Responsible Machine Learning

Lecture 7: Algorithm Auditing Overview

CS 4973-05

Fall 2023

Instructor: Jeffrey Gleason
gleason.je@northeastern.edu
Northeastern University, Boston, MA

Northeastern University

Agenda

1. What is algorithm auditing?
2. Case Study 1: Representation Bias on Google Images
3. Case Study 2: Rating/Ranking Bias on TaskRabbit and Fiverr
4. Design Brainstorm: TikTok

What is Algorithm Auditing?

History: Social Science Audits

- Social scientists and community organizers developed *empirical* methods to detect and measure housing discrimination in the 1970s [1]
    - Initial audits: participatory, accountability and reform oriented

[1] Vecchione, Briana, Karen Levy, and Solon Barocas. "Algorithmic auditing and social justice: Lessons from the history of audit studies." *Equity and Access in Algorithms, Mechanisms, and Optimization*. 2021. 1-9.

History: Social Science Audits

- Social scientists and community organizers developed *empirical* methods to detect and measure housing discrimination in the 1970s [1]
  - Initial audits: participatory, accountability and reform oriented
  - Over time: greater focus on statistical rigor (sample size, measurement precision)
  - Recent example: *“Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination”* [2]

[2] Bertrand, Marianne, and Sendhil Mullainathan. "Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination." *American economic review* 94, no. 4 (2004): 991-1013.

Auditing *Algorithms*

- Common Methods [3]:
    - Scraping Audit
        - What does the algorithm return in response to a specific input (e.g. search query)?

[3] Sandvig, Christian, et al. "Auditing algorithms: Research methods for detecting discrimination on internet platforms." *Data and discrimination: converting critical concerns into productive inquiry* 22.2014 (2014): 4349-4357.

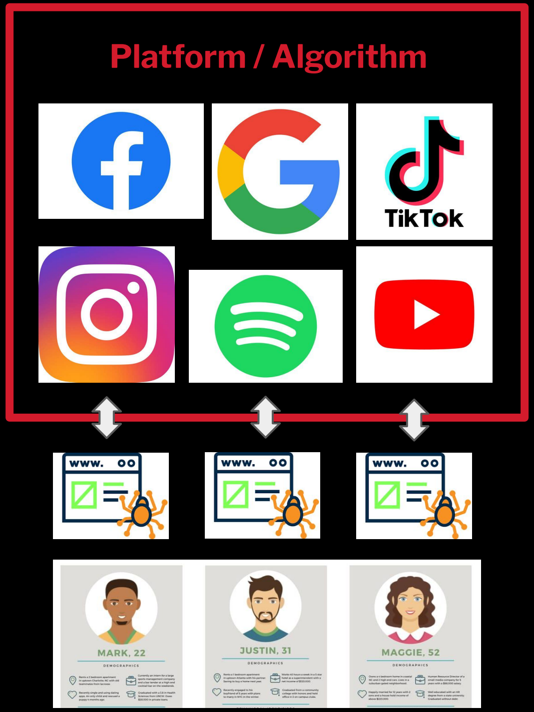
Auditing *Algorithms*

- Common Methods [3]:
  - Scraping Audit
    - What does the algorithm return in response to a specific input (e.g. search query)?

[3] Sandvig, Christian, et al. "Auditing algorithms: Research methods for detecting discrimination on internet platforms." *Data and discrimination: converting critical concerns into productive inquiry* 22.2014 (2014): 4349-4357.

Auditing *Algorithms*

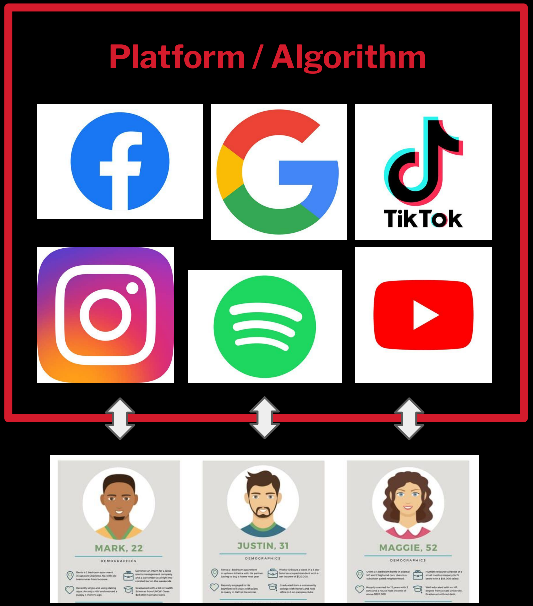
- Common Methods [3]:
    - Scraping Audit
    - Sock Puppet Audit
        - What does the algorithm return in response to a specific profile?



[3] Sandvig, Christian, et al. "Auditing algorithms: Research methods for detecting discrimination on internet platforms." *Data and discrimination: converting critical concerns into productive inquiry* 22.2014 (2014): 4349-4357.

Auditing *Algorithms*

- Common Methods [3]:
  - Scraping Audit
  - Sock Puppet Audit
  - Collaborative/Crowdsourced Audit
    - What does the algorithm return in response to real user behavior?



[3] Sandvig, Christian, et al. "Auditing algorithms: Research methods for detecting discrimination on internet platforms." *Data and discrimination: converting critical concerns into productive inquiry* 22.2014 (2014): 4349-4357.

Auditing *Algorithms*

- Common Methods [3]:
  - Scraping Audit
  - Sock Puppet Audit
  - Collaborative/Crowdsourced Audit
- External vs. Internal Auditing

Platform / Algorithm

TikTok

[3] Sandvig, Christian, et al. "Auditing algorithms: Research methods for detecting discrimination on internet platforms." *Data and discrimination: converting critical concerns into productive inquiry* 22.2014 (2014): 4349-4357.

Auditing *Algorithms*

- Common Methods [3]:
  - Scraping Audit
  - Sock Puppet Audit
  - Collaborative/Crowdsourced Audit
- External vs. Internal Auditing [4]
  - Tension: independence vs. access

Platform / Algorithm

[4] Raji, Inioluwa Deborah, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. "Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing." In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 33-44. 2020.

Case Study 1: An Image of Society: Gender and Racial Representation and Impact in Image Search Results for Occupations

Danaë Metaxa et al.

Research Questions

1. Do search results accurately represent the 2020 U.S. workforce in terms of representation of gender and race?

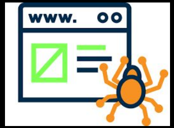
Research Questions

1. Do search results accurately represent the 2020 U.S. workforce in terms of representation of gender and race?
2. Does the representation of women and POC in search impact a participant's sense of belonging in an occupation?

Data Collection (RQ1)

1. Scrape Google Image search results

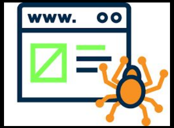
engineer
electrician
doctor
…
author
veterinarian
pilot



Data Collection (RQ1)

1. Scrape Google Image search results

engineer
electrician
doctor
…
author
veterinarian
pilot

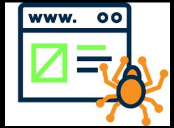


"engineer"

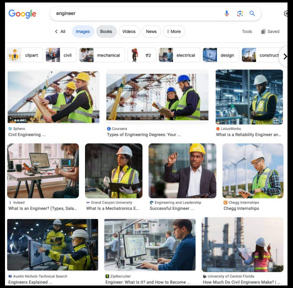
Data Collection (RQ1)

1. Scrape Google Image search results

engineer
electrician
doctor
…
author
veterinarian
pilot



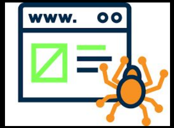
“engineer”



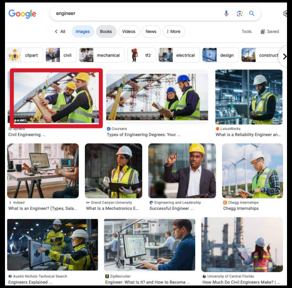
Data Collection (RQ1)

2. Hire crowd workers to label perceived race and gender

engineer
electrician
doctor
…
author
veterinarian
pilot



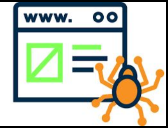
“engineer”



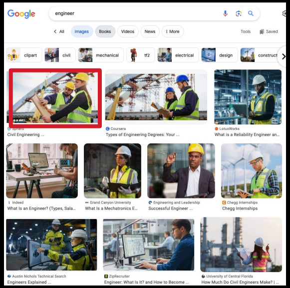
Data Collection (RQ1)

3. Compare perceived race and gender to BLS official statistics

engineer
electrician
doctor
…
author
veterinarian
pilot

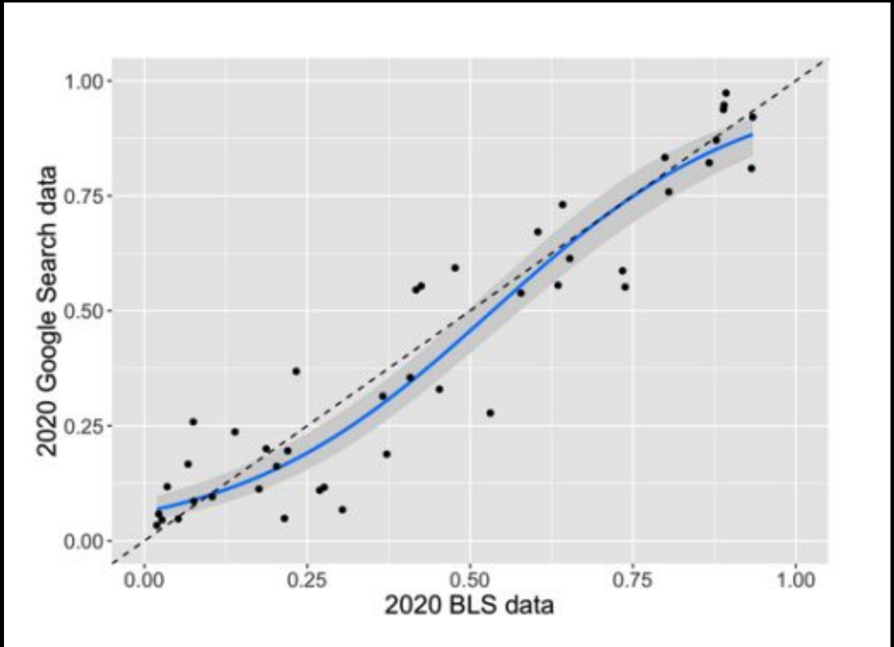


"engineer"



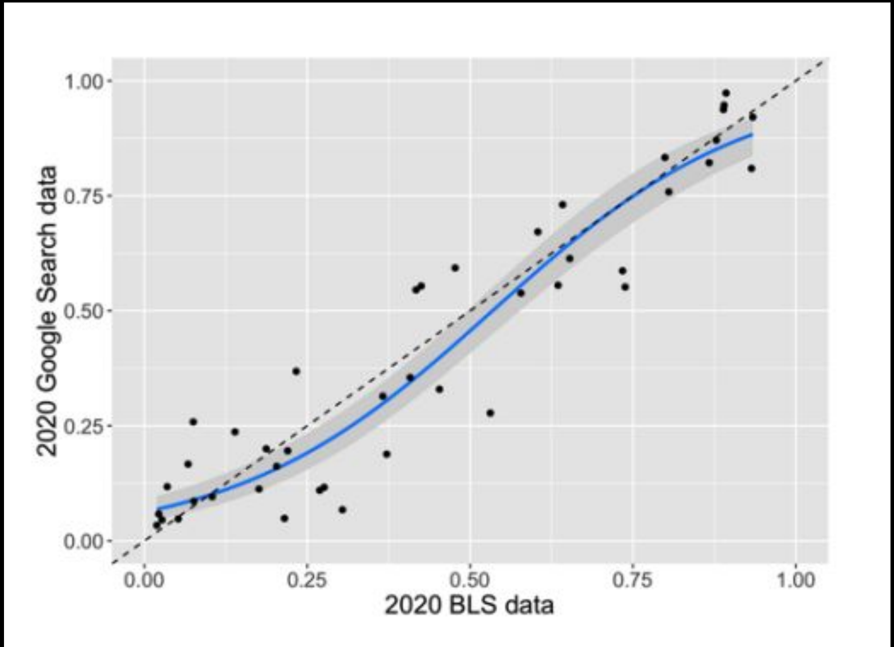
VS

Results: Representation (RQ1)

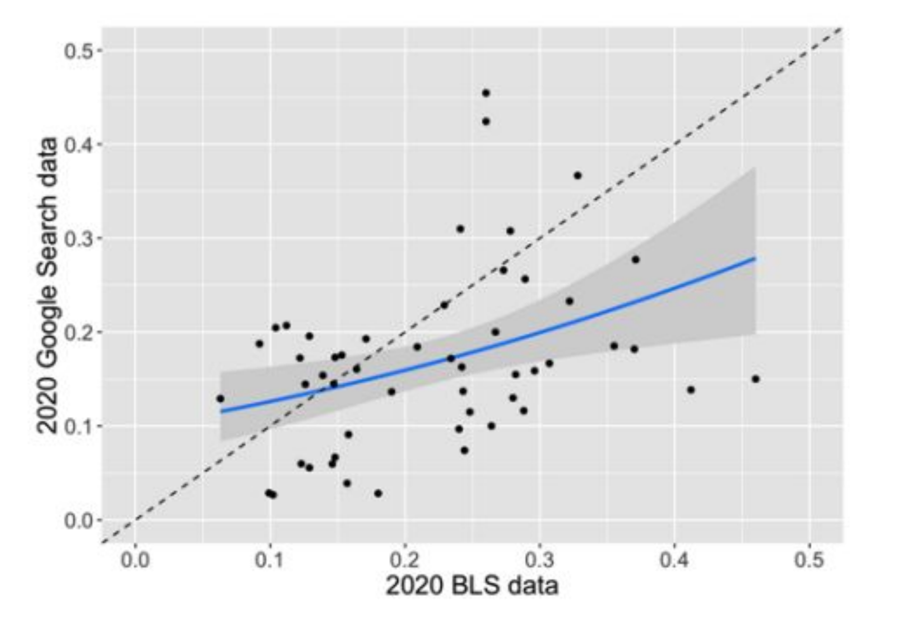


Representation of Women

Results: Representation (RQ1)

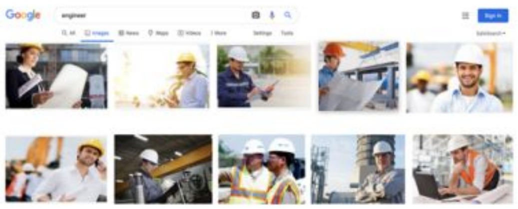


Representation of Women



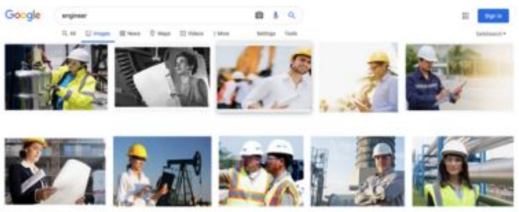
Representation of Non-white People

Data Collection (RQ2)



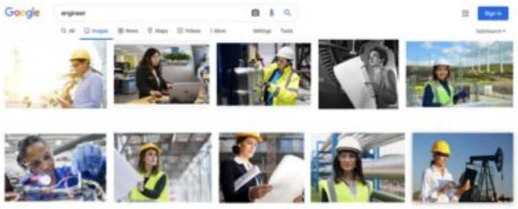
(a) 10% women

(b) 10% POC

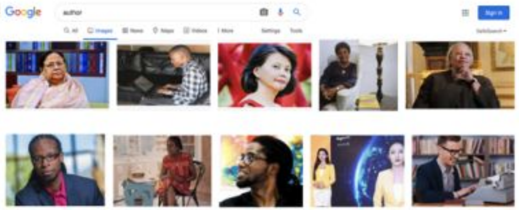


(c) 50% women

(d) 50% POC

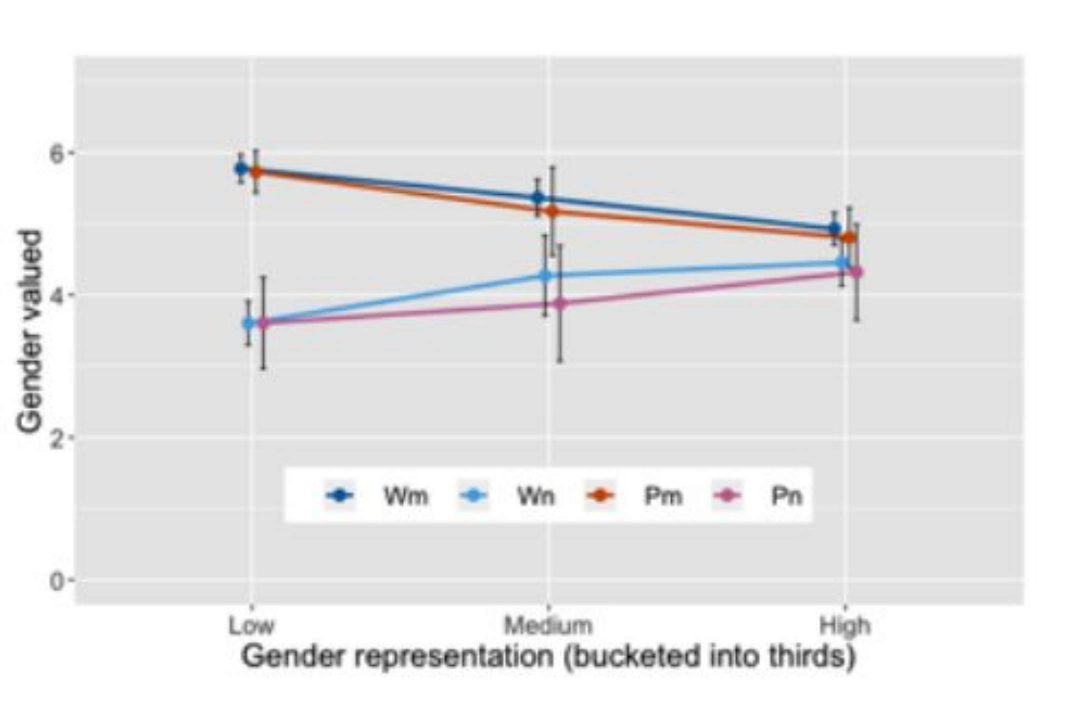


(e) 90% women



(f) 90% POC

Results: effects on feeling valued (RQ2)



Legend
1. Wm = White men
2. Wn = White non-men
3. Pm = POC men
4. Pn = POC non-men

Case Study 2: Bias in Online Freelance Marketplaces: Evidence from TaskRabbit and Fiverr

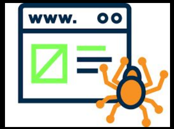
Anikó Hannák et al.

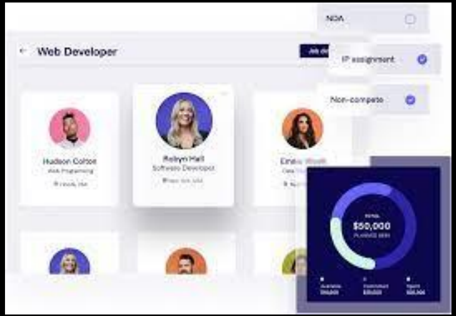
Research Questions

1. How do perceived gender, race, and other demographics influence the social feedback workers receive?
2. Do workers' perceived demographics correlate with their position in search results?

Data Collection

1. Scrape TaskRabbit and Fiverr

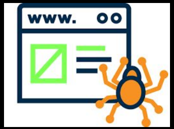




Collect:
1. Profile metadata
2. Profile picture: perceived demographics
3. Social feedback: ratings and reviews
4. Search result rank

Data Collection

1. Scrape TaskRabbit and Fiverr



Collect:
1. Profile metadata
2. Profile picture: perceived demographics
3. Social feedback: ratings and reviews
4. Search result rank

Website	Founded	# of Workers	# of Search Results	Unknown Demographics (%)	Gender (%) Female	Gender (%) Male	Race (%) White	Race (%) Black	Race (%) Asian
taskrabbit.com	2008	3,707	13,420	12%	42%	58%	73%	15%	12%
fiverr.com	2009	9,788	7,022	56%	37%	63%	49%	9%	42%

Results: Rating Bias (RQ1)

	Rating Score (w/o Interactions)
Completed Tasks	**0.002***[*]
Elite	**0.585*****[***]
Member Since	**−0.092***[*]
Number of Reviews	0.002
Recent Activity	**0.017*****[***]
Female	−0.041
Asian	−0.068
Black	**−0.306*****[***]
Asian Women	
Black Women	

TaskRabbit Rating Regression

Results: Rating Bias (RQ1)

	Rating Score (w/o Interactions)	Rating Score (w/ Interactions)
Completed Tasks	**0.002***	**−0.002***
Elite	**0.585*****	**0.587*****
Member Since	**−0.092***	**−0.100***
Number of Reviews	0.002	0.002
Recent Activity	**0.017*****	**0.017*****
Female	−0.041	−0.08
Asian	−0.068	−0.149
Black	**−0.306*****	**−0.347*****
Asian Women		0.206
Black Women		0.092

TaskRabbit Rating Regression

Results: Rating Bias (RQ1)

	Rating Score (w/o Interactions)	Rating Score (w/ Interactions)
Completed Tasks	**0.002***	**−0.002***
Elite	**0.585*****	**0.587*****
Member Since	**−0.092***	**−0.100***
Number of Reviews	0.002	0.002
Recent Activity	**0.017*****	**0.017*****
Female	−0.041	−0.08
Asian	−0.068	−0.149
Black	**−0.306*****	**−0.347*****
Asian Women		0.206
Black Women		0.092

TaskRabbit Rating Regression

	Rating Score (w/o Interactions)	Rating Score (w/ Interactions)
"About" Length	**0.013***	**0.002*****
Avg. Response Time	**0.002*****	**0.002*****
Facebook Profile	0.042	**0.193***
Google+ Profile	**0.355*****	**0.368*****
Member Since	**0.36*****	**0.422*****
Spoken Languages	**0.69****	0.014
No Image	**−0.608*****	
Not Human Image	−0.079	
Female	**0.175***	**0.203***
Asian	**−0.222****	**−0.377*****
Black	**−0.45*****	**−0.367***
Asian Female		0.15
Black Female		−0.156

Fiverr Rating Regression

Results: Ranking Bias (RQ2)

	Search Rank (w/o Interactions)
Avg. Rating	**0.003*** **
Completed Tasks	**0.003*** **
Member Since	**0.457*** **
Recent Activity	**0.105*** **
Reviews	-0.000
Female	-0.066
Asian	**0.283*** **
Black	**−0.076* **
Asian Female	
Black Female	

TaskRabbit Rank Regression

Results: Ranking Bias (RQ2)

	Search Rank (w/o Interactions)	Search Rank (w/ Interactions)
Avg. Rating	0.003***	0.003***
Completed Tasks	0.003***	0.003***
Member Since	0.457***	0.51***
Recent Activity	0.105***	0.089***
Reviews	-0.000	-0.004
Female	-0.066	−0.468***
Asian	0.283***	0.194*
Black	−0.076*	−0.428***
Asian Female		0.364*
Black Female		1.3***

TaskRabbit Rank Regression

Design Brainstorm: TikTok

Question 1: What research questions do you have about TikTok?

Question 2: What data do you want to collect to answer these questions? How?