Week 3: Causality and the Scientific Method

DSUA111: Data Science for Everyone, NYU, Fall 2020

TA Jeff, `jpj251@nyu.edu`

https://github.com/jpowerj/dsua111-sections
(https://github.com/jpowerj/dsua111-sections)

Zoom Room

https://nyu.zoom.us/j/6821254378 (https://nyu.zoom.us/j/6821254378)

Same URL for both [in case you want a reminder of your Lab/Section number]:

- Lab 007, Fridays 3:30pm-4:20pm
- Lab 004, Fridays 4:55-5:45pm

Outline

1. Doing HW1
2. Recap: Data? Science?
    - The Scientific Method
    - Measurement Issues
3. Correlation vs. Causation
    - Correlation Pitfalls
4. Causality and Counterfactuals
    - The Fundamental Problem of Causal Inference
5. Experiments: How Do They Help Us?

0. Doing HW1

- https://dsua-111-fall.rcnyu.org/hub (https://dsua-111-fall.rcnyu.org/hub)

1. Recap: Data? Science?

Data

- More than numbers (think "Symbolic Systems")
- It does not "speak for itself", but must be interpreted

In [9]:

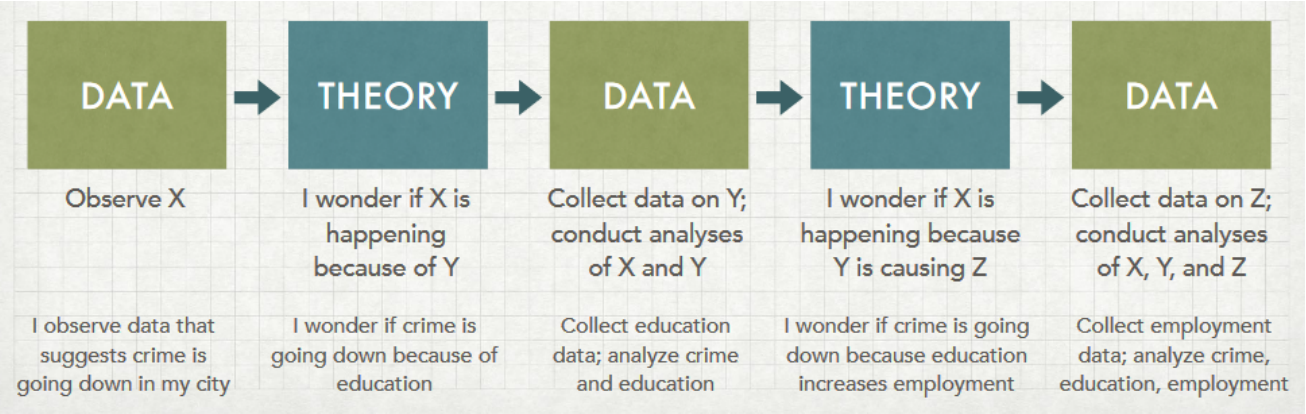
```
%%HTML
<video width="80%" controls>
      <source src="looked_at_the_data.mp4" type="video/mp4">
</video>
```

Science

- Doubt and skepticism
- **Dis**proving theories
    - "In so far as a scientific statement speaks about reality, it must be **falsifiable**.
    - And in so far as it is **not** falsifiable, it does not speak about reality." (Popper)
- Asking questions
- Humility that one study is a [usually very] **imperfect** snapshot

Scientific Method

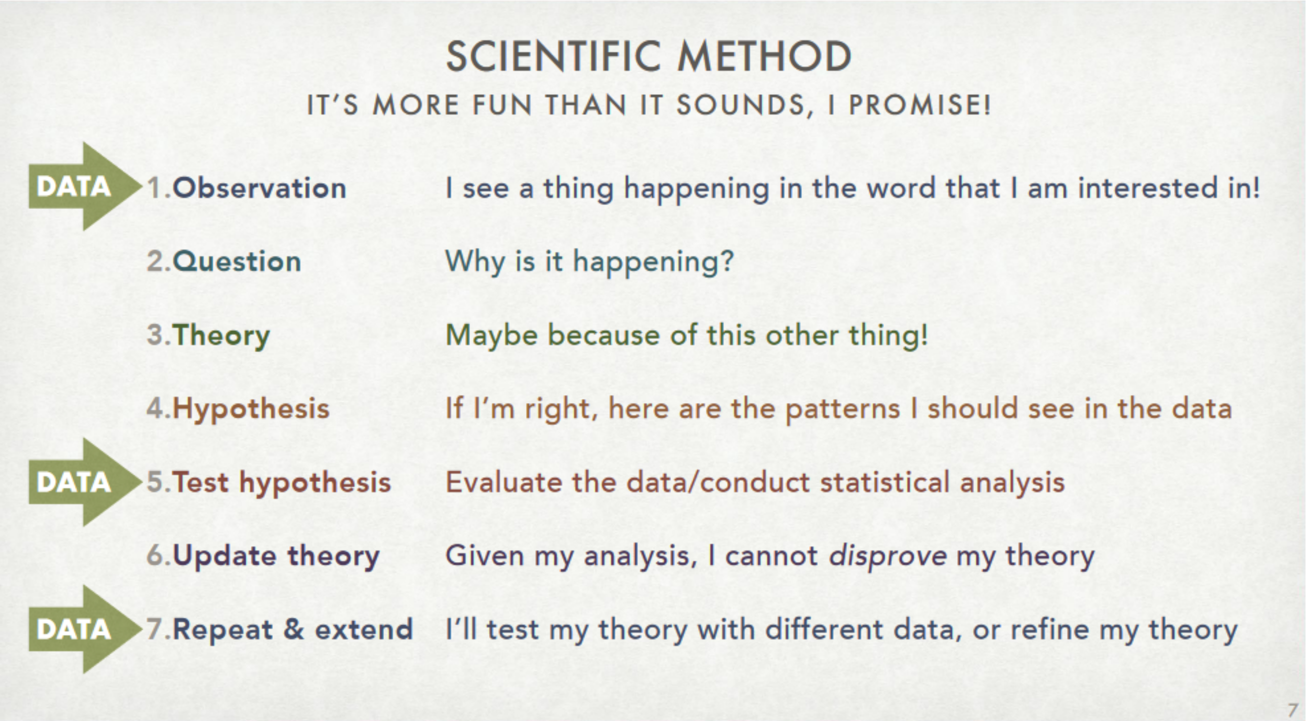
- Data <-> Theory
- Theory <-> Data



DATA	THEORY	DATA	THEORY	DATA
Observe X	I wonder if X is happening because of Y	Collect data on Y; conduct analyses of X and Y	I wonder if X is happening because Y is causing Z	Collect data on Z; conduct analyses of X, Y, and Z
I observe data that suggests crime is going down in my city	I wonder if crime is going down because of education	Collect education data; analyze crime and education	I wonder if crime is going down because education increases employment	Collect employment data; analyze crime, education, employment

- **No proving**!
- **Collective** endeavor, over decades/centuries!
  - "Standing on the shoulders of giants"

Zooming In



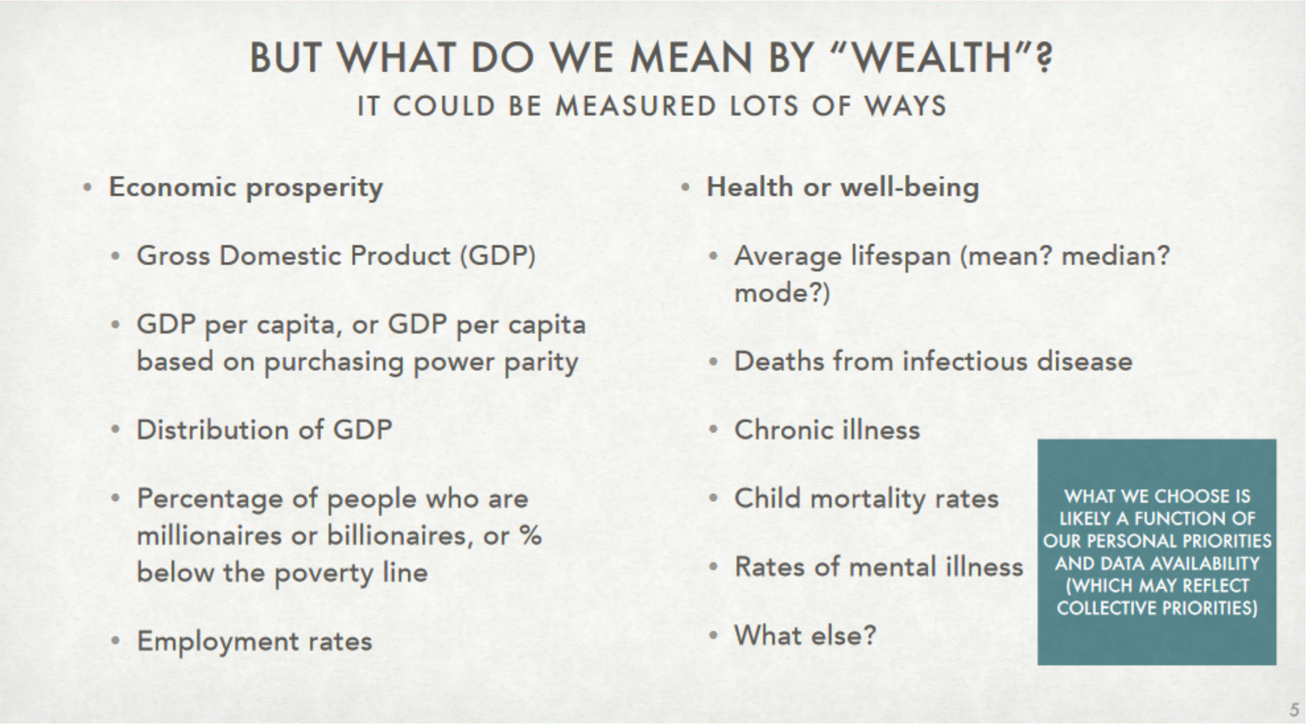
SCIENTIFIC METHOD

IT'S MORE FUN THAN IT SOUNDS, I PROMISE!

DATA	1. **Observation**	I see a thing happening in the word that I am interested in!
	2. **Question**	Why is it happening?
	3. **Theory**	Maybe because of this other thing!
	4. **Hypothesis**	If I'm right, here are the patterns I should see in the data
DATA	5. **Test hypothesis**	Evaluate the data/conduct statistical analysis
	6. **Update theory**	Given my analysis, I cannot *disprove* my theory
DATA	7. **Repeat & extend**	I'll test my theory with different data, or refine my theory

7

2. Measurement Issues



BUT WHAT DO WE MEAN BY "WEALTH"?

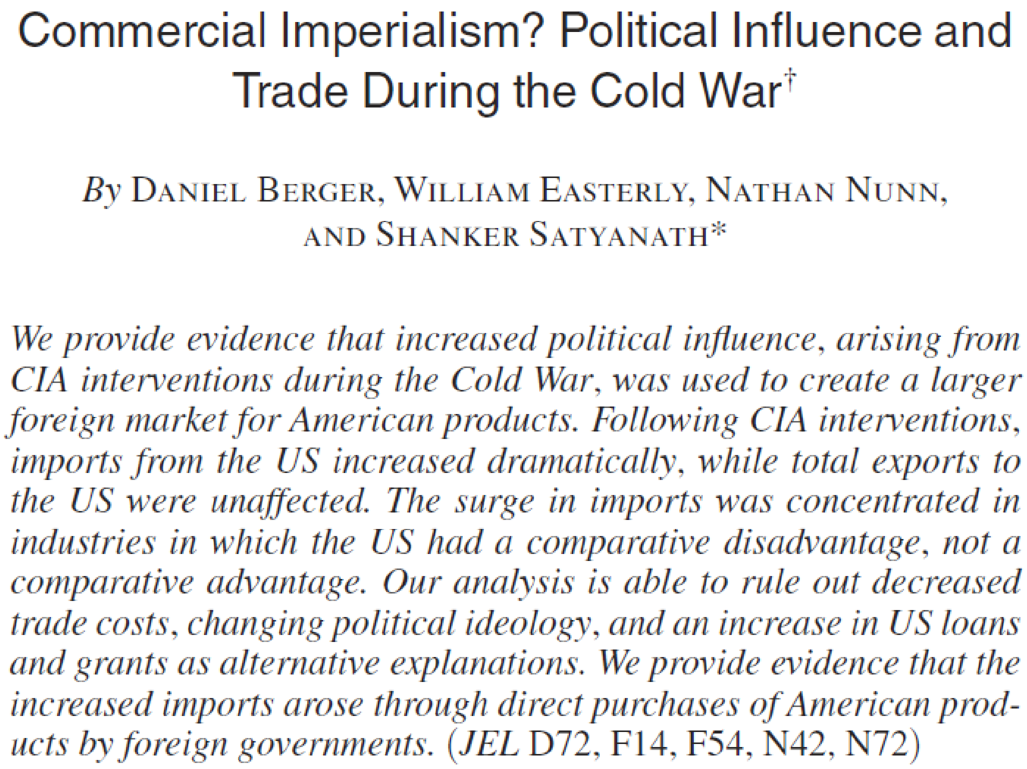
IT COULD BE MEASURED LOTS OF WAYS

- Economic prosperity
  - Gross Domestic Product (GDP)
  - GDP per capita, or GDP per capita based on purchasing power parity
  - Distribution of GDP
  - Percentage of people who are millionaires or billionaires, or % below the poverty line
  - Employment rates

- Health or well-being
  - Average lifespan (mean? median? mode?)
  - Deaths from infectious disease
  - Chronic illness
  - Child mortality rates
  - Rates of mental illness
  - What else?

WHAT WE CHOOSE IS LIKELY A FUNCTION OF OUR PERSONAL PRIORITIES AND DATA AVAILABILITY (WHICH MAY REFLECT COLLECTIVE PRIORITIES)

What does this look like in practice?



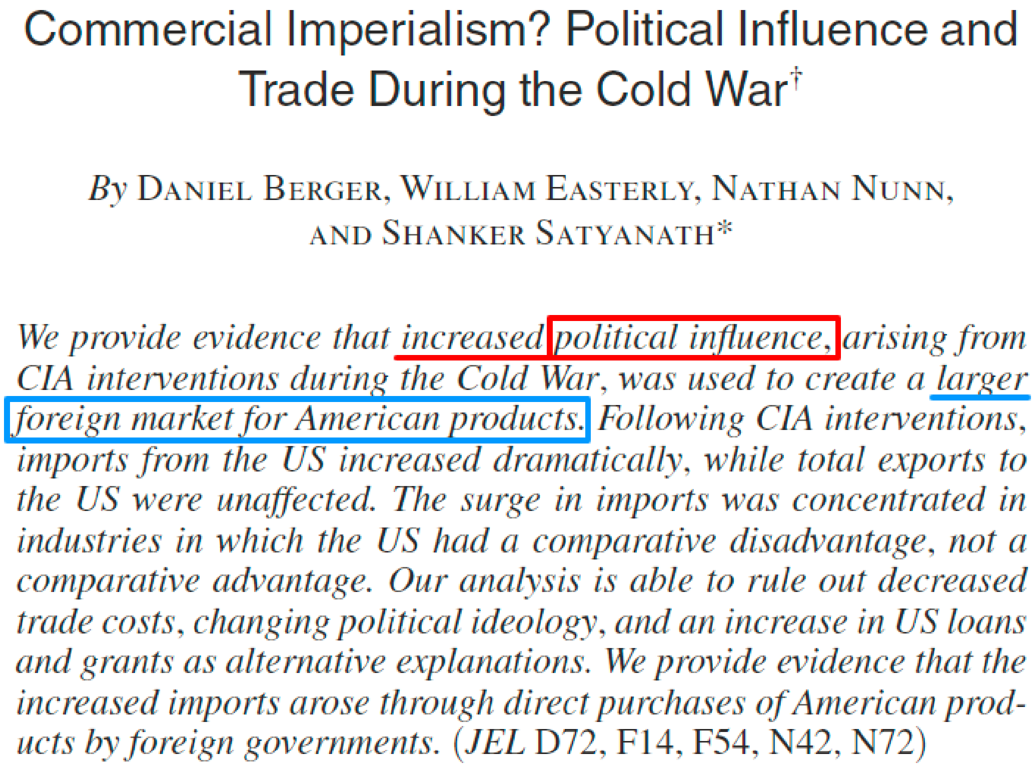
Commercial Imperialism? Political Influence and Trade During the Cold War[†]

By Daniel Berger, William Easterly, Nathan Nunn, and Shanker Satyanath*

We provide evidence that increased political influence, arising from CIA interventions during the Cold War, was used to create a larger foreign market for American products. Following CIA interventions, imports from the US increased dramatically, while total exports to the US were unaffected. The surge in imports was concentrated in industries in which the US had a comparative disadvantage, not a comparative advantage. Our analysis is able to rule out decreased trade costs, changing political ideology, and an increase in US loans and grants as alternative explanations. We provide evidence that the increased imports arose through direct purchases of American products by foreign governments. (*JEL* D72, F14, F54, N42, N72)

(Berger, Daniel, William Easterly, Nathan Nunn, and Shanker Satyanath. 2013. "Commercial Imperialism? Political Influence and Trade during the Cold War." (https://www.aeaweb.org/articles?id=10.1257/aer.103.2.863) American Economic Review, 103 (2): 863-96.)

The Variables

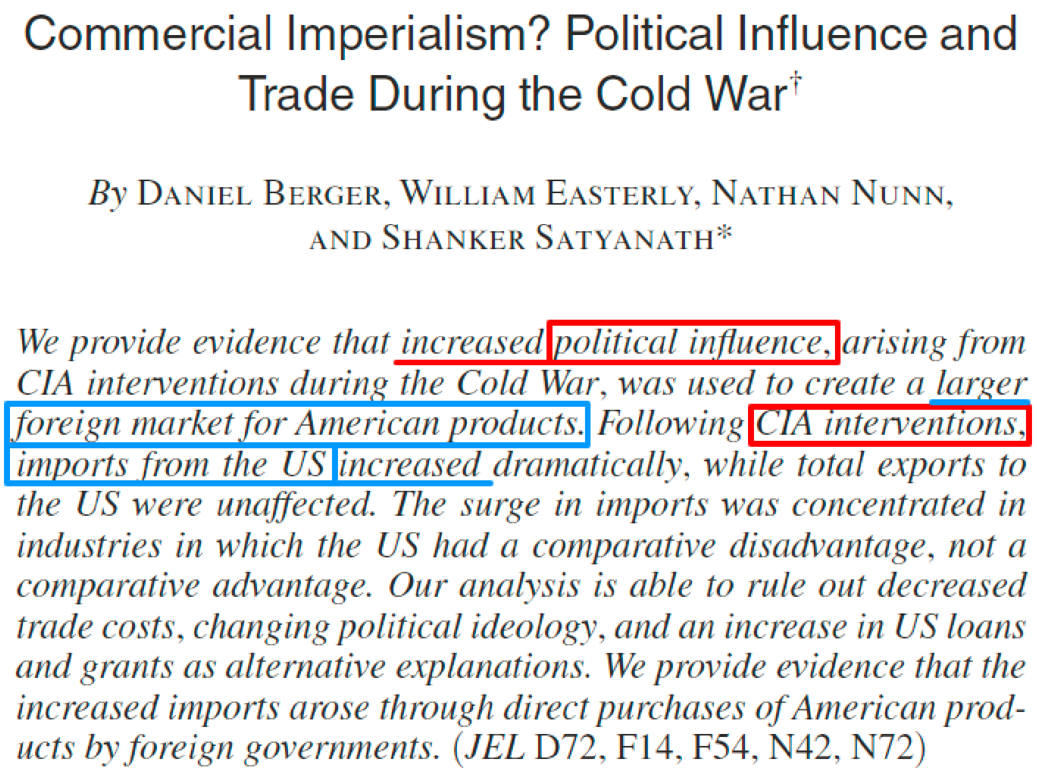


Commercial Imperialism? Political Influence and Trade During the Cold War[†]

By Daniel Berger, William Easterly, Nathan Nunn, and Shanker Satyanath*

We provide evidence that increased political influence, arising from CIA interventions during the Cold War, was used to create a larger foreign market for American products. Following CIA interventions, imports from the US increased dramatically, while total exports to the US were unaffected. The surge in imports was concentrated in industries in which the US had a comparative disadvantage, not a comparative advantage. Our analysis is able to rule out decreased trade costs, changing political ideology, and an increase in US loans and grants as alternative explanations. We provide evidence that the increased imports arose through direct purchases of American products by foreign governments. (*JEL* D72, F14, F54, N42, N72)

How Are They Measured?



Commercial Imperialism? Political Influence and Trade During the Cold War[†]

By Daniel Berger, William Easterly, Nathan Nunn, and Shanker Satyanath[*]

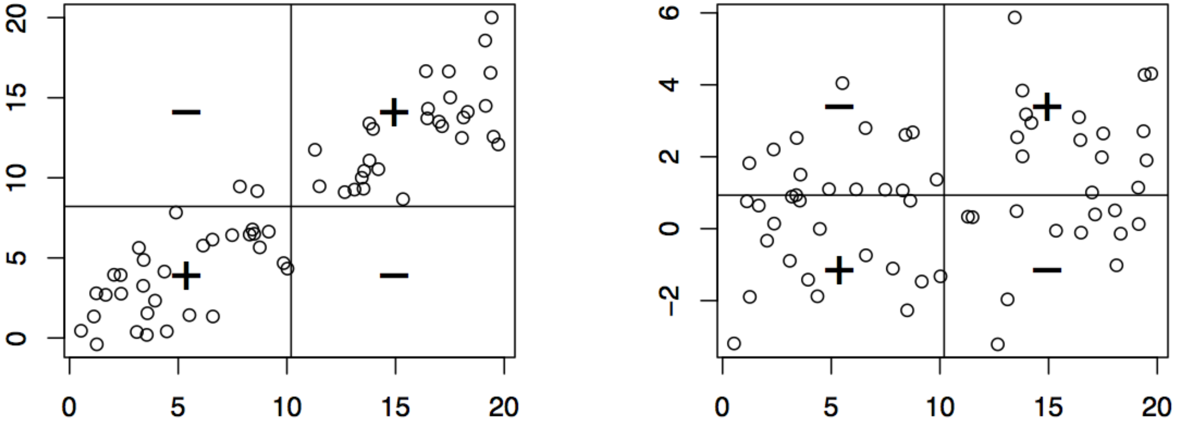
We provide evidence that increased political influence, arising from CIA interventions during the Cold War, was used to create a larger foreign market for American products. Following CIA interventions, imports from the US increased dramatically, while total exports to the US were unaffected. The surge in imports was concentrated in industries in which the US had a comparative disadvantage, not a comparative advantage. Our analysis is able to rule out decreased trade costs, changing political ideology, and an increase in US loans and grants as alternative explanations. We provide evidence that the increased imports arose through direct purchases of American products by foreign governments. (*JEL* D72, F14, F54, N42, N72)

Crucial point here (if you remember nothing else!):

- Before we can even start thinking about the *relationship* between two variables X and Y, we need to know exactly *how* each of them is measured
- So when you read/hear some claim like **"New study finds [*] causes [*]"**, *make sure you know exactly what's going in those blanks!*

3. Correlation vs. Causation

- **Correlation**: X and Y change "together" -- higher values of X tend to "co-occur" with higher values of Y

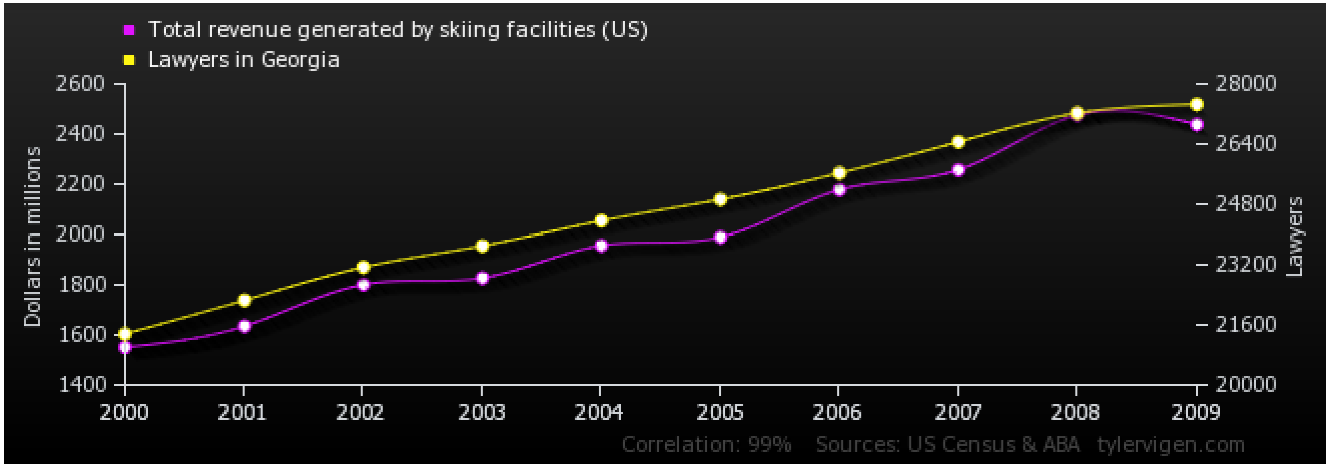


- **Causation**: ...it's trickier than this. Why? Let's find out.

4. Correlation Pitfalls

Spurious Correlations

- http://tylervigen.com/spurious-correlations (http://tylervigen.com/spurious-correlations)
- http://tylervigen.com/view_correlation?id=29272 (http://tylervigen.com/view_correlation?id=29272)



- (Monkeys on a typewriter)

Omitted variables

- **[Umbrellas]** cause **[car accidents]**! (Days with high umbrella use also have high car accident frequency!)

5. Causality and Counterfactuals

- Causality: the **holy grail** of science
- Causal statements require **counterfactuals**: What **would have** happened?
- "Easy" mode: an experiment (in the lab or "in nature")
- Hard mode: observational data

The Fundamental Problem of Causal Inference

- Remember John Snow: the **causal effect** of **[drinking "bad" water]** on the **[infection status]** for a particular person on a given day is the **difference** between:
    - (a) Their infection status after drinking the water, and
    - (b) The infection status they **would have had** on the same day had they not drunk "bad" water.
- **Fundamental Problem of Causal Inference**:
    - We never get to see **both** scenarios for the same unit (person) at the same time, and so
    - We can **never** know the causal effect with certainty!

So, what is to be done?!? Two options…

- Forget Everything And Run?

F.E.A.R.
HAS TWO MEANINGS
FORGET EVRYTHING AND RUN
FACE EVERYTHING AND RISE

Face Everything And Rise

- Find good **comparison** cases: **Treatment** and **Control**
- Without a **control** group, you **cannot** make inferences!
    - (Snow needed at least *some* people who did not drink the pump water... why?)

6. Controlled Experiments: How Do They Help Us?

- **Random Assignment**: Vietnam War/Second Indochina War Draft
    - Key point: makes treatment and control groups **similar**, on average, without us having to do any work!
    - (e.g., don't need to worry about "pairing up" similar treatment+control units)
- No more **Selection Effects**
- **Omitted variables** are in BOTH Treatment and Control groups

Complications: Selection

- Tl;dr **Why** did this person (unit) end up in the **treatment** group? **Why** did this other person (unit) end up in the **control** group?
    - Are there systematic differences?
- Vietnam/Indochina Draft: Why can't we just study **[men who join the military]** versus **[men who don't]**, and take the **difference** as a causal estimate?

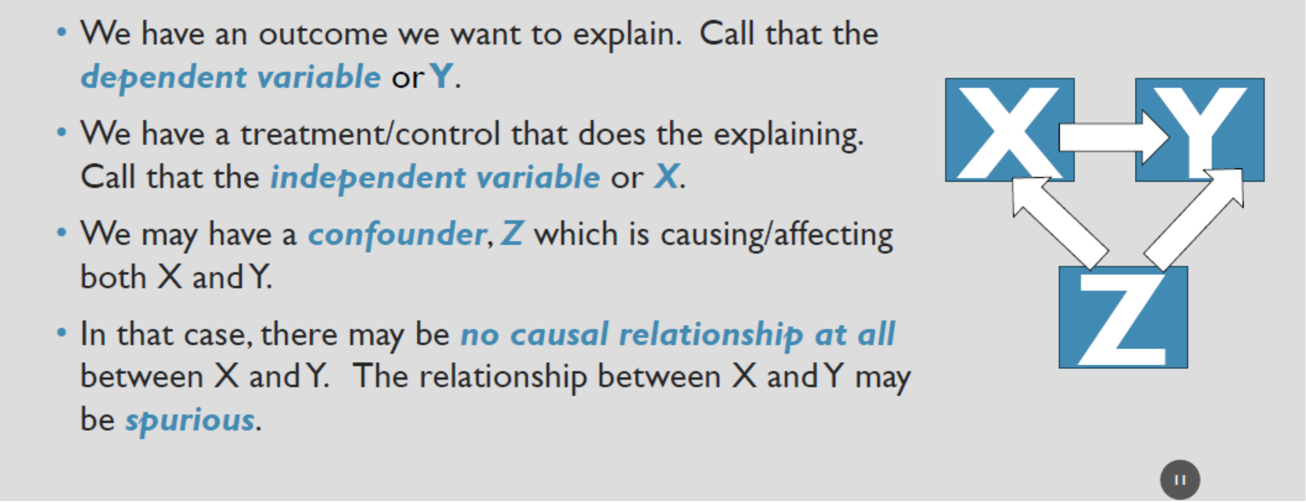
Complications: Compliance

- We ideally want people **assigned** to the treatment to **take** the treatment, and people **assigned** to the control to **take** the control.
- "Compliance" is the degree to which this is actually true in your experiment
    - High compliance = most people actually took what they were assigned
    - Low compliance = lots of people who were assigned to treatment actually took control, and vice-versa
- What problems might there be with **compliance** in the Draft example?

7. The Biggest Complication: Observational Data

- In observational studies, researchers have **no** control over assignment to treatment/control 😨
- On the one hand... Forget Everything And Run, if you can.
- On the other hand... statisticians over the last ~4 centuries have developed fancy causal inference tools/techniques to help us Face Everything And Rise

Causal Terminology for Observational Studies



- We have an outcome we want to explain. Call that the ***dependent variable*** or **Y**.
- We have a treatment/control that does the explaining. Call that the ***independent variable*** or ***X***.
- We may have a ***confounder***, ***Z*** which is causing/affecting both X and Y.
- In that case, there may be ***no causal relationship at all*** between X and Y. The relationship between X and Y may be ***spurious***.