Credible Prediction: Big Data, Machine Learning and the Credibility Revolution

Ryan Copus,* Ryan Hübert,† Hannah Laqueur‡

† To whom correspondence should be addressed.

The so-called "credibility revolution" changed empirical research (see Angrist and Pischke 2010). Before the revolution, researchers frequently relied on attempts to statistically model the world to make causal inferences from observational data. They would control for confounders, make functional form assumptions about the relationships between variables, and read regression coefficients on variables of interest as causal estimates. In essence, they would rely heavily on ex post *statistical analysis* to improve the validity of their causal inferences. The revolution centered around the idea that the only way to truly account for possible sources of bias is to remove the influence of all confounders ex ante through better *research design*. Thus, after the revolution, researchers have spent more effort to design studies around sources of random or as-if random variation, either with experiments or what have become known as "quasi-experimental" designs. This credibility revolution has increas-

*Harvard Law School, Griswold 107, 1525 Massachusetts Avenue, Cambridge, MA 02138; rwcopus@law.harvard.edu

†Department of Political Science, University of California, Davis, 469 Kerr Hall, One Shields Avenue, Davis, CA 95616; rhubert@ucdavis.edu

‡Department of Emergency Medicine, 2450 Stockton Boulevard, Sacramento, CA 95817; hslaqueur@ucdavis.edu

ingly brought quantitative researchers into agreement that, in the words of Donald Rubin, "design trumps analysis" (Rubin 2008).

However, the research landscape has changed dramatically in recent years. We are now in an era of "big data": at the same time as the internet vastly expanded the number of available data sources, sophisticated computational resources became widely accessible. This has opened up a whole new frontier for social scientists and empirical legal scholars: textual data. Indeed, most of the information we have about law, politics and society is contained in texts of one kind or another. These texts are now almost entirely digitized and available online. For example, in the 1990s, federal courts began to adopt online case records management—known as CM/ECF—where attorneys, clerks and judges file and access documents related to each case (United States Courts 2013). Using the federal government's PACER database (available at pacer.gov), researchers (both academic and professional) can now easily access the dockets and filings for each case that is filed in a federal court. LexisNexis, Westlaw and other companies have further improved access by providing raw text versions of a wide range of legal documents, along with expert-coded metadata to help researchers more easily find what they are looking for. And yet, despite the potential of these newly available resources, the sheer volume presents challenges for researchers. A core problem is how to draw substantively important inferences from a mountain of—often unstructured—digitized text. To deal with this challenge, researchers are turning their attention back toward the tools of statistical analysis. As many of the essays in this volume demonstrate, there is now a surging interest among researchers in one particularly powerful tool of statistical analysis: machine learning.

This essay addresses the place of machine learning in a post "credibility revolution" landscape. We begin with an overview of machine learning. Then, we make four main points. First, design still trumps analysis. The lessons of the credibility revolution should not be forgotten in the excitement around machine learning: machine learning does nothing

to address the problem of omitted variable bias. Nonetheless, machine learning can improve a researcher's data analysis. Indeed, with growing concerns about the reliability of even design-based research, perhaps we should be aiming for triangulation rather than design purism. Further, for some questions we do not have the luxury of waiting for a strong design, and we need a best approximation of answer in the meantime. Second, even design-committed researchers should not ignore machine learning: it can be used in service of design-based studies to make causal estimates less variable, less biased, and more heterogeneous. Third, there are important policy-relevant prediction problems for which machine learning is particularly valuable (e.g., predicting recidivism in the criminal justice system). Yet even with research questions centered around prediction, a focus on design is still essential. As with causal inference, researchers cannot simply rely on statistical models, but must also carefully consider threats to the validity of predictions. We briefly review some of these threats: GIGO ("garbage-in garbage out"), selective labels, and Campbell's law. Fourth, the predictive power of machine learning can be leveraged for descriptive research. Where possible, we illustrate these points using examples drawn from real-world research.

Learning with Machines

Machine learning is becoming increasingly popular among researchers. And yet, there is a great deal of ambiguity about what exactly it is. This is for good reason. Machine learning is not a specific research tool; it is a catch-all term that refers to any method that features *learning* by a *machine* about quantitative data. A unifying feature of these methods is that they leverage (ever increasing) computational power to apply (ever more complicated) techniques of statistical inference to (ever larger) datasets. As a result, one main distinction between "traditional" methods of statistical inference and machine learning methods is *scale*. In some sense, machine learning is statistical inference on steroids.

And yet, this distinction is too simplistic. The point of machine learning techniques is to delegate the statistical learning process to complex algorithms that are designed to generate the best predictions that are possible using a given dataset. To illustrate, consider a simple example. A state parole board may wish to have high quality predictions about the likelihood that a convicted felon will reoffend if they are released. Whether a potential parolee will reoffend is the **outcome** of interest to the board members. Outcome variables are also called "responses" or "dependent variables" and usually denoted mathematically by the letter Y.

Outcomes are determined by specific constellations of other factors. Variables that help predict outcomes are appropriately called **predictors** (also known as "independent variables"), which are usually denoted concisely by a vector $\mathbf{X} = [X_1, X_2, ..., X_k]$, where k is the number of predictors. The parole board members could rely on their intuitions about the factors that increase the likelihood of reoffending, which might generate decent predictions to guide their decision making. For example, a member may assume, based on their past experiences, that a potential parolee is much more likely to reoffend if they were convicted of a very serious offense. But more often than not, these kinds of intuitions will yield bad predictions. First of all, predictions based on these intuitions might be wildly inaccurate and lead to many high-risk prisoners being released and many low-risk prisoners being denied parole. Second, the predictions the members make might be very noisy. An individual member's assessments across many cases might be inconsistent, or members may disagree with one another.

Alternatively, the members of the parole board could assemble the data on all past parolees and delegate the task of predicting reoffending to a machine. Statistical analyses of relationships like this—where predictor variables are used to predict outcomes—are known as **supervised**

learning.[1] The machine will perform the task agnostically. It will search over all the possible ways to make statistical predictions using the data, and will return the best predictions it finds. Of course, while the machine may perform this task objectively, the quality of the predictions it produces depend critically on the data being used. In the section "Designing Good Predictions" below, we specifically address a set of concerns about the validity of predictive models when used in real-life applications.

Y-hats, Not Beta-hats

The machine doesn't care how or why it generates accurate predictions. This may be somewhat unsatisfying to a human analyst observing the machine learn from data: it will often be nearly impossible to make sense of the machine's learning process. As a result, and especially if she uses a sophisticated machine learning method, the analyst will usually be unable to make inferences about *why* certain variables were more or less helpful in generating accurate predictions. For the members of a parole board, this may not be such a problem. Their central concern is whether or not a potential parolee will reoffend. And with enough historical data on reoffending to guide the learning process, a machine can give them very accurate predictions about this outcome.

But social scientists are not like parole board members. Whereas a parole board member may find it unnecessary to probe into the determinants of the machine's predictions, a social scientist would approach the question of parolee reoffending with an eye toward explaining *why* reoffending occurs. At the very least, the social scientist would seek to

1. Supervised learning can take many forms, but it is always "supervised" by the researcher's expectation that outcomes can be predicted by the predictors. There is an entire class of statistical problems where there is no obvious relationship between an outcome variable and a set of predictors. In these cases, a researcher is often interested in making sense of the relationship among the independent variables. This is called **unsupervised learning**. A common example is clustering.

understand which of the variables in the dataset "explain" most of the reoffending behavior of past parolees. Indeed, social scientists don't generally want to predict what has happened or will happen in the world, except when it is useful for some auxiliary purpose on the way to explaining a larger phenomenon (e.g., using propensity scores to predict treatment in a matching analysis). This somewhat subtle point is fundamental to understanding the role of machine learning in social science, but it is often obscured by the fact that many of the tools social scientists use for explaining things are also useful for predicting things.

To illustrate, consider an example drawn from a dataset of all Ninth Circuit civil cases from 1995 to 2013, used in Copus and Hübert (2017). A perennial question among scholars who study U.S. courts is: what explains differential decision making by federal judges? A researcher studying the Ninth Circuit might speculate that that there is a linear relationship between the outcome—whether an appeal is reversed by the Ninth Circuit—and some case-specific and judge-specific predictors.[2] Accordingly, they might estimate an ordinary least squares (OLS) regression model, such as the following.[3]

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \varepsilon_i$$

For concreteness, suppose the researcher is specifically interested in whether panels with majority Republican appointees are more likely to reverse lower court decisions. Accordingly, she regresses reversal on an dummy variable indicating whether the appellate panel is majority Republican

2. More specifically, we examine whether cases are "treated negatively" by the Ninth Circuit, which encompasses reversals, remands and vacated cases. With some abuse of language, we will refer to the outcome variable from Copus and Hübert (2017) as "reversal."

3. For expositional purposes, we present a variant of OLS with binary outcome variables, known as a linear probability model. Since models like this can generate predicted probabilities greater than one or less than zero, an analyst may instead opt to impose additional functional form assumptions to constrain the predicted probabilities. This can be achieved with a logistic or probit regression.

appointees. Moreover, she expects this might also depend on whether the plaintiff won in the trial court, and so she includes that variable along with an interaction term. Using the data from Copus and Hübert (2017), she would get the following estimated model (with all coefficients statistically significant at 0.001 level):

$$\hat{Y}_i = 0.305 + 0.131 \cdot \text{Pltf}_i - 0.074 \cdot \text{MajRep}_i + 0.084 \cdot \text{Pltf}_i \cdot \text{MajRep}_i \quad (1)$$

The researcher ran this regression to try to detect whether politics affected case outcomes. Setting aside concerns about measurement and omitted variables bias, she would conclude that majority Republican panels are more deferential than majority Democratic panels to cases won by the defendant, but slightly less deferential than majority Democratic panels to cases won by the plaintiff. All else equal, this regression shows that majority Republican panels reverse about 23% of cases when the defendant wins at trial (as opposed to 31% for majority Democratic panels), but about 45% of cases when the plaintiff wins at trial (as opposed to 44% for majority Democratic panels).

In this example, the researcher is interested in the $\hat{\beta}$s—the marginal effects of *specific* variables on the outcome of interest. However, the regression in (1) doesn't just generate $\hat{\beta}$s, it also generates $\hat{Y}$s. That is, the regression can be used to generate a *prediction* about how a hypothetical case with certain characteristics would be decided. For example, suppose that there is a hypothetical case won by the plaintiff in the lower court and heard by a panel made up of mostly Republican appointees. Then, (1) generates a prediction about the probability that case will be reversed. Specifically, if we do the math, we see that:

$$\hat{Y}_i = 0.305 + 0.131 \cdot 1 - 0.074 \cdot 1 + 0.084 \cdot 1 \cdot 1 \approx 0.446$$

The model produced a prediction that this hypothetical case would be reversed with 44.6% probability.

OLS regression provides an analytically tractable way to recover both marginal effects of predictors on outcomes ($\hat{\beta}$s) as well as predictions

about outcomes themselves ($\hat{Y}$s). That said, most of the time, OLS regression is used by social scientists to explore specific relationships between variables, *not* to generate predictions. A social scientist might therefore ask "why do I care that the hypothetical case above would be reversed with 44.6% probability?" We are sympathetic to this concern, because much of social science is about understanding *why* the world works the way it does. But this concern is also too quick to dismiss the myriad ways that prediction exercises do strengthen social science research, even research that is causally oriented. In the sections below, we will provide specific examples to demonstrate how prediction alone can aid in social scientific research.

In the mean time, we need to fix ideas. Machine learning is not (yet) widely used in the social sciences, and so many readers will be unfamiliar with the ideas and jargon that motivate its applications. Moreover, understanding the promise of machine learning requires a relatively dramatic paradigm shift. As a methodological tool, analysts use machine learning precisely when they care about the $\hat{Y}$s, but not the $\hat{\beta}$s. This is because machine learning techniques generate very good $\hat{Y}$s, often *much* better $\hat{Y}$s than one can recover from simple regression-based methods familiar to most quantitative researchers. In the following subsection, we introduce the fundamental ideas motivating machine learning methods in a relatively non-technical way, focusing attention on explaining why it is so good at prediction.[4]

What Are We Predicting?

Suppose that we are interested in generating high quality predictions from an existing dataset. There are many potential ways to get predic-

4. There are several excellent primers on machine learning, which we highly recommend to interested readers: Hastie, Tibshirani, and Friedman (2008), James et al. (2013), and Grus (2015).

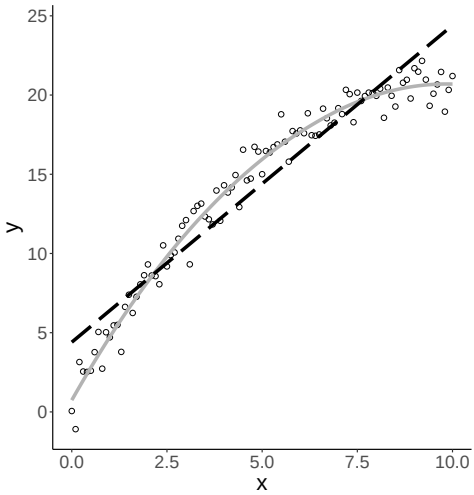

Figure 1. *In-Sample Prediction*

tions from data, and we have already seen one: OLS regression. Consider the specific prediction we generated from the Ninth Circuit data in Copus and Hübert (2017): a hypothetical case won by the plaintiff in the lower court and heard by an appellate panel of mostly Republican appointees has a 44.6% chance of being reversed by that panel. An obvious questions emerges: is this a "good" prediction of how the hypothetical case will be resolved?

Answering this question is more complicated than it initially appears. First, what exactly are we predicting? Sometimes our goal is to make predictions within the dataset we're analyzing. This is what traditional OLS regression does; it generates **in-sample predictions**. An analyst may wish to do this for the express purpose of summarizing existing data more clearly. For example, consider Figure 1, which presents a scatterplot of some simulated (i.e., fake) data. A researcher looking at the plotted data may struggle to see the relationship between the variables. To get a sense for the data, she runs a regression of Y on X, and gets the

following: $\hat{Y} = 4.42 + 1.99X$. These predictions—the $\hat{Y}$s—are represented by the black dashed line. After plotting, the analyst realizes that the predictions generated by this regression model do not fit the data very well. That is, she **misspecified** the regression model that she estimated. The solid grey line represents the predictions generated by an OLS regression with a quadratic term (the correctly specified model). Such in-sample predictions are thus useful for the researcher to figure out the most appropriate model of the data when she doesn't know ex ante what it is.

Alternatively, our goal might be to predict the outcome of a case from some other dataset, to which we may or may not have access. An obvious example would be predictions about future cases, about which we obviously do not have data. This is referred to as **out-of-sample prediction**. With out-of-sample prediction, the analyst is not interested in making predictions within their existing data except as a diagnostic for how well the model of the data will perform in other, similar datasets.[5] For example, suppose a researcher has access to Ninth Circuit data from 1995 to 2005, but not after 2005. They could build a model for predicting outcomes in the 1995-2004 time period and then use it to make a prediction about what happens from 2005 to 2010. One might expect the early data would provide decent (but not perfect) predictions for the latter period.

This is a somewhat subtle point, but we must emphasize that researchers using machine learning in contemporary research are almost always doing some form of out-of-sample prediction. We will discuss this in more depth below, but suffice to say, out-of-sample prediction is not only used to predict outcomes in data that is not available (such as forecasting the future). It turns out that the process of conducting out-of-sample predictions on subsamples of data not used in the initial machine learning analysis has important methodological benefits that im-

5. We use the terms "model" and "predictive model" loosely. As is common in machine learning applications, a "model" is an estimated relationship between variables which can be used to generate predictions in- or out-of-sample.

prove the quality of predictions. These subsamples, which are randomly chosen and held-out from the larger sample during the initial analysis are refereed to alternatively as **held-out samples**, **validation sets** or **test sets**. As these names imply, these subsamples are untouched until they are used to evaluate the quality of the machine learning predictions that are generated using the portion of the sample that is not held-out.

What Are "Good" Predictions?

We now ask what it means for predictions to be "good"? The challenge is that an analyst can choose among several (and a potentially infinite number of) methods for generating predictions. Even restricting attention to regression, an analyst can choose between various types of regression—such as linear OLS or logistic—and also can choose which variables to include in their estimation.

One trivial solution for maximizing the accuracy of one's predictions is to estimate a **saturated model**, which returns a specific prediction for Y for every combination of the predictors. And more specifically, for a specific combination of the predictors, the prediction a saturated model returns is simply the actual value of Y in the dataset for that combination of predictors. In this case, the predictions would be the least biased estimates of the actual outcomes, given the available data.[6] This is because they yield highly tailored and specific predictions. However, despite the fact that those models perform very well in the sample from which they are estimated, they will perform very badly on other samples drawn from the same population. This is because, in any given dataset, there will be a very small number of observations, or no observations at all, for each

6. Note that we use the term "bias" as in James et al. (2013): "bias refers to the error that is introduced by approximating a real-life problem, which may be extremely complicated, by a much simpler model" (p. 35).

combination of the predictors. This problem is often referred to as **overfitting**, and will lead to (out-of-sample) predictions with high variance. At the other extreme, an analyst could estimate a model that predicts Y by its mean. The predictions from this model will have very low variance, but will be very biased. Analysts working with limited data always face a choice between reducing the bias and reducing the variance in their predictions. This is referred to as the **bias-variance trade off**.

To find the optimal balance of bias and variance, one must evaluate a model's predictions outside of the sample used to generate the model. As a result, most supervised machine learning applications follow a specific recipe. First, the analyst randomly partitions the dataset into two subsets, a **training set** and a **test set**. The analyst uses the training set to estimate several predictive models of the outcomes in the training set. The analyst then takes each predictive model and assesses its performance in the test set in order to see whether it accurately predicts actual outcomes on data that was not used to build the model. One approach to doing this assessment is to calculate the **mean squared error** (MSE), which roughly speaking measures the difference between the model's predicted outcomes and the actual outcomes. If the MSE is high, then the predictive model generates "bad" predictions because it cannot accurately predict data that was not used to do build the model. The analyst chooses the model that performs best.

Validation, Validation, Validation

There is nothing inherently novel or difficult about splitting a dataset into training and test sets and performing model assessment via MSE. The advantage brought by machine learning techniques comes from the fact that computational power now allows us to quickly estimate complicated models and assess their quality repeatedly. This is advantageous in two ways. First, it allows analysts to specify a wide variety of candidate models, or even a combination of models (called an **ensemble learner**), and then let the machine choose which model (or combination

of models) optimizes on the bias-variance trade off. Second, for any given model, it allows the analyst to perform an especially robust form of evaluation known as **cross validation**, which relies on the idea that an analyst should repeatedly (and independently) divide the data into training and test sets, estimate models and assess model performance.

Cross validation allows an analyst to use as much data as possible to both build and evaluate predictive models, thus reducing bias in estimates. It also insulates an analyst against concerns that the initial split of the data unintentionally generated training and test sets that are unrepresentative.[7] The process proceeds as follows. First, the analyst randomly partitions the data into K subsets, each called a "fold" of the original dataset. Then, for each subset k, the analyst estimates a model on all of the data *excluding* k. Then, the analyst treats k as the test set, and calculates the MSE of the model generated without k. Once this is done for each of the K subsets, an analyst can calculate the average MSE across the K folds to either assess overall model performance, or to assess a particular model against an alternative one. The number of folds, i.e., K, is determined by the analyst. If K is low, then the benefits of cross validation are limited, and an analyst may tend to overstate the average MSE across the K-folds (i.e., it will be biased upward). However, if K is high, the procedure can become computationally difficult, which is a serious constraint in many applications. As a result, it is conventional for researchers to perform either 5- or 10-fold cross validation since "these values have been shown empirically to yield test error rate estimates that suffer neither from excessively high bias nor from very high variance" (James et al. 2013, p. 184).

As we have described in this section, machine learning offers a principled way to perform statistical analysis with the goal of generating ac-

7. For example, recall that at the 95% confidence level, randomization will fail to produce groups that are identical in expectation 5% of the time.

curate predictions. In later sections, we explore some of the ways this is useful for researchers. Before proceeding, however, we address its relationship to the credibility revolution.

Design Still Trumps a Machine Learning Analysis

The credibility revolution turned researchers' focus toward obtaining credible *causal* estimates. A causally credible estimate of the effect of some variable T (a **treatment**) on an outcome Y requires that it is only T, and no other confounding variable, that fully explains differences in outcomes across the subjects being studied. The most obvious way to establish this would be to examine an outcome for a particular subject with and without the treatment and compare those outcomes. Of course, the **fundamental problem of causal inference** rules out the possibility of observing an outcome when the subject was not exposed *and* when it was exposed (Holland 1986). In lieu of this, researchers compare average outcomes across "treated" and "untreated" (or "control") groups. In order for the effect on the outcome to be completely explained by the treatment, as required by our conception of causality, the two groups must be comparable and thus identical (on average) in all relevant characteristics.

With groups of research subjects formed "in the wild" it is often unknowable whether two groups are comparable since that conclusion may depend on information unavailable to the researcher. But if the researcher controls the construction of the groups, she can randomly create the groups to ensure comparability on average—a **randomized experiment**. Or, if she knows something about the world that guarantees two groups were "as-if" randomly generated in the wild, then she can assume they are comparable—a **quasi-experiment** or a **natural experiment**. Either way, a researcher's ability to make credible causal inferences depends on her outside knowledge of the comparability of her comparison groups. That is, the credibility of her measured effects is established by the **design** of her causal research study.

Estimates without such a design were largely discredited. The traditional approach of running an OLS regression, controlling for "other factors," and reading the coefficient on the variable of interest, frequently failed to generate accurate causal estimates. When researchers attempt to obtain credible estimates by statistical modeling rather then design—an **analysis-based approach**—the researcher must, at minimum, rely on the assumption that, conditional on observed variables, treatment is independent of the outcome (i.e., "selection on observables") to identify causal effects. With the credibility revolution, researchers largely lost faith in the plausibility of that assumption. For example, Robert LaLonde famously compared the results of an analysis-based approach with the results of a randomized experiment to assess whether an employment program had a causal effect on wages (LaLonde 1986). He showed that the experimental results could not be replicated by a standard analysis-based approach.

But with machine learning and the large-n, variable-rich datasets that text processing makes possible, it is tempting to believe that the selection on observables assumption can be revived. Hopes should be tempered. Even with an immense number of variables, it is exceedingly difficult to know if we have measured every possible confounder, and all it takes is one missing variable to violate the independence assumption and thus the credibility of an estimate. Moreover, even if all relevant variables have been measured, machine-learning algorithms may fail to recover the true form of the relationships between covariates and the dependent variable, and machine learning will not correct for a researcher's poorly-theorized selection of variables (e.g., if the researcher includes post-treatment or collider variables in the dataset she feeds to an algorithm, there is no reason to expect machine learning to correct for those mistakes).

... But Machine Learning Improves Analysis

At the same time, there is still arguably a place for observational studies to supplement experimental and quasi-experimental research, particularly given recent concerns regarding the reliability of design-based studies and the external validity limitations of experimental studies. We do not wish to wade too deeply in the debate among researchers about the appropriate standards to apply to evaluating the credibility of empirical research, especially research that does not rest on strong designs (in the sense we have defined it). But, we do think one thing is clear: insofar as researchers are proceeding without strong designs, they should be using machine learning in their analysis.

Machine learning techniques can improve traditional econometric methods for estimating average treatment effects under the conditional independence assumption by generating better estimates for any prediction component of a larger estimation problem. An obvious example is estimating the propensity score—the "conditional probability of assignment to a particular treatment given a vector of observed covariates" (Rosenbaum and Rubin 1983). At heart, the probability of treatment is a prediction question, and as such, it is well-suited to a machine learning approach where machine learning classification algorithms will often outperform standard parametric modeling techniques (Westreich, Lessler, and Funk 2010). In much of the published literature, the propensity score is estimated using maximum likelihood logistic regression models. However, the correct model of treatment is generally unknown, and a misspecified propensity score model can increase bias, even if the conditional independence assumption is valid (Drake 1993). A popular package for conducting matching in the R statistical program, GenMatch, is an example of a machine learning based matching method that requires no manual programming and performs better than propensity scores generated from logistic regressions (Diamond and Sekhon 2013).

More generally, model misspecification is a core problem for researchers working with observational data. Doubly robust estimation techniques

attempt to insulate an analyst from concerns about model misspecification by combining a propensity score model with a traditional model of the outcome, such as an OLS regression of Y on T and $\mathbf{X}$. The key advantage of doubly robust techniques is that they will generate consistent estimates of the average treatment effect *even if one of the models (but not both) is misspecified* (Bang and Robins 2005). Technical details aside, a doubly robust estimate is an average treatment effect derived from predictions generated by a propensity score model and predictions generated by an outcome model. It is unbiased if one is willing to believe there are no unmeasured confounding variables, and at least one of the models is correctly specified. Since machine learning itself reduces the possibility of model misspecification, doubly robust estimates generated via machine learning will be *even more* insulated against misspecification. One example of a doubly robust approach that incorporates data-adaptive machine learning is Targeted Maximum Likelihood Estimation (TMLE, see Van der Laan and Rose 2011).[8]

Machine Learning Improves Design-Based Causal Inference

Even design-committed researchers have something to gain by incorporating machine learning into their causal inference research. Research on the connection between machine learning and causal inference is rapidly expanding. For example, machine learning can be used to adjust for covariates to improve the precision of estimates of average treatment effects in randomized controlled trials, can increase power and reduce bias in instrumental variables research resulting from violations of the monotonicity assumption, and can help recover heterogeneous treatment effects. For extensive discussions, we refer readers to Athey and Imbens (2017) and Mullainathan and Spiess (2017). Here, we focus on an issue

8. Augmented inverse-probability weighting (AIPW) is another popular doubly robust estimator, introduced in the missing data literature by Robins, Rotnitzky, and Zhao (1994).

that is of particular interest to legal scholars: estimating the heterogeneous effects of decision-makers on case outcomes.

A core concern for the legal system, as well as policymakers in other parts of government, is whether front-line officials are resolving cases *consistently*. The empirical literature on inconsistency in adjudication is rapidly accumulating. We now have studies of inter-judge disparities in asylum cases, in social security disability awards, criminal sentencing in the federal courts, the Patent and Trademark Office, and in nursing home inspections. Some of the findings have been unsettling. Evidence of inconsistency in criminal sentencing, for example, facilitated the development of the Federal Sentencing Guidelines (Stith and Cabranes 1998). And large disparities in asylum grant rates have given rise to calls for institutional reform (Ramji-Nogales, Schoenholtz, and Schrag 2007), as have the findings of inconsistency in the Courts of Appeals (Tiller and Cross 1999).

Most empirical studies of inconsistency in decision making compare the decision rates of judges who have been randomly or as-if randomly assigned cases. Although the random assignment of cases allows for causal inference (e.g., the effect of Judge A on the reversal rate as compared to Judge B),[9] the traditional approach to studying inconsistency can dramatically understate inconsistency. By comparing decision rates on a single, intuitively specified dimension—such as whether a case was decided in a liberal direction or whether a case was reversed—scholars are potentially missing much of inter-judge disagreement (Fischman 2014).

9. Recent studies have cast doubt on the randomization assumption invoked by many legal and judicial politics scholars (for example, Chilton and Levy 2015). We note here that even if *judges* are not randomly or even as-if randomly empanelled in federal circuit courts, this does not mean that *cases* are not randomly assigned to panels. In Copus and Hübert (2017) we only rely on the assumption that groups of cases are randomly assigned. Even so, out of an abundance of caution, we also perform adjustments to guard against possible threats to randomization.

A judge's decision making may vary across different kinds of cases, and the nature of that variation might also depend on the judge she is being compared to. In essence, there are heterogeneous treatment effects. For example, two judges might have identical reversal rates overall but have very different reversal rates in subsets of cases: one judge may reverse more often when the plaintiff prevails, while the other may reverse more often when the defendant prevails. We could, as some researchers have done, subjectively code certain types of reversals differently from other types (e.g., in employment discrimination cases code a reversal as liberal/conservative if a defendant/plaintiff won at the trial level). And while we might fully capture the disagreement between Judge A and Judge B by changing the analyzed outcome to "liberal" rather than "reversed," that coding scheme might poorly describe the form of disagreement between Judge A and Judge C. Moreover, such a time-consuming effort may miss important distinctions or stress distinctions that are not actually relevant.

Machine learning can be used to aggressively search for inter-judge disagreement, minimizing the amount of undetected disagreement. Copus and Hübert (2017) provides a full and technical explanation, but we recount the core intuitions here. Identifying inconsistent decision making between any two judges is a ultimately a task of prediction: which kinds of cases is Judge A more likely to decide favorably than Judge B? We begin by using machine learning to generate a model of each judge's decision to reverse each case, as a function of various case-level characteristics. In this step, our goal is *not* to determine what factors "caused" judges to make their decisions, but rather to simply predict whether a given judge would affirm or reverse a given case. For each judge, this allowed us to generate a predicted probability that judge would reverse for every case in our dataset.

To illustrate, consider Figure 2, where we present plots comparing two pairs of judges: Judges Leavy and Reinhardt (left panel), and Judges Kleinfeld and Pregerson (right panel). On each axis, we plot a specific

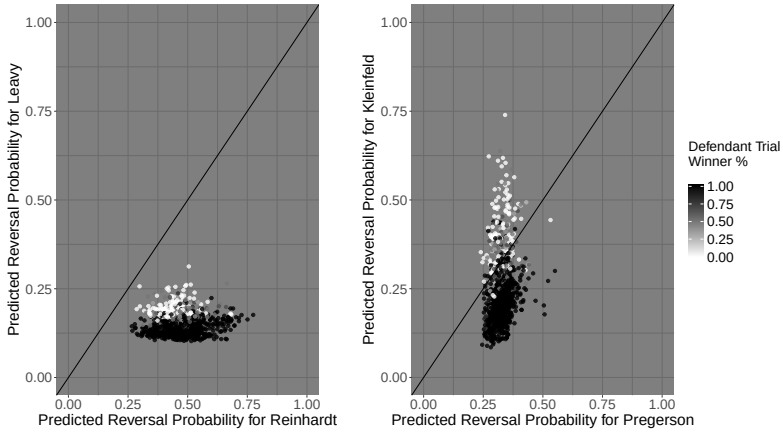

Figure 2. *Predicting the Votes of Ninth Circuit Judges*

judge's predicted probability of reversing a case. Each dot represents a single case, and the shading of the dot corresponds to whether the defendant or the plaintiff won in the district court. Darker dots indicate that the defendant won, whereas lighter dots indicate that the plaintiff won.[10] First, note that if two judges voted to reverse cases in similar ways, then all of the dots would be arranged on the 45 degree line. Judges Kleinfeld and Pregerson have roughly the same propensity to reverse (although Judge Kleinfeld's is noisier), while Judge Reinhardt is much more prone to reverse than Judge Leavy. However, while Judge Pregerson is equally prone to reverse cases won by the plaintiff or defendant, Judge Kleinfeld is much more likely to reverse cases won by the defendant. A similar, although much less pronounced pattern exists for Judge Leavy:

10. For reasons we do not need to recount here, the outcome in the trial court is expressed as the probability the defendant won, rather than a binary variable. That is why the dots are several shades of grey.

he is somewhat less likely to reverse these cases than those won by the plaintiff. To be clear, we had no ex ante theoretical expectation about the model that most accurately describes the way that case-level factors affected outcomes. Instead, we used machine learning to find the model that best predicted each judge's *actual* votes on cases, abstracting away from the precise reasons for those votes.

Using these models we are able to say, for each pair of judges, whether the outcome of each specific case in our data set is more consistent with Judge A's decision making, or more consistent with Judge B's decision making. Consider a hypothetical case from our dataset won by the defendant at trial and reversed on appeal. The predictive models depicted in Figure 2 would indicate that outcome is more "Judge Pregerson-ish" than "Judge Kleinfeld-ish" and more "Judge Reinhardt-ish" than "Judge Leavy-ish." Importantly, the former distinction is due to the fact that Kleinfeld has a pro-defendant tilt in his decision making, whereas the latter distinction is due to the fact that Reinhardt is simply more prone to reverse overall. By definition, the difference between the rates at which Judge A and Judge B make Judge A-ish decisions will best capture the disagreement rate between those two judges. Then, if Judge Reinhardt makes "Judge Reinhardt-ish" decisions in 80% of cases, while Judge Leavy makes "Judge Reinhardt-ish" decisions in 50% of cases, then they disagree 30% of the time.

This exercise in statistical modeling may strike some as overly complicated. Suppose instead that a researcher simply compared the reversal rates between these judges. She would estimate a large degree of disagreement between Judges Leavy and Reinhardt, but less so between Judges Kleinfeld and Pregerson. If instead, she compared the proportion of pro-plaintiff decisions between these judges,[11] she would estimate a

11. For example, affirming a pro-plaintiff lower court decision or reversing a pro-defendant lower court decision.

large degree of disagreement between Judges Kleinfeld and Pregerson, but less so between Judges Leavy and Reinhardt. Our procedure shows how researchers can use the predictive powers of machine learning to guide their coding choices instead of relying on intuitive, and possibly incorrect, guesses about which variables best capture disagreement among a heterogeneous set of judges. Indeed, because our procedure is designed to fully capture (as best as the data allows) the idiosyncratic differences between pairs of decision making, it gives a much clearer sense of how much inconsistency there is in large, decentralized decision making bodies.

The goal of Copus and Hübert (2017) is to *accurately measure* disagreement among judges, not to assess the reasons for that disagreement. Measuring disagreement among public officials—which is more feasible with machine learning—is important for policymakers who may be concerned about the quality and consistency of governance. This underscores the promise of prediction for helping resolve tough policy or management problems. But the lessons learned from predictions are only as good as the predictions themselves. In the following section, we describe how prediction alone can be useful for scholars and we emphasize that prediction-based researchers, just like causal inference researchers, should aim to generate highly credible estimates.

Designing Good Predictions

Machine learning methods were developed and optimized for prediction, and, unlike causal inference questions, few assumptions are required for off-the-shelf machine learning prediction techniques to work. As machine learning techniques are moving from the domain of computer science to empirical legal studies and the social sciences, a nascent body of work has pointed to causal-adjacent prediction problems that have important policy applications (Kleinberg et al. 2015). In what follows, we describe some examples of promising applications of prediction problems relevant to empirical legal scholars. At the same time, we also

emphasize that even if the problem is purely one of prediction, the researcher must still consider data and design issues that will impact the validity of the predictive model and the questions to which it is applied. As such, there are parallel lessons from the credibility revolution that should be heeded.

Prediction for Policy

Perhaps most relevant to the legal domain, machine learning prediction can be used to improve criminal justice decision making (see e.g., Berk, Sorenson, and Barnes 2016). Risk prediction is centrally embedded in every aspect of the criminal justice system—police target areas where crime is most likely, judges assess defendants' risk of flight and risk to public safety when determining bail, prison administrators segregate inmates according to risk scores, parole release determinations hinge on forecasts of inmates' future dangerousness. As such, better prediction via machine learning offers the potential to generate more efficient, effective, and equitable decisions and interventions. Kleinberg et al. (2018), for example, build a machine learning algorithm to predict criminal risk among defendants awaiting trial. They argue the use of such an algorithm in bail decisions could reduce crime by up to 25% without any change in jailing rates, reduce the the population jailed by 42%, without any increases in crime, and all while reducing the percentage of African-Americans and Hispanics jailed. Goel, Rao, and Shroff (2016) use machine learning methods to examine stop-and-frisk practices in New York City, arguing that if the police conducted only 6% of the stops that are statistically most likely to result in weapons seizure, they could recover the majority of weapons and mitigate racial disparities.

In addition to predicting external outcomes to help guide decisions, machine learning can also be used to predict decisions themselves—and those predictions can in turn be used to guide future decision making. In many legal contexts, an outcome variable like reoffending is not available: there often is no better indicator for the right decision than the

decision that a judge actually made. Federal and state judges, along with an army of front-line bureaucrats such as administrative law judges, food safety inspectors, and tax auditors, regularly interpret and apply centrally promulgated rules, but the application of those rules is often riddled with inconsistency. Some judges may make different types of decisions than other judges, and some judges may even be internally inconsistent, making different decisions based on their mood due to cognitive biases like the gambler's fallacy. Researchers can use prediction to distill decision signal and dispense with the noise, and such predictions can in turn be used to improve future decision making.

Laqueur and Copus (2016) explain how predictive models of decisions can pool the judgment of many decision makers and how that pooled judgment can be used to regulate and guide the decision making of individual decision makers, improving the consistency and overall quality of decisions. The key insight is that excluding factors that are statistically unrelated to the merits of cases (e.g., judicial identity and judicial mood) from a predictive model allows that model to smooth over and cancel out the influence of those arbitrary factors. The purified model of historical decision making can then promote more consistent and better decisions in the future. The authors use text parsing methods to extract a robust set of variables from the transcripts of all parole hearings conducted by the California Board of Parole Hearings. They then show that a predictive model of California parole decisions could be implemented to target the most abnormal judicial decisions for secondary review.

Prediction is also useful outside of the adjudication context. Kleinberg et al. (2015) point to a number of other policy-relevant prediction problems. For example, in health policy, there are resource allocation questions such as which elderly patients should receive hip replacement. Often a doctor may want to know whether a specific patient will respond positively to a new treatment, and it is less of a priority to know *why* that patient responds positively. In government regulatory policy, building or hygiene inspection problems are questions about *where* to inspect as op-

posed to *why*. In the context of criminal law, a parole board may wish to know whether a convict is more or less likely to reoffend before granting parole, but is less interested in what causes an inmate to reoffend.

Data and Design Considerations

Policy-oriented researchers can, should, and are focusing more on prediction problems. As compared to causal inference, prediction is easy. A turn toward prediction does not mean that researchers must abandon the policy issues they currently focus on, but they can alter their approach to leverage the clean power of prediction and avoid the messy complications of causal inference. With limited funding or time, a researcher primarily interested in a causal question might find it advantageous to convert their question to one of prediction. Consider, for example, a researcher interested in the relationship between probation services and violent reoffending. Without a source of randomization, any estimates of effects are likely to be unreliable. A researcher instead might put their efforts toward prediction: which probationers are most likely to commit a violent offense? Those predictions could then be used to direct more resources toward the high-risk individuals.

At the same time, even when the task is mere prediction, machine learning cannot be applied blindly. There are data and design considerations that pose potential threats to model validity. We now turn to discuss these potential threats, using forecasts of criminal risk as an illustrative example.

GIGO: The Data

Computer scientists popularized the term GIGO—garbage in, garbage out—and this is a paramount concern in any prediction policy question. Risk assessment instruments in the criminal justice system aim to help judges make decisions in bail, parole, and even sentencing, by predicting an offender's risk of return to crime. However, measuring whether a crime has occurred is not a straightforward matter. It requires relying

on officially recorded criminal justice events such as a crime report, an arrest, a conviction, or a return to prison, none of which may not be consistent proxies for criminal behavior. Take for example Pennsylvania's recent efforts to develop a Sentence Risk Assessment Instrument. The initial instrument design included *any* re-arrest or any re-incarceration, including for a technical violation, as the measure of recidivism that the model aimed to predict. This broad definition of recidivism results in an overly broad model that is not actually forecasting the outcome judges are most concerned about—serious crime and violence. Moreover, using any arrest or any technical violation as a measure of recidivism can compound racial disparities by producing artificially high scores for individuals in heavily policed and supervised minority communities. Indeed, recent review by Pennsylvania Commission on Sentencing has pointed to racial bias in the state's risk assessment instrument in its current form (Pennsylvania Commission on Sentencing 2018).

The Selective Labels Problem

The second, and perhaps the most crucial concern, is the problem of accurately evaluating algorithmic predictions in the presence of unobservables. Take, for example, a predictive model built to aid parole decisions. The model can only be built using data on individuals who are at risk of reoffending—those individuals whom a parole board has decided to release from prison. But the model aims to be applied to the full population of parole-eligible inmates. There is a potential mismatch between the dataset used to build a predictive model and the set of individuals to whom the model is applied. Information about paroled individuals may well not provide accurate forecasts for individuals that a judge would not have paroled. Judges do not, presumably, release observably similar individuals randomly, so there is reason to worry about the application of forecasts of paroled inmates to the entire population of parole-eligible inmates. The problem of unobservables invalidating predictive accuracy

parallels the problem that plagues valid causal inference with observational data.

The issue may be surmountable, but it requires careful attention and research design. For example, Lakkaraju et al. (2017) describe this "selective labels" concern and propose exploiting the heterogeneity of decision making as a means to accurately evaluate predictive performance of models in the presence of unobservables that influence the human decision and thus the observed outcome.

Campbell's Law

The psychologist Donald Campbell explained in 1979, "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." The adage applies to prediction for policy. A publicly available algorithm may alter the behavior of individuals to whom it is applied, rendering the algorithm less accurate. For example, consider a risk assessment instrument used in bail decisions. The presence of a defendant's family at the bail decision hearing has been found to be predictive of the defendant successfully returning to their next court date and not being rearrested in the meantime (Kleinberg et al. 2018). The presence of the family at the hearing is likely associated with unobservable characteristics of the defendant that are not included in the predictive model. If inmates know they will have a lower risk score if their families attend the hearing, more defendants may ensure their families join them. But given family presence is unlikely to be what *causes* the individual to reappear in court, but is merely predictive and associated with unobserved factors that make them lower risk, this could artificially lower the risk scores for inmates who otherwise would be classified as higher risk.

Prediction for Description

The credibility revolution has raised the bar for causal inference research and made it more difficult for researchers to identify credible causal effects. One promising, but underused, response is for researchers to engage in more descriptive research—research that eschews causal inference all together. Some of the most useful research is descriptive in nature (for a recent discussion of this, see Grimmer 2015). For example, the Martin-Quinn scores for Supreme Court justices (Martin and Quinn 2002) have been widely influential across several disciplines, as well as news reporting about the court. Machine learning is a valuable tool for building new measurements that can help researchers describe the world. In this section, we outline two particularly promising ways to use supervised machine learning techniques for measurement: classification and proxy variables.

Classification

Supervised classification involves sorting observations into known categories.[12] At its heart, classification is a kind of dimension reduction exercise that allows a researcher to generate aggregated variables based on fine-grained distinctions in the underlying dataset, which may be substantively useful. Classification can be used to solve more "technical" problems, such as identifying well-defined features from a large amount of unstructured text. Or, it can be used for more "interpretive" problems, such as applying an expert's judgment about conceptually complex issues to a large dataset. In either case, a researcher starts with a subsample of hand-coded observations and trains a machine learning model to

12. A more general version of this problem is **scoring**, where an analyst assigns observations to a continuous scale. The same ideas apply in those contexts.

Plaintiffs	**CALIFORNIA MEDICAL ASSOCIATION**
	CALIFORNIA DENTAL ASSOCIATION
	CALIFORNIA PHARMACISTS ASSOCIATION
	NATIONAL ASSOCIATION OF CHAIN DRUG STORES
	CALIFORNIA ASSOCIATION OF MEDICAL PRODUCT SUPPLIERS
	AIDS HEALTHCARE FOUNDATION
	AMERICAN MEDICAL RESPONSE WEST
	JENNIFER ARNOLD
Defendants	**TOBY DOUGLAS,** *Director* *Department of Health Care Services of the State of California*
	KATHLEEN SEBELIUS, *Secretary* *United States Department of Health and Human Services*

Figure 3. *Parties Listed in Ninth Circuit Docket Sheet (Case 12-55315)*

generate classifications for the remainder of the dataset.[13]

In Copus and Hübert (2017), we draw data from around 54,000 Ninth Circuit docket sheets. Despite the fact that most docket sheets follow a particular template, some of the data in docket sheets is unstructured text that is difficult to parse. For example, consider Figure 3, which presents a list of the parties to a specific Ninth Circuit case from 2012. Notice that there are many parties to this case, each of which could belong to a specific category of interest, such as "business," "government," "advocacy organization" or "private person." However, the format of these entries is inconsistent and thus difficult to categorize using a deterministic rule. For example, while one of the plaintiffs, the California Medical Association, is an advocacy organization, another one of the plaintiffs, Ameri-

13. To be clear, a researcher need not code his or her own training set. Samples of previously hand-coded observations exist in many places, such as LexisNexis's keywords. Moreover, researchers should be attuned to inconsistencies in hand-coding of a test set and account for them as best as possible.

can Medical Response West, is a business. Moreover, notice that Jennifer Arnold is a private person, whereas Toby Douglas is a government official. Without significant resources, it is not feasible for researchers to categorize each party to each case by hand. Instead, a researcher could draw a random sample of the total number of parties—say 1,000—and have a human coder classify each party into a specific category. Then, the researcher could use this set of human-coded parties as a test set to derive a classification model that can be applied to the rest of the data. There are currently models that are already trained to identify named entities and are ready for out-of-the-box use. One of the most famous examples is the Stanford Named Entity Recognizer[14] which is a Java-based tool that has a variety of interfaces to other programming languages, such as Python, Perl and Ruby.

Proxies

Another measurement-related use for machine learning is the creation of data-driven proxy variables. More specifically, a researcher may use high quality predictions from machine learning to serve directly as a proxy variable for some other quantity of interest. In Copus and Hübert (2017), we use case-level data to generate a predicted probability of reversal for each case and each possible panel of judges that could have been assigned to that case. We then use these predicted probabilities to generate a case-level disagreement score by measuring the spread in the predicted probability of reversal across the panel types. By using the predicted probabilities generated from our machine learning method, we were able to create a new, and substantively useful, measure of how much disagreement particular cases elicit among judges. We use this measure to test the conventional wisdom that judges are more likely to issue unpublished opinions on "easy" cases. (We find the opposite.) This measure has many

14. Available at https://nlp.stanford.edu/software/CRF-NER.shtml.

other potential applications. For example, a study of Supreme Court cert decisions might include this variable as a proxy for the salience and/or complexity of a case.

We wish to emphasize, however, that the primary benefit of using machine learning to generate new proxy variables is *not* its ability to help researchers determine which predictors are the most relevant. Machine learning methods seek to optimize predictions, not pin down which variables do most of the work. Making inferences about the most predictive variables using machine learning techniques will inevitably cause problems. Researchers are almost *always* constrained by their available data, and the fact that a variable is especially predictive in one particular machine learning application does not mean that, at a theoretical level, it is a suitable proxy. Rather than relying on intuition about what variables should be most predictive (even if guided by an estimated model), we suggest that researchers use the predictions directly.

Conclusion

In this essay, we have introduced the basic idea of machine learning and argued that its core contribution is its ability to make high quality predictions. We have also emphasized that while machine learning is not a solution to the fundamental problem of causal inference, it can be a powerful tool for aiding researchers with causally-oriented research. That said, ultimately the greatest promise of machine learning for the legal research community (and more broadly) is likely in its ability to solve prediction-policy issues and aid descriptive research. Many of the interesting questions confronting legal scholars lend themselves to high quality prediction tasks. How can we effectively classify a large set of documents, such as docket sheets or legal opinions, into useful categories? What is the probability that a convict will recidivate if granted parole? How do judges sitting in the same court differ in their decision making on individual cases? Indeed, as scholars explore the massive amount of new data made available through digitization of legal texts, they can and

should better exploit the power of machine learning to answer questions like these.

References

Angrist, Joshua, and Jörn-Steffen Pischke. 2010. "The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics." *Journal of Economic Perspectives* 24 (2): 3–30.

Athey, Susan, and Guido W. Imbens. 2017. "The State of Applied Econometrics: Causality and Policy Evaluation." *Journal of Economic Perspectives* 2 (31): 3–32.

Bang, Heejung, and James M. Robins. 2005. "Doubly Robust Estimation in Missing Data and Causal Inference Models." *Biometrics* 61:962–972.

Berk, Richard A., Susan B. Sorenson, and Geoffrey Barnes. 2016. "Forecasting Domestic Violence: A Machine Learning Approach to Help Inform Arraignment Decisions." *Journal of Empirical Legal Studies* 13 (1): 94–115.

Chilton, Adam S., and Marin K. Levy. 2015. "Challenging the Randomness of Panel Assignment in the Federal Courts of Appeals." *Cornell Law Review* 101:1–56.

Copus, Ryan, and Ryan Hübert. 2017. "Detecting Inconsistency in Governance." doi:10.2139/ssrn.2812914.

Diamond, Alexis, and Jasjeet S. Sekhon. 2013. "Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies." *Review of Economics and Statistics* 95 (3): 932–945.

Drake, Christiana. 1993. "Effects of Misspecification of the Propensity Score on Estimators of Treatment Effect." *Biometrics:* 1231–1236.

Fischman, Joshua B. 2014. "Measuring Inconsistency, Indeterminacy, and Error in Adjudication." *American Law and Economics Review* 16 (1): 40–85.

Goel, Sharad, Justin M. Rao, and Ravi Shroff. 2016. "Precinct or Prejudice? Understanding Racial Disparities in New York City's Stop-and-Frisk Policy." *The Annals of Applied Statisticss* 10 (1): 365–394.

Grimmer, Justin. 2015. "We Are All Social Scientists Now: How Big Data, Machine Learning, and Causal Inference Work Together." *PS: Political Science & Politics* 48 (1): 80–83.

Grus, Joel. 2015. *Data Science from Scratch: First Principles with Python.* Sebastopol, CA: O'Reilly Media.

Hastie, Trevor, Robert Tibshirani, and James Friedman. 2008. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* 2nd. New York: Springer.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81 (396): 945–960. doi:10.2307/2289064.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning.* New York: Springer.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. "Human Decisions and Machine Predictions." *The Quarterly Journal of Economics* 133 (1): 237–293.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. "Prediction Policy Problems." *The American Economic Review* 105 (5): 491–495.

Lakkaraju, Himabindu, Jon Kleinberg, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. "The Selective Labels Problem: Evaluating Algorithmic Predictions in the Presence of Unobservables." In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 275–284.

LaLonde, Robert J. 1986. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *American Economic Review* 76 (4): 604–620.

Laqueur, Hannah, and Ryan Copus. 2016. "Synthetic Crowdsourcing: A Machine-Learning Approach to the Problems of Inconsistency and Bias in Adjudication." doi:10.2139/ssrn.2694326.

Martin, Andrew D., and Kevin M. Quinn. 2002. "Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the U.S. Supreme Court, 1953-1999." *Political Analysis* 10 (2): 134–153.

Mullainathan, Sendhil, and Jann Spiess. 2017. "Machine Learning: An Applied Econometric Approach." *Journal of Economic Perspectives* 31 (2): 87–106.

Pennsylvania Commission on Sentencing. 2018. *Risk Assessment Update: Arrest Scales.* Technical report. http://www.hominid.psu.edu/specialty_programs/pacs/publications-and-research/research-and-evaluation-reports/risk-assessment/risk-assessment-update-february-2018-arrest-as-predictive-factor.

Ramji-Nogales, Jaya, Andrew I. Schoenholtz, and Phillip G. Schrag. 2007. "Refugee Roulette: Disparities in Asylum Adjudication." *Stanford Law Review* 60:295–412.

Robins, James M, Andrea Rotnitzky, and Lue Ping Zhao. 1994. "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed." *Journal of the American Statistical Association* 89 (427): 846–866.

Rosenbaum, Paul R., and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70 (1): 41–55.

Rubin, Donald B. 2008. "For Objective Causal Inference, Design Trumps Analysis." *The Annals of Applied Statistics* 2 (3): 808–840.

Stith, Kate, and José A. Cabranes. 1998. *Fear of Judging: Sentencing Guidelines in the Federal Courts*. Chicago: University of Chicago Press.

Tiller, Emerson H., and Frank B. Cross. 1999. "A Modest Proposal for Improving American Justice." *Columbia Law Review* 99 (1): 215–234.

United States Courts. 2013. *25 Years Later, PACER, Electronic Filing Continue to Change Courts*, December. http://www.uscourts.gov/news/2013/12/09/25-years-later-pacer-electronic-filing-continue-change-courts.

Van der Laan, Mark J., and Sherri Rose. 2011. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer Science & Business Media.

Westreich, Daniel, Justin Lessler, and Michele Jonsson Funk. 2010. "Propensity Score Estimation: Machine Learning and Classification Methods as Alternatives to Logistic Regression." *Journal of Clinical Epidemiology* 63 (8): 826–833.