Practical Software and Systems Measurement

Objective Information for Decision Makers

Causal Discovery with Observational Data Workshop
Sept. 19, 2019

Mike Konrad, Bob Stoddard, and Bill Nichols, SEI

Causal Discovery with Observational Data Workshop

Objectives and Intended Outputs of the Workshop

- 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**)
  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.

Workshop Format and 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

Workshop Background

The SEI is leading a three-year **research project** (SCOPE) 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.

Causal Discovery with Observational Data Workshop

Workshop Participants

- Will Hayes, SEI
- Cheryl Jones, US Army Futures Command
- Kim Roye, Galorath

Summary

- Evaluated **effectiveness** of three DoD policies:
  - Actions triggered by **Nunn-McCurdy** breaches
  - Promoting use of **COTS**
  - Promoting migration to the **Cloud**
  - Requiring use of **Ada**
  - Identified drivers of their effectiveness (+ and -)
- Discussed the **potential** of Causal Learning (CL):
  - Sometimes, CL works remarkably well, providing unique insights. But sometimes, results show too little causality. Could this be due to data quality?
- **No group statement yet**
  - The SEI will continue to research CL; would like PSM's help.

Drivers of Policy Effectiveness -1

- Nunn-McCurdy Policy
  - Triggers panic
  - Adds costs irrelevant to program mission but relevant to additional oversight
    - Trades exercised by programs based on external audience versus trades done internally for program benefit
  - How to evaluate effectiveness of policy?
    - Count of breaches with and without policy
    - Without policy, overruns would occur more frequently
    - With policy, needed corrective actions are less likely to be overlooked

Drivers of Policy Effectiveness -2

- COTS Policy
  - Unintended consequences: license costs going up exponentially
  - Wasted effort to: Upgrade to latest versions versus moving to other COTS vendors
  - Lifespans for COTS systems is shrinking down to a year
  - Much work to move to new product or new version
  - Contractor/government have little influence to get needed changes made to COTS
  - Implications for new vulnerabilities

Drivers of Policy Effectiveness -3

- Pushing to Cloud Policy
  - Thousands of products/decisions (thus costs are multiplied)
  - Vulnerability implications (in both directions: new threats vs. reduced existing threats)
  - Misconception: every data item needs to be stored and this is an additional cost incurred indefinitely
  - Ditto entire list for COTS (previous slide)
  - Support for fixing, migrating – are overlooked costs
  - Vendor lock-in (costs more to take data out)
  - Risks to unauthorized access, aggregating multiple sources

Drivers of Policy Effectiveness -4

- Ada Policy
  - Ada came with a heavy learning curve
    - No compiler, not enough tools to support Ada deployment
    - Documentation scant or misleading
  - Reading code to infer functionality was much easier (encapsulation and generics)
    - But Ada didn't support all principles for true OO
  - Compiler checks of interfaces ensured smoother integration
  - Deeper issues remained hidden; over-focused on syntax
    - Support for real-time was incompletely handled
  - Code optimization achieved very late
  - Hard to overcome demand for niche program'g languages
  - Programming cultures were like the "Wild West"

How would we evaluate Causal Learning? [+'s]

- Can identify **overlooked drivers** of program success
  - Kazman: architectural pattern violations can lead to vulnerabilities
    - Found causes for rarely-experienced events not found in an earlier analysis
  - Causal Learning can work remarkably well even when # of variables dwarfs # of cases
- Eliminates incorrect or less significant influences
  - Sheard: evidence for this causal pathway:
    # of stakeholders -> cognitive fog -> technical performance; and not for # of difficult requirements or quality of stakeholder relations
  - Overcomes Simpson's Paradox
    - Causal signals unique to a subpopulation can get buried
    - Acting on the wrong drivers can make things worse!
- Introduces a **new viewpoint** on what the data is revealing about underlying data generation process: what causes/influences what?
  - May **deepen** the conversation with the information-need recipient/decision maker

How would we evaluate Causal Learning? [–'s]

- Causal search is easy to execute, but resulting graphs are often **ambiguous** (sensitive to algorithm settings)
  - Analyzing COCOMO datasets, Size consistently found to drive Effort

    But not many other factors share this consistency!
  - Could be a limitation *not* of the *algorithms* but of the *data*; one that the algorithms cannot fully overcome (GiGo)
- Causal Learning *can* require **lots of data** if signals are weak
  - Covariance matrix, used by score-based CL algorithms (such as FGES), can require a lot of data to estimate accurately

❖ But systematic use of **bootstrapping** can reduce ambiguity of (and thus improve our confidence in) results

❖ Including for weak causal signals

Next Steps/Action Items

- SEI will continue **evaluation** of Causal Learning (CL) Technology: sharing insights gained
  - Continue **engagement** with the PSM Community:
  - Reporting CL experiences in identifying drivers of project success and costs, **successes and failures**
- We seek **joint collaboration opportunities** with those in **PSM Community**
  - Not necessary to share your data—only CL results
  - We can provide CL training at distance/in-person
  - Help develop new science—and better estimates!