Data Deserts, Data Accuracy and Inequality

Catherine Tucker

Outline

Outline

Speech Plan

1. Feel-Good Research Results
2. Not Feel-Good Research Results
3. Appropriately ‘Thought-Provoking’ Ranting

1: Digital Data Can Help Reduce Inequality

- Miller and Tucker (2011) showed that having patient data helps avoid babies dying in a hospital.
- Particular for black women and women for whom English was a second language.

2: Digital Technology Can Help Reduce Inequality

- Tucker et al. (2019) show that mobile technology helps improve access to government services
- By taking out the ability to ‘talk to a manager’ and use digital tools to standardize communications people in traditionally less privileged census blocks got faster access to government maintenance services.

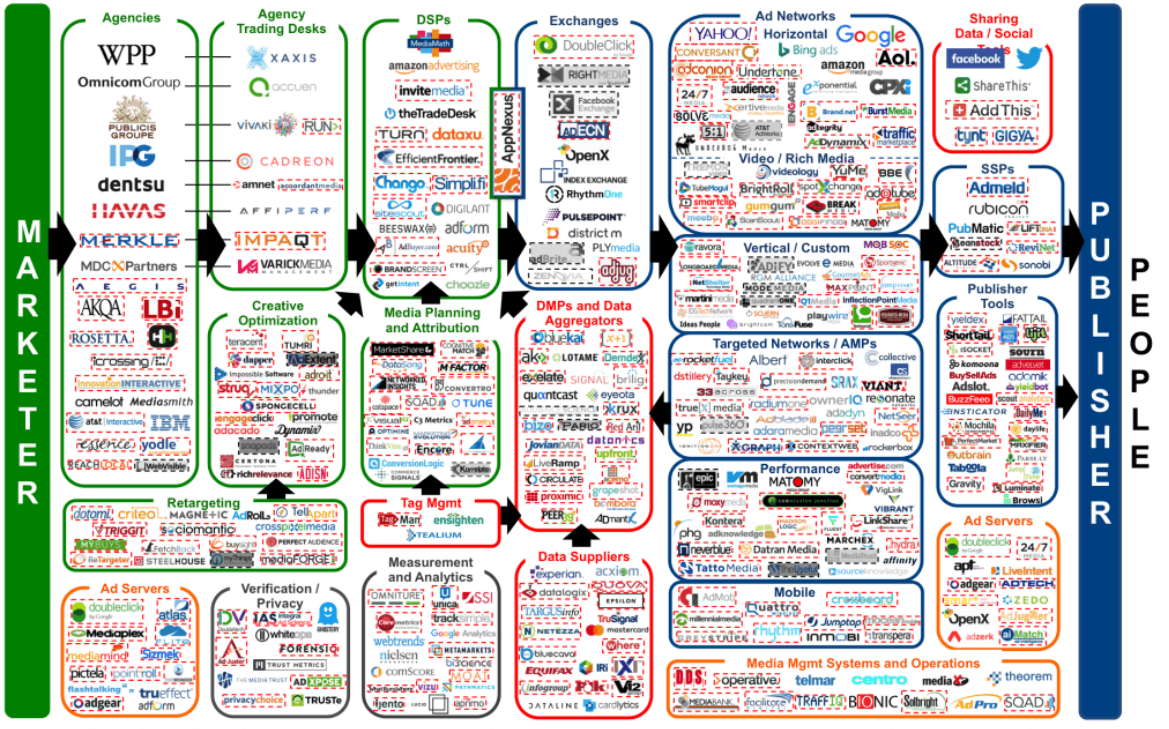
Research Question

‘How Effective Is Black-Box Digital Consumer Profiling And Audience Delivery?: Evidence from Field Studies’ Joint Work with Nico Neumann and Tim Whitfield

Research Question

How effective is big-data and ML profiling at delivering audience segments to advertisers?

DISPLAY LUMAscape



LUMA

Denotes acquired company

Denotes shuttered company

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

Data collection for profiling also raises privacy concerns

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)

In this paper we ask how effective are attempts at getting Age and Gender right

Outline

Drivers of the Economics of Personal Data

Feel-Good Results
- Empirical Studies

B2B?

Inequality and inaccuracy?

Provocative Conclusions

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

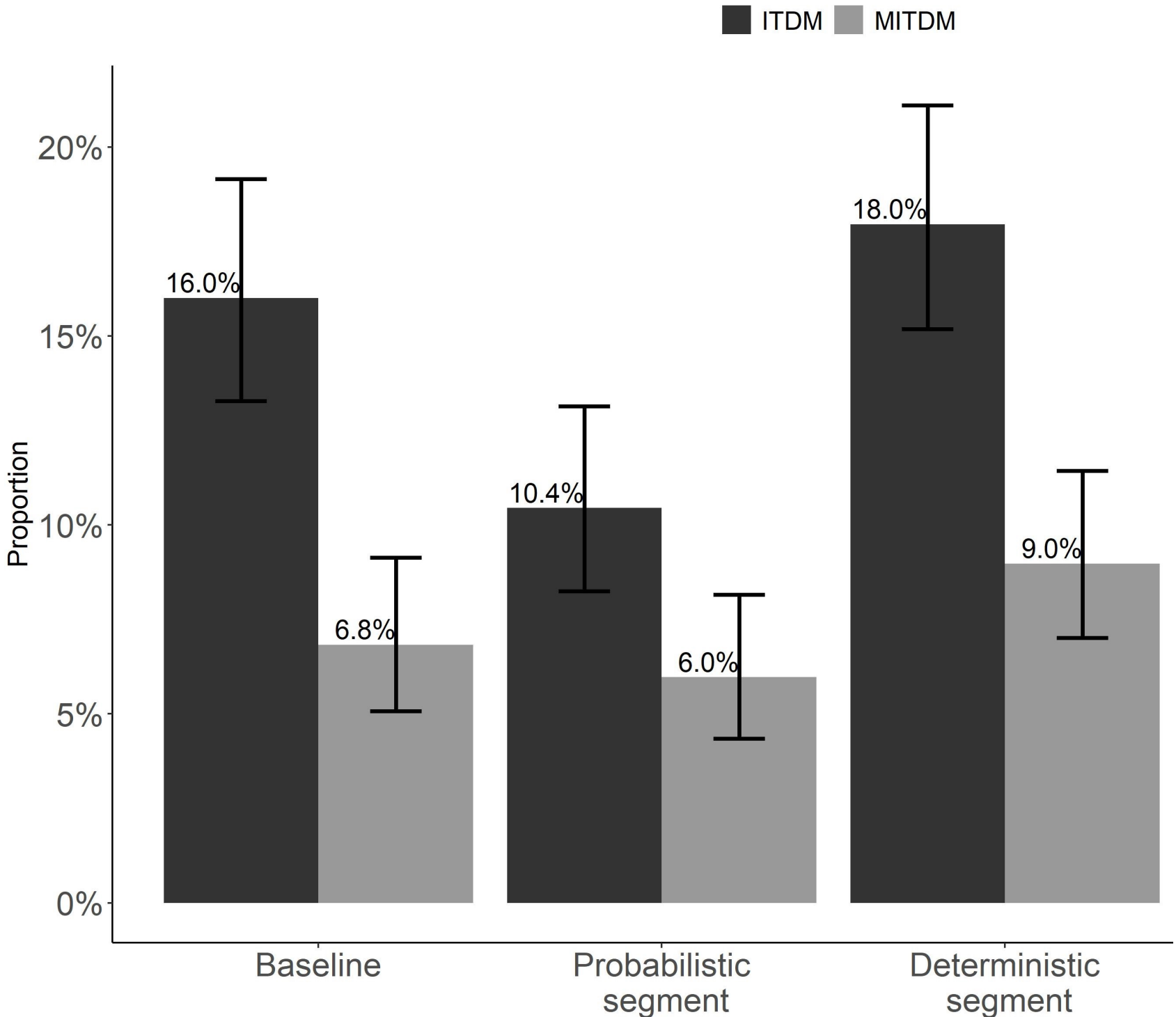
OK...B2C is bad...but surely B2B is better?

“What Type of Digital Advertising is Most Effective To Reach The ‘Right’ Customers: The Case of IT Decision-Makers” with Nico Neumann

Will repeat the exercise with IT Decision Makers

Ground Truth

1. IT needs identification: “I identify needs for new IT products in my company/department.”
2. IT vendor selection: “I select/shortlist vendors for IT purchases in my company.”
3. IT contract responsibility: “I sign contracts/make financial decisions for IT purchases in my company.”


ITDM
MITDM
20%
15%
10%
5%
0%
Proportion
16.0%
6.8%
10.4%
6.0%
18.0%
9.0%
Baseline
Probabilistic segment
Deterministic segment

Why is it so bad?

1. Probablistic: Mismatching of a profile to digital identifier when onboarding
2. Deterministic: Prospect lists - the names, titles, emails just don’t match reality.

The Result That HURTS

You would do better targeting middle-aged men than using any of these segments.

Outline

Research Question

‘Does accurate consumer profiling depend on who you are? An empirical investigation of what is driving audience profiling errors.’ Joint Work with Nico Neumann

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

- So now we now not only whether their age or gender or location or their hobbies were accurately profiled.
- But we also know what their socio-economic status is
- We also know something about the data broker and what they charged for the data
- We can investigate vendor-side vs consumer-side drivers of accuracy

Table: Observed Demographic and Socio-Economic Consumer Characteristics

Variable	Level Description	Observations	Mean	SD	Median	Min	Max
Demographic and socio-economic characteristics		3588 people					
Gender	Male	1186	0.33	0.47	0	0	1
	Female	2402	0.67	0.47	1	0	1
Age	Continuous variable	3588	44.37	12.94	43	16	104
Income (000s)	Continuous variable	3588	89.79	63.40	80	5	405
Education	Diploma	1503	0.42	0.49	0	0	1
	School	864	0.24	0.43	0	0	1
	College degree	1221	0.34	0.47	0	0	1
Job type	White-collar job	3132	0.87	0.33	1	0	1
	Blue-collar job	456	0.13	0.33	0	0	1
Employment	Employed	3383	0.94	0.23	1	0	1
	Not employed	205	0.06	0.23	0	0	1
Home ownership	Owns home	2337	0.65	0.48	1	0	1
	Rents home	1251	0.35	0.48	0	0	1
Has children	No	1241	0.35	0.48	0	0	1
	Yes	2347	0.65	0.48	1	0	1
Household type	Family	2073	0.58	0.49	1	0	1
	Shared	997	0.28	0.45	0	0	1
	Single	518	0.14	0.35	0	0	1

Does Price Moderate Accuracy?

	(1) correct	(2) correct
CPM Price	0.0975***	-0.00430
	(0.0106)	(0.00887)
Attribute Fixed Effects	No	Yes
Observations	12167	12167
R-Squared	0.00686	0.442
Mean Dep Var	0.578	0.578

Are Certain Data Brokers More Accurate?

	(1) correct	(2) correct	(3) correct
vendor.id=1	0.309		0.289
	(0.248)		(0.214)
vendor.id=2	-0.400		0.253
	(0.247)		(0.215)
vendor.id=3	0.241		0.281
	(0.248)		(0.214)
vendor.id=4	0.228		0.270
	(0.248)		(0.214)
vendor.id=5	-0.398		0.299
	(0.247)		(0.214)
vendor.id=6	-0.0890		0.283
	(0.247)		(0.213)
vendor.id=7	-0.407		0.254
	(0.247)		(0.215)
vendor.id=8	0.122		0.304
	(0.247)		(0.214)
vendor.id=9	0.274		0.276
	(0.247)		(0.214)
vendor.id=10	0.278		0.281
	(0.267)		(0.230)
vendor.id=11	-0.00952		0.336
	(0.257)		(0.222)
vendor.id=12	-0.0625		0.255
	(0.250)		(0.216)
vendor.id=13	-0.379		0.298
	(0.251)		(0.223)
vendor.id=14	0.0775		0.318
	(0.255)		(0.221)
vendor.id=15	0.0775		0.302
	(0.255)		(0.221)
vendor.id=16	0.0435		0.338
	(0.252)		(0.218)
vendor.id=17	0.246		0.260
	(0.263)		(0.227)
vendor.id=18	0.283		0.263
	(0.265)		(0.229)
vendor.id=19	0.0133		0.192
	(0.261)		(0.226)
vendor.id=20	0.333		0.339
	(0.295)		(0.255)
vendor.id=21	0		0
	(.)		(.)
Attribute Fixed Effects	No	Yes	Yes
Observations	12167	12167	12167
R-Squared	0.251	0.442	0.443
Mean Dep Var	0.578	0.578	0.578

Does Economic Background Drive Accuracy of Profiling?

	(1) correct	(2) correct	(3) correct	(4) correct	(5) correct
Income (000)	0.000326***			0.000244**	0.000264***
	(0.0000766)			(0.0000783)	(0.0000587)
College Graduate		0.0315***		0.0238*	0.0226**
		(0.00943)		(0.00960)	(0.00719)
Own House			0.0542***	0.0504***	0.0211**
			(0.00895)	(0.00899)	(0.00676)
Attribute Fixed Effects	No	No	No	No	Yes
Vendor Fixed Effects	No	No	No	No	Yes
Observations	12167	12167	12167	12167	12167
R-Squared	0.00148	0.000916	0.00300	0.00460	0.445
Mean Dep Var	0.578	0.578	0.578	0.578	0.578

Does Economic Background Drive Accuracy of Profiling?

	Interests		Background Demographics	
	(1)	(2)	(3)	(4)
	correct	correct	correct	correct
Income (000)	0.000187*	0.000151*	0.000337***	0.000351***
	(0.0000770)	(0.0000668)	(0.0000986)	(0.0000867)
College Graduate	0.0139	0.0119	0.0193	0.0281**
	(0.00946)	(0.00821)	(0.0121)	(0.0106)
Own House	-0.00685	-0.00593	0.0535***	0.0389***
	(0.00889)	(0.00773)	(0.0113)	(0.00995)
Attribute Fixed Effects	No	Yes	No	Yes
Vendor Fixed Effects	No	Yes	No	Yes
Observations	4787	4787	7380	7380
R-Squared	0.00213	0.256	0.00580	0.240
Mean Dep Var	0.896	0.896	0.372	0.372

Outline

Conclusions

- We show that black box profiling seems to be working poorly for advertisers and consumer
- Big Data does not appear to be analogous to ‘gold’
- Interaction with inequality appears important outside of advertising
- Instead it appears we need better algorithms

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 Bias dominates the AI fairness debate
- But think of our own research...and where most of our errors come from
- Trying to popularize the idea of ‘Algorithmic Exclusion’ to make people take this seriously

Provocative Conclusion 3: What does this tell us about Competition Data vs Analytics

- Widespread
- Little value without processing
- Implies complementarity in way not conceived of by work on data markets
- If processing is the key input to insight then how can we establish property rights towards data?

Thank you!

cetucker@mit.edu

Miller, A. and C. Tucker (2011). Can healthcare information technology save babies? *Journal of Political Economy* (2), 289–324.

Tucker, C. E., Y. Wang, and S. Yu (2019). Does it lead to more equal treatment? an empirical study of the effect of smartphone use on customer complaint resolution. *An Empirical Study of the Effect of Smartphone Use on Customer Complaint Resolution (June 12, 2019)*.