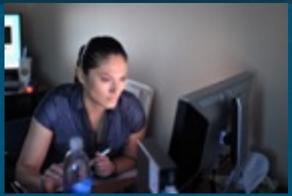
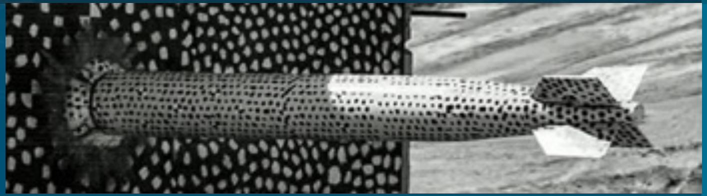
Sandia National Laboratories

Causal Discovery for Climate Science and the Energy Exascale Earth System Model





Presented by

J. Jake Nichol, 1461

U.S. DEPARTMENT OF ENERGY NNSA

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

What is Causality?

Cause → Effect

What is Causality?

Cause → Effect

Peter Spirtes
Clark Glymour
Richard Scheines
David Rubin
Judea Pearl

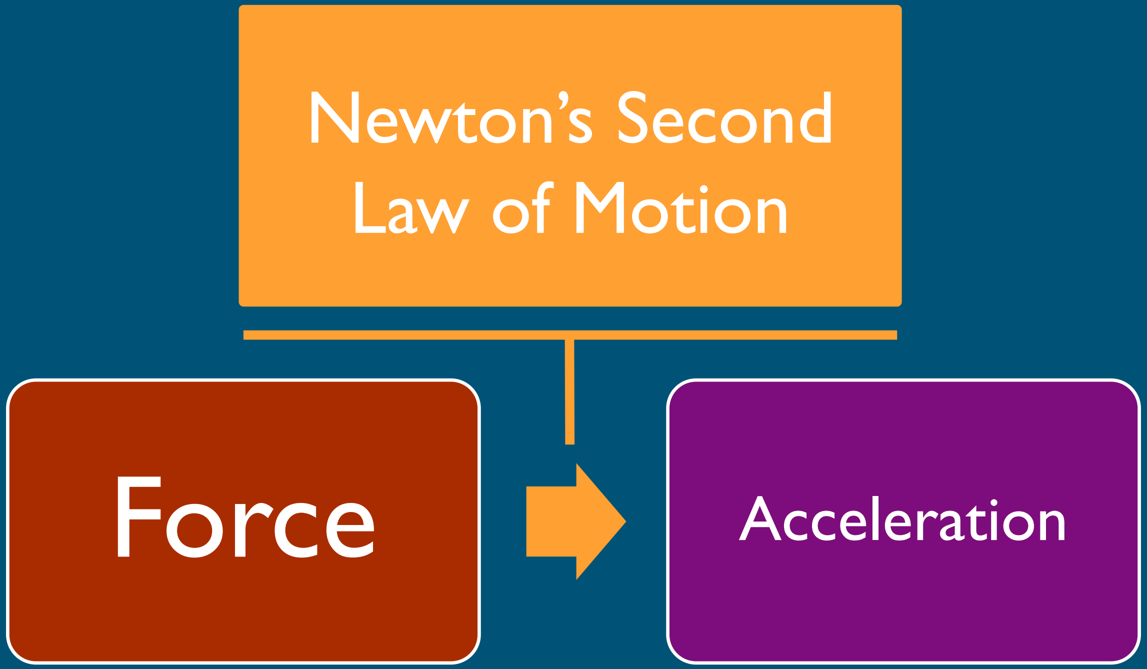
What is Causality?

Cause → Effect

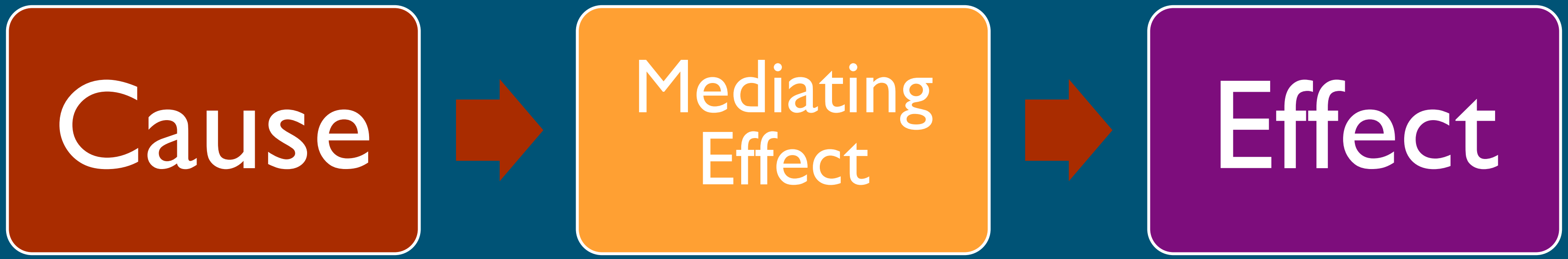
What does this arrow represent?

How can we determine it is there?

What is Causality?



What is Causality?



What is Causality?

Cause → Mediating Effect → Effect

Water plant

Plant grows flowers

Plant bears fruit

What is Causality?

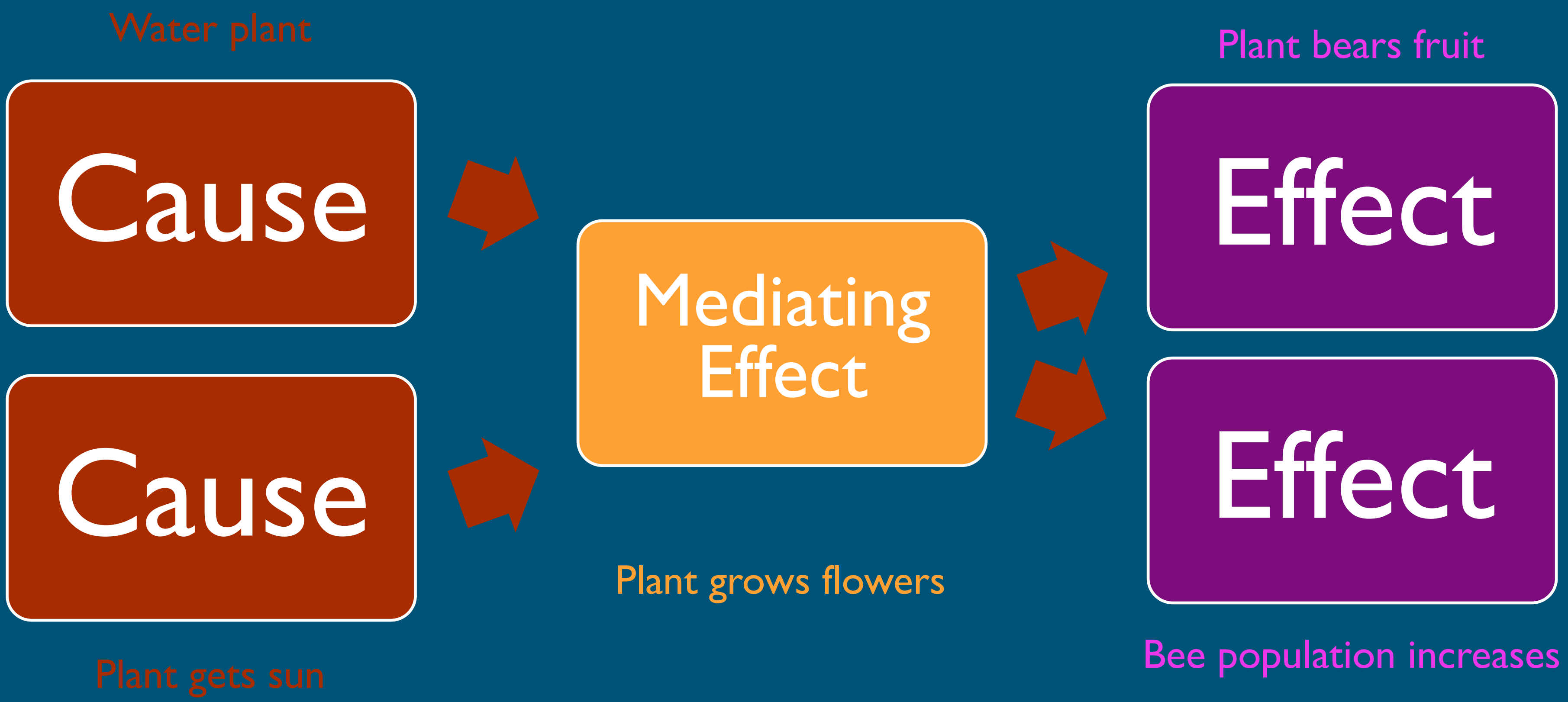
Cause → Mediating Effect → Effect

Water plant

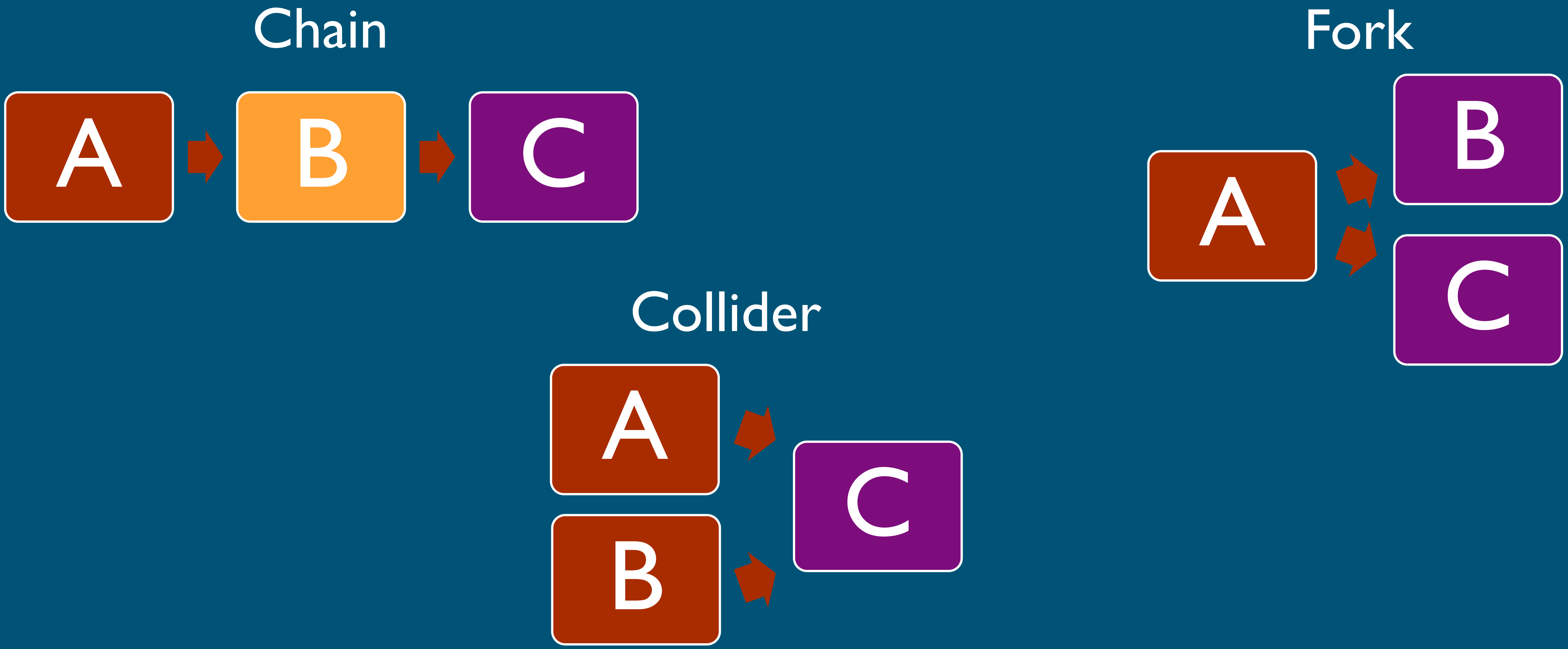
Plant grows flowers

Plant bears fruit

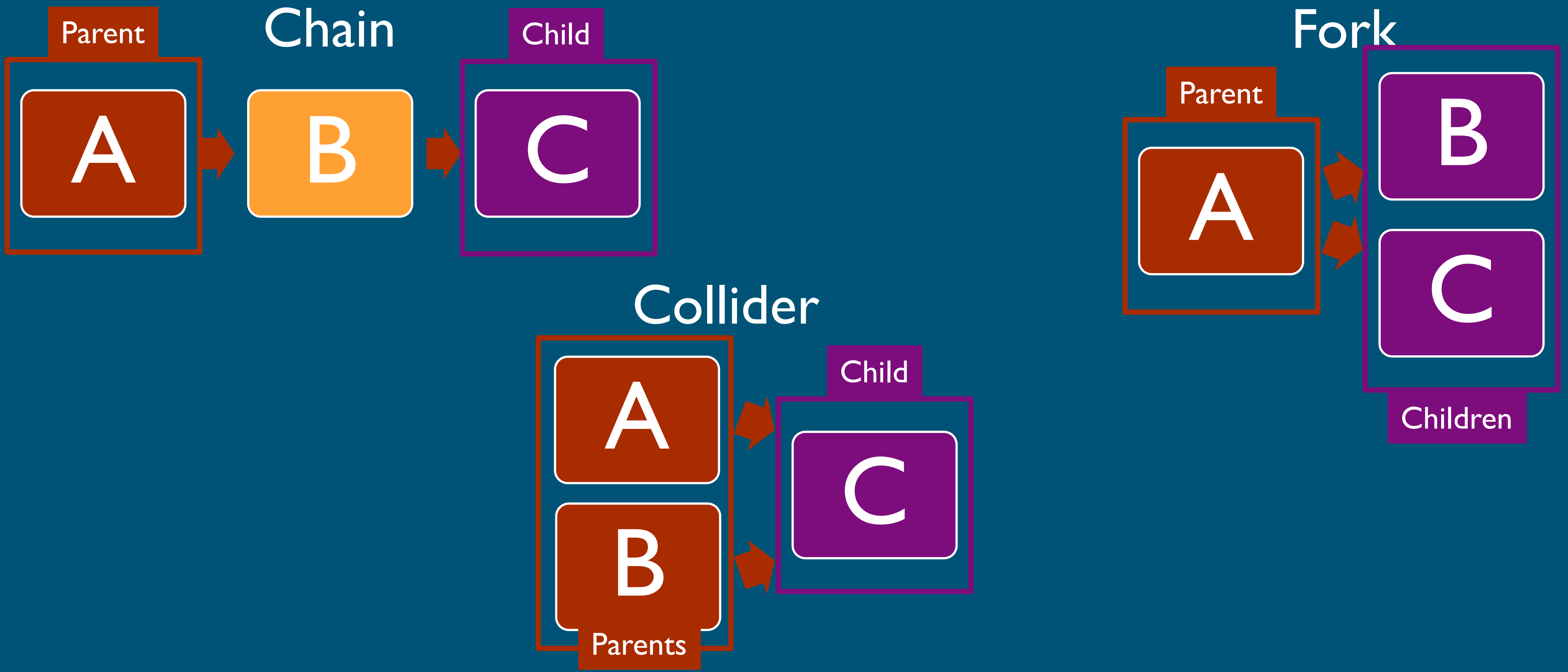
What is Causality?



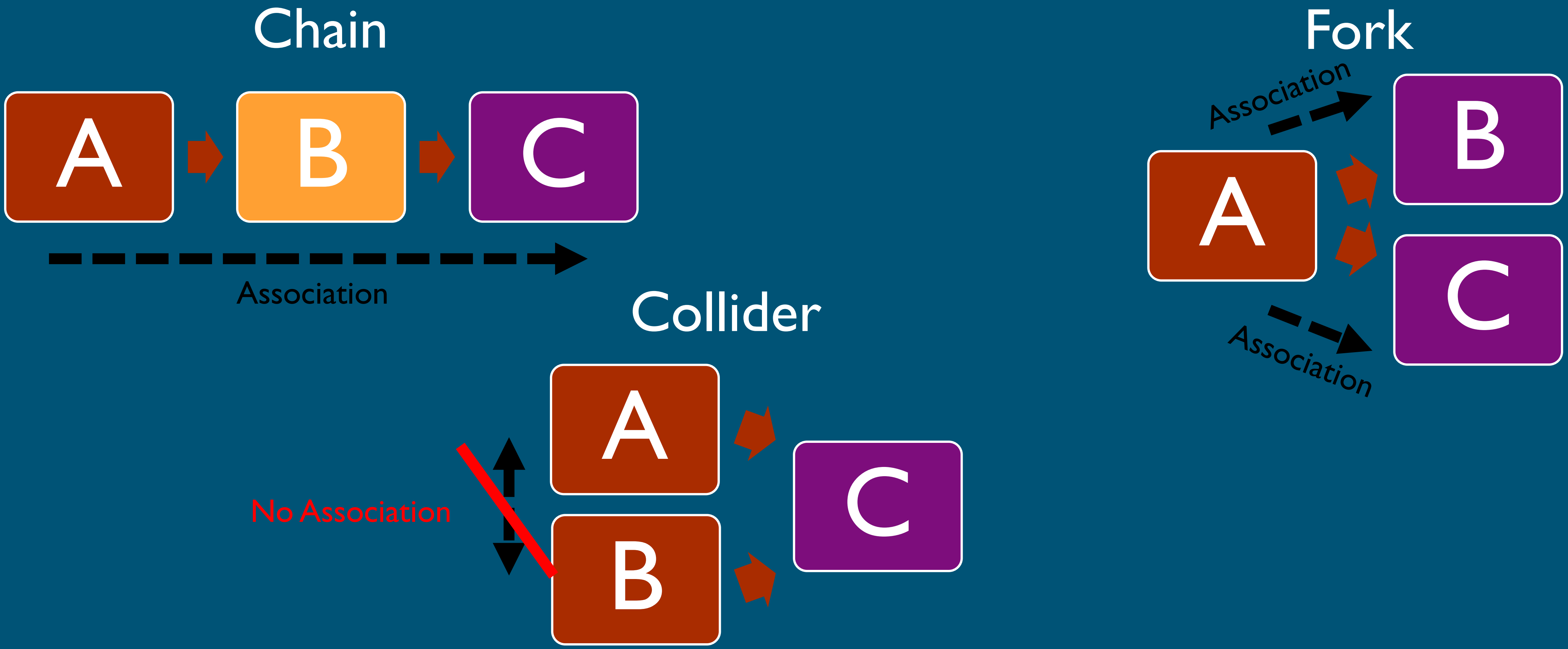
What is Causality?



What is Causality?



What is Causality?



Causal Discovery

A. Peter Spirtes & Clark Glymour (PC) algorithm
   - Causal network learning algorithm

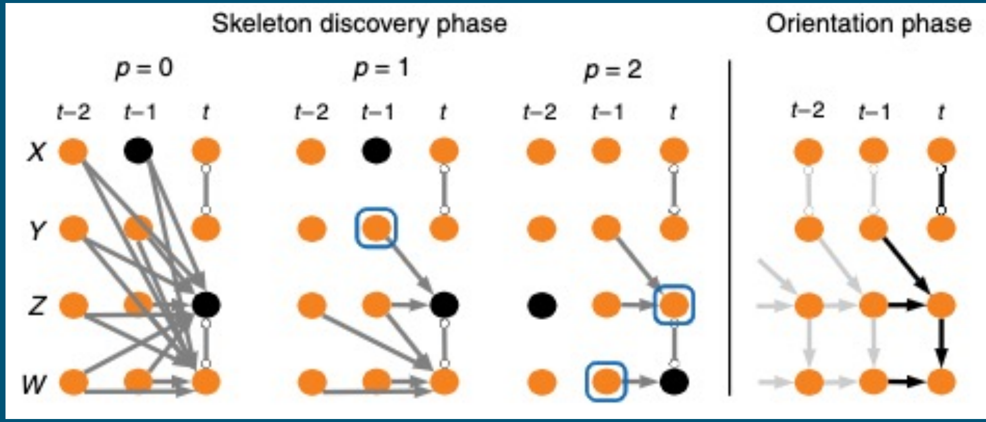
B. PC & Momentary Conditional Independence (PCMCI)
   - Extension to PC to handle false positives & high dimensionality

C. Fast Causal Inference (FCI) algorithm
   - Generalization of PC that does not require Causal Sufficiency

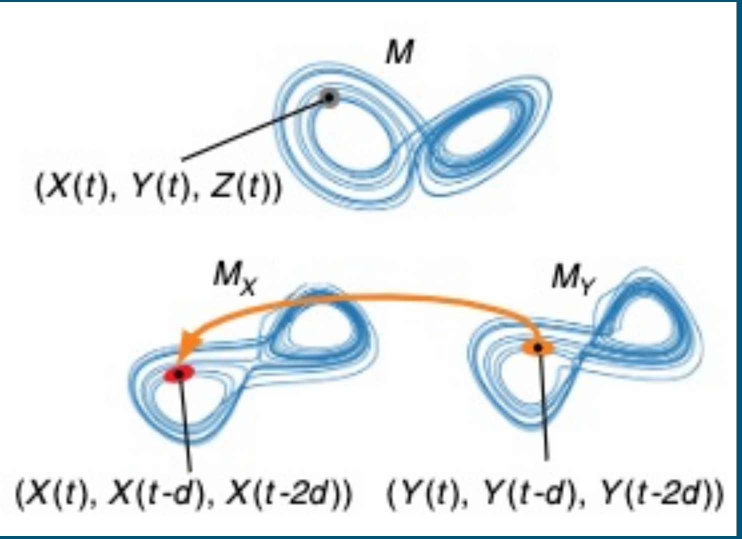
D. LiNGAM
   - For identifying Linear, Non-Gaussian, Acyclic causal Models based on purely observational, continuous-valued data
   - Structural Equation/Causal Modeling (SEM or SCM)

E. Convergent cross mapping
   - Uses Taken's theorem of Lorenz attractors to deconstruct a dynamical system's state space and infer causal pairs.

A.



E.



Jakob Runge, et al. 2019. Inferring causation from time series in Earth system sciences. *Nat Commun* 10, 1 (2019). DOI:https://doi.org/10.1038/s41467-019-10105-3

Causal Discovery

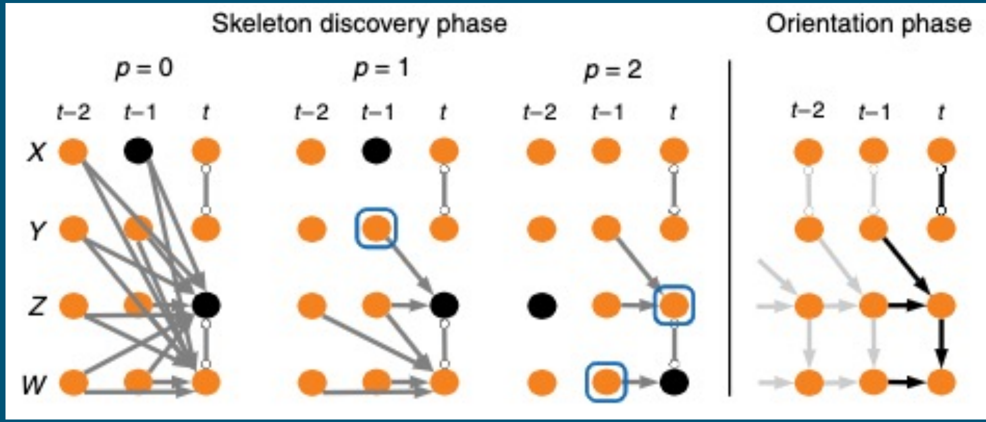
Independence/Constraint-Based Causal Network Learning

A. Peter Spirtes & Clark Glymour (PC) algorithm
   - Causal network learning algorithm

B. PC & Momentary Conditional Independence (PCMCI)
   - Extension to PC to handle false positives & high dimensionality

C. Fast Causal Inference (FCI) algorithm
   - Generalization of PC that does not require Causal Sufficiency

A.



Jakob Runge, et al. 2019. Inferring causation from time series in Earth system sciences. *Nat Commun* 10, 1 (2019). DOI:https://doi.org/10.1038/s41467-019-10105-3

Independence-Based Causal Discovery

Assumptions for independence-based causal discovery:

Causal Sufficiency: there are no unobserved confounders of any variables in the graph

Markov Assumption: $X \perp\!\!\!\perp_G Y \mid Z \implies X \perp\!\!\!\perp_P Y|Z$

- If X and Y are independent in a graph, G, given Z, then they must be statistically independent in their joint probabilities, given Z.

Faithfulness: $X \perp\!\!\!\perp_G Y \mid Z \impliedby X \perp\!\!\!\perp_P Y|Z$

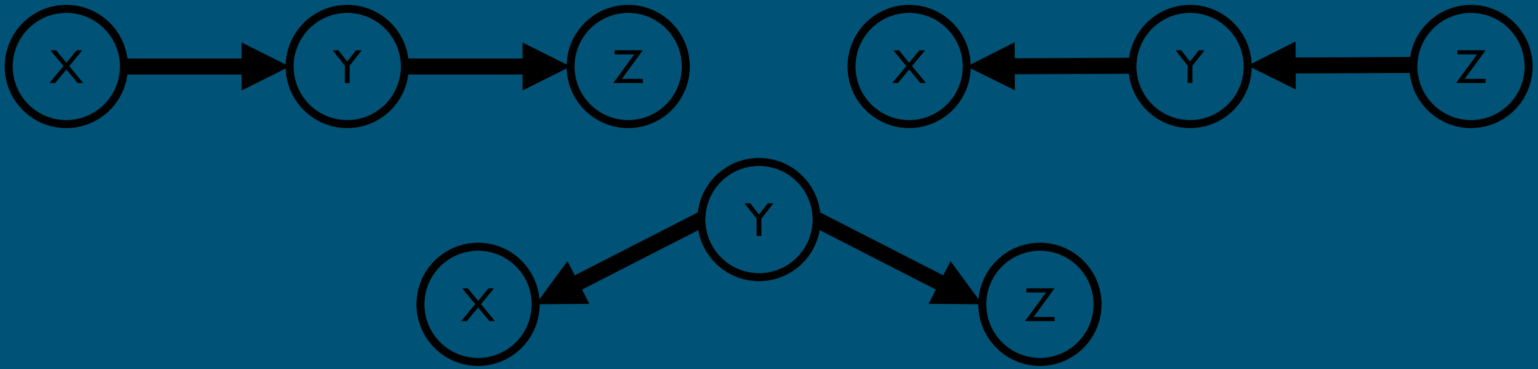
- If X and Y statistically independent in their joint probabilities, given Z, then they must be independent in the graph, G, conditioned on Z.

Acyclicity: assume there are no cycles in the graph

Independence-Based Causal Discovery

Markov Equivalence Classes

Chains and forks encode the same independencies:



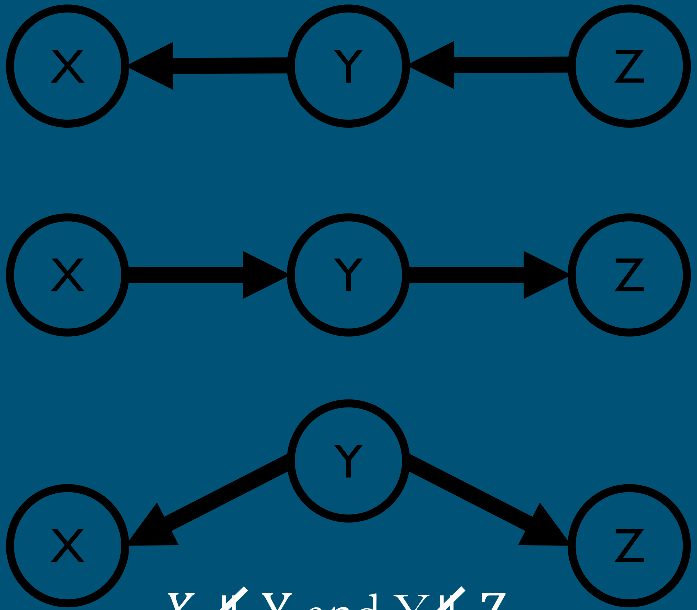
$X \not\perp\!\!\!\perp Y$ and $Y \not\perp\!\!\!\perp Z$

$X \not\perp\!\!\!\perp Z$

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

Independence-Based Causal Discovery

Markov Equivalence Classes



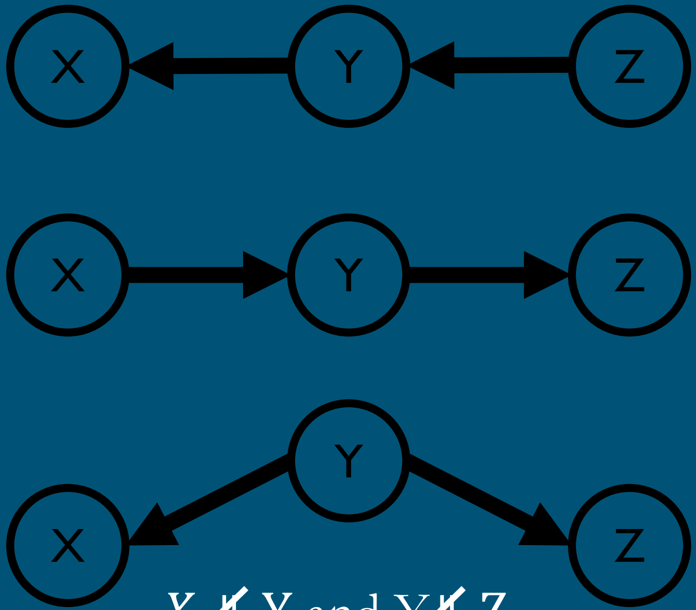
$X \not\perp\!\!\!\perp Y$ and $Y \not\perp\!\!\!\perp Z$

$X \not\perp\!\!\!\perp Z$

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

Independence-Based Causal Discovery

Markov Equivalence Classes

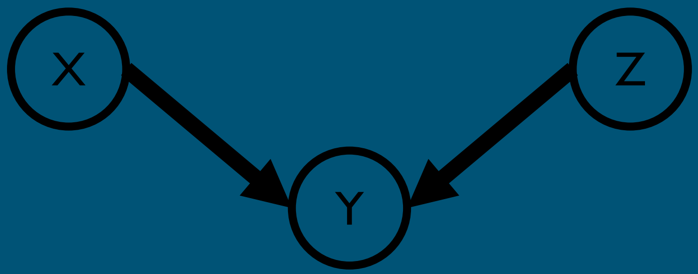


$X \not\perp\!\!\!\perp Y$ and $Y \not\perp\!\!\!\perp Z$

$X \not\perp\!\!\!\perp Z$

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

Colliders encode a unique independence relationship:

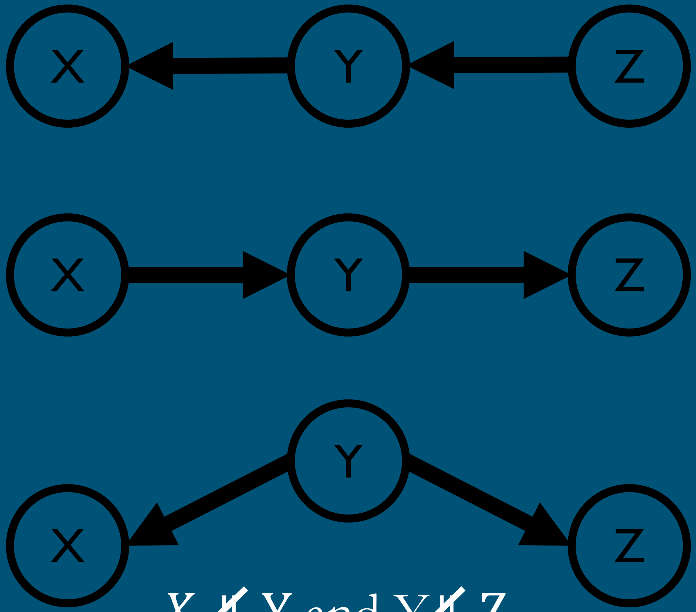


$X \perp\!\!\!\perp Z$ X is independent of Z, conditional on nothing

$X \not\perp\!\!\!\perp Z \mid Y$

Independence-Based Causal Discovery

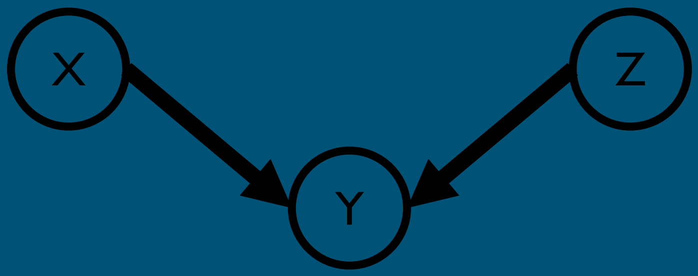
<u>Markov Equivalence Classes</u>



$X \not\perp\!\!\!\perp Y$ and $Y \not\perp\!\!\!\perp Z$

$X \not\perp\!\!\!\perp Z$

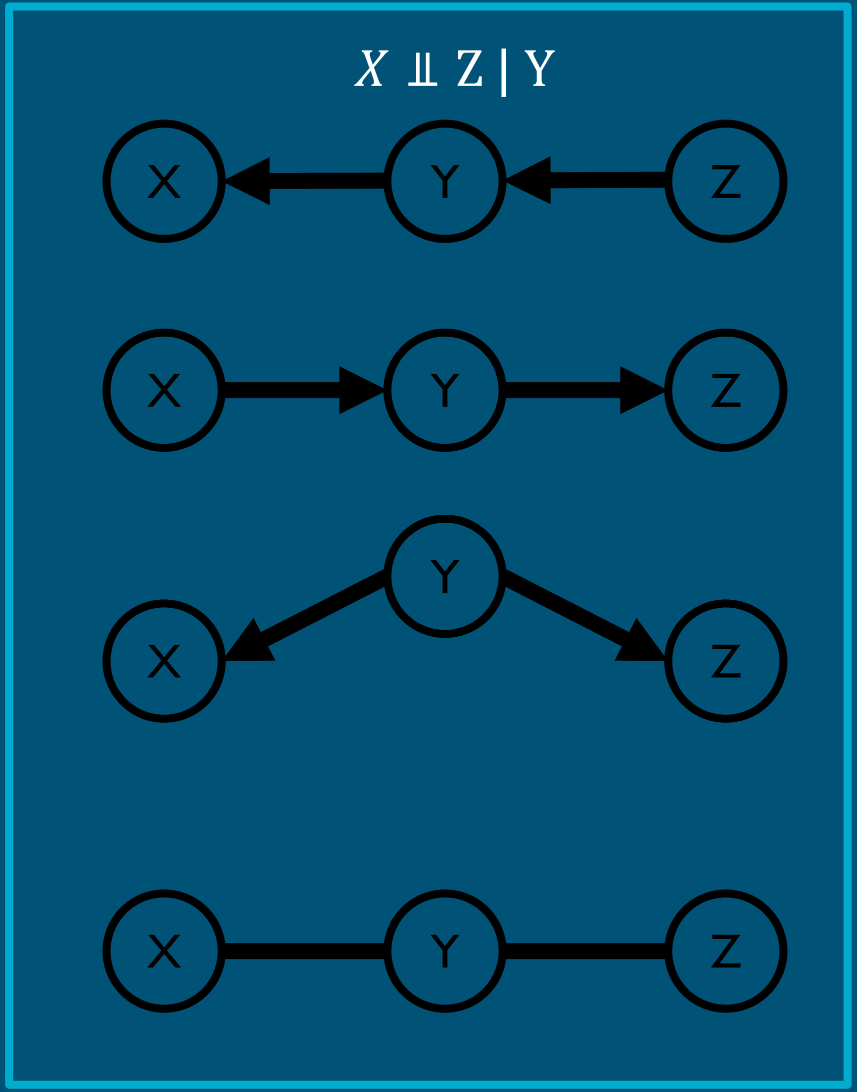
$X \perp\!\!\!\perp Z \mid Y$



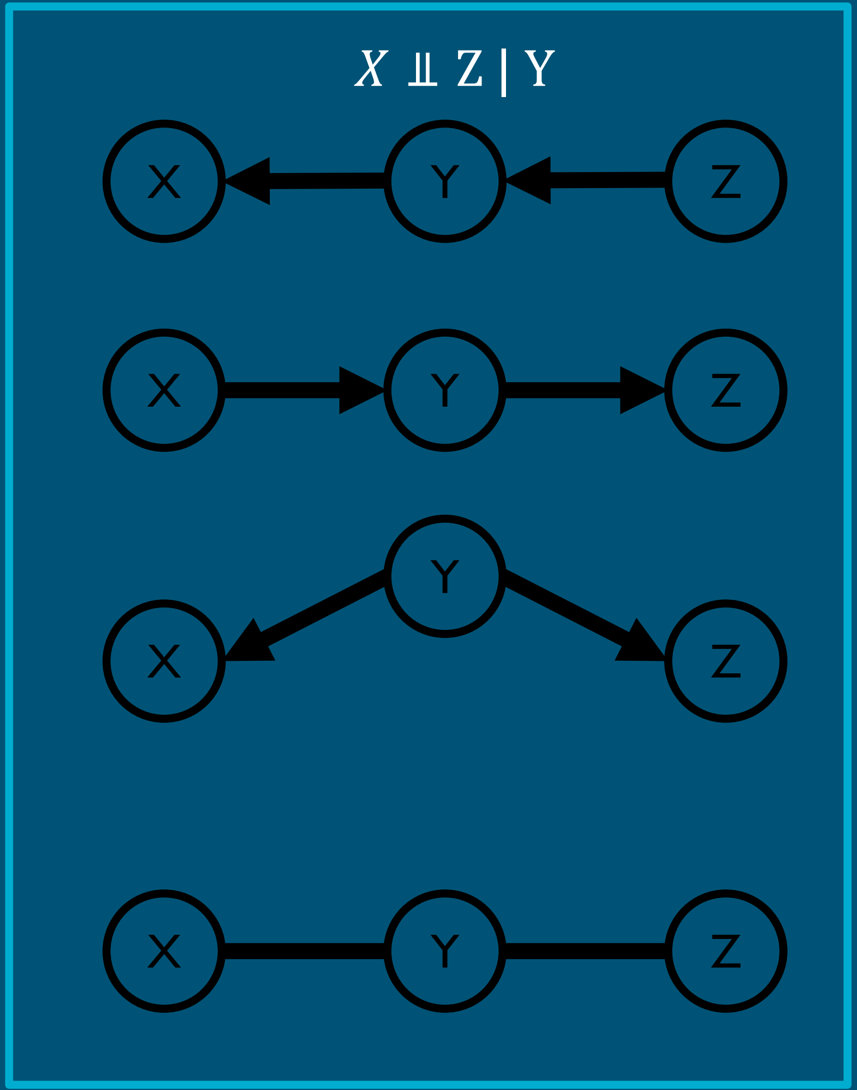
$X \not\perp\!\!\!\perp Z$ X is independent of Z, conditional on nothing

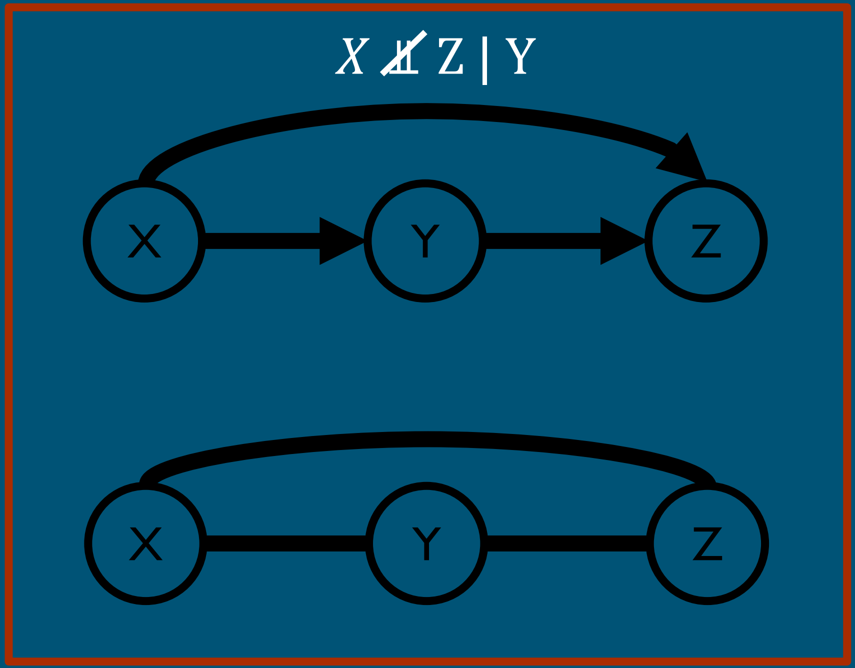
$X \perp\!\!\!\perp Z \mid Y$

Independence-Based Causal Discovery

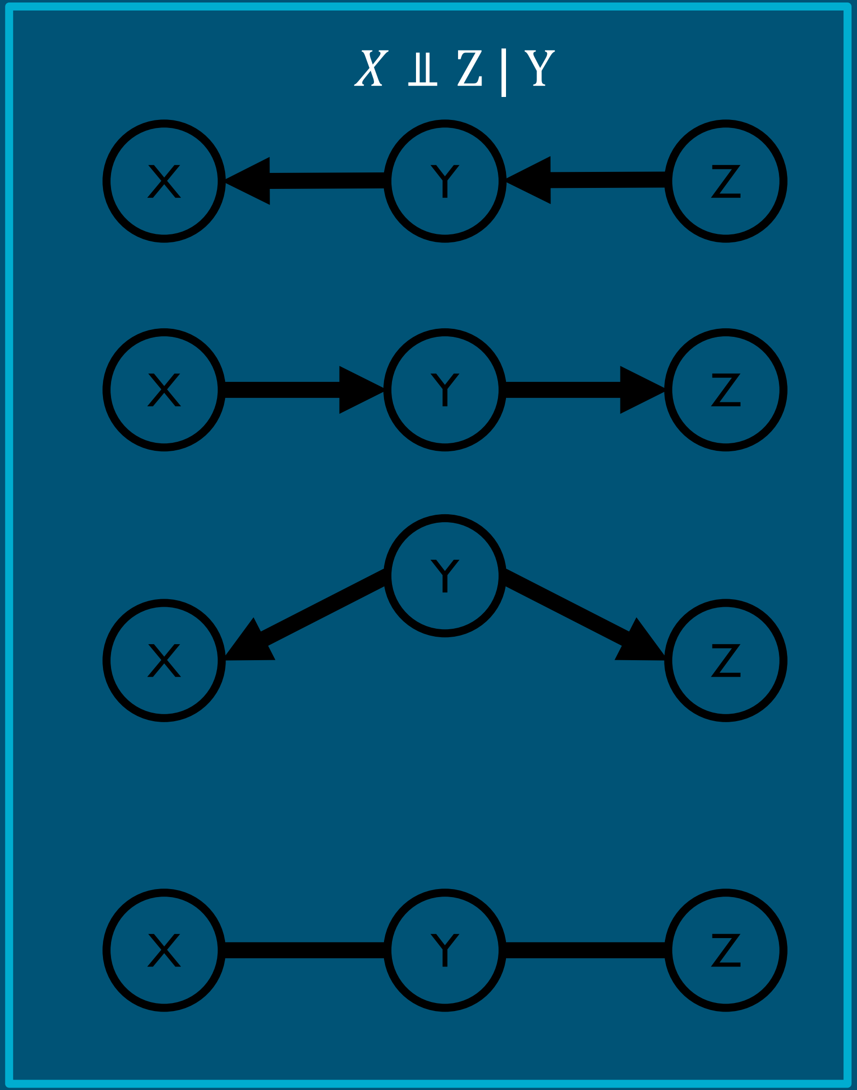


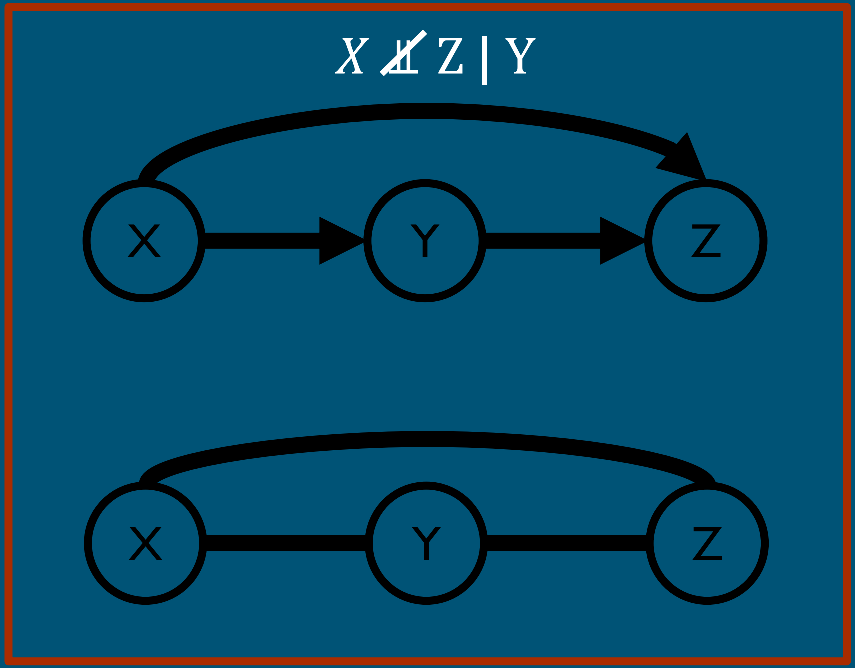
Independence-Based Causal Discovery





Independence-Based Causal Discovery



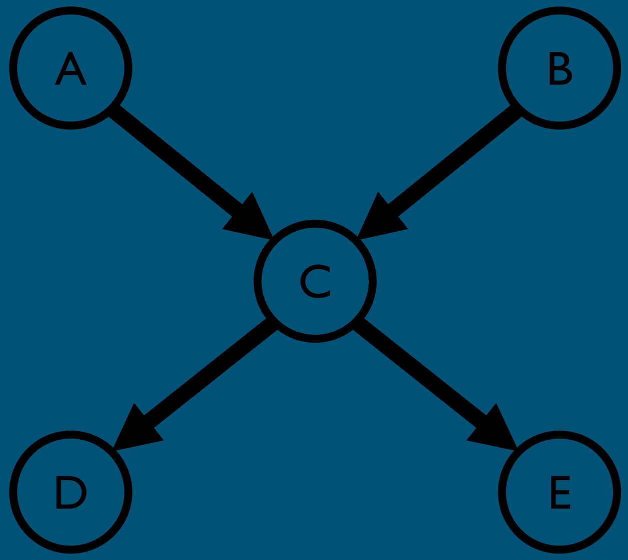


Markov equivalence can be found via colliders and skeletons

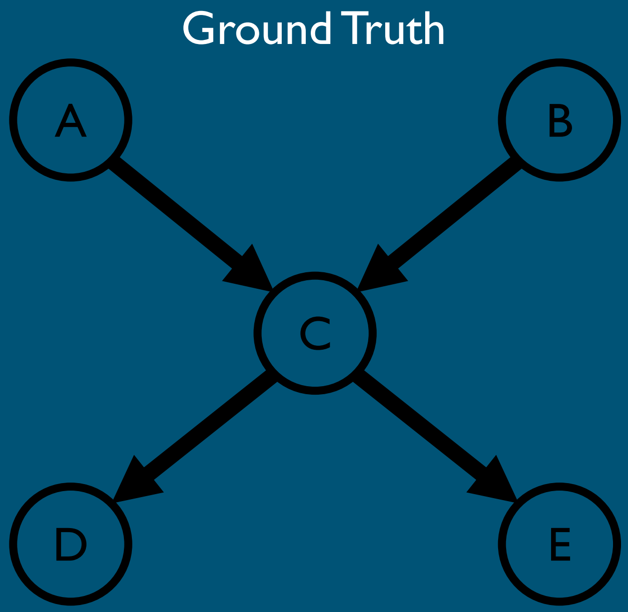
<u>Theorem</u>: two graphs are Markov equivalent if and only if they have the same skeleton and the same colliders (Verma and Pearl, 1990; Frydenburg, 1990)

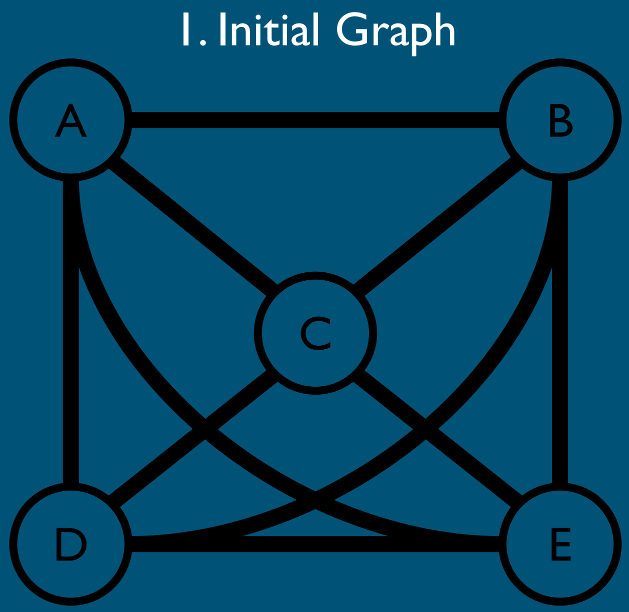
PC Algorithm Overview

Ground Truth



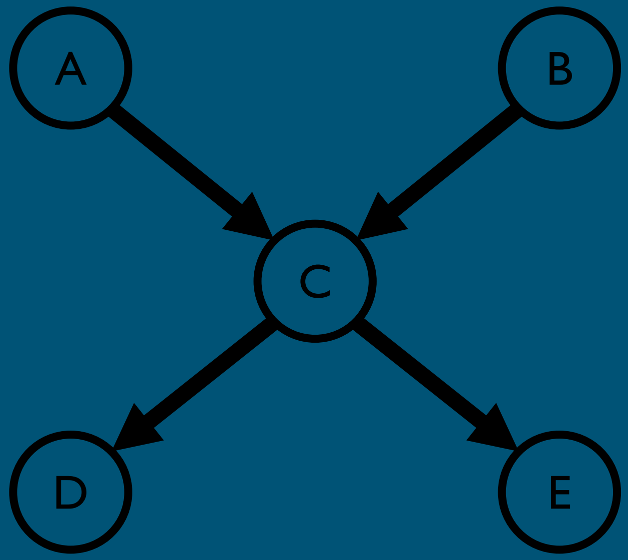
PC Algorithm Overview



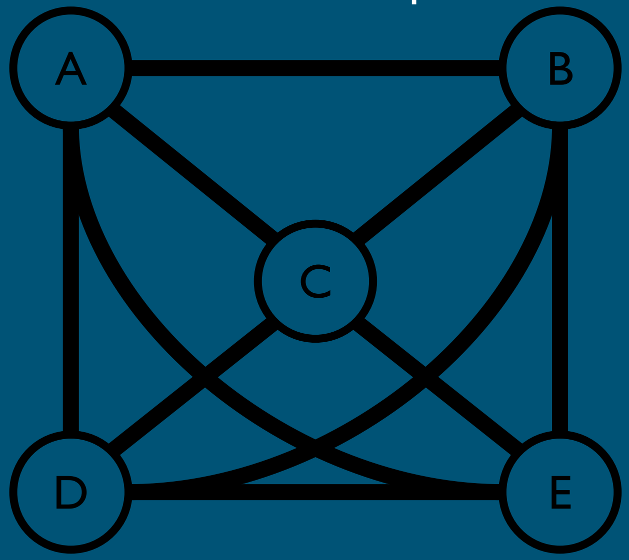


PC Algorithm Overview

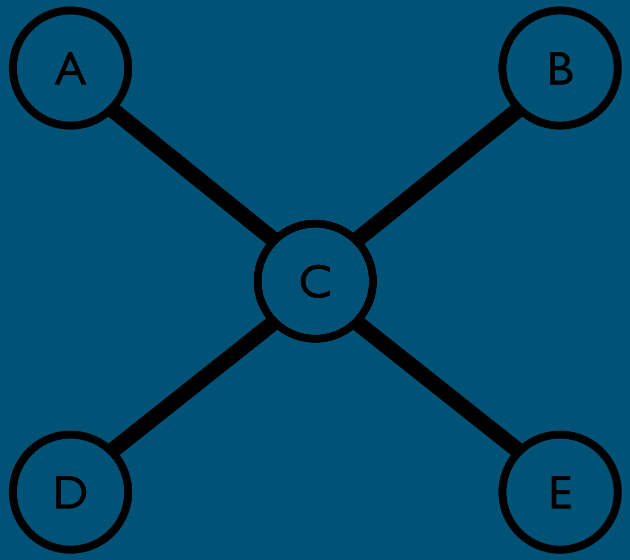
Ground Truth



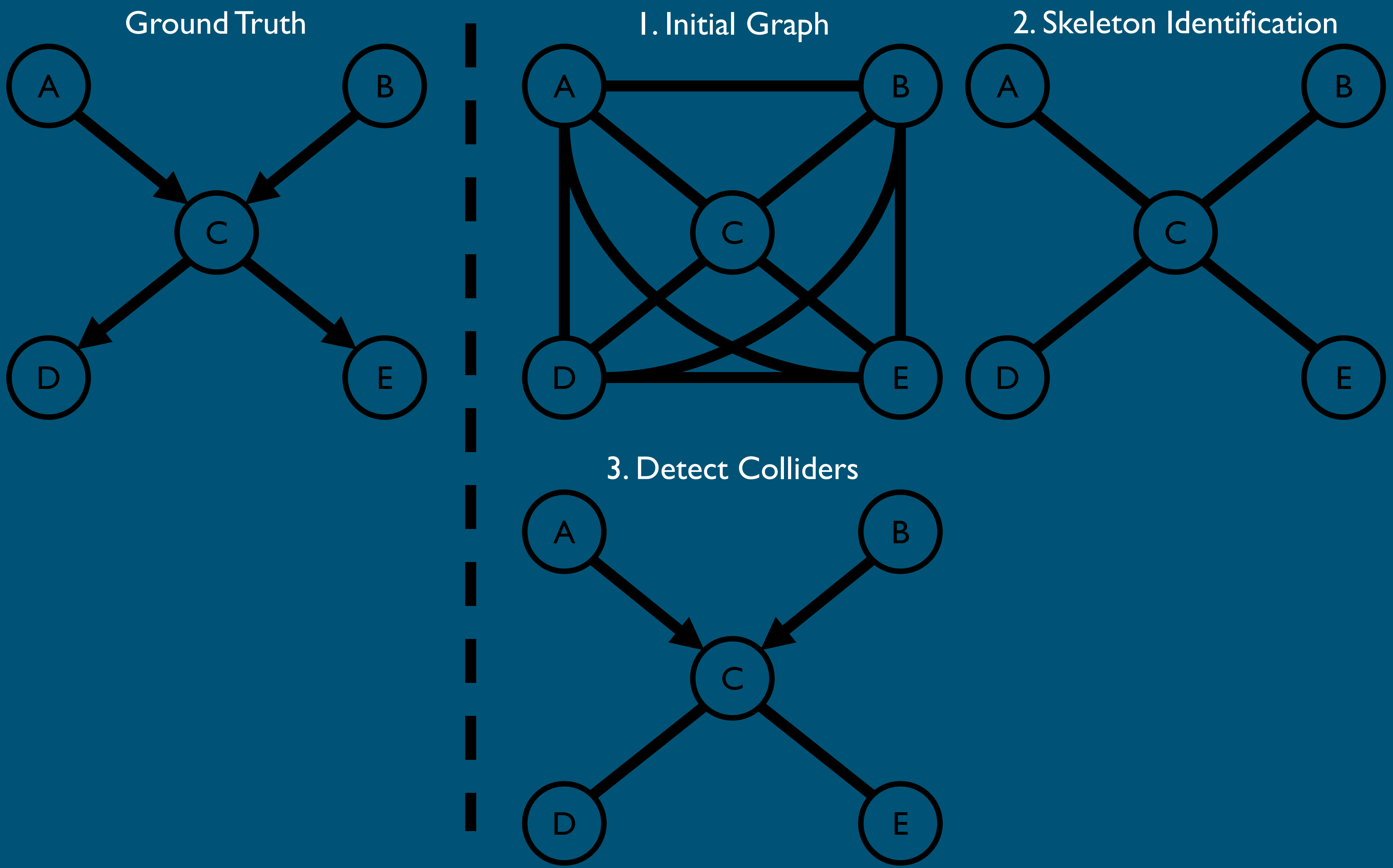
1. Initial Graph



2. Skeleton Identification

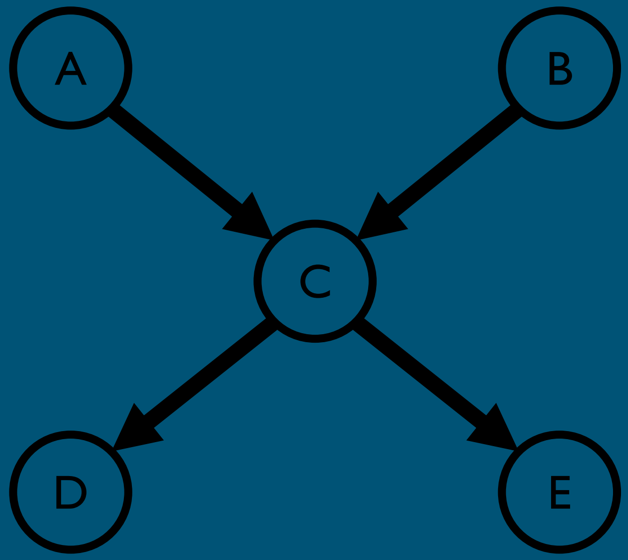


PC Algorithm Overview

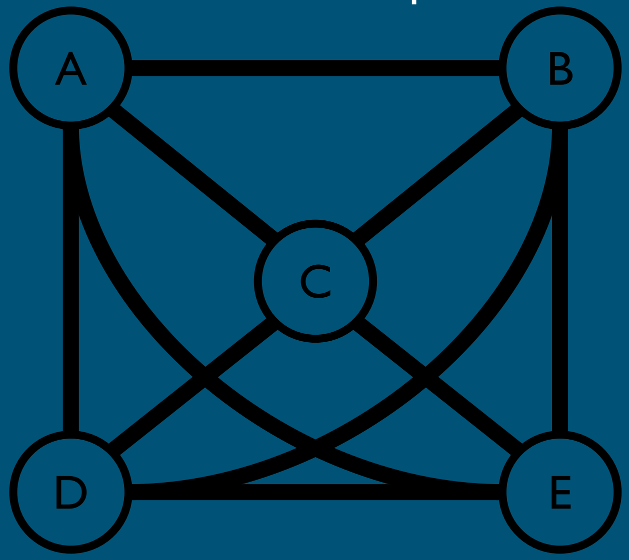


PC Algorithm Overview

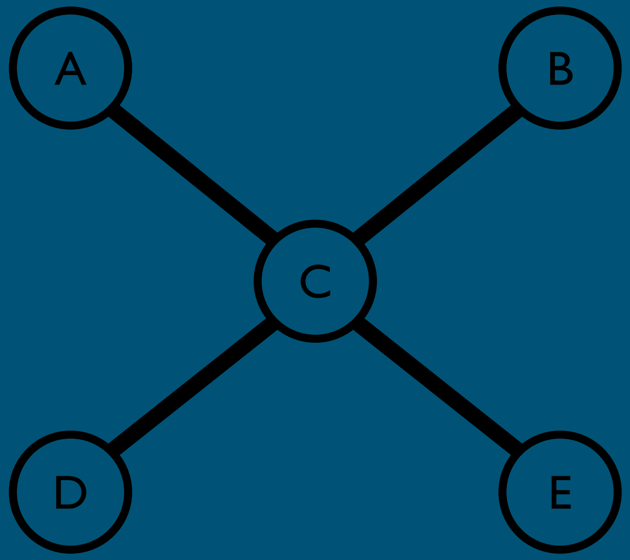
Ground Truth



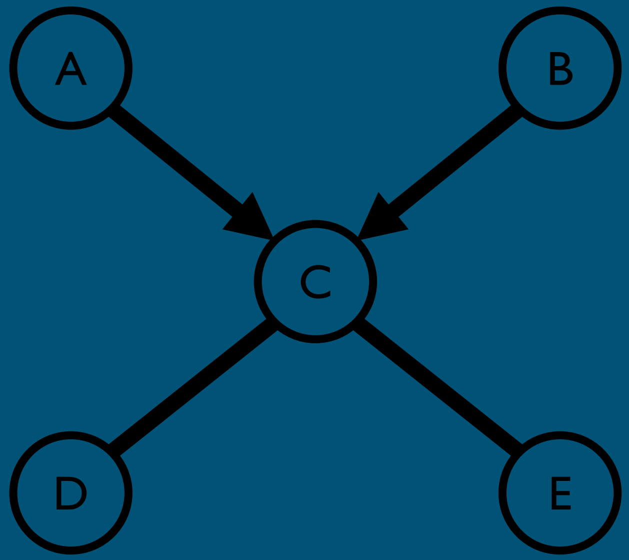
1. Initial Graph



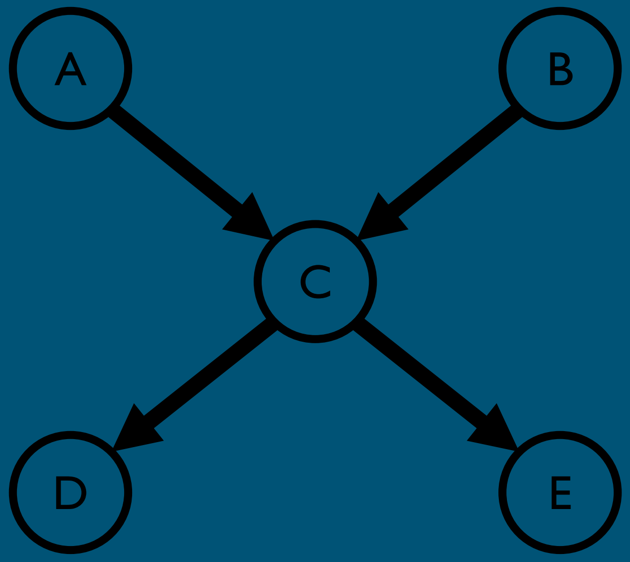
2. Skeleton Identification



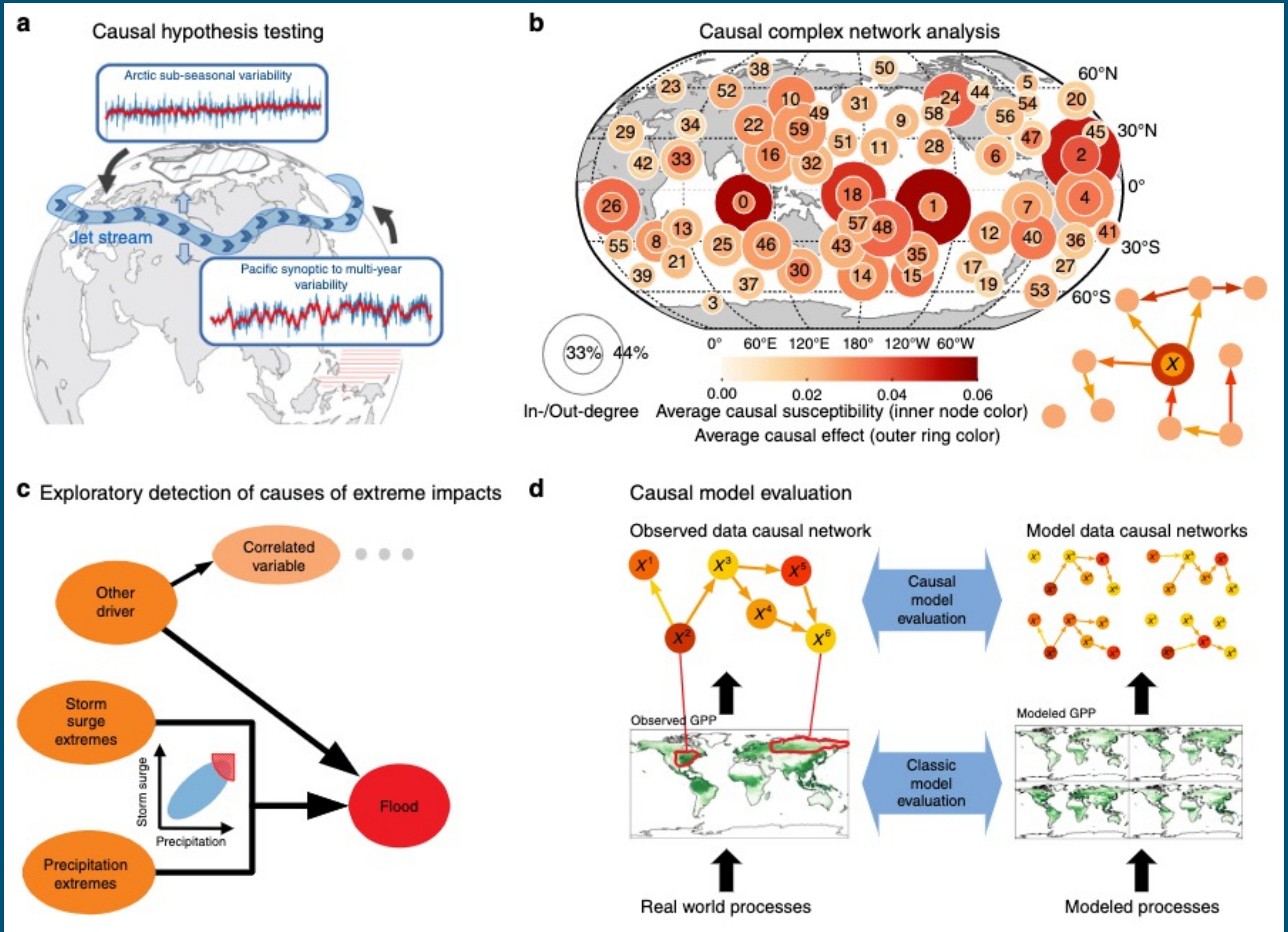
3. Detect Colliders



4. Orientation

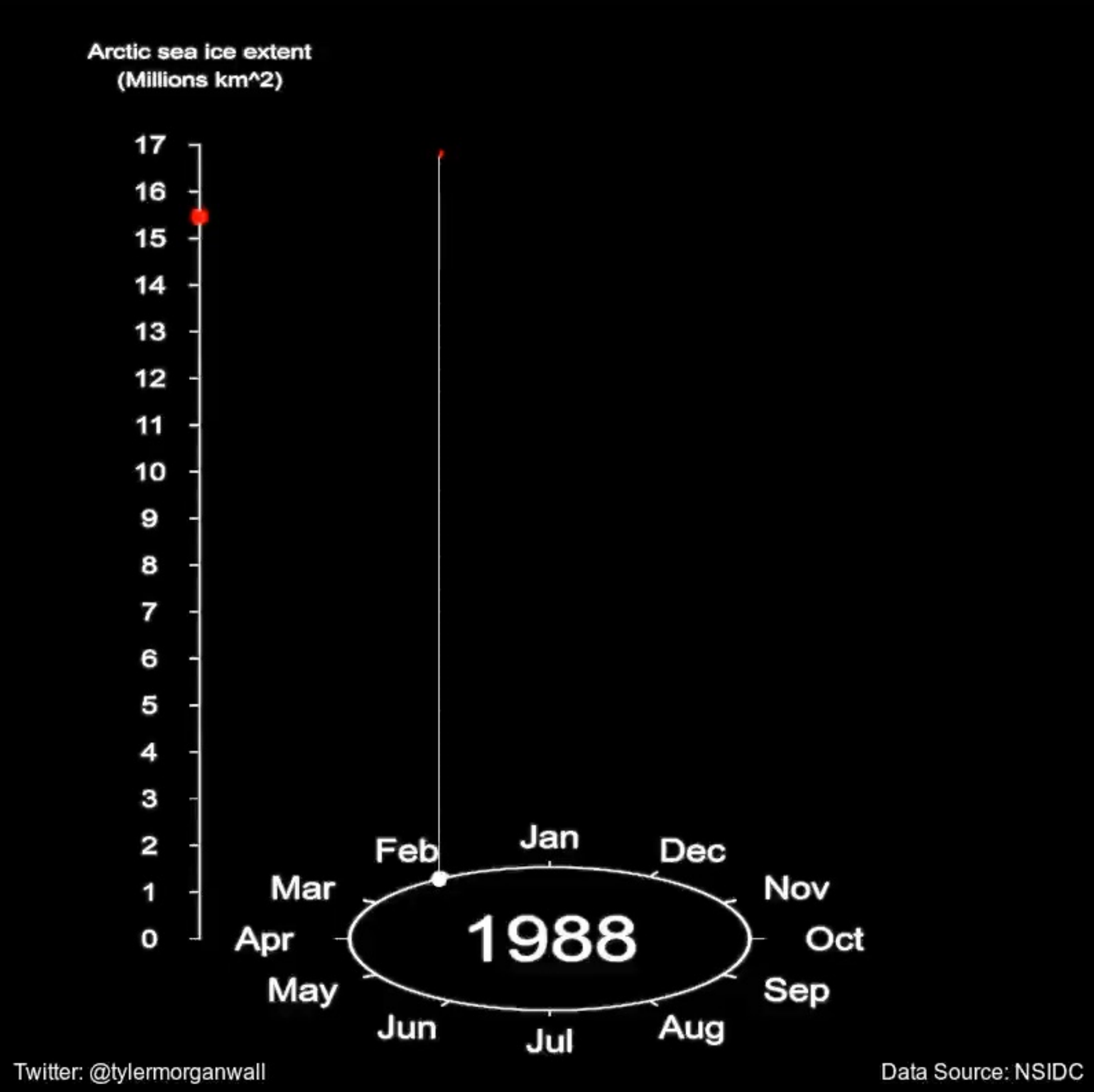


Causal Discovery



Jakob Runge, et al. 2019. Inferring causation from time series in Earth system sciences. *Nat Commun* 10, 1 (2019). DOI:https://doi.org/10.1038/s41467-019-10105-3

Daily Global Sea Ice Total Area with Monthly Polar Sea Ice Extent, 1988-2020



Permafrost Thaw

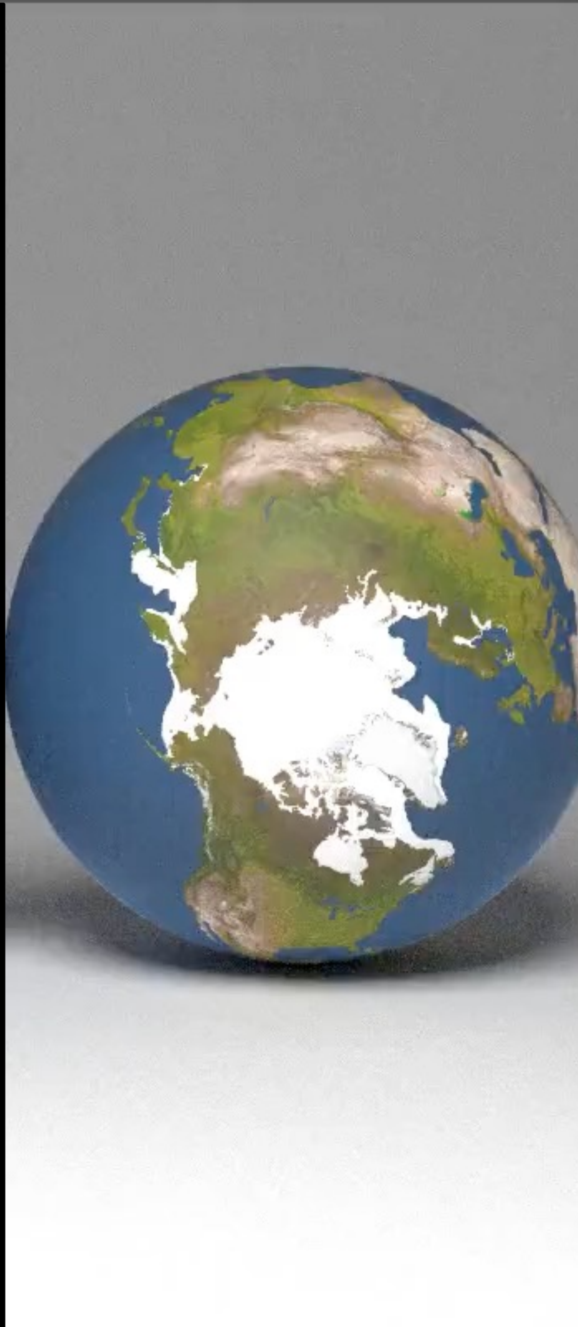
Impacts:
Significant greenhouse gas release; changes in hydrology; increased erosion

Shutdown of Atlantic Thermohaline Circulation

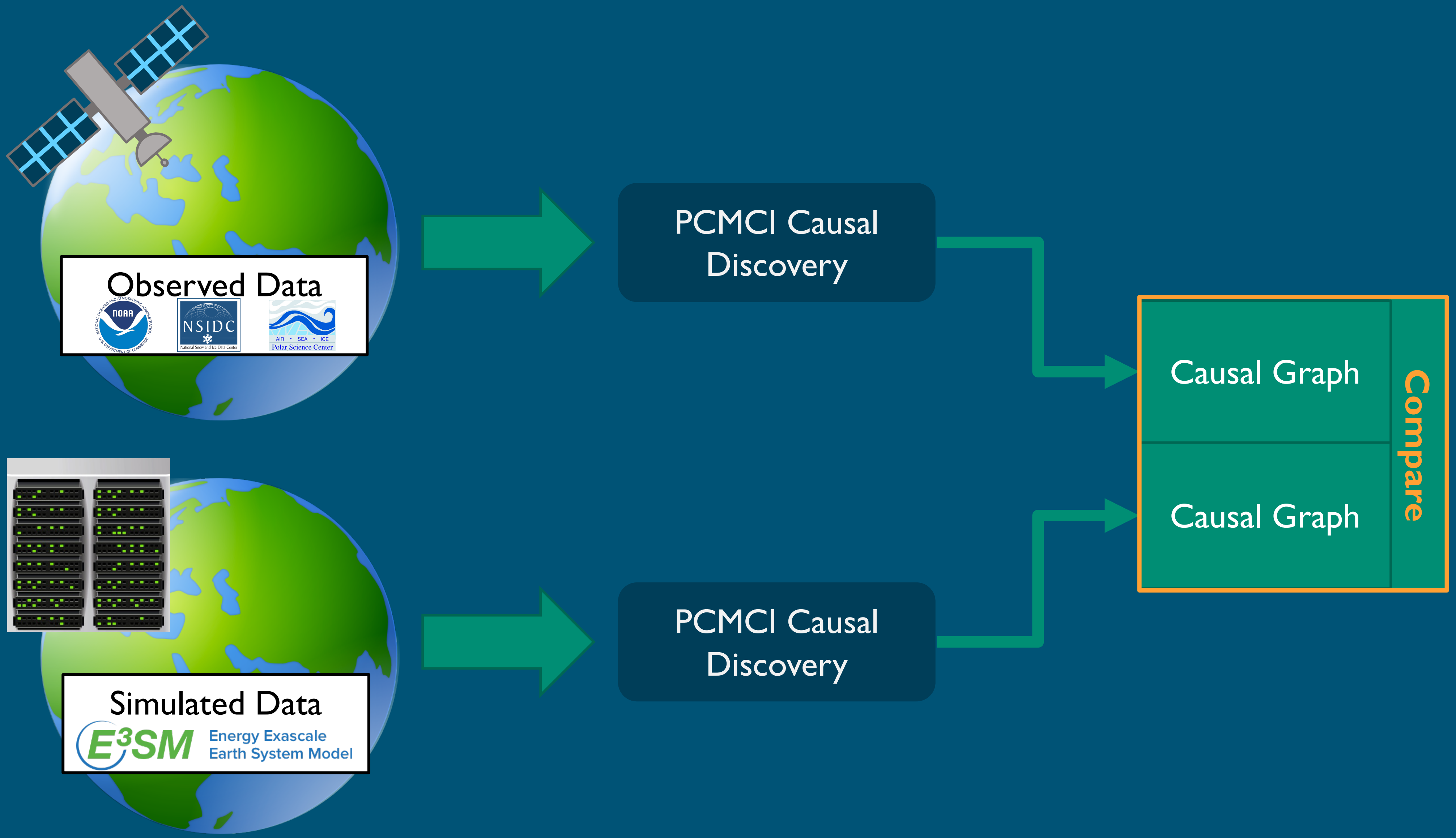
Impacts:
Regional cooling; significant weather shifts in the N. hemisphere

Loss of Summer Sea Ice

Impacts:
Mid-latitude weather changes; ocean current alterations



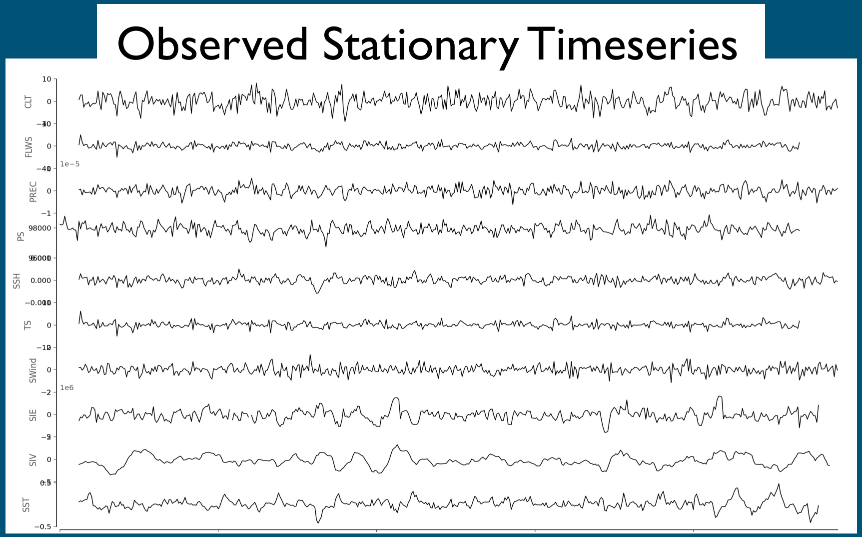
Specific Methodology

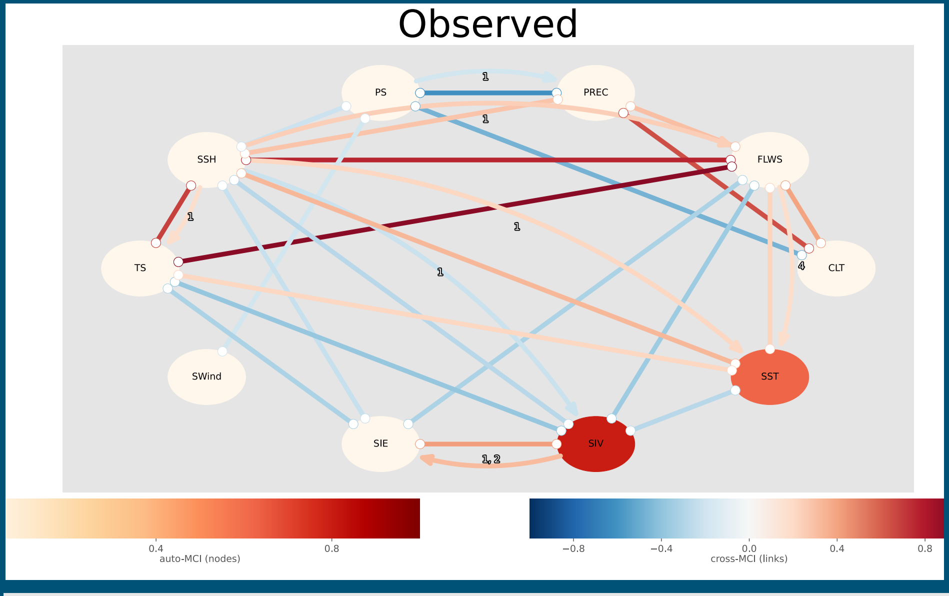


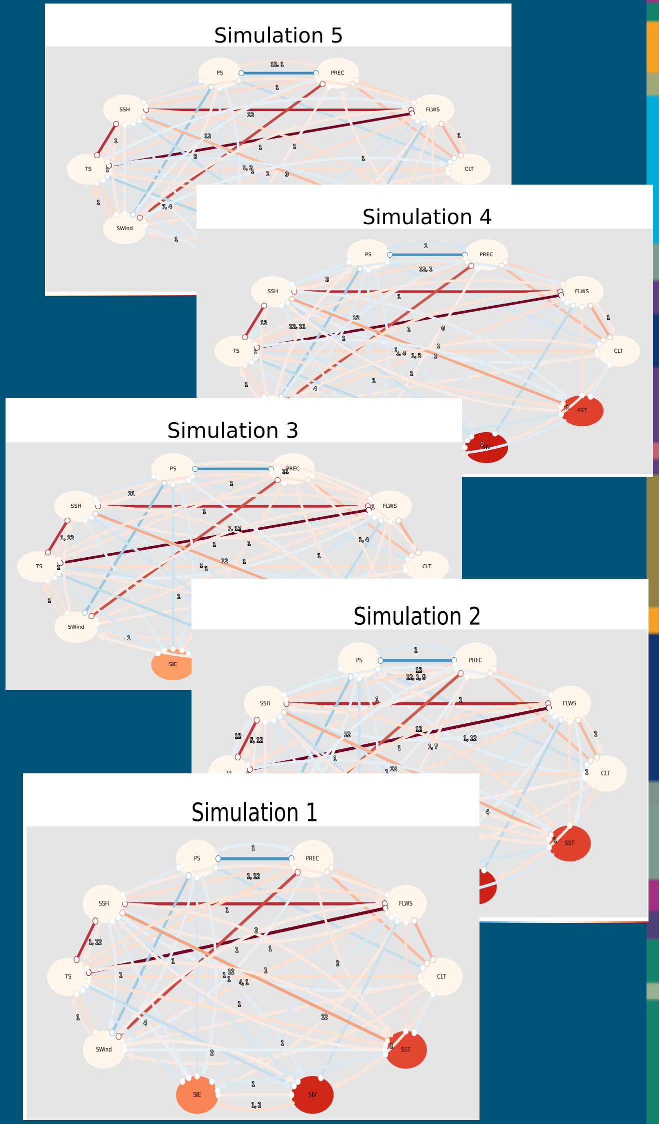
Results

Steps

- Preprocessing
  - Create a time series of each variable
  - Timeseries stationarity is needed because the algorithm must assume that deviations from the mean/variance are due to internal influences rather than some external seasonality or long-term trend
  - Transform time series to make them all stationary
- Parameterization Tuning
  - Choose a maximum lag to include
  - Choose the alpha significance threshold for independence tests
- Causal Network Learning
  - Fit the PCMCI [1] causal discovery algorithm to each dataset
  - Analyze resultant networks







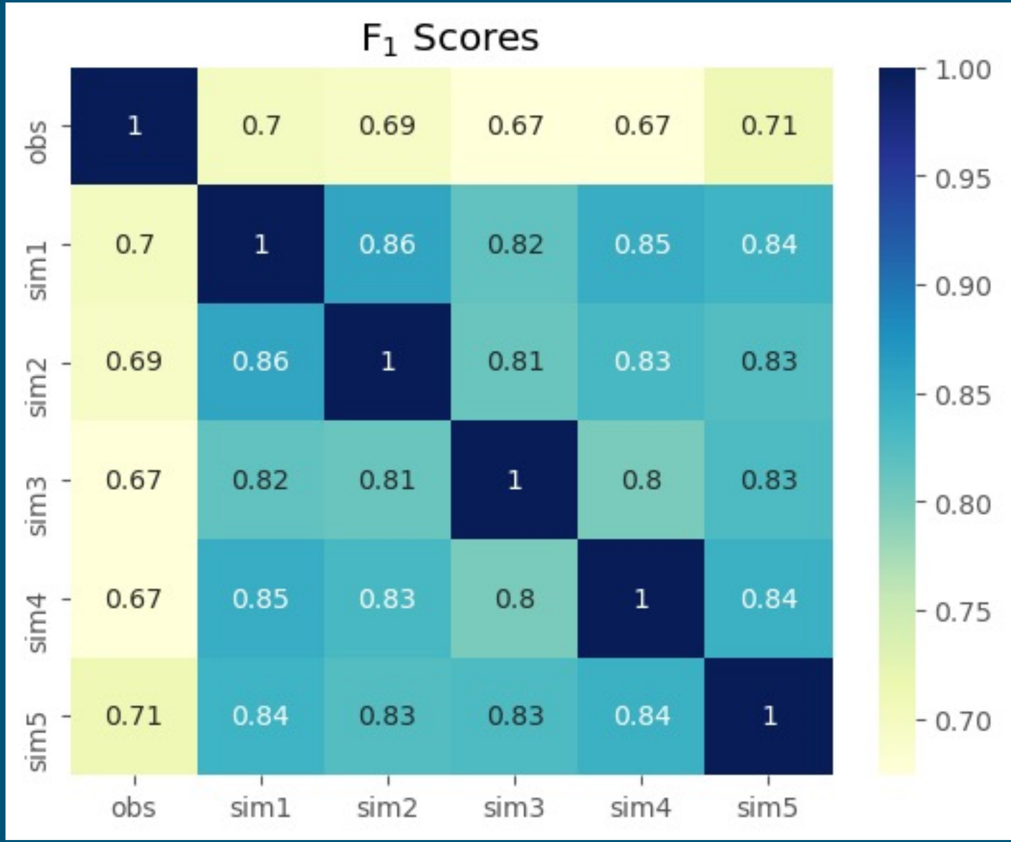
[1]Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. 2019. *Detecting and quantifying causal associations in large nonlinear time series datasets*. Retrieved from http://advances.sciencemag.org/

Results

The F_1 Score is a similarity metric computed from existence of edges in a pair of networks

Future work includes more metrics:

- Some node-node similarity metrics
  - Node-node F_1 Score
  - Others
- Node level metrics will identify where the differences occur and more meaningful inferences may be possible
- An average goodness of fit score for each network
  - Each edge has a goodness of fit and a significance value to determine if it should exist in the network
  - Combining these could be a good metric for overall fit
- Apply FCI and LPCMCI
  - Tolerance for latent or unobserved variables
  - Can sometimes discover latent variables



Questions?

Backups

Causal Inference

Causal Inference

Judea Pearl's three levels of causation

1. Seeing – associate quantities

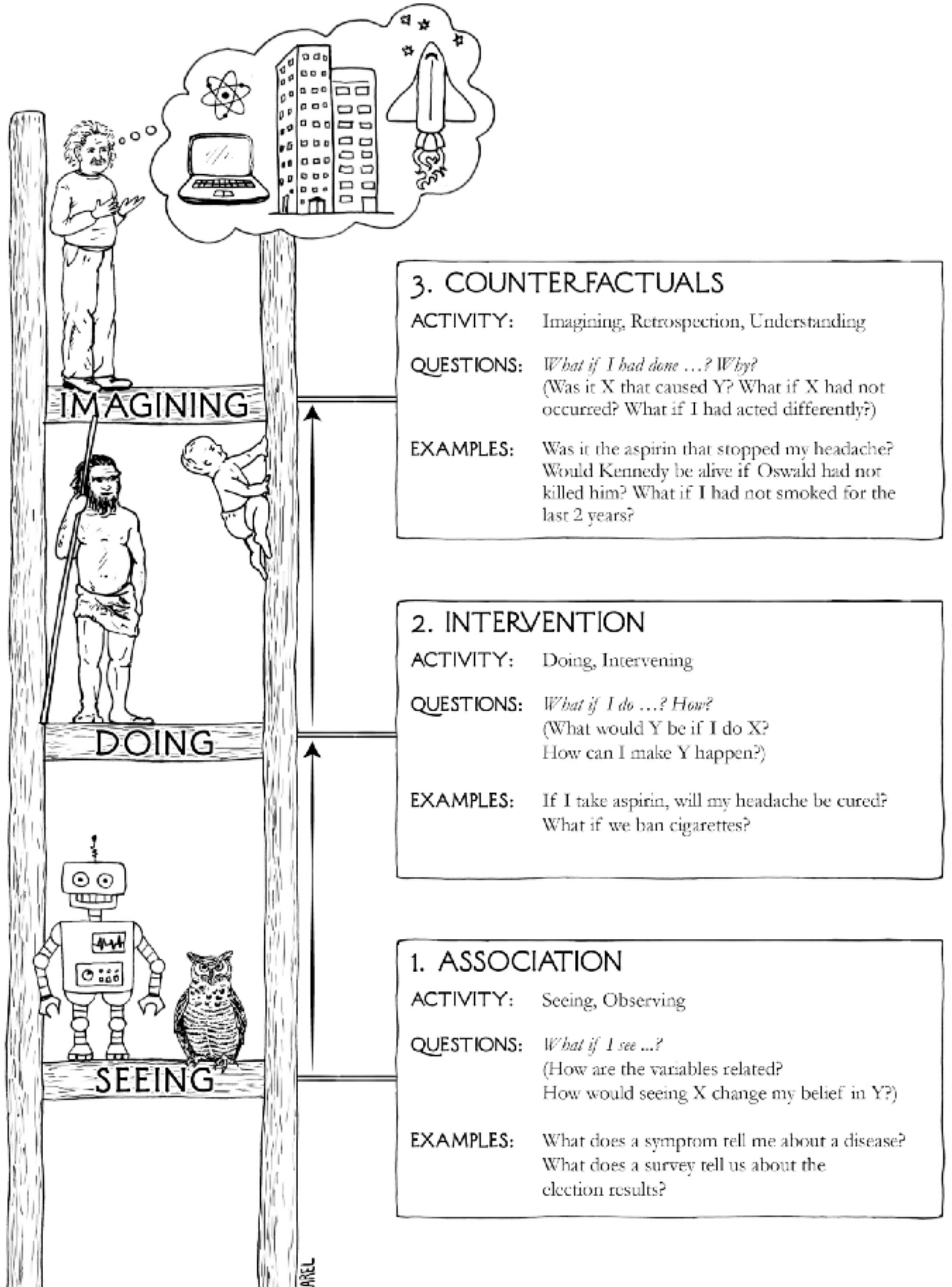
- What most animals and machines do
- What if X happens?
- Prediction

2. Doing – changing quantities

- Deliberate intervention/experimentation in a process
- What if I do X?

3. Imagining – retrospective analysis and understanding

- Counterfactual analysis
- What if I had done Y? Why did Z occur?



Judea Pearl's Ladder of Causation. Illustrator: Maayan Hare

What does imply causation?

Causal Inference

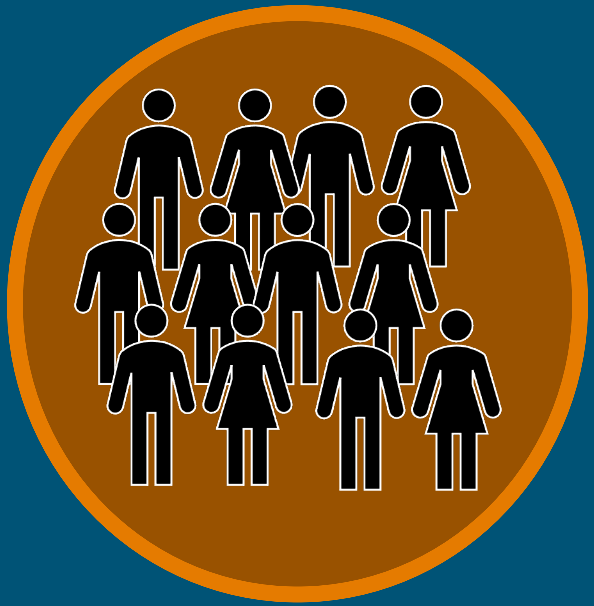
Randomized Control Trials

1. Take one sample population

Causal Inference

Randomized Control Trials

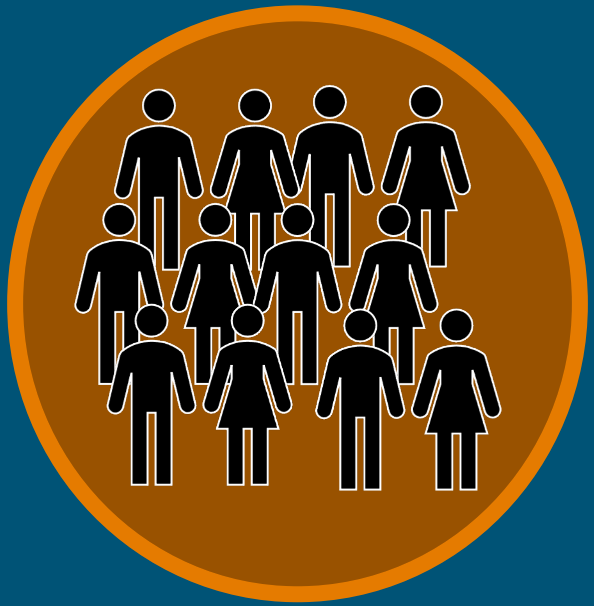
1. Take one sample population
2. Randomly divide them up into a treatment group and a control group



Causal Inference

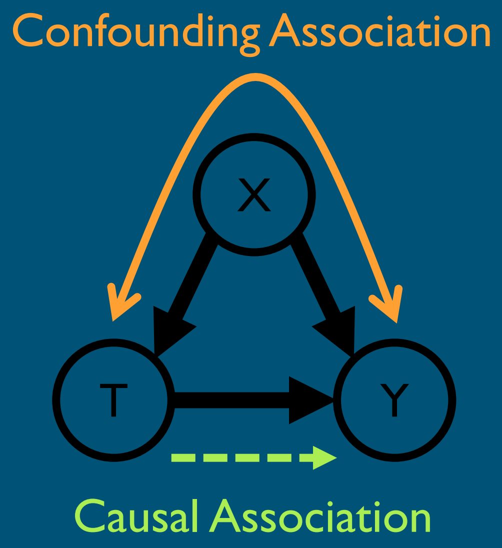
Randomized Control Trials

1. Take one sample population
2. Randomly divide them up into a treatment group and a control group



- Treatment is applied at random
- Confounding variables of individuals will not appear in the average treatment effect (ATE)

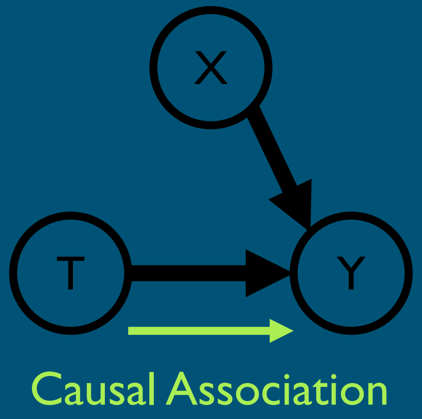
Causal Inference



Treatment T
Common Cause X
Potential Outcome Y

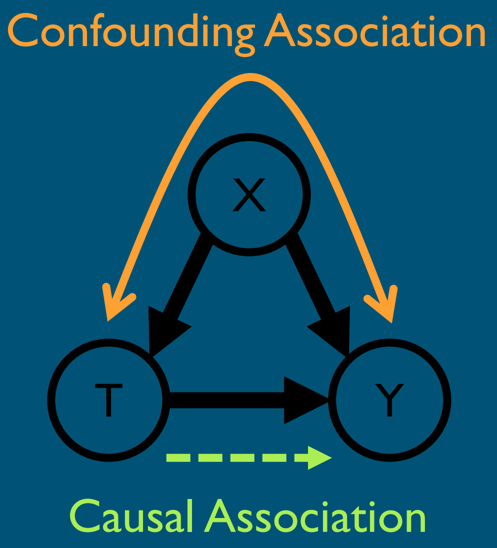
RCTs: experimenter randomizes subjects into control and treatment groups.

- Treatment group cannot have causal parents
- The groups are then comparable

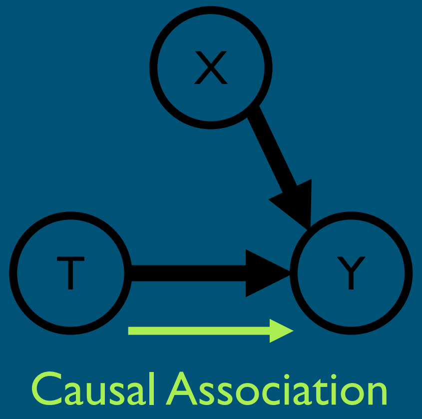


Causal Inference

Observational Studies

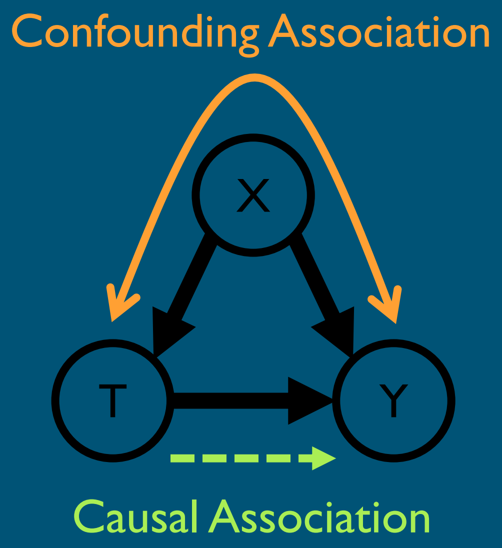


Ideal



Causal Inference - Observational Studies

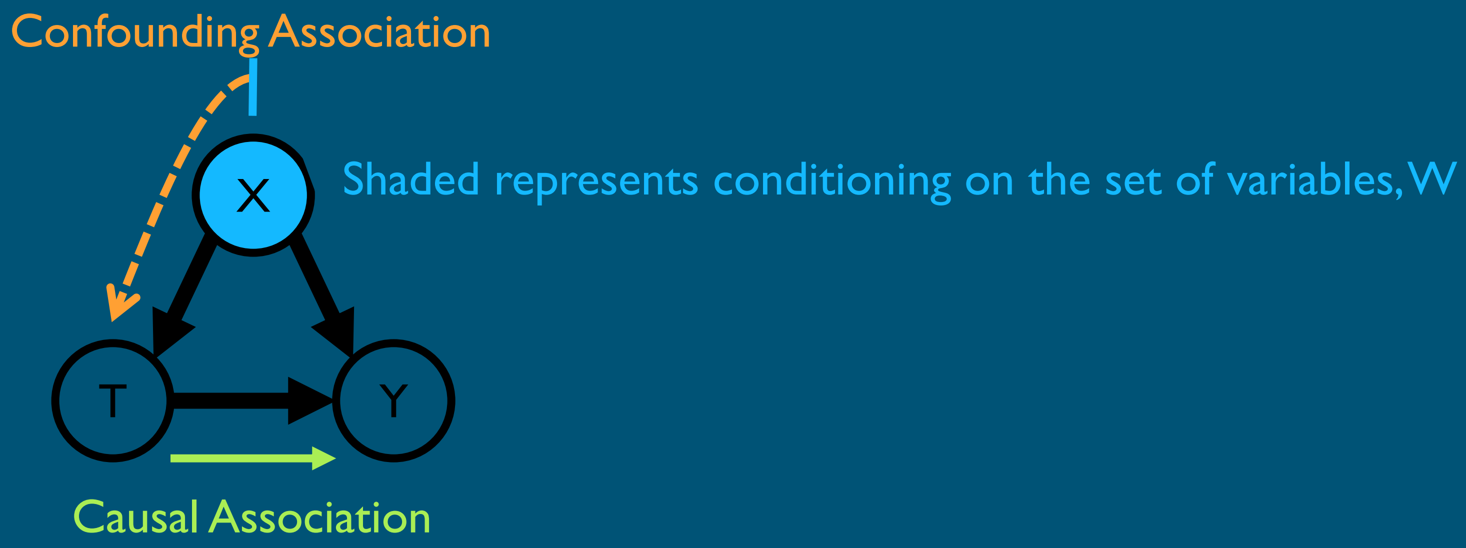
The solution is to adjust or control for confounders



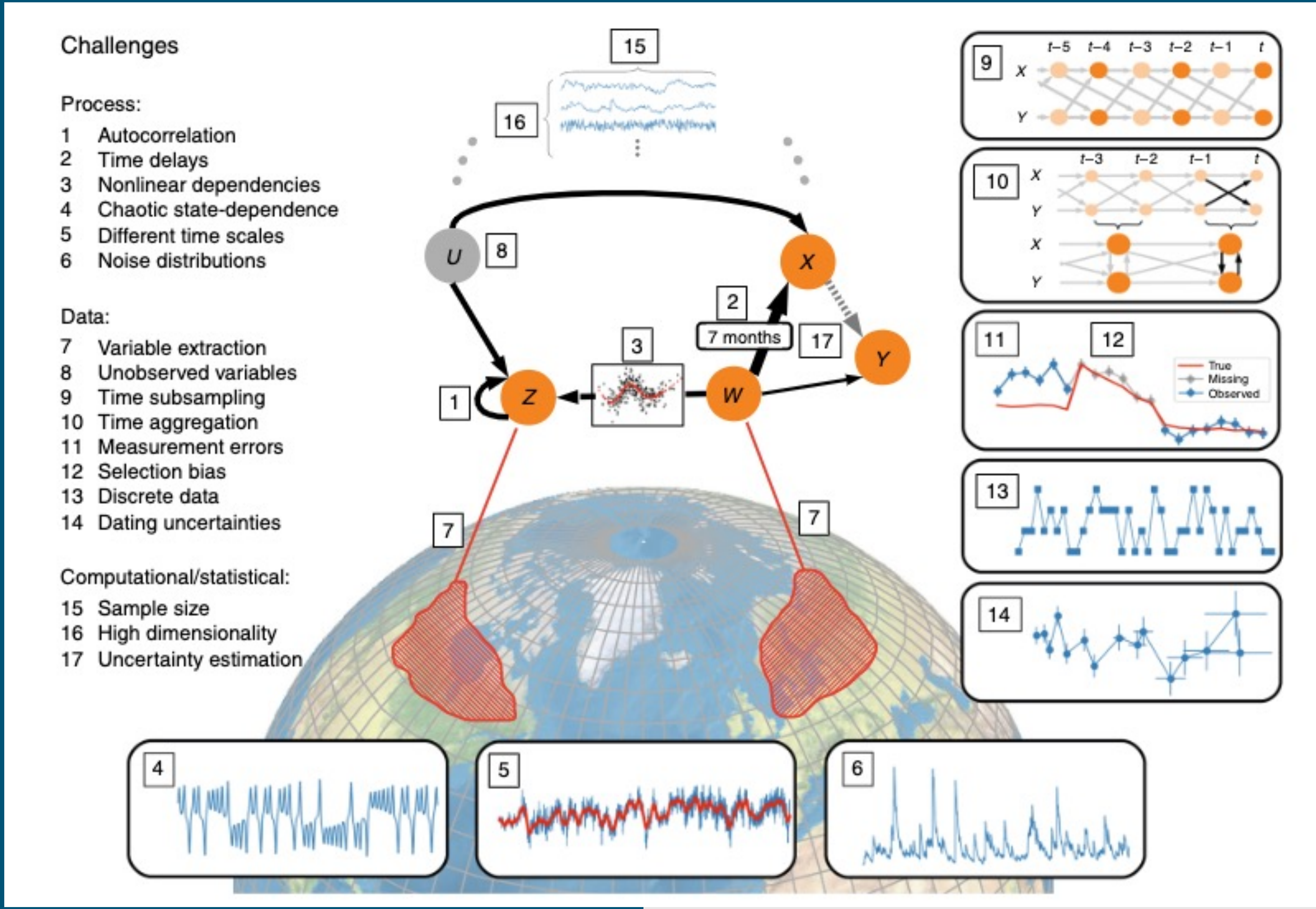
Causal Inference - Observational Studies

The solution is to adjust or control for confounders

If a set of variables, *W*, is a sufficient adjustment set, then we can block the confounding association and expose the causal association.



Causal Discovery

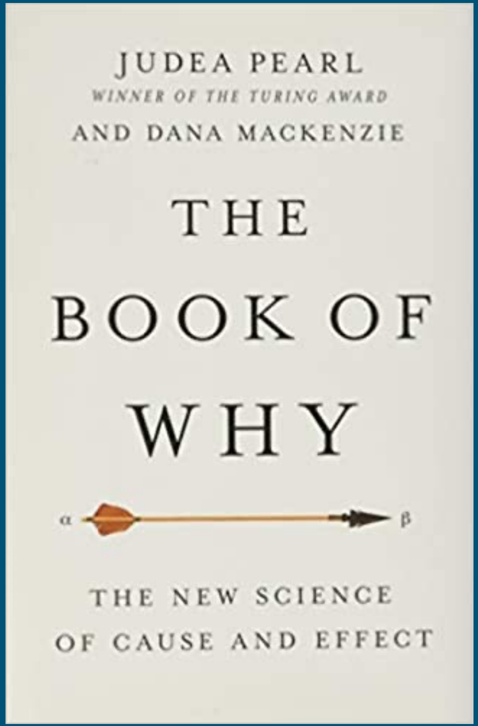


Jakob Runge, et al. 2019. Inferring causation from time series in Earth system sciences. *Nat Commun* 10, 1 (2019). DOI:https://doi.org/10.1038/s41467-019-10105-3

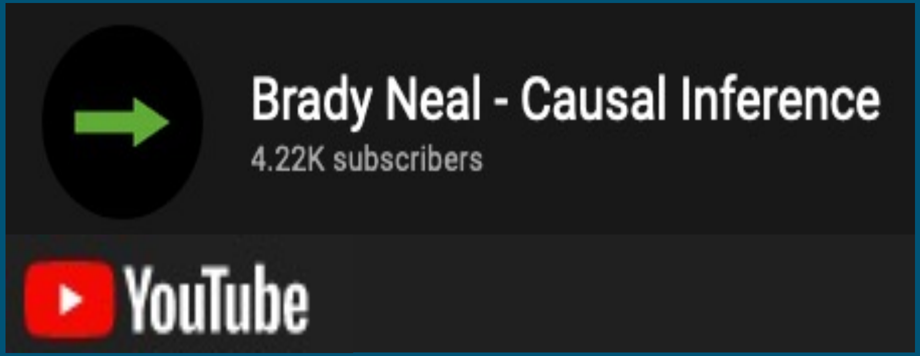
Helpful Resources

Jakob Runge, et al. 2019. Inferring causation from time series in Earth system sciences. *Nat Commun* 10, 1 (2019). DOI:https://doi.org/10.1038/s41467-019-10105-3

The Book of Why by Judea Pearl, Dana Mackenzie



Brady Neal – Causal Inference



CauseMe.net – Runge et al.

- "The CauseMe platform provides ground truth benchmark datasets featuring different real data challenges to assess and compare the performance of causal discovery methods."