SEI-CMU Collaboration Using Causal Learning

Robert W. Stoddard, Principal Researcher, SEI

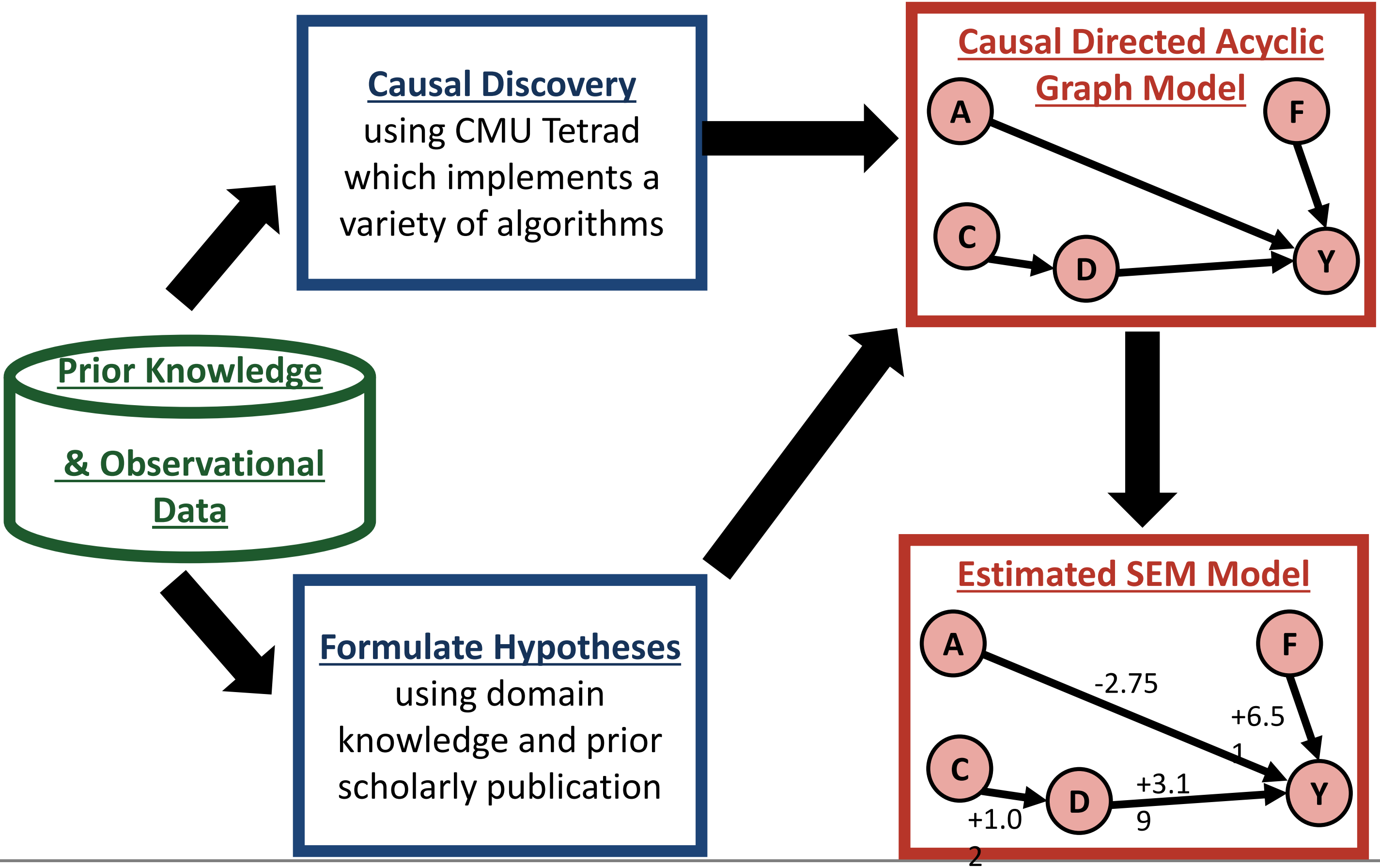
Dave Danks, Department Head & L.L. Thurstone Professor of Philosophy and Psychology

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. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

DM19-0321

The Causal Learning Landscape



Correlation vs. Causation

- **Correlation** = Using an *observation* about one factor to *predict* another
  Causation = Using an *intervention* on one factor to *influence* another

Predict warfighter performance given training?

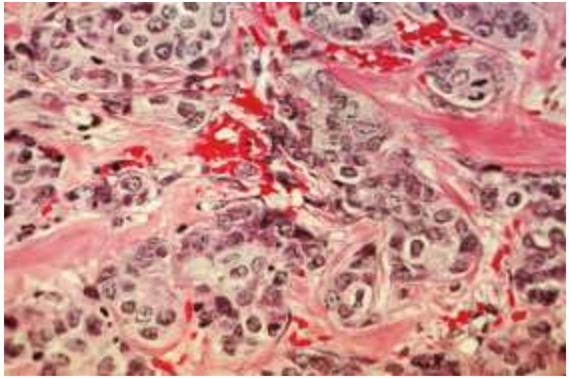
OR

Change training to improve performance?

Predict development of cancer cells?

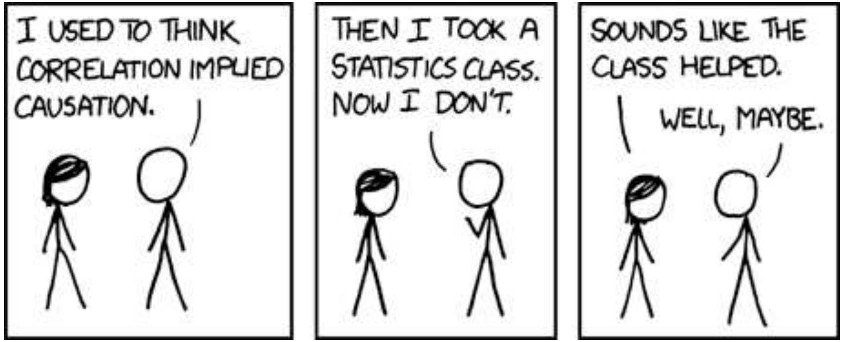
OR

Give treatment to slow growth of cancer?



Causation vs. correlation

Statistics slogan: *Causation ≠ Correlation*

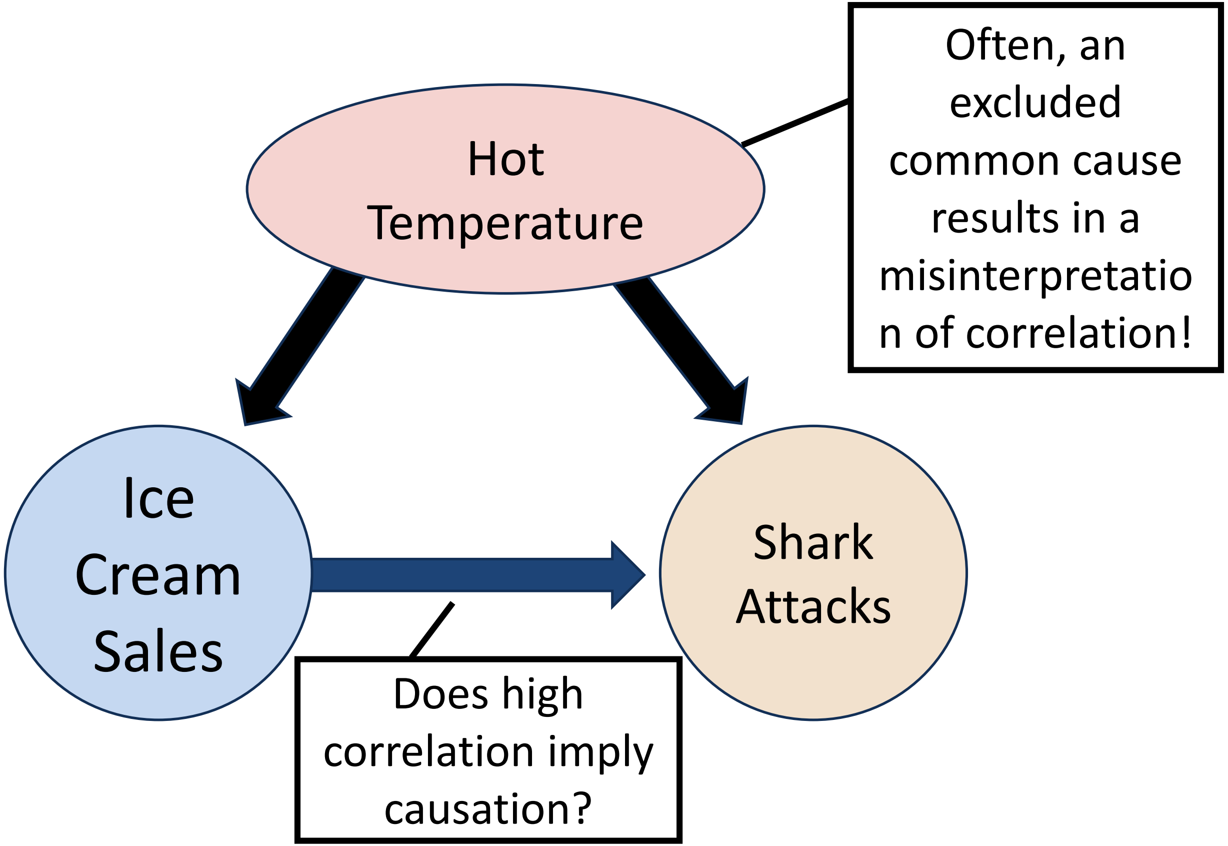


Credit: https://xkcd.com/552/

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

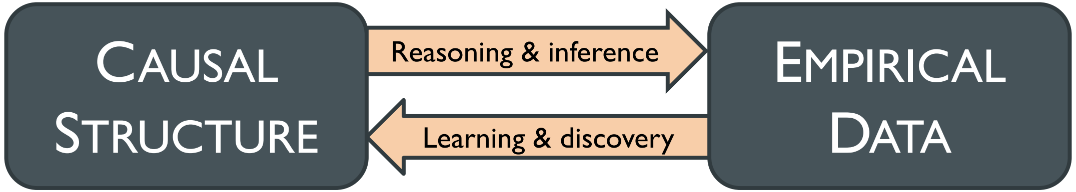
Prof. David Danks' summary: "Correlation is a noisy indicator of causation."

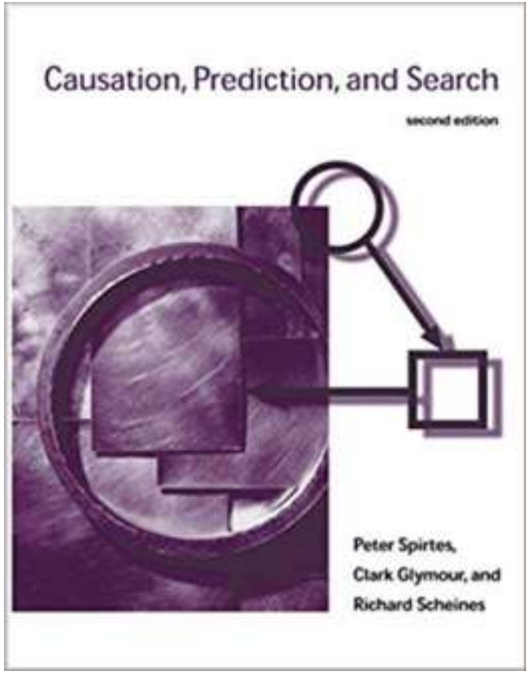
More about Misinterpreting Correlation!



CAUSAL DISCOVERY AT CMU

- 30 years of research at CMU, mostly in Philosophy
- *Basic strategy:*





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

Causal Learning has become practical

Sewall Wright Path Models (1920's)

Structural Equation Models (1930's)

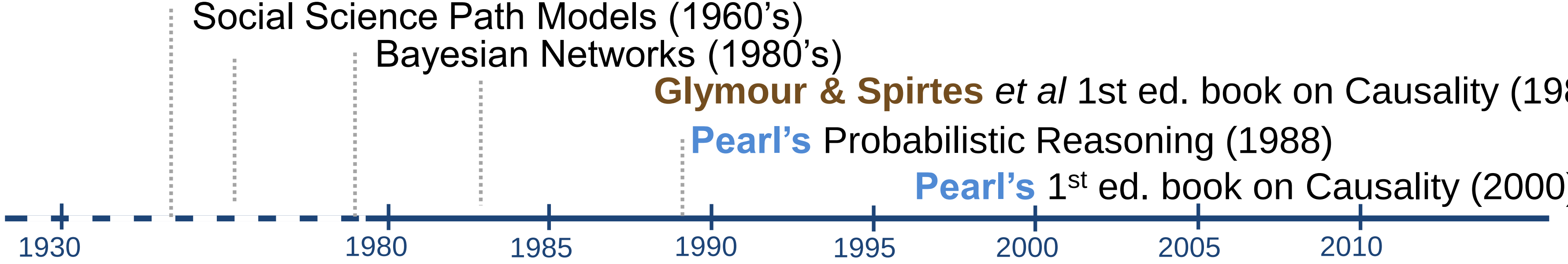
Social Science Path Models (1960's)

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)

Morgan Counterfactuals & Causality (2007)

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

Morgan Counterfactuals & Causality (2014)

Peters Elements of Causal Inference (2017)

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:

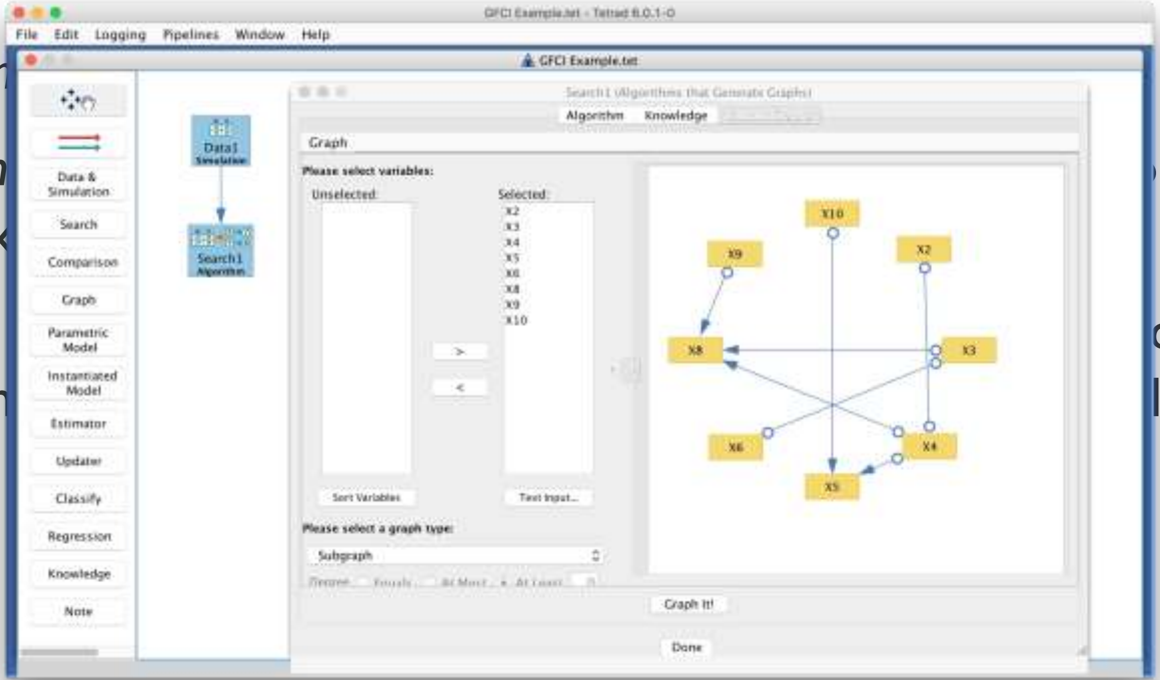
http://www.ccd.pitt.edu/training/presentation-videos/

CMU OLI - Causal and Statistical Reasoning

http://oli.cmu.edu/courses/future/causal-statistical-reasoning/

CAUSAL DISCOVERY AT CMU

- *Represen*
- *Reasonin* r complex
- *Learning* d plausible causal m ledge



https://www.ccd.pitt.edu/ http://www.phil.cmu.edu/projects/tetrad/

Causal discovery in action

- *Which genes regulate flowering time in Arabidopsis thaliana?*
- Causal discovery on observations of gene activation:
  9 (of 21,326) genes are good candidates
  - using only 47 samples…
- Greenhouse study that used knockout variants:
  4/9 were actual regulators

Context of SCOPE Research

-- PREDICTION --

(Correlation & Regression)

What are the subjective uncertainties to model in cost estimates?

WSARA 2009

QUELCE

Affordability

---------------------------- PRESCRIPTION ---------------------------------------

(Causal Search Collaboration **with CMU Dept of Philosophy**)

HOW TO CONTROL COST THROUGH …

… practitioner factors?

… team and leadership factors?

… organizational and enterprise factors?

… acquisition and technology factors?

Why Does Software Cost So Much? (LENS)

SCOPE LSI (3 Year LSI)

Pre-2017 | **2017** | **2018** | **2019** | **2020**

Why Do We Care About Causal Modeling?

Controlling costs requires knowing which "independent factors" actually *cause cost* outcomes, so that we may change cost in a predictable manner.

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

We must move beyond correlation to causation, if we want to make use of cause and effect relationships

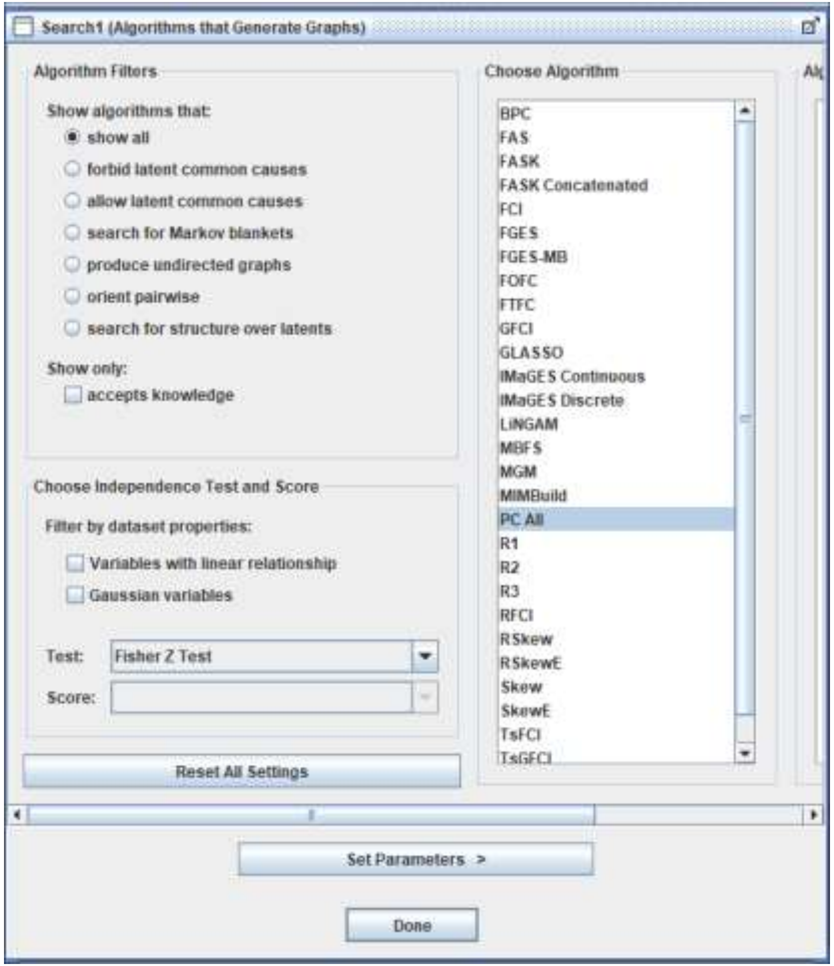
We can now evaluate causation without expensive and difficult experiments

Establishing causation with observational data remains a vital need and a key technical challenge, but is becoming more feasible and practical.

Causal Discovery Algorithms -1

Multiple approaches to searching a dataset:

1. **Constraint-based**: determined from independences in the data
2. **Score-based** (**Bayesian**): determined from likelihood calculation of different DAGs given the data
3. **Hybrid**



Initial SCOPE Causal Search Results

Controlling Size: Only 2 of 4 code size measures appear causal on effort and quality

Controlling Complexity: Only 1 of 3 factors appears causal on performance and quality

Controlling Architecture Violations: Only 1 of 4 violation factors appears causal on quality

Controlling Team Performance: Only 1 of 20+ factors appears causal on quality and cost

Causal search may provide useful feedback:

1) *Presence of causal links*
2) *Absence of causal links*

Moderate Future: Causal Learning for Simulation and Test

Problem
Lack of accredited simulators

Technical Challenge
Experts unsure of the expected result for a given simulated scenario

Research Questions
1. Scale up metamorphic testing to test very complex DoD systems?
2. Machine learning to identify metamorphic relations for testing?
3. Causal learning to drive metamorphic relations testing?

Moderate Future: Causal Learning for Sustainment

Problem
Unscheduled maintenance creates unacceptable costs

Technical Challenge
Traditional statistical approaches helpful, but insufficient

Research Questions
1. Machine learning of engine sensor and control data improve scheduled maintenance?
2. Causal learning integrated with machine learning add value?

Long Term Future: (Causal Learning Examples)

Affordable
- Acquisition practice improved using causal models
- Cost estimates and budget execution using causal models
- Simpler but more effective ROI models based on causal factors (e.g. Model Based Engineering, Architecture practice, Technical Debt)

Trustworthy
- Causal factors threatening cyber defenses
- Causal factors limiting resilience
- CL combined with ML tools for more affordable and trustworthy SW technologies (e.g. DOD initiative in Digital Engineering)
- Expected behavior from autonomous systems   (e.g. "*Explainable AI*"; Jensen, UMass)

Capable
- Causal drivers of workforce performance
- SW architecture strategies and tactics driving system performance
- More efficient experimentation of technical solutions
- Increased realism of complex system simulation
- Autonomous systems controlling consequences
- Machine learning with human-like intelligence (e.g. *"Strong AI"; Pearl, "The Book of Why"*)

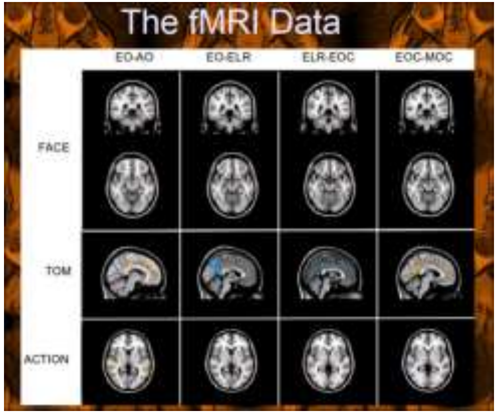
Timely
- Causal structures from DevOps information stream to control process and lifecycle
- Agile causal systems situationally prescribe practices aligned with goals
- Project risks controlled through causal structures of project parameters

Benefits of CMU Collaboration

1. World class expertise and coaching

2. Search algorithms & Tetrad updates
   (IMAGES, FASK, Multi-FASK, Bootstrapping, Cyclic search)

3. Sharing search approaches from other domains
   (classification approach with fMRI causal results)



4. Students test new algorithms and updates to Tetrad

5. Research-to-practice and practice-to-research cycles

Opportunities Everywhere!

1. Reproduce and confirm/clarify prior research
2. Create actionable models of performance
3. Root cause analyze problems using Big Data
4. Reduce time for experimentation through causal search of historical data
5. Use causal structures to better understand and test system behavior
6. Use causal structures to explain and assure AI solutions

Contact Information

Presenter / Point(s) of Contact

David Danks

Bob Stoddard

Email:

ddanks@cmu.edu
rws@sei.cmu.edu

Telephone:

+1 412.268.8047 (David)

+1 412.268.1121 (Bob)

Other SEI SCOPE Team Members

Mike Konrad

Bill Nichols

Sarah Sheard

Dave Zubrow

Other CMU SCOPE Contributors

Madelyn Glymour

Joe Ramsey

Kun Zhang

USC SCOPE Contributors

Jim Alstad

Barry Boehm

Anandi Hira