

Probabilistic Graphical Models

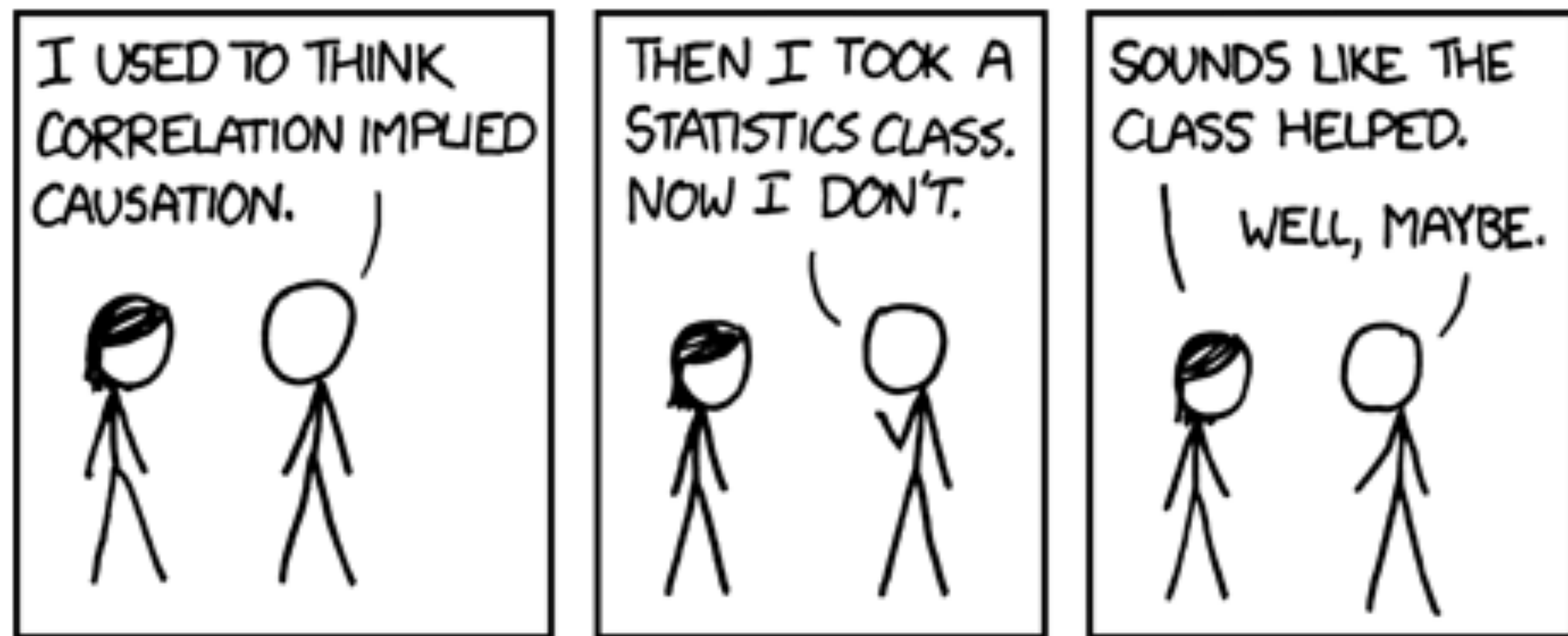
Causal Graphical Models and Causal Inference

Kun Zhang

Department of philosophy

Causality vs. Dependence

- Causality \rightarrow dependence ! Dependence \rightarrow causality



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

X and Y are **associated** iff

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

X is a **cause** of Y iff

$$\exists x_1 \neq x_2 \quad P(Y|\text{do}(X=x_1)) \neq P(Y|\text{do}(X=x_2))$$

What do you think?

Australian Open
2018

What can science tell us about grunting in tennis?


to support any of these suggestions.

Does grunting enhance hitting performance?

When the impact of a grunt is investigated, there is evidence that hitting performance is enhanced. Skilled university tennis players were found to hit with a 3.8% increase in groundstroke hitting velocity when they grunted.

For a serve, a 4.9% enhancement in velocity was found among players who grunted. This translated to “grunted serves” being hit 7kph faster than those that were not.

While the increase in hitting velocity came at no additional physiological cost, in relation to perception of effort and energy consumption, there was an increase in



One of tennis' perennial debates has ignited early at this year's Australian Open, after Belarusian player Aryna Sabalenka was accused of grunting too loudly during her first-round loss to Australian Ashleigh Barty.

Causality Examples



March, 2014

RESEARCH ARTICLES

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. Kitayama⁵

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.

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

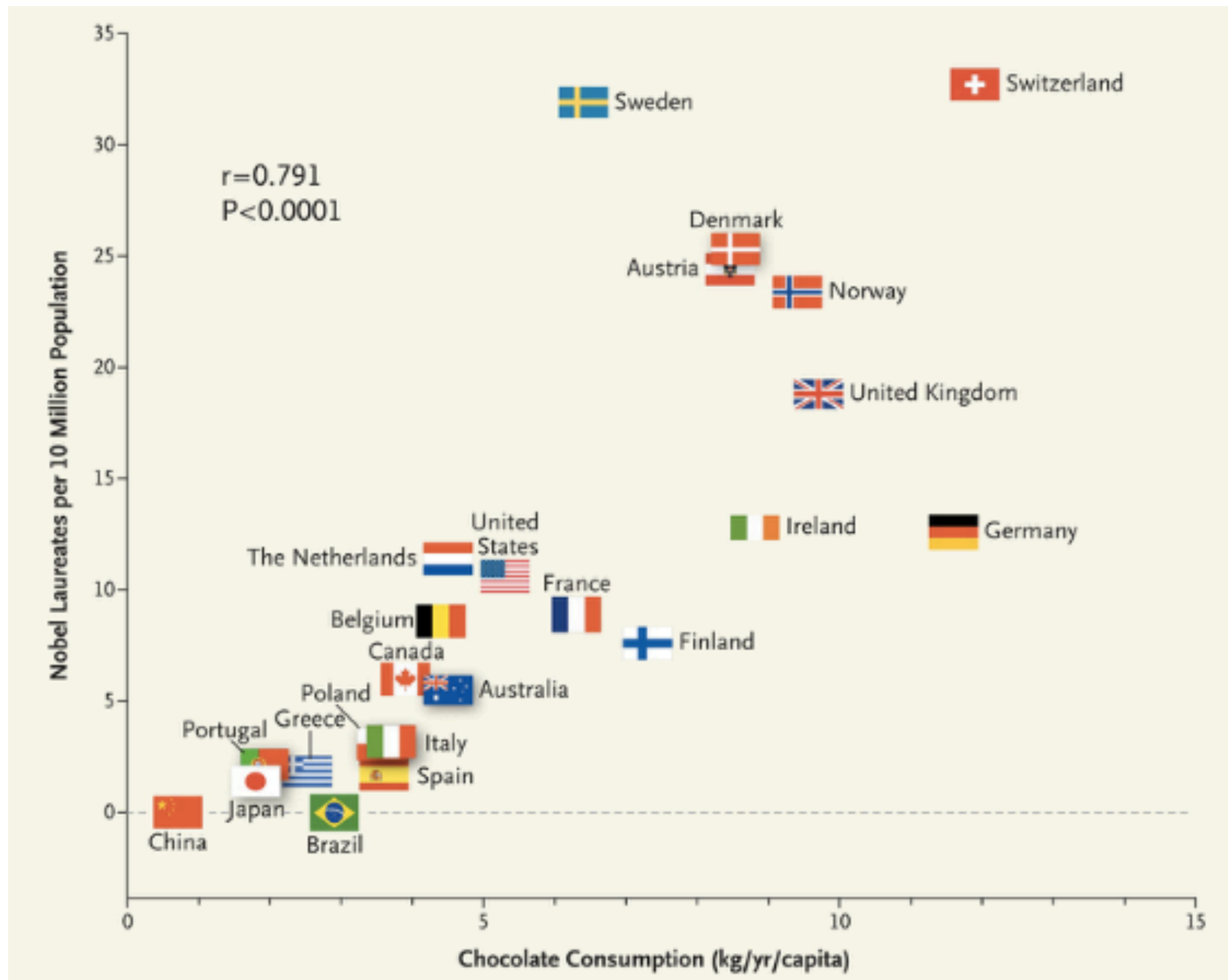
The Rice Theory

The rice theory is an extension of subsistence style theory, which argues that some forms of subsistence (such as farming) require more functional interdependence than other forms (such as herding). At the same time, ecology narrows the types of subsistence that are possible. For example, paddy rice requires a significant amount of water. Over time, societies that have to cooperate intensely become more interdependent, whereas societies that do not have to depend on each other as much become more individualistic.

In the past, most subsistence research has compared herders and farmers, arguing that the independence and mobility of herding make herding cultures individualistic and that the stability and high labor demands of farming make farming cultures collectivistic (*1*). We argue that subsistence theory is incomplete because it lumps all farming together. Two of the most common subsistence crops—rice and wheat—are very dif-

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

Causality Examples

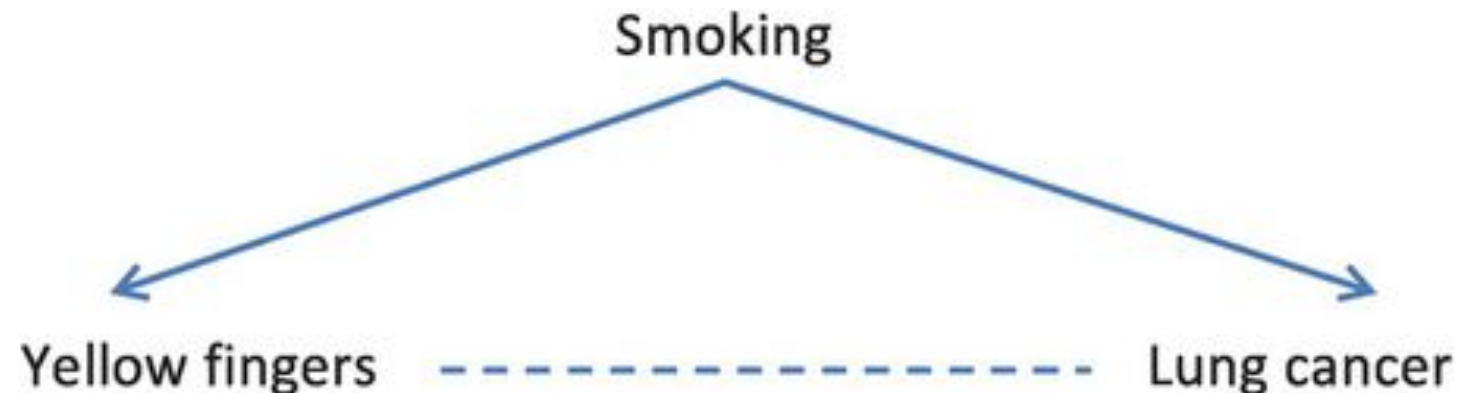


Outline

- Causal thinking
- Causal graphical models
 - Interventions
- Two main tasks
 - Identification of causal effects
 - *Causal discovery*
- Understanding cycles

Causal Thinking (1)

- Dependence vs. causality: Why do you want?

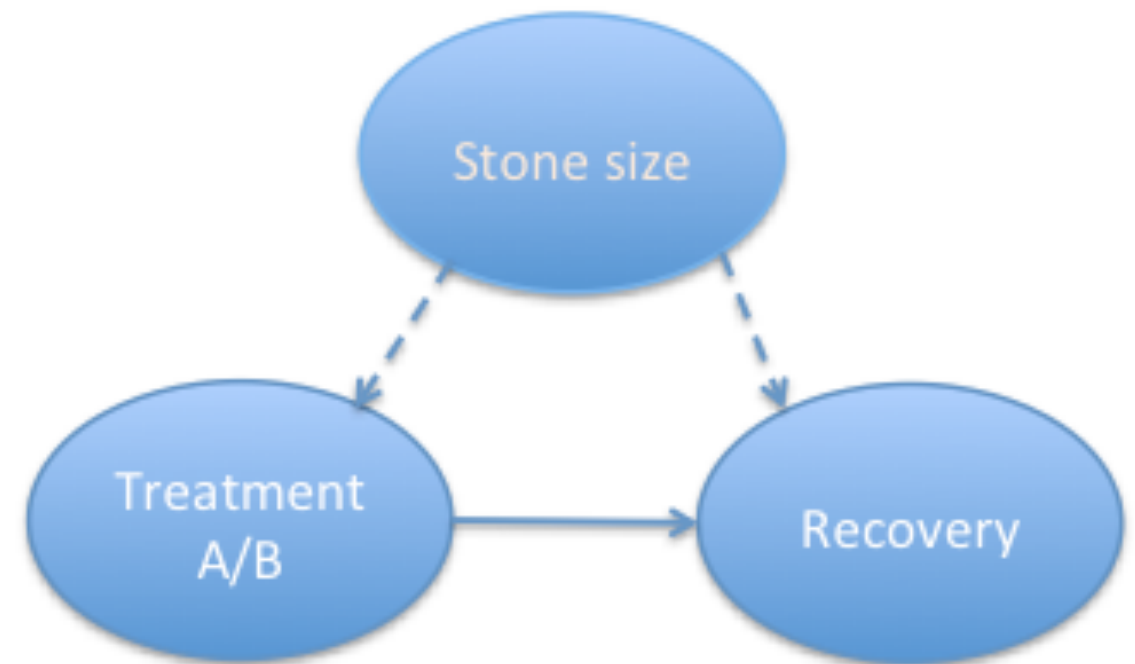


- Control, advertisements, recommender systems...
- learning (generalization), information fusion...
- What is the causal model for X , Y , and Z if $X \rightarrow Y$, $Y \rightarrow Z$ (expansion) or $X \rightarrow Z$, $Y \rightarrow Z$ (refinement)...
- According to Carl Jung, causal thinking gave rise to modern science

Causal Thinking (2)

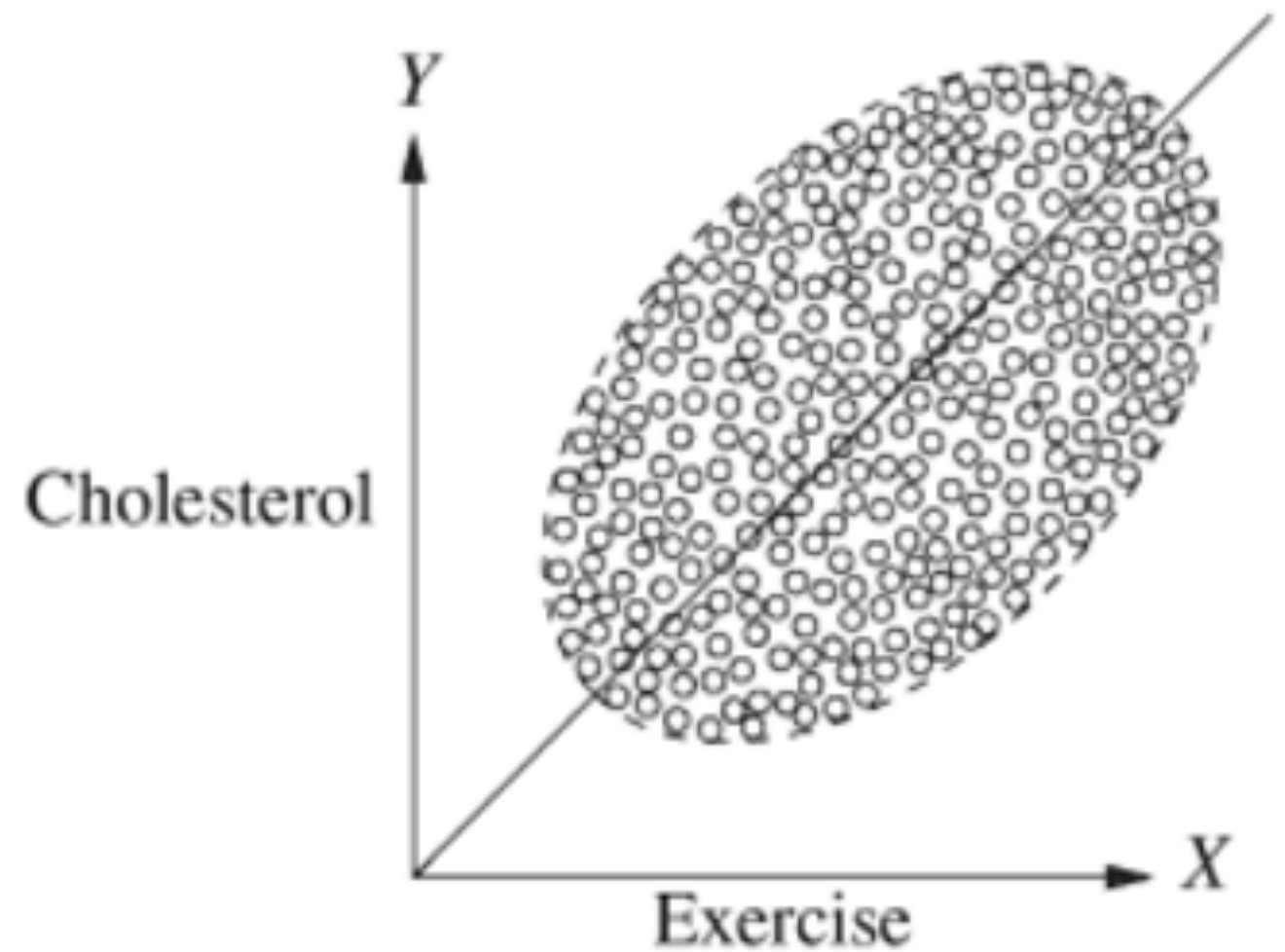
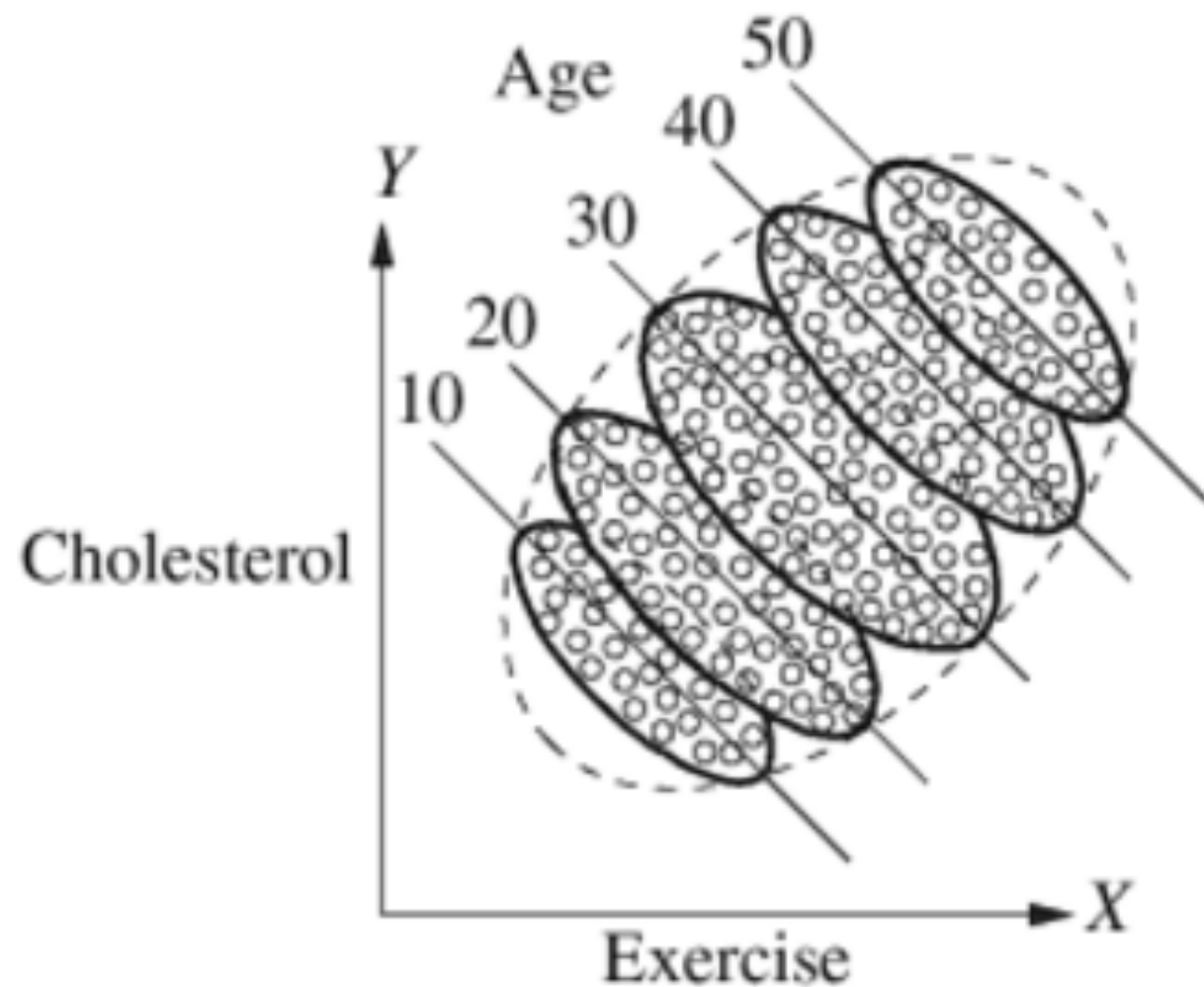
- Simpson's paradox

	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)



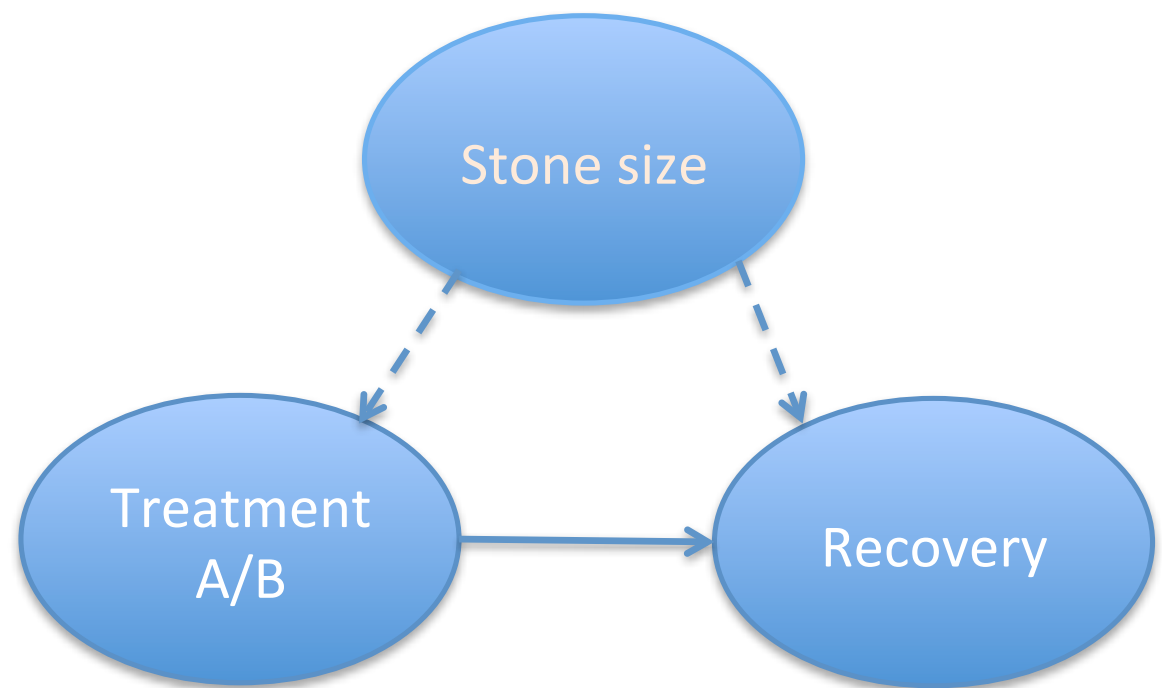
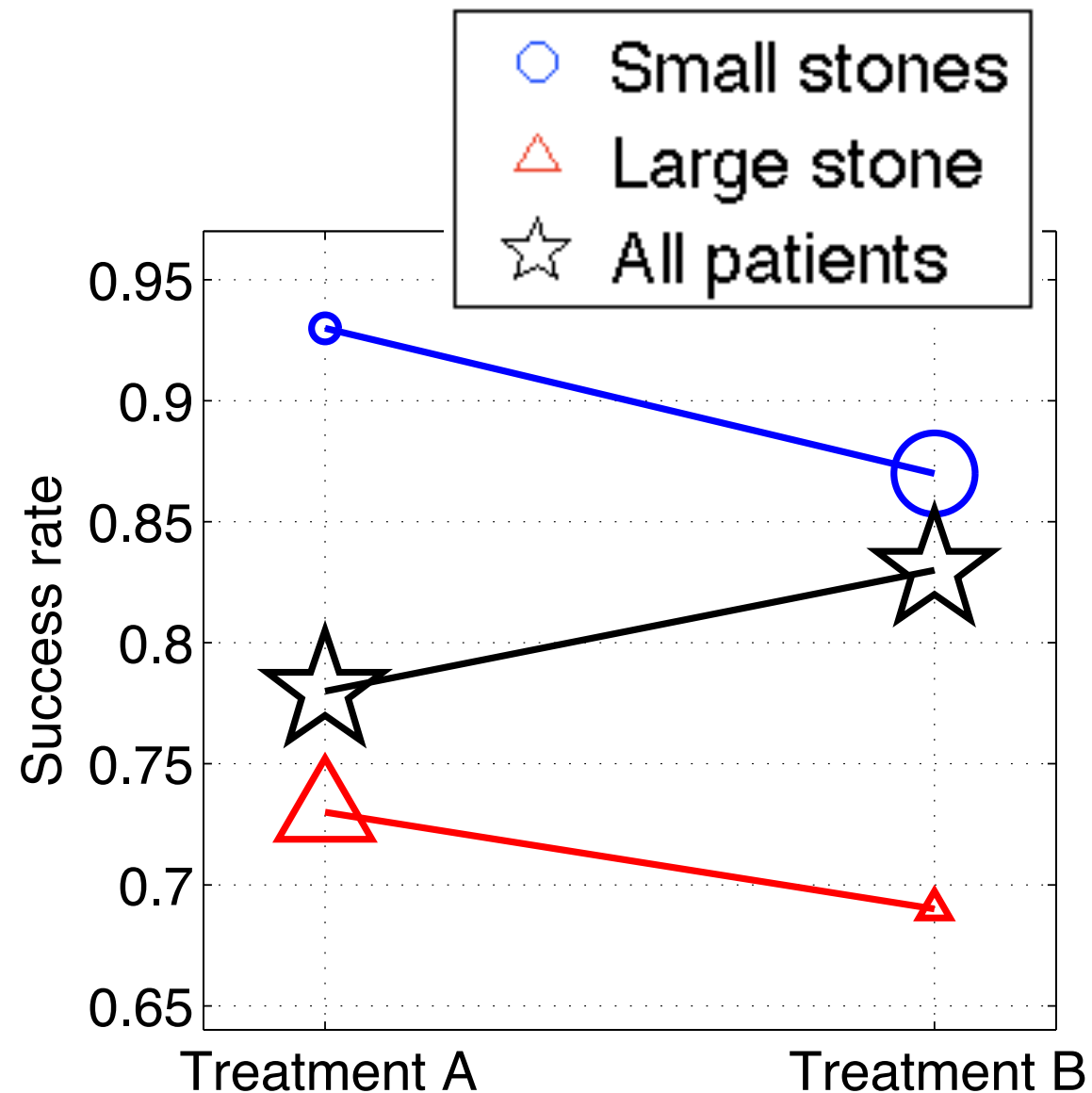
Simpson's Paradox: Another Example

- Exercise-cholesterol study



Simpson's Paradox: Why?

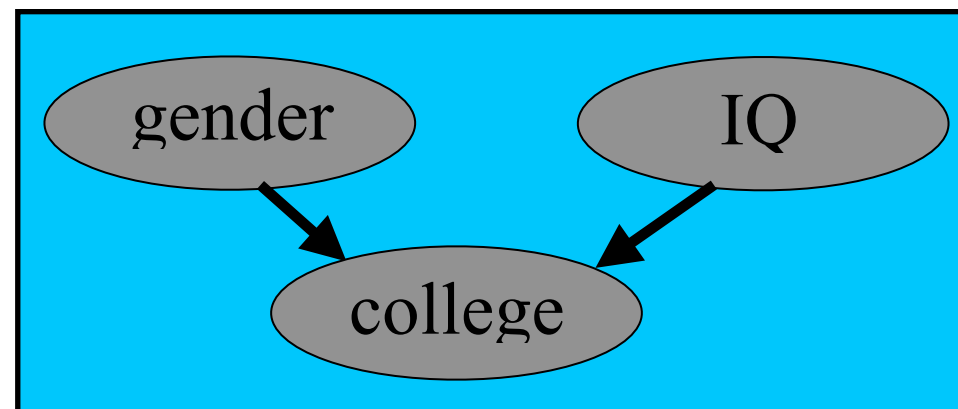
- Kidney stone treatment example:



*Would you make recommendations based on correlation or **something else?***

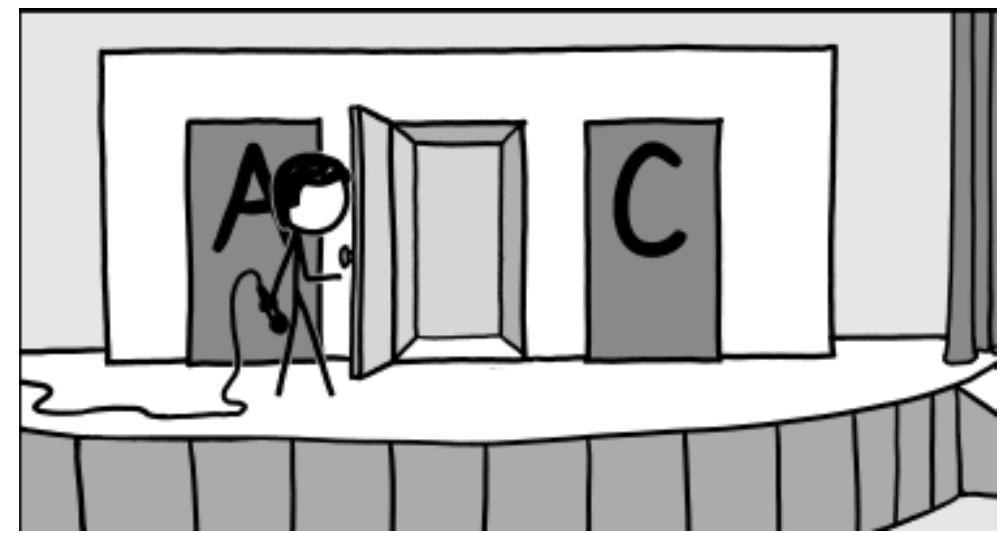
Causal Thinking (3)

- “Stranger” dependence
- Let’s go back 50 years; maybe you’ll find female college students are smarter than male ones on average. Why?



Question

Monty Hall Problem

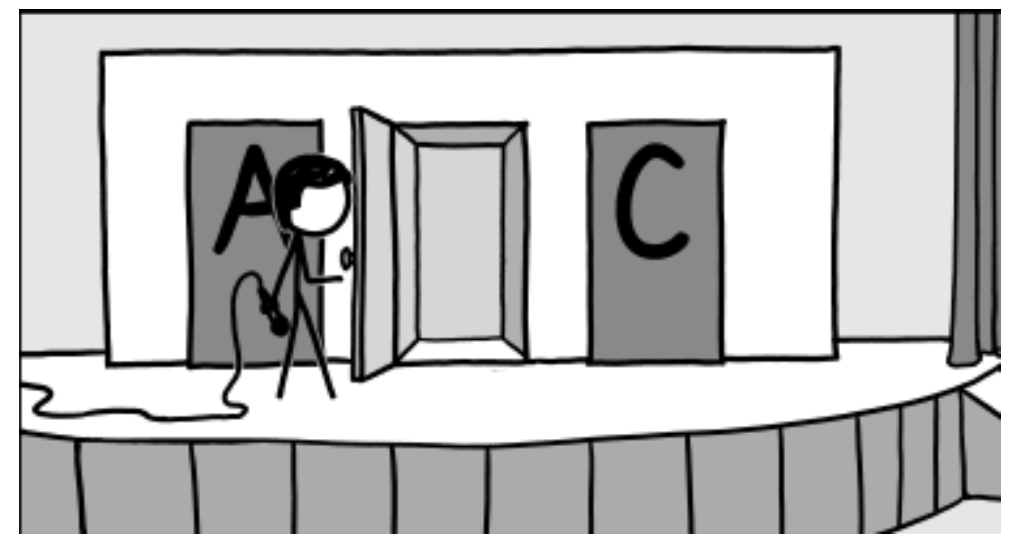


- You are a game show contestant. Before the game begins, the host, Monty Hall, has placed \$1,000 dollars behind one of three doors. Nothing is behind the other two doors. The game is played as followed. You, the contestant, choose one of the doors, say, door A. Then Monty opens a door that is not the 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"

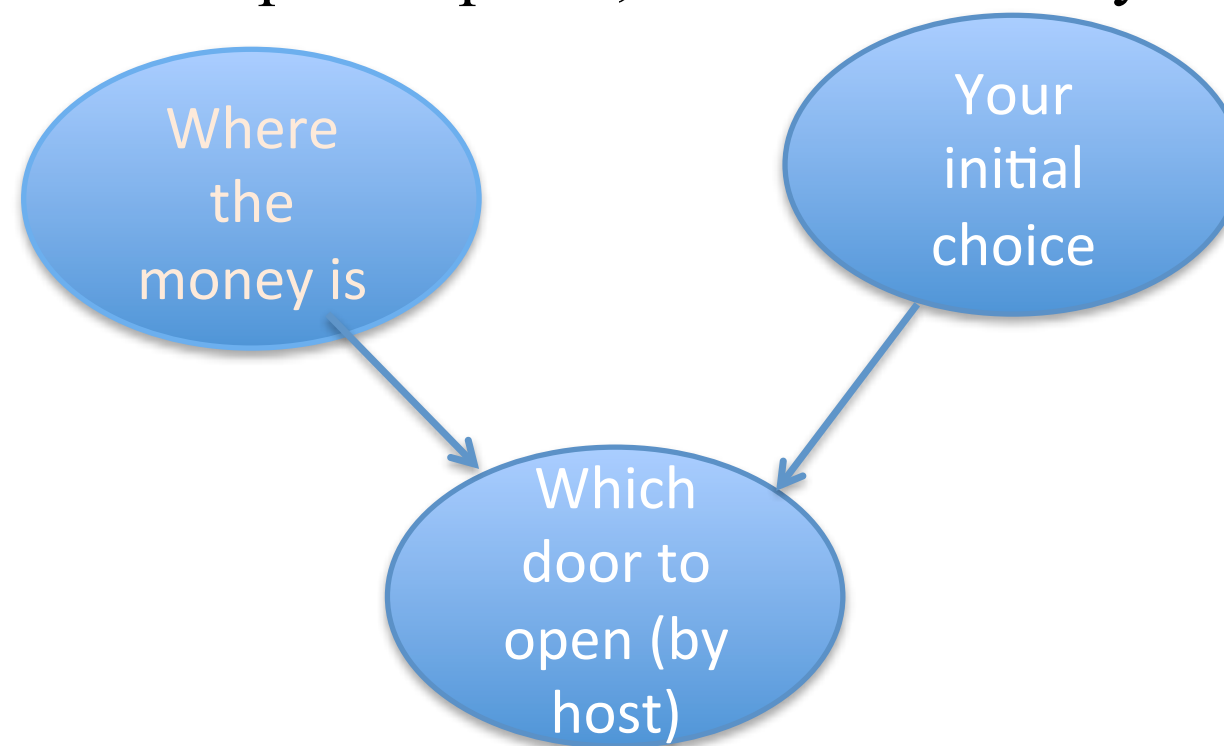
Question

Monty Hall Problem

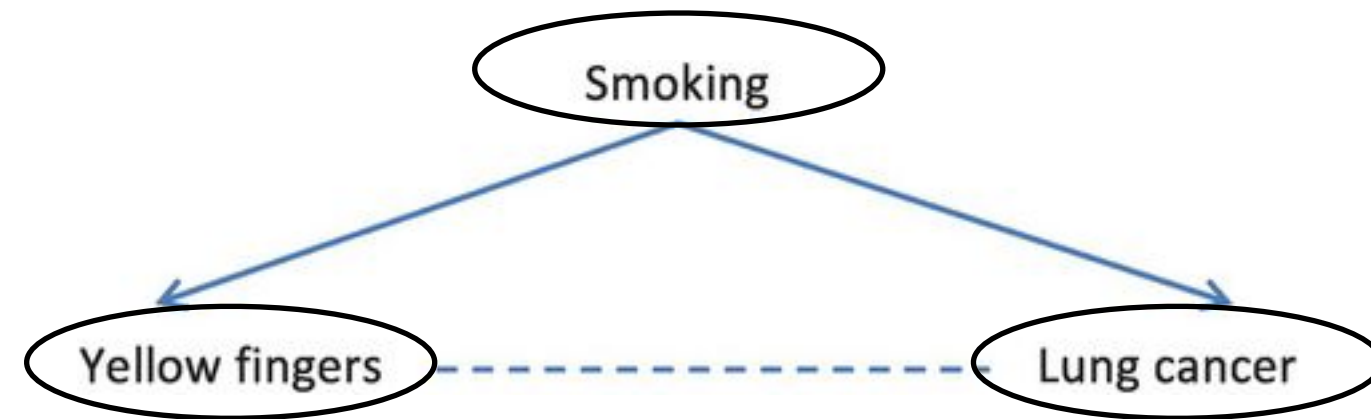


- You are a game show contestant. Before the game begins, the host, Monty Hall, has placed \$1,000 dollars behind one of three doors. Nothing is behind the other two doors. The game is played as followed. You, the contestant, choose one of the doors, say, door A. Then Monty opens a door that is not the 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?

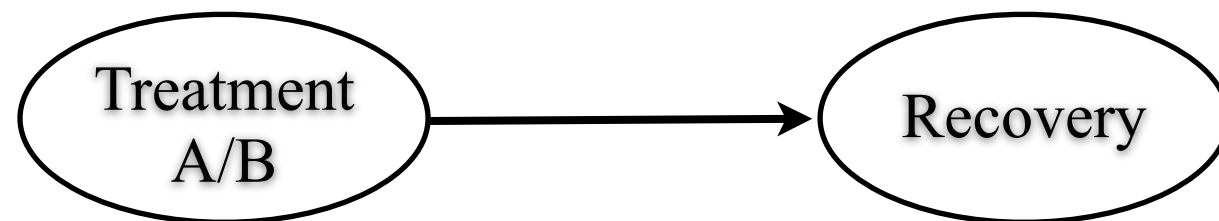
- A
- C



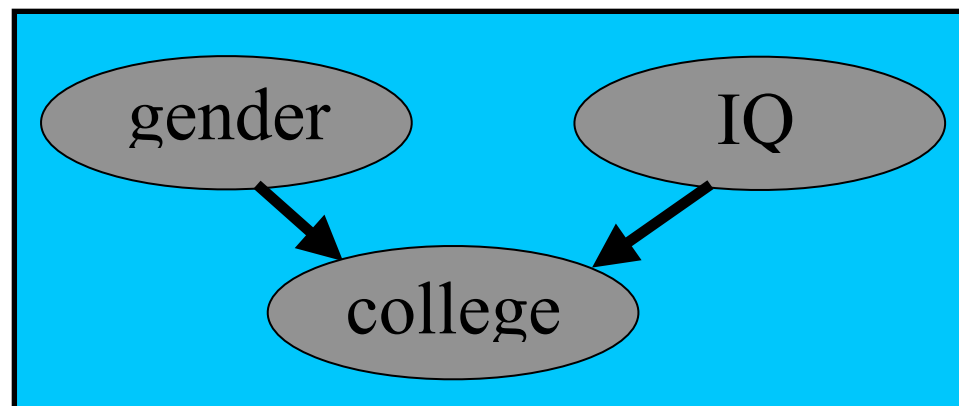
Ways to Produce Dependence



- Common cause



- Causal relation between them



- Conditional dependence
given common effect

Causality and Prediction (in Nonstationary Environments)

- Causality matters for prediction if distribution changes



Causal Thinking Makes a Difference

- Active manipulation / control vs. passive prediction
- Generalization / adaptation ability in new environments?
- Integration of causal information: what is the causal model for X , Y , and Z if
 - $X \rightarrow Y$, $Y \rightarrow Z$ (expansion) or $X \rightarrow Z$, $Y \rightarrow Z$ (refinement)...
- Creativity
 - Thoughts consist of the "What if?" and the "If I had only..." + knowledge integration + ...

Outline

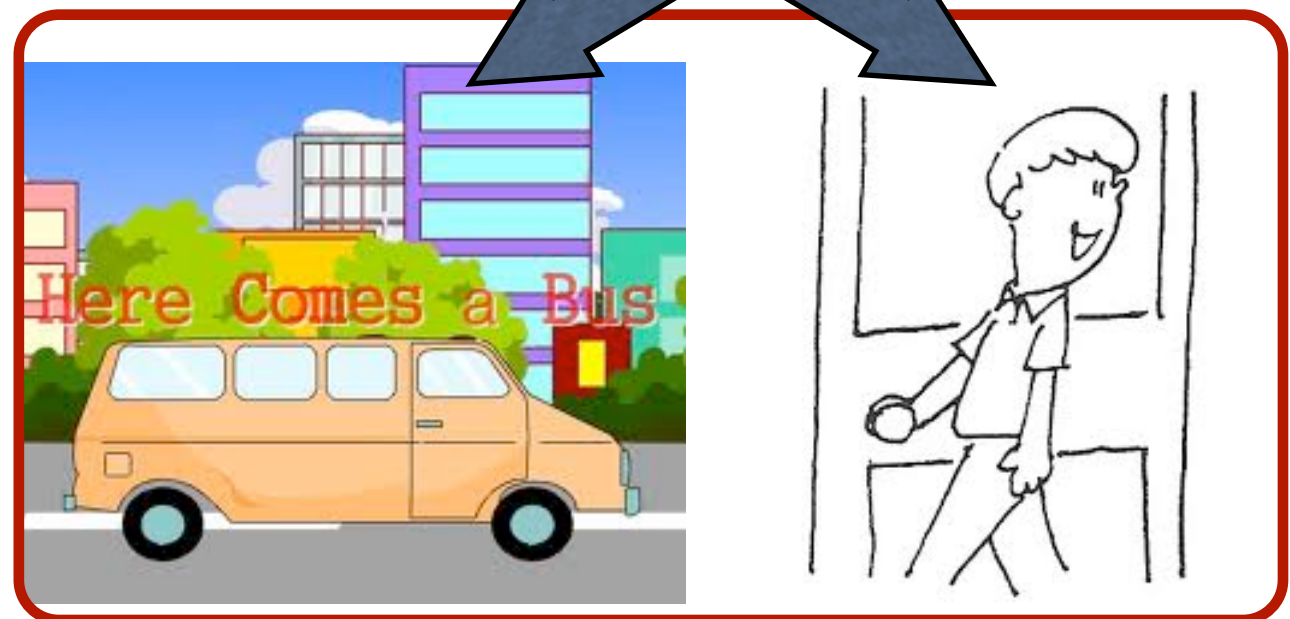
- Causal thinking
- Causal graphical models
 - Interventions
- Two main tasks
 - Identification of causal effects
 - *Causal discovery*
- Understanding cycles

Classic Ways to Find Causal Information

- What if X and Y are *dependent*?
- What if you *change* X and see Y also changes?
- An intervention directly changes only the target variable X



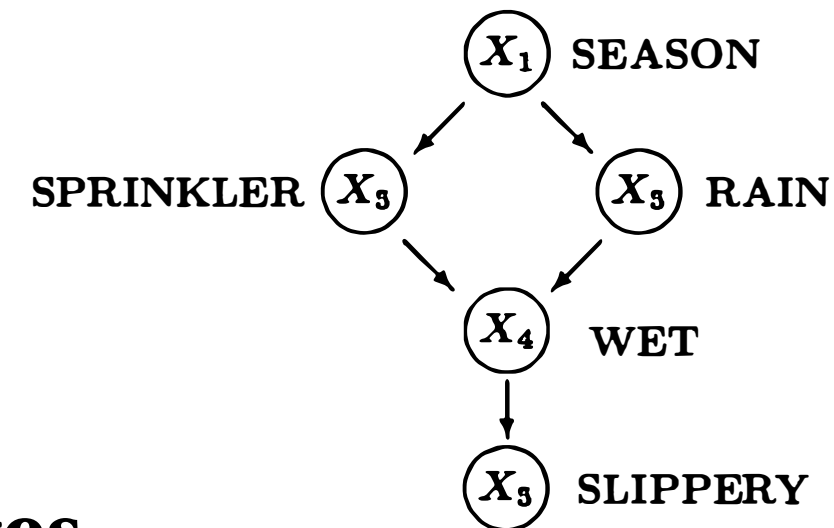
Timetable



** Definition of “interventions”*

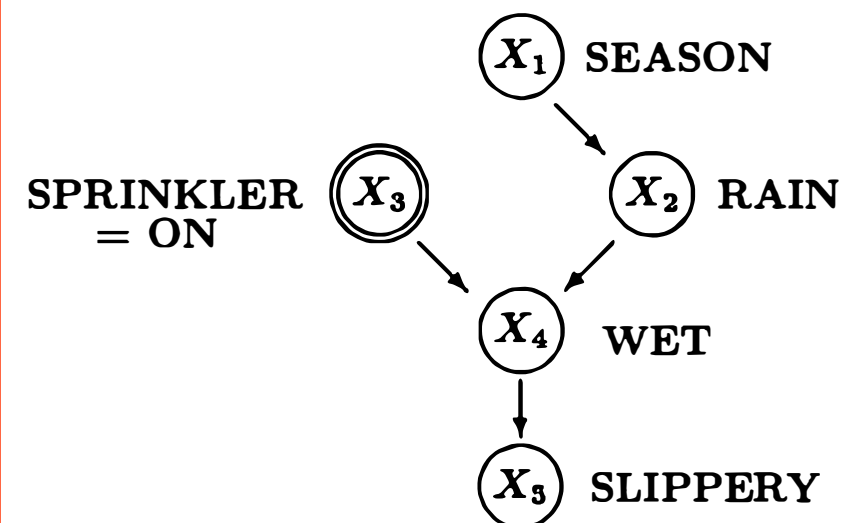
Causal DAGs

- Bayesian networks: DAGs
- Causal DAGs
 - More meaningful & 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. $P_x(V_i | PA_i) = P(V_i | PA_i)$ for all $V_i \notin X$, i.e., $P(V_i | PA_i)$ remains invariant to interventions not involving V_i .

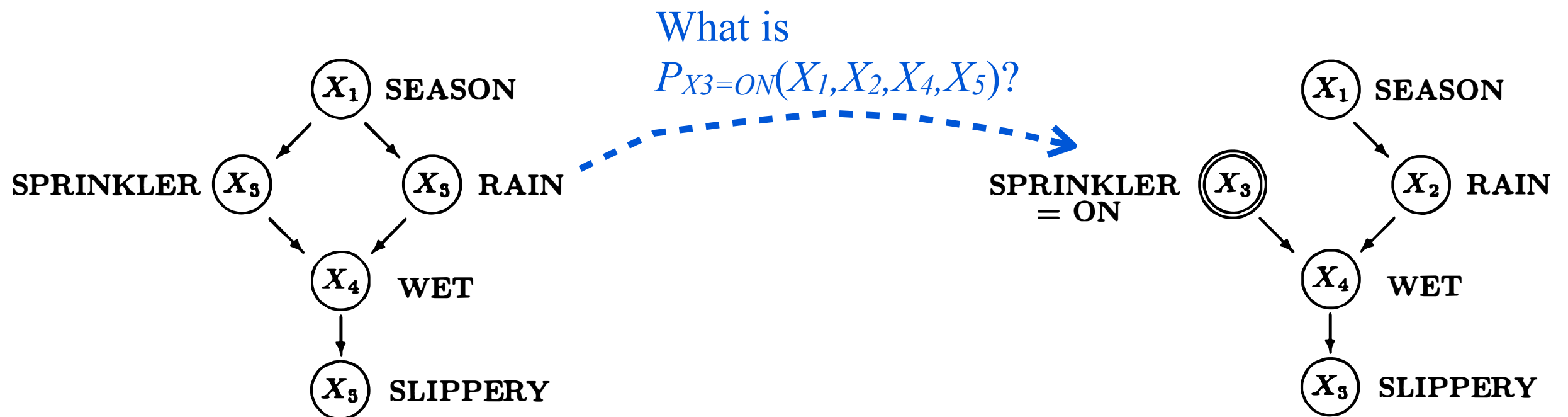


What is
 $P_{X3=ON}(X_1, X_2, X_4, X_5)$?

$$P_{X3=ON}(X_1, X_2, X_4, X_5) = P(X_1)P(X_2 | X_1)P(X_4 | X_2, X_3 = ON)P(X_5 | X_4)$$

Usage of 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

Causal structure $Y \rightarrow X$

Prob. model $P^{(1)}(X, Y), P^{(2)}(X, Y), P^{(3)}(X, Y), \dots, P^{(k)}(X, Y) \dots$

Conditioning, Manipulating, and Counterfactual Thinking



- Three questions:

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

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

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

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

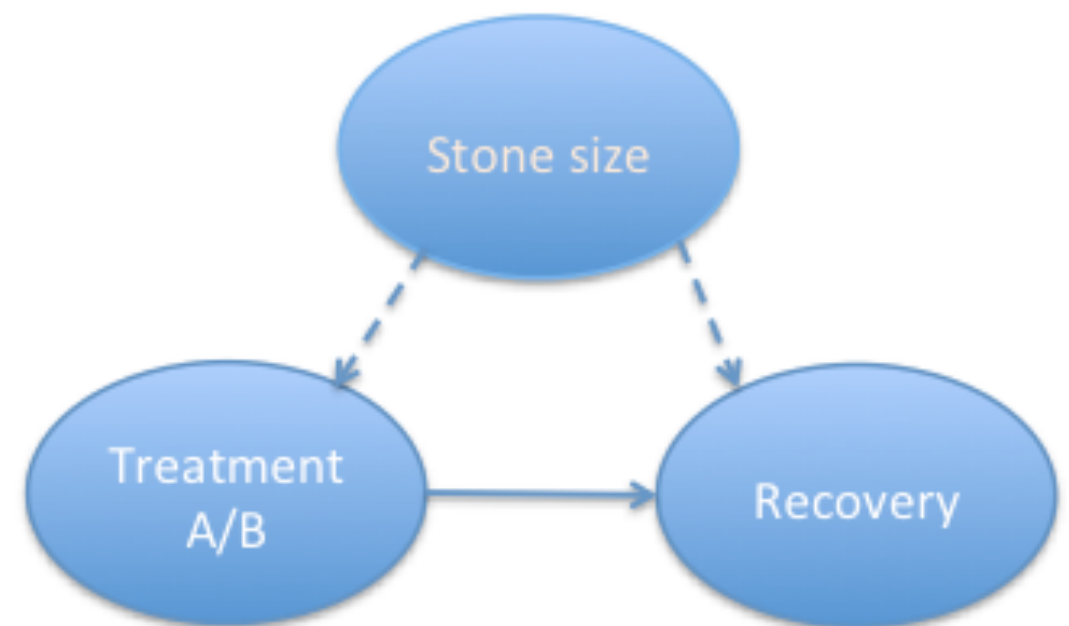
- **Counterfactual:** Would the pavement be slippery *had* the sprinkler been off, *given that the pavement is in fact not slippery and the sprinkler is on*?

$$P(\text{Slippery}_{\text{Sprinkler}=\text{off}} \mid \text{Sprinkler} = \text{on}, \text{Slippery} = \text{no})$$

Identification of Causal Effects

$P(\text{Slipperry} \mid \text{do}(\text{Sprinkler}=\text{off})) ?$

- “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

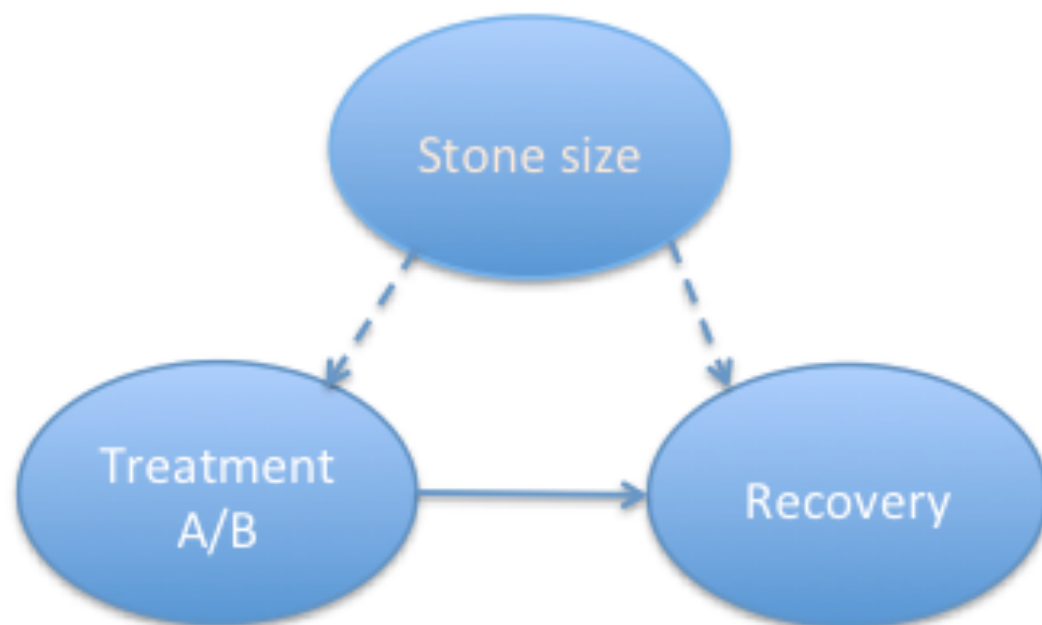


- Usually expensive or impossible to do!

Identification of Causal Effects: Example

	Treatment A	Treatment B
Small Stones	Group 1 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)

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

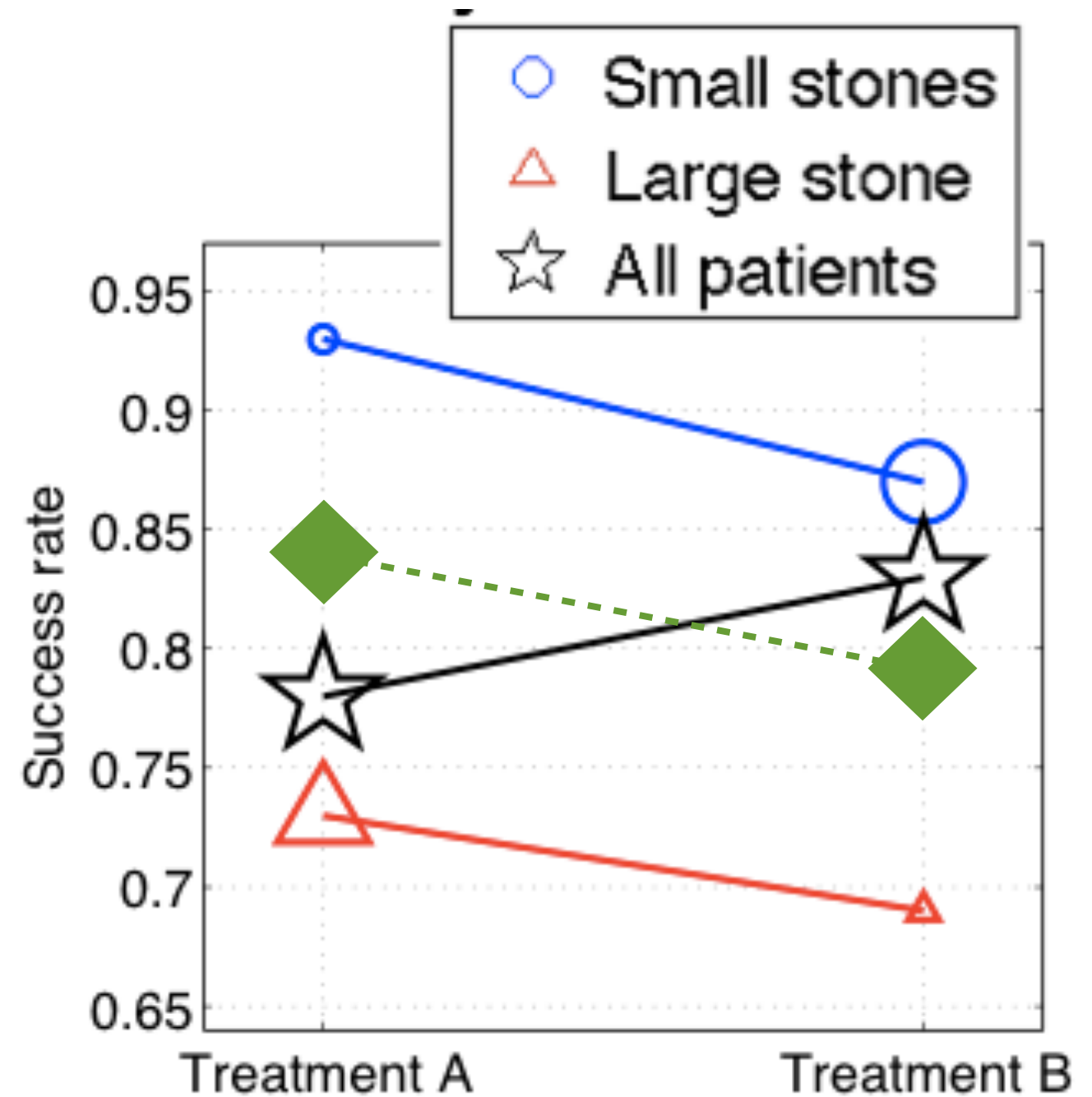
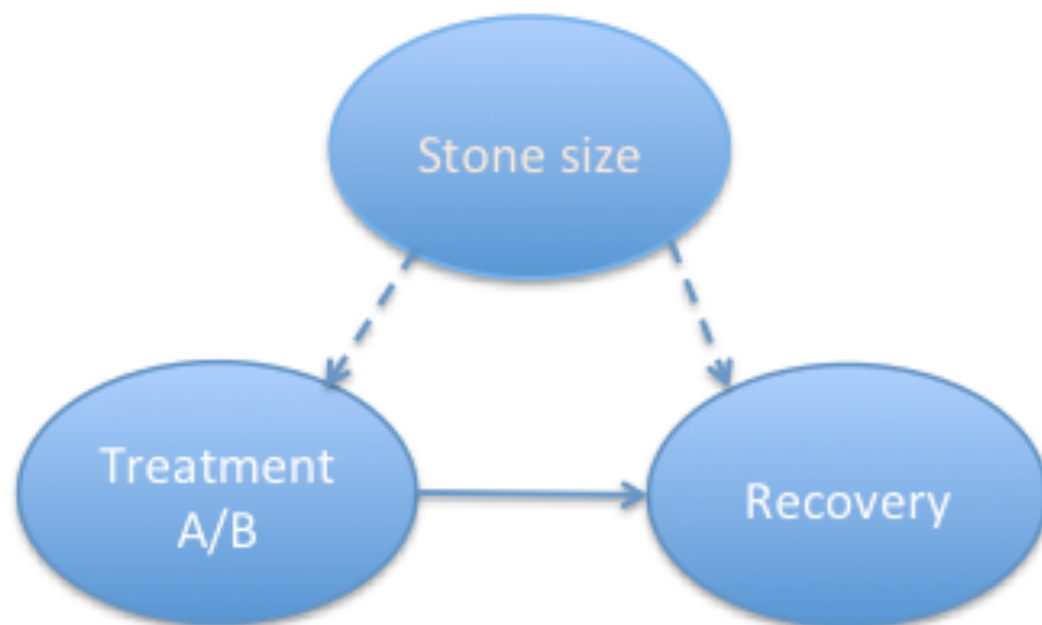


$$P(R|do(T)) = \sum_S P(R|T, S)P(S)$$

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

Identification of Causal Effects: Problems

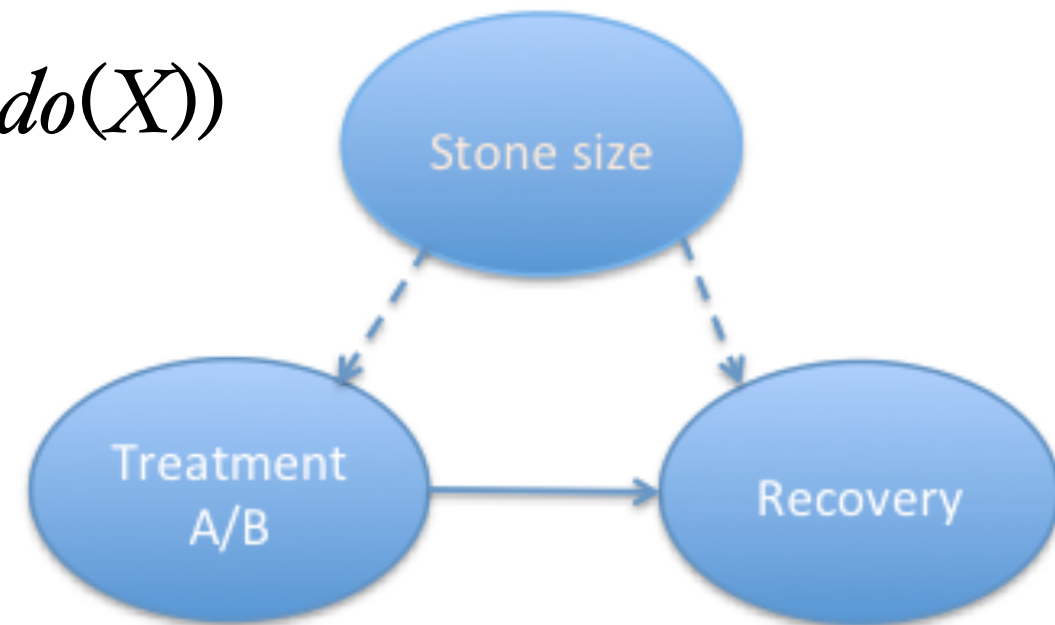
- Is $P(Y \mid do(X))$ identifiable given complete or partial causal knowledge?
- How?
 - A lot of work done by Pearl and Rubin...

Outline

- Causal thinking
- Causal graphical models
 - Interventions
- Two main tasks
 - Identification of causal effects
 - *Causal discovery*
- Understanding cycles

Causal Effects

- One definition of causal effect: $P(Y \mid do(X))$



* Definition 3.2.1 (Causal Effect)

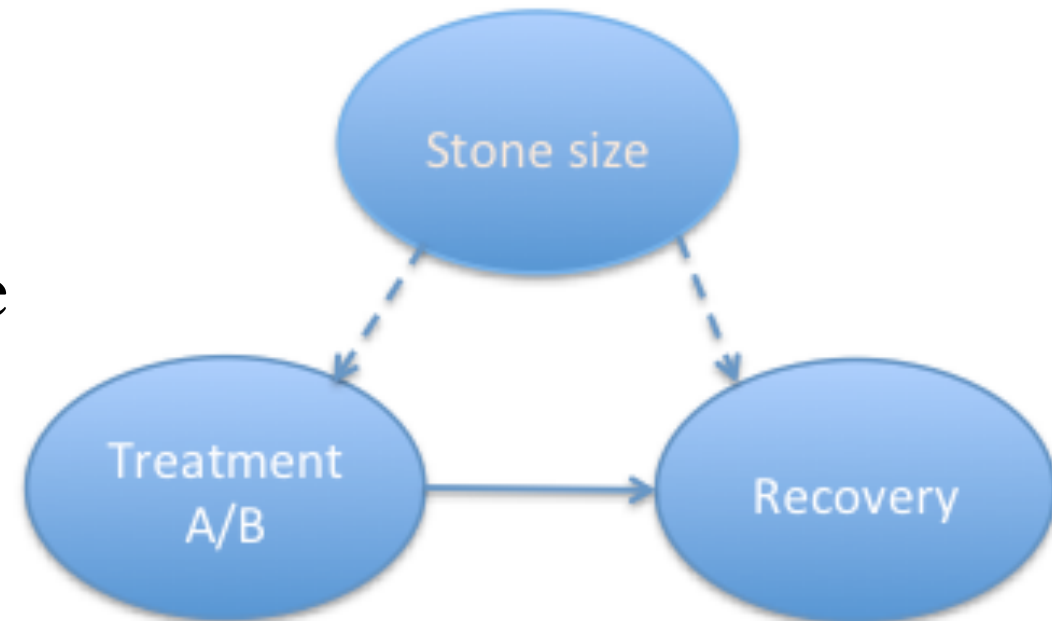
Given two disjoint sets of variables, X and Y , the causal effect of X on Y , denoted either as $P(y \mid \hat{x})$ or as $P(y \mid do(x))$, is a function from X to the space of probability distributions on Y . For each realization x of X , $P(y \mid \hat{x})$ gives the probability of $Y = y$ induced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

$$x_i = f_i(pa_i, u_i), \quad i = 1, \dots, n, \quad (3.4)$$

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)**

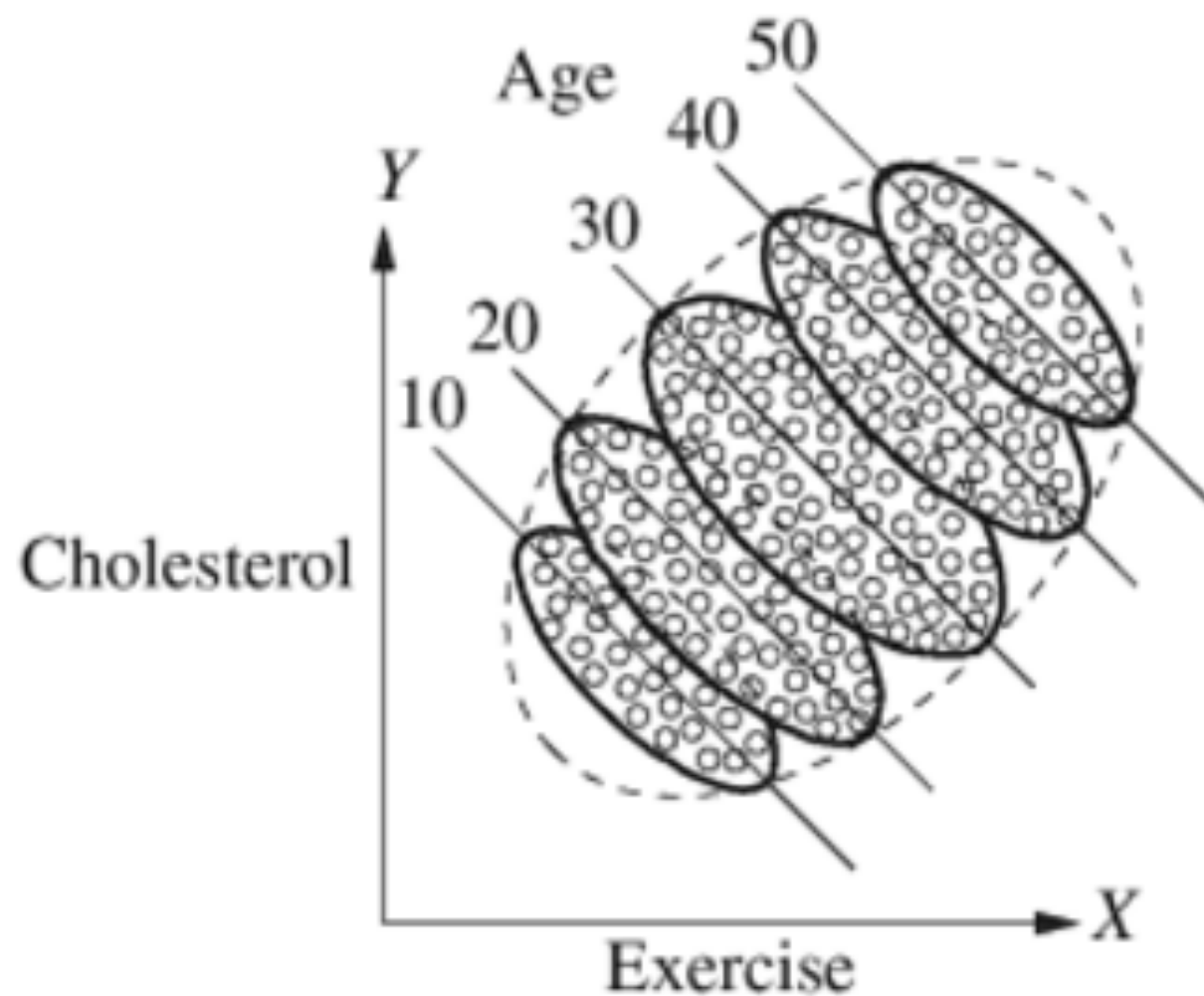
The causal effect of X on Y is identifiable from a graph G if the quantity $P(y \mid \hat{x})$ can be computed uniquely from any positive probability of the observed variables – that is, if

$P_{M_1}(y \mid \hat{x}) = P_{M_2}(y \mid \hat{x})$ 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

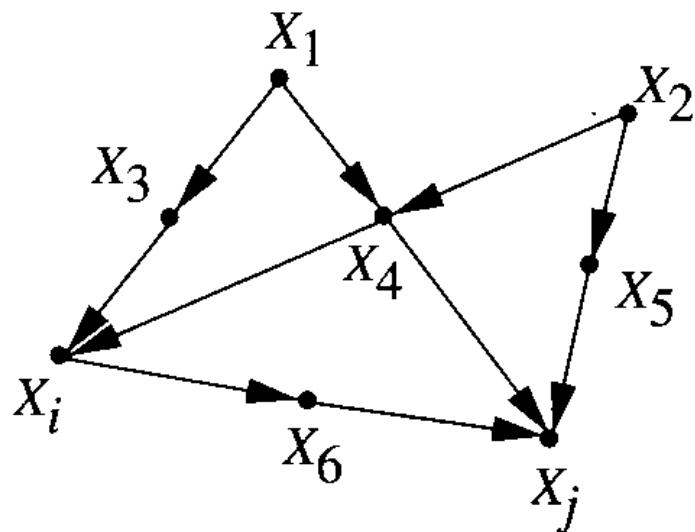


Back-Door Criterion

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?

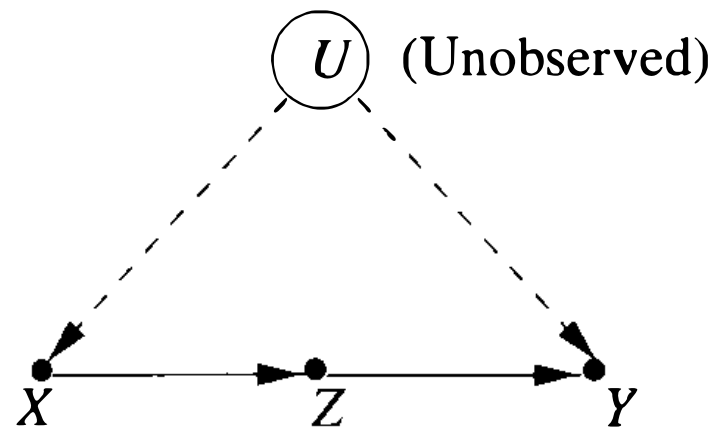
Theorem 3.3.2 (Back-Door Adjustment)

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 .

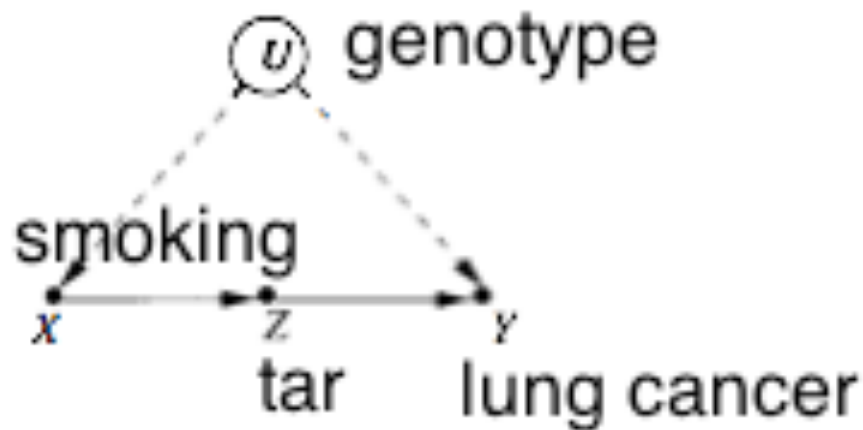
Theorem 3.3.4 (Front-Door Adjustment)

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

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



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

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

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

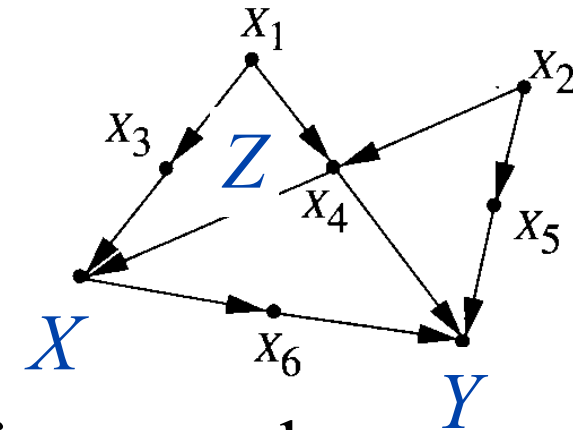


Relation to Ignorability (Potential Outcome Framework)

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:

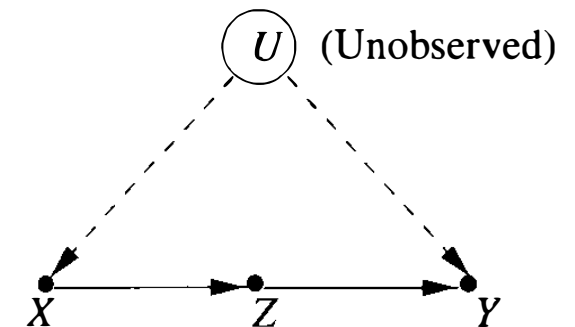
$$Y(x) \perp\!\!\!\perp X \mid Z.$$

$Y(x, u)$: the value attained by Y in unit u under intervention $\text{do}(x)$;
 $Y(x)$: 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 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 .



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

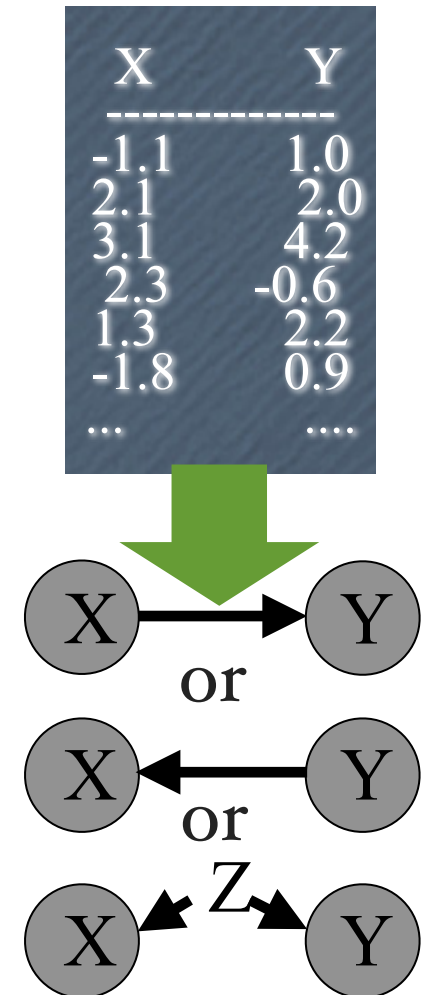
Causal Discovery: To Be Studied

Possible to

discover causal information
(*specific properties of the true process*)

from purely observational data ?

Can we go beyond the data?



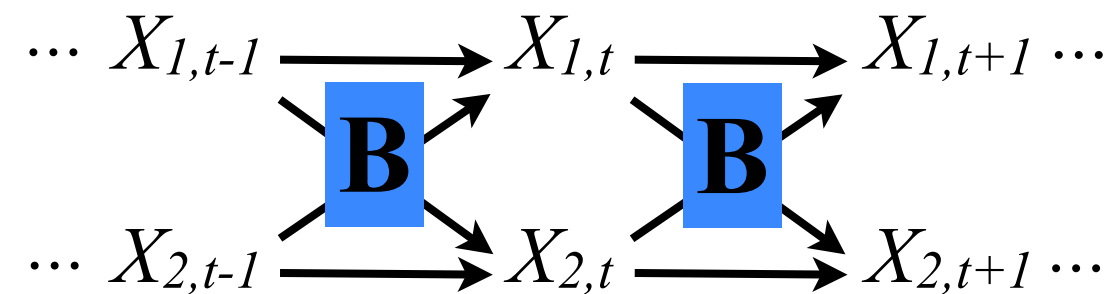
Outline

- Causal thinking
- Causal graphical models
 - Interventions
- Two main tasks
 - Identification of causal effects
 - *Causal discovery*
- Understanding cycles

Why Feedbacks?

$$X_1 \rightarrow X_2$$

- Some situations where we can recover cycles with ICA:
- Each process reaches its **equilibrium state** & we observe the equilibrium states of **multiple processes**



$$\mathbf{X}_t = \mathbf{B}\mathbf{X}_{t-1} + \mathbf{E}_t.$$

At convergence we have $X_t = X_{t-1}$ for each dynamical process, so

$$\mathbf{X}_t = \mathbf{B}\mathbf{X}_t + \mathbf{E}_t, \quad \text{or} \quad \mathbf{E}_t = (\mathbf{I} - \mathbf{B})\mathbf{X}_t.$$

- On **temporally aggregated** data

Suppose the underlying process is $\tilde{\mathbf{X}}_t = \mathbf{B}\tilde{\mathbf{X}}_{t-1} + \tilde{\mathbf{E}}_t$, but we just observe $\mathbf{X}_t = \frac{1}{L} \sum_{k=1}^L \tilde{\mathbf{X}}_{t+k}$. Since

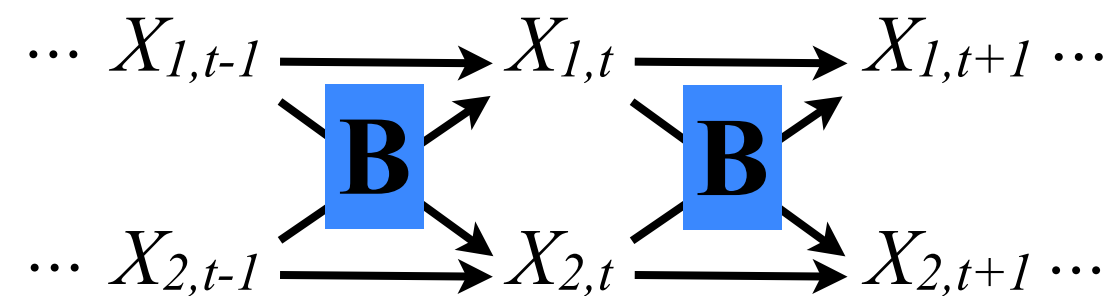
$$\frac{1}{L} \sum_{k=1}^L \tilde{\mathbf{X}}_{t+k} = \mathbf{B} \frac{1}{L} \sum_{k=1}^L \tilde{\mathbf{X}}_{t+k-1} + \frac{1}{L} \sum_{k=1}^L \tilde{\mathbf{E}}_{t+k}.$$

We have $\mathbf{X}_t = \mathbf{B}\mathbf{X}_t + \mathbf{E}_t$ as $L \rightarrow \infty$.

Examples

$$X_1 \overset{\curvearrowright}{\rightarrow} X_2$$

- Some situations where we can recover cycles with ICA:
- Each process reaches its **equilibrium state** & we observe the equilibrium states of **multiple processes**



Consider the price and demand of the same product in different states:

$$\begin{aligned} \text{price}_t &= b_1 \cdot \text{price}_{t-1} + b_2 \cdot \text{demand}_{t-1} + E_1 \\ \text{demand}_t &= b_3 \cdot \text{price}_{t-1} + b_4 \cdot \text{demand}_{t-1} + E_2 \end{aligned}$$

- On **temporally aggregated** data

Suppose the underlying process is $\tilde{\mathbf{X}}_t = \mathbf{B}\tilde{\mathbf{X}}_{t-1} + \tilde{\mathbf{E}}_t$, but we just observe $\mathbf{X}_t = \frac{1}{L} \sum_{k=1}^L \tilde{\mathbf{X}}_{t+k}$.

Consider the causal relation between two stocks: the causal influence takes place very quickly (~ 1 -2 minutes) but we only have daily returns.

Summary

- Causal thinking, causal representation, benefit from using causal graphs
- Causal discovery...?
- Causality-based learning?

Thanks to

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