AI and Exclusion: The Fragility of Prediction to Missing Data

Catherine Tucker

MIT

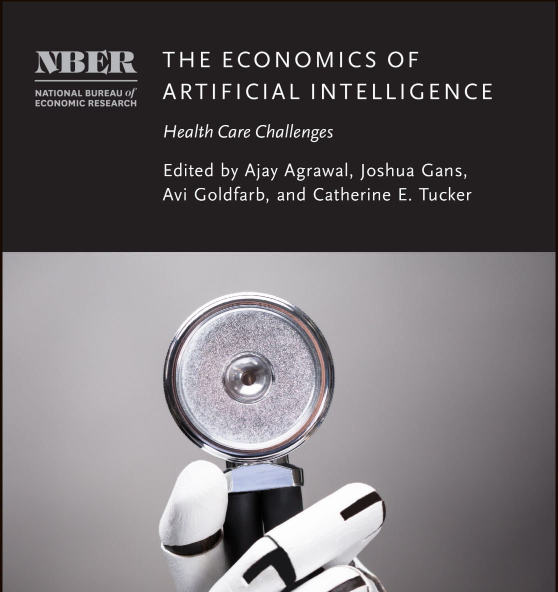
Agenda

1. Introduction
2. Algorithmic Exclusion?
3. Sparse Data
4. Fragmented Data
5. What are the effects?
6. Provocative Conclusions

My Plan for my Keynote

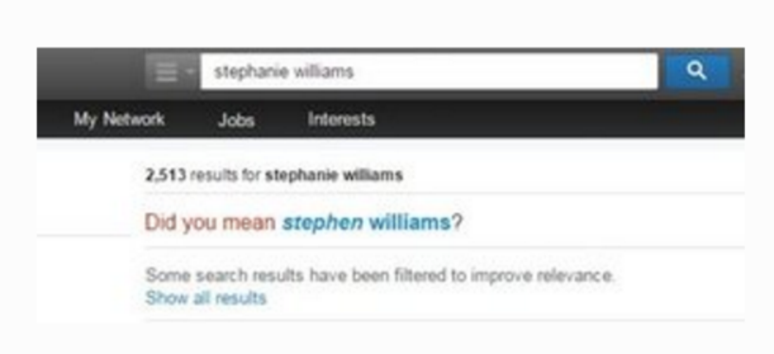
1. Some motivation
2. Some data
3. Some provocative conclusions

How do economists think about AI?



How This All Started

What is Algorithmic Bias?



Agenda

1. Introduction
2. Algorithmic Exclusion?
3. Sparse Data
4. Fragmented Data
5. What are the effects?
6. Provocative Conclusions

Outline

1. Introduction
2. Algorithmic Exclusion?
3. Sparse Data
4. Fragmented Data
5. What are the effects?
6. Provocative Conclusions

Algorithmic Exclusion

When Algorithms err because data is missing due to differences in privilege

- Sparsity
- Fragmentation

In equation form (this may be in the early morning but I am from MIT):

$$Y = X\beta + \epsilon$$

Agenda

1. Introduction
2. Algorithmic Exclusion?
3. Sparse Data
4. Fragmented Data
5. What are the effects?
6. Provocative Conclusions

Outline

1. Introduction
2. Algorithmic Exclusion?
3. Sparse Data
4. Fragmented Data
5. What are the effects?
6. Provocative Conclusions

Sparse Data

More general point that a broad digital footprint is a matter of privilege

- Computer Work
- Mobile Data
- Internet of Things

The idea of data deserts is neglected

Agenda

1. Introduction
2. Algorithmic Exclusion?
3. Sparse Data
4. Fragmented Data
5. What are the effects?
6. Provocative Conclusions

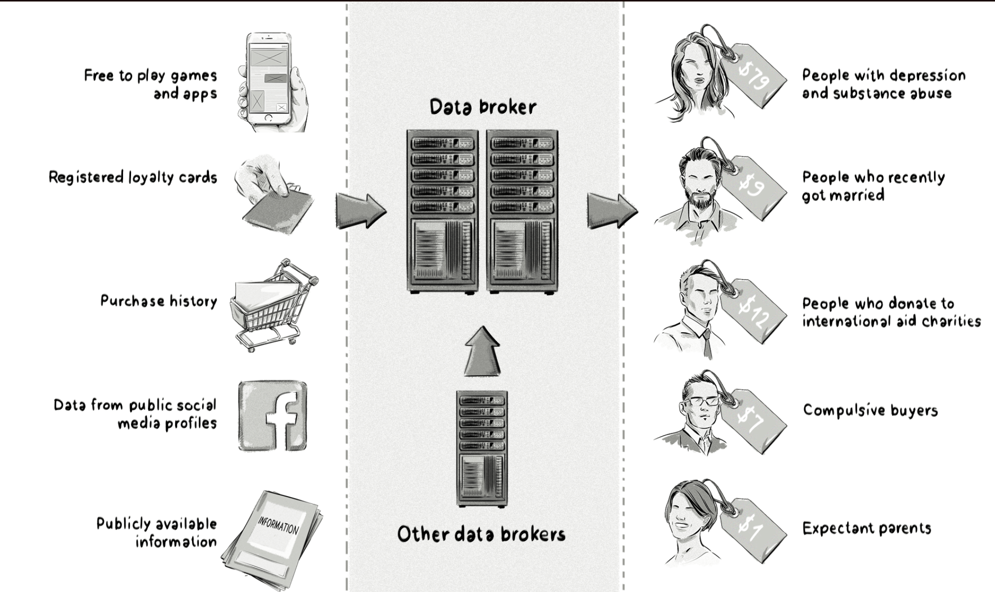
Fragmented Data

- Algorithmic data is not usually from single source
- Datasets have to be matched a
- How do you match? Cell phones..Email addresses...Names

Agenda

1. Introduction
2. Algorithmic Exclusion?
3. Sparse Data
4. Fragmented Data
5. What are the effects?
6. Provocative Conclusions

Based on Algorithms of Data Brokers



What Kind of Predictions are bought by data broker clients (Lotme)

- Age (76%)
- Gender (61%)
- Income (50%)
- Education (40%)
- Children (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)

We asked how good data brokers are at this

What we did

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

Results

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

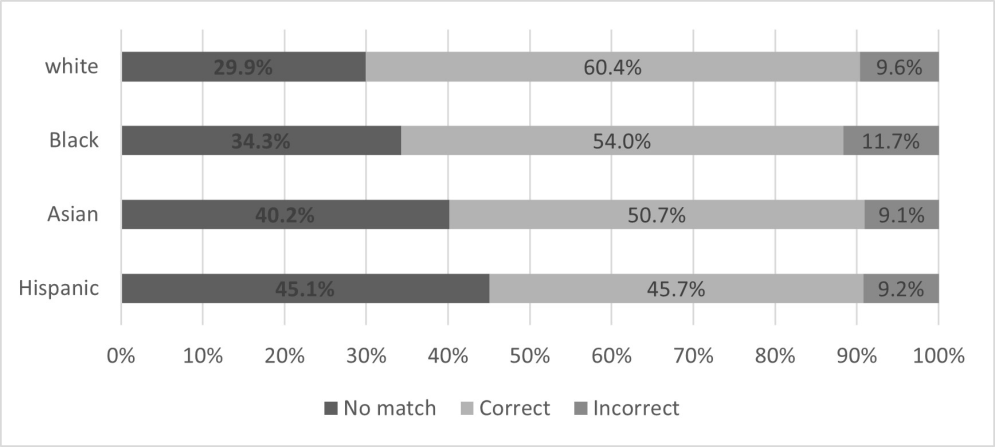
We went out and got new data on the people who were profiled

- 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

But should we care if people are poorly profiled by algorithms as they have missing data?

Summary

- Data is often sparse
- Data is often fragmented
- This leads to algorithmic exclusion where algorithms work poorly
- Interaction with inequality appears important outside of advertising

Agenda

1. Introduction
2. Algorithmic Exclusion?
3. Sparse Data
4. Fragmented Data
5. What are the effects?
6. Provocative Conclusions

Outline

1. Introduction
2. Algorithmic Exclusion?
3. Sparse Data
4. Fragmented Data
5. What are the effects?
6. Provocative Conclusions

Provocative Conclusion: 1

- Privacy 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?

Provocative Conclusion 2

- Algorithmic transparency or auditing doesn’t address this
- Instead we need to also think about data deserts in the way we think about food deserts