Data Deserts and Black Boxes: The Impact of Socio-Economic Status on Consumer Profiling

Nico Neumann and Catherine Tucker

Outline

Empirical Studies

Inequality and inaccuracy?

Research Question

How effective is big-data and ML profiling at delivering audience segments to advertisers and how does it interact with inequality?

Figure: People like saying that big data is like ‘gold’ or ‘oil’ in this economy

What Kind of Data Do Firms Buy (Lotme)

- Age (76%),
- Gender (61%)
- Household Income (50%)
- Education (40%)
- Number of Children in Household (32%).

But how do Data Brokers Know Age and Gender?

Simple prediction task

- Data on Browsing behavior
- May tell us whether someone is a female (if I browse sanitary products)
- May tell us age (if I browse retirement homes)

Outline

Empirical Studies

Inequality and inaccuracy?

What we did

- We identified cookies from ‘pureprofile’ panel survey.
- We asked data brokers to tell whether they were male or in the age bracket (25-34)

Results

Table: Study Three: Data Broker Accuracy at Profiling a Cookie They Have Data For

Data Broker	Attribute	Sample Size	Accuracy
Vendor A	Gender	1396	27.5
Vendor B	Gender	408	25.7
Vendor C	Gender	1777	35.2
Vendor D	Gender	495	56.4
Vendor E	Gender	527	48.8
Vendor F	Gender	480	47.9
Vendor G	Gender	562	46.8
Vendor H	Gender	1016	33.2
Vendor I	Gender	2336	33.6
Vendor J	Gender	14342	42.4
Vendor K	Gender	346	30.6
Vendor L	Gender	547	51.9
Vendor M	Gender	456	49.1
Vendor N	Gender	5099	62.7
Vendor A	Age	217	30.9
Vendor M	Age	296	20
Vendor G	Age	221	36.7
Vendor L	Age	141	15.6
Vendor N	Age	2825	28.8
Vendor K	Age	62	30.6
Vendor I	Age	33036	17.8
Vendor E	Age	211	32.2
Vendor J	Age	10935	18.7

Results

Table: Study Three: Data Broker Accuracy at Profiling a Cookie They Have Data For

Data Broker	Number of Cookies	Gender Accuracy
A	1396	27.5
B	408	25.7
C	1777	35.2
D	495	56.4
E	527	48.8
F	480	47.9
G	562	46.8
H	1016	33.2
I	2336	33.6
J	14342	42.4
K	346	30.6
L	547	51.9
M	456	49.1
N	5099	62.7

What we found

- Gender accuracy ranges from 25.7% to 62.7%. Chance 50%.
- Age bracket precision ranges from 17.8% to 36.7%. Chance 18%.
- Do a little bit better on age
- Regression analysis says they do better when no children, and person is in the UK (not Australia or New Zealand)

Outline

Empirical Studies

Inequality and inaccuracy?

Onto the new work

- Trying to understand why prediction is so poor
- But also trying to understand who inaccurate profiling affects

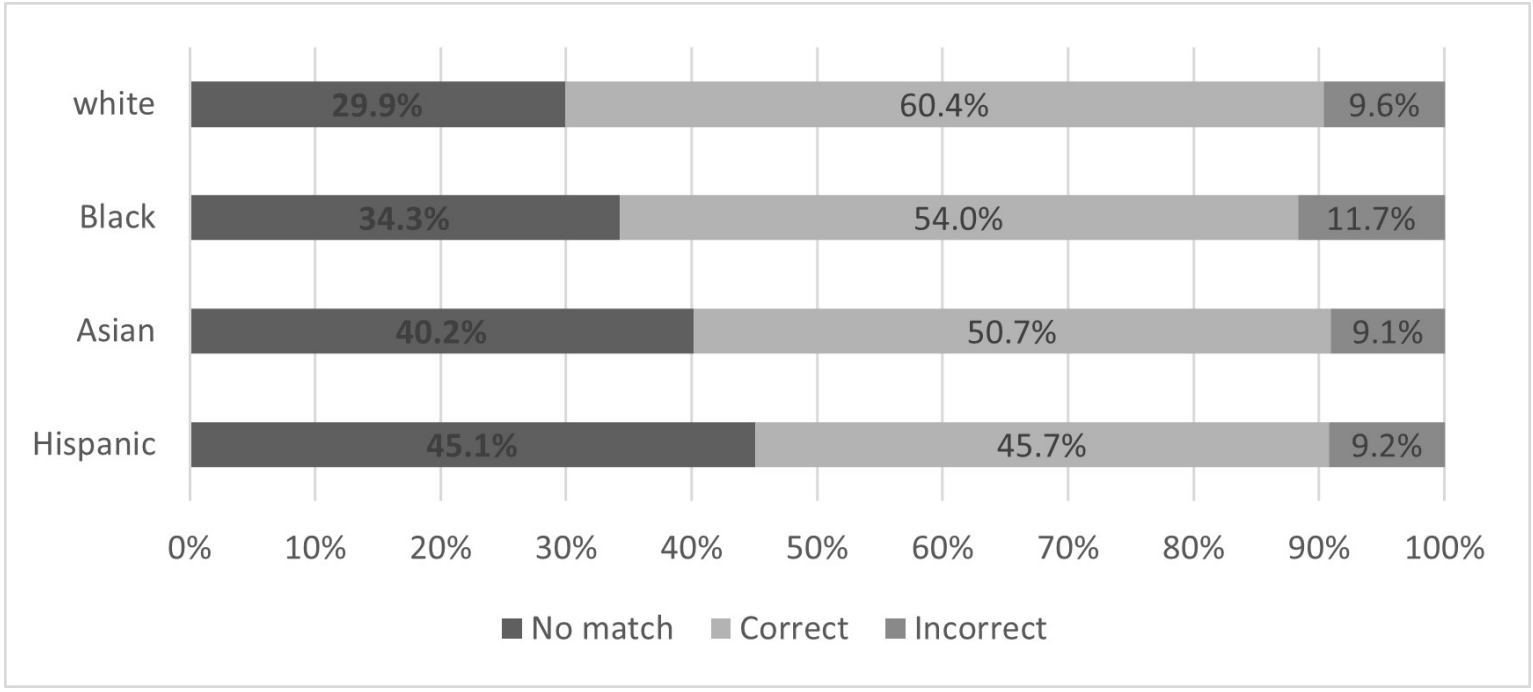
We went out and got new data on the people who were profiled

- We wanted to know if this was related to income inequality

What We Found

- Richer, more educated, home-owning people are more likely to be profiled accurately
- In particular, they are more likely to have accurate demographic information

And Race..


white
29.9%
60.4%
9.6%
Black
34.3%
54.0%
11.7%
Asian
40.2%
50.7%
9.1%
Hispanic
45.1%
45.7%
9.2%
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
No match
Correct
Incorrect

Provocative Conclusion

- Digital divide in data is not talked about enough
- Privacy (accurate profiling) is a ‘rich’ person’s concern
- Perhaps for low-income people data inaccuracy is a bigger concern
- Do we have the current privacy debate the right way around?

Thank you. Questions?

cetucker@mit.edu