The Impact of Machine Learning on Economics PRELIMINARY AND INCOMPLETE–CHECK WITH AUTHOR FOR LATEST DRAFT

Susan Athey
athey@stanford.edu

Current version September 2017

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. Next, we briefly review of some of the emerging econometric literature combining machine learning and causal inference. Finally, we overview a set of predictions about the future impact of machine learning on economics.

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

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

of the effects. For this reason, this article will both review recent research contributions and also make qualitative predictions about changes in the way economic research will be conducted in the future.

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 large subfield of computer science, but also to a narrower field that spans multiple disciplines. 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. The fact that many ML researchers seem unaware that applied statistics has been a field with numerous methodological and applied contributions for decades prior to their discovery of data analysis makes it tempting to go down this path. 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]

Starting from a relatively narrow definition of machine learning, machine learning is a field that develops algorithms to be used for prediction, classification, and clustering or grouping tasks with data. 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; it is commonly used for video, images and text. 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. 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.

Luca and Sendhil Mullainathan, who have also contributed to my thinking through their writing, lecture notes, and many conversations.

[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" ML, and will discuss it as such.

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. Another form of unsupervised learning is dimensionality reduction. Old methods such as principal components analysis fit into this category, while modern methods include matrix factorization (finding two low-dimensional matrices whose product well approximates a larger matrix), topic modeling, and neural networks.

In my view, these tools are very useful as an intermediate step in empirical work. 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 control for characteristics of the 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.

Unsupervised learning can also be used to create outcome variables. For example, Athey et al. (2017b) 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 netowrk theory.

There are a variety of techniques available for unsupervised learning, including k-means clustering, topic modeling, community detection methods for networks, and many more.

Supervised machine learning typically entails using a set of features or covariates (xs) to *p*redict 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 a test set based on the values of x. 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.

There are a variety of ML methods, such as regularized regression (LASSO, ridge and elastic net), random forest, regression trees, support vector machines, neural nets, matrix factorization, and many others, including model averaging and targeted maximum likelihood. See Varian (2014) for an overview of some of the most popular methods, and Mullainathan and Spiess (2017) for more details. 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.

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 lower 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 researcher might check robustness by looking at 2 or 3 alternatives. Researchers often checked dozens or even hundreds of alternative specifications behind the scenes, but rarely reported 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.

There are also other categories of ML models; 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).

Supervised and unsupervised ML can be contrasted with some of the more traditional types of questions considered by social scientists, who have been interested in conducting inference about causal effects and estimating the impact of counterfactual policies (that is, estimating the impact of policies that havent been tried yet, or estimating the counterfactual outcomes if a different policy had been used). Examples of questions economists often study are things like the effects of changing prices, or introducing price discrimination, or changing the minimum wage, or evaluating advertising effectiveness. In simple cases, this can boil down to estimating a single parameter, such as an average treatment effect for a particular population, or an average elasticity. In more complex cases, this might involve estimating a probability distribution over bidder values in an auction, For further discussions of the contrast between prediction and parameter estimation, see the recent review by Mullainathan and Spiess (2017).

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. Bjorkegren and Grissen (2015) use 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.

3 New 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–one where parameters of interest reflect the causal effect of an intervention. Such parameters by definition give the answer to a well-defined question, and so the magnitudes are straightforward to interpret. 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. 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, 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.

A third issue that arises is stability and robustness. 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, 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 Further Predictions

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 fo 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. Increased emphasis on model robustness and other supplementary analysis to assess credibility of studies
5. Adoption of new methods by empiricists at large scale
6. Revival and new lines of research in productivity and measurement
7. New methods for the design and analysis of large administrative data, including merging these sources
8. Increase in interdisciplinary research
9. Changes in organization, dissemination, and funding of economic research
10. Economist as engineer engages with firms, government to design and implement policies in digital environment
11. Design and implementation of digital experimentation, both one-time and as an ongoing process, in collaboration with firms and government
12. Increased use of data analysis in all levels of economics teaching; increase in interdisciplinary data science programs
13. Research on the impact of AI and ML on economy

I have already discussed the first two predictions in some detail; I will now discuss each of remaining predictions in turn, with the exceptoin of the third, which I will reivew in greater depth in the next Section of the paper.

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. 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. 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. 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.

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)).

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, etc. 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 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 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 with the university.

We will also see changes in how economists engage with government, industry, education, and health. The concept of the "economist as engineer" and even "economist as plumber" will move beyond its traditional homes in fields like 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, personalized resource allocation–all of these create opportunities for economists to design 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, reducing costs, and improving outcomes.

ML methods, when deployed in practice in industry, government, education and health, lend themselves to incremental improvement. Standard practice

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 machine learning algorithms every year. An emerging trend is to build the experimentation right into the algorithm, using "bandit" techniques. Multi-armed bandit is a term for algorithms that balance exploration and learning–trying new things that might be better than the status quo–against exploiting information that is already available about which alternative is best. Bandits can be dramatically faster than standard randomized controlled experiments (see, e.g., the description of bandits on Google's web site by Steve Scott: 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.

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 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 treated and not others).

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

[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.

of living for people providing services in cities, which could reduce effective inequality (which may otherwise continue to rise).

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 studey, including the issues experienced by parts of the workforce who need to transition jobs when their old jobs are eliminated due to automation.

5 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 or nothing 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 REDUCE 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, very recently a nascent 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 for this

reason, 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.

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. Experimental data
2. Unconfoundedness
3. Instrumental variables
4. Difference-in-difference methods
5. Panel data settings
6. Regression discontinuity designs
7. 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 optimal treatment assignment policies
4. 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 strategy 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.

5.1 Average Treatment Effects

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.

An early example of the application of ML methods to causal inference (see Belloni et al. (2014) and Chernozhukov et al. (2015) for reviews) concerns the question of estimating the average effect of a treatment in an environment with unconfoundedness, where the treatment assignment is as good as random conditional on covariates, but some covariates are correlated with both the treatment assignment and the outcome, so that 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." 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's (1988) nonparametric 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's original paper, 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 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 econometric efficiency, which provides strong guidance on what types of estimators are likely to be successful. A second 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.

5.2 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.

This paper builds on earlier work on "model-based recursive partitioning (Zeileis et al., 2008), which looked at recursive partitioning for more complex models, but did not provide statistical properties. More recently, Asher et al. (2016) and Athey et al. (2017c) have proposed methods to define subgroups based on more general method of moments models. For example, Athey et al. (2017c) considers any model that can be estimated using generalized method of moments, and builds trees tailored to discovering heterogeneity in parameter estimates.

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 never use the same data to construct a partition of the covariate space, and 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.

More recently, Athey et al. (2017c) 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. Recently, 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.

Finally, a motivating goal for understanding treatment effects is estimating optimal policy functions; this has been recently studied in economics by, e.g., Kitagawa and Tetenov (2015). The goal is to estimate a policy that maps from individual characteristics to treatment assignments, and select this to minimize the loss from failing to use the (infeasible) ideal 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. In Athey and Wager (2017), we show how bringing in insights from semi-parametric efficiency theory narrows down substantially the set of algorithms that might be efficient. We derive a bound on the loss from using a policy estimate relative to the ideal personalized policy, and show that in order to attain the bound, it is necessary to use an estimator that is semi-parameterically efficient.

5.3 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 supplementary analysis is a robustness measure. **?** 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. (2017a), 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.

5.4 Large Scale Bayesian Modeling

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. (2003) and Blei and M. (2012). I expect to see much closer synergies between these two literatures in the future. For example, **?** 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. 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 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. Thus, the paper again highlights the theme that for causal inference, the objective function

differs from standard prediction.

6 Conclusions

It is perhaps easier than one might think to make predictoins 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; 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

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

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

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

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

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.

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, 2003. ISSN ISSN 1533-7928. URL http://www.jmlr.org/papers/v3/blei03a.html.

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.

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.

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.

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.

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

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.

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.

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

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

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. 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.