Poli-301:

The Political Science Discipline

2

Dummy variables

Brief detour on dummy variables

Some categorical variables just take on 2 values

Ex: old/young, rich/poor, white/non-white, etc.

We typically represent these as 0/1 and call them *dummies*

Dummy variables in regression

$$\text{score} = 4.28 - 0.13(\text{rank}_{\text{tenure track}}) - 0.15(\text{rank}_{\text{tenured}}) + \epsilon$$

This is what R does with categorical variables in regression!

Tenure = 3 dummy variables, tenure-track (1/0), tenured (1/0), teaching (1/0)

Intuition and coding dummies

$$\text{score} = 4.28 - 0.13(\text{rank}_{\text{tenure track}}) - 0.15(\text{rank}_{\text{tenured}}) + \epsilon$$

```r
evals %>%
  mutate(old = case_when(age > 50 ~ 1,
                         TRUE ~ 0)) %>%
  select(age, old)
```

```
# A tibble: 463 x 2
     age   old
   <int> <dbl>
 1    36     0
 2    36     0
 3    36     0
 4    36     0
 5    59     1
 6    59     1
 7    59     1
 8    51     1
 9    51     1
10    40     0
```

TRUE ~ 0:
"everything that's not NA *and* doesn't meet the above condition, set to 0"

Or check out ifelse() or dummies package

3

Intro to Causality

And now…

1st half:
how to program + useful concepts

2nd half:
"How do we know if x causes y?"

Does School Suspension Work?

Between 2013-2014, 2.6 million public school students received at least one suspension

Suspension used to punish bad behavior, but it might exacerbate the problem

How do we know if suspensions —> crime?

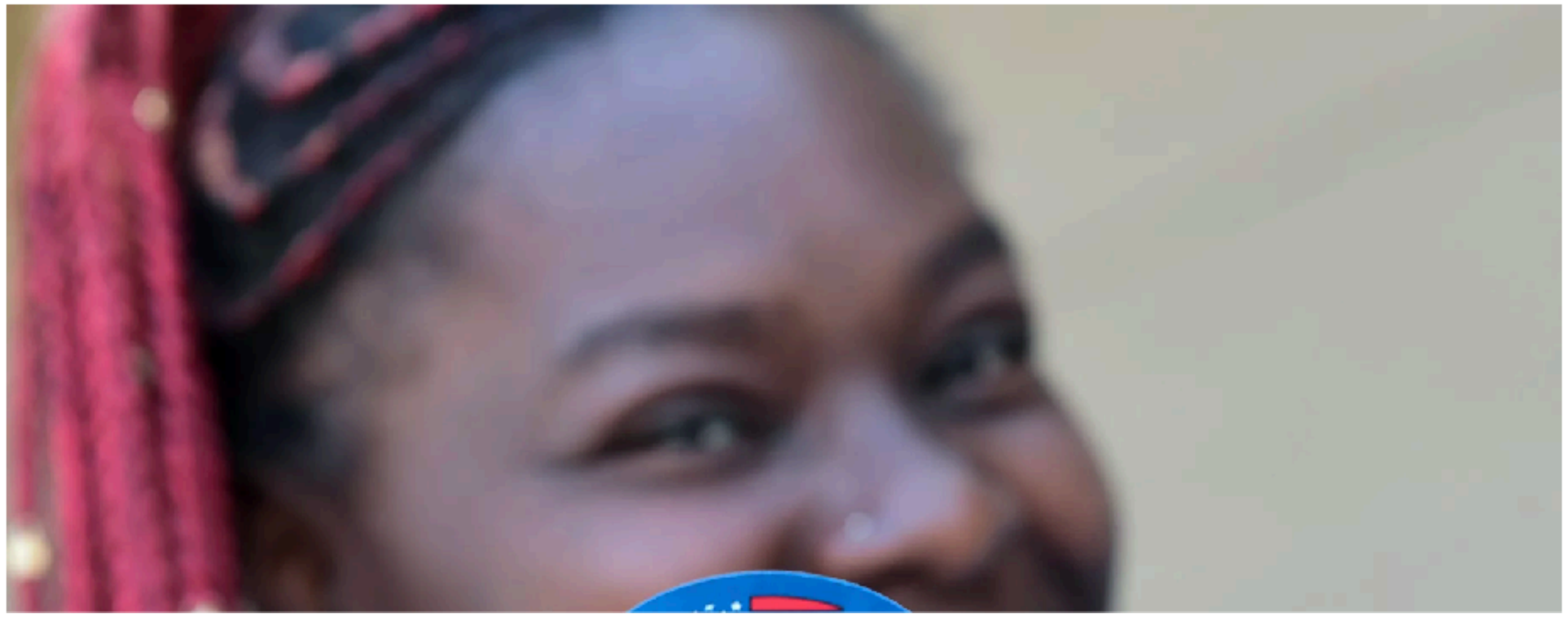
Do voter-ID laws suppress voter turnout?

A new study finds voter ID laws don't reduce voter fraud — or voter turnout

The laws don't seem to do what critics fear or proponents hope.

By German Lopez | @germanrlopez | german.lopez@vox.com | Feb 21, 2019, 8:00am EST

SHARE



Do minimum wage laws reduce employment?

82,514 views | Sep 28, 2018, 11:47am

How Higher Minimum Wages Impact Employment

Adam Millsap Contributor

Policy

I write about state and local policy and urban economics.

Data and causality

Most of the interesting questions we want to answer with data are causal

Some aren't:

Facebook might want to know: "is there a person in this photo?"

But not care about what factors *cause* the picture to be a photo of a person

Depends on the question; nearly every **WHY** question is causal

Value of causality

This is one of our comparative advantages

Not just academic

Companies, governments, international organizations need to answer “**WHY**” questions

Does this policy work or not? Did it do what was intended? How (in)effective was it?

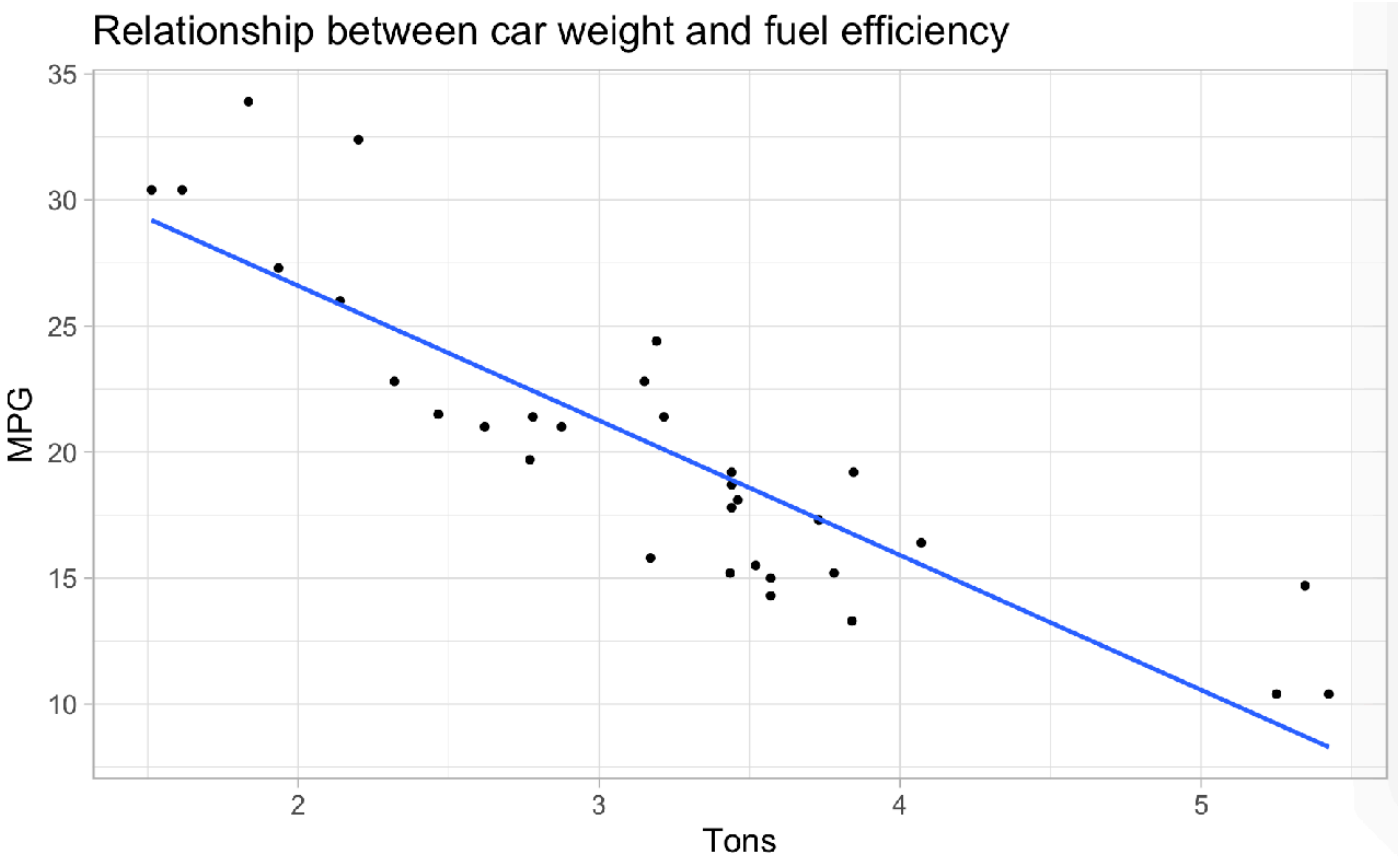
What is causality?

In this class we say ***X causes Y if…***

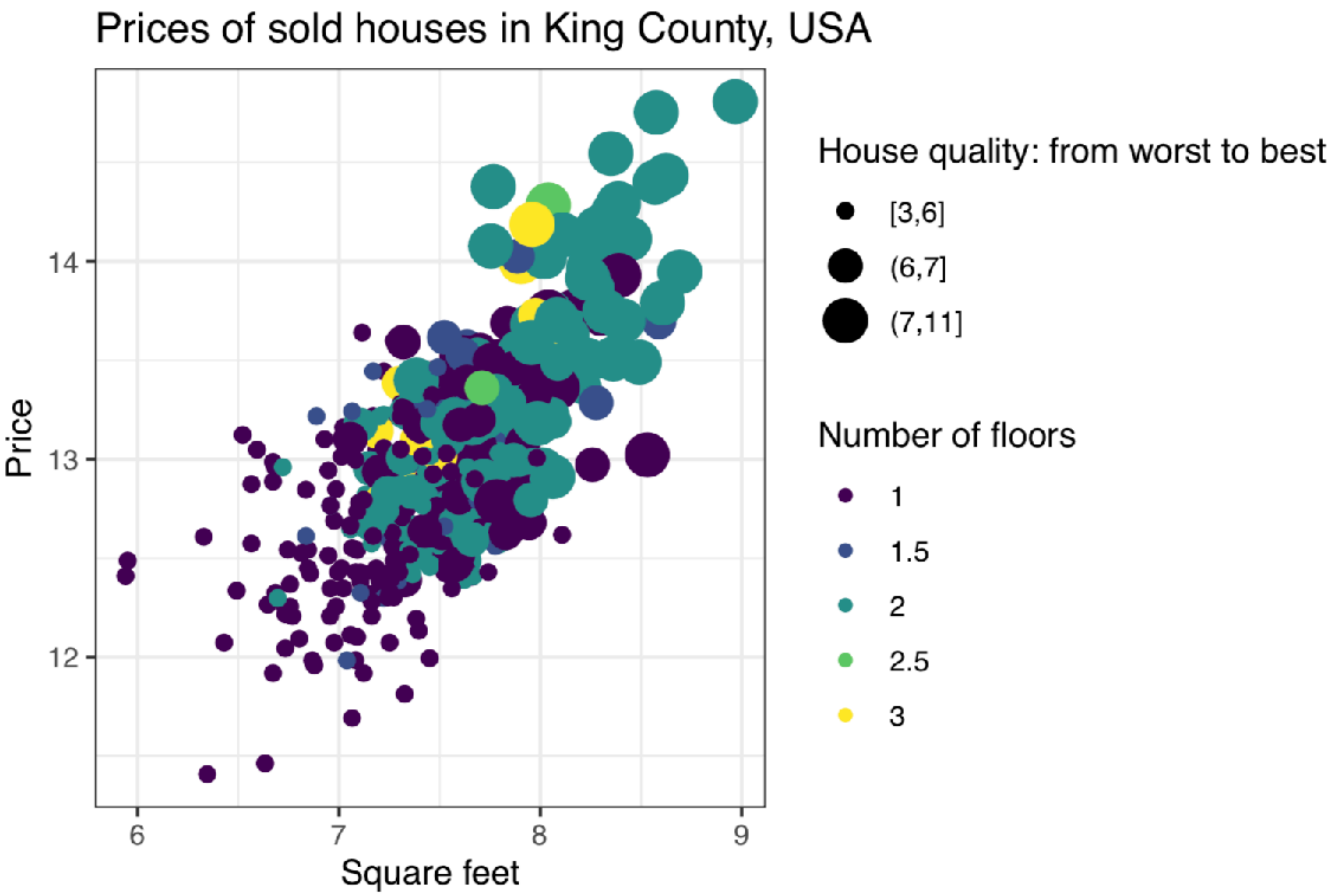
An intervention that **changes** the value of X (without changing anything else)

Produces a **change** in Y

Obviously causal relationships



Obviously causal relationships



Not obvious:

Data on display

Measuring the value of education

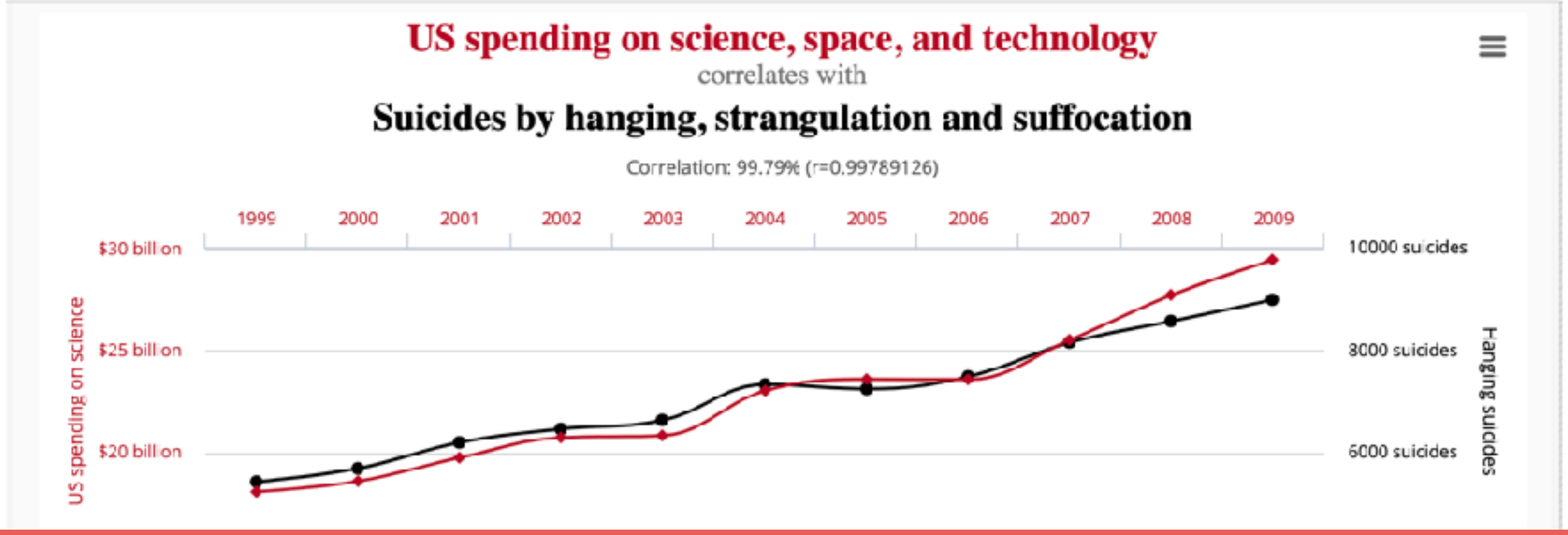
Elka Torpey | *April 2018*

It's hard to quantify the full value of an education. But U.S. Bureau of Labor Statistics (BLS) data consistently show that, in terms of dollars, education makes sense.

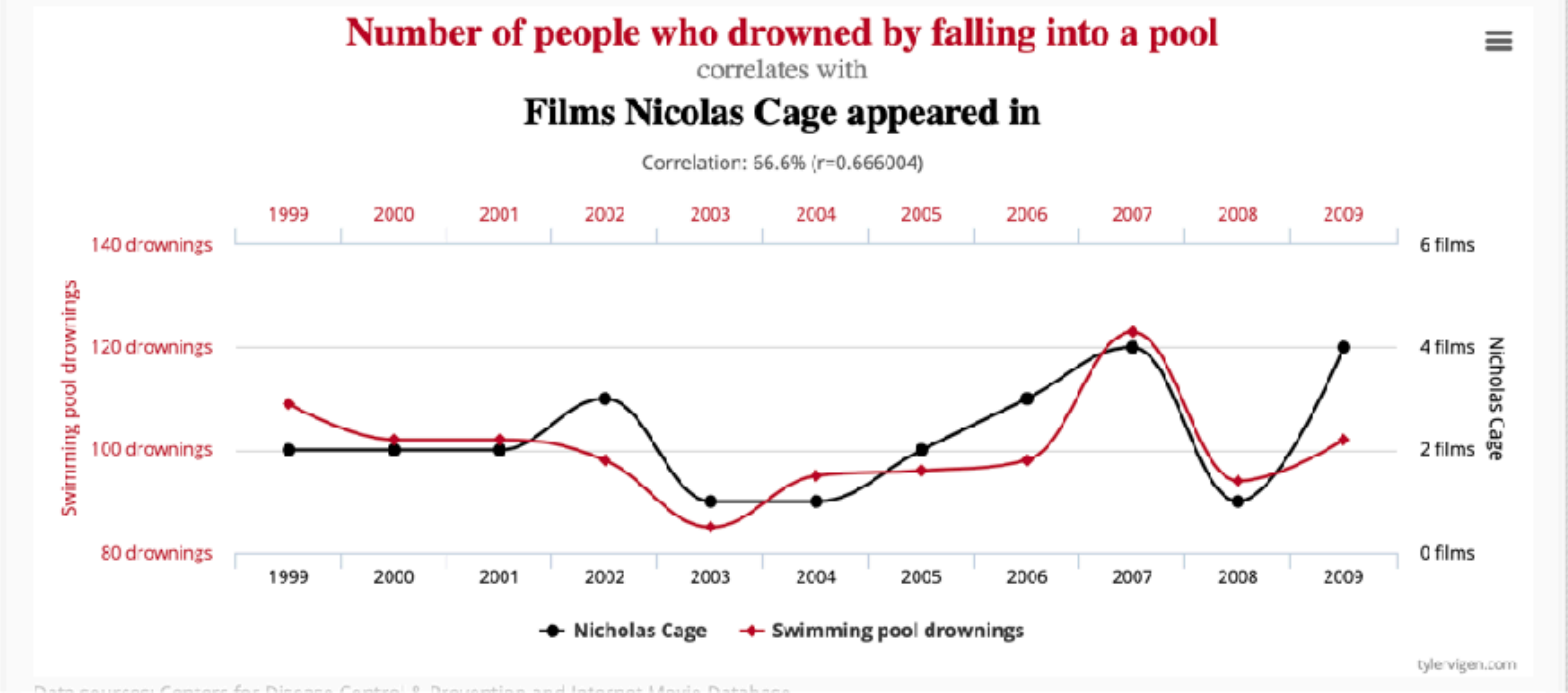
As the chart shows, the more you learn, the more you earn. Median weekly earnings in 2017 for those with the highest levels of educational attainment—doctoral and professional degrees—were more than triple those with the lowest level, less than a high school diploma. And workers with at least a bachelor's degree earned more than the $907 median weekly earnings for all workers.

Click the chart legend to see a second chart showing unemployment rates by educational attainment. As that chart shows, the higher the level of education, the lower the unemployment rate. Compare unemployment by education level in 2017 with the overall unemployment rate of 3.6 percent.

Spurious correlations



Problem:
correlation is common-place in the world!



Causation and correlation

Correlation does not imply causation

Causation does not imply correlation

Causation implies **conditional** correlation

X causes Y does not mean that X is the only thing that acuses Y

And it doesn't mean that there is a one-to-one correspondence between them

The important thing is that X changes the **probability** that Y happens

Causal inference

Some correlations are causal

But many aren’t

So how can we tell?

Does smoking cause cancer?

Imagine X and Y each only take two values (0, 1)

X here is **smoking** (don't smoke [0], smoke [1])
And Y is **cancer** (no cancer [0], has cancer [1])

One approach: take Joe, check Y when X = 0
And then check Y when X = 1

Do the same for everyone in sample, and average to know effect of X on Y

Potential outcomes

name	smokes	cancer
<chr>	<dbl>	<dbl>
Jamie	1	0.86
Jamie	0	0.5
Joey	1	0.82
Joey	0	0.85
Sarah	1	0.62
Sarah	0	0.4

The fundamental problem

What's the problem here?

We can only observe X at one value for each person

If Joey smokes, we can measure Y when X = 1, **but not** X = 0

If Sarah doesn't smoke, we can measure Y when X = 0, **but not** X = 1

What Y would have been if X took on a different value **is missing**,
And we don't know what it is

Observed outcomes

	name	smokes	cancer
	<chr>	<dbl>	<dbl>
1	Jamie	1	0.86
2	Jamie	NA	NA
3	Joey	NA	NA
4	Joey	0	0.85
5	Sarah	NA	NA
6	Sarah	0	0.4

Does School Suspension Work?

Between 2013-2014, 2.6 million public school students received at least one suspension

Suspension used to punish bad behavior, but it might exacerbate the problem

What's the analogy here?

The fundamental problem of causality

Well why don't we compare the Y's for people whose X = 0 and X = 1?

If Angela doesn't smoke and Joey does, let's compare them against each other (more generally = compare all smokers to non-smokers)

But Joey (smokers) and Angela (non-smokers) could differ in so many different ways!

Causal inference

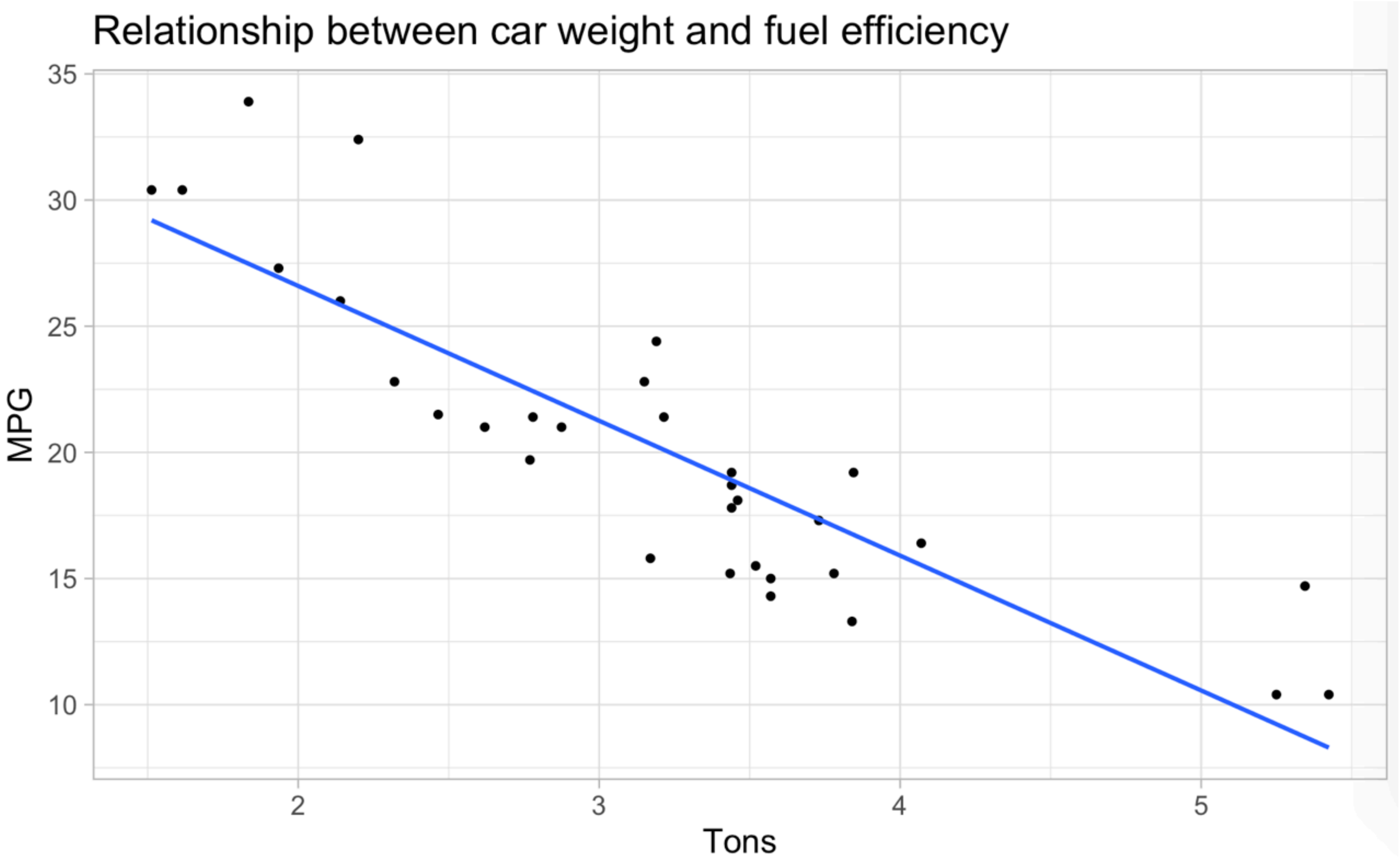
We essentially have missing data on what “would have been” had Joey not smoked

That “would have been” is called a *counterfactual*

Goal in causal inference is to make as good a guess as possible as to what Y would have been had X = 0

we want to think about two people/that are basically exactly the same except that one has X=0 and one has X=1

Obviously causal relationships



Experiments

One way is to use an experiment

If you *randomly assign* X then you know that, on average, people who get X = 0 mirror those for whom X = 1

Experiments

But these are infeasible/unethical in many, many settings

We have to use **models** to figure out what the counterfactual would have been

This **model** is our idea of what process generated the data

The **model** tells us what kinds of things might mess up our results so we can get closer to the right counterfactual