MACHINE LEARNING FOR ECONOMICS AND POLICY

Stephen HANSEN

Abstract

This chapter focuses on applications of machine learning algorithms for economic research and policymaking. It first introduces basic concepts in machine learning, whose main branches include supervised and unsupervised learning. The second half of the chapter discusses use cases and applications of machine learning algorithms. First, it discusses the quantification of unstructured data and how to recover information in a way that is useful for economists. The second application concerns new possibilities for measurement, where the combination of machine learning and new digital data, provides the opportunity to develop measures of objects like inflation and economic activity. The last two applications are related to forecasting and causal inference. The overall message of the chapter is that machine learning provides the tools needed to fully exploit the possibilities of rich new digital data sources.

Key words: Machine learning, digital data.

JEL classification: C55.

I. INTRODUCTION

Recent times have seen an astonishing growth in the production of data. More data was created in 2014 and 2015 than in the entire history of humankind beforehand, and by 2020 there will be approximately 44 zettabytes, or 44 trillion gigabytes, of data (Marr, 2015). Much of this explosion is due to digitization, as new technologies allow previously ephemeral human activities to be recorded. Messages and photos are now routinely sent via email or social media, which in turn allows them to be stored on servers indefinitely. Digital data of more direct economic relevance is also now increasingly available. Information from, for example, individual consumers' purchases, detailed product price histories, and rich administrative records is already beginning to transform empirical research in economics.

Along with the growth of data has come new empirical methods for analyzing it. The field of machine learning has developed rapidly in the past ten years in response to the digitization of data, and contributes many ideas to artificial intelligence, which is currently receiving much public attention. The relevance of these advancements for empirical research in economics is less clear. The bulk of machine learning methods have been developed by computer scientists, statisticians, and engineers, who typically have different goals than economists in conducting empirical work. This raises the question of what potential uses there are of machine learning in economics given its emphasis on causal inference and counterfactual prediction.

The first goal of this chapter is to introduce basics concepts in machine learning, and its first part focusses on this. The second goal is to reflect on its potential impact on economic research and public policy, and it does so through the discussion of several application areas. The discussion is non-technical and focused on broad ideas.[1]

There are several takeaway points. First, one important but sometimes underappreciated use of machine learning is the ability to use entirely new types of data. Modern econometrics typically uses data that is "regular": it can be represented in rectangular form with rows corresponding to individual observations and columns to variables. Moreover, variables are typically recorded as single, quantitative measurements like the total expenditure of households or the wages of employees. However, many of the newly available digital data sources do not have this format: text, satellite images, and web search profiles contain vast amounts of economically relevant information but

[1] Readers interested in a more technical, academic discussion can refer to several recent excellent surveys in the economics literature (for example, Einav and Levin, 2014; Varian, 2014; Mullainathan and Spiess, 2017).

have non-standard data structures. Machine learning can be used to extract the important information from these sources, and clean them for econometric analysis. The chapter illustrates several cases in which off-the-shelf approaches have been used to effectively do this.

Second, it is important to recognize that many machine learning methods are often not appropriate for the kinds of problems that economists confront. The chapter provides examples of this in forecasting and causal inference.

Third, despite the differing goals of machine learning and economics, specific ideas from machine learning can nevertheless be incorporated and extended to meet the needs of economic research. This process is only just beginning in economics, but is likely to hold the key for allowing economists and policymakers to fully exploit the potential of digital data.

II. WHAT IS MACHINE LEARNING?

There appears to be no single, agreed-upon definition of machine learning. A generic definition is the study of algorithms that allow machines to improve their performance in some given task as new data arrives. A more expansive definition from a popular textbook is that machine learning is “a set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainty” (Murphy, 2012). However, these definitions do not fully convey the differences between machine learning and econometrics. After all, the ordinary least squares regression model familiar to any undergraduate student in economics detects patterns in data, and has higher-quality estimates when estimated on larger datasets.

One area of difference between machine learning and econometrics is the role of statistical inference. Econometricians tend to focus on formal inference procedures. This involves estimating parameters of a given statistical model, and then deriving theoretical properties of the distributions of these estimates to do hypothesis testing. In contrast, machine learning is often less concerned about the “true” model that generates the data, and instead seeks out procedures that simply work well under some metric, such as predictive accuracy. This distinction is not black and white. For example, some (particularly Bayesian) machine learning algorithms begin from an assumed probability model for the data much like in econometrics, and these can in principle be used for inference. Even in these cases, though, the machine learning literature is typically less concerned with theoretical inference guarantees than is the econometrics literature. Breiman (2001) provides a good introduction to these "two cultures" of statistical modelling.

Another area of difference is computation. Econometric procedures are rarely assessed in terms of their computational complexity, whereas such considerations are at the heart of much of machine learning. Certain core algorithms are popular precisely because they are fast to compute and can scale well. This is largely due to the massive datasets that are used in many machine learning applications. Economists can afford to work with computationally inefficient algorithms given the much smaller datasets they typically analyze, but this will evolve as datasets grow.

There are also some semantic differences that sometimes obscure what are in fact similar ideas. Both fields write models that relate some variable of interest, denoted y, to some other variables potentially related to y, denoted x. Econometricians usually call y a "dependent variable", or "outcome" and the x variables "covariates", "explanatory variables", or "independent variables". In machine learning y is often called a "label", "response", or "target", while x are "features", or "predictors". Moreover, the process of building a model to relate x and y in econometrics is called "estimation" and in machine learning "learning". This chapter will adopt the standard language of econometrics.

Rather than debate the exact definition of machine learning, it is helpful to consider instead the specific tasks that machine learning is designed to solve. A typical division is between *supervised learning and unsupervised learning,* which we now turn to discuss.

1. Supervised Learning

Supervised learning is the task of building a model to explain an outcome variable given covariates. This is exactly what many econometric models do, but the metric to judge the quality of a model in machine learning is quite distinct. Essentially, the only goal is predictive accuracy. Achieving high predictive accuracy with a fixed dataset is trivial. A linear regression model in which one uses as many covariates as there are observations to explain an outcome will perfectly explain the data. Procedures like these, however, tend to over-fit the data and make predictions based on spurious relationships. Machine learning therefore targets *out-of-sample* predictive accuracy. The goal is to build a model that accurately predicts outcomes in new data that was not used in the building of the model in the first place. Models that are good at this task are deemed successful.

To take a concrete example, consider the case of spam email. The outcome variable is binary: either an email is spam or it is not. The covariates are the words in emails. Given some fixed set of emails, predicting spam is potentially as trivial as finding a single word that is only present in spam emails and never in non-spam emails. Suppose this word is ”xxx”. Then, the presence of ”xxx” is a perfect predictor of spam in this specific set of emails. But this model may not generalize well to new emails, for example spam emails requesting bank account details to receive the sender’s inheritance. Instead, we want a model that is likely to accurately classify *new* emails as spam or not.

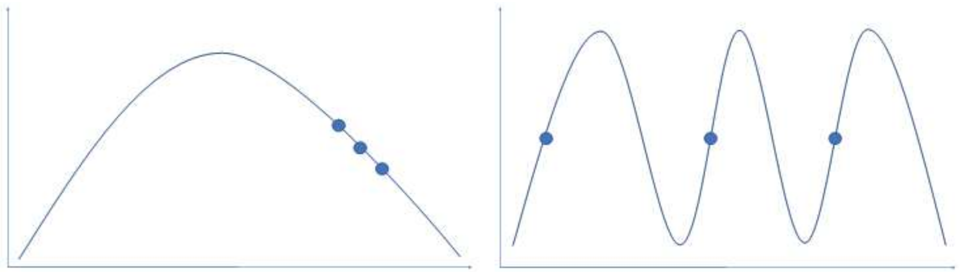
The machine learning literature has made enormous strides in building models with good predictive accuracy. Algorithms for face recognition in photos, speech recognition, and the aforementioned spam detection problem are now widely used across society, and are all applications of supervised learning.

Even if one wishes to use predictive accuracy as the benchmark to judge the success of a model, there are concerns with whether the way supervised learning algorithms are evaluated is sufficient. How can one evaluate the performance of an algorithm on out-of-sample data if such data is not available? The standard solution is to divide the data into two portions: a training sample and a test sample. The training sample is used to estimate a model. Then, for each observation in the test data, one can generate a predicted value for the outcome given the model estimated with the training sample, and then compare the prediction against the actual value in the test data. The test sample stands in for out-of-sample data since it is not used in training. However, there is often no guarantee that the *actual* out-of-sample data that an algorithm will confront in the real world corresponds to the data it confronts in the test set. Figure 1 below provides an illustration. Consider the situation on the left. Suppose the observed data are the three points on the curve, and we are trying to predict an outcome measured on the vertical axis given some covariate measured on the horizontal axis. The curve represents the real-world relationship between the covariate and the outcome. A supervised learning algorithm constructed only with the three observed points might go badly wrong *even if* it achieves high out-of-sample predictive accuracy on a test set. This is because all the data comes from a restricted part of the curve that behaves like a downward-sloping line, and a supervised algorithm will tend to estimate just this pattern. This pattern clearly does not generalize well to all possible covariate values since part of the real-world relationship involves an upward slope. Similarly, in the situation on the right the observed data again will give a misleading view on the true relationship. The problem is now that the observed data are too dispersed.[2]

[2] Many thanks to Bryan Pardo of Northwestern University who first made these points to the author.

FIGURE 1

DANGERS OF UNREPRESENTATIVE DATA FOR SUPERVISED LEARNING EVALUATION



These examples are simple and involve a single-dimensional covariate. In real applications of machine learning, one has hundreds, thousands, or even millions of different inputs, and determining whether the data on which supervised algorithms are evaluated gives a representative view of the world is extremely challenging. Economists and policymakers should bear this in mind. While mistakes in speech or image recognition can be annoying and embarrassing, they have low social costs. Mistakes in policymaking can be catastrophic.

Supervised learning algorithms are also generally constructed in environments that are quite different than those that economists face. First, they are data rich. Companies like Facebook and Google can draw on vast troves of data to train recommendation algorithms. In contrast, economists many times have very limited data to work with. For example, while predicting recessions is an important policy problem, recessions are relatively infrequent in historical time series. Second, the environments are stable in the sense that the future looks much like the past. Economies are often non-stationary, and often when predictive accuracy is most important, such as at the onset of a financial crises or the introduction of a disruptive technology. This brings into question whether off-the-shelf machine learning methods are appropriate for the kinds of prediction problems that economists are most interested in. The chapter returns to this issue in the discussion of applications below.

To better understand the differences between machine learning and traditional econometrics, it is instructive to consider a popular supervised algorithm called the LASSO (Least Absolute Shrinkage and Selection Operator), which was introduced by Tibshirani (1996) and has become increasingly popular in economics (see, for example, Belloni, Chernozhukov and Hansen, 2014). The

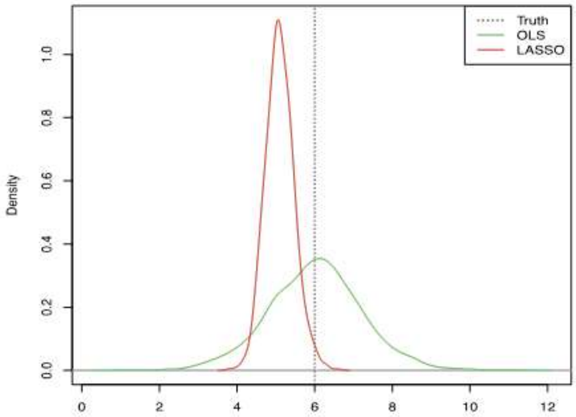
LASSO is a basic extension of the ordinary least squares (OLS) regression model that is the workhorse for applied economics. Both models relate an outcome y to covariates x by choosing coefficients for the x values that best explain y. For example, y might be income, and x might be composed of three variables: years of schooling, IQ, and eye color. We expect the first two covariates to relate to income, but the not third. The key difference between OLS and LASSO is that LASSO adds a penalty term that punishes large coefficient values. The idea behind this penalty is to assign a zero coefficient to unimportant variables and a non-zero coefficient to important variables. The hope is that the variables with non-zero coefficients have a true relationship with the outcome, and those with zero coefficients are noise variables that do not. In the previous example, this would mean that LASSO would give a positive coefficient to schooling and IQ, and a zero coefficient to eye color. Such an approach can be particularly fruitful when there are many variables relative to the number of observations. In fact, LASSO can even be estimated when there are many more variables than observations.

While the penalty term in LASSO can eliminate noise variables, this comes at a cost. The penalty term punishes large coefficient values for *all* covariates. This means that even the coefficients on the true variables are lower than they would be in the simple OLS model. In the technical language of econometrics, the coefficient values estimated from LASSO have a bias: the estimated effect of any covariate has on average a lower magnitude than whatever the true effect is. To continue with the example above, suppose that an extra year of schooling leads to an extra income of 600 EUR per year. The LASSO might estimate that the extra effect of a year of schooling is only 300 EUR per year. Why, then, would one want to use a model that intentionally introduced bias into its estimation procedure? The answer is that introducing bias reduces noise. The OLS model will estimate some coefficient value for eye color even though this has no relationship to income. On average this will be close to zero, but depending on the randomness in any specific dataset there may be some spurious correlation between eye color and income that OLS will pick up. This in turn introduces noise in predicted income. By contrast, LASSO will simply tend to drop eye color out of the model completely.

Figure 2 illustrates these properties. Suppose there is a person who has been to school for 5 years and has an IQ of 100. Moreover, suppose that an additional year of school increases income by 0.6 units, and an additional point of IQ increases income by 0.03 units. Thus, this individual’s true income is 5 * 0.6 + 100 * 0.03 = 6. Figure 2 plots the distributions of the values for predicted income produced by OLS and LASSO when there are also many noise variables in the model with no relationship to income. Here we see clearly what is known in the machine learning literature as the *bias-variance tradeoff.* OLS on

FIGURE 2

THE TRADEOFF BETWEEN BIAS AND VARIANCE



average generates the correct prediction since the distribution is centered at 6. But around this average we see large dispersion: there are predicted values as low as 0 and as high as 12. By contrast, LASSO is biased, since the distribution is centered around 5 rather than 6. But the predicted values are tightly centered around 5, without the extremes of OLS. Put another way, LASSO is wrong on average, but never too wrong; OLS is right on average, but often very wrong. One can show in this example that the average squared error–a popular metric for goodness-of-fit–is lower under LASSO than OLS.

What are implications of this example? Much of textbook econometrics restricts attention to models that are on average correct (unbiased), and then searches within such models for those with low variance. Machine learning shows us that this approach may be limited, especially when there are many variables and when the main goal is prediction, in which case biased models can perform well. At the same time, as discussed above, economists are interested in models with good inference properties: when deciding the amount to invest in public schools, it is crucial to know the true effect of an additional year of school on income (0.6 in the example above). Since supervised learning algorithms are designed for predictive accuracy, a natural question to ask is whether the two goals are in tension. In other words, can supervised learning algorithms be used for parameter inference even though they were not designed with this goal in mind? In many important cases the answer is "no", or perhaps more accurately, "not without modification". As we have seen for the LASSO, the coefficient estimates have a downward bias. Moreover, there is no guarantee that LASSO omits all noise variables. There is some theoretical work on statistical inference with the LASSO (interested readers can consult Bühlmann and van de Geer, 2011 or Hastie, Tibshirani and Wainwright, 2015), but in practice there are few reliable guarantees that are consistent across applications.

The main message here is that supervised learning has recently made enormous strides in accurate out-of-sample prediction in stable, data-rich environments. It often does so by introducing bias to reduce variance, which is crucial in models with vast numbers of variables. However, whether and when these models can be used for the inference problems many economists care about is still an open question undergoing active research. We will discuss recent contributions in the applications section below.

2. Unsupervised Learning

While unsupervised learning has received somewhat less attention in the literature, it is important in its own right. The goal of unsupervised learning is to uncover hidden structure in data. There is no notion of a dependent variable in unsupervised learning that one tries to explain with covariates. Each observation in a dataset simply has multiple recorded variables with potentially complex interdependencies that unsupervised learning tries to reveal. There may be several motivations for unsupervised learning. One may wish to describe the most prominent sources of variation within a vast array of covariates. Alternatively, unsupervised learning can provide a low-dimensional representation of a high-dimensional object that preserves most of the relevant information. Unsupervised learning can also group observations together based on similarity. None of these motivations should be wholly unfamiliar to economists. Clustering and factor analysis, for example, are examples of unsupervised learning tasks that are already rather common in empirical economics.

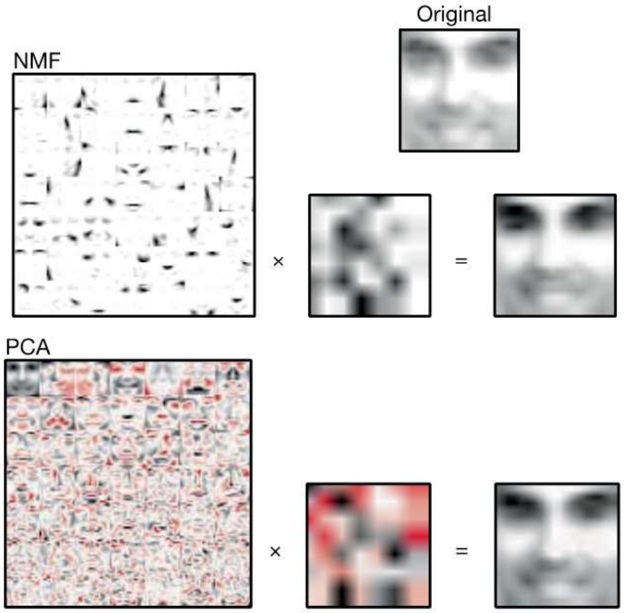
Unsupervised learning can be an end in itself if data exploration is the primary goal, or else be seen as a data preparation tool used to extract features to serve as inputs into supervised learning algorithms or econometric models. In applied economics research, this makes it arguably less controversial than supervised learning. For better or worse, even in economics issues of formal inference are often downplayed when the primary aim is data processing and preparation. In this sense, off-the-shelf methods for unsupervised learning can be applied more readily if they provide a richer description of data than existing methods. The rest of the discussion argues this is indeed the case.

Probably the most well-known unsupervised learning algorithm in economics is principal components analysis (PCA). The idea is to find common components across variables that explain how they move together. Observations are then represented as combinations of these common components rather than in terms of the original variables. Researchers typically use far fewer components to represent observations than there are variables, so there is dimensionality reduction. For example, such an approach is often used in macroeconomic time series to explain the co-movement of hundreds of different economic indices. The common components can be thought of as unobserved cyclical variables that drive the observed data.

PCA is also well-known in the machine learning literature, but machine learning has also developed additional algorithms that correct for some of PCA's limitations. Although economists are not generally aware of these, incorporating them into the econometric toolkit can be done at fairly low cost. One limitation of PCA is that the components it identifies can be difficult to interpret, and, in many instances, appear more like abstract objects that explain co-movement rather than objects with actual meaning. There has been work on alternative ways of constructing components in the machine learning literature that eases this problem in certain applications. An interesting example is from Lee and Seung (1999), who compare PCA with an alternative called non-negative matrix

FIGURE 3

PRINCIPAL COMPONENTS ANALYSIS VS. NON-NEGATIVE MATRIX FACTORIZATION



Source: Lee and Seung (1999).

factorization (NMF). NMF is similar to PCA, except it constrains the components to be made of only non-negative numbers. This seemingly technical distinction is in fact is substantive, because the components that NMF produces appear more like the elemental parts that each observation in the data is built from.

Figure 3 illustrates this idea for image data. The underlying dataset is a collection of photographs of human faces. The seven by seven, larger matrices on the left of the figure illustrate the 49 components that PCA and NMF uncover from the photos. Black shading indicates positive numbers, and red shading indicates negative numbers. The fascinating aspect of the example is that the NMF components appear to be elements of a face: there are eyes, mouths, noses, etc. A single photo in the data is then built by combining these elements into an individual face (the smaller matrices in the middle of the figures show the picture-specific weightings applied to the components to arrive at the observation on the right). The components of PCA are very different: the first component is essentially an average face, and the rest of the components add and subtract pixel intensity from this average face. A specific face is then represented as a weighted deviation from the average face, which is a less intuitive construction than NMF gives.

This example may seem like a curiosity, but it illustrates a deeper point that economists could potentially gain insight on latent structure from leveraging common algorithms in machine learning that have to date been almost entirely ignored. For example, one could apply NMF and related algorithms to individual product sales across consumers to learn archetypal shopping patterns and identify substitutes and complements, or to individual product prices to learn the underlying components in overall inflation.

Another limitation of PCA is that its foundations are most appropriate for data that varies continuously. One important example of data for which this is not the case is text. The most basic way of representing textual databases, also called corpora, is to count the occurrence of all unique terms in the vocabulary across all documents. The resulting data clearly has interdependencies, for example the word ‘labor’ will tend to co-occur with the word ‘wage’. But the data is fundamentally discrete, as a word cannot appear 1.5 times. Also, the vast majority of unique words in corpora do not occur in any specific document, and so the data is also populated by a large percentage of zeros. Such data calls for algorithms that model its specific features.

One of the most powerful and popular unsupervised learning models for text is Latent Dirichlet Allocation (LDA), introduced by Blei, Ng and Jordan, (2003). LDA is an example of a probabilistic topic model, which both identifies topics in corpora and then represents documents as combinations of those topics. More specifically, a topic is a probability distribution over all the unique words in the corpus. This probabilistic aspect of LDA is important. Suppose one imagines a topic about inflation and another about unemployment. Now consider the word ‘rate.’ *Prima facie* it is unclear into which topic ‘rate’ should go, since a topic about inflation or unemployment might feature ‘inflation rate’ or labor ‘participation rate’, respectively. Allowing probabilistic assignment of words to topics conveys this semantic flexibility. LDA is also a mixed-membership model because documents are not assigned a single topic. Instead, each document is allocated shares of all topics. So, a document can be 25% about unemployment, 10% about inflation, etc.

Figure 4 shows example output of LDA estimated on a corpus of verbatim transcripts of the discussions of the Federal Open Market Committee, which decides on monetary policy in the United States. The sample period for estimation is 1987-2009. The two word clouds represent two different estimated topics. The size of the word in the cloud is approximately equal to its probability in the topic.[3] Although the algorithm is not fed any information on the underlying content of the data, the topics are clearly interpretable: there is one about economic growth, and another about recession and recovery. The time series above the topics shows variation in the share of time that individual FOMC members spend discussing the respective topics (the blue dash is the maximum share in a given meeting, the solid black line is the median share, and the dashed red line is the minimum share). Periods of recession are shaded in gray. The series also shows very natural properties. Attention to growth systematically increases when the economy expands, then collapses at the beginning of recession periods. In contrast, attention towards recession spikes during contractions. Again, it is worth emphasizing that such patterns have been wholly captured by a machine learning algorithm, with no input from the researcher.

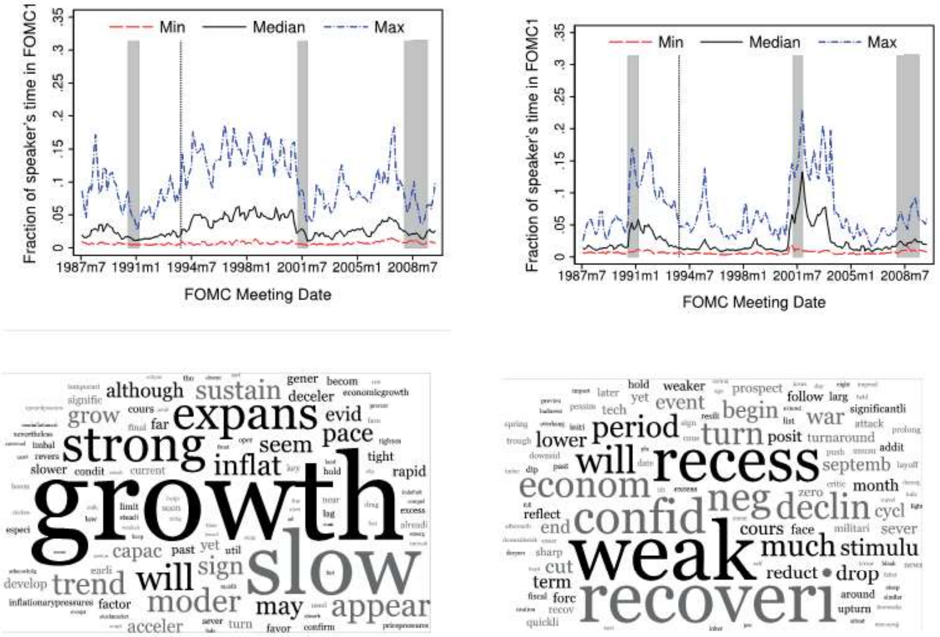
Another important point is that text is innately very high-dimensional. Even moderately-sized corpora contain thousands of unique terms. Overfitting such data is a serious problem, but the statistical structure of LDA guards against this. It is what is known as a Bayesian model, which means it places some initial likelihood on all possible combinations of words in topics. The observed data then changes these likelihoods but does not fully determine them. The transcript dataset above has roughly 10,000 unique terms, and yet LDA handles the dimensionality with ease.

These two examples show the power of unsupervised learning to reveal interesting patterns in data. Moreover, they also show how machine learning can convert what at first sight are unstructured, messy data–*i.e.,* image files and raw text data–into a tractable, quantitative forms that are suitable for

[3] Some of the terms are not English words because the data has been stemmed prior to estimation, a process whereby words are brought into their linguistic roots.

FIGURE 4

EXAMPLE OUTPUT OF LATENT DIRICHLET ALLOCATION



Source: Hansen, McMahon and Prat (2018).

traditional statistical analysis. This opens the possibility of not just having new techniques to use with existing data, but having access to new data itself. This point is discussed further in the applications section below.

One possible criticism of unsupervised learning algorithms, however, is that they have too little structure. Figure 4 shows that there are likely to be time-varying probabilities of topic coverage depending on the business cycle, but this is not built into LDA. One possible contribution of economists to the development of unsupervised learning algorithms is to introduce dependencies of interest into them to more directly link their outputs to quantities of interest. Such efforts will likely require collaboration across disciplines.

III. APPLICATIONS

Having set the foundations of basic concepts in machine learning, the rest of the chapter expands on potential applications in economics and policy. We begin with one of the most pragmatic applications: to quantify novel data in tractable forms. Next, we consider the role of machine learning in converting digital data into specific economic measures, followed by a discussion of machine learning in forecasting models. Finally, we reflect on possible applications in causal inference.

1. Quantification of Unstructured Data

Many firms and regulators are awash with unstructured data, and specifically text data. One leading example is the legal industry, in which much of the work of junior lawyers is taken up by trawling through documents to find relevant content from contracts, title deeds, prior judicial decisions, etc. Regulators too face a similar task when they initiate cases. For example, dawn raids on potential violators of competition law typically yield troves of documents, and sifting out the relevant material from the mass of irrelevant material is an important challenge. Automating the task of finding relevant information therefore has the potential to generate large efficiency gains in these contexts, and indeed this process is already well underway in the legal industry (Croft, 2017).

One of the most common ways of determining document relevance in economics is keyword searches. In this approach, a word or list of words is defined in advance, and then documents are flagged as containing these terms or not, or alternatively ordered according to the frequency with which terms appear. While simple and relatively easy to implement, keyword searches have limitations. Most basically, they require the definition in advance of the important words, which may require subjective judgments. For example, to measure economic activity, we might construct a word list which includes ‘growth’. But clearly other words are also used to discuss activity, and choosing these involves numerous subjective judgments. More subtly, ‘growth’ is also used in other contexts, such as in describing wage growth as a factor in inflationary pressures, and accounting for context with keyword searches is practically very difficult. In other cases, the academic or policymaker may simply have no idea how words relate to the content of interest. In litigation involving traders’ manipulation of market prices like the recent LIBOR rate-fixing scandal, much of the evidence comes from chat rooms in which traders make heavy use of jargon, slang, and code that make simple keyword searches difficult to implement.

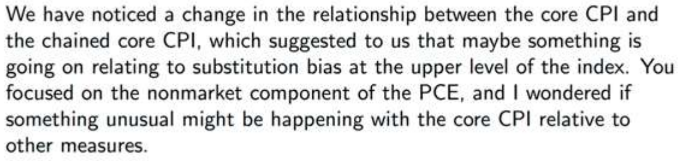
Unsupervised machine learning helps overcome some of these problems. Especially in environments with uncertainty about what content documents contain, and how words are used in different kinds of contexts, machine learning provides a powerful, data-driven approach for corpus exploration and information retrieval. The quantification of unstructured data might be an end in itself by, for example, allowing a regulator to quickly sift through documents and sort them into categories. Or it might be the first stage in extracting features from text data that then serve as inputs into further empirical studies.

To illustrate these points more concretely, consider the example data point in the FOMC transcript corpus discussed in the previous section represented in Figure 5. This is an utterance of Janet Yellen in March 2006 when she was President of the Federal Reserve Bank of San Francisco. This statement uses highly technical language, and determining its content manually would require a reader to have a high level of education in economics.

As an alternative to manual processing, one can use Latent Dirichlet Allocation (LDA), an unsupervised learning algorithm described above, to determine its content. The estimated LDA model associates this statement most with the topic on the left in Figure 6 below. This topic in turn places highest probability on 'inflation' (and other words that begin with the stem 'inflat'). The fascinating aspect of this illustration is that the example data point contains no occurrence of the word `inflation', and a keyword search for it would not flag this statement as relevant. Instead, Janet Yellen uses many words related to inflation (CPI is the consumer price index, PCE is personal consumption expenditure), and LDA learns from other documents in the corpus that the words Yellen uses are most often used in situation in which the word `inflation' is also used. This allows it to associate her statement with the inflation topic.

FIGURE 5

EXAMPLE DATA POINT IN FOMC TRANSCRIPTS



We have noticed a change in the relationship between the core CPI and the chained core CPI, which suggested to us that maybe something is going on relating to substitution bias at the upper level of the index. You focused on the nonmarket component of the PCE, and I wondered if something unusual might be happening with the core CPI relative to other measures.

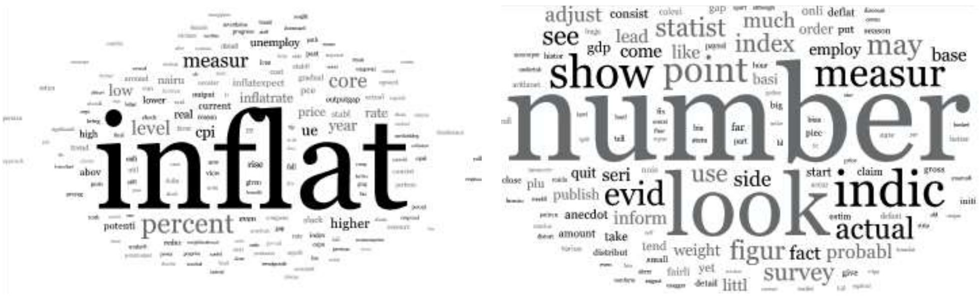
Source: Janet Yellen, March (2006).

Another point of interest is that LDA is able to place individual words within documents into their appropriate context. Consider the word 'measures' that Yellen uses in the example statement. While this word appears prominently in the inflation topic, it is also present with high probability in another topic about numerical indicators displayed on the right in Figure 6. LDA can resolve this ambiguity by looking at the other words that Yellen speaks. While the classification of ‘measures’ without any context is unclear, the fact that Yellen uses many words unambiguously associated with the inflation topic causes it to assign ‘measures’ to the inflation topic as well.

These features of LDA help explain its widespread popularity. One application familiar to many readers might be the indexing system in JSTOR, the popular repository for academic papers. LDA is used to alert readers to new articles of potential interest in the repository given the estimated content in previously viewed articles. The adoption of such systems for firms and regulators that also handle large textual corpora would potentially create large gains in automated information retrieval.

FIGURE 6

TOPIC MOST ASSOCIATED WITH EXAMPLE (LEFT) AND ALTERNATIVE TOPIC (RIGHT)



While the focus of the discussion so far has centered around text, similar points can be made for other kinds of unstructured data too. Policy institutes and marketing firms regularly collect survey data to measure attitudes, behaviors, and characteristics. This data is often analyzed in *ad hoc* ways, for example by computing some average response across a range of questions to obtain a single number. Again, unsupervised learning provides a way of modeling the full dependency structure in the data and extracting novel insights into the underlying ways in which respondents differ from each other. One example in the economics literature is Bandiera *et al.* (2017), who analyze detailed time-use surveys of over 1,000 CEOs across a range of countries. Using LDA, they find a novel behavioral distinction between CEO ‘leaders’ who spend time coordinating high-level functions in companies and `managers’ who spend time on more operational matters. Similar approaches have been used to measure health status (Erosheva *et al.,* 2007) and political ideology (Gross and Manrique-Vallier, 2014) from surveys.

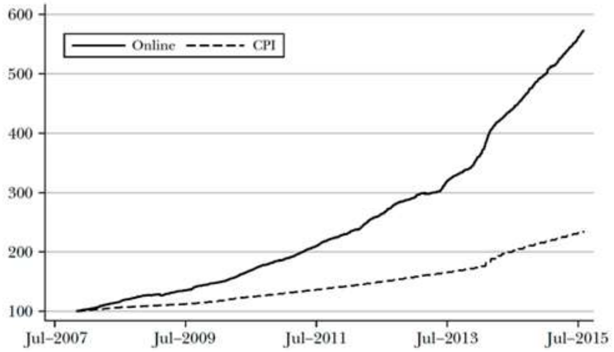
Another intriguing potential application of unsupervised learning is to network data, where the challenge is to identify groups of related nodes based on linkage patterns. There is a large literature on this so-called community detection problem outside economics, but hardly any economic applications. One exception is Nimczik (2017), who estimates the geographic extent of labor markets using data on worker flows in Austria using unsupervised learning.

2. New Data and New Measurement

The first application discussed above was simply to use machine learning to make sense of messy, difficult-to-interpret data while imposing minimal structure on the process of information retrieval. However, there is growing interest in not just describing such data, but also in using it to construct new measures of relevant economic variables. There are various ways in which traditional economic indicators are limited. They are often available at relatively infrequent intervals, as is the case with quarterly GDP measures. Furthermore, they are often constructed for aggregated geographical units like nation states with very little spatial granularity. Finally, in many regions of the world official economic statistics are either unavailable entirely, or else manipulated by governments to the extent that they contain very little information. For these reasons, there is demand for new sources of information. Recently there has been growing interest in digital data as a means of filling these gaps. Examples include:

FIGURE 7

ARGENTINIAN INFLATION AS MEASURED BY ONLINE PRICES AND OFFICIAL CPI MEASURE



Source: Cavallo and Rigobon (2016).

- In Argentina, the government actively manipulated official price statistics beginning in 2007. The Billion Prices Project at the Massachusetts Institute of Technology began as a means of providing an alternative, more accurate inflation index using prices posted by online retailers in Argentina and has since expanded to many other countries. While the universe of retailers for which one can obtain online prices is smaller than that surveyed by official government agencies, these prices are updated daily, have a low cost of extraction, and are free from government interference. Figure 7 below shows inflation measures using online prices and official statistics, and demonstrates the ability of digital data to capture the actual underlying dynamics in an economy when official data is unavailable or unreliable.

- Baker, Bloom, and Davis (2016) construct a popular and influential Economic Policy Uncertainty (EPU) Index (http://www.policyuncertainty.com/). While the impact of uncertainty on economic activity is acknowledged as important, historically there have been very few adequate measures of uncertainty. Financial-market based measures like VIX are based on option prices derived from US equity markets, which do not capture the full uncertainty that economic agents face. The EPU index instead measures uncertainty specifically about policymaking. It is constructed in large part based on the fraction of articles in a wide selection of newspapers that contain terms like ‘uncertain’, ‘economic’, ‘congress’, and ‘regulation’.

- Glaeser, Kim, and Luca (2017) construct a local activity index using the number of restaurants and businesses reviewed on the website Yelp. This index has predictive power for the much more aggregated and lagged data from the US Census Bureau on county business patterns, especially in more densely populated areas.

- SpaceKnow is a commercial company that produces numerous indices of economic activity using satellite image data. One such index is the China Satellite Manufacturing Index, which is based on 2.2 billion individual snapshots of more than 6,000 industrial sites in China (Wigglesworth 2018).

While activity indices are some of the most natural objects of interest that new data can provide, there are also less obvious but equally powerful possibilities. A good example comes from the work of Hoberg and Phillips (2010 and 2016) and has direct relevance to competition policy. The issue is how to measure the industry classification of firms. The often-used SIC or NAICS classification systems have several limitations. Firms typically do not receive different classifications over time even when their markets evolve. The classification systems also do not track the development of entirely new products particularly well. More generally, they provide a very coarse distinction of the ways in which firms differ from each other.

Hoberg and Phillips propose the use of text data to construct industry classifications that overcome some of these challenges. The idea is to use companies' product descriptions contained in their annual 10-K filings to the US Securities and Exchange Commission. For each pair of firms that make a filing in each year, one can compute a measure of linguistic similarity between descriptions and use it as a proxy for proximity in product space. Moreover, from these similarity measures, one can group firms into clusters to define industry categories. The resulting categorization provides a dynamic, continuous measure of firms' location in product space relative to all other firms in the data. Hoberg and Phillips show that their text-based categorization provides several new insights into why firms merge and how new products develop.

At this stage, it is useful to make the distinction between `big data' coming from digital sources on the one hand and machine learning on the other. While raw digital data no doubt contains information relevant for economic variables of interest, the exact mapping between the two is difficult to know. One possibility is to apply unsupervised learning algorithms to describe the data along the lines discussed in the first application, and then use the extracted features to build an index of interest. The problem is that these features will not have been chosen to have maximum predictive power for the economic variable, which implies a loss of information and thus usefulness.

Instead, the task of building new indices from vast data is in many ways a classic supervised learning problem, since the primary goal is to make the best possible prediction of the object of interest. Jean *et al.* (2016) is an example of research that combines vast digital data (satellite images) and state-of-the-art supervised machine learning algorithms to provide a new economic measurement (spatially granular poverty levels in several African countries). As the use of machine learning in economics becomes more widespread, many of the indices built from digital data will likely also be the output of targeted supervised algorithms.

3. Forecasting

As discussed above, supervised machine learning is at its heart the study of methods for achieving good out-of-sample prediction using high-dimensional or unstructured data. One area of high interest for policymakers is forecasting, or predicting the future based on past data. In fact, the idea that rich economic time-series data can be used to obtain better forecasts of the future predates the growth of interest in machine learning. Stock and Watson (1999) and Bernanke, Boivin and Eliasz (2005) are seminal contribution in the literature that show that augmenting standard macroeconomic forecasting models with many time series can improve future forecasts. These papers use methods like penalized regression and dimensionality reduction that are part of the standard machine learning toolkit.

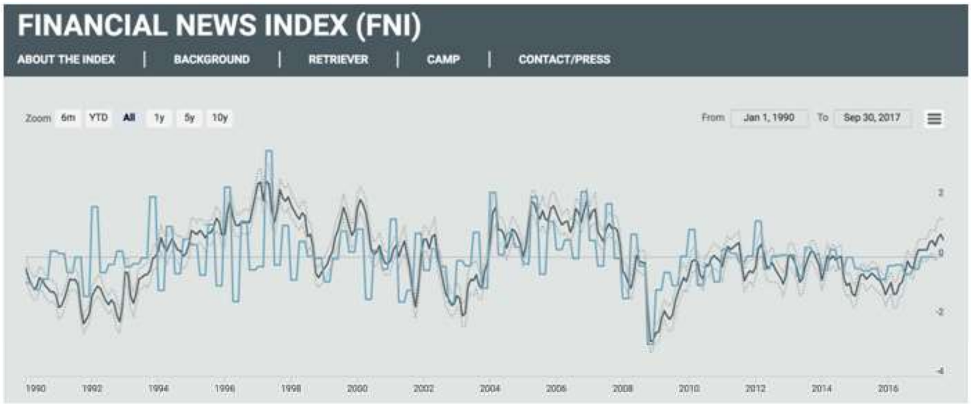
Before economists apply more modern supervised learning algorithms for forecasting, it is worth emphasizing again that the problem of economic forecasting differs in fundamental ways from the environments in which many machine learning algorithms are built and evaluated. First, a common assumption in machine learning is that the out-of-sample data has the same distribution as the training data. In a time-series context, this boils down to an assumption that the future looks like the past. While this may sometimes be true, in other cases it might not be if there are fundamental structural changes. For example, if there is shift in the productive capacity of the economy, then the historical relationship between unemployment and wage growth will change. While there is a well-established literature in econometrics on the detection of structural breaks, the machine learning literature in this area is much less developed. Second, in economics the data is often big on some dimensions but small on others. While there are hundreds of available time series for forecasting, many are observed only at a quarterly or even less frequent basis. Third, the so-called `signal-to-noise’ ratio in economic and financial data can be quite low, which means that fundamental relationships among variables can be hard to detect because there is a lot of randomness that affects all variables in the model. The overall challenge, then, is to find ways of employing supervised learning methods in situations for which they were not originally designed.

One possibility is the use of so-called *generative* models. These models construct a full statistical model for input and output data, in contrast to some recent tendencies in machine learning like deep learning that take a more agnostic stance on the model that generates the data. The main reason for the success of deep learning models is their remarkable predictive power in the presence of vast data. In smaller samples like the ones economists face, though, generative models have been proven to have better predictive power (Ng and Jordan, 2002). Davig and Smalter Hall (2017) make use of this insight, and show that a generative model better predicts US recessions than standard regressions models and the Survey of Professional Forecasters. Another advantage of generative models is that they are closer to the kinds of structural models that economists are already used to constructing and estimating.

Another approach to forecasting with large data is to first use unsupervised learning to extract features, and then use those features as inputs into an otherwise standard economic forecasting model. One example is Thorsrud (2016), who applies latent Dirichlet allocation to Norwegian media articles, and uses the extracted topics to predict evolution in the business cycle. Figure 8 below plots the derived index against actual Norwegian GDP. Clearly the two series co-move substantially, which illustrates the value of features extracted from unsupervised learning for forecasting.

FIGURE 8

FINANCIAL NEWS INDEX (BLACK) AND NORWEGIAN GDP GROWTH (BLUE)



Source: From https: //www.retriever-info.com/fni

Another example from outside macroeconomics is the prediction of conflict, which is important both for risk management of private sector companies and governments. Mueller and Rauh (2017) show that media data can help forecast the outbreak of political violence. They also use LDA to extract topics from text, and then show that variation in topic usage in newspapers' coverage of countries predicts conflict in those countries.

A general comment that applies to the approach of using extracted features as inputs into forecasting models is that they implicitly treat them as fixed data rather than estimated objects. While this has led to important advances in research, in the future one would expect the development of algorithms that jointly model high-dimensional data and whatever variable is being predicted. This is likely to lead to even better predictions, and also more rigorous statistical inference. Again, generative models can provide the backbone for such approaches.

4. Causal Inference

The applications discussed so far all represent important steps in empirical work in economics, but the profession is currently dominated by interest in causal inference, and more precisely in determining the effect of policy interventions. The usefulness of predictive models for this goal is not immediately obvious. Athey (2017) presents a nice illustration of this point. Suppose a hotel chain is interested in determining the effect on sales of rooms following an increase in the price of rooms. If one simply takes observed price and sales data, there is a positive relationship because as occupancy rates increase hotels raise the price of remaining rooms: during peak holiday periods rooms are scarce and prices are high, while during low season the reverse is true. Therefore, a purely predictive model would indicate higher sales following an unexpected increase in price. Of course, common sense dictates that exactly the reverse would occur, *i.e.,* a hotel would sell fewer rooms if it unexpectedly raised prices. The problem here is that a pure predictive model based on observed data fails to account for the unobserved underlying demand for hotel rooms. High occupancy rates are associated with high prices because high demand drives both. Methods for solving problems such as these have been the subject of a great deal of modern econometrics.

What, then, can machine learning offer for economists interested in estimating causal relationships? One important realization is that even causal inference procedures involve what are essentially pure prediction steps. One classic approach for causal inference is the use of so-called 'instrumental' variables. These are variables that are correlated with a treatment but not with the outcome of interest.[4] Replacing the treatment with the instruments allows one to isolate the causal impact of the treatment on the outcome. Instrumental variable estimation typically proceeds in two steps: first, one predicts the value of the treatment given the instruments; second, one uses the predicted value of the treatment as an independent variable in a regression on the outcome. The first step in this procedure can be viewed as a natural machine learning task as it involves making an optimal prediction of the treatment given the instruments. Machine learning methods for instrumental variables are particularly relevant when there are many potential instruments, or when one wants to estimate a flexible relationship between instruments and treatments. Several recent papers combine supervised machine learning methods with instrumental variables (Belloni *et al.,* 2012; Hartford *et al.,* 2017).

Another application of machine learning to causal inference is the problem of high-dimensional controls. Many potential *observable* variables can

[4] In the following discussion, a treatment will mean a variable that a researcher or policymaker intervenes to change, and an outcome will mean whatever target variable he or she is attempting to influence.

also affect the outcome of interest beyond just the treatment of interest. For example, the impact of worker training on productivity might depend on worker characteristics, firm characteristics, and the characteristics of the technology that the worker operates. Which control variables beyond the treatment to include in regression models is often unclear, especially in the absence of a relevant theory. A common approach is to run many different models, each of which includes different controls, and to examine how sensitive the relationship between a treatment and outcome is to the inclusion of a particular set of controls. One naïve machine learning approach would be to include all controls along with the treatment in a penalized regression model in order for the data to reveal which controls are relevant. In fact, this approach yields unreliable estimates of the treatment effect, but adjustments of off-the-shelf algorithms can help correct the problem (Belloni, Chernozhukov and Hansen, 2014).

Another approach to causal inference in economics is so-called structural modelling in which one takes a theoretical economic model, and then uses data to estimate the parameters of the theory. As models grow in complexity, the number of parameters can grow rapidly. For example, a consumer demand model could in theory involve cross-price elasticities between every possible pair of goods in a supermarket. Machine learning can also offer techniques for parameter estimation in large-scale structural models fit on large-scale data. Generative models with a Bayesian formulation again provide a natural framework for structural estimation in economics. While these have arguably lost favor in recent years in the machine learning community due to the rise of deep learning, their future in economics is promising. A recent example is Athey *et al.* (2018), but it is safe to say that this application of machine learning is probably the least developed of all those discussed.

As with the forecasting application, the broad point again arises that the context in which machine learning algorithms are often built is not necessarily directly applicable to empirical applications. This is not to say that machine learning has no relevance to causal inference, but in this area especially careful thinking is required to assess where machine learning techniques can add value.

IV. CONCLUSION

This chapter has reviewed basic concepts in machine learning and provided numerous examples of how machine learning might be useful to academic economists and policymakers. Some applications simply require off-the-shelf methods, while others require the development of new techniques to address the challenges specific to economics. While some of these techniques are already under development, there is much still to be done.

While this chapter has focused on the value policymaking authorities can derive from applying machine learning techniques to data, there are also new regulatory issues that are byproducts of the increased use of machine learning. One example is firms' use of pricing algorithms. When firms tailor prices to individual customers' traits and behavior, price discrimination almost necessarily increases. Whether this reduces consumer surplus is less clear. On the one hand, increasing prices while keeping quantity constant reduces surplus, but on the other pricing algorithms may allow firms to increase the quantity or variety of goods produced. A second issue is whether the use of pricing algorithms can increase tacit collusion by providing new opportunities for firms to link their prices to the prices their competitors post. This issue is the subject of recent academic (Salcedo, 2015) and policy (OECD, 2017) interest. While there is a growing awareness of these issues, determining the appropriate responses from competition authorities is still an open question, although there is a broad understanding that "the rise of pricing algorithms and AI software will require changes in our enforcement practices" (McSweeny, 2017). Of course, addressing these questions requires at least a basic understanding of the nature of machine learning algorithms, which is another important motivation for this chapter.

Another important regulatory issue is transparency. Firms are increasingly using machine learning to automate decisions that affect consumers in important ways, but in some cases this can increase opacity relative to human decision making. One example is the decision to grant credit. Financial institutions deploy machine learning algorithms to decide which kinds of consumer receive which types of loans, but consumers do not necessarily understand the key characteristics for predicting repayment risk. Regulators in this and other situations have a role to play in ensuring transparency and fairness.

Finally, much of the digital data valuable for machine learning applications is held by private sector companies whose main interest in exploiting it is commercial. To the extent that such data also has public value for research and policymaking, regulators will also be called upon to facilitate the transfer of data from the firms that directly collect it to a wider range of interested parties.

BIBLIOGRAPHY

ATHEY, S. (2017), "Beyond prediction: Using big data for policy problems," *Science,* 355: 483-485.

ATHEY, S.; BLEI, D.; DONNELLY, R.; RUIZ, F., and T. SCHMIDT (2018), "Estimating Heterogeneous Consumer Preferences for Restaurants and Travel Time Using Mobile Location Data," *American Economic Review Papers and Proceedings,* forthcoming.

Baker, S. R.; Bloom, N., and S. J. Davis (2016), "Measuring Economic Policy Uncertainty," *The Quarterly Journal of Economics,* 131(4): 1593-1636.

Bandiera, O.; Hansen, S.; Prat, A., and R. Sadun (2017), CEO Behavior and Firm Performance, *NBER Working Paper,* 23248.

Belloni, A.; Chen, D.; Chernozhukov, V., and C. Hansen (2012), "Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain," *Econometrica,* 80: 2369-2429.

— (2014), "High-Dimensional Methods and Inference on Structural and Treatment Effects," *Journal of Economic Perspective,* 28(2): 29-50.

Bernanke, B. S.; Boivin, J., and P. Eliasz (2005), "Measuring the Effects of Monetary Policy: A Factor-Augmented Vector Autoregressive (FAVAR) Approach," *Quarterly Journal of Economics,* 120(1): 387-422.

Blei, D.; Ng, A., and M. Jordan (2003), "Latent Dirichlet Allocation," *Journal of Machine Learning Research,* 3: 993-1022.

Breiman, L. (2001), "Statistical Modeling: The Two Cultures," *Statistical Science,* 16(3): 199-231.

Bühlmann, P., and S. van de Geer (2011), *Statistics for High-Dimensional Data: Methods, Theory, and Applications,* Springer Series in Statistics, Springer.

Cavallo, A., and R. Rigobon (2016), "The Billion Prices Project: Using Online Prices for Measurement and Research," *Journal of Economic Perspectives,* 30(2): 151-178.

Croft, J. (2017, May 4), "Artificial intelligence closes in on the work of junior lawyers," *Financial Times,* retrieved from www.ft.com

Davig, T., and A. Smalter Hall (2017), Recession Forecasting Using Bayesian Classification, *The Federal Reserve Bank of Kansas City Research Working Paper,* 16-06.

Einav, L., and J. Levin (2014), "Economics in the age of big data," *Science,* 346 (6210).

Erosheva, E. A.; Fienberg, S. E., and C. Joutard (2007), "Describing Disability through Individual-Level Mixture Models for Multivariate Binary Data," *The Annals of Applied Statistics,* 1(2): 502-537.

Glaeser, E. L.; Kim, H., and M. Luca (2017), Nowcasting the Local Economy: Using Yelp Data to Measure Economic Activity, *Harvard Business School Working Paper,* 18-022.

Gross, J. H., and D. Manrique-Vallier (2014), "A Mixed Membership Approach to the Assessment of Political Ideology from Survey Responses," in Airoldi, E. M.; D. Blei; E. A. Erosheva, and S. E. Fienberg, editors, *Handbook of Mixed Membership Models and Its Applications,* CRC Press.

Hansen, S.; McMahon, M., and A. Prat (2018), "Transparency and Deliberation on the FOMC: A Computational Linguistics Approach," *Quarterly Journal of Economics,* forthcoming.

Hartford, J.; Lewis, G.; Leyton-Brown, K., and M. Taddy (2017), Deep IV: A Flexible Approach for Counterfactual Prediction, *Proceedings of the 34th International Conference on Machine Learning.*

Hastie, T.; Tibshirani, R., and M. Wainwright (2015), *Statistical Learning with Sparsity: The Lasso and Generalizations,* Number 143 in Monographs on Statistics and Applied Probability. CRC Press.

Hoberg, G., and G. Phillips (2010), "Product Market Synergies and Competition in Mergers and Acquisitions: A Text-Based Analysis," *Review of Financial Studies,* 23(10): 3773-3811.

— (2016), "Text-Based Network Industries and Endogenous Product Differentiation," *Journal of Political Economy,* 124(5): 1423-1465.

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

Lee, D. D., and H. S. Seung (1999), "Learning the parts of objects by non-negative matrix factorization," *Nature,* 401(21 October 1999): 788-91.

Marr, B. (2015, September), "Big Data: 20 Mind-Boggling Facts Everyone Must Read," *Forbes,* retrieved from https://www.forbes.com/sites/bernardmarr/2015/09/30/big-data-20-mind-boggling-facts-everyone-must-read/#5741b5117b1e

McSweeny (2017), *Algorithms and Coordinated Effects. Remarks of Commissioner Terrell McSweeny,* University of Oxford Center for Competition Law and Policy, 22 May 2017.

MUELLER, H., and C. RAUH (2017), "Reading Between the Lines: Prediction of Political Violence Using Newspaper Text," *American Political Science Review,* forthcoming.

MULLAINATHAN, SENDHIL, and JANN SPIESS (2017), "Machine Learning: An Applied Econometric Approach," *Journal of Economic Perspectives,* 31(2): 87-106.

NG, A. Y., and M. I. JORDAN (2002), On Discriminative vs. Generative Classifiers: A comparison of Logistic Regression and Naive Bayes, *Neural Information Processing Systems.*

NIMCZIK, J. S. (2017), *Job Mobility Networks and Endogenous Labor Markets,* unpublished manuscript, Humboldt University Berlin.

OECD (2017), *Algorithms and Collusion: Competition Policy in the Digital Age,* www.oecd.org/competition/algorithms-collusion-competition-policy-in-the-digital-age.htm

SALCEDO, B. (2015), *Pricing Algorithms and Tacit Collusion,* unpublished manuscript, Pennsylvania State University.

STOCK, J. H., and M. W. WATSON (1999), "Forecasting inflation," *Journal of Monetary Economics,* 44: 293-335.

THORSRUD, L. A. (2016), Words are the new numbers: A newsy coincident index of business cycles, *Working Papers,* 4/2016, Centre for Applied Macro- and Petroleum economics (CAMP), BI Norwegian Business School.

TIBSHIRANI, R. (1996), "Regression Shrinkage and Selection via the LASSO," *Journal of the Royal Statistical Society Series B,* 58(1): 267-288.

VARIAN, H. R. (2014), "Big Data: New Tricks for Econometrics," *Journal of Economic Perspectives,* 28(2): 3-28.

WIGGLESWORTH, R. (2018, January 31), "Can big data revolutionise policymaking by governments?," *Financial Times,* retrieved from www.ft.com