



Synthesis of Causal Discovery and Machine Learning – Questions Posed

Robert Stoddard, Principal Researcher, SEI

Mike Konrad, Principal Researcher, SEI

Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213

Document Markings

Copyright 2018 Carnegie Mellon University. All Rights Reserved.

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.

The view, opinions, and/or findings contained in this material are those of the author(s) and should not be construed as an official Government position, policy, or decision, unless designated by other documentation.

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.

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

This material may be reproduced in its entirety, without modification, and freely distributed in written or electronic form without requesting formal permission. Permission is required for any other use. Requests for permission should be directed to the Software Engineering Institute at permission@sei.cmu.edu.

Carnegie Mellon® is registered in the U.S. Patent and Trademark Office by Carnegie Mellon University.

DM18-1275

Agenda

SEI SCOPE Research Focus

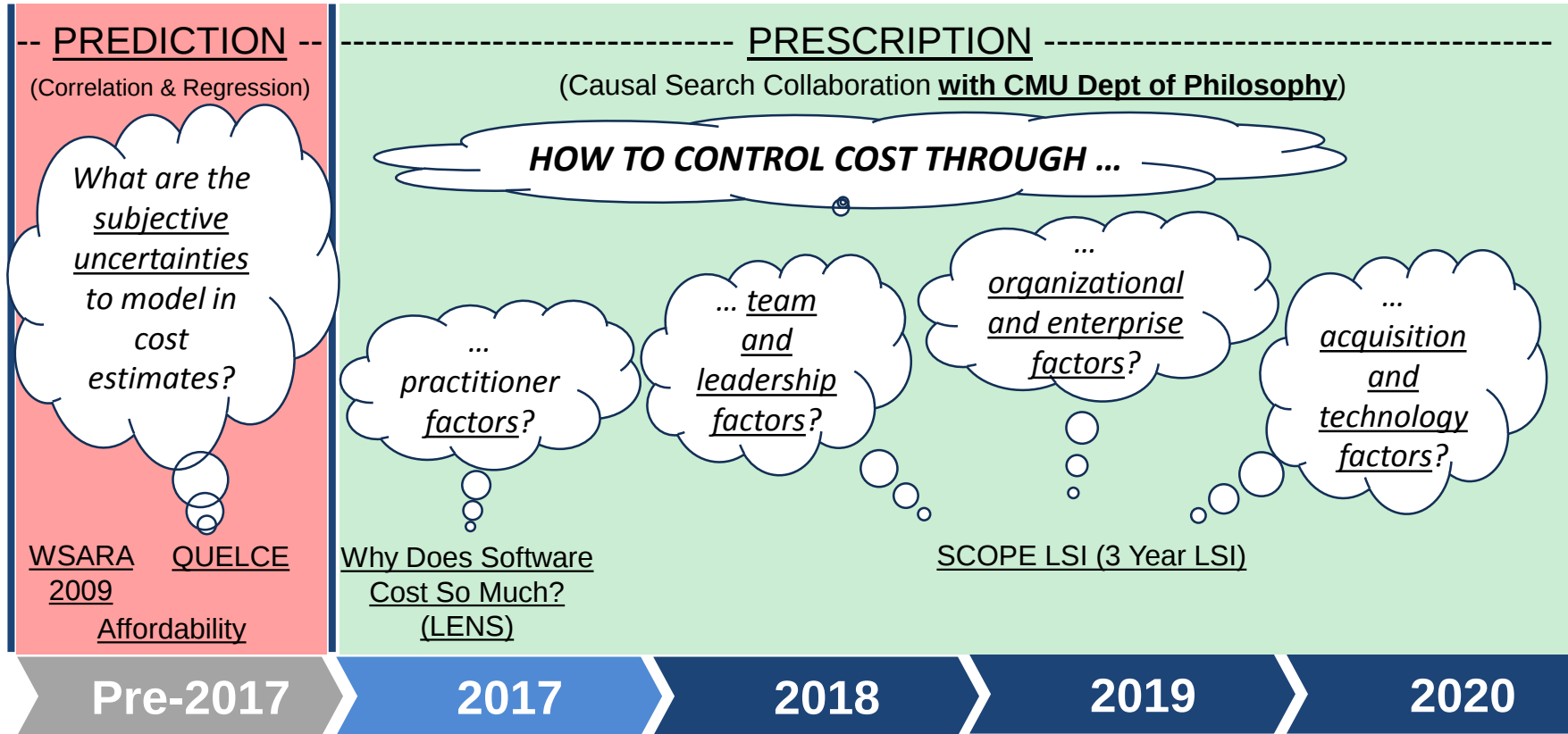
Use of BayesiaLab

Causal Learning

Comparison of ML and CL outputs

Questions Posed for Future Collaboration?

Context of Causal Models for Software Cost Control (SCOPE)



Agenda

SEI SCOPE Research Focus



Use of BayesiaLab

Causal Learning

Comparison of ML and CL outputs

Questions Posed for Future Collaboration?

Use of BayesiaLab

1. Supervised machine learning (ML) with cost, schedule and quality as targets
2. Multi-variate outlier analysis
 - a) Aid in data quality analysis
 - b) Possible data segmentation strategies
3. Data imputation, when needed
4. Prediction of “what-if” scenarios of factors against outcomes
5. Classifier to assign probability of a binary outcome (e.g. good vs bad outcomes)
6. Diagnostic of most likely factors associated with a given outcome
7. All in support of DoD cost estimation and affordability analysis

Agenda

SEI SCOPE Research Focus

Use of BayesiaLab

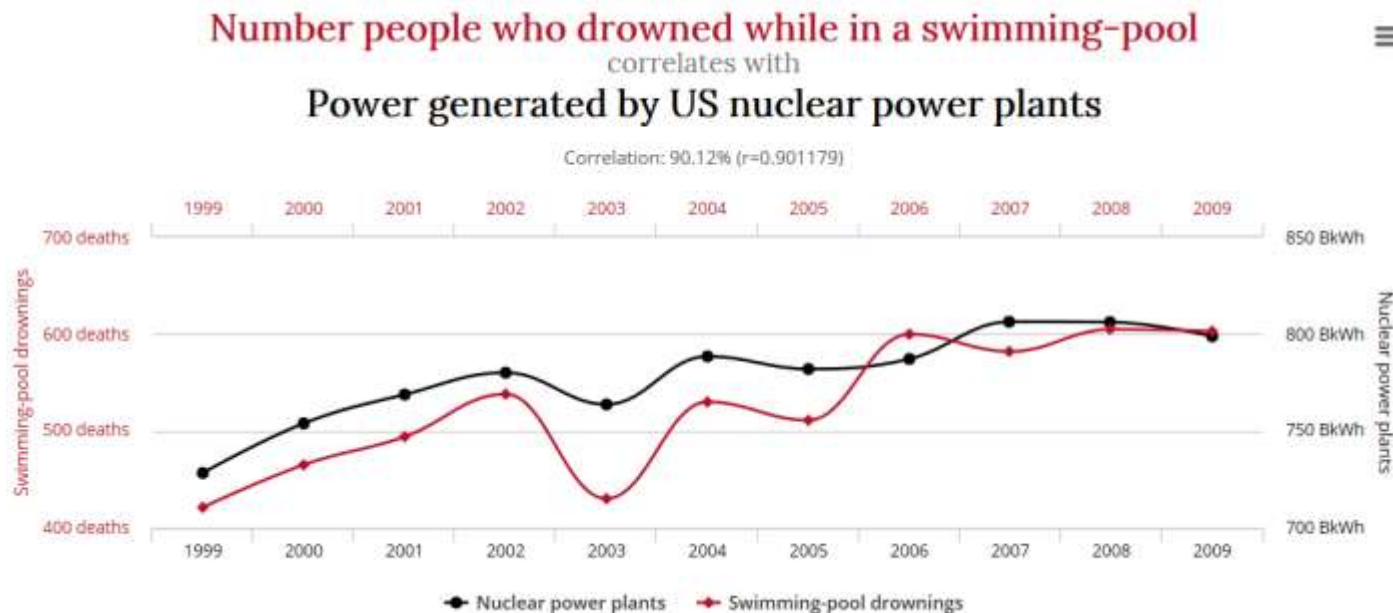


Causal Learning

Comparison of ML and CL outputs

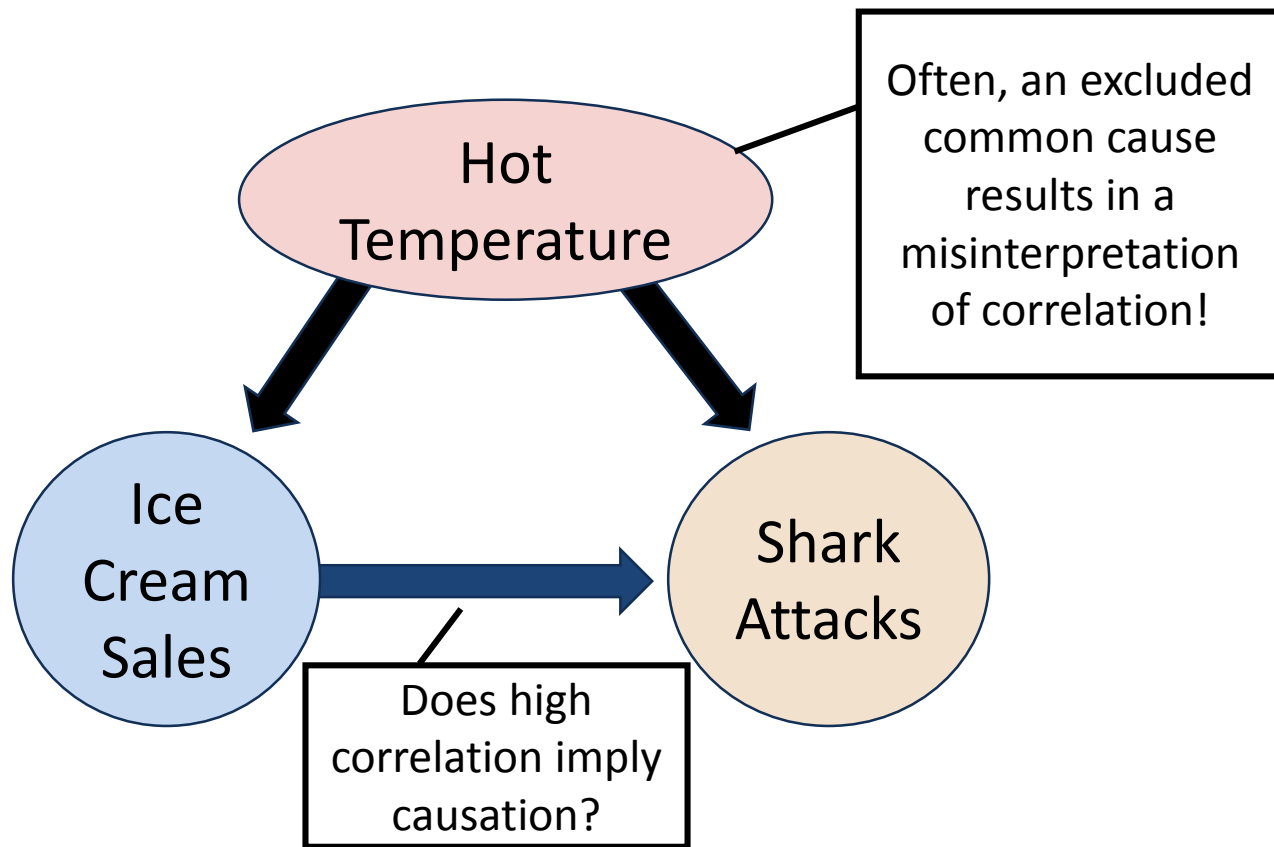
Questions Posed for Future Collaboration?

Why Do We Care about Causation?



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

More about Misinterpreting Correlation!



Regression & ML benefit from a Structural Causal Model!

Regression and ML may be fooled by spurious association!

Need a structural causal model (SCM) representing our theory and context

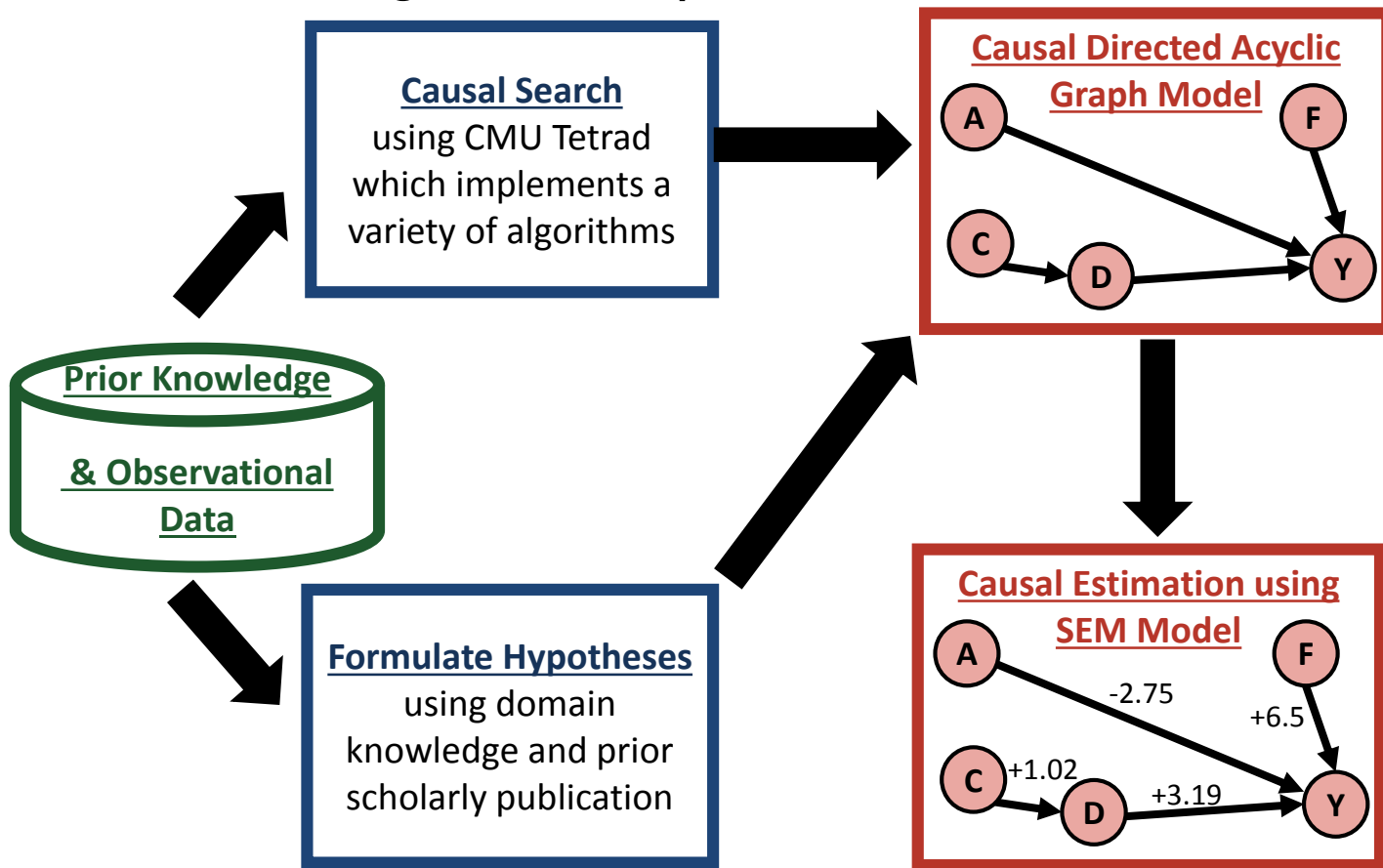
Need to determine which paths are causal versus non-causal

Must block non-causal paths

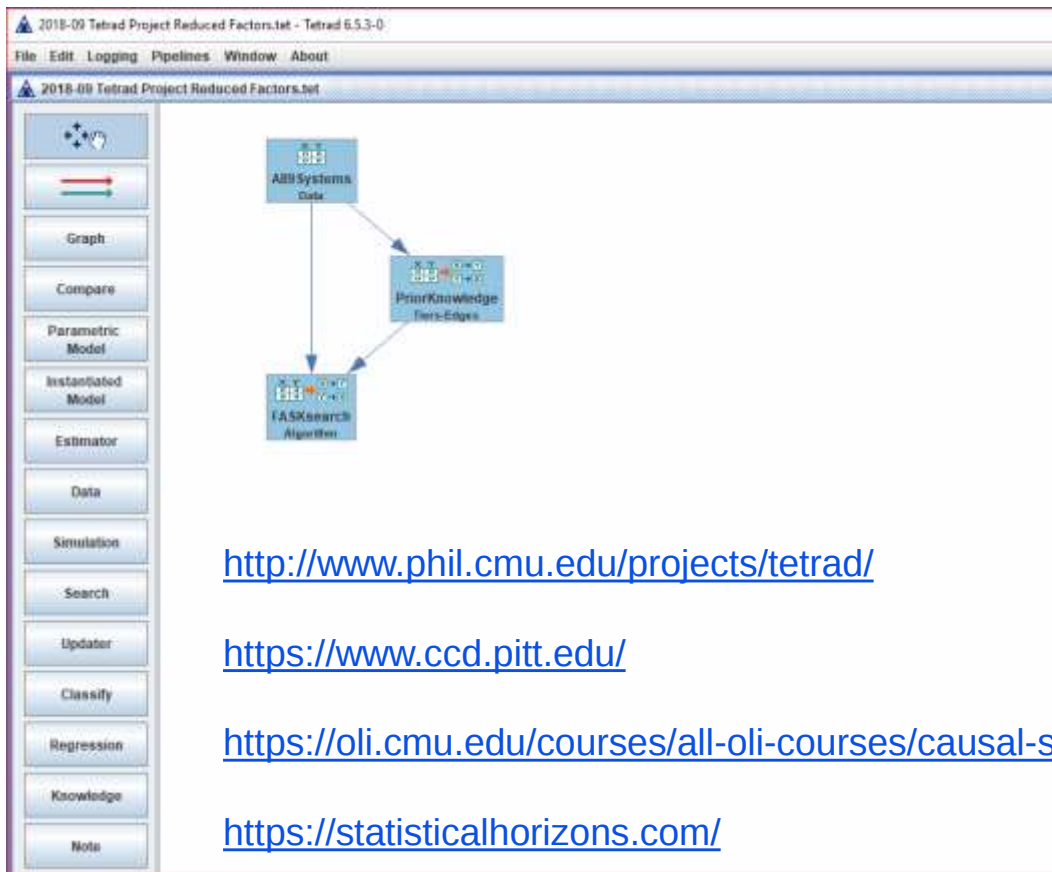
Then conduct regression and ML with the correct set of factors!

Suitability of the model depends on the SCM!

The Causal Learning Landscape



Conduct Causal Search using Tetrad



<http://www.phil.cmu.edu/projects/tetrad/>

<https://www.ccd.pitt.edu/>

<https://oli.cmu.edu/courses/all-oli-courses/causal-statistical-reasoning/>

<https://statisticalhorizons.com/>

A View of the Data File Loaded into Tetrad

All9 Systems (Data)

File Edit Tools

All 9 for Tetrad-v010.csv

	C1	C2	C3	C4	C5	C6	C7	C8	C9
	AgeMonths	NumDev	LOC	NumBugs	BugChurn	NumCyclic...	NumModul...	NumUnsta...	NumImpro...
1	71.0000	8.0000	491.0000	18.0000	241.0000	8.0000	2.0000	3.0000	1.0000
2	35.0000	5.0000	270.0000	10.0000	329.0000	167.0000	1.0000	1.0000	4.0000
3	52.0000	2.0000	58.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
4	42.0000	1.0000	47.0000	2.0000	13.0000	0.0000	0.0000	0.0000	0.0000
5	49.0000	1.0000	10.0000	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000
6	36.0000	2.0000	103.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000
7	54.0000	2.0000	29.0000	2.0000	0.0000	0.0000	0.0000	0.0000	0.0000
8	75.0000	8.0000	163.0000	13.0000	134.0000	0.0000	1.0000	3.0000	0.0000
9	74.0000	2.0000	15.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000
10	57.0000	2.0000	26.0000	1.0000	16.0000	22.0000	0.0000	0.0000	0.0000
11	48.0000	4.0000	81.0000	2.0000	6.0000	0.0000	1.0000	0.0000	0.0000
12	39.0000	1.0000	30.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
13	49.0000	2.0000	46.0000	3.0000	36.0000	0.0000	0.0000	0.0000	0.0000
14	46.0000	3.0000	34.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000
15	75.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000

Done

Prior Knowledge Entered into Tetrad

The screenshot shows the 'PriorKnowledge1 (Tiers and Edges)' window in the Tetrad software. The 'Edges' tab is selected. The window is divided into three sections for different tiers. At the top, there is a 'File' menu and a '# Tiers =' field set to 3. Below this, there is a 'Not in tier:' section. The first tier, 'Tier 1', has a 'Forbid Within Tier' checkbox checked and contains three items: 'AgeMonths', 'LOC', and 'NumDev'. The second tier, 'Tier 2', also has a 'Forbid Within Tier' checkbox checked and contains four items: 'NumCyclicDepend', 'NumImproperInherit', 'NumModularityViolations', and 'NumUnstableInterface'. The third tier, 'Tier 3', has a 'Forbid Within Tier' checkbox unchecked and contains two items: 'BugChurn' and 'NumBugs'. At the bottom, there is a 'Done' button and a note: 'Use shift key to select multiple items.'

PriorKnowledge1 (Tiers and Edges)

File

Tiers Other Groups Edges

Not in tier: # Tiers = 3

Tier 1 ☒ Forbid Within Tier

AgeMonths LOC NumDev

Tier 2 ☒ Forbid Within Tier

NumCyclicDepend NumImproperInherit NumModularityViolations

NumUnstableInterface

Tier 3 ☐ Forbid Within Tier

BugChurn NumBugs

Use shift key to select multiple items.

Done

Causal Learning Algorithms

Constraint-based: Calculate independences in the data and do “backwards inference”

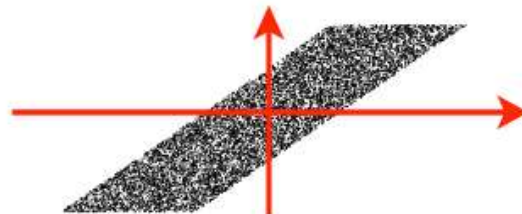
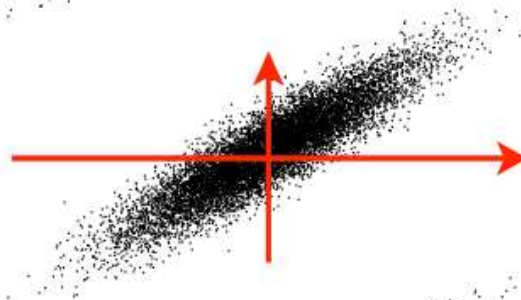
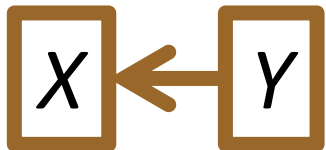
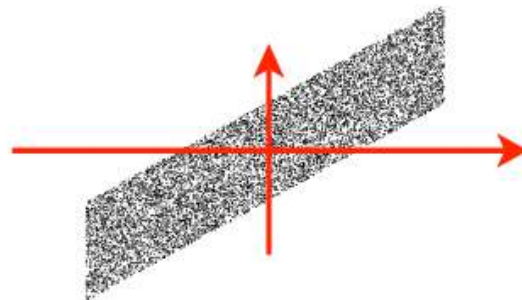
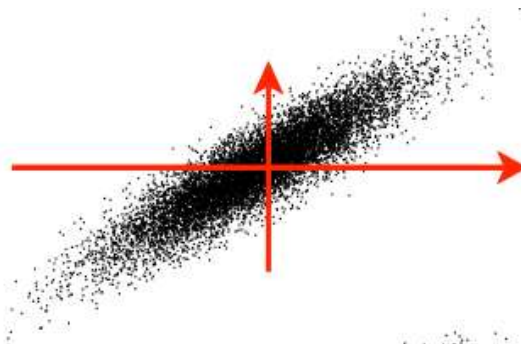
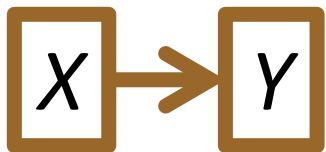
Score-based (Bayesian): Calculate the likelihood of different DAGs given the data

Hybrid: Use constraint-based to get “close,” then Bayesian search around neighborhood

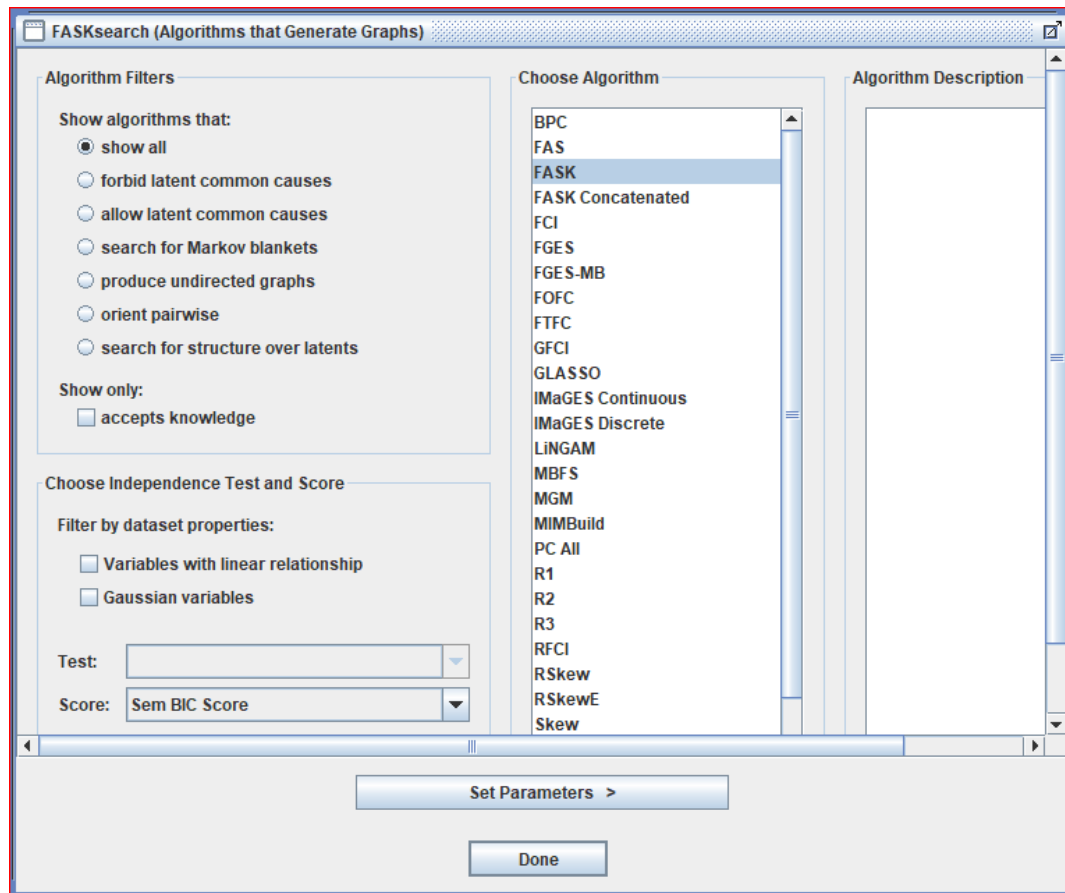
Some Algorithms Exploit Non-Gaussianity

Linear Gaussian

Linear non-Gaussian



Using FASK Search with Associated Parameters



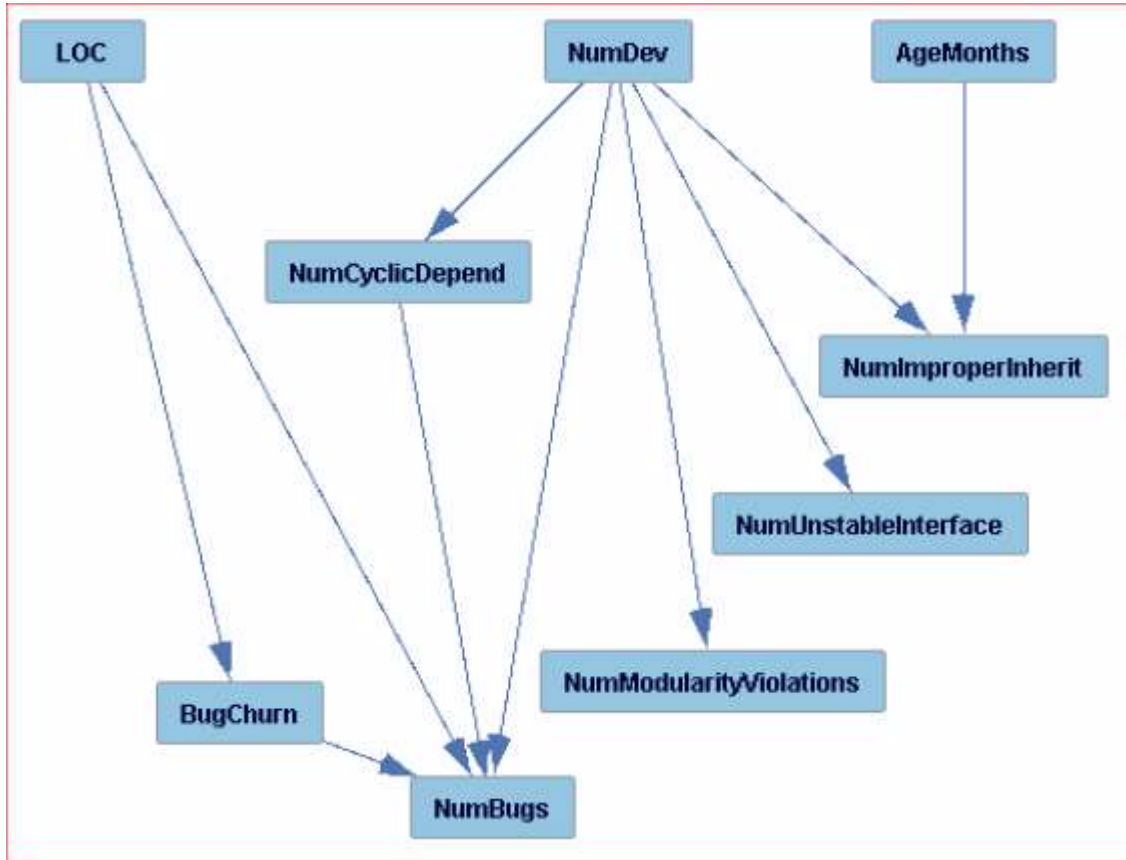
Additional FASK Search Parameter Settings

The screenshot shows a software window titled "FASKsearch (Algorithms that Generate Graphs)". Inside, there is a section labeled "FASK Parameters" with the following settings:

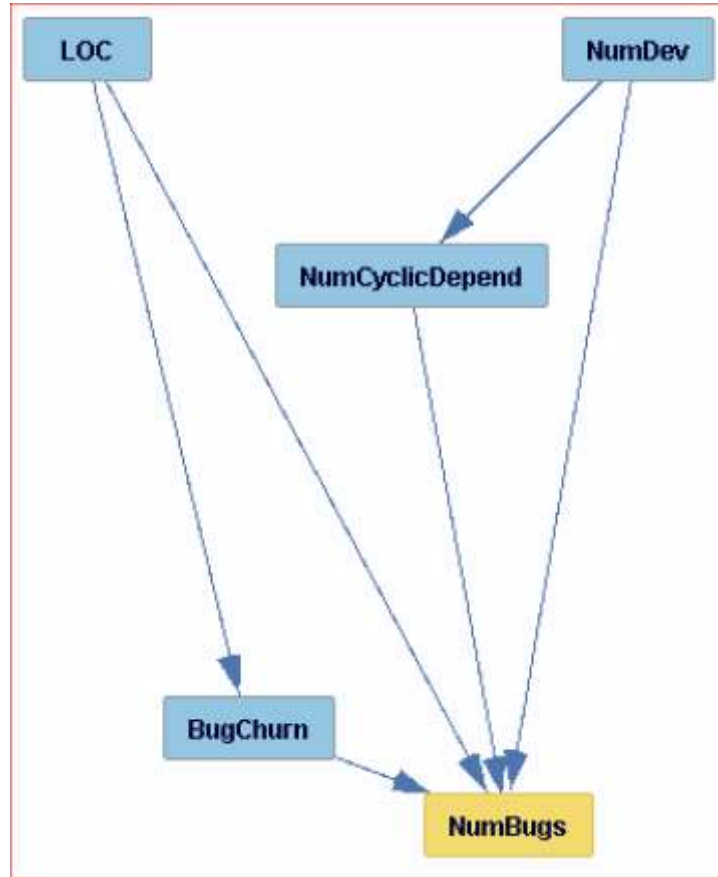
- Penalty discount (min = 0.0): 2
- Maximum size of conditioning set (unlimited = -1): -1
- Alpha orienting 2-cycles (min = 0.0): 1.0E-6
- Threshold for including extra edges: 0.3
- Threshold for judging negative coefficient edges as X->Y (range (-1, 0)): -0.2
- Yes if adjacencies from the FAS search should be used: ☒ Yes ☐ No
- Yes if adjacencies from conditional correlation differences should be used: ☒ Yes ☐ No
- The number of bootstraps (min = 0): 0
- Ensemble method: Preserved (0), Highest (1), Majority (2): 1
- Yes if verbose output should be printed or logged: ☐ Yes ☒ No

At the bottom of the window, there are three buttons: "< Choose Algorithm", "Run Search & Generate Graph >", and "Done".

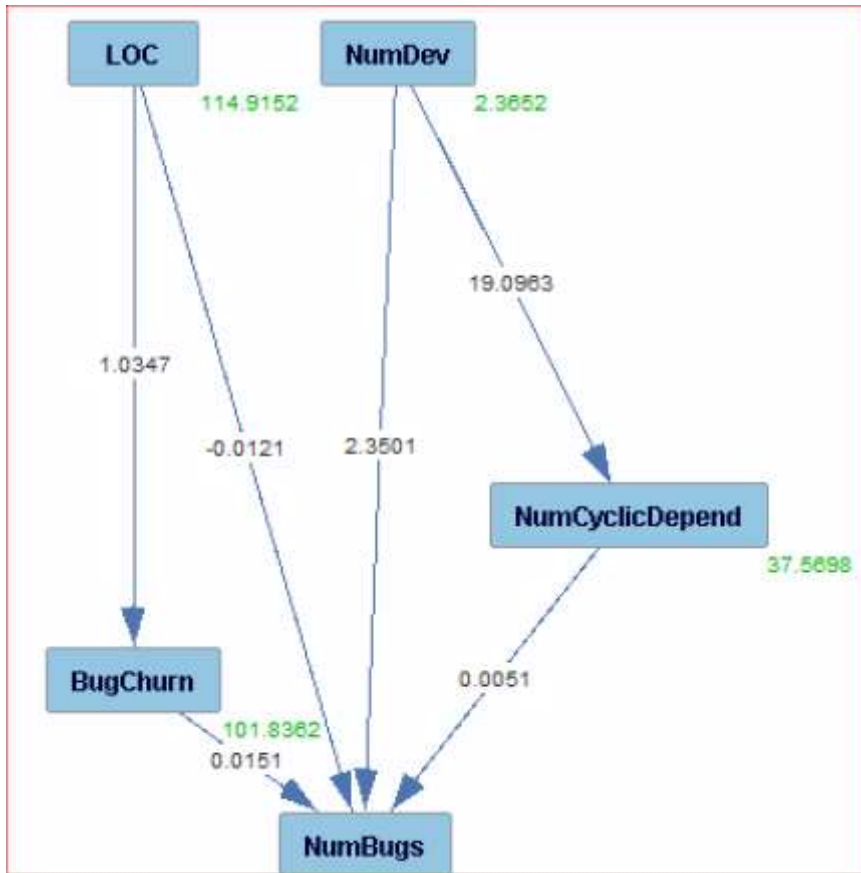
Causal Structure Graph Result



Markov Blanket of the NumBugs Factor



Traditional SEM Results from Tetrad



File Parameters Layout

Graphical Editor

Tabular Editor

Degrees of Freedom = 4

Chi Square = 2358.0099

P Value = 0.0000E0

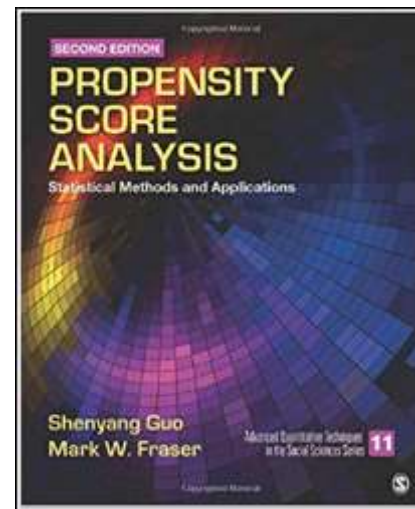
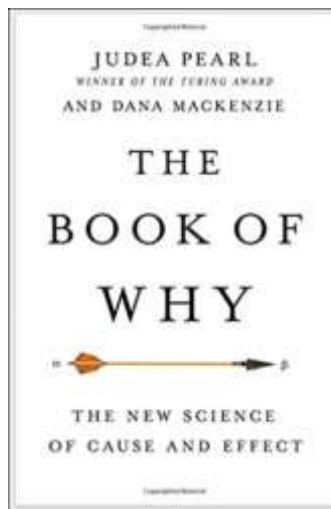
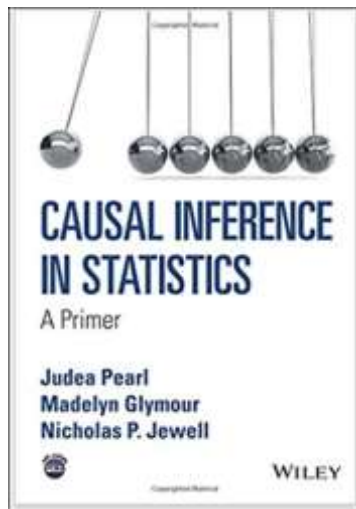
BIC Score = 2321.5678

CFI = 0.9907

RMSEA = 0.2550

Additional Causal Learning Topics

1. Algorithms operating on the Structural Causal Model (see Judea Pearl, 2018, “The Book of Why”)
2. Propensity Scoring (see Shenyang Guo and Mark W. Fraser, 2014, “Propensity Score Analysis”)
3. Instrumental Variables (see Felix Elwert, publications on Instrumental Variables)



Agenda

SEI SCOPE Research Focus

Use of BayesiaLab

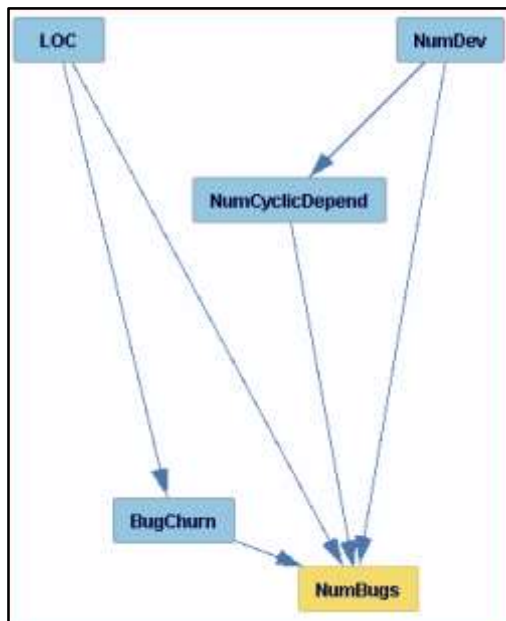
Causal Learning

 Comparison of ML and CL outputs

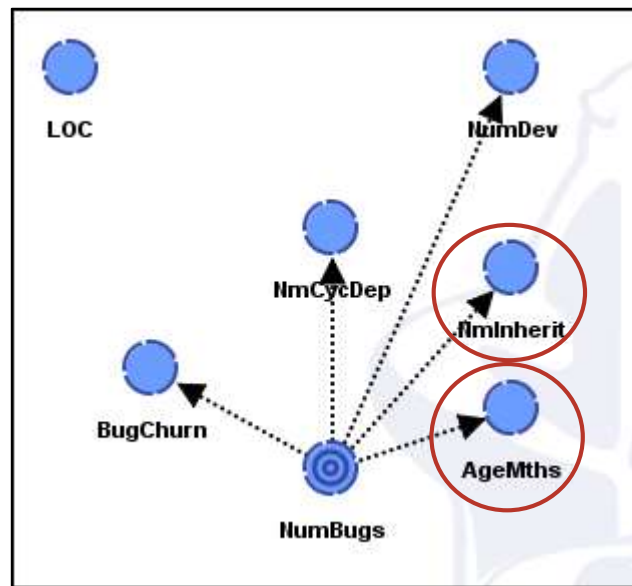
Questions Posed for Future Collaboration?

ML and CL Graph Structures May Be Different

CL Markov Blanket



ML Markov Blanket



Agenda

SEI SCOPE Research Focus

Use of BayesiaLab

Causal Learning

Comparison of ML and CL outputs

 Questions Posed for Future Collaboration?

When ML and CL Graph Structure Results Differ?

1. Choose to instantiate the Tetrad causal structure in BayesiaLab as a PSEM?
2. Use BayesiaLab to conduct Pearl graph surgery or Jouffe's likelihood matching for causal modeling?
3. Pursue metrics such as Average Causal Effect (ACE) and Total Causal Effect (TCE)?

Opportunities to Integrate ML & CL? - 01

1. Can a ML association graph structure result inform a CL causal search?
2. For extremely large datasets and # variables, would ML require significantly less computer time than a CL causal search? If so, could ML serve as a pre-screen of a CL causal search?
3. Could ML graph structure results inform opportunities for research into new CL causal search algorithms?
4. Could/should CL causal search be combined with ML graphical results for a new, superior output?

Opportunities to Integrate ML & CL? - 02

5. Is there a possible superior understanding obtainable from graphical structural results of both ML and CL?
 - a) Can differences between the two graphs provide insight?
 - b) Can commonality across the two graphs provide insight?
 - c) More generally, is there greater knowledge of combining Shannon Information Theory with Causal Theory?
6. Can combined use of ML and CL graphical structures enable an improved method of “stitching together” separate, but overlapping results towards a more holistic result?

Conclusion

We are seeking research collaboration in two ways:

1. Collaboration and data access for software project cost estimation and control, and
2. Collaboration to gain insight and answer the questions posed in this presentation

Contact Information

Presenter Contact Information



Dr. Mike Konrad
Principal Researcher,
SEI / CMU
mdk@sei.cmu.edu
1-412-268-5813



Robert Stoddard
Principal Researcher,
SEI / CMU
rws@sei.cmu.edu
1-412-268-1121