Big Data in Economics

Matthew Harding[1] and Jonathan Hersh[2]

Keywords: Big Data, machine learning, prediction, causal inference

Elevator Pitch

The success of modern economics results to a large extent from the availability of data sources which allowed us to quantify human behavior, from individual purchases in supermarkets to the degree of interconnectedness between global financial markets. As such many economists are already comfortable working with large datasets, from financial transactions to census datasets. Without new analytic methods Big Data fails to achieve its potential. We encourage economists to pay close attention to recent developments in machine learning techniques.

Key Findings

Pros

- Complex data is now available characterized by large volume, fast velocity, diverse varieties and the ability to link many datasets together (deep data).
- Powerful new analytic techniques derived from machine learning are increasingly part of the mainstream econometric toolbox.
- With Big Data we can predict economic phenomena better and improve causal inference.

Cons

- The use of machine learning techniques requires us to learn new models and paradigms. Predictions based on Big Data may have privacy concerns.
- Methods are computationally intensive, may not have unique solutions and require a high-degree of fine tuning for optimal performance.
- Big data is costly to collect and store and analyzing it requires investments in technology and human skill

Author's Main Message

Big Data is expanding the volume, variety and veracity of the data as economists we have to explore questions of economic importance. However, we believe it's necessary to learn new methods developed to handle data at this scale to properly harness the power of Big Data.

[1] Department of Economics and Department of Statistics, 3207 Social Science Plaza B, Irvine CA 9297; www.DeepDataLab.org; Email: harding1@uci.edu
[2] The George L. Argyros School of Business and Economics, Chapman University, One University Drive, Orange, CA 92866; Email: hersh@chapman.edu

1. Motivation

If Big Data were simply a statement about the size of the data, it would have had little impact on the profession. The term Big Data entered the mainstream vocabulary around 2010 when people primarily in the tech world became cognizant of the exponential rate at which data was generated, primarily through the use of social media (Harding, 2014).

Engineers and computer scientists quickly realized that Big Data cannot be defined purely in terms of size however. While, it is certainly true that the *Volume* of data available to use has increased by orders of magnitude over the past decade, other factors have changed the informational landscape as well. As economists, we might as well approach Big Data from a costs and benefits perspective in order to make sense of some of the recent developments that people refer to when talking about "Big Data".

While traditionally data was only collected for a specific purpose often by a national statistical agency, we now live in an increasingly quantified world where even the smallest company records detailed and sometimes individualized data. This is done through a vast ecosystem of software (apps) and hardware (sensors) embedded in our increasingly "smart" technology: from phones to wifi connected appliances, and from cars to satellites. This data avalanche has dramatically increased both the *Variety* of data and the *Velocity* at which the data is recorded. New opportunities abound for creating novel datasets from previously unstructured information, such as text (Gentzkow, Kelly, and Taddy, 2017) and satellite images (Engstrom, Hersh, Newhouse, 2017). It has opened new areas of economic query and questions which previous could only be answered many months or even years after-the-fact can now be addressed in real time. Economists have thus moved from *forecasting* to *nowcasting* (Scott and Varian, 2015). We now have the ability to use real time Google searches to predict changes in unemployment (d'Amuri and Marcucci, 2012), or Yelp data to predict county business patterns (Glaeser, Kim and Luca., 2017).

A second important realization was that *data depth* adds economic (and analytic) value. While individual datasets can deliver impressive results, the true *Value* of data comes from our ability to link many seemingly unrelated datasets together (deep data) in order to provide a comprehensive understanding of economic behavior. Thus, for example, Harding and Lovenheim (2017) simulate the impact of food taxes in the US by constructing a database of over 100 million food purchase transactions, detailed nutritional information on over 1 million food items identified at the barcode level, detailed consumer demographics on each household, and a precise mapping of the store environment capturing every single store where the household might have purchased food. Smaller datasets were previously available to estimate food demand, but the accuracy of a policy simulation depends on the availability of multiple linked datasets which can be analyzed jointly.

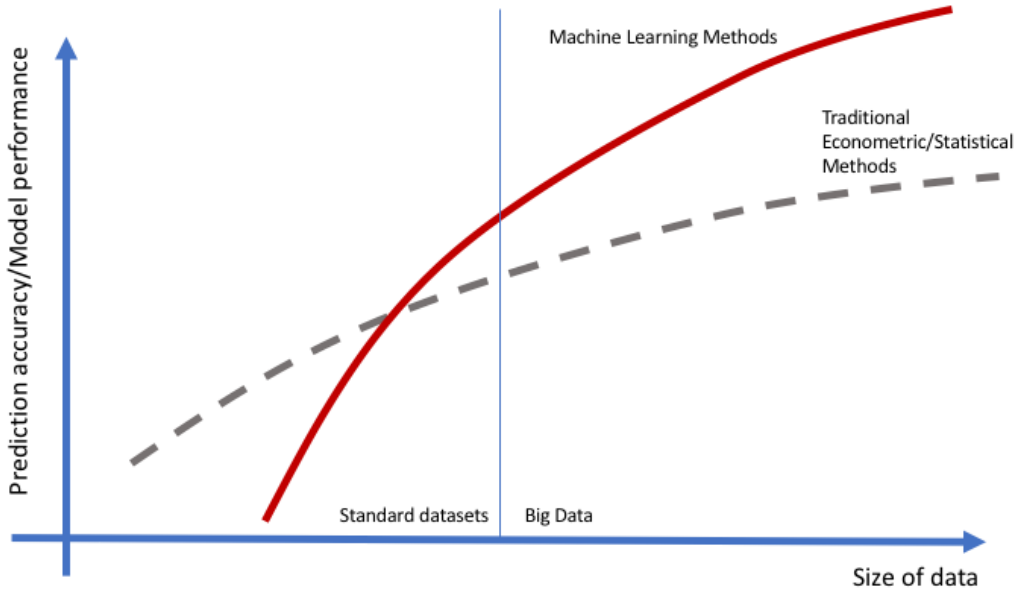

Figure 1: The use of machine learning techniques for Big Data analytics

In Figure 1 we attempt a stylized representation of the relationship between the size of available data and the expected value of different analytic approaches, where value is defined in terms of prediction accuracy or some related measure of model performance. In small datasets of the type commonly available in past decades, the traditional econometric or statistical tools tend to outperform more complex machine learning methods. The reason for this is that while standard statistical approaches are parsimonious and generally rely on optimizing a quadratic objective function, machine learning tools employ many parameters (often thousands to millions) and attempt to optimize complex objective functions with many local optima. Large datasets are thus necessary for machine learning tools to shine.

It is important to note that as the available data increases all methods will tend to improve in terms of their predictive accuracy. In recent years however, researchers have noticed that the performance of machine learning techniques tends to improve at a much faster rate. In the last couple of years even tasks once considered impossible for machines to perform (e.g. reading comprehension or playing complex games like Go) have been mastered by the latest generation of machine learning tools such as deep neural networks and their performance now exceeds that of expert humans (Brynjolfsson, Rock and Syverson, 2017). So where does this leave the average economics practitioner? If the data used tends to be small and relatively simple, we do not urge her to continue using existing methods and traditional software packages. If on the other hand she finds herself dealing with Big Data, we believe that learning new analytic paradigms from machine learning and investing in new software tools will lead to substantial performance improvement.

One common misconception about recent advances in Big Data tools is that they focus exclusively on prediction at the expense of causal inference. While, it is true that prediction is the major focus of machine learning as approached from a computer science perspective and causal inference has received comparatively less attention, this does not mean that the developments are irrelevant for causal inference. In fact, many econometricians have turned their attention to modifying machine learning algorithms to perform better causal inference (Athey and Imbens, 2017; Chernozhukov et al., 2017).

2. Discussion of Pros and Cons

Some basic machine learning terminology

We find that one of the deterrents to using machine learning is not the difference in conceptual approach of econometrics versus machine learning, but rather the unfamiliar terminology found in the machine learning literature.[3] What we call variables, machine learning refers to as features. Often the machine learning approach is similar to the econometric approach but because the terminology is different it falls on deaf economists' ears. Having taught machine learning courses to economists we find it instructive to cover some basic machine learning terminology before proceeding further.[4]

Machine learning is subdivided into *unsupervised* and *supervised* learning. Learning here is machine learning speak for fitting models to data. In supervised learning the goal is to fit a function to a target. Specifically, every data point x_i has an associated label y_i. The task of the supervised learning algorithm is to find a function $f(x_i)$ that finds a mapping between x_i and y_i. In econometrics, we just call this "regression", but machine learning researchers tend to use fancy names. Supervised learning occupies the lion's share of tasks that involve fitting models to data, in econometrics as well as currently in machine learning. If the goal is prediction -- that is learning a functional mapping between inputs and outputs and applying those out of sample -- machine learning methods such as random forest, LASSO, or deep neural networks will regularly beat econometric methods.

With unsupervised learning, in contrast, the goal is to find patterns in the data that reveal hidden structures, or interesting structures or patterns. In unsupervised learning, each data point, x_i does not have an underlying label y_i. The goal here is less well defined than with supervised learning. We might be trying to reduce the dimensionality of some very large object so that it fits in a smaller space (saving hard drives in the process.) Our goal might be to cluster observations into similar groups. (Think propensity score matching or creating a synthetic control.) We may be attempting to classify a large corpus of documents by topics, saving us (or our research assistants) the laborious task of having to read thousands of documents. Some unsupervised learning techniques may be familiar to the economist research. The reader may be familiar with Principle Component Analysis (PCA), a procedure that finds orthogonal variables that explain the maximal variance of underlying data, which is used in econometrics and forecasting models. The set of topics in unsupervised learning is large and growing, and much of it is unexplored in economics.

One key distinction between regression models in econometrics and supervised learning methods in machine learning is the type of model we are fitting to data. Machine learning methods were developed to handle terabytes of data, much larger than those we commonly encounter in economics. Therefore, the flexibility of the models that can be non-parametrically identified from data is usually greater than in economics. This creates a separate problem, however: how does the researcher know they are fitting true relationships to data and not those that have arisen spuriously from chance? Note the traditional null hypothesis significance testing here is of no use here. Given the sample sizes in the many millions of

[3] Mullainathan and Spiess (2017) provide a short and very readable introduction from an applied econometrics perspective.

[4] The best introduction to statistical machine learning is available in the monograph by James et. al. (2013). A more updated review focused on inference is provided by Efron and Hastie (2016). A slightly dated Bayesian perspective is provided by Murphy (2013). We should note that the role of Bayesian statistics is far from clear in the age of Big Data where the data strongly dominates any prior and Bayesian analysis is no different than maximum likelihood.

observations, achieving a threshold p-value of 0.05 is trivial. This is also why Big Data often implies different methods: we need fresh training and thinking to learn from datasets of this size.

The solution machine learning researchers developed to ensure their fitted model performs well out of sample is to approximate out of sample fit using a *testing-training-validation split* of the underlying data. In this approach, data (*test sample*) are completely reserved from the model fitting stage, left untouched until the very end of the analysis. The model is instead fit on a subject of the underlying data (*training sample*). Often in the model fitting process parameters of machine learning methods need to be calibrated or "tuned". To ensure this tuning process does not contaminate the fitting, a subset of the training data (*validation sample*) is reserved for to appropriately tune the parameter. Once the tuning parameter (or set of tuning parameters) have been selected the tuned model is applied to the validation sample to approximate how the model will perform in the wild. In economics, we are often concerned with losing observations, even if they are lost at random, because our datasets are relatively small. With Big Data, this isn't as much of a concern; we have ample data on which to train our models. Losing a few observations is worth it if in doing so we are able to more accurately test how our model will perform out of sample.

An important unmentioned step above is how to efficiently adjust the tuning parameter without overfitting or laboriously returning to the validation sample too often. The solution is a repeated sampling procedure known as *cross-validation*, which a clever way of choosing an optimal tuning parameter even without turning to the validation sample. In *k-fold cross-validation*, the training data are first partitioned into *k* distinct groups. Then a model is fit using all data *except* the data in partition 1. This fitted model is then applied to data in the first partition and predicted values are obtained for the first partition only. The process is repeated until we have predicted values for observations in every fold. The cleverness of this approach is that predictions of the model are never influenced by the value of that observation's dependent variables. Predictions really are "out of sample" or at least an approximation of it. One can and often does use cross validation to compare fitted values obtained for a variety of tuning parameters. The final selection is usually the one that minimizes a loss function, typically mean-squared error.

Example 1: Regularization and high-dimensional model in models of trade policy

One of the challenges of Big Data is having to manage larger datasets with many, even thousands, of variables. Without a clear prior of the underlying data generating process to direct your effort, time can be wasted inefficiently searching through options with nary a global optimum to be found. Fortunately, a few methods from machine learning may be of guidance to the intrepid economist lost in a sea of Big Data. In this section, we describe *regularization* and give a sample application in economics which will hopefully help you separate the wheat from the chaff.

Regularization is a statistical technique to adjust likelihoods such that when maximized they prefer sparse models (that is models with fewer variables or parameters) or models which shrink the value of coefficients towards zero.[5] Why do we prefer models with fewer variables? The reason is that we think sparser models are easier to interpret, to convey to co-authors and other researchers, and ultimately to policy-makers who might prefer the simple story to the more complex one. Given a choice between two models that perform equally well predicting out of sample, the one with fewer variables is usually preferred. Regularization has the added benefit that sparser models tend to have better out of sample predictive power than "dense" models, or models with many variables (James et. al. 2013). This has

[5] Regularization has a long tradition in econometrics. Jerry Hausman explored regularized regression in his graduate work at Oxford University, and which was published as Bacon and Hausman (1974). Applications of regularization in current econometric modeling is very common.

implications in the world of Big Data, where the number of possible variables include in a model can be measured in the hundreds or thousands. In some of our own research, we found that the gains to regularization increases with the size of the dataset (Afzal, Hersh, Newhouse, 2017). Once the dataset reaches a critical size, beyond the researcher's ability to guess at the data generating process perhaps, using regularization or another smart model selection technique is necessary for wringing additional insight from the data.

One popular statistical method of regularization is the LASSO estimator, or Least Absolute Selection and Shrinkage Operator (Tibshirani, 1996). LASSO looks a lot like ordinary least squares, which is part of the reason it's one of the more successful machine learning techniques to be integrated into economics. LASSO takes a standard regression squared loss function (other GLM functions may be used) and adds an ℓ_1 penalty on the magnitude of the coefficients to the likelihood. Under some mild conditions, this forces some of the estimated coefficients to be zero if they prove not useful in maximizing the likelihood. Depending on the degree of sparsity desires, more or less regularization may be used, which is carefully controlled by a smoothing parameter. To find the optimal amount of regularization, we use cross-validation to determine which parameter performs best.

One application of LASSO regularization applied in economics that may be instructive is the use by Baxter and Hersh (2017), who apply several machine learning techniques to understand country and policy determinants that could predict trade decline following the great recession. Data on trading patterns between countries really are Big Data. Bilateral trade patterns measure trade between a given country and all its trading partners, and a panel of these data over sufficient time can easily grow the number of observations in the many millions. Compounding the problem is that the choice of covariates to use can often be intractably large. Even using the gravity model of trade as a theoretical guideline, the inclusion of which policy and country level determinants of trade -- such as tariffs, monetary regimes, banking and other crises – is left to the discretion of the researcher. The authors employ LASSO regularization, along with random forests, to discipline the selection of variables and determine a model that predicts trade flows into the crisis.

The authors split the panel into training and test samples, covering the years pre-crisis (1970-2008) and post crisis (2009-2011) respectively. They fit models in the pre-crisis years, and after estimating coefficients, use these to predict into the post crisis years and compare model fit. The authors find that Lasso regularization "zeros out" many variables that are typically included in models of bilateral trade. Moreover, in Figure 2 we can view the so called "shrinkage page" of standardized coefficients and show how the magnitude of the coefficients (shown on the y-axis) change as we increase the amount of regularization (shown in the x-axis, increasing shrinkage penalty moving right). The shrinkage path reveals in some sense the ordering of importance of variables for predicting bilateral trade flows. Some variables, such as distance, GDP, common currency, WTO membership, and human capital) remain non-zero even as the shrinkage penalty is increased to very high levels. Other variables become zero as only a little shrinkage penalty is applied.

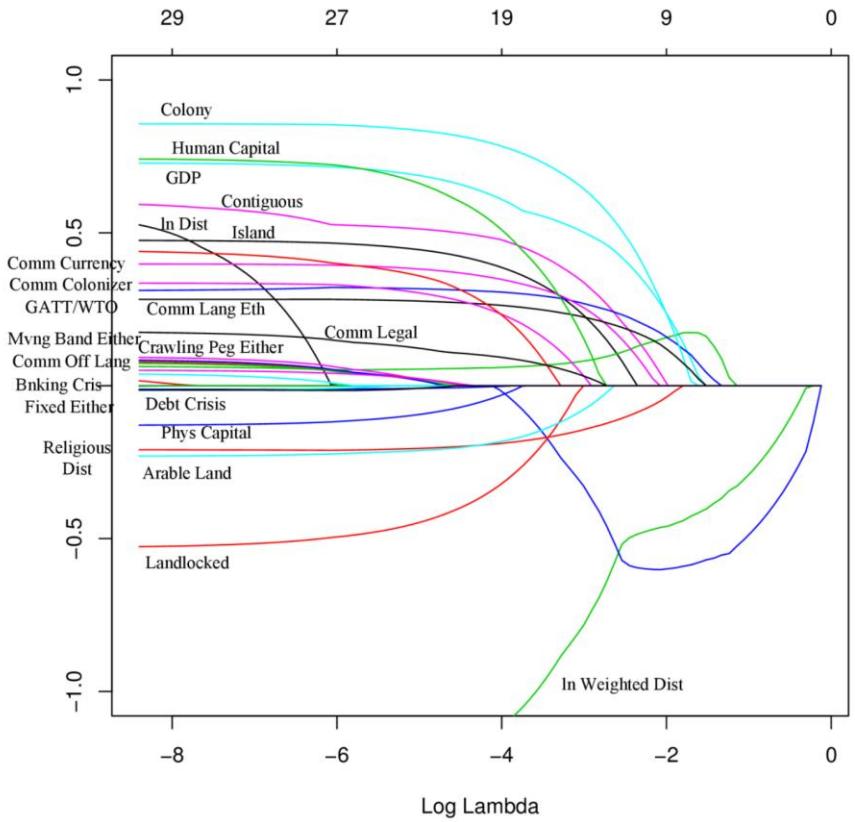

Figure 2: Variable selection as a function of the degree of regularization

Example 2: Clustering and demand modeling

It is important to note that using Big Data or machine learning does not mean that an economist needs to approach an analysis from a radically different viewpoint. Very often machine learning tools enhance the existing econometric methodology by grounding modeling decisions in data as opposed to unreliable human intuitions which manifest themselves as modeling choices. Let us consider the use of machine learning techniques in Harding and Lovenheim (2017), which was already mentioned earlier as an example of a paper employing Big Data. The aim of the paper is to construct a structural food demand model and simulate the impact of different product and nutrient taxes in the US. The existing food transaction data (scanner data) is vast and allows one to identify precise cross-price elasticities and account for detailed socio-demographic gradients and purchasing environments. The richness of the data also implies that food purchases are observed for over 1.1 million distinct food items (each identified by a unique barcode). Does this mean that we ought to estimate a demand system with 1.1 million equations (and an even larger set of covariates)? This is not currently feasible and even if it were feasible it would undoubtedly only obfuscate the policy implications by making the results uninterpretable. It would be equally obfuscating to aggregate the products at broad levels (e.g. a beverage category would conflate sugar sweetened beverages, diet soda, tea, bottled water etc.). Given 1.1 million products it would be hopeless to try and argue our way to an optimal grouping of the products.

This is where machine learning can help, and a more robust approach is to use an algorithm to cluster the products into distinct groups based on their detailed nutrition profiles (e.g. calories, fat, sugar etc). We can think of each product as a point in a high dimensional space where each coordinate measures the amount of a specific nutrient contained in that product. We would thus expect diet sodas to be close to

each other in this space, and quite far away from frozen pizzas. Numerous clustering algorithms were developed in machine learning for this task (as mentioned before this is an example of an unsupervised learning task since the algorithm is not predicting anything but rather attempts to find the underlying structure of the data). The analysis employs a popular algorithm called *k-medians clustering* which identifies the median of each cluster and labels all products as belonging to one of *k* such clusters.

It might be instructive to consider a simple pseudocode of the clustering algorithm (Figure 3).

Step 1: Initialize k cluster centers c_k at random.
Step 2: Assign each data point x_i to the cluster with the nearest center c_k. The assigned cluster is given by $\hat{k}(x_i) = argmin_k d(x, c_k)$, where $d(a, b)$ measures the distance between two points a and b
Step 3: For each cluster, re-compute the center

$$c_k = \frac{1}{n_k} \sum_{x_i \in C_k} x_i$$

Step 3: Convergence check: Have clusters moved?
Step 4: If not converged, go to iterate from Step 2.

Figure 3: Pseudo-code for the clustering algorithm

While this is probably one of the simplest algorithms available, it is often very effective. In the Harding and Lovenheim (2017) application each point $x_i = (x_{1i}, x_{2i}, \dots, x_{mi})$ corresponds to a unique product sold in US stores, and each product is an m-vector of attributes such as nutrient amounts. The algorithm was able to learn that diet soda is a separate category from regular soda, an important fact from a policy perspective, since a sugar sweetened beverage tax would shift consumption towards non-soda products but also encourage consumers to substitute diet soda for regular soda.

This algorithm has many properties of commonly encountered data science procedures, so it is worth reflecting on some of them. In particular,

1. The user has to choose the number of clusters k desired. As Harding and Lovenheim (2017) discuss, typically that is not known. The algorithm has to be re-run for many values of k and the user chooses the best fit.
2. The algorithm does not guarantee that the best fit is achieved or even that the solution provided is unique. Generically speaking, clustering is rather simple, and it requires one to compare all possible combinations. In practice, this would not be possible and algorithms such as this, while not guaranteeing a globally optimal or unique solution nevertheless provide a "good" solution which solves an otherwise insurmountable problem.
3. The user has to choose a metric $d(a, b)$ which measures the distance between two points a and b in an $m-$dimensional space. The Euclidean distance or the Mahalanobis distance are commonly used in Econometrics. Many hundreds of such metrics have been constructed and some may be more appropriate than others. Thus, while domain knowledge is not essential, it may be helpful. A computer scientist could run a clustering algorithm but an expert in say economics or biology may know that one distance metric is more suitable to a given problem than another one.
4. Successfully running this algorithm may require some additional choices, for example the initial allocation of the points to clusters ("choosing starting values") may lead to different outcomes. Similarly, stopping the iterations too soon may lead to poor performance. These choices are

often referred to as *fine tuning* and in modern data science they play an important role. Unfortunately, we have found that these choices are often poorly documented in the computational literature and thus the performance of machine learning algorithms is often hard to replicate from one paper to the next. Especially over the last few years as more and more cutting-edge research is being conducted in industry and away from academia in this area, many very important aspects of the best performing algorithms are proprietary and will probably never be made public.

5. While computationally intensive, the algorithm is easily implemented, fast and scales to large datasets. An attractive feature of many machine learning algorithms is that while they demand substantial computing resources, simplicity and scalability make them successful in Big Data applications.

Harding and Lovenheim (2017) then show that all the 1.1 million food products can in fact be categorized into 34 mutually exclusive and exhaustive groups. A structural system of equations is then fit to these 34 groups (after suitable aggregation and the construction of relevant price indexes). The structural parameters can then be estimated using a simple iterative non-linear least squares algorithm. While the model still has hundreds of parameters it can be estimated relatively fast on an average server. The ability to correctly estimate the full cross-price elasticity turns out to be crucial since taxing some products may be counterproductive from a public health perspective as consumers substitute towards equally unhealthy products. Furthermore, they find that a sugar tax has a much larger impact than a soda tax since it has a much larger tax base and by taxing all products proportionally to their sugar content it discourages substation towards unhealthy food items. As this example shows, machine learning can be used an enabling device to a more sophisticated economic analysis without it being the primary focus for an economist.

Causal inference

Traditionally machine learning has focused to a large degree on prediction problems. While many policy problems are at their core prediction problems (Kleinberg et. al., 2015), where for example the policy makers has to predict the duration of unemployment in order in order to best target job training programs, other policy problems require knowledge of counterfactuals and the estimation of causal treatment effects (Athey, 2017). The key insight is that a machine learning method which predicts very well may not necessarily provide unbiased estimates of the structural parameters. In fact, the parameters are often of secondary interest and may not even be identified even though a method has excellent the mean square prediction error properties.

The Neyman-Rubin causal model is the most common framework for causal inference in applied economics and most economists are familiar with methods such as matching, propensity score weighting or regression discontinuity.[6] Recent econometrics work on Big Data builds on this framework by asking how machine learning methods can be employed or modified in order to also provide unbiased estimates of key parameters such as average treatment effects. When approaching the Big Data literature economists might be surprised to find several competing notions of causality in the computer science literature. Not all these concepts are mutually consistent and may not be appropriate for economic policy evaluations. For example, graph theoretic notions of causality abound based on the work of Pearl (2009) which are a more restricted version of the simultaneous equations models popular in the earlier econometrics literature (Heckman and Pinto, 2015). Thus, we caution economists from applying Big Data algorithms labelled as "causal" in the computer science literature to policy analysis without fully

[6] See Athey and Imbens (2017) for a recent review article on this methodology and the integration of machine learning techniques.

understanding their theoretical underpinnings. Fortunately, the econometrics literature at the intersection of causal inference and machine learning is making rapid and very successful strides.

Double Machine Learning

A growing and very successful new econometric literature asks how unbiased estimates of key structural parameters such as average treatment effects can be obtained in Big Data problems. One simple example concerns the estimation of an average treatment effect in a high dimensional regression model where the econometrician has hundreds of potential control variables that can be included either on the right-hand side of the estimating equation or in the construction of the propensity score. Belloni, Chernozhukov and Hansen (2014) show that reducing the dimensionality of the problem through the use of standard penalty-based methods from machine learning such as LASSO leads to badly biased treatment effects. The problem is essentially one of misspecification where errors in the selection induce endogeneity. This problem has now been much studied and is well understood. Extensions include the use of machine learning methods from random forests to neural networks for the estimation of the propensity score or for the estimation of various nonlinear functions in generalized regression models. The solution involves the use of "double machine learning" (Chernozhukov et. al., 2017) which involves a two-step procedure. In the first step, any machine learning method can be employed to estimate the unknown functions of interest. In the second step residuals are used to construct orthogonal moment conditions which are then being solved to produce an unbiased estimate of the treatment effect. In order to avoid overfitting, the procedure also relies on sample splitting to remove finite sample biased. It is worth noting the extent to which Big Data lead to the blending of traditional econometric techniques with machine learning concepts.

Heterogeneous Treatment Effects

Another area where machine learning has been used for causal inference is recent work on using random forests to estimate heterogeneous treatment effects (Athey and Wager, 2017). The treatment of any intervention is likely to vary by participant characteristics. For example, a new cancer medicine might have a differential impact for older rather than younger patients. Estimating heterogeneous treatment effects has proven challenging. Using ex-post data where the treatment was randomized researchers may be accused of artificially choosing subgroups in which the treatment effect is largest. One solution is to provide pre-analysis plans, with subgroup analysis pre-determined prior to program initialization. With small number of participants, it may be impossible to sufficiently subdivide the population prior to treatment assignment. The solution is to use a training sample of the dataset to build regression trees to partition the predictor space into subgroups. Complexity of the tree is determined through cross-validation. On the remaining portion of the dataset, their estimator calculates the sample average treatment effect within a leaf. The key insight here is that the subgroup partitions were determined using data not used to identify the treatment effect. This prevents "data mining" to search for subgroups in which the treatment effect is highest, while still allowing for ex-post heterogeneous subgroup analysis.

Limitations

With high dimensional individualized data being generated at an unprecedented rate, and breakthroughs in abilities to process these data, Big Data poses a host of limitations and concerns. A primary concern is maintaining privacy. Surveys are typically carefully constructed to anonymize individuals so that sensitive information does not point to any given household. With Big Data, information in the wild that is de-identified may be ex-post identified given machine learning matching tools. To give a sense of the scope of the problem, de Montjoye et al., (2015) find that only four spatio-temporal credit card data points are sufficient to uniquely identify 90% of individuals in their database of credit card transactions. Policy

around data security needs to be designed with this new reality in mind – many more kinds of data should be considered sensitive and should have additional security considerations.

Another concern is how the broad use of machine learning predictions in policy and businesses may have unintended consequences. Racism may be unintentionally embedded into algorithms by using correlates of race as proxies (Caliskan, Bryson, and Narayanan, 2017). If these algorithms are sufficiently "black box", the racism may be unknown even to the algorithm builders themselves. We need strong checks to ensure that algorithmic predictions have their intended effect and are not unintentionally contributing to racial bias.

3. Summary and Policy Advice

The rise of Big Data is an exciting time for the ambitious economist. Never before has so much data been available to test our theories and develop new ones. As economists, we have a natural edge in that we are used to working with large datasets, however this edge is rapidly declining. Newer methods from machine learning are expanding a researcher's ability to handle Big Data at scale, and we risk being cut off from the frontier if these methods are not expanded into a researcher's toolkit. Don't forget economists have a natural advantage in understanding how to construct and test causal statements that make them rapidly valuable in a data saturated world. Learning how to implement these methods at scale remains a challenge. We suggest departments encourage more collaboration between computer scientists and economists to encourage two-way knowledge sharing. Eventually these methods should be taught as part of the core empirical sequence in PhD programs, just as Econometrics was expanded to include causal estimation methods. The future is Big (Data) and bright.

References

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

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

Brynjolfsson, E., Rock, D., & Syverson, C., 2017. *Artificial intelligence and the modern productivity paradox: A clash of expectations and statistics* (No. w24001). National Bureau of Economic Research.

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

d'Amuri, F. and Marcucci, J., 2012. The predictive power of Google searches in forecasting unemployment, Bank of Italy.

Afzal, M., Hersh, J. and Newhouse, D., 2017. *Building a Better Model: Variable Selection to Predict Poverty in Pakistan and Sri Lanka*. World Bank. Mimeo.

Bacon, R.W. and Hausman, J.A., 1974. The relationship between ridge regression and the minimum mean square error estimator of Chipman. *Oxford Bulletin of Economics and Statistics*, *36*(2), pp.115-124.

Baxter, M. and Hersh, J., 2017. Robust Determinants of Bilateral Trade. In *2017 Meeting Papers* (No. 591). Society for Economic Dynamics.

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

Caliskan, A., Bryson, J. J., & Narayanan, A,. 2017. Semantics derived automatically from language corpora contain human-like biases. Science, 356(6334), 183-186.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. and J. Robins, 2017. Double machine learning for treatment and causal parameters. *arXiv preprint arXiv:1608.00060*.

De Montjoye, Y. A., Radaelli, L., & Singh, V. K., 2015. Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science*, 347(6221), 536-539.

Efron, B. and Hastie, T., 2016. *Computer Age Statistical Inference* (Vol. 5). Cambridge University Press.

Engstrom, R., Hersh, J. and Newhouse, D., 2017. Poverty from space: using high resolution satellite imagery for estimating economic well-being and geographic targeting. *Unpublished paper*.

Gentzkow, M., Kelly, B.T. and Taddy, M., 2017. *Text as data* (No. w23276). National Bureau of Economic Research.

Glaeser, E. L., Kim, H., & Luca, M., 2017. *Nowcasting the Local Economy: Using Yelp Data to Measure Economic Activity* (No. w24010). National Bureau of Economic Research.

Harding, M., 2014. Good Data Public Policies, in *The Future of Data-Driven Innovation*, US Chamber of Commerce.

Harding, M. and Lovenheim, M., 2017. The effect of prices on nutrition: comparing the impact of product-and nutrient-specific taxes. *Journal of Health Economics*, *53*, pp.53-71.

Heckman, J. and Pinto, R., 2015. Causal analysis after Haavelmo. *Econometric Theory*, *31*(1), pp.115-151.

James, G., Witten, D., Hastie, T. and Tibshirani, R., 2013. *An introduction to statistical learning* (Vol. 112). New York: springer.

Kleinberg, J., Ludwig, J., Mullainathan, S. and Obermeyer, Z., 2015. Prediction policy problems. *The American economic review*, *105*(5), pp.491-495.

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

Murphy, K.P., 2012. *Machine learning: a probabilistic perspective*. MIT press.

Pearl, J., 2009. *Causality*. Cambridge university press.

Scott, S.L. and Varian, H.R., 2015. Bayesian variable selection for nowcasting economic time series. In *Economic analysis of the digital economy* (pp. 119-135). University of Chicago Press.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp.267-288.