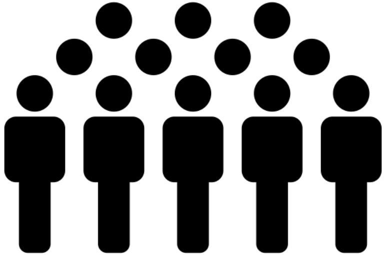
Causal Inference

Lecture 14

Xiaofei Shi

Where does bias come from?



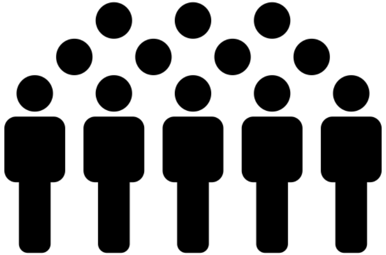
Created by Wilson Joseph
from Noun Project

Images from The Noun Project

Summary

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

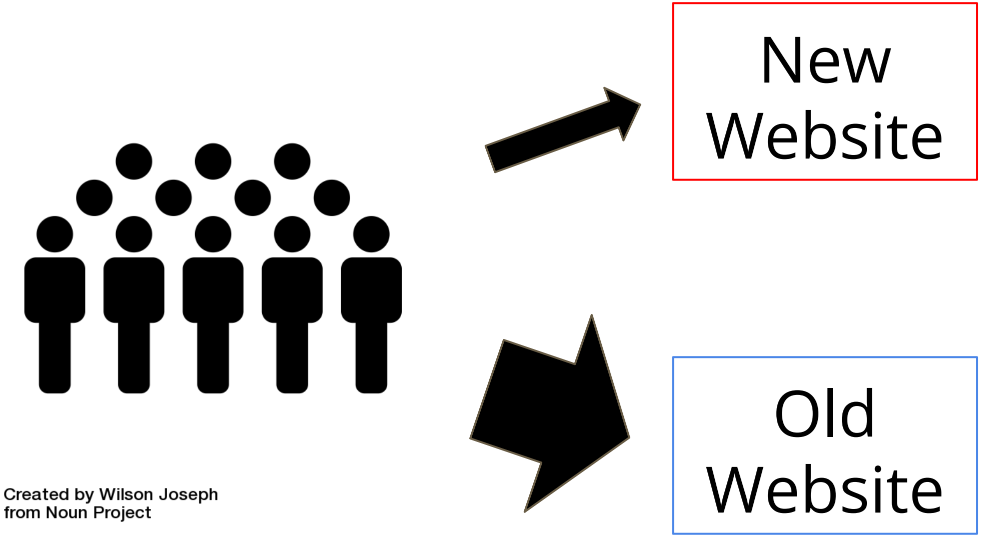
AB testing



Created by Wilson Joseph
from Noun Project

Images from The Noun Project

AB testing



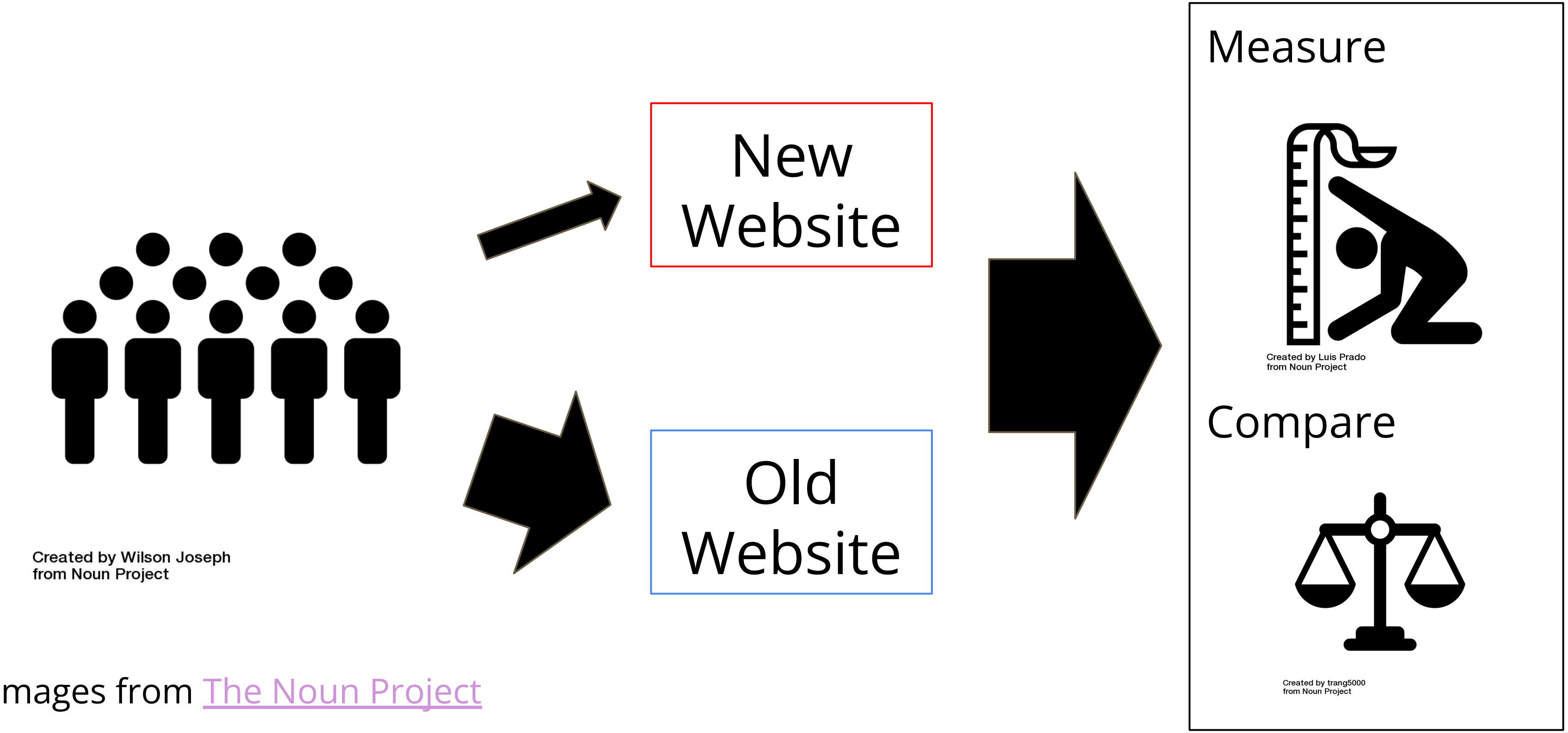
New Website

Old Website

Created by Wilson Joseph
from Noun Project

Images from The Noun Project

AB testing



Created by Wilson Joseph
from Noun Project

New Website

Old Website

Measure

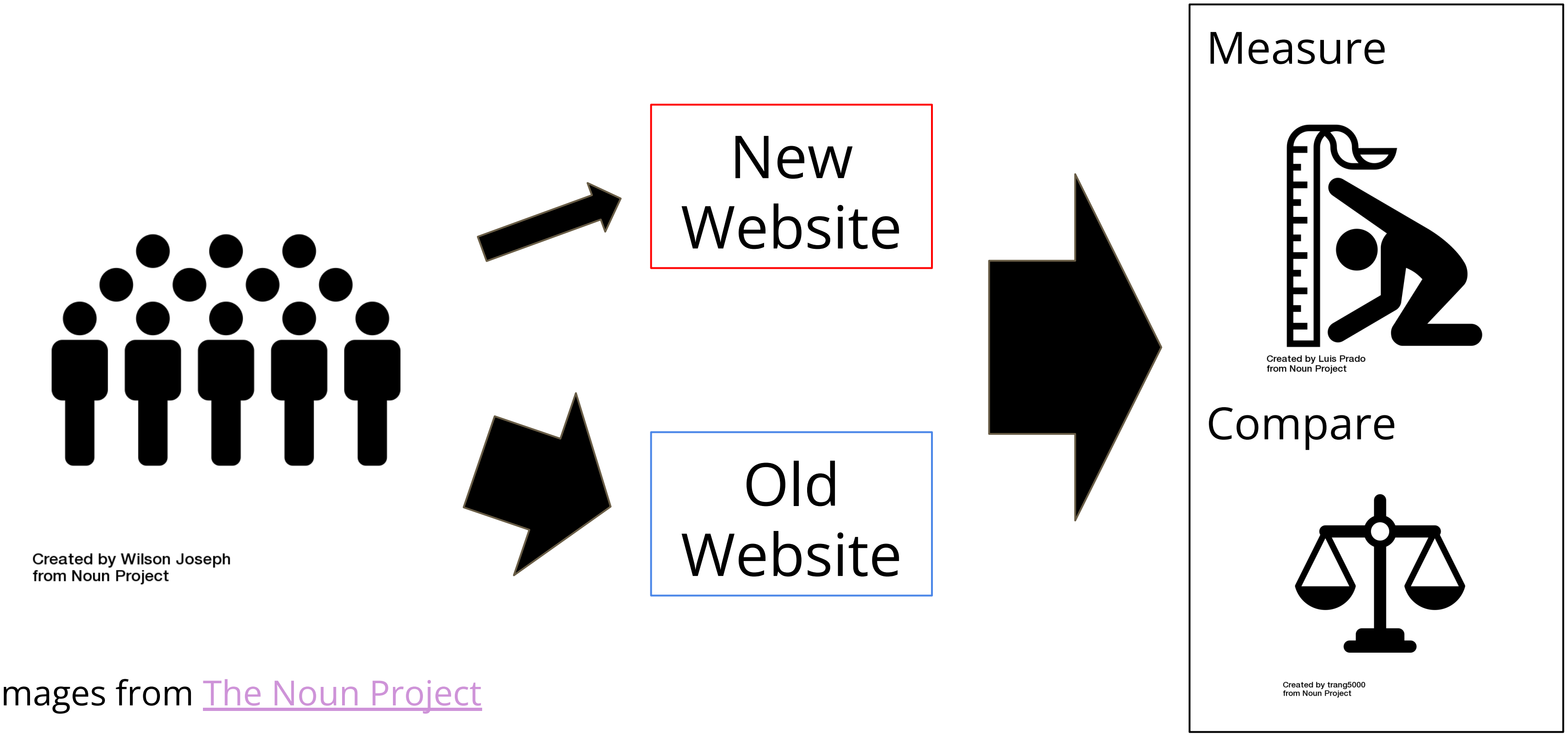
Created by Luis Prado
from Noun Project

Compare

Created by trang5000
from Noun Project

Images from The Noun Project

AB testing == Randomized controlled trials?



Created by Wilson Joseph
from Noun Project

New Website

Old Website

Measure

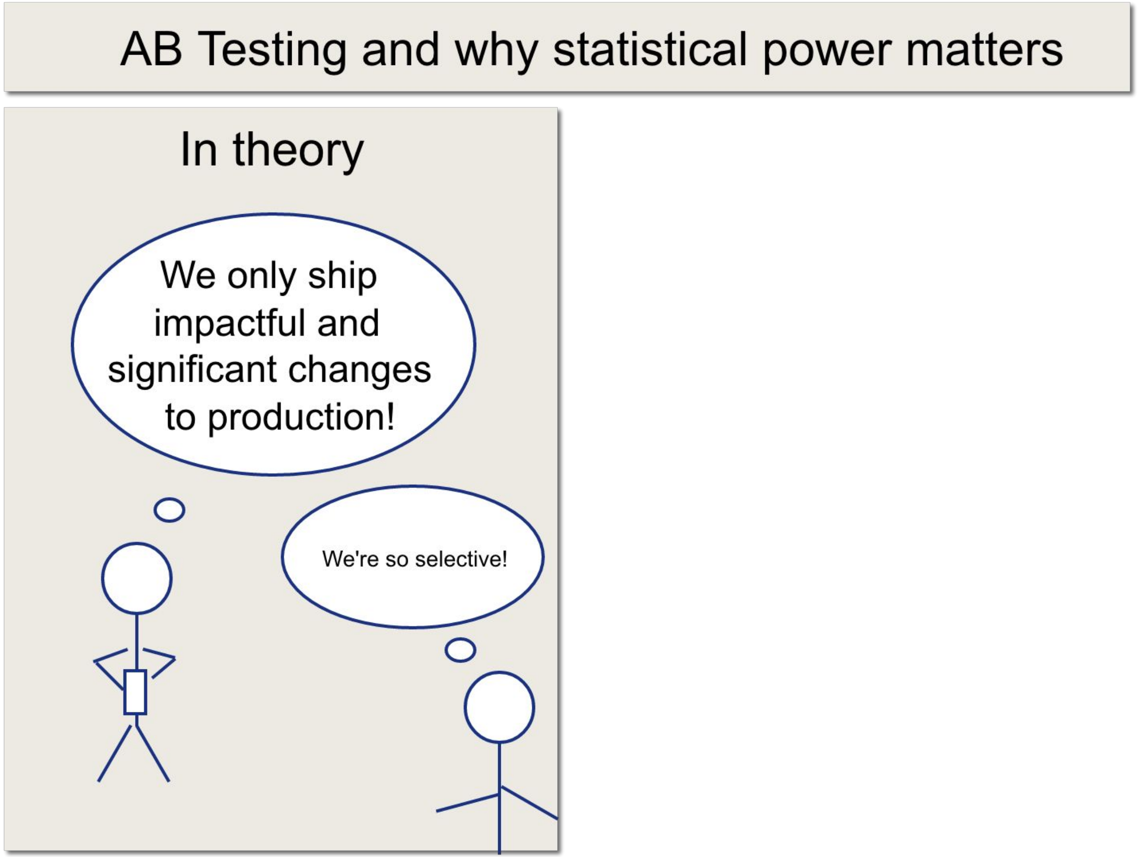
Created by Luis Prado
from Noun Project

Compare

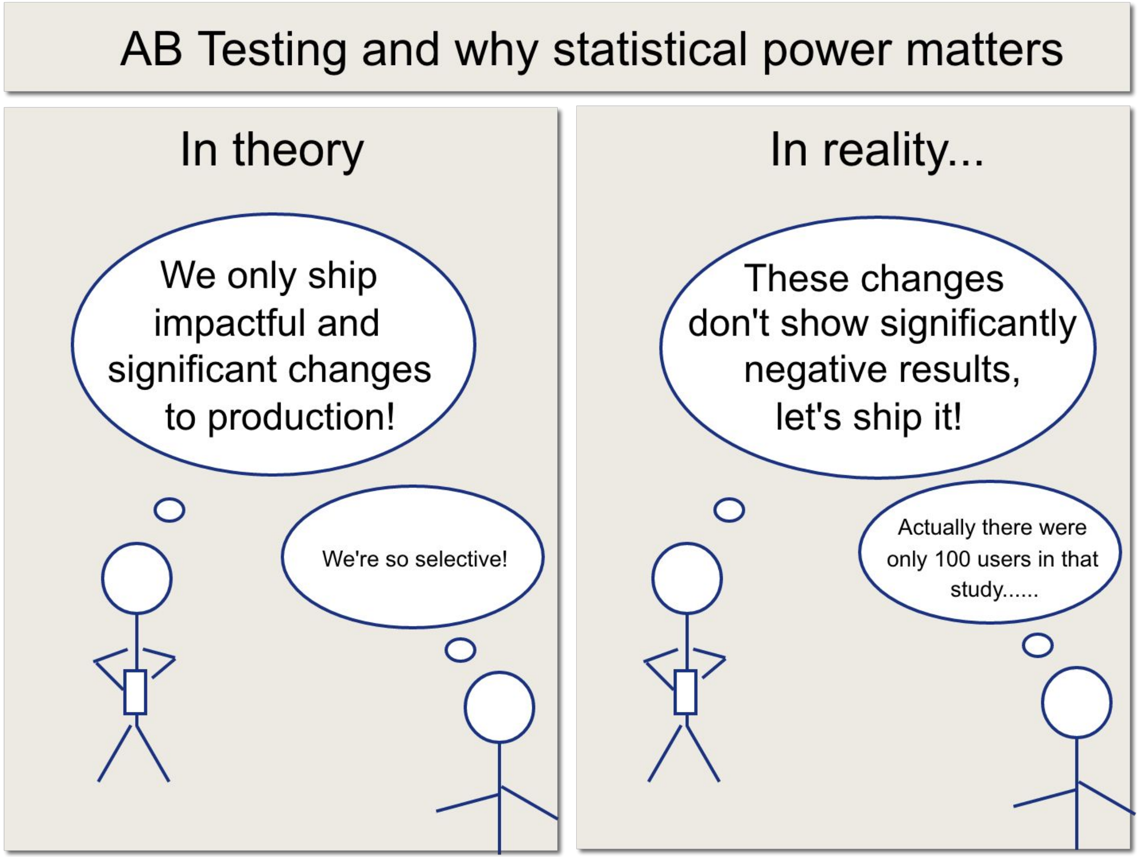
Created by trang5000
from Noun Project

Images from The Noun Project

Most tests are not significant


AB Testing and why statistical power matters
In theory
We only ship impactful and significant changes to production!
We're so selective!

Most tests are not significant



AB Testing and why statistical power matters

In theory

In reality...

We only ship impactful and significant changes to production!
We're so selective!
These changes don't show significantly negative results, let's ship it!
Actually there were only 100 users in that study......

What happens when tests are not significant

Created by Adrien Coquet
from Noun Project

Image from [The Noun Project]

What happens when tests are not significant

Created by Adrien Coquet
from Noun Project

- People want to understand **why**

Image from [The Noun Project]

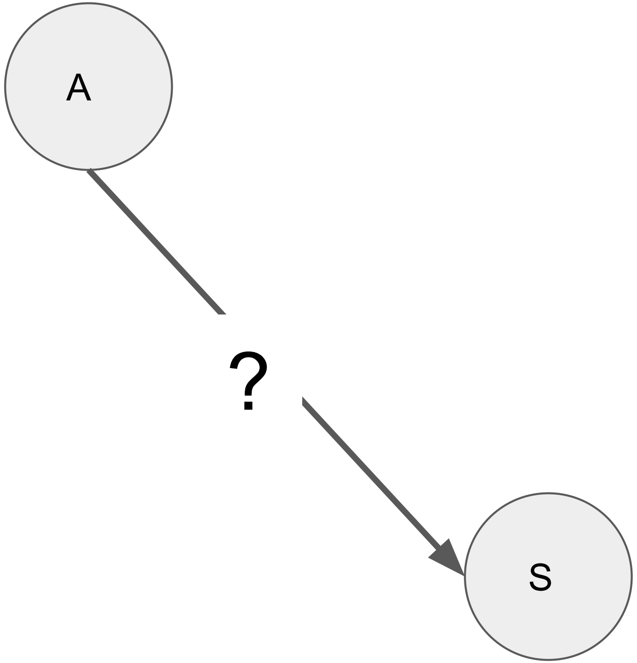
What happens when tests are not significant

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?

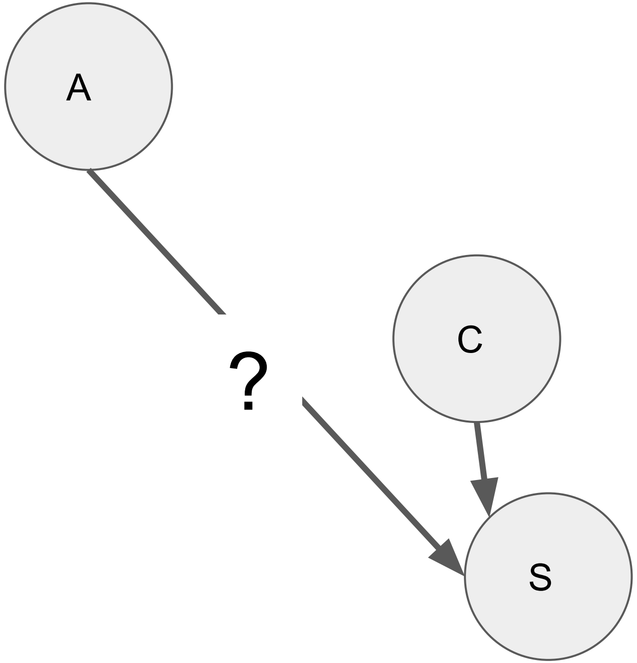
Image from The Noun Project

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



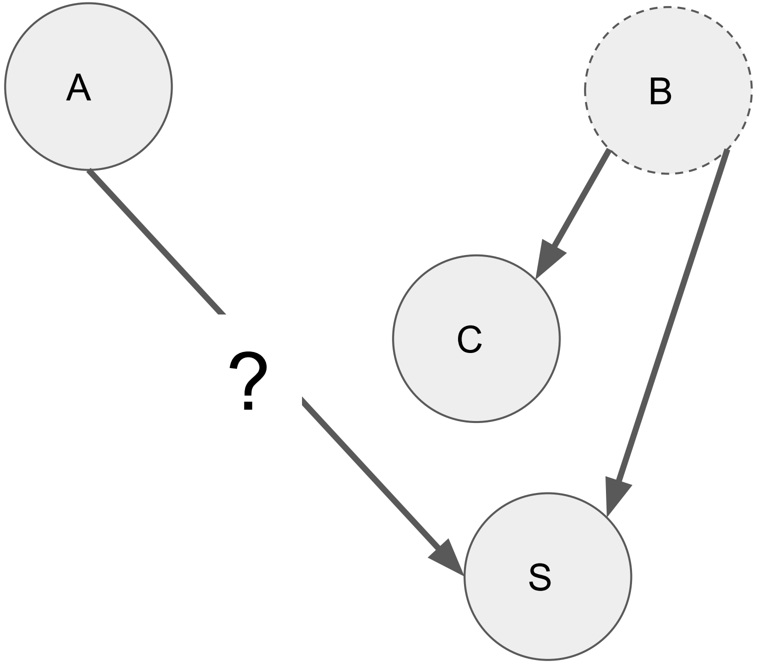
- A: exposure to an **A**dvertisement
- S: user **S**igned-up for the service

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



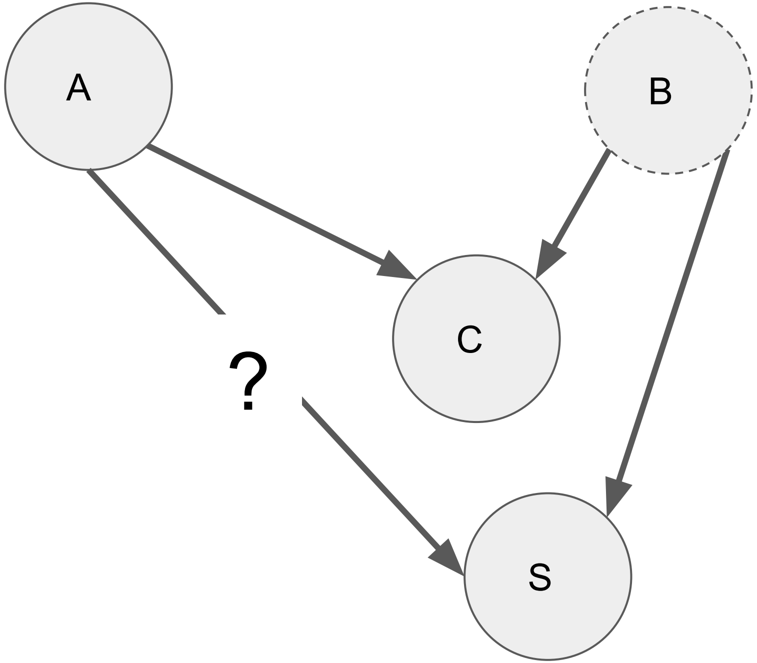
- A: exposure to an **A**dvertisement
- S: user **S**igned-up for the service
- C: user **C**licks

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



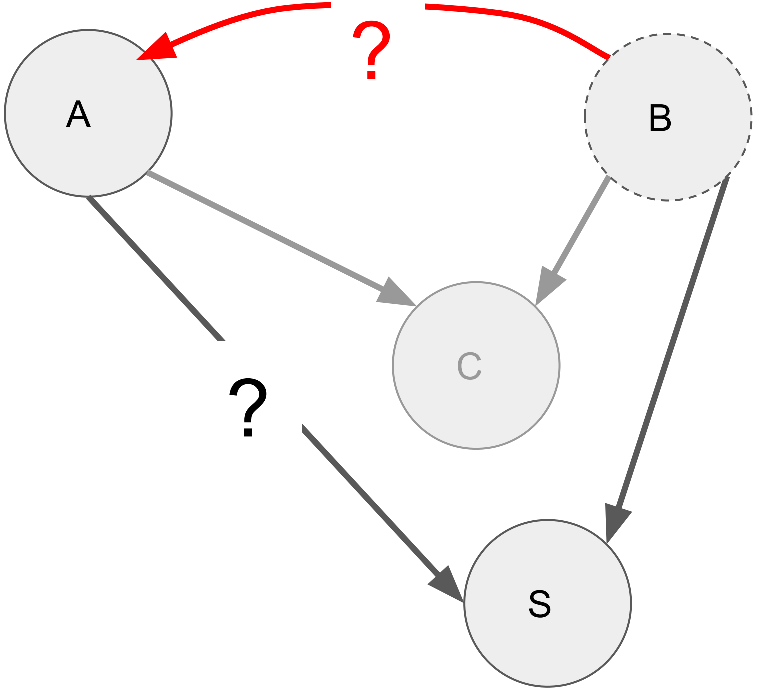
- A: exposure to an **A**dvertisement
- S: user **S**igned-up for the service
- C: user **C**licks
- B: unobserved **B**ackground

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



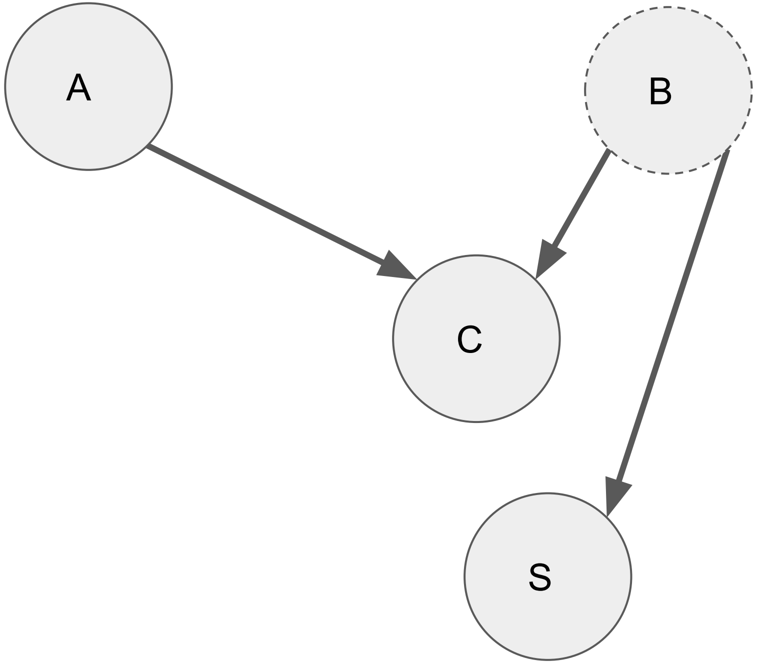
- A: exposure to an **A**dvertisement
- S: user **S**igned-up for the service
- C: user **C**licks
- B: unobserved **B**ackground

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



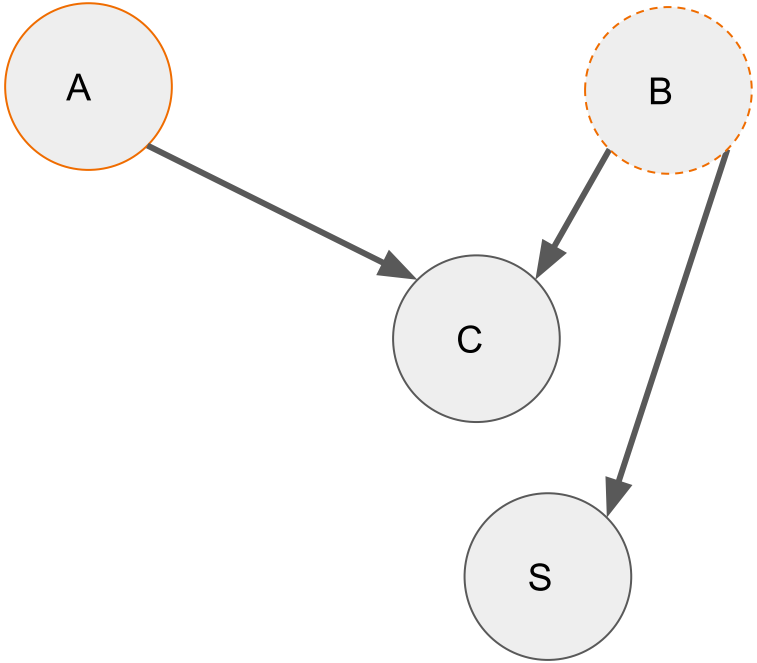
- A: exposure to an **A**dvertisement
- S: user **S**igned-up for the service
- C: user **C**licks
- B: unobserved **B**ackground

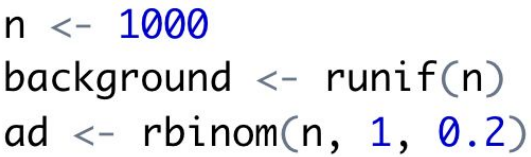
Turns out we can detect a significant effect from A on S even if A has no impact



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

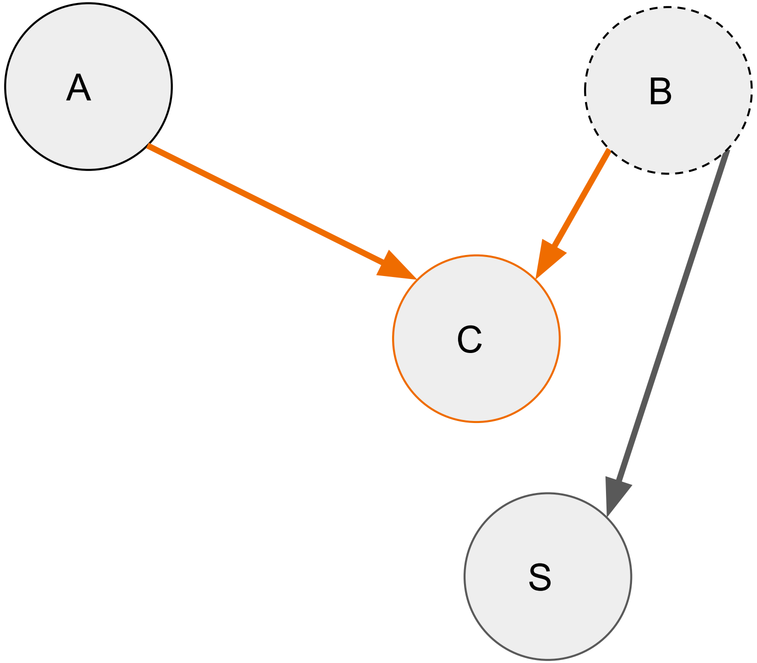
Let’s simulate this!

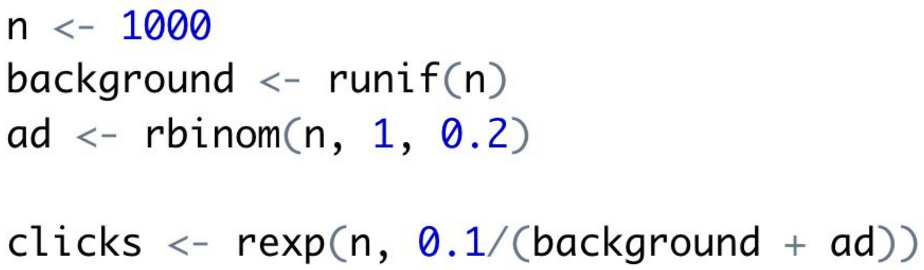




```
n <- 1000
background <- runif(n)
ad <- rbinom(n, 1, 0.2)
```

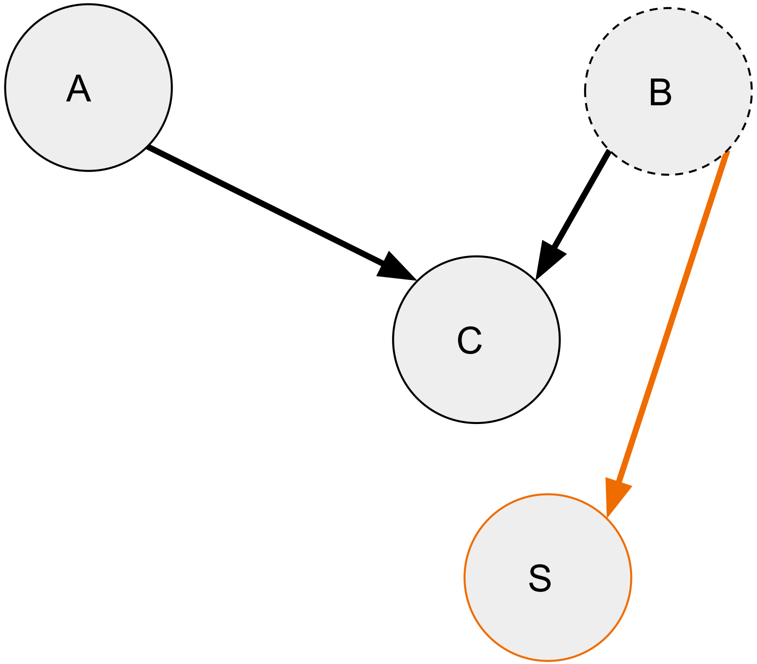
Let’s simulate this!

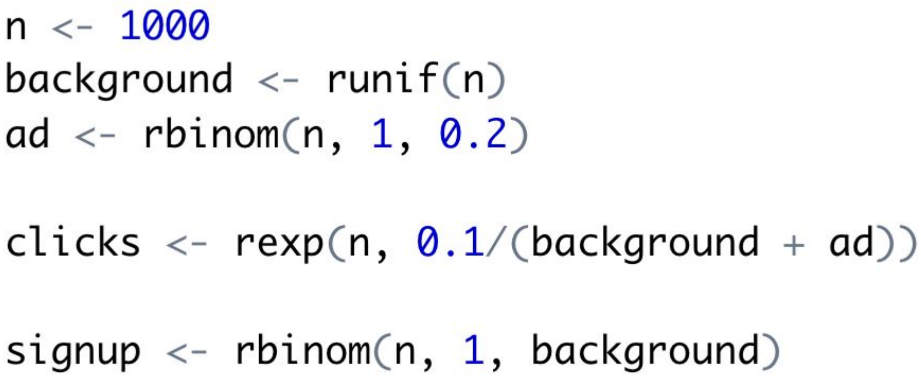




```r
n <- 1000
background <- runif(n)
ad <- rbinom(n, 1, 0.2)

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

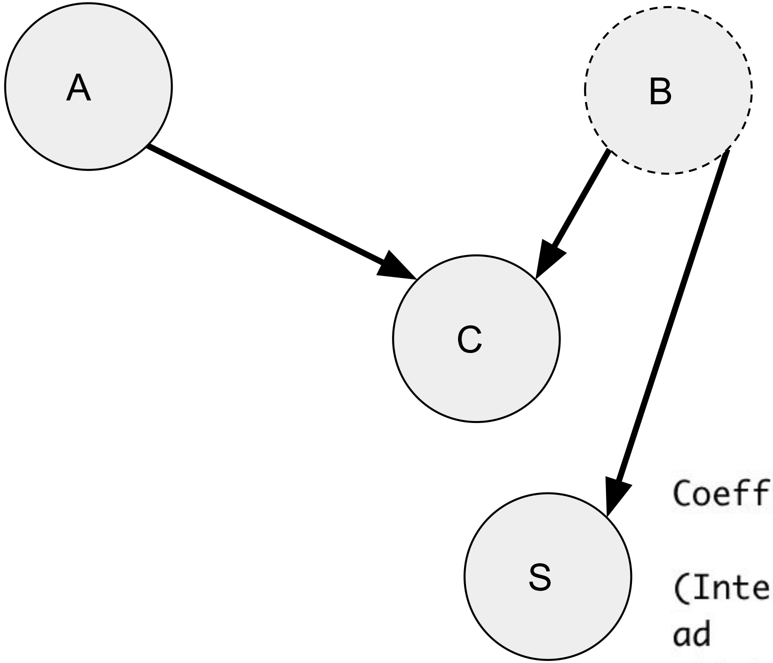
Let’s simulate this!





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

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

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

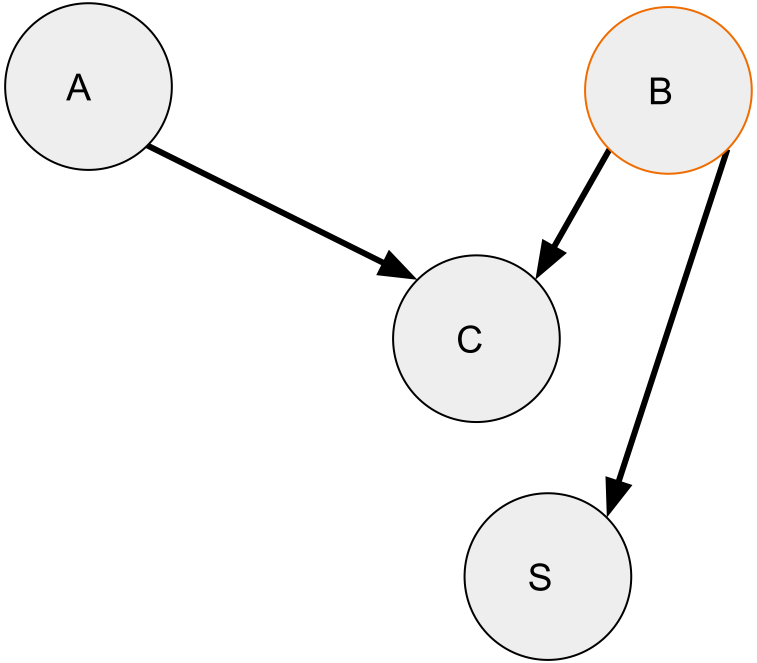
Use regression as a rough approximation

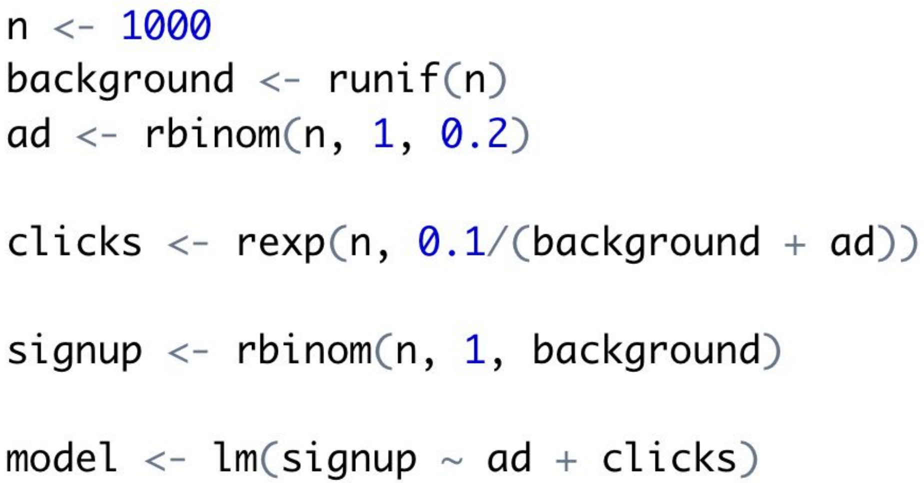


```
n <- 1000
background <- runif(n)
ad <- rbinom(n, 1, 0.2)

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 ***
```

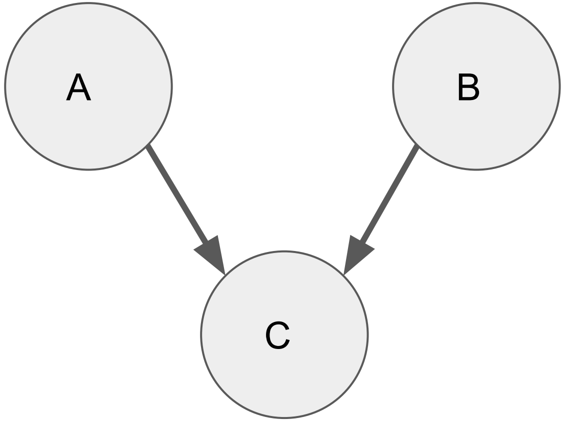
What if we observe the background?





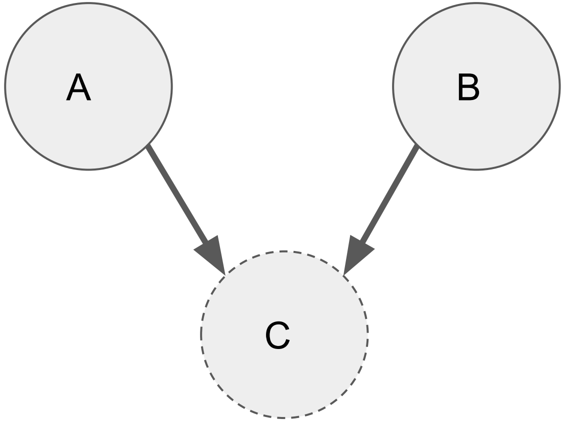
```r
n <- 1000
background <- runif(n)
ad <- rbinom(n, 1, 0.2)

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

signup <- rbinom(n, 1, background)

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

Collider’s in Graphical Models!



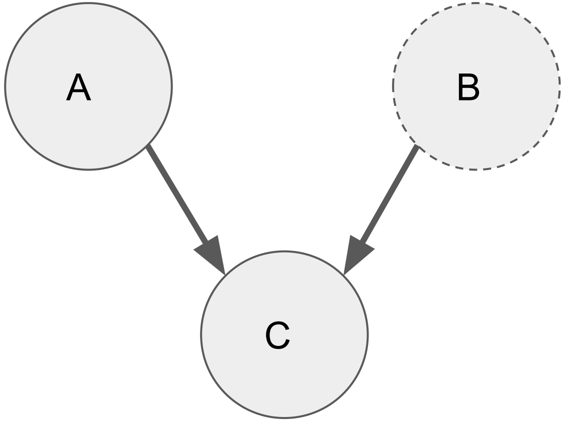
- Let A be a coin toss
- Let B be a separate coin toss
- Let C be “Were the coin toss outcomes from A and B the same?”
- C is a “collider”

A and B are independent



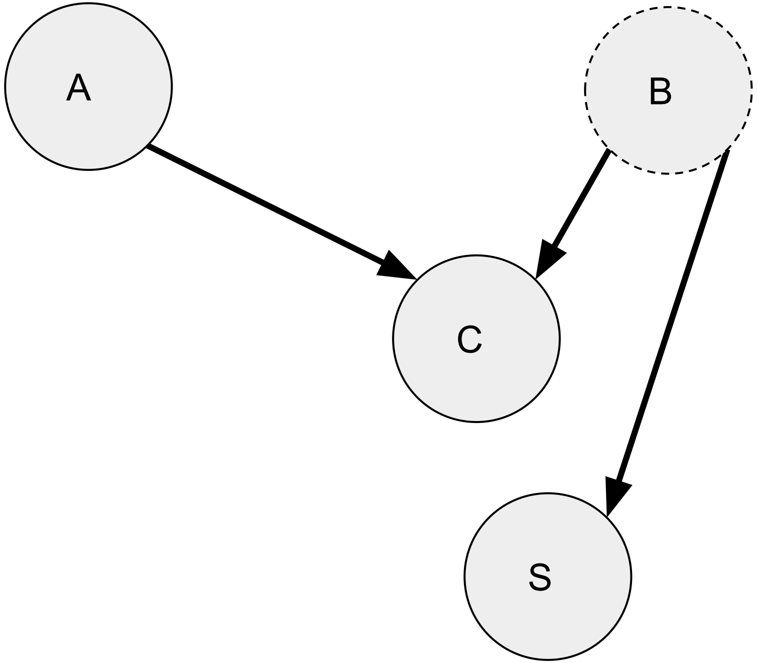
- Let A be a coin toss
- Let B be a separate coin toss
- Let C be “Were the coin toss outcomes from A and B the same?”

Conditioning on C, A and B are not independent

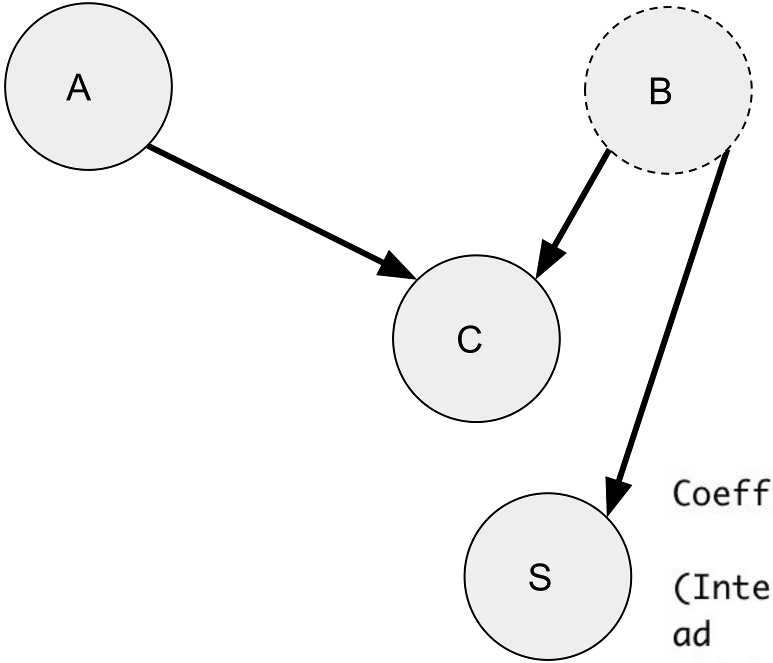


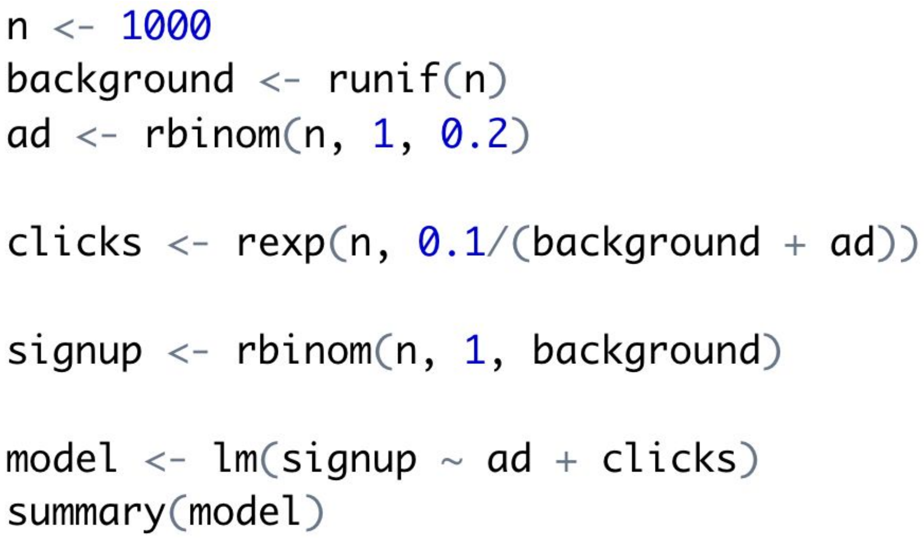
- Let A be a coin toss
- Let B be a separate coin toss
- Let C be “Were the coin toss outcomes from A and B the same?”

Where do we spot a collider?

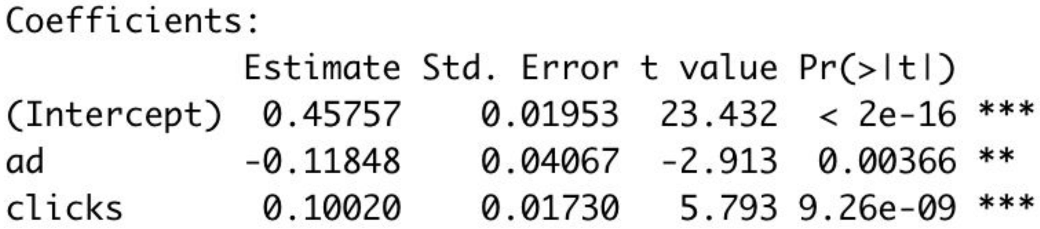


How would you spot this in real life?



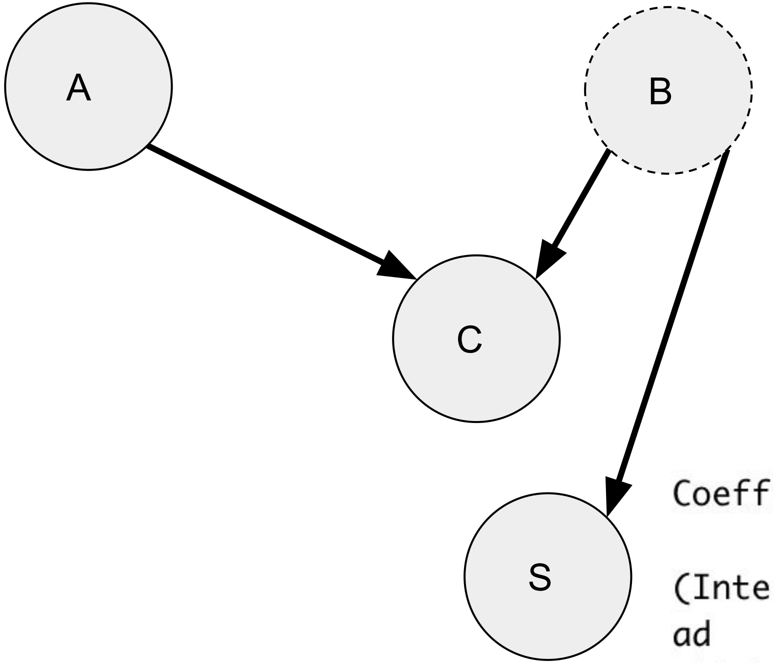


```
n <- 1000
background <- runif(n)
ad <- rbinom(n, 1, 0.2)

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 ***
```

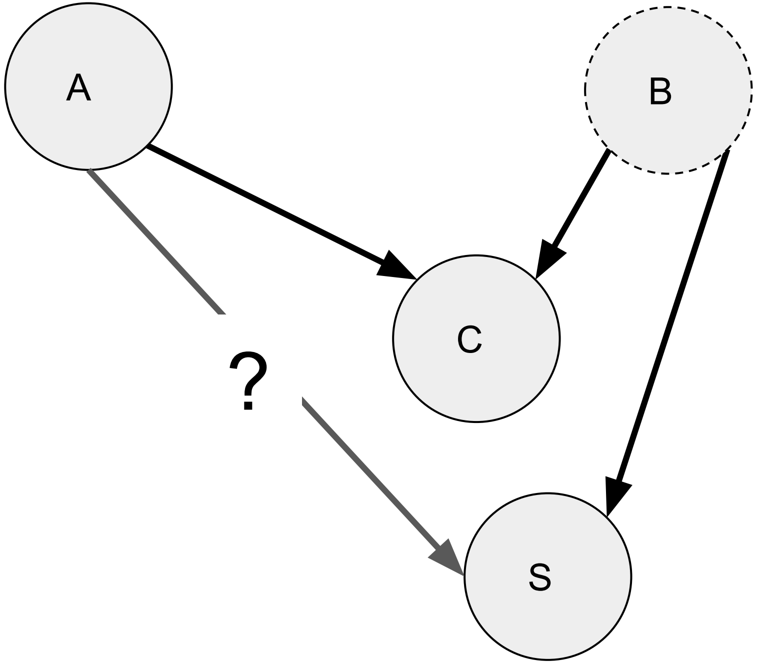
Can regularization or feature selection help?



```
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 ***
```

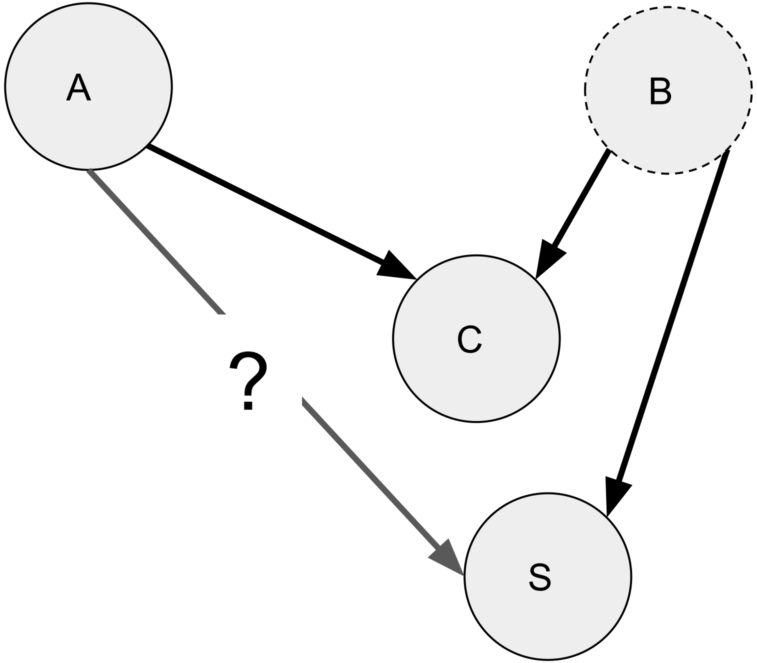
What is the fix?



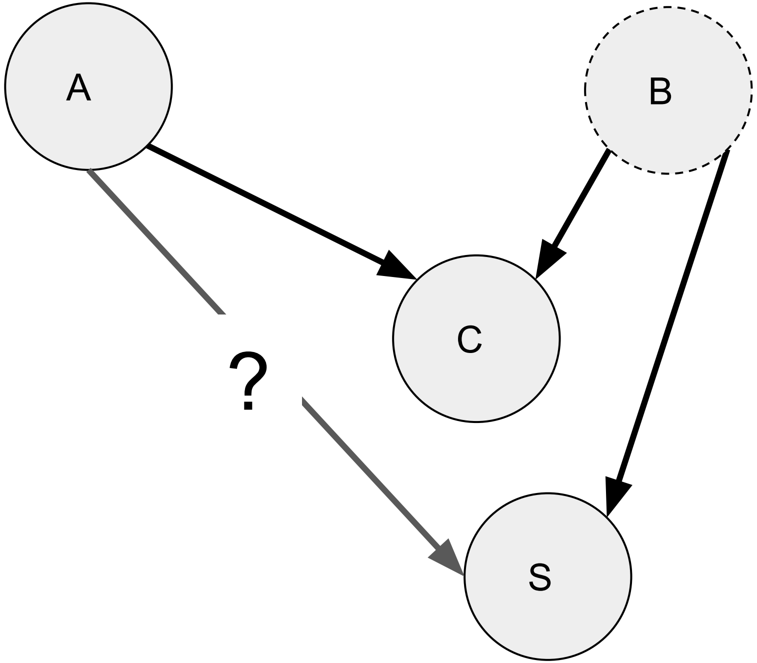
Takeaways

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

Bonus question: are colliders always bad?



Questions?



AB testing in Data Science is more than a calculation

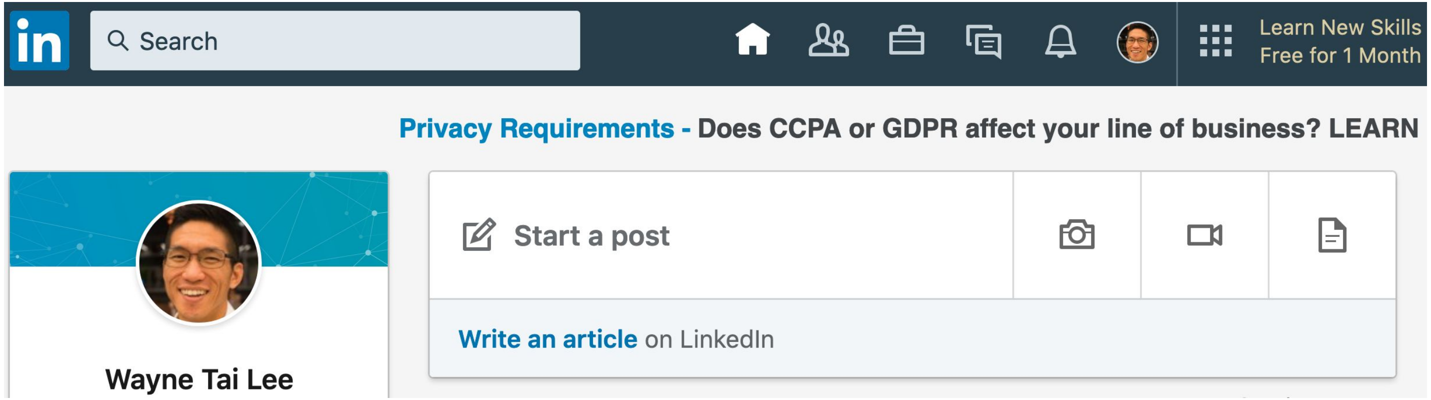
- Need to communicate the value as an internal product

- 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?



Privacy Requirements - Does CCPA or GDPR affect your line of business? LEARN

Wayne Tai Lee

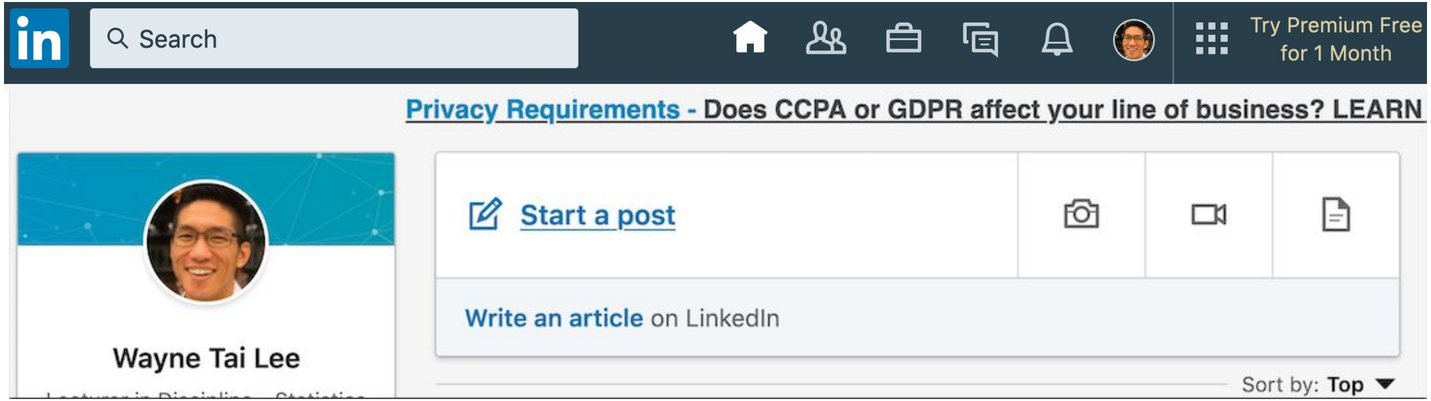
Start a post

Write an article on LinkedIn

Communicating AB testing as an internal product

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

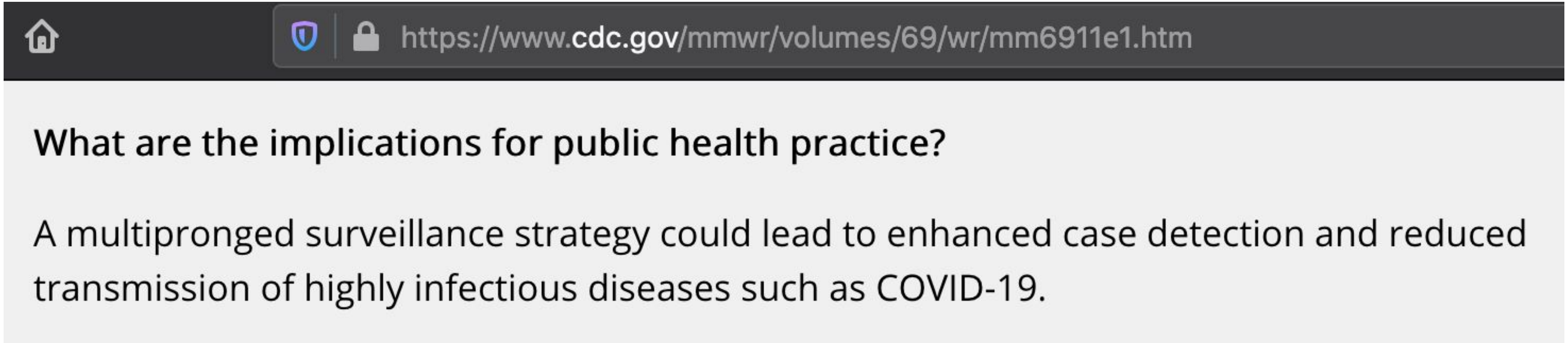
Can you see the difference?



Privacy Requirements - Does CCPA or GDPR affect your line of business? LEARN

Communicating AB testing as an internal product

- Value generation
  - Accelerates feedback
  - Align measurable objectives



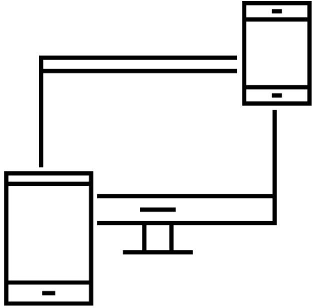
https://www.cdc.gov/mmwr/volumes/69/wr/mm6911e1.htm

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.

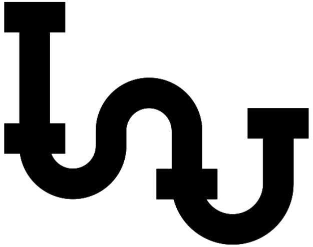
AB testing forces necessary changes in the company

- Teams need to agree on outcome definitions



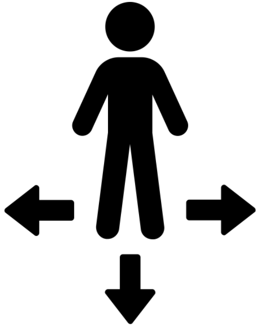
Created by Justin Blake
from Noun Project

- 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