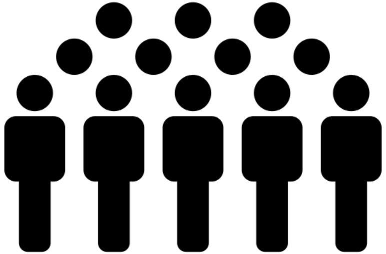
Causal Inference

Lecture 14

Xiaofei Shi

Where does bias come from?



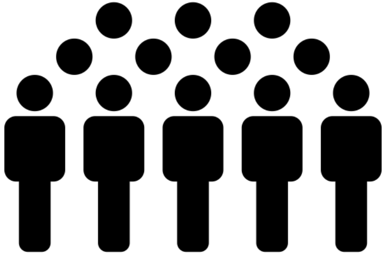
Created by Wilson Joseph
from Noun Project

Images from [The Noun Project]

Summary

- Trial data does not shield you from biased results
- Introducing graphical models

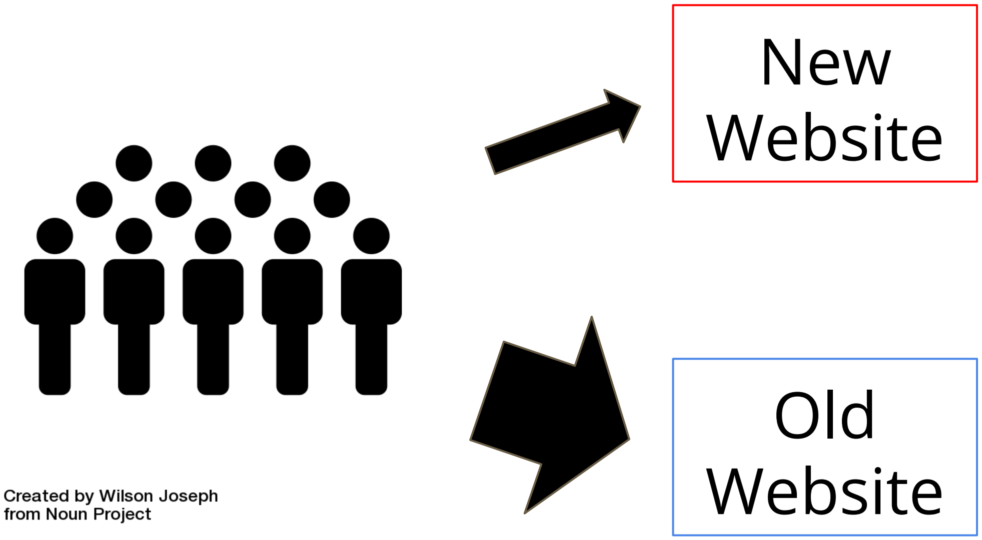
AB testing



Created by Wilson Joseph
from Noun Project

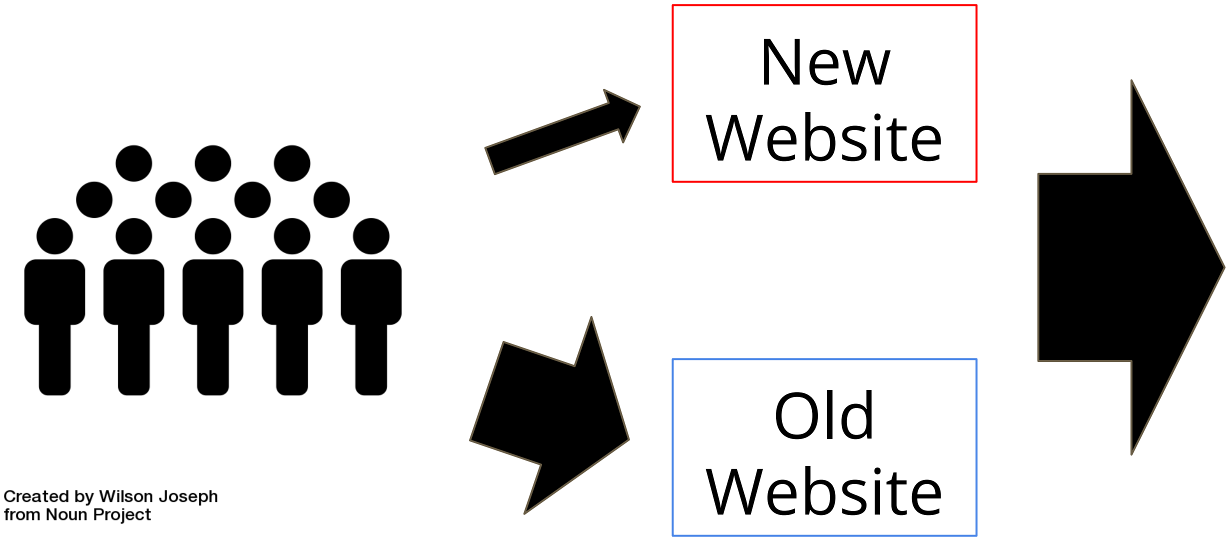
Images from [The Noun Project]

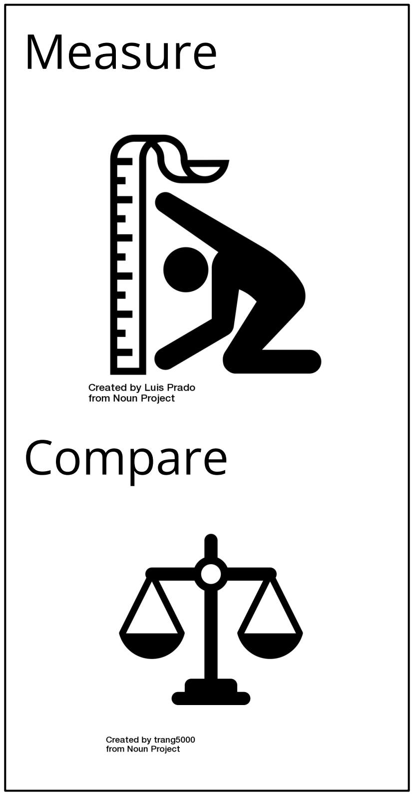
AB testing



Images from The Noun Project

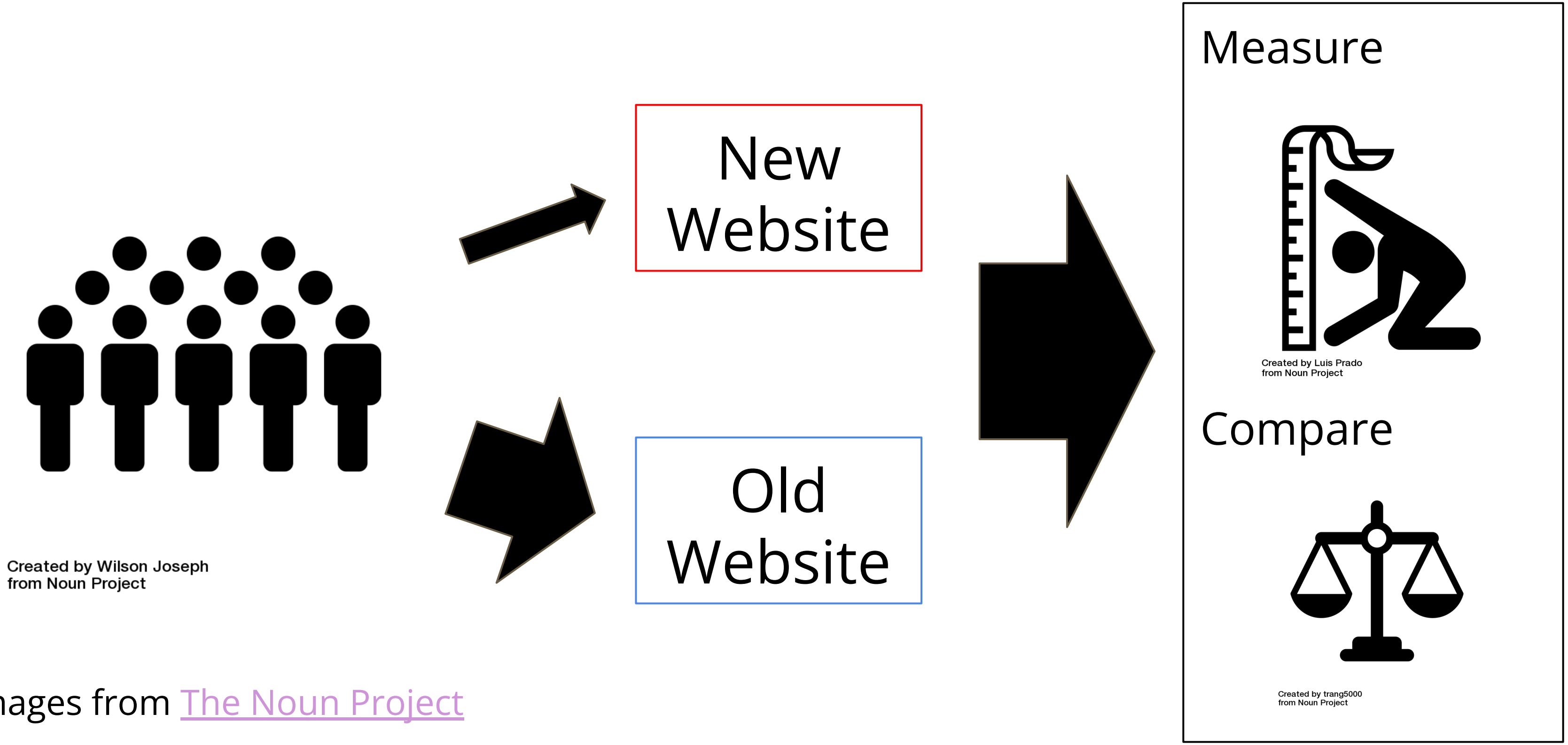
AB testing





Images from [The Noun Project]

AB testing == Randomized controlled trials?

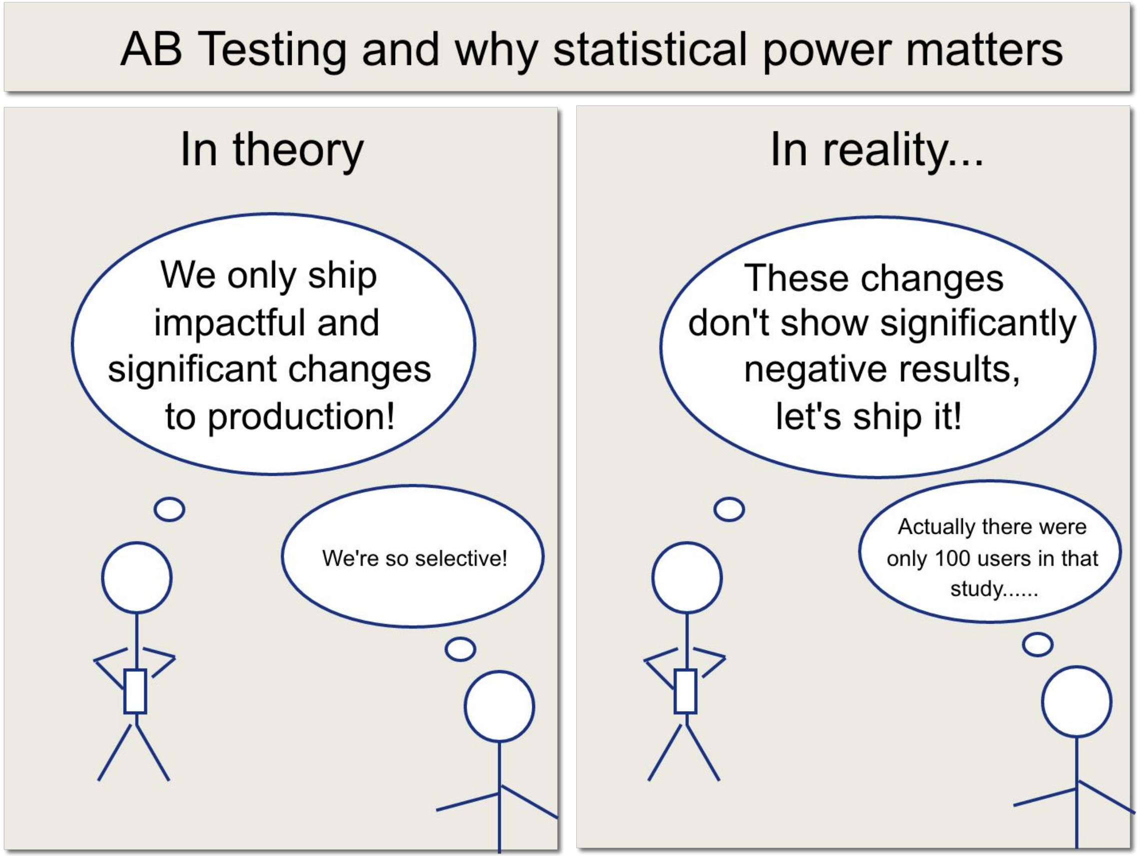


Images from [The Noun Project]

Most tests are not significant

AB Testing and why statistical power matters

Most tests are not significant



What happens when tests are not significant

Image from The Noun Project

What happens when tests are not significant

- People want to understand **why**

Image from [The Noun Project]

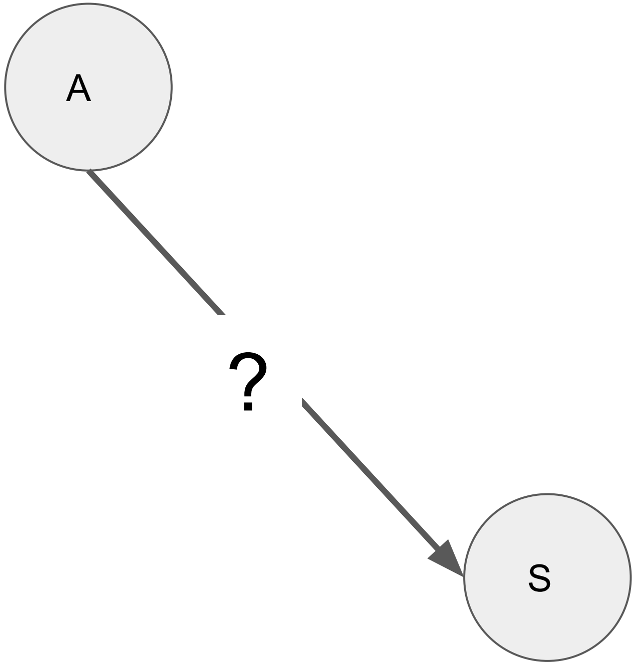
What happens when tests are not significant

Created by Adrien Coquet
from Noun Project

- People want to understand **why**
- Can data mining techniques help identify a group that would respond better to the new feature?

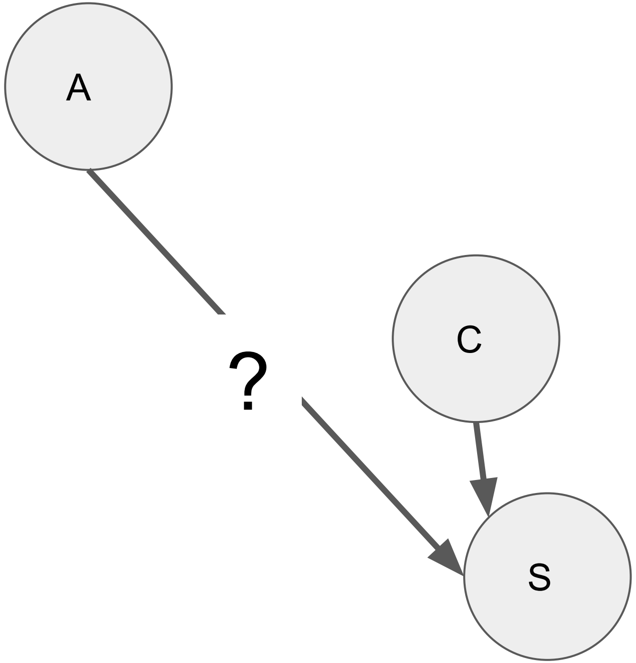
Image from The Noun Project

With 80% power, your feature had no significant impact from an AB test



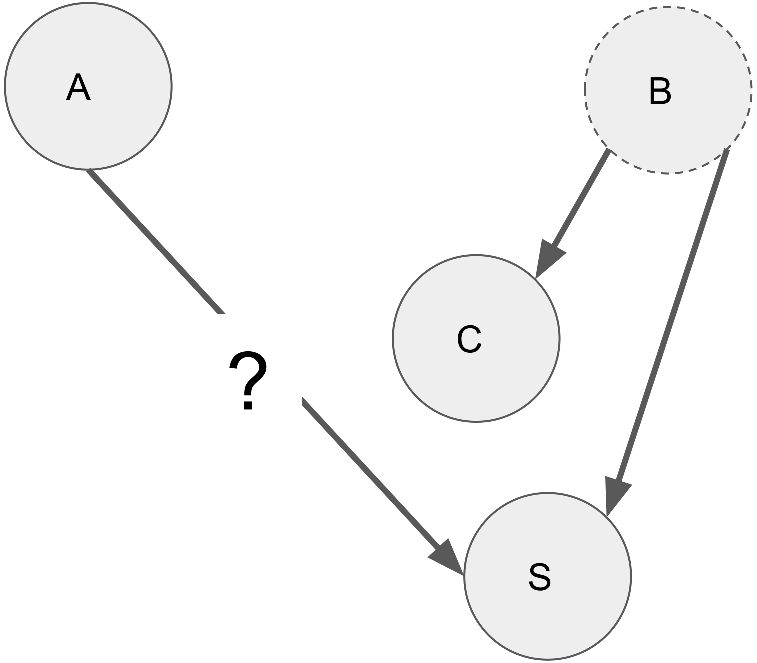
- A: exposure to an **A**dvertisement
- S: user **S**igned-up for the service

Perhaps the detectable effect was smaller than you thought, so you regress on user behavior



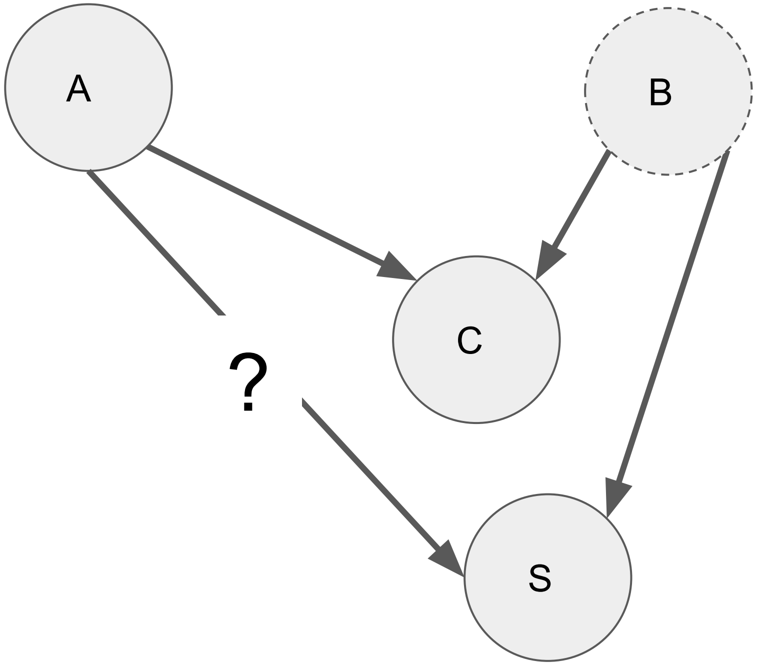
- A: exposure to an **A**dvertisement
- S: user **S**igned-up for the service
- C: user **C**licks

User behavior, however, are often proxy for unobserved user characteristics



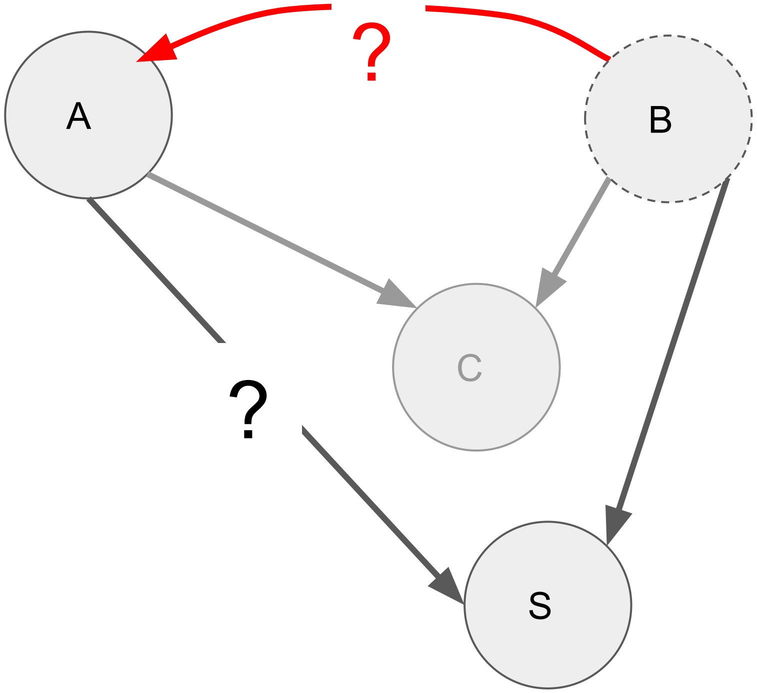
- A: exposure to an **A**dvertisement
- S: user **S**igned-up for the service
- C: user **C**licks
- B: unobserved **B**ackground

From data mining, you will likely find a feature that the ads impacts as well



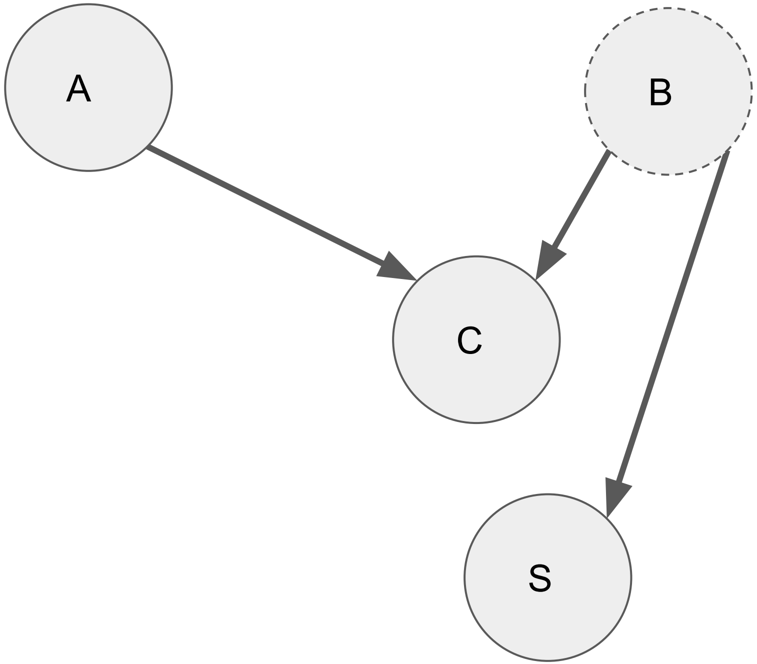
- A: exposure to an **A**dvertisement
- S: user **S**igned-up for the service
- C: user **C**licks
- B: unobserved **B**ackground

Is there any risk of different backgrounds being a confounder for our treatment?



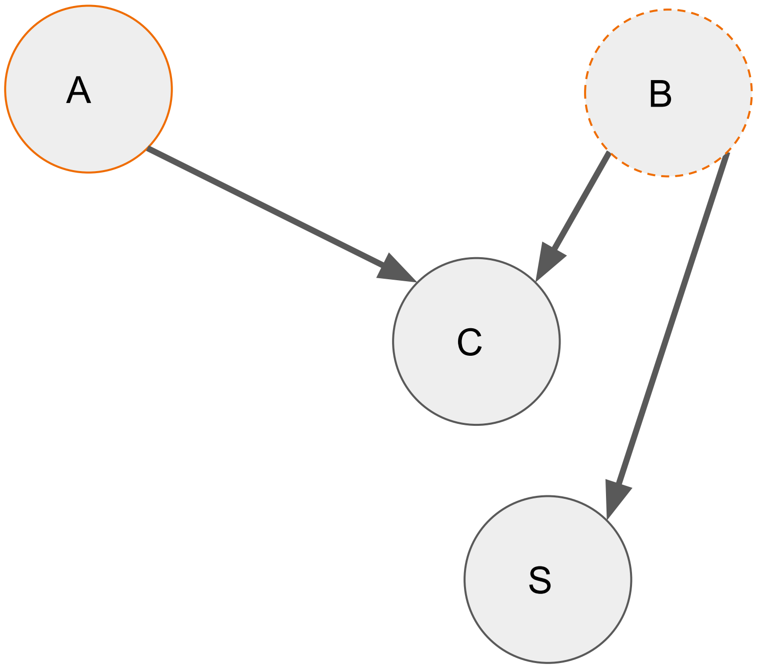
- A: exposure to an **A**dvertisement
- S: user **S**igned-up for the service
- C: user **C**licks
- B: unobserved **B**ackground

Turns out we can detect a significant effect from A on S even if A has no impact



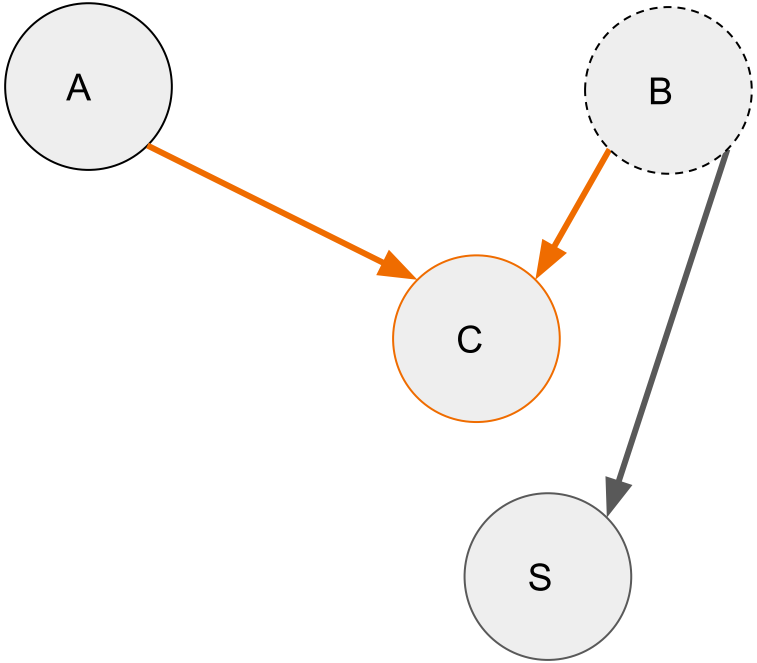
- A: exposure to an **A**dvertisement
- S: user **S**igned-up for the service
- C: user **C**licks
- B: unobserved **B**ackground

Let's simulate this!



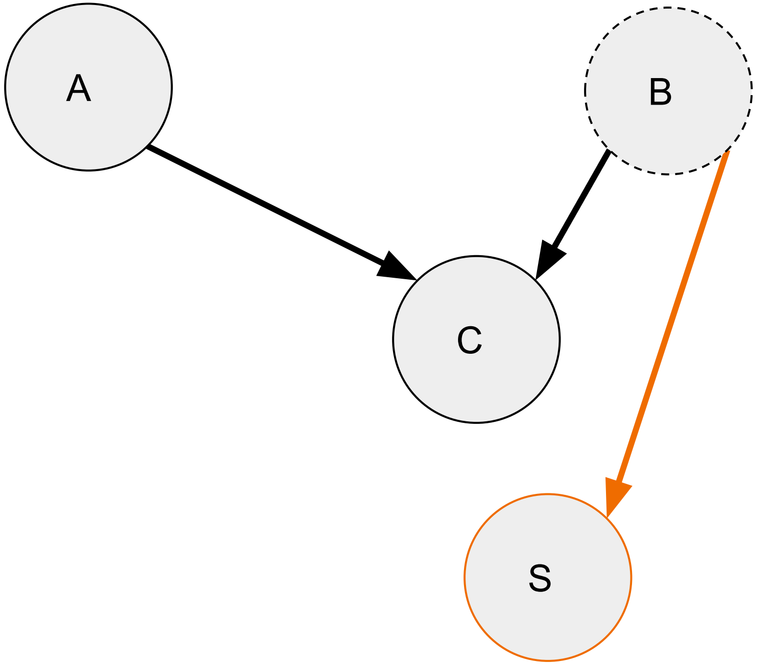

```r
n <- 1000
background <- runif(n)
ad <- rbinom(n, 1, 0.2)
```

Let's simulate this!



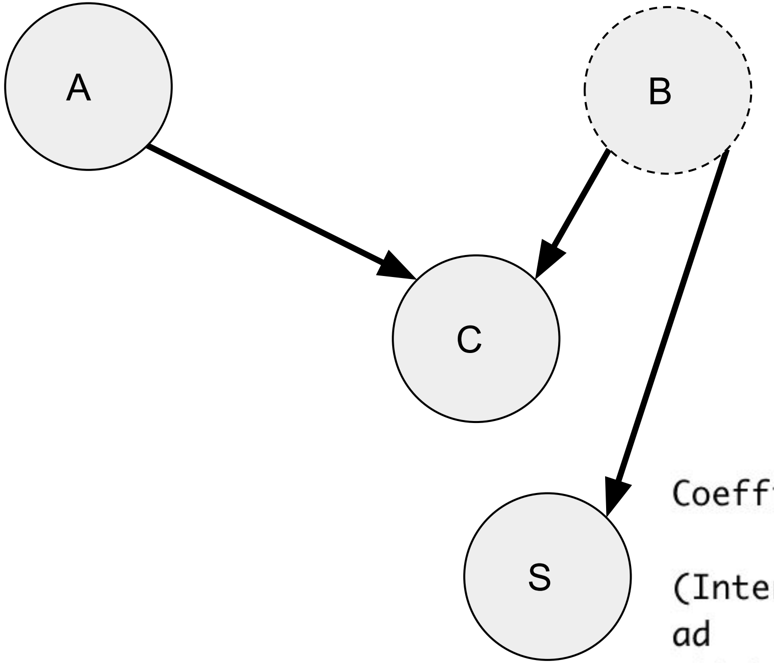

```r
n <- 1000
background <- runif(n)
ad <- rbinom(n, 1, 0.2)

clicks <- rexp(n, 0.1/(background + ad))
```

Let's simulate this!




```r
n <- 1000
background <- runif(n)
ad <- rbinom(n, 1, 0.2)

clicks <- rexp(n, 0.1/(background + ad))

signup <- rbinom(n, 1, background)
```

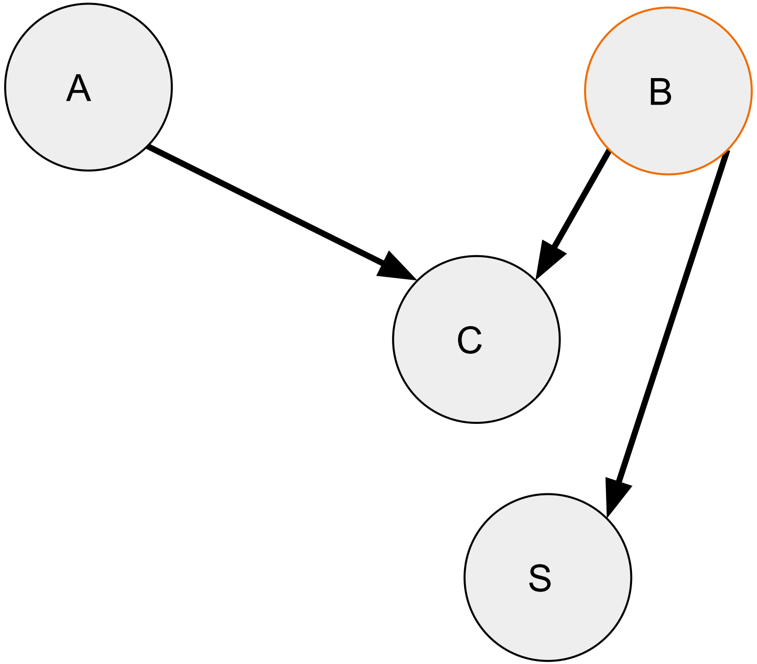
Use regression as a rough approximation




```r
n <- 1000
background <- runif(n)
ad <- rbinom(n, 1, 0.2)

clicks <- rexp(n, 0.1/(background + ad))

signup <- rbinom(n, 1, background)

model <- lm(signup ~ ad + clicks)
summary(model)
```

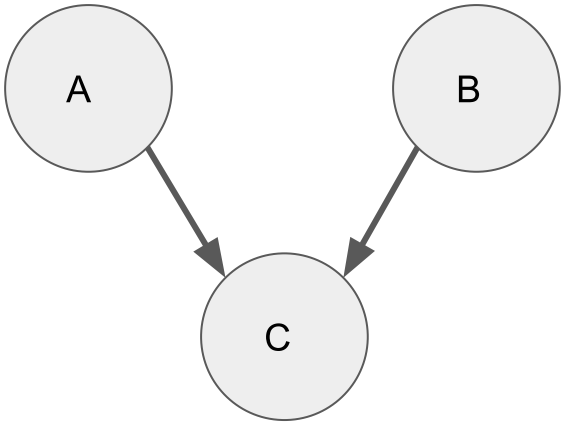
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.45757    0.01953  23.432  < 2e-16 ***
ad          -0.11848    0.04067  -2.913  0.00366 **
clicks       0.10020    0.01730   5.793 9.26e-09 ***
```

What if we observe the background?



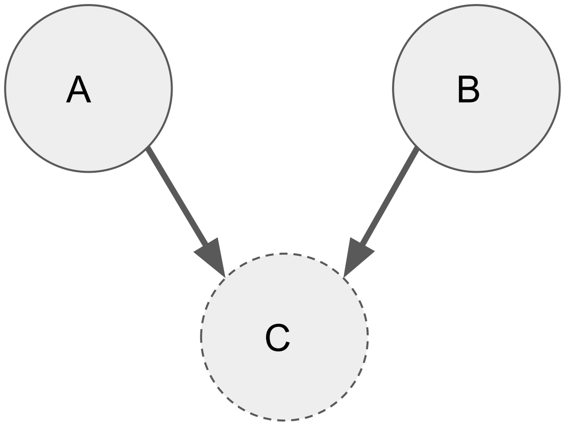

```r
n <- 1000
background <- runif(n)
ad <- rbinom(n, 1, 0.2)

clicks <- rexp(n, 0.1/(background + ad))

signup <- rbinom(n, 1, background)

model <- lm(signup ~ ad + clicks)
```

Collider's in Graphical Models!



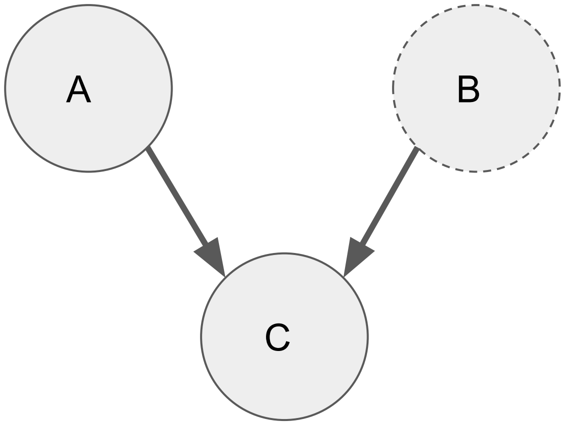
- Let A be a coin toss
- Let B be a separate coin toss
- Let C be "Were the coin toss outcomes from A and B the same?"
- C is a "collider"

A and B are independent



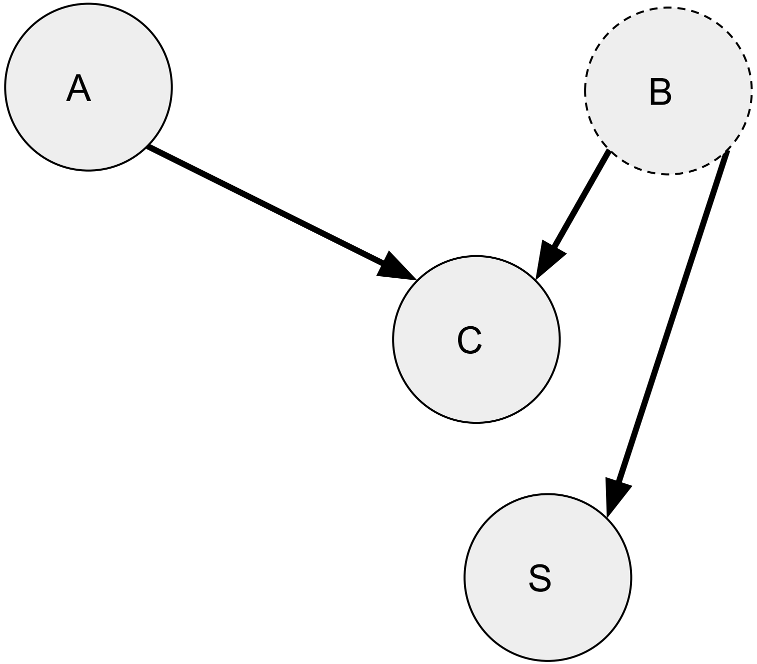
- Let A be a coin toss
- Let B be a separate coin toss
- Let C be "Were the coin toss outcomes from A and B the same?"

Conditioning on C, A and B are not independent

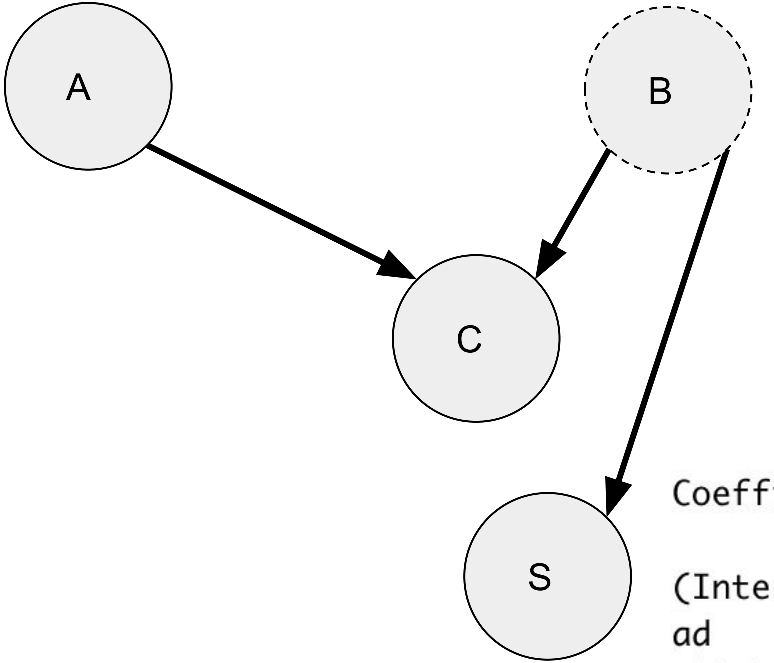


- Let A be a coin toss
- Let B be a separate coin toss
- Let C be “Were the coin toss outcomes from A and B the same?”

Where do we spot a collider?



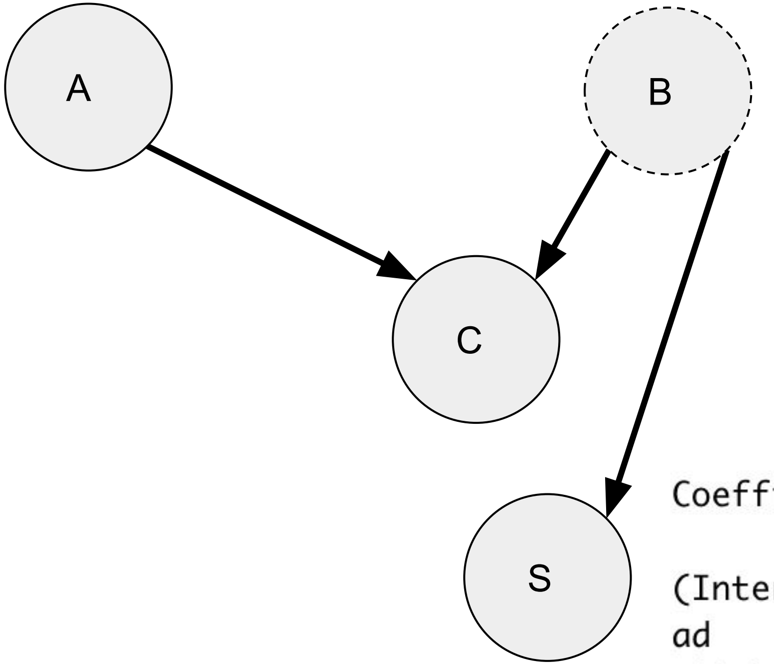
How would you spot this in real life?




```r
n <- 1000
background <- runif(n)
ad <- rbinom(n, 1, 0.2)

clicks <- rexp(n, 0.1/(background + ad))

signup <- rbinom(n, 1, background)

model <- lm(signup ~ ad + clicks)
summary(model)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.45757    0.01953  23.432  < 2e-16 ***
ad          -0.11848    0.04067  -2.913  0.00366 **
clicks       0.10020    0.01730   5.793 9.26e-09 ***
```

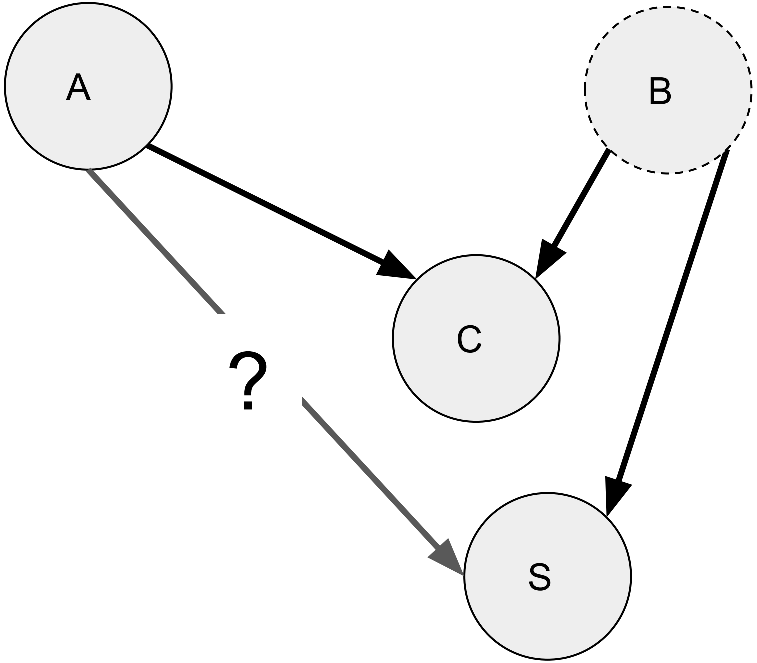
Can regularization or feature selection help?




```r
n <- 1000
background <- runif(n)
ad <- rbinom(n, 1, 0.2)

clicks <- rexp(n, 0.1/(background + ad))

signup <- rbinom(n, 1, background)

model <- lm(signup ~ ad + clicks)
summary(model)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.45757    0.01953  23.432  < 2e-16 ***
ad          -0.11848    0.04067  -2.913  0.00366 **
clicks       0.10020    0.01730   5.793 9.26e-09 ***
```

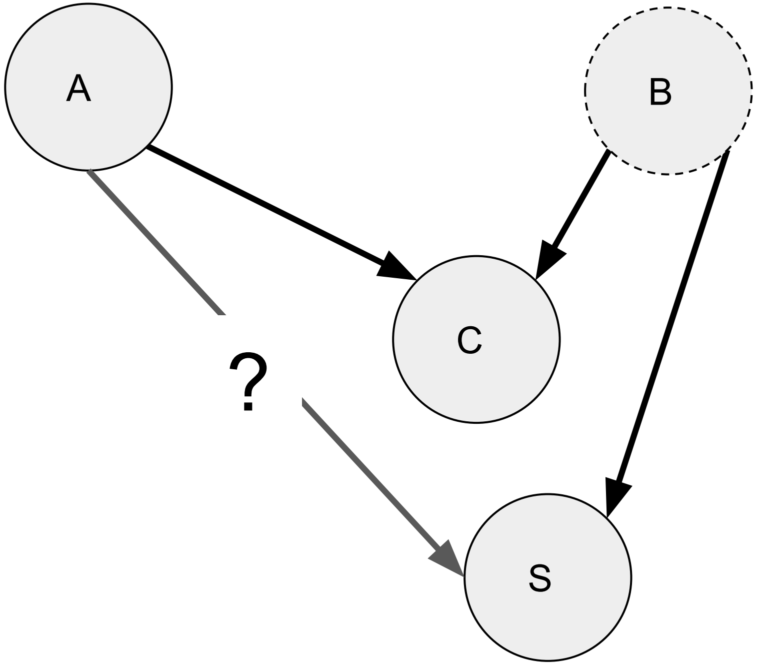
What is the fix?



Takeaways

- Trial data does not shield you from biased results
- Colliders are variables you need to be careful about adding to your model

Recall



AB testing in Data Science is more than a calculation

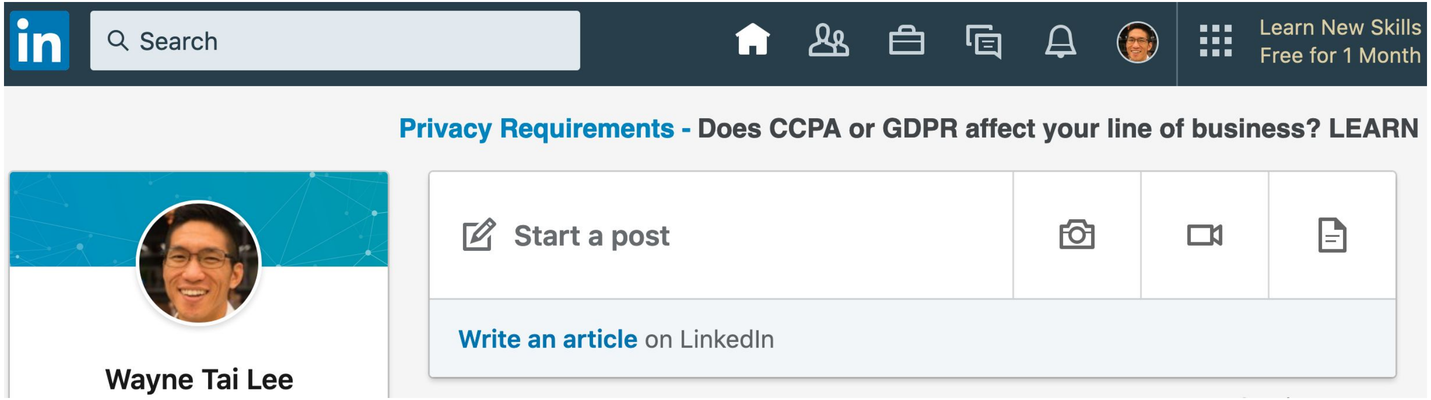
- Need to communicate the value as an internal product

- Forces a lot of necessary infrastructure and cultural changes

Communicating AB testing as an internal product

- Cost avoidance
  - Decrease “what about…?” arguments
  - Detect bugs early on

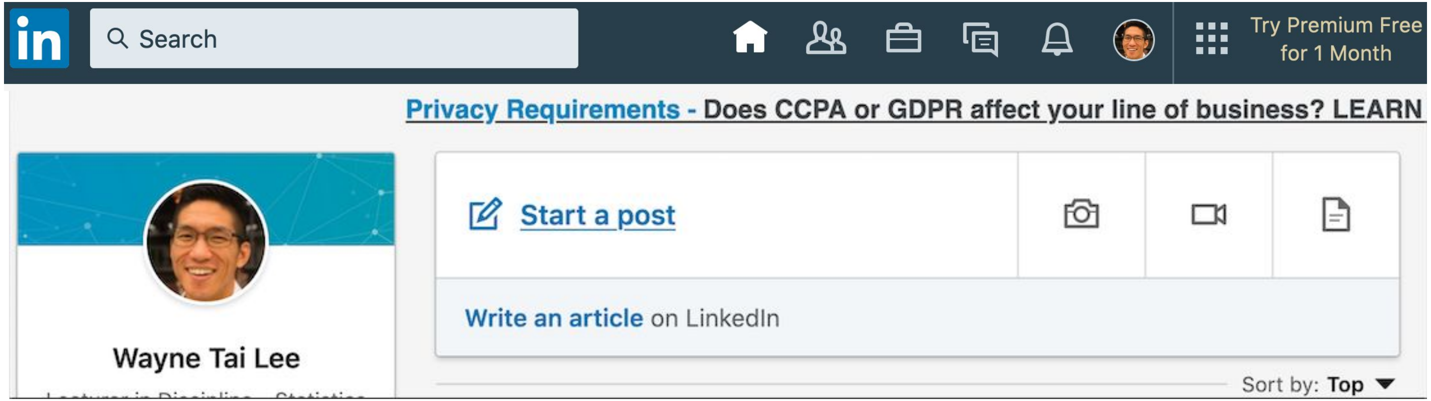
Can you see the difference?



Communicating AB testing as an internal product

- Cost avoidance
  - Decrease “what about...?” arguments
  - Detect bugs early on

Can you see the difference?



Communicating AB testing as an internal product

- Value generation
  - Accelerates feedback
  - Align measurable objectives

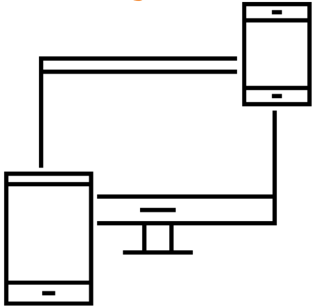
https://www.cdc.gov/mmwr/volumes/69/wr/mm6911e1.htm

What are the implications for public health practice?

A multipronged surveillance strategy could lead to enhanced case detection and reduced transmission of highly infectious diseases such as COVID-19.

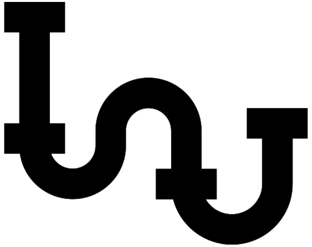
AB testing forces necessary changes in the company

- Teams need to agree on outcome definitions



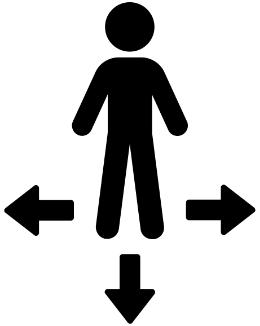
Created by Justin Blake
from Noun Project

- Data pipelines need to be transparent and consistent



by Francisco Javier Diaz Montejano
from Noun Project

- Make decisions based on AB testing results



Created by Adrien Coquet
from Noun Project

Images all from [The Noun Project]

Data Science is about end to end execution

- Communicate the importance of AB testing
- Provide the algorithm
- Test the execution
- Educate the company to proper AB testing

More about causal inference

- Prediction and causation are quite different. Typical questions are:

Prediction: Predict Y after **observing** $X = x$
Causation: Predict Y after **setting** $X = x$.

- Causation involves predicting the effect. For example:

Prediction: Predict health given that a person takes vitamin C
Causation: Predict health if I give a person vitamin C

Prediction v.s. Causation

- **Correlation is not causation!!!**
- Prediction is about passive observation;

  Causation is about active intervention
- Causal effects can be estimated consistently from randomized experiments.
- It is difficult to estimate causal effects from observational (non-randomized) experiments.
- All causal conclusions from observational studies should be regarded as very tentative.

Preliminaries

- 2 types of causal questions:

  - Causal inference: What is the causal effect of X on Y

  - Causal discovery: Determine the causal relationship among the variables.

Preliminaries

- 2 types of causal questions:

  - Causal inference: What is the causal effect of X on Y

  - Causal discovery: Determine the causal relationship among the variables.

- 2 types of data

  - Randomized: causal inference is straightforward

  - Observational (non-randomized): subject matter knowledge, stronger assumption

Preliminaries

- 2 types of causal questions:

  - Causal inference: What is the causal effect of X on Y

  - Causal discovery: Determine the causal relationship among the variables.
- 2 types of data

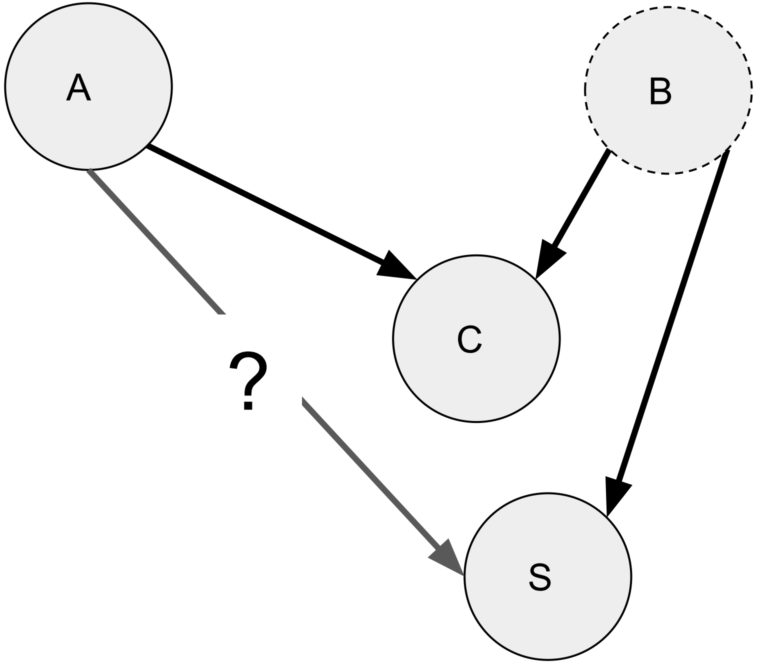
  - Randomized: causal inference is straightforward

  - Observational (non-randomized): subject matter knowledge, stronger assumption
- 2 languages of causation: mathematically they are equivalent!

  - Causal graphs: we have seen this last time!

  - Counterfactuals: we are going to see it today!

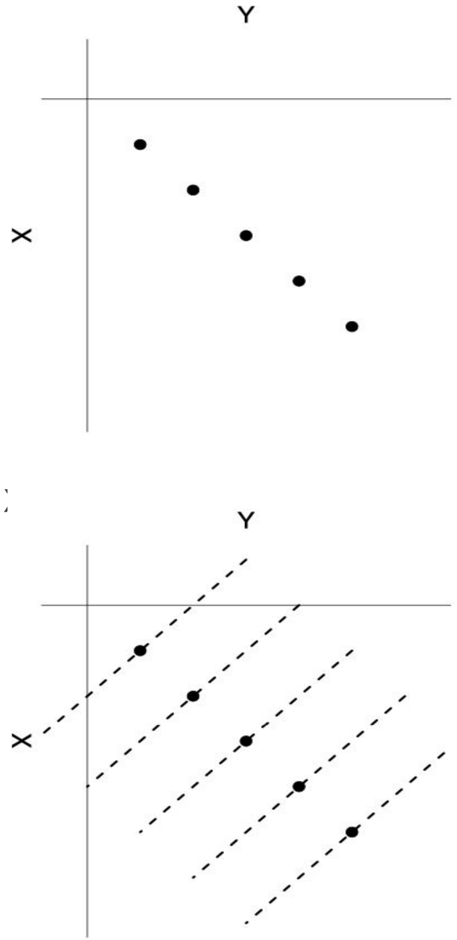
Causal graphs



Counterfactuals

- treatment X, outcome Y
- Counterfactuals:

what their value of Y would have been if we changed their value of X



Causal inference is tricky

- Causal effects can be estimated consistently from randomized experiments;
- It is difficult to estimate causal effects from observational (non-randomized) experiments;
- All causal conclusion from observational studies should be regarded as very tentative.