Causality I

Guest lecture for "Probabilisitic Graphical Models"

Kun Zhang

kunz1@cmu.edu

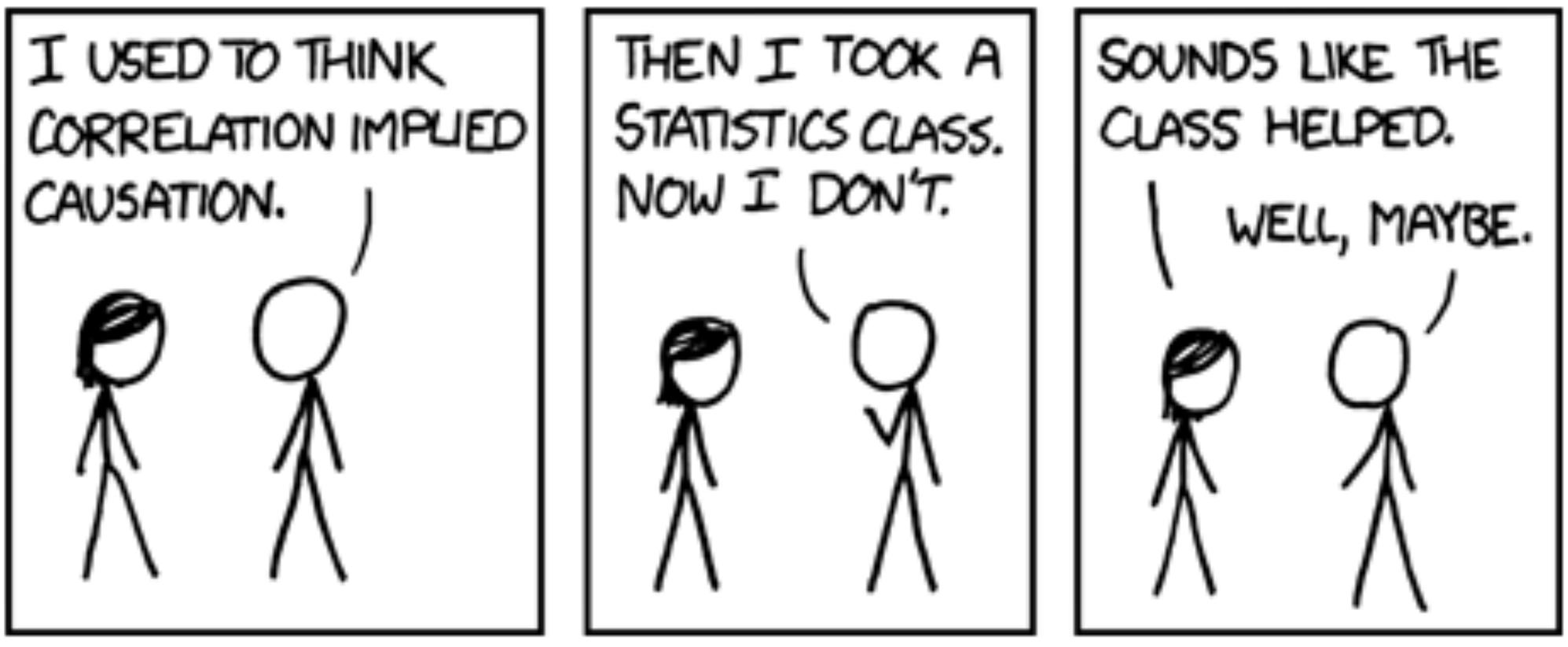
Carnegie Mellon University

Causality vs. Dependence: Example

Causality vs. Dependence: Example

Causality vs. Dependence

- Causality ➜ dependence ! Dependence ➜ causality



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

X and Y are associated iff

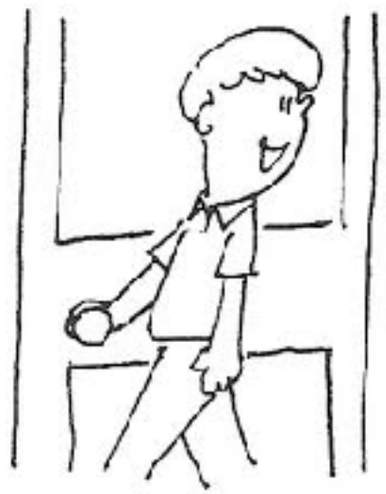
$\exists x_1 \neq x_2 \; P(Y|X=x_1) \neq P(Y|X=x_2)$

X is a cause of Y iff

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

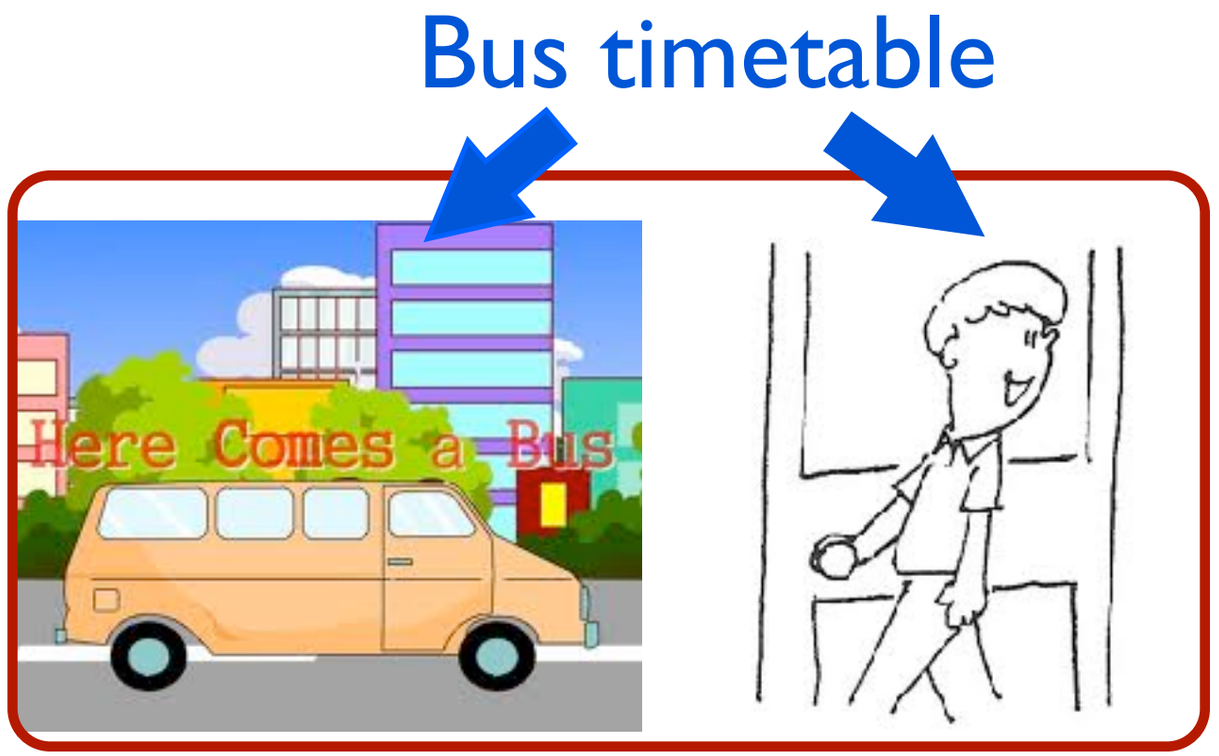
Classic Ways to Find Causal Information

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



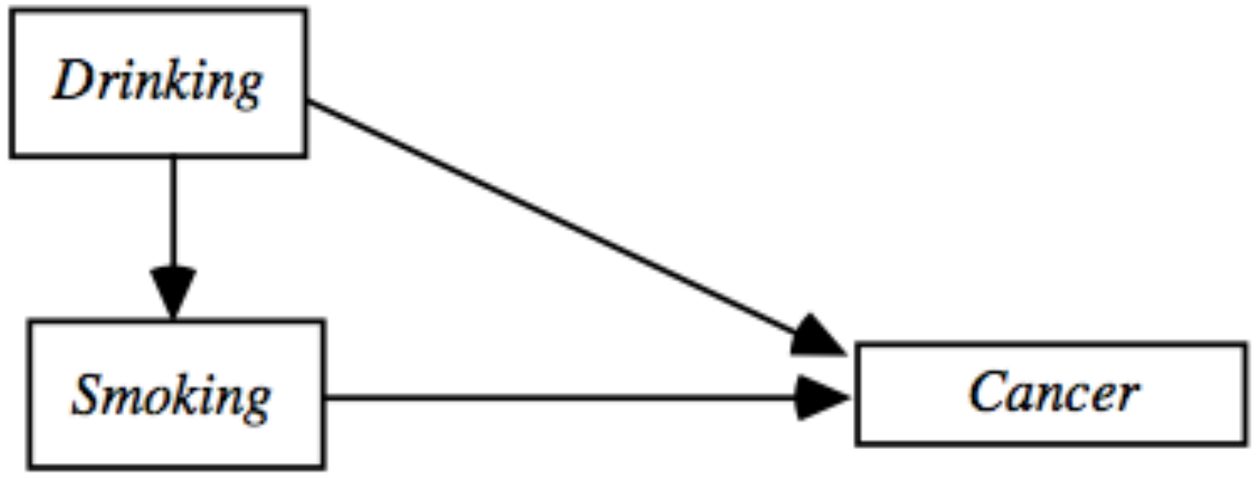
Classic Ways to Find Causal Information

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



Representing Causal Relations with Directed Graphs

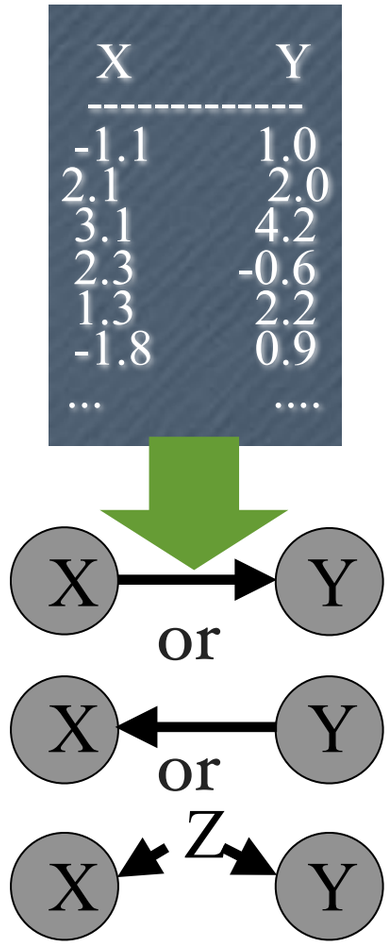
- A directed graph represents a causally sufficient causal structure



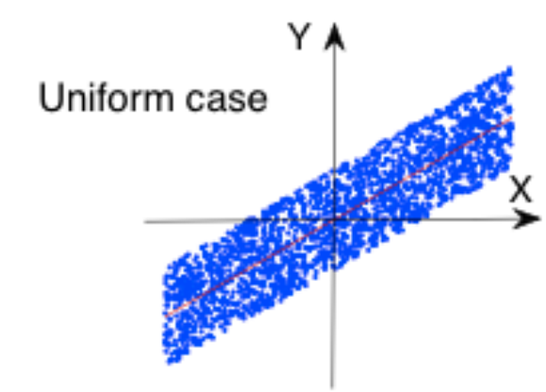
(adapted from "Causation, Prediction, and Search" by SGS, 1995)

- Directed edge from A to B means A is a direct cause of B relative to the given variable set $\boldsymbol{V}$

Outline

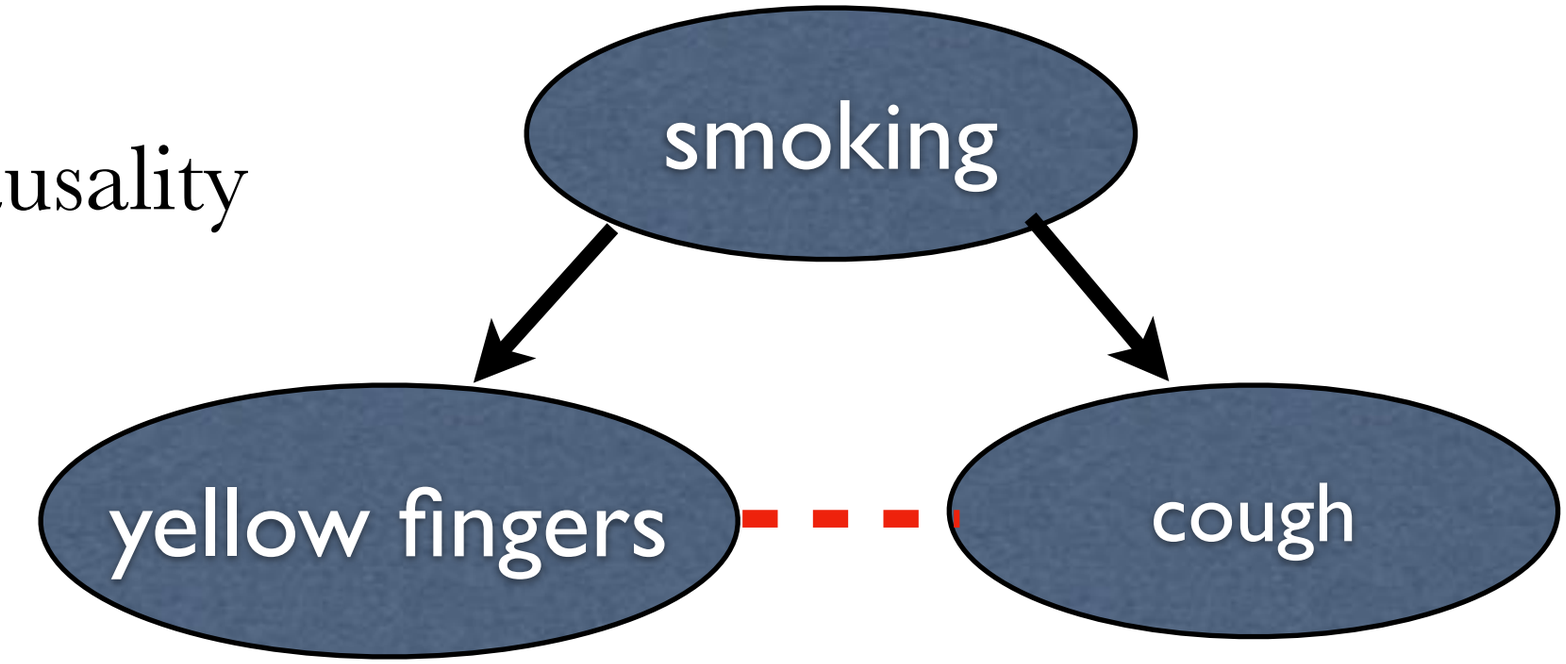


- Causality? Interventions? Causal thinking
- Causal graphical models
- Identification of causal effects
- Counterfactual reasoning
- Causal discovery
- Implications in machine learning



Causal Thinking (1)

- Dependence vs. causality



- Simpson's paradox

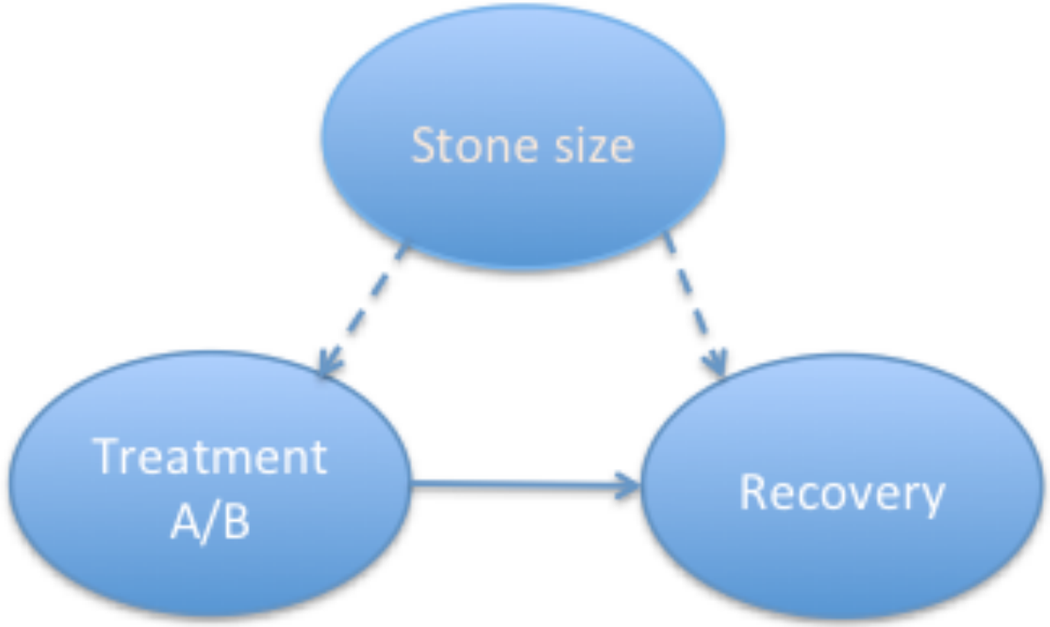
- "Strange" dependence

Causal Thinking (2)

- Dependence vs. causality

- 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)**

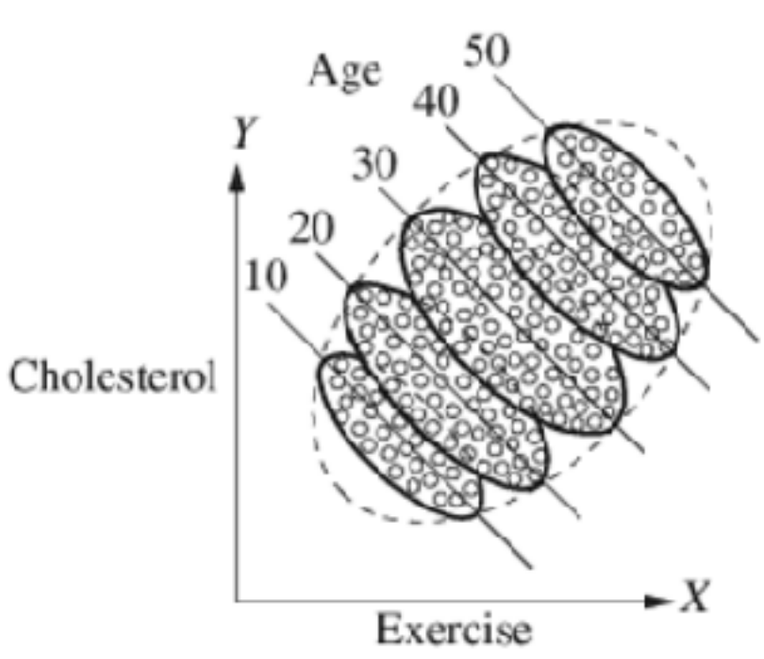


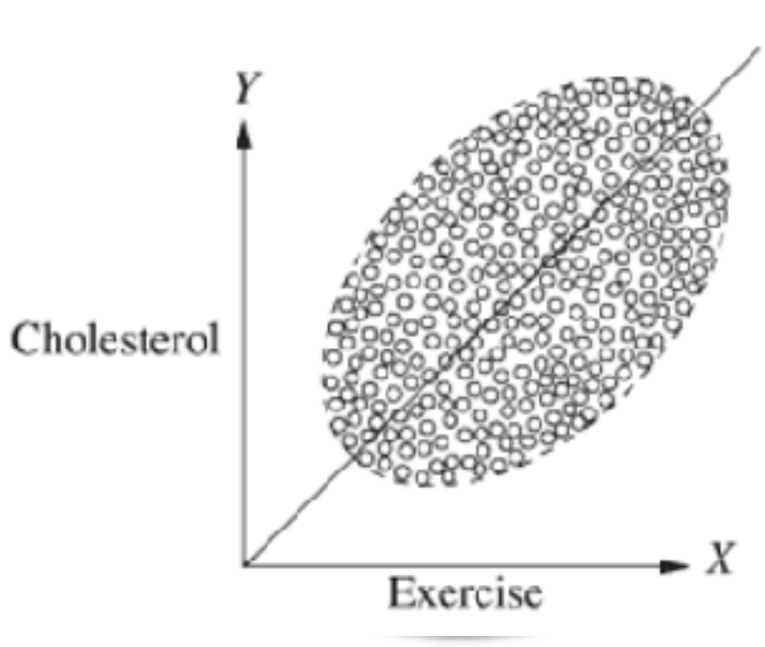
- "Strange" dependence

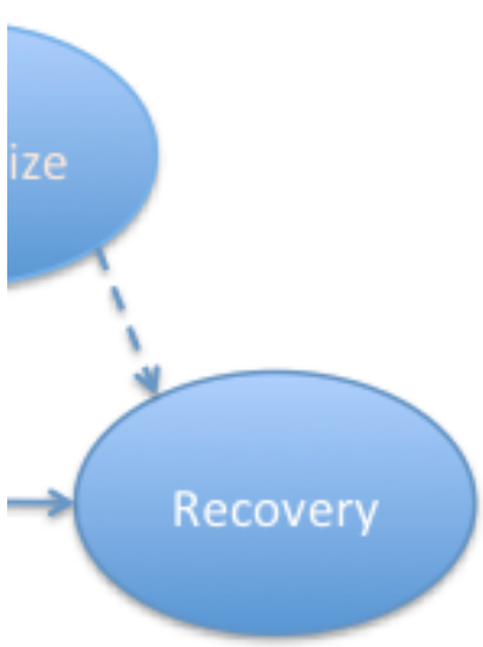
Causal Thinking (2)

- Dependence vs. causality

- Simpson's paradox







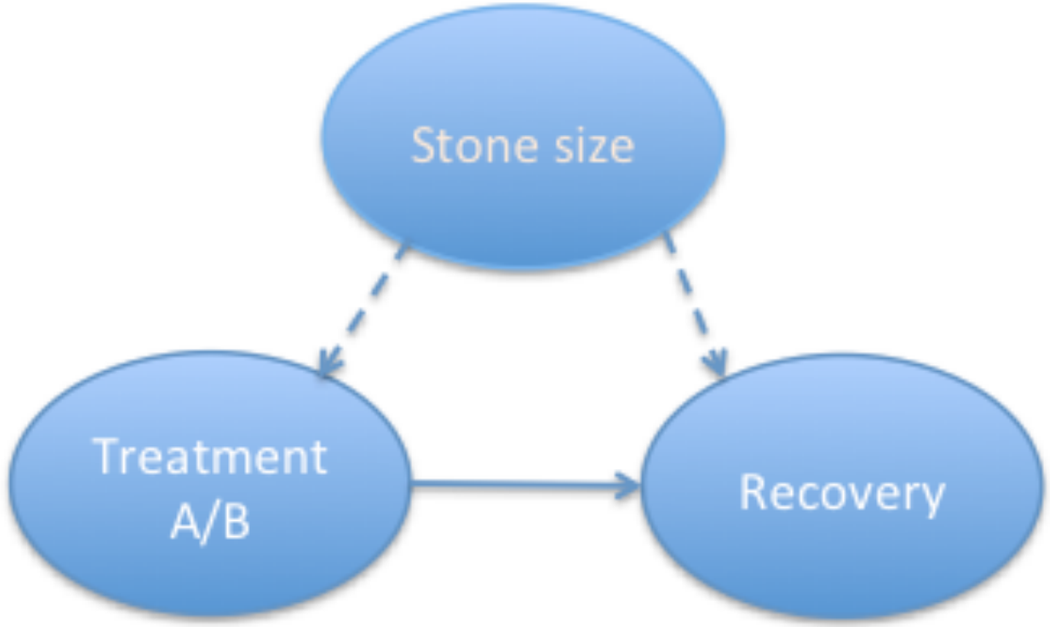
- "Strange" dependence

Causal Thinking (2)

- Dependence vs. causality

- 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)**



- "Strange" dependence

Causal Thinking (3)

- Dependence vs. causality

- Simpson's paradox

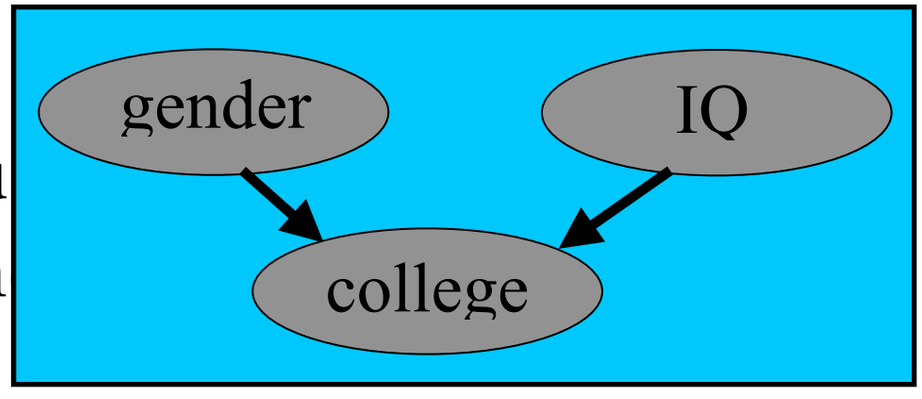
- "Stranger" dependence
  - Let's go back 50 years; maybe you'll find female college students are smarter than male ones on average. Why?

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

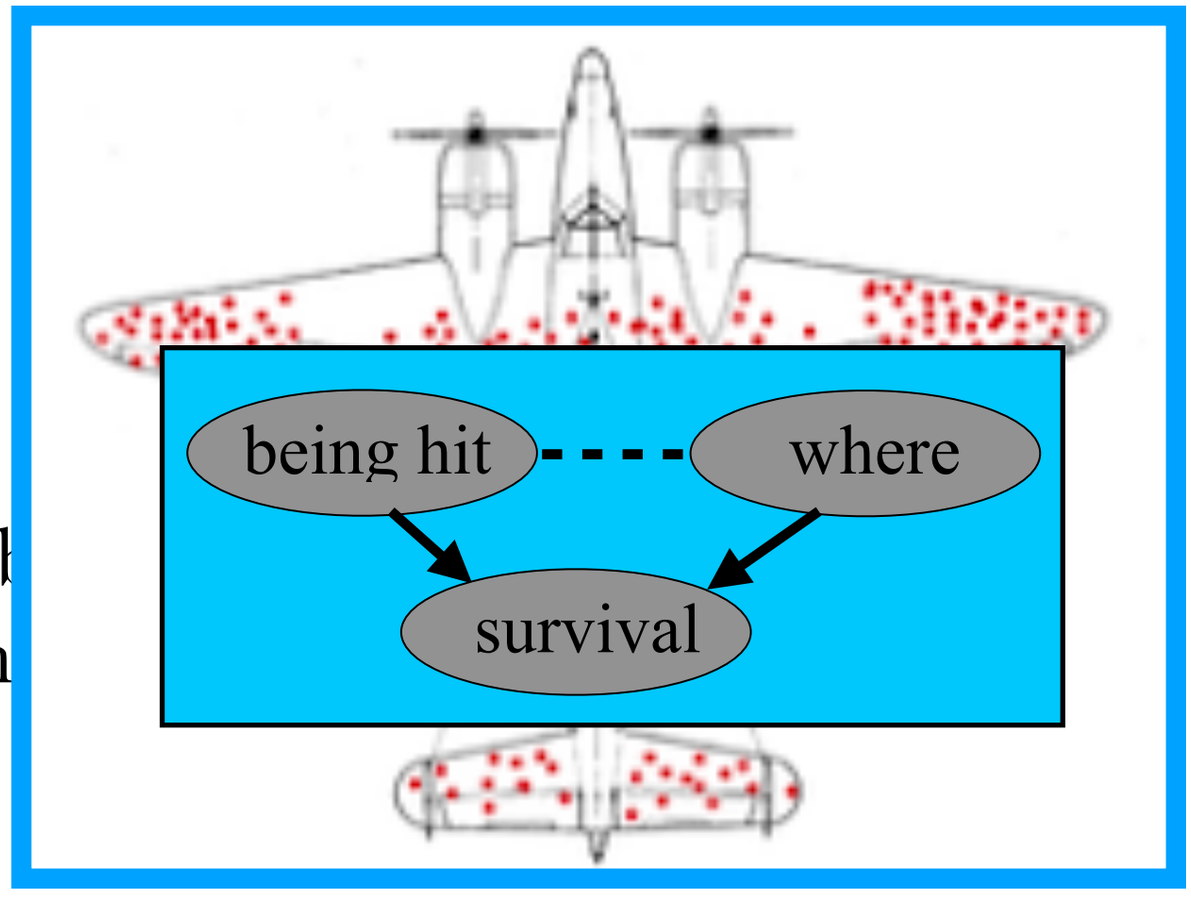


Causal Thinking (3)

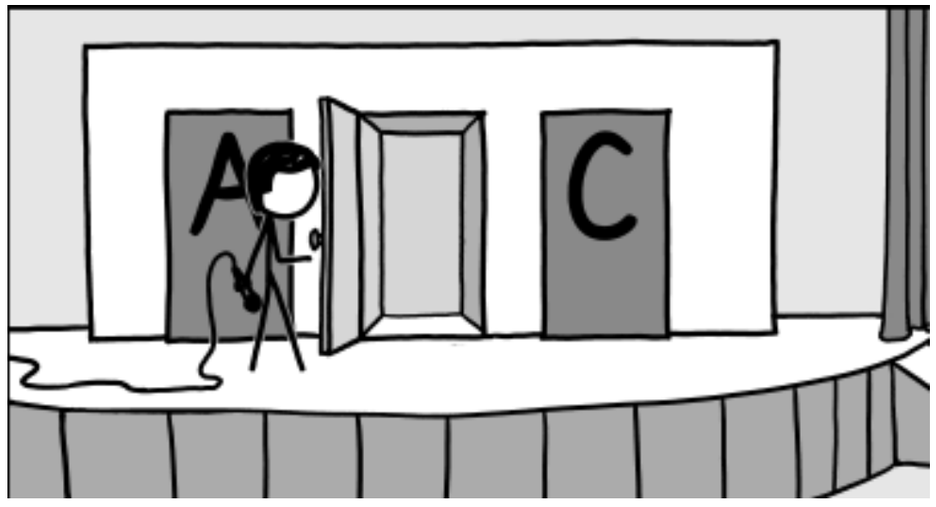
- Dependence vs. causality
- Simpson's paradox
- "Stranger" dependence
  - Let's go back 50 years; mayb
    students are smarter than m

Causal Thinking (3)

- Dependence vs. causality
- Simpson's paradox
- "Stranger" dependence
  - Let's go back 50 years; mayb
    students are smarter than m



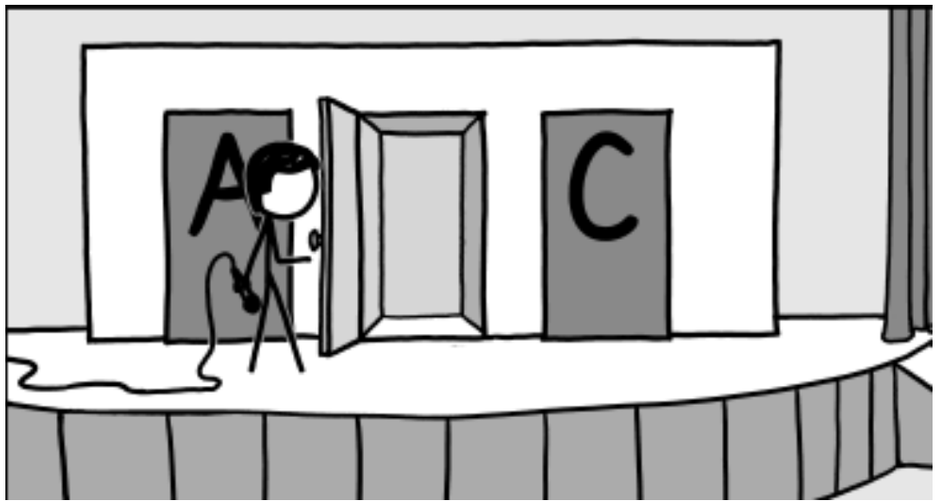
Question 3
Monty Hall Problem



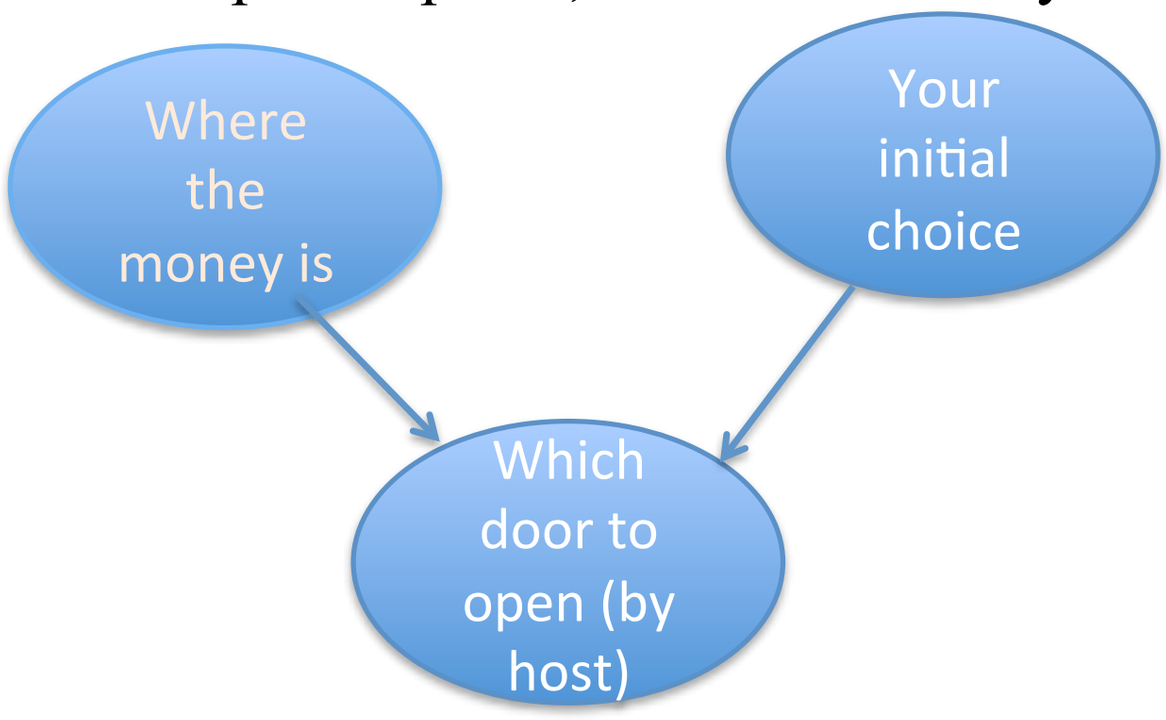
- You are a game show contestant. Before the game begins, the host, Monty Hall, has placed $1,000 dollars behind one of three doors. Nothing is behind the other two doors. The game is played as followed. You, the contestant, choose one of the doors, say, door A. Then Monty opens a door that is not the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

Question 2
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



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 + ...

“Causality” for Prediction: An Illustration

Understanding connections between different scenarios & *modeling* differences

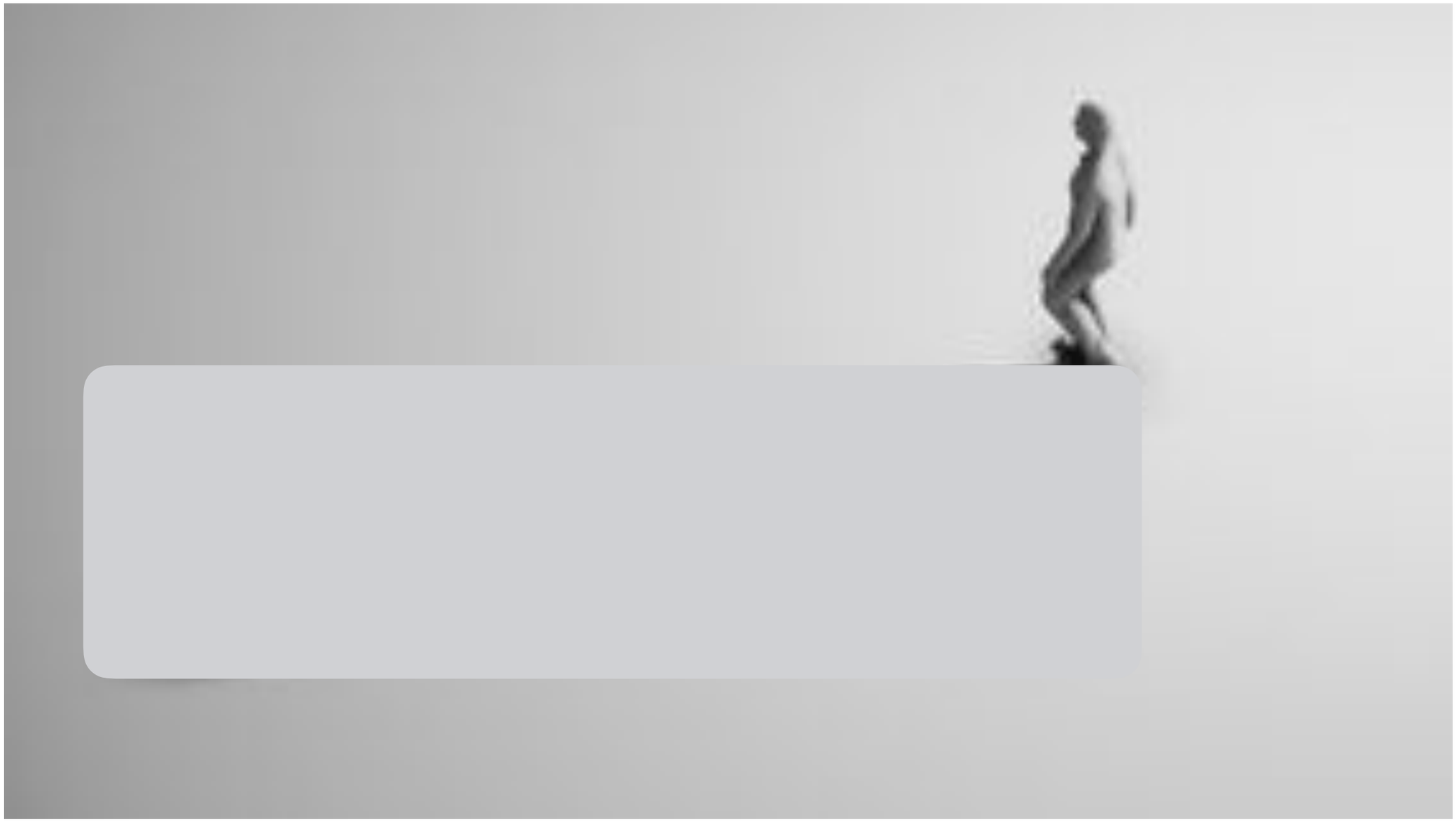
“Causality” for Prediction: An Illustration

Understanding connections between different scenarios & *modeling* differences

“Causality” for Prediction: An Illustration

Understanding connections between different scenarios & *modeling* differences

“Causality” for Prediction: An Illustration

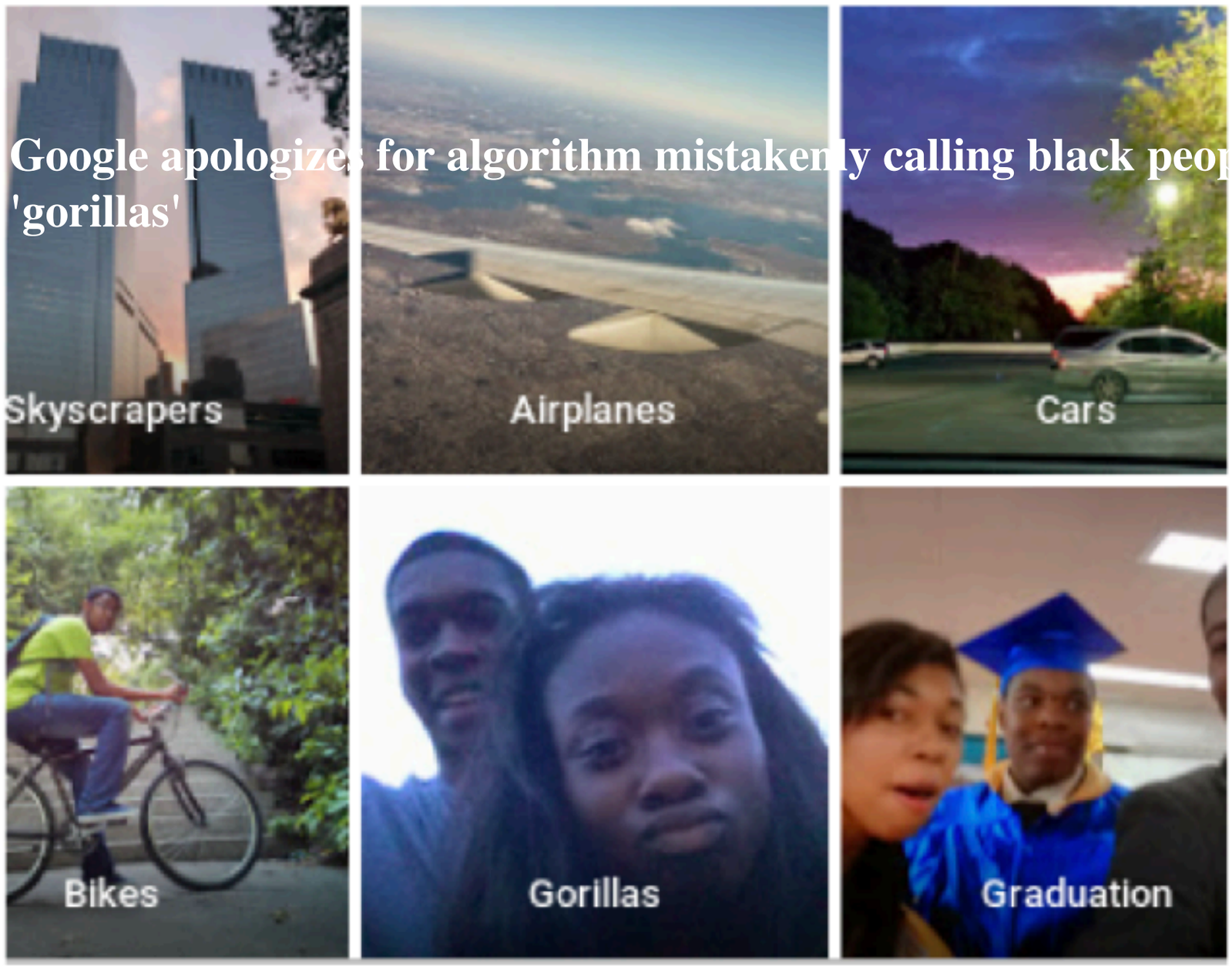


Understanding connections between different scenarios & *modeling* differences

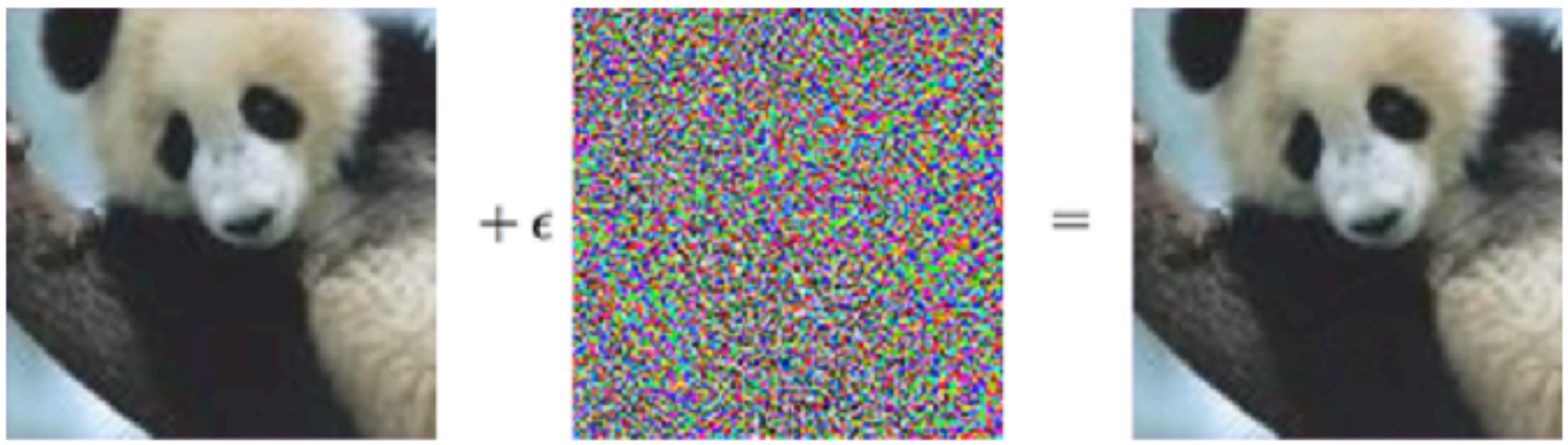
“Causality” for Prediction: An Illustration

Understanding connections between different scenarios & *modeling* differences

A Problem with Photo Categorization by Google Photos



Adversarial Attack



An adversarial input, overlaid on a typical image, can cause a classifier to miscategorize a panda as a gibbon.

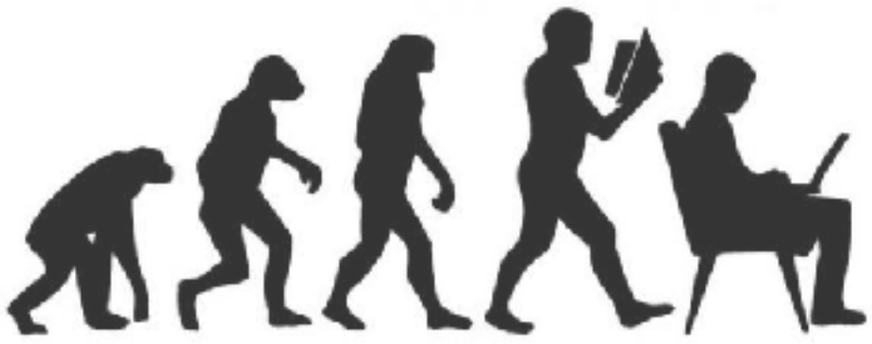
(Goodfellow, 2015)

Artificial “Intelligence”

- Traditional machine learning usually assumes a fixed data distribution; avoids overfitting

- Intelligence: understanding; control/intervention; decomposability; information fusion, learning with few examples, extrapolation

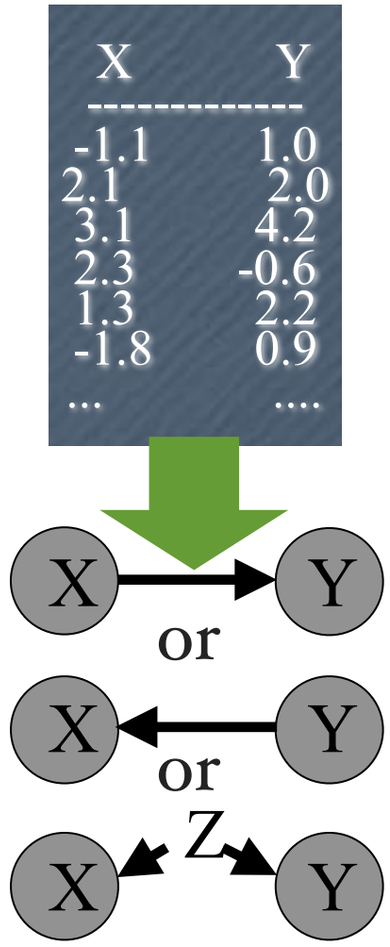
To Achieve "Artificial" Intelligence



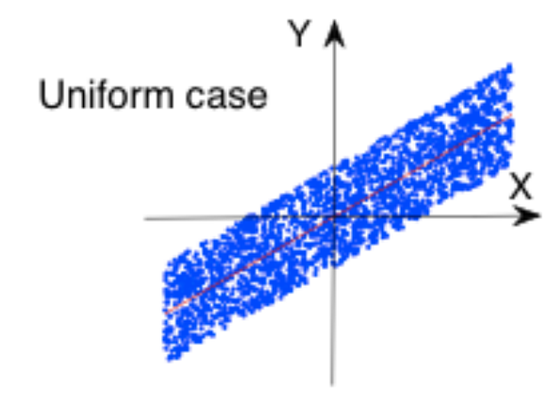
- Intelligence?
  - An evolution, selection, and growth perspective
- Two components
  - Survival: Good prediction/decision across scenarios
    - Inner compact representation...
  - Creativity
    - Causal representation

- The two types of representations seem consistent

Outline



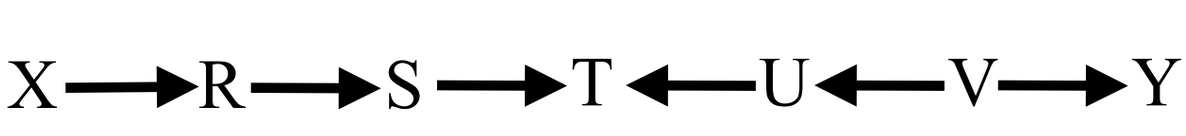
- Causality? Interventions? Causal thinking
- **Causal graphical models**
- Identification of causal effects
- Counterfactual reasoning
- Causal discovery
- Implications in machine learning

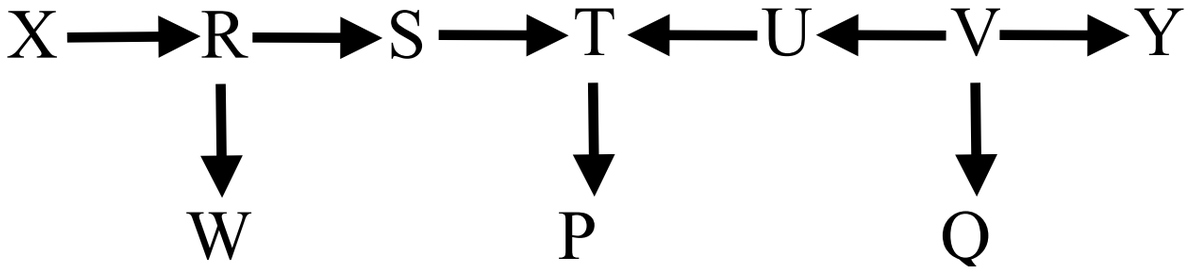


D-Separation

- A set of nodes **Z** d-separates two sets of nodes **X** and **Y** if every **path** from a node in **X** to a node in **Y** is blocked given **Z**.

- A path p is blocked by a set of nodes **Z** if
  - p contains a chain $i \rightarrow m \rightarrow j$ or a common cause $i \leftarrow m \rightarrow j$ such that the middle node m is in **Z**, **or**
  - p contains a collider $i \rightarrow m \leftarrow j$ such that the middle node m is in not **Z** and no descendant of m is in **Z**

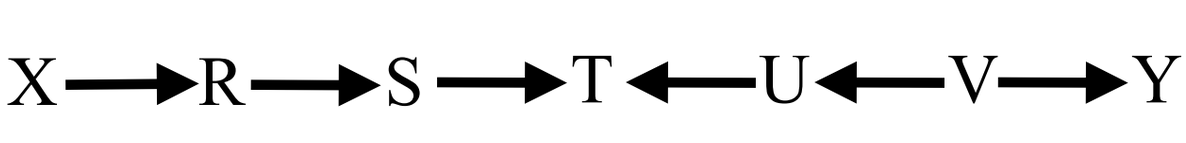




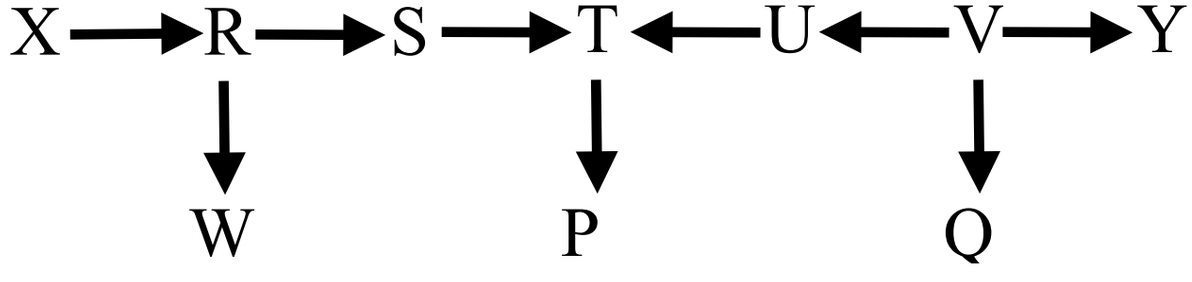
D-Separation

- A set of nodes **Z** d-separates two sets of nodes **X** and **Y** if every **path** from a node in **X** to a node in **Y** is blocked given **Z**.

- A path p is blocked by a set of nodes **Z** if
  - p contains a chain $i \rightarrow m \rightarrow j$ or a common cause $i \leftarrow m \rightarrow j$ such that the middle node m is in **Z**, **or**
  - p contains a collider $i \rightarrow m \leftarrow j$ such that the middle node m is in not **Z** and no descendant of m is in **Z**



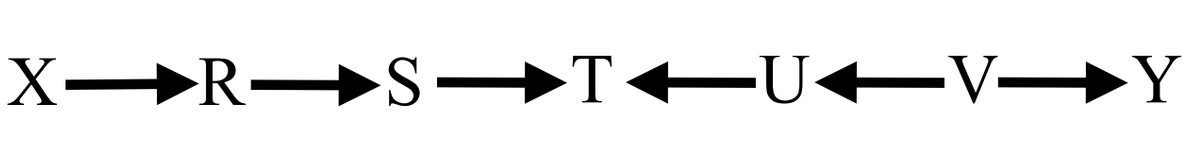
X and Y d-separated by {R,V}?
S and U d-separated by {R,V}?



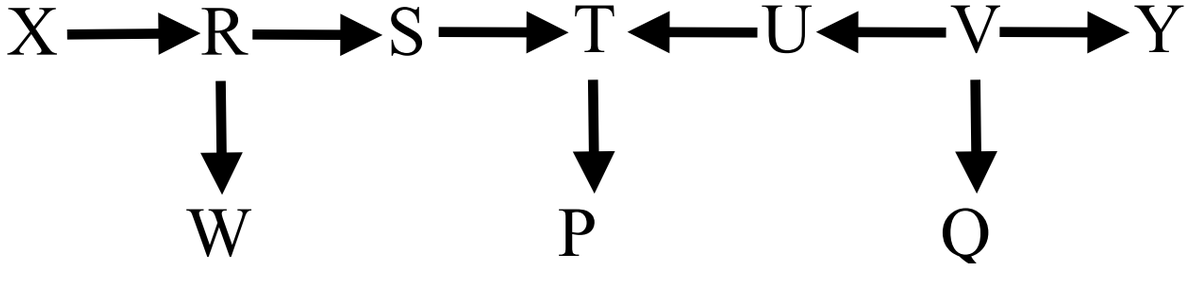
D-Separation

- A set of nodes **Z** d-separates two sets of nodes **X** and **Y** if every **path** from a node in **X** to a node in **Y** is blocked given **Z**.

- A path p is blocked by a set of nodes **Z** if
  - p contains a chain $i \rightarrow m \rightarrow j$ or a common cause $i \leftarrow m \rightarrow j$ such that the middle node m is in **Z**, **or**
  - p contains a collider $i \rightarrow m \leftarrow j$ such that the middle node m is in not **Z** and no descendant of m is in **Z**



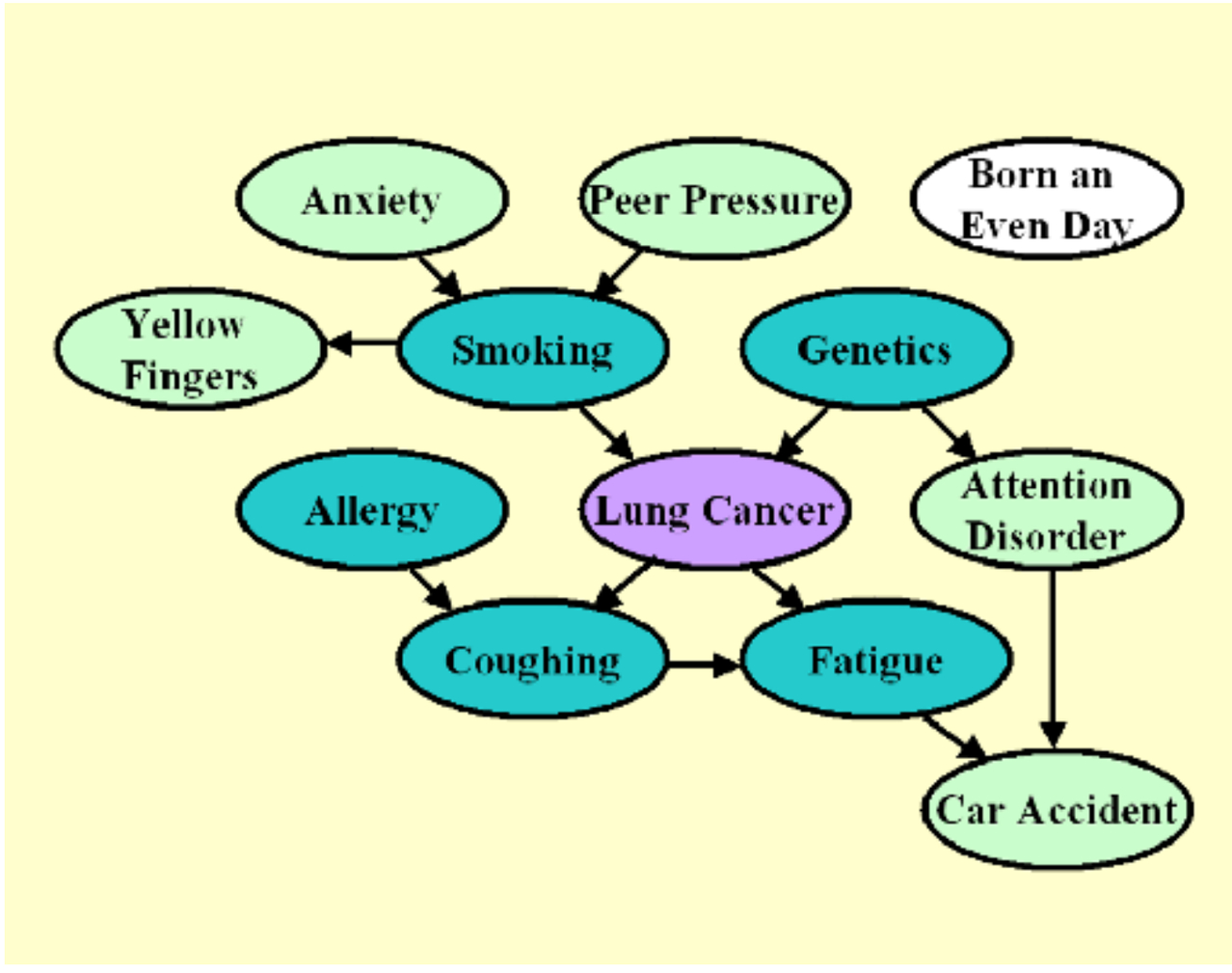
X and Y d-separated by {R,V}?
S and U d-separated by {R,V}?



X and Y d-separated by {R, P}?

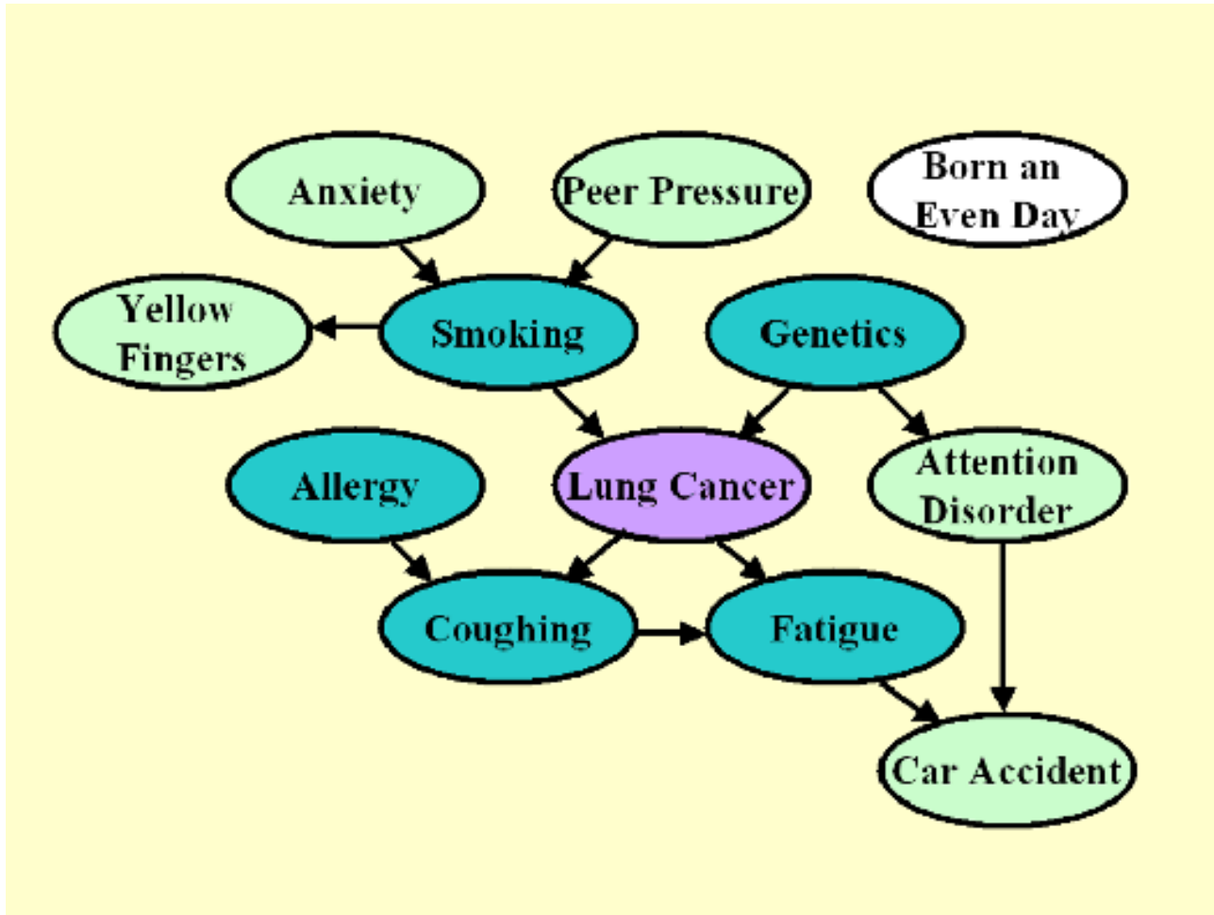
D-Separation: Intuition

- Suppose **X** and **Y** are d-separated by **Z**
- Then if you fix **Z**, **X** and **Y**
  - do not cause each other and
  - do not share a common cause
- **X** and **Y** are independent (conditional on **Z**)!



Local & Global Markov Conditions

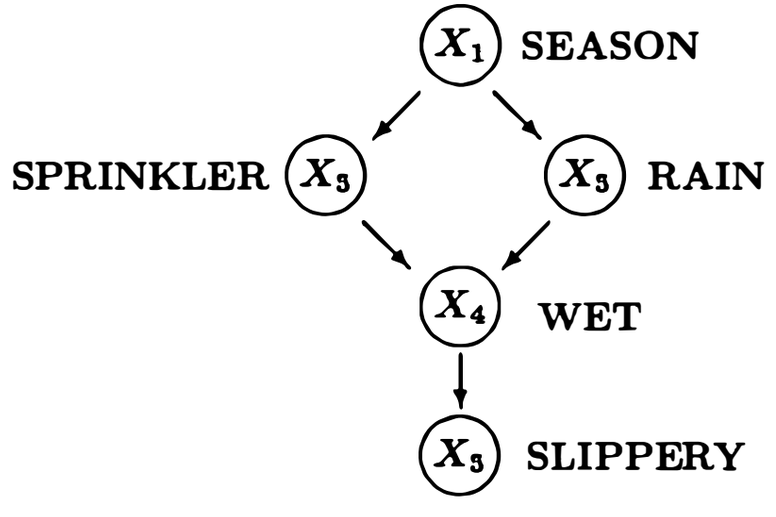
- **Local** Markov condition:
  - In a DAG, a variable X is independent of all its non-descendants given its parents

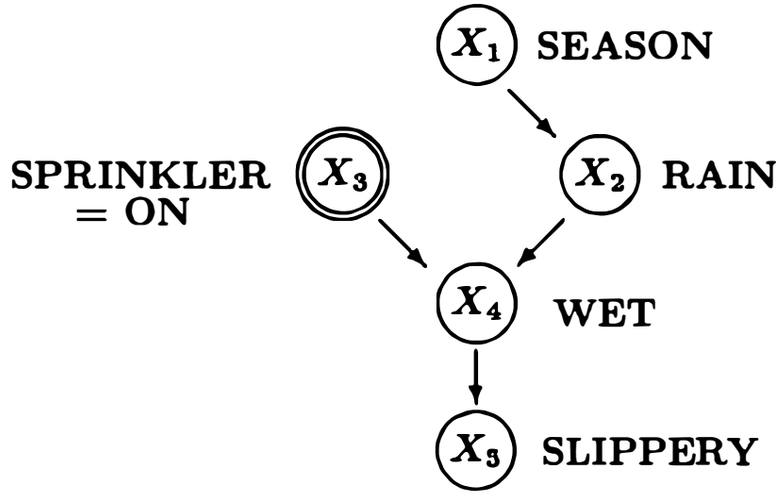


- **Global** Markov condition:
  - Given a DAG, let X and Y be two variables and $\mathbf{Z}$ be a set of variables that does not contain X or Y. If $\mathbf{Z}$ d-separates X and Y, then $X \perp\!\!\!\perp Y \mid \mathbf{Z}$.
- Actually equivalent on DAGs!

Causal Bayesian Networks (CBNs)

- Bayesian networks: DAGs
- Causal Bayesian networks
  - More meaningful & able to **represent and respond to external or spontaneous changes**

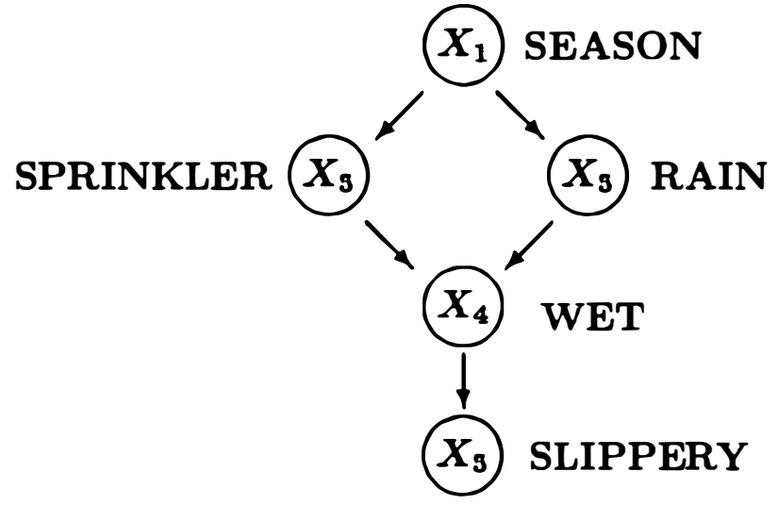




What is
$P_{X3=ON}(X_1,X_2,X_4,X_5)$?

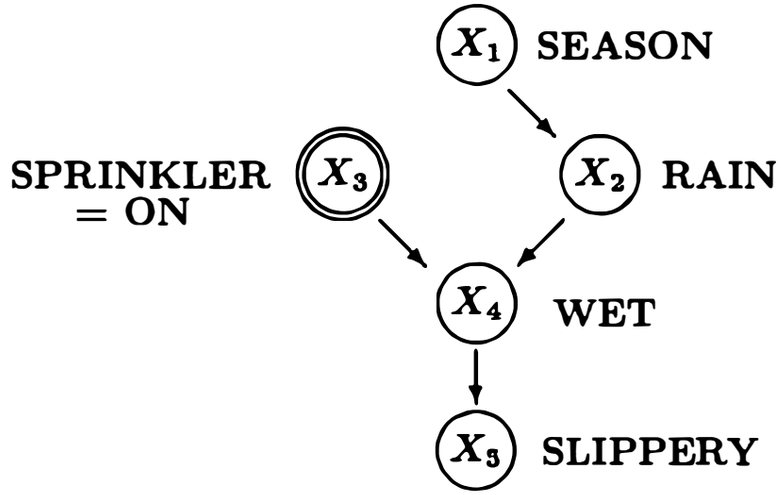
Causal Bayesian Networks (CBNs)

- Bayesian networks: DAGs
- Causal Bayesian networks
  - 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 CBN 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

$$PA_i \longrightarrow X_i$$

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

Structural Causal Models

$$PA_i \longrightarrow X_i$$

- $X_i = f_i(PA_i, E_i),\ i=1,\ldots,n$

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

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

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

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

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

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

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

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

Structural Causal Models

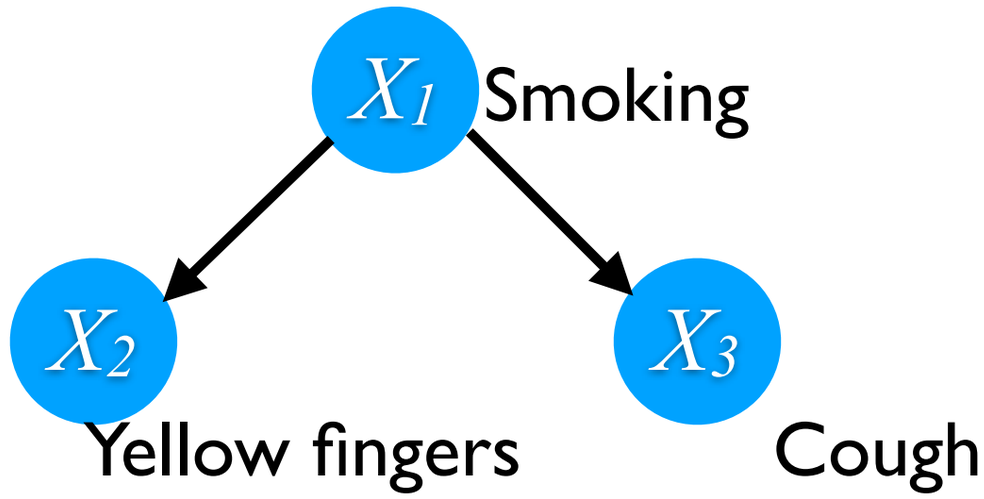
$$PA_i \longrightarrow X_i$$

- $X_i = f_i(PA_i, E_i),\ i=1,\ldots,n$
  - 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

Three Types of Problems in current AI

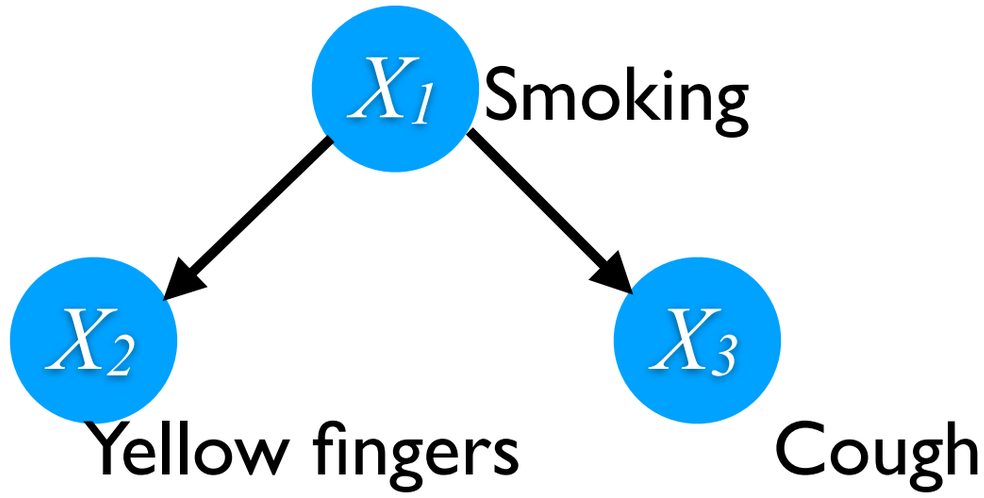


- Three questions:

X_1	X_2	X_3
1	0	0
0	0	1
0	1	1
1	1	1
0	0	0
0	1	0
1	1	1
1	1	1
0	0	0
1	0	0
...	...	...

- **Prediction**: Would the person cough if we *find* he/she has yellow fingers?
- **Intervention**: Would the person cough if we *make sure* that he/she has yellow fingers?
- **Counterfactual**: Would George cough *had* he had yellow fingers, *given that he does not have yellow fingers and coughs*?

Three Types of Problems in current AI



- Three questions:

X_1	X_2	X_3
1	0	0
0	0	1
0	1	1
1	1	1
0	0	0
0	1	0
1	1	1
1	1	1
0	0	0
1	0	0
...	...	...

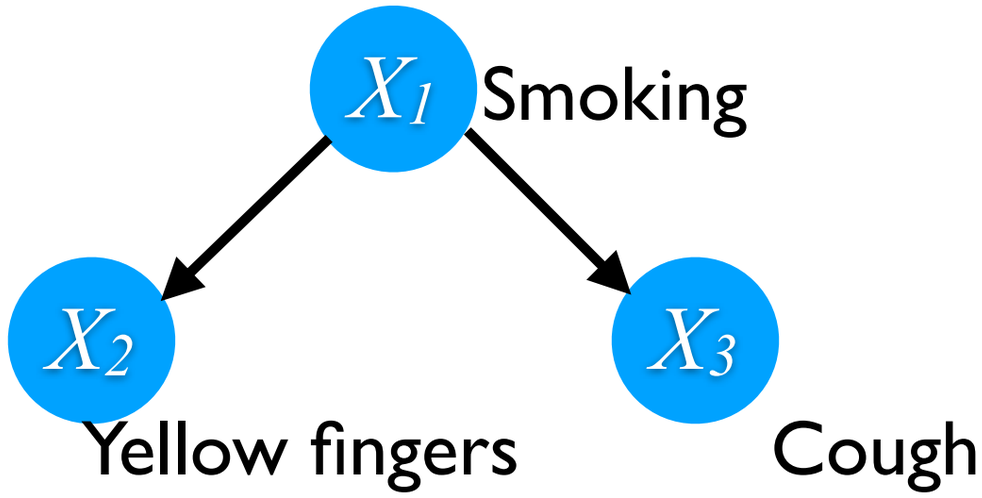
- **Prediction**: Would the person cough if we *find* he/she has yellow fingers?

$$P(X3 \mid X2=1)$$

- **Intervention**: Would the person cough if we *make sure* that he/she has yellow fingers?

- **Counterfactual**: Would George cough *had* he had yellow fingers, *given that he does not have yellow fingers and coughs*?

Three Types of Problems in current AI



- Three questions:

X_1	X_2	X_3
1	0	0
0	0	1
0	1	1
1	1	1
0	0	0
0	1	0
1	1	1
1	1	1
0	0	0
1	0	0
...	...	...

- **Prediction**: Would the person cough if we *find* he/she has yellow fingers?

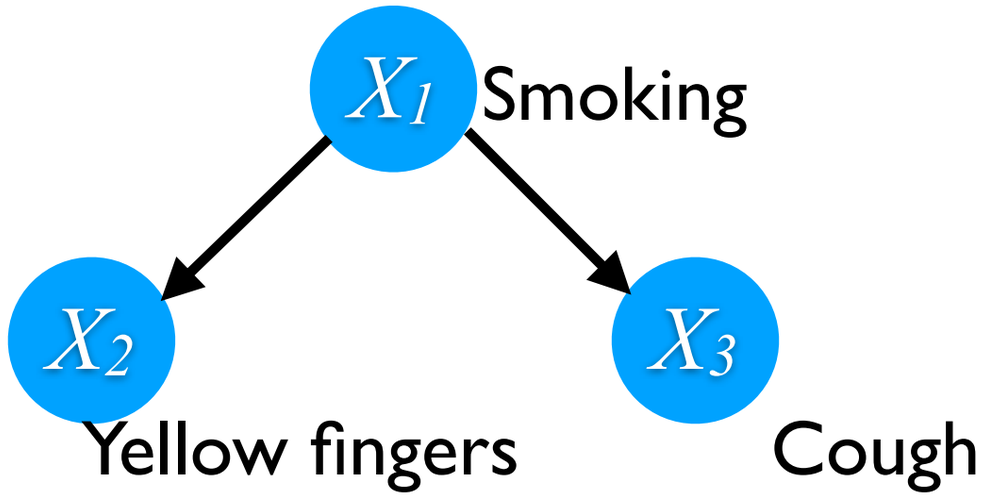
$$P(X3 \mid X2=1)$$

- **Intervention**: Would the person cough if we *make sure* that he/she has yellow fingers?

$$P(X3 \mid \text{do}\ (X2=1))$$

- **Counterfactual**: Would George cough *had* he had yellow fingers, *given that he does not have yellow fingers and coughs*?

Three Types of Problems in current AI



- Three questions:

X_1	X_2	X_3
1	0	0
0	0	1
0	1	1
1	1	1
0	0	0
0	1	0
1	1	1
1	1	1
0	0	0
1	0	0
...	...	...

- **Prediction**: Would the person cough if we *find* he/she has yellow fingers?

$$P(X3 \mid X2=1)$$

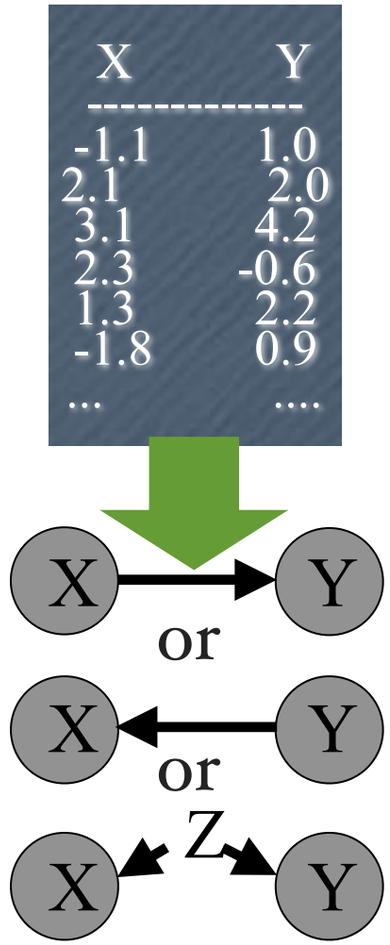
- **Intervention**: Would the person cough if we *make sure* that he/she has yellow fingers?

$$P(X3 \mid \text{do}\ (X2=1))$$

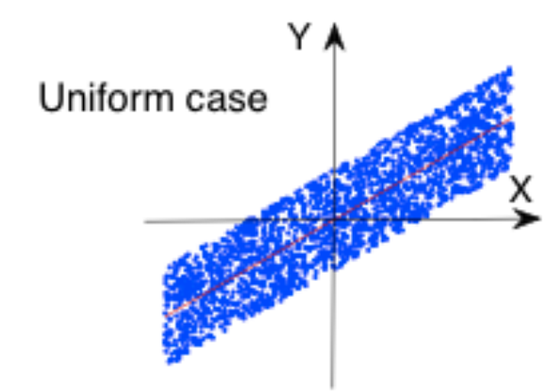
- **Counterfactual**: Would George cough *had* he had yellow fingers, *given that he does not have yellow fingers and coughs*?

$$P(X3_{\ X2=1} \mid X2 = 0, X3 = 1)$$

Outline

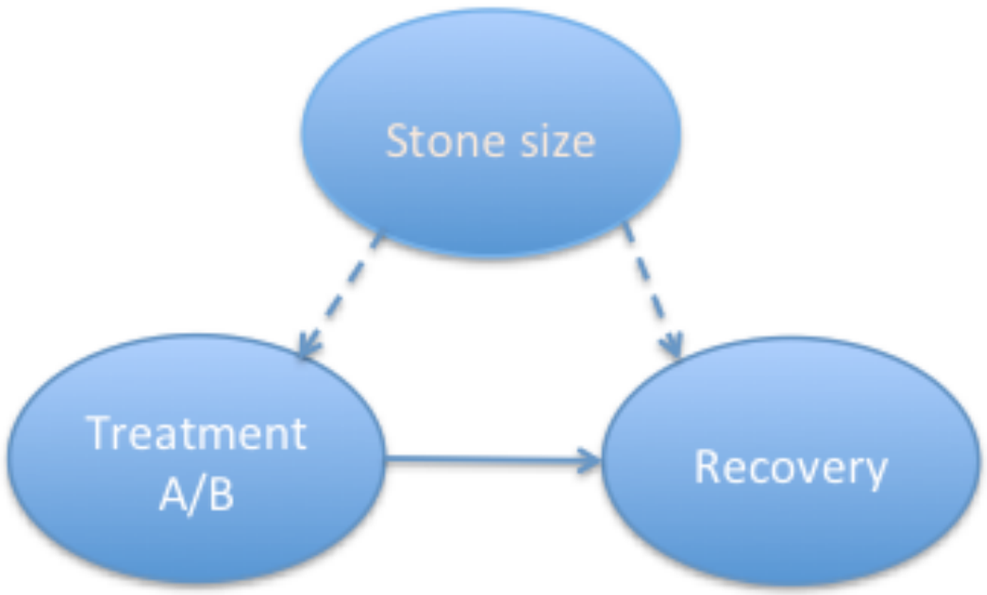


- Causality? Interventions? Causal thinking
- Causal graphical models
- **Identification of causal effects**
- Counterfactual reasoning
- Causal discovery
- Implications in machine learning



Identification of Causal Effects

- "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

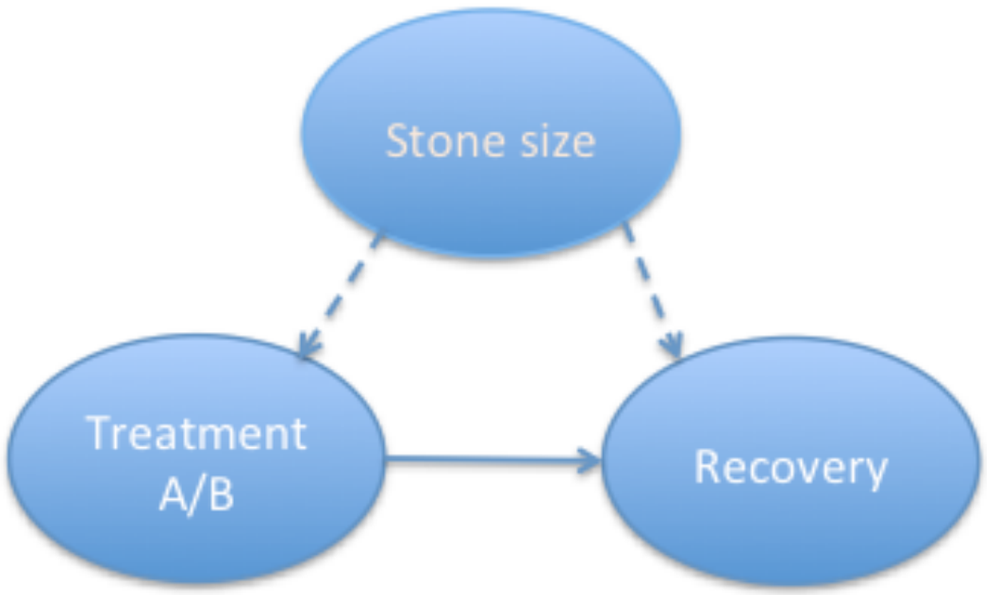


- Usually expensive or impossible to do!

Identification of Causal Effects

$$P(X3 \mid do\ (X2=1))$$

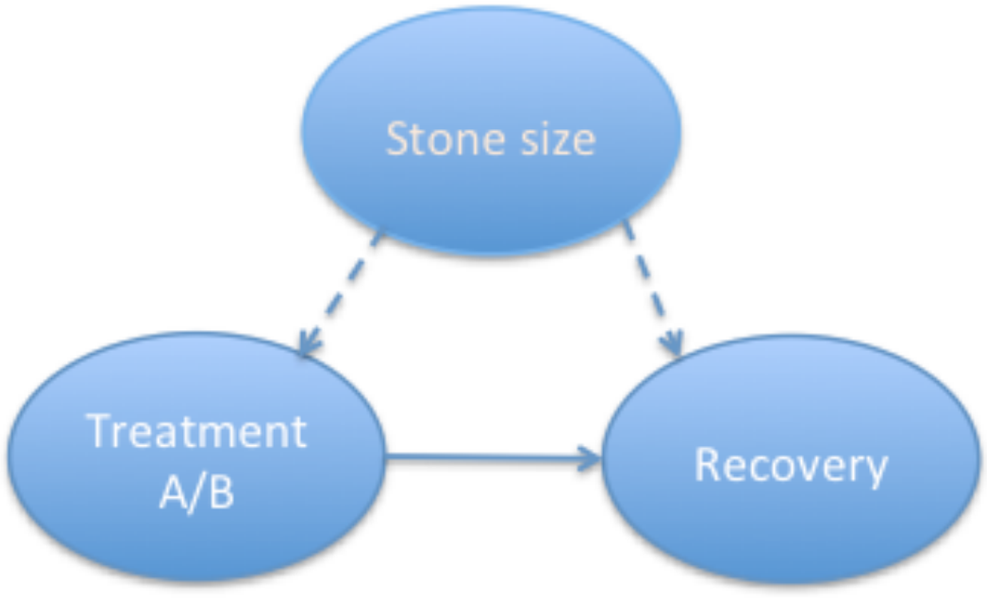
- "Golden standard": randomized controlled experiments
  - **All the other factors** that influence the outcome variable are either fixed or vary at random, so any changes in the outcome variable must be due to the controlled variable



- 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)**



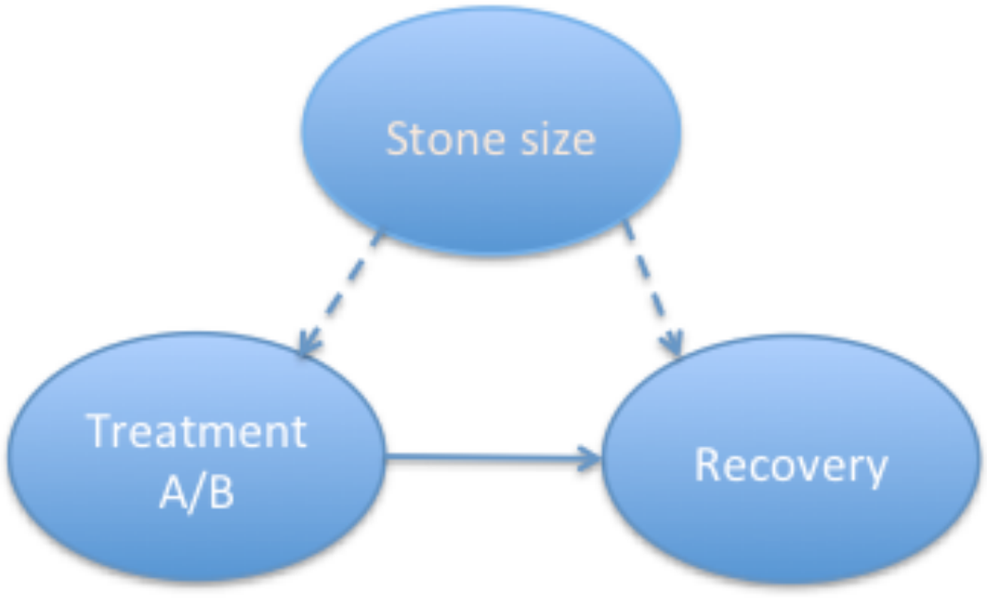
$$P(R|T) = \sum_S P(R|T,S)P(S|T)$$

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

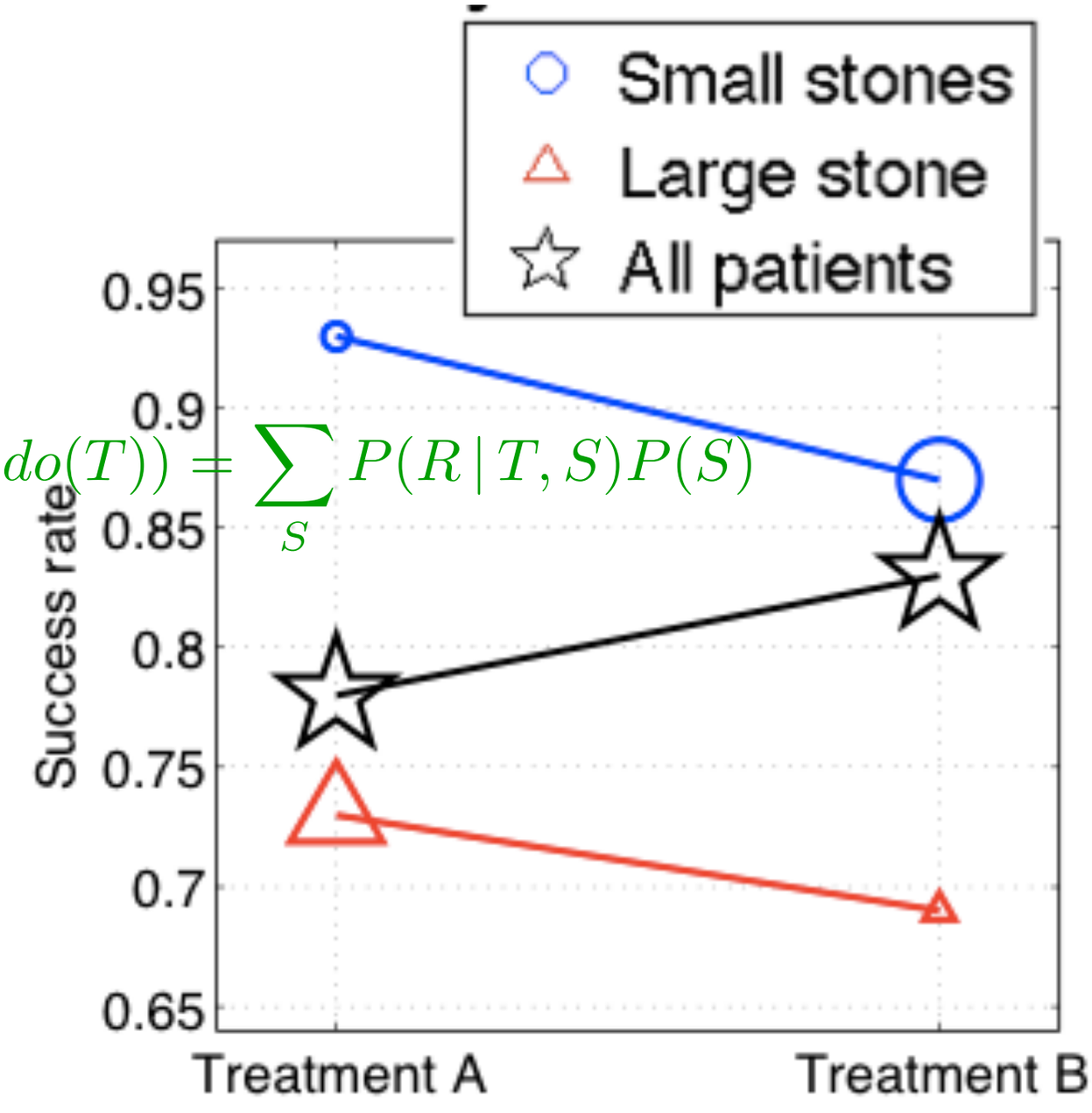
conditioning vs. manipulating

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)**



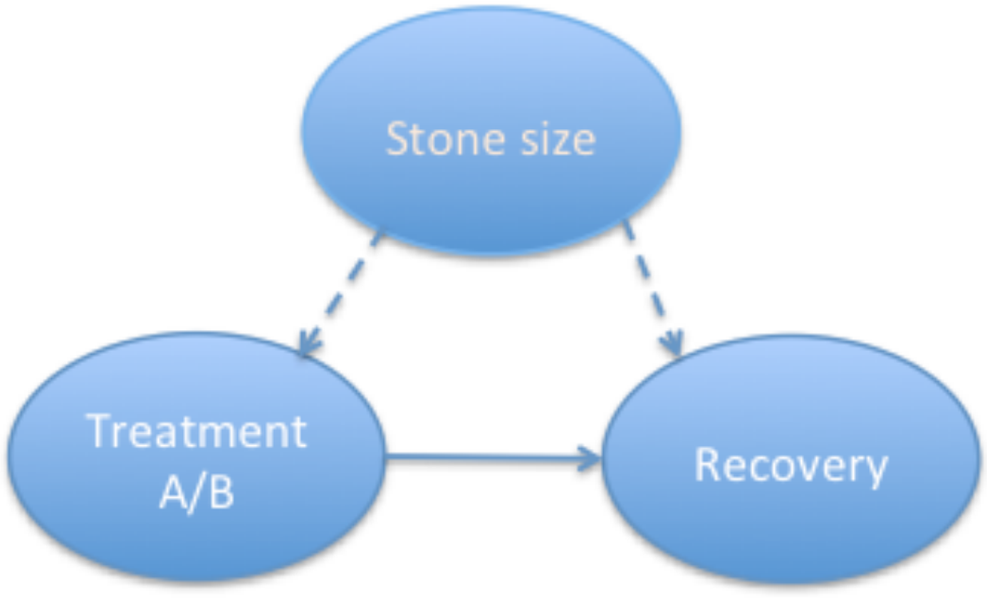
$$P(R \mid do(T)) = \sum_S P(R \mid T, S) P(S)$$



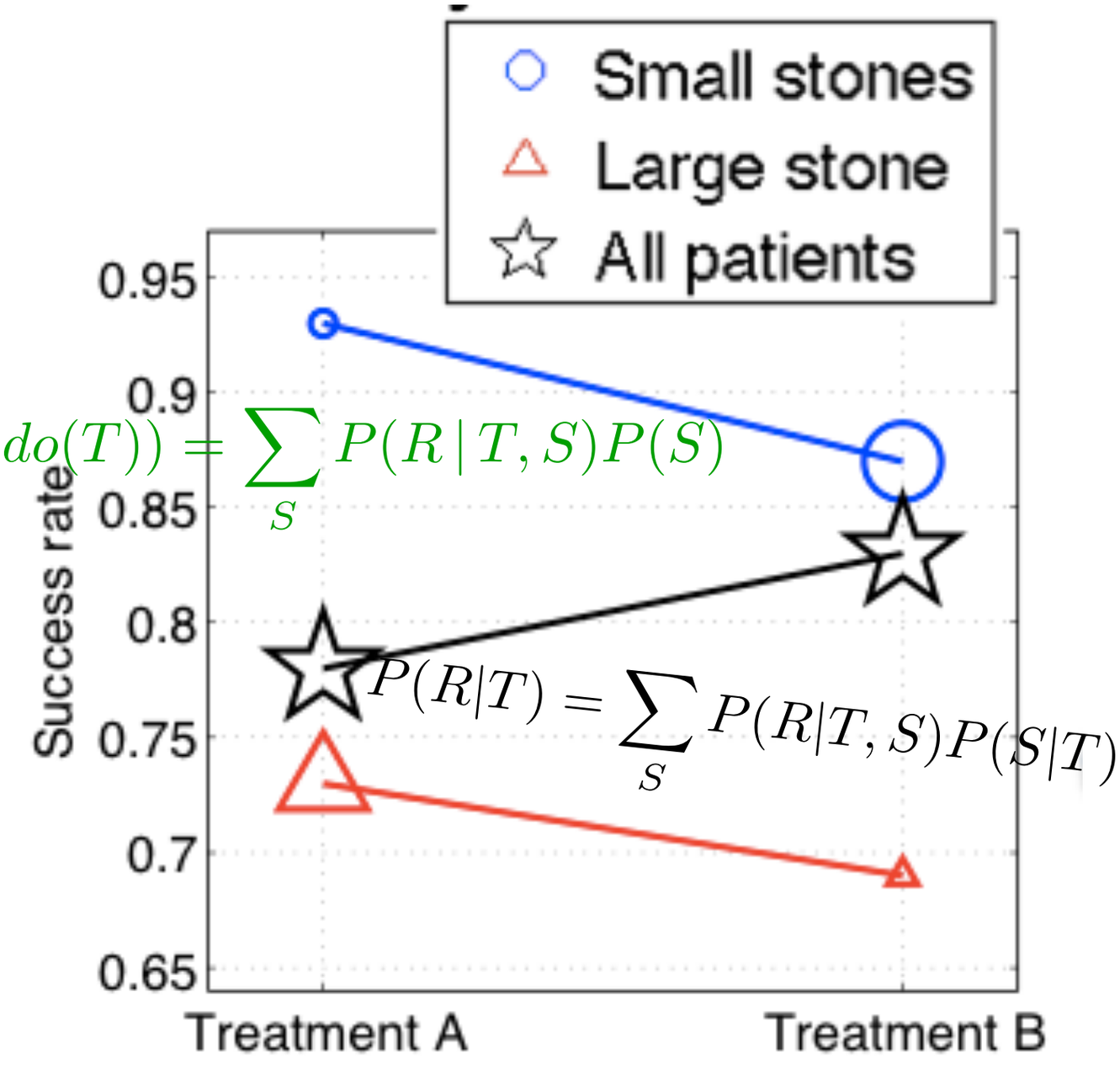
conditioning vs. manipulating

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)**



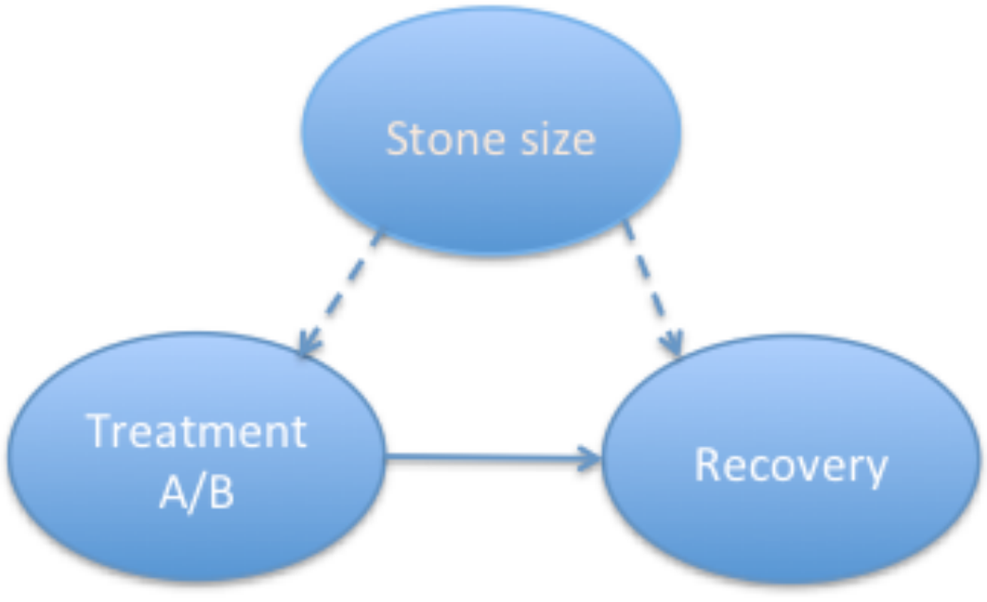
$$P(R \mid do(T)) = \sum_S P(R \mid T, S) P(S)$$



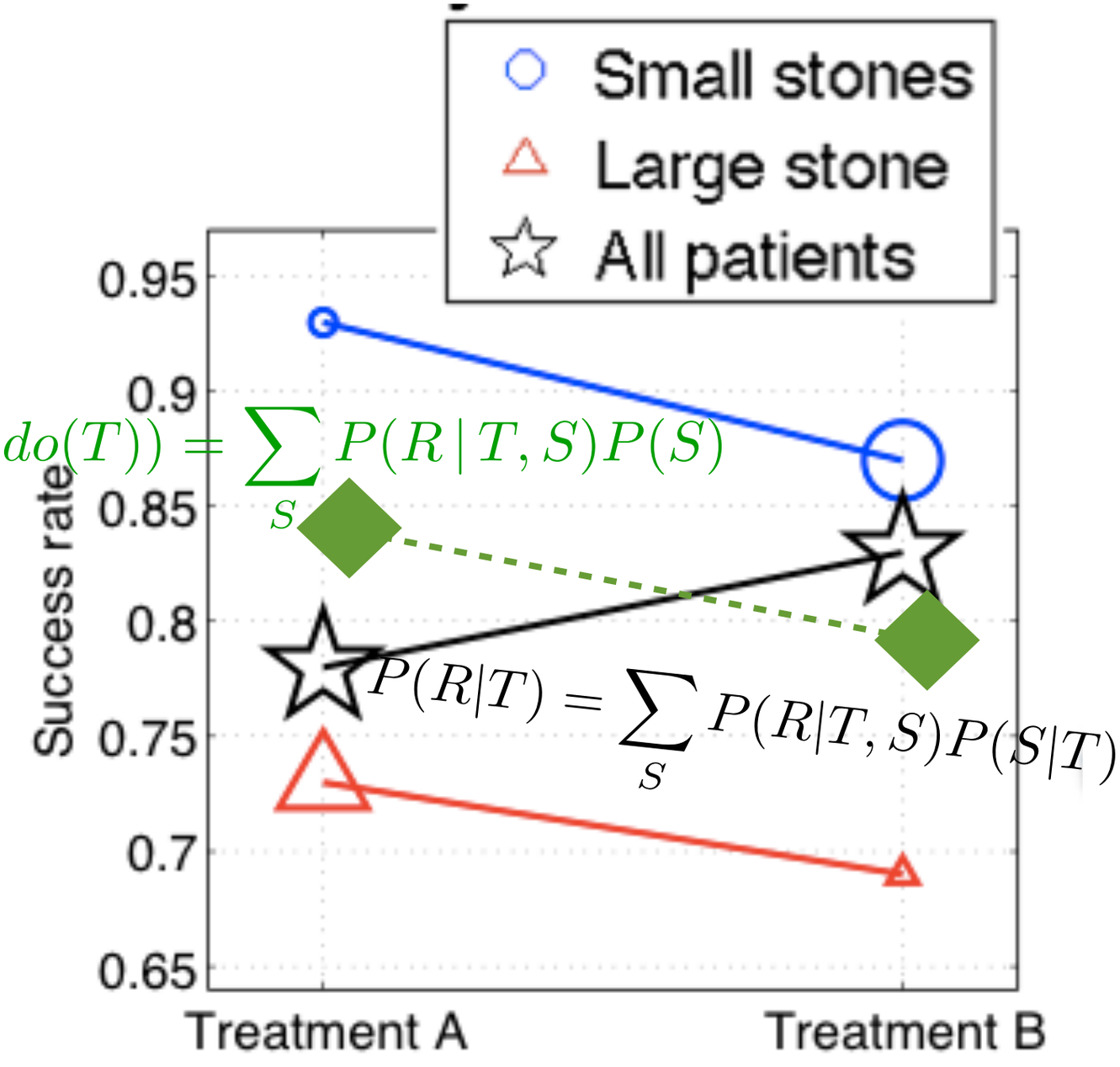
conditioning vs. manipulating

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)**



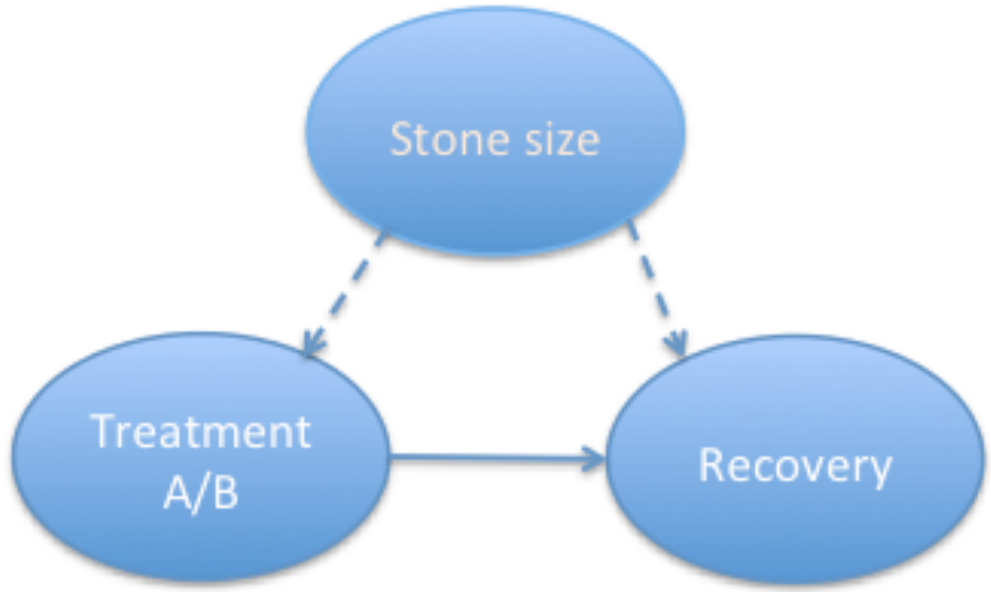
$$P(R \mid do(T)) = \sum_S P(R \mid T, S) P(S)$$



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)**

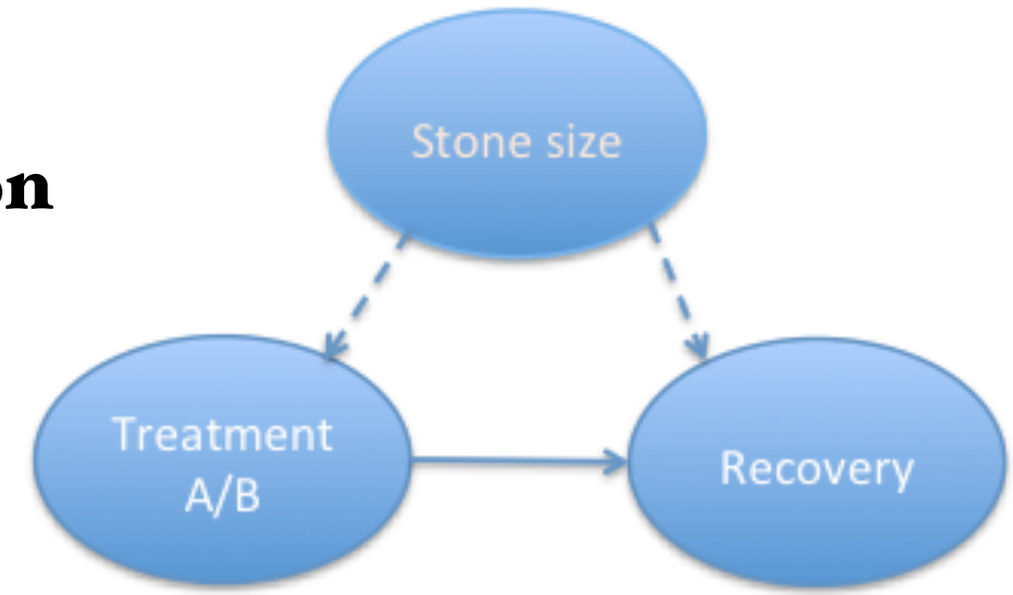
Given two disjoint sets of variables, X and Y, the causal effect *of X on Y, denoted either as* $P(y \mid \hat{x})$ *or as* $P(y \mid do(x))$, *is a function from X to the space of probability distributions on Y. For each realization x of X,* $P(y \mid \hat{x})$ *gives the probability of* $Y = y$ *induced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



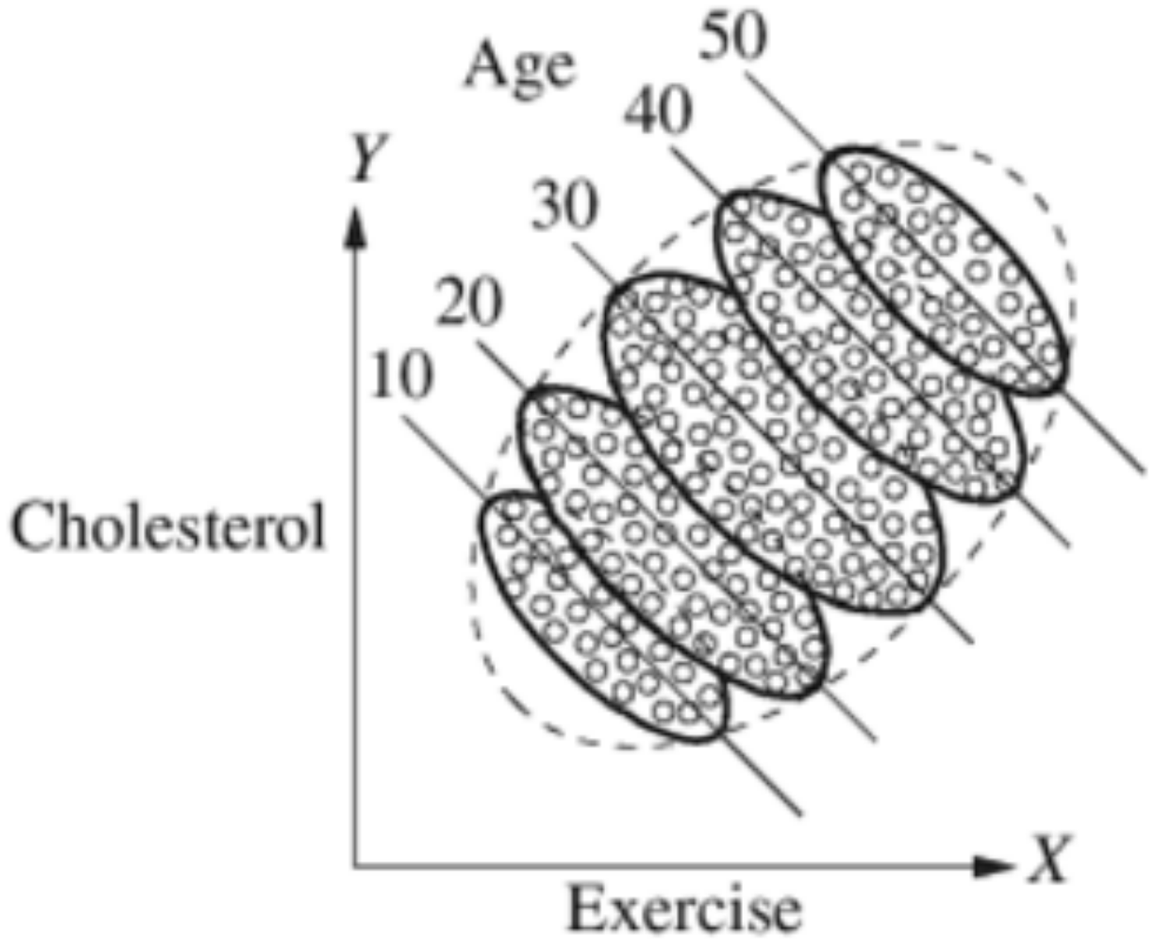
- How?

* **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

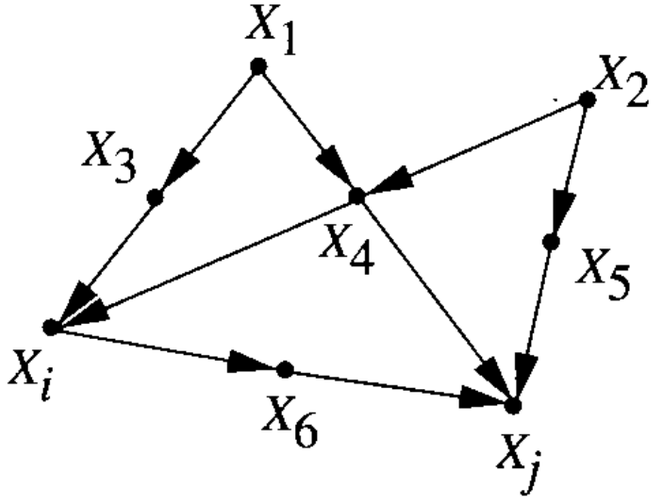


Back-Door Criterion

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.

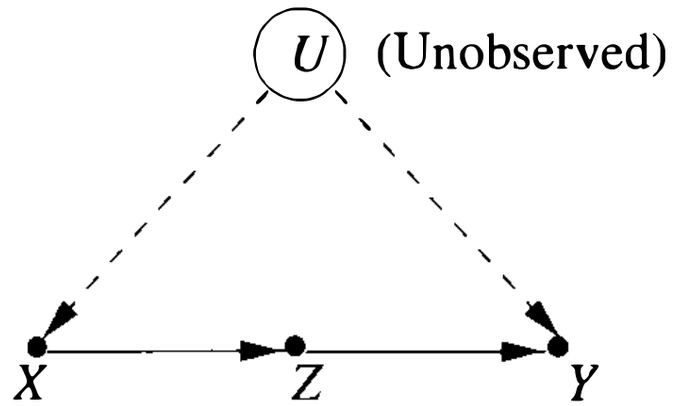


- What if $Z = \{X_3, X_4\}$?
  $Z = \{X_4, X_5\}$?
  $Z = \{X_4\}$?
- What if there is a confounder?

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*

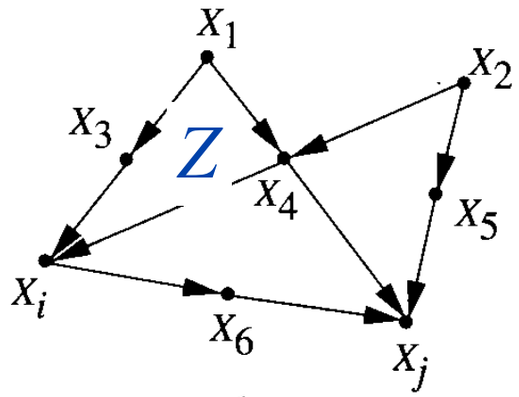
$$P(y \mid \hat{x}) = \sum_{z} P(z \mid x) \sum_{x'} P(y \mid x', z) P(x'). \tag{3.29}$$

* 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:

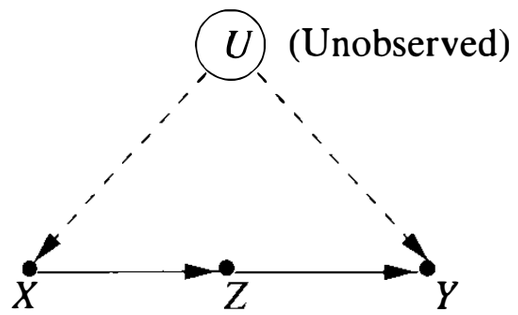
$$Y(x) \perp\!\!\!\perp X \mid Z.$$

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.*



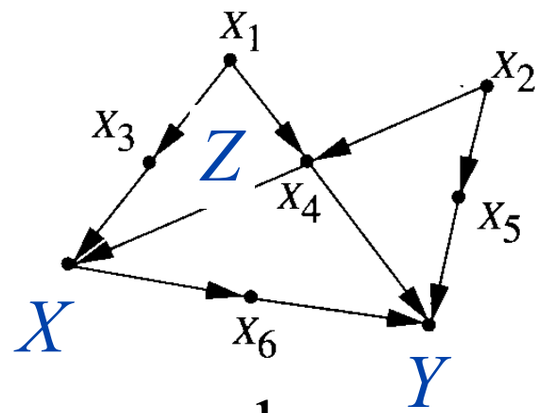
- $Y(z,x) = Y(z)$; $\{Y(z), X\} \perp\!\!\!\perp Z(x)$.

* 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:

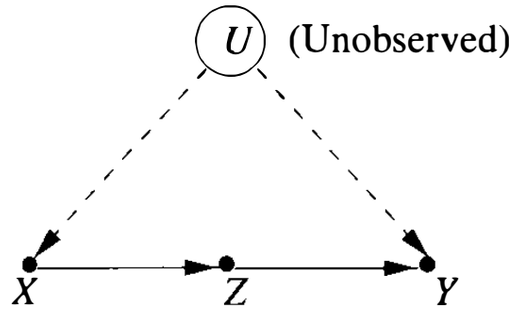
$$Y(x) \perp\!\!\!\perp X \mid Z.$$

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.*



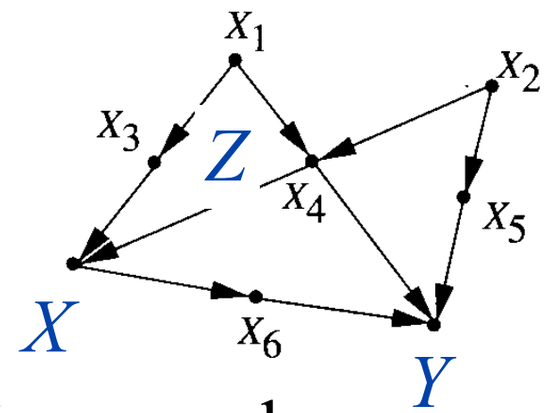
- $Y(z,x) = Y(z)$; $\{Y(z), X\} \perp\!\!\!\perp Z(x)$.

* 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:

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

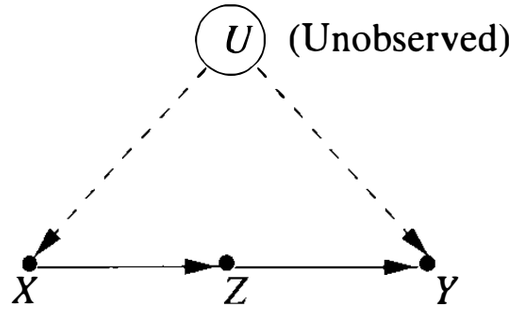
$Y(x,u)$: the value attained by Y in unit u under intervention do(x);
$Y(x)$: counterfactual variable (u is treated as a variable)

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*.



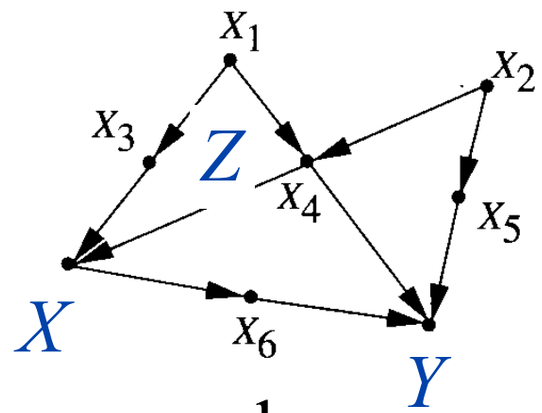
- $Y(z,x) = Y(z)$; $\{Y(z), X\} \perp\!\!\!\perp Z(x)$.

* 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:

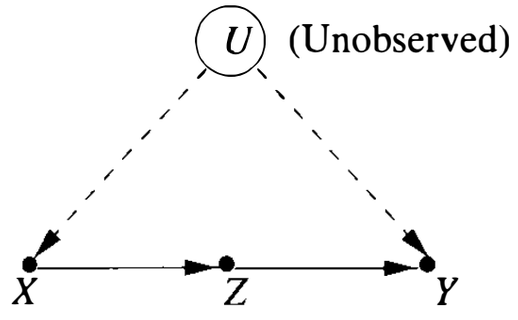
$$Y(x) \perp\!\!\!\perp X \mid Z.$$

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:

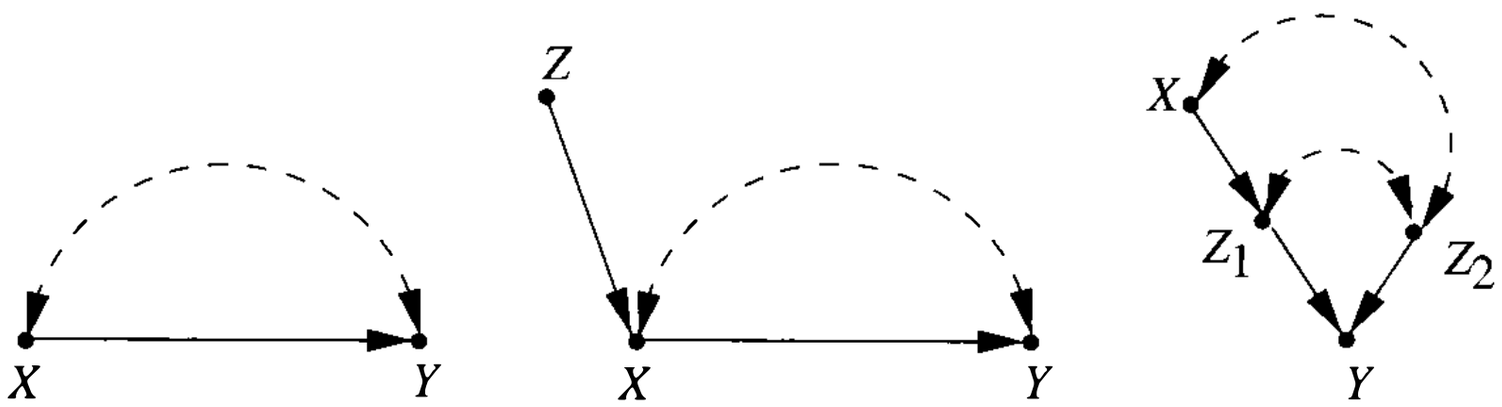
*

A 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:



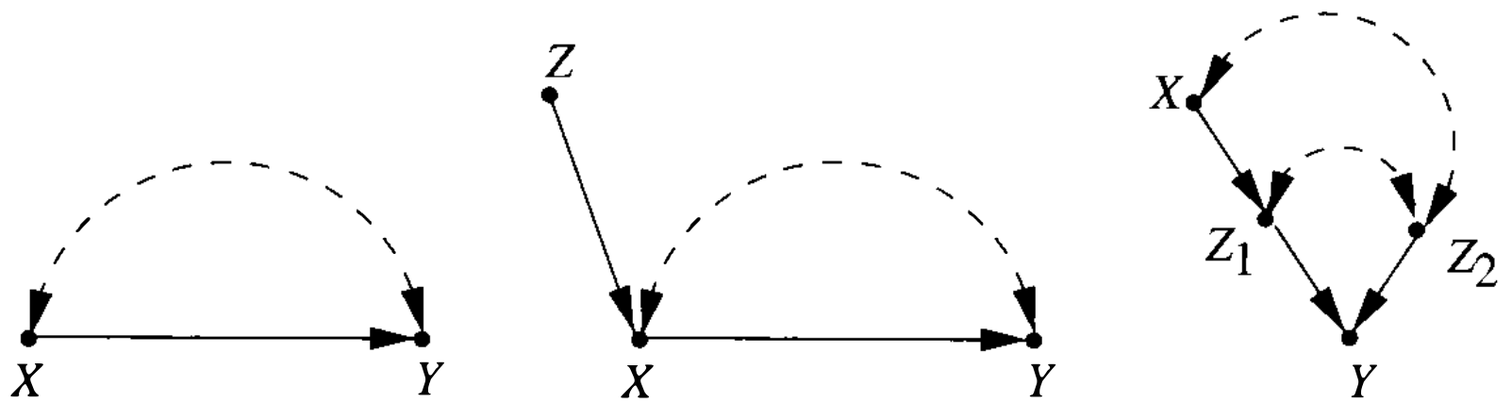
(a) (b) (c)

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 $P(y \mid \hat{x})$.

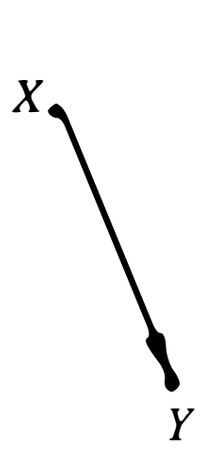
*

A Unification: Examples

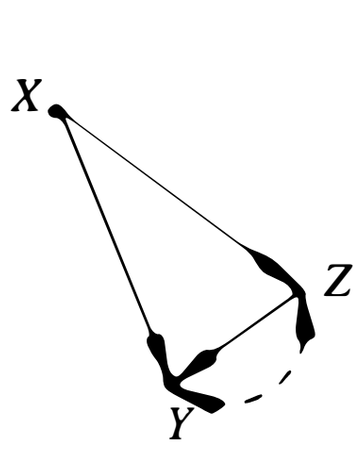
- Examples:

A Unification: Examples

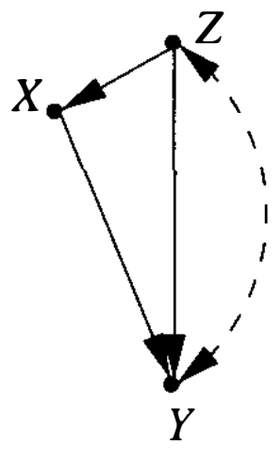
- Examples:



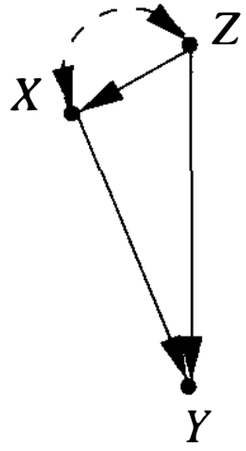
(a)



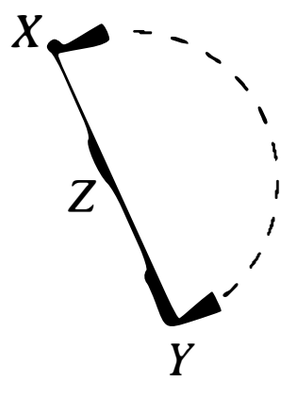
(b)



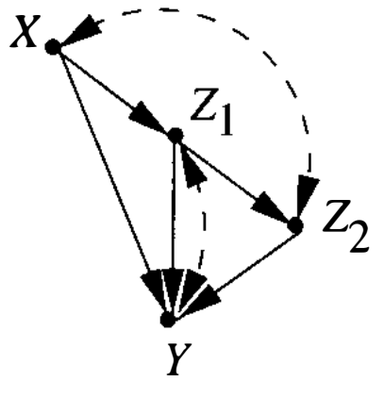
(c)



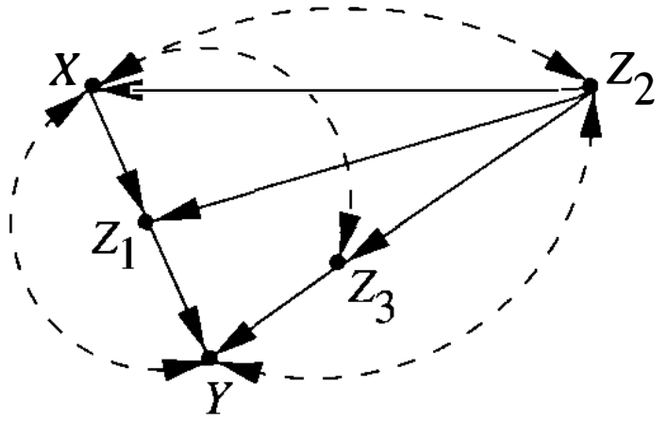
(d)



(e)



(f)



(g)

A Unification: Examples

- Examples:

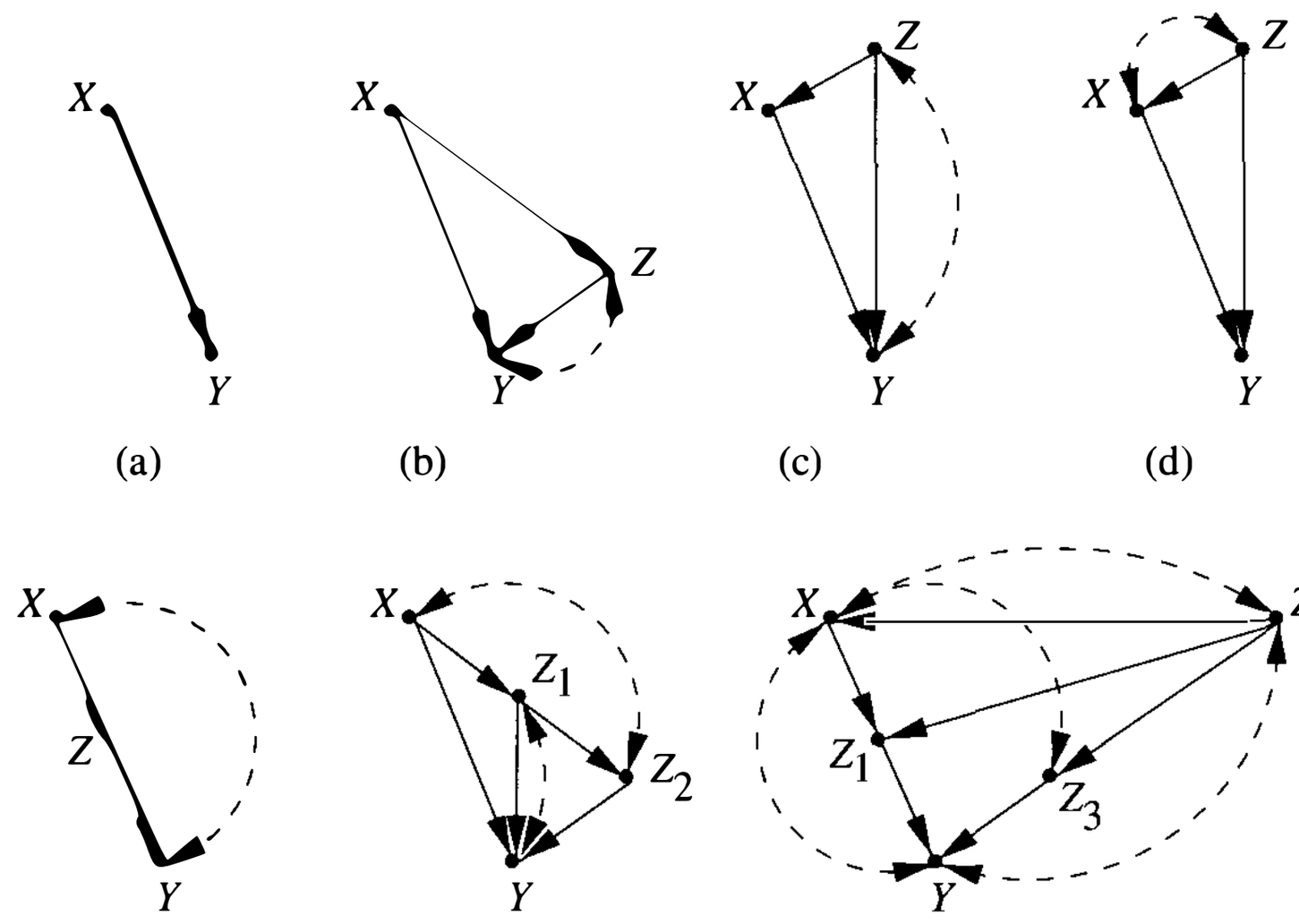


Figure 3.8 Typical models in which the effect of X on Y is identifiable. Dashed arcs represent confounding paths, and Z represents observed covariates.

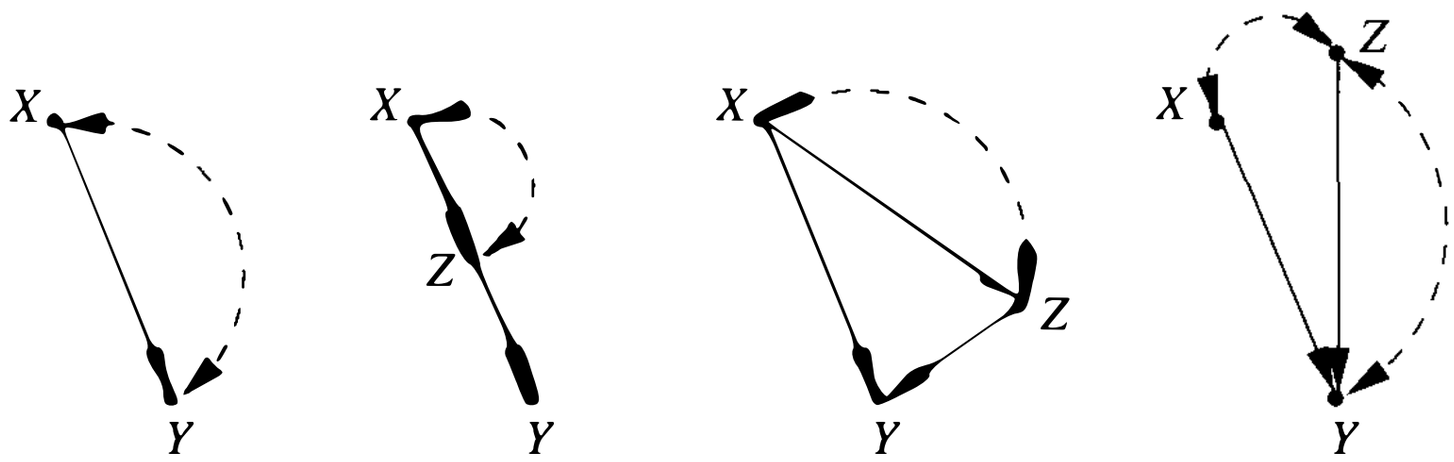
*

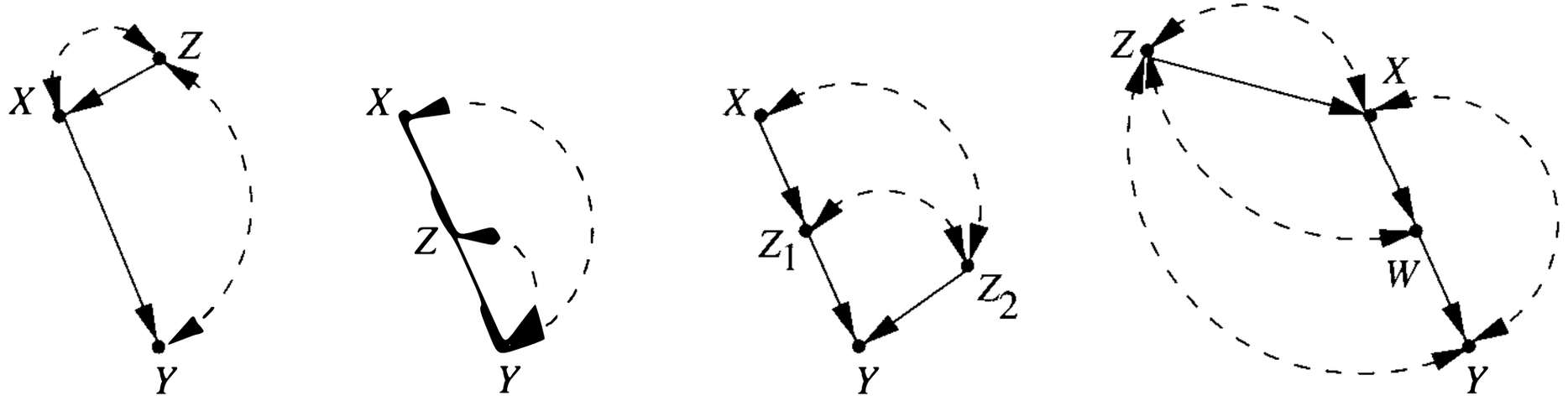
A Unification: Examples

- Examples:

A Unification: Examples

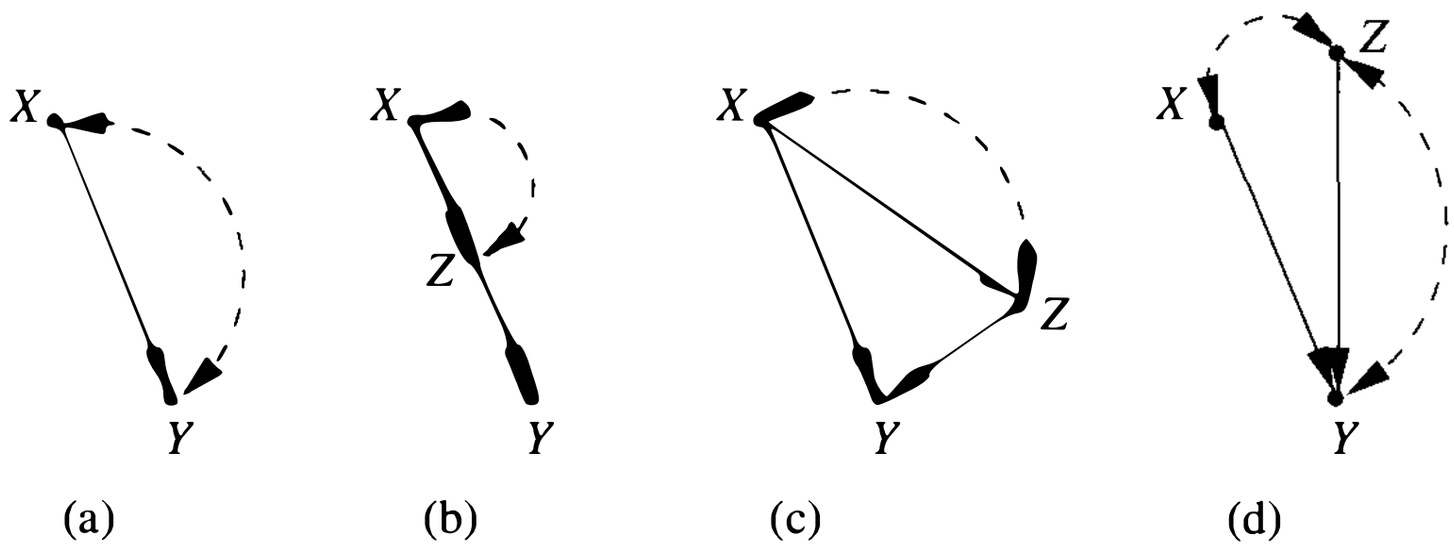
- Examples:





A Unification: Examples

- Examples:



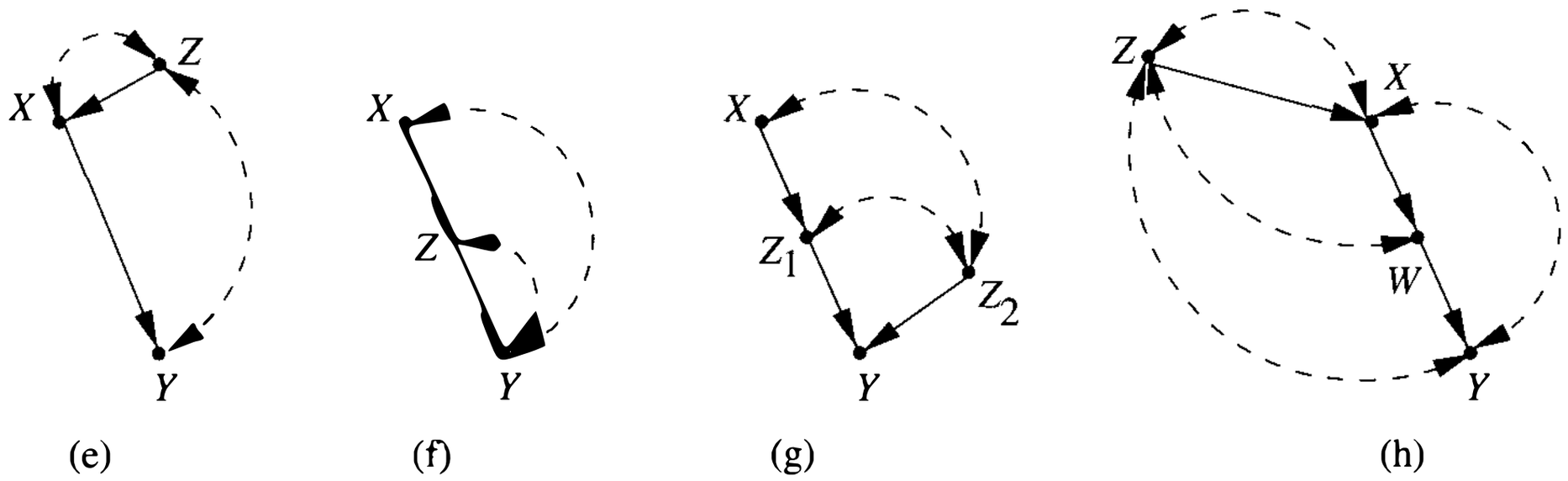
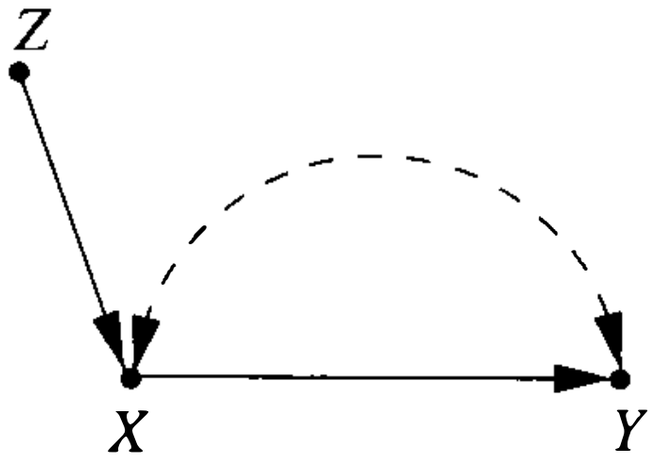
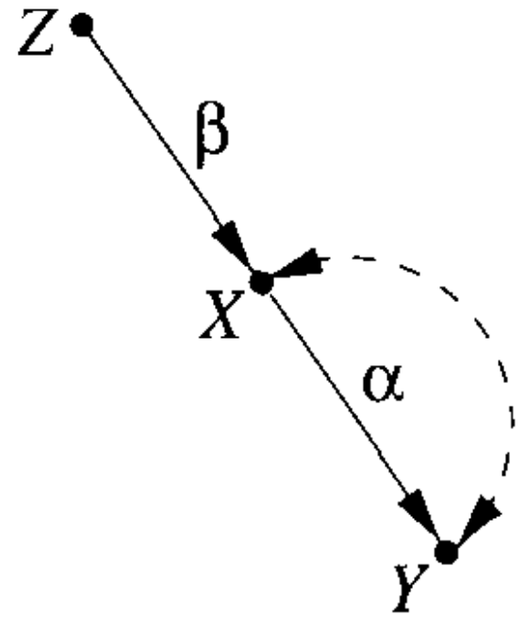


Figure 3.9 Typical models in which $P(y \mid \hat{x})$ is not identifiable.

Nonparametric vs. Parametric

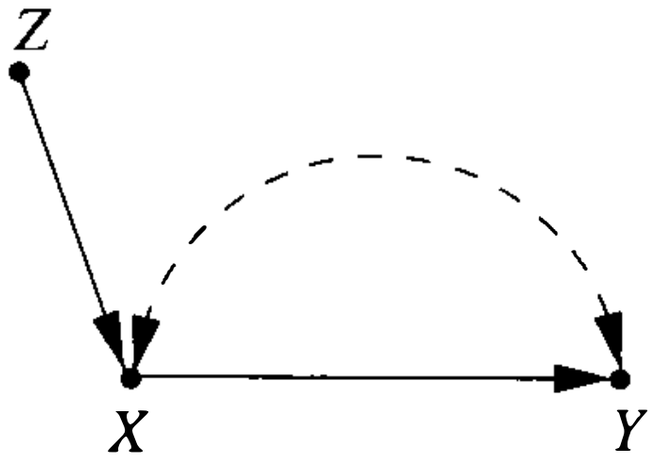


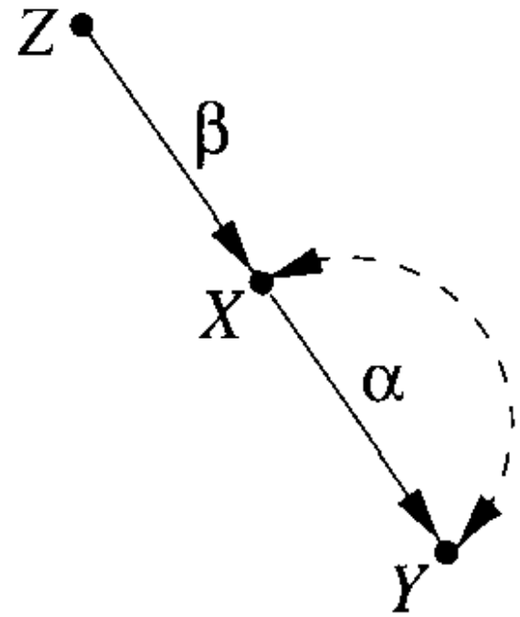


- What if the causal relations are linear?

α

Nonparametric vs. Parametric





- What if the causal relations are linear?

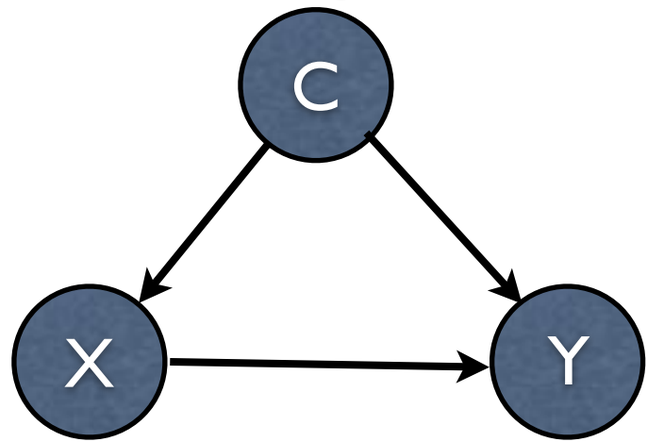
$\beta = r_{XZ}$ (regression coefficient of regressing X on Z)

$\alpha\beta = r_{YZ}$

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

Calculation of Causal Effects

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

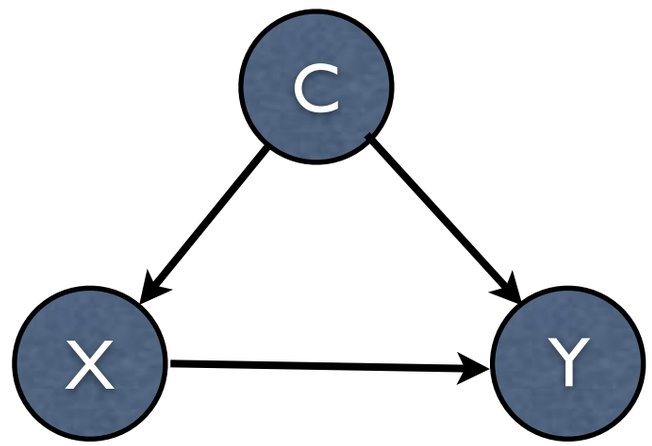


- ACE: $E[Y \mid do(x)] - E[Y \mid do(x')]$
  - x: the active treatment value; x': the baseline treatment value
  - The two groups do not necessary have the same $P(c)$

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

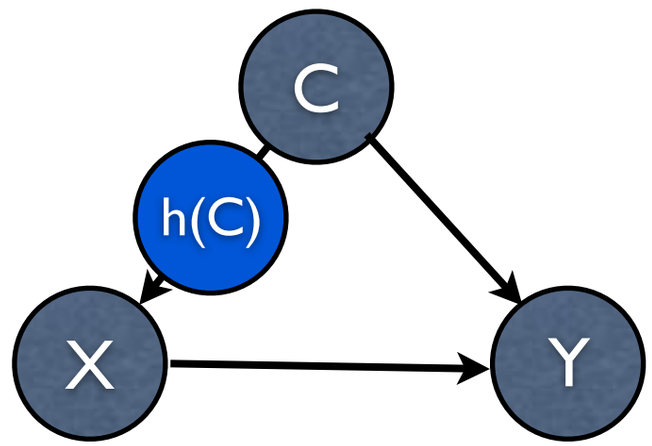
- One way is to match (usually high-dimensional) covariates C
- Alternatively, use the *propensity score*

*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:

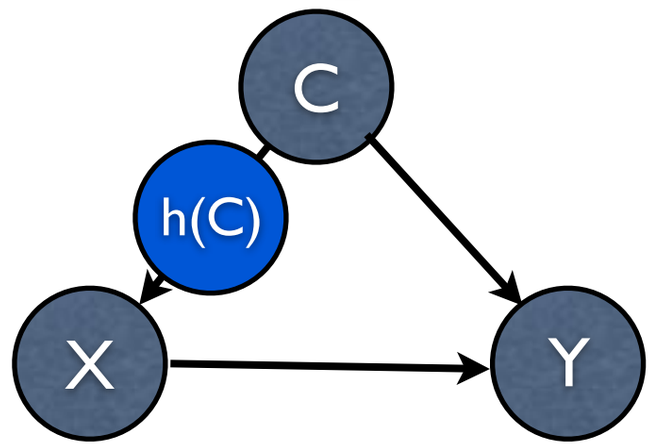
*Calculation of Causal Effects: Propensity Score



- 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

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

*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

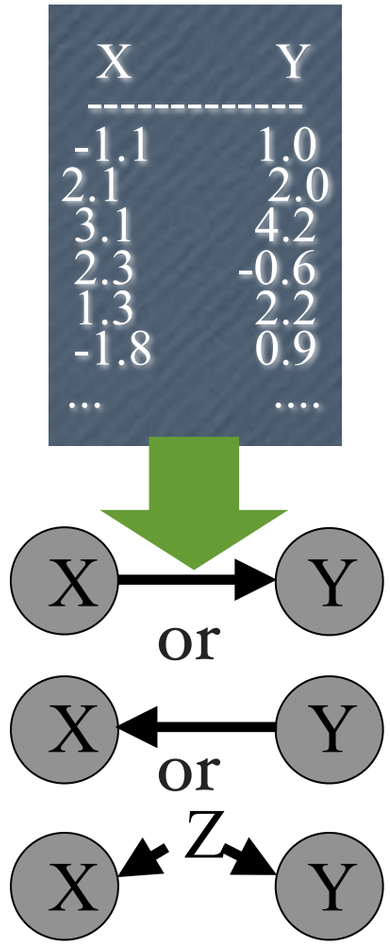
$$\sum_c P(Y|x,c)P(c) = \sum_c \sum_h P(Y|x,c)p(h)p(c|h)$$

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

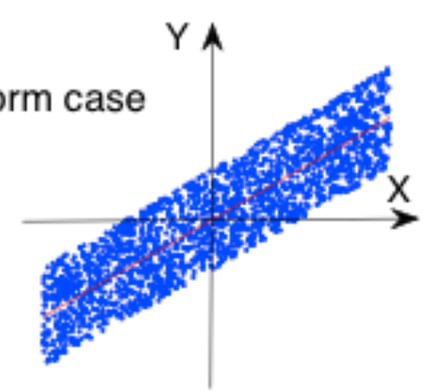
$$= \sum_h P(Y|x,h)P(h)$$

Advantage!?!

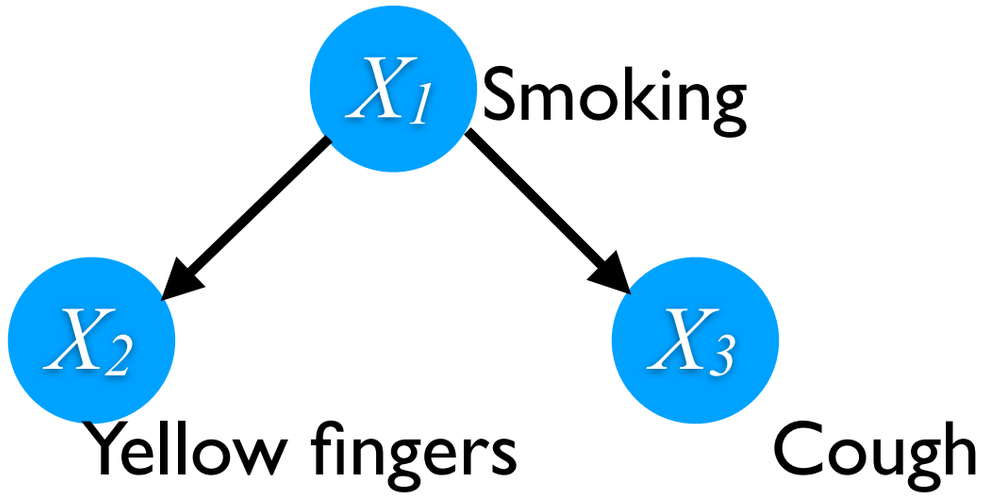
Outline



- Causality? Interventions? Causal thinking
- Causal graphical models
- Identification of causal effects
- **Counterfactual reasoning**
- Causal discovery
- Implications in machine learning



Three Types of Problems in current AI

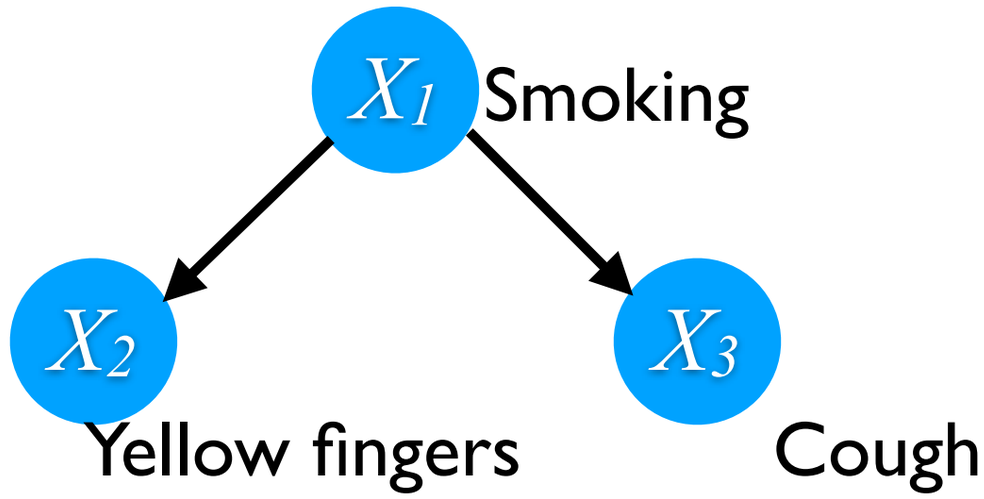


- Three questions:

X_1	X_2	X_3
1	0	0
0	0	1
0	1	1
1	1	1
0	0	0
0	1	0
1	1	1
1	1	1
0	0	0
1	0	0
...	...	...

- **Prediction**: Would the person cough if we *find* he/she has yellow fingers?
- **Intervention**: Would the person cough if we *make sure* that he/she has yellow fingers?
- ***Counterfactual***: Would George cough *had* he had yellow fingers, *given that he does not have yellow fingers and coughs*?

Three Types of Problems in current AI



- Three questions:

X_1	X_2	X_3
1	0	0
0	0	1
0	1	1
1	1	1
0	0	0
0	1	0
1	1	1
1	1	1
0	0	0
1	0	0
...	...	...

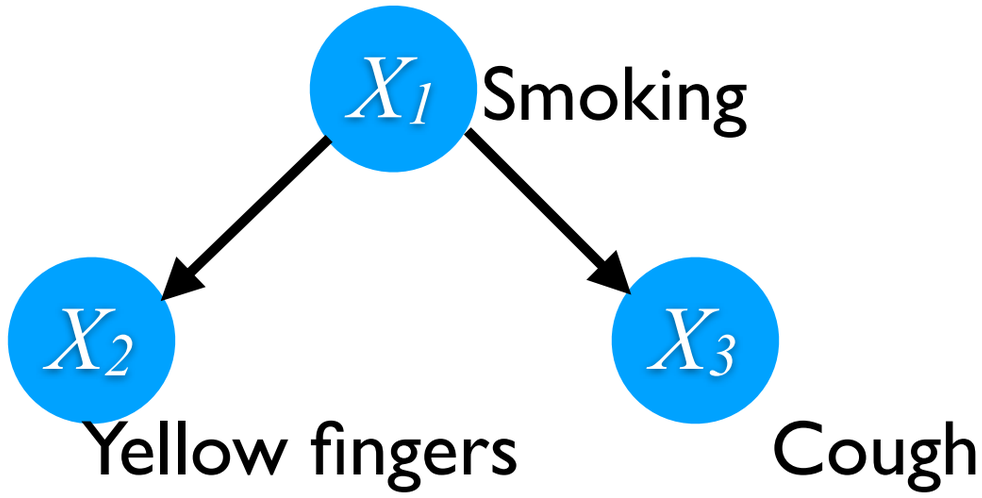
- **Prediction**: Would the person cough if we *find* he/she has yellow fingers?

$$P(X3 \mid X2=1)$$

- **Intervention**: Would the person cough if we *make sure* that he/she has yellow fingers?

- ***Counterfactual***: Would George cough *had* he had yellow fingers, *given that he does not have yellow fingers and coughs*?

Three Types of Problems in current AI



- Three questions:

X_1	X_2	X_3
1	0	0
0	0	1
0	1	1
1	1	1
0	0	0
0	1	0
1	1	1
1	1	1
0	0	0
1	0	0
...	...	...

- **Prediction**: Would the person cough if we *find* he/she has yellow fingers?

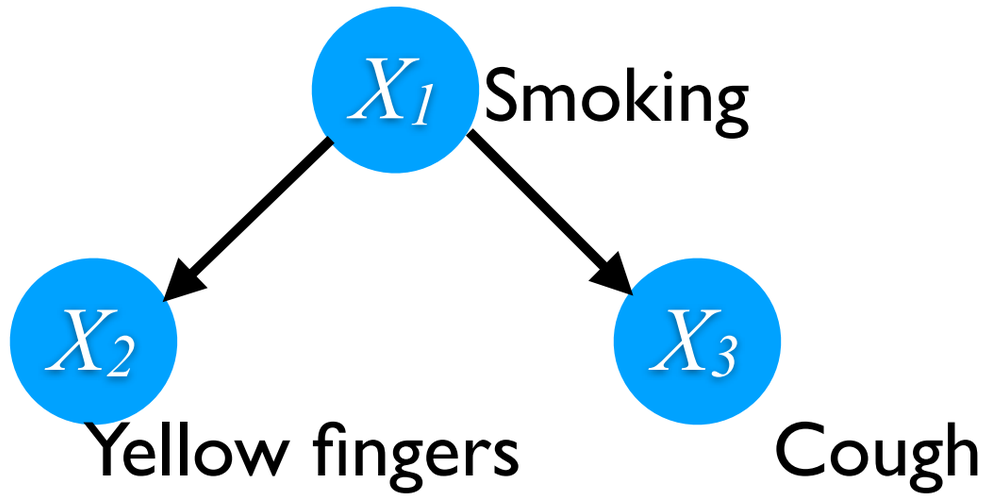
$$P(X3 \mid X2=1)$$

- **Intervention**: Would the person cough if we *make sure* that he/she has yellow fingers?

$$P(X3 \mid \text{do}\ (X2=1))$$

- ***Counterfactual***: Would George cough *had* he had yellow fingers, *given that he does not have yellow fingers and coughs*?

Three Types of Problems in current AI



- Three questions:

X_1	X_2	X_3
1	0	0
0	0	1
0	1	1
1	1	1
0	0	0
0	1	0
1	1	1
1	1	1
0	0	0
1	0	0
...	...	...

- **Prediction**: Would the person cough if we *find* he/she has yellow fingers?

$$P(X3 \mid X2=1)$$

- **Intervention**: Would the person cough if we *make sure* that he/she has yellow fingers?

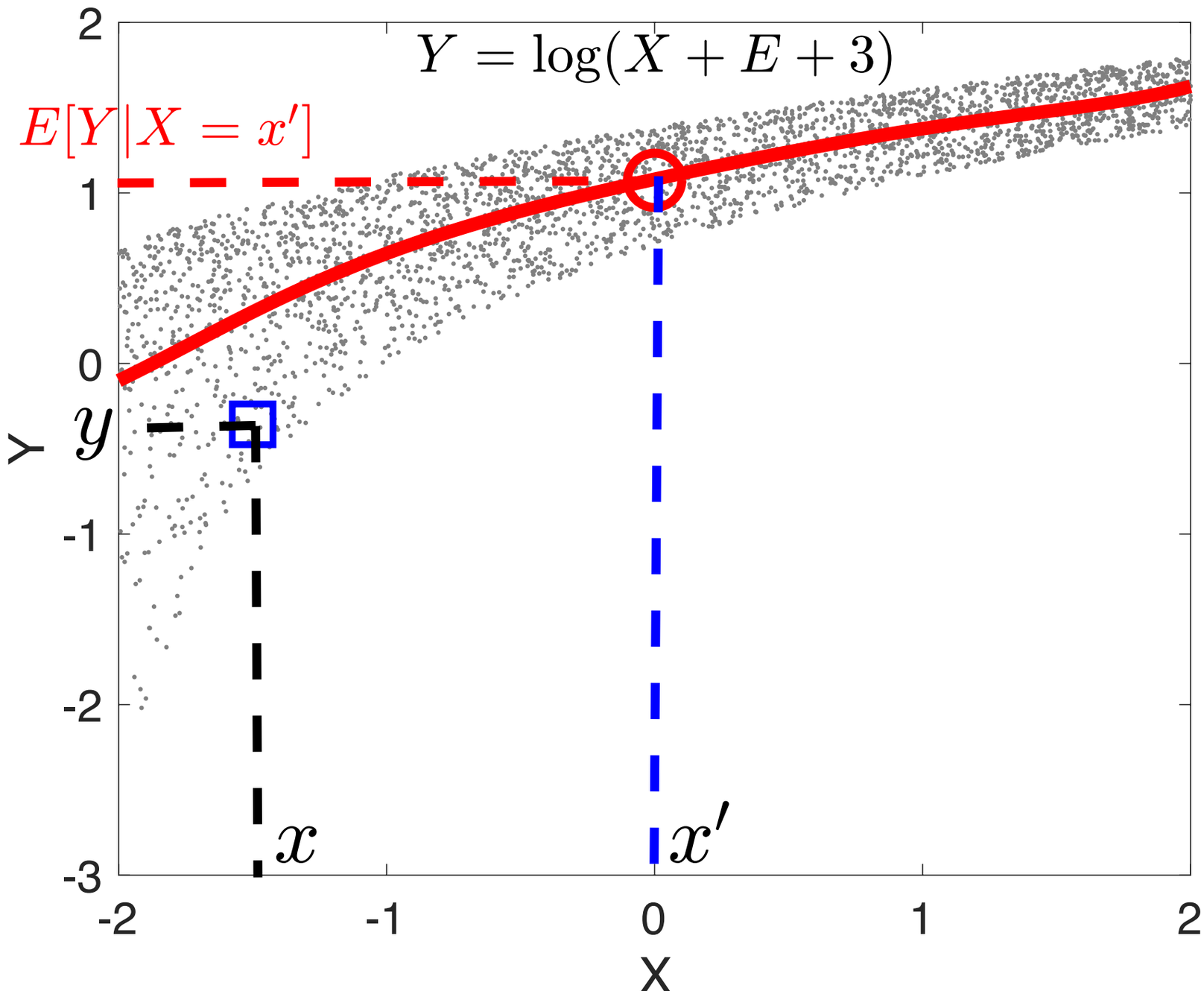
$$P(X3 \mid \text{do}\ (X2=1))$$

- ***Counterfactual***: Would George cough *had* he had yellow fingers, *given that he does not have yellow fingers and coughs*?

$$P(X3_{\,X2=1} \mid X2 = 0, X3 = 1)$$

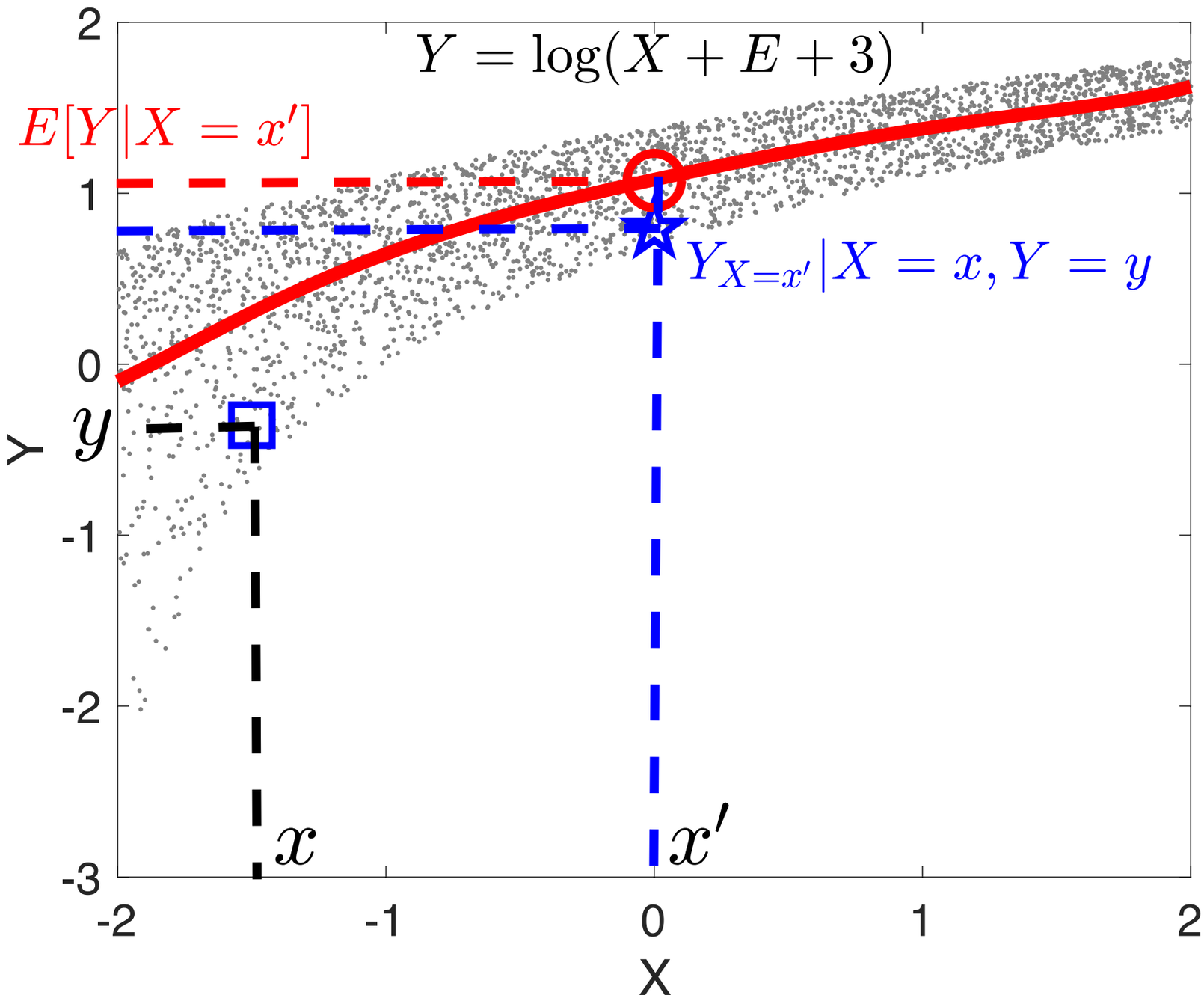
Counterfactual Inference vs. Prediction

- Suppose X→Y with Y = log(X + E + 3). For an individual with (*x*,*y*), what would Y be if X had been *x'* ?



Counterfactual Inference vs. Prediction

- Suppose X→Y with Y = log(X + E + 3). For an individual with (*x*,*y*), what would Y be if X had been *x'* ?



Standard Counterfactual Questions

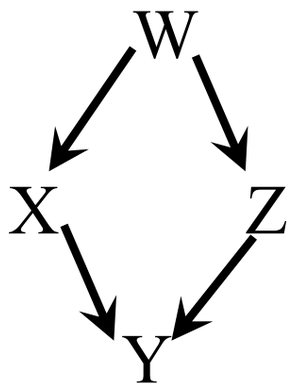
- We talk about a particular situation (or unit) U = *u*, in which X = *x* and Y = *y*

- What value would Y be had X been *x'* in situation *u*? I.e., we want to know $Y_{X=x'}(u)$, the value of Y in situation *u* if we do(X=*x'*)

- *u* is not directly observable, so $P(Y_{X=x'} \mid X = x,\ Y = y)$ instead

For identification of causal effects, U is randomized. It is fixed for counterfactual inference.

Counterfactual Inference



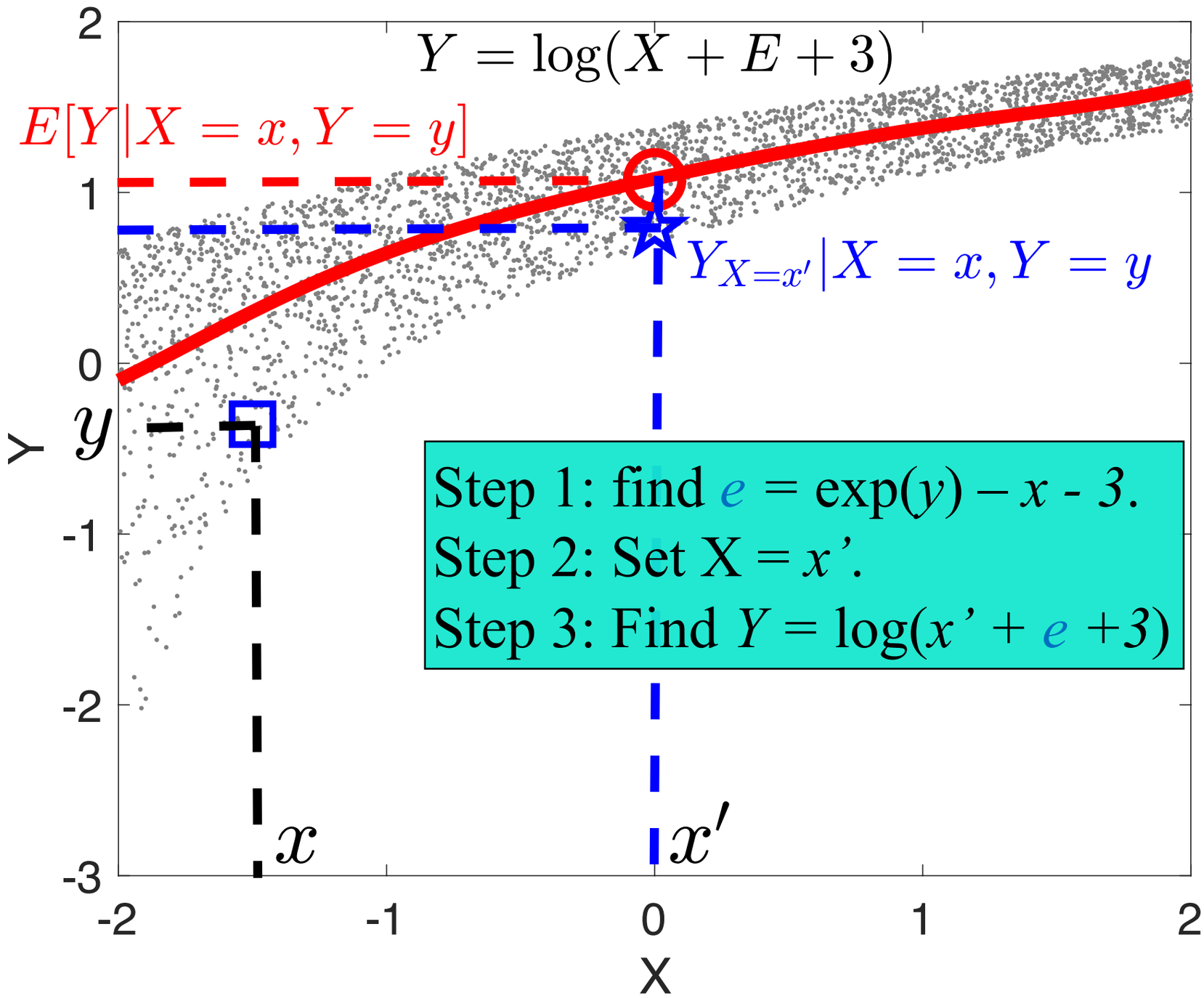
$$W = U_W$$
$$X = f_X(W, U_X)$$
$$Z = f_Z(W, U_Z)$$
$$Y = f_Y(X, Z, U_Z)$$

$$P(Y_{X=x'} \mid \underbrace{X = x,\ Y = y,\ W = w}_{\textit{evidence}})$$

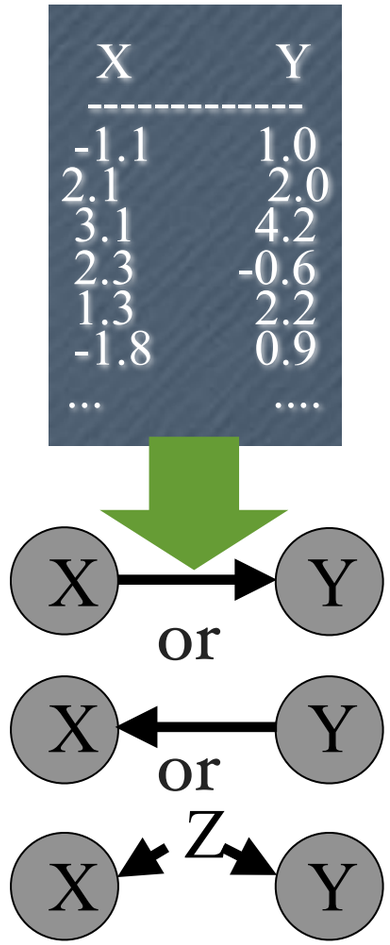
- Three steps
  - Abduction: find P(U | *evidence*)
  - Action: Replace the equation for X by X = *x'*
  - Prediction: Use the modified model to predict Y

Counterfactual Inference vs. Prediction

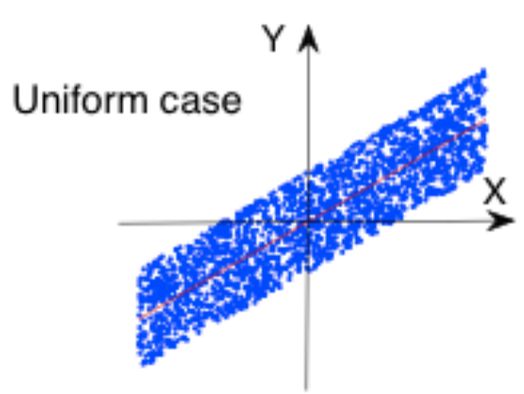
- Suppose X→Y with Y = log(X + E + 3). For an individual with (*x*,*y*), what would Y be if X had been *x'* ?



Next Class…



- Causality? Interventions? Causal thinking
- Causal graphical models
- Identification of causal effects



- Counterfactual reasoning
- **Causal discovery**
- **Implications in machine learning**