zalando

CAUSAL INFERENCE

ALICE LYNCH

RESEARCH ENGINEER

zalando

DATA SCIENCE FESTIVAL

MEETUP

12TH NOV 2019

Running

Any amount of running reduces risk of early death, study finds

Previous research suggested health benefits increased with greater volume of running

Nicola Davis
@NicolaKSDavis
Mon 4 Nov 2019
23.30 GMT

most viewed


Live Liverpool 3-1 Manchester City: Premier League - live reaction!


The five: exercises to help avoid an early death

Business

Want to live longer? Try getting a dog.

Dog ownership significantly lowers mortality risk; now researchers are trying to find out just how they keep people alive.

Researchers have attached a laundry list of health benefits to dog ownership. Dogs not only "offer companionship, reduce anxiety and loneliness, increase self-esteem, and improve overall mood," but also force their humans to exercise and spend more time outdoors. (iStock)

By Christopher Ingraham

Most Read Business

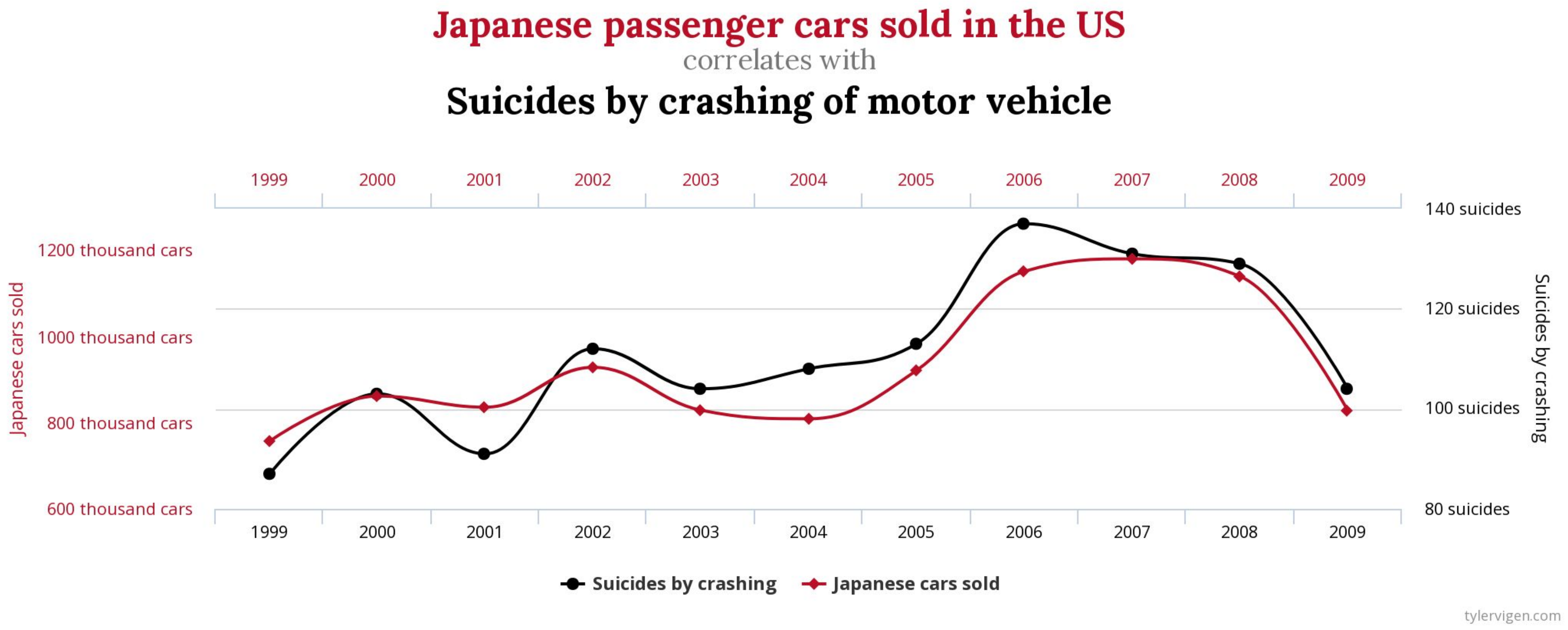
1 Analysis
Why Alabama and Mississippi have suddenly gone in opposite directions

2 The men behind GirlsDoPorn lured young women with modeling jobs, then tricked them into porn, FBI says

3 Perspective
Trust me: You need to start saving now so you can fly last-minute to be there for someone you love when the time comes

4 Perspective
Even doctors wonder how Medicare works

5 Inside the little-known world of flavorists, who are trying to make plant-based meat taste like the real thing

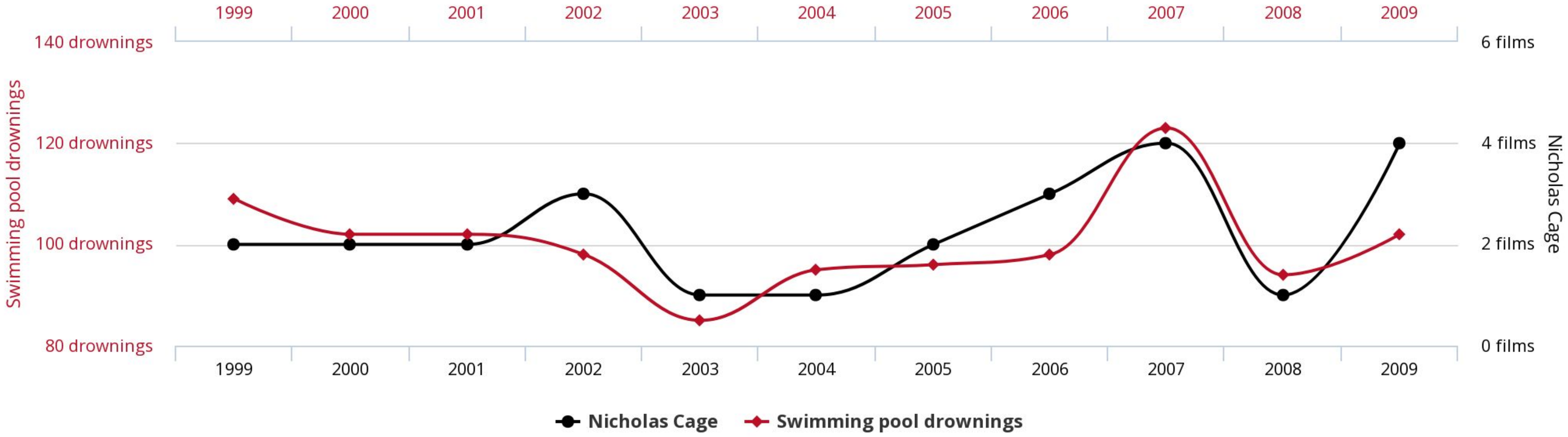


http://tylervigen.com/spurious-correlations

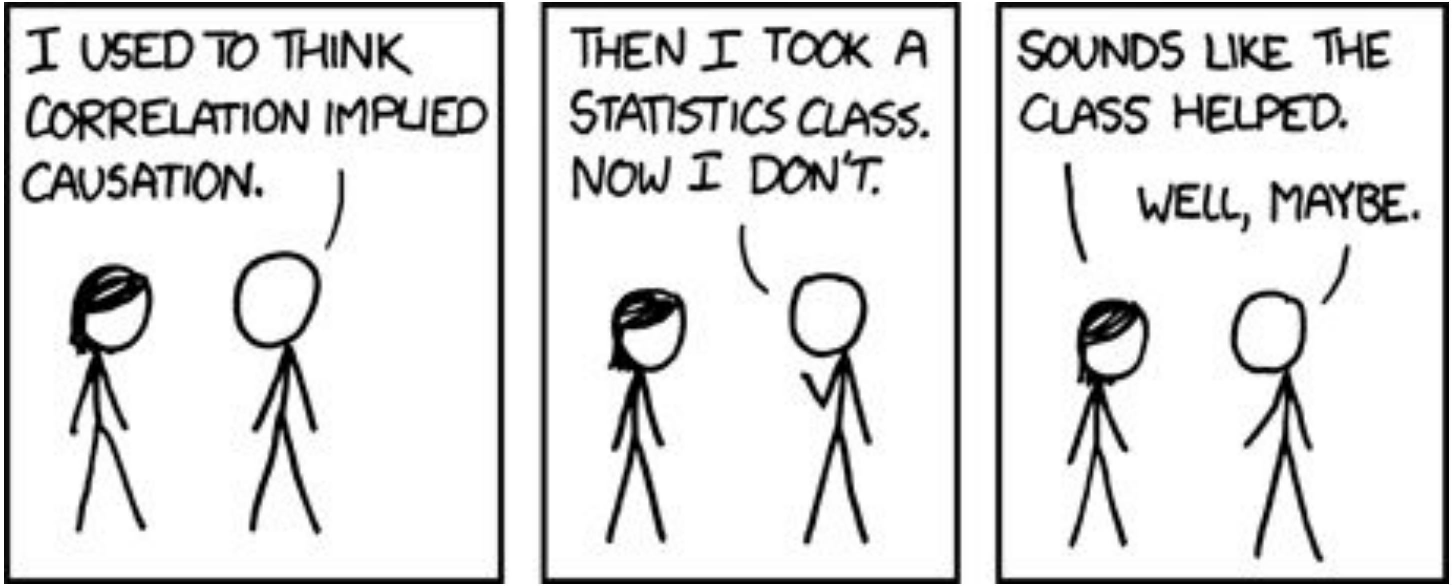
Number of people who drowned by falling into a pool

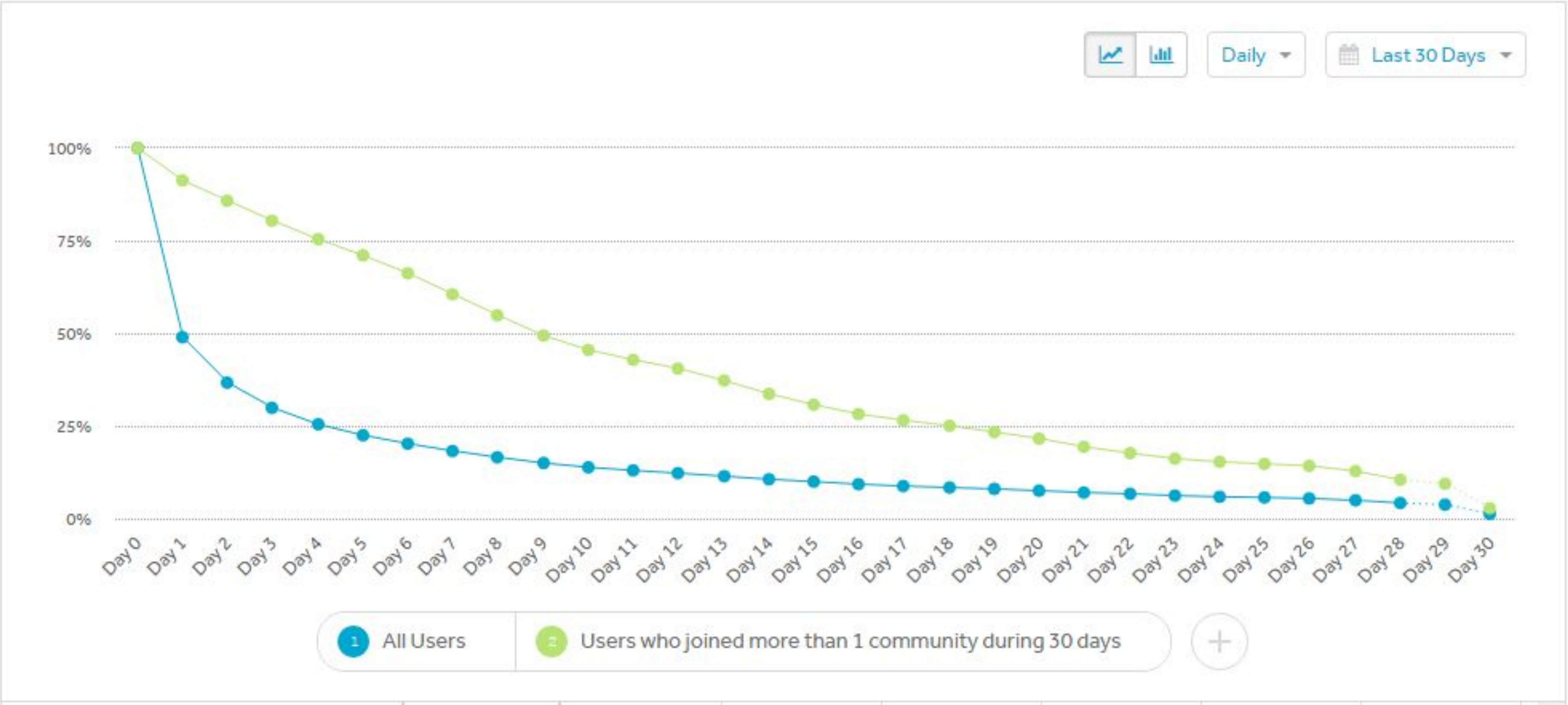
correlates with

Films Nicolas Cage appeared in

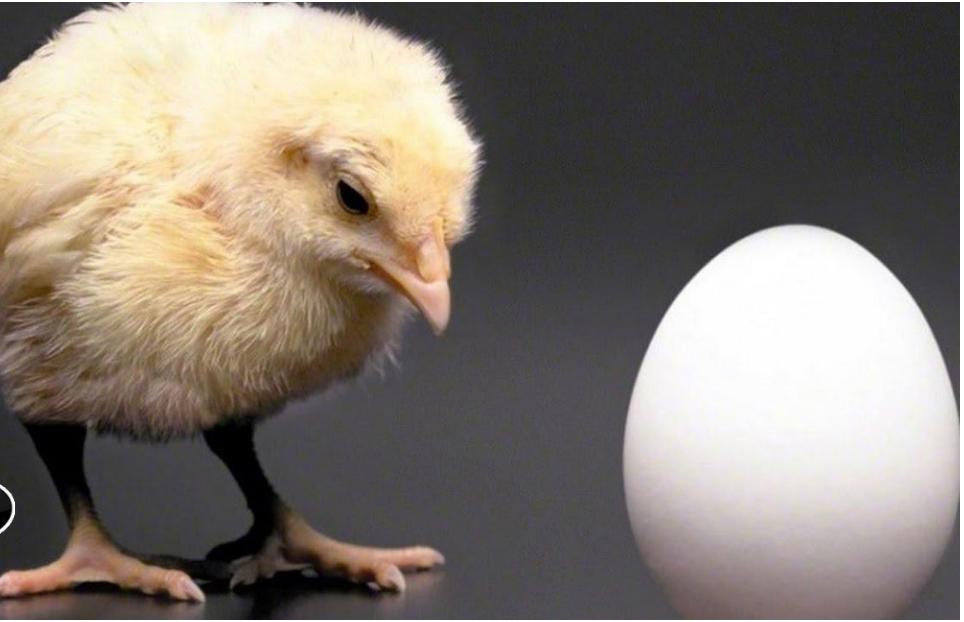


tylervigen.com


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.



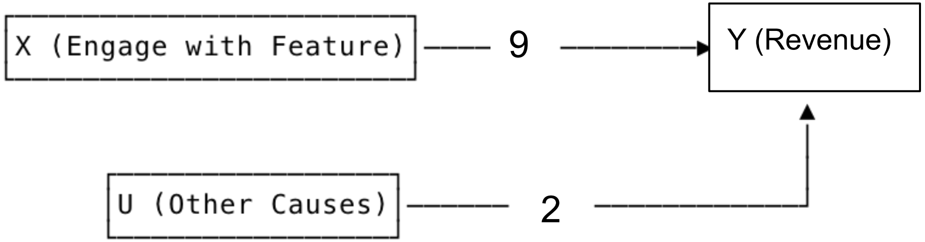
https://amplitude.com/blog/2017/01/19/causation-correlation



WHY CAUSALITY MATTERS

- Actions are taken based on their potential outcomes
- Can still make accurate predictions based on correlations

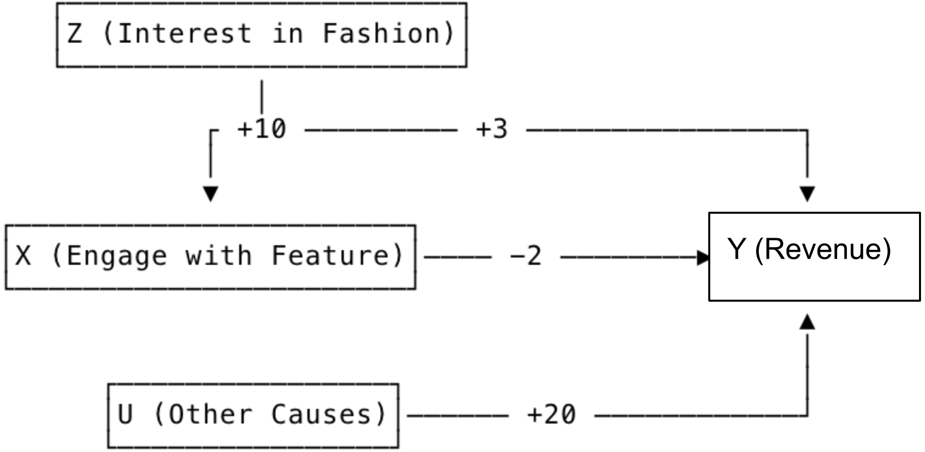
zalando



WHY CAUSALITY MATTERS

- A machine learning model may say that feature X has a statistically significant positive impact on Revenue (~9 euros more per hour) and the model has a very high R^2: ~0.96.
-

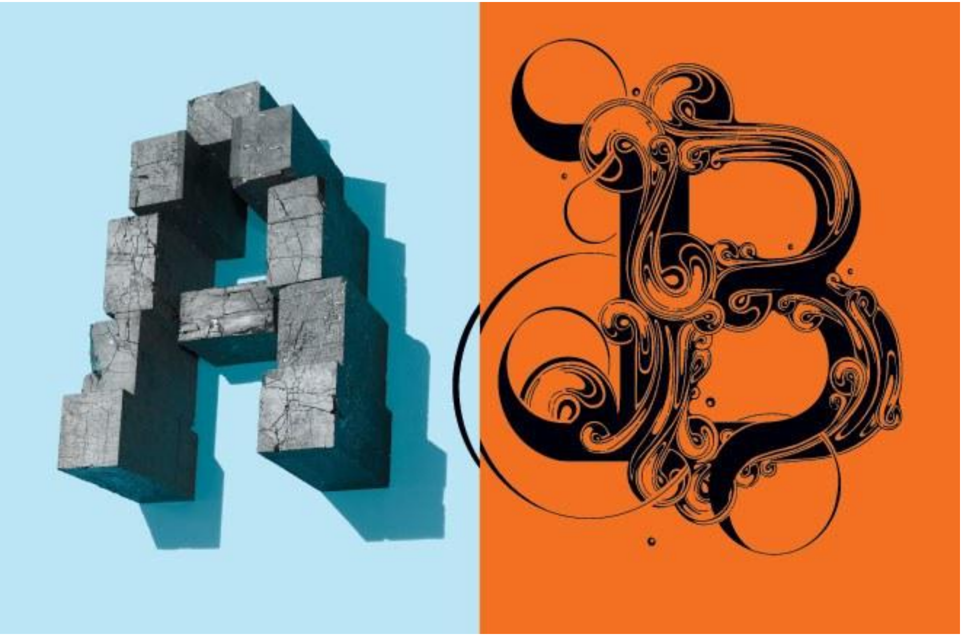
zalando



WHY CAUSALITY MATTERS

- Interest in Fashion is actually the cause of Engagement and Revenue
- The positive effect of Engagement with Revenue is actually just the positive effect of Interest in Fashion passing through (and actually decreased) by Time on Site.

zalando



ESTABLISHING CAUSALITY:

A/B TESTING

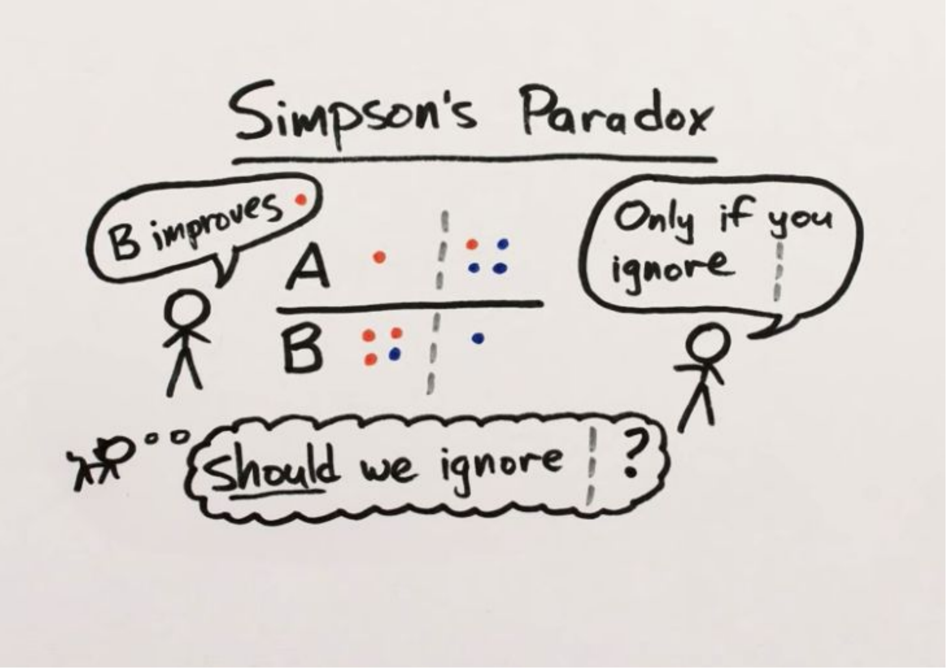
- Gold Standard: Randomised control trial
- Not always possible/ethical/affordable/in line with strategy

zalando

ESTABLISHING CAUSALITY: OBSERVATIONAL METHODS

- Data is obtained passively, without designing an experiment
- Care must be taken to control for Confounding variables and selection bias
-

zalando



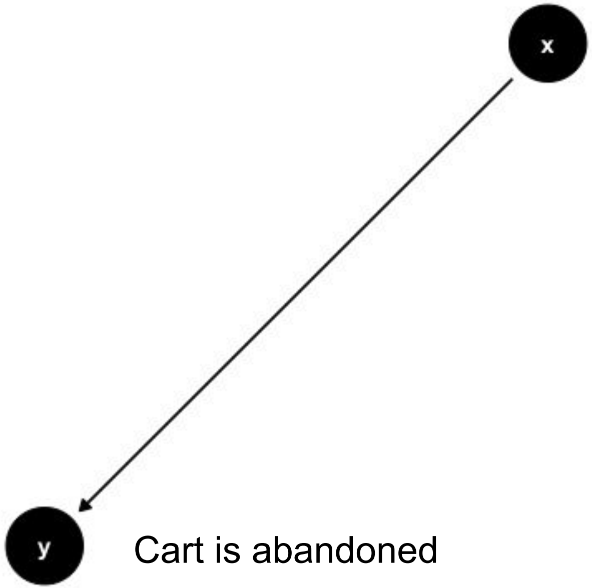
ESTABLISHING CAUSALITY: OBSERVATIONAL METHODS

- Simpson's Paradox
- https://www.youtube.com/watch?v=ebEkn-BiW5k

METHODS

- Directed Acyclic Graphs
- Instrumental Variables Analysis
- Matching

DIRECTED ACYCLIC GRAPHS - DAGS

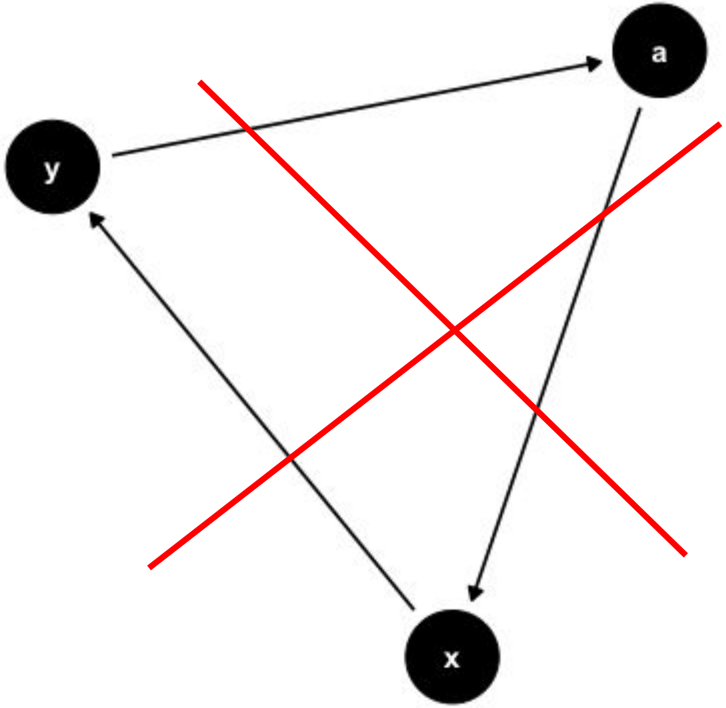


Too many steps in the checkout

Cart is abandoned

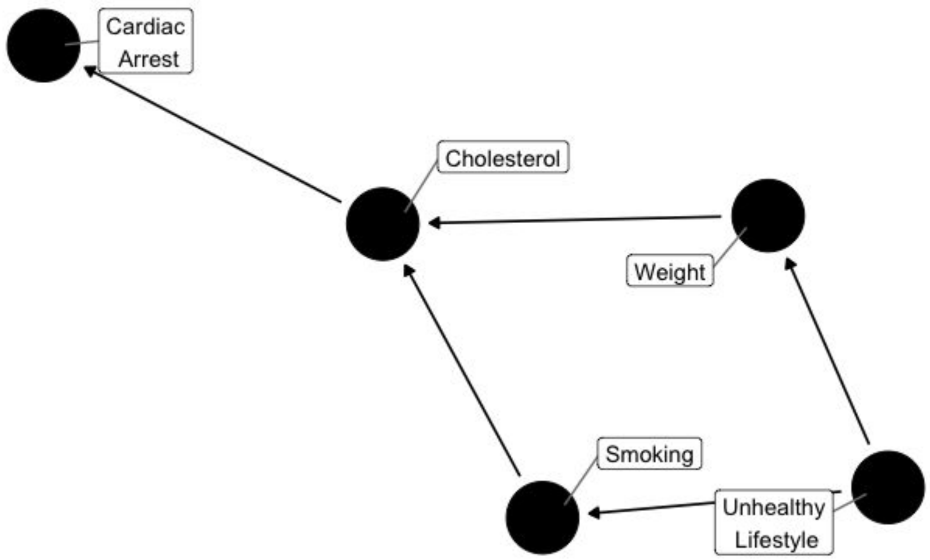
DIRECTED

DIRECTED ACYCLIC GRAPHS - DAGS



ACYCLIC

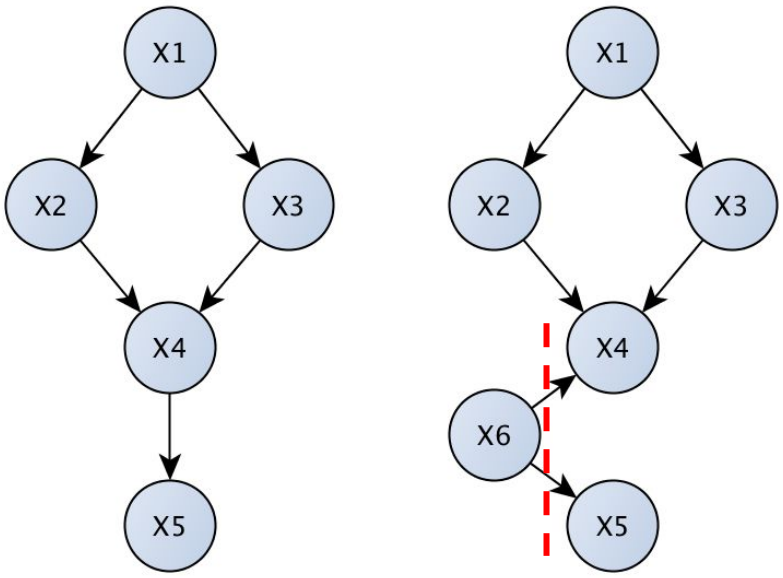
Structural Causal Graphs



- Great tool for conceptualising causal relationships
- Mathematically grounded
- Implies Linear relationships
- R package **ggdag**
- Python package **causality**

zalando

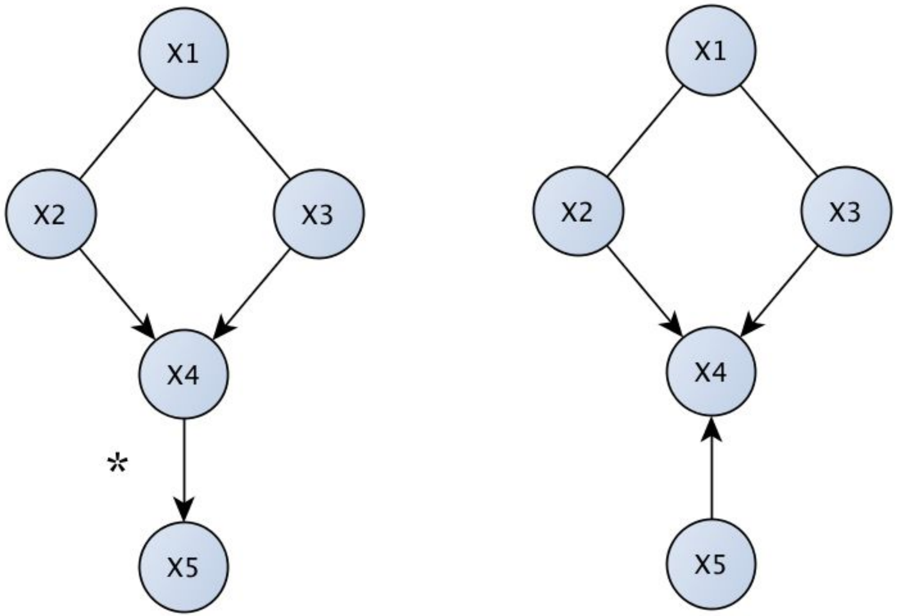
Structural Causal Graphs



- Two models specified here
- Left has genuine causal relationship between x4 and x5
- Right has a spurious correlation between x4 and x5
- They are in fact both caused by a common cause x6 - confounder
- graph inference

https://medium.com/@akelleh/causal-graph-inference-b3e3afd47110

Structural Causal Graphs



- Two models specified here
- Left has genuine causal relationship between x4 and x5
- Right has a spurious correlation between x4 and x5
- They are in fact both caused by a common cause x6 - confounder
- graph inference

https://medium.com/@akelleh/causal-graph-inference-b3e3afd47110

INSTRUMENTAL VARIABLE ANALYSIS

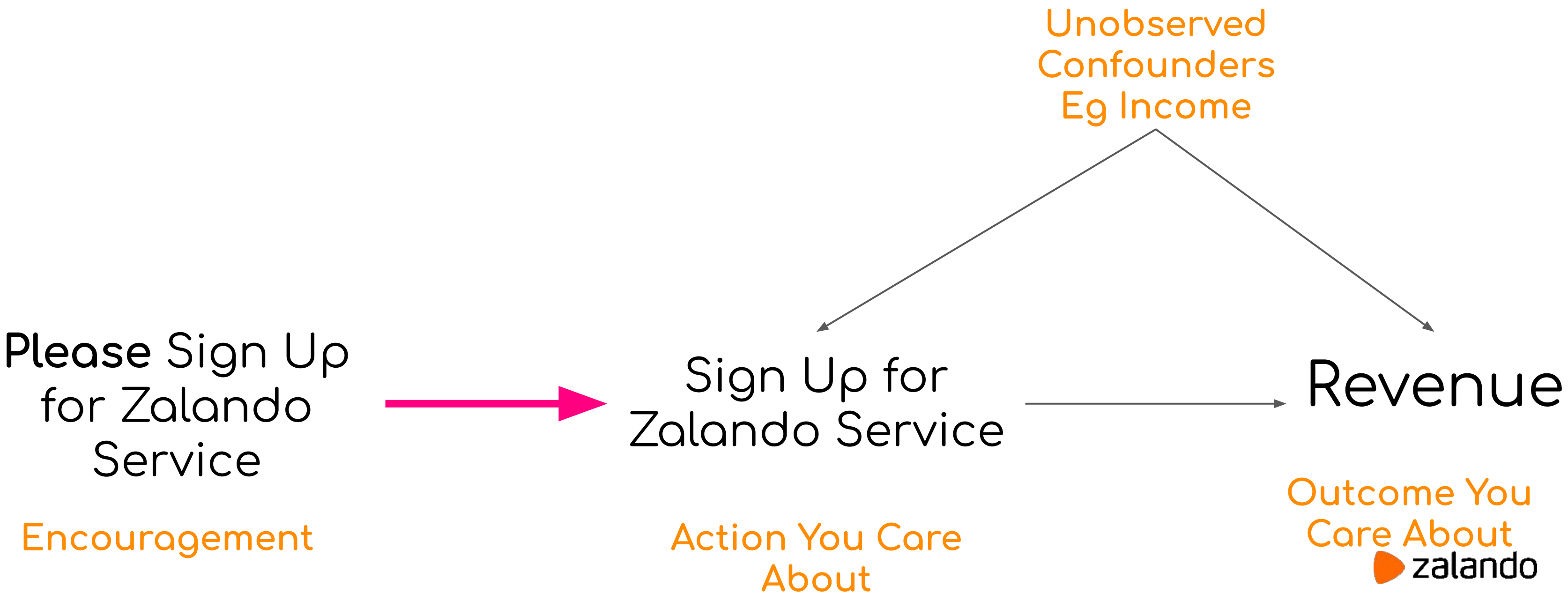
Q: What to do when you can only A/B test something that is associated with the thing you want to test ?

A: Instrumental Variable Analysis - Use a third variable that you can measure that is strongly associated with the one you actually want to measure

INSTRUMENTAL VARIABLE ANALYSIS

Instrumental Variables have three characteristics

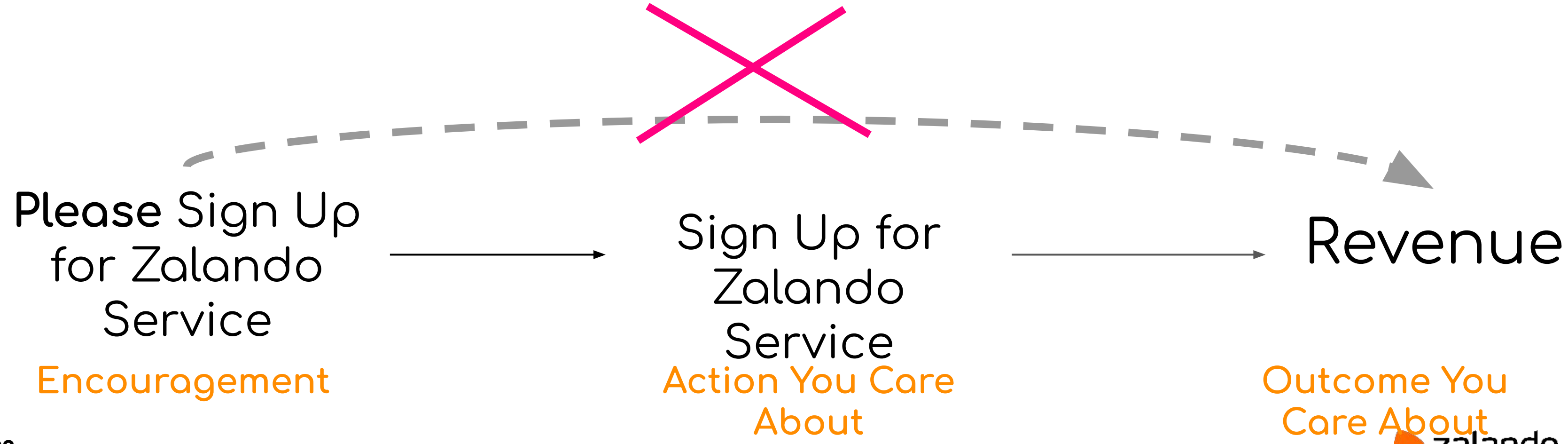
1. Associated/Correlated with variable whose impact we want to understand



INSTRUMENTAL VARIABLE ANALYSIS

Instrumental Variables have three characteristics

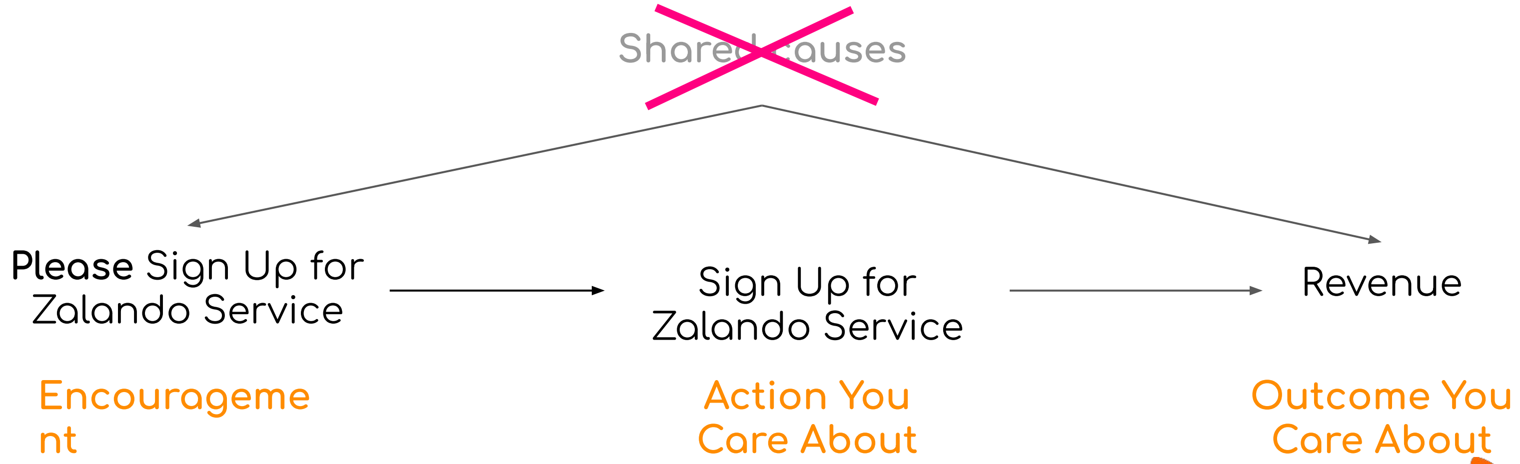
1. Associated/Correlated with variable whose impact we want to understand
2. Does not impact outcome, except via potential effect on the variable we want to understand (exclusion restriction)



INSTRUMENTAL VARIABLE ANALYSIS

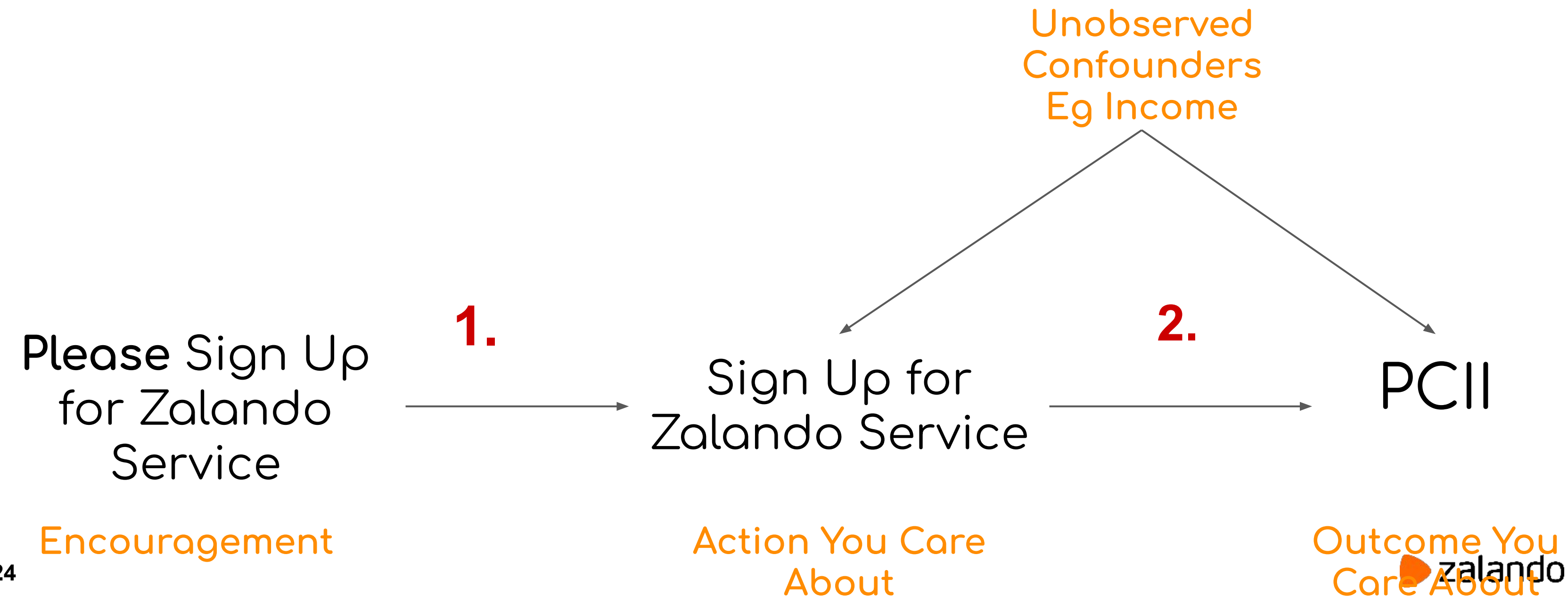
Instrumental Variables have three characteristics

1. Associated/Correlated with variable whose impact we want to understand
2. Does not impact outcome, except via potential effect on the variable we want to understand (exclusion restriction)
3. Outcome and instrument do not share causes



INSTRUMENTAL VARIABLE ANALYSIS

Instrumental Variables Analysis via Two Stage Least Squares



First Stage

$$\text{Sign up} = \alpha + \beta \text{Encouragement}$$

Second Stage

$$\text{Revenue} = \alpha + \beta \widehat{\text{Sign up}}$$

Prediction from first stage

INSTRUMENTAL VARIABLE ANALYSIS

Instrumental Variables Analysis via Two Stage Least Squares

- R package: AER package ivreg method
- Python package: Linear Models package IV2SLS method
- Book: Pearl & Mackenzie 2018, 249

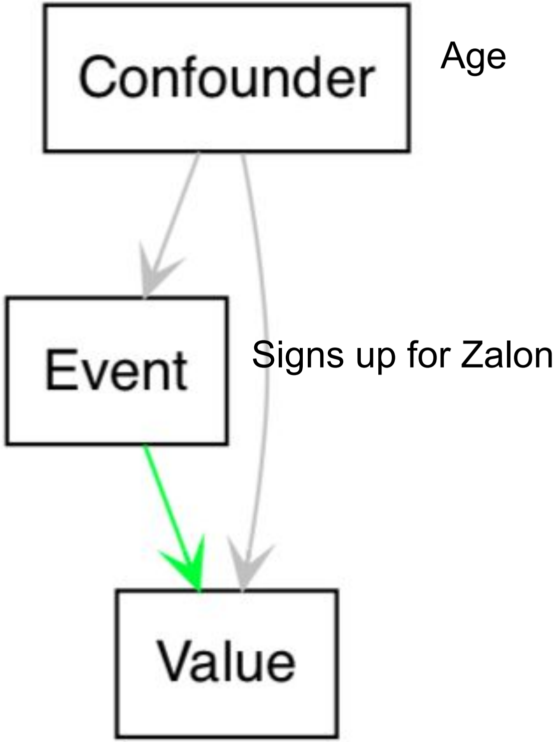
Limitations:

- Instrument must be justified - have a strong association with the variable of interest
- Instrument must not have direct impact on the outcome
- If you misuse instrumental variables, you can get even more biased results than if you had not used them at all.

MATCHING

- A method to estimate causality without an experiment, or in the presence of a “Natural” experiment
- E.g. Launch of a new product

MATCHING



- Must be able to measure the confounder
- Naive estimate of the Event and the Value is biased due to the confounder
- Instead measure all relationships
- “Control” for Age
- “Matching Estimator”
- https://cran.r-project.org/web/packages/Matching/Matching.pdf

SUMMARY

- Correlation is not Causation
- Understanding cause is important for business decision making
- Can still make accurate predictions without knowing cause
- Controlled Experiments are the Gold Standard
- Can still estimate cause from Observational Data
- Methods:
    - Directed Acyclic Graphs
    - Instrumental Variables
    - Matching

zalando

THANK YOU

QUESTIONS?

Reference Books

Causality: Models, Reasoning and Inference (2009) Judea Pearl

The Book of Why: The New Science of Cause and Effect (2018) Pearl & Mackenzie

Reference Blogs

https://medium.com/causal-data-science Adam Kelleher

My Github & Slides

https://github.com/alicelynch/meetup-talks