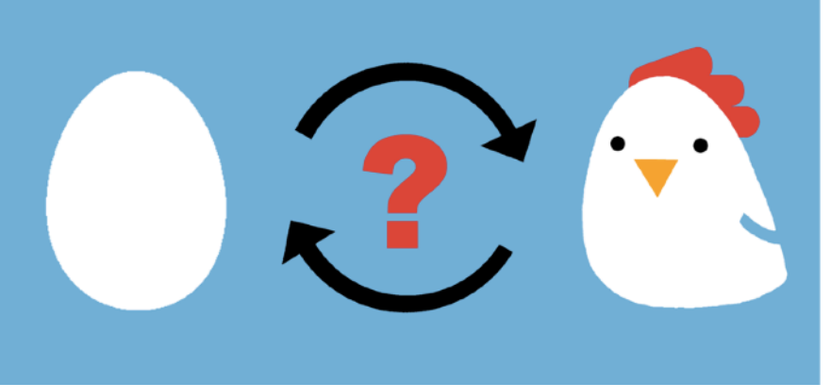


Causality and Machine Learning
(80-816/516)

Classes 10 (Feb 13, 2025)

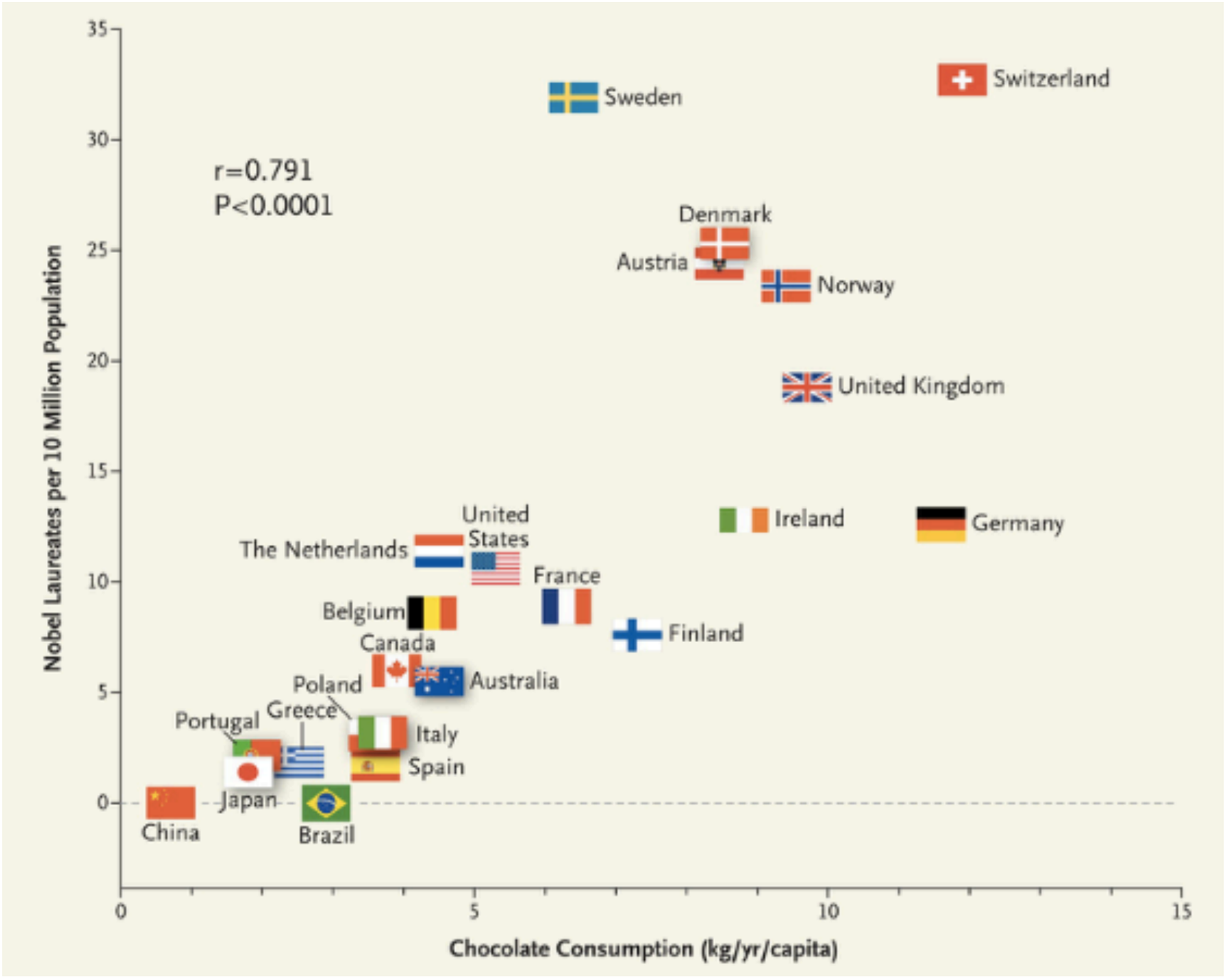
Identification of Causal Effects (Causal Inference)

Clark Glymour

Causality vs. Association

Causality vs. Association

Another Example

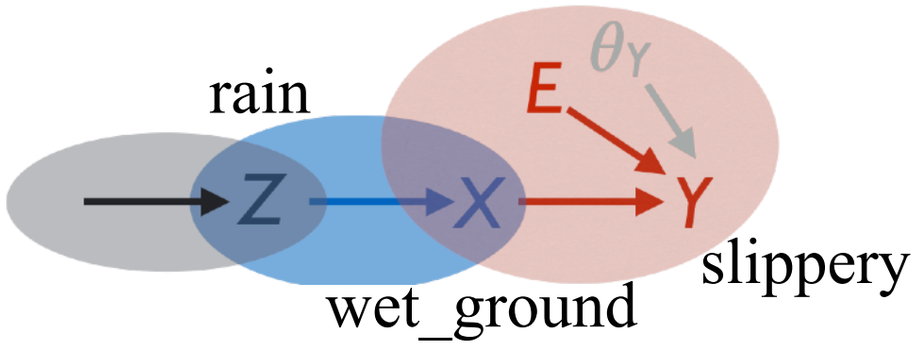


What Information Helps Find Causality?

- Connection between causal structure and statistical data under *suitable assumptions*

- Note this "irrelevance":

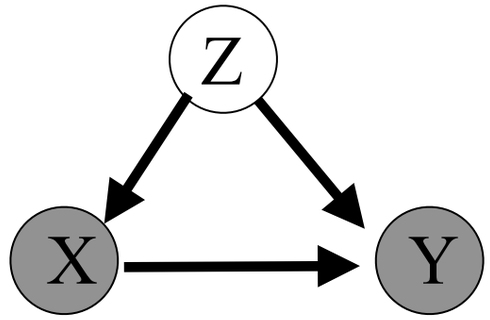
If there is no common cause of X and Y, the generating process for cause X is irrelevant to ("independent" from) that generates effect Y from X



- conditional independence among variables;
- independent noise condition;
- minimal (and independent) changes...

Causal Sufficiency

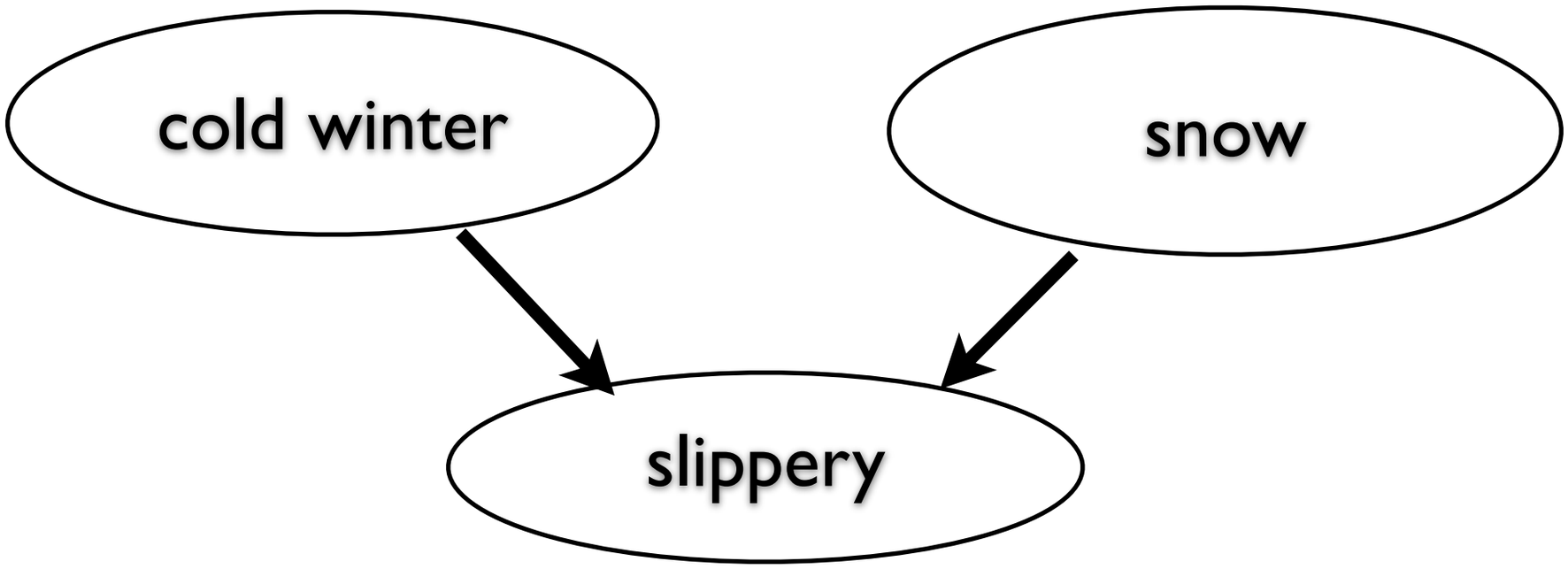
- A set of random variables ***V*** is causally sufficient if ***V*** contains every common cause (with respect to ***V***) of any pair of variables in ***V***



- $\boldsymbol{V} = \{X, Y, Z\}$: causally sufficient
- $\boldsymbol{V} = \{X, Y\}$: causally insufficient
- Methods exist in causally **insufficient** cases, e.g., FCI (*Chapter 6 of the SGS book*)

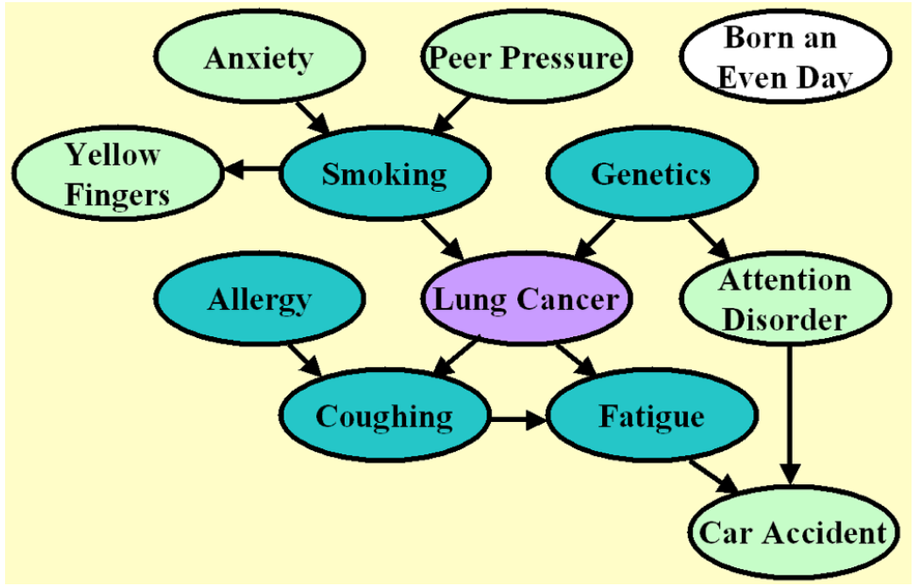
SGS Book, Chapter 5 (for causally sufficient structures); Chapter 6 (without causal sufficiency)

V-Structures



Why so interesting?

We can See CI Relations from DAGs...



- Local Markov condition
- Global Markov condition
- d-separation implies conditional independence:

$P(\mathbf{V})$, where $\mathbf{V}$ denotes the set of variables, obeys the global Markov condition (or property) according to DAG $\mathcal{G}$ if for any disjoint subsets of variables $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$, we have

$$\mathbf{X} \text{ and } \mathbf{Y} \text{ are d-separated by } \mathbf{Z} \text{ in } \mathcal{G} \implies \mathbf{X} \perp\!\!\!\perp \mathbf{Y} \,|\, \mathbf{Z}.$$

Going from CI to Graph?

$$\mathbf{X} \text{ and } \mathbf{Y} \text{ are d-separated by } \mathbf{Z} \text{ in } \mathcal{G} \implies \mathbf{X} \perp\!\!\!\perp \mathbf{Y} \,|\, \mathbf{Z}.$$

- Contrapositive:
  - Conditional dependence implies d-connection
  - What if variables are conditionally independent?
- Can we recover the property of the underlying graph from CI relations with Markov condition?
  - Arbitrary $P(\mathbf{V})$ would satisfy the global Markov condition according to G^{f} *in which there is an edge between each pair of variables*: trivial !
  - Under what assumptions can we have CI ⇒ d-separation?

Causal Structure vs. Statistical Independence (SGS, et al.)

Causal Markov condition: each variable is ind. of its non-descendants (non-effects) conditional on its parents (direct causes)

causal structure (causal graph)

$Y \rightarrow X \rightarrow Z$

Y -- X -- Z **?**

Statistical independence(s)

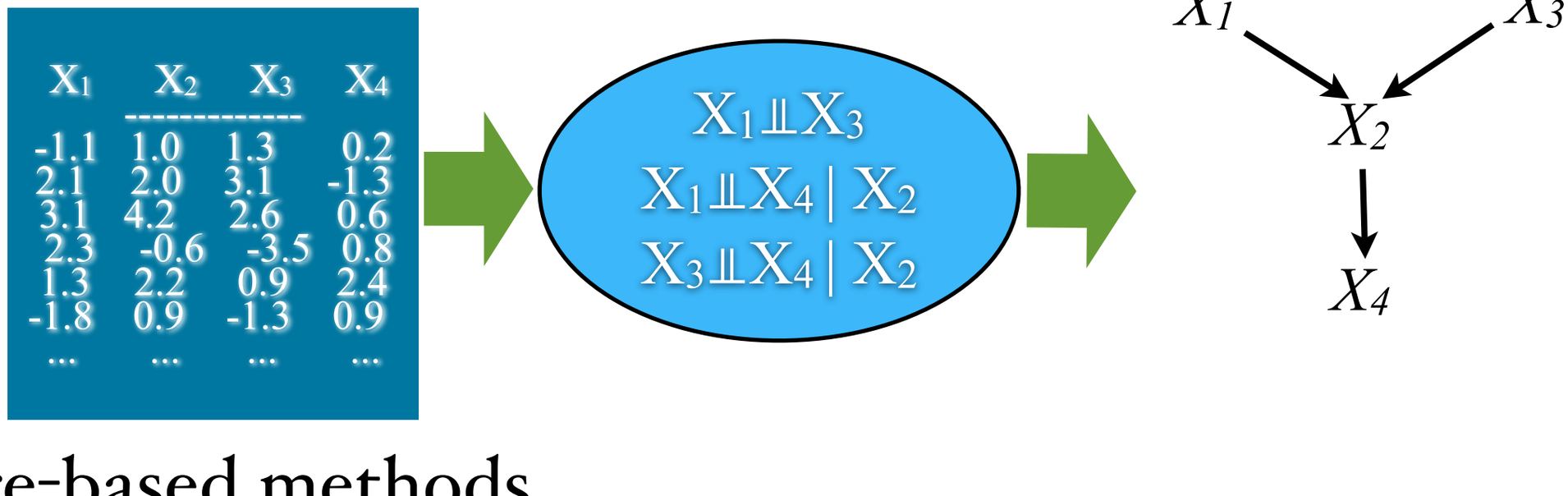
$Y \perp\!\!\!\perp Z \mid X$

Faithfulness: all observed (conditional) independencies are entailed by Markov condition in the causal graph

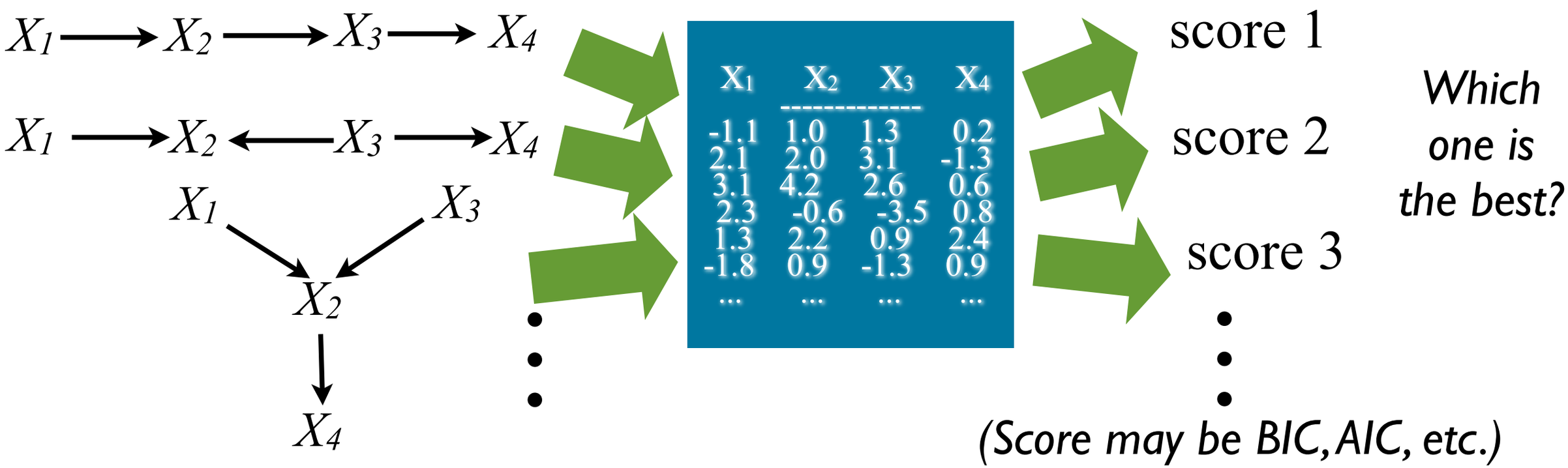
Recall: $Y \perp\!\!\!\perp Z \Leftrightarrow P(Y|Z)=P(Y)$; $Y \perp\!\!\!\perp Z|X \Leftrightarrow P(Y|Z,X)=P(Y|X)$

Constraint-Based vs. Score-Based

- Constraint-based methods



- Score-based methods



Constraint-Based vs. Score-Based

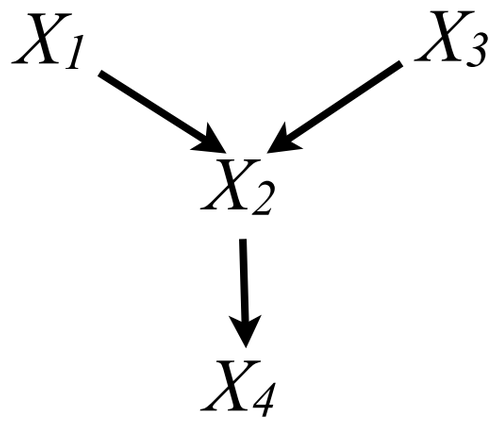
- Constraint-based methods

X_1	X_2	X_3	X_4
-1.1	1.0	1.3	0.2
2.1	2.0	3.1	-1.3
3.1	4.2	2.6	0.6
2.3	-0.6	-3.5	0.8
1.3	2.2	0.9	2.4
-1.8	0.9	-1.3	0.9
...	...	...	...

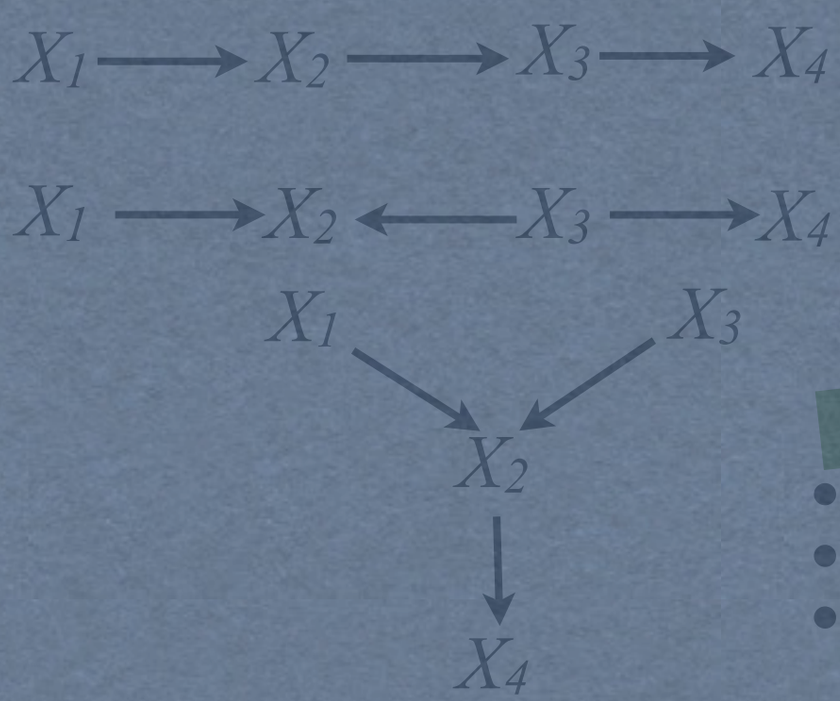
$X_1 \perp\!\!\!\perp X_3$

$X_1 \perp\!\!\!\perp X_4 \mid X_2$

$X_3 \perp\!\!\!\perp X_4 \mid X_2$



- Score-based methods



X_1	X_2	X_3	X_4
-1.1	1.0	1.3	0.2
2.1	2.0	3.1	-1.3
3.1	4.2	2.6	0.6
2.3	-0.6	-3.5	0.8
1.3	2.2	0.9	2.4
-1.8	0.9	-1.3	0.9
...	...	...	...

score 1

score 2

score 3

Which one is the best?

(Score may be BIC, AIC, etc.)

Discussion

- First, can we find the skeleton of the causal structure? If yes, how?

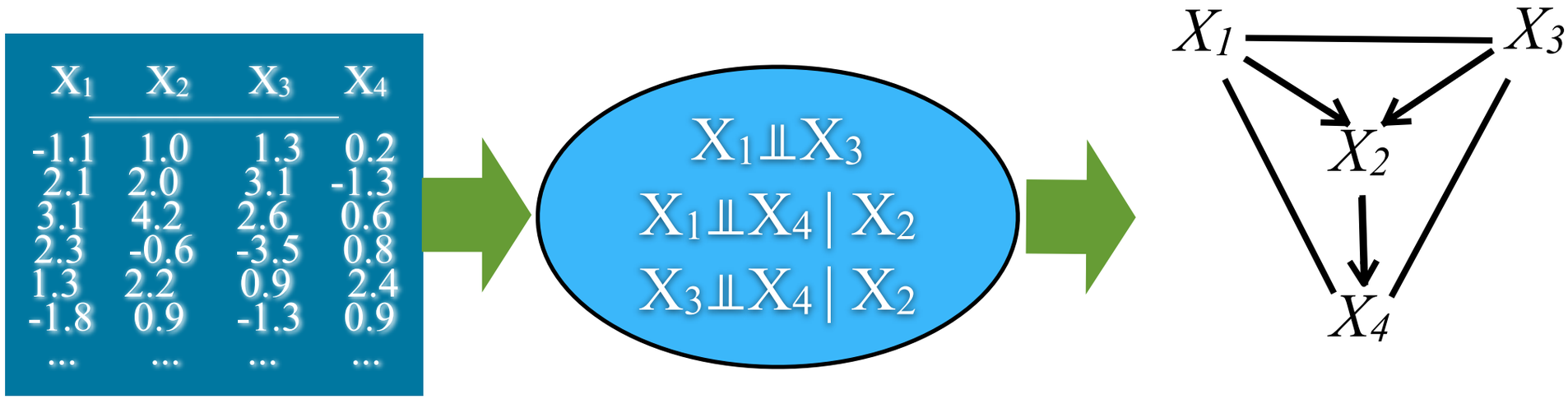
Causal Markov condition + faithfulness

- Second, can we determine the causal direction?

How?

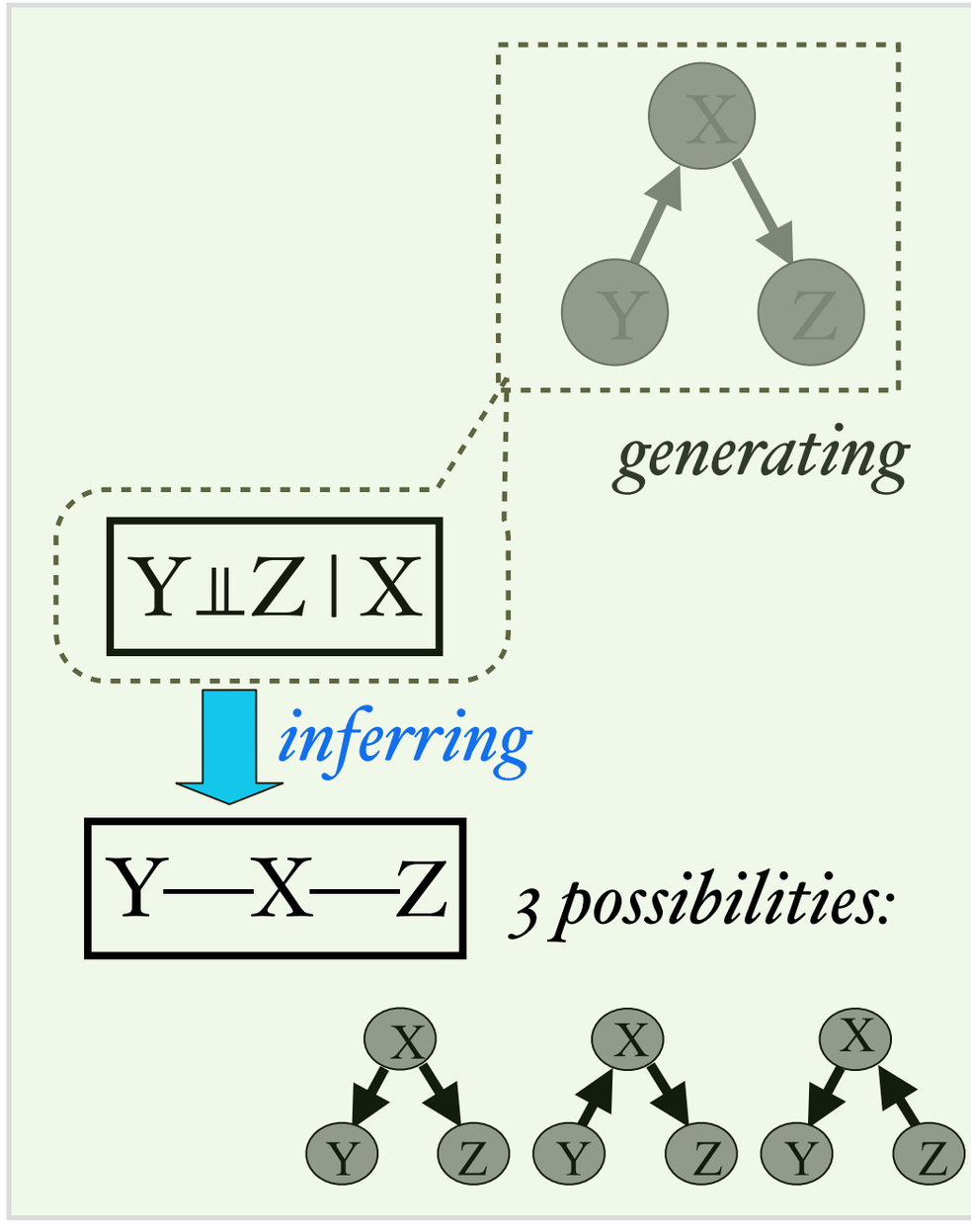
Constraint-Based Causal Discovery: Big Picture

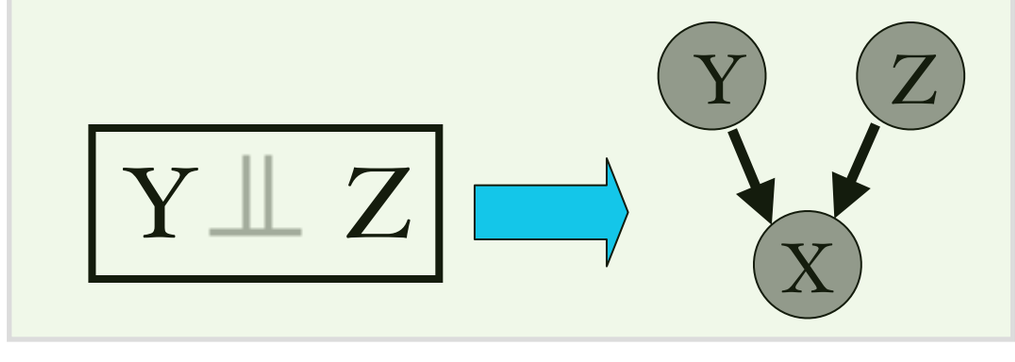
- Make use of conditional independence constraints
- Rely on causal Markov condition + faithfulness assumption



Constraint-Based Causal Discovery

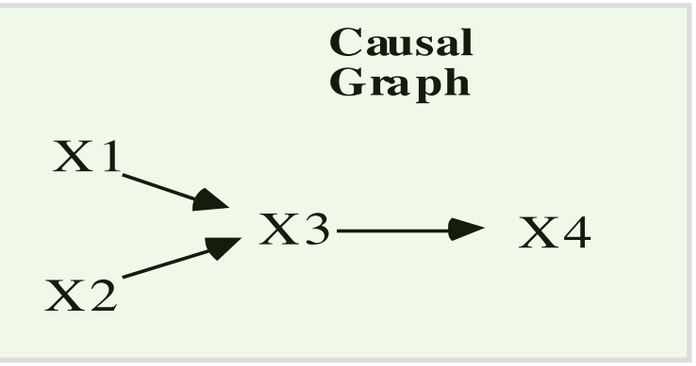
- (Conditional) independence constraints ⇒ candidate causal structures
  - Relies on causal Markov condition & faithfulness assumption
  - PC algorithm (Spirtes & Glymour, 1991)
    - *Step 1*: X and Y are adjacent iff they are dependent conditional on every subset of the remaining variables (SGS, 1990)
    - *Step 2*: Orientation propagation
- v-structure
- Markov equivalence class, represented by a pattern
  - same adjacencies; → if all agree on orientation; — if disagree





Example I

Step 1: finding skeleton



Independcies

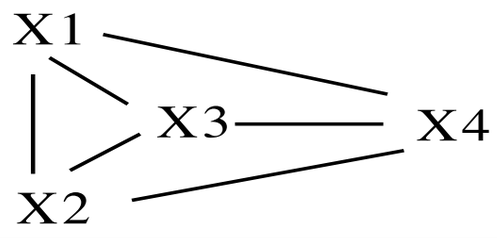
$X1 \perp\!\!\!\perp X2$

$X1 \perp\!\!\!\perp X4 \mid \{X3\}$

$X2 \perp\!\!\!\perp X4 \mid \{X3\}$

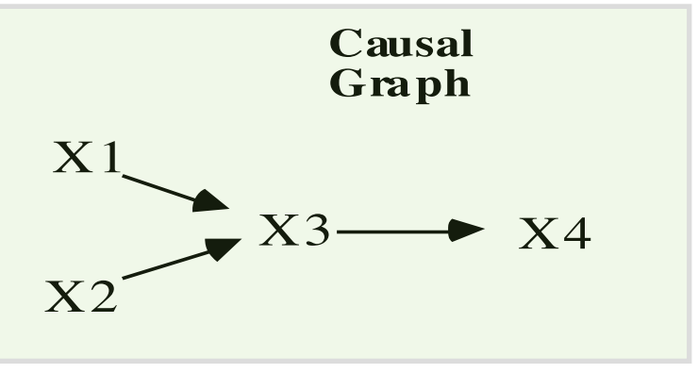
Step II: finding v-structure and doing orientation propagation

Begin with:



Example I

Step I: finding skeleton



Independcies

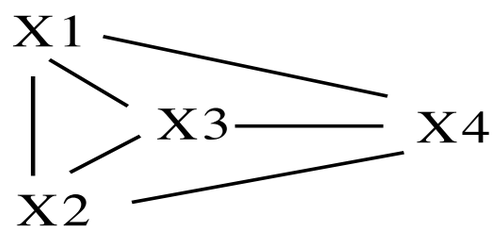
$X1 \perp\!\!\!\perp X2$

$X1 \perp\!\!\!\perp X4 \mid \{X3\}$

$X2 \perp\!\!\!\perp X4 \mid \{X3\}$

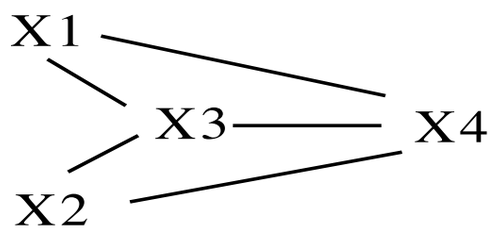
Step II: finding v-structure and doing orientation propagation

Begin with:



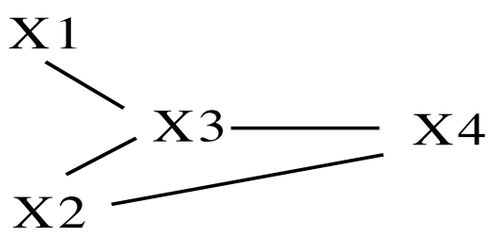
From

$X1 \perp\!\!\!\perp X2$



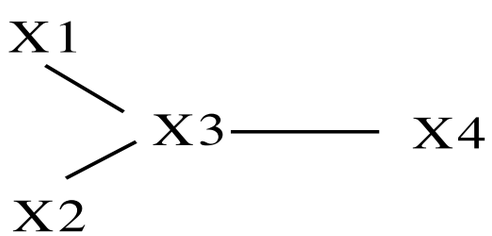
From

$X1 \perp\!\!\!\perp X4 \mid \{X3\}$

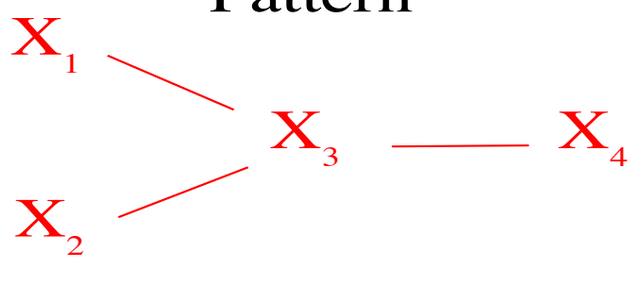


From

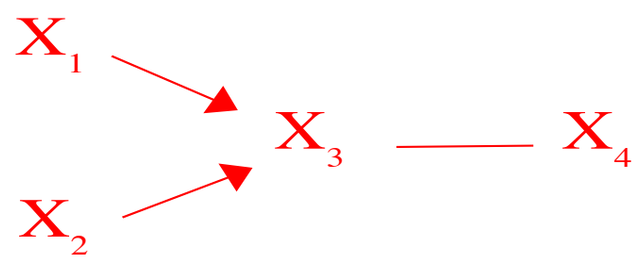
$X2 \perp\!\!\!\perp X4 \mid \{X3\}$

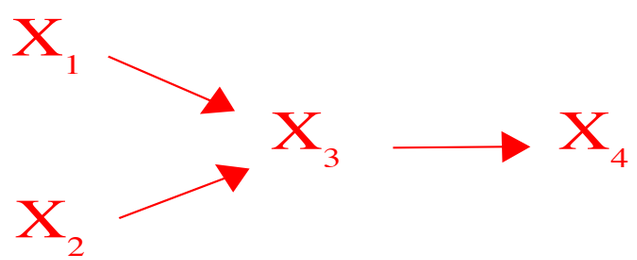


Pattern



$X_1 \perp\!\!\!\perp X_2$:





* (supplementary)

PC Algorithm

Test for (conditional) independence with an increased cardinality of the conditioning set

A.) Form the complete undirected graph C on the vertex set $\mathbf{V}$.

B.)

$n = 0$.

repeat

repeat

select an ordered pair of variables X and Y that are adjacent in C such that **Adjacencies**$(C,X)\backslash\{Y\}$ has cardinality greater than or equal to n, and a subset $\mathbf{S}$ of **Adjacencies**$(C,X)\backslash\{Y\}$ of cardinality n, and if X and Y are d-separated given $\mathbf{S}$ delete edge X - Y from C and record $\mathbf{S}$ in **Sepset**(X,Y) and **Sepset**(Y,X);

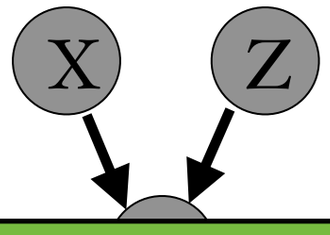
until all ordered pairs of adjacent variables X and Y such that **Adjacencies**$(C,X)\backslash\{Y\}$ has cardinality greater than or equal to n and all subsets $\mathbf{S}$ of **Adjacencies**$(C,X)\backslash\{Y\}$ of cardinality n have been tested for d-separation;

$n = n + 1$;

until for each ordered pair of adjacent vertices X, Y, **Adjacencies**$(C,X)\backslash\{Y\}$ is of cardinality less than n.

C.) For each triple of vertices X, Y, Z such that the pair X, Y and the pair Y, Z are each adjacent in C but the pair X, Z are not adjacent in C, orient X - Y - Z as X -> Y <- Z if and only if Y is not in **Sepset**(X,Z).

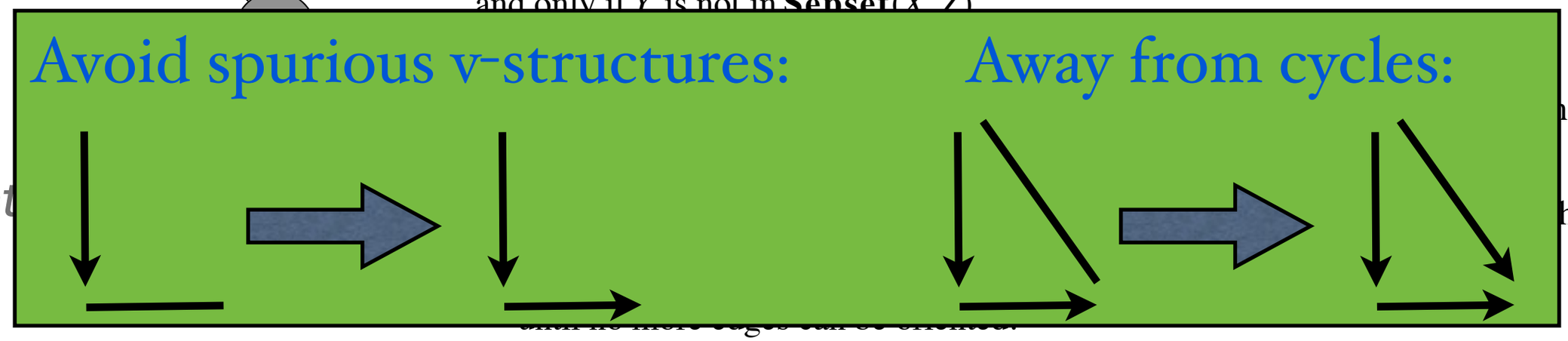
Finding V-structures



Orient

here is no

hen orient

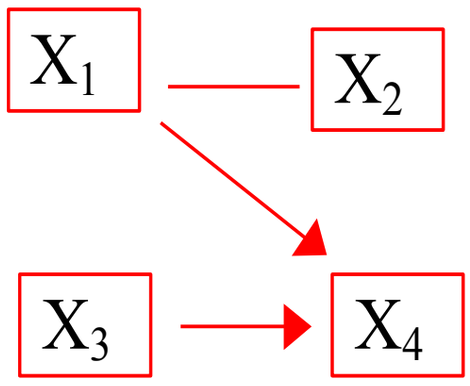


(Independence) Equivalent Classes: Patterns

- Two DAGs are (independence) equivalent if and only if they have the same skeletons and the same v-structures (Verma & Pearl, 1991)
- Patterns or CPDAG (Completed Partially Directed Acyclic Graph): graphical representation of (conditional) independence equivalence among models with no latent common causes (i.e., causally sufficient models)

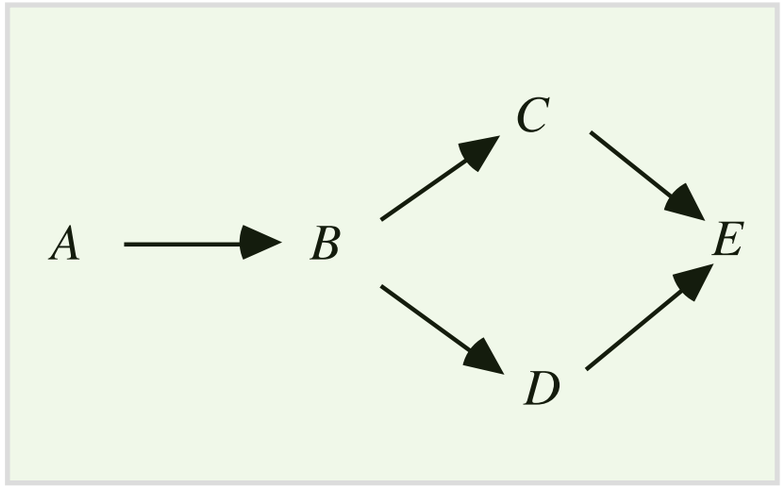
	Possible Edges
X_1 and X_2 are not adjacent in any member of the equivalent class	X_1 X_2
$X_1 \rightarrow X_2$ in some members of the equivalent class, and $X_1 \leftarrow X_2$ in some others	X_1 — X_2
$X_1 \rightarrow X_2$ in every member of the equivalent class	$X_1 \rightarrow X_2$

Example



How many DAGs in this class?

Demonstration with Tetrad

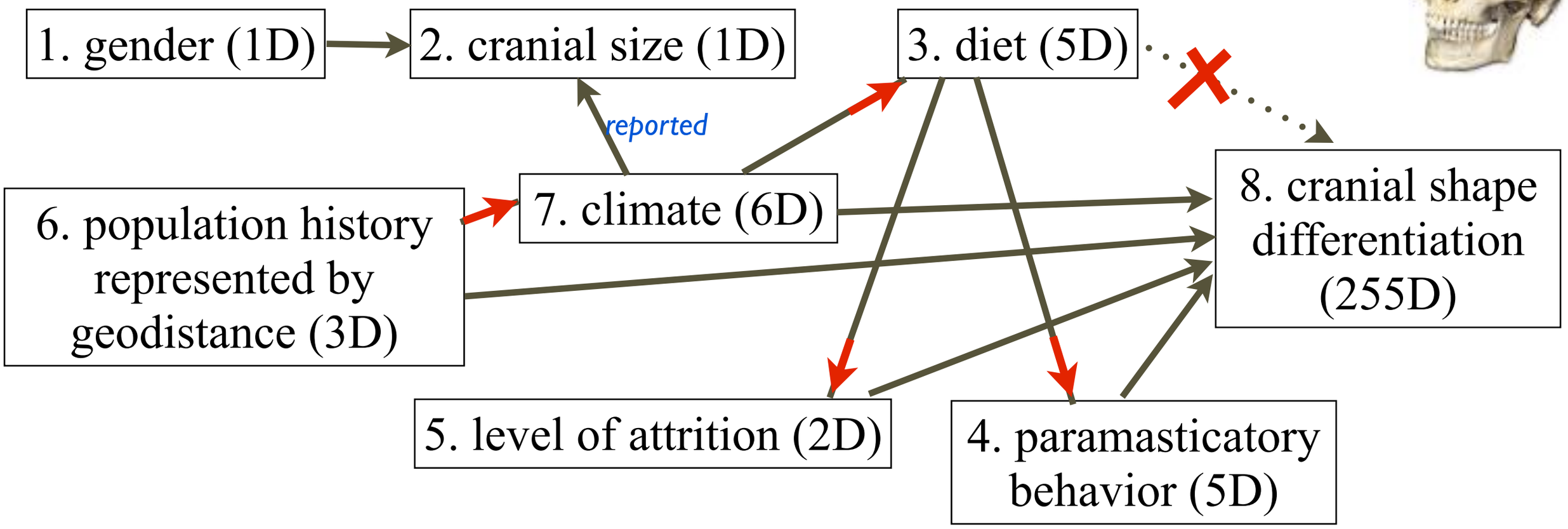


- To see the finite sample size effect, we generate linear-Gaussian data according to the graph with $T = 50$ & 1000

Example I: Result on the Archeology Data

Thanks to collaborator Marlijn Noback

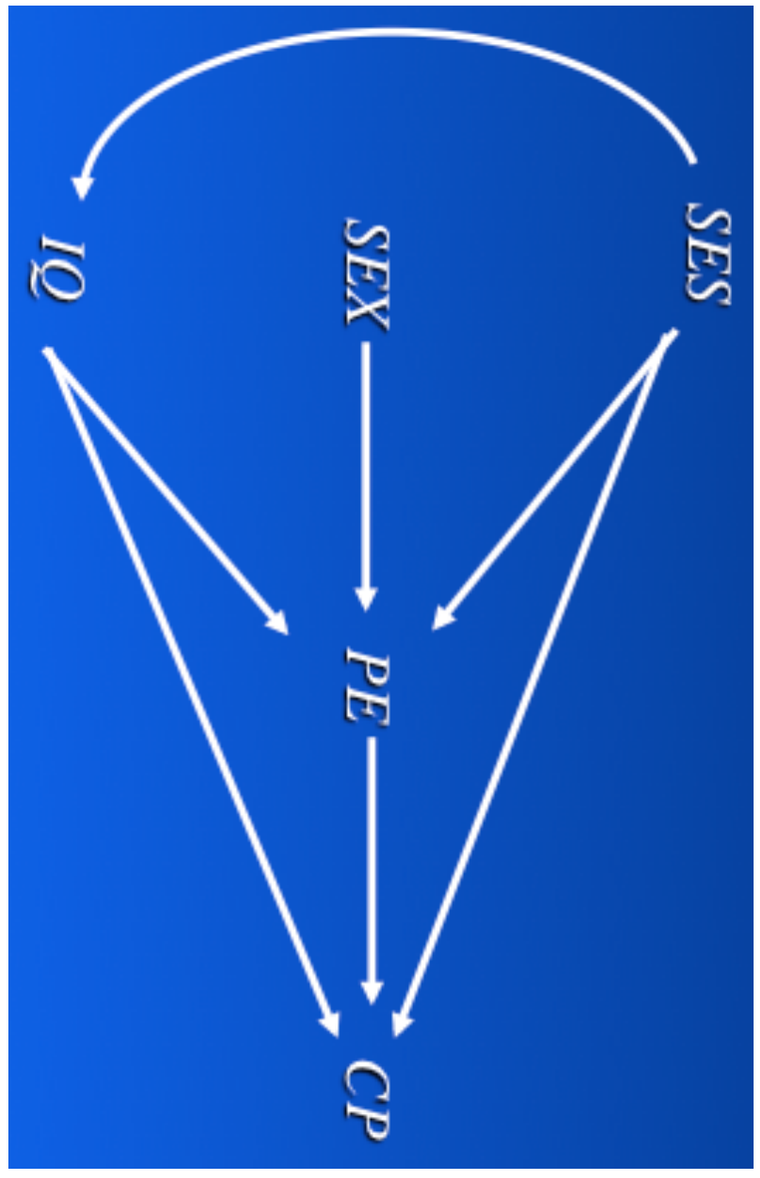
- 8 variables of 250 skeletons collected from different locations
- Different dimensions (from 1 to 255) with nonlinear dependence
- By PC algorithm + kernel-based conditional independence test (Zhang et al., 2011)



Example II: College Plans

Sewell and Shah (1968) studied five variables from a sample of 10,318 Wisconsin high school seniors.

SEX [male = 0, female = 1]
IQ = Intelligence Quotient [lowest = 0, highest = 3]
CP = college plans [yes = 0, no = 1]
PE = parental encouragement [low = 0, high = 1]
SES = socioeconomic status [lowest = 0, highest = 3]



Dealing with Confounders?

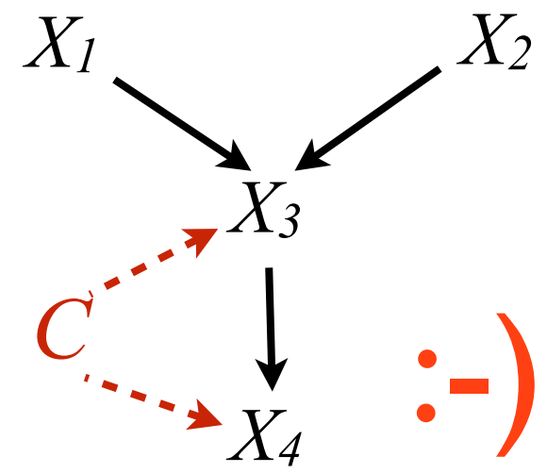
Example I

$X_1 \perp\!\!\!\perp X_2;$

$X_1 \perp\!\!\!\perp X_4 \,|\, X_3;$

$X_2 \perp\!\!\!\perp X_4 \,|\, X_3.$

Possible to have confounders behind X_3 and X_4?



E.g., X_1: Raining; X_3: wet ground; X_4: slippery.

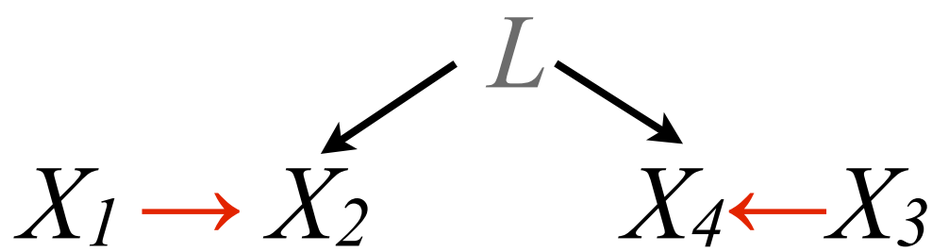
Example II

$X_1 \perp\!\!\!\perp X_3;$

$X_1 \perp\!\!\!\perp X_4;$

$X_2 \perp\!\!\!\perp X_3.$

Are there confounders behind X_2 and X_4?



E.g., X_1: I am not sick; X_2: I am in this lecture room; X_4: you are in this lecture room; X_3: you are not sick.

(See the FCI algorithm)

I know There Is No Confounder: Example

- In the 1970s, the Edison Electric Company in North Carolina was concerned about the effects on plant growth of acid rain produced by emissions from its electric generators.

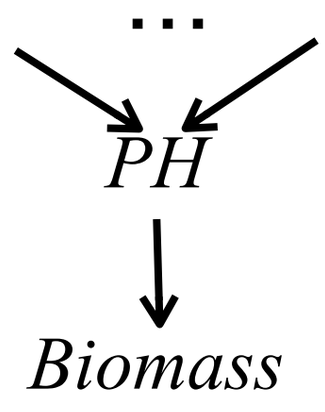
- The investigators chose samples from the Cape Fear estuary, where the Cape Fear River flows into the Atlantic Ocean.

- obtained 45 samples of Spartina grass up and down the estuary, and measured 13 variables in the samples, including **concentrations of various minerals, acidity (pH), salinity, and the outcome variable, the biomass of each sample**

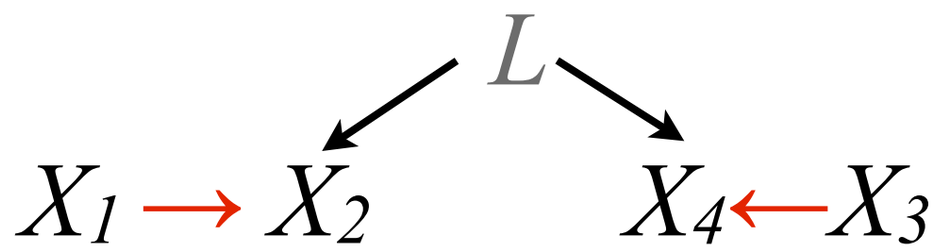
- The PC algorithm found that among **the measured variables the only *direct* cause of biomass was pH**.

- Y-structure: no confounder!

- Later verified by intervention-based analysis

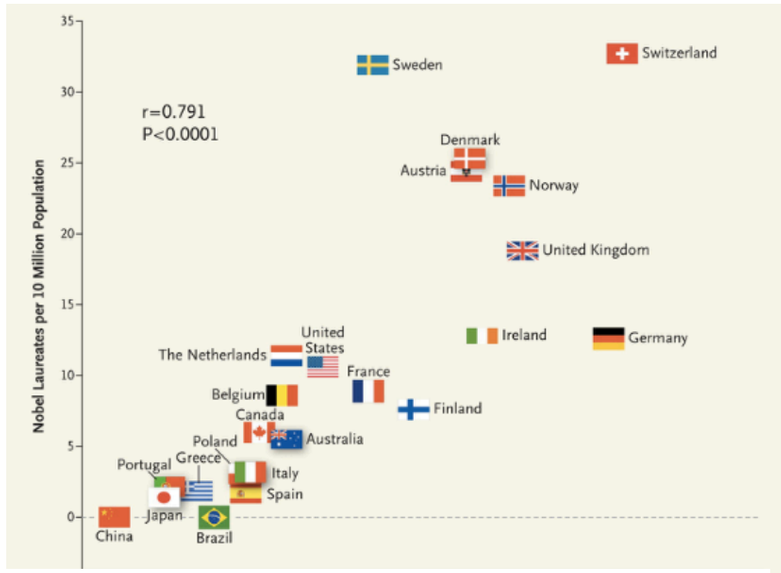


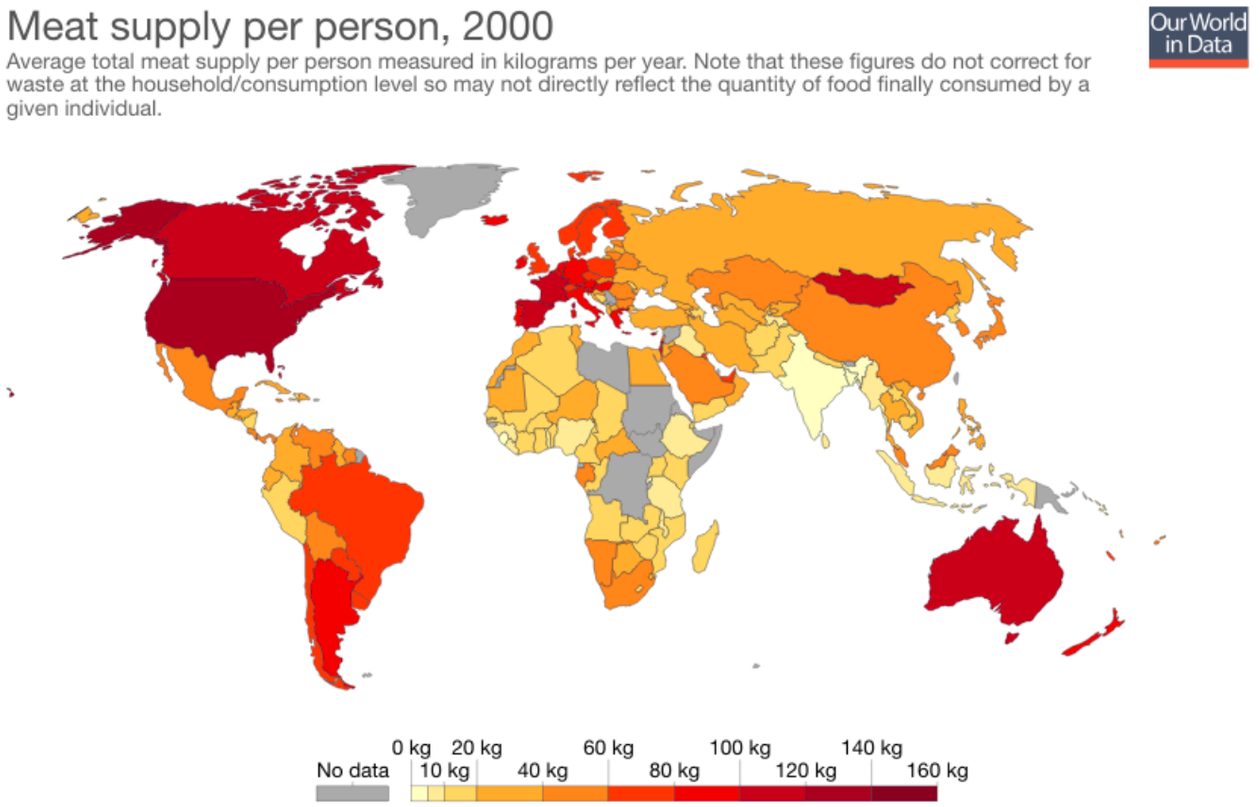
* I know There must Be Confounder



- X_1: I am not sick; X_2: I am in class; X_4: you are in class; X_3: you are not sick

- X_1: European/South American country; X_2: leading in science; X_4: Chocolate consumption; X_3: meat supply per person



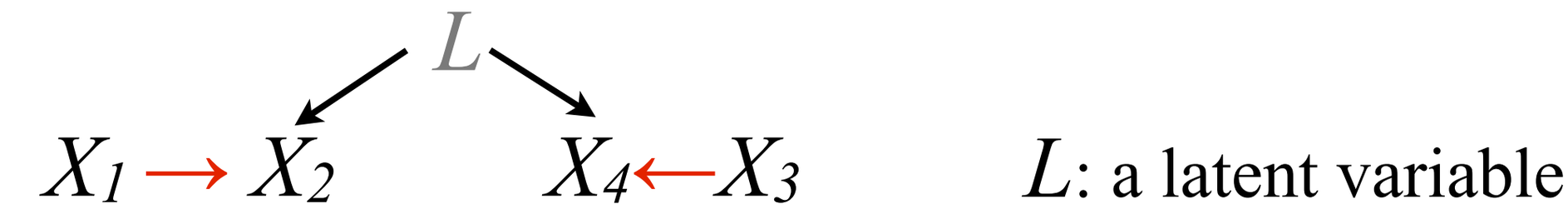


Example II...

$$X_1 \perp\!\!\!\perp X_3;$$

$$X_1 \perp\!\!\!\perp X_4;$$

$$X_2 \perp\!\!\!\perp X_3.$$



- There must exist some confounder for X_2 and X_4.

- In the presence of latent variables, the causal process over measured variables **O** is not necessarily a DAG. How can we represent (independence) equivalence classes over **O** ?

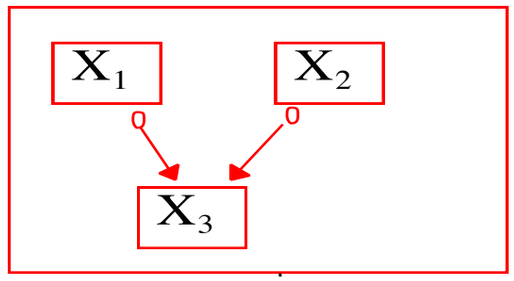
Remember the Output of PC?
(Independence) Equivalent Classes: Patterns

- Two DAGs are (independence) equivalent if and only if they have the same skeletons and the same v-structures (Verma & Pearl, 1991)
- Patterns or CPDAG (Completed Partially Directed Acyclic Graph): graphical representation of (conditional) independence equivalence among models with no latent common causes (i.e., causally sufficient models)

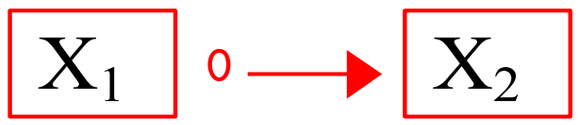
	Possible Edges	Example
X_1 and X_2 are not adjacent in any member of the equivalent class	X_1 X_2	X_1 — X_2
$X_1 \rightarrow X_2$ in some members of the equivalent class, and $X_1 \leftarrow X_2$ in some others	X_1 — X_2	$X_3 \rightarrow X_4$
$X_1 \rightarrow X_2$ in every member of the equivalent class	$X_1 \rightarrow X_2$	

How many DAGs in this class?

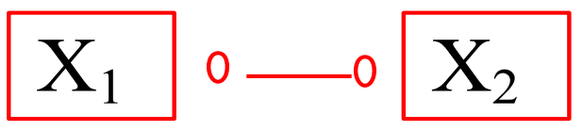
PAGs: What Edges Mean?



X_1 and X_2 are not adjacent

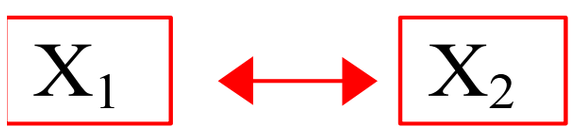


X_2 is not an ancestor of X_1



No set d-separates X_2 and X_1

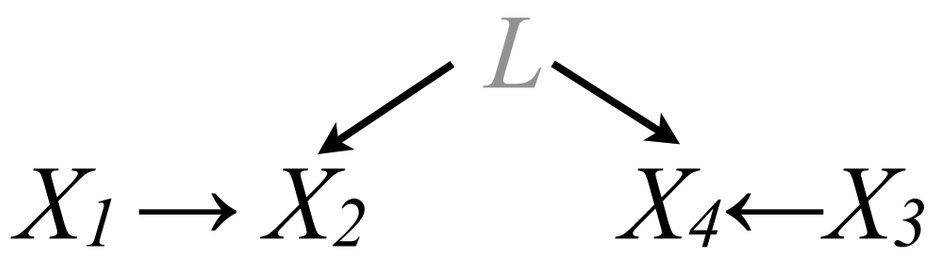
X_1 is a cause of X_2



There is a latent common cause of X_1 and X_2

FCI (Fast Causal Inference) Allows Confounders

- Assume the distribution over measured variables **O** is the marginal of a distribution satisfying the Markov and faithfulness conditions for the true graph
- Results represented by PAGs (Partial Ancestral Graphs)



What's FCI's output?

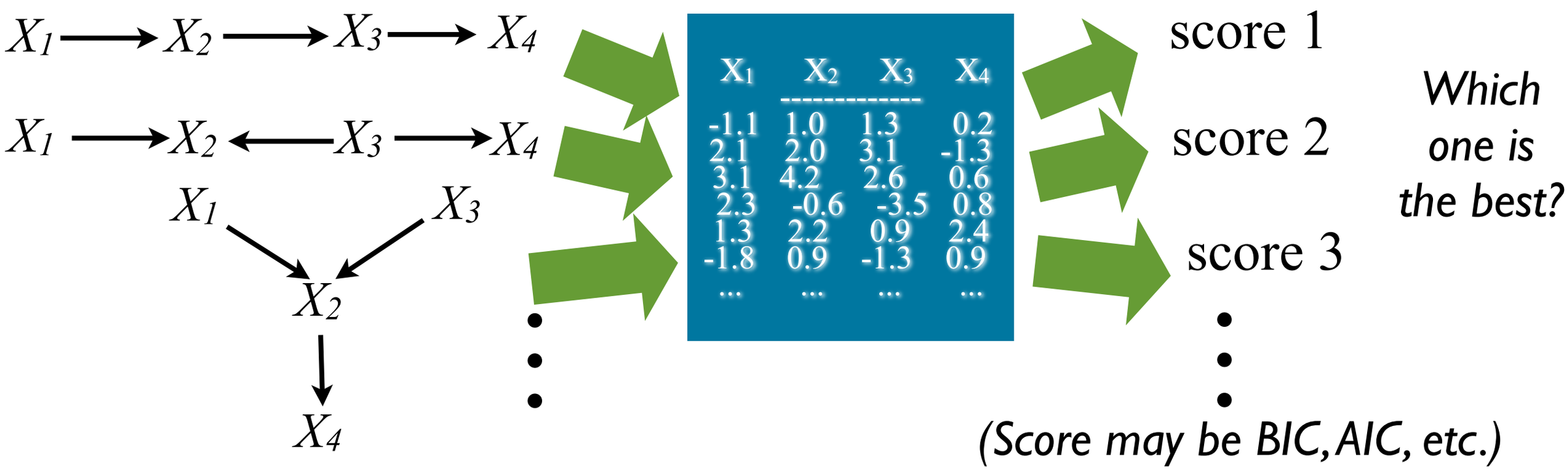
Data available in
'Illust_FCI_4variables.txt'

Spirtes et al., Causal inference in the presence of latent variables and selection bias, 1997

Constraint-Based vs. Score-Based

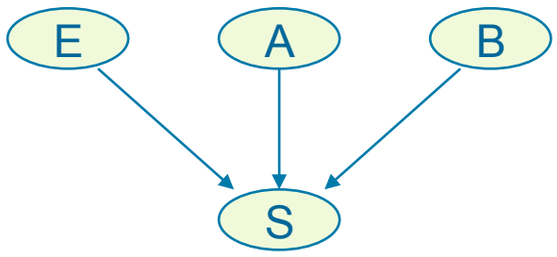
- Constraint-based methods

- Score-based methods

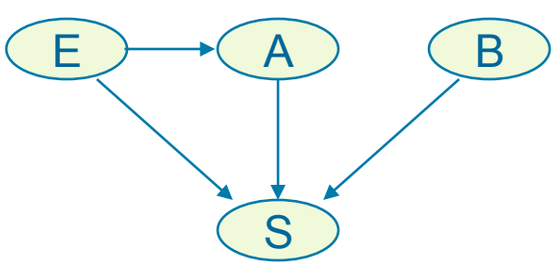


Why Is It Possible?

"True" structure



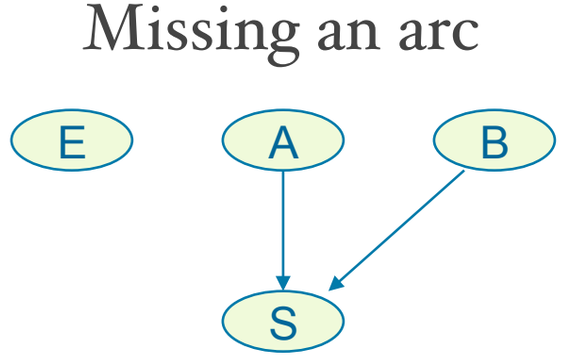
Adding an arc



- Increases the number of parameters to be fitted;

Wrong assumptions about causality and domain structure

Missing an arc

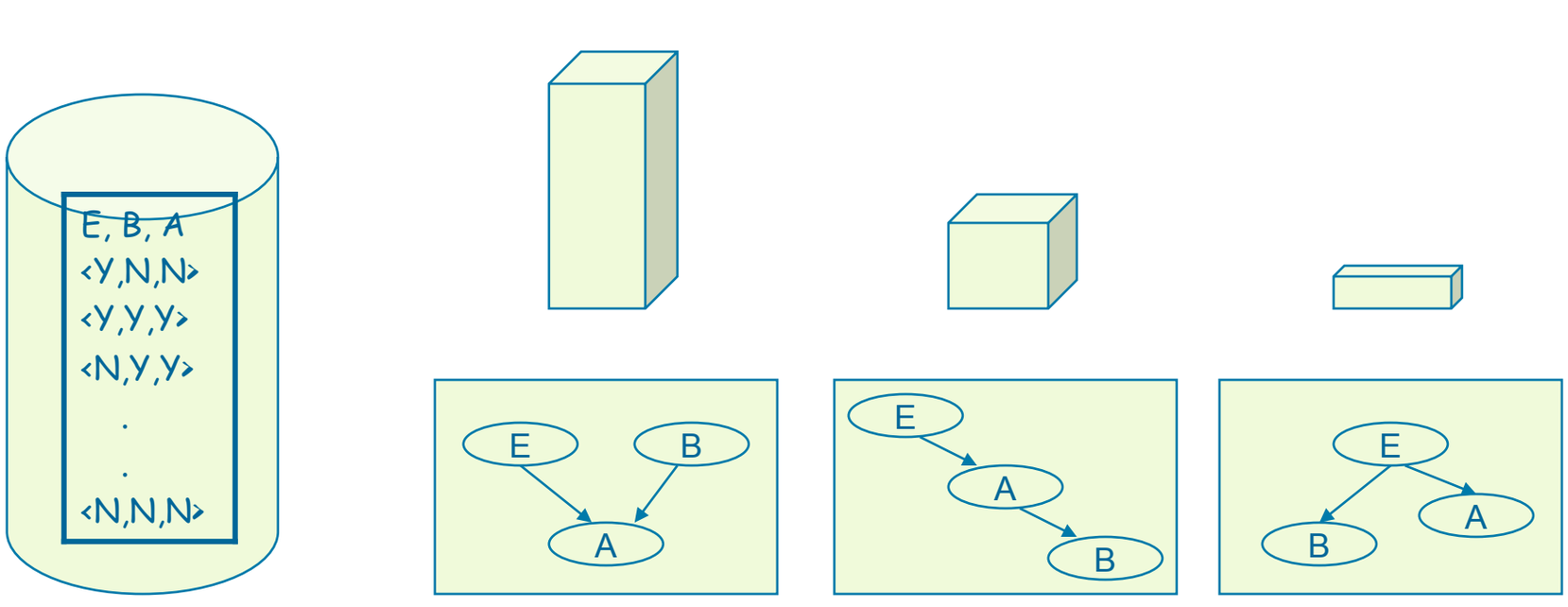


- Cannot be compensated by accurate fitting of parameters;

Also misses causality and domain structure

Score-Based Learning

- Score: evaluates how well a structure matches the data + how simple the structure is



- Search for a structure that maximizes (or minimizes) the score

GES (Greedy Equivalence Search): Score Function

- Assumptions: The score is
  - score equivalent (i.e., assigning the same score to equivalent DAGs)
  - locally consistent: score of a DAG increases (decreases) when adding any edge that eliminates a false (true) independence constraint
  - decomposable: $Score(\mathcal{G}, \mathbf{D}) = \sum_{i=1}^{n} Score(X_i, \mathbf{Pa}_i^{\mathcal{G}})$
- E.g., BIC: $S_B(\mathcal{G}, \mathbf{D}) = \log p(\mathbf{D}|\hat{\boldsymbol{\theta}}, \mathcal{G}^h) - \frac{d}{2} \log m$

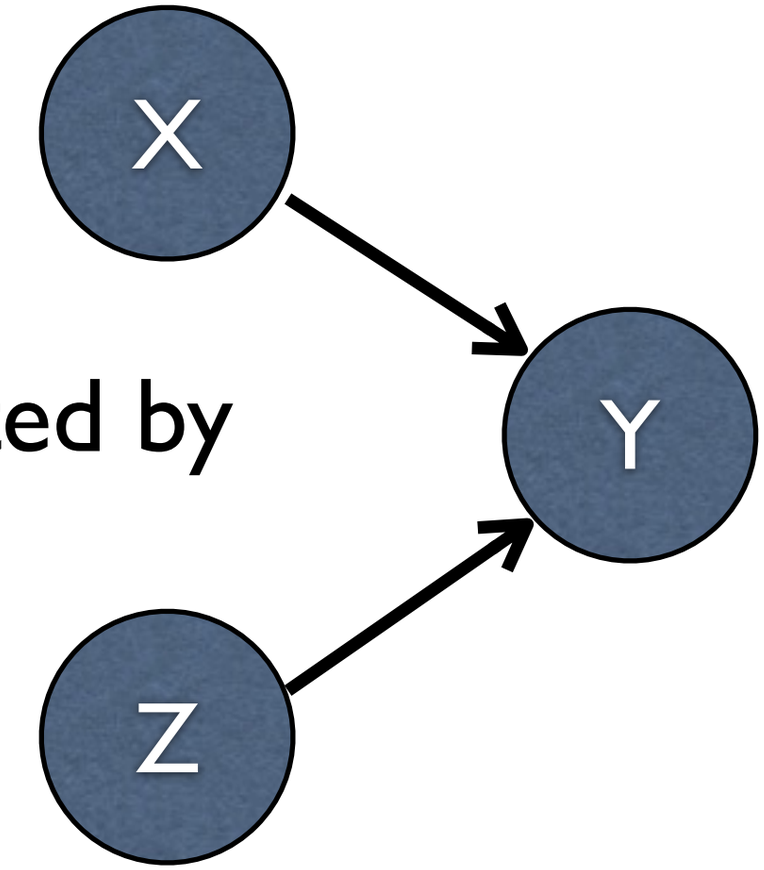
Chickering, Optimal Structure Identification With Greedy Search, Journal of Machine Learning Research, 2002

GES: Search Procedure

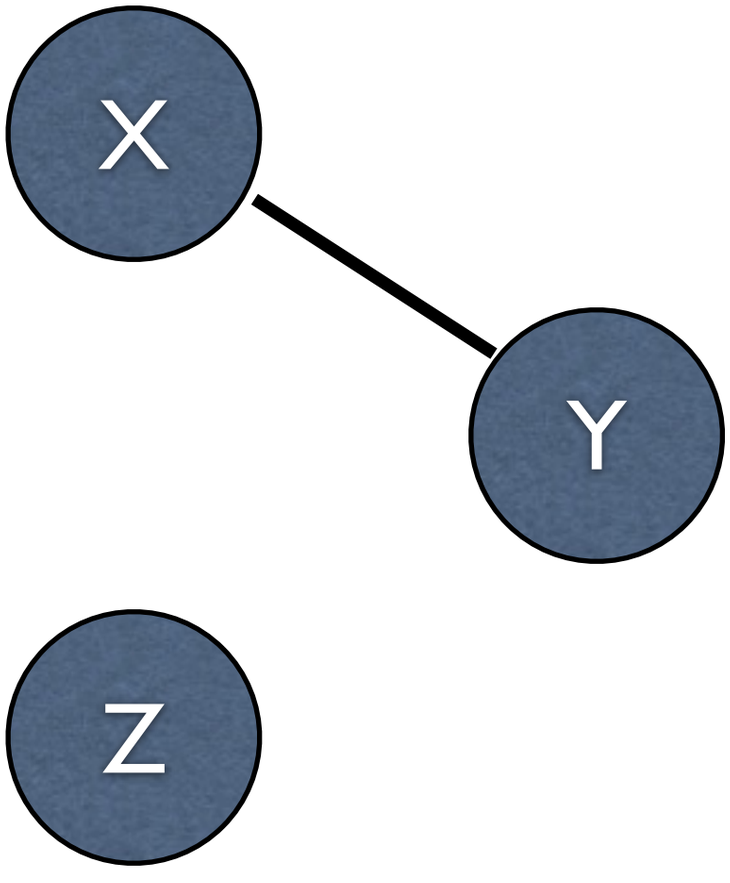
- Performs forward (addition) / backward (deletion) equivalence search through the space of DAG equivalence classes
  - Forward Greedy Search (FGS)
    - Start from some (sparse) pattern (usually the empty graph)
    - Evaluate all possible patterns with one more adjacency that entail strictly fewer CI statements than the current pattern
    - Move to the one that increases the score most
    - Iterate until a local maximum
  - Backward Greedy Search (BGS)
    - Start from the output of Stage (1)
    - Evaluate all possible patterns with one fewer adjacency that entail strictly more CI statements than the current pattern
    - Move to the one that increases the score most
    - Iterate until a local maximum

GES

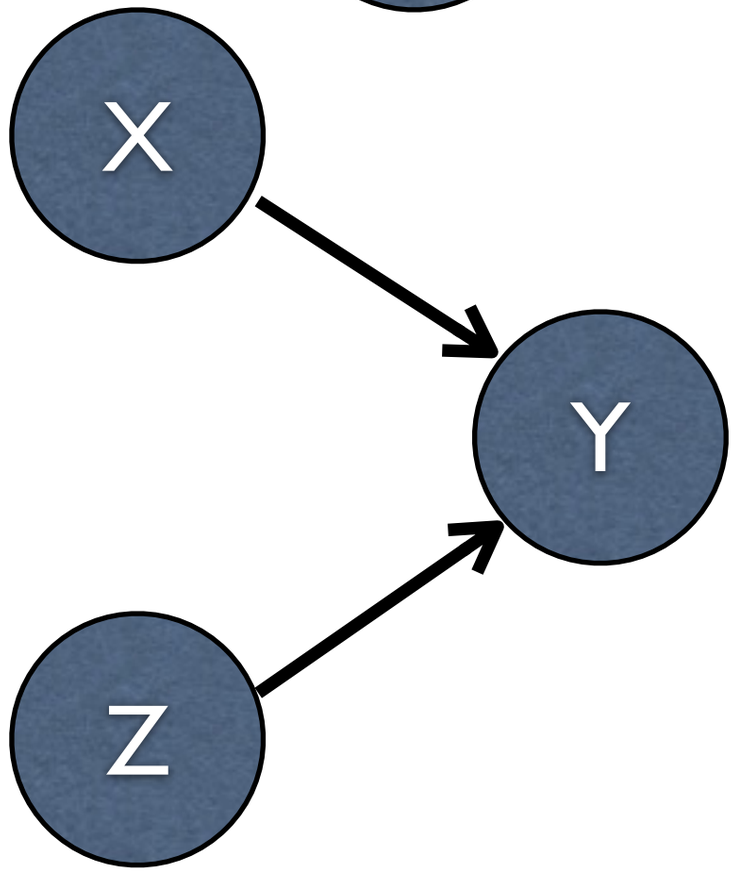
Suppose data were generated by



(1)

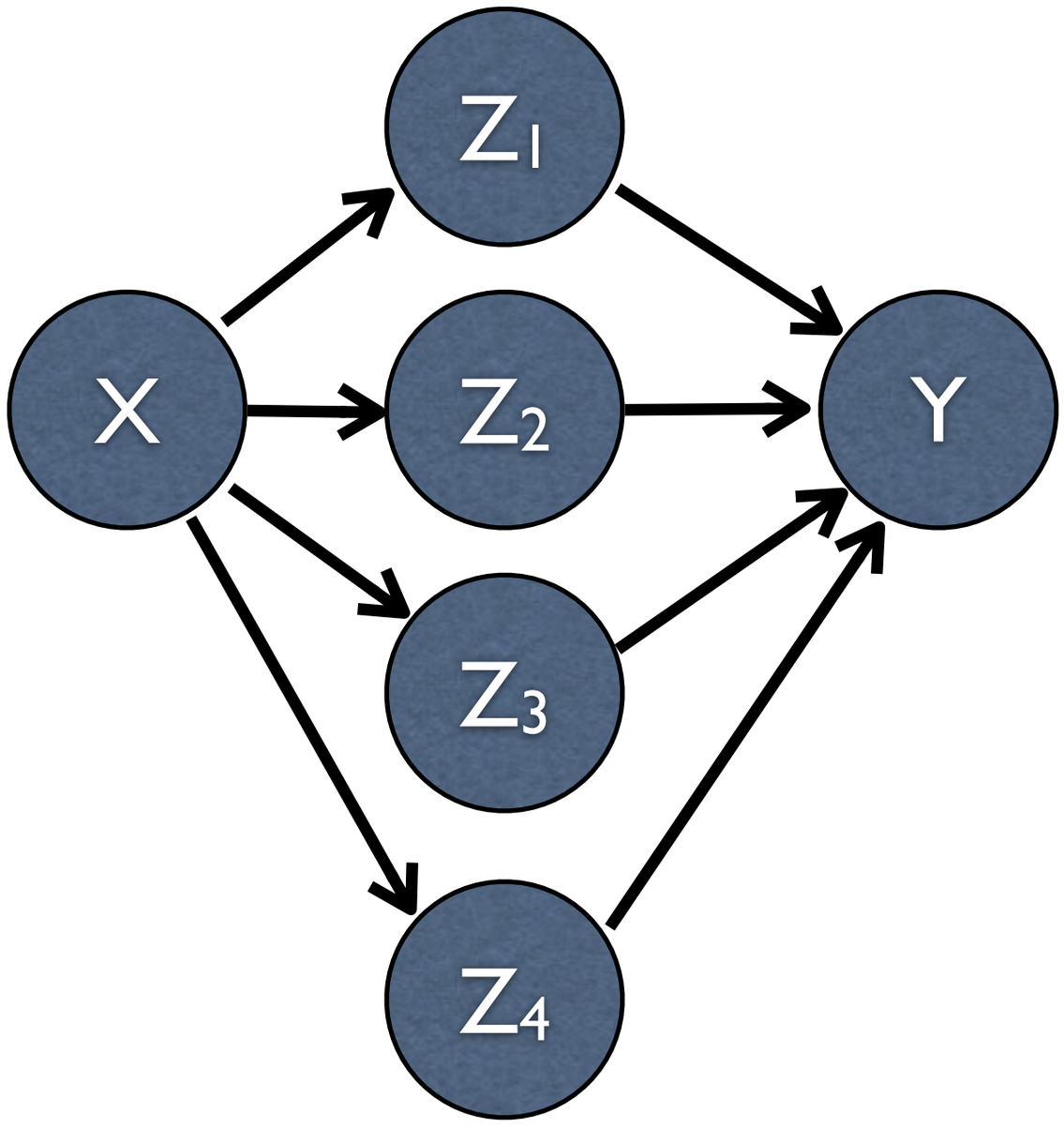


(2)



GES

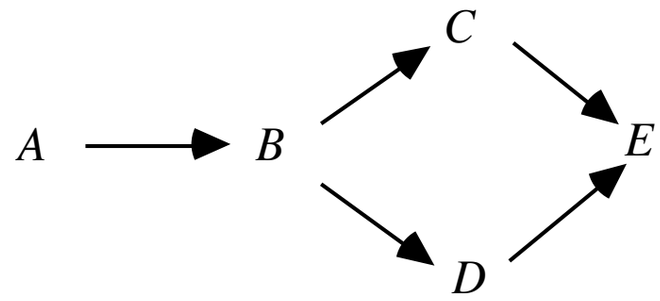
Suppose data were generated by



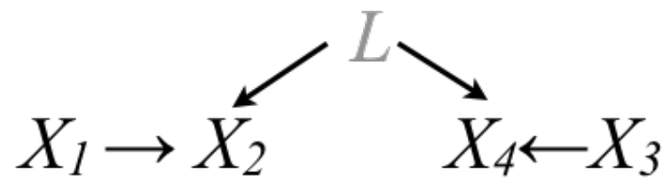
Imagine the GES procedure...

Demonstrations with Tetrad

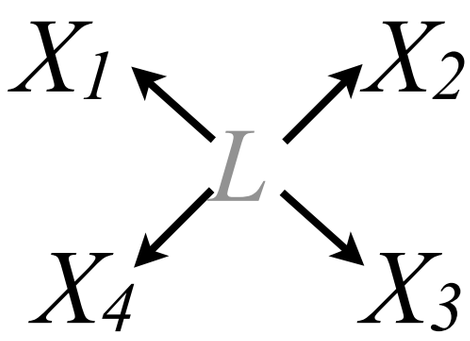
- 1. sample size effect (T = 1000 & 50)

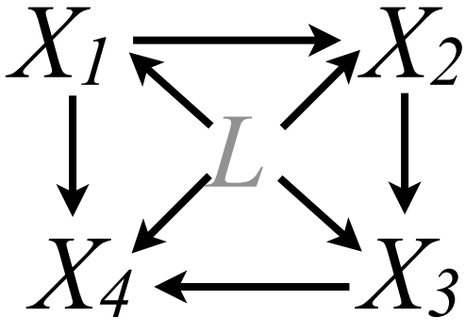


- 2. FCI (simple structure)



- 3 & 4. FCI (more complex structure)



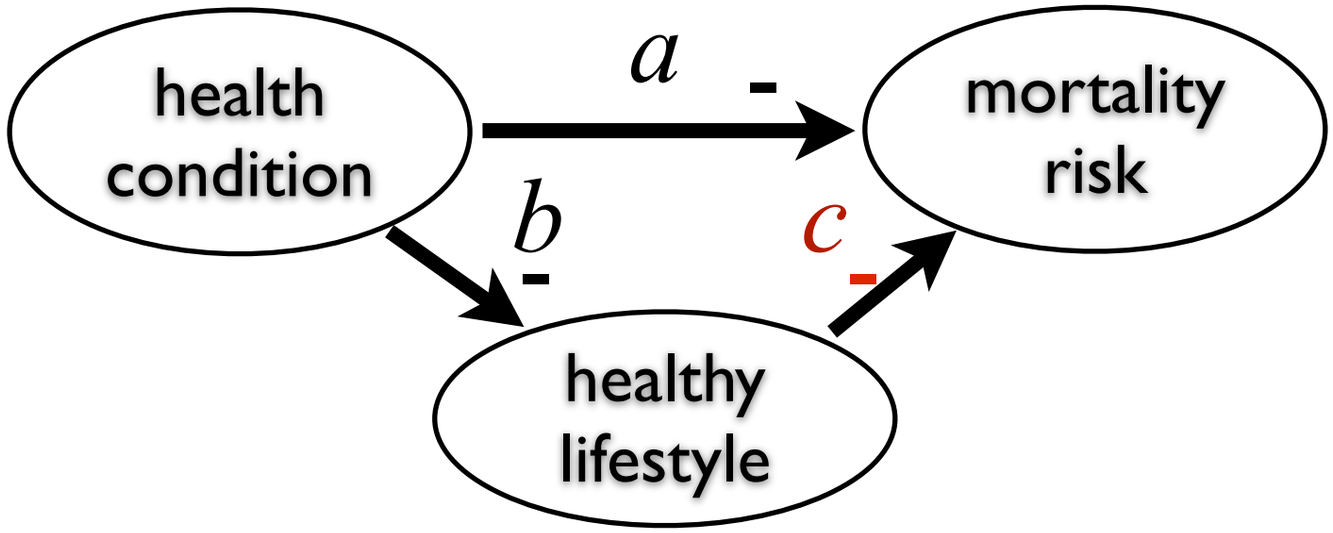


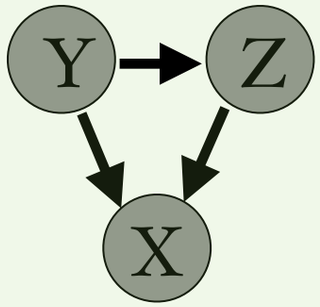
- 5. LiNGAM

$$X_1 \rightarrow X_2 \rightarrow X_3$$

On the Faithfulness Assumption

- One might find independence between health condition & risk of mortality. Why?





Possible to have $Y \perp\!\!\!\perp Z \mid X$?

- E.g., if $a=-bc$, then *health_condition* $\perp\!\!\!\perp$ *mortality_risk*, which cannot by seen from the graph!
- Faithfulness assumption eliminates this possibility!
  - Weak or strong?
  - Possible to be avoided?

Summary: The PC Algorithm

- "Process independence" implied by causal models
- Causal Markov condition
- Faithfulness Assumption
- Relating conditional independence relations to properties of causal DAG
- The PC algorithm?
- What if there may exist confounders?