



Analytic Tools/Tips for your Agile DevSecOps Measurement Toolkit

Robert W. Stoddard

Principal Researcher, SEI

ASQ Fellow

Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213

Document Markings

Copyright 2019 Carnegie Mellon University.

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

The view, opinions, and/or findings contained in this material are those of the author(s) and should not be construed as an official Government position, policy, or decision, unless designated by other documentation.

References herein to any specific commercial product, process, or service by trade name, trade mark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by Carnegie Mellon University or its Software Engineering Institute.

NO WARRANTY. THIS CARNEGIE MELLON UNIVERSITY AND SOFTWARE ENGINEERING INSTITUTE MATERIAL IS FURNISHED ON AN "AS-IS" BASIS. CARNEGIE MELLON UNIVERSITY MAKES NO WARRANTIES OF ANY KIND, EITHER EXPRESSED OR IMPLIED, AS TO ANY MATTER INCLUDING, BUT NOT LIMITED TO, WARRANTY OF FITNESS FOR PURPOSE OR MERCHANTABILITY, EXCLUSIVITY, OR RESULTS OBTAINED FROM USE OF THE MATERIAL. CARNEGIE MELLON UNIVERSITY DOES NOT MAKE ANY WARRANTY OF ANY KIND WITH RESPECT TO FREEDOM FROM PATENT, TRADEMARK, OR COPYRIGHT INFRINGEMENT.

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

This material may be reproduced in its entirety, without modification, and freely distributed in written or electronic form without requesting formal permission. Permission is required for any other use. Requests for permission should be directed to the Software Engineering Institute at permission@sei.cmu.edu.

Carnegie Mellon® is registered in the U.S. Patent and Trademark Office by Carnegie Mellon University.

DM19-1082

Agenda



Challenge

Selection/Survivor Bias in Data

Simpson's Paradox

Opportunity for Latent Factor Modeling

Causal Learning and Counterfactual Questions

Questions

Challenge

Agile measurement challenges:

- What questions are we trying to answer?
- Is our data biased?
- Are we aware of the deficiencies in our analytic methods?
- Are we aware of late-breaking analytic methods?
- Would we benefit from answering counterfactual questions?

Agenda

Challenge



Selection/Survivor Bias in Data

Simpson's Paradox

Opportunity for Latent Factor Modeling

Causal Learning and Counterfactual Questions

Questions

Using causal learning to deal with selection bias in torpedo data

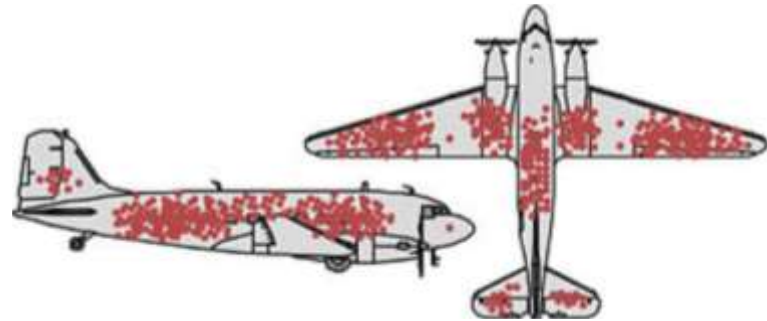
Selection bias in data can be difficult to anticipate and identify

Survivor bias is a form of selection bias in which not all failures are represented in the data to be analyzed, resulting in incorrect analysis and recommendations

According to Abraham Wald, the statisticians were looking at the planes that came back, meaning that the damage was not critical.

Wald pointed out that they should do the exact opposite of what the Navy was planning to do.

According to him, they should understand that the undamaged areas on the diagram were the reason that the aircraft was able to make it back.



Recent publications on Selection Bias (doubleclick to open)

To appear in Proceedings of AAAI-14.

TECHNICAL REPORT
14-023
April 2014

Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-14)

MLR Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS),
June 22, pp. 100-108, 2013.

TECHNICAL REPORT
14-301
April 2012

Recovering from Selection Bias in Causal and Statistical Inference

Elias Bareinboim

Cognitive Systems Laboratory
Computer Science Department
University of California, Los Angeles
Los Angeles, CA, 90095
eb@cs.ucla.edu

Jin Tian

Department of Computer Science
Iowa State University
Ames, IA, 50011
jtian@iastate.edu

Judea Pearl

Cognitive Systems Laboratory
Computer Science Department
University of California, Los Angeles
Los Angeles, CA, 90095
judea@cs.ucla.edu

Abstract

Selection bias is caused by preferential exclusion of units from the sampler and represents a major obstacle to valid causal and statistical inferences; it cannot be removed by randomized experiments and can rarely be detected in either experimental or observational studies. In this paper, we provide complete graphical and algorithmic conditions for recovering conditional probabilities from selection biased data. We also provide graphical conditions for recoverability when selection data is available over a subset of the variables. Finally, we provide a graphical condition that generalizes the backdoor criterion and serves to correct causal effects when the data is collected under post-treatment selection.

Introduction

Selection bias is induced by preferential selection of units for data analysis, usually governed by unknown factors including treatment, outcome, and their common-cause, and represents a major obstacle to valid causal and statistical inferences. It cannot be removed by randomized experiments and can rarely be detected in either experimental or observational studies. For instance, in a typical study of the effect of training programs on earnings, subjects achieving higher income tend to report their earnings more frequently than those who earn less. The data-gathering process in this case will reflect this distortion in the sample proportions and, since the sample is no longer a faithful representation of the population, biased estimates will be produced regardless of how many samples were collected.

This preferential selection challenges the validity of inferences in several tasks in AI (Geyer 1995; Hahn 2001; Zidek 2004; Corrao et al. 2005) and statistics (Wallerstein 1978; Little and Rubin 1988; Dawid 1990; Kuroki and Cai 2006) as well as in the empirical sciences (e.g., Genovese (Pharm., Donnelly, and Spencer 2012; Murtugudde and Wain 2012), Economics (Rosen 1970; Angrist 1997), and Epidemiology (Robins 2004; Glymour and Greenland 2008)).

To illuminate the nature of preferential selection, consider

Copyright © 2014, Association for the Advancement of Artificial Intelligence. All rights reserved.

*Remarkably, there are experiments in which selection bias can be detected even from observations, as in the tests of a new child-selected component (Zhang 2008).

the data-generating model in Fig. 1(a) in which X represents an action, Y represents an outcome, and Z represents a binary indicator of entry into the data pool ($Z = 1$ means that the unit is in the sample, $Z = 0$ otherwise). If our goal is to compute the population-level conditional distribution $P(Y|X)$, and the samples available are collected under selection, only $P(Y|X, Z = 1)$ is accessible for us.¹ Given that to pinpoint these two distributions are just loosely connected, the natural question to ask is under what conditions $P(Y|X)$ can be recovered from data coming from $P(Y|X, Z = 1)$. In this specific example, both action and outcome affect the entry in the data pool, which will be shown not to be recoverable (see Corollary 1).² In fact, there is no method capable of accurately estimating the population-level distribution using data gathered under this selection process.

The bias arising from selection differs fundamentally from the one due to confounding, though both constitute threats to the validity of causal inferences. The former bias is due to treatment or outcome (or common-cause) affecting the inclusion of the subject in the sample (Fig. 1(a)), while the latter is the result of treatment X and outcome Y being affected by a common causal mechanism U (Fig. 1(b)). In both cases, we have unblocked common-cause³ information between treatment and outcome, which appear under the rubric of “spurious correlation,” since it is not what we seek to estimate.

It is instructive to understand selection graphically, as in Fig. 1(a). The preferential selection that is encoded through conditioning on Z causes spurious association between X and Y through two mechanisms. First, given that Z is a collider, conditioning on it induces spurious association between its parents, X and Y (Pearl 1988). Second, Z is also a descendant of a “virtual collider” V , whose parents are X and the secret node U (also called “hidden variable”) which always appears, though often not shown in the graph.⁴

Related work and Our contributions

There are three sets of assumptions that are enlightening to researchers if we want to understand the problems arising

¹In a typical AI task such as classification, we could have X being a collection of features and Y the class to be predicted, and Z could be the class of units that are available for training.

²See (Pearl 2000, pp. 139-141) and (Pearl 2011) for further explanation of this line of thinking.

Abstract

Confounding for selection and confounding biases are two of the most challenging problems in the empirical sciences as well as in artificial intelligence tasks. Common adjustment (e.g., backdoor Adjustment) is the most pervasive technique used for controlling confounding bias, but the same is difficult to reuse for selection bias. In this paper, we summarize a generalized version of common adjustment that simultaneously controls for both confounding and selection biases. We first derive a sufficient and necessary condition for recovering causal effects using common adjustment from observational data collected under post-treatment selection. We then start by trying to consider cases when additional, unobserved mechanisms exist and contain an available form such as e.g., the age and gender distribution obtained from census data. Finally, we present a complete algorithm with polynomial delay to find all sets of admissible covariates for adjustment when confounding and selection biases are simultaneously present and adjusted data is available.

Introduction

One of the central challenges in data-driven fields is to compute the effect of interventions – for instance, how increasing the education level will affect violence rates in a city, whether issuing patents with a certain drug will help their recovery, or how increasing the product price will change monthly sales? These questions are commonly reformulated as the problem of identification of causal effects. There are two types of treatment bias that pose obstacles in this kind of inference, namely confounding bias and selection bias. The former refers to the presence of a set of factors that affect both the action (also known as treatment) and the outcome (Pearl 1993), while the latter arises when the action, outcome, or other factors differentially affect the inclusion of subjects in the data sample (Bareinboim and Pearl 2014).

The goal of our study is to produce an adjusted estimate of the causal effect, typically, the probability distribution of the outcome when an action is performed by an autonomous agent (e.g., FDA, editor, regardless of how the decision would naturally occur (Pearl 2000, Ch. 1)). For simplicity, consider the graph in Fig. 1(a) in which X represents

Copyright © 2014, Association for the Advancement of Artificial Intelligence. All rights reserved.

Causal Effect Identification by Adjustment under Confounding and Selection Biases

Junji H. Corrao

Public Health Institute
corrao@public.ia.edu

Elias Bareinboim

Public Health Institute
eb@public.ia.edu

a treatment (e.g., taking or not a drug), Y represents an outcome (health status), and Z is a factor (e.g., gender, age) that affects both the propensity of being treated and the outcome. The edges (Z, X) and (Z, Y) may encode the facts “gender effects how the drug is being prescribed” and “gender affects health recovery” respectively – for example, females may be more health conscious, so they seek for treatment more frequently than their male counterparts and at the same time are less likely to develop large complications for the particular disease. Intuitively, the causal effect represents the variation of X that brings about change to Y regardless of the influence of Z on Y , which is graphically represented in Fig. 1(b). Multifaceted is the graphical operation of removing arrows representing a decision made by an autonomous agent of setting a variable to a certain value. The mathematical counterpart of selection is the do() operator and the average causal effect of X on Y is usually written in terms of the do-distribution $P(Y|do(X))$ (Pearl 2000, Ch. 1).

The gold standard for obtaining the do-distribution is through the use of randomization, where the treatment assignment is selected by a randomized device (e.g., coin flip) regardless of any other set of covariates (Z). In fact, this operation physically transforms the study of the underlying population (Fig. 1(a)) into the corresponding standard world (Fig. 1(b)). The effect of Z on X is neutralized once randomization is applied. Despite its effectiveness, randomization seldom can be practically expensive, and even possible to create crises, either for technical, ethical, or technical reasons – e.g., we cannot randomize the cholesterol level of a patient and record if it causes the heart to stop, one may try to assess the effect of cholesterol level on cardiac health.

An alternative way of computing causal effects is trying to relate non-experimentally collected samples drawn from $P(X, Y)$ with the experimental distribution $P(Y|do(X))$. Non-experimental (either called observational) data relate to the model in Fig. 1(a) where subjects decide by themselves to take or not the drug (X) while influenced by other factors (Z). There is a number of techniques developed for this task, when the most general one to know is do-calculus (Pearl 1995). In particular, one particular strategy from do-calculus called adjustment is used the most. It consists of removing the effect of X on Y over the different levels of Z , resulting

Controlling Selection Bias in Causal Inference

Elias Bareinboim

Cognitive Systems Laboratory
Department of Computer Science
University of California, Los Angeles
Los Angeles, CA, 90095
eb@cs.ucla.edu

Judea Pearl

Cognitive Systems Laboratory
Department of Computer Science
University of California, Los Angeles
Los Angeles, CA, 90095
judea@cs.ucla.edu

Abstract

Selection bias, caused by preferential exclusion of units from the data, is a major obstacle to valid causal and statistical inferences; it cannot be removed by randomized experiments and can hardly be detected in either experimental or observational studies. This paper highlights several graphical and algebraic methods capable of mitigating and correcting selection bias. These non-parametric methods generalize previously reported results, and identify the type of knowledge that is needed for recovering in the case of selection bias. Specifically, we derive a general condition together with a procedure for checking recoverability of the stable units (OU) from n -biased data. We show that recoverability is feasible if and only if our condition holds. We further offer a new method of controlling selection bias using instrumental variables that permits the recovery of other effect measures besides OI.

1 Introduction

Selection bias is induced by preferential selection of units for data analysis, usually governed by unknown factors including treatment, outcome and their common-cause. Case-control studies in epidemiology are particularly susceptible to such bias, e.g., since only the reported only when the outcome (disease or complication) is observed, while non-cases remain unreported (Glymour and Greenland 2008; Rubin et al., 2008; Rubin, 2001; Rosen et al., 2004).

Appearing in Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS) 2012, La Palma, Canary Islands, Vol. 22 of MLCP, 10:100-108, 2012. Copyright 2012 by the authors.

To illuminate the nature of this bias, consider the model of Fig. 1 (a) in which X is a variable affected by both X (treatment) and Y (outcome), influencing entry into the data pool. Such preferential selection to the good accounts to conditioning on Z , which causes spurious association between X and Y . Through two mechanisms. First conditioning on Z induces spurious association between its parents, X and Y . Second, Z is also a descendant of a “virtual collider” V , whose parents are X and the secret node U (also called “hidden factor”) which is always present, though often not shown in the diagram.⁵

A useful example of selection bias was reported in (Oberweis and Finkenauer, 2008), and subsequently studied in (Hornik et al., 2008; Genovese et al., 2009), in which it was argued that the effect of Omeprazole (X) on Endothelial Chaperon (Y) was overestimated in the data studied. One of the symptoms of the case of Omeprazole is spurious bleeding (Fig. 1(a)), and the hypothesis was that doctors noticing bleeding are more likely to visit their doctors, causing wrong rates among Omeprazole to be overrepresented in the study.

In causal inference studies, the two most common sources of bias are confounding (Fig. 1(b)) and selection (Fig. 1(a)). The former is a result of treatment X and outcome Y being affected by a common unobserved variable U , while the latter is due to treatment or outcome (or its descendants) affecting the inclusion of the subject in the sample (induced by Z). In both cases, we have unblocked common-cause⁶ of influence between treatment and outcome, which appear under the rubric of “spurious correlation.” It is called spurious because it is not part of what we seek to estimate – the causal effect of X on Y in the target population – the cause of confounding, bias occurs because we cannot condition on the unobserved common-cause, while in selection, the distribution is always conditioned on Z .

⁵See (Pearl, 2008, pp. 100-101) for further explanation of this line of thinking.

Agenda

Challenge

Selection/Survivor Bias in Data



Simpson's Paradox

Opportunity for Latent Factor Modeling

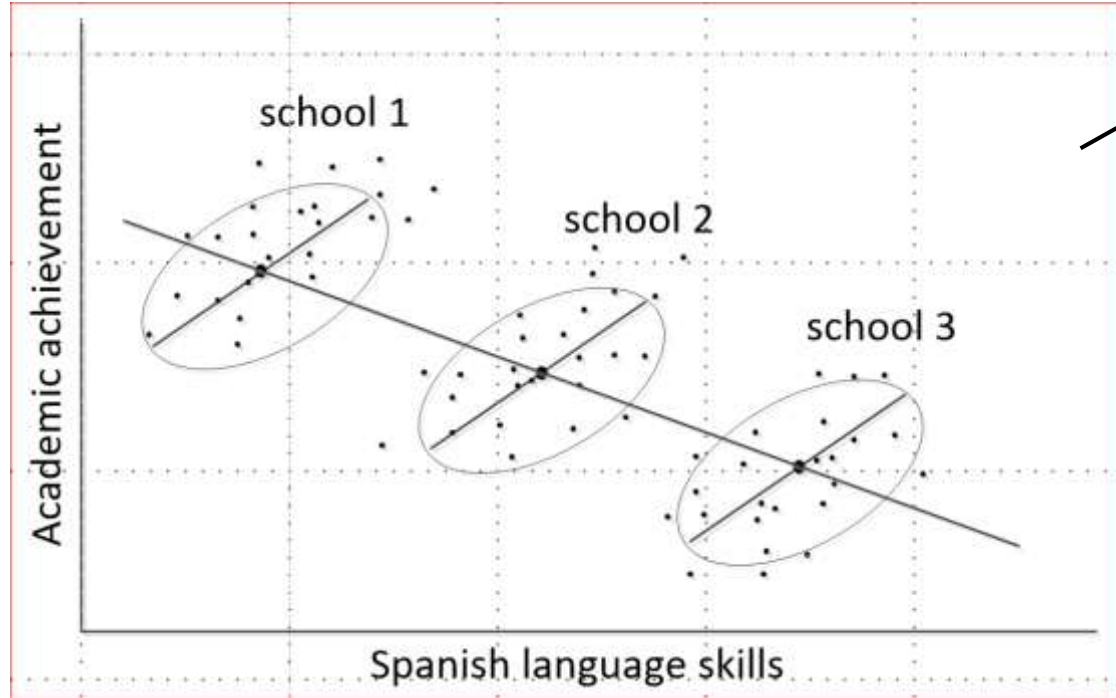
Causal Learning and Counterfactual Questions

Questions

Motivation to Look at Multi-Level SEM Models (MSEM)

Within schools, students with better Spanish skills had higher academic achievement.

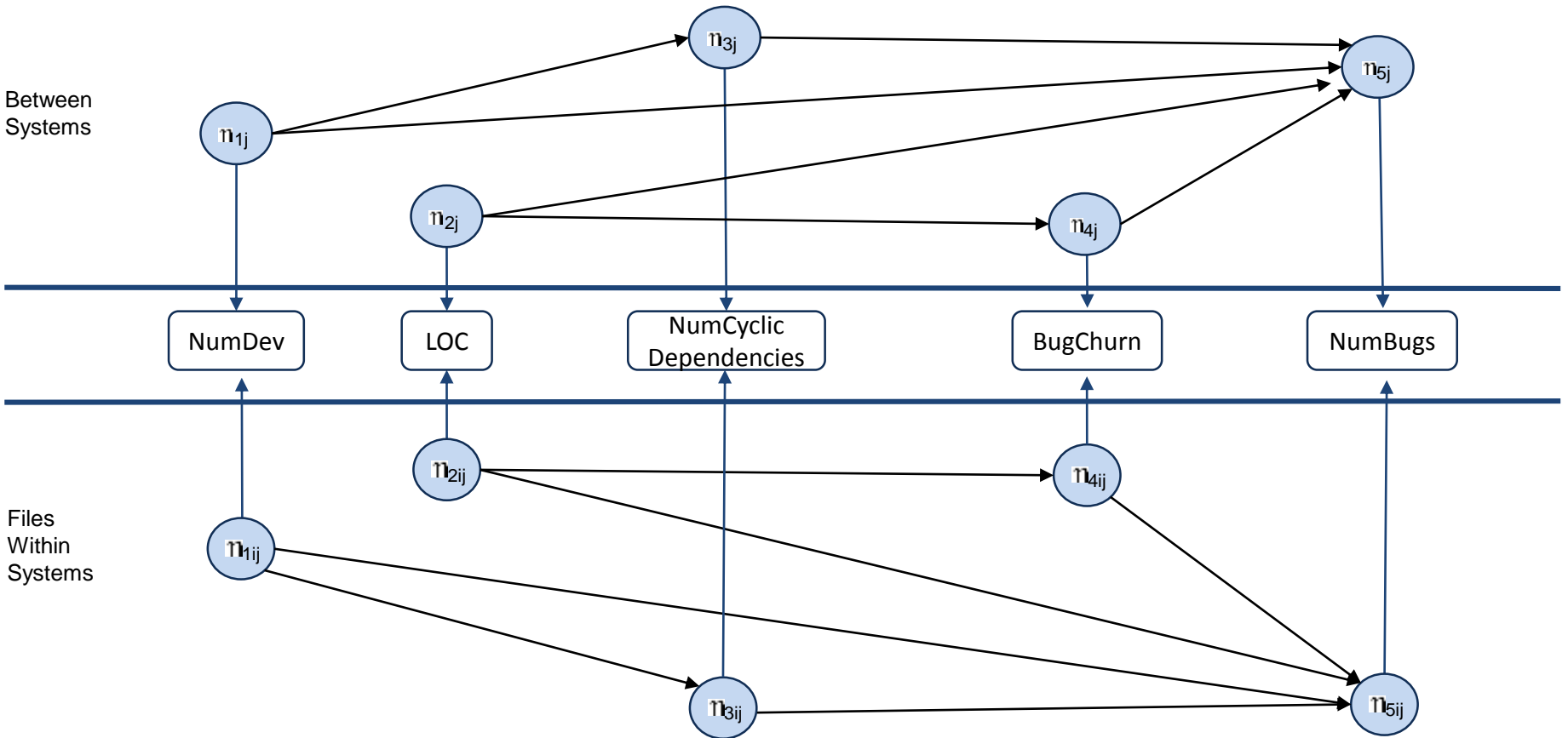
Yet, schools with highest proportion of Spanish speakers performed poorest.



Also called
Simpson's
Paradox
and the
Ecological
Fallacy

Kris Preacher, 2018

Mplus Multi-Level Structural Equation Model-01



Mplus MSEM Results

SUMMARY OF DATA

Number of clusters 9

Average cluster size 1005.556

Estimated Intraclass Correlations for the Y Variables

| Variable | Intraclass Correlation | Variable | Intraclass Correlation | Variable | Intraclass Correlation |
|----------|------------------------|----------|------------------------|----------|------------------------|
| NUMBUGS | 0.052 | NUMDEV | 0.084 | LOC | 0.008 |
| CYCLES | 0.039 | BUGCHURN | 0.026 | | |

Takeaways for MSEM and Simpson's Paradox

1. We use MSEM modeling to be sensitive to the “between” and “within” variation components of all the factors
2. We want to guard against Simpson's paradox
3. We use the Mplus MSEM analysis, specifically the Intraclass Correlation measure, to assess whether we need to perform MSEM with two levels
4. Traditional regression would have been ignorant of the above

Agenda

Challenge

Selection/Survivor Bias in Data

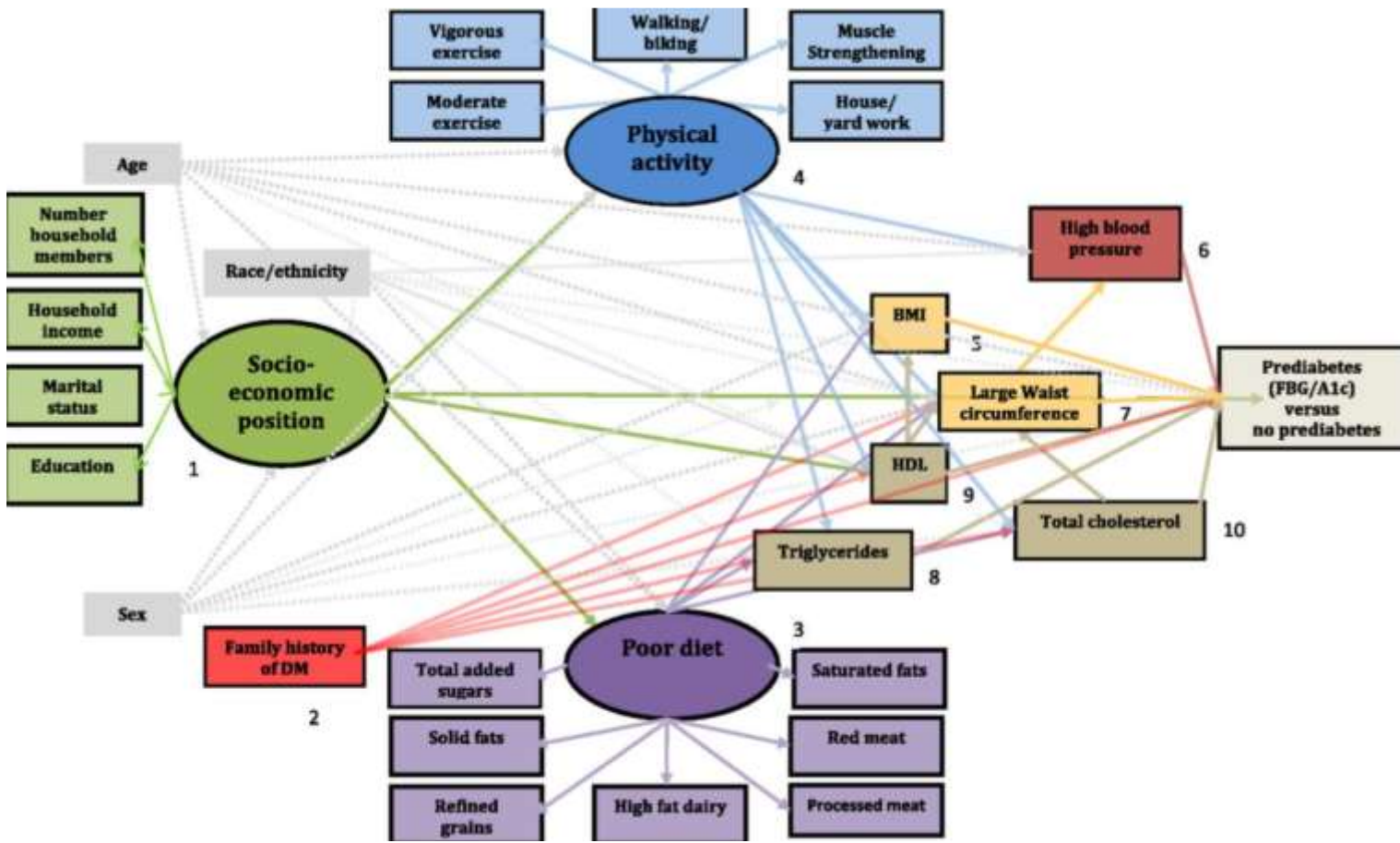
Simpson's Paradox



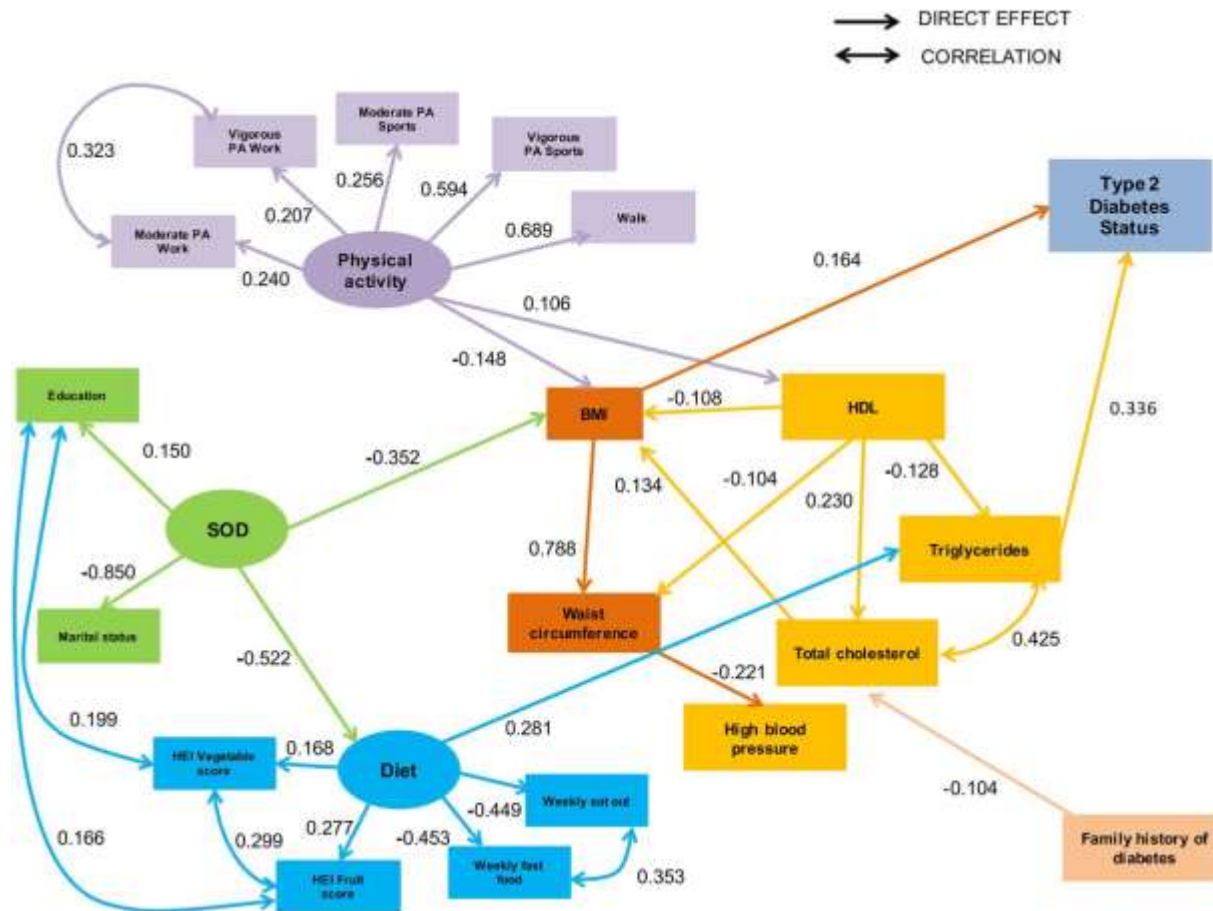
Opportunity for Latent Factor Modeling

Causal Learning and Counterfactual Questions

Questions



<https://www.bing.com/images/search?q=Sem+Model+for+Diabetes&form=IRMHRS&first=1&cw=1129&ch=557>



Andres Roman-Urrestarazu et al. *BMJ Open Diab Res Care* 2016;4:e000231

Agenda

Challenge

Selection/Survivor Bias in Data

Simpson's Paradox

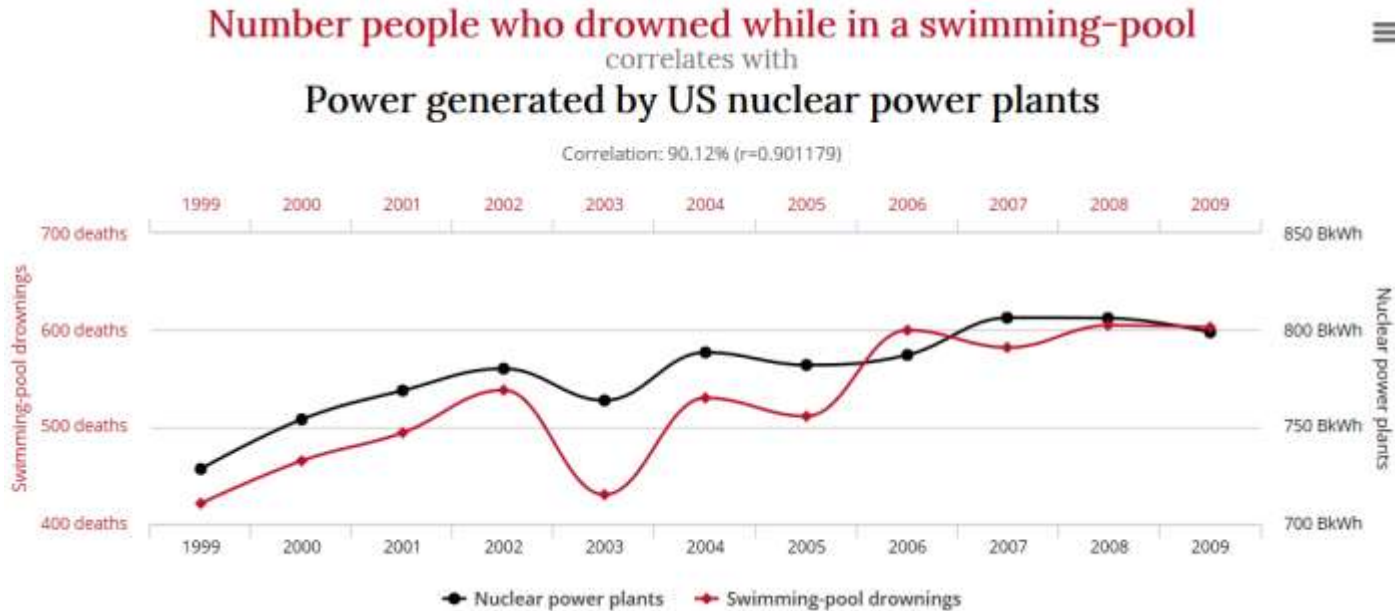
Opportunity for Latent Factor Modeling



Causal Learning and Counterfactual Questions

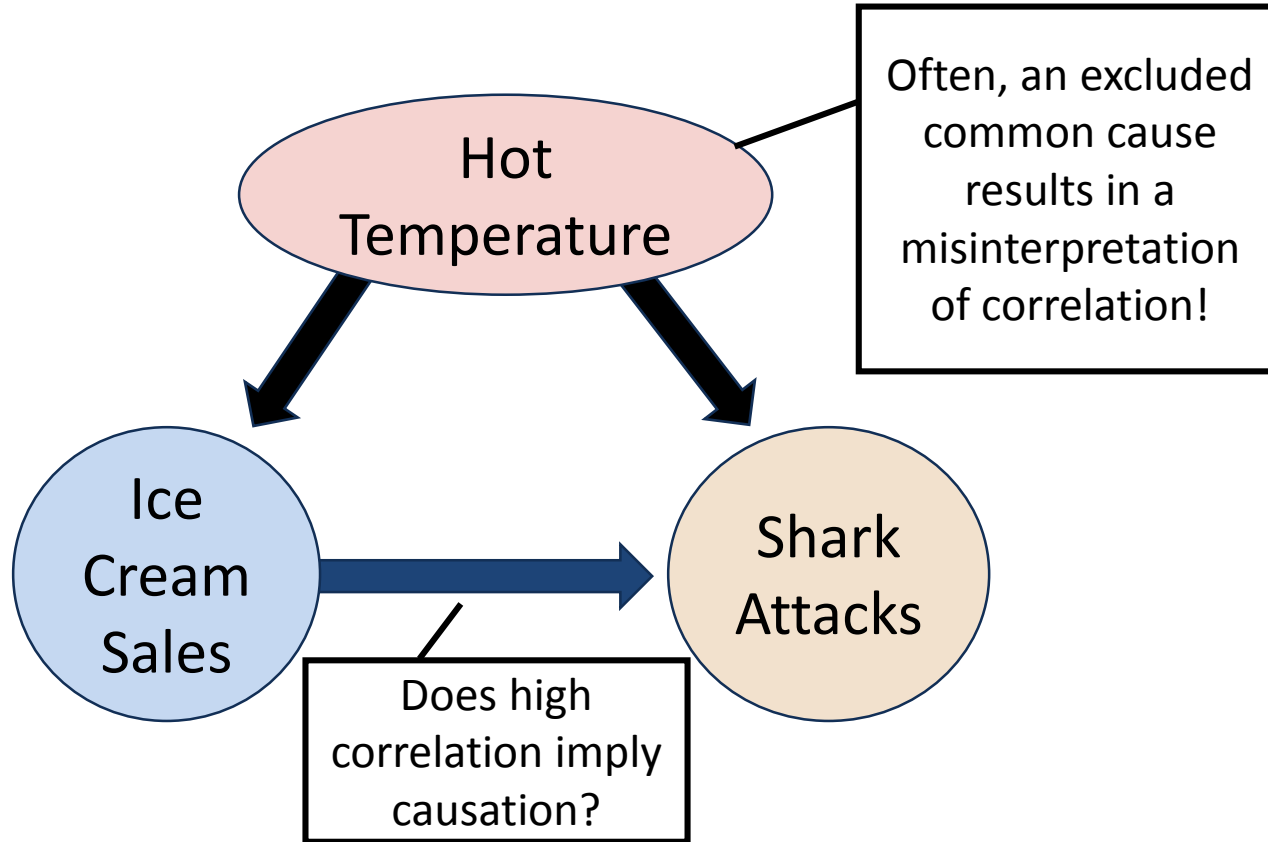
Questions

Why Do We Care about Causation?



<http://www.tylervigen.com/spurious-correlations>

More about Misinterpreting Correlation!



Regression must be interpreted in context of a DAG!

Correlation, hence regression, may be fooled by spurious association!

Before jumping into regression, we need a Directed Acyclic Graph (DAG) representing our context

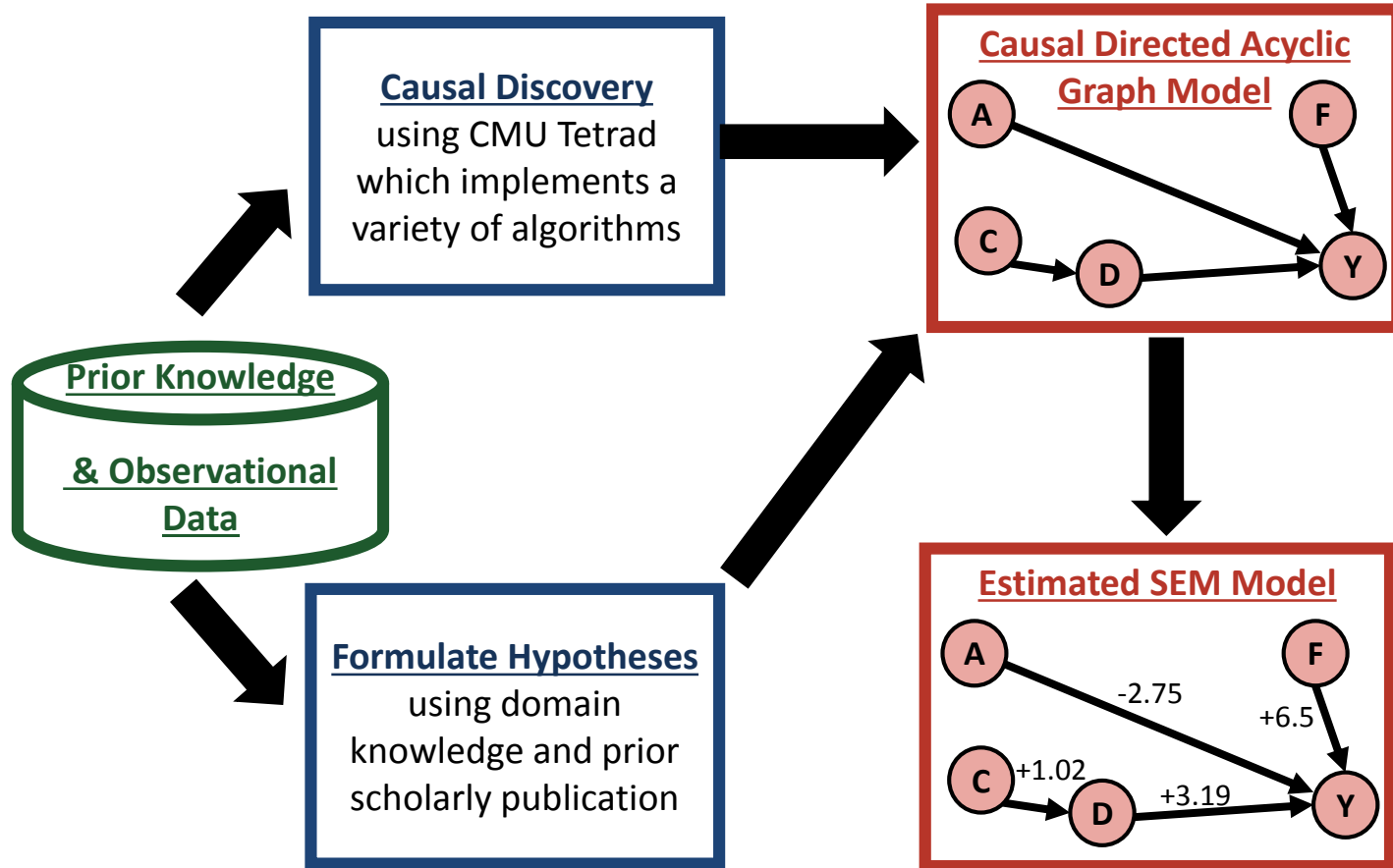
We then need to determine which paths are causal and which are spurious.

We then must block spurious correlation paths.

Lastly, we then conduct regression with the correct set of factors!

***Remember, context of the DAG
determines the suitability of the regression model!***

The Causal Learning Landscape



Use Causal Learning to Answer Counterfactual Questions

The most robust method to answer causal and counterfactual questions from a causal inference standpoint is a causal algebra called “Do-Calculus”

From Wikipedia: “Counterfactual history, also sometimes referred to as virtual history, is a form of historiography that attempts to answer "what if" questions known as counterfactuals. Black and MacRaild provide this definition: "It is, at the very root, the idea of conjecturing on what did not happen, or what might have happened, in order to understand what did happen.””

Example Counterfactual Questions

1. What would the project outcome be if agile practice xyz was not used?
2. What would the quality level be if test type xyz was not performed?
3. What would the delivery date be if a different number of sprints were employed?
4. What would the project outcome be if the customer interaction and feedback were doubled in intensity?
5. What would the project outcome be if the software teams were co-located rather than geographically separated?

Recent publications on Counterfactual Reasoning (doubleclick to open)

Forthcoming section in The handbook of Rationality, MIT press.

TECHNICAL REPORT
R-402
October 2018

Causal and Counterfactual Inference

Judea Pearl
University of California, Los Angeles
Computer Science Department
Los Angeles, CA, 90095-1596, USA
judea@cs.ucla.edu

October 2, 2018

Abstract

All accounts of rationally praiseworthy knowledge of how actions affect the state of the world and how the world would change had alternative actions been taken. The paper presents a framework called Structural Causal Model (SCM) which operationalizes this knowledge and explains how it can be derived from both theories and data. In particular, we show how counterfactuals are computed and how they can be embedded in a calculus that solves critical problems in the empirical sciences.

1 Introduction - Actions, Physical, and Metaphysical

If the options available to an agent are specified in terms of their immediate consequences, as in “make him laugh,” “paint the wall red,” “raise taxes” or, in general, $do(X=x)$, then a rational agent is instructed to maximize the expected utility

$$EU(x) = \sum_y P_x(y)U(y) \quad (1)$$

over all options x . Here, $U(y)$ stands for the utility of outcome $Y=y$ and $P_x(y)$ - the focus of this paper - stands for the (subjective) probability that outcome $Y=y$ would prevail, had action $do(X=x)$ been performed so as to establish condition $X=x$.

It has long been recognized that Bayesian conditionalization, i.e., $P_x(y) = P(y|x)$, is inappropriate for serving in Eq. (1), for it leads to paradoxical results of several kinds (see (Skyrms, 1980; Pearl, 2000a, pp. 108-9)). For example, patients would avoid going to the doctor to reduce the probability that one is seriously ill; barometers would be manipulated to reduce the chance of storms; doctors would recommend a drug to male and female patients, but not to patients with undiagnosed gender, and so on. Yet the question of what function should substitute for $P_x(y)$, despite decades of thoughtful debates (Jeffrey, 1965; Harper et al., 1981; Cartwright, 1983) seems to still baffle philosophers in the 21st century (Alic-Costa, 2007; Weirich, 2008; Woodward, 2003).

1

Causal inference and the data-fusion problem

Elan Berman^{1,2,3} and Judea Pearl¹

¹Department of Computer Science, University of California, Los Angeles, CA 90095, and ²Department of Computer Science, Purdue University, West Lafayette, IN 47907

Received by Richard M. Jeffrey, Indiana University, Bloomington, IN, and approved March 15, 2018 (received for review June 26, 2018)

We review concepts, principles, and tools that unify current approaches to multiple causes and aimed to raise challenges generated by big data. In particular, we address the problem of data fusion—joining multiple datasets collected under heterogeneous conditions (i.e., different populations, regimes, and sampling methods) to obtain valid answers to queries of interest. The availability of multiple heterogeneous datasets presents new opportunities to big data analysis, because the knowledge that can be acquired from combined data would not be possible from any individual source alone. However, the biases that emerge in heterogeneous evidence events require new analytical tools. Some of these biases, including confounding, sampling selection, and non-population biases, have been addressed by solutions largely restricted to pairwise models. We have present a general, nonparametric framework for handling these biases and, ultimately, a theoretical solution to the problem of data fusion in causal inference tasks.

causal inference | counterfactuals | causal stability | selection bias | transportation

The exponential growth of electronically accessible information has led some to conjecture that data alone can replace subjective knowledge in the process of making and testing scientific hypotheses. In this paper, we argue that traditional scientific methodologies that have been successful in the natural and biomedical sciences would still be necessary for big data applications, albeit faced with new challenges: to go beyond prediction and, using information from multiple sources, provide users with reasoned recommendations for actions and policies. The feasibility of meeting these challenges is demonstrated here using specific data fusion tasks, followed by formal derivations in the structural causal model (SCM) framework (3-5).

Introduction—Causal Inference and Big Data

The SCM framework invoked in the paper constitutes a synthesis between the counterfactual (or potential outcomes) framework of Dawid, Rubin, and Holland with the econometric tradition of Heckman, Imbens, and Hoxby (1). In this synthesis, counterfactuals are viewed as properties of structural equations and new to formally articulate such questions of interest. Graphical models on the other hand, are used to encode scientific assumptions in a qualitative (i.e., nonparametric) and transparent language as well as to do the logical formalization of these assumptions, in particular, their testable implications and how they shape behavior under intervention.

One unique feature of the SCM framework, essential in big data applications, is the ability to encode mathematically the method by which data are acquired, often referred to generically as the “design.” This usability to design, which can be used procedurally as “one set of data are created equal,” is illustrated schematically through a series of scenarios depicted in Fig. 1. Each design shown in Fig. 1, however, represents a valid sampling of the population, the target (counterfactual) set experimentally, and the sampling method by which each dataset is generated. This formal encoding allows us to distinguish the information that may be drawn from each design to answer the query of interest (Fig. 1, Top).

Consider the task of predicting the distribution of outcomes Y after intervening on a variable X , written $Q = P(Y=y|do(X=x))$.

Assume that the information available to us comes from an observational study, in which Z, X , and Y are measured, and samples are selected at random. We ask the conditional under which the query Q can be inferred from the information available, which takes the form $P(Y=y, z)$, where Z and Y are sets of observed covariates. This represents the standard task of policy evaluation, where controlling for confounding bias is the major issue (Fig. 1, task 1).

Consider now Fig. 1, task 2, in which the goal is again to estimate the effect of the intervention $do(X=x)$ but the data available to the investigator were collected in an experimental study in which variable Z were accessible to manipulation but X , a confounded (instrumental variable Z), an special case of the task 1. The general question in this scenario is under what conditions can randomization of variable Z be used to infer how the population would react to interventions over X . Formally, our problem is to infer $P(Y=y|do(X=x))$ from $P(Y=y, z|do(Z=z))$. A nonparametric solution to these two problems is presented in Policy Evaluation and the Problem of Confounding.

In each of the two previous tasks we assumed that a perfect random sample from the underlying population was drawn, which may not always be realistic. Task 3 in Fig. 1 represents a randomized dataset that was collected on a nonrepresentative sample of the population. Here, the information available takes the specific form $P(Y=y, z|do(X=x), do(Z=z))$, where Z is a sample selection indicator, with $do(Z=z)$ indicating inclusion in the sample. This design appears to estimate the effect of interest from the imperfect sampling conditions. Formally, we ask when the target quantity $P(Y=y|do(X=x))$ is derivable from the available information (i.e., using causal-based distributions). Sample Definition thus presents a solution to this problem.

Finally, the previous examples assumed that the population from which data were collected is the same as the one for which inferences was intended. This is often not the case (Fig. 1, task 4). For example, biological experiments often use animals as substitutes for human subjects. Or, in a less obvious example, data may be available from an experimental study that took place several years ago, and the current population has changed in a set S of possibly unmeasured variables. Our task here is to infer the causal effect at the target population, $P(Y=y|do(X=x), S=s)$ from the information available, which now takes the form $P(Y=y, z|do(X=x), do(Z=z), do(S=s))$. The second equation represents information obtained from nonexperimental studies on the current population, where $S=s$.

The problem represented in these two last examples can be seen as confounding bias (Fig. 1, task 5), and sample-selection bias (Fig. 1, task 6), and transportation bias (Fig. 1, task 6). The

This paper results from the author’s Institute Collaboration of the National Academy of Sciences’ Program on Causal Inference from Big Data” (June 2016–March 2018), which is a blend of traditional and modern approaches to causal inference. The complete program and other information on our program are available on our website at www.nas.edu/igbig.

Author contributions: E.B. and J.P. conceived the paper.

The authors declare no conflict of interest.

This article is a U.S. Government work.

No other competing interests should be declared. © 2018 copyright holder.

<https://doi.org/10.1101/393713>

PNAS | July 4, 2018 | vol. 115 | no. 27 | 2845-2850

Agenda

Challenge

Selection/Survivor Bias in Data

Simpson's Paradox

Opportunity for Latent Factor Modeling

Causal Learning and Counterfactual Questions



Questions

Contact Information

Presenter / Point(s) of Contact



Robert Stoddard

Email:

rws@cmu.edu

Telephone:

+1 412.268.1121

Other SEI Causal Research Team Members

Mike Konrad

Chris Miller

Bill Nichols

Dave Zubrow

Other CMU Causal Research Contributors

David Danks

Madelyn Glymour

Joe Ramsey

Kun Zhang

USC Causal Research Contributors

Jim Alstad

Barry Boehm

Anandi Hira