Why Does Software Cost So Much? Towards a Causal Model

Bob Stoddard, Principal Researcher and PI

Mike Konrad, Principal Researcher

Why Does Software Cost So Much? Towards a Causal Model

Carnegie Mellon University
Software Engineering Institute

Why Does Software Cost So Much? Towards a Causal Model
© 2017 Carnegie Mellon University

[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution.

2

Why Does Software Cost So Much? Towards a Causal Model

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

Carnegie Mellon University
Software Engineering Institute

Why Does Software Cost So Much? Towards a Causal Model
© 2017 Carnegie Mellon University

[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution.

3

SEI's Continuing Focus on Improving Cost Estimation

Research & Development

2011-2016: QUELCE reducing uncertainty in early lifecycle software cost estimation

2013-14: Investment model for software sustainment

2016-17: Why Does SW Cost So Much? Causality

The SEI has a long track record of cost related research.

Proof of Concept

2014: QUELCE workshop with JSpOC Space Program

Exploit causal modeling on observational data.

Program

2012-13: F-22: SEI-led "should cost" analysis of software modernization

2014-15 NAVAIR adoption of investment model for sustainment

2016-17 SRDR Factbook published

Future Activity:

FY18-20: An Integrated Causal Model for Software Cost Prediction and Control (SCoPe)

Causal Sensitivity Dashboard Prototype

Service & DoD Cost Centers; NAVAIR and USAF Logistic Sites

| 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 - 2020 |

Why Does Software Cost So Much? Towards a Causal Model
© 2017 Carnegie Mellon University

[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution.

4

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.

Carnegie Mellon University
Software Engineering Institute

Why Does Software Cost So Much? Towards a Causal Model
© 2017 Carnegie Mellon University

[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution.

5

Significant Progress Toward Practicality

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)

The science supporting our research has just emerged.

| 1930 | 1980 | 1985 | 1990 | 1995 | 2000 | 2005 | 2010 |

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/

The SEI has connections to these leading researchers.

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)
Morgan Handbook Social Science Causal Inference (2014)

Research Methodology

Causal modeling will drive science-based cost estimation.

Causal Discovery

using Tetrad, which implements a variety of algorithms

Prior Knowledge

Observational Data

Data: size, task, duration, defects, skills, experience, process, tools

Formulate Hypotheses
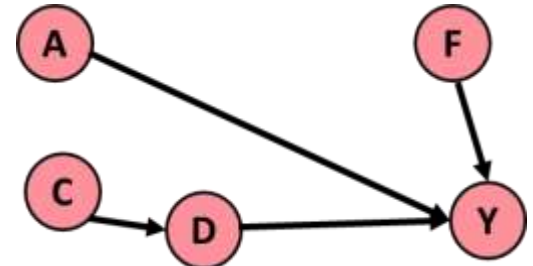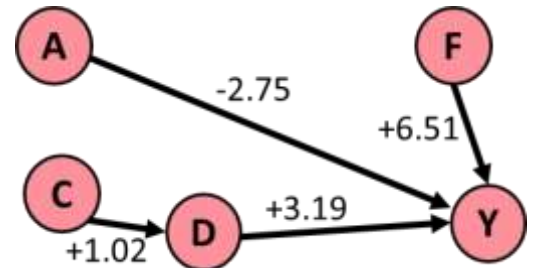
using domain knowledge and prior scholar publication

Causal Model (DAG)

Estimated Model (SEM)

Carnegie Mellon University
Software Engineering Institute

Why Does Software Cost So Much? Towards a Causal Model
© 2017 Carnegie Mellon University

[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution.

7

Integrating a Full Causal Model

COCOMO Data

Vendor 1 Data

Vendor 2 Data

Vendor 3 Data

SRDR Data

TSP/PSP Data

CSIAC Data

Tetrad Learning
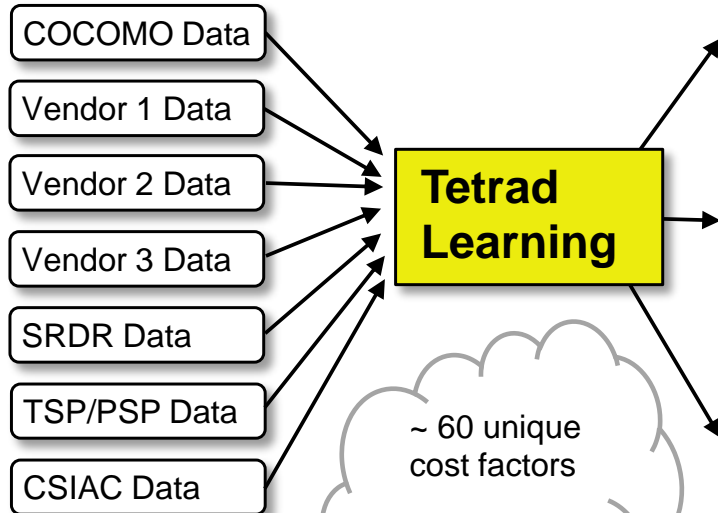
~ 60 unique cost factors

15+ cost relationships to evaluate

The SEI will collaborate with key parties to help them analyze their data.

New algorithms from campus for stitching models, enabling collaboration

Compressed Schedule → Experience
Tools
Language → Effort

Compressed Schedule → Experience
Tools
Language → Effort

Compressed Schedule → Experience
Tools
Language → Effort

Refocusing cost estimation requires a community effort.

Compare
⬇
Integrate
⬇
Estimate Strength

Actionable Sub-Causal Models

Module Effort = f(factor1, factor2, factor3)
Module Post-Development Quality = g(factor1, factor4, factor5)
High-Reliability Module Cost = h(factor4, factor6, factor7)

Key Activities

Engaged with University of Southern California COCOMO research team

- Kickoff and collaborating since August 2017
- Applying causal learning to various datasets including original COCOMO 81 dataset (corrected)
- Supporting two Ph.D. students in their dissertation work

Initial evaluation of SRDR dataset (next two slides)

- Dataset was basis for just-published DoD Software Factbook

Initial evaluation of PSP datasets from SEI TSP Team

- Includes information on every error committed and caught toward program completion

Reanalyzing publicly available datasets

- Comparing our results with results from prior researcher analyses

Carnegie Mellon University
Software Engineering Institute

Why Does Software Cost So Much? Towards a Causal Model
© 2017 Carnegie Mellon University

[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution.

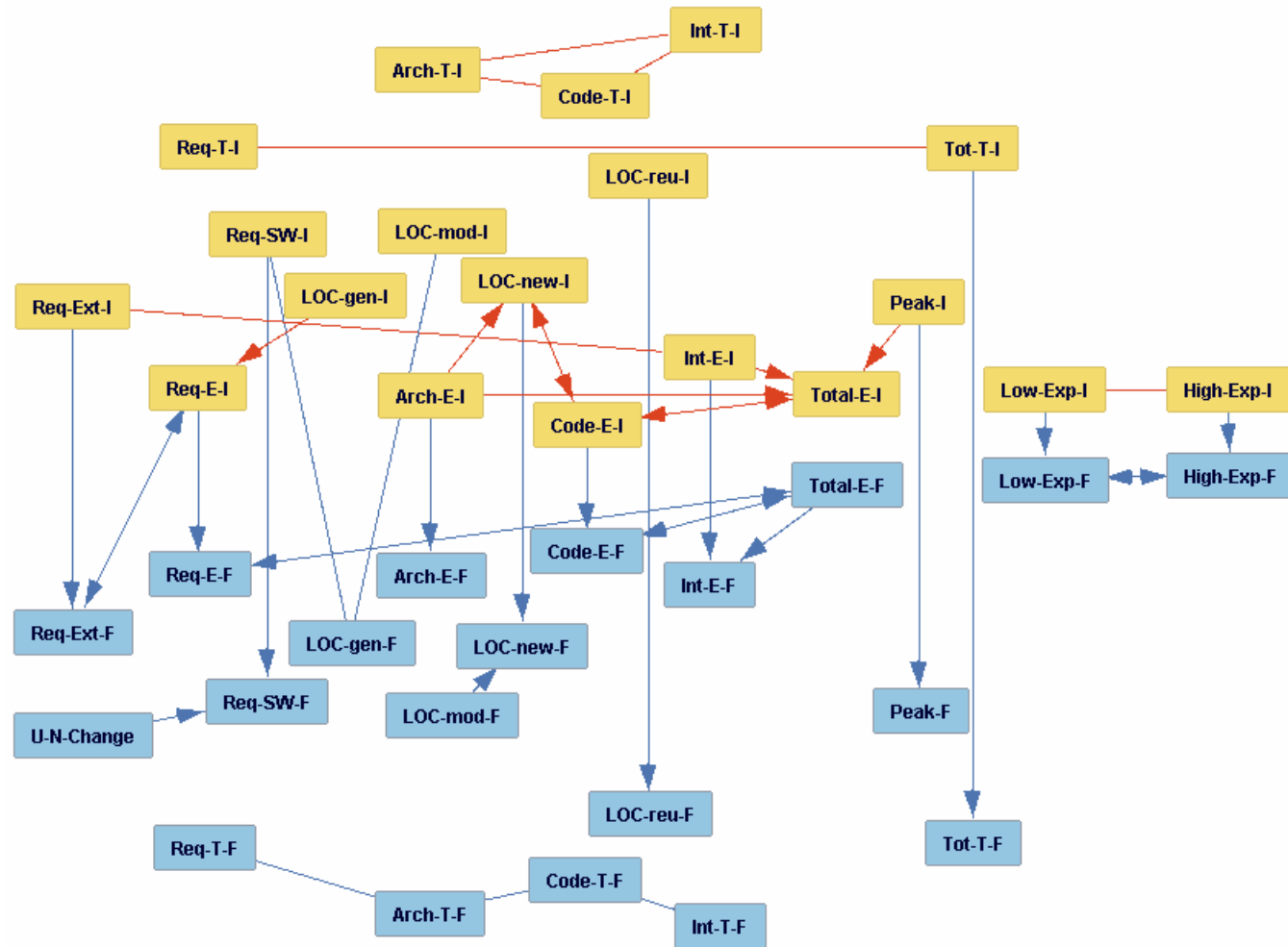**9**

Initial Results: Explaining Final Effort and Duration[1]

181 pairs of matched initial-final SRDR reports reduced to 134 (complete Req…INT data).

Analyzed with PC with Alpha set to .001.

Key:

Req-Ext-I	estimate (initial report)
Req-T-F	actual (final report)

Carnegie Mellon University
Software Engineering Institute

Why Does Software Cost So Much? Towards a Causal Model
© 2017 Carnegie Mellon University

[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution.

10

Initial Results: Explaining Final Effort and Duration²

Note 6 distinct islands:
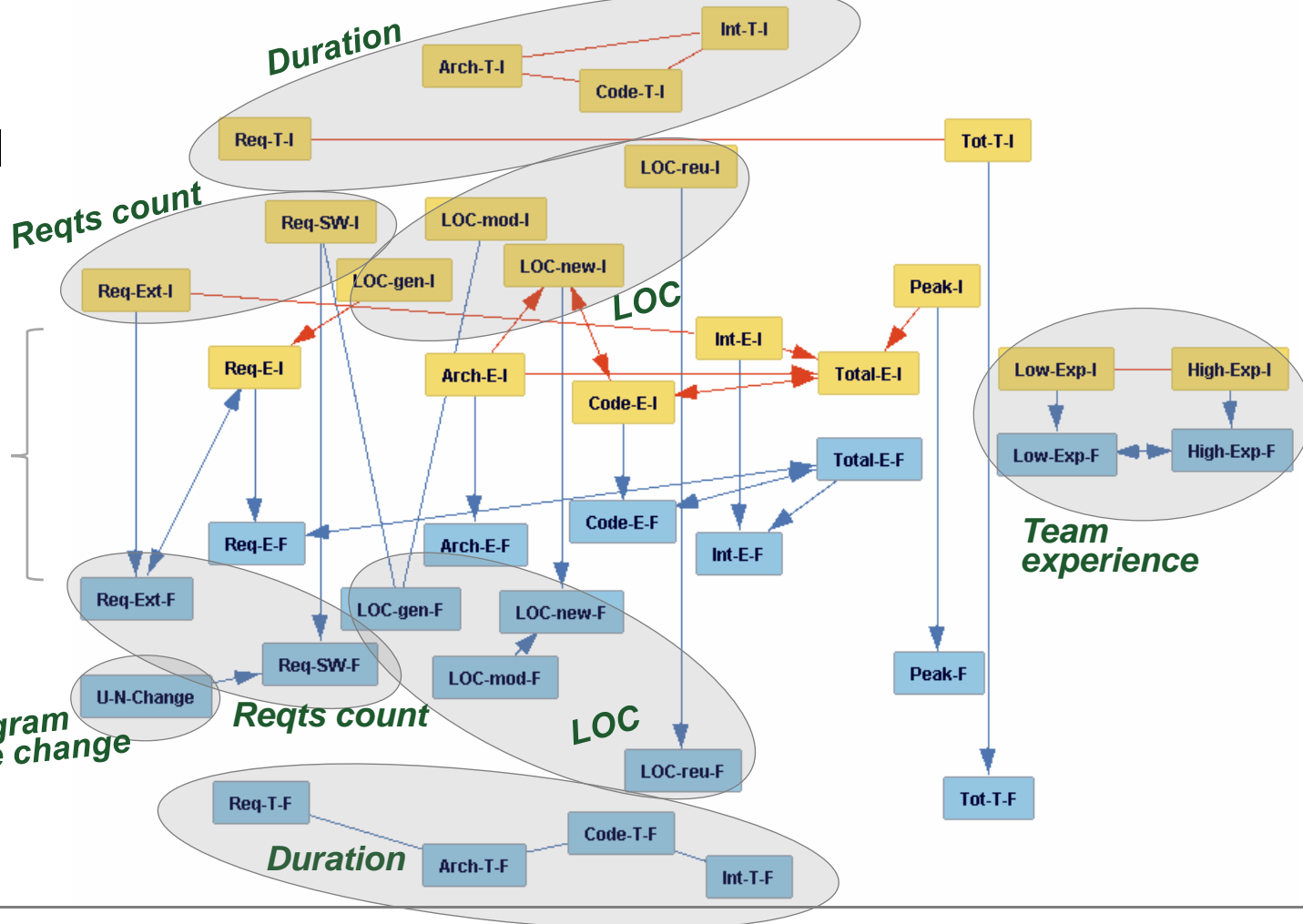
1. Effort and size (2):
   - effort: ***-E-***
   - size: **LOC-***
   - requirements: **Req-***
   - peak time size: **Peak-***
   - program changed (upgrade <-> new): **U-N-Change**
   - code reuse is its own island

2. Duration (3): ***-T-***

3. Team experience: ***-Exp-***

Carnegie Mellon University
Software Engineering Institute

Why Does Software Cost So Much? Towards a Causal Model
© 2017 Carnegie Mellon University

[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution.

11

Early Results Summary

Conventional Wisdom	Not yet checked for generalizability…
TSP/PSP Data Productivity and attention to quality are persistent programmer characteristics.	**TSP/PSP Data** Indeed, past productivity and defect injection rates do have causal effects on effort and defect density of next program.
SRDR Data Effort expended in Req'ts, Architecture, Coding, and Integration drives Total Effort.	**SRDR Data** Architecture effort does not appear to have a causal effect on Total Effort.
COCOMO81 Approximately 25 factors are driving cost.	**COCOMO81** Only software size (ESLOC) has causal effect on cost.
Architecture Data Four architecture pattern violations drive effort and quality.	**Architecture Data** Three of the four architecture pattern violations have a direct causal effect on effort and quality.
USC Cost Data Six key factors were shown to have strong correlation with Total Effort.	**USC Cost Data** Four of six factors have a causal effect on Total Effort. Two new causal factors were also identified.

Carnegie Mellon University
Software Engineering Institute

Why Does Software Cost So Much? Towards a Causal Model
© 2017 Carnegie Mellon University

[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution.

12

Artifacts and Accomplishments

Causal models of the factors that drive software effort

- 40-60 factors related to software effort, quality, and schedule
  - Domain, lifecycle, technology
  - Programmer capability
- Estimates of causal relationships for predicting interim and final costs

Causal Learning and Tetrad tooling and training materials

Presentations

- Software & IT-CAST Workshop, Arlington, VA, August 2017
- Invitation to speak at the 2018 ICEAA conference (based on above presentation)

Use and early results obtained by USC research collaborators

Why Does Software Cost So Much? Towards a Causal Model
© 2017 Carnegie Mellon University

[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution.

Bottom Line and Future Vision

The time is right for applying causal learning to improve cost estimation
- Causal learning has come of age from both a theoretical and tooling standpoint
- Causal models lend themselves to actionable intelligence better than models based on correlation

May lead to ways to improve control of software programs
- Identifying practices, methods, and tools that improve how software is built

Our research provides a rare opportunity
- Bring a diverse community of DoD programs, contractors, and cost estimation researchers together in a joint effort to improve understanding of software costs
- Working with world-class causal learning researchers

SEI will extend this research in FY18-20 to other SW cost-related factors
- Programmer, team, technology, organization, contractor, and acquisition risk factors
- In collaboration with multiple other cost estimation researchers

Contact Information

Presenter / Point(s) of Contact

Bob Stoddard, PI

Mike Konrad

Principal Researchers

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

Telephone: +1 412.268.1121 (Bob)
+1 412.268.5813 (Mike)

Other SEI Team Members

Bill Nichols

Dave Zubrow

CMU Contributors

David Danks

Madelyn Glymour

Kun Zhang

USC Contributors

Jim Alstad

Barry Boehm

Anandi Hira