' work as "the second birth of the perceptron."[64] By the 1960s, computers made it possible to address much more complex, high-dimensional functions, which are not easily approximated by a small set of known distributions where models could be selected on the basis of the investigator's familiarity with the problem. Beyond the number of parameters, these problems are characterized by severe epistemic limitations: "one does not have reliable a priori information about the statistical law underlying the problem or about the function that one would like to approximate."[65] As computing technologies allowed researchers to address these new types of problems, Vapnik proposed new inductive criteria to evaluate solutions. For instance, "empirical risk minimization" states that the best function is the one that minimizes errors on the training set of paired inputs and desired outputs [**Fig. 2**]. Later, Vapnik expanded this into a more theoretically robust (but arguably less practically effective) structural risk minimization principle, which adds a further requirement of finding a function with a small capacity.[66]

How should we choose the best such function? By measuring "the *loss* or discrepancy between the responses of the supervisor to a given input and the responses provided by the learning machine."[67] By now this criterion should sound familiar, echoing ideas found in the work of Hinton, Rumelhart, and Williams, and earlier in that of Rosenblatt and even Wald. Vapnik gave these functions a further theoretical interpretation. Different types of learning tasks can be characterized not by specifying a probability distribution in advance, as the error theorists and their descendants had, but rather by specifying a loss function: 0–1 loss for pattern recognition tasks, squared loss for regression

62 Vapnik, *Statistical Learning Theory*, 3.
63 Ibid., 4.
64 Ibid., 12. For more on Vapnik and support vector machines, see Mackenzie, *Machine Learners*, 142–48.
65 Vapnik, *Statistical Learning Theory*, 3.
66 Ibid., 7–10
67 Vapnik, "An Overview," 988.

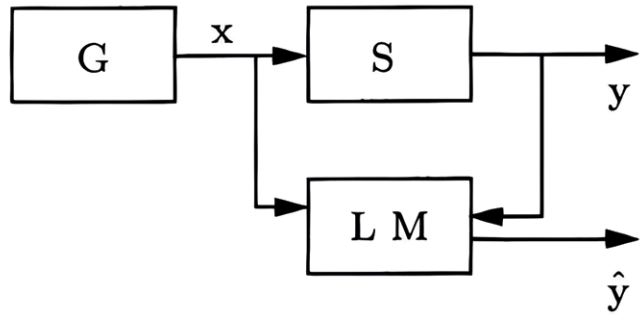

Figure 2. Vapnik's more general representation of the (machine) learning problem. The generator G produces data (x) according to some distribution for which the supervisor (S) returns responses (y). The learning machine (LM) observes a set of these desired (x, y) pairs and constructs a function to return similar responses to the supervisor. Source: Vapnik, *Statistical Learning Theory*, 20. © 1998 John Wiley and Sons.

tasks, and log loss for density estimation tasks.[68] Thus, in addition to providing a practical means for updating model weights through backpropagation, loss functions can provide a theoretical means for distinguishing between different learning tasks.

While not all machine learning researchers are interested in Vapnik's philosophy of inductive inference—and it should be said that many advances in machine learning have occurred precisely by breaking with earlier philosophies of statistical inference—all of them do use loss functions to refine their models. And the existence of such theories in the first place points beyond a purely pragmatic understanding of machine learning as engineering, endowed with a transhistorical imperative to minimize error. Rather, at critical junctures in the recent history of machine learning, error was the object of sophisticated reflection. Von Neumann pointed away from a logical-symbolic toward a statistical-probabilistic understanding of computing, drawing on physics and information theory, with error given a temporal dimension. Rosenblatt deployed this statistical understanding in a neurophysiological context, where computers recognize patterns in environments to *improve* performance on tasks over time. Techniques for measuring loss—divergences between a model's predicted output and known, desired outputs—gave errors a positivity

68 Vapnik, *Statistical Learning Theory*, 25–32. See also Williamson, "Loss Functions."

when they were propagated in a feedback system to recalibrate the model's weights. Later, these loss functions took on a conceptual character in characterizing these objectives and tasks.

Error from Types to Tasks

What can we now say after comparing two moments in the history of statistical treatment of error? The earlier astronomical conception was rooted in problems of observation and measurement. This statistical treatment of error showed, remarkably, that aggregation would not lead further into error, to unintelligibility or chaos, but could converge toward the truth—a single most probable value. Large numbers of independent events converged to a lawlike form; a distribution was understood to govern "error"—in the singular—itself. The success of these methods, first for the practical problem of delivering reliable estimates, spilled into many areas of both knowledge and politics. Of these, the most profound was the transformation, analyzed by Hacking, by which the reality of the value of a measurement—the true position of a star in the sky—was transferred to any social or biological variable whose distribution resembled the normal one.[69] Populations could be characterized by this single, typical feature, and individuals *evaluated* as normal or abnormal by their proximity to this type. More broadly, the appearance and apparent stability of these types out of apparent chaos and error were interpreted as comforting evidence of the reality of an underlying social order in a society undergoing the revolutionary changes of modernity.

During the twentieth century, the coordinates of this statistical understanding of error shifted. First, the singularity of the law of error and the univocity of the normal distribution gave way to a plurality of probability distributions and data models. Later, the spatialized context of measurement was temporalized, as statistical treatments of measurement were applied in different processes, ranging from the design and evaluation of experiments, to sampling, to the governance of quality control procedures in manufacturing, stimulated by mobilization for the Second World War. The development of digital computing, initially conceptualized in logical terms, was reinterpreted by von Neumann and Rosenblatt through a processual, probabilistic lens, where errors were used as a source of feedback for a machine learning from its environment. Here it is not the measurement or observation process that is problematic, subject to error. On the contrary, a set of known, exemplary pairs of inputs and outputs is given. What is instead at issue is the formulation of learning *tasks* in such a way that the gradual, self-directed, feedback-driven minimization of error during a training or learning process makes it possible for a model to "generalize" to instances not included in this set, which is taken to indicate that some essential

69 Hacking, *Taming of Chance*, 107.

(but not directly knowable) feature of the underlying data-generating process has been captured by the model.

This comparative history of error can attune us to the shifting valences of a concept that oscillates between factual and normative senses. The normativity of error in machine learning is no longer predetermined by the formal symmetry around the mean value of a normal distribution. Ethical and political interpretations of the earlier error laws took on a new force when they treated this mean as an objective social essence—analogous to the true position of a star—that then served as a reference point for moral valuations of normality and moderation. We have only recently entered a machine learning age. However, certain directions are becoming intelligible. One concerns the imperative of continuous improvement, progress, or even teleology implied in machine learning's temporalized sense of error. How does the use of error as a feedback signal in a training process create distinctive temporal valuations?[70]

Another critical direction would attend to the normative relationship between inputs and *desired* outputs that are meant to represent different learning tasks. Beyond questions of construct validity, these tasks take on their own social reality through the collective scientific enterprise of benchmarking, where a common error metric is used to rank models and drive further development.[71] Rather than the appearance of the error law indicating some real, normative tendency or even essence of a population, improvements in model performance are interpreted as evidence that some human-like "capability" exists.[72] But these performance metrics tell us as much about the types of tasks or capability we *value* as any abstract comparison between humans and machines. Analyzing the formulation of learning tasks illuminates the slippery movement from training objectives such as predicting the next most likely word in a sequence to more ambiguous evaluations and interpretations, such as the degree to which such models can be said to understand language at a human level.[73] This normative difference between desired and obtained response in a task thus takes on a further ethical sense when it is used as a loss signal to direct the behaviors of both models and the people governed by them.

Acknowledgements

The author would like to acknowledge Louise Amoore, Benjamin Jacobsen, Ludovico Rella, and Cindy Kaiying Lin who provided valuable comments.

70 Campolo, "State-of-the-Art."
71 Luitse et al., "AI Competitions."
72 Grill, "Constructing Capabilities."
73 Hendrycks et al., "Measuring Multitask Language Understanding."

Funding Information

The research has received funding from the European Research Council (ERC) under Horizon 2020, Advanced Investigator Grant ERC-2019-ADG-883107-ALGOSOC "Algorithmic Societies: Ethical Life in the Machine Learning Age."

About the Author

Alexander Campolo is a postdoctoral research associate on the "Algorithmic Societies" project in the Department of Geography at Durham University. His current research draws from the history of science and technology to explore epistemological and political implications of machine learning. He received his PhD from New York University and has previously worked at the Institute on the Formation of Knowledge at the University of Chicago and the AI Now Institute.

Bibliography

Aldrich, John. "R. A. Fisher and the Making of Maximum Likelihood 1912–1922." *Statistical Science* 12, no. 3 (1997): 162–76. https://doi.org/10.1214/ss/10300379064.

Amoore, Louise. "Machine Learning Political Orders." *Review of International Studies* 49, no. 1 (2023): 20–36. https://doi.org/10.1017/S0260210522000031.

Anderson, James A., and Edward Rosenfeld, eds. *Talking Nets: An Oral History of Neural Networks*. MIT Press, 2000.

Annany, Mike. "Seeing Like an Algorithmic Error: What Are Algorithmic Mistakes, Why Do They Matter, How Might They Be Public Problems?" *Yale Journal of Law and Technology* 24 (2022): 342–64.

Aradau, Claudia, and Tobias Blanke. "Algorithmic Surveillance and the Political Life of Error." *Journal for the History of Knowledge* 2 (2021): 1–13. https://doi.org/10.5334/jhk.42.

Aspray, William. *John von Neumann and the Origins of Modern Computing*. MIT Press, 1990.

Boden, Margaret. *Mind As Machine: A History of Cognitive Science*. Clarendon Press, 2006.

Breiman, Leo. "Statistical Modeling: The Two Cultures." *Statistical Science* 16, no. 3 (2001): 199–231. https://doi.org/10.1214/ss/1009213726.

Campolo, Alexander. "State-of-the-Art: The Temporal Order of Benchmarking Culture." *Digital Society* 4, no. 35 (2025): 1–10. https://doi.org/10.1007/s44206-025-00190-x.

Campolo, Alexander, and Katia Schwerzmann. "From Rules to Examples: Machine Learning's Type of Authority." *Big Data & Society* 10, no. 2 (2023): 1–13. https://doi.org/10.1177/20539517231188725.

Canales, Jimena. *Bedeviled: A Shadow History of Demons in Science*. Princeton University Press, 2020.

Canguilhem, Georges. *The Normal and the Pathological*. Translated by Carolyn R. Fawcett. Zone Books, 1991.

Daston, Lorraine. "Scientific Error and the Ethos of Belief." *Social Research* 72, no. 1 (2005): 1–28. https://doi.org/10.1353/sor.2005.0016.

Dick, Stephanie. "Of Models and Machines: Implementing Bounded Rationality." *Isis* 106, no. 3 (2015): 623–34. https://doi.org/10.1086/683527.

Donoho, David. "Data Science at the Singularity." *Harvard Data Science Review* 6, no. 1 (2024). https://doi.org/10.1162/99608f92.b91339ef.

Dupuy, Jean-Pierre. *The Mechanization of the Mind: On the Origins of Cognitive Science*. Translated by M. B. DeBevoise. Princeton University Press, 2000.

Edwards, A. W. F. "The History of Likelihood." *International Statistical Review* 42, no. 1 (1974): 9–15. https://doi.org/10.2307/1402681.

Eisenhart, Churchill. "Laws of Error." In *Encyclopedia of Statistical Sciences*. Vol. 4, *Icing the Tails to Limit Theorems*, edited by Samuel Kotz, Norman Lloyd Johnson, and Campbell B. Read. Wiley, 1983.

Fisher, Ronald A. "On the Mathematical Foundations of Theoretical Statistics." *Philosophical Transactions of the Royal Society Series A, Containing Papers of a Mathematical or Physical Character* 222 (1922): 309–68. https://doi.org/10.1098/rsta.1922.0009.

———. "Statistical Methods and Scientific Induction." *Journal of the Royal Statistical Society. Series B (Methodological)* 17, no. 1 (1955): 69–78. https://doi.org/10.1111/j.2517-6161.1955.tb00180.x.

Fourier, Joseph. "Extrait d'un Mémoire sur la Théorie Analytique des Assurances." *Annales de Chimie et de Physique* 2, no. 10 (1819): 177–89.

Galison, Peter. "The Ontology of the Enemy: Norbert Wiener and the Cybernetic Vision." *Critical Inquiry* 21, no. 1 (1994): 228–66. https://doi.org/10.1086/448747.

———. "Author of Error." *Social Research* 72, no. 1 (2005): 63–76. https://doi.org/10.1353/sor.2005.0032.

Gigerenzer, Gerd, Zeno Swijtink, Theodore Porter, Lorraine Daston, John Beatty, and Lorenz Krüger. "The Inference Experts." In *The Empire of Chance: How Probability Changed Science and Everyday Life*. Cambridge University Press, 1989. https://doi.org/10.1017/CBO9780511720482.005.

Grill, Gabriel. "Constructing Capabilities: The Politics of Testing Infrastructures for Generative AI." *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (2024): 1838–49. https://doi.org/10.1145/3630106.3659009.

Grosman, Jérémy, and Tyler Reigeluth. "Perspectives on Algorithmic Normativities: Engineers, Objects, Activities." *Big Data & Society* 6, no. 2 (2019): 1–12. https://doi.org/10.1177/2053951719858742.

Hacking, Ian. *The Taming of Chance*. Cambridge University Press, 1990.

Halpern, Orit. "The Future Will Not Be Calculated: Neural Nets, Neoliberalism, and Reactionary Politics." *Critical Inquiry* 48, no. 2 (2022): 334–59. https://doi.org/10.1086/717313.

Hendrycks, Dan, Collin Burns, Steven Basart, et al. "Measuring Massive Multitask Language Understanding." Preprint, ArXiv, January 12, 2021. https://doi.org/10.48550/arXiv.2009.03300.

Jones, Matthew L. *Reckoning with Matter: Calculating Machines, Innovation, and Thinking about Thinking from Pascal to Babbage*. University of Chicago Press, 2016.

———. "How We Became Instrumentalists (Again): Data Positivism Since World War II." *Historical Studies in the Natural Sciences* 48, no. 5 (2018): 673–84. https://doi.org/10.1525/hsns.2018.48.5.673.

Joque, Justin. *Revolutionary Mathematics: Artificial Intelligence, Statistics and the Logic of Capitalism*. Verso, 2022.

Klein, Judy L. "Economics for a Client: The Case of Statistical Quality Control and Sequential Analysis." *History of Political Economy* 32, (2000): 25–70. https://doi.org/10.1215/00182702-32-Suppl_1-25.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks." *Proceedings of the 26th International Conference on Neural Information Processing Systems* (2012): 1097–105.

Law, Harry. "Bell Labs and the 'Neural' Network, 1986–1996." *BJHS Themes* 8 (2023): 143–54. https://doi.org/10.1017/bjt.2023.1.

LeCun, Yann. "Learning Process in an Asymmetric Threshold Network." In *Disordered Systems and Biological Organization*, edited by E. Bienenstock, F. Fogelman Soulié, and G. Weisbuch. Springer, 1986. https://doi.org/10.1007/978-3-642-82657-3_24.

Lin, Cindy Kaiying, and Steven J. Jackson. "From Bias to Repair: Error as a Site of Collaboration and Negotiation in Applied Data Science Work." *Proceedings of the ACM on Human-Computer Interaction* 7 (2023): 131:1–32. https://doi.org/10.1145/3579607.

Luitse, Dieuwertje, Tobias Blanke, and Thomas Poell. "AI Competitions as Infrastructures of Power in Medical Imaging." *Information, Communication & Society* (2024): 1–22. https://doi.org/10.1080/1369118X.2024.2334393.

Mackenzie, Adrian. *Machine Learners: Archaeology of a Data Practice*. MIT Press, 2017.

Mayo, Deborah G. *Error and the Growth of Experimental Knowledge*. University of Chicago Press, 1996.

McCorduck, Pamela. *Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence*. AK Peters, 2004.

Minsky, Marvin, and Seymour Papert. *Perceptrons: An Introduction to Computational Geometry*. MIT Press, 1969.

Neumann, John von. "Probabilistic Logics and the Synthesis of Reliable Organisms from Unreliable Components." In *Automata Studies*, edited by Claude E. Shannon and John McCarthy. Princeton University Press, 1956.

Neyman, Jerzy, and Egon S. Pearson. "On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference: Part I." *Biometrika* 20, no. 1/2 (1928): 175–240. https://doi.org/10.2307/2331945.

Olazaran, Mikel. "A Sociological Study of the Official History of the Perceptrons Controversy." *Social Studies of Science* 26, no. 3 (1996): 611–59. https://doi.org/10.1177/030631296026003005.

Porter, Theodore M. *The Rise of Statistical Thinking: 1820–1900*. Princeton University Press, 1986.

Rosenblatt, Frank. "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain." *Psychological Review* 65, no. 6 (1958): 386–408. https://doi.org/10.1037/h0042519.

———. "Two Theorems of Statistical Separability in the Perceptron." In *Mechanisation of Thought Processes: Proceedings of a Symposium Held at the National Physical Laboratory*. H.M. Stationary Office, 1959.

———. "Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms." Cornell Aeronautical Laboratory, March 16, 1961.

Rosenblueth, Arturo, Norbert Wiener, and Julian Bigelow. "Behavior, Purpose and Teleology." *Philosophy of Science* 10, no. 1 (1943): 18–24. https://doi.org/10.1086/286788.

Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. "Learning Internal Representations by Error Propagation." In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations*. Edited by David E. Rumelhart and James L. McClelland. MIT Press, 1986. https://doi.org/10.7551/mitpress/5236.003.0012.

Savage, Leonard J. *The Foundations of Statistics*. John Wiley & Sons, 1954.

Shannon, Claude E. "A Mathematical Theory of Communication." *Bell System Technical Journal* 27, no. 3 (1948): 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.

Sheynin, Oscar. "Density Curves in the Theory of Errors." *Archive for History of Exact Sciences* 49, no. 2 (1995): 163–96. https://doi.org/10.1007/BF00376546.

Simpson, Thomas. "XIX. A Letter to the Right Honourable George Earl of Macclesfield, President of the Royal Society, on the Advantage of Taking the Mean of a Number of Observations, in Practical Astronomy." *Philosophical Transactions of the Royal Society of London* 49 (1755): 82–93. https://doi.org/10.1098/rstl.1755.0020.

Skinner, B.F. *The Behavior of Organisms: An Experimental Analysis*. B.F. Skinner Foundation, 1991.

Stigler, Stephen M. *The History of Statistics: The Measurement of Uncertainty Before 1900*. Belknap Press, 1986.

Talcott, Samuel. *Georges Canguilhem and the Problem of Error*. Palgrave Macmillan, 2019.

Tversky, Amos, and Daniel Kahneman. "Judgment under Uncertainty: Heuristics and Biases." *Science* 185, no. 4157 (1974): 1124–31. https://doi.org/10.1126/science.185.4157.1124.

Vapnik, Vladimir N. "An Overview of Statistical Learning Theory." *IEEE Transactions on Neural Networks* 10, no. 5 (1999): 988–99. https://doi.org/10.1109/72.788640.

———. *Statistical Learning Theory*. Wiley, 1998.

———. *The Nature of Statistical Learning Theory*. Springer, 1995.

Vapnik, Vladimir N., and Alexey Y. Chervonenkis. "On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities." Translated by B. Seckler. *Theory of Probability & Its Applications* 16, no. 2 (1971): 264–80. https://doi.org/10.1137/1116025.

Wald, Abraham. "Contributions to the Theory of Statistical Estimation and Testing Hypotheses." *The Annals of Mathematical Statistics* 10, no. 4 (1939): 299–326.

Wallis, W. Allen. "The Statistical Research Group, 1942–1945." *Journal of the American Statistical Association* 75, no. 370 (1980): 320–30. https://doi.org/10.1080/01621459.1980.10477469.

Werbos, Paul. "Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences." PhD diss., Harvard University, 1974.

Williamson, Robert C. "Loss Functions." In *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*, edited by Bernhard Schölkopf, Zhiyuan Luo, and Vladimir Vovk. Springer, 2013. https://doi.org/10.1007/978-3-642-41136-6_8.

Zhang, Chiyuan, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. "Understanding Deep Learning (Still) Requires Rethinking Generalization." *Communications of the ACM* 64, no. 3 (2021): 107–15. https://doi.org/10.1145/3446776.