A/B Testing II: Cases

THE INTELLIGENT FUTURE

Mini 4 / Spring 2024

Carnegie Mellon University
Tepper School of Business

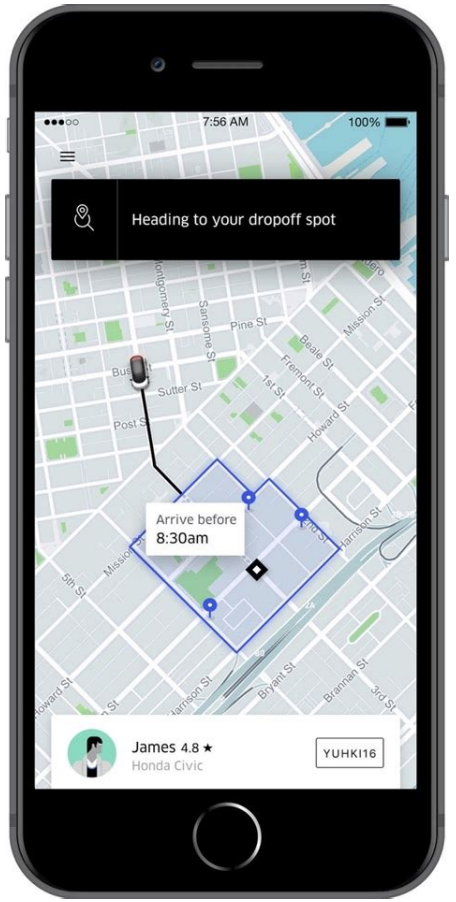
UBER POOL CASE

LEARNING OBJECTIVES

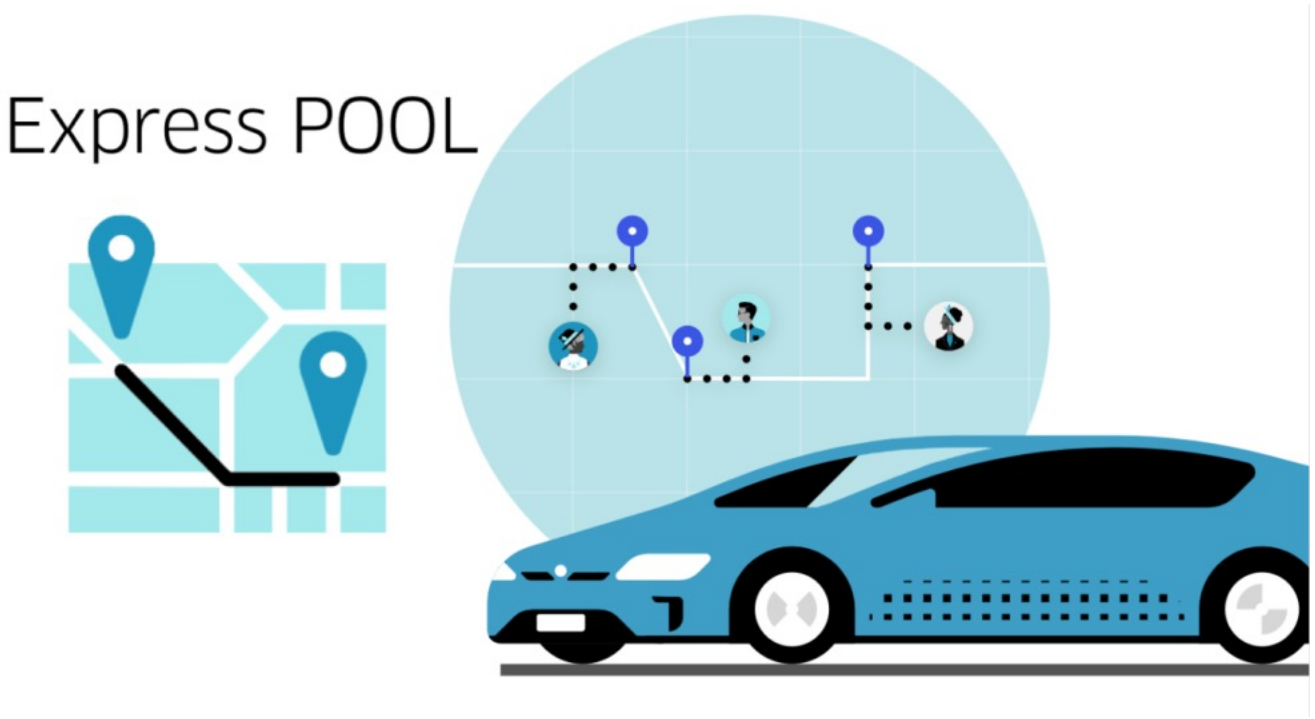
- **Product Development Process / Feature Rollout at Uber**
  - Can take a long time!
  - Lots of testing and validation pre-launch (conjoint, market simulation, etc.)

- **(Correct) Randomization into A and B is critical**
  - Stable Unit Treatment Value Assumption (SUTVA)
  - Marketplace A/B Testing
  - Tradeoffs to User, Time, and Market-level randomization

- **Hypotheses Definition, Metrics, and Experimentation (A/B Testing)**
  - Critical to define in “pre-launch” using best estimates from other pre-launch methods (qualitative surveys, quantitative methods like Conjoint Analysis)



Topic Presentation: Team 5 (Section A) & Team 4 (Section E)

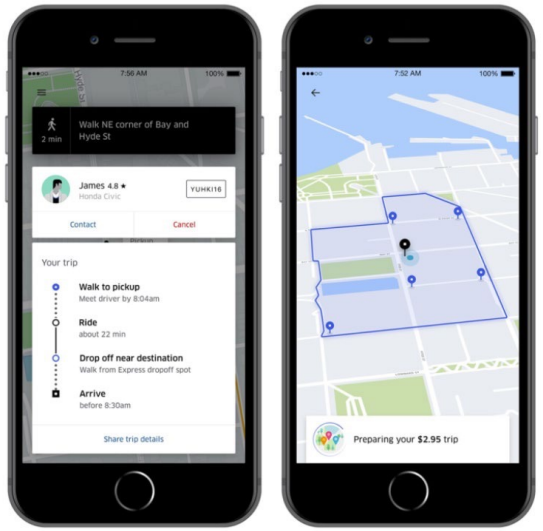


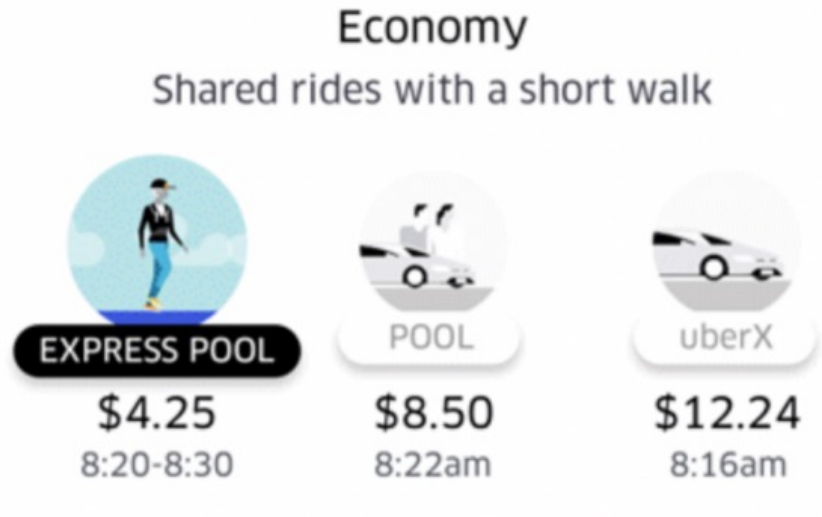
- Product launch using experiments

WHAT IS UBER EXPRESS POOL?

- **Uber**
  - UberCab (2009)
  - Uber (2010)
  - Uber POOL (2014)

- **Ride-sharing service**
  - 2-sided marketplace
  - **Question:** are there high or low barriers to customers (drivers) or customers (riders) switching to another product (e.g., Lyft)?





QUESTIONS FOR US (INITIAL)

- **Question 1: What was key "value" proposition of Uber Express POOL for customers (riders and drivers)? For Uber?**
  - **What are expected tradeoffs for customers (e.g., wait time with ___)?**

Uber Express POOL (2018)

"riders wait a few minutes before their trips begin, and then walk a short distance to a nearby spot for pick up and drop off"

	Type	Launched	Description
Carpool Options	Express	2018	Matches riders going in the same direction; Requires riders to walk to/from their pick-up and drop-off points and wait a few minutes to be matched
	Pool	2014	Matches riders going in the same direction; Offers door-to-door rides with no walking or waiting
Economy Options	UberX	2012	Provides private, affordable rides for 1 to 4 people; Uber's core economy product
	UberXL	2014	Provides private, affordable rides for up to 6 people
	UberSelect	2015	Provides private rides for 1 to 4 people with a driver who has been consistently highly rated
Premium Options	UberBLACK	2010	Uber's original ride option; Provides private rides in high-end black cars with professional drivers for 1 to 4 people
	UberSUV	2015	Provides private rides in luxury SUVs for up to 6 people
	UberLUX	2014	Uber's most luxurious option; Provides private rides in high-end cars for 1 to 4 people

Source: Case Study

QUESTIONS FOR US (INITIAL)

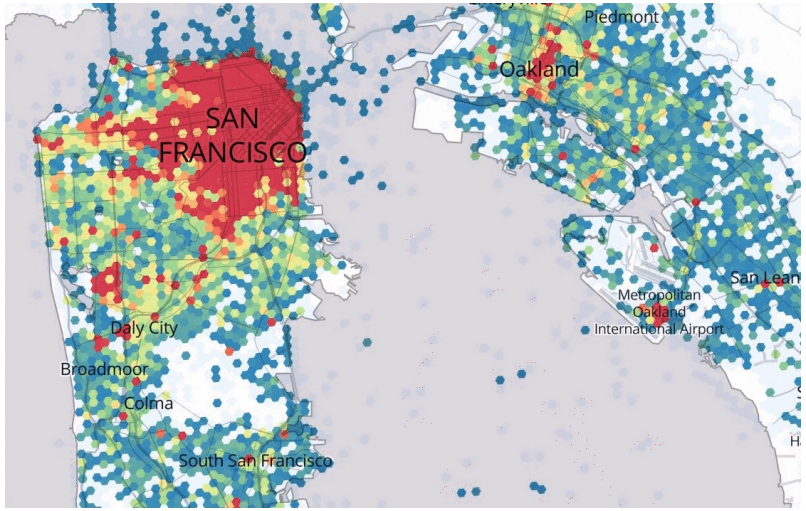
- **Question 1: What did the product development process / feature rollout look like?**
  - **How long did it take? Why did it take that long?**

- **Question 2: What was the role of "conjoint analysis", "demand forecasting", and "market simulation?**

Pre-Launch Consumer Needs and Product Feature Definition (Early 2017)	**Pre-Launch** Conjoint Analysis, Demand Forecasting, and Market Simulation (Sept. 2017)	**Post-Launch (Pilot 1)** SF / Boston (Nov. 2017)	**Post-Launch (Pilot 2)** 12 City Launch 6 Treatment/6 Control (Feb, 2018)

IN-CLASS DISCUSSION: RANDOMIZATION AND A/B TESTING

- **Q1: Why didn't Uber launch Express POOL and "simply" randomize users into A and B?** (i.e., randomly assign users that use the app to either A or B?)
  - **For example, why not launch Express to 50% of users in a city and leave remaining 50% with existing product offerings to see effects?"**



IN-CLASS DISCUSSION: RANDOMIZATION AND A/B TESTING

- **What were the 3 types of randomization Uber used?**

- **From the case: 3 types of randomization that Uber used.**
  - User level
  - Time level (called “switchbacks” in Case Study)
  - Market level (called “synthetic controls” in Case Study)

- **Q2: What are the pros and cons of each randomization? What would you do if you were a PM at Uber?**

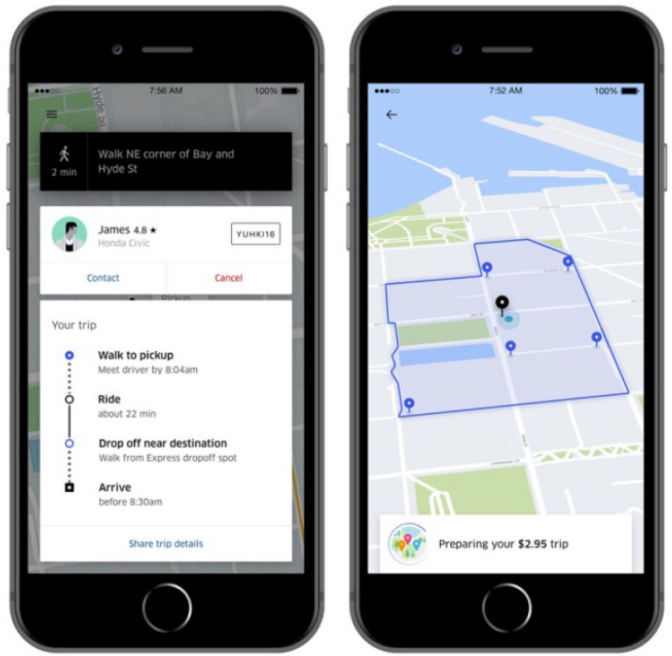
IN-CLASS DISCUSSION: RANDOMIZATION AND A/B TESTING

- **From the case:** Uber hoping to test the effect of **2 minute vs 5 minute** wait.

- **Q3. Hypotheses Definition:** How this might affect (1) Customer Experience and (2) Uber matching efficiency?
  - What are your hypotheses?
  - What are the tradeoffs for consumers (riders and drivers) and Uber?

In-class discussion: randomization and A/B testing

- **What metrics was Uber tracking and A/B testing?**
- **Trip Metrics**
  - Ridesharing Trips (Overall)
  - Uber POOL Trips
  - Uber Express POOL Trips
  - Cancellations
- **Revenue Metrics**
  - Cost per trip
- **Matching Metrics**
  - Match rate
  - Double Match Rate (Question: why?)
- **Q4:** What other Metrics might you consider?



IN-CLASS DISCUSSION: RANDOMIZATION AND A/B TESTING

- **Q1:** Why didn't Uber launch Express POOL and "simply" randomize users into A and B? (i.e., randomly assign users that use the app to either A or B?)

- **Q2:** What are the pros and cons of each randomization (A/B, switchbacks, synthetic control)? What would you do if you were a PM at Uber?

- **Q3.** What are your hypotheses? What are the tradeoffs for consumers (riders and drivers) and Uber?

- **Q4:** What other Metrics might you consider?

Q2: What are the pros and cons of each randomization?

		Time Randomization	Market / City Randomization
Experimental Considerations	User A/B	Switch backs	Synthetic Controls
1. Ability to reduce contamination between treatment & control groups			
2. Ability to measure effects on market equilibrium (e.g. average prices and total matches)			
3. Number of observations			
4. Ability to detect small effects with precision			
5. Experiment duration			
6. Number of experiments that can be run concurrently			
7. Ability to perfectly randomize			
8. Economic costs (potential losses)			

Q3. What are your hypotheses? What are the tradeoffs for consumers (riders and drivers) and Uber?

Q4. Metrics?

Customer Experience Metrics	Efficiency Metrics

- Cancellations. Substitute towards
  - POOL
  - UberX
  - Lyft
  - Other means of transportation
- Total Shared trips:
  - Express POOL trips
  - POOL trips
- Ratings
- Complaints
	- Cost per trip
	- Earning per trip
	- Match rate
	- Double match rate
	- Seat utilization during trips

Question for us: Would you increase wait times from 2 to 5 minutes?

- Would you increase wait times from 2 to 5 minutes in the 6 cities (treatment) in market-level randomization? Why or why not?

Question for us: Would you increase wait times from 2 to 5 minutes?

- Would you increase wait times from 2 to 5 minutes in the 6 cities (treatment) in market-level randomization? Why or why not?

- **Uber's decision:** do not increase from 2-5 minutes in the 6 cities

- This turned out to be a particularly smart decision ex-post,
  - In the winter of 2018, several snowstorms pummeled the northeast U.S., which affected overall ridership so significantly in some large cities that Uber was forced to discard much of the data collected during the synthetic control experiment, thereby decreasing the number of observations available and thus precision of the comparison between treatment and control cities.

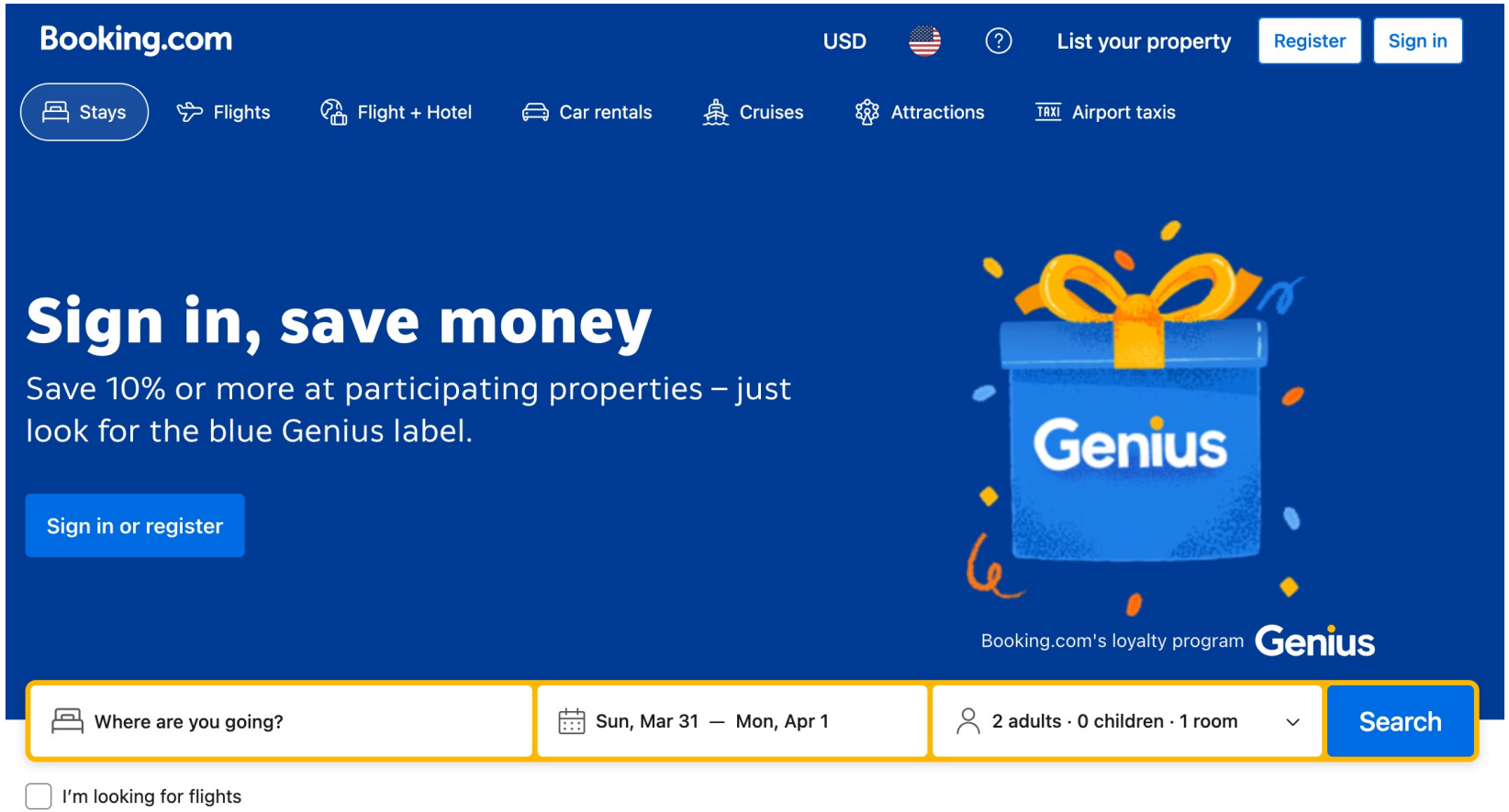
Booking.com case:

How to experiment at large scale

Importance of culture and management

BOOKING.COM EXPERIMENT

- How to optimize touch points with customers



CONTEXT

- How many of you have used Booking.com? Why?
- How do you use their landing page? For booking? Research on prices and/or reviews?
- Do you check other platforms? Do you book directly with hotels? Why or why not?
- Did anybody know that all users, including you, participate in Booking's experiments? If so, how?

CONTEXT

- A competitive advantage: A/B testing at scale
  - E.g., 1000 concurrent experiments, more than 25,000 tests a year

- Key metric: user conversion (measured by bookings per day) & post-booking behaviors
- Hypotheses developed from the user experience lab and the customer service department.
- In-house experimentation platform to execute the tests and to search for past experiments
- Most experiments get null results (but it doesn't mean failure!). E.g., quality of wifi in a hotel room

IN-CLASS DISCUSSION: A/B TESTING

- **Q1:** The design director wants to run the radical landing page experiment on millions of customers - people like you. Should the CEO get involved in the test? Why or why not?

- If so, what are the risks of a CEO micro-managing the design of an experiment?
- If not, how would you as a CEO explain the experiment to a frustrated customers, who is trying to book travel? How would you, as a customer, react to the landing page?

IN-CLASS DISCUSSION: A/B TESTING

- **Q2:** Is the Blue Screen landing page a good experiment? Why or why not? Consider the objective/intent; variables the experiment tries to change; and the financial risk associated with the experiment.

IN-CLASS DISCUSSION: A/B TESTING

- **Q3:** Booking's large-scale experimentation engine clearly gives it a competitive advantage and makes the company very profitable. Why don't competitors just copy the same model? Why is this so hard in most companies?

IN-CLASS DISCUSSION: A/B TESTING

- **Q4:** The case doesn't tell you how many experiments Booking runs annually. But please estimate the scale using Little's Law (Average Output rate = Average Work-in-process / Average Lead time). Does the scale change your recommendation for the CEO and the Blue Screen experiment?

Next Week

- Topic 1: A/B Testing 3
  - Guest speaker: **Sachal Lakhavani (AI Product Manager @ Meta; Tepper MBA alum)**
  - Everyone is expected to prepare for at least one question. (I may cold call!)
  - 30 min Q&A about A/B testing & Product management (Section A: 5-5:30pm, Section E: 6:30-7pm)
  - 15 min 1:1 Q&A (Section A: 5:30-5:45pm, Section E: 7-7:15pm)
- Topic 2: Product Strategy (Acquisition, Retention, Growth)
  - Pre-class reading: Choosing the Right Platform for Your Brand
  - Case assignment: Blue Apron

Upcoming Guests

- Mon, April 8
  - Guest speaker: **Sachal Lakhavani (AI Product Manager @ Meta; Tepper MBA alum)**
  - 30 min Q&A about A/B testing & Product management (Section A: 5-5:30pm, Section E: 6:30-7pm)
  - 15 min 1:1 Q&A (Section A: 5:30-5:45pm, Section E: 7-7:15pm)
- Mon, April 15
  - Guest speaker: **Michael Degnan (Senior VP, Head of Enterprise Innovation @ PNC)**
  - 1 hour lecture on New Product Development & Innovation
  - 30 min Q&A about Innovation & Product management