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

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.

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
April 2014

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

Jin Tian
Department of Computer Science
Iowa State University
Ames, IA. 50011

Judea Pearl
Cognitive Systems Laboratory
Computer Science Department
University of California, Los Angeles
Los Angeles, CA. 90095

Abstract

Introduction

Related work and Our contributions


Causal Effect Identification by Adjustment under Confounding and Selection Biases

Juan D. Correa
Purdue University

Elias Bareinboim
Purdue University

Abstract

Introduction


TECHNICAL REPORT
April 2012

Controlling Selection Bias in Causal Inference

Elias Bareinboim
Cognitive Systems Laboratory
Department of Computer Science
University of California, Los Angeles
Los Angeles, CA. 90095

Judea Pearl
Cognitive Systems Laboratory
Department of Computer Science
University of California, Los Angeles
Los Angeles, CA. 90095

Abstract

1 Introduction

Agenda

Challenge

Selection/Survivor Bias in Data

Simpson’s Paradox

Opportunity for Latent Factor Modeling

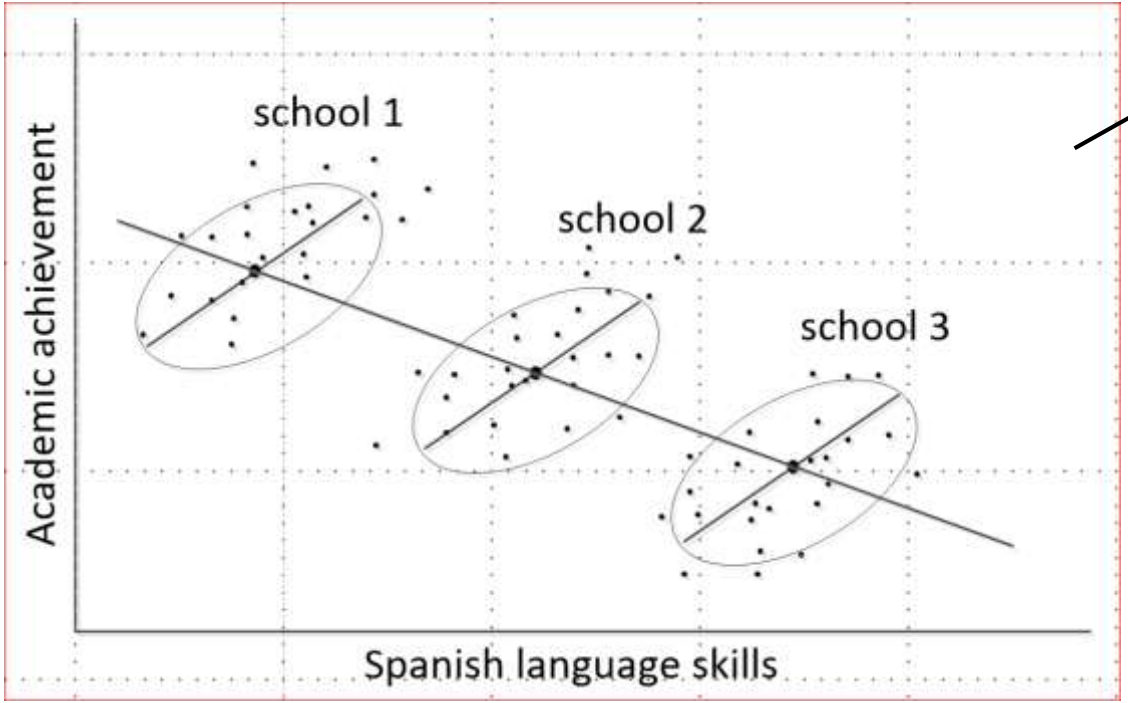
Causal Learning and Counterfactual Questions

Questions

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

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

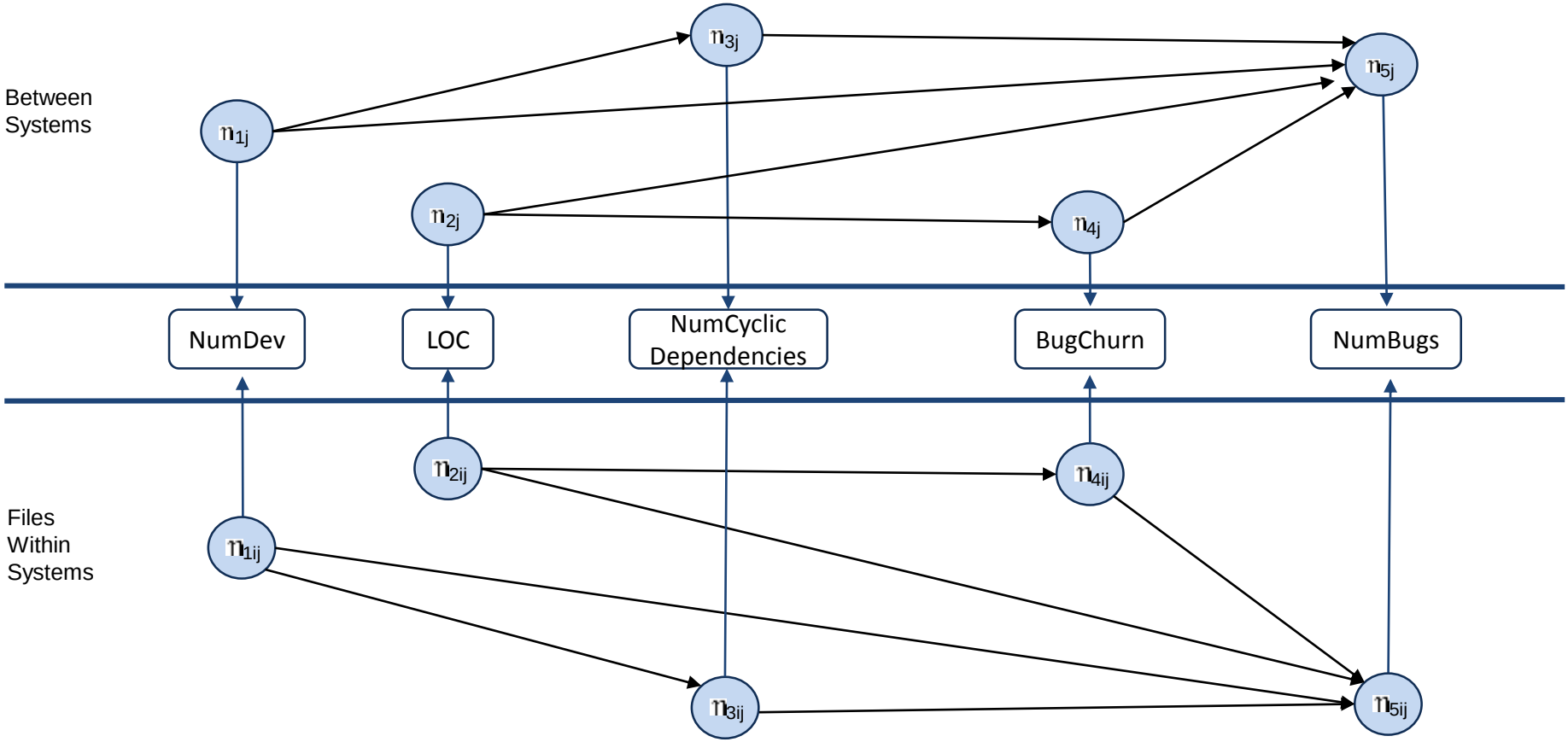
Yet, schools with highest proportion of Spanish speakers performed poorest.



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

                Intraclass                  Intraclass                  Intraclass
     Variable   Correlation   Variable      Correlation   Variable      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

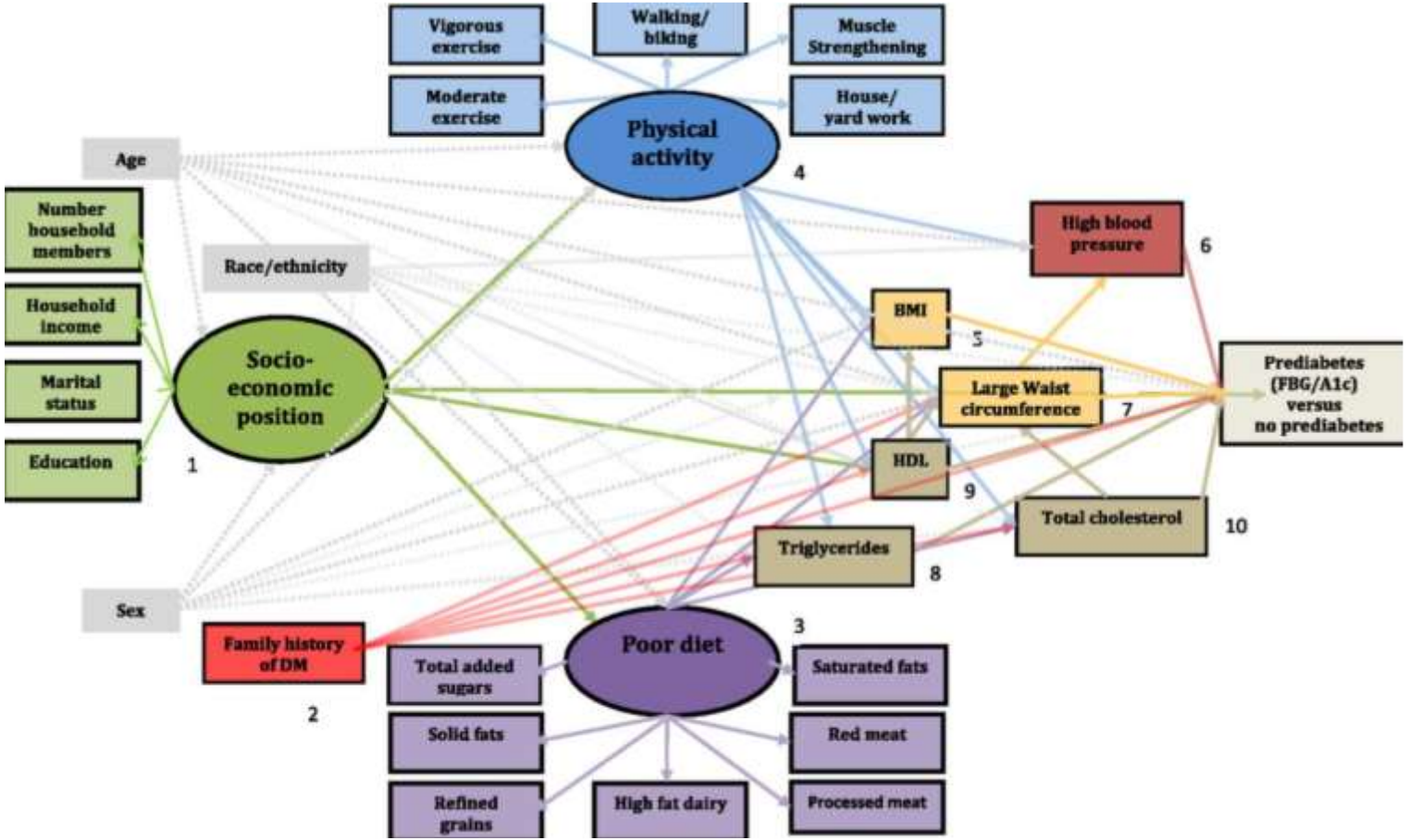
Selection/Survivor Bias in Data

Simpson's Paradox

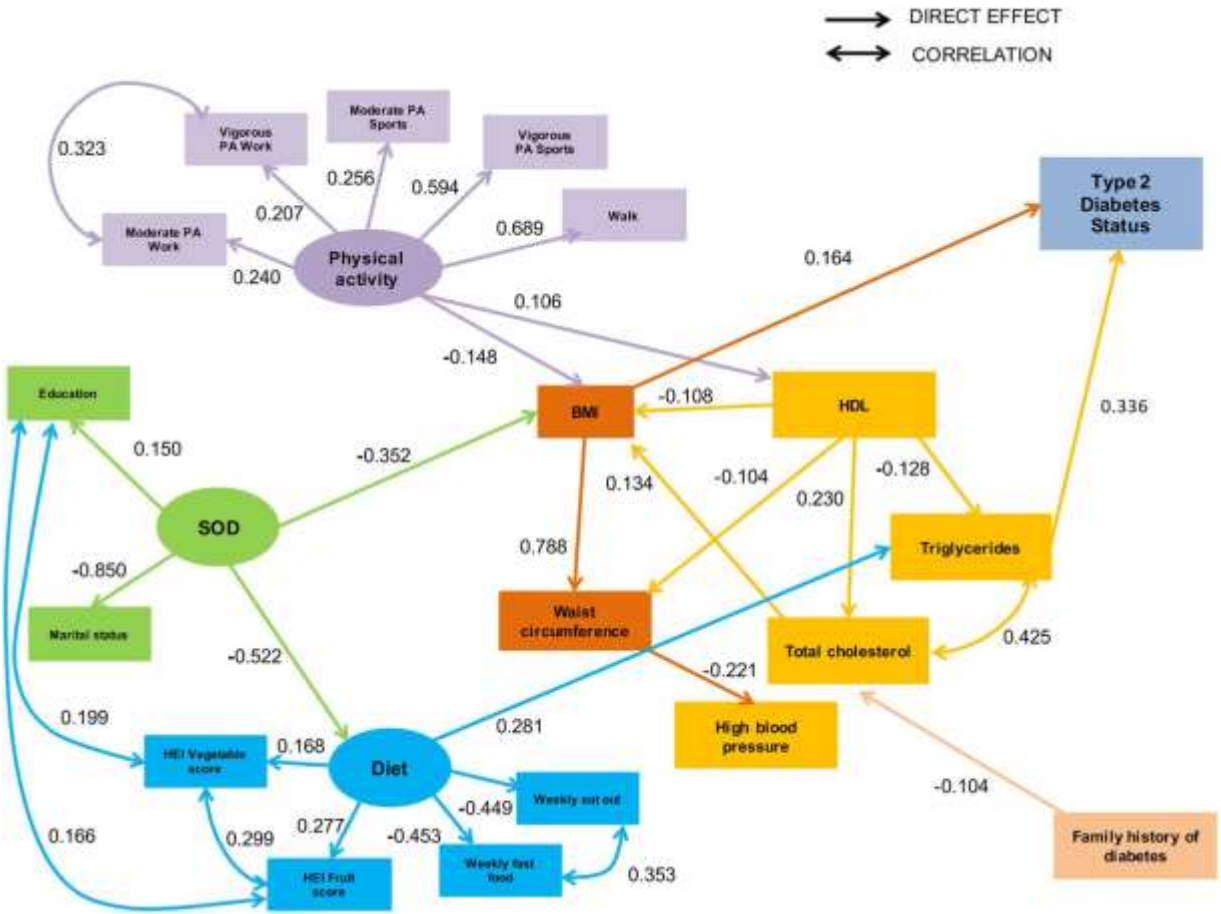
Opportunity for Latent Factor Modeling

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

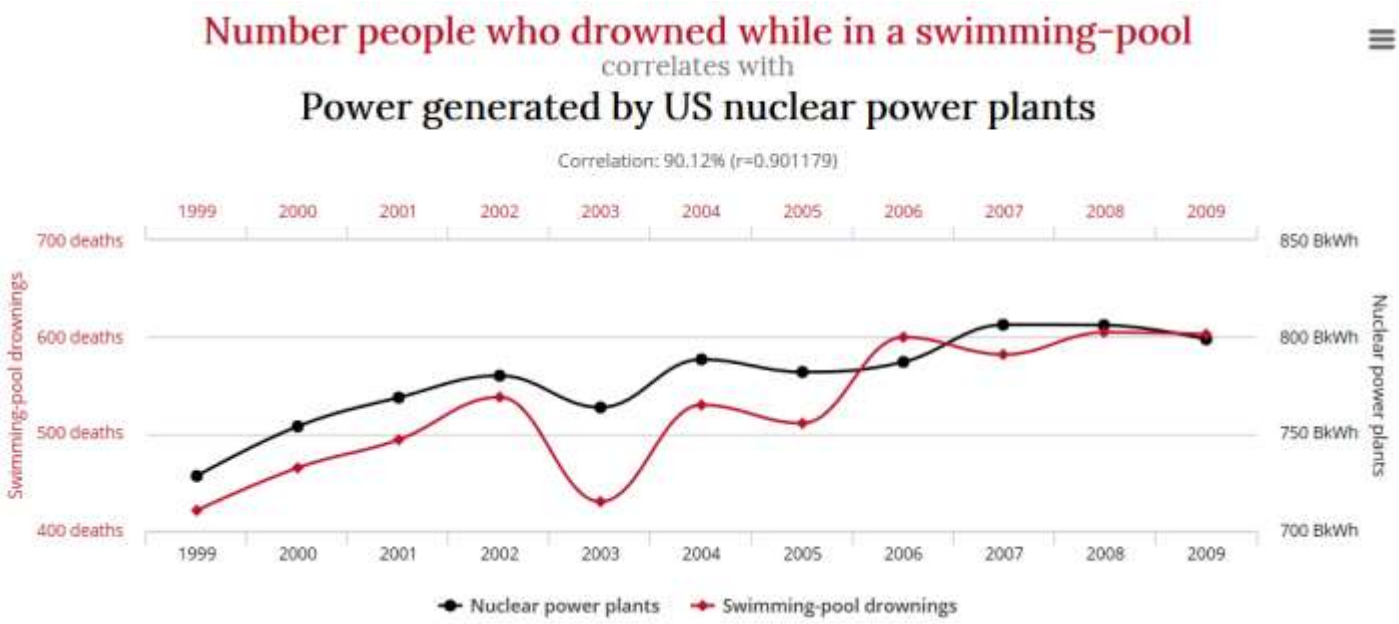
Simpson's Paradox

Opportunity for Latent Factor Modeling

Causal Learning and Counterfactual Questions

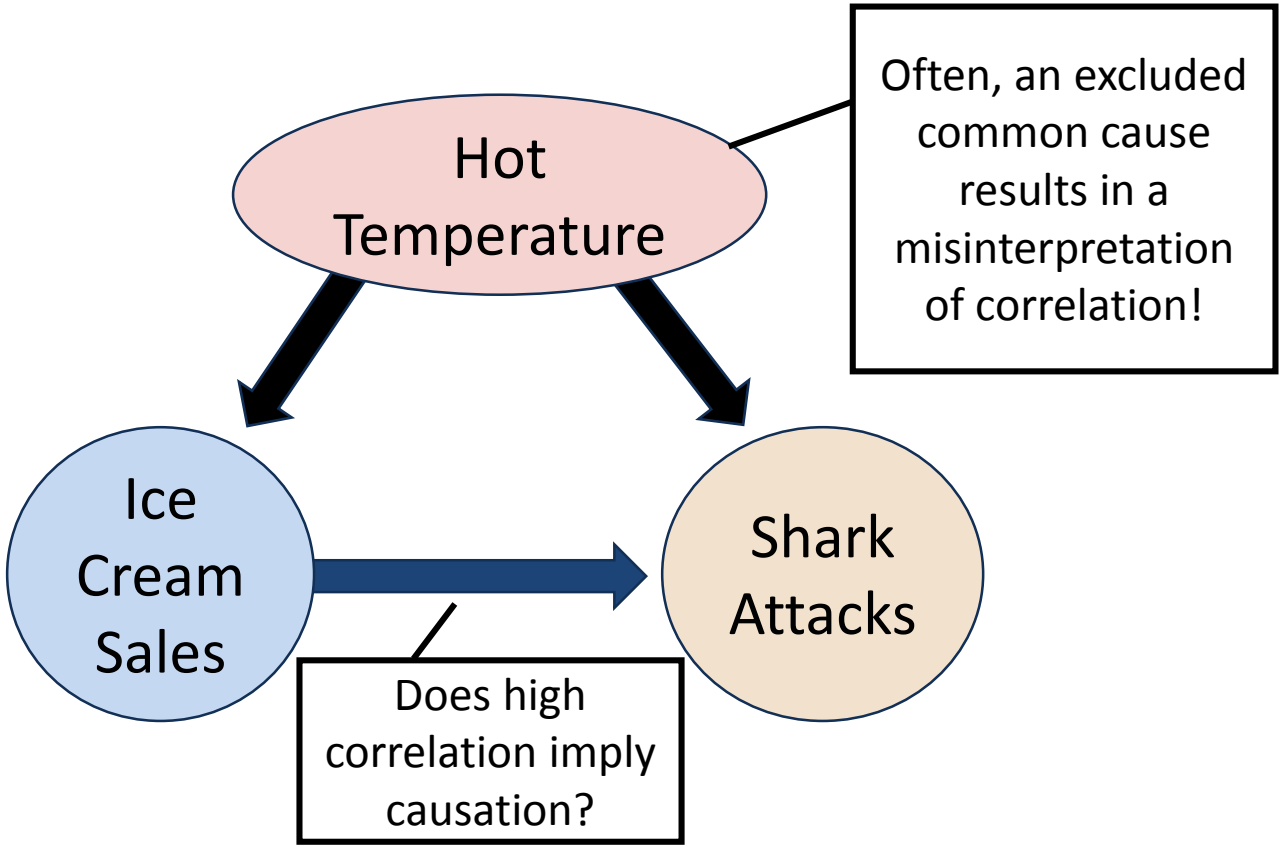
Questions

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

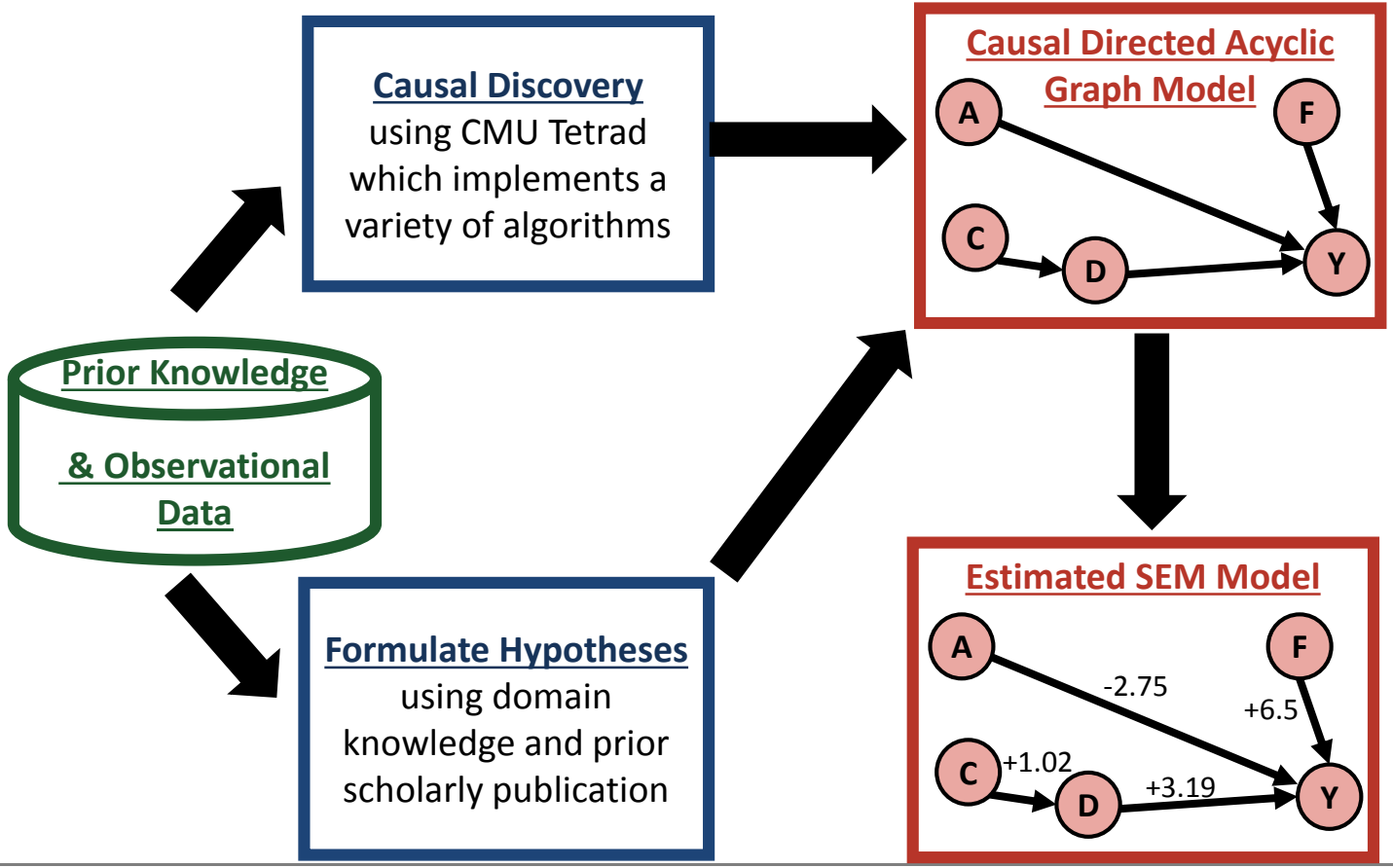
We then need to determine which paths are causal and which are spurious.

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 Counterfactural 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 Counterfactural 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-485
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 rationality presuppose 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 explicates 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 undisclosed 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 (Arlo-Costa, 2007; Weirich, 2008; Woodward, 2003).

1

PNAS

TECHNICAL REPORT
R-450
July 2016

Causal inference and the data-fusion problem

Elias Bareinboim and Judea Pearl

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