19th PSMUG 2018

Causal Search in Observational Data Workshop

Mike Konrad

Bob Stoddard

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

DM18-1041

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:

- Enlighten the practical measurement 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 **next steps** to enlighten the full community on causal learning and encourage adoption

We intend to accomplish the above by engaging in **working discussions** in **small groups** followed by a final large group summary of:

a. Research questions and hypotheses to be investigated through causal learning (confirm/debunk conventional wisdom)
b. Data sources helpful in causal learning research
c. Next Steps and discussion of participants' datasets

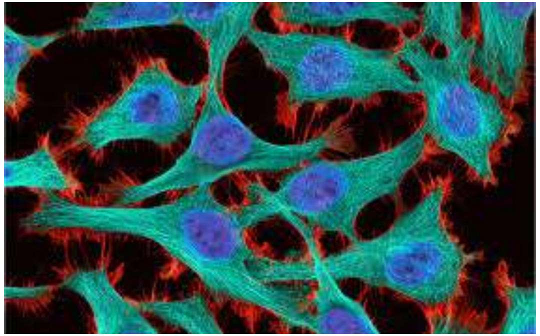
Bottom-line: a clearer understanding of causal discovery and the unique role it can play in conducting research using observational data.

Outline

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

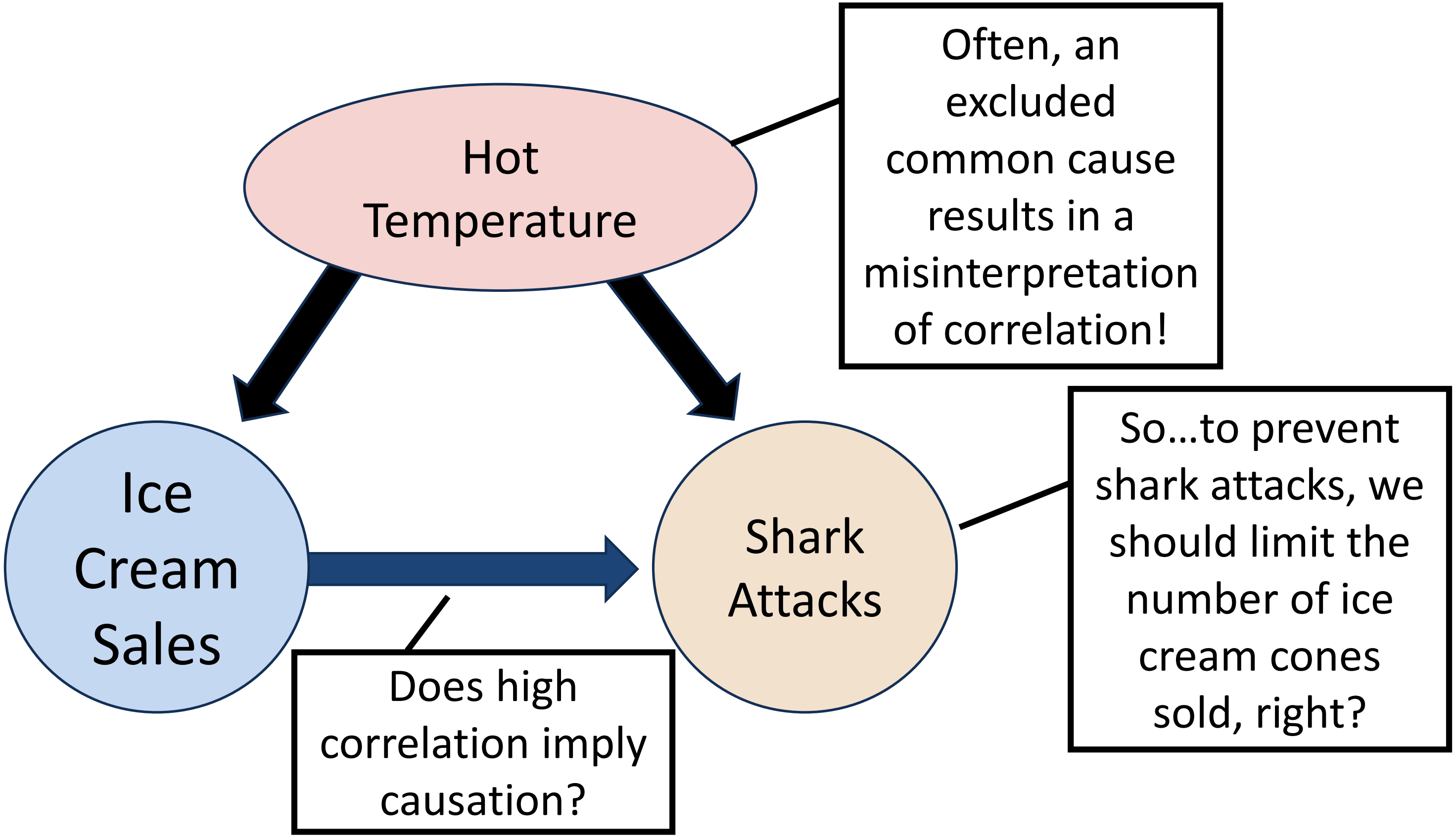
Correlation ➛ things tend to go together (or in opposite directions)

- Learning about one is informative about other

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

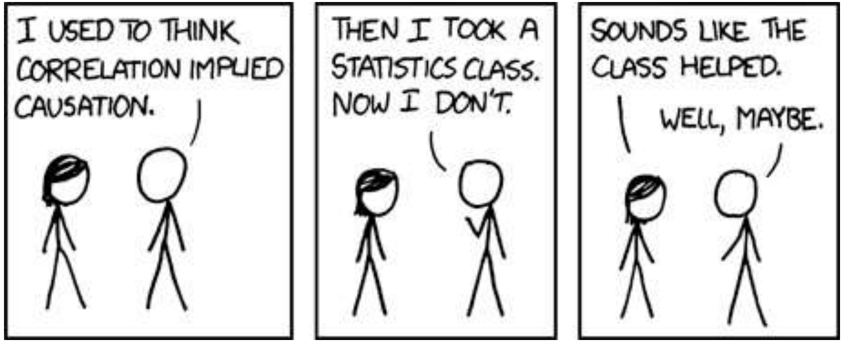
- Manipulation of one leads (probabilistically) to variation in the other

More about Misinterpreting Correlation!



Causation vs. correlation

Statistics slogan: *Correlation ≠ Causation*



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

Better 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:

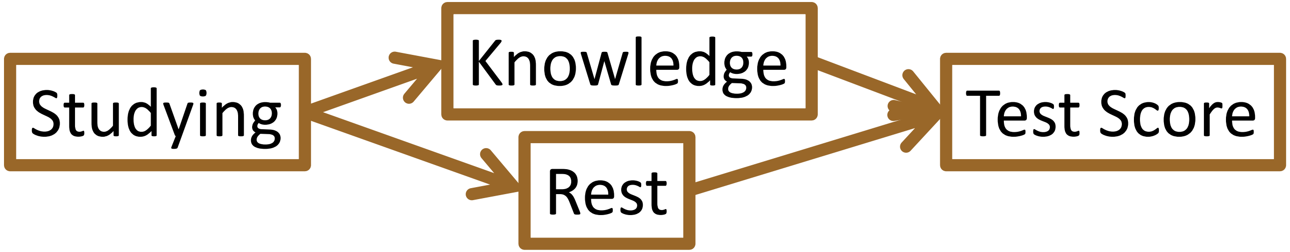
1. Make sure you know which you have
2. Make sure you know what you want to do

Causal learning: Framework

Causal graphical models

Graph ➛ qualitative (direct) causation

- Directed Acyclic Graph 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

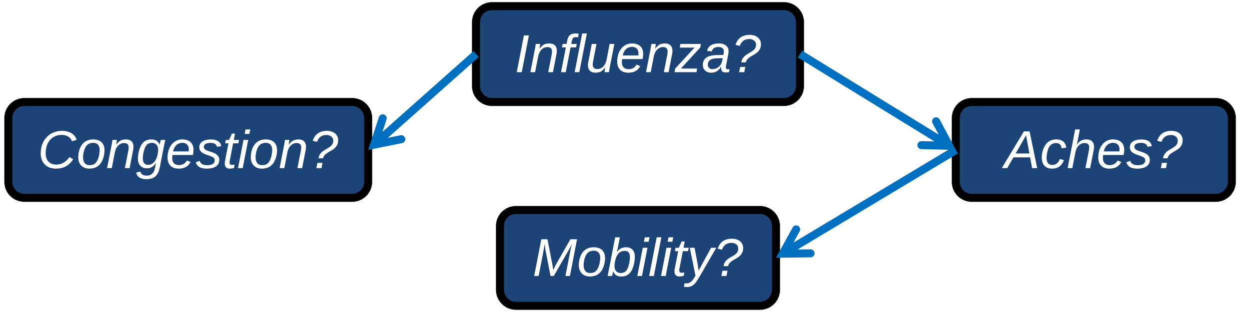
Given multiple causal models, you can:

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

Using causal knowledge

Basic idea of actions/interventions:

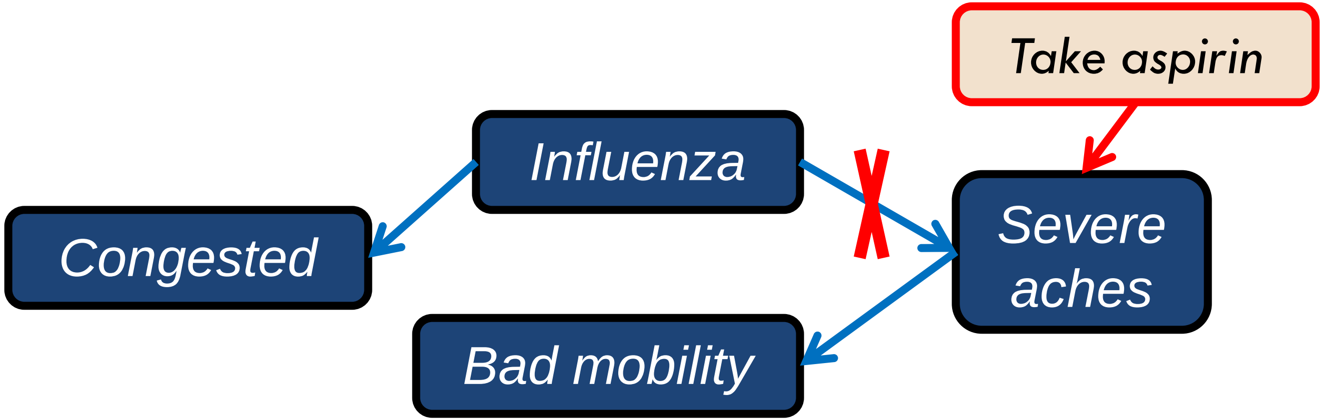
- Often, can "take control" of a node
- A *manipulation* that *changes* the causal system from "outside"
  - In contrast with merely observing the system



Using causal knowledge

Basic idea of actions/interventions:

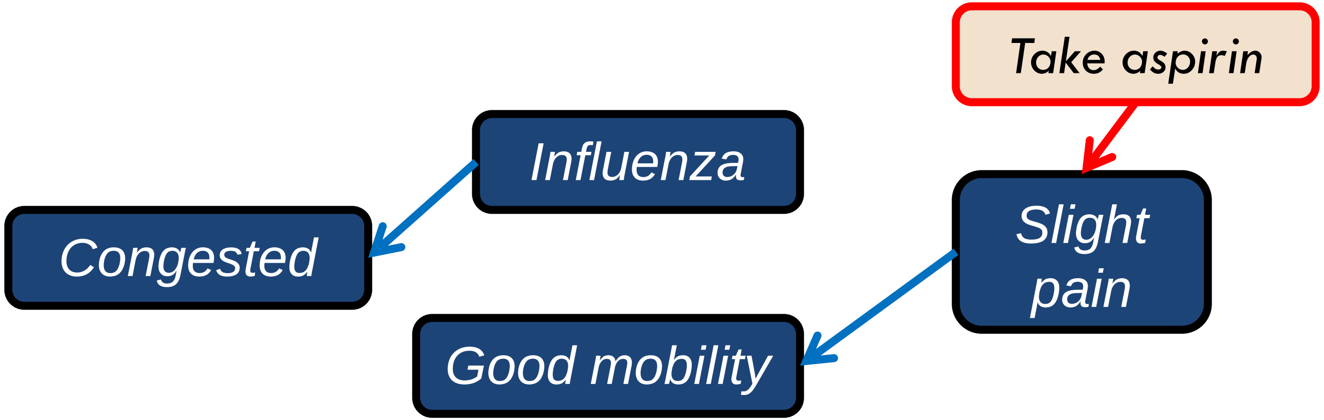
- Often, can “take control” of a node
- A *manipulation* that *changes* the causal system from “outside”
  - In contrast with merely observing the system



Using causal knowledge

Basic idea of actions/interventions:

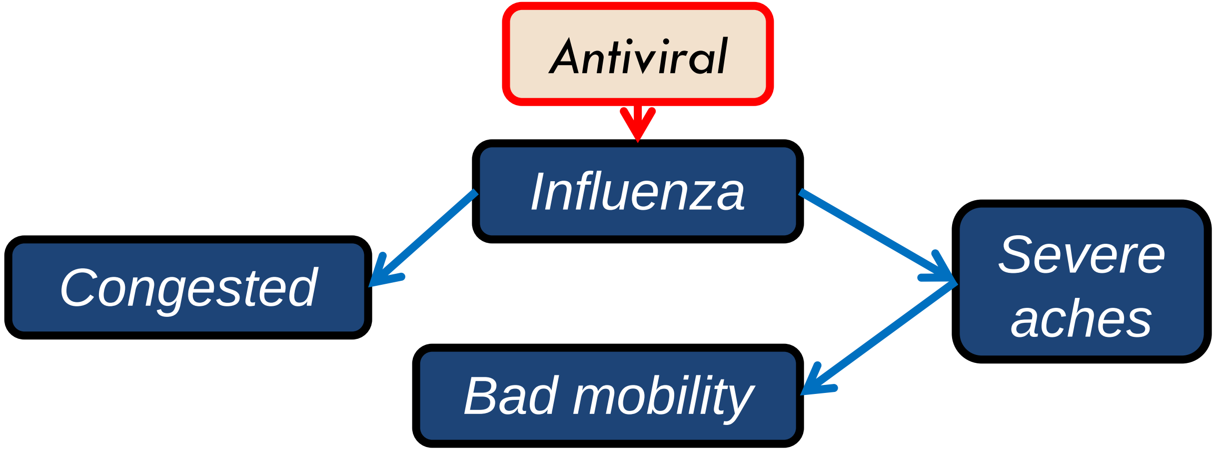
- Often, can “take control” of a node
- A *manipulation* that *changes* the causal system from “outside”
  - In contrast with merely observing the system



Using causal knowledge

Basic idea of actions/interventions:

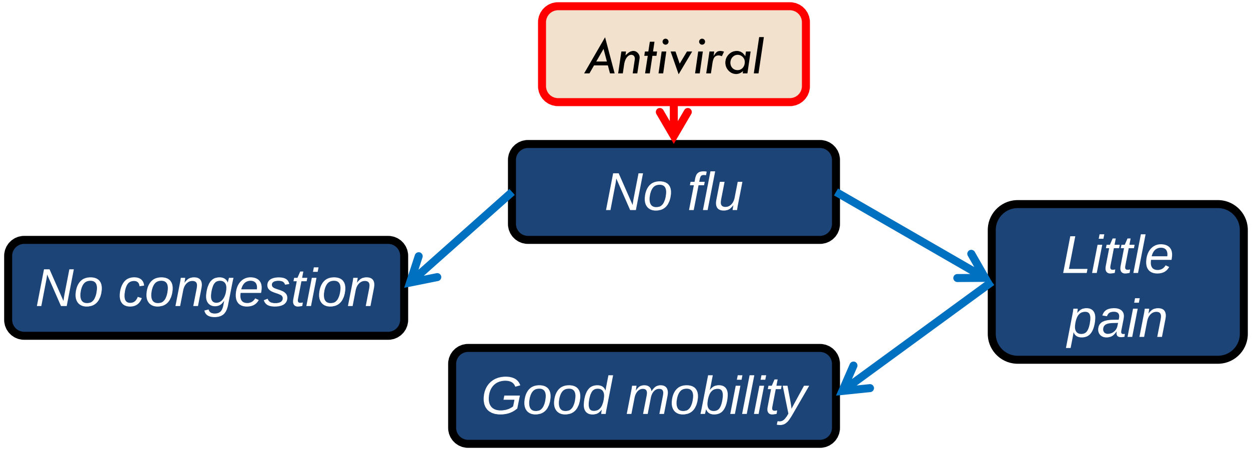
- Often, can “take control” of a node
- A *manipulation* that *changes* the causal system from “outside”
  - In contrast with merely observing the system



Using causal knowledge

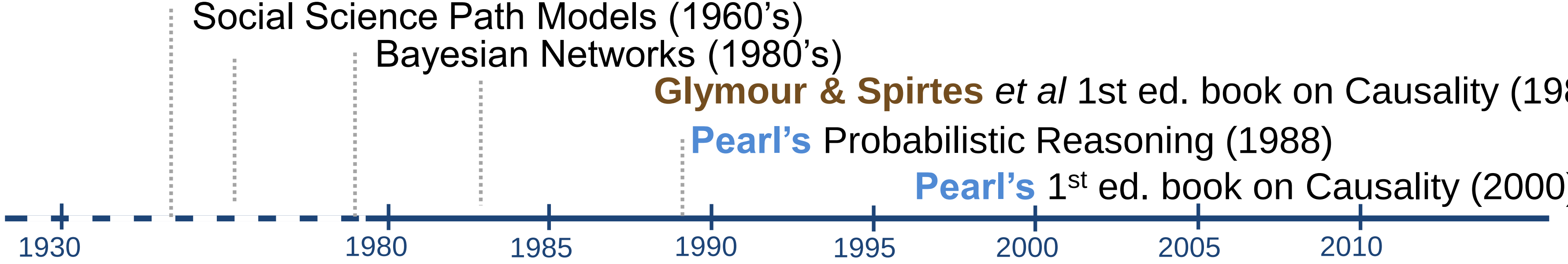
Basic idea of actions/interventions:

- 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/

Activity 1: Identify Research Questions or Factors

In separate groups (15 minutes):

- Brainstorm 2-3 topics of relevance to the US DoD and its goals
- Brainstorm what research questions or factors should be investigated to establish the causal knowledge needed to ensure effective policy making

Outputs: a text document that identifies:

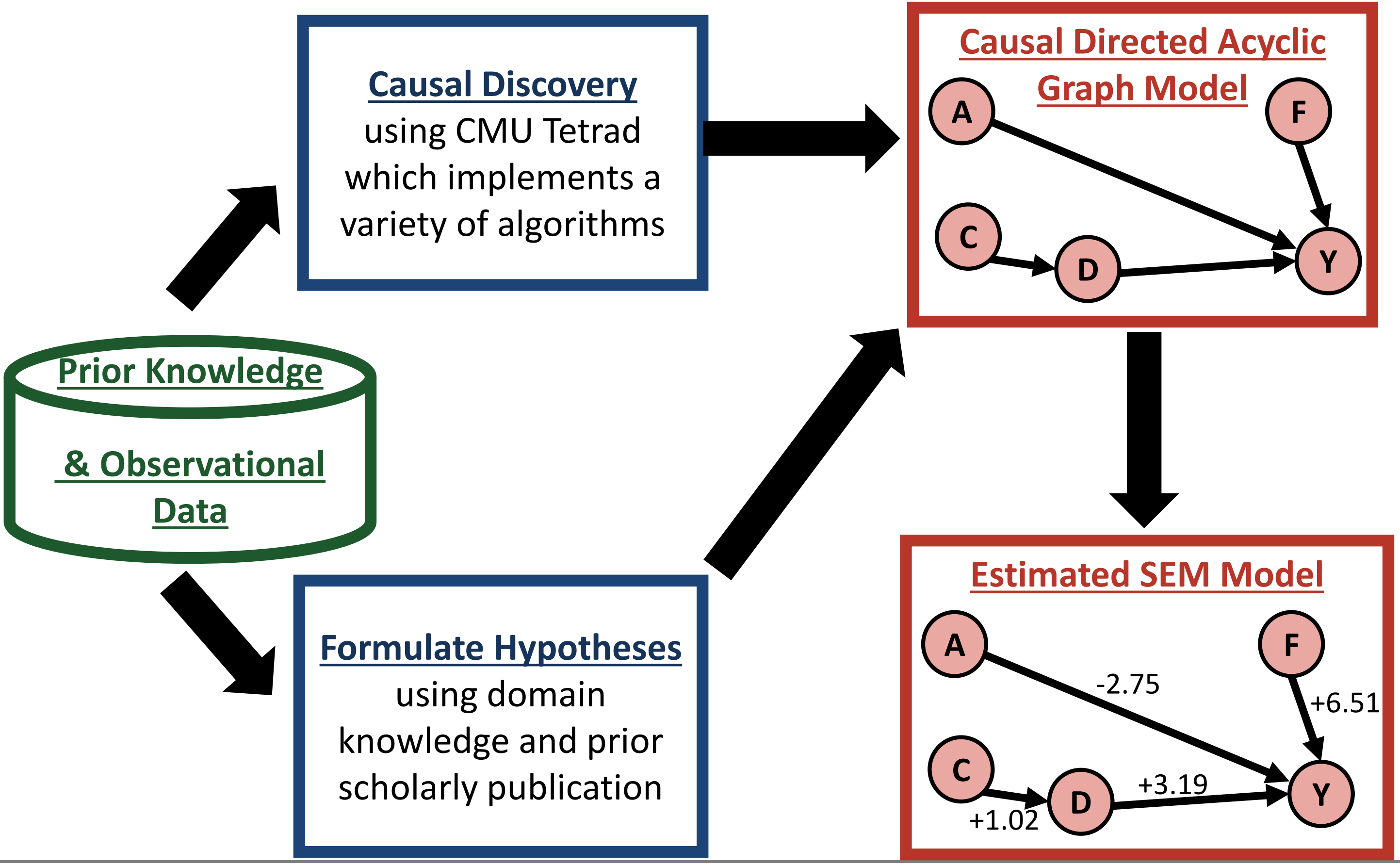
(1) Topic/policy of interest?

(2) Research questions to investigate to guide DoD policy making?

(3) Factors to analyze to help answer the research questions?

Takeaway: A scientific approach to policy definition and deployment requires investigating causes and effects relevant to achieving the US DoD's goals.

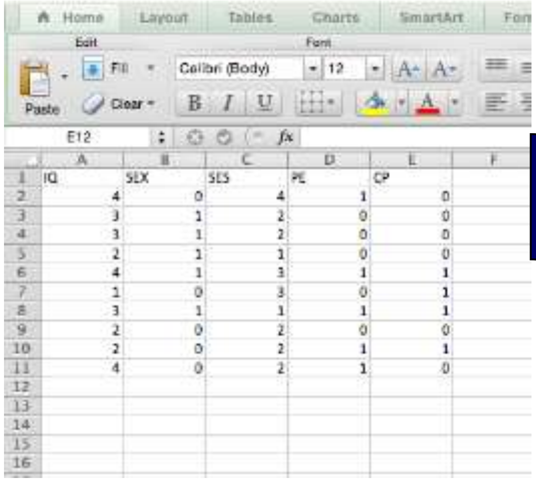
The Broader Causal Learning Landscape



Using a Causal Discovery Algorithm

Person XYZ

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



	A	B	C	D	E
1	IQ	SEX	SES	PE	CP
2	4	0	4	1	0
3	3	1	2	0	0
4	3	1	2	0	0
5	2	1	1	0	0
6	4	1	3	1	1
7	1	0	3	0	1
8	3	1	1	1	1
9	2	0	2	0	0
10	2	0	2	1	1
11	4	0	2	1	0

(Optional) Background Knowledge +

PC (or other) Algorithm

Pattern

Causal learning: Algorithms

Multiple types of methods for this idea:

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

Example Constraint-Based and Score-Based Algorithms

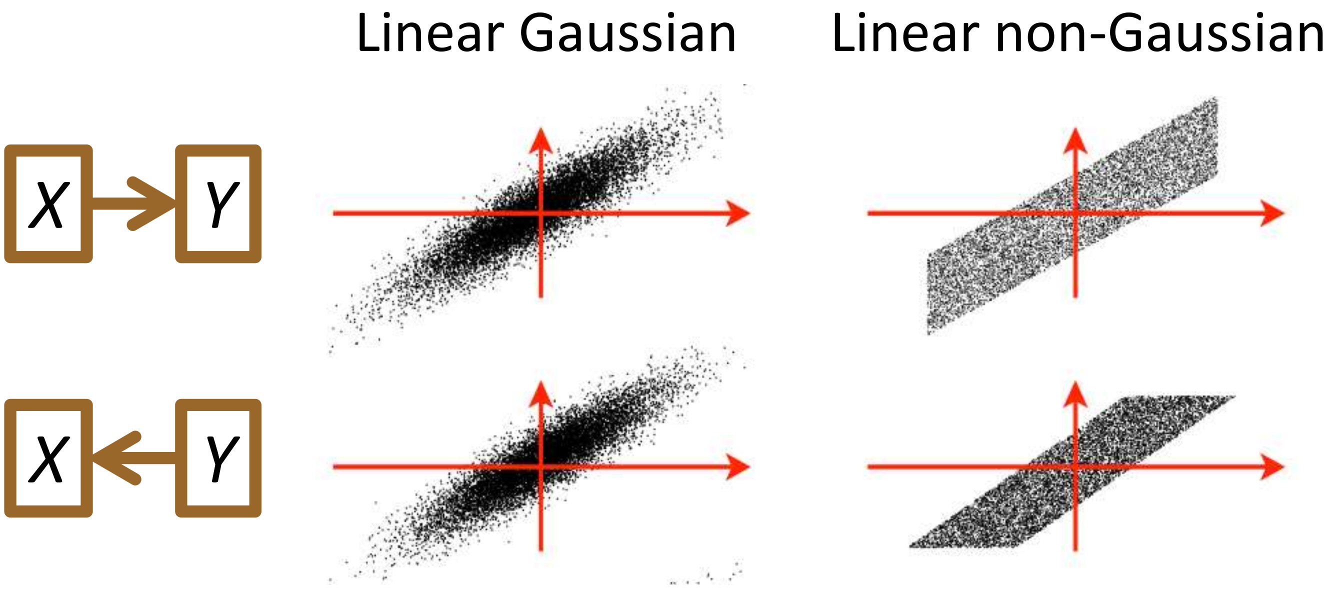
PC Stable, constraint-based search algorithm

- 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**: for example, Chi Square Test
  - **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**: for example, BIC Score
  - **Penalty Discount**: the default is 2; higher values lead to sparser graphs

Some Algorithms Exploit Non-Gaussianality



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.
3. Dataset has both categorical and continuous variables: use the newer algorithms. Discretizing continuous variables is generally not a good idea.
4. Missing values: a challenge to many statistical methods, not just Causal Discovery. In general, you will want to address these before applying discovery algorithms.
5. Selection bias? Measurement error? Consider algorithms designed to address these issues: FCI, GFCI, RFCI. For unmeasured common causes, consider Two-Step.
6. Incorporate knowledge: factors known to cause (or to not cause) other factors.
7. Search procedures generally have no confidence intervals for their results. Model fit statistics are also problematic unless one has very large samples.
8. Consider bootstrapping as a way of assessing how much trust to place on an output.

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).

Activity 2: Analyze a dataset

In separate groups (20 minutes work + 10-15 minutes debrief):

Load your dataset into a Tetrad **Data box**

Brainstorm (< 5 minutes) what causal associations you'd expect to see

Select a causal discovery algorithm (FASK, PC Stable, or FGES) and configure a Tetrad **Search box** and perform the search

Repeat with a second algorithm

Compare results using a Tetrad **Compare box**

Document:

- How the two graphs differ from each other
- How the graphs differ from your group's expectations
- What your team learned

Outputs:

Written briefing (2-4 minutes) on what your group did and learned

Case Study 1 (Complexity Drivers and Project Success) -1

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

Research question: what 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

Two overlapping datasets formed from 81 survey responses

- **AeroDefense**: the 61 survey responses indicating the domain as AeroSpace or Defense
- **AllDomains**: the full set of 81 survey responses, which included representation from Civil Government and Consumer domains

Case Study 1 (Complexity Drivers and Project Success) -2

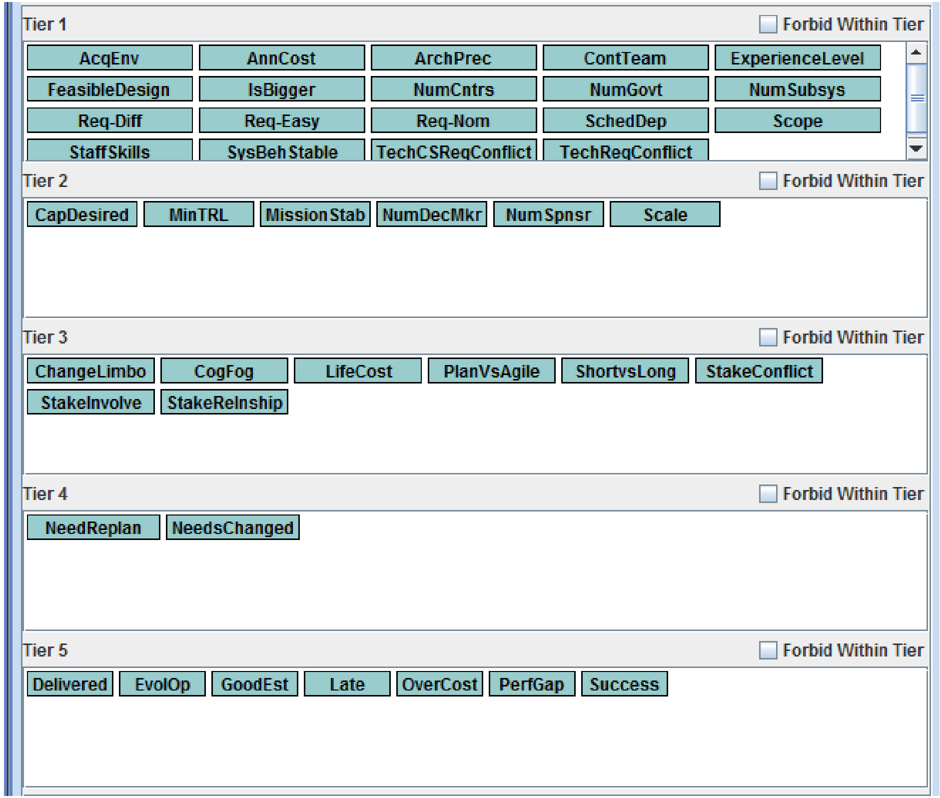
Original result: Three of the complexity variables strongly predicted all outcomes:

Req-Diff	**Difficult requirements are considered difficult to implement or engineer, are hard to trace to source, and 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 your project fit in the following eight attributes, on a scale of (1)Traditional, (2)Transitional, or (3)Messy Frontier? [Translating for] Stakeholder relationships: (1)Relationships stable (2)New relationships (3)Resistance to changing relationships**

Case Study 1 (Complexity Drivers and Project Success) -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.



Case Study 1 (Complexity Drivers and Project Success) -4

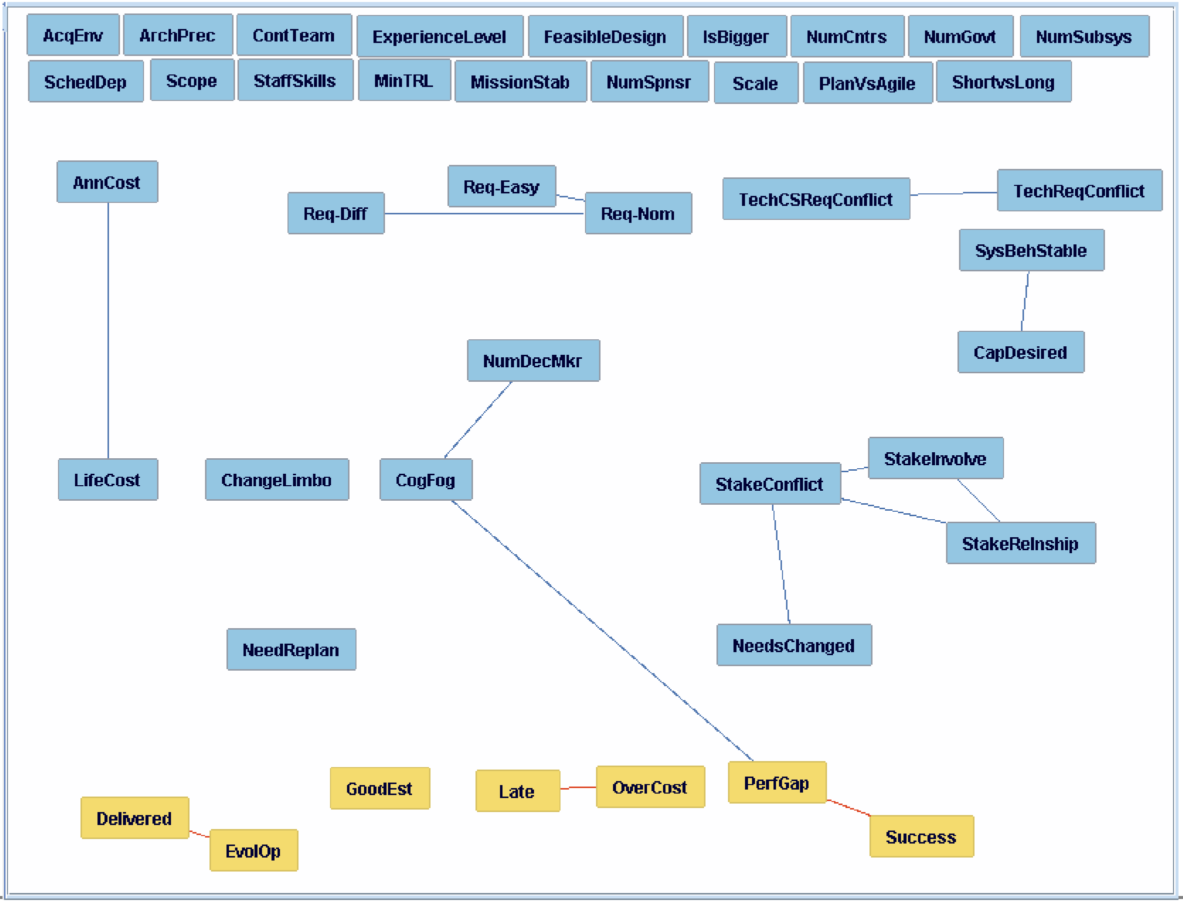
PC-Stable and FGES algorithms were applied to both the Aerospace/Defense projects (61 in number) and the full dataset (81 projects).

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

Outcome (Tier 5) variables are highlighted in yellow.

Note CogFog relationships.

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



Case Study 1 (Complexity Drivers and Project Success) -5

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

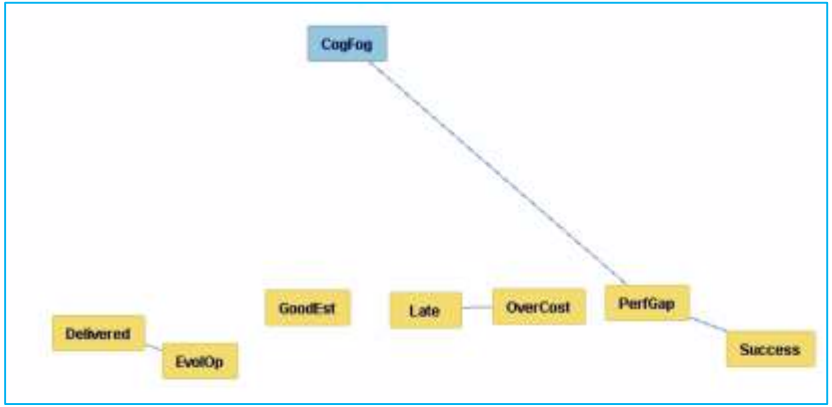
- NumDecMkr directly causes **CogFog**, which directly causes PerfGap, which directly causes (or is caused by) Success.
  - Note that the last two of these are project outcomes.
- The three **stakeholder variables** (StakeConflict, StakeRelnship, StakeInvolve) relate to each other and cause program needs 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.

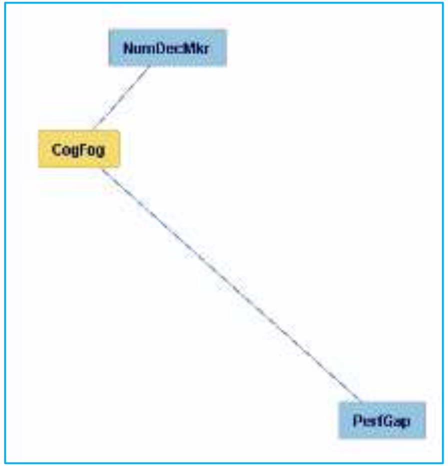
We can further express the causal roles for NumDecMkr and CogFog in terms of text and Markov blankets. See next slide.

Case Study 1 (Complexity Drivers and Project Success) -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:

- 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.

Case Study 1 (Complexity Drivers and Project Success) -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 we predict a low likelihood of meeting project outcomes, one thing we could do is **intervene** to **reduce/streamline the number of decision makers**.
  - This in turn should help **reduce** the amount of **cognitive fog**, which should help **reduce the performance gap** (specified mission-critical features vs. what was actually achieved) and possibly improve project **success**.
- However, from our causal analyses, we found **no evidence** that taking action to:
  - improve stakeholder relationships
  - reduce the number of difficult requirements

  would improve project outcomes.
  - These factors correlate with outcomes (Sheard, 2012), but there's no evidence of causality

These negative results may be simply due to having a relatively **small sample**.

Activity 3: Identify Promising Sources for Datasets

In separate groups (15 minutes):

Select a topic, policy, or research question considered back in Activity 1

Brainstorm:

- What datasets might exist on which to perform causal discovery?
- What sources might already have such datasets or could help in developing them?

Repeat with a second topic, policy, or research question

Document:

- What topics, policies, or research questions were selected
- What datasets exist or might exist; and possible sources

Takeaway: Answering a research question requires good quality data obtained by measuring attributes inferred from the research question. Obtaining good quality data can be hard.

Case Study 2 (Team Dynamics and Project Success) -1

Source: SEI Client, 2014

Research question: what team dynamics factors drive software project success?

Dataset: weekly surveys issued randomly to 30 software staff

- 33 items on a binary scale (Yes / No) representing independent team variables
  - The subset of the 120+ team factors identified by Watts Humphrey that reasonably could change on a weekly basis
- 3 items on a 4 point ordinal scale representing dependent project outcomes:
  - Project Quality, Schedule and Cost

Rationale for Binary Data:

- Staff were overworked; informal piloting indicated survey must not exceed 2-3 minutes of response time
- Staff wanted to point and click with minimal scrolling
- We achieved 90% response rates

Case Study 2 (Team Dynamics and Project Success) -2

Traditional correlation results:

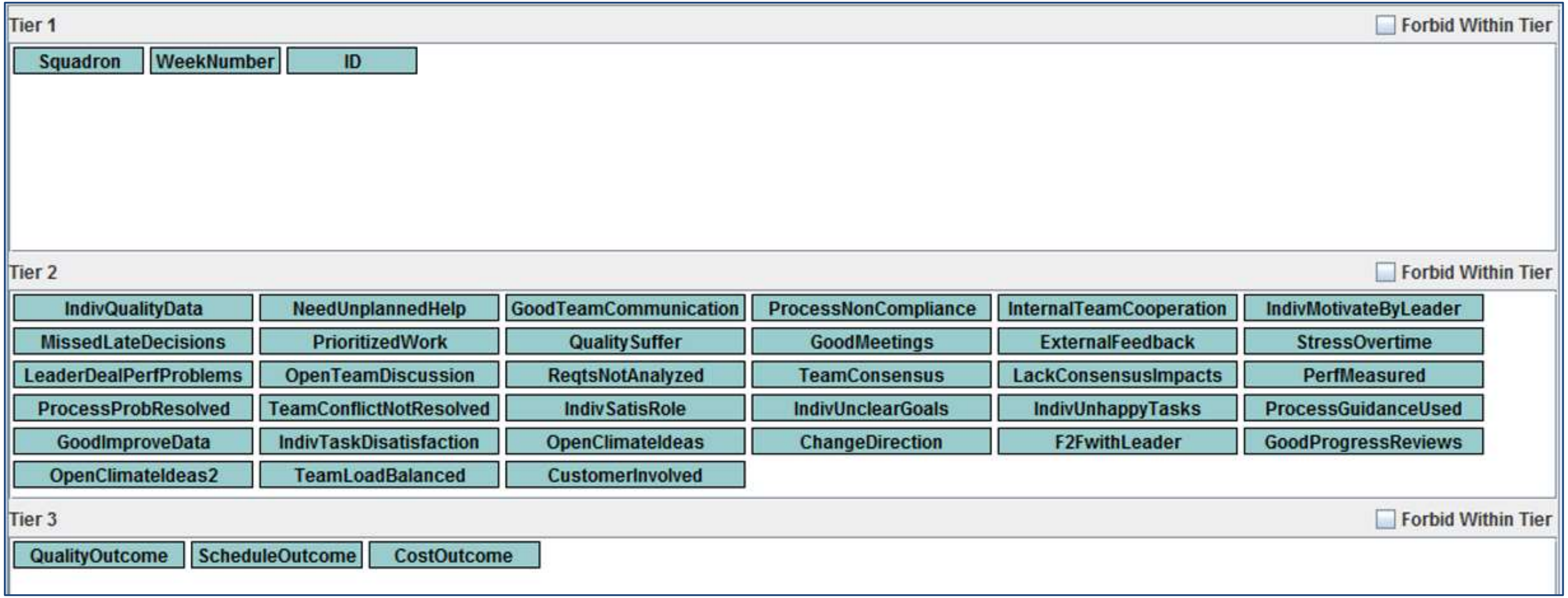
Correlation measures used included Kendall tau-b, Kendall tau-c, Gamma and Spearman's. All were in agreement using the 0.05 cutoff for significance (blue highlighted cells).

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

	QualityOutcome	CostOutcome	ScheduleOutcome
ID			
WeekNumber			
Squadron			
IndivUnclearGoals			
IndivMotivateByLeader			
LeaderDealPerfProblems			
TeamConflictNotResolved			
PerfMeasured			
PrioritizedWork			
ChangeDirection			
QualitySuffer			
IndivUnhappyTasks			
MissedLateDecisions			
IndivSatisRole			
GoodMeetings			
ProcessNonCompliance			
TeamConsensus			
LackConsensusImpacts			
GoodProgressReviews			
GoodImproveData			
OpenClimateIdeas			
ExternalFeedback			
TeamLoadBalanced			
ReqtsNotAnalyzed			
NeedUnplannedHelp			
CustomerInvolved			
ProcessGuidanceUsed			
ProcessProbResolved			
IndivQualityData			
IndivTaskDisatisfaction			
GoodTeamCommunication			
StressOvertime			
OpenClimateIdeas2			
OpenTeamDiscussion			
InternalTeamCooperation			
F2FwithLeader			

Case Study 2 (Complexity Drivers and Project Success) -3

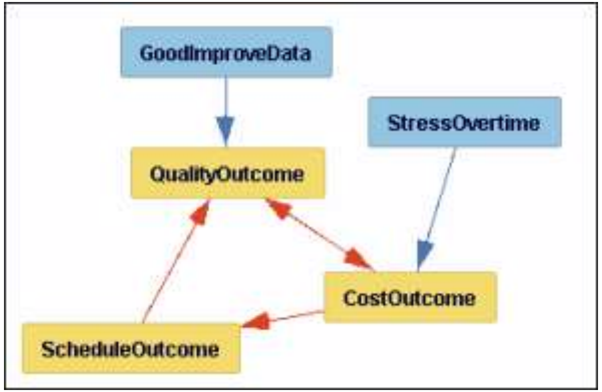
The Tetrad knowledge box for this case study comprised a simple 3-tier approach with three exogenous factors in tier 1, three outcome factors in tier 3 and all remaining factors in tier 2.



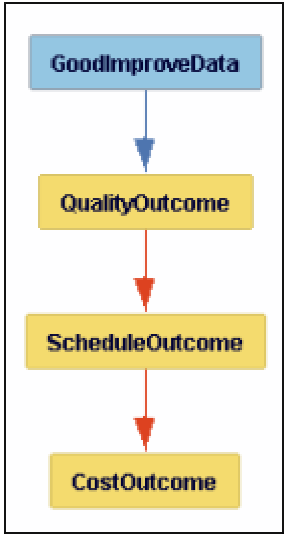
Case Study 2 (Team Dynamics and Project Success) -4

On this slide we show two Markov blankets for the **set of three outcomes**:

1) using PC-Stable, and



2) using FGES.



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.

Case Study 2 (Team Dynamics and Project Success) -5

Although 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,

the causal search discovered:

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

Reflections from a year of causal learning

Work cycle is similar to that for machine/deep learning:

- Pose/revise research **questions**
  - Ensure there are variables representing the outcomes of interest
- Obtain, review, prepare, and analyze **dataset**
  - May require some feature engineering
- Learn more about the **algorithms** to guide in selection and interpretation
  - What assumptions are made and are they met by the dataset?
    - Identically and independently distributed (i.i.d.) data? Censored data? Missing values?
    - Latent confounders? Distributional assumptions?
- Revise approach and **repeat** previous three steps, et cetera.

What has helped:

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

Activity 4: Causal Learning Adoption Roadmap

Purpose: define a causal learning adoption roadmap to: (1) further enlighten the practical measurement community; (2) encourage adoption of causal learning

Before breaking into groups, define:

- Adoption objectives

In separate groups (20 minutes), consult the objectives and brainstorm:

- What activities will help promote understanding and adoption of causal learning?
- Who should perform those activities (assume willingness)?
- By when should those activities be performed (what calendar year; what quarter)?

Document roadmap and missing objectives

Takeaway: A roadmap will help remind us of what we can collectively do to improve the quality of research in the broader communities of which we are a part.

Activity 5: Establish a Causal Learning Adoption User Group

Purposes: (1) identify the group responsible for implementing the CL Adoption Roadmap; (2) provide coaching support for early adopters of causal learning in their research

Brainstorm as a single larger group:

- Timeline for CLAUG
- Charter: intended outcomes, vision, roles and responsibilities
- Reporting results of activities out to larger PSMUG meeting on Friday

Document results and be ready to support the outbriefer for the outbrief on Friday

Conclusion

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

The practical software measurement community has an opportunity to lead the way by adopting causal learning algorithms as part of their research toolkit.

Your inputs and artifacts from this workshop can help facilitate this move forward.

This won't happen without your continued support!

Contact Information

Presenter / Point(s) of Contact

Mike Konrad

Bob Stoddard

Email:

mdk@sei.cmu.edu,
rws@sei.cmu.edu,

Telephone:

+1 412.268.5813 (Mike)

+1 412.268.1121 (Bob)

Other SEI Team Members

Bill Nichols

Sarah Sheard

Dave Zubrow

CMU Contributors

David Danks

Madelyn Glymour

Joe Ramsey

Kun Zhang

USC Contributors

Jim Alstad

Barry Boehm

Anandi Hira