Causality & Experiments

MODELING SOCIAL DATA

JAKE HOFMAN

COLUMBIA UNIVERSITY

Prediction

Seeing: Make a forecast, leaving the world as it is

vs.

Causation

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

Prediction

Seeing: Make a forecast, leaving the world as it is

(seeing my neighbor with an umbrella might predict rain)

vs.

Causation

Doing: 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.

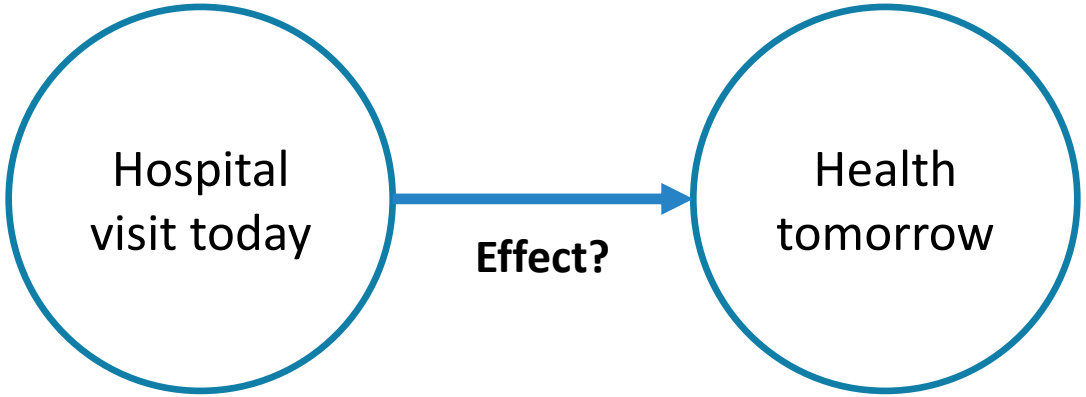
- How does education impact future earnings?
- What is the effect of advertising on sales?
- How does hospitalization affect health?

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

John Stuart Mill (1843)

Example: Hospitalization on health

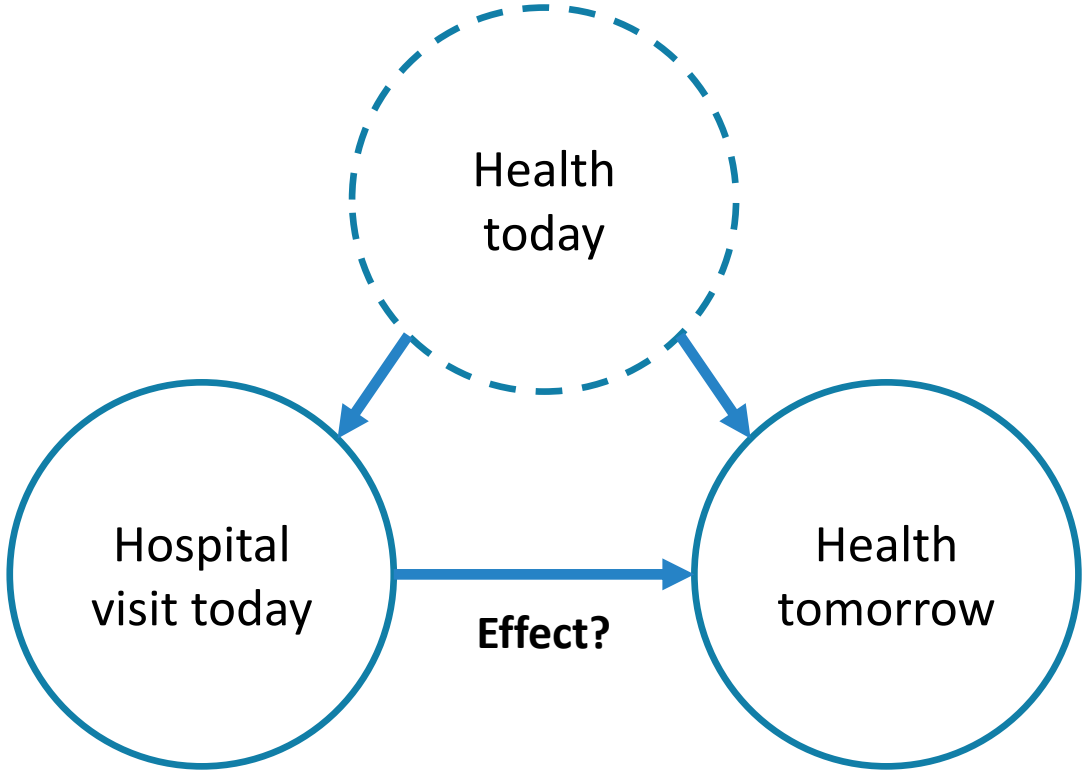
What's wrong with estimating this model from observational data?



Arrow means "X causes Y"

Confounds

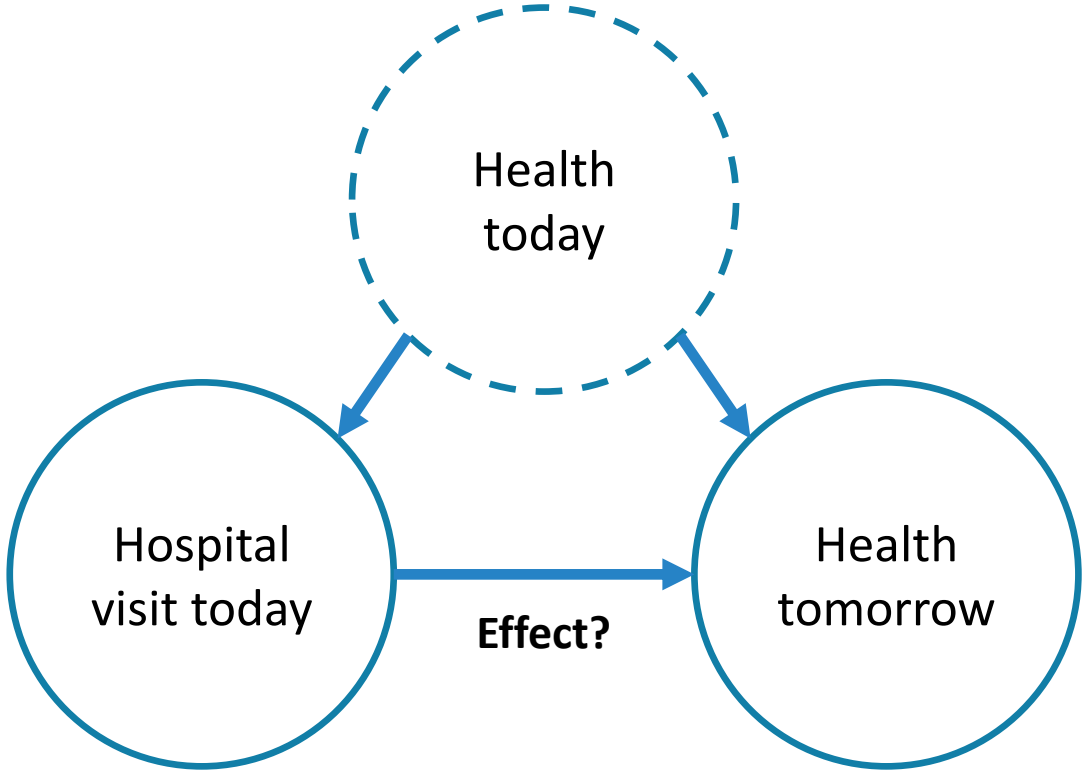
The effect and cause might be *confounded* by a common cause, and be *changing together* as a result



Dashed circle means “unobserved”

Confounds

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



A counterfactual (what-if) definition

What if you would have acted differently?

E.g., how does the health of a hospitalized patient compare to their health if they would have stayed home?

We only get to observe one of these outcomes, which is the *fundamental problem of causal inference*

How does this differ from an observational estimate?

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

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

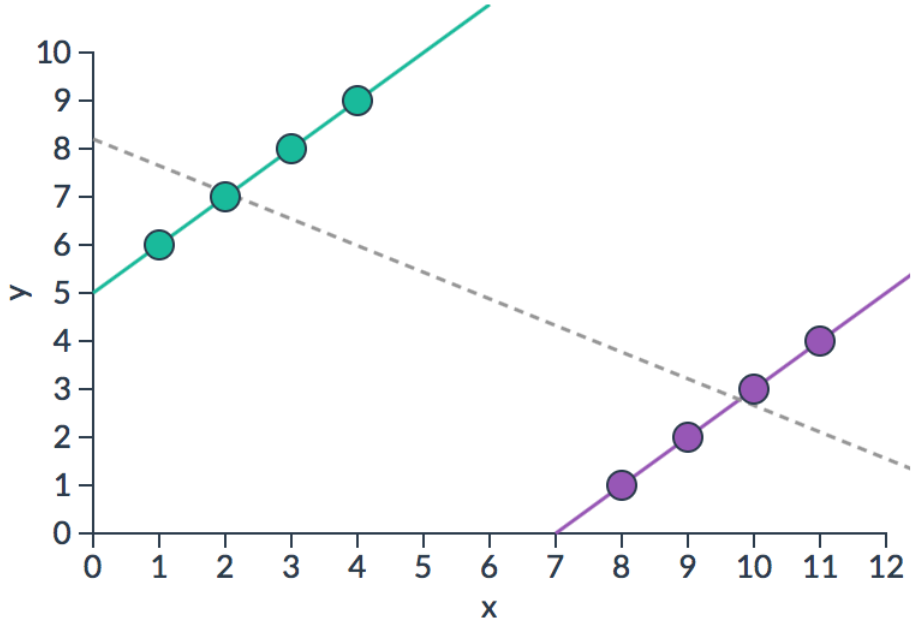
The observed difference in health tomorrow is:

Observed difference = Causal effect – Selection bias

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

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

Simpson's paradox

So which is right, the aggregated or the partitioned?

It depends on the causal mechanism

	Men		Women	
	Applicants	Admitted	Applicants	Admitted
Total	8442	**44%**	4321	35%

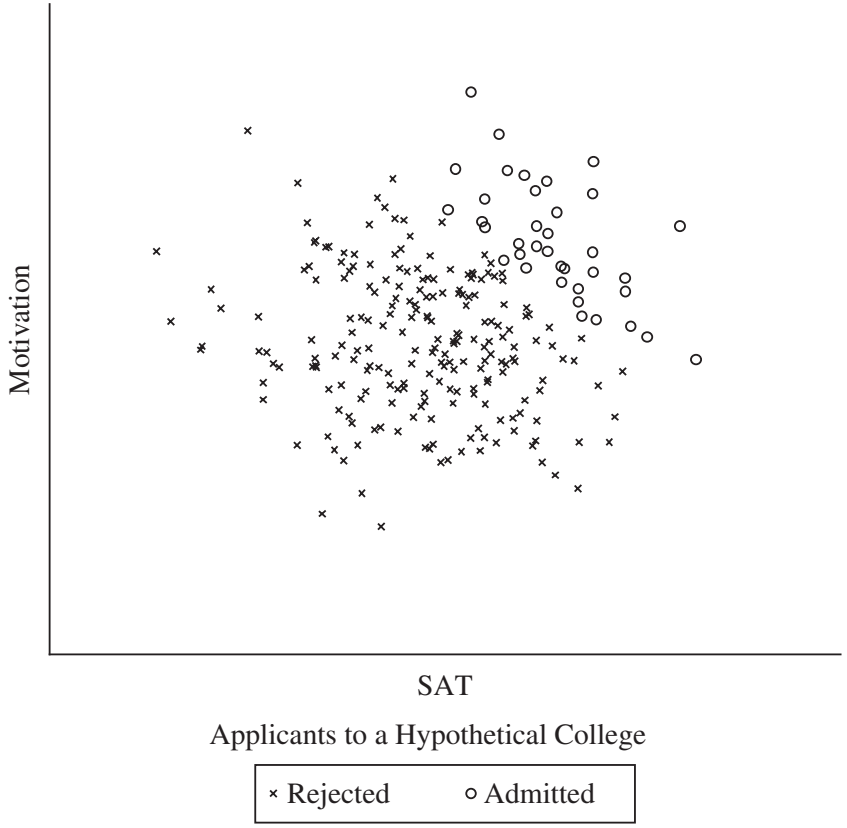
Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	***825***	62%	108	**82%**
B	***560***	63%	25	**68%**
C	325	**37%**	***593***	34%
D	417	33%	375	**35%**
E	191	**28%**	***393***	24%
F	373	6%	341	**7%**

https://en.wikipedia.org/wiki/Simpson%27s_paradox

Simpson's paradox

So which is right, the aggregated or the partitioned?

It depends on the causal mechanism



Morgan and Winship (2015)

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

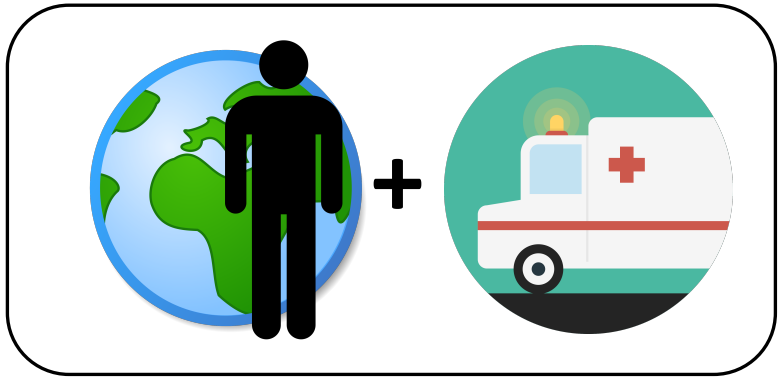
-GEORGE BOX

Controlled experiments

Counterfactuals

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

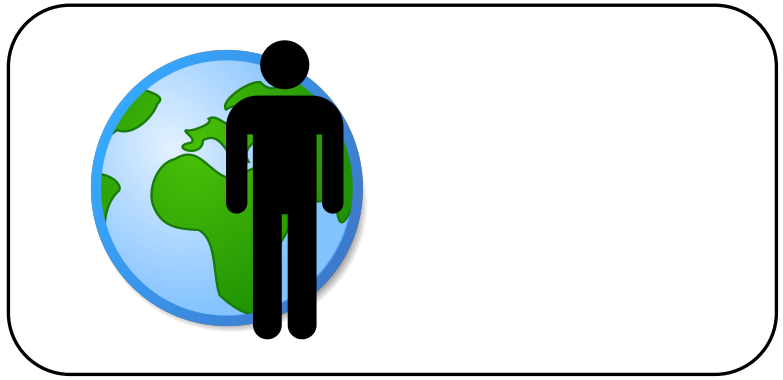
Reality



(what happened)

vs

Counterfactual



(what would have happened)

The ideal causal estimate

CLONE EACH PERSON

SEND ONE COPY TO THE HOSPITAL, MAKE THE OTHER STAY HOME

MEASURE THE DIFFERENCE IN HEALTH BETWEEN THE COPIES

Scott Kelly Spent a Year in Orbit. His Body Is Not Quite the Same.

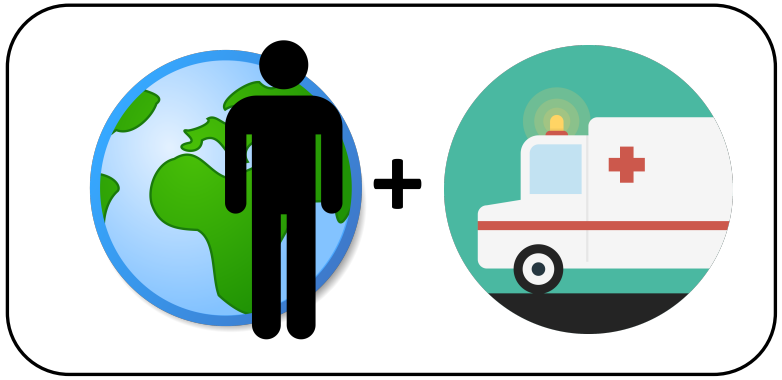
NASA scientists compared the astronaut to his earthbound twin, Mark. The results hint at what humans will have to endure on long journeys through space.

But this might be confounded for various reasons---e.g., Mark has a different diet than Scott

Counterfactuals

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

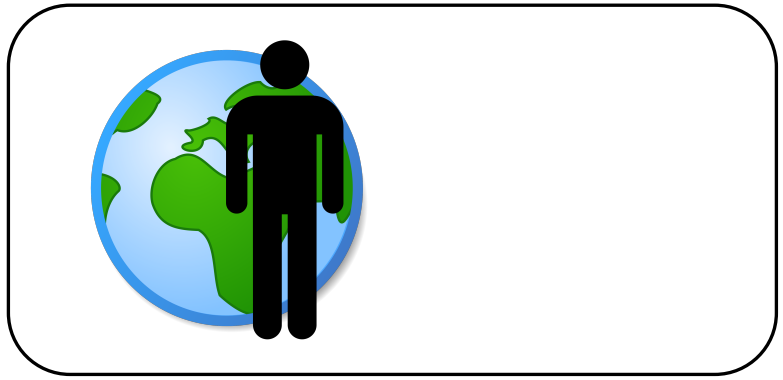
Reality



(what happened)

vs

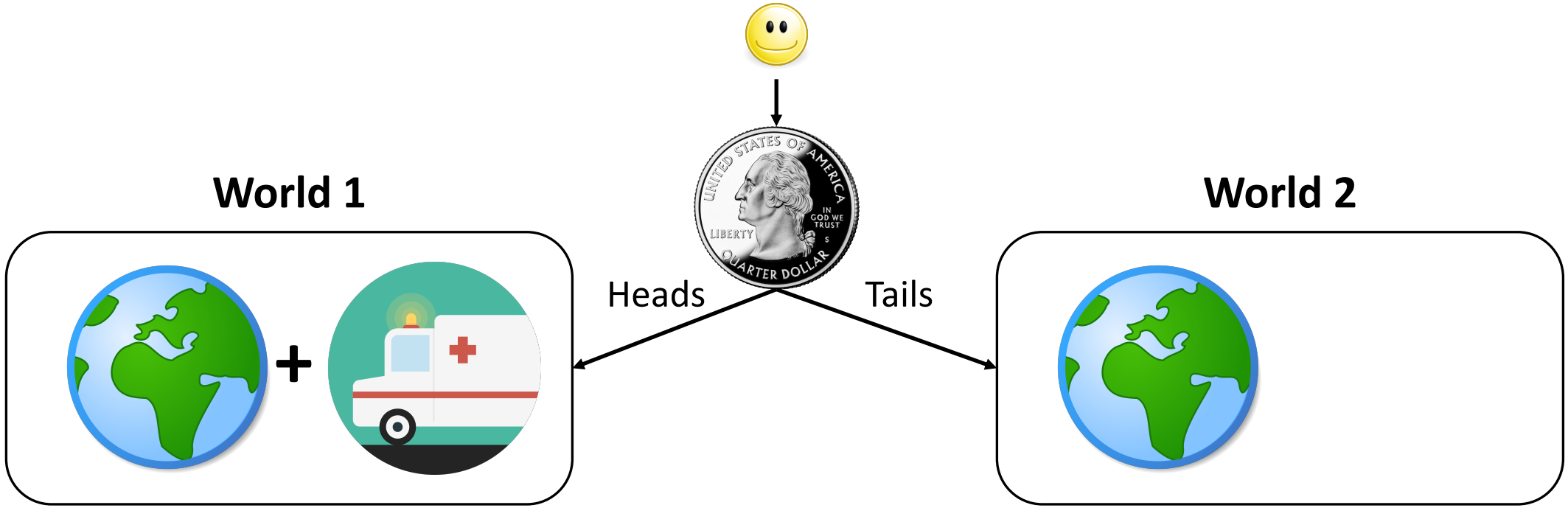
Counterfactual



(what would have happened)

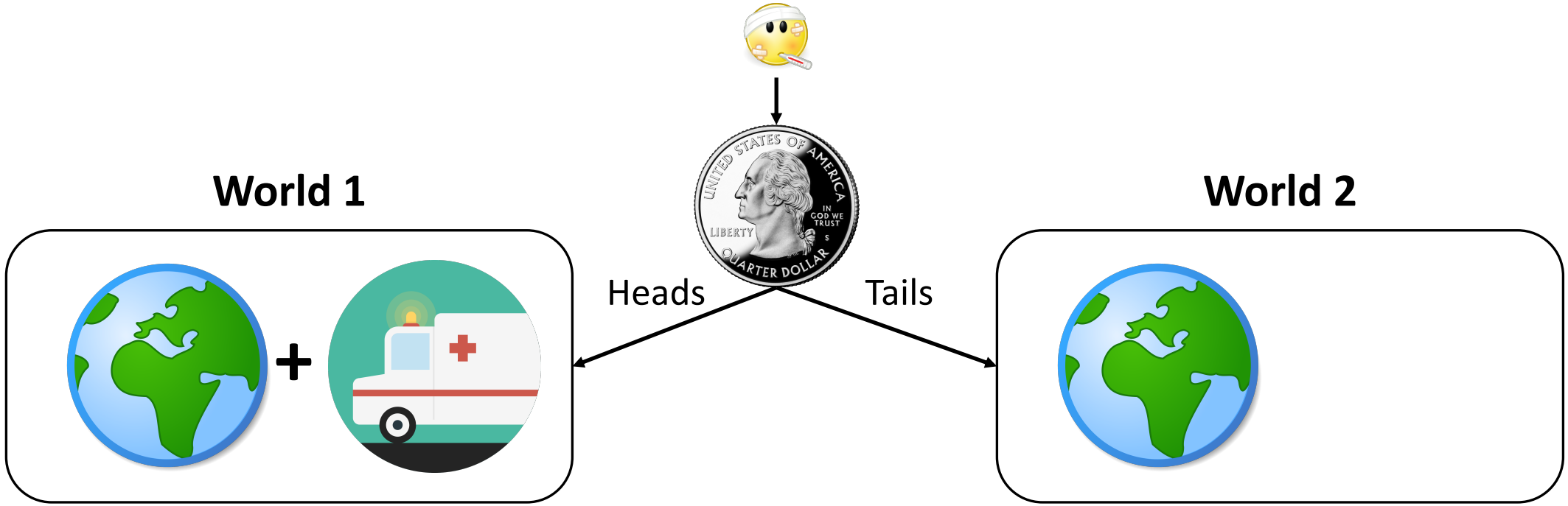
Random assignment

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



Random assignment

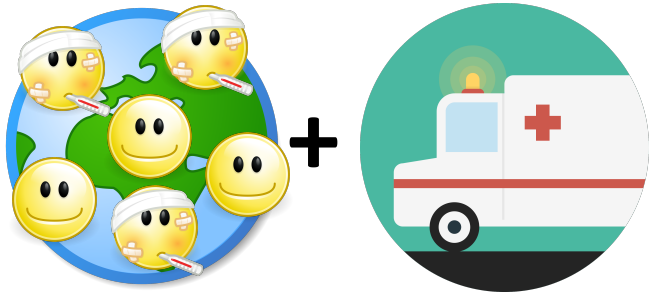
We can use randomization to create two groups that differ only in which treatment they receive, restoring symmetry



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:

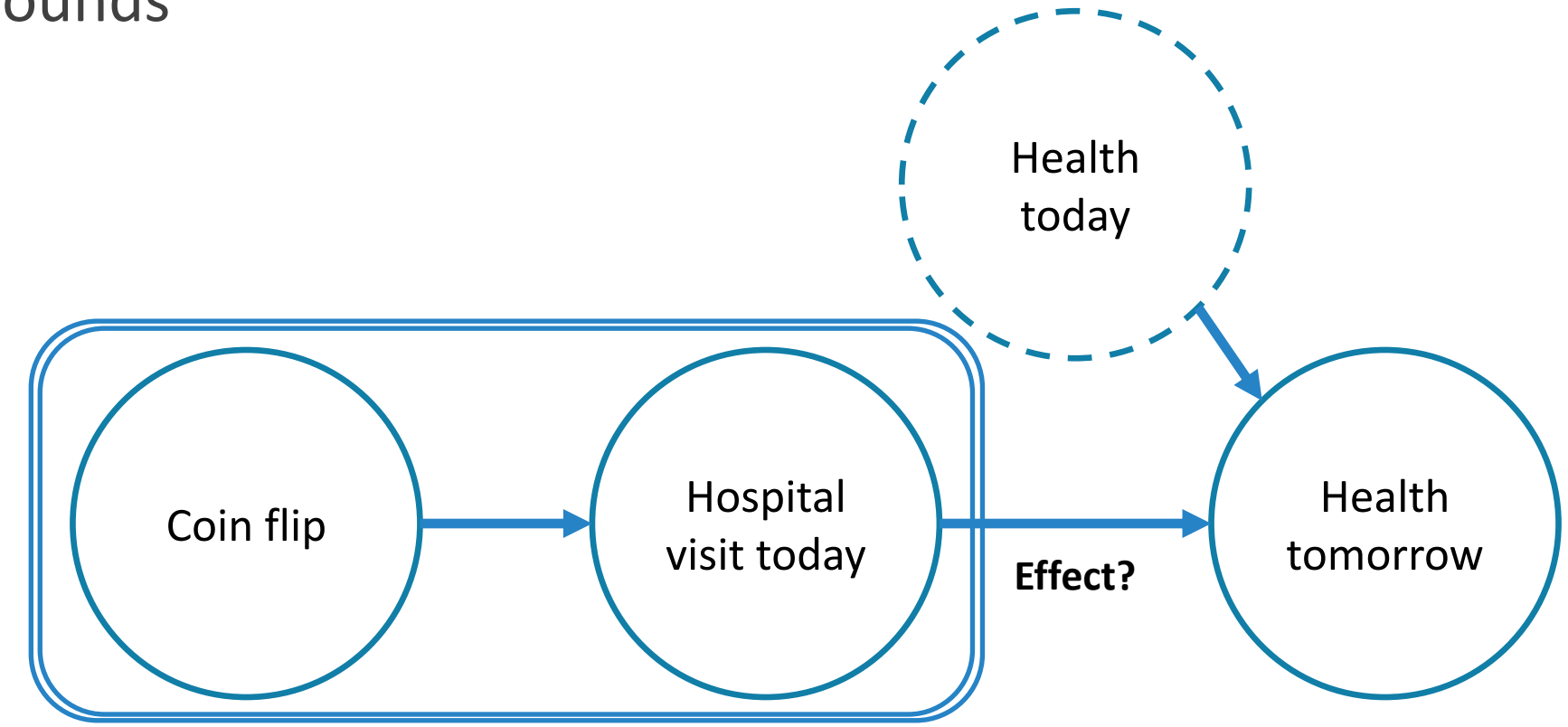
Observed difference = Causal effect – ~~Selection bias~~

= Causal effect

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

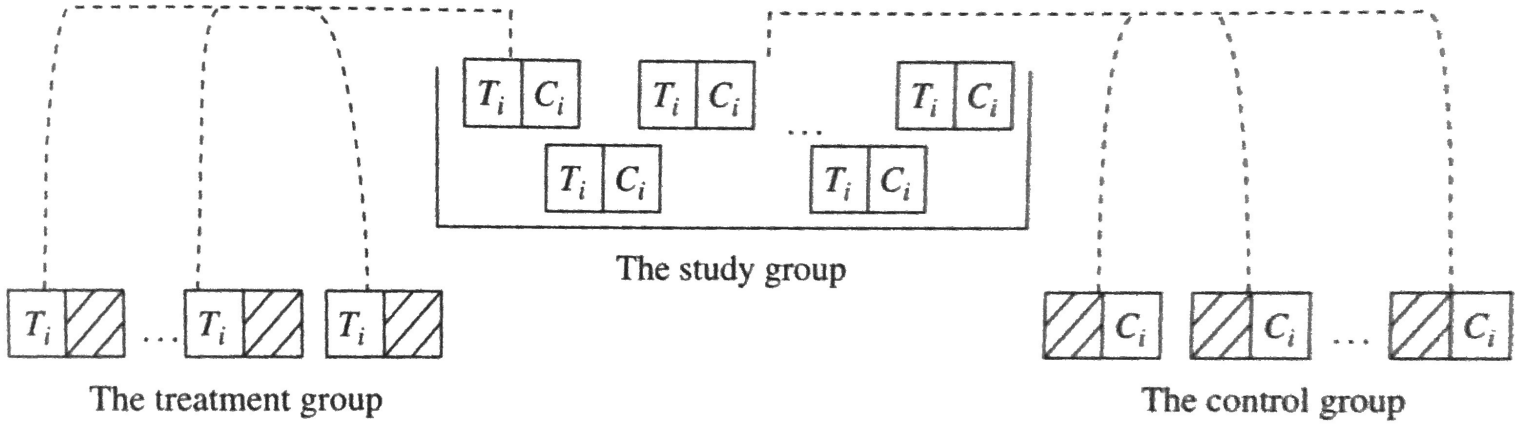
Random assignment

Random assignment determines the treatment independent of any confounds



Double lines mean "intervention"

Random assignment



The Neyman model.

Here, we are drawing at random from a box with N tickets. Each ticket represents one unit in the natural-experimental study group. Here, T_i and C_i are the potential outcomes under treatment and control, respectively. If unit i is sampled into treatment, we observe T_i but not C_i; if unit i is assigned to control, we observe C_i but not T_i. The average of the T_is in the treatment group estimates the average of all the T_is in the box, while the average of the C_is in the control group estimates the average of all the C_is.

Dunning (2012)

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
- Effects in the lab can differ from real-world effects
- Inevitably people deviate from their random assignments

Validity of experiments

INTERNAL VALIDITY

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

Was the study double-blind? Did doctors give the experimental drug to some especially sick patients (breaking randomization) hoping that it would save them? Or treat patients differently based on whether they got the drug or not?

EXTERNAL VALIDITY

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 or when tried on a different population of patients?

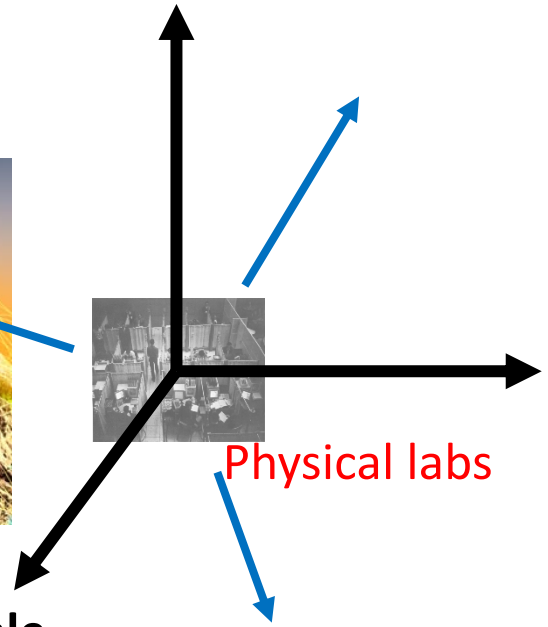
Slide thanks to Andrew Mao

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

Slide thanks to Andrew Mao

Practical Guide to Controlled Experiments on the Web: Listen to Your Customers not to the HiPPO

Ron Kohavi
Microsoft
One Microsoft Way
Redmond, WA 98052
ronnyk@microsoft.com

Randal M. Henne
Microsoft
One Microsoft Way
Redmond, WA 98052
rhenne@microsoft.com

Dan Sommerfield
Microsoft
One Microsoft Way
Redmond, WA 98052
dans@microsoft.com

5.2 Trust and Execution

5.2.1 Run Continuous A/A Tests

Run A/A tests (see Section 3.1) and validate the following.

1. Are users split according to the planned percentages?
2. Is the data collected matching the system of record?
3. Are the results showing non-significant results 95% of the time?

Continuously run A/A tests in parallel with other experiments.

5.2.2 Automate Ramp-up and Abort

As discussed in Section 3.3, we recommend that experiments ramp-up in the percentages assigned to the Treatment(s). By doing near-real-time analysis, experiments can be auto-aborted if a treatment is statistically significantly underperforming relative to the Control. An auto-abort simply reduces the percentage of users assigned to a treatment to zero. By reducing the risk in exposing many users to egregious errors, the organization can make bold bets and innovate faster. Ramp-up is quite easy to do in online environments, yet hard to do in offline studies. We have seen no mention of these practical ideas in the literature, yet they are extremely useful.

5.2.3 Determine the Minimum Sample Size

Decide on the statistical power, the effect you would like to detect, and estimate the variability of the OEC through an A/A test. Based on this data you can compute the minimum sample size needed for the experiment and hence the running time for your web site. A common mistake is to run experiments that are underpowered. Consider the techniques mentioned in Section 3.2 point 3 to reduce the variability of the OEC.

5.2.4 Assign 50% of Users to Treatment

One common practice among novice experimenters is to run new variants for only a small percentage of users. The logic behind that decision is that in case of an error only few users will see a bad treatment, which is why we recommend Treatment ramp-up. In order to maximize the power of an experiment and minimize the running time, we recommend that 50% of users see each of the variants in an A/B test. Assuming all factors are fixed, a good approximation for the multiplicative increase in running time for an A/B test relative to 50%/50% is $1/(4p(1-p))$ where the treatment receives portion p of the traffic. For example, if an experiment is run at 99%/1%, then it will have to run about 25 times longer than if it ran at 50%/50%.

5.2.5 Beware of Day of Week Effects

Even if you have a lot of users visiting the site, implying that you could run an experiment for only hours or a day, we strongly recommend running experiments for at least a week or two, then continuing by multiples of a week so that day-of-week effects can be analyzed. For many sites the users visiting on the weekend represent different segments, and analyzing them separately may lead to interesting insights. This lesson can be generalized to other time-related events, such as holidays and seasons, and to different geographies: what works in the US may not work well in France, Germany, or Japan.

Experimental evidence of massive-scale emotional contagion through social networks

Adam D. I. Kramer[a,1], Jamie E. Guillory[b,2], and Jeffrey T. Hancock[b,c]

[a]Core Data Science Team, Facebook, Inc., Menlo Park, CA 94025; and Departments of [b]Communication and [c]Information Science, Cornell University, Ithaca, NY 14853

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved March 25, 2014 (received for review October 23, 2013)

Significance

We show, via a massive ($N = 689,003$) experiment on Facebook, that emotional states can be transferred to others via emotional contagion, leading people to experience the same emotions without their awareness. We provide experimental evidence that emotional contagion occurs without direct interaction between people (exposure to a friend expressing an emotion is sufficient), and in the complete absence of nonverbal cues.

Editorial Expression of Concern and Correction

PSYCHOLOGICAL AND COGNITIVE SCIENCES

PNAS is publishing an Editorial Expression of Concern regarding the following article: "Experimental evidence of massive-scale emotional contagion through social networks," by Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock, which appeared in issue 24, June 17, 2014, of *Proc Natl Acad Sci USA* (111:8788–8790; first published June 2, 2014; 10.1073/pnas.1320040111). This paper represents an important and emerging area of social science research that needs to be approached with sensitivity and with vigilance regarding personal privacy issues.

Questions have been raised about the principles of informed consent and opportunity to opt out in connection with the research in this paper. The authors noted in their paper, "[The work] was consistent with Facebook's Data Use Policy, to which all users agree prior to creating an account on Facebook, constituting informed consent for this research." When the authors prepared their paper for publication in PNAS, they stated that: "Because this experiment was conducted by Facebook, Inc. for internal purposes, the Cornell University IRB [Institutional Review Board] determined that the project did not fall under Cornell's Human Research Protection Program." This statement has since been confirmed by Cornell University.

Obtaining informed consent and allowing participants to opt out are best practices in most instances under the US Department of Health and Human Services Policy for the Protection of Human Research Subjects (the "Common Rule"). Adherence to the Common Rule is PNAS policy, but as a private company Facebook was under no obligation to conform to the provisions of the Common Rule when it collected the data used by the authors, and the Common Rule does not preclude their use of the data. Based on the information provided by the authors, PNAS editors deemed it appropriate to publish the paper. It is nevertheless a matter of concern that the collection of the data by Facebook may have involved practices that were not fully consistent with the principles of obtaining informed consent and allowing participants to opt out.

Inder M. Verma
Editor-in-Chief

PSYCHOLOGICAL AND COGNITIVE SCIENCES

Correction for "Experimental evidence of massive-scale emotional contagion through social networks," by Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock, which appeared in issue 24, June 17, 2014, of *Proc Natl Acad Sci USA* (111:8788–8790; first published June 2, 2014; 10.1073/pnas.1320040111).

The authors note that, "At the time of the study, the middle author, Jamie E. Guillory, was a graduate student at Cornell University under the tutelage of senior author Jeffrey T. Hancock, also of Cornell University (Guillory is now a postdoctoral fellow at Center for Tobacco Control Research and Education, University of California, San Francisco, CA 94143)." The author and affiliation lines have been updated to reflect the above changes and a present address footnote has been added. The online version has been corrected.

The corrected author and affiliation lines appear below.

Adam D. I. Kramer[a,1], Jamie E. Guillory[b,2], and Jeffrey T. Hancock[b,c]

[a]Core Data Science Team, Facebook, Inc., Menlo Park, CA 94025; and Departments of [b]Communication and [c]Information Science, Cornell University, Ithaca, NY 14853

[1]To whom correspondence should be addressed. Email: akramer@fb.com.

[2]Present address: Center for Tobacco Control Research and Education, University of California, San Francisco, CA 94143.

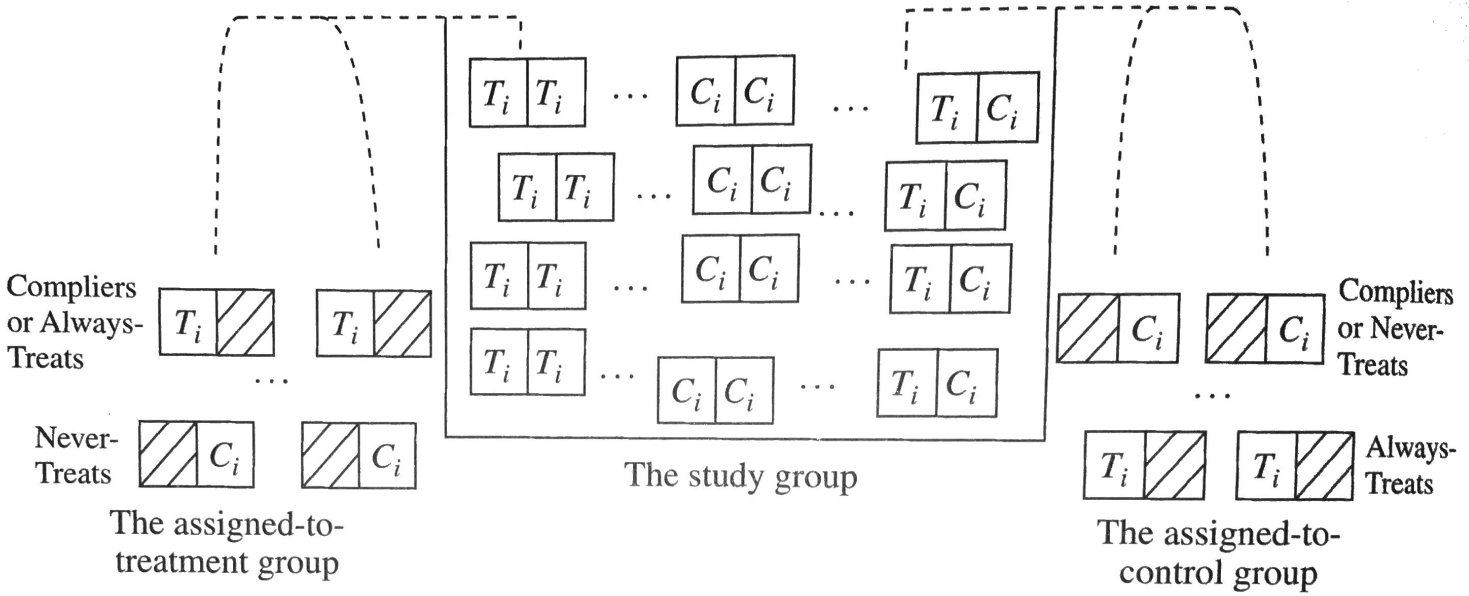
www.pnas.org/cgi/doi/10.1073/pnas.1412583111

Experiments with non-compliance

Table 1: Deaths from Breast Cancer and Other Causes (HIP study).

	Group size	Deaths from breast cancer	Death rate per 1,000 women	Deaths from other causes	Death rate from other causes, per 1,000 women
Assigned to treatment:					
Accepted Screening	20,200	23	1.14	428	21.19
Refused Screening	10,800	16	1.48	409	37.87
Total	31,000	39	1.26	837	27.00
Assigned to control:					
Would have accepted screening	20,200	–	–	–	–
Would have refused screening	10,800	–	–	–	–
Total	31,000	63	2.03	879	28.35

The table is adapted from Freedman (2005: 4, Table 1).



Noncompliance under the Neyman model.

A model of natural-experimental crossover. Each ticket in the box represents one unit in the natural-experimental study group. Every ticket has two fields, one representing potential outcomes under assignment to treatment and the other potential outcomes under assignment to control. Tickets with T_i in both fields are "Always-Treats." Tickets with C_i in both fields are "Never-Treats." Tickets with T_i in one field and C_i in the other are "Compliers." (Defiers are ruled out by assumption.) Here, we draw at random without replacement from a box with N tickets, placing $n < N$ tickets in the assigned-to-treatment group and $m = N - n$ tickets in the assigned-to-control group. The assigned-to-treatment groups and assigned-to-control groups contain a mixture of Always-Treats, Compliers, and Never-Treats; the mixture should be the same in both groups, up to random error, because both groups are random samples of the tickets in the box.

Dunning (2012)

Experiments with non-compliance

Table 1: Deaths from Breast Cancer and Other Causes (HIP study).

	Group size	Deaths from breast cancer	Death rate per 1,000 women	Deaths from other causes	Death rate from other causes, per 1,000 women
Assigned to treatment:					
Accepted Screening	20,200	23	1.14	428	21.19
Refused Screening	10,800	16	1.48	409	37.87
Total	31,000	39	1.26	837	27.00
Assigned to control:					
Would have accepted screening	20,200	47	2.33	–	–
Would have refused screening	10,800	16	1.48	–	–
Total	31,000	63	2.03	879	28.35

The table is adapted from Freedman (2005: 4, Table 1).

Natural experiments

Natural experiments

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

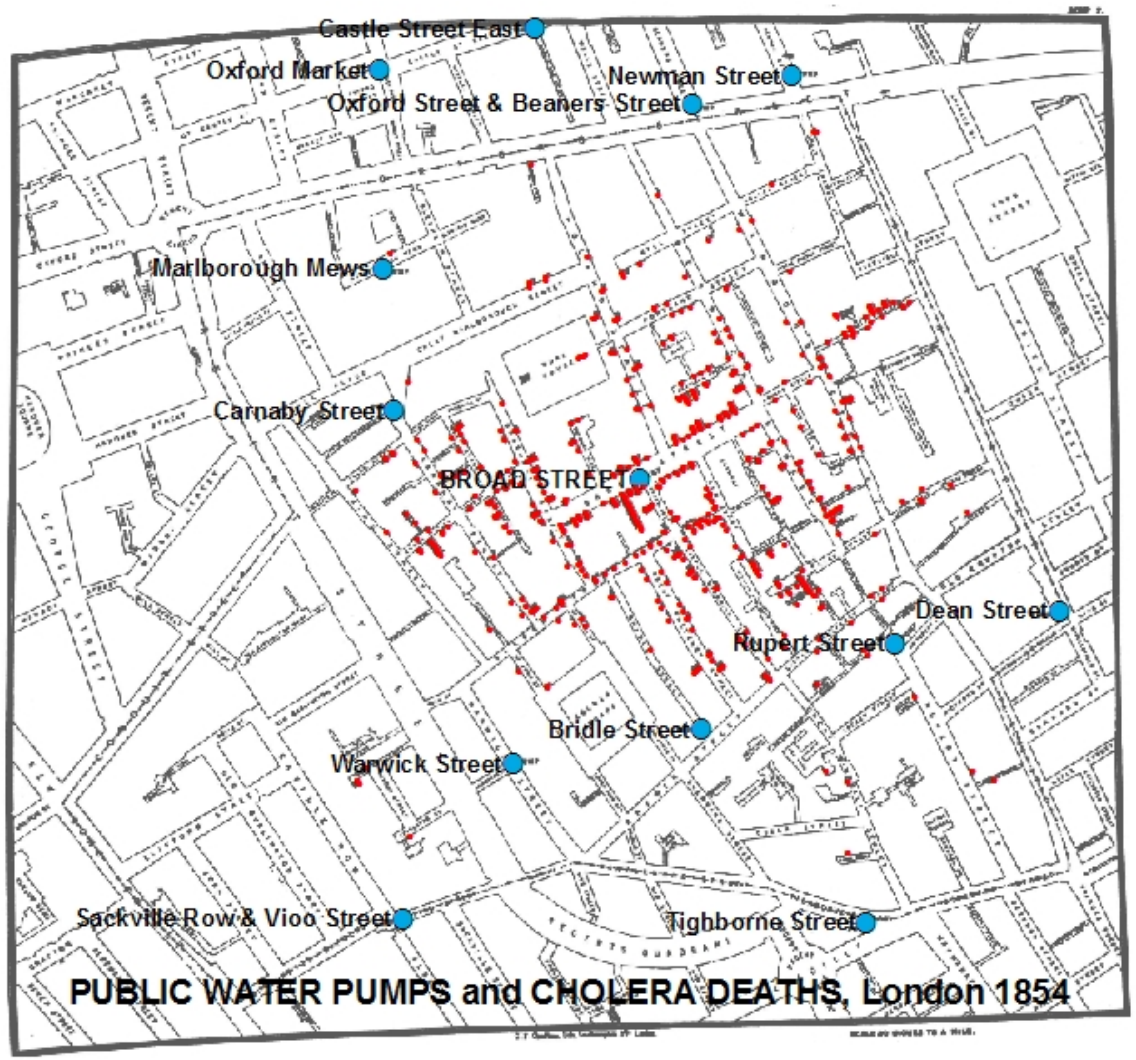
- As-if random: People are randomly exposed to water sources
- Instrumental variables: A lottery influences military service
- Discontinuities: Star ratings get arbitrarily rounded
- Difference in differences: Minimum wage changes in just one state

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



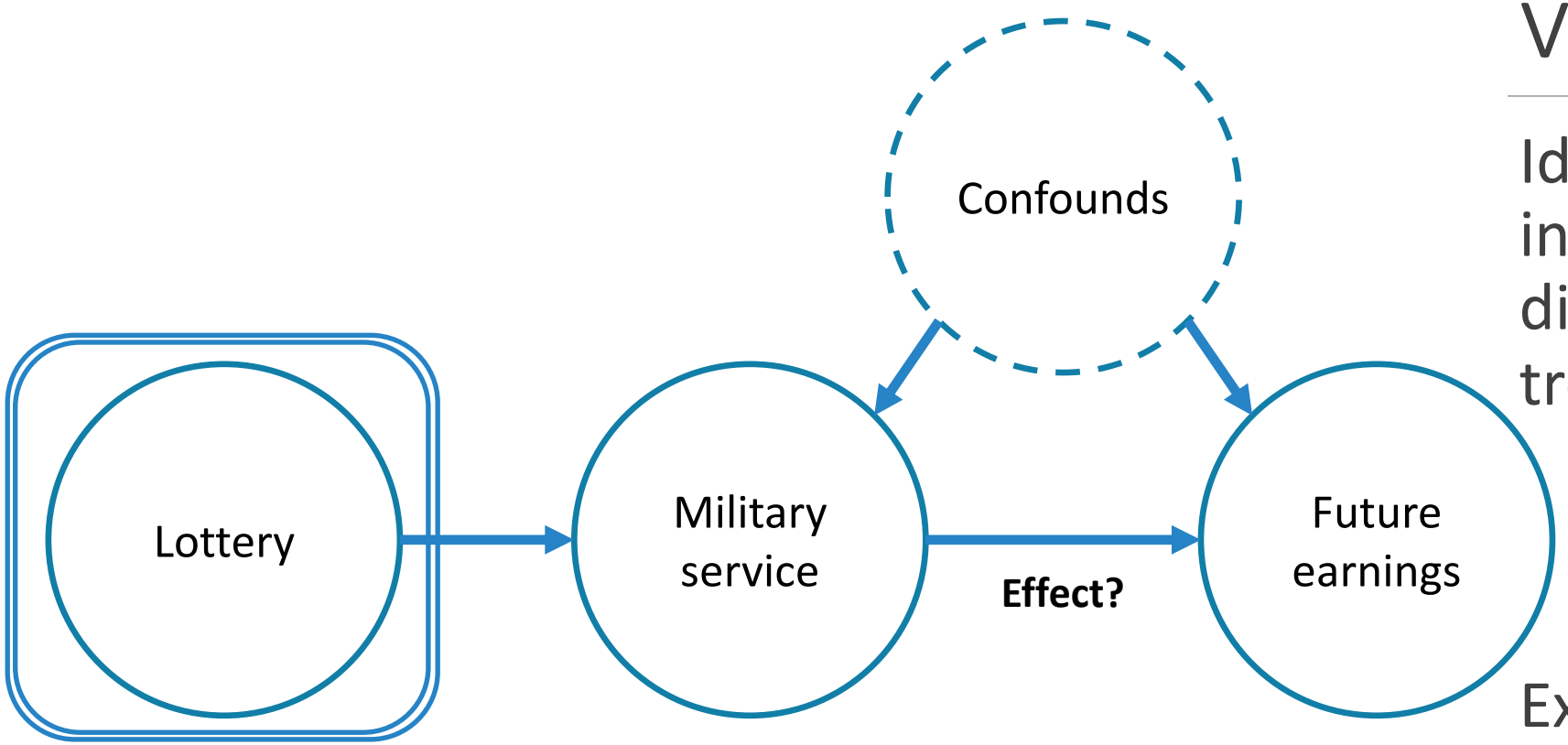
As-if random

Idea: Nature randomly assigns conditions

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

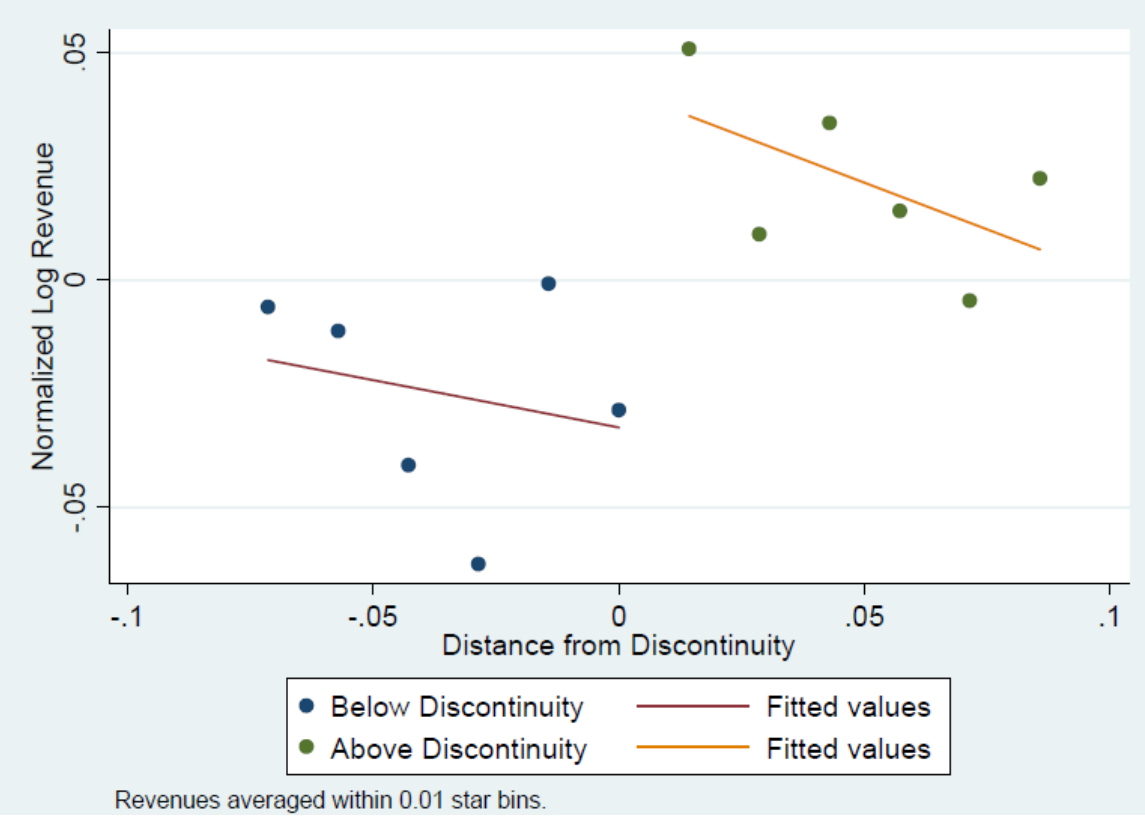
http://bit.ly/johnsnowmap

Instrumental variables



Idea: An instrument independently shifts the distribution of a treatment

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

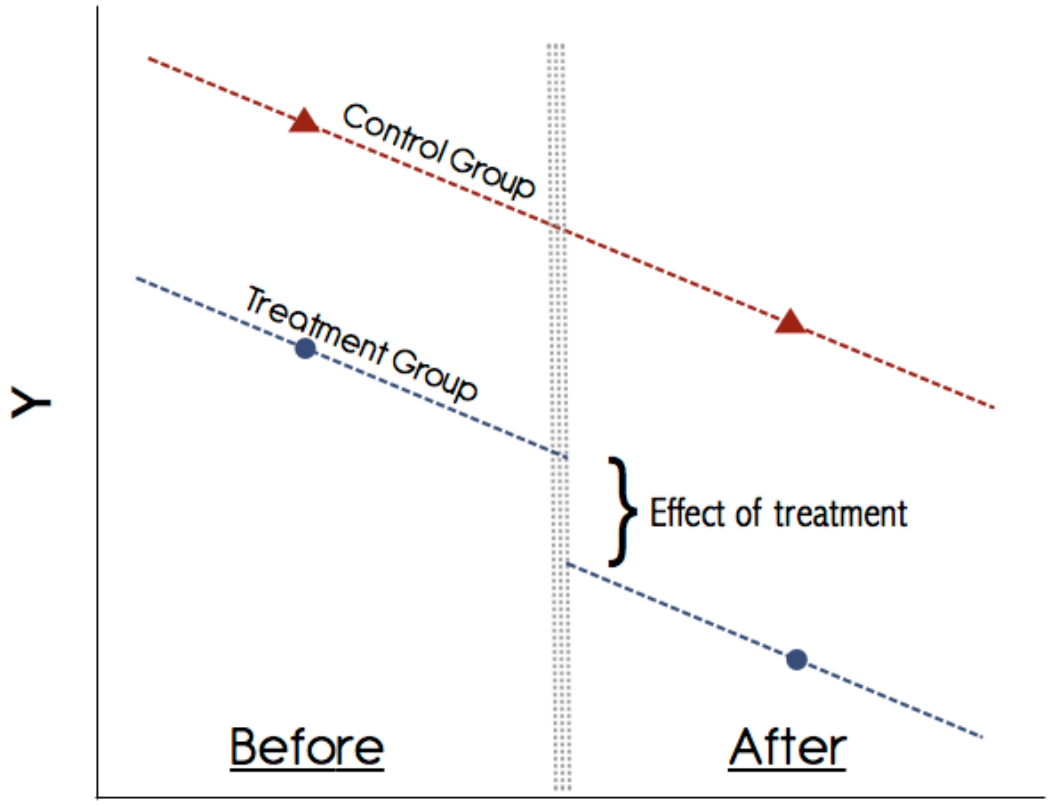


Regression discontinuities

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

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

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

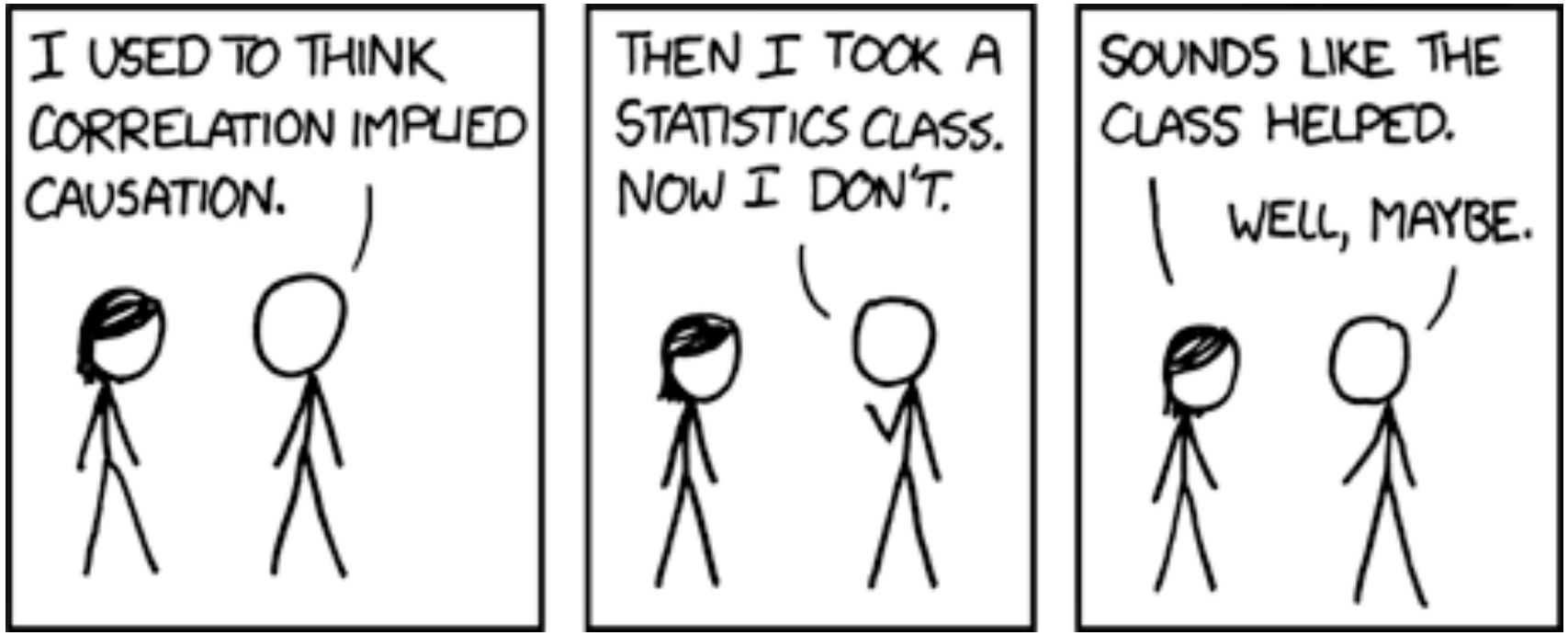
Closing thoughts

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/