ORIE 5355: Applied Data Science - Decision-making beyond Prediction

Lecture 4: Weighting + Uncertainty

Nikhil Garg

Announcements

- HW1 due Tuesday 9/15
- Quiz 1 will also be that week

https://orie5355.github.io/Fall_2026/attendance/

Questions from last time?

Plan for today

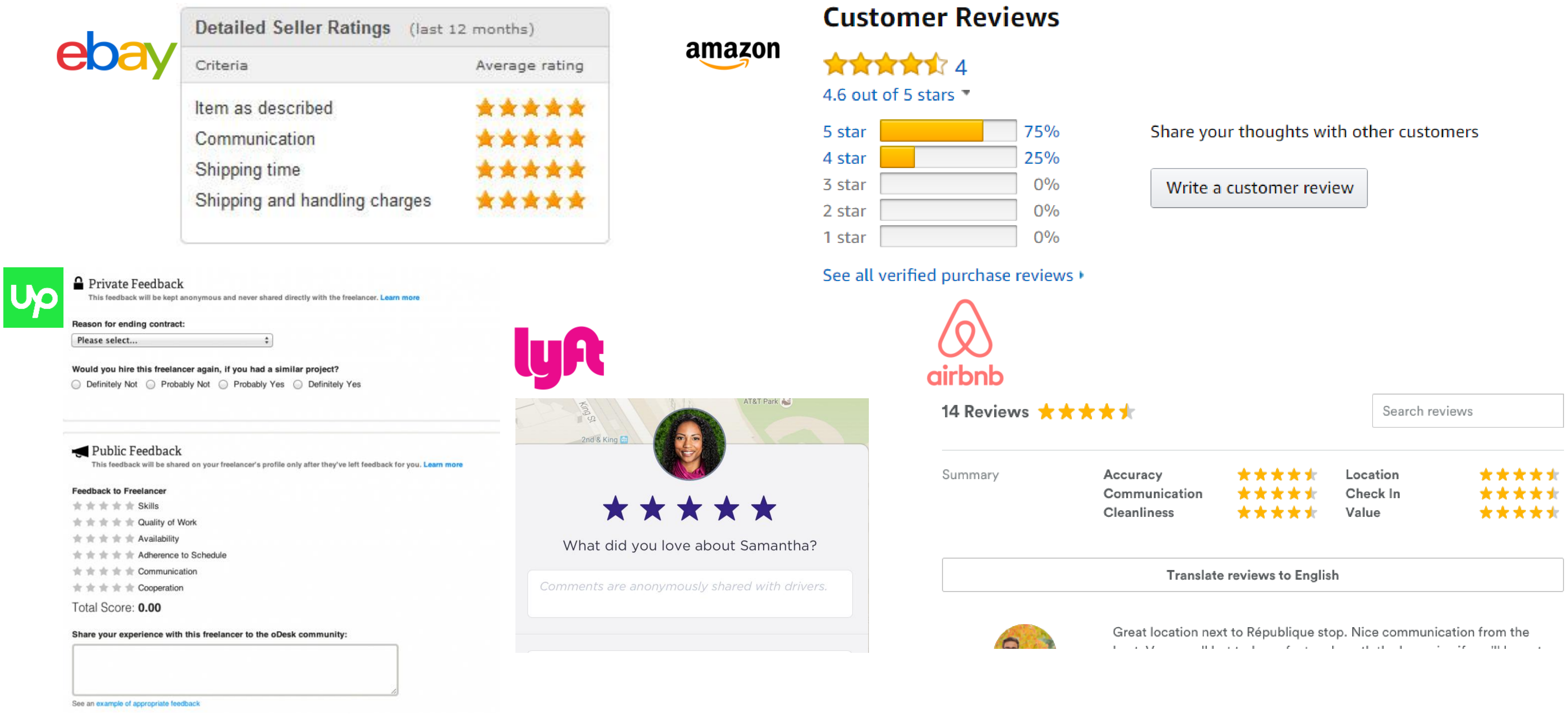
- Stratification
- Weighting techniques
- Quantifying uncertainty
- Other topics in data collection (time permitting)

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

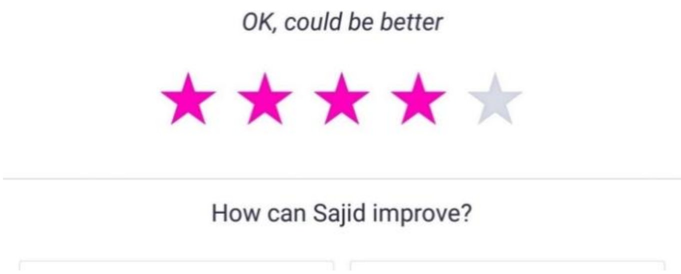
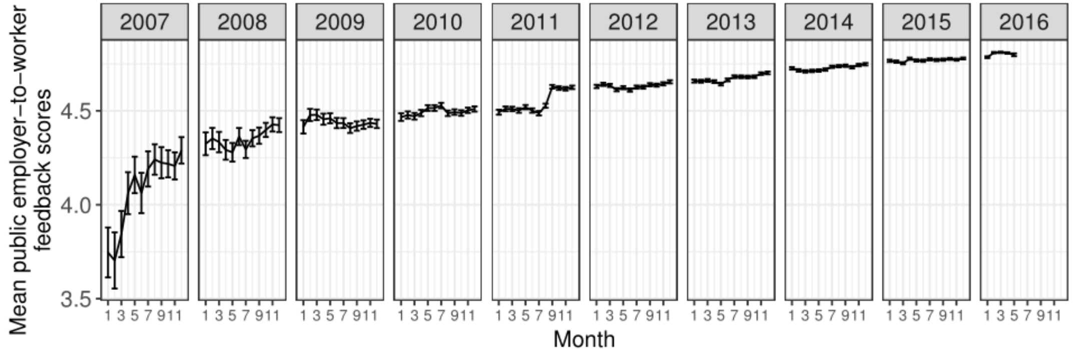


Figure 2: Monthly average public feedback scores assigned to workers by employers on completed projects.



[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]

Experiment Description

Status quo: Clients hire freelancers, rate them at contract end

Form includes a numeric rating from 0 to 10, with avg >8/10

Challenge: Can we induce different (non-inflated) ratings by changing the question we ask on the rating form?

Experiment design

- Add additional question to private portion of the form (6 treatments)

  Randomization at the *client* level

- Observe ratings for 3 months (180k jobs, 60k clients, 80k freelancers)

“Designing Informative Rating Systems: Evidence from an Online Labor Market” Nikhil Garg and Ramesh Johari

Treatment groups

Treatment	Question Phrasing	Answer choices
Numeric	How would you rate this freelancer overall?	0 – 5

Treatment groups

Treatment	Question Phrasing	Answer choices
Numeric	How would you rate this freelancer overall?	0 – 5
Adjectives	How would you rate this freelancer overall?	Terrible Mediocre Good Great Phenomenal Best possible freelancer!

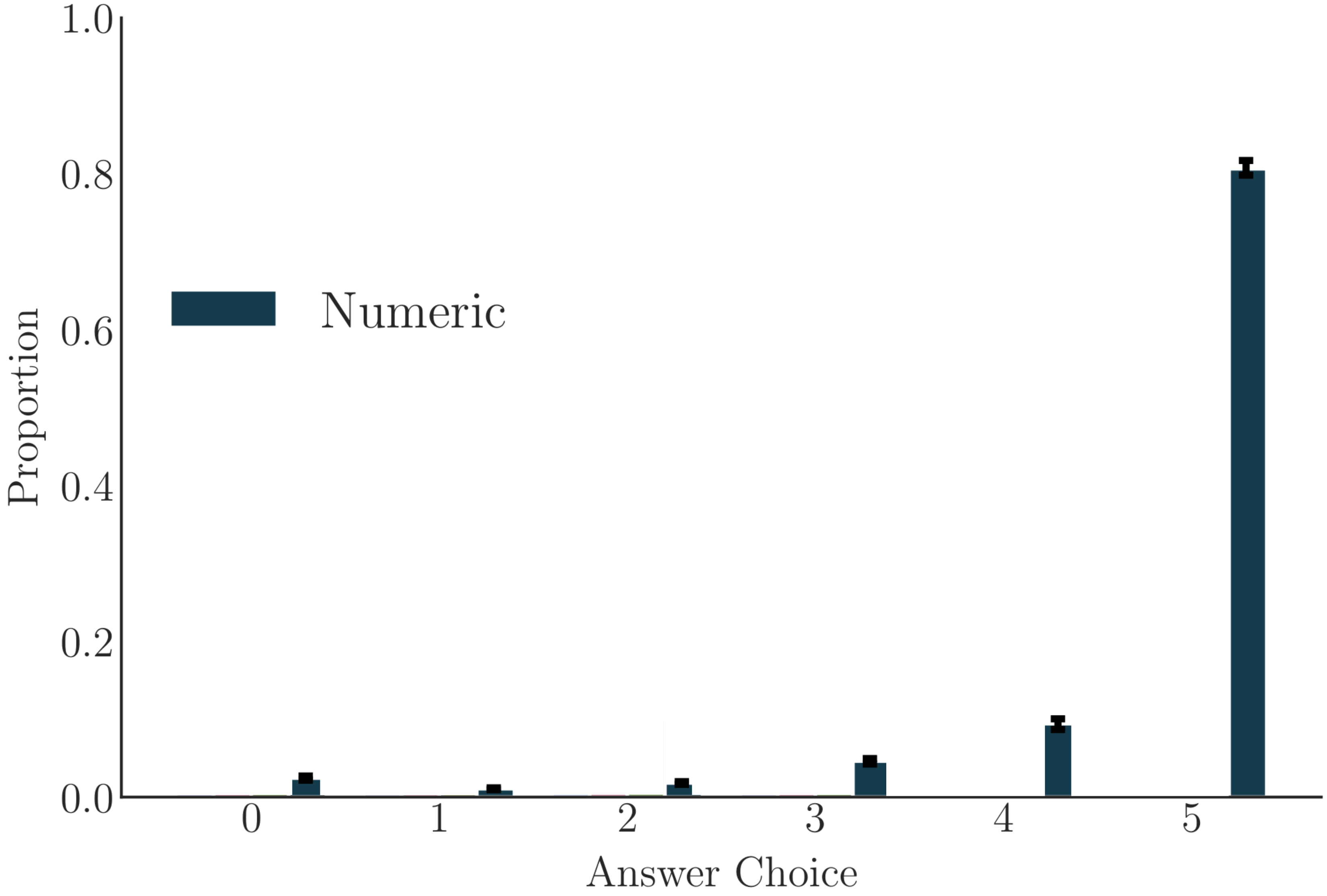
“Designing Informative Rating Systems: Evidence from an Online Labor Market” Nikhil Garg and Ramesh Johari

Treatment groups

Treatment	Question Phrasing	Answer choices
Numeric	How would you rate this freelancer overall?	0 – 5
Adjectives	How would you rate this freelancer overall?	Terrible Mediocre Good Great Phenomenal Best possible freelancer!
Expectations	How did this freelancer compare to your expectations?	Much worse than I expected … Beyond what I could have expected
Average	How does this freelancer compare to others you have hired?	Worst Freelancer I’ve Hired Below Average Average Above Average Well Above Average Best Freelancer I’ve hired
Average, random order		
Average, not affect score	How does this freelancer compare to others you have hired? (This will not impact the freelancer’s score)	

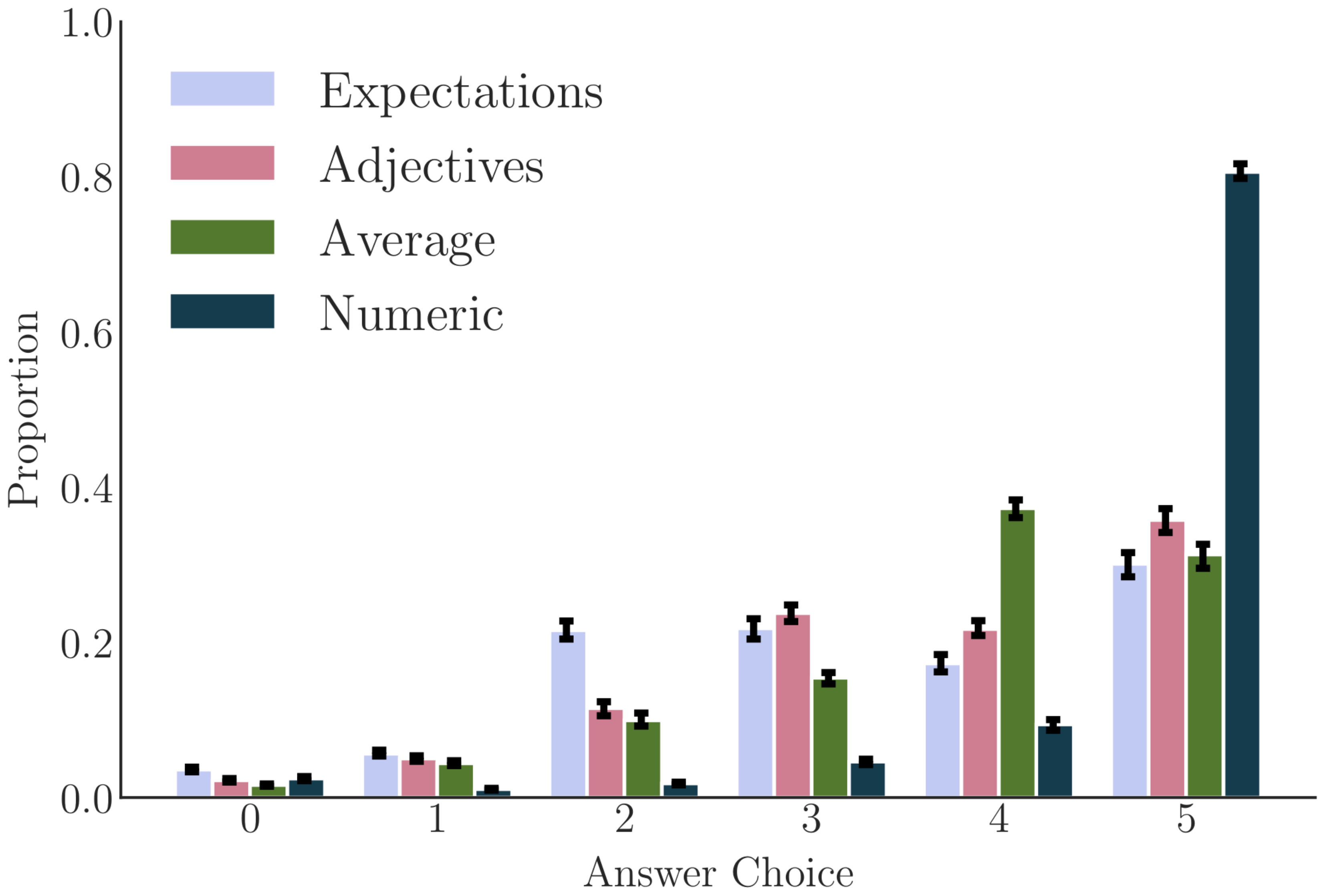
“Designing Informative Rating Systems: Evidence from an Online Labor Market” Nikhil Garg and Ramesh Johari

Result: marginal rating distributions



"Designing Informative Rating Systems: Evidence from an Online Labor Market"
Nikhil Garg and Ramesh Johari

Result: marginal rating distributions



"Designing Informative Rating Systems: Evidence from an Online Labor Market"
Nikhil Garg and Ramesh 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?

Case study 2: Crowdsourcing

Government service allocation

Local government manages many services

- ~8k miles of streets in NYC
- ~700k trees lining streets in NYC
- Housing, sanitation, transportation, etc.

Operational tasks

- [Learning] What problems are there?
- [Allocation] Which ones to address?
- [Auditing] Did we do a good job?

Desiderata: Efficiency & Equity

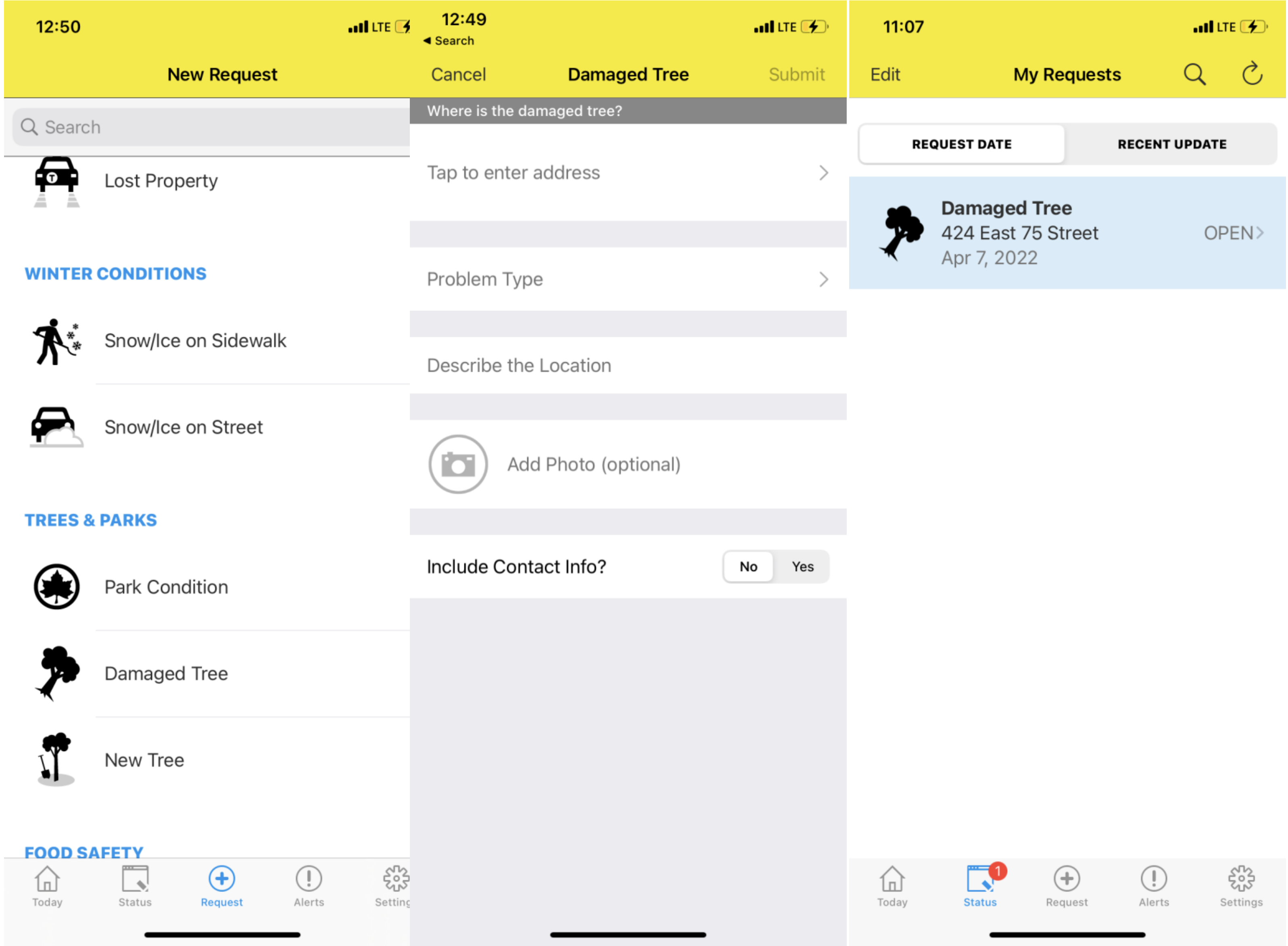
Street trees on Upper East Side in NYC

311 (crowdsourcing) systems

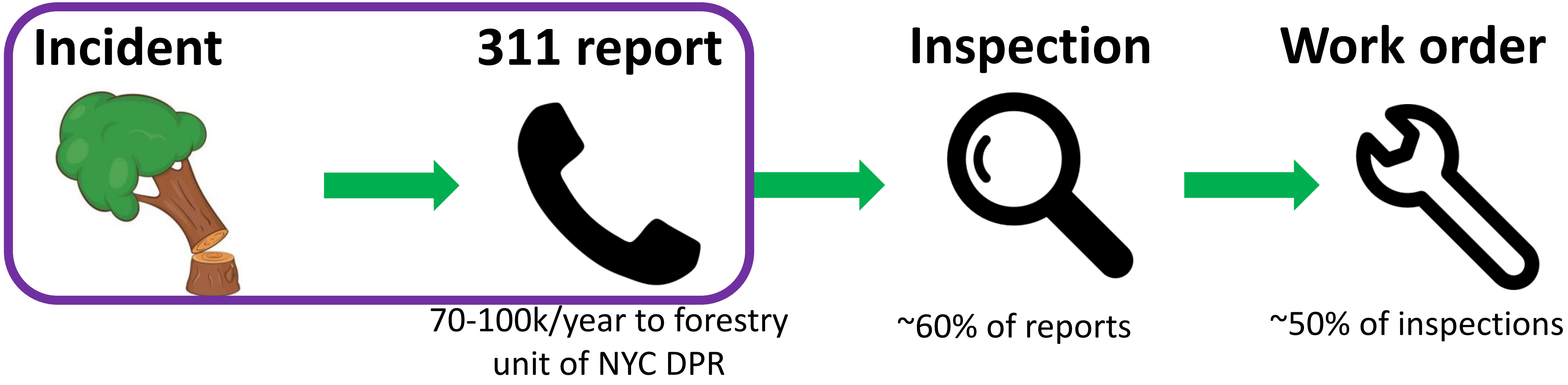
Cities have a phone number & app to complain to the local government

NYC’s 311 system received about 2.7 million requests 2021

These are the primary way the government learns about problems



Pipeline: from incident to work orders



Why is this hard? Uncertainty, heterogeneous + strategic behavior, distribution shifts over time, capacity constraints, pipelined decisions

Reporting behavior: If there are crowdsourcing differences (who reports what), then there will be downstream differences in decision-making

Possible data collection heterogeneity

Underreporting: The same number of problems exist in 2 neighborhoods, but one neighborhood reports more problems, faster.

Mis-reporting: Same types of problems in 2 neighborhoods, but people in one tend to exaggerate incident type/risk to get faster service.

In each case, we'll have disparities in what work gets done! (bad allocation of government services)!

Research agenda: How do we understand these reporting differences and then correct for them?

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

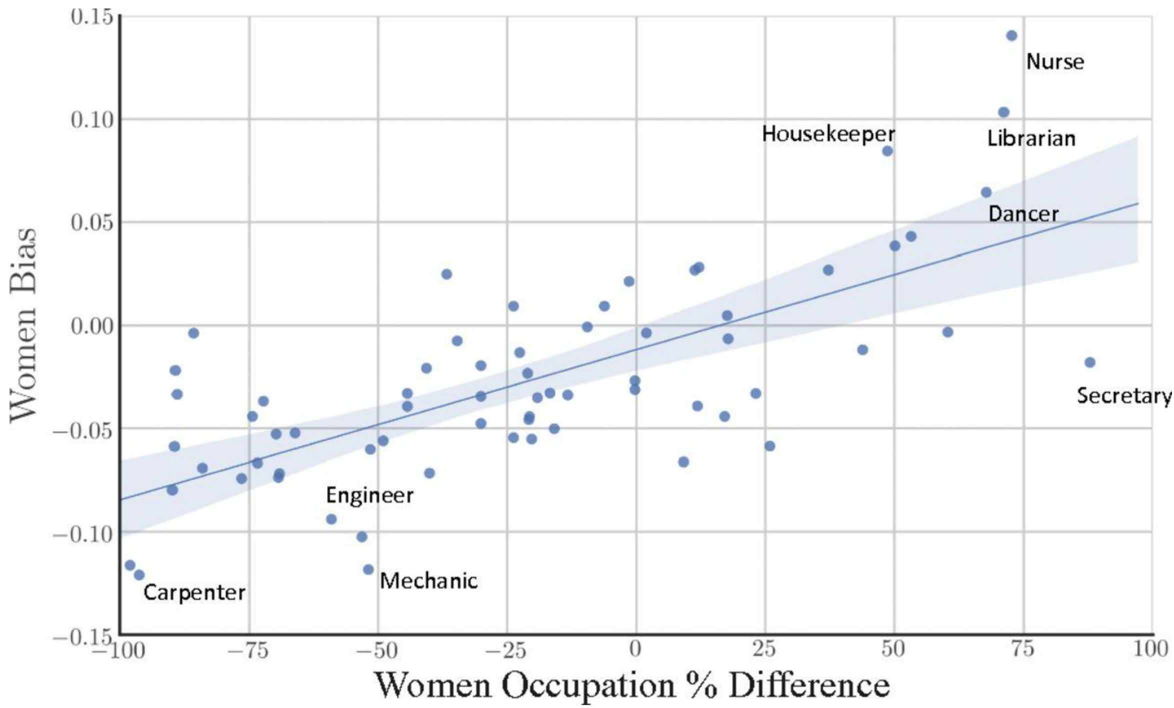
Eliciting complex opinions

- 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

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

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

(3) What am I going to *do* with the data, and how does that affect data collection?

Will show up in the rest of the course!

Questions?