20th PSMUG 2019

Causal Discovery with Observational Data Workshop

Mike Konrad

Bob Stoddard

Bill Nichols

Dave Zubrow

SEI's SCOPE Project: Towards a Causal Model for Software Cost

Problem

- DoD leadership continues to ask "Why does software cost so much?"
- DoD program offices need to know where to intervene to control software costs

Solution

- An actionable, full **causal** model of software cost factors immediately useful to DoD programs and contract negotiators

Actionable intelligence

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

Purpose of Workshop

The SEI is leading a **three-year research project (SCOPE**—previous slide**)** that seeks to:

- Apply modern advances in **causal learning** (search and estimation)
- Go beyond traditional correlation and regression analyses and accurately identify the causal relations among software process factors and product outcomes

With this **workshop**, we intend to **continue to**:

- Inform and update the PSM community
- Encourage **joint collaboration** in the **early adoption of causal learning** to improve the quality of systems engineering and software engineering research.

Goals/Products of Workshop

The workshop will produce the following:

Group statement to the PSM community on what **evidence** is needed to **support an endorsement** on:

- Use of causal learning for conducting research in systems and software engineering

Recommendations on:

a. What research questions should be a focus for causal learning to confirm/debunk conventional wisdom (a continuation from last year's PSM 2018 workshop)

b. Next steps to build awareness of the need for and benefit from causal learning

Bottom-line: a clearer understanding of causal learning and the unique role it can play in conducting research using observational data.

Outline

What is **causal learning**?

Activity 1: Identify research **questions** for evaluating a policy

What are causal discovery **algorithms**?

Activity 2: Identify **policies** that became dysfunctional

What is an **example** application of causal learning? (Case Study)

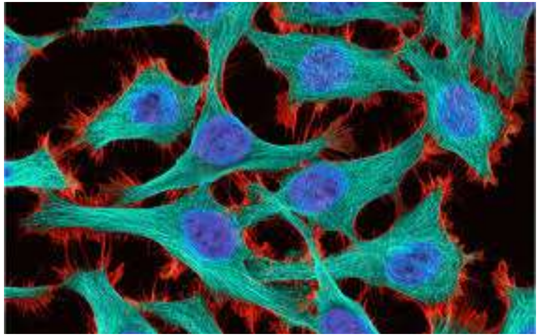
Activity 3: Formulate a group statement on next steps for PSM Community

Conclusion

Attribution

A portion of the presentation that follows was **adapted from** "AN INTRODUCTION TO CAUSAL MODELING AND DISCOVERY USING GRAPHICAL MODELS" **by David Danks**, Head of Philosophy Department at CMU: http://www.andrew.cmu.edu/user/ddanks/pubs.html.

Correlation doesn't inform us about causes



How do cancer cells differ from non-cancerous cells?

If we just want to predict which cells are cancerous, then correlations are sufficient.

If we want to change cancerous cells into non-cancerous ones (or at least, not dangerously cancerous), then we need causal knowledge.

Causation vs. correlation

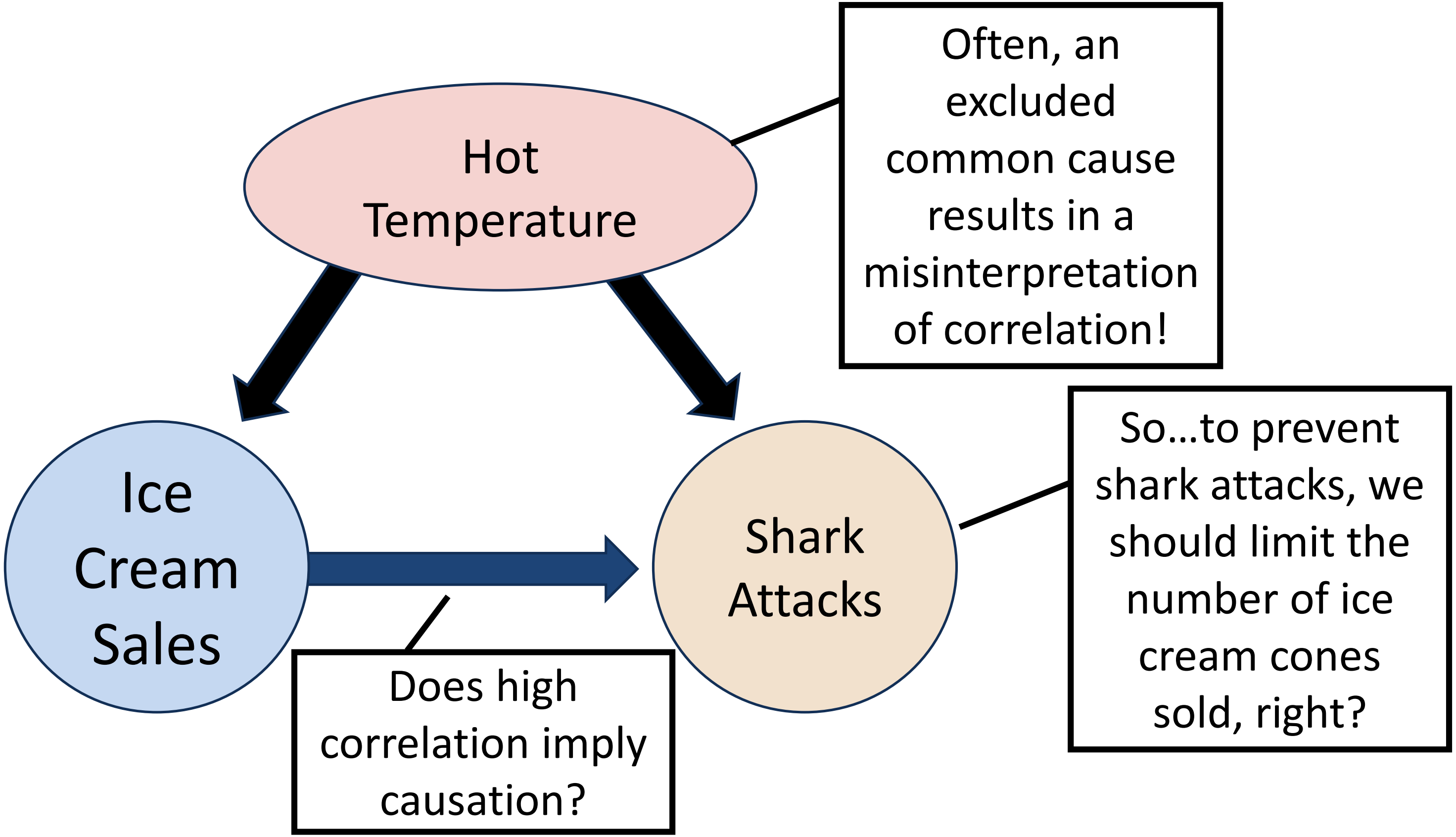
Correlation ➛ things tend to go together (or in opposite directions)

- Learning about one is informative about other

Causation ➛ changing one (from the outside) tends to change the other

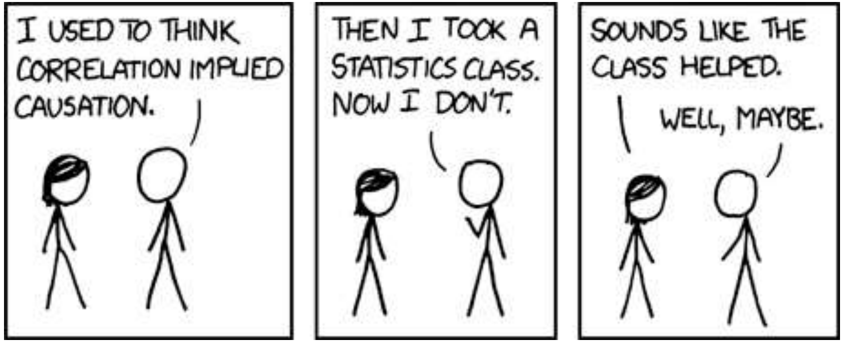
- Manipulation of one leads (probabilistically) to variation in the other

More about Misinterpreting Correlation!



Causation vs. correlation

Statistics slogan: *Correlation ≠ Causation*



Credit: https://xkcd.com/552/

Sarah Sheard slogan: "Correlation doesn't *cause* causation, but is *correlated* with causation."

Prof. David Danks' summary: "Correlation is a noisy indicator of causation."

Causation vs. correlation

Different uses for each:

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

Causation vs. correlation

Caution: don’t conclude that one is better than the other…

Moral is two-fold:

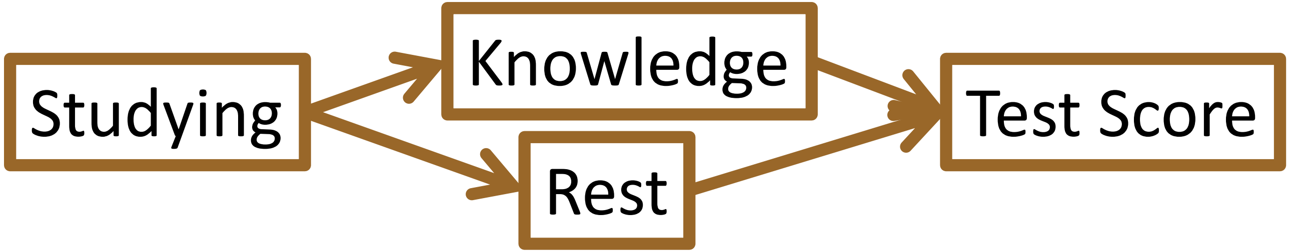
1. Make sure you know which you have
2. Make sure you know what you want to do

Causal Learning: Framework

Causal graphical models

Graph ➛ qualitative (direct) causation

- Directed Acyclic Graph (DAG) over variables
- Many variations (time-indexing, context variables, …)



Using causal knowledge

Given a causal model, you can:

- Predict given evidence or information
- Construct explanations & troubleshoot
- Design actions/policies to achieve specific outcomes

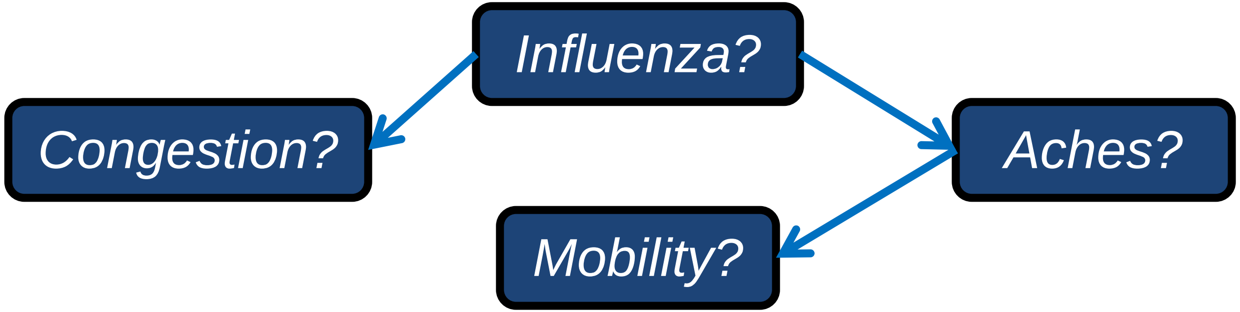
Given multiple causal models, you can:

- Find distinguishing experiments or evidence
- Determine which is better supported
- Compute “expected” outcomes

Using causal knowledge

Basic idea of actions/interventions:

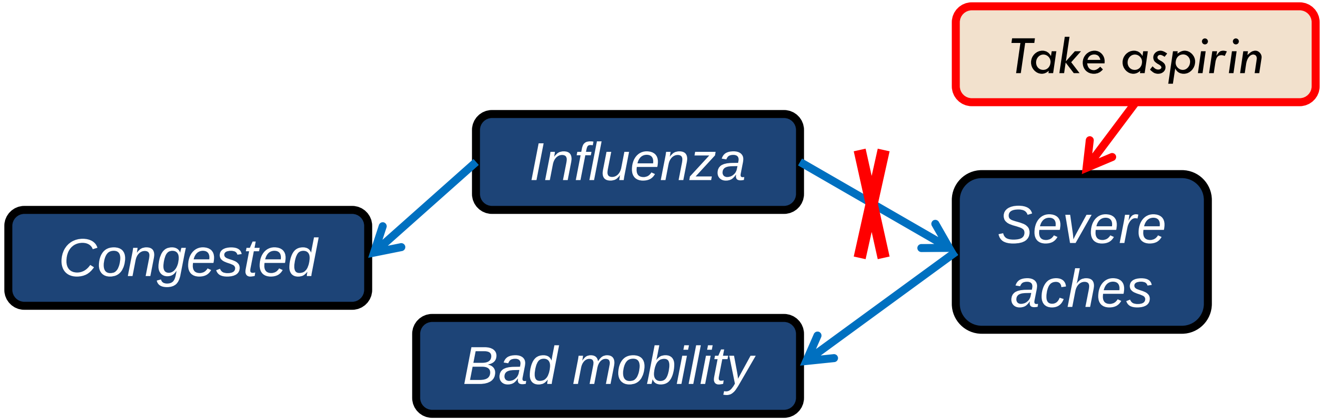
- Often, can “take control” of a node
- A *manipulation* that *changes* the causal system from “outside”
  - In contrast with merely observing the system



Using causal knowledge

Basic idea of actions/interventions:

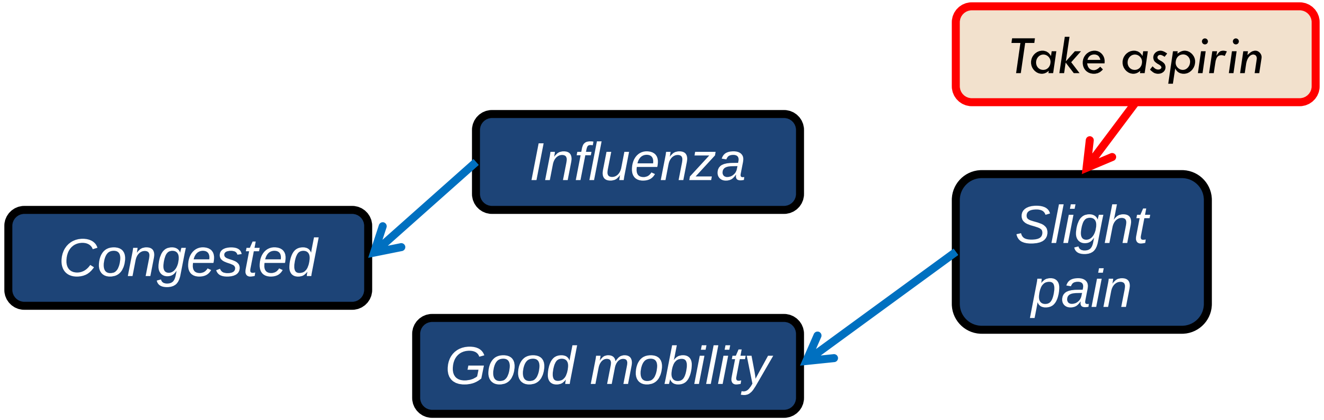
- Often, can “take control” of a node
- A *manipulation* that *changes* the causal system from “outside”
  - In contrast with merely observing the system



Using causal knowledge

Basic idea of actions/interventions:

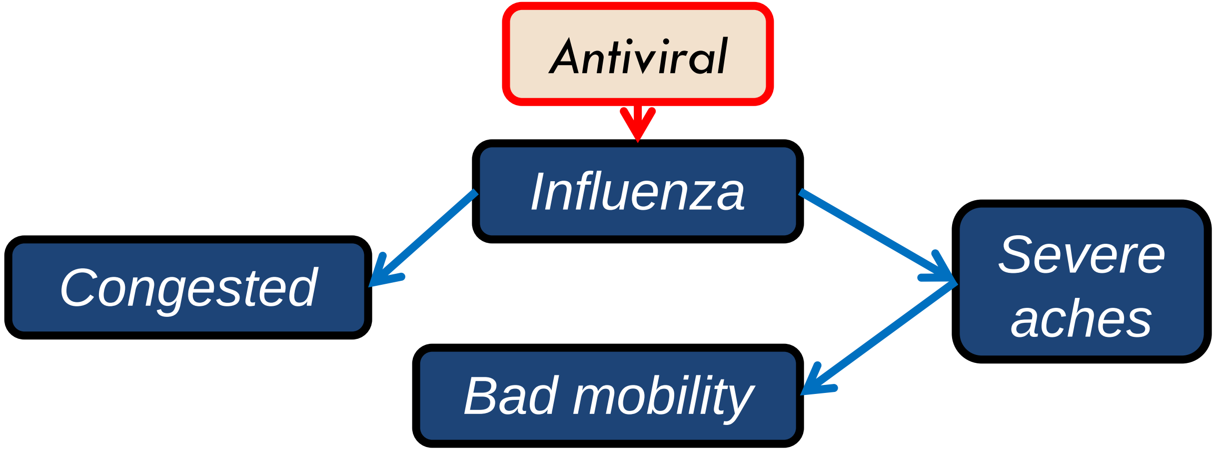
- Often, can “take control” of a node
- A *manipulation* that *changes* the causal system from “outside”
  - In contrast with merely observing the system



Using causal knowledge

Basic idea of actions/interventions:

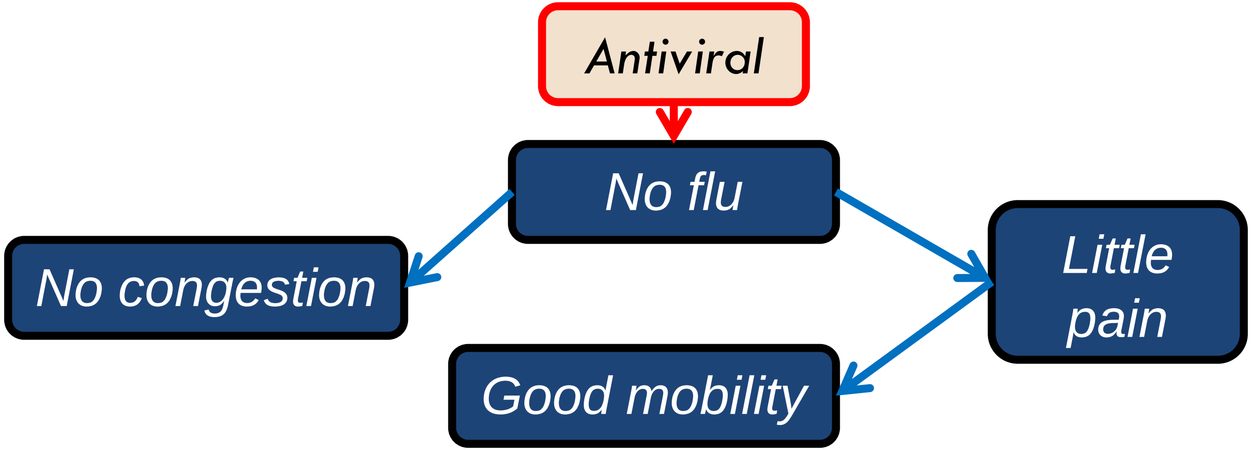
- Often, can "take control" of a node
- A *manipulation* that *changes* the causal system from "outside"
  - In contrast with merely observing the system



Using causal knowledge

Basic idea of actions/interventions:

- Often, can “take control” of a node
- A *manipulation* that *changes* the causal system from “outside”
  - In contrast with merely observing the system



Resurgence of Causal Learning in the Past 30 Years

Sewall Wright Path Models (1920's)
Structural Equation Models (1930's)
Social Science Path Models (1960's)
Bayesian Networks (1980's)
Glymour & Spirtes *et al* 1st ed. book on Causality (1988)
Pearl's Probabilistic Reasoning (1988)
Pearl's 1st ed. book on Causality (2000)

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

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

Morgan Counterfactuals & Causality (2014)

Peters Elements of Causal Inference (2017)

Pearl The Book of Why (2018)

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/

Outline

What is **causal learning**?

Activity 1: Identify research **questions** for evaluating a policy

What are causal discovery **algorithms**?

Activity 2: Identify **policies** that became dysfunctional

What is an **example** application of causal learning? (Case Study)

Activity 3: Formulate a group statement on next steps for PSM Community

Conclusion

Activity 1: Identify Research Questions for evaluating a policy

Approach:

- Select 1-3 US DoD policy **goals**
- What research **questions** will help us evaluate the policy's **effectiveness**?
- What **results** would convince a researcher that causal learning (CL) provided an insight beyond what could be obtained by another analytic approach?

Outputs: a text document that identifies:

(1) Policy goal(s) relevant to the US DoD

(2) Research questions and datasets to potentially investigate

(3) Example CL results that convey unique insight not obtainable in other ways

Takeaway: A scientific approach to policy definition and deployment requires investigating causes and effects relevant to achieving its goals.

Outline

What is **causal learning**?

Activity 1: Identify research **questions** for evaluating a policy

What are causal discovery **algorithms?**

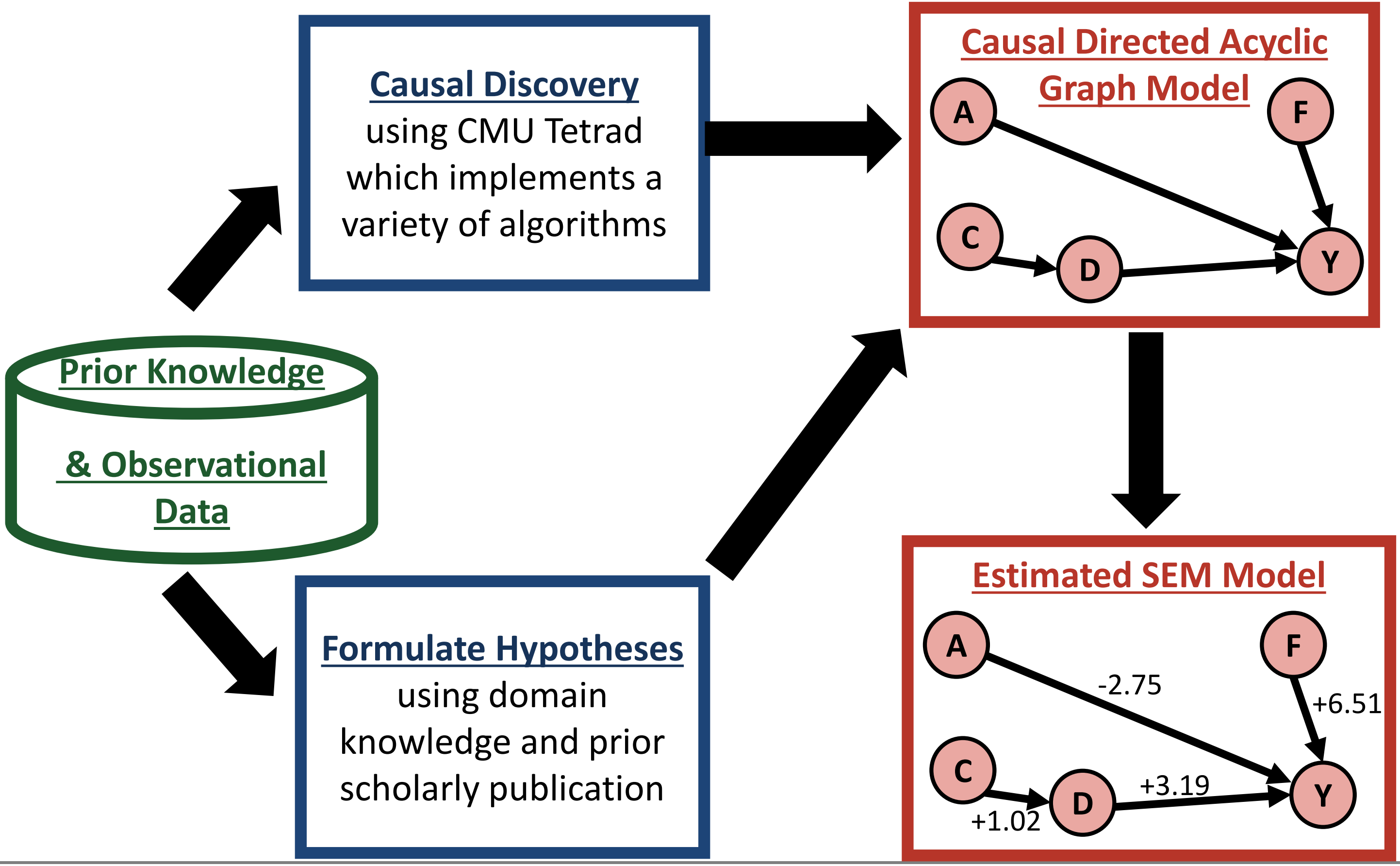
Activity 2: Identify **policies** that became dysfunctional

What is an **example** application of causal learning? (Case Study)

Activity 3: Formulate a group statement on next steps for PSM Community

Conclusion

The Broader Causal Learning Landscape



The Output of a Causal Discovery Algorithm is a DAG

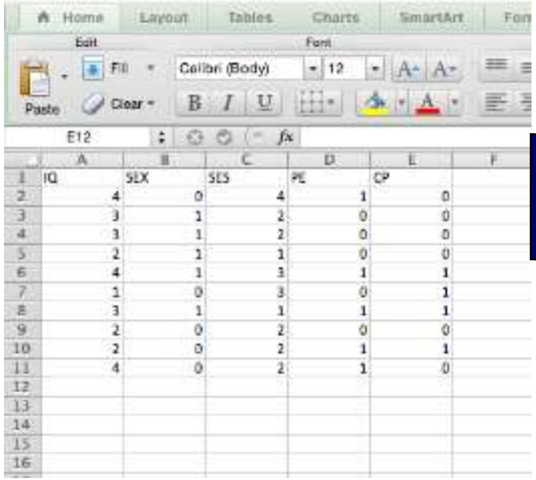
The output of a causal discovery algorithm is a **graph**

- Nodes are the variables in the dataset
- **Directed** (oriented) edges indicate direct cause-and-effect relationships
- Resulting graph has no cycles (e.g., A→B→C→A) and is called **acyclic**
- Thus a **Directed Acyclic Graph** (DAG)
- Not all edges need to be oriented
  - An undirected edge means that there's more than one graph representing the same data generation process and that for some of these graphs, the edge is oriented one way, while in others it is oriented the other way.
- Thus X → Y means:
  - A change in the value of X may result in a change in the value (or probability) of Y
    - (Whether or not you hold other variables constant)
  - The reverse is not true

Using a Causal Discovery Algorithm

Person XYZ

1. IQ: _____
2. Socio-Economic-Status: _____
3. Parental Encouragement: _____
4. College Plans: _____
5. Sex: _____



IQ	SEX	SES	PE	CP
4	0	4	1	0
3	1	2	0	0
3	1	2	0	0
2	1	1	0	0
4	1	3	1	1
1	0	3	0	1
3	1	1	1	1
2	0	2	0	0
2	0	2	1	1
4	0	2	1	0

(Optional) Background Knowledge +

PC (or other) Algorithm

Pattern (DAG)

Causal Learning: Algorithms

Multiple types of methods for identifying the direct cause-and-effect relationships in a dataset:

1. **Constraint-based**: Calculate independences in the data and do "backwards inference"
2. **Score-based** (**Bayesian**): Calculate the likelihood of different DAGs given the data
3. **Hybrid**: Use constraint-based to get "close," then Bayesian search around neighborhood

How Constraint-Based Algorithms Work (in a Nutshell)

Start with a **complete undirected** graph formed from all the variables in your dataset

First stage: determine **adjacencies** (not-yet directed edges) by iteratively:

- Selecting a pair of nodes connected by an edge, say A and B
- If A and B are independent conditioned on some set of the other variables (including the empty set), then *remove* that edge from the graph.

Continue until you can remove no more edges. The result is a graph of un-oriented edges.

Second stage: determine **orientation** by iteratively identifying *colliders*:

- Select an *unshielded triple* A --- C --- B or A → C --- B
  (“unshielded” means that there’s no edge connecting A and B)
- The fact that A and B have no edge means that there was a set **S** that when conditioned on, made A and B independent. If **S** does not contain C then orient the triple as A → C ←B.
- When there are no more colliders, apply Meek’s Rules to finish orienting some edges.

How Score-Based Algorithms Work (in a Nutshell)

Start with an **empty** graph formed from all the variables in your dataset

First stage: add edges to the graph by iteratively:

- For each pair of nodes not already connected by an edge, say A and B, determine the relative improvement in score (SEM BIC) from adding the edge A→B.
  - In other words, how much more probable is the data we have, if the graph describes the data-generation process?
- For that edge A→B best improving the score, add that edge to the graph.

Continue until you can add no more edges to improve the score.

Second stage: remove edges from the graph by iteratively:

- Removing the edge that best improves the score (SEM BIC).

Continue until you can remove no more edges to improve the score.

The result is a graph of oriented edges (*all* edges are oriented).

Example Constraint-Based and Score-Based Algorithms

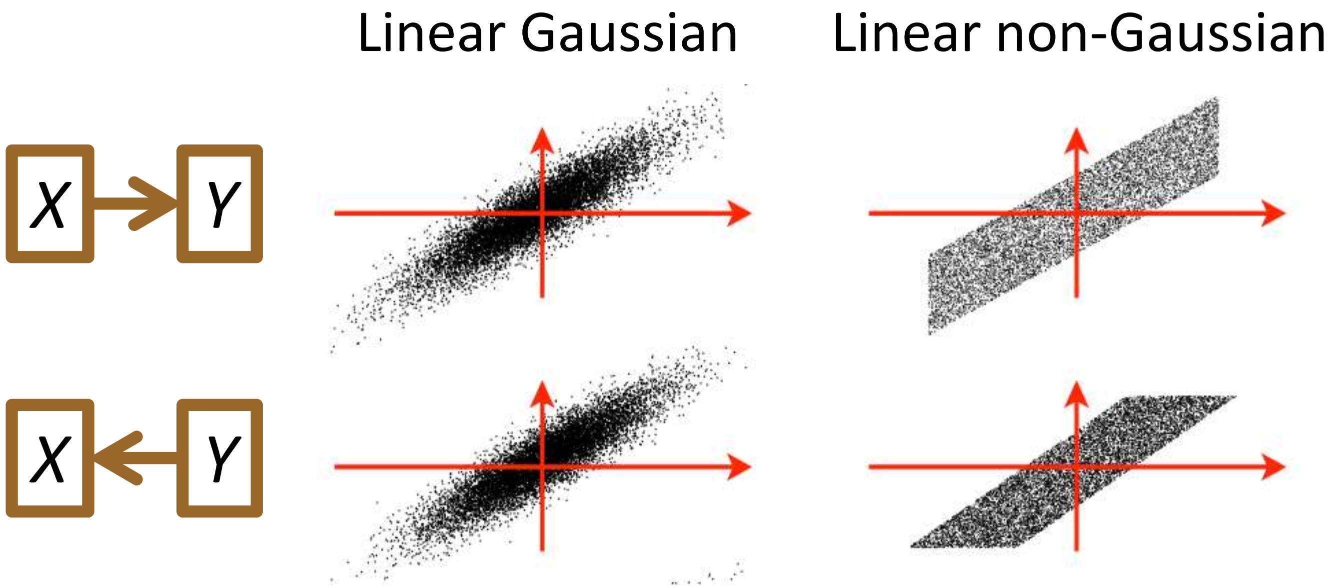
PC Stable, constraint-based search algorithm

- Variant of PC, the most widely used algorithm
- Resulting search graph does not depend on the order of the variables
- Parameters to tune (settings for running the algorithm):
  - **Independence Test type**: Fisher Z (Continuous), Chi Square Test (Discrete)
  - **Alpha**: cutoff for p-values in independence testing; for small datasets, choose higher Alpha
  - **Collider discovery and conflicts**: Max-P and Orient bi-directed
  - **Maximum size of conditioning set**: when sample size is small, chose value in range 1..3

FGES (Fast Greedy Equivalent Search), score-based search algorithm

- Parameters to tune (settings for running the algorithm):
  - **Scoring method**: SEM BIC Score
  - **Penalty Discount**: the default is often 1 or 2; higher values lead to sparser graphs; lower values (e.g., 0.5) may be used to reduce risk of false negatives (e.g., small datasets)

Some Algorithms Exploit Non-Gaussianity



Guidelines for Causal Discovery

1. Examine distributions: are they **Gaussian**? Do scatterplots suggest **linearity**?
2. Continuous variables: are they mixtures of different causal systems? Be aware of **Simpson's paradox**. Consider using algorithms such as IMaGES or MultiFASK.
3. Dataset has **both categorical and continuous** variables: use Conditional Gaussian model. Discretizing continuous variables is generally not a good idea.
4. **Bootstrapping** is a very useful way of assessing how much trust to place in a DAG.
5. **Missing values**: a challenge to many statistical methods, not just causal discovery. In general, you will want to address these before applying discovery algorithms.
6. **Selection bias? Measurement error?** Consider algorithms designed to address these issues: FCI, GFCI, RFCI. For unmeasured common causes, consider Two-Step.
7. Incorporate **knowledge**: factors known to cause (or to not cause) other factors.
8. Search procedures generally have no confidence intervals for their results. For estimation, **Model fit** statistics are problematic unless one has very large samples.

Advanced Topics

Unobserved common causes & selection bias

Measurements of proxies, not underlying causal factors

Time series causal structures

Equilibrated systems with feedback

Unobserved intermediate “mechanism” variables

Datasets with multiple (overlapping) sets of variables

Non-stationary causal *structure*

Similar-but-varying causal structures across individuals

Undersampled time series with missing, causally relevant variables

Massive numbers of variables (> 1M)

…

Where to Learn More

Pearl J, Glymour M, Jewell NP. Causal Inference in Statistics – A Primer (John Wiley & Sons, 2016).

Spirtes Peter, "Introduction to causal inference." Journal of Machine Learning Research 11 (2010) 1643-1662. http://jmlr.org/papers/volume11/spirtes10a/spirtes10a.pdf

The Tetrad Project. http://www.phil.cmu.edu/tetrad/

Jonas Peters, Dominik Janzing, Bernhard Schölkopf. Elements of Causal Inference: Foundations and Learning Algorithms. (Adaptive Computation and Machine Learning series, 2017).

Clark Glymour, Kun Zhang, and Peter Spirtes. A Brief Review of Causal Discovery Methods. (Frontiers, 2018).

Malinsky D, Danks D. Causal discovery algorithms: A practical guide. (Philosophy Compass, 2018). https://doi.org/10.1111/phc3.12470

Raghu VK, Poon A, Benos P. Evaluation of Causal Structure Learning Methods on Mixed Data Types. (JMLR 2018).

Outline

What is **causal learning**?

Activity 1: Identify research **questions** for evaluating a policy

What are causal discovery **algorithms**?

Activity 2: Identify **policies** that became dysfunctional

What is an **example** application of causal learning? (Case Study)

Activity 3: Formulate a group statement on next steps for PSM Community

Conclusion

Activity 2: Identify policies that became dysfunctional

Approach: brainstorm, perhaps based on the results of Activity 1:

- 1-3 policies that maybe became dysfunctional and perhaps why

Outputs: a text document that identifies:

1. US DoD policies that maybe became dysfunctional and in what way
2. How much of that dysfunctionality can be attributed to:
   - A lack of causal knowledge
   - Embracing a myth (as causal)
3. What causal knowledge was lacking or falsely embraced?

Takeaway: Though a policy may have laudable goals, what it prescribes or how it is deployed may be based on an incorrect or incomplete understanding of cause-and-effect relationships, resulting in unintended or undesirable consequences.

Outline

What is **causal learning**?

Activity 1: Identify research **questions** for evaluating a policy

What are causal discovery **algorithms**?

Activity 2: Identify **policies** that became dysfunctional

What is an **example** application of causal learning? (Case Study)

Activity 3: Formulate a group statement on next steps for PSM Community

Conclusion

Complexity Drivers and Project Success Case Study -1

Source: Sarah Sheard's Ph.D. dissertation, 2012

Research question: which complexity factors, determinable early in life of a program, impact project outcomes such as cost overrun, late delivery, performance shortfall?

Dataset: survey covering complexity factors and project success

- 41 items on a 3-point or larger ordinal scale
- 1 item (Delivered) on a binary scale (yes/no)
- 7 items representing project outcomes:
  - Delivered, EvolOp, GoodEst, Late, OverCost, PerfGap, Success

There were 81 survey responses, most indicating their domain as AeroSpace or Defense, but some indicating Civil Government and Consumer domains.

Complexity Drivers and Project Success Case Study -2

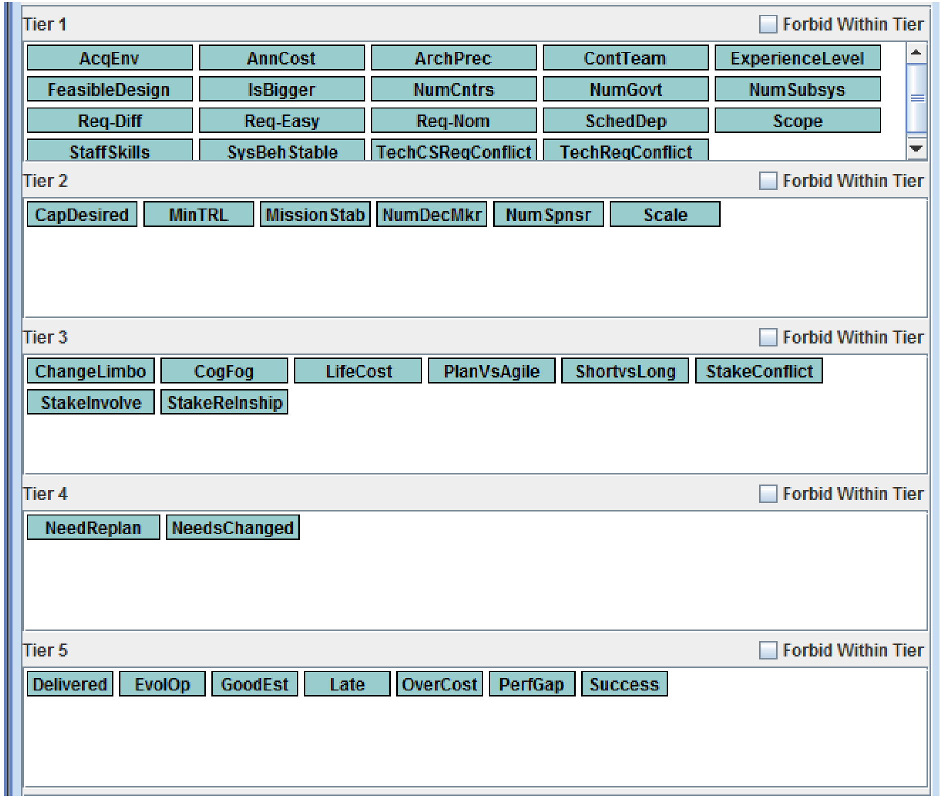
Original result: Three of the complexity variables strongly predicted all outcomes:

Req-Diff	**Difficult requirements are considered:** **- difficult to implement or engineer** **- hard to trace to source** **- to have a high degree of overlap with other requirements.** **How many system requirements were there that were Difficult?** **(1) 1-10 (2) 10-100 (3) 100-1000 (4) 1000-10,000 (5) Over 10,000**
CogFog	**"The project frequently found itself in a fog of conflicting data and cognitive overload". Do you agree with this statement?** **(1) Strongly Agree (2) Agree (3) Neutral (4) Disagree (5) Strongly Disagree**
StakeRelnship	**Where did project stakeholder relationships fit:** **(1)Relationships stable (Traditional frontier)** **(2)New relationships (Transitional frontier)** **(3)Resistance to changing relationships (Messy frontier)**

Complexity Drivers and Project Success Case Study -3

In Sarah's dissertation, the goal was to find factors that could be measured at the beginning or middle of a program that would indicate the need to take corrective action.

- At right, we see how the variables might be organized according to when they might be available to be measured in a program.
- Tier 1 represents program beginning.
- Tier 5 represents program outcomes.



Complexity Drivers and Project Success Case Study -4

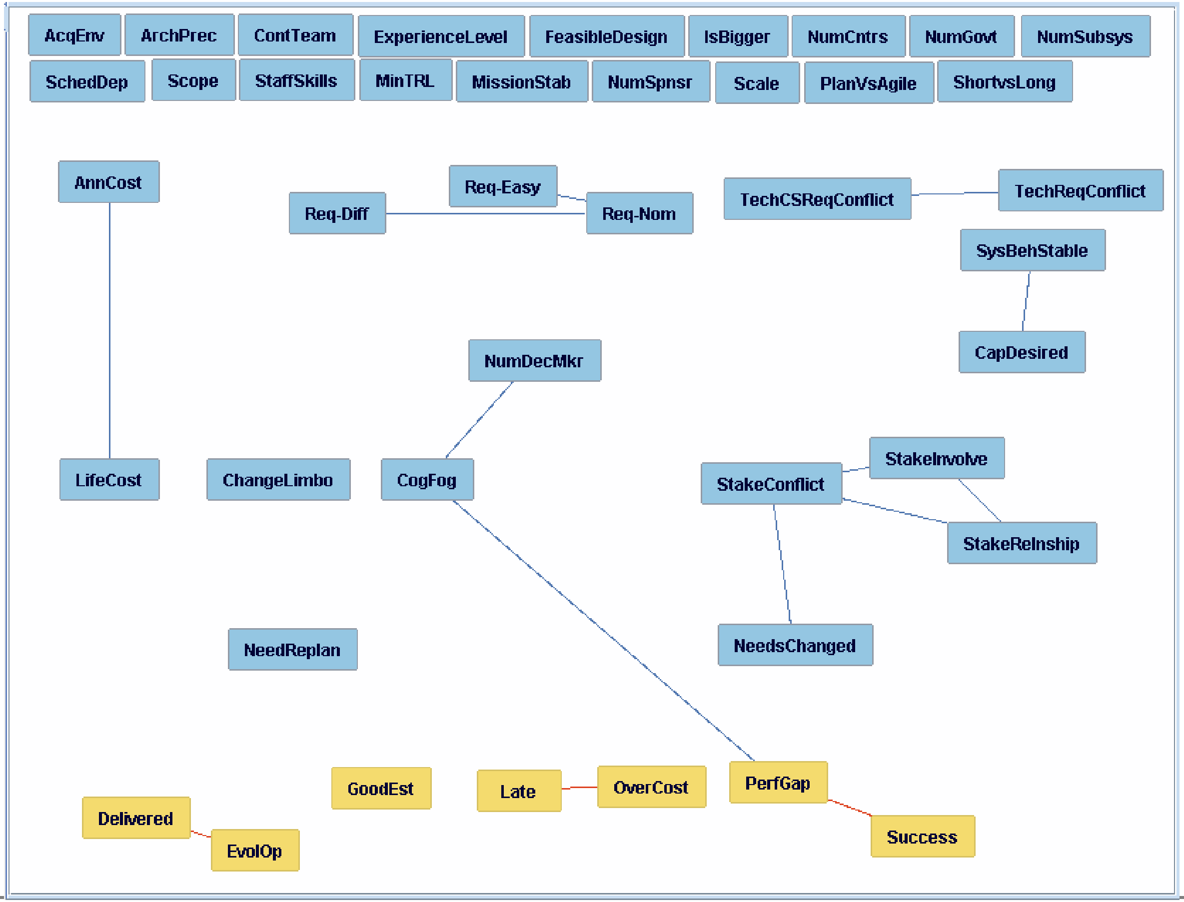
PC-Stable and FGES algorithms were applied to the full dataset (81 projects).

Here is an example search result from applying PC-Stable (Alpha=.10) to the full dataset.

Outcome (Tier 5) variables are highlighted in gold.

Note the CogFog-PerfGap relationship.

Variables without causal relationships were moved to the very top to help highlight direct causal relationships.



Complexity Drivers and Project Success Case Study -5

Regarding the three predictors identified in (Sheard, 2012), we would interpret the causal search result presented on the previous slide as saying **there is evidence that**:

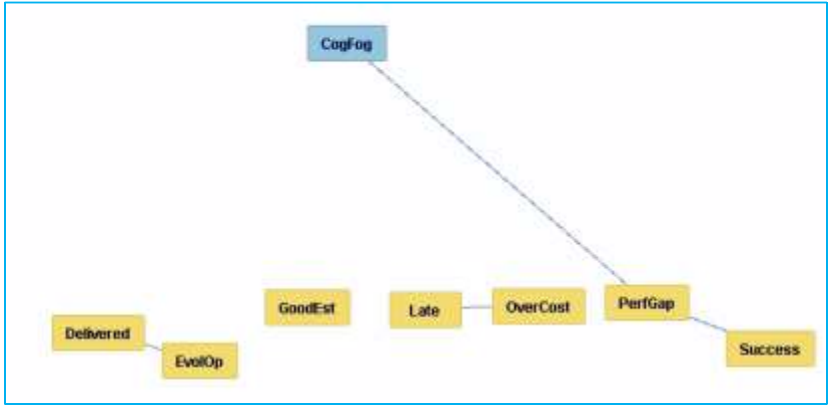
- NumDecMkr directly causes **CogFog**, which directly causes PerfGap, which directly causes (or is caused by) Success.
  - Note that the last two of these are project outcomes.
- The three **stakeholder variables** (StakeConflict, StakeRelnship, StakeInvolve) relate to each other and cause program needs to change
  - But there is no evidence for a causal path from stakeholder variables to any project outcome.
- There is also no evidence of a causal path from the **number of difficult requirements** (Req-Diff) to any project outcome.

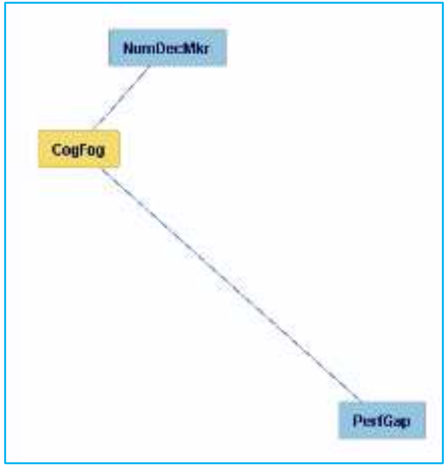
We can further express the causal roles for NumDecMkr and CogFog in terms of Markov blankets. See next slide.

Complexity Drivers and Project Success Case Study -6

On this slide we show two **Markov blankets**: (1) for all project outcomes; (2) for CogFog.

A **Markov blanket** is a node, its parents, its children, and its children's parents. The Markov blanket of a node is the only knowledge needed to predict the behavior of that node. (Wikipedia)





So according to the above search graphs, and for this dataset:

- The only knowledge that will help predict project outcomes is amount of cognitive fog.
- The only knowledge that will help predict cognitive fog is number of decision makers.

Complexity Drivers and Project Success Case Study -7

Summary of what we learned from this limited causal analysis:

- We have evidence for this causal path: NumDecMkr → CogFog → PerfGap
- **Early in a program**, if the likelihood of meeting project outcomes seems low, consider how to reduce/streamline the **number of decision makers**.
  - Will help **reduce** the amount of **cognitive fog**
  - Will help **reduce the performance gap** (specified mission-critical features vs. what was actually achieved)
  - Will help improve project **success**.
- However, from our causal analyses, we found **no evidence** that taking the following type of actions (Sheard, 2012) would improve project outcomes:
  - improve stakeholder relationships
  - reduce the number of difficult requirements
- Taking the above two actions might not improve project outcomes at all!

Why Couldn't We Identify Other Cause-Effect Relationships?

Our expectations are wrong: the purported effects of particular development and acquisition approach might not be so significant

- Or project and acquisition management are capable of quickly mitigating these effects

Data, measurement, and algorithm issues

- We are **not measuring the right behaviors** ("proxies" *might not always* help)
  - **Missing measures** of factors that may be the real drivers of program success
- Small sample (insufficient to detecting some cause-and-effect relationships)
- Data quality (e.g., do survey respondents accurately recollect/represent a project?)
- Assumptions for algorithm correctness are not met (e.g., linearity of relationships)

Mixture of causal systems (e.g., architectures, suppliers, processes, platforms, people)

- Simpson's paradox

A single study, even causal learning-based, will not answer all the questions we have.

Reflections from two years of causal learning

Causal Learning involves an iterative work cycle similar to that for machine/deep learning:

- Pose/revise research **questions**
  - Ensure there are variables representing the outcomes of interest and *context*
- Obtain, review, prepare, and analyze **dataset**
  - We maintain a **methodology** capturing lessons learned for: using two different search algorithms on a dataset, bootstrapping, imputing missing data, etc.
- Learn more about the **algorithms** and their assumptions to guide interpreting results
- Use **Bootstrapping** to: (1) reduce model overfitting; (2) improve confidence in results

What has helped:

- Patient, curious coworkers: Dave Zubrow, Sarah Sheard, Anandi Hira, Jim Alstad
- Expert assistance: David Danks, Kun Zhang, Madelyn Glymour, Joe Ramsey (CMU)

Outline

What is **causal learning**?

Activity 1: Identify research **questions** for evaluating a policy

What are causal discovery **algorithms**?

Activity 2: Identify **policies** that became dysfunctional

What is an **example** application of causal learning? (Case Study)

Activity 3: Formulate a group statement on next steps for PSM Community

Conclusion

Activity 3: Formulate group statement on next steps for PSM Community

Thesis: Given measures of the right project attributes, Causal Learning (CL) can help substantiate key cause-and-effect relationships that are the basis for US DoD policies

Approach: brainstorm, what actions the PSM Community and the SEI should take next.

Outputs: a text document identifying candidate next steps, **<u>for example</u>**:

1. Continue researching the application of CL to software
   - Identify significant but latent project factors.
2. Brainstorm, specify, and pilot measures for these factors
3. Offer training in CL methods.
4. Continue to build awareness in CL methods
5. Address issues that arise in the adoption and use of CL methods

Takeaway: There is much that we can collectively do to improve the quality of research in the broader communities of which we are a part.

Conclusion

Progress in systems engineering and software engineering can be accelerated through use of causal learning.

Causal learning also helps identify the need for **measuring new project attributes**.

The practical software measurement community has an opportunity to lead the way:

- Building awareness of CL methods
- Evaluating usefulness CL methods
- Improving usefulness of CL methods by:
  - Providing forums for indicator and measure specification and piloting
  - Supporting adoption of the new measures
- Adopting causal learning as part of your research toolkit

This won't happen without your continued support!

Contact Information

Presenter / Point(s) of Contact

Mike Konrad

Bob Stoddard

Bill Nichols

Email:

mdk@sei.cmu.edu

rws@sei.cmu.edu

wrn@sei.cmu.edu

Telephone:

+1 412.268.5813 (Mike)

+1 412.268.1121 (Bob)

+1 412.268.1727 (Bill)

Other SEI Team Members

Sarah Sheard

Dave Zubrow

Rhonda Brown

Chris Miller

CMU Contributors

David Danks

Madelyn Glymour

Joe Ramsey

Kun Zhang

USC Contributors

Anandi Hira, Jim Alstad, Barry Boehm