Computational Sociology

Machine learning challenges

Dr. Thomas Davidson

Rutgers University

April 12, 2021

Plan

1. Course updates
2. The (un)predictability of social life
3. Biased predictions
4. Mitigating bias

Course updates

Homework

- Final homework on machine learning will be released on Friday
  - Content will be similar to that covered in lectures 10 and 11
  - Due 4pm on **4/26**

Course updates

Project: Data collection

- Initial data collection was due today at 4pm
  - Please follow instructions on Slack to submit
    - Make sure to add me to your Github repo and complete the form
- You will use the same Github repository for the next two phases for the project
  - For future submissions you will just need to update the contents

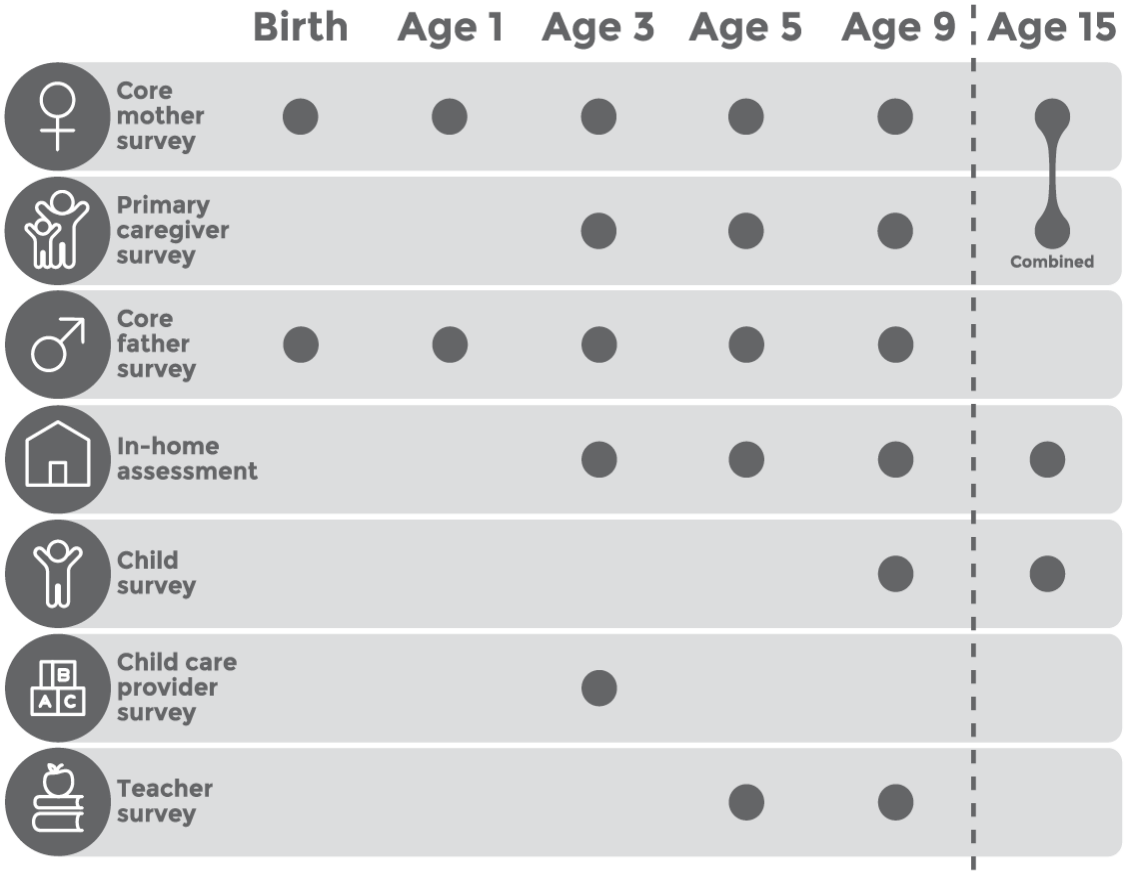
Course updates

Project: Timeline

- Preliminary analyses deadline extended, now due **5/3** at 4pm
- In-class presentations on **5/3**
  - 5 slides in 5 minutes
    - Title
    - Motivation
    - Data
    - Methodology
    - Results
  - 5 minutes of Q&A
- Final paper *new* deadline, **5/10** at 5pm ET

The (un)predictability of social life

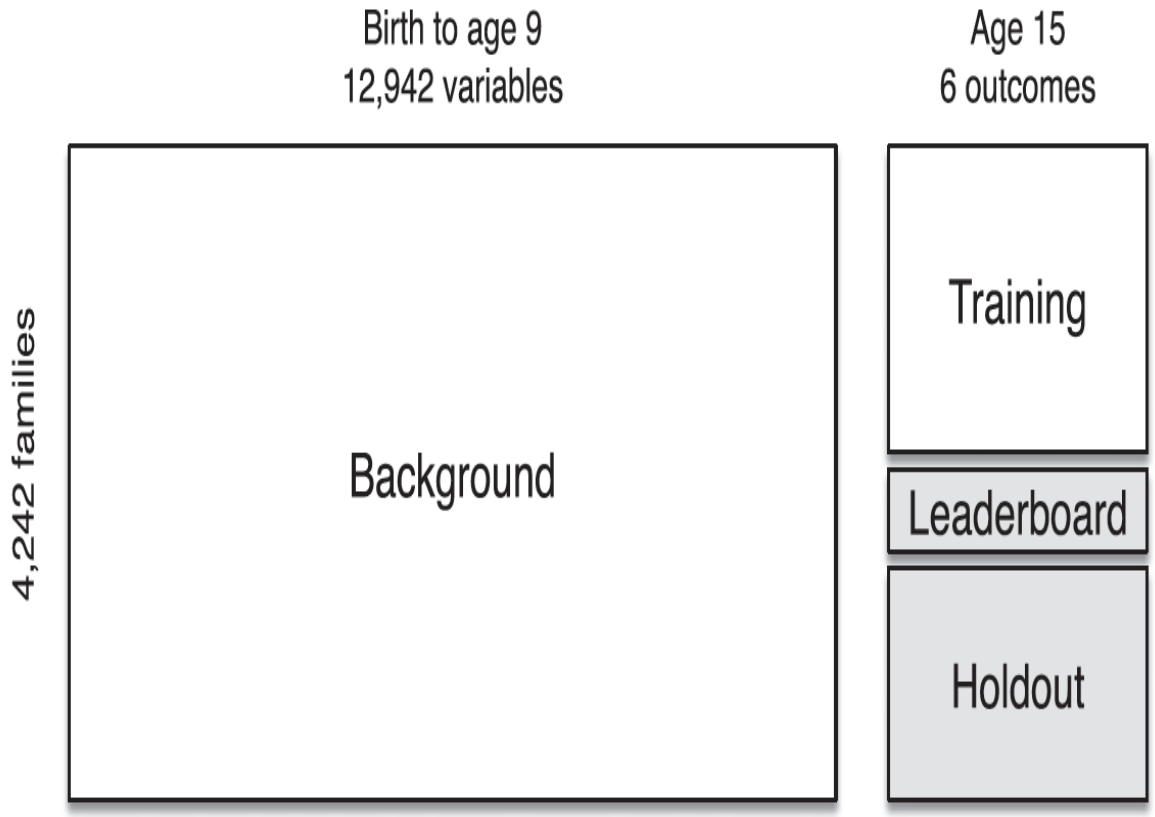
The Fragile Families Challenge



Salganik et al. 2020.

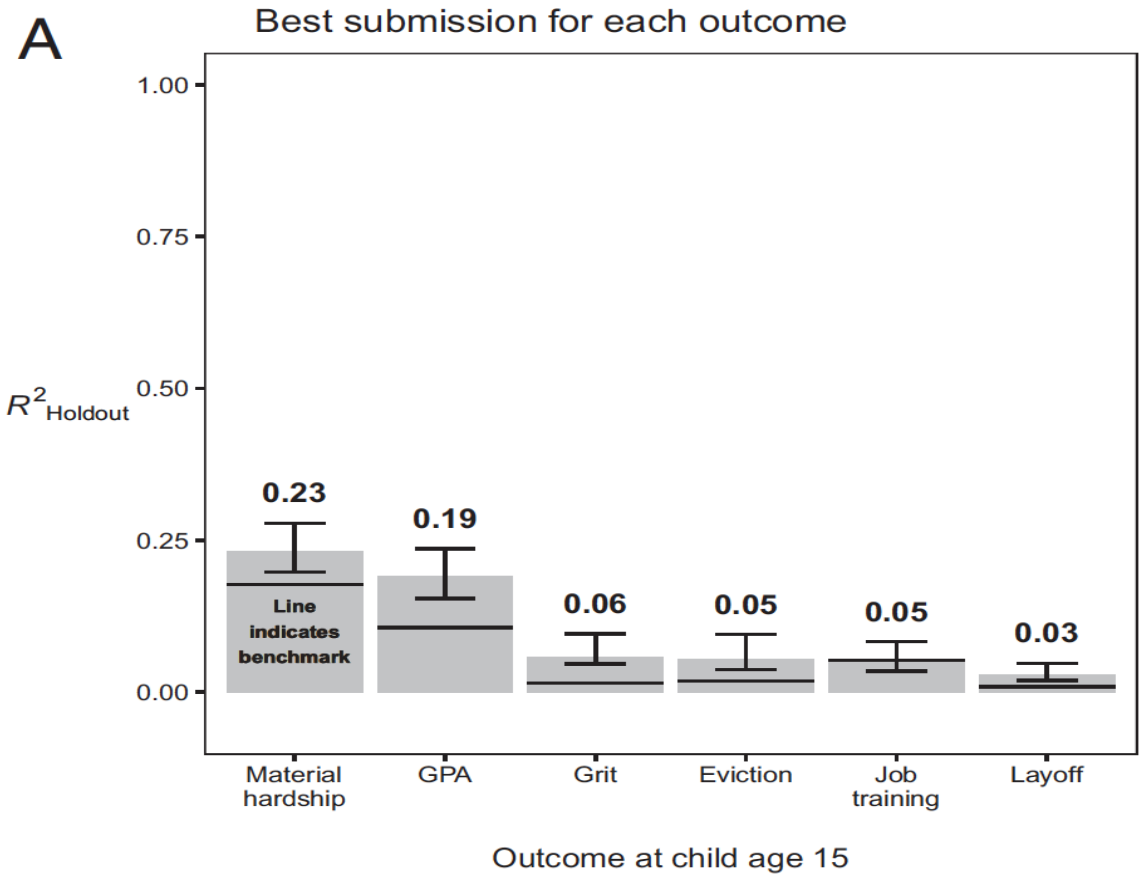
The (un)predictability of social life

The Fragile Families Challenge



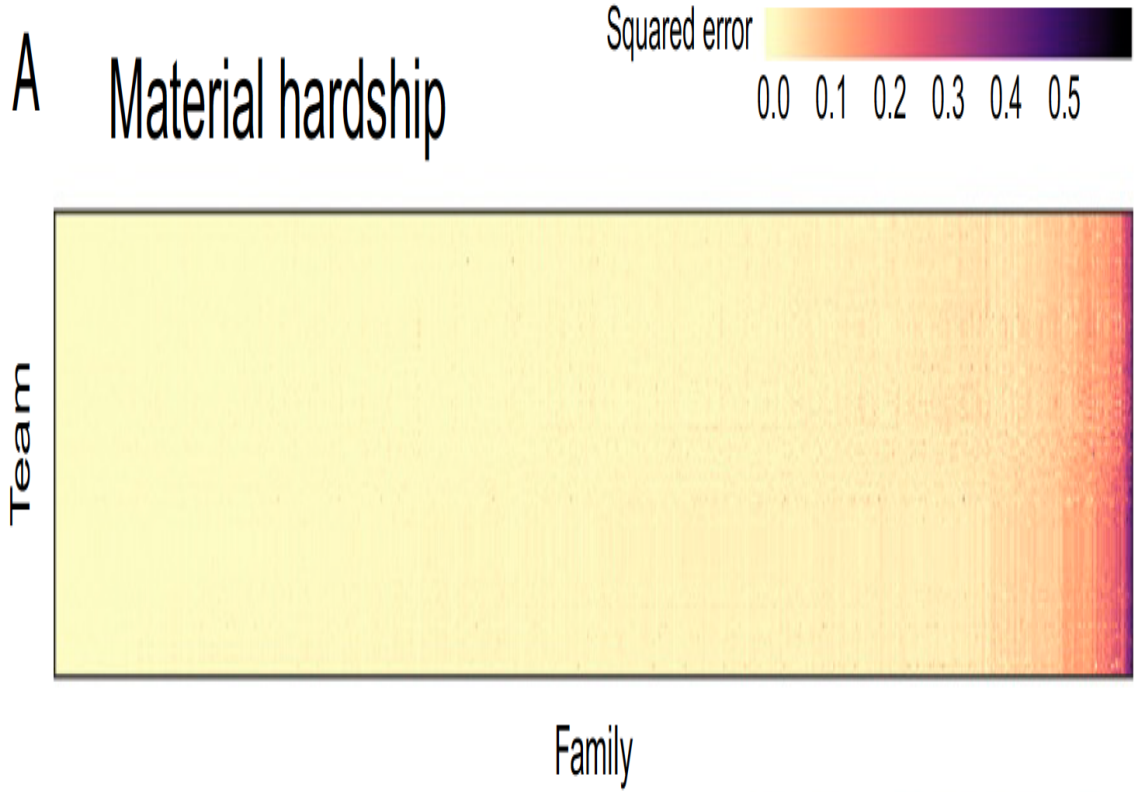
The (un)predictability of social life

The Fragile Families Challenge



The (un)predictability of social life

The Fragile Families Challenge



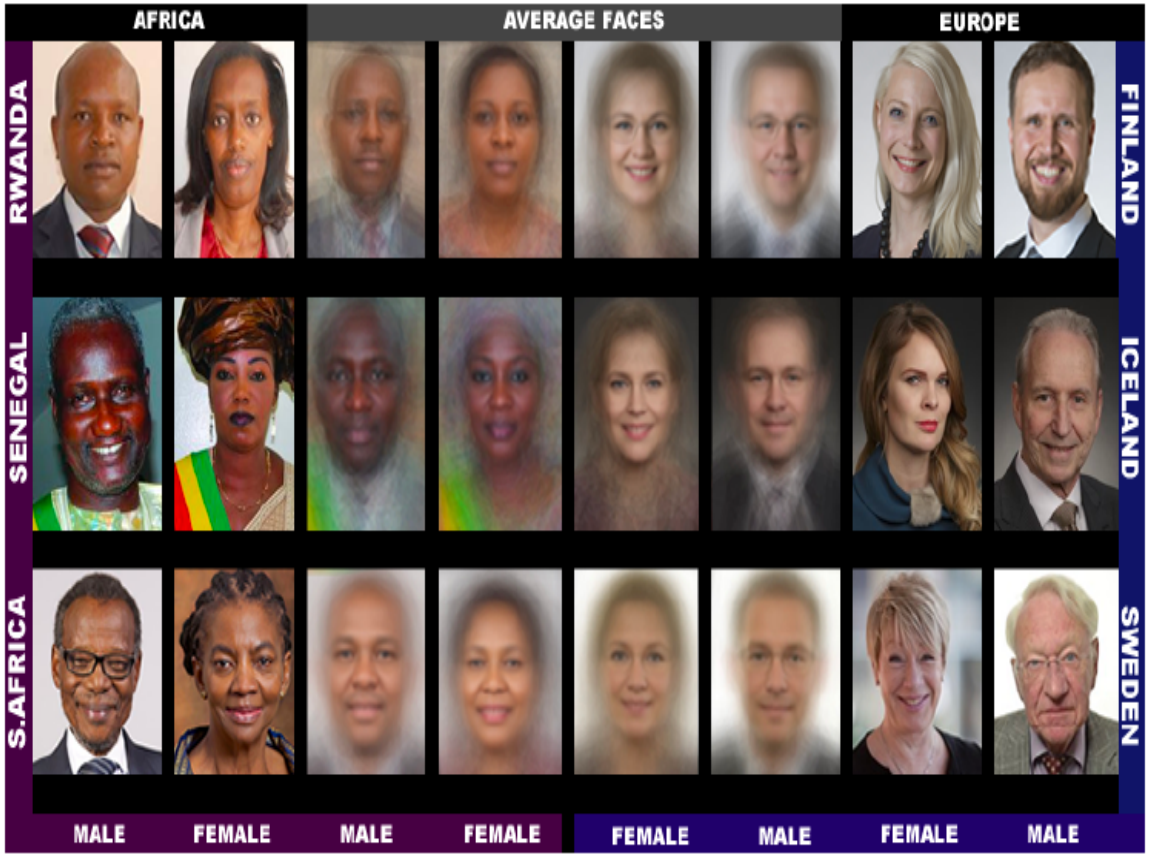
The (un)predictability of social life

Discussion

- Is social life inherently unpredictable or do you think we could predict outcomes better if we improved our measurements?

Biased predictions

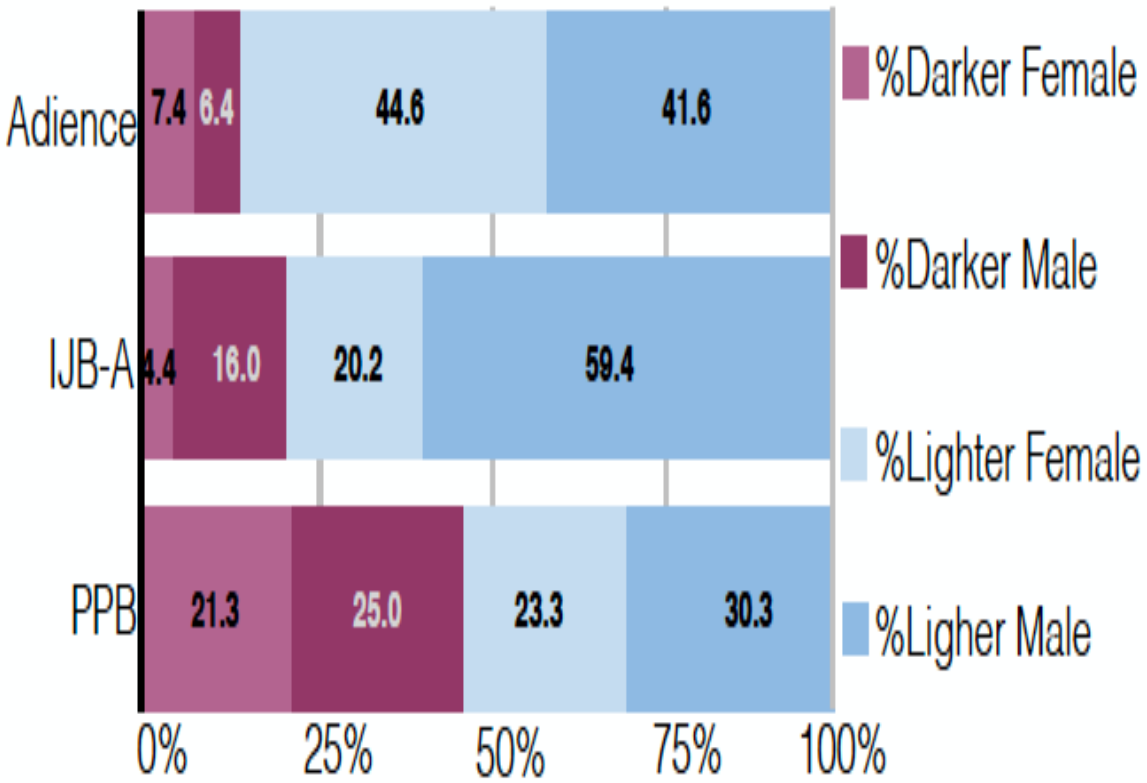
Facial recognition datasets



Buolamwini and Gebru 2018.

Biased predictions

Dataset distributions



Biased predictions

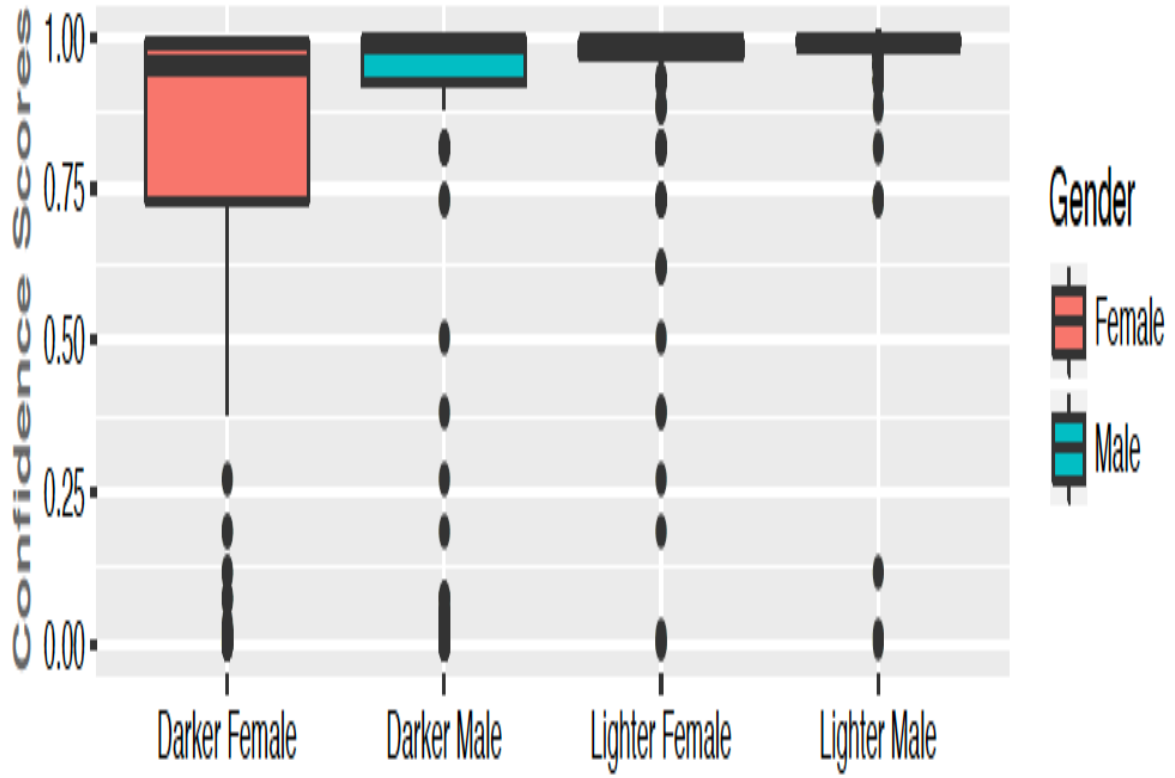
Digital audit studies

Classifier	Metric	DF	DM	LF	LM
MSFT	PPV(%)	76.2	**100**	**100**	**100**
	Error Rate(%)	**23.8**	0.0	0.0	0.0
	TPR(%)	**100**	84.2	**100**	**100**
	FPR(%)	**15.8**	0.0	0.0	0.0
Face++	PPV(%)	64.0	99.5	**100**	**100**
	Error Rate(%)	**36.0**	0.5	0.0	0.0
	TPR(%)	99.0	77.8	**100**	96.9
	FPR(%)	**22.2**	1.03	3.08	0.0
IBM	PPV(%)	66.9	94.3	**100**	98.4
	Error Rate(%)	**33.1**	5.7	0.0	1.6
	TPR(%)	90.4	78.0	96.4	**100**
	FPR(%)	**22.0**	9.7	0.0	3.6

Table 5: Gender classification performance as measured by the positive predictive value (PPV), error rate (1-PPV), true positive rate (TPR), and false positive rate (FPR) of the 3 evaluated commercial classifiers on the South African subset of the PPB dataset. Results for South Africa follow the overall trend with the highest error rates seen on darker-skinned females.

Biased predictions

Digital audit studies



Biased predictions

Coded Bias (2020) documentary

Biased predictions

Discussion

- What are the implications of this study for sociological research?
- Do you think we should use facial recognition software as a research tool?
- What are the other applications of digital audit studies?

Biased predictions

Digital audit studies of text classifiers

Term	Toxic	Overall
atheist	0.09%	0.10%
queer	0.30%	0.06%
gay	3%	0.50%
transgender	0.04%	0.02%
lesbian	0.10%	0.04%
homosexual	0.80%	0.20%
feminist	0.05%	0.05%
black	0.70%	0.60%
white	0.90%	0.70%
heterosexual	0.02%	0.03%
islam	0.10%	0.08%
muslim	0.20%	0.10%
bisexual	0.01%	0.03%

Table 1: Frequency of identity terms in toxic comments and overall.

Dixon et al. 2018.

Biased predictions

Digital audit studies of text classifiers

Template Examples	Label
I am <IDENTITY>	Non-Toxic
I am a <IDENTITY> person, ask me anything	Non-Toxic
<IDENTITY> people are just like everyone else	Non-Toxic
I hate all <IDENTITY>	Toxic
I am a <IDENTITY> person and I hate your guts and think you suck	Toxic
<IDENTITY> people are gross and universally terrible	Toxic

Table 2: Phrase template examples.

Biased predictions

Digital audit studies of text classifiers

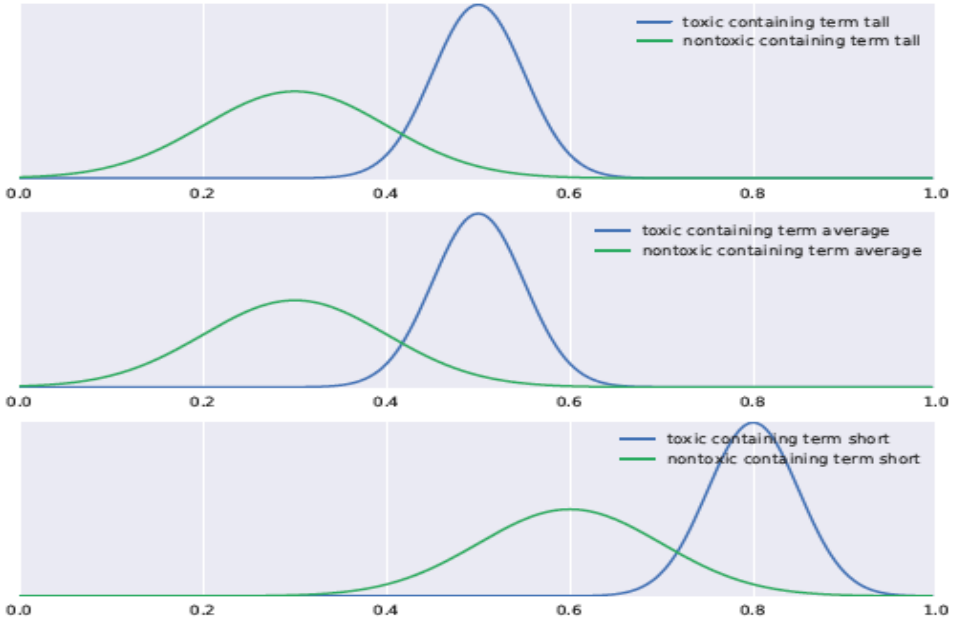


Figure 2: Distributions of toxicity scores for three groups of data, each containing comments with different identity terms, "tall", "average", or "short".

Biased predictions

Racial bias in hate speech detection classifiers

Dataset	Class	$\widehat{p_{i_{black}}}$	$\widehat{p_{i_{white}}}$	t	p	$\frac{\widehat{p_{i_{black}}}}{\widehat{p_{i_{white}}}}$
Waseem and Hovy	Racism	0.001	0.003	-20.818	***	0.505
	Sexism	0.083	0.048	101.636	***	1.724
Waseem	Racism	0.001	0.001	0.035		1.001
	Sexism	0.023	0.012	64.418	***	1.993
	Racism and sexism	0.002	0.001	4.047	***	1.120
Davidson et al.	Hate	0.049	0.019	120.986	***	2.573
	Offensive	0.173	0.065	243.285	***	2.653
Golbeck et al.	Harassment	0.032	0.023	39.483	***	1.396
Founta et al.	Hate	0.111	0.061	122.707	***	1.812
	Abusive	0.178	0.080	211.319	***	2.239
	Spam	0.028	0.015	63.131	***	1.854

Davidson, Bhattacharya, and Weber 2019.

Biased predictions

Racial bias in hate speech detection classifiers

Dataset	Class	$\widehat{p_{i_{black}}}$	$\widehat{p_{i_{white}}}$	t	p	$\frac{\widehat{p_{i_{black}}}}{\widehat{p_{i_{white}}}}$
Waseem and Hovy	Racism	0.010	0.010	-0.632		0.978
	Sexism	0.963	0.944	20.064	***	1.020
Waseem	Racism	0.011	0.011	-1.254		0.955
	Sexism	0.349	0.290	28.803	***	1.203
	Racism and sexism	0.012	0.012	-0.162		0.995
Davidson et al.	Hate	0.017	0.015	4.698	***	1.152
	Offensive	0.988	0.991	-6.289	***	0.997
Golbeck et al.	Harassment	0.099	0.091	6.273	***	1.091
Founta et al.	Hate	0.074	0.027	46.054	***	2.728
	Abusive	0.925	0.968	-41.396	***	0.956
	Spam	0.010	0.010	0.000		1.000

This table shows results for tweets containing the word "b***h". This word was used in ~1.7% of AAE and 0.5% of SAE tweets.

Biased predictions

Gender bias in occupation classifiers

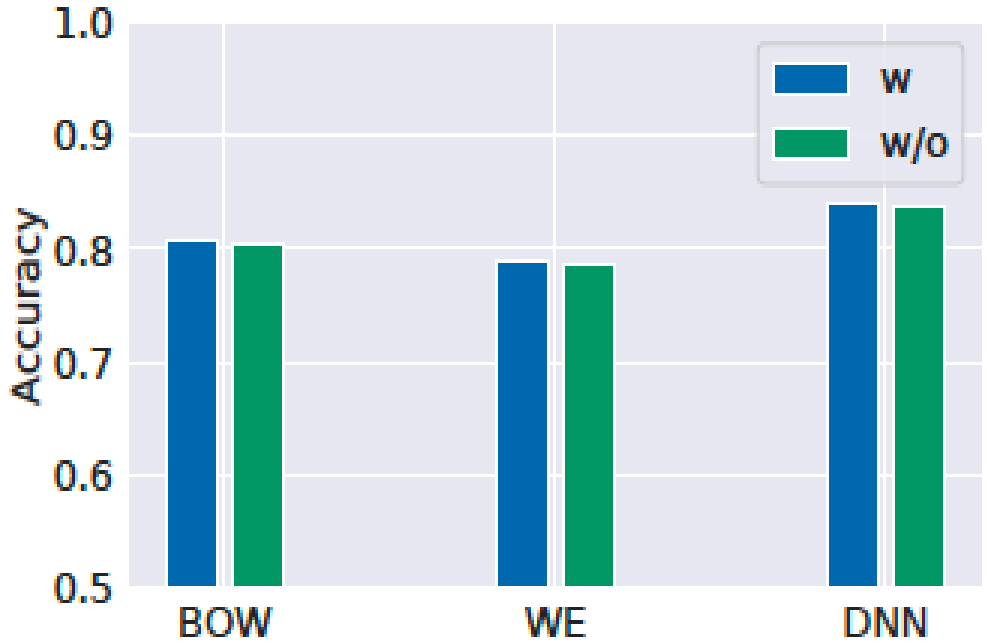


Figure 2: Occupation classifier accuracy for each semantic representation, with and without explicit gender indicators.

De-Arteaga et al. 2019.

Biased predictions

Gender bias in occupation classifiers

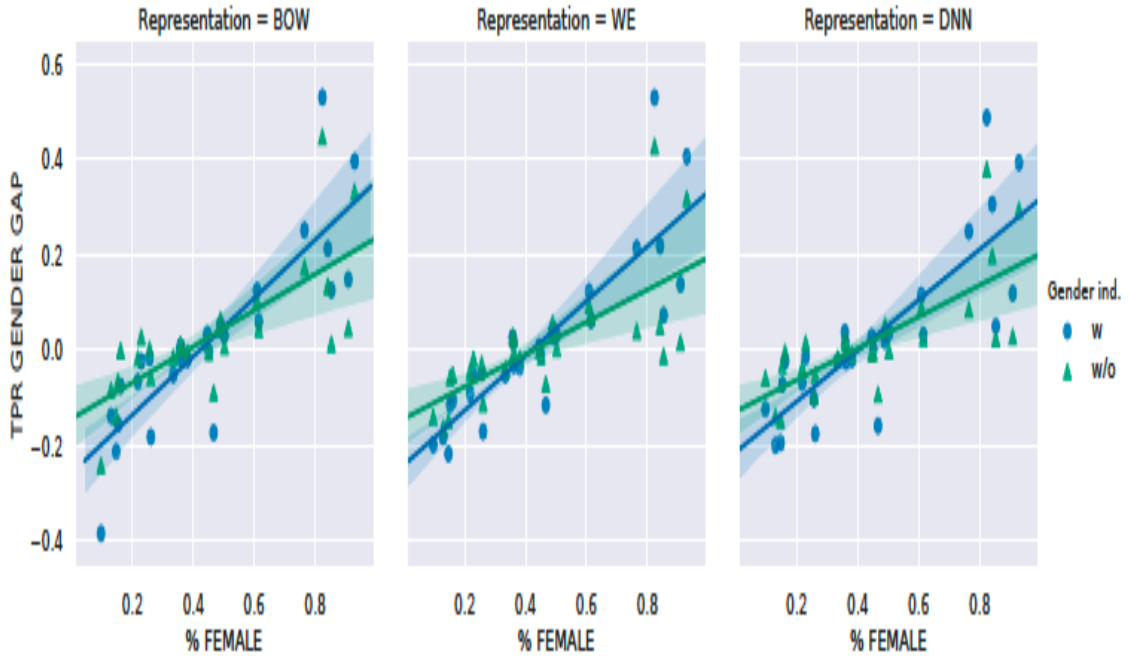
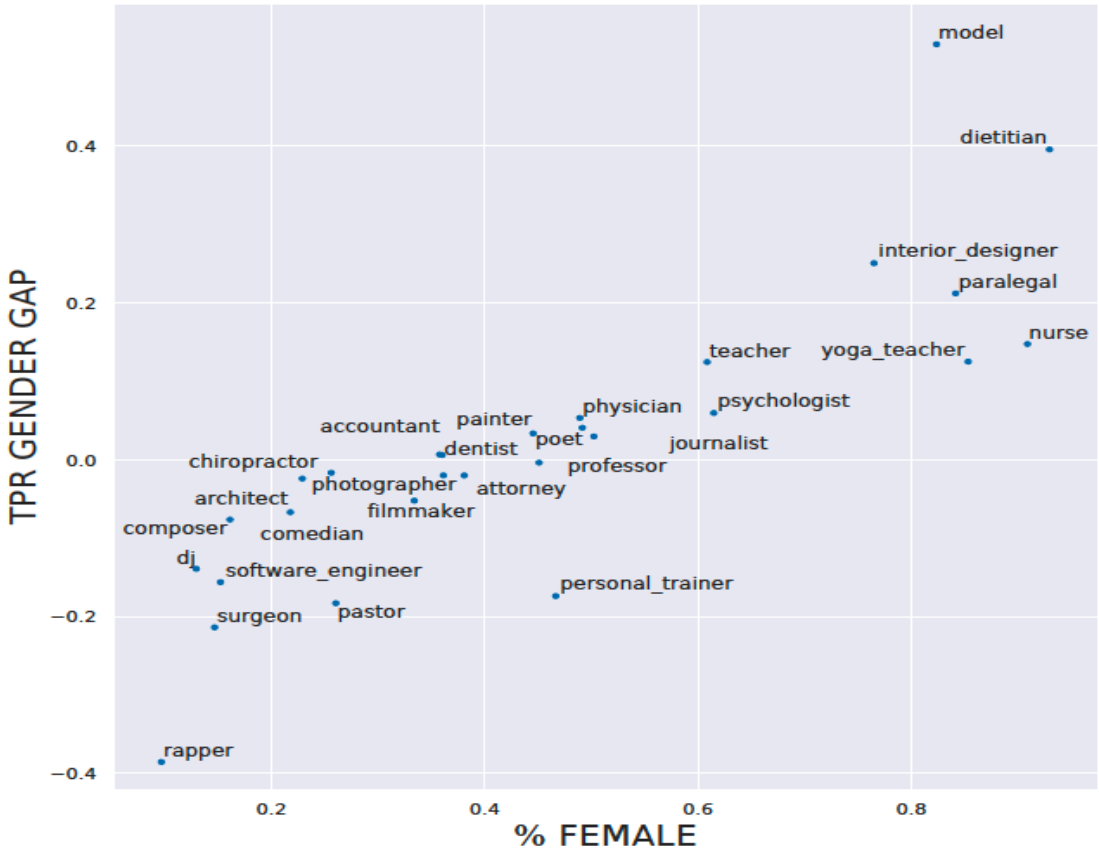


Figure 4: $\text{Gap}_{\text{female},y}$ versus $\pi_{\text{female},y}$ for each occupation y for all three semantic representations, with and without explicit gender indicators. Correlation coefficients: BOW-w 0.85; BOW-wo 0.74; WE-w 0.86; WE-wo 0.71; DNN-w 0.82, DNN-wo 0.74.

Biased predictions

Gender bias in occupation classifiers



Biased predictions

Gender bias in occupation classifiers

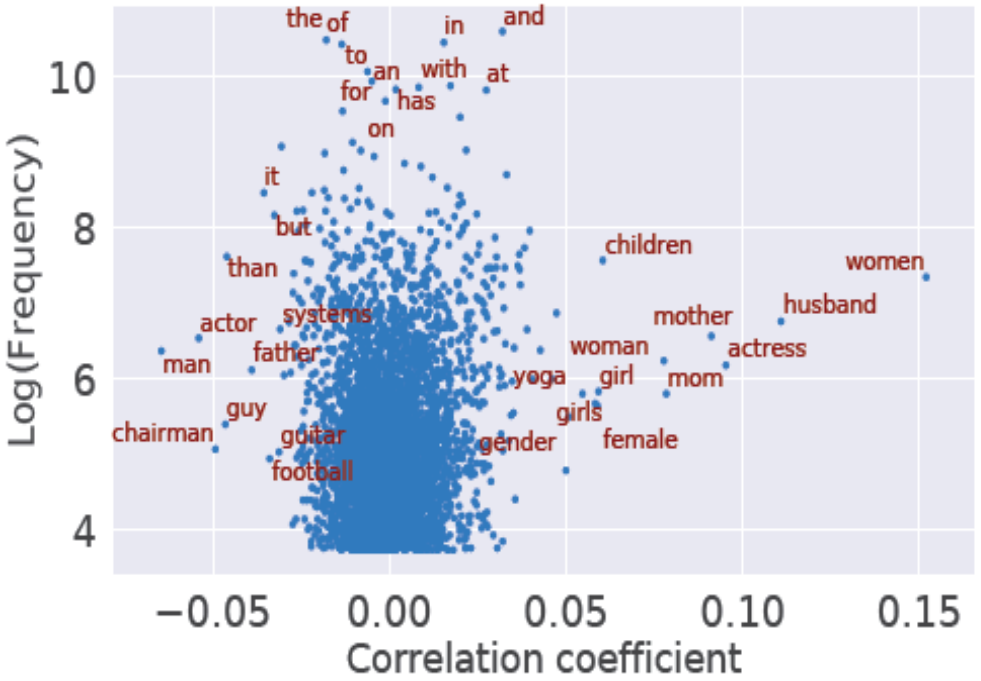


Figure 5: Scatterplot of log frequency versus correlation with G = female for each word type in the vocabulary.

Biased predictions

Gender bias in occupation classifiers

william henry gates iii (born october 28 , 1955) is an american business magnate , investor , author , philanthropist , humanitarian , and principal founder of microsoft corporation . during his career at microsoft , gates held the positions of chairman , ceo and chief software architect , while also being the largest individual shareholder until may 2014 . in 1975 , gates and paul allen launched microsoft , which became the world 's largest pc software company . gates led the company as chief executive officer until stepping down in january 2000 , but he remained as chairman and created the position of chief software architect for himself . in june 2006 , gates announced that he would be transitioning from full-time work at microsoft to part-time work and full-time work at the bill & melinda gates foundation , which was established in 2000 .

Figure 7: Visualization of the DNN's per-token attention weights. Predicted label (i.e., occupation): *software engineer*.

Biased predictions

Bias in language models

On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜

Emily M. Bender*
ebender@uw.edu
University of Washington
Seattle, WA, USA

Timnit Gebru*
timnit@blackinai.org
Black in AI
Palo Alto, CA, USA

Angelina McMillan-Major
aymm@uw.edu
University of Washington
Seattle, WA, USA

Shmargaret Shmitchell
shmargaret.shmitchell@gmail.com
The Aether

ABSTRACT

The past 3 years of work in NLP have been characterized by the development and deployment of ever larger language models, especially for English. BERT, its variants, GPT-2/3, and others, most recently Switch-C, have pushed the boundaries of the possible both through architectural innovations and through sheer size. Using these pretrained models and the methodology of fine-tuning them for specific tasks, researchers have extended the state of the art on a wide array of tasks as measured by leaderboards on specific

alone, we have seen the emergence of BERT and its variants [39, 70, 74, 113, 146], GPT-2 [106], T-NLG [112], GPT-3 [25], and most recently Switch-C [43], with institutions seemingly competing to produce ever larger LMs. While investigating properties of LMs and how they change with size holds scientific interest, and large LMs have shown improvements on various tasks (§2), we ask whether enough thought has been put into the potential risks associated with developing them and strategies to mitigate these risks.

We first consider environmental risks. Echoing a line of recent

Biased predictions

Discussion

- What other kinds of biases might arise in supervised text classification models?
- What are the implications of these results for social scientific research?

Mitigating bias

Causes

- Sampling
- Data annotation
- Modeling

Mitigating bias

Sampling

- Buolamwini and Gebru 2018 argue that bias results underrepresentation of certain groups in training data (*undersampling*)
- Davidson et al. 2019 argue that bias in hate speech detection systems is due to the overrepresentation of AAE in training data (*oversampling*)
- This is also a consequence of *class imbalance*, as Dixon et al. 2017 demonstrate
  - Even if we have a representative dataset we might have differences between groups with respect to the annotations.
- Bender et al. 2021 caution that large samples do not guarantee diversity or reduce the risk of bias

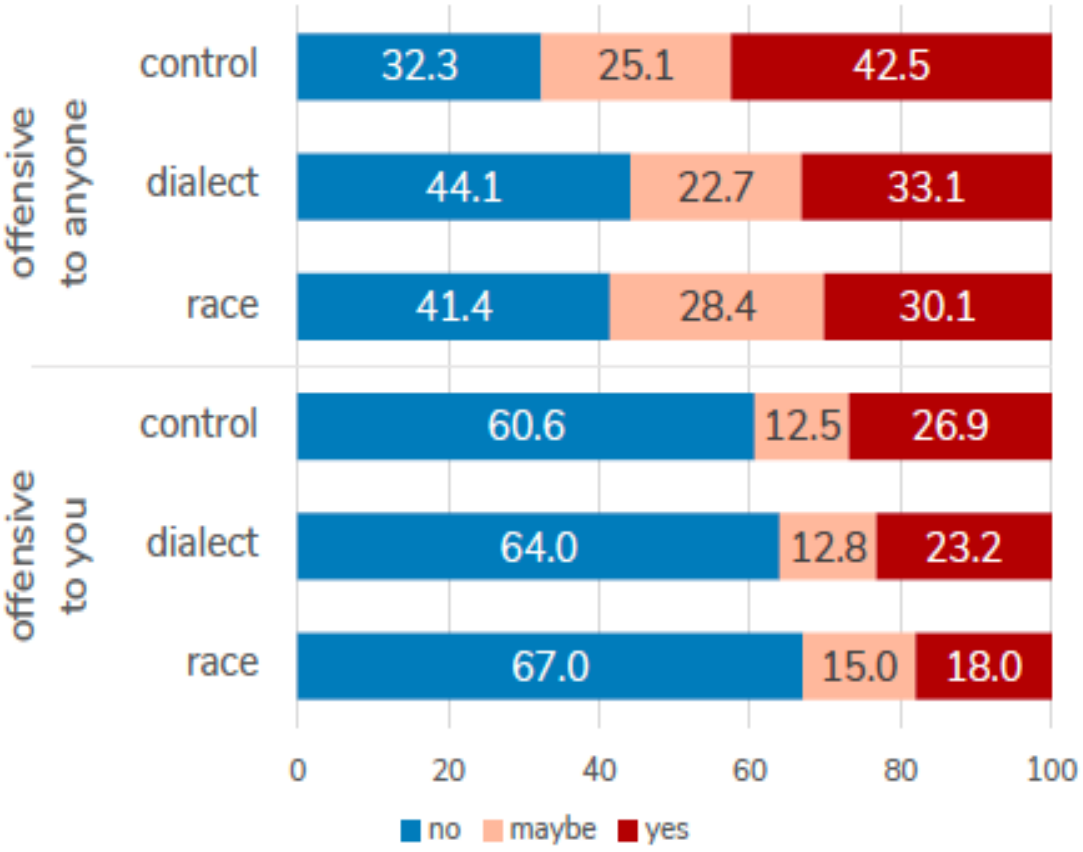
Mitigating bias

Sampling

- Solutions
  - Develop a clear sampling frame
  - Account for subgroup differences
    - e.g. Are certain groups more active and thus likely to be oversampled?
  - Consider the pitfalls of using simple heuristics for sampling
    - e.g. Many hate speech datasets use keyword sampling as an initial step

Mitigating bias

Data annotation



Sap, Maarten, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. "The Risk of Racial Bias in Hate Speech Detection." *ACL*.

Mitigating bias

Data annotation

- Define a relevant population of annotators
- Develop procedures to identify potential biases and to train annotators accordingly
- Experiment with different annotation protocols

Mitigating bias

Modeling

- Dixon et al. 2017 show how adjusting training data to account for class imbalance can reduce bias
- Many computer scientists are working on adjustments for bias at the modeling stage
  - But Gonen and Goldberg 2019 argue that bias reduction methods for word embeddings can be akin to putting "lipstick on a pig" because they simply conceal bias, rather than removing it.

Mitigating bias

"Data cascades"

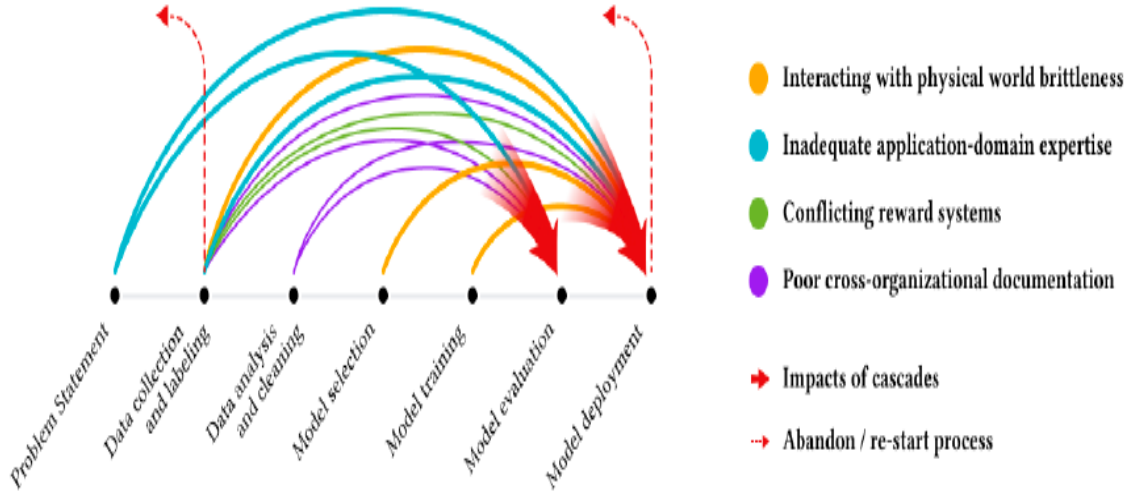


Figure 1: Data cascades in high-stakes AI. Cascades are opaque and protracted, with multiplied, negative impacts. Cascades are triggered in the upstream (*e.g.*, data collection) and have impacts on the downstream (*e.g.*, model deployment). Thick red arrows represent the compounding effects after data cascades start to become visible; dotted red arrows represent abandoning or re-starting of the ML data process. Indicators are mostly visible in model evaluation, as system metrics, and as malfunctioning or user feedback.

Sambasivan, Nithya, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora Aroyo. 2021. "Everyone Wants to Do the Model Work, Not the Data Work": Data Cascades in High-Stakes AI." *CHI*.

Concluding remarks

Directions for sociological research

- What causes bias in predictive models?
- How can we assess bias in existing systems?
- What are the implications of this bias?
- How can we make less biased systems?

Summary

- The results of the Fragile Families Challenge show how social outcomes are difficult to predict
  - Simple models perform as well as complex ones
  - Unclear whether outcomes inherently unpredictable or if we are not measuring them well enough
- Bias in machine learning
  - Evidence of gender and racial bias in image and text classification systems
- Causes of bias
  - Sampling, annotation, modeling
- Scholars are currently working on solutions to these issues at different stages of the research process