THE UNIVERSITY OF SYDNEY

Advanced Machine Learning

(COMP 5328)

Week 12 Tutorial: Causal Inference

Anjin Liu
anjin.liu@sydney.edu.au

Tutorial Contents

THE UNIVERSITY OF SYDNEY

- Review (20min):
  - Lecture 10: Causal Inference
- Tutorial exercise & QA (40min):

Key points

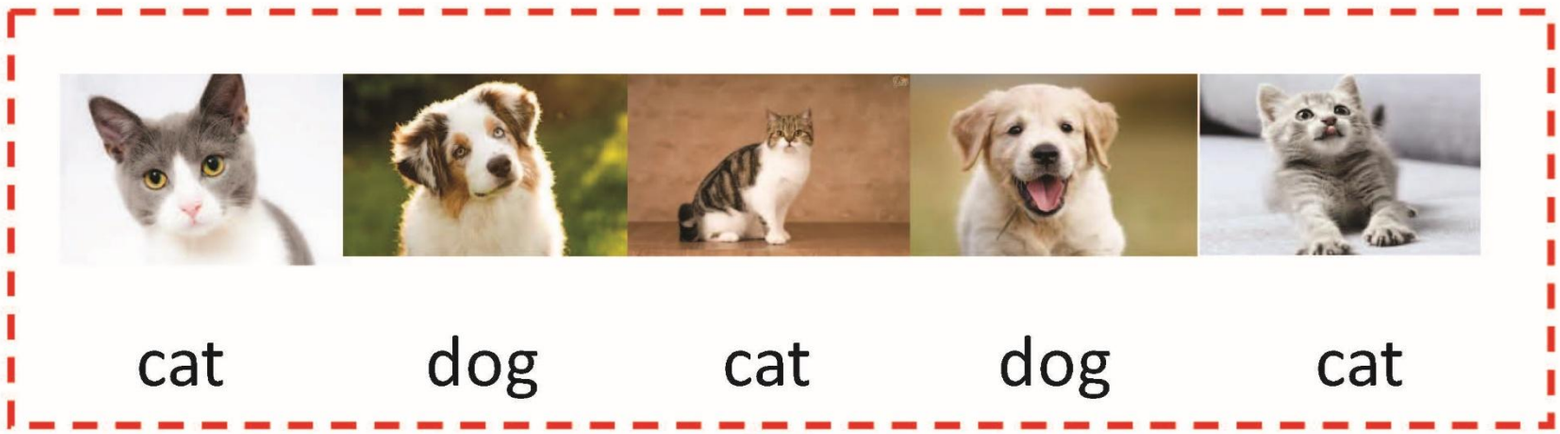
- Dependence
- Causal Thinking
- Causal Representation

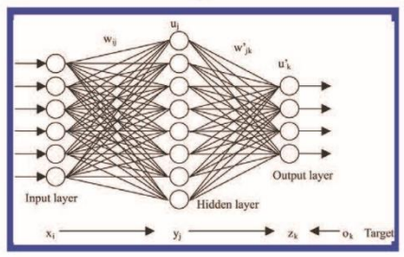
THE UNIVERSITY OF SYDNEY

Dependence

THE UNIVERSITY OF SYDNEY

Machine learning systems are usually driven by statistical dependence. But not causality





dog

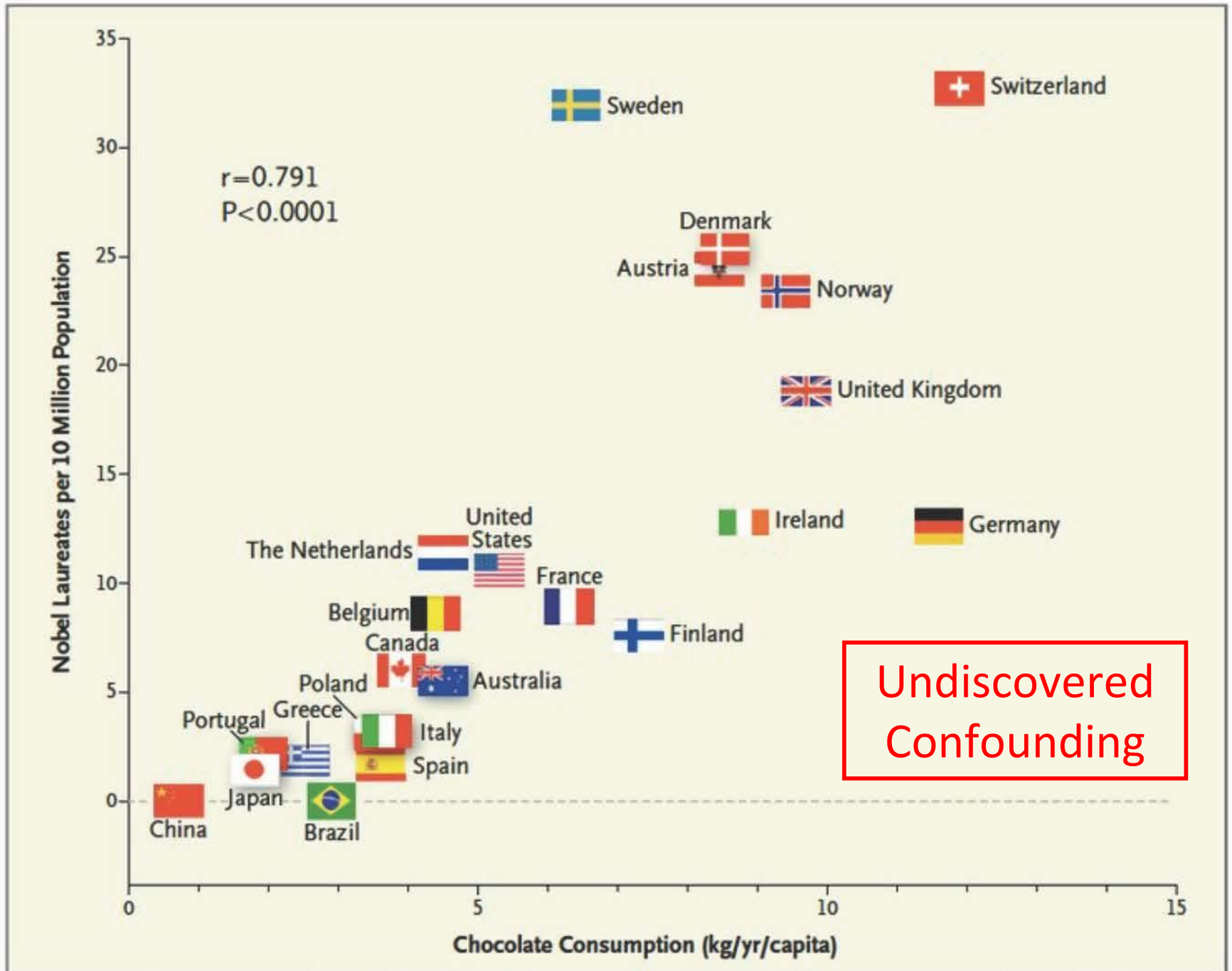


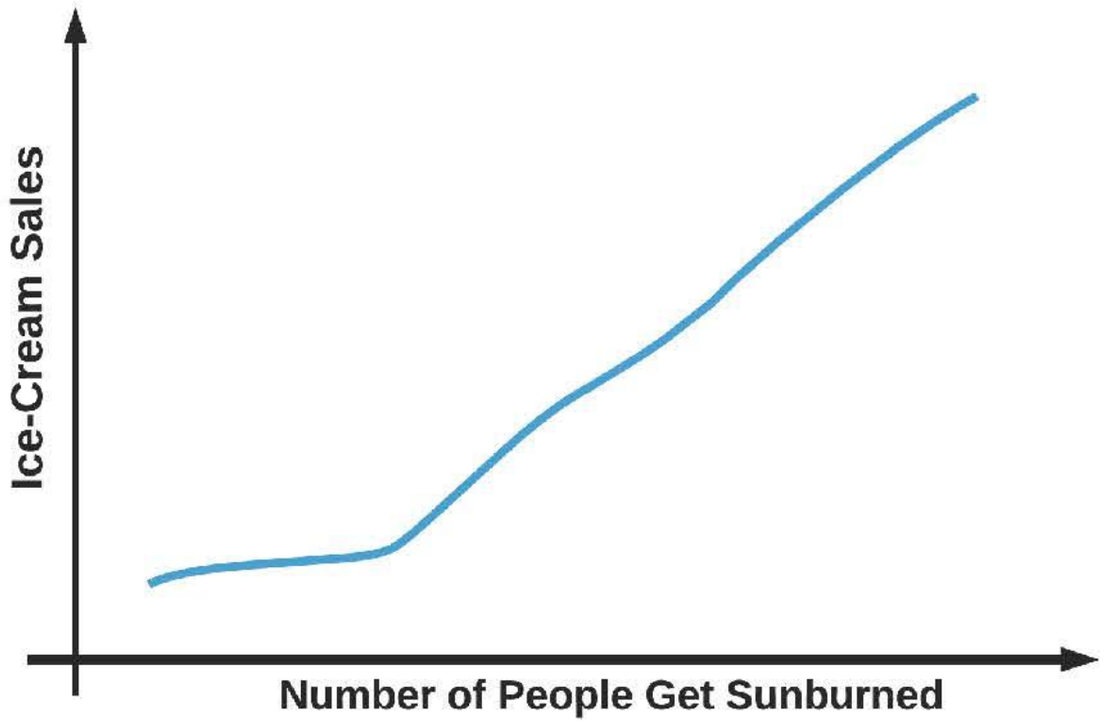
Figure 1. **Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.**

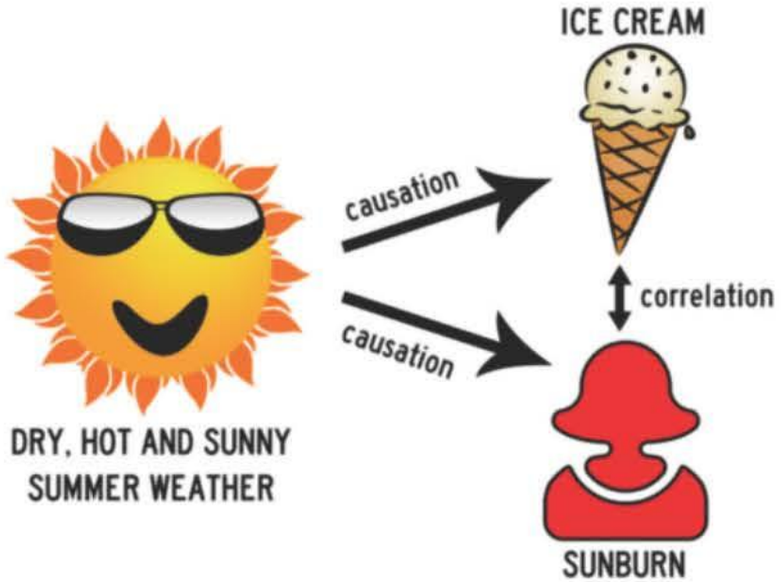
Messerli, Franz H. "Chocolate Consumption, Cognitive Function, and Nobel Laureates." (2012).

THE UNIVERSITY OF SYDNEY

Causal thinking

"Strange" Dependence

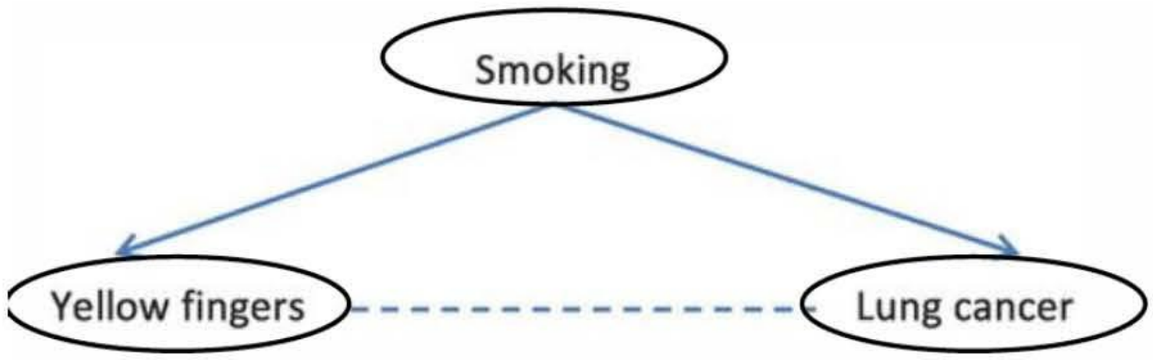




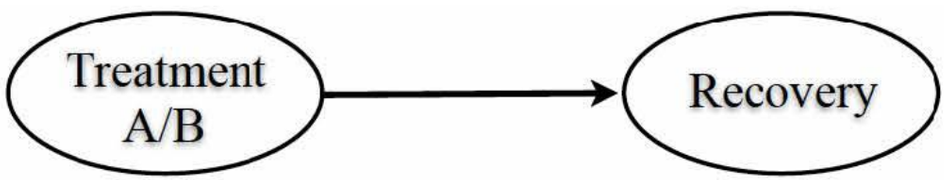
Now, a task is to predict the likelihood of a person got sunburn. Do you think the number of Ice-cream consumed would be a good feature?

Ways to Produce Dependence

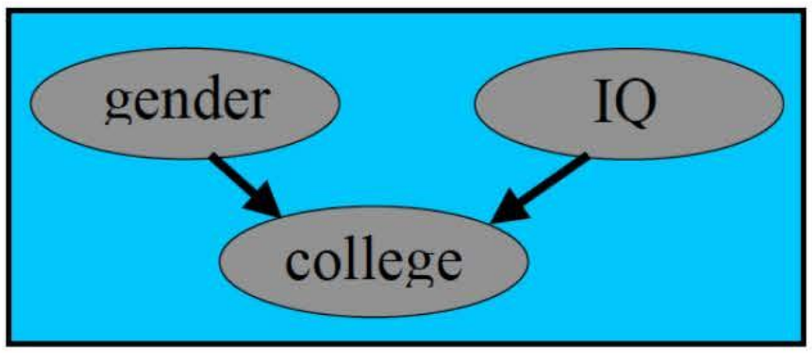
THE UNIVERSITY OF SYDNEY



Common Cause



Causal relation



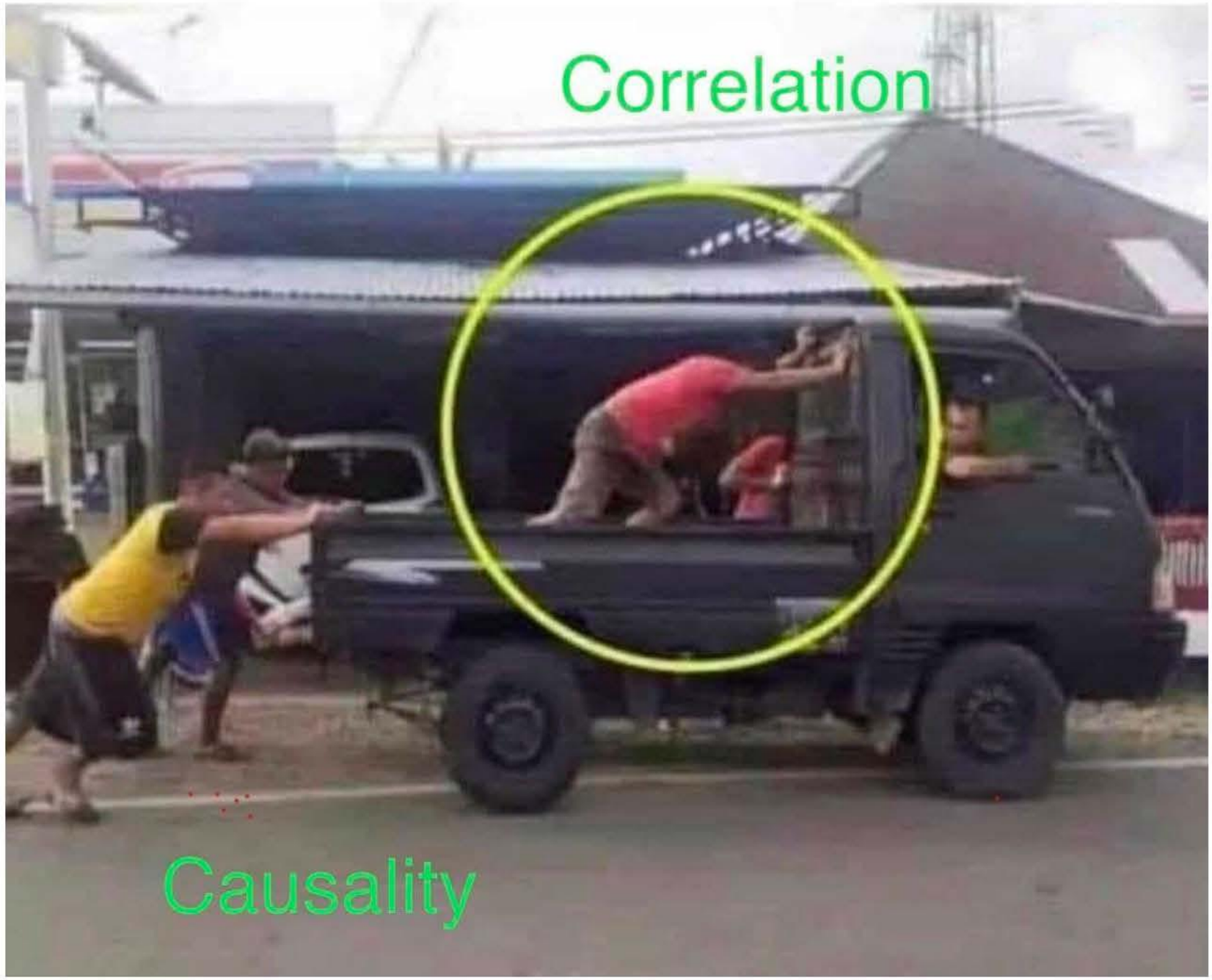
Conditional dependence given common effect

THE UNIVERSITY OF SYDNEY

Concept	Formal Criterion	What It Captures	Detectable from Data Alone?	Example
Statistical Dependence	$P(X, Y) \neq P(X)P(Y)$	Any relationship (linear or nonlinear)	✅ Yes	Rain ↔ Umbrella sales
Correlation	$\rho_{XY} = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}$	Linear dependence only	✅ Yes	Height ↔ Weight
Causation	$P(Y \mid do(X)) \neq P(Y)$	Direct causal influence	❌ No, requires intervention or causal model	Smoking → Lung cancer

- **Dependence** is the broadest concept: any statistical relationship.
- **Correlation** is a *numeric measure* of *linear* dependence.
- **Causation** implies a *mechanism* — not just co-occurrence, but *one variable producing change in another*.

Causation vs Dependence



Why casual?

THE UNIVERSITY OF SYDNEY

Objective:

- Discover causal relationships

  Identify the *direction* of influence between variables (who causes whom).
- Estimate the strength of causal effects

  Quantify *how much* one variable changes when another changes
- **Enable causal interventions and counterfactual reasoning**

  Predict outcomes under *interventions and* reason about ***what-if* scenarios**.

Basic Assumptions Causal Bayesian Net:

- Markov Condition: Once you know the ***direct causes*** of something, **everything else** in the past **becomes irrelevant** to it.
- Faithfulness Condition: If two things ***look independent* in the data**, it's **because the graph truly says they're independent** — not because of a mathematical coincidence.

Concept	Simple Intuition	Everyday Meaning	What It Guarantees
Markov Condition	"Once you know the direct causes, the rest doesn't matter."	The world works locally — each thing depends only on its direct causes.	Allows us to factorize and predict using local structure.
Faithfulness Condition	"Observed independencies reflect the real causal structure."	There are no lucky coincidences that make connected things look independent.	Lets us learn the causal graph from data.

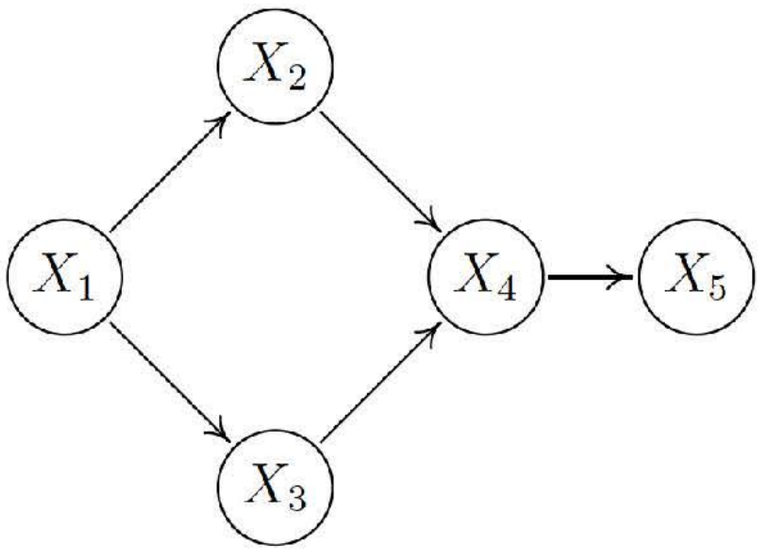
THE UNIVERSITY OF SYDNEY

Causal representation

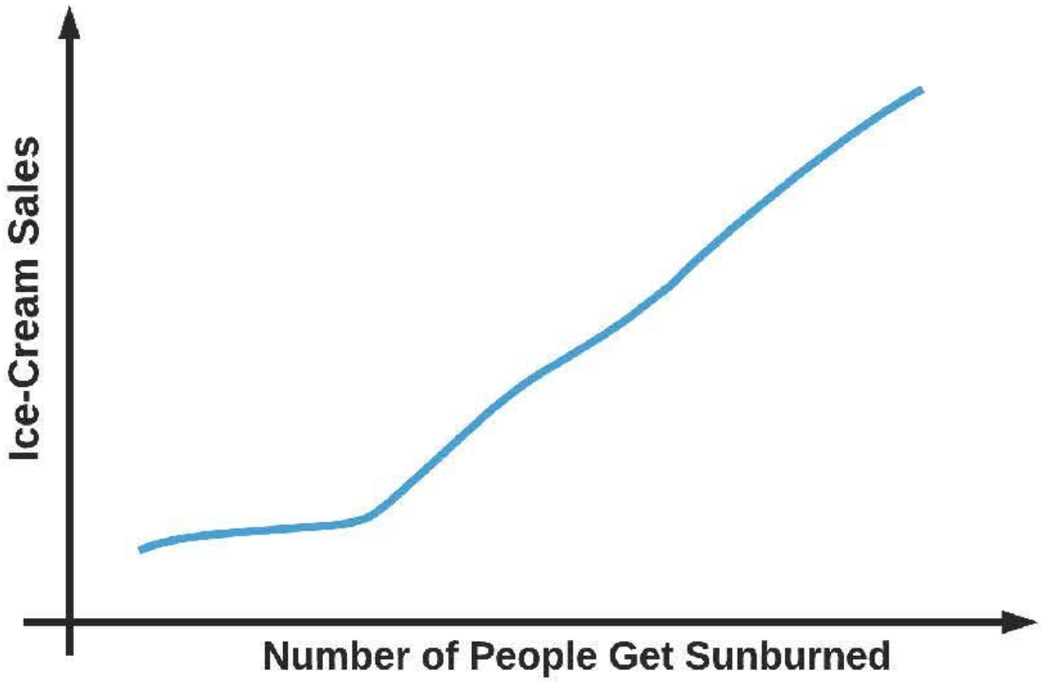
Directed Acyclic Graph (DAG)

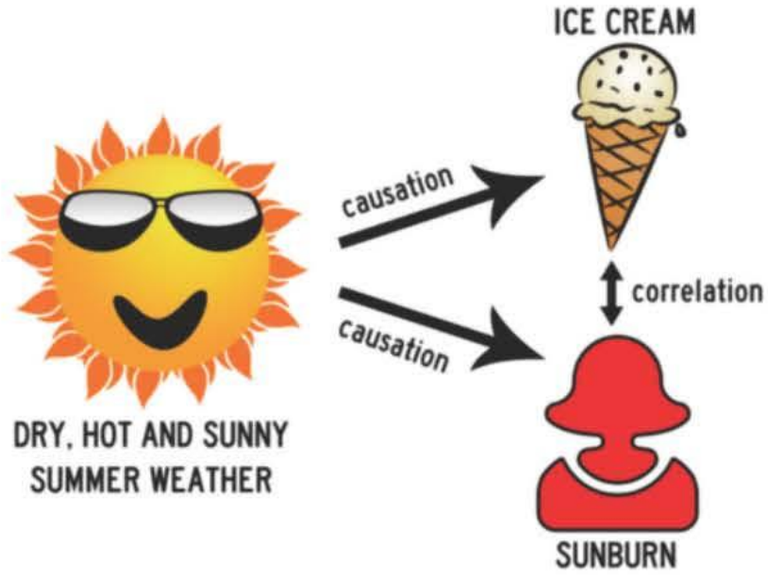
THE UNIVERSITY OF SYDNEY

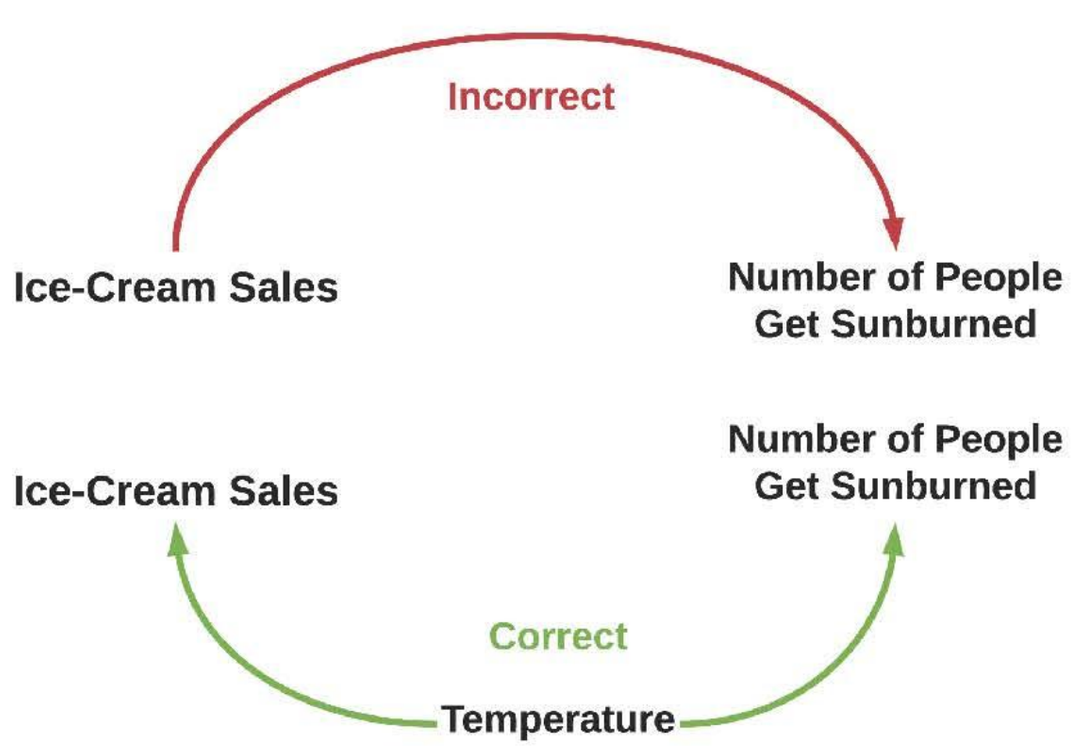
- **DAG**
  - graph with directed edges, no cycles
- **Path**
  - a sequence of connected vertices
  - X1 - X2 - X4 - X5
  - X1 - X2 - X4 - X3



"Strange" Dependence







Markov Conditions

THE UNIVERSITY OF SYDNEY

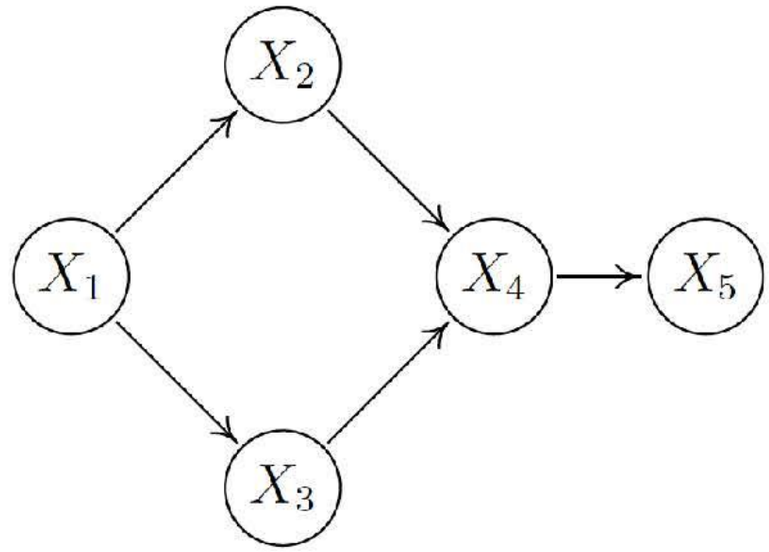
- Markov conditions state that if a certain graph property holds true, then a certain statistically independence holds true.

- There are local Markov condition and global Markov condition.

Local Markov Condition

THE UNIVERSITY OF SYDNEY

Every variable X_i, in a directed acyclic graph, is independent of its non-descendant Y conditional on its parents, i.e., $(X_i \perp\!\!\!\perp Y \mid PA_i)_G$, which also implies $(X_i \perp\!\!\!\perp Y \mid PA_i)_p$.



For example:
$(X_4 \perp\!\!\!\perp X_1 \mid \{X_2, X_3\})_G \implies (X_4 \perp\!\!\!\perp X_1 \mid \{X_2, X_3\})_p$.

By the local Markov condition,
we could obtain a casual factorisation of the joint distribution as follows

$$P(X_1, ..., X_5) = P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_1)P(X_4 \mid X_2, X_3)P(X_5 \mid X_4)$$

Global Markov Conditions (D-Separation)

For three disjoint sets of variables $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{S}$, $\mathbf{X}$ is d-separated from $\mathbf{Y}$ conditional on $\mathbf{S}$ if and only if all paths between any member of $\mathbf{X}$ and any member of $\mathbf{Y}$ are *blocked* by $\mathbf{S}$.

A path q is said to be blocked by the set $\mathbf{S}$ if

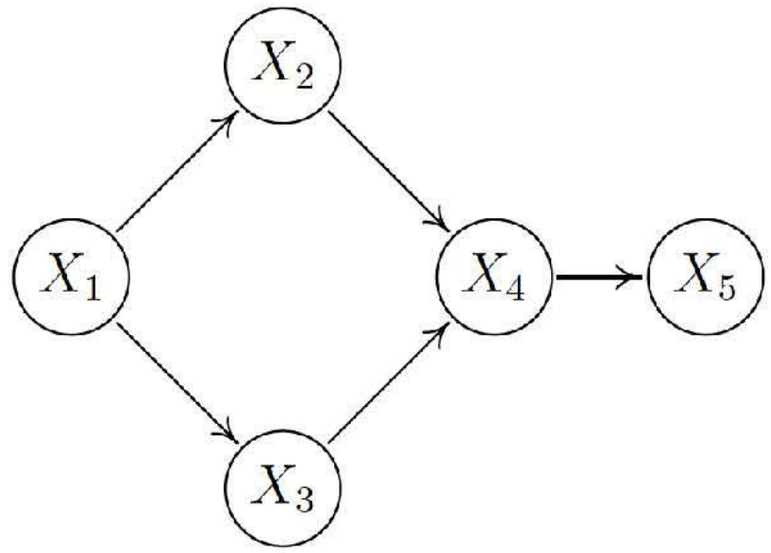
- q contains a chain $i \to m \to j$ or a fork $i \leftarrow m \to j$ such that the middle node m is in $\mathbf{S}$, or
- q contains a collider $i \to m \leftarrow j$ such that the middle node m is not in $\mathbf{S}$, and no descendant of m is in $\mathbf{S}$.

Formally, we use $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \,|\, \mathbf{S})_G$ to denote that $\mathbf{S}$ d-separates $\mathbf{X}$ and $\mathbf{Y}$ in the DAG G, which also implies $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \,|\, \mathbf{S})_p$.

Global Markov Conditions (D-Separation)

THE UNIVERSITY OF SYDNEY

- Let $\mathbf{X} = \{X_1\}, \mathbf{Y} = \{X_4, X_5\}$, then $\mathbf{X}$ is d-separated from $\mathbf{Y}$ conditional on $\mathbf{S} = \{X_2, X_3\}$ (Because all paths from $\mathbf{X}$ to $\mathbf{Y}$ are blocked by conditioning on $\mathbf{S}$).



- Let $\mathbf{X} = \{X_2\}, \mathbf{Y} = \{X_3\}$, then $\mathbf{X}$ is not d-separated from $\mathbf{Y}$ conditional on $\mathbf{S} = \{X_1, X_4\}$.

- Let $\mathbf{X} = \{X_2\}, \mathbf{Y} = \{X_3\}$, then $\mathbf{X}$ is d-separated from $\mathbf{Y}$ conditional on $\mathbf{S} = \{X_1\}$.

Casual faithfulness Assumption

THE UNIVERSITY OF SYDNEY

The probability distribution may have additional conditional independence relations that are not entailed by d-separation applied to a graph. When no such extra conditional independence relations hold the distribution is said to be faithful to the graph, i.e.,

$$(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \,|\, \mathbf{S})_p \implies (\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \,|\, \mathbf{S})_G,$$

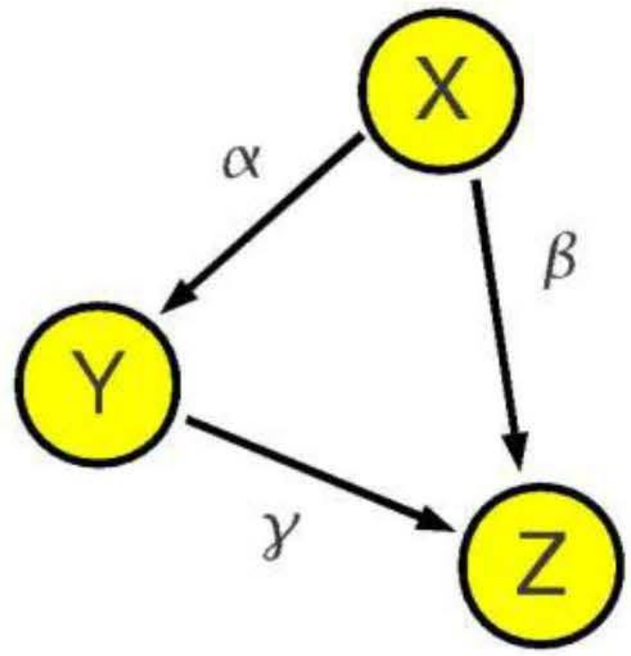
then the distribution p is said to be faithful to the graph.

Casual faithfulness Assumption

THE UNIVERSITY OF SYDNEY

Here is an example of a unfaithful distribution

$$X = U_x,$$
$$Y = \alpha X + U_y,$$
$$Z = \beta X + \gamma Y + U_z.$$

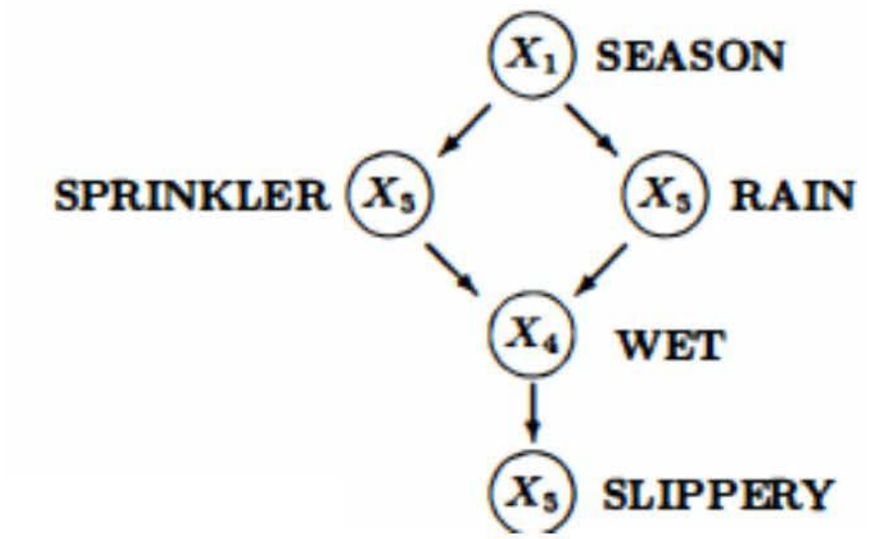


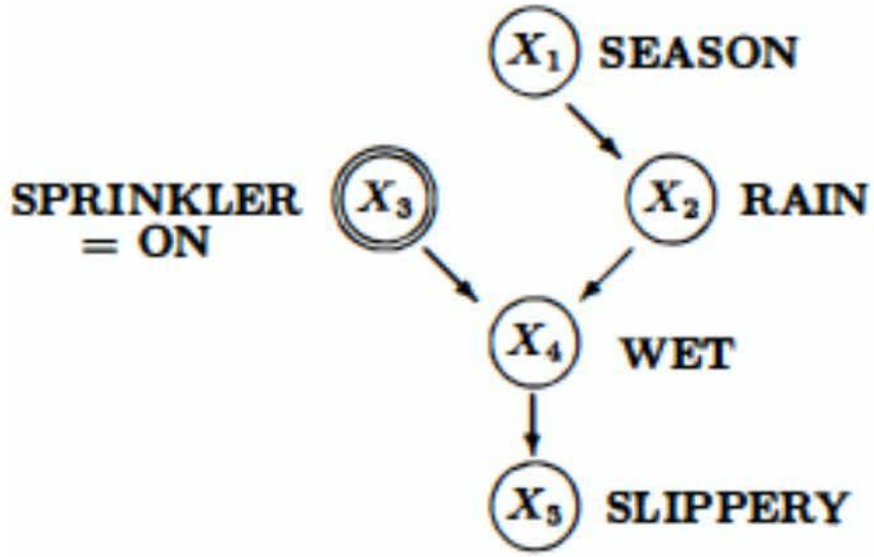
From the graph $(\{X\} \not\perp\!\!\!\perp \{Z\})_G$, however, if $\beta + \alpha\gamma = 0$, then $(\{X\} \perp\!\!\!\perp \{Z\})_p$, and $(\{X\} \perp\!\!\!\perp \{Z\})_p$ does not imply $(\{X\} \perp\!\!\!\perp \{Z\})_G$.

Causal Bayesian Net

THE UNIVERSITY OF SYDNEY

- **Causal Bayesian Network**
  - directed edges representation causal direction, causal DAG
  - more meaningful and represent external changes





Conditioning, Intervention, Counterfactual

THE UNIVERSITY OF SYDNEY

- **Prediction/Conditioning**
  - would the pavement be slippery if we find the sprinkler off?

$$P(slippery \mid Sprinkler = off)$$

- **Intervention**
  - would the pavement be slippery if we turn off the sprinkler?

$$P(slippery \mid do(Sprinkler = off))$$

- **Prediction/Counterfactual reasoning**
  - would the pavement be slippery, had the sprinkler been off, given that the pavement is in fact not slippery and the sprinkler is on?

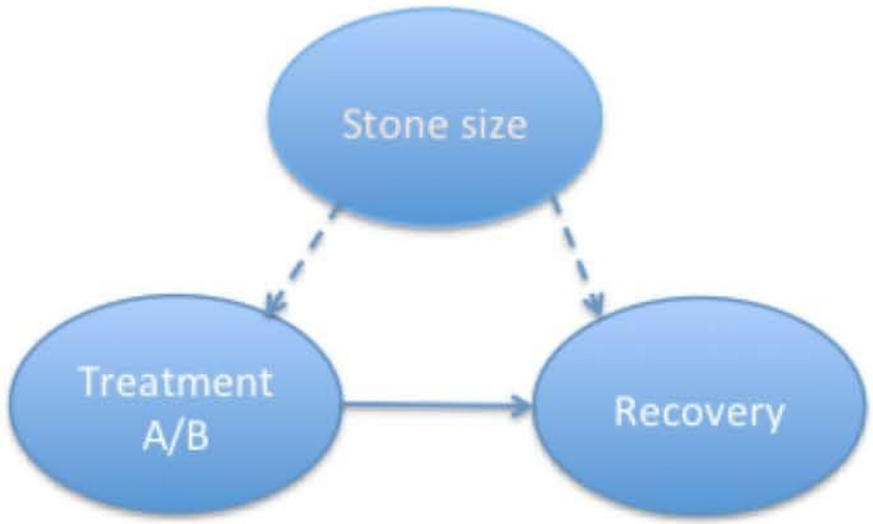
$$P(slippery_{sprinkler=off} \mid Sprinkler = on, Slippery = no)$$

Identification of Causal Effects

THE UNIVERSITY OF SYDNEY

$$P(\text{Recovery} \mid do(Treatment = A))$$

- **"Golden standard": randomised 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



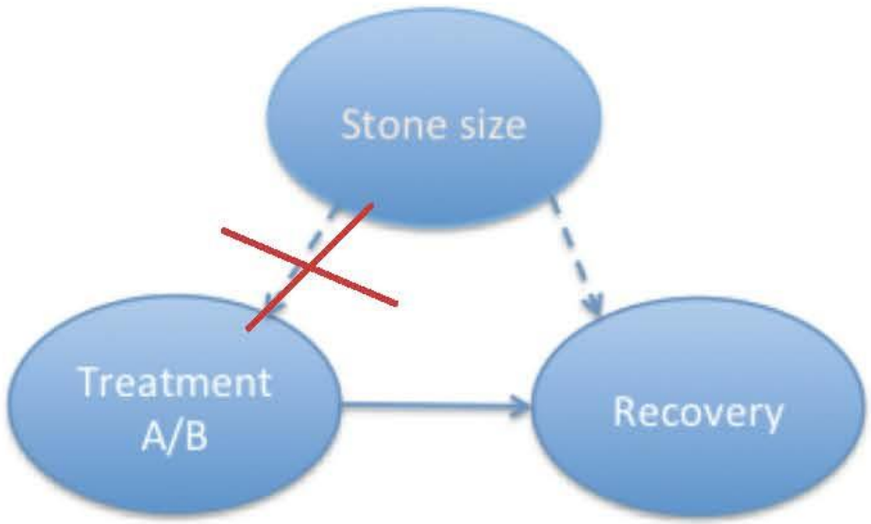
- Usually expensive or infeasible to do!

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)**

T - Treatment
R - Recovery

$$P(\mathrm{R} \mid T) = \sum_S P(R \mid T, S) P(S \mid T)$$



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

Intervention vs. conditioning

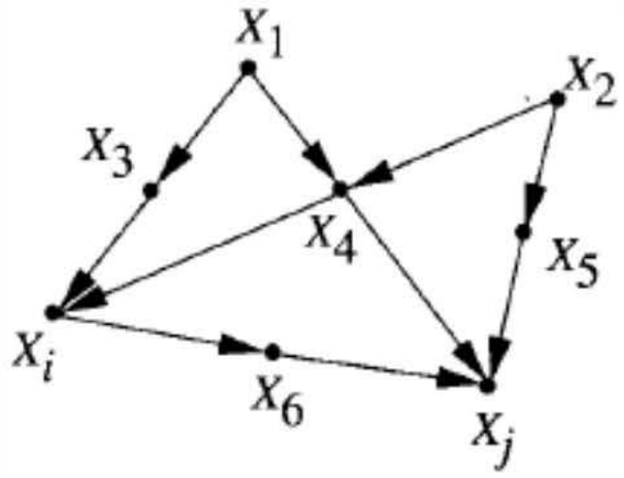
Back-Door Criterion

THE UNIVERSITY OF SYDNEY

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*

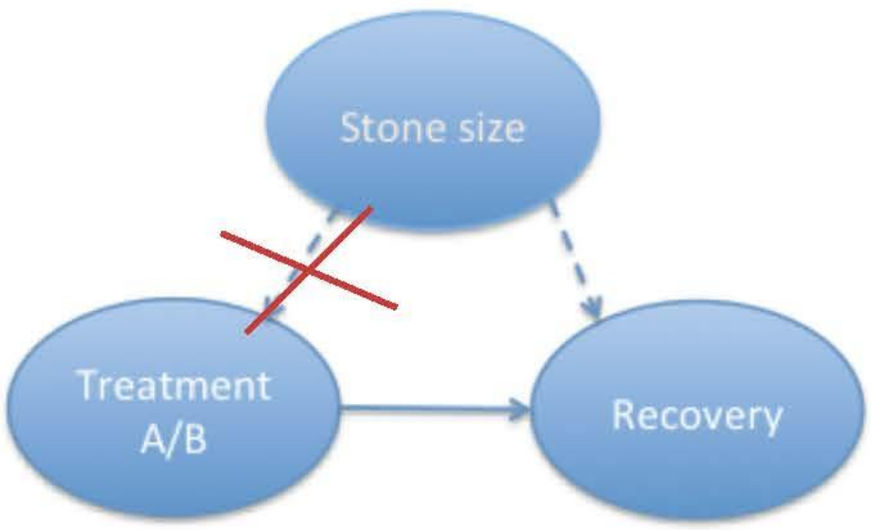
$$P(y \mid \hat{x}) = \sum_z P(y \mid x, z) P(z).$$

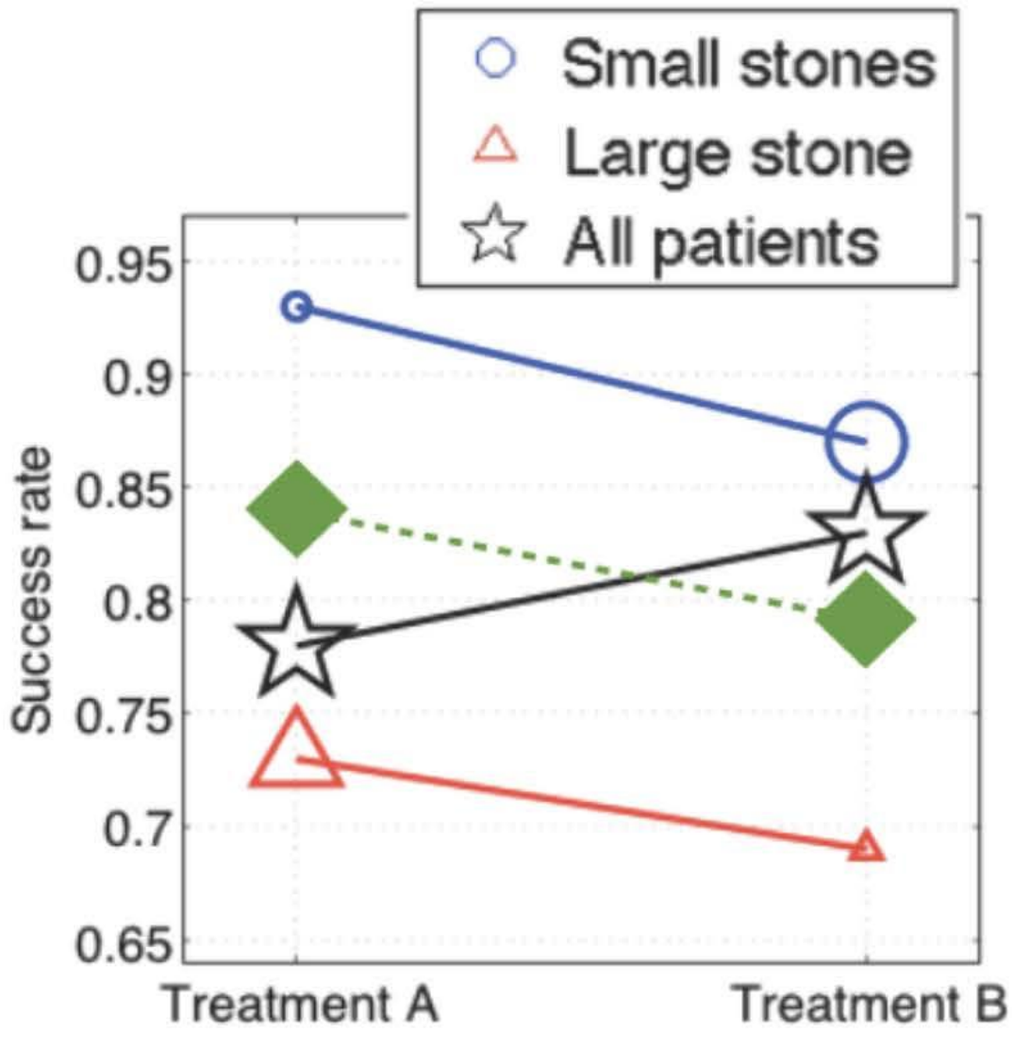
$$P(y \mid \hat{x}) = \sum_z P(y \mid x, z) P(z).$$

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)**

T - Treatment
R - Recovery





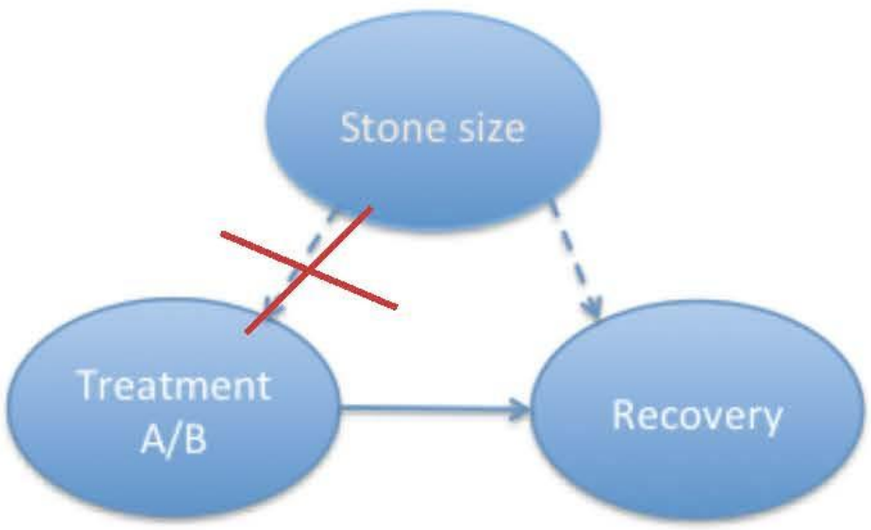
$$P(y \mid \hat{x}) = \sum_z P(y \mid x, z) P(z).$$

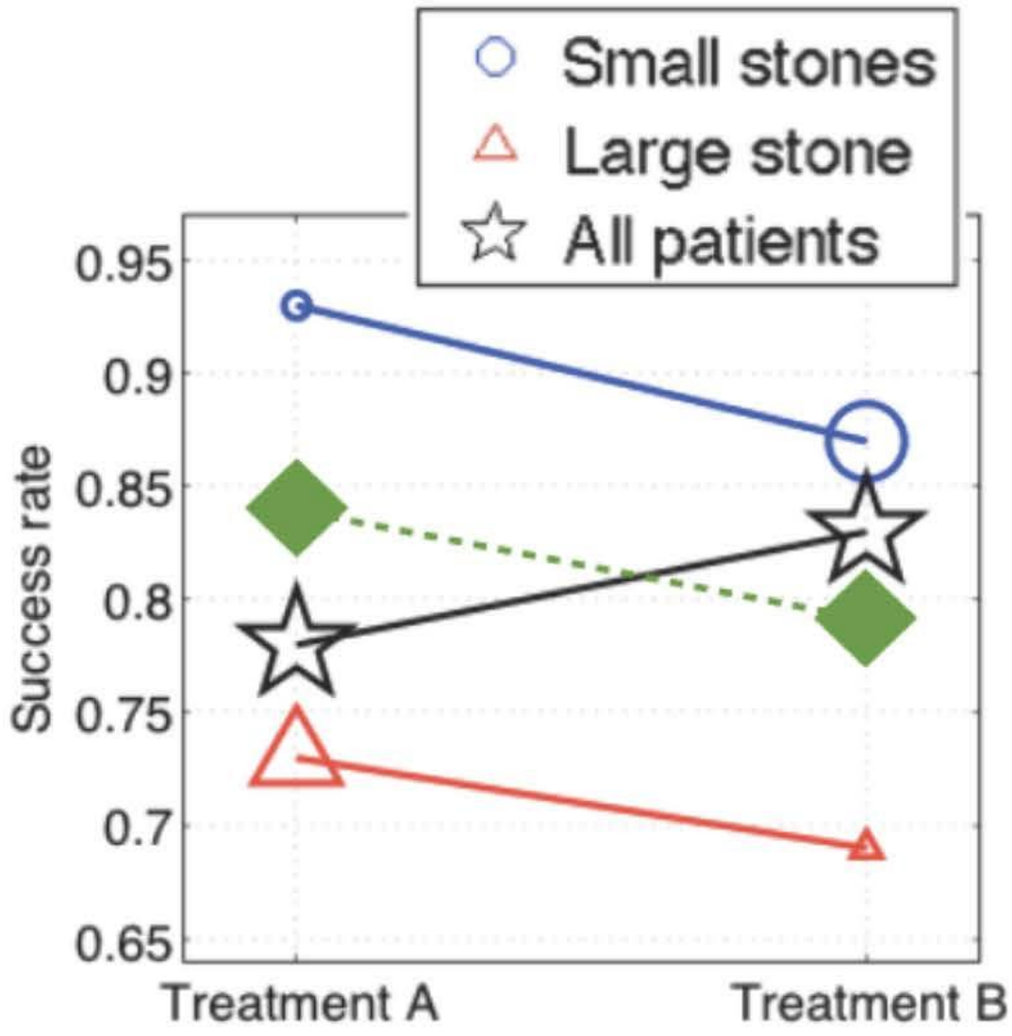
(87 + 270) / (350 + 350) x 93% + (263 + 80) / (350 + 350) x 73% = 83.2%

Example

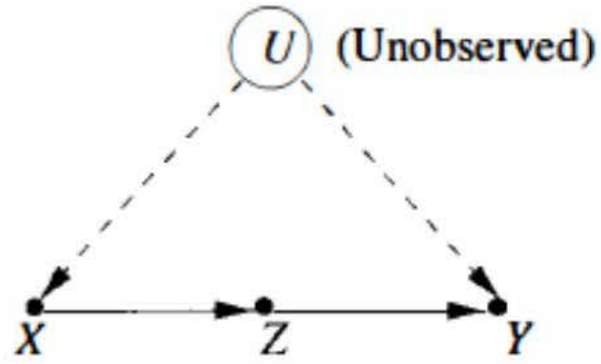
	Treatment A	Treatment B
Small Stones	*Group 1* **93% (81/87)**	*Group 2* 87% (234/270)
Large Stones	*Group 3* **73% (192/263)**	*Group 4* 69% (55/80)
Both	78% (273/350)	**83% (289/350)**

T - Treatment
R - Recovery





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;*

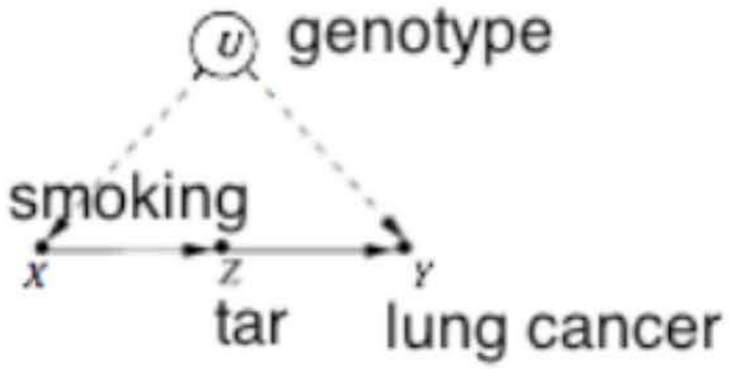
(ii) *there is no back-door path from X to Z; and*

(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



	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

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

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

Summary

- Dependence does not equal causality!!!
- Causality is essential in interventional studies.
- Given a causal graph, causal effects can be estimated from observational data.
- Causal graph can be learned from observational data.

Key points

- Dependence
- Causal Thinking
- Causal Representation