



**Software Engineering Institute**  
Carnegie Mellon University

# Causal Analysis and Software and Systems Engineering

Anandi Hira, Jim Alstad

[A.Hira@usc.edu](mailto:A.Hira@usc.edu), [Alstad@acm.org](mailto:Alstad@acm.org)

Tecolete Research; University of Southern California (retired)

Work with Mike Konrad of the Software Engineering  
Institute (SEI) at Carnegie-Mellon [1]

Orange County ACM Meeting  
May 19, 2021



# Presentation Outline

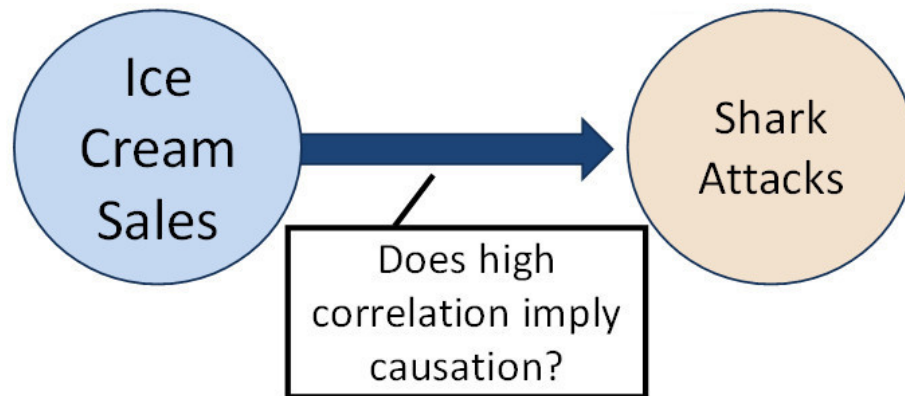
Note: We use “Causal Inference” in our talk to mean “Causal Analysis”

- Introduction to the Theory of Causal Inference
  - Intro to Causal Inference
  - How direction of causality is determined
  - Aspects of Causal Inference: Causal Search, Causal Estimation
  - Tetrad Demo
  - Summary
- Application to 2 Datasets
  - Description of Datasets
  - Causal Search Results
  - Causal Estimation Results
- Conclusions
- How to get started with Causal Inference
- End Matter
  - Bibliography
  - Backup slides

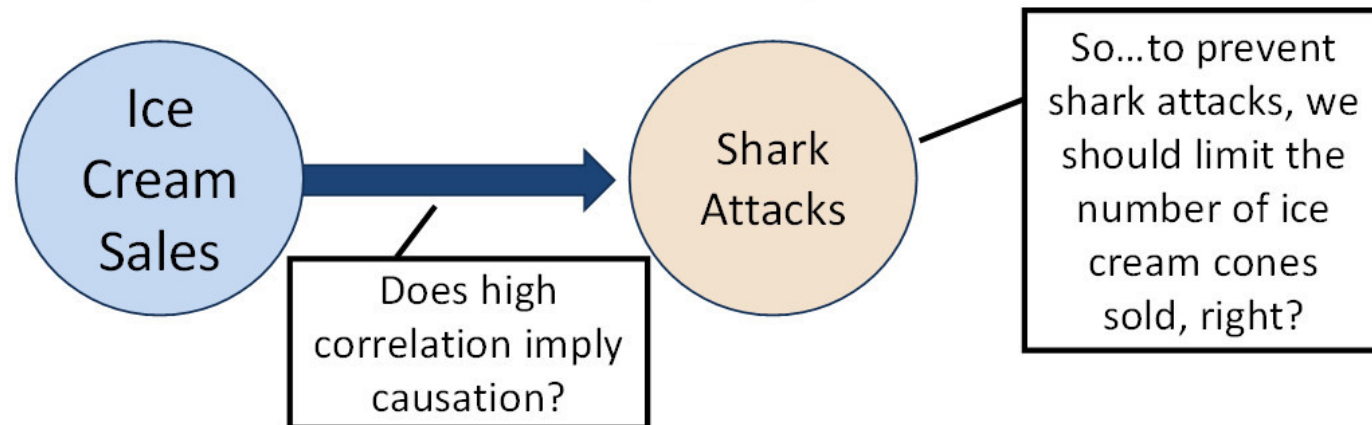
# Causal Inference: Introduction to the Theory

Following slides adapted from Carnegie Mellon University (CMU)

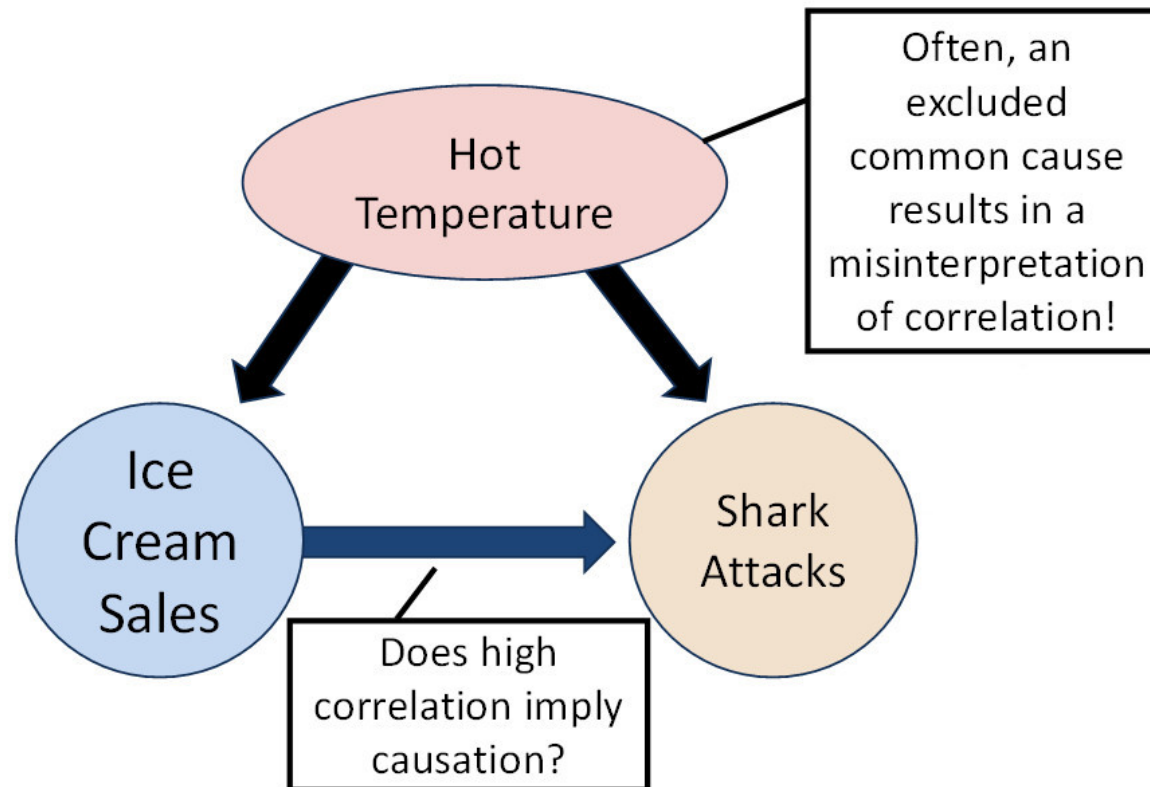
# Causal Model – An Example



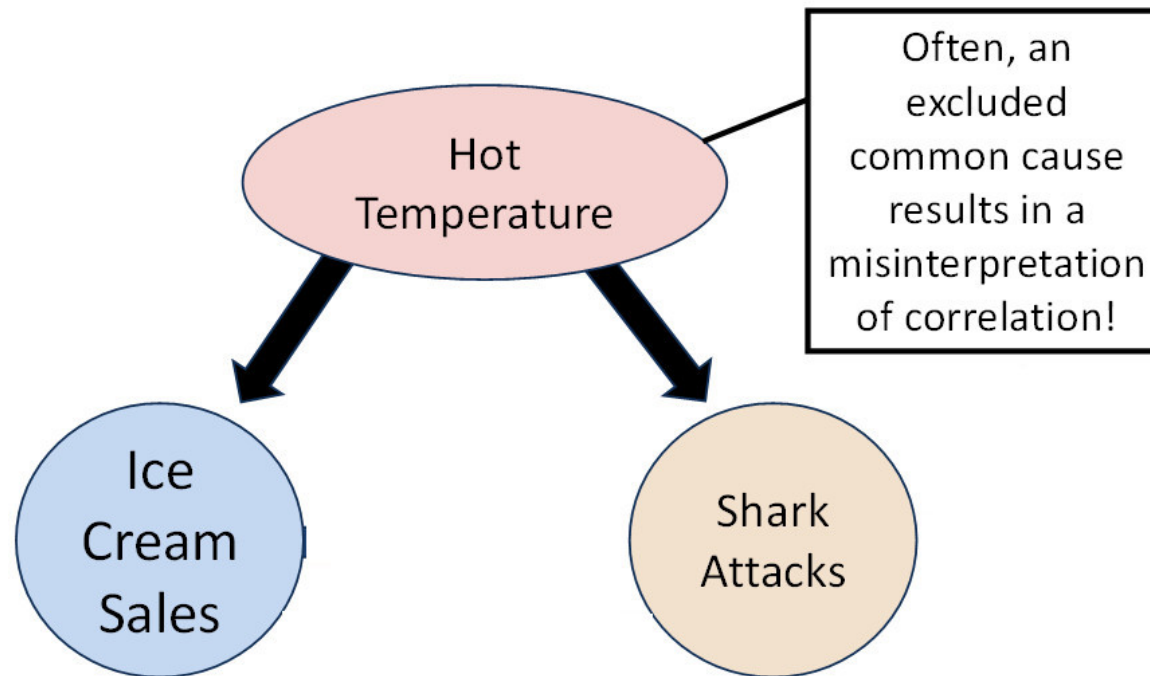
# Causal Model – An Example Cont.



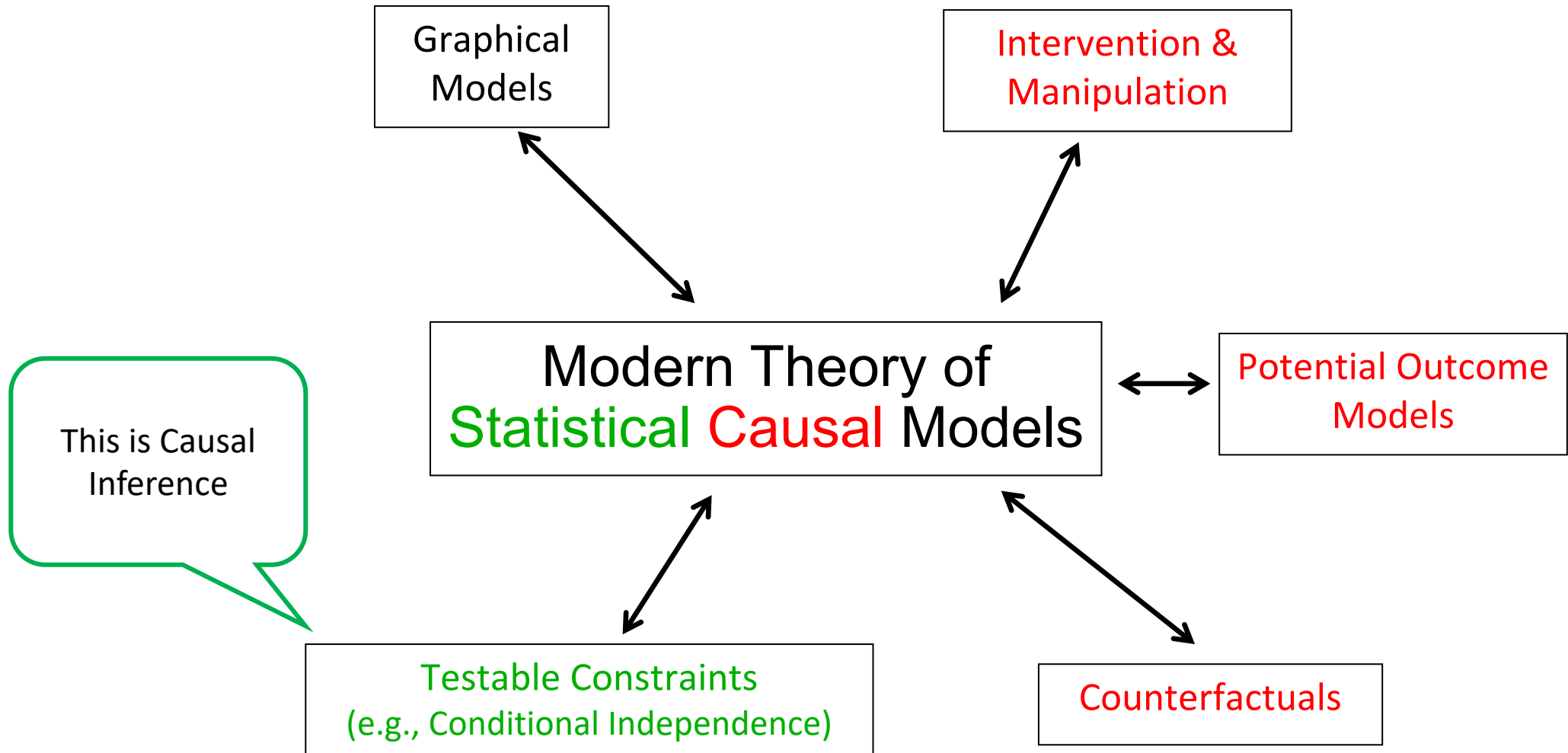
# Causal Model – An Example Cont.



# Causal Model – An Example Cont.



# Heritage of Causal Inference



# Heritage of Causal Inference (Cont.)

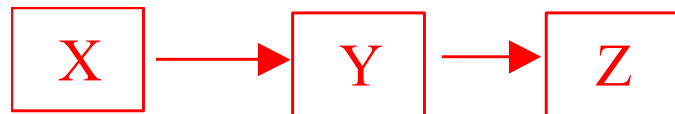
Causal  
Structure



Testable  
Statistical  
Predictions

*Causal Graphs*

e.g., Conditional Independence



$X \perp\!\!\!\perp Z \mid Y$

Now we can go from Testable Statistical Predictions to Causal Structure

# Definition of Conditional Independence

## Independence

- $X \perp\!\!\!\perp Z$  means “X and Z are independent”
- Statistically, this means that X and Z are not correlated

## Conditional Independence

- $X \perp\!\!\!\perp Z \mid Y$  means “Given any value of Y, X and Z are independent”
- Statistically, this means that, for each value of Y, X and Z are not correlated
  - If you take a subset of data for any single value of Y, X and Z are not correlated

# Conditional Independence: Implication of being Causally Disconnected

Weak Causal Markov Assumption

$X, Z$  causally disconnected  $\Rightarrow X \perp\!\!\!\perp Z$

$X, Z$  causally disconnected  $\Leftrightarrow$  No trek between  $X$  and  $Z$ , i.e.,

- i.  $X$  not a cause of  $Z$ , and
- ii.  $Z$  not a cause of  $X$ , and
- iii. There is no common cause  $Y$  of  $X$  and  $Z$

# Causal Graph Edge Adjacency: Variables Not Adjacent

*Per the previous slide,  
X and Z are not adjacent if they  
are *independent conditional* on any subset of variables  
that doesn't include X and Z*

# Presentation Outline

Note: We use “Causal Inference” in our talk to mean “Causal Analysis”

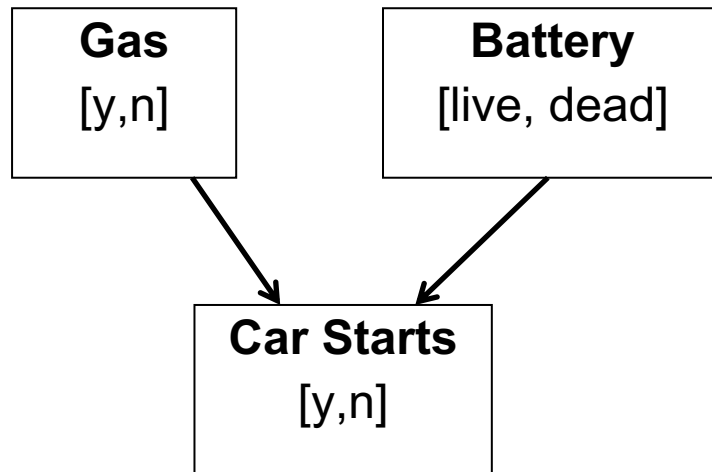
- Introduction to the Theory of Causal Inference
  - Intro to Causal Inference
  - How direction of causality is determined
  - Aspects of Causal Inference: Causal Search, Causal Estimation
  - Tetrad Demo
  - Summary
- Application to 2 Datasets
  - Description of Datasets
  - Causal Search Results
  - Causal Estimation Results
- Conclusions
- How to get started with Causal Inference
- End Matter
  - Bibliography
  - Backup slides

# Causal Graph Edge Adjacency: Interaction of 3 (Sets of) Variables

- Chain:  $X \rightarrow Y \rightarrow Z$ 
  - X and Z are conditionally independent given Y
- Fork:  $X \leftarrow Y \rightarrow Z$ 
  - X and Z are independent conditional on Y
  - As long as there is no other path between X and Z
- Collider:  $X \rightarrow Y \leftarrow Z$ 
  - X and Z are unconditionally independent, but dependent conditional on Y and any effects of Y
  - As long as there is only 1 path between X and Z

# Causal Graph Edge Adjacency: Collider and Chain Examples

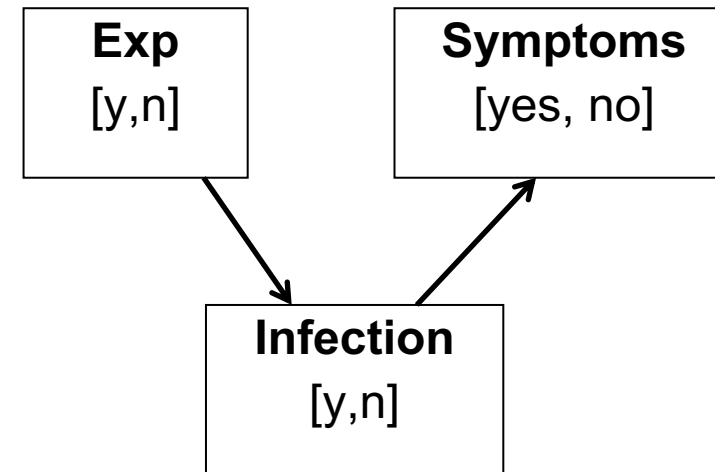
Conditioning on Colliders  
induce Association



Gas  $\perp\!\!\!\perp$  Battery

Gas  $\not\perp\!\!\!\perp$  Battery | Car starts = no

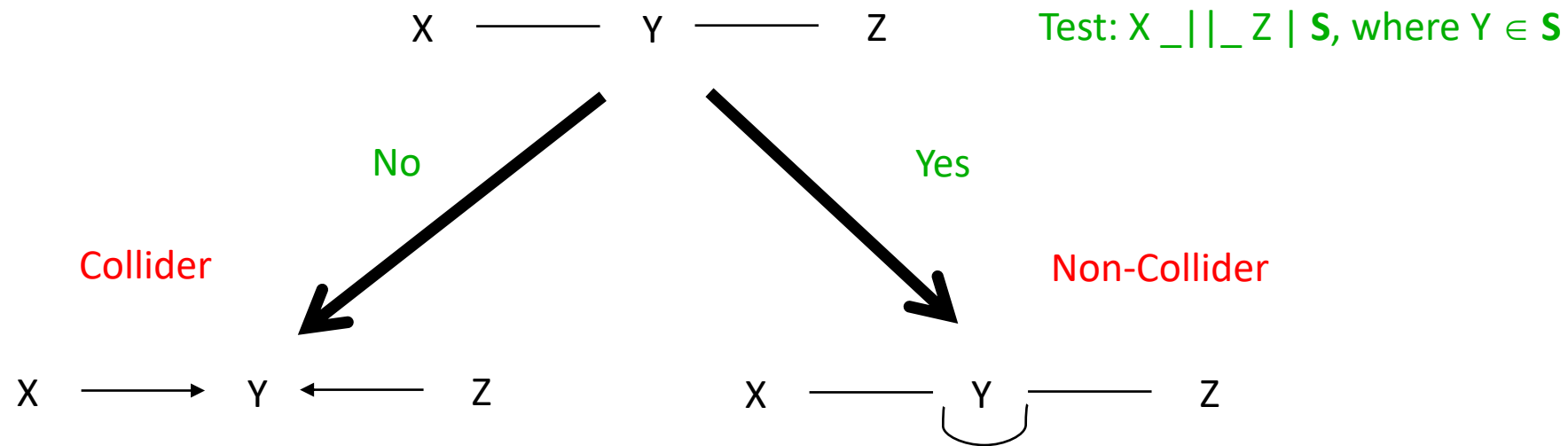
Conditioning on Non-Colliders  
screen-off Association



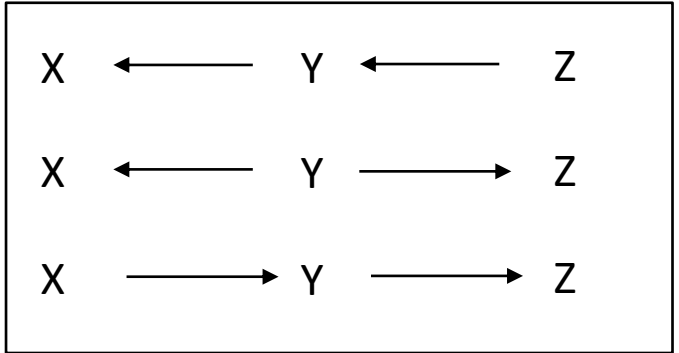
Exp  $\not\perp\!\!\!\perp$  Symptoms

Exp  $\perp\!\!\!\perp$  Symptoms | Infection

# Causal Graph Edge Adjacency: Trying To Statistically Determine Orientation



In the collider case, have determined the orientation of 2 edges



In the non-collider case, have not determined any orientation of edges

# Causal Graph Edge Adjacency: Conclusion

$X_1$        $X_2$        $X_1$  and  $X_2$  are not **adjacent**

---

$X_1$   $\longrightarrow$   $X_2$        $X_1 \rightarrow X_2$  ( $X_1$  is a **cause** of  $X_2$ )

---

$X_1$  —  $X_2$        $X_1 \rightarrow X_2$  or  $X_2 \rightarrow X_1$ ; orientation  
could not be determined.

# Presentation Outline

Note: We use “Causal Inference” in our talk to mean “Causal Analysis”

- Introduction to the Theory of Causal Inference
  - Intro to Causal Inference
  - How direction of causality is determined
  - Aspects of Causal Inference: Causal Search, Causal Estimation
  - Tetrad Demo
  - Summary
- Application to 2 Datasets
  - Description of Datasets
  - Causal Search Results
  - Causal Estimation Results
- Conclusions
- How to get started with Causal Inference
- End Matter
  - Bibliography
  - Backup slides

# Causal Inference

## Causal Search/Discovery

Algorithms and domain knowledge on observational data

## Causal Estimation

Algorithms to quantify causal influence; structural equation modeling (SEM)

# 2 Types of Causal Search Algorithms

## Constraint-Based

- Example: PC [4]
- Uses conditional independence to determine which causal relationships exist and their orientations

## Score-Based

- Example: FGES [2]
- Uses a step-by-step search to pick the next edge that best captures the causal aspects of the data

# Applying Constraints on the Algorithms

## Constraint-Based

- Alpha (usually 0.1, 0.05, or 0.01)
- = Allowed level of confidence for causal relationship
- Larger value is more lenient

## Score-Based

- Penalty Discount (usually 1)
- = Amount of penalty applied for cases where assumed graph does not match changes in data
- Smaller discount is more lenient

# Presentation Outline

Note: We use “Causal Inference” in our talk to mean “Causal Analysis”

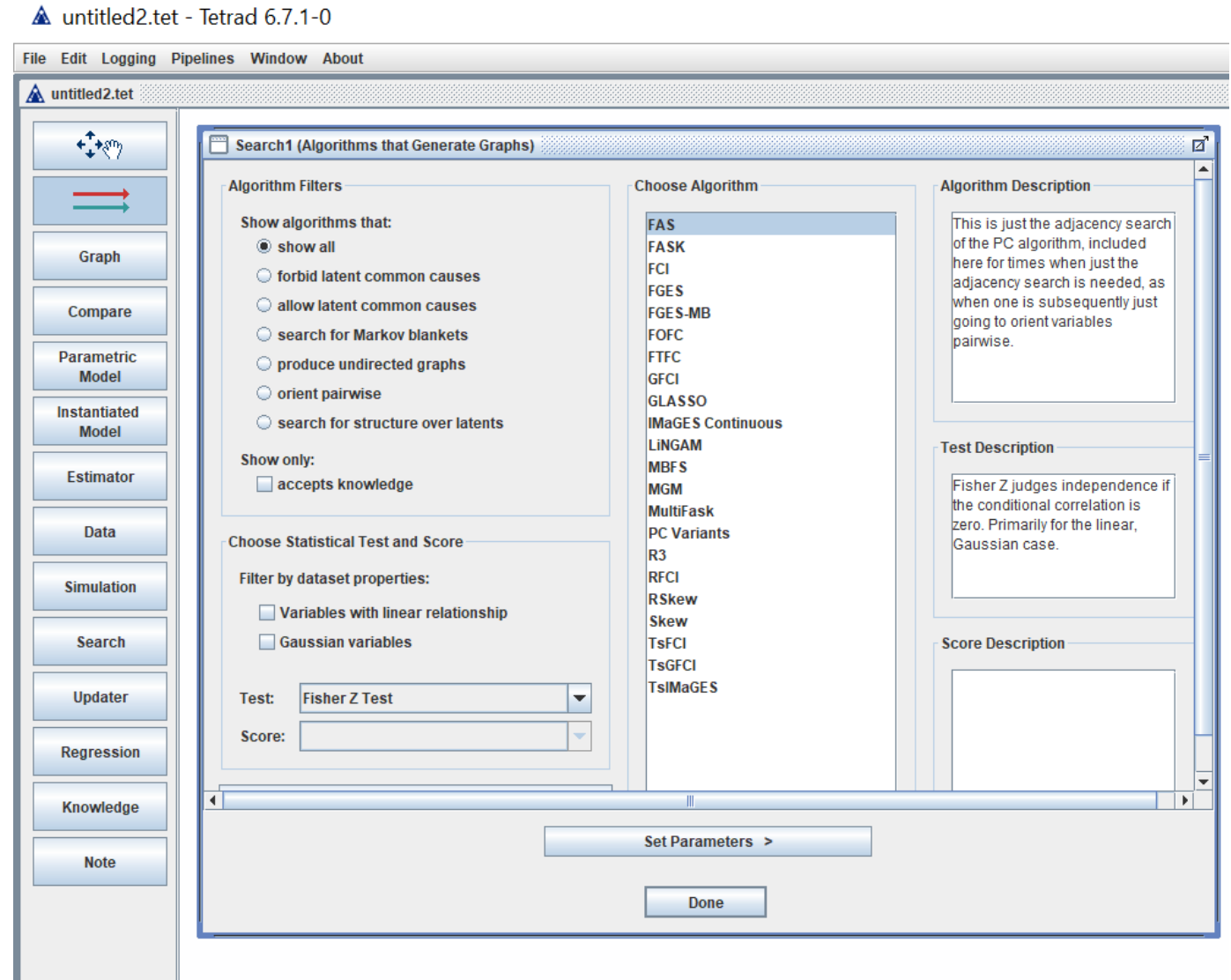
- Introduction to the Theory of Causal Inference
  - Intro to Causal Inference
  - How direction of causality is determined
  - Aspects of Causal Inference: Causal Search, Causal Estimation
  - Tetrad Demo
  - Summary
- Application to 2 Datasets
  - Description of Datasets
  - Causal Search Results
  - Causal Estimation Results
- Conclusions
- How to get started with Causal Inference
- End Matter
  - Bibliography
  - Backup slides

# Tetrad Demo

- Loading a dataset (Data)
- Running Causal Search algorithms (Algorithm)
- Running Bootstrap (part of Algorithm)
- Including Knowledge to improve results (Knowledge)
- Generating an Estimate (Estimator)

# Tetrad

- Implements the causal search algorithms [3]

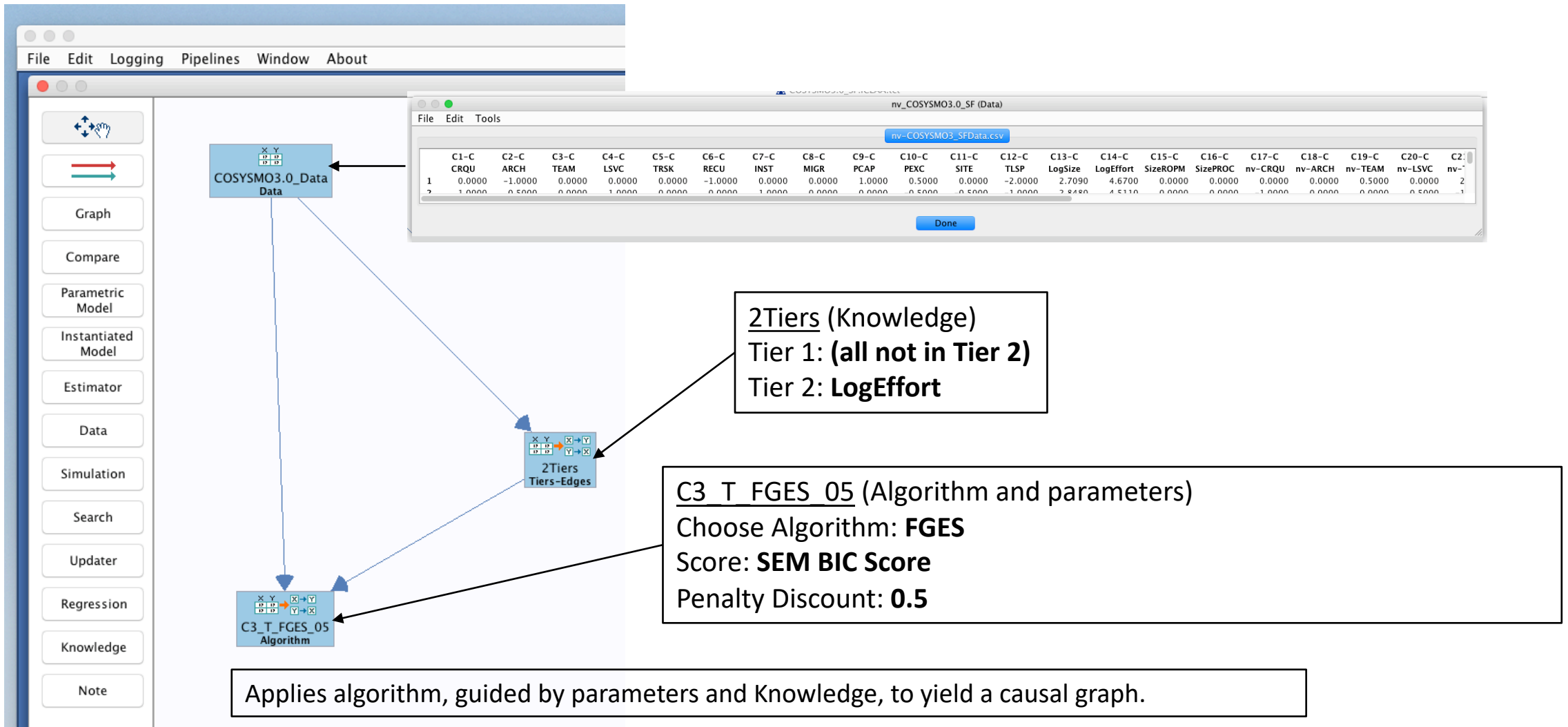


# Presentation Outline

Note: We use “Causal Inference” in our talk to mean “Causal Analysis”

- Introduction to the Theory of Causal Inference
  - Intro to Causal Inference
  - How direction of causality is determined
  - Aspects of Causal Inference: Causal Search, Causal Estimation
  - Tetrad Demo
  - Summary
- Application to 2 Datasets
  - Description of Datasets
  - Causal Search Results
  - Causal Estimation Results
- Conclusions
- How to get started with Causal Inference
- End Matter
  - Bibliography
  - Backup slides

# Tool Chain for Simple Tetrad Session



# Bootstrap Results: Edge Probability Table

Node 1	Interaction	Node 2	Ensemble	No edge
GrndSEITPM_BurnRate	-->	TotSEITPM_BurnRate	1.0000	0.0000
ATP_Yr	-->	EndPoP_Yr	0.9524	0.0000
EndPoP_Yr	<--	PoP_months	0.9524	0.0000
GOnPC_SEPM_TYcost	-->	SEcost	0.9048	0.0000
ELOCperHr	<--	ELOCperSM	0.8571	0.0000
GOnPC_SEPM_BYcost	<--	GOnPC_SEPM_TYcost	0.8095	0.0000
SysPMP_BYcost	<--	SysPMPcost	0.7619	0.0476
SysSEIT_BYcost	-->	TotSEITPM_BurnRate	0.7619	0.2381
HWCost	-->	HW_TYcost	0.7143	0.0000
SysSEITPMcost	-->	TotSEITPM_BurnRate	0.7143	0.2857
PMP_BYcost	-->	PMPnoITnCO_BYcost	0.6667	0.2381
SWDev_BYperHr	<--	SWDev_BYperSM	0.6667	0.0000
SwHwRatio	<--	SwTotRatio	0.6667	0.0952
SysPMPcost	-->	TotCost	0.6667	0.2381
SysSEITPMcost	-->	SysSEITPMratio	0.6667	0.1905
ATP_Yr	-->	GrndSEITPM_BurnRate	0.6190	0.0476

- Randomly sample dataset multiple times
  - Results less sensitive to outliers
- Results in an **edge probability table (EPT)**
  - Percentage of times edge found, reflecting the fraction of data points that has this direct-causal relationship
- Filter by probability of no edge (PNE) to keep only reasonable edges

# Causal Estimation

- **Causal estimation** involves parameterizing the relationships appearing in the causal search graph and then determining what values to assign to these parameters.
  - Enables making predictions about the values that variables will attain as a result of hypothesized events; i.e., allows making an estimating model.
  - Causal estimation, when applied to just a single variable and its direct causes works like ordinary linear regression: **coefficients** are assigned to each edge
  - A *one-unit* change in a direct cause, with all other variables held constant, results in a change in the child of *coefficient* units.
- The resulting model is then evaluated for **model fit**.
  - **Model fit statistics** include: Chi square (per degrees of freedom), Bayesian Information Criterion (BIC), Comparative Fit Index (CFI), and Root Mean Square Error of Approximation (RMSEA).
- More information can be found in [4, 5]

# Causal Inference Summary

1

- Allows analysts find causal relationships from observational data vs running experiments

2

- Uses interactions among variables to determine causal relationships

3

- Tetrad allows us to run these causal algorithms relatively easily and efficiently

# Presentation Outline

Note: We use “Causal Inference” in our talk to mean “Causal Analysis”

- Introduction to the Theory of Causal Inference
  - Intro to Causal Inference
  - How direction of causality is determined
  - Aspects of Causal Inference: Causal Search, Causal Estimation
  - Tetrad Demo
  - Summary
- Application to 2 Datasets
  - Description of Datasets
  - Causal Search Results
  - Causal Estimation Results
- Conclusions
- How to get started with Causal Inference
- End Matter
  - Bibliography
  - Backup slides

# Datasets

## **COCOMO<sup>®</sup> II Calibration Dataset**

- 16 organizations, various application types
- Variability in all 26 variables
- 161 projects
- See [6] for more details

## **COSYSMO 3.0 Calibration Dataset**

- Covers various types of systems
  - > 2 orders of magnitude size variation
- Variability in all 18 variables
- 68 projects
- See [7] for more details

Each dataset is reasonably representative of projects of its type

# Presentation Outline

Note: We use “Causal Inference” in our talk to mean “Causal Analysis”

- Introduction to the Theory of Causal Inference
  - Intro to Causal Inference
  - How direction of causality is determined
  - Aspects of Causal Inference: Causal Search, Causal Estimation
  - Tetrad Demo
  - Summary
- Application to 2 Datasets
  - Description of Datasets
  - Causal Search Results
  - Causal Estimation Results
- Conclusions
- How to get started with Causal Inference
- End Matter
  - Bibliography
  - Backup slides

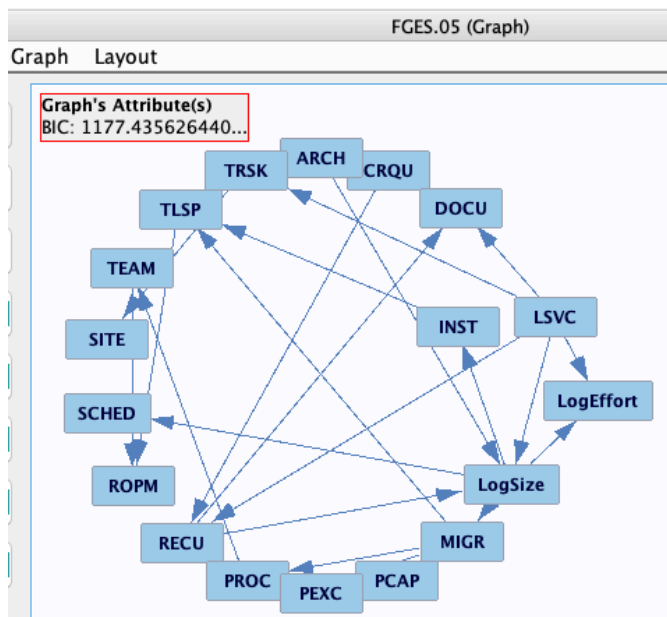
# Our Methodology for Causal Search on Small Samples

- **Problem:** Search did not produce structures that were very informative
  - Using strict parameter settings, few edges were found
    - Frequently only one: Size --> Effort
  - Using looser parameter settings, too many edges were found
    - Found additional plausible causes of Effort, but found non-plausible edges also
    - Such edges might be spurious (due to accidental correlations)—how would we know?
  - A consequence of having relatively few data points (projects)
- **Solution:** We invented an approach called weak-signal analysis (WSA), which consists of these steps, some based on the PNE (probability of no edge):
  1. Inject **null variables**: For each original variable, add a “null variable” column, a copy of the original variable values randomly sorted.
  2. Do causal search with **bootstrap**: determine for each edge terminating on a null variable (a “random edge”) its PNE
  3. Set a **trim threshold** at the **10<sup>th</sup> percentile** of random edge PNEs (i.e., 90% of random edges will have a higher PNE)
  4. Discard all edges among original variables whose PNE > trim threshold. Also, discard all null variables and random edges.

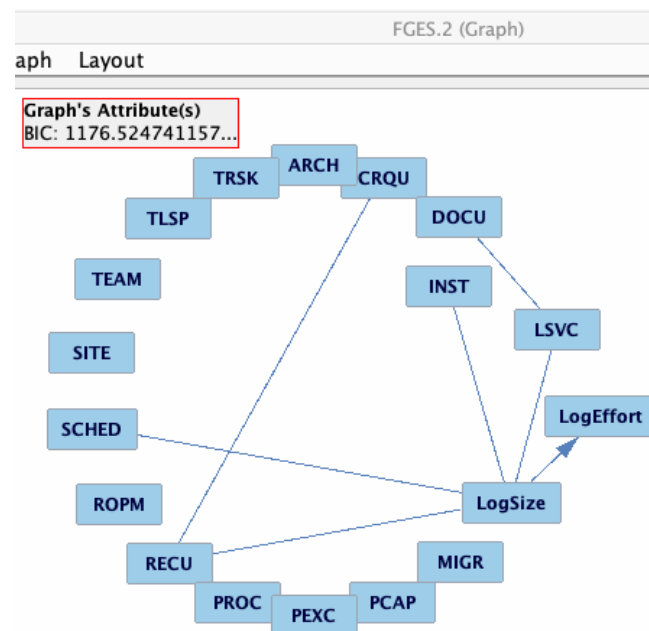
# Unsatisfactory Simple Tetrad Results

## Using COSYSMO 3.0 Dataset as an Example

Results of FGES Algorithm,  
with Penalty Discount = 0.5  
(less strict): 24 edges, with  
23 directed, & 2 direct  
causes of Effort



Results of FGES Algorithm,  
with Penalty Discount = 2.0  
(more strict): 9 edges, with  
1 directed, and 1 direct  
cause of Effort

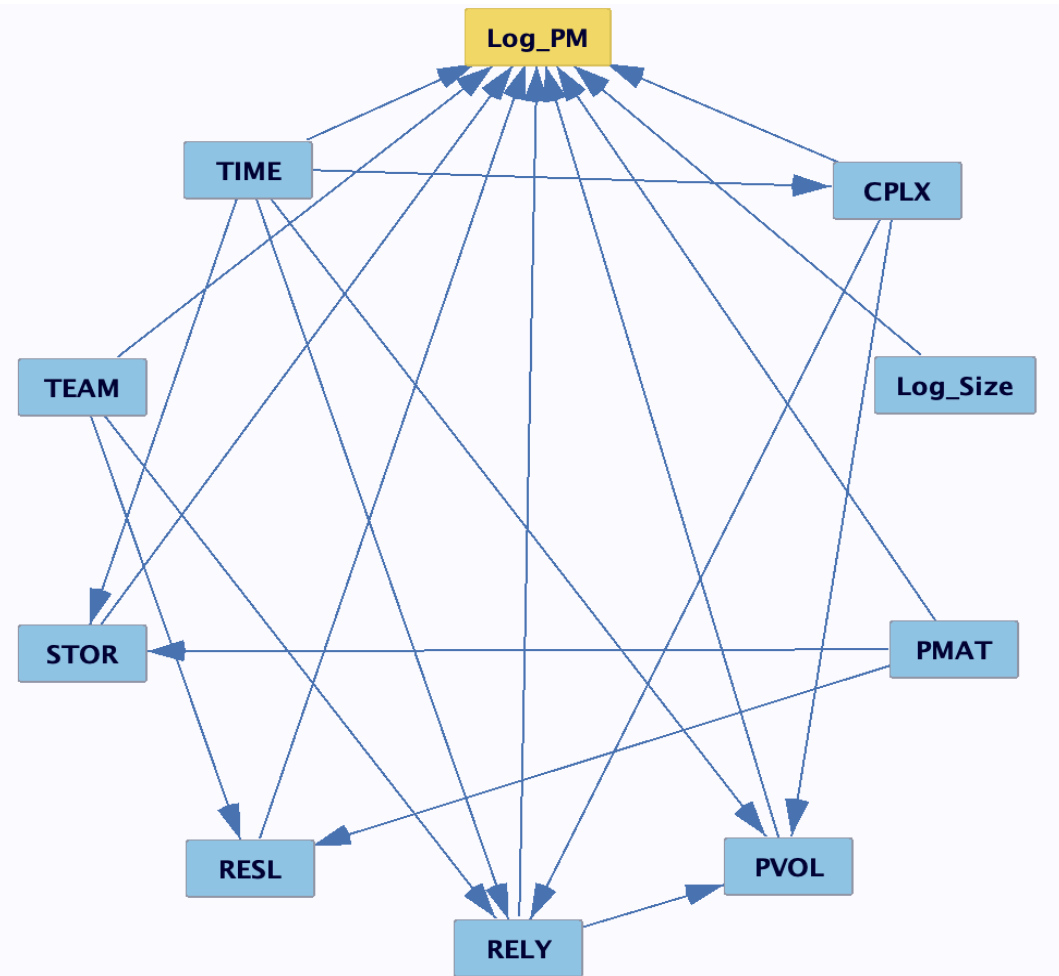


**Less strict simple search produces many spurious edges, while more strict simple search produces fewer useful results (direct causes of effort).**

# Direct Causes Using WSA of Software Engineering Effort

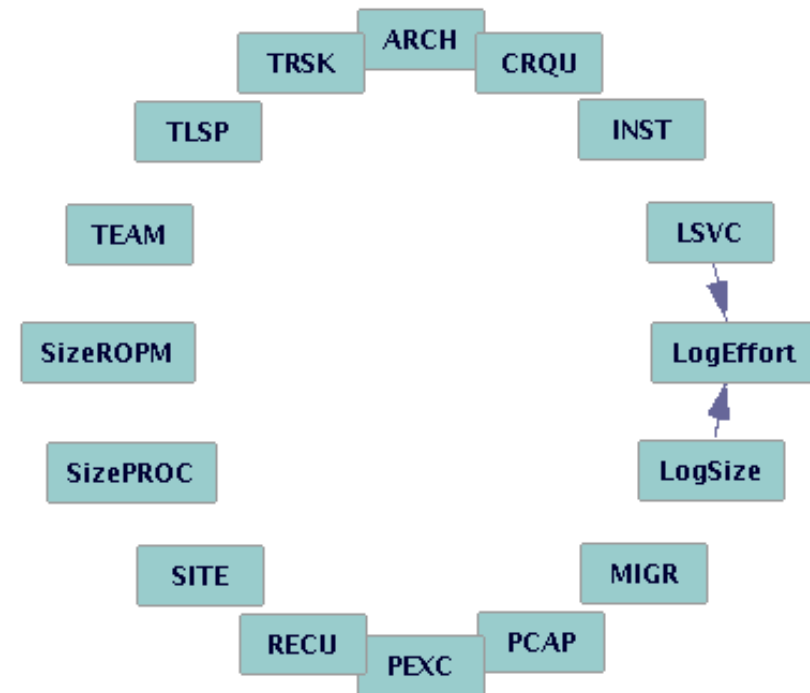
## Intervening on These in a Project May Improve Outcomes

- Size (SLOC)
- Team Cohesion (TEAM)
- Platform Volatility (PVOL)
- Reliability (RELY)
- Storage Constraints (STOR)
- Time Constraints (TIME)
- Product Complexity (CPLX)
- Process Maturity (PMAT)
- Risk and Architecture Resolution (RESL)



# Direct Causes Using WSA of Systems Engineering Effort Intervening on These in a Project May Improve Outcomes

- Size
- Level of Service Requirements (LSVC)



# Presentation Outline

Note: We use “Causal Inference” in our talk to mean “Causal Analysis”

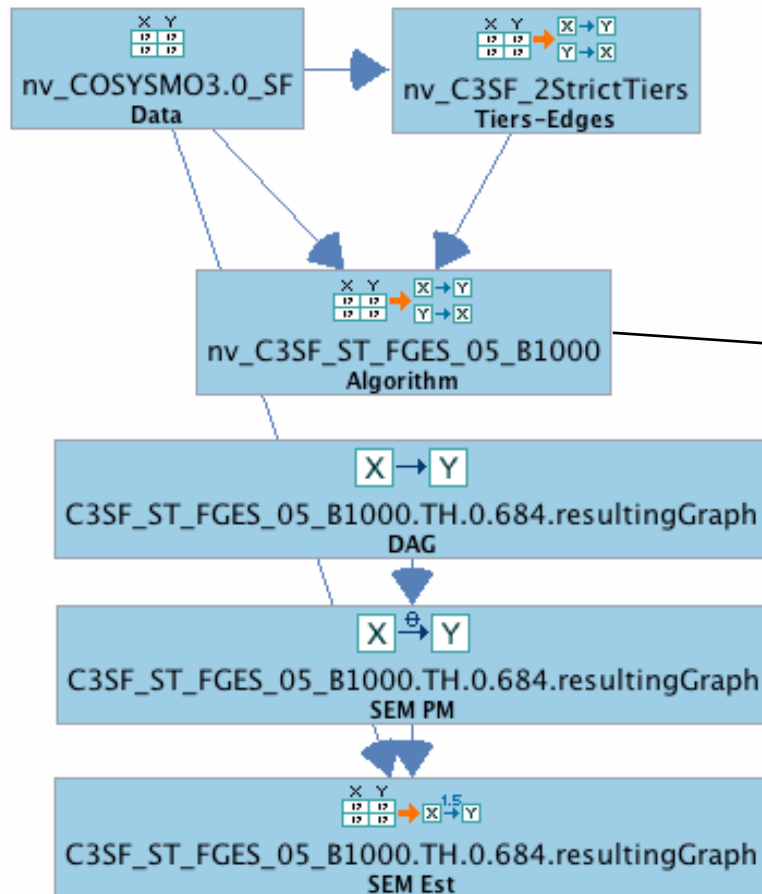
- Introduction to the Theory of Causal Inference
  - Intro to Causal Inference
  - How direction of causality is determined
  - Aspects of Causal Inference: Causal Search, Causal Estimation
  - Tetrad Demo
  - Summary
- Application to 2 Datasets
  - Description of Datasets
  - Causal Search Results
  - Causal Estimation Results
- Conclusions
- How to get started with Causal Inference
- End Matter
  - Bibliography
  - Backup slides

# Using Tetrad to Derive “Mini-Models” to Produce Plausible Cost Estimates

We were guided by the existing COCOMO® II and COSYSMO 3.0 estimating models' structure.

1. The structure of the estimating models does not directly conform to that needed by Tetrad. We therefore transformed the structure of each estimating equation:
  - We took the logarithm of the equation (Size -> LogSize, etc)
  - Cost drivers and scale factors are represented differently in the linear mini model.
  - Cost drivers are additive variables, which we directly included in the mini model.
  - Scale factors are multipliers of LogSize, we replaced each with the scale factor times LogSize.
2. We forced cost predictors to be independent of each other (with the Knowledge box in Tetrad).
3. We applied WSA to obtain a plausible causal graph. We discarded any variables that have no edges.
4. We used the Tetrad Estimation capability to obtain coefficients and intercepts on the resulting graph. The mini-model was obtained by extracting the mini-estimating equation from the resulting graph.
5. However, intercepts need further work (below).

# Tool Chain for Applying WSA



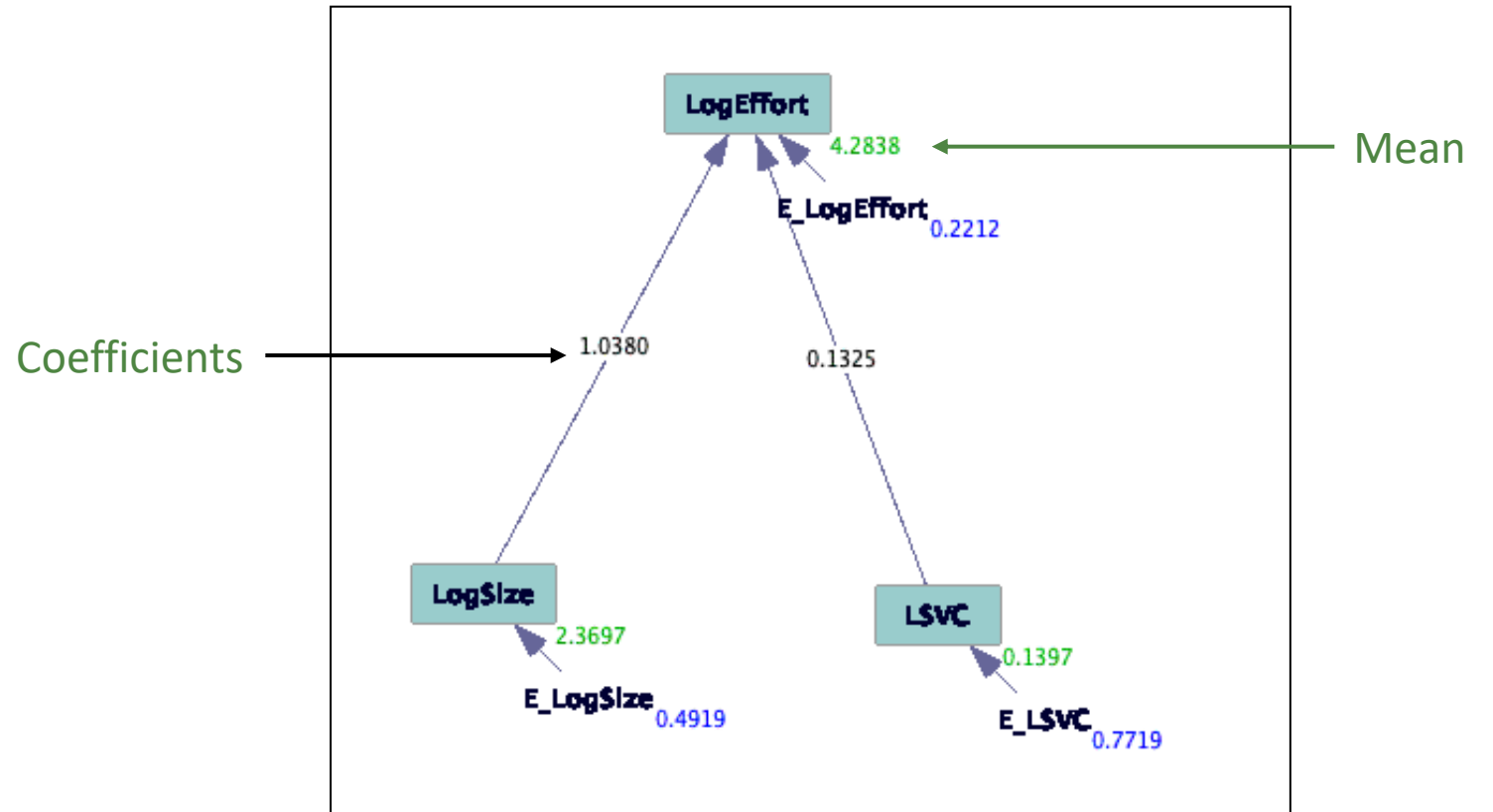
Top 3 boxes: per previous Tool Chain (slide 23)

Outside Tetrad, perform WSA analysis (per slide 33) on the PNE table (slide 27): Analyze the distribution of random edge PNEs to determine the 10<sup>th</sup>-percentile (trim) threshold; all edges with lower probability are deleted. (Therefore only those edges among the original variables whose PNE < trim threshold are retained.) Also discard all null variables, and all variables without edges. Import the resulting graph back into a Tetrad DAG box.

The SEM PM (Parametric Model) and SEM Est (Estimator) boxes, together with the Data, produce a graph annotated with estimation results.

# Resulting COSYSMO 3.0 Estimation

Model fit statistics (listed on slide 28)  
lead to the conclusion that:  
**This model fit is Poor-to-Fair.**

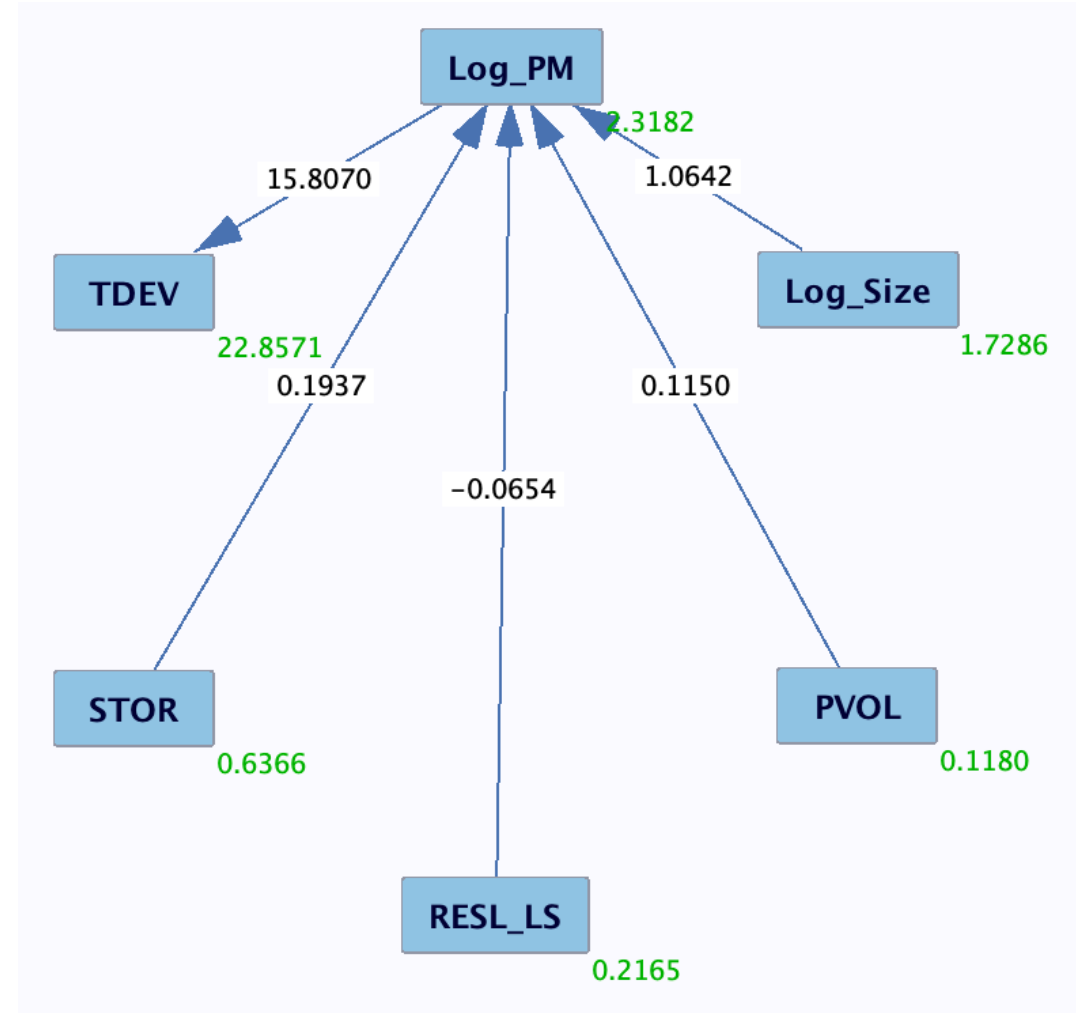


# How to Get a Mini-Model from that Tetrad Estimation Model

- Reading off from the Tetrad estimation model, a mini-model would be:  
$$\text{LogEffort} = 1.0380 * \text{LogSize} + 0.1325 * \text{LSVC} + 4.2838$$
- First attempt at Effort estimation:  
$$\text{Effort} = \text{Size}^{1.0380} * 1.357^{\text{LSVC}} * 10^{4.2838}$$
- That, however, doesn't work
  - The problem is that 4.2838 is the mean of the LogEffort values; however, raising 10 to that power does not yield the mean of the Effort values.
- One has to do a separate linear regression of LogEffort against LogSize and LSVC
  - That yields an exponent for 10 of 1.805, which gives this estimating equation:  
$$\text{Effort} = 63.834 * \text{Size}^{1.0380} * 1.357^{\text{LSVC}}$$

# Resulting COCOMO<sup>®</sup> II Estimation

Model fit statistics (listed on slide 28)  
lead to the conclusion that:  
**This model fit is Fair-to-Good.**



# Presentation Outline

Note: We use “Causal Inference” in our talk to mean “Causal Analysis”

- Introduction to the Theory of Causal Inference
  - Intro to Causal Inference
  - How direction of causality is determined
  - Aspects of Causal Inference: Causal Search, Causal Estimation
  - Tetrad Demo
  - Summary
- Application to 2 Datasets
  - Description of Datasets
  - Causal Search Results
  - Causal Estimation Results
- Conclusions
- How to get started with Causal Inference
- End Matter
  - Bibliography
  - Backup slides

# Conclusion – Causal Search

- Straightforward use of causal discovery algorithms may result in little information about cost-causing factors
  - Relatively small datasets (# of cases) compared to # of variables
- Weak-Signal Approach (WSA) enhanced results
  - Identified additional causes of effort and duration, while minimizing spurious correlations
  - Established a principled approach (methodology) to determining what cutoff to use for trimming results of a bootstrapped search (based on null variables and EPT)
- We identify (on slides 35 & 36) specific direct causes, where action has been shown statistically to cause the cost or schedule
  - The data we used considered multiple application types and multiple organizations
  - We also investigated choice of Tetrad search algorithm and parameter values

# Conclusion – Causal Estimation

- We developed a methodology (slide 33) for generating cost estimation mini-models based on datasets that deliver plausible results
  - Based mostly on features built in to Tetrad
  - Tetrad can deliver off-the-shelf models, if logarithms don't need to be applied to data
- Observation
  - Modestly fitting with somewhat inferior predictions compared to original model

# Presentation Outline

Note: We use “Causal Inference” in our talk to mean “Causal Analysis”

- Introduction to the Theory of Causal Inference
  - Intro to Causal Inference
  - How direction of causality is determined
  - Aspects of Causal Inference: Causal Search, Causal Estimation
  - Tetrad Demo
  - Summary
- Application to 2 Datasets
  - Description of Datasets
  - Causal Search Results
  - Causal Estimation Results
- Conclusions
- How to get started with Causal Inference
- End Matter
  - Bibliography
  - Backup slides

# How You Can Get Started with Causal Inference

- We encourage you to try Tetrad on your data
  - Find out what the most important (i.e., causal) factors are
  - If results are unsatisfactory, consider using WSA
- Applicable to an organizational project database of moderate or larger size
- Get Tetrad [3] and the Tetrad Manual [8]
- Obtain training in Causal Discovery and Tetrad [9, 10]

# Presentation Outline

Note: We use “Causal Inference” in our talk to mean “Causal Analysis”

- Introduction to the Theory of Causal Inference
  - Intro to Causal Inference
  - How direction of causality is determined
  - Aspects of Causal Inference: Causal Search, Causal Estimation
  - Tetrad Demo
  - Summary
- Application to 2 Datasets
  - Description of Datasets
  - Causal Search Results
  - Causal Estimation Results
- Conclusions
- How to get started with Causal Inference
- End Matter
  - Bibliography
  - Backup slides

# Bibliography

1. Hira, A., Alstad, JP., Konrad, M. (2020). *Discovering Causal Effects of Software and Systems Engineering Effort*. Joint IT and Software Cost Forum 2020. Department of Homeland Security.
2. Center for Causal Discovery (CCD). Fast Greedy Search (FGES) Algorithm for Continuous Variables. [https://www.ccd.pitt.edu/wp-content/uploads/2018/10/FGES1c-user-documentation-5\\_21\\_2016-sample-size.pdf](https://www.ccd.pitt.edu/wp-content/uploads/2018/10/FGES1c-user-documentation-5_21_2016-sample-size.pdf)
3. Center for Causal Discovery (CCD), a partnership among data scientists from the [University of Pittsburgh](#) (Pitt), [Carnegie Mellon University](#) (CMU), and the [Pittsburgh Supercomputing Center](#) (PSC), Tetrad Software. <https://www.ccd.pitt.edu/tools/>
4. Spirtes, P. (2010). Introduction to causal inference. *Journal of Machine Learning Research*. 11, May (2010), 1643–1662.
5. Kline, R. (2015). Principles and Practice of Structural Equation Modeling, Fourth Edition. Series: Methodology in the Social Sciences. The Guilford Press.
6. Boehm, B. W., et al. (2000). *Software Cost Estimation with COCOMO II*. Upper Saddle River, NJ: Prentice-Hall, Inc.
7. Alstad, JP. (2019). “Development of COSYSMO 3.0”, *Procedia Computer Science* 153
8. Center for Causal Discovery (CCD), Tetrad Manual: <http://cmu-phil.github.io/tetrad/manual/>
9. Center for Causal Discovery (CCD), Training in Causal Discovery from those who pioneered Tetrad: <https://www.ccd.pitt.edu/video-tutorials/>
10. SEI, Training in Causal Discovery from the SEI: contact Mike Konrad (also the source for WSA Python scripts): [MDK@SEI.CMU.edu](mailto:MDK@SEI.CMU.edu)

# Backup Slides

# Prediction Accuracy: Mini-Models vs Estimating Models

## COSYSMO 3.0 - Effort

	Mini-Model	Original
<b>Max MRE</b>	285.4%	234.8%
<b>MMRE</b>	45.9%	57.3%
<b>PRED(25)</b>	41.2%	23.5%
<b>PRED(30)</b>	48.5%	23.5%

## COCOMO<sup>®</sup> II - Effort

	Mini-Model	Original
<b>Max MRE</b>	455.4%	229.41%
<b>MMRE</b>	38.64%	25.67%
<b>PRED(25)</b>	44.72%	67.08%
<b>PRED(30)</b>	52.8%	74.53%