Mara Lederman

Discussion of "The Impact of Machine Learning on Economics" (Susan Athey, 2018)

Athey (2018) provides a comprehensive, accessible and exciting summary of the impact that machine learning (ML) is having – and will continue to have – on the field of economics. It is a thorough, thoughtful and optimistic paper which makes clear the unique strengths of ML and the unique strengths of traditional econometrics-based techniques for causal inference and highlights both the opportunities to combine these approaches as well as the sorts of tasks and problems that are likely to remain in each domain. The paper contains a several useful and practical examples that illustrate the application of ML techniques to questions and problems that are of interest to economists including allocating health care procedures, pricing and measuring the impact of advertising.

At a broad level, the paper has four main sections. The paper begins by offering straightforward definitions of unsupervised and supervised ML. Athey puts it quite simply: unsupervised ML uses algorithms to identify observations that are similar in their covariates while supervised ML uses algorithms to predict an outcome variable from observations on covariates. It is important to emphasize, and I will return to this below, that the observations and variables that ML algorithms can handle often do not look like the typical quantitative data that economists use in empirical analysis. Both unsupervised and supervised machine learning techniques can be applied to text, images and video. For example, unsupervised ML algorithms can be used to identify similar videos (without needing to specify in advance what makes these videos similar) or similar restaurant reviews (again, without needed to specify which reviews are positive or negative or what words or phrases makes a review positive or negative). Supervised ML algorithms can be used to predict variables such as the sentiment of a tweet or the slant of a newspaper article, without having to specify ex ante what the relevant covariates are.

The paper then discusses a number of ways in which off-the-shelf ML techniques can be directly integrated into traditional economics research. For example, both unsupervised and supervised ML can be used to create variables that can be used in standard econometric analyses. In addition, ML techniques can be directly applied to what Kleinberg et al (2015) call "prediction policy problems". These are policy problems or decisions that inherently involve a prediction component and, in these cases, ML techniques may be superior to other statistical methodologies. These problems may involve novel sources of so-called "big data" – such as satellite image data used in Glaeser et al (2016) – but need not. They are simply policy problems in which the predicted value of an unknown variable acts an input into a decision.

The third and most substantial section of the paper discusses the growing literature at the intersection of machine learning, statistics and econometrics. As Athey puts it, this literature is developing novel methodologies that "harass the strengths of ML algorithms to solve causal inference problems" (Athey (2018) pp. 10). Athey provides details on a number of recent contributions in this area, highlighting the parts of the estimation approaches that are improved by ML and the parts that continue to rely on traditional econometric approaches and assumptions. Athey predicts that these techniques will soon become commonly used in applied empirical work in economics.

Finally, the paper concludes with a discussion of some of the broader effects that ML might have on the economics profession, beyond the impact on the way we do empirical research, including the types of questions economists will ask, the degree of cross-disciplinary collaboration, the production function for research and the emergence as the "economist as an engineer", working with business and government to implement policies and experiments in a digital environment.

Athey's paper lays out an exciting future for empirical work in economics. It makes clear that there are real complementarities between ML techniques and econometric techniques and she and others are working to develop the relevant methodological tools and make them available to applied researchers. Athey also points out that the growth of ML and ML-based decision-making raises a number of new questions – such as, how to avoid "gaming" of the algorithms as they become known and how to ensure algorithms are fair and non-discriminatory – and that economists and other social scientists seem particularly well-suited to shed light on these types of issues.

While Athey (2018) discusses the current opportunities for economists to utilize "off-the-shelf" ML methodologies in their research – for example, to systematize model selection and robustness checks, to create variables or to carry-out prediction exercises - I believe this point deserves even greater emphasis. The opportunities for researchers to integrate ML techniques into traditional reduced-form or structural empirical work seem enormous. This is because ML, at a fundamental level, takes inputs that do not look like data and turns them into an output that looks very much like the type of data that we can include in traditional econometric analyses. ML is a machinery for prediction. Sometimes that prediction exercise looks like the kind of prediction exercise we might carry out with a simple logit or probit model. For example, we might have data on which students graduate college along with a number of their attributes upon admission and we might use this data to develop a model that predicts that probability of graduation for each new college applicant.

However, much of the excitement around ML algorithms is that they can handle datasets that are "unstructured" – that do not contain a set of neatly labeled covariates in a series of columns. Indeed, ML does not even require the covariates to be specified or labelled. The algorithm determines what the relevant covariates are. Consider, text. Text doesn't look like data. We cannot easily put text—whether long bodies of texts or short fragments of text - into regression models. But what ML can do is take text as an input and predict a variety of things about that text - its content, its sentiment, its political leaning – and these can be used as variables in traditional empirical analyses. As a very simple example, in Gans, Goldfarb and Lederman (2017), we use a sentiment analysis algorithm to classify the sentiment of over 4 million unique tweets to or about a major U.S. airline. This allows us to construct a variety of variables that measure not only the quantity but also the sentiment of "voice" to an airline on a given day which can be used in our empirical analysis. Absent the algorithm, we would be able to count up the number of tweets but would have a much harder time classifying the sentiment of the tweet for anything other than a sample small enough to code by hand.

Tweets are only one example. There are many potentially interesting and informative sources of text that, with ML, can be now be exploited in empirical research. For example, other types of

social media posts, online reviews, patent applications, job descriptions, newspaper articles, commercial contracts, court transcripts, research papers, email communications, customer service logs, performance evaluations, and financial filings to name just a few. Indeed, some of these examples have been discussed by others in this volume. ML technologies literally open the door to novel sources of data that economists can use to answer important questions in a variety of fields.

Finally, in addition to thinking about how we as researchers might integrate ML techniques into our own work, it seems critical to also think about how organizations' integration of ML into their decision-making may impact our research. Despite the growing use of randomized experiments, most research in applied economics still relies on observational data. Observational data, of course, creates challenges for causal identification because the data-generating process is unlikely to be random. We believe that observed equilibrium prices are the result of the interaction of supply and demand and we therefore cannot regress quantity on price to estimate the slope of a demand curve. Or, to use an example from organizational economics, we believe that organizational forms are chosen optimally, to maximize performance, including economizing on transaction costs, and therefore we cannot simply regress performance on organizational form in order to estimate the performance implications of firm boundary decisions. We develop theoretical models to help us understand the data-generating process which, in turn, informs both our concerns about causality as well as the identification strategies that we develop.

As organizations increasingly allocate decisions to ML-based algorithms we need to ask what implications this will have for the variation we observe and exploit in the data we use for research. There are a number of factors to consider. First, ML-based decisions are generally opaque. Thus, even the organizations deploying the ML may not be able to explain how certain decisions were made and so we may not be to understand the data-generating process in some cases. Second, to the extent that organizations use ML to optimize decisions – for example, to target advertising towards those for which it will have the largest impact or to admit the MBA students who are predicted to be the most successful upon graduation – the use of ML may exacerbate selection problems. The treated and non-treated groups that we observe in our data may be even more different on unobservables when those two groups are the result of ML-based decisions. On the other hand, in some instances, ML-based decisions may come closer to the behavioral models we specify. For example, many structural papers in industrial organization specify complicated pricing or entry models. ML-based algorithms may come closer to solving these problems than individual decision-makers within a firm. Finally, as ML and other artificial intelligence technologies diffuse across organizations, they are likely to diffuse at different rates. This means that, at least in some datasets, we are likely to observe a mix of ML-based and traditional decision-making which creates another potentially important source of unobserved heterogeneity. Overall, as applied researchers working with real-world datasets, we need to recognize that increasingly the data we are analyzing is going the be the result of decisions that are made by algorithms in which the decision-making process may or may not resemble the decision-making processes we model as social scientists.

References:

Athey, Susan (2018), “The Impact of Machine Learning on Economics,” this volume.

Gans, Joshua S, Avi Goldfarb and Mara Lederman (2017), “Exit, Tweets and Loyalty,” mimeo.

Glaeser, Edward L., Scott Duke Kominers, Michael Luca, and Nikhil Naik. "Big Data and Big Cities: The Promises and Limitations of Improved Measures of Urban Life." *Economic Inquiry* 56, no. 1 (January 2018)

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. "Prediction Policy Problems."*American Economic Review*, 105(5): 491-95.