Software Solutions Symposium 2017

March 20–23, 2017

Why Does Software Cost So Much? Toward a Causal Model

Robert Stoddard, SEI

Mike Konrad, SEI

Bill Nichols, SEI

David Danks, CMU

Kun Zhang, CMU

Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8721-05-C-0003 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the United States Department of Defense.

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.

DM-0004586

Agenda

✓ Introduction

Why do We Care about Causal Modeling?

What is Causal Learning and Modeling?

Case Studies exemplifying Causal Learning

Practical Tooling

Conclusions

Questions

FY17 Project Approach and Objectives

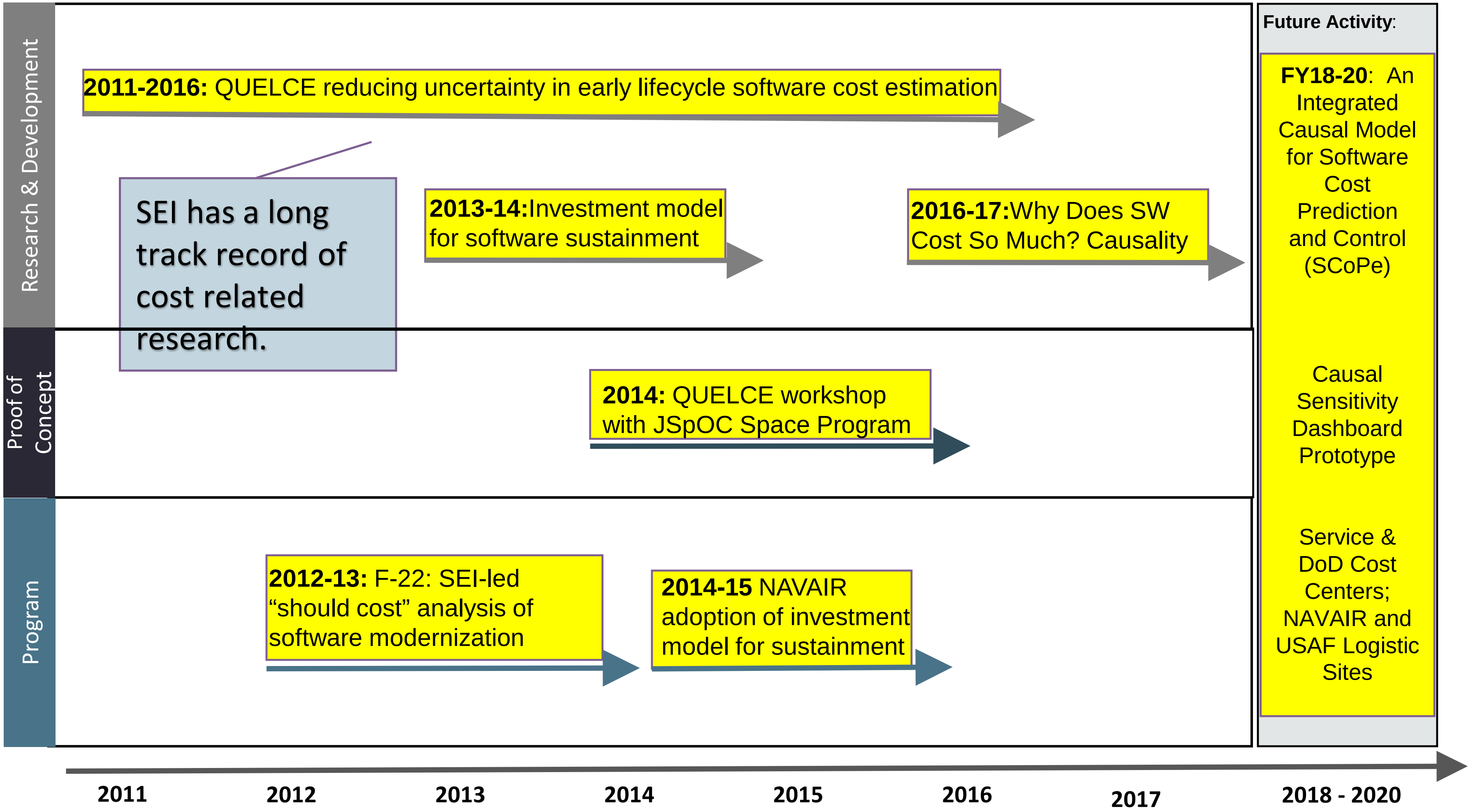
To answer: *"Why does software cost so much?"*

- We will quantify causal drivers of Software Size and Effort that are more applicable across programs
- We will reconfirm, within the software cost domain, that causal models are more accurate and efficient than traditional models

This project will help:

- Improve contract incentives for software intensive programs
- Increase competition using effective criteria related to software cost
- Inform "could/should cost" analysis and price negotiations
- Enhance program control of software cost throughout the development and sustainment lifecycles

SEI Long-Term Initiative Focused on Cost



Research & Development

2011-2016: QUELCE reducing uncertainty in early lifecycle software cost estimation

SEI has a long track record of cost related research.

2013-14: Investment model for software sustainment

2016-17: Why Does SW Cost So Much? Causality

Proof of Concept

2014: QUELCE workshop with JSpOC Space Program

Program

2012-13: F-22: SEI-led "should cost" analysis of software modernization

2014-15 NAVAIR adoption of investment model for sustainment

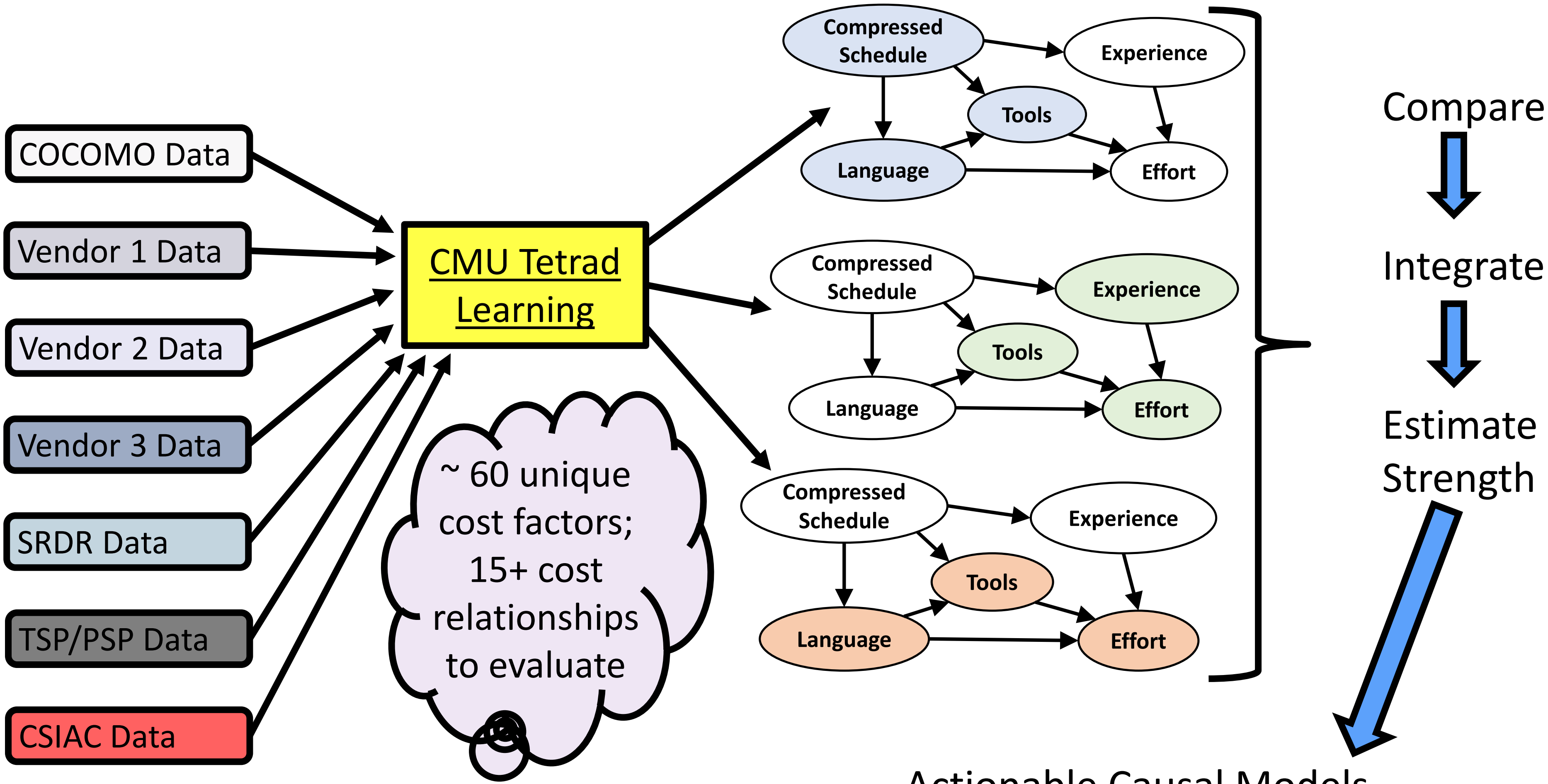
Future Activity:

FY18-20: An Integrated Causal Model for Software Cost Prediction and Control (SCoPe)

Causal Cost Sensitivity Dashboard Prototype

Service & DoD Cost Centers; NAVAIR and USAF Logistic Sites

Method, Approach, Validity



COCOMO Data

Vendor 1 Data

Vendor 2 Data

Vendor 3 Data

SRDR Data

TSP/PSP Data

CSIAC Data

CMU Tetrad Learning

~ 60 unique cost factors; 15+ cost relationships to evaluate

Compare

Integrate

Estimate Strength

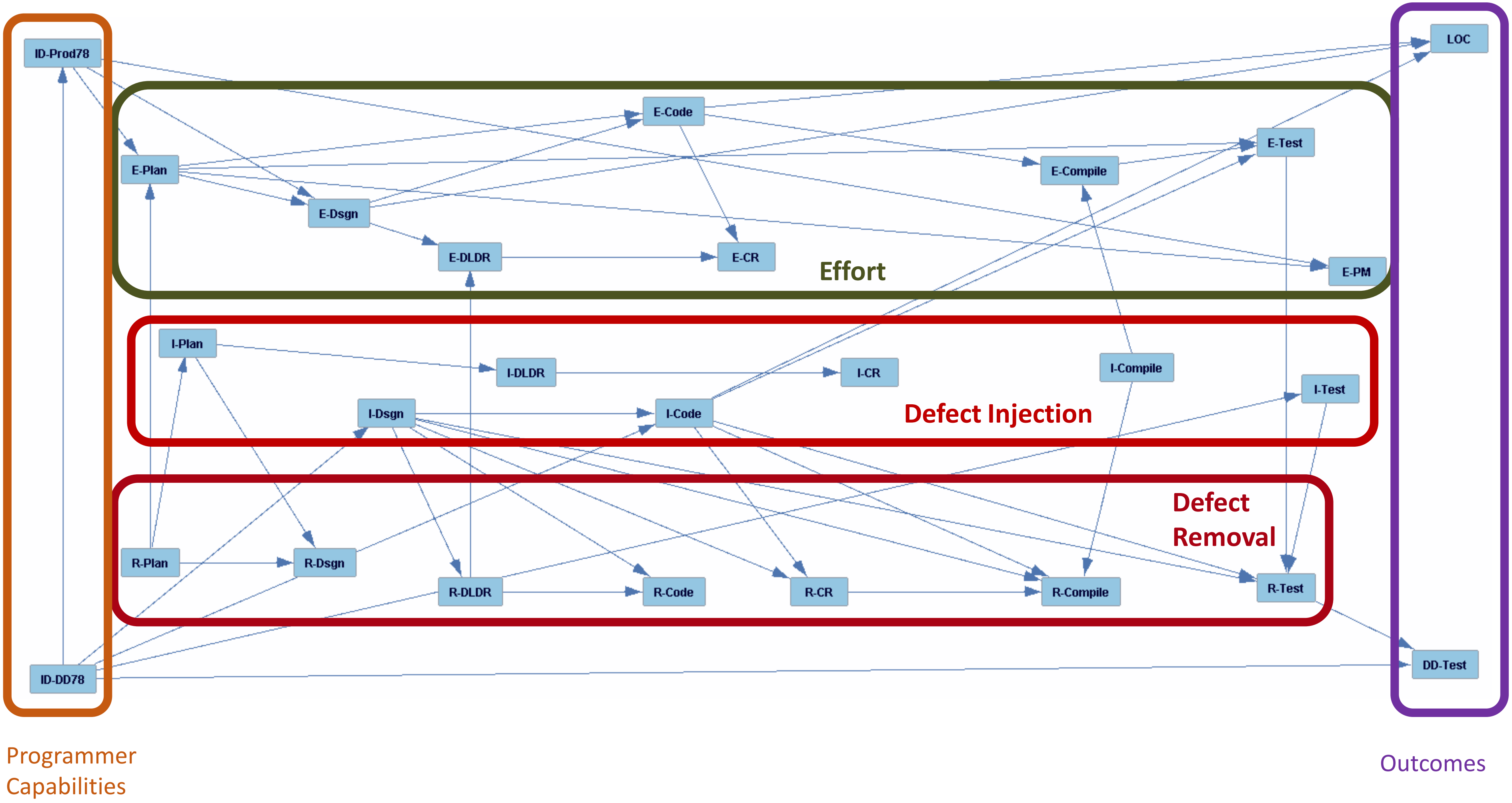
Actionable Causal Models

Agile Cost = f(factor1, factor2, factor3)

Incremental Cost = g(factor1, factor4, factor5)

Waterfall Cost = h(factor4, factor6, factor7)

Initial Results: PSP Program 9 (975 programmers)



Agenda

Introduction

✓ Why do We Care about Causal Modeling?

What is Causal Learning and Modeling?

Case Studies exemplifying Causal Learning

Practical Tooling

Conclusions

Questions

Why Do We Care about Causal Modeling?

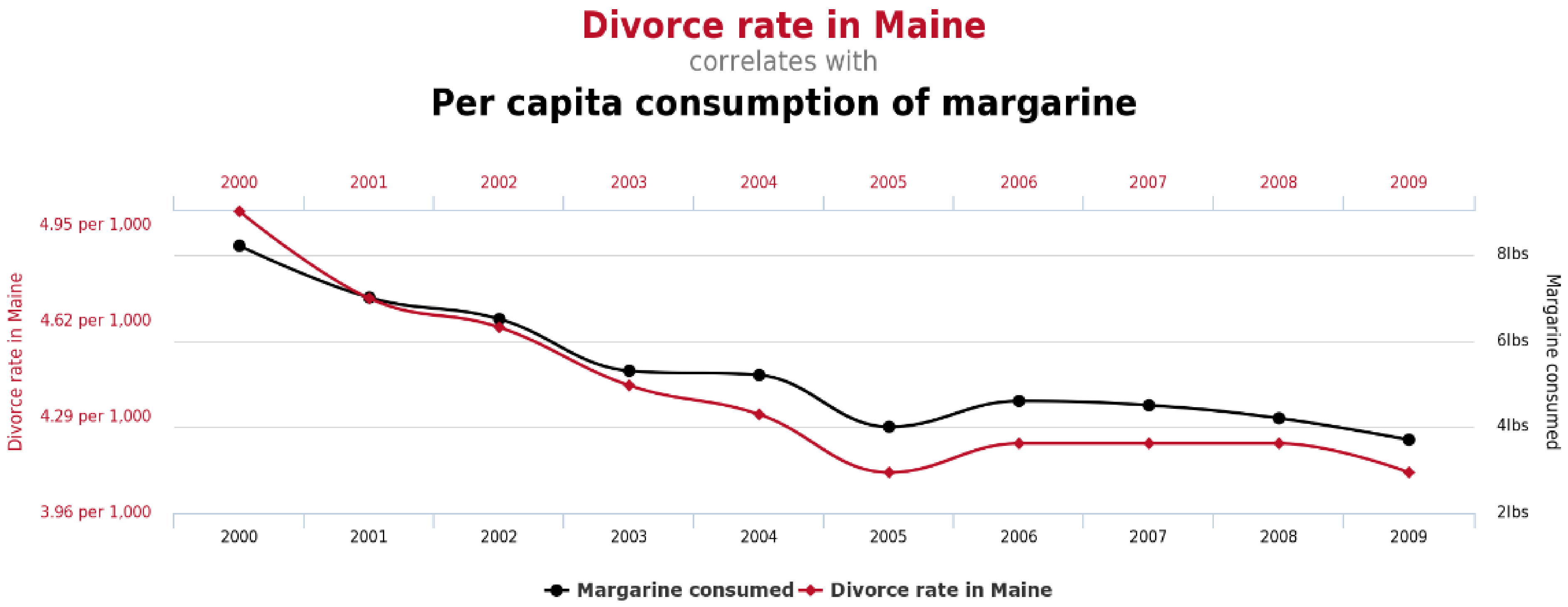
If we want to proactively control outcomes, it would be safer if we knew our independent factors actually caused the outcomes

To do this, we must move beyond correlation & regression to causation

We want to establish causation without the expense and challenge of conducting a controlled experiment

Establishing causation with observational data remains a vital need and a key technical challenge!

Correlation is Insufficient



From TylerVigen.com. Retrieved March 14, 2017. Available under a Creative Commons Attribution NonCommercial Share Alike 4.0 International License

Correlation and Regression Can Mislead!

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

Traditional Regression Result

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)

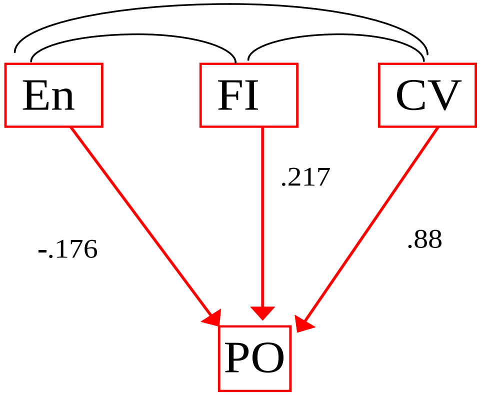
CV is lack of civil liberties

EN is energy consumption per capita (economic development)

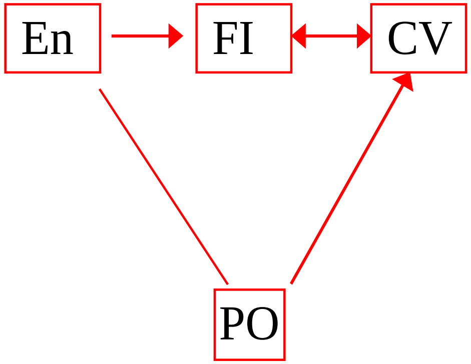
FI is level of foreign investment

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

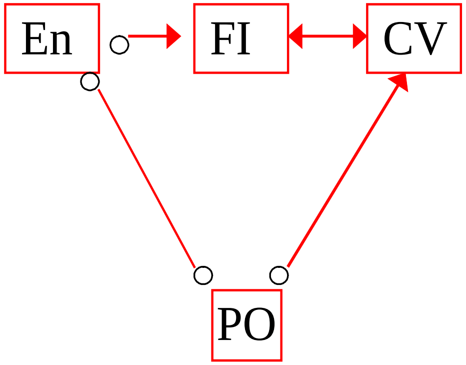
Causal Modeling refutes Regression Result!



Regression

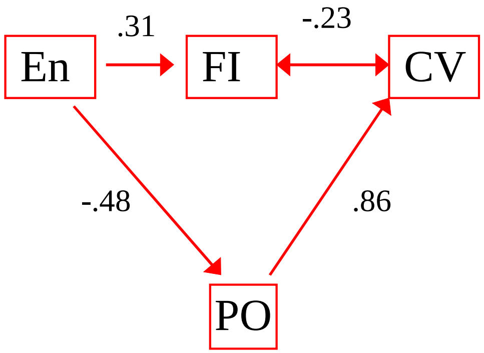


Tetrad - PC



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

Regression Cannot be Trusted without a DAG!

Just as correlation may be fooled by spurious association, so can regression!

Before jumping into regression, we need to discover causal relationships or formulate our hypotheses in the form of a Directed Acyclic Graph (DAG)

Then we need to apply causal identification and determine which paths are causal and which are spurious.

Lastly, we must block spurious paths and develop a resulting regression with the causal factors.

Remember, the same regression equation mapped to two different Directed Acyclic Graphs can have very different suitability!

Agenda

Introduction

Why do We Care about Causal Modeling?

✓ What is Causal Learning and Modeling?

Case Studies exemplifying Causal Learning

Practical Tooling

Conclusions

Questions

[Distribution Statement A. This material has been approved for public release and unlimited distribution]

Causality via Matching and Experimentation

Prior to 1935, causal conclusions were made by matching data of different conditions and comparing the outcomes

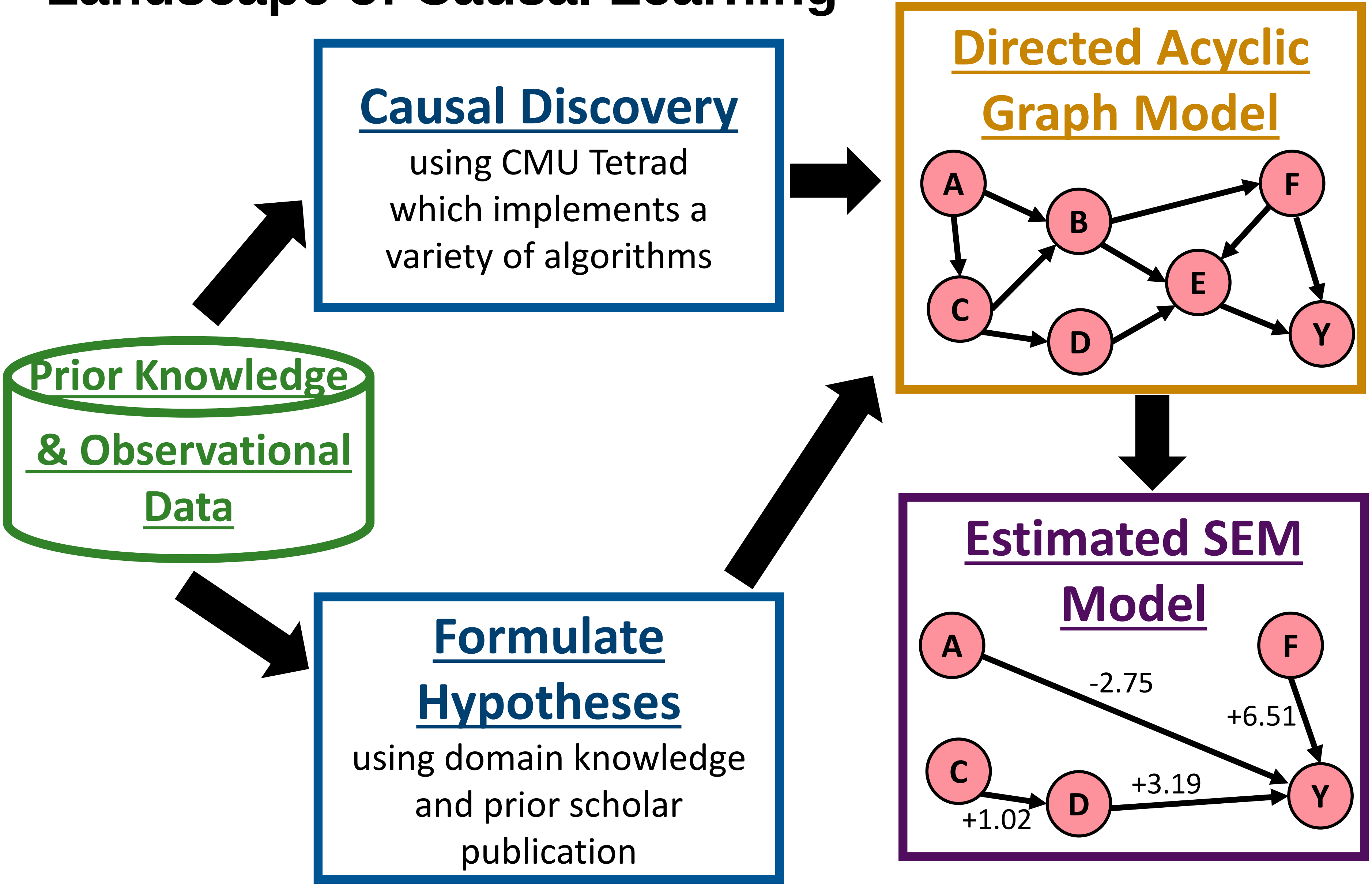
This was deemed too expensive, slow and often prohibitive

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.

So, we now have causal methods other than controlled experiments!

Landscape of Causal Learning



Prior Knowledge & Observational Data

Causal Discovery

using CMU Tetrad which implements a variety of algorithms

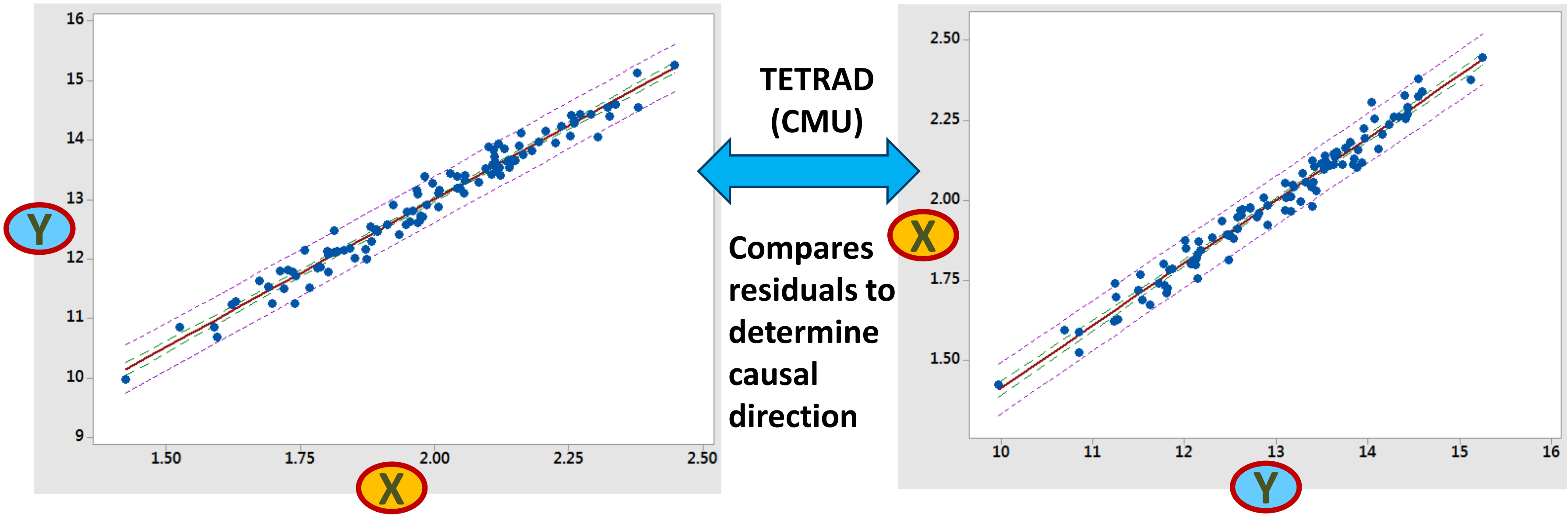
Formulate Hypotheses

using domain knowledge and prior scholar publication

Directed Acyclic Graph Model

Estimated SEM Model

Causal Discovery using Data & Prior Knowledge



TETRAD (CMU)

Compares residuals to determine causal direction

Does the correlation reflect causal association?

Can we ascertain the direction of causation?

Do we just have spurious correlation?

- If so, what other known factors are the real cause(s)?
- Do we have an unknown causal factor(s) to continue hunting for?

Causal Estimation Using Counterfactuals

An example counterfactual question is:

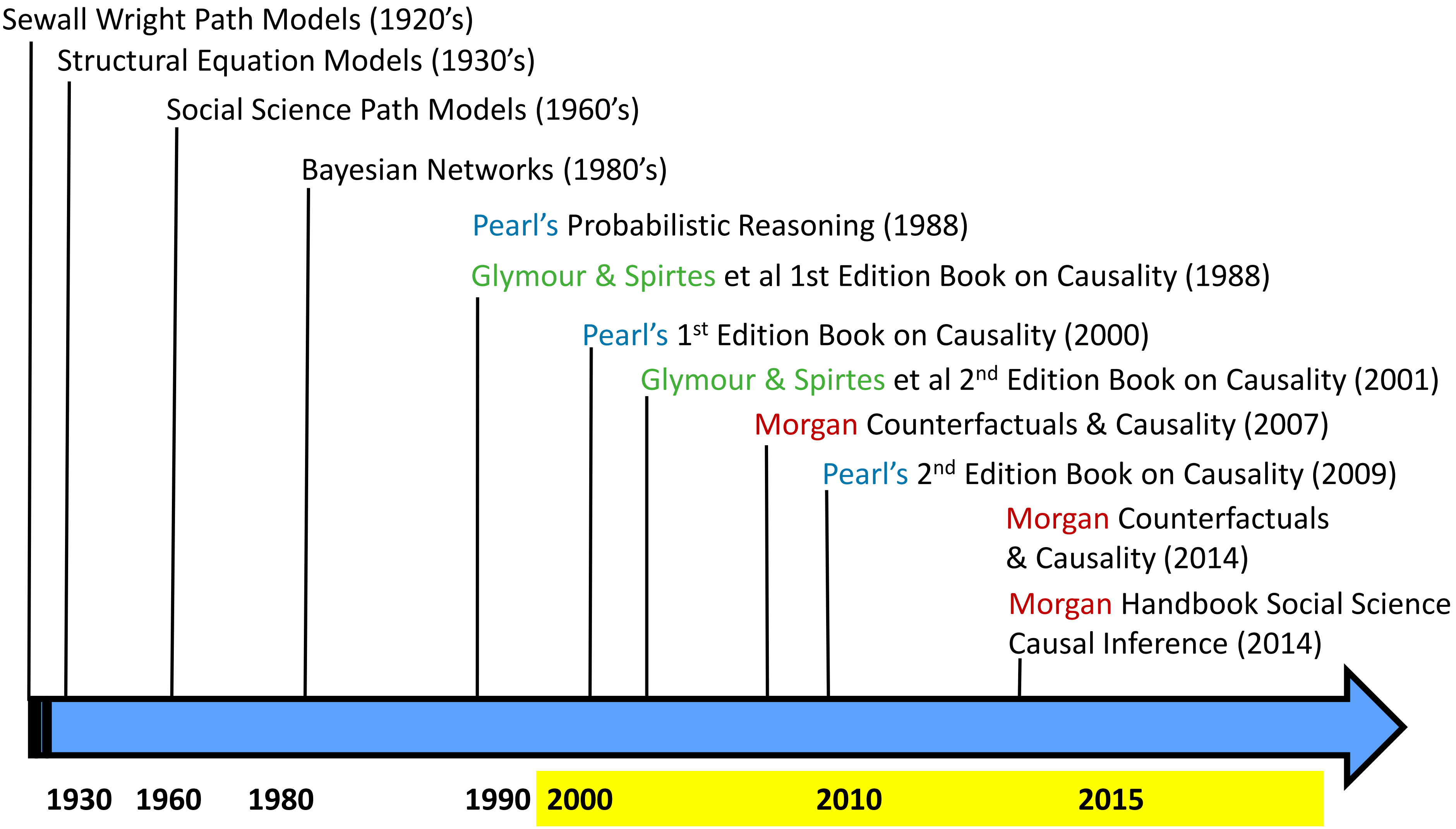
“Had event A alternatively occurred, what would be the potential outcome?”

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)

Amazing Progress in the Past 17 Years!



Sewall Wright Path Models (1920’s)

Structural Equation Models (1930’s)

Social Science Path Models (1960’s)

Bayesian Networks (1980’s)

Pearl’s Probabilistic Reasoning (1988)

Glymour & Spirtes et al 1st Edition Book on Causality (1988)

Pearl’s 1st Edition Book on Causality (2000)

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)

Morgan Handbook Social Science
Causal Inference (2014)

Agenda

Introduction

Why do We Care about Causal Modeling?

What is Causal Learning and Modeling?

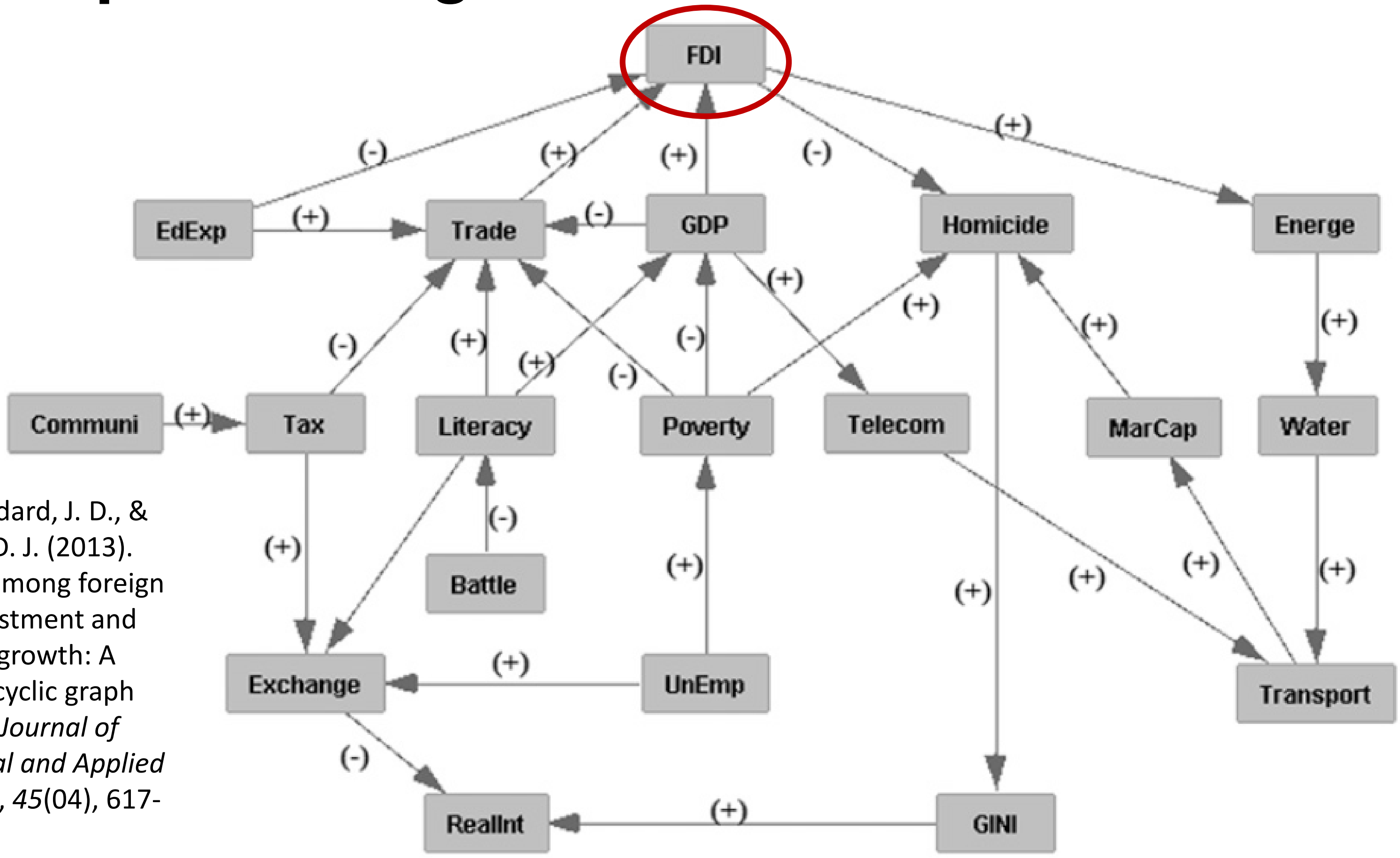
✓ Case Studies exemplifying Causal Learning

Practical Tooling

Conclusions

Questions

Example: Foreign Investment Benefits



Li, Y., Woodard, J. D., & Leatham, D. J. (2013). Causality among foreign direct investment and economic growth: A directed acyclic graph approach. *Journal of Agricultural and Applied Economics*, *45*(04), 617-637.

Causal pattern of 19 FDI related variables, estimated with Tetrad IV for 61 Developing Countries; GDP, trade, and educational expenditure are the three direct causal variables to FDI. GDP and trade have a positive impact on FDI, whereas educational expenditure affects FDI negatively. GDP is a typical indicator used to measure a country's economic health and thus higher GDP signals opportunity and attracts FDI.

Example: Energy Economics

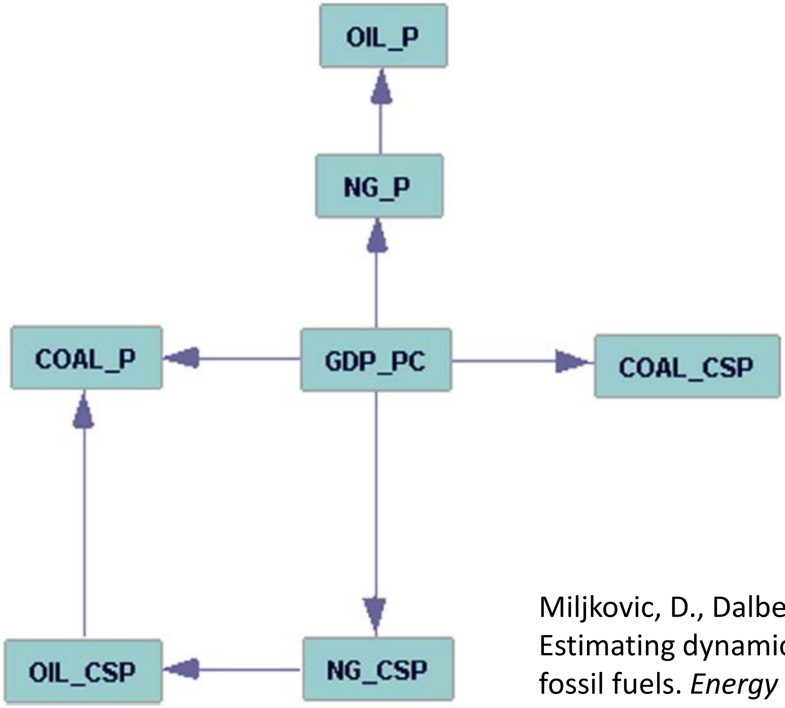


Fig. 2. PC graph.

Miljkovic, D., Dalbec, N., & Zhang, L. (2016). Estimating dynamics of US demand for major fossil fuels. *Energy Economics*, *55*, 284-291.

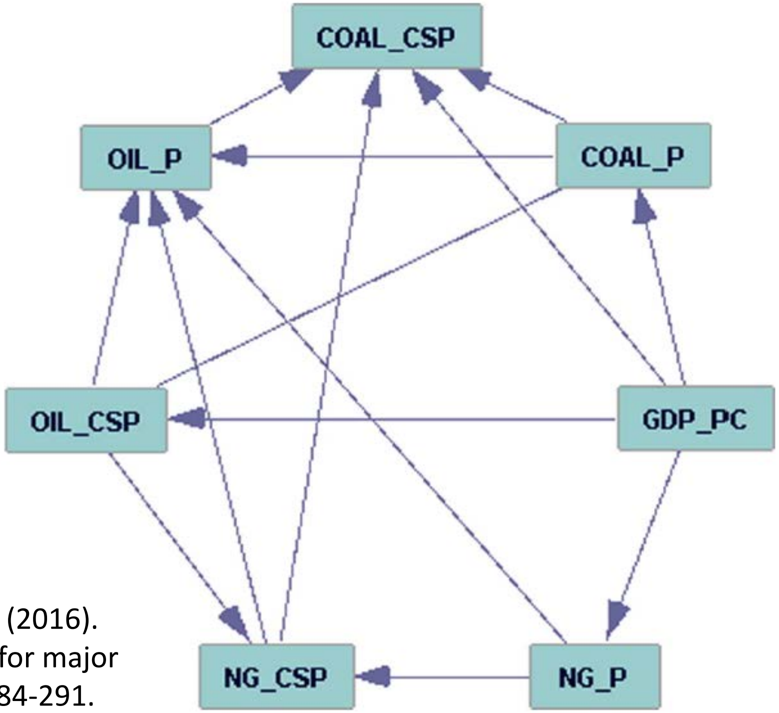


Fig. 3. GES graph.

US governments …favor …different energy modes and fossil fuels, due to different interest group pressures. Yet this lack of substitutability among main fossil fuels points to potentially very distortionary impacts of such policies since no obvious replacements emerge in energy use when one source of energy becomes overregulated. Hence …advisable to link, from policy standpoint, the sources of energy with the users based on source..

Example: Energy/Agriculture Linkage in Context of Renewable Fuel Policy

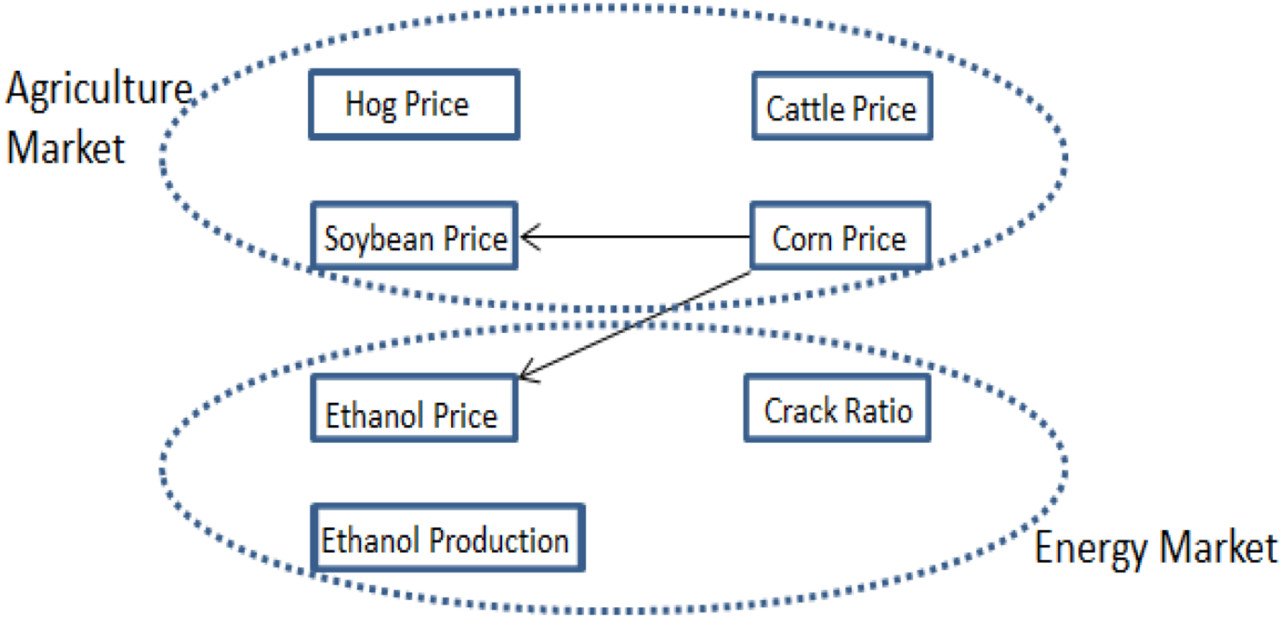


Figure 4: Directed acyclic graphs at 0.7 prune factor using LiNGAM.

“This study shows a significant linkage between agriculture and energy market through corn. The results …suggest that at contemporaneous time corn causes soybeans price and ethanol price. Renewable fuel standards requirements and rise in ethanol for demand affect the agriculture market. Corn price and soybeans price together count for about 12 and 18 percent of change in cattle and hog prices.”

Examples All Around!!!

Do you hear of preliminary research outcomes with surprising, if not provocative results? (e.g. journals, NPR)

Have you heard of research results that can not be replicated?

Do you hear results that seem to be correlated but no one is sure of the causal direction?

Do you hear conclusions about what is going on but you suspect an unmeasured, third factor is really the culprit?

Drawing correct inferences in such situations requires rigorous causal modeling; testing hypotheses in context of a Probabilistic Graphical Model!

Agenda

Introduction

Why do We Care about Causal Modeling?

What is Causal Learning and Modeling?

Case Studies exemplifying Causal Learning

✓ Practical Tooling

Conclusions

Questions

Tetrad is a CMU Open Source Product

May be found at:

http://www.phil.cmu.edu/tetrad/

Code located at github:

https://github.com/cmu-phil/tetrad

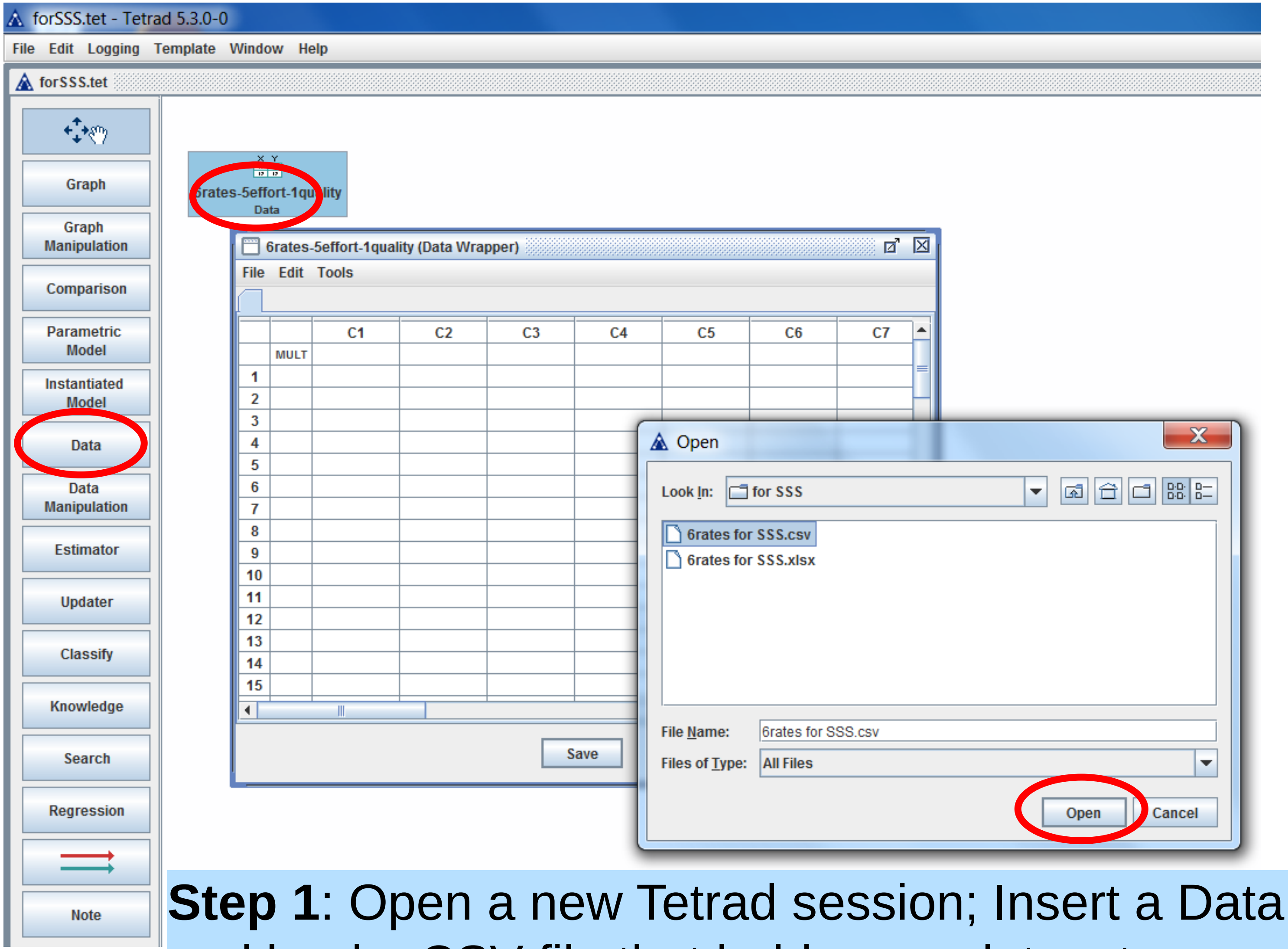
Video Tutorials from 2016 Summer Short Course

http://www.ccd.pitt.edu/training/presentation-videos/

User Manual

Adobe Acrobat Document

Using Tetrad for Causal Discovery and Estimation -1



Look In: for SSS

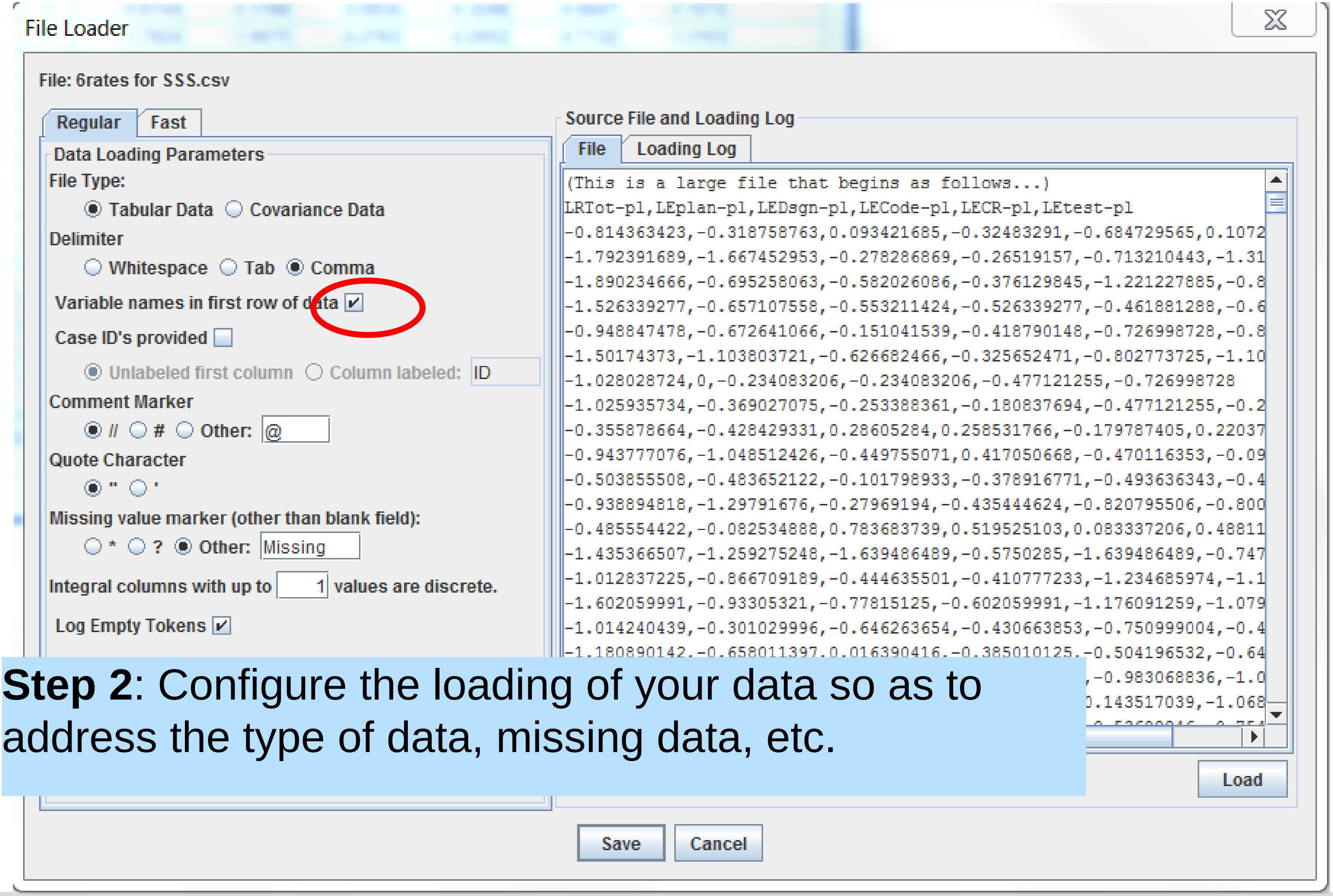
File Name: 6rates for SSS.csv

Files of Type: All Files

Step 1: Open a new Tetrad session; Insert a Data box to open and load a CSV file that holds your dataset.

[Distribution Statement A. This material has been approved for public release and unlimited distribution]

Using Tetrad for Causal Discovery and Estimation -2



File Loader

File: 6rates for SSS.csv

Regular | Fast

Data Loading Parameters

File Type:

● Tabular Data ○ Covariance Data

Delimiter

○ Whitespace ○ Tab ● Comma

Variable names in first row of data ☑

Case ID's provided □

● Unlabeled first column ○ Column labeled: ID

Comment Marker

● // ○ # ○ Other: @

Quote Character

● " ○ '

Missing value marker (other than blank field):

○ * ○ ? ● Other: Missing

Integral columns with up to 1 values are discrete.

Log Empty Tokens ☑

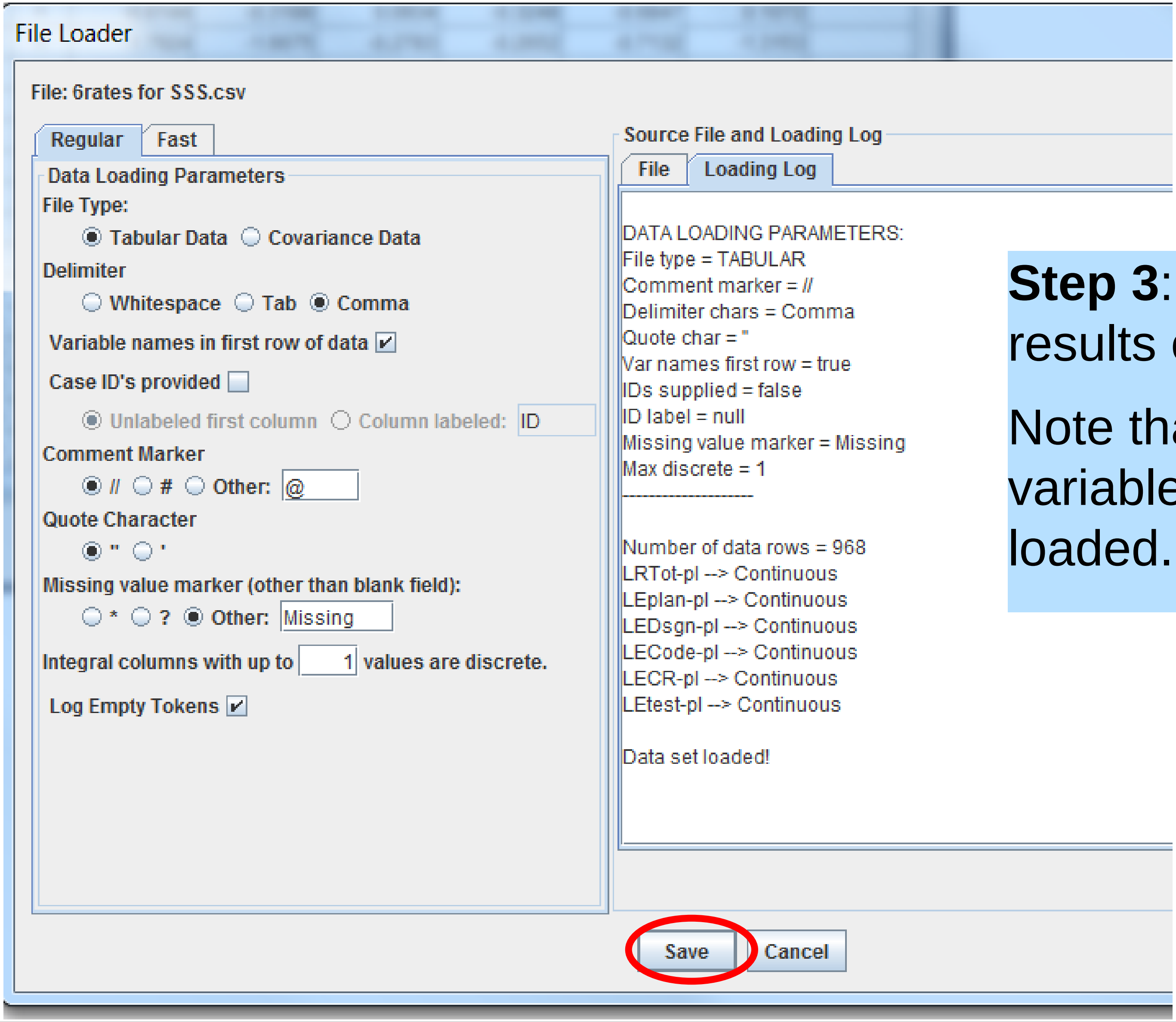
Source File and Loading Log

File | Loading Log

```
(This is a large file that begins as follows...)
LRTot-pl,LEplan-pl,LEDsgn-pl,LECode-pl,LECR-pl,LEtest-pl
-0.814363423,-0.318758763,0.093421685,-0.32483291,-0.684729565,0.1072
-1.792391689,-1.667452953,-0.278286869,-0.26519157,-0.713210443,-1.31
-1.890234666,-0.695258063,-0.582026086,-0.376129845,-1.221227885,-0.8
-1.526339277,-0.657107558,-0.553211424,-0.526339277,-0.461881288,-0.6
-0.948847478,-0.672641066,-0.151041539,-0.418790148,-0.726998728,-0.8
-1.50174373,-1.103803721,-0.626682466,-0.325652471,-0.802773725,-1.10
-1.028028724,0,-0.234083206,-0.234083206,-0.477121255,-0.726998728
-1.025935734,-0.369027075,-0.253388361,-0.180837694,-0.477121255,-0.2
-0.355878664,-0.428429331,0.28605284,0.258531766,-0.179787405,0.22037
-0.943777076,-1.048512426,-0.449755071,0.417050668,-0.470116353,-0.09
-0.503855508,-0.483652122,-0.101798933,-0.378916771,-0.49363643,-0.4
-0.938894818,-1.29791676,-0.27969194,-0.435444624,-0.820795506,-0.800
-0.485554422,-0.082534888,0.783683739,0.519525103,0.083337206,0.48811
-1.435366507,-1.259275248,-1.639486489,-0.5750285,-1.639486489,-0.747
-1.012837225,-0.866709189,-0.444635501,-0.410777233,-1.234685974,-1.1
-1.602059991,-0.93305321,-0.77815125,-0.602059991,-1.176091259,-1.079
-1.014240439,-0.301029996,-0.646263654,-0.430663853,-0.750999004,-0.4
-1.180890142,-0.658011397,0.016390416,-0.385010125,-0.504196532,-0.64
                                                ,-0.983068836,-1.0
                                                .143517039,-1.068
```

Step 2: Configure the loading of your data so as to address the type of data, missing data, etc.

Using Tetrad for Causal Discovery and Estimation -3



File: 6rates for SSS.csv

Data Loading Parameters

File Type:

● Tabular Data ○ Covariance Data

Delimiter

○ Whitespace ○ Tab ● Comma

Variable names in first row of data ☑

Case ID's provided □

● Unlabeled first column ○ Column labeled: ID

Comment Marker

● // ○ # ○ Other: @

Quote Character

● " ○ '

Missing value marker (other than blank field):

○ * ○ ? ● Other: Missing

Integral columns with up to 1 values are discrete.

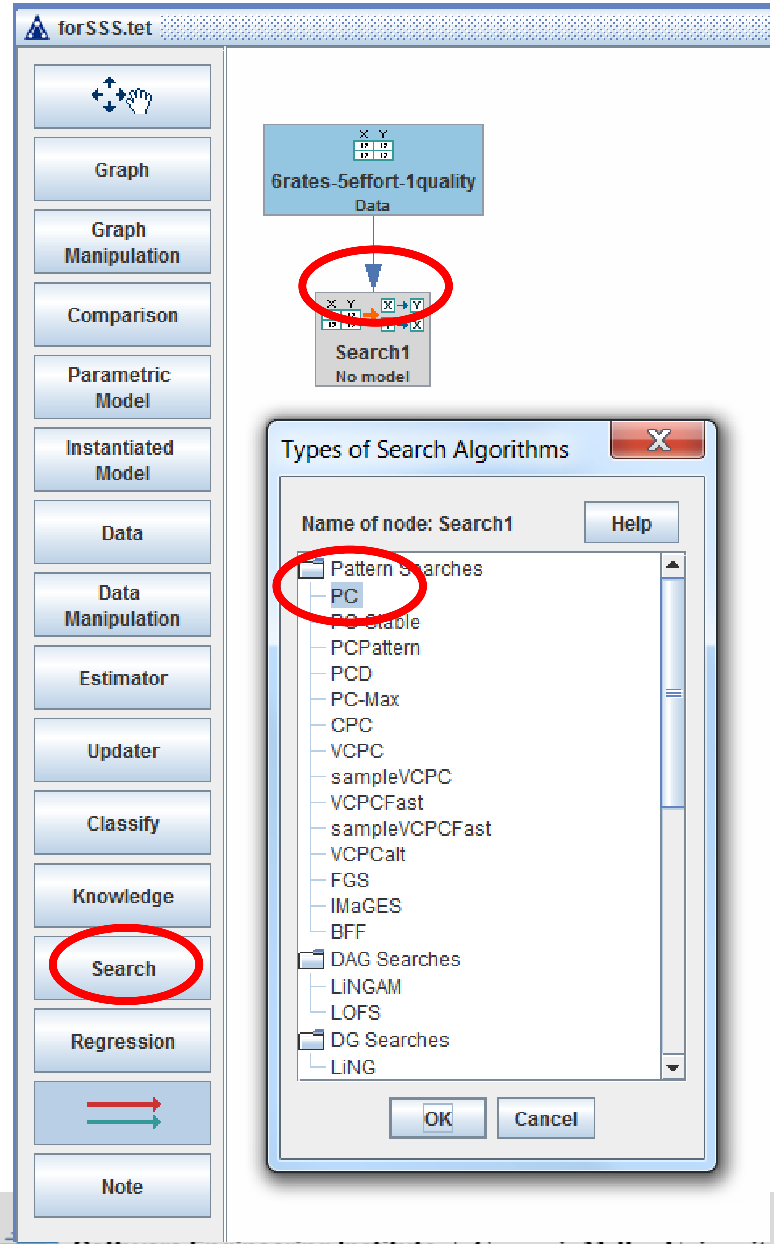
Log Empty Tokens ☑

```
DATA LOADING PARAMETERS:
File type = TABULAR
Comment marker = //
Delimiter chars = Comma
Quote char = "
Var names first row = true
IDs supplied = false
ID label = null
Missing value marker = Missing
Max discrete = 1
--------------------

Number of data rows = 968
LRTot-pl --> Continuous
LEplan-pl --> Continuous
LEDsgn-pl --> Continuous
LECode-pl --> Continuous
LECR-pl --> Continuous
LEtest-pl --> Continuous

Data set loaded!
```

Step 3: Check the results of the loading.

Note that 6 continuous variables have been loaded.

Using Tetrad for Causal Discovery and Estimation -4



Types of Search Algorithms

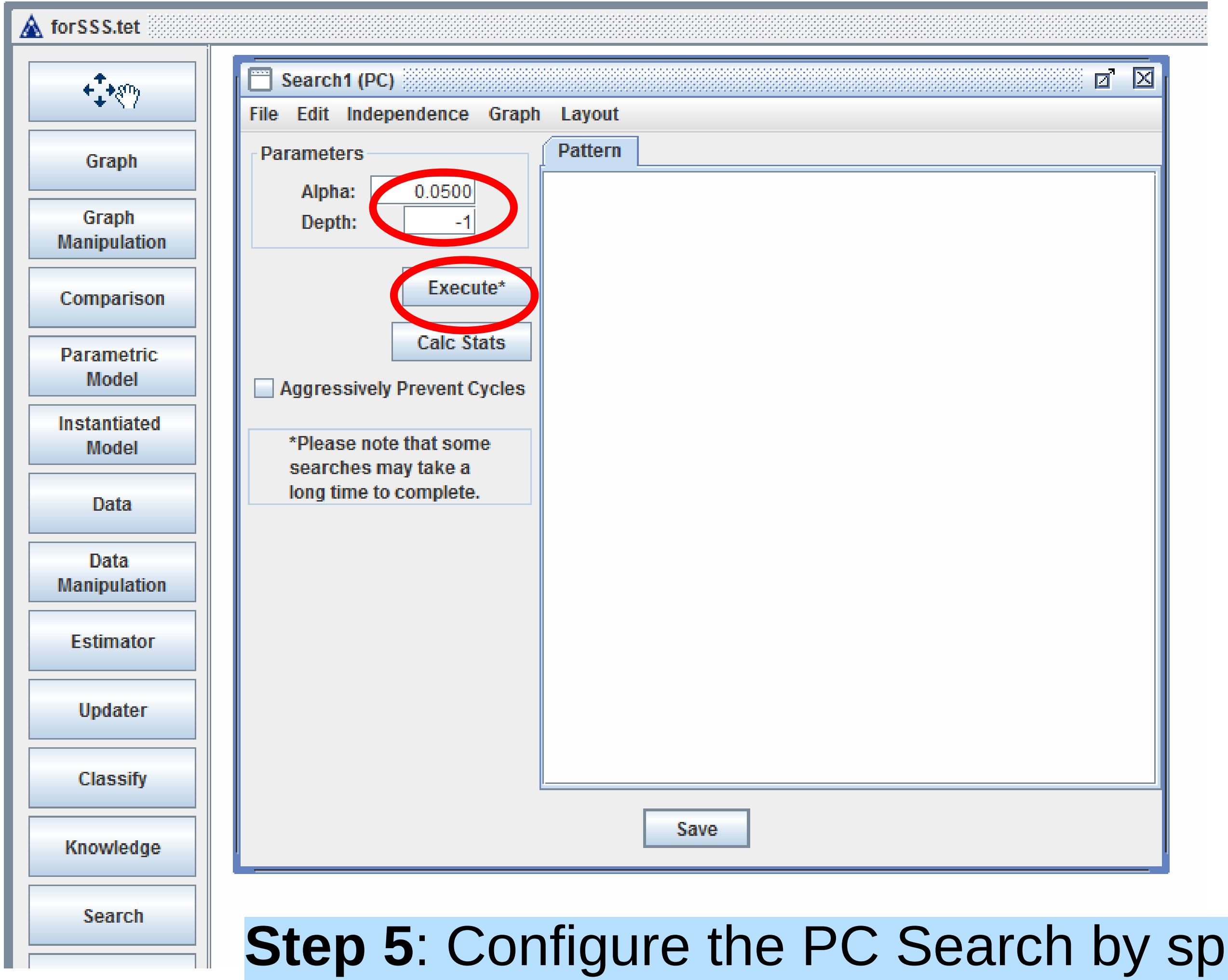
Name of node: Search1

Step 4: Initiate a causal search by inserting a Search box with a connection from the Data box.

Select from over a dozen different algorithms for causal discovery.

The PC algorithm, named for Peter Spirtes and Clark Glymour (CMU), its creators, is one of the most widely used and known causal search algorithms.

Using Tetrad for Causal Discovery and Estimation -5



Parameters

Alpha: 0.0500

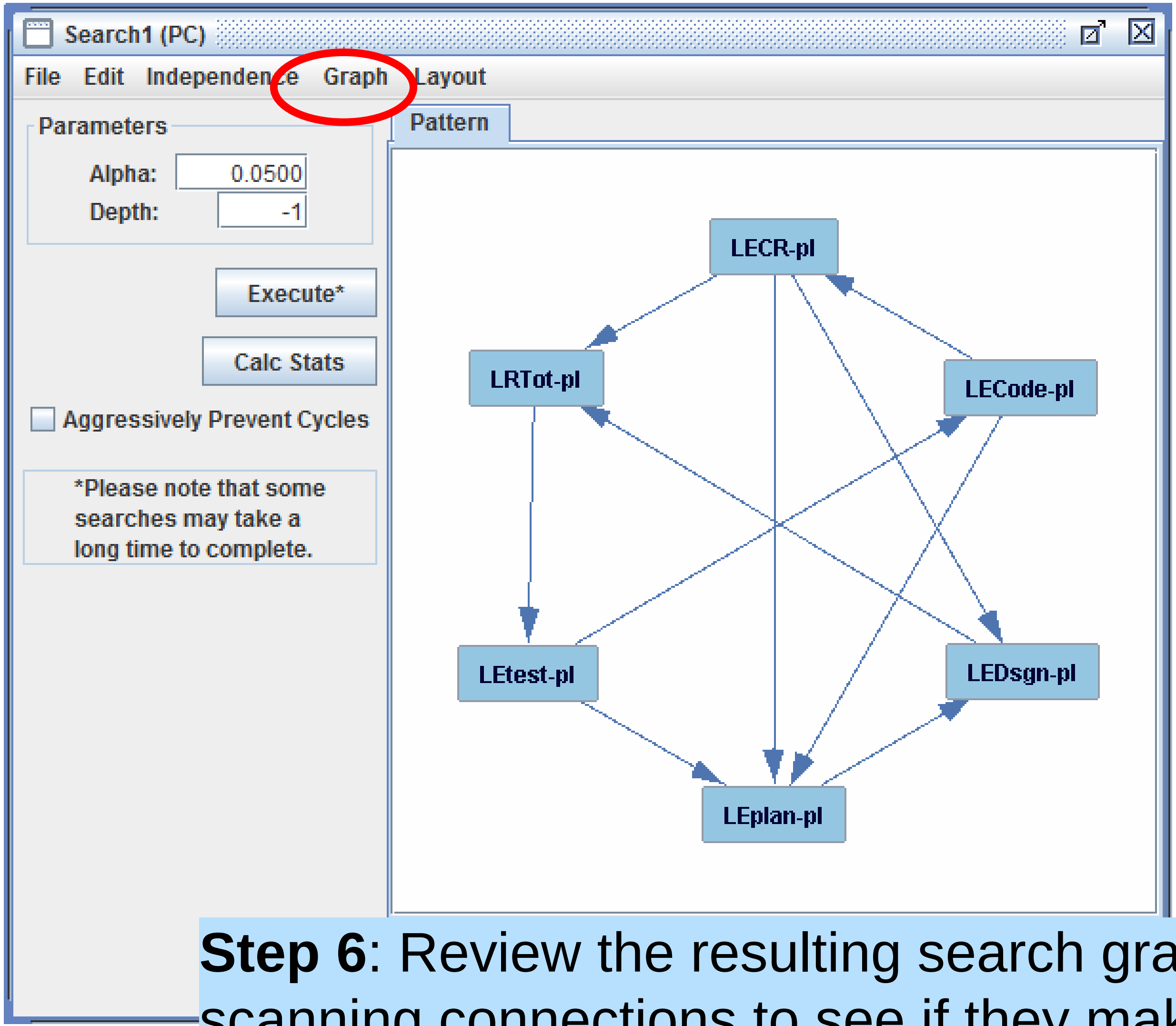
Depth: -1

□ Aggressively Prevent Cycles

*Please note that some searches may take a long time to complete.

Step 5: Configure the PC Search by specifying an appropriate Alpha and Depth (we’ll use the default settings .05 and -1), and then click on Execute.

Using Tetrad for Causal Discovery and Estimation -6



Parameters

Alpha: 0.0500

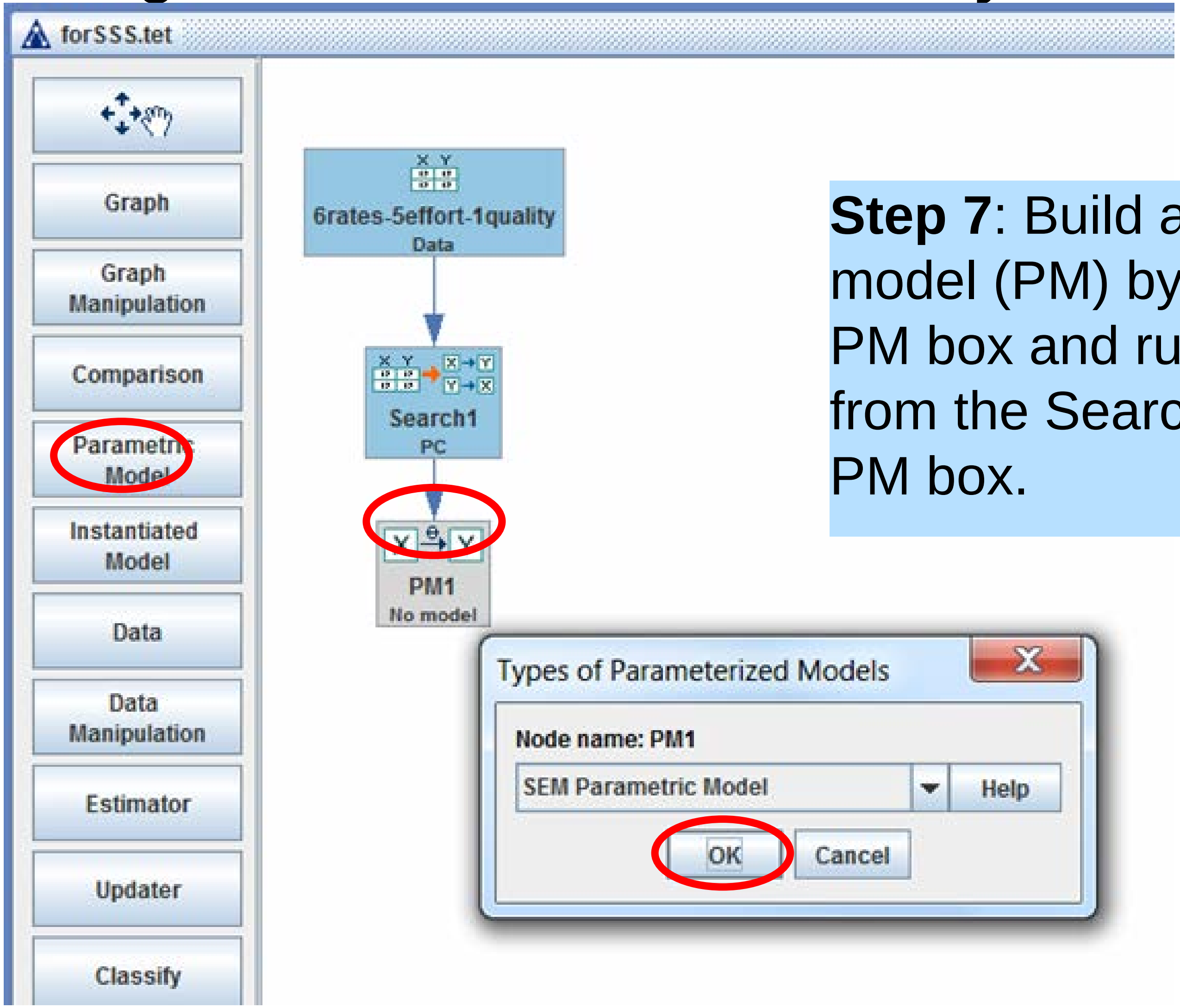
Depth: -1

□ Aggressively Prevent Cycles

*Please note that some searches may take a long time to complete.

Step 6: Review the resulting search graph by scanning connections to see if they make sense.

Using Tetrad for Causal Discovery and Estimation -7



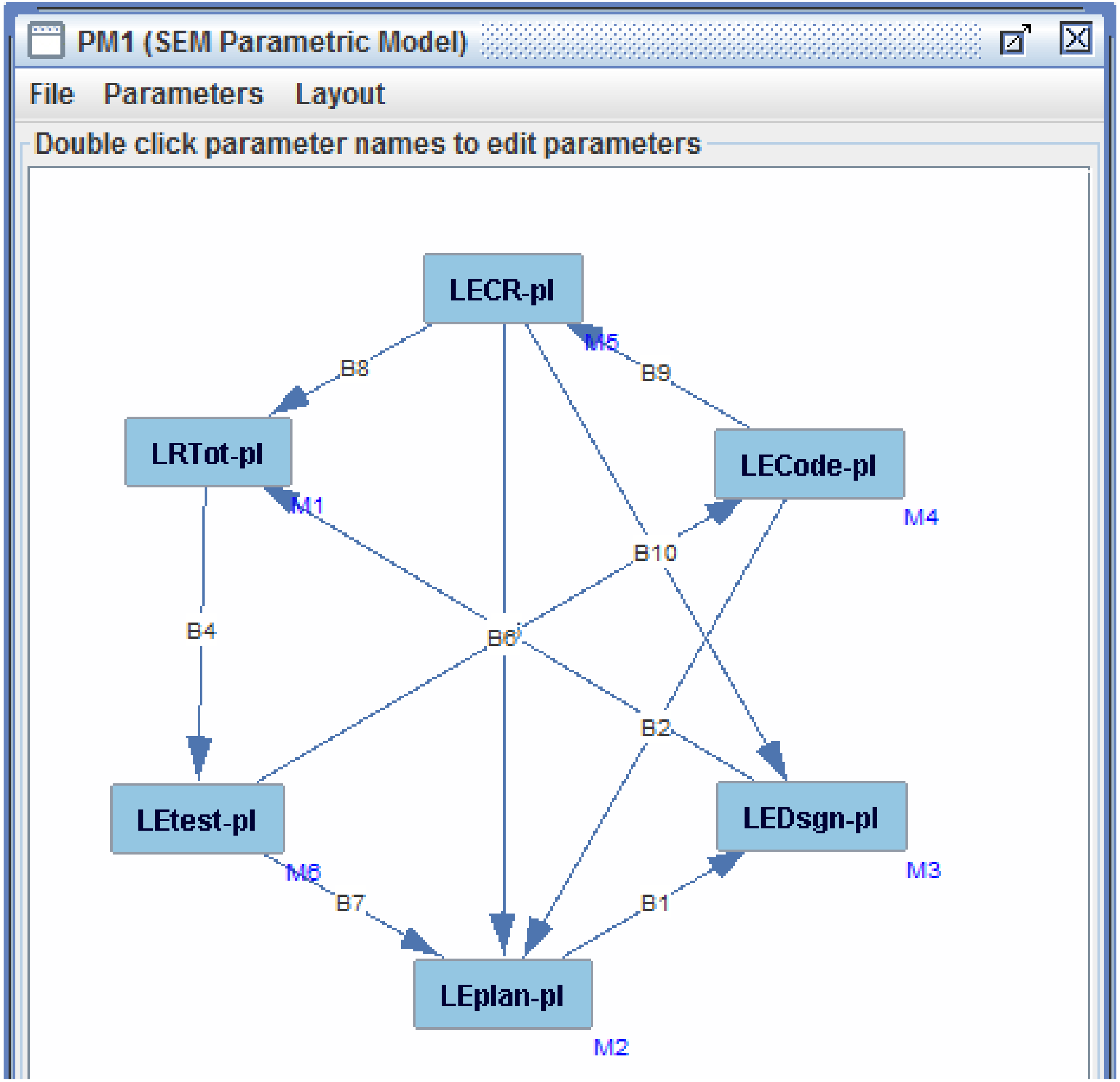
Types of Parameterized Models

Node name: PM1

SEM Parametric Model

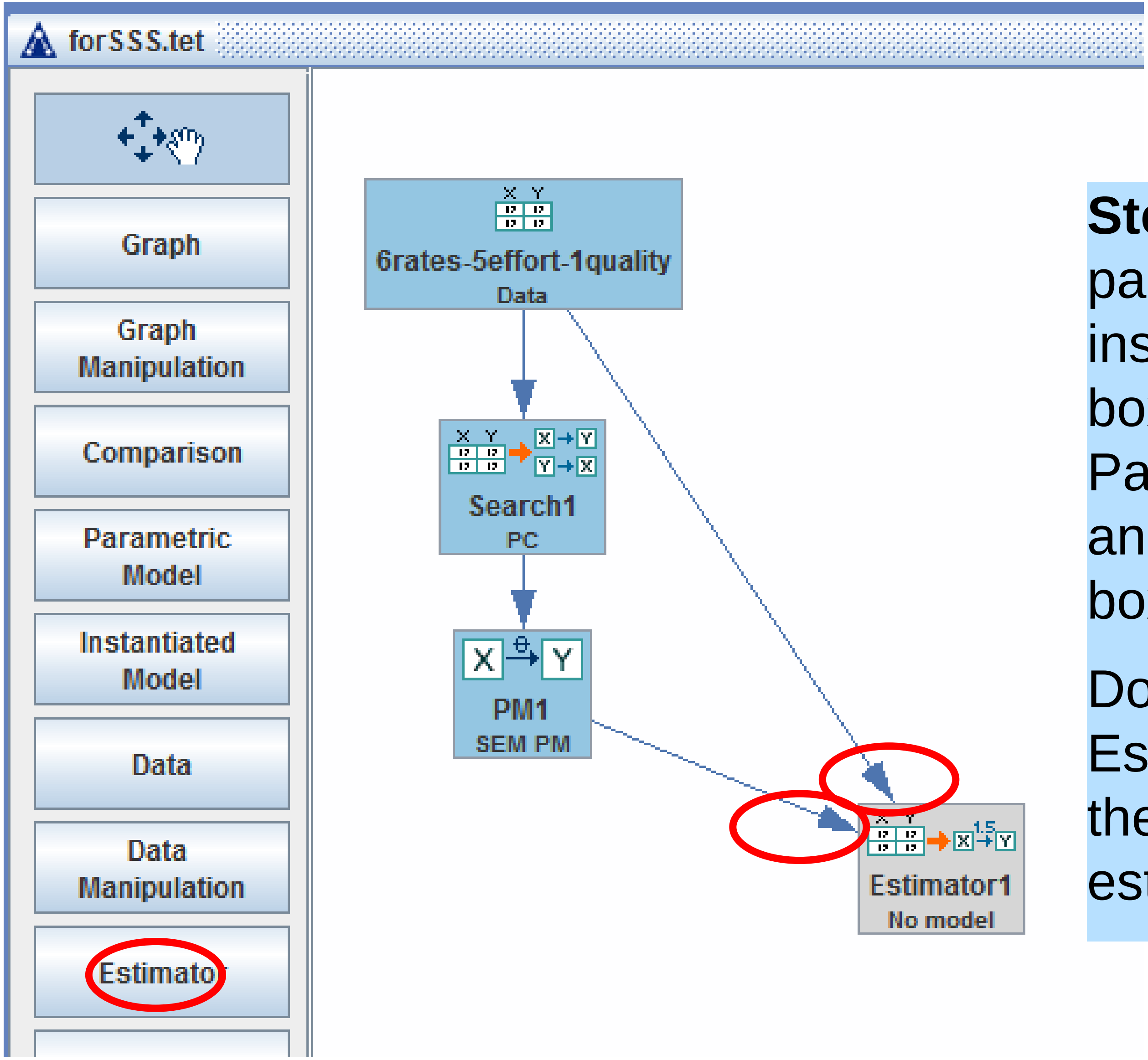
Step 7: Build a parametric model (PM) by inserting a PM box and running an arrow from the Search box into the PM box.

Using Tetrad for Causal Discovery and Estimation -8



Step 8: Note the numbered parameters of the parametric model, and click on Save.

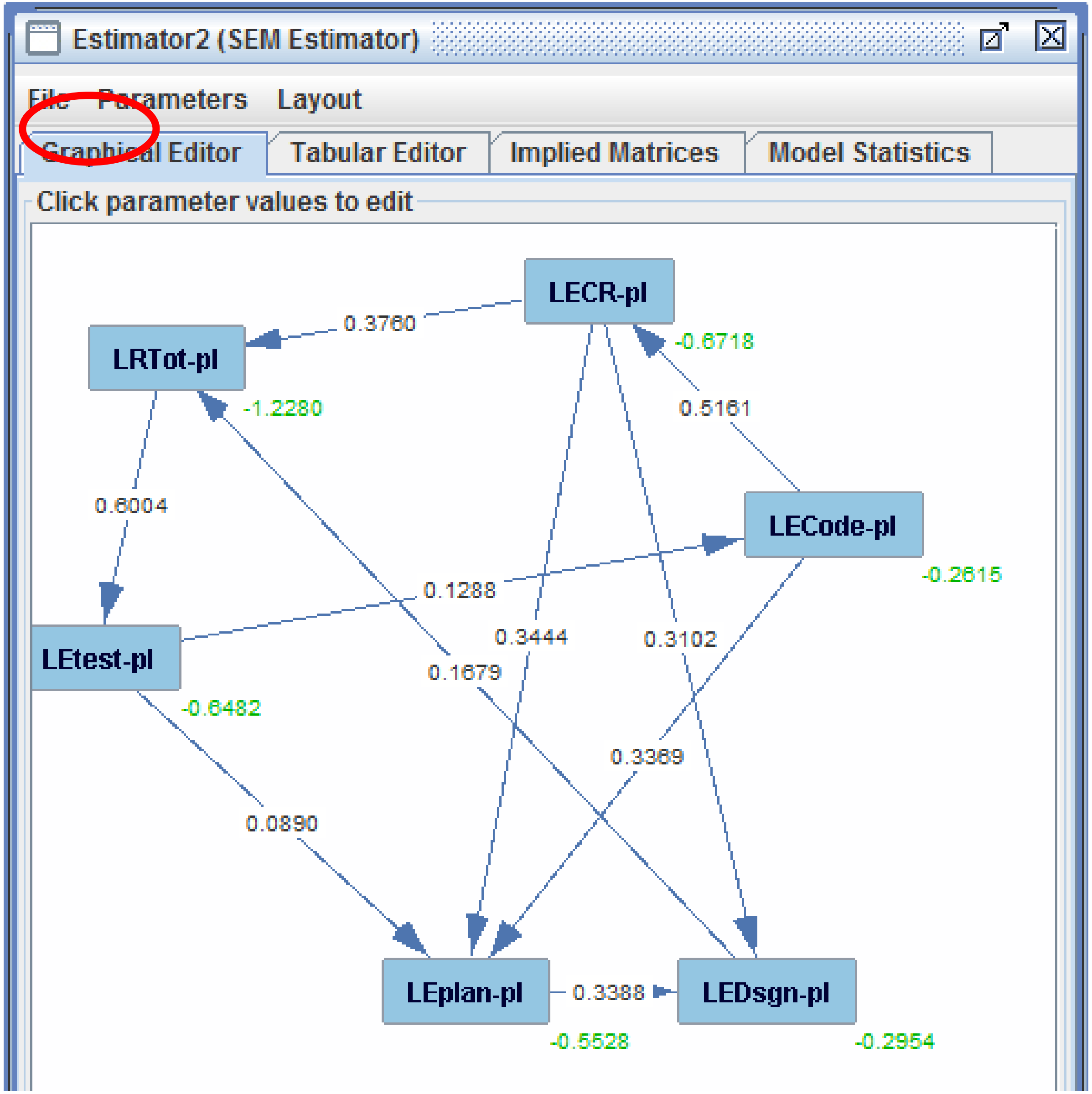
Using Tetrad for Causal Discovery and Estimation -9



Step 9: Instantiate the parametric model by inserting an Estimator box, hanging it off the Parametric Model box and the original Data box.

Double click the Estimator box to initiate the appropriate estimation algorithm.

Using Tetrad for Causal Discovery and Estimation -10



Click parameter values to edit

Step 10: Examine the resulting estimated model to assess model fit.

Agenda

Introduction

Why do We Care about Causal Modeling?

What is Causal Learning and Modeling?

Case Studies exemplifying Causal Learning

Practical Tooling

✓ Conclusions

Questions

Conclusions

Causal learning has come of age from both a theoretical and practical tooling standpoint

Causal learning may be performed on data whether it be derived from experimentation or passive observation

Causal models help separate true causes from spuriously-correlated factors

Causal models help identify when unknown causes may likely exist

Causal models have been shown to be more accurate than traditional models that contain spuriously-correlated factors

We seek to drive cost estimation to a new level with actionable, controllable causal models, and position the SEI as a preferred source for inquiries on “should cost” and “could cost” estimation.

Agenda

Introduction

Why do We Care about Causal Modeling?

What is Causal Learning and Modeling?

Case Studies exemplifying Causal Learning

Practical Tooling

Conclusions

✓ Questions

Contact Information

Points of Contact

SEMA Measurement Team

Robert Stoddard
rws@sei.cmu.edu

Mike Konrad
mdk@sei.cmu.edu

William Nichols
wrn@sei.cmu.edu

David Danks
ddanks@cmu.edu

Kun Zhang
kunz1@cmu.edu

U.S. Mail

Software Engineering Institute
Customer Relations
4500 Fifth Avenue
Pittsburgh, PA 15213-2612, USA

Web

www.sei.cmu.edu
www.sei.cmu.edu/contact.cfm

Customer Relations

Email: info@sei.cmu.edu
Telephone: +1 412-268-5800
SEI Phone: +1 412-268-5800
SEI Fax: +1 412-268-6257

Software Solutions Symposium 2017

Backup Slides

Example: Environmental Causes

Which environmental factors directly influence growth/biomass of Spartina grass (in Cape Fear)?

Chemical, mechanical, radiation, etc.

Causal learning yields: only pH matters

Other factors indirectly matter, mediated by pH levels

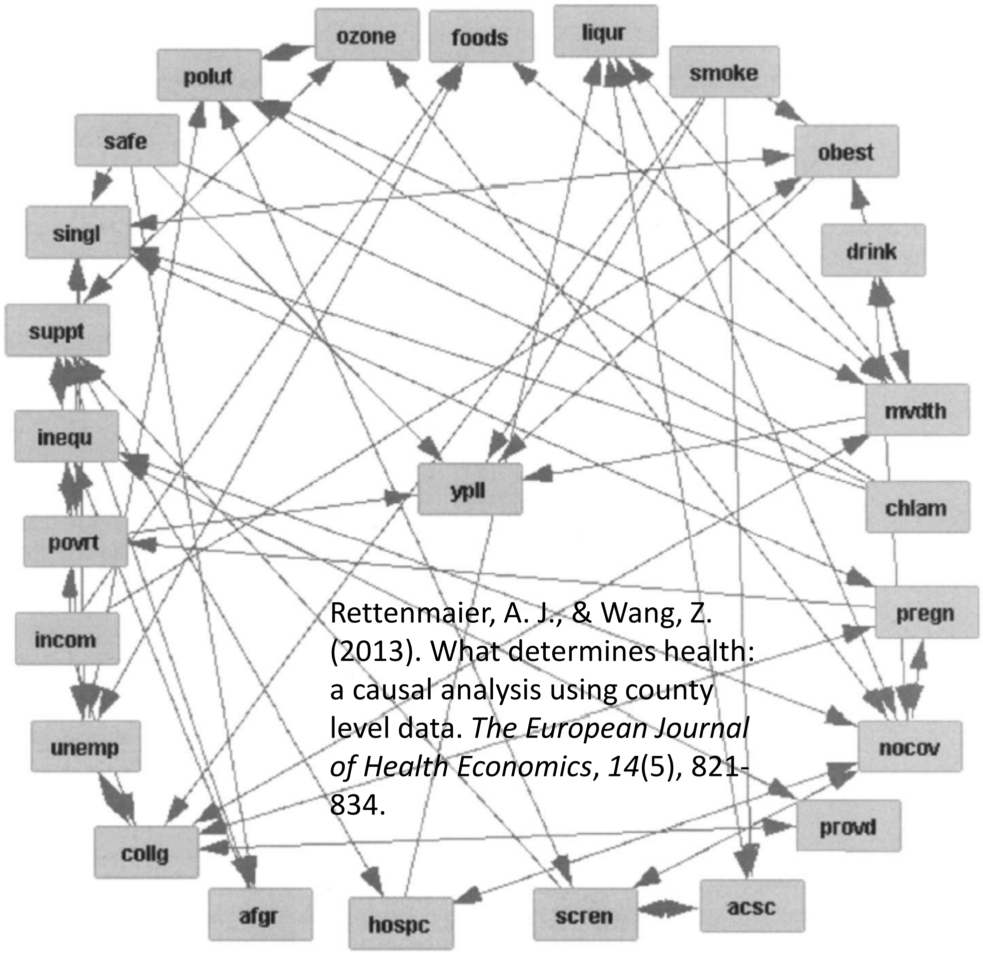
Subsequent greenhouse study randomly varying pH & salinity yields: only pH matters

Taken from “Why Causal Discovery and Modeling Should be in Your Research Design”, SEI Presentation, by Dr. David Danks, CMU, 2016.

Example: Premature Death

Figure 1; Final directed graph for YPLL. Note: …Obtained with the PC search algorithm using as input the correlations between premature death rate (YPLL) and 24 potential risk factors.

"…only (factors), adult smoking, obesity, the motor vehicle crash death rate, the percent of children in poverty, and violent crime rate, are causal factors of premature death.



Rettenmaier, A. J., & Wang, Z. (2013). What determines health: a causal analysis using county level data. *The European Journal of Health Economics*, *14*(5), 821-834.

"Chlamydia rate, violent crime or homicide, and liquor store density are three important factors that cause low birth weight, though in our discussion we note that two latter factors may proxy for correlated unobserved variables."

Basic Concepts of Directed, Acyclic Graphs

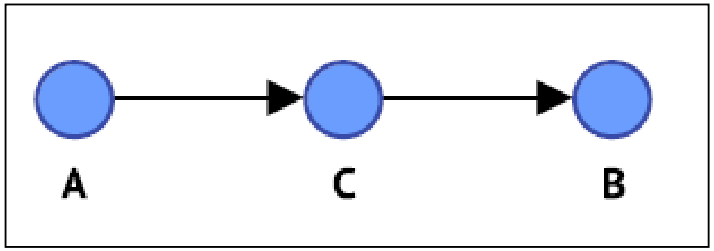
1. DAGs consist of:
   a) nodes (variables),
   b) directed arrows (possible causal relationships ordered by time), and
   c) missing arrows (confident assumptions about absence of causal effects)

2. DAGs are nonparametric
   a) No distributional assumptions
   b) Linear and/or nonlinear

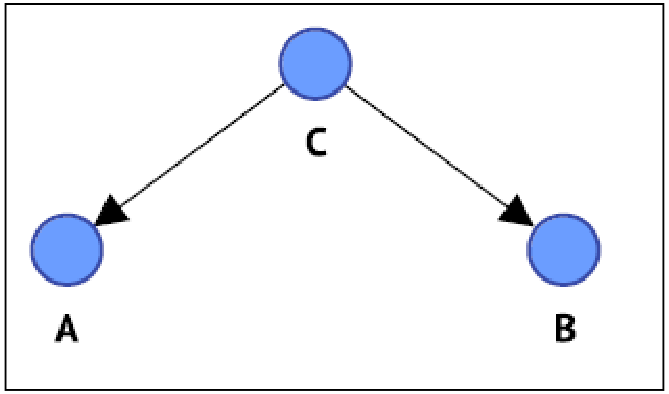
3. DAGs have both causal paths and non-causal (spurious) paths

Fundamental Structures in a Directed, Acyclic Graph

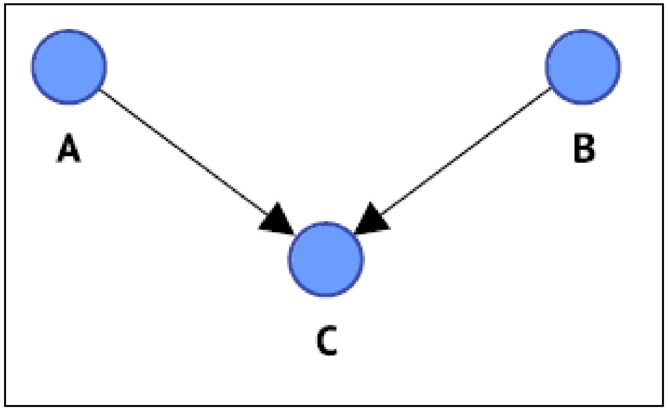
1. Indirect Connection



2. Common Cause



3. Common Effect (Collider)



Use of Directed, Acyclic Graphs

1. Derive testable implications of a causal model to evaluate if the model is correct

2. Understand causal identification requirements to confirm whether causality may be inferred from the data
   - Separating causal from spurious associations in the data

3. Inform use of traditional statistical techniques such as regression
   - Deciding which control variables to include versus not to include in the analysis to achieve identification of causality

Deriving Testable Implications

1. Uses a technique such as d-Separation
   a) Algorithm to help determine which paths are causal versus non-causal
   b) Uses concept of **blocking a path (see next slide)** to stop transmission of non-causal association

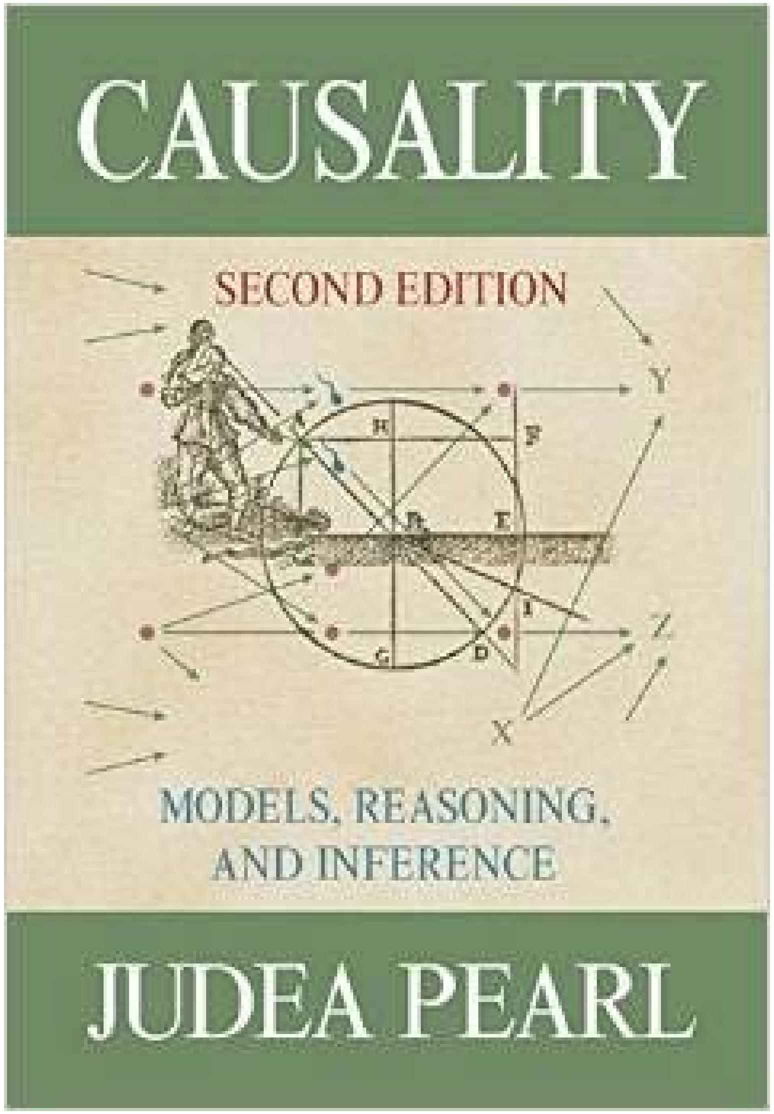
2. Additional techniques that may be employed include
   a) Graphical identification
   b) Adjustment Criterion
   c) Backdoor Criterion
   d) Frontdoor Criterion
   e) Pearl's do-Calculus

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!

Causal Modeling is the Real Destination



Quotes by Judea Pearl

*"… 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)

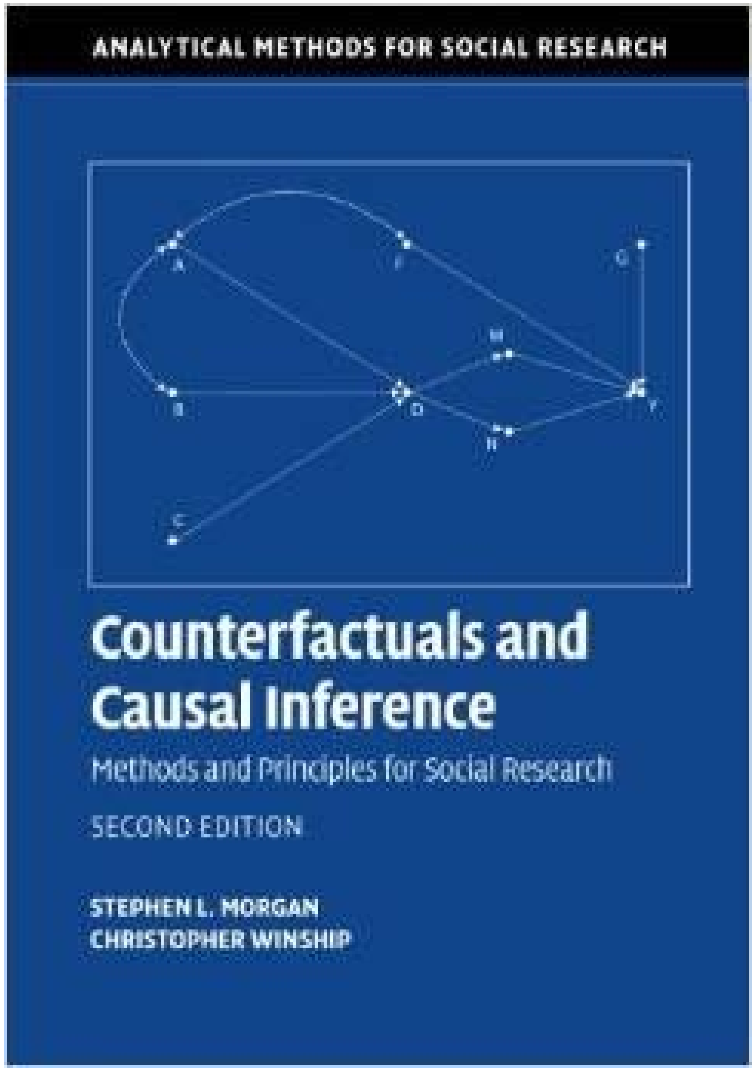
*"The development of **Bayesian Networks**, so people tell me, marked a turning point in the way uncertainty is handled in computer systems. For me, this development was a stepping stone towards a more profound transition, **from reasoning about beliefs to reasoning about causal and counterfactual relationships**."*

Judea Pearl: From Bayesian Networks to Causal and Counterfactual Reasoning

Keynote Lecture at the 2014 BayesiaLab User Conference
Recorded on September 24, 2014, in Los Angeles.

[Distribution Statement A. This material has been approved for public release and unlimited distribution]

Causal Modeling – Dr. Stephen Morgan



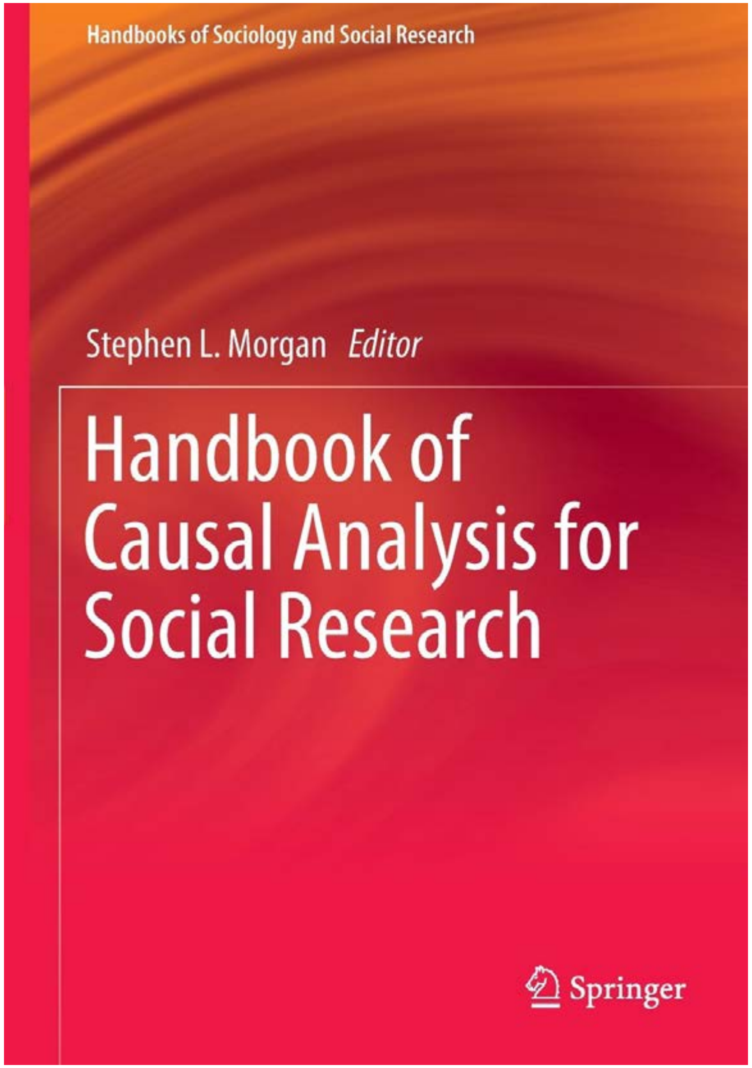
ANALYTICAL METHODS FOR SOCIAL RESEARCH

Counterfactuals and Causal Inference

Methods and Principles for Social Research

SECOND EDITION

STEPHEN L. MORGAN
CHRISTOPHER WINSHIP



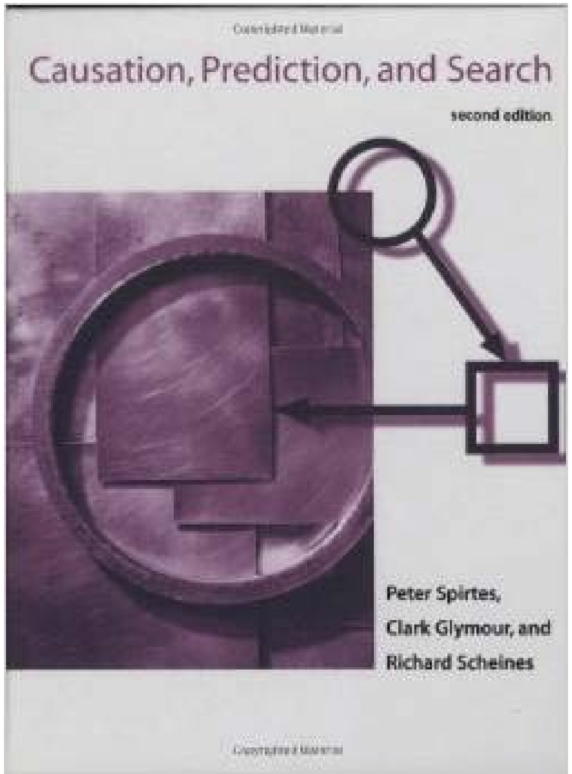
Handbooks of Sociology and Social Research

Stephen L. Morgan *Editor*

Handbook of Causal Analysis for Social Research

Springer

CMU Causal Modeling Researchers-01



Copyrighted Material

Causation, Prediction, and Search

second edition

Peter Spirtes,
Clark Glymour, and
Richard Scheines

Copyrighted Material

Richard Scheines (Dean, Dietrich College of Humanities & Social Sciences)

Professor of Philosophy, Machine Learning Department, and HCII

- Graphical and Statistical Causal Inference
- Philosophy of Social Science
- Foundations of Causation
- Educational Technology & Online Courses

Baker Hall 154
412.268.2831
scheines[at]cmu.edu

David Danks (Department Head)

Professor of Philosophy and Psychology
Department Head

- Causal learning (human and machine)
- Cognitive Science
- Philosophy of Psychology
- Philosophy of Science
- Bounded Rationality
- Decision-making

Baker Hall 161D
412.268.8047
ddanks[at]cmu.edu

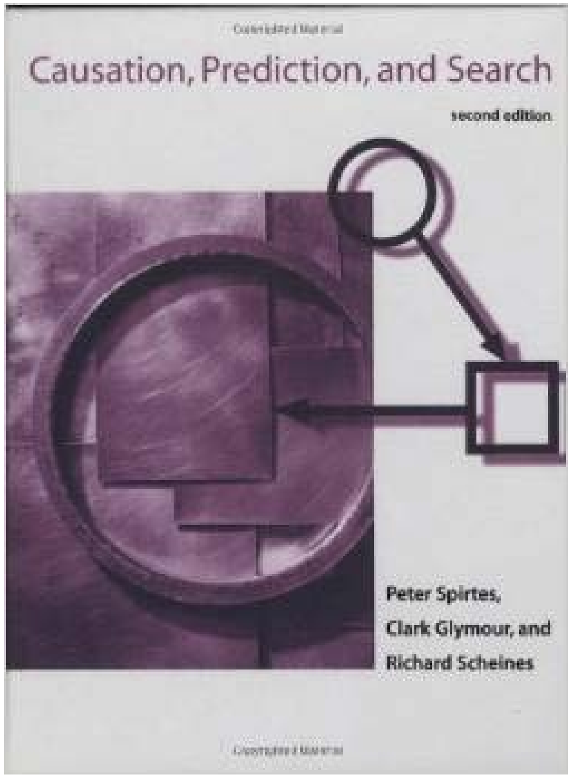
Clark Glymour

Alumni University Professor

- Philosophy of Science
- Causal Modeling
- Cognitive Science
- Machine Learning
- Automated Genomics

Baker Hall 135L
412.268.2933
cg09[at]andrew.cmu.edu

CMU Causal Modeling Researchers-02



Copyrighted Material

Causation, Prediction, and Search

second edition

Peter Spirtes,
Clark Glymour, and
Richard Scheines

Copyrighted Material

Joseph Ramsey

Director of Research Computing

Baker Hall 143
412.268.8063
jdramsey[at]andrew.cmu.edu

- Computational Causal Inference
- Automatic Proof Search
- Automated Genomics
- Online Courseware

Peter Spirtes

Professor

Baker Hall 135D
412.268.8460
ps7z[at]andrew.cmu.edu

- Graphical and Statistical Modeling of Causes
- Causation in the Social Sciences
- Philosophy of Physics

Causal Inference with Directed Graphs and Treatment Effects

Dr. Felix Elwert, Univ of Wisconsin

Available through two channels:

Statistical Horizons
www.statisticalhorizons.com

BayesiaLab
http://www.bayesia.us/causal-inference-course-fairfax

Notes on Tetrad Screenshots -1

Step 1: A few words about Tetrad: Tetrad is free and maintained by an open-source community led Dr. Joe Ramsey at CMU. Like many statistical tools, Tetrad takes CSV files as input, so you if work in Excel, you should Save As Type: CSV (Comma Delimited).

Step 2: Note that you can specify other types of input, delimiters, comment tags, missing value tags, etc. Also, note that there is a default way to specify which variables are categorical vs. ordinal; but if a column has numbers and at least one of them is a decimal number, it is treated as Continuous by default. When you have configured the loading the way you want, click on “Load.”

Step 5 (Using the default settings): When there aren’t too many cases in your dataset, these default settings are very reasonable. As the number of cases increase, your tolerance for False Positives drops, or the graph gets too bushy, you might want to reduce Alpha by one or more magnitudes. Also, if you have a very large dataset and a lot of variables (e.g., tens of thousands) so that execution is slow, you might want to limit the size of subsets entering conditional independence testing by changing the Depth from -1 (representing unlimited size) to something small (e.g., positive < 10).

Notes on Tetrad Screenshots -2

Step 5 (Executing search): It is tempting at this stage to move the variables around so that causation moves in the same direction, say left to right, or top to bottom.

But the default organization of the variables around a circle is a pretty good default as it makes each edge potentially visible and gauge graph bushiness.

If you find some edge orientations to be incorrect, you can specify such information about orientation (or temporality) in a **Knowledge** box, which will then guide/constrain the causal search.

You connect a Knowledge box into your Tetrad session by running an arrow from the Data Box to the Knowledge box. You then specify the information by opening the box to map the variables of your dataset into tiers (layers) and white-list/black-list particular dependencies. Then having specified such knowledge, you connect the Knowledge box to the Search box (which must separately be connected to the Data box) and then run your search. The resulting graph will conform to the knowledge you specified.

Notes on Tetrad Screenshots -3

Step 7: Had the Search box indicated that there were undirected edges, you would first need to modify the search graph before connecting the Search box to the Parametric Model box by inserting a Graph Manipulation box between the two to convert undirected edges into bi-directed edges.

Selecting the Generalized SEM option allows specification of more complex functions and error terms than ordinary SEM.

You can directly build a parametric model for a graph that includes bi-directed edges. For a bi-directed edge, the assumption is that there is a common latent cause and thus the error terms at both ends must be correlated to each other. To address a bi-directed edge, the graph produced by the Parametric Model box will show the edge as running between the associated error terms rather than the variables themselves.

Notes on Tetrad Screenshots -4

Step 8: Error terms are assumed to be independent, but in the case of a bidirectional edge, the edge is between the error terms, as mentioned. If our variables were Discrete rather than Continuous, the Parametric Model box menu would offer a Bayesian parametric model.

Note the default labeling of Bn for edge weights and Mn for mean value of a variable (node). There’s also a parameter for the variance (standard deviation) of the error term.

Step 9: Depending on the dataset, complexity of the search graph (number of bi-directed edges), etc., the estimation algorithm might take a while.

Note that from the same dataset, you can hang other search boxes (e.g., with different Alpha settings), parametric models, and estimator boxes, all stemming from the same dataset. You can also hang a Data Manipulation box to subset the data further before proceeding with search, etc.

Notes on Tetrad Screenshots -5

Step 10: Examine the resulting estimated model (now with instantiated parameters) for anything unusual, specifically: unnatural edge weights, stability of solution, quality of model fit.

Because our original graph was cyclic, a statistical regression (each variable against its parents) is not an appropriate optimizer.

One of the iterative approximating algorithms (e.g., Powell) had to be used instead. Click on Estimate Again, to see whether the instantiated model is sensitive somehow to the random seed. Repeat several times and in the Model Statistics view, if the chi square is about the same value each time, then the convergence to the particular configuration of parameter estimates is pretty robust convergence.

This particular result, from our PSP dataset subset, turned out to be robust.