# ORIE 5355: People, Data, & Systems

## Lecture 4: Other topics in Data Collection

Nikhil Garg

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

# Questions from last time?

# Plan for today

- Unmeasured confounding and quantifying uncertainty
- Data collection case studies beyond polling
  Ratings + Recommendations
- Other topics in data collection
  - Differential privacy
  - Bias
  - Eliciting complex opinions
  - Modeling opinion dynamics
- Module summary + questions

# Unmeasured confounding and quantifying uncertainty

# The challenge

- In the last lecture, weighting helped us deal with *measured* selection bias/differential non-response

  Response rates and political opinions both correlate with educational status;

  (1) Education status can be asked for during the poll

  (2) We can roughly guess at voter distribution by education status
- What if response rates & opinions depend on a covariate that we don't observe, or that we don't know the population distribution of?
- Very little we can do to recover "point-estimate" of population opinion
- However, we can *quantify the uncertainty* under *assumptions* on how bad the problem is

# Setup

- Suppose there is a (binary) covariate $u_j$ that correlates with both the opinion of interest $Y_j$ and whether people respond $A_j$.
- You don't observe $u_j$ for any individual $j$
- $u$ is the only unmeasured confounding: $A_j$ is uncorrelated with true opinion $Y_j$ given $u_j$
- You have an estimate $\hat{y}$ (raw average of responses)
- Idea: Make assumptions on "how bad" the unmeasured confounding can get to derive uncertainty regions for your estimate of interest.

# Notation and Insight

- True population fractions of $u$: $\mathrm{P}^1 = \mathrm{Pr}(u_j = 1), 1 - \mathrm{P}^1 = \mathrm{Pr}(u_j = 0)$
- Response fractions: $\tilde{\mathrm{P}}^\ell = \mathrm{Pr}(u_j = \ell \mid A_j = 1)$
- $\bar{y} \stackrel{\text{def}}{=} E[Y_j] = P^1 E[Y_j \mid u_j = 1] + (1 - P^1) E[Y_j \mid u_j = 0]$
- $\hat{y} \rightarrow E[Y_j \mid A_j = 1] = \tilde{\mathrm{P}}^1 E[Y_j \mid u_j = 1, A_j = 1]$
$$+ (1 - \tilde{P}^1) E[Y_j \mid u_j = 0, A_j = 1]$$
- Insight:

$$E[Y_j \mid u_j = \ell, A_j = 1] = E[Y_j \mid u_j = \ell]$$

"Conditional on what group the respondent belongs to, their opinion does not correlate with whether they respond" ← We assumed this on last slide!

# Quantifying uncertainty in math

$$\bar{y} = P^1 E[Y_j \mid u_j = 1] + (1 - P^1) E[Y_j \mid u_j = 0]$$
$$\hat{y} \rightarrow \tilde{P}^1 E[Y_j \mid u_j = 1] + (1 - \tilde{P}^1) E[Y_j \mid u_j = 0]$$

Rearrange:

$$\hat{y} \rightarrow \bar{y} + \left(\tilde{P}^1 - P^1\right) E[Y_j \mid u_j = 1] + \left(P^1 - \tilde{P}^1\right) E[Y_j \mid u_j = 0]$$
$$= \bar{y} + \left(\tilde{P}^1 - P^1\right) \left(E[Y_j \mid u_j = 1] - E[Y_j \mid u_j = 0]\right)$$

Then, make assumptions on *whether respond* and *opinion* differences to quantify how far $\hat{y}$ can be from $\bar{y}$

If *either* response fractions or opinions between groups are similar, effect of unmeasured confounding is small!

# Unmeasured confounding in ML

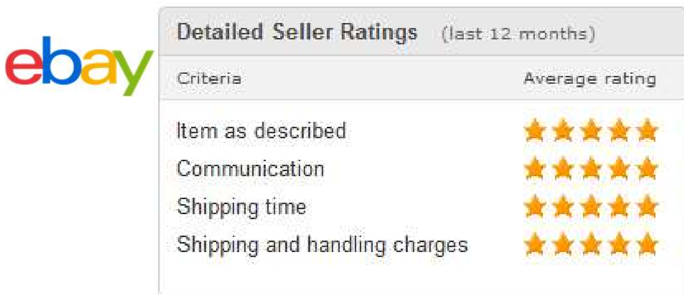
- In data science, we often care about *causal inference* (later in semester)
  - “What is the causal effect of going to a private high school on college success?”
  - Problem: In the US, private HS attendance correlated with parents’ wealth
- Unmeasured confounding (you don’t know parents’ wealth) would mess up your *inference* of the relationship in a regression
- You can also quantify unmeasured confounding and range of effects in such cases

# Case study: Ratings and recommendations

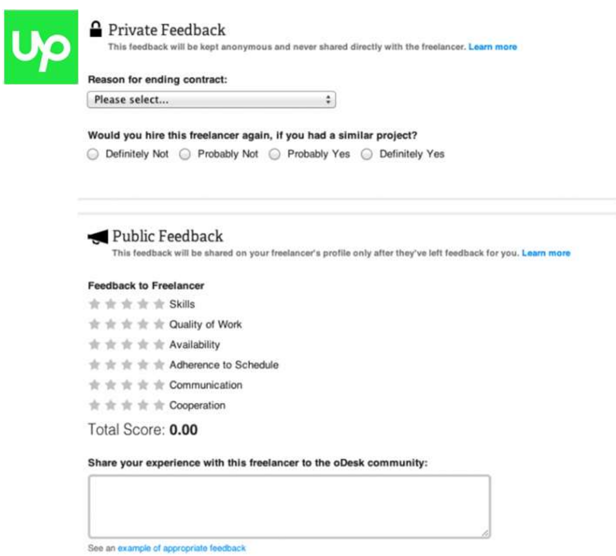
# Overview

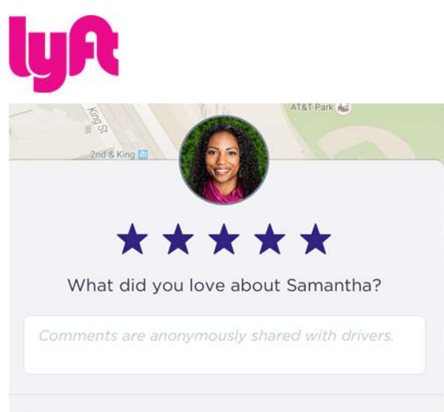
- So far, we’ve talked about explicit opinion collection in polling
- The same challenges apply in other settings
- Some differences
  - Often we don’t care about “absolute” opinion but “relative” opinions
  - We care a lot about “heterogeneous” opinions
  - We often have other “implicit” data on people’s opinions
- Briefly discuss some of these challenges in context of ratings and recommendations

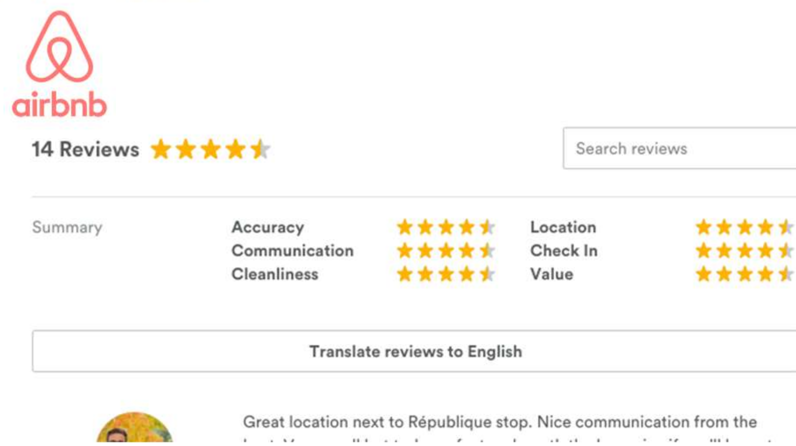
# Rating systems







# Measurement error: Ratings Inflation



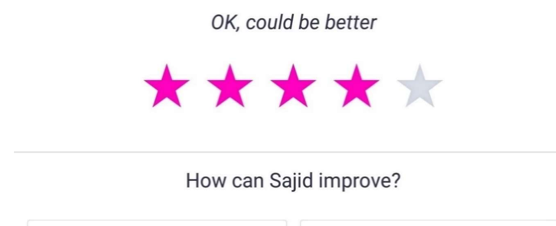
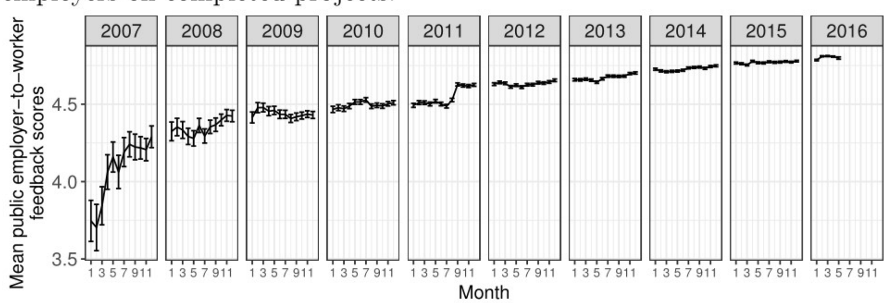



Figure 2: Monthly average public feedback scores assigned to workers by employers on completed projects.




[Filippas, Horton, Golden 2017]

**When 4.3 Stars Is Average: The Internet's Grade-Inflation Problem**

The Wall Street Journal April 5, 2017


https://xkcd.com/1098/

# Why ratings inflation & what to do about it?

- Many hypotheses for why ratings inflate
  - Explicit pressure from sellers – worry about retaliation
  - Implicit pressure – don’t want to hurt people’s livelihoods

  → Either misreport, or selection – less likely to report after bad experience
- Inflation is a type of measurement error:
  - The “quality” scale doesn’t match well to the “rating” scale
  - Inflation over time – mapping from quality to rating changes over time
  - Why does it matter? We ask you this in the homework
- What to do about it:
  - Try to reduce some of the pressure
  - Weighting to tackle selection: paper in the homework [Nosko & Tadelis]
  - Change the rating scale: [Garg and Johari]

# Ratings heterogeneity

- There is much ratings “heterogeneity”
  - Different people have different opinions on the same item
  - Different ‘categories’ of items might have different average ratings
- Why does this matter?
  - You want to give each person a personalized “rating” or recommendation
  - You want to compare items across categories
- What to do about it?
  - Personalized recommendations → starting next time
  - “Standardize” ratings across categories
  - Communicate to customers – e.g., “relative” ratings instead of “absolute” ones

# Implicit data collection in recommendations

- You have many implicit signals about people's opinions
  - Do they finish watching the show, or start watching the next episode?
  - Do they keep coming back and buying other things
  - Did they browse other items instead of putting something in their cart?
  - Do they re-hire the same freelancer/work with the same client again?
- These give *different* information than do explicit ratings
  - From a different population of users
  - Often more numerous, but harder to analyze
  - "revealed preference" – might be more predictive of future behavior
- Using such data
  - Train models to predict different future behavior, using various signals
  - Might take away "user agency" – what if they want to change their behavior?

# Miscellaneous topics in data and data collection

# (Differential) Privacy

- What if you’re asking about a sensitive attribute?

  For example, an insurance company wants to estimate the percentage of their policy holders who smoke

- Goal: collect data in a way such that you learn very little about any individual person, but you are accurate across population
- How? Add noise to each response
- Example: Tell each person, “roll a 6-sided dice. If it’s 1 or 2, lie about whether you smoke. Otherwise, tell the truth.” If fraction $Y$ people tell you that they smoke, then we know that the truth $X$ satisfies:

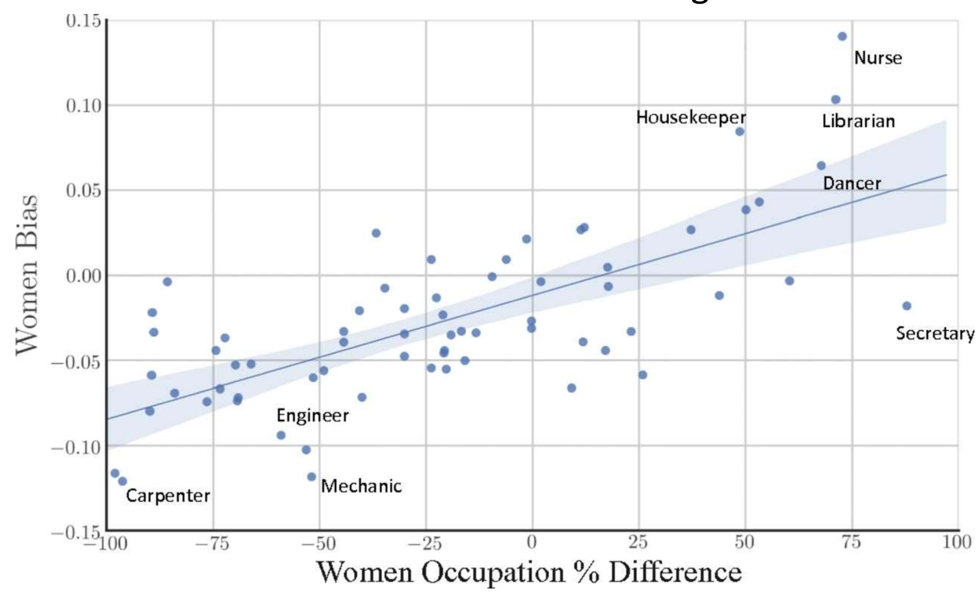
$$Y = \frac{4}{6}X + \frac{2}{6}(1 - X)$$

- Similar ideas used to collect and share data at Apple and the US Census

# Using biased data

- The world is full of historic inequities
  - Some neighborhoods are over-policed compared to others → data will have more "crimes there"
  - Every possible opinion expressed on forums like Reddit
  - Who succeeded at a university
- Models trained using this data will *reflect* and *amplify* these biases
- Many techniques to audit and mitigate such biases in models

Bias in word embeddings




"Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes" by Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou

# Eliciting complex opinions

- So far, we’ve talked about soliciting “low-dimensional” opinions
  - Binary opinions, or one of a small number of options
- What if we want to solicit opinions on complicated things?
  - How your town should spend $2M budget across parks, sports teams, art festivals, etc.
  - When should we schedule these five events over 10 time slots?
- You can’t ask people to rank every option
- Several standard techniques
  - Participatory budgeting
  - Pairwise comparisons
- More generally, many cool techniques in crowdsourcing

# Data dynamics

- The world is not static
  - Opinions change with external events
  - Your startup is growing and attracting new kinds of customers
  - Weekends are different than weekdays, except on holidays…
- Similar problem as “Problem 1” in survey weighting – if you don’t share data across time, then you don’t have enough data. But if you do share data, then suddenly your dataset differs from what you care about
- Techniques to model opinion dynamics – “smooth” over time
- Some related challenges covered in pricing module

# Module Summary

- Measurement error: The construct you care about is never perfectly captured by the data that you have
- Selection effects/differential non-response happens everywhere you're collecting opinions from people
- You can use stratification and weighting to mitigate selection effects *on known covariates*
- On unknown covariates, quantify uncertainty!

Never take opinion data at face value. Always ask:

(1) What did I measure, versus what did I care to measure?

(2) who answered versus what's the population of interest

# Announcements

- HW1 due Sunday evening
  - Don’t wait until the last minute!
  - We are unlikely to provide much help on EdStem over the weekend, but we will be active throughout the week
  - Go to office hours
- Guest lecture next Monday – please attend in person if possible

# Questions?