Hersh, Jonathan; Harding, Matthew

Article

Big Data in economics

IZA World of Labor

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Hersh, Jonathan; Harding, Matthew (2018) : Big Data in economics, IZA World of Labor, ISSN 2054-9571, Institute of Labor Economics (IZA), Bonn, https://doi.org/10.15185/izawol.451

This Version is available at:
https://hdl.handle.net/10419/193433

MATTHEW HARDING
University of California, Irvine, USA
JONATHAN HERSH
Chapman University, USA

I Z A
World of Labor
Evidence-based policy making

Big Data in economics

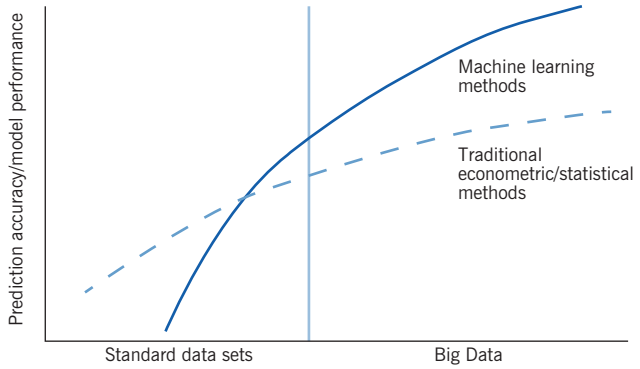
New sources of data create challenges that may require new skills

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

ELEVATOR PITCH

Big Data refers to data sets of much larger size, higher frequency, and often more personalized information. Examples include data collected by smart sensors in homes or aggregation of tweets on Twitter. In small data sets, traditional econometric methods tend to outperform more complex techniques. In large data sets, however, machine learning methods shine. New analytic approaches are needed to make the most of Big Data in economics. Researchers and policymakers should thus pay close attention to recent developments in machine learning techniques if they want to fully take advantage of these new sources of Big Data.

The use of machine learning techniques for Big Data analytics



Source: Author's own compilation.

I Z A
World of Labor

KEY FINDINGS

Pros

- Complex data are now available, characterized by large volume, fast velocity, diverse varieties, and the ability to link many data sets together.
- Powerful new analytic techniques derived from machine learning are increasingly part of the mainstream econometric toolbox.
- Big Data allows for better prediction of economic phenomena and improves causal inference.
- Machine learning techniques allow researchers to create simple models that describe very large, complex data sets.
- Machine learning methods and Big Data also allow for the complex modeling of relationships that predict well beyond the sample.

Cons

- Predictions based on Big Data may have privacy concerns.
- Machine learning methods are computationally intensive, may not have unique solutions, and may 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.
- Big Data may suffer from selection bias depending on how and by whom data are being generated.
- Access to these data may involve partnering with firms who limit researcher freedom.

AUTHOR'S MAIN MESSAGE

Due to the prevalence of connected digital devices, observational data sets are now available that are much larger and of higher frequency than traditional surveys: so-called Big Data. This has created opportunities for economists and policymakers to learn about economic systems and choices with a higher degree of precision. However, new methods, particularly those related to machine learning, are needed to take full advantage of Big Data. Furthermore, policymakers should consider a broader range of data as sensitive, researchers need checks to avoid unintentional bias, and economists should learn general-purpose coding languages.

MOTIVATION

The term "Big Data" entered the mainstream vocabulary around 2010 when people became cognizant of the exponential rate at which data were being generated, primarily through the use of social media [1]. Engineers and computer scientists quickly realized that Big Data cannot be defined purely in terms of size; while it is certainly true that the volume of data has increased by orders of magnitude over the past decade, other factors have changed the informational landscape as well.

While data were traditionally only collected for a specific purpose, often by a national statistical agency, the world is becoming increasingly quantified, where even the smallest company collects and records detailed and sometimes individualized data. This is done through a vast ecosystem of software (apps) and hardware (sensors) embedded in the vast sea of "smart" technology, including phones, Wi-Fi connected appliances, cars, and satellites. This data avalanche has dramatically increased both the variety of data and the velocity at which the data are recorded. New opportunities abound for creating novel data sets from previously unstructured information, such as text [2] and satellite images [3]. This development has opened new areas of economic query; questions which previously 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. For instance, it is now possible to use real time Google searches to predict changes in unemployment [4], or Yelp data to predict local business patterns [5].

DISCUSSION OF PROS AND CONS

It is important to note that as the amount of 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 has tended to improve at a much faster rate. Tasks once considered impossible for machines to perform (e.g. reading comprehension or playing complex games, such as Go) have now been mastered by the latest generation of machine learning tools, such as deep neural networks, and their performance now exceeds that of expert humans. So, where does this leave the average economics practitioner? If the available data tend to be small and relatively simple, then existing methods and traditional software packages should suffice. If, on the other hand, researchers find themselves dealing with Big Data, then 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, from a computer science perspective, prediction is the major focus of machine learning, and that 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 [6].

Some basic machine learning terminology

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 latter's literature. Often, the machine learning approach is similar to the econometric approach, but because the terminology is different it falls on deaf ears. What economists call "variables," for instance, machine learning refers to as "features." It is thus instructive to cover some basic machine learning terminology before proceeding further.

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 has an associated label or target. The task of the supervised learning algorithm is to find a function that finds a mapping between each data point and its associated label. In econometrics, this is simply called "regression." 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, least absolute selection and shrinkage operator (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, data points do not each have an underlying (associated) label. The goal here is less well defined than with supervised learning. For instance, one might try to reduce the dimensionality of some very large object (i.e. a very large data set) so that it fits into a smaller dimensional space (saving hard drive space in the process). The goal might be to cluster observations into similar groups, or to classify a large corpus of documents by topics, saving researchers the laborious task of having to read thousands of documents. 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 being fit to the data. Machine learning methods were developed to handle terabytes of data, much larger than those commonly encountered in economics. Therefore, the flexibility of the models that can be nonparametrically 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 is of limited use here. Given sample sizes in the many millions of observations, magnitude is more important than statistical significance. This is also why Big Data often implies different methods; as such, researchers need fresh training and thinking to learn from data sets of this size.

The solution that machine learning researchers developed to ensure their fitted models perform 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 excluded from the model fitting process, and left untouched until the very end of the analysis. The model is instead fitted 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 to appropriately tune the parameter. Once the tuning parameter (or set of tuning parameters) has been selected, the tuned model is applied to the test sample to approximate how the model will perform in the wild.

Economists are often concerned with losing observations, even if they are lost at random, because the data sets are relatively small. With Big Data, this is not as much of a concern; there are ample data on which to estimate models. Losing a few observations is therefore not a huge price to pay if it is ultimately possible to more accurately test how the model will perform out of sample.

An important step not mentioned 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 is 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 that partition only. The process is repeated until one has predicted values for observations in every fold. The cleverness of this approach is that the model's predictions 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. Cross-validation can be and often is used to compare fitted values obtained for a variety of tuning parameters. The final selection is usually the one that minimizes error between observed and predicted values.

Regularization to assist with variable selection in high-dimensional trade policy models

One of the challenges of Big Data is having to manage larger data sets with many, even thousands of, variables. Without a clear prior understanding of the underlying data generating process to direct one's effort, time can be wasted inefficiently searching through options. Fortunately, a few methods from machine learning may be of guidance to the intrepid researcher lost in a sea of Big Data.

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 toward zero. Why are models with fewer variables preferable? The reason is that sparser models are generally easier to interpret and to convey to researchers and, ultimately, policymakers. Given a choice between two models that perform equally well, 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. This has implications in the world of Big Data, where the number of possible variables included in a model can be measured in the hundreds or thousands. Indeed, recent research has found that the gains from regularization increase with the size of the data set [7]. Once the data set reaches a critical size, beyond the researcher's ability to guess at the data generating process, 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 [8]. LASSO looks a lot like a traditional econometric method (namely ordinary least squares (OLS)), which is part of the reason it is one of the more successful machine learning techniques to be integrated into economics. LASSO takes a standard regression and adds a cost ("shrinkage penalty") to increasing the magnitude of the regression coefficients. Under

some mild conditions, this forces some of the estimated coefficients to be zero if they do not prove useful in explaining the dependent variable. Depending on the degree of sparsity desired, more or less regularization may be used, which is carefully controlled by a smoothing parameter. To find the optimal amount of regularization, cross-validation is used to determine which parameter performs best.

One application of LASSO regularization in economics that may be instructive is its use in helping to understand country and policy determinants that 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 into the many millions. Compounding the problem is the choice of which covariates to use, which can often be intractably large. Moreover, deciding which policy- and country-level determinants of trade to include—such as tariffs, monetary regimes, banking and other crises—is left to the discretion of the researcher. One research team employs LASSO regularization, along with random forests, to discipline the selection of variables and determine a model that predicts trade flows during the crisis [9].

The authors split the panel into training and test samples, covering the years pre- (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, the method 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. In other words, the most reliable predictors of trade flows are the variables that remain non-zero after regularization shrinkage is applied. These variables may more reliably predict trade flows during crisis periods, while others may just be adding statistical noise.

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. Consider the use of machine learning techniques in a recent study from 2017 [10]. The aim was 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 (i.e. the change in demand of one good in reaction to the price change of another good) and to account for detailed differences in socio-demographic characteristics 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 a demand system with 1.1 million equations (and an even larger set of covariates) should be estimated? This is not currently feasible, and even if it were, it would undoubtedly only obfuscate the policy implications by making the results uninterpretable. It would be equally confusing to aggregate the products at broad

levels (e.g. a beverage category would conflate sugar-sweetened beverages, diet soda, tea, and bottled water). Given 1.1 million products, it would be hopeless to try and argue what would be an optimal grouping of the products.

This is where machine learning can help by providing a more robust approach using an algorithm to cluster products into distinct groups based on their detailed nutrition profiles (e.g. calories, fat, and sugar). 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. Diet sodas would be expected 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 (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.

While this is probably one of the simplest algorithms available, it is often also very effective. For instance, it was able to learn that diet soda is a separate category from regular soda. This is an important fact from a policy perspective, since a sugar-sweetened beverage tax would shift consumption toward non-soda products, and would also encourage consumers to substitute diet soda for regular soda. Furthermore, the study finds that a sugar tax has a much larger impact than a soda tax because it has a much larger tax base; by taxing all products proportionally to their sugar content, it discourages substitution toward unhealthy food items. 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 toward equally unhealthy products.

The *k*-medians clustering algorithm has many properties of commonly encountered data science procedures, so it is worth reflecting on some of them. In particular:

- The user has to choose the number of clusters *k* desired, but this is typically not known. The algorithm has to be re-run for many values of *k* and the user chooses the best fit.
- 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 a comparison of all possible combinations. In practice, this would not be possible; algorithms such as this, while not guaranteeing a globally optimal or unique solution, nevertheless provide a good solution, which solves an otherwise insurmountable problem.
- The user has to choose a metric which measures the distance between two points in the dimensional space. 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.
- Successfully running the algorithm may require some additional choices; for example, the initial allocation of 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, 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. Over the last few years especially, more and more cutting-edge research is being conducted in industry and away from academia, this means that many very important aspects of the best performing algorithms are proprietary and will probably never be made public.

- While computationally intensive, the algorithm is easily implemented, fast, and scales to large data sets. 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.

As this example shows, machine learning can be used as an enabling device to a more sophisticated economic analysis, without it being the primary focus of the research output.

Machine learning methods to improve causal inference

Traditionally, machine learning has focused to a large degree on prediction problems. While many policy problems are, at their core, prediction problems (where, for example, policymakers have to predict the duration of unemployment in order to best target job training programs), others require knowledge of counterfactuals and the estimation of causal treatment effects. It is important to note that machine learning methods may not necessarily provide unbiased estimates of the structural parameters, despite being very good at prediction. For example, LASSO coefficients are biased toward zero, which will under or overstate the regressor's impact on the dependent variable if it is taken at face value. However, because correlation does not imply causation, one should seldom interpret even unbiased OLS coefficients as causal. Understanding causal impact of policy variables requires a separate framework such as instrumental variables, matching estimators, or other methods.

The Neyman-Rubin causal model is the most common framework for causal inference in applied economics. Recent 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. Researchers might be surprised to find several competing notions of causality in the computer science literature on Big Data. Not all these concepts are mutually consistent, and may not be appropriate for economic policy evaluations. Thus, economists should be wary of applying Big Data algorithms labeled as "causal" in the computer science literature to policy analysis without fully understanding their theoretical underpinnings. Fortunately, the econometrics literature at the intersection of causal inference and machine learning is making rapid and very successful strides.

A growing and successful new branch of 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 to consider. A 2014 study shows that reducing the dimensionality of the problem through the use of standard machine learning methods

such as LASSO leads to badly biased treatment effects [11]. The problem is essentially one of misspecification, where errors in the selection induce endogeneity. This problem is now well understood. The solution involves a two-step procedure called "double machine learning." In the first step, any machine learning method can be employed to estimate the unknown functions of interest. In the second step, errors in the regression are used to construct restrictions on the data, which are then 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 biases. It is worth noting the extent to which Big Data lead to the blending of traditional econometric techniques with machine learning concepts. Machine learning methods can construct more complex data relationships with ease. Sample splitting ensures models are not fitting noise in the data. Big Data has provided a wealth of information to exploit, and machine learning is the tool that disciplines the cavalcade of data.

Another area where machine learning has been used for causal inference is in estimating heterogeneous treatment effects, which has traditionally proven challenging [12]. 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 patients against younger ones. 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, in which subgroup analysis is pre-determined prior to program initialization. However, in cases with small numbers of participants, it may be impossible to sufficiently subdivide the population prior to treatment assignment. The random forests method in machine learning allows researchers to partition subgroups based on data that are 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 AND GAPS

The unprecedented rate at which high-dimensional individualized data are being generated, along with breakthroughs in the processing of these data, means that 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 that is de-identified may be ex-post identified using machine learning matching tools. To give a sense of the scope of the problem, one study finds that as few as four spatio-temporal credit card data points are sufficient to uniquely identify 90% of individuals in a database containing supposedly de-identified credit card transactions [13]. 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. If these algorithms are sufficiently opaque, the racism may be unknown even to the algorithm builders themselves. Strong checks are therefore needed to ensure that algorithmic predictions have their intended effect and are not unintentionally contributing to racial bias.

Depending on how the data are generated, Big Data may suffer from selection bias. Not everyone has the same propensity to use digital devices, or any given app or website, resulting in possible biases, particularly when making generalizations about subgroups which happen to be represented in the data. Appropriate weighting and selection methods must be used to ensure results are representative without strong caveats. Another concern is that many valuable sources of Big Data are often controlled by private firms that limit researcher freedom. It is common to see veto rights in data sharing agreements, which could result in publication bias in terms of the results that are allowed to be published. Firms are encouraged to be transparent and allow researchers freedom in this respect, and policymakers should consider policy to address this issue.

SUMMARY AND POLICY ADVICE

The rise of Big Data is an exciting time for the ambitious economist, policymaker, or social scientist. Never before has so much data been available to test existing theories and develop new ones. Economists have a natural edge in this endeavor, as they are used to working with complex data. However, this edge is rapidly declining. Newer methods from machine learning are expanding the ability to handle Big Data at scale, and researchers risk being cut off from the frontier if these methods are not incorporated into their toolkit. Economists understand how to construct and test causal statements, which makes their skills highly valuable in a data-saturated world. The challenge lies in learning how to implement these methods at scale.

Researchers and policymakers should take seriously the limitations and gaps of Big Data, but also its potential. Issues of privacy, selection bias, and private firms controlling research output derived from their data remain serious concerns going forward. Beyond the limited advice in dealing with these issues discussed above, further collaboration should be encouraged between computer scientists and data scientists to develop two-way knowledge sharing. Economists are also encouraged to "get their hands dirty" and learn to code in more general-purpose languages, such as Python or R, to develop a theoretical and practical understanding of machine learning methods. Eventually, these methods should be taught as part of the core empirical sequence in graduate programs, just as econometrics was expanded to include causal estimation methods, and policy organizations should consider offering training courses in these methods.

Acknowledgments

The authors thank an anonymous referee and the IZA World of Labor editors for many helpful suggestions on earlier drafts.

Competing interests

The IZA World of Labor project is committed to the *IZA Guiding Principles of Research Integrity*. The authors declare to have observed these principles.

REFERENCES

Further readings

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

Efron, B., and T. Hastie. *Computer Age Statistical Inference*. Volume 5. Cambridge: Cambridge University Press, 2016.

Key references

[1] Harding, M. "Good data public policies." In: *The Future of Data-Driven Innovation*. Washington, DC: US Chamber of Commerce, 2014; pp. 43–53.

[2] Gentzkow, M., B. T. Kelly, and M. Taddy. *Text as Data*. NBER Working Paper No. 23276, March 2017.

[3] Engstrom, R., J. Hersh, and D. Newhouse. *Poverty from Space: Using High-Resolution Satellite Imagery for Estimating Economic Well-Being*. World Bank Group Policy Research Working Paper No. 8284, December 2017.

[4] D'Amuri, F., and J. Marcucci. *The Predictive Power of Google Searches in Forecasting Unemployment*. Bank of Italy Economic Working Papers No. 891, November 2012.

[5] Glaeser, E. L., H. Kim, and M. Luca. *Nowcasting the Local Economy: Using Yelp Data to Measure Economic Activity*. NBER Working Paper No. 24010, November 2017.

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

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

[8] Tibshirani, R. "Regression shrinkage and selection via the LASSO." *Journal of the Royal Statistical Society. Series B (Methodological)* 58:1 (1996): 267–288.

[9] Baxter, M., and J. Hersh. *Robust Determinants of Bilateral Trade*. Society for Economic Dynamics Meeting Paper No. 591, 2017.

[10] Harding, M., and M. Lovenheim. "The effect of prices on nutrition: Comparing the impact of product- and nutrient-specific taxes." *Journal of Health Economics* 53 (2017): 53–71.

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

[12] Wager, S., and S. Athey. "Estimation and inference of heterogeneous treatment effects using random forests." *Journal of the American Statistical Association* (Forthcoming).

[13] de Montjoye, Y. A., L. Radaelli, V. K. Singh, and A. Pentland. "Unique in the shopping mall: On the reidentifiability of credit card metadata." *Science* 347:6221 (2015): 536–539.

Online extras

The **full reference list** for this article is available from:

https://wol.iza.org/articles/big-data-in-economics

View the **evidence map** for this article:

https://wol.iza.org/articles/big-data-in-economics/map