Extended Case Study of Causal Learning within Architecture Research (preliminary results)

Robert Stoddard, SEI
Mike Konrad, SEI
Rick Kazman, SEI
David Danks, CMU

Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213

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

Goal of the Authors

Robert Stoddard
SEI

Dr. Mike Konrad
SEI

Dr. David Danks
CMU

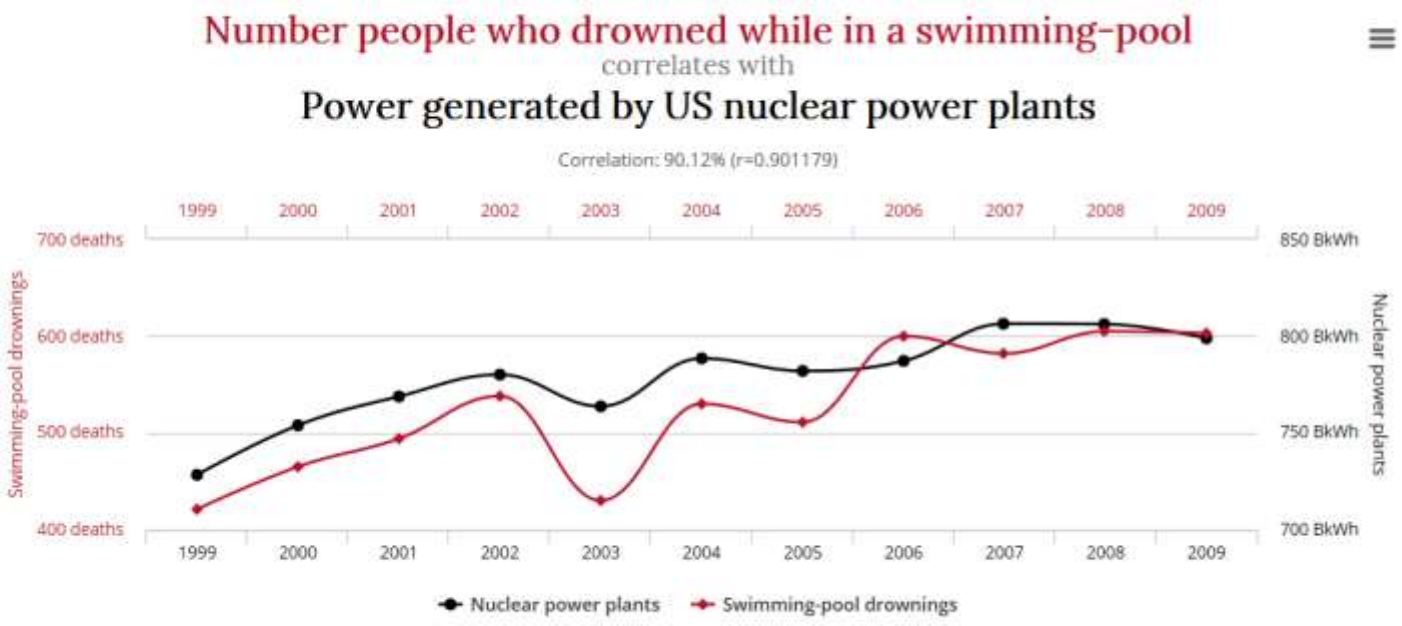
Goal is to motivate adoption of Causal Learning within software measurement!

Dr. Rick Kazman
SEI

© 2018 Carnegie Mellon University

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

Why Do We Care about Causation?

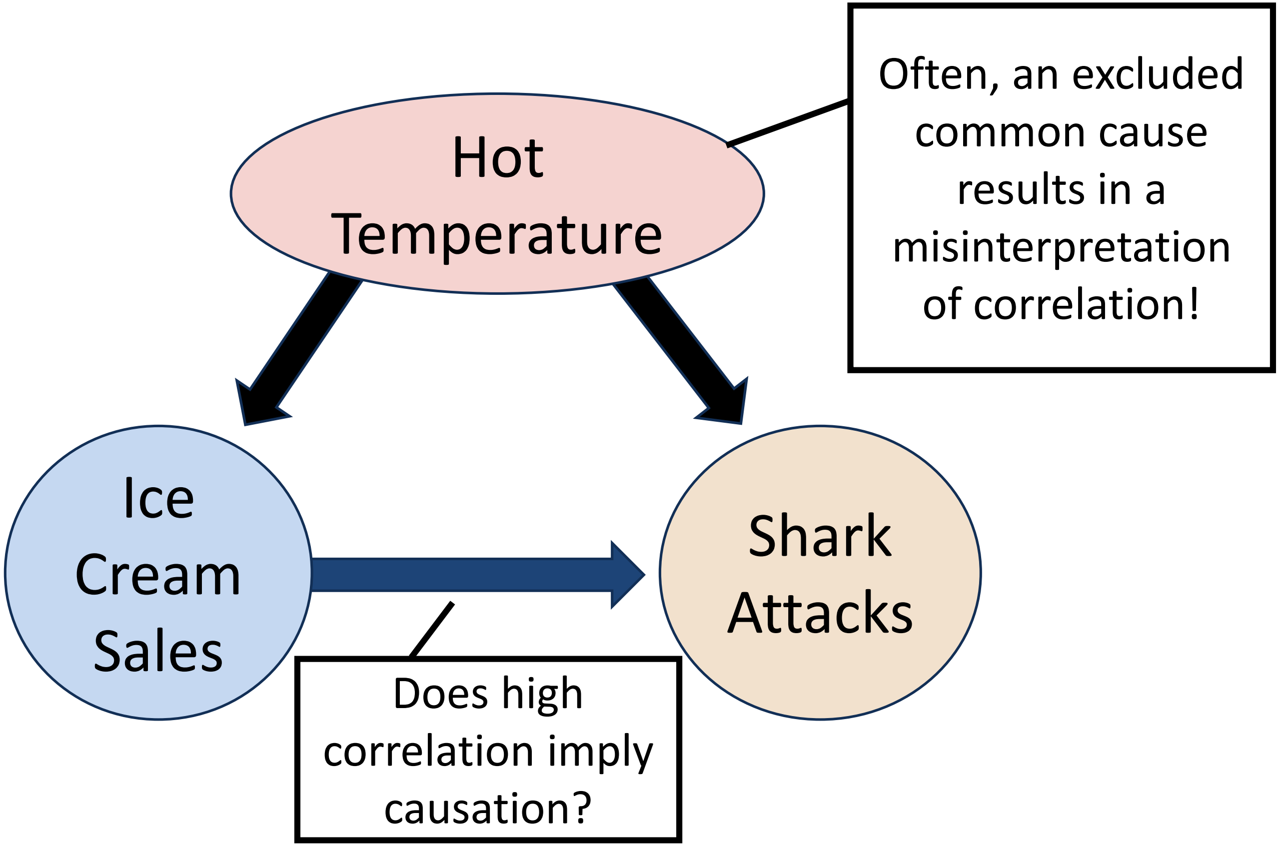


http://www.tylervigen.com/spurious-correlations

© 2018 Carnegie Mellon University

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

More about Misinterpreting Correlation!



© 2018 Carnegie Mellon University
[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution.]

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.

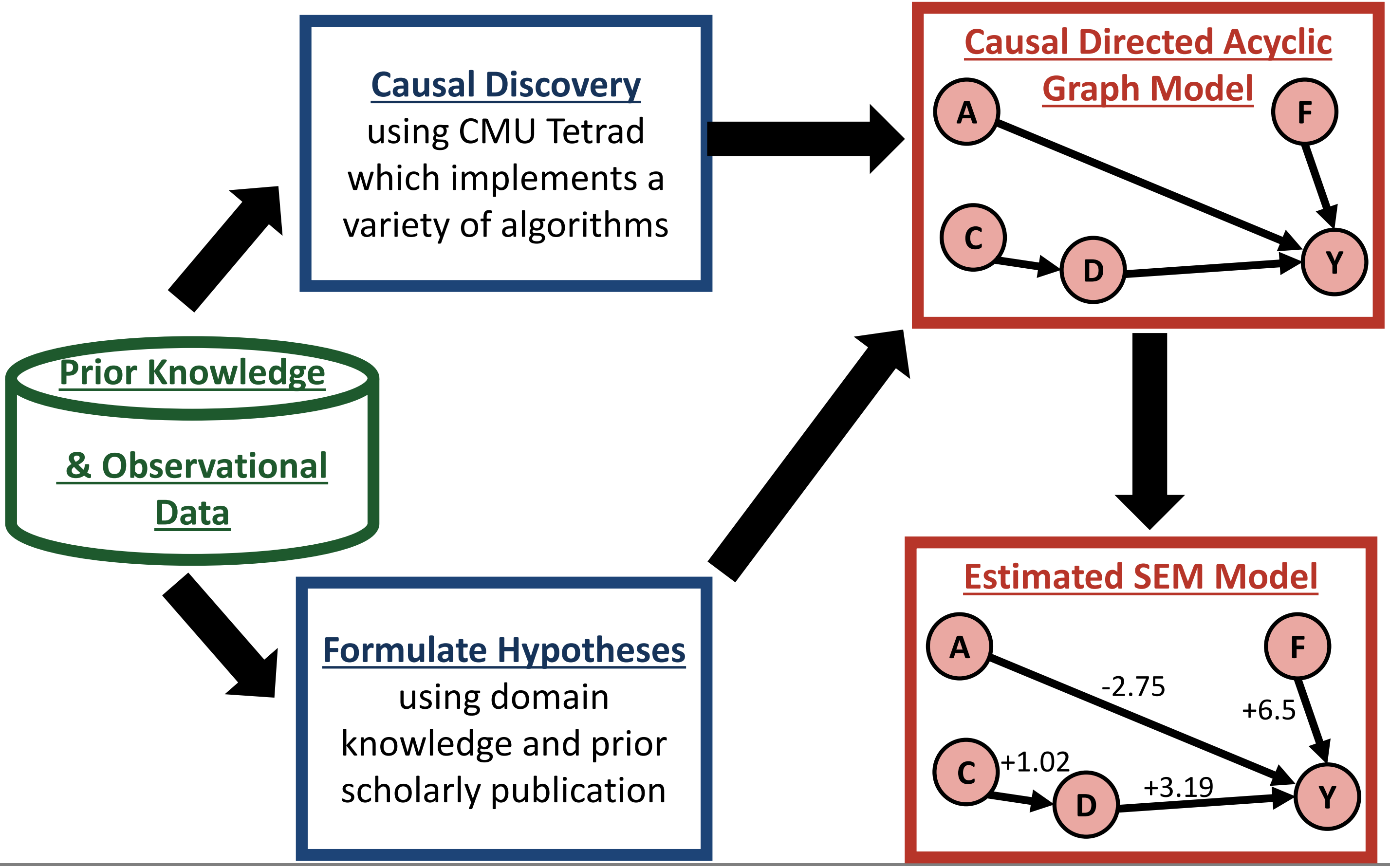
Lastly, we then conduct regression with the correct set of factors!

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

© 2018 Carnegie Mellon University

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

The Causal Learning Landscape



© 2018 Carnegie Mellon University

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

Preliminary Architecture Research Causal Findings

Nine open source systems analyzed using static code analysis (> 9000 files)

Four architecture pattern violations studied for impact on quality

Each file had the following attributes measured:

- Age in Months
- Number of Developers touching each file
- Size in Lines of Code
- Number of times the file participated in a pattern violation of:
  - the cyclic dependency
  - Improper inheritance
  - Unstable interface
  - Lack of modularity
- Quality outcome of Number of Bugs associated with each file
- Bug churn associated with each file

R. Mo, Y. Cai, R. Kazman and L. Xiao, "Hotspot Patterns: The Formal Definition and Automatic Detection of Architecture Smells," *2015 12th Working IEEE/IFIP Conference on Software Architecture*, Montreal, QC, 2015, pp. 51-60. doi: 10.1109/WICSA.2015.12

© 2018 Carnegie Mellon University

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

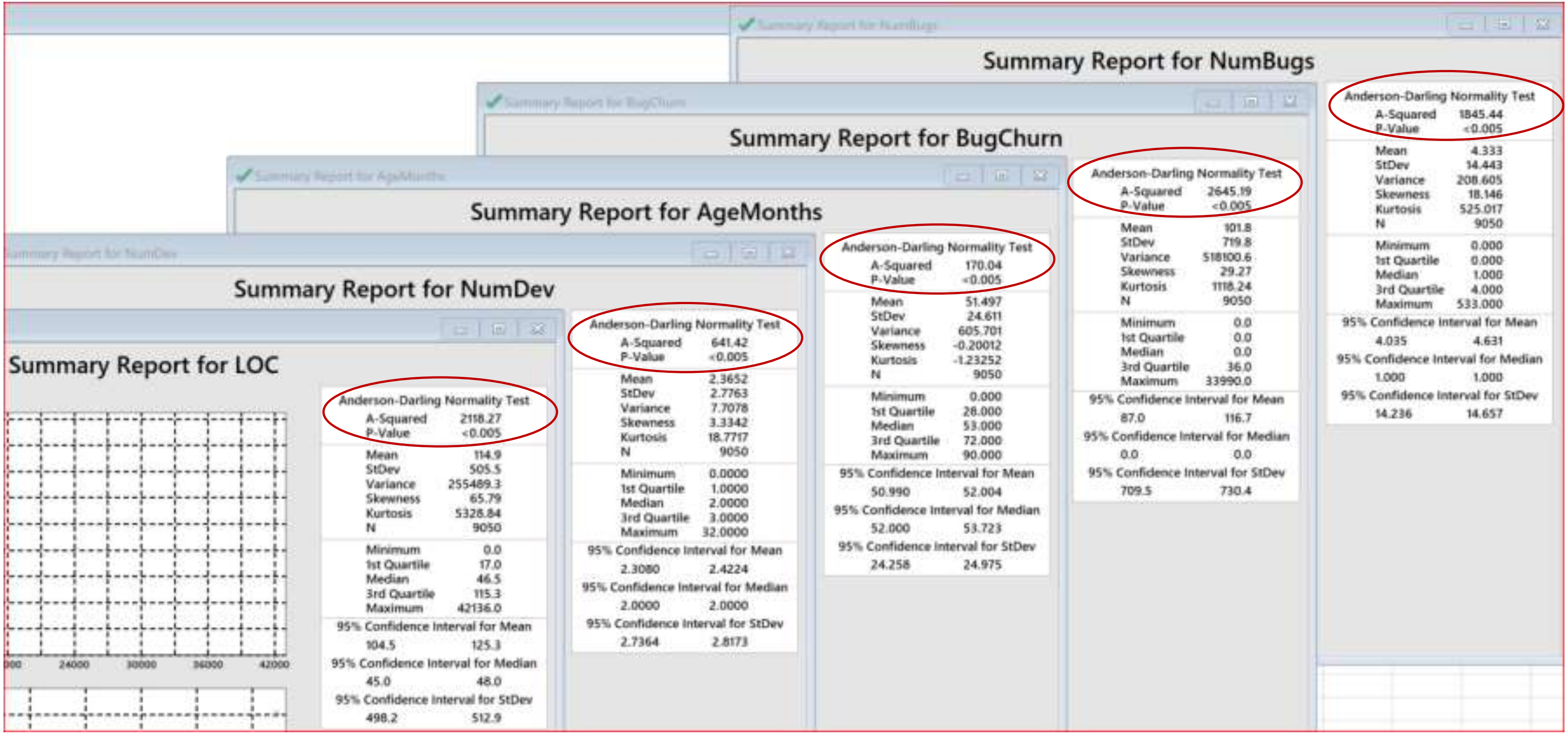
Correlation Matrix of All Factors

	AgeMonths	NumDev	NumCommits	LOC	NumBugs	NumChanges	BugChurn	ChangeChurn	NumCyclicDepend	NumModularityVio	NumUnstableInter
NumDev	0.1790										
	0.0000										
NumCommits	0.0930	0.6890									
	0.0000	0.0000									
LOC	0.0460	0.2640	0.2720								
	0.0000	0.0000	0.0000								
NumBugs	0.1160	0.6540	0.9330	0.2570							
	0.0000	0.0000	0.0000	0.0000							
NumChanges	0.0960	0.6880	0.9990	0.2720	0.9340						
	0.0000	0.0000	0.0000	0.0000	0.0000						
BugChurn	0.0380	0.3920	0.5810	0.7270	0.6390	0.5820					
	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000					
ChangeChurn	0.0180	0.2980	0.4180	0.9400	0.4120	0.4180	0.8300				
	0.0880	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000				
NumCyclicDepend	0.0340	0.1520	0.2920	0.1000	0.2430	0.2900	0.1240	0.1080			
	0.0010	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000			
NumModularityVio	0.0490	0.3270	0.2100	0.1070	0.1590	0.2100	0.0980	0.1000	0.0130		
	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.2140		
NumUnstableInter	0.0390	0.5400	0.4820	0.1580	0.3940	0.4810	0.2220	0.2000	0.1420	0.2670	
	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	
NumImproperInher	0.1280	0.2060	0.2110	0.1040	0.1850	0.2120	0.1150	0.0740	0.1620	0.0020	0.1330
	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.8540	0.0000

NumBugs, NumChanges, and NumCommits are highly correlated; Will keep NumBugs only in the modeling;
Likewise, ChangeChurn and LOC highly correlated, so kept only LOC in the modeling

© 2018 Carnegie Mellon University
[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution.]

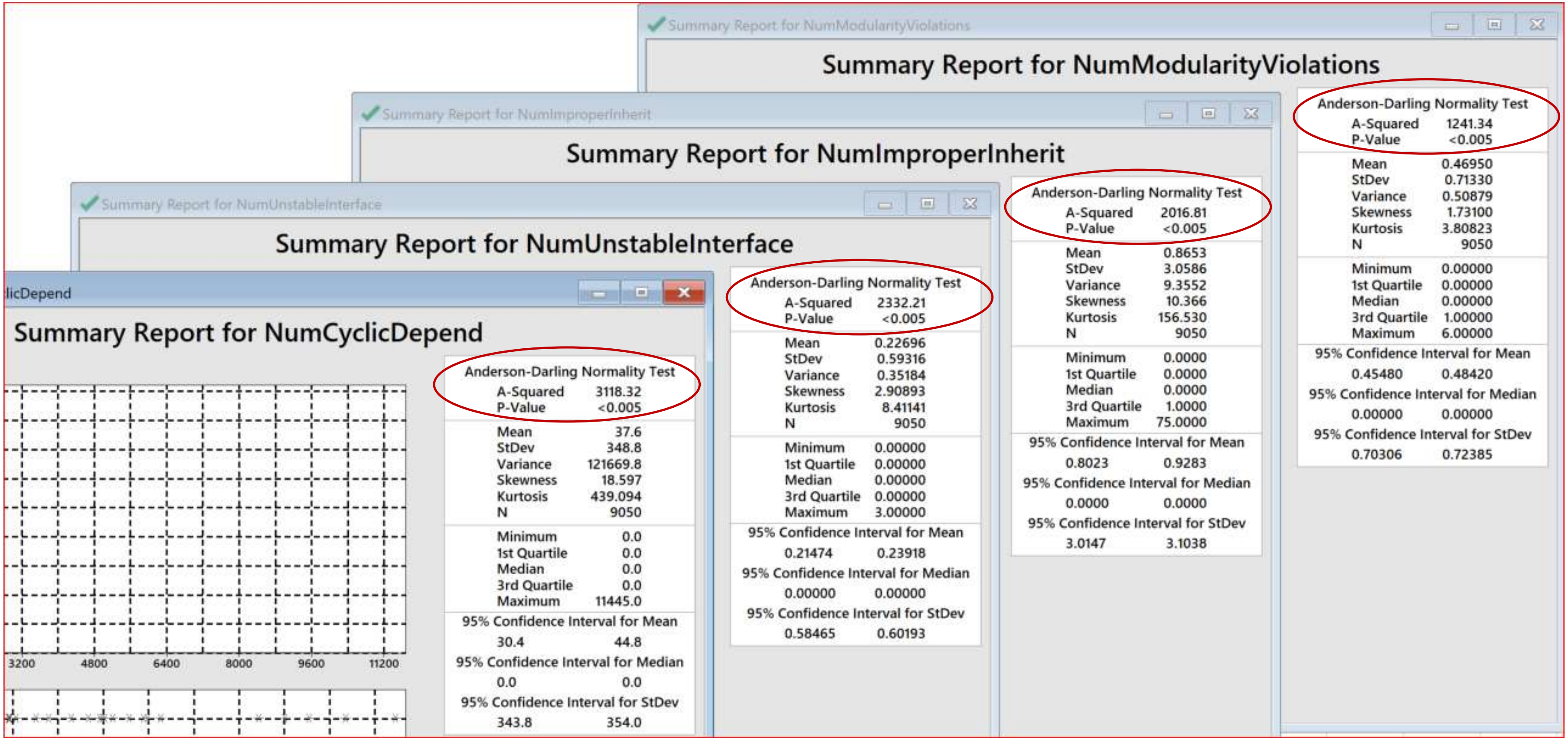
All Remaining Factors are Non-Normal - 01



© 2018 Carnegie Mellon University

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

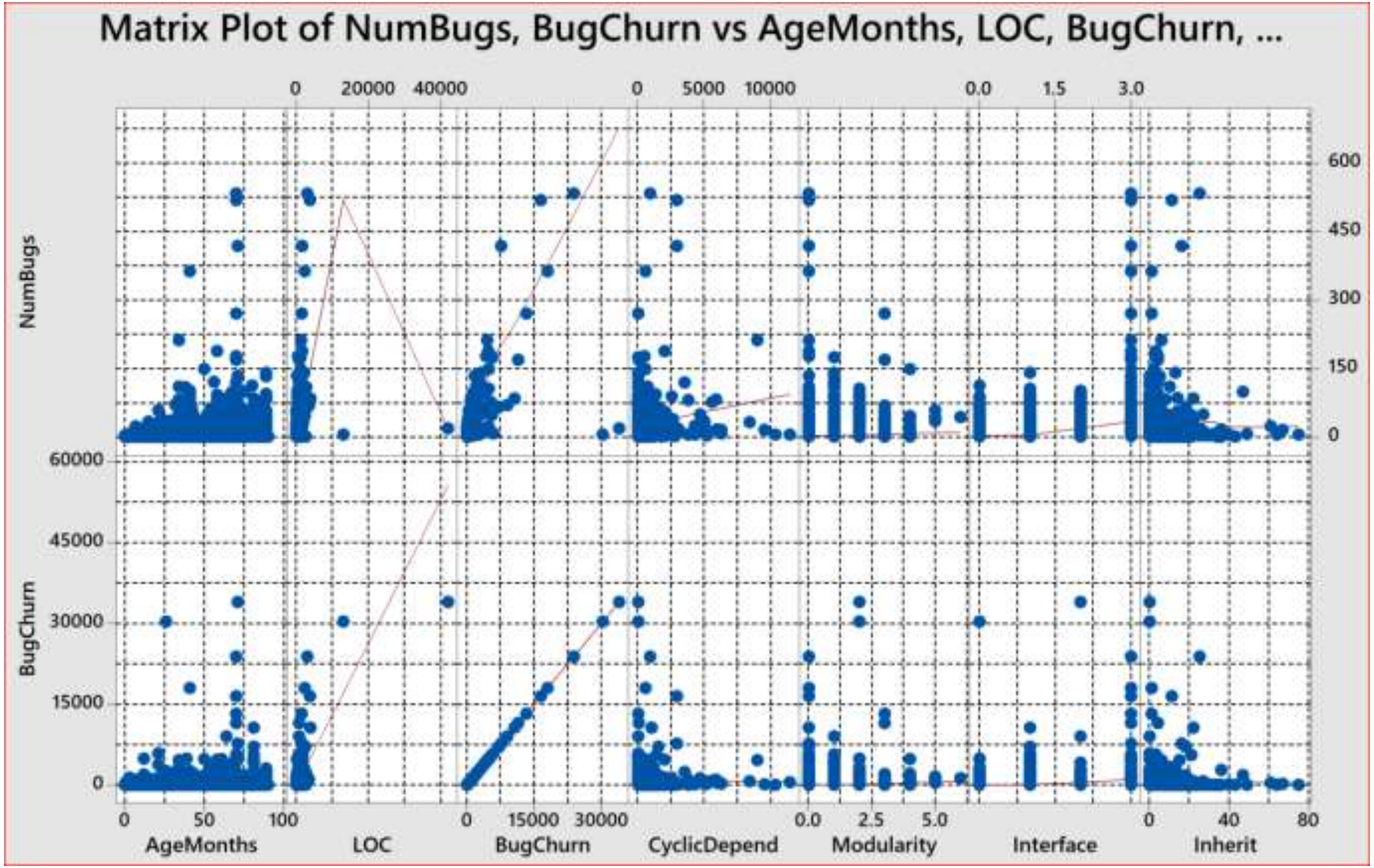
All Remaining Factors are Non-Normal - 02



© 2018 Carnegie Mellon University

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

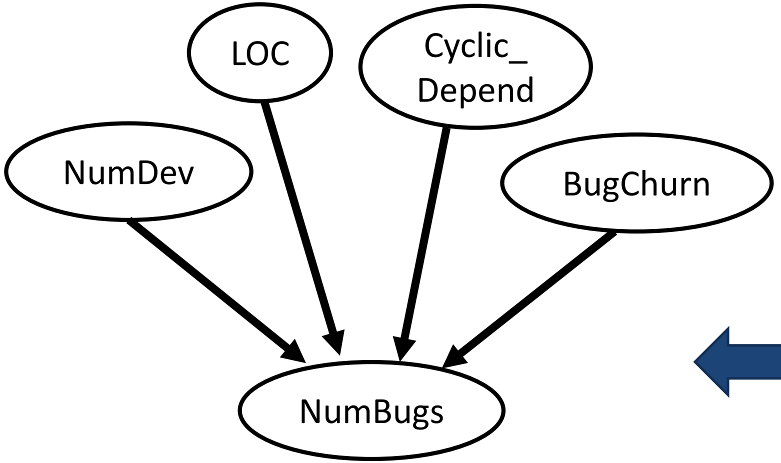
Eyeballing Bivariate Relationships



© 2018 Carnegie Mellon University

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

Best Subsets Regression



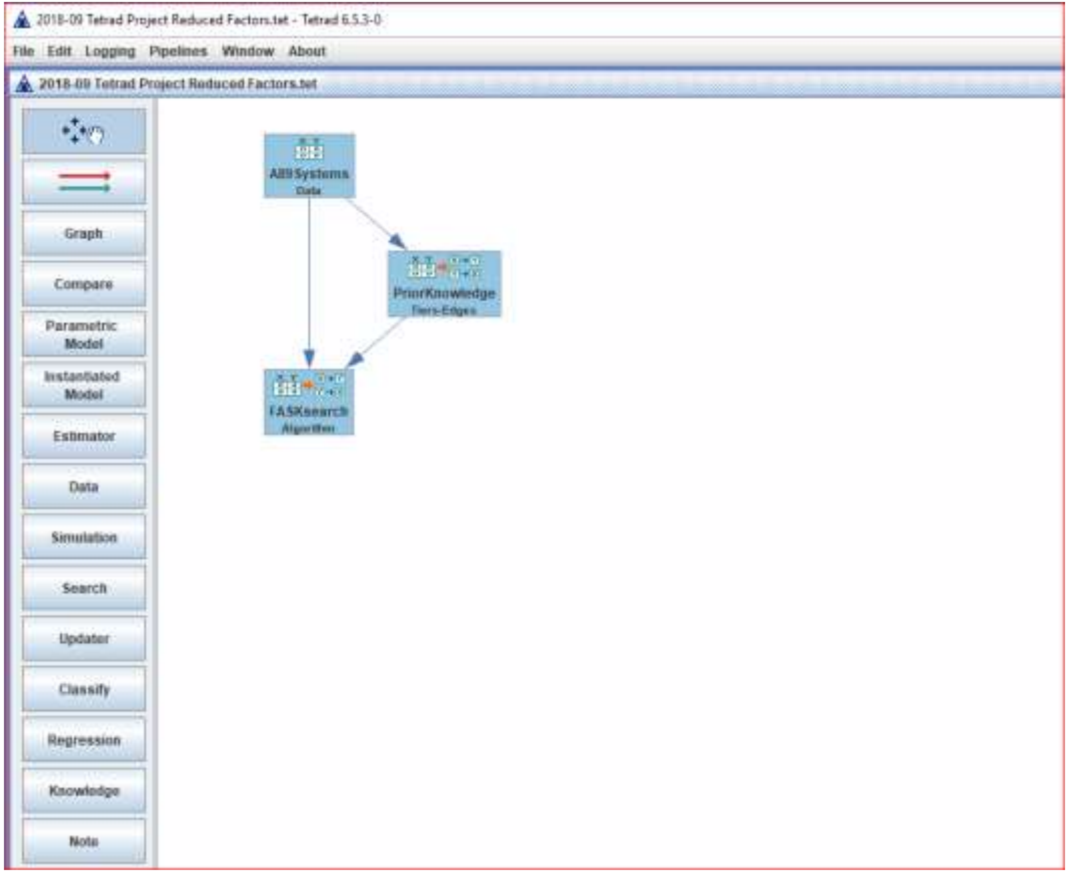
Response is NumBugs

Vars	R-Sq	R-Sq (adj)	R-Sq (pred)	Mallows Cp	S	AgeMonths	NumDev	LOC	BugChurn	CyclicDepend	Modularity	Interface	Inherit
1	42.8	42.8	42.1	8301.1	10.921		X						
1	40.9	40.9	27.7	8891.2	11.106				X				
2	60.1	60.1	50.8	3051.3	9.1201		X		X				
2	50.0	50.0	27.0	6132.0	10.216			X	X				
3	68.4	68.4	63.0	544.7	8.1199		X	X	X				
3	61.5	61.5	52.1	2647.3	8.9662		X		X	X			
4	69.9	69.9	64.7	101.9	7.9297		X	X	X	X			
4	68.6	68.6	63.2	480.3	8.0921		X	X	X			X	
5	70.0	70.0	64.8	59.3	7.9108		X	X	X	X		X	
5	70.0	69.9	64.8	77.0	7.9184		X	X	X	X			X
6	70.1	70.1	64.9	35.5	7.9000		X	X	X	X		X	X
6	70.1	70.1	64.9	42.5	7.9030		X	X	X	X	X	X	
7	70.2	70.1	65.0	21.6	7.8935		X	X	X	X	X	X	X
7	70.2	70.1	65.0	23.0	7.8941	X	X	X	X	X		X	X
8	70.2	70.2	65.0	9.0	7.8875	X	X	X	X	X	X	X	X

© 2018 Carnegie Mellon University

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

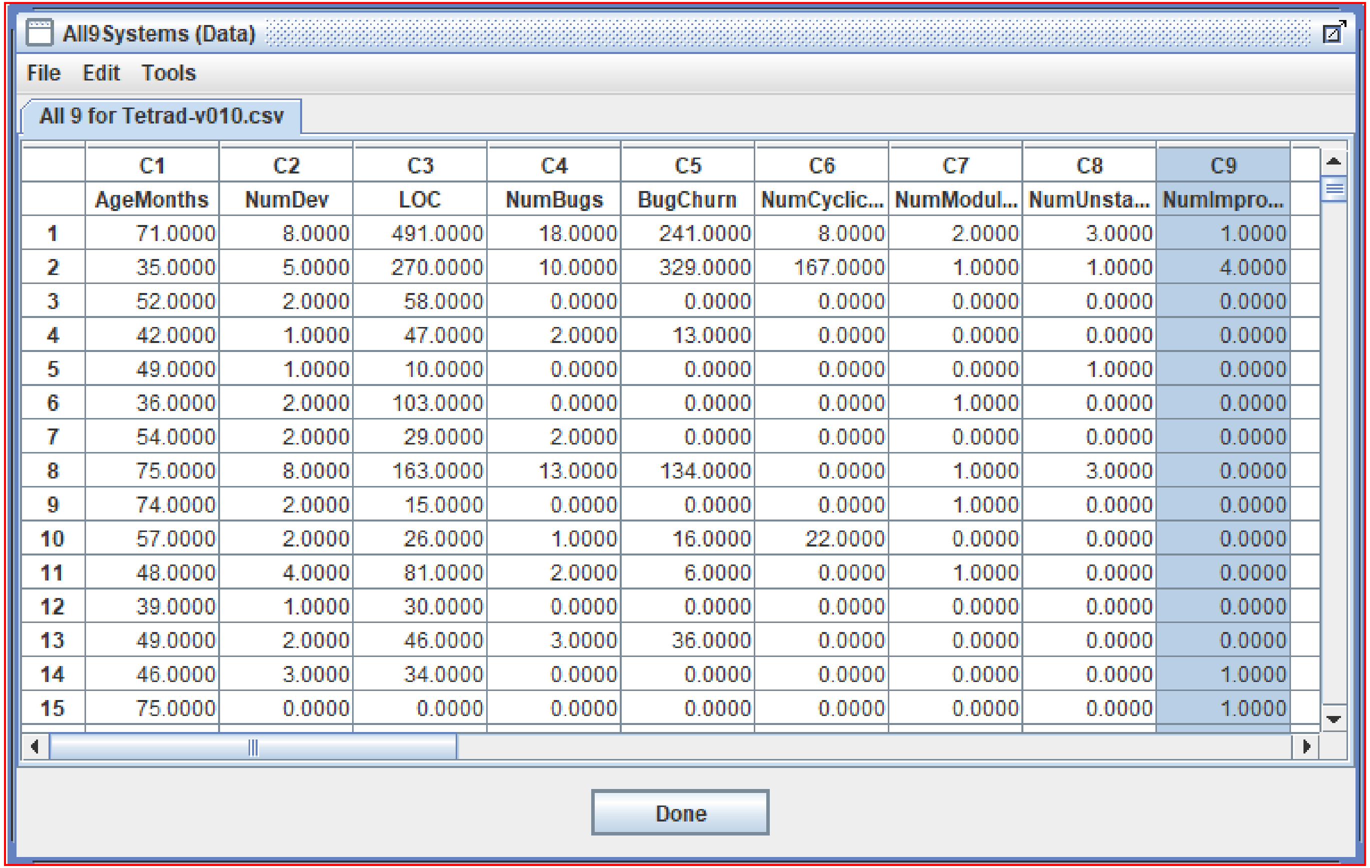
Conduct Causal Search using Tetrad



© 2018 Carnegie Mellon University

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

A View of the Data File Loaded into Tetrad

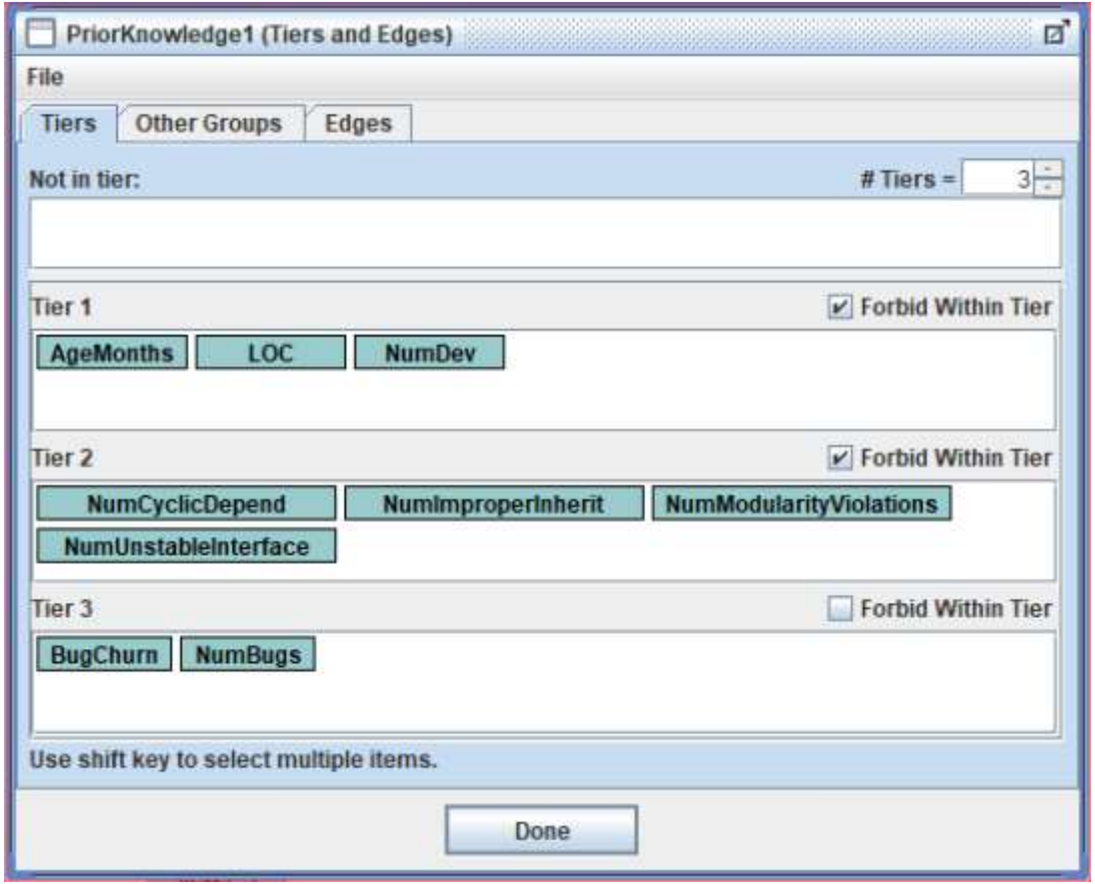


	C1	C2	C3	C4	C5	C6	C7	C8	C9
	AgeMonths	NumDev	LOC	NumBugs	BugChurn	NumCyclic...	NumModul...	NumUnsta...	NumImpro...
1	71.0000	8.0000	491.0000	18.0000	241.0000	8.0000	2.0000	3.0000	1.0000
2	35.0000	5.0000	270.0000	10.0000	329.0000	167.0000	1.0000	1.0000	4.0000
3	52.0000	2.0000	58.0000	0.0000	0.0000	0.0000	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

© 2018 Carnegie Mellon University

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

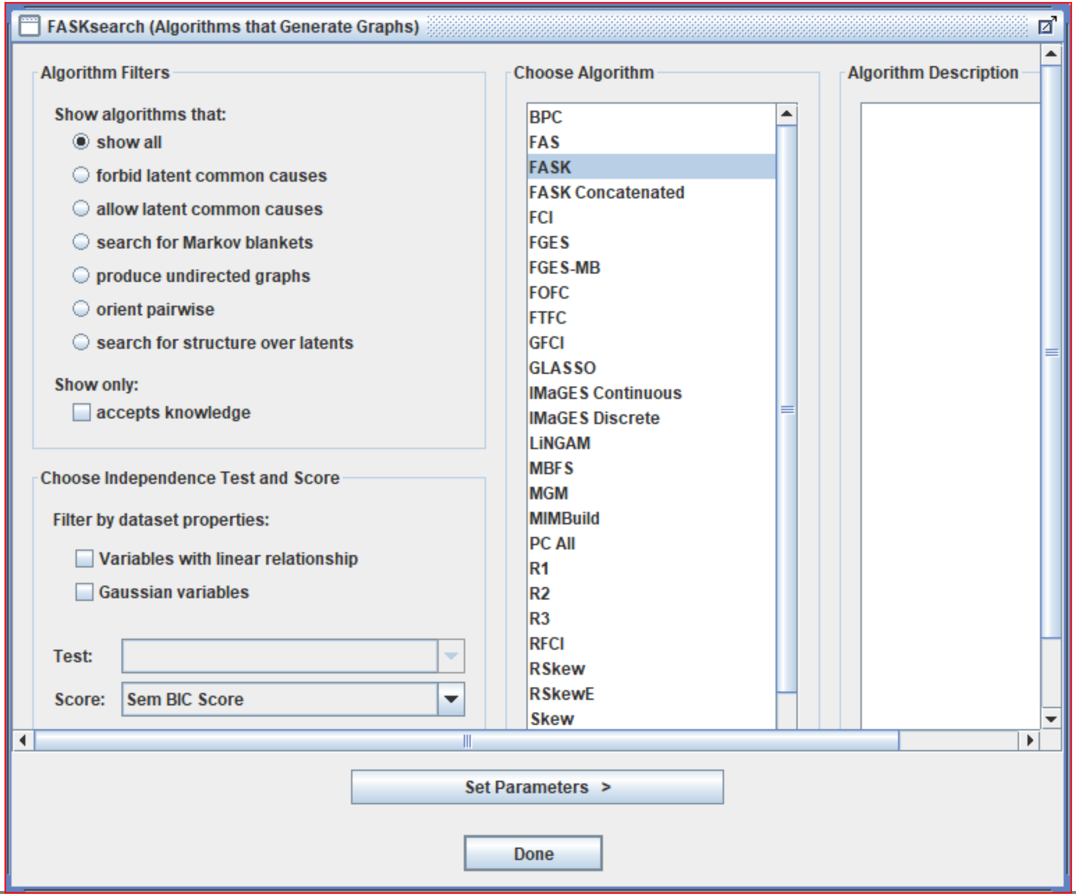
Prior Knowledge Entered into Tetrad



© 2018 Carnegie Mellon University

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

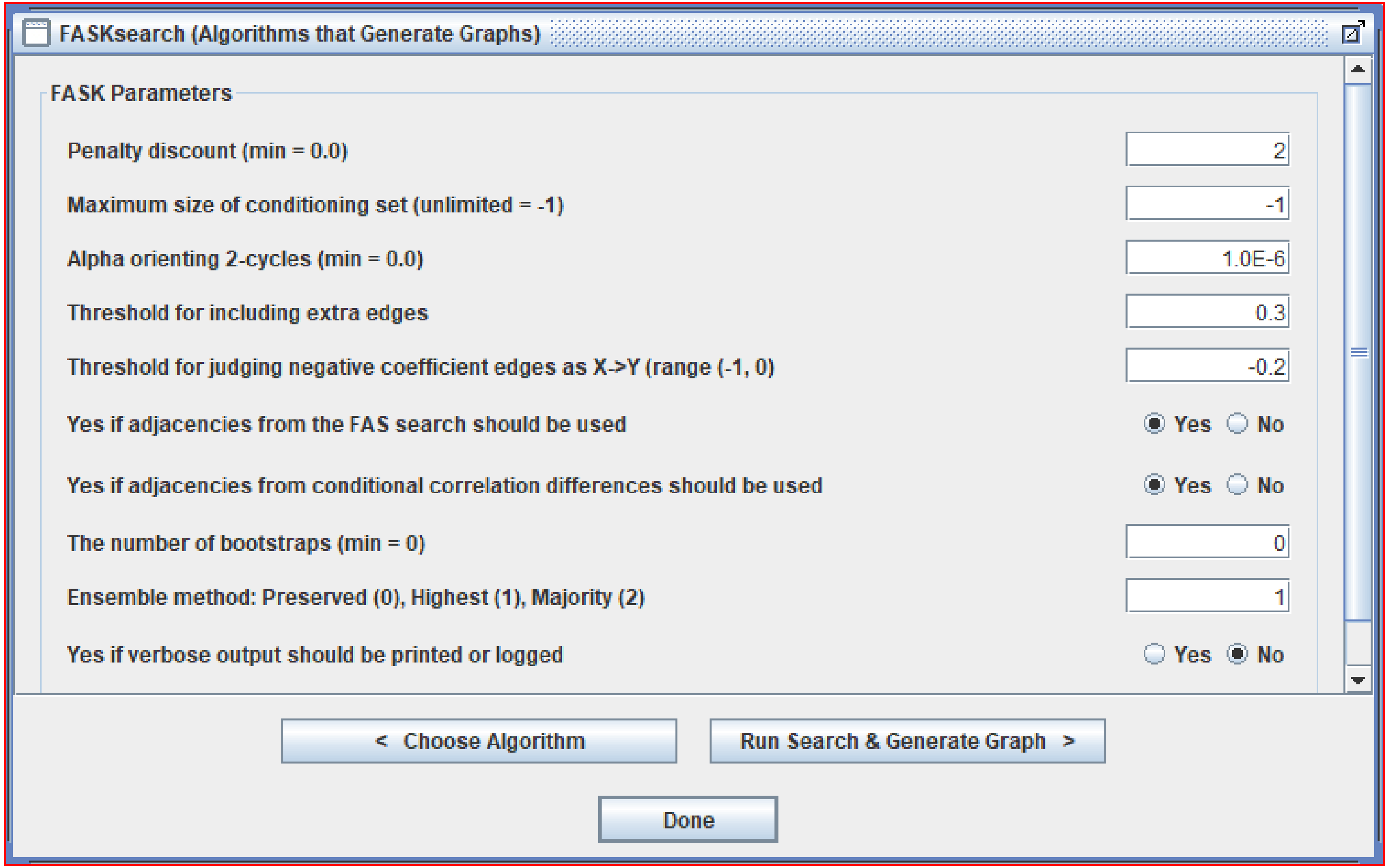
Using FASK Search with Associated Parameters



© 2018 Carnegie Mellon University

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

Additional FASK Search Parameter Settings



© 2018 Carnegie Mellon University

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

Causal Search Algorithms

Constraint-based: Calculate independences in the data and do “backwards inference”; used to minimize the degree of false negative edges

Score-based (Bayesian): Calculate the likelihood of different DAGs given the data; used to minimize the degree of false positive edges

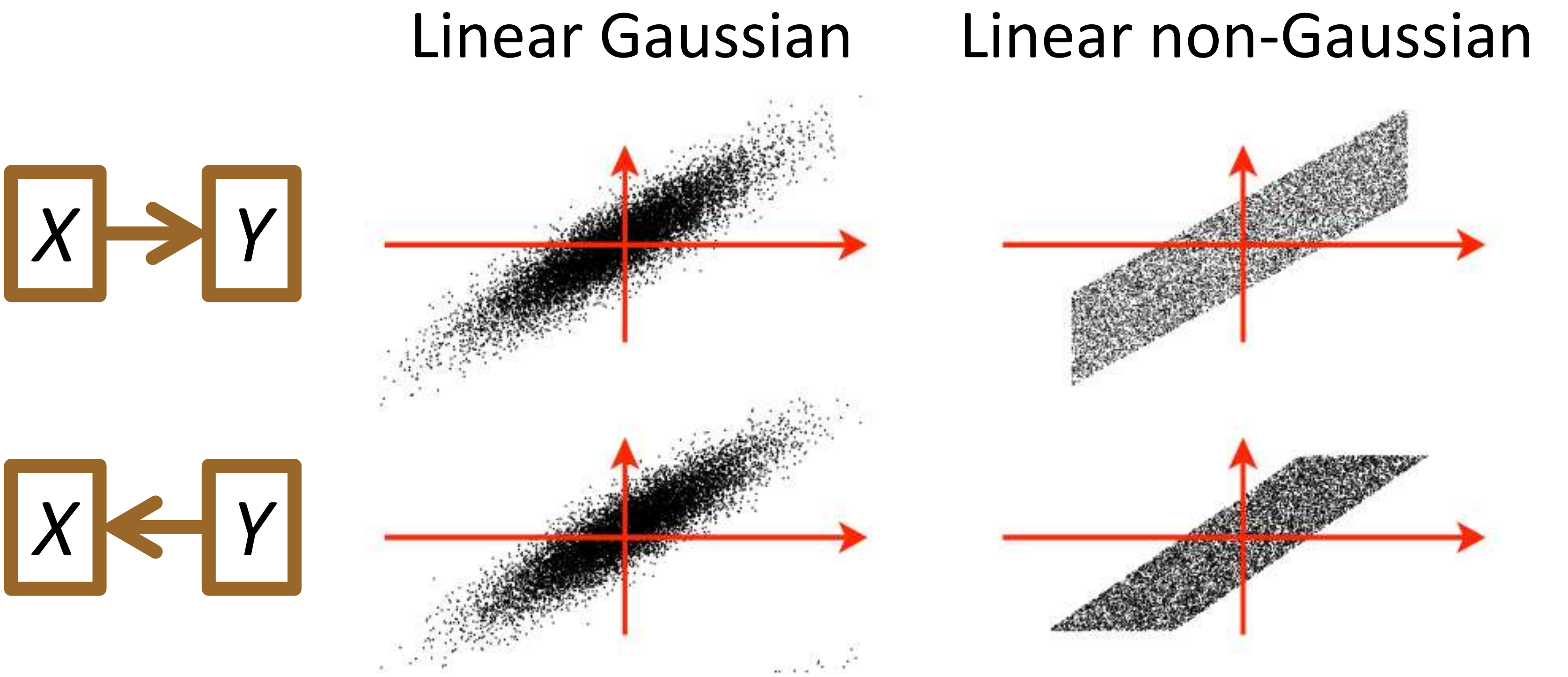
Hybrid: Use constraint-based to get “close,” then Bayesian search around neighborhood

A	B	No evidence of a causal link
A	$\longrightarrow$ B	Evidence of a causal link from A to B
A $\longleftarrow$	B	Evidence of a causal link from B to A
A	$\longleftrightarrow$ B	Evidence of an unmeasured confounder

© 2018 Carnegie Mellon University

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

Some Algorithms Exploit Non-Gaussianality



© 2018 Carnegie Mellon University

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

Causal Search Capable with Small Data

<u>**Challenge**</u>: Which genes regulate flowering time in Arabidopsis thaliana?

Using only 47 observations, causal search identified 9 out of 21,326 genes as causal on gene activation

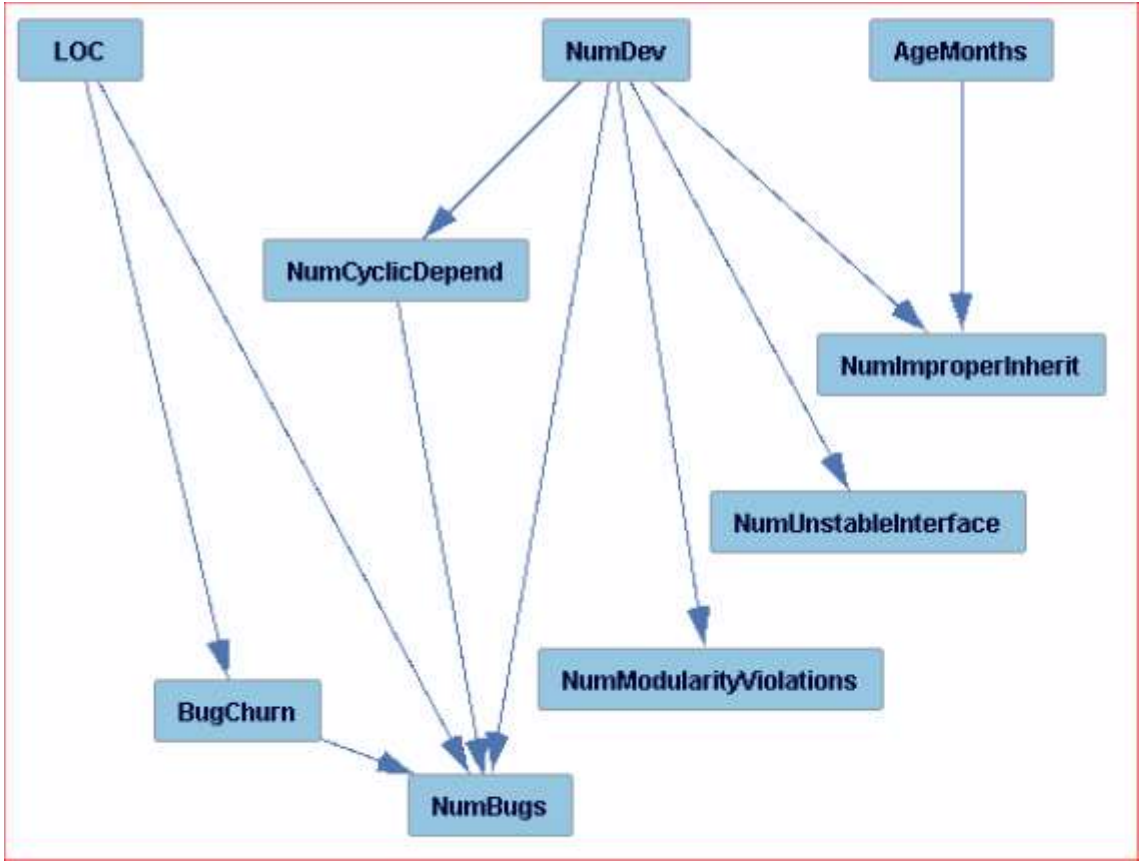
Subsequent greenhouse study, that used knockout variants, confirmed that 4 of the 9 were actual regulators

Taken from Dave Danks, 2016 Summer Causal Workshop

© 2018 Carnegie Mellon University

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

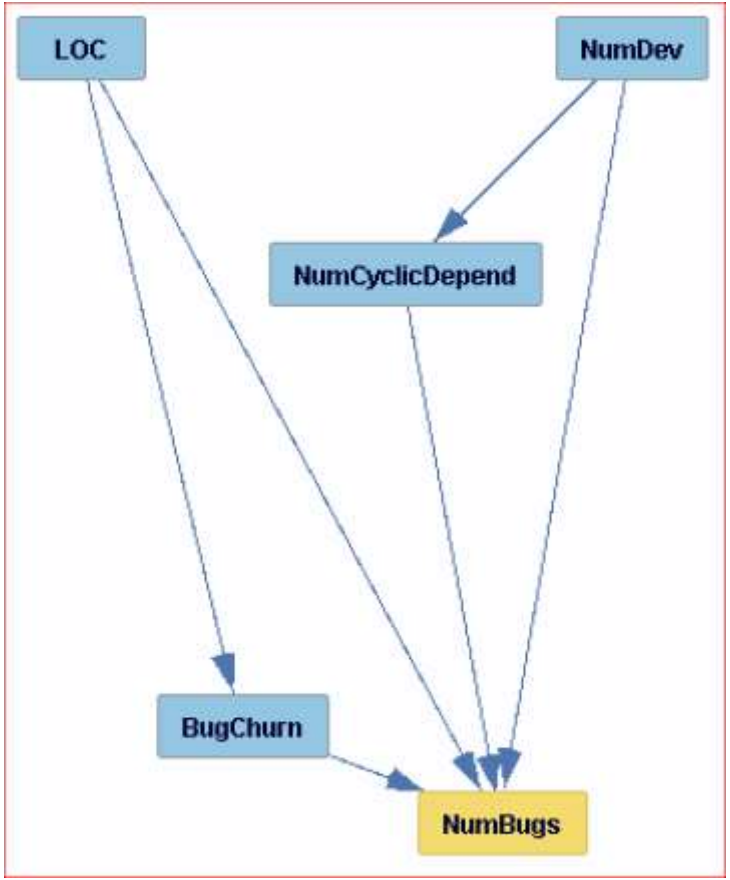
Causal Structure Graph Result



© 2018 Carnegie Mellon University

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

Markov Blanket of the NumBugs Factor



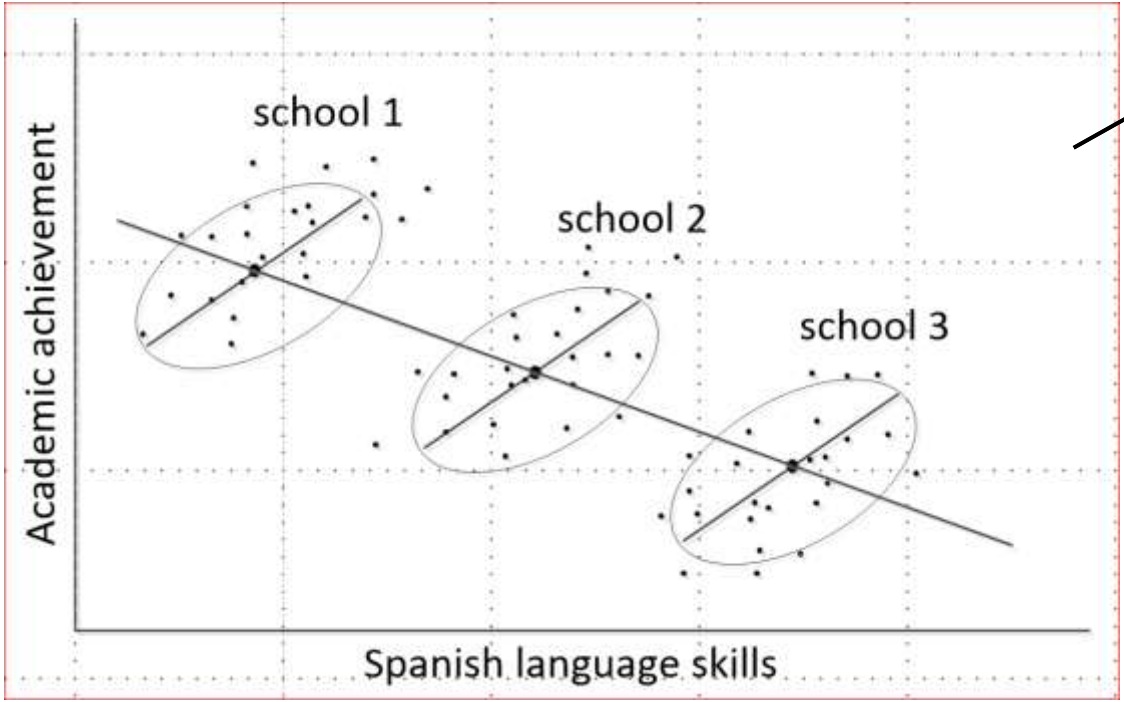
© 2018 Carnegie Mellon University

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

Motivation to Look at Multi-Level SEM Models (MSEM)

Within schools, students with better Spanish skills had higher academic achievement.

Yet, schools with highest proportion of Spanish speakers performed poorest.



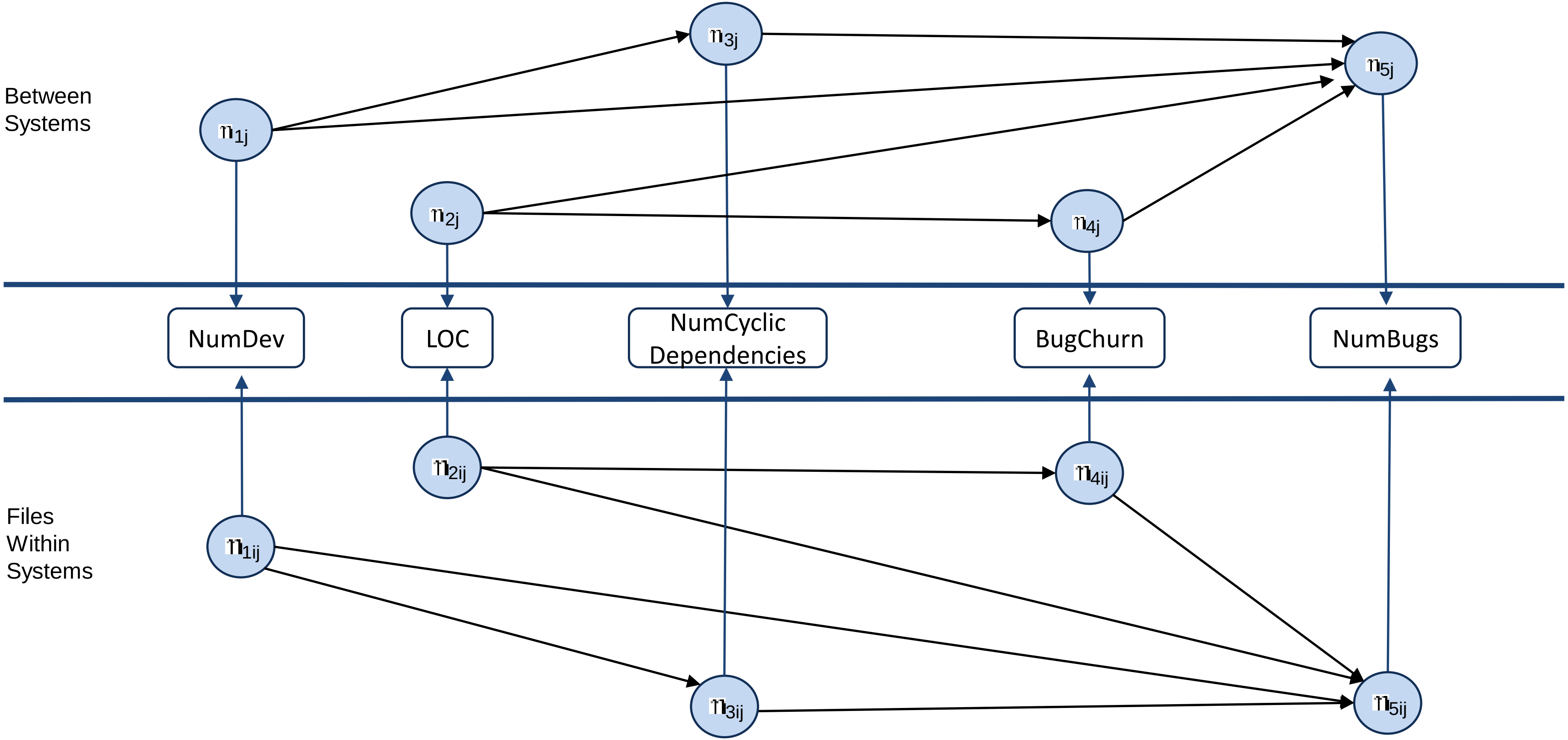
Also called Simpson's Paradox and the Ecological Fallacy

Kris Preacher, 2018

© 2018 Carnegie Mellon University

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

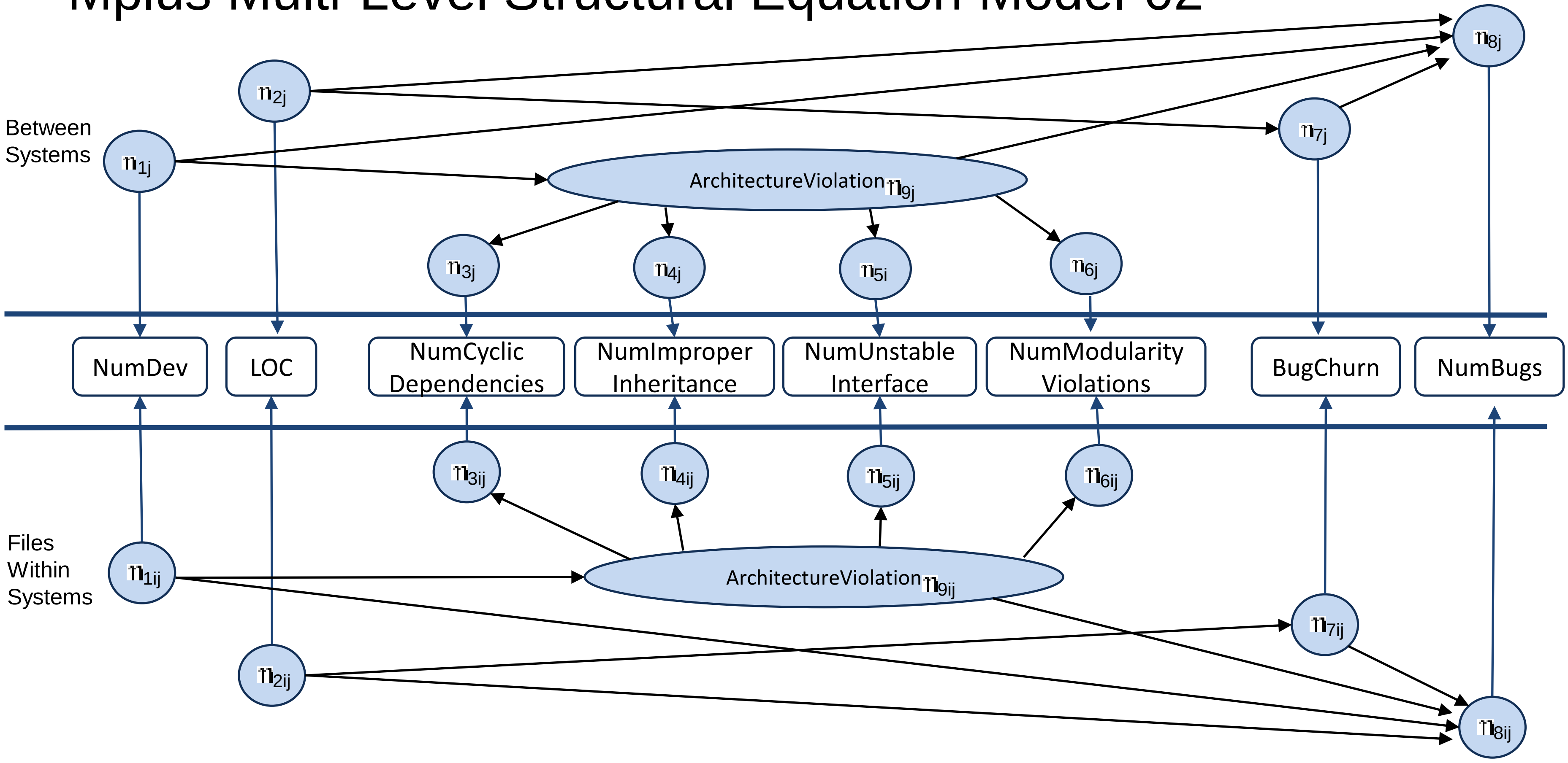
Mplus Multi-Level Structural Equation Model-01



© 2018 Carnegie Mellon University

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

Mplus Multi-Level Structural Equation Model-02



© 2018 Carnegie Mellon University

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

Mplus Code

```
TITLE: Basic Model of NumBugs Markov Blanket;

DATA: FILE IS All9forMplus.csv;

VARIABLE: NAMES ARE AgeMos NumDev LOC Cycles Inherit Interfac Modular BugChurn NumBugs
System;

USEVARIABLES ARE NumDev LOC Cycles BugChurn NumBugs System;

CLUSTER IS System;

ANALYSIS: TYPE IS TWOLEVEL;

MODEL:

%BETWEEN%
NumBugs ON BugChurn LOC NumDev Cycles;
NumBugs; BugChurn; LOC; NumDev; Cycles;
[NumBugs]; [BugChurn]; [LOC]; [NumDev]; [Cycles];

%WITHIN%
NumBugs ON BugChurn LOC NumDev Cycles;

OUTPUT: SAMPSTAT STDYX;
```


© 2018 Carnegie Mellon University

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

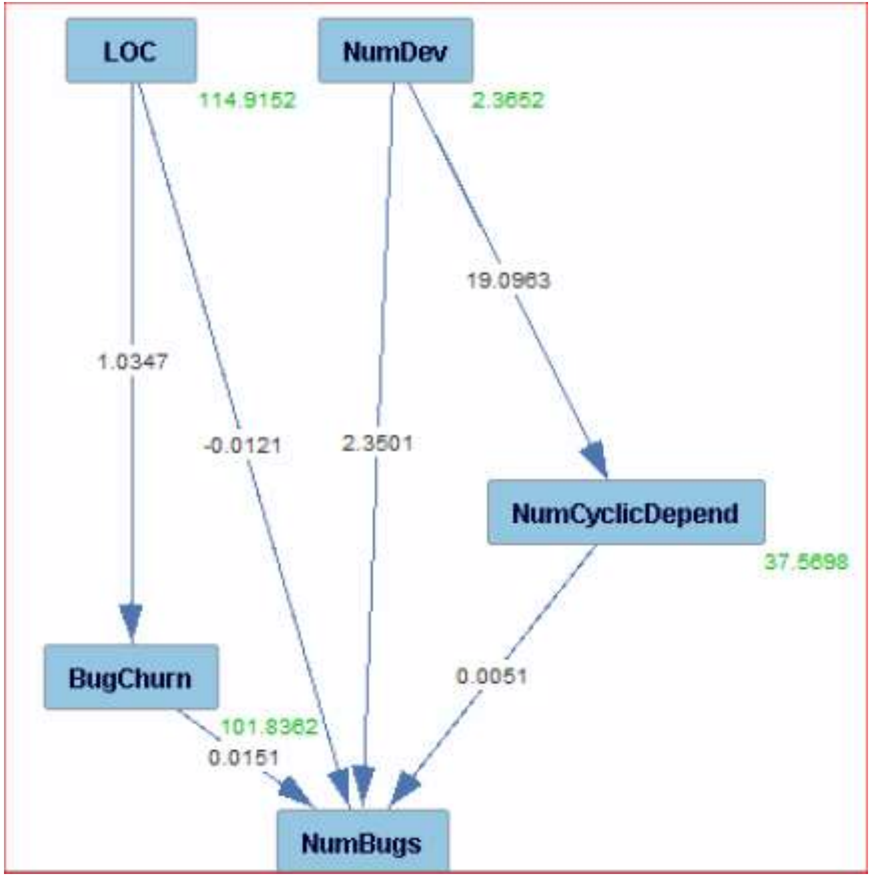
Mplus MSEM Results

```
SUMMARY OF DATA

     Number of clusters                              9

     Average cluster size     1005.556

     Estimated Intraclass Correlations for the Y Variables

               Intraclass              Intraclass              Intraclass
     Variable  Correlation   Variable  Correlation   Variable  Correlation

     NUMBUGS      0.052      NUMDEV       0.084      LOC           0.008
     CYCLES       0.039      BUGCHURN     0.026
```


© 2018 Carnegie Mellon University

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

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

© 2018 Carnegie Mellon University

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

Conclusions

1. We attempted MSEM modeling to be sensitive to the "between" and "within" variation components of all the factors

2. We also wanted to guard against Simpson's paradox

3. The Mplus MSEM analysis, via the Intraclass Correlation measures, showed that in this data situation, we do not need to perform MSEM with two levels

4. We then conducted a single level, univariate SEM within Tetrad

5. We achieved regression coefficients that take into account the mediation effects occurring on the outcome, NumBugs

6. Traditional regression would have been ignorant of the above

© 2018 Carnegie Mellon University

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

Next Steps

Perform more causal searches

- Additional algorithms
- Sensitivity analysis of algorithm parameters
- Using bootstrapping to get confidence intervals on causal edges

Perform additional multilevel structural equation models:

- Investigate more factors associated with attributes of the open source system
- Evaluate whether a latent factor representing the “voice” of any architecture pattern might be helpful

Publish results:

- Comparison of different models
- Distinguish the causal influence of factors at both the file level and within a system

Convince others in the community to adopt Causal Learning and MSEM

© 2018 Carnegie Mellon University

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

Questions?

Robert Stoddard, SEI
rws@sei.cmu.edu

Mike Konrad, SEI
mdk@sei.cmu.edu

Rick Kazman, SEI
rkazman@sei.cmu.edu

David Danks, CMU
ddanks@cmu.edu

Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213

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