# Getting Started in R & Causal Relationships

# What is R

- R is a programming language & a program
- R is free and widely used
- There are lots of functions and programs that are already written in R that we can use for free
- We will run regressions, calculate means, plot data, etc.

# Some examples of maps

- https://www.youtube.com/watch?v=adm3RB4ieXU
- https://www.youtube.com/watch?v=I9aLRsMTk_o

# The Fundamental Problem of Causal Inference

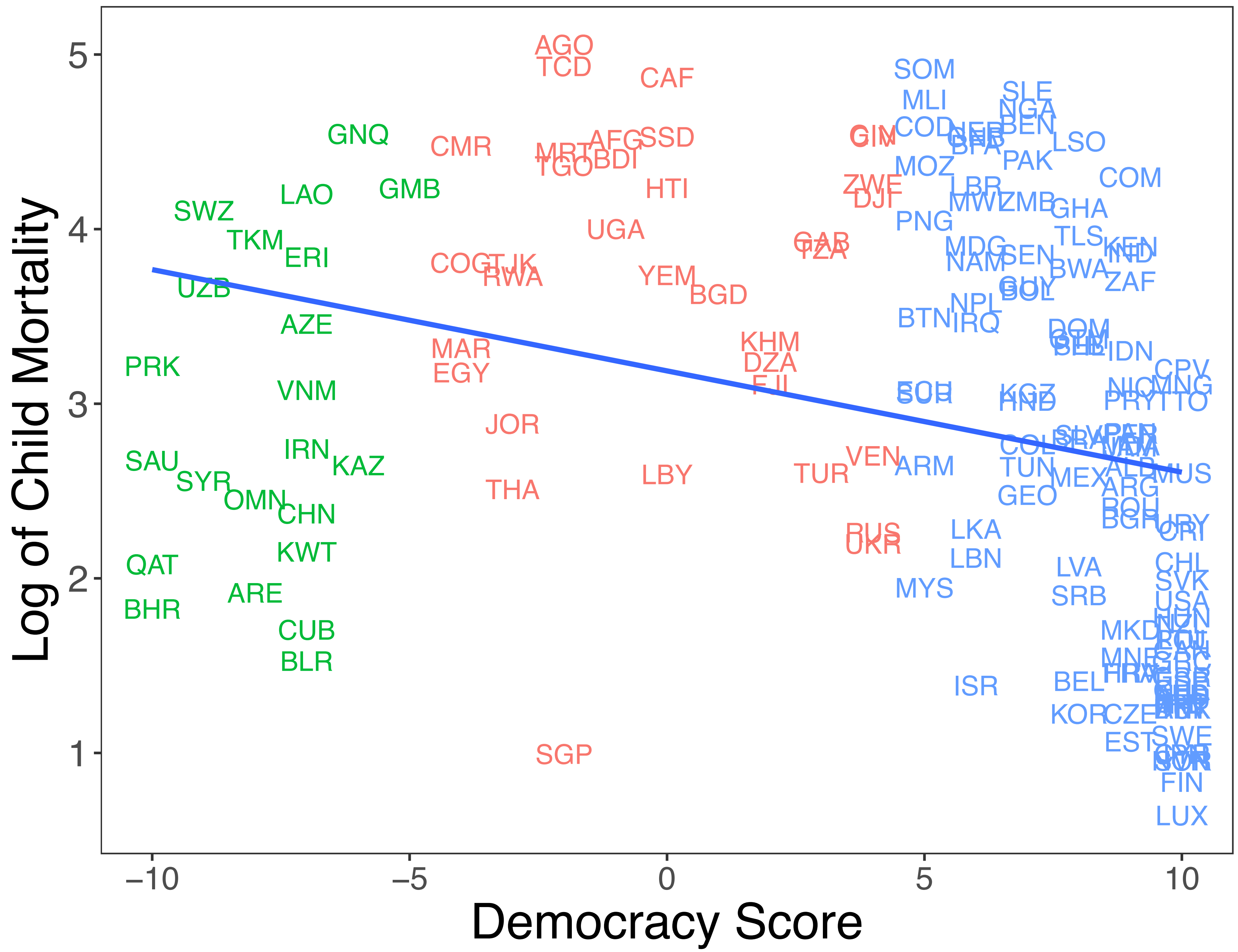
- Many (most) of the questions we are interested in answering in political science are causal questions

- What is the effect of X in Y?

- How much does Y change if X changes by 1?

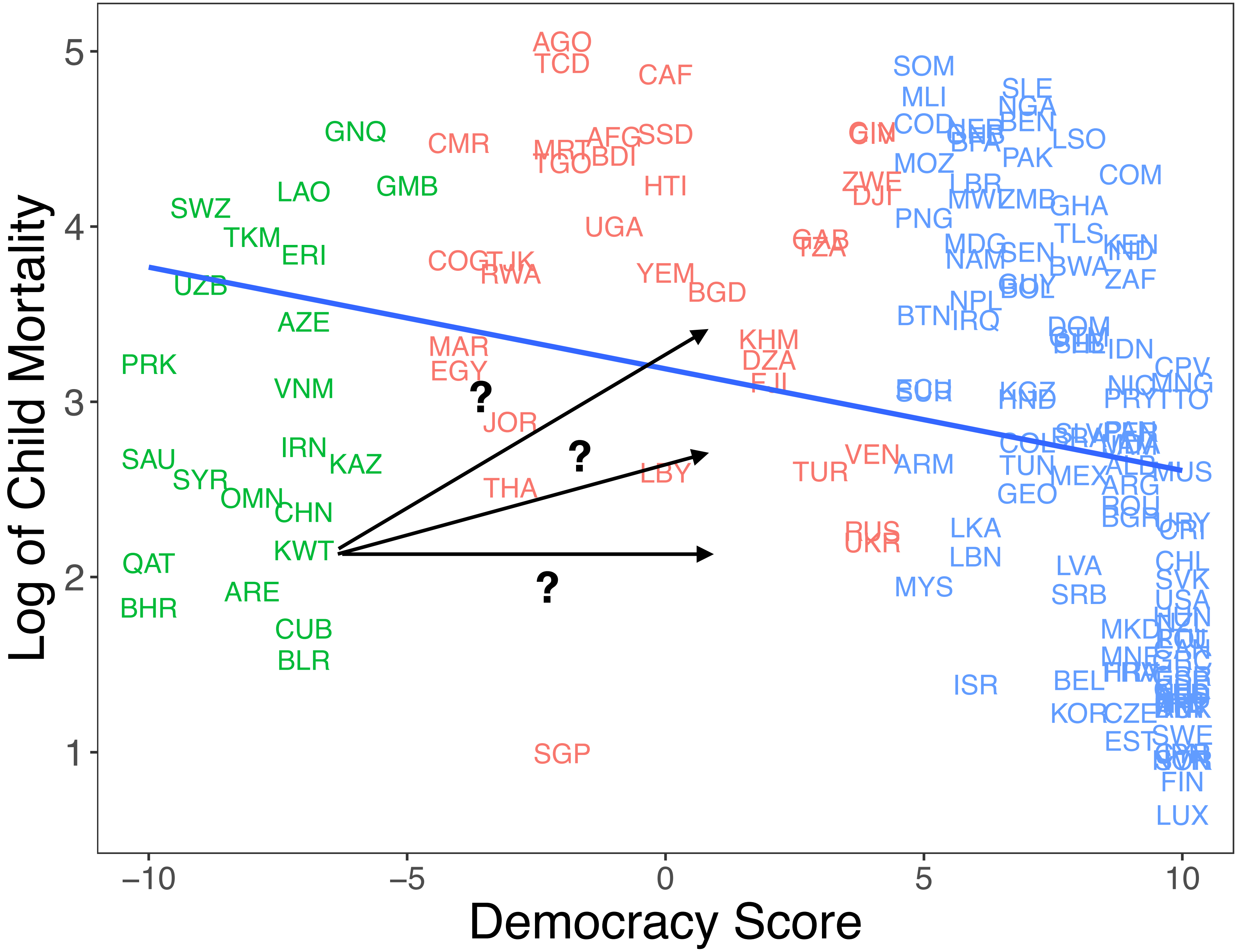
Think of the causal effect as the difference between what happened and what could have happened with/without a “treatment” (or change in X)

How do we measure the causal effect?




# Example

Democracy and child mortality




# What if Kuwait was (more of) a democracy?

Independent (explanatory) Variable

Dependent (outcome) Variable

$$X \longrightarrow Y$$

**Generally we think that for X to be a cause of Y they have to go together**

**If X happens and causes Y, then Y should also occur**

$$X \longrightarrow Y$$

Generally we think that for X to be a cause of Y they have to go together

If X happens and causes Y, then Y should also occur

In statistics speak, if X and Y occur together, they are correlated

**In statistics speak, if X and Y occur together, they are correlated**

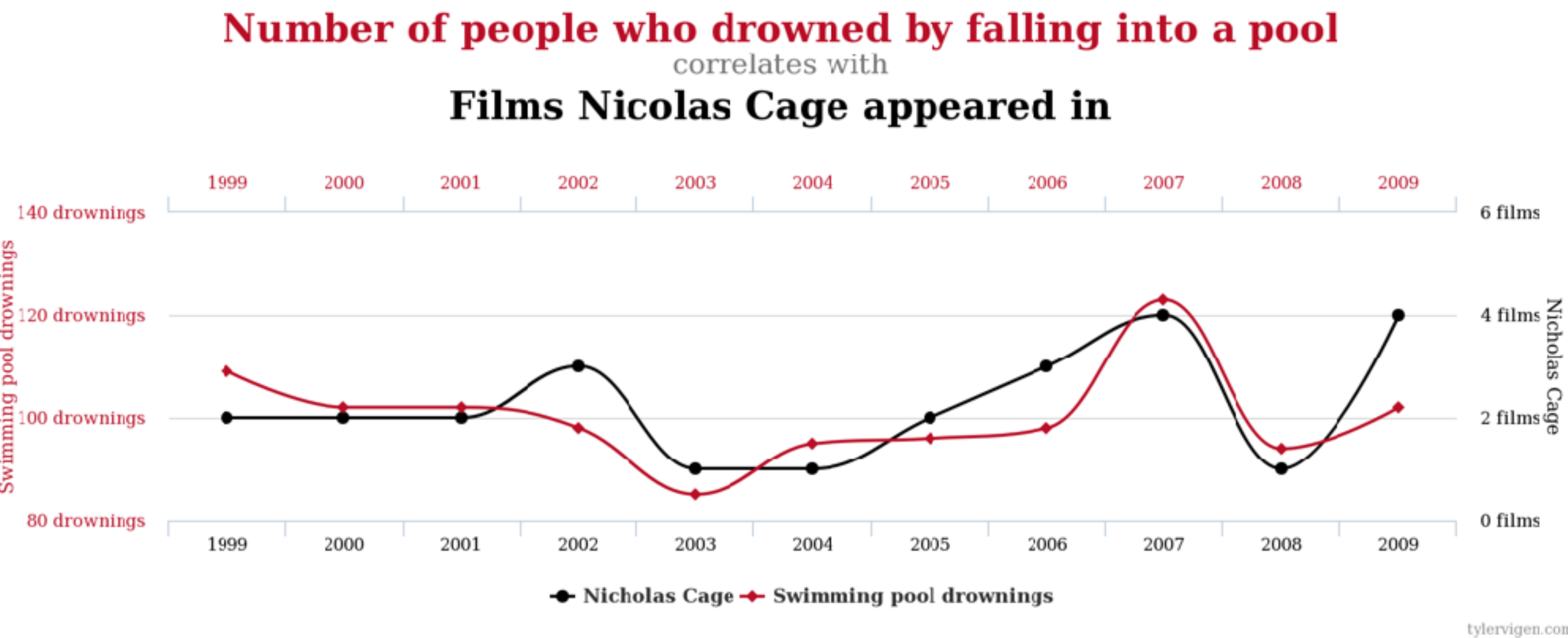
**if X goes up, Y goes up -> positive correlation**

**if X goes up, Y goes down -> negative correlation**

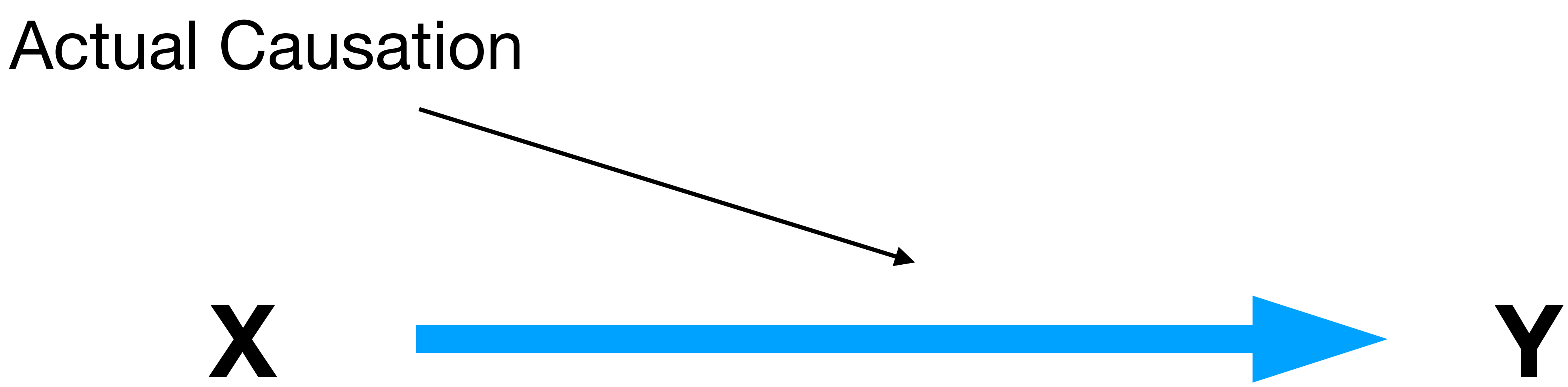
**But does X and Y being correlated mean X causes Y?**

**But does X and Y
being correlated mean X causes Y?**

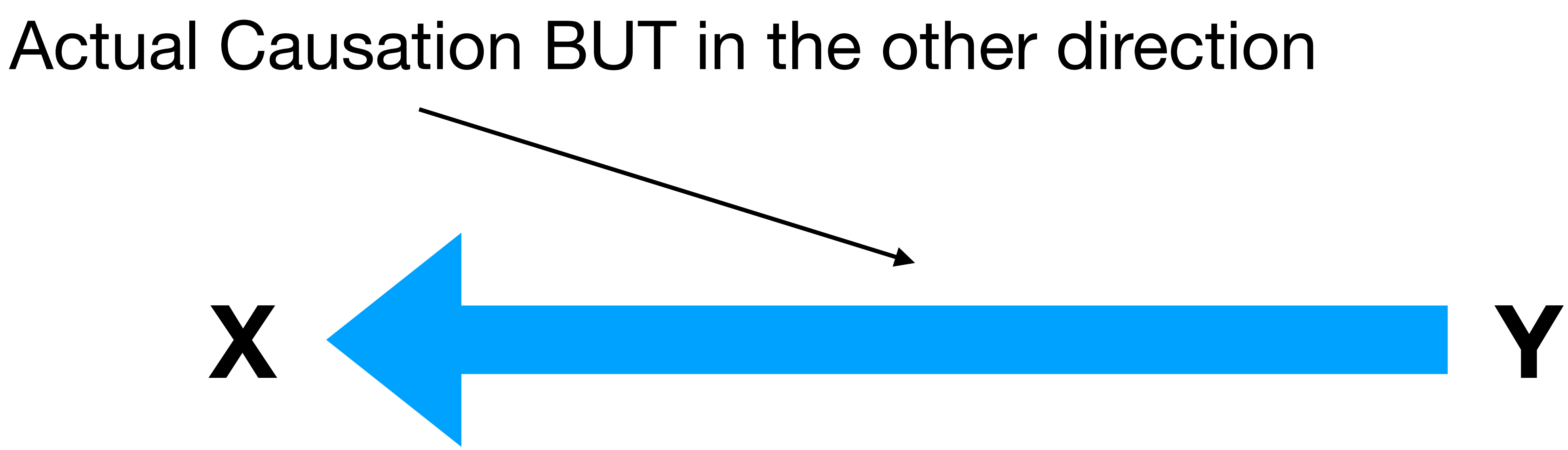
NO

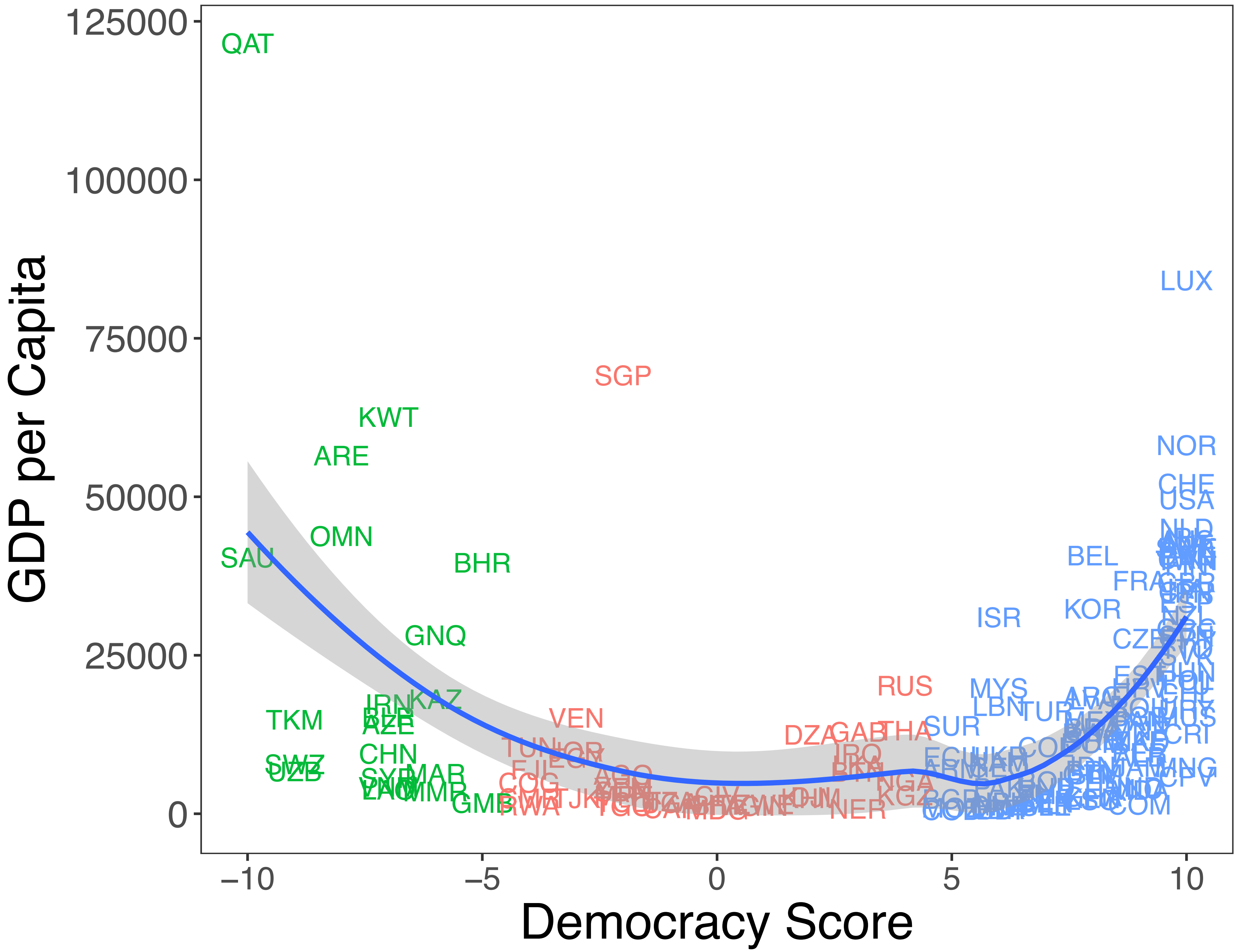

Number of people who drowned by falling into a pool
correlates with
Films Nicolas Cage appeared in
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
140 drownings
120 drownings
100 drownings
80 drownings
Swimming pool drownings
6 films
4 films
2 films
0 films
Nicholas Cage
Nicholas Cage
Swimming pool drownings
tylervigen.com

# Possible Causes of Correlations between X & Y



# Possible Causes of Correlations between X & Y



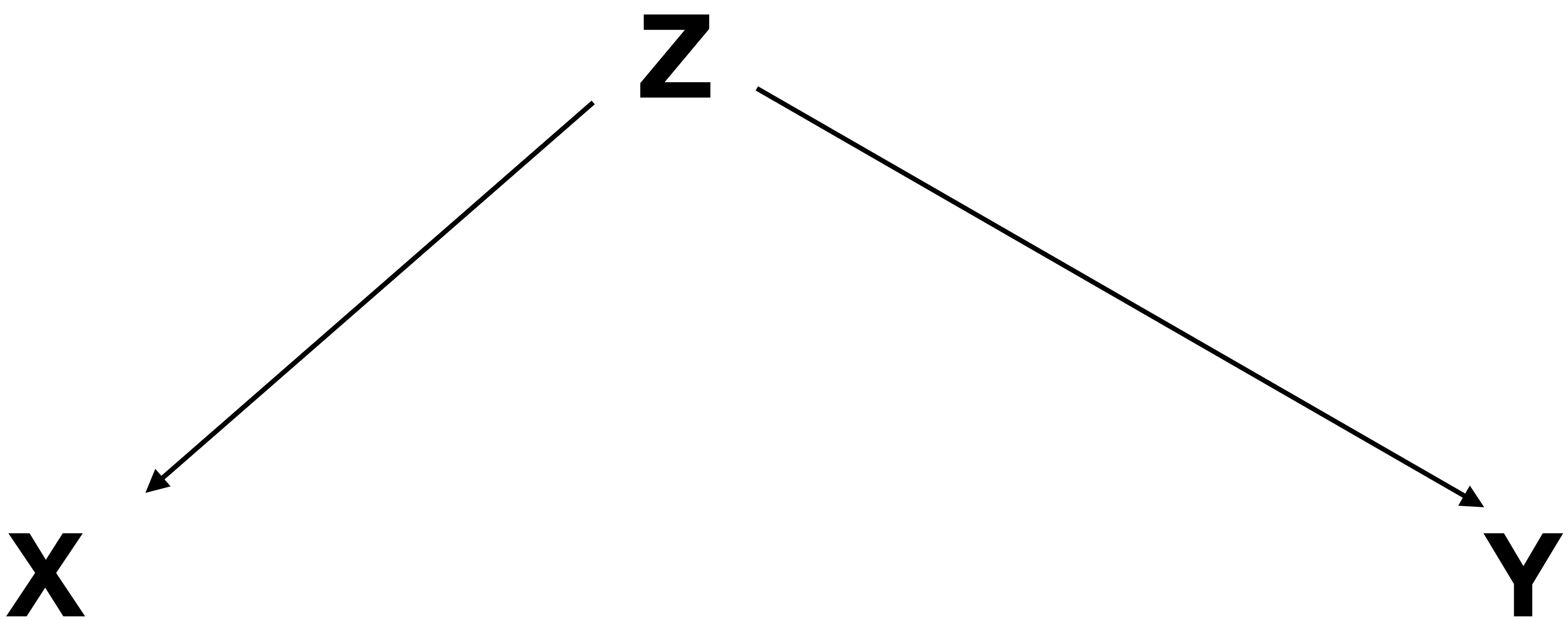



**How can we know whether democratic countries grow faster?**

What if richer countries are more likely to transition to democracy?

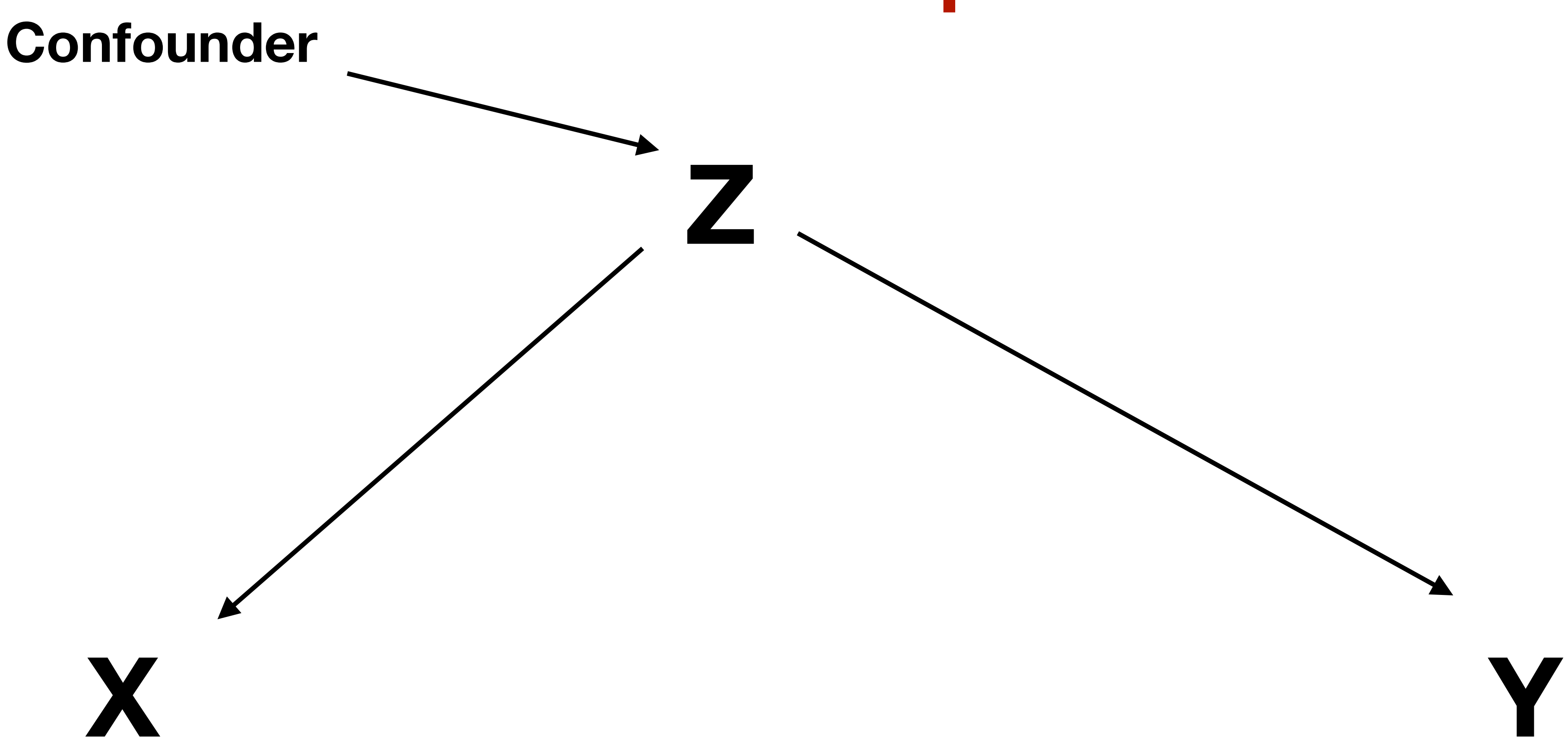
# Possible Causes of Correlations between X & Y

**Spurious correlation**



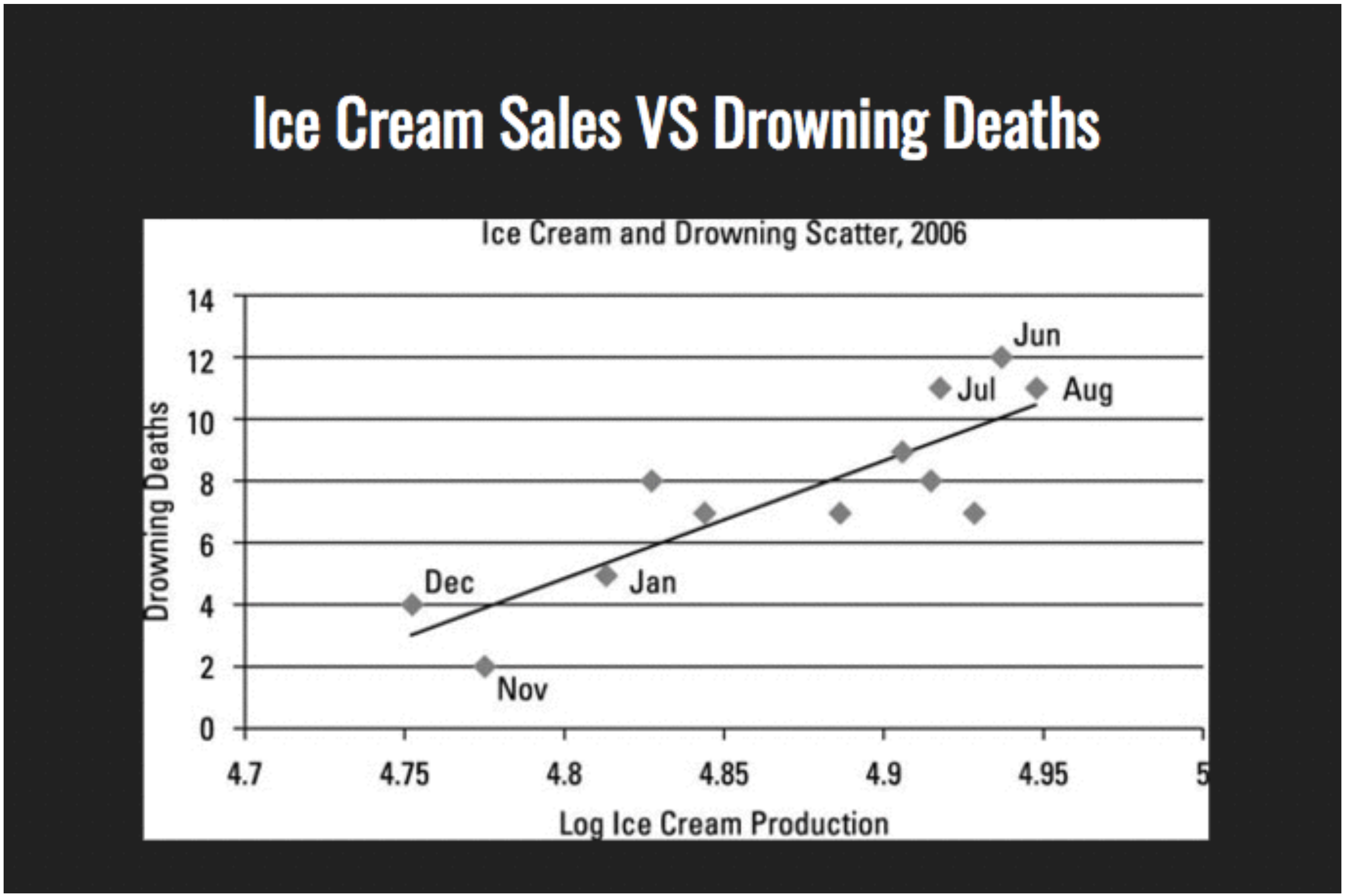
# Possible Causes of Correlations between X & Y

**Spurious correlation**




**The confounder Z, causes both X and Y**

# Example of Spurious Correlation



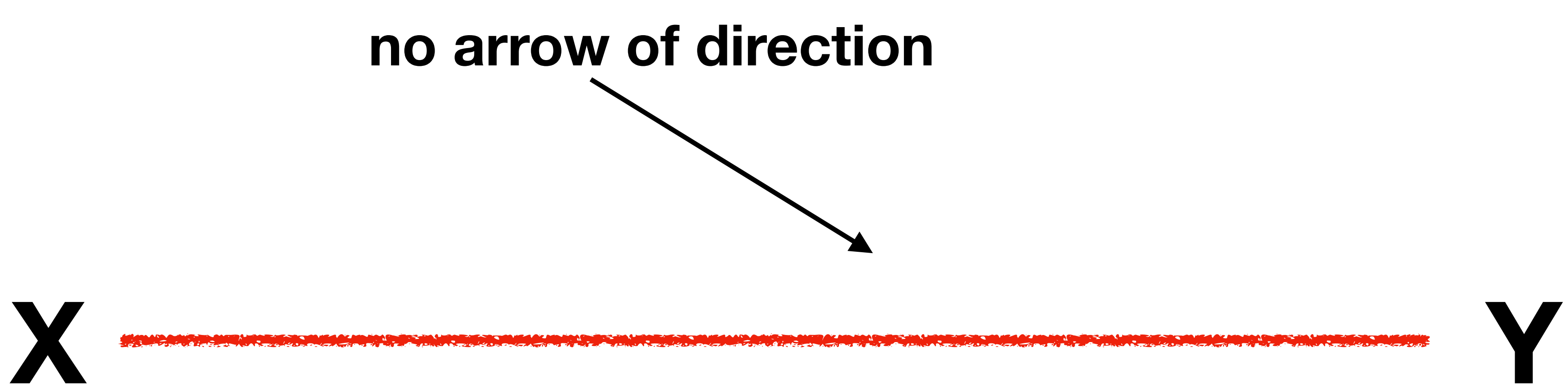
# Possible Causes of Correlations between X & Y

**Chance**

**X** ——————————— **Y**

# Possible Causes of Correlations between X & Y

## Chance

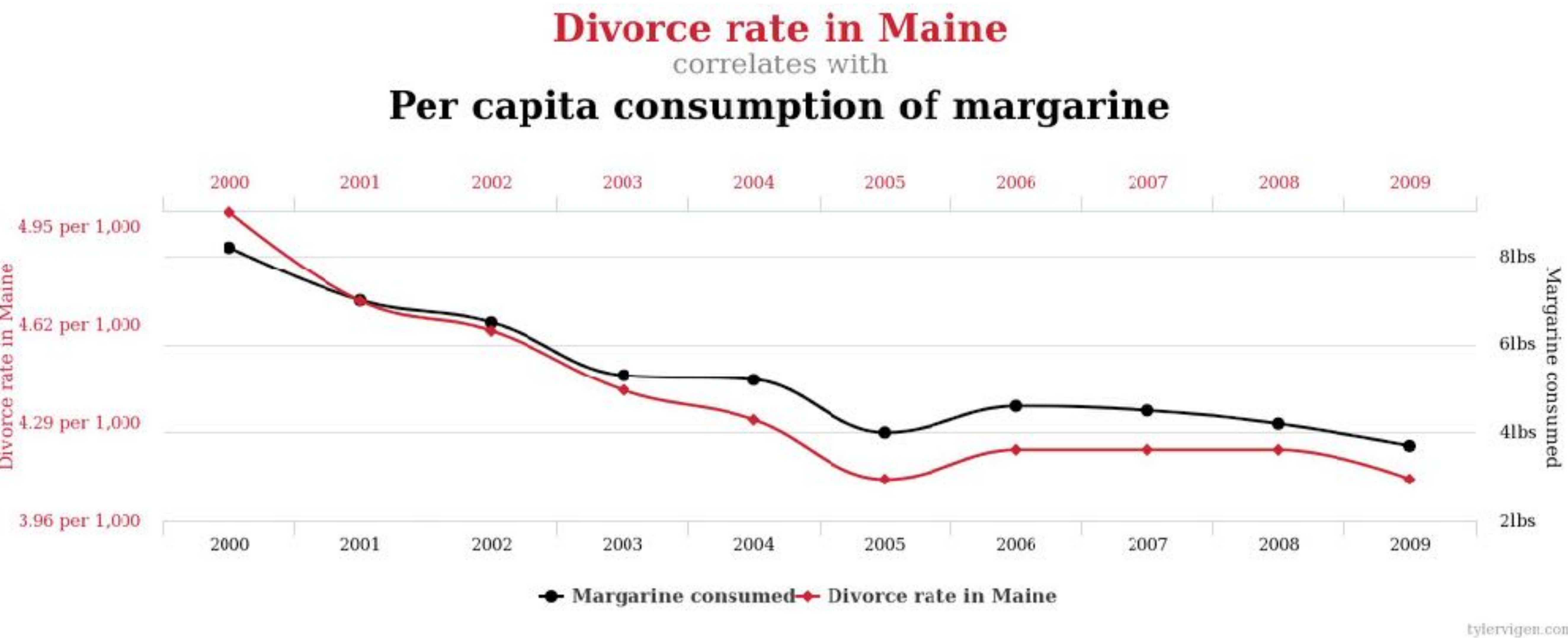


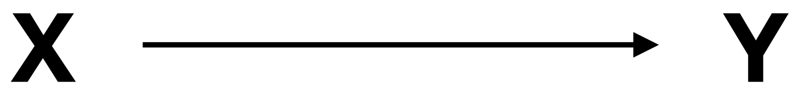

**x and y occur at similar times/similar instances by random chance**

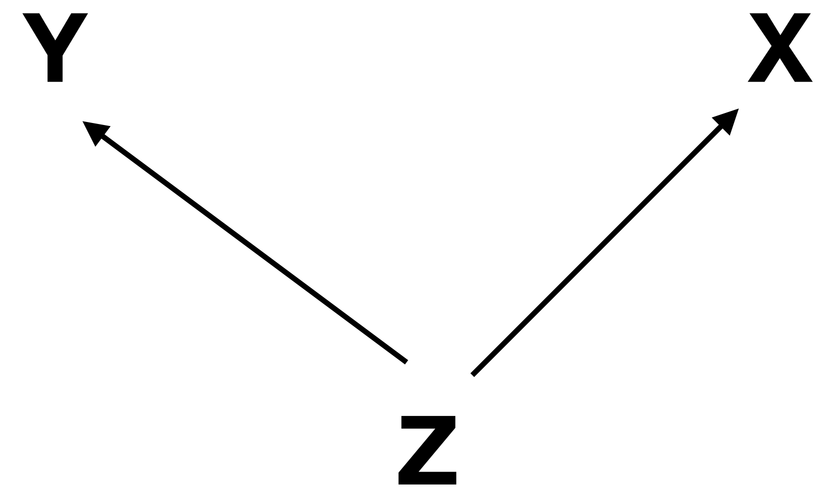
# Possible Causes of Correlations between X & Y

## Chance

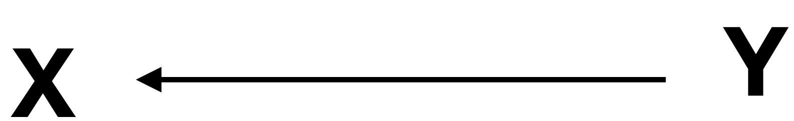
**sometimes thins are correlated due to random chance, even though no systematic relationship exists**


Divorce rate in Maine
correlates with
Per capita consumption of margarine
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
4.95 per 1,000
4.62 per 1,000
4.29 per 1,000
3.96 per 1,000
Divorce rate in Maine
8lbs
6lbs
4lbs
2lbs
Margarine consumed
Margarine consumed
Divorce rate in Maine
tylervigen.com

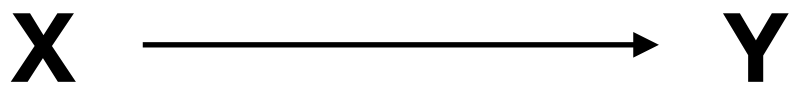


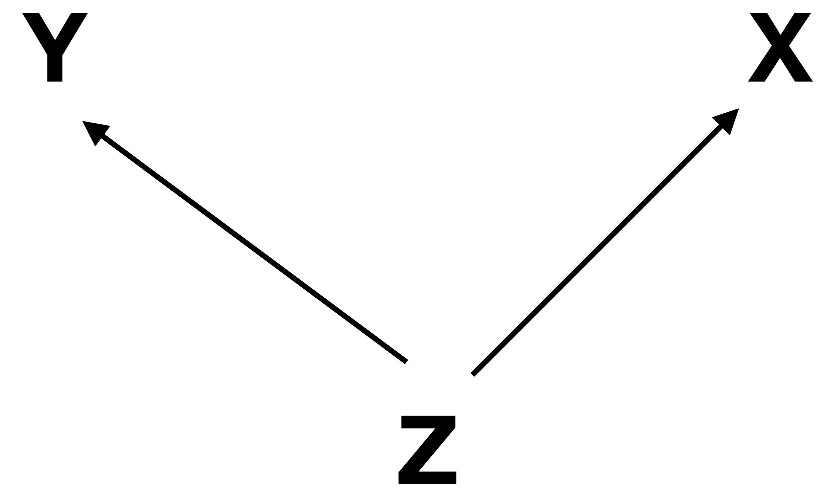

# Correlations




## Causality

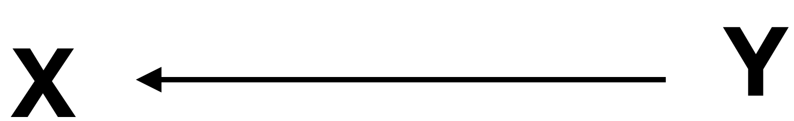



## Spuriousness




# Correlations

## Reverse Causality




## Random Chance

How do we find out if X causes Y?

How do we rule out chance, reverse causality, and spuriousness?

# Theoretical Model

- One way to refute claims of reverse causality or spuriousness is with a strong theoretical model
- Does it even make sense that Y could cause X?
- Are there other variables that could be confounders?
- It is your task as a scholar to make a case against these claims

# Controls

- To determine causal effects of X on Y we like to “control” for all other possible factors
- We want to compare very similar groups, that only differ on X:
    - if everything else is the same except for X, then only X can be the causal factor

# Randomization as the gold standard of causal inference

**Experiments are research designs in which the researchers controls and randomly assigns values of the independent variable to subjects.**

Can reverse causality occur in an experiment?

if not, why not?

**Can reverse causality occur in an experiment?**

the treatment is assigned randomly -> ergo independent of Y

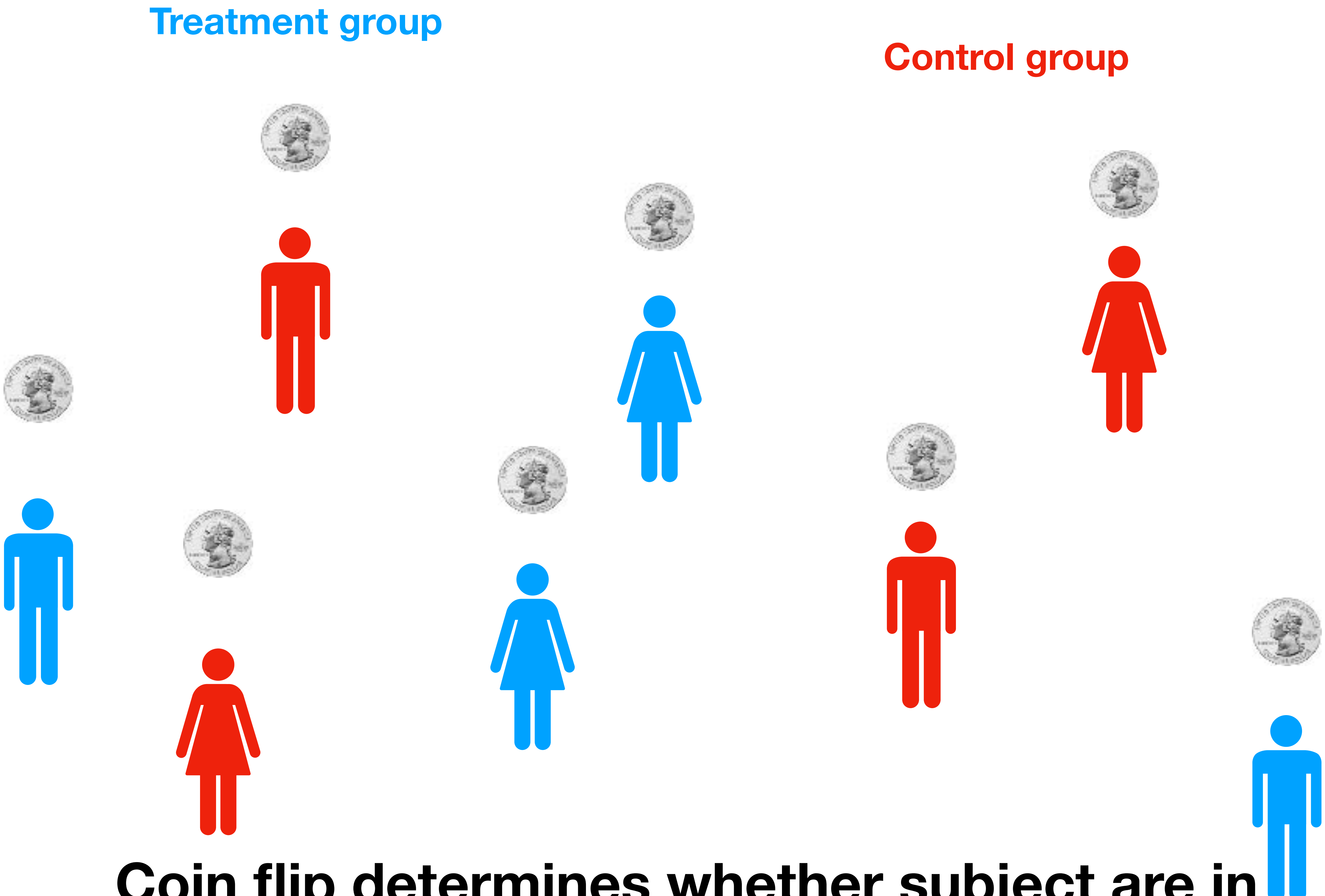
## Can confounding occur in an experiment?

# Can confounding occur in an experiment?

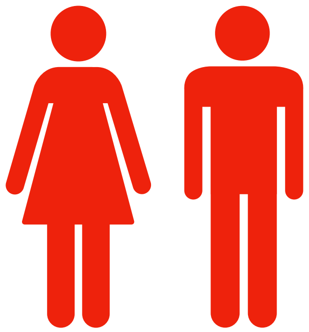
**Not if done correctly**

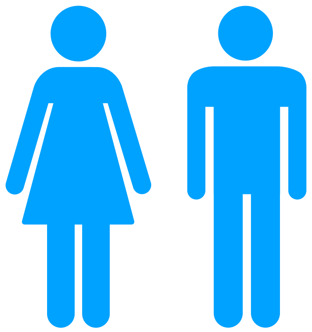
random assignment into treatment and control groups means we, by design, control ***for all possible confounding*** factors

In a **randomized controlled trial (RCT)**, each unit is randomly assigned either to the treatment or control group. The randomization of treatment assignment guarantees that the average difference in outcome between the treatment and control groups can be attributed solely to the treatment because the two groups are on average identical to each other in all other pre-treatment characteristics.




**Coin flip determines whether subject are in treatment or control group**





**Because assignment to each group is random, in expectation, the groups should be very similar or the same**

- But if subjects know they are in the treatment group, they might behave differently

- That is why we use placebos, i.e. subjects don't know which group they are in (single blind)

- Similarly, researchers might bias results/measurements if they know which subjects are in which group

- Double-blind: neither subject nor evaluators know who is in control/treatment group

## Example from polisci

# Causal effect of intergroup contact on exclusionary attitudes

Ryan D. Enos[1]

The effect of intergroup contact has long been a question central to social scientists. As political and technological changes bring increased international migration, understanding intergroup contact is increasingly important to scientific and policy debates. Unfortunately, limitations in causal inference using observational data and the practical inability to experimentally manipulate demographic diversity has limited scholars' ability to address the effects of intergroup contact. Here, I report the results of a randomized controlled trial testing the causal effects of repeated intergroup contact, in which Spanish-speaking confederates were randomly assigned to be inserted, for a period of days, into the daily routines of unknowing Anglo-whites living in homogeneous communities in the United States, thus simulating the conditions of demographic change. The result of this experiment is a significant shift toward exclusionary attitudes among treated subjects. This experiment demonstrates that even very minor demographic change causes strong exclusionary reactions. Developed nations and politically liberal subnational units are expected to experience a politically conservative shift as international migration brings increased intergroup contact.

The experiment leveraged the tendency for commuters to ride the same train every day. I treated certain trains by assigning pairs of Spanish-speaking persons to visit the same train stations at the same time every day. Within each train station, these experimental confederates were the same persons every day. Other trains were randomly assigned to the control condition and had no intervention at the stations. Subjects were surveyed about their socio-political attitudes before and after the treatment. With this design, subjects were exposed to the same Spanish-speaking persons in a location near their homes for an extended period, as would be the situation if immigrants had moved into their neighborhood and used the public transportation. With this design, I experimentally manipulated the conditions of demographic change and, by comparing changes in survey responses before and after the treatments, I identified the effect of exposure to these Spanish-speaking persons.

| Condition | Control | Treatment |
|---|---|---|
| Liberal† | 0.47 | 0.47 |
| Republican | 0.17 | 0.19 |
| Obama disapprove | 0.27 | 0.29 |
| Ride MBTA every day | 0.85 | 0.90 |
| Voted 2010 | 0.77 | 0.66 |
| Romney voter | 0.24 | 0.22 |
| Hispanic threat | 0.06 | 0.05 |
| Age | 44.66 | 40.43 |
| Residency year | 8.22 | 7.07 |
| College | 0.89 | 0.86 |
| Male | 0.60 | 0.60 |
| Hispanic | 0.03 | 0.05 |
| White | 0.91 | 0.83 |
| Income | 146,236 | 140,103 |
| *n* | 117 | 103 |

Treated subjects were far more likely to advocate a reduction in immigration from Mexico and were far less likely to indicate that illegal immigrants should be allowed to remain in this country. The ATEs and associated SEs allow me to reject the Null Hypothesis of no effect with a high degree of confidence. The ATE on favoring English as an official language, although in the same exclusionary direction, is smaller and does not allow me to reject the Null Hypothesis. However, baseline rates for this question are considerably higher (0.610, 0–1 scale) than for the other questions, indicating relatively high support for English as an official language, regardless of treatment.

# Problems with RCTs

- Limits the questions we can ask and theories we can test
- low degree of external validity (generalization)
- try to learn about generalization through replication
- does the treatment apply the same in the real world?

# Problems with RCTs

- Ethical questions
- It is hard to compare different causal effects
- Hawthorne effect - do people behave differently because they are watched?
- Costs