Annual Review of Sociology

Machine Learning for Sociology

Mario Molina and Filiz Garip

Department of Sociology, Cornell University, Ithaca, New York 14853, USA;
email: mm2535@cornell.edu, fgarip@cornell.edu

Annu. Rev. Sociol. 2019. 45:27–45

First published as a Review in Advance on May 13, 2019

The *Annual Review of Sociology* is online at soc.annualreviews.org

https://doi.org/10.1146/annurev-soc-073117-041106

Keywords

supervised learning, unsupervised learning, causal inference, prediction, heterogeneity, discovery

Abstract

Machine learning is a field at the intersection of statistics and computer science that uses algorithms to extract information and knowledge from data. Its applications increasingly find their way into economics, political science, and sociology. We offer a brief introduction to this vast toolbox and illustrate its current uses in the social sciences, including distilling measures from new data sources, such as text and images; characterizing population heterogeneity; improving causal inference; and offering predictions to aid policy decisions and theory development. We argue that, in addition to serving similar purposes in sociology, machine learning tools can speak to long-standing questions on the limitations of the linear modeling framework, the criteria for evaluating empirical findings, transparency around the context of discovery, and the epistemological core of the discipline.

INTRODUCTION

Algorithm: a set of instructions telling a computer what to do

Supervised machine learning: methods that use training data of paired input (X) and output (Y) samples to learn parameters that predict Y from X in new data

Unsupervised machine learning: methods to summarize or characterize input data (X) without reference to a ground-truth output (Y)

Machine learning (ML) seeks to automate discovery from data. It represents a breakthrough in computer science, where past intelligent systems typically involved fixed algorithms (logical sets of instructions) that coded the desired output for all possible inputs. Now, intelligent systems learn from data and estimate complex functions that discover representations of some input (X), or link the input to an output (Y) in order to make predictions on new data (Jordan & Mitchell 2015). ML can be viewed as an offshoot of nonparametric statistics (Kleinberg et al. 2015).

We can classify ML tools by how they learn (extract information) from data. Different varieties of ML use different algorithms that invoke different assumptions about the principles underlying intelligence (Domingos 2015). We can also categorize ML tools by the kind of experience they are allowed to have during the learning process (Goodfellow et al. 2016), and we use this latter categorization here.

In supervised machine learning (SML), the algorithm observes an output (Y) for each input (X). That output gives the algorithm a target to predict and acts as a teacher. In unsupervised machine learning (UML), the algorithm only observes the input. It needs to make sense of the data without a teacher providing the correct answers—in fact, there are often no correct answers.[1]

We start with a brief (and somewhat technical) description of SML and UML and follow with examples of social science applications. We cannot give a comprehensive account, given the sprawl of the topic, but we hope to provide enough coverage to allow readers to follow up on different ideas. Our concluding remarks state why ML matters for sociology and how these tools can address some long-standing questions in the field.

SUPERVISED MACHINE LEARNING

SML involves searching for functions, $f(X)$, that predict an output (the dependent variable, Y) given an input (the explanatory or independent variable, X). In SML, the prediction task is called classification when the output is discrete, and regression when it is continuous. One can consider different classes of functions, such as linear models, decision trees, or neural networks. Let us take the linear model as a tool for prediction.[2] We have an input vector, X, and want to make a prediction on the output, Y, denoted as $\hat{Y}$, with the model

$$Y = f(X) = X^T\beta,$$

where X^T is the vector transpose and β is the vector of coefficients.

Suppose we use ordinary least squares (OLS)—the most commonly used method in sociology—to estimate the function, $f(X)$, from data. We pick the coefficients, β, that minimize the sum of squared residuals—one of many possible loss functions in ML—from data with n observations:

$$\sum_{i=1}^{n}[y_i - f(x_i)]^2. \qquad 1.$$

[1] Supervised and unsupervised learning are not formally defined terms (Goodfellow et al. 2016). Many ML algorithms can be used for both tasks. Scholars have proposed alternative labels, such as predictive and representation learning (Grosse 2013). There are other kinds of learning not captured with a binary categorization. In so-called reinforcement learning, the algorithm observes only some indication of the output (e.g., the end result of a chess game but not the rewards/costs associated with each move) (Jordan & Mitchell 2015).

[2] Uppercase letters (e.g., X or Y) denote variable vectors, and lowercase letters refer to observed values (e.g., x_i is the ith value of X).

CLASSICAL STATISTICS VERSUS MACHINE LEARNING

Breiman (2001b) describes two cultures of statistical analysis: data modeling and algorithmic modeling. Donoho (2017) updates the terms as generative modeling and predictive modeling. Classical statistics follows generative modeling. The central goal is inference, that is, to understand how an outcome is related to inputs. The analyst proposes a stochastic model that could have generated the data, and estimates the parameters of the model from the data. Generative modeling leads to simple and interpretable models but often ignores model uncertainty and out-of-sample performance. ML follows predictive modeling. The central goal is prediction, that is, to forecast the outcome for future inputs. The analyst treats the underlying generative model for the data as unknown and considers the predictive accuracy of alternative models on new data. Predictive modeling favors complex models that perform well out of sample, but can produce black-box results that offer little insight on the mechanism linking the inputs to the output.

This strategy ensures estimates of β that give the best fit in sample, but not necessarily the best predictions out of sample (i.e., on new data) (see sidebar titled Classical Statistics Versus Machine Learning).

To see that, consider the generalization error of the OLS model, that is, the expected prediction error on new data. This error comprises two components: bias and variance (Hastie et al. 2009). A model has bias if it produces estimates of the outcome that are consistently wrong in a particular direction (e.g., a clock that is always an hour late). A model has variance if its estimates deviate from the expected values across samples (e.g., a clock that alternates between fast and slow) (Domingos 2015). OLS minimizes in-sample error (Equation 1), but it can still have high generalization error if it yields high-variance estimates (Kleinberg et al. 2015).

To minimize generalization error, SML makes a trade-off between bias and variance—that is, unlike OLS, the methods allow for bias in order to reduce variance (Athey & Imbens 2017).[3] For example, an SML technique is to minimize

$$\sum_{i=1}^{n}[y_i - f(x_i)]^2 + \lambda R(f), \qquad 2.$$

that is, in-sample error plus a regularizer, $R(f)$, that penalizes functions that create variance (Kleinberg et al. 2015, Mullainathan & Spiess 2017). An important decision is to select λ, which sets the relative price for variance (Kleinberg et al. 2015). In OLS, that price is set to zero. In SML methods, the price is determined using the data (more on that later).

For example, in linear models, larger coefficients yield more variance in predictions. A popular SML technique called lasso (least absolute shrinkage and selection operator) introduces a regularizer,

$$R(f) = \sum_{j=1}^{p}|\beta_j|, \qquad 3.$$

that equals the sum of the absolute values of the coefficients, β_j $(j = 1, \ldots, p)$ (Tibshirani 1996). The optimal function, $f(X)$, now needs to select coefficients that minimize the sum of squared residuals while also yielding the smallest absolute coefficient sum.

Generalization error: the prediction error of a model on new data (also known as test error)

Model: a formal representation of our assumptions about the world

Regularizer: a term that penalizes estimation variance in out-of-sample predictions

[3] One can find a similar approach in multilevel models popular in sociology where cluster parameters are deliberately biased (Gelman & Hill 2007).

Training error: the error of a model on training data (e.g., sum of squared residuals)

Overfitting: the concept of a model fitting the data at hand well, but not generalizing to new data

Empirical tuning: using the data to optimize model design, including the choice of the regularization weight

SML techniques seek to achieve an ideal balance between reducing the in-sample and out-of-sample error (i.e., training and generalization error, respectively). This goal helps avoid two pitfalls of data analysis: underfitting and overfitting. Underfitting occurs when a model fits the data at hand poorly: As a simple example, an OLS model with only a linear term linking an input to output offers a poor fit if the true relationship is quadratic. Overfitting occurs when a model fits the data at hand too well and fails to predict the output for new inputs; for example, an OLS model with N inputs (plus a constant) will perfectly fit N data points, but it will likely not generalize well to new observations (Belloni et al. 2014).

Underfitting means we miss part of the signal in the data; we remain blind to some of its patterns. Overfitting means we capture not just the signal, but also the noise—the idiosyncratic factors that vary from sample to sample—and hallucinate patterns that are not there (Domingos 2015).

Through regularization, SML effectively searches for functions that are sufficiently complex to fit the underlying signal without fitting the noise. One way to regularize is to restrain model parameters. Let us consider lasso. The regularizer in Equation 3 puts a bound on the sum of absolute values of the coefficients. It can be shown that lasso favors sparse models, where a small number of inputs have nonzero coefficients, and effectively restrains model complexity (Tibshirani 1996).

Now consider regression trees, another function class in SML. The method proceeds by partitioning the inputs into separate regions in a tree-like structure and returning a separate output estimate ($\hat{Y}$) for each region. Say we want to predict whether someone migrates, using individual attributes of age and education. A tree might first split into two branches by age (young and old), and then each branch might split into two by education (college degree or not). Each terminal node (leaf) corresponds to a migration prediction (e.g., 1 for young college graduates). With enough splits in the tree, one can perfectly predict each observation in sample. To prevent overfitting, a typical regularizer controls the tree depth and, thus, makes us search not for the best fitting tree overall, but the best fitting tree among those of a certain depth (Mullainathan & Spiess 2017).

How do we select the model that offers the right compromise between in-sample and out-of-sample fit? To answer this question, we need to decide, first, on how to regularize [measure model variance/complexity, $R(f)$] and, second, on how much to regularize [set the price for variance/complexity, λ, in Equation 2].

In SML, we start the analysis by picking a function class and a regularizer. There are many function classes and many associated regularizers (see the sidebar titled Some Supervised Machine Learning Techniques). The so-called no free lunch theorem proves that no ML method (or no form of regularization) is universally better than any other (Wolpert & Macready 1997); the task is not to seek the best overall method, but the best method for the particular question at hand (Goodfellow et al. 2016, but see Domingos 2015 for a counterargument). The general recommendation is to use the substantive question at hand to guide these choices.[4] With the function class and regularizer in hand, we turn to data to choose the optimal model complexity. Put differently, in SML, we use the data not just to estimate the model parameters (e.g., coefficients, β, in lasso), but also for tuning regularization parameters (e.g., the price for variance/complexity, λ).

What sets SML apart from classical statistical estimation, then, are two essential features: regularization and the data-driven choice of regularization parameters (also known as empirical tuning) (Athey & Imbens 2017, Kleinberg et al. 2015, Mullainathan & Spiess 2017). These features allow

[4]Hastie et al. (2009, table 10.1) compare different methods on several criteria (e.g., interpretability, predictive power, ability to deal with different kinds of data). Athey & Imbens (2016), Athey (2017), and Abadie & Kasy (2017) link SML methods to traditional tools and questions in economics. Olson et al. (2018) offer an empirical comparison on bioinformatics data.

SOME SUPERVISED MACHINE LEARNING TECHNIQUES

- **Penalized regression:** a linear model of output (Y) as a function of inputs ($X^T\beta$). Regularizers include $\sum_{j=1}^{p} |\beta_j|$ for lasso, $\sum_{j=1}^{p} \beta_j^2$ for ridge regression, and $\alpha \sum_{j=1}^{p} |\beta_j| + (1-\alpha) \sum_{j=1}^{p} \beta_j^2$ for elastic net regression. Penalized regression shrinks coefficients toward zero; estimates need to be interpreted with caution (Athey & Imbens 2016).
- **Classification and regression trees:** a tree-like model that describes a sequence of splits in the input space (X) that predict an output (Y) at the end node. Regularizers include tree depth and number of leaves. This model captures nonlinearities and interactions in inputs. A version called random forests averages over multiple trees (Breiman 2001a), leading to more accurate predictions but less interpretable relationships of X to Y.
- **Nearest neighbor:** a method that relies on user-defined distances to average k nearest neighbors of a new input (X) to predict output (Y). The number of neighbors (k) is a regularizer. It offers black-box predictions with little insight into the relationship between X and Y.
- **Neural networks/deep learning:** a multilayer setup that models the output (Y) as a concatenation of simple nonlinear functions of the linear combinations of inputs (X) (neurons). Regularizers include the number of layers and the number of neurons per layer.

researchers to consider complex functions and more inputs (polynomial terms, high-order interactions, and, in some cases, more variables than observations) without overfitting the data. This flexibility contrasts sharply with classical statistics, where one typically selects a small number of inputs and a simple functional form to relate the inputs to the output.

One way SML uses data, therefore, is for model selection, that is, to estimate the performance of alternative models (functions, regularization parameters) to choose the best one. This process requires solving an optimization problem. Another way SML uses data is for model assessment, that is, after settling on a final model, estimating its generalization (prediction) error on new data (Hastie et al. 2009).

A crucial step in SML is to separate the data used for model selection from the data used for model assessment. In fact, in an idealized setup, one creates three, not two, separate data sets. Training data are used to fit the model; validation data are put aside to select among different models (or to select among the different parameterizations of the same model); and finally, test (or hold-out) data are kept in the vault to compute the generalization error of the selected model. There is no generic rule for determining the ideal partition, but typically, a researcher can reserve half of the data for training, and a quarter each for validation and testing (Hastie et al. 2009).

Splitting the data in this way comes at a cost, however. By reserving a validation and test set, we reduce the chance of overfitting but now run the risk of underfitting because there are fewer data left for estimation (Yarkoni & Westfall 2017). To achieve a middle ground, we can reserve the test data but combine training and validation sets into one, especially if the data are small. We can then recycle the training data for validation purposes (e.g., to select the optimal degree of complexity). One version of this process, called k-fold cross-validation, involves randomly splitting the data into k subsets (folds), and then successively fitting the data to $k-1$ of the folds and evaluating the model performance on the kth fold.

Training data: sample used to fit the model

Cross-validation: a method to estimate and validate a model on different partitions of data

Consider the regression-tree example above. We can divide the training data into $k = 5$ folds, use four of the folds to grow a tree with a particular depth (complexity), and then predict the output (migration) separately on the excluded fold, repeating for each of the five folds. We can then repeat the same process with a different tree depth and pick the complexity level that minimizes the average prediction error across the left-out folds (see Varma & Simon 2006 for more

sophisticated nested cross-validation). In the final step, we can use the test data to compute the predictive accuracy (generalization error) of the selected model.

Validation data: sample used to select among different models

Test data: Sample reserved to compute the generalization error of the selected model (hold–out data)

In SML, there are model-averaging techniques to improve predictive performance. For example, bagging involves averaging across models estimated on different bootstrap samples (where one draws N observations with replacement from a sample of size N). Boosting involves giving more weight to misclassified observations over repeated estimation (Hastie et al. 2009).

SUPERVISED MACHINE LEARNING FOR POLICY PREDICTIONS, CAUSAL INFERENCE, AND DATA AUGMENTATION

SML uses flexible functions of inputs to predict an output. Some SML tools, such as nearest neighbor, have no parameters at all. Other methods, such as lasso, give parameter estimates, $\hat{\beta}$, but those estimates are not always consistent (that is, they do not converge to the true value as N grows) (Knight & Fu 2000).

Social scientists are used to working with statistical models that produce parameter estimates with particular properties (unbiased and consistent). But SML is not designed for recovering $\hat{\beta}$. Instead, SML is good at solving what Mullainathan & Spiess (2017, p. 88) call $\hat{Y}$ tasks. Social scientists (mostly economists) have identified three classes of $\hat{Y}$ tasks: predictions for policy and theory development, certain procedures for causal inference, and data augmentation [for reviews, see Mullainathan & Spiess (2017) for predictive modeling in economics, Cranmer & Desmarais (2017) for political science, and Yarkoni & Westfall (2017) for psychology].

Predictions for Policy and Theory Development

SML is a useful tool for policy predictions if the researcher is not immediately interested in understanding the relationship between X and Y, but rather in using X to predict Y in new data. Policy predictions impose a clear goal ($\hat{Y}$) and performance metric (difference between Y and $\hat{Y}$), and they allow for a common-task framework where different teams can compete on the same question (Donoho 2017). For example, the company Kaggle hosts competitions (**http://www.kaggle.com/competitions**) where contestants train models on shared data and compete on predictive accuracy.

Economist Glaeser and colleagues (2016) used this idea to set up a competition to produce predictive algorithms for city governments. Salganik and collaborators started a challenge to predict educational (and other) outcomes in the Fragile Families and Child Wellbeing Study data (**http://www.fragilefamilieschallenge.org/**). The organizing team judged the submissions from 150 multidisciplinary teams on predictive accuracy on test (hold-out) data. In the ongoing second phase, the team plans to conduct an in-depth study of the discrepant cases in the winning model (e.g., students who beat the odds) and, thus, envisions the predictions as a first step to generating new insights and theory, not as an end goal.

Scholars apply SML to various questions in economics, demographics, political science, and criminology. Kleinberg et al. (2015) use a lasso model to predict which patients would benefit most from joint replacement surgery among Medicare beneficiaries. Billari et al. (2006) rely on decision trees to discriminate between Italians and Austrians in terms of the timing, sequencing, and quantum of life-course events. Cederman & Weidmann (2017) discuss how SML can predict deadly conflict. Beck et al. (2000) use neural networks to forecast militarized international disputes. Brandt et al. (2011) employ automated coding of news stories to predict Palestinian-Israeli conflicts, and Perry (2013) applies random forests to predict violent episodes in Africa. Berk (2012) reviews his extensive work that uses SML for predictions of criminal risk. These scholars use their predictions as a starting point for disentangling the process in question and for pushing existing theory.

Kleinberg et al. (2017), for example, illustrate how machine predictions can help us understand the process underlying judicial decisions. The authors first train a gradient-boosted decision tree model to predict judges' bail-or-release decisions in New York City, and then they use the quasi-random assignment of judges to cases to explain the sources of the discrepancy between model predictions and actual decisions. Their findings show that judges overweight the current charge, releasing high-risk cases if their present charge is minor and detaining low-risk ones if the present charge is serious. These findings reveal important insights on human decision-making and carry the potential to inspire new theory. From a policy standpoint, the authors' predictive model, if used in practice, promises significant welfare gains over human decisions without eroding important social values (e.g., racial equality): reducing the reoffending rate by 25% with no increase in jailing rate or, alternatively, pulling down the jailing rate by 42% with no increase in reoffending rate.

An important discussion in the literature is on how SML tools should weight different kinds of prediction errors. Berk et al. (2016), for example, apply random forests to forecast repeat offenses in domestic violence cases. In consultation with stakeholders, the authors weight false negatives (where the model predicts no repeat offense when there is one) 10 times more heavily than false positives (where the model predicts repeat offense when there is none). Their model, consequently, produces highly accurate predictions of no-offense cases (which require very strong evidence), but less accurate forecasts of repeat offenses (many of which do not end up occurring).

There are legitimate concerns that SML predictions (and the data on which they are based) can perpetuate social inequalities (Barocas & Selbst 2016, Harcourt 2007, Starr 2014). What if predicted offenders are disproportionately drawn from minority groups? What if predicted beneficiaries of health interventions are mostly high-status individuals?

Scholars now acknowledge an inherent trade-off between predictive accuracy and algorithmic fairness (Berk et al. 2018, Hardt et al. 2016, Kleinberg et al. 2016). An open question is how to define fairness. While most definitions relate to treatment of protected groups, one can operationalize fairness in many different ways (Berk et al. 2018, Narayanan 2018).

To see the complexity of the problem, consider a predictive algorithm that outputs loan decisions ($\hat{Y}$) from credit scores (X) (Hardt et al. 2016). Assume the algorithm produces more accurate predictions for men than women and recommends more loans to be given to men. One way to make the algorithm fair is to exclude applicants' gender from the data, but this solution fails if gender is correlated with another input, such as income. Another way is to seek demographic parity, that is, to constrain the model so that gender has no correlation with the loan decision. But this constraint might generate disparity in some other characteristic (Dwork et al. 2012). Yet another way to define fair is to impose equal opportunity (Hardt et al. 2016), that is, to force the model to make men and women equally likely to qualify for loans within a given subpopulation (e.g., individuals who pay back their loans).

Different definitions of fairness yield different outcomes. And it is difficult (if not impossible) to implement multiple definitions at the same time (Berk et al. 2018). Addressing algorithmic fairness is not just a technical issue in ML; it requires us—as a society—to consider difficult trade-offs.[5]

Causal Inference

Social scientists are often interested in identifying the causal effect of an input (treatment) on an output. SML tools can help in certain causal inference procedures that involve prediction tasks.

[5]Similar moral dilemmas abound in the use of ML in new technologies, such as self-driving cars (Greene 2016). Survey experiments show that while people agree that an algorithm should minimize casualties, they are not thrilled with the prospect of riding in a utilitarian car that can sacrifice its driver for the greater good (Bonnefon et al. 2016).

We provide some basic intuition and examples from this rather technical literature and refer the readers to Athey & Imbens (2017) and Mullainathan & Spiess (2017) for comprehensive reviews, and to Pearl & Mackenzie (2018) and Peters et al. (2017) for general frameworks that link ML to causality.

As a primer, consider the fundamental problem of causal inference: We observe an individual (or any unit of analysis) in one condition alone (treatment or control) and cannot measure individual-level variation in the effect of the treatment (for authoritative reviews of causal inference in the social sciences, see Morgan & Winship 2007, 2014). We instead focus on an aggregate average effect that we treat as homogeneous across the population (Xie 2013). In experimental design, we randomly assign individuals to treatment and control groups and directly estimate the average causal effect by comparing the mean output between the groups (Imbens & Rubin 2015).

Social scientists now use SML to identify heterogeneous treatment effects in subpopulations in existing experimental data. For example, Imai & Ratkovic (2013) discover groups of workers differentially affected by a job training program. They interact the treatment (i.e., being in the program) with different inputs and use a lasso model to select the inputs that are most important in predicting increases in worker earnings. Similarly, Athey & Imbens (2016) develop causal trees to estimate treatment effects for subgroups. Different from standard regression trees in ML (where one seeks to minimize the error in predictions, $\hat{Y}$), causal trees focus on minimizing the error in treatment effects. One can then obtain valid inference for each leaf (subgroup) with honest estimation, that is, by using half the sample to build the tree (select the optimal partition of inputs), and the other half to estimate the treatment effects within the leaves. Wager & Athey (2018) extend the method to random forests that average across many causal trees and allow for personalized treatment effects (where each individual observation gets a distinct estimate). Similarly, Grimmer et al. (2017) propose ensemble methods that weight several ML models and discover heterogeneous treatment effects in data from two existing political science experiments.

Most empirical work in sociology relies on observational data where we do not control assignment to treatment. One way to estimate the causal effect in this case is to assume the potential output to be independent of assignment to treatment, conditional on other observed inputs. Under this so-called selection-on-observables assumption, we can estimate a causal effect by matching treatment and control groups on their propensity score (that is, the likelihood of being in the treatment group conditional on inputs). Estimation of this score is well suited to SML as it involves a prediction task (where the effects of inputs are not of interest). Recent work uses boosting (McCaffrey et al. 2004), neural networks (Setoguchi et al. 2008, Westreich et al. 2010), and regression trees for this task (Diamond & Sekhon 2013, Hill 2011, Lee et al. 2010, Wyss et al. 2014) as alternatives to traditional logistic regression.

In some cases, the selection-on-observables assumption does not hold, and we suspect that some unobserved inputs are correlated with both the assignment to treatment and the output, creating omitted variable bias in estimation. Regularization in SML could lead to exclusion of some inputs from the model, amplifying this bias. Similarly, with many inputs, one generally runs the risk of model misspecification (Belloni et al. 2014, Ho et al. 2007, King & Nielsen 2019, Muñoz & Young 2018, Raftery 1995, Young & Holsteen 2017). Athey & Imbens (2015) develop a measure of sensitivity to misspecification. Belloni et al. (2017) propose double selection of inputs to address potential omitted variable bias. This procedure involves solving two prediction tasks to determine, first, the inputs correlated with the treatment and, second, those correlated with the output. The union of these two sets of inputs enters an OLS regression of the output, leading to parameter estimates with improved properties (Belloni et al. 2014, 2017; Chernozhukov et al. 2017).

Another way to address the omitted variable bias is to find an instrument—an input that is correlated with the output only through its correlation with assignment to treatment (Angrist et al. 1996). We can then regress the treatment (a given input, X) on the instrument (Z), and then use the predicted values ($\hat{Z}$) as an input in the output (Y) regression. Because the first stage in this instrumental variables regression involves a prediction task, we can use SML tools. There are now many examples of this application in the econometrics literature. Belloni et al. (2012) use lasso to produce first-stage predictions in data with many potential instruments, while Carrasco (2012) and Hartford et al. (2016) turn to ridge regression and neural networks, respectively.

Data Augmentation and Imputation

Scholars use SML for data linking and augmentation.[6] Feigenbaum (2015), for example, inputs human-coded data to train SML algorithms to link individuals across census waves. Abramitzky et al. (2019) develop a fully automated method to estimate probabilities of matches across census waves, and then measure intergenerational occupational mobility. Using a nested design, Bernheim et al. (2013) recruit a subset of survey respondents to participate in a lab experiment and use their responses in the lab as training data to impute responses for the remaining sample. Blumenstock et al. (2015) collect survey responses from a subset of cell phone users in Rwanda as training data to predict the wealth and well-being of one million phone users.

Scholars are similarly turning to supervised topic modeling (Blei & McAuliffe 2010) to use human-identified topics as training data to classify a larger set of documents (Hopkins & King 2010, Mohr et al. 2013). For instance, Chong et al. (2009) apply this approach successfully to predict topics for image labels and annotations.

Researchers are also using SML for missing data imputation. Farhangfar et al. (2008) investigate the performance of different ML classifiers in fifteen data sets and find that, although no method is universally best, naive-Bayes and support vector machine classifiers perform particularly well in imputing missing values. More recently, Sovilj et al. (2016) use Gaussian mixture models to estimate the underlying distribution of data and an extreme learning machine (a type of one-layer neural network) for data imputation. Their approach, evaluated in six different data sets, yields more accurate values compared with conditional mean imputation.

UNSUPERVISED MACHINE LEARNING

UML searches for a representation of the inputs that is more useful than X itself (Goodfellow et al. 2016). Some UML tools reduce the dimensionality of the data (e.g., principal component analysis, factor analysis, topic modeling). Other methods partition the data into groups (e.g., cluster analysis, latent class analysis, sequence analysis, community detection) (see the sidebar titled Some Unsupervised Machine Learning Techniques).[7] There is no target output to predict, no teacher showing the algorithm what it should aim for, and no immediate measure of success. Researchers use heuristic tools to evaluate the results.

[6] In the ML community, researchers use the term data augmentation to also refer to the technique of artificially increasing your training data in order to improve the predictive performance of ML classifiers. This strategy is widely used in deep neural networks for image recognition (e.g., Wong et al. 2016) but remains outside the scope of our review.

[7] There are excellent reviews of latent class analysis (Bollen 2002), sequence analysis (Abbott & Tsay 2000, Cornwell 2015), and community detection (Fortunato 2010, Fortunato & Hric 2016, Watts 2004).

SOME UNSUPERVISED MACHINE LEARNING TECHNIQUES

- **Principal component analysis:** discovers a small number of linear combinations of the inputs that are uncorrelated with one another and capture most of the variability in the data. These linear combinations (principal components) can be used as inputs in subsequent analysis (e.g., in regression to predict some output)
- **Factor analysis:** discovers latent (unobserved) factors that account for the correlation in inputs; returns factor loadings for each input that can be used to interpret the factors
- **Cluster analysis:** groups observations into a given number of clusters so that observations in a cluster are more similar to one another than to observations in other clusters; returns cluster membership for each observation
- **Latent class analysis:** discovers latent classes of observations that can account for the correlations in observed categorical inputs; returns probability of class membership for each observation
- **Sequence analysis:** compares sequences (ordered elements or events) with optimal matching to discover groups of observations with similar patterns (typically with cluster analysis)
- **Topic modeling:** discovers latent topics in text data based on co-occurrence of words across documents
- **Community detection:** identifies communities in networks (graphs) based on structural position of nodes

Unsupervised Machine Learning for Measurement and Discovery

Social scientists can use UML for measurement and discovery. The output from UML (data partitioned or projected onto a lower dimension) typically becomes an input that allows subsequent analysis or theorizing. In the absence of a ground truth, researchers need to pay particular attention to model checking, and validate their results using statistical, substantive, or external criteria.

Generating measures from complex data. UML can produce measures from data to be used in subsequent statistical analysis. Sociologists have long used principal components and factor analysis to reduce many inputs into a smaller set. Social scientists now use UML to process new kinds of data (images or text). Economists, for example, classify satellite images with UML to generate measures (deforestation, pollution, night lights, and so on) that relate to economic outputs (see Donaldson & Storeygard 2016 for a review). Sociologists categorize text to develop proxies for discourse in the media (DiMaggio et al. 2013), state documents (Mohr et al. 2013) and academic publications (McFarland et al. 2013). For more information about text analysis, readers are directed to articles by Bail (2014), Blei (2012), Evans & Aceves (2016), Grimmer & Stewart (2013), and Mohr & Bogdanov (2013).

Following a long tradition, sociologists also use UML to group social network data. Earlier applications, such as block models, employed structural equivalence (sharing neighbors) to evaluate similarity and to then partition the network into subgroups (Breiger et al. 1975, White et al. 1976). Recent improvements involve using centrality (instead of equivalence) measures to discover communities (Girvan & Newman 2002), assuming generative probabilistic distributions (Nowicki & Snijders 2001) that help in model selection (Handcock et al. 2007), allowing for mixed membership in communities (Airoldi et al. 2008), and considering temporal dynamics (Matias & Miele 2017, Xing et al. 2010, Yang et al. 2011) and latent social structure (Hoff et al. 2002).

Characterizing population heterogeneity. UML can help characterize population heterogeneity. For example, Bail (2008) applies fuzzy cluster analysis (which allows cases to belong to multiple groups) to discover three configurations of symbolic boundaries between immigrants and natives

in Europe. Bonikowski & DiMaggio (2016) employ latent class analysis to characterize four types of popular nationalism in the United States. Frye & Trinitapoli (2015) use sequence analysis to discover five distinct event sequences that characterize discrepancies between women's ideal and experienced prelude to sex in Malawi. Killewald & Zhuo (2018) employ the same method to identify four maternal employment patterns of American mothers. Garip (2012, 2016) uses cluster analysis to identify four distinct groups among first-time Mexico-US migrants. Goldberg (2011) develops relational class analysis that considers associations between individuals' survey responses (rather than responses themselves) to discover three separate logics of cultural distinction around musical tastes. Baldassarri & Goldberg (2014) apply the same tool to identify three configurations of political beliefs among Americans.

These examples use a variety of methods, but share a common goal. They search for the hidden structure in a population that would be presumed homogeneous under the traditional statistical approach (Xie 2007, 2013; Duncan 1982). This approach often yields new hypotheses that emerge from data.

Model checking. Unlike prediction problems, there is often no ground truth in UML; therefore, model checking is an important step. Researchers use statistical validation techniques that involve some heuristic measure to capture whether, for example, clusters (Garip 2012, Killewald & Zhuo 2018), latent classes (Bonikowski & DiMaggio 2016), or topics (DiMaggio et al. 2013) are well separated. Scholars employ substantive validation to see if the produced partitions cohere with existing typologies or, more generally, with human judgement. Grimmer & King (2011) offer a method for computer-assisted clustering. The method allows researchers to explore and select from thousands of partitions produced by different clustering methods and, thus, puts their domain knowledge at the center (Grimmer & Stewart 2013).

Researchers also resort to external validation that brings new data to evaluate whether identified patterns confirm expectations. Bail (2008), for example, shows that three types of symbolic boundaries emerging from attitudinal data are associated with country-level immigration patterns and integration philosophies in Europe. Bonikowski & DiMaggio (2016) find that four varieties of nationalism in the United States correlate with social and policy attitudes that were not used in the identification of the typology. DiMaggio et al. (2013) check that topics identified in the news coverage of government assistance to the arts respond to other news events in hypothesized ways. Garip (2016) confirms that four migrant types, obtained by clustering survey responses alone, relate differently to macrolevel economic and political indicators.

MACHINE LEARNING: NEW ANSWERS TO OLD QUESTIONS

While many of the ML tools are quite new to sociology, the problems they address are not. Below we discuss how ML can speak to some long-standing concerns in our field, and point to promising directions for future research.

Supervised Machine Learning Helps Us Break Away from General Linear Reality

In quantitative sociology, we often follow the classical statistics approach: Assume a distribution of the data, select a few inputs, and specify a parametric (typically linear) model to relate the inputs to an output (Breiman 2001a, Donoho 2017). We tend to favor models that seem to align with common sense (Watts 2014). We consider some alternative specifications (for example, nested models that gradually introduce controls) but do not exhaust all possibilities (Varian 2014) or fully take into account model uncertainty (Western 1996, Young 2009).

SML allows us to include many inputs (including higher-order terms and interactions) and complex functions that connect inputs to the output. It helps break away from the linear model imposed by OLS (Abbott 2001). It helps us avoid underfitting (missing part of the signal) and mine the data effectively without overfitting (capturing the noise as well as the signal). This gain comes at a cost. Predictive tools in SML typically do not yield reliable estimates of the effects of particular inputs ($\hat{\beta}$), and indeed, some methods only produce black-box results.

Sociologists can identify pure prediction ($\hat{Y}$) problems where different research teams can potentially compete in a common-task framework (Donoho 2017). Economists, for example, are already using SML to make policy predictions (Kleinberg et al. 2015). Sociologists can further use predictions as a starting point to understand underlying social process and to develop theory. Sociologists can also use their expertise in processes of stratification to inform debates on the ethics of predictive modeling, and its fairness to different social groups (Berk et al. 2018).

Another direction for sociologists is to use SML to improve classical statistical techniques. Economists now apply SML to prediction tasks within the causal-inference framework, for example, estimation of the propensity score in matching (Westreich et al. 2010) or the first-stage equation in instrumental variables (Belloni et al. 2012), and identification of heterogeneous treatment effects in existing experimental data (Athey & Imbens 2016). One particularly fruitful application (and one that is highly relevant to sociologists given our typical attention to omitted variable bias) involves using SML for model selection (Belloni et al. 2014, 2017).

Machine Learning Allows Us to Study Population Heterogeneity

Quantitative sociology often takes a deductive approach, where the researcher derives hypotheses from a theory to test on data. This approach, inspired by classical physics, can act as a straitjacket that limits the questions we can ask and the methods we can use (Lieberson & Lynn 2002).

To fit our work into the mold of hypothesis testing, we flatten social theories into a few variables and estimate the average effect of each variable in some given population. We neglect that most theories offer sometimes-true statements (Coleman 1964) that hold under specific conditions and for specific groups of individuals. We also pit multiple theories against one another to determine the best fit empirically. We ignore the possibility that different mechanisms might be simultaneously at work [what Goldberg (2011) calls equifinality or what Watts (2014) refers to as the indeterminacy problem]. We rule out heterogeneity in explanation a priori.

It is these concerns about causal complexity that have led Ragin (1987) to develop a toolbox (qualitative comparative analysis) to identify different causal bundles (configurations of various conditions) that underlie some historical phenomenon, and Abbott (1995) to advocate for sequence analysis as a way to characterize configurations of events that inform social outcomes.

ML offers new tools to characterize population heterogeneity. Economists use SML to uncover heterogeneous treatment effects in experimental data (Athey & Imbens 2016). Sociologists use UML to discover subgroups in populations and then link the emergence of each subgroup to various external factors (Bail 2008, Bonikowski & DiMaggio 2016, Garip 2012). This latter approach is akin to searching for ideal types (Weber 1978) as a first step to developing theory (Swedberg 2014). Indeed, Muller et al. (2016) and Baumer et al. (2017) make an insightful connection of ML to inductive reasoning in the social sciences (and the grounded theory approach in particular).

By expanding their tool kit to include ML, sociologists can better consider heterogeneity and close the gap between their pluralistic stance when it comes to embracing different theories and their monism when it comes to testing those theories with data.

Supervised Machine Learning Makes Us Sensitive to Researcher Degrees of Freedom and Replication

In sociology, we commonly estimate and evaluate a model on the same sample and run the risk of overfitting. SML gives us the crucial idea that we need to validate our results on new data (or with efficient partitioning of the original data, also known as cross-validation).

When we test a model out of sample, not only do we minimize the risk of overfitting (to which models with low R^2—share of explained variation—are especially vulnerable), but also we evaluate the overall performance of a model in explaining an output. We get more information on the strength of underlying theory, in other words, than is typically available with in-sample estimates (e.g., coefficients in an OLS model) (Watts 2014).

Out-of-sample testing can also help address what Yarkoni & Westfall (2017) call procedural overfitting (also known as *p*-hacking) that can occur during data cleaning or model selection. There are many choices available to us (researcher degrees of freedom) that might influence the results (Simmons et al. 2011, King & Nielsen 2019).[8] Any time we use the data to optimize over these degrees of freedom (for example, choosing variables that give the best fit), we need to conduct an out-of-sample test (or cross-validation) to evaluate the true performance of our choices. A related activity at the research community level is to encourage independent replication studies, which would serve as out-of-sample tests (Freese 2007).

Machine Learning Offers Tools for Exploration and Discovery

In quantitative sociology, we mostly engage in exploratory work, but couch it in the language of hypothesis testing. We often use flexible research designs and statistical models until we learn something new and interesting, but present our results as if we were confirming a hypothesis that we knew all along. We give our readers the context of justification, but not the context of discovery (Popper 1935). This practice makes it difficult to teach our students research design or encourage creative theorizing (Swedberg 2014).

ML gives us a vast array of tools to explore and learn from data, but for these tools to be useful in sociology, we first need to distinguish exploratory work from confirmatory research. Conducting confirmatory research requires minimizing researcher degrees of freedom, ideally by preregistering hypotheses and other design choices in a public forum (e.g., the Open Science Framework) (Baldassarri & Abascal 2017, Hofman et al. 2017, Ioannidis & Doucouliagos 2013, Simmons et al. 2011, Watts 2014).

Many of us do not conduct confirmatory work in this strict sense. Instead we go back and forth between data, statistical models, and theory until we gain a novel insight. Presenting such efforts as exploratory allows us to truthfully describe where our ideas come from. It frees us to use ML (and other) tools for discovery and creative conceptualization. It helps us generate novel hypotheses for subsequent confirmatory work. Recognition of exploratory work, however, requires support from journals and an expansion of scientific values.

Machine Learning Provides a Diverse Set of Tools That Can Inform a Diverse Set of Questions

In sociology, we rely largely on a hypothesis testing framework and classical statistical approach. We routinely fit our questions to this setup and use data to estimate the effects of some input on

[8]This issue has led to heated debates in psychology where researchers have been unable to replicate some well-known experimental findings (Simmons et al. 2011, Open Sci. Collab. 2015).

an output. ML not only helps us improve parts of this strategy, but also gives us tools that can inspire new questions. How well does a set of inputs, for example, predict an output? How do these predictions deviate from observed outcomes and why? Or what is the underlying structure of some input? How is that structure related to external factors? Answering these questions can help us push theory forward or generate new hypotheses. Indeed, in some of the best social science applications, the results from ML provide not an end goal, but the starting point for further analysis and conceptualization. As such, ML tools complement, but do not replace, existing methods in sociology.

SUMMARY POINTS

1. Classical statistics focuses on inference (estimating parameters, β, that link the output, Y, to inputs, X); supervised machine learning (SML) aims at prediction (using inputs X to forecast unobserved output $\hat{Y}$).
2. SML balances in-sample and out-of-sample fit through regularization (i.e., penalizing model complexity and estimation variance) and empirical tuning (i.e., data-driven choice) of regularization parameters.
3. Unsupervised machine learning (UML) discovers underlying structure in data (e.g., principal components, clusters, latent classes) that needs to be validated with statistical, substantive, or external evidence.
4. Sociologists can apply SML to predict outputs, to use the predictions as a starting point to understand underlying social process, or to improve classical statistical techniques.
5. Sociologists can use UML to describe and classify inputs, and to conceptualize on the basis of the descriptions.

FUTURE ISSUES

1. What are the prediction ($\hat{Y}$) questions in sociology?
2. What can the deviations from predictions reveal about the underlying social process?
3. What are the criteria for evaluating predictive fairness?
4. How can we use predictions given by SML or descriptions produced by UML to theorize?
5. How can we validate the findings of ML applications?

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 offer our apologies to scholars whose work could not be appropriately cited due to space constraints. We extend our thanks to Thomas Davidson, Joscha Legewie, Karen Levy, Samir Passi,

Mert Sabuncu, Florencia Torche, and Cristobal Young for their thoughtful feedback on our earlier drafts. We also thank an anonymous reviewer and the editors. All errors are our own.

LITERATURE CITED

Abadie A, Kasy M. 2017. The risk of machine learning. arXiv:1703.10935 [stat.ML]

Abbott A. 1995. Sequence analysis: new methods for old ideas. *Annu. Rev. Sociol.* 21:93–113

Abbott A. 2001. *Time Matters: On Theory and Method*. Chicago: Univ. Chicago Press

Abbott A, Tsay A. 2000. Sequence analysis and optimal matching methods in sociology. *Sociol. Methods Res.* 29:3–33

Abramitzky R, Mill R, Perez S. 2019. Linking individuals across historical sources: a fully automated approach. *Hist. Methods*. In press. **https://doi.org/10.1080/01615440.2018.1543034**

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

Angrist JD, Imbens GW, Rubin DB. 1996. Identification of causal effects using instrumental variables. *J. Am. Stat. Assoc.* 91:444

Athey S. 2017. Beyond prediction: using big data for policy problems. *Science* 355:483–85

Athey S, Imbens G. 2015. A measure of robustness to misspecification. *Am. Econ. Rev.* 105:476–80

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

Athey S, Imbens GW. 2017. The state of applied econometrics: causality and policy evaluation. *J. Econ. Perspect.* 31:3–32

Bail CA. 2008. The configuration of symbolic boundaries against immigrants in Europe. *Am. Sociol. Rev.* 73:37–59

Bail CA. 2014. The cultural environment: measuring culture with big data. *Theor. Soc.* 43:465–82

Baldassarri D, Abascal M. 2017. Field experiments across the social sciences. *Annu. Rev. Sociol.* 43:41–73

Baldassarri D, Goldberg A. 2014. Neither ideologues nor agnostics: alternative voters' belief system in an age of partisan politics. *Am. J. Sociol.* 120:45–95

Barocas S, Selbst A. 2016. Big data's disparate impact. *Calif. Law Rev.* 104:671–732

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

Beck N, King G, Zeng L. 2000. Improving quantitative studies of international conflict: a conjecture. *Am. Political Sci. Rev.* 94:21–35

Belloni A, Chen D, Chernozhukov V, Hanse C. 2012. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica* 80:2369–429

Belloni A, Chernozhukov V, Fernandez-Val I, Hansen C. 2017. Program evaluation and causal inference with high-dimensional data. *Econometrica* 85:233–98

Belloni A, Chernozhukov V, Hansen C. 2014. Inference on treatment effects after selection among high-dimensional controls. *Rev. Econ. Stud.* 81:608–50

Berk R. 2012. *Criminal Justice Forecasts of Risk*. New York: Springer

Berk R, Heidari H, Jabbari S, Kearns M, Roth A. 2018. Fairness in criminal justice risk assessments: the state of the art. *Sociol. Methods Res.* **https://doi.org/10.1177/0049124118782533**

Berk RA, Sorenson SB, Barnes G. 2016. Forecasting domestic violence: a machine learning approach to help inform arraignment decisions. *J. Empir. Legal Stud.* 13:94–115

Bernheim BD, Bjorkegren D, Naecker J, Rangel A. 2013. *Non-choice evaluations predict behavioral responses to changes in economic conditions.* NBER Work. Pap. 19269

Billari FC, Fürnkranz J, Prskawetz A. 2006. Timing, sequencing, and quantum of life course events: a machine learning approach. *Eur. J. Popul.* 22:37–65

Blei DM. 2012. Probabilistic topic models. *Commun. ACM* 55:77–84

Blei DM, McAuliffe JD. 2010. Supervised topic models. arXiv:1003.0783 [stat.ML]

Blumenstock J, Cadamuro G, On R. 2015. Predicting poverty and wealth from mobile phone metadata. *Science* 350:1073–76

Bollen KA. 2002. Latent variables in psychology and the social sciences. *Annu. Rev. Psychol.* 53:605–34

Overview of applied econometrics and the place of machine learning tools in the field.

Consideration of omitted variable bias in ML.

Bonikowski B, DiMaggio P. 2016. Varieties of American popular nationalism. *Am. Sociol. Rev.* 81:949–80
Bonnefon JF, Shariff A, Rahwan I. 2016. The social dilemma of autonomous vehicles. *Science* 352:1573–6
Brandt PT, Freeman JR, Schrodt PA. 2011. Real time, time series forecasting of inter- and intra-state political conflict. *Conflict Manag. Peace* 28:41–64
Breiger RL, Boorman SA, Arabie P. 1975. An algorithm for clustering relational data with applications to social network analysis and comparison with multidimensional scaling. *J. Math. Psychol.* 12:328–83
Breiman L. 2001a. Random forests. *Mach. Learn.* 45:5–32
Breiman L. 2001b. Statistical modeling: the two cultures (with comments and a rejoinder by the author). *Stat. Sci.* 16:199–231
Carrasco M. 2012. A regularization approach to the many instruments problem. *J. Econom.* 170:383–98
Cederman LE, Weidmann NB. 2017. Predicting armed conflict: Time to adjust our expectations? *Science* 355:474–76
Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W. 2017. Double/debiased/Neyman machine learning of treatment effects. *Am. Econ. Rev.* 107:261–65
Chong W, Blei D, Li FF. 2009. Simultaneous image classification and annotation. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1903–10. New York: IEEE
Coleman J. 1964. *Introduction to Mathematical Sociology*. New York: Free Press
Cornwell B. 2015. *Social Sequence Analysis: Methods and Applications*. Cambridge, UK: Cambridge Univ. Press
Cranmer SJ, Desmarais BA. 2017. What can we learn from predictive modeling? *Political Anal.* 25:145–66
Diamond A, Sekhon JS. 2013. Genetic matching for estimating causal effects: a general multivariate matching method for achieving balance in observational studies. *Rev. Econ. Stat.* 95:932–45
DiMaggio P, Nag M, Blei D. 2013. Exploiting affinities between topic modeling and the sociological perspective on culture: application to newspaper coverage of U.S. government arts funding. *Poetics* 41:570–606
Domingos P. 2015. *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*. New York: Basic Books
Donaldson D, Storeygard A. 2016. The view from above: applications of satellite data in economics. *J. Econ. Perspect.* 30(4):171–98
Donoho D. 2017. 50 years of data science. *J. Comput. Graph. Stat.* 26:745–66
Duncan OD. 1982. *Rasch measurement and sociological theory*. Hollingshead Lecture presented at Yale University, New Haven, CT, April 20
Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ed. S Goldwasser, pp. 214–26. New York: ACM
Evans JA, Aceves P. 2016. Machine translation: mining text for social theory. *Annu. Rev. Sociol.* 42:21–50
Farhangfar A, Kurgan L, Dy J. 2008. Impact of imputation of missing values on classification error for discrete data. *Pattern Recognit.* 41:3692–705
Feigenbaum JJ. 2015. *Automated census record linking: a machine learning approach*. Work. Pap., Harvard Univ., Cambridge, MA. **https://scholar.harvard.edu/jfeigenbaum/publications/automated-census-record-linking**
Fortunato S. 2010. Community detection in graphs. *Phys. Rep.* 486:75–174
Fortunato S, Hric D. 2016. Community detection in networks: a user guide. *Phys. Rep.* 659:1–44
Freese J. 2007. Replication standards for quantitative social science. *Sociol. Methods Res.* 36:153–72
Frye M, Trinitapoli J. 2015. Ideals as anchors for relationship experiences. *Am. Sociol. Rev.* 80:496–525
Garip F. 2012. Discovering diverse mechanisms of migration: the Mexico–US stream. 1970–2000. *Popul. Dev. Rev.* 38:393–433
Garip F. 2016. *On the Move: Changing Mechanisms of Mexico–U.S. Migration*. Princeton, NJ: Princeton Univ. Press
Gelman A, Hill J. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge, UK: Cambridge Univ. Press
Girvan M, Newman MEJ. 2002. Community structure in social and biological networks. *PNAS* 99:7821–6
Glaeser EL, Hillis A, Kominers SD, Luca M. 2016. Crowdsourcing city government: using tournaments to improve inspection accuracy. *Am. Econ. Rev.* 106:114–18

Goldberg A. 2011. Mapping shared understandings using relational class analysis: the case of the cultural omnivore reexamined. *Am. J. Sociol.* 116:1397–436

Goodfellow I, Bengio Y, Courville A. 2016. *Deep Learning*. Cambridge, MA: MIT Press

Greene JD. 2016. Our driverless dilemma. *Science* 352:1514–5

Grimmer J, King G. 2011. General purpose computer-assisted clustering and conceptualization. *PNAS* 108:2643–50

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, Stewart BM. 2013. Text as data: the promise and pitfalls of automatic content analysis methods for political texts. *Political Anal.* 21:267–97

Grosse R. 2013. Predictive learning vs. representation learning. *Laboratory for Intelligent Probabilistic Systems Blog*, Feb. 4. **https://lips.cs.princeton.edu/predictive-learning-vs-representation-learning/**

Handcock MS, Raftery AE, Tantrum JM. 2007. Model-based clustering for social networks. *J. R. Stat. Soc.* 170:301–54

Harcourt BE. 2007. *Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age*. Chicago: Univ. Chicago Press

Hardt M, Price E, Srebro N. 2016. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, ed. DD Lee, U von Luxburg, R Garnett, M Sugiyama, I Guyon, pp. 3323–31. Red Hook, NY: Curran

Hartford J, Lewis G, Leyton-Brown K, Taddy M. 2016. Counterfactual prediction with deep instrumental variables networks. arXiv:1612.09596 [stat.AP]

Hastie T, Tibshirani R, Friedman J. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer. 2nd ed.

Hill JL. 2011. Bayesian nonparametric modeling for causal inference. *J. Comput. Graph. Stat.* 20:217–40

Ho DE, Imai K, King G, Stuart EA. 2007. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Anal.* 15:199–236

Hoff PD, Raftery AE, Handcock MS. 2002. Latent space approaches to social network analysis. *J. Am. Stat. Assoc.* 97:1090–98

Hofman JM, Sharma A, Watts DJ. 2017. Prediction and explanation in social systems. *Science* 355:486–88

Hopkins DJ, King G. 2010. A method of automated nonparametric content analysis for social science. *Am. J. Political Sci.* 54:229–47

Imai K, Ratkovic M. 2013. Estimating treatment effect heterogeneity in randomized program evaluation. *Ann. Appl. Stat.* 7:443–70

Imbens GW, Rubin DB. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge, UK: Cambridge Univ. Press

Ioannidis J, Doucouliagos C. 2013. What's to know about the credibility of empirical economics? *J. Econ. Surv.* 27:997–1004

Jordan MI, Mitchell TM. 2015. Machine learning: trends, perspectives, and prospects. *Science* 349:255–60

Killewald A, Zhuo X. 2018. U.S. mothers' long-term employment patterns. *Demography* 56:285–320

King G, Nielsen R. 2019. Why propensity scores should not be used for matching. *Political Anal.* In press

Kleinberg J, Liang A, Mullainathan S. 2017. The theory is predictive, but is it complete? An application to human perception of randomness. arXiv:1706.06974 [cs.LG]

Kleinberg J, Ludwig J, Mullainathan S, Obermeyer Z. 2015. Prediction policy problems. *Am. Econ. Rev.* 105:491–95

Kleinberg J, Mullainathan S, Raghavan M. 2016. Inherent trade-offs in the fair determination of risk scores. arXiv:1609.05807 [cs.LG]

Knight K, Fu W. 2000. Asymptotics for lasso-type estimators. *Ann. Stat.* 28:1356–78

Lee BK, Lessler J, Stuart EA. 2010. Improving propensity score weighting using machine learning. *Stat. Med.* 29:337–46

Lieberson S, Lynn FB. 2002. Barking up the wrong branch: scientific alternatives to the current model of sociological science. *Annu. Rev. Sociol.* 28:1–19

Basic introduction to machine learning (with emphasis on deep learning toolbox).

Overview of applied econometrics, the place of machine learning tools in the field.

Explains importance of prediction in social sciences.

Uses machine learning predictions to understand human decision-making.

Matias C, Miele V. 2017. Statistical clustering of temporal networks through a dynamic stochastic block model. *Stat. Methodol.* 79:1119–41
McCaffrey DF, Ridgeway G, Morral AR. 2004. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychol. Methods* 9(4):403–25
McFarland DA, Ramage D, Chuang J, Heer J, Manning CD, Jurafsky D. 2013. Differentiating language usage through topic models. *Poetics* 41:607–25
Mohr JW, Bogdanov P. 2013. Introduction—topic models: what they are and why they matter. *Poetics* 41:545–69
Mohr JW, Wagner-Pacifici R, Breiger RL, Bogdanov P. 2013. Graphing the grammar of motives in national security strategies: cultural interpretation, automated text analysis and the drama of global politics. *Poetics* 41:670–700
Morgan SL, Winship C. 2007. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge, UK: Cambridge Univ. Press. 1st ed.
Morgan SL, Winship C. 2014. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge, UK: Cambridge Univ. Press. 2nd ed.

Introduction to machine learning methods to economists in relation to econometric toolbox.

Mullainathan S, Spiess J. 2017. Machine learning: an applied econometric approach. *J. Econ. Perspect.* 31:87–106
Muller M, Guha S, Baumer EP, Mimno D, Shami NS. 2016. Machine learning and grounded theory method: convergence, divergence, and combination. In *Proceedings of the 19th International Conference on Supporting Group Work*, pp. 3–8. New York: ACM
Muñoz J, Young C. 2018. We ran 9 billion regressions: eliminating false positives through computational model robustness. *Sociol. Methodol.* 48:1–33
Narayanan A. 2018. Tutorial: 21 fairness definitions and their politics. *YouTube*. **https://www.youtube.com/watch?v=jIXIuYdnyyk**
Nowicki K, Snijders TAB. 2001. Estimation and prediction for stochastic blockstructures. *J. Am. Stat. Assoc.* 96:1077–87
Olson RS, La Cava W, Mustahsan Z, Varik A, Moore JH. 2018. Data-driven advice for applying machine learning to bioinformatics problems. *Pac. Symp. Biocomput.* 23:192–203
Open Sci. Collab. 2015. Estimating the reproducibility of psychological science. *Science* 349:aac4716
Pearl J, Mackenzie D. 2018. *The Book of Why: The New Science of Cause and Effect*. New York: Basic Books
Perry C. 2013. Machine learning and conflict prediction: a use case. *Stab. Int. J. Secur. Dev.* 2(3):56
Peters J, Janzing D, Scholkopf B. 2017. Elements of causal inference: foundations and learning algorithms. Cambridge, MA: MIT Press
Popper K. 1935. *Logik der Forschung*. Vienna: Julius Springer
Raftery AE. 1995. Bayesian model selection in social research. *Sociol. Methodol.* 25:111–63
Ragin C. 1987. *The Comparative Method: Moving Beyond Qualitative and Quantitative Strategies*. Berkeley: Univ. Calif. Press
Setoguchi S, Schneeweiss S, Brookhart MA, Glynn RJ, Cook EF. 2008. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiol. Drug Saf.* 17:546–55
Simmons JP, Nelson LD, Simonsohn U. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* 22:1359–66
Sovilj D, Eirola E, Miche Y, Björk KM, Nian R, et al. 2016. Extreme learning machine for missing data using multiple imputations. *Neurocomputing* 174:220–31
Starr SB. 2014. Evidence-based sentencing and the scientific rationalization of discrimination. *Stanford Law Rev.* 66:803
Swedberg R. 2014. *The Art of Social Theory*. Princeton, NJ: Princeton Univ. Press
Tibshirani R. 1996. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B* 58:267–88

Advice on software to analyze big data, to apply machine learning to social science questions.

Varian HR. 2014. Big data: new tricks for econometrics. *J. Econ. Perspect.* 28:3–27
Varma S, Simon R. 2006. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics* 7:91
Wager S, Athey S. 2018. Estimation and inference of heterogeneous treatment effects using random forests. *J. Am. Stat. Assoc.* 113:1228–42

Watts DJ. 2004. The new science of networks. *Annu. Rev. Sociol.* 30:243–70
Watts DJ. 2014. Common sense and sociological explanations. *Am. J. Sociol.* 120:313–51
Weber M. 1978. *Economy and Society*. Berkeley: Univ. Calif. Press
Western B. 1996. Vague theory and model uncertainty in macrosociology. *Sociol. Methodol.* 26:165–92
Westreich D, Lessler J, Funk MJ. 2010. Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *J. Clin. Epidemiol.* 63:826–33
White HC, Boorman SA, Breiger RL. 1976. Social structure from multiple networks. I. Blockmodels of roles and positions. *Am. J. Sociol.* 81:730–80
Wolpert D, Macready W. 1997. No free lunch theorems for optimization. *IEEE Trans. Evol. Comput.* 1:67–82
Wong SC, Gatt A, Stamatescu V, McDonnell MD. 2016. Understanding data augmentation for classification: when to warp? In *2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, ed. AW Liew, B Lovell, C Fookes, J Zhou, Y Gao, et al., pp. 1–6. New York: IEEE
Wyss R, Ellis AR, Brookhart MA, Girman CJ, Jonsson Funk M, et al. 2014. The role of prediction modeling in propensity score estimation: an evaluation of logistic regression, bCART, and the covariate-balancing propensity score. *Am. J. Epidemiol.* 180:645–55
Xie Y. 2007. Otis Dudley Duncan's legacy: the demographic approach to quantitative reasoning in social science. *Res. Soc. Strat. Mobil.* 25:141–56
Xie Y. 2013. Population heterogeneity and causal inference. *PNAS* 110:6262–8
Xing EP, Fu W, Song L. 2010. A state-space mixed membership blockmodel for dynamic network tomography. *Ann. Appl. Stat.* 4:535–66
Yang T, Chi Y, Zhu S, Gong Y, Jin R. 2011. Detecting communities and their evolutions in dynamic social networks—a Bayesian approach. *Mach. Learn.* 82:157–89
Yarkoni T, Westfall J. 2017. Choosing prediction over explanation in psychology: lessons from machine learning. *Perspect. Psychol. Sci.* 12:1100–22
Young C. 2009. Model uncertainty in sociological research: an application to religion and economic growth. *Am. Sociol. Rev.* 74:380–97
Young C, Holsteen K. 2017. Model uncertainty and robustness. a computational framework for multimodel analysis. *Sociol. Methods Res.* 46:3–40

Explains relevance of machine learning to psychological research, to preventing *p*-hacking.

RELATED RESOURCES

Bishop CM. 2016. *Pattern Recognition and Machine Learning*. New York: Springer
Murphy KP. 2012. *Machine Learning: A Probabilistic Perspective*. Cambridge, MA: MIT Press
Salganik MJ. 2017. *Bit by Bit: Social Research in the Digital Age*. Princeton, NJ: Princeton Univ. Press
Summer Institute in Computational Social Science. 2017. Online resources. *SICSS.* **https://compsocialscience.github.io/summer-institute/2017/#schedule**