Modeling Social Data: Causality and Experiments

Guest lecturers: Andrew Mao, Amit Sharma

Slides by: Jake Hofman, Andrew Mao, Amit Sharma

Intro to Causal Inference

Amit #1

Prediction

Make a forecast, leaving the world as it is

(seeing my neighbor with an umbrella might predict rain)

vs.

Causation

Anticipate what will happen when you make a change in the world

(but handing my neighbor an umbrella doesn’t cause rain)

“Causes of effects”

It’s tempting to ask “what caused Y”, e.g.

- What makes an email spam?
- What caused my kid to get sick?
- Why did the stock market drop?

This is ”reverse causal inference”, and is generally quite hard

John Stuart Mill (1843)

“Effects of causes”

Alternatively, we can ask “what happens if we do X?”, e.g.

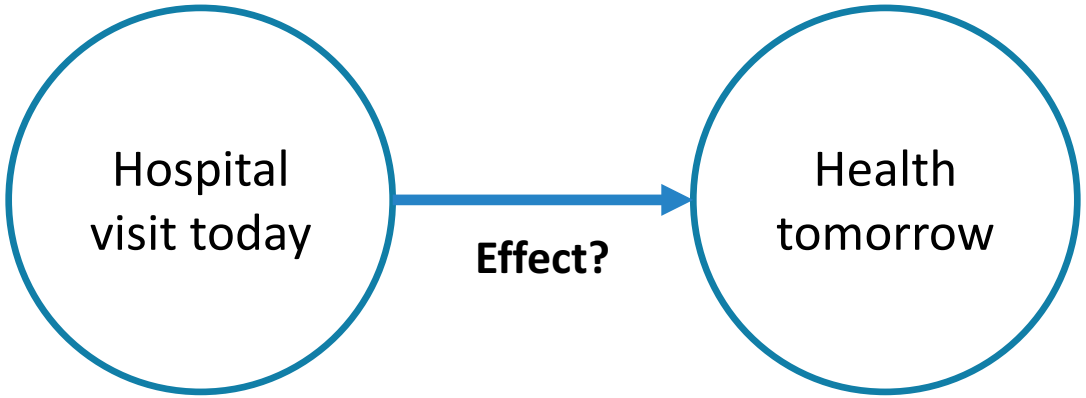
- How does education impact future earnings?
- What is the effect of advertising on sales?
- How does hospitalization affect health?

This is “forward causal inference”: still hard, but less contentious!

John Stuart Mill (1843)

Example: Hospitalization on health

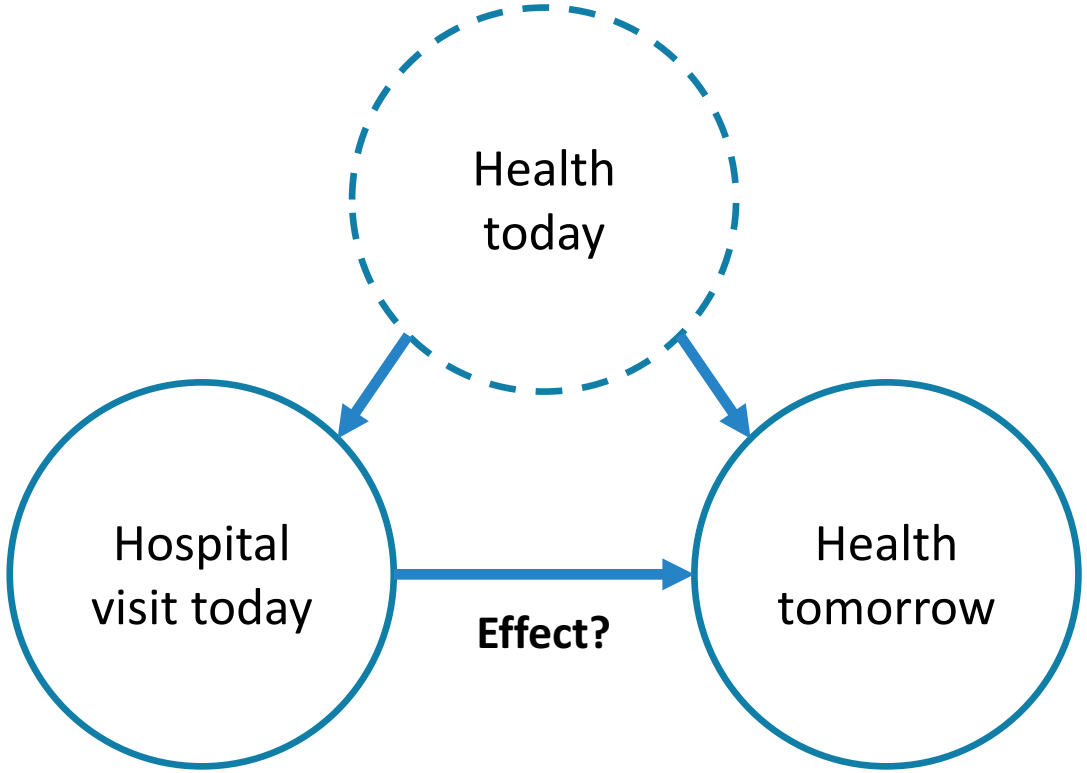
What's wrong with estimating this model from observational data?



Arrow means "X causes Y"

Confounds

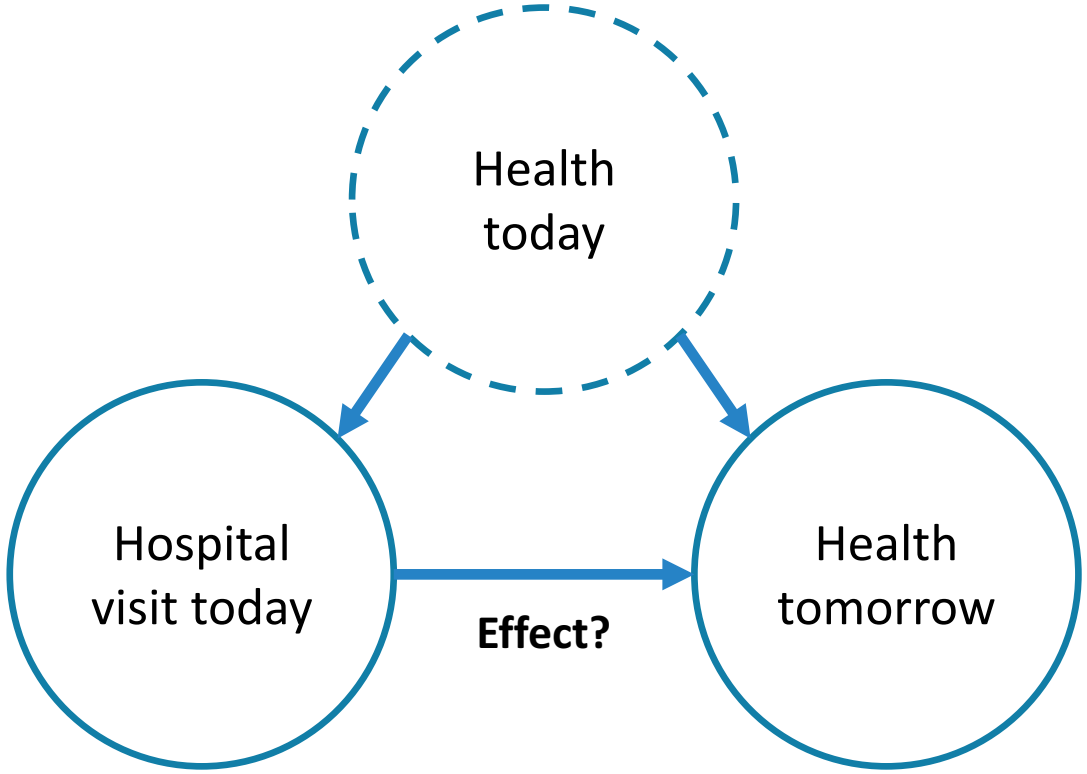
The effect and cause might be *confounded* by a common cause, and be *changing together* as a result



Dashed circle means “unobserved”

Confounds

If we *only get to observe them changing together*, we can't estimate the effect of hospitalization changing alone



Observational estimates

Let's say all sick people in our dataset went to the hospital today, and healthy people stayed home

The observed difference in health tomorrow is:

$$\Delta_{obs} = (\text{Sick and went to hospital}) - (\text{Healthy and stayed home})$$

Observational estimates

Let's say all sick people in our dataset went to the hospital today, and healthy people stayed home

The observed difference in health tomorrow is:

$$\Delta_{obs} = [(\text{Sick and went to hospital}) - (\text{Sick if stayed home})] +$$
$$[(\text{Sick if stayed home}) - (\text{Healthy and stayed home})]$$

Selection bias

Let's say all sick people in our dataset went to the hospital today, and healthy people stayed home

The observed difference in health tomorrow is:

Causal effect

$$\Delta_{obs} = [(\text{Sick and went to hospital}) - (\text{Sick if stayed home})] +$$

$$[(\text{Sick if stayed home}) - (\text{Healthy and stayed home})]$$

Selection bias

(Baseline difference between those who opted in to the treatment and those who didn't)

Basic identity of causal inference[1]

Let's say all sick people in our dataset went to the hospital today, and healthy people stayed home

The observed difference in health tomorrow is:

Observed difference = Causal effect – Selection bias

Selection bias is likely negative here, making the observed difference an underestimate of the causal effect

[1]Varian (2016)

A fundamental problem across science and industry

- Knowledge (Science)
  - What is the effect of doing X?
  - Social sciences
  - Biology and medicine
- Solving problems (Industry)
  - When should I do X?
  - Predictive models in practice
  - Policies for better outcomes

There is a gender gap in earnings for the alumni at every top university, although the size of the difference varies greatly.Credit Matt Rourke/Associated Press

Gaps in Earnings Stand Out in Release of College Data

SEPT. 13, 2015
nytimes.com | Sept. 13, 2015

How do predictive systems work?

Aim: Predict future activity for a user.



...

We see data about their user profile and past activity.

E.g., for any user, we might see their age, gender, past activity and their social network.

From data to prediction

Higher Activity

Lower Activity

Use these correlations to make a predictive model.

Future Activity ->

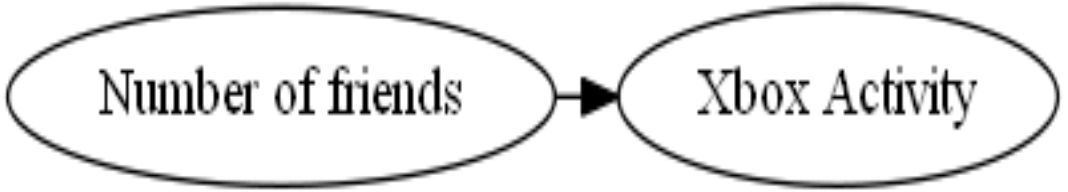
f(number of friends, logins in past month)

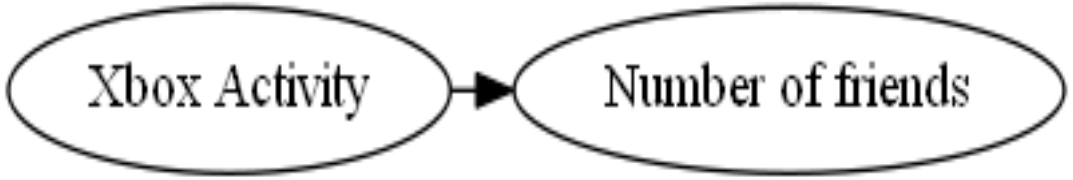
From data to “actionable insights”

Number of friends can predict activity with high accuracy.

How do we increase activity of users?

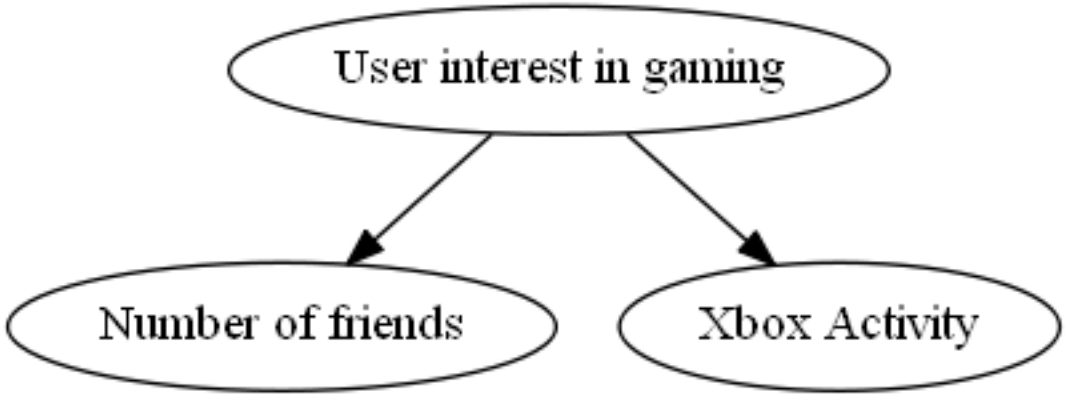
Different explanations are possible





How do we know

what causes what?



Decision: To increase activity, would it make sense to launch a campaign to increase friends?

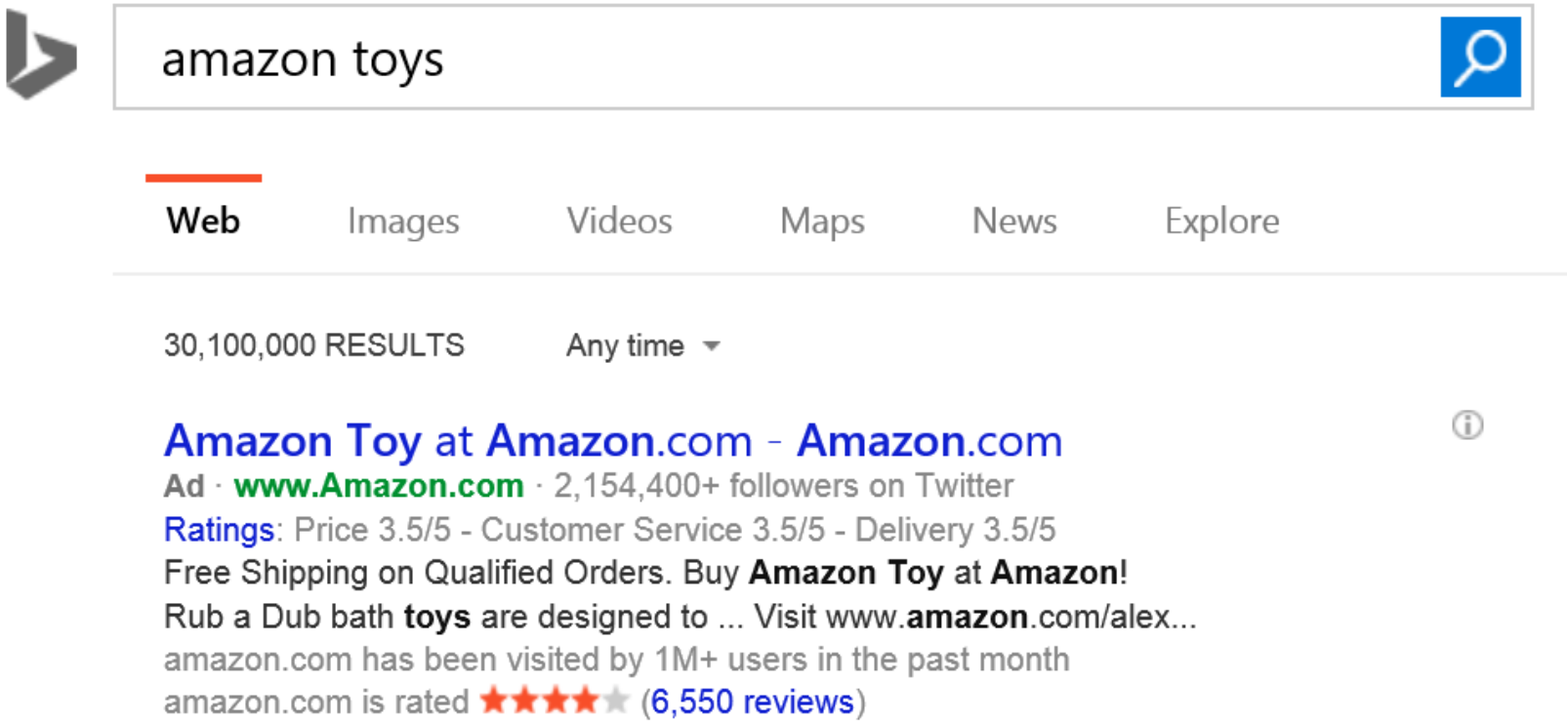
Another example: Search Ads

Search engines uses ad targeting to show relevant ads.

Prediction model based on user's search query.

Search Ads have the highest click-through rate (CTR) in online ads.

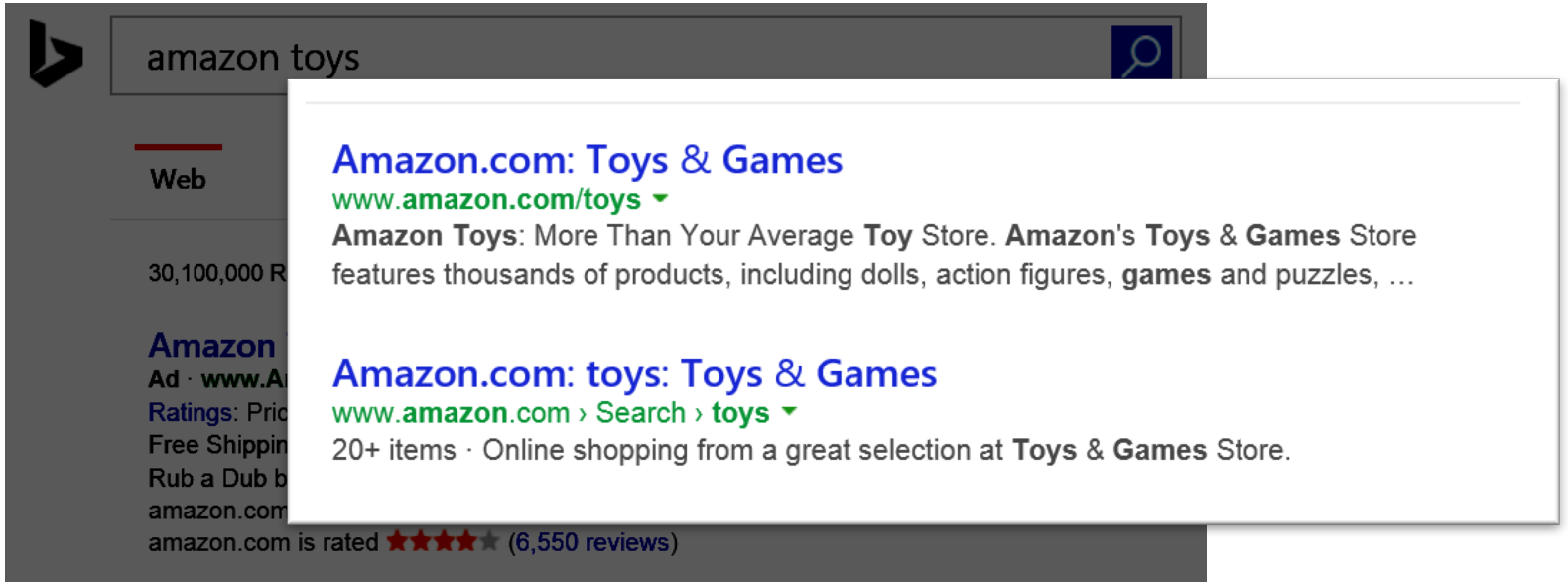
Are search ads really that effective?



Ad targeting was highly accurate.

Blake-Tadelis-Noskos (2014)

But search results point to the same website

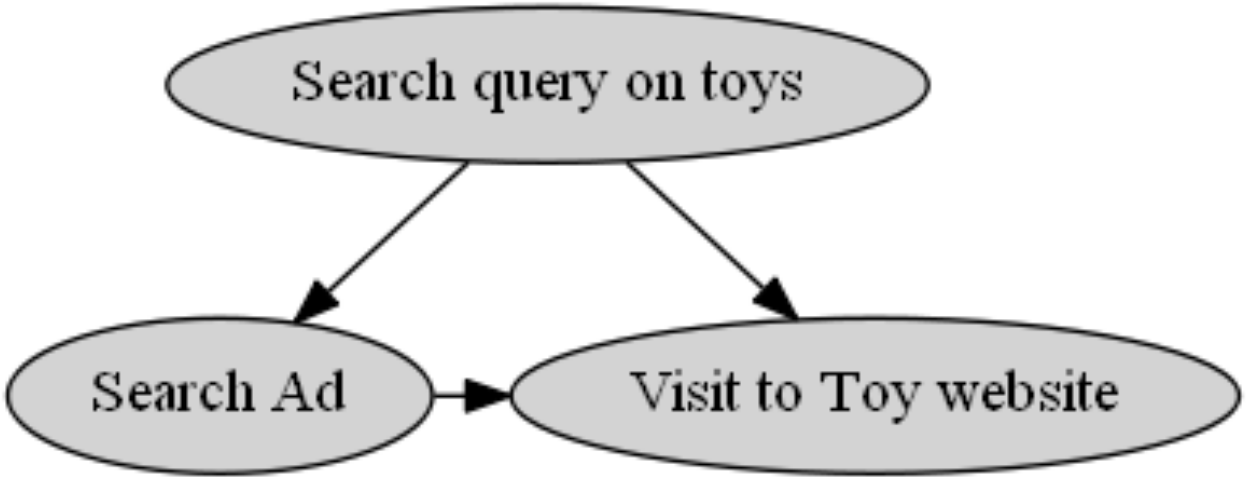


Counterfactual question: Would I have reached Amazon.com anyways, without the ad?

Without reasoning about causality, may overestimate effectiveness of ads

x% of ads shown are effective

<*x%* of ads shown are effective

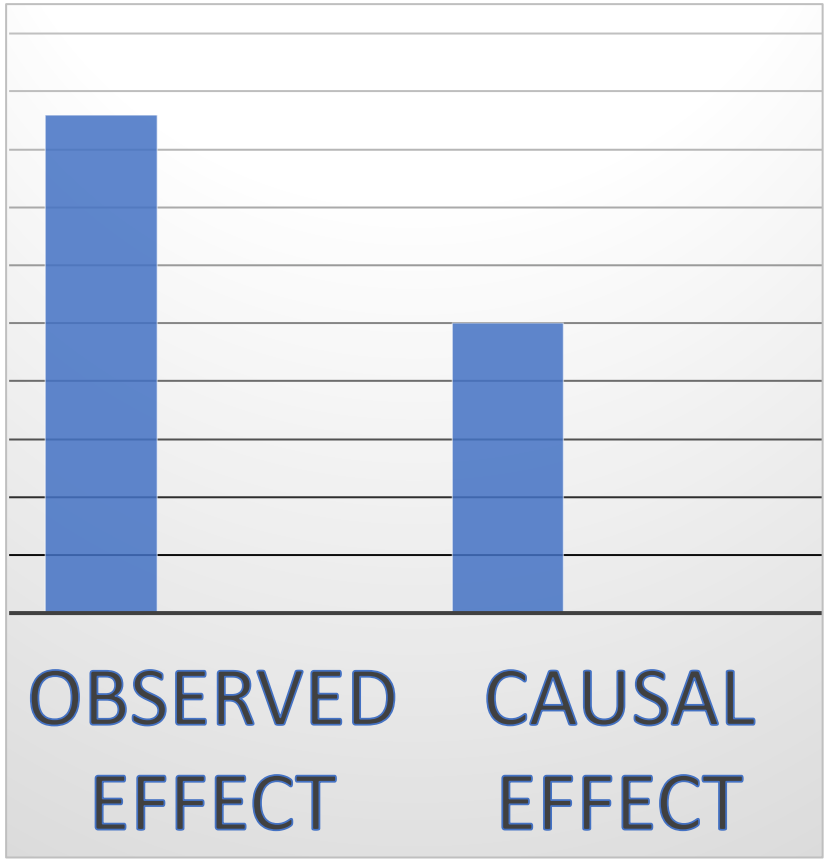


Okay, search ads have an explicit intent. Display ads should be fine?

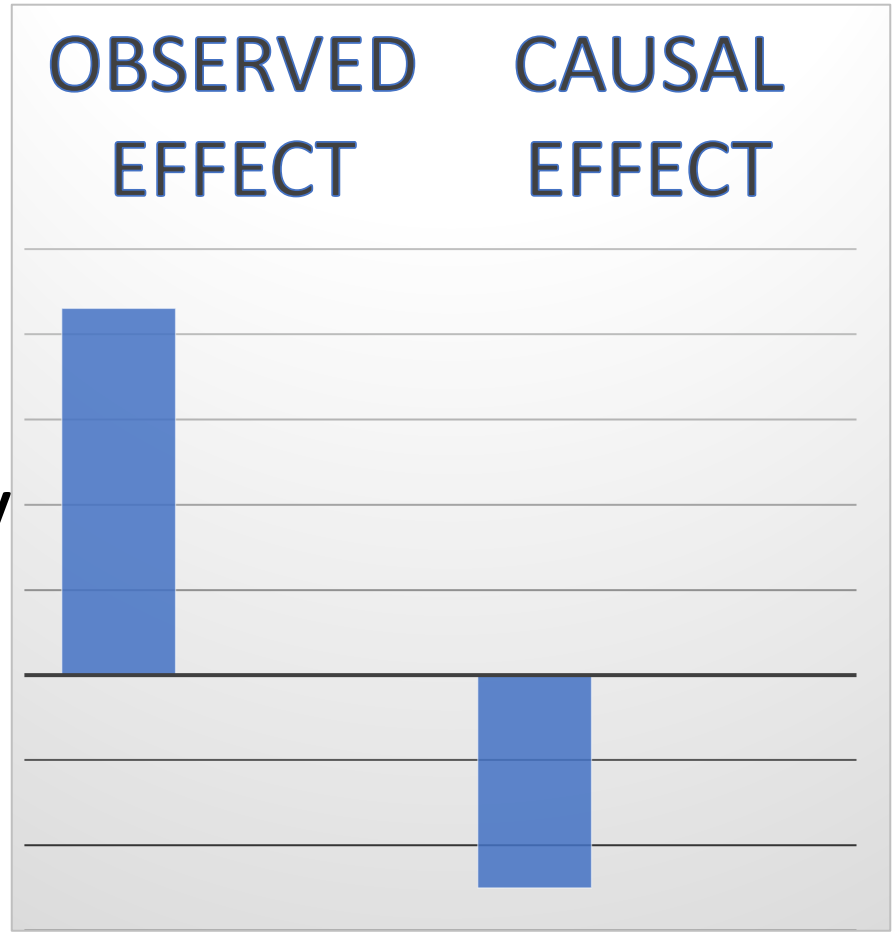
Probably not.

There can be many hidden causes for an action, some of which may be hard to quantify.

So far, so good. Be mindful of hidden causes, or else we might overestimate causal effects.

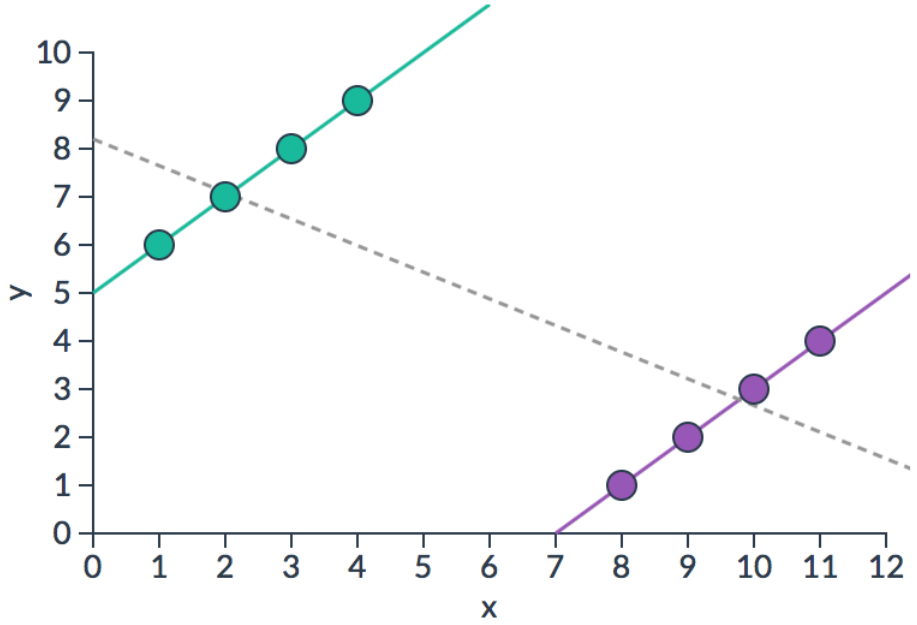


(But)
Ignoring hidden causes can also lead to completely wrong conclusions.



Simpson's paradox

Selection bias can be so large that *observational and causal estimates give opposite effects* (e.g., going to hospitals makes you less healthy)

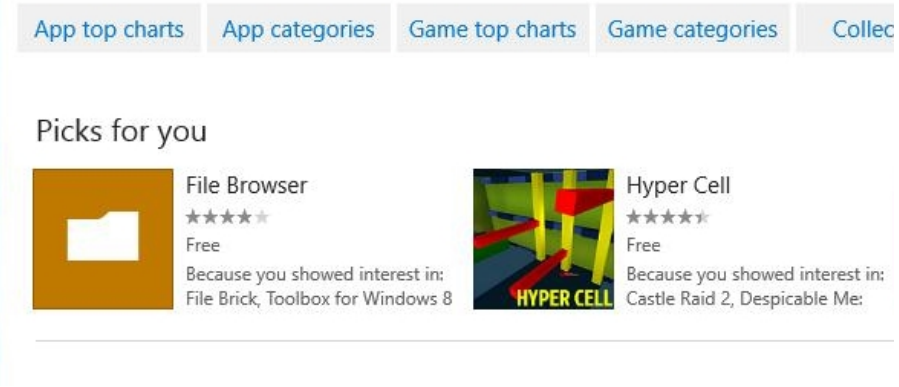


http://vudlab.com/simpsons

Example: Which algorithm is better?

Have a current production algorithm. Want to test if a new algorithm is better.

Say recommendations on app store.



Algorithm A

?

Algorithm B

Comparing old versus new algorithm

Two algorithms, A (production) and B (new) running on the system.

From system logs, collect data for 1000 sessions for each. Measure CTR.

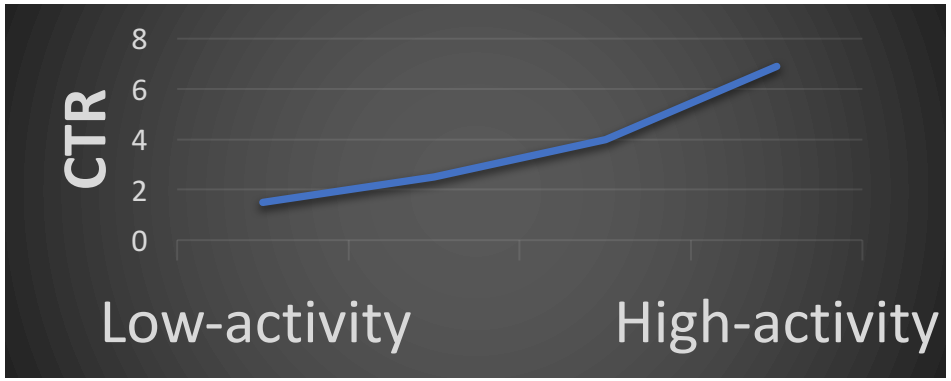
Old Algorithm (A)	New Algorithm (B)
50/1000 **(5%)**	54/1000 **(5.4%)**

New algorithm is better?

Frequent users of the Store tend to be different from new users

So let us look at

CTR separately.



Old Algorithm (A)	New Algorithm (B)
10/400 **(2.5%)**	4/200 **(2%)**

Low-activity Users

Old Algorithm (A)	New Algorithm (B)
40/600 **(6.6%)**	50/800 **(6.2%)**

High-activity Users

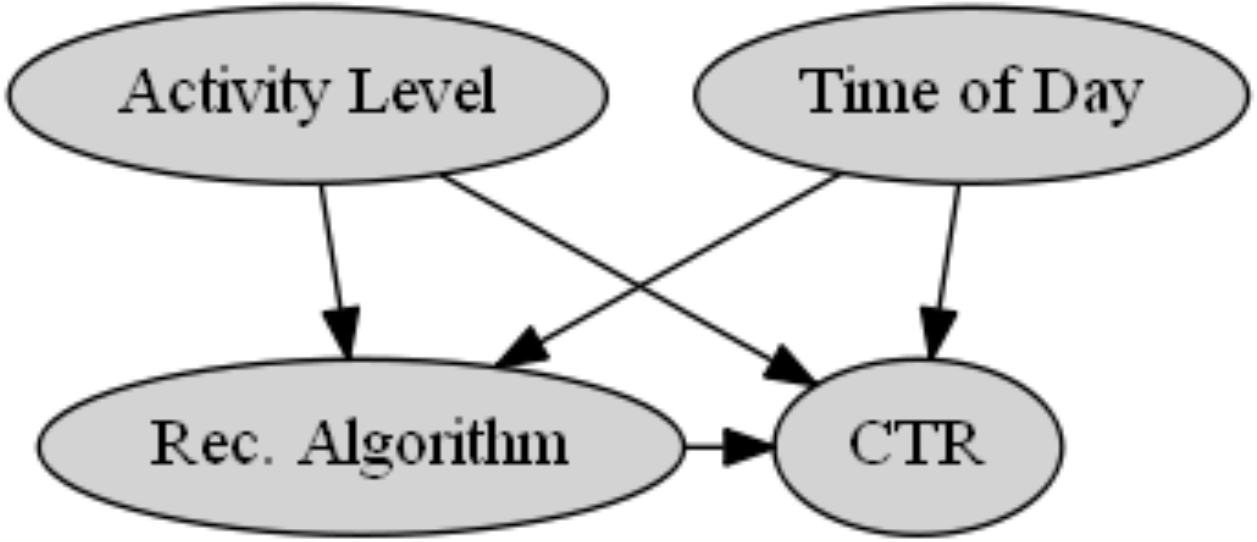
The Simpson's paradox

	Old algorithm (A)	New Algorithm (B)
CTR for Low-Activity users	10/400 (2.5%)	4/200 (2%)
CTR for High-Activity users	40/600 (6.6%)	50/800 (6.2%)
Total CTR	**50/1000 (5%)**	**54/1000 (5.4%)**

- Is Algorithm A better?

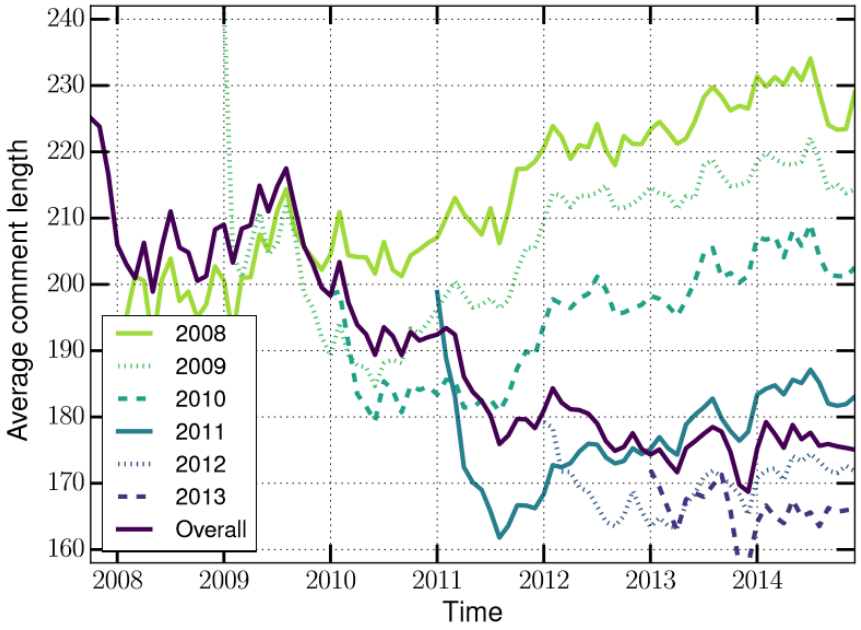
Simpson (1951)

Answer (as usual): May be, may be not.



- E.g., Algorithm A could have been shown at different times than B.
- There could be other hidden causal variations.

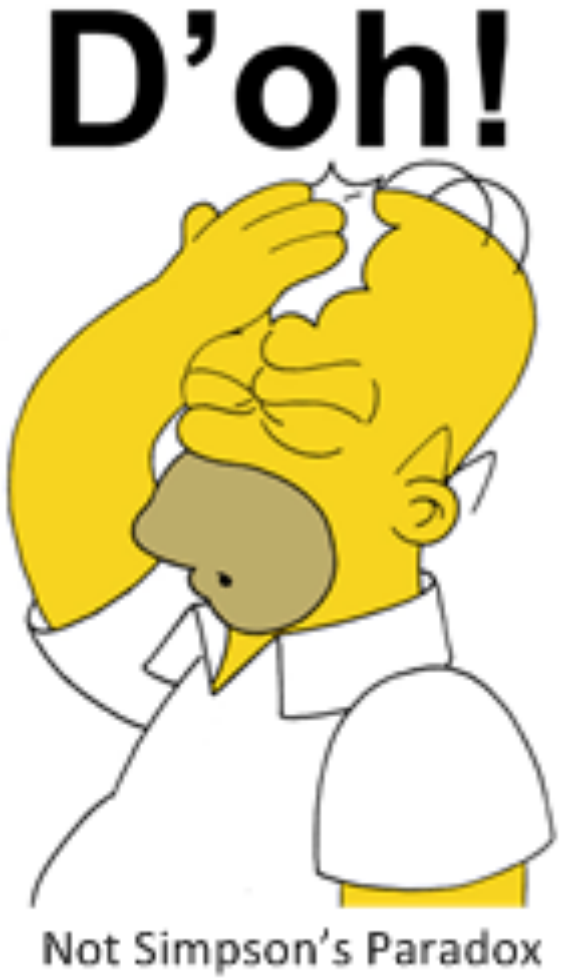
Example: Simpson's paradox in Reddit



- Average comment length decreases over time.

Barbosa-Cosley-Sharma-Cesar (2016)

Making sense of such data can be too complex.



Intro to Experiments

Andrew #1

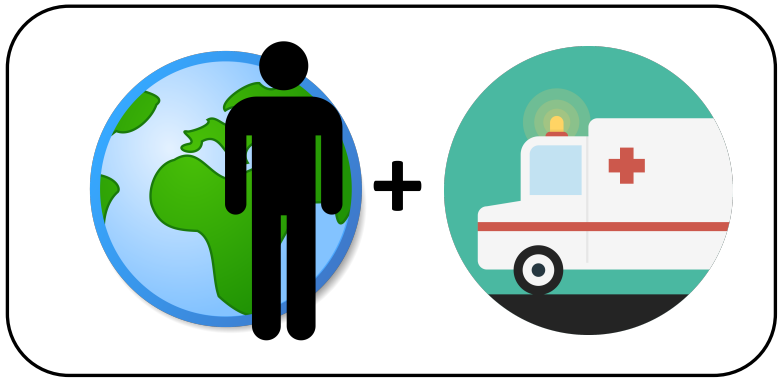
“To find out what happens when you change something, it is necessary to change it.”

-GEORGE BOX

Counterfactuals

To isolate the causal effect, we have to *change one and only one thing* (hospital visits), and compare outcomes

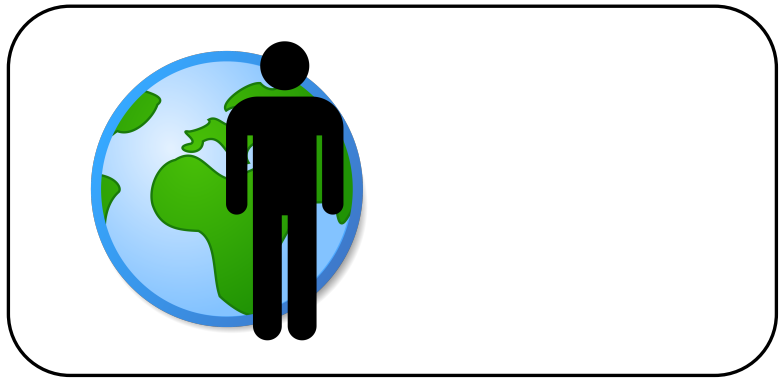
Reality



(what happened)

vs

Counterfactual

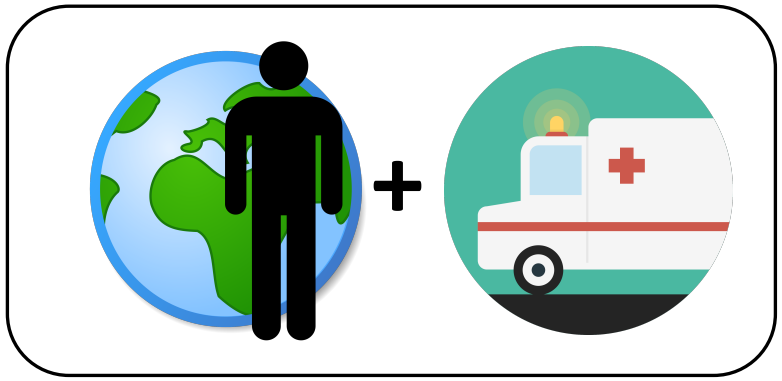


(what would have happened)

Counterfactuals

We never get to observe *what would have happened if we did something else*, so we have to estimate it

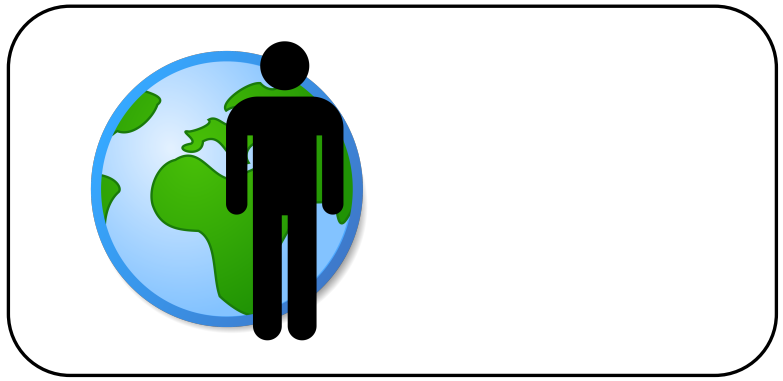
Reality



(what happened)

vs

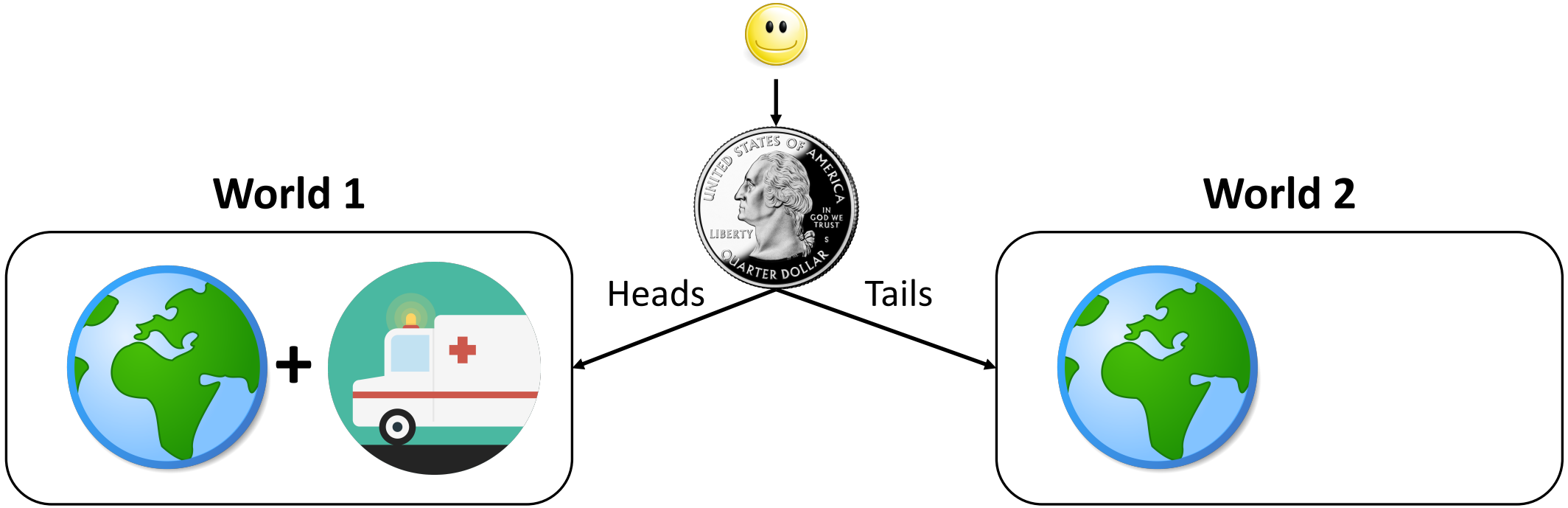
Counterfactual



(what would have happened)

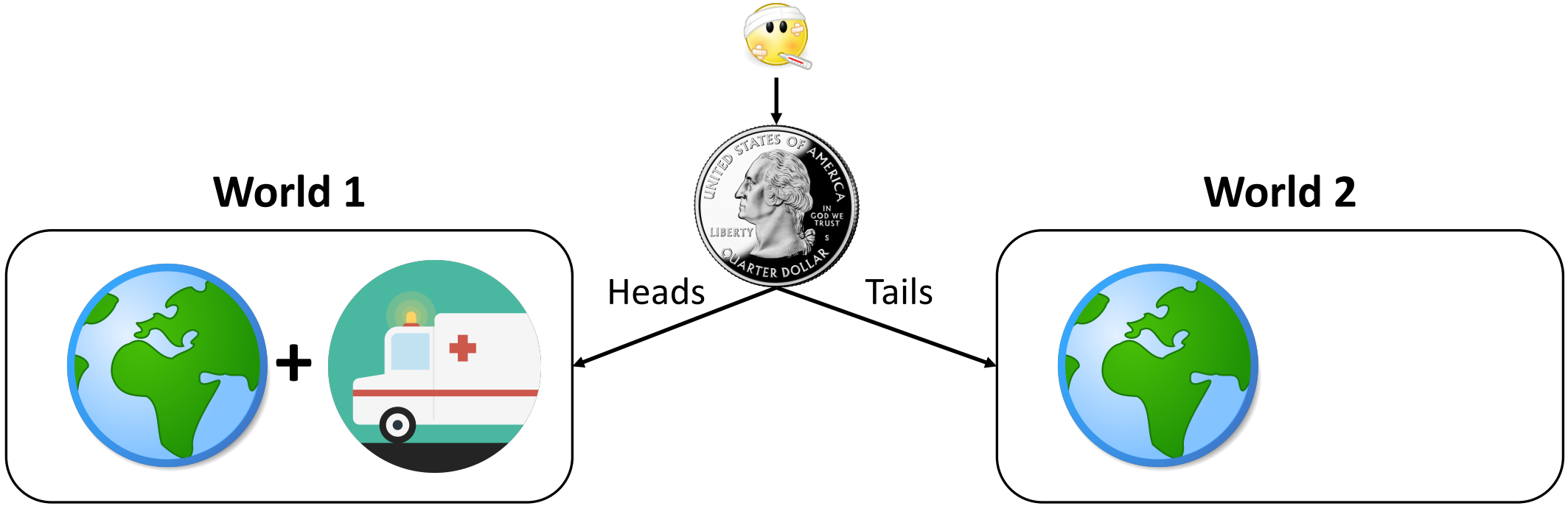
Random assignment

We can use randomization to create two groups that differ only in which treatment they receive, restoring symmetry



Random assignment

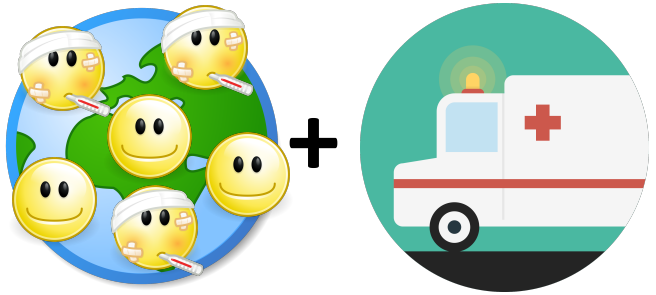
We can use randomization to create two groups that differ only in which treatment they receive, restoring symmetry



Random assignment

We can use randomization to create two groups that differ only in which treatment they receive, restoring symmetry

World 1



World 2

Basic identity of causal inference

The observed difference is now the causal effect:

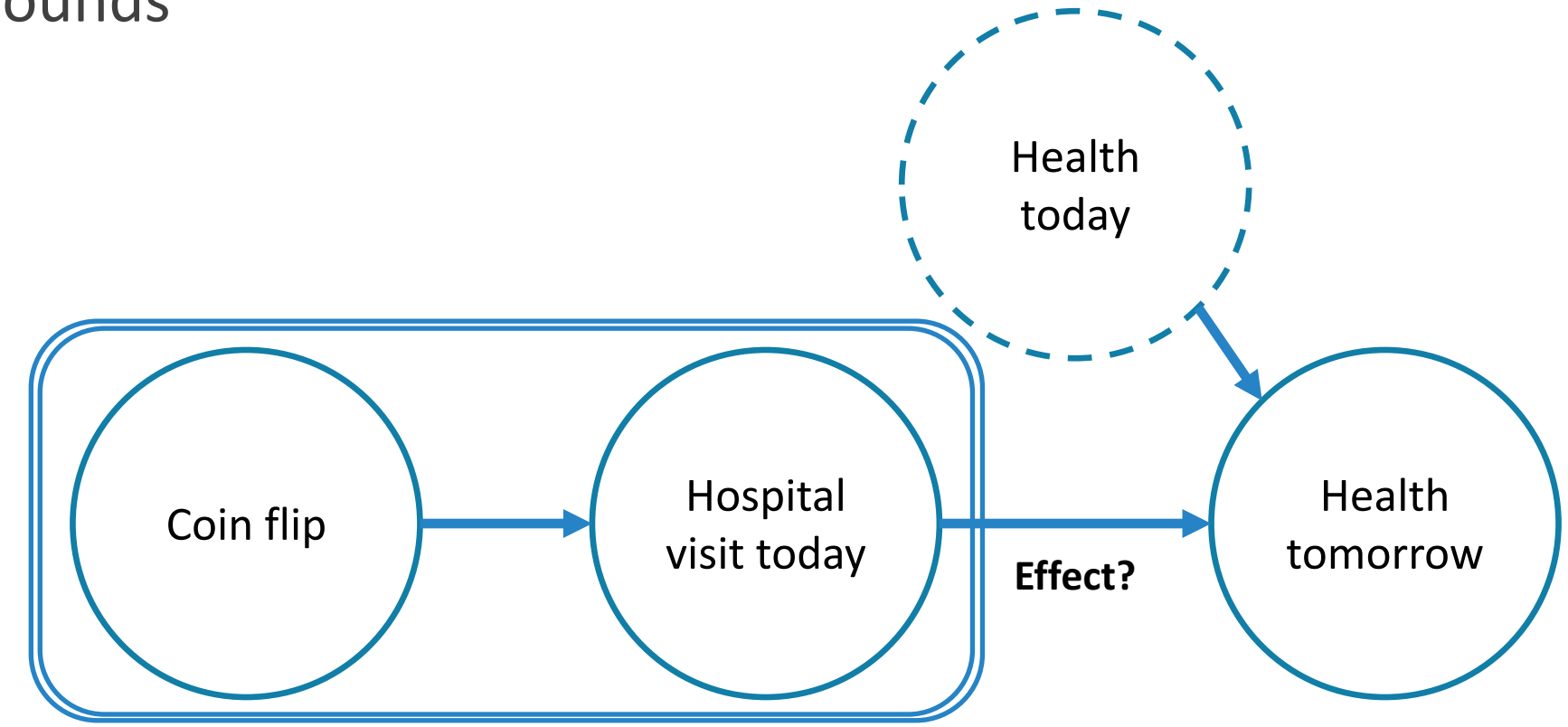
Observed difference = Causal effect – ~~Selection bias~~

= Causal effect

Selection bias is zero, since there's no difference, on average, between those who were hospitalized and those who weren't

Random assignment

Random assignment determines the treatment independent of any confounds



Double lines mean "intervention"

Experiments: Caveats / limitations

Random assignment is the “gold standard” for causal inference, but it has some limitations:

- Randomization often isn’t feasible and/or ethical
- Experiments are costly in terms of time and money
- It’s difficult to create convincing parallel worlds
- Inevitably people deviate from their random assignments

Anyone can flip a coin, but it’s difficult to create convincing parallel worlds

Two goals for experiments

Internal validity:

Could anything other than the treatment (i.e. a confound) have produced this outcome?

Did doctors give the experimental drug to some especially sick patients (breaking randomization) hoping that it would save them?

External validity (Generalization)

Do the results of the experiment hold in settings we care about?

Would this medication be just as effective outside of a clinical trial, when usage is less rigorously monitored?

How we conduct behavioral experiments

Lab Experiment

Field Experiment

Better internal validity (correctness):

- Greatest procedural control
- Can carefully curate situations

But less external validity (generalization):

- Artificial context, simple tasks
- Demand effects
- Homogeneous (WEIRD) subject pools
- Time/scale limitations

Better generalization

- Experiment findings apply to at least one real-world setting

But:

- Less control, more potential confounds
- Demand of experiment conflict with goals of real organizations
- More effort to conduct and manage
- More room for error

Examples of Causal Inference

Amit #2

Natural experiments

Sometimes we get lucky and nature effectively runs experiments for us, e.g.:

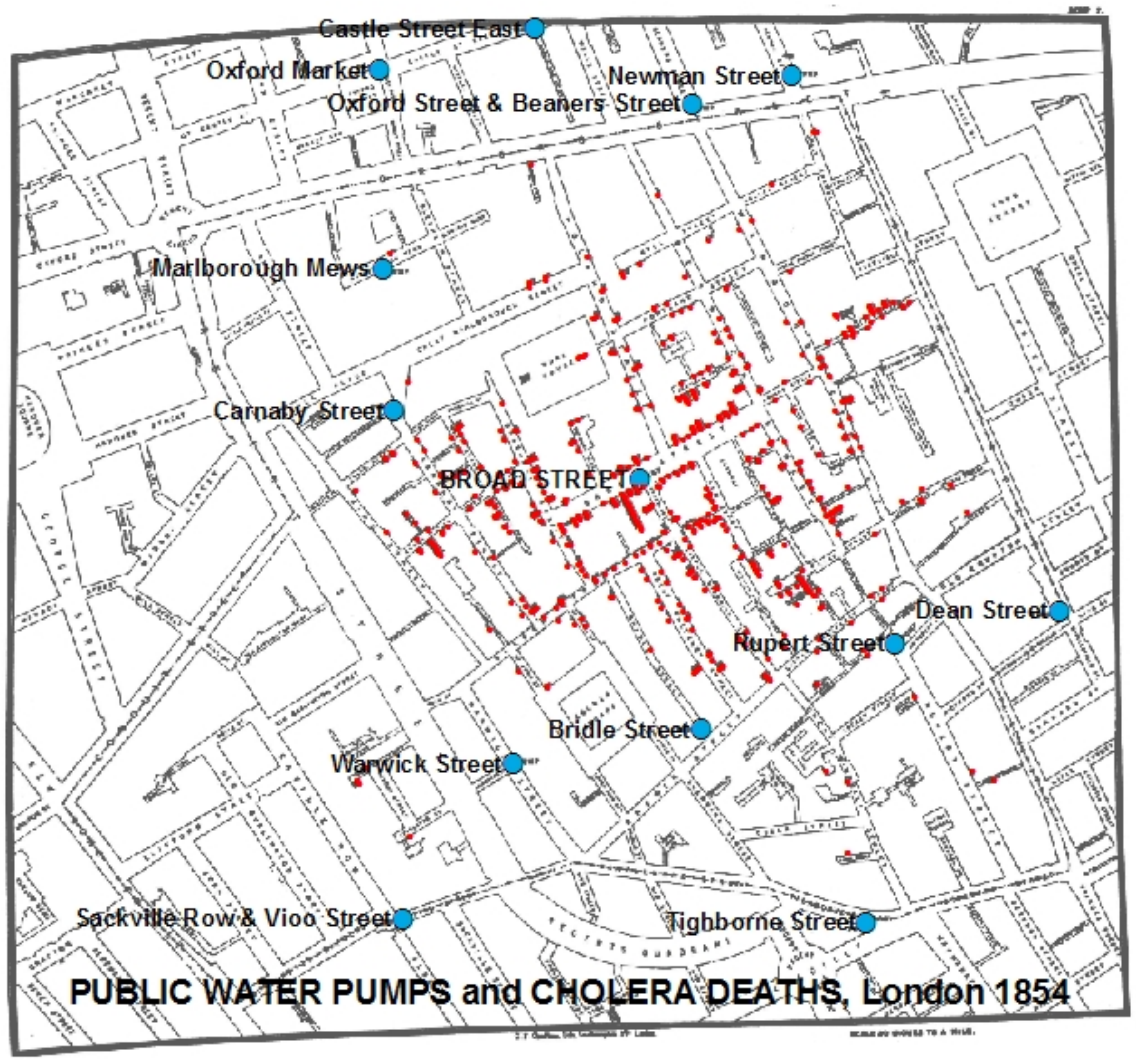
- As-if random: People are randomly exposed to water sources
- Instrumental variables: A lottery influences military service
- Discontinuities: Star ratings get arbitrarily rounded
- Difference in differences: Minimum wage changes in just one state

Natural experiments

Sometimes we get lucky and nature effectively runs experiments for us, e.g.:

- As-if random: People are randomly exposed to water sources
- Instrumental variables: A lottery influences military service
- Discontinuities: Star ratings get arbitrarily rounded
- Difference in differences: Minimum wage changes in just one state

Experiments happen all the time, we just have to notice them



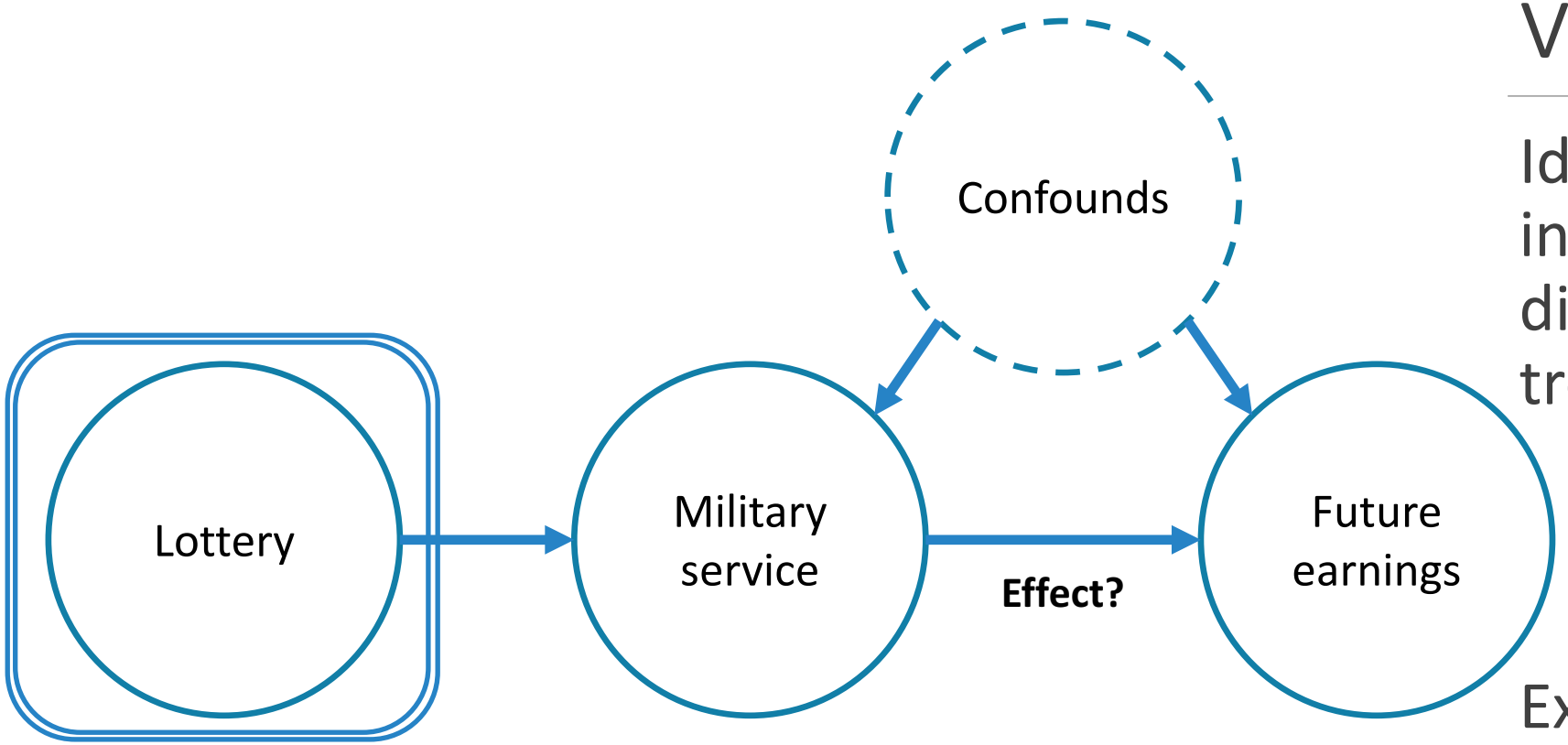
As-if random

Idea: Nature randomly assigns conditions

Example: People are randomly exposed to water sources (Snow, 1854)

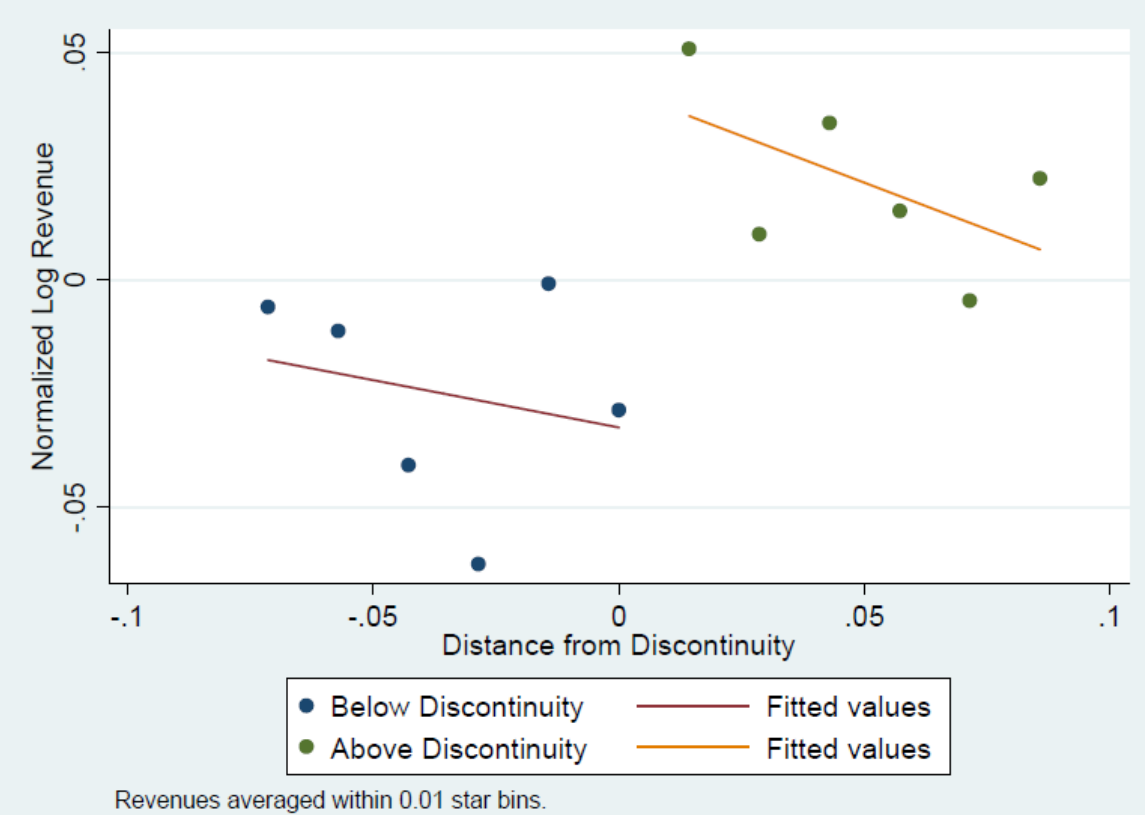
http://bit.ly/johnsnowmap

Instrumental variables



Idea: An instrument independently shifts the distribution of a treatment

Example: A lottery influences military service (Angrist, 1990)

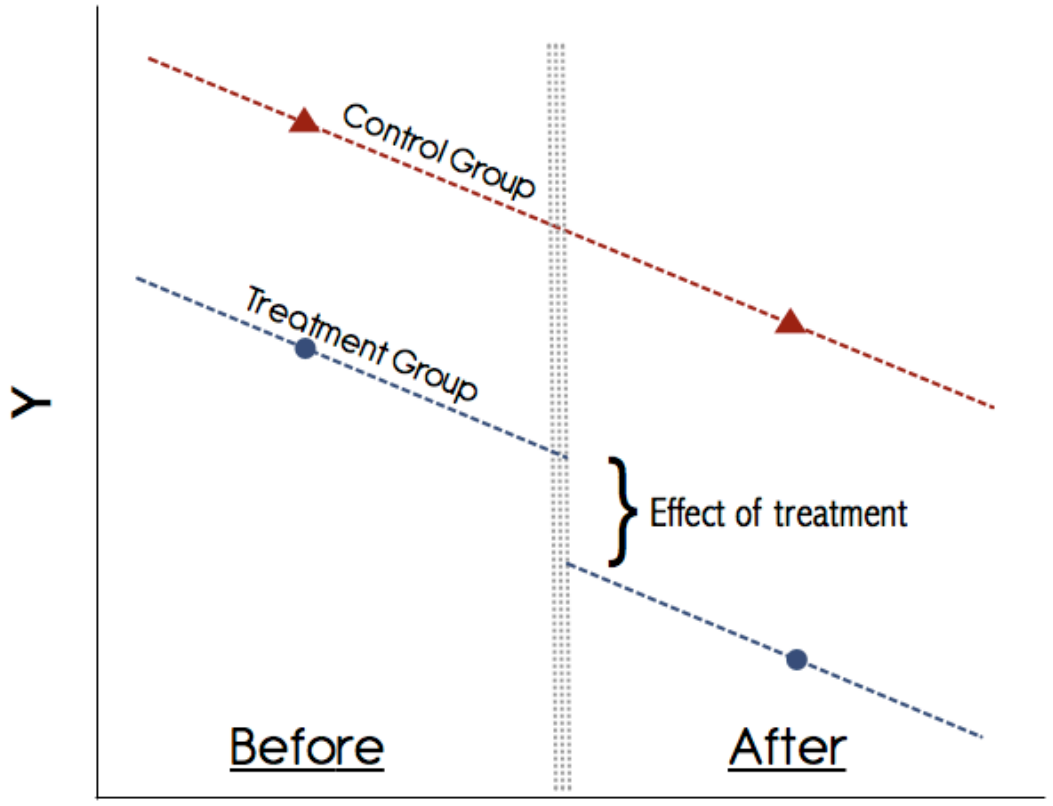


Regression discontinuities

Idea: Things change around an arbitrarily chosen threshold

Example: Star ratings get arbitrarily rounded (Luca, 2011)

http://bit.ly/yelpstars



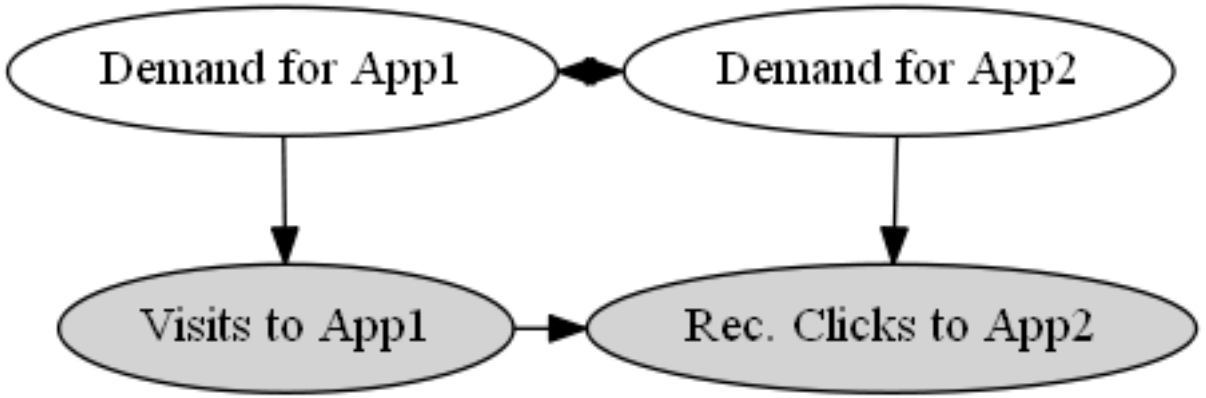
Difference in differences

Idea: Compare differences after a sudden change with trends in a control group

Example: Minimum wage changes in just one state (Card & Krueger, 1994)

http://stats.stackexchange.com/a/125266

Cont. example: Effect of store recommendations

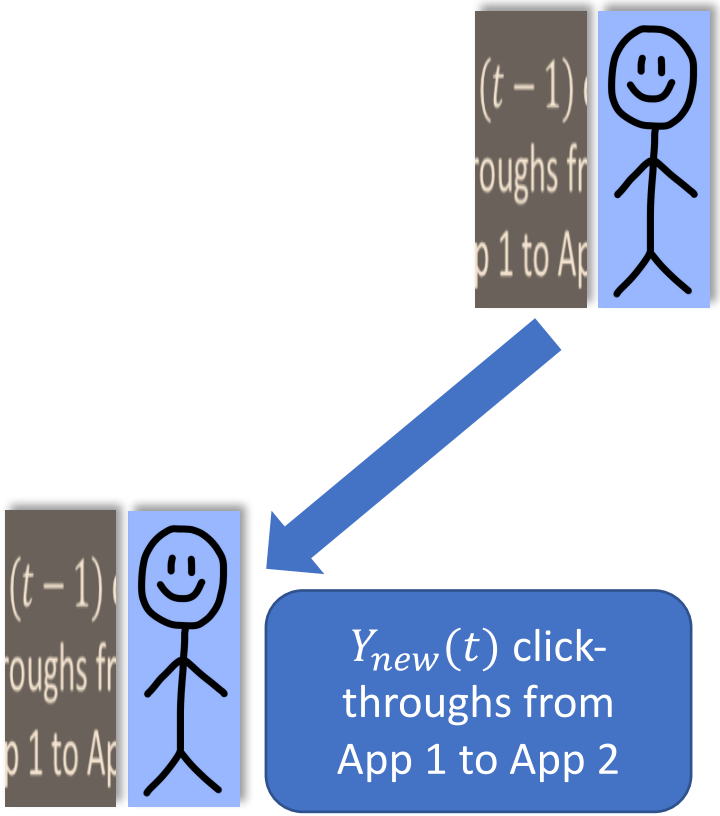


How many new visits are *caused* by the recommender system?

Demand for App 1 is correlated with demand for App 2.

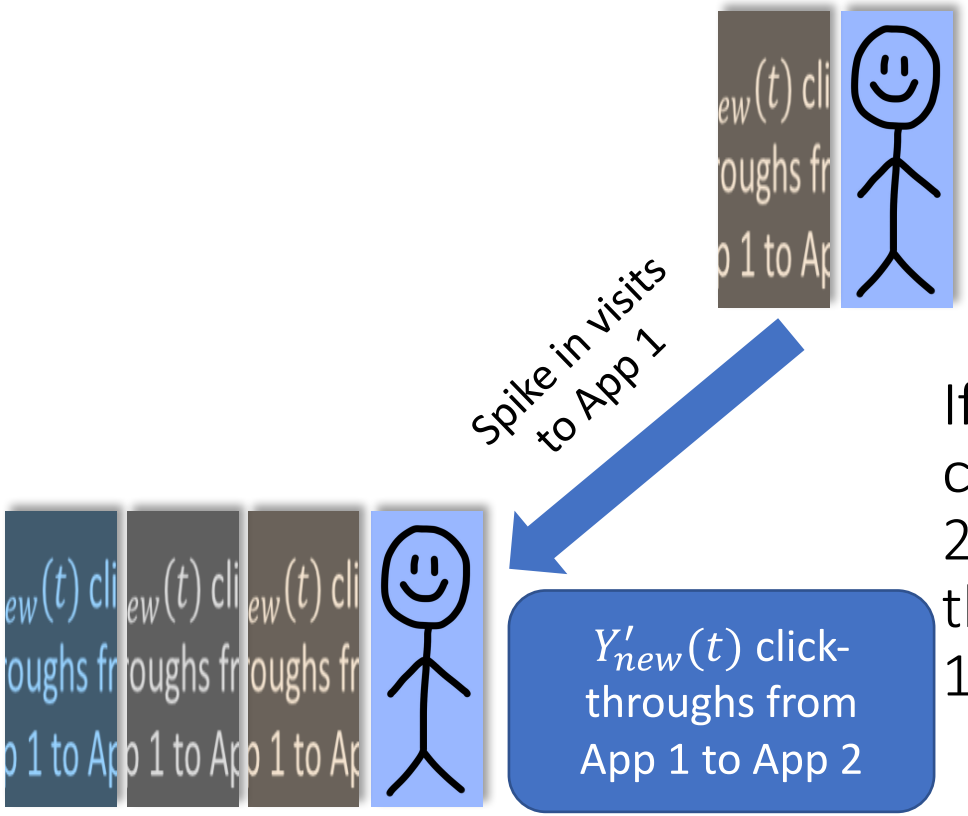
⇒ Users would most likely have visited App 2 even without recommendations.

Traffic on normal days to App 1



Cannot say much about the causal effect of recommendations from App 1.

External shock brings as-if random users to App1



$Y_{old}(t-1)$ click-throughs from App 1 to App 2

Spike in visits to App 1

$Y'_{new}(t)$ click-throughs from App 1 to App 2

If demand for App 2 remains constant, additional views to App 2 would not have happened had these new users not visited App 1.

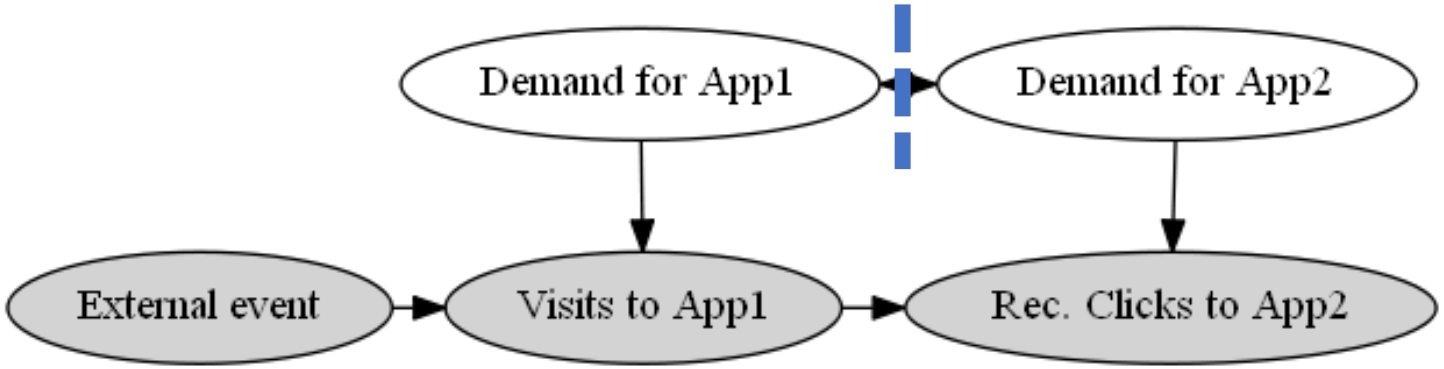
$$Causal\ clicks = Y'_{new}(t) - Y_{old}(t-1)$$

Sharma-Hofman-Watts (2015)

Exploiting sudden variation in traffic to App 1

To compute Causal CTR of Visits to App1 on Visits to App2:

- Compare observed effect of external event separately on Visits to App1, and on Rec. Clicks to App2.
- Causal click-through rate $= \frac{\Delta(\text{Rec. Click−throughs from App1 to App2})}{\Delta(\text{Visits to App1})}$



Natural experiments: Caveats

Natural experiments are great, but:

- Good natural experiments are hard to find
- They rely on many (untestable) assumptions
- The treated population may not be the one of interest

Natural experiments: Caveats

Natural experiments are great, but:

- Good natural experiments are hard to find
- They rely on many (untestable) assumptions
- The treated population may not be the one of interest

Sometimes we can use *additional data + algorithms* to *automatically find* natural experiments

What can we do without an experiment or natural experiment?

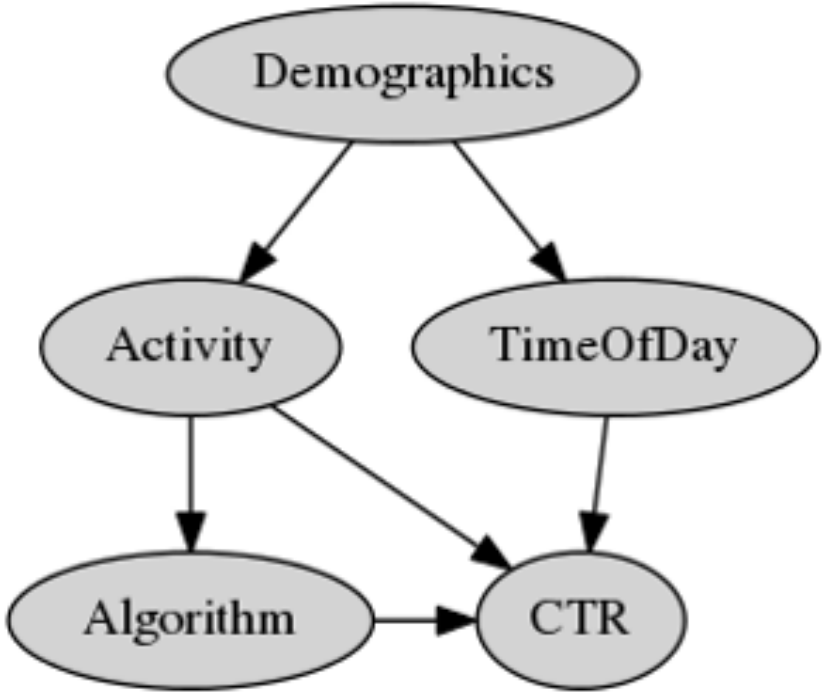
- Condition on the “right” variables

- How to find the right variables?
  - No good answer
  - Depends on domain knowledge
  - Many different frameworks (e.g. Pearlian graphical models and do-calculus)

General method: Conditioning on variables

Intuition: Compare effect of algorithm on similar users.

⇒Compare users with the same activity level.



Steps:

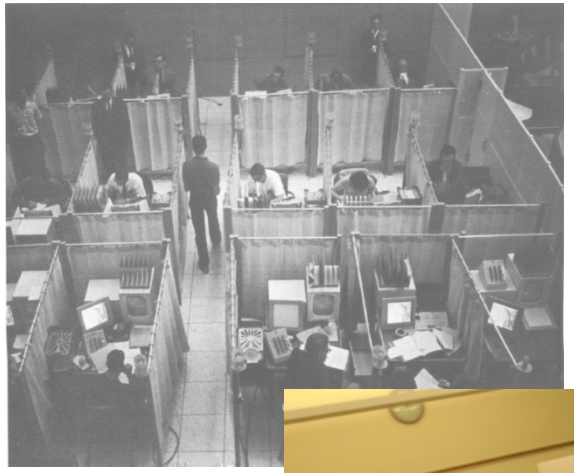
1. Stratify log data based on activity levels.
2. Compare CTR of different algorithms within these strata.

High-dimension? Can derive many complicated methods from this intuition.

Examples of Experiments and Demo

Andrew #2

Behavioral science labs are very limiting



ca. 1960s

ca. 2000s

- High degree of procedural control
- **Optimized for causal inference**

But, **many limitations**:

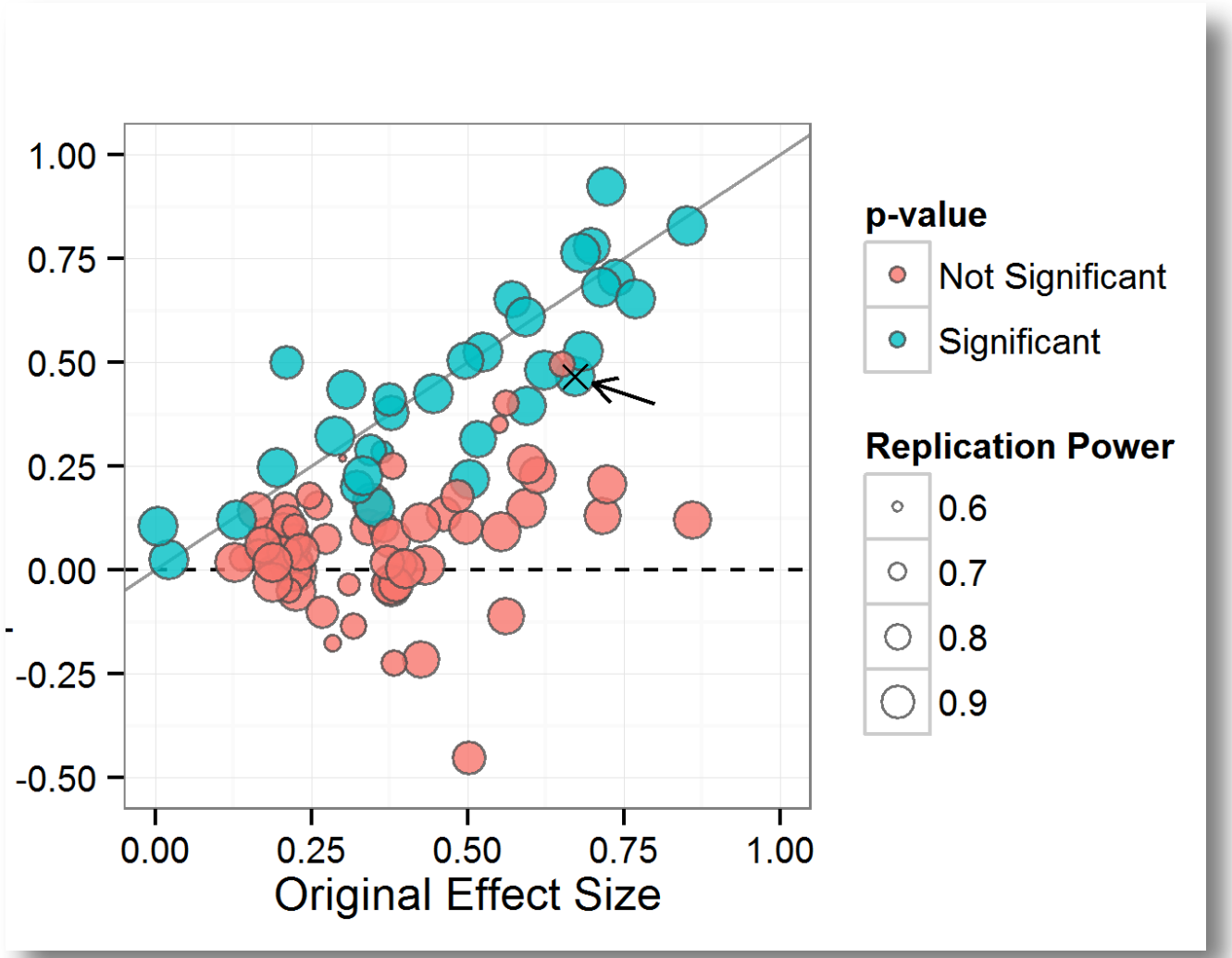
- Artificial environment
- Simple tasks, demand effects
- Homogeneous (WEIRD)* subject pools
- Time/scale limitations
- Expensive, difficult to set up

Poor generalization, expensive, slow

* Western, Educated, Industrialized, Rich, Democratic [Henrich et al. 2010]

Experiments are underpowered

Two-thirds of psychology studies don’t replicate!



Estimating the Reproducibility of Psychological Science (2015)

Most social science experiments aren't social

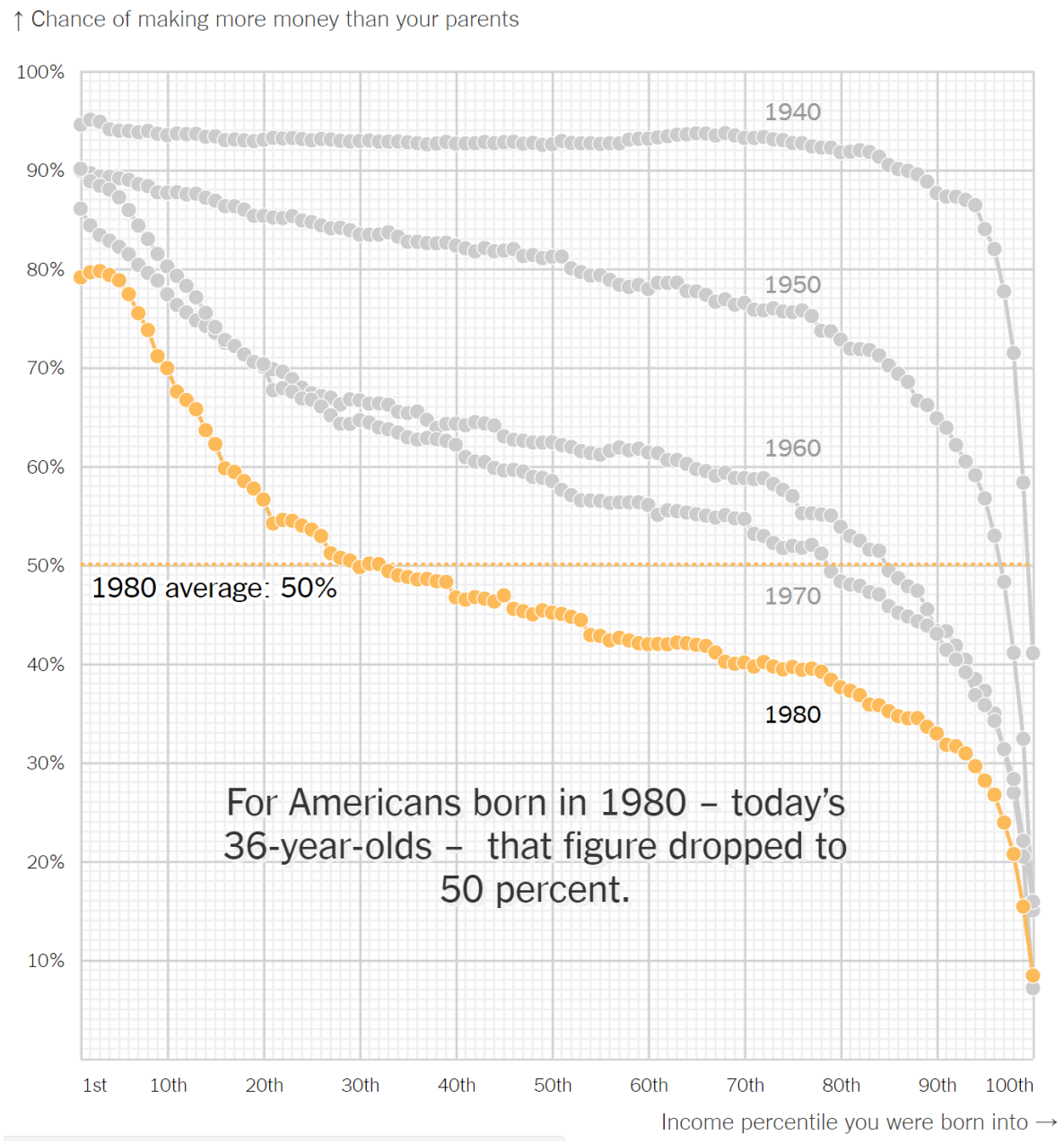
- Vast majority of experiments done with individuals, focused on individual behavior: logistically, it's just easier

- **We actually don't know the causal effects of policies in many large-scale, collective behavior settings!**

Economic Inequality

Social mobility: if you work hard, can you be more successful than your parents?

- How do social safety nets, economic redistribution, etc. affect social mobility and resulting inequality?
- What **policies can we enact** that will improve social mobility or reduce inequality?



The American Dream, Quantified at Last. NYTimes, Dec. 8 2016.

Systems of Governance

I think we should look to countries like Denmark, like Sweden and Norway, and learn from what they have accomplished for their working people

What would happen if we suddenly replaced our current political system with a Scandinavian “socialist state”?

Would everyone be happier?

Or would it be culturally untenable?

The **black box of** macroeconomic policy

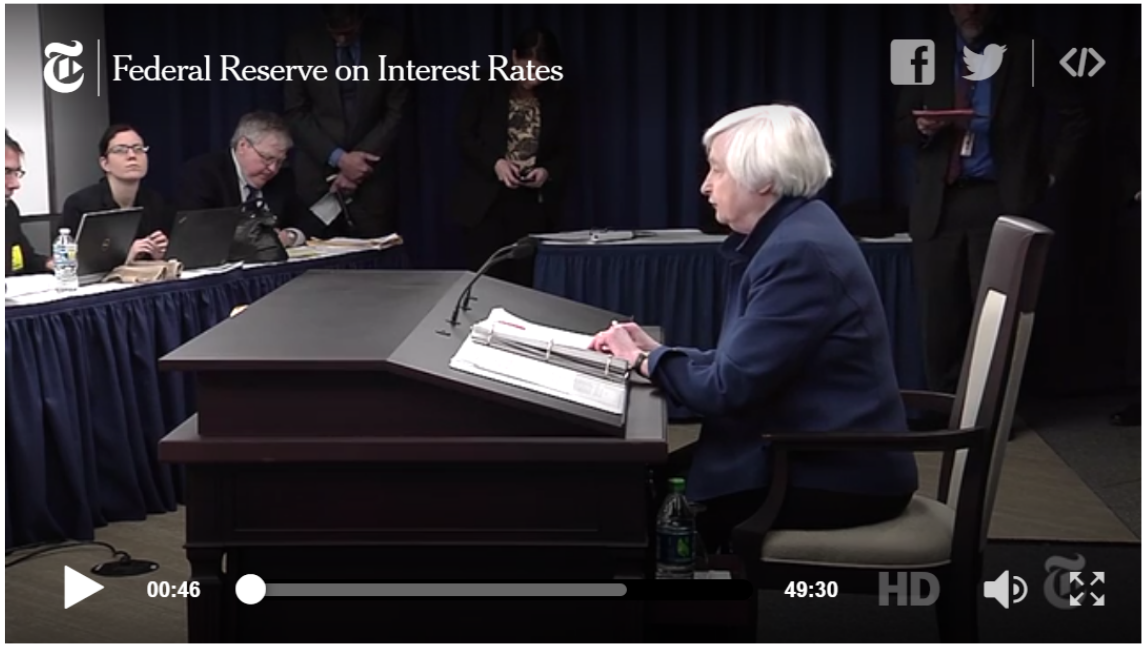
There are no clear theories behind whether **lowering** interest rates, **tightening** interest rates, or **quantitative easing** work as intended.

Given a sufficiently large simulated economy, can we study these experimentally?

ECONOMY

Fed Raises Interest Rates for Third Time Since Financial Crisis

By **BINYAMIN APPELBAUM** MARCH 15, 2017



Janet Yellen, the Federal Reserve chairwoman, announced the board's decision on interest rates. March 15, 2017. Photo by Al Drago/The New York Times. Watch in Times Video »

Embed

The Federal Reserve, which raised its benchmark rate on Wednesday for the

So, about them experiments…

- They’re costly to run
- Yet limited in the types of questions they can answer
- Published experimental research is probably wrong
- They’re far from answering many big important questions

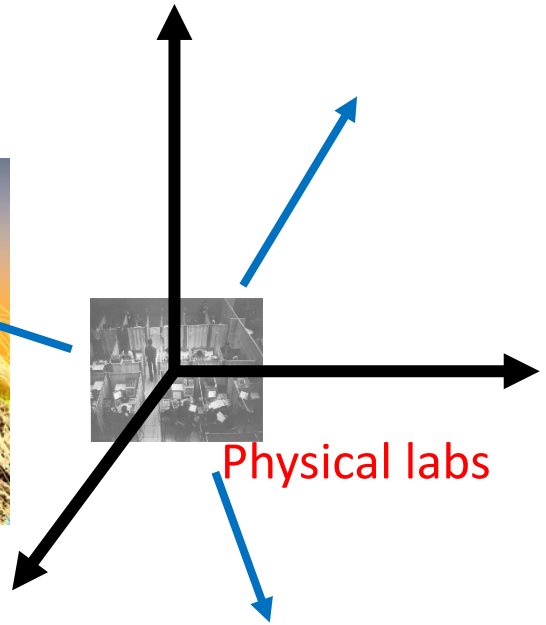
How do we fix this?

Expanding the experiment design space

Complexity, Realism

- Realistic *vs.* abstract, simple tasks
- More precise instrumentation

A software-based "virtual lab" with online participants



Duration, Participation

- Longer periods of time
- Fewer constraints on location

Size, Scale

- More samples of data
- Large-scale social interaction

Behavioral experiments + the Internet

Control

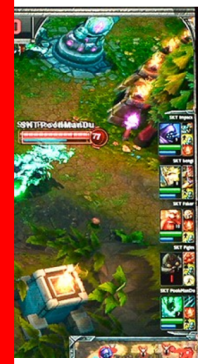
Generalization

The Virtual Lab

Volunteer web labs

Inside the Largest Virtual Psychology Lab in the World

Riot Games wants you to behave yourself when you play League of Legends, so it's turned the game into a virtual lab



Unwitting labs in the field

Prisoner's Dilemma

Cooperate

Defect

Split – Split	50% - 50%
Steal – Split	100% - 0%
Steal – Steal	0% - 0%

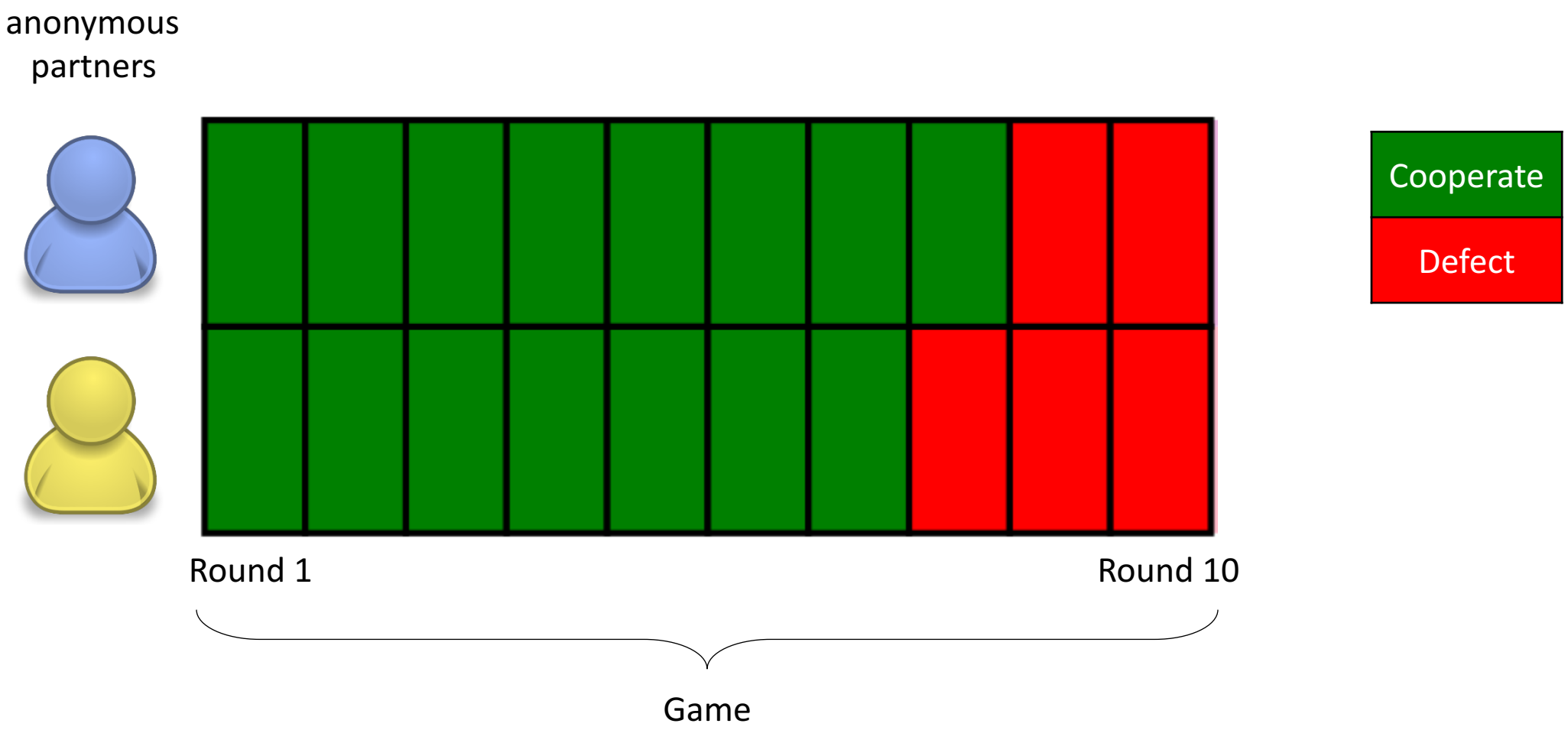
Q: What happens to people's behavior when they are in a social dilemma for a long time?

Also known as: real life.

A: According to lab experiments, we don't know. It seems like people start stabbing each other in the back when they play this game

…but the world seems to (mostly) be doing okay.

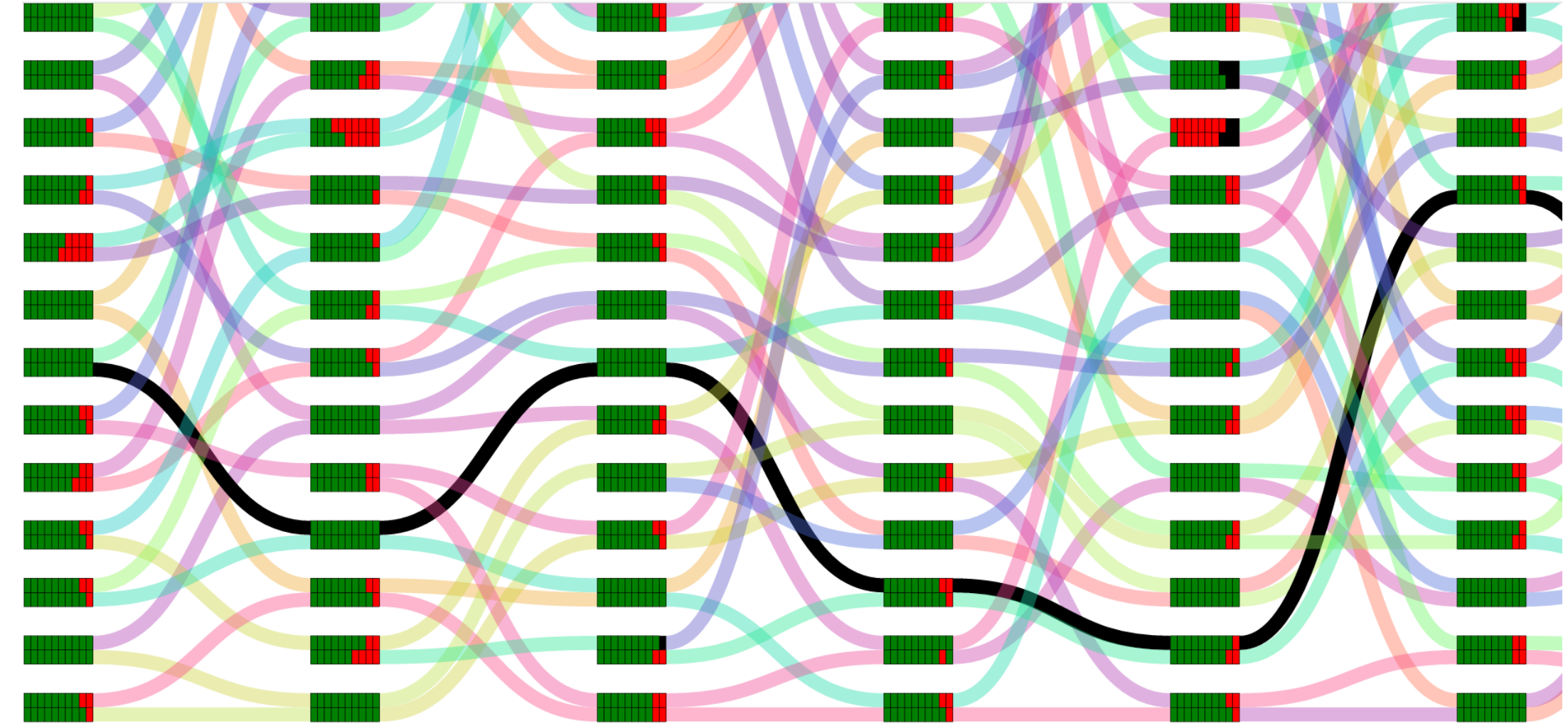
Longitudinal behavior in a social dilemma



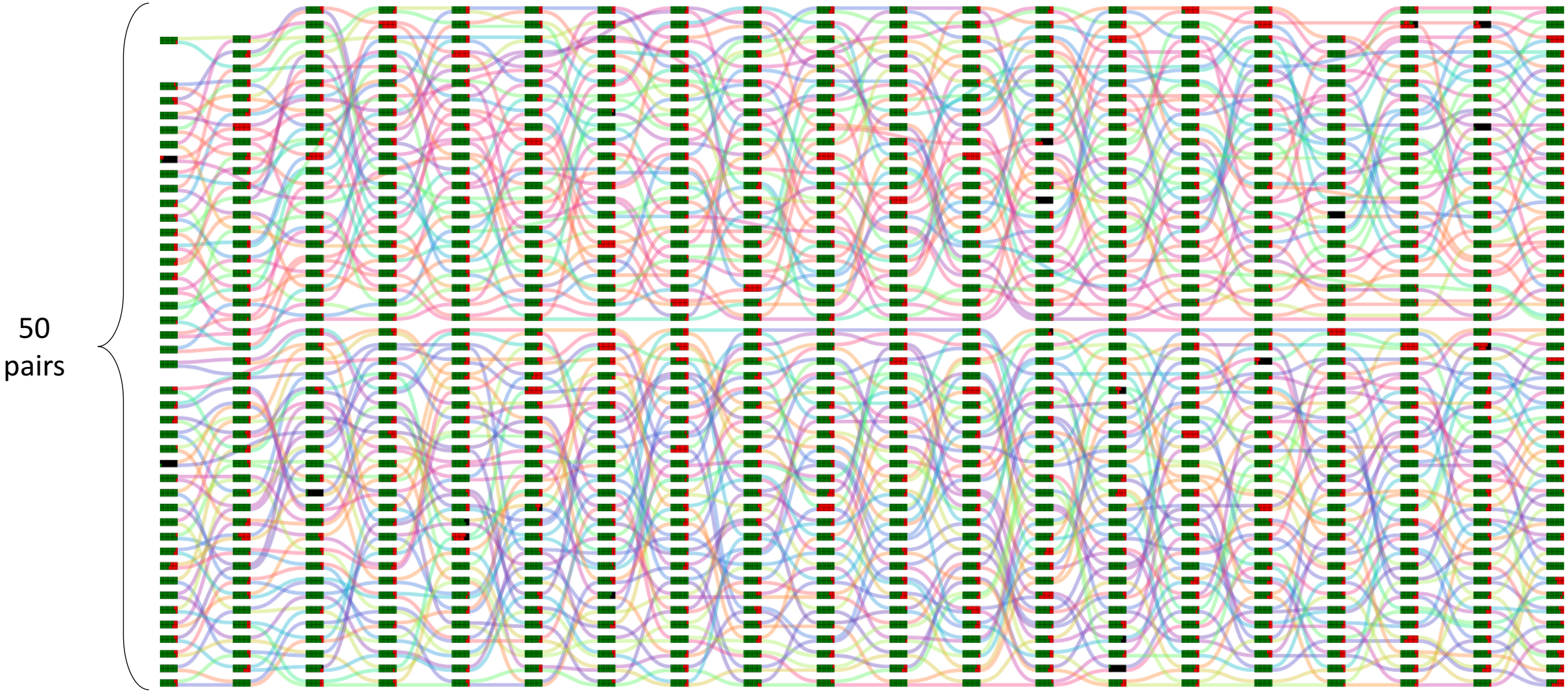
Random rematching across games

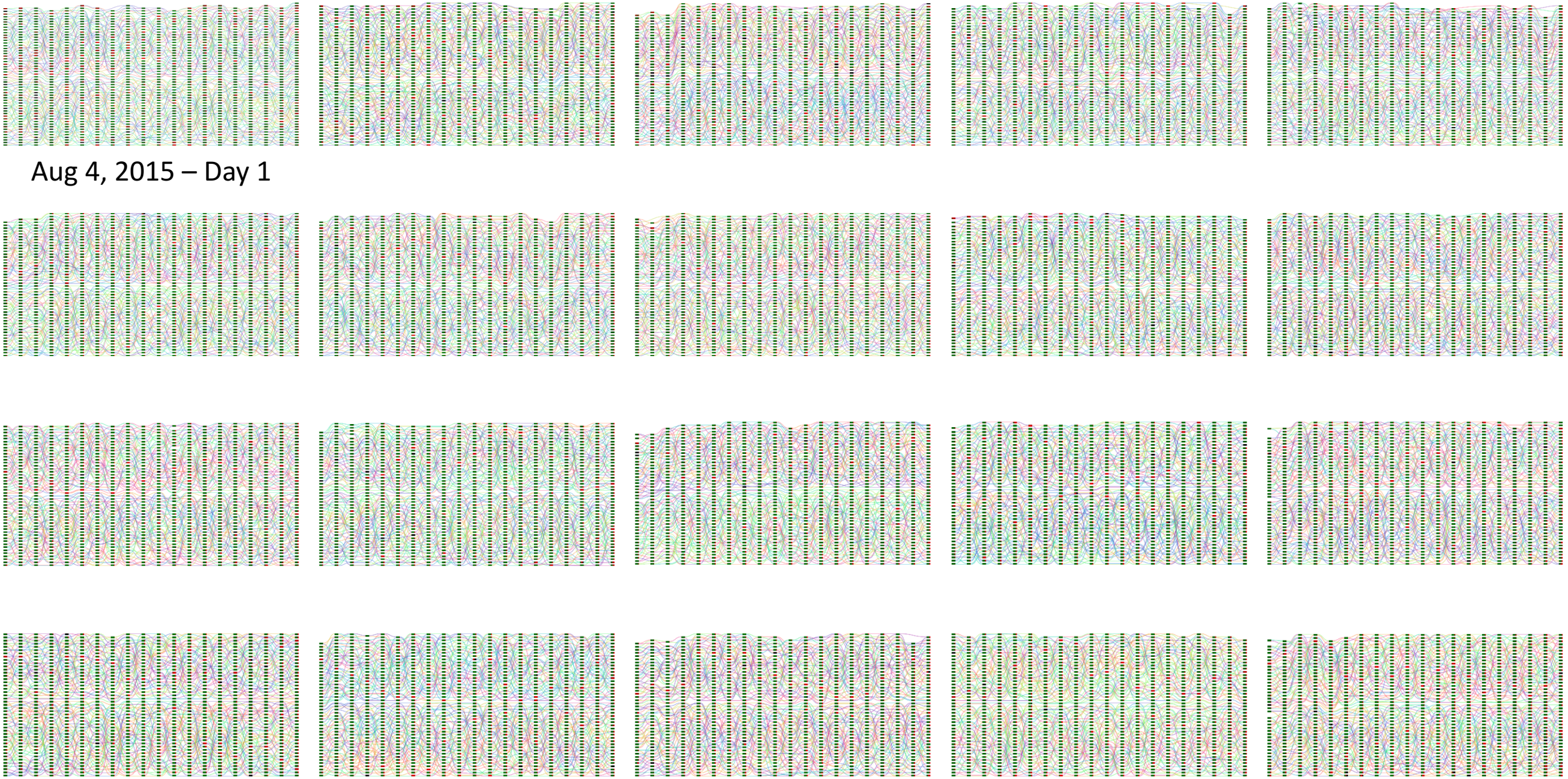
Game 1

Random rematching across games



One experiment session – 20 games





Aug 4, 2015 – Day 1

Aug 31, 2015 – Day 20

Demo time!

You get to play prisoner’s dilemma with each other!

The person with the most money at the end wins.

Navigate your laptop or mobile browser to:

http://skygeirr.andrewmao.net:3000

Demo

- You will be the participants in this prisoner's dilemma experiment!
- Log in using the menu, choose "main"
- Instructions: you will be playing IPD with repeated partners
  - Loses some context, so I will explain. (wait a bit)
  - Press begin when you are ready
- Admin console; shows all user connections
- Lobby, starting a game
  - Play some rounds with a partner; timeouts and stuff
- Live experiments timer and graphs
- Disconnections, slow players
- One-way mirror of slow games – very useful to understand behavior
- Visualization of data at end

A “supercollider” for social science

Realistic settings for **collective** behavior…

with **precise instrumentation** and measurement…

studied **longitudinally** over periods of time…

with a high degree of **experimental control**.

Closing thoughts

Large-scale *observational data* is useful for building *predictive models* of a *static world*

Closing thoughts

But without appropriate *random variation*, it's hard to *predict what happens when you change something* in the world

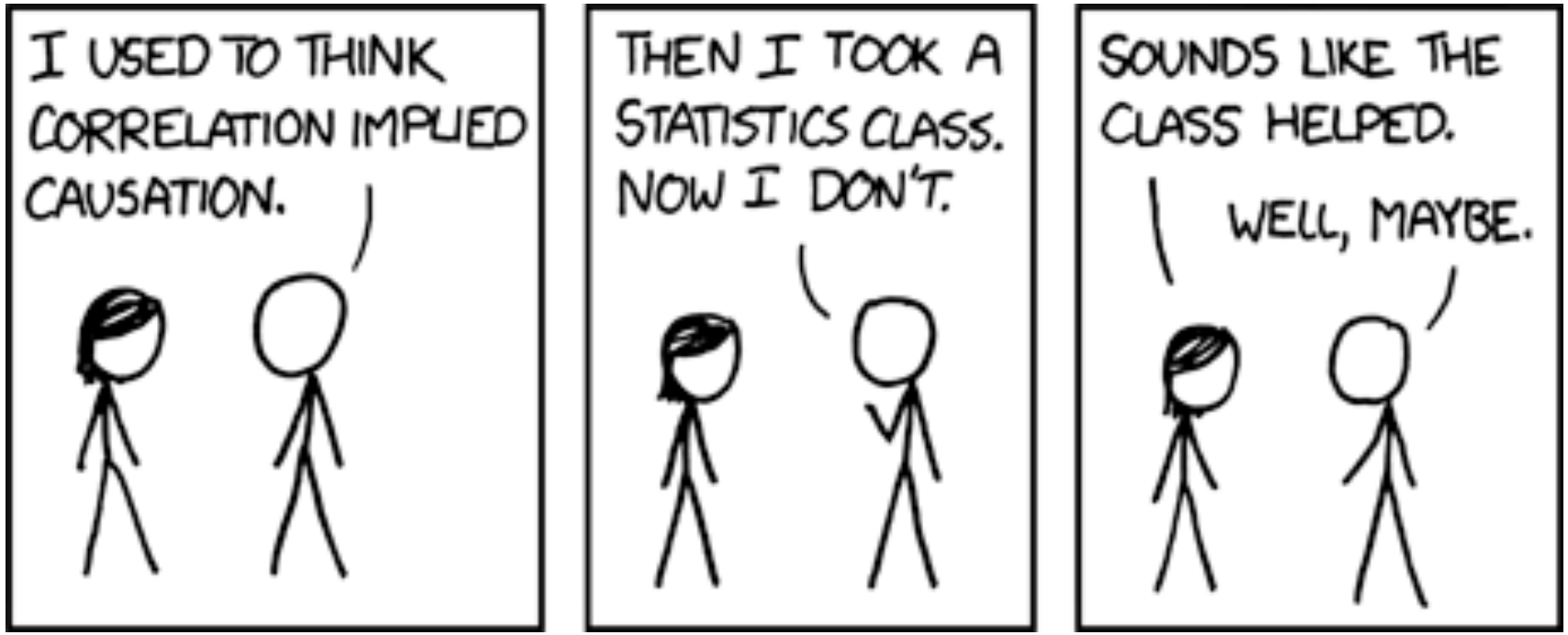
Closing thoughts

Randomized experiments are like *custom-made datasets* to answer a specific question

Closing thoughts

Additional data + algorithms can help us *discover* and *analyze* these examples in the wild

Causality is tricky!



"Correlation doesn't imply causation, but it does waggle its eyebrows suggestively and gesture furtively while mouthing 'look over there'"

https://www.xkcd.com/552/