CHANCE

ISSN: 0933-2480 (Print) 1867-2280 (Online) Journal homepage: https://www.tandfonline.com/loi/ucha20

A Second Chance to Get Causal Inference Right: A Classification of Data Science Tasks

Miguel A. Hernán, John Hsu & Brian Healy

To cite this article: Miguel A. Hernán, John Hsu & Brian Healy (2019) A Second Chance to Get Causal Inference Right: A Classification of Data Science Tasks, CHANCE, 32:1, 42-49, DOI: 10.1080/09332480.2019.1579578

To link to this article: https://doi.org/10.1080/09332480.2019.1579578

A Second Chance to Get Causal Inference Right: A Classification of Data Science Tasks

Miguel A. Hernán, John Hsu, and Brian Healy

For much of the recent history of science, learning from data was the academic realm of statistics,[1,2] but in the early 20th century, the founders of modern statistics made a momentous decision about what could and could not be learned from data: They proclaimed that statistics could be applied to make causal inferences when using data from randomized experiments, but not when using nonexperimental (observational) data.[3,4,5] This decision classified an entire class of scientific questions in the health and social sciences as not amenable to formal quantitative inference.

Not surprisingly, many scientists ignored the statisticians' decree and continued to use observational data to study the unintended harms of medical treatments, health effects of lifestyle activities, or social impact of educational policies. Unfortunately, these scientists' causal questions often were mismatched with their statistical training. Perplexing paradoxes arose; for example, the famous "Simpson's paradox" stemmed from a failure to recognize that the choice of data analysis depends on the causal structure of the problem.[6] Mistakes occurred. For example, as a generation of medical researchers and clinicians believed that postmenopausal hormone therapy reduced the risk of heart disease because of data analyses that deviated from basic causal considerations. Even today, confusions generated by a century-old refusal to tackle causal questions explicitly are widespread in scientific research.[7]

To bridge science and data analysis, a few rogue statisticians, epidemiologists, econometricians, and computer scientists developed formal methods to quantify causal effects from observational data. Initially, each discipline emphasized different types of causal questions, developed different terminologies, and preferred different data analysis techniques. By the beginning of the 21st century, while some conceptual discrepancies remained, a unified theory of quantitative causal inference had emerged.[8,9]

We now have a historic opportunity to redefine data analysis in such a way that it naturally accommodates a science-wide framework for causal inference from observational data. A recent influx of data analysts, many not formally trained in statistical theory, bring a fresh attitude that does not a priori exclude causal questions. This new wave of data analysts refer to themselves as data scientists and to their activities as data science, a term popularized by technology companies and embraced by academic institutions.

Data science, as an umbrella term for all types of data analysis, can tear down the barriers erected by traditional statistics; put data analysis at the service of all scientific questions, including causal ones; and prevent unnecessary inferential mistakes. We may miss our chance to successfully integrate data analysis into all scientific

[1]Tukey, J.W. 1962. The future of data analysis. *Annals of Mathematical Statistics* 33:1-67.
[2]Donoho, D. 2017. 50 years of data science. *Journal of Computational and Graphical Statistics* 26(4):745–66.
[3]Pearl, J. 2009. *Causality: Models, Reasoning, and Inference* (2nd edition). New York: Cambridge University Press.
[4]Fisher, R.A. 1925. *Statistical Methods for Research Workers*, 1st ed. Edinburgh: Oliver and Boyd.
[5]Pearson, K. 1911. *The Grammar of Science*, 3rd ed. London: Adam and Charles Black.
[6]Hernán, M.A., Clayton, D., and Keiding, N. 2011. The Simpson's paradox unraveled. *International Journal of Epidemiology* 40(3):780–5.
[7]Hernán, MA. 2018. The C-word: Scientific euphemisms do not improve causal inference from observational data (with discussion). *American Journal of Public Health* 108(5): 616–9.
[8]Hernán, M.A., Robins J.M. 2018 (forthcoming). *Causal Inference*. Boca Raton: Chapman & Hall/CRC.
[9]Pearl, J. 2018. *The Book of Why*. New York: Basic Books.

questions, though, if data science ends up being defined exclusively in terms of technical[10] activities (management, processing, analysis, visualization…) without explicit consideration of the scientific tasks.

A Classification of Data Science Tasks

Data scientists often define their work as "gaining insights" or "extracting meaning" from data. These definitions are too vague to characterize the scientific uses of data science. Only by precisely classifying the "insights" and "meaning" that data can provide will we be able to think systematically about the types of data, assumptions, and analytics that are needed. The scientific contributions of data science can be organized into three classes of tasks: description, prediction, and counterfactual prediction (see table for examples of research questions for each of these tasks).

Description is using data to provide a quantitative summary of certain features of the world. Descriptive tasks include, for example, computing the proportion of individuals with diabetes in a large healthcare database and representing social networks in a community. The analytics employed for description range from elementary calculations (a mean or a proportion) to sophisticated techniques such as unsupervised learning algorithms (cluster analysis) and clever data visualizations.

Prediction is using data to map some features of the world (the inputs) to other features of the world (the outputs). Prediction often starts with simple tasks (quantifying the association between albumin levels at admission and death within one week among patients in the intensive care unit) and then progresses to more-complex ones (using hundreds of variables measured at admission to predict which patients are more likely to die within one week). The analytics employed for prediction range from elementary calculations (a correlation coefficient or a risk difference) to sophisticated pattern recognition methods and supervised learning algorithms that can be used as classifiers (random forests, neural networks) or predict the joint distribution of multiple variables.

Counterfactual prediction is using data to predict certain features of the world as if the world had been different, which is required in *causal inference* applications. An example of causal inference is the estimation of the mortality rate that would have been observed if all individuals in a study population had received screening for colorectal cancer vs. if they had not received screening.

The analytics employed for causal inference range from elementary calculations in randomized experiments with no loss to follow-up and perfect adherence (the difference in mortality rates between the screened and the unscreened) to complex implementations of g-methods in observational studies with treatment-confounder feedback (the plug-in g-formula).[11]

Note that, contrary to some computer scientists' belief, "causal inference" and "reinforcement learning" are not synonyms. Reinforcement learning is a technique that, in some simple settings, leads to sound causal inference. However, reinforcement learning is insufficient for causal inference in complex settings (discussed below).

Statistical inference is often required for all three tasks. For example, one might want to add 95% confidence intervals for descriptive, predictive, or causal estimates involving samples of target populations.

As in most attempts at classification, the boundaries between the above categories are not always sharp. However, this trichotomy provides a useful starting point to discuss the data requirements, assumptions, and analytics necessary to successfully perform each task of data science. A similar taxonomy has traditionally been taught by data scientists from many disciplines, including epidemiology, biostatistics,[12] economics,[13] and political science.[14] Some methodologists have referred to the causal inference task as "explanation,"[15] but this is a somewhat-misleading term because causal effects may be quantified while remaining unexplained (randomized trials identify causal effects even if the causal mechanisms that explain them are unknown).

Sciences are primarily defined by their questions rather than by

[10]Cleveland, W. 2001. Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics. *Intenational Statistical Review* 69(1):21-6.
[11]Robins, J.M. 1986. A new approach to causal inference in mortality studies with a sustained exposure period—Application to the healthy worker survivor effect. *Mathematical Modelling* 7:1,393–512 (1987. errata, Mathematicall Modelling 14:917–21).
[12]Vittinghoff, E., Glidden, D.V., Shiboski, S.C., and McCulloch, C.E. 2012. *Regression Methods in Biostatistics*. New York: Springer.
[13]Mullainathan, S., and Spiess, J. 2017. Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives* 31(2):87–106.
[14]Toshkov, D. 2016. *Research Design in Political Science*. London: Palgrave McMillan.
[15]Schmueli, G. 2010. To explain or to predict? *Statistical Science* 25(3):289–310.

their tools: We define astrophysics as the discipline that learns the composition of the stars, not as the discipline that uses the spectroscope. Similarly, data science is the discipline that describes, predicts, and makes causal inferences (or, more generally, counterfactual predictions), not the discipline that uses machine learning algorithms or other technical tools. Of course data science certainly benefits from the development of tools for the acquisition, storage, integration, access, and processing of data, as well as from the development of scalable and parallelizable analytics. This data engineering powers the scientific tasks of data science.

Prediction vs. Causal Inference

Data science has excelled at commercial applications, such as shopping and movie recommendations, credit rating, stock trading algorithms, and advertisement placement. Some data scientists have transferred their skills to scientific research with biomedical applications such as Google's algorithm to diagnose diabetic retinopathy[16] (after 54 ophthalmologists classified more than 120,000 images), Microsoft's algorithm to predict pancreatic cancer months before its usual diagnosis[17] (using the online search histories of 3,000 users who were later diagnosed with cancer), and Facebook's algorithm to detect users who may be suicidal[18] (based on posts and live videos).

All these applications of data science have one thing in common: They are predictive, not causal. They map inputs (an image of a human retina) to outputs (a diagnosis of retinopathy), but they do not consider how the world would look like under different courses of action (whether the diagnosis would change if we operated on the retina).

Mapping observed inputs to observed outputs is a natural candidate for automated data analysis because this task only requires: 1) a large data set with inputs and outputs, 2) an algorithm that establishes a mapping between inputs and outputs, and 3) a metric to assess the performance of the mapping, often based on a gold standard.[19] Once these three elements are in place, as in the retinopathy example, predictive tasks can be automated via data-driven analytics that evaluate and iteratively improve the mapping between inputs and outputs without human intervention.

More precisely, the component of prediction tasks that can be automated easily is the one that does not involve any expert knowledge. Prediction tasks require expert knowledge to specify the scientific question—what to input and what outputs—and to identify/generate relevant data sources.[20] (The extent of expert knowledge varies with different prediction tasks.[21]) However, no expert knowledge is required for prediction after candidate inputs and the outputs are specified and measured in the population of interest. At this point, a machine learning algorithm can take over the data analysis to deliver a mapping and quantify its performance. The resulting mapping may be opaque, as in many deep learning applications, but its ability to map the inputs to the outputs with a known accuracy in the studied population is not in question.

The role of expert knowledge is the key difference between prediction and causal inference tasks. Causal inference tasks require expert knowledge not only to specify the question (the causal effect of what treatment on what outcome) and identify/generate relevant data sources, but also to describe the causal structure of the system under study. Causal knowledge, usually in the form of unverifiable assumptions,[22,23] is necessary to guide the data analysis and to provide a justification for endowing the resulting numerical estimates with a causal interpretation. In other words, the validity of causal inferences depends on structural knowledge, which is usually incomplete, to supplement the information in the data. As a consequence, no algorithm

[16]Gulshan, V., Peng, L., Coram, M., et al. 2016. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA* 316(22):2,402–10.

[17]Paparrizos, J., White, R.W., and Horvitz, E. 2016. Screening for Pancreatic Adenocarcinoma Using Signals From Web Search Logs: Feasibility Study and Results. *Journal of Oncological Practice* 12(8):737–44.

[18]Rosen, G. 2017. Getting Our Community Help in Real Time. *https://newsroom.fb.com/news/2017/11/getting-our-community-help-in-real-time/* (accessed April 26, 2018).

[19]Brynjolfsson, E., and Mitchell, T. 2017. What can machine learning do? Workforce implications. *Science* 358(6370):1,530–4.

[20]Conway, D. 2010. The Data Science Venn Diagram. Accessed October 9, 2018. *http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram.*

[21]Beam, A.L., and Kohane I.S. 2018. Big Data and Machine Learning in Health Care. *JAMA* 319(13):1,317–8.

[22]Robins, J.M. 2001, Data, design, and background knowledge in etiologic inference. *Epidemiology* 11:313–20.

[23]Robins, J.M., and Greenland, S. 1986. The role of model selection in causal inference from nonexperimental data. *American Journal of Epidemiology* 123(3):392–402.

can quantify the accuracy of causal inferences from observational data. The following simplified example helps fix ideas about the different role of expert knowledge for prediction versus causal inference.

Example

Suppose we want to use a large health records database to predict infant mortality (the output) using clinical and lifestyle factors collected during pregnancy (the inputs). We have just applied our expert knowledge to decide what the output and candidate inputs are, and to select a particular database in the population of interest. The only requirement is that the potential inputs must precede the outputs temporally, regardless of the causal structure linking them. At this point of the process, our expert knowledge will not be needed any more: An algorithm can provide a mapping between inputs and outputs at least as good as any mapping we could propose and, in many cases, astoundingly better.

Now suppose we want to use the same health records database to determine the causal effect of maternal smoking during pregnancy on the risk of infant mortality. A key problem is confounding: Pregnant women who do and do not smoke differ in many characteristics (including alcohol consumption, diet, access to adequate prenatal care) that affect the risk of infant mortality. Therefore, a causal analysis must identify and adjust for those confounding factors which, by definition, are associated with both maternal smoking and infant mortality.

However, not all factors associated with maternal smoking and infant mortality are confounders that should be adjusted for. For example, birthweight is strongly associated with both maternal smoking and infant mortality, but adjustment for birthweight induces bias because birthweight is a risk factor that is itself causally affected by maternal smoking. In fact, adjustment for birthweight results in a bias often referred to as the "birthweight paradox": Low birthweight babies from mothers who smoked during pregnancy have a lower mortality than those from mothers who did not smoke during pregnancy.[24]

An algorithm devoid of causal expert knowledge will rely exclusively on the associations found in the data and is therefore at risk of selecting features, like birthweight, that increase bias. The "birthweight paradox" is indeed an example of how the use of automatic adjustment procedures may lead to an incorrect causal conclusion. In contrast, a human expert can readily identify many variables that, like birthweight, should not be adjusted for because of their position in the causal structure.

A human expert also may identify features that should be adjusted for, even if they are not available in the data, and propose sensitivity analyses[25] to assess the reliability of causal inferences in the absence of those features. In contrast, an algorithm that ignores the causal structure will not issue an alert about the need to adjust for features that are not in the data.

Given the central role of (potentially fallible) expert causal knowledge in causal inference, it is not surprising that researchers look for procedures to alleviate the reliance of causal inferences on causal knowledge. Randomization is the best such procedure.

When a treatment is randomly assigned, we can unbiasedly estimate the average causal effect of treatment assignment *in the absence of detailed causal knowledge about the system under study*. Randomized experiments are central in many areas of science where relatively simple causal questions are asked.[26] Randomized experiments are also commonly used, often under the name A/B testing, to answer simple causal questions in commercial web applications. However, randomized designs are often infeasible, untimely, or unethical in the extremely complex systems studied by health and social scientists.[26]

A failure to grasp the different role of expert knowledge in prediction and causal inference is a common source of confusion in data science (the confusion is compounded by the fact that predictive analytic techniques, such as regression, can also be used for causal inference when combined with causal knowledge).

Both prediction and causal inference require expert knowledge to formulate the scientific question i, but only causal inference requires causal expert knowledge to answer the question. As a result,

[24]Hernández-Díaz, S., Schisterman, E.F., and Hernán, M.A. 2006. The birth weight "paradox" uncovered? *American Journal of Epidemiology* 164(11):1,115–20.

[25]Robins, J.M., Rotnitzky, A., and Scharfstein, D.O. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In Halloran E., and Berry D., eds. *Statistical Methods in Epidemiology: The Environment and Clinical Trials*. New York: Springer Verlag; 1999:1–92.

[26]Hernán, M.A. 2015. Invited commentary: Agent-based models for causal inference-reweighting data and theory in epidemiology. *American Journal of Epidemiology* 2181(2):103–5.

the accuracy of causal estimates cannot be assessed by using metrics computed from the data, even if the data were perfectly measured in the population of interest.

Implications for Decision-making

A goal of data science is to help people make better decisions. For example, in health settings, the goal is to help decision-makers—patients, clinicians, policy-makers, public health officers, regulators—decide among several possible strategies. Frequently, the ability of data science to improve decision-making is predicated on the basis of its success at prediction.

However, the premise that predictive algorithms will lead to better decisions is questionable. An algorithm that excels at using data about patients with heart failure to predict who will die within the next five years is agnostic about how to reduce mortality. For example, a prior hospitalization may be identified as a useful predictor of mortality, but nobody would suggest that we stop hospitalizing people to reduce mortality. Identifying patients with bad prognoses is very different from identifying the best course of action for preventing or treating a disease. Worse, predictive algorithms, when incorrectly used for causal inference, may lead to incorrect confounder adjustment and therefore conclude, for example, that maternal smoking appears to be beneficial for low birthweight babies.

Predictive algorithms inform us that decisions have to be made, but they cannot help us make the decisions. For example, a predictive algorithm that identifies patients with severe heart failure does not provide information about whether heart transplant is the best treatment option. In contrast, causal analyses are designed to help us make decisions because they tackle "what if" questions. A causal analysis will, for instance, compare the benefit-risk profile of heart transplant versus medical treatment in patients with certain severity of heart failure.

Interestingly, the distinction between prediction and causal inference (counterfactual prediction) becomes unnecessary for decision-making when the relevant expert knowledge can readily be encoded and incorporated into the algorithms. A purely predictive algorithm that learns to play Go can perfectly predict the counterfactual state of the game under different moves, and a predictive algorithm that learns to drive a car can accurately predict the counterfactual state of the car if, say, the brakes are not operated.

Because these systems are governed by a set of known game rules (in the case of games like Go) or physical laws with some stochastic components (in the case of engineering applications like self-driving cars), an algorithm can eventually predict the behavior of the entire system under a hypothetical intervention.

Take the game of Go, which has been mastered by an algorithm "without human knowledge."[27] When making a move, the algorithm has access to all information that matters: game rules, current board position, and future outcomes fully determined by the sequence of moves. Further, a reinforcement learning algorithm can collect an arbitrary amount of data by playing more games (conducting numerous experiments), which allows it to learn by trial and error. In this setting, a cleverly designed algorithm running on a powerful computer can spectacularly outperform humans—but this form of causal inference has, at this time in history, a restricted domain of applicability.

Many scientists work on complex systems with partly known and nondeterministic governing laws (the "rules of the game"), with uncertainty about whether all necessary data are available, and for which learning by trial and error—or even conducting a single experiment—is impossible. Even when the laws are known and the data available, the system may still be too chaotic for exact long-term prediction. For example, it was impossible to predict when and where the Chinese space station,[28] while in orbit at an altitude of about 250 km, would fall to Earth.

Consider a causal question about the effect of different epoetin strategies on the mortality of patients with renal disease. We do not understand the causal structure by which molecular, cellular, individual, social, and environmental factors regulate the effect of epoetin dose on mortality risk. As a result, it is currently impossible to construct a predictive model based on electronic health records to reproduce the behavior of the system under a hypothetical intervention on an individual. Some widely publicized disappointments in causal applications of data science, like "Watson for Oncology,"

[27]Silver, D., Schrittwieser, J., and Simonyan, K., et al. 2017. Mastering the game of Go without human knowledge. *Nature* 550(7676):354–9.
[28]The Data Team. 2018. An out-of-control Chinese space station will soon fall to Earth. *The Economist* March 19, 2018.

Table 1—Examples of Tasks Conducted by Data Scientists Working with Electronic Health Records

	Data Science Task		
	Description	**Prediction**	**Causal inference**
Example of scientific question	How can women aged 60–80 years with stroke history be partitioned in classes defined by their characteristics?	What is the probability of having a stroke next year for women with certain characteristics?	Will starting a statin reduce, on average, the risk of stroke in women with certain characteristics?
Data	• Eligibility criteria • Features (symptoms, clinical parameters …)	• Eligibility criteria • Output (diagnosis of stroke over the next year) • Inputs (age, blood pressure, history of stroke, diabetes at baseline)	• Eligibility criteria • Outcome (diagnosis of stroke over the next year) • Treatment (initiation of statins at baseline) • Confounders • Effect modifiers (optional)
Examples of analytics	Cluster analysis …	Regression Decision trees Random forests Support vector machines Neural networks …	Regression Matching Inverse probability weighting G-formula G-estimation Instrumental variable estimation …

have arguably resulted from trying to predict a complex system that is still poorly understood and for which a sound model to combine expert causal knowledge with the available data is lacking.[29]

The striking contrast between the cautious attitude of most traditional data scientists (statisticians, epidemiologists, economists, political scientists…) and the "can do" attitude of many computer scientists, informaticians, and others seems to be, to a large extent, the consequence of the different complexity of the causal questions historically tackled by each of these groups. Epidemiologists and other data scientists working with extremely complex systems tend to focus on the relatively modest goal of designing observational analyses to answer narrow causal questions about the average causal effect of a variable (such as epoetin treatment), rather than try to explain the causal structure of the entire system or identify globally optimal decision-making strategies.

On the other hand, newcomers to data science have often focused on systems governed by known laws (like board games or self-driving cars), so it is not surprising that they have deemphasized the distinction between prediction and causal inference. Bringing this distinction to the forefront is, however, urgent as an increasing number of data scientists address the causal questions traditionally asked by health and social scientists. Sophisticated prediction algorithms may suffice to develop unbeatable Go software and, eventually, safe self-driving vehicles, but causal inferences in complex systems (say, the effects of clinical strategies to

[29]Ross, C., and Swetlitz, I. 2017. IBM pitched its Watson supercomputer as a revolution in cancer care. It's nowhere close. *STAT*. *https://www.statnews.com/2017/09/05/watson-ibm-cancer/*.

[30]Pearl, J. 2018. Theoretical Impediments to Machine Learning With Seven Sparks from the Causal Revolution. *Technical Report* R-475 (*http://ftp.cs.ucla.edu/pub/stat_ser/r475.pdf*. Accessed April 26, 2018.

treat a chronic disease) need to rely on data analysis methods equipped with causal knowledge.[30]

Processes and Implications for Teaching

The training of data scientists tends to emphasize mastering tools for data management and data analysis. While learning to use these tools will continue to play a central role, it is important that the technical training of data scientists makes it clear that the tools are at the service of distinct scientific tasks—description, prediction, and causal inference.

A training program in data science can, therefore, be organized explicitly in three components, each devoted to one of the three tasks of data science. Each component would describe how to articulate scientific questions, data requirements, threats to validity, data analysis techniques, and the role of expert knowledge (separately for description, prediction, and causal nference). This is the approach that we adopted to develop the curriculum of the Clinical Data Science core at the Harvard Medical School, which three cohorts of clinical investigators have now learned.

Our students first learn to differentiate between the three tasks of data science, then how to generate and analyze data for each task, as well as the differences between tasks. They learn that description and prediction may be affected by selection and measurement biases, but that only causal inference is affected by confounding. After learning predictive algorithms, teams of students compete against each other in a machine learning competition to develop the best predictive model (in an application of the Common Task Framework[2]).

By contrast, after learning causal inference techniques, students understand that a similar competition is not possible because their causal estimates cannot be ranked automatically. Teams with different subject-matter knowledge may produce different causal estimates, and there often is no objective way to determine which one is closest to the truth using the existing data.[31]

Then students learn to ask causal questions in terms of a contrast of interventions conducted over a fixed time period as would be specified in the protocol of a (possibly hypothetical) experiment, which is the target of inference.

For example, to compare the mortality under various epoetin dosing strategies in patients with renal failure, students use subject-matter knowledge to 1) outline the design of the hypothetical randomized experiment that would estimate the causal effect of interest—the target trial, 2) identify an observational database with sufficient information to approximately emulate the target trial, and 3) emulate the target trial and therefore estimate the causal effect of interest using the observational database. We discuss why causal questions that cannot be translated into target experiments are not sufficiently well-defined,[31] and why the accuracy of causal answers cannot be quantified using observational data. In parallel, the students also learn computer coding and the basics of statistical inference to deal with the uncertainty inherent to any data analyses involving description, prediction, or causal inference.

A data science curriculum along the three dimensions of description, prediction, and causal inference facilitates interdisciplinary integration. Learning from data requires paying attention to the different emphases, questions, and analytic methods developed over several decades in statistics, epidemiology, econometrics, computer science, and others. Data scientists without subject-matter knowledge cannot conduct causal analyses in isolation: They don't know how to articulate the questions (what the target experiment is) and they don't know how to answer them (how to emulate the target experiment).

Conclusion

Data science is a component of many sciences, including the health and social ones. Therefore, the tasks of data science are the tasks of those sciences—description, prediction, causal inference. A sometimes-overlooked point is that a successful data science requires not only good data and algorithms, but also domain knowledge (including causal knowledge) from its parent sciences.

The current rebirth of data science is an opportunity to rethink data analysis free of the historical constraints imposed by traditional statistics, which have left scientists ill-equipped to handle causal questions. While the clout of statistics in scientific training and publishing impeded the introduction of a unified formal framework for causal inference in data

[31]Hernán, MA. 2019 (in press). Spherical cows in a vacuum: Data analysis competitions for causal inference. *Statistical Science*.

analysis, the coining of the term "data science" and the recent influx of "data scientists" interested in causal analyses provides a once-in-a-generation chance of integrating all scientific questions, including causal ones, in a principled data analysis framework. An integrated data science curriculum can present a coherent conceptual framework that fosters understanding and collaboration between data analysts and domain experts.

On the other hand, if the definitions of data science currently discussed in mainstream statistics take hold, causal inference from observational data will be once more marginalized, leaving health and social scientists on their own. The American Statistical Association statement on "The Role of Statistics in Data Science" (August 8, 2015) makes no reference to causal inference. A recent assessment of data science and statistics[2] did not include the word "causal" (except when mentioning the title of the course "Experiments and Causal Inference"). Heavily influenced by statisticians, many medical editors actively suppress the term "causal" from their publications.[33]

A data science that embraces causal inference must (1) develop methods for the integration of sophisticated analytics with expert causal expertise, and (2) acknowledge that, unlike for prediction, the assessment of the validity of causal inferences cannot be exclusively data-driven because the validity of causal inferences also depends on the adequacy of expert causal knowledge. Causal directed acyclic graphs[34,35] may play an important role in the development of analytic methods that integrate learning algorithms and subject-matter knowledge. These graphs can be used to represent different sets of causal structures that are compatible with existing causal knowledge and thus to explore the impact of causal uncertainty on the effect estimates.

Large amounts of data could make expert knowledge irrelevant for prediction and for relatively simple causal inferences involving games and some engineering applications, but expert causal knowledge is necessary to formulate and answer causal questions in more-complex systems. Affirming causal inference as a legitimate scientific pursuit is the first step in transforming data science into a reliable tool to guide decision-making.

Finally, the distinction between prediction and causal inference is also crucial to defining artificial intelligence (AI). Some data scientists argue that "the essence of intelligence is the ability to predict," and therefore that good predictive algorithms are a form of AI. From this point of view, large chunks of data science can be rebranded as AI (and that is exactly what the tech industry is doing). However, mapping observed inputs to observed outputs barely qualifies as intelligence. Rather, a hallmark of intelligence is the ability to predict *counterfactually* how the world would change under different actions by integrating expert knowledge and mapping algorithms. No AI will be worthy of the name without causal inference. C

About the Authors

Miguel Hernán conducts research to learn what works for the treatment and prevention of cancer, cardiovascular disease, and HIV infection. With his collaborators, he designs analyses of healthcare databases, epidemiologic studies, and randomized trials. He teaches clinical data science at the Harvard Medical School, clinical epidemiology at the Harvard-MIT Division of Health Sciences and Technology, and causal inference methodology at the Harvard T.H. Chan School of Public Health, where he is the Kolokotrones Professor of Biostatistics and Epidemiology.

John Hsu is director of the Program for Clinical Economics and Policy Analysis in the Mongan Institute, Massachusetts General Hospital, and Harvard Medical School. He studies innovations in healthcare financing and delivery, and their effects on medical quality and efficiency. He primarily uses large automated and electronic health record data sets, often exploiting natural experiments from both clinical and behavioral economics perspectives.

Brian Healy is an assistant professor of neurology at the Harvard Medical School and an assistant professor in the Department of Biostatistics at the Harvard T.H. Chan School of Public Health. He is the primary biostatistician for the Partners MS Center at Brigham and Women's Hospital and a member of the Massachusetts General Hospital (MGH) Biostatistics Center. He teaches introductory statistics in several programs and co-directs the clinical data science sequence in the master of medical science and clinical investigation with Miguel Hernán.

[32]Ruich, P. 2017. The Use of Cause-and-Effect Language in the JAMA Network Journals. *AMA Style Insider. http://amastyleinsider.com/2017/09/19/use-cause-effect-language-jama-network-journals/*. Accessed May 25, 2018.

[33]Hernán, M.A, Hernández-Díaz, S., Werler, M.M., and Mitchell, A.A. 2002. Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. *American Journal of Epidemiology* 155:176–84.

[34]Greenland, S., Pearl, J, and Robins, J.M. Causal diagrams for epidemiologic research. *Epidemiology* 1999; 10(1):37–48.