ANNUAL REVIEWS

Annual Review of Political Science

Machine Learning for Social Science: An Agnostic Approach

Justin Grimmer,[1] Margaret E. Roberts,[2] and Brandon M. Stewart[3]

[1]Department of Political Science and Hoover Institution, Stanford University, Stanford, California 94305, USA; email: jgrimmer@stanford.edu

[2]Department of Political Science and Halıcıoğlu Data Science Institute, University of California San Diego, La Jolla, California 92093, USA; email: meroberts@ucsd.edu

[3]Department of Sociology and Office of Population Research, Princeton University, Princeton, New Jersey 08540, USA; email: bms4@princeton.edu

Annu. Rev. Political Sci. 2021. 24:395–419

First published as a Review in Advance on March 5, 2021

The *Annual Review of Political Science* is online at polisci.annualreviews.org

https://doi.org/10.1146/annurev-polisci-053119-015921

Keywords

machine learning, text as data, research design

Abstract

Social scientists are now in an era of data abundance, and machine learning tools are increasingly used to extract meaning from data sets both massive and small. We explain how the inclusion of machine learning in the social sciences requires us to rethink not only applications of machine learning methods but also best practices in the social sciences. In contrast to the traditional tasks for machine learning in computer science and statistics, when machine learning is applied to social scientific data, it is used to discover new concepts, measure the prevalence of those concepts, assess causal effects, and make predictions. The abundance of data and resources facilitates the move away from a deductive social science to a more sequential, interactive, and ultimately inductive approach to inference. We explain how an agnostic approach to machine learning methods focused on the social science tasks facilitates progress across a wide range of questions.

1. INTRODUCTION

Benchmark: established quantitative measure to assess performance of some procedure

Agnostic: a position of skepticism that a given machine learning method can capture the underlying data-generating process

For much of its history, empirical work in the social sciences has been defined by scarcity. Data were hard to find, surveys were costly to field, and record storage was close to impossible. Computation was an even more pressing bottleneck with limited and expensive computing time. The consequence of this scarcity was that social scientists developed and relied on statistical techniques that enabled progress with few data and even less computing power.

Abundance now defines the social sciences. The rapid expansion of available data has shifted the evidence base. Election scholars used to rely on occasional surveys administered around national elections; now, researchers use voter files with millions of records. International relations scholars can bolster careful reading of archives with the analysis of millions of declassified state department cables. The difference is not just a matter of scale. New forms of data can fundamentally change our ability to measure phenomena; for example, tracking the removal of social media posts in real time provides a new window into how authoritarian regimes control information available to the public. Computing power has also exploded, with personal computers able to analyze millions of rows of data and more powerful cloud computing services readily available.

Social scientists increasingly rely on machine learning methods to make the most of this new abundance. Machine learning is a class of flexible algorithmic and statistical techniques for prediction and dimension reduction. The machine learning community largely prioritizes performance on established quantitative benchmarks. This includes not only explicitly predictive tasks, such as classifying emails as spam or predicting who will click on an advertisement, but also other tasks amenable to quantitative feedback, such as compressing information in an image or maneuvering a robot in an environment. By focusing on optimizing performance on such tasks, the community has made astonishingly rapid progress. The results include not only more accurate spam filters but also algorithms that can generate realistic fake images, write near-human-quality prose, and defeat world champion human players in games of strategy.

Just as they have transformed so many other areas of life, machine learning methods have transformative potential in social science. Unlocking this potential involves reconsidering the conventions of machine learning and reapplying these techniques to accomplish social science tasks such as discovery, measurement, and causal inference. Likewise, the introduction of machine learning methods also invites us to reevaluate the typical model of social science. In this article, we argue that the current abundance of data allows us to break free from the deductive mindset that was previously necessitated by data scarcity. Instead, we adopt a more inductive approach, which involves sequential and iterative inferences—a reality that characterizes much of social science work, but that is difficult to talk about because it conflicts with the dominant deductive framing of research.

This article provides an overview of how social scientists have used machine learning methods, how they have evaluated the performance of models, and what is distinctive about a social science approach to machine learning. We describe our approach to machine learning as agnostic because we avoid assuming that the data emerge from a process that matches our machine learning method.[1] Our position arises from an underlying skepticism in which we doubt our models but trust our validations. We summarize our view of how to apply machine learning techniques in the social sciences in **Figure 1**. There, we describe the move from task to approach and then to evaluation according to the goals of the task. We argue that, when applied to the social sciences, the core tasks are discovery, measurement, causal inference, and prediction. Most of our perspective on the core tasks extends beyond machine learning and draws on the long history of techniques

[1] See also the agnostic regression framework of Lin (2013) and Aronow & Miller (2019), which focuses on the properties of ordinary least squares regression without assuming the classical linear model holds.

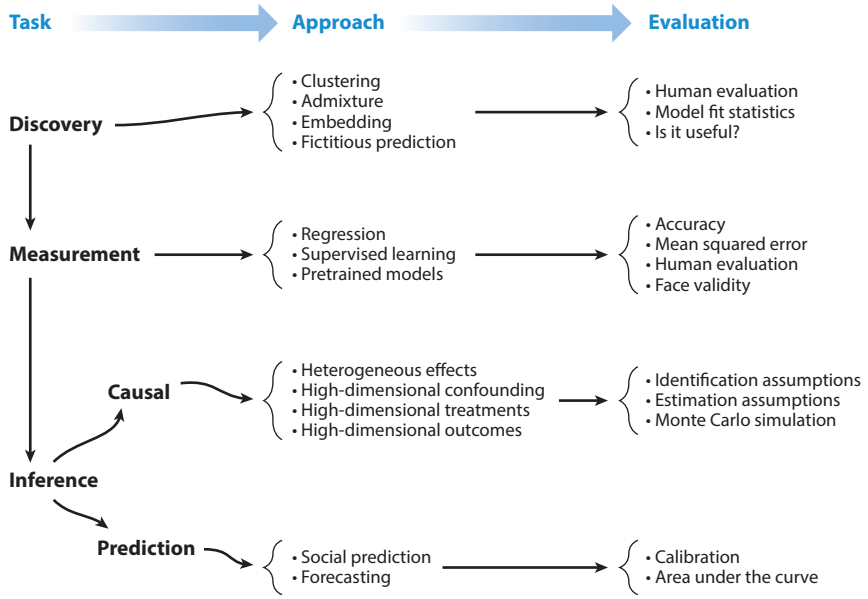

Figure 1

Our approach to machine learning in the social sciences. We reframe the tasks to ones relevant to social science: discovery, measurement, causal inference, and prediction.

in the social sciences that predate modern machine learning. The availability of machine learning tools only heightens the importance of reexamining these foundational issues in research design.

In the next section, we provide an overview of the core tools that define the machine learning approach to the world. After that, we argue that integrating machine learning methods into social science tasks invites us to reconsider the traditional deductive model of social science. We then address our major tasks—discovery, measurement, causal inference, and prediction. We conclude with a brief look at the frontiers of machine learning for social science.

Because we focus primarily on a framework for applying machine learning to social science tasks, we do not discuss the machine learning techniques that are already in use in political science, such as tree-based models (Stewart & Zhukov 2009, Hill & Jones 2014, Montgomery & Olivella 2018, Kaufman et al. 2019, Acharya et al. 2021), naïve Bayes (Nielsen 2017, Rashkin et al. 2017), neural networks (Beck et al. 2000, Williams et al. 2020), and support vector machines (Hillard et al. 2008). Instead, we point the reader to accessible textbooks (Bishop 2006, Murphy 2012, Goodfellow et al. 2016) and the excellent reviews of machine learning for economists (Mullainathan & Spiess 2017, Athey & Imbens 2019) and sociologists (Molina & Garip 2019).

2. THE CULTURE OF MACHINE LEARNING

Ask five researchers what machine learning is and you will likely get five different answers. Most agree that certain methods—for example, deep neural networks—are machine learning, but other techniques—for example, linear regression and the least absolute shrinkage and selection operator (LASSO)—originate in statistics even if they are taught in nearly all machine learning courses. We argue that machine learning is as much a culture defined by a distinct set of values and tools as it is a set of algorithms. Breiman (2001) made a similar point 20 years ago in his seminal piece "Statistical

Modeling: The Two Cultures," which drew a contrast between stochastic data-generating process modeling and algorithmic modeling cultures.

Prediction is explicitly the goal in many areas of machine learning—not within the data set, as might be approximated by something like R^2, but prediction in new, unseen data. This is in contrast to most areas of social science. Mullainathan & Spiess (2017) characterize this as a difference between a focus on $\hat{y}$ (the prediction of the outcome and the focus of machine learning) and $\hat{\beta}$ (the parameter of the model and the focus of social scientists). The focus on $\hat{\beta}$ arises from an interest in mapping the (often causal) relationships between variables. Broadly speaking, machine learning can be helpful in this area as well—particularly if we abandon the idea that the relationship between two variables needs to be represented by a single parameter in a linear model (Lundberg et al. 2021)—but it requires refocusing on the specific tasks relevant to social science.

The focus on prediction has led machine learning researchers to adopt substantially more complex models than the traditional linear and generalized-linear models that are common in the social sciences. There are some good reasons that social scientists have been slow to adopt these models—the predictions can be a more opaque function of the inputs, the models may not easily provide meaningful estimates of uncertainty, and the estimation routines are frequently more computationally intensive. These methods do, however, provide much better predictive performance. In the remainder of this article, we consider how to approach other tasks that are common in social science while harnessing this improved predictive power.

Before we do so, we describe seven broad tools in the machine learning toolkit that we will return to throughout the article. Given the emphasis on prediction, we start with how to assess performance of a prediction method in a new data set.

2.1. Assessing Performance

When we fit a model to data—particularly a very flexible model—prediction error within that data set is no longer a good indicator of prediction in new data. Intuitively, this is because a complex model will always be able to fit a pattern to the available data, but we have no guarantee it is picking up more signal than noise. Sample splitting, the first tool in the machine learning toolkit, addresses this problem directly.

> **Tool 1: Sample splitting**. Partitioning the data into multiple disjoint sets so that separate pieces can be used for different purposes. Most commonly, a training set is used for model fitting; a smaller validation set is used for provisionally evaluating performance when making model choices; and a test set is used exactly once, at the very end of the process, for reporting the final accuracy metric.

Sample splitting plays a crucial role in the machine learning pipeline because (when used only once!) the test set provides an unbiased estimate of the performance of the classifier on unseen data drawn from the same distribution as the test set. This guards against the excessive optimism of the training set.

The disadvantage of sample splitting is that data must be distributed across the different sets, leaving fewer data to train on. Because data are often the most valuable resource, this can be a high price to pay. In these settings, a useful technique is V-fold cross validation (Efron & Gong 1983, Hastie et al. 2013).

> **Tool 2: V-fold cross validation**. Partition the data randomly into V folds (groups) and fit the model on all but one fold, holding the last one out. Predictions are generated on the held-out fold. This process is repeated V times, such that every fold of data points is held out

exactly one time and used in training $V-1$ times. The collected set of held-out predictions—one per observation in the data—can then be used to evaluate predictive accuracy.

For large V, the number of observations the model is trained on approaches the size of the data set (while still preserving the held-out set), but at the cost that the model must be fit many times, which is often computationally prohibitive.

Cross validation is only reliable when all decisions about the procedure are made only on the basis of the data in the training folds (not the full data set). Thus, the pragmatic limitation is that all decisions have to be automated. By comparison, with a standard sample split, an analyst can examine the training set, check against the validation set, and make judgment calls about what decisions will be most useful. For this reason, it is common to see cross validation used as a technique within a training set to make choices about model specification or hyperparameters.

While both sample splitting and cross validation are powerful, it is important to remember that they provide an accurate estimate of the prediction error of interest only if the data are a random sample from the population of interest. Sometimes this is straightforward, and other times it is fundamentally impossible—for example, when our goal is to predict the future. This is a core challenge for predictive machine learning—measuring the intended generalization performance. This limitation notwithstanding, tools to detect overfitting play an essential role in the machine learning toolkit by opening up the possibility of adopting more complex models than we might otherwise use.

Hyperparameter: a parameter that controls some aspect of the learning algorithm, such as the model complexity or level of regularization

Generalization performance: performance of the model in a sample separate from the training set

Loss function: a function that takes in both the model's predictions and the true values to be predicted and returns a measure of performance

Features: the observed properties of the unit of analysis, e.g., the machine learning terminology for covariates in a linear regression

2.1.1. Fitting models. At the core, fitting machine learning models involves optimization of some parameters with respect to an objective function defined over the data—for example, the sum of squared residuals loss function that defines the ordinary least squares (OLS) regression covered in all standard social science statistics courses. What tends to distinguish machine learning methods is that they can fit substantially more flexible functions. While sample splitting and cross validation help detect an overly optimistic view of predictive performance, they do not provide a direct way to fight against the overfitting phenomenon. For this we turn to regularization.

To make this concrete, consider the OLS model. We have n data points indexed by i, each of which has an outcome we want to predict, y_i, and a column vector of p features, $\boldsymbol{x}_i$. We can estimate this model by minimizing the sum of squared residuals:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i}^{n} (y_i - \boldsymbol{\beta}'\boldsymbol{x}_i)^2.$$

However, when the dimensionality of the predictors is high relative to the sample size, the estimator may provide high variance estimates or may not even be defined at all (in the case that $p > n$). Regularization is a valuable tool for stabilizing estimates.

Tool 3: Regularization. Regularization is the process of adding non-data information or a constraint to the parameters in order to prevent overfitting. The most common form of regularization directly adds a penalty to the objective function we are optimizing. For example, in the context of OLS, we might add a regularizer by minimizing

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \underbrace{\sum_{i}^{n} (y_i - \boldsymbol{\beta}'\boldsymbol{x}_i)^2}_{\text{sum of squared residuals}} + \overbrace{\lambda \sum_{j=1}^{p} |\beta_j|^q}^{\text{complexity penalty}}$$

for some positive regularization penalty λ and a value q that defines the type of regularizer (Hastie et al. 2013).

Note that as $\lambda \to 0$, the regularized regression collapses on the OLS solution, while as $\lambda \to \infty$, the optimal solution is for each coefficient to be zero. Regularization encodes a trade-off between more accurately fitting the data and using a simpler function. Changing q results in penalties with different properties: Setting $q = 2$ produces ridge regression (which tends to draw coefficients smoothly toward zero), and setting $q = 1$ produces the LASSO (which induces sparsity—setting many coefficients to exactly zero).

Implicit here is a general idea in machine learning that it is better to use a very flexible model constrained by regularization than to constrain the model ex ante by using fewer predictors.[2] The downsides are that regularization biases the estimates of the parameters and that the complex model may be less interpretable. Nevertheless, these methods can produce much more accurate predictions than standard methods such as OLS across a range of problems.

The form of regularization that we have shown here is the most iconic representation; however, many techniques in machine learning have been shown to work because they correspond to a form of regularization. Bayesian priors, data augmentation, dropout in neural networks, and early-stopping have all been shown to correspond to regularization of different kinds. Many modern machine learning systems make use of both explicit regularization (e.g., ridge penalties) and implicit regularization (e.g., early-stopping and dropout).

Regularization provides various ways to tune our empirical models to improve their performance. We can choose different penalties (by changing q) or set the strength of the penalty (by setting λ). More broadly, there is a huge range of different machine learning models, most of which have some parameters that must be specified when the model is fit. In practice, these parameters are often set via our fourth tool, hyperparameter search.

> **Tool 4: Hyperparameter search.** For a few hyperparameters that control the behavior of our models (e.g., λ in ridge regression or LASSO), we can set their value by evaluating their performance in data separate from the data they were trained on (via a validation set or cross validation).

The core intuition is that rather than making choices a priori, we simply choose the model or parameter that maximizes generalization performance. This is often done by performing a grid search over possible values of the parameter and choosing the value that provides the best performance.

2.1.2. Automatic feature engineering. In the early days of artificial intelligence, most systems were rule-based mapping between features and outcomes—that is, an expert would write down an explicit decision rule. Although this is an intuitive idea, these systems ultimately proved difficult to construct and fragile. One of the major innovations of machine learning was demonstrating that a more effective strategy was automatically learning the mapping between features and outcomes using many data points. One of the most important parts of this strategy is the selection of features to use in the model. Thus, much of the human effort was built around this idea of feature engineering—constructing features that would be highly predictive of the outcome, a process familiar to any social scientist who has carefully selected a particular measure of a latent concept such as ideology or economic growth. Deep learning upended decades of such research in

[2] Regularization in the regression example above can equivalently be written as a literal constraint on the norm of the coefficients (instead of as a penalty to the loss function). In other words, the regularization ensures that the solution does not exceed some degree of complexity as measured by the norm (see, for example, Hastie et al. 2013).

natural language processing and computer vision in just a few years by demonstrating that learning features from very large collections of data often beats crafting them by hand.

Tool 5: Automatic feature engineering. Many machine learning models reduce to a linear or logistic regression model on a set of inputs $\boldsymbol{x}_i$ that have been transformed to a new set of basis functions $\phi(\boldsymbol{x}_i)$. We refer to the process of learning the transformations $\phi(\cdot)$ with the model as automatic feature engineering.

To give a concrete example, a neural network with a single hidden layer is a logistic regression where each feature is itself a logistic regression of all the original inputs. High-performing methods such as classification and regression trees, boosting, and generalized additive models can all be described as adaptive basis function models because they learn transformations of the inputs in the process of estimating the model (Murphy 2012).[3]

The largest gains of automatic feature engineering are in fields like computer vision, where the atomic feature—a single pixel of an image—is close to meaningless on its own. Only through interactions with other features around it does the pixel take on meaning, necessitating more complex features. Historically, this has been unnecessary for social science data as scholars have focused on a small number of data sets with carefully constructed survey questions and hand-engineered measures. However, in the modern era, data sets are larger and messier with many more features (Salganik 2018), making the automatic feature engineering approach more relevant than ever.[4]

A relatively recent insight is that, in many settings (notably the analysis of images and text), we can exploit similarity across data sets by using large preexisting data sets to learn features in advance. This is the core insight of transfer learning, which relies on the idea that the features that help perform a range of image and text classification tasks are roughly the same across tasks. Transfer learning has proven to be an invaluable strategy when the number of data points in our training set is too low to use the most complex forms of automatic feature engineering.

Basis function: function of the inputs that allows us to represent nonlinear functions using a linear model, e.g., X^2, $\log(X)$

Transfer learning: using information from one problem to help solve a different problem

Supervised: in a supervised task, the goal is to learn a mapping between inputs and an output based on examples of the input–output pairs

Unsupervised: in an unsupervised task, the goal is to discover patterns in a set of inputs. Unlike a supervised task, an unsupervised task has no labeled outputs

2.1.3. Reducing dimensionality.

One of the characteristics of the age of data abundance is that data sets have a lot of variables. Implicitly or explicitly, a common assumption in machine learning is that these high-dimensional data are approximately low dimensional. The successes in both the social science and machine learning communities with finding low-dimensional structure in even highly complex data sources (e.g. images and text) suggest that the approximately low-dimensional assumption is often reasonable. The idea of automatic feature engineering is to produce a supervised dimension reduction where only the information in the inputs relevant to the outputs is preserved. However, there has also been great success with unsupervised dimensionality reduction, which attempts to preserve as much information as possible from the inputs under the constraint of a low-dimensional representation.

Tool 6: Unsupervised dimensionality reduction. Given some data $\boldsymbol{x}_i$ with dimension p, a dimensionality reduction method generates a representation π_i with dimensionality K,

[3]Most methods not covered by this description use one of two related ideas: (*a*) They explicitly create an enormous number of nonlinear transformations and use regularization to select among them (Hansen et al. 1997, Ratkovic & Tingley 2021) or (*b*) they implicitly enumerate a large (possibly infinite) number of transformations using kernel methods and regularize (Shawe-Taylor & Cristianini 2004, Hainmueller & Hazlett 2014). As in the case of automatic feature engineering, we have shifted the role of feature selection from a human-driven process to a data-driven one.

[4]The move toward automatic feature engineering does not preclude expert knowledge. When some knowledge about the optimal feature transformation is available, it will always be more efficient to use that knowledge.

which is generally much smaller than p. This representation is chosen to minimize a loss function between the original data $\boldsymbol{x}_i$ and a learned function of the representation. These methods are called unsupervised because they do not involve a labeled outcome y_i.

Clustering: a technique for grouping units into inductively determined, mutually exclusive, and exhaustive sets

K-means: an iterative algorithm for clustering that partitions units into K clusters by minimizing the distance between a unit and its assigned cluster. Each cluster is characterized by the mean of observations assigned to it

Admixture: similar to cluster analysis but representing each unit with a set of proportions (nonnegative weights that sum to one) that represent membership across all clusters

Cosine distance: a particular measure of similarity based on the cosine of the angle between two vectors. It is often used because it is invariant to the magnitude of the vectors

There are many unsupervised dimensionality reduction methods, but they vary in four broad ways: (*a*) the form of the low-dimensional representation (π_i), (*b*) the function of the representation that produces the reconstruction, (*c*) the loss function that is minimized, and (*d*) the algorithm used to do the minimization.

The low-dimensional representation π_i is often constrained to take a particular form. Making it a K-dimensional vector consisting of a single one and the rest zeroes results in a clustering model, where π_i indicates the cluster to which unit i belongs, as in K-means. Constraining π_i to be nonnegative and sum to one (i.e., a set of proportions) leads to an admixture model—perhaps familiar to text analysts from latent Dirichlet allocation (LDA) (Blei et al. 2003). Finally, π_i can take on any real value, which leads to models like principal components analysis (PCA) or variational autoencoders (Kingma & Welling 2019).

For many simple dimensionality reduction models, including K-means, PCA, and LDA, the reconstruction function is a linear function parametrized by some weights (which may have some constraints of their own). Newer methods such as variational autoencoders can have substantially more complex, nonlinear reconstruction functions, which allow them to represent the data with high accuracy using a smaller K. The trade-off of these more complex models is that they are both harder to fit and harder to interpret.

The loss function defines what it means for a data point to be reconstructed well and influences the degree to which extreme dimensions of $\boldsymbol{x}_i$ can affect the representation. For instance, suppose we are performing a K-means clustering. We might define the distance between an observation and its reconstruction using squared Euclidean distance (which uses a squared error loss as in linear regression). However, if we are clustering documents using the counts of words they contain, we might care less about the length of the document than about the types of words that are typical. By using a different metric such as cosine distance, which measures only the angle between the reconstruction and observation, we can prioritize a different aspect of the representation (in this case, downweighting the role of document length). Loss functions control many other aspects of the reconstruction, including sensitivity to extreme data points and asymmetry of certain types of errors. After the loss function is selected, it is minimized by some algorithm, the choice of which can matter quite a lot for the representation when the implied optimization problem is very difficult (e.g., LDA) or matter very little when the implied optimization problem is comparatively easy (e.g., PCA).

While there are many dimensionality reduction methods, they are all based on the bottleneck principle. This is the idea that by forcing a high-dimensional observation to be reconstructed from a very low-dimensional latent variable (i.e., the bottleneck), we can distill the essence of the data. From a statistical perspective, this can be seen as removing noise or smoothing. From a social science perspective, we often think of the low-dimensional space as representing an underlying property of the observation.

2.1.4. Making progress as a field. Machine learning has made astonishingly rapid progress as a field over the last 30 years. There are many reasons for this—funding, public interest, the conference structure for publication, the adoption of the laboratory model of research. However, a key aspect is what Donoho [(2017), crediting Liberman (2010)] calls the "the secret sauce" of predictive culture, the common task framework.

Tool 7: The common task framework. A common task framework has three components: a publicly available training data set, a set of competitors who share the common task of

predicting some outcome with the data, and a referee who judges submissions in an objective and automatic way.

In practice, many areas of machine learning are organized around making progress on a comparatively small number of benchmark prediction tasks. The advantage of benchmarks is that they can focus attention and innovation on particular tasks and provide an obvious way to compare across methods and thus adjudicate progress. For example, the ImageNet Challenge, a prediction competition that involved classifying images into a thousand classes, provided a comparable target metric that demonstrated the transformative potential of deep convolutional neural networks (Russakovsky et al. 2015). Here again machine learning contrasts sharply with the social sciences, where most papers are studying different questions and answers are rarely comparable.

There is a fundamental challenge in benchmarks—emulating the true target task. This could be for (at least) two reasons: (*a*) a lack of appropriate test data and (*b*) a narrow performance metric. Sometimes we lack appropriate test data because they are hard to collect, but in other cases, data collection is fundamentally impossible. Consider, for example, causal inference. A core tenet—the fundamental problem of causal inference—is that we never get to observe a causal effect. This makes it extremely difficult to create a realistic benchmark for causal effect estimation because the thing we are trying to predict is not observable by construction. Some have tried to construct benchmarks by randomized experiments (Lalonde 1986) or simulation (Dorie et al. 2019), but these approaches may not reflect real applications. Even in systems where prediction is the goal, the system of interest may be constantly changing in a way that rules out obtaining representative test data.

Even when appropriate test data do exist, the structure of benchmark tasks encourages maximization of the scored performance metric to the exclusion of all other considerations. In the last few years, the machine learning community has begun to come to grips with other values—such as fairness, interpretability, and privacy—that are important to the application of machine learning systems but may not be easily formulated into a scorable performance metric. Inattention to these values can lead to pernicious effects, such as exacerbating existing racial inequalities in society (Benjamin 2019). These values are not necessarily in tension with traditional predictive accuracy metrics—Rudin (2019) argues that simpler, naturally interpretable models can be as accurate as more complex and opaque ones. However, preserving these values requires looking beyond any one benchmark metric.

3. A TASK-BASED APPROACH TO MACHINE LEARNING IN SOCIAL SCIENCE

In this section, we revisit the deductive model of social science, where quantitative tests and variables must be defined a priori. Rather than supposing that our theories are completely developed before looking at data, we emphasize that machine learning offers many tools to help generate the research questions, concepts, and hypotheses that can later be rigorously tested with new data. We advocate an agnostic approach to these machine learning tools—focusing on selecting the method that optimizes performance for a given research task rather than seeking a true model of the data.

3.1. Reconsidering the Model of Deductive Social Science

The most common process in the social sciences—evident in published research and conveyed to graduate students in research seminars—is that before viewing or collecting any data, authors must have a clear theory from which they derive a set of testable propositions. In this linear view, researchers must divine the concepts that structure their variables of interest; then, use a

strategy to measure the prevalence of those concepts; and finally, develop a set of hypotheses and a research design to test the observable implications of their stated theory (King et al. 1995). This understanding of the research process is so prevalent that it is often synonymous with principled research. An extreme version of this approach supposes that the theory and observable implications are determined before examining any original data that are collected for a project. Indeed, each of us has written several papers that follow this research model.

This standard deductive approach has many virtues both inside and outside academia and can be particularly powerful when there are known or established theories that have testable implications. It encourages analysts to reflect on their beliefs about the mechanistic processes that underlie the phenomenon they seek to understand. If followed explicitly, the deductive approach helps to reduce false discoveries that can result from researcher discretion. This is the rationale behind preregistration of hypotheses and analysis procedures before running an experiment (Humphreys et al. 2013). Most importantly, in an era when data were scarce, this approach was the most efficient use of that precious resource.

However, forcing researchers to use data to test theories that were developed before the data arrive also leaves potential insights on the table. Qualitative scholars in the social sciences have long acknowledged the importance of more inductive forms of analysis in qualitative research, including full-cycle research design, grounded theory, nested analysis, and abductive analysis (Glaser & Strauss 1967, Chatman & Flynn 2005, Lieberman 2005, Tavory & Timmermans 2014). Researchers often discover new directions, questions, and measures within quantitative data as well, and, as we detail below, machine learning can facilitate those discoveries. Under the standard deductive procedure, researchers might miss the opportunity to refine their concepts, develop new theories, and assess new hypotheses. We learn a great deal while analyzing data, and, as we show in the sections below, machine learning can be used to facilitate this learning. If researchers are up front about the role of quantitative discovery in research, we can better innovate methods for discovery and ensure that our discoveries are tested on new or held-out data.

3.2. An Agnostic Approach to Machine Learning Methods

We advance an agnostic approach to using machine learning to make social science inferences. By agnostic, we mean that our general view is that in many instances there is no correct or true model that we target with machine learning methods, and there is no one best method that can be used for all applications of a data set. Instead, we advocate that researchers use the method that optimizes performance for their particular research task. Unlike in many computer science applications, this task will often not be prediction but may instead be discovery, measurement, description, or causal inference. How we evaluate the quality of a model and compare models will be specific to the research question.

Our agnostic view contrasts most sharply with a structural approach to applying machine learning methods. As the name suggests, a structural approach posits the existence of an underlying true data-generating process, even in the discovery phase. The clearest and most influential statement of the structural view is found in Denny & Spirling's (2018) insightful paper on how to think about different ways to preprocess texts. As we will see, if we assume that there is some underlying true model or organization of data, problems like model selection are much more straightforward. But this approach inevitably involves strong and often unrealistic assumptions about the data-generating process. For example, we may have to assume that there is a true underlying set of categories or a right organization of our data, which might make little sense in some applications.

Consider a simple example: Suppose we want to organize campaign advertisements for a study of how political candidates present themselves to the public (Vavreck 2009). For some

applications, we might be interested in the tone of the advertisement—positive or negative. Other times, we might be interested in the topic of the advertisement. Or we might find that a nested conceptualization—with coarse topics composed of more granular organizations of the text—is most useful to our question.

There is no sense in which any of these organizations are correct unless we have more information beforehand about the goal of the research question. Critically, machine learning methods can help researchers to discover and identify different organizations, which in turn could lead to new research questions, hypotheses, and insights into political campaigns. During the discovery phase, researchers might prefer methods that fail to be consistent or unbiased because they yield the most useful insights.

Once the research question is defined and the researcher is performing measurement, we can assess how accurate a method is at recovering the specific measure of interest to the researcher. In other words, we are agnostic about which particular model is used for measurement, as long as the model can accurately and reliably measure the concept of interest. This perspective leads us to place a heavy emphasis on validations that demonstrate the connection between the inferred representation and the concept we claim it is measuring. These validations often involve careful reading of the documents, codebooks with category definitions, and/or application-specific criteria rather than model-fit diagnostics because—consistent with our agnostic perspective—the model is only an approximation.

Embedding: mapping of units to a low-dimensional, real-valued vector that contains information about the unit

4. DISCOVERY

Social science research is often presented as though the basic concepts we use to organize the empirical world are given. Consider an example from recent research. McGhee et al. (2014) examine how moving from a closed primary (where only party members can vote) to an open primary (where any eligible voter can cast a ballot) affects ideological polarization in state legislatures. On its face, this question is clear and corresponds with important questions about how to reform American politics, but it also assumes a particular organization of the world. Primary elections have to be categorized according to who is eligible to vote, and members of the legislature are organized by their placement in an ideological space.

Conceptualizations—organizations of the data—are often taken as given within empirical work or treated as objects to be created separate from the empirical research process. Quantitative empirical work tends to ignore the process of concept formation, even though the qualitative methods field has a rich literature on the creation of new concepts. Grounded theory (Glaser & Strauss 1967) and abductive analysis (Tavory & Timmermans 2014), to give two examples, provide iterative frameworks for moving, in a principled way, from initial field notes and interviews toward the generation of organizations, explanations, and new hypotheses.

Machine learning procedures expand our tools for engaging in data-driven discovery of new concepts underlying data. Several scholars in sociology have offered frameworks for using quantitative methods to enhance these qualitative methodologies, scoping out a place for quantitative methods in what has traditionally been a qualitative task (Baumer et al. 2017, Nelson 2017, Karell & Freedman 2019). By grouping together similar observations or features in a data set, unsupervised dimensionality reduction and automated feature engineering can suggest organizations of large data sets that might otherwise be hard to find or costly to produce manually. In this section, we detail machine learning methods—including clustering, admixtures, and embeddings—that can help social science researchers identify new ways of organizing and conceptualizing their data by compressing the high-dimensional data into a low-dimensional latent representation.

The goal of discovery, obtaining a useful organization of the text, is vague, which can make evaluation challenging. Model fit or predictive performance on held-out data—criteria often used in computer science—do not always correspond with human evaluations of good or interesting organizations of data (Chang et al. 2009). Instead, we suggest evaluating new organizations and conceptualizations discovered with quantitative methods according to their utility for researchers. This often requires clever customized validation checks and human-in-the-loop evaluations.

Taking an inductive approach to research can lead researchers in new directions and help them uncover unexpected but substantively important political dynamics. To see how, consider an unlikely starting point: congressional credit claiming about fire department grants. A surprising fact about congressional press releases is that approximately 2% are about small grants made to local fire departments through the Assistance to Firefighters Grants Program (AFGP). To discover this fact, Grimmer (2010) applied a topic model, the expressed agenda model, to a collection of press releases from US Senate offices. This revealed a category of press releases about fire departments, referencing "AFGP" along with the regular use of the word "announce." No conceptualization of congressional speech had ever proposed this as a relevant category of speech, but Grimmer was able to validate this category as referring to a real program and found that classification of press releases into this category was surprisingly accurate.

But why were there so many press releases about fire departments? After replicating the finding in House press releases, Grimmer et al. (2014) decided to learn more about the origins of fire department grant press releases. Their first finding only increased the intrigue: The AFGP was designed to be impervious to congressional influence, with grants awarded on a competitive basis. To find out how legislators were able to write press releases that claimed credit for the grants, Grimmer et al. (2014) interviewed officials who oversaw the distribution of grants, one of whom unraveled the mystery. She explained that the officials at the program alerted congressional officials about the grants before the official agency announcement, creating an opportunity for elected officials to receive credit for the grants. Grimmer et al. (2014) confirmed this pattern in the aggregate by using their press release data and information about the timing of AFGP announcements.

This led to a new and deeper question: If legislators cannot influence the allocation of the grants, what do they say in their press releases that effectively brings them credit? Grimmer et al. (2014) discovered that they merely imply that they are responsible without ever explicitly claiming credit. Time and again, in the fire department press releases, legislators "announce" new money to the district and speak in broad terms about continuing to fight for funding like this. This is in contrast to credit claiming for expenditures that legislators are actually directly responsible for, where they explain that they were able to "secure" funding for the district directly. Grimmer et al. (2014) show that this implicature is an effective strategy with constituents. In their experiment, when legislators implied that they deserved credit they received credit, but when legislators explained their actual contribution, constituents were reluctant to reward them.

This example shows the power of being explicitly inductive in our work. Rather than expecting their hypotheses to be present beforehand, Grimmer et al. (2014) assembled evidence inductively, learning about the phenomenon under study. Where appropriate, they applied more deductive approaches—such as an experiment to demonstrate the effectiveness of implicature. Not only is it false to pretend that this discovery followed from a deductive approach—it also obscures the process of science.

Methods for unsupervised dimension reduction can be important tools for data-driven discoveries of this sort. Next, we describe three such unsupervised techniques—clustering, admixtures, and embeddings—that can aid discovery. We then describe how supervised approaches can aid in discovery using an approach we call the fictitious prediction problem.

4.1. Clustering

Clustering methods partition observations into mutually exclusive categories, or clusters, using the principles of unsupervised dimension reduction. Specifically, the goal of clustering methods is to place each observation into one of K clusters. Formally, for every observation i, we constrain the low-dimensional representation π_i to be a K-length vector, with a one indicating the cluster to which the observation belongs and zeroes for the rest. Crucially, the number of clusters K is usually chosen beforehand, but what constitutes those clusters is not predetermined. The methods will also provide an estimate of the cluster center $\boldsymbol{\mu}_k$, which can often be interpreted as an exemplar observation for each cluster. There are many clustering methods across fields, with methods based on algorithms (Frey & Dueck 2007), statistical models (Fraley & Raftery 2002), graphical models (Ng et al. 2002), and either hierarchical (Fraley 1998) or flat (Park & Jun 2009) clustering. Clustering methods have been applied in political science to study the types of democracies (Ahlquist & Breunig 2012) and the structure of public opinion (Blaydes & Linzer 2008), to categorize countries (Wolfson et al. 2004), and even to identify observations that conform to particular theories (Imai & Tingley 2012).

4.2. Admixtures

Admixture models are closely related to clustering models and have received even more recent attention. An admixture model assumes that each unit has proportional membership in a set of latent categories, in contrast to clustering methods, which assume that each unit belongs to only one cluster. Thus, the latent representation π_i is still a K-length vector but is now constrained to be nonnegative and sum to one.

Admixture models have been adapted for many types of data, including discovering communities in networks (Airoldi et al. 2008) and types of respondents in surveys (Erosheva et al. 2007), but admixture models exploded in popularity after the introduction of LDA for text analysis (Blei et al. 2003). Political scientists initially extended LDA by adding covariates in an ad hoc way (Grimmer 2010, Quinn et al. 2010). The structural topic model generalizes these models, providing a default method for incorporating covariates into models (Roberts et al. 2016). Topic models have been applied to a wide array of data sets across the social sciences, including tweets (Jamal et al. 2015), newspapers (Jacobi et al. 2016), and discourse around climate change (Tvinnereim & Fløttum 2015). Similar models are used for studying culture across countries (Blaydes & Grimmer 2020).

4.3. Factor Models and Embeddings

The earliest machine learning methods in political science were factor models: methods that seek to find a low-dimensional and continuous representation of texts to approximate a high-dimensional data set. In a factor model, each observation is located in a continuous space, so for each unit $\pi_i \in \mathbb{R}^K$, where K is the dimension to be chosen. For over 60 years, factor models have been used as a tool to study the structure of survey responses (Visser et al. 2000). Factor models have also been used to study the structure of preferences in legislatures and the public, without specifying beforehand the structure of the underlying preferences in the data. By far the most successful research agenda has focused on estimating the structure of preferences within American political institutions. The most widely used method to estimate those preferences is NOMINATE (Rosenthal & Poole 1985), which has been extended to an item response theory framework (Clinton et al. 2004). Similar methods have been applied to study preferences using donation data (Bonica 2013), survey responses (Shor & McCarty 2011, Tausanovitch & Warshaw

2013), and even the words that politicians speak (Slapin & Proksch 2008). The output from these models has been used in thousands of papers to study political competition within the United States and across the world.

Fictitious prediction problem: a prediction problem where the goal is to learn the features that enable prediction; the predictions themselves are not of interest

4.4. Fictitious Prediction Problems

We call the last class of approaches to discovery fictitious prediction problems because they convert the discovery task into a supervised prediction problem where the prediction itself is not of interest. While the idea is broadly relevant, it is most easily illustrated through applications in text analysis. For example, we might ask what words predict whether a member of Congress is a Democrat or a Republican (Gentzkow et al. 2019). The actual interest, though, is discovering features that convey a particular category. For example, Gentzkow & Shapiro (2008) and Monroe et al. (2008) use the output from a fictitious prediction problem to identify "Republican" and "Democratic" words. Nelson (2017) uses a fictitious prediction problem to discover distinctive words used by feminist movements in different geographic locations as part of an explicitly inductive theoretical exercise. The core idea unifying these approaches is that in finding the features that predict some piece of information that is already known (hence the fiction), we learn something about what distinguishes the categories themselves.

4.5. What Do Different Conceptualizations Mean and How Do We Evaluate Them?

The use of machine learning methods for discovery has led to important insights across a wide range of fields. These methods provide useful insights into data by suggesting new organizations, but estimating conceptualizations using rigorous statistical models does not imply that we have uncovered the true conceptualization. In fact, it makes little sense to discuss true conceptualizations, because different organizations of the data can be useful for different purposes. Machine learning methods can aid in discovering new organizations that may be useful in some circumstances but not in others.

This implies that we should not base our assessment of organizations on whether they come from a model or are hand crafted; rather, we should assess them for how useful they are for research. Researchers have argued that model-based approaches to conceptualization are more trustworthy because they are data driven. For example, Ahlquist & Breunig (2012) apply model-based clustering methods to the varieties of capitalism data and find a different organization than the original authors uncovered. Ahlquist & Breunig (2012) interpret this to imply that the empirical foundation for varieties of capitalism may require empirical revision. Similarly, Quinn et al. (2010) argue that their model "estimates, rather than assumes" the underlying topics. In our agnostic view of machine learning, there isn't a right conceptualization for any data set. Both model-based and hand-engineered clusters require assumptions about similarity that can give rise to different organizations. Whether these organizations are useful to researchers will depend on the particular research question.

Similarly, when embedding models are used to estimate preferences, scholars often ask what the true dimensionality of the underlying preferences is (Koford et al. 1991, Levine et al. 1999). This is a theoretically compelling question, but it is a difficult question to answer empirically. Indeed, answers that are proffered depend not only on the data but also on the underlying modeling assumptions and an implicit decision rule that distinguishes between additional structure and noise. As a result, there can be no model-free determination of dimensionality.

Rather than place our trust fully in models and fit statistics, we argue that human feedback is essential for judging the quality of model results used for discovery. Researchers can assess the

model and determine the insights that come from a particular organization. Some organizations are useful because they explain another related phenomenon; some are useful because they reveal a previously unknown category, or a coarser organization that allows more generalization, or finer distinctions that increase explanatory power. It might be the case that better insights come from models that have better fit statistics, but there is no technical or theoretical reason why this must be the case.

Tying discovery methods to researchers underscores the utility of thinking about research as iterative and cumulative. Iteration implies that researchers can improve their conceptualization through refinements of the data. Of course, we might worry that we have overfit, obtaining a rule for a conceptualization found only in one particular data set. But examining many data sets or splitting data sets before engaging in discovery provides important safeguards from overfitting.

5. MEASUREMENT

Machine learning methods have allowed an era of custom measurement to flourish, as scholars can easily generate enormous quantitative data sets designed specifically for their own projects. Stylistically, these new measurement strategies sit somewhere between quantitative and qualitative traditions. They provide us with systematic, quantifiable summaries of empirical phenomena, but they require qualitative interpretation and considerable care in application.

This democratization of measurement is a large change from how measurement occurred in social science in the late twentieth century. Measurement during this time period was built around a relatively small number of large, centrally maintained data sets such as the General Social Survey, the American National Election Study, Militarized Interstate Disputes, and the Panel Study on Income Dynamics. The use of existing data is powerful because it lowers the cost of research by allowing scholars to piggyback off earlier work. However, this convenience comes at a price; these surveys had to be sufficiently broad that they could answer a wide range of questions, and occasionally the number of studies using their data is greater than the number of respondents. Large centralized data sets enshrine a particular conceptualization of empirical phenomena that can eventually exert a hegemonic influence on the field, shaping the nature of the questions that are posed. For example, gross domestic product (GDP) measurements are not collected for a specific academic study, but they are often used as a surrogate of economic strength. While GDP might be the right measure for some projects, it is not appropriate for all settings. The choice to use GDP can often go uninterrogated for lack of better options.

The advantage of having a few focal data sources is in validation. Large research teams and hundreds of scholars using the data ensure that the data and measures are carefully checked and validated. The democratization of measurement is powerful, but it thrusts more responsibility into the hands of the individual analysts, who must now focus more extensively on validation. Rather than relying on a shared knowledge of how to use a particular data set and a common understanding of its limitations, new measurement strategies require that researchers provide evidence that their measures capture the theoretical concepts of interest and are accurate.

Machine learning methods are useful for a wide range of measurement strategies. For example, machine learning methods can improve the classification of observations into categories, even if that classification is done primarily by hand. And increasingly, machine learning methods are used to extrapolate the hand coding decisions of coders to other data, greatly reducing the cost of using otherwise difficult-to-use data, such as text, images, and video. Machine learning methods have also been useful to make regression more flexible. We highlight how measurement methods can be used to improve hand coding and how machine learning can facilitate both classification and other descriptive exercises.

Crowd-sourced: collected by using a large number of nonexpert workers or volunteers

Kernel methods: a class of algorithms that represent units through their similarity to other units rather than their own features

5.1. Hand Coding

One of the most widely used tools in the social sciences is hand coding: manually classifying observations into a set of categories that are determined before the analysis begins. The usual procedure for hand coding involves training coders to reach a level of agreement and then subsequently ignoring any remaining disagreement when making inferences with the sample. This is problematic because any remaining error is guaranteed to bias inferences; neutral measurement error is not possible with categorical variables. There has been substantial recent work on developing methods to improve the results of hand coding. One method for hand coding is the Dawid–Skene algorithm (Dawid & Skene 1979), a flexible model that assumes only that the coders are making independent errors. Recently, scholars have improved upon Dawid–Skene, adapting it for social science applications (Tyler 2020), and have extended it to settings with many coders (Tian et al. 2019). Other scholars have focused on reliable methods for reducing the dependence on experts by leveraging crowd-sourced annotations (Benoit et al. 2016, Carlson & Montgomery 2017).

5.2. Regression and Supervised Learning

Collecting labeled data is expensive and time consuming. It often requires hiring coders, training them, monitoring their performance, and potentially retraining. One approach to circumvent this cost is to use machine learning regression methods to extrapolate from a relatively small training set to a much larger set of unlabeled data (for a useful practical guide in the text analysis case, see Barberá et al. 2021).

This process can be automated by turning the measurement problem into a prediction problem—tasking the machine learning system with predicting the label that a human would assign. Viewed in this way, all of the tools in the machine learning toolkit can be brought over to measure quantities of interest in large data sets even if these measures are only coded in a small subsample of data.

5.3. Improving Model Fit with Machine Learning Regression

Machine learning approaches are also widely used to make regressions more flexible, even without classification as an explicit goal. For example, Hainmueller & Hazlett (2014) introduce kernel regularized least squares (KRLS), a method that builds upon kernel methods, to flexibly estimate the regression surface. Scholars have made similar arguments for LASSO (Chen et al. 2019) and random forest (Hill & Jones 2014).

Perhaps the most widely used application of machine learning regression in the social sciences is to extrapolate from national surveys to lower-level geographies (Ghitza & Gelman 2013). A popular method to make this extrapolation is MrP. The "Mr" stands for a multilevel regression using survey responses and respondent characteristics. The multilevel regression is used to regularize the coefficients, enabling researchers to estimate sample means for groups that have few respondents in the survey. Then, the results of this regression are combined with census data to "poststratify" the results and estimate opinion levels in geographic units where there were few responses (Ghitza & Gelman 2013). MrP has been broadly applied to facilitate the study of public opinion on policies and representation across various research areas (Lax & Phillips 2009, Warshaw & Rodden 2012). A growing literature examines how different machine learning methods, such as gradient boosted decision trees, can improve upon the regularization from multilevel regression (Bisbee 2019).

5.4. How Do We Evaluate Measurement Models?

When validating a measure, we are showing that we are right about what a measure is capturing about the world. Validation of measurement models is more straightforward when a gold standard exists. By a gold standard, we mean that for a subset of units we observe the true value of the quantity that we want to measure.

Gold standard: perfectly labeled or otherwise idealized annotations/labels

Gold standard evaluations can be particularly valuable for assessing measurement models, even if they are produced at substantial cost. For example, one method for assessing the output of MrP measures is to run a survey of a large number of individuals in a geographic area and then compare the average opinion in the survey to the one produced by MrP. In a similar approach, Warshaw & Rodden (2012) assess the performance of MrP models using output from state-level referenda.

Validation without a gold standard is inherently more complicated. We might approximate gold standards using other data (Quinn et al. 2010), use human coders to assess the performance of measures, examine the properties of our models, and assess our measure's face validity—the extent to which it aligns with external data or conforms with broad expectations from subject area experts. While the best validations are generally application specific, crowd-sourced human judgments can provide one general approach to validation (Ying et al. 2019).

Regardless of how measures are validated within any one study, scientific trust in measures is built through iteration, cumulation, and openness. Iteratively, we have to refine models repeatedly or improve coder performance. Further, the best measures will be valid across different contexts, with different source materials, and when similar methods are used on similar data sets. These assessments can be made only if the measures are applied in new settings and with new data—a requirement that is in tension with customized measurement schemes that machine learning methods enable. The only way for scholars to truly assess whether measures perform as researchers claim is if the original source material is released, so researchers can understand how the underlying data were used to form the measurements and assess whether inferences from the model follow logically from the underlying data.

6. INFERENCE: PREDICTION AND CAUSAL INFERENCE

Discovery and measurement are essential goals in research, but scholars are most often interested in inferential tasks. One task is prediction, predicting an outcome that has been or will be realized from a set of inputs. Previously, we had assumed that we could randomly sample the population of interest and thus, on average, the mapping between inputs and outputs in the sample would hold among new cases from the population.

However, in social science we often cannot draw a random sample from the population of interest—for example, when forecasting the future. To assess performance, we then need to assume that our currently available data are not too different from the data we want to predict—an assumption that is ex ante unverifiable. In practice, we can creatively approximate similar scenarios, using data at some time $t-1$ to predict the outcomes at the current time t, but we will never know for sure that this pattern will continue to hold when we reach the future of $t+1$.

To see why this can be a strong assumption, consider a popular forecasting task: predicting presidential elections. Building on a 30-year-old literature in political science, a large number of forecasters followed Nate Silver's meteoric rise to predict presidential elections. The statistical models rely on a basic insight: Public opinion polls about presidential vote choice are predictive of election results. Using data from prior elections, forecasters model the relationship between opinion polls and the prior election results. They then use the relationship found in prior elections and current polling results to predict the result of current elections. The strong assumption is that

the relationship between the polls and outcomes found in prior elections will also be found in the current election. This assumption is unverifiable; the accuracy of forecasts in prior elections is no guarantee that they will be accurate in subsequent elections.

Counterfactual: an unobservable outcome that represents what would have occurred under a well-defined intervention

Making accurate predictions requires identifying features that are simultaneously predictive of the outcome of interest, available when the prediction is made, and likely to have the same relationship with the outcomes in the available sample and the population of interest. To go back to the election forecast, a large literature in American politics has noted that the economic conditions of the country are related to incumbent presidents' reelection chances. And yet, in the run-up to the 2020 election, forecasters were uncertain about how to handle the economic downturn in 2020 due to the COVID-19-related shutdowns. While an economic downturn usually portends disaster for the incumbent party, there was no guarantee that this would be the case in the 2020 election. This inherent uncertainty means that there were no data available to directly test whether the relationship between the economy and incumbent performance in 2020 would be similar to what it was in prior elections.

In contrast, consider the task of causal inference: How do outcomes differ if we intervene to change some feature of the world? The applications of causal inference are wide ranging: It is used in assessing the effect of taking a drug, the effect of implementing a policy in a country, the effect of an advertisement on the purchase decisions of consumers, and more. The key to a causal statement is that scholars ask how the world would have been different if only the intervention were different.

At first glance, we might think that causal inference is just another prediction problem. After all, the response of a unit to an intervention is an example of predicting a response outside of our data. Yet the characteristic feature of prediction—that the true outcome eventually is, or could be, revealed—does not hold for causal inference. This is the fundamental problem of causal inference (Holland 1986): We assess causal effects with counterfactual statements, but we observe only one version of history. Conflating prediction and causal inference has caused considerable confusion in the computational social science literature.

This conflation and the subsequent confusion represent a serious obstacle to progress on computational questions. To see why, consider three of the most salient differences between prediction and causal inference. First, their goals are fundamentally different. With prediction, our goal is to predict the value of the outcome for a unit from a different observable population, but with causal inference we ask how the outcomes differ under different counterfactual states of the world. Causal questions are necessarily comparative in nature: We ask what the effect of an intervention is on a set of units, and computing this effect necessarily requires comparing different counterfactuals for the same unit. Causal inference questions require a well-defined intervention that explains what is different between the two states of the world.

Second, evaluating performance is more complex for causal inference because we will never actually observe the counterfactual data that are essential for computing causal effects. Contrast this with prediction problems, where the truth is revealed at some point in the future.

Third, the features in prediction and causal inference serve fundamentally different roles. In prediction, the features are merely tools used to make a prediction for unobserved units. In causal inference, however, we differentiate between treatments and covariates. Treatments represent the intervention we are assessing the effect of. Covariates enable the more accurate, unbiased, and precise estimation of causal effects. This difference between prediction and causal inference leads to dramatically different advice about what information to include in our models. In prediction, we seek out whatever covariates we can use to predict the outcome well, are available when we make our prediction, and are likely to have a stable relationship with the dependent variable. In

causal inference, we have to be attentive to exclude variables that are post treatment and to avoid other variables (e.g., colliders) that can induce more bias than they mitigate.

The differences between prediction and causal inference can be subtle and require careful delineation. Suppose, for example, we are hired to assess the effectiveness of a company's internet advertising campaign. We can imagine the company might ask us to predict who will buy their product. To make this prediction, we would likely use a host of demographic, socioeconomic, and proprietary data. We can then imagine a separate question: Does this ad campaign cause individuals to buy the product? This is a causal question. It asks how the behavior of an individual who sees an advertisement differs from the same individual had they not been shown the advertisement—a question about counterfactual states of the world.

The company may ask more subtle questions that blend causal inference and prediction. For example, our client may ask, "Will this individual buy the product if we show them the advertisement?" Taken literally, this is a prediction question. It merely asks us to assess whether an individual (with a set of characteristics) will buy a product given that we also serve them with an advertisement. We might answer that there is a high probability of purchase even if the advertisement has little to no effect on the individual. In contrast, we can ask, "Who is most responsive to the advertisement?" This is a well-defined causal inference question that focuses on estimating conditional average treatment effects and identifying strata of individuals where the advertisement is found to be very effective.

For both prediction and causal inference, we draw a distinction between two components of the process: identification and estimation. Heuristically, we will say that parameters are identified if data sets map uniquely to a set of parameter values. The term identification refers to what we can possibly learn from our data sets with infinite data. For prediction, an identification assumption might be that the data that we have in hand contain the same relationships between inputs and outputs as the out-of-sample data. For causal inference, the identification assumption is that the treatment assignment is unconfounded (uncorrelated with the counterfactual outcomes). The term estimation refers to how we actually estimate the effects from our data set. Estimation assumptions might be about functional forms used for estimation or the way particular statistical models are fit. Both identification and estimation are important, but scholars of causal inference have tended to focus on identification as a first priority. This is because identification assumptions are unverifiable from the data, and without identification, no estimation routine could possibly find the correct parameters.

Identification: the process of assessing whether a quantity can be learned with infinite data. In causal inference, the process of establishing equivalence between observable data and the target counterfactual quantity

Estimation: the process of using data to approximate the value of some unknown parameter

6.1. Causal Inference and Machine Learning

In this section, we provide three examples of how causal inference and machine learning are used together: the use of machine learning to estimate heterogeneous treatment effects, to adjust for high-dimensional confounders, and to assess the effects of high-dimensional treatments or the effects of treatments on high-dimensional outcomes.

6.1.1. Heterogeneous treatment effects. The distinction between causal inference and prediction does not mean that the tools from prediction are unhelpful in performing causal inferences. For example, a growing literature uses machine learning methods to estimate heterogeneous treatment effects: how the effect of a particular intervention differs across characteristics of individuals (Imai & Ratkovic 2013, Athey & Imbens 2016, Grimmer et al. 2017, Wager & Athey 2018, Künzel et al. 2019). This information can be used both to more effectively target treatments and as indirect evidence of the mechanism through which the treatment operates. Mechanically, these problems

involve using machine learning to estimate the average treatment effects within strata defined by the covariates.

6.1.2. Machine learning methods for high-dimensional confounding. Scholars are often interested in adjusting for confounding with high-dimensional covariates. For example, Roberts et al. (2020) introduce text matching, a procedure to adjust for background confounders from text, which often uses a high-dimensional representation (see also Veitch et al. 2019, Keith et al. 2020, Mozer et al. 2020). Machine learning has also been applied to adjustment of standard covariates either because the covariates are high-dimensional or to allow greater flexibility in the functional form (Hill 2011, Johansson et al. 2016, Chernozhukov et al. 2017). While these approaches can make estimation assumptions more tractable, D'Amour et al. (2020) notes that in high-dimensional settings, the overlap assumption becomes considerably stronger. High-dimensional adjustment is an area of active research and will likely continue to develop quickly.

6.1.3. Assessing the effects of high-dimensional treatments or the effects on high-dimensional outcomes. Scholars are often interested in assessing the effects of images, advertisements, texts, and speeches on an individual's behavior. These treatments are fundamentally high-dimensional, with many potential facets that can affect outcomes. Fong & Grimmer (2020) provide a set of necessary and sufficient conditions for isolating the effect of one component of the treatments, and Fong & Grimmer (2016) provide a method for discovering underlying treatments. Approaches based on stochastic interventions can enable causal inferences in high-dimensional spatiotemporal settings as well (Papadogeorgou et al. 2020). High-dimensional data can also be an outcome, such as when scholars want to understand the effect of an intervention on the words that individuals say or the videos that news agencies produce (Roberts et al. 2014, Knox & Lucas 2021). In each case, analysis is predicated on the idea that the apparently high-dimensional data (whether treatment or outcome) can be represented in a low-dimensional way (Egami et al. 2018).

7. CONCLUSION: THE FRONTIERS OF MACHINE LEARNING

The era of abundance creates an opportunity for social scientists to reevaluate their approach to research. With ample data, deduction no longer needs to be the default. Sequential and interactive approaches to data analysis are now possible because the abundance of data allows researchers to avoid overfitting to a single sample.

Machine learning methods are ideally suited to help researchers use the abundance of data to its best capacity. Machine learning tools are just tools, though, and not a magic new method that resolves the long-standing problems that befall social scientists. Machine learning will perform best when it is applied appropriately to research problems. Researchers need to adopt a task focus when applying machine learning methods. Focusing on tasks not only clarifies to researchers which machine learning methods and tools are most appropriate to apply in a particular setting but also guides researchers on how to evaluate those methods.

While we focus on empirical work, the abundance of data and new empirical applications has made theory—and formal theory in particular—more important for the social sciences. Theory provides crucial guidance on how to design and interpret the results of analyses, guidance that the data can sometimes fail to provide (Ashworth et al. 2015, Slough 2019, de Marchi & Stewart 2020).

Abundance does not guarantee progress. There is a risk that researchers are seduced by the size of their data sets and forget hard-learned lessons from when data were scarce. But with careful

thought, a focus on designs, and the appropriate application of methods, we are optimistic that the era of abundance will lead to an era of insights.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We thank Chad Hazlett and David Stasavage for comments on an earlier draft and Naoki Egami and Christian Fong for collaboration on related work. Research reported in this publication was supported by the Eunice Kennedy Shriver National Institute of Child Health and Human Development of the National Institutes of Health under award number P2CHD047879.

LITERATURE CITED

Acharya A, Bansak K, Hainmueller J. 2021. Combining outcome-based and preference-based matching: the g-constrained priority mechanism. *Political Anal.* In press

Ahlquist JS, Breunig C. 2012. Model-based clustering and typologies in the social sciences. *Political Anal.* 20:92–112

Airoldi EM, Blei DM, Fienberg SE, Xing EP. 2008. Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.* 9:1981–2014

Aronow PM, Miller BT. 2019. *Foundations of Agnostic Regression*. Cambridge, UK: Cambridge Univ. Press

Ashworth S, Berry CR, De Mesquita EB. 2015. All else equal in theory and data (big or small). *PS: Political Sci. Politics* 48:89–94

Athey S, Imbens GW. 2016. Recursive partitioning for heterogeneous causal effects. *PNAS* 113:7353–60

Athey S, Imbens GW. 2019. Machine learning methods that economists should know about. *Annu. Rev. Econ.* 11:685–725

Barberá P, Boydstun AE, Linn S, McMahon R, Nagler J. 2021. Automated text classification of news articles: a practical guide. *Political Anal.* 29:19–42

Baumer EPS, Mimno D, Guha S, Quan E, Gay GK. 2017. Comparing grounded theory and topic modeling: extreme divergence or unlikely convergence? *J. Assoc. Inform. Sci. Technol.* 68:1397–410

Beck N, King G, Zeng L. 2000. Improving quantitative studies of international conflict. *Am. Political Sci. Rev.* 94:21–36

Benjamin R. 2019. *Race After Technology: Abolitionist Tools for the New Jim Code*. Cambridge, UK: Wiley

Benoit K, Conway D, Lauderdale B, Laver M, Mikhaylov S. 2016. Crowd-sourced text analysis: reproducible and agile production of political data. *Am. Political Sci. Rev.* 110:278–95

Bisbee J. 2019. BARP: improving Mister P using Bayesian additive regression trees. *Am. Political Sci. Rev.* 113:1060–65

Bishop C. 2006. *Pattern Recognition and Machine Learning*. New York: Springer

Blaydes L, Grimmer J. 2020. Political cultures: measuring values heterogeneity. *Political Sci. Res. Methods* 8:571–79

Blaydes L, Linzer DA. 2008. The political economy of women's support for fundamentalist Islam. *World Politics* 60:576–609

Blei DM, Ng AY, Jordan MI. 2003. Latent Dirichlet allocation. *J. Mach. Learn. Res.* 3:993–1022

Bonica A. 2013. Ideology and interests in the political marketplace. *Am. J. Political Sci.* 57:294–311

Breiman L. 2001. Statistical modeling: the two cultures. *Stat. Sci.* 16:199–215

Carlson D, Montgomery JM. 2017. A pairwise comparison framework for fast, flexible, and reliable human coding of political texts. *Am. Political Sci. Rev.* 111: 835–43

Chang J, Gerrish S, Wang C, Boyd-Graber JL, Blei DM. 2009. Reading tea leaves: how humans interpret topic models. In *Proceedings of the 22nd International Conference on Neural Information Processing Systems*, pp. 288–96. Red Hook, NY: Curran Assoc.

Chatman JA, Flynn FJ. 2005. Full-cycle micro-organizational behavior research. *Organ. Sci.* 16:434–47
Chen JKT, Valliant RL, Elliott MR. 2019. Calibrating non-probability surveys to estimated control totals using LASSO, with an application to political polling. *J. R. Stat. Soc. Ser. C Appl. Stat.* 68:657–81
Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, et al. 2017. Double/debiased machine learning for treatment and structural parameters. *Econom. J.* 21:C1–68
Clinton J, Jackman S, Rivers D. 2004. The statistical analysis of roll call data. *Am. Political Sci. Rev.* 98:355–70
D'Amour A, Ding P, Feller A, Lei L, Sekhon J. 2020. Overlap in observational studies with high-dimensional covariates. *J. Econom.* 221:644–54
Dawid AP, Skene AM. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *J. R. Stat. Soc. Ser. C Appl. Stat.* 28:20–28
de Marchi S, Stewart BM. 2020. Computational and machine learning models: the necessity of connecting theory and empirics. In *SAGE Handbook of Research Methods in Political Science and International Relations*, ed. L Curini, R Franzese, pp. 289–310. London: SAGE
Denny MJ, Spirling A. 2018. Text preprocessing for unsupervised learning: why it matters, when it misleads, and what to do about it. *Political Anal.* 26:168–89
Donoho D. 2017. 50 years of data science. *J. Comput. Graph. Stat.* 26:745–66
Dorie V, Hill J, Shalit U, Scott M, Cervone D, et al. 2019. Automated versus do-it-yourself methods for causal inference: lessons learned from a data analysis competition. *Stat. Sci.* 34:43–68
Efron B, Gong G. 1983. A leisurely look at the bootstrap, the jackknife, and cross-validation. *Am. Stat.* 37:36–48
Egami N, Fong CJ, Grimmer J, Roberts ME, Stewart BM. 2018. How to make causal inferences using texts. arXiv:1802.02163 [stat.ML]
Erosheva EA, Fienberg SE, Joutard C. 2007. Describing disability through individual-level mixture models for multivariate binary data. *Ann. Appl. Stat.* 1:502–37
Fong C, Grimmer J. 2016. Discovery of treatments from text corpora. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1600–9. Stroudsburg, PA: Assoc. Comput. Ling.
Fong CJ, Grimmer J. 2020. *Causal inference with latent treatments*. Work. Pap., Dep. Political Sci., Stanford Univ., Stanford, CA
Fraley C. 1998. Algorithms for model-based Gaussian hierarchical clustering. *SIAM J. Sci. Comput.* 20:270–81
Fraley C, Raftery A. 2002. Model-based clustering, discriminant analysis, and density estimation. *J. Am. Stat. Assoc.* 97:611
Frey BJ, Dueck D. 2007. Clustering by passing messages between data points. *Science* 315:972–76
Gentzkow M, Shapiro JM. 2008. Competition and truth in the market for news. *J. Econ. Perspect.* 22:133–54
Gentzkow M, Shapiro JM, Taddy M. 2019. Measuring polarization in high-dimensional data: method and application to congressional speech. *Econometrica* 87:1307–40
Ghitza Y, Gelman A. 2013. Deep interactions with MRP: election turnout and voting patterns among small electoral subgroups. *Am. J. Political Sci.* 57:762–76
Glaser BG, Strauss AL. 1967. *The Discovery of Grounded Theory: Strategies for Qualitative Research*. New York: Aldine de Gruyter
Goodfellow I, Bengio Y, Courville A, Bengio Y. 2016. *Deep Learning*. Cambridge, MA: MIT Press
Grimmer J. 2010. A Bayesian hierarchical topic model for political texts: measuring expressed agendas in Senate press releases. *Political Anal.* 18:1–35
Grimmer J, Messing S, Westwood SJ. 2017. Estimating heterogeneous treatment effects and the effects of heterogeneous treatments with ensemble methods. *Political Anal.* 25:413–34
Grimmer J, Westwood SJ, Messing S. 2014. *The Impression of Influence: Legislator Communication, Representation, and Democratic Accountability*. Princeton, NJ: Princeton Univ. Press
Hainmueller J, Hazlett C. 2014. Kernel regularized least squares: reducing misspecification bias with a flexible and interpretable machine learning approach. *Political Anal.* 22:143–68
Hansen MH, Kooperberg C, Truong YK, Stone CJ. 1997. Polynomial splines and their tensor products in extended linear modeling: 1994 Wald memorial lecture. *Ann. Stat.* 25:1371–470
Hastie T, Tibshirani R, Friedman J. 2013. *The Elements of Statistical Learning*. New York: Springer

Hill DW Jr., Jones ZM. 2014. An empirical evaluation of explanations for state repression. *Am. Political Sci. Rev.* 108:661–87
Hill JL. 2011. Bayesian nonparametric modeling for causal inference. *J. Comput. Graph. Stat.* 20:217–40
Hillard D, Purpura S, Wilkerson J. 2008. Computer-assisted topic classification for mixed-methods social science research. *J. Inf. Technol. Politics* 4:31–46
Holland PW. 1986. Statistics and causal inference. *J. Am. Stat. Assoc.* 81:945–60
Humphreys M, Sanchez de la Sierra R, Van der Windt P. 2013. Fishing, commitment, and communication: a proposal for comprehensive nonbinding research registration. *Political Anal.* 21:1–20
Imai K, Ratkovic M. 2013. Estimating treatment effect heterogeneity in randomized program evaluation. *Ann. Appl. Stat.* 7:443–70
Imai K, Tingley D. 2012. A statistical method for empirical testing of competing theories. *Am. J. Political Sci.* 56:218–36
Jacobi C, Van Atteveldt W, Welbers K. 2016. Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital J.* 4:89–106
Jamal AA, Keohane RO, Romney D, Tingley D. 2015. Anti-Americanism and anti-interventionism in Arabic Twitter discourses. *Perspect. Politics* 13:55–73
Johansson F, Shalit U, Sontag D. 2016. Learning representations for counterfactual inference. In *33rd International Conference on Machine Learning, ICML 2016*, Vol. 6, ed. KQ Weinberger, MF Balcan, pp. 4407–18. New York: Int. Machine Learning Soc.
Karell D, Freedman M. 2019. Rhetorics of radicalism. *Am. Sociol. Rev.* 84:726–53
Kaufman AR, Kraft P, Sen M. 2019. Improving Supreme Court forecasting using boosted decision trees. *Political Anal.* 27:381–87
Keith KA, Jensen D, O'Connor B. 2020. Text and causal inference: a review of using text to remove confounding from causal estimates. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5332–44. Stroudsburg, PA: Assoc. Comput. Linguist.
King G, Keohane RO, Verba S. 1995. The importance of research design in political science. *Am. Political Sci. Rev.* 89:454–81
Kingma DP, Welling M. 2019. An introduction to variational autoencoders. *Found. Trends Machine Learn.* 12:307–92
Knox D, Lucas C. 2021. A dynamic model of speech for the social sciences. *Am. Political Sci. Rev.* In press
Koford K, Poole KT, Rosenthal H. 1991. On dimensionalizing roll call votes in the US Congress. *Am. Political Sci. Rev.* 85:955–75
Künzel SR, Sekhon JS, Bickel PJ, Yu B. 2019. Metalearners for estimating heterogeneous treatment effects using machine learning. *PNAS* 116:4156–65
Lalonde R. 1986. Evaluating the econometric evaluations of training programs. *Am. Econ. Rev.* 76:604–20
Lax JR, Phillips JH. 2009. How should we estimate public opinion in the states? *Am. J. Political Sci.* 53:107–21
Levine J, Carmines EG, Sniderman PM. 1999. The empirical dimensionality of racial stereotypes. *Public Opin. Q.* 63:371–84
Liberman M. 2010. Fred Jelinek. *Comput. Linguist.* 36:595–99
Lieberman ES. 2005. Nested analysis as a mixed-method strategy for comparative research. *Am. Political Sci. Rev.* 99:435–52
Lin W. 2013. Agnostic notes on regression adjustments to experimental data: reexamining Freedman's critique. *Ann. Appl. Stat.* 7:295–318
Lundberg I, Johnson R, Stewart BM. 2021. What is your estimand? Defining the target quantity connects statistical evidence to theory. *Am. Sociol. Rev.* In press
McGhee E, Masket S, Shor B, Rogers S, McCarty N. 2014. A primary cause of partisanship? Nomination systems and legislator ideology. *Am. J. Political Sci.* 58:337–51
Molina M, Garip F. 2019. Machine learning for sociology. *Annu. Rev. Sociol.* 45:27–45
Monroe B, Colaresi M, Quinn K. 2008. Fightin' words: lexical feature selection and evaluation for identifying the content of political conflict. *Political Anal.* 16:372–403
Montgomery JM, Olivella S. 2018. Tree-based models for political science data. *Am. J. Political Sci.* 62:729–44
Mozer R, Miratrix L, Kaufman AR, Anastasopoulos LJ. 2020. Matching with text data: an experimental evaluation of methods for matching documents and of measuring match quality. *Political Anal.* 28:445–68

Mullainathan S, Spiess J. 2017. Machine learning: an applied econometric approach. *J. Econ. Perspect.* 31:87–106
Murphy KP. 2012. *Machine Learning: A Probabilistic Perspective*. Cambridge, MA: MIT Press
Nelson LK. 2017. Computational grounded theory: a methodological framework. *Sociol. Methods Res.* 49:3–42
Ng A, Jordan M, Weiss Y. 2002. On spectral clustering: analysis and an algorithm. In *Advances in Neural Information Processing Systems*, ed. T Dietterich, S Becker, Z Ghahramani, pp. 849–56. Cambridge, MA: MIT Press
Nielsen RA. 2017. *Deadly Clerics: Blocked Ambition and the Paths to Jihad*. Cambridge, UK: Cambridge Univ. Press
Papadogeorgou G, Imai K, Lyall J, Li F. 2020. Causal inference with spatio-temporal data: estimating the effects of airstrikes on insurgent violence in Iraq. arXiv:2003.13555 [stat.ME]
Park HS, Jun CH. 2009. A simple and fast algorithm for K-medoids clustering. *Expert Syst. Appl.* 36:3336–41
Quinn K, Monroe BL, Colaresi M, Crespin MH, Radev DR. 2010. How to analyze political attention with minimal assumptions and costs. *Am. J. Political Sci.* 54:209–28
Rashkin H, Choi E, Jang JY, Volkova S, Choi Y. 2017. Truth of varying shades: analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, ed. H Rashkin, E Choi, JY Jang, S Volkova, Y Choi, pp. 2931–37. Stroudsburg, PA: Assoc. Comput. Linguist.
Ratkovic M, Tingley D. 2021. *Estimation and inference on nonlinear and heterogeneous effects*. Work. Pap., Harvard Univ., Cambridge, MA. **https://scholar.harvard.edu/files/dtingley/files/mdei.pdf**
Roberts ME, Stewart BM, Airoldi EM. 2016. A model of text for experimentation in the social sciences. *J. Am. Stat. Assoc.* 111:988–1003
Roberts ME, Stewart BM, Nielsen RA. 2020. Adjusting for confounding with text matching. *Am. J. Political Sci.* 64:887–903
Roberts ME, Stewart BM, Tingley D, Lucas C, Leder-Luis J, et al. 2014. Structural topic models for open-ended survey responses. *Am. J. Political Sci.* 58:1064–82
Rosenthal H, Poole K. 1985. A spatial model for legislative roll call analysis. *Am. J. Political Sci.* 29:357–84
Rudin C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Machine Intel.* 1:206–15
Russakovsky O, Deng J, Su H, Krause J, Satheesh S, et al. 2015. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* 115:211–52
Salganik M. 2018. *Bit by Bit: Social Research in the Digital Age*. Princeton, NJ: Princeton Univ. Press
Shawe-Taylor J, Cristianini N. 2004. *Kernel Methods for Pattern Analysis*. New York: Cambridge Univ. Press
Shor B, McCarty N. 2011. The ideological mapping of American legislatures. *Am. Political Sci. Rev.* 105:530–51
Slapin JB, Proksch SO. 2008. A scaling model for estimating time-series party positions from texts. *Am. J. Political Sci.* 52:705–22
Slough T. 2019. *On theory and identification: when and why we need theory for causal identification*. Work. Pap., Dep. Politics, New York Univ., New York, NY
Stewart BM, Zhukov Y. 2009. Use of force and civil-military relations in Russia: an automated content analysis. *Small Wars Insurg.* 20:319–43
Tausanovitch C, Warshaw C. 2013. Measuring constituent policy preferences in Congress, state legislatures, and cities. *J. Politics* 75:330–42
Tavory I, Timmermans S. 2014. *Abductive Analysis: Theorizing Qualitative Research*. Chicago: Univ. Chicago Press
Tian T, Zhu J, Qiaoben Y. 2019. Max-margin majority voting for learning from crowds. *IEEE Trans. Pattern Anal. Mach. Intell.* 41:2480–94
Tvinnereim E, Fløttum K. 2015. Explaining topic prevalence in answers to open-ended survey questions about climate change. *Nat. Climate Change* 5:744–47
Tyler M. 2020. *Getting the most out of human coding*. Work. Pap., Dep. Political Sci., Stanford Univ., Stanford, CA
Vavreck L. 2009. *The Message Matters: The Economy and Presidential Campaigns*. Princeton, NJ: Princeton Univ. Press

Veitch V, Wang Y, Blei D. 2019. Using embeddings to correct for unobserved confounding in networks. In *Advances in Neural Information Processing Systems*, ed. H Wallach, H Larochelle, A Beygelzimer, F Alché-Buc, E Fox, R Garnett, pp. 13792–802. Red Hook, NY: Curran Assoc.

Visser PS, Krosnick JA, Lavrakas PJ. 2000. Survey research. In *Handbook of Research Methods in Social and Personality Psychology*, ed. HT Reis, CM Judd, pp. 223–52. Cambridge, UK: Cambridge Univ. Press

Wager S, Athey S. 2018. Estimation and inference of heterogeneous treatment effects using random forests. *J. Am. Stat. Assoc.* 113:1228–42

Warshaw C, Rodden J. 2012. How should we measure district-level public opinion on individual issues? *J. Politics* 74:203–19

Williams NW, Casas A, Wilkerson JD. 2020. *Images as Data for Social Science Research: An Introduction to Convolutional Neural Nets for Image Classification*. Cambridge, UK: Cambridge Univ. Press

Wolfson M, Madjd-Sadjadi Z, James P. 2004. Identifying national types: a cluster analysis of politics, economics, and conflict. *J. Peace Res.* 41:607–23

Ying L, Montgomery JM, Stewart BM. 2019. *Inferring concepts from topics: towards procedures for validating topics as measures*. Presented at the 36th Annual Meeting of the Society for Political Methodology (PolMeth XXXVI), July 18–20, Cambridge, MA

Annual Review of
Political Science

Volume 24, 2021

Contents

Errata

An online log of corrections to *Annual Review of Political Science* articles may be found at http://www.annualreviews.org/errata/polisci