*Probabilistic Graphical Models*

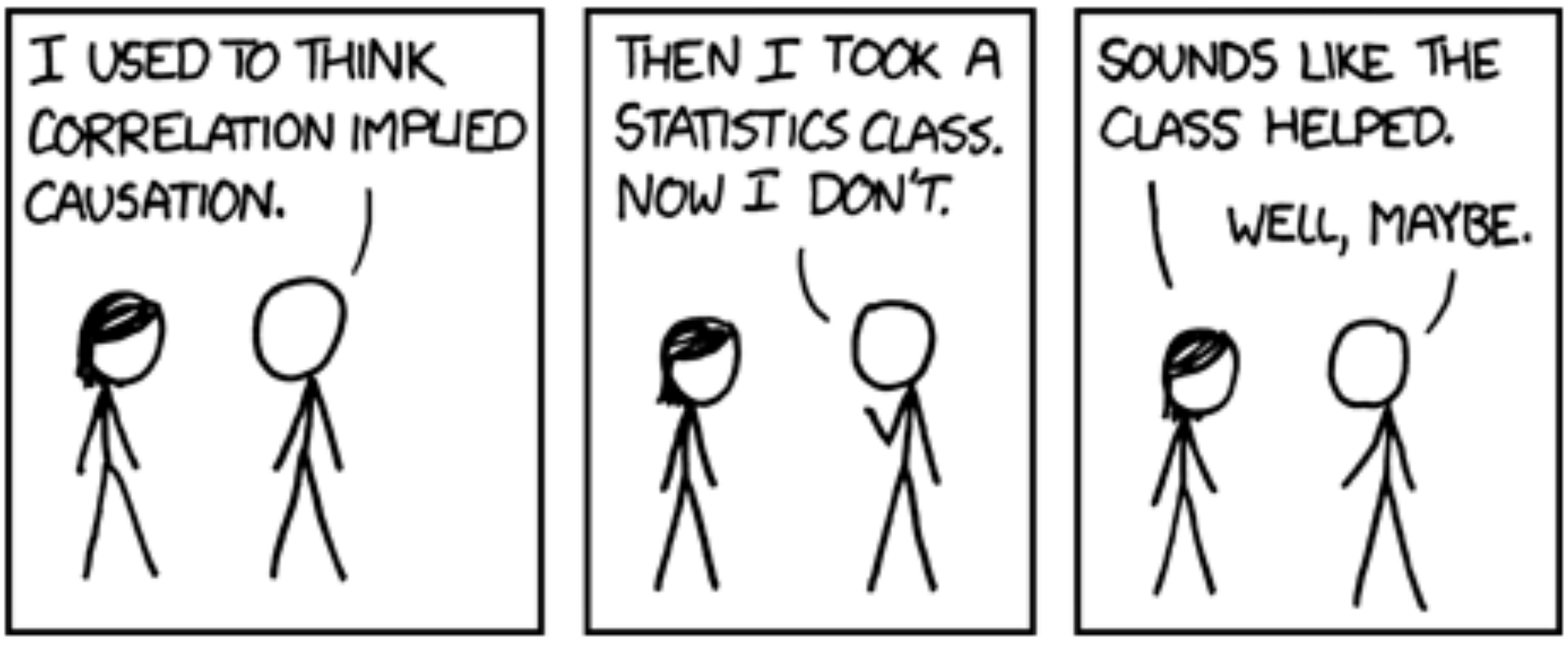
# Causal Discovery and Inference

Kun Zhang

Department of philosophy
CMU

# Causality vs. Dependence

- Causality ➜ dependence ! Dependence ➜ causality




(http://imgs.xkcd.com/comics/correlation.png)

X and Y are associated iff

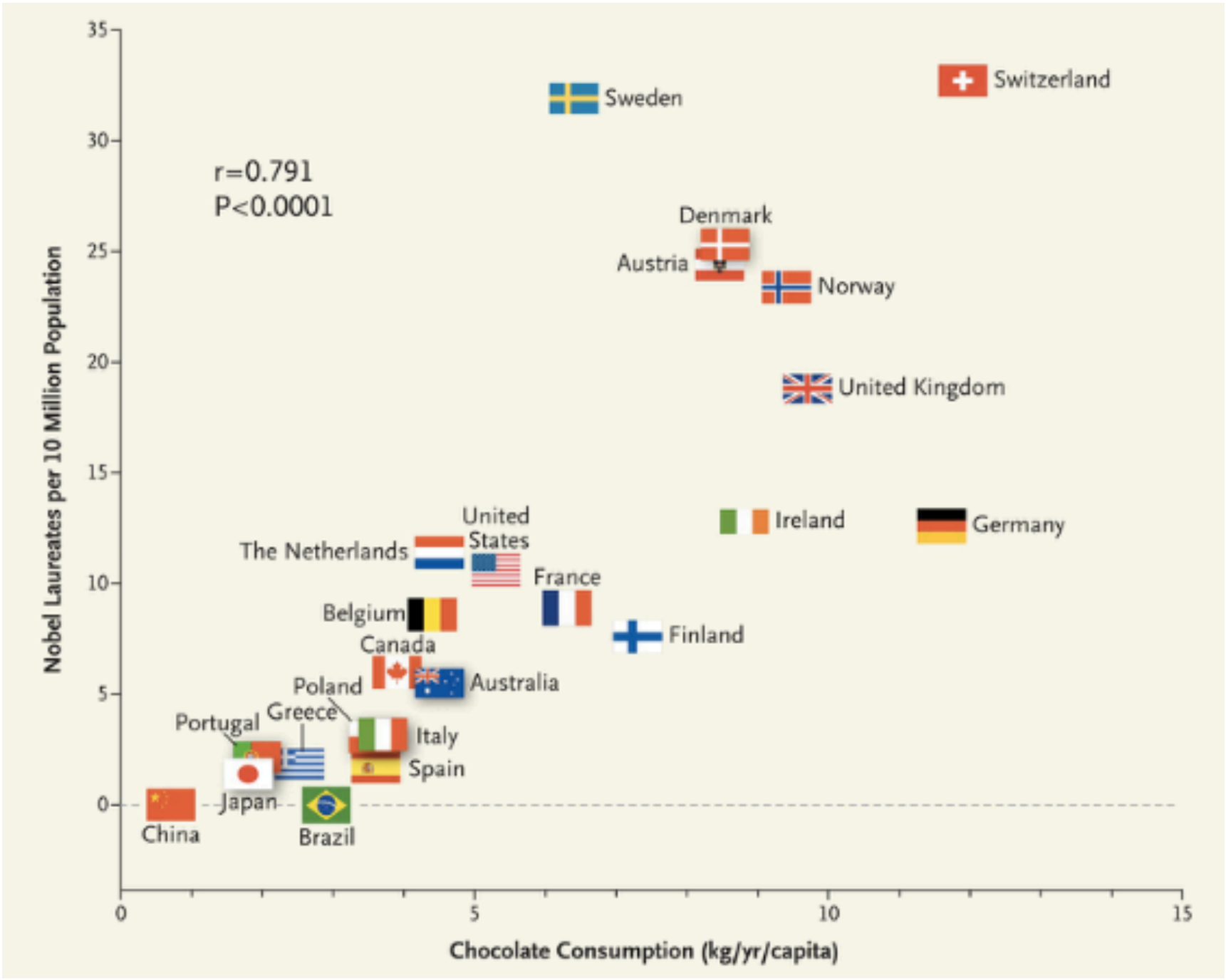
$\exists x_1 \neq x_2 \; P(Y|X=x_1) \neq P(Y|X=x_2)$

X is a cause of Y iff

$\exists x_1 \neq x_2 \; P(Y|do\ (X=x_1)) \neq P(Y|do\ (X=x_2))$

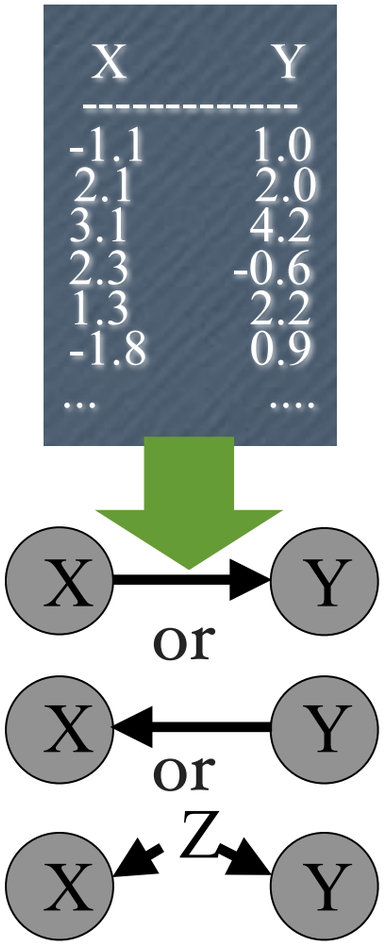
# Causal Discovery from Data: Example

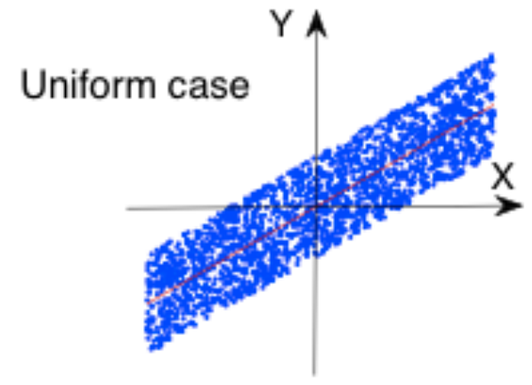
# Causal Discovery from Data: Example



# Outline

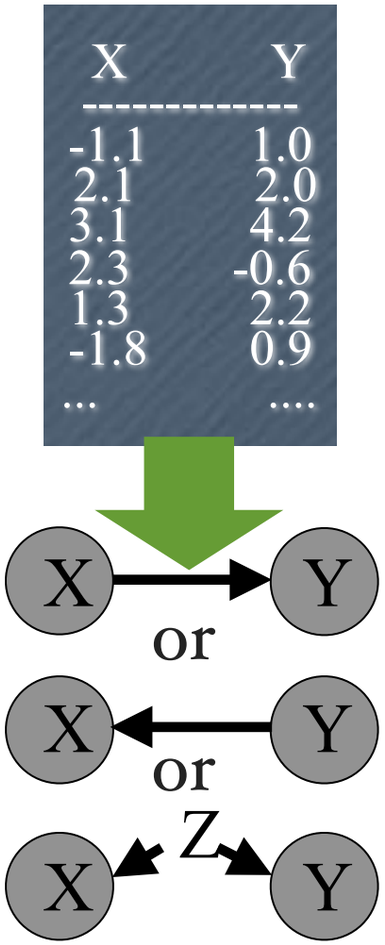
- Causal thinking
- Identification of causal effects
- Causal discovery
  - Constraint-based approach
  - Non-Gaussian or nonlinear methods
  - Extensions

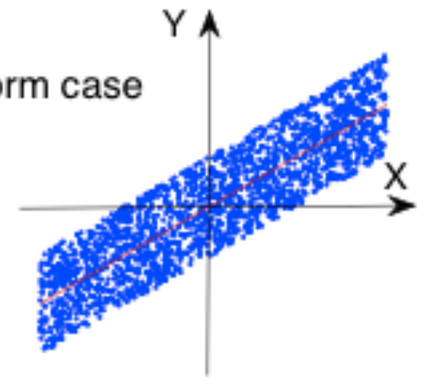



# Outline

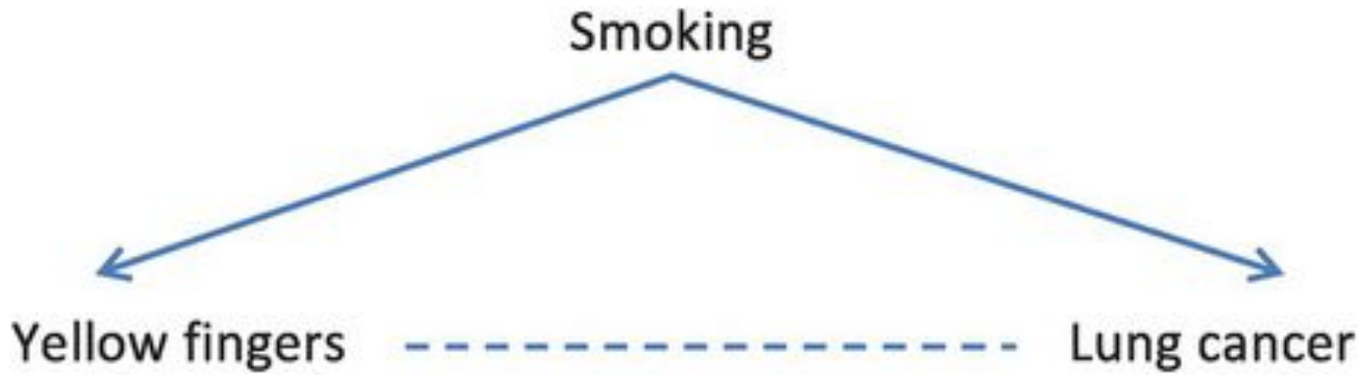
- **Causal thinking**
- Identification of causal effects
- Causal discovery
  - Constraint-based approach
  - Non-Gaussian or nonlinear methods
  - Extensions




# Causal Thinking (1)

- Dependence vs. causality



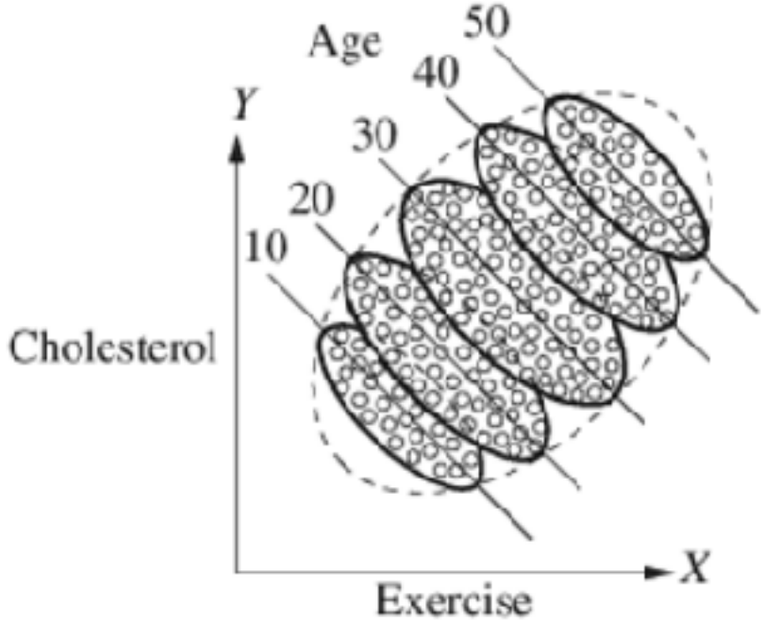

- Simpson's paradox

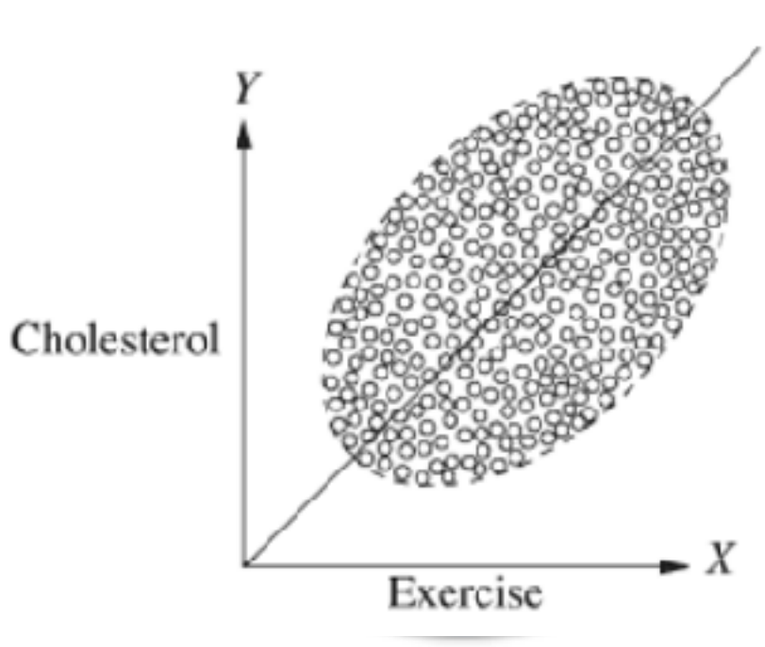
- "Strange" dependence

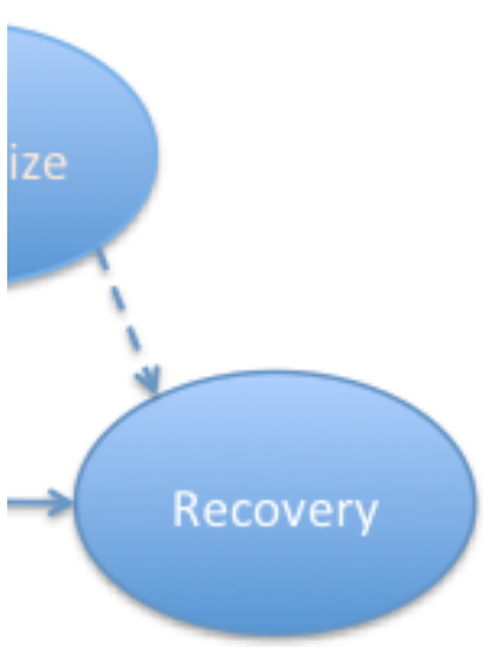
# Causal Thinking (2)

- Dependence vs. causality

- Simpson's paradox



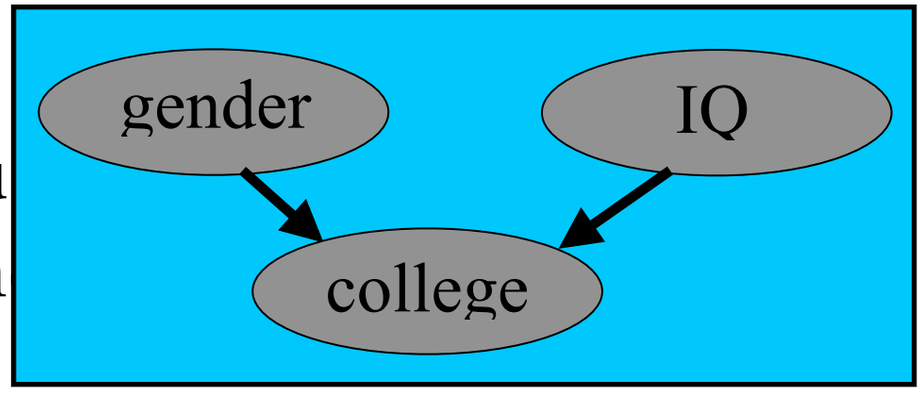

- "Strange" dependence

# Causal Thinking (3)

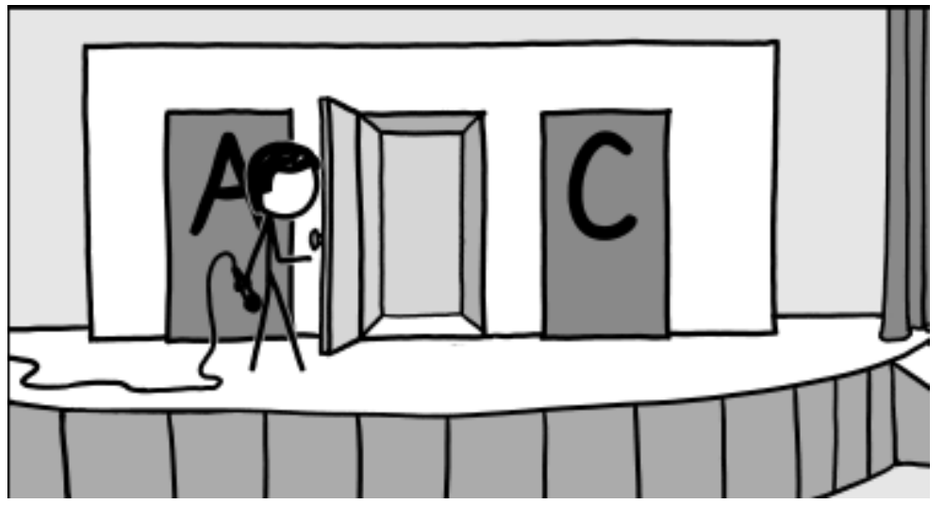
- Dependence vs. causality

- Simpson's paradox

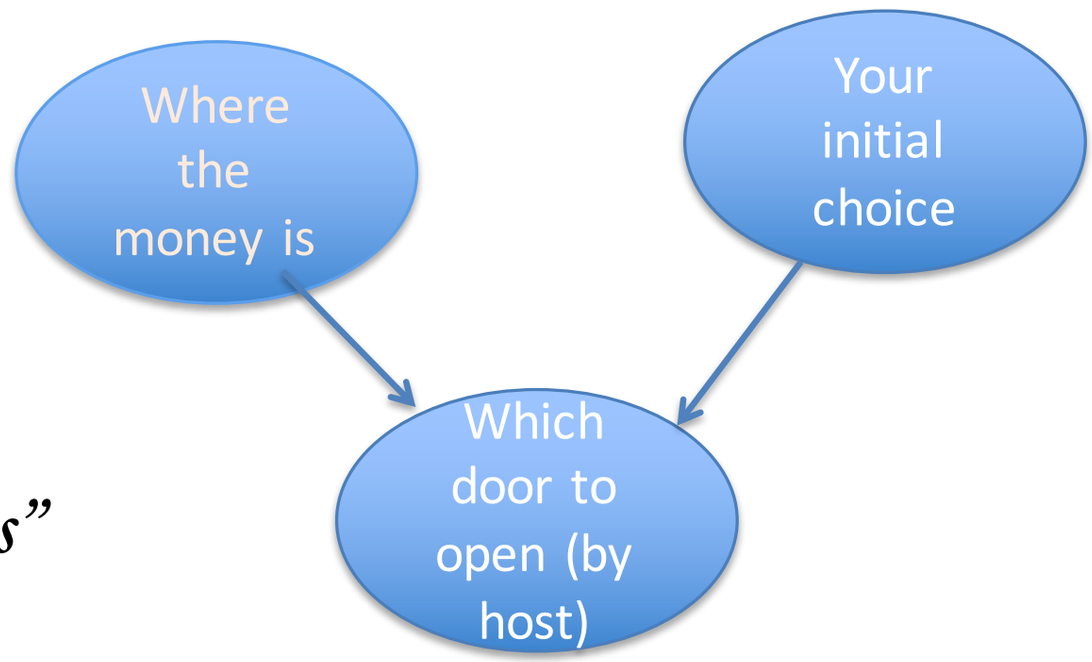
- "Stranger" dependence
  - Let's go back 50 years; maybe you students are smarter than male on



# Question 3
# Monty Hall Problem



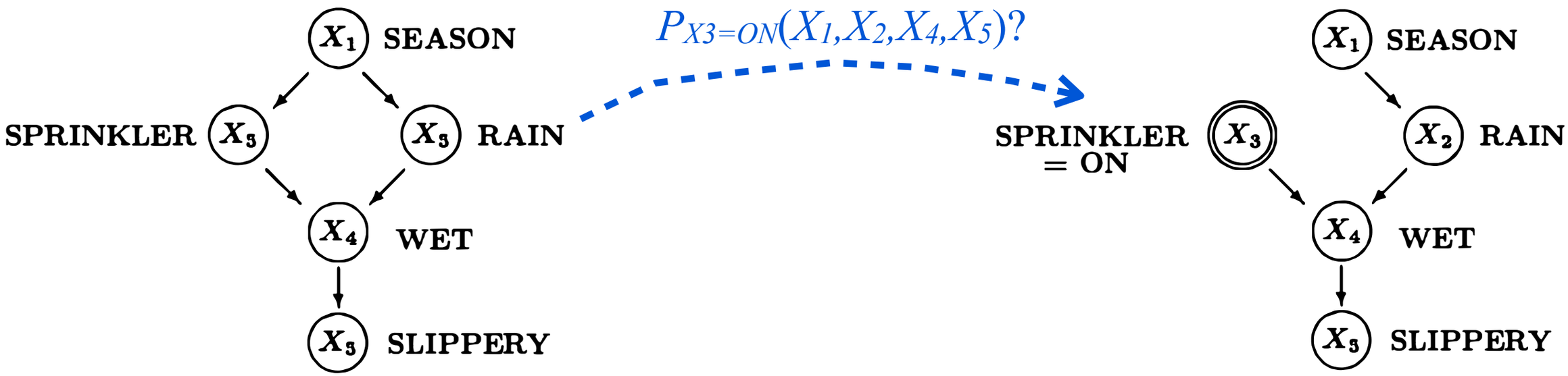

- You are a game show contestant. Before the game begins, the host, Monty Hall, has placed $1,000 dollars behind one of three doors. Nothing is behind the other two doors. The game is played as followed. You, the contestant, choose one of the doors, say, door A. Then Monty opens a door that is not the door you chose and does not have the money behind it, say B. If you want to maximize the expected profit, which door will you finally choose?
  - A
  - C



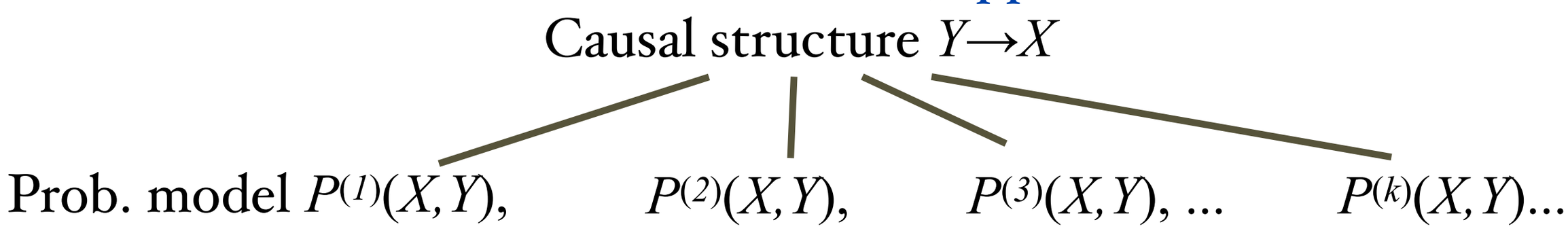

*Excerpt from "The Mind's Arrows"*

# Why Causal Models?

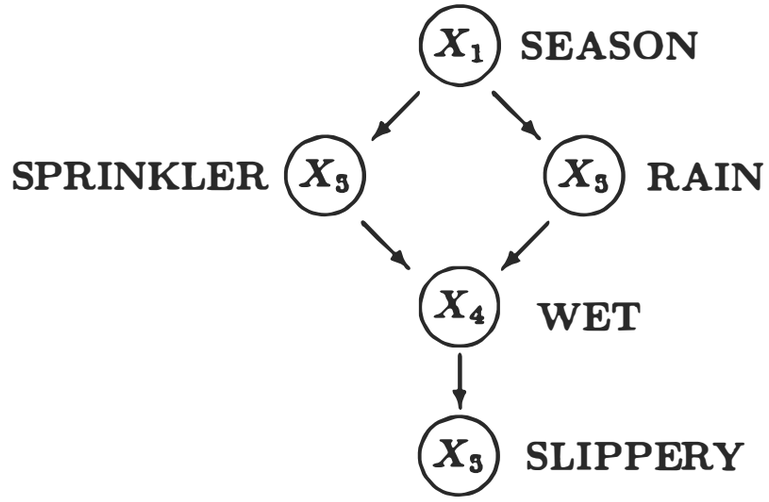
- Infer effect of interventions:




- Causal model: compact description of the properties of the joint distribution
- Derived from one or few distributions; applied to other scenarios



# Conditioning, Manipulating, and Counterfactual Thinking




- Three questions:
  - **Prediction**: Would the pavement be slippery if we *find* the sprinkler off?

$$P(Slippery \mid Sprinkler{=}off)$$

  - **Intervention**: Would the pavement be slippery if we *make sure* that the sprinkler is off?

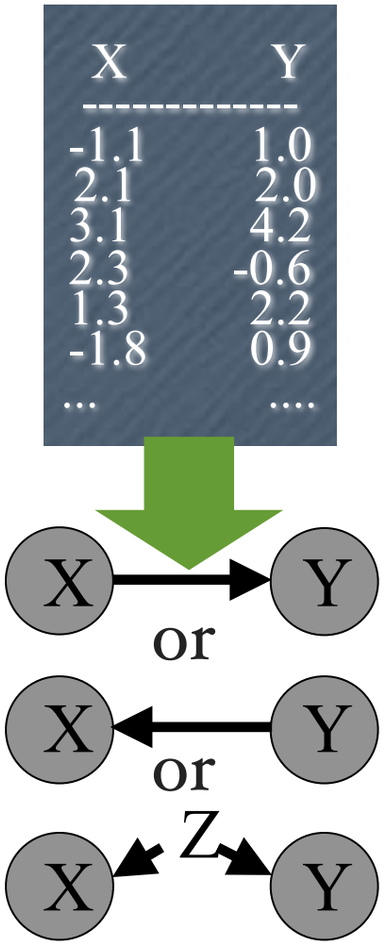
$$P(Slippery \mid do(Sprinkler{=}off))$$

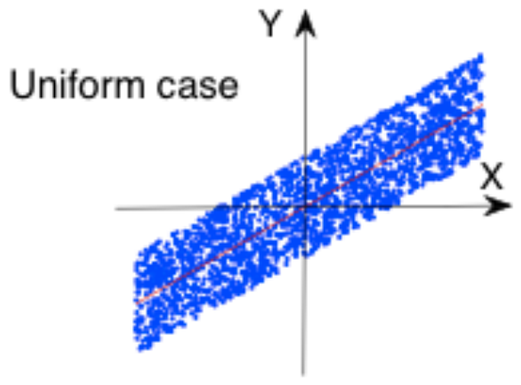
  - **Counterfactual**: Would the pavement be slippery *had* the sprinkler been off, *given that the pavement is in fact not slippery and the sprinkler is on*?

$$P(Slippery_{\,Sprinker=off} \mid Sprinkler = on, Slippery = no)$$

# Outline

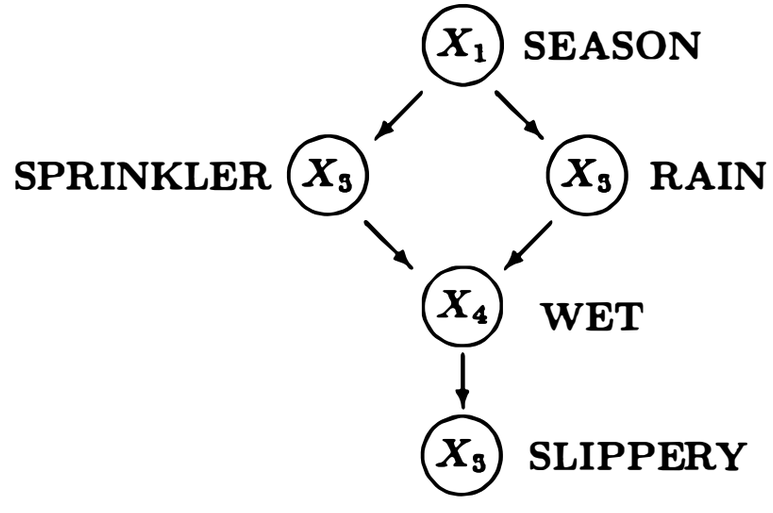
- Causal thinking
- **Identification of causal effects**
- Causal discovery
  - Constraint-based approach
  - Non-Gaussian or nonlinear methods
  - Extensions



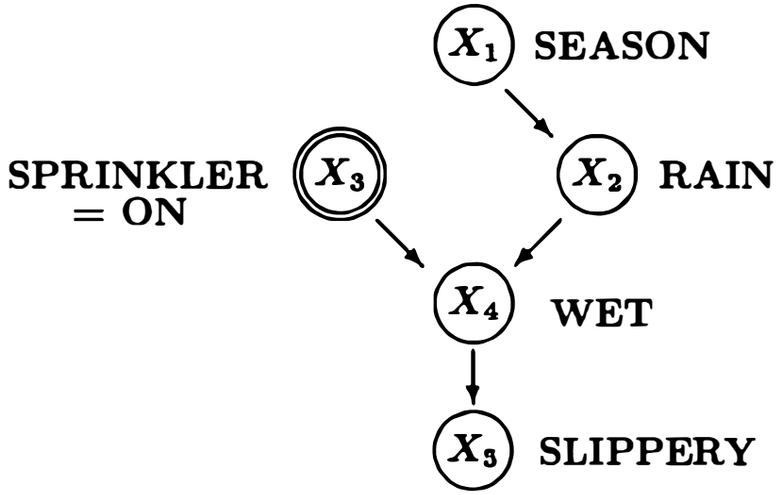

# Causal DAGs

- Causal DAGs
  - Able to **represent and respond to external or spontaneous changes**




Let $P_x(V)$ be the distribution of $V$ resulting from intervention $do(X{=}x)$. A DAG $G$ is a causal DAG if

1. $P_x(V)$ is Markov relative to $G$;
2. $P_x(V_i{=}v_i){=}1$ for all $V_i \in$X and $v_i$ consistent with $X{=}x$;
3. $P_x(V_i \mid PA_i) = P(V_i \mid PA_i)$ for all $V_i \notin$X, i.e., $P(V_i \mid PA_i)$ remains invariant to interventions not involving $V_i$.




What is $P_{X3=ON}(X_1,X_2,X_4,X_5)$?

# Structural Causal Models

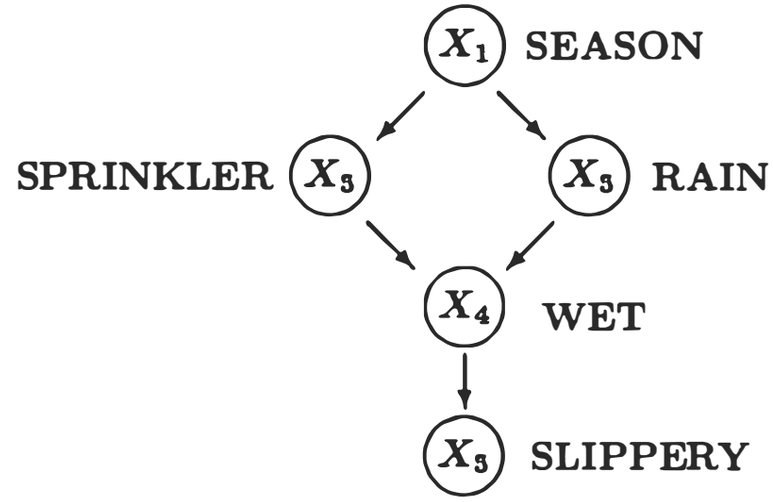
$$PA_i \longrightarrow X_i$$

- $X_i = f_i(PA_i, E_i)$, $i=1,\ldots,n$
  - $E_i$: exogenous variables / errors / disturbances
  - Each equation represents an *autonomous* mechanism
    - Describes how nature assigns values to variables of interest

$$\begin{aligned} X_1 &= E_1, \\ X_2 &= f_2(X_1, E_2), \\ X_3 &= f_3(X_1, E_3), \\ X_4 &= f_2(X_3, X_2, E_4), \\ X_5 &= f_5(X_4, E_5) \end{aligned}$$

- Distinction between structural equations & algebraic equations
- Associated with graphical causal models

# The Three Questions




- Three questions:
  - **Prediction**: Would the pavement be slippery if we *find* the sprinkler off?

$$P(Slippery \mid Sprinkler{=}off)$$

  - **Intervention**: Would the pavement be slippery if we *make sure* that the sprinkler is off?

$$P(Slippery \mid do(Sprinkler{=}off))$$

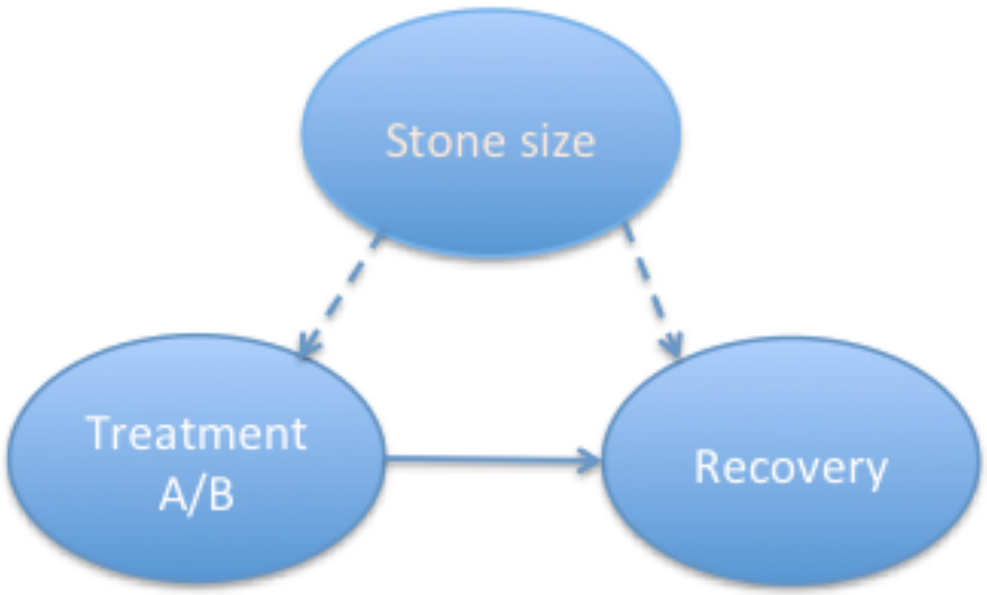
  - **Counterfactual**: Would the pavement be slippery *had* the sprinkler been off, *given that the pavement is in fact not slippery and the sprinkler is on*?

$$P(Slippery_{\,Sprinker=off} \mid Sprinkler = on, Slippery = no)$$

# Identification of Causal Effects

*P(Recovery | do(Treatment=A)) ?*

- "Golden standard": randomized controlled experiments
  - **All the other factors** that influence the outcome variable are either fixed or vary at random, so any changes in the outcome variable must be due to the controlled variable

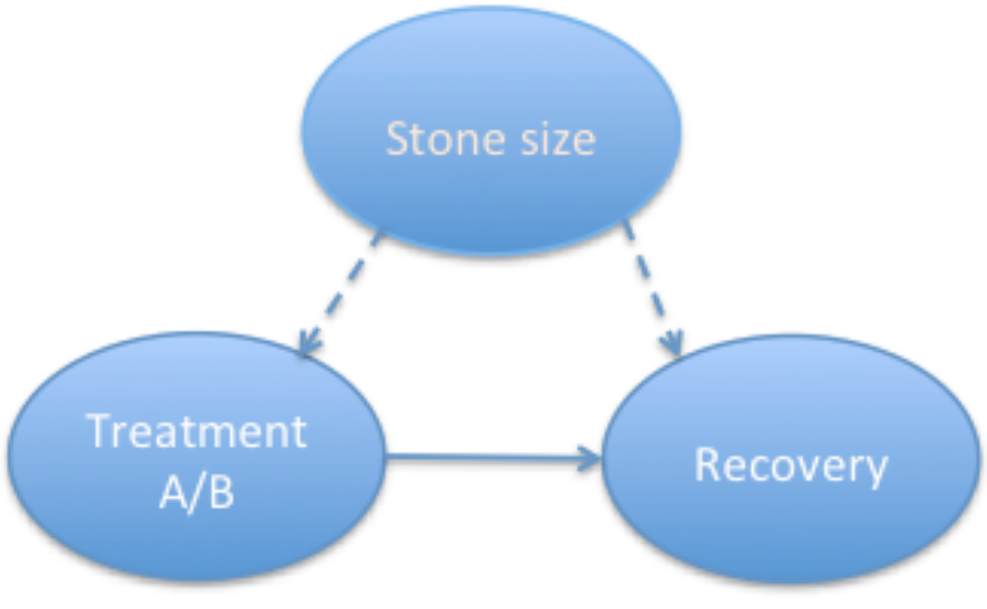



- Usually expensive or impossible to do!

# Identification of Causal Effects: Example

| | Treatment A | Treatment B |
|---|---|---|
| **Small Stones** | *Group 1*<br>**93% (81/87)** | *Group 2*<br>87% (234/270) |
| **Large Stones** | *Group 3*<br>**73% (192/263)** | *Group 4*<br>69% (55/80) |
| **Both** | 78% (273/350) | **83% (289/350)** |

$$P(R|T) = \sum_S P(R|T,S)P(S|T)$$

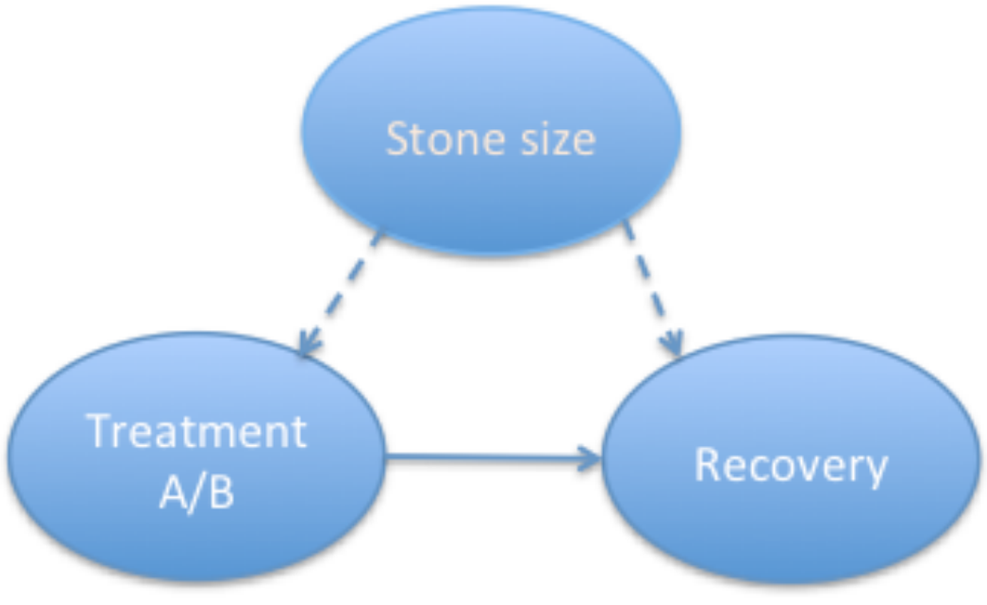


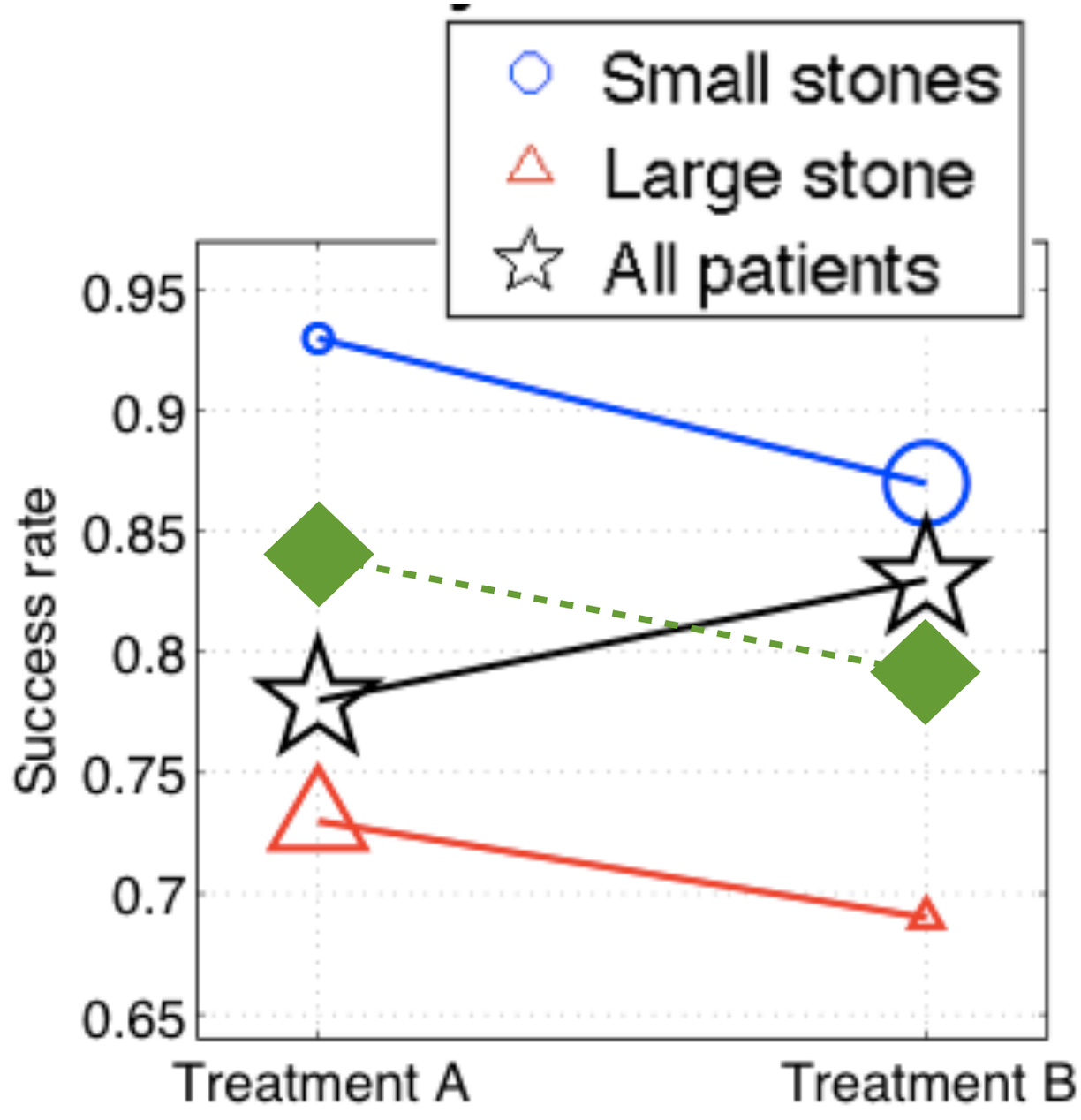

$$P(R|do(T)) = \sum_S P(R|T,S)P(S)$$

conditioning vs. manipulating

# Identification of Causal Effects: Example

| | Treatment A | Treatment B |
|---|---|---|
| **Small Stones** | *Group 1* **93% (81/87)** | *Group 2* 87% (234/270) |
| **Large Stones** | *Group 3* **73% (192/263)** | *Group 4* 69% (55/80) |
| **Both** | 78% (273/350) | **83% (289/350)** |

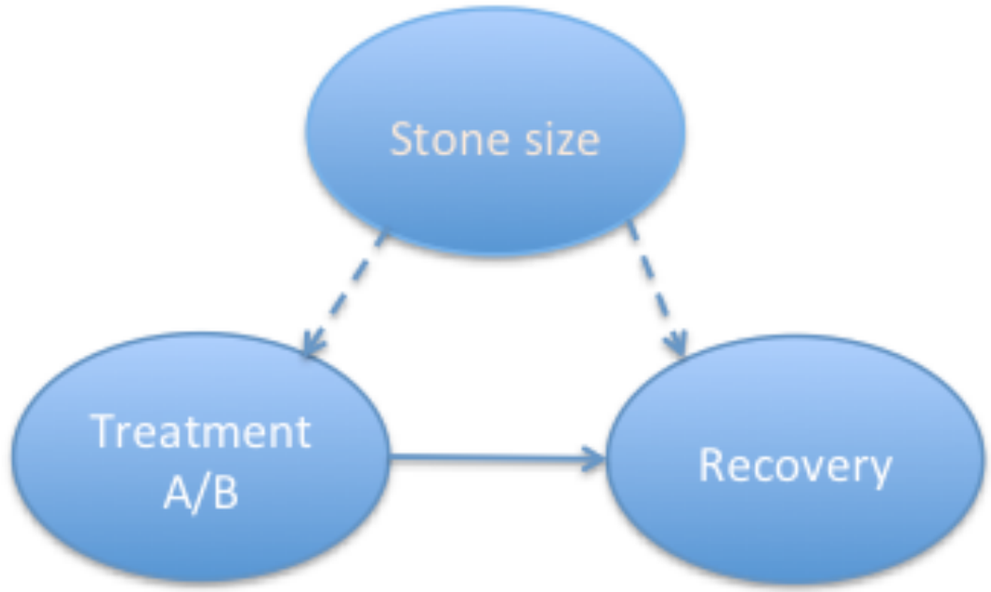



conditioning vs. manipulating

# Causal Effects

- Is causal effect, denoted by $P(Y \mid do(X))$, identifiable given complete or partial causal knowledge?
- How?




✳ **Definition 3.2.1 (Causal Effect)**

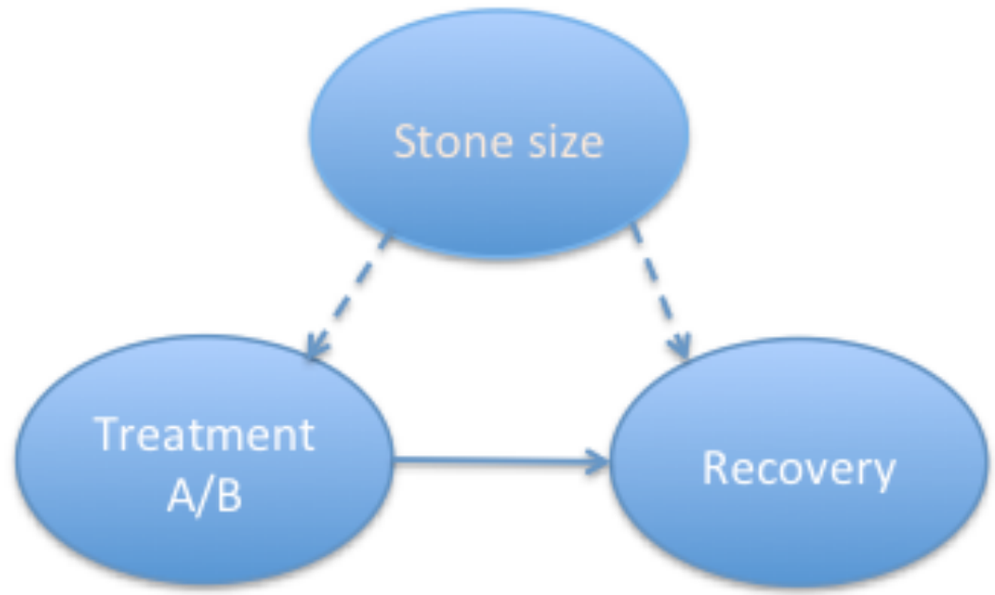
*Given two disjoint sets of variables, X and Y, the* causal effect *of X on Y, denoted either as* $P(y \mid \hat{x})$ *or as* $P(y \mid do(x))$, *is a function from X to the space of probability distributions on Y. For each realization x of X,* $P(y \mid \hat{x})$ *gives the probability of* $Y = y$ *induced by deleting from the model of* (3.4) *all equations corresponding to variables in X and substituting* $X = x$ *in the remaining equations.*

$$x_i = f_i(pa_i, u_i), \quad i = 1, \ldots, n, \tag{3.4}$$

*Examples: Average causal effect (ACE)...*

# Identifiability of Causal Effects

- Is causal effect, denoted by $P(Y \mid do(X))$, identifiable given complete or partial causal knowledge?
  - Two models with the same causal structure and the same distribution for the observed variables give the same causal effect?



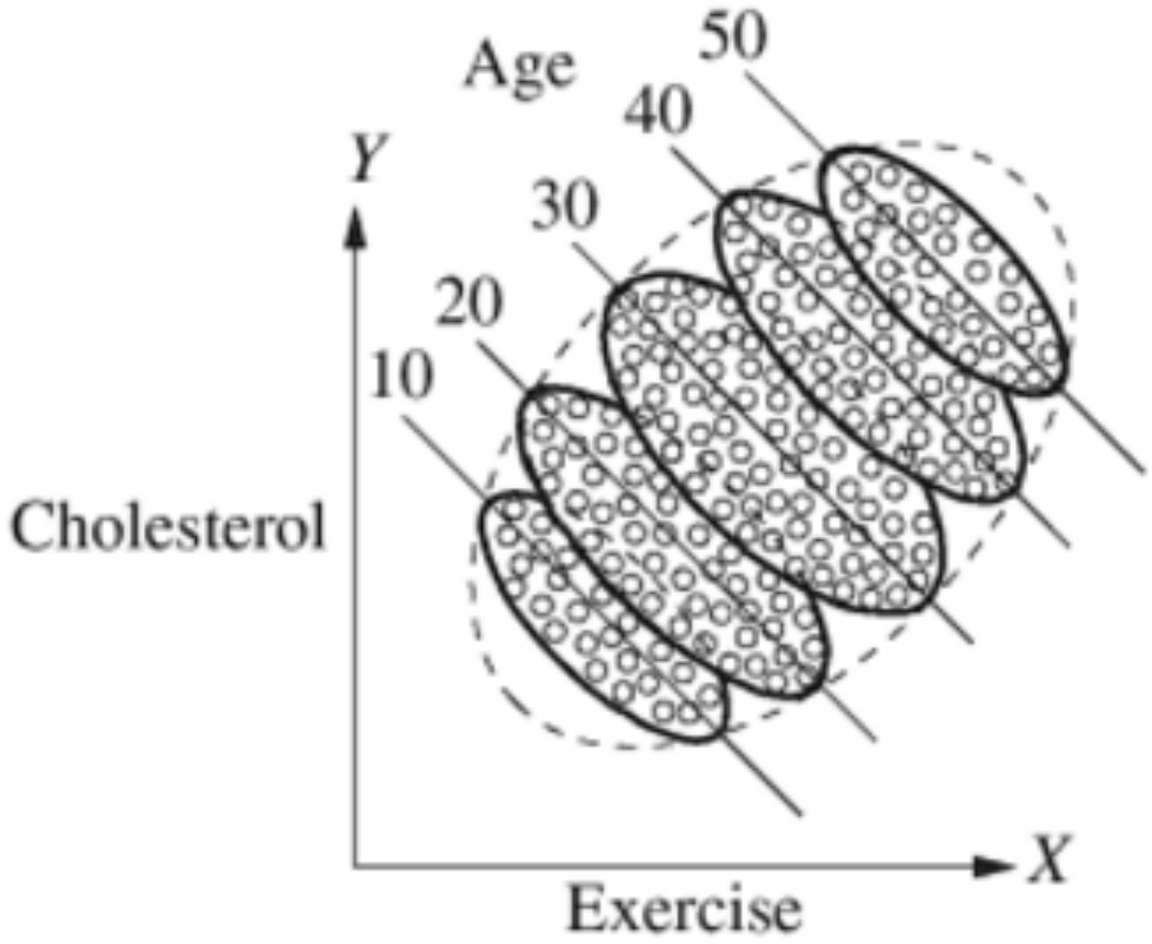

- How?

* **Definition 3.2.4 (Causal Effect Identifiability)**
*The* causal effect *of X on Y is* identifiable *from a graph G if the quantity* $P(y \mid \hat{x})$ *can be computed uniquely from any positive probability of the observed variables – that is, if* $P_{M_1}(y \mid \hat{x}) = P_{M_2}(y \mid \hat{x})$ *for every pair of models* $M_1$ *and* $M_2$ *with* $P_{M_1}(v) = P_{M_2}(v) > 0$ *and* $G(M_1) = G(M_2) = G$.

*Examples: Average causal effect (ACE)...*

# Key Issue: Controlling Confounding Bias

- Exercise-cholesterol study

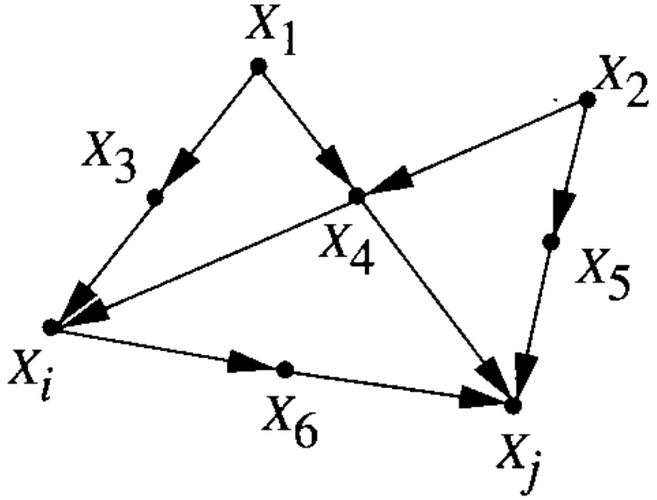


# Back-Door Criterion

**Definition 3.3.1 (Back-Door)**
*A set of variables Z satisfies the* back-door *criterion relative to an ordered pair of variables* $(X_i, X_j)$ *in a DAG G if:*

(i) *no node in Z is a descendant of* $X_i$; *and*

(ii) *Z blocks every path between* $X_i$ *and* $X_j$ *that contains an arrow into* $X_i$.

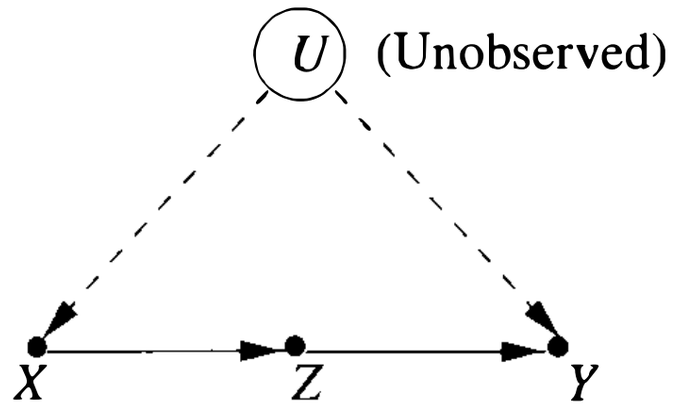



- What if $Z = \{X_3, X_4\}$?
  $Z = \{X_4, X_5\}$?
  $Z = \{X_4\}$?
- What if there is a confounder?

**Theorem 3.3.2 (Back-Door Adjustment)**
*If a set of variables Z satisfies the back-door criterion relative to* $(X, Y)$, *then the causal effect of X on Y is identifiable and is given by the formula*

$$P(y \mid \hat{x}) = \sum_z P(y \mid x, z) P(z).$$

# Front-Door Criterion




**Definition 3.3.3 (Front-Door)**
*A set of variables Z is said to satisfy the* front-door *criterion relative to an ordered pair of variables* $(X, Y)$ *if:*

(i) *Z intercepts all directed paths from X to Y;*

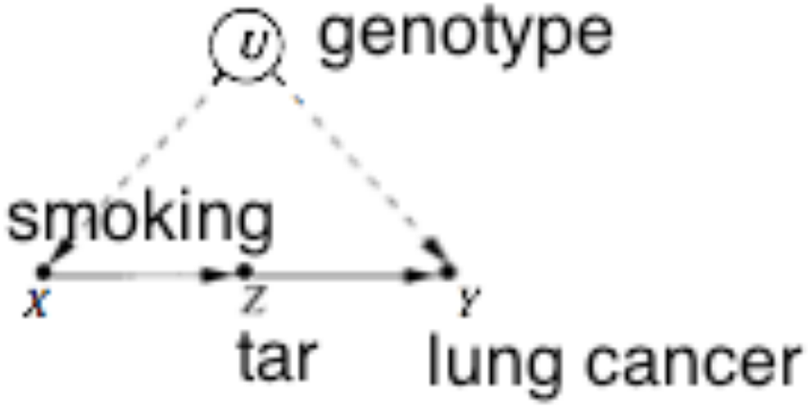
(ii) *there is no back-door path from X to Z; and*

(iii) *all back-door paths from Z to Y are blocked by X.*

**Theorem 3.3.4 (Front-Door Adjustment)**
*If Z satisfies the front-door criterion relative to* $(X, Y)$ *and if* $P(x, z) > 0$, *then the causal effect of X on Y is identifiable and is given by the formula*

$$P(y \mid \hat{x}) = \sum_{z} P(z \mid x) \sum_{x'} P(y \mid x', z) P(x'). \tag{3.29}$$

# Example: Smoking & Genotype Theory




| | Group Type | $P(x, z)$ Group Size (% of Population) | $P(Y = 1 \mid x, z)$ % of Cancer Cases in Group |
|---|---|---|---|
| $X = 0,\ Z = 0$ | Nonsmokers, No tar | 47.5 | 10 |
| $X = 1,\ Z = 0$ | Smokers, No tar | 2.5 | 90 |
| $X = 0,\ Z = 1$ | Nonsmokers, Tar | 2.5 | 5 |
| $X = 1,\ Z = 1$ | Smokers, Tar | 47.5 | 85 |

$$
\begin{aligned}
P(Y = 1 \mid do(X = 1)) &= .05(.10 \times .50 + .90 \times .50) \\
&\quad +.95(.05 \times .50 + .85 \times .50) \\
&= .05 \times .50 + .95 \times .45 = .4525,
\end{aligned}
$$

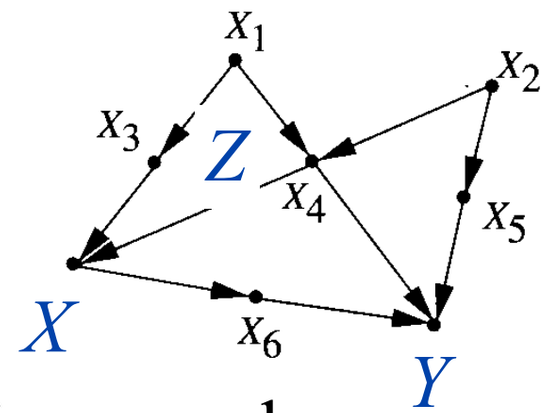
$$
\begin{aligned}
P(Y = 1 \mid do(X = 0)) &= .95(.10 \times .50 + .90 \times .50) \\
&\quad +.05(.05 \times .50 + .85 \times .50) \\
&= .95 \times .50 + .05 \times .45 = .4975.
\end{aligned}
$$

# * Relation to Ignorability (Potential Outcome Framework)

**Definition 3.3.1 (Back-Door)**
*A set of variables Z satisfies the* back-door *criterion relative to an ordered pair of variables* $(X_i, X_j)$ *in a DAG G if:*

(i) *no node in Z is a descendant of* $X_i$; *and*

(ii) *Z blocks every path between* $X_i$ *and* $X_j$ *that contains an arrow into* $X_i$.




- (Conditional) ignorability assumption in the potential outcome framework:

$Y(x) \perp\!\!\!\perp X \mid Z$.

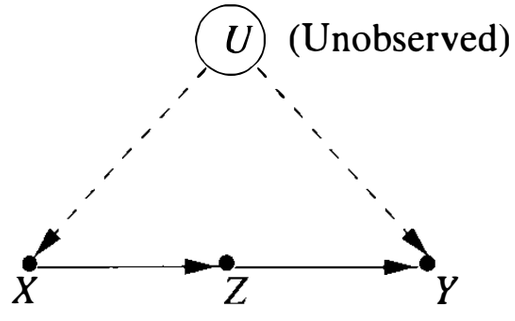
$Y(x,u)$: the value attained by $Y$ in unit $u$ under intervention do($x$);
$Y(x)$: counterfactual variable ($u$ is treated as a variable)

**Definition 3.3.3 (Front-Door)**
*A set of variables Z is said to satisfy the* front-door *criterion relative to an ordered pair of variables* $(X, Y)$ *if:*

(i) *Z intercepts all directed paths from X to Y*;

(ii) *there is no back-door path from X to Z*; *and*

(iii) *all back-door paths from Z to Y are blocked by X*.



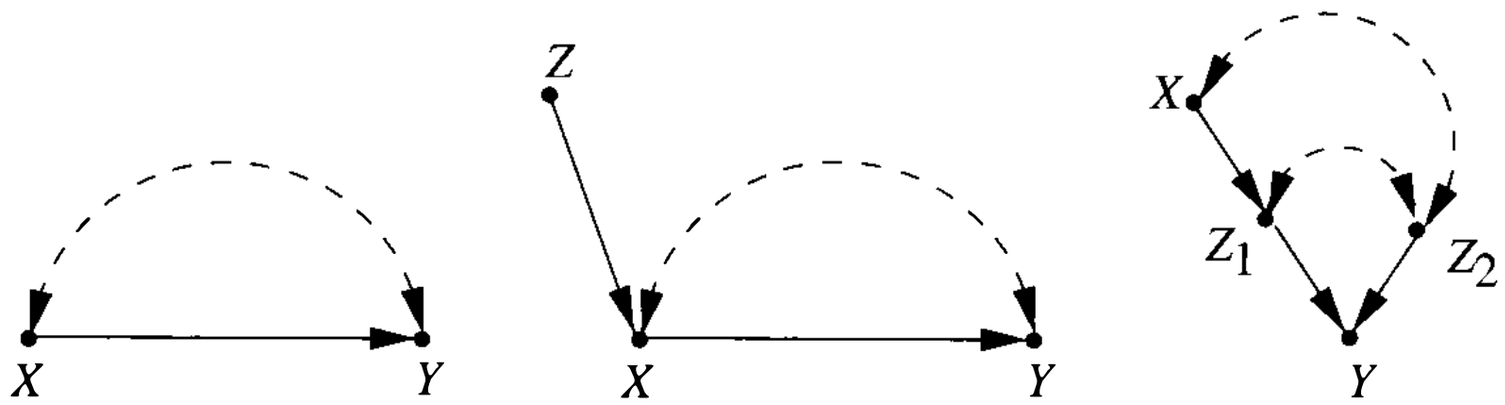

- $Y(z,x) = Y(z)$; $\{Y(z), X\} \perp\!\!\!\perp Z(x)$.

# A Unification

- (Pear & Tian, 2002) A sufficient condition for identifying the causal effect $P(y \mid do(x))$ is that there exists no bi-directed path (i.e., a path composed entirely of bi-directed arcs) between $X$ and any of its children.

- Necessary & sufficient conditions also exist...

- Examples:

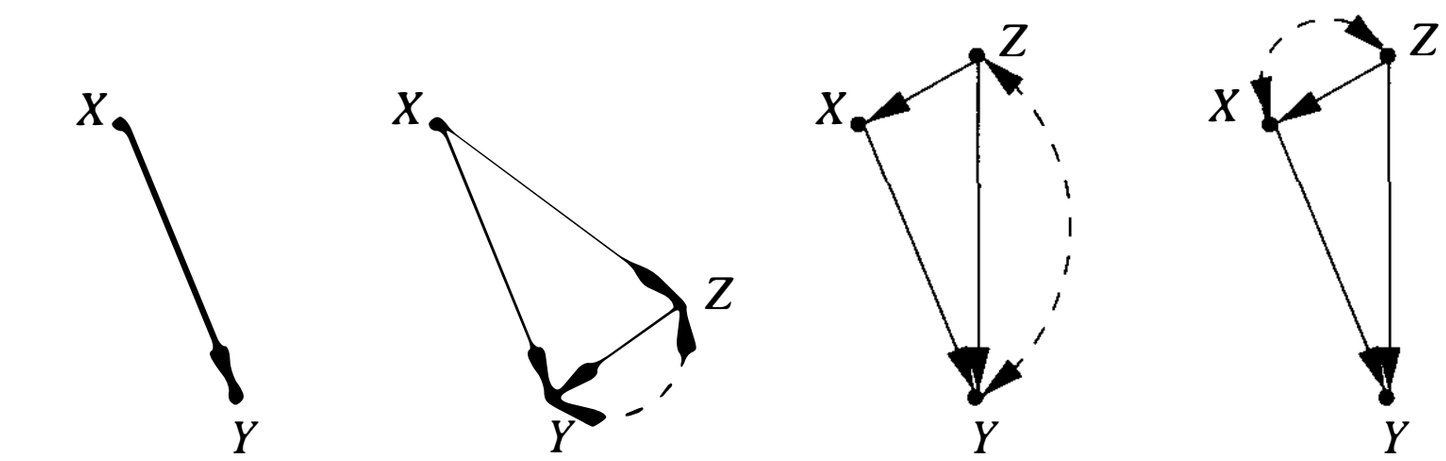



(a) (b) (c)

**Figure 3.7** (a) A bow pattern: a confounding arc embracing a causal link $X \rightarrow Y$, thus preventing the identification of $P(y \mid \hat{x})$ even in the presence of an instrumental variable $Z$, as in (b). (c) A bowless graph that still prohibits the identification of $P(y \mid \hat{x})$.

# A Unification: Examples

- Examples:



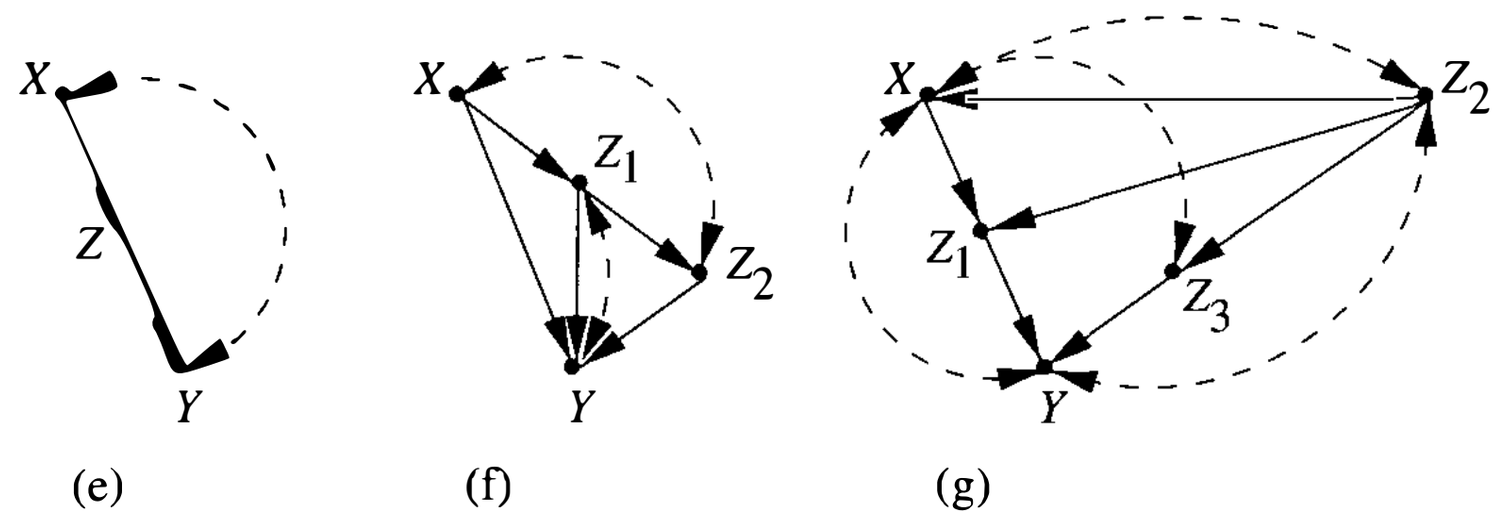
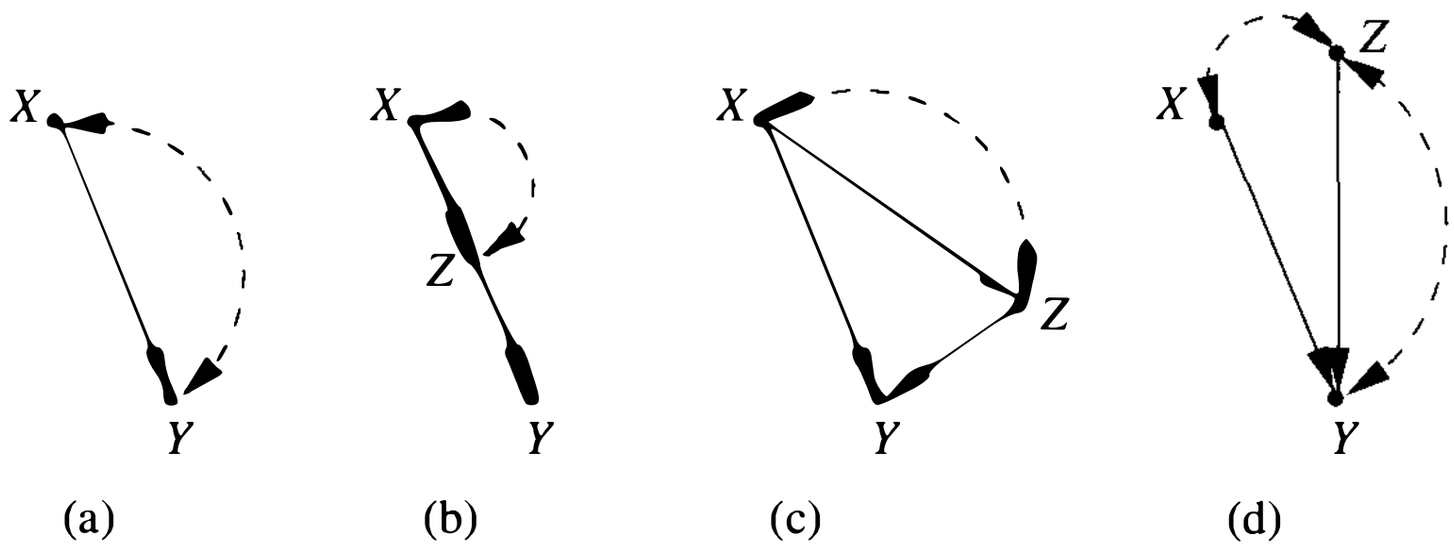




**Figure 3.8** Typical models in which the effect of $X$ on $Y$ is identifiable. Dashed arcs represent confounding paths, and $Z$ represents observed covariates.

# A Unification: Examples

- Examples:



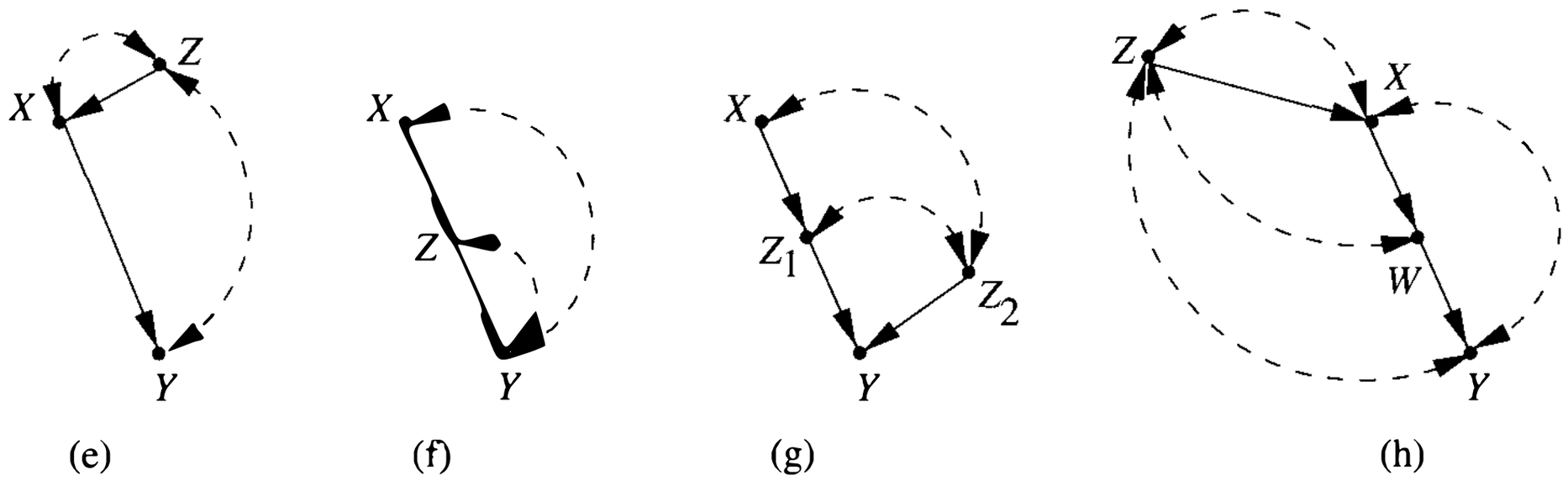
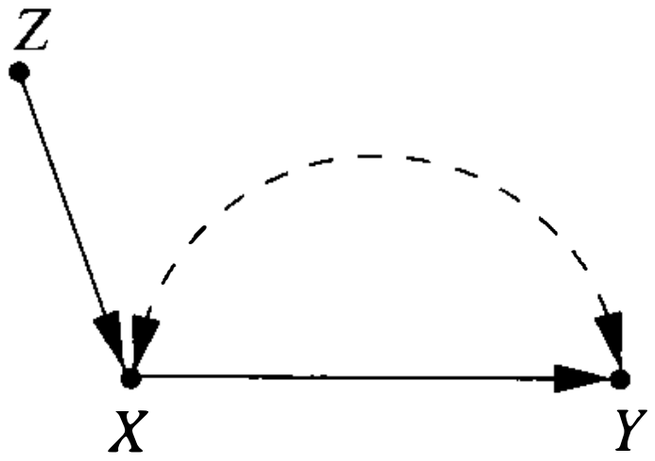
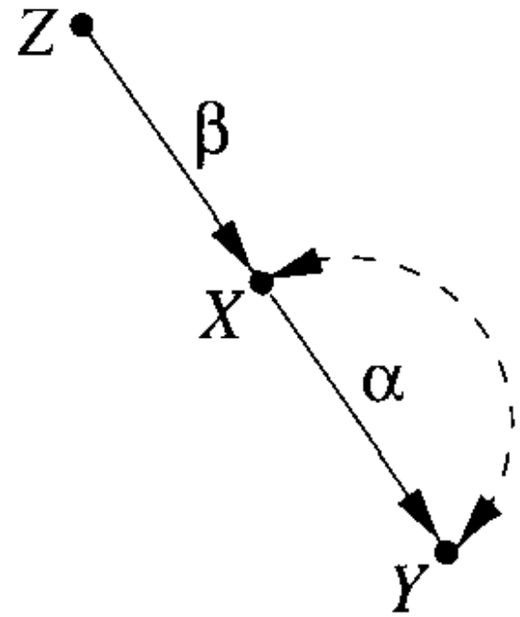




**Figure 3.9** Typical models in which $P(y \mid \hat{x})$ is not identifiable.

# Nonparametric vs. Parametric




- What if the causal relations are linear?

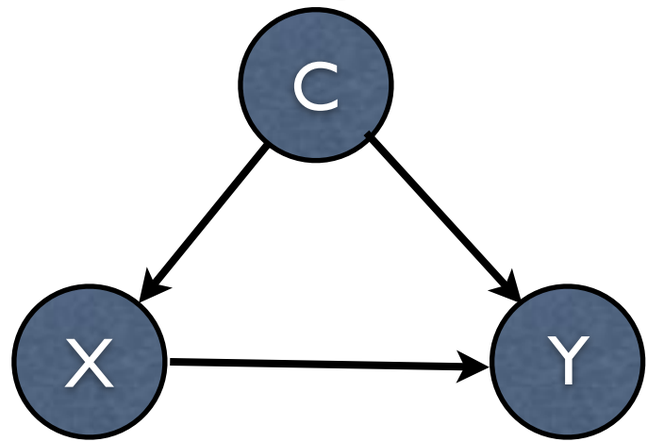
$\beta = r_{XZ}$ (regression coefficient of regressing X on Z)

$\alpha\beta = r_{YZ}$

so $\alpha = r_{YZ}/r_{XZ}$.

# Calculation of Causal Effects

- Suppose the back-door criterion (or conditional ignorability) holds

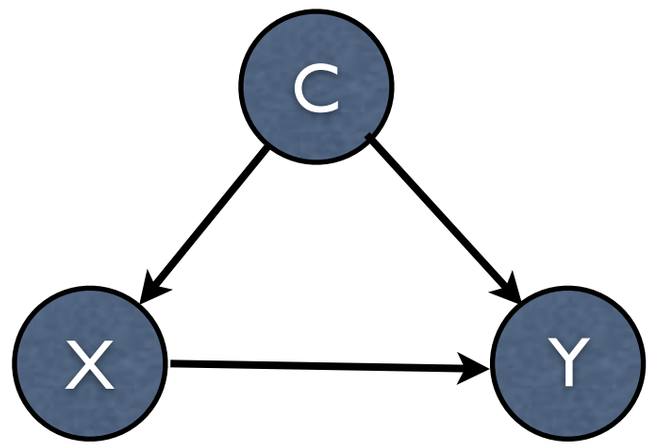



- ACE: $E[Y \mid do(x)] - E[Y \mid do(x')]$
  - $x$: the active treatment value; $x'$: the baseline treatment value
  - The two groups do not necessary have the same $P(c)$

$$P(Y|do(x)) = \sum_c P(Y|x,c)P(c)$$

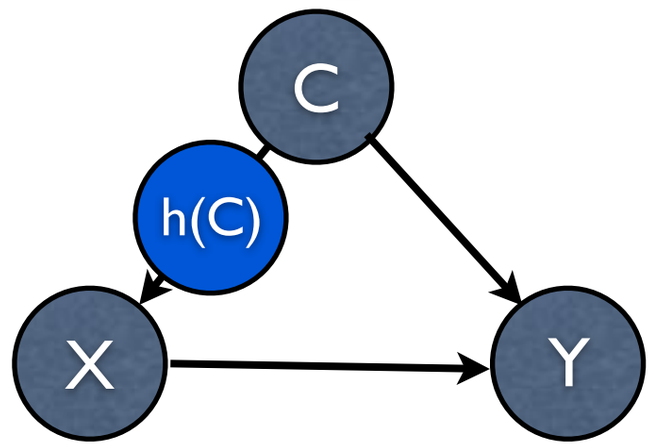
- One way is to match (usually high-dimensional) covariates $C$
- Alternatively, use the *propensity score*

# *Calculation of Causal Effects: Propensity Score




- ACE: $E[Y \mid do(x)]$ - $E[Y \mid do(x')]$
  - $x$: the active treatment value; $x'$: the baseline treatment value $$P(Y|do(x)) = \sum_c P(Y|x,c)P(c)$$
    - The two groups do not necessary have the same $P(c)$
- One way is to match (usually high-dimensional) covariates $C$
- Propensity Score
  - Let $h(C) = P(X=1 \mid C)$; $X \perp\!\!\!\perp C \mid h(C)$
  - Then $h(C)$ and $C$ are (confounding)-equivalent:

# *Calculation of Causal Effects: Propensity Score




- ACE: $E[Y \mid do(x)]$ - $E[Y \mid do(x')]$
  - $x$: the active treatment value; $x'$: the baseline treatment value

$$P(Y|do(x)) = \sum_c P(Y|x,c)P(c)$$

  - The two groups do not necessary have the same $P(c)$

- One way is to match (usually high-dimensional) covariates $C$
- Pro
  - I
  - T

$$\sum_c P(Y|x,c)P(c) = \sum_c \sum_h P(Y|x,c)p(h)p(c|h)$$

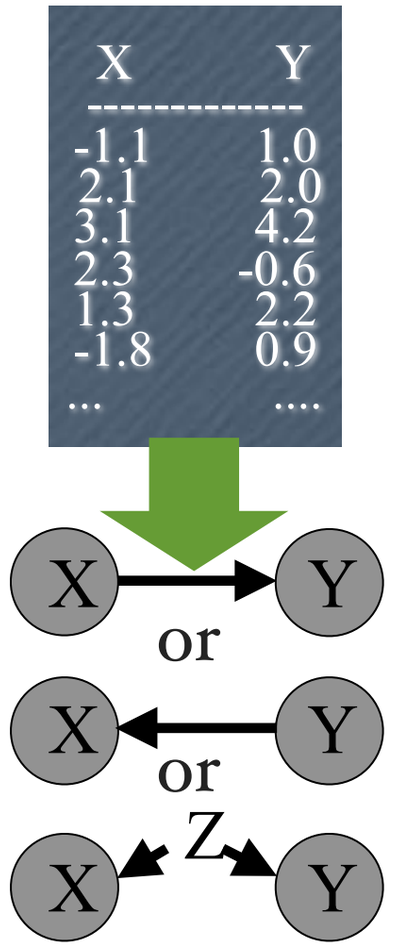
$$= \sum_c \sum_h P(Y|x,c,h)p(h)p(c|h,x) = \sum_c \sum_h P(Y,c|x,h)P(h)$$

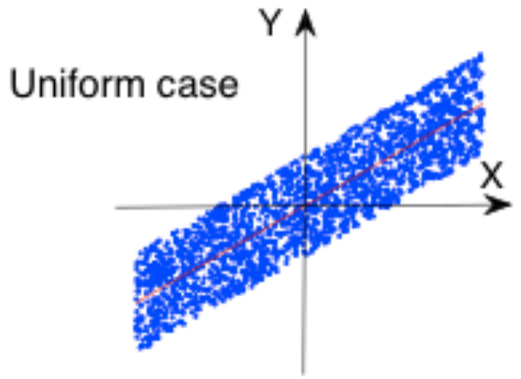
$$= \sum_h P(Y|x,h)P(h)$$

Advantage!?!

# Outline

- Causal thinking
- Identification of causal effects
- **Causal discovery**
  - **Constraint-based approach**
  - Non-Gaussian or nonlinear methods
  - Extensions



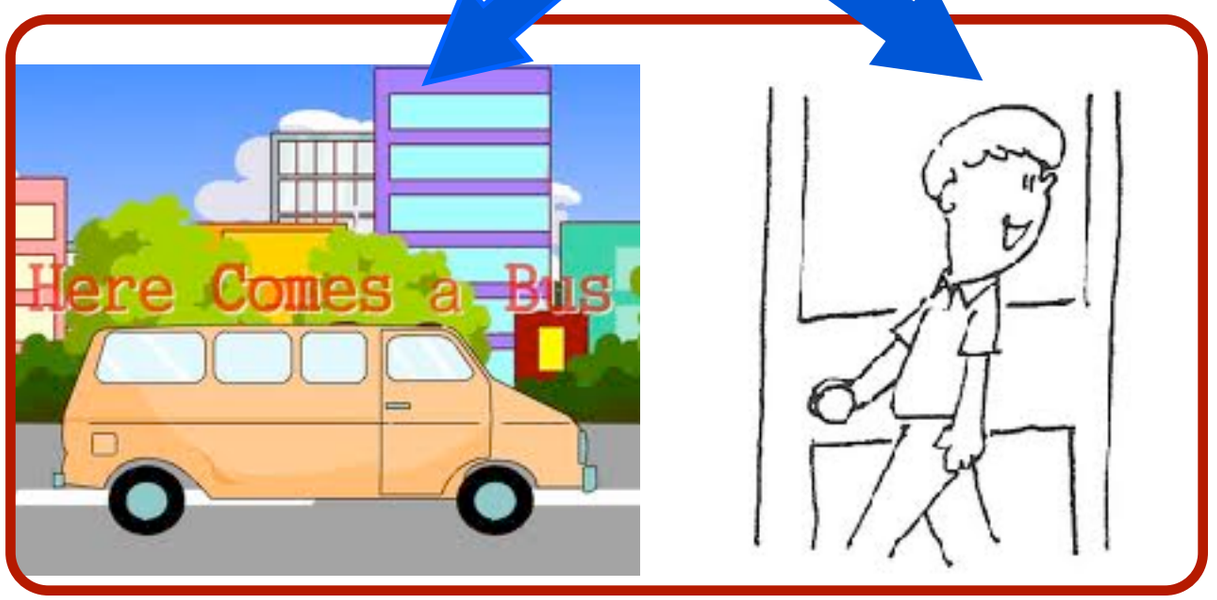




# Classic Ways to Find Causal Information



- What if you manipulate *X* and see *Y* also *changes*?
  - A manipulation directly changes only the target variable *X*
- Manipulate vs. change

Bus timetable



# Causal Discovery from Data: Examples

Science AAAS March, 2014



## Large-Scale Psychological Differences Within China Explained by Rice Versus Wheat Agriculture

T. Talhelm,[1*] X. Zhang,[2,3] S. Oishi,[1] C. Shimin,[4] D. Duan,[2] X. Lan,[5] S. Kitayama[5]

Cross-cultural psychologists have mostly contrasted East Asia with the West. However, this study shows that there are major psychological differences within China. We propose that a history of farming rice makes cultures more interdependent, whereas farming wheat makes cultures more independent, and these agricultural legacies continue to affect people in the modern world. We tested 1162 Han Chinese participants in six sites and found that rice-growing southern China is more interdependent and holistic-thinking than the wheat-growing north. To control for confounds like climate, we tested people from neighboring counties along the rice-wheat border and found differences that were just as large. We also find that modernization and pathogen prevalence theories do not fit the data.

Over the past 20 years, psychologists have cataloged a long list of differences be- more insular and collectivistic (*6*). Studies have found that historical pathogen prevalence

founded with rice—a possibility that prior research did not control for.

X: rice/wheat agriculture;
Y: culture;
Z: climate etc.:

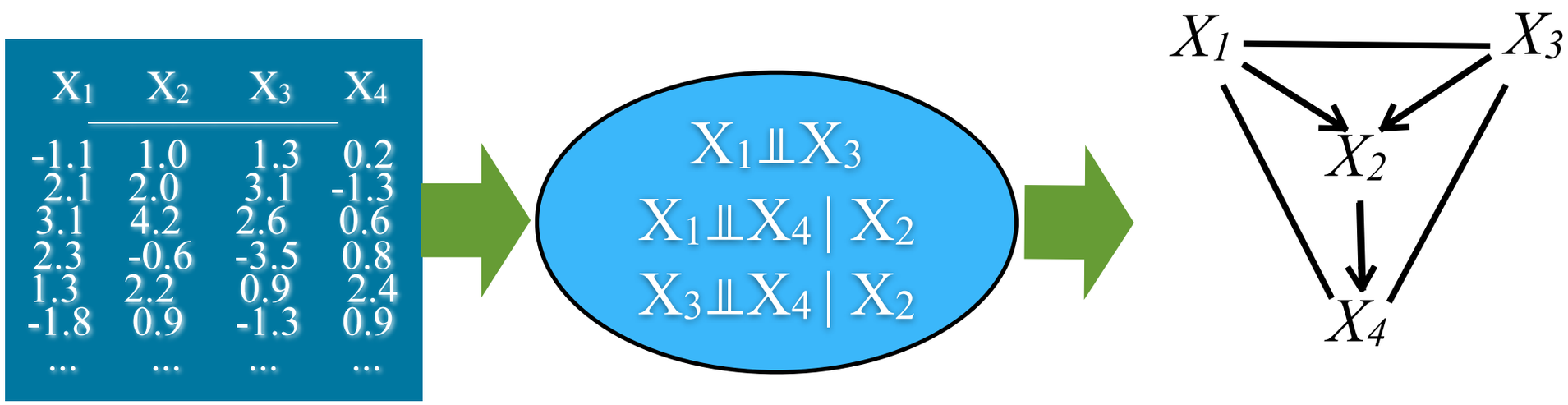
$X \not\perp\!\!\!\perp Y$;
$X \not\perp\!\!\!\perp Y \mid Z$.

Under what conditions can we say
$X \rightarrow Y$ ?

subsistence crops—rice and wheat—are very dif-

# Constraint-Based Causal Discovery: Big Picture

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



# 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 $\mathsf{G}^{\mathsf{f}}$ *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)

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

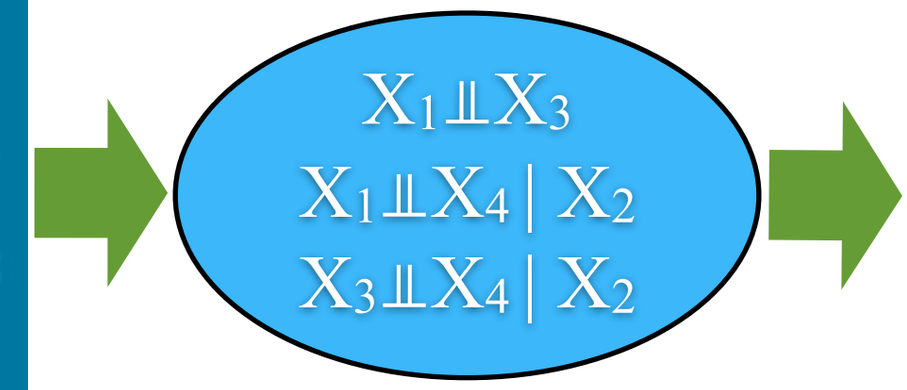
**Faithfulness**: all observed (conditional) independencies are entailed by Markov condition in the causal graph

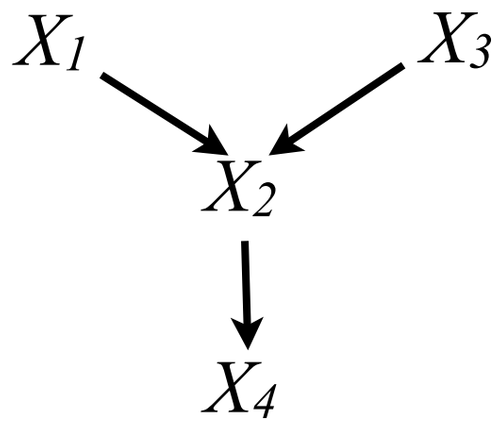
Recall: $Y \perp\!\!\!\perp Z \Leftrightarrow P(Y|Z)=P(Y)$; $Y \perp\!\!\!\perp Z|X \Leftrightarrow P(Y|Z,X)=P(Y|X)$

# Constraint-Based vs. Score-Based

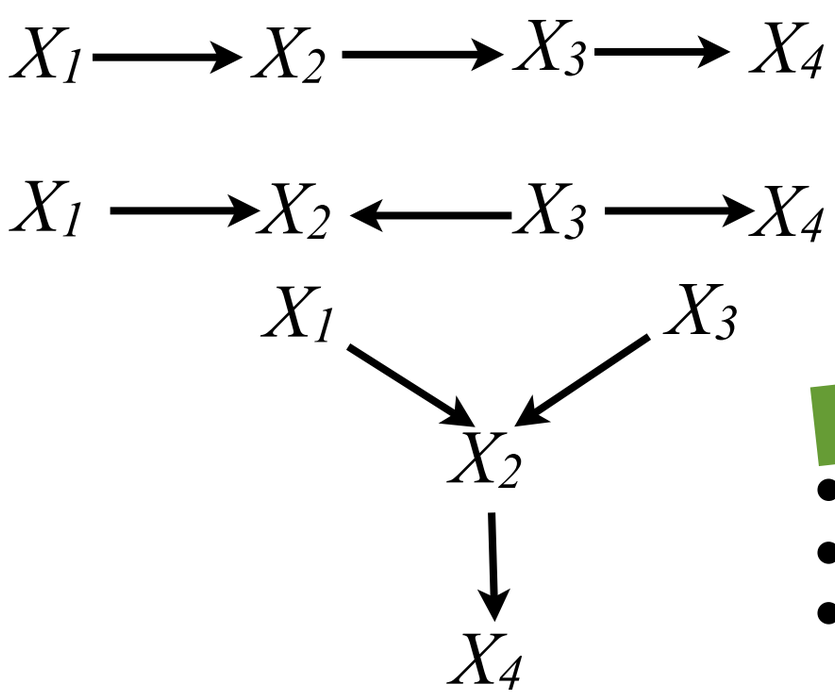
- Constraint-based methods

| $X_1$ | $X_2$ | $X_3$ | $X_4$ |
|---|---|---|---|
| -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-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.)*

# Constraint-Based vs. Score-Based

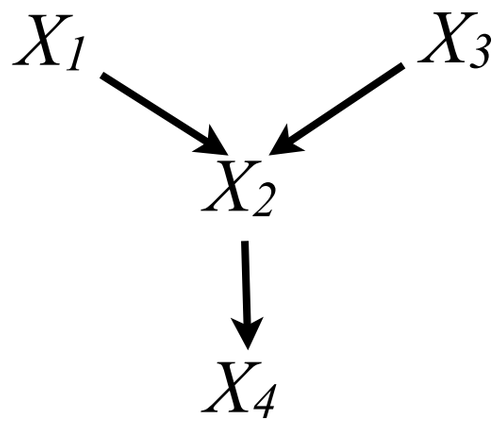
- Constraint-based methods

| $X_1$ | $X_2$ | $X_3$ | $X_4$ |
|---|---|---|---|
| -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 |
| ... | ... | ... | ... |

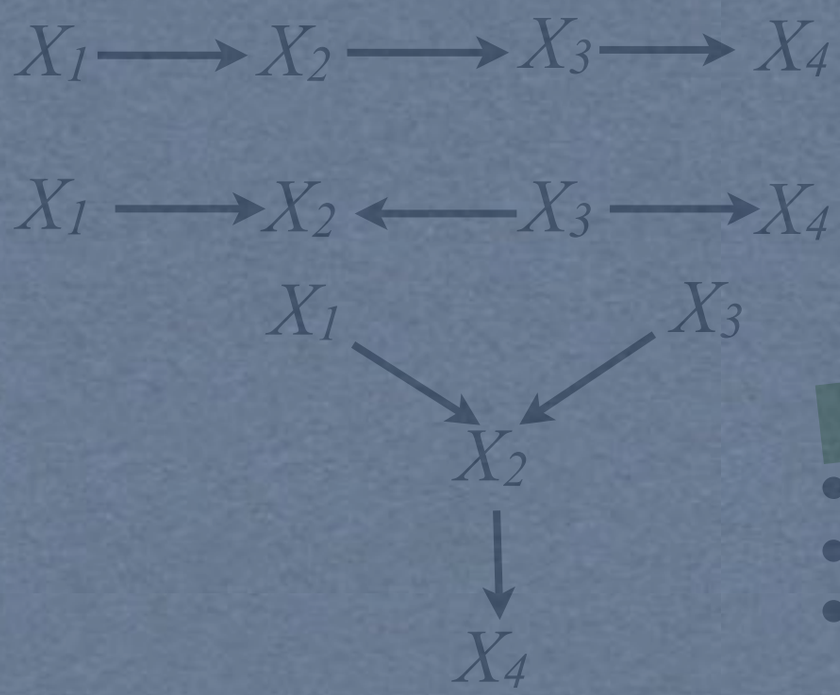
$X_1 \perp\!\!\!\perp X_3$

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

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




- 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?

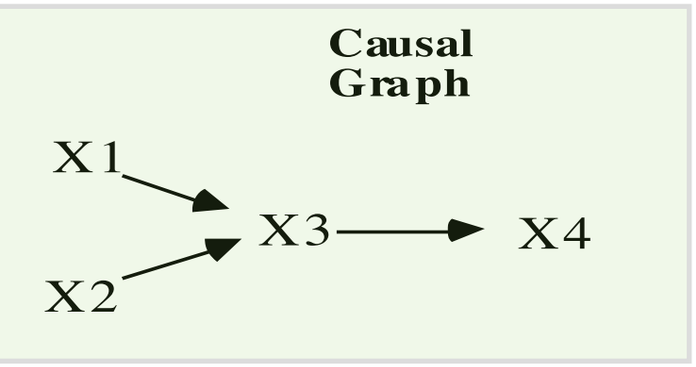
Causal Markov condition + faithfulness

- Second, can we determine the causal direction?

How?

# Example

## *Step 1: finding skeleton*




**Independcies**

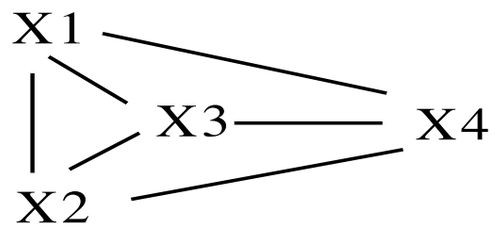
$X1 \perp\!\!\!\perp X2$

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

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

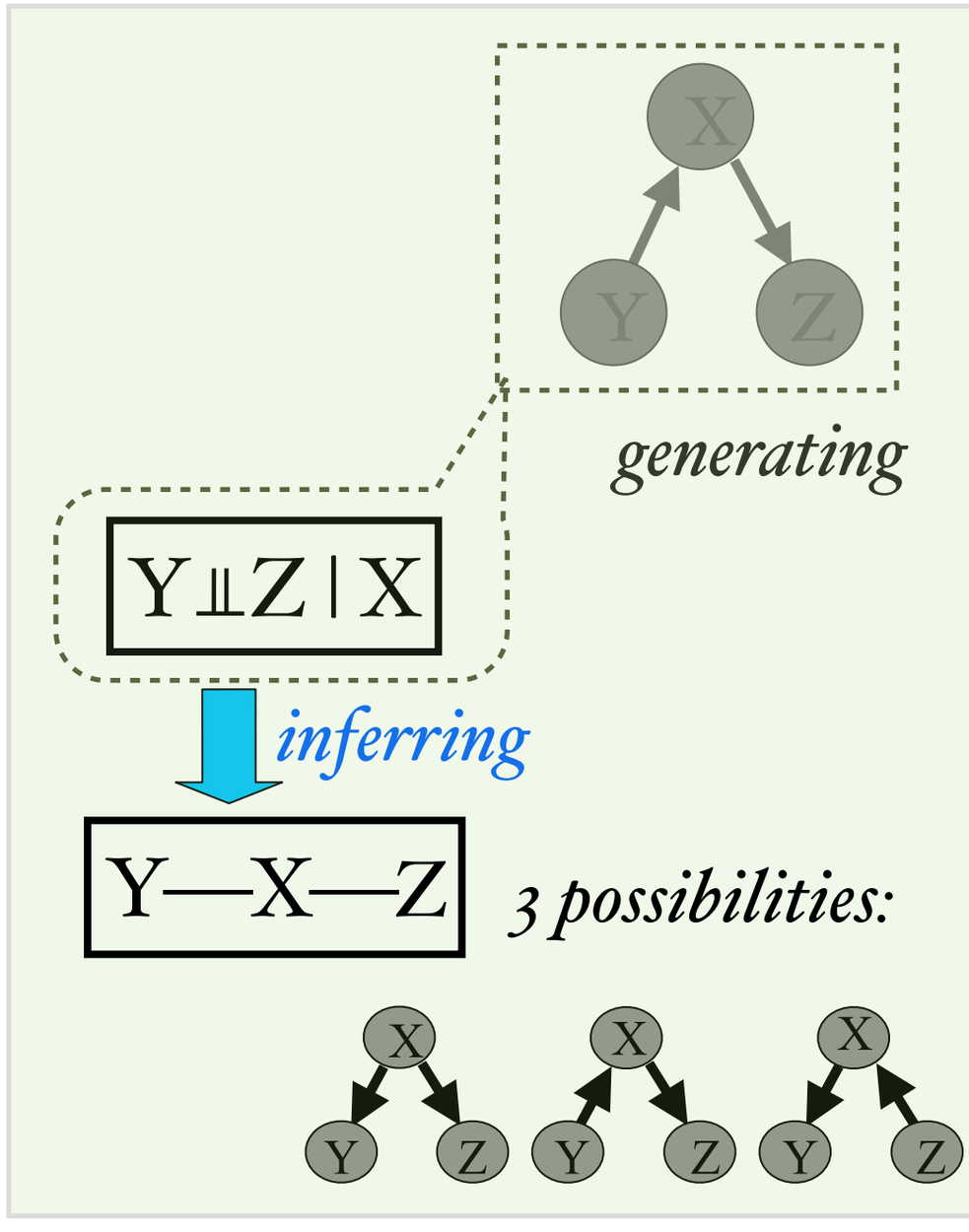
## *Step 11: finding v-structure and doing orientation propagation*

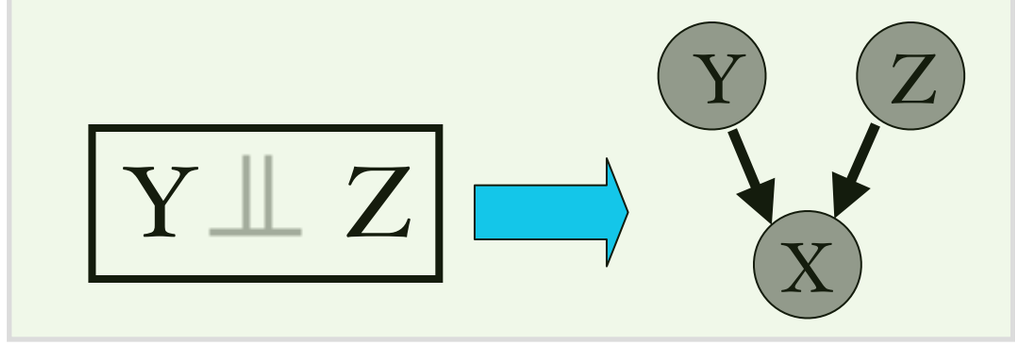
Begin with:



# 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



# 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**$(Y,X)$;

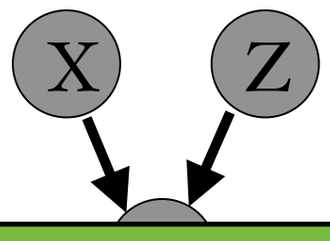
until all ordered pairs of adjacent variables $X$ and $Y$ such that **Adjacencies**$(C,X)\backslash\{Y\}$ has cardinality greater than or equal to $n$ and all subsets $\mathbf{S}$ of **Adjacencies**$(C,X)\backslash\{Y\}$ of cardinality $n$ have been tested for d-separation;

$n = n + 1$;

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**$(X,Z)$.

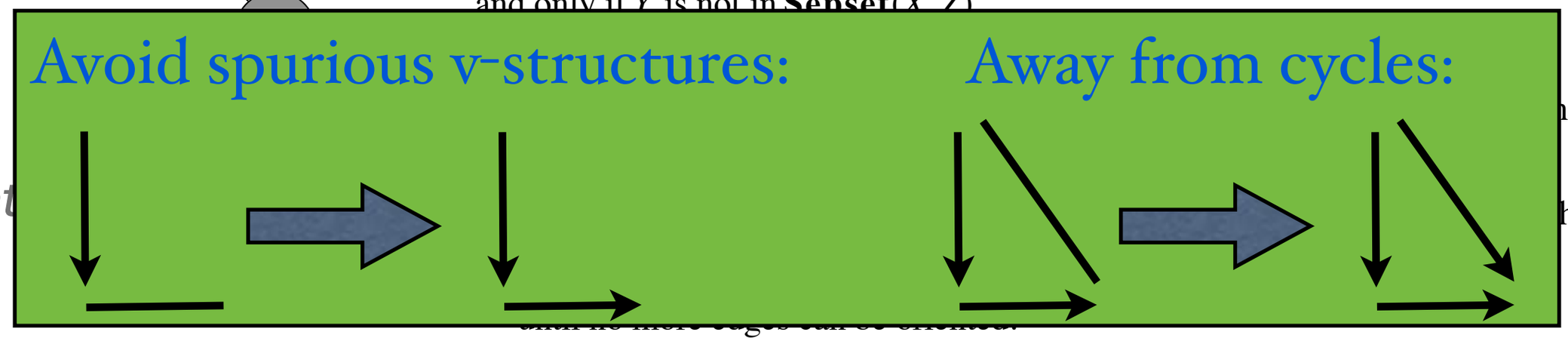
*Finding V-structures*




*Orient*

here is no

hen orient

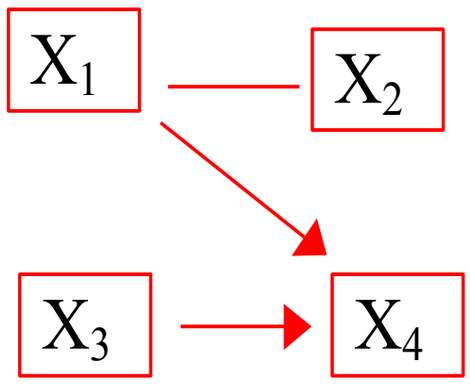


# (Independence) Equivalent Classes: Patterns

- 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

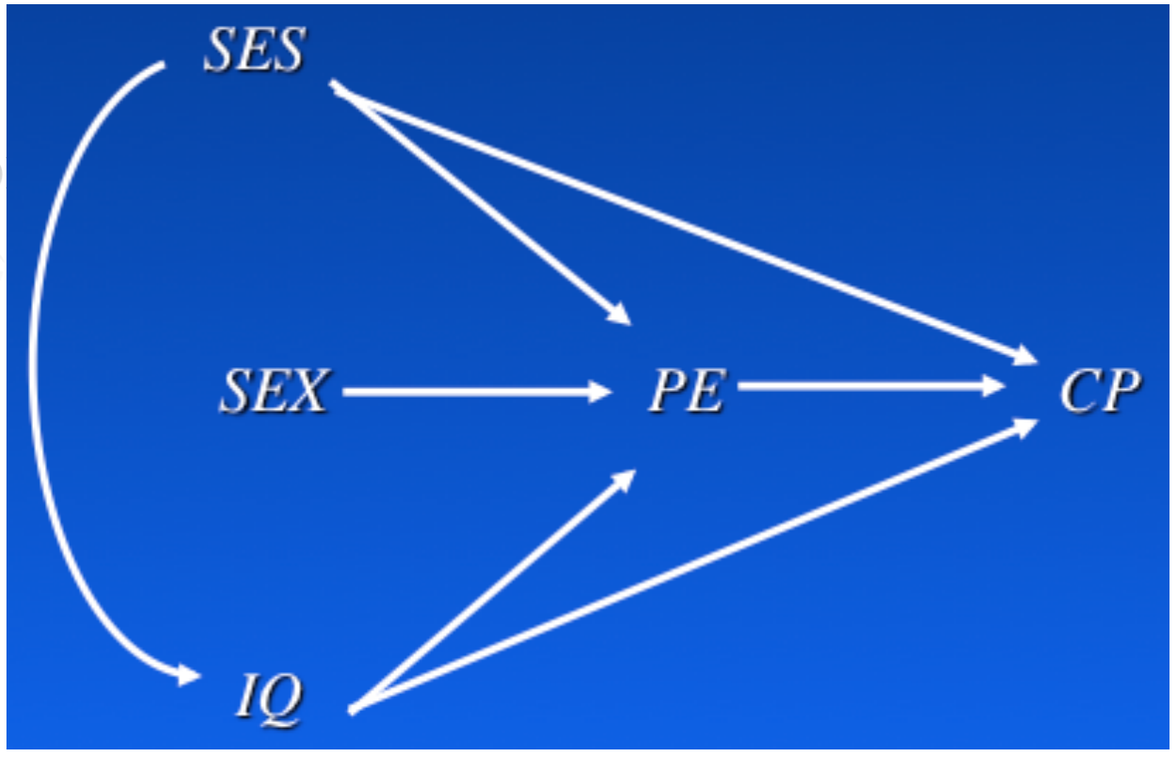



*How many DAGs in this class?*

# Example 1: 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]



# Example II: Causal analysis of archeology data

*Thanks to collaborator Marlijn Noback*

- 8 variables of 250 skeletons collected from different locations

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | ID | Population | Sex | Cranial size | Diet or subsistence | | | | | Paramastic | Dental wear | | Geographic location per population | | | Climate per population | | | | | |
| 2 | | | (Male, fem | (Centroid S | Gathering | Hunting | Fishing | Pastoralism | Agriculture | Yes=1, no=( | Average att | Attrition pe | Distance to | Longitude | Latitude | Tmean | Tmin | Tmax | Vpmean | Vpmin | Vpmax |
| 3 | ANU3L_1 | Ainu | Unknown | 713.2542 | 2 | 3 | 4 | 0 | 1 | 0 | 1.5 | 2 | 15464 | 43.548548 | 142.639159 | 2.86 | -11.19 | 17.01 | 7.43 | 2.27 | 15.83 |
| 4 | ANU7_1 | Ainu | Unknown | 676.148 | 2 | 3 | 4 | 0 | 1 | 0 | 1.5 | 1 | 15464 | 43.548548 | 142.639159 | 2.86 | -11.19 | 17.01 | 7.43 | 2.27 | 15.83 |
| 5 | ANU7_2 | Ainu | Unknown | 675.4924 | 2 | 3 | 4 | 0 | 1 | 0 | 1.5 | 1 | 15464 | 43.548548 | 142.639159 | 2.86 | -11.19 | 17.01 | 7.43 | 2.27 | 15.83 |
| 6 | ANU_L016 | Ainu | Male | 684.3304 | 2 | 3 | 4 | 0 | 1 | 0 | 1.5 | 2.5 | 15464 | 43.548548 | 142.639159 | 2.86 | -11.19 | 17.01 | 7.43 | 2.27 | 15.83 |
| 7 | ANU_L016 | Ainu | Female | 686.285 | 2 | 3 | 4 | 0 | 1 | 0 | 1.5 | 4 | 15464 | 43.548548 | 142.639159 | 2.86 | -11.19 | 17.01 | 7.43 | 2.27 | 15.83 |
| 8 | AUSM245 | Australia | Male | 673.8749 | 6 | 4 | 0 | 0 | 0 | 1 | 2.5 | 1 | 20164 | -24.287027 | 135.615234 | 22.46 | 13.33 | 30.27 | 11.10 | 7.55 | 15.96 |
| 9 | AUSM246 | Australia | Male | 647.4586 | 6 | 4 | 0 | 0 | 0 | 1 | 2.5 | 4 | 20164 | -24.287027 | 135.615234 | 22.46 | 13.33 | 30.27 | 11.10 | 7.55 | 15.96 |
| 10 | AUSM8217 | Australia | Male | 658.6616 | 6 | 4 | 0 | 0 | 0 | 1 | 2.5 | 2 | 20164 | -24.287027 | 135.615234 | 22.46 | 13.33 | 30.27 | 11.10 | 7.55 | 15.96 |
| 11 | AUSM8177 | Australia | Male | 667.5444 | 6 | 4 | 0 | 0 | 0 | 1 | 2.5 | 4 | 20164 | -24.287027 | 135.615234 | 22.46 | 13.33 | 30.27 | 11.10 | 7.55 | 15.96 |
| 12 | AUSM8173 | Australia | Male | 629.7138 | 6 | 4 | 0 | 0 | 0 | 1 | 2.5 | 3.5 | 20164 | -24.287027 | 135.615234 | 22.46 | 13.33 | 30.27 | 11.10 | 7.55 | 15.96 |
| 13 | AUSM8173 | Australia | Male | 648.7064 | 6 | 4 | 0 | 0 | 0 | 1 | 2.5 | 3.5 | 20164 | -24.287027 | 135.615234 | 22.46 | 13.33 | 30.27 | 11.10 | 7.55 | 15.96 |
| 14 | AUSM8171 | Australia | Male | 648.0378 | 6 | 4 | 0 | 0 | 0 | 1 | 2.5 | 2 | 20164 | -24.287027 | 135.615234 | 22.46 | 13.33 | 30.27 | 11.10 | 7.55 | 15.96 |
| 15 | AUSM8165 | Australia | Male | 616.55 | 6 | 4 | 0 | 0 | 0 | 1 | 2.5 | 3.5 | 20164 | -24.287027 | 135.615234 | 22.46 | 13.33 | 30.27 | 11.10 | 7.55 | 15.96 |
| 16 | AUSM8154 | Australia | Male | 635.0605 | 6 | 4 | 0 | 0 | 0 | 1 | 2.5 | 2 | 20164 | -24.287027 | 135.615234 | 22.46 | 13.33 | 30.27 | 11.10 | 7.55 | 15.96 |
| 17 | AUSM8153 | Australia | Male | 650.6559 | 6 | 4 | 0 | 0 | 0 | 1 | 2.5 | 3 | 20164 | -24.287027 | 135.615234 | 22.46 | 13.33 | 30.27 | 11.10 | 7.55 | 15.96 |
| 18 | AUSF1412 | Australia | Female | 618.4781 | 6 | 4 | 0 | 0 | 0 | 1 | 2.5 | 1 | 20164 | -24.287027 | 135.615234 | 22.46 | 13.33 | 30.27 | 11.10 | 7.55 | 15.96 |
| 19 | AUSF8179 | Australia | Female | 634.3122 | 6 | 4 | 0 | 0 | 0 | 1 | 2.5 | 3.5 | 20164 | -24.287027 | 135.615234 | 22.46 | 13.33 | 30.27 | 11.10 | 7.55 | 15.96 |
| 20 | AUSF8175 | Australia | Female | 605.1759 | 6 | 4 | 0 | 0 | 0 | 1 | 2.5 | 1.5 | 20164 | -24.287027 | 135.615234 | 22.46 | 13.33 | 30.27 | 11.10 | 7.55 | 15.96 |
| 21 | AUSF8177 | Australia | Female | 613.8424 | 6 | 4 | 0 | 0 | 0 | 1 | 2.5 | 4 | 20164 | -24.287027 | 135.615234 | 22.46 | 13.33 | 30.27 | 11.10 | 7.55 | 15.96 |
| 22 | AUSF8169 | Australia | Female | 610.1206 | 6 | 4 | 0 | 0 | 0 | 1 | 2.5 | 2.5 | 20164 | -24.287027 | 135.615234 | 22.46 | 13.33 | 30.27 | 11.10 | 7.55 | 15.96 |
| 23 | AUSF8157 | Australia | Female | 628.2814 | 6 | 4 | 0 | 0 | 0 | 1 | 2.5 | 2 | 20164 | -24.287027 | 135.615234 | 22.46 | 13.33 | 30.27 | 11.10 | 7.55 | 15.96 |
| 24 | AUSF8155 | Australia | Female | 628.4609 | 6 | 4 | 0 | 0 | 0 | 1 | 2.5 | 3.5 | 20164 | -24.287027 | 135.615234 | 22.46 | 13.33 | 30.27 | 11.10 | 7.55 | 15.96 |
| 25 | AUSF1578 | Australia | Female | 640.6311 | 6 | 4 | 0 | 0 | 0 | 1 | 2.5 | 2 | 20164 | -24.287027 | 135.615234 | 22.46 | 13.33 | 30.27 | 11.10 | 7.55 | 15.96 |
| 26 | AUSF243 | Australia | Female | 606.164 | 6 | 4 | 0 | 0 | 0 | 1 | 2.5 | 2.5 | 20164 | -24.287027 | 135.615234 | 22.46 | 13.33 | 30.27 | 11.10 | 7.55 | 15.96 |
| 27 | AUSF8158 | Australia | Female | 631.6258 | 6 | 4 | 0 | 0 | 0 | 1 | 2.5 | 2 | 20164 | -24.287027 | 135.615234 | 22.46 | 13.33 | 30.27 | 11.10 | 7.55 | 15.96 |
| 28 | DENM1432 | Denmark | Male | 653.6198 | 0 | 0 | 1 | 3 | 6 | 0 | 2.1 | 2 | 10440 | 55.717055 | 11.711426 | 8.01 | -0.02 | 16.66 | 9.67 | 5.59 | 15.27 |
| 29 | DENM1011 | Denmark | Male | 651.4847 | 0 | 0 | 1 | 3 | 6 | 0 | 2.1 | 3 | 10440 | 55.717055 | 11.711426 | 8.01 | -0.02 | 16.66 | 9.67 | 5.59 | 15.27 |
| 30 | DENM1709 | Denmark | Male | 646.9841 | 0 | 0 | 1 | 3 | 6 | 0 | 2.1 | 1.5 | 10440 | 55.717055 | 11.711426 | 8.01 | -0.02 | 16.66 | 9.67 | 5.59 | 15.27 |
| 31 | DENM116 | Denmark | Male | 642.9192 | 0 | 0 | 1 | 3 | 6 | 0 | 2.1 | 3 | 10440 | 55.717055 | 11.711426 | 8.01 | -0.02 | 16.66 | 9.67 | 5.59 | 15.27 |
| 32 | DENM116_ | Denmark | Male | 646.6609 | 0 | 0 | 1 | 3 | 6 | 0 | 2.1 | 2.5 | 10440 | 55.717055 | 11.711426 | 8.01 | -0.02 | 16.66 | 9.67 | 5.59 | 15.27 |
| 33 | DENM116 | Denmark | Male | 674.9799 | 0 | 0 | 1 | 3 | 6 | 0 | 2.1 | 2 | 10440 | 55.717055 | 11.711426 | 8.01 | -0.02 | 16.66 | 9.67 | 5.59 | 15.27 |
| 34 | DENM7_77 | Denmark | Male | 666.53 | 0 | 0 | 1 | 3 | 6 | 0 | 2.1 | 2.5 | 10440 | 55.717055 | 11.711426 | 8.01 | -0.02 | 16.66 | 9.67 | 5.59 | 15.27 |
| 35 | DENM1_58 | Denmark | Male | 627.4583 | 0 | 0 | 1 | 3 | 6 | 0 | 2.1 | 1.5 | 10440 | 55.717055 | 11.711426 | 8.01 | -0.02 | 16.66 | 9.67 | 5.59 | 15.27 |
| 36 | DENM903 | Denmark | Male | 662.5553 | 0 | 0 | 1 | 3 | 6 | 0 | 2.1 | 2 | 10440 | 55.717055 | 11.711426 | 8.01 | -0.02 | 16.66 | 9.67 | 5.59 | 15.27 |
| 37 | DENM901 | Denmark | Male | 677.8408 | 0 | 0 | 1 | 3 | 6 | 0 | 2.1 | NaN | 10440 | 55.717055 | 11.711426 | 8.01 | -0.02 | 16.66 | 9.67 | 5.59 | 15.27 |
| 38 | DENF1552 | Denmark | Female | 624.4664 | 0 | 0 | 1 | 3 | 6 | 0 | 2.1 | 0.5 | 10440 | 55.717055 | 11.711426 | 8.01 | -0.02 | 16.66 | 9.67 | 5.59 | 15.27 |

INFO | Shape_rawcoordinates | Shape_symmProcCoord. | Factors | Background | Geodistance_extra

# Example II: Result

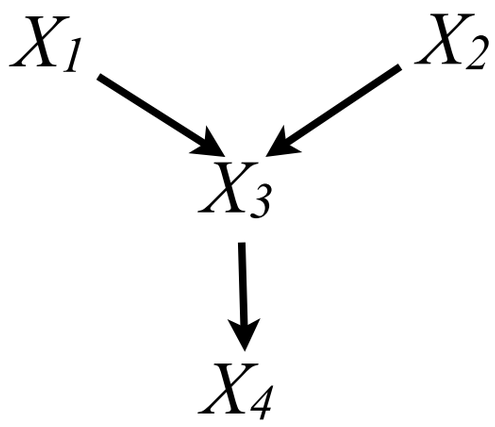
*Thanks to collaborator Marlijn Noback*

- 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)

# How about This Case?

$$X_1 \perp\!\!\!\perp X_2;$$
$$X_1 \perp\!\!\!\perp X_4 \,|\, X_3;$$
$$X_2 \perp\!\!\!\perp X_4 \,|\, X_3.$$

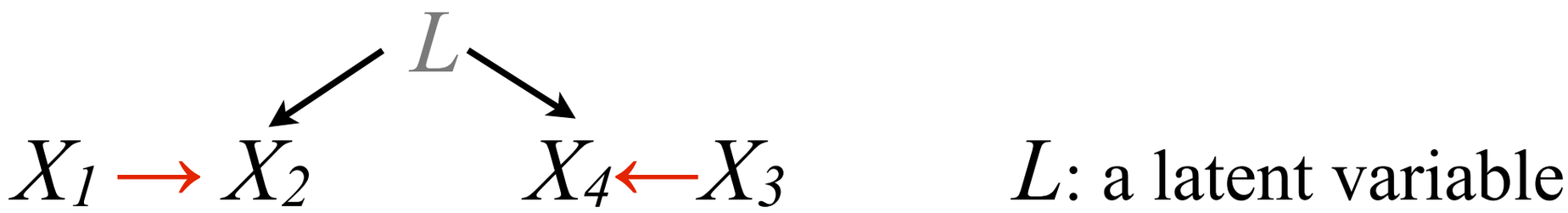
*What is corresponding causal structure? Possible to have confounders behind $X_3$ and $X_4$?*




:-)

# How about This Case?

$$X_1 \perp\!\!\!\perp X_3;$$
$$X_1 \perp\!\!\!\perp X_4;$$
$$X_2 \perp\!\!\!\perp X_3.$$

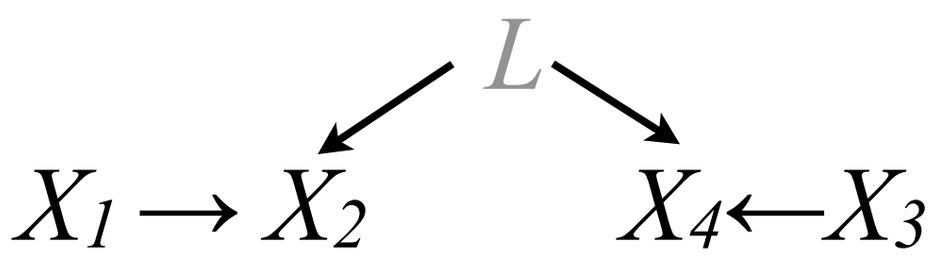



- There must exist some confounder for $X_2$ and $X_4$.

- 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** ?

# 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




*What's FCI's output?*

Data available in
'Illust_FCI_4variables.txt'

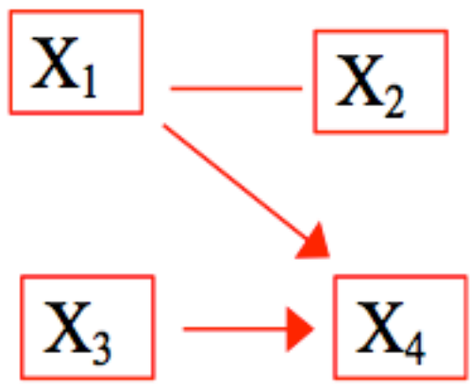
*Spirtes et al., Causal inference in the presence of latent variables and selection bias, 1997*

# (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)

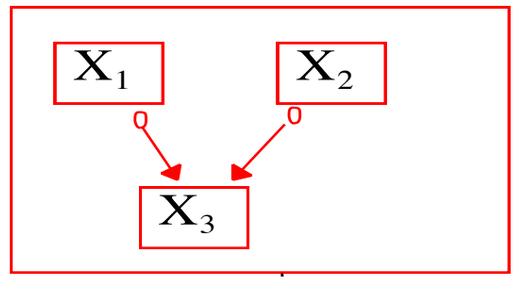
| | Possible Edges |
|---|---|
| $X_1$ and $X_2$ are not adjacent in any member of the equivalent class | $X_1$ $X_2$ |
| $X_1 \rightarrow X_2$ in some members of the equivalent class, and $X_1 \leftarrow X_2$ in some others | $X_1$ — $X_2$ |
| $X_1 \rightarrow X_2$ in every member of the equivalent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

Example




*How many DAGs in this class?*

# PAGs: What Edges Mean?




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

$X_1$ $\circ\!\!\rightarrow$ $X_2$ $X_2$ is not an ancestor of $X_1$

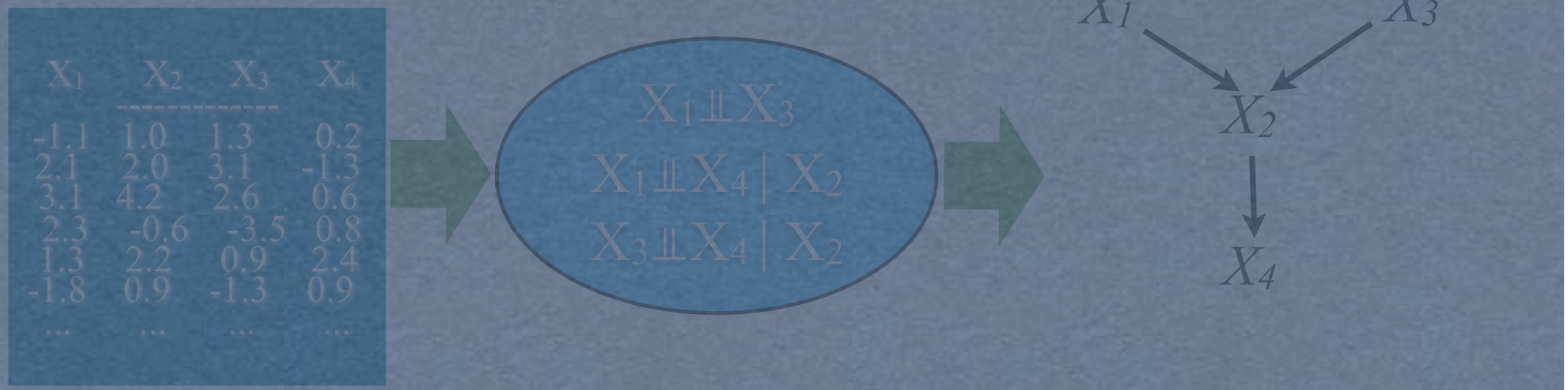
$X_1$ $\circ\!\!-\!\!\circ$ $X_2$ No set d-separates $X_2$ and $X_1$

$X_1$ $\rightarrow$ $X_2$ $X_1$ is a cause of $X_2$

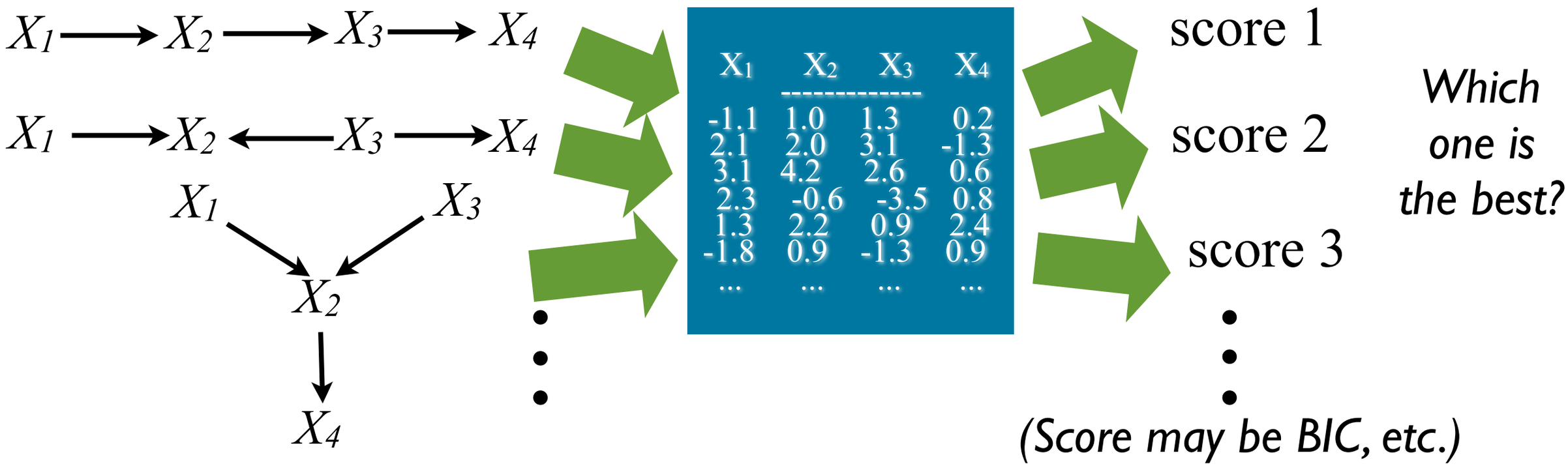
$X_1$ $\leftrightarrow$ $X_2$ There is a latent common cause of $X_1$ and $X_2$

# Constraint-Based vs. Score-Based

- Constraint-based methods




- Score-based methods



# Key Issues

- What score to use?
- How to traverse the search space of the graph?
  - DAGs? Equivalence classes?
  - How to do optimization?

# 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: $S_B(\mathcal{G}, \mathbf{D}) = \log p(\mathbf{D}|\hat{\boldsymbol{\theta}}, \mathcal{G}^h) - \frac{d}{2} \log m$

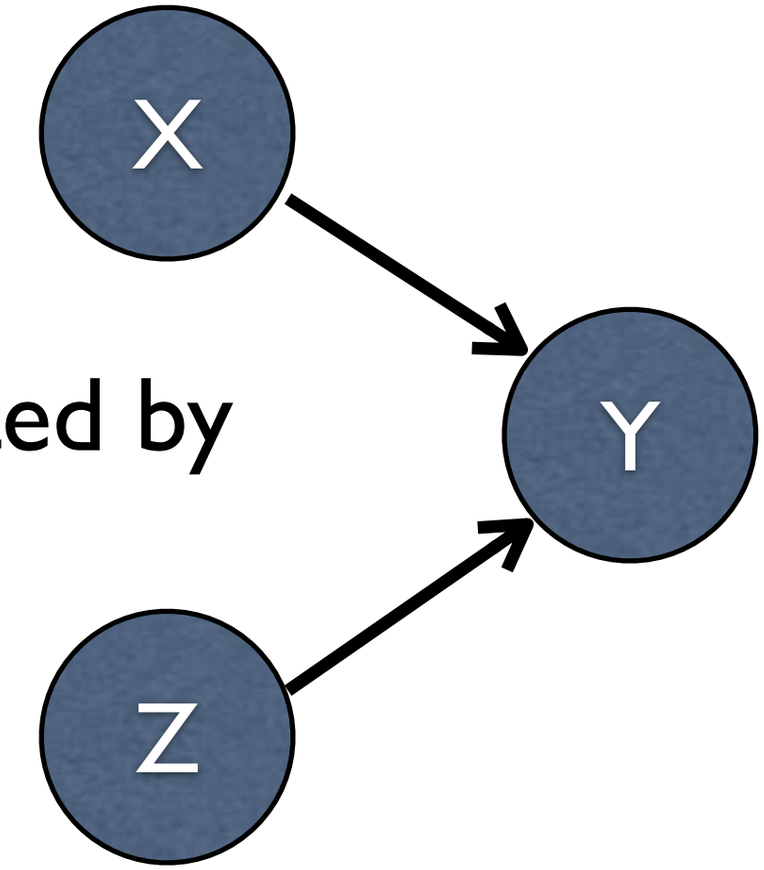
*Chickering, Optimal Structure Identification With Greedy Search, Journal of Machine Learning Research, 2002*

# GES: Search Procedure

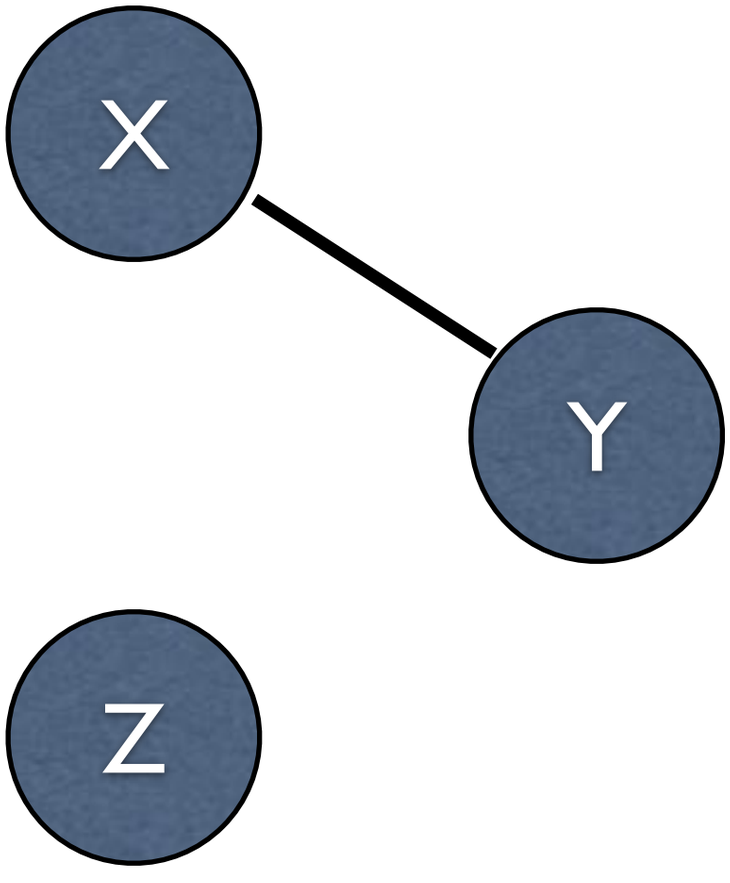
- Performs forward (addition) / backward (deletion) equivalence search through the space of DAG equivalence classes
  - Forward Greedy Search (FGS)
    - Start from some (sparse) pattern (usually the empty graph)
    - Evaluate all possible patterns with one more adjacency that entail strictly fewer 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 the Forward Stage
    - 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

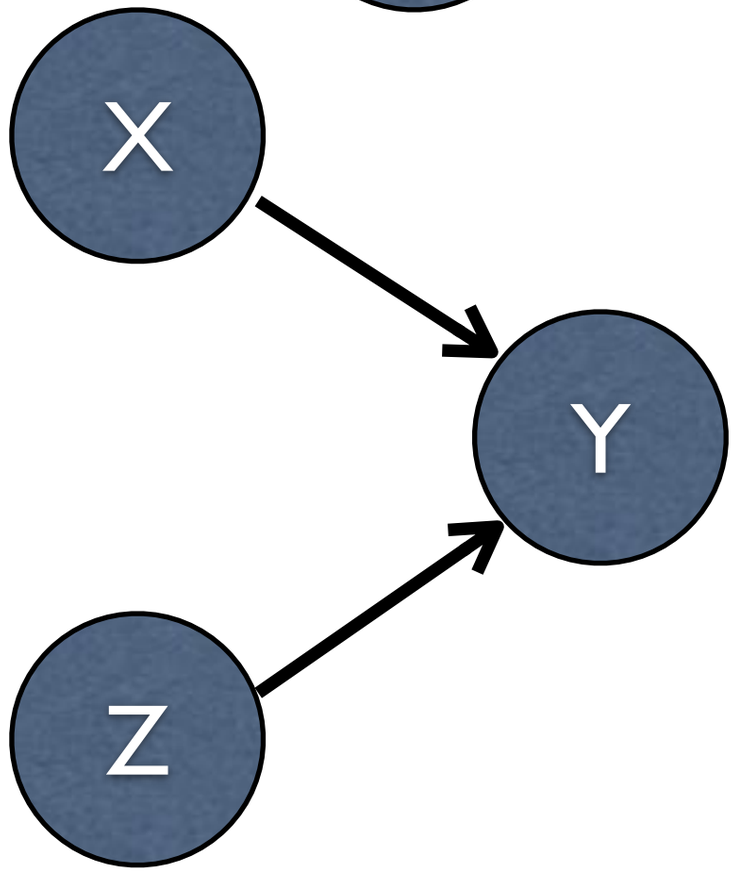
Suppose data were generated by




(1)

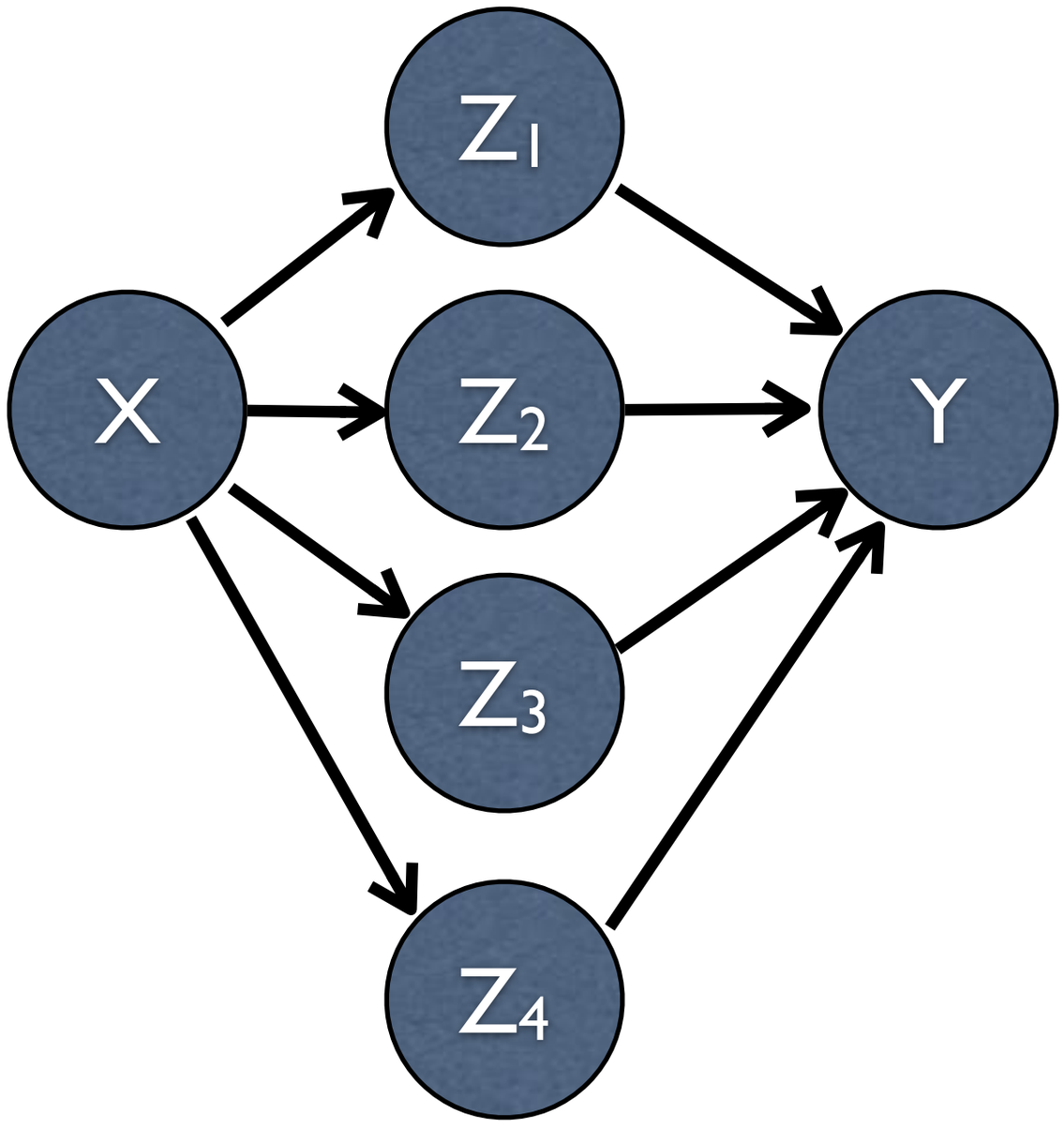



(2)



# GES

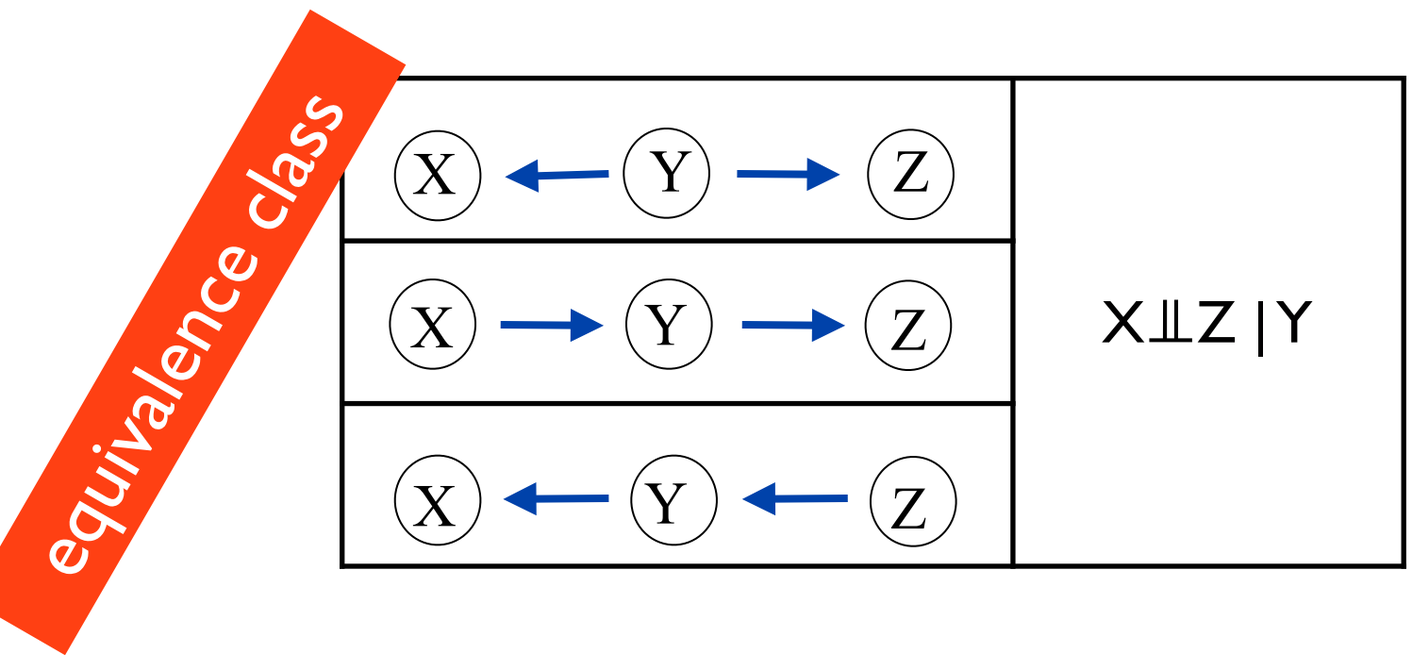
Suppose data were generated by

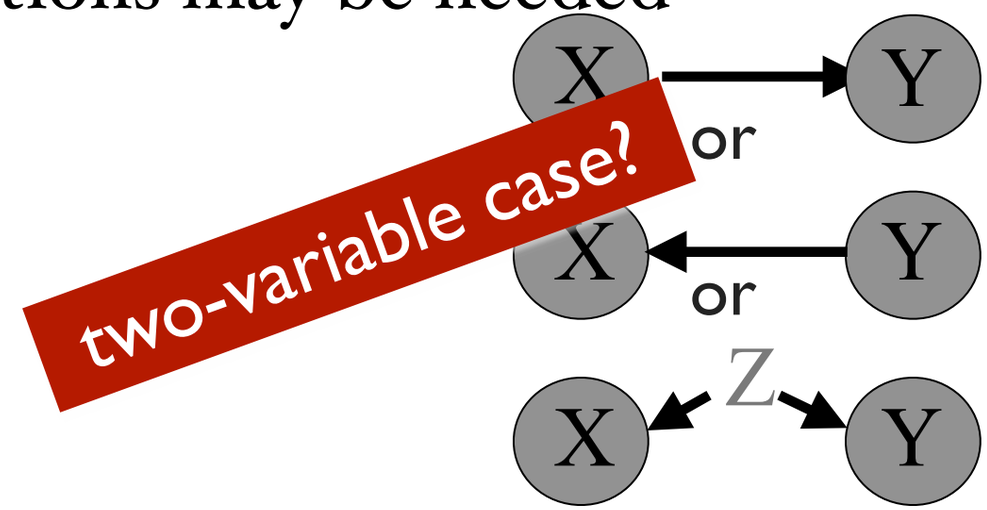



*Imagine how the GES procedure works…*

# Constraint-based Causal Discovery: Advantages and Limitations

- Nonparametric; widely applicable given reliable conditional independence tests
- Recovering {causal relations} from {conditional independences}: bounded by the equivalence class
- Directly characterize and recover cause-effect relationships?
- additional weak and reasonable assumptions may be needed

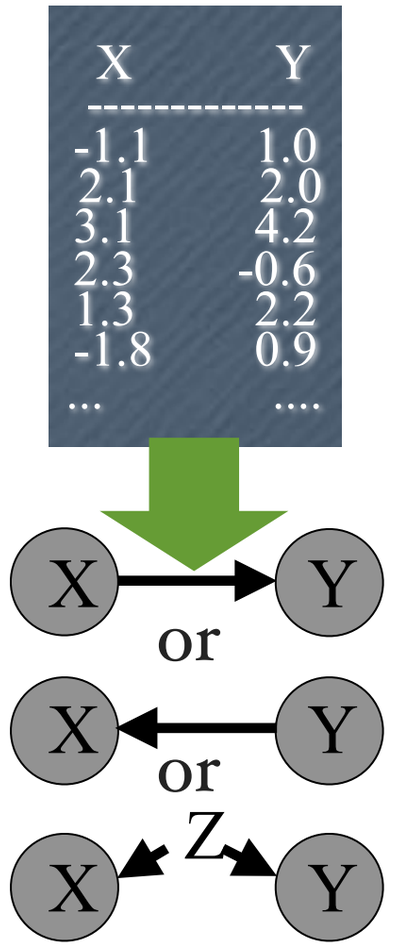


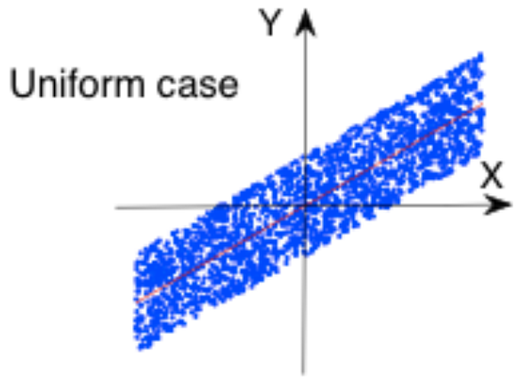




- Instead, try to directly identify local causal structures with functional causal models/structural equation models

# Outline

- Causal thinking
- Identification of causal effects
- **Causal discovery**
  - Constraint-based approach
  - **Non-Gaussian or nonlinear methods**
  - Extensions

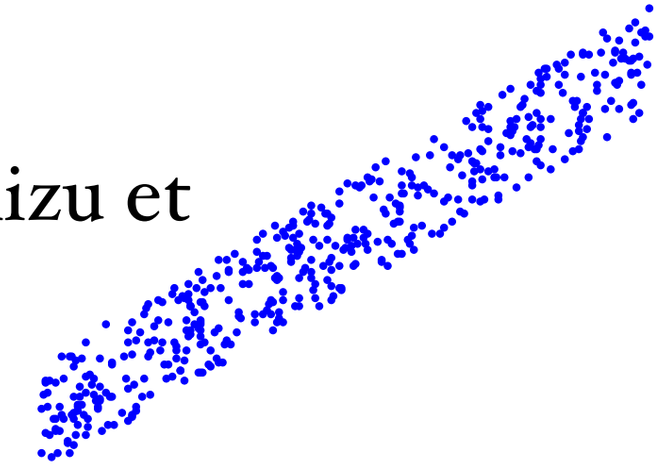



# To Fully Identify Causal Structure with Functional Causal Models

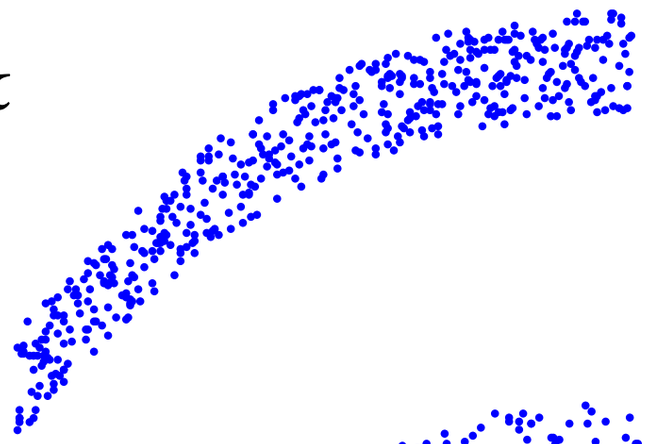
- Linear non-Gaussian acyclic causal model (Shimizu et al., '06)

$$Y = a \cdot X + E$$



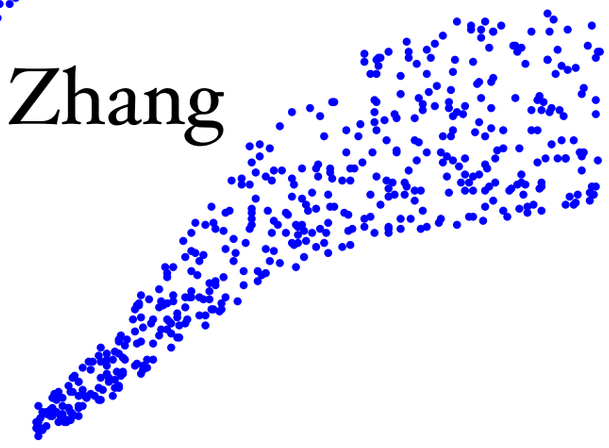
- Additive noise model (Hoyer et al., '09; Zhang & Hyvärinen, '09b)

$$Y = f(X) + E$$



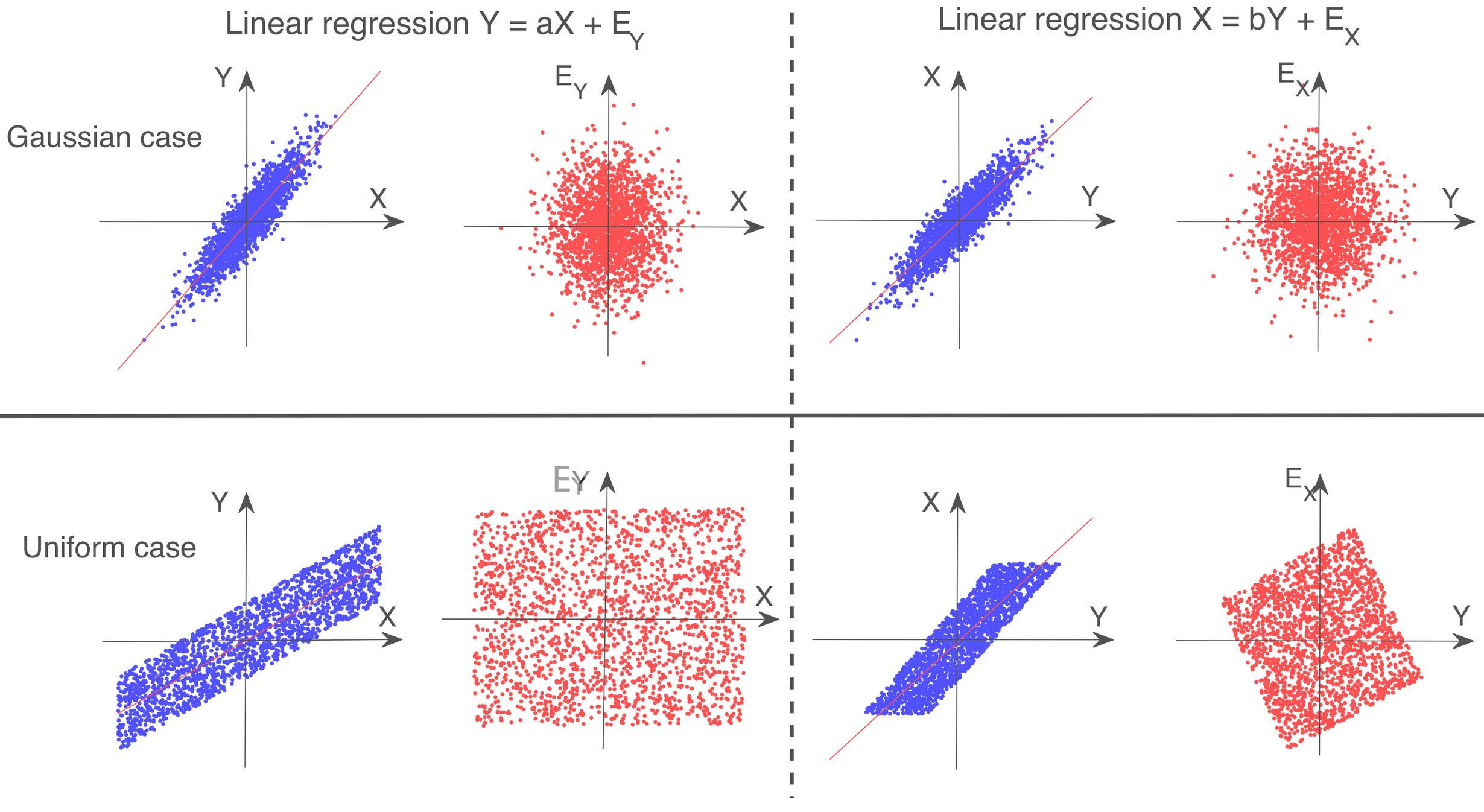
- Post-nonlinear causal model (Zhang & Chan, '06; Zhang & Hyvärinen, '09a)

$$Y = f_2 ( f_1(X) + E )$$



# Causal Asymmetry in the Linear Case: Illustration

Data generated by $Y = aX + E$ (i.e., $X \rightarrow Y$):



# More Generally, LiNGAM Model

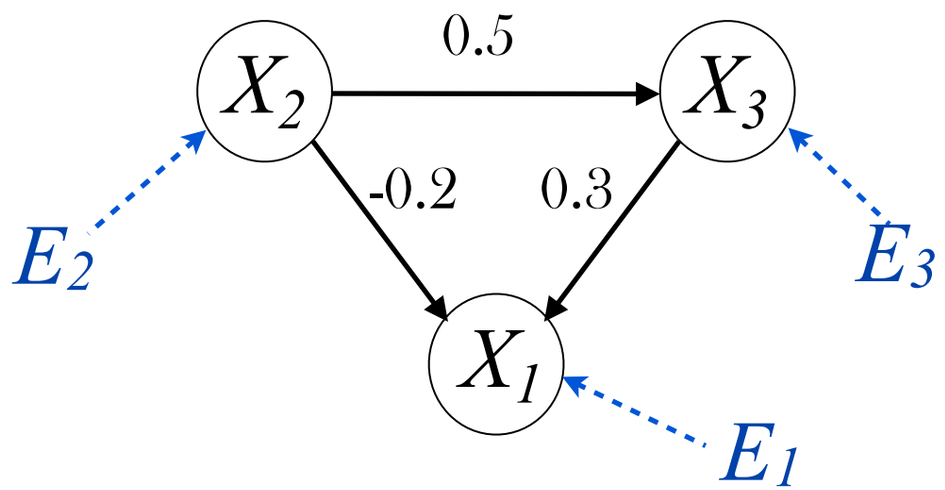
- Linear, non-Gaussian, acyclic causal model (LiNGAM) (Shimizu et al., 2006):

$$X_i = \sum_{j:\ \text{parents of } i} b_{ij} X_j + E_i \quad or \quad \mathbf{X} = \mathbf{B}\mathbf{X} + \mathbf{E}$$

- Disturbances (errors) $E_i$ are non-Gaussian (or at most one is Gaussian) and mutually independent

- Example:

$$X_2 = E_2,$$
$$X_3 = 0.5X_2 + E_3,$$
$$X_1 = -0.2X_2 + 0.3X_3 + E_1.$$




*Shimizu et al. (2006). A linear non-Gaussian acyclic model for causal discovery. Journal of Machine Learning Research, 7:2003–2030.*

# Darmois-Skitovich Theorem & Identifiability of Causal Direction in Two-Variable Case

**Darmois-Skitovitch theorem**: Define two random variables, $Y_1$ and $Y_2$, as linear combinations of independent random variables $S_i$, $i = 1, ..., n$:

$$Y_1 = \alpha_1 S_1 + \alpha_2 S_2 + ... + \alpha_n S_n,$$
$$Y_2 = \beta_1 S_1 + \beta_2 S_2 + ... + \beta_n S_n.$$

If $Y_1$ and $Y_2$ are statistically independent, then all variables $S_j$ for which $\alpha_j \beta_j \neq 0$ are Gaussian.

Generated by $Y = aX + E$ ($X \to Y$):

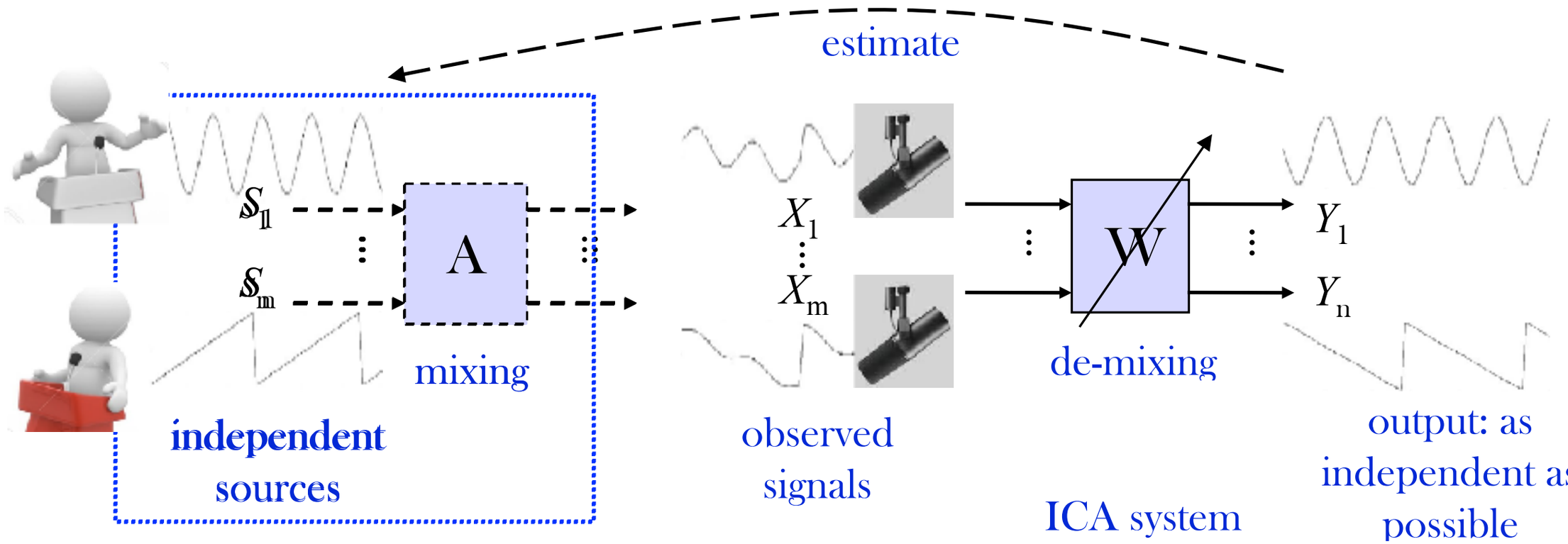
$$\begin{bmatrix} 0 & 1 \\ 1 & a \end{bmatrix} \cdot \begin{bmatrix} E \\ X \end{bmatrix} = \begin{bmatrix} X \\ Y \end{bmatrix} \Rightarrow \begin{bmatrix} E_Y \\ Y \end{bmatrix} = \begin{bmatrix} 1 - ab & -b \\ a & 1 \end{bmatrix} \cdot \begin{bmatrix} E \\ X \end{bmatrix}$$

Assuming $Y \to X$ (fitting $X = bY + E_Y$):

$$\begin{bmatrix} 1 & b \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} E_Y \\ Y \end{bmatrix} =$$

*Kagan et al., Characterization Problems in Mathematical Statistics. New York: Wiley, 1973*

# Independent Component Analysis




$$\mathbf{X} = \mathbf{A} \cdot \mathbf{S}$$

$$\mathbf{Y} = \mathbf{W} \cdot \mathbf{X}$$

$$\begin{matrix} X_1 \\ X_2 \end{matrix} \begin{bmatrix} .5 & .3 & 1.1 & -0.3 & \dots \\ .8 & -.7 & .3 & .5 & \dots \end{bmatrix} = \begin{bmatrix} ? & ? \\ ? & A & ? \end{bmatrix} \cdot \begin{bmatrix} ? & ? & ? & ? & \dots \\ ? & ? & ? & ? & \dots \end{bmatrix} \begin{matrix} s_1 \\ s_2 \end{matrix}$$

- Assumptions in ICA
  - At most one of $S_i$ is Gaussian
  - #Source >= # Sensor, and **A** is of full column rank

Then A can be estimated up to column scale and permutation indeterminacies

*Hyvärinen et al., Independent Component Analysis, 2001*

# Gaussianity or Non-Gaussianity?

- Non-Gaussianity is actually ubiquitous
  - Linear closure property of Gaussian distribution: If the sum of any finite independent variables is Gaussian, then all summands must be Gaussian (Cramér, 1970)
  - Gaussian distribution is "special" in the linear case
- Practical issue: How non-Gaussian they are?

# LiNGAM Analysis by ICA

- LiNGAM: $X_i = \sum_{j:\text{ parents of } i} b_{ij} X_j + E_i \quad or \quad \mathbf{X} = \mathbf{BX} + \mathbf{E} \Rightarrow \mathbf{E = (I-B)X}$
  - **B** has special structure: acyclic relations
- ICA: **Y = WX**
- **B** can be seen from **W** [illegible] and re-scaling
- Faithfulness assumption avoided

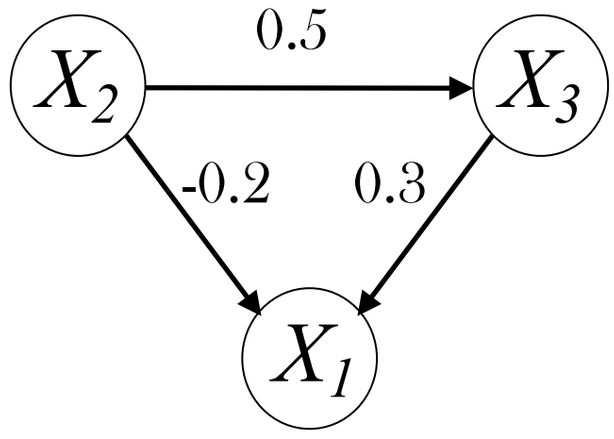
Question 1. How to find W?

Question 2. How to see B from W?

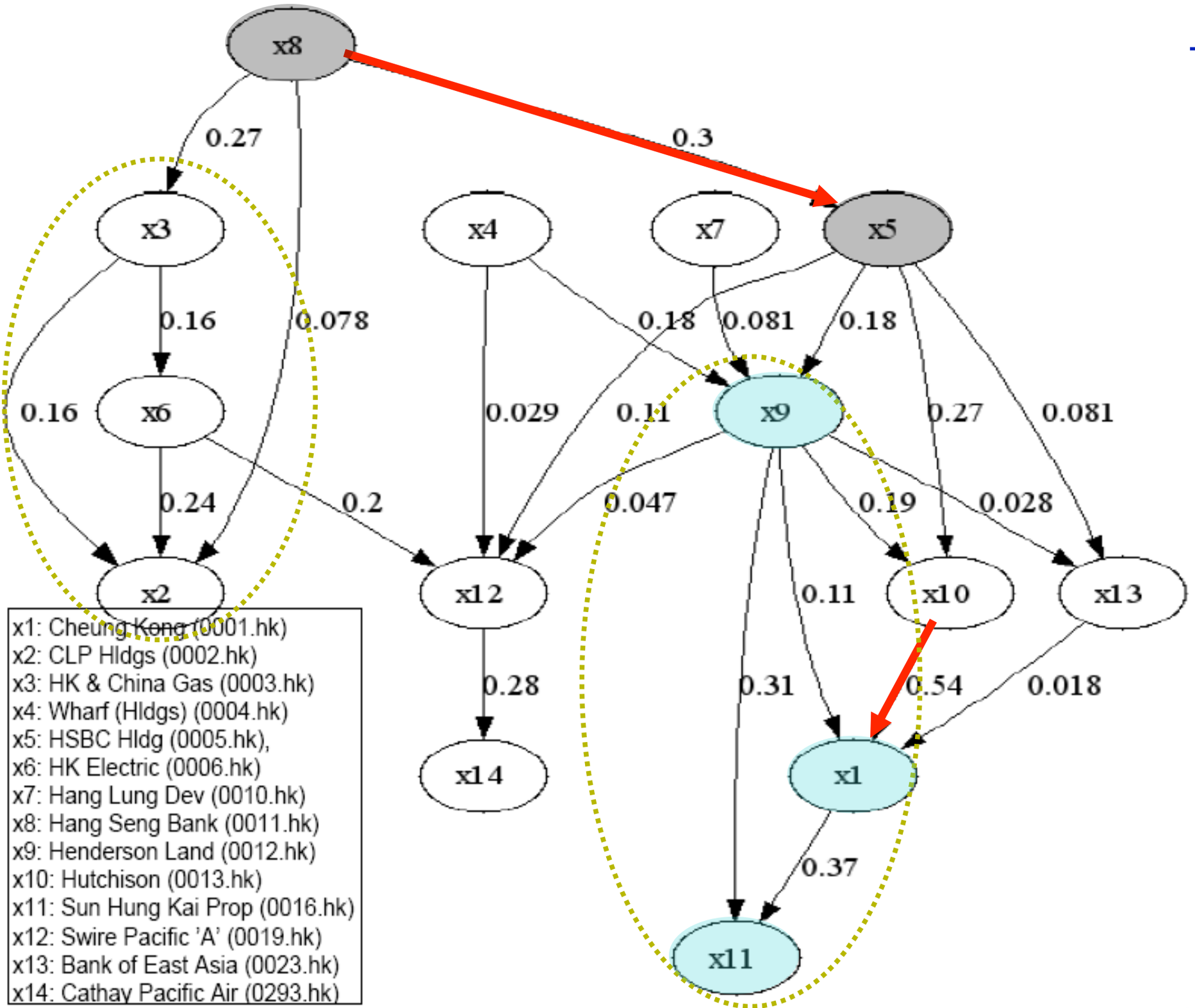
- E.g., $\begin{bmatrix} E_1 \\ E_3 \\ E_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ -0.5 & 1 & 0 \\ 0.2 & -0.3 & 1 \end{bmatrix} \cdot \begin{bmatrix} X_2 \\ X_3 \\ X_1 \end{bmatrix}$

$$\Leftrightarrow \begin{cases} X_2 = E_1 \\ X_3 = 0.5X_2 + E_3 \\ X_1 = -0.2X_2 + 0.3X_3 + E_2 \end{cases}$$

So we have the causal relation:



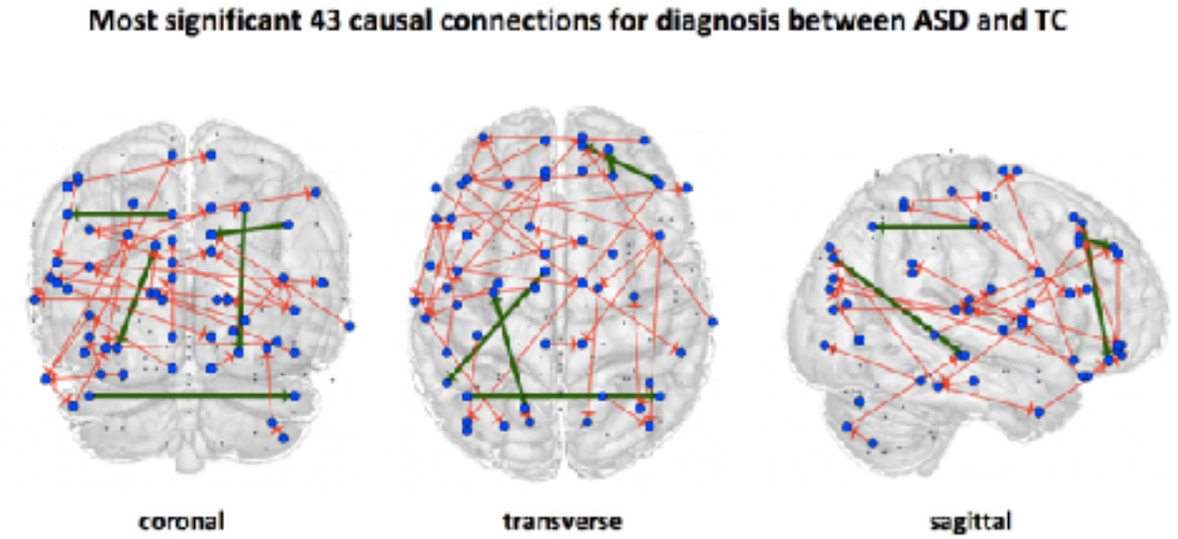
# Application: Causal diagram in HK Stock Market (Zhang & Chan, 2006)




1. Ownership relation: x5 owns 60% of x8; x1 holds 50% of x10.
2. Stocks belonging to the same subindex tend to be connected.
3. Large bank companies (x5 and x8) are the cause of many stocks.
4. Stocks in Property Index (x1, x9, x11) depend on many stocks, while they hardly influence others.

# Causality-Based ASD Diagnosis of ASD from fMRI

- Autism spectrum disorder (high rate: 1/68 children)
- Resting-state fMRI signal : ABIDE dataset (16 international sites)
- Causal graph & strength estimation
  - High dimension, small sample size, heterogeneity, etc.
- Neuropathology diagnostic biomarkers
- Prediction: 81% out-of-sample accuracy

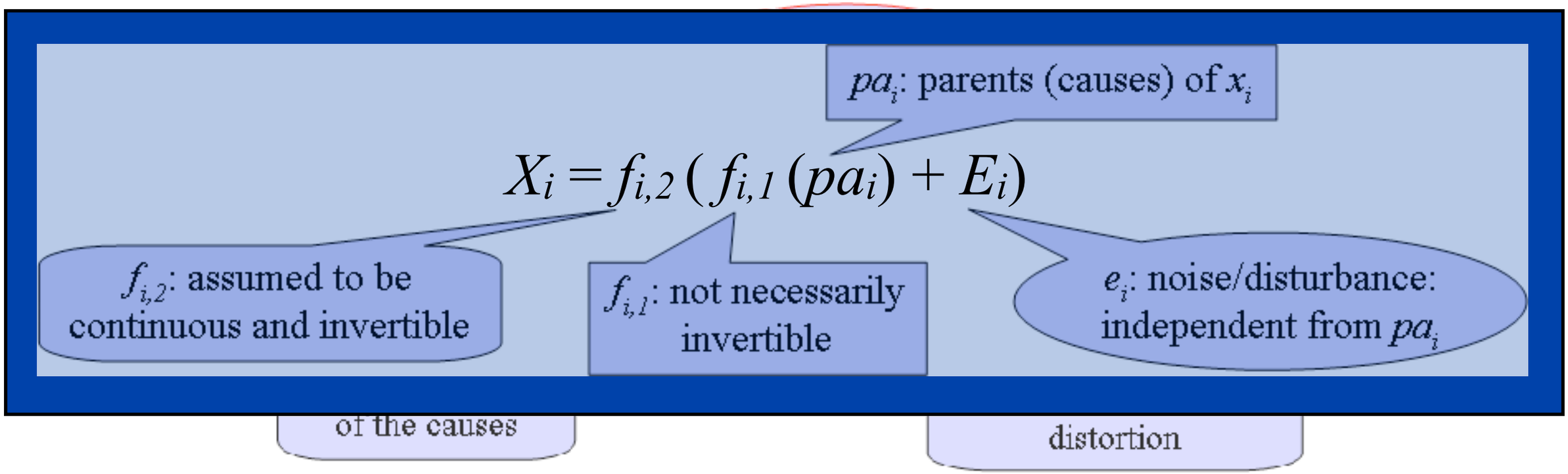



Huang, Zhang et al., Diagnosis of ASD by causal connectivity strength from fMRI, Neuroscience 2017

# Post-Nonlinear (PNL) Causal Model

(Zhang & Chan, 2006; Zhang & Hyvärinen, '09a)

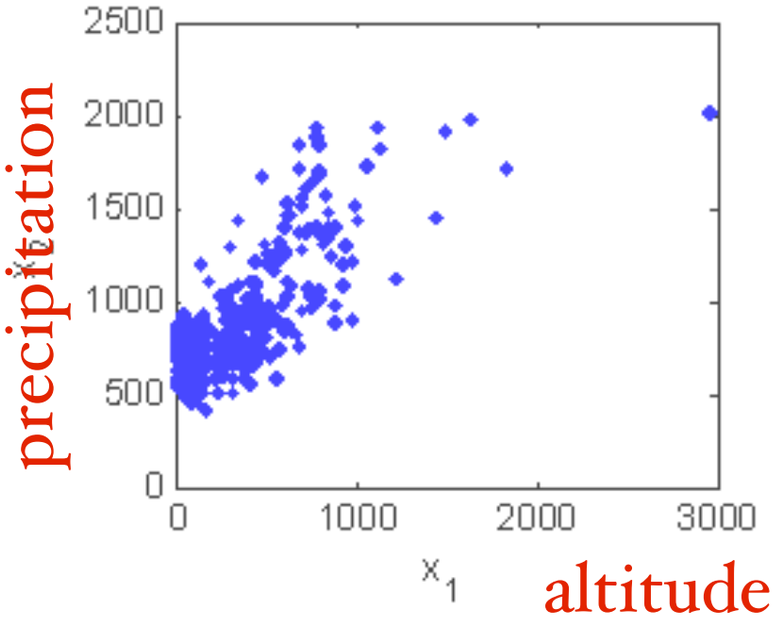
- Without prior knowledge, the assumed model is expected to be
  - general enough: adapt to approximate the true generating process
  - identifiable: asymmetry in causes and effects

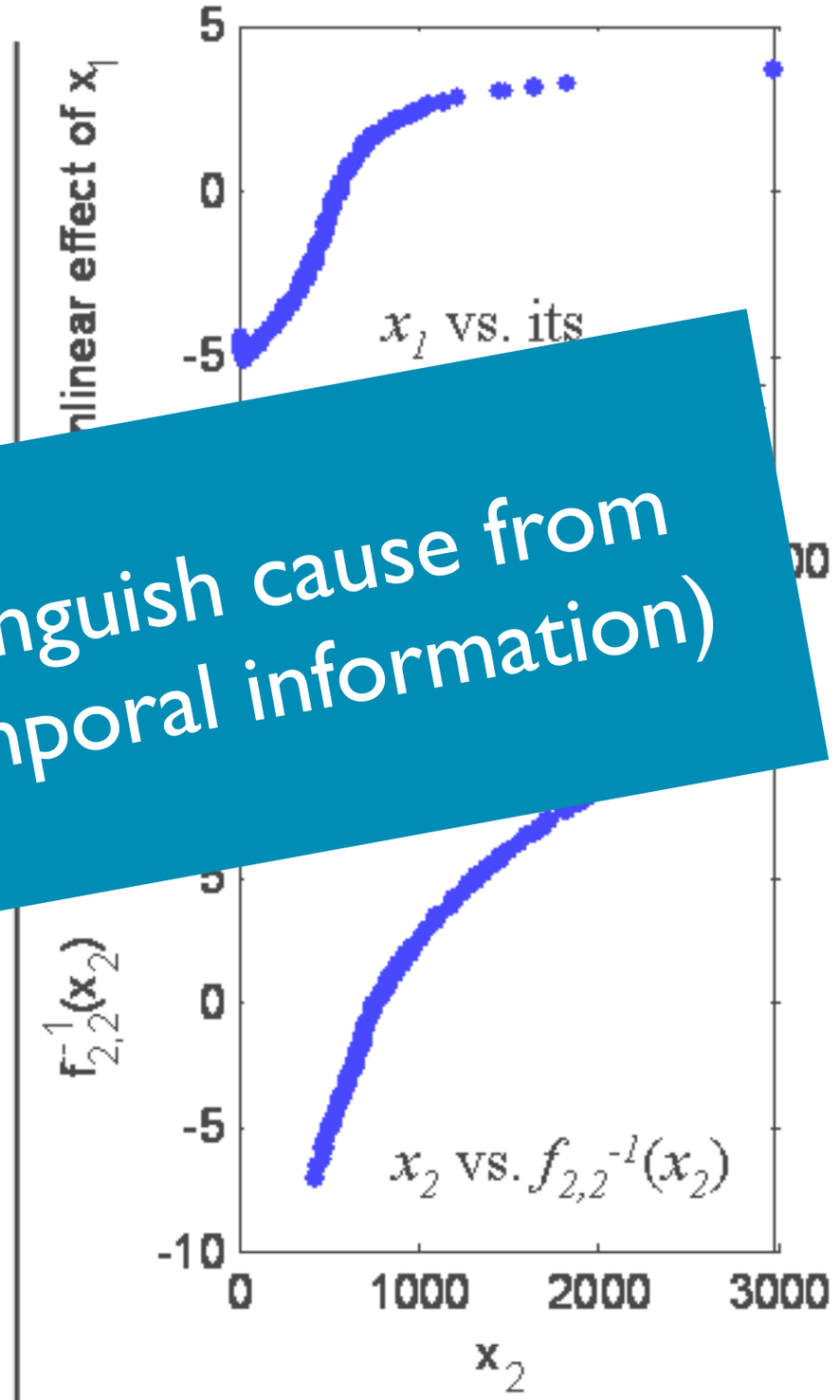



- Special cases: linear models; nonlinear additive noise models; multiplicative noise models: $Y = X \cdot E = \exp\big(\log(X) + \log(E)\big)$

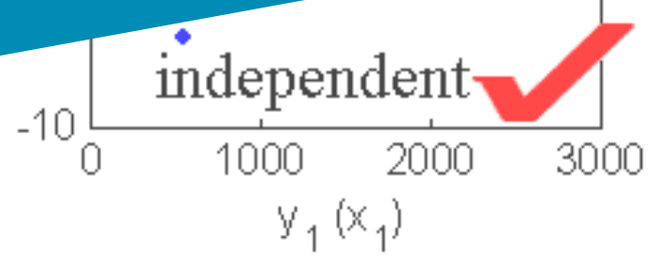
# Data Set 2

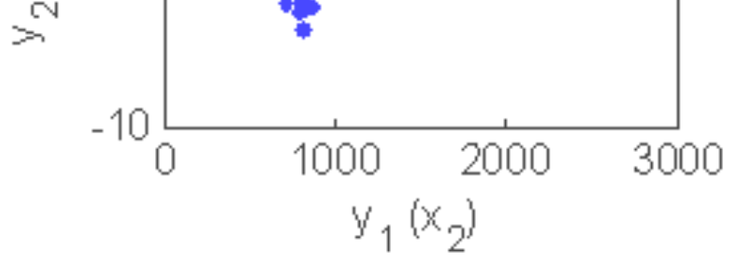
*with PNL Model*




(a) $y_1$ vs $y_2$ under hypothesis




Generally speaking, possible to distinguish cause from effect with FCMs (without using temporal information)

# Identifiability in Two-variable Case: Theoretical Results

$$X_i = f_{i,2}\left(f_{i,1}(pa_i) + E_i\right)$$

$pa_i$: parents (causes) of $x_i$

$f_{i,2}$: assumed to be continuous and invertible

$f_{i,1}$: not necessarily invertible

$e_i$: noise/disturbance: independent from $pa_i$

- Two-variable case: if $X_1 \rightarrow X_2$, then $X_2 = f_{2,2}\left(f_{2,1}(X_1) + E_2\right)$
- Is the causal direction implied by the model unique?
- By a proof of contradiction
  - Assume both $X_1 \rightarrow X_2$ and $X_2 \rightarrow X_1$ satisfy PNL model
  - One can then find all non-identifiable cases

# Identifiability: A Mathematical Result

- **Theorem 1**
  - Assume $x_2 = f_2(f_1(x_1) + e_2)$,
    $x_1 = g_2(g_1(x_2) + e_1)$,
  - Further suppose that involved densities and nonlinear functions are third-order differentiable, and that $p_{e2}$ is unbounded,
  - For every point satisfying $\eta_2'' h' \neq 0$, we have

$$\eta_1''' - \frac{\eta_1'' h''}{h'} = \left(\frac{\eta_2' \eta_2'''}{\eta_2''} - 2\eta_2''\right) \cdot h' h'' - \frac{\eta_2'''}{\eta_2''} \cdot h' \eta_1'' + \eta_2' \cdot \left(h''' - \frac{h''^2}{h'}\right).$$

**Notation**

$t_1 \triangleq g_2^{-1}(x_1), \quad z_2 \triangleq f_2^{-1}(x_2),$
$h \triangleq f_1 \circ g_2, \quad h_1 \triangleq g_1 \circ f_2.$
$\eta_1(t_1) \triangleq \log p_{t_1}(t_1), \quad \eta_2(e_2) \triangleq \log p_{e_2}(e_2).$

- Obtained by using the fact that the Hessian of the logarithm of the joint density of independent variables is diagonal everywhere (Lin, 1998)
- It is not obvious if this theorem holds in practice…

# All Non-Identifiable Cases

$$x_2 = f_2(f_1(x_1) + e_2)$$

$$x_1 = g_2(g_1(x_2) + e_1)$$

Log-mixed-linear-and-exponential:

$$\log p_v = c_1 e^{c_2 v} + c_3 v + c_4$$

$(\log p_v)' \to c \ (c \neq 0)$,

as $v \to -\infty$ or as $v \to +\infty$

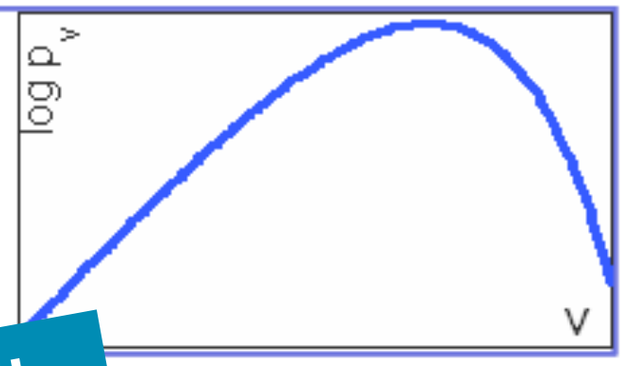



Table 1: All situations in wh… not identifiable.

| | $p_{e_2}$ | Remark |
|---|---|---|
| I | Gaussian | $h_1$ also linear |
| II | log-mix-lin-exp | $h_1$ strictly monotonic, and $h_1' \to$ …, as $z_2 \to +\infty$ or as $z_2 \to -\infty$ |
| III | log-mix-lin-exp | |
| IV | log-mix-lin-exp | |
| V | generalized mixture of two exponentials | |

$$p_v \propto (c_1 e^{c_2 v} + c_3 e^{c_4 v})^{c_5}$$

Causal direction is generally identifiable if the data were generated according to

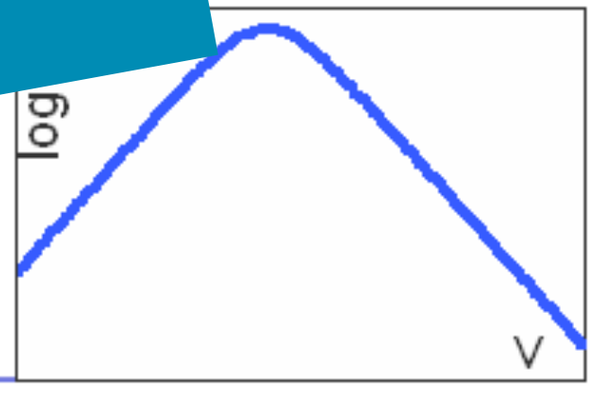
$$X_2 = f_2(f_1(X_1) + E).$$

Special cases: $X_2 = a \cdot X_1 + E$ and $X_2 = g(X) + E$.

… and

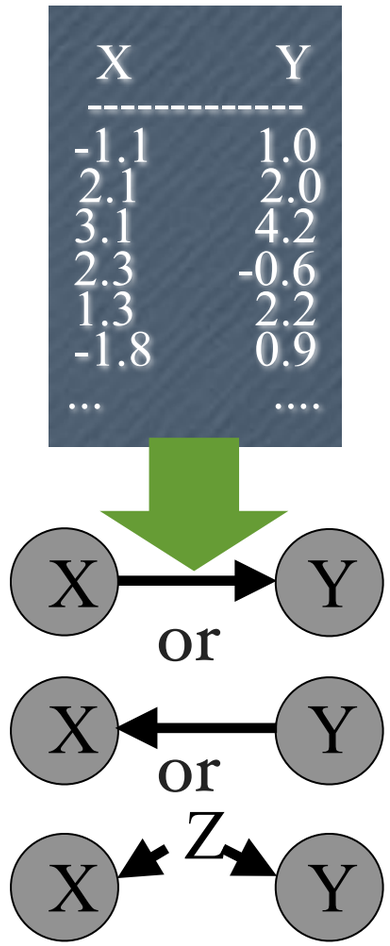
$(\log p_v)' \to c_2 \ (c_2 \neq 0)$,

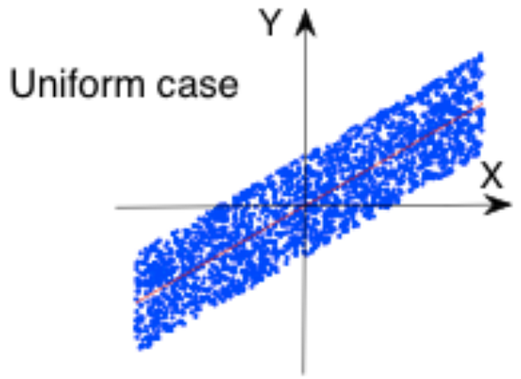
as $v \to +\infty$



# Outline

- Causal thinking
- Identification of causal effects
- **Causal discovery**
  - Constraint-based approach
  - Non-Gaussian or nonlinear methods
  - **Extensions**



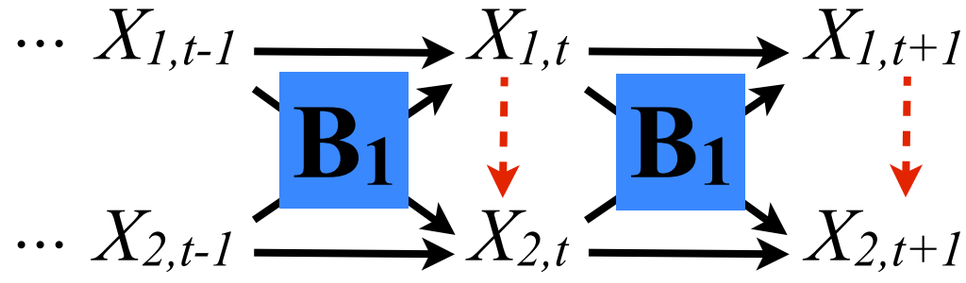




# Practical Issues in Causal Discovery...

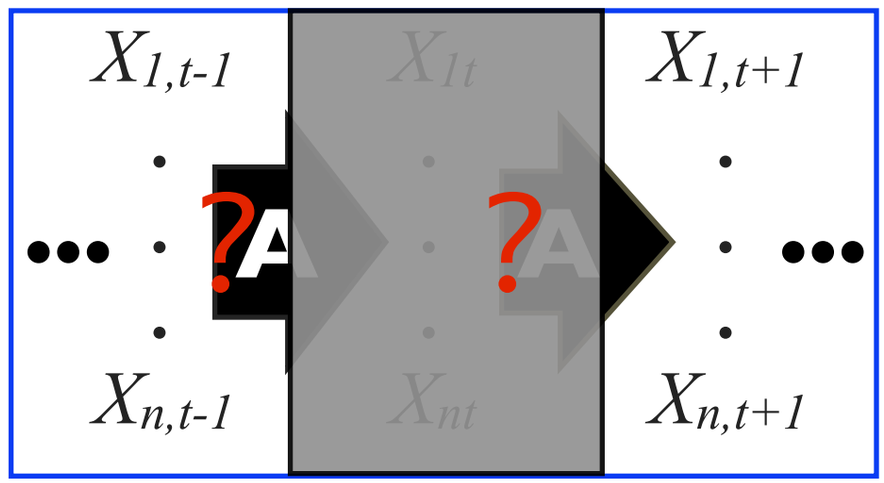
- Nonlinearities (Zhang & Chan, ICONIP'06; Hoyer et al., NIPS'08; Zhang & Hyvärinen, UAI'09; Huang et al., KDD'18)
- Categorical variables or mixed cases (Huang et al., KDD'18; Cai et al., NIPS'18)
- Nonstationary/heterogeneous data (Zhang et al., IJCAI'17; Huang et al, ICDM'17, Ghassami et al., NIPS'18)
- Measurement error (Zhang et al., UAI'18; PSA'18)
- Selection bias (Zhang et al., UAI'16)
- Confounding (SGS 1993; Hoyer et al., 2008; Zhang et al., 2018c)
- Missing values (Tu et al., AISTATS'19)
- Causality in time series
  - Time-delayed + instantaneous relations (Hyvarinen ICML'08; Zhang et al., ECML'09; Hyvarinen et al., JMLR'10)
  - Subsampling / temporally aggregation (Danks & Plis, NIPS WS'14; Gong et al., ICML'15 & UAI'17)
  - From partially observable time series (Geiger et al., ICML'15)
- Application in recommender systems (Wang et al., AAAI'18; Wang et al., NIPS'18)

# Extension 1: Causality in Time Series

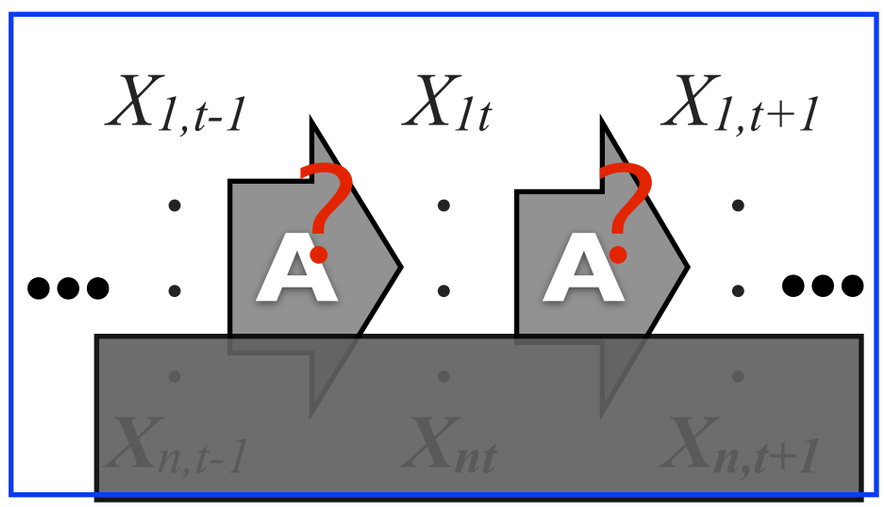
- Functional causal models in time series
  - Time-delayed causality + instantaneous relations




- Causal discovery from subsampled or temporally aggregated data




- From partially observable time series




*Zhang & Hyvärinen, ECML 2009;*
*Hyvärinen , Zhang et al., JMLR 2010;*
*Gong, Zhang, Schölkopf, Tao, Geigere, ICML 2015; UAI 2017;*
*Geiger, Zhang, Gong, Janzing, Schölkopf, ICML 2015*

# Two Schemes of Temporal Aggregation

- **Subsampling (syst**
  **sampling)**

$$\mathbf{X}$$

- T

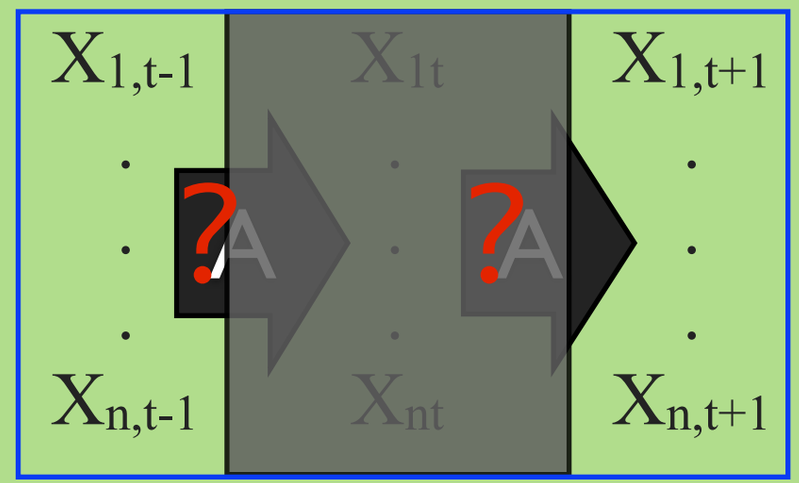
$$\mathbf{x}_1,$$

$$\tilde{\mathbf{x}}_1 = \frac{1}{k}$$

Can we recover the causal influence matrix *A*?

- Examples: temperature data, stock daily returns, GDP, fMRI...

Assume $\mathbf{X}_t = \mathbf{A}\mathbf{X}_{t-1} + \mathbf{E}_t$



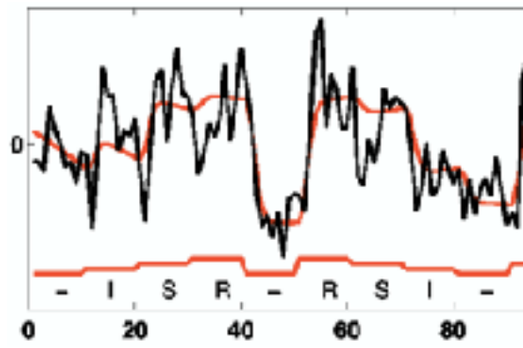

*Causal info tends to disappear as* $k \to \infty$

*Causal info tends to be instantaneous as* $k \to \infty$ :

$$\tilde{\mathbf{X}}_t \approx \mathbf{A}\tilde{\mathbf{X}}_t + \tilde{\mathbf{E}}_t$$

# Extension 2: Causal Discovery from Nonstationary/Heterogeneous Data

- Ubiquity of nonstationary/heterogeneous data
  - Nonstationary time series (brain signals, climate data…)
  - Multiple data sets under different conditions

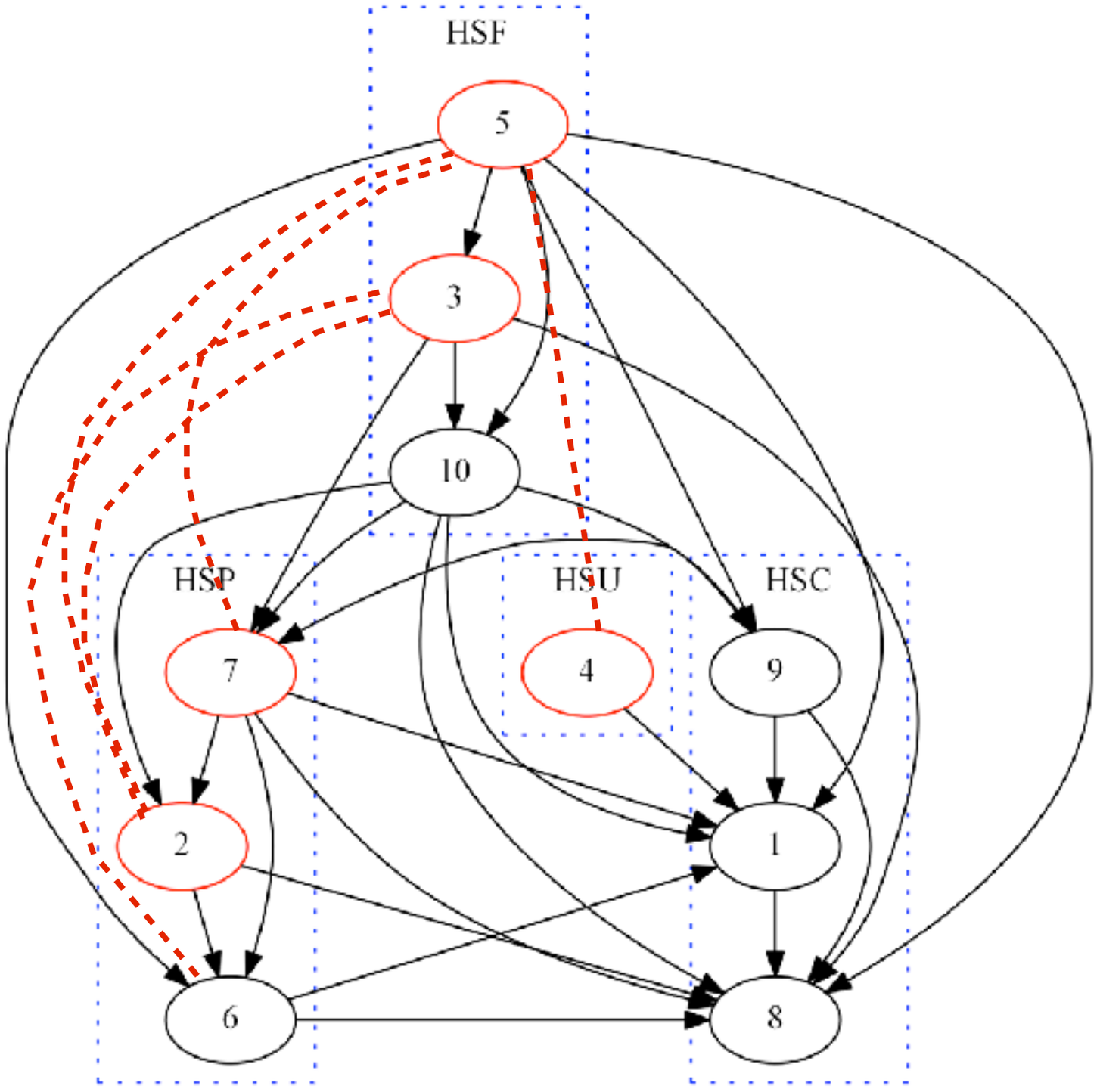



- Causal modeling & distribution shift heavily coupled
- A framework to
  - determine changing causal modules + recover skeleton
  - find causal direction based on changes
  - estimate driving force of nonstationarity

*Zhang et al., Discovery and visualization of nonstationary causal models, arxiv 2015*
*Zhang et al., Causal discovery in the presence of nonstatioarity/heterogeneity: Skeleton estimation and orientation determination, IJCAI 2017*

# Causal Analysis of Major Stocks in Hong Kong Market (10/09/2006 - 08/09/2010)

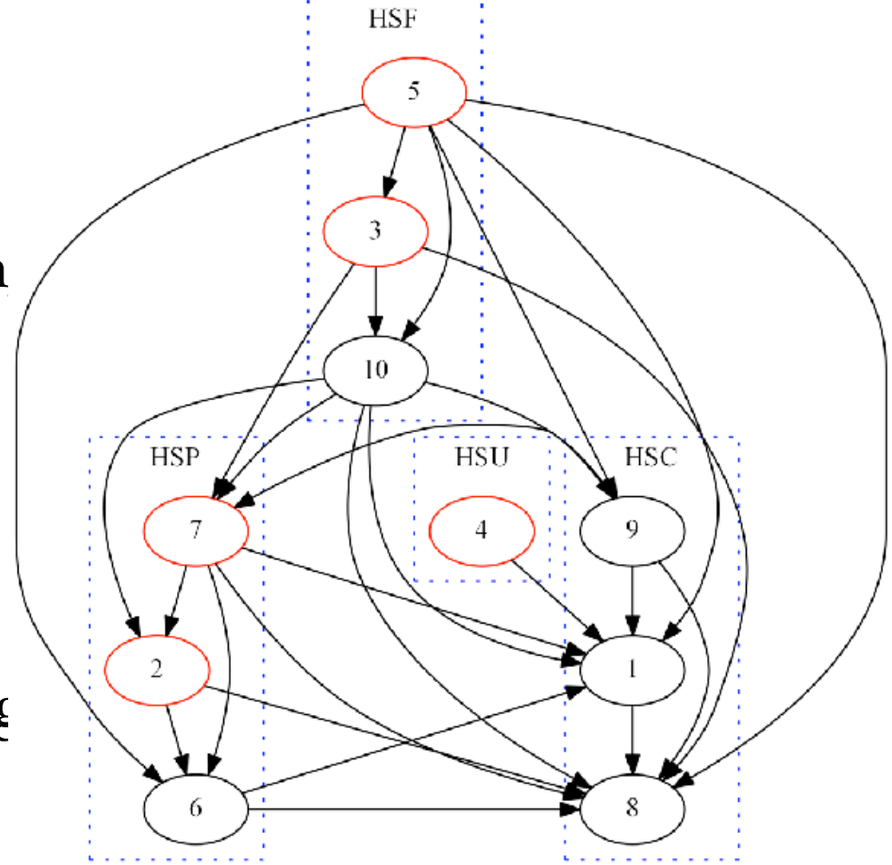
1. Cheng Kong Holdings,
2. Wharf (Holdings),
3. HSBC,
4.Hong Kong Electric Holdings,
5. Hang Seng Bank,
6. Henderson Land Dev.,
7. Sun Hung Kai Properties,
8. Swire Group,
9. Cathay Pacific Airways
10. Bank of China Hong Kong

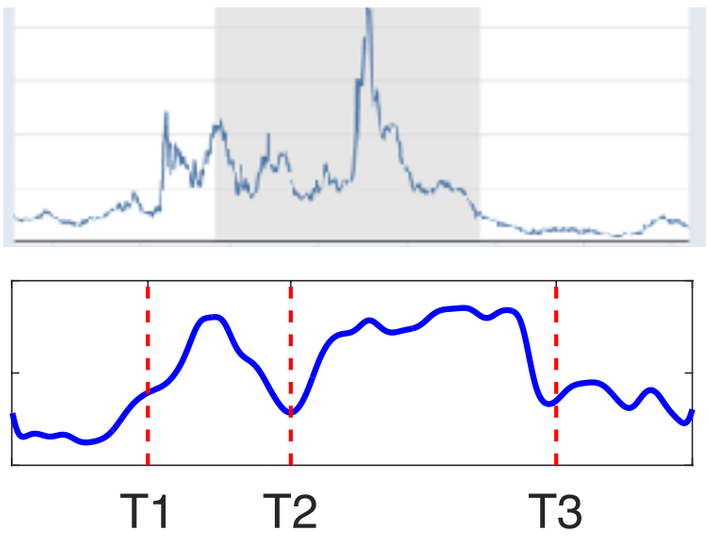



- HSF and HSP usually have nonstationary confounders

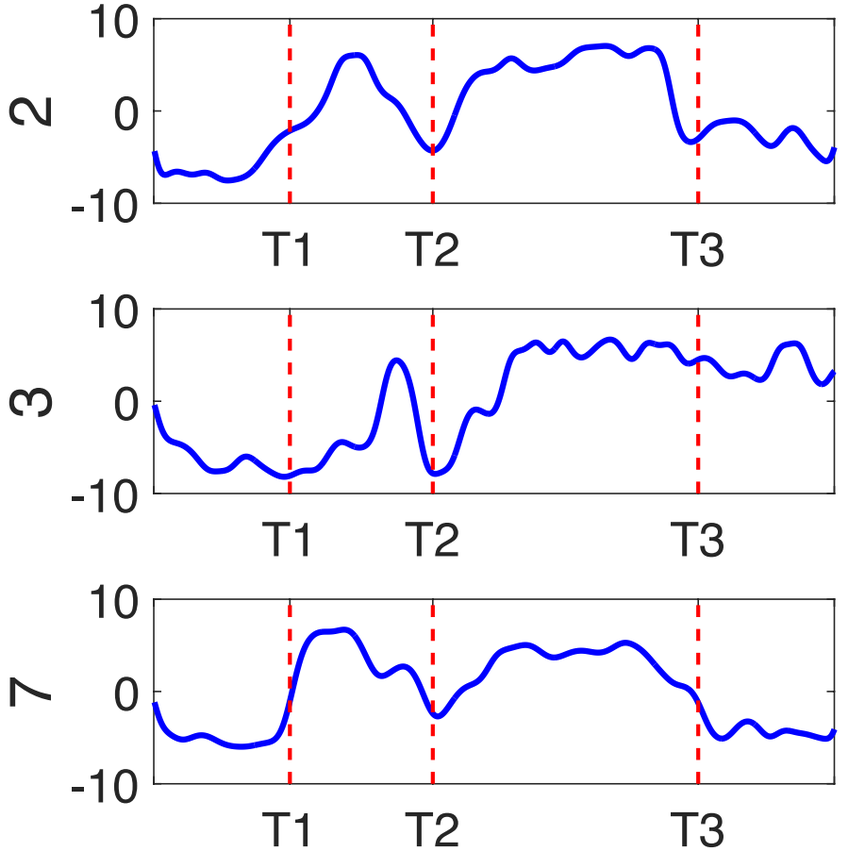
# Nonstationarity Driving Force

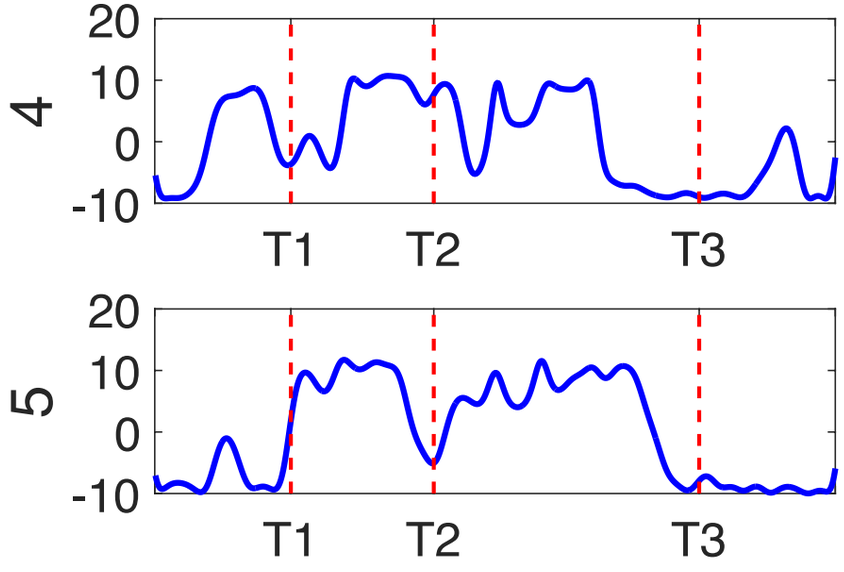
1. Cheng Kong Holdings,
2. Wharf (Holdings),
3. HSBC,
4. Hong Kong Electric Holdin
5. Hang Seng Bank,
6. Henderson Land Dev.,
7. Sun Hung Kai Properties,
8. Swire Group,
9. Cathay Pacific Airways
10. Bank of China Hong Kong



(Curve of TED spread; https://research.stlouisfed.org/fred2/series/TEDRATE)



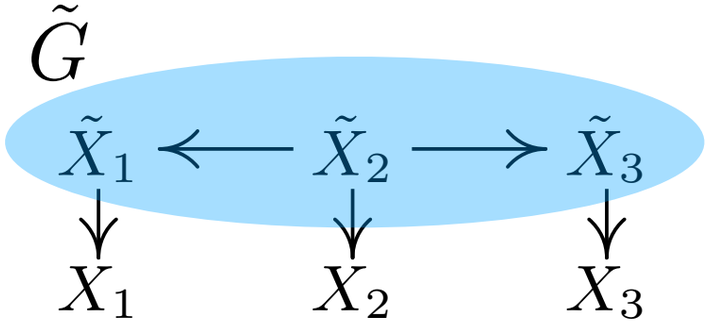

*$T_1$: 07/16/2007,*
*$T_2$: 06/30/2008,*
*$T_3$: 02/11/2009*

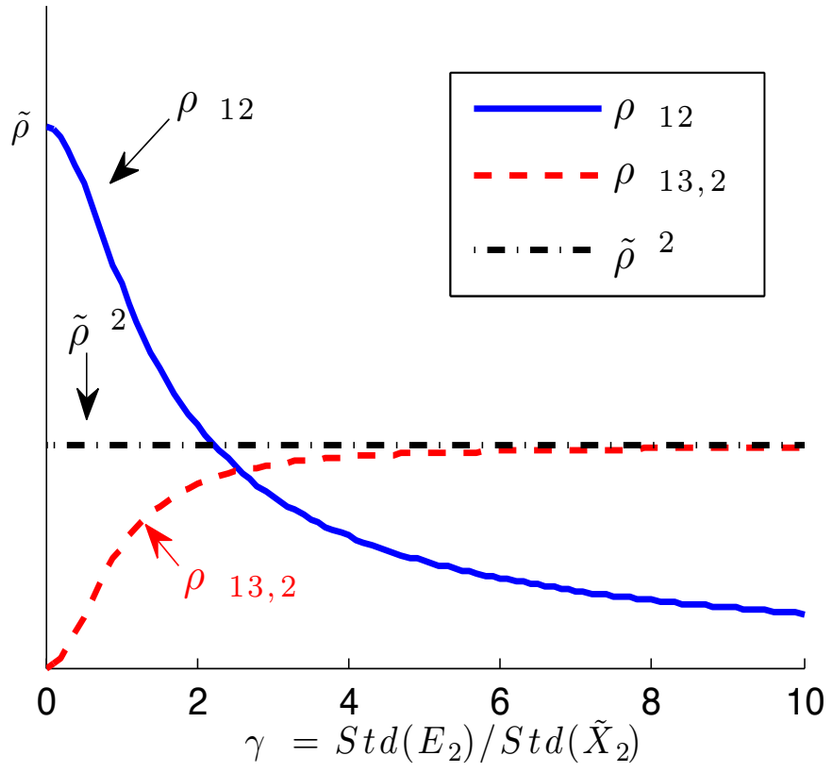
# Extension 3: Causal Discovery in the Presence of Measurement Error

- To estimate $\tilde{G}$ over variables $\tilde{X}_i$ from noisy observations $X_i = \tilde{X}_i + E_i$.
- Conditional independence/dependence relations among $X_i$ different from those among $\tilde{X}_i$

Measurement error changes causal discovery results!

- Illustration: Correlation($X_1$, $X_2$) & partial_correlation($X_1$, $X_3$ | $X_2$)



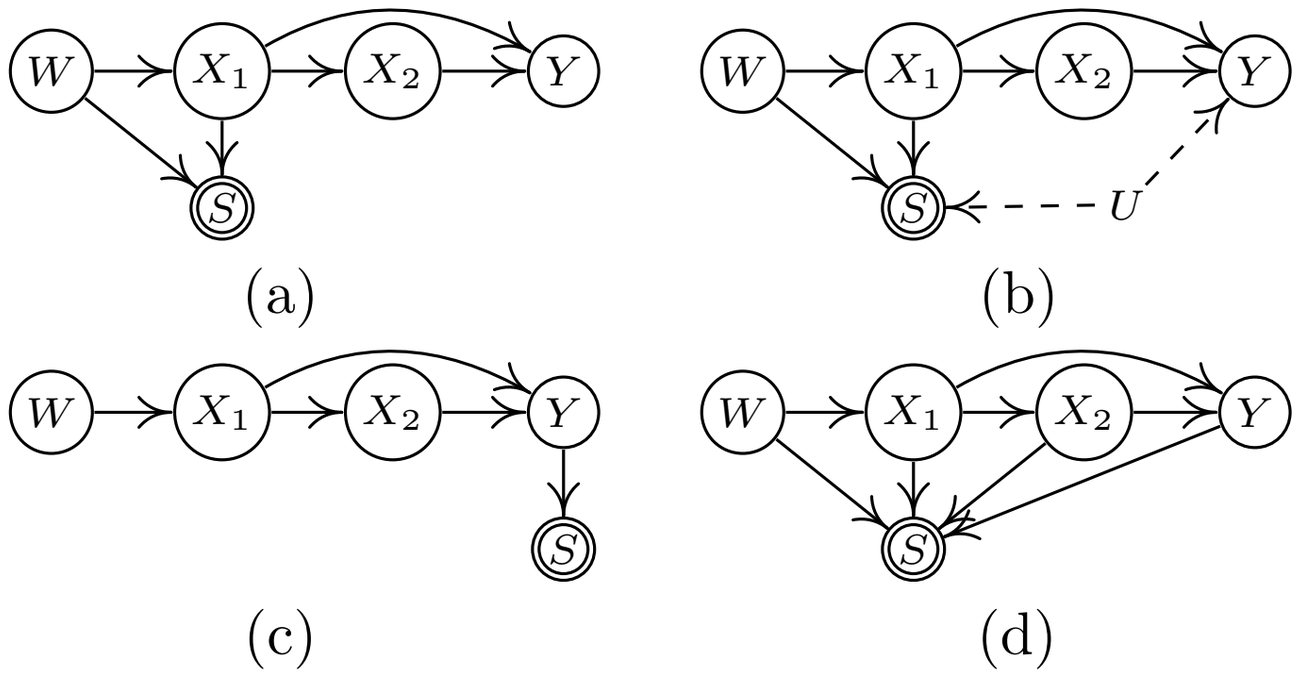




*Zhang, Gong, Ramsey, Batmanghelich, Spirtes, Glymour, "Causal Discovery in the Presence of Measurement Error: Identifiability Conditions," UAI 2018 (plenary session)*

# Extension 4: Selection Bias

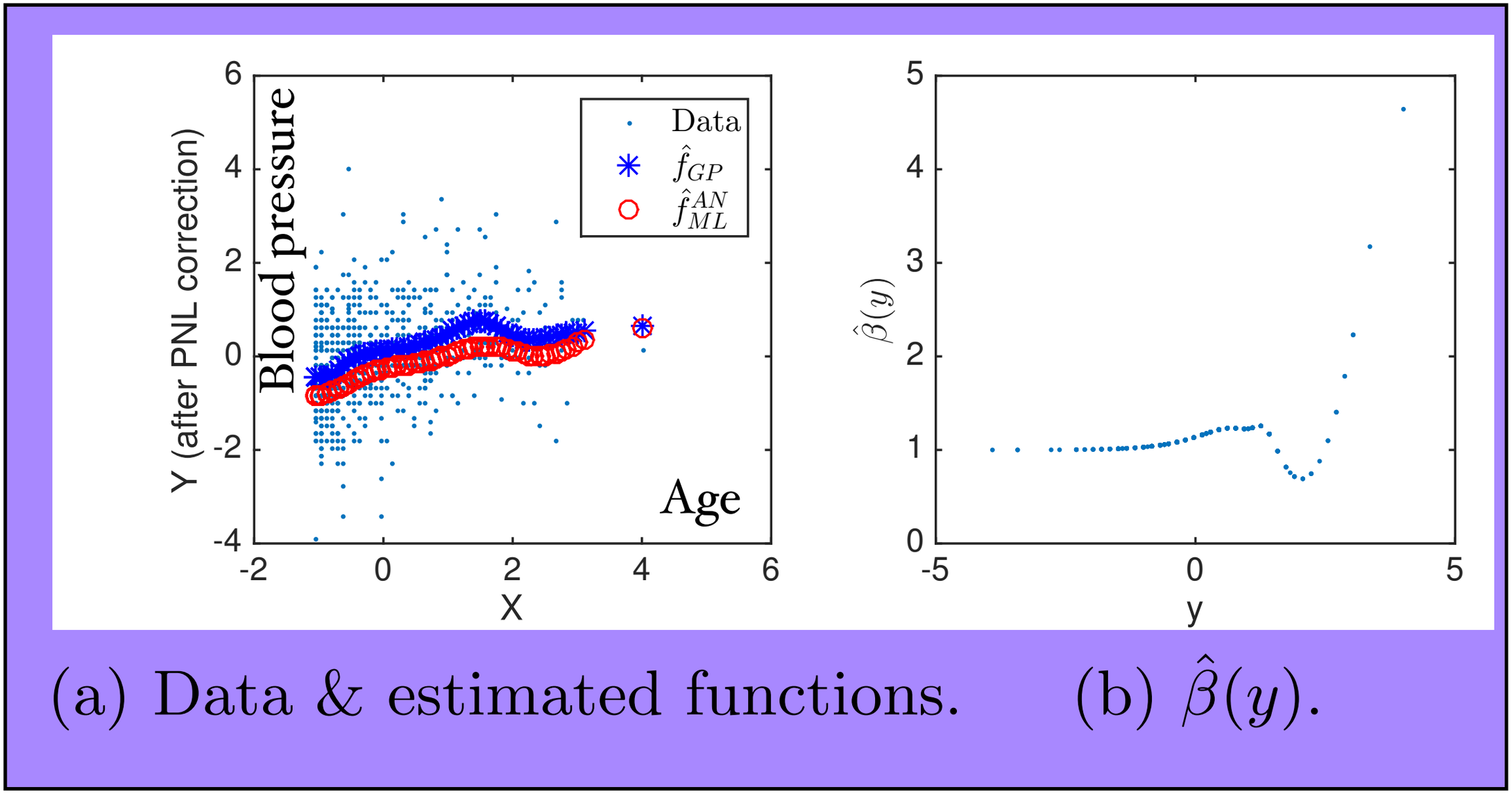
- Selection bias: The chance of including a data point in the sample depends on some attributes of the point
- Example: hospital-based disease research


- Both learning causal structures and estimating causal mechanisms become more difficult




*Zhang, Zhang, Schölkopf, Glymour, "On the Identifiability and Estimation of Functional Causal Models in the Presence of Outcome-Dependent Selection Bias," UAI 2016 (plenary session)*

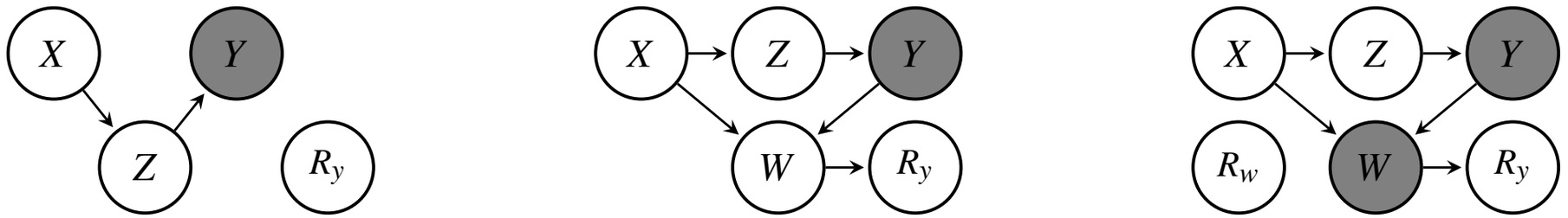
# Causal Discovery and Inference under Output-Dependent Selection




(a) Data & estimated functions. (b) $\hat{\beta}(y)$.

# Extension 5: Causal Discovery in the Presence of Missing Data

| X1 | X2 | X3 | X4 | X5 | X6 |
|---|---|---|---|---|---|
| -9.4653403e-01 | | 6.6703495e-01 | 8.2886922e-01 | -1.3695521e+00 | -3.2675465e-02 | 1.8634806e-01 |
| -9.4895568e-01 | | | | -4.6381657e-01 | -1.8280031e+00 | |
| | | 5.1435422e-01 | 6.7338326e-01 | 4.3403559e-01 | 9.4535076e-01 | 7.5164028e-01 |
| 7.2489037e-01 | | | 5.1325341e-01 | 8.3567780e-01 | 2.9825903e-01 | 7.7796018e-02 |
| | | | -1.3440612e+00 | | | -7.3325009e-01 |
| 1.3261794e+00 | | -6.1971037e-01 | -1.0498756e-01 | 1.4171149e+00 | 1.6251026e+00 | 3.7478050e-01 |
| -2.1128404e+00 | | 1.3359744e-02 | -2.0209600e+00 | -1.7172659e+00 | -2.4746799e+00 | -2.8026586e+00 |
| 1.5453163e+00 | | -5.3986972e-01 | 4.5157367e-01 | 1.5566262e+00 | 9.3882105e-01 | -4.3382982e-01 |
| 6.5974086e-02 | | 5.5826895e-01 | 6.5247930e-01 | -5.7895322e-01 | 5.0062743e-01 | 1.0183537e+00 |
| 8.9772858e-01 | | 2.6752870e-01 | -4.9204975e-01 | 7.7933358e-02 | 8.3467624e-01 | 9.2744311e-01 |
| 1.1349917e+00 | | 3.5184872e-01 | -5.6961669e-01 | -4.8335698e-01 | 9.3747444e-01 | 2.2763032e-03 |



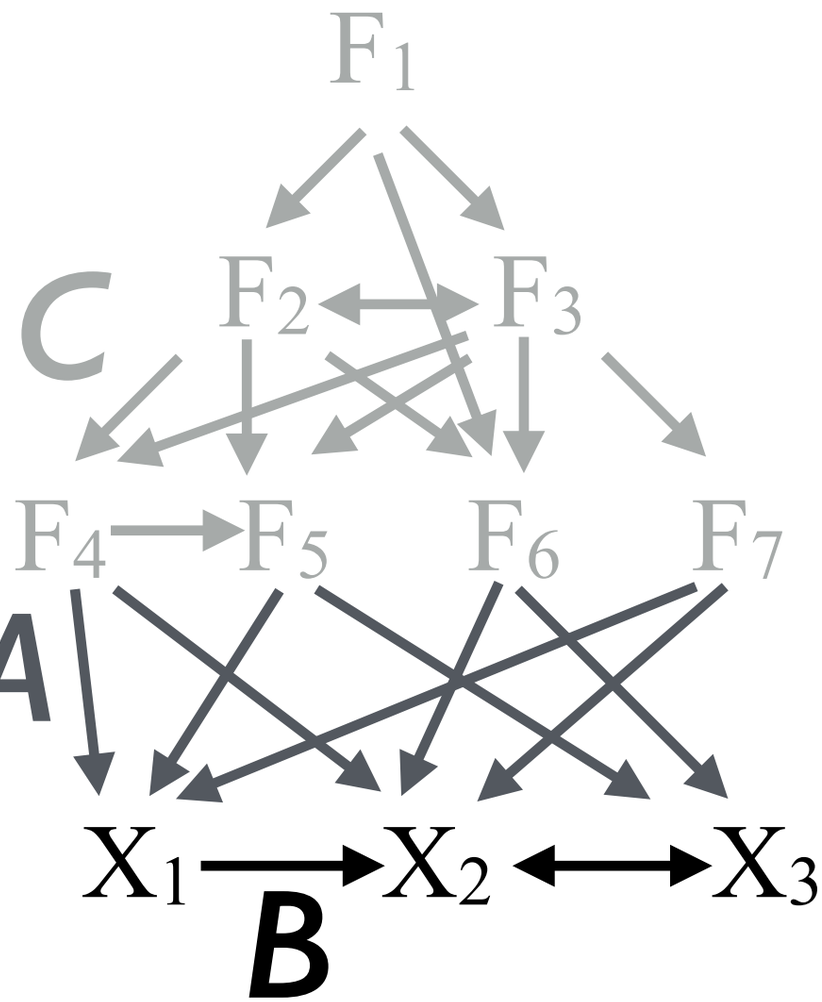

(a) An MCAR graph (b) An MAR graph (c) An MNAR graph

- Conditional independence relations in the data are sensitive to the missingness mechanism
- Key issue: Recover conditional independence relations in the original population from incomplete data

*R. Tu, C. Zhang, P. Ackermann, K. Mohan, H. Kjellström, C. Glymour, K. Zhang, "Causal discovery in the presence of missing data," AISTATS 2019*

# Extension 6: With Causally Related Latent Common Causes

- Only observe $X_i$, generated by
  - causal relations between them and
    - a large number of causally related latent variables
- Under what conditions is the whole structure identifiable?




$$\mathbf{F} = C\mathbf{F} + \mathbf{E}_F,$$
$$\mathbf{X} = B\mathbf{X} + A\mathbf{F} + \mathbf{E}_X$$

*Zhang, Adams, Glymour, et al., "Causal discovery in the presence of causally related confounders," to be submitted*

# Summary

- Why causality?
- Different types of "independence" helps in causal discovery:
  - Conditional independence: constraint-based approach
  - Cause ⫫ noise in constrained FCMs ⇒ causal asymmetry
  - Independent changes in P(cause) and P(effect | cause)
- Go beyond the data!

Thanks to
- Biwei Huang, Mingming Gong, Jiji Zhang, Philipp Geiger, Jeff Adams
- Aapo Hyvarinen, Bernhard Schölkopf, Clark Glymour, Peter Spirtes
- Judea Pearl, Lei Xu, Laiwan Chan, Dominik Janzing, Zhi-Hua Zhou, Shohei Shimizu
- Zhikun Wang, Jonas Peters, Joris Mooij, Patrik Hoyer