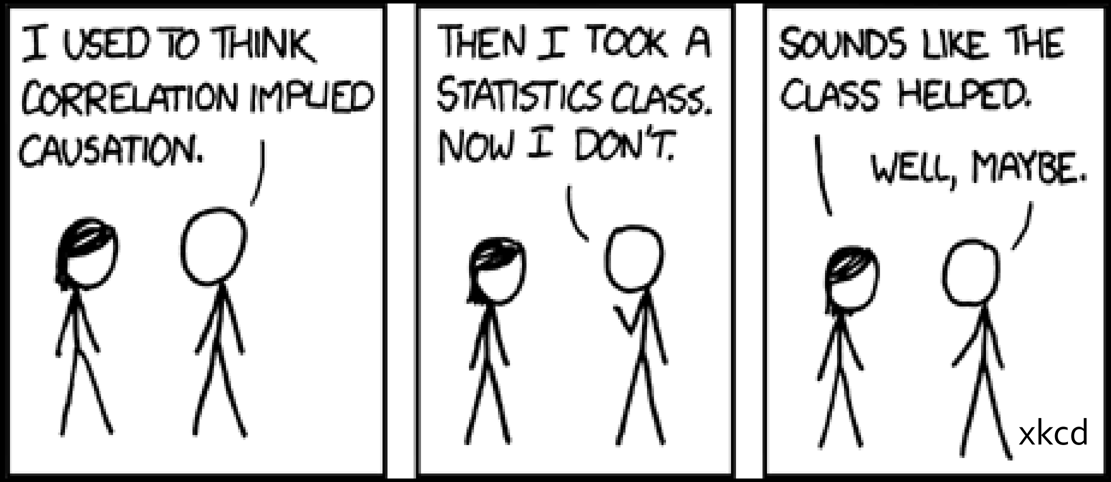

I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.
xkcd

INFO 251: Applied Machine Learning

Experimental Methods (a.k.a. Impact Evaluation 101)

Announcements

- PS1 due on Tuesday
- Enrollment
  - Please drop if you are not planning on taking the course. Encourage your friends to drop too!
  - Enrollment is still in flux. ~10 people dropped since last lecture. Waitlist is down to ~30.

Course Outline

- Causal Inference and Research Design
  - **Experimental methods**
  - Non-experiment methods
- Machine Learning
  - Design of Machine Learning Experiments
  - Linear Models and Gradient Descent
  - Non-linear models
  - Neural models
  - Unsupervised Learning
  - Practicalities, Fairness, Bias
- Special topics

Key Concepts

- Random selection and assignment
- Internal and external validity
- Counterfactuals
- Identifying assumptions
- Control groups
- Statistical Power
- Multiple testing, and adjustments

- Single difference design
- Pre vs. Post research design
- Difference-in-Difference (Double Difference) design
- Differential Trends
- Encouragement designs

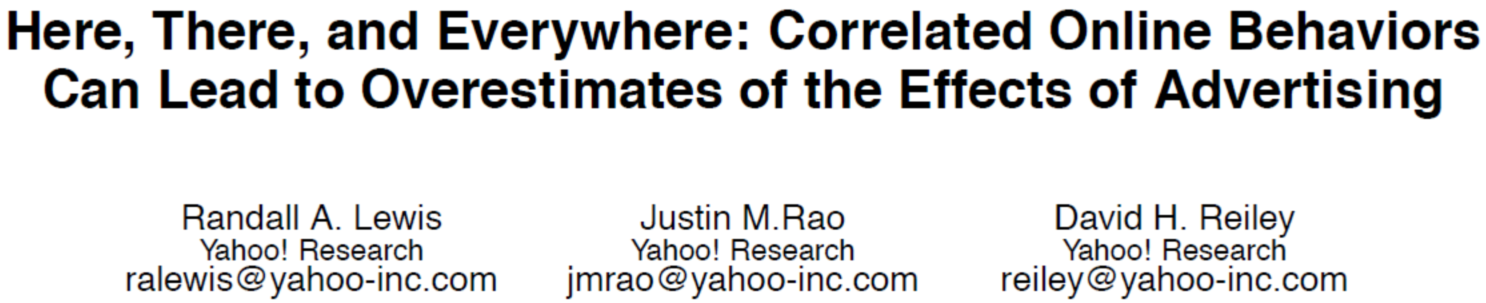
Outline

- **Why experiment?**
- How to experiment: The basics
- Measuring impact
- Important considerations

Why experiment?

- "Of course" correlation doesn't imply causation…
  - … But humans aren't very smart
- Example 1: **Bloodletting**
  - "the withdrawal of blood from a patient to prevent or cure illness and disease"
  - Prevalent from 1st century BC to 1800's
  - It made patients feel calmer!
  - Famous anecdotes: George Washington (1799); France 1833;
  - Until 1836: Dr. Pierre Louis randomized the timing at which 77 pneumonia patients were bled. 44% of early group died, vs. 25% of the late group

Why experiment?



Here, There, and Everywhere: Correlated Online Behaviors Can Lead to Overestimates of the Effects of Advertising

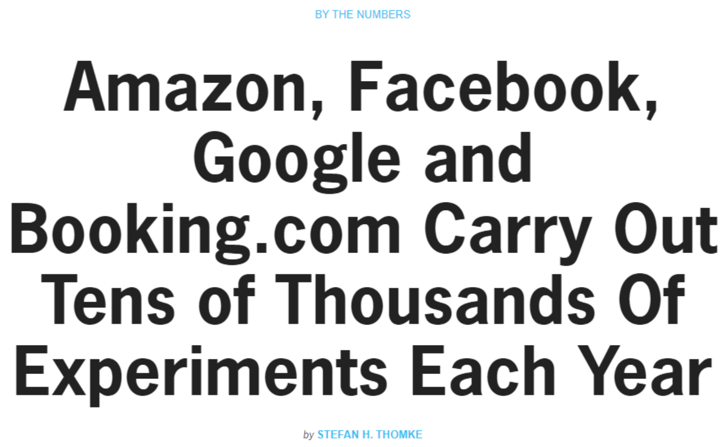
Randall A. Lewis
Yahoo! Research
ralewis@yahoo-inc.com

Justin M.Rao
Yahoo! Research
jmrao@yahoo-inc.com

David H. Reiley
Yahoo! Research
reiley@yahoo-inc.com

- Example 2 (from 2011 – not that long ago!): *Does a display ad increase the number of searches for keywords related to the brand shown?*
  - Large study with 50 million users at Yahoo
  - Basic observational study suggests massive impact = 1198%
  - Regression with some control variables: 894%
  - Regression with more control variables: 871%
    - Confidence intervals on above estimates are +/-10%
  - Randomized controlled experiment… 5.4%

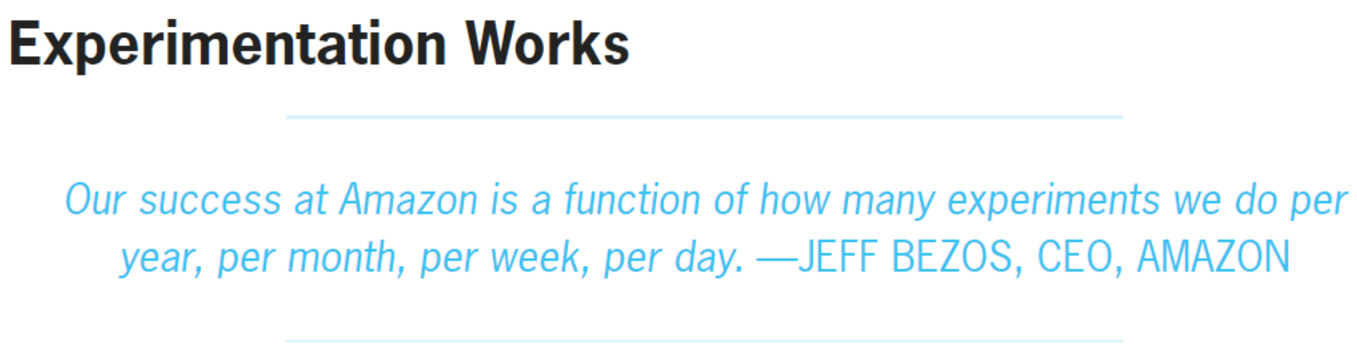
Experiments: A/B testing



BY THE NUMBERS

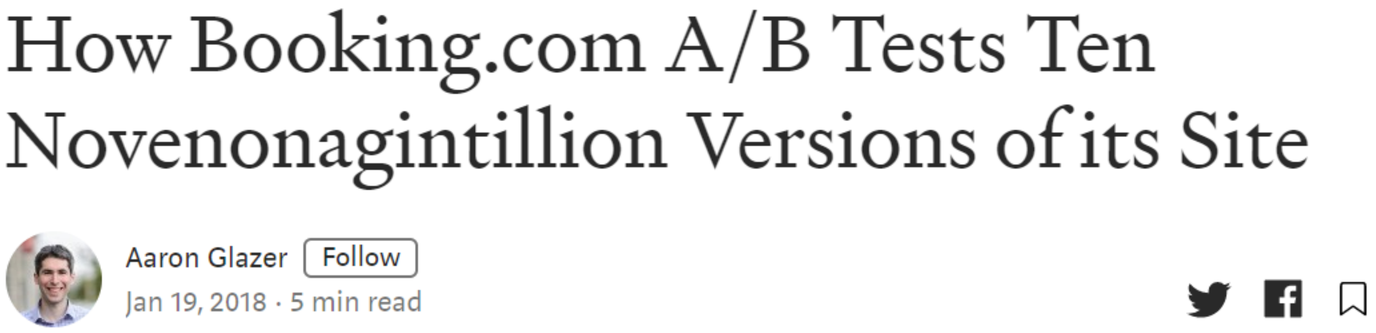
Amazon, Facebook, Google and Booking.com Carry Out Tens of Thousands Of Experiments Each Year

by STEFAN H. THOMKE



Experimentation Works

Our success at Amazon is a function of how many experiments we do per year, per month, per week, per day. —JEFF BEZOS, CEO, AMAZON



How Booking.com A/B Tests Ten Novenonagintillion Versions of its Site

Aaron Glazer

Jan 19, 2018 · 5 min read

- Zuckerberg: “At any given point in time, there’s not just one version of Facebook running the world, there’s probably tens of thousands of versions running” (2016)

Outline

- Why experiment?
- **How to experiment: The basics**
- Measuring impact
- Important considerations

What is an Experiment?

- A way to determine the **causal impact** of a **treatment (T)** on an **outcome (Y)**
  - Outcome: something you can measure, and something you care about
    - For example…
  - Treatment: an intervention
    - For example…
- (No, this has nothing to do with machine learning)

What is “impact”?

- Impact is the difference between:
  - A: what happened (with treatment) and
  - B: what would have happened (without treatment)
  - A – B = Impact
- Sounds easy, right?

The elusive counterfactual

- The *counterfactual* represents the state of the world that treated participants would have experienced in the absence of treatment
- *Problem*: Counterfactual cannot be observed
  - Why not?
  - "Fundamental problem of causal inference"

    - Paul Holland (1986)
- *Solution*: We need to "mimic" or construct the counterfactual. What is the easiest way?

Randomization

- Randomize the treatment status
- Ensures that attributes (observable and unobservable) of treated and untreated individuals are the same, on average

- *Under randomized treatment, a simple difference between outcomes in treated and control units gives an unbiased estimate of impact*

The identifying assumption

- Informally, an identifying assumption is an assumption which, if true, makes it possible to interpret a model's estimates causally
- With randomization, we typically assume treatment assignment is statistically independent of all observed and unobserved characteristics
  - If randomization is done correctly, this assumption should hold
  - (We also assume no interference between units – more on this soon)

Key advantage of randomization

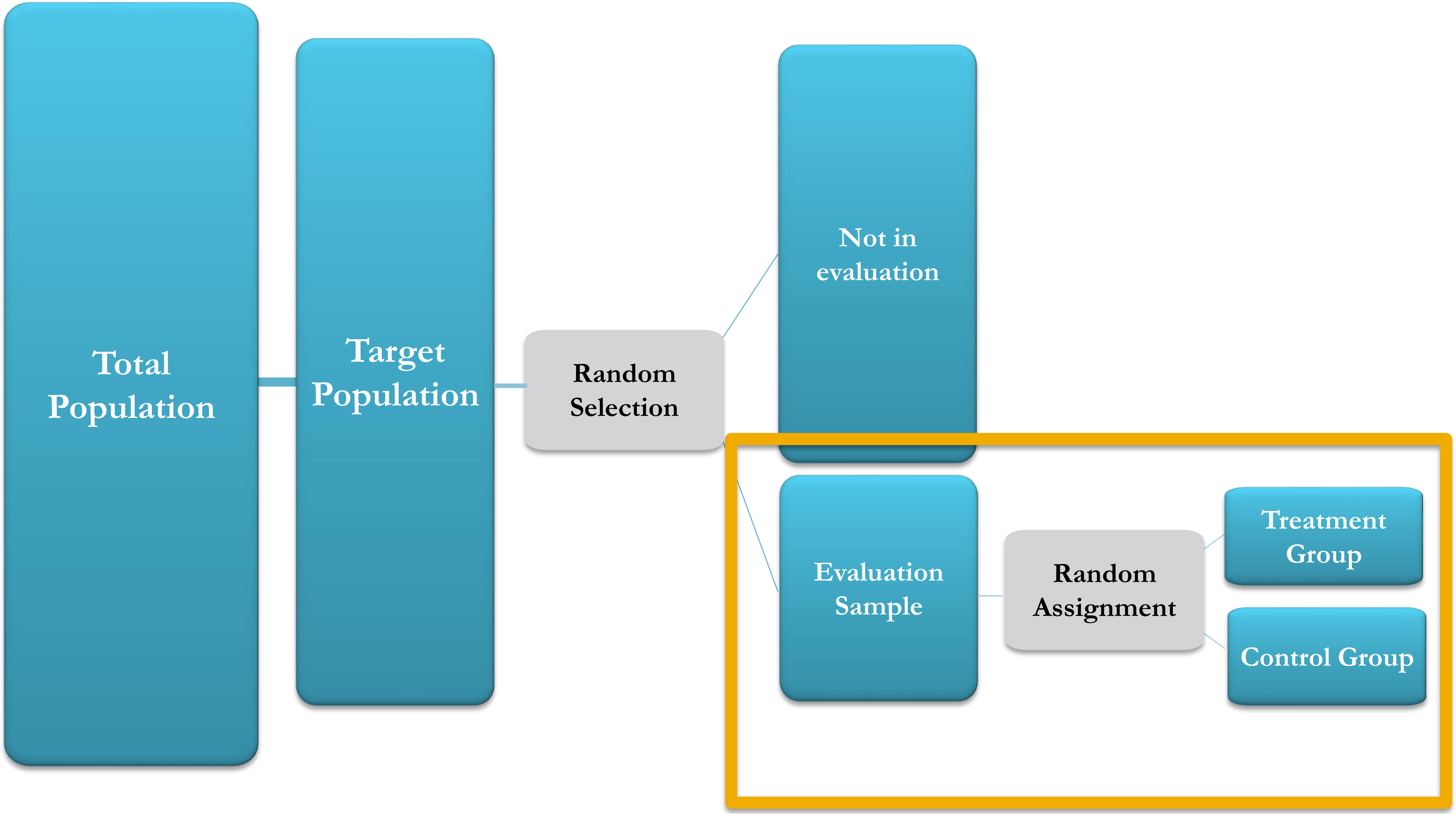
- With randomization:
  - Members of the groups should not differ systematically at the outset of the experiment…
  - …Any difference that subsequently arises between them can be attributed to the treatment rather than to other factors
- Absent randomization:
  - Members of the groups (treatment and control) may differ systematically at the outset of the experiment
  - Differences in groups may be erroneously attributed to treatment

What do we mean by “Random”

- In practice
  - Use software with a random number generator
  - Set a “seed”
  - Assign each individual/unit a random number
  - Sort by random number
  - Assign the first *m* subjects to treatment

- Bad ideas:
  - Subjects whose last name begins with A-K
  - Every other subject

Random Selection vs. Random Assignment



Internal and External Validity

- Internal validity
  - "The ability of a study to estimate causal effects within the study population"

    - Athey & Imbens (2016)

  - "The observed covariance between a treatment and an outcome reflects a causal relationship… in which the variables were manipulated [by the experimenter]"

    - Shadish, Cook, and Campbell (2002)

  - Random assignment helps with internal validity

Internal and External Validity

- External Validity
  - "External validity is concerned with generalizing causal inferences, drawn for a particular population and setting, to others…"

    - Athey & Imbens (2016)
  - "External validity concerns inferences about the extent to which a causal relationship holds over variation in persons, settings, treatments, and outcomes."

    - Shadish, Cook, and Campbell (2002)
  - Random sampling helps ensure that evaluation sample is representative of the population

Quick discussion

- What’s the difference between a control group and a counterfactual?
  - Give an example where the control group might be an inappropriate counterfactual (note: example doesn’t need to be a randomized experiment)

Outline

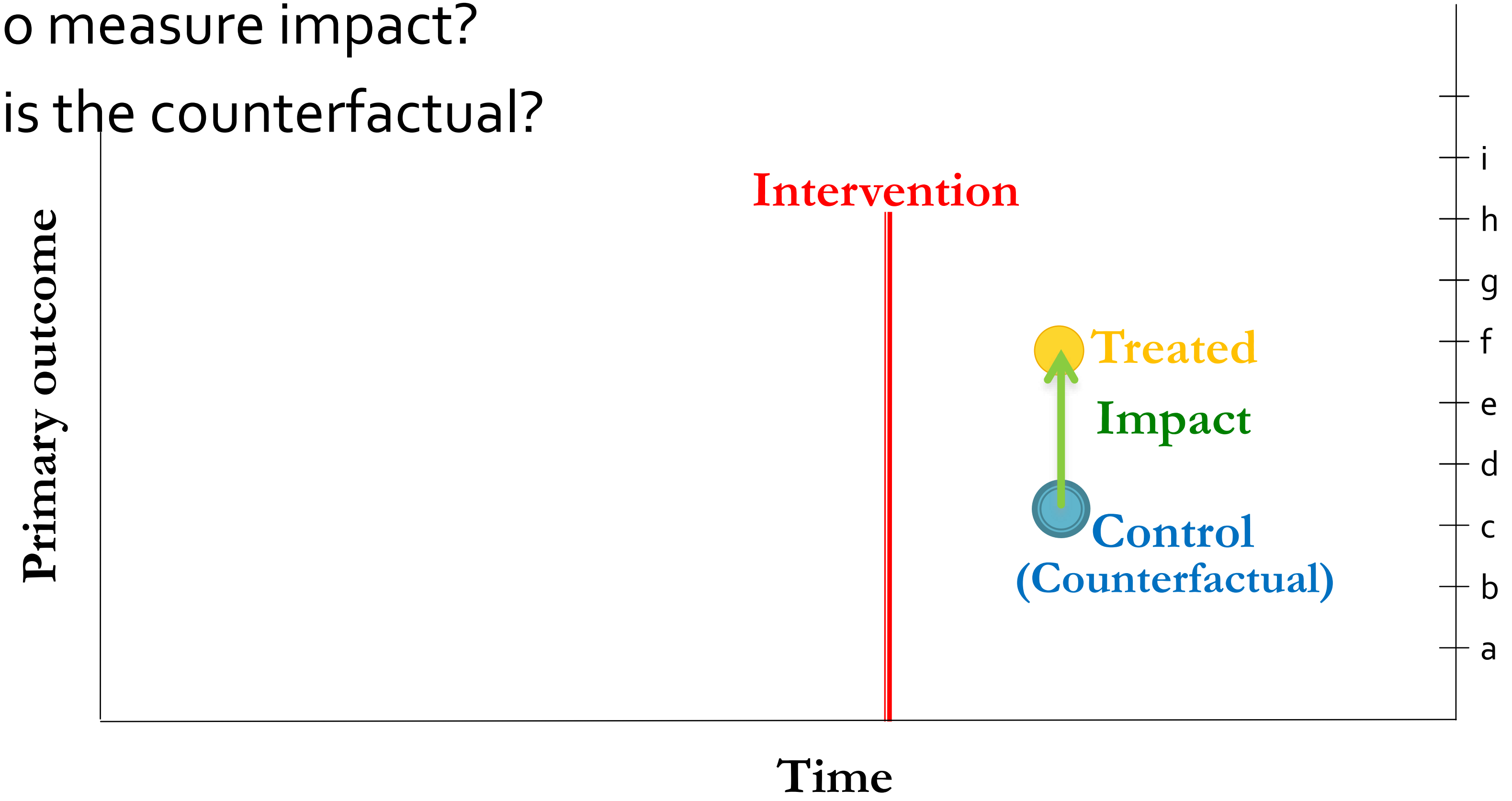
- Why experiment?
- How to experiment: The basics
- **Measuring impact**
- Important considerations

Methods for measuring impact

- Single Difference
  - Treatment vs. Control
  - Pre vs. Post
- Double Difference (Difference-in-Difference)
- Regression analysis
- Additional methods (coming soon!)
  - Instrumental Variables, Regression Discontinuity, Matching, Propensity scores, Randomization inference

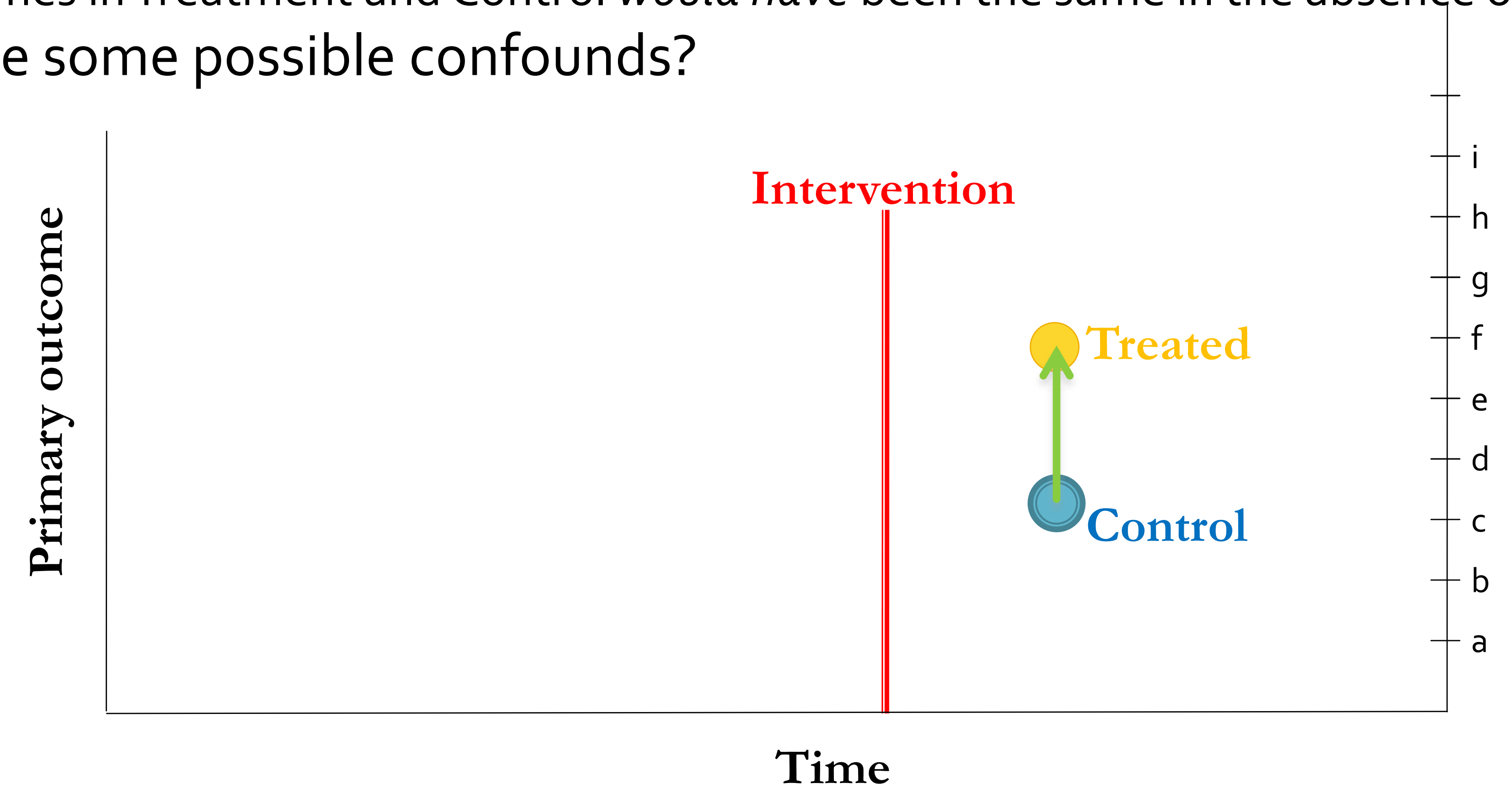
Single Difference: T vs. C

- Treatment vs. Control
  - Simple randomized experiment
  - How to measure impact?
  - What is the counterfactual?



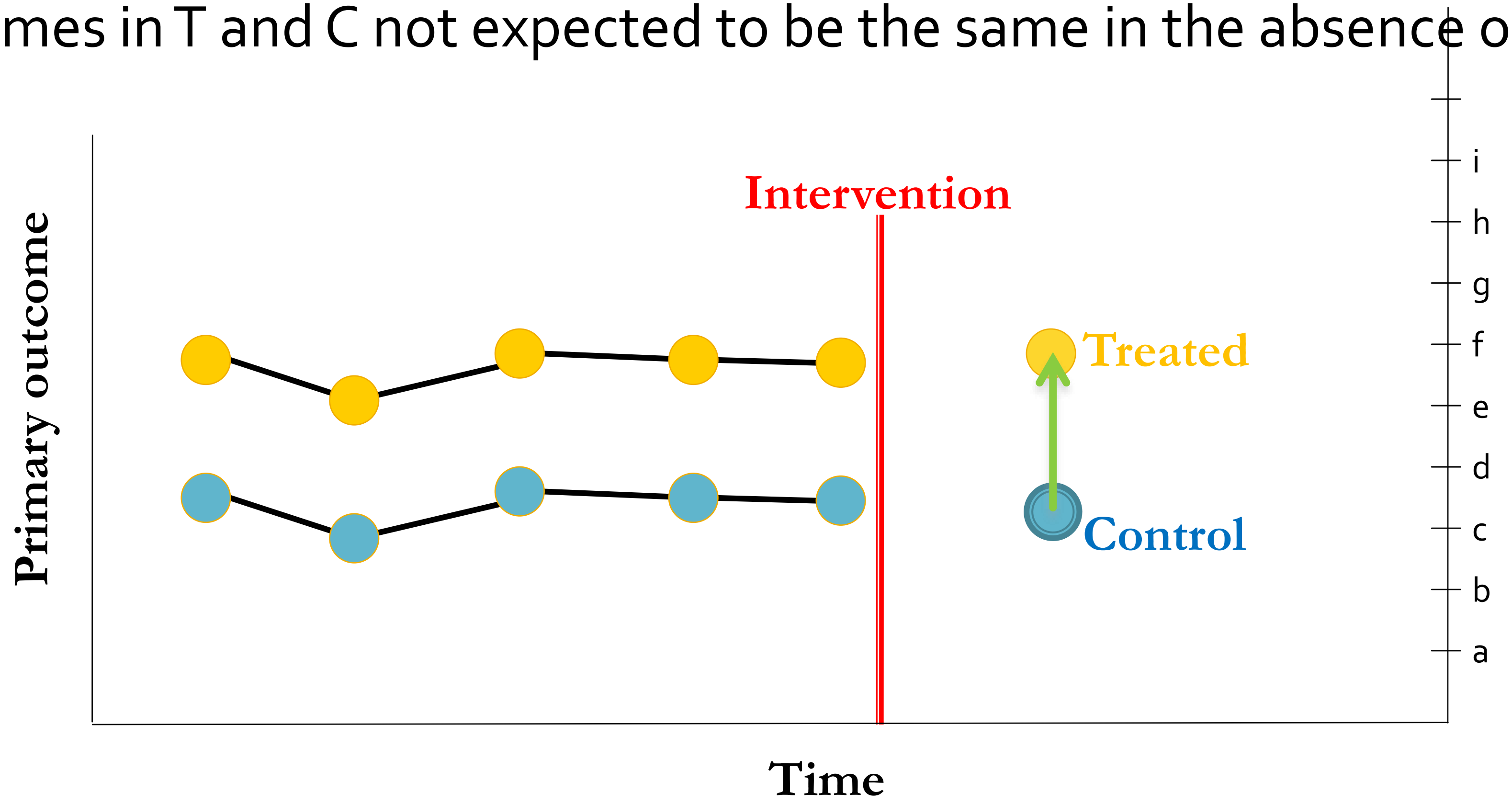
Single Difference: T vs. C

- What is the "identifying" assumption (i.e., what is the key assumption required to draw a causal inference about the impact of the intervention)?:
  - Outcomes in Treatment and Control *would have* been the same in the absence of treatment
- What are some possible confounds?



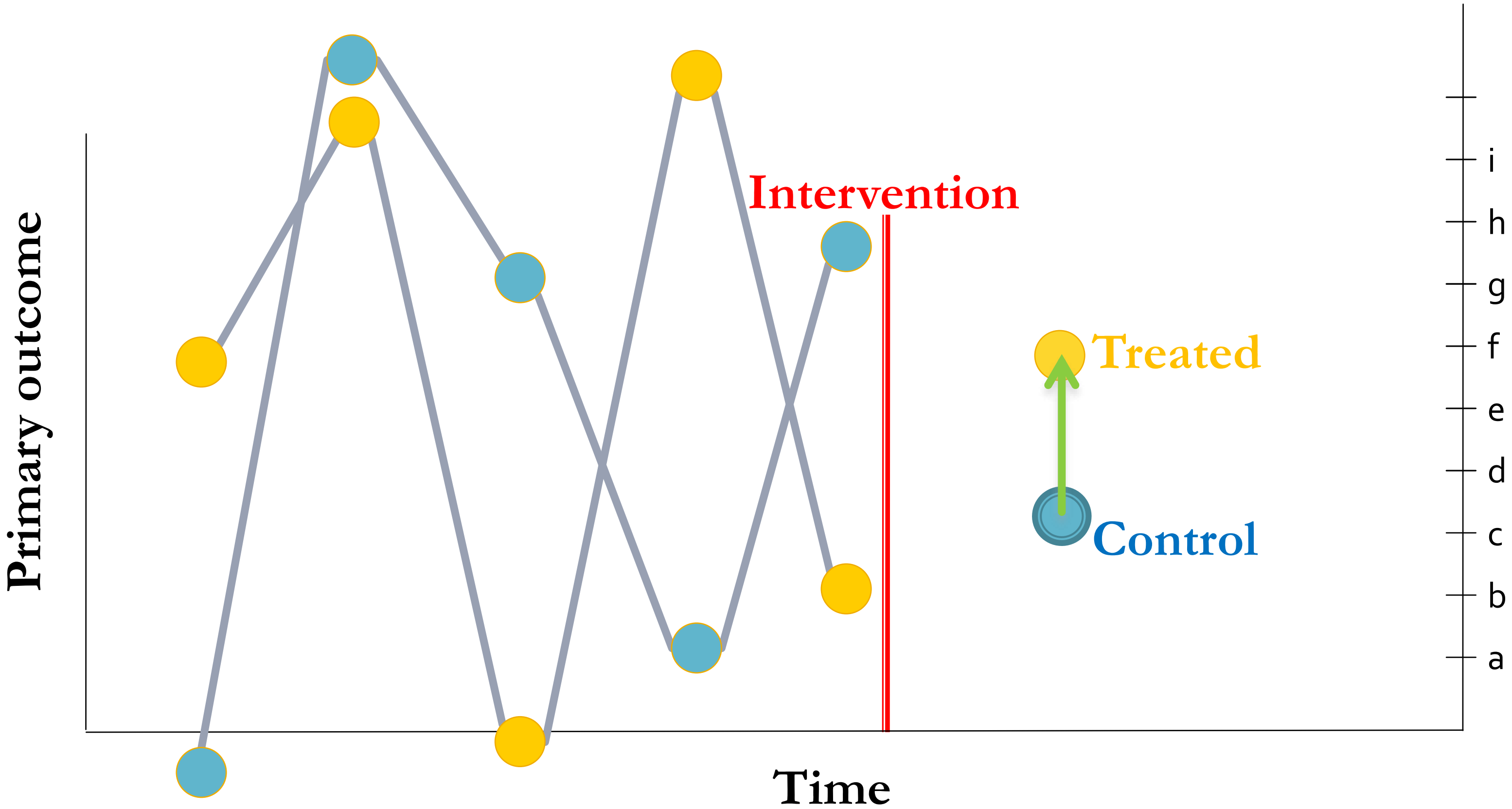
Single Difference: T vs. C

- Randomization messed up or impossible
  - e.g., selection into treatment, non-compliance
  - Outcomes in T and C not expected to be the same in the absence of treatment



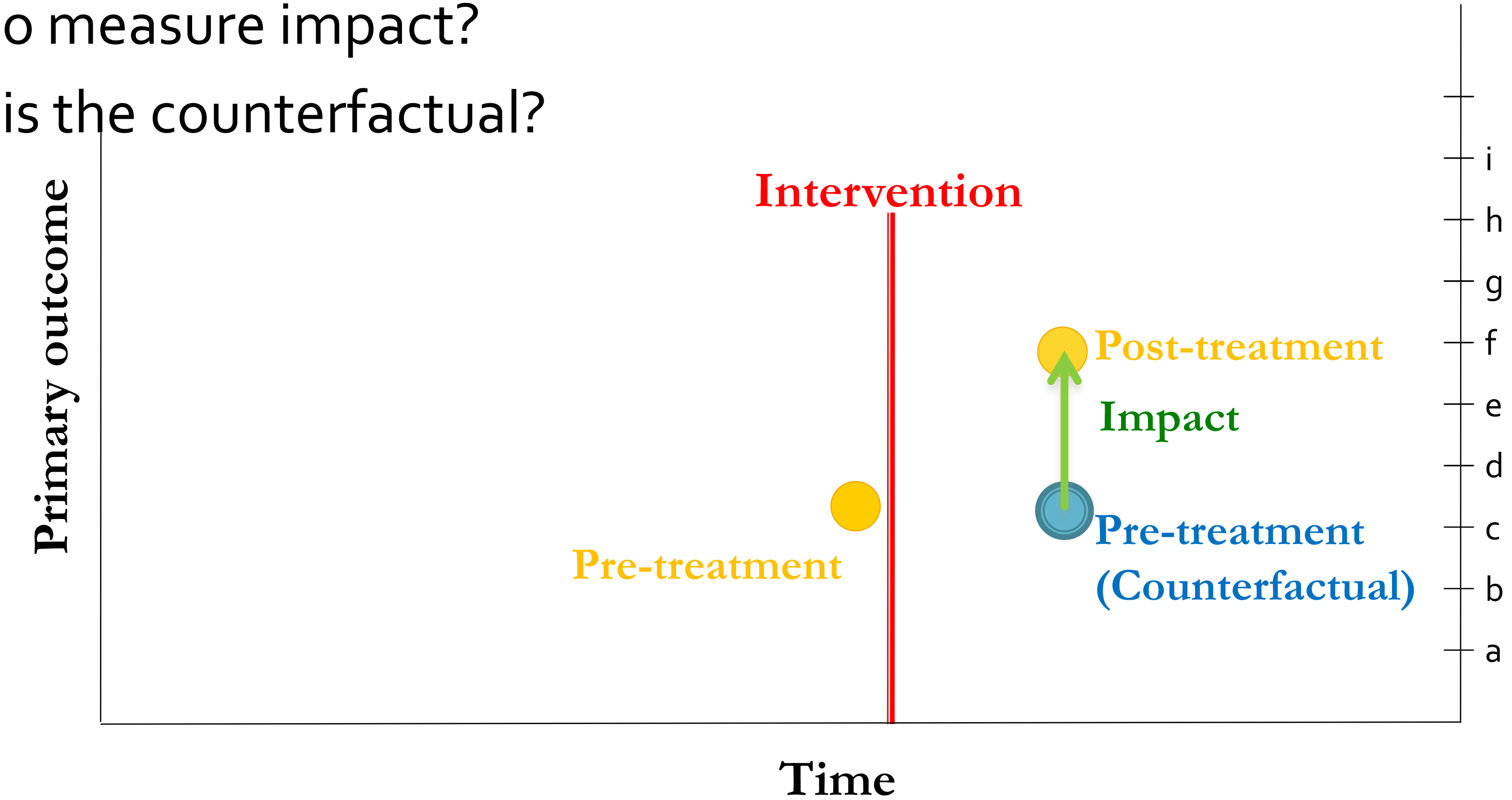
Single Difference: T vs. C

- (Too much noise)



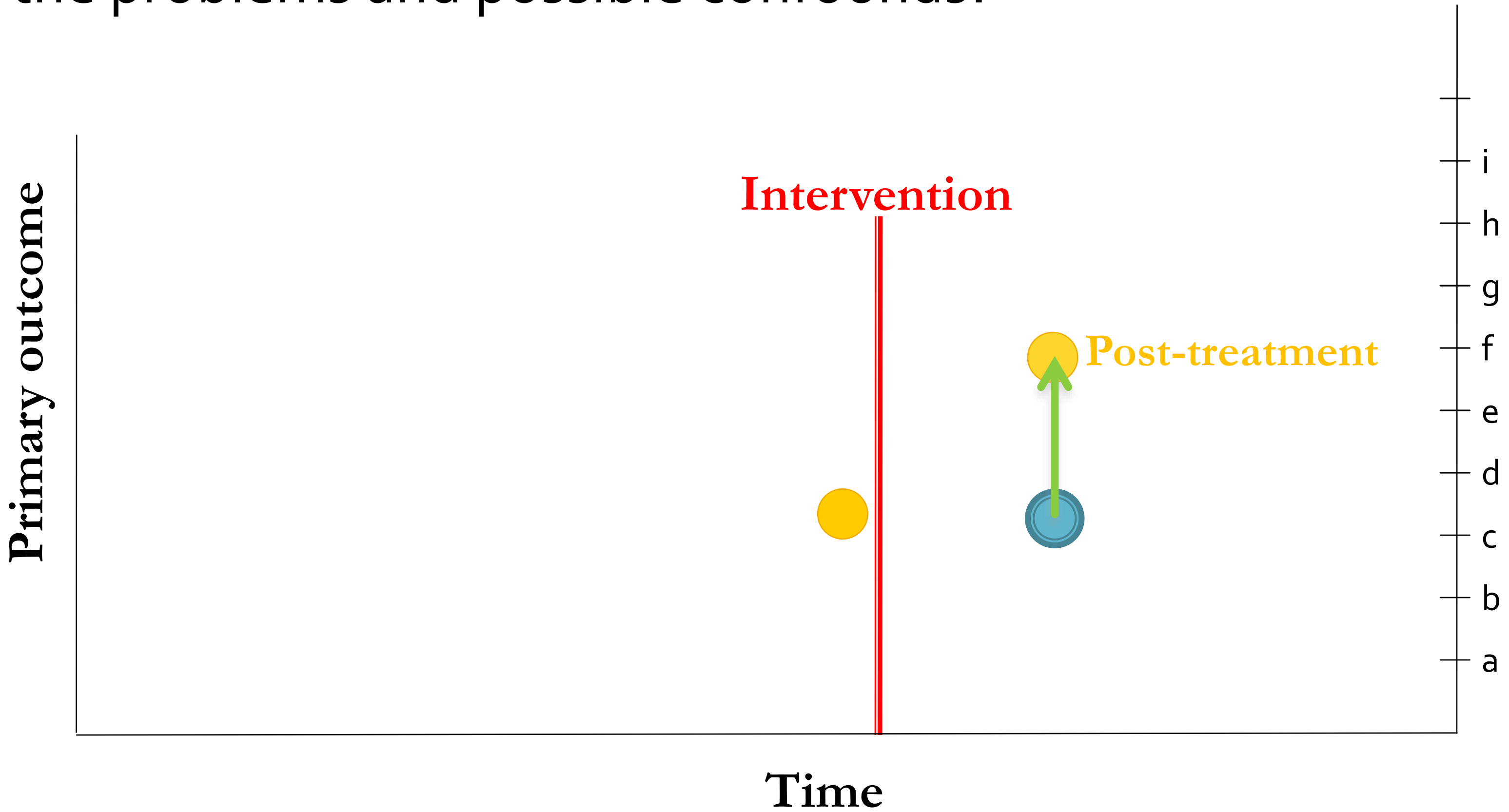
Single Difference: Pre vs. Post

- “Pre vs. Post”
  - Non-experimental, observe changes over time
  - How to measure impact?
  - What is the counterfactual?



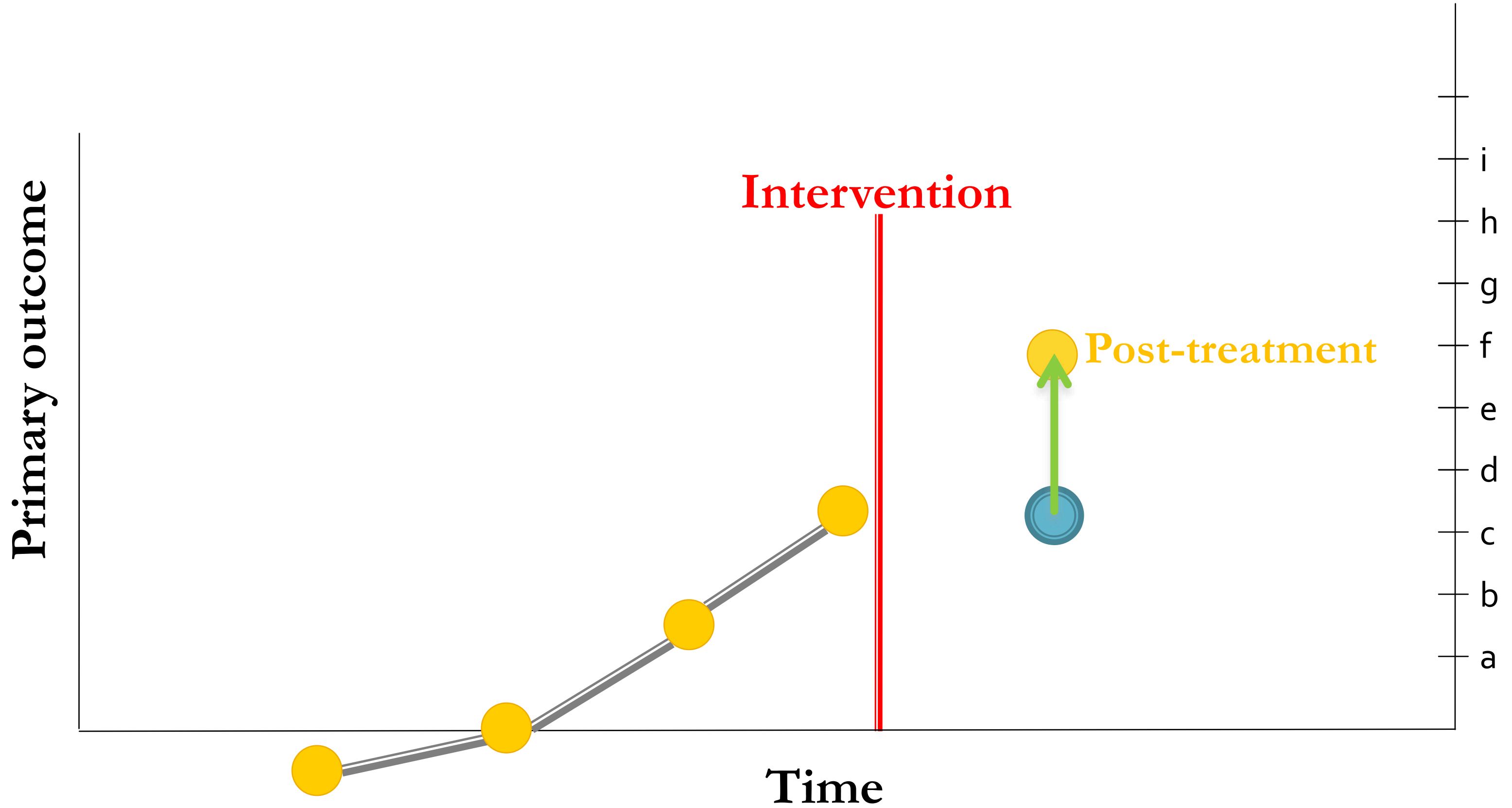
Single Difference: Pre vs. Post

- What is the key identifying assumption?
  - Outcomes pre- and post-treatment would have been the same in the absence of treatment
- What are the problems and possible confounds?



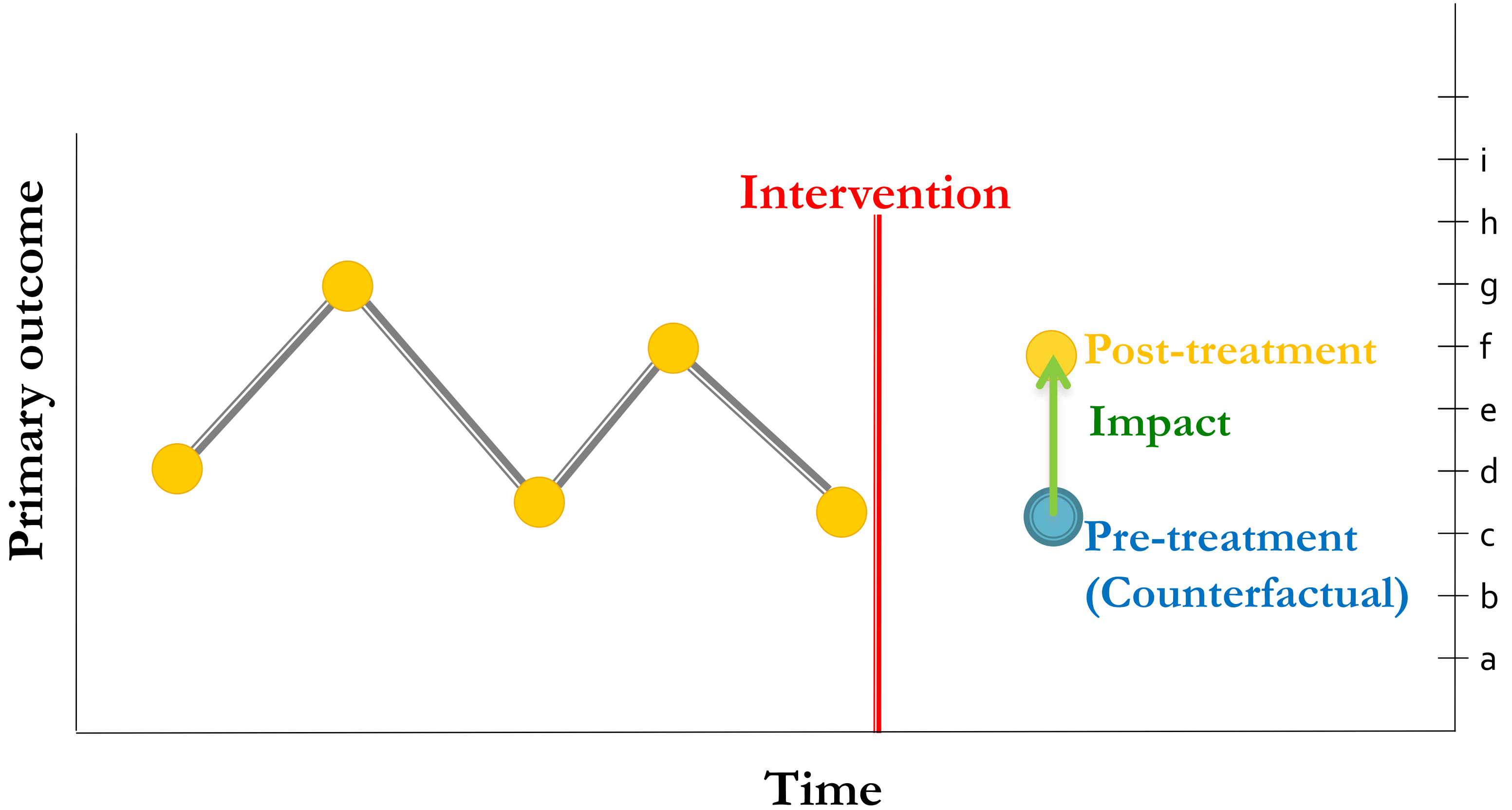
Single Difference: Pre vs. Post

- Temporal trends
  - Outcomes pre- and post-treatment not expected to be the same in the absence of treatment



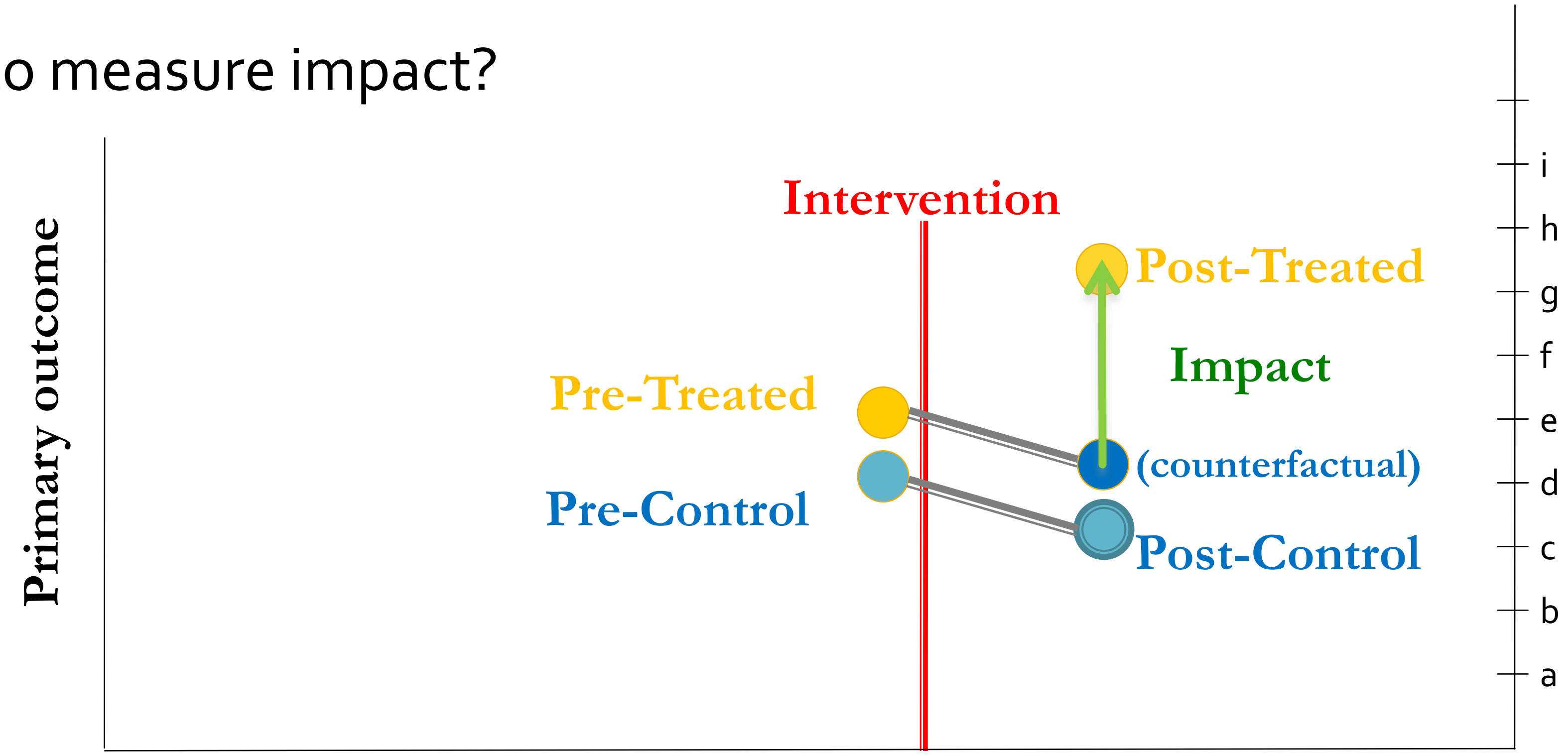
Single Difference: Pre vs. Post

- Seasonality
  - Outcomes pre- and post-treatment not expected to be the same in the absence of treatment



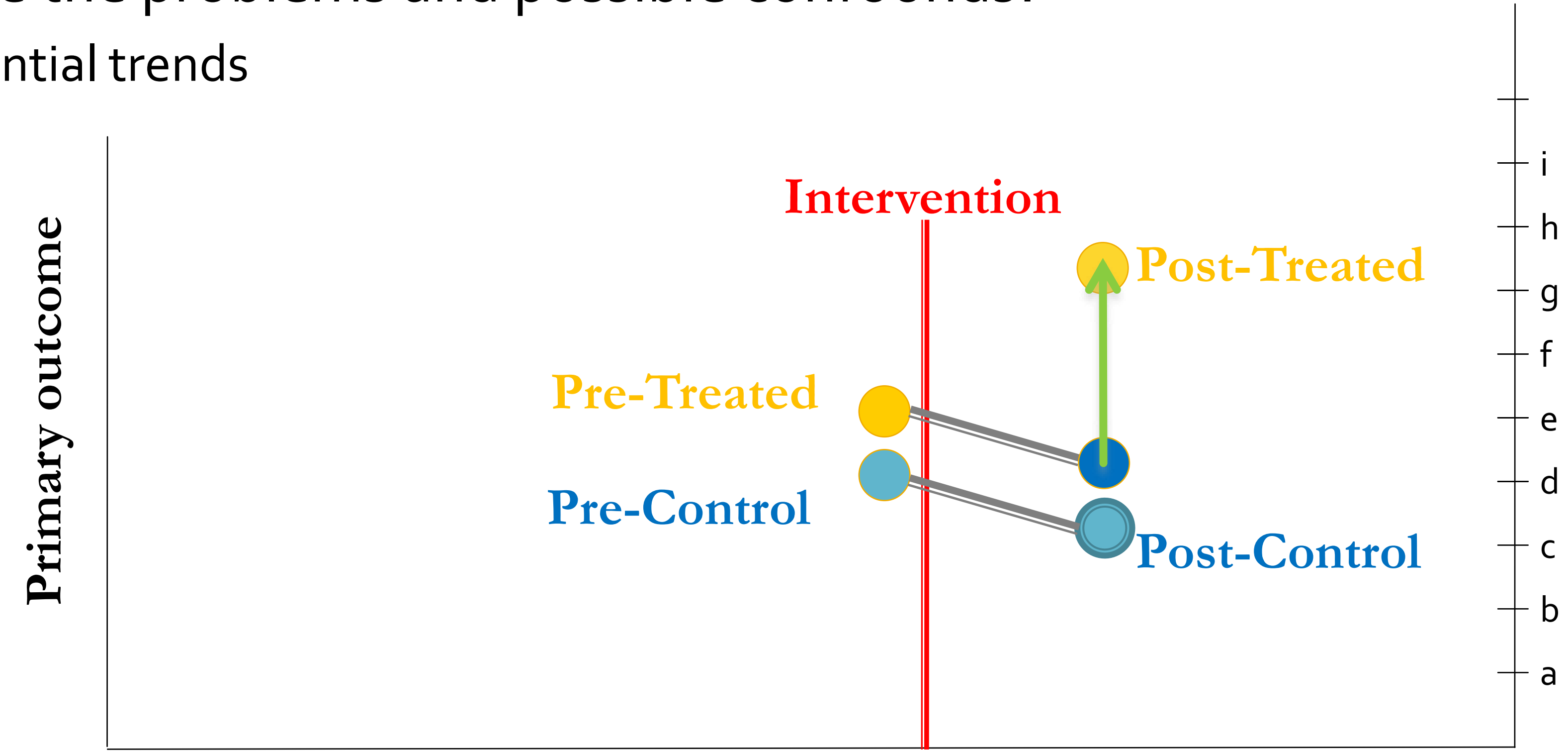
Double Difference

- Double Difference:
  - Requires pre-post *and* T vs. C
  - Here, the counterfactual is more subtle
  - How to measure impact?



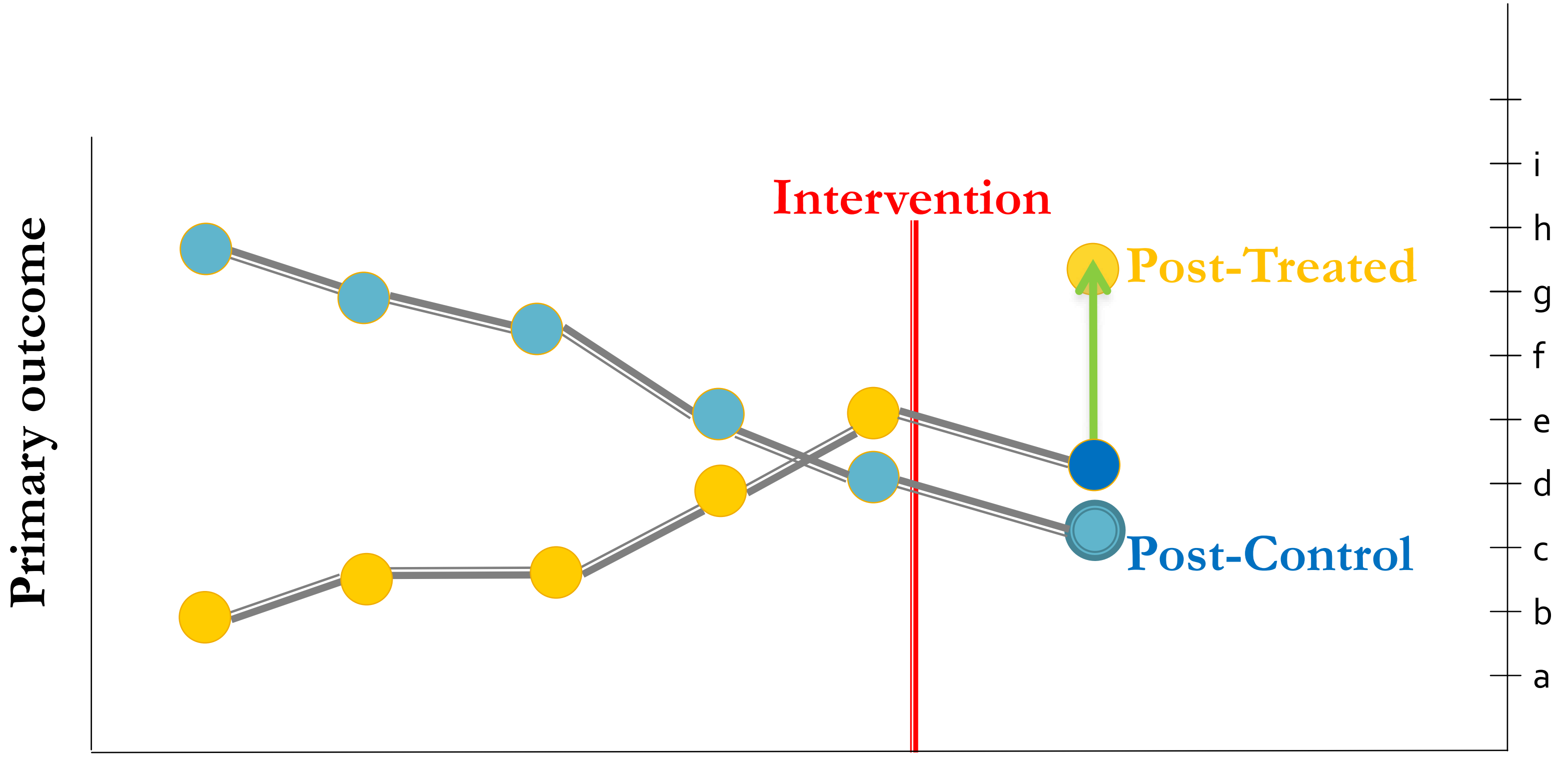
Double Difference

- What is the key identifying assumption?
  - *Trends* (changes over time) in treatment group and control group would have been the same in the absence of treatment
- What are the problems and possible confounds?
  - Differential trends



Double Difference

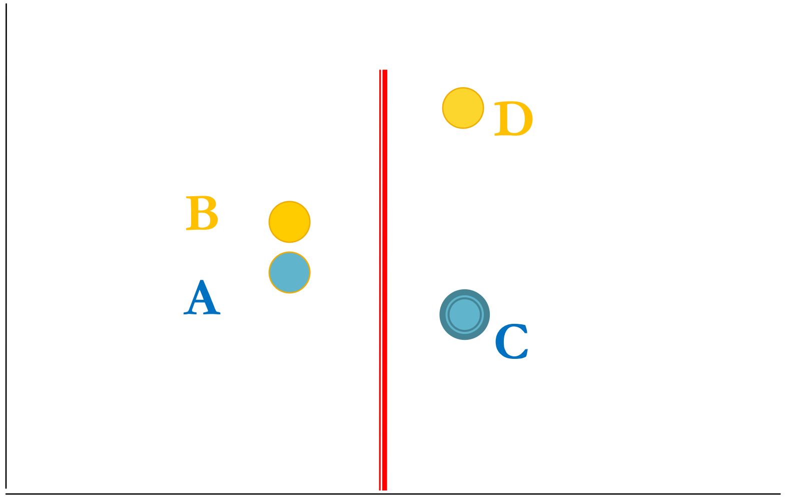
- Differential trends
  - *Trends* in T and C groups would have *not* been the same in the absence of treatment



Double Difference

- How to measure impact? (of treatment on the treated)
- Pre vs. Post
  - D – B
- Treat vs. Control
  - D – C
- Double Difference
  - (D-B) – (C-A)
  - [equivalent to (D-C) – (B-A)]

	Control Group	Treatment Group
Before Treatment	A	B
After Treatment	C	D



Double Difference

- What is the impact?
  - Before intervention, Treatment group outcome = 80
  - After intervention, Treatment group outcome = 74
  - Before intervention, Control group outcome = 82
  - After intervention, Control group outcome = 92

DD Example: Jensen (2007 QJE)

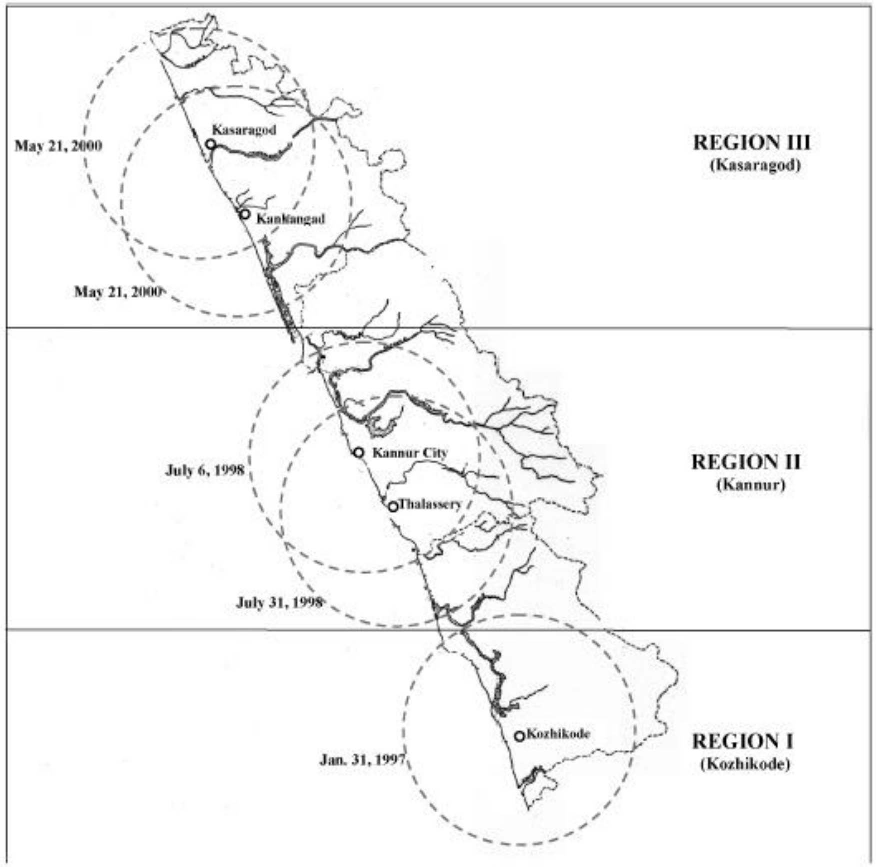


FIGURE II
Spread of Mobile Phone Coverage in Kasaragod, Kannur, and Kozhikode Districts

DD Example: Jensen (2007 QJE)

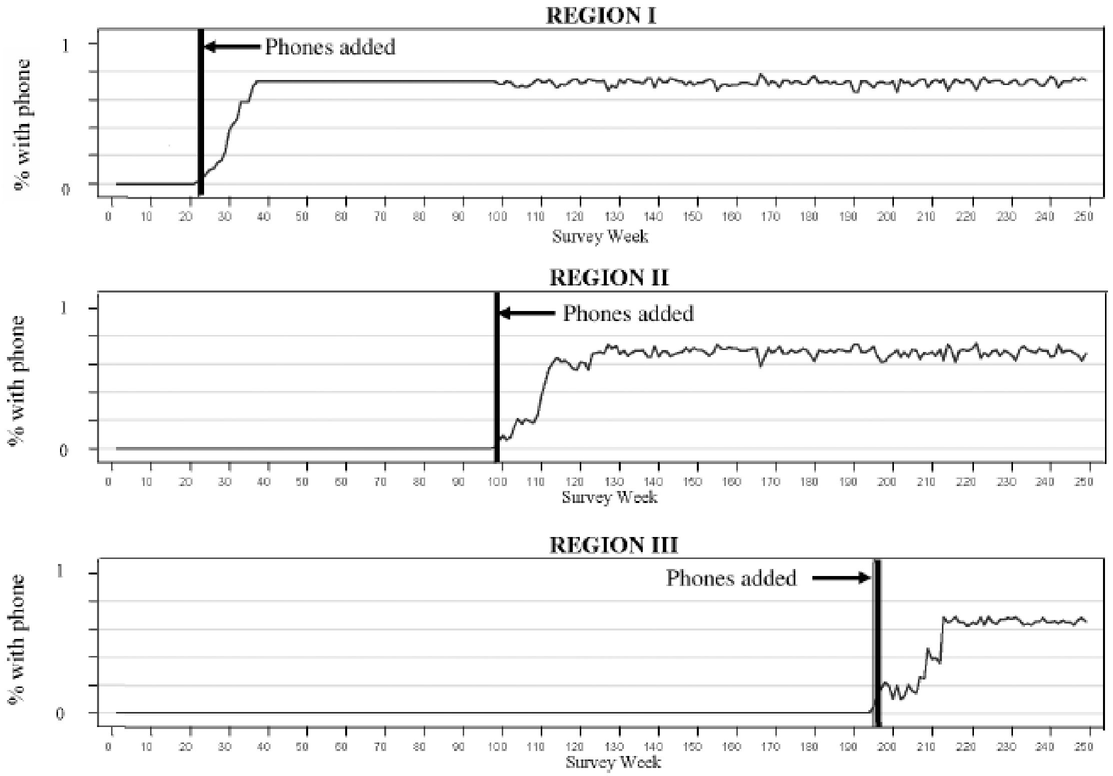


FIGURE III
Mobile Phone Adoption by Fishermen
Data from the Kerala Fisherman Survey conducted by the author.

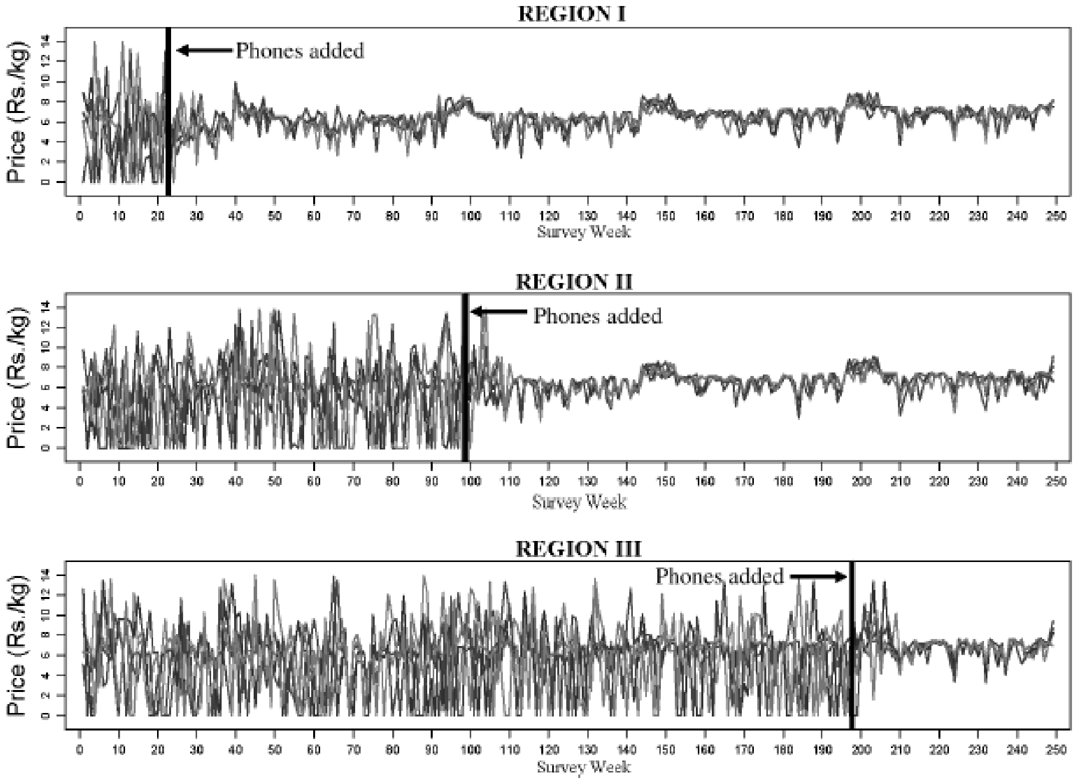


FIGURE IV
Prices and Mobile Phone Service in Kerala
rala Fisherman Survey conducted by the author. The price series represent the average 7:30–8
s. All prices in 2001 Rs.

Outline

- Why experiment?
- How to experiment: The basics
- Measuring impact
- **Important considerations**
  - Power calculations
  - Encouragement designs
  - Pitfalls

How large should the sample be?

- Intuitively
  - More people = more power
  - More people, more "ways to slice data"
  - But larger samples come at cost

Power Calculations

- Power calculations allow you to determine the **sample size** necessary to detect a **minimum detectable effect** given assumptions about:
  - Standard deviation of outcome in each population
  - Average value of outcome in each population
  - Statistical threshold for significance (e.g. 0.10)
  - Power/Sensitivity – more on this later
- See Duflo et al. (2006), section 4

Encouragement Design

- What if you can’t randomize the treatment you really care about?
  - E.g. I want to know effect of attendance on final grades
  - Can’t randomize attendance
  - Why not compare product final grades for people with high vs. low attendance?

- Encouragement design
  - When you can’t randomize adherence to the treatment you care about
  - Randomly **encourage** a treatment group to be treated

Encouragement Design

- We're in a world with "two treatments"
  - The encouragement
  - The actual treatment of interest
- Easy to measure the impact of encouragement on outcomes
- Also possible to measure the effect of the actual treatment on outcomes!
  - Instrumental Variables

Randomization: Pitfalls

- Common threats to internal and external validity
  - Spillovers, externalities, and interference
    - An indirect or unintended effect which often does not comply with treatment assignment
    - Can be positive and negative
    - Examples?
  - Non-compliance
  - Attrition (esp. differential attrition)

Multiple testing concerns

- Joseph Rhine was a parapsychologist in the 1950's
  - Founder of Journal of Parapsychology, affiliate of AAAS
- He ran an experiment where subjects had to guess whether 10 hidden cards were red or blue
- He found that about 1 person in 1000 had ESP, i.e. they could guess the color of all 10 cards
  - He called back the "psychic" subjects and had them do the same test again. They all failed.
  - He concluded that the act of telling psychics that they have psychic abilities causes them to lose it…

John Canny/Jeff Ullman/Anand Rajaraman

Publication and "Results" Bias

- Most scientific tests are probabilistic (due to measurement error, sampling etc), and are only true to some "significance" value (the false positive rate), e.g. 0.05 or 1/20.
- Both experimenters and journals have a strong bias to publish only results that "succeed", i.e. which were significant at 0.05
- Suppose 20 experimenters try to show "coffee causes dementia"

John Canny/Jeff Ullman/Anand Rajaraman

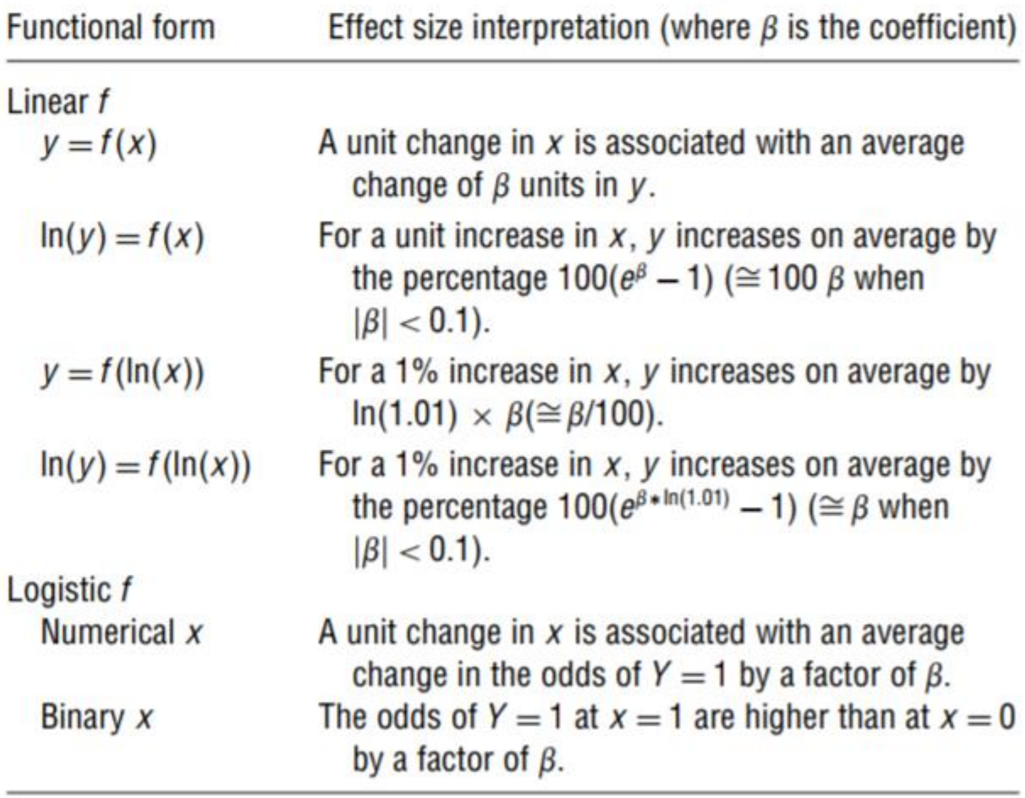
“Big” data doesn’t help

- Data mining technologies and large datasets make it incredibly easy to ask questions and test hypotheses – you can effectively run hundreds of experiments in a few hours
- You will get many “positives” by chance
- Its very important to know what the false positive rate is, and whether a result you see is really “unusual” relative to chance

John Canny/Jeff Ullman/Anand Rajaraman

What can you do?

- Report everything you tried, not just successes
- Present & interpret effect size
  - "Statistical significance is the least interesting thing about the results. You should describe the results in terms of measures of magnitude – not just, does a treatment affect people, but how much does it affect them."
    - Gene V. Glass
  - "The primary product of a research inquiry is one or more measures of effect size, not P values."
    - Jacob Cohen



Functional form	Effect size interpretation (where β is the coefficient)
Linear f	
$y = f(x)$	A unit change in x is associated with an average change of β units in y.
$\ln(y) = f(x)$	For a unit increase in x, y increases on average by the percentage $100(e^{\beta} - 1)$ ($\cong 100\,\beta$ when $\lvert\beta\rvert < 0.1$).
$y = f(\ln(x))$	For a 1% increase in x, y increases on average by $\ln(1.01) \times \beta (\cong \beta/100)$.
$\ln(y) = f(\ln(x))$	For a 1% increase in x, y increases on average by the percentage $100(e^{\beta * \ln(1.01)} - 1)$ ($\cong \beta$ when $\lvert\beta\rvert < 0.1$).
Logistic f	
Numerical x	A unit change in x is associated with an average change in the odds of $Y = 1$ by a factor of β.
Binary x	The odds of $Y = 1$ at $x = 1$ are higher than at $x = 0$ by a factor of β.

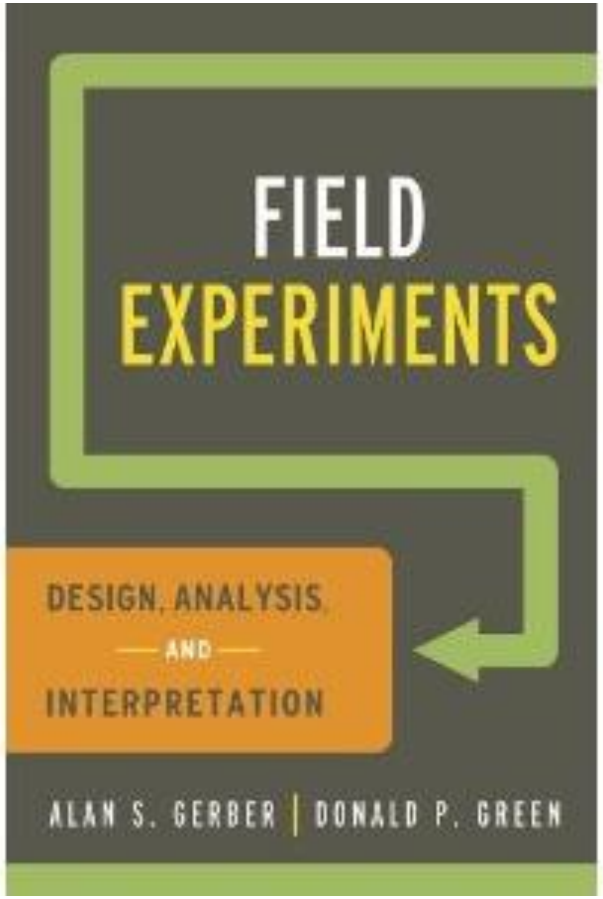
Interpreting regression coefficients

What can you do?

- Bonferroni Adjustments
  - With *k* tests, reduce the significance threshold for each test to 0.05 / *k*
- Dunn-Sidak
  - With *k* tests, reduce the significance threshold to $1-(.95)^{1/k}$
- Other options:
  - Family error rates
  - Pre-analysis plans
  - Randomization inference
  - Canny et al.: Partition the total significance into unequal-sized bins applied to different tests
    - E.g. if you have a test that you think is important but may not succeed, you could assign 0.025 out of 0.05 to it, and distribute the remaining 0.025 among other tests. All of this should happen before you run any experiments.

For Next Class:

- Read about Progresa
- Review basics of regression and econometrics
- Further reading:
  - Take a course in Field Experiments
  - INFO 290 “Experiments and Causal Inference”
  - UGBA 196.7 “Applied Impact Evaluation”


FIELD EXPERIMENTS
DESIGN, ANALYSIS, AND INTERPRETATION
ALAN S. GERBER | DONALD P. GREEN