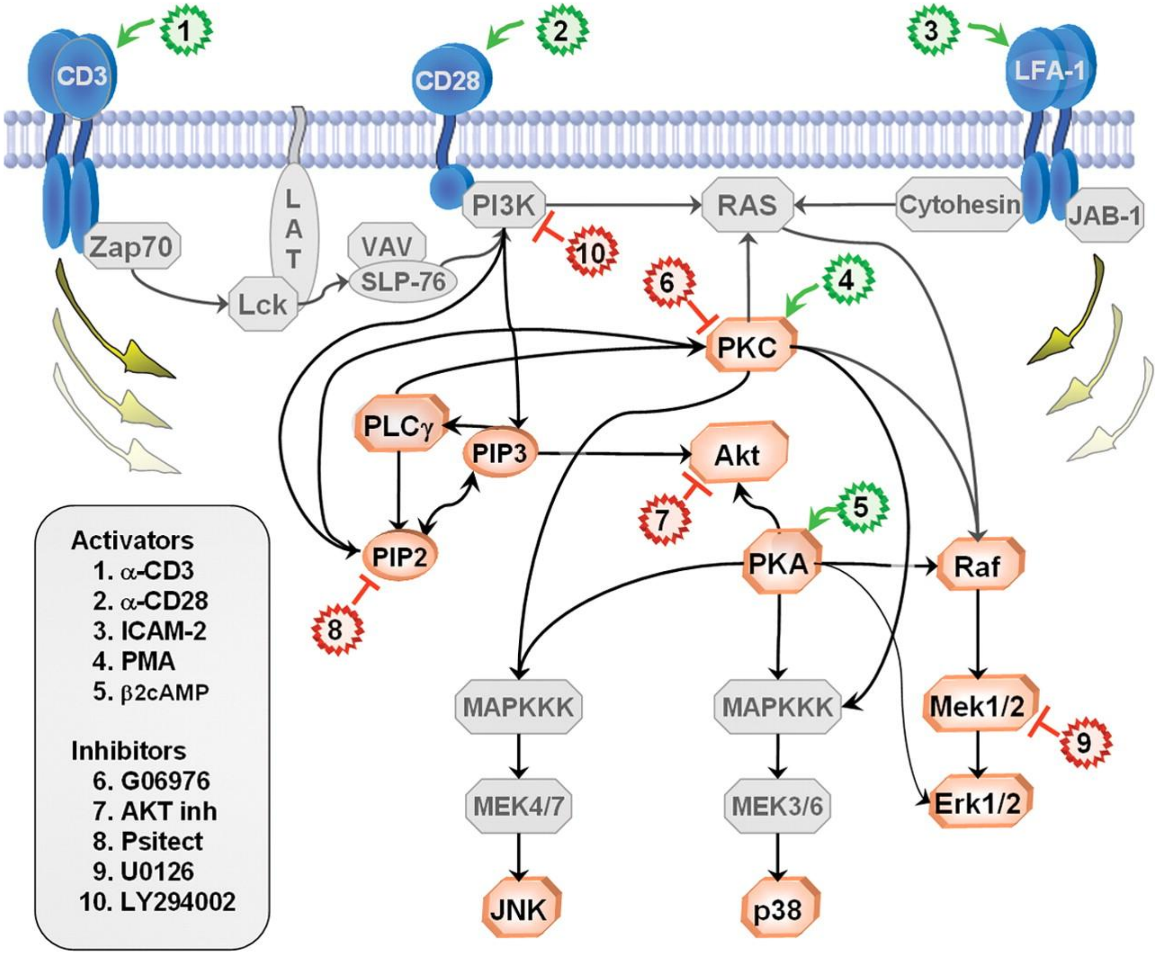
Using Experiments to Discover Cyclic Causal Structures with Latent Variables

Frederick Eberhardt
Philosophy-Neuroscience-Psychology Program,
Department of Philosophy;
Washington University in St. Louis
eberhardt@wustl.edu
www.artsci.wustl.edu/~feberhar

Causal & Statistical Inference

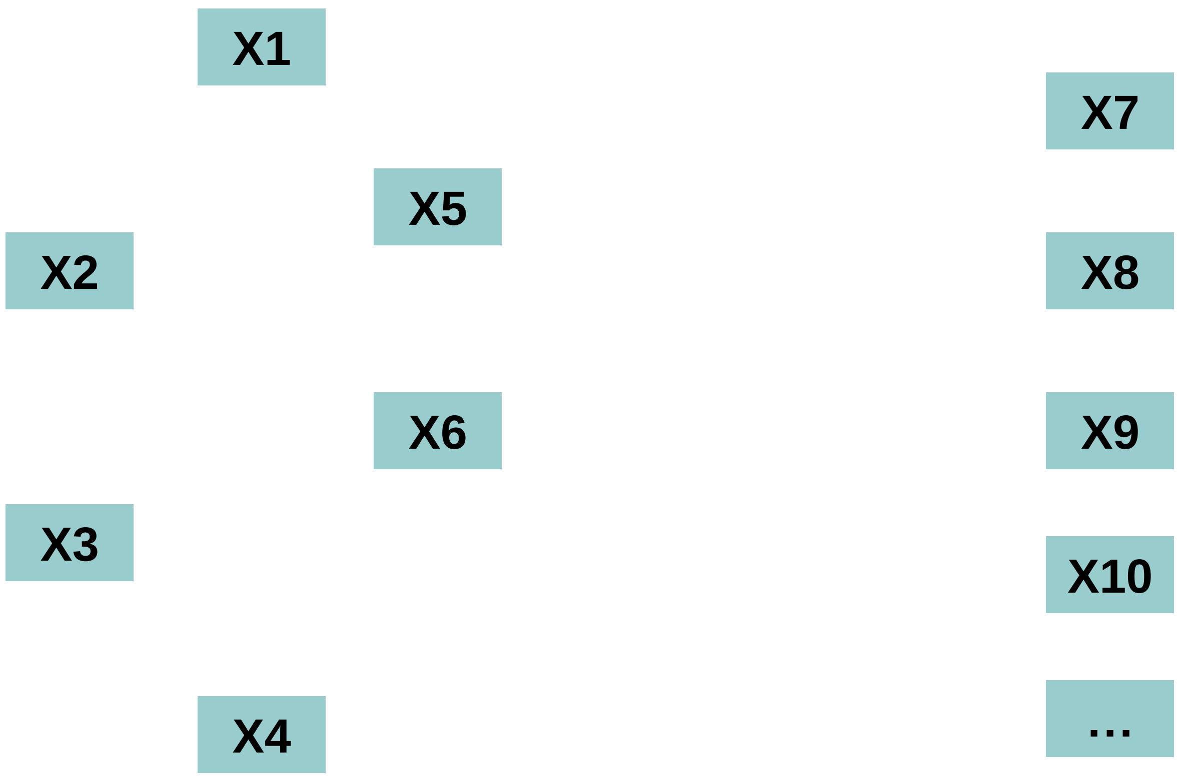
- Conditions for causal discovery
- Learning from experimental data
  - Algorithms
  - Different types of experimental interventions
- Optimal sequences of experiments
  - Normative: analyzing discovery strategies
  - Descriptive: causal learning in humans and animals
- Impact of fundamental and methodological assumptions

Causal Discovery

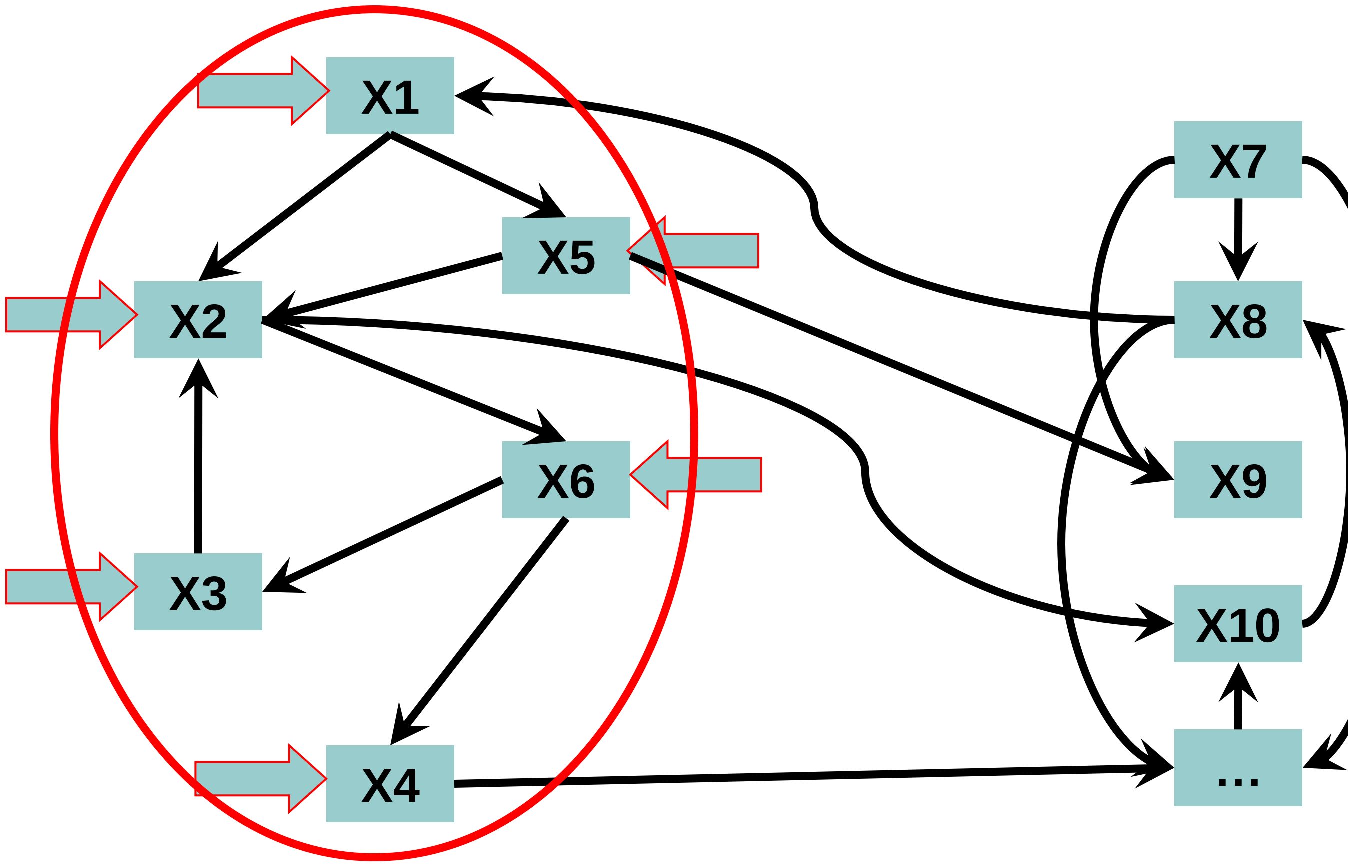


Sachs et al. (2005)

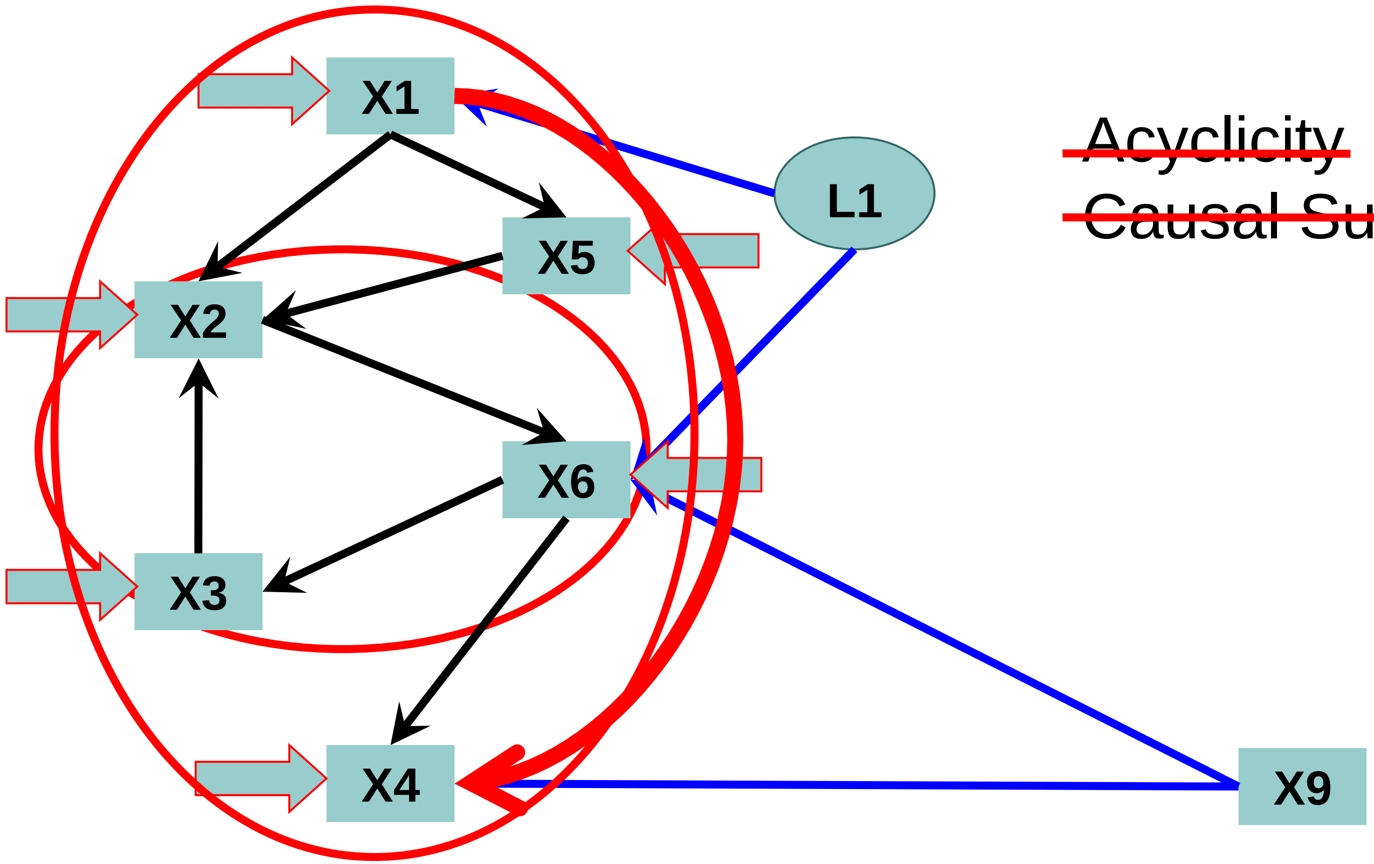
Signaling Network



Signaling Network



Model Assumptions

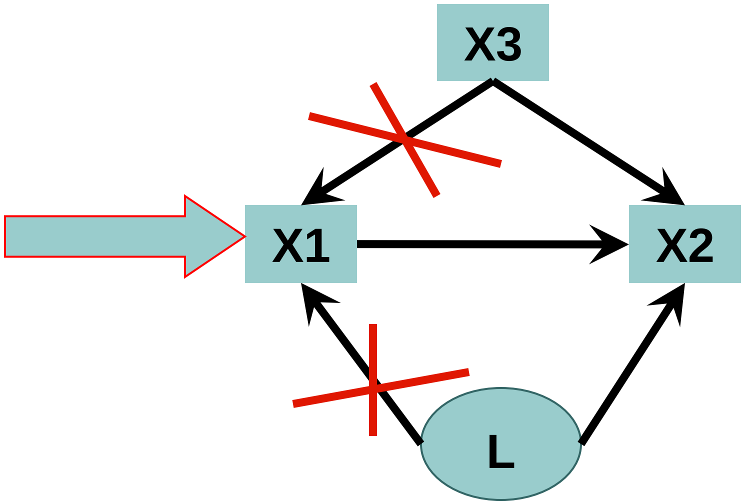


~~Acyclicity~~

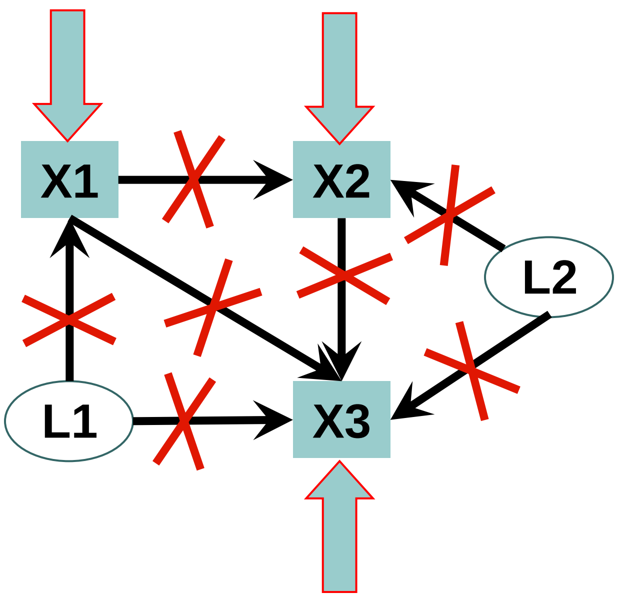
~~Causal Sufficiency~~

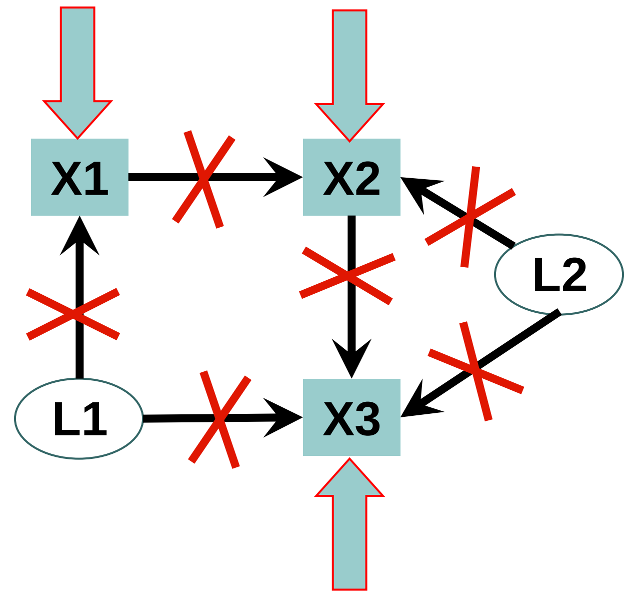
Experimental Interventions

- Determines causal order
- Breaks confounding
- Confounding due to latent variables



Latent Variables

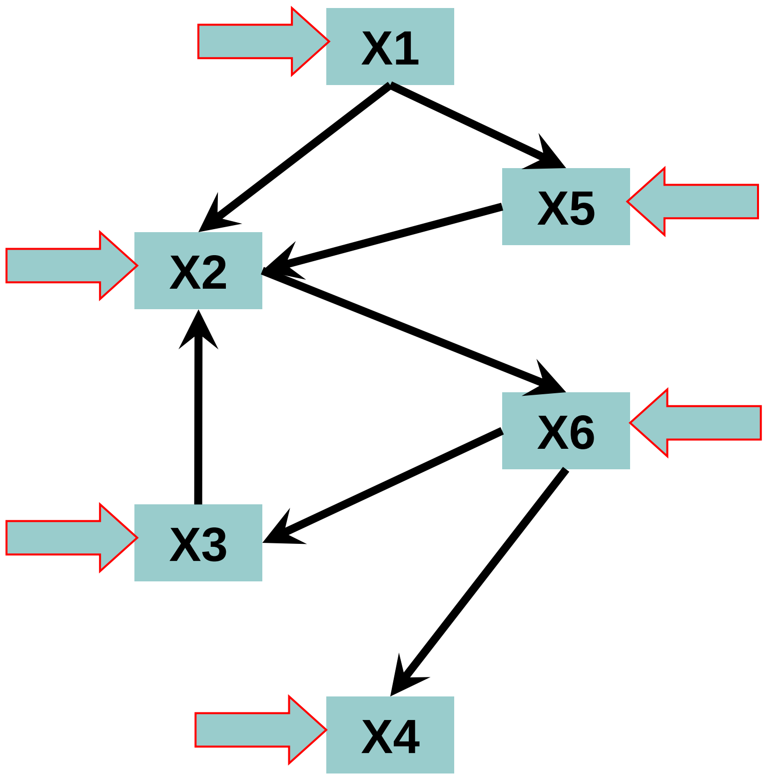




X1 _||_ X3 | {set(X1), set(X2)} vs. X1 _||_ X3 | {set(X1), set(X2)}

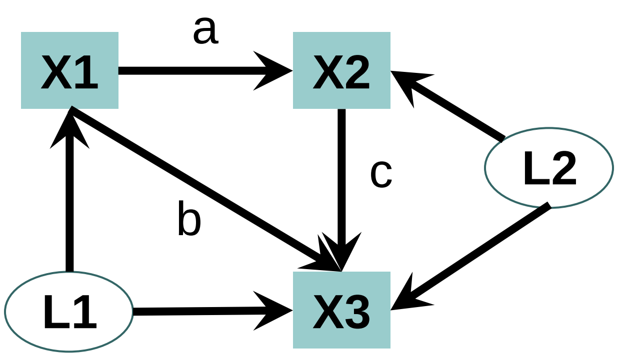
Using only independence and dependence constraints, these two graphs are indistinguishable by *any* sequence of experiments involving single (or no) interventions per experiment.

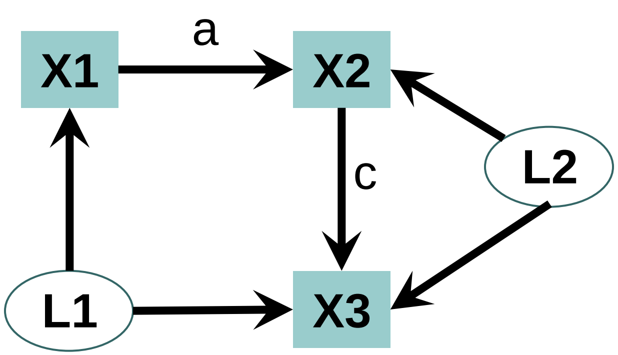
Given only Independence Constraints…



- In the worst case N experiments, each intervening on up to N-1 variables are necessary to discover the causal structure.

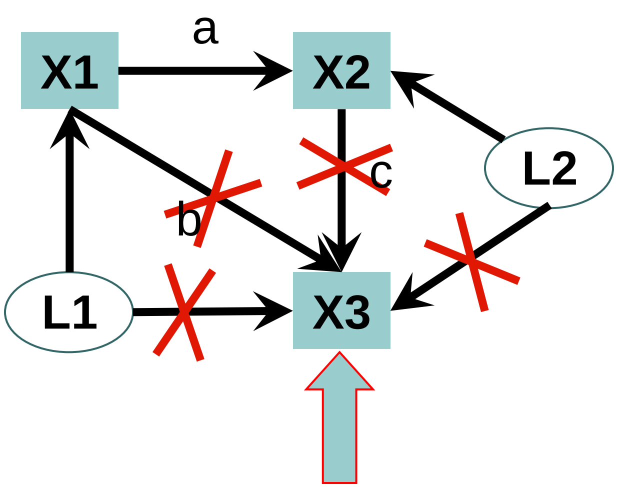
Linearity

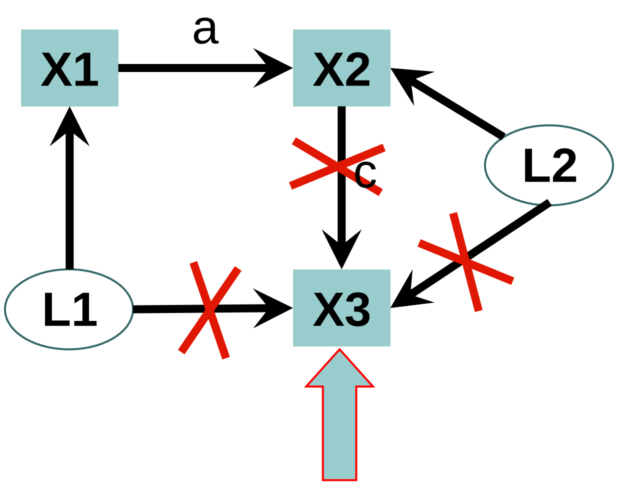




$$X_i = \sum_{X_j \in pa(X_i)} k_j X_j + e_i$$

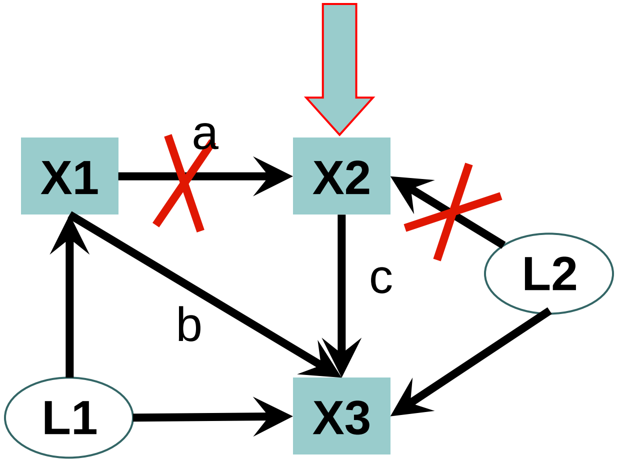
Linearity

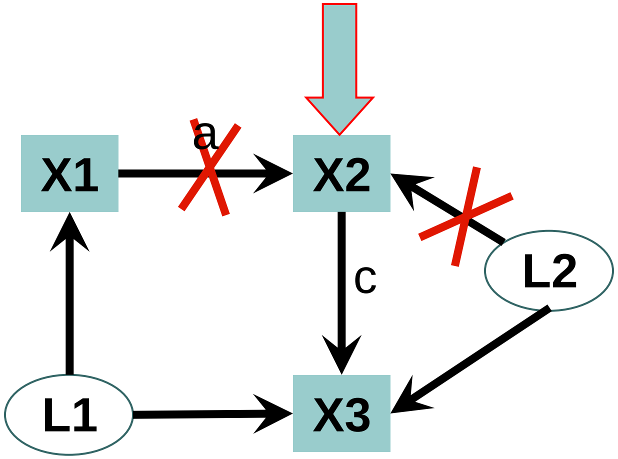




covariance	X1	X2	X3
X1			0
X2			0
X3			1

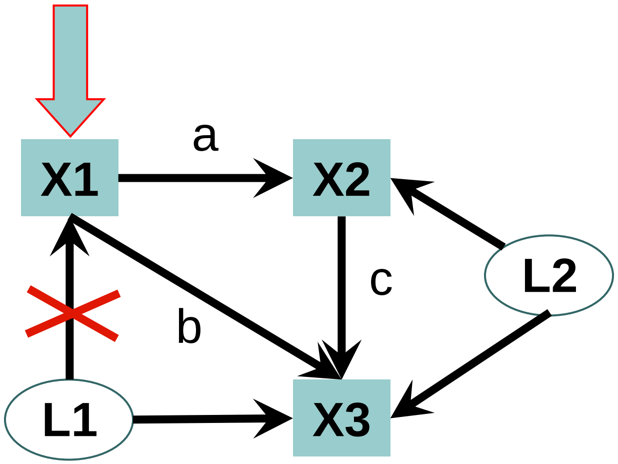
Linearity

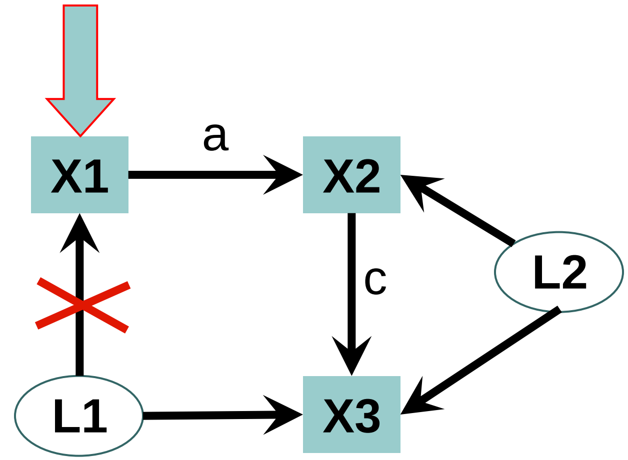




covariance	⇨ X1	⇨ X2	⇨ X3
X1		0	0
X2		1	0
X3		c	1

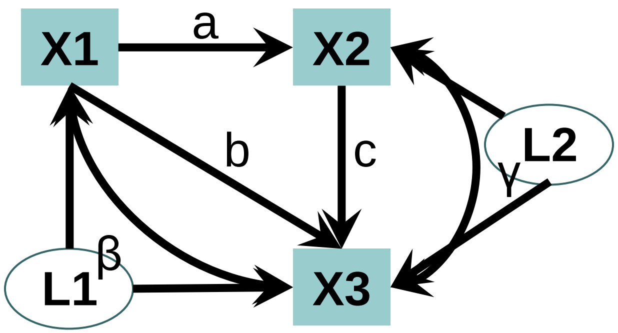
Linearity

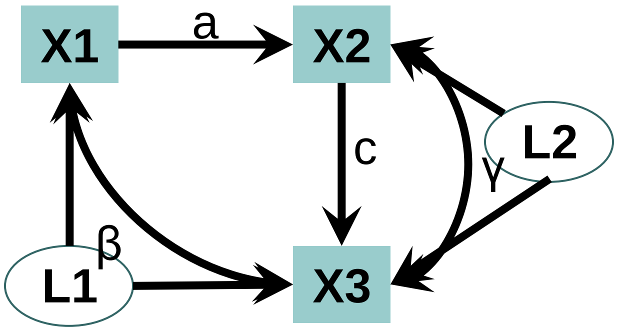




covariance	⇒ X1	⇒ X2	⇒ X3
X1	1	0	0
X2	a	1	0
X3	**ac+b** vs. **ac**	c	1

Linear Model



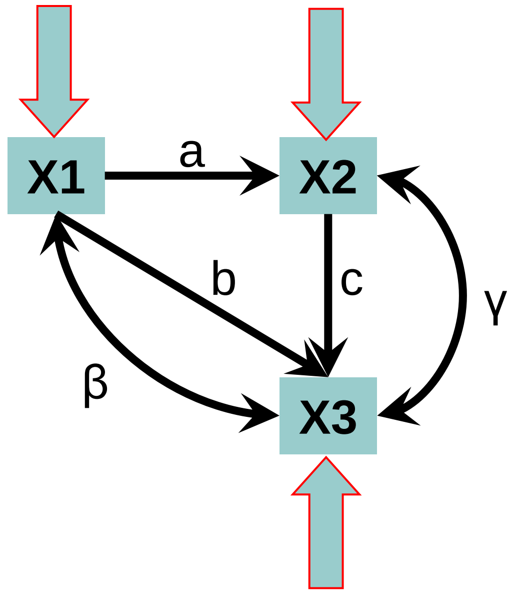


$$\check{x}_{t+1} = B\check{x}_t + \check{e}$$

$$\check{e} \sim N(0, \Sigma)$$

$$B = \begin{pmatrix} 0 & 0 & 0 \\ a & 0 & 0 \\ b & c & 0 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} e1 & 0 & \beta \\ 0 & e2 & \gamma \\ \beta & \gamma & e3 \end{pmatrix} \qquad B = \begin{pmatrix} 0 & 0 & 0 \\ a & 0 & 0 \\ 0 & c & 0 \end{pmatrix}$$

Inference



$$\vec{x}_{t+1} = B\vec{x}_t + \vec{e}$$

$$\vec{e} \sim N(0, \Sigma)$$

$$B = \begin{pmatrix} 0 & 0 & 0 \\ a & 0 & 0 \\ b & c & 0 \end{pmatrix} \qquad A = \begin{pmatrix} 1 & 0 & 0 \\ a & 1 & 0 \\ ac+b & c & 1 \end{pmatrix}$$

Total effect to direct effect: $(I - B)^{-1} = A$

Covariance in terms of correlated errors: $C_x = E(xx^t) = E(Aee^tA^t)$

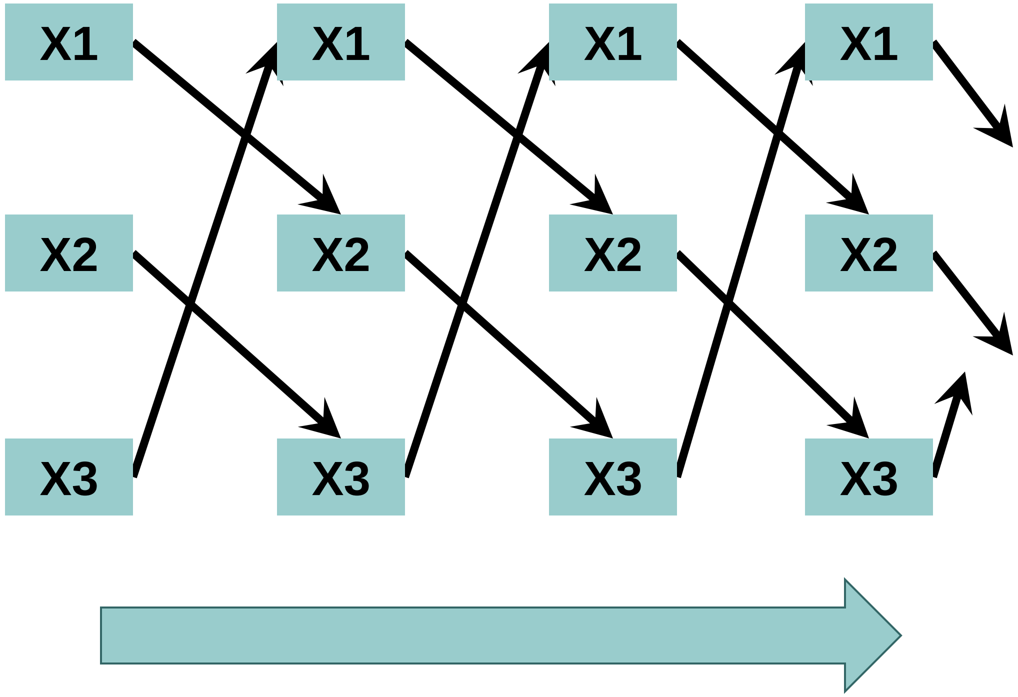
$$= A\Sigma A^t$$

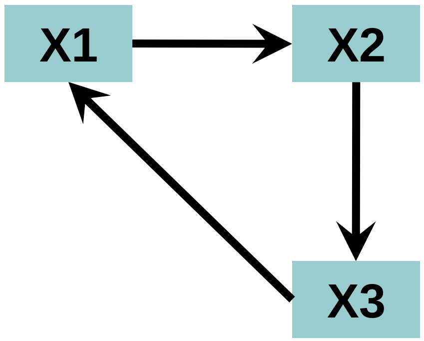
Algorithm

- For each intervention on X_i, compute the total effect of X_i on every other variable (i.e. fill column i of total effect matrix A)
- Given A, compute $B = I – A^{-1}$
- Determine the passive observational covariance matrix C_x
- Determine the error covariance matrix $\Sigma = A^{-1}C_x(A^t)^{-1}$

➔ Structure among observed variables and location of latent variables

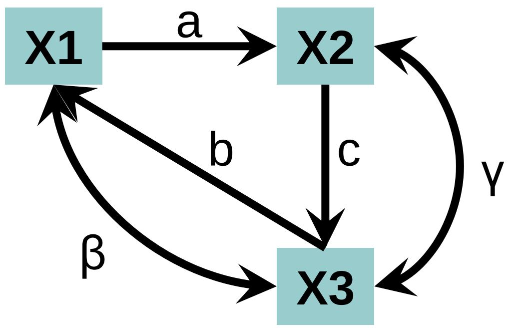
Cyclic Models





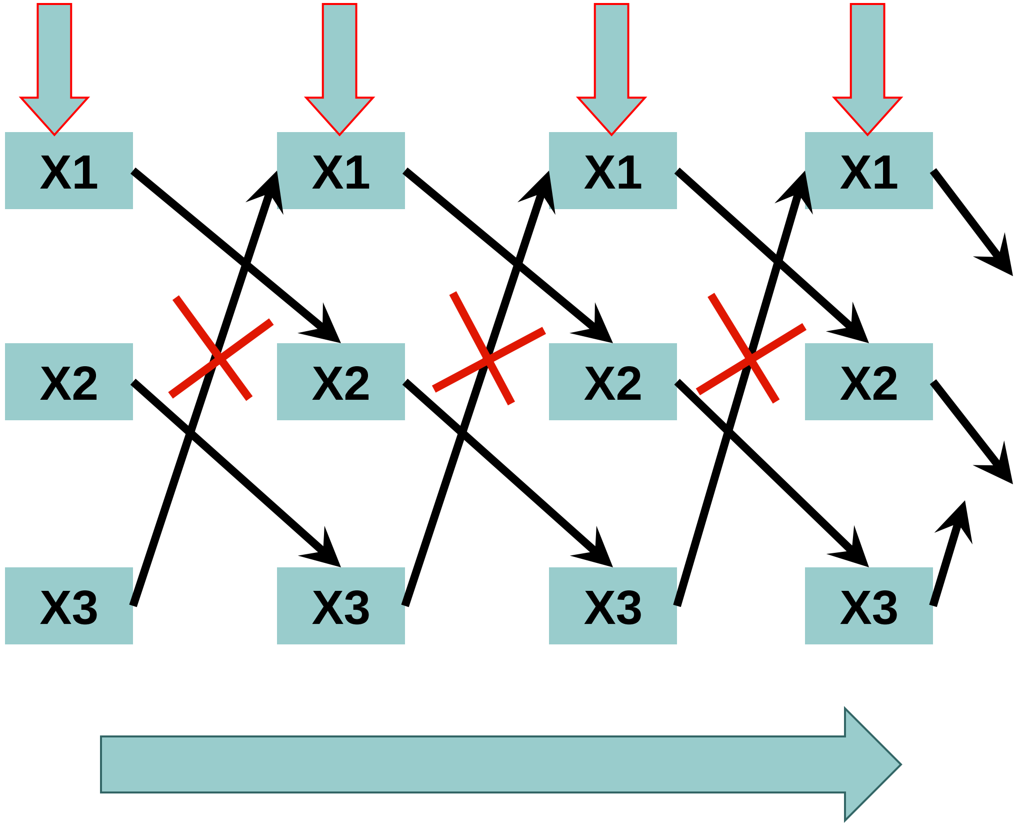
Cyclic Models

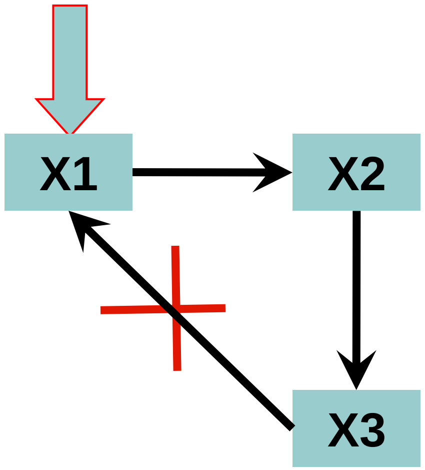
$$B = \begin{pmatrix} 0 & 0 & b \\ a & 0 & 0 \\ 0 & c & 0 \end{pmatrix}$$



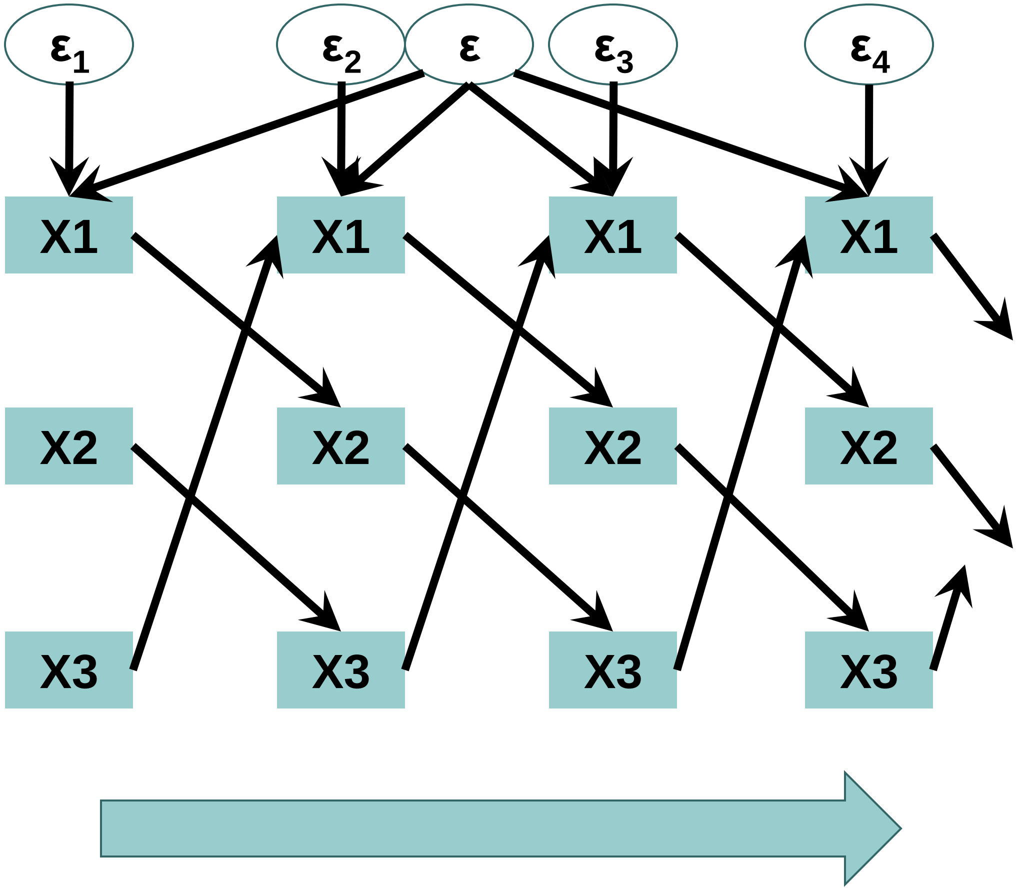
Equilibrium Guarantee: $\left|eigenv(B)\right| < 1$

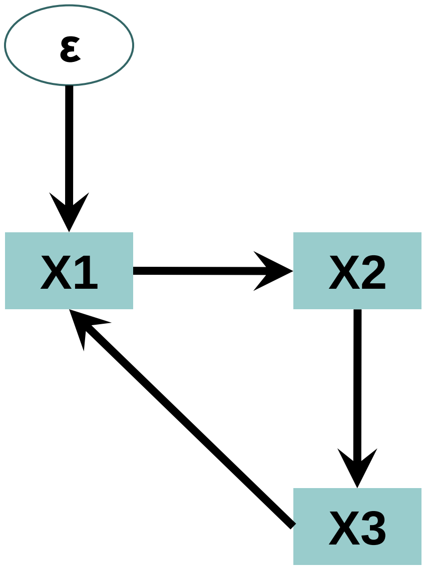
Cycles





Cycles





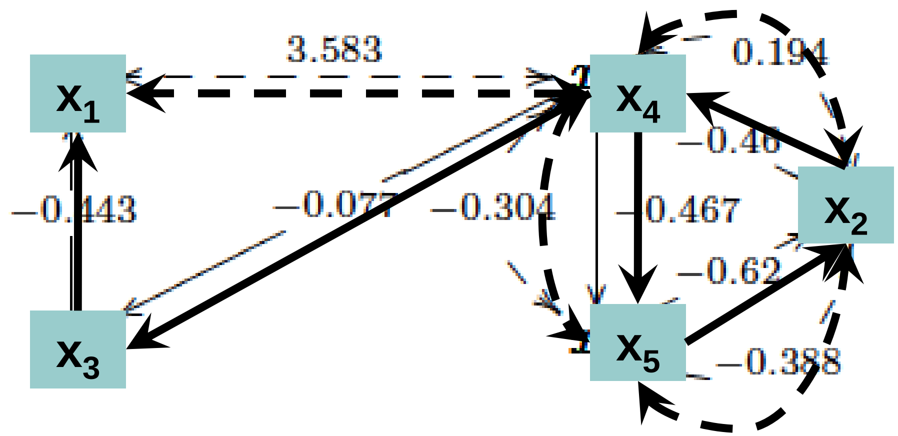
Algorithm

- For each intervention on X_i, compute the total effect of X_i on every other variable (i.e. fill column i of total effect matrix A)
- Given A, compute $B = I – A^{-1}$
- Determine the passive observational covariance matrix C_x
- Determine the error covariance matrix Σ as a function of C_x and B.

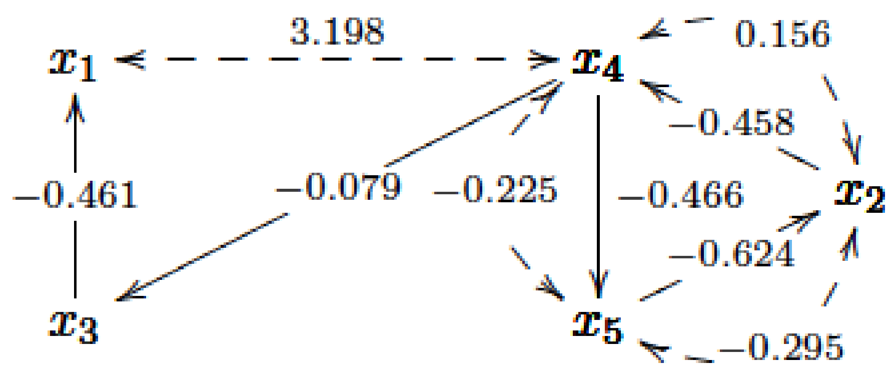
➔ Structure among observed variables and location of latent variables

Example

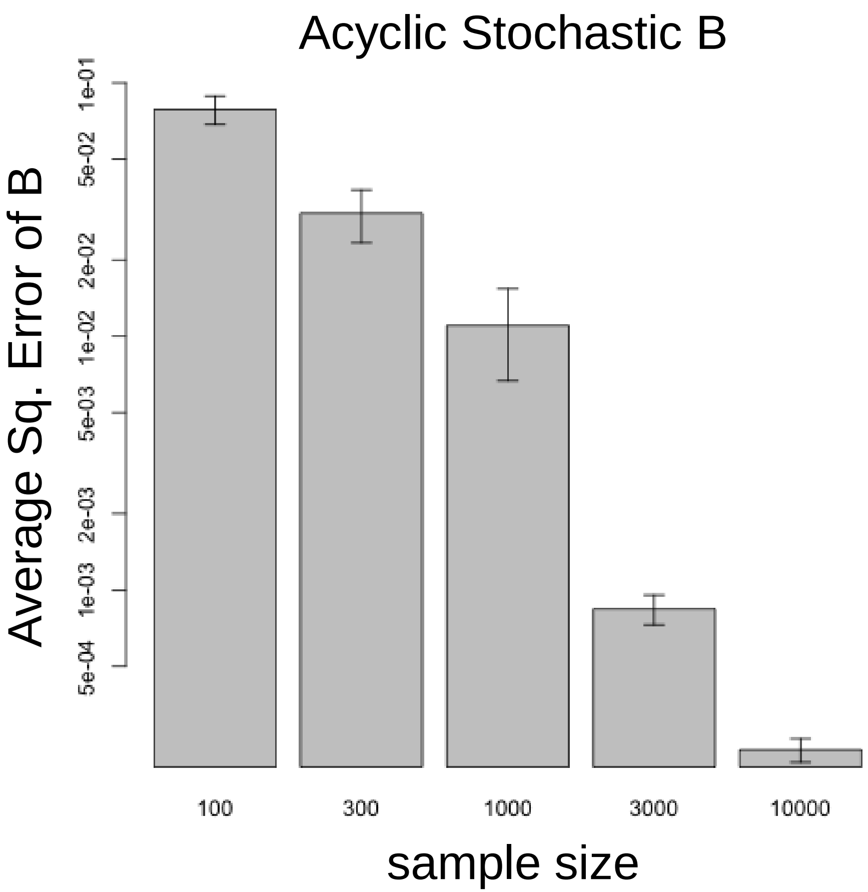
Ground Truth



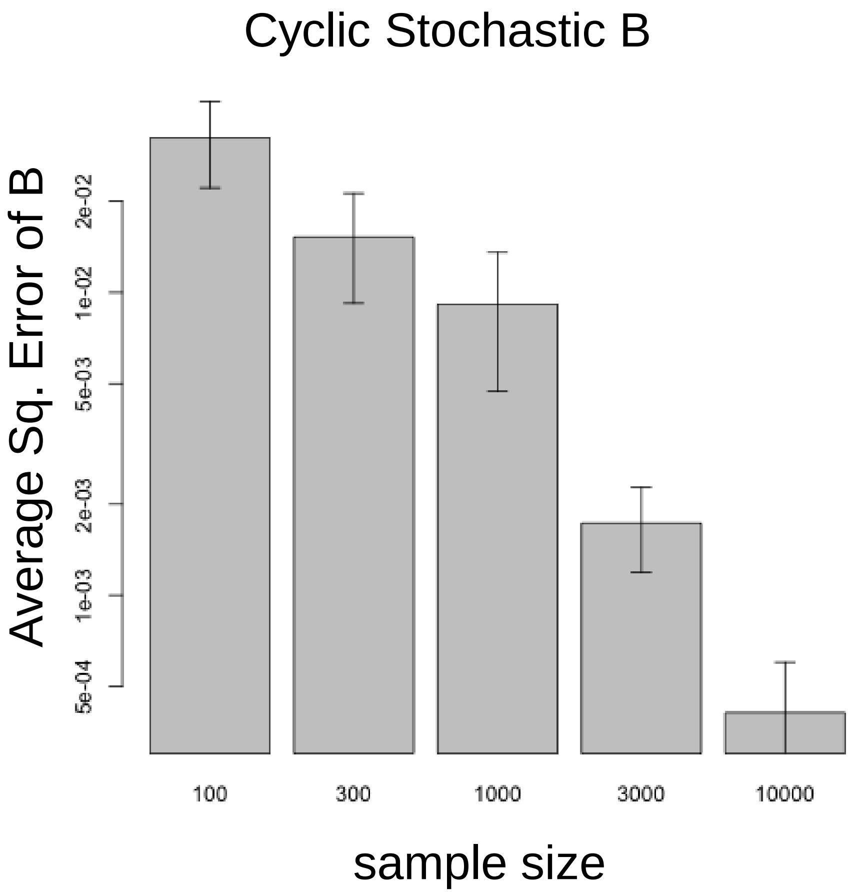
Estimated Model



Simulation



Acyclic Stochastic B
Average Sq. Error of B
1e-01
5e-02
2e-02
1e-02
5e-03
2e-03
1e-03
5e-04
100
300
1000
3000
10000
sample size



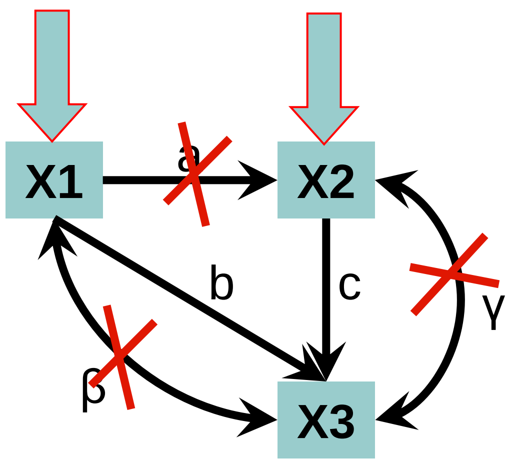
Cyclic Stochastic B
Average Sq. Error of B
2e-02
1e-02
5e-03
2e-03
1e-03
5e-04
100
300
1000
3000
10000
sample size

Results

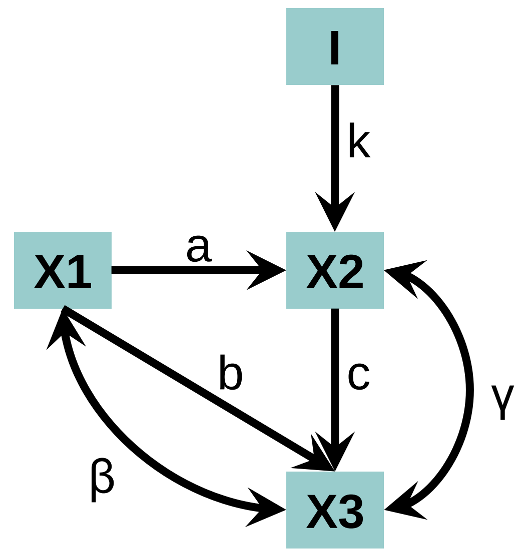
- Given an experimental intervention on each of a set of variables, and assuming the causal relations are linear, we can determine
- → The exact causal connection among the intervened variables
- → Presence and location of unmeasured common causes
- → for a large variety of acyclic and cyclic models
- → with stochastic or deterministic causal relations

Other Types of Interventions

Multiple Simultaneous Interventions

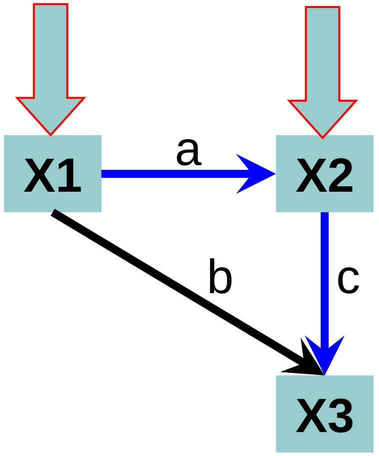


Soft Interventions / Instrumental Variables



Weakening Linearity

- Compositionality
  - Trek computation



Bigger Picture

- Optimal Sequences of Experiments
  - Different types of interventions
  - Different types of assumptions
  - Algorithms that combine results from different experiments
  - [trading off discovery and its cost]

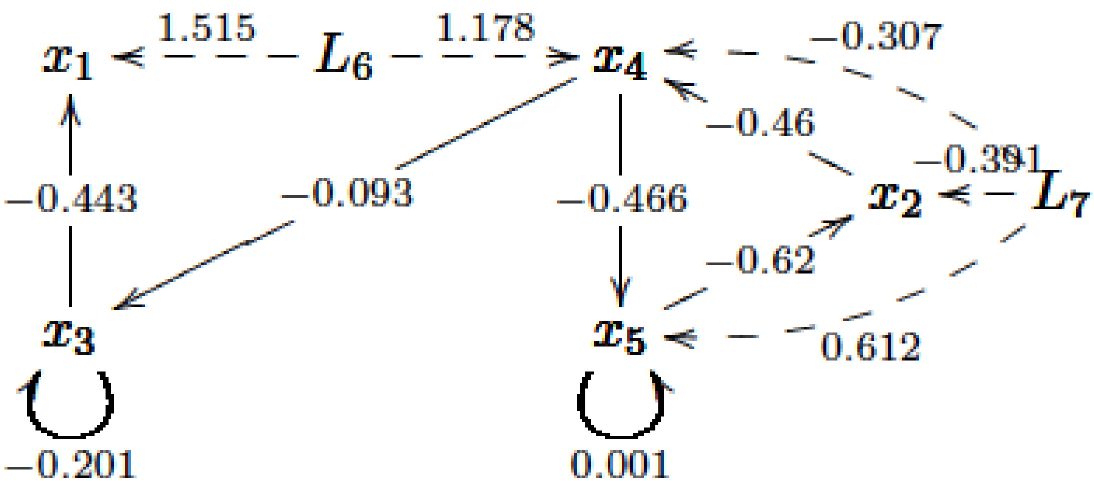
Collaborators

- Patrik Hoyer (University of Helsinki)
- Richard Scheines (Carnegie Mellon University)

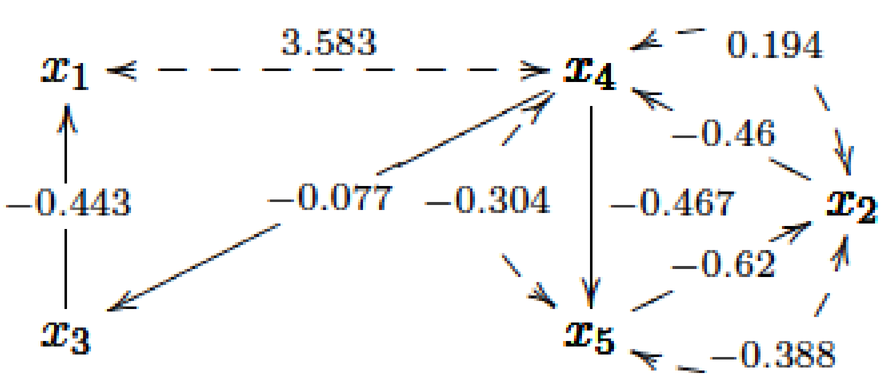
Thank you!

Frederick Eberhardt
Philosophy-Neuroscience-Psychology Program,
Department of Philosophy;
Washington University in St. Louis
eberhardt@wustl.edu
www.artsci.wustl.edu/~feberhar

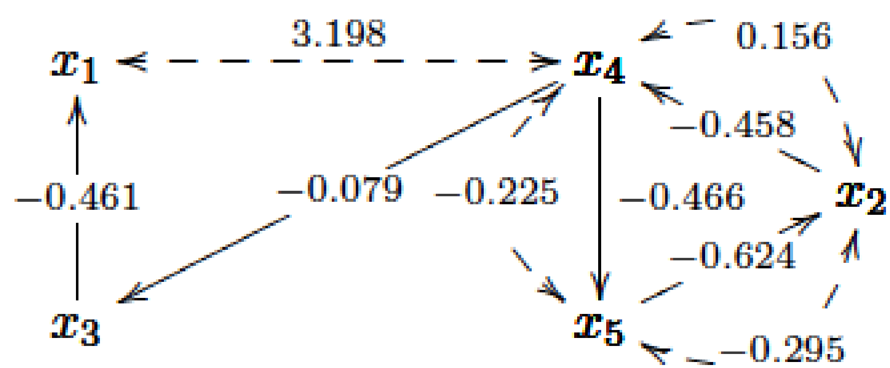
Self Loops



x_1
1.515
L_6
1.178
x_4
−0.307
−0.46
−0.391
x_2
L_7
−0.443
−0.093
−0.466
−0.62
x_3
x_5
0.612
−0.201
0.001



x_1
3.583
x_4
0.194
−0.46
−0.443
−0.077
−0.304
−0.467
x_2
−0.62
x_3
x_5
−0.388



x_1
3.198
x_4
0.156
−0.458
−0.461
−0.079
−0.225
−0.466
x_2
−0.624
x_3
x_5
−0.295

Bounds for Independence Constraints

Interventions per experiment	Strength of Intervention	Number of experiments	
		No Latents	Latents
Single	Hard	N-1	impossible
Single	Soft	N-1	impossible
Multiple	Hard	$\log_2(N)+1$	N
Multiple	Soft	1!!	impossible

Bounds for **Linear** Models

Bonus:
Presence and location of latent variables and structure among latent variables can be discovered.

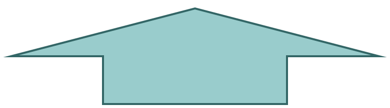
Interventi per experime		mber of experiments	
		Latents	Latents
Single	Hard	N-1	impossible / N
Single	Soft	N-1	impossible / N
Multiple	Hard	$\log_2(N)+1$	$< 2\log_2(N)$ / N
Multiple	Soft	1	impossible / 1

Bridge Principles

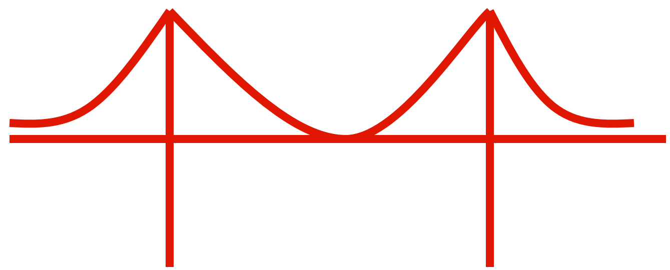
Statistical Properties

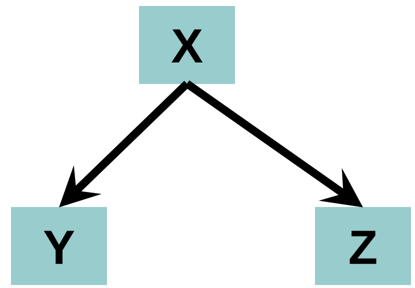
Causal Structure

$$P(X,Y,Z)$$



X	Y	Z
- 1.01	- 0.71	5.18
- 0.85	0.16	- 3.31
0.11	- 0.59	- 1.00
0.20	1.47	- 3.21
...	...	...





Causal Markov Condition

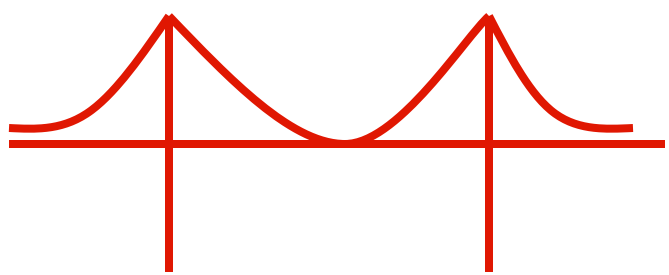
Each causal variable is independent of its non-descendents given its causes.

Probability

- Probabilistic Dependence
- Probabilistic Independence

Causality

- Causal Connection
- Causal Separation



Causal Faithfulness Condition

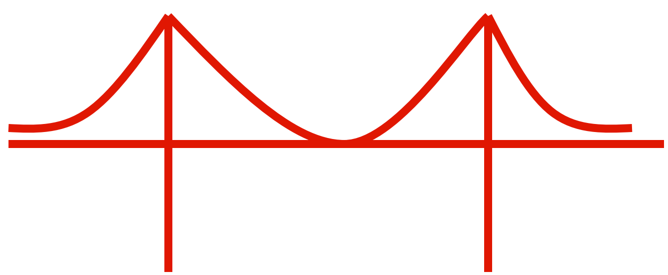
All the independence relations in the probability distribution are due to the Markov condition.

Probability

- Probabilistic Dependence
- Probabilistic Independence

Causality

- Causal Connection
- Causal Separation



Using Markov and Faithfulness

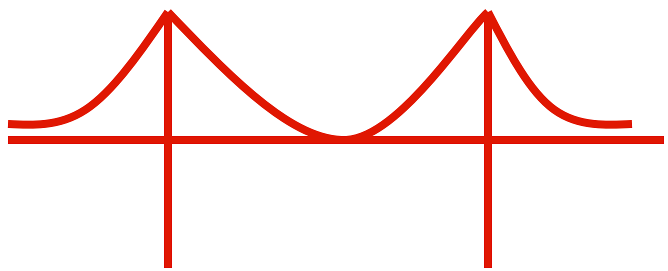
Statistical Properties **Causal Structure**

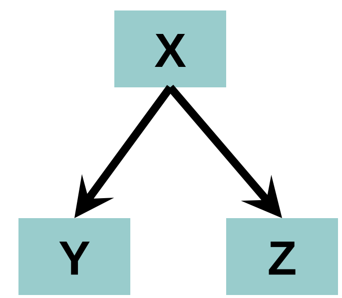
$X \not\perp\!\!\!\perp Y$

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

$Y \not\perp\!\!\!\perp Z$

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





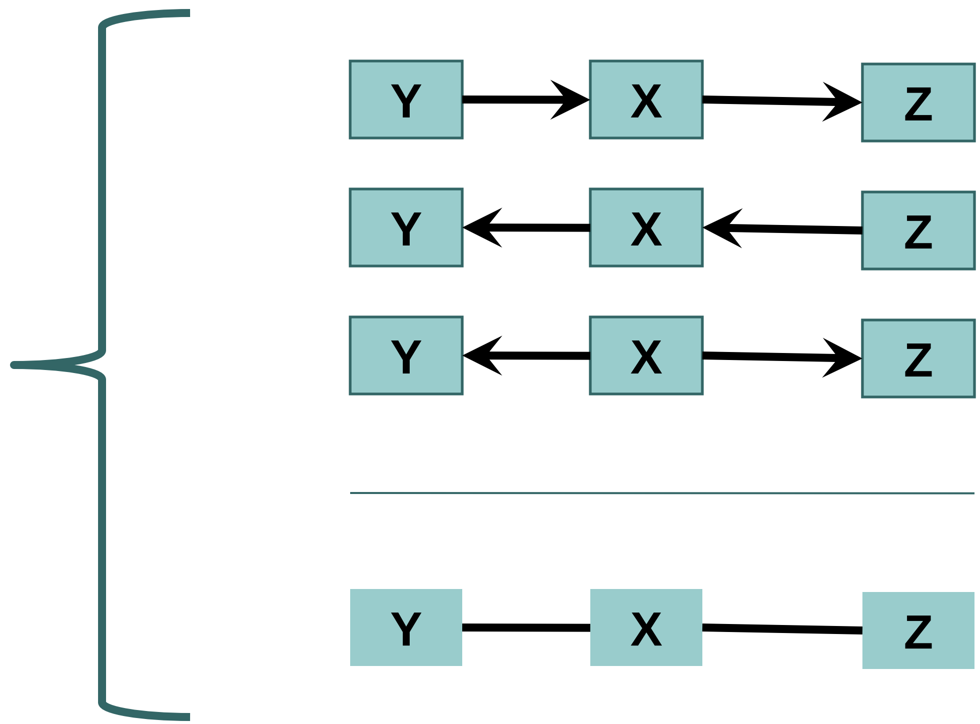
$$P(Y)P(X \mid Y)P(Z \mid X)$$

Underdetermination

Statistical Properties

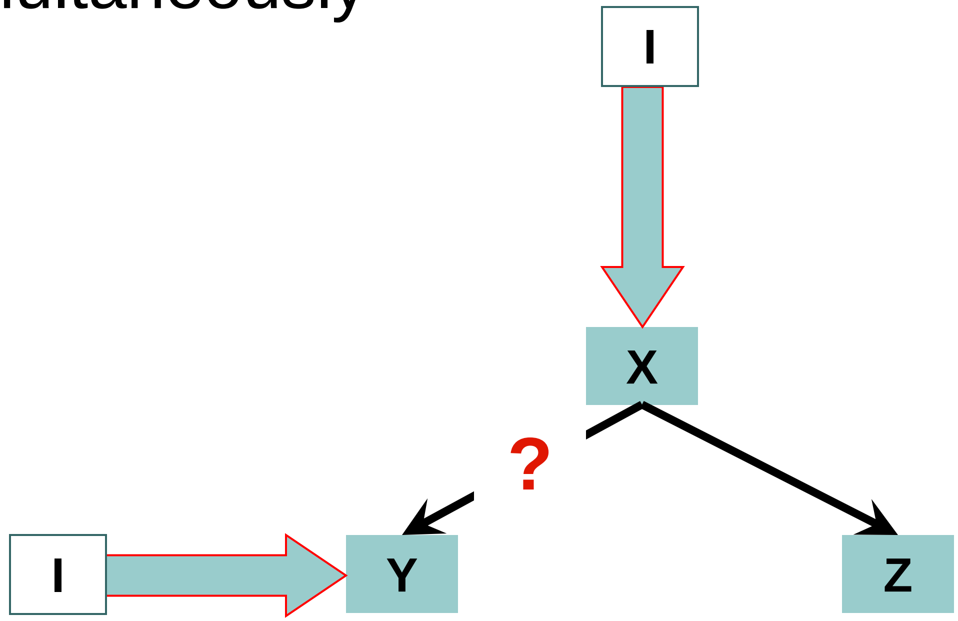
$X \not\perp\!\!\!\perp Y$
$X \not\perp\!\!\!\perp Z$
$Y \not\perp\!\!\!\perp Z$
$Y \perp\!\!\!\perp Z \mid X$

Equivalence Classes



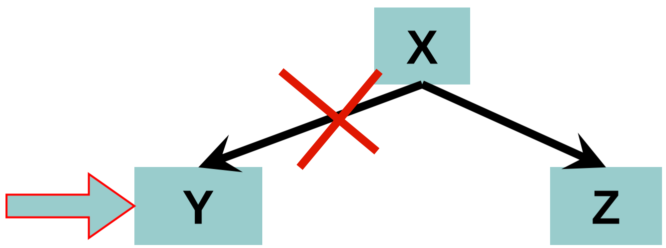
Intervention and Discovery (2)

- One at a time
- Multiple Simultaneously



Example with 3 Variables

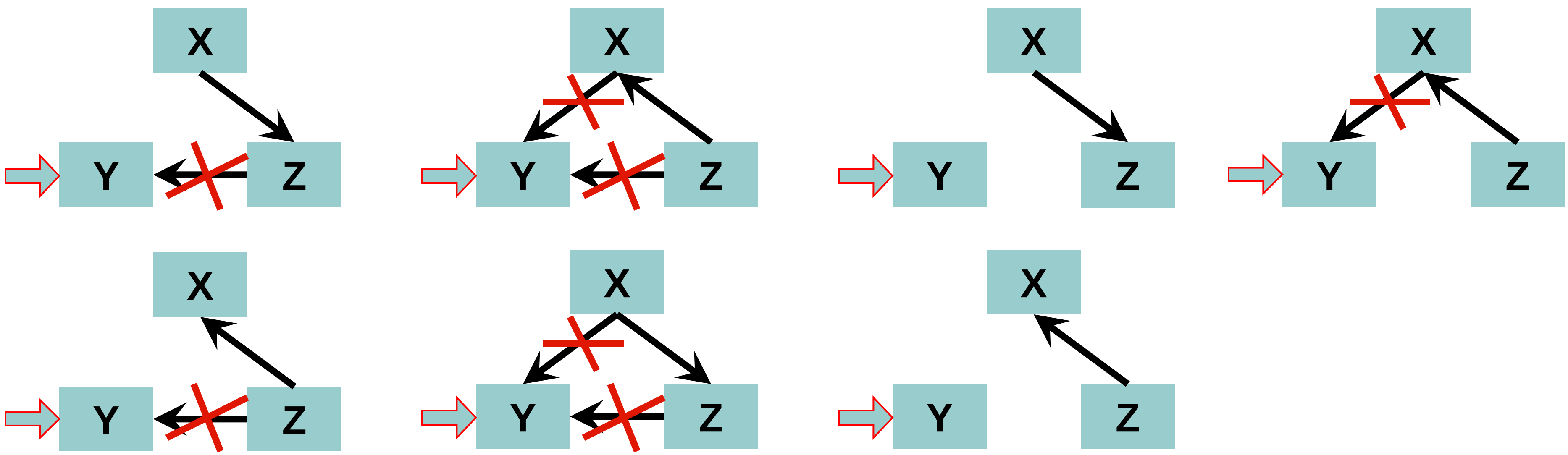
Truth (unknown)



X _||_ Y   X _||_ Y | Z

X _~~||~~_ Z   X _||_ ~~Z~~ | Y

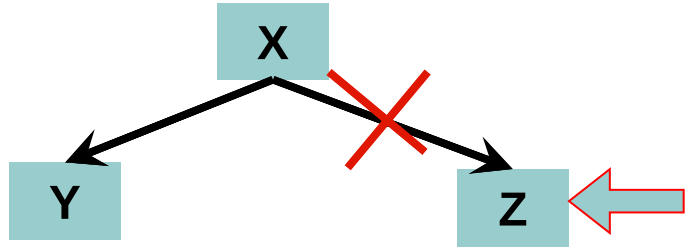
Y _||_ Z   Y_||_Z | X



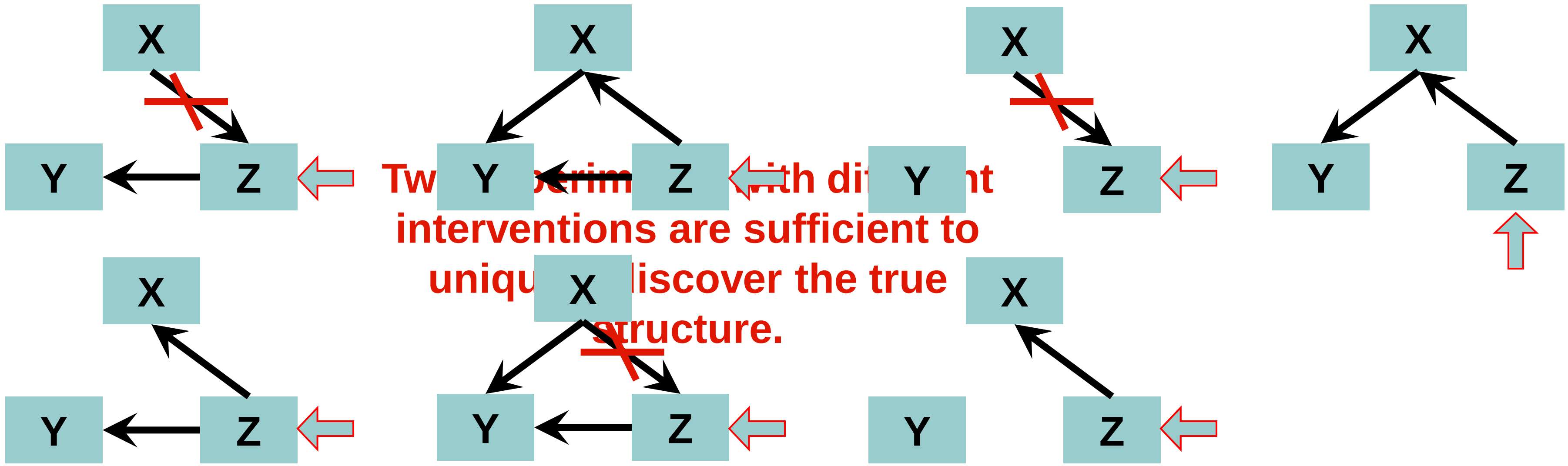
Another intervention is needed to uniquely identify the true structure!

Example with 3 Variables

Truth (unknown)

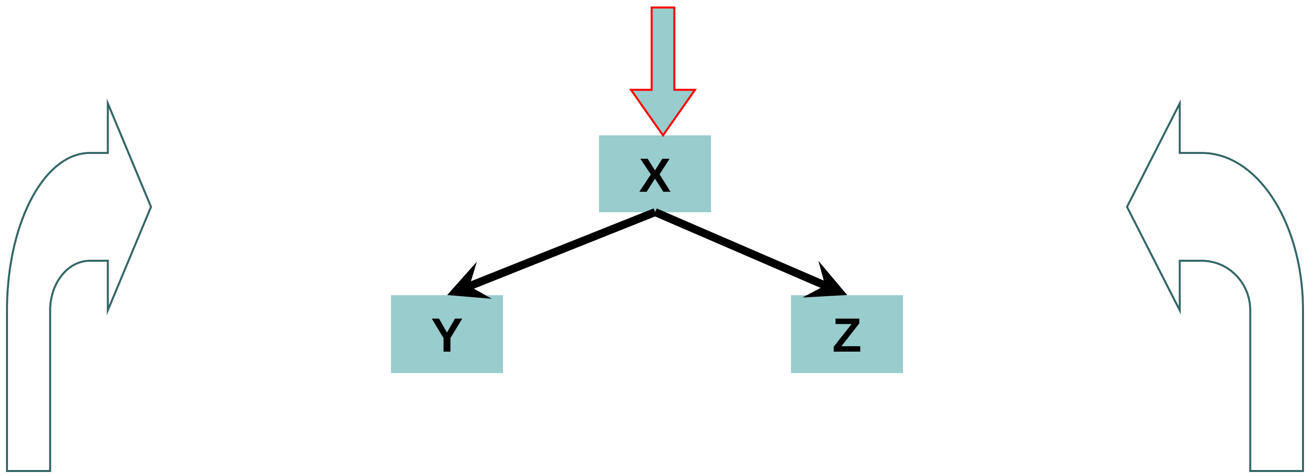


$X \not\perp\!\!\!\perp Y$	$X \not\perp\!\!\!\perp Y \mid Z$
$X \perp\!\!\!\perp Z$	$X \perp\!\!\!\perp Z \mid Y$
$Y \perp\!\!\!\perp Z$	$Y \perp\!\!\!\perp Z \mid X$



Tw… perim… with dif… nt interventions are sufficient to uniqu… liscover the true structure.

But we could have been lucky…



- X ~~_||_~~ Y
- X ~~_||_~~ Z
- Y ~~_||_~~ Z

- X ~~_||_~~ Y | Z
- X ~~_||_~~ Z | Y
- Y _||_ Z | X

- Intervention on X

Worst Case Sequences

Interventions per Experiment	Worst case number of experiments
Single	N-1 (for N>2; 2 if N=2)
Multiple	$\log_2(N) + 1$ (for N>2; 2 for N=2)

- Worst case is not representative
- What about the expected case?
  - Expected case depends on distribution of possible hypotheses
  - Use an uninformative distribution (?) – flat?

Discovery and its cost

- Some experiments might be impossible or unethical
- Some experiments might be more expensive than others
- Collecting data passive observationally may be cheap, while experimental data is expensive

➔ **Cost Functions**

Discovery as a Game against Nature

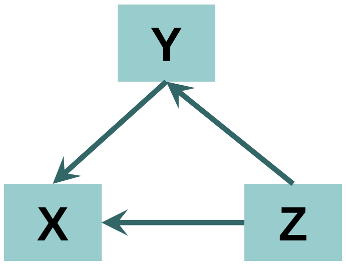
- Sequences of experiments as strategies in a sequential game
- Cost functions represented in the pay-off structure
- Optimal trade-off between discovery and its cost is solution to the game
- Expected result of a strategy can be characterized as the expected outcome of a rational response to an opponent (Nature) in a game.

Three Different Strategies

Fixed Strategy:

Ex1: X

Ex2: Y

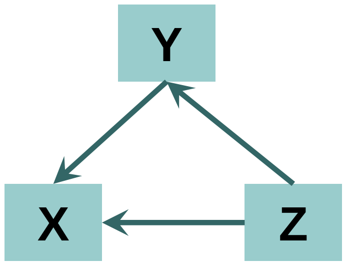


Adaptive Strategy:

Ex1: X

If resolved, no further experiment,

otherwise Ex2: Y

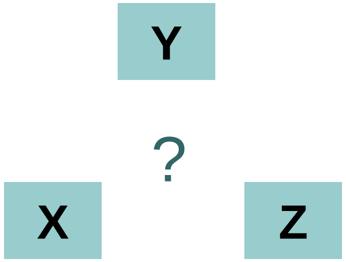


Mixed Strategy:

Ex1: K where K is sampled from {X,Y,Z} with 1/3.

If resolved, no further experiment,

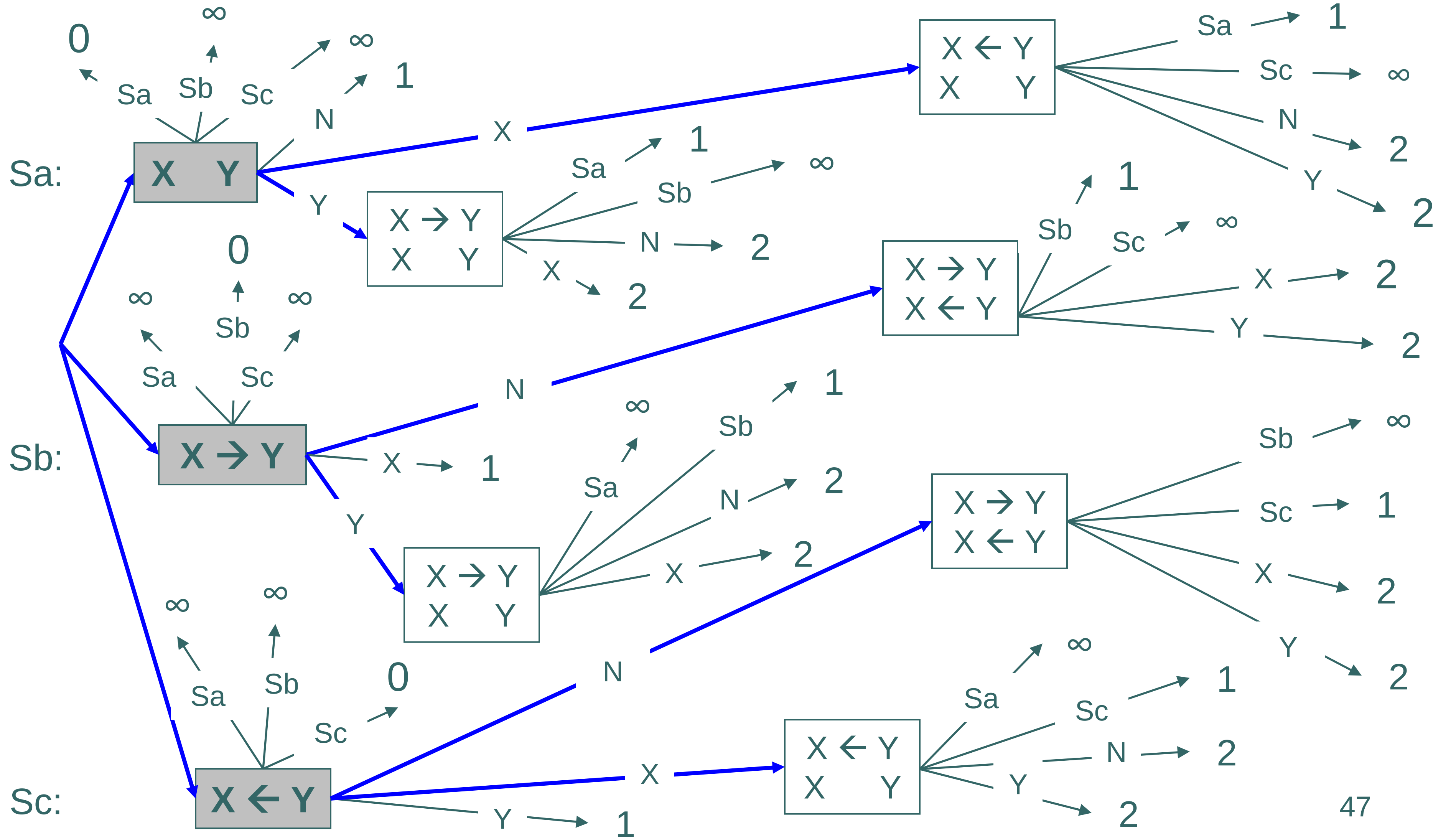
otherwise Ex2: M where M is sampled with ½ from {X,Y,Z}\{K}



Rules of the Game

- Two player (Nature & Scientist) zero-sum game
- Nature chooses the true graph
- Scientist chooses and performs experiments
- Independence relations true in the manipulated distribution are revealed in each experiment
- Nature may not change true graph after initial setting
- Pay-off to Nature is the number of experiments that are performed
- Scientist may end game at any point by announcing a graph. If it is false, the loss is infinity.

Game with 2 Variables



Analyzing the Game

- 2-person zero-sum game implies that the Nash equilibrium corresponds to the mini-max solution

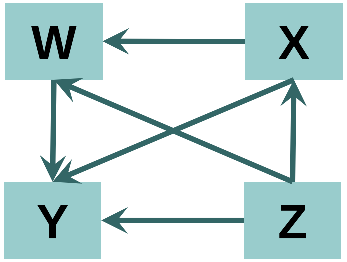
- ➔ Nash equilibrium identifies cost of optimal strategy when Nature is an adversarial opponent
- ➔ Nash equilibrium specifies distribution over structures that is most costly to learn
- ➔ Nash equilibrium describes a sequence of experiments that is globally optimal

Results for Games

- For single (or no) interventions per experiment:
  - N=2: max(E(#exp)) = 5/3
  - N=3: max(E(#exp)) = 2 =N-1!!!
  - N>3: max(E(#exp)) = 2/3 N – 1/3
- Nature's distribution: uniform over ***complete*** structures
- Scientist's strategy: uniform sampling (without replacement) of variables for intervention

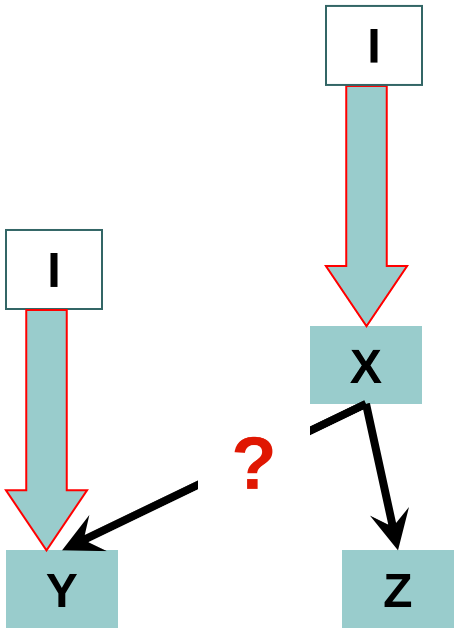
Nature's distribution over structures

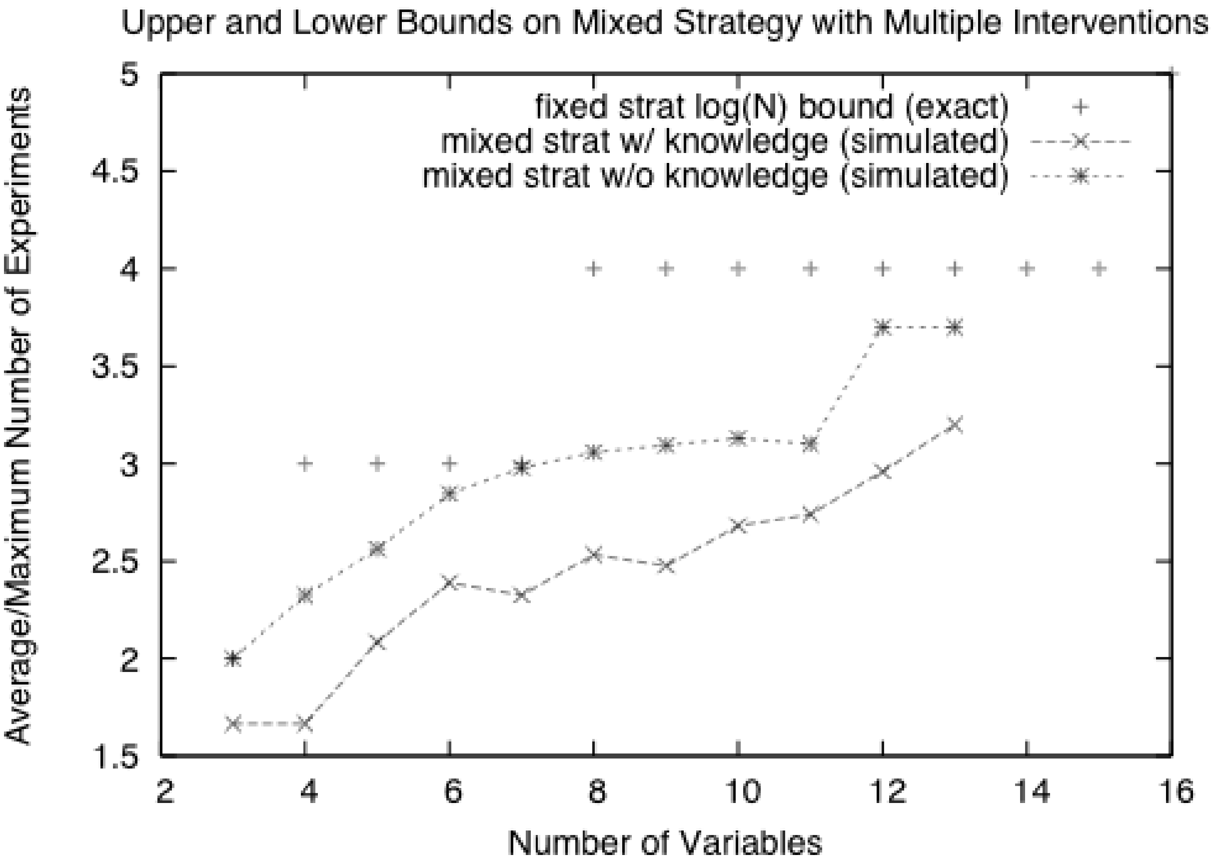
- The most costly discovery problem occurs when Nature samples from a distribution that is uniform over complete causal structures.



- ➔ This is not the "uninformative" distribution over all possible causal structures.
- ➔ Once multiple simultaneous interventions are permitted in one experiment, even the above no longer holds.

Multiple Simultaneous Interventions per Experiment





Results for Strategies

Theorem 1: No adaptive search strategy is better than a static strategy for the worst case expectation.

Theorem 2: Adaptive strategies are better than static strategies in expectation for most distributions over graphs.

Theorem 3: There is a mixed strategy that strictly dominates any static or adaptive strategy in expectation, but only for N>3 variables.

…

Many open problems…

- Robustness of search strategies
  - off-equilibrium play by Nature
- Is nature an adversarial opponent?
  - If not, we can now analyze the game accordingly, without an a priori commitment to a particular distribution over structures
- Changing the rules of the game
  - Perhaps Nature is not the “constant gardener”
- Change the cost functions
  - Experiments with different costs
  - Cost per sample, etc.

Bigger Problems

- Analyzing the “cost” of an assumption
  - What effect does the assumption of acyclicity have?
- Representing large games

Strategy / Structure	⇨ X ⇨ Y	⇨ X ⇨ Z	⇨ Y ⇨ Z	…
X Y				
X → Y				
X ← Y				

Super-exponential in N

> Factorial in N

On the shoulders of…

- Abraham Wald (1902 – 1950)
- *Statistical Decision Functions* (1950)
- Here: Extension of Wald’s framework to sequences of experiments.