Causality: Inferencing and Learning

Antoine Cornuéjols

AgroParisTech – INRAE MIA Paris-Saclay

EKINOCS research group

AgroParisTech

INRAE
la science pour la vie, l'humain, la terre

A **first route** towards causality

An empirical motivation

1. Does **aspirin** reduce risk for **cardiovascular disease**?

2. If so, **how large** is this effect?

3. Is the size of this effect large **relative to the effects** of other possible causes?

1. Does **glyphosate** increase the risk of **cancers**?

2. Is there an impact of **neonicotinoids** on **honey bees**?

3. Is the **socioeconomic status** a cause of **health disparities**?

Event

Variable of interest

Binary variable
(e.g. treatment, no treatment)

Real value
(e.g. temperature)

Binary variable
(e.g. ill, not ill)

Real value
(e.g. yield)

- Causation is not (linear) correlation
  - Illustration with [Aleksander Molak]
- Correlation is not causation
  - The Monty Hall paradox ?

Correlation is not causation

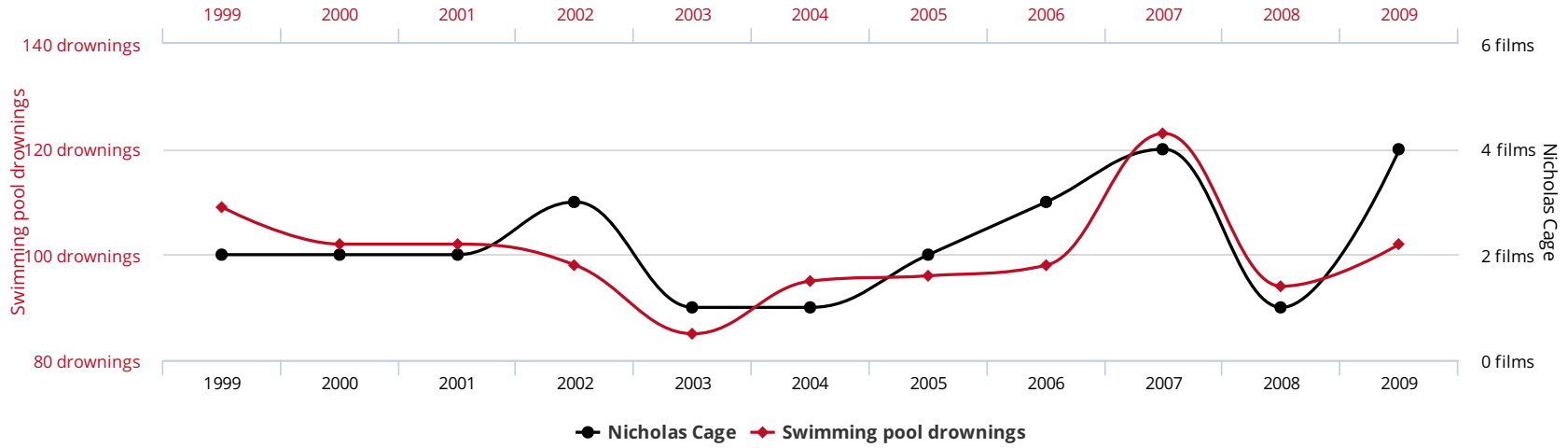


Figure 1.3: The yearly number of movies Nicolas Cage appears in correlates with the yearly number of pool drownings [1].

From [Brady Neal (2020). **Introduction to Causal Inference from a Machine Learning Perspective.** (unfinished manuscript)]

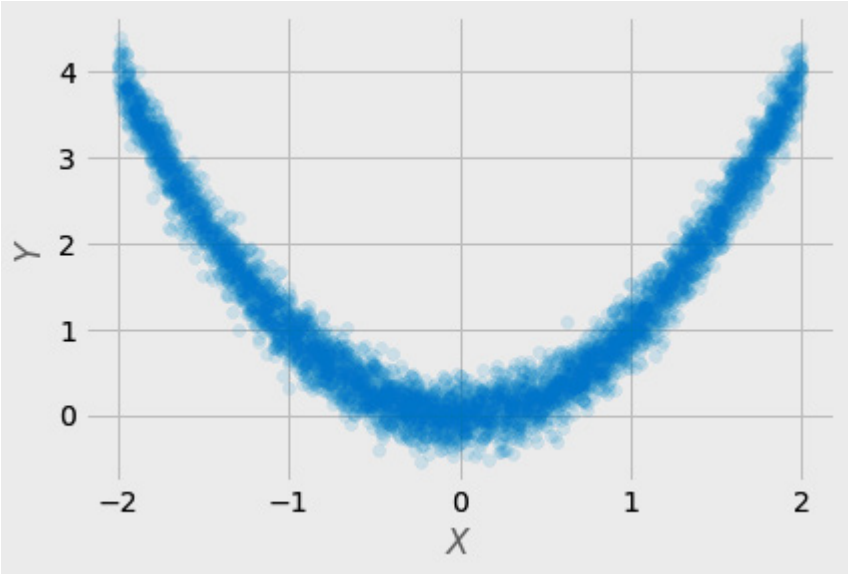
Can there be **causation without correlation**?

- It depends on the measurement of correlation

$$X := \mathcal{U}(-2, 2)$$

$$Y := X^2 + 0.2 \times \mathcal{N}(0, 1)$$

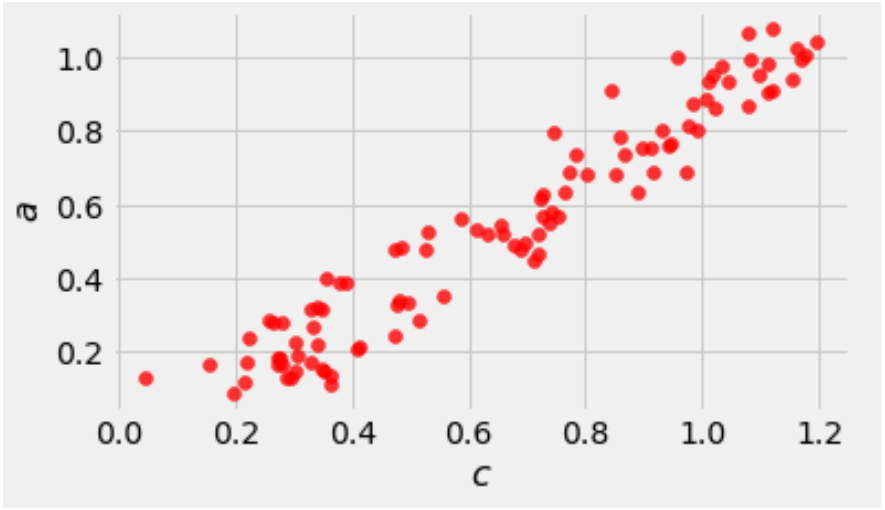
Popular correlation metrics such as **Pearson**'s *r* or **Spearman**'s *rho* give a coefficient essentially equal to 0.
They are not appropriate to measure non-linear correlation.



Other correlation metrics such as the **Maximal Information Coefficient** (MIC) or the **Hilbert-Schmidt Independence Criterion** (HSIC) **work for non-linear, non-monotonic data**.

Confounding and statistics

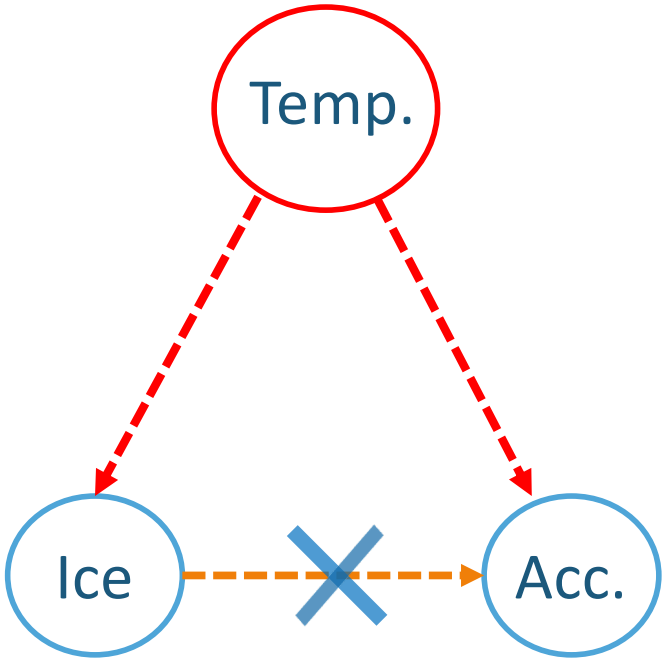
- **Ice** : buying ice-cream
- **Acc.** : having a swimming accident (e.g. drowning)



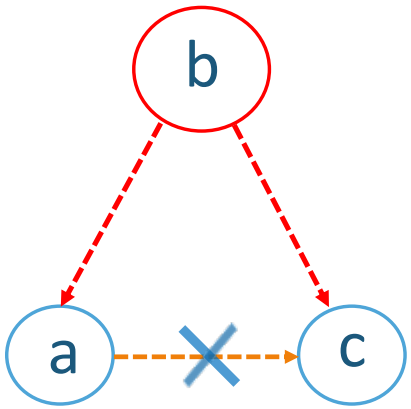
- There is a strong linear correlation between the two variables
- Does buying ice-cream **cause** more swimming accidents?

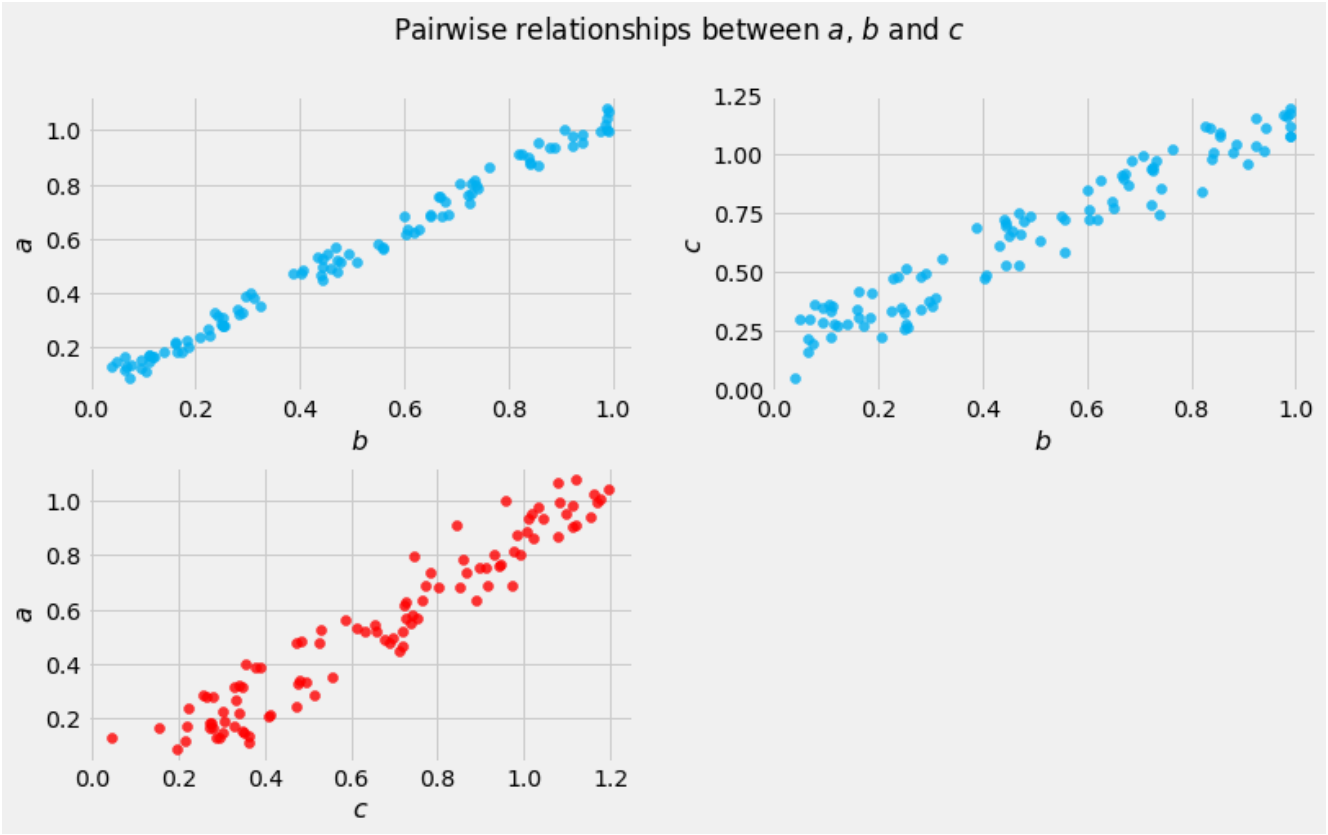
Confounding and statistics

- **Ice** : buying ice-cream
- **Acc**. : having a swimming accident (e.g. drowning)



- High temperature
  - Makes people more likely to **buy ice-cream**
  - Makes people more likely to **go swimming**





- Most of the time, we cannot distinguish the spurious (red) and non-spurious (blue) relationships from statistical data alone

Simpson's paradox

- Suppose we have **two treatments**, A and B, with the following **success rates** in the population:

	Treatment **A**	Treatment **B**
Population	78%	83%

- **Which one** would you choose to administer?

Simpson's paradox

- Suppose we have **two treatments**, A and B,
- with the following **success rates** given the **condition**, either 'small stones' or 'large stones':

	Treatment **A**	Treatment **B**
Small stones	**93**%	87%
Large stones	**73**%	69%
Population	78%	**83**%

- **Which one** would you choose to administer?

Simpson's paradox

- Suppose we have **two treatments**, A and B,
- with the following **success rates** given the **condition**, either 'small stones' or 'large stones':

	Treatment **A**	Treatment **B**	
Small stones	**93**% (81/87)	87% (234/270)	357
Large stones	**73**% (192/263)	69% (55/80)	343
Population	78% (273/350)	**83**% (289/350)	700

- **Which one** would you choose to administer?

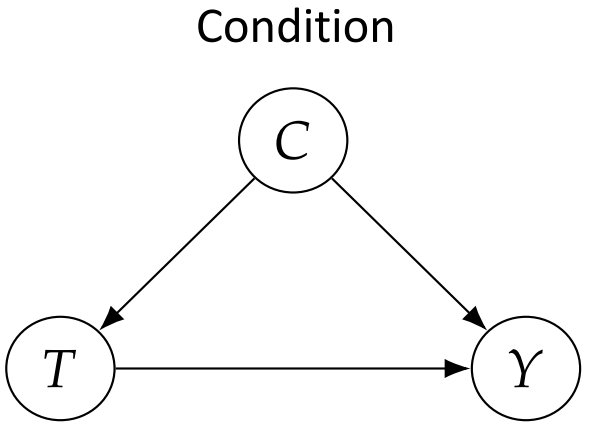
Simpson's paradox

	Treatment **A**	Treatment **B**	
Small stones	**93**% (81/87)	87% (234/270)	357
Large stones	**73**% (192/263)	69% (55/80)	343
Population	78% (273/350)	**83**% (289/350)	700

- Higher rate of success for 'small stones' than 'large stones' **independently** of treatment
- Treatment **B** is given **predominantly for the 'small stones'** condition: 270 / (270 + 80) = 77.1%
- Treatment **A** is given to 'small stones' **only** 87 / (87 + 263) = 23.1%

Simpson's paradox

	Treatment A	Treatment B	
Small stones	**93**% (81/87)	87% (234/270)	357
Large stones	**73**% (192/263)	69% (55/80)	343
Population	78% (273/350)	**83**% (289/350)	700



- Higher rate of success for 'small stones' than 'large stones' **independently** of treatment
- Treatment **B** is given **predominantly for the 'small stones'** condition: 270 / (270 + 80) = 77.1%
- Treatment **A** is given to 'small stones' **only** 87 / (87 + 263) = 23.1%

→ Treatment **A** is **better**

- What challenges our understanding is that, intuitively, **we assum**e that the treatment is **independent** of the condition.

- But wait a minute, what if somebody else gives you a **further segmented data** that show, say, for both small stone and large stone cases, when segmented by **gender**, *treatment B is better than treatment A for all gender*?!

- We can ask this kind of questions on and on and reach to a conclusion that **no conclusion can be made** based on any type of.
  Now we see the real and unsettling problem.

Can we **learn** anything **from data?**

Simpson's paradox

Suppose a new disease COVID-27

- Two treatments: A and B
- You have **data** on the **percentage of people who die** from COVID-27, given the treatment they were assigned and given their condition at the time the treatment was decided.
- **You are in charge** of choosing which treatment your country will use

Simpson's paradox

Suppose a new disease COVID-27

- Two treatments: A and B
- You have **data** on the **percentage of people who die** from COVID-27, given the treatment they were assigned and given their condition at the time the treatment was decided.
- **You are in charge** of choosing which treatment your country will use

		Condition		
		Mild	Severe	Total
Treatment	A	15% (210/1400)	30% (30/100)	**16%** (240/1500)
	B	**10%** (5/50)	**20%** (100/500)	19% (105/550)

Which one do you **choose**?

Treatment	Condition: Mild	Condition: Severe	Condition: Total
A	15% (210/1400)	30% (30/100)	**16%** (240/1500)
B	**10%** (5/50)	**20%** (100/500)	19% (105/550)

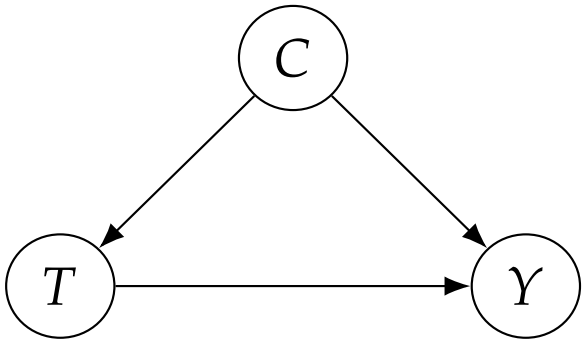
		Condition		
		Mild	Severe	Total
Treatment	A	15% (210/1400)	30% (30/100)	**16%** (240/1500)
	B	**10%** (5/50)	**20%** (100/500)	19% (105/550)

Well, … it depends

Treatment	Condition: Mild	Condition: Severe	Condition: Total
A	15% (210/1400)	30% (30/100)	**16%** (240/1500)
B	**10%** (5/50)	**20%** (100/500)	19% (105/550)

Well, … it depends

Scenario 1



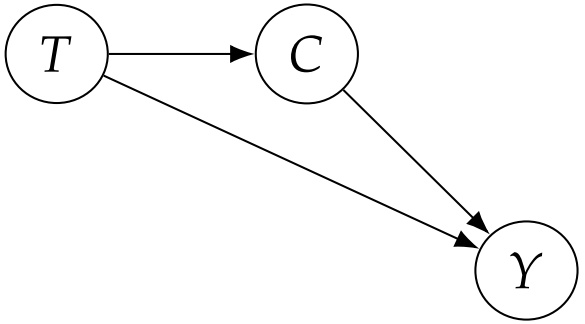
Doctors decide to give **treatment A** to most people who have **mild** conditions.

Thus, treatment B is associated with a higher mortality rate simply because **condition is a common cause** of both treatment and mortality.

The best treatment is the one that yields lower mortality in each of the subpopulations: **treatment B**

		Condition: Mild	Condition: Severe	Condition: Total	
Treatment	A	15% (210/1400)	30% (30/100)	**16%** (240/1500)	73%
	B	**10%** (5/50)	**20%** (100/500)	19% (105/550)	27%

Well, … it depends



Scenario 2

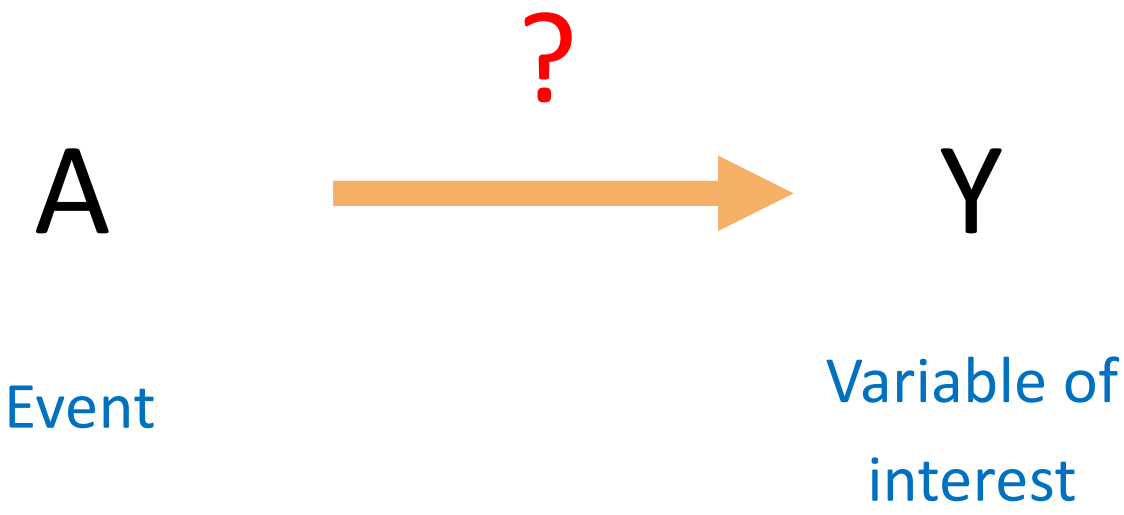
Here, treatment B is so scarse that it requires patients to wait a long time before they can receive the treatment. And the condition of the patient worsens over time, causing a higher mortality rate.

Even if treatment B is more effective that treatment A *once administered* (positive effect condition -> effect), it causes worse conditions in total. One should choose **treatment A**.

Lessons

- The causal relationships **brings information** that is not present in the conditional probabilities

- **Knowledge of the causal graph** changes the decision here



We would like to **measure** the **effect** of A on Y

Individual causal effect

- *A*: treatment variable (either 0 or 1 here)
- *Y*: outcome for an individual

A has a **causal effect** on an individual's outcome *Y* if

$$Y^{a=1} \neq Y^{a=0}$$

for the individual

Individual causal effect

- *A*: exposition to neonicotinoid (either 0 or 1)
- *Y*: bee alive, bee dead (0, 1)

Neonicotinoid has a **causal effect** on the *bee survival* if

$$Y^{a=1} \neq Y^{a=0}$$

for the individual bee

Individual causal effect

- *A*: disease occurrence (either 0 or 1)
- *Y*: crop yield in a given site-year.

  E.g. a given wheat field in Saclay in 2023

Disease A has a **causal effect** on *the crop yield* if

$$Y^{a=1} \neq Y^{a=0}$$

for the crop field considered

Individual causal effects

- But it is **impossible** to subject the same individual to two different treatments

Average causal effects

- There is an **average causal effect** in the population if

$$\mathbb{E}[Y^{a=1}] \neq \mathbb{E}[Y^{a=0}]$$

Average causal effect

- *A*: disease occurrence (either 0 or 1)
- *Y*: crop yield for all wheat sites in France in 2023

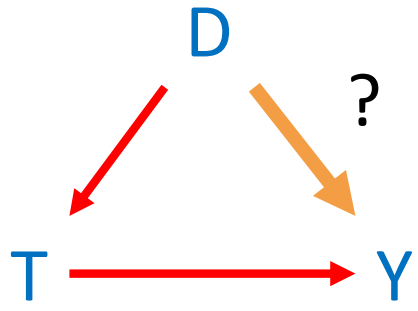
Disease A has an **average causal effect** on *the crop yield* if

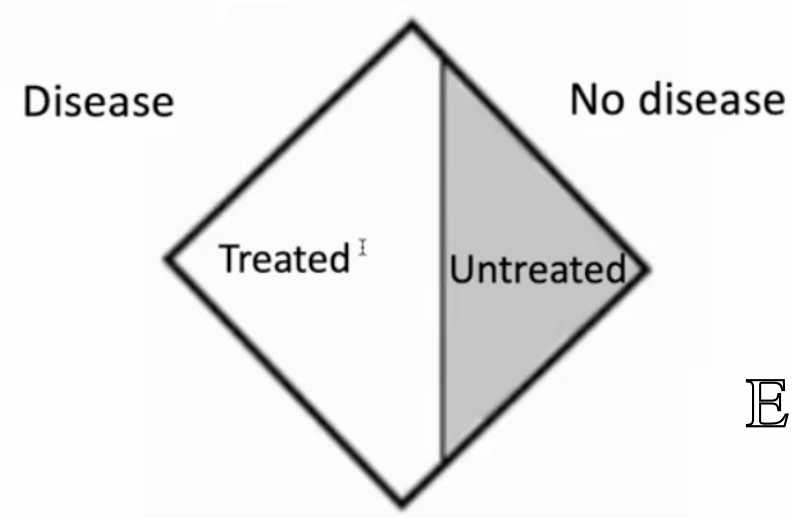
$$\mathbb{E}[Y^{a=1}] \neq \mathbb{E}[Y^{a=0}]$$

for the crop fields considered

Risk of confounding

- But **how** to *measure average causal effects*?
- Risk of confounding
  - If treatment is given when disease



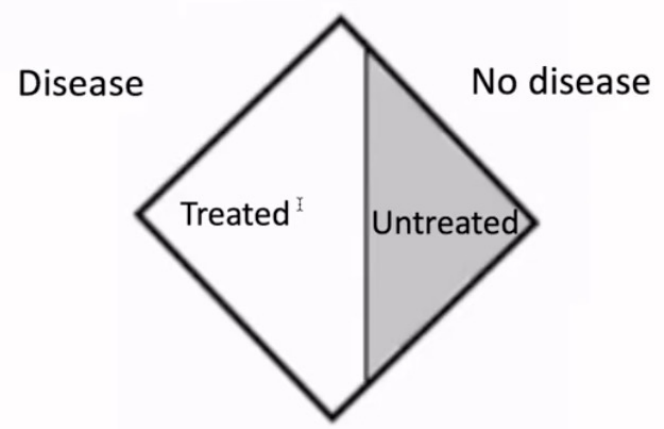


$$\mathbb{E}[Y^{NoDisease,Untreated}] = \mu_0$$

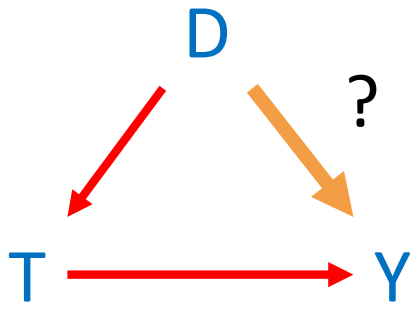
$$\mathbb{E}[Y^{Disease,Treated}] = \mu_0 - \alpha_D + \beta_T$$

Risk of confounding

- Risk of confounding
  - If treatment is given when disease



D is a confounder



$$\mathbb{E}[Y^{Disease,Treated}] - \mathbb{E}[Y^{NoDisease,Untreated}] = -\alpha_D + \beta_T$$

The effect of the disease will be **underestimated** because of the treatment

Two fundamental questions:

1. An entity cannot be both **treated** and **not treated**:

   How to estimate a causal effect?

2. Is it possible to **estimate causal effects** from **observational** data alone?

   And if yes, under **which conditions**, and **how**?

Randomized Controlled Trials

- We look for the estimation of the **individual causal effect** (for individual or unit *i*)

$$\tau_i = Y_i^{a=1} - Y_i^{a=0} \qquad \text{(ITE)}$$

- Since this is impossible, we look for the **average causal** (or treatment) **effect** (ATE)

$$\tau = \mathbb{E}[Y_i^{a=1}] - \mathbb{E}[Y_i^{a=0}] \qquad \text{(ATE)}$$

$$= \mathbb{E}[Y^{a=1} - Y^{a=0}] \qquad \text{(by linearity of the expectation over the mean)}$$

Randomized Controlled Trials

- We look for the estimation of the **individual causal effect** (for individual or unit *i*

$$\tau_i = Y_i^{a=1} - Y_i^{a=0}$$

- Since this is impossible, we look for the **average causal** (or treatment) **effect**

$$\tau = \mathbb{E}[Y_i^{a=1}] - \mathbb{E}[Y_i^{a=0}]$$
$$= \mathbb{E}[Y^{a=1} - Y^{a=0}] \qquad \text{(by linearity of the expectation over the mean)}$$

i	T	Y	$Y(1)$	$Y(0)$	$Y(1) - Y(0)$
1	0	0	?	0	?
2	1	1	1	?	?
3	1	0	0	?	?
4	0	0	?	0	?
5	0	1	?	1	?
6	1	1	1	?	?

But how can we do that?

A special missing value problem

i	T	Y	$Y(1)$	$Y(0)$	$Y(1) - Y(0)$
1	0	0	?	0	?
2	1	1	1	?	?
3	1	0	0	?	?
4	0	0	?	0	?
5	0	1	?	1	?
6	1	1	1	?	?
			2/3	1/3	1/3

Why can't we just compute

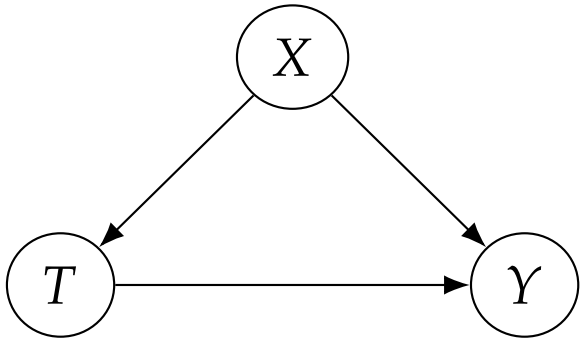
$$\mathbb{E}[Y^{a=1}] - \mathbb{E}[Y^{a=0}] = \mathbb{E}[Y|a = 1] - \mathbb{E}[Y|a = 0]$$

i	T	Y	$Y(1)$	$Y(0)$	$Y(1) - Y(0)$
1	0	0	?	0	?
2	1	1	1	?	?
3	1	0	0	?	?
4	0	0	?	0	?
5	0	1	?	1	?
6	1	1	1	?	?

Why can't we just compute 2/3 1/3 1/3

$$\mathbb{E}[Y^{a=1}] - \mathbb{E}[Y^{a=0}] = \mathbb{E}[Y|a=1] - \mathbb{E}[Y|a=0]$$

What if



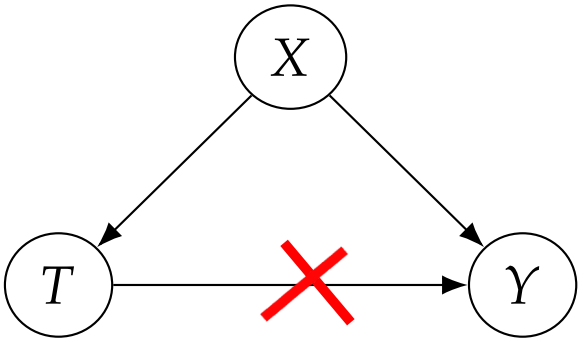
i	T	Y	$Y(1)$	$Y(0)$	$Y(1) - Y(0)$
1	0	0	?	0	?
2	1	1	1	?	?
3	1	0	0	?	?
4	0	0	?	0	?
5	0	1	?	1	?
6	1	1	1	?	?

Why can't we just compute 2/3 1/3 1/3

$$\mathbb{E}[Y^{a=1}] - \mathbb{E}[Y^{a=0}] = \mathbb{E}[Y|a=1] - \mathbb{E}[Y|a=0]$$

What if

X explains both *T* and *Y* ?



In general

$$\mathbb{E}[Y^{a=1}] - \mathbb{E}[Y^{a=0}] \neq \mathbb{E}[Y|a = 1] - \mathbb{E}[Y|a = 0]$$

- Causality is NOT simple associations

- We need to **prevent** the effects of **confounders**

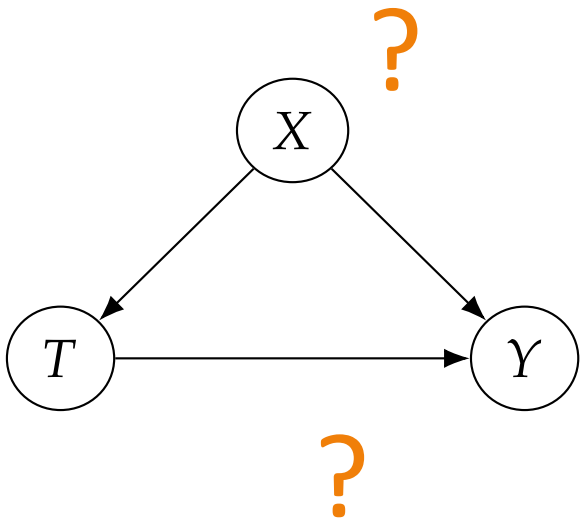
In general

$$\mathbb{E}[Y^{a=1}] - \mathbb{E}[Y^{a=0}] \neq \mathbb{E}[Y|a = 1] - \mathbb{E}[Y|a = 0]$$

- Causality is NOT simple associations
- We need to **prevent** the effects of **confounders**

Remove the effect of X on T

Randomly assigns the individuals to the treatment



The treatment groups are the same in all **relevant aspects** other than the treatment

One central question is

How to circumvent **confounding**?

if possible

- So that **causal effects** can be computed from **observational** data (conditional probabilities) alone

One central question is

How to identify **confounders**?

A **second route** towards causality

A foundational motivation

Classical data science

… is associational

- From data, we seek $P(y|x)$ or a relationship $y = h(x)$
  - *How likely is a customer who bought toothpaste to also buy dental floss?*
  - Collect data about the customers and what they bought

But …

- What if the question is:

  *What will happen to the floss sales **if we double** the price of toothpaste?*

  – Calls for a **new kind of knowledge** which is absent from the data

But …

- What if the question is:

Intervention

*What will happen to the floss sales **if we double** the price of toothpaste?*

  - Calls for a **new kind of knowledge** which is absent from the data

But …

- What if the question is:

Intervention

*What will happen to the floss sales **if we double** the price of toothpaste?*

  - Calls for a **new kind of knowledge** which is absent from the data

Why?

The science of interventions

- Even if in the past the price of toothpaste was double than the price now, **the context may have been different**
  - E.g. because of an insufficient supply, the price was doubled **everywhere**

    While now, we want to know what would happen if we double the price **only in our store**

The science of interventions

- One question is:

  Under which conditions can only **passive observations**

  allow to answer questions about **interventions**?

- We would like to have **a formal framework** for encoding and reasoning with cause and effect relationships

  - Semi-formal studies have been proposed for centuries
  - This is **only during the last three decades**, that a formal and deep understanding has begun to emerge

Causation is **not** reducible to probabilities

- Philosophers: *X **causes** Y if X **raises** the **probability** of Y*
  - E.g. Reckless driving causes accidents

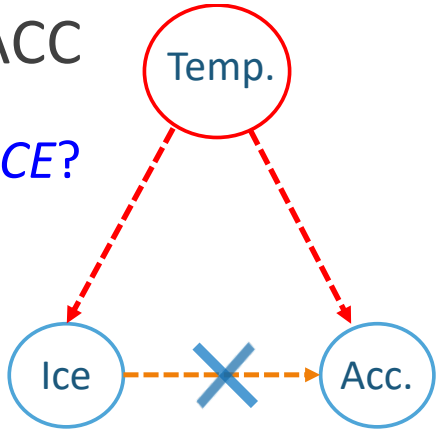
- Why is this misleading?
  - *X raises the probability of Y*

    does **not** equal to $P(Y|X) > P(Y)$ (which means: if we see *X*, then the probability of *Y* increases)

    But this increase may come about for **other reasons**!
    - *Y* being the cause of *X*
    - Another factor *Z* being the cause of both *X* and *Y*
    - …

Intervention is not observation

- We want to **estimate the causal effect** of ICE on ACC
  - What the change would be in *ACC* if we **intervened** on *ICE*?
    - P(*ACC* = *acc* | do(*ICE*) = *ice*)
  - From the causal graph:



$$P(ACC = acc|do(ICE = ice)) = \sum_{tmp} P(ACC = acc|ICE = ice, TMP = tmp)\, P(TMP = tmp)$$

  - More generally

$$P\big(Y = y|do(X = x)\big) = \sum_{z} P(Y = y|X = x, Pa = z)\, P(Pa = z)$$

  - So that:
    - P(ACC = acc | do(ICE) = ice) ≠ P(ACC = acc | ICE = ice)

« While probabilities encode our beliefs about a static world, **causality** tells us whether and how probabilities change when the world changes, be it by **intervention** or by act of **imagination** »

[PEARL, Judea & MACKENZIE, Dana. **The book of why: the new science of cause and effect**. Basic books, 2018.], p.51

Pearl's causal hierarchy

Ring	Layer	Typical **activity**	Typical **question**	Example
1	**Associational** $P(y \mid x)$	Seeing	What is? How would seeing X **change my belief in** Y?	*What does a symptom tell us about the disease?*
2	**Interventional** $P(y \mid do(x),c)$	Doing	What if? What if **I do** X?	*What if I take aspirin, will my headache be cured?*
3	**Counterfactual** $P(y_x \mid x', y')$	Imagining	Why? What if I **had acted differently**?	*What if I had not taken this pill?*

Pearl's causal hierarchy

Ring	Layer	Typical **activity**	Typical **question**	Example
1	**Associational** $P(y\|x)$	**Seeing** Supervised /unsup. learning	What is? How would seeing X **change my belief in** Y?	*What does a symptom tell us about the disease?*
2	**Interventional** $P(y\|do(x),c)$	**Doing** Reinforcement learning	What if? What if **I do** X?	*What if I take aspirin, will my headache be cured?*
3	**Counterfactual** $P(y_x\|x', y')$	**Imagining**	Why? What if I **had acted differently**?	*What if I had not taken this pill?*

How can **machines** acquire **causal knowledge**?

[PEARL, Judea & MACKENZIE, Dana. **The book of why: the new science of cause and effect**. Basic books, 2018.], p.37

How can **machines** acquire **causal knowledge**?

Traditional AI: **Representation** first, **acquisition** second

[PEARL, Judea & MACKENZIE, Dana. **The book of why: the new science of cause and effect**. Basic books, 2018.], p.38

Three ways towards causality

1. Randomized Control Trials

2. Potential outcomes framework

   – Donald B. Rubin. '**Estimating causal effects of treatments in randomized and nonrandomized studies**.' In: *Journal of educational Psychology* 66.5 (1974)

3. Causal graphs and Pearl's approach

   – Judea Pearl. '**Causal inference in statistics: An overview'**. In: *Statist. Surv.* 3 (2009), pp. 96–146

   – Judea Pearl. ***Causality***. Cambridge University Press, 2009

Causal diagrams

and Structural Causal Models (SCM)

Causal diagrams

- An arrow from *X* to *Y* implies that a rule states how *Y* would change if *X* were to change
  - Very often the **structure** of the diagram itself enables us to estimate all sorts of causal and counterfactual relationships
  - Sometimes, we have to consider **probabilities**

Causal diagrams

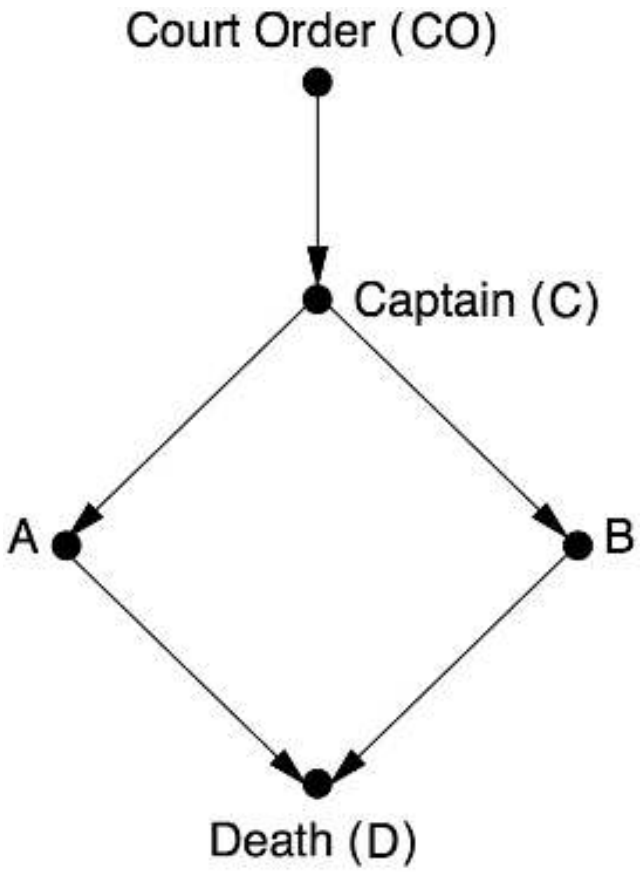
CO : court order

C : captain of the firing squad

A : first soldier

B : second soldier

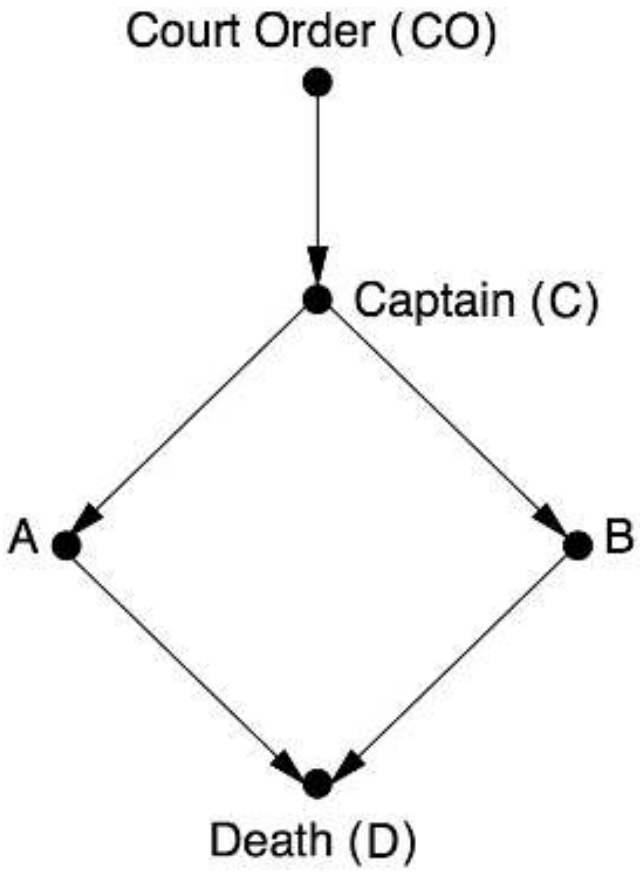
D : state of the prisoner
(True = dead, False = alive)



Causal diagrams

Question (associational) that we can ask:

If the prisoner is dead, does that mean the court order was given?



The soldiers wouldn't have fired without the captain's order

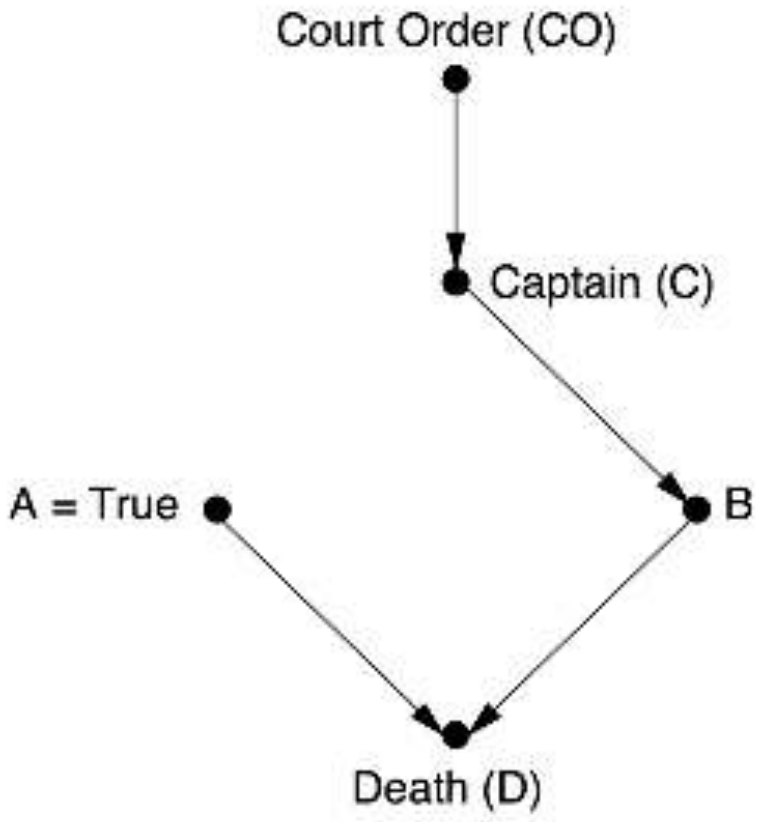
The captain would not have given the command without the order in his possession

➔ Answer = **yes**

Causal diagrams

Question (intervention) that we can ask:

What if Soldier A decides on his own initiative to fire, without waiting for the captain's command?



Remove the arrows that point to the intervention node.

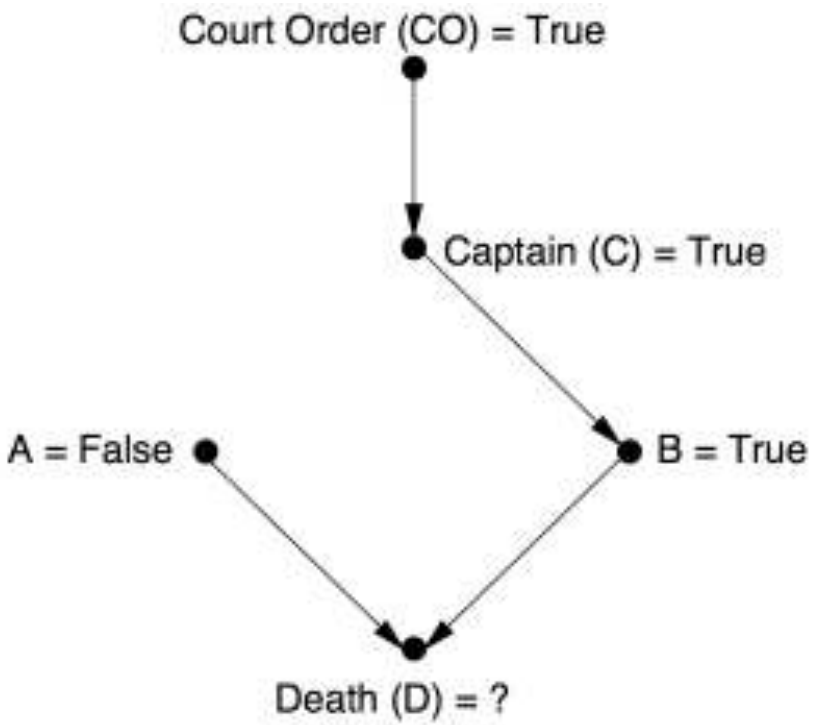
The prisoner will be shot

➔ Answer **Dead = True**

Causal diagrams

Question (counterfactual) that we can ask:

What if Soldier A had decided not to shoot?



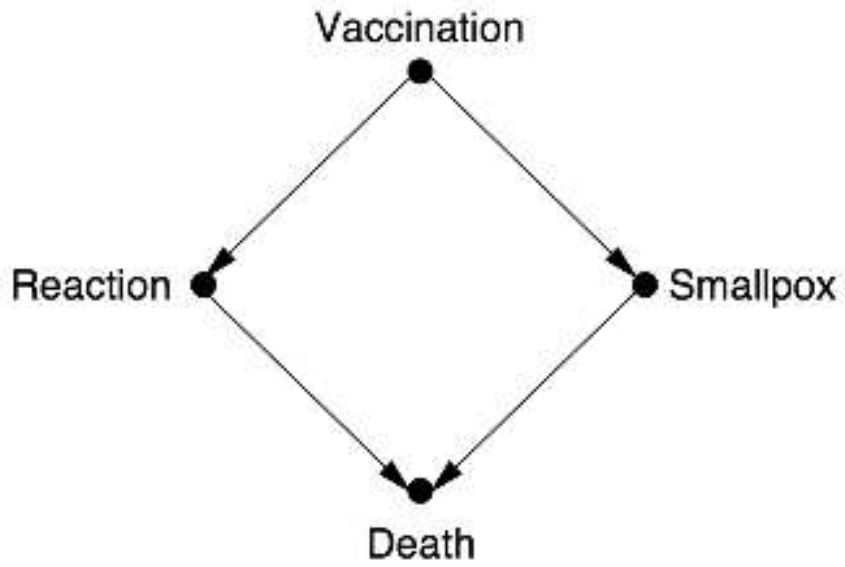
Remove the arrows that point to the intervention node.

Soldier B would have obeyed the order and shoot

➔ Answer **Dead = True**

Causal diagrams

- Is vaccination beneficial or harmful?



Statistical independencies and graphs

Methods to eliminate the counterfactual variables

1. Randomized Controlled Trials (RCT)
2. The « back-door » criterion
3. The « front-door » criterion
4. The « do-calculus »
5. The Instrumental Variables (IV)

The “back-door” criterion

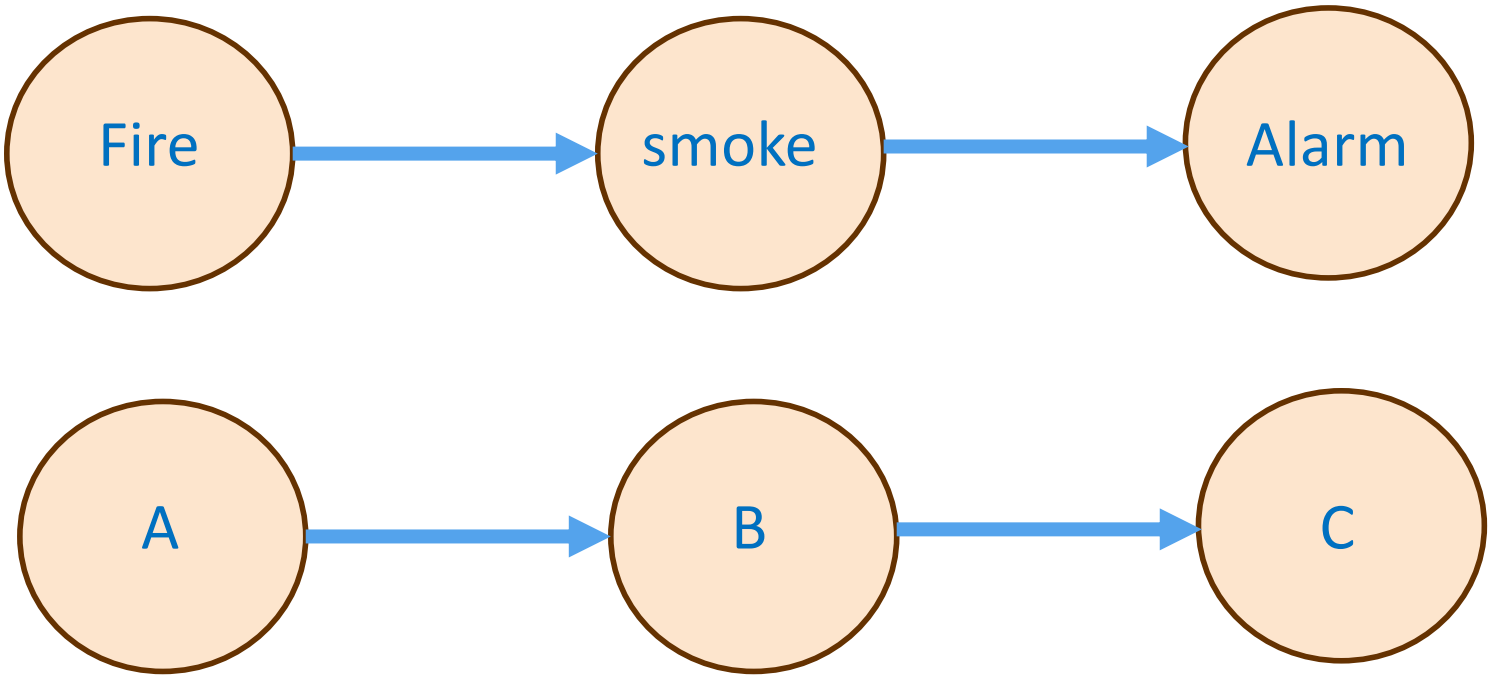
- The non causal paths are the **sources of confounding** (anything that make $P(Y \mid do(X)) \neq P(Y \mid X)$)

- The **do-operator** erases all the arrows that come into X
  - it prevents any information about X from flowing in the non causal direction.
  - Randomization has the same effect.

Causal graphs

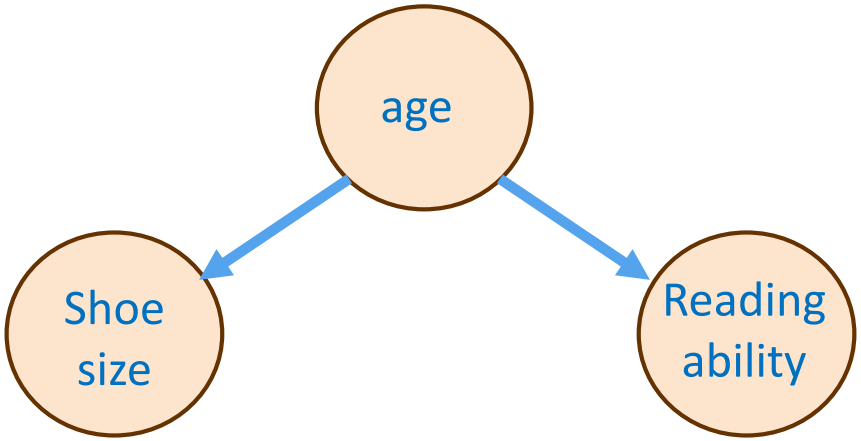
- Chains
- Forks
- Colliders
- d-separation
- The back-door criterion (causal inference)

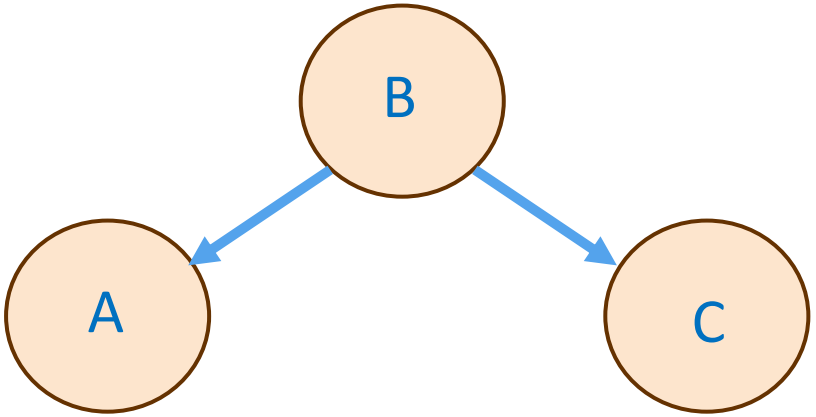
Chains



- A and C are **independent** in the graph **given** B: $A \perp\!\!\!\perp_G C \mid B$

Forks

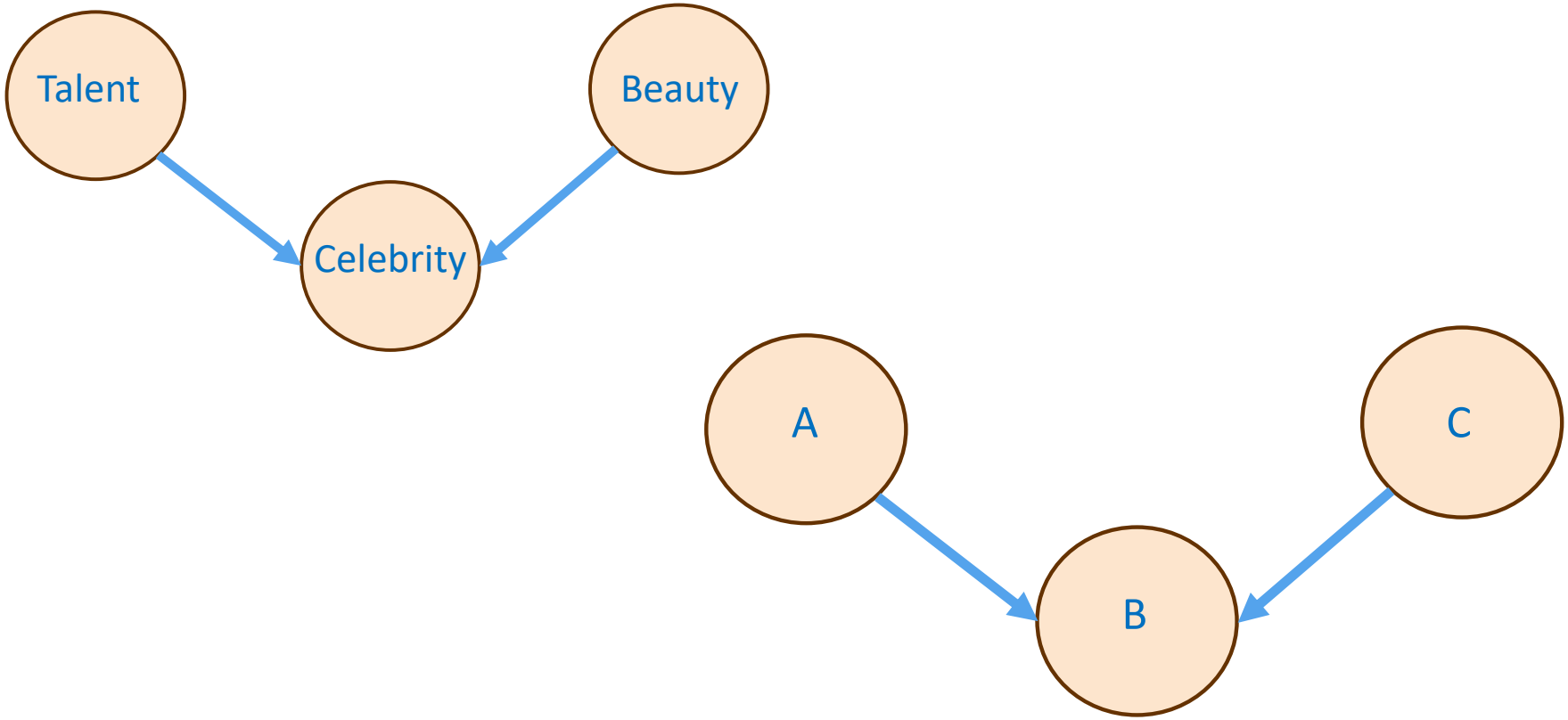




- A and C are **dependent** when we do **not control** B (confounding factor)

$$A \not\perp\!\!\!\perp_G C \mid B$$

Colliders



- A and C are **dependent** in the graph **given** B: $A \not\perp C \mid B$

The “back-door” criterion

(a) In a “chain junction” A -> B -> C, **controlling** for B **prevents** information about A from getting to C or vice versa

(b) In a “fork” A <- B -> C, **controlling** for B **prevents** information about A from getting to C or vice versa

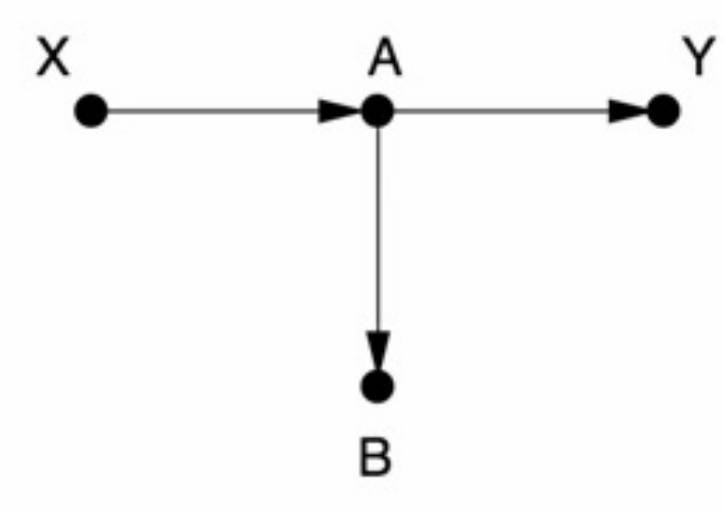
(c) In a “collider” A -> B <- C, A and C are independent, but if you **control** for B then **information flows** between A and C

(d) **Controlling for descendants** of a variable is like partially controlling for the variable itself.

- Controlling for a descendant of a **mediator** partly closes the pipe;
- Controlling for a descendant of a **collider** partly opens the pipe

The “back-door” criterion

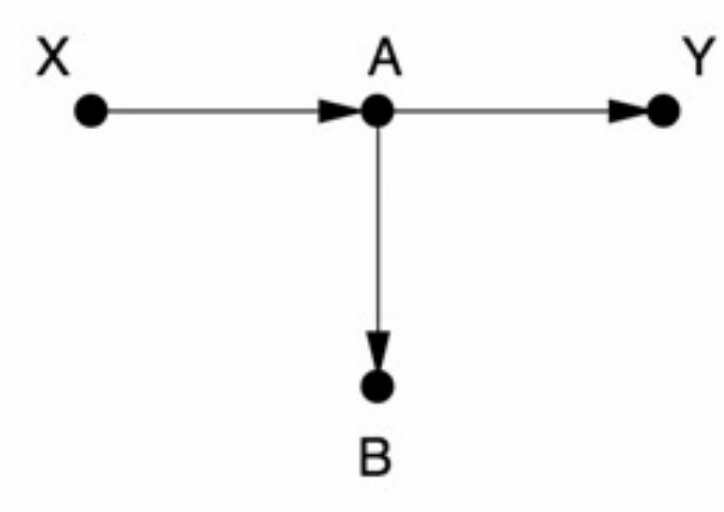
Game 1



- Identify **backdoor paths** between X and Y if any exists

The “back-door” criterion

Game 1



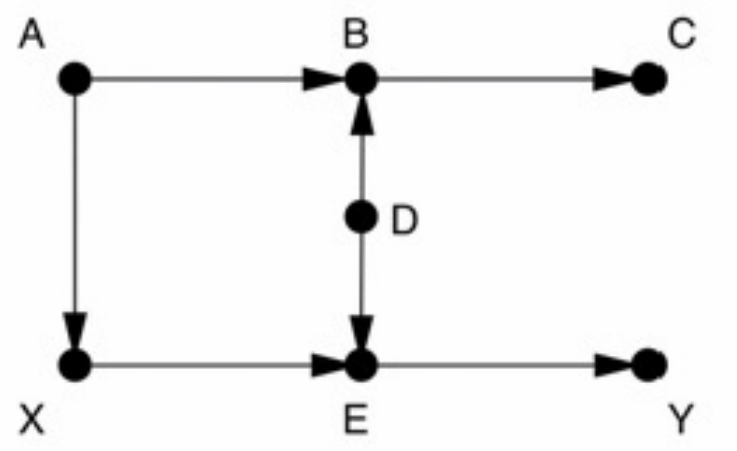
- Identify **backdoor paths** between X and Y if any exists

- No arrow leading to X, therefore **no backdoor path**, and no need to control for anything

The “back-door” criterion

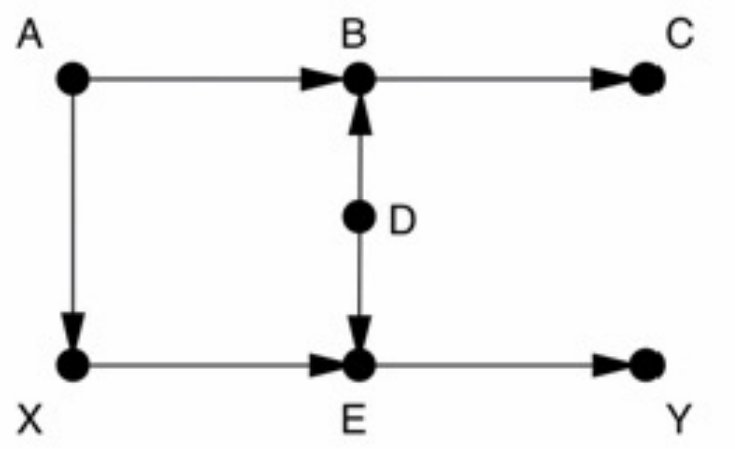
Game 2

- Identify **backdoor paths** between X and Y if any exists



The “back-door” criterion

Game 2



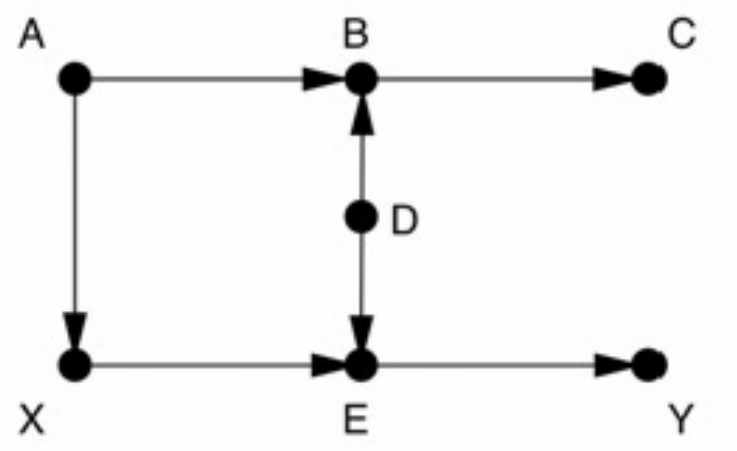
- Identify **backdoor paths** between X and Y if any exists

- A, B, C and D can be considered as pretreatment variables
- One **backdoor path**: X <- A -> B <- D -> E -> Y
- Already blocked by the collider at B: **no need to control** for anything

The “back-door” criterion

Game 3

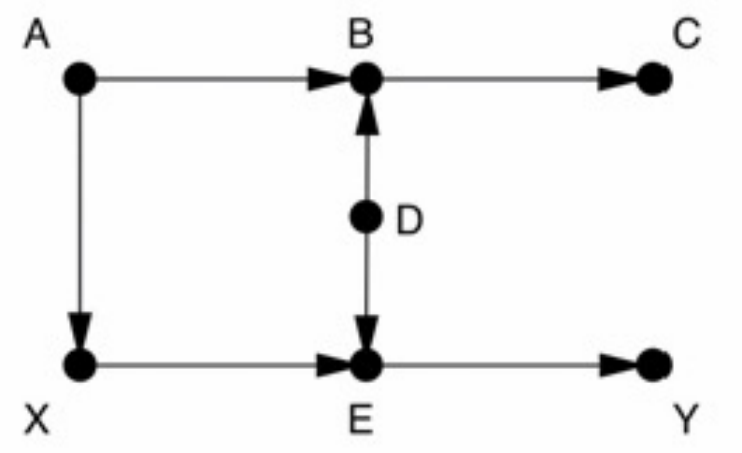
- Identify **backdoor paths** between X and Y if any exists



The “back-door” criterion

Game 3

- Identify **backdoor paths** between X and Y if any exists

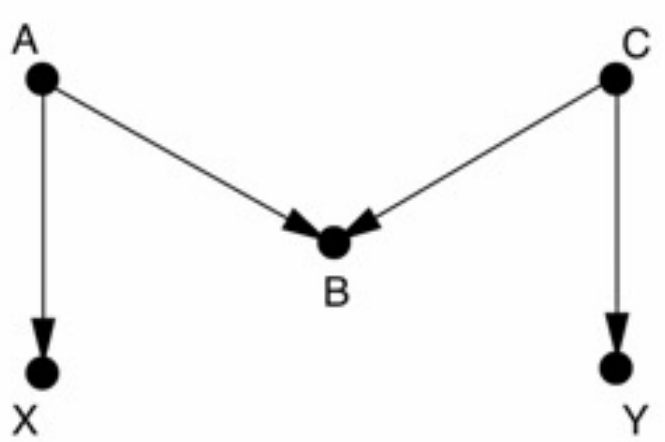


- One **backdoor path**: X <- B -> Y
- Can only be blocked by **controlling** for **B**
- If **B is not observable**, there is no way of estimating the effect of X on Y without running a randomized controlled experiment

The “back-door” criterion

Game 4

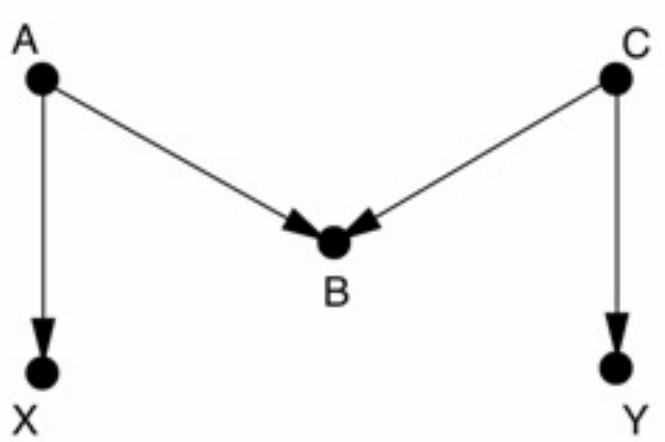
- Identify **backdoor paths** between X and Y if any exists



The “back-door” criterion

Game 4

- Identify **backdoor paths** between X and Y if any exists

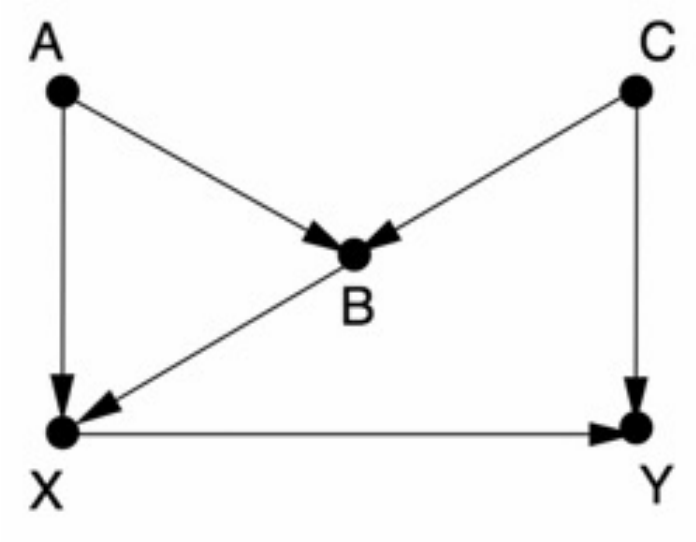


- One **backdoor path**: X <- A -> B <- C -> Y
- **Already blocked** by a collider at B
- No need to control for anything

The “back-door” criterion

Game 5

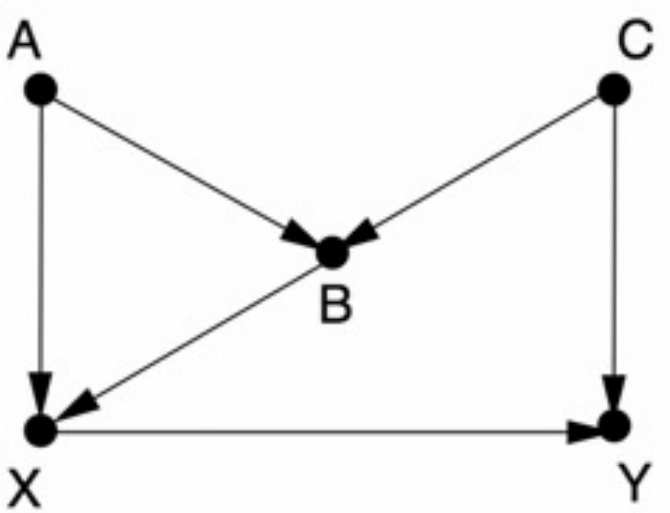
- Identify **backdoor paths** between X and Y if any exists



The “back-door” criterion

Game 5

- Identify **backdoor paths** between X and Y if any exists

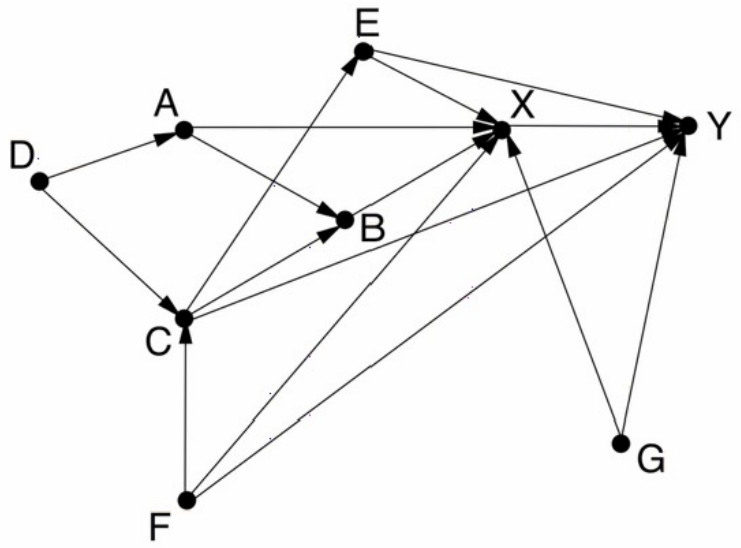


- One **backdoor path**: X <- A -> B <- C -> Y
- One **backdoor path**: X <- B <- C -> Y
- If we block the **second path** by controlling for **B**, we must now block the **1st path** by controlling for **A** or for **C** as well

The “back-door” criterion

Game 5

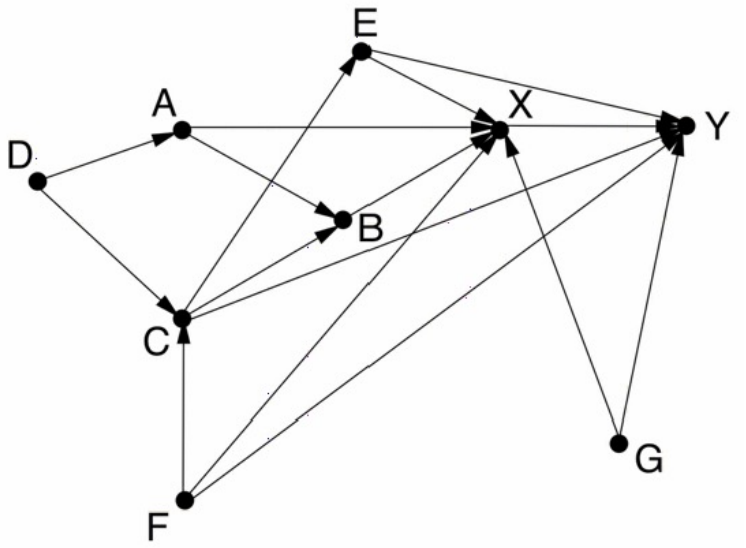
- Identify **backdoor paths** between X and Y if any exists



The “back-door” criterion

Game 6

- Game 5 is embedded in game 6



- One **backdoor path**: X <- A -> B <- C -> Y
- One **backdoor path**: X <- B <- C -> Y
- We have to control for A and B or for A and C. But C is an unobservable and therefore uncontrollable variable
- In addition, there are 4 new confounding variables: D, E, F and G
- We must control for E, F and G, but not for D

A: parental smoking
B: having asthma as a child
C: predisposition toward asthma (unobservable)
D: parental asthma
E: chronic bronchitis
F: Gender
G: socio-economic status
X: smoking behavior
Y: having asthma as an adult

Methods to eliminate the counterfactual variables

1. Randomized Controlled Trials (RCT)
2. The « back-door » criterion
3. The « front-door » criterion
4. The « do-calculus »
5. The Instrumental Variables (IV)

Not covered in this class

Impact on the

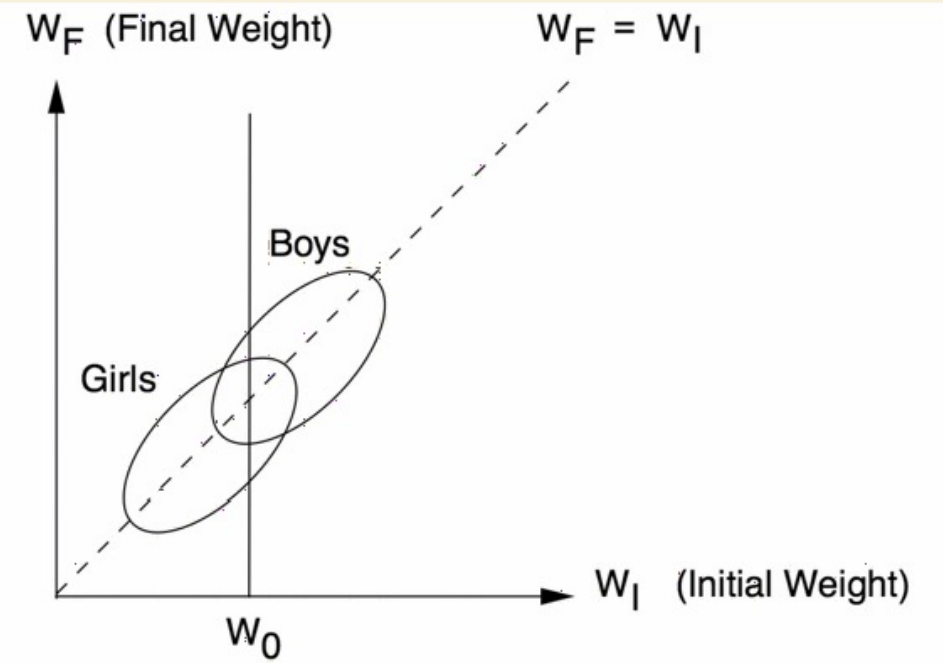
interpretation of **observational data**

Lord's paradox

[Frederic Lord, 1967]

- A school wants to **study the effects of the diet** it is providing in its dining halls and in particular whether it has different effects on **girls** and **boys**

- The students' weight is measured in September and again the following June

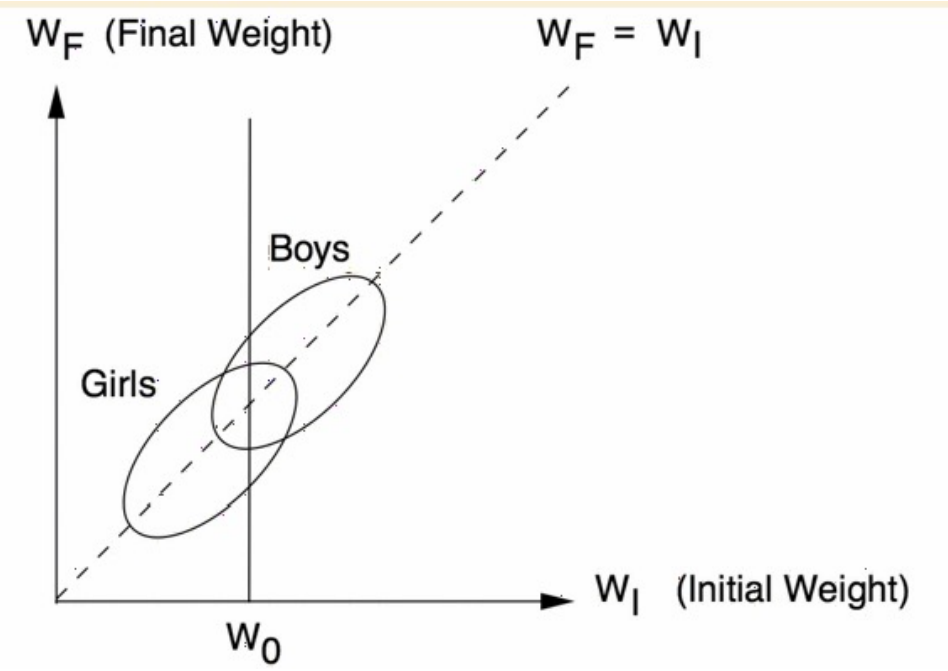


Ellipses represent scatter plots of data.

Lord's paradox

First statistician:

- The average weight of the girls is the same in June as in September (symmetry of the scatter plot around the diagonal)
- Same for the boys
- Therefore the diet has **no differential effect** on boys and girls

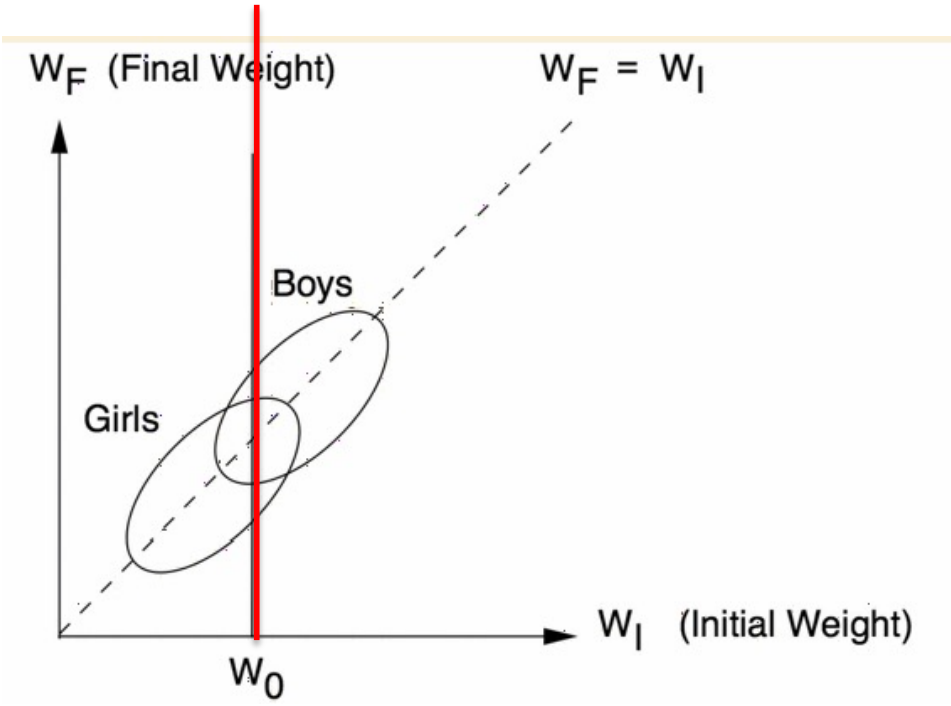


Ellipses represent scatter plots of data.

Lord's paradox

Second statistician:

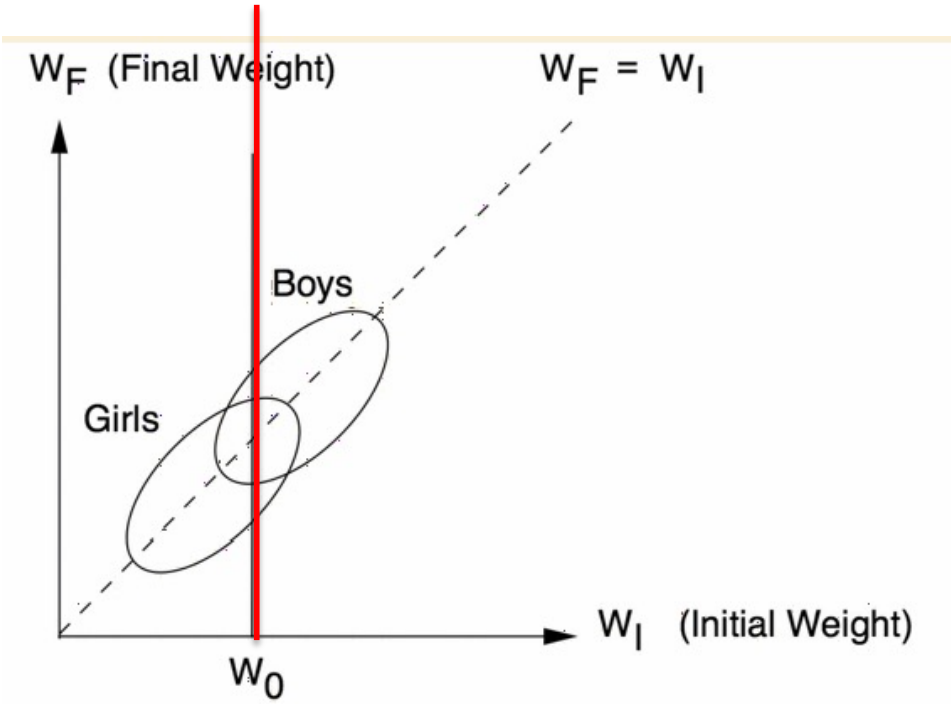
- Because the final weight of a student is strongly influenced by his/her initial weight, **we should stratify** the students **by initial weight**
- If vertical slice through both ellipses (e.g. W_0), **the vertical line intersects the boys ellipse higher up than it does for the girls ellipse**
- Therefore the diet **has a differential effect** on boys and girls



Ellipses represent scatter plots of data.

Lord's paradox

Who is right?

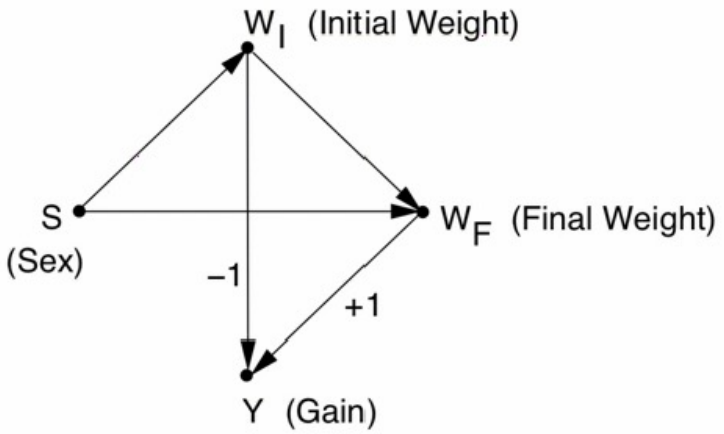


Ellipses represent scatter plots of data.

Lord's paradox

Let us look at the **underlying causal graph**

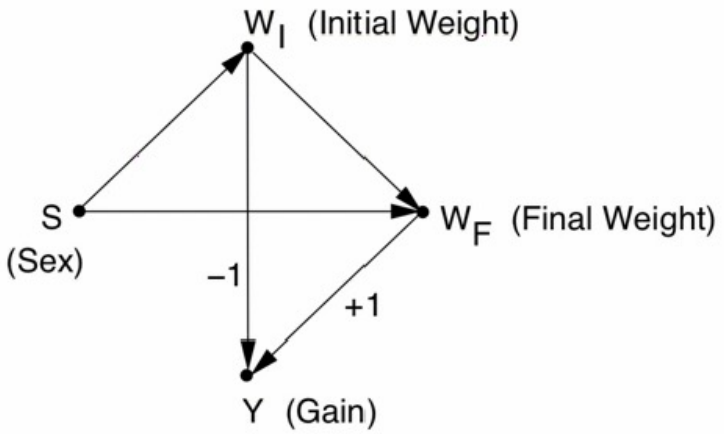
- Sex (**S**) is cause of initial weight ($\mathbf{W_i}$) and final weight ($\mathbf{W_f}$)
- $\mathbf{W_i}$ affects final weight $\mathbf{W_f}$ independently of gender (students of either gender who weight more at the beginning of the year tend to weight more at the end of the year)
- The variable of interest is the weight gain:
  $\mathbf{Y} = W_f - W_i$



Lord's paradox

Let us look at the **underlying causal graph**

- Sex (**S**) is cause of initial weight ($\mathbf{W_i}$) and final weight ($\mathbf{W_f}$)
- $\mathbf{W_i}$ affects final weight $\mathbf{W_f}$ independently of gender (students of either gender who weight more at the beginning of the year tend to weight more at the end of the year)
- The variable of interest is the weight gain:
  $\mathbf{Y} = W_f - W_i$



No backdoor path between S and Y

Therefore no need to control

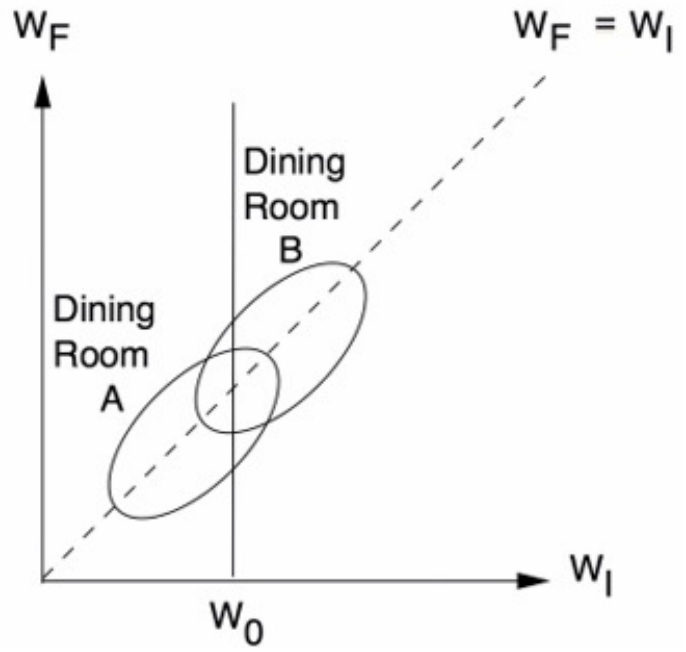
The **1st statistician is right**:

No differentiating effect on gender

Lord's paradox (2)

Now the school wants to study the effect of diet (not gender) on weight gain

- The students eat in one of two **dining halls** with **different diets**
- The students who **weigh more** in the beginning tend to eat in dining **hall** B, while the ones who **weigh less** eat in dining **hall A**

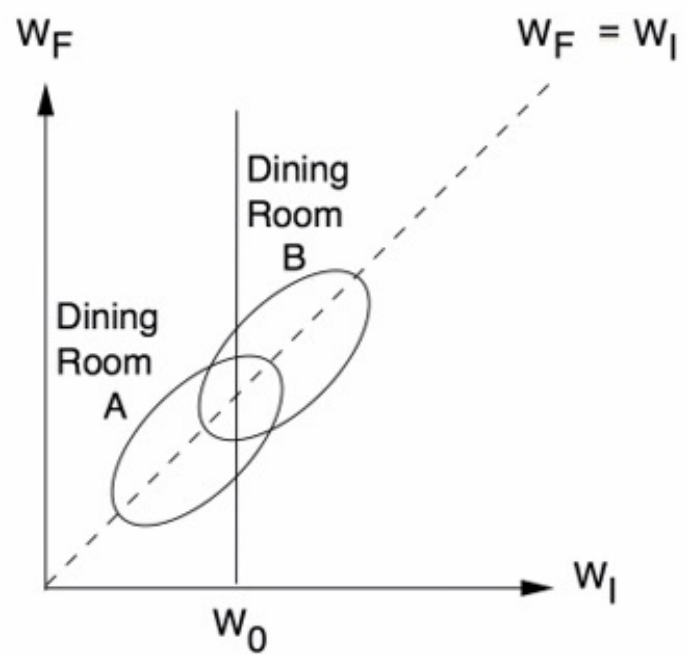


Lord's paradox (2)

First statistician:

- Based on symmetry consideration, switching from diet A to B has no effect on weight gain (the difference W_f- W_i has the same distribution in both ellipses)

- Therefore the diet has **no differential effect** on weight gain

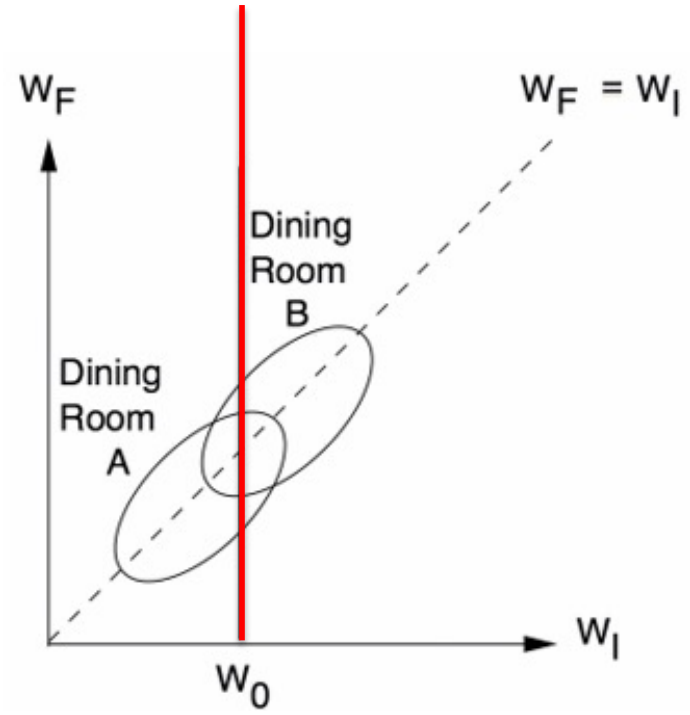


Ellipses represent scatter plots of data.

Lord's paradox (2)

Second statistician:

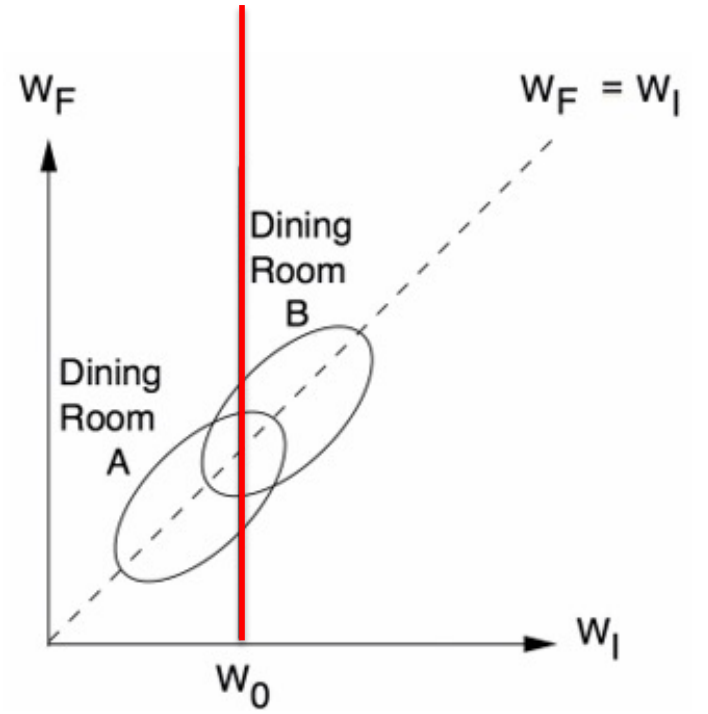
- Compares the final weights under diet A to those of diet B for a group of students starting with weight W_0
- Therefore the diet **has** a **differential effect** on weight gain



Ellipses represent scatter plots of data.

Lord's paradox (2)

Who is right?

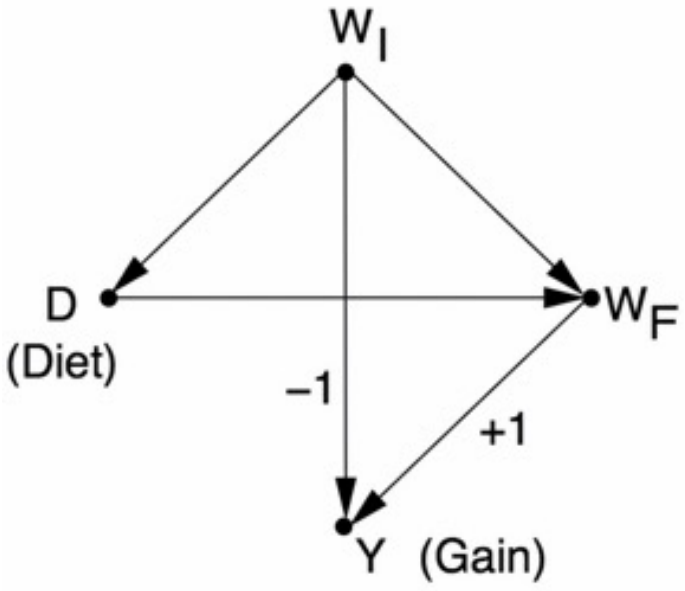


Ellipses represent scatter plots of data.

Lord's paradox (2)

Let us look at the **underlying causal graph**

- The causal variable is D (diet), not S
- The arrow that pointed from S to W*i* now reverses direction. The initial weight now affects the diet
- $\mathbf{W_i}$ is a confounder of **D** and $\mathbf{W_f}$, not a mediator



Backdoor path between D and Y

Therefore need to control $\mathbf{W_i}$

The **2nd statistician is right**:

Clear differentiating effect of diet

Conclusion

- The scatterplots by themselves are **ambiguous**
- Assumptions on the underlying causal graph is **necessary** to interpret the findings

Do-calculus

- The goal of *do* calculus is to **turn** an expression with *do* operation **into** a ***do-free*** expression involving **only observable variables** so we can come up with an estimator.

The « do-calculus »

Règle 1

- Quand on observe une variable *W* non pertinente pour *Y* (possiblement conditionnée sur une autre variable *Z*), alors la distribution de probabilités de *Y* est inchangée

$$P(Y \mid do(X), Z, W) = P(Y \mid do(X), Z)$$

- Du moment que la variable *Z* bloque tous les chemins de *W* à *Y* après que les flèches pointant vers *X* aient été éliminées
- Exemple :
  - *W* = Feu -> *Z* = fumée -> *Y* = alarme
  - *Z* bloque tous les chemins de *W* à *Y*

Le « do-calculus »

Règle 2

- Si un ensemble de variables Z bloque tous les chemins back-door de X à Y, alors $do(X)$ est équivalent à $see(X)$ conditionné sur Z

$$P(Y \mid do(X), Z) = P(Y \mid X, Z)$$

- Du moment que la variable Z satisfait la condition back-door

The « do-calculus »

Rule 3

- On peut retirer $do(X)$ de $P(Y \mid do(X))$ s'il n'y a pas de chemins causaux de X à Y

$$P(Y \mid do(X)) = P(Y)$$

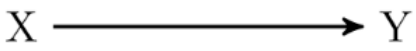
- Du moment qu'il n'y a pas de chemin de X à Y avec seulement des flèches dirigées vers l'avant.

The do-calculus

Rule 1 (Insertion/deletion of observations)

$$P(y|\text{do}(x), z, w) = P(y|\text{do}(x), w) \text{ if } Y \perp\!\!\!\perp Z|X, W \text{ in } G_{\overline{X}}$$

In words, this tells us that we can remove a variable z from our expression if z is independent of y, given x and potentially other variables w, in the DAG in which we remove all arrows going into x.



(a) A very simple DAG

are no arrows connecting X and Y in $G_{\underline{X}}$, that is, if we remove all the arrows going out of X. Hence, X and Y are independent in $G_{\underline{X}}$, which means that we can apply rule 2 of *do*-calculus:

$$P(y|\text{do}(x)) = P(y|x)$$

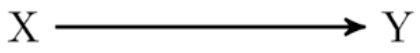
...

The do-calculus

Rule 2 (Action/observation exchange)

$$P(y|\text{do}(x), \text{do}(z), w) = P(y|\text{do}(x), z, w) \text{ if } Y \perp\!\!\!\perp Z|X, W \text{ in } G_{\overline{X},\underline{Z}}$$

In word, this tells us that we can replace the action $do(z)$ with the variable z observed in the data if y and z are independent, given x and potentially other variables w, in the DAG in which we remove the arrow going into x and out of z.



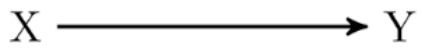
(a) A very simple DAG

are no arrows connecting X and Y in $G_{\underline{X}}$, that is, if we remove all the arrows going out of X. Hence, X and Y are independent in $G_{\underline{X}}$, which means that we can apply rule 2 of *do*-calculus:

$$P(y|\text{do}(x)) = P(y|x)$$

...

The do-calculus



(a) A very simple DAG

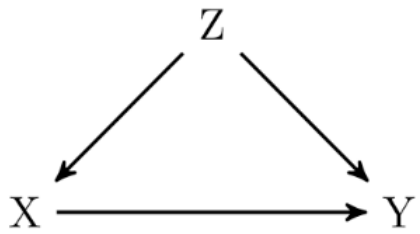
are no arrows connecting X and Y in $G_{\underline{X}}$, that is, if we remove all the arrows going out of X. Hence, X and Y are independent in $G_{\underline{X}}$, which means that we can apply rule 2 of *do*-calculus:

$$P(y|\text{do}(x)) = P(y|x)$$

Success! Using *do*-calculus we could replace all the *do*-statements with observed variables, which now allows us to estimate the causal effect of changing X on Y based on our observed data. This was quite an easy example. But before we continue with the next example, let's take a closer look at $G_{\underline{X}}$ again. The reason why we are interested in looking at the graph in which we remove all arrows going out from X is that we want to make sure that X is only affecting Y directly or through causes that are caused by X, i.e., we are interested in the total effect of X on Y. Thus, if we remove all arrows going out of X or going into Y, and we find that in this submodel there is no open causal path between X and Y, we can be sure, that in the whole model G, all causal paths between X and Y must be direct paths, i.e., paths that we want to include in our estimation.

...

The do-calculus



(b) A DAG with a confounder (Z)

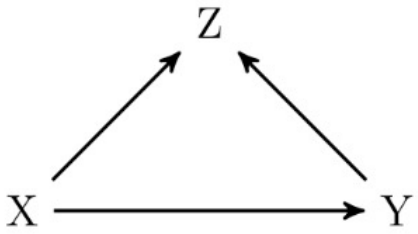
Figure 1 (b) includes a classical example of confounding, in which the variable Z confounds the effect of X on Y. If we remove all arrows going out of X, we find that X is still associated with Y through the fork $X \leftarrow Z \rightarrow Y$. Hence, we cannot directly disentangle the direct effect of X on Y and the association between X and Y that is due to the confounding of Z. However, as stated in rule 2 we can also condition on other variables to render X and Y independent in $G_{\underline{X}}$.

$$
\begin{aligned}
P(y|\text{do}(x)) &= \sum_z P(y|\text{do}(x), z)P(z) \\
&= \sum_z P(y|x, z)P(z) \qquad \text{Rule 2: } Y \perp\!\!\!\perp X|Z \text{ in } G_{\underline{X}}
\end{aligned}
$$

Ok, let's go through this in more detail. The first step we need to do is to condition our analysis on the variable Z. This renders X and Y independent in $G_{\underline{X}}$. After this, we can now replace $P(y|\text{do}(x), z)$ with $P(y|x, z)$ as X and Y are independent when conditioning on Z.

...

The do-calculus

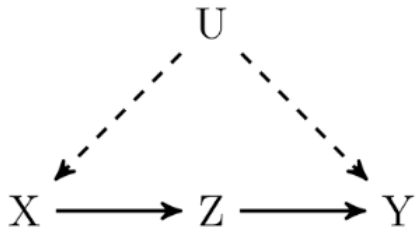


(c) A DAG with a collider (Z)

Figure 1 (c) again is a more simple example. In this graph X and Y are independent in $G_{\underline{X}}$ because Z is a collider on the path $X \rightarrow Z \leftarrow Y$. Hence, we can just calculate $P(y|\mathrm{do}(x))$ based on our observed data $P(y|x)$.

...

The do-calculus



(d) A DAG with an unmeasured confounder (U)

Figure 1 (d) is a tricky one and in contrast to the graphs before, we cannot only rely on rule 2 in order to identify the causal effect of X on Y. Using only the back-door criteria would not allow us to identify the causal effect of X on Y in this graph, but using *do*-calculus we actually can identify this effect. For this, let's first take a look at the effect that we would like to estimate:

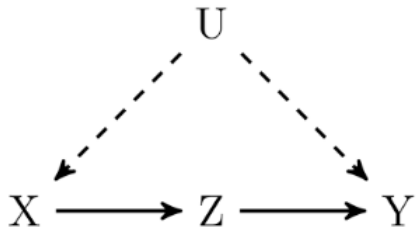
$$P(y|\text{do}(x)) = \sum_z P(y|\text{do}(x), z)P(z|x) \tag{1}$$

Unfortunately, we cannot estimate the first part of the right hand hand side directly using only observed data, but we can achieve this with the help of both rule 2 and 3.

$$\begin{aligned} P(y|\text{do}(x), z) &= P(y|\text{do}(x), \text{do}(y)) && \text{Rule 2: } Y \perp\!\!\!\perp Z \text{ in } G_{\overline{X}\underline{Z}} \\ &= P(y|\text{do}(y)) && \text{Rule 3: } Y \perp\!\!\!\perp X \text{ in } G_{\overline{XZ}} \\ &= \sum_x P(y|x, z)P(x) && \text{Rule 2: } Y \perp\!\!\!\perp Z|X \text{ in } G_{\underline{Z}} \end{aligned} \tag{2}$$

...

The do-calculus



(d) A DAG with an unmeasured confounder (U)

Now, we yielded an expression for the first part of the right hand site that only includes observed variables. Let's do the same for the second part of the right hand side in Equation 1. Translating this part of the equation to an expression, only including observed variables, is actually a lot easier, as Y is a collider on the path $X \leftarrow U \rightarrow Y \leftarrow Z$ which renders Z and Y independent in $G_{\underline{X}}$.

$$P(z|\text{do}(x)) = P(z|x) \qquad \text{Rule 2: } Z \perp\!\!\!\perp X \text{ in } G_{\underline{X}} \tag{3}$$

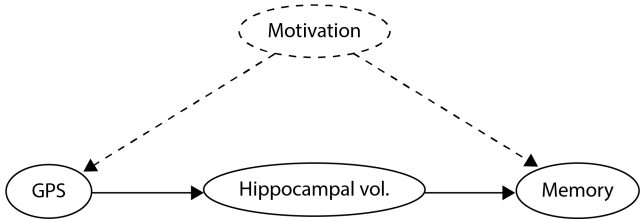
Now, we have all pieces that we need in order to translate Equation 1 into an expression only including observed variables. Let's substitute Equation 1 with Equation 2 and Equation 3:

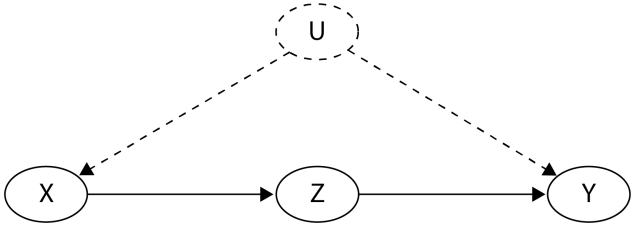
$$P(y|\text{do}(x)) = \sum_z P(z|x) \sum_{x'} P(y|x', z)P(x') \tag{4}$$

...

Do-calculus: For the **frontdoor** criterion

- Does the use of GPS diminish spatial memory?





- First, let's take a look at the **relationship between X and Z**.
  - There's a one back-door path between them, X ← U → Y ← Z, but it's already blocked. Because there's a collider, U → Y ← Z, that blocks the flow of information.

$$P(Z = z|do(X = x)) = P(Z = z|X = x)$$

- How about the **effect of Z on Y**?
  - There's one open back-door path, *Z*← *X*← *U* → *Y*. There's no collider on this path and *U* is unobserved, so we cannot control for it. Fortunately, we can control for the other variable, *X*. A valid estimand of the causal effect of *Z* on *Y* is therefore the following:

$$P(Y = y|do(Z = z)) = \sum_x P(Y = y|Z = z, X = x)\, P(X = x)$$

- Now, **combine** both estimands together:

$$P(Y = y|do(X = x)) = \sum_z P(Y = y|do(Z = z))\, P(Z = z|do(X = x)) = \sum_z P(Y|do(Z))\, P(Z|do(X))$$
$$= \sum_z P(Z|X) \sum_{x'} P(Y|X, Z)\; P(X)$$

Illustration

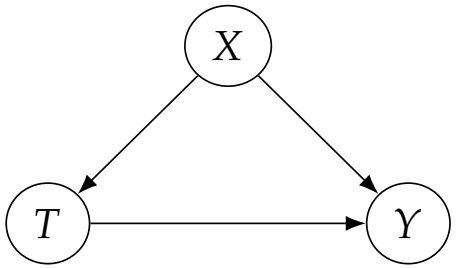


Figure 4.5: Simple causal structure where X counfounds the effect of T on Y and where X is the only confounder.

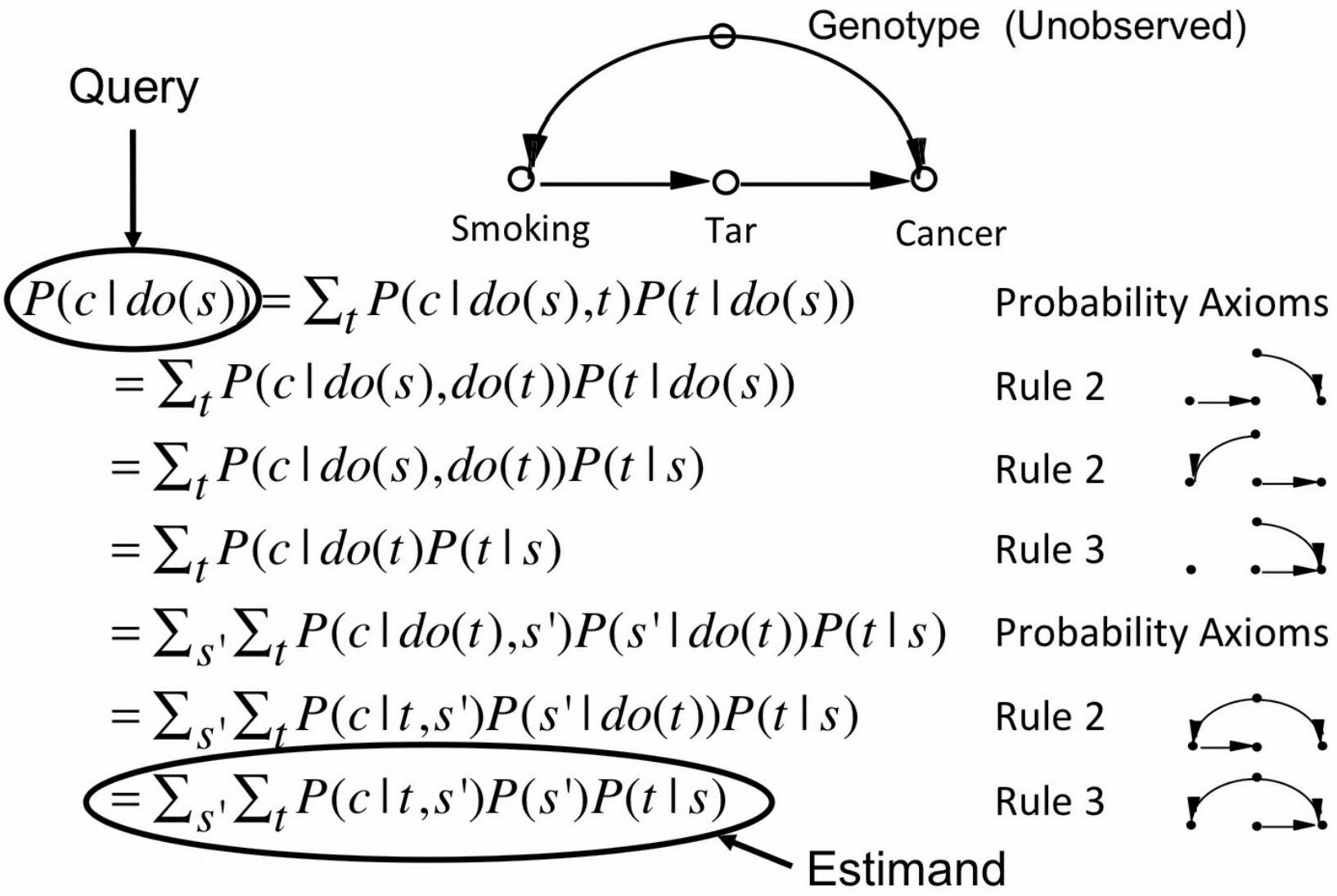
$$P(y, x \mid do(t)) = P(x)\, P(y \mid t, x)$$

$$P(y \mid do(t)) = \sum_x P(y \mid t, x)\, P(x)$$

$$\sum_x P(y \mid t, x)\, P(x \mid t) = \sum_x P(y, x \mid t)$$
$$= P(y \mid t)$$

...

DO-CALCULUS AT WORK



From [Pearl (2018), « The book of why », p.236]

Theorem

- Whenever a causal effect is **estimable from data**, a sequence of steps using the three rules of **the do-calculus can eliminate** the *do*-operator

  and thus allows **estimation** of causal effects **from observational data alone**

How to validate causal models

How to ~~validate~~ causal models

refute

Types of refutation tests

The **basic idea** behind refutation tests is to **modify** an element of either

- the **model**
- or a **dataset**

and see how it impacts the results.

- For instance, a *random common cause refuter* adds a new confounding variable to the dataset and controls for it.

  If the original model is correctly specified, we expect that such an addition will *not lead to significant changes* in the model estimates.

  (This test belongs to the *invariant transformations* category).

Types of transformations available in the DoWhy library

- **Invariant** transformations
  - Change the data in such a way that **the result should not change** the **estimate**.
  - **If** the **estimate changes** significantly, the **model fails** to pass the test.

- **Nullifying** transformations
  - Change the data in a way that it should **cause the estimated effect** to be **zero**.
  - **If** the result significantly differs from zero, the **model fails** the test.

- Experiments

Conclusions

1. Inferencing causality relationships **requires knowledge** beyond statistics

2. One approach is the **Pearl**'s one, using causal graphs and the do-calculus

3. **In some situations**, it is possible to estimate causal effects from observational data alone

4. There is a **ladder of causality**, with three levels ending in counterfactuals

(counterfactuals not covered in this class)

5. Techniques for **learning** causal relationships are still in their infancy

(learning not covered in this class)