The Impact of Machine Learning on Economics

Susan Athey
athey@stanford.edu

Current version January 2018

Abstract

This paper provides an assessment of the early contributions of machine learning to economics, as well as predictions about its future contributions. It begins by briefly overviewing some themes from the literature on machine learning, and then draws some contrasts with traditional approaches to estimating the impact of counterfactual policies in economics. Next, we review some of the initial "off-the-shelf" applications of machine learning to economics, including applications in analyzing text and images. We then describe new types of questions that have been posed surrounding the application of machine learning to policy problems, including "prediction policy problems," as well as considerations of fairness and manipulability. We present some highlights from the emerging econometric literature combining machine learning and causal inference. Finally, we overview a set of broader predictions about the future impact of machine learning on economics, including its impacts on the nature of collaboration, funding, research tools, and research questions.

1 Introduction

I believe that machine learning (ML) will have a dramatic impact on the field of economics within a short time frame. Indeed, the impact of ML on economics is already well underway, and so it is perhaps not too difficult to predict some of the effects.

The paper begins by stating the definition of ML that I will use in this paper, describing its strengths and weaknesses, and contrasting ML with traditional econometrics tools for causal inference, which is a primary focus of the empirical economics literature. Next, I review some applications of ML in economics where ML can be used off-the-shelf: the use case in economics is essentially the same use case that the ML tools were designed an optimized for. I then review "prediction policy" problems (Kleinberg et al., 2015), where prediction tools have been embedded in the context of economic decision-making. Then, I provide an overview of the questions considered and early themes of the the emerging literature in econometrics and statistics combining machine learning and causal inference, a literature that is providing insights and theoretical results that are novel from the perspective of both ML and statistics/econometrics. Finally, I step back and describe the implications of the field of economics as a whole. Throughout, I make reference to the literature broadly, but do not attempt to conduct a comprehensive survey or reference every application in economics.

The paper highlights several themes.

A first theme is that ML does not add much to questions about identification, which concern when the object of interest, e.g. a causal effect, can be estimated with infinite data, but rather

I am grateful to David Blei, Guido Imbens, Denis Nekipelov, Francisco Ruiz, and Stefan Wager, with whom I have collaborated on many projects at the intersection of machine learning and econometrics and who have shaped my thinking, as well as to Hal Varian, Mike Luca and Sendhil Mullainathan, who have also contributed to my thinking through their writing, lecture notes, and many conversations.

yields great improvements when the goal is semi-parametric estimation or when there are a large number of covariates relative to the number of observations. ML has great strengths in using data to select functional forms flexibly.

A second theme is that a key advantage of ML is that ML views empirical analysis as "algorithms" that estimate and compare many alternative models. This approach constrasts with economics, where (in principle, though rarely in reality) the researcher picks a model based on principles and estimates it once. Instead, ML algorithms build in "tuning" as part of the algorithm. The tuning is essentially model selection, and in an ML algorithm that is data-driven. There are a whole host of advantages of this approach, including improved performance as well as enabling researchers to be systematic and fully describe the process by which their model was selected. Of course, cross-validation has also been used historically in economics, for example for selecting the bandwidth for a kernel regression, but it is viewed as a fundamental part of an algorithm in ML.

A third, closely related theme is that "outsourcing" model selection to algorithm works very well when the problem is "simple"–for example, prediction and classification tasks, where performance of a model can be evaluated by looking at goodness of fit in a held-out test set. Those are typically not the problems of greatest interest for empirical researchers in economics, who instead are concerned with causal inference, where there is typically not an unbiased estimate of the ground truth available for comparison. Thus, more work is required to apply an algorithmic approach to economic problems. The recent literature at the intersection of ML and causal inference, reviewed in this paper, has focused on providing the conceptual framework and specific proposals for algorithms that are tailored for causal inference.

A fourth theme is that the algorithms also have to be modified to provide valid confidence intervals for estimated effects when the data is used to select the model. Many recent papers make use of techniques such as sample splitting, leave-one-out estimation, and other similar techniques to provide confidence intervals that work both in theory and in practice. The upside is that using ML can provide the best of both worlds: the model selection is data driven, systematic, and a wide range of models are considered; yet, the model selection process is fully documented, and confidence intervals take into account the entire algorithm.

Finally, the combination of ML and newly available datasets will change economics in fairly fundamental ways, ranging from new questions, to new approaches to collaboration (larger teams and interdisciplinary interaction), to a change in how involved economists are in the engineering and implementation of policies.

2 What is Machine Learning and What are Early Use Cases?

It is harder than one might think to come up with an operational definition of ML. The term can be (and has been) used broadly or narrowly; it can refer to a collections of subfields of computer science, but also to a set of topics that are developed and used across computer science, engineering, statistics, and increasingly the social sciences. Indeed, one could devote an entire article to the definition of ML, or to the question of whether the thing called ML really needed a new name other than statistics, the distinction between ML and AI, and so on. However, I will leave this debate to others, and focus on a narrow, practical definition that will make it easier to distinguish ML from the most commonly used econometric approaches used in applied econometrics until very recently.[1] For readers coming from a machine learning background, it is also important to note

[1] I will also focus on the most popular parts of ML; like many fields, it is possible to find researchers who define themselves as members of the field of ML doing a variety of different things, including pushing the boundaries of ML with tools from other disciplines. In this article I will consider such work to be interdisciplinary rather than "pure"

that applied statistics and econometrics have developed a body of insights on topics ranging from causal inference to efficiency that have not yet been incorporated in mainstream machine learning, while other parts of machine learning have overlap with methods that have been used in applied statistics and social sciences for many decades.

Starting from a relatively narrow definition of machine learning, machine learning is a field that develops algorithms designed to be applied to datasets, with the main areas of focus being prediction (regression), classification, and clustering or grouping tasks. These tasks are divided into two main branches, supervised and unsupervised ML. Unsupervised ML involves finding clusters of observations that are similar in terms of their covariates, and thus can be interpreted as "dimensionality reduction"; it is commonly used for video, images and text. There are a variety of techniques available for unsupervised learning, including k-means clustering, topic modeling, community detection methods for networks, and many more. For example, the Latent Dirichlet Allocation model (Blei et al., 2003b) has frequently been applied to find "topics" in textual data. The output of a typical unsupervised ML model is a partition of the set of observations, where observations within each element of the partition are similar according to some metric; or, a vector of probabilities or weights that describe a mixture of topics or groups that an observation might belong to. If you read in the newspaper that a computer scientist "discovered cats on YouTube," that might mean that they used an unsupervised ML method to partition a set of videos into groups, and when a human watches the the largest group, they observe that most of the videos in the largest group contain cats. This is referred to as "unsupervised" because there were no "labels" on any of the images in the input data; only after examining the items in each group does an observer determine that the algorithm found cats or dogs. Not all dimensionality reduction methods involve creating clusters; older methods such as principal components analysis can be used to reduce dimensionality, while modern methods include matrix factorization (finding two low-dimensional matrices whose product well approximates a larger matrix), regularization on the norm of a matrix, hierarchical Poisson factorization (in a Bayesian framework) (Gopalan et al., 2015), and neural networks.

In my view, these tools are very useful as an intermediate step in empirical work in economics. They provide a data-driven way to find similar newspaper articles, restaurant reviews, etc., and thus create variables that can be used in economic analyses. These variables might be part of the construction of either outcome variables or explanatory variables, depending on the context. For example, if an analyst wishes to estimate a model of consumer demand for different items, it is common to model consumer preferences over characteristics of the items. Many items are associated with text descriptions as well as online reviews. Unsupervised learning could be used to discover items with similar product descriptions, in an initial phase of finding potentially related products; and it could also be used to find subgroups of similar products. Unsupervised learning could further be used to categorize the reviews into types. An indicator for the review group could be used in subsequent analysis without the analyst having to use human judgement about the review content; the data would reveal whether a certain type of review was associated with higher consumer perceived quality, or not. An advantage of using unsupervised learning to create covariates is that the outcome data is not used at all; thus, concerns about spurious correlation between constructed covariates and the observed outcome are less problematic. Despite this, Egami et al. (2016) have argued that researchers may be tempted to fine-tune their construction of covariates by testing how they perform in terms of predicting outcomes, thus leading to spurious relationships between covariates and outcomes. They recommend the approach of sample splitting, whereby the model tuning takes place on one sample of data, and then the selected model is applied on a fresh sample of data.

ML, and will discuss it as such.

Unsupervised learning can also be used to create outcome variables. For example, Athey et al. (2017d) examine the impact of Google's shutdown of Google News in Spain on the types of news consumers read. In this case, the share of news in different categories is an outcome of interest. Unsupervised learning can be used to categorize news in this type of analysis; that paper uses community detection techniques from network theory. In the absence of dimensionality reduction, it would be difficult to meaningfully summarize the impact of the shutdown on all of the different news articles consumed in the relevant time frame.

Supervised machine learning typically entails using a set of features or covariates (X) to *predict* an outcome (Y). When using the term prediction, it is important to emphasize that the framework focuses not on forecasting, but rather on a setting where there are some labelled observations where both X and Y are observed (the training data), and the goal is to predict outcomes (Y) in an independent test set based on the realized values of X for each unit in the test set. In other words, the goal is to construct $\hat{\mu}(x)$, which is an estimator of $\mu(x) = E[Y|X = x]$, in order to do a good job predicting the true values of Y in an independent dataset The observations are assumed to be independent, and the joint distribution of X and Y in the training set is the same as that in the test set. These assumptions are the only substantive assumptions required for most machine learning methods to work.

In the case of classification, the goal is to accurately classify observations. For example, the outcome could be the animal depicted in an image, the "features" or covariates are the pixels in the image, and the goal is to correctly classify images into the correct animal depicted. A related but distinct estimation problem is to estimate $Pr(Y = k|X = x)$ for each of $k = 1, .., K$ possible realizations of Y.

It is important to emphasize that the ML literature does not frame itself as solving estimation problems – so estimating $\mu(x)$ or $Pr(Y = k|X = x)$ is not the primary goal. Instead, the goal is to achieve goodness of fit in an independent test set by minimizing deviations between actual outcomes and predicted outcomes. In applied econometrics, we often wish to understand an object like $\mu(x)$ in order to perform exercises like evaluate the impact of changing one covariate while holding others constant. This is not an explicit aim of ML modeling.

There are a variety of ML methods for supervised learning, such as regularized regression (LASSO, ridge and elastic net), random forest, regression trees, support vector machines, neural nets, matrix factorization, and many others, such as model averaging. See Varian (2014) for an overview of some of the most popular methods, and Mullainathan and Spiess (2017) for more details. (Also note that White (1992) attempted to popularize neural nets in economics in the early 1990s, but at the time they did not lead to substantial performance improvements and did not become popular in economics.) What leads us to categorize these methods as "ML" methods rather than traditional econometric or statistical methods? First is simply an observation: until recently, these methods were neither used in published social science research, nor taught in social science courses, while they were widely studied in the self-described ML and/or "statistical learning" literatures. One exception is ridge regression, which received some attention in economics; and LASSO had also received some attention. But from a more functional perspective, one common feature of many ML methods is that they use data-driven model selection. That is, the analyst provides the list of covariates or features, but the functional form is at least in part determined as a function of the data, and rather than performing a single estimation (as is done, at least in theory, in econometrics), so that the method is better described as an algorithm that might estimate many alternative models and then select among them to maximize a criterion.

There is typically a tradeoff between expressiveness of the model (e.g. more covariates included in a linear regression) and risk of over-fitting, which occurs when the model is too rich relative to the sample size. (See Mullainathan and Spiess (2017) for more discussion of this.) In the latter case, the

goodness of fit of the model when measured on the sample where the model is estimated is expected to be much better than the goodness of fit of the model when evaluated on an independent test set. The ML literature uses a variety of techniques to balance expressiveness against over-fitting. The most common approach is cross-validation whereby the analyst repeatedly estimates a model on part of the data (a "training fold") and then evaluates it on the complement (the "test fold"). The complexity of the model is selected to minimize the average of the mean-squared error of the prediction (the squared difference between the model prediction and the actual outcome) on the test folds. Other approaches used to control over-fitting include averaging many different models, sometimes estimating each model on a subsample of the data (one can interpret the random forest in this way).

In contrast, in much of cross-sectional econometrics and empirical work in economics, the tradition has been that the researcher specifies one model, estimates the model on the full dataset, and relies on statistical theory to estimate confidence intervals for estimated parameters. The focus is on the estimated effects rather than the goodness of fit of the model. For much empirical work in economics, the primary interest is in the estimate of a causal effect, such as the effect of a training program, a minimum wage increase, or a price increase. The researcher might check robustness of this parameter estimate by reporting two or three alternative specifications. Researchers often check dozens or even hundreds of alternative specifications behind the scenes, but rarely report this practice because it would invalidate the confidence intervals reported (due to concerns about multiple testing and searching for specifications with the desired results). There are many disadvantages to the traditional approach, including but not limited to the fact that researchers would find it difficult to be systematic or comprehensive in checking alternative specifications, and further because researchers were not honest about the practice, given that they did not have a way to correct for the specification search process. I believe that regularization and systematic model selection have many advantages over traditional approaches, and for this reason will become a standard part of empirical practice in economics. This will particularly be true as we more frequently encounter datasets with many covariates, and also as we see the advantages of being systematic about model selection. As I discuss below, however, this practice must be modified from traditional ML and in general "handled with care" when the researcher's ultimate goal is to estimate a causal effect rather than maximize goodness of fit in a test set.

To build some intuition about the difference between causal effect estimation and prediction, it can be useful to consider the widely used method of instrumental variables. Instrumental variables are used by economists when they wish to learn a causal effect, for example the effect of a price on a firm's sales, but they only have access to observational (non-experimental) data. An instrument in this case might be an input cost for the firm that shifts over time, and is unrelated to factors that shift consumer's demand for the product (such demand shifters can be referred to as "confounders" becaues the affect both the optimal price set by the firm and the sales of the product). The instrumental variables method essentially projects the observed prices onto the input costs, thus only making use of the variation in price that is explained by changes in input costs when estimating the impact of price on sales. It is very common to see that a predictive model (e.g. least squares regression) might have very high explanatory power (e.g. high R^2), while the causal model (e.g. instrumental variables regression) might have very low explanatory power (in terms of predicting outcomes). In other words, economists typically abandon the goal of accurate prediction of outcomes in pursuit of an unbiased estimate of a causal parameter of interest.

Another difference derives from the key concerns in different approaches, and how those concerns are addressed. In predictive models, the key concern is the tradeoff between expressiveness and overfitting, and this tradeoff can be evaluated by looking at goodness of fit in an independent test set. In contrast, there are several distinct concerns for causal models. The first is whether the

parameter estimates from a particular sample are spurious, that is, whether estimates arise due to sampling variation, so that if a new random sample of the same size was drawn from the population, the parameter estimate would be substantially different. The typical approach to this problem in econometrics and statistics is to prove theorems about the consistency and asymptotic normality of the parameter estimates, propose approaches to estimating the variance of parameter estimates, and finally to use those results to estimate standard errors that reflect the sampling uncertainty (under the conditions of the theory). A more data-driven approach is to use bootstrapping, and estimate the empirical distribution of parameter estimates across bootstrap samples. The typical ML approach of evaluating performance in a test set does not directly handle the issue of the uncertainty over parameter estimates, since the parameter of interest is not actually observed in any test set. The researcher would need to estimate the parameter again in the test set.

A second concern is whether the assumptions required to "identify" a causal effect are satisfied, where in econometrics we say that a parameter is identified if we can learn it eventually with infinite data (where even in the limit, the data has the same structure as in the sample considered). It is well known that the causal effect of a treatment is not identified without making assumptions, assumptions which are generally not testable (that is, they cannot be rejected by looking at the data). Examples of identifying assumptions include the assumption that the treatment is randomly assigned, or that treatment assignment is "unconfounded." In some settings, these assumptions require the analyst to observe all potential "confounders" and control for them adequately; in other settings, the assumptions require that an instrumental variable is uncorrelated with the unobserved component of outcomes. In many cases it can be proven that even with a data set of infinite size, the assumptions are not testable–they can not be rejected by looking at the data, and instead must be evaluated on substantive grounds. Justifying assumptions is one of the primary components of an observational study in applied economics. If the "identifying" assumptions are violated, estimates may be biased (in the same way) in both training data and test data. Testing assumptions usually requires additional information, like multiple experiments (designed or natural) in the data. Thus, the ML approach of evaluating performance in a test set does not address this concern at all. Instead, ML is likely to help make estimation methods more credible, while maintaining the identifying assumptions: in practice, coming up with estimation methods that give unbiased estimates of treatment effects requires flexibly modeling a variety of empirical relationships, such as the relationship between the treatment assignment and covariates. Since ML excels at data-driven model selection, it can be useful in systematizing the search for the best functional forms when implementing an estimation technique.

Economists also build more complex models that incorporate both behavioral and statistical assumptions in order to estimate the impact of counterfactual policies that have never been used before. A classic example is McFadden's methodological work in the early 1970s (e.g. McFadden et al. (1973)) analyzing transportation choices. By imposing the behavioral assumption that consumers maximize utility when making choices, it is possible to estimate parameters of the consumer's utility function, and estimate the welfare effects and market share changes that would occur when a choice is added or removed (e.g. extending the BART transportation system), or when the characteristics of the good (e.g. price) are changed. Another example with more complicated behavioral assumptions is the case of auctions. For a dataset with bids from procurement auctions, the "structural" approach involves estimating a probability distribution over bidder values, and then evaluating the counterfactual effect of changing auction design (e.g. Laffont et al. (1995), Athey et al. (2011), Athey et al. (2013) or the review Athey and Haile (2007)). For further discussions of the contrast between prediction and parameter estimation, see the recent review by Mullainathan and Spiess (2017). There is a small literature in ML referred to as "inverse reinforcement learning" (Ng et al., 2000) that has a similar approach to the structural estimation literature economics; this

ML literature has mostly operated independently without much reference to the earlier econometric literature. The literature attempts to learn "reward functions" (utility functions) from observed behavior in dynamic settings.

There are also other categories of ML models; for example, anomaly detection focuses on looking for outliers or unusual behavior, and is used, for example, to detect network intrusion, fraud, or system failures. Other categories that I will return to below are reinforcement learning (roughly, approximate dynamic programming) and multi-armed bandit experimentation (dynamic experimentation where the probabiity of selecting an arm is chosen to balance exploration and exploitation). These literatures often take a more explicitly causal perspective and thus are somewhat easier to relate to economic models, and so my general statements about the lack of focus on causal inference in ML must be qualified when discussing the literature on bandits.

Before proceeding, it is useful to highlight one other contribution of the ML literature. The contribution is computational rather than conceptual, but it has had such a large impact that it merits a short discussion. The technique is called stochastic gradient descent (SGD), and it is used in many different types of models, including the estimation of neural networks as well as large scale Bayesian models (e.g. Ruiz et al. (2017), discussed in more detail below). In short, stochastic gradient descent is a method for optimizing an objective function, such as a likelihood function or a generalized method of moments objective function, with respect to parameters. When the objective function is expensive to compute (e.g. because it requires numerical integration), stochastic gradient descent can be used. The main idea is that if the objective is the sum of terms, each term corresponding to a single observation, the gradient can be approximated by picking a single data point and using the gradient evaluated at that observation as an approximation to the average (over observations) of the gradient. This estimate of the gradient will be very noisy, but unbiased. The idea is that it is more effective to "climb a hill" taking lots of steps in a direction that is noisy but unbiased, than it is to take a small number of steps, each in the right direction, which is what happens if computational resources are focused on getting very precise estimates of the gradient of the objective at each step. SGD can lead to dramatic performance improvements, and thus enable the estimation of very complex models that would be intractable using traditional approaches.

3 Using Prediction Methods in Policy Analysis

3.1 Applications of Prediction Methods to Policy Problems in Economics

There have already been a number of successful applications of prediction methodology to policy problems. Kleinberg et al. (2015) have argued that there is a set of problems where off-the-shelf ML methods for prediction are the key part of important policy and decision problems. They use examples like deciding whether to do a hip replacement operation for an elderly patient; if you can predict based on their individual characteristics that they will die within a year, then you should not do the operation. Many Americans are incarcerated while awaiting trial; if you can predict who will show up for court, you can let more out on bail. ML algorithms are currently in use for this decision in a number of jurisdictions. Another natural example is credit scoring; an economics paper by Bjorkegren and Grissen (2015) uses ML methods to predict loan repayment using mobile phone data.

In other applications, Goel et al. (2016) use ML methods to examine stop-and-frisk laws, using observables of a police incident to predict the probability that a suspect has a weapon, and they show that blacks are much less likely than whites to have a weapon conditional on observables and being frisked. Glaeser et al. (2016a) helped cities design a contest to build a predictive model

that predicted health code violations in restaurants, in order to better allocate inspector resources. There is a rapidly growing literature using machine learning together with images from satellites and street maps to predict poverty, safety, and home values (see, e.g., Naik et al. (2017)). As Glaeser et al. (2016b) argue, there are a variety of applications of this type of prediction methodology. It can be used to compare outcomes over time at a very granular level, thus making it possible to assess the impact of a variety of policies and changes, such as neighborhood revitalization. More broadly, the new opportunities created by large-scale imagery and sensors may lead to new types of analyses of productivity and well-being.

Although prediction is often a large part of a resource allocation problem – people who will almost certainly die soon should not receive hip replacement surgery, and rich people should not receive poverty aid– Athey (2017) discusses the gap between identifying units that are at risk and those for whom intervention is most beneficial. Determining which units should receive a treatment is a causal inference question, and answering it requires different types of data than prediction. Either randomized experiments or natural experiments may be needed to estimate heterogeneous treatment effects and optimal assignment policies. In business applications, it has been common to ignore this distinction and focus on risk identification; for example, as of 2017, the Facebook advertising optimization tool provided to advertisers optimizes for consumer clicks, but not for the causal effect of the advertisement. The distinction is often not emphasized in marketing materials and discussions in the business world, perhaps because many practitioners and engineers are not well versed in the distinction between prediction and causal inference.

3.2 Additional Topics in Prediction for Policy Settings

Athey (2017) summarizes a variety of research questions that arise when prediction methods are taken into policy applications. A number of these have attracted initial attention in both ML and the social sciences, and interdisciplinary conferences and workshops have begun to explore these issues.

One set of questions concerns interpretability of models. There are discussions of what interpretability means, and whether simpler models have advantages. Of course, economists have long understood that simple models can also be misleading. In social sciences data, it is typical that many attributes of individuals or locations are positively correlated–parents' education, parents' income, child's education, and so on. If we are interested in a conditional mean function, and estimate $\hat{\mu}(x) = E[Y_i|X_i = x]$, using a simpler model that omits a subset of covariates may be misleading. In the simpler model, the relationship between the omitted covariates and outcomes is loaded onto the covariates that are included. Omitting a covariate from a model is not the same thing as controlling for it in an analysis, and it can sometimes be easier to interpret a partial effect of a covariate controlling for other factors, than it is to keep in mind all of the other (omitted) factors and how they covary with those included in a model. So, simpler models can sometimes be misleading; they may seem easy to understand, but the understanding gained from them may be incomplete or wrong.

One type of model that typically is easy to interpret and explain is a causal model. As reviewed in Imbens and Rubin (2015), the causal inference framework typically makes the estimand very precise–e.g. the average effect if a treatment were applied to a particular population, the conditional average treatment effect (conditional on some observable characteristics of individuals), or the average effect of a treatment on a subpopulation such as "compliers" (those whose treatment adoption is affected by an instrumental variable). Such parameters by definition give the answer to a well-defined question, and so the magnitudes are straightforward to interpret. Key parameters of "structural" models are also straightforward to interpret–they represent parameters of consumer

utility functions, elasticities of demand curves, bidder valuations in auctions, marginal costs of firms, and so on. An area for further research concerns whether there are other ways to mathematically formalize what it means for a model to be interpretable, or to analyze empirically the implications of interpretability. Yeomans et al. (2016) study empirically a related issue of how much people trust ML-based recommender systems, and why.

Another area that has attracted a lot of attention is the question of fairness and nondiscrimination, e.g. whether algorithms will promote discrimination by gender or race when used in settings like hiring, judicial decisions, or lending. There are a number of interesting questions that can be considered. One is, how can fairness constraints be defined? What type of fairness is desired? For example, if a predictive model is used to allocate job interviews based on resumes, there are two types of errors, type I and type II. It is straightforward to show that it is in general impossible to equalize both type I and type II errors across two different categories of people (e.g. men and women), so the analyst must choose which to equalize (or both). See Kleinberg et al. (2016) for further analysis and development of the inherent tradeoffs in fairness in predictive algorithms. Overall, the literature on this topic has grown rapidly in the last two years, and we expect that as ML algorithms are deployed in more and more contexts, the topic will continue to develop. My view is that it is more likely that ML models will help make resource allocation more rather than less fair; algorithms can absorb and effectively use a lot more information than humans, and thus are less likely than humans to rely on stereotypes. To the extent that unconstrained algorithms do have undesirable distributional consequences, it is possible to constrain the algorithms. Generally, algorithms can be trained to optimize objectives under constraints, and thus it may be easier to impose societal objectives on algorithms than on subjective decisions by humans.

A third issue that arises is stability and robustness, e.g. in response to variations in samples or variations in the environment. There are a variety of related ideas in machine learning, including domain adaptation (how do you make a model trained in one environment perform well in another environment), "transfer learning," and others. The basic concern is that ML algorithms do exhaustive searches across a very large number of possible specifications looking for the best model that predicts Y based on X. The models will find subtle relationships bewteen X and Y, some of which might not be stable across time or across environments. For example, for the last few years, there may be more videos of cats with pianos than dogs with pianos. The presence of a piano in a video may thus predict cats. However, pianos are not a fundamentnal feature of cats that holds across environments, and so if a fad arises where dogs play pianos, performance of an ML algorithm might suffer. This might not be a problem for a tech firm that re-estimates its models with fresh data daily, but predictive models are often used over much longer time periods in industry. For example, credit scoring models may be held fixed, since changing them makes it hard to assess the risk of the set of consumers who accept credit offers. Scoring models used in medicine might be held fixed over many years. There are many interesting methodological issues involved in finding models that have stable performance and are robust to changing circumstances.

Another issue is that of manipulability. In the application of using mobile data to do credit scoring, a concern is that consumers may be able to maniplate the data observed by the loan provider (Bjorkegren and Grissen, 2015). For example, if certain behavioral patterns help a consumer get a loan, the consumer can make it look like they have these behavioral patterns, for example visiting certain areas of a city. If resources are allocated to homes that look poor via satellite imagery, homes or villages can possibly modify the aerial appearance of their homes to make them look poorer. An open area for future research concerns how to constrain ML models to make them less prone to manipulability; Athey (2017) discusses some other examples of this.

There are also other considerations that can be brought into ML when it is taken to the field, including computational time, the cost of collecting and maintaining the "features" that are used

in a model, and so on. For example, technology firms sometimes make use of simplified models in order to reduce the response time for real-time user requests for information.

Overall, my prediction is that social scientists (and computer scientists at the intersection with social science), particularly economists and other social scientists, will contribute heavily to defining these types of problems and concerns formally, and proposing solutions to them. This will not only provide for better implementations of ML in policy, but will also provide rich fodder for interesting research.

4 A New Literature on Machine Learning and Causal Inference

Despite the fascinating examples of "off-the-shelf" or slightly modified prediction methods, in general ML prediction models are solving fundamentally different problems from much empirical work in social science, which instead focuses on causal inference. A prediction I have is that there will be an active and important literature combining ML and causal inference to create new methods, methods that harness the strengths of ML algorithms to solve causal inference problems. In fact, it is easy to make this prediction with confidence, because the movement is already well underway. Here I will highlight a few examples, focusing on those that illustrate a range of themes, while emphasizing that this is not a comprehensive survey or a thorough review.

To see the difference between prediction and causal inference, imagine that you have a data set that contains data about prices and occupancy rates of hotels. Prices are easy to obtain through price comparison sites, but occupancy rates are typically not made public by hotels. Imagine first that a hotel chain wishes to form an estimate of the occupancy rates of competitors, based on publicly available prices. This is a prediction problem: the goal is to get a good estimate of occupancy rates, where posted prices and other factors (such as events in the local area, weather, and so on) are used to predict occupancy. For such a model, you would expect to find that higher posted prices are predictive of higher occupancy rates, since hotels tend to raise their prices as they fill up (using yield management software). In contrast, imagine that a hotel chain wishes to estimate how occupancy would change if the hotel raised prices across the board (that is, if it reprogrammed the yield management software to shift prices up by 5% in every state of the world). This is a question of causal inference. Clearly, even though prices and occupancy are positively correlated in a typical dataset, we would not conclude that raising prices would increase occupancy. It is well known in the causal inference literature that the question about price increases cannot be answered simply by examining historical data without additional assumptions or structure. For example, if the hotel previously ran randomized experiments on pricing, the data from these experiments can be used to answer the question. More commonly, an analyst will exploit natural experiments or instrumental variables, where the latter are variables that are unrelated to factors that affect consumer demand, but that shift firm costs and thus their prices. Most of the classic supervised ML literature has little to say about how to answer this question.

To understand the gap between prediction and causal inference, recall that the foundation of supervised ML methods is that model selection (through, e.g., cross-validation) is carried out to optimize goodness of fit on a test sample. A model is good if and only if it predicts outcomes well in a test set. In contrast, a large body of econometric research builds models that substantially *r*educe the goodness of fit of a model in order to estimate the causal effect of, say, changing prices. If prices and quantities are positively correlated in the data, any model that estimates the true causal effect (quantity goes down if you change price) will not do as good a job fitting a test dataset that has the same joint distribution of prices and quantities as the training data. The place where the econometric model with a causal estimate would do better is at fitting what happens if the firm

actually changes prices at a given point in timeat doing counterfactual predictions when the world changes. Techniques like instrumental variables seek to use only some of the information that is in the data the clean or exogenous or experiment-like variation in pricesacrificing predictive accuracy in the current environment to learn about a more fundamental relationship that will help make decisions about changing price.

However, a new but rapidly growing literature is tackling the problem of using ML methods for causal inference. This new literature takes many of the strengths and innovations of ML methods, but applies them to causal inference. Doing this requires changing the objective function, since the ground truth of the causal parameter is not observed in any test set. Also as a consequence of the fact that the truth is not observed in a test set, statistical theory plays a more important role in evaluating models, since it is more difficult to directly assess how well a parameter estimates the truth, even if the analyst has access to an independent test set. Indeed, this discussion highlights one of the key ways in which prediction is substantially simpler than parameter estimation: for prediction problems, a prediction for a given unit (given its covariates) can be summarized in a single number, the predicted outcome, and the quality of the prediction can be evaluated on a test set without further modeling assumptions. Although the average squared prediction error of a model on a test set is a noisy estimate of the expected value of the mean squared error on a random test set (due to small sample size), the law of large numbers applies to this average and it converges quickly to the truth as the test set size increases. Since the standard deviation of the prediction error can also be easily estimated, it is straightforward to evaluate predictive models without imposing additional assumptions.

There are a variety of different problems that can be tackled with ML methods. An incomplete list of some that have gained early attention is given as follows. First, we can consider the type of identification strategy for identifying causal effects. Some that have received attention in the new ML/causal inference literature include:

1. Treatment randomly assigned (experimental data)
2. Treatment assignment unconfounded (conditional on covariates)
3. Instrumental variables
4. Panel data settings (including difference-in-difference designs)
5. Regression discontinuity designs
6. Structural models of individual or firm behavior

In each of those settings, there are different problems of interest:

1. Estimating average treatment effects (or a low-dimensional parameter vector)
2. Estimating heterogeneous treatment effects in simple models or models of limited complexity
3. Estimating heterogeneous treatment effects non-parametrically
4. Estimating optimal treatment assignment policies
5. Identifying groups of individuals that are similar in terms of their treatment effects

Although the early literature is already too large to summarize all of the contributions to each combination of identification strategty and problem of interest, it is useful to observe that at this

point there are entries in almost all of the "boxes" associated with different identification strategies, both for average treatment effects and heterogeneous treatment effects. Here, I will provide a bit more detail on a few leading cases that have received a lot of attention, in order to illustrate some key themes in the literature.

It is also useful to observe that even though the last four problems seem closely related, they are distinct, and the methods used to solve them as well as the issues that arise are distinct. These distinctions have not traditionally been emphasized as much in the literature on causal inference, but they matter more in environments with data-driven model selection, because each has a different objective and the objective function can make a big difference in determining the selected model in ML-based models. Issues of inference are also distinct, as we will discuss further below.

4.1 Average Treatment Effects

A large and important branch of the literature on causal inference focuses on estimation of average treatment effects under the unconfoundedness assumption. This assumption requires that potential outcomes (the outcomes a unit would experience in alternative treatment regimes) are independent of treatment assignment, conditional on covariates. In other words, treatment assignment is as good as random after controlling for covariates.

From the 1990s through the 2000s, a literature emerged about using semi-parametric methods to estimate average treatment effects (e.g. Bickel et al. (1993), focusing on an environment with a fixed number of covariates that is small relative to the sample size. The methods are semi-parametric in the sense that the goal is to estimate a low-dimensional parameter–in this case, the average treatment effect–without making parametric assumptions about the way in which covariates affect outcomes (e.g. Hahn (1998)). See Imbens and Wooldridge (2009); Imbens and Rubin (2015) for reviews. In the mid-2000s, Mark van der Laan and coauthors introduced and developed a set of methods called "targeted maximum likelihood" (van der Laan and Rubin, 2006). The idea is that maximum likelihood is used to estimate a low-dimensional parameter vector in the presence of high-dimensional nuisance parameters. The method allows the nuisance parameters to be estimated with techniques that have less well established properties or a slower convergence rate. This approach can be applied to estimate an average treatment effect parameter under a variety of identification assumptions, but importantly, it is an approach that can be used with many covariates.

An early example of the application of ML methods to causal inference in economics (see Belloni et al. (2014) and Chernozhukov et al. (2015) for reviews) uses regularized regression as an approach to deal with many potential covariates, in an environment where the outcome model is "sparse," meaning that only a small number of covariates actually affect mean outcome (but there are many observables, and the analyst does not know which ones are important). In an environment with unconfoundedness, since some covariates are correlated with both the treatment assignment and the outcome, if the analyst does not condition on them, the omission of the confounder will lead to a biased estimate of the treatment effect. BCH propose a double-selection method based on the LASSO. The LASSO is a regularized regression procedure, where a regression is estimated using an objective function that balances in-sample goodness of fit with a penalty term that depends on the sum of the magnitude of regression coefficients. This form of penalty leads many covariates to be assigned a coefficient of zero, effectively dropping them from the regression. The magnitude of the penalty parameter is selected using cross-validation. The authors observe that if LASSO is used in a regression of the outcome and both the treatment indicator and other covariates, the coefficient on the treatment indicator will be a biased estimate of the treatment effect, because confounders that have a weak relationship with the outcome but a strong relationship with the treatment assignment may be zeroed out by an algorithm whose sole objective is to select variables

that predict outcomes.

A variety of other methods have been proposed for combining machine learning and traditional econometric methods for estimating average treatment effects under the unconfoundedness assumption. Athey et al. (2016c) propose using a method they refer to as "residual balancing," building on work on balancing weights by Zubizarreta (2015). Their approach is similar to a "doubly-robust" method for estimating average treatment effects that proceeds by taking the average of the efficient score, which involves an estimate of the conditional mean of outcomes given covariates as well as the inverse of the estimated propensity score; however, the residual balancing replaces inverse propensity score weights with weights obtained using quadratic programming, where the weights are designed to achieve balance between the treatment and control group. The conditional mean of outcomes is estimated using LASSO. The main result in the paper is that this procedure is efficient and achieves the same rate of convergence as if the outcome model was known, under a few key assumptions. The most important assumption is that the outcome model is linear and sparse, although there can be a large number of covariates and the analyst does not need to have knowledge of which ones are important. The linearity assumption, while strong, allows the key result to hold in the absence of any assumptions about the structure of the process mapping covariates to the assignment, other than overlap (propensity score bounded strictly between 0 and 1, which is required for identification of average treatment effects). No other approach has been proposed that is efficient without assumptions on the assignment model. In settings where the assignment model is complex, simulations show that the method works better than alternatives, without sacrificing much in terms of performance on simpler models. Complex assignment rules with many weak confounders arise commonly in technology firms, where complex models are used to map from a user's observed history to assignments of recommendations, advertisements, and so on.

More recently, Chernozhukov et al. (2017) propose "double machine learning," a method analogous to Robinson (1988), using a semi-parametric residual-on-residual regression as a method for estimating average treatment effects under unconfoundedness. The idea is to run a non-parametric regression of outcomes on covariates, and a second non-parametric regression of the treatment indicator on covariates; then, the residuals from the first regression are regressed on the residuals from the second regression. In Robinson (1988), the non-parametric estimator was a kernel regression; the more recent work establishes that any ML method can be used for the non-parametric regression, so long as it is consistent and converges at the rate $n^{\frac{1}{4}}$.

A few themes are common to the latter two approaches. One is the importance of building on the traditional literature on statistical efficiency, which provides strong guidance on what types of estimators are likely to be successful, as well as the particular advantages of doubly robust methods for average treatment effect estimation. A second theme is that orthogonalization can work very well in practice–using machine learning to estimate flexibly the relationship between outcomes and treatment indicators, and covariates–and then estimating average treatment effects using residualized outcomes and/or residualized treatment indicators. The intuition is that in high dimensions, mistakes in estimating nuisance parameters are likely, but working with residualized variables makes the estimation of the average treatment effect orthogonal to errors in estimating nuisance parameters. I expect that this insight will continue to be utilized in the future literature.

4.2 Heterogeneous Treatment Effects and Optimal Policies

Another area of active research concerns the estimation of heterogeneity in treatment effects, where here we refer to heterogeneity with respect to observed covariates. For example, if the treatment is a drug, we can be interested in how the drug's efficacy varies with individual characteristics. Athey

and Imbens (2017) provides a more detailed review of a variety of questions that can be considered relating to heterogeneity; we will focus on a few here.

Treatment effect heterogeneity can be of interest either for basic scientific understanding (that can be used to design new policies or understand mechanisms), or as a means to the end of estimating treatment assignment policies that map from a user's characteristics to a treatment.

Starting with basic scientific understanding of treatment effects, another question concerns whether we wish to discover simple patterns of heterogeneity, or whether a fully nonparametric estimator for how treatment effects vary with covariates is desired. One approach to discovering simpler patterns is provided by Athey and Imbens (2016). This paper proposes to create a partition of the covariate space, and then estimate treatment effects in each element of the partition. The splitting rule optimizes for finding splits that reveal treatment effect heterogeneity. The paper also proposes sample splitting as a way to avoid the bias inherent in using the same data to discover the form of heterogeneity, and to estimate the magnitude of the heterogeneity. One sample is used to construct the partition, while a second sample is used to estimate treatment effects. In this way, the confidence intervals built around the estimates on the second sample have nominal coverage no matter how many covariates there are. The intuition is that since the partition is created on an independent sample, the partition used is completely unrelated to the realizations of outcomes in the second sample. In addition, the procedure used to create the partition penalizes splits that increase the variance of the estimated treatment effects too much. This, together with cross-validation to select tree complexity, ensures that the leaves don't get too small, and thus the confidence intervals have nominal coverage.

There have already been a wide range of applications of "causal trees" in applications ranging from medicine to economic field experiments. The methods allow the researcher to discover forms of heterogeneity that were not specified in a pre-analysis plan, without invalidating confidence intervals. The method is also easily "interpretable," in that for each element of the partition, the estimator is a traditional estimate of a treatment effect. However, it is important for researchers to recognize that just because, say, three covariates are used to describe an element of a partition (e.g. male individuals with income between $100,000 and $120,000 and 15 to 20 years of schooling), the average of all values of covariates will vary across partition elements. So, it is important not to draw conclusions about what covariates are *n*ot associated with treatment effect heterogeneity. This paper builds on earlier work on "model-based recursive partitioning (Zeileis et al., 2008), which looked at recursive partitioning for more complex models (general models estimated by maximum likelihood), but did not provide statistical properties (nor suggest the sample splitting which is a focus of Athey and Imbens (2016)). Asher et al. (2016) provide another related example of building classification trees for heterogeneity in GMM models.

In some contexts, a simple partition of the covariate space is most useful. In other contexts, it is desirable to have a fully non-parametric estimate of how treatment effects vary with covariates. In the traditional econometrics literature, this could be accomplished through kernel estimation or matching techniques; these methods have well-understood statistical properties. However, even though they work well in theory, in practice matching methods and kernel methods break down when there are more than a handful of covariates.

In Wager and Athey (2017), we introduce the idea of a "causal forest." Essentially, a causal forest is the average of a lot of causal trees, where trees differ from one another due to subsampling. Conceptually, a causal forest can be thought of as a version of a nearest neighbor matching method, but one where there is a data-driven approach to determine which dimensions of the covariate space are important to match on. The main technical results in this paper establish the first asymptotic normality results for random forests used for prediction; this result is then extended to causal inference. We also propose an estimator for the variance and prove its consistency, so that

confidence intervals can be constructed.

A key requirement for our results about random forests is that each individual tree is "honest," that is, we use different data to construct a partition of the covariate space from the data used to estimate treatment effects within the leaves. That is, we use sample splitting, similar to Athey and Imbens (2016). In the context of a random forest, all of the data is used for both "model selection" and estimation, as an observation that is in the partition-building subsample for one tree may be in the treatment effect estimation sample in another tree.

Athey et al. (2017e) extended the framework to analyze nonparametric parameter heterogeneity in any model where the parameter of interest can be estimated via GMM. The idea is that the random forest is used to construct a series of trees. Rather than estimating a model in the leaves of every tree, the algorithm instead extracts the weights implied by the forest. In particular, when estimating treatment effects for a particular value of X, we estimate a "local GMM" model, where observations close to X are weighted more heavily. How heavily? The weights are determined by the fraction of time an observation ended up in the same leaf during the forest creation stage. A subtlety in this project is that it is difficult to design general purpose, computationally light-weight "splitting rules" for constructing partitions according to the covariates that predict parameter heterogeneity. We provide a solution to that problem, and also provide a proof of asymptotic normality of estimates as well as an estimator for confidence intervals. The paper highlights the case of instrumental variables, and how the method can be used to find heterogeneity in treatment effect parameters estimated with instrumental variables. An alternative approach to estimating parameter heterogeneity in instrumental variables models was proposed by Hartford et al. (2016), who use an approach based on neural nets. General nonparametric theory is more challenging for neural nets.

The method of Athey et al. (2017e), "generalized random forests," can be used as an alternative to "traditional" methods such as local generalized method of moments or local maximum likelihood (Tibshirani and Hastie, 1987). Local methods such as local linear regression typically target a particular value of covariates, and use a kernel weighting function to weight nearby observations more heavily when running a regression. The insight in Athey et al. (2017e) is that the random forest can be re-interpreted as a method to generate a weighting function, and the forest-based weighting function can substitute for the kernel weighting function in a local linear estimation procedure. The advantages of the forest weighting function are that is data-adaptive as well as model-adaptive. It is data-adaptive in that covariates that are important for heterogeneity in parameters of interest are given more importance in determining what observations are "nearby." It is model-adaptive in that it focuses on heterogeneity in parameter estimates in a given model, rather than hetereogeneity in predicting the conditional mean of outcomes, as in a traditional regression forest.

The insight of Athey et al. (2017e) is more general and I expect it to reappear in other papers in this literature: anyplace in traditional econometrics where a kernel function might have been used, ML methods that perform better than kernels in practice may be substituted. However, the statistical and econometric theory for the new methods needs to be established in order to ensure that the ML-based procedure has desired properties such as asymptotic normality of parameter estimates. Athey et al. (2017e) does this for their generalized random forests for estimating heterogeneity in parameter estimates, and Hartford et al. (2016) use neural nets instead of kernels for semi-parametric instrumental variables; Chernozhukov et al. (2017) does this for their generalization of Robinson (1988) semi-parametric regression models.

There are also other possible approaches to estimating conditional average treatment effects when the structure of the heterogeneity is assumed to take a simple form, or when the analyst is willing to understand treatment effects conditioning only on a subset of covariates rather than attempting to condition on all relevant covariates. Targeted maximum likelihood (van der Laan and

Rubin, 2006) is one approach to this; more recently, Imai et al. (2013) proposed using LASSO to uncover heterogeneous treatment effects, while Künzel et al. (2017) proposes an ML approach using "meta-learners." It is important to note, however, that if there is insufficient data to estimate the impact of all relevant covariates, a model such as LASSO will tend to drop covariates (and their interactions) that are correlated with other included covariates, so that the included covariates "pick up" the impact of omitted covariates.

Finally, a motivating goal for understanding treatment effects is estimating optimal policy functions, that is, functions that map from the observable covariates of individuals to policy assignments. This problem has been recently studied in economics by, e.g., Kitagawa and Tetenov (2015), who focus on estimating the optimal policy from a class of potential policies of limited complexity. The goal is to select a policy function to minimize the loss from failing to use the (infeasible) ideal policy, referred to as the "regret" of the policy. Despite the general lack of research about causal inference in the ML literature, the topic of optimal policy estimation has received some attention. However, most of the ML literature focuses on algorithmic innovations, and does not exploit insights from the causal inference literature. An exception is that a line of research has incorporated the idea of propensity score weighting or doubly robust methods, although often without much reference to the statistics and econometrics literature. Examples of papers from the ML literature focused on policy learning include Strehl et al. (2010); Dudik et al. (2011); Li et al. (2012); Dudik et al. (2014); Li et al. (2014); Swaminathan and Joachims (2015); Jiang and Li (2016); Thomas and Brunskill (2016); Kallus (2017). One type of result in that literature establishes bounds on the regret of the algorithm. In Athey and Wager (2017), we show how bringing in insights from semi-parametric efficiency theory allows us to establish a tighter "regret bound" than the existing literature, thus narrowing down substantially the set of algorithms that might achieve the regret bound. This highlights the fact that the econometric theory literature has added value that has not been fully exploited in ML. Another unrelated observation is that, perhaps surprisingly, the econometrics of the problem of estimating optimal policy functions within a class of potential policies of limited complexity is quite different from the problem of estimating conditional average treatment effects, although of course the problems are related.

4.3 Contextual Bandits: Estimating Optimal Policies using Adaptive Experimentation

Above, I reviewed methods for estimating optimal policies mapping from individual covariates to treatment assignments. A growing literature based primarily in ML studies the problem of "bandits," which are algorithms that actively learn about which treatment is best. Online experimentation works yields large benefits when the setting is such that it is possible to quickly measure outcomes, and when there are many possible treatments. In the basic bandit problem when all units have identical covariates, the problem of "online experimentation," or "multi-armed bandits," asksthe question, how can experiments be designed to assign individuals to treatments as they arrive, using data from earlier individuals to determine the probabilities of assigning new individuals to each treatment, balancing the need for exploration against the desire for exploitation. That is, bandits balance the need to learn against the desire to avoid giving individuals suboptimal treatments. This type of online experimentation has been shown to yield reliable answers orders of magnitude faster than traditional randomized controlled trials in cases where there are many possible treatments (see e.g. Scott (2010)); the gain comes from the fact that treatments that are doing badly are effectively discarded, so that newly arriving units are instead assigned to the best candidates. When the goal is to estimate an optimal policy, it is not necessary to continue to allocate units to treatments that are fairly certain not to be optimal. Further, it is also not important

from the perspective of expected payoffs to statistically distinguish two very similar treatments. The literature has developed a number of heuristics for managing the explore-exploit tradeoff; for example, "Thompson sampling" allocates units to treatment arms in proportion to the estimated probability that each treatment arm is the best.

There is much less known about the setting where individuals have observed attributes, in which case the goal is to construct and evaluate personalized treatment assignment policies. This problem has been termed the "contextual bandit" problem, since treatment assignments are sensitive to the "context" (in this case, user characteristics). At first, the problem seems very challenging, because the space of possible policies is large and complex (each policy maps from user characteristics to the space of possible treatments). However, if the returns to each of the actions can be estimated as a function of individual attributes, a policy can be constructed by finding the action whose return is estimated to be highest, balanced against the need for exploration. Although there are a number of proposed methods for the contextual bandit problem in the literature already, there is relatively little known about how to select among methods and which ones are likely to perform best in practice. For example, the literature on optimal policy estimation suggests that particular approaches to policy estimation may work better than others.

In particular, there are a variety of choices a researcher must make when selecting a contextual bandit algorithm. These include the choice of the model that maps user characteristics to expected outcomes (where the literature has considered alternatives such as Ridge regression (Li et al., 2010), ordinary least squares (OLS) (Goldenshluger and Zeevi, 2013), generalized linear model (GLM) (Li et al., 2017), LASSO (Bastani and Bayati, 2015), and random forests (Dimakopoulou et al., 2017; Feraud et al., 2016)). Another choice concerns the heuristic used to balance exploration versus exploitation, with leading choices Thompson Sampling and Upper Confidence Bounds (UCB) Chapelle and Li (2011).

Dimakopoulou et al. (2017) highlights some issues that arise uniquely in the contextual bandit and that relate directly to the estimation issues that have been the focus of the literature on estimation of treatment effects (Imbens and Rubin, 2015). For example, the paper highlights the comparison between non-contextual bandits, where there will be many future individuals arriving with exactly the same context (since they all share the same context), and contextual bandits, where each unit is unique. The assignment of a particular individual thus contributes to learning for the future indirectly indirectly, since the future individuals will have different contexts (characteristics). The fact that the exploration benefits the future through a model of how contexts relates to outcomes changes the problem.

This discussion highlights a further theme for the connection between ML and causal inference: estimation considerations matter even more in the "small sample" settings of contextual bandits, where the assumption is that there is not enough data available to the policy maker to estimate perfectly the optimal assignment. However, we know from the econometrics literature that the small sample properties of different estimators can vary substantially across settings (Imbens and Rubin, 2015), making it clear that the best contextual bandit approach is likely to also vary across settings.

4.4 Robustness and Supplementary Analysis

In a recent review paper, Athey and Imbens (2017) highlights the importance of "supplementary analyses" for establishing the credibility of causal estimates in environments where crucial assumptions are not directly testable without additional information. Examples of supplementary analyses include placebo tests, whereby the analyst assses whether a given model is likely to find evidence of treatment effects even at times where no treatment effect should be found. One type of sup-

plementary analysis is a robustness measure. Athey and Imbens (2015) proposes to use ML-based methods to develop a range of different estimates of a target parameter (e.g. a treatment effect), where the range is created by introducing interaction effects between model parameters and covariates. The robustness measure is defined as the standard deviation of parameter estimates across model specifications. This paper provides one possible approach to ML-based robustness measures, but I predict that more approaches will develop over time as ML methods become more popular.

Another type of ML-based supplementary analysis, proposed by Athey et al. (2017c), uses ML-based methods to construct a measure of how challenging the confounding problem is in a particular setting. The proposed measure constructs an estimated conditional mean function for the outcome as well as an estimated propensity score, and then estimates the correlation between the two.

There is much more potential for supplementary analyses to be further developed; the fact that ML has well-defined, systematic algorithms for comparing a wide range of model specifications makes ML well suited for constructing additional robustness checks and supplementary analyses.

4.5 Panel Data and Difference-in-Difference Models

Another commonly used approach to identifying causal effects is to exploit assumptions about how outcomes vary across units and over time in panel data. In a typical panel data setting, units are not necessarily assigned to a treatment randomly, but all units are observed prior to some units being treated; the identifying assumption is that one or more untreated units can be used to provide an estimate of the counterfactual time trend that would have occurred for the treated units in the absence of the treatment. The simplest "difference-in-difference" case involves two groups and two time periods; more broadly, panel data may include many groups and many periods. Traditional econometric models for the panel data case exploit functional form assumptions, for example, assuming that a unit's outcome in a particular time period is an additive function of a unit effect, a time effect, an independent shock. The unit effect can then be inferred for treated units in the pre-treatment period, while the time effect can be inferred from the untreated units in the periods where some units receive the treatment. Note that this structure implies that the matrix of mean outcomes (with rows associated with units and columns associated with time) has a very simple structure: it has rank two.

There have been a few recent approaches bringing ML tools to the panel data setting. Doudchenko and Imbens (2016) develop an approach inspired by synthetic controls (pioneered by Abadie et al. (2010)), where a weighted average of control observations is used to construct the counterfactual untreated outcomes for treated units in treated periods. Doudchenko and Imbens (2016) propose using regularized regression to determine the weights, with the penalty parameter selected via cross-validation.

4.5.1 Factor Models and Matrix Completion

Another way to think about causal inference in a panel data setting is to consider a matrix completion problem; Athey et al. (2017a) propose taking such a perspective. In the ML literature, a matrix completion problem is one where there is an observed matrix of data (in our case, units and time periods), but some of the entries are missing. The goal is to provide the best possible prediction of what those entries should be. For the panel data application, we can think of the units and time periods where the units are treated as the missing entries, since we dont observe the counterfactual outcomes of those units in the absence of the treatment (this is the key bit of missing information for estimating the treatment effect).

Athey et al. (2017a) propose using a matrix version of regularized regression to find a matrix that well approximates the matrix of untreated outcomes (a matrix that has missing elements corresponding to treated units and periods). Recall that LASSO regression minimizes sum of squared errors in sample, plus a penalty term that is proportional to the sum of the magnitudes of the coefficients in the regression. We propose matrix regression that minimizes the sum of squared errors of all elements of the matrix, plus a penalty term proportional to the nuclear norm of the matrix. The nuclear norm is the sum of absolute values of the singular values of the matrix. A matrix that has a low nuclear norm is well approximated by a low rank matrix.

How do we interpret the idea that a matrix can be well approximated by a low rank matrix? A low rank matrix can be "factored" into the product of two matrices. In the panel data case, we can interpret such a factorization as incorporating a vector of latent characteristics of for each unit and a vector of latent characteristics of each period. The outcome of a particular unit in a particular period, if untreated, is approximately equal to the inner product of the unit's characteristics and the period characteristics. For example, if the data concerned employment at the county level, we can think of the counties as having outcomes that depend on the share of employment in different industries, and then each industry has common shocks in each period. So a county's latent characteristic would be the vector of industry shares, and the time characteristics would be industry shocks in a given period.

Athey et al. (2017a) show that the matrix completion approach reduces to commonly employed techniques in the econometrics literature when the assumptions needed for those approaches hold, but the matrix completion approach is able to model more complex patterns in the data, while allowing the data (rather than the analyst) to indicate whether time-series patterns within units, or cross-sectional patterns within a period, or a more complex combination, are more useful for predicting counterfactual outcomes.

The matrix completion approach can be linked to a literature that has grown in the last two decades in time series econometrics on factor models (see, e.g., Bai et al. (2008) for a review). The matrix factorization approach is similar, but rather than assuming that the true model has a fixed but unknown number of factors, the matrix completion approach simply looks for the best fit while penalizing the norm of the matrix. The matrix is well approximated by one with a small number of factors, but does not need to be exactly represented that way. Athey et al. (2017a) describe a number of advantages of the matrix completion approach, and also show that it performs better than existing panel data causal inference approaches in a range of settings.

4.6 Factor Models and Structural Models

Another important area of connection between machine learning and causal inference concerns more complex structural models. For decades, scholars working at the intersection of marketing and economics have built structural models of consumer choice, sometimes in dynamic environments, and used Bayesian estimation to estimate the model, often Markov Chain Monte Carlo. Recently, the ML literature has developed a variety of techniques that allow similar types of Bayesian models to be estimated at larger scale. These have been applied to settings such as textual analysis and consumer choices of, e.g., movies at Netflix. See, e.g., Blei et al. (2003a) and Blei and M. (2012). I expect to see much closer synergies between these two literatures in the future. For example, Athey et al. (2017b) builds on models of hierarchical Poisson factorization to create models of consumer demand, where a consumer's preference over thousands of products are considered simultaneously, but the consumer's choices in each product category are independent of one another. The model reduces the dimensionality of this problem by using a lower-dimensional factor representation of a consumer's mean utility as well as the consumer's price sensitivity for each product. The paper

establishes that substantial efficiency gains are possible by considering many product categories in parallel; it is possible to learn about a consumer's price sensitivity in one product using behavior in other products. The paper departs from the pure prediction literature in ML by evaluating and tuning the model based on how it does at predicting consumer responses to price changes, rather than simply on overall goodness of fit. In particular, the paper highlights that different models would be selected for the "goodness of fit" objective as opposed to the "counterfactual inference" objective. In order to achieve this goal, the paper analyzes goodness of fit in terms of predicting changes in demand for products before and after price changes, after providing evidence that the price changes can be treated as natural experiments after conditioning on week effects (price changes always occur mid-week). The paper also demonstrates the benefits of personalized prediction, versus more standard demand estimation methods. Thus, the paper again highlights the theme that for causal inference, the objective function differs from standard prediction.

With more scalable computational methods, it becomes possible to build much richer models with much less prior information about products. Ruiz et al. (2017) analyzes consumer preferences for bundles selected from over 5000 items in a grocery store, *w*ithout incorporating information about which items are in the same category. Thus, the model uncovers whether items are substitutes or complements. Since there are 2^{5000} bundles when there are 5000 products, in principle each individual consumer's utility function has 2^{5000} parameters. Even if we restrict the utility function to have only pairwise interaction effects, there are still millions of parameters of a consumer's utility function over bundles. Ruiz et al. (2017) uses a matrix factorization approach to reduce the dimensionality of the problem, factorizing the mean utilities of the items, the interaction effects among items, and the user's price sensitivity for the items. Price and availability variation in the data allows the model to distinguish correlated preferences (some consumers like both coffee and diapers) from complementarity (tacos and taco shells are more valuable together). In order to further simplify the analysis, the model assumes that consumers are boundedly rational when they make choices, and consider the interactions among products as the consumer sequentially adds items to the cart. The alternative–that the consumer considers all 2^{5000} bundles and optimizes among them–does not seem plausible. Incorporating human computational constraints into structural models thus appears to be another potential fruitful avenue at the intersection of ML and economics. In the computational algorithm for Ruiz et al. (2017), we rely on a technique called variational inference to approximate the posterior distribution, as well as the technique stochastic gradient descent (described in detail above) to find the parameters that provide the best approximation.

In another application of similar methodology, Athey et al. (forthcoming) analyzes consumer choices over lunchtime restaurants using data from a sample of several thousand mobile phone users in the San Francisco Bay Area. The data is used to identify users typical morning location, as well as their choices of lunchtime restaurants. We build a model where restaurants have latent characteristics (whose distribution may depend on restaurant observables, such as star ratings, food category, and price range), and each user has preferences for these latent characteristics, and these preferences are heterogeneous across users. Similarly, each item has latent characteristics that describe users willingness to travel to patronize the restaurant, and each user has individual-specific preferences for those latent characteristics. Thus, both users willingness to travel and their base utility for each restaurant vary across user-item pairs. To make the estimation computationally feasible, we build on the methods of Ruiz et al. (2017). We show that our model performs better than more standard competing models such as multinomial logit and nested logit models, in part due to the personalization of the estimates. We demonstrate in particular that our model performs better when predicting consumer responses to restaurant openings and closings, and we analyze how consumers re-allocate their demand after a restaurant closes to nearby restaurants versus more distant restaurants with similar characteristics. Since there are several hundred restaurant openings

and closings in the data, we are able to use the large number of "natural experiments" in the data to assess performance of the model. Finally, we show how the model can be used to analyze questions involving counterfactuals such as what type of restaurant would attract the most consumers in a given location.

Another recent paper that makes use of factorization in the context of a structural model of consumer demand is Wan et al. (2017). This paper builds a model of consumer choice that includes choices over categories, purchases within a category, and quantity to purchase. The model allows for individual heterogeneity in preferences, and uses factorization techniques to estimate the model.

5 Broader Predictions About the Impact of Machine Learning on Economics

My prediction is that there will be substantial changes in how empirical work is conducted; indeed, it is already happening, and so this prediction already can be made with a high degree of certainty. I predict that a number of changes will emerge, summarized as follows:

1. Adoption of off-the-shelf ML methods for their intended tasks (prediction, classification, and clustering, e.g. for textual analysis)
2. Extensions and modifications of prediction methods to account for considerations such as fairness, manipulability, and interpretability
3. Development of new econometric methods based on machine learning designed to solve traditional social science estimation tasks
4. No fundamental changes to theory of identification of causal effects
5. Incremental progress to identification and estimation strategies for causal effects that exploit modern data settings including large panel datasets and environments with many small experiments
6. Increased emphasis on model robustness and other supplementary analysis to assess credibility of studies
7. Adoption of new methods by empiricists at large scale
8. Revival and new lines of research in productivity and measurement
9. New methods for the design and analysis of large administrative data, including merging these sources and privacy-preserving methods
10. Increase in interdisciplinary research
11. Changes in organization, dissemination, and funding of economic research
12. Economist as engineer engages with firms, government to design and implement policies in digital environment
13. Design and implementation of digital experimentation, both one-time and as an ongoing process, including "multi-armed bandit" experimentation algorithms, in collaboration with firms and government

14. Research on developing high-quality metrics that can be measured quickly, in order to facilitate rapid incremental innovation and experimentation

15. Increased use of data analysis in all levels of economics teaching; increase in interdisciplinary data science programs

16. Research on the impact of AI and ML on economy

This article has al discussed the first three predictions in some detail; I will now discuss each of remaining predictions in turn.

First, as emphasized in the discussion about the benefits from using ML, ML is a very powerful tool for data-driven model selection. Getting the best flexible functional form to fit data is very important for many reasons; for example, when the researcher assumes that treatment assignment is unconfounded, it is still crucial to flexibly control for covariates, and a vast literature has documented that modeling choices matter. A theme highlighted in this paper is that ML can be used any time that semi-parametric methods might have been used in the traditional econometrics literature. However, finding the best functional form is a distinct concern from whether an economic parameter would be identified with sufficient data. Thus, there is no obvious benefit from ML in terms thinking about identification issues.

However, the types of datasets that are becoming widely available due to digitization suggest new identification questions. For example, it is common for there to be frequent changes in algorithms in ecommerce platforms. These changes in algorithms create variation in user experiences (as well as in seller experiences in platforms and marketplaces). Thus, a typical user or seller may experience a large number of changes, each of which has modest effects. There are open questions about what can be learned in such environments. From an estimation perspective, there is also room to develop ML-inspired algorithms that take advantage of the many sources of variation experienced by market participants. In my 2012 Fisher Schultz lecture, I illustrated the idea of using randomized experiments conducted by technology firms as instruments for estimating position effects for sponsored search advertisements. This idea has since been exploited more fully by others, e.g. Goldman and Rao (2014), but many open questions remain about the best ways to use the information in such datasets.

Digitization is also leading to the creation of many panel datasets that record individual behavior at relatively high frequency over a period of time. There are many open questions about how to make the best use of rich panel data. Above, we discussed several new papers at the intersection of ML and econometrics that made use of panel data (e.g. Athey et al. (2017a)), but I predict that this literature will grow dramatically over the next few years.

There are many reasons that empiricists will adopt ML methods at scale. First, many ML methods simplify a variety of arbitrary choices analysts needed to make. In larger and more complex datasets, there are many more choices. Each choice must be documented, justified, and serves at a potential source of criticism of a paper. When systematic, data-driven methods are available, research can be made more principled and systematic, and there can be objective measures against which these choices can be evaluated. Indeed, it would really be impossible for a researcher using traditional empirical methods to fully document the process by which the model specification was selected; in contrast, algorithmic selection (when the algorithm is given the correct objective for the problem) has superior performance while simultaneously being reproducible. Second, one way to conceptualize ML algorithms is that they perform like automated research assistants–they work much faster and more effectively than traditional research assistants at exploring modeling choices, yet the methods that have been customized for social science applications also build in protections so that, for example, valid confidence intervals can be obtained. Although it is crucial to consider

carefully the objective that the algorithms are given, in the end they are highly effective. Thus, they help resolve issues like "p-value hacking" by giving researchers the best of both worlds–superior performance as well as correct p-values that take into account the specification selection process. Third, in many cases, new results can be obtained. For example, if an author has run a field experiment, there is no reason not to search for heterogeneous treatment effects using methods such as those in Athey and Imbens (2016). The method ensures that valid confidence intervals can be obtained for the resulting estimates of treatment effect heterogeneity.

Alongside the adoption of ML methods for old questions, new questions and types of analyses will emerge in the fields of productivity and measurement. Some examples of these have already been highlighted, such as the ability to measure economic outcomes at a granular level over a longer period of time, through, e.g. imagery. Glaeser et al. (2018) provides a nice overview of how big data and ML will affect urban economics as a field as well as the operational efficiency of cities. More broadly, as governments begin to absorb high-frequency, granular data, they will need to grapple with questions about how to maintain the stability of official statistics in a world where the underlying data changes rapidly. New questions will emerge about how to architect a system of measurement that takes advantage of high-frequency, noisy, unstable data but yields statistics whose meaning and relationship with a wide range of economic variables remains stable. Firms will face similar problems as they attempt to forecast outcomes relevant to their own businesses using noisy, high-frequency data. The emerging literature in academics, government, and industry on "now-casting" in macroeconomics (e.g. (Banbura et al., 2013) and ML begins to address some, but not all, of these issues. We will also see the emergence of new forms of descriptive analysis, some inspired by ML. Examples of these include techniques for describing association, e.g., people who do A also do B; as well as interpretations and visualizations of the output of unsupervised ML techniques such as matrix factorization, clustering, and so on. Economists are likely to refine these methods to make them more directly useful quantiatively, and for business and policy decisions.

More broadly, the ability to use predictive models to measure economic outcomes at high granularity and fidelity will change the types of questions we can ask and answer. For example, imagery from satellites or Google's street view can be used in combination with survey data to train models that can be used to produce estimates of economic outcomes at the level of the individual home, either within the U.S. or in developing countries where administrative data quality can be problematic (e.g. Jean et al. (2016), Engstrom et al. (2017), Naik et al. (2014)).

Another area of transformation for economics will be in the design and analysis of large-scale administrative data sets. We will see attempts to bring together disparate sources to provide a more complete view of individuals and firms. The behavior of individuals in the financial world, the physical world, and the digital world will be connected, and in some cases ML will be needed simply to match different identities from different contexts onto the same individual. Further, we will observe behavior of individuals over time, often with high-frequency measurements. For example, children will leave digital footprints throughout their education, ranging from how often they check their homework assignments, the assignments themselves, comments from teachers, and so on. Children will interact with adaptive systems that change the material they receive based on their previous engagement and performance. This will create the need for new statistical methods, building on existing ML tools, but where the methods are more tailored to a panel data setting with significant dynamic effects (and possibly peer effects as well; see for some recent statistical advances designed around analyzing large scale network data Ugander et al. (2013), Athey et al. (2016b), Eckles et al. (2016)).

Another area of future research concerns how to analyze personal data without compromising user privacy. There is a literature in computer science around querying data while preserving privacy; the literature is referred to as "differential privacy." Some recent research has brought

together the computer science literature with questions about estimating statistical models; see, e.g., Komarova et al. (2015).

I also predict a substantial increase in interdisciplinary work. Computer scientists and engineers may remain closer to the frontier in terms of algorithm design, computational efficiency, and related concerns. As I will expand on further in a moment, academics of all disciplines will be gaining a much greater ability to intervene in the environment in a way that facilitates measurement and caual inference. As digital interactions and digital interventions expand across all areas of society, from education to health to government services to transportation, economists will collaborate with domain experts in other areas to design, implement, and evaluate changes in technology and policy. Many of these digital interventions will be powered by ML, and ML-based causal inference tools will be used to estimate personalized treatment effects of the interventions and design personalized treatment assignment policies.

Alongside the increase in interdisciplinary work, there will also be changes to the organization, funding, and dissemination of economics research. Research on large datasets with complex data creation and analysis pipelines can be labor intensive, and also require specialized skills. Scholars who do a lot of complex data analysis with large datasets have already begun to adopt a "lab" model more similar to what is standard today in computer science and many natural sciences. A lab might include a post-doctoral fellow, multiple Ph.D. students, pre-doctoral fellows (full-time research assistants between their bachelors and Ph.D.), undergraduates, and possibly full-time staff. Of course, labs of this scale are expensive, and so the funding models for economics will need to adapt to address this reality. One concern is inequality of access to resources required to do this type of research, given that it is expensive enough that it cannot be supported given traditional funding pools for more than a small fraction of economists at research universities.

Within a lab, we will see increased adoption of collaboration tools such as those used in software firms; tools include GitHub (for collaboration, version control, and dissemination of software) as well as communication tools; for example, my generalized random forest software is available as an open source package on github at `http://github.com/swager/grf`, and users report issues through the GitHub, and can submit request to pull in proposed changes or additions to the code.

There will also be an increased emphasis on documenation and reproducibility, which are necessary to make a large lab function. This will happen even as some data sources remain proprietary. "Fake" data sets will be created that allow others to run a lab's code and replicate the analysis (except not on the real data). As an example of institutions created to support the lab model, both Stanford GSB and the Stanford Institute for Economic Policy Research have "pools" of predoctoral fellows that are shared among faculty; these programs provide mentorship, training, the opportunity to take one class each quarter, and they also are demographically more diverse than graduate student populations. The predoctoral fellows have a special form of student status within Stanford. Other public and private sector research groups have also adopted similar programs, with Microsoft Research-New England an early innovator in this area, while individual researcheres at universities like Harvard and MIT have also been making use of predoctoral research assistants for a number of years.

We will also see changes in how economists engage with government, industry, education, and health. The concept of the "economist as engineer" promoted by market design experts including Robert Wilson, Paul Milgrom, and Al Roth (Roth, 2002) and even "economist as plumber" (Duflo, 2017) will move beyond the fields of market design and development. As digitization spreads across application areas and sectors of the economy, it will bring opportunities for economists to develop and implement policies that can be delivered digitially. Farming advice, online education, health information and information, government service provision, government collections, personalized resource allocation–all of these create opportunities for economists to propose policies, design the

delivery and implementation of the policy including randomization or staggered roll-outs to enable evaluation, and to remain involved through successive rounds of incremental improvement for adopted policies. Feedback will come more quickly and there will be more opportunities to gather data, adapt, and adjust. Economists will be involved in improving operational efficiency of government and industry, reducing costs, and improving outcomes.

ML methods, when deployed in practice in industry, government, education and health, lend themselves to incremental improvement. Standard practice in the technology industry is to evaluate incremental improvements through randomized controlled trials. Firms like Google and Facebook do 10,000 or more randomized controlled trials of incremental improvements to ML algorithms every year. An emerging trend is to build the experimentation right into the algorithm, using "bandit" techniques. As described in more detail above, "multi-armed bandit" is a term for an algorithm that balances exploration and learning against exploiting information that is already available about which alternative treatment is best. Bandits can be dramatically faster than standard randomized controlled experiments (see, e.g., the description of bandits on Google's web site: https://support.google.com/analytics/answer/2844870?hl=en), because they have a different goal: the goal is to learn what the best alternative is, not to accurately estimate the average outcome for each alternative, as in a standard randomized controlled trial.

Implementing bandit algorithms requires the statistical analysis to be embedded in the system that delivers the treatments. For example, a user might arrive at a web site. Based on the user's characteristics, a contextual bandit might randomize among treatment arms in proportion to the current best estimate of the probability that each arm is optimal for that user. The randomization would occur "on the fly" and thus the software for the bandit needs to be integrated with the software for delivering the treatments. This requires a deeper relationship between the analyst and the technology than a scenario where an analyst analyzes historical data "offline" (that is, not in real time).

Balancing exploration and exploitation involves fundamental economic concepts about optimization under limited information and resource constraints. Bandits are generally more efficient and I predict they will come into much more widespread use in practice. In turn, that will create opportunities for social scientists to optimize interventions much more effectively, and to evaluate a large number of possible alternatives faster and with less inefficiency. More broadly, statistical analysis will come to be commonly placed in a longer-term context where information accumulates over time.

Beyond bandits, other themes include combining experimental and observational data to improve precison of estimates (see, e.g., Peysakhovich and Lada (2016)), and making use of large numbers of related experiments when drawing conclusions.

Optimizing ML algorithms require an objective or an outcome to optimize for. In an environment with frequent and high-velocity experimentation, measures of success that can be obtained in a short time frame are needed. This leads to a substantively challenging problem: what are good measures that are related to long-term goals, but can be measured in the short term, and are responsive to interventions? Economists will get involved in helping define objectives and constructing measures of success that can be used to evaluate incremental innovation. One area of research that is receiving renewed attention is the topic of "surrogates," a name for intermediate measures that can be used in place of long-term outcomes; see, e.g., Athey et al. (2016a). Economists will also place renewed interest on designing incentives that counterbalance the short-term incentives created by short-term experimentation.

All of these changes will also affect teaching. Anticipating the digital transformation of industry and government, undergraduate exposure to programming and data will be much higher than it was ten years ago. Within 10 years, most undergraduates will enter college (and most MBAs will

enter business school) with extensive coding experience obtained from elementary through high school, summer camps, online education, and internships. Many will take coding and data analysis in college, viewing these courses as basic preparation for the workforce. Teaching will need to change to complement the type of material covered in these other classes. In the short run, more students may arrive at econometrics classes thinking about data analysis from the perspective that all problems are prediction or classification problems. They may have a cookbook full of algorithms, but little intuition for how to use data to solve real-world problems or answer business or public policy questions. Yet, such questions are prevalent in the business world: firms want to know the return on investment on advertising campaigns,[2], the impact of changing prices or introducing products, and so on. Economic education will take on an important role in educating students in how to use data to answer questions. Given the unique advantages economics as a discipline has at these methods and approaches, many of the newly created data science undergraduate and graduate programs will bring in economists and other social scientists, creating an increased demand for teaching from empirical economists and applied econometricians. We will also see more interdisciplinary majors; Duke and MIT both recently announced joint degrees between computer science and economics. There are too many newly created data science master's programs to mention, but a key observation is that while early programs most commonly have emerged from computer science and engineering, I predict that these programs will over time incorporate more social science, or else adopt and teach social science empirical methods themselves. Graduates entering the workforce will need to know basic empirical strategies like difference-in-differences that often arise in the business world (e.g. some consumers or areas are exposed to a treatment and not others, and there are important seasonality effects to control for).

A final prediction is that we will see a lot more research into the societal impacts of machine learning. There will be large-scale, very important regulatory problems that need to be solved. Regulating the transportation infrastructure around autonomous vehicles and drones is a key example. These technologies have the potential to create enormous efficiency. Beyond that, reducing transportation costs substantially effectively increases the supply of land and housing in commuting distance of cities, thus reducing housing costs for people who commute into cities to provide services for wealthier people. This type of reduction in housing cost would be very impactful for the cost of living for people providing services in cities, which could reduce effective inequality (which may otherwise continue to rise). But there are a plethora of policy issues that need to be addressed, ranging from insurance and liability, to safety policy, to data sharing, to fairness, to competition policy, and many others. Generally, the problem of how regulators approach algorithms that have enormous public impact is not at all worked out. Are algorithms regulated on outcomes, or on procedures and processes? How should regulators handle equilibrium effects, for example, if one autonomous vehicle system makes a change to its driving algorithms, how is that communicated to others? How can we avoid problems that have plagued personal computer software, where bugs and glitches are common following updates? How do we deal with the fact that having an algorithm used by 1% of cars does not prove it will work when used by 100% of cars, due to interaction effects?

Another industry where regulation of ML is already becoming problematic is financial services. Financial service regulation traditionally concerned processes, rules, and regulations. There is not currently a framework for cost-benefit analysis, or deciding how to test and evaluate algorithms, and determining an acceptable error rate. For algorithms that might have an effect on the economy, how do we assess systematic risks? These are fruitful areas for future research as well. And of

[2]For example, several large technology companies employ economists with PhD's from top universities who specialize in evaluating and allocating advertising spend for hundreds of millions of dollars of expenditures; see Lewis and Rao (2015) for a description of some of the challenges involved.

course, there are crucial questions about how ML will affect the future of work, as ML is used across wider and wider swaths of the economy.

We will also see experts in the practice of machine learning and AI collaborate with different subfields of economics in evaluating the impact of AI and ML on the economy.

Summarizing, I predict that economics will be profoundly transformed by AI and ML. We will build more robust and better optimized statistical models, and we will lead the way in modifying the algorithms to have other desirable properties, ranging from protection against over-fitting and valid confidence intervals, to fairness or non-manipulability. The kinds of research we do will change; in particular, a variety of new research areas will open up, with better measurement, new methods, and different substantive questions. We will grapple with how to re-organize the research process, which will have increased fixed costs and larger scale research labs, for those who can fund it. We will change our curriculum and take an important seat at the table in terms of educating the future workforce with empirical and data science skills. And, we will have a whole host of new policy problems created by ML and AI to study, including the issues experienced by parts of the workforce who need to transition jobs when their old jobs are eliminated due to automation.

6 Conclusions

It is perhaps easier than one might think to make predictions about the impact of ML on economics, since many of the most profound changes are well underway. There are exciting and vibrant research areas emerging, and dozens of applied papers making use of the methods. In short, I believe there will be an important transformation.

At the same time, the automation of certain aspects of statistical algorithms does not change the need to worry about the things that economists have always worried about: is a causal effect really identified from the data; are all confounders measured; what are effective strategies for identifying causal effects; what considerations are important to incorporate in a particular applied setting; defining outcome metrics that reflect overall objectives; constructing valid confidence intervals; and many others. As ML automates some of the routine tasks of data analysis, it becomes all the more important for economists to maintain their expertise at the art of credible and impactful empirical work.

References

A. Abadie, A. Diamond, and J. Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of californias tobacco control program. *Journal of the American statistical Association*, 105(490):493–505, 2010.

S. Asher, D. Nekipelov, P. Novosad, and S. Ryan. Classification Trees for Heterogeneous Moment-Based Models. Technical report, National Bureau of Economic Research, Cambridge, MA, dec 2016. URL http://www.nber.org/papers/w22976.pdf.

S. Athey. Beyond prediction: Using big data for policy problems. *Science*, 355(6324):483–485, 2017.

S. Athey and P. A. Haile. Nonparametric approaches to auctions. *Handbook of econometrics*, 6: 3847–3965, 2007.

S. Athey and G. Imbens. A measure of robustness to misspecification. *The American Economic Review*, 105(5):476–480, 2015.

S. Athey and G. Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

S. Athey and G. W. Imbens. The state of applied econometrics: Causality and policy evaluation. *The Journal of Economic Perspectives*, 31(2):3–32, 2017.

S. Athey and S. Wager. Efficient policy estimation. *arXiv preprint arXiv:1702.02896*, 2017. URL https://arxiv.org/abs/1702.02896.

S. Athey, J. Levin, and E. Seira. Comparing open and sealed bid auctions: Evidence from timber auctions. *The Quarterly Journal of Economics*, 126(1):207–257, 2011.

S. Athey, D. Coey, and J. Levin. Set-asides and subsidies in auctions. *American Economic Journal: Microeconomics*, 5(1):1–27, 2013.

S. Athey, R. Chetty, G. Imbens, and H. Kang. Estimating treatment effects using multiple surrogates: The role of the surrogate score and the surrogate index. *arXiv preprint arXiv:1603.09326*, 2016a.

S. Athey, D. Eckles, and G. W. Imbens. Exact p-values for network interference. *Journal of the American Statistical Association*, (just-accepted), 2016b.

S. Athey, G. W. Imbens, and S. Wager. Approximate residual balancing: De-biased inference of average treatment effects in high dimensions. *arXiv preprint arXiv:1604.07125*, 2016c.

S. Athey, M. Bayati, N. Doudchenko, G. Imbens, and K. Khosravi. Matrix completion methods for causal panel data models. *arXiv preprint arXiv:1710.10251*, 2017a.

S. Athey, D. Blei, R. Donnelly, and F. Ruiz. Counterfactual inference for consumer choice across many product categories. 2017b.

S. Athey, G. Imbens, T. Pham, and S. Wager. Estimating average treatment effects: Supplementary analyses and remaining challenges. *American Economic Review*, 107(5):278–81, 2017c.

S. Athey, M. M. Mobius, and J. Pál. The impact of aggregators on internet news consumption. 2017d.

S. Athey, J. Tibshirani, and S. Wager. Generalized random forests. *arXiv preprint arXiv:1610.01271*, 2017e. URL https://arxiv.org/abs/1610.01271.

S. Athey, D. M. Blei, R. Donnelly, F. J. Ruiz, and T. Schmidt. Estimating heterogeneous consumer preferences for restaurants and travel time using mobile location data. forthcoming.

J. Bai, S. Ng, et al. Large dimensional factor analysis. *Foundations and Trends® in Econometrics*, 3(2):89–163, 2008.

M. Banbura, D. Giannone, M. Modugno, and L. Reichlin. Now-casting and the real-time data flow. 2013.

H. Bastani and M. Bayati. Online decision-making with high-dimensional covariates. 2015.

A. Belloni, V. Chernozhukov, and C. Hansen. High-dimensional methods and inference on structural and treatment effects. *The Journal of Economic Perspectives*, 28(2):29–50, 2014.

P. J. Bickel, C. A. Klaassen, , Y. Ritov, J. Klaassen, J. A. Wellner, and Y. Ritov. *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins University Press Baltimore, 1993.

D. Bjorkegren and D. Grissen. Behavior revealed in mobile phone usage predicts loan repayment. 2015.

D. M. Blei and D. M. Probabilistic topic models. *Communications of the ACM*, 55(4):77, apr 2012. ISSN 00010782. doi: 10.1145/2133806.2133826. URL http://dl.acm.org/citation.cfm?doid=2133806.2133826.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003a. ISSN ISSN 1533-7928. URL http://www.jmlr.org/papers/v3/blei03a.html.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003b.

O. Chapelle and L. Li. An empirical evaluation of thompson sampling. *Conference on Neural Information Processing Systems*, 2011.

V. Chernozhukov, C. Hansen, and M. Spindler. Valid Post-Selection and Post-Regularization Inference: An Elementary, General Approach. jan 2015. doi: 10.1146/annurev-economics-012315-015826. URL http://arxiv.org/abs/1501.03430http://dx.doi.org/10.1146/annurev-economics-012315-015826.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, and W. Newey. Double/Debiased/Neyman Machine Learning of Treatment Effects. jan 2017. URL http://arxiv.org/abs/1701.08687.

M. Dimakopoulou, S. Athey, and G. Imbens. Estimation considerations in contextual bandits. *arXiv*, 2017.

N. Doudchenko and G. W. Imbens. Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. Technical report, National Bureau of Economic Research, 2016.

M. Dudik, J. Langford, and L. Li. Doubly robust policy evaluation and learning. *International Conference on Machine Learning*, 2011.

M. Dudik, D. Erhan, J. Langford, and L. Li. Doubly robust policy evaluation and optimization. *Statistical Science*, 2014.

E. Duflo. The economist as plumber. Technical report, National Bureau of Economic Research, 2017.

D. Eckles, B. Karrer, J. Ugander, L. Adamic, I. Dhillon, Y. Koren, R. Ghani, P. Senator, J. Bradley, and R. Parekh. Design and Analysis of Experiments in Networks: Reducing Bias from Interference. *Journal of Causal Inference*, 0(0):1–62, jan 2016. ISSN 2193-3677. doi: 10.1515/jci-2015-0021. URL https://www.degruyter.com/view/j/jci.ahead-of-print/jci-2015-0021/jci-2015-0021.xml.

N. Egami, C. Fong, J. Grimmers, M. Roberts, and B. Stewart. How to Make Causal Inferences Using Text. 2016. URL https://polmeth.polisci.wisc.edu/Papers/ais.pdf.

R. Engstrom, J. Hersh, and D. Newhouse. Poverty from space. 2017.

R. Feraud, R. Allesiardo, T. Urvoy, and F. Clerot. Random forest for the contextual bandit problem. *International Conference on Artificial Intelligence and Statistics*, 2016.

E. L. Glaeser, A. Hillis, S. D. Kominers, and M. Luca. Predictive cities crowdsourcing city government: Using tournaments to improve inspection accuracy. *The American Economic Review*, 106 (5):114–118, 2016a.

E. L. Glaeser, S. D. Kominers, M. Luca, and N. Naik. Big data and big cities: The promises and limitations of improved measures of urban life. *Economic Inquiry*, 2016b.

E. L. Glaeser, S. D. Kominers, M. Luca, and N. Naik. Big data and big cities: The promises and limitations of improved measures of urban life. *Economic Inquiry*, 56(1):114–137, 2018.

S. Goel, J. M. Rao, R. Shroff, et al. Precinct or prejudice? understanding racial disparities in new york citys stop-and-frisk policy. *The Annals of Applied Statistics*, 10(1):365–394, 2016.

A. Goldenshluger and A. Zeevi. A linear response bandit problem. *Stochastic Systems*, 2013.

M. Goldman and J. M. Rao. Experiments as instruments: heterogeneous position effects in sponsored search auctions. 2014.

P. Gopalan, J. M. Hofman, and D. M. Blei. Scalable recommendation with hierarchical poisson factorization. In *UAI*, pages 326–335, 2015.

J. Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, pages 315–331, 1998.

J. Hartford, G. Lewis, and M. Taddy. Counterfactual Prediction with Deep Instrumental Variables Networks. 2016. URL https://arxiv.org/pdf/1612.09596.pdf.

K. Imai, M. Ratkovic, et al. Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443–470, 2013.

G. W. Imbens and D. B. Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015.

G. W. Imbens and J. M. Wooldridge. Recent developments in the econometrics of program evaluation. *Journal of economic literature*, 47(1):5–86, 2009.

N. Jean, M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794, 2016.

N. Jiang and L. Li. Doubly robust off-policy value evaluation for reinforcement learning. *International Conference on Machine Learning*, 2016.

N. Kallus. Balanced policy evaluation and learning. *arXiv*, 2017.

T. Kitagawa and A. Tetenov. Who should be treated? Empirical welfare maximization methods for treatment choice. Technical report, Centre for Microdata Methods and Practice, Institute for Fiscal Studies, 2015.

J. Kleinberg, J. Ludwig, S. Mullainathan, and Z. Obermeyer. Prediction policy problems. *American Economic Review*, 105(5):491–95, 2015.

J. Kleinberg, S. Mullainathan, and M. Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

T. Komarova, D. Nekipelov, and E. Yakovlev. Estimation of treatment effects from combined data: Identification versus data security. In *Economic Analysis of the Digital Economy*, pages 279–308. University of Chicago Press, 2015.

S. Künzel, J. Sekhon, P. Bickel, and B. Yu. Meta-learners for estimating heterogeneous treatment effects using machine learning. *arXiv preprint arXiv:1706.03461*, 2017.

J.-J. Laffont, H. Ossard, and Q. Vuong. Econometrics of first-price auctions. *Econometrica: Journal of the Econometric Society*, pages 953–980, 1995.

R. A. Lewis and J. M. Rao. The unfavorable economics of measuring the returns to advertising. *The Quarterly Journal of Economics*, 130(4):1941–1973, 2015.

L. Li, W. Chu, J. Langford, and R. Schapire. A contextual-bandit approach to personalized news article recommendation. *International World Wide Web Conference*, 2010.

L. Li, W. Chu, J. Langford, T. Moon, and X. Wang. An unbiased offline evaluation of contextual bandit algorithms with generalized linear models. *Journal of Machine Learning Research Workshop and Conference Proceedings*, 2012.

L. Li, S. Chen, J. Kleban, and A. Gupta. Counterfactual estimation and optimization of click metrics for search engines. *CoRR*, 2014.

L. Li, Y. Lu, and D. Zhou. Provably optimal algorithms for generalized linear contextual bandits. *International Conference on Machine Learning*, 2017.

D. McFadden et al. Conditional logit analysis of qualitative choice behavior. 1973.

S. Mullainathan and J. Spiess. Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106, 2017.

N. Naik, J. Philipoom, R. Raskar, and C. Hidalgo. Streetscore-predicting the perceived safety of one million streetscapes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 779–785, 2014.

N. Naik, S. D. Kominers, R. Raskar, E. L. Glaeser, and C. A. Hidalgo. Computer vision uncovers predictors of physical urban change. *Proceedings of the National Academy of Sciences*, 114(29): 7571–7576, 2017.

A. Y. Ng, S. J. Russell, et al. Algorithms for inverse reinforcement learning. In *Icml*, pages 663–670, 2000.

A. Peysakhovich and A. Lada. Combining observational and experimental data to find heterogeneous treatment effects. nov 2016. URL http://arxiv.org/abs/1611.02385.

P. M. Robinson. Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, pages 931–954, 1988.

A. E. Roth. The economist as engineer: Game theory, experimentation, and computation as tools for design economics. *Econometrica*, 70(4):1341–1378, 2002.

F. J. Ruiz, S. Athey, and D. M. Blei. Shopper: A probabilistic model of consumer choice with substitutes and complements. *arXiv preprint arXiv:1711.03560*, 2017.

S. L. Scott. A modern bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry*, 26(6):639–658, 2010.

A. Strehl, J. Langford, L. Li, and S. Kakade. Learning from logged implicit exploration data. *Conference on Neural Information Processing Systems*, 2010.

A. Swaminathan and T. Joachims. Batch learning from logged bandit feedback through counterfactual risk minimization. *Journal of Machine Learning Research*, 2015.

P. Thomas and E. Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. *International Conference on Machine Learning*, 2016.

R. Tibshirani and T. Hastie. Local likelihood estimation. *Journal of the American Statistical Association*, 82(398):559–567, 1987.

J. Ugander, B. Karrer, L. Backstrom, and J. Kleinberg. Graph cluster randomization. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '13*, page 329, New York, New York, USA, 2013. ACM Press. ISBN 9781450321747. doi: 10.1145/2487575.2487695. URL http://dl.acm.org/citation.cfm?doid=2487575.2487695.

M. J. van der Laan and D. Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1), 2006.

H. R. Varian. Big data: New tricks for econometrics. *The Journal of Economic Perspectives*, 28 (2):3–27, 2014.

S. Wager and S. Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, (just-accepted), 2017.

M. Wan, D. Wang, M. Goldman, M. Taddy, J. Rao, J. Liu, D. Lymberopoulos, and J. McAuley. Modeling consumer preferences and price sensitivities from large-scale grocery shopping transaction logs. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1103–1112. International World Wide Web Conferences Steering Committee, 2017.

H. White. *Artificial neural networks: approximation and learning theory*. Blackwell Publishers, Inc., 1992.

M. Yeomans, A. K. Shah, and J. Kleinberg. Making Sense of Recommendations. 2016. URL http://goo.gl/8BjhMN.

A. Zeileis, T. Hothorn, and K. Hornik. Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, 17(2):492–514, 2008.

J. R. Zubizarreta. Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110(511):910–922, 2015. doi: 10.1080/01621459.2015.1023805.