Machine Learning for Sociology[a]

Mario Molina (mm2535@cornell.edu)*
Filiz Garip (fgarip@cornell.edu)*
**Department of Sociology, Cornell University*,
Ithaca, NY, 14853, USA

January 9, 2019

[a]This manuscript is forthcoming in the *Annual Review of Sociology*, vol. 45, 2019.

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 into this vast toolbox, and illustrate its current uses in 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. In addition to providing similar use in sociology, we argue that ML 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.

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

Introduction

Machine learning (ML) seeks to automate discovery from data. It represents a breakthrough in computer science where intelligent systems typically involved fixed algorithms (logical set of instructions) that code 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 off-shoot of non-parametric statistics (Kleinberg et al., 2015).

We can classify ML tools by how they learn (extract information) from data. Different 'tribes' 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). 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 (X). 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 example social science applications. We cannot give a comprehensive account given the sprawl of the topic, but we hope to provide enough coverage to allow the 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

Supervised machine learning (SML) involves searching for functions, $f(X)$, that predict an output (Y) given an input (X).[2] One can consider different classes of functions, such as linear models, decision trees, or neural networks. Let's take the linear model as a tool for prediction.[3] We have an input vector, X, and want to make a prediction on the output, Y, denoted as $\hat{Y}$ ('y-hat') with the model

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

where X^T is the vector transpose and β ('beta') is the vector of coefficients.

[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 the 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]In SML, the dependent variable is referred to as the 'output' while the explanatory or independent variables are called 'inputs' or 'features'. The prediction task is called **classification** when the output is discrete, and **regression** when it is continuous.

[3]Uppercase letters, such as X or Y, denote variable vectors, and lowercase letters refer to observed values (e.g., x_i is the i-th value of X).

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 from data with n observations:[4]

$$\sum_{i=1}^{n}(y_i - f(x_i))^2 \tag{1}$$

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 of 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).[5] For example, an SML technique is to minimize

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 (Y) is related to inputs (X). 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. Machine learning follows predictive modeling. The central goal is prediction, that is, to forecast the outcome (Y) for future inputs (X). 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.

$$\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda R(f) \tag{2}$$

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 ('lambda'), which sets the relative 'price' for variance (Kleinberg et al., 2015).

[4]The sum of squared residuals is only one among many possible "loss functions" in ML.

[5]One can find a similar approach in multi-level models popular in sociology where cluster parameters are deliberately biased (Gelman & Hill, 2007).

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,

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

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 offer a compromise between minimizing the sum of squared residuals, and yielding the smallest absolute coefficient sum.

SML techniques seek to achieve an ideal balance between reducing the in-sample and out-of-sample error (aka **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. Take a simple example. An OLS model with only a linear term linking an input (X) to output (Y) 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. Consider an extreme case. An OLS model with N inputs (plus a constant) will perfectly fit N data points, but 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, that is, the idiosyncratic factors that vary from sample to sample. We 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. To see that, note that a complex function will typically have low bias but high variance (Hastie et al., 2009). And recall that the regularizer, $R(f)$, penalizes functions that create variance; it often does so by expressing model complexity.

Let's go back to 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 (X) have non-zero coefficients, and effectively restrains model complexity (Tibshirani, 1996).[6]

Now consider regression trees, another function class in SML. The method proceeds by partitioning the inputs (X) 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

[6]Regularization can be understood as putting a prior on the final solution, β. To illustrate, assume we have the optimization problem in linear regression: $\max_\beta \; P(\beta|y, x)$, for some random variables $Y = y$ and $X = x$. If we apply Bayes' rule (and omit the normalizing term for simplicity), we get $P(\beta|y, x) = P(y, x|\beta) \times P(\beta)$. We further express $P(y, x|\beta)$ as $P(y|x, \beta) \times P(x|\beta)$ by applying chain rule, where $P(y|x, \beta)$ is the likelihood function. Then, the optimization problem becomes: $\max_\beta \; P(\beta|y, x) = \max_\beta P(y|x, \beta) \times P(x|\beta) \times P(\beta) = \max_\beta \prod_{i=1}^{n} [P(y_i|x_i, \beta) \times P(x_i|\beta) \times P(\beta)]$, assuming n independent and identically distributed observations. If we take the logarithm, we obtain the log-likelihood function: $\max_\beta \sum_{i=1}^{n} [\log(P(y_i|x_i, \beta)) + \log(P(x_i|\beta)) + \log(P(\beta))]$. The term $P(x_i|\beta)$ can be dropped from this function given that x is not a function of β (and therefore $P(x_i|\beta) = P(x_i)$), leading to: $\max_\beta \sum_{i=1}^{n} [\log(P(y_i|x_i, \beta)) + \log(P(\beta))]$ for some regularizer $R(f) = P(\beta)$.

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 within 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, first, decide 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).

Some SML 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 (least absolute shrinkage and selection operator), $\sum_{j=1}^{p} \beta_j^2$ 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 ('leaf'). Regularizers include tree depth and number of leaves. 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 neighbors

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 multi-layer set-up that models the output (Y) as a concatenation of simple non-linear functions of the linear combinations of inputs (X) ('neurons'). Regularizers include number of layers and number of neurons per layer.

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 SML Techniques).[7] The general recommendation is to use the substantive question at hand to

[7]The 'no free lunch' theorem proves that no ML method (or no form of regularization) is universally

guide these choices.[8] 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 (*aka* **empirical tuning**) (Mullainathan & Spiess, 2017; Athey & Imbens, 2017; Kleinberg et al., 2015). These features allow 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 (X), and a simple functional form to relate the inputs to the output (Y).

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, having settled on a final model, to estimate its generalization (prediction) error on new data (Hastie et al., 2009).[9]

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 set-up, one creates three, not two, separate data sets. **Training data** is used to fit the model; **validation data** is 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** is 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 with less 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 re-cycle 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 k-th fold.

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

better than any other (Wolpert & Macready, 1997). The task, then, 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 counter argument).

[8] 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), Abadie & Kasy (2017) link SML methods to traditional tools and questions in economics. Olson et al. (2018) offer an empirical comparison on bioinformatics data.

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

pick the complexity level that minimizes the average prediction error across the left-out folds.[10] In the final step, we can use the test data to compute the predictive accuracy (generalization error) of the selected model.

SML for Policy Predictions, Causal Inference and Data Augmentation

SML uses flexible functions of inputs (X) to predict an output (Y). Some SML tools, such as nearest neighbor, have no parameters at all. Other methods, such as LASSO, give parameter estimates, $\hat{\beta}$ ('beta-hat'), but those estimates are not always consistent (that is, 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.[11]

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 allow for a "common-task framework" where different teams can compete on the same question (Donoho, 2017).[12]

Economist Ed Glaesar and colleagues (2016) used this idea to set up a competition to produce predictive algorithms for city governments. Sociologist Matt Salganik and collaborators started a challenge to predict educational (and other) outcomes in the Fragile Families data.[13] The organizing team judged the submissions from 150 multi-disciplinary teams on predictive accuracy on test (hold-out) data. In the on-going second phase, the team plans to conduct 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, 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. Cederman & Weidmann (2017) discuss how SML can predict and prevent 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

[10] See Varma & Simon (2006) for more sophisticated 'nested cross-validation'.

[11] Mullainathan & Spiess (2017) review predictive modeling in economics, Cranmer & Desmarais (2017) in political science, and Yarkoni & Westfall (2017) in psychology.

[12] The company Kaggle hosts competitions (www.kaggle.com/competitions) where contestants train models on shared data and compete on predictive accuracy.

[13] http://www.fragilefamilieschallenge.org/

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 regression-tree model to predict judges' bail-or-release decisions in New York City, and then 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: reducing reoffending rate by 25 percent with no increase in jailing rate, or alternatively, pulling down jailing rate by 42 percent 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 a random-forest model 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, like 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 sub-population (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.[14]

[14]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

Causal Inference

Social scientists are often interested in identifying the causal effect of an input ('treatment') (X) on an output (Y). SML tools can help in certain causal inference procedures that involve prediction tasks. 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.[15] 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 (X), and use a LASSO model to select the inputs that are most important in predicting increase in worker earnings. Similarly, Athey & Imbens (2016) develop 'causal trees' to estimate treatment effects for sub-groups. 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' (sub-group) 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 output (Y) 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, 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 assignment to treatment and the output. Regularization in SML could lead to exclusion of such inputs from the model,

casualties, they are not thrilled with the prospect of riding in 'utilitarian cars' that can sacrifice its driver for the greater good (Bonnefon et al., 2016)

[15] Morgan & Winship (2007, 2014) offer an authoritative review of causal inference in social sciences.

for example, leading to omitted variable bias in estimation. Similarly, with many inputs, one generally runs the risk of model misspecification (Belloni et al., 2014; Ho et al., 2007; King & Nielsen, 2016; Raftery, 1995; Young & Holsteen, 2017; Muñoz & Young, 2018). 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 enter 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 assignment to treatment but not directly with the output (Angrist et al., 1996). We can then regress the treatment (a given input, X) on the instrument (Z), and then use the predicted values ($\hat{X}$) as an input in the output (Y) regression. Because the first stage in this 'instrumental variables' (IV) 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.[16] Feigenbaum (2015), for example, input human-coded data to train SML algorithms to link individuals across census waves. Abramitzky et al. (2018) develop a fully-automated method to estimate probabilities of matches across census waves, and then measured intergenerational occupational mobility. Using a nested design, Bernheim et al. (2013) recruited a subset of survey respondents to participate in a lab experiment, and used their responses in the lab as training data to impute responses for the remaining sample. Blumenstock et al. (2015) collected 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) applied this approach successfully to predict topics for image labels and annotations.

Researchers are also using SML for missing data imputation. Farhangfar et al. (2008) investigated the performance of different ML classifiers in fifteen datasets and find that, although no method is universally best, naïve-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 datasets, yields more accurate values compared to conditional mean imputation.

[16] 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.

UNSUPERVISED MACHINE LEARNING (UML)

Unsupervised machine learning (UML) searches for a representation of the inputs (X) 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 sidebar titled Some UML Techniques).[17] There is no target output (Y) 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.

UML 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 and 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).[18]

Following a long tradition, sociologists also use UML to group social network data. Earlier applications, such as 'blockmodels', employed *structural equivalence* (sharing neighbors) to evaluate similarity, and to then partition the network into sub-groups (White et al., 1976; Breiger et al., 1975). 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).

[17] 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).

[18] To learn more about text analysis, see Blei (2012), Grimmer & Stewart (2013), Mohr & Bogdanov (2013), Bail (2014), Evans & Aceves (2016).

Some UML Techniques

Principal components analysis

Discovers a small number of linear combinations of the inputs (X) 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, Y).

Factor analysis

Discovers latent (unobserved) factors that account for the correlation in inputs (X); 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 (X); 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.

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 women's 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-U.S. 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' are well separated (DiMaggio et al., 2013). 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' (CAC). 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 bring 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 macro-level economic and political indicators.

ML: New Answers to Old Questions

There are two broad categories of machine learning (ML) – an off-shoot of computer science and statistics. Supervised machine learning (SML) builds a model of inputs (X) to predict an output (Y) in new data. Unsupervised machine learning (UML) discovers patterns in inputs (X) without a target (Y) to predict. 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.

SML 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), and 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 (X) to the output (Y). It helps break away from the 'general linear reality' 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).

ML 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, or 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 sub-groups in populations, and then link the emergence of each sub-group 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 'grounded theory' approach in particular).

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

SML 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' (capturing the idiosyncrasies of the sample at hand). SML, if nothing else, gives us the crucial idea that we need to validate our results on new data (or, with efficient partitioning of the original data, *aka* 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 we also evaluate the overall performance of a model in explaining an output (Y). 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' (*aka* '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, 2016).[19] Any time we use the data to optimize over these degrees of freedom (for example, choose 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).[20]

ML 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

[19]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 Science Collaboration, 2015). In sociology, Teplitskiy et al. (2018), for example, re-estimated hundreds of published models on the General Social Data (GSS) with slight perturbations, and found reduced number of significant coefficients, standardized coefficient sizes, and share of explained variation (R^2).

[20]Social change makes it difficult to conduct out-of-sample tests with data from different time periods. Teplitskiy et al.'s (2018) replication of GSS-based studies, for example, shows weaker results when models are estimated on 'future' data.

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 pre-registering hypotheses and other design choices in a public forum (e.g., the Open Science Framework) (Baldassarri & Abascal, 2017; Hofman et al., 2017; Simmons et al., 2011; Ioannidis & Doucouliagos, 2013; Watts, 2014). Instead we go back and forth between data, statistical models, and theory until we gain a novel insight.

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.

ML 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 set-up, and use data to estimate the effects of some input (X) on an output (Y). ML not only helps us improve parts of this strategy, but also give us tools that can inspire new questions. How well do a set of inputs (X), for example, predict output (Y)? How do these predictions deviate from observed outcomes and why? Or what is the underlying structure of some input (X)? How is that structure related to external factors (Z)? Answering these questions can help us push theory 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, 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 aims at prediction (use inputs X to forecast unobserved output $\hat{Y}$).
2. Supervised machine learning (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 (X), 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?

Related Resources

Murphy, Kevin P., 2012. Machine learning. A probabilistic perspective. MIT Press, Cambridge.

Bishop, C,M., 2016. Pattern recognition and machine learning. Springer-Verlag, New York.

Salganik, M.J., 2017. Bit by bit: Social research in the digital age. Princeton University Press, Princeton.

Summer Institute in Computational Social Science, Online Resources. `https://compsocialscience.github.io/summer-institute/2017/#schedule`

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.

Acknowledgements

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.

Bibliography

Abadie A, Kasy M. 2017. The Risk of Machine Learning. `arxiv.org/abs/1703.10935`

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: University of 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, 3. 2018. Linking Individuals Across Historical Sources: a Fully Automated Approach. *Hist. Methods* (forthcoming)

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* 355:483–485

Athey S, Imbens G. 2015. A Measure of Robustness to Misspecification. *Am. Econ. Rev.* 105:476–480

Athey S, Imbens G. 2016. Recursive partitioning for heterogeneous causal effects. *PNAS* 113:7353–7360

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. Soc. Rev.* 73:37–59

Bail CA. 2014. The cultural environment: measuring culture with big data. *Theor. Soc.* 43:465–482

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. Tech.* 68:1397–1410

Beck N, King G, Zeng L. 2000. Improving Quantitative Studies of International Conflict: A Conjecture. *Am. Polit. 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* 80:2369–2429

Belloni A, Chernozhukov V, Fernandez-Val I, Hansen C. 2017. Program Evaluation and Causal Inference With High-Dimensional Data. *Econometrica* 85:233–298

Belloni A, Chernozhukov V, Hansen C. 2014. Inference on Treatment Effects after Selection among High-Dimensional Controls. *Rev. Econ. Stud.* 81:608–650

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. Method. Res.*

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. `http://www.nber.org/papers/w19269`

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

Blei DM, McAuliffe JD. 2010. Supervised Topic Models. `https://arxiv.org/abs/1003.0783`

Blumenstock J, Cadamuro G, On R. 2015. Predicting poverty and wealth from mobile phone metadata. *Science* 350:1073–1076

Bollen KA. 2002. Latent Variables in Psychology and the Social Sciences. *Annu. Rev. Psychol.* 53:605–634

Bonikowski B, DiMaggio P. 2016. Varieties of American Popular Nationalism. *Am. Soc. Rev.* 81:949–980

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.* 12:328–383

Breiman L. 2001b. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Stat. Sci.* 16:199–231

Breiman L. 2001a. Random Forests. *Mach. Learn.* 45:5–32

Carrasco M. 2012. A regularization approach to the many instruments problem. *J. Econometrics* 170:383–398

Cederman LE, Weidmann NB. 2017. Predicting armed conflict: Time to adjust our expectations? *Science* 355:474–476

Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W. 2017. Double/Debiased/Neyman Machine Learning of Treatment Effects. *Am. Econ. Rev.* 107:261–265

Chong W, Blei D, Li FF. 2009. Simultaneous image classification and annotation, In *Proc. CVPR. IEEE*. Queensland, Australia

Coleman J. 1964. *Introduction to Mathematical Sociology*. New York: Free Press

Cornwell B. 2015. *Social sequence analysis : methods and applications*. New York: Cambridge University Press

Cranmer SJ, Desmarais BA. 2017. What Can We Learn from Predictive Modeling? *Polit. Anal.* 25:145–166

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.* 95:932–945

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

Donoho D. 2017. 50 Years of Data Science. *J. Comput. Graph. Stat.* 26:745–766

Duncan OD. 1982. *Rasch measurement and sociological theory (Hollingshead Lecture)*

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*. New York, USA: ACM Press

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.* 41:3692–3705

Feigenbaum JJ. 2015. Automated Census Record Linking: A Machine Learning Approach. 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.* 36:153–172

Frye M, Trinitapoli J. 2015. Ideals as Anchors for Relationship Experiences. *Am. Soc. 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*. New Jersey: Princeton University Press

Gelman A, Hill J. 2007. *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University 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.* 106:114–118

Goldberg A. 2011. Mapping Shared Understandings Using Relational Class Analysis: The Case of the Cultural Omnivore Reexamined. *Am. J. Sociol.* 116:1397–1436

Goodfellow I, Bengio Y, Courville A. 2016. *Deep learning*. Cambridge: 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. *Polit. Anal.* 25:413–434

Grimmer J, Stewart BM. 2013. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Polit. Anal.* 21:267–297

Grosse R. 2013. Predictive learning vs. representation learning. https://hips.seas.harvard.edu/blog/2013/02/04/predictive-learning-vs-representation-learning/

Handcock MS, Raftery AE, Tantrum JM. 2007. Model-Based Clustering for Social Networks. *J. R. Stat. Soc.* 170:301–354

Harcourt BE. 2007. *Against prediction. Profiling, policing, and punishing in an actuarial age*. Chicago: University of 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*. Curran Associates Inc.

Hartford J, Lewis G, Leyton-Brown K, Taddy M. 2016. Counterfactual Prediction with Deep Instrumental Variables Networks. `https://arxiv.org/abs/1612.09596`

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.* 20:217–240

Ho DE, Imai K, King G, Stuart EA. 2007. Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Polit. Anal.* 15:199–236

Hoff PD, Raftery AE, Handcock MS. 2002. Latent Space Approaches to Social Network Analysis. *J. Am. Stat. Assoc.* 97:1090–1098

Hofman JM, Sharma A, Watts DJ. 2017. Prediction and explanation in social systems. *Science* 355:486–488

Hopkins DJ, King G. 2010. A Method of Automated Nonparametric Content Analysis for Social Science. *Am. J. Polit. Sci.* 54:229–247

Imai K, Ratkovic M. 2013. Estimating treatment effect heterogeneity in randomized program evaluation. *Ann. Appl. Stat.* 7:443–470

Imbens GW, Rubin DB. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences. An Introduction*. New York: Cambridge University 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* 349:255–260

Killewald A, Zhuo X. 2018. Mothers' Long-Term Employment Patterns. `https://scholar.harvard.edu/xiaolinzhuo/publications/mothers\%E2\%80\%99-long-term-employment-patterns`

King G, Nielsen R. 2016. Why Propensity Scores Should Not Be Used for Matching. `https://gking.harvard.edu/publications/why-propensity-scores-should-not-be-used-formatching`

Kleinberg J, Liang A, Mullainathan S. 2017. The Theory is Predictive, but is it Complete? An Application to Human Perception of Randomness. `https://arxiv.org/abs/1706.06974`

Kleinberg J, Ludwig J, Mullainathan S, Obermeyer Z. 2015. Prediction Policy Problems. *Am. Econ. Rev.* 105:491–495

Kleinberg J, Mullainathan S, Raghavan M. 2016. Inherent Trade-Offs in the Fair Determination of Risk Scores. https://arxiv.org/abs/1609.05807

Knight K, Fu W. 2000. Asymptotics for lasso-type estimators. *Ann. Stat.* 28:1356–1378

Lee BK, Lessler J, Stuart EA. 2010. Improving propensity score weighting using machine learning. *Stat. Med.* 29:337–346

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

Matias C, Miele V. 2017. Statistical clustering of temporal networks through a dynamic stochastic block model. *Stat. Methodol.* 79:1119–1141

McFarland DA, Ramage D, Chuang J, Heer J, Manning CD, Jurafsky D. 2013. Differentiating language usage through topic models. *Poetics* 41:607–625

Mohr JW, Bogdanov P. 2013. IntroductionTopic models: What they are and why they matter. *Poetics* 41:545–569

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*. New York: Cambridge University Press, 1st ed.

Morgan SL, Winship C. 2014. *Counterfactuals and Causal Inference. Methods and Principles for Social Research*. New York: Cambridge University Press, 2nd ed.

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*. New York, New York, USA: ACM Press

Muñoz J, Young C. 2018. We Ran 9 Billion Regressions: Eliminating False Positives through Computational Model Robustness. *Sociol. Methodol.*

Narayanan A. 2018. 21 fairness definitions and their politics. https://www.youtube.com/watch?v=jIXIuYdnyyk

Nowicki K, Snijders TAB. 2001. Estimation and Prediction for Stochastic Blockstructures. *J. Am. Stat. Assoc.* 96:1077–1087

Olson RS, La Cava W, Mustahsan Z, Varik A, Moore JH. 2018. Data-driven advice for applying machine learning to bioinformatics problems., In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, vol. 23. NIH Public Access

Open Science Collaboration. 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

Peters J, Janzing D, Scholkopf B. 2017. Elements of causal inference: foundations and learning algorithms. Cambridge: The MIT Press

Popper K. 1935. *Logik der Forschung*. Vienna: Julius Springer

Raftery AE. 1995. Bayesian Model Selection in Social Research. *Sociol. Methodol.* 25:111–163

Ragin C. 1987. *The comparative method: Moving beyond qualitative and quantitative strategies*. Berkeley and Los Angeles: University of California 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. *Pharmacoepidem. Dr. S.* 17:546–555

Simmons JP, Nelson LD, Simonsohn U. 2011. False-Positive Psychology. Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychol. Sci.* 22:1359–1366

Sovilj D, Eirola E, Miche Y, Björk KM, Nian R, et al. 2016. Extreme learning machine for missing data using multiple imputations. *Neurocomputing* 174:220–231

Starr SB. 2014. Evidence-Based Sentencing and the Scientific Rationalization of Discrimination. *Stanford Law Rev.* 66

Swedberg R. 2014. *The Art of Social Theory*. Princeton: Princeton University Press

Teplitskiy M., St. Onge J., Evans J. 2018. How Firm is Sociological Knowledge? Reanalysis of GSS findings with alternative models and out-of-sample data, 1972-2012. *Working paper*

Tibshirani R. 1996. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc.. Series B (Methodological)* 58:267–288

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

Wager S, Athey S. 2018. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *J. Am. Stat. Assoc.*

Watts DJ. 2004. The New Science of Networks. *Annu. Rev. Sociol.* 30:243–270

Watts DJ. 2014. Common Sense and Sociological Explanations. *Am. J. Sociol.* 120:313–351

Weber M. 1978. *Economy and Society*. Berkeley and L.A.: University of California Press

Western B. 1996. Vague theory and model uncertainty in macrosociology. *Sociol. Methodol.* 26:165–192

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.* 63:826–833

White HC, Boorman SA, Breiger RL. 1976. Social Structure from Multiple Networks. I. Blockmodels of Roles and Positions. *Am. J. Sociol.* 81:730–780

Wolpert D, Macready W. 1997. No free lunch theorems for optimization. *IEEE T. Evolut. Comput.* 1:67–82

Wong SC, Gatt A, Stamatescu V, McDonnell MD. 2016. Understanding Data Augmentation for Classification: When to Warp?, In *Proc. CVPR IEEE* I.5.2, I.4.7

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.* 180:645–655

Xie Y. 2007. Otis Dudley Duncan's legacy: The demographic approach to quantitative reasoning in social science. *Res. Soc. Strat. Mobil.* 25:141–156

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.* 4:535–566

Yang T, Chi Y, Zhu S, Gong Y, Jin R. 2011. Detecting communities and their evolutions in dynamic social networksa Bayesian approach. *Mach. Learn.* 82:157–189

Yarkoni T, Westfall J. 2017. Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspect. Psychol. Sci.* 12:1100–1122

Young C. 2009. Model Uncertainty in Sociological Research: An Application to Religion and Economic Growth. *Am. Soc. Rev.* 74:380–397

Young C, Holsteen K. 2017. Model Uncertainty and Robustness. A Computational Framework for Multimodel Analysis. *Sociol. Methods Res.* 46:3–40