SYMPOSIUM

Can Big Data Solve the Fundamental Problem of Causal Inference?

Rocío Titiunik, *University of Michigan*

The access to "big data" has opened up new possibilities for research. Among other things, we now can collect a person's geographic location, genetic information, blog posts, product preferences, and social media interactions, giving scientists access to vast amounts of information that was previously unavailable. Understandably, this has caused great excitement, and some scholars believe that big data holds the key to many of the scientific answers that so far have eluded us. I believe that only part of this optimism is justified. Regarding our ability to make causal inferences, the relevant question is whether big data has the potential to uncover causal relationships that could not be discovered with "small" data. I argue that only rarely is the amount of data to which we have access the binding constraint in our ability to make causal inferences. Instead, the bottleneck often is the lack of a solid research design and a credible theory, both of which are essential to develop, test, and accumulate causal explanations. This does not mean that big data has no benefits. The access to more and new information leads to richer descriptive analysis, opens up new possibilities for exploratory analysis and hypothesis generation, and increases the number of theoretical implications that may be empirically tested.

In the view of causation that I adopt, units have different potential outcomes that would occur for each possible level of a given treatment, but we are able to see only one of those outcomes—that which is realized when the unit chooses a specific level of the treatment. The fundamental problem of causal inference is that for every unit, we fail to observe the value that the outcome would have taken if the chosen level of the treatment had been different (Holland 1986). Therefore, the search for causal inferences is a search for assumptions under which we can infer the values of these unobserved counterfactual outcomes from observed data.[1] The question at the center of my argument is whether access to big data fundamentally increases the likelihood that those assumptions will hold.

The term "big data" may be ascribed various meanings. In their contribution to this symposium, Patty and Penn understand big data as a highly multidimensional and complex body of information that is not inherently ordered and that necessarily must go through a process of data reduction before it can be analyzed. In contrast, in the definitions of big data that I consider, "big data" refers to a rectangular array of information with n rows and p columns. Thus, my notion of big data as a rectangle presupposes that the reduction discussed by Patty and Penn already has occurred and that appropriate decisions have been made to transform complex, non-ordered data into variables that take particular values for each observation—decisions that, as Patty and Penn point out in this issue, can be crucially consequential.

Another view of big data, and one that Ashworth, Berry, and Bueno de Mesquita adopt in their contribution to this symposium, refers to the statistical and computational tools of machine learning that typically are used to "mine" or extract structured patterns of information from large datasets.[2] Although this is a valuable definition for many purposes, I do not engage it directly because my argument centers on whether the access to vast amounts of new information per se will solve the most fundamental obstacles to causal inference, and I consider this question to be largely independent of the available techniques to detect and characterize empirical regularities in big datasets.[3]

Instead, my focus is on two other interpretations of big data: big data "as large n" and big data "as large p." The former refers to data sources with several observations—that is, records collected for hundreds of thousands or millions of units of analysis. In this interpretation, big data represents a situation in which the number of observations, n, is extremely large relative to the number of variables available in the dataset. In contrast, the latter interpretation refers to the availability of a large amount of information per observation—that is, data sources in which the number of variables, p, is very large relative to n. In the following discussion, I argue that big data is no "silver bullet" for causal inference under either of these two interpretations.

BIG DATA AS LARGE N

Access to big data in the sense of large n rarely translates into a fundamentally improved ability to make causal inferences. Larger n is helpful for increasing the precision of estimates or the power of hypothesis tests, and it can allow for a wider range of estimation methods that would be unreliable with few observations (e.g., nonparametric methods). However, large n does not automatically remove or even alleviate what is often the most challenging step of empirical research: our ability to estimate consistently the parameters of interest and to make valid and robust statistical inferences.

An estimator that is inconsistent remains inconsistent regardless of how large the number of observations used. Indeed, the definition of an inconsistent estimator is precisely that even after we allow the sample size to go to infinity, the estimator is unlikely to be near the true value of the parameter. For example, no increase in the number of observations,

doi:10.1017/S1049096514001772

no matter how large, will cause the omitted variable bias in a mis-specified linear regression model to disappear. This is a rather obvious but important point.

Perhaps a more subtle point is that even when consistent estimators or pivotal test statistics are available, increasing the sample size will not necessarily lead to more accurate approximations if we consider all relevant data-generating processes. For example, given a distribution generating the data, confidence intervals will have correct asymptotic coverage for a causal parameter if the probability that these intervals contain the true parameter is arbitrarily close to a predetermined confidence level when n is very large. However, this does not necessarily imply that there will be an n large enough to guarantee approximately correct coverage for all possible distributions that could have generated the data. In other words, even if fixing a data-generating process we can obtain confidence intervals with approximately correct coverage for some large n_o, there is no guarantee that n_o will be large enough to ensure an equally accurate or better approximation for a different potential data-generating process.

This is the well-known distinction between pointwise and uniform convergence for large n. With only pointwise convergence approximations, increasing the sample size does not guarantee that we get closer to the true value of a parameter or the true coverage of confidence intervals in any meaningful practical sense. Consider, for example, post-model-selection inference. As discussed by Leeb and Pötscher (2005), with consistent model-selection procedures, the probability of selecting the true model generating the data tends to 1 as the sample size increases. However, this asymptotic result is pointwise, not uniform, which implies that for any sample size—regardless of how large—the probability of selecting the true model will tend to zero when some of the parameters in the true model are sufficiently small. Thus, although it is possible to construct consistent estimators for the finite sample distribution of the post-model-selection estimator, these estimators are of very limited practical use because the lack of uniformity implies that this finite-sample distribution can be arbitrarily far from its pointwise asymptotic limit for any sample size.

BIG DATA AS LARGE P

A related but different understanding of big data centers on the larger universe of models that researchers can explore. The increasing availability of information has led to datasets that, instead of (or in addition to) having a large number of observations (n), have a very large number of variables (p) for each observation. In this interpretation, "big data" refers to datasets with large p, which in turn may refer to cases in which p is either smaller than n but very large relative to it or strictly larger than n.

The fundamental obstacle that we encounter when attempting to make causal inferences is that our units of analysis (e.g., voters, politicians, or businesses) typically choose the courses of action that they expect will be most beneficial. As a result, units that do or do not take certain actions may be very different from one another in potentially unknown but systematic ways. This problem, sometimes referred to as "selection bias," is particularly pervasive in the social sciences due to the strategic nature of most social interactions.

The increasing availability of information has led to datasets that, instead of (or in addition to) having a large number of observations (n), have a very large number of variables (p) for each observation.

As argued previously, big data as large n per se cannot solve the problem of self-selection. However, big data as large p brings more information. A large p dataset may contain information on a person's preferred newspapers, political registration, recent travel history, fitness routine, social media contacts, campaign contributions, business transactions, and blog entries. Such a dataset surely will provide a more complete picture of the individual than a typical "small" dataset with only demographic and socioeconomic information. It is tempting to imagine that such massive quantities of new information will dramatically improve our ability to make causal inferences by allowing us to "control for" a very large set of variables—so large that the systematic preexisting differences among units will all but disappear.

Consider the following example. We want to understand whether the high rates of reelection of incumbent members of the US House of Representatives arise because they are the highest quality candidates or because, once in office, they have access to resources that scare off strong challengers and increase their popularity among voters. To establish which explanation is true, we may seek a measure of incumbency advantage net of intrinsic incumbent quality. A possible strategy is to measure the vote share of incumbent politicians and compare it to that of non-incumbent candidates of the same party. However, what if the incumbent anticipated a defeat and this anticipation prompted his retirement? And what if the anticipation of a tough battle makes it difficult for the party to recruit a strong non-incumbent candidate? The result might be that in some districts in which no incumbent is running, the party's candidate may be weaker and the electoral circumstances more adverse than in incumbent-held seats—which likely would overestimate the true advantage enjoyed by incumbents.

This is when big data as large p might help. It may be implausible that candidate quality and retirement decisions are exogenous conditional on only party and previous vote share, but what if we had access to big data on politicians and

could condition on all previous political speeches, public text messages, press reports, election forecasts, social media interactions, and travel history? Perhaps then we would capture enough information so that after conditioning on it our inferences would be valid. This idea is promising indeed because it implies that big data as large p eventually will allow us to solve automatically the fundamental problem of causal inference by "controlling for" massive amounts of information using sophisticated algorithms, computers, and statistical assumptions—all of which likely would be necessary to address the complications of large p inferences (see the following discussion).

Could an automatic and sophisticated "kitchen-sink" approach eliminate the need to rely on theories and research designs? I suspect not because there is one major catch. When the goal is causal inference rather than prediction or description, inference methods for large p datasets rely on the crucial assumption that the large set of variables included in the model is approximately equal to or, at least, includes all of the variables needed to make exogeneity plausible. In other words, given a large set of baseline variables, high-dimensional methods can be used to perform dimension reduction or to adjust for the potential biases induced by including many regressors. However, these methods will work only under the assumption that the original set of candidate variables includes all variables necessary for exogeneity to hold and excludes all variables affected by the treatment.

How do we know if our large p dataset contains all relevant controls and excludes all posttreatment variables?

And how do we know if our large p dataset contains all relevant controls and excludes all posttreatment variables? We need either a theory about how the treatment affects the outcome of interest and the other variables in the model or a research design that justifies focusing on a particular subset of variables—and, if we want to maximize our chances of accumulating knowledge, we need both.

Moreover, even if we could identify a potentially large subset of our original large p dataset that includes those variables that make exogeneity possible and excludes those affected by the treatment (a big if), crucial restrictions and additional assumptions are needed to avoid the inferential challenges that otherwise would arise. For example, it is well known that when p is smaller than n but very large relative to it, the large number of incidental parameters poses considerable complications. Increasing the number of fixed-effects in nonlinear panel data models may lead to inconsistent parameter estimates (see, e.g., Hahn and Newey 2004), and even classical heteroskedasticity robust standard errors may be inconsistent in linear panel models (Stock and Watson 2008). Similar issues may arise in cross-sectional linear models with many regressors (see, e.g., Cattaneo, Jansson, and Newey 2012; Koenker 1988) and instrumental variable models with many instruments (see, e.g., Andrews and Stock 2007). These issues may be addressed by imposing appropriate rate restrictions or bias corrections, but the important point is that increasing p leads to inferential challenges that cannot be ignored.

In addition, recent advances in econometrics have focused on developing uniformly valid methods for causal inference with high-dimensional datasets in which p is allowed to be much larger than n. These approaches begin with an extremely large number of variables, perform model selection to choose only those that are needed, and develop conditions under which valid inferences can be made after the model-selection step (see, e.g., Belloni et al. 2013; 2014; Farrell 2014).[4] However, these methods crucially require some type of data-reduction structure, such as sparsity—that is, the assumption that only some of the included variables appear in the true model. In other words, although these methods allow for a massive (i.e., larger than n) number of variables in the conditioning set, valid causal inferences cannot be obtained unless we impose restrictions on the way those variables affect the outcome of interest—restrictions that can be justified only by providing a strong research design or a strong theory.

Therefore, the need for theory and research design with big data as large p is not fundamentally altered, even if large p methods provide more flexibility by allowing researchers to use a very large number of variables. Causal inferences based on large p datasets still require the assumption that we have not omitted important variables and have not included posttreatment variables, the same assumption that is required when datasets are "small."

Returning to our example, is the collection of all previous political speeches, public text messages, press reports, social media interactions, and international travel records sufficient to capture a politician's inherent quality? What about the way he kisses babies at rallies, how she pronounces vowels, his ability to connect emotionally with voters, and her intelligence? The list could be expanded endlessly, and it is likely that many items on that list would be inherently unobservable. Without a theory and a research design, it is not possible to know when to stop adding to the list.

BIG DATA FOR DESCRIPTION, EXPLORATION, AND HYPOTHESIS GENERATION

That big data is no substitute for theory and research design does not mean that it has no benefits. At the very least, the ability to collect and analyze big datasets creates numerous possibilities for description, exploration, and hypothesis generation, all of which are crucial elements of scientific inquiry.

At the most basic level, big data allows us to systematically and quantitatively analyze phenomena that were previously unavailable. This leads to an increasing availability of "dependent" and "independent" variables that, together with modern machine learning tools, enhances our ability to provide a richer description of phenomena of interest and allows us to uncover empirical regularities previously

unobservable. Because the ability to engage in systematic descriptive analysis often is a first step in developing scientific theories and explanations, this is one significant way in which access to big data can contribute to scientific progress.

A related benefit of big data is that as more phenomena become quantifiable, the range of implications of scientific theories that can be tested empirically is expanded significantly. For example, until recently, it was difficult to test systematically those theories that predict shifts in spoken or written political discourse. However, the increasing ability to convert text into data expands the number of predictions that can be tested (see, e.g., Grimmer 2013). This represents another contribution of big data to scientific progress.

Moreover, the availability of big data—in particular, big data as large p—allows researchers to engage in systematic, large-scale statistical inference (i.e., the simultaneous exploration of hundreds and even thousands of hypotheses) with the goal of detecting a small subset that holds the most promise for a specific scientific explanation (see, e.g., Efron 2010). The promise of this type of large-scale exploratory analysis is the ability to accelerate the pace of hypothesis generation. For example, in genetics, researchers may test hundreds or thousands of genes to detect a few that are most likely associated with a particular disease. After this small set of genes has been detected, efforts then can be focused on developing theories and designing specific clinical trials to develop a deeper scientific understanding of the causes of the disease. In the absence of large-scale exploratory analysis, each hypothesis would be studied with the same amount of detail, which may result in a prohibitively costly and slow process.

Finally, for these benefits of big data to materialize, we must ensure that the quality of big data satisfies certain standards, particularly in those cases in which it is created for direct business purposes. Many of the commercial sources of big data (e.g., Google, Twitter, and Facebook) are appealing in that they record detailed information about individuals' social interactions, preferences, consumption decisions, and so forth—all of which can be valuable for scientific purposes. However, as noted by Lazer et al. (2014), a major challenge in using commercial data is that the companies that produce it constantly make modifications to their data-collection algorithms in order to increase profits and support their business model. As a consequence, commercial data often are endogenously affected by a company's business decisions rather than exogenously determined—which compromises their validity as a source of information. For example, the number of social media connections per individual might be used as a proxy for a person's social network. However, if these connections are affected by the company's algorithms to suggest new contacts and the algorithms evolve over time in ways unknown to the scientific community, this type of data might prove highly unreliable for scientific purposes. These potential challenges must be considered as commercial data becomes more common in scientific research.

To summarize, in combination with theory and research design (and precautions about commercial sources), big data can enhance and deepen our ability to explore and develop scientific explanations. However, the availability of big data per se does not constitute a structural breakthrough in our ability to make causal inferences. This is not because big data is limited but rather because the accumulation of scientific knowledge ultimately requires a theory of how and why phenomena occur as well as a research design to make valid causal inferences about the theory's empirical implications. There are no algorithmic or automatic shortcuts to scientific discovery. In fact, the need for critical thinking will be stronger the more we become inundated with ever-larger amounts of new information.

There are no algorithmic or automatic shortcuts to scientific discovery. In fact, the need for critical thinking will be stronger the more we become inundated with ever-larger amounts of new information.

ACKNOWLEDGMENTS

I am grateful to Matt Golder for organizing the roundtable "Big Data, Causal Inference, and Formal Theory: Contradictory Trends in Political Science" at the 2014 Annual Meeting of the Midwest Political Science Association, which led to this symposium. This paper benefited greatly from discussions with the members of that roundtable (i.e., Ethan Bueno de Mesquita, Matt Golder, Justin Grimmer, Luke Keele, John Patty, and Josh Tucker), as well as the audience, and with Matias Cattaneo, Max Farrell, Josh Clinton, and Bill Yeaton. I gratefully acknowledge financial support from the National Science Foundation (SES 1357561). ■

NOTES

1. For various perspectives from different disciplines, see, for example, Gerber and Green (2012); Heckman and Vytlacil (2007); Imbens and Wooldridge (2009); Pearl (2009); Rosenbaum (2010); Rubin (2005); Sekhon (2009); and Van der Laan and Robins (2003).
2. For a detailed treatment of machine learning tools, see, for example, Hastie, Tibshirani, and Friedman (2009); Michalski, Carbonell, and Mitchell (1983); and Witten and Frank (2005). For a classic treatment of data mining and specification searches from an econometric perspective, see Leamer (1978).
3. Of course, these techniques increase our ability to quantify previously inaccessible phenomena and, in that sense, can be helpful to the causal-inference goal (see the distinction made in this issue by Ashworth, Berry, and Bueno de Mesquita between machine learning for datafication and for inference-oriented data analysis). However, the fundamental problem of causal inference remains regardless of how accurately and how many empirical regularities are detected.
4. In contrast, Berk et al. (2013) adopt a view of model selection that allows p to be larger than n when the goal is not to select a model in which the parameters have a causal interpretation but rather in which the targets of inference are "submodel" parameters that describe the association between covariates and outcome variables. Thus, in this approach, the meaning of a particular parameter depends on which other covariates are included in the model.

REFERENCES

Andrews, Donald W. K., and James H. Stock. 2007. "Inference with Weak Instruments." In *Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress, Vol. III*, ed. Richard Blundell, Whitney K. Newey and Torsten Persson, 122–73. New York: Cambridge University Press.

Belloni, Alexandre, Victor Chernozhukov, Iván Fernández-Val, and Christian Hansen. 2013. "Program Evaluation with High-Dimensional Data." Available at arXiv preprint arXiv:1311.2645.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2014. "High-Dimensional Methods and Inference on Structural and Treatment Effects." *Journal of Economic Perspectives* 28 (2): 29–50.

Berk, Richard, Lawrence Brown, Andreas Buja, Kai Zhang, Linda Zhao, *et al.* 2013. "Valid Post-Selection Inference." *The Annals of Statistics* 41 (2): 802–37.

Cattaneo, Matias D., Michael Jansson, and Whitney K. Newey. 2012. "Alternative Asymptotics and the Partially Linear Model with Many Regressors." Working paper. Ann Arbor: University of Michigan.

Efron, Bradley. 2010. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*, Volume 1. Cambridge: Cambridge University Press.

Farrell, Max H. 2014. "Robust Inference on Average Treatment Effects with Possibly More Covariates than Observations." Working paper. Chicago: University of Chicago Press.

Gerber, Alan S., and Donald P. Green. 2012. *Field Experiments: Design, Analysis, and Interpretation*. New York: W. W. Norton.

Grimmer, Justin. 2013. "Appropriators Not Position Takers: The Distorting Effects of Electoral Incentives on Congressional Representation." *American Journal of Political Science* 57 (3): 624–42.

Hahn, Jinyong, and Whitney Newey. 2004. "Jackknife and Analytical Bias Reduction for Nonlinear Panel Models." *Econometrica* 72 (4): 1295–319.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning*, second edition. New York: Springer.

Heckman, James J., and Edward J. Vytlacil. 2007. "Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation." *Handbook of Econometrics* 6:4779–874.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81 (396): 945–60.

Imbens, Guido W., and Jeffrey M. Wooldridge. 2009. "Recent Developments in the Econometrics of Program Evaluation." *Journal of Economic Literature* 47 (1): 5–86.

Koenker, Roger. 1988. "Asymptotic Theory and Econometric Practice." *Journal of Applied Econometrics* 3 (2): 139–47.

Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. "The Parable of Google Flu: Traps in Big Data Analysis." *Science* 343 (6176): 1203–5.

Leamer, Edward E. 1978. *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. New York:Wiley.

Leeb, Hannes, and Benedikt M. Pötscher. 2005. "Model Selection and Inference: Facts and Fiction." *Econometric Theory* 21 (01): 21–59.

Michalski, Ryszard S., Jaime G. Carbonell, and Tom M. Mitchell. 1983. "An Overview of Machine Learning." In *Machine Learning: An Artificial Intelligence Approach*, ed. Ryszard S. Michalski, Jaime G. Carbonell, and Tom M. Mitchell, 3–23. Palo Alto, CA: Tioga Publishing Co.

Pearl, Judea. 2009. *Causality: Models, Reasoning and Inference*, Volume 29, second edition. Cambridge: Cambridge University Press.

Rosenbaum, Paul R. 2010. *Design of Observational Studies*. New York: Springer.

Rubin, Donald B. 2005. "Causal Inference Using Potential Outcomes." *Journal of the American Statistical Association* 100 (469): 322–31.

Sekhon, Jasjeet S. 2009. "Opiates for the Matches: Matching Methods for Causal Inference." *Annual Review of Political Science* 12: 487–508.

Stock, James H., and Mark W. Watson. 2008. "Heteroskedasticity: Robust Standard Errors for Fixed Effects Panel Data Regression." *Econometrica* 76 (1): 155–74.

Van der Laan, Mark J., and James M. Robins. 2003. *Unified Methods for Censored Longitudinal Data and Causality*. New York: Springer.

Witten, Ian H., and Eiber Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*, second edition. Burlington, MA: Morgan Kaufmann.