Landscape of Causal Learning and DIA Analytical Use Cases

Robert W. Stoddard

Principal Researcher

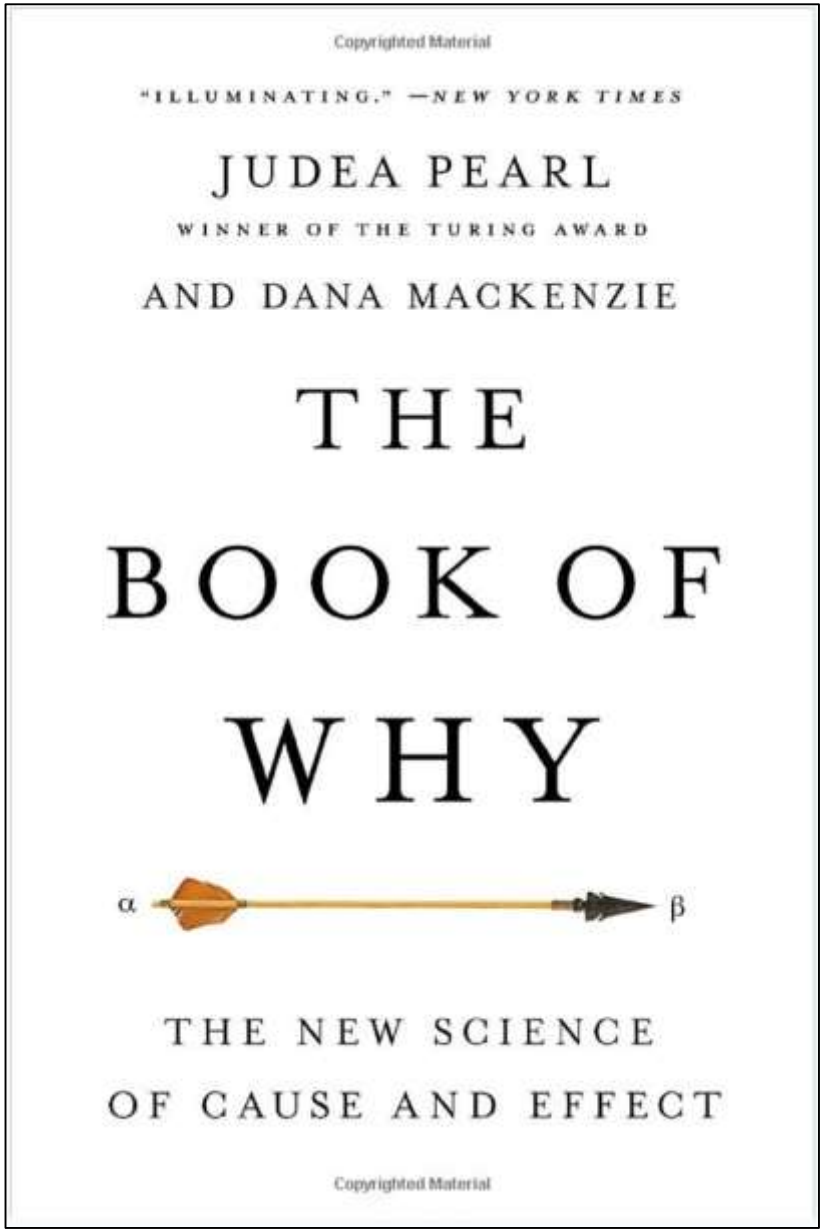
Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213

The New Science -01

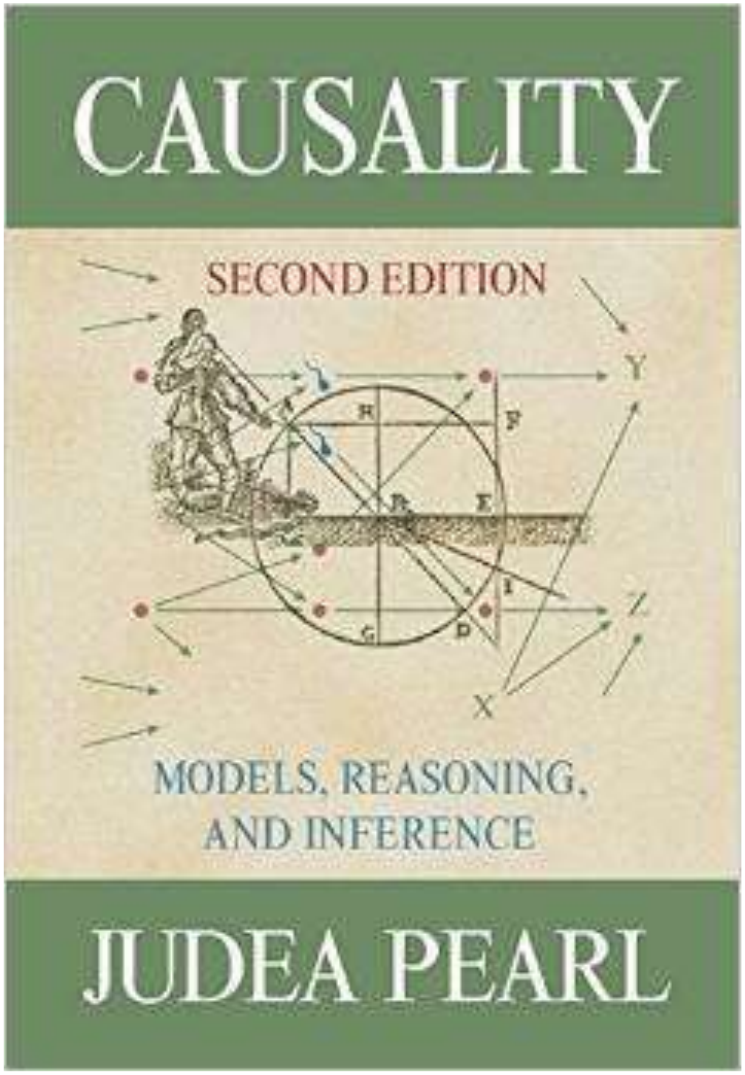
"… I see **no greater impediment to scientific progress** than the prevailing practice of focusing all of our mathematical resources on probabilistic and statistical inferences while **leaving causal considerations to the mercy of intuition and good judgment.**"

Pearl, J. (2009). *Causality*. Cambridge university press. (Preface to 1st Edition)

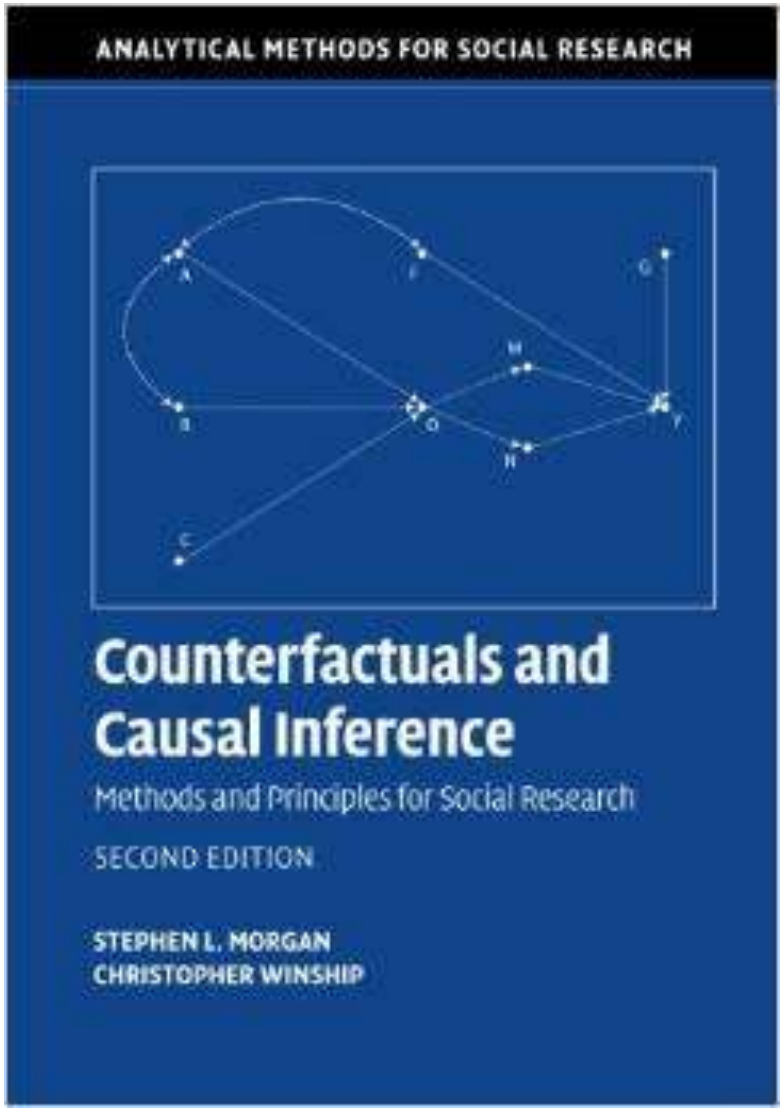
Dr Judea Pearl won the 2012 Turing Award based on this work

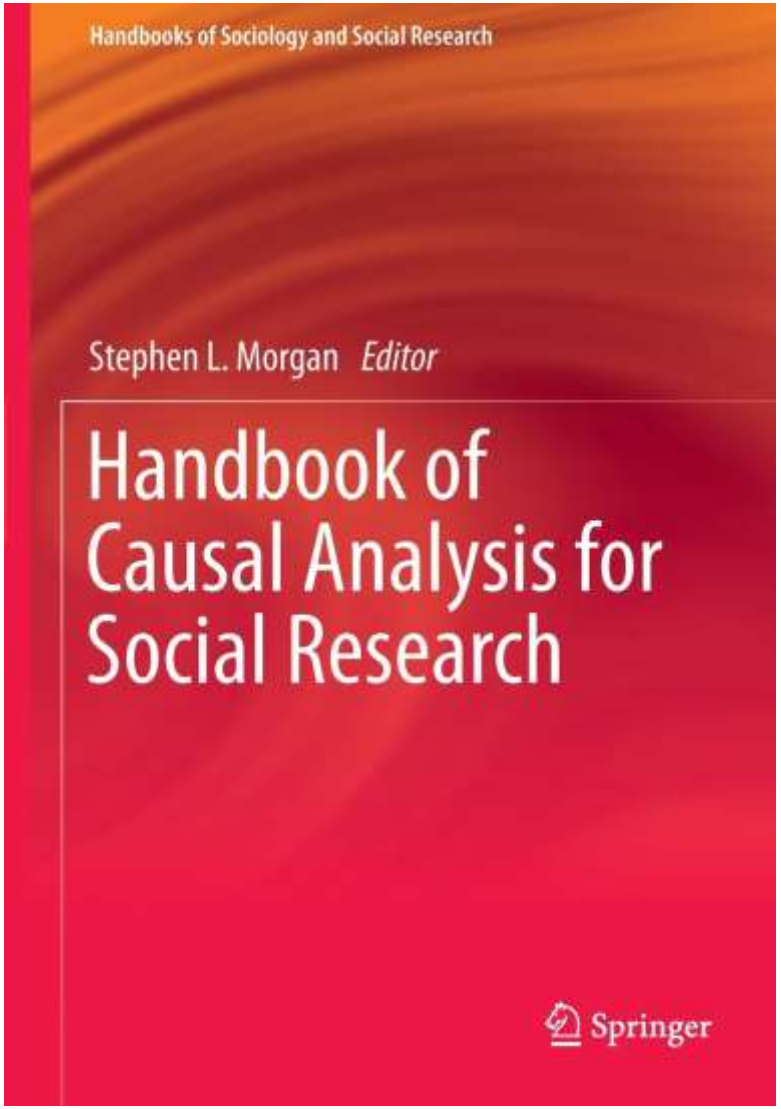


The New Science -02

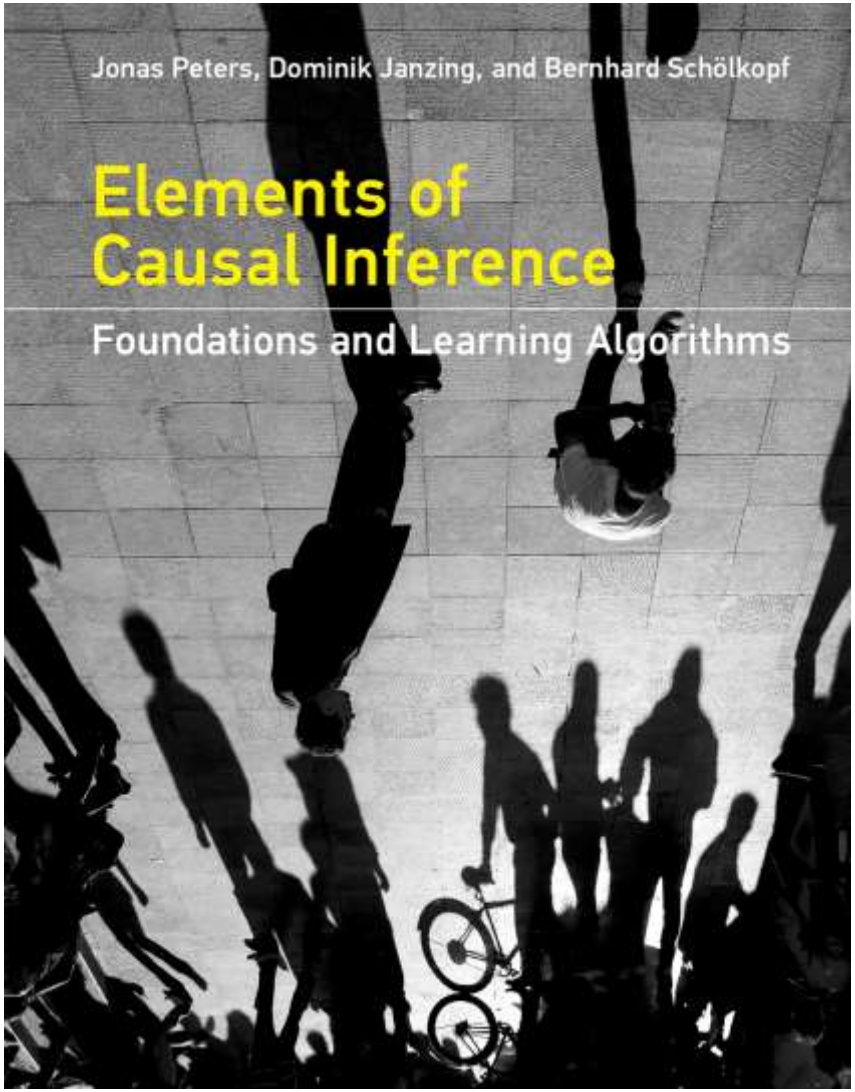


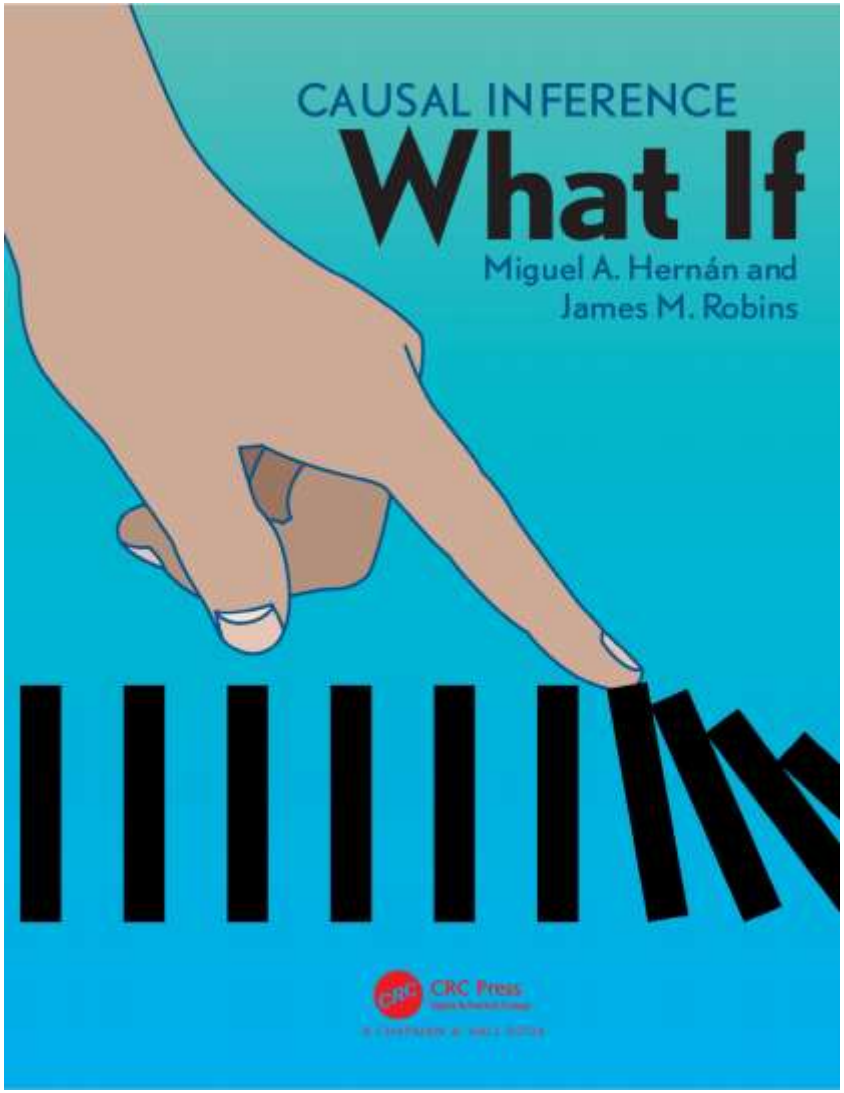
The New Science -03





The New Science -04





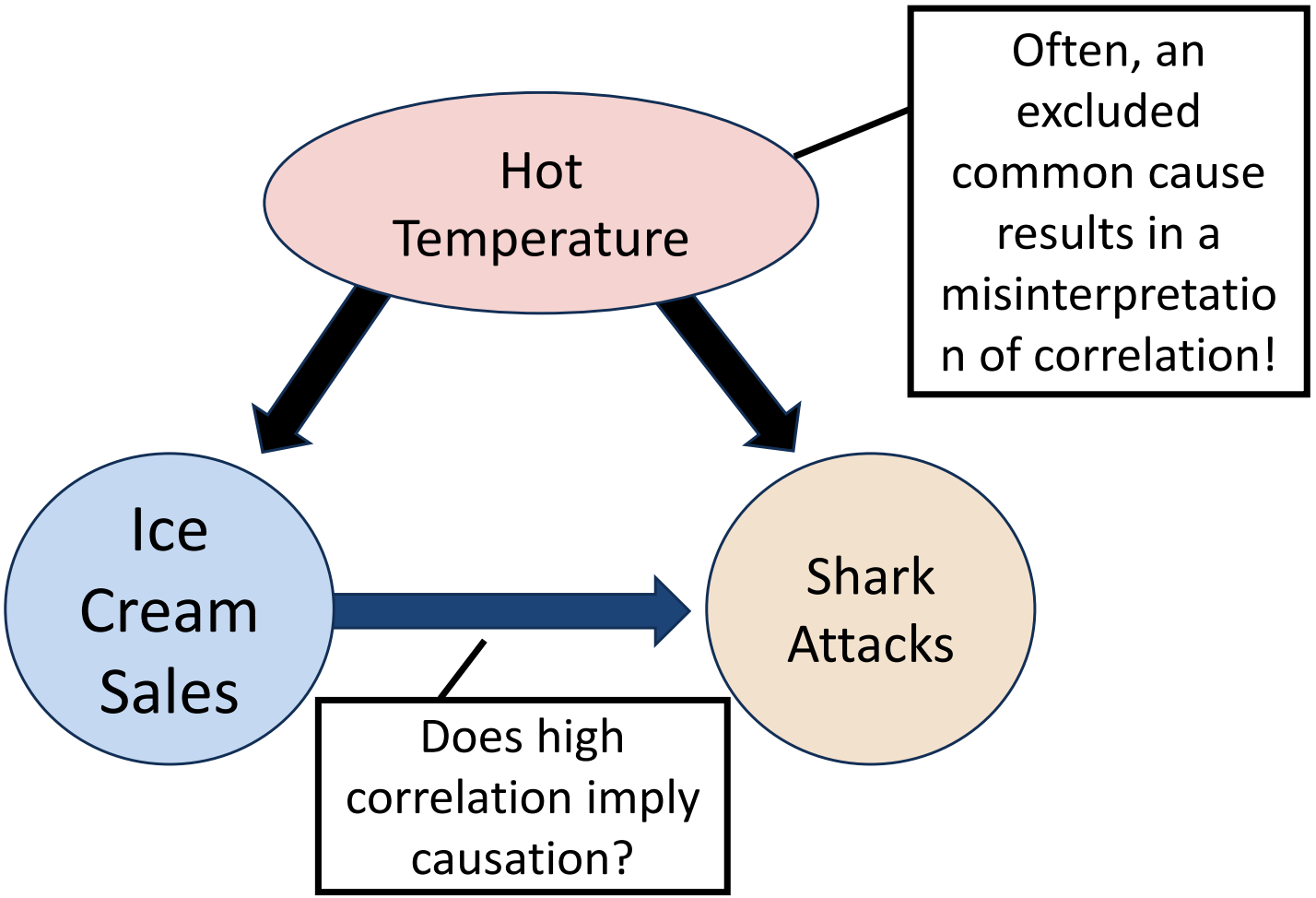
Traditional Science Rooted in Experimental Design

Prior to 1935, causal conclusions made by matching data of different conditions and comparing the outcomes.
Deemed too expensive, slow and often prohibitive.
As a result, Sir Ronald Fisher devised statistically-designed experiments, with randomization and orthogonal arrays, to more quickly intervene and draw conclusions about causality.

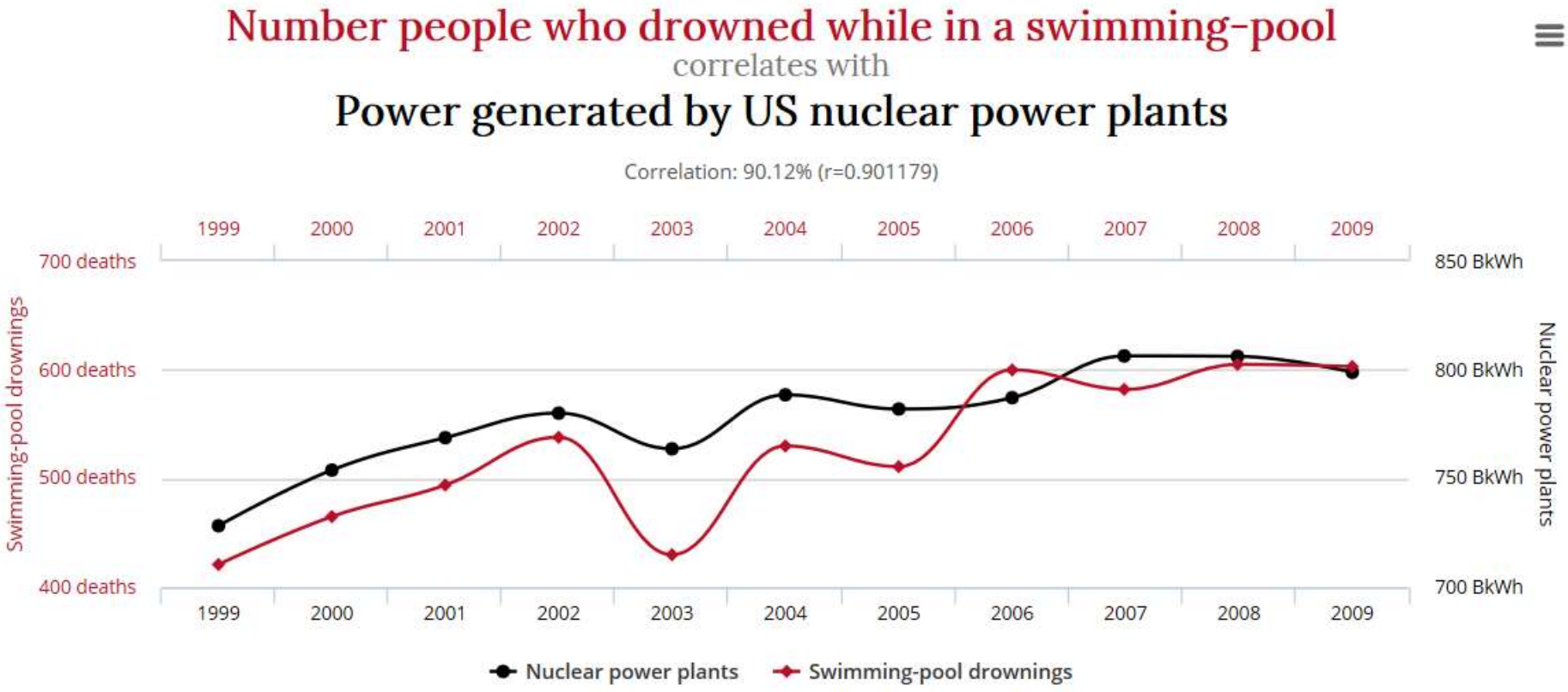
Since 1935, an evolving body of knowledge surrounding matching matured to what we now know as Directed Acyclic Graphs Causal Modeling, Counterfactual Reasoning, Instrumental Variables and Propensity Scoring.

We now have causal methods other than controlled experiments!

Misinterpreting Correlation!



Why Correlational Studies and Regression are Not Enough -01



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

Why Correlational Studies and Regression are Not Enough -02

Does Foreign Investment in 3rd World Countries inhibit Democracy?

Timberlake, M. and Williams, K. (1984). Dependence, political exclusion, and government repression: Some cross-national evidence. American Sociological Review 49, 141-146.

N = 72 data points

PO is degree of political exclusivity (repression outcome)

CV is lack of civil liberties

EN is energy consumption per capita (economic development)

FI is level of foreign investment

Reused from Dr. Richard Schienes, Center for Causal Discovery: Summer Short Course/Datathon, June 13-18, 2016, Carnegie Mellon University

Why Correlational Studies and Regression are Not Enough -03

PO = .227***FI** - .176***EN** + .880***CV**

Traditional Interpretation: Foreign Investment (**FI**) and lack of civil liberties (CV) increase political repression (**PO**)

while energy consumption (EN) reduces political repression (**PO**)

PO is degree of political exclusivity (repression outcome)

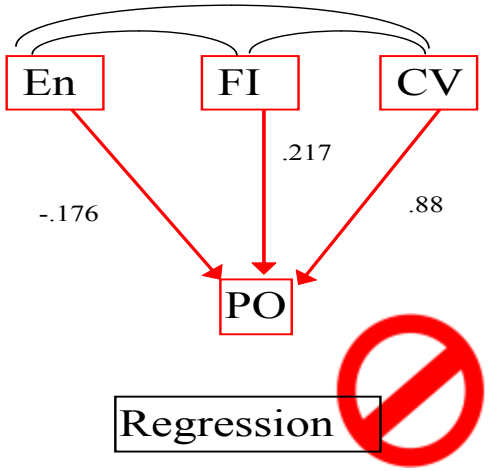
CV is lack of civil liberties

EN is energy consumption per capita (economic development)

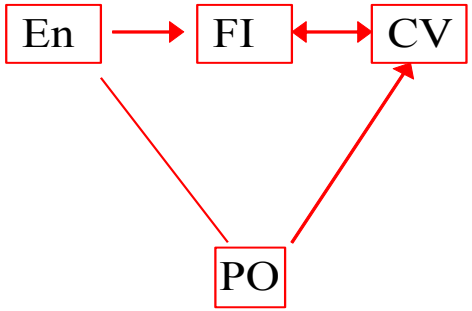
FI is level of foreign investment

Reused from Dr. Richard Schienes, Center for Causal Discovery: Summer Short Course/Datathon, June 13-18, 2016, Carnegie Mellon University

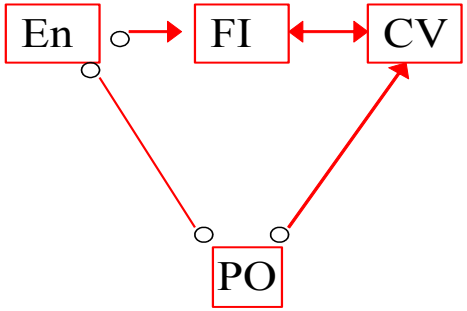
Why Correlational Studies and Regression are Not Enough -04



Regression

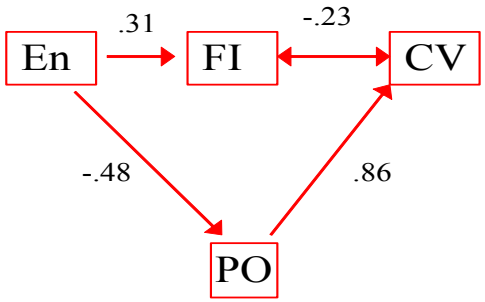


Tetrad - PC



Tetrad - FCI

There is no model with testable constraints (df > 0) that is not rejected by the data, in which FI has a positive effect on PO.



Fit: df=2, $\chi 2$=0.12, p-value = .94

Reused from Dr. Richard Schienes, Center for Causal Discovery: Summer Short Course/Datathon, June 13-18, 2016, Carnegie Mellon University

Why Correlational Studies and Regression are Not Enough -05

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

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

We determine which paths are causal and which are spurious.

We must block spurious correlation paths.

Lastly, we conduct regression with the causal set of factors!

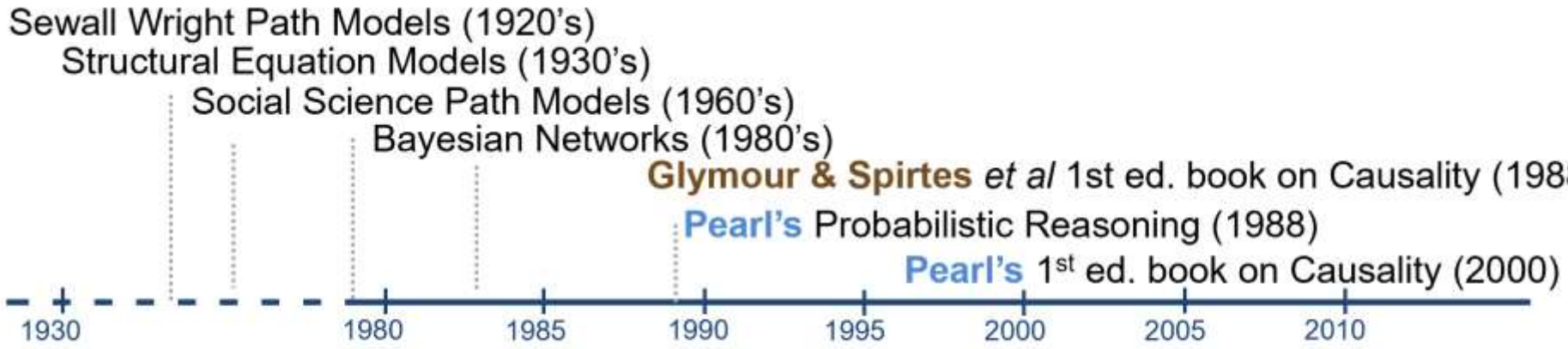
Remember, suitability of the regression model depends on the context of a DAG!

Different Uses for Correlation / Causation

Correlation	Causation
Classifying & identifying	Influencing & acting
Informational value of different evidence	Using evidence to guide policy or actions
Prediction & reasoning given observations	Prediction & reasoning given interventions
Probable explanations for some event or issue	Ways to produce or prevent an event or problem

Reused from Dr. David Danks, Chair Dept of Philosophy, Carnegie Mellon University

Evolution of Causal Inferential Techniques



TETRAD – An Open Source Tool for Causal Learning

Carnegie Mellon University
http://www.phil.cmu.edu/tetrad/

University of Pittsburgh
http://www.ccd.pitt.edu/

For video tutorials from 2016 summer short course:
http://www.ccd.pitt.edu/training/presentation-videos/

CMU OLI - Causal and Statistical Reasoning
http://oli.cmu.edu/courses/future/causal-statistical-reasoning/

Glymour & Spirtes *et al* 2nd Edition Book on Causality (2001)

Morgan Counterfactuals & Causality (2007)

Pearl's 2nd Edition Book on Causality (2009)

Morgan Counterfactuals & Causality (2014)

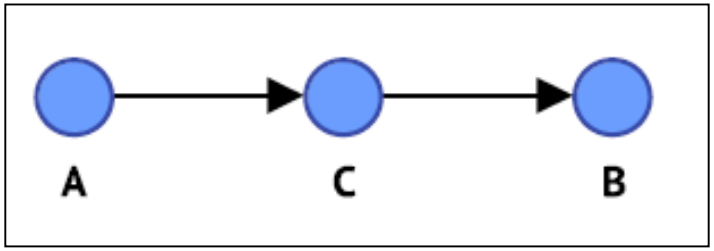
Peters Elements of Causal Inference (2017)

Focus is on Directed Acyclic Graphs - 1

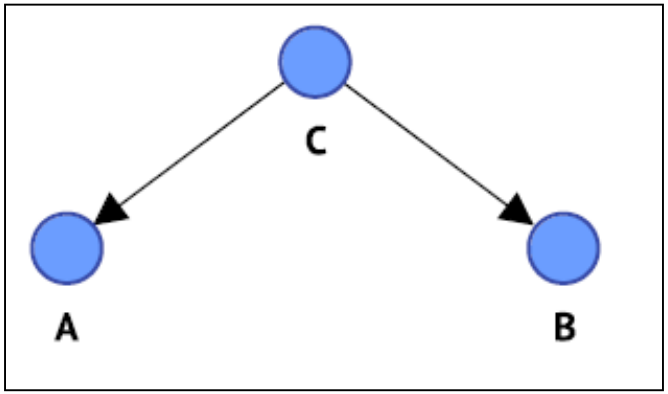
1. DAGs consist of:
   a) nodes (variables),
   b) directed arrows (possible causal relationships ordered by time), and
   c) missing arrows (confident assumptions about absence of causal effects)
2. DAGs are nonparametric
   a) No distributional assumptions
   b) Linear and/or nonlinear
3. DAGs have both causal paths and non-causal (spurious) paths

Focus is on Directed Acyclic Graphs - 2

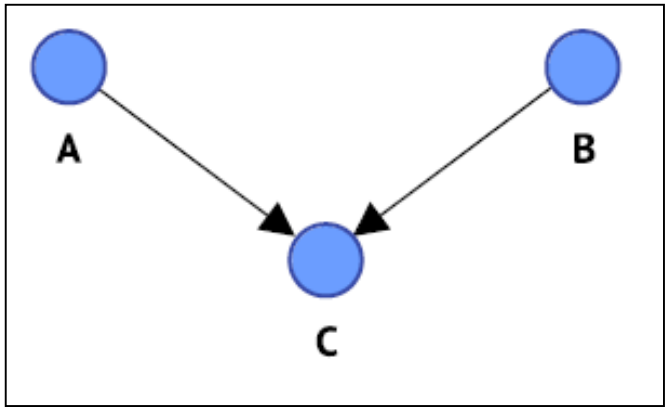
a) Indirect Connection



b) Common Cause



c) Common Effect (Collider)



Blocking or Adjusting Paths

1. Controlling a variable
2. Stratifying a variable
3. Setting evidence on a variable
4. Observing (or conditioning on) a variable
5. Matching a variable (eg making distributions of sub-populations as similar as possible for estimating effect size)

These techniques critically inform which factors to put into the regression equation and which ones to keep out!

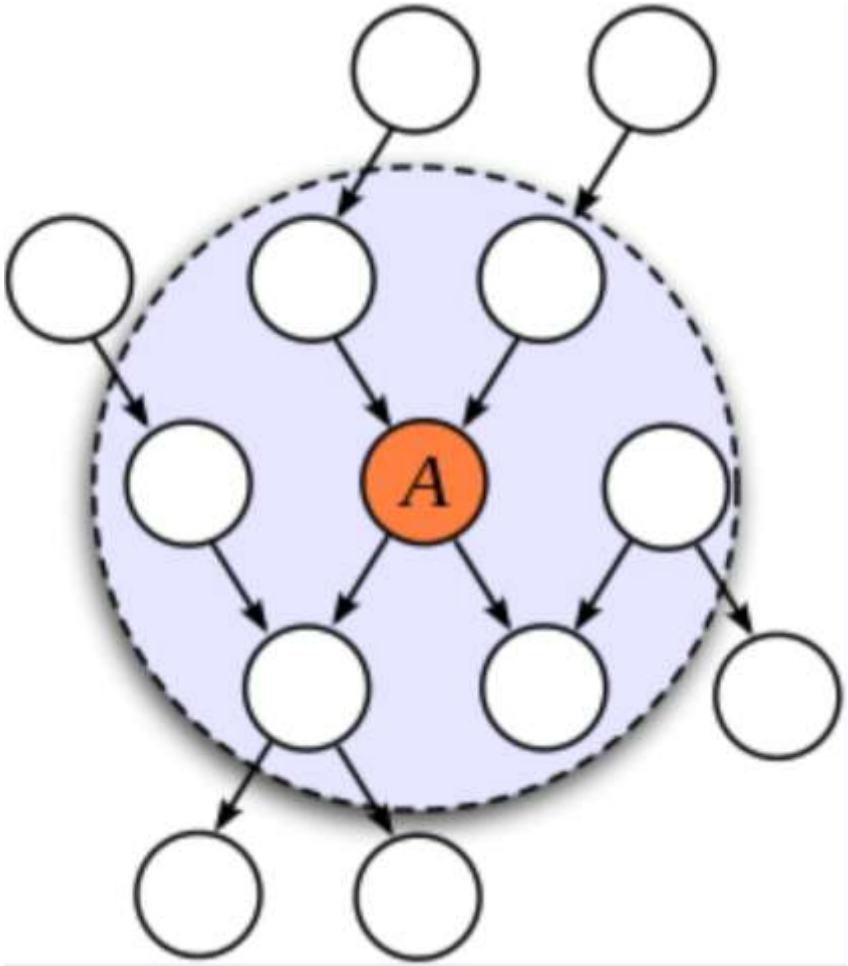
Focus on the Markov Blanket

The Markov blanket for a node contains all the variables that shield the node from the rest of the network.

This means that the Markov blanket of a node is the only knowledge needed to predict or cause the behavior of that node and its children.

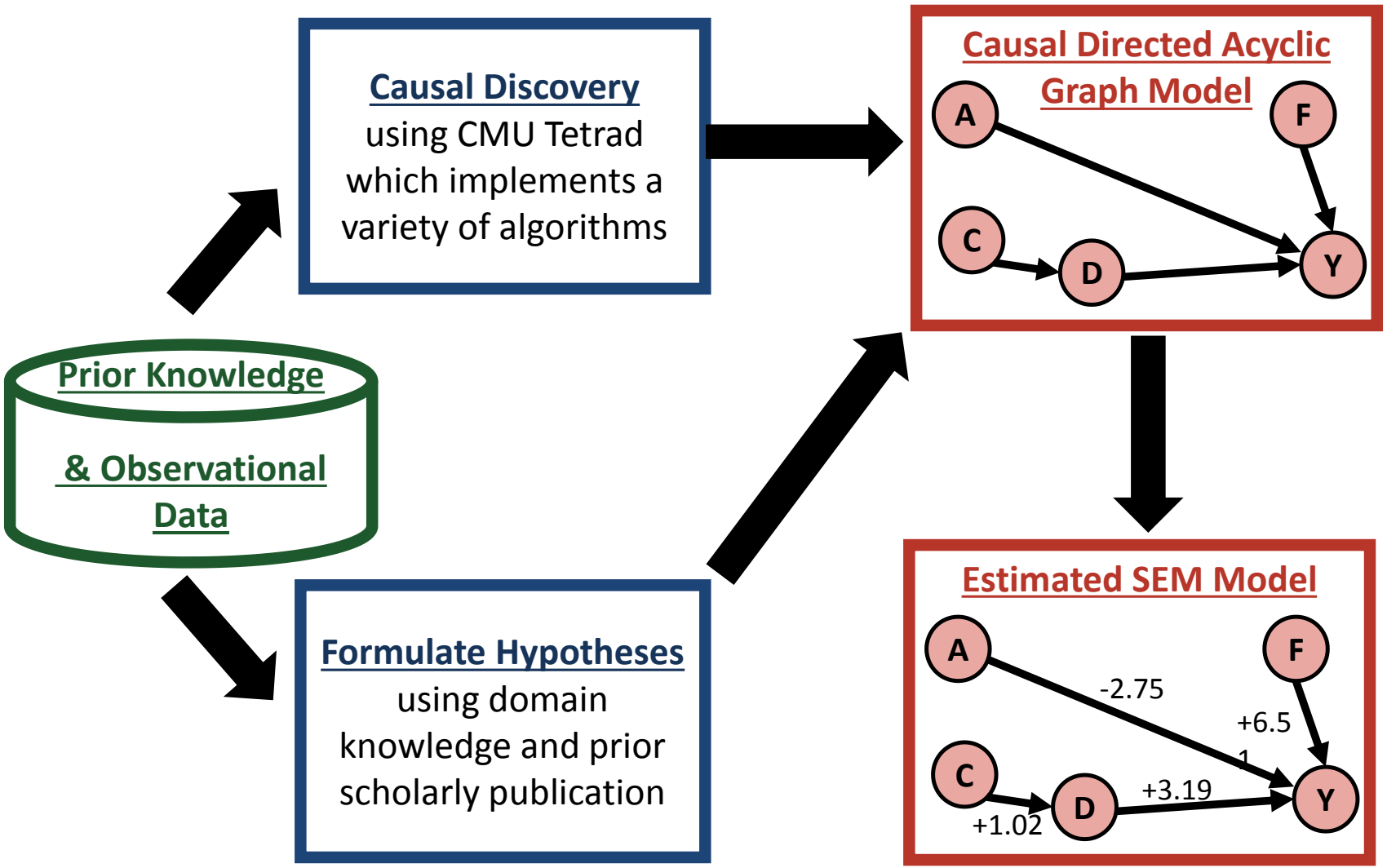
The term was coined by Judea Pearl in 1988.

The blanket consists of the parent nodes, the children nodes, and the other parents of the children nodes.



https://en.wikipedia.org/wiki/Markov_blanket

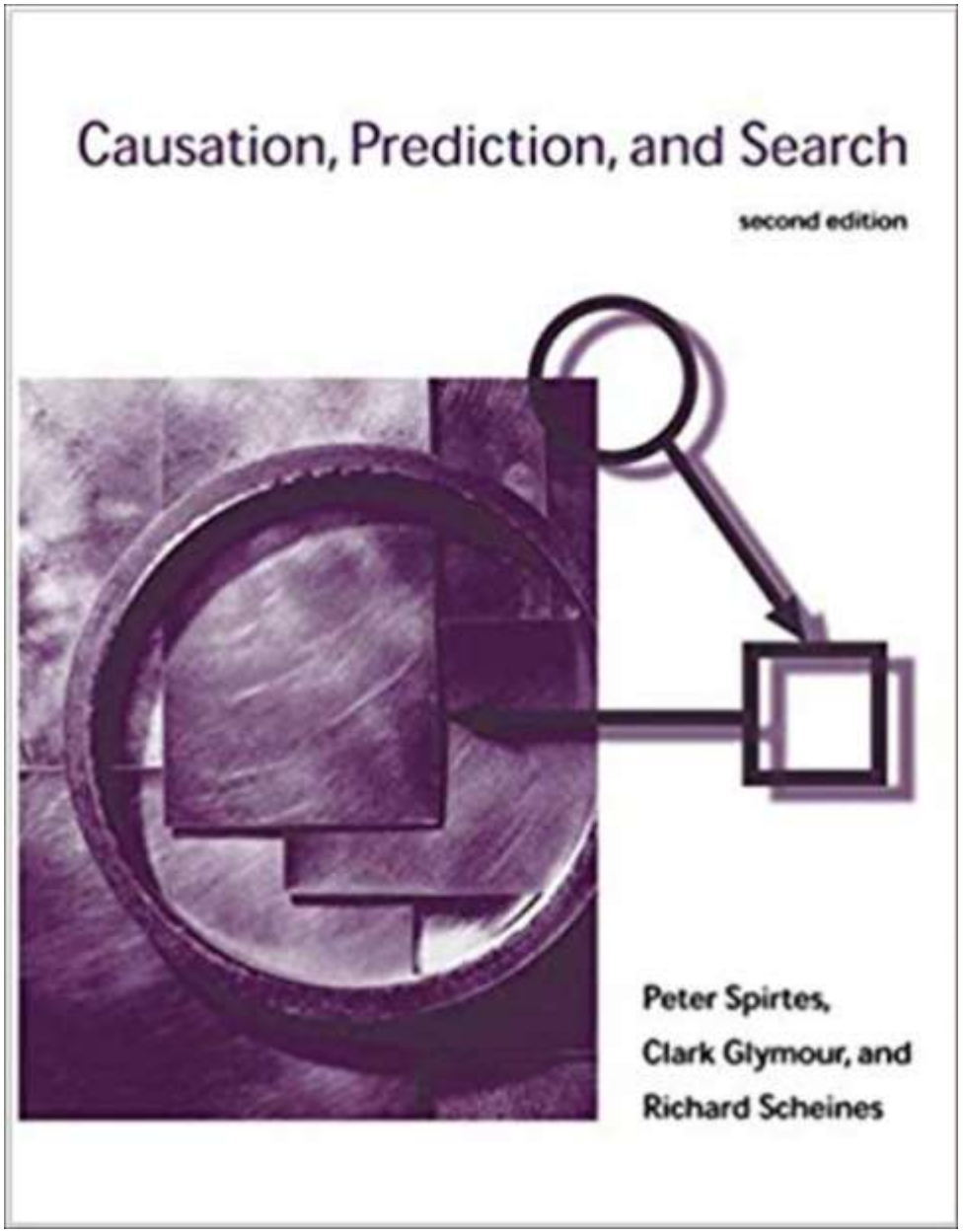
The Causal Learning Landscape



Causal Search



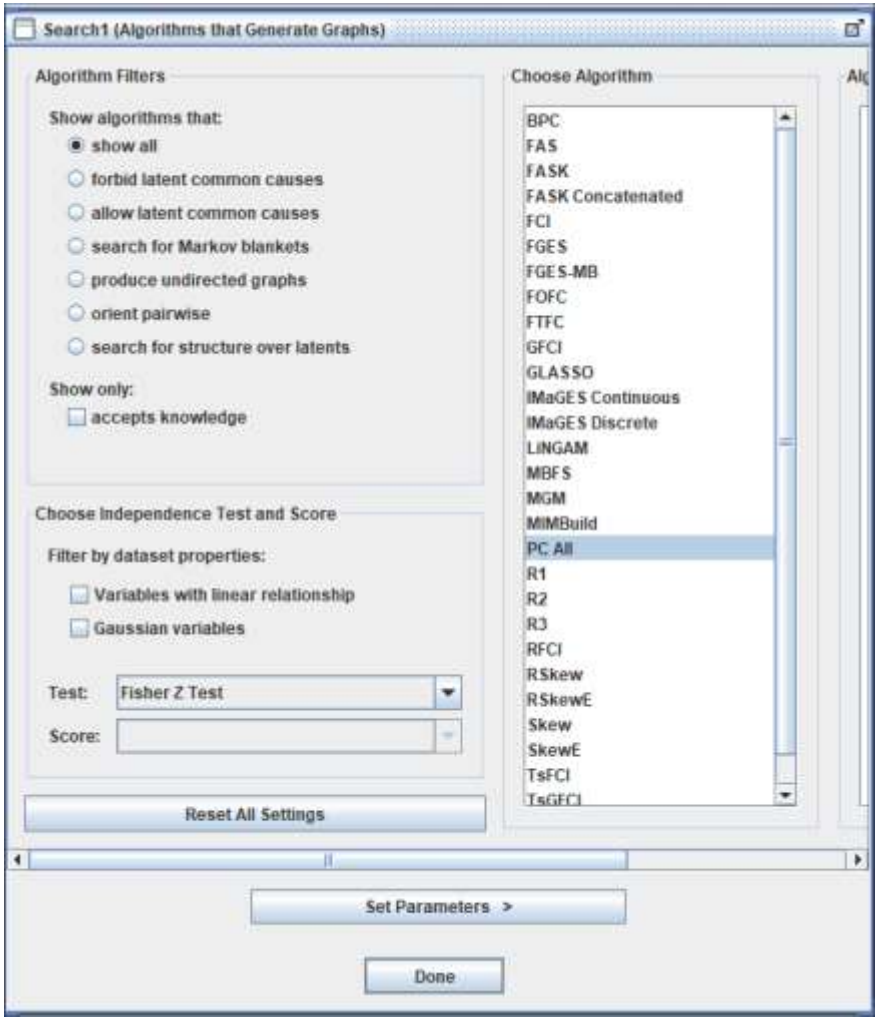


Practical Toolkit for Causal Search and Estimation

Multiple approaches to searching a dataset:

1. **Constraint-based**: determined from independences in the data
2. **Score-based** (**Bayesian**): determined from likelihood calculation of different DAGs given the data
3. **Hybrid**



Causal Search in Small and Large Data

- *Which genes regulate flowering time in Arabidopsis thaliana?*
- Causal discovery on observations of gene activation:
  9 (of 21,326) genes are good candidates
  - using only 47 samples…
- Greenhouse study that used knockout variants:
  4/9 were actual regulators

Reused from Dr. David Danks, Chair Dept of Philosophy, Carnegie Mellon University

Causal Estimation

Estimating the quantitative direct and indirect causal effects of factors on specific outcomes

Mediation causal effects: A causes B causes C

Moderated causal effects: A and B both cause C with an interaction effect

And the real fun begins with…

Moderated mediation
Mediated moderation

New Causal Metrics Replace p and r^2

Causal metrics are based on answering counterfactual questions such as:

"Had event A alternatively occurred, what would be the potential outcome Y?"

Counterfactual reasoning helps extend the logic of randomized experiments to observational data

Counterfactual reasoning involves Causal Measures:

- Total Causal Effect (TCE)
- Individual Level Causal Effect (ICE)
- Average Causal Effect (ACE)

Causal Estimation Techniques

Structural Equation Modeling

Propensity Scoring

Instrumental Variables

Front Door Adjustment

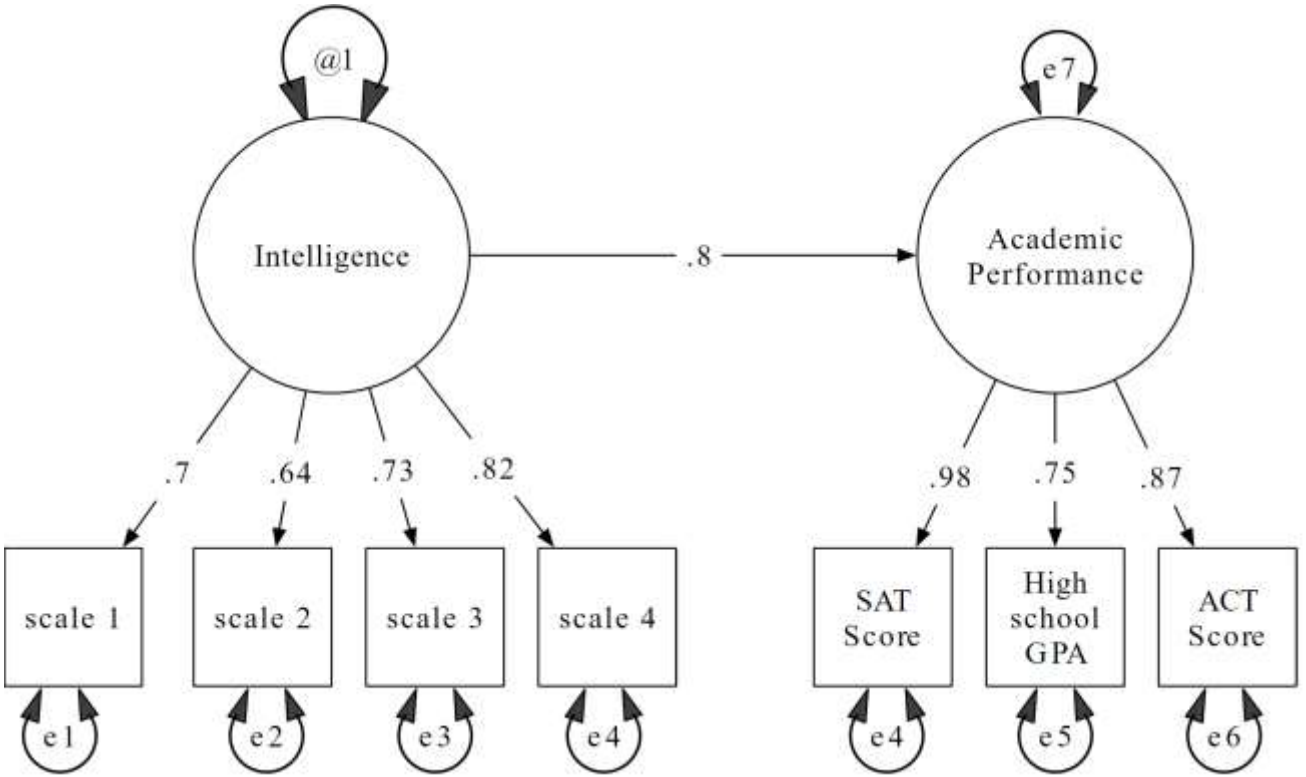
Back Door Adjustment

Do-Calculus Causal Algebra

Sigma-Calculus Algebra

Structural Equation Modeling

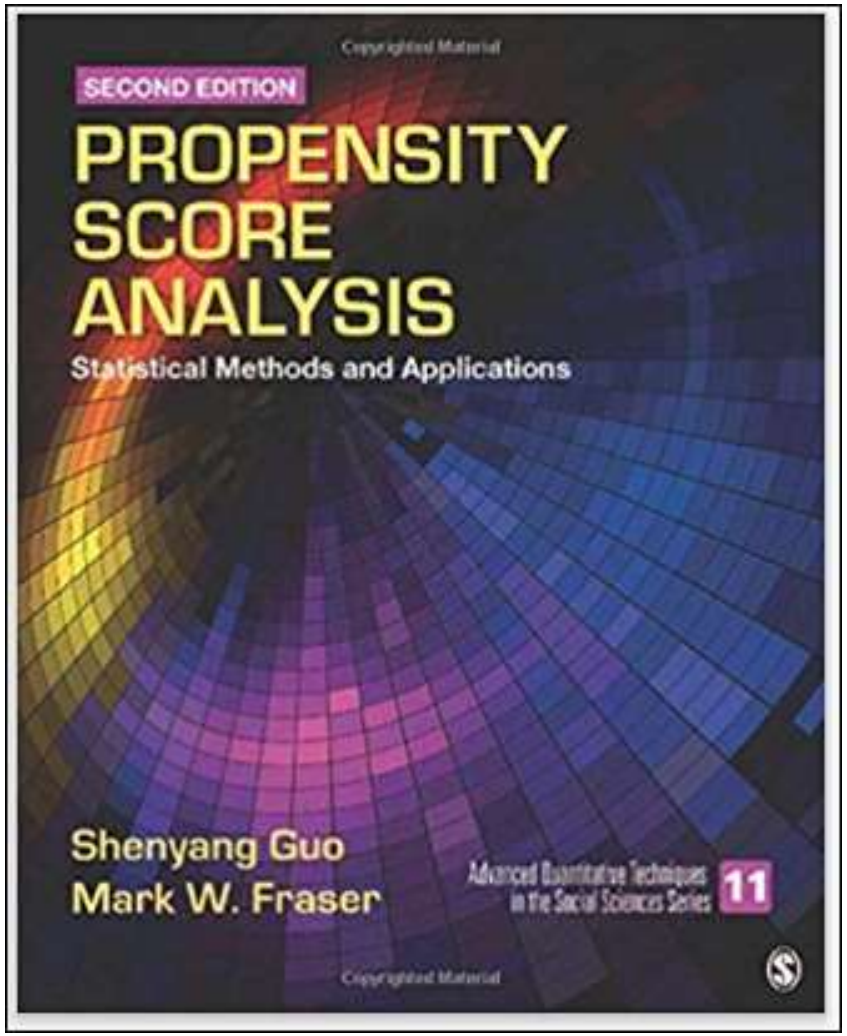
https://en.wikipedia.org/wiki/Structural_equation_modeling



Simultaneous multiple regression models in which factors (nodes) can serve as both independent and dependent factors; latent factors may exist

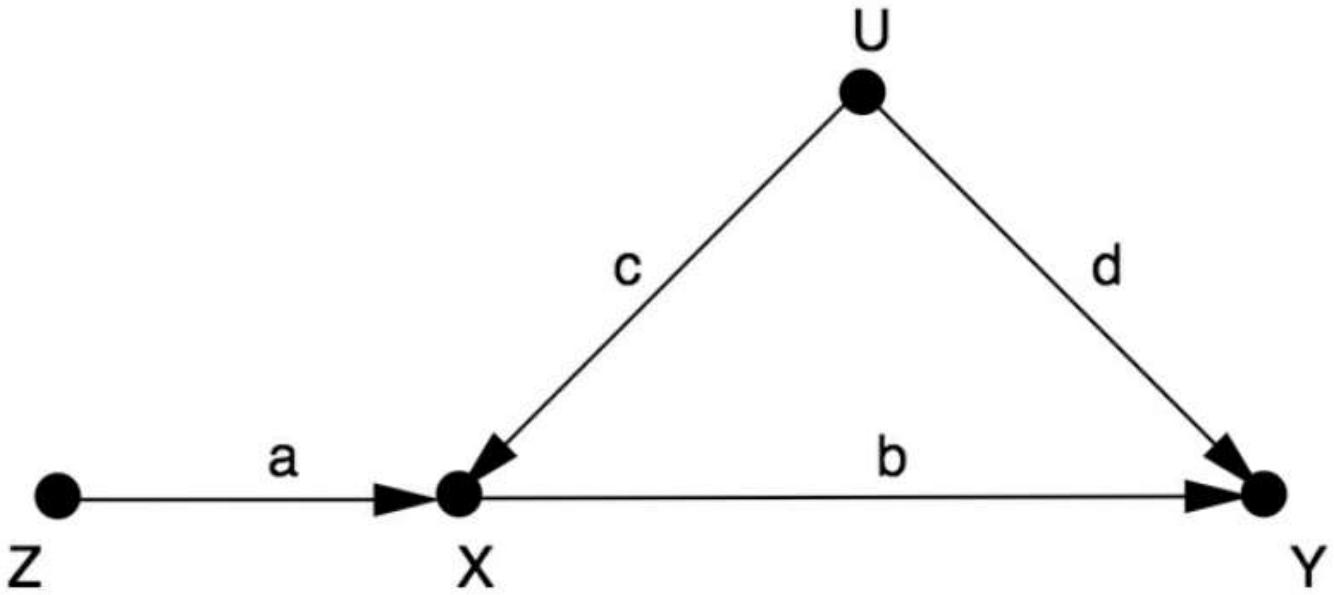
Propensity Scoring

1. Logistic regression (probability of x level treatment) called a Propensity Score
2. Use Propensity Scores to match data records of treatment and control
3. Drop unmatched data records
4. Conduct causal analysis



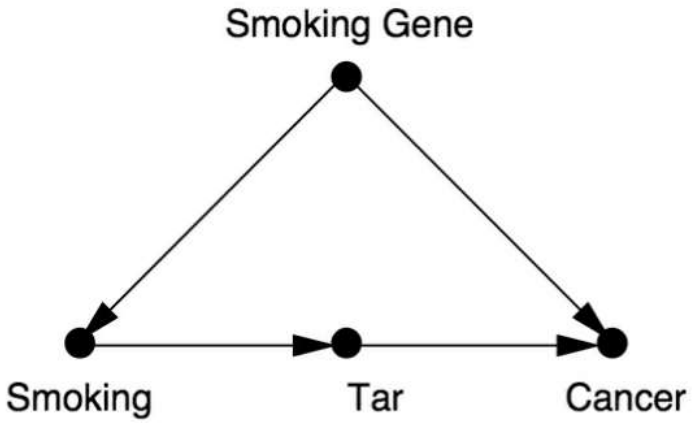
See Shenyang Guo and Mark W. Fraser, 2014, "Propensity Score Analysis"

Instrumental Variables



Causal effect (a) is the slope (1) of regression of X = f(Z)
Causal effect (a * b) is the slope (2) of regression of Z = f(Y)
Thus, b = slope(2) / slope(1) and is unbiased even though we don't collect data on the possible confounder (U)

Front Door Adjustment



Evaluating smoking effect on cancer

We can't block Back Door path thru Smoking Gene as we don't measure Smoking Gene

We must try Front Door criterion of Smoking to Tar to Cancer

Using rules of Do-Calculus, we adjust on both Smoking and Tar and eventually compute the P(Cancer|Smoking) in terms of Total Causal Effect (TCE) and Average Causal Effect (ACE)

Taken from The Book of Why, 2017

Back Door Adjustment

A Back Door path between X and Y is any path between the two that begins with an arrow pointing to X

X and Y can be **de-confounded** if all Back Door paths are **blocked** because such paths allow spurious correlation

If we block on a set of variables Z, we **must ensure** no member of Z sits on a causal path of X to Y

Do-Calculus Causal Algebra

Rule 1: Variables that are irrelevant to Y, possibly conditional on other variables, may be removed from the equation

Rule 2: If a set of variables (Z) blocks all back door paths from X to Y, then conditional on Z, do(X) may be replaced with P(X)

Rule 3: We can remove do(X) from P(Y | do(X)) if there are no causal paths from X to Y

Taken from The Book of Why, 2017

Sigma-Calculus Causal Algebra

Do-Calculus only handles hard interventions
e.g. set specific factors in model to a constant setting

Sigma-Calculus handles soft interventions
e.g. set specific factors in model to uncertain settings such as a possible range of values

Practical Toolkit for Causal Learning

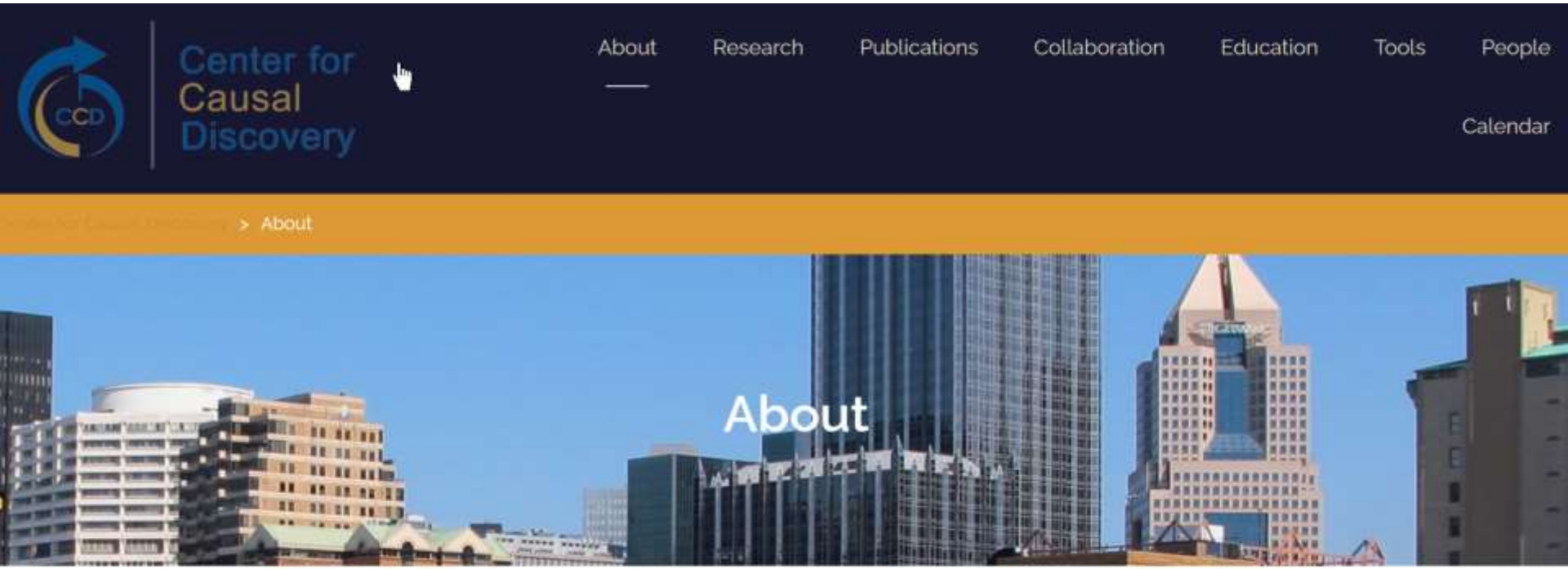
Tetrad for Causal Search (open source, GitHub)

Mplus for Structural Equation Modeling (commercial)

Stata for unique Propensity Scoring routines (commercial)

Tetrad - 1

www.ccd.pitt.edu/tools/

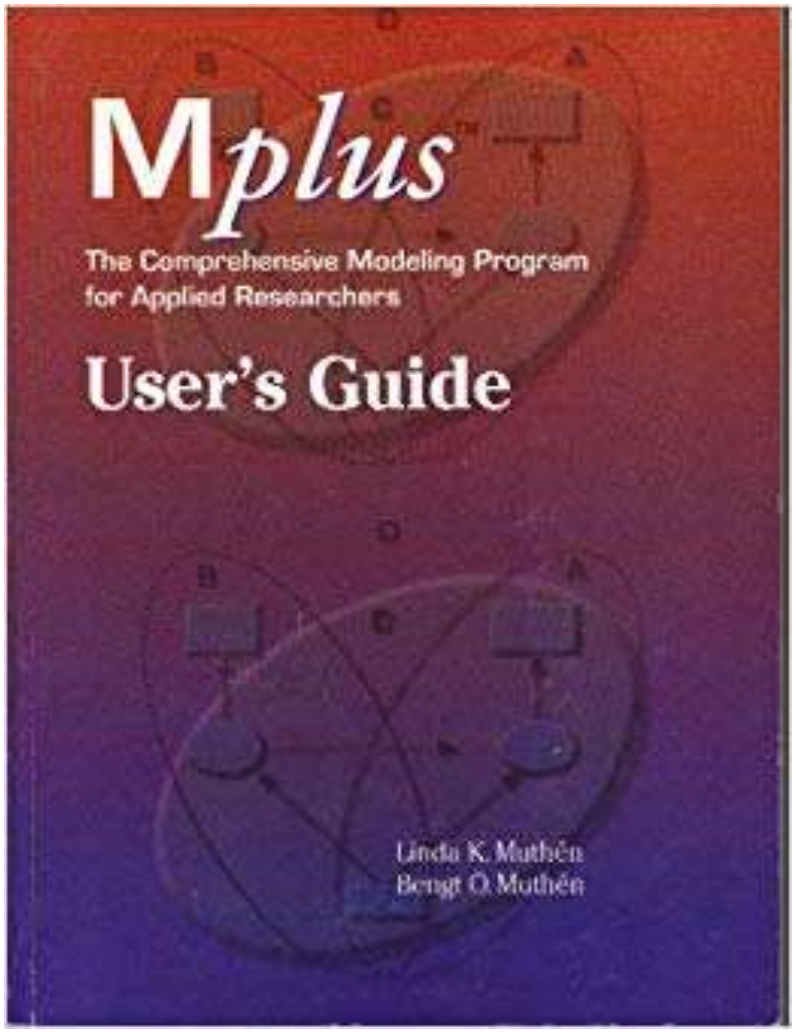


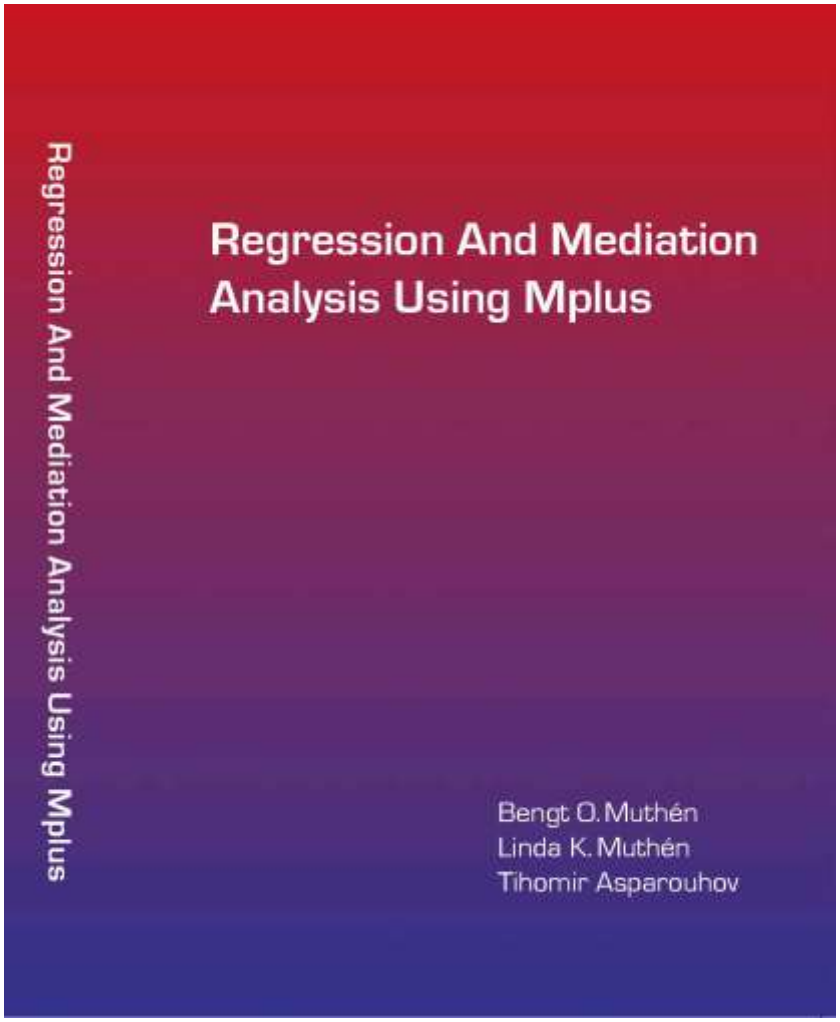
Tetrad - 2

- **Causal Web** – web application w/Super Computer
- **Causal Command** – Java library with shell scripts
- **Py Causal** – Python module for causal search
- **R-Causal** – R module for causal search
- **Tetrad** – desktop application with GUI interface

Mplus

https://www.statmodel.com/





Use Cases of The New Science

Generic Use Cases of Causal Learning

Specific Use Cases

Generic Use Cases of Causal Learning

- Reproduce and confirm/clarify prior research
- Create actionable models of performance
- Root cause analyze problems using Big Data
- Reduce time for experimentation through causal search of historical data
- Use causal structures to better understand and test system behavior
- Use causal structures to explain and assure AI solutions

Multivariate Causal Models of Outcomes

- Go beyond single factor correlation and regression
- Develop multiple regression models based on causal factors

Alternative to Experiments

- Experiments can be almost impossible to perform
- Often is not enough time or resources
- Depending on factors, experiments may be unethical
- Maintaining experimental discipline may be extremely challenging
- Environmental conditions of the development process, people, technology, etc… cause results to expire in applicability

Digest Observational Data

- As opposed to experimental science, causal learning enables the analytics and modeling of purely observational data

- Ability to digest observational data dramatically increases the opportunity to analyze different types of data from many different sources or organizations

Recover Poor Experimental Design and/or Execution

- Causal estimation techniques have been published to help recover from poor experimental design and poor experimental execution

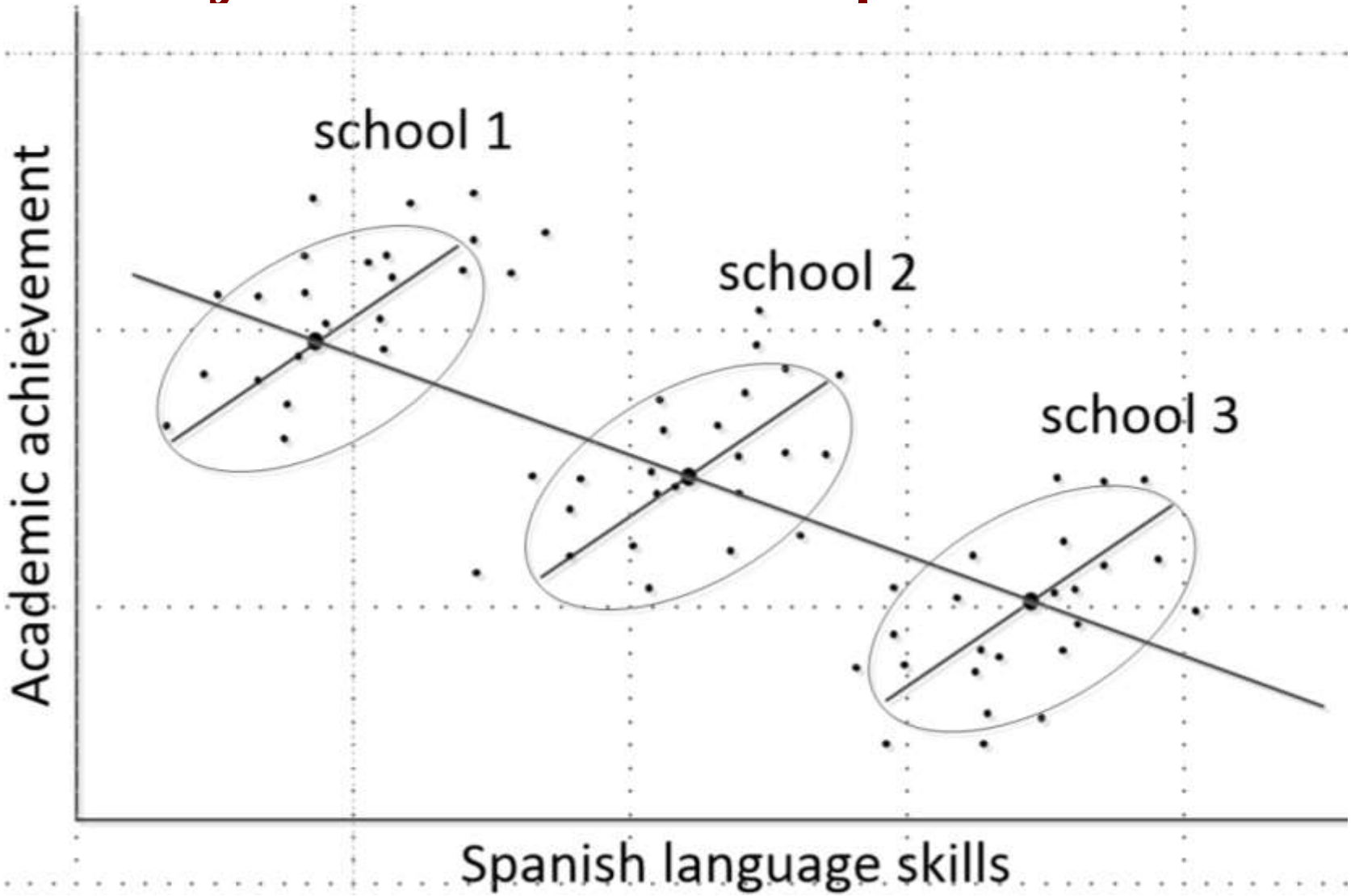
- Partial experimental data can still be analyzed via causal estimation

Account for Sample Data Bias

During WWII, statistician Abraham Wald was asked to help the British decide where to add armor to their bombers. After analyzing the records, he recommended adding more armor to the places where there was no damage!

This seems backward at first, but Wald realized his data came from bombers that survived. That is, the British were only able to analyze the bombers that returned to England; those that were shot down over enemy territory were not part of their sample. These bombers' wounds showed where they could afford to be hit. Said another way, the undamaged areas on the survivors showed where the lost planes must have been hit because the planes hit in those areas did not return from their missions. https://www.johndcook.com/blog/2008/01/21/selection-bias-and-bombers/

Identify and Deal with Simpson's Paradox



Determine Robustness of Models and AI Solutions

Causal inferential techniques have been published demonstrating ability to evaluate robustness of models and AI solutions to changes in environment, etc…

Causal techniques are proposed and practical in conducting domain transportability as well as transfer learning using causal inference

Causal techniques are also being integrated into Reinforcement Learning to produce more robust AI solutions (Causal Reinforcement Learning)

Conduct Exploratory Models

Explore causal relationships in the data via causal search

We have found up to 80% of statistically significant correlation is spurious and not based on causal relationships!

Combining causal search results with Exploratory Structural Equation Modeling can produce enlightening thoughts on more robust and actionable models

Conduct Confirmatory Models

Causal search results combined with confirmatory structural equation modeling can begin to make major leaps forward in the nature of needed measurement models

This approach includes both formative and reflective structural equation modeling, as well as covariance-based and partial least squares SEM modeling

This type of modeling confirms the significant measurement factors to include related to latent factors (discussed on next slide)

Support Feature Engineering with Latent Factor Modeling

Structural equation modeling enables the use of unmeasured latent factors along with manifest factors (aka measured factors)

In reflective structural equation modeling, the latent factors are assumed to be the truth signal and the manifest factors are the observable result of the truth signal. Hence, the causal arrows point from the latent factor to the manifest factors.

Using latent factors in models achieves much less noise from measurement error.

Data Fusion Enhanced with Causal Inference

"Piecing together multiple datasets collected under heterogeneous conditions (i.e., different populations, regimes, and sampling methods) to obtain valid answers to queries"

"Availability of multiple heterogeneous datasets presents new opportunities to big data analysts, because the knowledge that can be acquired from combined data would not be possible from any individual source alone"

"Biases that emerge in heterogeneous environments require a new analytical tool; a general, nonparametric framework for handling these biases and, ultimately, a theoretical solution to the problem of data fusion in causal inference tasks"

Bareinboim, Elias; Pearl, Judea; "Causal inference and the data fusion problem", www.pnas.org/cgi/doi/10.1073/pnas.1510507113

Let's Not Take 50 Years to Adopt!!!

Dr. Judea Pearl analyzed the history of major statistical innovations and discovered that on average, it **takes 50 years for an innovation to be widely adopted** by the statistical community.

His desire is to speed up the adoption time for causal learning. He separately muses over the **millions of smoker deaths that could have been prevented** had causal learning evolved 30-40 years sooner to establish link of smoking to cancer!

Become and Educated Consumer

To accelerate adoption of both causal search and causal estimation, we must all become educated consumers of causal learning results.

If you see a research study result or a newly-developed model, ask what the causal basis is for the model.

Demand more than a correlational result; Demand a causal structure and the testing of that structure in context of the real world data.

Contact Information

Robert W. Stoddard
Principal Researcher
Software Engineering Institute
Carnegie Mellon University
4500 Fifth Avenue (Office 3110)
Pittsburgh, PA 15213

rws@sei.cmu.edu
724-263-7113 cell
www.sei.cmu.edu

Backup Slides

Early Software Causal Studies

Controlling Size: Only 2 of 4 code size measures appear causal on effort and quality

Controlling Complexity: Only 1 of 3 factors appears causal on performance and quality

Controlling Architecture Violations: Only 1 of 4 violation factors appears causal on quality

Controlling Team Performance: Only 1 of 20+ factors appears causal on quality and cost

Causal search may provide useful feedback:

1) *Presence of causal links*
2) *Absence of causal links*

Controlling Size

	Direct cause of total effort							
Algorithm	**ESLOC**	**FP**	**SNAP**	**CFP**	**CPLX**	**ACAP**	**PCAP**	**DOCU**
Stepwise Regression (Adj R^2=.84)				Yes			Yes	
PC				Yes			Yes	
PC-Stable				Yes				
FGES				Yes			Yes	
FASK		Yes	Yes	Yes				Yes

Size estimators compared:

Equivalent SLOC (ESLOC), IFPUG Function Points (FP), IFPUG Software Non-functional Assessment Process (SNAP), COSMIC Function Points (CFP)

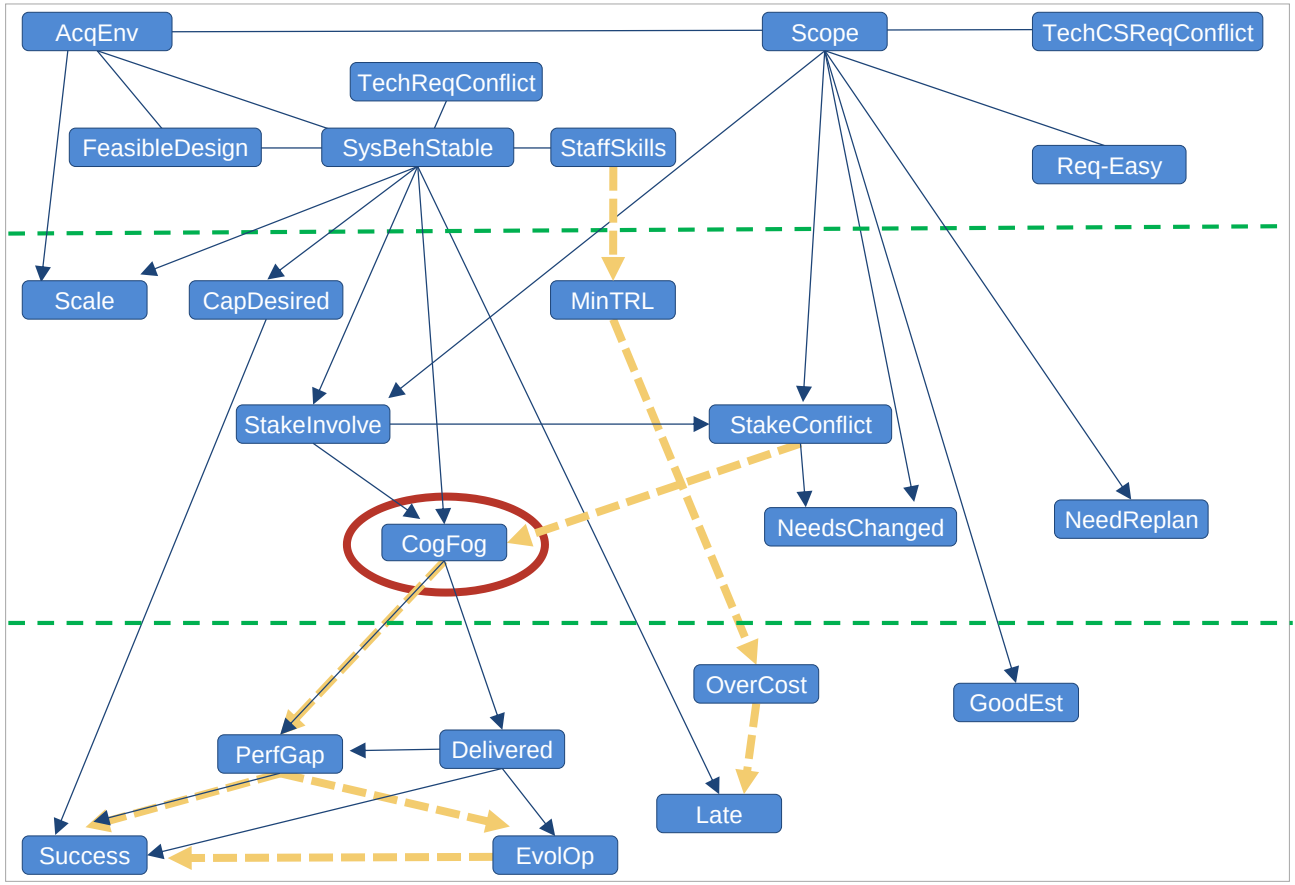
Other variables: Applications Experience, Platform Experience, Use of Software Tools, Personnel Continuity, Documentation Match to Needs (DOCU), Analyst Capability (ACAP), Programmer Capability (PCAP), Product Complexity (CPLX)

From "Integrated Causal Model for Software Cost Prediction & Control (SCOPE)", Feb, 2019 SEI Research Review

Controlling Complexity

In 2012, Sheard found that 3 of 40 measures of complexity correlated highly with 7 measures of success:
1) difficult requirements
2) stakeholder relationship
3) cognitive fog

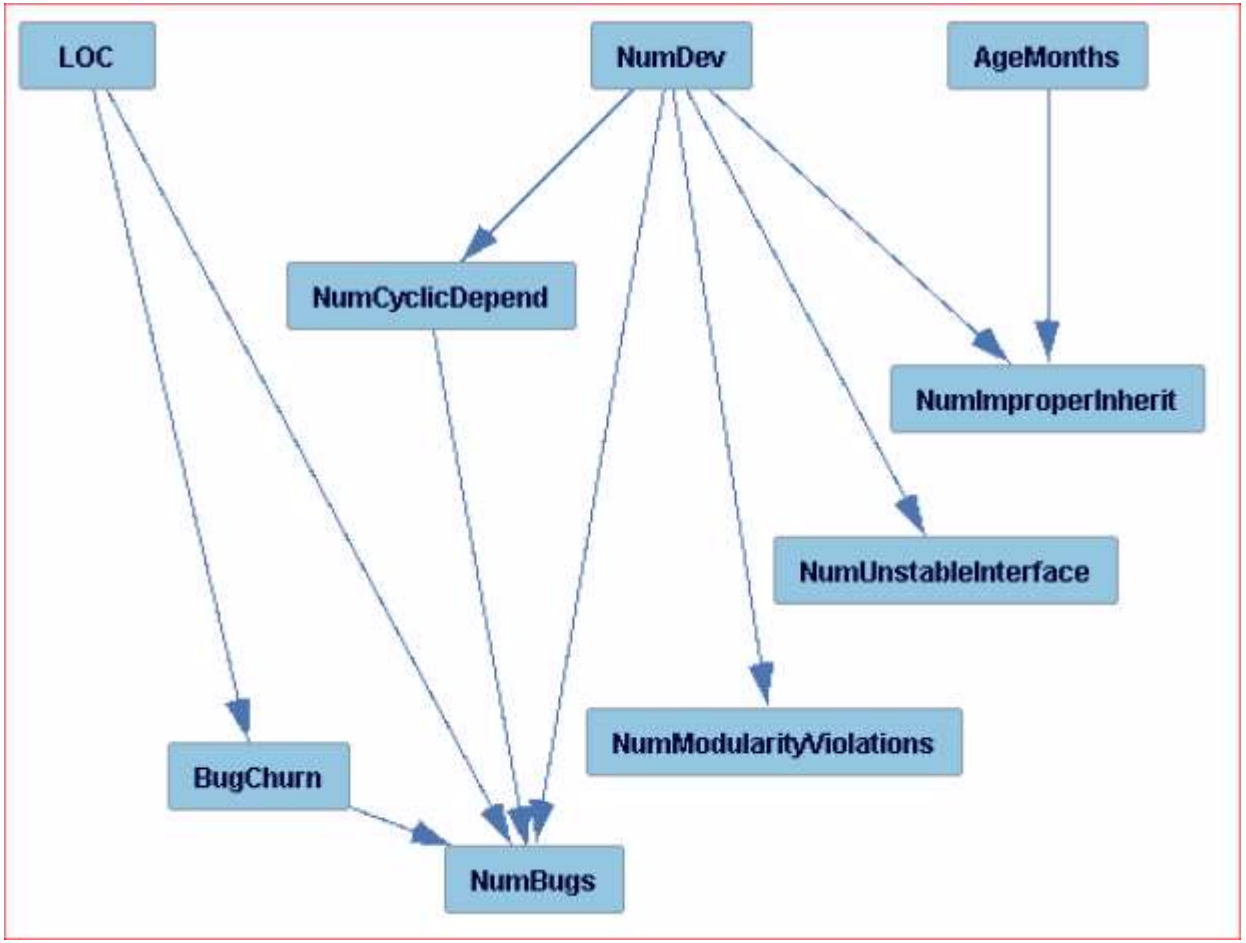
But causal learning found
- no evidence for 1,
- consistent evidence of 2 but only mediated through 3, and
- consistent evidence for 3



From “Causal Models for Software Cost Prediction and Control (SCOPE)”, 2019 SEI Annual Research Review

Controlling Architecture Violations

Of 4 architecture pattern violations previously showing significant correlation to resulting software quality (NumBugs), only NumCyclicDependency demonstrated a causal effect. The other 3 were deemed spurious correlation.



From "Synthesis of Causal Discovery and Machine Learning – Questions Posed", 2018 BayesiaLab Conference

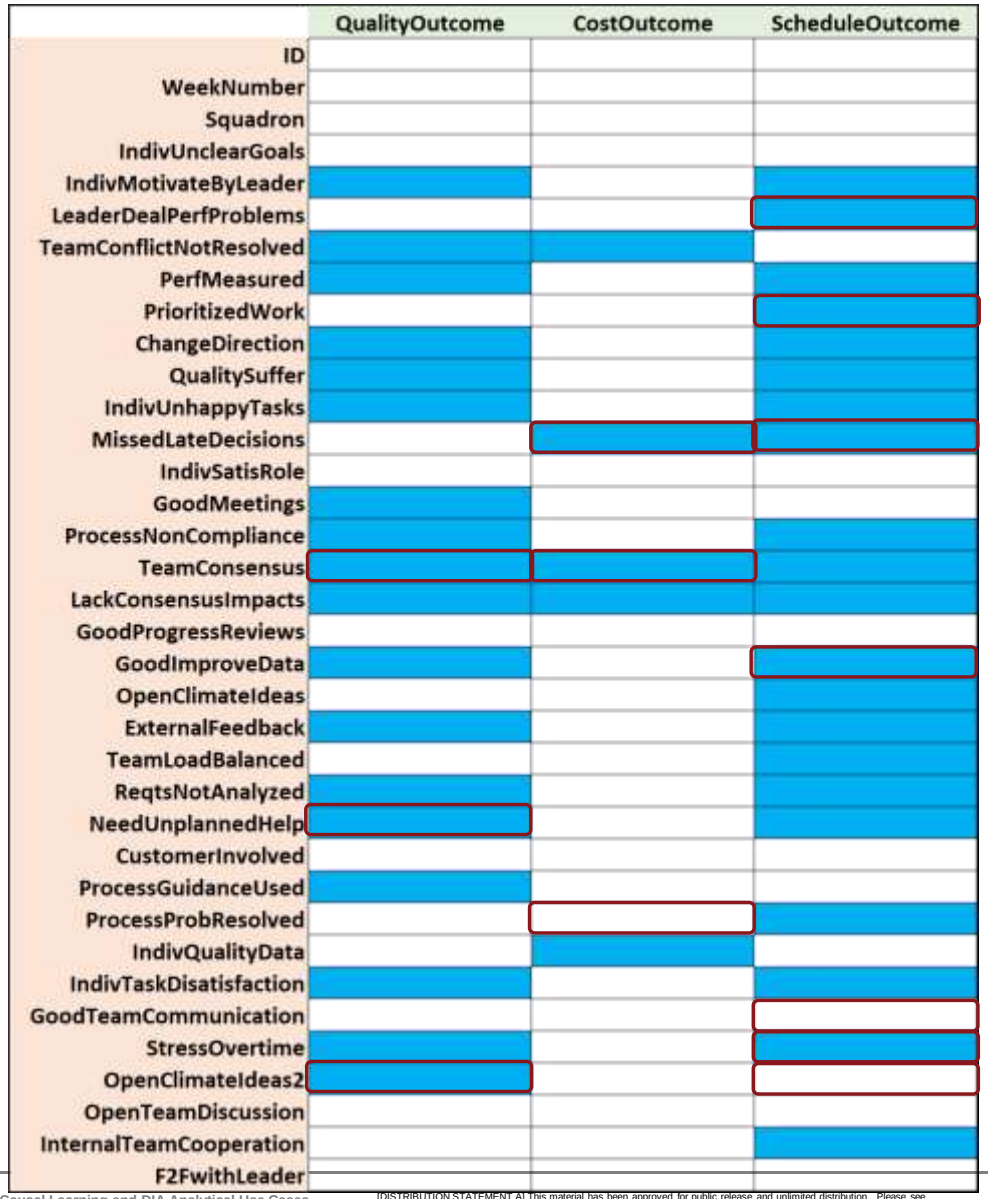
Controlling Team Performance - 1

Traditional correlation results:

Correlation measures used included Kendall tau-b, Kendall tau-c, Gamma and Spearman's. All were in agreement using the 0.05 cutoff for significance (blue highlighted cells).

Ordinal logistic regression:

using 0.05 alpha for significance and McFadden pseudo Rsquare indicated significant factors (red borders).



Controlling Team Performance - 2

Traditional statistical correlation depicted:

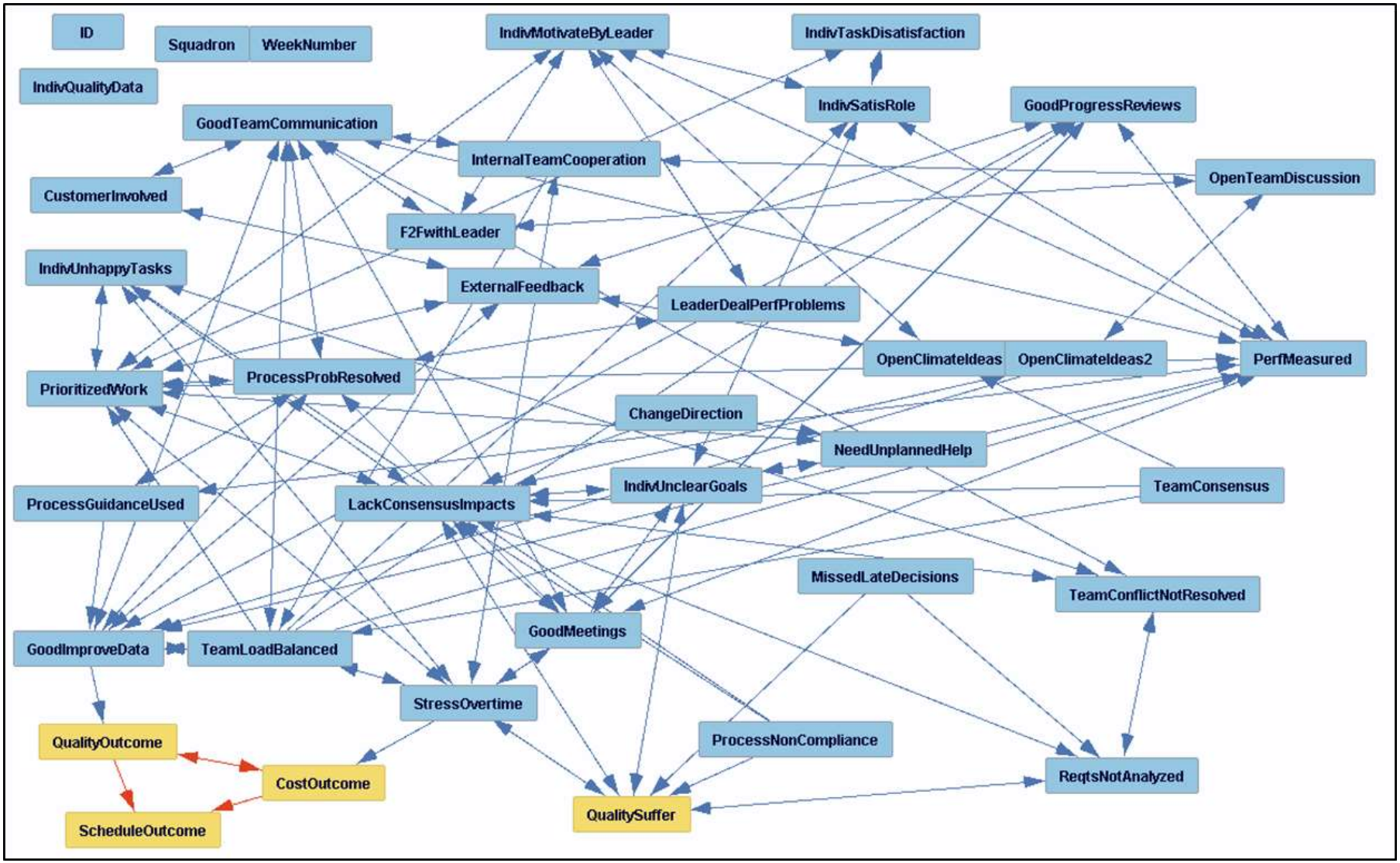
- 18 factors highly correlated with Quality [2 confirmed with Logistic Regression]
- 5 factors highly correlated with Cost, and
- 21 factors highly correlated with Schedule,

Causal search discovered evidence of:

- 1 factor (GoodImproveData) causing Quality performance,
- 1 factor (StressOvertime) causing Cost performance,
- No independent factors appear to cause Schedule performance.

From “Why Does Software Cost So Much? Towards a Causal Model: Initial Results Looking at Project Datasets”, ICEAA Conference, 2018

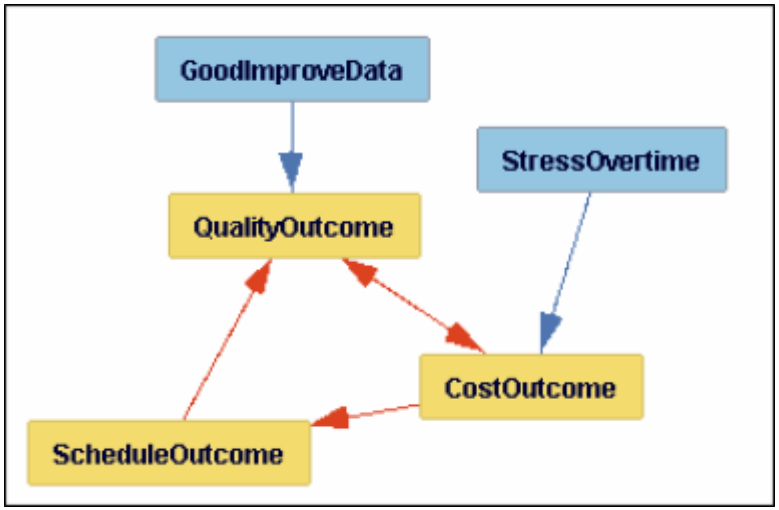
Controlling Team Performance - 3



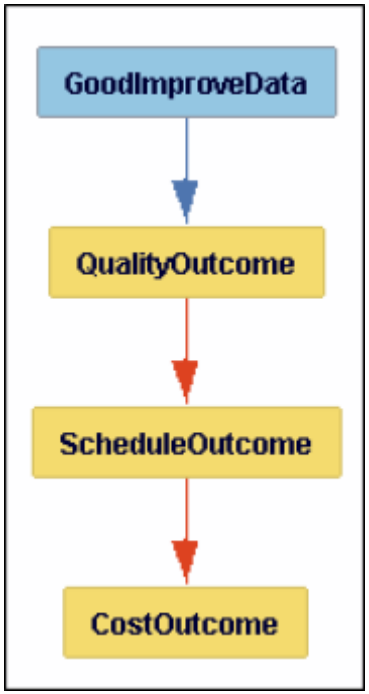
Controlling Team Performance - 4

We show two Markov blankets for the set of three outcomes:

1) using PC-Stable search



2) using FGES search



Although the two algorithms differ on the directed edges among the three outcomes, there is agreement on GoodImproveData causing QualityOutcome. PC-Stable adds StressOvertime as a cause of CostOutcome.

From “Why Does Software Cost So Much? Towards a Causal Model: Initial Results Looking at Project Datasets”, ICEAA Conference, 2018

Do Architecture Pattern Violations Cause Vulnerabilities?