Why GBSD Should Consider Machine Learning (ML) and Causal Learning (CL)

Robert W. Stoddard

Principal Researcher, SEI, CMU

ASQ Fellow

Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213

Document Markings

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.

DM19-1231

Agenda

BLUF

The Opportunity of Machine Learning

The Opportunity of Causal Learning

Questions

BLUF

1. GBSD is definitely a software system
2. As such, many large streams of sensor and other data will be available
3. This software system offers many benefits but also faces critical challenges and threats
4. Machine learning will digest and model these large streams of data, if GBSD captures and stores the data for later modeling
5. Causal learning can actually determine cause-effect relationships in data as opposed to spurious correlation, thereby offering results not possible via traditional statistics and machine learning
6. SEI is poised to contribute to the GBSD Program Office in specifying the data capture and storage needed, and the adoption of these new technologies with oversight and leadership with the contractors
7. Almost all “ilities” represent low-hanging fruit opportunities for ML and CL

Agenda

BLUF

The Opportunity of Machine Learning

The Opportunity of Causal Learning

Questions

What is Machine Learning?

Basically a more sophisticated form of correlation, association and pattern recognition

Can accommodate and needs Big Data, e.g. large volumes of streams of data

Forms include:

1. Unsupervised machine learning, e.g. to explore relationships between factors
2. Supervised machine learning, e.g. to predict outcome(s)
3. Deep Learning (DL), e.g. a layered network to better identify and learn patterns
4. Reinforcement Learning (RL), e.g. a network that helps learn actions to maximize a reward, sort of an optimization approach
5. Generative Adversarial Networks (GANs), e.g. a set of networks that can interact with each other to generate additional data based on what each network learns from the other

GBSD Data Stream Opportunities

Machine learning could help increase understanding and prediction of the following data streams:

1. Sensor and failure data for next generation safety, reliability, security, resilience, testability, maintainability, and almost all the other pertinent “ilities”
2. Data streams captured from both weapon system, ground systems, command and control systems
3. Would be superior to traditional linear regression and multiple regression modeling
4. Uses non-parametric approaches that are not suspect due to statistical assumptions
5. More capable for earlier feedback, warning and prediction ability based on the extensive multivariate space

REQUIREMENT: Even if GBSD does not immediately implement ML models, care must be taken to specify the capture and storage of data streams for future use

Example Usage Abounds!

SEI has used ML in many scenarios to include but not limited to:

- Analysis of flight data recorder parametric data to predict engine health and anticipate degradation leading to catastrophic failure with the intent to enable predictive and scheduled maintenance (Tinker AFB Propulsion Directorate ~ 65K jet engines)

- Analysis of US Navy torpedo simulation and in-water testing to:
  - 1) understand and predict torpedo behavior and performance under different conditions, and
  - 2) help accredit simulators under complex scenarios where the experts cannot confidently identify the oracle for test cases

Agenda

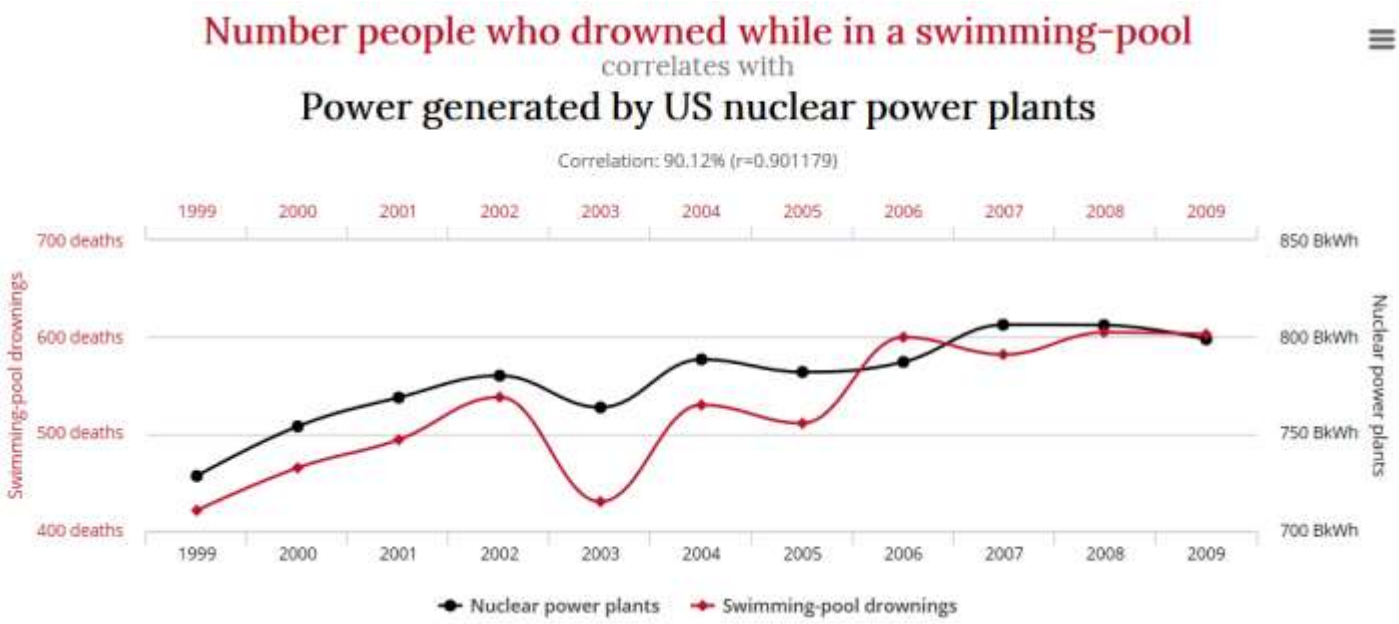
BLUF

The Opportunity of Machine Learning

The Opportunity of Causal Learning

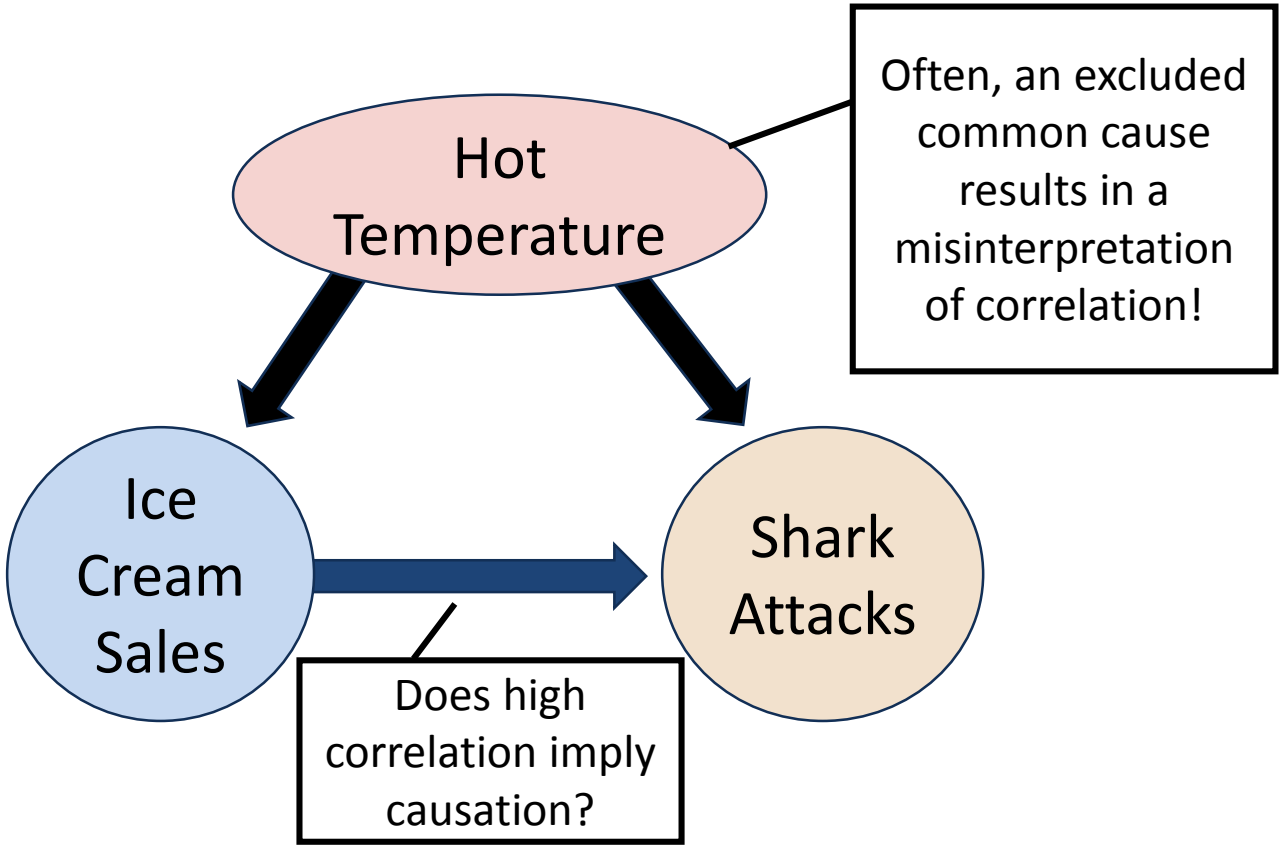
Questions

Why Do We Care about Causation?



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

More about Misinterpreting Correlation!



Regression must be interpreted in context of a DAG!

Correlation, hence regression, may be fooled by spurious association!

Before jumping into regression, we need a Directed Acyclic Graph (DAG) representing our context

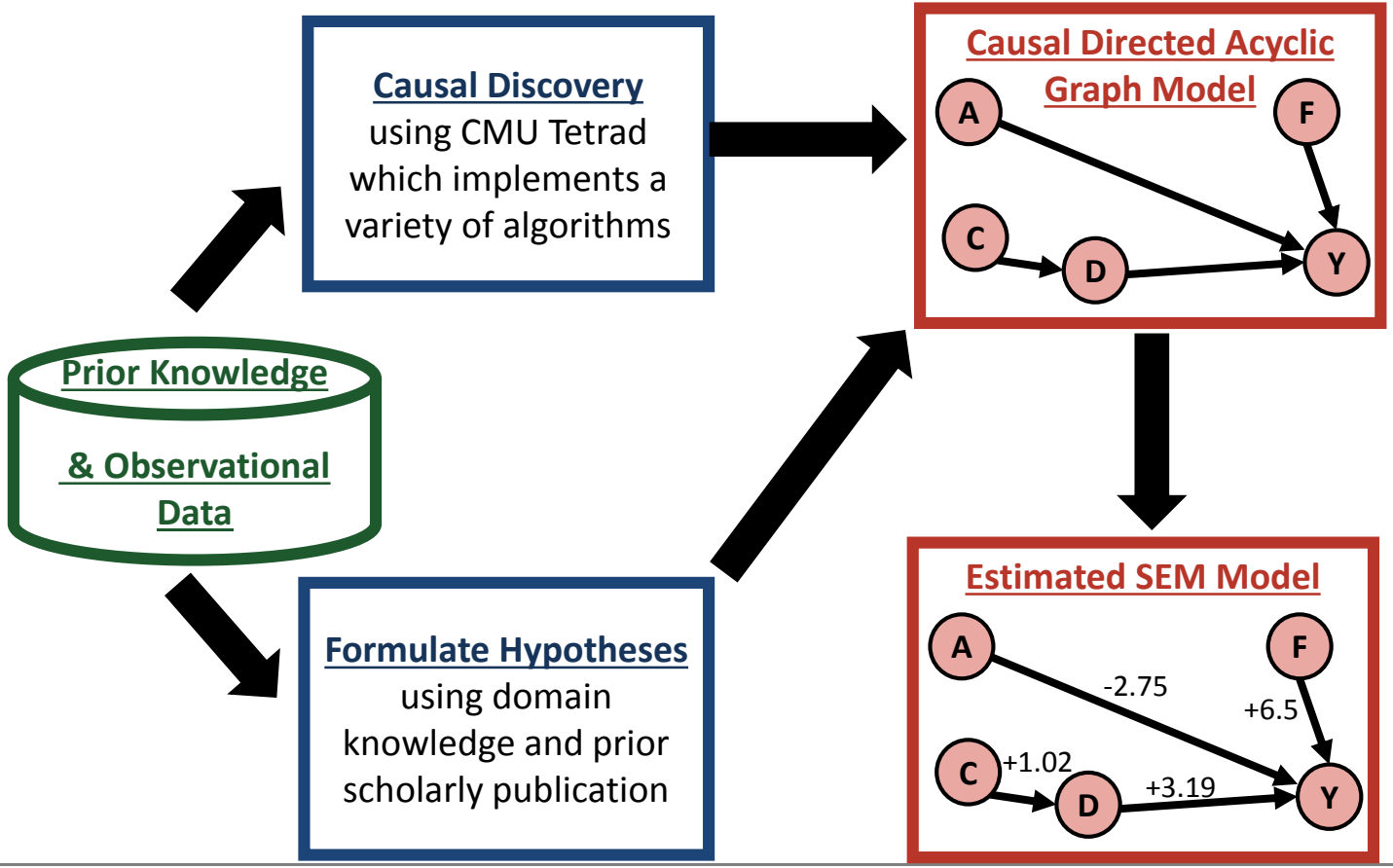
We then need to determine which paths are causal and which are spurious.

We then must block spurious correlation paths.

Lastly, we then conduct regression with the correct set of factors!

Remember, context of the DAG determines the suitability of the regression model!

The Causal Learning Landscape



Use Causal Learning to Answer Counterfactural Questions

The most robust method to answer causal and counterfactual questions from a causal inference standpoint is a causal algebra called “Do-Calculus”

From Wikipedia: “Counterfactual history, also sometimes referred to as virtual history, is a form of historiography that attempts to answer "what if" questions known as counterfactuals. Black and MacRaild provide this definition: "It is, at the very root, the idea of conjecturing on what did not happen, or what might have happened, in order to understand what did happen.“”

Example Counterfactural Questions (Agile Process Example)

1. What would the project outcome be if agile practice xyz was not used?
2. What would the quality level be if test type xyz was not performed?
3. What would the delivery date be if a different number of sprints were employed?
4. What would the project outcome be if the customer interaction and feedback were doubled in intensity?
5. What would the project outcome be if the software teams were co-located rather than geographically separated?

Example Usage Abounds!

SEI has recently applied causal learning in the following applications:

- Applied causal learning to distinguish spuriously correlated parameters from true cause-effect relationships in jet engine performance and failure data as a way to leap forward in reliability, performance and safety engineering
- Enabled causal footprints of healthy versus unhealthy jet engines which proved quite helpful in predicting engine health
- Enabled identification of the causal chain of events in the data which led to a specific engine failure
- Offered insight to intervention tactics to increase engine life and inform scheduled maintenance

Future Opportunities from Causal Research

1. Answer counterfactual questions, reducing the need for expensive if not prohibitive experimentation
2. Downsize the set of factors that might be experimented on
3. Identify when existing models or machine learning solutions can be trusted versus not trusted, thereby identifying when models should be “retrained”
4. Handle and recover from selection and survivor bias in data
5. Inform different data fusion questions in real-time, embedded systems
6. Enable semi-autonomous and autonomous systems to be more intelligent by understanding causal relationships in the real world
7. Capture and model much rich knowledge from experts in the face of a “retirement bubble” occurring in the workforce
8. Better support cybersecurity, reliability, and nuclear surety needs

SEI Collaborates with Causal Experts from Universities

CMU (Dr. Richard Scheines, Dr. David Danks, Dr. Kun Zhang, Dr. Peter Spirtes, Dr. Clark Glymour, Dr. Joe Ramsey)

UCLA (Dr. Judea Pearl)

Colombia University (Dr. Elias Bareinboim)

Univ of Wisconsin at Madison (Dr. Felix Elwert)

USC (Dr. Barry Boehm and PhD students)

Agenda

BLUF

The Opportunity of Machine Learning

The Opportunity of Causal Learning

Questions

Contact Information

Presenter / Point(s) of Contact

Robert Stoddard

Email:

rws@cmu.edu

Telephone:

+1 412.268.1121 desk

+1 724.263.7113 cell

Other SEI Causal Research Team Members

Mike Konrad

Bill Nichols

Dave Zubrow

Other CMU Causal Research Contributors

David Danks

Madelyn Glymour

Joe Ramsey

Kun Zhang

USC Causal Research Contributors

Jim Alstad

Barry Boehm

Anandi Hira