ORIE 5355: People, Data, & Systems

Lecture 14: Experimentation in marketplaces

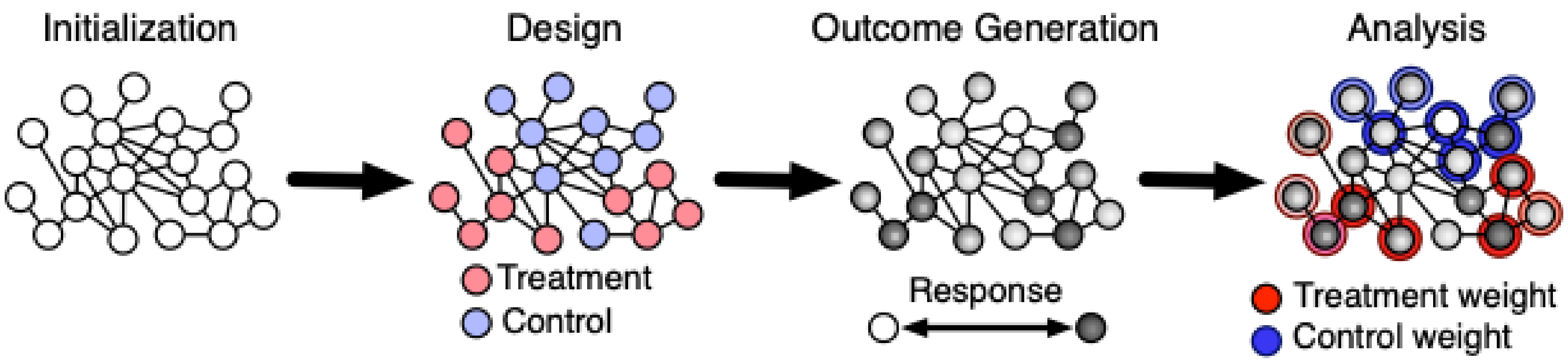
Nikhil Garg

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

Announcements

- HW4 released; due next week
- Quiz 4 next week
- Project details released in next week
  - Project partner form on EdStem
- OHs
  - Mine today – 2-3 (In Person + Zoom)
  - Mine Friday (1 – 1:45, zoom only)
  - Zhi Friday (1:30 – 2:30, same zoom link as mine)
  - Zhi Monday
  - No Wednesday office hours next week

Last time: Network Experimentation

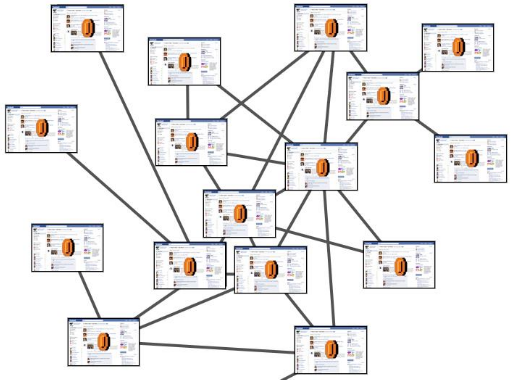


- Initialization: An empirical graph or graph model
- Design: Graph cluster randomization
- Outcome generation: Observe behavior (or observe model)
- Analysis: Discerning effective treatment

Slide credit: Johan Ugander, Stanford

General lesson: “unit” of randomization

- If you randomize at the “individual” level (each individual is its own “unit”), then treatment and control units can interfere with each other
- Solution is often to change the *unit of randomization*: randomize “clusters” instead of individuals
  - Hope: clusters are *close to independent*
  - If independent, experiment is *unbiased*
- Downside: Experiment “variance” goes down with sample size of experiment
  - Before: Sample size is *millions* (of users)
  - Now: Sample size is *hundreds* (of clusters)
- Same bias-variance trade-off we’ve seen before!



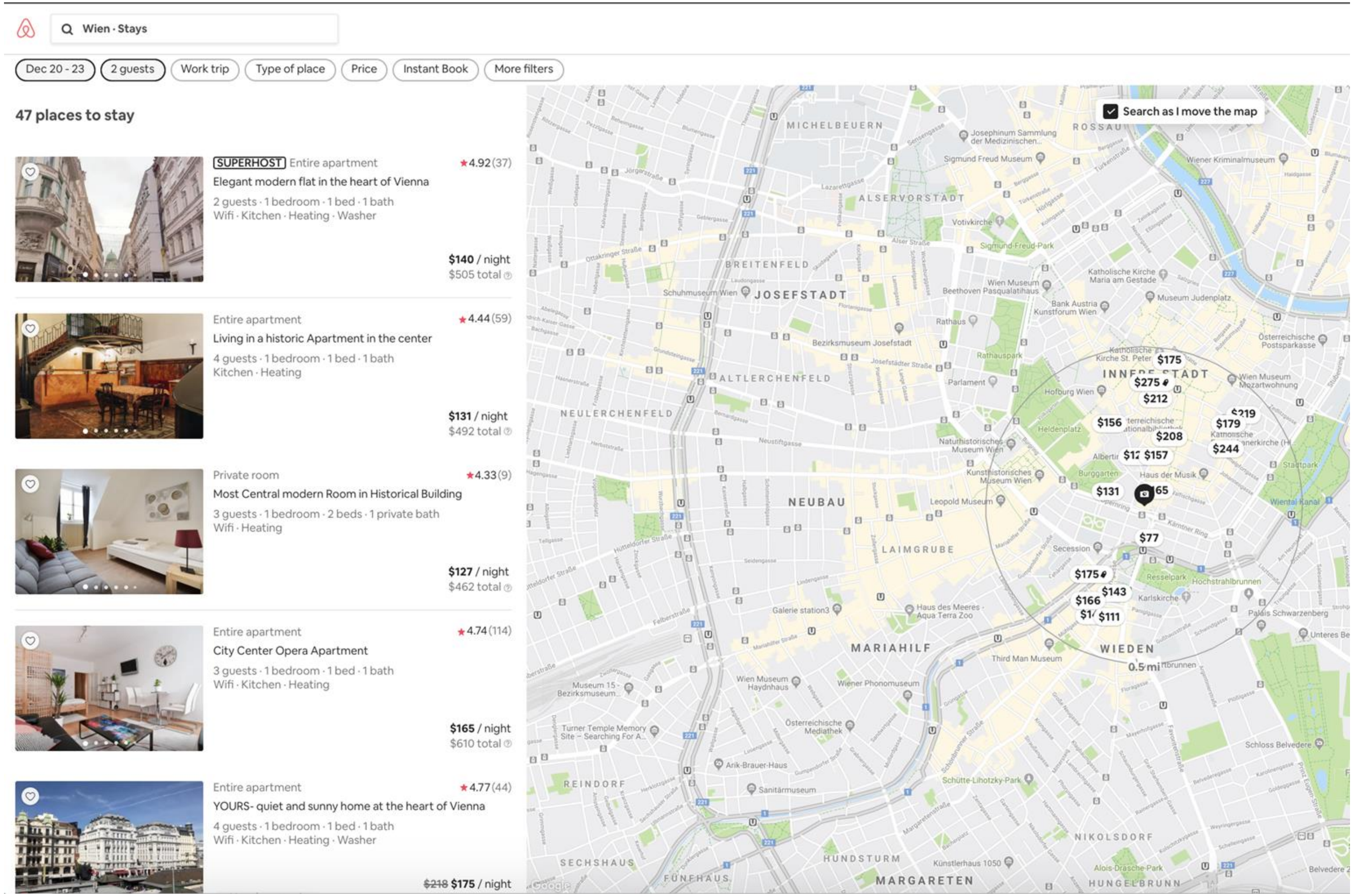
Interference in marketplaces

- Interference between treatment and control also arises in marketplaces
- In social networks: Interference because use case is *social* – me getting video messaging doesn't matter if none of my friends get it
- In markets, interference rises from *competition and capacity constraints*
- If I make half the products cheaper, customers will *increase* their purchases of the cheaper items…why?
    - Go from not purchasing at all, to buying the now cheaper item (**new customer**)
    - *Decrease* their purchases of the more expensive items (**cannibalization**)
- *Not* representative of what would happen if I make *all* my products cheaper

  Cannibalization effect would not occur; only attraction of new customers
- Hannah's lecture and today: experimentation in marketplaces under interference

Graph cluster randomization in marketplaces

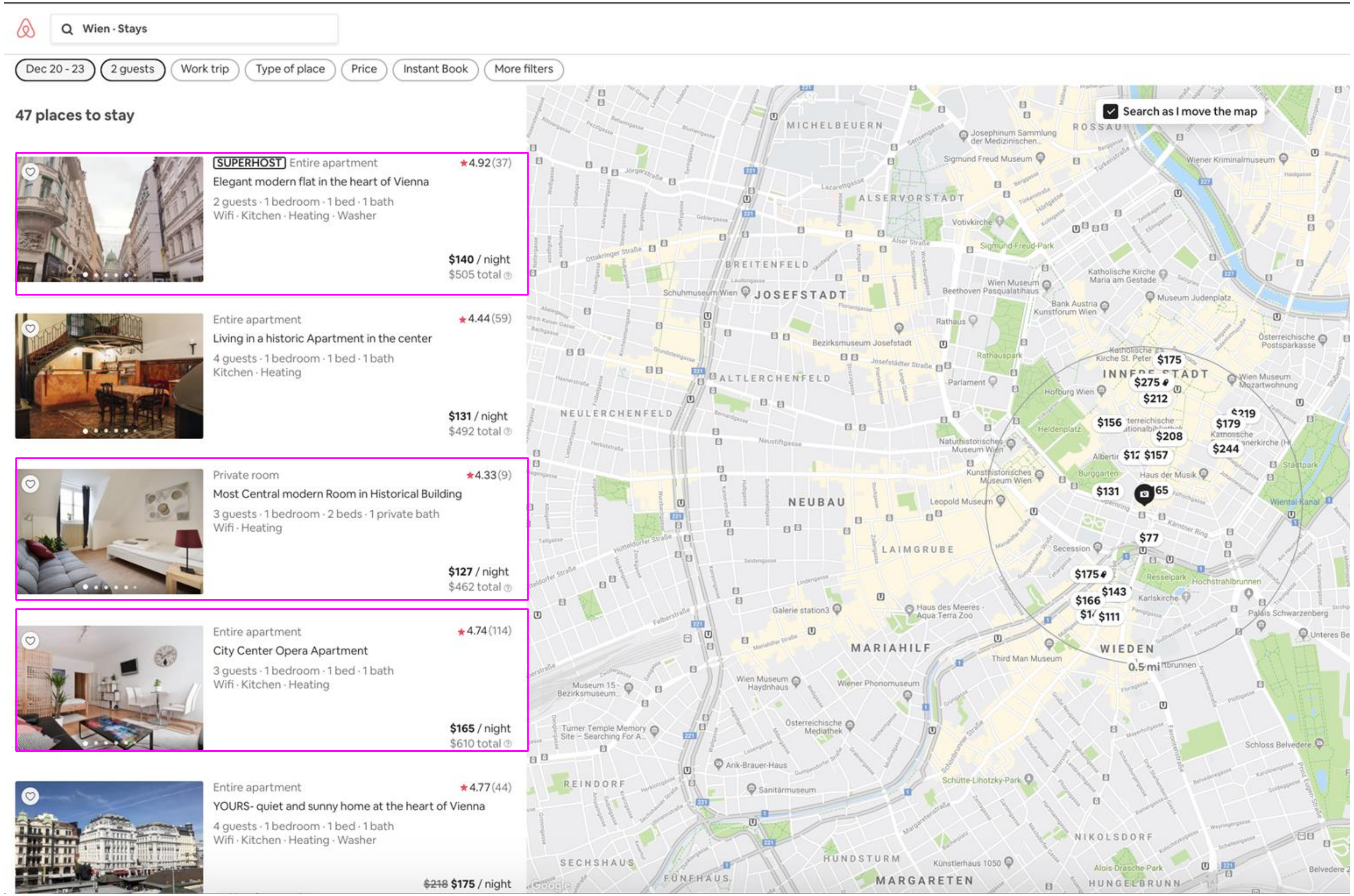
Example 1: price change experiment on Airbnb



Slide credit: Dave Holtz, UC Berkeley

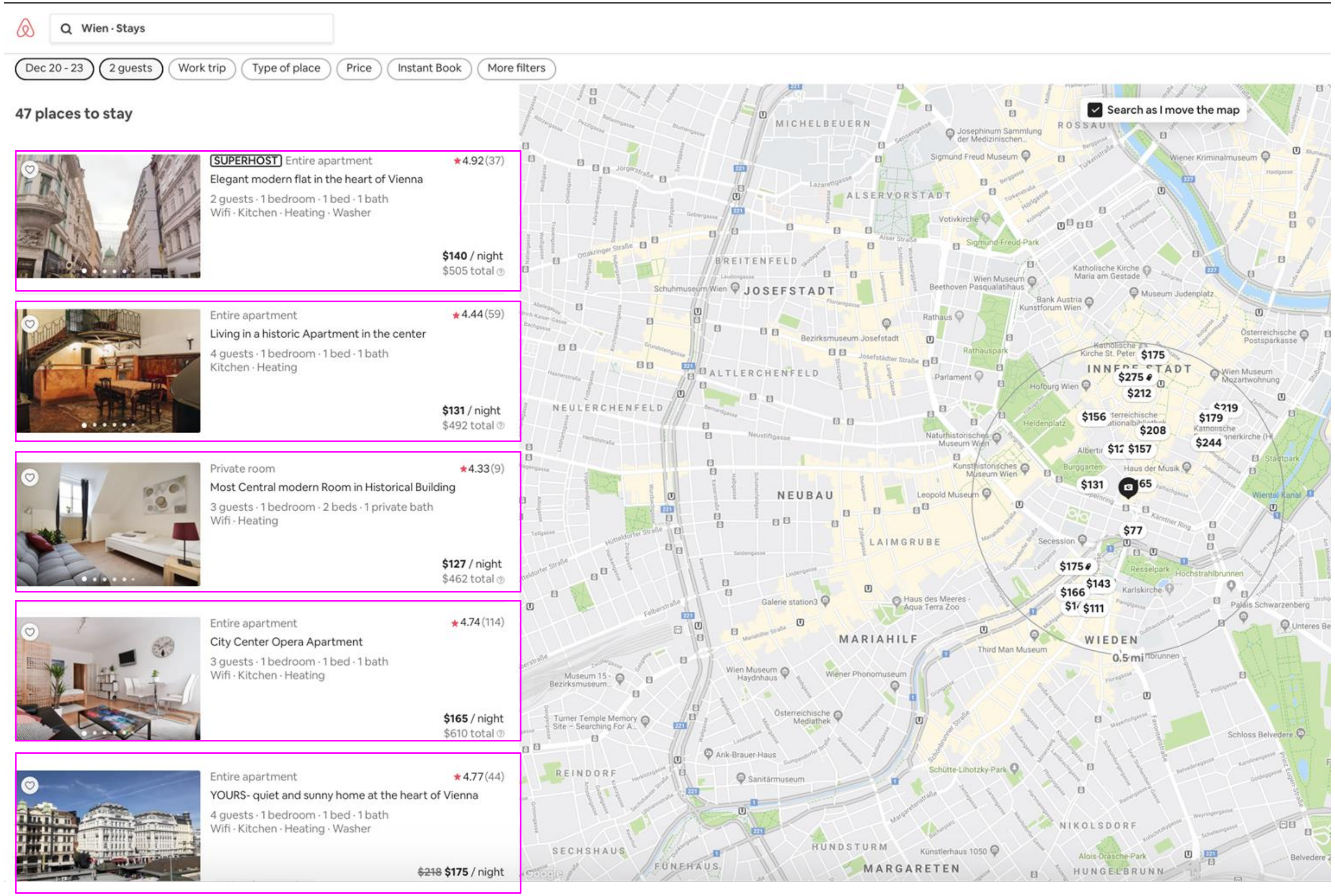
Example 1: price change experiment on Airbnb

If lower fees on half of the listings, bookings for those listings ↑ 3% ☺



Slide credit: Dave Holtz, UC Berkeley

Example 1: price change experiment on Airbnb

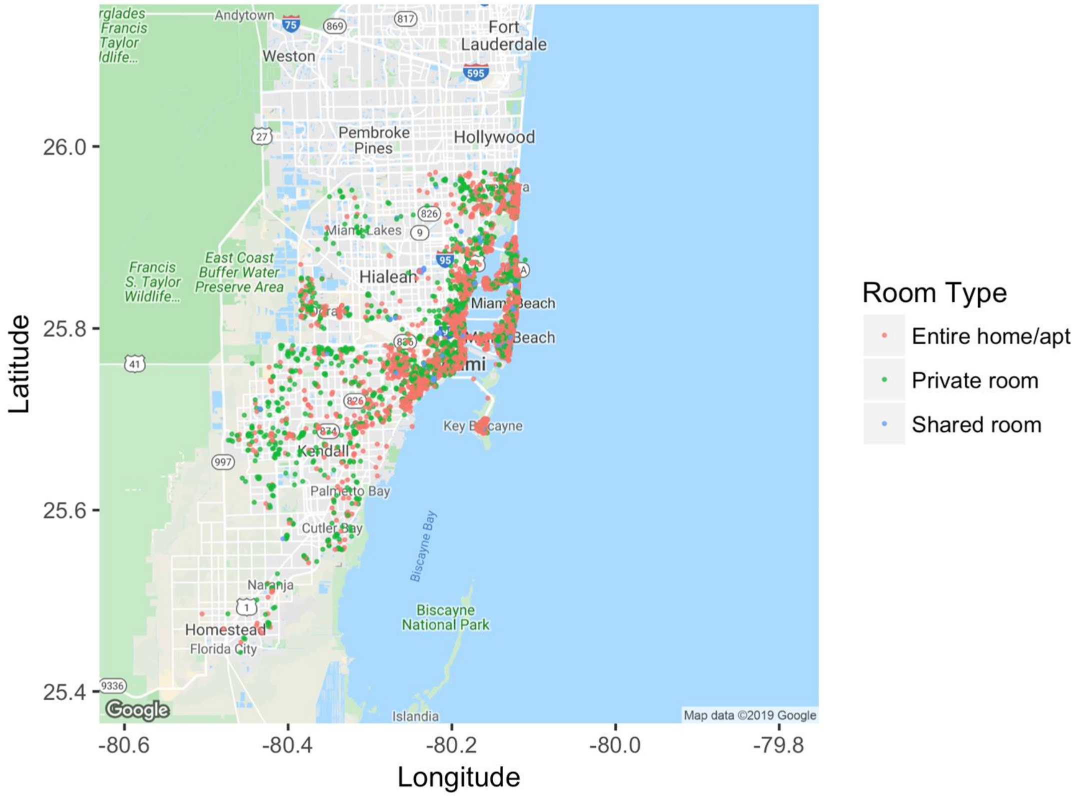


If lower fees on all the listings, Overall bookings flat ☹

Slide credit: Dave Holtz, UC Berkeley

Approach 1: transform the marketplace into a network

Slide credit:
Dave Holtz,
UC Berkeley



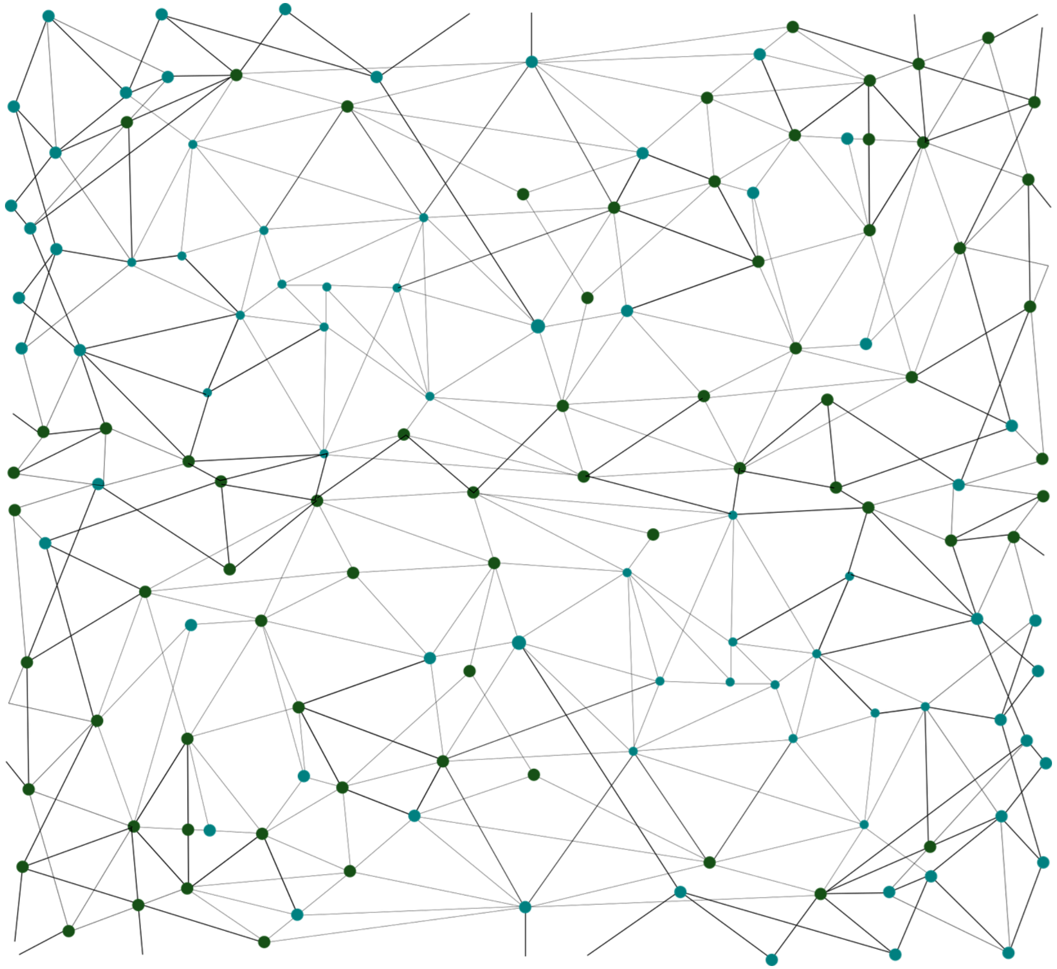


Image source

Network experiment designs + analysis techniques

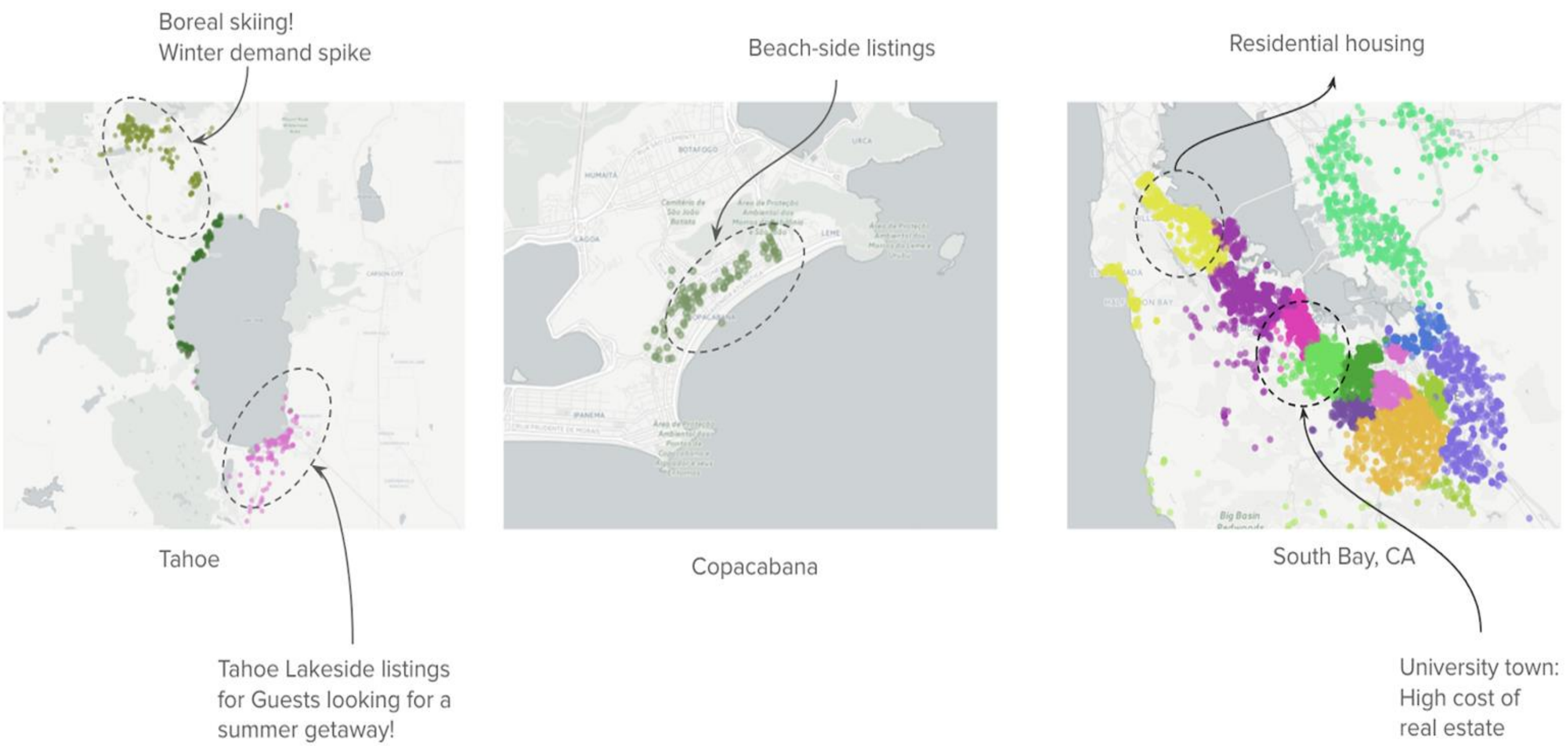
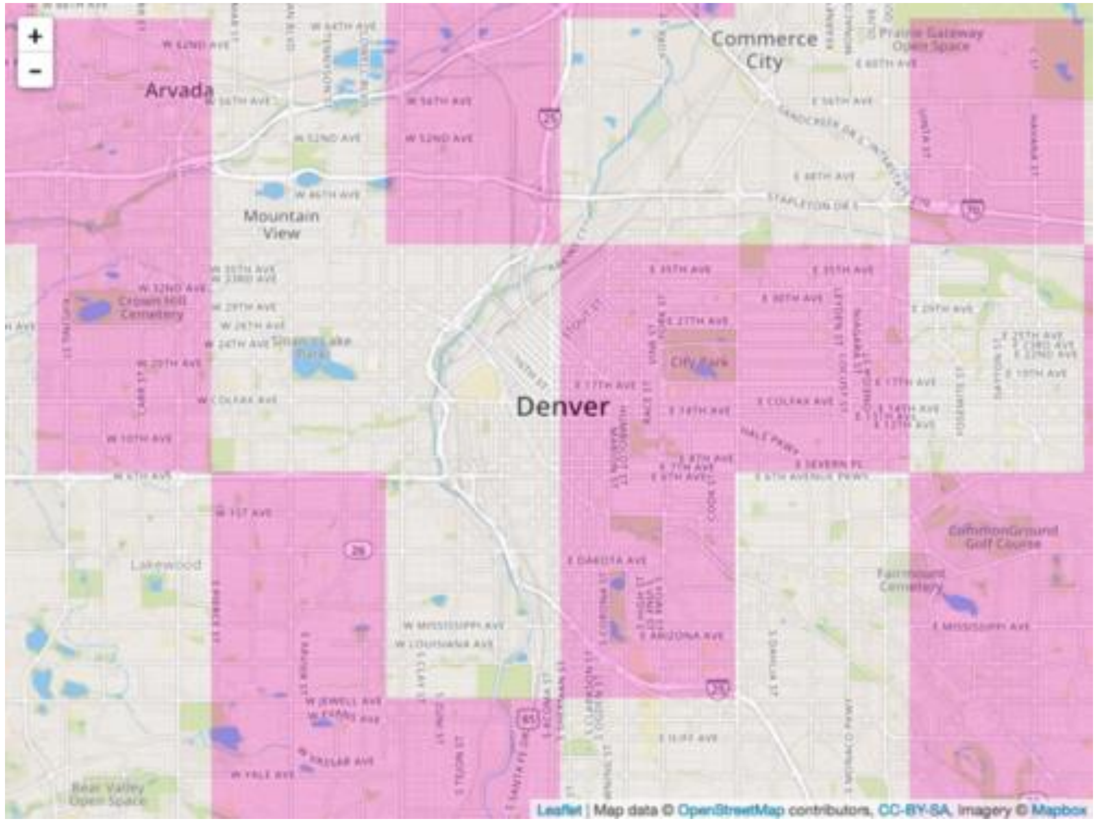


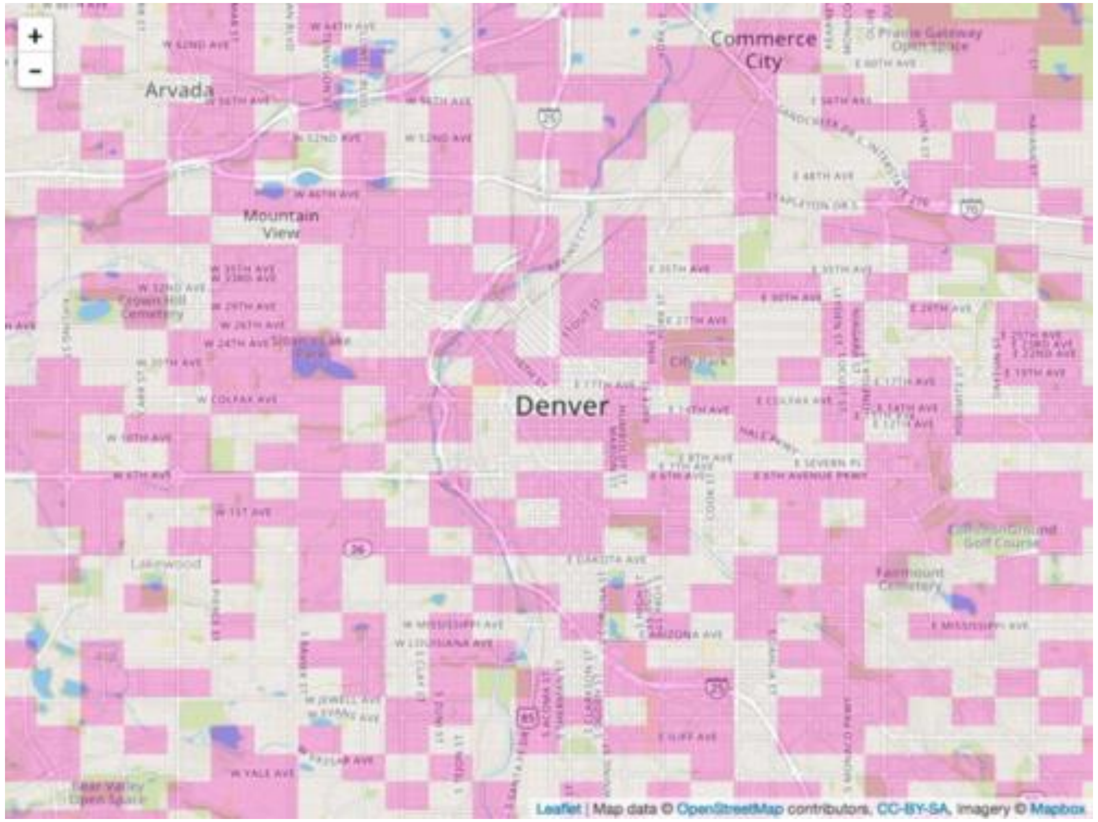
Image credit: Dave Holtz, UC Berkeley

- Now, listings are connected if they tend to be *substitutes*
- Much more complicated to learn the network structure
- Once have network structure, use cluster randomization techniques
- Challenge: "graph" might be too interconnected

See Hannah's lecture on Monday for more discussion on this

Spatial randomization in ride-hailing





[Experimentation in a Ridesharing Marketplace | by Nicholas Chamandy | Lyft Engineering]

Beyond spatial (and graph cluster) randomization: experimenting over time

Switchbacks

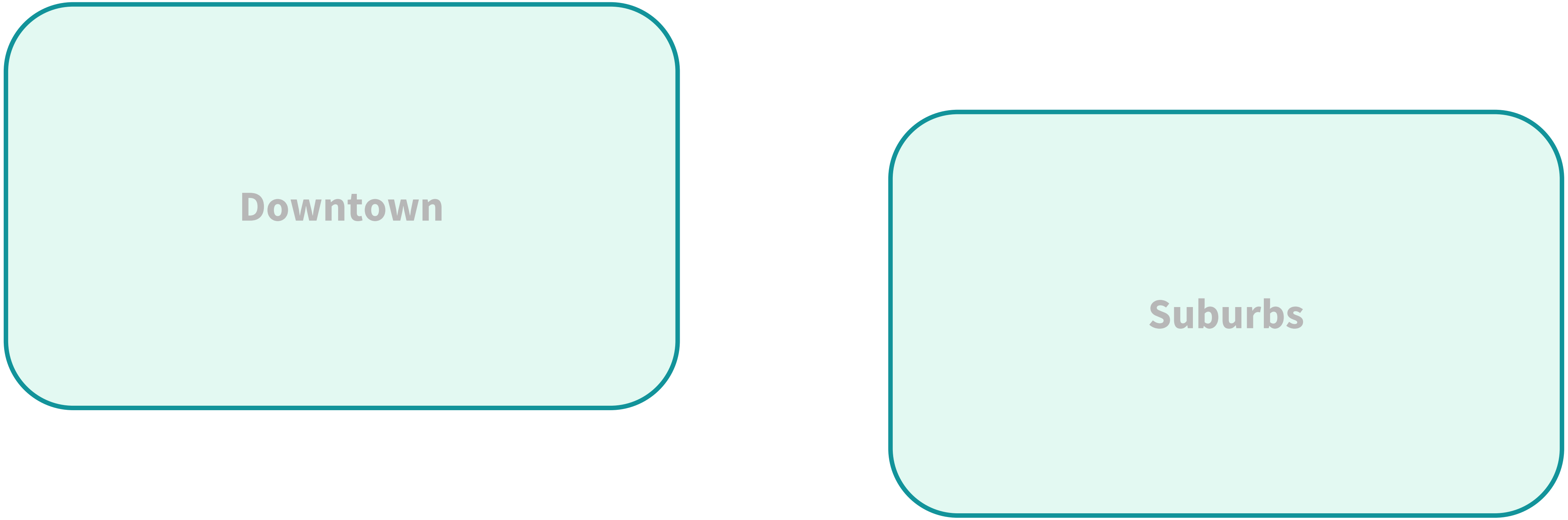
Why is cluster randomization not enough?

- Often difficult to define the clusters
- There legitimately might not be enough “clusters” that don’t interfere with one another
  - In AirBnB, rentals near Disney Land (in Los Angeles) might compete with rentals near Disney World (in Orlando)
  - In ride-hailing, a driver in a suburb could be instead choose to drive in the city

Slide credit:
Uber MX team

Driver Positioning Example

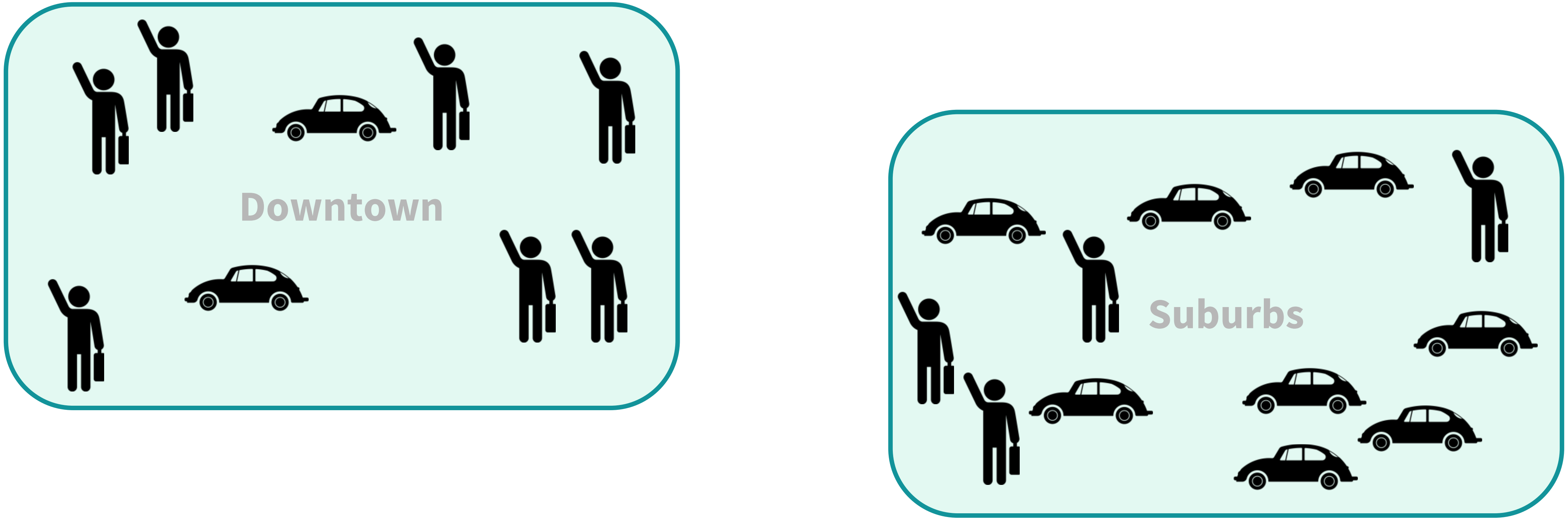
Suppose our city has two geos: downtown and the suburbs



Driver Positioning Example

Slide credit:
Uber MX team

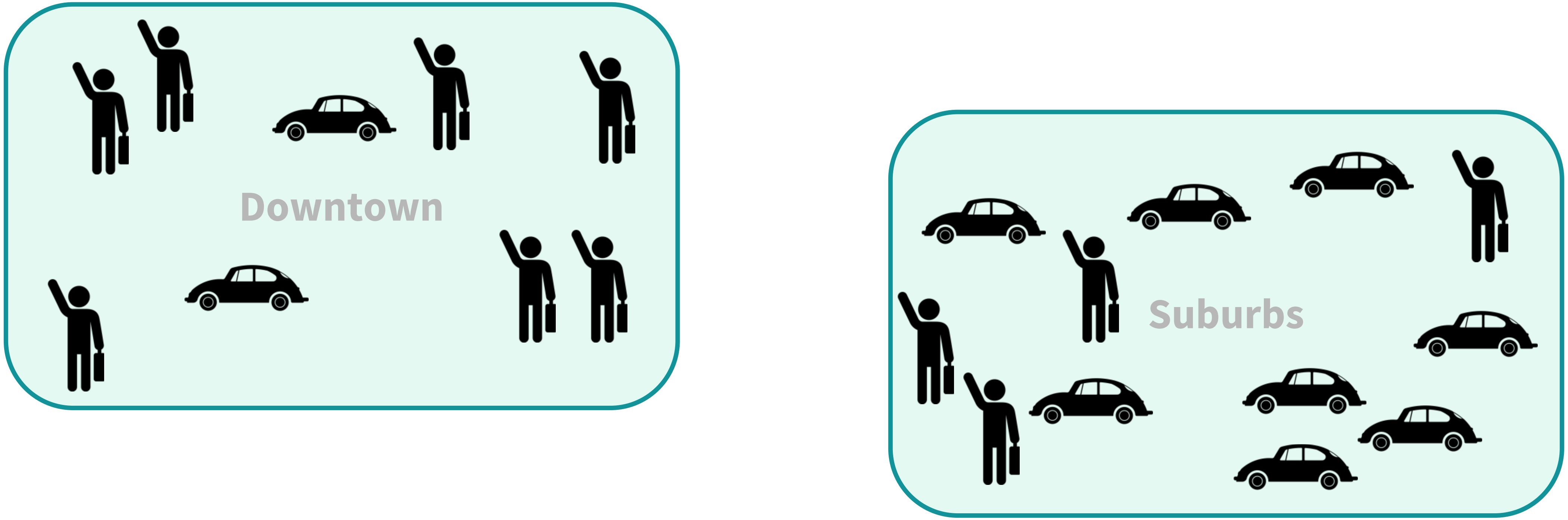
We notice that we are chronically undersupplied in downtown and oversupplied in the suburbs. Uber is concerned that this adversely impacts driver earnings.



Slide credit:
Uber MX team

Driver Positioning Example

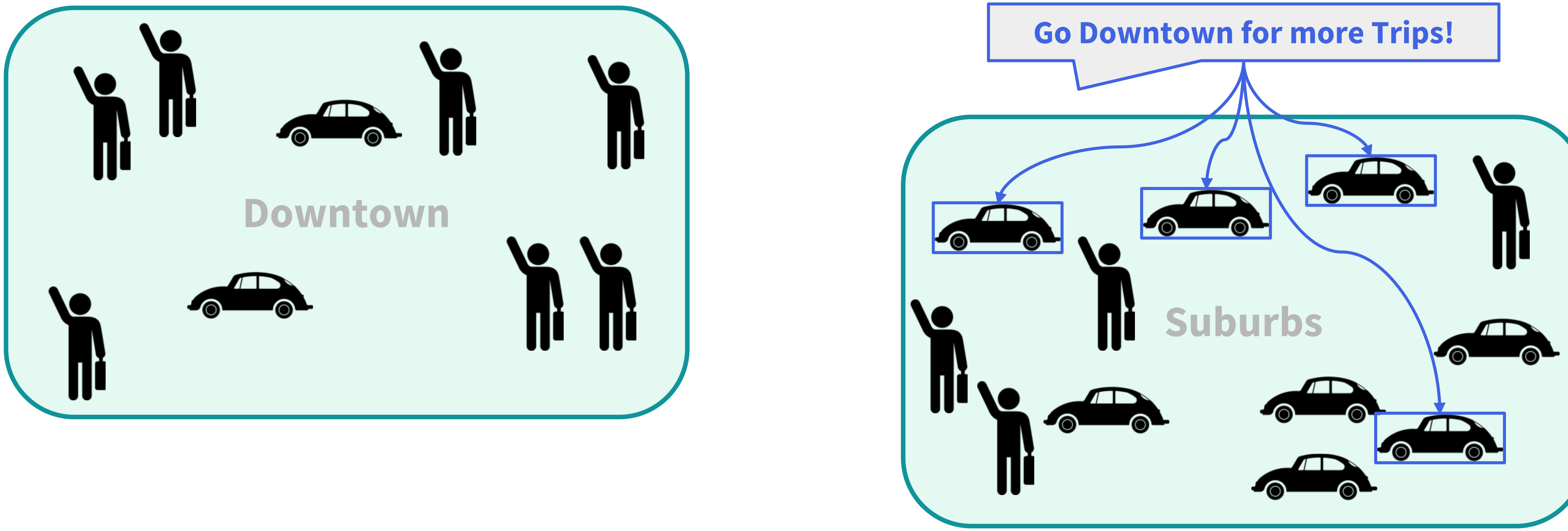
Tech builds a product that dynamically identifies over- and under-supplied areas and sends repositioning recommendations to drivers in over-supplied areas.



Driver Positioning Example

Slide credit:
Uber MX team

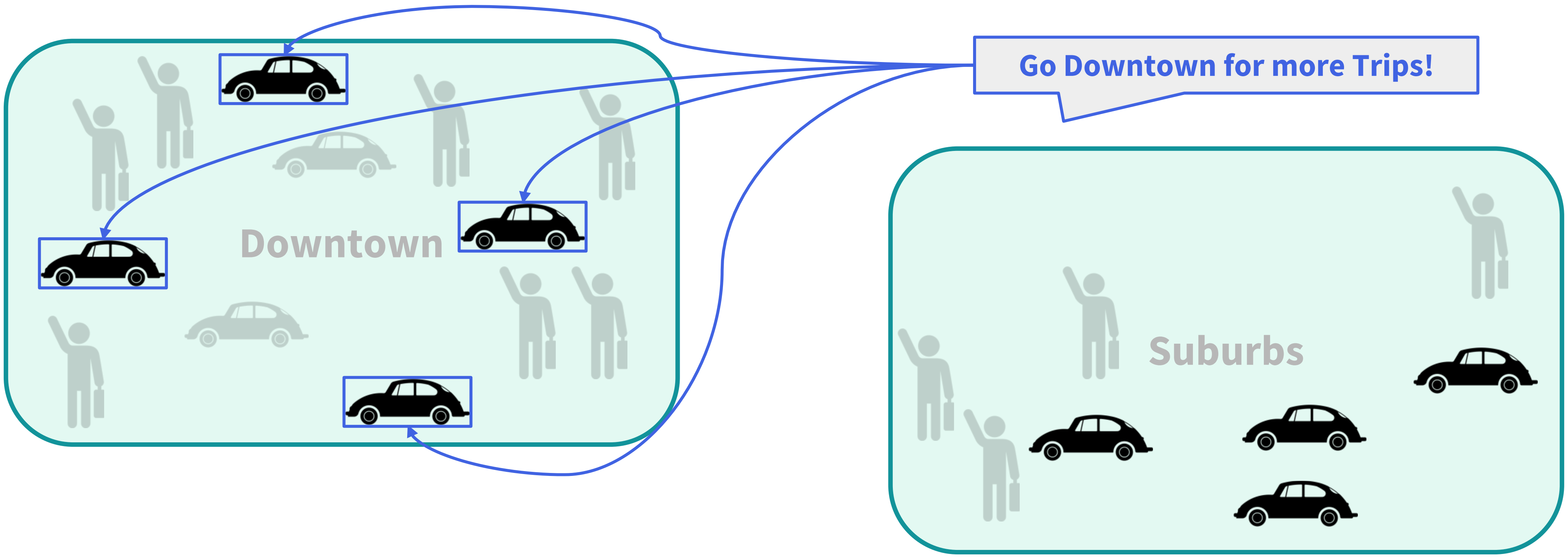
To test this, Uber runs a driver A/B experiment where 50% of drivers in the Suburbs are asked to relocate to Downtown. (The other 50% do not get recommendations.)



Slide credit:
Uber MX team

Driver Positioning Example

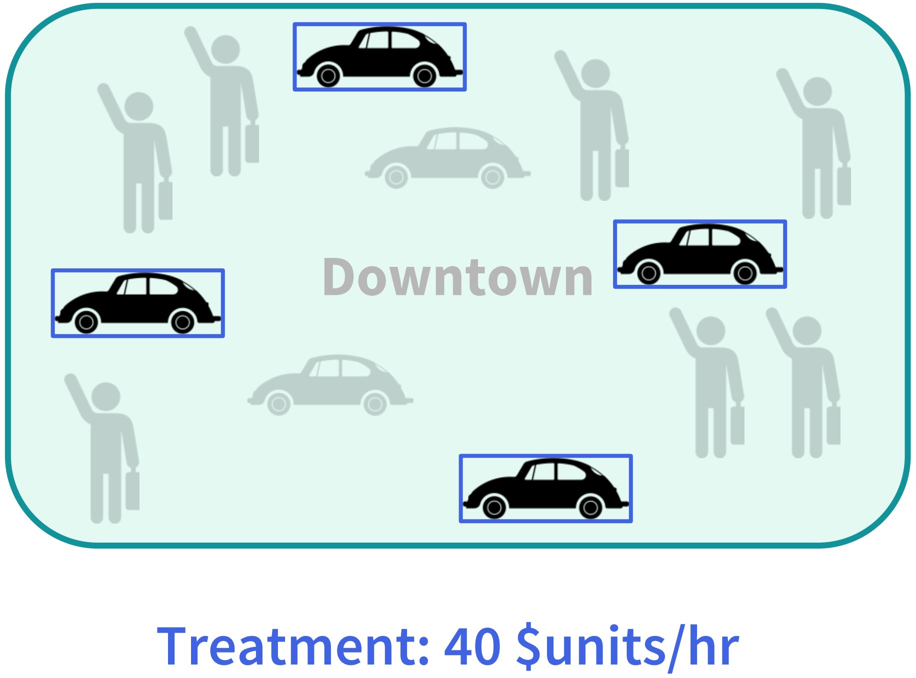
Suppose the drivers follow the recommendation and relocate

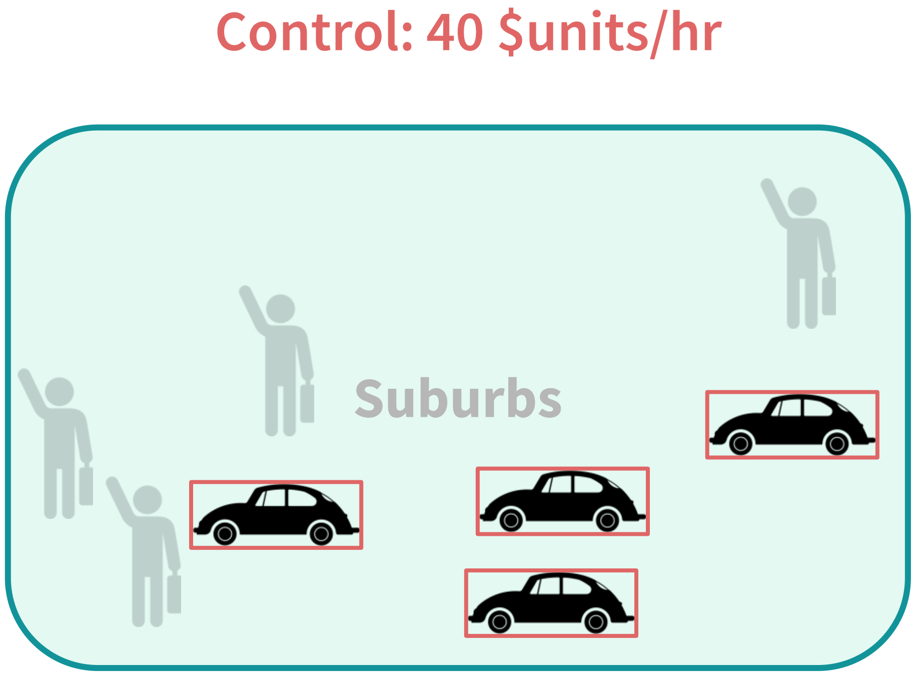


Driver Positioning Example

Slide credit:
Uber MX team

Suppose we find that drivers who got the repositioning message (and relocated) had the same earnings per hour as drivers who didn't get the message!

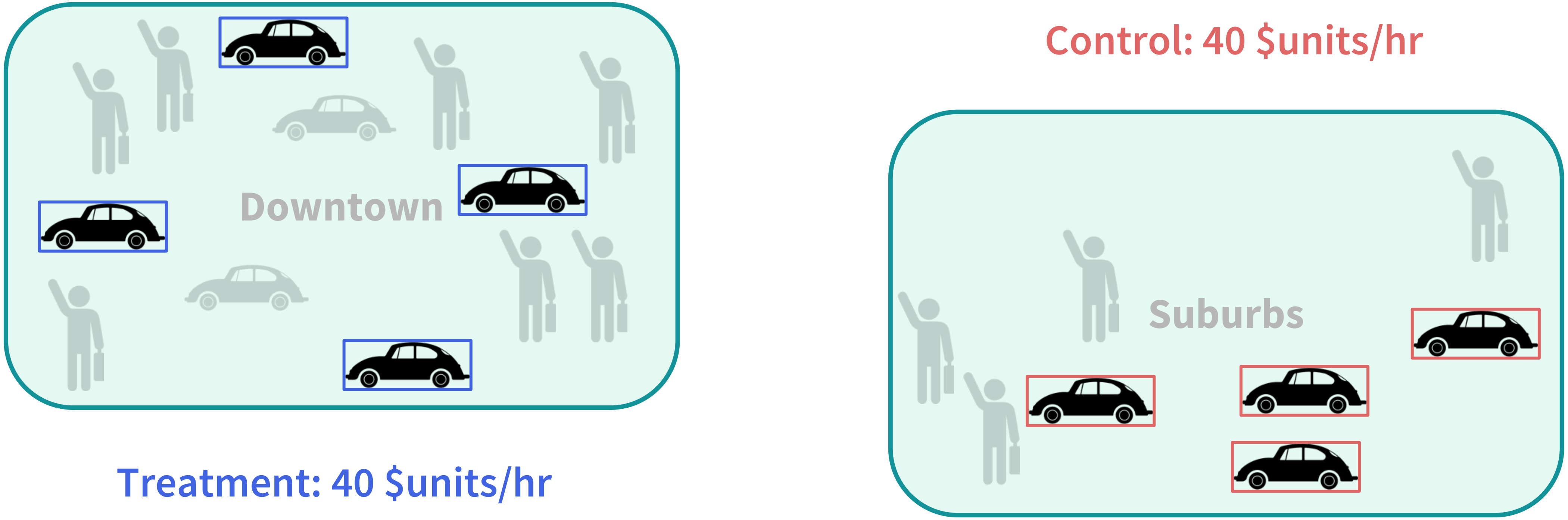




Driver Positioning Example

Slide credit:
Uber MX team

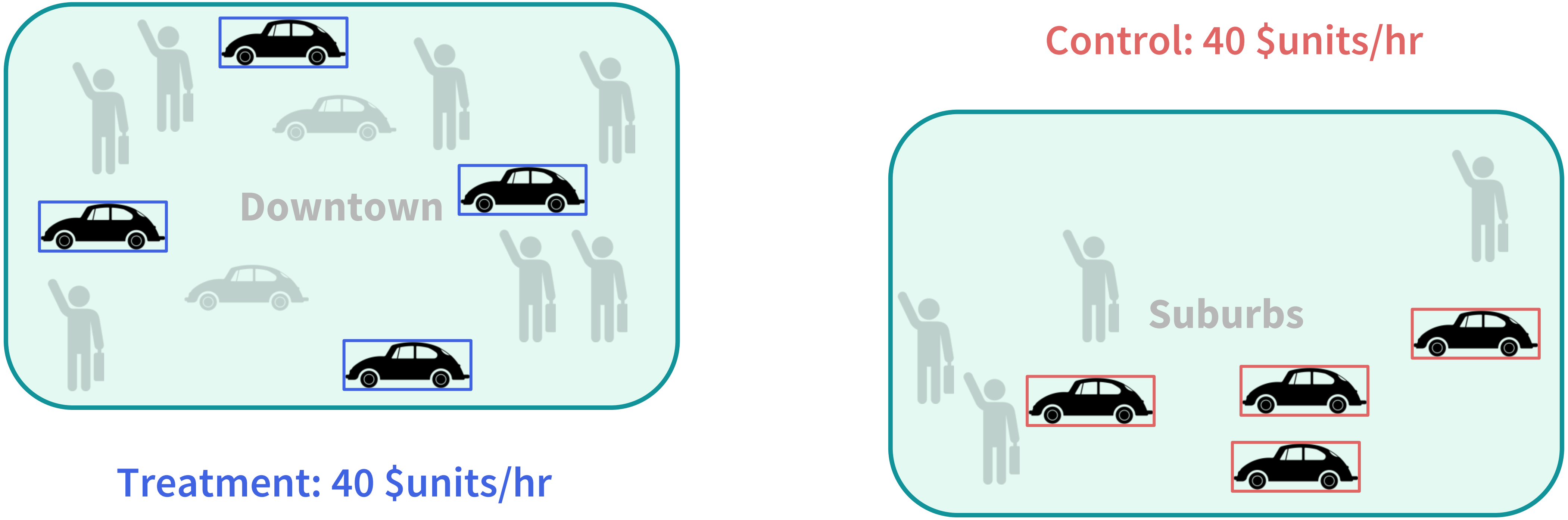
On the basis of this A/B earnings comparison, we might conclude that this product did **nothing** to raise driver earnings.



Driver Positioning Example

Slide credit:
Uber MX team

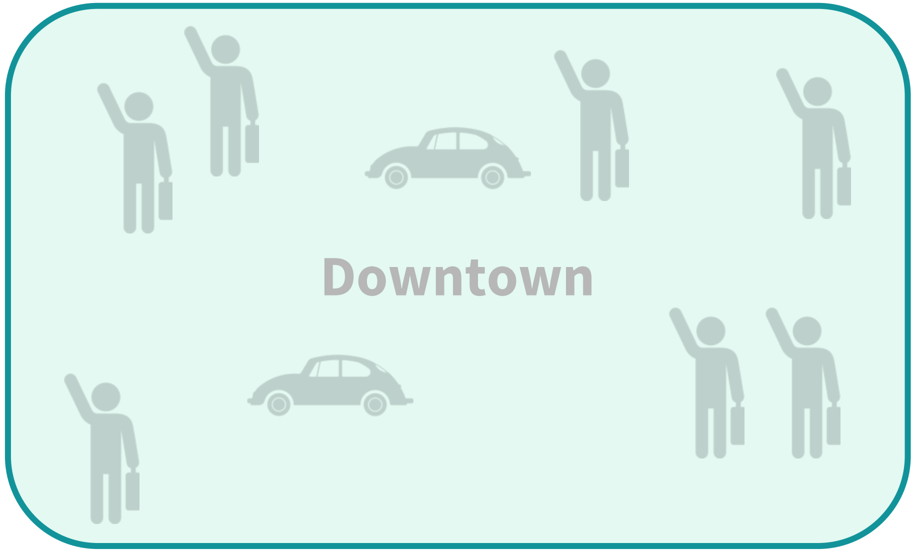
The mistake here is that by moving drivers out of the Suburbs, we increased the earnings opportunities of the Control drivers. Control was **contaminated**.



Driver Positioning Example

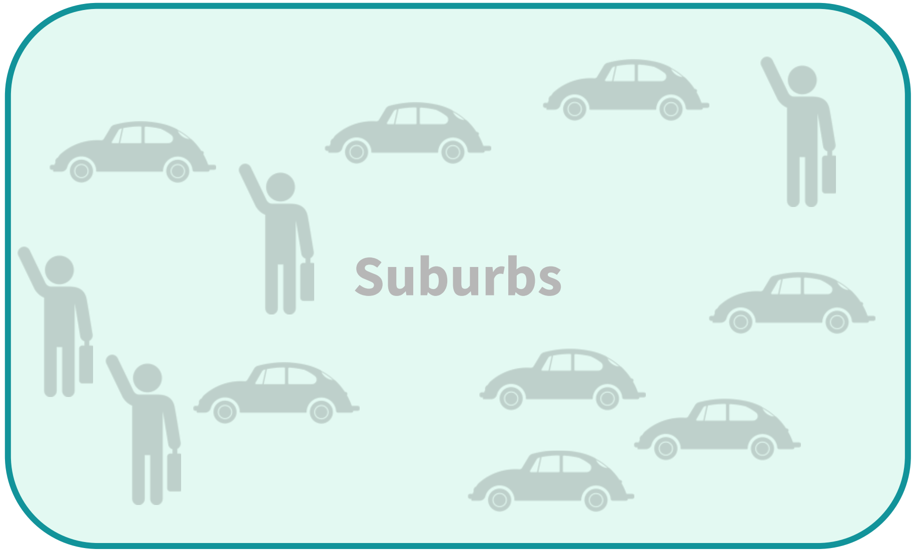
Slide credit:
Uber MX team

Counterfactually, *had we not sent the repositioning messages*, we might have seen the following driver earnings:



Counterfactual
Downtown: 40 $units/hr

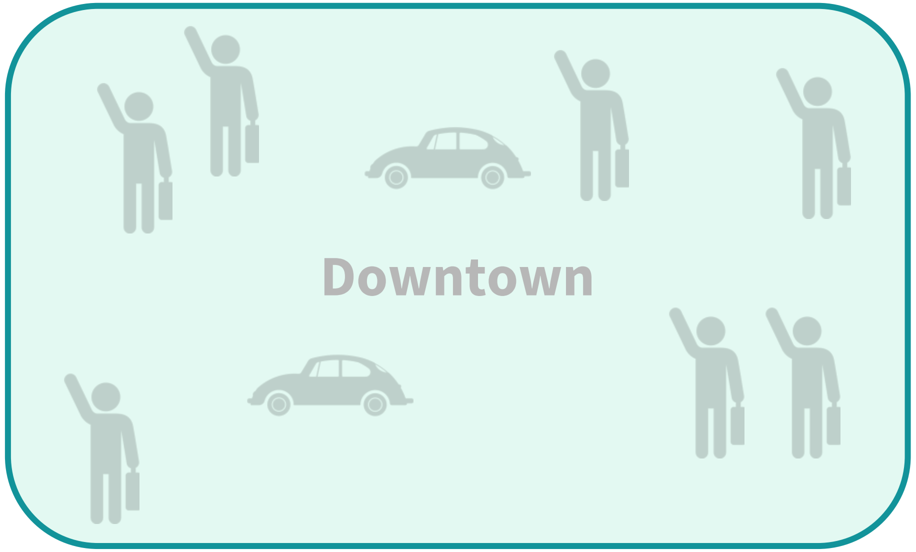
Counterfactual
Suburbs: 30 $units/hr



Driver Positioning Example

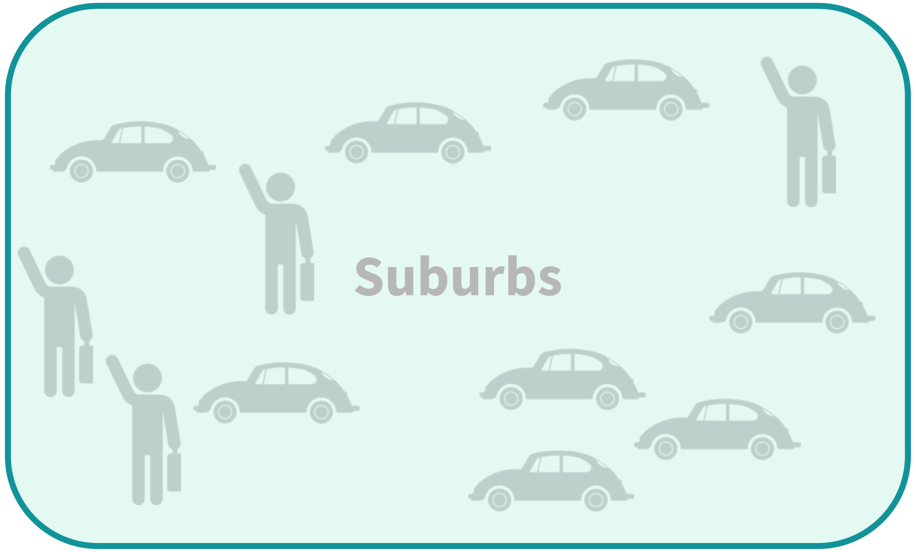
Slide credit:
Uber MX team

So in fact, the supply repositioning product increased earnings by **10 $units/hr** for both the treatment *and* the control group!



Counterfactual
Downtown: 40 $units/hr

Counterfactual
Suburbs: 30 $units/hr



Why is cluster randomization not enough?

- Often difficult to define the clusters
- There legitimately might not be enough “clusters” that don’t interfere with one another
  - In AirBnB, rentals near Disney Land (in Los Angeles) might compete with rentals near Disney World (in Orlando)
  - In Uber, a driver in a suburb could be instead choose to drive in the city
- What happened?
  - Giving the treatment to (some) drivers in the suburbs *decreased* competition for other drivers in the suburb, and *increased* competition for drivers in downtown
  - Both driver-level A/B testing and graph-cluster randomization would learn biased estimates
- We’d have to cluster at the city-level to prevent such interference
  - Still might not be enough: drivers commute from Sacramento to SF to work

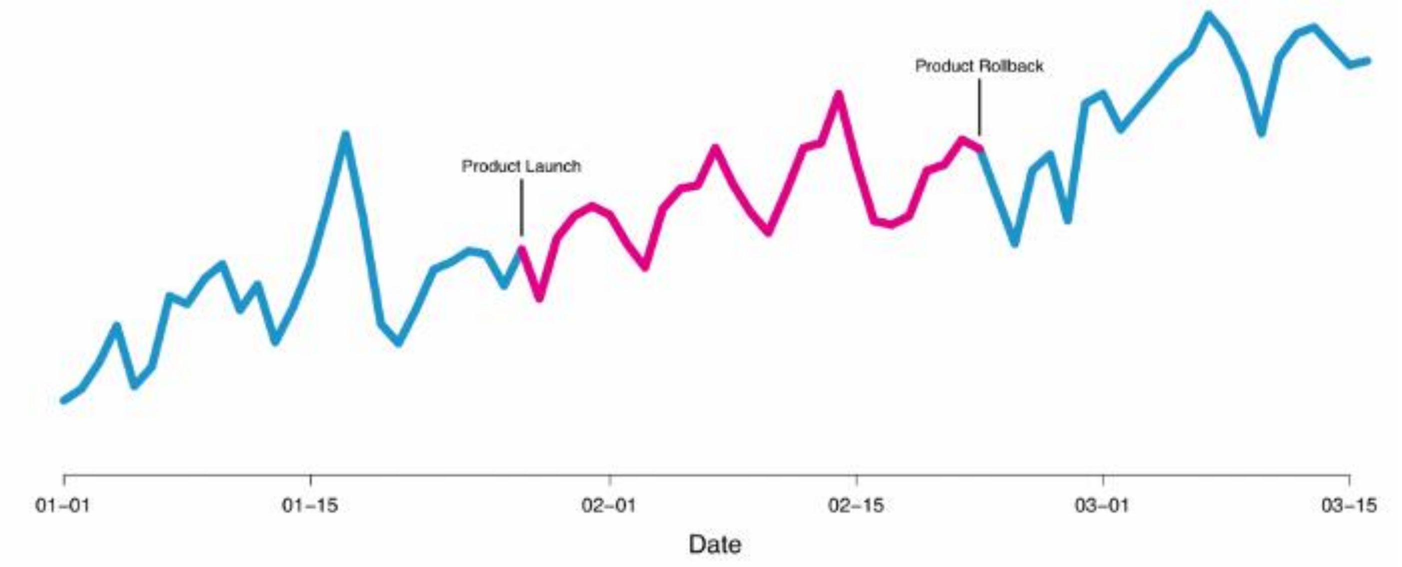
A solution: what about time?

- So far, we’ve thought about partitioning user clusters (often geographically correlated), or literally partitioning space (New Zealand; listings in Palo Alto)
- This is problematic when there isn’t enough unique space clusters
- Time to the rescue! Allocate the *same* set of users (same city, same region of space…) to treatment or control, at different times
- Most naïve: allocate entire city to control up to time T, and then entire city to treatment after that, to time $2T$
  - Compare your metric from the control and treatment periods

Challenge with naïve solution: time-varying marketplace

“The outside world often has a much larger effect on metrics than product changes do” – AirBnb, (Jan Overgoor) [Experiments at Airbnb | by AirbnbEng | The Airbnb Tech Blog | Medium]

If you compare the control period (earlier), to the experiment period (later), are changes because of the product or because of underlying marketwide changes, like seasonality?



Switchbacks

- For each region (city, graph cluster, neighborhood, etc), simply *switch back and forth* on whether that region is assigned to treatment or control
- For each unit of space-time, randomly assign it treatment or control
- Hope: that different units of space-time don't interfere with one another

  Then, analyze like you do a simple A/B test or graph cluster randomization test
- Sometimes interference still happens; need to deal with that in analysis

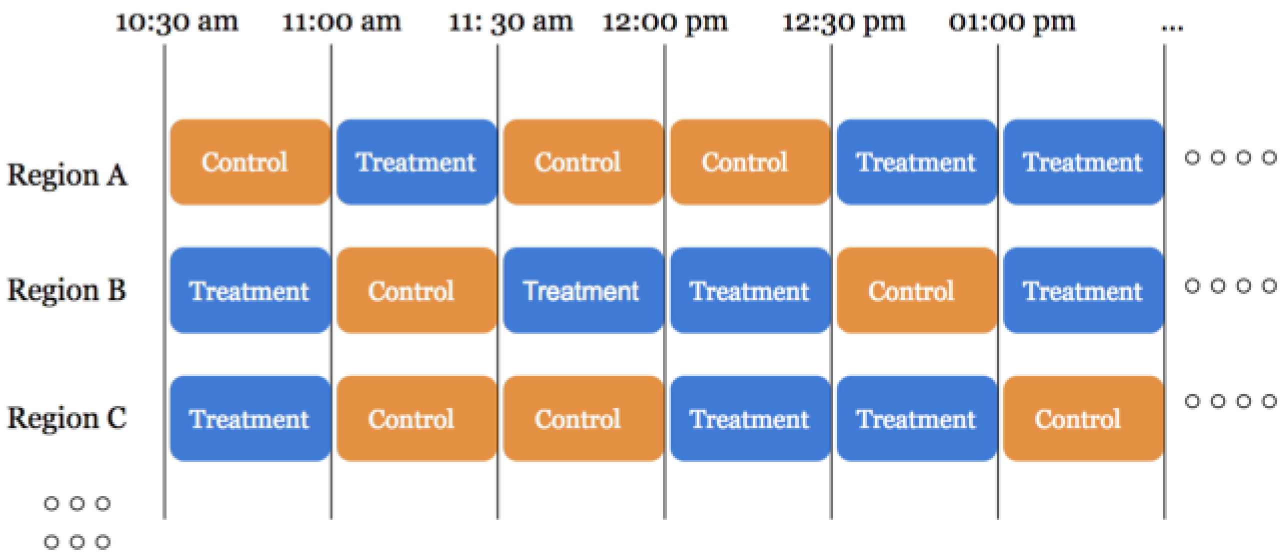


Image credit: [Switchback Tests and Randomized Experimentation Under Network Effects at DoorDash | by DoorDash | Medium] (David Kastelman, Data Scientist & Raghav Ramesh, Machine Learning Engineer)

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

Reminder 1: Bias-variance trade-off

- Bias-variance trade-off:
  - Smaller clusters (units) => more likely to interfere => more *bias*
  - Bigger clusters (units) => fewer clusters (units) => more *variance*
- What does each mean?

  Variance: If you run multiple experiments, each gives you a different answer

  Bias: If you run multiple experiments: each gives you the same wrong answer

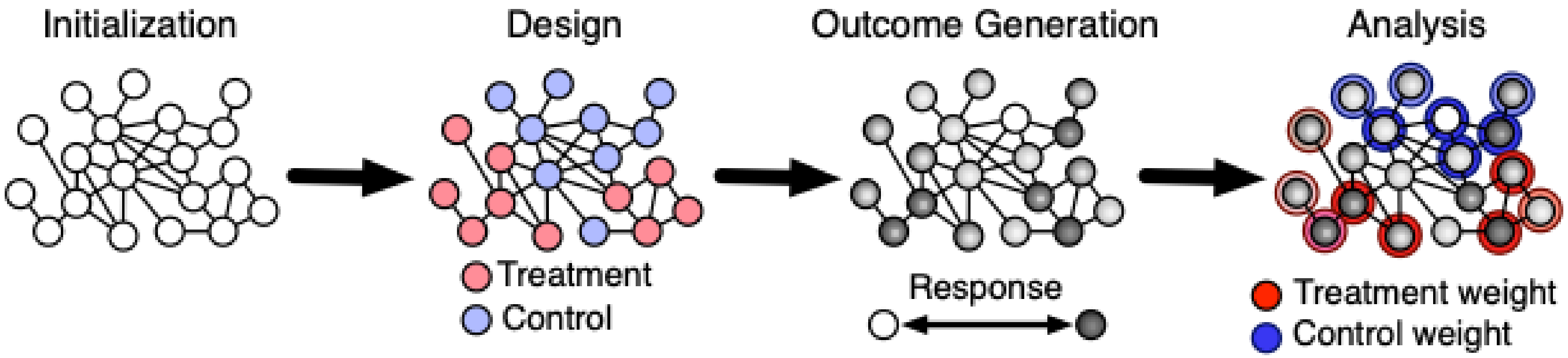
Randomization unit	Bias axis	Variance axis
User sessions		
Users		
Fine spatial units (geohash)		
Time interval (hour)		
Coarse spatial units (city)		

Table 1. Different choices of experimental units correspond to different points on the bias-variance tradeoff spectrum. In the context of network experiments, bias comes from interference effects; variance comes from decreasing unit set cardinality, and from between-unit heterogeneity.

[Experimentation in a Ridesharing Marketplace | by Nicholas Chamandy | Lyft Engineering]

Reminder 2: Design & Analysis

Two parts of running a good experiment: design and analysis



Design: Who gets assigned to treatment, who gets assigned to control

Analysis: Given the assignments and metrics for each unit, how do we calculate the Global Treatment Effect?

We have focused on design: **good design** simplifies analysis, **bad design** makes analysis impossible

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
- Next time: “synthetic control”

  Launch in just a few cities. Then, create a model for how that city would have behaved without the treatment, based on other how control cities actually behaved.

Announcements

- HW4 released; due next week
- Quiz 4 next week
- Project details released in next week
    - Project partner form on EdStem
- OHs
    - Mine today – 2-3 (In Person + Zoom)
    - Mine Friday (1 – 1:45, zoom only)
    - Zhi Friday (1:30 – 2:30, same zoom link as mine)
    - Zhi Monday
    - No Wednesday office hours next week