ORIE 5355: People, Data, & Systems

Lecture 15: Synthetic control, experimentation culture, and miscellaneous topics

Nikhil Garg

Course webpage: https://orie5355.github.io/Fall_2021/

Announcements

- In-person guest lecture (Lily Xu) on Wednesday – please attend
- Fill out the “miscellaneous topics” survey
- HW4 due tomorrow
- Quiz 4 released Wednesday, due Sunday
- Project details released this week
- OHs
  - Zhi today (regular time; Zoom)
  - No Nikhil Wednesday office hours this week
  - Friday
    - Nikhil 12:30 – 1:30
    - Zhi 1:30 – 2:30

Experimentation summary so far

Several different experimental designs

- Classic, individual level A/B testing
- Graph cluster randomization
  - More generally, *spatial* randomization
- Switchbacks: randomization over time

These experimental techniques are not workable sometimes

- Product is “public-facing” – hard to roll back
- Interference really network/city wide, so spatial randomization less effective
- Sensitive change, so can’t launch in many cities at once
- It takes a long time for effect to occur

Today: “synthetic control”

(and experimentation culture, communication, and miscellaneous)

Synthetic control

Synthetic control is important

"The gold standard for drawing inferences about the effect of a policy is a randomized controlled experiment."

"However, in many cases, experiments remain difficult or impossible to implement, for financial, political, or ethical reasons, or because the population of interest is too small."

Synthetic control "is arguably the most important innovation in the policy evaluation literature in the last 15 years"

"The State of Applied Econometrics: Causality and Policy Evaluation." Susan Athey and Guido Imbens, 2017

Recently used to study..

In academia: gun laws, immigration policy, minimum wage, taxation, organized crime…

- What is the effect of a minimum wage increase?
- What is the effect of more immigration into a city?

Outside academia: “multi-lateral organizations, think tanks, business analytics units, governmental agencies, and consulting firms”

- Bill & Melinda Gates Foundation
- Tech companies

“Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspect.” Alberto Abadie, 2021

Synthetic control: Setting

Suppose running standard experiments is hard:

- Product is “public-facing” – hard to roll back
- Interference really network/city wide, so spatial randomization less effective
- Sensitive change, so can’t launch in many cities at once
- It takes a long time for effect to occur

So, you decide to launch in just 1 city

- This is the treatment city… so what is the control?
- Why do we need a control?

  want to measure global treatment effect: $Y_1 - Y_0$

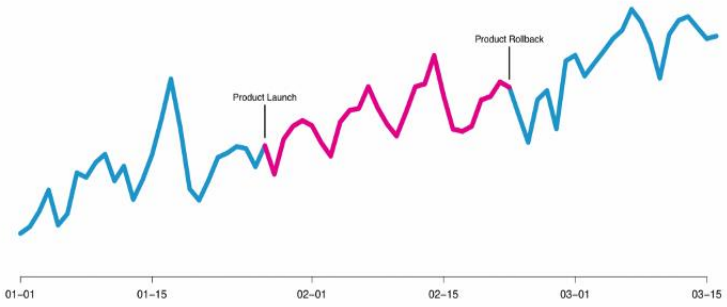
  “what would have happened to this city if I didn’t launch the treatment?”

Possible controls

Example: Suppose you're launching a product at time T in Miami, and your platform also has users in Houston, Atlanta, and Orlando

What are possible controls? (Y_0)

- The same city (Miami) from time 0 to time T
  - Problem: seasonality!
- Seasonality-adjusted Miami
  - Problem: unforeseen events (Covid)
- Pick one of: Houston, Atlanta, Orlando from time T to time 2T
  - Hope: Seasonality/Covid affects control city similar to how it affects Miami
  - Challenge: Which city do I pick?
  - Problem: No city is perfect analogue



AirBnb, (Jan Overgoor) Experiments at Airbnb | by AirbnbEng | The Airbnb Tech Blog | Medium

Main Idea

Problem with using past Miami (from time 0 to time T): Time matters; future might be different than past, for reasons nothing to do with treatment

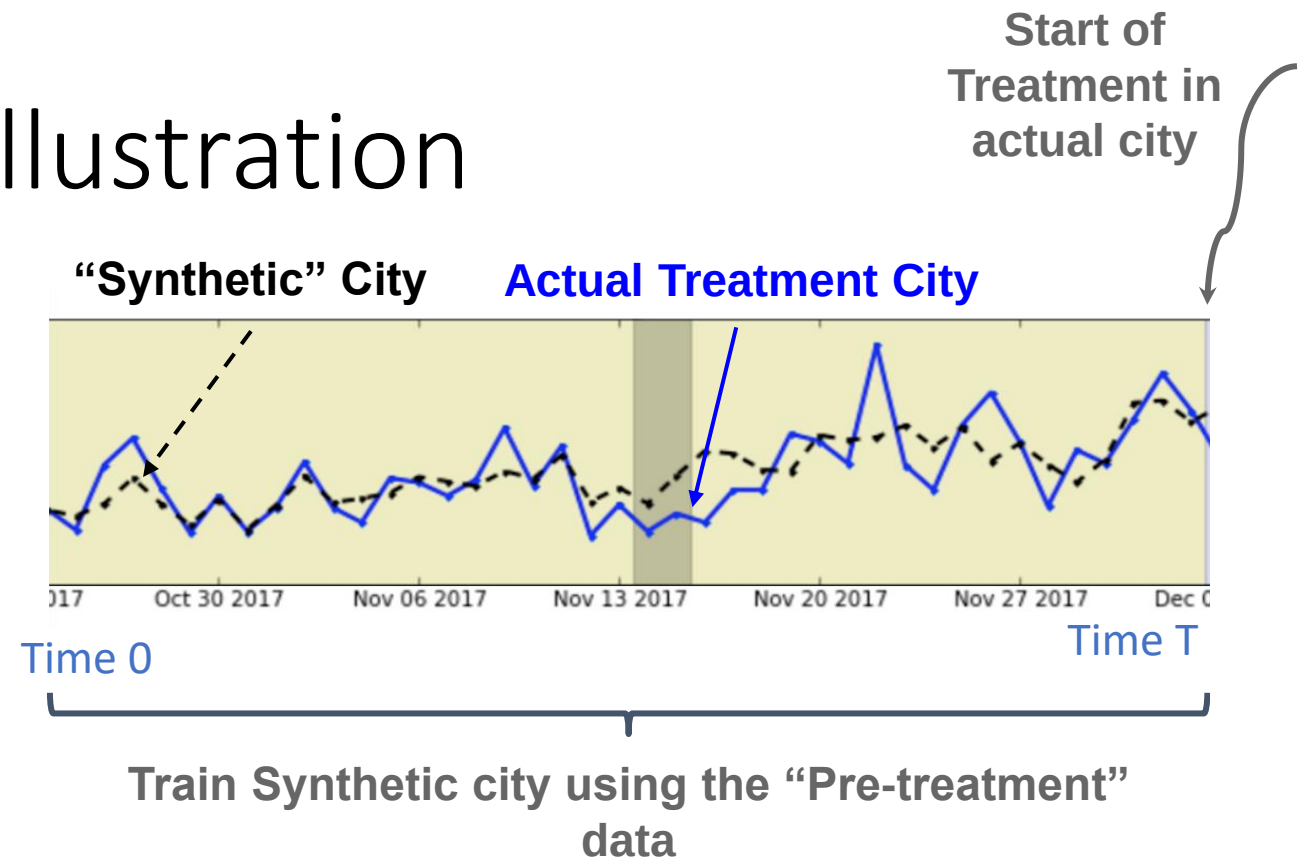
Problem with using Houston (from time T to time 2T): Geography also matters; Houston is not Miami

(differs from Miami even in observed past, from time 0 to T)

Synthetic control main idea: Use a weighted average of Houston/Orlando/Atlanta, from time T to time 2T

- Design weights such that "Synthetic Miami" matches real Miami in past (0 to T)
- Hope: That Synthetic Miami in future (T to 2T) behaves like what Miami would have behaved like if you didn't launch the treatment

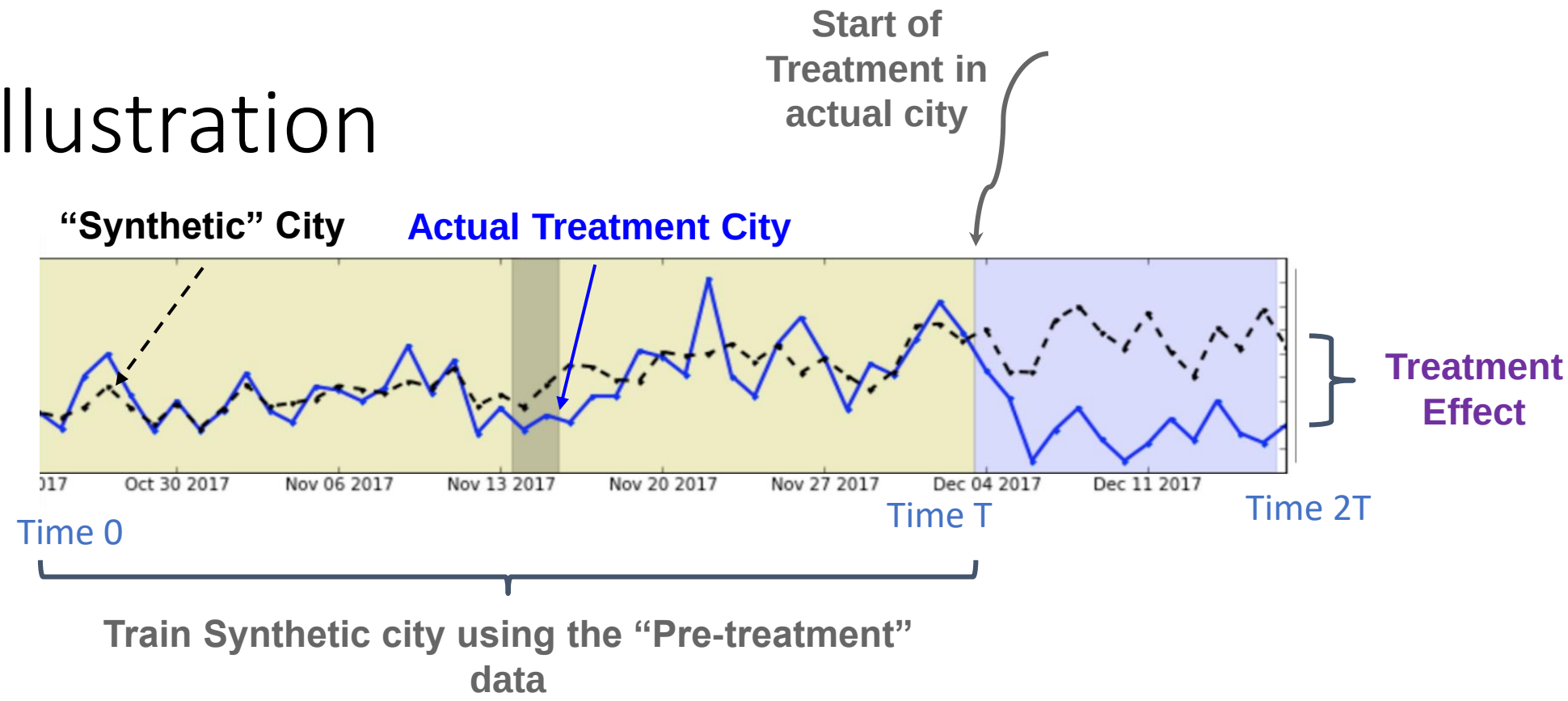
Illustration



Step 1: Train a synthetic city using past data, such that synthetic city and actual treatment city match when *neither* is in treatment

Initial Plot credit:
Uber MX team

Illustration

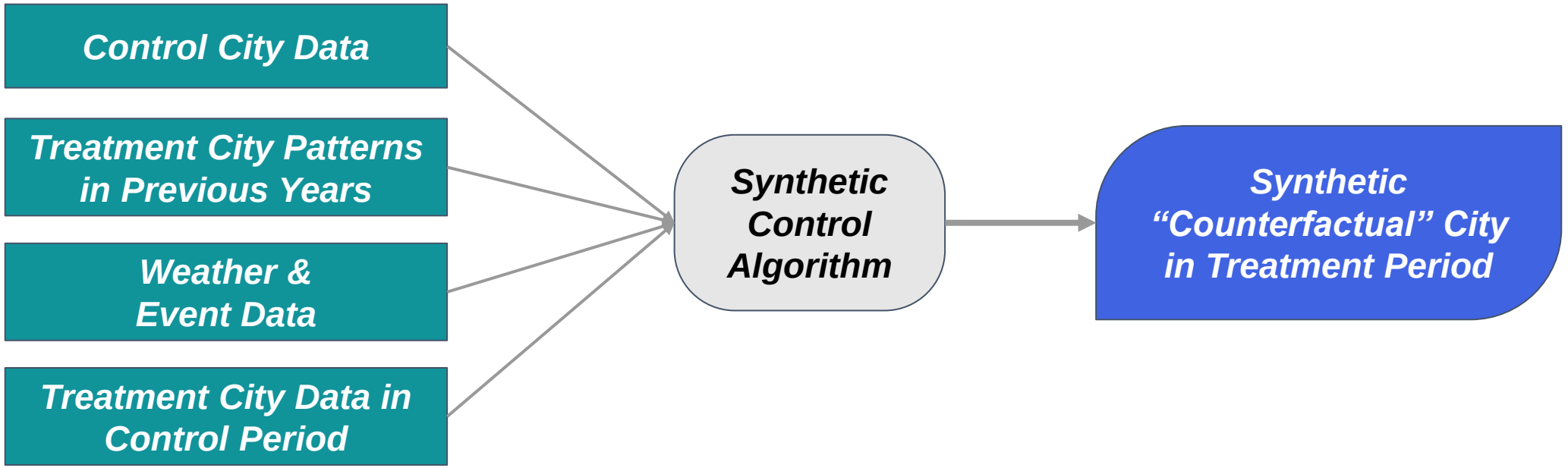


Step 2: Launch experiment in just 1 city, and calculate treatment effect as difference between synthetic city and the actual measurement in treatment city

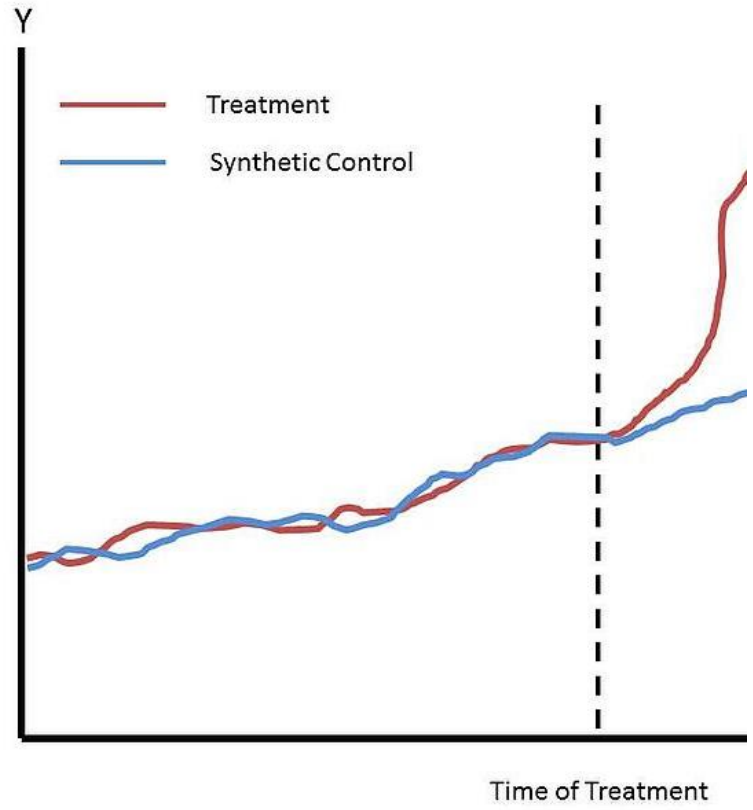
Initial Plot credit:
Uber MX team

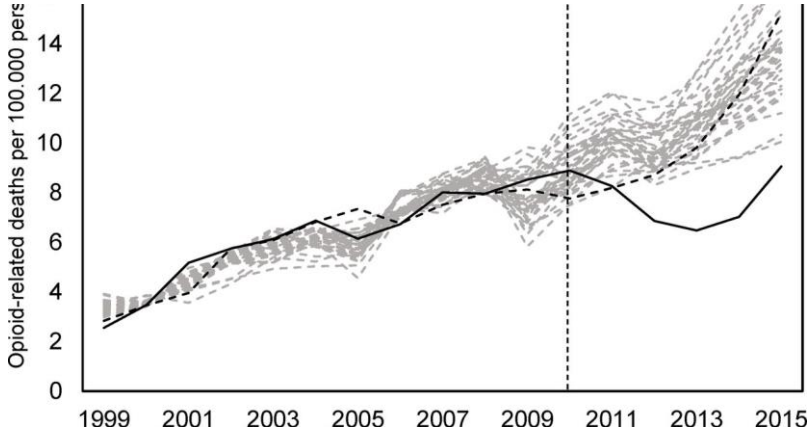
How does Uber form their synthetics?

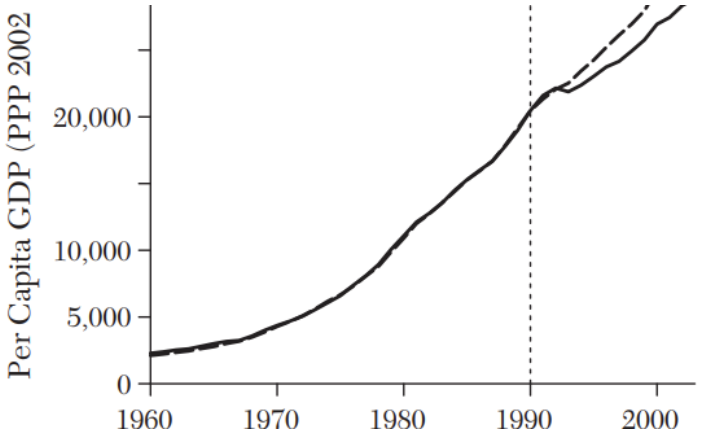
They use a model to “forecast” outcomes in our target treatment city, using only data from control cities, weather and events, and previous year’s data in that city:

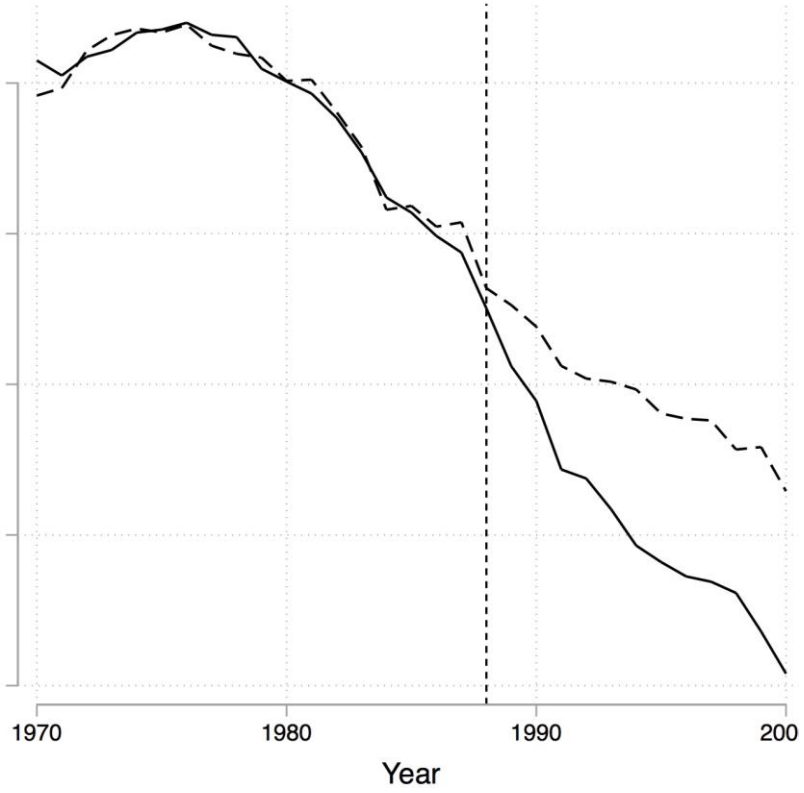


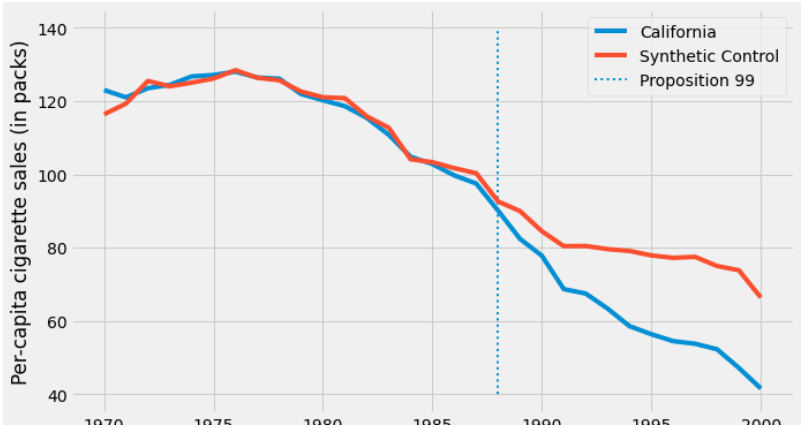
Slide credit:
Uber MX team











Classic synthetic control plot

Warning

"At the same time, the validity of synthetic control estimators depends on important practical requirements. Perfunctory applications that ignore the context of the empirical investigation and the characteristics of the data may miss the mark, producing misleading estimates."

"Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspect." Alberto Abadie, 2021

Challenges

- Overfitting during training period (have a "validation" period)
- What if something happens during treatment? (hurricane in Miami)

Experimentation culture

Classical power analyses

- In a past statistics class, you might have learned “power analysis”
  - If the “true effect” is at least as big as X, then an experiment on N samples will *reject the null hypothesis* at least Z% of the time.
  - If the true effect is 0 (null hypothesis is true), the experiment will falsely reject it no more than α% of the time.
  - Given X, Z%, and α%, easy to calculate fixed sample size N
  - You run an experiment, with N samples
- This reflects a “scientific” approach to experiments: an experiment that rejects false hypotheses and accepts true hypotheses
- This is ***a wrong approach*** in practice

  You don’t care about doing good science

Problems with classical approach

Your goal: you want to quickly launch amazing products (large $Y_1 - Y_0$)

- It's ok to not launch an "ok" product that would reject the null (small $Y_1 - Y_0$)
- Also ok to sometimes launch "useless" products (zero $Y_1 - Y_0$)
- Never want to launch a product that hurts your metrics (very negative $Y_1 - Y_0$)

Your advantage: You have *many* possible products to launch/experiments to run; limitation is sample size

Classical approach:

- Sample size N optimized to find small effects (small $X = Y_1 - Y_0$)
- Wastes samples & time that could be spent on other experiments

“Discovery-driven” experimentation

Insight: You have *many* experiments. If one product looks mediocre early in the experiment, just move on

Run an experiment *just* long enough to determine if it’s an amazing product (large $Y_1 - Y_0$) or if it’s a dud

- “Peek”, but smartly this time, based on $\hat{Y}_1 - \hat{Y}_0$
- Upper threshold $\mathrm{u}(n)$ to stop experiment and *declare victory;* decrease with more samples n
- *Lower* threshold $\ell(\mathrm{n})$ to stop experiment and *declare loss;* increases with more samples n

Result

- Bad science: you’ll often reject small, positive products
- But you’ll find amazing products as quickly as possible

More generally, *adaptive experimentation* when have many different arms of a treatment (for example, 41 shades of blue); remove poor arms quickly, focus on best ones

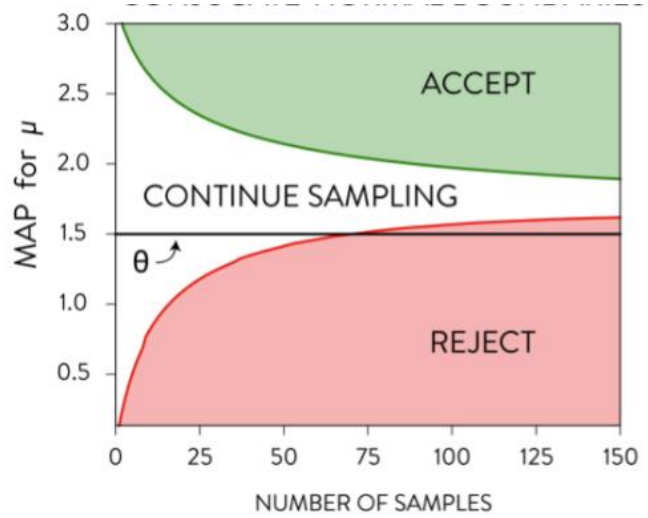


Image credit: Large scale experimentation | Stitch Fix Technology – Multithreaded, Sven Schmit, Brian Coffey

Paper: “Optimal Testing in the Experiment-rich Regime” Sven Schmit, Virag Shah, Ramesh Johari

Simulation

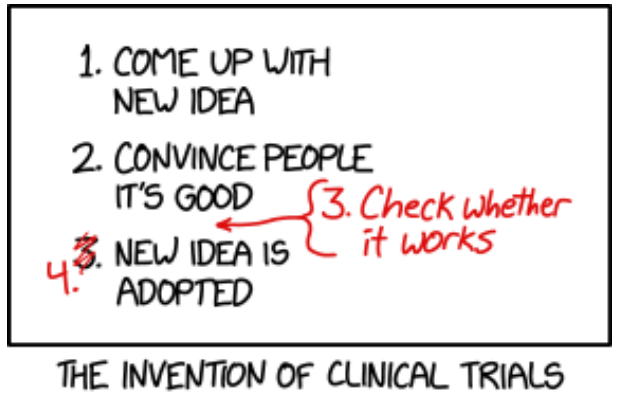
Build a “simulator” for how your market performs

Lyft blog post: have drivers drive around in simulator, matched with riders using their algorithms

- Can simulate how different matching algorithms perform
- Also can simulate pricing algorithms
  - Need assumptions on how individual riders will respond (big assumption)
  - Under these assumptions, can learn market-wide effects of algorithm
    i.e., simulate interference patterns, if we know “first-order” effects of product
- Also can simulate different *experimentation* methods
  - Know the ground truth, simulate what different protocols would find
  - For example: in homework we simulated different experiment designs using the same historical data from an A/B test

General pipeline of launching a product

- Come up with idea, iterate on design
- Code it up, and evaluate on simulator
- Test in real experiment in one city/market
- If that goes well over time, roll out in multiple markets
- Continue rolling out in more and more markets
- Eventually, will have rolled out everywhere



THE INVENTION OF CLINICAL TRIALS

[xkcd: Clinical Trials]

Universal holdout

Downsides of standard approaches:

- Test one product at a time
- Usually enroll as few users as possible (don't want to waste sample size)
- Experiments are usually short → Don't observe long-term metrics

What if you want to know, "What is the total effect on everything I launched last quarter on customer retention?"

Solution: Universal holdout

- Each quarter (or month or year…), hold out same set of users from *every product* you launch that quarter
- End of quarter, compare metrics for that group to all other users; re-enroll a new set of universal holdout for next quarter

Universal Holdout Groups at Disney Streaming | by Tian Yang | disney-streaming | Oct, 2021 | Medium

Miscellaneous topics in experimentation

Ethics and Communication

When is an experiment unethical to run?

What if your strong intuition is that the new product is bad?

- Challenge trials in medicine: very controversial, especially during Covid
- Can you purposely degrade your product to evaluate how it usually performs?

What if Uber purposely broke surge during New Years?

What if your strong intuition is that the old product is bad?

Should you launch the replacement product immediately, or can you experiment first?

"Objecting to experiments that compare two unobjectionable policies or treatments" PNAS, Michelle Meyer et al. 2019

Various treatment effects

- So far we've discussed the "Global Average Treatment Effect" (GATE)
  - How does the world where *everyone* receives the treatment, compare to the world where *everyone* receives the control?
- If there is no interference (and 1 more condition; SUTVA), this is equal to the "Average Treatment Effect" (ATE)
  - On average, if *I* receive the treatment, how does that compare to if *I* received the control?
- "Local Average Treatment Effect" (LATE) or "Complier ATE" (CATE)
  - Example: if treatment is *access to vaccine*, CATE only counts as treated those who *take* the vaccine; ATE would count everyone given *access*
- *Heterogeneous treatment effect:*
  - *What if the treatment effect differs for different sub-populations?*
  - Example: Giving students coupons to Broadway shows vs giving professors coupons

Experimentation summary

- Classic A/B Test
- In social networks and marketplaces, *interference* ruins tests
  - Social networks: Social effect; Me getting treatment effects you
  - Marketplaces: Competition and scarcity introduces interference
- Experiments in the face of interference:
  - (Spatial or Graph-based) Cluster randomized assignment
  - Time-based experimentation: Switchbacks
- Causal inference without experiment: Synthetic control
- Naïve peeking in experimentation is bad, but can be done smartly

Announcements

- In-person guest lecture (Lily Xu) on Wednesday – please attend
- Fill out the “miscellaneous topics” survey
- HW4 due tomorrow
- Quiz 4 released Wednesday, due Sunday
- Project details released this week
- OHs
  - Zhi today (regular time; Zoom)
  - No Nikhil Wednesday office hours this week
  - Friday
    - Nikhil 12:30 – 1:30
    - Zhi 1:30 – 2:30