



Published in final edited form as:

*Soc Sci Res.* 2022 November ; 108: 102807. doi:10.1016/j.ssresearch.2022.102807.

## Researcher reasoning meets computational capacity: Machine learning for social science

Ian Lundberg<sup>a,\*</sup>, Jennie E. Brand<sup>b</sup>, Nanum Jeon<sup>b</sup>

<sup>a</sup>Department of Information Science, Cornell University, USA

<sup>b</sup>Department of Sociology, Department of Statistics, California Center for Population Research, UCLA, USA

### Abstract

Computational power and big data have created new opportunities to explore and understand the social world. A special synergy is possible when social scientists combine human attention to certain aspects of the problem with the power of algorithms to automate other aspects of the problem. We review selected exemplary applications where machine learning amplifies researcher coding, summarizes complex data, relaxes statistical assumptions, and targets researcher attention to further social science research. We aim to reduce perceived barriers to machine learning by summarizing several fundamental building blocks and their grounding in classical statistics. We present a few guiding principles and promising approaches where we see particular potential for machine learning to transform social science inquiry. We conclude that machine learning tools are increasingly accessible, worthy of attention, and ready to yield new discoveries for social research.

### 1. Introduction

Advances in statistics and machine learning have the potential to rapidly expand the toolkit available to social scientists. The pace of change will depend on how social scientists weigh the costs and benefits of adopting new tools. Our review emphasizes four benefits to adoption: machine learning can amplify researcher coding, summarize complex data, relax some statistical assumptions, and target researcher attention. But many social scientists have yet to adopt machine learning tools. One reason machine learning methods have appeared infrequently thus far may be the appearance of high adoption costs, such as the time needed to learn new methods and the difficulties that arise when interpreting a complex model. Yet the increasing availability of open-source software and pedagogical materials means that these costs are quickly falling. One aim of our review is to contribute to the reduction in these costs by making new methods accessible; in this respect, we build on the excellent guidance provided by other recent review papers (e.g., Molina and Garip 2019; Grimmer et al., 2021; Athey and Imbens 2019). A theme of our review is that the benefits of machine learning are likely to substantially outweigh the costs over time.

---

\*Corresponding author. Direct correspondence to Ian Lundberg, Cornell University, Department of Information Science, 223 Gates Hall, 107 Hoy Road, Ithaca, NY 14853, [ilundberg@cornell.edu](mailto:ilundberg@cornell.edu).

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.ssresearch.2022.102807>.

Related to assumed costs, some social scientists may have a preconception that the adoption of machine learning methods requires a qualitative shift away from classical statistical methods. A second theme of our review is that there is no such qualitative shift. While the fields of “statistics” and “machine learning” have at times differed in their emphasis on various aspects of data analysis (Breiman, 2001b), many of the key advances occur when these perspectives are brought together. What unites these fields is far greater than what divides them. For example, a generalized linear model is a standard statistical tool. Yet one could say that such a model “learns” a set of coefficients from data. A Least Absolute Shrinkage and Selection Operator (LASSO) version of that regression “learns” which of the covariates should enter the prediction function. As one moves from methods considered “classical statistics” toward methods considered “machine learning,” one axis of change is away from imposed structure and toward a greater role for the data in learning. But this is a difference of degree rather than a difference of kind. Indeed, when a social scientist uses a statistical method, they can conceptualize that method as a specific case of a machine learning tool (perhaps a highly structured version). We emphasize these connections and ground our review in classical statistics.

Hesitancy about the use of machine learning also stems from concerns that these methods are “black box,” involving many parameters that are difficult to interpret. This concern may loom especially large among social scientists who are familiar with estimating regression models, placing the coefficients in a table, and interpreting those coefficients. Two responses address this concern. First, some machine learning (ML) methods (e.g., ridge regression) still involve coefficients. A second (and more radical) response is that social scientists’ comfort with “interpretable” regression coefficients is often misplaced. For example, researchers might interpret the coefficient as the “effect” of a particular variable. But such an “effect” may not correspond to any causal effect in the absence of additional assumptions. And if those assumptions hold, any machine learning prediction function can yield a similarly interpretable average effect estimator: predict the outcome for all units as observed, add one to the key predictor and make another prediction, difference the two, and average. Both approaches rely on the same causal assumptions, and under those assumptions both can yield an interpretable estimate of an average effect. An advantage of some machine learning methods is that the statistical assumptions may be more credible (e.g., allowing additional interaction terms). This example illustrates a general point: a researcher who is precise about the quantity to be estimated can often engineer a machine learning approach to yield an interpretable estimate of that quantity.

A final hesitancy may stem from the belief that social science should be theory-driven, and machine learning seems to be data-driven. This hesitancy is misplaced; machine learning in social science requires both theory and data. To answer social science questions with supervised machine learning, a researcher typically begins by arguing theoretically for the importance of the quantity to be estimated and then for the credibility of required assumptions. Only then can the data speak to the theory-driven question. With unsupervised learning, the quantity to be estimated may emerge from the data, but then the researcher makes a theoretical argument to label and interpret what has been learned. In either case, both theory and data play essential roles.

Our argument proceeds as follows. We first emphasize several benefits of machine learning by reviewing its use in existing social science research. Having motivated machine learning from its applicability in social science, we then provide a pedagogical introduction to some of the central building blocks of machine learning. We place special emphasis on their connection to standard statistical approaches. Third, we discuss some frontiers of machine learning research which are increasingly fruitful for social science research. Finally, we conclude with a discussion of how machine learning can contribute to social science knowledge moving forward.

## 2. What you can do with machine learning: Exemplary applications in social science

Machine learning is already yielding new insights within social science. We illustrate this point by discussing select papers published from 2016 to 2021 in a set of journals drawn from sociology, political science, and economics.<sup>1</sup> We do not review all uses of machine learning in these journals, nor do we review all classes of machine learning methods. Instead, we highlight cases that illustrate high-level ways that machine learning can transform the research process for measurement, estimation, discovery, and dimension reduction. Table 1 points toward introductory texts on these topics. The end of this section closes with a word of warning: while there is much that machine learning can do, there are also aspects of research (e.g., causal claims) for which machine learning is useful only in combination with assumptions about data that cannot be observed (e.g., counterfactuals).

### 2.1. Measurement: Supervised machine learning can amplify researcher coding

One characteristic of the digital age is the high volume unstructured text, audio, and video data. These data pose a challenge for measurement: it is not straightforward to convert them into a small set of categories or numeric summaries relevant to a research question. In small samples, researchers can carry out measurement by hand coding: examine the data and manually label. Yet hand coding becomes prohibitive in massive digital samples. Supervised learning methods can amplify the researcher's judgment: the researcher manually labels a random sample, uses an algorithm to learn patterns mapping the high-dimensional data to the low-dimensional labels in that training sample, and then predicts the unknown labels in the much larger population (see Fig. 1).<sup>2</sup>

For example, King et al. (2017) studied government involvement in the social media ecosystem in China. They examined 43,757 social media posts made by individuals employed by the Chinese government to spread propaganda. This volume of digital data would be extremely costly to analyze by hand. Instead, the authors drew a random sample of 200 posts and hand-coded them into a set of categories chosen by the authors (e.g.,

<sup>1</sup>We specifically searched for applications in *American Sociological Review*, *American Journal of Sociology*, *American Political Science Review*, *American Economic Review*, and *Social Science Research*. We also searched three methodological journals: *Sociological Methodology*, *Political Analysis*, and *Econometrica*.

<sup>2</sup>Unsupervised methods can also be used for measurement (e.g., topic models, Blei et al., 2003). With unsupervised measurement, the researcher does not manually code any cases and all cases are labeled by the algorithm. Unsupervised dimension reduction is summarized in Sec 2.2. Here we emphasize supervised methods because of the clear distinction between domain expertise (defining the categories) and algorithmic amplification (labeling many cases).

whether the post engages in arguments about the Communist party). Using these 200 posts, they learned the statistical patterns linking the words used in the posts (high-dimensional predictors) to the categories defined by the researchers (low-dimensional labels). Finally, they estimated the prevalence of each category in the entire set of 43,757 posts (using a pre-established procedure available in open-source R software; see Hopkins and King 2010 and Jerzak et al., 2022). Roughly 80% of the posts did not engage in arguments about the Communist Party but instead simply involved cheerleading for China and for the Party. This descriptive evidence was made possible by researcher expertise (defining the categories of posts and labeling a sample) amplified by the power of machine learning (labeling in a massive data set).<sup>3</sup>

In a sociological application, Friedman and Reeves (2020) explored patterns of cultural distinction in recreational activities by studying the lives of 71,393 British elites over the 19th and 20th centuries who appeared in *Who's Who*, a book cataloguing their lives. They manually coded 600 entries into three categories of recreational activities—aristocratic, highbrow, or ordinary—and then used supervised learning to estimate the prevalence of each type of recreation in the full set of 71,393 entries. The authors then summarized how patterns of elite portrayal of their recreational activities changed over time, an exercise which was only possible by combining researcher decisions (categorizing text into these three categories) amplified to a new scale by machine learning.

Beyond text, new forms of audio and visual data also become amenable to analysis through a strategy of amplified researcher coding. Knox and Lucas (2021) observe that political scientists transform audio files of political speech into transcripts. But they emphasize that doing so discards information such as vocal tone and flow of speech. Instead, the authors analyzed a sample of audio files from Supreme Court hearings. They manually labeled some speech patterns as demonstrating skepticism, and then they developed methods to predict skepticism in unlabeled utterances as a function of the audio profile of those utterances. Similar to the study of audio data, supervised learning methods developed for computer vision (Szeliski, 2010) could amplify the analysis of images in social science. For example, Cantú (2019) studied fraud in a Mexican election by examining 53,249 images of vote tally sheets. They labeled a random sample of 900 images for whether alterations were present. By amplifying that coding to the full set using a convolutional neural network classifier, Cantú (2019) revealed the extent of fraud.

Sometimes, the human coding necessary for supervised machine learning already exists and does not have to be carried out by the researcher. Gentzkow et al. (2019) worked with the text of Congressional speeches, which are already labeled by party, and used this labeled set to learn a one-dimensional reduction of the text to a continuous score for the degree to which a given piece of text was typical of one party versus the other. The algorithm allowed them to track this score over time in Congressional speeches. Zhang and Pan (2019) identified protest events in 9.5 million Chinese social media posts using a classifier trained on a subset of text and image data that was already hand-coded by activists. The use of a pre-labeled

---

<sup>3</sup>In a study with similar structure, Su and Meng (2016) manually categorized the topics of 1,000 messages from citizens to Chinese provincial officials and then used supervised learning methods to make predictions for the topics in the full set of 207,554 messages.

set can save researcher time but also can bring the costs of validating the human labels and assessing the generalizability of the potentially non-random training sample.

Amplified human coding is powerful because it draws on the strengths of both machine learning and social science. The social scientist excels at defining categories: taking a high-dimensional predictor  $\vec{x}$  (text, audio, or video data) and mapping that observation into a category  $y$  among a few discrete choices constructed for the theoretical question. The machine learning algorithm excels at the amplification task: given training samples, learn the patterns in those samples and predict for new cases. Machine learning tools thus amplify a human labeling task to apply at a new scale.

## 2.2. Dimension reduction: Unsupervised machine learning can summarize complex data

In contrast to supervised learning, there exist other *unsupervised* settings (see Fig. 2) which do not begin with labeled outcomes. Instead, they take a high-dimensional input (e.g., text data) and reduce that to a low-dimensional summary (e.g., topics). The key to success in unsupervised learning is first for the researcher to formalize what would make a pattern interesting. Then, an algorithm searches for those patterns. Finally, the researcher draws on theory to interpret the result and argue for its usefulness. We join past reviews (e.g., Molina and Garip 2019) in taking the division between supervised and unsupervised methods to be an important distinction within machine learning.

One example for unsupervised learning is the latent Dirichlet allocation algorithm (LDA, Blei et al., 2003), an unsupervised method to take documents (the unit of analysis) and convert the words they contain (a high-dimensional predictor) into a set of topics discovered inductively (low-dimensional labels). Each topic is a vector of probabilities over word frequencies, and the algorithm learns the topics to maximize the probability of the observed word frequencies in the documents subject to regularization by Bayesian priors. LDA is entirely inductive—the researcher never labels a training sample. Like many methods for discovery, LDA implicitly assumes a definition of what is “interesting” in data: a topic is interesting to the degree that it captures a distribution of words which recurs across many documents. Yet whether this algorithmic definition produces topics which are substantively useful in social science is a determination that falls to the human researcher, who should examine each topic and argue for its substantive meaning (Grimmer and Stewart, 2013). For example, sociologists often use topics learned inductively as measures of culture (Mohr and Bogdanov, 2013; DiMaggio et al., 2013; Bail, 2014). This path forward is promising but also entails risks, because there is no guarantee that an algorithm operating on word frequencies will arrive at a meaningful definition of a cultural category. For this reason, unsupervised methods to summarize text data always place the burden on the researcher to justify their chosen interpretation and validate the utility of the topics learned (Grimmer and Stewart, 2013; Ying et al., 2021; Grimmer et al., 2022). LDA thus illustrates a key idea that applies more broadly to unsupervised methods: while these methods may appear to inductively discover insights from the data alone, they actually involve extensive theoretical work on the part of the researcher to justify and interpret the result.

High-dimensional data exist in many social science settings beyond text as data. For example, Frye and Trinitapoli (2015) examine the sequence of events that precede sexual

intercourse for young women in Malawi. Each respondent's data consist of an ordered set of experiences they had with their partner leading up to sex—events such as giving presents, meeting to chat in private, talking about contraception, and getting married. Although all respondents worked from the same set of possible events, the sheer number of unique experience trajectories makes it difficult to generate theory from the raw data alone. Instead, Frye and Trinitapoli (2015) defined a distance between the experience trajectories of every pair of women, based on the number of insertions, deletions, and substitutions needed to convert one woman's trajectory into that of the other woman. The authors then used a hierarchical clustering algorithm to inductively place the observed relationship trajectories into five clusters. While the raw data are insurmountably large for theory generation, the low-dimensional set of clusters reveals commonalities and differences across these inductive groupings. The authors highlight differences across clusters such as typical length of the sequence of events, whether marriage precedes sex, and the inclusion or exclusion of socially embedded experiences such as attending a community event together. The data reduction is largely data-driven, yet it supports a heavily theory-driven interpretation of the meaningful differences that appear across the learned clusters. The problem is analogous to how LDA produces a data-driven reduction of complex textual data into a small set of topics which requires theory-driven interpretation. Unsupervised methods for sequence analysis are thus similar to unsupervised methods for text as data: they convert high-dimensional data to a low-dimensional representation, thus enabling new theorizing about that low-dimensional summary. This interplay between data and theory corresponds to a promise of unsupervised learning more generally.

### 2.3. Estimation: Supervised machine learning can relax statistical assumptions

While both supervised and unsupervised learning hold promise in social science, supervised learning is particularly straightforward as a plug-in substitute in settings where researchers might have used parametric regression. Given a set of predictors  $\vec{X}$  and an outcome  $Y$  chosen for conceptual reasons, machine learning can help with the step of estimation: learning the statistical mapping from  $\vec{X}$  to  $Y$ . In classical statistics, this step would involve choices such as whether to include interactions and squared terms. In machine learning, this step might involve consideration of nearest neighbors, random forests, and other methods to pool information across units. A machine learning perspective to estimation grounds these choices in empirical evidence: choose the estimator that fits the data well. The use of model fit to select an estimator has long been standard in social science (e.g., in the field of social mobility, see e.g. Hauser et al., 1975, 1983). Machine learning methods lean particularly heavily on empirical evidence for model selection, often assessing candidate estimators by their ability to predict the outcomes of out-of-sample cases not used for learning the model. Doing so not only yields empirically-grounded modeling choices, but it may also improve the predictive power of the estimated model.

For example, Dube et al. (2020) scraped data on Human Intelligence Tasks (HITs) posted online on Amazon Mechanical Turk (MTurk), including tasks such as placing labels on images or completing short questionnaires. Before deciding whether to complete a task, workers could see information about the financial reward offered for completion as well as other aspects of the task. The authors studied the causal effect of the reward amount

on the duration of time that the HIT remained posted before achieving its desired number of responses, taken as a metric of how quickly workers signed up and completed the task. But there was a problem: whether each worker chose to complete a task might also have been a function of other aspects of that task, such as the title, keywords, and time allotted by the requester. Those variables produced confounding, so that the marginal association between reward amount and posted time arose only in part due to a causal effect of reward amount. The authors assumed that the measured variables blocked all confounding, which is a conceptual rather than a statistical assumption (Pearl, 2009; Imbens and Rubin, 2015). But under that assumption, the authors needed to predict the outcome as a function of the treatment and confounders. To do so, Dube et al. (2020) use double machine learning (Chernozhukov et al., 2018) to adjust for confounding by learning an ensemble that averages over several learning algorithms to predict the treatment (reward amount) and the outcome (duration of posting). This machine learning strategy thus handles difficult statistical choices automatically, allowing the authors to focus their attention on the definition of the research question and the conceptual choices about variables to include so that causal assumptions will be valid.

As another example, researchers may seek to draw inference about a population using a non-representative sample. Gelman and Little (1997) proposed to accomplish this task by a parametric method: estimate a multilevel model for the survey responses as a function of measured variables (e.g., race, age), predict the outcome in each subgroup defined by those variables, and post-stratify by the known population distribution of the predictors. The validity of this procedure relies not only on an identification assumption (ignorable sample inclusion within strata of covariates), but also on the assumed functional form of the regression model. Bisbee (2019) relaxed the latter assumption with a nonparametric machine learning approach (Bayesian Additive Regression Trees, Chipman et al., 2010). This extension illustrates a key principle: once researchers make conceptual assumptions about the relevant variables, nonparametric machine learning methods can be used in place of classical models to learn statistical patterns with flexible functional forms.

#### 2.4. Discovery: Supervised machine learning can target researcher attention

Supervised machine learning can involve a complex mapping between predictors and the outcome. How to summarize that mapping may not be straightforward—the researcher could focus on many aspects of the learned relationships. Supervised learning for discovery automates the process of finding the interesting patterns in the data. We discuss discovery with two examples: the causal effect of a high-dimensional treatment and the heterogeneous causal effects of a binary treatment across a high-dimensional set of pre-treatment variables.

Conjoint experiments in political science assess how participants' perceptions of a hypothetical candidate respond to a series of randomized attributes about that candidate (Hainmueller et al., 2014). In one concrete example, Breitenstein (2019) presented voters with profiles of hypothetical mayoral candidates and randomly varied signals of the candidates' sex, party affiliation, experience qualities, economic performance under their leadership, and evidence of corruption. The space of possible treatment conditions is high-dimensional, with  $2 \times 4 \times 2 \times 2 \times 3 = 96$  unique profiles possible by combining these

attributes. The randomized design identifies the average potential outcome under each of the 96 treatment conditions, but absorbing that much information would be difficult for a reader: a low-dimensional summary of the effects is needed. Breitenstein (2019) used ordinary least squares to produce a low-dimensional summary: the average effect of each component of the profiles, marginalized over the other components. But this is far from the only possible summary. In a reanalysis, Incerti (2020) used a decision tree to search for combinations of randomized signals that interact to produce particularly disparate effects. Decision trees recursively split the data into subsets where outcomes are increasingly homogeneous, pointing toward new interactive estimands discovered inductively from the data: voters were most likely to support a non-corrupt politician's profile (72% in support) but also demonstrated high support for profiles involving corruption as long as the candidate was of the same party as the respondent and had a history of good economic performance under their leadership (67% in support). Meanwhile, a corrupt candidate of a different political party from the respondent garnered only 36% support. This substantively interesting interactive partition emerged inductively from the data. Importantly, both the original study and the re-analysis yielded valid and useful findings. This illustrates how machine learning can take existing evidence and marshal it in new ways to reveal new insights, in this case targeting researcher attention toward treatment combinations with particularly interesting outcomes.

In other settings, there is one binary treatment  $T$  and a high-dimensional set of confounders  $\vec{X}$ . The conditional average causal effect  $\tau(\vec{X})$  might take unique values at each value of the confounders. With so many average causal effects that could be reported, a question arises: for which population subgroups should we report the average causal effect? A researcher could pre-register a hypothesis comparing the average causal effect across two subgroups motivated by theory, such as racial categories. But the researcher may not know a priori which strata of confounders will show interesting effect heterogeneity. Machine learning with sample splitting provides a principled solution in this setting: discover interesting subgroups in a training sample and then estimate their effects in a new test sample. For example, Athey and Imbens (2016) developed causal trees, an extension of decision trees specifically designed to uncover effect heterogeneity. In one application, Brand et al. (2021) assessed variation in the effects of college completion on low-wage work. They found that college completion reduced low-wage work most for individuals whose mothers had less than a high school degree, who grew up in large families, and who had low social control. Not only does the use of causal trees allow researchers to uncover subgroups not previously considered, it also transparently depicts the analyses that led researchers to focus on particular subgroups. When a researcher chooses manually to highlight the outcomes of a particular subgroup, it is difficult to know how they came to that decision. When a causal tree highlights a particular subgroup, the algorithm that determines the highlighted result is fully transparent.<sup>4</sup>

---

<sup>4</sup>Causal trees do not always discover effect heterogeneity. Sometimes, they reveal a surprising lack of effect heterogeneity. Handel and Kolstad (2017) analyzed a randomized health intervention and found almost no evidence of heterogeneity across the measured variables. Davis and Heller (2017) found that a randomized youth intervention in Chicago had roughly the same effect on arrests in all subpopulations studied. In general, a lack of evidence for effect heterogeneity does not mean that effects are constant for everyone, but only reveals a lack of evidence for heterogeneity as a function of the measured variables.



## 2.5. A word of caution: Machine learning, causal inference, and policy

To make the most of machine learning, social scientists must recognize what it can and cannot do. In particular, machine learning can describe the world as it exists but does not inform policy (what would happen under an intervention to change the world) in the absence of additional assumptions. Often, one needs an assumption that the statistical association between two variables derives from a causal relationship rather than from confounding. This assumption may go under various names, such as unconfoundedness or conditional independence of the potential outcomes from treatment assignment (Imbens and Rubin, 2015; Pearl, 2009; Hernán and Robins, 2021).

To illustrate causal assumptions, we first discuss an example from Kleinberg et al. (2015) about carrying an umbrella in the rain. When it rains, the people who carry an umbrella remain dry. Consider two explanations: (1) those who like carrying umbrellas are really good at dodging between raindrops and (2) an umbrella causes a person to stay dry. Explanation (1) corresponds to confounded treatment assignment—if this explanation were true, then handing an umbrella to an umbrella-less person would not keep them dry because they still would not know how to dodge the raindrops. Explanation (2) corresponds to a causal effect—handing an umbrella to an umbrella-less person will cause them to be dry. Causal claims require us to exclude the confounding explanations in favor of a causal explanation—often called an assumption of unconfoundedness.

Kleinberg et al. (2015) address the question of when to carry an umbrella. They argue that carrying an umbrella is a “prediction policy problem” where one only needs to predict whether it will rain in order to inform policy (whether or not to carry an umbrella). We would argue that the reason this is such a good example of a prediction policy problem is because the causal structure of the problem is so very straightforward: we all agree that carrying an umbrella causes people to stay dry, rather than that those who carry umbrellas are good at dodging raindrops.

But in other settings, the unconfoundedness assumption is not nearly so straightforward. For example, Chalfin et al. (2016) consider whether firing some police officers and replacing them with other officers could reduce the rate of police shootings in Philadelphia. For this policy, the central question is causal: if we took a given encounter between a police officer and a civilian but counterfactually changed the officer involved, would the probability of a police shooting decrease? The question is difficult to answer. Perhaps (1) some officers shoot more because they are assigned to particularly dangerous encounters or (2) some officers shoot more because they are simply more prone to shooting in any given encounter. The former is a confounded explanation—if Officer A moved to Officer B’s encounters, their rate of shooting would change to match that of Officer B. The latter is a causal explanation: which officer is involved causes a difference in the rate of shooting. The authors coin the term “task confounding” for scenario (1). To conclude that differences across officers are caused by differences in the officers rather than the tasks, the authors assume the absence of task confounding. Under this causal assumption, Chalfin et al. (2016) draw a causal conclusion: firing the 10% of officers with the highest propensity to shoot and replacing them with officers of average propensities to shoot would reduce shootings by

4.81 percent. Importantly, while the authors emphasize the use of predictive modeling, the conclusion rests critically on a causal assumption (the absence of task confounding).

In both the umbrella problem and the officer shooting problem, we want to know about outcomes that would be realized under a counterfactual treatment. Would I have stayed dry if I had remembered my umbrella? Would I have stayed alive if I had encountered Officer A instead of Officer B? But we only get to see one of the outcomes (Holland, 1986), thus requiring an assumption about the distribution of unobserved outcomes. In the umbrella problem, this assumption is highly plausible. In the officer shooting problem, it is much less clear. In general, the more we can confidently assume about the causal structure of a problem, the more we can rightfully focus on the predictive side of the problem. And when our causal assumptions are doubtful, it becomes all the more important to give them close attention. One of the most exciting areas of machine learning is its intersection with causal inference; for reviews and introductions, see Athey and Imbens (2017), Athey and Imbens (2019), Van der Laan and Rose (2018), and Brand et al. (2022). Broadly, the dichotomy between prediction problems and causal problems is misleading: new tools for prediction are best deployed in tandem with careful attention to underlying causal assumptions.

### 3. Conceptual building blocks: The statistical foundations of machine learning

To realize the benefits discussed above does not require years of training in machine learning. Rather, researchers trained in classical statistics already possess knowledge of the fundamental building blocks that support machine learning. This pedagogical section links machine learning to classical statistics by presenting a set of core concepts: task clarity, the bias-variance trade-off, data-driven estimator selection, interpretation, and tasks involving a new (and possibly counterfactual) target population.

#### 3.1. Task clarity: Define a precise goal

Every statistical problem begins with a task—the goal that we hope to accomplish. For instance, we might wish to make predictions in a particular setting or estimate a mean in some population. A precise statement of the task is essential in all quantitative research, and it takes on renewed importance in the context of machine learning, which can often be tailored specifically to the task at hand. For example, consider a task which has been well-studied in both statistics and machine learning: drawing inference about a target population from a sample. We discuss this task from two perspectives: estimation of unknown population parameters and prediction of out-of-sample cases.

Suppose a researcher studies academic performance for students nested within classrooms. Each student  $i$  has a test score  $Y_i$  capturing their academic performance. We would like to understand how test scores vary across classrooms,

$$\theta_j = \mathbf{E}(Y_i \mid J_i = j) \tag{1}$$

where  $J_i = j$  means we are taking the expectation among students in classroom  $j$ . Equivalently, we can conceptualize  $\theta_j$  as a prediction rule: if we see a new student in class  $j$ , we would predict that student's unknown test score to be  $\theta_j$  (Fig. 3).

If we observed all students in every classroom, we could calculate each  $\theta_j$  directly by the classroom mean. If we only observe a random sample of students, then we need an estimator for this unknown parameter. For instance, we could estimate by the sample mean,

$$\hat{\theta}_j^{\text{Mean}} = \bar{y}_j = \frac{1}{|\mathcal{S}_j|} \sum_{i \in \mathcal{S}_j} Y_i \quad (2)$$

where the term beneath the summation sign indicates that we are summing over all students  $i$  in the sample  $\mathcal{S}_j$  from classroom  $j$ . The sample mean is a consistent and unbiased estimator, yet it may not be the optimal estimator in a finite sample. We discuss this issue in the next section.

Estimation of class-specific means is a useful example because it bears resemblance to both statistics and machine learning. Social scientists and statisticians could easily study this problem without conceptualizing it as a machine learning problem. Yet it also contains several hallmarks of machine learning. Machine learning estimators often involve a very large set of parameters to be estimated (e.g., many classes) and apply in settings where the sample size seems large (e.g., many students) but in fact is small given the large number of predictors (e.g., few students in each classroom). Prediction and estimation are in one sense mathematically identical: if you only know the class identifier and you want to guess with low mean squared error the test score of a previously-unseen student (prediction for an individual), then the best you can do is to choose the population mean score in that class (which requires estimation for a population subgroup). Yet the emphases in the two cases often differ. For example, if test scores vary considerably within each class, then you might accurately estimate the class mean while making poor predictions for each individual student. One can achieve good estimation despite poor prediction. Yet the mathematical connection between the two nonetheless means that tools designed for prediction may be useful for estimation, and vice versa.

### 3.2. The bias-variance trade-off: Choose a biased estimator

Continuing the example of students in classrooms, suppose that the sample size  $|\mathcal{S}_j|$  in class  $j$  is small (e.g., 5 students). In this case, the sample mean may be a poor estimator of the population mean in the classroom because the sample size is so small. For every statistical and machine learning estimator, a first-order concern is how well we can expect that estimator to accomplish our task. We want to choose an estimator which will be close to the truth on average when applied to hypothetical samples we could take from the population. Counterintuitively, to produce an estimator which is close to the truth on average, one might be well-advised to choose an estimator which has low variance but is slightly wrong on average—a biased estimator. Many of the best statistical estimators and nearly every estimator that would be considered “machine learning” accepts some bias in

order to improve performance. To make the most of machine learning, social scientists will need to come to appreciate the benefits that bias can bring. We illustrate this point through an example which is standard in statistics: a hierarchical linear model (HLM).

To better estimate the classroom mean in a small sample, the researcher could add a shrinkage term to produce a hierarchical linear model estimator (Bryk and Raudenbush, 1992),

$$\hat{\theta}_j^{\text{HLM}} = \bar{y}_j - \underbrace{\frac{\frac{1}{n_j} \hat{\sigma}_j^2}{\frac{1}{n_j} \hat{\sigma}_j^2 + \hat{\delta}^2}}_{\substack{\text{Shrinkage Term} \\ \text{(creates bias)}}} (\bar{y}_j - \bar{y}) \quad (3)$$

where  $\hat{\sigma}_j^2$  is the empirical variance of test scores across students within class  $j$ ,  $\hat{\delta}^2$  is the empirical variance of classroom-level mean test scores across all classrooms, and  $\bar{y}$  is the mean test score in the entire sample.<sup>5</sup> The multilevel estimator  $\hat{\theta}_j^{\text{HLM}}$  is a partial pooling estimator because it pools information from class  $j$  together with other information about the mean test score in the sample overall. The consequence of partial pooling is that the estimator for each class is biased toward the overall mean—the greater the shrinkage, the more the bias. Yet shrinking toward the overall mean also yields the benefit of reduced variance. Fig. 4 shows that the amount of shrinkage in  $\hat{\theta}_j^{\text{HLM}}$  is the amount that minimizes the expected squared error of the estimator: across repeated samples, the average squared distance between the estimated mean and the truth.

The notion of accepting some optimal amount of bias in order to reduce the variance of an estimator is an idea that is much broader than multilevel models. In particular, the expected squared error of any estimator can be decomposed into components corresponding to bias and variance.

$$\text{Bias-Variance Trade-Off: } \underbrace{\mathbf{E}((\hat{\theta} - \theta)^2)}_{\text{Expected Squared Error}} = \underbrace{\mathbf{E}(\hat{\theta}) - \theta}_{\text{Bias Squared}} + \underbrace{\mathbf{E}((\hat{\theta} - \mathbf{E}(\hat{\theta}))^2)}_{\text{Variance}} \quad (4)$$

If we want our estimator to be close to the truth on average (low expected squared error), then it is often worthwhile to accept some bias in order to reduce the variance of the estimator.

The bias-variance trade-off is especially relevant in settings where the variance of an unbiased estimator is high. High-variance estimators are common when the number of parameters to be estimated is large (e.g., many class-specific means), because the amount of

<sup>5</sup>This estimator is sometimes called the Best Linear Unbiased Predictor, although that name is misleading because the estimator is biased.

data relevant to each parameter (e.g., the students in a particular class) may be small even if the overall sample size is very large. Beyond the setting of students in classrooms, the bias-variance trade-off plays a central role in other statistical problems characterized by large sample sizes but also many parameters to be estimated, such as in small-area estimation (Rao, 2003). In machine learning, the bias-variance trade-off is especially important because machine learning estimators often involve many parameters, such that variance is a serious concern even in big-data settings. Machine learning estimators resolve this problem by accepting some bias in order to reduce variance and improve expected squared error. Social scientists applying these methods should be comfortable with this acceptance of bias just as they are already comfortable with bias in classical statistical settings, such as multilevel models (Bryk and Raudenbush, 1992) or any setting in which we regularize estimates to reduce variance. The existence of bias should not be a barrier to the adoption of machine learning.

### 3.3. Data-driven estimator selection: Automate what can be automated

Analytic choices abound in quantitative social science. For example, the choice of a model specification is a central question in classical statistics. Researchers have traditionally approached this question by some combination of conceptual argument paired with empirical metrics of model fit, such as  $R^2$ . A machine learning perspective transfers the weight of these choices in the direction of empirical evidence. To the degree that data can inform the choice of estimator, machine learning approaches allow the data to speak.

Fig. 5 illustrates data-driven estimator selection in a simulated setting. The predictor variable  $X$  is related to the outcome  $Y$  by a complicated conditional mean function, as is likely to be the case in many realistic settings. Not knowing this function in advance, the researcher might consider several possible estimators with different assumed functional forms (e.g., various OLS specifications) or different procedures to learn the functional form from the data (e.g., a regression tree and a Generalized Additive Model). A social scientist might report the results of all these specifications. Despite the inclusion of machine learning estimators like regression trees, this overall research approach could be considered “classical” in the sense that it involves choosing the estimator or estimators for conceptual rather than data-driven reasons. An approach more inspired by machine learning might instead seek to empirically score the performance of the estimators in order to make a data-driven choice. The metric by which an estimator is evaluated is often called a *loss function*, which formalizes what it means for an estimator to perform poorly (and by extension, what it means to perform well). For instance, one loss function would take an estimator  $\hat{\theta}_s$  estimated in a sample  $\mathcal{S}$  and score it by its mean squared error when predicting new observations from the population.

$$\text{Loss Function: } \mathcal{L}(\hat{\theta}_s) = \mathbf{E}_{i: i \notin \mathcal{S}} \left( (\hat{\theta}_s(x_i) - y_i)^2 \right) \quad (5)$$

In practice, we do not observe the full population and thus must rely on an estimate  $\widehat{\mathcal{L}}()$  of the loss function. Suppose we take our sample  $\mathcal{S}$  and randomly assign observations into two

equally-sized samples: a training sample  $\mathcal{S}_{\text{Training}}$  and a test sample  $\mathcal{S}_{\text{Test}}$  (Fig. 5 Panel C). We then learn the prediction function in the training sample and estimate the loss function in the test sample.

$$\text{Estimated Loss Function: } \widehat{\mathcal{L}}(\widehat{\theta}_s) = \frac{1}{|\mathcal{S}_{\text{Test}}|} \sum_{i \in \mathcal{S}_{\text{Test}}} (\widehat{\theta}_{\mathcal{S}_{\text{Training}}}(x_i) - y_i)^2 \quad (6)$$

Finally, we can choose the estimator for which the estimated loss function  $\widehat{\mathcal{L}}(\widehat{\theta}_s)$  is as close as possible to zero. In the simulated example of Fig. 5, this procedure selects the Generalized Additive Model estimator. In this setting, it is visually apparent in Fig. 5 Panel B that this is the best estimator. But in non-simulated settings, the true conditional mean function (the gray curve in Fig. 5 Panel B) is unknown and out-of-sample predictive performance can still help the researcher choose an estimator which comes as close as possible to that unknown function.

While data-driven estimator selection is a hallmark of machine learning, it is also in full alignment with standard statistical procedures. Social scientists already compare models by empirical scores such as  $R^2$ , likelihood ratios, the Akaike information criterion (AIC, Akaike, 1973), the Bayesian information criterion (BIC, Schwarz 1978), and numerous other scores. Each of these can be interpreted as a loss function for data-driven model selection. When carried out within machine learning, the loss function is typically evaluated on data not used to estimate the model in order to assess the ability of the model to generalize to new observations.

We have taken care to distinguish the true loss function  $\mathcal{L}()$  from the estimated loss function  $\widehat{\mathcal{L}}()$  because the estimated loss function may be statistically uncertain, especially if it evaluated on a small sample. An estimator which is inferior in the population may outperform another estimator in the test sample because of the chance of which cases from the population happen to appear in the test sample. One way to improve the precision of  $\widehat{\mathcal{L}}()$  is to conduct cross validation, a procedure in which the full sample  $\mathcal{S}$  is partitioned into a set of  $k$  folds  $\mathcal{S}_1, \dots, \mathcal{S}_k$ , each of which plays the role of  $\mathcal{S}_{\text{Test}}$  in turn.

$$\text{Cross-Validated Estimate: } \widehat{\mathcal{L}}_{\text{CV}}(\widehat{\theta}_s) = \underbrace{\frac{1}{k} \sum_{f=1}^k}_{\text{Average over } k \text{ folds}} \left( \underbrace{\frac{1}{|\mathcal{S}_f|} \sum_{i \in \mathcal{S}_f} (\widehat{\theta}_{\{\mathcal{S}_f\}^c}(x_i) - y_i)^2}_{\text{Average squared error in fold } \mathcal{S}_f \text{ from estimates that pool all other folds}} \right) \quad (7)$$

Cross-validation has long been used in statistics (e.g., Stone 1974) and is common in machine learning today. There are two advantages of cross-validation over a single split into training and test samples. First, each fold-specific error estimate is trained on a sample with a fraction  $\frac{k-1}{k}$  of all observations. For large  $k$ , the training sample size in each cross-validated fold thus approximates the full sample size  $|\mathcal{S}|$ , which is useful if the

researcher will ultimately draw inference by training a model on the full sample. Second, in cross-validation all observations play the role of the holdout at some point, which potentially improves the statistical precision of the estimated loss function.

Both sample splitting and cross validation yield a signal about the empirical merits of various candidate models. One word of caution is that those signals can themselves be noisy. We might like to know how Model A and Model B would perform when predicting an arbitrarily large set of previously-unseen observations. But we estimate their out-of-sample performance using an actual sample, just as one would estimate a population mean from a sample. And just as a population mean estimated in a sample ought to be accompanied by a confidence interval, a mean squared error estimated in a sample perhaps also ought to be accompanied by a confidence interval. Regardless of whether one reports confidence intervals for model performance metrics, one should be aware that those metrics themselves have statistical uncertainty arising from sampling. Just because Model A is best in our observed out-of-sample data does not necessarily imply that Model A would be best if assessed on an entire population of out-of-sample data, even if our sample comes from the same data generating process as that population.

#### 3.4. Interpretation: Report the target estimate, not a model parameter

A barrier to the adoption of machine learning methods is that they are “black box.” That is, researchers perceive a loss of interpretability if they use these methods. This barrier may loom particularly large for social scientists who ordinarily summarize their models by tables of regression coefficients, which is not possible for many machine learning methods. We emphasize two responses to this concern. First, for machine learning methods there exists a single-number summary analogous to a regression coefficient: the average partial effect. Second, social scientists’ comfort with regression coefficients may be misplaced: coefficients are more difficult to interpret than some may believe.

A regression coefficient is easy to present: it summarizes a relationship between two variables with a single number. Fig. 6 Panel A presents a simulated example with 10 data points, where the relationship between the predictor  $X$  and the outcome  $Y$  could be summarized by the slope of a regression line:  $\hat{\beta} = 1.061$ . Conventionally, one might say that a unit increase in  $X$  is associated with a 1.061 increase in  $Y$ . By the assumption of a line, the slope 1.061 applies to every data point, regardless of the value of  $X$ . But what if the line is a poor approximation? Fig. 6 Panel B presents a smooth curve estimated by a thin-plate spline (Wood, 2017). Unlike a regression line, no single coefficient summarizes the shape of the thin-plate spline. Yet we can arrive at a summary which is analogous to the regression coefficient. At each data point, estimate the slope of the curve at that data point. Then, average over all data points. In this example, the average partial effect estimate is 1.513, substantially higher than the regression coefficient. Beyond smooth curves, the notion of an average partial effect or average first difference applies broadly as a tool to summarize prediction functions. Take a key predictor, add a small  $\Delta$  to each data point, record the average change in the predicted values, and divide by  $\Delta$  to yield an estimate of the average responsiveness of the outcome per unit change in the key predictor. Almost any prediction

function can thus be summarized with one number which is as interpretable as a regression coefficient.

Importantly, an average partial effect from a flexible estimator may actually correspond more closely to what researchers aim to know. The heuristic interpretation of a regression coefficient—on average over units, the change in  $Y$  for a small change in  $X$ —corresponds to the average partial effect but may not correspond to the regression coefficient. In our example, the pattern is nonlinear and the data are skewed, with a higher density of observations at lower values of  $X$ . The regression coefficient is heavily influenced by the sparsely-populated high values of  $X$  (where the slope is flatter) while most of the data is in the densely-populated low values of  $X$  (where the slope is steeper). Thus, the regression coefficient is closer to 0 than the slope averaged over points where the data exist. When a line is a poor approximation to a statistical pattern, an average partial effect may better correspond to the quantity that social scientists want. More generally, the approach above can generalize to multivariate settings where other variables are held constant, in which one can make predictions at the observed values of confounders  $\vec{X}$  while examining the average partial effect of a change in a treatment variable  $T$ .

Average partial effects offer single-number summaries of statistical patterns analogous to regression coefficients. But there is a further reason why the familiarity of regression coefficients should not hold us back: familiar interpretations of regression coefficients are often misleading in two ways. First, it is rare that an entire table of regression coefficients is interpretable. When a regression model includes as predictors a treatment variable as well as pre-treatment covariates sufficient to block confounding, then only the coefficient on the treatment is interpretable in causal terms—the coefficients on the pre-treatment covariates are not causal effects at all. Second, even the coefficient on the treatment may be misleading if the assumed functional form is incorrect. Fig. 7 gives a simple example. Suppose we are interested in the causal effect of precipitation on outdoor exercise. We sample 5,000 residents on random days in January in each of two locations: Vail, Colorado and Phoenix, Arizona. We then record whether it precipitated on the day in question and whether they engaged in outdoor exercise. When it does not precipitate, 50% of respondents in each city exercise outdoors. When it does precipitate, the outcomes are very different. In Vail, the precipitation is snow and outdoor exercise jumps to 90% as residents hit the ski slopes. In Phoenix, the rare days of precipitation are rain and residents stay indoors: only 10% exercise outside. Assuming that location is the only confounder of precipitation in this example, we could say that precipitation increases outdoor exercise by 40 percentage points in Vail but reduces it by 40 percentage points in Phoenix. Because the two balance out, the average causal effect in our sample is zero.

Now suppose that we estimate an OLS regression model with precipitation and city entered additively as predictors. The model would be misspecified due to the omission of an important interaction: the effect of precipitation is very different in Vail and in Phoenix. With this interaction omitted, the coefficient on precipitation does not estimate the average causal effect. In fact, the resulting estimate suggests that precipitation increases outdoor exercise by 18 percentage points. The reason the coefficient is much closer to the Vail effect



than to the Phoenix effect is because an OLS coefficient in this setting estimates a weighted average causal effect with weights proportional to the variance of the treatment variable (precipitation) within strata of the confounders (city, see Elwert and Winship 2010, Brand and Thomas 2013, and Aronow and Samii 2016). Because it precipitates on 48% of days in Vail but only 10% of days in Phoenix, there is more information about the effect of rain in Vail. OLS therefore maximizes efficiency by placing 76% of the weight on Vail and only 24% of the weight on Phoenix, so the coefficient is much closer to the effect in Vail. This example is extreme—the treatment variance and causal effects both differ dramatically across strata. While the problem may be smaller in other settings, it remains the case that regression coefficients do not estimate average causal effects in the absence of strong faith in the assumed functional form. Further, one can avoid this problem by estimating a more flexible model including the interaction term. The average causal effect could then be estimated by a discrete version of the average partial effect: take every observation, predict the outcomes under precipitation and no precipitation, difference, and average over the sample. That procedure is equally valid for regression and machine learning strategies, as long as the underlying causal assumptions (required for both) are valid.

More broadly, the appeal of regression coefficients is that these model parameters might equal the quantities of interest under simplifying assumptions. That is no longer the case with machine learning estimators, which may not involve coefficients. Yet if one assumes that many regression models were misspecified to begin with, then coefficients were not guaranteed to correspond to the quantities of interest. Machine learning therefore presents an opportunity to estimate a realistic model, complete with interaction terms and nonlinearities which would complicate the interpretation of regression coefficients. Then, one can use the model to predict the data needed to estimate the target quantity, thus producing an interpretable estimate of the target parameter.

### **3.5. A return to task clarity: Prediction in new populations and prediction for causal inference**

Our first conceptual building block was task clarity—being precise about the goal of the quantitative exercise. To re-emphasize the importance of task clarity, we now turn from standard out-of-sample prediction tasks to a range of more complex tasks. We discuss two settings that demonstrate the importance of task clarity: prediction in a new target population (where statistical patterns may differ from the training population) and prediction for causal inference.

To consider prediction in a new target population, suppose we study a cohort of students entering Statsville West High School in 2017. For each student, we observe many variables about academic performance in 8th grade and we observe whether they drop out of high school over the next four years. Using a machine learning algorithm, we learn a function to predict high school dropout. Impressed by our model, the principal of Statsville West suggests that for the entering cohort of 2022 we predict the likelihood of dropping out for each student, so that the principal can target extra counseling resources to those students. Perhaps the principal of Statsville East High School also hears about our model and wants to deploy it in that context as well. For each of these use cases, there is a danger: the population

that entered Statsville West in 2017 is not the same as the population entering in 2022, and is surely different from the population entering Statsville East in 2022. The mapping between the predictors and the outcome in these new populations may not be the same as the mapping in the original population on which the algorithm was learned (i.e., Statsville West, entering in 2017). The problem of Statsville West and Statsville East is ubiquitous across real applications of machine learning. Researchers routinely learn things in one context in the past, and then apply what they have learned in the future and possibly in new contexts. To use statistics and machine learning responsibly, one must be aware when there is a leap when we extrapolate to a new target population (see Fig. 8).

Prediction in a new target population is especially relevant for causal inference (Fig. 9). Suppose the principal of Statsville West had already implemented a program to offer extra counseling to some students in the 2017 entering cohort. After observing whether those students dropped out, the principal wants to predict whether those who did not receive the program would have benefited if the program had been available to them. But those who did not receive the extra counseling are by definition not part of the learning population from whom we drew the sample. In fact, it is impossible to sample people who did not receive counseling and observe the outcome they would have realized if they had received the counseling (i.e., this is the fundamental problem of causal inference, Holland 1986). To learn about what would have happened if other students had received extra counseling, the principal is necessarily requiring the researcher to make predictions in a new population. Absent additional assumptions, prediction for causal questions *a/ways* involves a target population which is different from the learning population. Only by an assumption can we generalize from the learning population to the target population (Hartman et al., 2015; Pearl and Bareinboim, 2011; Stuart et al., 2015; Xie, 2013). For instance, we might assume that the potential outcome under treatment  $Y_i(1)$  follows the same distribution among the treated units as among the untreated units, within each subpopulation defined by a set of predictor values. By this assumption, any mapping  $\vec{X}_i \rightarrow Y_i(1)$  learned in the learning population will still be valid in the target population.

Yet even in the best-case scenario, causal inference for policy prescriptions often involves an additional leap to a new target population (see Fig. 10). Suppose the principal randomly assigned counseling to students entering Statsville West in 2017. But then, the principal wants to use these results to justify the expansion of counseling support for the cohort entering in 2022. Despite strong internal validity for the causal effect estimate in the 2017 cohort, the principal still must leap to a new population to deploy the policy in the 2022 cohort. The leap from the training population to the target population is therefore particularly relevant to causal policy prescriptions.<sup>6</sup>

In fact, there is often a trade-off between internal and external validity, where one can study a population less like the target population in a randomized design (high internal validity)

---

<sup>6</sup>The assumption to draw causal inference in the target population is  $\{Y(0), Y(1)\} \perp \{P, S, T\} | \vec{X}$ , where  $P$  indicates membership in the learning versus target population,  $S$  indicates inclusion in the sample of cases,  $T$  indicates treatment assignment, and  $\vec{X}$  denotes the vector of pre-treatment predictors. One setting where this would hold is if the target population is the learning population and  $S$  and  $T$  are randomly assigned.

or a population more like the target population in an observational study (high external validity). Perhaps the Statsville principal has a randomized experiment from a very old cohort that entered in 2000 and an observational study on the cohort that entered in 2017. It would not be clear which study would be more informative for a policy prescription applying to the cohort entering in 2022. Every study has limitations, and the leap from a learning population to a different target population is a limitation of which one must always be aware.

#### 4. The future: Guiding principles and promising approaches

We proceed cautiously in predicting how machine learning will be used in the future. In this section, we nonetheless offer a few guiding principles and approaches which we believe hold promise for resolving particular issues in social science research. These guiding principles are relevant to both descriptive and causal claims, but our discussion often emphasizes the relevance to causal claims because we see an especially promising space of applications in that domain.

##### 4.1. Resolving p-hacking: The promise of automated model selection

The replication crisis raises questions as to the validity of common quantitative social science research practice (Freese and Peterson, 2017; Simmons et al., 2011). A key source of concern is the practice in which researchers iterate between model fitting and interpretation until arriving at a chosen specification which is reported to the reader. This procedure creates many opportunities for a well-meaning researcher to select the model for which the results most align with the researcher's preexisting beliefs (Gelman and Loken, 2014). The (possibly unintentional) practice of choosing an estimator based on the results undermines the validity of  $p$ -values and confidence intervals, which are designed under the assumption that the researcher follows a single procedure that would be applied the same way in any hypothetical sample.

Machine learning may seem to amplify this problem: with more candidate estimators, researchers who stay the course will simply have more opportunities to select their preferred result. Yet automated model selection offers a way out of this problem. Before analyzing the data, researchers can specify a single decision rule for choosing among many candidate estimators. For instance, we might choose the one with the lowest cross-validated mean squared error. By defining the decision rule before viewing any results, researchers can remove the danger of choosing a result based on their preferred specification. Beyond selection of the single best model, there are numerous possible decision rules to combine several candidate learners into one aggregate prediction function, including stacking (Wolpert, 1992), boosting (Schapire and Freund, 2012), and Bayesian model averaging (Raftery et al., 1997). One promising ensemble method is Super Learner (Van der Laan et al., 2007), which accepts a dataset and a set of candidate learners as arguments and returns a single prediction function which is a weighted average of those learners with weights learned through cross-validation. Super Learner is available in open-source software for R, both in the `SuperLearner` package (Van der Laan et al., 2007) and in the `s13` package (Coyle et al., 2021) which is part of the `tlverse`.

An important caveat to automated model selection is that not all analytical choices can, or should, be automated. The choice of a target quantity requires an argument from the researcher about why that target quantity would matter for theory (Lundberg et al., 2021). For causal target quantities in observational data, the set of variables needed for adjustment cannot be chosen from data because the choice necessarily involves unobserved counterfactual quantities; the set of variables must instead be chosen by argument about the underlying causal model (Pearl, 2009). For example, a variable that is a consequence of the treatment may itself affect the outcome. Yet, such a variable should not be held constant when estimating the average causal effect of the treatment, because conditioning on it would block a causal pathway from the treatment to the outcome and may unblock other non-causal pathways. Our theory about the causal ordering of the variables, not statistical evidence, informs the decision to exclude this covariate from the adjustment set.

There exist some settings where empirical evidence can inform the causal model—for example, if a pre-treatment variable is independent of treatment given other confounders, then that variable may not be needed for confounding adjustment.<sup>7</sup> But as a general principle, the choice of adjustment variables is often better decided by theoretical argument rather than empirical evidence. Two pieces of the research process—defining the question and defending assumptions about counterfactuals—rest heavily on the researcher’s conceptual argument. The piece which can be more easily automated involves questions about estimation, such as whether or not an interaction between two predictors should be included (Brand et al., 2021).<sup>8</sup> For a given research question and a given set of variables, a researcher who defines a decision rule can use data to automate the choice of a prediction function using those variables. By doing so, the researcher removes some of the opportunities for *p*-hacking.

#### 4.2. Resolving approximate models: The promise of an agnostic perspective

Researchers often begin by assuming a model. All statistical properties (parameter definitions, standard errors, confidence intervals, etc.) are then valid only by the assumption that the model is correct. Yet social scientists and statisticians have long accepted that “all models are wrong” (Box, 1976:792). If we believe that all models are wrong, then under the standard perspective all statistical properties of those models are thus unreliable as well. One resolution to this conundrum is a more balanced perspective known as an agnostic approach: any statistical analysis is understood as an approximation to more complex phenomena. For example, past work in this perspective has formalized the statistical properties of ordinary least squares as a best linear approximation to a nonlinear or interactive function (see e.g. Lin 2013; Buja et al., 2019a,b; Aronow and Miller 2019). From an agnostic perspective the question is not whether we have the correct model, but rather whether one can provide evidence and argument that the chosen model is likely to usefully approximate the target quantity of interest (see Grimmer et al., 2021). A researcher who defines this target quantity

<sup>7</sup>Imbens and Rubin (2015) Ch. 13 discusses an iterative procedure for selecting confounders based on their statistical relationship with the treatment. The Stata package *ITPSCORE* (Moore et al., 2021) implements this type of iterative propensity score specification. Researchers still select the input variables. With *ITPSCORE*, researchers can choose to bypass allowing the algorithm to eliminate any input covariates. For a LASSO-based approach to variable selection, see Belloni et al. (2014).

<sup>8</sup>The Imbens and Rubin (2015) iterative procedure and *ITPSCORE* (Moore et al., 2021) also selects higher order and interaction terms.

or estimand outside of the model can then begin to reason about the relative merits of various approaches (Lundberg et al., 2021).

While the agnostic perspective is not new, it gains new importance with machine learning methods. With many machine learning methods, it is difficult to argue on conceptual grounds that the learned function is “correct.” For example, while one could reason about the relative merits of an assumed functional form for ordinary least squares, it is more difficult to reason conceptually about whether a random forest has been given a large enough sample to arrive at a “correct” representation of the response surface. These types of flexible learners have asymptotic guarantees, but one never knows if one’s sample is large enough to lean heavily on those guarantees. Here the notion of a useful approximation becomes essential: for at least some target quantities, one can marshal evidence about out-of-sample predictive performance to indirectly support the claim that the chosen algorithm will yield a good approximation to the target quantity.

The agnostic approach is related to a difference in worldview between classical statistics and machine learning. Classical approaches often emphasize the conceptual merits of a model, relying heavily on the assumption that the model generates the observed data. Machine learning approaches may never seek to learn the true model, but rather emphasize empirical evidence that the model performs well at a given task. In fact, Donoho (2017) argued that the “secret sauce” of machine learning is to precisely specify a task, open that task to all approaches, and select the best algorithm by a well-defined metric involving out-of-sample prediction. The best algorithm performs the task well, and it may or may not accurately represent the entire process that generated the data (Breiman, 2001b).<sup>9</sup> Of course, what it means to “perform well” is part of the definition of the task, and is subject to researchers’ theoretical and normative choices. Social scientists may benefit from adopting elements of an algorithmic machine learning perspective. One element of that perspective is an agnostic approach: instead of seeking the correct model, we seek an approximation for which there is good reason or empirical evidence to expect the empirical properties that are desirable for our research question.

#### 4.3. Resolving extrapolation: The promise of local estimators

Extrapolation is an ever-present danger in globally parametric models like ordinary least squares.<sup>10</sup> Extrapolation occurs when a data point to be predicted is far from the mass of the training data, so that the predicted value may depend heavily on the assumed functional form (e.g., a line). Another side of the problem is influence. Influence is the converse, when a training point far from the mass of the data heavily shapes the fitted prediction function. Extrapolation and influence are two consequences with the same source: the assumption of global parametric models (e.g., the assumption of a line). Local estimators offer a solution to the problem: only allow each unit  $j$  to contribute to the estimate for unit  $i$  to the degree that

<sup>9</sup>As an example of the algorithmic modeling culture, data scientists often emphasize algorithms as a set of procedures applied to data and the conditions under which those procedures perform well. See for example Wu et al. (2008).

<sup>10</sup>As used here, a globally parametric model is one where information is shared along an assumed functional form such that statistical patterns in one part of the space are taken to be informative about patterns in a very far removed part of the space. For instance, ordinary least squares is a globally parametric model where the pattern at the lowest values of a predictor informs the estimated pattern at the highest values of that predictor via the assumption of linearity, thus producing a substantial possibility of extrapolation.

unit  $j$  is “near” to unit  $i$ . For every local estimator, the central question is what it means for two units to be “near” each other. New advances in local estimation are thus most powerful paired with conceptual social science argument for the chosen definition of “near.”

Propensity score matching for causal inference is one example of a local estimator (Imbens, 2015; Morgan and Harding, 2006). Suppose we know the probability of treatment  $p_i$  (also known as the propensity score) given the values of confounding variables for unit  $i$ . If unit  $i$  is treated, we might estimate the potential outcome under control  $Y_i(0)$  by the outcome of the untreated unit  $j$  with propensity score  $p_j$  closest to unit  $i$ . This is a local estimator because only the nearest untreated unit contributes to the estimate for unit  $i$ . Propensity score matching is a nearest neighbors estimator (Fix and Hodges, 1989): the unit or units nearest to the focal unit contribute to the estimate.

Nearest neighbors and other local estimators depend crucially on the definition of “near.” There are many ways to define what it means to be near. In propensity score matching, the distance between any pair of units may be defined as the difference in their probabilities of treatment  $p_i - p_j$ , which is a univariate summary of the difference in their confounder sets  $\vec{X}_i$  and  $\vec{X}_j$ . Or we could also define nearness as a function of the confounders  $\vec{X}_i$  and  $\vec{X}_j$  directly, as is the case for Manhattan distance (sum of absolute differences over all covariate values), Euclidean distance (sum of squared differences), or Mahalanobis distance (a generalization of Euclidean distance which incorporates the covariance among  $\vec{X}$ , Mahalanobis 1936). For each of these distances, one can define a local estimator by averaging across units which are “near” the focal unit by the chosen distance metric.

The definition of nearness is consequential: units that are “near” by one metric may be far apart by another metric.<sup>11</sup> Future research with local estimators will need to reason carefully about the definition of “near” that is relevant to the problem at hand. For instance, the covariate balancing propensity score (CBPS) (Imai and Ratkovic, 2014) modifies the propensity score to optimize balance along the covariates. Entropy balancing (Hainmueller, 2012) optimizes matches such that first, second, or higher moments of the covariates are similar across matched units. No distance metric is inherently superior to another outside of a specific application—they are all different definitions of what it means to be “near.”

Machine learning tools offer new ways to define the distance between any pair of observations. Random forests (Breiman, 2001b) are one example. A random forest is an algorithm which repeatedly (1) randomly samples a subset of predictors from the data, (2) randomly samples observations from the data with replacement, and (3) partitions the resulting sample into a set of “leaves” which are cells defined by the predictor variables and for which the outcome  $Y$  is relatively homogeneous. Each iteration produces a tree, and the average of all the trees is a forest. As highlighted by Lin and Jeon (2006), the random forest can be interpreted as a weighted nearest neighbors estimator, where units  $i$

<sup>11</sup>In fact, when the predictor set  $\vec{X}$  is high-dimensional (i.e., containing many unique values), it is possible that every unit is in some sense quite far from all other units. In causal inference, this can create a setting where arguably there is no untreated unit which is comparable to any given treated unit (D’Amour et al., 2021).

and  $j$  are “near” to each other proportional to the frequency that they fall in the same leaf. The connection is powerful because it connects random forests (a machine learning tool) to a setting well-studied in classical statistics (weighted means). Wager and Athey (2018) exploit this connection to derive asymptotically-valid confidence intervals for estimates from random forests, drawing on results from classical statistics (Hájek, 1968; Hoeffding, 1948). The notion of random forests as adaptive nearest neighbors estimator generalizes to many problems (Athey et al., 2019), such as using random forests to define nearness for weighted local linear regression (Friedberg et al., 2021).

Local estimators hold great promise for future social science research. The barriers to adoption are low: many of the advances discussed above are implemented in open-source R software, including `cbps` for the covariate balancing propensity score (Fong et al., 2021), `eba1` for entropy balancing (Hainmueller, 2014), `ranger` for random forests (Wright and Ziegler, 2017), and `grf` for generalized random forests (Tibshirani et al., 2018). The open task for social scientists is to motivate the chosen definition of “near” with respect to their substantive problem.

#### 4.4. Resolving poor convergence: The promise of targeted learning

Flexible machine learning estimators such as random forests can approximate unknown conditional mean functions  $E(Y | \vec{X})$  without the strong parametric assumptions common to classical methods like generalized linear models. Yet flexibility comes at a cost: the rate at which adaptive estimators converge toward the conditional mean is slower than the rate achieved by parametric methods. Targeted learning (Van Der Laan and Rubin, 2006; Van der Laan and Rose, 2018) resolves this convergence problem. While one cannot generally achieve fast convergence for the full conditional mean function, it is often possible to target the estimator and achieve fast convergence rates for a low-dimensional parameter of social science interest.

For concreteness, suppose we are interested in the population-average potential outcome  $E(Y(t))$  that would be realized if a treatment variable  $T$  were assigned to the value  $t$ . We make the causal assumption that a set of measured variables  $\vec{X}$  is sufficient to block all confounding, and we proceed by predicting the outcome  $Y$  as a function of the treatment  $T$  and confounders  $\vec{X}$ . Our causal target parameter can be rewritten as a particular aggregation of a statistical function.

Expected outcome  
under treatment  $t$   
(**low-dimensional**)

Conditional mean  
(**high-dimensional**)

$$\mathbf{E}(Y(t)) = \mathbf{E} \left( \mathbf{E} \left( Y \mid T = t, \vec{X} \right) \right)$$

(8)

The target  $\mathbf{E}(Y(t))$  is just one number. But the internal conditional expectation  $\mathbf{E}(Y \mid T = t, \vec{X})$  is high-dimensional because the confounders  $\vec{X}$  have many unique values. For this reason, a flexible machine learning estimator of  $\mathbf{E}(Y \mid T = t, \vec{X})$  may achieve good properties only at an extremely large sample size.

Targeted learning (Fig. 11) is a strategy to improve performance by incorporating the ultimate target into the process: the full response function  $\mathbf{E}(Y \mid T = t, \vec{X})$  is only relevant to our question insofar as we ultimately will aggregate over  $\vec{X}$  to estimate the target parameter. Begin by estimating a prediction function  $\hat{g}(T, \vec{X}) \approx \mathbf{E}(Y \mid T, \vec{X})$  to predict outcomes. Ultimately, we would like to make predictions for all units under the (possibly counterfactual) treatment  $T = t$ . Yet here is a problem: suppose a stratum  $\vec{X} = \vec{x}$  of the confounders contains only a few treated units ( $T = t$ ) but also contains many untreated units ( $T \neq t$ ). A naive prediction function will not optimize for prediction in this stratum, because few treated units are observed in the stratum. Yet, in the target population this stratum may be very important because it is home to many untreated units. Ideally, we would modify the prediction function  $\hat{g}(T, \vec{X})$  to place greater weight on accurate prediction in the spaces of  $\{T, \vec{X}\}$  where we will be making predictions. A targeted learning estimator achieves that ideal by incorporating a model for the conditional probability of treatment  $\hat{m}(t, \vec{X}) \approx \mathbf{P}(T = t \mid \vec{X})$ . For each unit with factual treatment  $T = t$ , the inverse  $\frac{1}{\hat{m}(t, \vec{X})}$  is a weight which captures the prevalence of units like this one in the target population relative to their prevalence in the available sample. There are many methods to incorporate the inverse probability of treatment to improve performance for an estimated parameter, including augmented inverse probability weighting (Robins et al., 1994; Robins and Rotnitzky, 1995) and double machine learning (Chernozhukov et al., 2018).<sup>12</sup> In targeted learning, the inverse probability weight is used as a covariate in a new regression model to predict the outcome, with the initial prediction included as an offset term (a known intercept). To the degree that

<sup>12</sup>Appendix Fig. 13 presents double machine learning (Chernozhukov et al., 2018), and Appendix Fig. 12 presents targeted learning for a continuous outcome to support direct comparisons with double machine learning.



outcomes trend upward or downward as a function of the inverse probability of treatment, that trend can be corrected to produce a new prediction function optimized for prediction in the counterfactual target population.

Targeted learning should be more widely applied in social science due to its several advantages. First, the applicability of machine learning in social science is limited by the need for massive samples for flexible estimators to perform well. Targeted learning brings down this barrier by improving convergence for the target parameter (in this case,  $E(Y(t))$ ). Second, targeted learning is superior to other methods such as augmented inverse probability weighting (Robins et al., 1994; Robins and Rotnitzky, 1995) and double machine learning (Chernozhukov et al., 2018) because targeting (Step 4 in Fig. 11) can involve a generalized linear model with a link function (e.g., logit). Thus, targeted learning never makes predictions outside the support of the outcome variable. Third, targeted learning comes with statistical guarantees of consistency and asymptotic normality (Van der Laan and Rose, 2018). Finally, targeted learning is accessible: for common target parameters, the `tlverse` suite of R packages supports this approach.

## 5. Conclusion

Machine learning creates abundant opportunities in social science. Where researchers might have coded a few documents by hand, we can now amplify that coding to tens of thousands of documents. Where researchers might have summarized a vast dataset by a set of coefficients, we can now use methods for discovery to target attention toward interesting patterns we may have missed. Where researchers might have assumed a regression specification, we can now learn interactions and nonlinearities directly from the data.

Social science is an inherently human endeavor. A researcher has to carefully consider what questions to ask, how data have been measured, and what assumptions are needed to answer those questions. Machine learning empowers researchers to do more; it does not replace us. The most effective uses of machine learning are likely to be in settings where social scientists can define a clear aspect of the problem to usefully outsource to an algorithm. One subfield that has already seen this synergy is causal inference, where researchers use argument to translate causal questions to statistical parameters (Pearl, 2009) before outsourcing the estimation of those parameters to flexible machine learning tools (Van der Laan and Rose, 2018). Machine learning is likely to have the most impact on problems for which researchers successfully partition the human and machine components of research.

Aspects of machine learning are not really new, since at its foundation it is just new developments in statistics. After all, the basics are the same: variables, assumptions, and patterns in data. Yet we have argued throughout this manuscript that there are advantages to machine learning that extend beyond traditional approaches. Indeed, technological improvements increasingly render machine learning tools attractive alternatives to classical methods. Conversely, some social scientists may perceive that the technical barriers to using machine learning remain too high to surmount. Yet barriers to adoption are plummeting with

accessible, open-source software. We predict that the benefits will rapidly outpace the costs to widespread adoption of machine learning tools.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

Replication code is available on Dataverse: <https://doi.org/10.7910/DVN/UVO6Z3>. For helpful discussions and feedback relevant to this project, we thank Brandon Stewart, Xiang Zhou, and the Social Inequality Data Science (SIDS) Lab at UCLA. Research reported in this publication was supported by the National Science Foundation under Award Number 2104607 and by the Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health under Award Number P2CHD041022.

## References

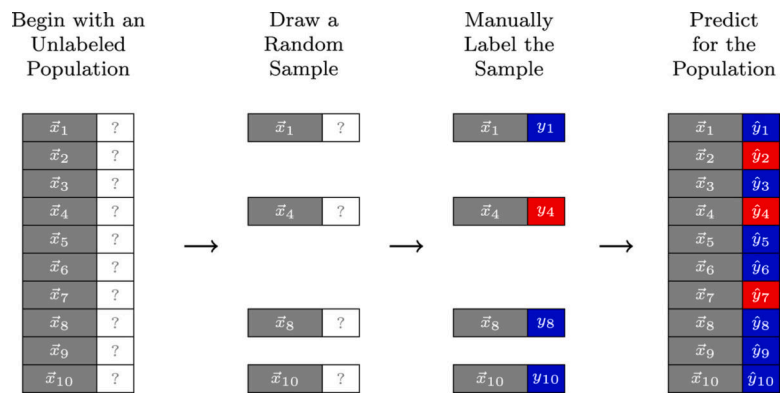
- Ahrens A, Hansen CB, Schaffer ME, 2018. PDSLASSO: Stata Module for Post-selection and Post-regularization OLS or IV Estimation and Inference. Statistical Software Components. Boston College Department of Economics.
- Akaike H, 1973. Information theory and the maximum likelihood principle. In: Petrov BN, Csaki F (Eds.), 2nd International Symposium on Information Theory. Akademiai Kiado, Budapest.
- Aronow PM, Miller BT, 2019. Foundations of Agnostic Statistics. Cambridge University Press.
- Aronow PM, Samii C, 2016. Does regression produce representative estimates of causal effects? *Am. J. Polit. Sci.* 60 (1), 250–267.
- Athey S, Imbens G, 2016. Recursive partitioning for heterogeneous causal effects. *Proc. Natl. Acad. Sci. USA* 113 (27), 7353–7360. [PubMed: 27382149]
- Athey S, Imbens GW, 2017. The state of applied econometrics: causality and policy evaluation. *J. Econ. Perspect.* 31 (2), 3–32. [PubMed: 29465214]
- Athey S, Imbens GW, 2019. Machine learning methods that economists should know about. *Annual Review of Economics* 11, 685–725.
- Athey S, Tibshirani J, Wager S, 2019. Generalized random forests. *Ann. Stat.* 47 (2), 1148–1178.
- Bail CA, 2014. The cultural environment: measuring culture with big data. *Theor. Soc.* 43 (3), 465–482.
- Belloni A, Chernozhukov V, Hansen C, 2014. Inference on treatment effects after selection among high-dimensional controls. *Rev. Econ. Stud.* 81 (2), 608–650.
- Bisbee J, 2019. BARP: improving Mister P using Bayesian additive regression trees. *Am. Polit. Sci. Rev.* 113 (4), 1060–1065.
- Bishop CM, 2006. *Pattern Recognition and Machine Learning*, ume 4. Springer.
- Blei DM, Ng AY, Jordan MI, 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.* 3 (Jan), 993–1022.
- Box GE, 1976. Science and statistics. *J. Am. Stat. Assoc.* 71 (356), 791–799.
- Brand JE, Koch B, Xu J, 2020. *Machine Learning*. SAGE, London.
- Brand JE, Thomas JS, 2013. Causal effect heterogeneity. In: *Handbook of Causal Analysis for Social Research*. Springer, pp. 189–213.
- Brand JE, Xu J, Koch B, Geraldo P, 2021. Uncovering sociological effect heterogeneity using tree-based machine learning. *Socio. Methodol.* 51 (2), 189–223.
- Brand JE, Zhou X, Xie Y, 2022. *Developments in Causal Inference and Machine Learning*. Working Paper.
- Breiman L, 2001a. Random forests. *Mach. Learn.* 45 (1), 5–32.
- Breiman L, 2001b. Statistical modeling: the two cultures. *Stat. Sci.* 16 (3), 199–231.
- Breitenstein S, 2019. Choosing the crook: a conjoint experiment on voting for corrupt politicians. *Research & Politics* 6 (1), 2053168019832230.

- Bryk AS, Raudenbush SW, 1992. Hierarchical Linear Models: Applications and Data Analysis Methods. Sage Publications, Inc.
- Buja A, Brown L, Berk R, George E, Pitkin E, Traskin M, Zhang K, Zhao L, 2019a. Models as approximations I: consequences illustrated with linear regression. *Stat. Sci.* 34 (4), 523–544.
- Buja A, Brown L, Kuchibhotla AK, Berk R, George E, Zhao L, 2019b. Models as approximations II: a model-free theory of parametric regression. *Stat. Sci.* 34 (4), 545–565.
- Cantú F, 2019. The fingerprints of fraud: evidence from Mexico’s 1988 presidential election. *Am. Polit. Sci. Rev.* 113 (3), 710–726.
- Cerulli G, 2020. Machine Learning Using Stata. <https://sites.google.com/view/giovannicerulli/machine-learning-in-stata>.
- Chalfin A, Danieli O, Hillis A, Jelveh Z, Luca M, Ludwig J, Mullainathan S, 2016. Productivity and selection of human capital with machine learning. *Am. Econ. Rev.* 106 (5), 124–127.
- Chernozhukov V, Chetverikov D, Demirer M, Dufo E, Hansen C, Newey W, Robins J, 2018. Double/debiased machine learning for treatment and structural parameters. *Econom. J.* 21 (1), C1–C68.
- Chipman HA, George EI, McCulloch RE, 2010. BART: Bayesian additive regression trees. *Ann. Appl. Stat.* 4 (1), 266–298.
- Coyle J, Hejazi N, Malenica I, Phillips R, Sofrygin O, 2021. sl3: Pipelines for Machine Learning and Super Learning. R package version 1.4.4.
- Davis J, Heller SB, 2017. Using causal forests to predict treatment heterogeneity: an application to summer jobs. *Am. Econ. Rev.* 107 (5), 546–550.
- DiMaggio P, Nag M, Blei D, 2013. Exploiting affinities between topic modeling and the sociological perspective on culture: application to newspaper coverage of us government arts funding. *Poetics* 41 (6), 570–606.
- Donoho D, 2017. 50 years of data science. *J. Comput. Graph Stat.* 26 (4), 745–766.
- Dube A, Jacobs J, Naidu S, Suri S, 2020. Monopsony in online labor markets. *Am. Econ. Rev.: Insights* 2 (1), 33–46.
- D’Amour A, Ding P, Feller A, Lei L, Sekhon J, 2021. Overlap in observational studies with high-dimensional covariates. *J. Econom.* 221 (2), 644–654.
- Efron B, Hastie T, 2016. *Computer Age Statistical Inference*. Cambridge University Press.
- Efron B, Tibshirani RJ, 1994. *An Introduction to the Bootstrap*. CRC press.
- Elwert F, Winship C, 2010. Effect Heterogeneity and Bias in Main-Effects-Only Regression Models. *Heuristics, Probability and Causality: A Tribute to Judea Pearl*, pp. 327–336.
- Ferwerda J, Hainmueller J, Hazlett CJ, 2017. Kernel-based regularized least squares in R (KRLS) and Stata (krls). *J. Stat. Software* 79 (3), 1–26.
- Fix E, Hodges JL, 1989. Discriminatory analysis. nonparametric discrimination: consistency properties [1951] *International Statistical Review/Revue Internationale de Statistique* 57 (3), 238–247.
- Fong C, Ratkovic M, Imai K, Hazlett C, Yang X, Peng S, Lee I, 2021. CBPS: Covariate Balancing Propensity Score. R package version 0.23.
- Freese J, Peterson D, 2017. Replication in social science. *Annu. Rev. Sociol.* 43, 147–165.
- Friedberg R, Tibshirani J, Athey S, Wager S, 2021. Local linear forests. *J. Comput. Graph Stat.* 30 (2), 503–517.
- Friedman J, Hastie T, Tibshirani R, 2010. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Software* 33 (1), 1.
- Friedman S, Reeves A, 2020. From aristocratic to ordinary: shifting modes of elite distinction. *Am. Socio. Rev.* 85 (2), 323–350.
- Frye M, Trinitapoli J, 2015. Ideals as anchors for relationship experiences. *Am. Socio. Rev.* 80 (3), 496–525.
- Gelman A, Little TC, 1997. Poststratification into many categories using hierarchical logistic regression. *Surv. Methodol.* 23 (2), 127–135.
- Gelman A, Loken E, 2014. The statistical crisis in science. *Am. Sci.* 102 (6), 460.
- Gentzkow M, Shapiro JM, Taddy M, 2019. Measuring group differences in high-dimensional choices: method and application to congressional speech. *Econometrica* 87 (4), 1307–1340.

- Grimmer J, Roberts ME, Stewart BM, 2021. Machine learning for social science: an agnostic approach. *Annu. Rev. Polit. Sci.* 24, 395–419.
- Grimmer J, Roberts ME, Stewart BM, 2022. *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton University Press.
- Grimmer J, Stewart BM, 2013. Text as data: the promise and pitfalls of automatic content analysis methods for political texts. *Polit. Anal.* 21 (3), 267–297.
- Hainmueller J, 2012. Entropy balancing for causal effects: a multivariate reweighting method to produce balanced samples in observational studies. *Polit. Anal.* 20 (1), 25–46.
- Hainmueller J, 2014. Ebal: Entropy Reweighting to Create Balanced Samples. R package version 0, pp. 1–6.
- Hainmueller J, Hopkins DJ, Yamamoto T, 2014. Causal inference in conjoint analysis: understanding multidimensional choices via stated preference experiments. *Polit. Anal.* 22 (1), 1–30.
- Hájek J, 1968. Asymptotic Normality of Simple Linear Rank Statistics under Alternatives. *The Annals of Mathematical Statistics*, pp. 325–346.
- Handel B, Kolstad J, 2017. Wearable technologies and health behaviors: new data and new methods to understand population health. *Am. Econ. Rev.* 107 (5), 481–485. [PubMed: 29553625]
- Hartman E, Grieve R, Ramsahai R, Sekhon JS, 2015. From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate population treatment effects. *J. Roy. Stat. Soc.* 178 (3), 757–778.
- Hastie T, Tibshirani R, Friedman JH, Friedman JH, 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, ume 2. Springer.
- Hauser RM, Koffel JN, Travis HP, Dickinson PJ, 1975. Temporal change in occupational mobility: evidence for men in the United States. *Am. Socio. Rev.* 279–297.
- Hauser RM, Tsai S-L, Sewell WH, 1983. A Model of Stratification with Response Error in Social and Psychological Variables. *Sociology of Education*, pp. 20–46.
- Healy K, 2018. *Data Visualization: A Practical Introduction*. Princeton University Press.
- Hernán MA, Robins JM, 2021. *Causal Inference: what if*. Chapman & Hall/CRC, Boca Raton.
- Hoeffding W, 1948. A class of statistics with asymptotically normal distribution. *Ann. Math. Stat.* 19 (3), 293–325.
- Holland PW, 1986. Statistics and causal inference. *J. Am. Stat. Assoc.* 81 (396), 945–960.
- Hopkins DJ, King G, 2010. A method of automated nonparametric content analysis for social science. *Am. J. Polit. Sci.* 54 (1), 229–247.
- Imai K, Ratkovic M, 2014. Covariate balancing propensity score. *J. Roy. Stat. Soc. B* 76 (1), 243–263.
- Imbens GW, 2015. Matching methods in practice: three examples. *J. Hum. Resour.* 50 (2), 373–419.
- Imbens GW, Rubin DB, 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.
- Incerti T, 2020. Corruption information and vote share: a meta-analysis and lessons for experimental design. *Am. Polit. Sci. Rev.* 114 (3), 761–774.
- Jerzak CT, King G, Strezhnev A, 2022. An Improved Method of Automated Nonparametric Content Analysis for Social Science. *Political Analysis*, pp. 1–17.
- King G, Pan J, Roberts ME, 2017. How the Chinese government fabricates social media posts for strategic distraction, not engaged argument. *Am. Polit. Sci. Rev.* 111 (3), 484–501.
- Kleinberg J, Ludwig J, Mullainathan S, Obermeyer Z, 2015. Prediction policy problems. *Am. Econ. Rev.* 105 (5), 491–495. [PubMed: 27199498]
- Knox D, Lucas C, 2021. A dynamic model of speech for the social sciences. *Am. Polit. Sci. Rev.* 115 (2), 649–666.
- Lin W, 2013. Agnostic notes on regression adjustments to experimental data: reexamining freedman’s critique. *Ann. Appl. Stat.* 7 (1), 295–318.
- Lin Y, Jeon Y, 2006. Random forests and adaptive nearest neighbors. *J. Am. Stat. Assoc.* 101 (474), 578–590.
- Lundberg I, Johnson R, Stewart BM, 2021. What is your estimand? Defining the target quantity connects statistical evidence to theory. *Am. Socio. Rev.* 86 (3), 532–565.

- Mahalanobis PC, 1936. On the Generalized Distance in Statistics. National Institute of Science of India.
- Mohr JW, Bogdanov P, 2013. Introduction—topic models: what they are and why they matter. *Poetics* 41 (6), 545–569.
- Molina M, Garip F, 2019. Machine learning for sociology. *Annu. Rev. Sociol.* 45, 27–45.
- Moore R, Brand JE, Shinkre T, 2021. ITPSCORE: Stata Module to Implement Iterative Propensity Score Logistic Regression Model Search Procedure. Statistical Software Components S459018, Boston College Department of Economics.
- Morgan SL, Harding DJ, 2006. Matching estimators of causal effects: prospects and pitfalls in theory and practice. *Socio. Methods Res.* 35 (1), 3–60.
- Mullainathan S, Spiess J, 2017. Machine learning: an applied econometric approach. *J. Econ. Perspect.* 31 (2), 87–106.
- Murphy KP, 2012. *Machine Learning: A Probabilistic Perspective*. MIT press.
- Pearl J, 2009. *Causality*. Cambridge University Press.
- Pearl J, Bareinboim E, 2011. Transportability of causal and statistical relations: a formal approach. In: *Twenty-Fifth AAAI Conference on Artificial Intelligence*.
- Pearl J, Mackenzie D, 2018. *The Book of Why: the New Science of Cause and Effect*. Basic books.
- Raftery AE, Madigan D, Hoeting JA, 1997. Bayesian model averaging for linear regression models. *J. Am. Stat. Assoc.* 92 (437), 179–191.
- Rao JN, 2003. *Small Area Estimation*. John Wiley & Sons.
- Roberts ME, Stewart BM, Tingley D, Lucas C, Leder-Luis J, Gadarian SK, Albertson B, Rand DG, 2014. Structural topic models for open-ended survey responses. *Am. J. Polit. Sci.* 58 (4), 1064–1082.
- Robins JM, Rotnitzky A, 1995. Semiparametric efficiency in multivariate regression models with missing data. *J. Am. Stat. Assoc.* 90 (429), 122–129.
- Robins JM, Rotnitzky A, Zhao LP, 1994. Estimation of regression coefficients when some regressors are not always observed. *J. Am. Stat. Assoc.* 89 (427), 846–866.
- Schapire RE, Freund Y, 2012. *Boosting: Foundations and Algorithms*. MIT Press.
- Schwarz G, 1978. Estimating the Dimension of a Model. *The Annals of Statistics*, pp. 461–464.
- Simmons JP, Nelson LD, Simonsohn U, 2011. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* 22 (11), 1359–1366. [PubMed: 22006061]
- StataCorp, 2021. *Stata: Release 17*. StataCorp LLC, College Station, TX.
- Stone M, 1974. Cross-validatory choice and assessment of statistical predictions. *J. Roy. Stat. Soc. B* 36 (2), 111–133.
- Stuart EA, Bradshaw CP, Leaf PJ, 2015. Assessing the generalizability of randomized trial results to target populations. *Prev. Sci.* 16 (3), 475–485. [PubMed: 25307417]
- Su Z, Meng T, 2016. Selective responsiveness: online public demands and government responsiveness in authoritarian China. *Soc. Sci. Res.* 59, 52–67. [PubMed: 27480371]
- Szeliski R, 2010. *Computer Vision: Algorithms and Applications*. Springer Science & Business Media.
- Textor J, Hardt J, Knüppel S, 2011. DAGitty: a graphical tool for analyzing causal diagrams. *Epidemiology* 22 (5), 745.
- Therneau T, Atkinson B, Ripley B, Ripley MB, 2015. Package ‘rpart’. Available online: <https://cran.r-project.org/web/packages/rpart/index.html>.
- Tibshirani J, Athey S, Friedberg R, Hadad V, Hirshberg D, Miner L, Sverdrup E, Wager S, Wright M, Tibshirani MJ, 2018. Package ‘grf’.
- Townsend W, 2017. *Elasticregress*. <https://github.com/wilburtownsend/elasticregress>.
- Van der Laan MJ, Polley EC, Hubbard AE, 2007. Super learner. *Stat. Appl. Genet. Mol. Biol.* 6 (1).
- Van der Laan MJ, Rose S, 2018. *Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies*. Springer.
- Van Der Laan MJ, Rubin D, 2006. Targeted maximum likelihood learning. *Int. J. Biostat.* 2 (1).

- Wager S, Athey S, 2018. Estimation and inference of heterogeneous treatment effects using random forests. *J. Am. Stat. Assoc.* 113 (523), 1228–1242.
- Wickham H, 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer.
- Wolpert DH, 1992. Stacked generalization. *Neural Network.* 5 (2), 241–259.
- Wood SN, 2017. *Generalized Additive Models: an Introduction with R*. CRC press.
- Wright MN, Ziegler A, 2017. ranger: a fast implementation of random forests for high dimensional data in C++ and R. *J. Stat. Software* 77 (1), 1–17.
- Wu X, Kumar V, Ross Quinlan J, Ghosh J, Yang Q, Motoda H, McLachlan GJ, Ng A, Liu B, Yu PS, et al. , 2008. Top 10 algorithms in data mining. *Knowl. Inf. Syst.* 14 (1), 1–37.
- Wurm MJ, Rathouz PJ, Hanlon BM, 2017. Regularized Ordinal Regression and the Ordinalnet R Package. arXiv preprint arXiv:1706.05003.
- Xie, Y., 2013. Population heterogeneity and causal inference. *Proc. Natl. Acad. Sci. USA* 110 (16), 6262–6268.
- Ying L, Montgomery JM, Stewart BM, 2021. Topics, concepts, and measurement: a crowdsourced procedure for validating topics as measures. *Polit. Anal.* 1–20.
- Zhang H, Pan J, 2019. CASM: a deep-learning approach for identifying collective action events with text and image data from social media. *Socio. Methodol.* 49 (1), 1–57.



**Fig. 1. Supervised machine learning for measurement amplifies researcher coding.**

One setting which is particularly promising for machine learning exists when social scientists have many observations, each of which contains some high-dimensional predictor set  $\vec{x}_i$  (e.g., the text of document  $i$ ) but the researcher is interested in some low-dimensional, unobserved categorization  $Y_j$  (e.g., the topic of document  $i$ , here represented by colors). A researcher who manually codes a random sample of the observations into categories can use machine learning tools to amplify that coding by predicting for the full population. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Begin with  
high-dimensional  
data

$\vec{x}_1$
$\vec{x}_2$
$\vec{x}_3$
$\vec{x}_4$
$\vec{x}_5$
$\vec{x}_6$
$\vec{x}_7$
$\vec{x}_8$
$\vec{x}_9$
$\vec{x}_{10}$

Each row has  
**many** columns

Learn a  
low-dimensional  
representation

$\hat{z}_1$
$\hat{z}_2$
$\hat{z}_3$
$\hat{z}_4$
$\hat{z}_5$
$\hat{z}_6$
$\hat{z}_7$
$\hat{z}_8$
$\hat{z}_9$
$\hat{z}_{10}$

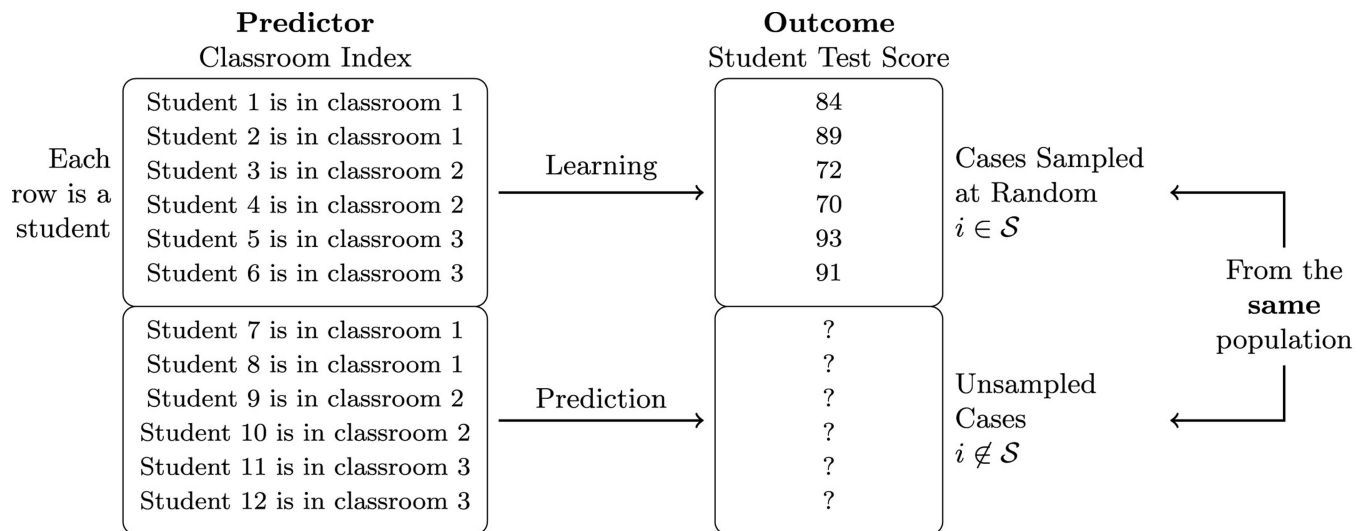
Each row has  
**few** columns

Inductive  
→

**Fig. 2. Unsupervised machine learning inductively summarizes complex data.**

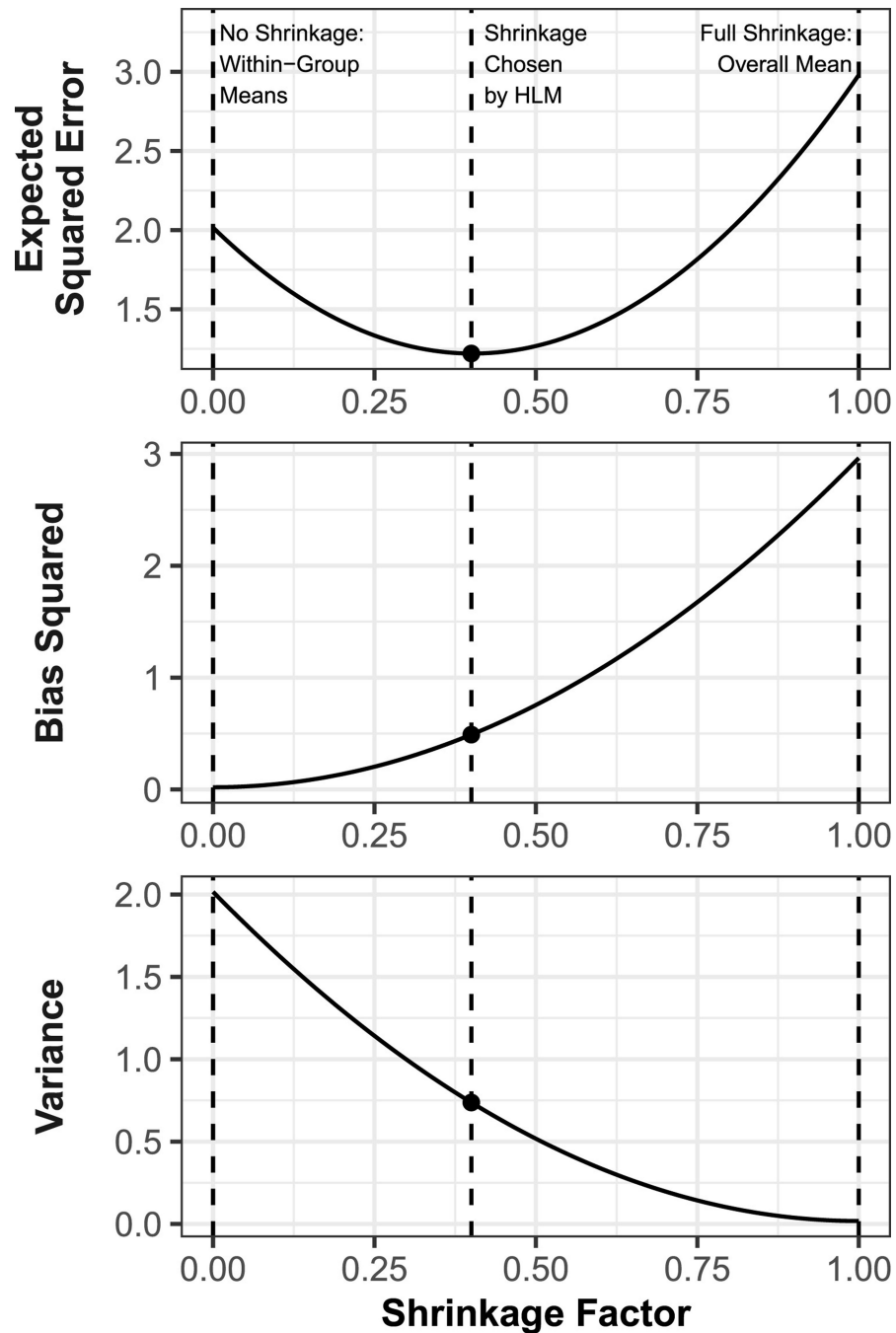
In some settings, each observation contains a high-dimensional feature set  $\vec{x}_i$  (e.g., unstructured text data) with no labeled outcomes. Unsupervised machine learning reduces those features to a low-dimensional representation  $\vec{z}_i$ . Clustering is one example. The representation could be a scalar (e.g., assignment to a cluster) or a vector (e.g., a set of probabilities over many possible clusters). With unsupervised methods, the resulting representation depends entirely on the objective function that the algorithm seeks to optimize. Unsupervised learning can discover representations of data automatically, but the researcher must then validate those representations and argue for their substantive usefulness.





**Fig. 3. Task clarity: Out-of-sample prediction.**

A well-studied machine learning task involves a random sample  $\mathcal{S}$  taken from a target population, where the goal is to learn a prediction function to predict the outcomes of new samples from that same target population. For instance, we might use classroom indices to predict the test scores of individual students who were not observed in the training sample.

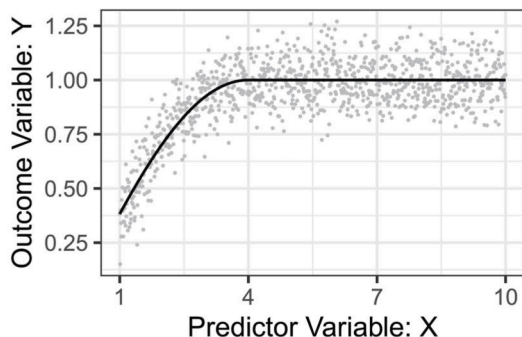


**Fig. 4. Simulation: Balancing the bias-variance trade-off.**

In this simulation, there are 100 classes with class-level mean test scores normally distributed with variance 3. Within classes, student scores are normally distributed with variance 10. In each of 100 simulated samples, we estimate from a sample of 5 students from each class. The estimator partially pools the class-specific mean with the overall mean according to a shrinkage factor:  $\hat{\theta}_j^{(\text{shrinkagefactor})} = \bar{y}_j - (\text{shrinkagefactor})(\bar{y}_j - \bar{y})$ . A shrinkage factor of 0 involves no pooling so that the estimate is the sample mean within each class, and a shrinkage factor of 1 involves complete shrinkage so that the estimate for every class equals

the overall sample mean. A hierarchical linear model selects a shrinkage factor equal to the variance of the within-class means divided by that variance plus the variance of the means across classes. The center dashed line takes those variances as known and shows that the multilevel shrinkage minimizes the expected squared error. To create each curve, we first calculate the statistic over simulations within classes, and then we report the average of the statistic over all classes.

A) Data generating process in this simulation. The conditional mean function  $\mu(x)$  (black curve) is intentionally chosen to not correspond to any of the functional forms assumed by the estimators.



Conditional mean function:

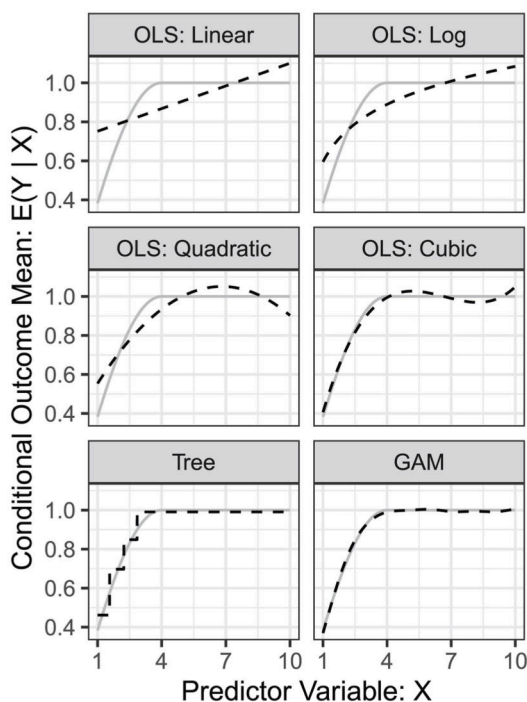
$$\mu(x) \equiv \begin{cases} \sin\left(\frac{\pi}{8}x\right) & \text{if } x \leq 4 \\ 1 & \text{if } x > 4 \end{cases}$$

For  $i = 1, \dots, 1000$ :

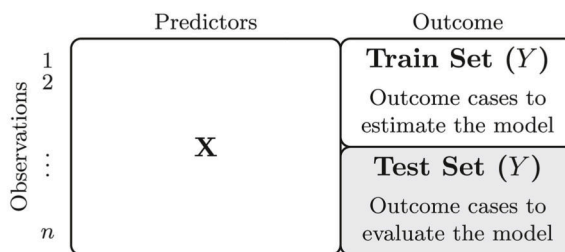
$$X_i \sim \text{Uniform}(1, 10)$$

$$Y_i \sim \text{Normal}(\mu(X_i), 0.1)$$

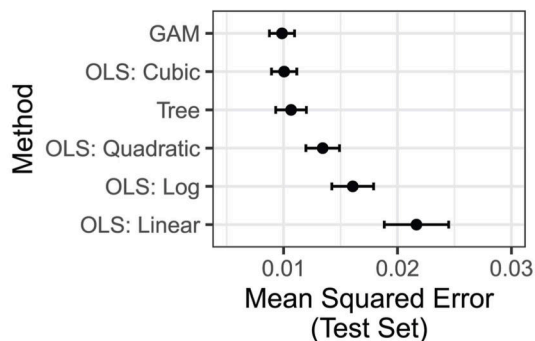
B) Performance of six estimators (dashed black) for the simulated conditional mean function (solid gray).



C) We estimated the dashed functions in a train set and then evaluated them in a test set.



D) This yields a data-driven procedure to select an estimator: the one with the best performance in the test set.



**Fig. 5. Simulation: Data-driven estimator selection.**

We consider six estimators: OLS with linear, log, quadratic, and cubic specifications, a regression tree following defaults in the `rpart` package (Therneau et al., 2015), and a Generalized Additive Model (GAM, Wood, 2017) following defaults in the `mgcv` package. Visually, the GAM comes closest to the true response function (Panel B). Panel C depicts how we randomly assigned observations to two equally-sized subsamples: the train set and test set. We then estimated each function on the train set and estimated its mean squared error when predicting the new cases in the test set. Panel D shows that the GAM achieves

the best performance. This exercise illustrates a building block of machine learning: instead of arguing conceptually for a particular estimator (e.g. OLS with a particular form), empirically evaluate the performance of many candidate estimators.

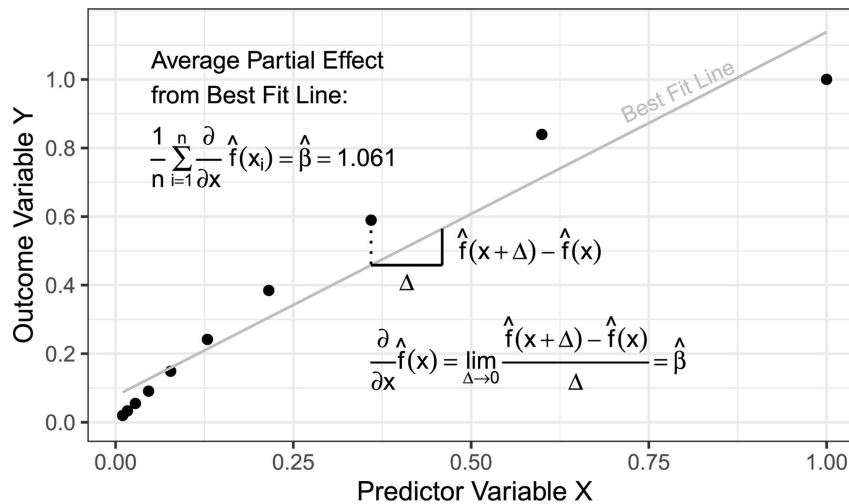
Author Manuscript

Author Manuscript

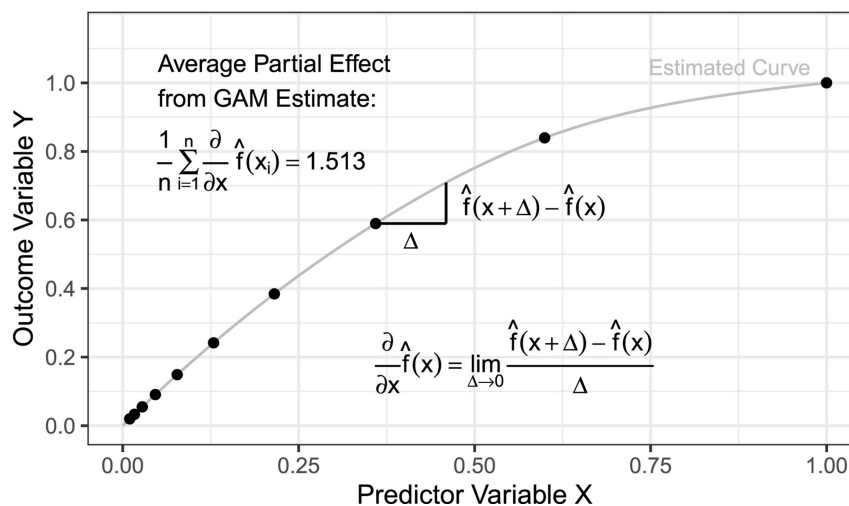
Author Manuscript

Author Manuscript

- A) The coefficient of a best-fit line is an estimate of the average partial effect: the average rate of change in  $Y$  as a function of  $X$ .



- B) A flexible curve can also be summarized by an average partial effect estimate.
- 1) Predict at the observed  $X$ .
  - 2) Predict at  $X + \Delta$  for a small nudge  $\Delta$ .
  - 3) Difference and take the mean, divided by  $\Delta$ .



**Fig. 6. Partial effects: The interpretability of regression is not lost.**

In this simulation, the true curve is the quadratic function  $Y = 1 - (1 - X)^2$ . The predictor values  $x_1, \dots, x_{10}$  are spaced on an exponential scale. The true average partial effect is greater than the OLS coefficient estimate because the data are denser at the far left of the plot, where the slope is steeper, but the points at the far right of the plot have high leverage over the OLS estimate. The estimated curve is a thin-plate spline estimated by the `gam` function in the `mgcv` package in R (Wood, 2017). The simulation illustrates that a key benefit of OLS—a single-number summary of a relationship—is also available for flexible machine

learning methods. Further, machine learning methods can yield superior estimates of the average partial effect.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

An OLS coefficient is misleading: Illustrated by a hypothetical example.  
The effect of precipitation on outdoor exercise in two cities.

### Vail, Colorado

Sample of 5,000 residents  
on random days in January  
48% chance of precipitation  
Probability of outdoor exercise:  
— 90% if it precipitates (snow)  
— 50% if it does not


### Phoenix, Arizona

Sample of 5,000 residents  
on random days in January  
10% chance of precipitation  
Probability of outdoor exercise:  
— 10% if it precipitates (rain)  
— 50% if it does not

Assumption: Given location, the event “precipitation occurs” is  
unconfounded with respect to potential exercise

Conditional Average Causal Effect  
Precipitation **increases** outdoor  
exercise by 40 percentage points

Conditional Average Causal Effect  
Precipitation **decreases** outdoor  
exercise by 40 percentage points

  
 Average Causal Effect  
 On average, precipitation causes  
 a **0 percentage point**  
 change in outdoor exercise.

What would we get with an OLS model?

$$P(\text{Outdoor Exercise} \mid \text{Precipitation, City}) = \alpha + \beta(\text{Precipitates}) + \gamma(\text{Vail})$$

This misspecified model assumes the same precipitation effect  $\beta$  in Vail and Phoenix.  
But there is more information in Vail, where the chance of precipitation is closer to 50%.  
Because OLS has assumed the effect is the same value  $\beta$  in both places, it gives greater  
weight to Vail (where the effect is more precise).

OLS puts 76% of the weight on Vail and 24% of the weight on Phoenix.

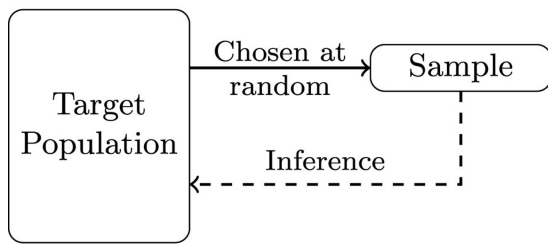
$$\text{OLS estimate: } \hat{\beta} = 0.18$$

When effects are heterogeneous, OLS does not estimate the average causal effect.

### **Fig. 7. A regression coefficient is hard to interpret.**

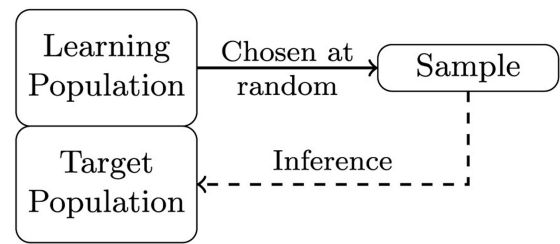
An apparent barrier to the adoption of machine learning is the perceived interpretability of standard regression models. Regression coefficients may seem familiar, but they are not as easy to interpret as many scholars may believe. As one example, the coefficient of a misspecified OLS model cannot be interpreted as the average causal effect, even when that model includes all confounding variables. For a longer discussion, see Elwert and Winship (2010), Brand and Thomas (2013), and Aronow and Samii (2016). In the example above, the probability of January precipitation in each city is true; all else is simulated.



**Task: Prediction in the same population**

Example:

- 1) Sample students entering Statsville West High School in **2017**
- 2) Observe if they drop out
- 3) Learn a prediction function
- 4) Predict for all students entering Statsville West in **2017**

**Task: Prediction in a new population**

Example:

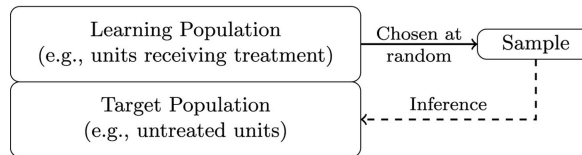
- 1) Sample students entering Statsville West High School in **2017**
- 2) Observe if they drop out
- 3) Learn a prediction function
- 4) Predict for all students entering Statsville West in **2022**

**Fig. 8. Caution: Prediction in a new target population.**

A standard machine learning task is to learn about a target population using a sample of cases selected at random from that population. In practice, however, algorithms are often deployed to make predictions in new populations from which no training cases were available. For example, a function to predict high school dropout learned in a cohort entering high school in 2017 might be used to target resources to at-risk students entering high school in 2022. But if the mapping between the predictors and outcome changes across cohorts, that prediction function may no longer be useful. To the extent that prediction functions are learned in one population and applied in a new target population, the validity of predictions may be uncertain.

Prediction in a **counterfactual** population.

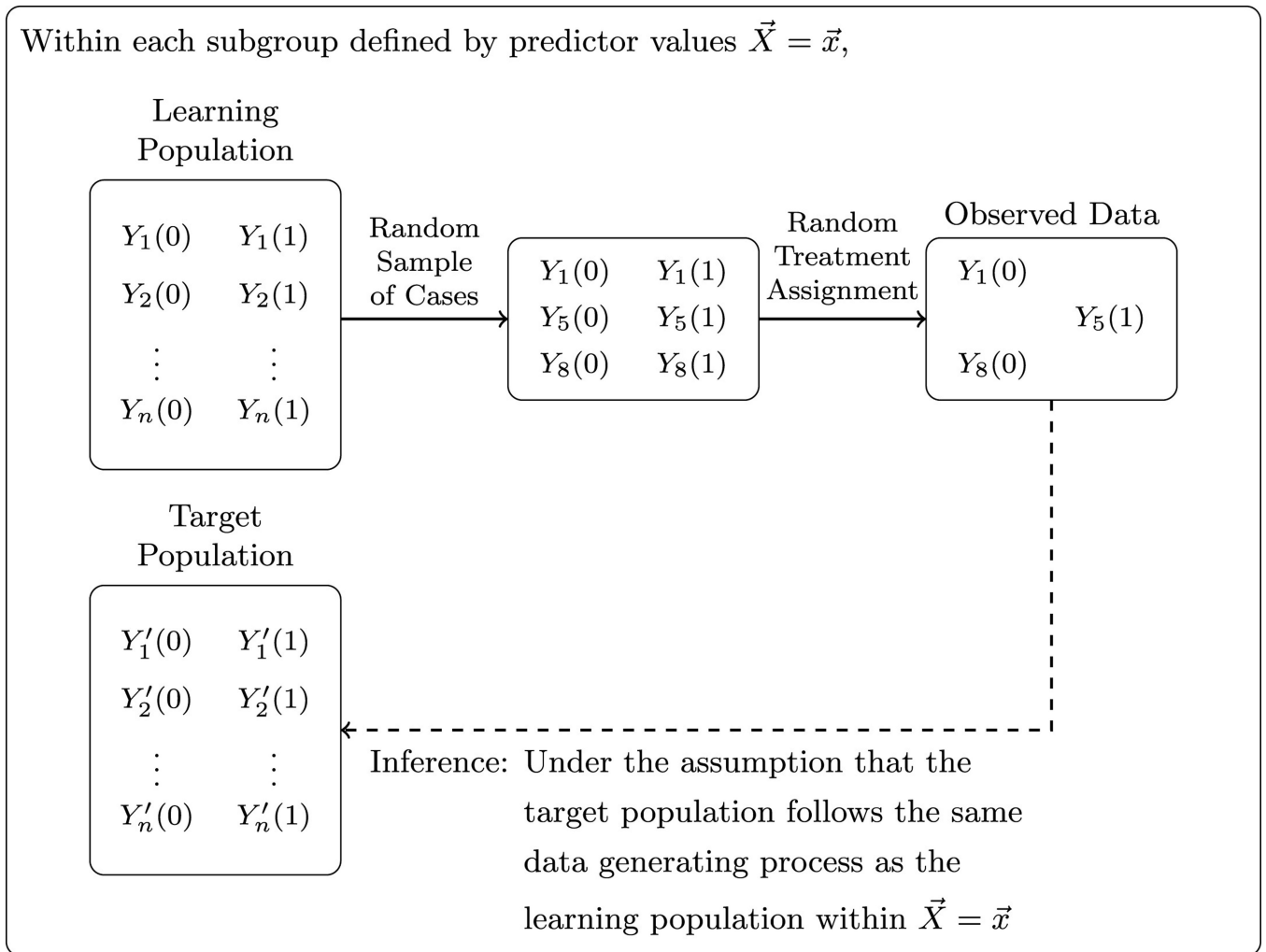
**Task:** Predict the outcome that would be realized under treatment.



**Required Assumption:** Treatment (which determines membership in the learning versus target population) is independent of the outcome  $Y_i(1)$  that would be realized under treatment, within sub-populations defined by the predictor variables.

**Fig. 9. Causal inference: A task that involves a new target population.**

Suppose we observe a set of units who receive a treatment of interest (e. g., extra counseling in high school). After learning a prediction function in a sample of treated units, we wish to predict the outcome that untreated units would have realized if they had received treatment. For instance, we might predict whether those who did not receive counseling would not have dropped out if they had received counseling. Causal questions of this form require assumptions because in the absence of a randomized treatment it is impossible to draw a simple random sample from the target population.



**Fig. 10. Causal inference for policy prescriptions: A particular leap to a new target.**

Suppose there is a learning population of  $n$  units, each of whom has a potential outcome that would be realized under the control condition  $Y_i(0)$  and under the treatment condition  $Y_i(1)$ . Suppose we take a random sample from the learning population and then randomly assign treatments to units in that sample. For each unit in the sample, we observe one of the two potential outcomes. Under randomization, a prediction function learned in the observed data can be used to make predictions in the learning population. But when designing policy, we generally want to predict treatment effectiveness in a new population who have not yet received the treatment. To predict in a new target population, we would have to additionally assume that the same data generating process holds in the target population as in the learning population. In observational settings, machine learning can be used for causal inference if the assumptions of random sampling and random treatment assignment are credible within subgroups defined by the observed predictors  $\vec{X}$ .

- 1) Estimate initial prediction functions

$$\hat{g}(t, \vec{x}) \approx P(Y = 1 | T = t, \vec{X} = \vec{x})$$

$$\hat{m}(t, \vec{x}) \approx P(T = t | \vec{X})$$

- 2) Define a new covariate

$$\hat{H}(T, \vec{X}) = \frac{\mathbb{I}(T=t)}{\hat{m}(t, \vec{X})}$$

**Sample split:** Optionally, carry out step 3 in a different sample from steps 1 and 2

- 3) Regress  $Y$  on the new covariate with an offset

$$\text{logit} \left( P(Y = 1 | T, \vec{X}) \right) \approx \underset{\substack{\uparrow \\ \text{Offset term} \\ \text{(from 1)}}}{\text{logit} \left( \hat{g}(T, \vec{X}) \right)} + \underset{\substack{\nearrow \\ \text{Clever covariate} \\ \text{(from 2)}}}{\hat{H}(T, \vec{X})} \underset{\substack{\uparrow \\ \text{Coefficient to estimate} \\ \text{here}}}{\beta}$$

- 4) Target the prediction function

$$\hat{g}'(T, \vec{X}) = \underset{\substack{\uparrow \\ \text{Targeted prediction rule} \\ \text{optimized for the way we} \\ \text{will aggregate predictions}}}{\text{logit}^{-1}} \left( \underset{\substack{\uparrow \\ \text{Original prediction rule} \\ \text{optimized for} \\ \text{disaggregate prediction}}}{\text{logit}(\hat{g}(T, \vec{X}))} + \hat{H}(T, \vec{X}) \hat{\beta} \right)$$

- 5) Estimate using the targeted prediction function

$$\hat{\mathbf{E}}(Y(t)) = \frac{1}{n} \sum_{i=1}^n \hat{g}'(t, \vec{X}_i)$$

**Fig. 11. Targeted learning with a binary outcome.**

An important advantage of targeted learning (Van der Laan and Rose, 2018) over double machine learning (Chernozhukov et al., 2018) is that targeted learning can accommodate a link function (the logit in steps 3 and 4) which can guarantee that predicted values fall within the support of the outcome. For targeted learning with a continuous outcome, see Supplemental Fig. 12. For double machine learning, see Supplemental Fig. 13.

**Table 1**  
**Where to go next: Introductory material for interested readers.**

These are a few of the growing number of pedagogical introductions to machine learning, many of which are accompanied by software packages in R or Stata.

Category	Reference	R package	Stata package
General references			
Statistical foundations	Hastie et al. (2009)		
	Efron and Hastie (2016)		
	Bishop (2006)		
Bayesian perspective	Murphy (2012)		
Measurement			
With text	Grimmer et al. (2022)		
	Hopkins and King (2010)	<code>readme</code>	
	Jerzak et al. (2022)	<code>readme2</code>	
With audio	Knox and Lucas (2021)	<code>communication</code>	
With images	Szeliski (2010)		
Dimension reduction			
For text analysis	Roberts et al. (2014)	<code>stm</code>	
	Blei et al. (2003)	<code>topicmodels</code>	
Estimation			
For smooth functions	Wood (2017)	<code>mgcv</code>	
For interactive functions	Breiman (2001a)		
	Wright and Ziegler (2017)		
	Wright and Ziegler (2017)	<code>ranger</code>	
	Cerulli (2020)	<code>r_ml_stata_cv</code>	
For penalized generalized linear models	Friedman et al. (2010)	<code>glmnet</code>	
	Wurm et al. (2017)	<code>ordinalNet</code>	
	StataCorp (2021)		<code>lasso</code>
	Townsend (2017)		<code>elasticregress</code>
Related topics			
Causal inference	Van der Laan and Rose (2018)	<code>tlverse</code>	
	Imbens and Rubin (2015)		
	Hernán and Robins (2021)		
	Pearl (2009)		
	Pearl and Mackenzie (2018)		
	Textor et al. (2011)	<code>dagitty</code>	
	Athey and Imbens (2016)	<code>causalTree</code>	
	Athey et al. (2019)	<code>grf</code>	
	Brand et al. (2021)	<code>htetree</code>	
	Ahrens et al. (2018)		<code>pdslasso</code>

Category	Reference	R package	Stata package
	Ferwerda et al. (2017)	krls	krls
Quantification of uncertainty	Efron and Tibshirani (1994)	bootstrap	
Visualization	Wickham (2016)Healy (2018)	ggplot2	
Social science reviews	Grimmer et al. (2021)		
	Brand et al. (2020)		
	Athey and Imbens (2019)		
	Molina and Garip (2019)		
	Mullainathan and Spiess (2017)		

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript