COLLOQUIUM INTRODUCTION

Drawing causal inference from Big Data

Richard M. Shiffrin[a,1]

Human society has found the means to collect and store vast amounts of information about every subject imaginable, and is archiving this information in attempts to use it for scientific, utilitarian (e.g., health), and business purposes. These large databases are colloquially termed Big Data. How big Big Data are is of course a matter of perspective, and can range, for example, from the "tiny" amount of data sociologists and climate scientists dealt with many years ago to everything being posted on the World Wide Web; Big Data can arise from relatively well-controlled experiments or from uncontrolled sets of natural observations.

The advent of the age of Big Data poses enormous challenges, because collecting and storing the data are only a minimal first step and this step is not by itself helpful. The challenges can be divided into stages: finding potentially interesting patterns in the data, explaining those patterns (possibly with the help of experimental manipulations of some variables coupled with additional data collection), and then using the patterns and explanations for a variety of purposes. Finding interesting patterns is itself a daunting task, because a hallmark of Big Data is the fact that it vastly exceeds human comprehension. Imagine a relatively small dataset a terabyte in size classified along 50 identifiable dimensions. One might define an interesting pattern as a significant correlation of any subset of these variables with any other. This would not be a useful definition because the numbers of potentially significant correlations would be far too large to search with any existing computational techniques, and because most of the correlations, even if found, would be too complex to be useful and interesting to humans (e.g., if some 23-way interaction were somehow identified). Thus, finding interesting patterns is an enormous field of its own and involves the generation of efficient machine search algorithms, interactive and dynamic visualizations, and a large dose of informed human judgment. A first step along these lines was the subject of a 2003 Sackler Colloquium and a subsequent PNAS special issue "Mapping Knowledge Domains," that I organized with Katy Borner (1).

Finding patterns of correlations in the data is a necessary first step, but not in itself the goal: one must explain and use the information. Human explanation is almost always couched in terms of causal forces; that is the way we try to understand the infinite complexities of the world we inhabit. That was the subject of the recent Sackler Colloquium and is the focus of the present set of articles that emerged from the colloquium. However, there are enormous difficulties facing researchers trying to draw causal inference from or about some pattern found in Big Data: there are almost always a large number of additional and mostly uncontrolled confounders and covariates with correlations among them, and between them and the identified variables. This is particularly the case given that most Big Data are formed as a nonrandom sample taken from the infinitely complex real world: pretty much everything in the real world interacts with everything else, to at least some degree.

Thus, the challenges lie on every front: How does one define causality and degrees of causality in ways that make sense for large recurrent interacting systems? This issue is at least implicit and in a few cases explicit in the present articles. How does one judge what is a significant pattern or correlation? How does one explain the causal pathways? These questions and their answers are to a large extent a matter of statistical practice and implementation. However, the standard ways to use statistics that were developed to deal with, say, two-by-two tables of the result of crop plantings, are nowhere close to being applicable to the complexities of Big Data. Thus, many of the present articles are aimed at solving some of the difficult statistical problems raised by Big Data.

Going further, we can ask: How does one find ways to use the patterns and their explanations for a variety of utilitarian purposes, like health and business? Some of the present articles deal with this aspect of Big Data.

In many ways, the problems of Big Data are those of science generally, but writ small. In science we find ways to measure the environment we inhabit, identify interesting patterns ("interesting" being assessed in

[a]Department of Psychological and Brain Sciences, Indiana University, Bloomington, IN 47405

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "Drawing Causal Inference from Big Data," held March 26–27, 2015, at the National Academies of Sciences in Washington, DC. The complete program and video recordings of most presentations are available on the NAS website at www.nasonline.org/Big-data.

Author contributions: R.M.S. wrote the paper.

The author declares no conflict of interest.

[1]Email: shiffrin@indiana.edu.

www.pnas.org/cgi/doi/10.1073/pnas.1608845113

part by reference to previous causal explanations), try to come up with explanations of the patterns, test those explanations in mostly controlled experiments, and use the explanations to further scientific progress, to guide new research, and to accomplish a variety of utilitarian goals. It is important to note that testing proposed explanations with interventions may be difficult or impossible for many forms of Big Data. In scientific settings it is usually possible to manipulate carefully at least some of the many variables of Big Data. However, Big Data collected via natural observations may not lend itself to controlled manipulation, although there is a counterexample included in the present set of articles. In cases where intervention is not feasible, one can try to take advantage of unplanned and accidental interventions that are in the current set of Big Data, albeit these mostly occur in conjunction with many other uncontrolled covariates.

Drawing causal inference from Big Data is a daunting task, one requiring new development and novel thinking. There are many different aspects to this task, and they are presently being pursued actively and vigorously by many individuals and groups worldwide, because even partial advances can produce immense payoffs for society in such forms as scientific understanding, health, and business. The present articles represent a tiny sample of these efforts, but serve to illustrate the present state of the art.

Not unlike Big Data itself, the content of the present articles is highly multidimensional, not lending itself to any one linear ordering. Thus, the following brief statements about the contributions follow the order in which they appear in print, an order that is largely arbitrary.

Hal R. Varian (2) presents "Causal inference in economics and marketing," a survey of econometric methods based on counterfactual reasoning, and argues that machine learning methods can be used in this context.

Eckles, et al. (3) present "Estimating peer effects in networks with peer encouragement designs," describing a large randomized experiment (on Facebook) aimed at demonstrating the effects of receiving additional feedback from networked friends on individual behavior.

Levitt et al. (4) present "Quantity discounts on a virtual good: The results of a massive pricing experiment at King Digital Entertainment," a natural very large-scale field study about the results of discounts on behavior and resultant revenue.

Hripcsak et al. (5) present "Characterizing treatment pathways at scale using the OHDSI network," an analysis of a very large health database (Observational Health Data Sciences and Informatics) to improve understanding of treatment choices and to produce better designs for clinical inference.

Hawrylycz et al. present (6) "Project MindScope: Inferring cortical function in the mouse visual system," describing methods used in the project aimed to map and understand mammalian neocortex and its action.

Elias Bareinboim and Judea Pearl (7) present "Causal inference and the data-fusion problem," describing the counter-factual basis for causal inference and methods for combining data from heterogeneous databases to allow the drawing of causal inference.

Susan Athey and Guido Imbens (8) present "Recursive partitioning for heterogeneous causal effects," giving methods for estimating heterogeneity in causal effects and testing differences, using an "honest" data-driven estimation approach based on regression trees.

Meinshausen et al. (9) present "Methods for causal inference from gene perturbation experiments and validation," giving a new statistical method called "invariant causal prediction" for assigning probabilities to inferred casual structures, with applications to biology.

Higgins et al. (10) present "Improving massive experiments with threshold blocking," giving a new and sophisticated form of blocking in large-scale experimentation that should produce data much more amenable to causal interpretation.

Heckerman et al. (11) present "Linear mixed model for heritability estimation that explicitly addresses environmental variation," describing a way to take spatial location into account when using Big Data to tackle the age-old causal question about nature versus nurture.

Bloniarz et al. (12) present "Lasso adjustments of treatment effect estimates in randomized experiments," providing a "Lasso" method for overcoming the limitations of linear multivariate regression for dealing with large numbers of covariates.

Finally, Schölkopf et al. (13) present "Modeling confounding by half-sibling regression," showing how to remove the effect of confounders in large-scale data, and give an application to astronomy.

Acknowledgments

I give special thanks to Jill Sackler, who gave generous support in honor of her husband Arthur M. Sackler that enabled an outstanding series of colloquia, including the present one. A thank you is also due to the National Science Foundation for funding travel grants to the colloquium for many young researchers. Thanks are due to the staff of PNAS for the difficult task of organizing the assessments and finalization of the present set of submissions, particularly David Stopak and Elizabeth Huhn. Finally, very special thanks are to be given to my co-organizers, Susan Dumais from Microsoft, Michael Hawrylycz from the Allen Foundation, Jennifer Hill from New York University, Michael Jordan from UC Berkeley, Bernhard Schölkopf from the Max Planck at Tuebingen, and Jasjeet S. Sekhon from the University of California, Berkeley. The co-organizers not only gave talks at the colloquium and contributed papers to this special issue, but also acted as action editors for the PNAS submissions.

1 Shiffrin RM, Börner K (2004) Mapping knowledge domains. *Proc Natl Acad Sci USA* 101(Suppl 1):5183–5185.
2 Varian HR (2016) Causal inference in economics and marketing. *Proc Natl Acad Sci USA* 113:7310–7315.
3 Eckles D, Kizilcec RF, Bakshy E (2016) Estimating peer effects in networks with peer encouragement designs. *Proc Natl Acad Sci USA* 113:7316–7322.
4 Levitt SD, List JA, Neckermann S, Nelson D (2016) Quantity discounts on a virtual good: The results of a massive pricing experiment at King Digital Entertainment. *Proc Natl Acad Sci USA* 113:7323–7328.
5 Hripcsak G, et al. (June 6, 2016) Characterizing treatment pathways at scale using the OHDSI network. *Proc Natl Acad Sci USA* 113:7329–7336.
6 Hawrylycz M, et al. (2016) Inferring cortical function in the mouse visual system through large-scale systems neuroscience. *Proc Natl Acad Sci USA* 113:7337–7344.
7 Bareinboim E, Pearl J (2016) Causal inference and the data-fusion problem. *Proc Natl Acad Sci USA* 113:7345–7352.
8 Athey S, Imbens G (2016) Recursive partitioning for heterogeneous causal effects. *Proc Natl Acad Sci USA* 113:7353–7360.
9 Meinshausen N, et al. (2016) Methods for causal inference from gene perturbation experiments and validation. *Proc Natl Acad Sci USA* 113:7361–7368.
10 Higgins MJ, Sävje F, Sekhon JS (2016) Improving massive experiments with threshold blocking. *Proc Natl Acad Sci USA* 113:7369–7376.
11 Heckerman D, et al. (2016) Linear mixed model for heritability estimation that explicitly addresses environmental variation. *Proc Natl Acad Sci USA* 113:7377–7382.
12 Bloniarz A, Liu H, Zhang C-H, Sekhon JS, Yu B (2016) Lasso adjustments of treatment effect estimates in randomized experiments. *Proc Natl Acad Sci USA* 113:7383–7390.
13 Schölkopf B, et al. (2016) Modeling confounding by half-sibling regression. *Proc Natl Acad Sci USA* 113:7391–7398.