



THE UNIVERSITY OF
SYDNEY

Advanced Machine Learning

(COMP 5328)

Causal Inference

Tongliang Liu



THE UNIVERSITY OF
SYDNEY

Review

RL Framework (Procedure)



THE UNIVERSITY OF
SYDNEY

An agent interacts with its environment according to following procedure.

- At each time step t , the agent is in some state $S_t = s_t$.
- It chooses an action $A_t = a_t$.
- Based on s_t and a_t , it receives reward $R_t = r_t$ and goes into state $S_{t+1} = s_{t+1}$.

In general, the action A_t , reward R_t and next state S_{t+1} have the following probability distributions,

$$\pi(A_t|S_t = s_t), \quad P(S_{t+1}|A_t = a_t, S_t = s_t), \quad P(R_t|A_t = a_t, S_t = s_t).$$

RL Framework (Objective)



THE UNIVERSITY OF
SYDNEY

Given current state is s_0 , we want to find the optimal policy (best strategy) π^* that maximises the expected cumulative reward, i.e.,

$$\pi^* = \arg \max_{\pi} \mathbb{E} \left[\sum_{t \geq 0} \gamma^t r_t | S_0 = s_0, \pi \right],$$

With

$$a_t \sim \pi(A_t | S_t = s_t), \quad s_{t+1} \sim P(S_{t+1} | A_t = a_t, S_t = s_t),$$

and a discounted factor

$$\gamma \in (0, 1)$$

Q-value function

Q-value function $Q : S \times A \rightarrow \mathbb{R}$ determines the expected cumulative reward by following the policy π given current state is s_0 , and action is a_0 , specifically,

$$\underbrace{Q(s_0, a_0)}_{\text{ }} = \mathbb{E} \left[\sum_{t \geq 0} \gamma^t r_t | S_0 = s_0, A_0 = a_0, \pi \right],$$

with

$$a_t \sim \pi(A_t | S_t = s_t), \quad s_{t+1} \sim P(S_{t+1} | A_t = a_t, S_t = s_t),$$

$$r_t \sim P(R_t | A_t = a_t, S_t = s_t)$$

and γ is the discounted factor.



Optimal Q-value function

The optimal Q-value (denoted as Q^*) is the maximum expected cumulative reward achievable by following the optimal policy π^* given that the current state is s_0 and action is a_0 , specifically,

$$Q^*(s_0, a_0) = \max_{\pi} \mathbb{E} \left[\sum_{t \geq 0} \gamma^t r_t | S_0 = s_0, A_t = a_0, \pi \right]$$

with

$$a_t \sim \pi(A_t | S_t = s_t), \quad s_{t+1} \sim P(S_{t+1} | A_t = a_t, S_t = s_t)$$

and r^t is the immediate reward for the state-action pair (s_t, a_t)

Update Q-value function

Empirically, we could use the following equation to update Q-value function

$$Q_{i+1}(s, a) = Q_i(s, a) + \eta \left(r + \gamma \max_{a'} Q_i(s', a') - Q_i(s, a) \right),$$

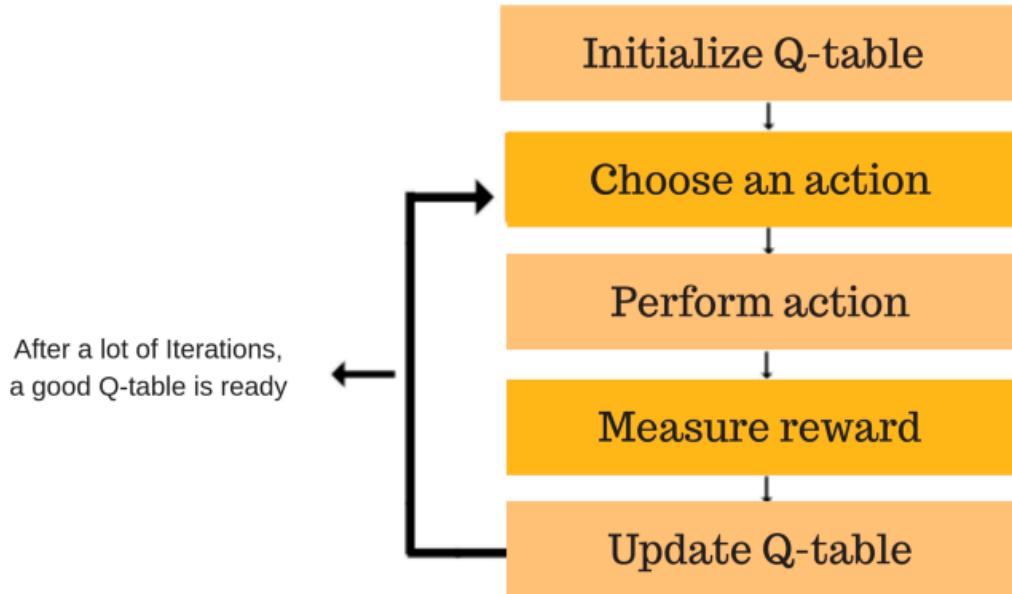
where $\eta \in (0, 1)$, and i is the number of iterations.

Intuitively, $r + \gamma \max_{a'} Q_i(s', a')$ is a more accurate approximation of the Q-value compared to $Q_i(s, a)$ because it contains a true value r .

We add a hyper-parameter η because the estimation $r + \gamma \max_{a'} Q_i(s', a')$ can not be fully trusted. If i goes to infinity, Q_i will converge to Q^* (see proof [I]).



Q-Table



Credit: <https://www.freecodecamp.org/news/a-brief-introduction-to-reinforcement-learning-7799af5840db/>



Deep Q-learning Network (DQN)

Deep Q-learning Network

For some tasks both state and action could be continuous. It is infeasible to store Q-value for every state-action pair into the Q table. Therefore, we use a neural network as a function approximator to store Q-values.

Deep Q-learning Network

We parameterise Q by weight w of neural network, i.e, Q_w .
The objective function is to iteratively minimise

$$(r + \gamma \max_{a'} Q_w(s', a') - Q_w(s, a))^2,$$

where state s' is observed by substituting action a and s into the environment.

Note that gradient is applied only to $Q_w(s, a)$, not to $Q_w(s', a')$

Policy Gradients

- Sometimes Q-function can be very complicated!
For example, a robot grasping an object has a very high-dimensional state (image), and it is hard to learn exact value of every (state, action) pair.
However the policy can be much simpler.

Why don't we directly learning the optimal policy?

- we can use the similar way as modelling Q-values, i.e., using a neural network to model policy π_θ , where, θ is the parameter of the network.

Policy Gradients

The expected reward of a policy θ is:

$$J(\theta) = \mathbb{E} \left[\sum_{t \geq 0} \gamma^t r_t | \pi_\theta \right],$$

and we want to find the optimal policy (parameters)

$$\theta^* = \arg \max_{\theta} J(\theta).$$



Calculate Gradient on θ

Could we use gradient ascent on policy parameters?

Let $r(\tau)$ be the reward of a trajectory:

$$\tau = (s_0, a_0, r_0, s_1, a_1, r_1, \dots)$$

mathematically, we could write the expected reward $J(\theta)$ as:

$$J(\theta) = \mathbb{E} \left[\sum_{t \geq 0} \gamma^t r_t | \pi_\theta \right] = \mathbb{E}_{\tau \sim p(\tau; \theta)} [r(\tau)] = \int_{\tau \sim p(\tau; \theta)} r(\tau) p(\tau; \theta) d\tau.$$

*each trajectory has probability
calculate the weighted cumulative reward*



Calculate Gradient on θ

Suppose we have some trajectories $\tau_1, \tau_2, \dots, \tau_n$, could we approximate:

$$\nabla_{\theta} J(\theta) = \int_{\tau \sim p(\tau; \theta)} r(\tau) \nabla_{\theta} p(\tau; \theta) d\tau,$$

by summation (Monte Carlo methods)

$$\nabla_{\theta} \hat{J}(\theta) = \frac{1}{n} \sum_{i=1}^n r(\tau_i) \nabla_{\theta} p(\tau_i; \theta)?$$

$$\xrightarrow{n \rightarrow \infty} \underbrace{\int [r(\tau) \nabla_{\theta} p(\tau; \theta)] p(\tau; \theta) d\tau}_{\text{Monte Carlo approximation}}$$

No! The integration is not an expected value! Because none of $r(\tau)$ and $\nabla_{\theta} p(\tau; \theta)$ are distributions! Then we can not approximate it by empirical mean!

Calculate Gradient on θ (Log Derivative Trick)

$$\begin{aligned}
 \nabla_{\theta} J(\theta) &= \int_{\tau \sim p(\tau; \theta)} r(\tau) \nabla_{\theta} p(\tau; \theta) d\tau, \\
 &= \int_{\tau \sim p(\tau; \theta)} r(\tau) p(\tau; \theta) \frac{\nabla_{\theta} p(\tau; \theta)}{p(\tau; \theta)} d\tau, \\
 &= \int_{\tau \sim p(\tau; \theta)} r(\tau) p(\tau; \theta) \nabla_{\theta} \log p(\tau; \theta) d\tau, \\
 &= \mathbb{E}_{\tau \sim p(\tau; \theta)} [r(\tau) \nabla_{\theta} \log p(\tau; \theta)].
 \end{aligned}$$

Then, we can proximate the expectation by

$$\nabla_{\theta} \hat{J}(\theta) = \frac{1}{n} \sum_{i=1}^n r(\tau_i) \nabla_{\theta} \log p(\tau_i; \theta).$$



Calculate Gradient on θ

Assume the Markov condition holds, i.e.,

$$P(s_{t+1}, a_t | s_t, s_{t-1}, \dots, s_0) = P(s_{t+1}, a_t | s_t),$$

then $p(\tau; \theta) = \prod_{t \geq 0} p(s_{t+1} | s_t, a_t) \pi_\theta(a_t | s_t)$. We have that,

$$\begin{aligned}\nabla_\theta \log p(\tau; \theta) &= \nabla_\theta \sum_{t \geq 0} \log p(s_{t+1} | s_t, a_t) + \log \pi_\theta(a_t | s_t), \\ &= \sum_{t \geq 0} \nabla_\theta \log p(s_{t+1} | s_t, a_t) + \nabla_\theta \log \pi_\theta(a_t | s_t), \\ &= \sum_{t \geq 0} \nabla_\theta \log \pi_\theta(a_t | s_t).\end{aligned}$$

Finally, we have a gradient estimator:

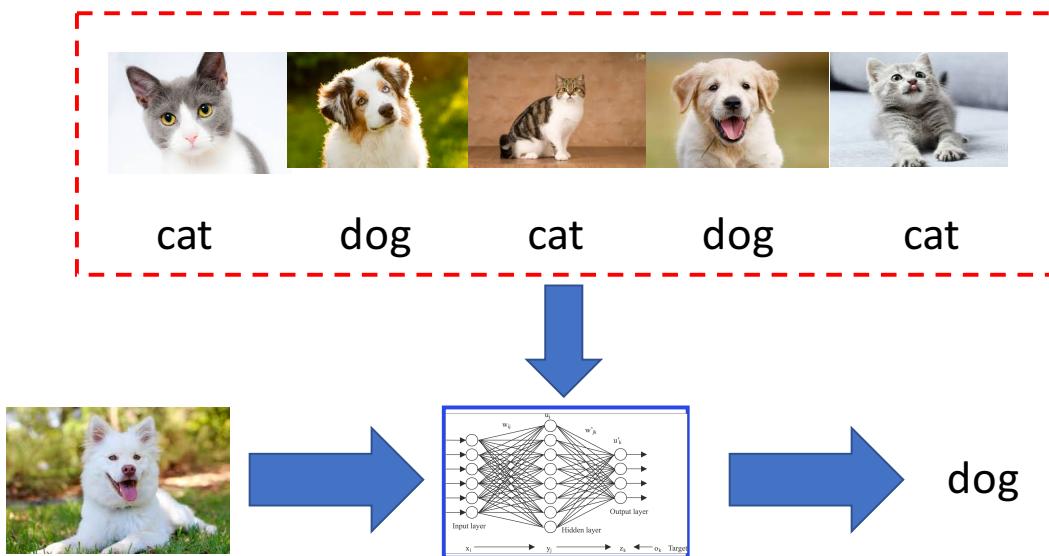
$$\nabla_\theta \hat{J}(\theta) = \frac{1}{n} \sum_{i=1}^n r(\tau_i) \sum_{t \geq 0} \nabla_\theta \log \pi_\theta(a_t | s_t) = \frac{1}{n} \sum_{i=1}^n \sum_{t \geq 0} r(\tau_i) \nabla_\theta \log \pi_\theta(a_t | s_t).$$



THE UNIVERSITY OF
SYDNEY

Dependence

Machine learning systems are usually driven by statistical dependence.



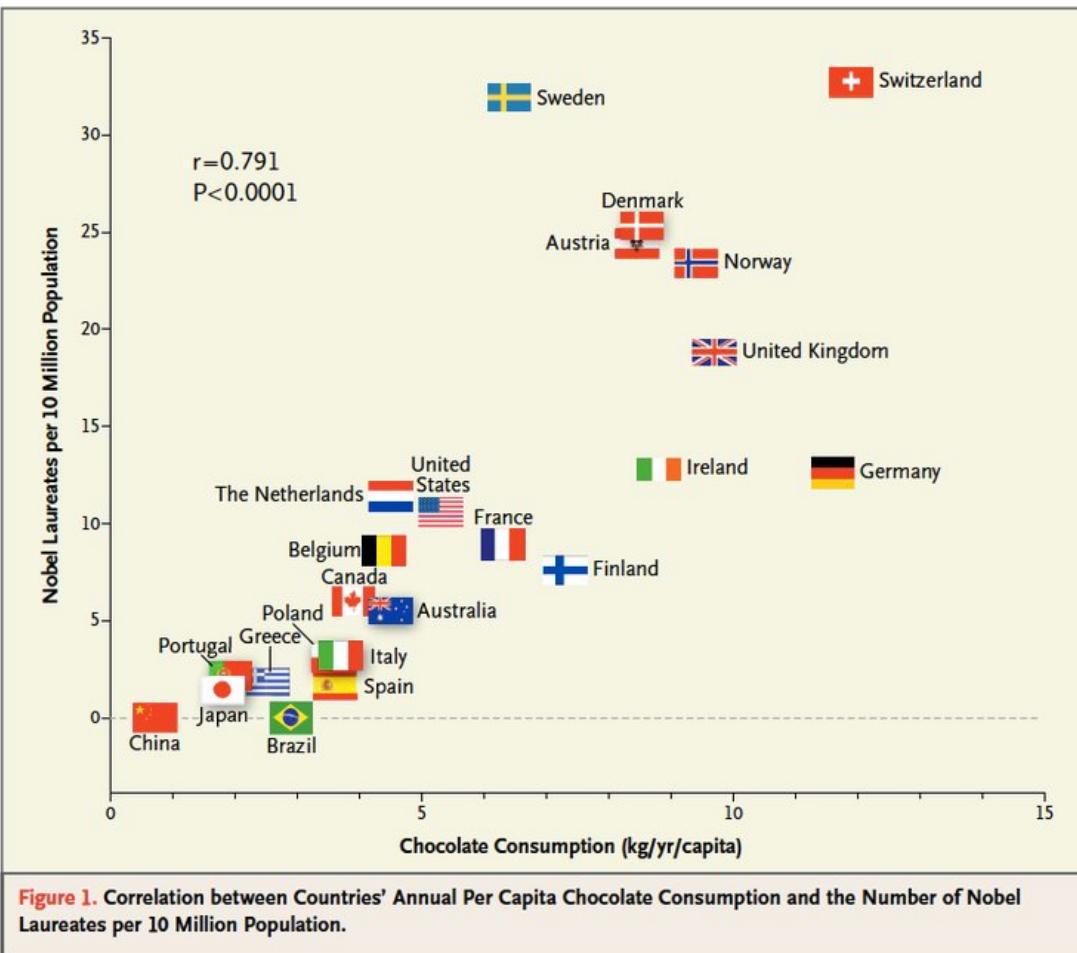


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

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



THE UNIVERSITY OF
SYDNEY



The NEW ENGLAND JOURNAL of MEDICINE

HOME ARTICLES & MULTIMEDIA ▾ ISSUES ▾ SPECIALTIES & TOPICS ▾ FOR AUTHORS ▾ CME ▾

ORIGINAL ARTICLE

A Correction Has Been Published ▶

Association of Coffee Drinking with Total and Cause-Specific Mortality

Neal D. Freedman, Ph.D., Yikyung Park, Sc.D., Christian C. Abnet, Ph.D., Albert R. Hollenbeck, Ph.D., and Rashmi Sinha, Ph.D.
N Engl J Med 2012; 366:1891-1904 | May 17, 2012 | DOI: 10.1056/NEJMoa1112010



can use dependence to make prediction
..... to achieve some aims (need causality)
can't

- The risk of death was increased among coffee drinkers
- Coffee drinkers were also more likely to smoke
- After adjustment for smoking, there was a significant inverse association between coffee assumption and mortality



THE UNIVERSITY OF
SYDNEY

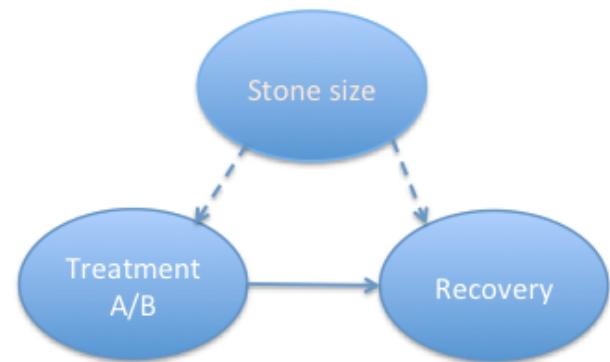
Causal thinking

Simpson's Paradox

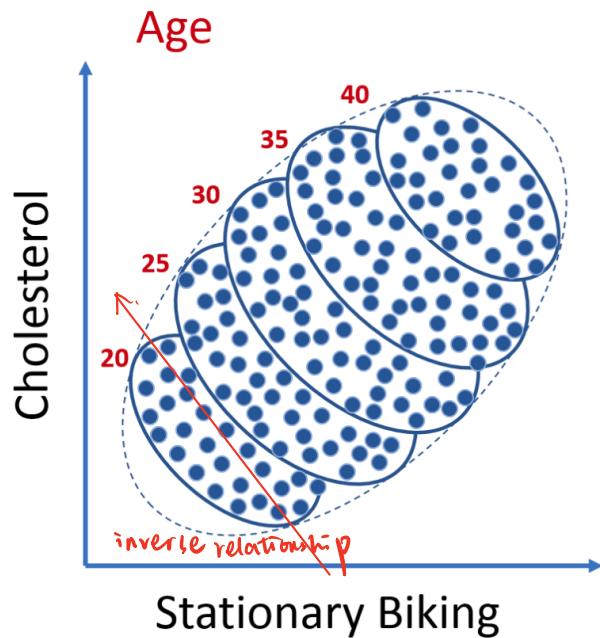
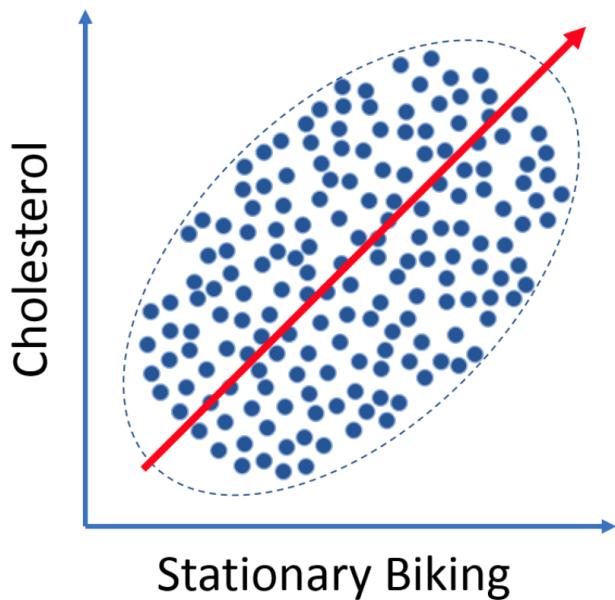
	Treatment A	Treatment B
Small Stones	Group 1 93% (81/87)	Group 2 87% (234/270)
Large Stones	Group 3 73% (192/263)	Group 4 69% (55/80)
Both	78% (273/350)	83% (289/350)

overall B better
w.r.t stone size A better

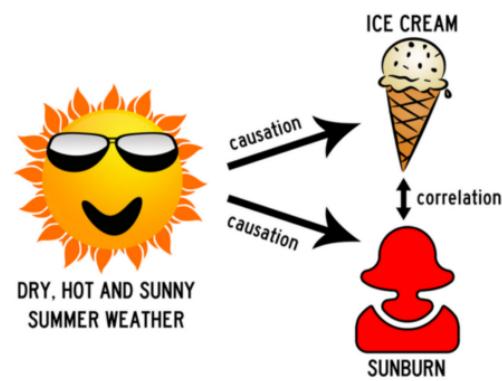
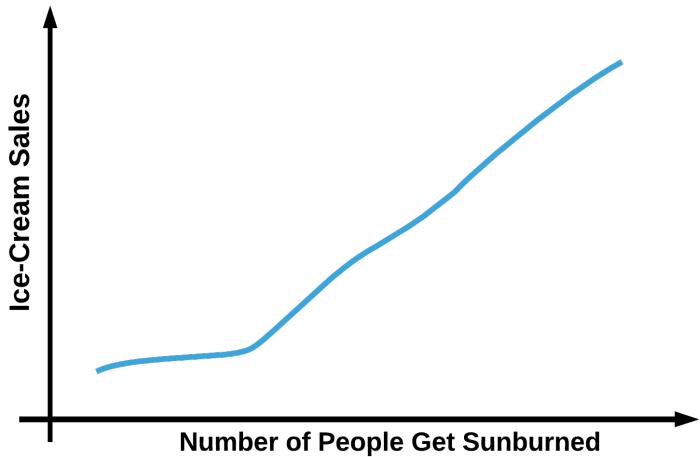
Would you make recommendations based on correlation or something else? A



“Strange” Dependence

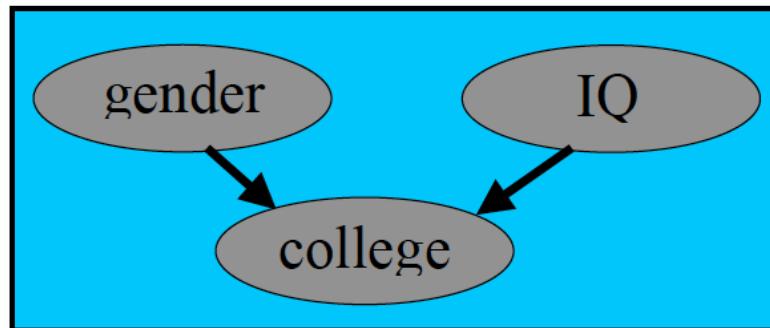


“Strange” Dependence



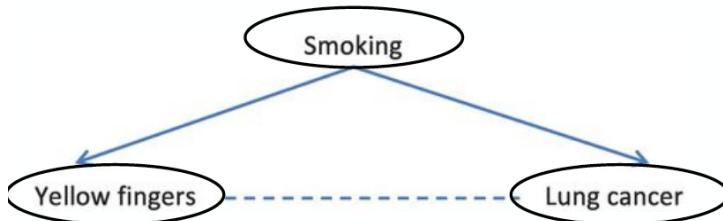
“Strange” Dependence

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



Selection bias!

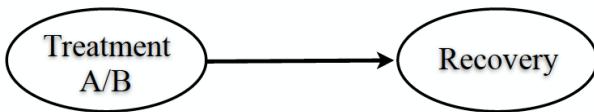
Ways to Produce Dependence



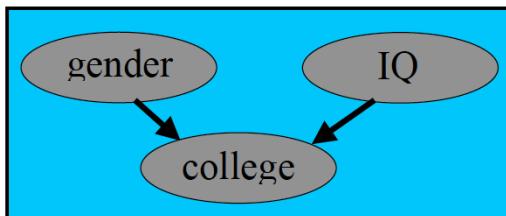
high dependence
but different causal structure

causal contains more information
→ more reliable for constructing ml models.

Common Cause



Causal relation



Conditional dependence
given common effect

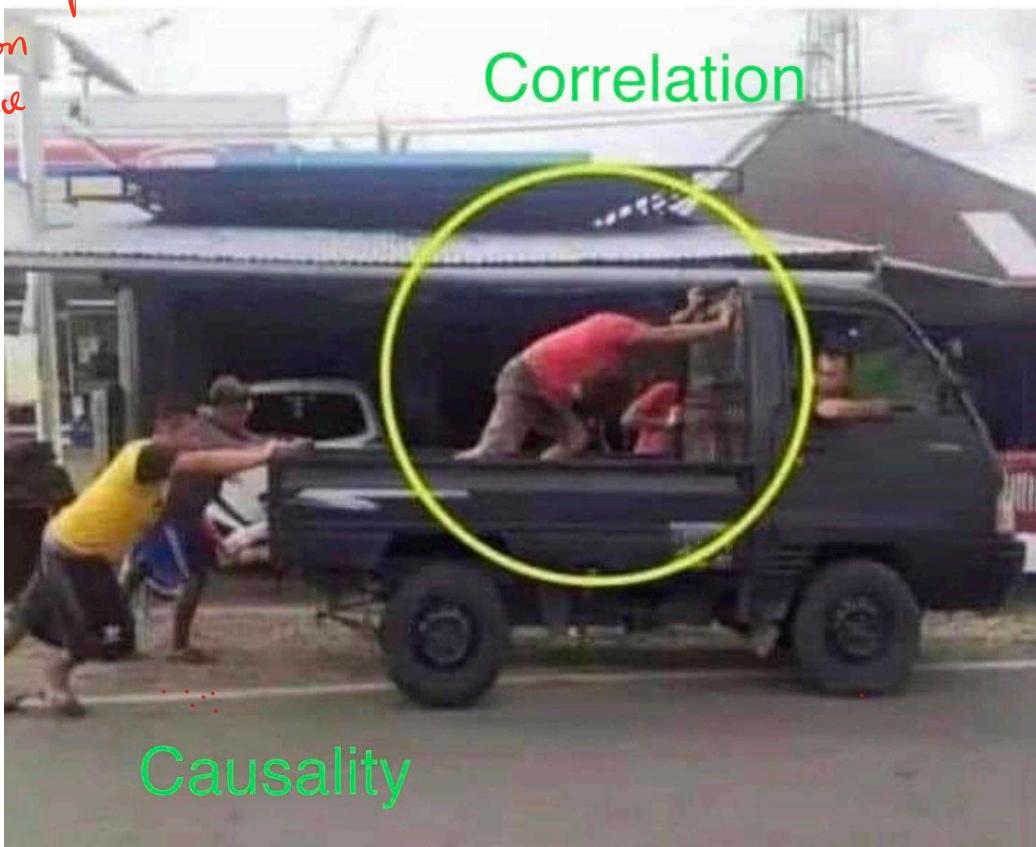
Causation vs Dependence

correlation ↔ dependence

high correlation
high dependence

Correlation

Causality



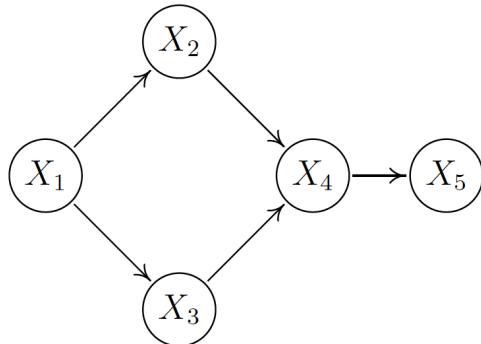


THE UNIVERSITY OF
SYDNEY

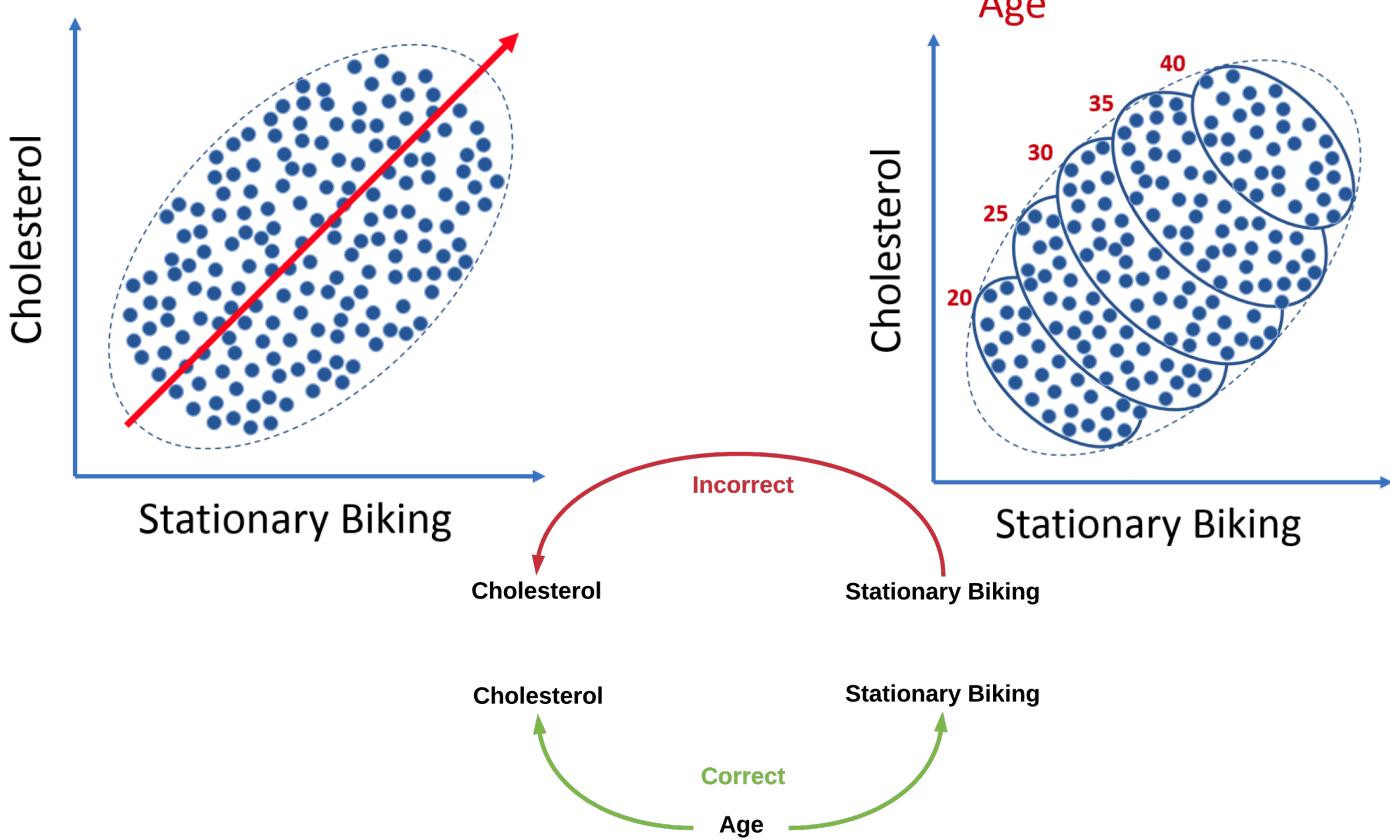
Causal representation

Directed Acyclic Graph (DAG)

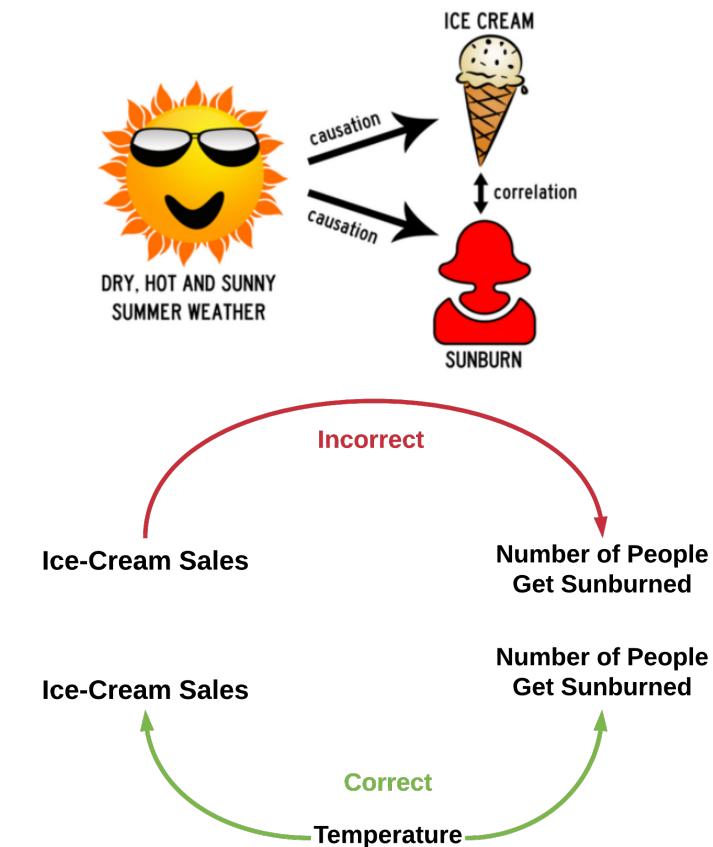
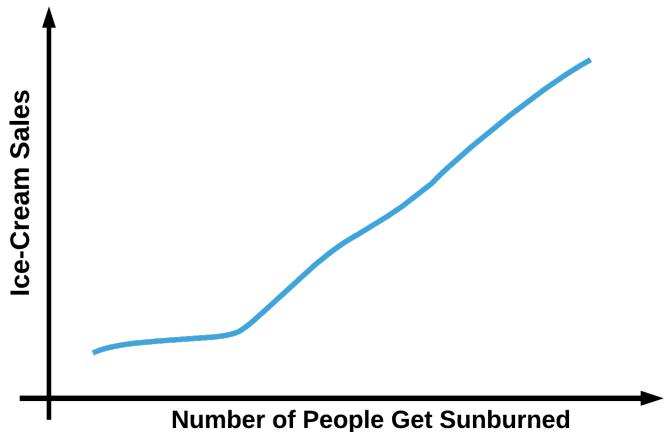
- **DAG**
 - graph with directed edges, no cycles
 - **Path**
 - a sequence of connected vertices
 - $X_1 - X_2 - X_4 - X_5$
 - $X_1 - X_2 - X_4 - X_3$
- require connection
not require direction*



“Strange” Dependence

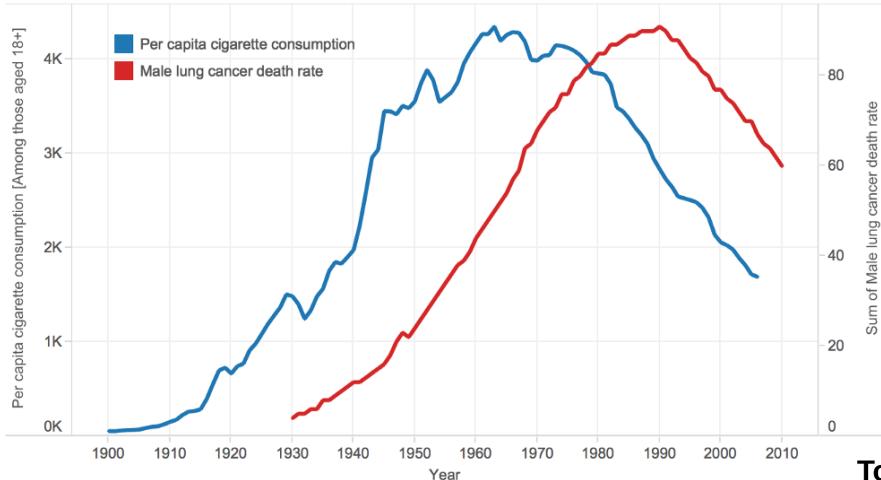


“Strange” Dependence



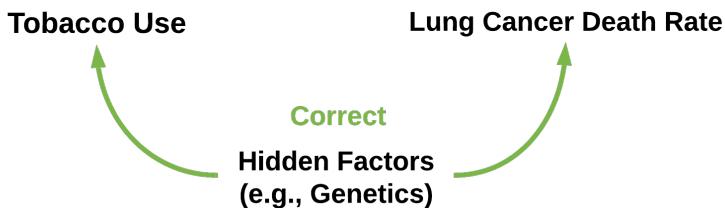
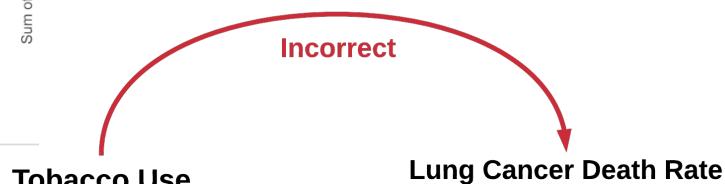
“Strange” Dependence

Trends in Tobacco Use and Lung Cancer Death Rates in the U.S.



Death rates source: US Mortality Data, 1960-2010, US Mortality Volumes, 1930-1959, National Center for Health Statistics, Centers for Disease Control and Prevention.

Cigarette consumption source: US Department of Agriculture, 1900-2007.





Markov Conditions

- 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.



$X \perp\!\!\!\perp Y$: X and Y are independent
 $X \not\perp\!\!\!\perp Y$ = X and Y are not independent
(dependent)

ancestor vs. descendant
parent vs. children

Local Markov Condition

Every variable X_i , in a directed acyclic graph, is independent of its non-descendant Y conditional on its parents, i.e., $(X_i \perp\!\!\!\perp Y | PA_i)_G^{w.r.t \text{graph}}$, which also implies $(X_i \perp\!\!\!\perp Y | PA_i)_p$.

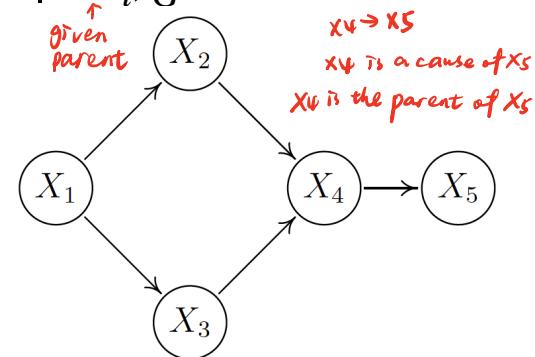
from the perspective of statistics

For example:

$$(X_4 \perp\!\!\!\perp X_1 | \{X_2, X_3\})_G \implies (X_4 \perp\!\!\!\perp X_1 | \{X_2, X_3\})_p.$$

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

$$\begin{aligned} P(X_1, \dots, X_5) &= P(X_1)P(X_2 | X_1)P(X_3 | X_1)P(X_4 | X_2, X_3)P(X_5 | X_4) \\ &= p(x_5 | x_1, x_2, x_3, x_4) \cdot p(x_1, x_2, x_3, x_4) \\ &= p(x_5 | x_4) \cdot p(x_4 | x_1, x_2, x_3) \cdot p(x_1, x_2, x_3) = \dots \end{aligned}$$





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 $i \rightarrow m \rightarrow j$ or a fork $i \leftarrow m \rightarrow j$ such that the middle node m is in \mathbf{S} , or
- q contains a collider $i \rightarrow m \leftarrow j$ 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 $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} | \mathbf{S})_p$.

X influence Y		
若X与Y相互独立，就说明X不会影响(influence) Y。将上述4种情况总结起来，见下图：		
$X \rightarrow W \rightarrow Y$	$W \notin Z$	\checkmark
$X \leftarrow W \leftarrow Y$	$W \in Z$	\times
$X \leftarrow W \rightarrow Y$	$W \in Z$	\times
$X \rightarrow W \leftarrow Y$	$W \notin Z$	\checkmark
http://blog.csdn.net/ybdeite		

Active tail.

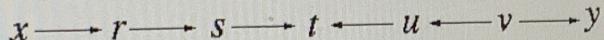
$X \rightarrow X \subseteq X$ X 或 X 的子事件属于 Z 。
没有 X 属于 Z (Z 不为空)。

不 Active，就说明 X 和 Y 关于 Z 是 d-separation

① 碰撞点 (Collider)

没有条件集的独立性

规则1：如果 x 到 y 的任一 path(路径) 都经过 collider(碰撞点)，则 x 和 y 独立。注意，这里的路径是忽略边的方向的，而碰撞点是指有多个箭头指向它的节点，即类似于下图的 $s \rightarrow t \leftarrow u$ 。



现在我们看看这个图变量之间的独立性是怎样的。先考虑 x 和 t 的路径： $x \rightarrow r \rightarrow s \rightarrow t$ 。这条路径中并不存在碰撞点，所以 x 和 t 不独立，同样地， t 和 y 、 u 、 v 都不独立。然而，对于变量 x 到 y 而言， x 、 y 之间在碰撞点 t 处必然会经过碰撞点 t ，所以 x 和 y 是独立的，同样地，在碰撞点两侧的变量都是独立的，比如 x 和 v 、 s 和 u ， r 和 u 也是独立的。

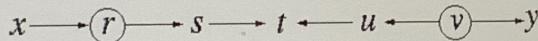
② block (chain / fork)

一般的条件独立

分享

那么如果一条路径中没有碰撞点，怎么才能让他们独立呢？我们还能使用条件独立的性质，只要条件集 Z 能够将一条路径 block 掉，那么就可以让两个变量独立。注意的是，当碰撞点或碰撞点的子代出现在条件集中时候要小心，很有可能会导致不同的结果，这个问题我们留到规则3。现在我们先假设条件集中不存在碰撞点。

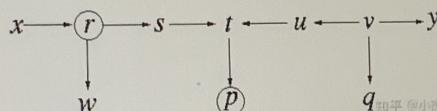
规则2：当 x 到 y 之间的任一路径都经过 Z 中的节点，且 Z 并不包含碰撞点或碰撞点的子代，则 x 和 y 独立。



如图，设结点集 $Z = \{r, v\}$ （图中画圈的节点），根据规则2， x 和 s 是条件独立的，因为 x 和 s 之间的路径被 Z block 掉了，同样地， u 和 y 也是条件独立的。

③

规则3：当碰撞点或碰撞点的子孙节点为集合 Z 的成员时，该碰撞点不再截断路径。



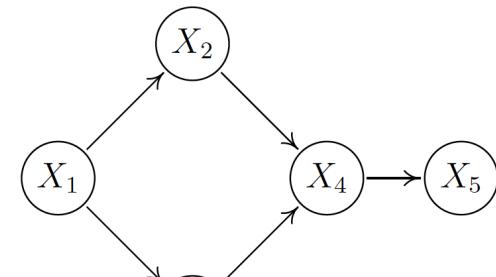
设结点集 $Z = \{r, p\}$ （图中画圈的节点），根据规则3，在给定 Z 的情况下， s 和 y 不独立，因为 t 的孩子节点 p 在集合 Z 中，所以 t 没办法像规则1一样截断路径 $s \rightarrow t \rightarrow u \rightarrow v \rightarrow y$ ，与此相反，条件集 p 使得 s 和 u 变得不独立了。然而在这里， x 和 u 是独立的，虽然 t 不能截断它，但是 r 可以截断它（根据规则2）。

Global Markov Conditions (D-Separation)

eg $X = \{X_1\}$
 $Y = \{X_4, X_5\}$
 $S = \{X_2, X_3\}$
 d-separated

- 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 $\mathbf{S} = \{X_1\}$.

contain A chain
 not all chain



folk
 $X_2 \leftarrow X_1 \rightarrow X_3$

collider
 $X_2 \rightarrow X_4 - X_3$



Causal faithfulness

use statistical dependence
to find causal structure

Assumption

machine learning → statistical dependence → causal structure
causal
faithfulness assumption

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.,

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

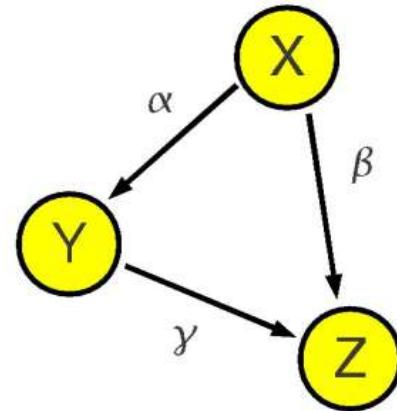
then the distribution p is said to be faithful to the graph.

Causal faithfulness Assumption

Here is an example of a **unfaithful distribution**

$$\begin{aligned} X &= U_x, \\ Y &= \alpha X + U_y, \\ Z &= \beta X + \gamma Y + U_z. \end{aligned}$$

$$\begin{aligned} p + \alpha r &= 0 \\ z &= -\alpha x + r y + u_z \\ &= -\alpha x + \alpha x + \alpha y + u_z \\ &= \alpha y + u_z \end{aligned}$$



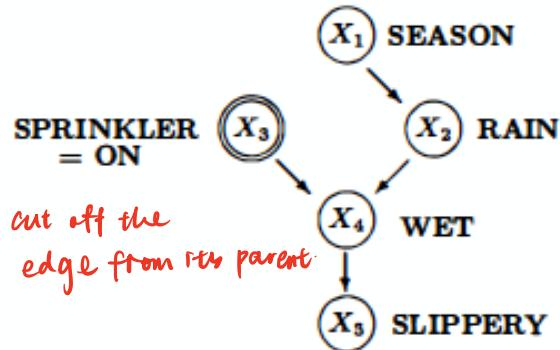
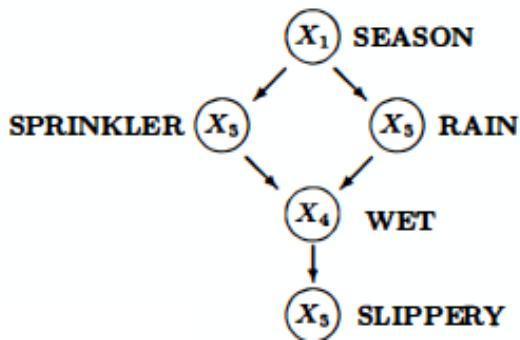
From the graph $(\{X\} \perp\!\!\!\perp \{Z\})_G$, however, if $\beta + \alpha\gamma = 0$, then $(\{X\} \perp\!\!\!\perp \{Z\})_p$, and $(\{X\} \perp\!\!\!\perp \{Z\})_p$ does not imply $(\{X\} \perp\!\!\!\perp \{Z\})_G$.

Causal Bayesian Net

- **Causal Bayesian Network**

- directed edges representation causal direction, causal DAG
- more meaningful and represent external changes

if directed edge = causal relationship \Rightarrow causal DAG





Conditioning, Intervention, Counterfactual

- **Prediction/Conditioning**
 - 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 turn off the sprinkler?

$$P(\text{slippery} \mid \text{sprinkler} = \text{off}) \neq P(\text{slippery} \mid \text{do}(\text{Sprinkler} = \text{off}))$$

manually cutoff

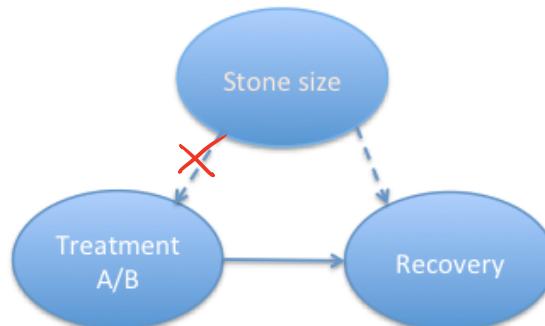
- **Prediction/Counterfactual reasoning**
 - would the pavement be slippery, had the sprinkler been off, given that the pavement is in fact not slippery and the sprinkler is on?

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

Identification of Causal Effects

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

- “**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!

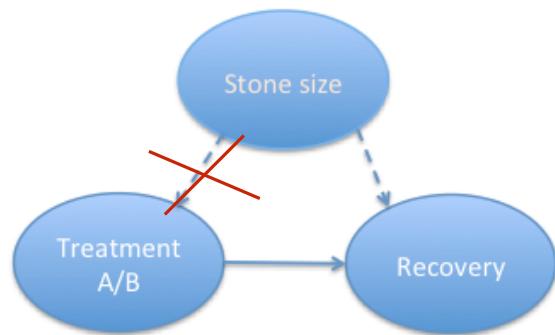
Example

	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

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

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



Intervention vs. conditioning



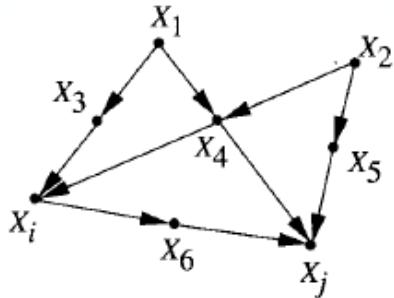
Back-Door Criterion

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:

* 要求

- (i) no node in Z is a descendant of X_i ; and Z 阻断 (X_i, X_j) 之间的所有指向 X_i 的路
- (ii) Z blocks every path between X_i and X_j that contains an arrow into X_i .

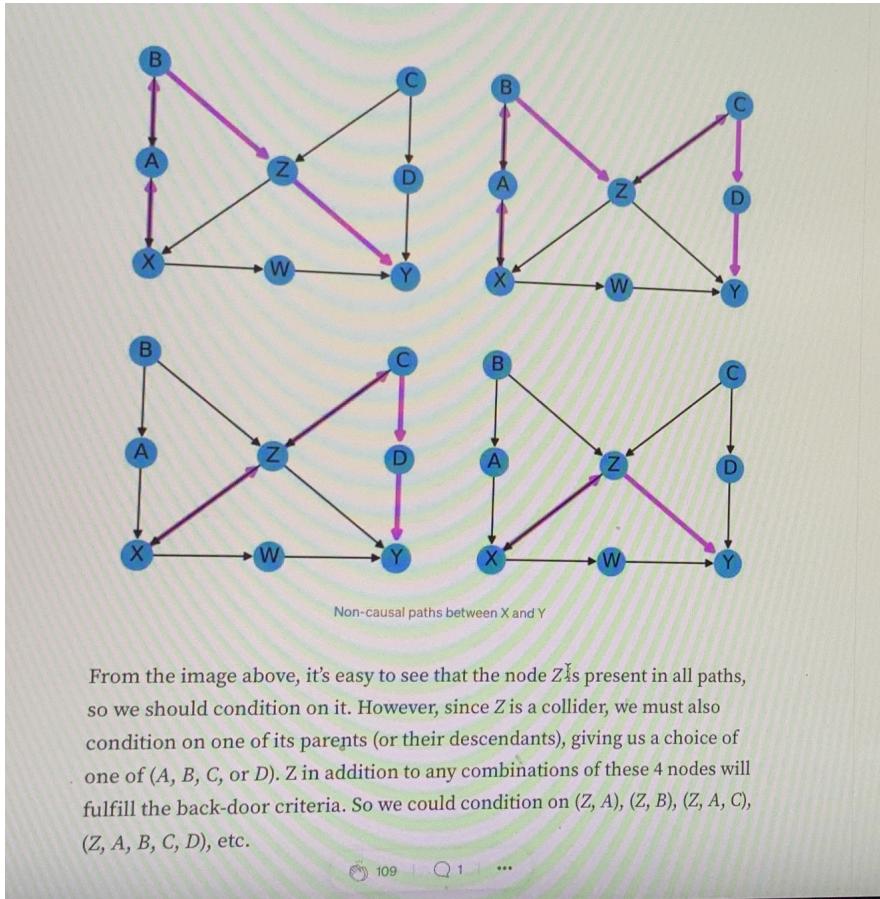


- What if $Z = \{X_3, X_4\}$?
 $Z = \{X_4, X_5\}$?
 $Z = \{X_4\}$?
- What if there is a confounder?

Theorem 3.3.2 (Back-Door Adjustment)

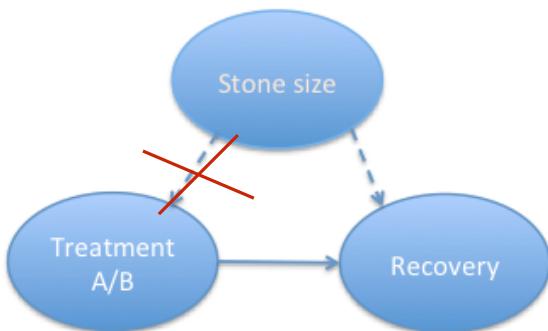
If a set of variables Z satisfies the back-door criterion relative to (X, Y) , then the causal effect of X on Y is identifiable and is given by the formula

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

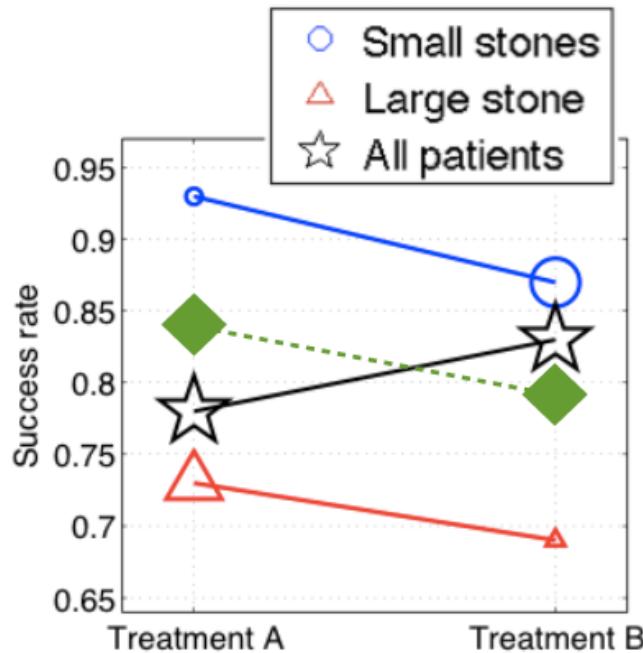


Example

	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)

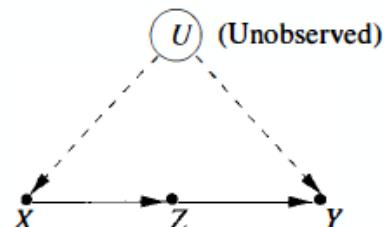


T - Treatment
R - Recovery





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:

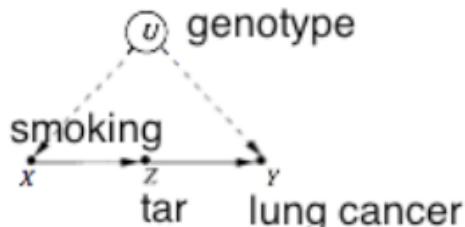
- (i) Z intercepts all directed paths from X to Y ;
- (ii) there is no back-door path from X to Z ; and
- (iii) all back-door paths from Z to Y are blocked by X .

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

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

Example: Smoking



Group Type	$P(x, z)$ Group Size (% of Population)	$P(Y = 1 x, z)$ % 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

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

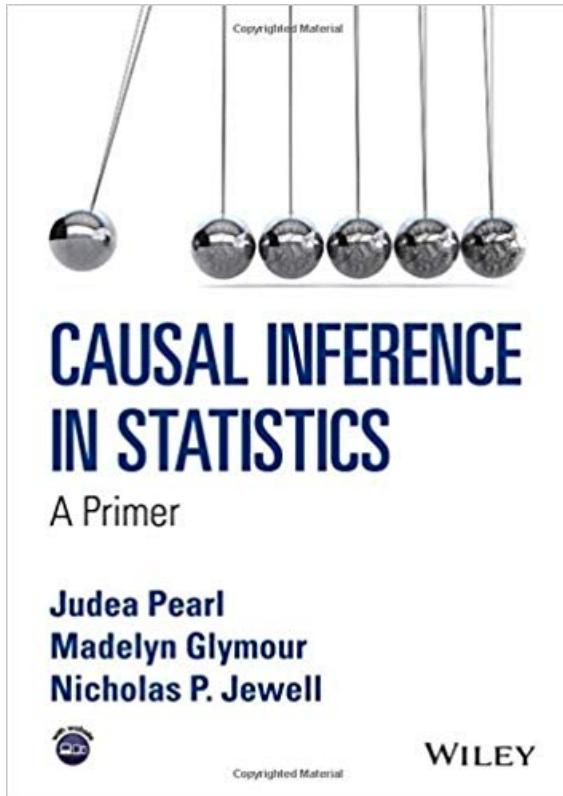
$$\begin{aligned}
 P(Y = 1 | do(X = 0)) &= .95(.10 \times .50 + .90 \times .50) \\
 &\quad + .05(.05 \times .50 + .85 \times .50) \\
 &= .95 \times .50 + .05 \times .45 = .4975.
 \end{aligned}$$

$$\begin{aligned}
 & P(Y = 1|do(X = 1)) \\
 &= P(Z = 0|X = 1) [P(Y = 1|X = 1, Z = 0)P(X = 1) + P(Y = 1|X = 0, Z = 0)P(X = 0)] \\
 &\quad + P(Z = 1|X = 1) [P(Y = 1|X = 1, Z = 1)P(X = 1) + P(Y = 1|X = 0, Z = 1)P(X = 0)] \\
 &= 0.05 \times (0.10 \times 0.50 + 0.90 \times 0.50) + 0.95 \times (0.05 \times 0.50 + 0.85 \times 0.50) \\
 &= 0.05 \times 0.50 + 0.95 \times 0.45 \\
 &= 0.4525
 \end{aligned}$$
$$\begin{aligned}
 & P(Y = 1|do(X = 0)) \\
 &= P(Z = 0|X = 0) [P(Y = 1|X = 1, Z = 0)P(X = 1) + P(Y = 1|X = 0, Z = 0)P(X = 0)] \\
 &\quad + P(Z = 1|X = 0) [P(Y = 1|X = 1, Z = 1)P(X = 1) + P(Y = 1|X = 0, Z = 1)P(X = 0)] \\
 &= 0.95 \times (0.10 \times 0.50 + 0.90 \times 0.50) + 0.05 \times (0.05 \times 0.50 + 0.85 \times 0.50) \\
 &= 0.95 \times 0.50 + 0.05 \times 0.45 \\
 &= 0.4975
 \end{aligned}$$



THE UNIVERSITY OF
SYDNEY

Causal Inference



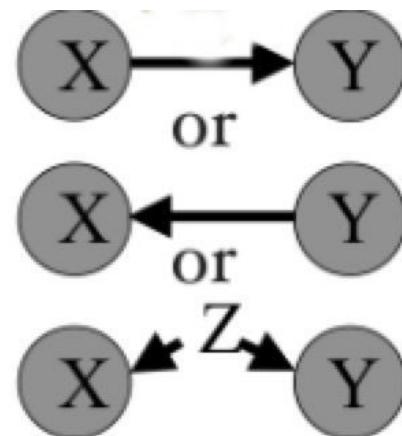
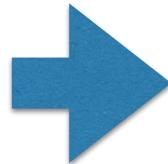


Causal Discovery

Is it possible to discover causal structure from purely observational data?

Yes . If causal faithfulness assumption holds true

X	Y
-1.1	1.0
2.1	2.0
3.1	4.2
2.3	
-0.6	
1.3	2.2
-1.8	0.9
...





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.