A Hierarchy of Limitations in Machine Learning

Momin M. Malik

Berkman Klein Center for Internet & Society at Harvard University
momin_malik@cyber.harvard.edu

4 February 2020*

Abstract

"All models are wrong, but some are useful," wrote George E. P. Box (1979). Machine learning has focused on the *usefulness* of probability models for prediction in social systems, but is only now coming to grips with the ways in which these models are *wrong*—and the consequences of those shortcomings. This paper attempts a comprehensive, structured overview of the specific conceptual, procedural, and statistical limitations of models in machine learning when applied to society. Machine learning modelers themselves can use the described hierarchy to identify possible failure points and think through how to address them, and consumers of machine learning models can know what to question when confronted with the decision about if, where, and how to apply machine learning. The limitations go from commitments inherent in quantification itself, through to showing how unmodeled dependencies can lead to cross-validation being overly optimistic as a way of assessing model performance.

Introduction

There is little argument about whether or not machine learning models are *useful* for applying to social systems. But if we take seriously George Box's dictum, or indeed the even older one that "the map is not the territory' (Korzybski, 1933), then there has been comparatively less systematic attention paid within the field to how machine learning models are *wrong* (Selbst et al., 2019) and seeing possible harms in that light. By "wrong" I do not mean in terms of making misclassifications, or even fitting over the 'wrong' class of functions, but more fundamental mathematical/statistical assumptions, philosophical (in the sense used by Abbott, 1988) commitments about how we represent the world, and sociological processes of how models interact with target phenomena.

This paper takes a particular model of machine learning research or application: one that its creators and deployers think provides a reliable way of interacting with the social world (whether that is through understanding, or in making predictions) without any intent to cause harm (McQuillan, 2018) and, in fact, a desire to not cause harm and instead improve the world,[1] for example as most explicitly in the various "{Data [Science], Machine Learning, Artificial Intelligence} for [Social] Good" initiatives. It focuses on the almost entirely statistical modern version of machine learning, rather than eclipsed older visions (see section 3). While many of the limitations I discuss apply to the use of machine learning in any domain, I focus on applications to the social world in order to explore the domain where limitations are strongest and stickiest. I consider limitations in machine learning such that, contrary to the expectations and intentions of creators and deployers, machine

---

[1] I thank John Basl for encouraging me to make clear that I consider both methodological and ethical limitations.

learning can fail to be reliable and/or can cause harm. I do so in a systematic review, structured along four decisions that are implicitly made when deciding to use machine learning:

1. To use quantitative analysis over qualitative analysis;

2. To use probabilistic modeling over other mathematical modeling or simulation;

3. To use predictive modeling over explanatory modeling;

4. To rely on cross-validation to evaluate of model performance.[2]

This is conceptually illustrated in fig. (1). None of these decisions, leading up to the choice of using machine learning and cross-validation, are inevitable; it is possible to *not* use modeling, or machine learning, or any computer-based approach (Baumer and Silberman, 2011)—and so these are the points at which we should consider consequences and weight against alternatives.
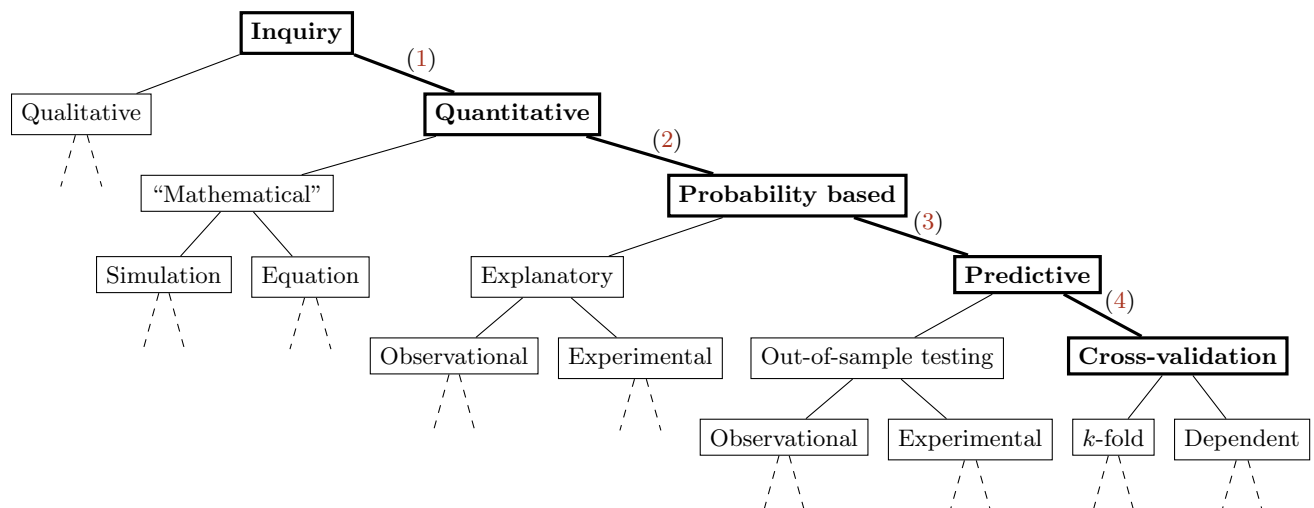

Figure 1: A diagram of the hierarchy explained in this paper, with each bold branch corresponding to a section.

On the four decision points, I argue:

1. When choosing quantitative analysis, we sacrifice the ability to have narrative understandings of meaning-making, commit to working with proxies rather than the actual constructs of interest, and risk delegitimizing lived experience.

2. When choosing probability-based modeling, we restrict our view to the world as consisting of fixed entities with properties, and prioritize the use of a central tendency over engaging with individuality.

3. When prioritizing prediction, we make models that are not reliable for getting insight into the underling system, and are sensitive to changes and reactions.

---

[2]Cross-validation is a method of model *selection*, but also a method of model *evaluation* (Hastie et al., 2009) by which to determine the validity of a machine learning claim. These two uses of cross-validation can have very different theoretical properties from each other (Wager, 2019); I am concerned exclusively with the latter. Two other ways of evaluating model performance that I will consider are experiments and true out-of-sample testing, but cross-validation is the dominant method, at least in research papers.

4. When relying on cross validation, we risk being overly optimistic about a model's generalizability.

The drawbacks of the four decision points form a hierarchy: the problems of prediction supersede all problems of cross-validation, the problems of the central tendency supersede all problems of prediction, and the problems of quantification supersede those of the central tendency. The hierarchy also be seen as a *necessary chain of custody*;[3] the results of cross-validation will only be as meaningful as the setup of a machine learning model, which will only be as meaningful as the way the world is put into observations and properties, which will only be as meaningful as way in which a phenomenon is quantified and measurement is done. Equivalently, limitations propagate through this hierarchy, with problems with quantification affecting everything downstream, and so on.

I treat the levels in sequence, although interesting questions remain to ask relating these levels to each other, such as what limitations there may be in hermeneutic/interpretive (rather than modeling, especially based on central tendency amidst variability) approaches to data (e.g., Gibbs and Owens, 2013),[4] or in dimension reduction techniques that see the need for qualitative interpretation of latent dimensions as a distinct strength (Desrosières, 2012).[5]

Guide to reading this work

The four sections of this work require progressively more technical background to follow, but they are also relatively self-contained. As a consequence, readers from different disciplinary backgrounds may refer only to specific sections covering topics of interest.

Machine learning readers most interested in my technical argument of how to theorize the effect of dependencies on cross-validation should go to section (4). Social scientists and policymakers considering whether and how to adopt machine learning (especially in place of social statistics) and the ethical implications of doing so should consider section (3); a background in statistical modeling (notions of estimation, inference, model specification) are necessary for later parts of this section, which get into limitations of causal inference and experiments, but earlier parts are more historical, sociological, and philosophical. Section (2) identifies some frequently overlooked but fundamental assumptions of statistics and machine learning, and considers their consequences, which useful for all audiences. Section (1) is a useful review for machine learning researchers about what qualitative research can do that machine learning cannot, and conversely, it can help qualitative researchers understand key gaps in machine learning and how qualitative methods can fill these.

Scope

There are four important aspects of the scope.

First, the limitations I identify are associated with using *only* machine learning, which can therefore be seen as more a cultural problem among people who do machine learning than intrinsic limitations attributable to machine learning itself. But such machine learning-only approaches exist: for example, computer scientist Hanna Wallach (2018) talks candidly about her own process of realizing that "machine learning is not a be-all and end-all solution." By using mixed methods, and including other approaches alongside machine learning for a given problem, many of these problems can be avoided.[6] That is: if machine learning is done with experimental validation via randomized

---

[3]I thank Beau Sievers for the term "logical chain of custody."

[4]I thank Elena Fernández Fernández for pointing me to this work.

[5]I thank Nick Seaver for pointing me to this work.

[6]I thank Berkman Klein Center fellows, and particularly Apryl Williams and Beau Sievers, for encouraging me to make this point explicit.

control trials, then we will not suffer from over-optimism about performance. If we accompany predictive modeling with causal/statistical analysis, we can gain a better understanding of underlying processes, and better hope to make predictions that are robust to changes and reactions. And if we include qualitative analysis, we can understand constructs, and engage with individuality, narrative understandings, and lived experience.

I do include several examples of mixed methods overcoming the limitations of machine learning alone (e.g., Patton et al., 2019a; Cardoso et al., 2016, respectively for validating lived experience and experimental validation). But mixed methods research is not something that can done casually: it requires collaboration between researchers not only with different toolsets but potentially with vastly different understandings of the world, or else it requires individuals with multiple training in methods that individually can take a lifetime to master. So far, there is little work on on how to systematically do mixed methods machine learning research. Such work is needed to both precisely articulate what the difficulties are, and then understand how to overcome them. This is beyond the scope of this work, although in the conclusion I will return to this point as the most promising possibility for how to proceed.

Second, I do not mean to imply that alternative approaches, on their own, are superior to machine learning on its own (Shapiro, 2014, writes: "Methods are like people: if you focus only on what they can't do, you will always be disappointed."). Each method has its own assumptions and resulting limitations, appropriate for different goals. The purpose of this paper is to highlight the specific limitations of *machine learning*, and point out what are available methodological alternatives worth considering for any point at which machine learning may not be fit for purpose.

So, if a researcher's goal is to understand meaning-making or identify underlying constructs, qualitative analysis is superior to machine learning, or indeed to any quantitative methodology. If the goal is to quantitively understand underlying data-generating processes (especially in the midst of "haphazard variability," and in a way that quantifies uncertainty; Cox, 1990), statistics is superior to machine learning. If the goal is strong guarantees about generalizability, then experimental design or another validation methodology is superior to the use of cross-validation. But for pre-testing the design of a system for detection/classification by a given schema (especially if developed by or with relevant stakeholders), representing an arguably enormous space of use cases neglected prior to the development of machine learning (Breiman, 2001; Mullainathan and Spiess, 2017), machine learning with cross-validation—for all its limitations—is very likely the appropriate approach.

Third, for the most part I deliberately avoid discussing consideration around the *data* that go into probabilistic models. The decision of what data to bring to a particular problem supersedes all other limitations and questions (Mallows, 1998), but data has also been covered extensively, most recently around social media data in comprehensive review by Olteanu et al. (2019). The problems that can lead analysis astray include demographic sampling bias, limited data access, filtering decisions (especially selection on the dependent variable), and behavior driven by norms or constraints rather than "naturalistic" preferences. But there has been far less about properties of the probabilistic *models* into which such data are fed, which this article seeks to address. Two aspects of data I do address, as they relate to what such models can do: what kinds of variability are captured in non-randomly sampled data, and the possibility of unobserved confounders.

Last, this hierarchy is also limited to systematic inquiry. There are levels even higher up not included here, such as ways of knowing (which need not be systematic) and ways of being (which need not be structured around the pursuit of knowledge), as per Linda Tuhiwai Smith (2012).[7] That is, being a researcher or a developer is itself a choice that need not have been made; to borrow the language of Colin Mallows (1998), that is the "zeroth decision" within this hierarchy.

---

[7]I thank Jonnie Penn for this point and reference.

Purpose and contribution

The paper draws together critiques from anthropology, sociology, science and technology studies, statistics, and machine learning itself, for the first time putting them into a comprehensive sequence. Then, the treatment of the effect of dependencies on biasing cross-validation using the idea of *optimism* (Efron, 2004) is a simple but novel extension of this statistical theory that has the promise of being an elegant unifying framework for multiple disconnected efforts on how to think through properly structuring data splits.

There are two intended audiences. First is machine learning modelers, to provide a comprehensive review of assumptions deeper than the mathematical ones made when doing modeling and intrinsic limitations induced by assumptions, as well as some examples of the assumptions breaking down to show concrete consequences. The second are audiences, consumers, and analysts of machine learning to better place the discipline among others, and to have a sense of contingencies they can consider when deciding if, where, and how to adopt machine learning. Critics in these audiences can hopefully use this to be *specific* about their objections to machine learning, identifying the points at which they disagree with assumptions.

Then, in addition to the purpose of serving the above two target audiences simultaneously (rather than make differentiated products), it is also to put social theory and statistical theory in conversation with each other. It agrees with Matteo Pasquinelli (2019), who writes, "the discussion about AI's limits may be inaccurate if technical limits are divorced from social limits, and vice versa," although as compared to that work (which addresses a non-expert audience and so spends time explaining core concepts and terms in machine learning), I take these for granted in addressing a machine learning audience and so can go far deeper into technical aspects. In methodological terms, this paper relates to "*in-situ* hybridization" (Mackenzie, 2017), treating different disciplinary forms of knowledge as coincident[8] in order to relate the internal understandings of modeling to the larger social contexts of both the models and the understandings thereof. Unlike *in-situ* hybridization, however, my approach is ultimately not an *analytical* one, but an attempt to intervene as an actor (Collins, 2008).[9] I advocate for specific interpretations and forgo interpretive flexibility, making this a work ultimately of practice and of a practitioner rather than of analysis. My work also relates to the call of Philip E. Agre (1997) for a *critical technical practice* to add critical reflection to the development of technical knowledge, although this paper (in itself) lacks a praxis or demonstrated application and so "critical technical theory" might be more apt. Then, to preserve accessibility for different audiences, I do not pursue the same extent of co-incidence as Adrian Mackenzie (2017), but divide the paper into modular sections over a *hierarchy* of priorities. Mathematical portions are confined to the final section, and should be of particular interest to machine learning modelers to consider how to systematically think about cross-validation.

The ideal effect of this paper paper would not simply be for those within machine learning to engage in methodological hedging—although being more humble about claim-making will be a positive outcome—or for those outside machine learning to retroactively justify an already-made decision to reject it. Even though mixed methods are not a part of this paper, by pointing out the limitations of machine learning as compared to other approaches, I hope to encourage cultural practices of putting machine learning in context with other possible approaches, and help develop systematic responses to limitations.[10] This paper can be seen as a follow-up to Selbst et al. (2019): that work was about general limitations of mathematical abstractions, and this one is specifically about the structural and social limitations of the abstractions used in machine learning. I intend

---

[8]I thank Nick Couldry for reminding me to make explicit the link to social theory.

[9]I thank Sheila Jasanoff for pointing me to this work.

[10]I thank Beau Sievers for this framing.

this to be a definitive work identifying and detailing these limitations that, going forward, can be the basis by which to plan out how to overcome these limitations and on which to structure efforts for doing so.

1 Quantitative versus qualitative analysis

Much has been written about how qualitative and quantitative methods are not mutually exclusive, and how they can complement and inform one another. This is certainly the case, although there are serious and deep philosophical conflicts (Bellotti, 2015; Erikson, 2013). But maintaining my focus on the branching paths that would lead to a machine learning-*only* approach, here I focus on four failure points of a quantitative-*only* approach. The first relates to the impossibility of quantifying meaning-making. The second is the difficulty of measurement in social science. The third relates to experience and personal knowledge. The fourth relates to how quantification can succeed by imposing its logic on the world, totally separate from any notion of correspondence or having empirical adequacy.[11]

1.1 Meaning

On the first point, qualitative researcher Michael Quinn Patton (2015)[12] gives an argument for the "nature, niche, value, and fruit of qualitative inquiry":

> "During the writing of this book, my first grandchild was born, and this book is dedicated to her. The hospital records document her weight, height, health, and *Apgar* score – activity (muscle tone), pulse, grimace (reflex response), appearance, and respiration. The mother's condition, length of labor, time of birth, and hospital stay are all documented... But nowhere in the hospital records will you find anything about what the birth of Calla Quinn *means*. Her name is recorded but not why it was chosen by her parents and what it means to them. Her existence is documented but not what she means to our family, what decision-making process led up to her birth, the experience and meaning of the pregnancy, the family experience of the birth process, and the familial, social, cultural, political, and economic context that is essential to understanding what her birth means to family and friends in this time and place. A qualitative case study of Calla's birth would capture and interpret the story and meaning of her entry into the world from the perspectives of those involved in and touched by her coming into our lives."

What Patton describes here is "thick description": including the specificity and larger context, and getting at *meanings* behind behavior and expressions. Theodore Porter (2012) links the converse, "thin description," to the kind of descriptions provided by quantification and modeling: "In a thin world, surfaces should be valid and deeper meanings superfluous."[13] He sees anthropologist Clifford Geertz's classic championing of "thick description" as "a battle cry against an idea of social science

---

[11]I am treating qualitative analysis here as something more systematic and specific than qualitative *reasoning*; qualitative reasoning is universal, pervading every step of data analysis, such as picking which possible directions of data analysis to follow in what Andrew Gelman and Loken (2014) call the "garden of forking paths." And, Donald Campbell (1975) points out that scientific results are uninterpretable without a "narrative history" portion that situates a work and the results.

[12]I thank Maya Randolph for pointing me to this work.

[13]Here, in modeling terms, I would not take "surface" to be synonymous with *observables*: while latent variables can be a way to express deeper meanings than what is directly observed, they might also only go only one layer deeper than the surface and so not be much less superficial.

harmonized, if not unified, by a shared commitment to the external and observable."[14] Quantitative inquiry requires standardization and generalization; like any abstraction, this can succeed at its aims (Stinchcombe, 2001) and be seductive in its power (Selbst et al., 2019), in what Porter (2012) calls the "siren song of thinness," where "the sacrifice of human meaning seem[s]... a price worth paying for solid results."

In referencing what his granddaughter's birth *means*, Patton's quote also connects with the concept of *meaning-making*. This concept, originating from psychiatrist and holocaust survivor Viktor Frankl and also connecting to the sociological tradition of symbolic interactionism, is about the linguistic categories through which people make up their reality and define, justify, and interpret their actions. These are theorized as the most fundamental aspect of human social settings (Krauss, 2005). Choosing quantitative analysis gives solidity but precludes the study of meaning-making—even as quantification relies on existing meanings.

This is one standard by which qualitative inquiry is superior to quantitative: quantitative inquiry is *unable to account for itself*. Quantitative analysis is also at the mercy of the quantification process; all exploration or discovery is strictly limited to variables that have been measured, whereas qualitative analysis has no such constraints (Campbell, 1975).

While, again, I am not going into the drawbacks of alternatives to the path that leads to machine learning, here it is worth noting that the quote above points to some limitations of qualitative analysis: as Patton (2015) writes, the "physiological and institutional metrics" around birthing "provide trend data about the beginning of life... when aggregated across many babies and mothers." If our goal is such trend data, then measurements such as the *Agpar* score are appropriate, while the meaning of Patton's granddaughter Calla Quinn's birth is not. In terms of drawbacks, while the "subjective" and contingent nature of qualitative inquiry should not be seen as a drawback since the studied phenomena are themselves contingent, the usual objections of a lack of scalability and generalizability apply.

Furthermore, qualitative inquiry is not necessarily more ethical. In practice, qualitative inquiry has developed strong ethical frameworks (Costanza-Chock et al., 2018), but this has been in response to both historic and contemporary failures to commit to principles of reciprocity and sharing knowledge that Linda Tuhiwai Smith (2012) argues are necessary for non-exploitative research. After all, exposing the inner meanings that people hold can be more invasive than any surveillance, and indeed, ethnography and "thick" descriptions have both historically played and continue today to play a role in furthering oppression and domination (Smith, 2012; Price, 2011).[15] In another critique, medical anthropologists Kleinman and Kleinman (1995) write how ethnographers frequently treat human suffering with detached analysis which is "every bit as dehumanizing as that of [medical] colleagues who unreflectively draw upon the tropes of tropical medicine or behaviorism to create their subject matter."

Similarly, the principle of reflexivity—self-awareness by researchers about both their effect on the research, and the effect of the research on them (Attia and Edge, 2017), which is not just a subjective process but an *intersubjective* one that considers relationships and interconnectedness between different people's subjectivities (Cunliffe, 2016)[16]—does not guarantee responsible research. One can be perfectly aware of being an exploiter. Nor is it impossible to be reflexive in quantitative research. However, it is considerably harder when working with layers of mathematical forms that require years of specialized training to understand and internalize as natural and intuitive, abstractions not shared by the potential subjects of study. Perhaps this distance makes the need for reciprocity even more pressing: Dan McQuillan (2018) suggests that, because "Machine learning

---

[14]I thank Rodrigo Ochigame for pointing me to this piece.

[15]Again, I thank Rodrigo Ochigame for pushing me to this point.

[16]I thank Maya Randolph for this reference.

represents one of the highest historical forms of the abstraction of social relationships," it "needs to be counterbalanced by the unmediated relationships of popular assemblies."

1.2 Measurement

Second is the question of measurement. Qualitative research is able to get more directly at processes of interest, engaging directly with how they are multifaceted, variable, and context-dependent. But when moving to quantification, what is measurable is almost never what is of interest in social systems (DeVellis, 2017); in 1939, John Maynard Keynes (1939) expressed skepticism towards mathematical modeling in early econometrics partly based on the difficulty of measuring "political, social and psychological factors, including such things as government policy, the progress of invention and the state of expectation." John Stuart Mill had similar concerns even earlier, in 1871, about mathematics in economics, which Peter Spiegler (2015) summarizes as the worry that "mathematical language might generate a purely quantitative conceptual map of the subject matter it purported to outline, with no way of telling whether the outlines on the map corresponded to the subject's own contours."

No quantification captures everything; like modeling more generally, a measurement is a reduction that solidifies some meanings while excluding others. Yarden Katz (2017) explains how machine learning (and the "connectionist" tradition in artificial intelligence) echoes the old behavioralist paradigm in a focus "on an input-output relationship that's learned by tuning the model using error signals in data, without worrying about any internal states that govern that relationship." This prioritization is a bargain that machine learning makes; it has achieved successes far beyond those of previous behavioralist research, but this section is about where, how, and for what goals this bargain can fail.

Psychometrics, in particular, has developed extensive theory around the idea of an underlying *construct*, which is the entity of actual interest, versus the proxies used to measure it (although these efforts are not without serious problems, e.g., Michell, 1999). "Constructs" are things like well-being (Alexandrova, 2005), user engagement, creditworthiness, or 'intelligence,' which we attempt to measure, respectively, with proxies like self-report, clickthrough rates, correlates of loan repayments, and standardized test responses.[17] Even "crime" is a construct, insofar as (even given a particular legal code *defining* crime) there are no measures of crime: there are only data of arrest records, or of crime reports, or of incarceration, none of which is one-to-one with actual crime (Ochigame et al., 2018; Elliott, 1995). Pierce (2008) analogizes proxies to asking, "Can we really weigh an iceberg by measuring its tip?" If quantitative research does not deal with the problems of proxies, it is akin to thinking that the tip of the iceberg is its entirety. And, following the analogy, predicting to what is available is akin to predicting to the tip of the iceberg. Sometimes this will be sufficient, but many times, it will not be.

Less metaphorically, constructs (the "inner states") can be formulated in a latent variable model (DeVellis, 2017); that is, not every latent variable represents a construct, but constructs can be expressed as latent variables. If, conceptually, the typical supervised learning task looks like fig. (2), then the place of underlying construct can be illustrated as in fig. (3). Unlike "ground truth," which is something measurable, a construct (which is closer to an actual underlying "truth") is something latent and potentially unmeasurable but that produces observable measures (with some

---

[17]This is not to say that we should treat every abstract social idea as a construct; we could distinguish constructs from, say, values. For example, for the World Press Freedom Index, we could say that "press freedom" is a construct which we can measure by proxy. But it would be more precise to identify the construct as a normative standard of "how professional journalists should act and exist in society," and the goal to be to measure the obstacles to them meeting this standard. I thank Joshua Kroll for a critique of treating values as constructs, from where I got this point.

noise, and/or with other complicating inputs). We can potentially use this to estimate latent values.
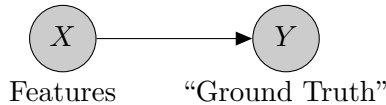

Figure 2: A graphical model (Pearl, 2009) that gives the conceptual setup of a typical supervised learning task. Nodes in gray are observed, and nodes in white (not present here) are unobserved, and arrows show causal directionality. Generally, we imagine that the features, X, "produce" or cause the "ground truth" label, Y, but the causal direction is not important for a machine learning application; for prediction (see section 3), the only relevant factor is that the two are correlated. This can happen identically whether it is X that causes Y, Y that causes X (as in a detection task), or an unobserved common cause that causes both X and Y.
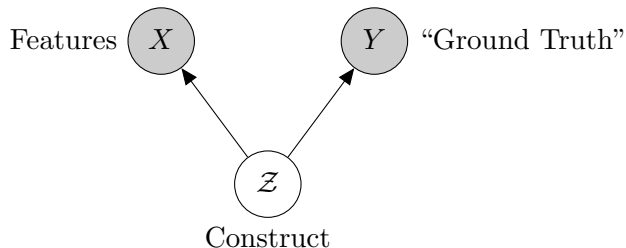

Figure 3: A graphical model showing the place of a construct in the supervised learning task illustrated in fig. (2). There is an unobserved, potentially unmeasurable, underlying construct $\mathcal{Z}$ that produces both the labels Y and the associated features X, such that X and Y are correlated.

To take a more concrete example, imagine a supervised learning task to recognize an image as being of a cat or not being of a cat. Here, the underlying construct is something we might call "cat-ness" (fig. 4), which is not directly measured; what is available is the *human labeling of images* as being of cats or not of cats.
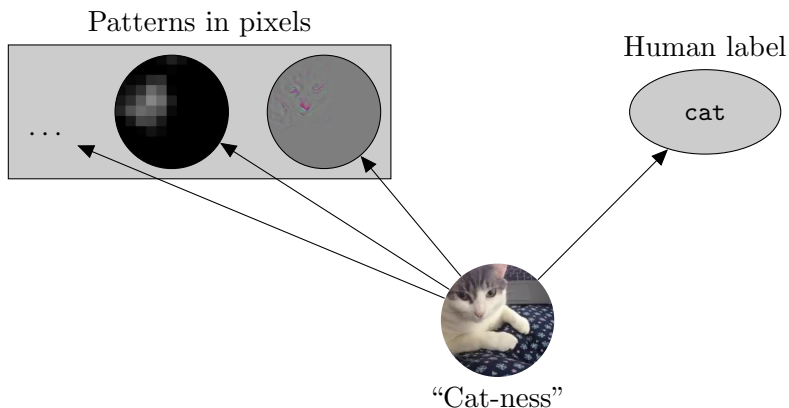

Figure 4: The way in which a construct underlies the specific task of image recognition. Images from Yosinski et al. (2015), used under a CC BY-NC-SA 3.0 license.

That is, what is used as "ground truth" is not cat-ness itself; rather, cat-ness both produces

both the human labels, and the groupings of pixels (which is why the two are correlated). As far as we currently know, the specific "patterns of pixels" identified by neural nets are quite far from how human perception works, even if there are overlaps (Zhou and Firestone, 2019). But also, the construct cat-ness is not some Platonic form; we have to decide what we want a notion of underlying cat-ness to accomplish. Depending on our goal, even human perception is not a foolproof proxy. Perhaps we might want to identify cat-ness with phylogeny, of being of the family Felidae; in such a case, the left image in fig. (5) lacks cat-ness, since it is Charles R. Knight's 1904 illustration of Dinictis (a genus of the extinct family Nimravidae, known as "false saber-tooth cats") and not a cat. Alternatively, we might want to define the cat-ness of an image as whether or not people perceive a cat; in that case, the image on the left of fig. (5) would possess cat-ness, whereas the image of the right of fig. (5) would lack it.[18]
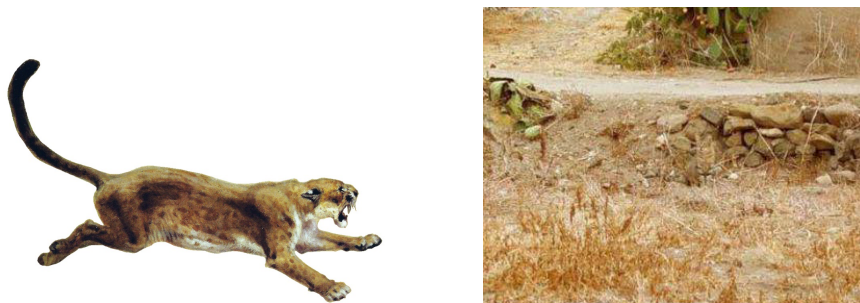

Figure 5: Left: not an image of a cat, but potentially possessing cat-ness. Right: an image of a cat, but potentially lacking cat-ness.

As we can imagine from this example, it is an extremely difficult undertaking to decide how to set up the hypothesized latent structure in order to estimate or even quantify latent values, decide what observable measures to use for which hypothesized latent entities, and decide how to determine success of inference to something unobservable.

My focus here is on limitations of not *accounting* for constructs; for a comprehensive overview of how machine learning can use ideas of constructs and validation, I point readers to Jacobs and Wallach (2019). But briefly, psychometrics validates proposed models relating measurements to constructs multiple ways. For example:

- External validity is whether associations seen in the study generalize beyond it; it is synonymous with generalizability; I discuss it further in section (3) as the sole focus of machine learning.

- Criterion-related validity[19] is when a measure should predict some auxiliary outcome that the construct is hypothesized to cause (e.g., a measure of well-being should, say, correlate with health indicators, preferably before those indicators are manifested if there is a theoretical model that well-being causes health).

- Internal validity is whether conclusions about causal relationships within the study generalize.

- Content validity relates to whether the measurement can cover the total possible range of the construct (or effectively sample within it).

---

[18]We should, however, be careful to specify who we mean by "people," as people who are primed with the information that there is a camouflaged cat in the picture (Rosenfeld et al., 2018) and who are determined enough will find the cat.

[19]This is also called "predictive" validity, but I uses criterion-related to distinguish it from "prediction" in the sense of machine learning, which is about external validity.

- Face validity is if something makes sense based on known relationships. So, for example, a quantitative model that can show that patterns in tea leaves in a cup of hot water predict real-world events would lack face validity, even if it succeeds empirically.

As uncertain as this process, is, *any quantitative research that ignores it can only make pronouncements about what is directly measurable.* Or, when the goal is not study but simply to construct systems, "concept drift" (Schlimmer and Granger, 1986) may be understood as a consequence of using what is measurable rather than identifying the underlying construct. The "hidden context" is the actual target, rather than a sort of complicating factor to be managed (Widmer and Kubat, 1996) using measurables.

For example, if not considering issues of measurement, studies of online social networks are only able to say something about ties on those networks and not of the construct of 'friendship' that those ties supposedly capture. Or, if there are untrained annotators applying their own heuristic understandings to produce labels, the results may be meaningless and useless (Patton et al., 2019a) regardless of how how high the "accuracy" is of a model to recover those annotations.

This also relates to what David Hackett Fischer (1970) names as "quantitative fallacy": relying on what is measurable to the exclusion of all else. This was also later named the "McNamara fallacy" after the disastrous reliance of the US Secretary of Defense during the Vietnam War, Robert McNamara, solely on reported measures of enemy body counts as his measure of success (O'Mahony, 2017) without even any validation. This gave incentives to those reporting to him to simply lie about the numbers (see also Sec. 3.2.2). A recent result by Obermeyer et al. (2019) can be interpreted in terms of constructs: they detail a machine learning model used by a major healthcare company that de-prioritized care for black patients versus equivalently sick non-black patients. This was because model optimized to healthcare *costs*, which are easy to measure, but the actual target of interest is the hard to define and measure construct of *illness*. Costs are not a solely a proxy solely of illness; they also capture structural racism, as less money is spent caring for Black patients. Consequently, the use of what is available and measurable led to a system that formalizes discrimination against Black patients.

Arguably, machine learning can defer responsibility for any problems with quantification to whatever is given to it as the "ground truth." That is often fair as a division of labor. But many times in practice, when machine learning makes claims, it forgets to recognize and communicate that the measurements serving as "ground truth" are a black box that hide problems of measurement, and that the ground truth is not construct itself. This is especially true when, as documented by Geiger et al. (2020) for social computing, machine learning researchers frequently report doing the labeling themselves, or else do not even report where labels come from. At other times (e.g., when a machine learning researcher is working with haphazardly collected or legacy data, or when the data collectors are themselves irresponsible about measurement, constructs, and validation), there may be nobody to defer the responsibility to; in such cases, if a machine learning researcher wants their research to be meaningful and effective, they must themselves take up responsibility of the quantification process. Or, at the very least, they should maintain some awareness of the contingencies and uncertainties of quantification and measurement, and communicate appropriate caveats.

Measurement can also be mutli-layered, with high-level problems compounding on underlying ones. Consider Google Ngrams.[20] First, those *n*-grams are built from a measurement technique: scans run through optical character recognition (OCR). For over a hundred years' worth of data in the corpus, lowercase 's' was typeset in a way that made OCR classify it as 'f' (Zhang, 2015). Second, *n*-grams are a frequency measurement, but as Pechenick et al. (2015) argue, these frequencies do not

---

[20]I thank Jürgen Pfeffer for this example.

work as a measure of popularity: the Google Books corpus contains one copy of each book, which equally weights blockbuster hits along with books that never sold a single copy. One of the original applications of n-grams was as a measure of censorship in Nazi Germany; but Koplenig (2015) documents how the lack of metadata make it questionable whether the results reflect censorship or are even linked to World War II. Such measurement problems are over and above problems coming from an unclear sampling frame (Pechenick et al., 2015). Consequently, Koplenig (2015) recommends we "explicitly restrict the results to linguistic or cultural change 'as it is represented in the Google Ngram data';" but such a total lack of generalizability was hardly the intention of the corpus.

1.3 Experience and legitimacy

One last consequence of quantitative analysis is that it centralizes power in the analyst. This is not intrinsically bad, but can enable oppressive outcomes. And, this can happen in qualitative analysis as well, but for those methodologies there exist methods of co-creation, and specifically, participatory action research (Costanza-Chock et al., 2018) that seek to avoid this.

With quantification locking in one possible set of meanings, a consequence is the risk of delegitimizing the experiences of individuals. For example, why should quantitative evidence (e.g., Gelman et al., 2007), rather than people's lived experience, be the arbiter of whether or not policing is biased? In a classic example of Simpson's paradox—graduate admissions at UC Berkeley—it was determined that, after controlling for departments, there was no evidence of bias towards men (Bickel et al., 1975). But the authors of this analysis recognized that this required a specific assumption, that men and women had equal academic qualifications, and that the real process of interest, discrimination, is the *unobservable* "exercise of decision influenced by the sex of the applicant when that is immaterial to the qualifications for entry." Even if internal decision-making of admissions officers could be observed, as Deborah Hellman (2008) also points out, the criteria for defining "merit" are themselves subjective and endlessly contingent (how would we compare one person to another who worked twice as hard to achieve, say, just less than half as much?) and could be discriminatory. Contrast this to a possible qualitative approach, not considered by the authors, which would be to examine the *experience* of women in the graduate programs and their differential treatment, such as in a study from shortly after by Sandler and Hall (1986).

Structures of power may commit to using quantitative evidence as a neutral ground. Then, just as Gayatri Spivak (1985) theorizes about "strategic essentialism"—the adoption of essentializing identities by marginalized peoples as a tactic for gaining recognition—we can discuss "strategic quantification," or "strategic modeling."[21]

Quantitative evidence may also be a tactic for majority groups to combat their own prejudice (Wise, 2019). And, the Clark Doll Experiments of the 1940s were cited as evidence about the psychological harms of segregation in the Brown v. Board of Education case, in a famous example of quantitative analysis successfully contributing to real change (Tomes, 2004). However, this is perhaps an exception. Ruha Benjamin (2019) notes that there is a long history (Wells-Barnett, 1895; Du Bois, 1899, 1900; Battle-Baptiste and Rusert, 2018) of making suffering and injustice quantitatively visible, but such efforts have often proved ineffective. She cites some experimental evidence suggesting efforts can even be counterproductive, and herself advocates *narrative tools* as the more effective strategy for change.

When not chosen strategically but instead imposed, insisting that people from marginalized groups quantitatively prove that their experiences—specially when this demand is made only to

---

[21]I thank Os Keyes for making this connection between strategic essentialism and data/modeling.

members of those groups—is unjust (Lanius, 2015). And effort spent towards doing this also falls under a critique by Toni Morrison (1975) pointing out the perversity and fecundity of quantitative (and other) standards:[22]

> "The function, the very serious function of racism, is distraction. It keeps you from doing your work. It keeps you explaining over and over again, your reason for being. Somebody says you have no language and so you spend 20 years proving that you do. Somebody says your head isn't shaped properly so you have scientists working on the fact that it is. Somebody says that you have no art so you dredge that up. Somebody says that you have no kingdoms and so you dredge that up. None of that is necessary. There will always be one more thing."

Indeed, when model outcomes do not benefit structures of power, those structures can, and do, called the modeling into question. One example is a reply from advocates of the auto lending industry to a report from the Consumer Financial Protection Bureau (CFPB) statistically arguing for existence of discrimination by indirect auto lenders, and citing those lenders for fines (Witkowski, 2015). The reply calls the CFPB methodology (specifically, inferring unobserved traits using proxies) "essentially sophisticated guesses": but so much modeling, including the modeling used by the cited industry to benefit itself, could be called the same.

Harry Collins (1981) introduced the idea of "experimenter's regress" to discuss how the outcome of experiments depend on proper apparatus and experimenter competence, such that it is always possible to challenge experimental results (MacKenzie, 1989). It is similar to the earlier idea of *confirmation holism*, where all claims rely on supporting claims that can also be challenged (and so multiple claims have to be accepted simultaneously even if they are logically sequential), and has also been extended to the idea of theoretician's regress (Kennefick, 2000).[23] We can similarly extend this to choices around probabilistic modeling and propose "modeler's regress"—which is how the limitations and uncertainties of modeling can always be used to strategically dispute the legitimacy of a particular model when its conclusions are unfavorable. R. A. Fisher himself, an avid smoker and paid consultant for the tobacco industry, ended his life trading in his authority to vigorously argue against the emerging consensus about the causal link between smoking and lung cancer (Stolley, 1991). The essential variability of the modeling process itself is on display in the amazing study of Silberzahn et al. (2018), where twenty-nine teams of statistical analysts given the same data set and the same target quantity to estimate, and not a single team used the same statistical model! Reassuringly, the confidence intervals of most of the studies overlapped, giving an overall robust result, but this shows how enormously variable the modeling process is and how much room this suggests there is for making challenges.

What are ways to mitigate the limitations of quantitative-only research? One is to center narrative, rather than quantitative tools, when trying to carry out advocacy. Another is to make qualitative inquiry an integral part of any rigorous research project. For example, developing labeling schemes is an area where qualitative researchers have extensive training and experience and which can entirely change the results of a machine learning classification task. Frey et al. (2018) and Patton et al. (2019a,b) argue for this; but these works also embody something even more important, which is for qualitative research can frame overall projects. That way, qualitative research can also determine the meaningfulness of final quantitative results to the actual systems. For quantitative-only research or quantitative components, any claims about unmeasurable constructs based only on analysis of measurable proxies should recognize the uncertainty (whether quantitatively, such as in terms of measurement error, or otherwise) that comes from this distance.

---

[22]Maya Randolph connected this statement to efforts quantitatively, statistically "prove" racism.
[23]I thank Nick Seaver for this connection and the extension.

1.4 Performativity

A pragmatic objection is to say: so long as we can reliably anticipate system outputs, do the limitations of the sorts outlined above really matter? The response is to emphasize again that it may be perfectly possible to build a model that anticipates with high accuracy, precision, and recall how an ignorant annotator would label a given instance, but such modeling will not provide meaningful statement about the system or underlying constructs.

But further, what if we use a model outputs to successfully take action? Would this not show that the limitations do not seriously affect modeling? In response, we must consider the *performativity* critique: that models do not work because they are "true," but because they (or rather, their creators and the loci of power that deploy modeling) *impose the logic of the models on the world.*[24] This critique was originally developed around economic models, where it is easy to see how, once people begin optimizing to something like the Black-Scholes model, it becomes true regardless of whether or not it successfully described the system beforehand. But it applies much more widely (Healy, 2015), including open opportunities to apply to machine learning—even if, unlike the strong claims of economics (Syll, 2018), machine learning is generally agnostic about mechanisms.

While there are clear-cut empirical cases of a model working by transforming the world, rather than necessarily by reflecting it (Malik and Pfeffer, 2016; Malik, 2018), a strong form of the performativity critique is to say that models *only* work by transforming the world. An intermediate form of performativity is that people react to being modeled, creating nonstationary environments and leading to concept drift.[25] And a light form of performativity is to note that self-driving cars have not exactly been successful in navigating an open-ended social and natural environment: rather, they require both the social and natural aspects of the environment to be manipulated to succeed. The ultimate success of integrating fully autonomous cars alongside human drivers will likely require training human drivers as well. Thus, even if not committing to the strongest form, performativity can be a powerful analytic frame for generating hypotheses about sources of modeling success that can be empirically investigated (including through modeling itself).

2 Probability-based modeling

When it comes to social systems, the probability-based modeling of statistics and machine learning is far more popular that other approaches. It has two attractive features—namely, being able to account for variability and being able to make use of data, properties that competitors do not have. However, it requires making commitments in terms how it structures the world, and how it reduces data to usable information.

---

[24]More precisely, Kieran Healy (2015) summarizes: "Originally articulated by Callon (1998) and refined by MacKenzie and Millo (2003), and MacKenzie (2006), the *performativity thesis* is that economics produces a body of formal models and transportable techniques that, when carried out into the world by its professionals and popularizers, reformats and reorganizes the phenomena the models purport to describe... The success of economics is not just a matter of a particular conception of rationality serving as a ceremonial gloss on social action; nor is it a simple instance of ideological indoctrination... Rather, tools implementing formal models of action—'calculative devices'—are put in the hands of social agents by the model-builders or their representatives. These devices act as 'cognitive prostheses' that enable actors to accomplish calculative tasks previously beyond their reach, but which are required by the theoretical models. When incorporated into the everyday work of market agents, these devices allow real settings to better approximate the original models, and their assumptions." I thank Abby Jacobs for pointing me to this work.

[25]Thanks to Joshua Simons for suggesting the connection of concept drift to performativity. Connecting to the earlier mention of concept drift around constructs, here we would say that there is some underling, stationary but probably unmeasurable, construct consisting of how people react to certain incentives.

2.1 The data matrix

In a critique of regression modeling in social science, Andrew Abbott (1988) identified how such modeling assumes the world can be divided into entities (the observations, or instances), and properties of those entities (the variables, or features). He continues,

> "...it is striking how absolutely these assumptions contradict those of the major theoretical traditions of sociology. Symbolic interactionism rejects the assumption of fixed entities and makes the meaning of a given occurrence depend on its location – within an interaction, within an actor's biography, within a sequence of events. Both the Marxian and Weberian traditions deny explicitly that a given property of a social actor has one and only one set of causal implications. Marx's dialectical causality makes events produce an opposite as well as a direct outcome, while Weber and the various hermeneutic schools treat attributes as infinitely nuanced and ambiguous. Marx, Weber, and work deriving from them in historical sociology all approach social causality in terms of stories, rather than in terms of variable attributes."

That is, the very formulation of the world into a data matrix is neither inevitable nor natural. Analytically, it limits what we can imagine and describe, and in terms of power and control, Mark Poster (1995) wrote about how databases fragment people by "grids of specification," both multiplying and decentering them, which allows them to be recalled and reassembled for the purposes of the database owners and users.[26] In inheriting this formulation of the world, machine learning creates the same sorts of reconfigurations of subjects (Mackenzie, 2017).

Abbott notes that some of these assumptions can be relaxed—for example, a time series allows an entity to change over time—but this does not substantially change the formulation of the world (at a given point in time, an entity is still fixed, and there is still no notion of multiple interpretations).

As an example of this formulation leading to conceptual problems downstream in statistics and econometrics (and, in this case, in applications of those in law as well), Issa Kohler-Hausmann (2019) critiques the "counterfactual" view of race. This view defines discrimination by asking if a person would have been treated if everything was the same (i.e., *ceteris paribus*) *except* for their race. From a sociologically rigorous point of view, the idea that one could hold everything but somebody's race constant is incoherent: aspects of people are so deeply tied into how their racial identity is constructed, and the "effects" of race so suffused into so many other variables, that a hypothetical "race-switched" world is meaningless (also, within a Rubin causal framework, race is immutable and not manipulable hence cannot have a causal effect). We would need to "backtrack" and revise all aspects of the person to make the counterfactual meaningful, but then there would be no pure effect to estimate.[27]

Lily Hu (2019a) extends this to machine learning, noting that auditing for fairness based on posing counterfactual in terms of model inputs suffers the same conceptual incoherence. She argues that this framing reduces discrimination to irrationality (i.e., discrimination relies on "irrelevant" characteristics in decision-making), egregiously neglecting the ways in which racism is very rational for maintaining power structures that bring benefits to certain groups. This problem can somewhat be alleviated by recognizing the "right" kind of dependencies between race and other variables, and structuring graphical models along those terms by which analysis is done, but so long as the

---

[26] This reference is from Raval (2019).

[27] One solution is to change the causal effect from race itself to perceptions of race; while clever, such a shift does not account for differences in life trajectories. That is, "audit studies do not measure the objective-isolated treatment effect of race and race alone because there is no such thing to measure" (Kohler-Hausmann, 2019). Actual engagement with discrimination would come through traditions of antisubordination and cultural reconstruction.

model allows for statements to be made based on a generative case of two people who are alike in all aspects but race, the fundamental problem remains (Hu, 2019b).

This ties also into a larger problem identified by Philip Agre (1997) of technical disciplines trying to read social science and humanities literature as descriptions of mechanisms (or potential model formulations). Doing so fails to understand their content and, in the above case for race, can even be inconsistent with it.

2.2 Central tendency

The central mechanism of all statistics and machine learning is the use of a central tendency (justified by reference to probability distributions, and the "sufficient statistics" of a given distribution). Both statistics and machine learning are fundamentally unable to do anything with one observation of one entity; there either have to be multiple observations of that entity, and/or observations of multiple entities (along with an assumption or definition of what constitutes an "entity," and which such entities are comparable in a distribution of some property). While there are interpretations of probability statements as applying to individuals rather than to groups (Dawid, 2017),[28] estimates of probability are still only able to be calculated from multiple observations.

Predictions are of the central tendency; for continuous variables, this would be the mean or median, or for categorical variables, this would be the majority class. The central tendency might be of a subset of the range of values (e.q., quantile regression, or mixture models), but is still a central tendency within that region.[29] Outlier-based tasks like anomaly detection that identify individual observations still do so by reference to a central tendency. And estimators for supervised learning tasks or used for statistical inference are formed from the *conditional* central tendencies at various levels of features/covariates. Despite promises of "personalized" medicine or "individualized" risk assessments,[30] treatment or scores are not really done by the individual, but by making finer and finer bins (which is possible given more and more data).[31] There are questions of justice in applying inferences about a group to any individual member of that group (Faigman et al., 2014); and at the point where the bin contains only one person, statistics and machine learning are incapable of doing anything but restating the data. Similarly, Katz (2017) talks about the limits of image recognition, selecting famous historical photos (e.g., around segregation in the US) to make a point about how image recognition not being able to understand the context of those photos: "The presumption is that mappings from images to labels are sufficient; that the 'information' is there, and it's only a matter of finding the right model to decode it," but one-off information is not what "Statistical pattern recognizers are suited for." Alternatively, correlates of one-off contextual information are simply not present in the loss landscape of the model, so no optimization approach over a feature space derived from pixels could ever recover them.[32]

Some sort of central tendency will invariably be the minimizer of a loss function, so thinking in terms of loss functions *aggregated across multiple observations* leads inevitably to central tendency.

---

[28]I thank Joshua Simons for pointing me to this reference.

[29]The canonical parameters of a uniform distribution could be characterized as an "extremal" tendency, but such a distribution could also be described its mean and radius.

[30]Specifically, Philip Dawid (2017) says that the foundational philosophical question of of "individual*ized* risk" is a notion of "individual risk." For an individual, Sam, a notion of indiviudal risk requires, for a Frequentist interpretation of probability, the assumption that "the chosen attributes capture 'all relevant characteristics' of the individuals." A Personalist interpretation of probability (a variety of a Bayesian interpretation) would require an assumption of "no relevant additional information about Sam (or any of the other individuals in the data), and can properly assume exchangeability—conditional on the limited information that is being taken into account." Dawid notes, "Neither of these requirements is fully realistic." Again, thanks to Joshua Simons for this reference.

[31]This phrasing came Elena Esposito, in a conversation. My thanks to her for this.

[32]I thank Beau Sievers for encouraging me to think about the loss landscape.

But such aggregate loss is a commitment to efficiency, and of a certain notion of utilitarianism (where the utility to be maximized is something measureable). It, too, is neither natural nor inevitable, and perhaps is quite unnatural: when the idea of characterizing a population by its central tendency first arose, the idea was met with great skepticism and opposition (Donnelly, 2016; Hacking, 1990), and represented a large shift in basic ideas about the world (Daston, 1989).

Even if unnatural, central tendency was a highly effective solution to a problem of too much data. In the time of R. A. Fisher (1922), data sets (in the dozens of observations) were already too massive to interpret manually. He wrote,

> "briefly, and in its most concrete form, the object of statistical methods is the reduction of data. A quantity of data, which usually by its mere bulk is incapable of entering the mind, is to be replaced by relatively few quantities which shall adequately represent the whole, or which, in other words, shall contain as much as possible, ideally the whole, of the relevant information contained in the original data."

But extending the notion of minimizing loss, or to summarize information via a central tendency, from experimental design contexts to tasks of social management and *governance* was a slow process for which certain commercial interests had to put enormous effort to succeed (Bouk, 2015).

Using the central tendency to determine outcomes has consequences: namely, it invariably punishes outliers (Keyes, 2018; Costanza-Chock, 2018). Is it possible to treat people as individuals at scale? Todd Rose (2015) gives an example at least of an amelioration: he describes the US Air Force designing plane seats around the "average pilot," but then discovering that these averages did not fit a single pilot (and the design leading to deaths). Instead, *adjustable seats* were developed as a way to account for variability.

Incorporating variability may often improve over using the mean or median, but is not something fundamentally different. Variance—defined as the second central moment of a probability density/mass function—is itself a central tendency, and therefore still describing something about a population rather than an individual.[33]

The problem may fundamentally be one of scale: perhaps scale inevitably requires standardization, and standardization is most effectively achieved through managing via central tendencies. Perhaps the only way to treat people as individuals requires giving up scale and centralization, and allowing systems and analysis to be local. This, however, is a very different picture of civilization.

2.3   Alternative forms of modeling

Other forms of quantitative modeling may avoid these limitations, although they come with their own problems. There is a style of modeling of social systems, found in "social physics," microeconomics, and mathematical sociology, that has been called "mathematical" (Kolaczyk and Csárdi, 2014). It consists of using linear algebra and/or calculus to express core assumptions (e.g., as equations) and then making derivations from them to reach conclusions. They potentially also use distributions to get a range of results, or use simulations to get plots of functional forms that cannot be calculated analytically, but in all cases, the focus is on the mathematical expression of processes and relationships. But this style also divides the world into entities and properties of those entities—just as mechanics, for example, has objects that have have properties like mass or

---

[33]We might try to address problems with variability by estimating skewness (the third central moment), and problems with skew may be solved by estimating kurtosis (the fourth central moment), and so on; but perhaps the only way to account for all ways in which individuals are different is to have as many moments as observations—in which case we have done nothing but restate the data. I thank Beau Sievers for a discussion that led to this point.

momentum.[34] And, as compared with statistics, the arguments of this literature are not necessarily applicable to understanding any specific real-world system (Borgatti et al., 2009). Indeed, it is a whole separate question of whether a given structural property and the kind of generating processes implied by it—such as "power law distributions"—could even apply in a specific case, or indeed ever applies to meaningful processes in the social world (Clauset et al., 2009).

Simulation modeling, especially as realized in agent-based modeling or agent-based social simulation, is a distinct alternative to the formulations of probabilistic modeling (note that simulations use statistics, and statistics uses simulation, but as forms of *modeling* they have very different goals and assumptions; Gilbert and Troitzsch, 2005). Here, modeling is of interactions between software agents, offering an alternative to relationships between variables (Macy and Willer, 2002), whether done through statistics or equations. It is useful for proposing causal mechanisms, and can be useful precisely when experimentation is unethical or infeasible. However, it does not involve fitting to data, only initializing with data and then perhaps qualitatively comparing simulation distributions to observed ones; and it is nearly useless for specific, real-world prediction (Pfeffer and Malik, 2017). It is therefore best being used as a method of *theory development* (Gilbert and Troitzsch, 2005), and is not applicable for most cases where statistics and machine learning are used. To believe otherwise, and think that such an abstract simulation might give insights into a specific real-world system, is an unjustified leap of logic that risks dangerously distracting from processes at play in the world (Shalizi, 2011), such as thinking that the Shelling model of segregation might describe segregation in Chicago rather than identifying redlining as the cause.

3   Prediction and explanation

Today, statistics and machine learning have enormous, if not complete, overlap in their underlying machinery. While machine learning started off trying to understand cognitive processes of learning, from the 1980s on, a strand of machine learning that used statistical models (Boden, 2006) in (arguably) instrumentalist ways (Jones, 2018) to mimic aspects of "learning"—namely, improving with "experience" (Mitchell, 1997) (which became operationalized as "data")—proved successful at many target tasks that had otherwise eluded artificial intelligence researchers (Norvig, 2010) and became the overwhelmingly dominant form of machine learning and, indeed, AI.[35] As part of this shift, machine learning adopted advanced statistical concepts in the span of only a few years (Wasserman, 2014). The result is that the use of models like linear regression, or logistic regression, could either be statistics or it could be machine learning, depending on *how* it is used, and for what purpose.[36] Also as a consequence, the "learning" of machine learning is now a metaphor for optimization, bearing almost no resemblance to what is studied under socio-cultural learning theory.

---

[34]Even network models do not escape this reductionism; they move what counts as an entity from individual people to dyads, triads, or even whole networks, but they are still reductionist to those (Erikson, 2013), and still do not allow for multiple interpretations or mutable entities.

[35]While other versions and visions of machine learning and AI exist, I do not discuss them here.

[36]Confusion around this is shown in a recent exchange of Christodoulou et al. (2019); Bian et al. (2019); Van Calster et al. (2019). I thank Muhammad "Tuan" Amith for these references. As Bian et al. (2019) rightfully point out, just because something uses logistic regression does not make it statistics (especially if it includes regularization). The difference, to use the articulation of Mullainathan and Spiess (2017), is in which part of the model matters; if a fitted linear regression is $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$, then statistics cares about $\widehat{\boldsymbol{\beta}}$ (which captures relationships between inputs and outputs), whereas machine learning cares about $\widehat{\mathbf{y}}$ (the match to the target system). That is: the model class is the same, but the final models include different terms based on the differing goals. And it is surprising that there is a trade-off between these goals, but (partially based on definitions of success; see below) there is (Shmueli, 2010). I thank Doaa Abu Elyounes for noting the lack of clarity on this point previously. There are other stylistic differences as well, such as statistics focusing on quantifying uncertainty, and machine learning paying heed to efficiency in implementation.

And unlike the goals of earlier machine learning, it arguably has vanishing relevance even for what is studied under cognitive learning theory (Watson, 2019; Marcus, 2018).

The goal of statistics, traditionally, has been to use data to understand something about underlying processes (Fisher, 1922). For this, quantifying uncertainty has been central; how well a model generalizes to new data (or, mimicking having new data by use of a held-out set) can be a good argument in support of the model's validity, but it makes perfect sense to, say, model data within a unique system (Kass, 2011). In contrast, while using much of the mathematical machinery and statistical models, machine learning seeks to design reliable input-output mappings, what Leo Breiman (2001) called "algorithmic modeling," and the emergence of which represented a distinct historical development (Jones, 2018).[37] Most machine learning research is "supervised learning," which is only meaningful if the systems or procedures demonstrated were actually to be deployed in an out-of-sample setting. Without demonstrations of success in applications settings, machine learning predictions "are not prediction at all!" but simply post-hoc claims that a prediction "could have been made" (Gayo-Avello, 2012).[38]

Machine learning having a narrow standard of success does provide many benefits: it means machine learning can arguably ignore the many problems that have been the focus of criticism within statistics, such as problems with quantifying uncertainty (Mallows, 1998; Freedman, 2009), interpreting probability (Freedman, 2009; Cox, 1990), and determining what the "inferences" of statistical inference are to (Kass, 2011), especially given model misspecification (Buja et al., 2014); so long as out-of-sample success can be demonstrated, none of that matters. The same applies for different types of validity; only external validity (generalizability) matters (Domingos, 2012). Even face validity can be ignored. If we successfully showed that astrology (or rather, a machine learning model fitting astronomical data to real-world outcomes) had robust out-of-sample success, then for machine learning purposes, it would be perfectly valid to use.

While both machine learning and statistics hold that there is an external world that takes primacy (i.e., over actors' interpretations and renegotiations of it; Payne and Payne, 2004), they differ in how they treat this. Also distancing machine learning from statistical critique is how statistics—including the mainstream of Bayesian statistics (Gelman and Shalizi, 2012)—has a strongly realist bent, hypothesizing an underlying truth from which data are generated, and seeing the task of statistics as recovering this truth (Lavine, 2019). The role of statistical theory is to understand how well a method recovers this underlying truth. In contrast, machine learning's internal definition of "improving with experience" (Mitchell, 1997), and indeed the very metaphorical

---

[37]Nonparametric statistics is frequently much more like machine learning, in often being much more useful for input-output mappings than for understanding underlying processes (although it is still distinguished from machine learning in usually including a quantification of uncertainty). Consider even a bivariate nonparametric regression, applied to data generated from a sinusoidal process: it would fit a curve close to the sinusoidal functional form automatically, but the model would not tell us in substantive terms that the functional form is, indeed, sinusoidal. Consequently, the fit would be mainly useful for prediction, and not insight into the underlying process. In this simple case, the data and fit could be plotted to perhaps visually identify the specific functional form (which is powerful if the data cannot be modeled as coming from a simple, known functional form), but this becomes harder and harder in higher dimensions. There, nonparametric models really only become useful for prediction, or exploratory analysis.

[38]Unsupervised learning, which corresponds to clustering within statistics, is usually not associated with claims of "prediction." It might be associated with claims about uncovering underlying structure, but there is ample work (von Luxburg et al., 2012) about the difficulty of validating clusters beyond the face validity of "making sense." Sometimes, clustering might be used to find structure that then is used as an input into a predictive model (e.g., dimension reduction techniques); but in this case, the final task of prediction is the standard of success. Rarely, an unsupervised approach might be taken to prediction (e.g., outlier-based anomaly detection for fraud): in those cases, a "ground truth" set of labels exists (or can be determined, such as by manual investigation) against which cluster assignments can be validated. Then, claims of success are again post-hoc statements about success that could have been achieved if the system were applied. I thank Doaa Abu Elyounes for encouraging me to clarify this.

use of learning, is more instrumental (Jones, 2018) and even agnostic about truth; a machine learning model simply improves its performance with more data, and does not necessarily uncover some underlying truth (with the exception of the subfield of causal discovery and causal learning, and sometimes claims from clustering). Machine learning theory, then, is focused on what the maximum possible *performance* is, and whether/how it might be achieved.

But the focus on prediction brings with it as many or more problems than it avoids. There are two issues. The first is that predictions, by virtue of being correlations, cannot provide explanations of the system, as most (but not all; see Woodward, 2003, for a review) philosophical treatments of "explanation" relate it to *causality* (Alexandrova, 2009; Reiss, 2012; Alexandrova and Northcott, 2013; Reiss, 2013; Miller, 2019),[39] where causality can be defined in terms of manipulating, controlling, or changing things in the world (Woodward, 2003).[40] The second is that, even our only goal is prediction, correlations are fragile basis for doing so.

3.1   Prediction is not explanation

As Zachary Lipton (2018) notes, "The optimization objective for most supervised learning models... is simply to minimize error, a feat that might be achieved in a purely correlative fashion." In fact, this is the *only* way in which they are achieved. Defining the output of optimized models as 'predictions' is to define predictions as extrapolation (or perhaps interpolation) from correlations. Defining prediction thusly is not universal. Indeed, in other disciplines (like physics, or economics) and perhaps in lay understanding as well, 'prediction' is assumed to be referring to causality: consider how Friedman (1953) talks about prediction *in the presence of change*, which is not the machine learning sense of the word.

Consider a case discussed by Peña and Varon (2019), where an anti-abortion conservative Argentinian politician hailed a system using public databases and Microsoft Azure system to predict teenage pregnancies (as part of his opposition to a proposed law legalizing abortion). Of the system, the politician said, "With technology, based on name, surname and address, you can predict five or six years ahead which girl, or future teenager, is 86% predestined to have a teenage pregnancy."[41] This was in support of a frequently advanced anti-abortion activist position that Peña and Varon (2019) articulate as: "if [governments] have enough information from poor families, conservative public policies can be deployed to predict and avoid abortions by poor women."

It is striking to see prediction explicitly equated with *predestination*, which is far beyond what machine learning predictions actually mean or do. And, it is also striking to note that this confusion about how certain "predictions" allowed machine learning to be rhetorically deployed in support of a regressive political agenda.

Talking about "prediction" can obscure how it is based on correlations. After all, how could a 'wrong' model predict well? Even further, we can use the bias-variance tradeoff to explicitly derive 'false' (misspecified) models that predict better than 'true' models (i.e., the specification of the data-generating process fit back to the data it generated; Shmueli, 2010). This is true even asymptotically, although (if the estimator is consistent, and within the same model class as the true model) the gap in performance becomes negligible. Practically, though, for finite-sample prediction (as all prediction ultimately is) there is a "conflict between the aims of optimizing finite-sample prediction and of finding the true data-generating model class" (Kunst, 2008) of a magnitude large

---

[39]I thank Hemank Lamba for the latter reference.

[40]I thank Lily Hu for this reference.

[41]In Spanish, "Con la tecnología vos podés prever cinco o seis años antes, con nombre, apellido y domicilio, cuál es la niña que está un 86% predestinada a tener un embarazo adolescente." The Spanish *predestinada* has the same metaphysical connotations as the English predestined. I thank Joana Varon for pointing me to the original quote.

enough to matter. Elsewhere, Mullainathan and Spiess (2017) discuss the division as "$\widehat{y}$" problems and "$\widehat{\beta}$" problems, noting that very different models, with very different causal implications (and hence, implications for public policy interventions) can achieve minimum loss equally well (see also what Breiman, 2001, terms the "Rashomon effect").

Philosophically, Nicholas Rescher (1998) theorizes "trend projection" and "curve fitting" (which Pearl and Mackenzie, 2018, critique mainstream machine learning as being limited to) as *rudimentary* or *elementary* types of prediction, ranking underneath the *scientific* or *sophisticated* types of prediction of "indicator coordination," "law derivation," or "phenomenological modeling." We can take or leave the normative implication of this progression, but either way, Rescher's levels are a useful way to relate what predictions mean in machine learning from they mean in physics or econometrics.

While it may seem like we can gain substantive knowledge about a phenomenon or entity based on which feature sets give superior predictive performance, this is essentially interpreting causation from correlation. It is a worthwhile *exploratory* exercise to hypothesize what causal mechanisms might generated observed correlations, but it is only that, and should not be presented as definitive (Lipton, 2018). Machine learning papers with titles like "Understanding..." or "Explaining..." or even "Studying..." or "Characterizing..." are thus potentially misrepresenting or misunderstanding what can be concluded from the methods used.

3.1.1   Explainable is not the same as explanatory

Furthermore, separating out the explainability or interpretability of a *model* from the idea that the model explains the underlying *system* is a difficult task. Leo Breiman (2001), in hailing the interpretability of decision trees, also provides an example of this breaking down:

> "A project I worked on in the late 1970s was the analysis of delay in criminal cases in state court systems... The dependent variable for each criminal case was the time from arraignment to the time of sentencing. All of the other information in the trial history were the predictor variables. A large decision tree was grown, and I showed it on an overhead and explained it to the assembled Colorado judges. One of the splits was on District N which had a larger delay time than the other districts. I refrained from commenting on this. But as I walked out I heard one judge say to another, 'I knew those guys in District N were dragging their feet.'"

Breiman does not remark further on this, but I would say: was District N actually dragging its feet, i.e. were faster criminal trials within their ability but simply not done? To properly answer this question would be to provide an estimate of the effect of being in District N, after controlling for other factors (such as, say, amount of resources). But a decision tree might use an indicator for being in District N as a proxy for the actual causal factors but achieve the same final predicted values. In this story, the judge interpreting the explainable decision tree arrived at an unjustified conclusion. That is, maybe the conclusion was actually accurate, but using the decision tree was not a valid way to make that determination.[42] This also relates to notions of model "correctness": as a description, a bivariate correlation is simply an observation and (within its way of framing the world, and dependent on the meaningfulness of the data) is "correct," but this examples shows that

---

[42]Consider also Letham et al. (2015), who present interpretable decision rules for classification using the *Titanic* dataset as an example. They write, "For example, we predict that a passenger is less likely to survive than not *because* he or she was in the 3rd class" [emphasis original]. The "because" refers here to the *model's* behavior, not to causality (which the model is not attempting to capture), although in this simple case they likely coincide.

this notion of correctness is deeply counter-intuitive, and, as I discuss below, perhaps too narrow and fragile to be an acceptable standard.

This is a running, but backgrounded, tension in the interpretability/explanability literature. Doshi-Velez and Kim (2017) write, "one can provide a feasible explanation that fails to correspond to a causal structure, exposing a potential concern." Caruana et al. (2015) write, "Because the models in this paper are intelligible, it is tempting to interpret them causally. Although the models accurately explain the predictions they make, they are still based on correlation." And Lipton (2015) writes, "Another problem is that such an interpretation might explain the behavior of the model but not give deep insight into the causal associations in the underlying data... The real goal may be to discover potentially causal associations that can guide interventions."

3.2 Correlations can fail

3.2.1 Fragility of non-causal correlations

Non-causal correlations are traditionally known as "spurious correlations," i.e. a correlation that empirically exists and is robust not because of a direct causal link, but by some structure like a common cause (Pearl, 2009). Murders and ice cream sales may indeed correlate, but the underlying cause is hot weather. Per capita, Nobel prize wins and chocolate consumption are correlated (Messerli, 2012), but (despite Messerli's joking post-hoc explanation about certain chocolate proteins stimulating brain function) the underlying cause is likely access to resources. However, "spurious correlations" are also used to refer to correlations that are the result of improper data mining, i.e. that are *not* robust do not generalize (Fan and Zhou, 2016).[43] For clarity, I will refer to these robust correlations as "non-causal" rather than spurious.

The danger of using non-causal correlations alone for prediction, as is also a central argument in causal learning literature (Spirtes and Zhang, 2016), is that predictions will not be robust to changes in context and over time. In such cases, underlying causal processes will lead to "distribution shifts." Bernhard Schölkopf (2015) references old behavioralist language (Katz, 2017) in analogizing this to "superstitious behavior," "in which statistical associations may be misinterpreted as causal" (although I would say that just because a model *uses* correlation for prediction, it does not mean we should speak anthropomorphically of the model as "interpreting" the correlation as causal; Watson, 2019). And Jonathan Zittrain (2019) proposes the idea of "intellectual debt" to describe the dangers of coming to rely on systems that work through correlations whose underlying causal structure we don't understand.

An excellent example of this is in the case of Google Flu Trends. Here, the cross-validation was done correctly (as I discuss in section 4), in temporal blocks (Ginsberg et al., 2009). Nevertheless, the system failed when deployed: as it turned out, the model was relying on a strong correlation between winter-related features and incidence of flu, not having observed the variability represented in cases where these two did not correlate. Lazer et al. (2014) phrase this as a problem of overfitting, and indeed we can think of models that fit to anything other than an underlying causal process

---

[43]While the "robust but not causal" sense of spuriousness is more common, I note that the famous article of Udny Yule (1926) talks about "nonsense-correlations" in the other sense: "if we had or could have experience of the two variables over a very much longer period of time we would not find any appreciable connection between them." This is also the sense used by the book of Vigen (2015), although the book does not get into why these correlations are spurious. Yule's explanations were ultimately getting at what we would today call temporal autocorrelation (and cyclic behavior); two time series can appear correlated with each other when in fact they only have similar autocorrelation or cyclic structure. One solution to this is to think in terms of effective sample size; if we calculated confidence intervals for correlations that took the effective sample size into account, the correlations may be large but still not significant, even before including a correction for multiple comparisons.

as "overfitting," but it is more strongly a lesson about causality ([Harford](), [2014]()). As more data are gathered, capturing more of the total range of variability, the system would likely improve; but there might be some other causal processes because of which the system fails later on.

3.2.2   Correlations are proxies and can be gamed

Another failure point is in how correlations, like any proxy, can be gamed. And, if there are any stakes associated with a proxy, it is nearly a guarantee that they *will* be gamed in a market system. "Goodhart's law"[44] is that "When a measure becomes a target, it ceases to be a good measure" ([Strathern](), [1997]()). This relates also to earlier work by [Ravetz]() ([1971]()), and to the "corrupting effect of quantitative indicators," about which [Campbell]() ([1975]()) wrote (emphasis original), "*The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures, and the more apt it will be to distort and corrupt the social processes it is intended to monitor.*"

Take the example from above (fig. [5]()) of image recognition: neural net models find groupings of pixels that correlate with laboriously collected human-given labels (labor that [Gray and Suri](), [2019](), refer to as "automation's last mile"). The groupings of pixels are only a proxy for the image content and neither the content itself nor analogous to human perception; as a consequence, adversarial examples can exploit this ([Goodfellow et al.](), [2015](); [Eykholt et al.](), [2018]()) and add noise imperceptible to human eyes but that can cause an image (or, indeed, an object from any orientation) to be arbitrarily misclassified as anything an attacker wants.

Indeed, one reason for keeping credit scoring systems secret is because knowing their operations would make it easy to game them by optimizing to the proxies that they ultimately use, which is far easier than optimizing the underlying construct (not so much creditworthiness—which implies a moral judgement and should probably be done subjectively—as whether or not somebody will repay a loan, which is "objective" but which only time can reveal). Cynthia [Rudin]() ([2019]()) writes, "The reason a system may be gamed is because it most likely was not designed properly in the first place, leading to a form of Goodhart's law if it were revealed... transparency could help improve the quality of the system, whereby attempting to game it would genuinely align with the overall goal of improvement. For example, improving one's credit score should actually correspond to an improvement in creditworthiness." But that does not address the issue of *constructs*; so long as we use proxies, Goodhart's law will invariably kick in. Creditworthiness is a construct that does not have a physical reality on which improvement can be known. The physical quantity of "whether or not a loan a given loan will be repaid," which is what a measure of creditworthiness is supposedly useful for predicting, cannot be known in advance. Yes, for potential borrowers it would be ideal that they only be judged on the actions they take that make loan repayment more likely; but repayment is likely heavily influenced by factors outside of a potential borrower's control, such that what minimizes loss for lenders is to use circumstantial information. This means that lenders will prefer using non-causal, potentially black-box correlations that they will keep secret, rather than trying to develop a "properly designed system" by the above criteria.

3.3   Correlations are not a reliable basis for intervention

Another consideration of what predicts, under this notion of prediction, not corresponding to causal processes is that predictions are unable to guide *interventions*. [Barabas et al.]() ([2018]()) discuss how risk assessments in the criminal justice system are also used for *diagnostic* purposes and in deciding how

---

[44]Originally, "Any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes" ([Goodhart](), [1975]()).

the criminal justice system intervenes in the lives of offenders; but this is unreliable. Mullainathan and Spiess (2017) give an example of applying the lasso to random subsets of the American Housing Survey data. In each subset, a very different set of variables are selected in, all giving equally good predictive performance[45] but, as they point out, giving very different implications for intervening. Indeed, the lasso has no guarantees that it will select causal variables (if they are even among the measurable and measured variables); consistency guarantees are to the *oracle* predictor, the predictor giving the best possible predictive performance within the given model class (Zhao and Yu, 2006). As above (Shmueli, 2010), the "true" model is not necessarily the best-predicting model, asymptotically and certainly in finite-sample prediction (Kunst, 2008) even if the estimator is consistent and in the correct model class. Of course, all bets are off if the estimator is not of the true model class, if the causal variables are not measurable or not measured or, more philosophically, if the true data-generating process cannot be described by *any* functional form, let alone one that is smooth enough to reliably estimate from available amounts of data.

3.4 What about size?

What is the relationship of the size of available data to these problems? Not much; size has little to do with causality. In statistical terms, the main benefit from size is that it allows inference for characteristics of smaller and smaller subpopulations (Cox, 2015). For machine learning, it allows greater and greater accuracy for predictions on such subpopulations.

Causality aside, size tending to the entire population does not even overcome selection bias, as Xiao-Li Meng (2018) shows in recent work. He breaks down the estimation error (the difference between an estimate and a true underlying population parameter) as product of three quantities: the correlation between whether people respond and what their response is (a "quality" index, e.g. people who vote for an unpopular candidate are less likely to report their choice honestly), the square root of the difference between the sample size and population size (a "quantity" index), and the standard deviation of the underlying quantity (a "difficulty" index). With this, he shows that there are three ways to make the expected error be zero: have the entire population, have a homogenous population (the standard deviation is zero), or have no correlation between whether people respond and what their response is. Conversely, while randomization guarantees the correlation between responding and response is zero, this term is otherwise $O(1)$. It will never go to zero, even in bigger and bigger populations.

3.5 Prediction policy problems?

Are there cases where correlation alone is sufficient? Alongside the obviously wrong claim in the 2008 "End of Theory" that with enough data we don't need "models,"[46] is the claim that "correlation is enough." There is a legitimate version of this argument: *sometimes*, under specific conditions, correlations *may* be sufficient (although, again, this has nothing to do with the size of the data). Breiman (2001) argued this, writing about "*prediction problems*," where focus on trying to approximate a data-generating process was unhelpful for the application; problems like detecting toxicity of chemicals in water, or speech recognition. He argued that statisticians have neglected these problems, and development of techniques for these settings "has occurred largely outside

---

[45]I.e., a case where there is no "stable set" (Meinshausen and Bühlmann, 2010).

[46]Any use of data requires summarizing or reducing, e.g. into a prediction. And any summary of data is necessarily a model (Fisher, 1922). "End of Theory" also starts with the Box (1979) quote, but completely misunderstands what a model is, or the relationships of models to data.

statistics in a new community—often called machine learning—that is mostly young computer scientists."

While not citing Breiman, Kleinberg et al. (2015) make a nearly identical argument fourteen years later, and to an audience of econometrics and public policy rather than statistics: that "there are also many policy applications where causal inference is not central, or even necessary... Not only are these prediction problems neglected [in public policy], machine learning can help us solve them more effectively." And, similarly to how Breiman argues that predictive solutions can nevertheless provide "information" (which I read as exploratory insight) about the underlying system even if that is not necessary or an explicit goal, Kleinberg et al. (2015) argue that the solutions to prediction policy problems can provide "not just policy impact but also theoretical and economic insights."[47]

They systematize the conditions for such "prediction policy problems," metaphorically labeled "umbrella problems" (in contrast to causal problems, metaphorically labeled—in move that is unfortunately more primitivist than it is humble about causal inference—"rain dance" problems): our intervention must have no effect on the outcome, such as how umbrellas have no effect on rain.

Are there really such problems? Breiman (2001) gives "sources of delay in criminal trials" as an example of a "prediction problem," but "sources" is a causal ("$\widehat{\beta}$") question and as I argue above, what the clients (the judges) really wanted was causal knowledge! Correlations might still be *useful* if used as exploratory, but they would not be sufficient for guiding interventions. Notably, while he gives other candidate tasks that are promising, they are of physical and not social systems (using mass spectra, radar, sonar).[48] And, as referenced earlier, Barabas et al. (2018) discuss how, in practice, risk assessments in criminal justice are used for deciding on interventions; if this is how they are being used, risk assessment is not a prediction policy problem.

Kleinberg et al. (2015) propose five examples that are also questionable. A full reconsideration of these examples is beyond the scope of this paper, but I mention some problems with a few of them.

Their first example is "predicting which teacher will have the greatest value added" in education, which I would see as a causal question (what *causes* teachers to add value?), but would also point out that the study they cite measures "value" by students' improvement on standardized test scores. This is an unreliable measure of the underlying construct of "having learned" (Jacobs and Wallach, 2019; O'Neil, 2016), and the uncertainty of this carries through to the potential prediction policy problem.

Their third example is "targeting health inspections" in restaurants using Yelp review data,[49] but this setup is (potentially) adversarial, with any proxy being susceptible to gaming and manipulation. While clouds don't determine whether or not they rain based on the presence of umbrellas, restaurants could make compliance decisions based on how inspections are directed.

Their fifth example is "lenders identifying the underlying credit-worthiness of potential borrowers." But this is both about an underlying construct (credit-worthiness), and a construct for which the historically and currently used proxies have caused and continue to cause oppression and increasing

---

[47]Similar papers, arguing for the neglected importance of prediction, are recently appearing for many other fields as well such as Yarkoni and Westfall (2017) for psychology and Bzdok et al. (2018) for biology. Of these two, only the former is aware of and cites Breiman as precedent for this argument, but both conflate prediction with understanding, failing to acknowledge how a model can "predict" well without corresponding to a causal process.

[48]Another example he gives, "On-line prediction of the cause of a freeway traffic breakdown," has the word "cause" but presumably the cause of a given breakdown would be a human-coded label, which would be what a model would try to recover as a mutliclass problem, rather than something inferred from data that measures aspects of a traffic breakdown.

[49]While the paper they cite has recently been disproved in a re-analysis by Altenburger and Ho (2019), who show reviews provide no stronger a correlation than restaurant characteristics, those restaurant characteristics could still be the basis for this being a prediction policy problem.

inequality (Lauer, 2017; Poon, 2007; Fourcade and Healy, 2013), and which I further discuss below. There might yet be examples of "prediction policy problems," where the data and system are sufficient to apply correlation-only models to questions of public policy and the social world; but cases claimed so far, upon closer examination, have many problems.

3.5.1 Injustice of correlations

Conversely, what happens when we apply a correlation-only approach to problems that are *not* prediction policy ones? Predictions have consequences: as McQuillan (2018) writes, "The predictive nature of machine learning promotes preemption, that is, action that attempts to anticipate or prevent the predicted outcome," which Pasquinelli (2019) interprets as transforming correlations into a "political apparatus of preemption."

For a case discussion, I take the last example of Kleinberg et al. (2015) of creditworthiness. In recent work, Rodrigo Ochigame (2020) connects the history of credit to machine learning, particularly around "actuarial fairness," and here I summarize and expand his arguments. Note that risk assessment in criminal justice has very much the same story, with many historical developments happening parallel to (and not independent of) those of credit; this connection is summarized by Barabas et al. (2018), who also show how the ethical issues and sociological consequences of risk assessment are instructive for how to think about machine learning.

For credit, being given a loan is not decided by things like people's intention or effort towards repayment, which are the only things over which people have control; instead they are decided by things over which people potentially have no control, namely the central tendency of previous repayment behavior by those who share attributes with the individual. For the US, Josh Lauer (2017) details how creditworthiness decisions in the 1960s were made on the basis of judgements of "character"; this was far from fair, but there could be notions of (and methods of making determinations about) character that would be a fair standard against which to hold individuals accountable. This approach, though, was gradually replaced by one based on optimization and correlation that was far more empirically successful, just as in risk assessment at the same time (Hannah-Moffat, 2013).

Having such data-driven decisions, based on correlations that become how lending "risk" is estimated, is certainly more efficient and effective for lenders; but it effectively rewards and punishes people, shaping what Fourcade and Healy (2013) call "life chances," for things over which they potentially have no control.

While the shift in the insurance industry happened in the 1960s, attempts at empirical justifications for differential treatment goes back further. Dan Bouk (2015) details how, in the 1880s and 1890s, insurance companies were charging African Americans more for life insurance (often without their knowledge), which was noticed by Massachusetts state representative Julius C. Chappelle, "an African American born in antebellum South Carolina and a janitor for the state of Massachusetts by trade," who proposed a bill to ban the practice.

> "Frederick Homer Williams, a lawyer from Brookline, chaired the committee that opposed the bill... Williams cited statistics from around the nation showing shorter life spans for blacks, including 1870 census figures showing a 17.28 death rate for 'colored people' against 14.74 for whites. These numbers, Williams argued, and not any 'discrimination on the ground of color' motivated insurers' rates. It was a 'matter of business,' and any interference, he warned ominously and presciently, 'would probably cut off insurance entirely from the colored race.'

> "Chappelle's allies noted that Williams's statistics, while bleak enough, answered the

wrong question. The question was not whether blacks in slavery or adjusting to freedom were poor insurance risks, or even whether southern blacks were poor risks. The question was African Americans' *potential* for equality and specifically the present and future state of Massachusetts' African Americans—about whom no statistics had been offered by either side."

While the mortality statistics were perhaps *empirically* true, as a basis for justifying charging more, it exacerbated inequality and was therefore rejected. Here, the argument for equality won the day, with the bill passing and inspiring similar bills in other states.

The move from a moral basis to an empirical basis that happened in the 1960s provoked similar opposition. In an argument summarized by Ochigame (2020), Caley Horan (2011) details how, in the 1970s and 1980s, civil rights and feminist campaigners charged the US private insurance industry with discrimination; the industry responded by promoting the idea of "actuarial fairness," which Ochigame et al. (2018) summarize as "the antiredistributive principle that 'each person should pay for his own risk,' in opposition to the principle of solidarity and mutual aid (Stone, 1993)." Horan (2011) writes, "Admitting that fairness might mean one thing socially and another 'actuarially' allowed pragmatically minded activists a way out of sticky philosophical discussions and even stickier debates with insurers over highly technical aspects of industry. Yet this concession, at the same time, had a crippling effect on attempts at regulation. Once insurers were granted their own unique definition and claim to fairness separate from that of society writ large, the application of social legislation became nearly impossible." Continuing with their summary of Horan (2011), Ochigame et al. (2018) write, "The industry not only... defeated the efforts of civil rights and feminist activists [and] evaded legal regulation but also cemented an understanding of private insurance as essential, natural, and detached from questions of social justice."

In another facet of this debate, Barbara Kiviat (2019) details resistance from consumers and local lawmakers against the auto insurance industry adopting certain measures, namely credit scores; these indeed *correlated* with previous higher insurance payouts, but seemed to industry outsiders to lack a causal connection to being a good driver, or be something to which it was fair to hold drivers morally accountable (and therefore fairly rewarded with lower insurance premiums or punished with higher premiums). And Devin Fergus (2013) details how US auto insurance is based more on postal code than driving record, which in effect is a "ghetto tax" that punishes female, black, and Latinx drivers, those least able to afford a higher burden. Postal code (and indeed, ultimately, race itself) may indeed correlate more highly with costs to insurance companies than driving records, because structural factors are frequently more important for outcomes than factors under people's individual control.

Today, credit decisions being made on an expanded set of types of data that can be correlated with previous repayments (Nopper, 2019) are the continuation of this historical legacy, with equally troubling implications.

A legal and philosophical analysis is given by Deborah Hellman (2008) in her book, *When is discrimination wrong?* (in an argument later also presented to a machine learning audience; Hellman, 2018). She develops a theory of when discrimination (i.e., drawing distinctions between people and treating them differently on that basis) is morally permissible, versus when it is morally impermissible. She identifies a trait being "relevant" or "irrelevant" as an insufficient standard, since "relevant" could be defined simply as correlated, and "sex is likely correlated with work schedules and the costs of childbearing" yet such reasoning has been rejected as contributing to marginalization and overall undesirable inequality. Her solution to when a distinction is impermissible is when it "demeans," which she defines as treating another in a way that, in a given context, denies equal moral worth (regardless of intention, and regardless of whether or not an affected person feels demeaned,

stigmatized, or harmed). If we were to adopt this standard, we could combine this argument with that of Fourcade and Healy (2013) that ties being considered morally worthy to being given positive life chances, and say: classifications that negatively affect the ability of people to exist, survive, thrive, and change in positive ways of their own choosing are morally wrong, regardless of their empirical accuracy.

That is, much of the "fairness" literature in machine learning perhaps concedes too much in presuming that we will, or should, be using the correlation-maximizing models of machine learning, and privileging efficiency and the ability of, say, insurance companies to minimize uncertainty around payouts and thereby increase profits. For many of the applications currently being discussed, perhaps what would be actually "fair" is to collectivize risk, or even to deliberately redistribute resources to those whose structural position puts them at the highest "risk." This would be a completely different optimization objective, or perhaps a political task like prison abolitionism (Benjamin, 2019; Ochigame et al., 2018) that falls outside of any optimization framework.

3.6 Optimizing to the status quo

Carr (2014) critiques big-data approaches as ultimately encouraging us to "optimize the status quo" rather than challenge it, channelling the point of Paulo Freire (1970) critiquing teaching to the status quo. If the only available data reflect heavy social biases, taking previous success as the optimization objective will only reproduce those biases (Noble, 2018), especially in machine learning where we rely on what correlates most optimally. Perhaps if we identify possible underlying causal structures, we can try to see how we might divert things from their current course. And, outside of empirical research or anything we can extrapolate from existing data, we can imagine wholly different futures through fiction and social theory. Within cultural products, optimizing to previously commercially successful music or to film scripts (Barnes, 2013) prevents investment in new art and risks making cultural products stale. And, outliers are by definition observations that fall far outside of a central tendency, making it difficult to use them as exemplars on which to fit a machine learning model.

3.7 Problems in implementation

This paper is devoted to articulating the implicit assumptions made when using machine learning, and the problems of these assumptions. But I want to acknowledge that problems of *implementation* are an additional layer.

First, there are issues with implementation in software (Merali, 2010). Documented examples of mistakes in scientific results mostly have to do with data handling, prior to going into software that does analysis and model-building (Miller, 2006; Karraker and Latham, 2015), especially when Excel is involved (Panko, 1998; Hutson, 2010; Herndon et al., 2013). But mistakes happen in analysis as well (Qiu et al., 2019). There are concerns about using out-of-date software libraries; Irpan (2018) gives the anecdotal example of finding a warning message "saying my NumPy version was too old," and being "told that sufficiently old versions of NumPy led to incorrect physics calculations that broke some baseline algorithms. No one understood why this happened, and no one thought it was worth debugging (and rightfully so) [versus just updating NumPy]."

Difficulty in doing something correctly can be considered a limitation, and indeed the difficulties of reproducible code and sensitivity to initial conditions in machine learning (Tian et al., 2019) is an increasing source of doubt for the overall reliability of machine learning research (Hutson, 2018).

But manual or mechanical computation as an alternative to software implementation is simply not feasible for the models used by machine learning—and, at this point, even for the models used

in statistics (Efron and Hastie, 2016). For standard models (e.g., KDE, GLM, SVM, random forests, etc.), there is a choice about whether to implement them from scratch or to use standard software packages. But if available, in general there is no drawback to relying on standard packages that are often extensively verified, field-tested, and optimized. For more specialized models or field-specific applications, rewriting model implementations where feasible may provide needed robustness checks (Bhandari Neupane et al., 2019), but this can be seen as a part of best practice rather than a decision or assumption with limitations.

For deep learning models, there may be a decision to be made about how much computational budget to use on searching through the solution space, as this has a huge impact on the quality of results (Dodge et al., 2019) but better results come at the cost of potentially massive carbon emissions for the energy the computation requires (Strubell et al., 2019).

But the biggest point of implementation is not in software or even computational infrastructure; it is in institutions and power dynamics. Even if we had a perfect prediction policy problem, reliable software, and competent implementers with the best of intentions, the system might still go awry. Eubanks (2018) describes three cases of computer systems for optimizing and automating decisions, detections, and resource allocations. While one of these cases seemed to be driven by bad faith (organized around a desire to see as much denials of benefits as possible), the other two projects (one implementing a matching algorithm and the other a machine learning model) had system designers with genuinely good intentions. Yet despite these, she describes the dehumanization felt by people subject to the systems, and the myriad ways in which the systems were not clearly better than what had come before, and were in some ways worse.

We also cannot assume that models outcomes alone determine how people are treated; the "human-in-the-loop" can present another layer. For example, work looking at how judges use recidivism prediction systems for decision support (Cowgill, 2018) shows that judges do not ignore the predictions (if they did, what would be the point?), but nor do they adopt them entirely, and this selective adoption is nonrandom. How people treat such recommendations can be affected by factors like the race of the defendant, as suggested by MTurk testing that shows higher predicted risks overriding priors more often for black defendants (Green and Chen, 2019). This is an emerging, and important area of research: how end users react to and use recommendations should be studied (and potentially modeled), and is central to the question of whether or not a model is effective and successful.

3.8 Alternatives: Modeling causality?

What is the alternative? Of course, inferring causal relationships from correlations alone is unjustified (Freedman, 2005, 1997); further tools are needed. One option is to return to statistical modeling, and specifically, the kind of causal inference done in econometrics. This addresses many problems in prediction-only approaches, but this raises its own host of problems of statistical limitations and conceptual inconsistency, as summarized in a review by Syll (2018). One problem is the "fundamental problem of causal inference" (Holland, 1986), that we can never observe the same unit undergoing more than one treatment: so we need to make assumptions about how two units with different treatments are comparable (such as random assignment, or matching, or using structural relationships that can be interpreted as natural experiments such as with instrumental variables or regression discontinuity). This also means that cross-validation is a poor way of validating causal estimates (Wager and Athey, 2018).

Gelman (2009) notes that doing observational causal inference using natural experiments is limited to what natural experiments econometricians manage to find. And, even perfect application of techniques can break down in the presence of a latent confounder. This is the case in a

fantastic example by Arceneaux et al. (2010), who show that the seemingly innocuous confounder of "reachability by phone" in a test of a get-out-the-vote campaign, is so powerful that it defeats perfect nonparametric matching on all of a rich set of observed covariates (including age, gender, party registration, and past voting, all which fail to control for reachability) in a sample size of over a million.

The language of graphical models and do-calculus (Pearl, 2009) are a great advance in the ability to make precise causal statements, analyze relationships between such statements, and connect causality to probability distributions and statistical inference. But as a basis for *estimating* causal graphical structure, Humphreys and Freedman (1996) argue that the guarantees rely on strong, untestable, and probably false assumptions. More recently, Freedman (2004) remained skeptical about the assumptions needed to do causal learning, noting that substantial prior causal knowledge is needed for these models to work, but there are few causal pathways that can be excluded *a priori*.

And even as a representational tool, Richardson (2020) notes that causal graphs are capable only of representing short-term, immediate causal relationships; diffuse, long-term processes like structural racism (Richardson et al., 2016) don't fit easily into a directed acyclic graph. If we rely on such representations, we risk constraining our thinking away from the solutions that are actually needed, just as Barabas et al. (2018) talk about how prominent researchers in criminal justice risk assessment tools "have explicitly discounted the relevance of sociological factors, such as ethnicity and socioeconomic status, because those factors are viewed as static and challenging to intervene upon."

Where possible, experimental manipulations can introduce the full range of variability of inputs that allows observation of how they covary with outputs. But, even within the bounds of statistical theory, in order to interpret the coefficients of, say, a linear model as causal and generalizable, one would need *random assignment* of treatments in a *representative sample* of a target population (Aronow and Samii, 2016)[50] and an experimental setup which includes all *key variables* (whether controlled or manipulated) without uncontrolled interference (Rosenbaum, 2007). This may be demanding too much—for example, noncompliance ("Student," 1931), imperfect randomization, and participant crossover (Teele, 2014b) are persistent problems. On the latter point, omitting or controlling for some key variable or variables in an experiment could lead to a lack of ecological validity: that is, the causal processes that the experiment allows us to identify may not actually happen in the larger world.

It is an open debate about how to balance these concerns, detailed in the collection edited by Teele (2014a).[51] Randomized control trials are the "gold standard" for causal estimation, but without ecological validity and a representative sample (which is harder in random assignment), an estimate may mean relatively little. Another option is field experiments, which pursue greater ecological validity, but they sacrifice random assignment and the ability to control for all variables.

Even with any type of experiment, there remain profound philosophical questions about whether causality, (as defined in terms of manipulation, control, or change; Woodward, 2003) is knowable with any certainty: is it sufficient that we observe that doing a certain manipulation produces a certain outcome in some majority of cases to declare that we have found a causal relationship? How can we ever truly rule out some hidden confounder as being the true cause? As in the thought experiment of Descartes' malicious demon, there could be a universal confounder that magically changes things to fool us into thinking there are causal relationships where there are none. We can also interpret this as what a "black swan" event does (treating it as something more specific than just an extremal value in a distribution): reveal some causal process that we did not know (or

---

[50]I thank Baobao Zhang for this reference

[51]I thank Baobao Zhang for this reference as well.

assumed away) beforehand, changing everything we thought we knew about a system.

This also relates to the earlier critiques of quantitative versus qualitative analysis I gave; things that are difficult or impossible to quantify are just as important as things that can be quantified, creating blindspots in modeling, whether that modeling is causal or predictive. It also relates to the central tendency. Causal estimates being of average treatment effects again are of a central tendency; treatment effects may be very heterogenous within a population (Deaton, 2014), and preferencing the average may disadvantage those who fall far from it. Estimating heterogenous treamtent effects partially ameliorates this and is superior insofar as it may be possible, but still is fundamentally only making finer and finer bins. Probabilistic models can say nothing once a bin has only one observation.

4   Assessing model performance

Machine learning can claim to avoid all issues of interpretation, and of construct validity, internal validity, criterion-related validity (DeVellis, 2017), and even face validity, if it can demonstrate *external validity*: that is, generalizability (although the argument of Jacobs and Wallach, 2019, is precisely that these issues cannot be ignored for achieving certain gaols). Within academic research, machine learning does so almost exclusively with the use of cross-validation: splitting data into a training set and a test set, and using performance on the test set as the basis for claims about performance on out-of-same data, i.e. about generalizability. Thus, the ability of cross-validation to establish generalizability is key for the overall validity of current machine learning research.

Rescher (1998) writes about how each prediction necessarily entails a metaprediction: a prediction about how likely is that prediction to hold. As cross-validation gives "predictions of the prediction error" (or, estimates of the generalization error), it is a metapredictive exercise, but we can also investigate one level higher: how likely is it for cross-validation to hold?

There are a few distinct processes under which cross-validation can fail to hold. There are, of course, upstream problems. Bad labeling schemes, as discussed above, make everything following it meaningless. Filtering only to observations that have an observable signal (Tufekci, 2014) also makes models not generalizable, as in Cohen and Ruths (2013); in their case, the showed how going from standard filtering practices to a (more) representative sample led to classifier performance dropping from over 90% to barely 65%. If there is a distribution shift coming from the causal structure of the system that introduces never-before-seen covariance, as in Google Flu Trends, correlation-based modeling alone will fail. Not maintaining strict separation of the test set (e.g., extracting features before data splitting) can "contaminate" evaluations of performance (Domingos, 2012). But beyond those problems, with cross-validation *itself*, there are three major threats to validity:

4.1 Publication bias.

4.2 Overfitting to the test set.

4.3 Dependencies among observations.

4.1   Publication bias

Gayo-Avello (2012) notes that the "predictions" of machine learning research in social media are really "claim[s] that a prediction could have been made, but the analysis is post hoc. And needless to say, negative results are rare." By this, he is referencing how negative results are not published, such that some positive findings may be the result of luck rather than something generalizable. That is: just because machine learning doesn't use p-values doesn't mean it escapes the problems of getting

positive findings just by chance. This also ties into the increasing attention to a "reproducibility crisis" in machine learning (Hutson, 2018), just like in biomedical science or psychology, with publication bias being a concern alongside inaccessible data, undocumented tuning parameters and random seeds, and unusable code.

4.2 Overfitting to the test set

Overfitting to the test set is a recognized problem (Rao et al., 2008; Cawley and Talbot, 2010; Reunanen, 2003; Dwork et al., 2015a), but perhaps under-appreciated in larger discourse. Dwork et al. (2015b) write, "in practice even data allocated for the sole purpose of testing is frequently reused... Such abuse of the holdout set is well known to result in significant overfitting on the holdout or cross-validation set." They give Kaggle competitions as an example of this. Kaggle is cleverly set up: competitions have a "public" leaderboard, where contestants can see their rankings through the duration of the competition, but also a "private" leaderboard, the results of which are not revealed until the end, and rankings on which are the basis for actually determining winners. The existence of this second, private, test set allows Kaggle to catch overfitting to the public test set. Kaggle forums contain discussions of people finding their ranking dropping enormously from the public to private leaderboard,[52] and Park (2012) details how repeatedly submitting to rise on public leaderboards actually led to him unknowingly falling on the private leaderboard.

In *any* case where data is re-used for purposes other than testing, it can bias performance evaluation (Cawley and Talbot, 2010). This is a drawback to benchmark data sets; while it is incredibly important to have common data and measures of success to make sure that competing methods are compared on the same grounds, after a certain point any improvements to the state of the art may be overfitting (Gijsbers et al., 2019). This is a place where significance testing may be good practice machine learning; McNemar's test (McNemar, 1947) is a version of the χ^2 test for contingency tables applicable to paired comparisons, and therefore appropriate for the confusion matrices used to determine performance in machine learning (Dietterich, 1998). This test is one way (Raschka, 2018) to test if an improvement in accuracy of one classifier versus another (or over a baseline of the majority class rate) is greater than might be expected by random chance, with other work offering appropriate tests for other metrics like precision (Gondara, 2016). However, we should also make a correction for multiple comparisons; but without knowing how many attempts have been made across all researchers to improve on a certain baseline, we would not know what correction to make (or, equivalently, what would be a scientifically appropriate level at which to test).

The third problem, dependencies, is my focus here, as it has received no systematic attention. First, I set up the context by introducing the idea within statistical theory of optimism, and then present an extension of the argument to connect to the issue of dependencies among observations.[53]

4.3 Dependencies

In statistics, the problems of data that are not independent and identically distributed (iid) are well-known; dependencies shrink standard errors, and potentially cause bias in estimates. In causal inference and even experimental design, dependencies create interference (Rosenbaum, 2007) that can confound causal estimates.

---

[52] "What is up with the final leaderboard?" *Kaggle*, StumbleUpon Evergreen Classification Challenge. https://www.kaggle.com/c/stumbleupon/discussion/6185.

[53] Depending on the area, this problem is also known as non-iid data, autocorrelation (for certain types of dependencies), endogeneity, and 'pseudoreplication.' I thank Jeremy Koster for pointing me to the literature on pseudoreplication.

Machine learning, in not doing statistical inference, does not need to be concerned with estimating standard errors at all, and does not care about unbiased estimators; but it does not escape the consequences of the fundamental need of probability-based modeling to have multiple observations in order to accomplish anything. If machine learning fails to understand how dependencies change the effective sample size and affect modeling, it too suffers.

Trivially, if the test set is an exact copy of the training set, obviously the performance on the test set is unreliable as a guide to out-of-sample performance. But duplicated observations are only the most simple case of dependent observations; partial dependencies will still affect training error.

The main review of cross-validation by Arlot and Celisse (2010) has only a cursory treatment of cross-validation with dependent data, and focuses on the one example of dependencies of time-series. Solutions that are appropriate mainly for that setting, such as "modified CV" (Chu and Marron, 1991), similar to the *hv*-block cross validation of Racine (2000), require that observations are uniquely ordered in sequence, that we have a way of characterizing distances between observations, and that there is a certain distance beyond which pairs are independent. While the general principle is good, the case of observations being ordered is fairly limited one. And even though probabilistic graphical models have a rich language for describing dependencies, it is not extended to metaprediction: the canonical textbook (Koller and Friedman, 2009) does not connect these dependency structures to how to do data splitting for cross validation.

As a way of understanding this in more generality, I propose using the framework of *optimism*, defined as the amount by which the training error—the empirical loss of the model under the same data used to fit the model—departs from the true, out-of-sample error (Efron, 2004).

There is a classic argument, for why the expected optimism will be positive, making the training error less than the true out-of-sample error (and hence, 'optimistic').

We imagine a data-generating process $Y \sim f$ with $\mathbb{E}_f(Y) = \mu$ and $\mathrm{Cov}_f(Y, Y) = \mathrm{Var}_f(Y) = \Sigma$. We use a draw from this process, $\mathbf{y} \in \mathbb{R}^n$, to form an estimator $\widehat{\boldsymbol{\mu}}_{\mathbf{y}}$ (perhaps using correlations between $\mathbf{y}$ and a set of features $\mathbf{X}$, not included here in the notation). For a set of unknown values $\mathbf{y}^*$, the predicted values would be $\widehat{\boldsymbol{\mu}}_{\mathbf{y}}(\mathbf{y}^*) = \widehat{\mathbf{y}}^*$. We can also apply this estimator back to make predictions of the same $\mathbf{y}$ that produced it, and we call these the *fitted values* of $\mathbf{y}$,[54] $\widehat{\boldsymbol{\mu}}_{\mathbf{y}}(\mathbf{y}) := \widehat{\mathbf{y}}$. For theoretical analysis, we can analyze $\hat{\mu}_Y(Y) = \widehat{Y}$ as a random quantity, representing the *procedure* of transforming a draw of Y into an estimator (Tibshirani and Rosset, 2019).

While the bias-variance tradeoff can for the most part be generalized (Friedman, 1997) to arbitrary loss functions (although things can get strange for 0/1 loss; Domingos, 2000), the bias-variance tradeoff is most easily illustrated with squared error loss. Squared error loss is empirically calculated by $\frac{1}{n}\|\mathbf{y} - \widehat{\mathbf{y}}\|_2^2$. We can see this as an estimator, $\widehat{\mathrm{err}}$, of the *expected* in-sample error or *expected* training error, to which we can apply the standard bias-variance argument (involving introducing a positive and negative version of certain terms in order to use probability identities),

---

[54]These may also be called the 'predicted values,' but I believe it makes sense to say 'fitted values' if and only if talking about training/in-sample data.

given here in matrix form:

$$\text{err}(\hat{\mu}) = \frac{1}{n}\mathbb{E}_f\|Y - \widehat{Y}\|_2^2 \tag{1}$$

$$= \frac{1}{n}\left[\mathbb{E}_f\|Y\|_2^2 + \mathbb{E}_f\|\widehat{Y}\|_2^2 - 2\mathbb{E}_f(Y^T\widehat{Y})\right] \tag{2}$$

$$= \frac{1}{n}\left[\mathbb{E}_f\|Y\|_2^2 + \mathbb{E}_f\|\widehat{Y}\|_2^2 - 2\,\text{tr}\,\mathbb{E}_f(Y\widehat{Y}^T)\right]$$

$$+ \frac{1}{n}\left[-\mu^T\mu - \mathbb{E}_f(\widehat{Y})^T\mathbb{E}_f(\widehat{Y}) + 2\mu^T\mathbb{E}_f(\widehat{Y})\right]$$

$$+ \frac{1}{n}\left[\ \ \mu^T\mu + \mathbb{E}_f(\widehat{Y})^T\mathbb{E}_f(\widehat{Y}) - 2\,\text{tr}\,\mu\mathbb{E}_f(\widehat{Y})^T\right] \tag{3}$$

$$= \frac{1}{n}\left[\text{tr}\,\Sigma + \|\mu - \mathbb{E}(\widehat{Y})\|_2^2 + \text{tr}\,\text{Var}_f(\widehat{Y}) - 2\,\text{tr}\,\text{Cov}_f(Y,\widehat{Y})\right] \tag{4}$$

The error of an estimator $\widehat{\mu}$ in the *test* set can be seen, again, as itself an estimator of the *out-of-sample* error, also called the generalization error, the prediction error, or the test error. This quantity is defined as the error on a new draw from the data-generating process, Y^*, and is

$$\text{Err}(\hat{\mu}) = \frac{1}{n}\mathbb{E}_f\|Y^* - \widehat{Y}\|_2^2 \tag{5}$$

$$= \frac{1}{n}\left[\text{tr}\,\Sigma + \|\mu - \mathbb{E}(\widehat{Y})\|_2^2 + \text{tr}\,\text{Var}_f(\widehat{Y}) - 2\,\text{tr}\,\text{Cov}_f(Y^*,\widehat{Y})\right] \tag{6}$$

The only difference between the expected training error and the true out-of-sample error is in the respective covariance terms. $\widehat{Y}$ is dependent on Y since it is formed from it (e.g., see eqn. 8), so the covariance term in eqn. (4) is nonzero. On the other hand, $\widehat{Y}$ is formed without any knowledge of Y^*, so they should be independent; and if they are, the covariance term in (6), $\text{tr}\,\text{Cov}_f(Y^*,\widehat{Y})$ is zero. Then, the difference between the training and test error is called the *optimism* (Efron, 2004):

$$\text{Opt}(\hat{\mu}) = \text{Err}(\hat{\mu}) - \text{err}(\hat{\mu}) = \frac{2}{n}\,\text{tr}\,\text{Cov}_f(Y,\widehat{Y}). \tag{7}$$

This is the expected amount by which the training error will under-estimate the true error.[55] For a linear smoother, e.g. $\mathbf{Hy}$, this covariance term reduces to $\frac{2}{n}\,\text{tr}\,\mathbf{H}$; thus, an alternative to cross-validation is to estimate this covariance penalty directly (Efron, 2004), through parametric or nonparametric means.

Note that recently, Rosset and Tibshirani (2019) have extended the analysis of optimism to the random-X case, more appropriate to observational data, whereas previously a simplifying fixed-X assumption has been made.[56] They show that randomness in X adds an additional bias term and an additional variance term to the above derivations, although they make the iid assumption so it is possible that relaxing that assumption would also produce an additional covariance term. But it is unclear that, even if there is an additional covariance penalty, if this would suggest anything different for how to think about designing cross-validation procedures to guard against optimism in the test set. This is an opportunity for future investigation.

Dependencies between the training set and test set means that the covariance term in the optimism may not be zero, making the test *set* error a downwardly biased (overly optimistic) estimator of the true out-of-sample error. In this sense, defining the "test error" as synonymous with the generalizability error may be unwise.

I provide two examples of how this might happen.

---

[55] While no general proof exists that the trace of this covariance is always positive (i.e., that optimism is always positive), Zhang (2004) shows that this only happens if either we use a nonconvex loss function (i.e., so the estimator is *rewarded* the further it is from the training values), or if the estimator has *negative* sensitivity to each data point (the more a given data point changes, the less the estimator changes; this is equivalent to the estimator having negative degrees of freedom). Tibshirani and Rosset (2019) note that they cannot think of a way an estimator could have negative degrees of freedom, but they do not rule out the possibility such as in some pathological case.

[56] I thank an anonymous reviewer for pointing me to this work.

4.4   AR(1) time series

Consider a homoskedastic, mean-zero, stationary first-order autoregressive process, $Y_t = \phi Y_{t-1} + \varepsilon_t$, with $|\phi| < 1$ and $\varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$. A 2-nearest neighbor regression estimator would, before data splitting, be $\hat{\mu}_Y(Y_t) = \widehat{Y}_t = \frac{1}{2}(Y_{t-1} + Y_{t+1})$. If we randomly partitioned the data set into training and test, and the point Y_t happened to fall into the test set whereas Y_{t-1} and Y_{t+1} were both in training, then the decomposition of the test error at point t would include a term for the covariance between the point Y_t and the point $\widehat{Y}_t$ (rather than between some independent copy Y_t^* and $\widehat{Y}_t$). The autocovariance function of a mean-zero AR(1) process is $\gamma(h) = \mathbb{E}(Y_{t+h} Y_t) = (\sigma^2 \phi^{|h|})/(1 - \phi^2)$ (Shumway and Stoffer, 2011), and the t-th diagonal element of the covariance matrix will be the average of the autocovariance between two sets of adjacent points:

$$\begin{aligned}
\mathrm{Cov}(Y_t, \widehat{Y}_t) &= \tfrac{1}{2}\mathbb{E}(Y_t(Y_{t-1} + Y_{t+1})) \\
&= \tfrac{1}{2}\left[\gamma(1) + \gamma(-1)\right] \\
&= \tfrac{\sigma^2 \phi}{1 - \phi^2}.
\end{aligned}$$

If ϕ is positive, then the test set error would be providing a downwardly biased, overly optimistic estimate of the true out-of-sample error. (Alternatively, if ϕ were negative, then the test set would provide an overly pessimistic, and upwardly biased estimate of the true out-of-sample error; but also, a 2-nearest-neighbor regression would not be a very good estimator.) A 4-nearest neighbor estimator would similarly give, for a held-out point Y_t amidst four training points, the average of the covariances of two sets of adjacent points and two sets of points once removed, $\sigma^2(\phi + \phi^2)/(1 - \phi^2)$, and so on.[57]

For forecasting, randomly partitioning a time series into training and test has, in terms of test set error serving as an assessment of real-world performance, been critiqued as "time-traveling."[58]

In a series of simulations, Bergmeir et al. (2018) consider different cross-validation schemes on an autoregressive series. They consider k-fold cross-validation, "non-dependent cross validation" where partitions do not include adjacent time periods (resulting in many observations left out), leave-one-out cross validation, and temporal block cross-validation (which they call "out of sample," although as we saw for Google Flu Trends in Ginsberg et al., 2009, such blocks may still fail to correspond to out-of-sample performance). For a linear fit to a low-order autoregression series, temporal block cross-validation gives a good estimate of error, whereas k-fold and LOO cross validation schemes give error rates half as high (non-dependent CV gives too-large estimates of error, probably because it has less data available for fitting).

4.5   Linear regression

In a ordinary least squares fit, the predicted values for an unknown $\mathbf{y}^*$ are made from the corresponding feature values $\mathbf{X}^*$ and weights estimated from known $\mathbf{y}$ and corresponding feature values $\mathbf{X}$.

$$\widehat{\boldsymbol{\mu}}_{\mathbf{y};\mathbf{X}}(\mathbf{y}^*; \mathbf{X}^*) = \mathbf{X}^*(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \widehat{\mathbf{y}}^* \tag{8}$$

---

[57]Of course, what counts as nonzero optimism depends on what the "generalized" application is; if we had an imputation problem, random cross-validation would not be overly optimistic in its estimate of error. And for a one-ahead forecasting task, the whole point would be to know how well forecasts can do by leveraging any autocorrelation that is present, perhaps through a 2-nearest-left-neighbor estimator.

[58]This comes from the pseudonymous critique of the paper "Twitter mood predicts the stock market" and similar work in "No limits to garbatrage," *Buy the Hype* blog, August 29, 2013, https://sellthenews.tumblr.com/post/59720892780/no-limits-to-garbatrage.

When we use $\mathbf{X}$ for $\mathbf{X}^*$, we can express the fitted values as $\widehat{\mathbf{y}} = \mathbf{H}\mathbf{y}$ using the hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$.

Normally, we assume $Y \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$. But now let us assume instead that all observations have a correlation ρ, and $Y \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}_{ii} = \sigma^2$ and $\boldsymbol{\Sigma}_{ij} = \rho\sigma^2$ for $i \neq j$, $0 < |\rho_{ij}| < 1$, and $\mathbf{X} \in \mathbb{R}^{n \times (p+1)}$ has a column of all ones.

In order to both keep Y random and to simply the algebra comparing across training and test sets, take the model

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{X} \\ \mathbf{X} \end{bmatrix}\boldsymbol{\beta}, \begin{bmatrix} \boldsymbol{\Sigma} & \rho\sigma^2\mathbf{1}\mathbf{1}^T \\ \rho\sigma^2\mathbf{1}\mathbf{1}^T & \boldsymbol{\Sigma} \end{bmatrix}\right). \tag{9}$$

That is, let there be two exact copies of the feature levels $\mathbf{X}$, what Rosset and Tibshirani (2019) call the "fixed-X" setting. Then we can assign observations $1, ..., n$ to the training set and observations $n+1, ..., 2n$ to the test set, such that the i-th training observation and the i-th test observation will have the same mean, $\mathbf{x}_i^T\boldsymbol{\beta}$. Given Y, the OLS fitted values for the training set are $\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^TY_1 = \mathbf{H}Y_1$. Then, the optimism of the training set error is:

$$\tfrac{2}{n}\operatorname{tr}\operatorname{Cov}_f(Y_1, \widehat{Y}_1) = \tfrac{2}{n}\operatorname{tr}\operatorname{Cov}_f(Y_1, \mathbf{H}Y_1) = \tfrac{2}{n}\operatorname{tr}\mathbf{H}\operatorname{Var}_f(Y_1) = \tfrac{2}{n}\operatorname{tr}\mathbf{H}\boldsymbol{\Sigma} \tag{10}$$

But the test set error, too, is overly optimistic, with optimism:

$$\tfrac{2}{n}\operatorname{tr}\operatorname{Cov}_f(Y_2, \widehat{Y}_1) = \tfrac{2}{n}\operatorname{tr}\operatorname{Cov}_f(Y_2, \mathbf{H}Y_1) = \tfrac{2\rho\sigma^2}{n}\operatorname{tr}\mathbf{H}\mathbf{1}\mathbf{1}^T = 2\rho\sigma^2, \tag{11}$$

with the last step coming from how the row sums of a hat matrix are all one,[59] and the hat matrix times a matrix of all ones is a vector of its row sums repeated column-wise.

Then, the other components of the expected out-of-sample error are the variance of the estimator, $\tfrac{1}{n}\operatorname{tr}\operatorname{Var}_f(\widehat{Y}) = \tfrac{1}{n}\operatorname{tr}\mathbf{H}\boldsymbol{\Sigma}\mathbf{H}^T$, and the irreducible error, $\tfrac{1}{n}\operatorname{tr}\boldsymbol{\Sigma} = \sigma^2$. For the test set error, the contribution of dependencies would be comparable in magnitude to the contribution of the irreducible error (e.g., for $\rho = .5$, the irreducible error would be seemingly canceled out, and as ρ approaches 1, the irreducible error would seemingly be *subtracted* from the test set error as opposed to added), making the test set error a biased estimator of the true out-of-sample error. Alternatively, if ρ were negative (observations are anticorrelated), the test set error would be *upwardly* biased, i.e. an overly *pessimistic* estimator of the true out-of-sample error (the expected error on a new draw of data from the same distribution). But for a multivariate normal distribution of size n with uniform diagonal elements and uniform off-diagonal elements, the correlation ρ is constrained below by $\rho \geq -1/(n-1)$ to ensure the variance-covariance matrix only has positive eigenvalues and the distribution is well-defined. (Intuitively, as n rises, it is less and less possible for all the dimensions of a multivariate normal to mutually have negative linear correlation.) So, negative correlation leading to possible pessimism is not a significant possibility, and does not balance concerns around possible optimism.

Comparing the magnitude of optimism to the variance of the estimator is more difficult. Since $\boldsymbol{\Sigma}$ is not diagonal, $\operatorname{tr}\mathbf{H}\boldsymbol{\Sigma}\mathbf{H}^T = \operatorname{tr}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Sigma}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ does not reduce to $\sigma^2\operatorname{tr}(\mathbf{X}^T\mathbf{X})^{-1}$ as usual, and it is harder to compare the magnitude of this term and determine the impact of the non-zero off-diagonals without having a specific form of $\mathbf{X} \in \mathbb{R}^{n \times (d+1)}$. But, as an example, if $\mathbf{X}$ is a matrix of uncorrelated features,[60] then $\sigma^2\operatorname{tr}(\mathbf{X}^T\mathbf{X})^{-1} = \sigma^2(d + \tfrac{1}{n})$, but $\operatorname{tr}\mathbf{H}\boldsymbol{\Sigma}\mathbf{H}^T = \sigma^2((1-\rho)(d+1) + \rho n)$,

---

[59]In a with-intercept model, the first column of $\mathbf{X}$ is all ones, and $\mathbf{H}$ is the orthogonal projection onto the column space of $\mathbf{X}$. Hence $\mathbf{H}\mathbf{1} = \mathbf{1}$

[60]E.g., as achieved by orthogonal polynomials, where orthogonality for the jth and kth polynomials are defined as $\sum_{i=1}^n p_j(x_i)p_k(x_i) = 0$ and $\sum_{i=1}^n p_j^2(x_i) \neq 0$, and hence the correlation of polynomial feature j and polynomial feature k will be zero. The algorithm for computing such polynomials implemented in R is given in Kennedy and Gentle (1980).

i.e. the variance of the estimator increases with the sample size proportionally to the amount of correlation between observations. This is reasonable, considering how dependencies might "fool" an estimator (consider the variability of a model that is fit to draws like those shown in fig. 6).

Important to note is that, in a multivariate normal response with nonzero positive correlation, *there is no cross-validation scheme that would not lead to an under-estimate of the true error.* More generally, allowing arbitrary variance-covariance structure (i.e., $\Sigma_{ij} = \rho_{ij}\sigma_i\sigma_j$), and also covering the case of a no-intercept model, it would be difficult to say what would happen. The entries of the hat matrix are constrained, with positive diagonal entries generally larger in magnitude than off-diagonal entries (Mohammadi, 2016), but there are negative off-diagonal entries that could interact with arbitrary correlations in arbitrary ways. Then, $\mathbf{X}$ could first be allowed to be random, but across a fixed set of values (like $\mathbf{X}$ being integer-valued within a given range) such that the levels of covariates may be matched up exactly between training and test, what (Tibshirani and Rosset, 2019) call the "same-X" setting, or vary completely (like $\mathbf{X}$ being real-valued rather than integer-values), what they call the "random-X" setting where covariate levels do not match up. Both of these complicate analysis considerably. Working out the non-iid results in the same-X and random-X settings is a task for future analysis, and may contribute to understanding the amount by which we can decrease the bias in test error by designing cross-validation schemes that respect dependency structure.

Simulating from the above model (eqn. 9), with $\mathbf{X} \in \mathbb{R}^{100 \times (20+1)}$ being the orthogonal polynomial features (Kennedy and Gentle, 1980) from degree 0 to degree 20 on $n = 100$ points equally spaced in the interval $[0, 1)$ with gaps of 0.01, $\boldsymbol{\beta} = \mathbf{10}$, $\rho = .5$, and $\sigma^2 = 1$, we can see the effect of correlated observations. Fig. (6) shows what it looks like to have correlated observations split into training and test sets be potentially far from an independent draw.

Fig. (7) shows that this model indeed produces a true out-of-sample error higher in expectation than the average test set error. In particular, the out-of-sample error has only a slightly higher mode, but is right-skewed with large variance,[61] leading to a substantially higher mean squared error than those of the training and test sets: 1.606 versus, respectively, 0.396 and 0.607. Calculating from the above, we have:

- Irreducible error $\operatorname{tr}\boldsymbol{\Sigma}/n = 1$;

- Variance of the estimator $\operatorname{tr}\mathbf{H}\boldsymbol{\Sigma}\mathbf{H}^T/n = 0.605$;

  - Expected true out-of-sample error $1 + 0.605 = 1.605$;

- Optimism in the training error $2\operatorname{tr}\mathbf{H}\boldsymbol{\Sigma}/n = 1.21$;

  - Expected training error $1.605 - 1.21 = 0.395$;

- Optimism of test set $2\rho\sigma^2 = 1$; and

  - Expected test set error $1.605 - 1 = 0.605$,

which agrees with the simulation to two significant figures.

Raising ρ towards 1 makes the distribution of out-of-sample error flatten out even more and have an even longer tail; and, ρ cannot be lowered more that slightly below zero (for $n = 100$, ρ cannot be less than $-1/99$, around 0.01) to keep $\boldsymbol{\Sigma}$ positive definite, a point at which there is not much difference between the test set error and out-of-sample error.

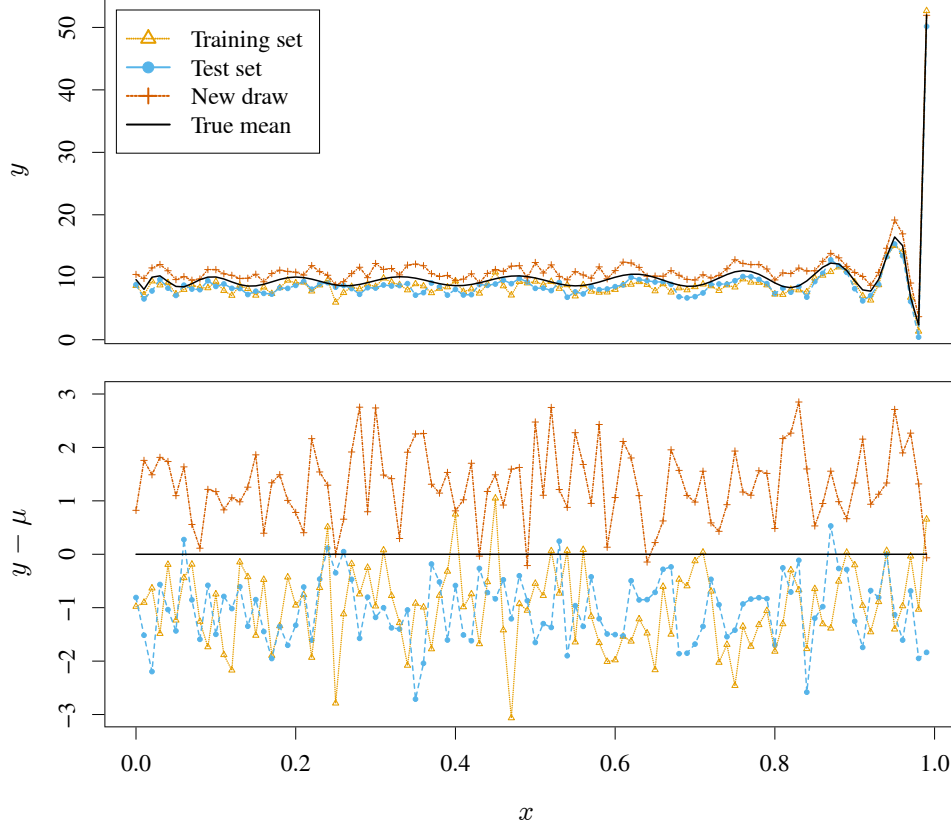[61]The mean squared error here is the OLS residual sum of squares.

Figure 6: Top: the curve generated from polynomial features, along with one draw from eqn. (9) used as a training/test split, and an independent second draw. Bottom: for greater visibility, a plot of the difference between the draws and the true curve. Correlated observations can produce training and test sets close to one another but potentially far from an independent draw. Note that while the correlations between observations in one draw may bias a particular estimate, OLS is still an unbiased estimat*or* because, across draws, the average bias is still zero.

While it would not necessarily give insight into the relative contribution of dependencies to the MSE, it may be possible to derive the theoretical distributions that correspond to those shown in fig. (7), based on results for quadratic forms (Mathai and Provost, 1992) for $\|Y_1 - \mathbf{H}Y_1\|_2^2 = Y_1^T(\mathbf{I} - \mathbf{H})Y_1$, which follows a generalized non-central chi-square distribution (Mohsenipour, 2012), and bilinear forms for $\|Y_2 - \mathbf{H}Y_1\|_2^2$ and $\|Y^* - \mathbf{H}Y_1\|_2^2$ (Mathai and Provost, 1992; Mathai et al., 1995). Unfortunately, there is no closed-form expression for the generalized non-central chi-square distribution, only an MGF; and similarly, for the bilinear forms, there are only expression for calculating cumulants. While these may give more detail about theoretical quantities relating these distributions, this is a fairly narrow example, so I avoid the exercise.

4.6   Consequences

Based on what points are partitioned into a test set, from a known data-generating process we could compute the optimism of the test set error. Better yet, we could try to design cross-validation schema to minimize this optimism.

Temporal autocorrelation, as in the first example above, is the simplest form of dependency. To avoid "time traveling," the best practice is to use temporal block cross-validation (Bergmeir
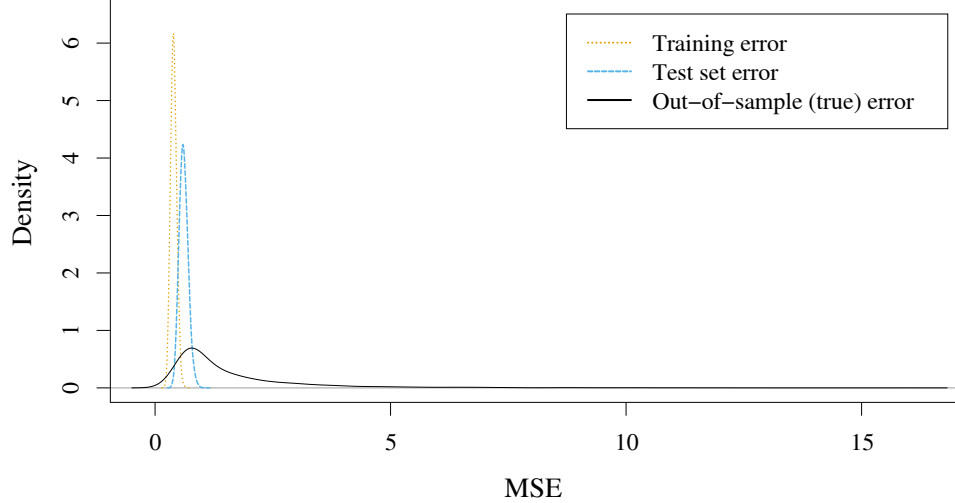
Figure 7: Distribution of training error, test set error, and out-of-sample error over 10,000 replications (5,000 were not sufficient for the average out-of-sample to error converge to its expected value) for the model described in eqn. (9).

and Benítez, 2012), where the test set is a block chronologically after the entire training set. More specifically, Racine (2000) recommends leaving out a block between the training and test sets in order to break the autocorrelation that would persists from having adjacent training and test blocks.

For other simple, regular forms of dependency, it is also possible to combat optimism in the test set error. Hammerla and Plötz (2015) summarize how assessments of model performance in ubiquitous computing activity recognition has been led astray by dependencies that exist between multiple observations of the activity signatures of the same individual; in one case, test set error dropped from over 95% to about 78.2% when going from random data splitting to "leave-one-subject-out" validation.

Network autocorrelation (Dow et al., 1982) gets more complex. If predictions are made about individuals who are connected in a network that causes network neighbors to be similar, having neighboring observations split across a training and test split will create optimism in the test set error. For example, label propagation for nodes on the boundary of any test/train split would have nonzero test set-estimator covariance structure very similar to that of nearest-neighbor regression and the AR(1) process described above. Even if an estimator did not use network features directly (or, indeed, if the network structure is unknown), the network structure may nonetheless make neighboring points more similar, making comparisons of neighbors across the train/test split unfairly positive, and we may not get a reliable picture of how well the estimator will generalize to new points outside of the observed network.

If predictions are made about the network ties themselves, as in link prediction, there are nonetheless dependencies; network processes are dependencies *between* ties, like reciprocity (A_{ij} and A_{ji}), friend-of-a-friend connections (A_{ij}, A_{ik}, and A_{jk}), and cumulative advantage or preferential attachment (A_{ij} and A_{ik}). For either node-level predictions or link prediction, if a network is partitioned into a training subgraph and a testing subgraph, the ties that are excluded are a means of "sharing" information between either the nodes or the ties of the training and rest sets. If there are no dependencies between ties other than, say, latent class membership of the nodes, then such a cross-validation scheme would be valid (Chen and Lei, 2018); otherwise, the test set error undoubtedly suffers from optimism as well, perhaps unavoidably as in the non-independent

multivariate normal case in the second example above.

But in general, we would have to know or estimate the dependency structure to know the covariance between a test set and the estimator. And for a multivariate normal distribution with correlation as in the second example above, even if homoskedastic and with uniform correlations, there is no cross-validation scheme that would allow the test set error to be an unbiased estimate of the true out-of-sample error. If we knew we had such an intractable case, we would have to accept that there would be no avoiding some amount of optimism, or else we would need to collect additional data (rather than do data splitting).

And as Opsomer et al. (2001) point out, without assuming a parametric form for either the mean or covariance functions, models are not identifiable. So, *no amount of additional data is ever enough*; without some independence assumptions about the structure of time, space, individuals, networks, or whatever else, additional observations cannot actually be used to estimate the covariance, because new observations themselves would have covariance that would need further observations to estimate *ad infinitum.*

And unlike wrong assumptions about, say, the model specification or functional form/function class (which machine learning can claim are irrelevant so long as performance is good), being wrong about the dependency structure can bias assessments of that performance itself.

4.7 Alternatives

For making claims about machine learning systems, real-world testing is key.

4.7.1 Experimental testing

An excellent example of testing a machine learning system is given in Cardoso et al. (2016). This was based on an earlier paper, van't Veer et al. (2002), which trained a classifier (which, from the supplementary information, seems like a sort of custom-made decision tree, validated with leave-one-out CV) to find a set of gene expressions out of a set of 5,000 that correlated most highly with developing breast cancer. The 2016 paper took this system and subjected it to not only real out-of-sample testing, but did so in a randomized control trial to test the effect of relying on machine learning for decision-making. Specifically, patients were assessed for their "clinical risk" (traditionally made diagnosis) and their "genomic risk" (based on correlations found by the machine learning model). Patients who were high-risk on both criteria were always assigned chemotherapy, and patients who were low-risk on both criteria were never assigned chemotherapy. But those who were high-risk for one and low-risk for the other (the two tests disagreed) were randomly assigned to chemotherapy. The results: in cases where the "clinical" risk was low but the "genomic" risk was high, chemotherapy actually led to *worse* outcomes! However, in cases where the "clinical" risk was high and the "genomic" risk was low, chemotherapy had no effect on outcomes, in which case avoiding a painful therapy is preferred. That is, *relying only on the originally trained model alone for making chemotherapy decisions would have actually led to worse outcomes.* But, with the results of the experiment, as the authors point out (Cardoso et al., 2016), we know how to use the model: as a second pass by which to catch false positives. Using it thusly, the study suggests, we will be able to avoid subjecting about 46% of women with high clinical breast cancer risk to chemotherapy unlikely to be helpful.

This study is not perfect: the original classifier was trained with the data of only 98 purposively sampled people, compared with the randomized control trial with 6,693 subjects (2,142 of which had discordant results to qualify for being randomly assigned). More data, representatively capturing as diverse a patient pool as possible, would expose the classifier to more variability over which to do

better for non-majority cases. And, as large as was the sample for the experimental study—and done over 112 institutions in nine countries—all of those countries were European, potentially not sufficiently representing non-European populations. Then, rather than relying on a bespoke classifier written and trained in 2002, a better classifier might have been trained using tried-and-tested models in standard modern implementations, perhaps a random forest (Fernández-Delgado et al., 2014; Caruana et al., 2008).

Even though requirements for and norms around medical testing are why there is even such a careful validation of the machine learning, the same regulations are likely part of what is behind the above shortcomings. An updated model would need to be validated before being approved for medical use, making continuous or even periodic updates very difficult.

Ultimately, developing scientific understanding of which genes are causally linked to developing breast cancer would be a more solid basis on which to classify and then take action. But in terms of the use of machine learning itself, this is a good example of a meaningful, valid, and responsible approach.

4.7.2   Careful cross-validation schema

For cross-validation, machine learning should begin to think systematically about how dependencies in data might be affecting the reliability of cross-validation, and develop and adopt cross-validation schemes to address this. Such efforts could draw on existing work on different forms of dependencies (Hammerla and Plötz, 2015; Bergmeir and Benítez, 2012; Chen and Lei, 2018), mixing them and adding in new forms of dependencies.

4.7.3   Rhetorical change

But true out-of-sample testing, and experimentally testing the result of *intervening* on the basis of a result for machine learning, may not be possible, or preliminary results might be needed in order to determine whether such testing is worthwhile. In such a case, there are effective alternatives in terms of communicating the results of machine learning models. There are some basic issues of communications, such as in explaining the "accuracy paradox."[62] The technical meanings of precision and recall are not likely to be colloquially known, and conversely, the technical meaning of prediction in terms of correlations do not correspond to lay understandings of "prediction," suggesting a strategy of giving an explanation of the meanings of these terms along with public presentation of results.

The problem of over-optimism in assessments of model generalizability on the basis of cross-validation can similarly be addressed somewhat just by providing caveats about results being preliminary until experimental testing is done. Indeed, over-optimism is misleading not only to the public, but potentially to researchers as well, who may take their own performance claims as "an assertion about the performance of the system under general conditions" (Cohen and Ruths, 2013) rather than as preliminary claims. Without understanding how out-of-sample testing is the ultimate arbiter, claims may enable the adopting of machine learning systems that turn out to not perform well. Worse, if there is no feedback mechanism for checking model performance in the wild (e.g., if human annotation was necessary for model building, and no further annotation is done once the system is deployed), we may never know that a model is failing to perform well.

What are alternative to current practices? For publication bias, it might be pre-registration as has been adopted in other areas. For overfitting to test data, there are promising technical proposals

---

[62]If the base rate is 1%, a classifier that always guesses the majority class would be useless, but would achieve 99% accuracy.

(Dwork et al., 2015a), but like any procedure there will be trade-offs we may not want to make and assumptions that might fail. A purists' solution would be to set aside data and use it for testing only *once*. That one test is what must be reported; if the test performance turns out to be too low to be publishable (outside of a pre-registration database) or applied, for example, then the whole project must be abandoned (or else a new test set must be gathered). But this is likely neither feasible nor desirable. Held-out data is already wasteful in terms of getting better estimates; the goal of getting a good estimate of the out-of-sample error is important enough to make this sacrifice, but if held-out data was only usable once, it would be too high a price. Here I agree with Irpan (2018): certain costs of reproducibility may be higher than its benefits. A better solution, then, is simply to be more honest about the possibility of overfitting even with data splitting and cross-validation.

Overall, research can acknowledge that statements about "predictions" from cross-validation are post-hoc statements about the strength of correlations (Gayo-Avello, 2012) that are only partially reliable guides to generalizability. In fact, predictions are probably better called "backtesting" or "retrodiction" to avoid colloquial confusion (although the use of these terms alone is not sufficient to prevent modelers from overfitting, as Bailey et al., 2014, argue happens in financial modeling). The use of 'prediction' to refer to fitted values is quite old and inherited from statistics, and so this is not an example of machine learning engaging in "wishful mnemonics" (McDermott, 1976) or misusing language via "choosing terms of art with colloquial connotations" (Lipton and Steinhardt, 2018), but it is nonetheless misleading to the public and something that machine learning practitioners and researchers can change.

In all cases, cross-validation evaluations of model performance should be taken not as a guarantee, but as preliminary. Only true out-of-sample testing, in a setting mimicking what an actual application would be (including, where appropriate or possible, experimental testing), is what can truly determine a machine learning model's success.

5   Summary and conclusion

In this paper, I have attempted to systematically review grounds for the ways in which machine learning models are limited and can fail as applied to social systems. The described hierarchy can be used both for challenging models, for example when they lead to unjust ends, and for modelers to understand the limitations of machine learning, in order to try and mitigate potential failures and to better communicate what modeling actually demonstrates and achieves.

In summarizing the commitments of machine learning, I draw on work from Science and Technology Studies, theories of measurement, and critiques of statistics still applicable to the way machine learning sees the world. Summarizing work on the difference between prediction and explanation (Shmueli, 2010; Breiman, 2001), I make connections both to philosophy (Rescher, 1998) and to current debates about explainability and interpretability (Doshi-Velez and Kim, 2017; Caruana et al., 2015; Lipton, 2015). Lastly, I bring together materials calling into question the reliability of cross-validation estimates of generalizability error, including unifying discussions of how to properly do cross-validation in a framework of non-iid data, demonstrating the effect on test set optimism and showing how issues identified by statisticians in the past can arise in new forms for machine learning.

As mentioned in the introduction, systematically developing ways of integrating other methods into machine learning project pipelines, or even making true mixed-methods pipelines, represents the most promising way forward of overcoming limitations and failure points of machine learning-only research. At every point where machine learning can fail, there are alternatives that will not fail;

conversely, places where other methods fail may be places where machine learning can provide a way forward.

In terms of alternatives, my work suggests the following tasks:

- Systematically incorporate qualitative methods;

- Seek ways of including variability, or higher moments, as one way of mitigating a reliance only on the central tendency corresponding to the first moment;

- Split data for cross-validation in ways that take dependencies into account;

- Do out-of-sample, real-world testing before making strong claims about model performance (and, conversely, demand such testing before accepting strong claims);

- Be aware, open, and humble about the limitations of machine learning modeling.

That is, machine learning modelers should also consider how some tasks might be better addressed with statistics, other forms of mathematical modeling, or even qualitative research.

But returning to the cautionary point in the introduction, mixed methods are not easy to do. A researcher trained in machine learning should not think there is some one book they can read and then be qualified to undertake an ethnography, or even to design survey questions, any more than the converse. Seawright (2016)[63] points to another subtlety of mixed methods: specifically, he argues *against* a parallel or linear "triangulation" approach to mixed methods, which tries to answer the same question with different methods in order to try and get convergence, noting that since "qualitative and statistical approaches produce results that are different in kind, it is only possible to assess such convergence very abstractly." He instead advocates for an integrative approach of putting different methods in sequence (potentially iteratively), such as statistically estimating coefficients "from a model designed, refined, and tested in light of serious qualitative analysis." Similarly, machine learning could try to train predictive models designed, refined, and tested in light of serious qualitative inquiry. Or, perhaps taking into account the history of ethical failing in qualitative research that was not based on reciprocal exchange (Smith, 2012), models could be designed, refined, and tested in light of meaningful participation of human subjects (meaningful in the sense of having veto power, and not just serving perfunctorily as non-binding consultation) over what labels are used, for building which models, to accomplish what ends. McQuillan (2018) proposes popular assemblages, in the model of patients' councils, as an example of the institutional form such input might take for machine learning. Dunning (2008)[64] also talks about the role of qualitative methods in the design of experiments, which we could also apply to experimental testing of machine learning like that of Cardoso et al. (2016).

In the short term, attempts at machine learning that engage in good faith with the limitations of modeling are better positioned to identify and avoid pitfalls of unjust development and application. They will also avoid contributing to unrealistic public expectations that could harm the vitality of the field. A healthy skepticism in approaching modeling can therefore be a central part of developing a community of responsible practice. In the long term, developing frameworks, pipelines, best practices, and experience in actually incorporating alternative methods into uses of machine learning holds promise for reliably interacting with the world to improve the human condition, free of unintentional or unanticipated harms.

---

[63]I thank Baobao Zhang for this reference.
[64]I thank Baobao Zhang for this reference as well.

Conflict of Interest Statement

Author Contributions

MMM did all work enclosed, with inputs from other individuals given in footnotes.

Funding

Acknowledgments

References

Abbott, Andrew. Transcending general linear reality. *Sociological Theory*, 6(2):169–186, 1988. https://dx.doi.org/10.2307/202114.

Agre, Philip E. Towards a critical technical practice: Lessons learned from trying to reform AI. In Geoffrey C. Bowker, Susan Leigh Star, Will Turner, and Les Gasser, editors, *Social science, technical systems, and cooperative work: Beyond the great divide*, pages 131–158. Lawrence Erlbaum Associates, 1997. https://web.archive.org/web/20040203070641/http://polaris.gseis.ucla.edu/pagre/critical.html.

Alexandrova, Anna. Subjective well-being and Kahneman's 'objective happiness'. *Journal of Happiness Studies*, 6(3):301–324, 2005. https://dx.doi.org/10.1007/s10902-005-7694-x.

Alexandrova, Anna. When analytic narratives explain. *Journal of the Philosophy of History*, 3(1): 1–24, 2009. https://dx.doi.org/10.1163/187226309X408776.

Alexandrova, Anna and Robert Northcott. It's just a feeling: Why economic models do not explain. *Journal of Economic Methodology*, 20(3):262–267, 2013. https://dx.doi.org/10.1080/1350178X.2013.828873.

Altenburger, Kristen M. and Daniel E. Ho. Is Yelp actually cleaning up the restaurant industry? A re-analysis on the relative usefulness of consumer reviews. In *Proceedings of the 2019 World Wide Web Conference*, WWW '19, pages 2543–2550, 2019. https://dx.doi.org/10.1145/3308558.3313683.

Arceneaux, Kevin, Alan S. Gerber, and Donald P. Green. A cautionary note on the use of matching to estimate causal effects: An empirical example comparing matching estimates to an experimental benchmark. *Sociological Methods & Research*, 39(2):256–282, 2010. https://dx.doi.org/10.1177/0049124110378098.

Arlot, Sylvain and Alain Celisse. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79, 2010. https://dx.doi.org/10.1214/09-SS054.

Aronow, Peter M. and Cyrus Samii. Does regression produce representative estimates of causal effects? *American Journal of Political Science*, 60(1):250–267, 2016. https://dx.doi.org/10.1111/ajps.12185.

Attia, Mariam and Julian Edge. Be(com)ing a reflexive researcher: A developmental approach to research methodology. *Open Review of Educational Research*, 4(1):33–45, 2017. https://dx.doi.org/10.1080/23265507.2017.1300068.

Bailey, David H., Jonathan M. Borwein, Marcos López de Prado, and Qiji Jim Zhu. Pseudo-mathematics and financial charlatanism: The effects of backtest overfitting on out-of-sample performance. *Notices of the AMS*, 61(5):458–471, 2014. https://dx.doi.org/10.1090/noti1105.

Barabas, Chelsea, Madars Virza, Karthik Dinakar, Joichi Ito, and Jonathan Zittrain. Interventions over predictions: Reframing the ethical debate for actuarial risk assessment. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, Proceedings of Machine Learning Research 81, pages 62–76, 2018. http://proceedings.mlr.press/v81/barabas18a.html.

Barnes, Brooks. Solving equation of a hit film script, with data. *The New York Times*, 5 May 2013. https://nyti.ms/17GKrA7.

Battle-Baptiste, Whitney and Britt Rusert, editors. *W. E. B. Du Bois's data portraits: Visualizing Black America.* Princeton Architectural Press, 2018.

Baumer, Eric P. S. and M. Six Silberman. When the implication is not to design (technology). In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 2271–2274, 2011. https://dx.doi.org/10.1145/1978942.1979275.

Bellotti, Elisa. *Qualitative networks: Mixed methods in sociological research.* Routledge, 2015.

Benjamin, Ruha. *Race after technology: Abolitionist tools for the New Jim Code.* Polity Press, 2019.

Bergmeir, Christoph and José M. Benítez. On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191:192–213, 2012. https://dx.doi.org/10.1016/j.ins.2011.12.028.

Bergmeir, Christoph, Rob J. Hyndman, and Bonsoo Koo. A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, 120:70–83, 2018. https://dx.doi.org/https://doi.org/10.1016/j.csda.2017.11.003.

Bhandari Neupane, Jayanti, Ram P. Neupane, Yuheng Luo, Wesley Y. Yoshida, Rui Sun, and Philip G. Williams. Characterization of leptazolines A-D, polar oxazolines from the cyanobacterium *Leptolyngbya* sp., reveals a glitch with the "Willoughby-Hoye" scripts for calculating NMR chemical shifts. *Organic Letters*, 21(20):8449–8453, 2019. https://dx.doi.org/10.1021/acs.orglett.9b03216.

Bian, Jiang, Iain Buchan, Yi Guo, and Mattia Prosperi. Statistical thinking, machine learning. *Journal of Clinical Epidemiology*, 116:136–137, 2019. https://dx.doi.org/https://doi.org/10.1016/j.jclinepi.2019.08.003.

Bickel, Peter J., Eugene A. Hammel, and J. W. O'Connell. Sex bias in graduate admissions: Data from Berkeley. *Science*, 187(4175):398–404, 1975. https://dx.doi.org/10.1126/science.187.4175.398.

Boden, Margaret A. *Mind as machine: A history of cognitive science*. Oxford University Press, 2006.

Borgatti, Stephen P., Ajay Mehra, Daniel J. Brass, and Giuseppe Labianca. Network analysis in the social sciences. *Science*, 323(5916):892–895, 2009. https://dx.doi.org/10.1126/science.1165821.

Bouk, Dan. *How our days became numbered: Risk and the rise of the statistical individual*. The University of Chicago Press, 2015.

Box, George E. P. Robustness in the strategy of scientific model building. Technical Report #1954, Mathematics Research Center, University of Wisconsin-Madison, May 1979.

Breiman, Leo. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231, 2001. https://dx.doi.org/10.1214/ss/1009213726.

Buja, Andreas, Richard Berk, Lawrence Brown, Edward George, Emil Pitkin, Mikhail Traskin, Linda Zhao, and Kai Zhang. Models as approximations: How random predictors and model violations invalidate classical inference in regression. arXiv:1404.1578v2, 2014. https://arxiv.org/abs/1404.1578v2.

Bzdok, Danilo, Naomi Altman, and Martin Krzywinski. Statistics versus machine learning. *Nature Methods*, 15(4):233–234, 2018. https://dx.doi.org/10.1038/nmeth.4642.

Callon, Michel. The embeddedness of economic markets in economics. In Michel Callon, editor, *The laws of the market*, pages 1–57. Blackwell, 1998.

Campbell, Donald T. Assessing the impact of planned social change. In Gene M. Lyons, editor, *Social research and public policies: The Dartmouth / OECD Conference*, pages 3–45. The Public Affairs Center, Dartmouth College, 1975.

Cardoso, Fatima, Laura J. van't Veer, Jan Bogaerts, Leen Slaets, Giuseppe Viale, Suzette Delaloge, Jean-Yves Pierga, Etienne Brain, Sylvain Causeret, Mauro DeLorenzi, Annuska M. Glas, Vassilis Golfinopoulos, Theodora Goulioti, Susan Knox, Erika Matos, Bart Meulemans, Peter A. Neijenhuis, Ulrike Nitz, Rodolfo Passalacqua, Peter Ravdin, Isabel T. Rubio, Mahasti Saghatchian, Tineke J. Smilde, Christos Sotiriou, Lisette Stork, Carolyn Straehle, Geraldine Thomas, Alastair M. Thompson, Jacobus M. van der Hoeven, Peter Vuylsteke, René Bernards, Konstantinos Tryfonidis, Emiel Rutgers, and Martine Piccart. 70-gene signature as an aid to treatment decisions in early-stage breast cancer. *New England Journal of Medicine*, 375(8): 717–729, 2016. https://dx.doi.org/10.1056/NEJMoa1602253.

Carr, Nicholar. The limits of social engineering. *MIT Technology Review*, May/June, 16 April 2014. https://www.technologyreview.com/s/526561/the-limits-of-social-engineering/.

Caruana, Rich, Nikos Karampatziakis, and Ainur Yessenalina. An empirical evaluation of supervised learning in high dimensions. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 96–103, 2008. https://dx.doi.org/10.1145/1390156.1390169.

Caruana, Rich, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 1721–1730, 2015. https://dx.doi.org/10.1145/2783258.2788613.

Cawley, Gavin C. and Nicola L. C. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11:2079–2107, 2010. http://www.jmlr.org/papers/v11/cawley10a.html.

Chen, Kehui and Jing Lei. Network cross-validation for determining the number of communities in network data. *Journal of the American Statistical Association*, 113(521):241–251, 2018. https://dx.doi.org/10.1080/01621459.2016.1246365.

Christodoulou, Evangelia, Jie Ma, Gary S. Collins, Ewout W. Steyerberg, Jan Y. Verbakel, and Ben Van Calster. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology*, 110:12–22, 2019. https://dx.doi.org/https://doi.org/10.1016/j.jclinepi.2019.02.004.

Chu, Chih-Kang and James Stephen Marron. Comparison of two bandwidth selectors with dependent errors. *The Annals of Statistics*, 19(4):1906–1918, 1991. https://dx.doi.org/10.1214/aos/1176348377.

Clauset, Aaron, Cosma Rohilla Shalizi, and Mark E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009. https://dx.doi.org/10.1137/070710111.

Cohen, Raviv and Derek Ruths. Classifying political orientation on Twitter: It's not easy! In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, ICWSM-13, pages 91–99, 2013. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6128.

Collins, Harry. Actors' and analysts' categories in the social analysis of science. In Peter Meusburger, Michael Welker, and Edgar Wunder, editors, *Clashes of knowledge: Orthodoxies and heterodoxies in science and religion*, pages 101–110. Springer, 2008.

Collins, Henry M. Son of seven sexes: The social destruction of a physical phenomenon. *Social Studies of Science*, 11(1):33–62, 1981. https://dx.doi.org/10.1177/030631278101100103.

Costanza-Chock, Sasha. Design justice, A.I., and escape from the matrix of domination. *Journal of Design and Science*, July 2018. https://dx.doi.org/10.21428/96c8d426.

Costanza-Chock, Sasha, Maya Wagoner, Berhan Taye, Caroline Rivas, Chris Schweidler, Georgia Bullen, and the Tech for Social Justice Project. #MoreThanCode: Practitioners reimagine the landscape of technology for justice and equity. Technical report, Research Action Design & Open Technology Institute, 2018. https://morethancode.cc.

Cowgill, Bo. The impact of algorithms on judicial discretion: Evidence from regression discontinuities, 2018. http://www.columbia.edu/~bc2656/papers/RecidAlgo.pdf.

Cox, David R. Role of models in statistical analysis. *Statistical Science*, 5(2):169–174, 1990. https://dx.doi.org/10.1214/ss/1177012165.

Cox, David R. Big data and precision. *Biometrika*, 102(3):712–716, 2015. https://dx.doi.org/10.1093/biomet/asv033.

Cunliffe, Ann L. "On becoming a critically reflexive practitioner," redux: What does it mean to be reflexive? *Journal of Management Education*, 40(6):740–746, 2016. https://dx.doi.org/10.1177/1052562916668919.

Daston, Lorraine. *Classical probability in the Enlightenment*. Princeton University Press, 1989.

Dawid, Philip. On individual risk. *Synthese*, 194(9):3445–3474, 2017. https://dx.doi.org/10.1007/s11229-015-0953-4.

Deaton, Angus. Instruments, randomization, and learning about development. In Dawn Langan Teele, editor, *Field experiments and their critics: Essays on the uses and abuses of experimentation in the social sciences*, pages 141–184. Yale University Press, 2014.

Desrosières, Alain. Mapping the social world: From aggregates to individual. *Limn*, 2, 2012. https://limn.it/articles/mapping-the-social-world-from-aggregates-to-individuals/.

DeVellis, Robert F. *Scale development: Theory and applications*. SAGE, 4th edition, 2017.

Dietterich, Thomas G. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1923, 1998. https://dx.doi.org/10.1162/089976698300017197.

Dodge, Jesse, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, EMNLP 2019, pages 2185–2194, 2019. https://dx.doi.org/10.18653/v1/D19-1224.

Domingos, Pedro. A unified bias-variance decomposition and its applications. In *Proceedings of the 17th International Conference on Machine Learning*, ICML 2000, pages 231–238, 2000.

Domingos, Pedro. A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87, 2012. https://dx.doi.org/10.1145/2347736.2347755.

Donnelly, Kevin. *Adolph Quetelet, social physics, & the average men of science, 1796-1874*. University of Pittsburgh Press, 2016. http://www.jstor.org/stable/j.ctt1dfnth2.

Doshi-Velez, Finale and Been Kim. Towards a rigorous science of interpretable machine learning. arXiv:1702.08608, 2017. https://arxiv.org/abs/1702.08608.

Dow, Malcolm M., Michael L. Burton, and Douglas R. White. Network autocorrelation: A simulation study of a foundational problem in regression and survey research. *Social Networks*, 4 (2):169–200, 1982. https://dx.doi.org/10.1016/0378-8733(82)90031-4.

Du Bois, W. E. B. *The Philadelphia Negro: A social study*. Schocken Books, 1899. https://archive.org/details/philadelphianegr001901mbp.

Du Bois, W. E. B. The Georgia Negro: A social study, 1900.
https://www.loc.gov/pictures/item/2005679642. Charts and graphs showing the condition
of African Americans at the turn of the century exhibited at the Paris Exposition Universelle in
1900.

Dunning, Thad. Natural and field experiments: The role of qualitative methods. *Qualitative &
Multi-Method Research*, 6(2), 2008. https://dx.doi.org/10.5281/zenodo.996336.

Dwork, Cynthia, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth.
The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638,
2015a. https://dx.doi.org/10.1126/science.aaa9375.

Dwork, Cynthia, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Leon
Roth. Preserving statistical validity in adaptive data analysis. In *Proceedings of the
Forty-Seventh Annual ACM Symposium on Theory of Computing*, STOC '15, pages 117–126,
2015b. https://dx.doi.org/10.1145/2746539.2746580.

Efron, Bradley. The estimation of prediction error: Covariance penalties and cross-validation.
Journal of the American Statistical Association, 99(467):619–632, 2004.
https://dx.doi.org/10.1198/016214504000000692.

Efron, Bradley and Trevor Hastie. *Computer age statistical inference: Algorithms, evidence, and
data science.* Cambridge University Press, 2016.

Elliott, Delbert S. Lies, damn lies and arrest statistics. Technical Report CSPV-015, Center for the
Study and Prevention of Violence, Institute of Behavioral Science, University of Colorado,
Boulder, 1995.

Erikson, Emily. Formalist and relationalist theory in social network analysis. *Sociological Theory*,
31(3):219–242, 2013. https://dx.doi.org/10.1177/0735275113501998.

Eubanks, Virginia. *Automating inequality: How high-tech tools profile, police, and punish the poor.*
St. Martin's Press, 2018.

Eykholt, Kevin, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul
Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning
visual classification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern
Recognition*, pages 1625–1634, 2018. https://dx.doi.org/10.1109/CVPR.2018.00175.

Faigman, David L., John Monahan, and Christopher Slobogin. Group to individual (G2I) inference
in scientific expert testimony. *University of Chicago Law Review*, 81(1):417–480, 2014.
https://chicagounbound.uchicago.edu/uclrev/vol81/iss2/1.

Fan, Jianqing and Wen-Xin Zhou. Guarding against spurious discoveries in high dimensions.
Journal of Machine Learning Research, 17(1):7123–7156, 2016.
http://jmlr.org/papers/v17/16-068.html.

Fergus, Devin. The ghetto tax: Auto insurance, postal code profiling, and the hidden history of
wealth transfer. In Fredrick C. Harris and Robert C. Lieberman, editors, *Beyond discrimination:
Racial inequality in a postracial era*, pages 277–316. Russel Sage Foundation, 2013.

Fernández-Delgado, Manuel, Eva Cernadas, Senén Barro, and Dinani Amorim. Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, 15(1):3133–3181, 2014. http://jmlr.org/papers/v15/delgado14a.html.

Fischer, David Hackett. *Historians' fallacies: Toward a logic of historical thought.* Harper & Row, 1970.

Fisher, Ronald A. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 222:309–368, 1922. https://dx.doi.org/10.1098/rsta.1922.0009.

Fourcade, Marion and Kieran Healy. Classification situations: Life-chances in the neoliberal era. *Accounting, Organizations and Society*, 38(8):559–572, 2013. https://dx.doi.org/10.1016/j.aos.2013.11.002.

Freedman, David A. From association to causation via regression. *Advances in Applied Mathematics*, 18(1):59–110, 1997. https://dx.doi.org/10.1006/aama.1996.0501.

Freedman, David A. Graphical models for causation, and the identification problem. *Evaluation Review*, 28(4):267–293, 2004. https://dx.doi.org/10.1177/0193841X04266432.

Freedman, David A. Linear statistical models for causation: A critical review. In Brian S. Everitt and David C. Howell, editors, *Encyclopedia of statistics in behavioral science*, pages 1061–1073. John Wiley & Sons, Ltd, 2005. https://dx.doi.org/10.1002/0470013192.bsa598.

Freedman, David A. *Statistical models and causal inference: A dialogue with the social sciences.* Cambridge University Press, 2009.

Freire, Paulo. *Pedagogy of the oppressed.* Continuum, 1970.

Frey, William R., Desmond U. Patton, Michael B. Gaskell, and Kyle A. McGregor. Artificial intelligence and inclusion: Formerly gang-involved youth as domain experts for analyzing unstructured Twitter data. *Social Science Computer Review*, 38(1):42–56, 2018. https://dx.doi.org/10.1177/0894439318788314.

Friedman, Jerome H. On bias, variance, 0/1-loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery*, 1(1):55–77, 1997. https://dx.doi.org/10.1023/A:1009778005914.

Friedman, Milton. *Essays in positive economics.* University of Chicago Press, 1953.

Gayo-Avello, Daniel. No, you cannot predict elections with Twitter. *IEEE Internet Computing*, 16 (6):91–94, 2012. https://dx.doi.org/10.1109/MIC.2012.137.

Geiger, R. Stuart, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang. Garbage in, garbage out? Do machine learning application papers in social computing report where human-labeled training data comes from? In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '20)*, 2020. https://dx.doi.org/10.1145/3351095.3372862.

Gelman, Andrew. A statistician's perspective on *Mostly harmless econometrics: An empiricist's companion*, by Joshua D. Angrist and Jörn-Steffen Pischke. *Stata Journal*, 9(2):315–320, 2009. http://www.stata-journal.com/article.html?article=gn0046.

Gelman, Andrew and Eric Loken. The statistical crisis in science. *American Scientist*, 102(6):460, 2014. https://dx.doi.org/10.1511/2014.111.460.

Gelman, Andrew and Cosma Rohilla Shalizi. Philosophy and the practice of Bayesian statistics in the social sciences. *British Journal of Mathematical and Statistical Psychology*, 66(1):8–38, 2012. https://dx.doi.org/10.1111/j.2044-8317.2011.02037.x.

Gelman, Andrew, Jeffrey Fagan, and Alex Kiss. An analysis of the New York City Police Department's "stop-and-frisk" policy in the context of claims of racial bias. *Journal of the American Statistical Association*, 102(479):813–823, 2007. https://dx.doi.org/10.1198/016214506000001040.

Gibbs, Fred and Trevor Owens. The hermeneutics of data and historical writing. In Jack Dougherty and Kristen Nawrotzki, editors, *Writing history in the digital age*, pages 159–170. The University of Michigan Press, 2013. https://dx.doi.org/10.3998/dh.12230987.0001.001.

Gijsbers, Pieter, Erin LeDell, Janek Thomas, Sébastien Poirier, Bernd Bischl, and Joaquin Vanschoren. An open source AutoML benchmark. In *6th ICML Workshop on Automated Machine Learning*, 2019. https://www.automl.org/wp-content/uploads/2019/06/automlws2019_Paper45.pdf.

Gilbert, Nigel and Klaus G. Troitzsch. *Simulation for the social scientist.* Open University Press, 2nd edition, 2005.

Ginsberg, Jeremy, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457: 1012–1015, 2009. https://dx.doi.org/10.1038/nature07634.

Gondara, Lovedeep. Classifier comparison using precision. arXiv:1609.09471, 2016. https://arxiv.org/abs/1609.09471.

Goodfellow, Ian, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. arXiv:1412.6572, 2015. https://arxiv.org/abs/1412.6572.

Goodhart, Charles A. E. *Problems of monetary management: The UK experience*, chapter 4, pages 1–20. Reserve Bank of Australia, 1975. https://dx.doi.org/10.1007/978-1-349-17295-5_4.

Gray, Mary L. and Siddharth Suri. *Ghost work: How to stop Silicon Valley from building a new global underclass.* Houghton Mifflin Harcourt, 2019.

Green, Ben and Yiling Chen. Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 90–99, 2019. https://dx.doi.org/10.1145/3287560.3287563.

Hacking, Ian. *The taming of chance.* Cambridge University Press, 1990.

Hammerla, Nils Y. and Thomas Plötz. Let's (not) stick together: Pairwise similarity biases cross-validation in activity recognition. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, pages 1041–1051, 2015. https://dx.doi.org/10.1145/2750858.2807551.

Hannah-Moffat, Kelly. Actuarial sentencing: An "unsettled" proposition. *Justice Quarterly*, 30(2): 270–296, 2013. https://dx.doi.org/10.1080/07418825.2012.682603.

Harford, Tim. Big data: Are we making a big mistake? *Significance*, December 2014:14–29, 2014.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: Data mining, inference and prediction*. Springer, 2nd edition, 2009. https://dx.doi.org/10.1007/b94608.

Healy, Kieran. The performativity of networks. *European Journal of Sociology*, 56(2):175–205, 2015. https://dx.doi.org/10.1017/S0003975615000107.

Hellman, Deborah. *When is discrimination wrong?* Harvard University Press, 2008.

Hellman, Deborah. What is discrimination, when is it wrong and why? In *Proceedings of Machine Learning Research 81*, Conference on Fairness, Accountability, and Transparency, page 1, 2018.

Herndon, Thomas, Michael Ash, and Robert Pollin. Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff. *Cambridge Journal of Economics*, 38(2): 257–279, 2013. https://dx.doi.org/10.1093/cje/bet075.

Holland, Paul W. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986. https://dx.doi.org/10.1080/01621459.1986.10478354.

Horan, Caley. *Actuarial age: Insurance and the emergence of neoliberalism in the postwar United States*. PhD thesis, University of Minnesota, 2011.

Hu, Lily. Disparate causes, pt. I. *Phenomenal World*, 11 October 2019a. https://phenomenalworld.org/analysis/disparate-causes-i.

Hu, Lily. Disparate causes, pt. II. *Phenomenal World*, 17 October 2019b. https://phenomenalworld.org/analysis/disparate-causes-ii.

Humphreys, Paul and David Freedman. The grand leap. *The British Journal for the Philosophy of Science*, 47(1):113–123, 1996. https://dx.doi.org/10.1093/bjps/47.1.113.

Hutson, Matthew. Artificial intelligence faces reproducibility crisis. *Science*, 359(6377):725–726, 2018. https://dx.doi.org/10.1126/science.359.6377.725.

Hutson, Stu. Data handling errors spur debate over clinical trial. *Nature Medicine*, 16(6):618, 2010. https://dx.doi.org/10.1038/nm0610-618a.

Irpan, Alex. Some papers don't reproduce. Should we care? In *ICML 2018: The Debates, at the 35th International Conference on Machine Learning*, ICML 2018, 2018.

Jacobs, Abigail Z. and Hanna Wallach. Measurement and fairness. arXiv:1912.05511, 2019. https://arxiv.org/abs/1912.05511.

Jones, Matthew L. How we became instrumentalists (again): Data positivism since World War II. *Historical Studies in the Natural Sciences*, 48(5):673–684, 2018. https://dx.doi.org/10.1525/hsns.2018.48.5.673.

Karraker, Amelia and Kenzie Latham. Authors' explanation of the retraction. *Journal of Health and Social Behavior*, 56(3):417–419, 2015. https://dx.doi.org/10.1177/0022146515595817.

Kass, Robert E. Statistical inference: The big picture. *Statistical Science*, 26(1):1–9, 2011. https://dx.doi.org/10.1214/10-STS337.

Katz, Yarden. Manufacturing an Artificial Intelligence revolution, November 2017. https://ssrn.com/abstract=3078224.

Kennedy, William J., Jr. and James E. Gentle. *Statistical computing*. Marcel Dekker, Inc., 1980.

Kennefick, Daniel. Star crushing: Theoretical practice and the theoreticians' regress. *Social Studies of Science*, 30(1):5–40, 2000. https://dx.doi.org/10.1177/030631200030001001.

Keyes, Os. The misgendering machines: Trans/HCI implications of automatic gender recognition. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):88:1–88:22, 2018. https://dx.doi.org/10.1145/3274357.

Keynes, John Maynard. Professor Tinbergen's method. *The Economic Journal*, 49(195):558–577, 1939. https://dx.doi.org/10.1093/ej/49.195.558.

Kiviat, Barbara. *Prediction and the moral order: Contesting fairness in consumer data capitalism*. PhD thesis, Department of Sociology, Harvard University, 2019.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. Prediction policy problems. *American Economic Review*, 105(5):491–95, 2015. https://dx.doi.org/10.1257/aer.p20151023.

Kleinman, Arthur and Joan Kleinman. Suffering and its professional transformation: Toward an ethnography of interpersonal experience. In *Writing at the margin: Discourse between anthropology and medicine*, pages 95–119. University of California Press, 1995.

Kohler-Hausmann, Issa. Eddie Murphy and the dangers of counterfactual causal thinking about detecting racial discrimination. *Northwestern University Law Review*, 113(5):1163–1228, 2019. https://scholarlycommons.law.northwestern.edu/nulr/vol113/iss5/6/.

Kolaczyk, Eric D. and Gábor Csárdi. *Statistical analysis of network data with R*. Use R! Springer, 2014. https://dx.doi.org/10.1007/978-1-4939-0983-4.

Koller, Daphne and Nir Friedman. *Probabilistic graphical models: Principles and techniques*. The MIT Press, 2009.

Koplenig, Alexander. The impact of lacking metadata for the measurement of cultural and linguistic change using the Google Ngram data sets—Reconstructing the composition of the German corpus in times of WWII. *Digital Scholarship in the Humanities*, 32(1):169–188, 2015. https://dx.doi.org/10.1093/llc/fqv037.

Korzybski, Alfred. *Science and sanity: An introduction to non-Aristotelian systems and general semantics*. International Non-Aristotelian Library, 1933.

Krauss, Steven Eric. Research paradigms and meaning making: A primer. *The Qualitative Report*, 10(4):758–770, 2005. https://nsuworks.nova.edu/tqr/vol10/iss4/7/.

Kunst, Robert M. Cross validation of prediction models for seasonal time series by parametric bootstrapping. *Austrian Journal of Statistics*, 37(3&4):271–284, 2008. https://dx.doi.org/10.17713/ajs.v37i3&4.308.

Lanius, Candice. Fact check: Your demand for statistical proof is racist. *Cyborgology* [blog], 12 January 2015. https://thesocietypages.org/cyborgology/2015/01/12/fact-check-your-demand-for-statistical-proof-is-racist/.

Lauer, Josh. *Creditworthy: A history of consumer surveillance and financial identity in America.* Columbia University Press, 2017.

Lavine, Michael. Frequentist, Bayes, or other? *The American Statistician*, 73(sup1):312–318, 2019. https://dx.doi.org/10.1080/00031305.2018.1459317.

Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani. The parable of Google Flu: Traps in big data analysis. *Science*, 343(6176):1203–1205, 2014. https://dx.doi.org/10.1126/science.1248506.

Letham, Benjamin, Cynthia Rudin, Tyler H. McCormick, and David Madigan. Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics*, 9(3):1350–1371, 2015. https://dx.doi.org/10.1214/15-AOAS848.

Lipton, Zachary C. The myth of model interpretability. *KDnuggets*, 15(13), 2015. https://www.kdnuggets.com/2015/04/model-interpretability-neural-networks-deep-learning.html.

Lipton, Zachary C. The mythos of model interpretability. *Communications of the ACM*, 61(10): 36–43, 2018. https://dx.doi.org/10.1145/3233231.

Lipton, Zachary C. and Jacob Steinhardt. Troubling trends in machine learning scholarship. arXiv:1807.03341, 2018. https://arxiv.org/abs/1807.03341.

Mackenzie, Adrian. *Machine learners: Archaeology of a data practice.* The MIT Press, 2017.

MacKenzie, Donald A. From Kwajalein to Armageddon? Testing and the social construction of missile accuracy. In David Gooding, Trevor Pinch, and Simon Schaffer, editors, *The uses of experiment: Studies in the natural sciences*, pages 409–435. Cambridge University Press, 1989.

MacKenzie, Donald A. *An engine, not a camera: How financial models shape markets.* MIT Press, 2006.

MacKenzie, Donald A. and Yuval Millo. Constructing a market, performing theory: The historical sociology of a financial derivatives exchange. *American Journal of Sociology*, 109(1):107–145, 2003. https://dx.doi.org/10.1086/374404.

Macy, Michael W. and Robert Willer. From factors to actors: Computational sociology and agent-based modeling. *Annual Review of Sociology*, 28(1):143–166, 2002. https://dx.doi.org/10.1146/annurev.soc.28.110601.141117.

Malik, Momin M. *Bias and beyond in digital trace data.* PhD thesis, Carnegie Mellon University, August 2018. http://reports-archive.adm.cs.cmu.edu/anon/isr2018/abstracts/18-105.html.

Malik, Momin M. and Jürgen Pfeffer. Identifying platform effects in social media data. In *Proceedings of the Tenth International AAAI Conference on Web and Social Media*, ICWSM-16, pages 241–249, 2016. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13163.

Mallows, Colin. The zeroth problem. *The American Statistician*, 52(1):1–9, 1998. https://dx.doi.org/10.1080/00031305.1998.10480528.

Marcus, Gary. Deep learning: A critical appraisal. arXiv:1801.00631, 2018. https://arxiv.org/abs/1801.00631.

Mathai, A. M. and Serge B. Provost. *Quadratic forms in random variables: Theory and applications*. Marcel Dekker, Inc., 1992.

Mathai, A. M., Serge B. Provost, and Takesi Hayakawa. *Bilinear Forms and Zonal Polynomials*. Springer, 1995. https://dx.doi.org/10.1007/978-1-4612-4242-0.

McDermott, Drew. Artificial intelligence meets natural stupidity. *SIGART Bulletin*, (57):4–9, April 1976. https://dx.doi.org/10.1145/1045339.1045340.

McNemar, Quinn. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947. https://dx.doi.org/10.1007/BF02295996.

McQuillan, Dan. People's councils for ethical machine learning. *Social Media + Society*, 4(2): 2056305118768303, 2018. https://dx.doi.org/10.1177/2056305118768303.

Meinshausen, Nicolai and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010. https://dx.doi.org/10.1111/j.1467-9868.2010.00740.x.

Meng, Xiao-Li. Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. *The Annals of Applied Statistics*, 12(2): 685–726, 2018. https://dx.doi.org/10.1214/18-AOAS1161SF.

Merali, Zeeya. Computational science: ...error. *Nature*, 467(7317):775–777, 2010. https://dx.doi.org/10.1038/467775a.

Messerli, Franz H. Chocolate consumption, cognitive function, and Nobel laureates. *New England Journal of Medicine*, 367(16):1562–1564, 2012. https://dx.doi.org/10.1056/NEJMon1211064.

Michell, Joel. *Measurement in psychology: Critical history of a methodological concept*. Cambridge University Press, 1999.

Miller, Greg. A scientist's nightmare: Software problem leads to five retractions. *Science*, 314 (5807):1856–1857, 2006. https://dx.doi.org/10.1126/science.314.5807.1856.

Miller, Tim. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019. https://dx.doi.org/10.1016/j.artint.2018.07.007.

Mitchell, Tom M. *Machine learning*. McGraw-Hill, 1997.

Mohammadi, Mohammad. On the bounds for diagonal and off-diagonal elements of the hat matrix. *REVSTAT – Statistical Journal*, 14(1):75–87, 2016. https://www.ine.pt/revstat/pdf/rs160104.pdf.

Mohsenipour, Ali Akbar. *On the distribution of quadratic expressions in various types of random vectors*. PhD thesis, The University of Western Ontario, 2012. https://ir.lib.uwo.ca/etd/955.

Morrison, Toni. A humanist view, 30 May 1975. https://mackenzian.com/blog/2014/07/07/transcript-morrison-1975/. Black Studies Center public dialogue, Pt. 2. Portland State University, Oregon Public Speakers Collection.

Mullainathan, Sendhil and Jann Spiess. Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106, 2017. https://dx.doi.org/10.1257/jep.31.2.87.

Noble, Safiya Umoja. *Algorithms of oppression: How search engines reinforce racism*. NYU Press, 2018.

Nopper, Tamara K. Digital character in "The Scored Society": FICO, social networks, and competing measurements of creditworthiness. In Ruha Benjamin, editor, *Captivating technology: Race, carceral technoscience, and liberatory imagination in everyday life*, pages 170–187. Duke University Press, 2019.

Norvig, Peter. On Chomsky and the two cultures of statistical learning, 2010. http://norvig.com/chomsky.html.

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019. https://dx.doi.org/10.1126/science.aax2342.

Ochigame, Rodrigo. The long history of algorithmic fairness. *Phenomenal World*, 30 January 2020. https://phenomenalworld.org/analysis/long-history-algorithmic-fairness.

Ochigame, Rodrigo, Chelsea Barabas, Karthik Dinakar, Madars Virza, and Joichi Ito. Beyond legitimation: Rethinking fairness, interpretability, and accuracy in machine learning. In *ICML 2018: The Debates, at the 35th International Conference on Machine Learning*, ICML 2018, 2018.

Olteanu, Alexandra, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, 2, 2019. https://dx.doi.org/10.3389/fdata.2019.00013.

O'Mahony, Seamus. Medicine and the McNamara fallacy. *Journal of the Royal College of Physicians of Edinburgh*, 47(3):281–287, 2017. https://dx.doi.org/10.4997/JRCPE.2017.315.

O'Neil, Cathy. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown, 2016.

Opsomer, Jean, Yuedong Wang, and Yuhong Yang. Nonparametric regression with correlated errors. *Statistical Science*, 16(2):134–153, 2001. https://dx.doi.org/10.1214/ss/1009213287.

Panko, Raymond R. What we know about spreadsheet errors. *Journal of End User Computing*, 10 (2):15–21, 1998. http://panko.shidler.hawaii.edu/SSR/Mypapers/whatknow.htm.

Park, Greg. The dangers of overfitting: A Kaggle postmortem, 6 July 2012. http://gregpark.io/blog/Kaggle-Psychopathy-Postmortem/.

Pasquinelli, Matteo. How a machine learns and fails — a grammar of error for Artificial Intelligence. *spheres: A Journal for Digital Cultures*, 5, November 2019. http://spheres-journal.org/how-a-machine-learns-and-fails-a-grammar-of-error-for-artificial-intelligence/.

Patton, Desmond, Philipp Blandfort, William Frey, Michael Gaskell, and Svebor Karaman. Annotating social media data from vulnerable populations: Evaluating disagreement between domain experts and graduate student annotators. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*, HICSS-52, pages 2142–2151, 2019a. http://hdl.handle.net/10125/59653.

Patton, Desmond U., William R. Frey, and Michael B. Gaskell. Guns on social media: Complex interpretations of gun images posted by chicago youth. *Palgrave Communications*, 5(119), 2019b. https://dx.doi.org/10.1057/s41599-019-0330-x.

Patton, Michael Quinn. *Qualitative research & evaluation methods: Integrating theory and practice.* SAGE, 4th edition, 2015.

Payne, Geoff and Judy Payne. Positivism and realism. In *Key concepts in social research*, pages 171–174. SAGE, 2004. https://dx.doi.org/10.4135/9781849209397.

Peña, Paz and Joana Varon. Decolonising AI: A transfeminist approach to data and social justice. In *Artificial intelligence: Human rights, social justice and development*, Global Information Society Watch, pages 28–32. Association for Progressive Communications (APC), 2019.

Pearl, Judea. *Causality: Models, reasoning, and inference.* Cambridge University Press, 2nd edition, 2009.

Pearl, Judea and Dana Mackenzie. *The book of why: The new science of cause and effect.* Basic Books, 2018.

Pechenick, Eitan Adam, Christopher M. Danforth, and Peter Sheridan Dodds. Characterizing the Google Books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PLOS ONE*, 10(10):1–24, 2015. https://dx.doi.org/10.1371/journal.pone.0137041.

Pfeffer, Jürgen and Momin M. Malik. Simulating the dynamics of socio-economic systems. In Betina Hollstein, Wenzel Matiaske, and Kai-Uwe Schnapp, editors, *Networked governance: New research perspectives*, pages 143–161. Springer International Publishing, 2017. https://dx.doi.org/10.1007/978-3-319-50386-8_9.

Pierce, Roger. Qualitative versus quantitative methods: A relevant argument? In *Research methods in politics*, pages 40–50. Sage Publications Ltd, 2008. https://dx.doi.org/10.4135/9780857024589.

Poon, Martha. Scorecards as devices for consumer credit: The case of Fair, Isaac & Company Incorporated. *The Sociological Review*, 55(2 supplement):284–306, 2007. https://dx.doi.org/10.1111/j.1467-954X.2007.00740.x.

Porter, Theodore M. Thin description: Surface and depth in science and science studies. *Osiris*, 27 (1):209–226, 2012. https://dx.doi.org/10.1086/667828.

Poster, Mark. Databases as discourse, or electronic interpellations. In *The second media age*, pages 78–94. Polity Press, 1995.

Price, David H. *Weaponizing anthropology: Social science in service of the militarized state.* CounterPunch, 2011.

Qiu, Xiaoyan, Diego F. M. Oliveira, Alireza Sahami Shirazi, Alessandro Flammini, and Filippo Menczer. Retraction note: Limited individual attention and online virality of low-quality information. *Nature Human Behaviour*, 3(1):102, 2019. https://dx.doi.org/10.1038/s41562-018-0507-0.

Racine, Jeff. Consistent cross-validatory model-selection for dependent data: *hv*-block cross-validation. *Journal of Econometrics*, 99(1):39–61, 2000. https://dx.doi.org/10.1016/S0304-4076(00)00030-0.

Rao, R. Bharat, Glenn Fung, and Romer Rosales. On the dangers of cross-validation: An experimental evaluation. In *Proceedings of the 2008 SIAM International Conference on Data Mining*, pages 588–596, 2008. https://dx.doi.org/10.1137/1.9781611972788.54.

Raschka, Sebastian. Model evaluation, model selection, and algorithm selection in machine learning. arXiv:1811.12808, 2018. https://arxiv.org/abs/1811.12808.

Raval, Noopur. An agenda for decolonizing data science. *spheres: A Journal for Digital Cultures*, 5, November 2019. http://spheres-journal.org/an-agenda-for-decolonizing-data-science/.

Ravetz, Jerome R. *Scientific knowledge and its social problems*. Oxford University Press, 1971.

Reiss, Julian. The explanation paradox. *Journal of Economic Methodology*, 19(1):43–62, 2012. https://dx.doi.org/10.1080/1350178X.2012.661069.

Reiss, Julian. The explanation paradox redux. *Journal of Economic Methodology*, 20(3):280–292, 2013. https://dx.doi.org/10.1080/1350178X.2013.828874.

Rescher, Nicholas. *Predicting the future: An introduction to the theory of forecasting*. State University of New York Press, 1998.

Reunanen, Juha. Overfitting in making comparisons between variable selection methods. *Journal of Machine Learning Research*, 3:1371–1382, 2003. http://jmlr.org/papers/v3/reunanen03a.html.

Richardson, Eugene T. *Epidemic illusions: On the coloniality of global public health*. MIT Press, 2020.

Richardson, Eugene T., Carl D. Morrow, Theodore Ho, Nicole Fürst, Rebekkah Cohelia, Khai Hoan Tram, Paul E. Farmer, and Robin Wood. Forced removals embodied as tuberculosis. *Social Science & Medicine*, 161:13–18, 2016. https://dx.doi.org/10.1016/j.socscimed.2016.05.015.

Rose, Todd. *The end of average: How we succeed in a world that values sameness*. HarperCollins Publishers, 2015.

Rosenbaum, Paul R. Interference between units in randomized experiments. *Journal of the American Statistical Association*, 102(477):191–200, 2007. https://dx.doi.org/10.1198/016214506000001112.

Rosenfeld, Amir, Mahdi Biparva, and John K. Tsotsos. Priming neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2124–2133, June 2018. http://openaccess.thecvf.com/content_cvpr_2018_workshops/w39/html/Rosenfeld_Priming_Neural_Networks_CVPR_2018_paper.html.

Rosset, Saharon and Ryan J. Tibshirani. From fixed-X to random-X regression: Bias-variance decompositions, covariance penalties, and prediction error estimation. *Journal of the American Statistical Association*, 2019. https://dx.doi.org/10.1080/01621459.2018.1424632.

Rudin, Cynthia. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019. https://dx.doi.org/10.1038/s42256-019-0048-x.

Sandler, Bernice R. and Roberta M. Hall. The campus climate revisited: Chilly for women faculty, administrators, and graduate students. Association of American Colleges, Project on the Status and Education of Women, 1986. https://eric.ed.gov/?id=ED282462.

Schlimmer, Jeffrey C. and Richard H. Granger, Jr. Beyond incremental processing: Tracking concept drift. In *Proceedings of the 5th National Conference on Artificial Intelligence*, pages 502–507, 1986. https://www.aaai.org/Papers/AAAI/1986/AAAI86-084.pdf.

Schölkopf, Bernhard. Learning to see and act. *Nature*, 518(7540):486–487, 2015. https://dx.doi.org/10.1038/518486a.

Seawright, Jason. *Multi-method social science: Combining qualitative and quantitative tools.* Cambridge University Press, 2016.

Selbst, Andrew D., danah boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 59–68, 2019. https://dx.doi.org/10.1145/3287560.3287598.

Shalizi, Cosma R. Connecting the dots. *American Scientist*, 99(4), 2011. https://dx.doi.org/10.1511/2011.91.335.

Shapiro, Ian. Methods are like people: If you focus only on what they can't do, you will always be disappointed. In Dawn Langan Teele, editor, *Field experiments and their critics: Essays on the uses and abuses of experimentation in the social sciences*, pages 228–241. Yale University Press, 2014.

Shmueli, Galit. To explain or to predict? *Statistical Science*, 25(3):289–310, 2010. https://dx.doi.org/10.1214/10-STS330.

Shumway, Robert H. and David S. Stoffer. *Time series analysis and its applications with R examples.* Springer Texts in Statistics. Springer, 3rd edition, 2011. https://dx.doi.org/10.1007/978-1-4419-7865-3.

Silberzahn, Raphael, Eric L. Uhlmann, Dan P. Martin, Pasquale Anselmi, Frederik Aust, Eli Awtrey, Štěpán Bahník, Feng Bai, Colin Bannard, Evelina Bonnier, Rickard Carlsson, Felix Cheung, Garret Christensen, Russ Clay, Maureen A. Craig, Anna Dalla Rosa, Lammertjan Dam, Mathew H. Evans, Ismael Flores Cervantes, Nathan Fong, Monica Gamez-Djokic, Andreas Glenz, Shauna Gordon-McKeon, Tim J. Heaton, Karin Hederos, Moritz Heene, Alicia J. Mohr, Fabia Högden, Kent Hui, Magnus Johannesson, Jonathan Kalodimos, Erikson Kaszubowski, Deanna M. Kennedy, Ryan Lei, Thomas A. Lindsay, Silvia Liverani, Christopher R. Madan, Daniel Molden, Eric Molleman, Richard D. Morey, Laetitia B. Mulder, Bernard R. Nijstad, Nolan G. Pope, Bryson Pope, Jason M. Prenoveau, Floor Rink, Egidio Robusto, Hadiya Roderique, Anna Sandberg, Elmar Schlüter, Felix D. Schönbrodt, Martin F. Sherman, S. Amy Sommer, Kristin Sotak, Seth Spain, Christoph Spörlein, Tom Stafford, Luca Stefanutti, Susanne Tauber, Johannes Ullrich, Michelangelo Vianello, Eric-Jan Wagenmakers, Maciej Witkowiak, Sangsuk Yoon, and Brian A. Nosek. Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, 1(3):337–356, 2018. https://dx.doi.org/10.1177/2515245917747646.

Smith, Linda Tuhiwai. *Decolonizing methodologies: Research and indigenous peoples.* Zed Books, 2nd edition, 2012.

Spiegler, Peter. *Behind the model: A constructive critique of economic modeling.* Cambridge University Press, 2015.

Spirtes, Peter and Kun Zhang. Causal discovery and inference: Concepts and recent methodological advances. *Applied Informatics*, 3(3):1–28, 2016. https://dx.doi.org/10.1186/s40535-016-0018-x.

Spivak, Gayatri Chakravorty. Subaltern studies: Deconstructing historiography. In Ranajit Guha, editor, *Subaltern Studies IV*, pages 330–363. Oxford University Press, 1985.

Stinchcombe, Arthur L. *When formality works: Authority and abstraction in law and organizations.* University of Chicago Press, 2001.

Stolley, Paul D. When genius errs: R. A. Fisher and the lung cancer controversy. *American Journal of Epidemiology*, 133(5):416–425, 1991. https://dx.doi.org/10.1093/oxfordjournals.aje.a115904.

Stone, Deborah A. The struggle for the soul of health insurance. *Journal of Health Politics, Policy and Law*, 18(2):287–317, 1993. https://dx.doi.org/10.1215/03616878-18-2-287.

Strathern, Marilyn. 'Improving ratings': Audit in the British university system. *European Review*, 5(3):305–321, 1997. https://dx.doi.org/10.1002/(SICI)1234-981X(199707)5:3<305::AID-EURO184>3.0.CO;2-4.

Strubell, Emma, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in NLP. In *the 57th Annual Meeting of the Association for Computational Linguistics*, ACL 2019, pages 3645–3650, 2019. https://dx.doi.org/10.18653/v1/P19-1355.

"Student" (William Sealy Gosset). The Lanarkshire Milk Experiment. *Biometrika*, 23(3/4): 398–406, 1931. https://dx.doi.org/10.2307/2332424.

Syll, Lars. The main reason why almost all econometric models are wrong. *WEA Commentaries*, 8 (3):5–10, 2018.

Teele, Dawn Langan, editor. *Field experiments and their critics: Essays on the uses and abuses of experimentation in the social sciences.* Yale University Press, 2014a.

Teele, Dawn Langan. Introduction. In Dawn Langan Teele, editor, *Field experiments and their critics: Essays on the uses and abuses of experimentation in the social sciences*, pages 1–8. Yale University Press, 2014b.

Tian, Yuandong, Jerry Ma, Qucheng Gong, Shubho Sengupta, Zhuoyuan Chen, James Pinkerton, and Larry Zitnick. ELF OpenGo: An analysis and open reimplementation of AlphaZero. In *Proceedings of the 36th International Conference on Machine Learning*, Proceedings of Machine Learning Research 97, pages 6244–6253, 2019. http://proceedings.mlr.press/v97/tian19a.html.

Tibshirani, Ryan J. and Saharon Rosset. Excess optimism: How biased is the apparent error of an estimator tuned by SURE? *Journal of the American Statistical Association*, 114(526):697–712, 2019. https://dx.doi.org/10.1080/01621459.2018.1429276.

Tomes, Henry. In the public interest: The case–and the research–that forever connected psychology and policy. *Monitor on Psychology*, 35(7):28, 2004. https://www.apa.org/monitor/jun04/itpi.

Tufekci, Zeynep. Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, ICWSM-14, pages 505–514, 2014. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8062.

Van Calster, Ben, Jan Y. Verbakel, Evangelia Christodoulou, Ewout W. Steyerberg, and Gary S. Collins. Statistics versus machine learning: definitions are interesting (but understanding, methodology, and reporting are more important). *Journal of Clinical Epidemiology*, 116:137–138, 2019. https://dx.doi.org/10.1016/j.jclinepi.2019.08.002.

van't Veer, Laura J., Hongyue Dai, Marc J. van de Vijver, Yudong D. He, Augustinus A. M. Hart, Mao Mao, Hans L. Peterse, Karin van der Kooy, Matthew J. Marton, Anke T. Witteveen, George J. Schreiber, Ron M. Kerkhoven, Chris Roberts, Peter S. Linsley, René Bernards, and Stephen H. Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530—536, 2002. https://dx.doi.org/10.1038/415530a.

Vigen, Tyler. *Spurious correlations*. Hachette Books, 2015.

von Luxburg, Ulrike, Robert C. Williamson, and Isabelle Guyon. Clustering: Science or art? In *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, Proceedings of Machine Learning Research 27, pages 65–79, 2012. http://proceedings.mlr.press/v27/luxburg12a.html.

Wager, Stefan. Cross-validation, risk estimation, and model selection. arXiv:1909.11696, 2019. https://www.arxiv.org/abs/1909.11696.

Wager, Stefan and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018. https://dx.doi.org/10.1080/01621459.2017.1319839.

Wallach, Hanna. Computational social science $\neq$ computer science $+$ social data. *Communications of the ACM*, 61:42–44, 2018. https://dx.doi.org/10.1145/3132698.

Wasserman, Larry. Rise of the machines. In Xihong Lin, Christian Genest, David Banks, Geert Molenberghs, David Scott, and Jane-Ling Wang, editors, *Past, Present and Future of Statistical Science*, pages 525–536. CRC Press, 2014.

Watson, David. The rhetoric and reality of anthropomorphism in Artificial Intelligence. *Minds and Machines*, 29(3):417–440, 2019. https://dx.doi.org/10.1007/s11023-019-09506-6.

Wells-Barnett, Ida B. The Red Record: Tabulated statistics and alleged causes of lynching in the United States, 1895. https://www.gutenberg.org/files/14977/14977-h/14977-h.htm.

Widmer, Gerhard and Miroslav Kubat. Learning in the presence of concept drift and hidden contexts. *Machine Learning*, 23(1):69–101, 1996. https://dx.doi.org/10.1023/A:1018046501280.

Wise, Tim. 'Forget Facts, I've Had Experiences!' Racism and the problem with anecdotes, bad data, and dog breed analogies. Medium, 19 June 2019. https://medium.com/@timjwise/forget-facts-ive-had-experiences-2fb0d6d89baf.

Witkowski, Rachel. CFPB overestimates potential discrimination, documents show. *American Banker*, 17 September 2015. https://www.americanbanker.com/news/cfpb-overestimates-potential-discrimination-documents-show.

Woodward, James. *Making things happen: A theory of causal explanation*. Oxford University Press, 2003.

Yarkoni, Tal and Jacob Westfall. Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, 12(6):1100–1122, 2017. https://dx.doi.org/10.1177/1745691617693393.

Yosinski, Jason, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. In *Deep Learning Workshop at the 31st International Conference on Machine Learning (ICML'14)*, 2015. https://arxiv.org/abs/1506.06579.

Yule, G. Udny. Why do we sometimes get nonsense-correlations between time-series?—A study in sampling and the nature of time-series. *Journal of the Royal Statistical Society*, 89(1):1–63, 1926. https://dx.doi.org/10.2307/2341482.

Zhang, Chunming. Comment. *Journal of the American Statistical Association*, 99(467):637–640, 2004. https://dx.doi.org/10.1198/016214504000000890.

Zhang, Sarah. The pitfalls of using Google Ngram to study language. *Wired*, 12 October 2015. https://www.wired.com/2015/10/pitfalls-of-studying-language-with-google-ngram/.

Zhao, Peng and Bin Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7(Nov):2541–2563, 2006. http://www.jmlr.org/papers/v7/zhao06a.html.

Zhou, Zhenglong and Chaz Firestone. Humans can decipher adversarial images. *Nature Communications*, 10(1334):1–9, 2019. https://dx.doi.org/10.1038/s41467-019-08931-6.

Zittrain, Jonathan. The hidden costs of automated thinking. *The New Yorker*, 23 July 2019. https://www.newyorker.com/tech/annals-of-technology/the-hidden-costs-of-automated-thinking.