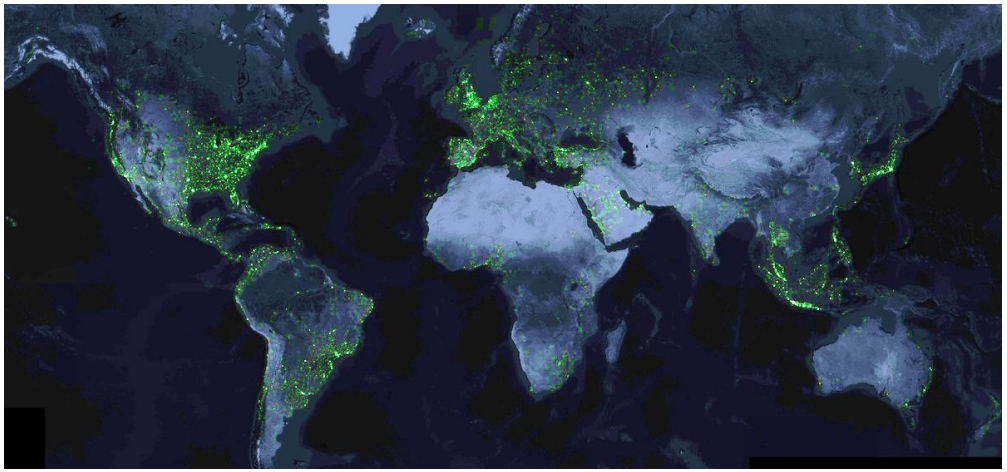
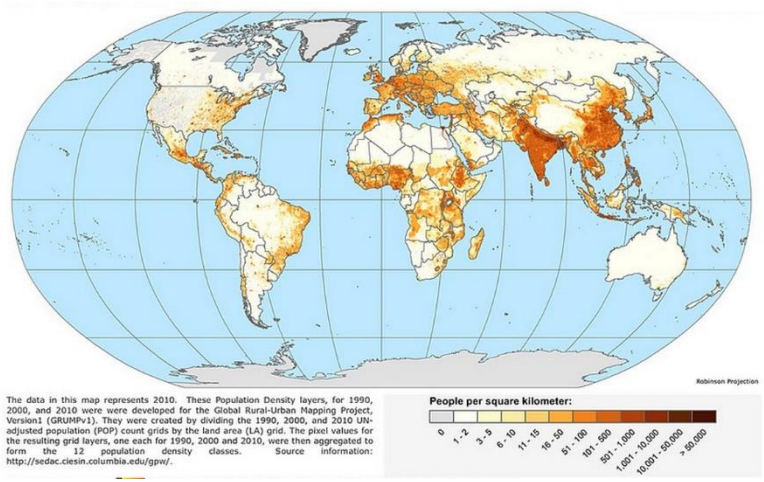
Understanding Demographic Bias and Representation in Social Media Health Data

Nina Cesare *Biostatistics & Epidemiology Data Analytics Center, Boston University*
Elaine Nsoesie *Department of Global Health, Boston University*
Christan Grant *Department of Computer Science, University of Oklahoma*

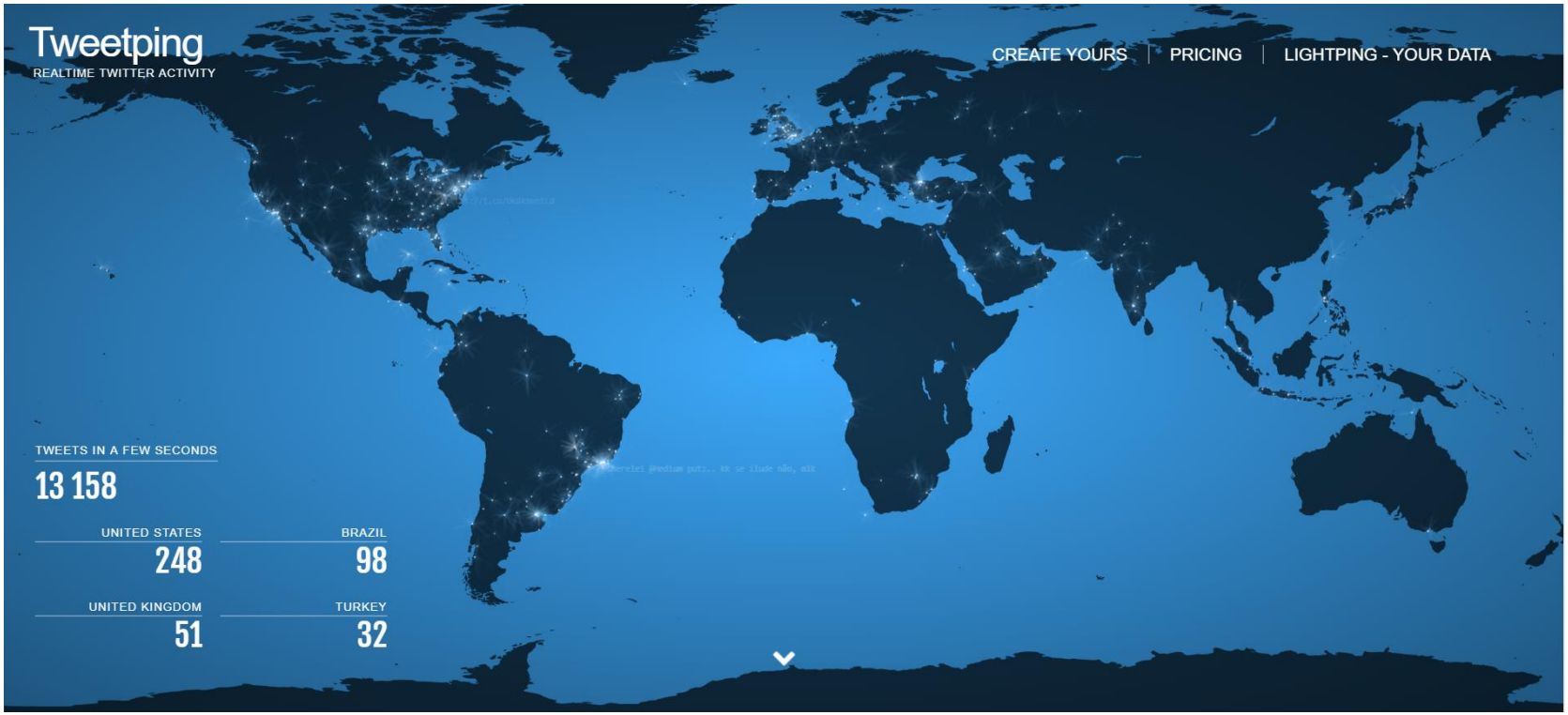
Social Media for Population Studies





Every geotagged tweet from mid 2011- 2014, compared with 2010 global population density. Map by Eric Fischer. Source: https://blog.mapbox.com/making-the-most-detailed-tweet-map-ever-b54da237c5ac

Social Media for Population Studies



Geotagged tweets visualized in real time via tweetping.net (developer: Franck Ernewein)

Using social media for health research

Using social media for health research

Temporally granular

- Foodborne illness
- Influenza
- Other communicable/infectious disease

Geographically granular

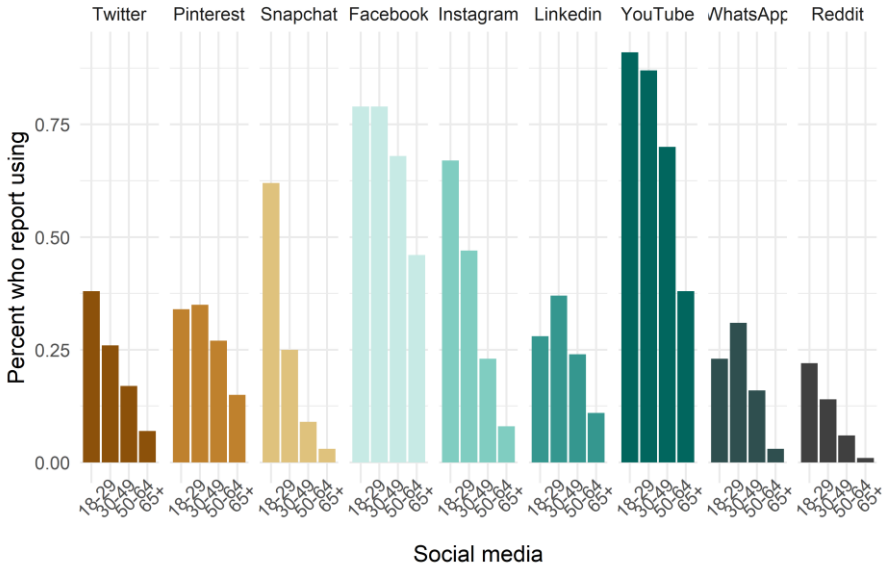
- Obesity
- Physical inactivity
- Other acute/chronic health outcomes

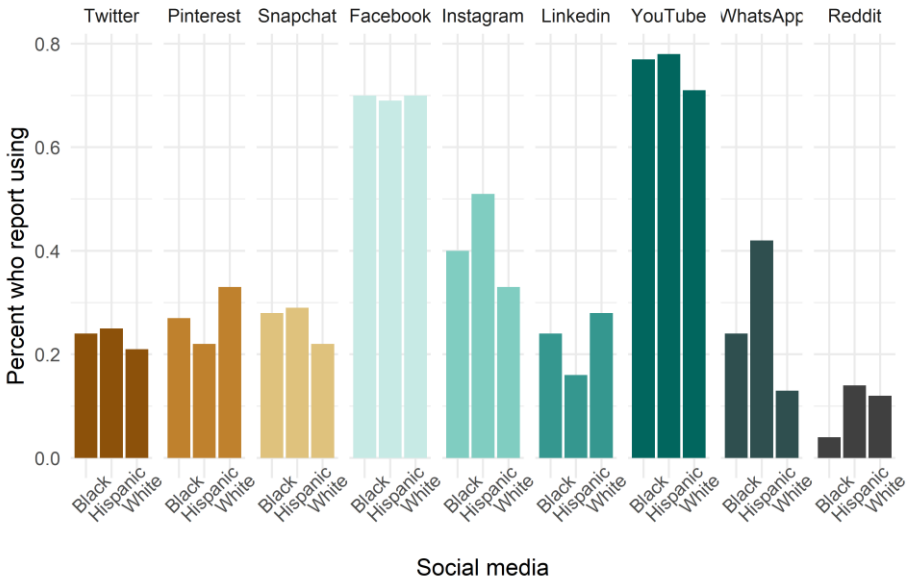
Qualitative detail

- Attitudes toward health services (e.g. vaccines)
- Attitudes toward health behaviors (e.g. physical activity)

Challenges of this work

- Under/over representation of groups across platforms



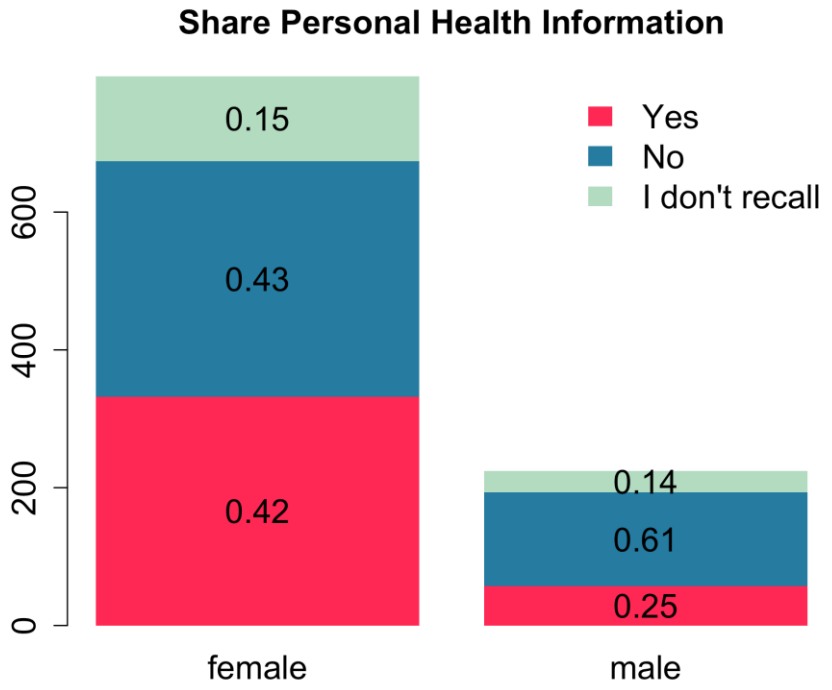


Data from the Pew Research Center Social Media Fact Sheet. Updated June 12th, 2019.
https://www.pewinternet.org/fact-sheet/social-media/

Challenges of this work

Patterns of social media use:

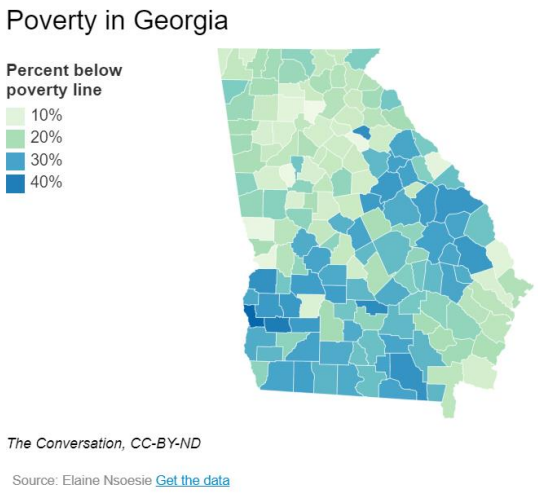
- May vary between demographically divided peer groups
- May be driven by unobserved social/cultural expectations

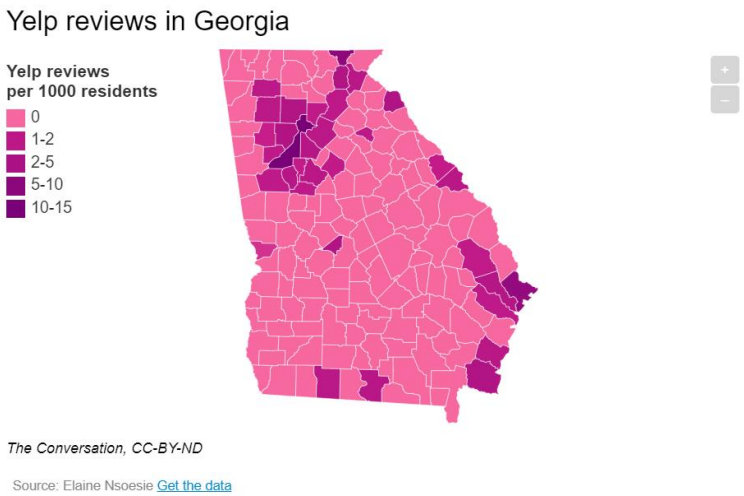


Results from a survey distributed by Nsoesie et al. via Pollfish. Results not nationally representative.

Demographic bias within data

- Sampling from specific geographies != capturing communities of interest

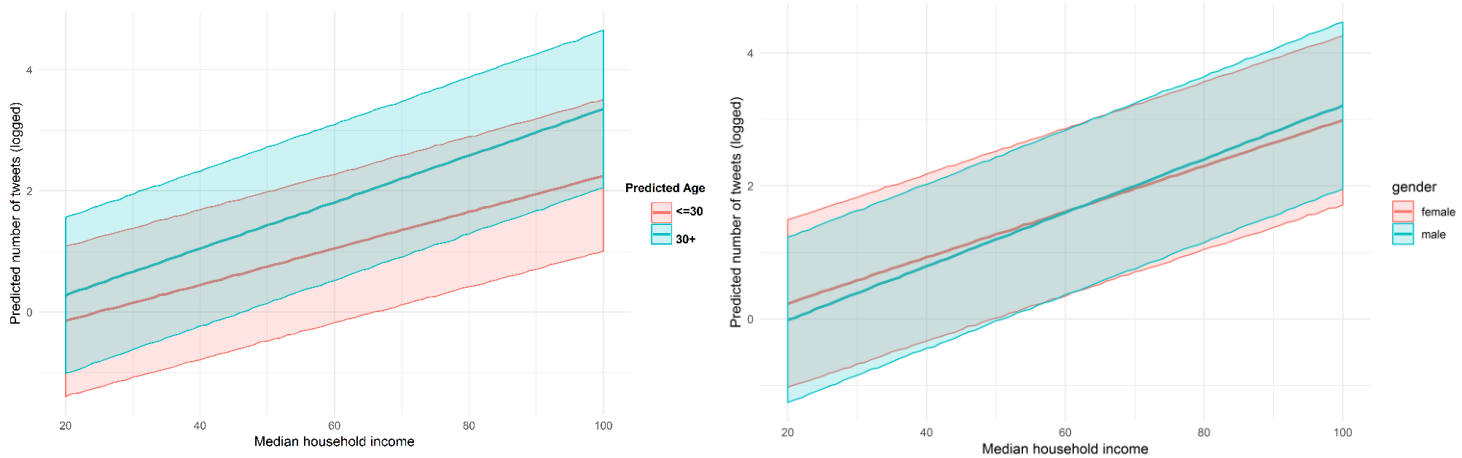




Source: Elaine Nsoesie. “Social media helps officials spot public health threats - but only for the rich?” *The Conversation.* http://theconversation.com/social-media-helps-officials-spot-public-health-threats-but-only-for-the-rich-76364

Demographic bias within data

- Sampling from specific geographies != capturing communities of interest



Predicted number of observed foodborne illness tweets per county, varying median income, estimated demographics of Twitter users

Demographic bias within data

- Bias in actual sample composition
- Bias in estimated sample composition

	All	Female	Male	Darker	Lighter	DF	DM	LF	LM
MSFT									
Error Proportion	-	76.9%	23.1%	93.6%	6.4%	70.5%	23.1%	6.4%	0.0%
Error Count	78	60	18	73	5	55	18	5	0
FACE++									
Error Proportion	-	95.9%	4.1%	74.0%	26.0%	72.4%	1.6%	23.6%	2.4%
Error Count	123	118	5	91	32	89	2	29	3
IBM									
Error Proportion	-	74.7%	25.3%	85.7%	14.3%	61.0%	24.7%	13.6%	0.6%
Error Count	154	115	39	132	22	94	38	21	1

Table 1: The error proportion and count contribution of each subgroup to the overall error rates for the 3 evaluated commercial classifiers.

Source: Joy Buolamwini and Timnit Gebru. “Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Published in FAT 2018.

Demographic bias within data

"In computer science literature, bias is defined in terms of the dataset—bias in labeling, bias in sample selection, bias in the task, or bias in model structure, favoring certain types of error over others….In medical and social science research, bias has been defined as any tendency that prevents unprejudiced consideration of a question or advances prejudice in favor of or against one group compared with another."

- Kadija Ferryman and Mikaela Pitcan. *Fairness in Precision Medicine.* Data & Society

Demographics, Ethics and Privacy

Data-driven approaches toward estimating, controlling for demographic bias within samples

What **can** we estimate, and what **should** we estimate?

What is acceptable/ethical data use?

Moving forward

- Interdisciplinary collaboration
  - Researchers developing prediction models + domain experts + ethics experts = more reliable, ethical work?
- Involve stakeholders
  - Take administrative steps to ensure people are in control of their own metadata
  - Small-scale surveys of social media preference, comfort with domain specific research

Moving forward

- Keep the conversation rolling!
- Interpret bias broadly!
- Hand the microphone to marginalized groups!
- Your ideas?...

Thank you!

Contact: ncesare@bu.edu
@nlcesare