ODI Workshop

November 22, 2022

The Opportunities of Big Data in Social Science Research

Milena Tsvetkova

London School of Economics and Political Science

Outline

1. Overview
    1. Big data
    2. Data science methods
2. Examples
    1. Contagion and epidemics
    2. Political misinformation and polarization
    3. Poverty and inequality
3. Demo: Measuring socioeconomic differences in consumption patterns using the Facebook Marketing API
4. Problems, debates, and next steps

Big Data

- Big data
  - High volume
  - High velocity
  - High variability/complexity
- New formats
  - Networks
  - Text
  - Images
  - Audio
  - Video

Big Data

- New sources
  - Digital trace data
  - Satellite imagery
  - Digitized historical archives
  - Census/register data
  - A/B testing
  - Crowdsourced data donation
  - Browser tracking data
  - Surveillance cameras

Data Science Methods

- Collecting data
  - Web scraping
  - APIs (Application Programming Interface)
- Storing and managing data
  - Relational databases – e.g., SQL
  - Non-relational databases – document/file collections
- Processing data
  - Cleaning – e.g., lemmatization
  - Labeling – e.g., crowdsourced image tagging and tweet classification
  - Matching – e.g., named entity recognition with regular expressions

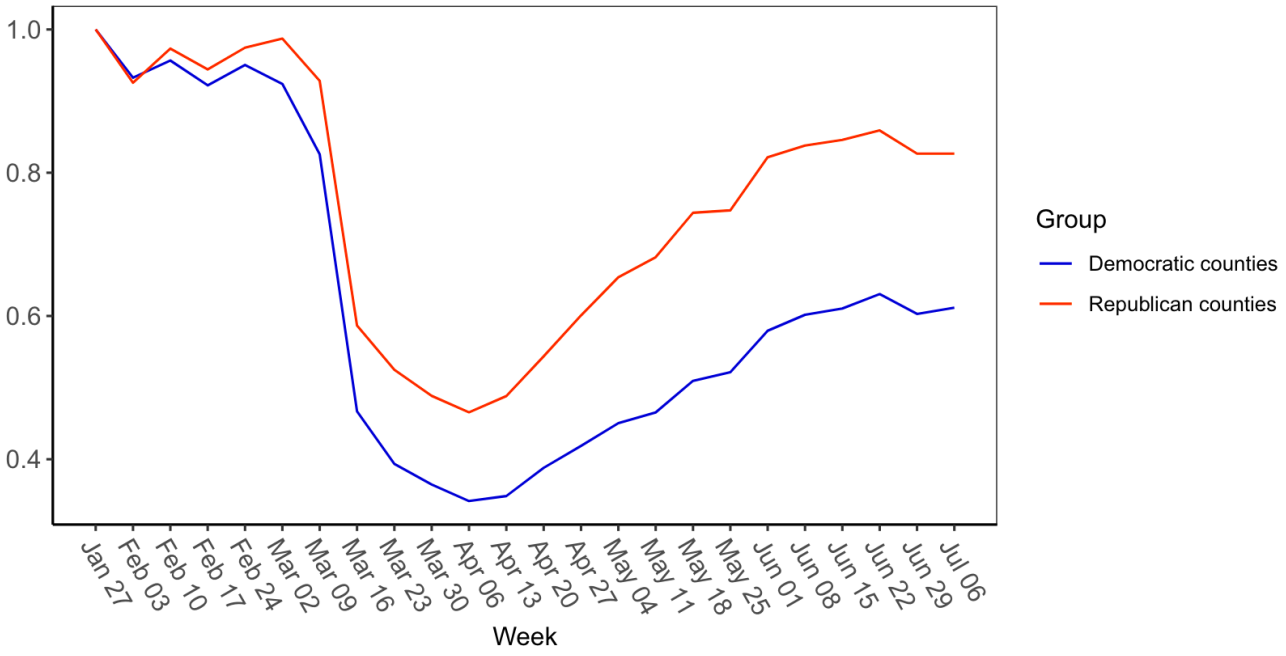
Data Science Methods

- Analyzing and visualizing data
  - **Data mining** – discovering patterns
  - **Machine learning** – making predictions
  - **Causal inference** – identifying causal relationships
  - **Natural language processing** – understanding text
  - **Network analysis** – understanding relational patterns

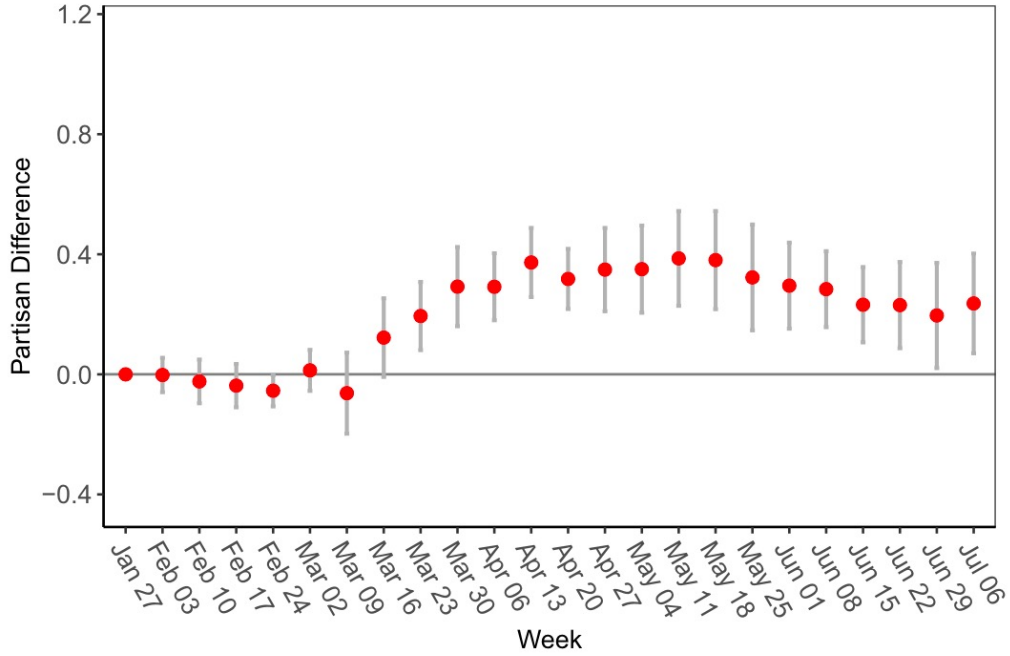
Example: Contagion and Epidemics

- Were there partisan differences in social distancing during COVID?
- Mobile location data from SAFEGRAPH

POI visits



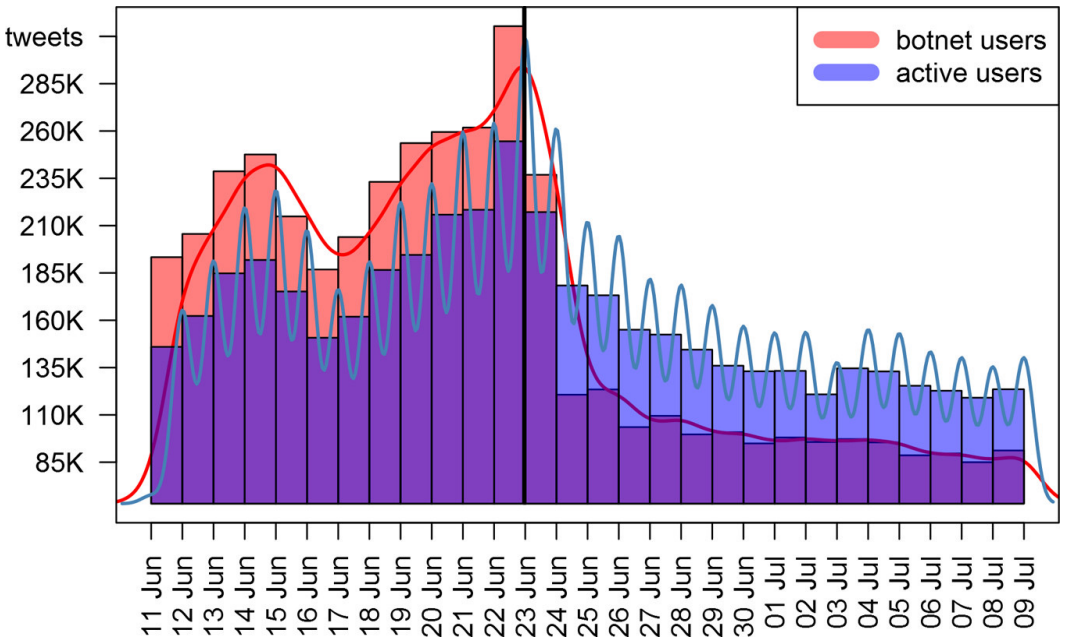
Estimated coefficient for effect of Republican vote share in county for log(POI visits)

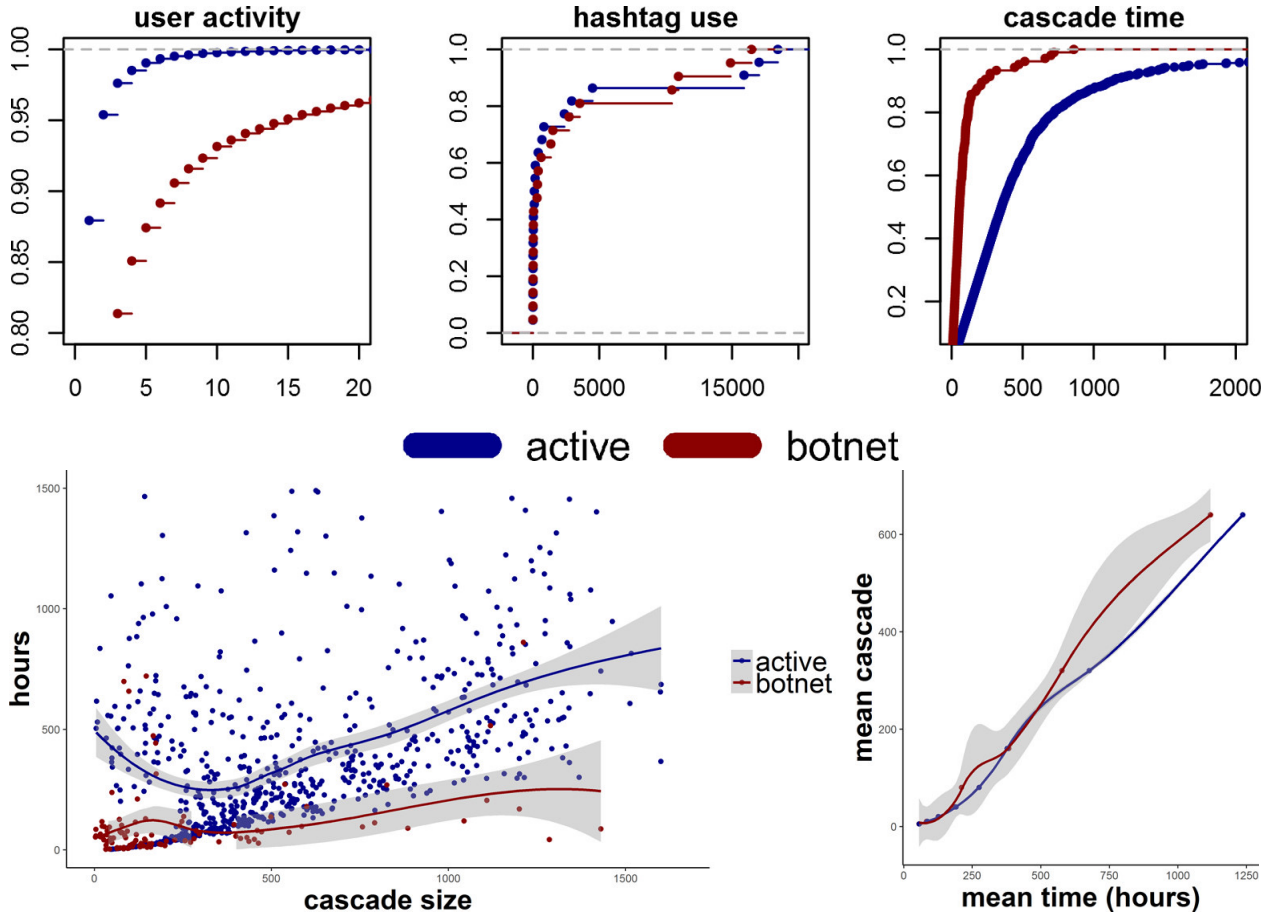


Allcott, H., Boxell, L., Conway, J., Gentzkow, M., Thaler, M., & Yang, D. (2020). Polarization and public health: Partisan differences in social distancing during the coronavirus pandemic. *Journal of Public Economics*, 191, 104254.

Example: Political Misinformation

- Can bots on Twitter influence political opinion?

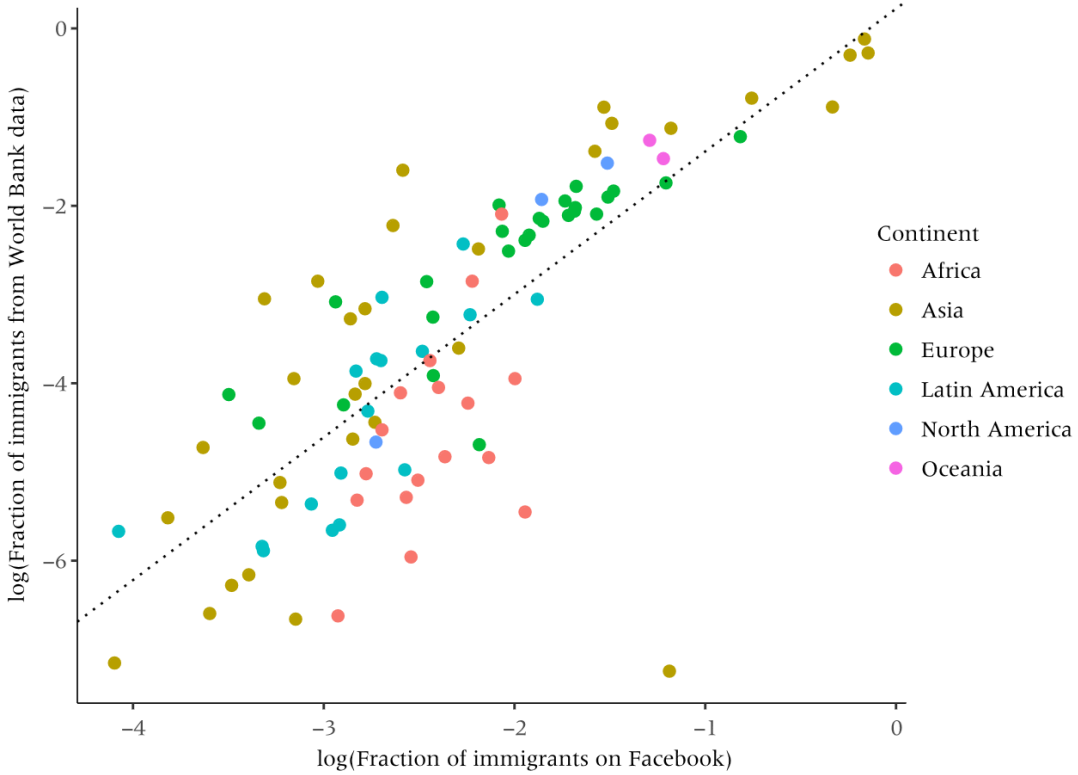


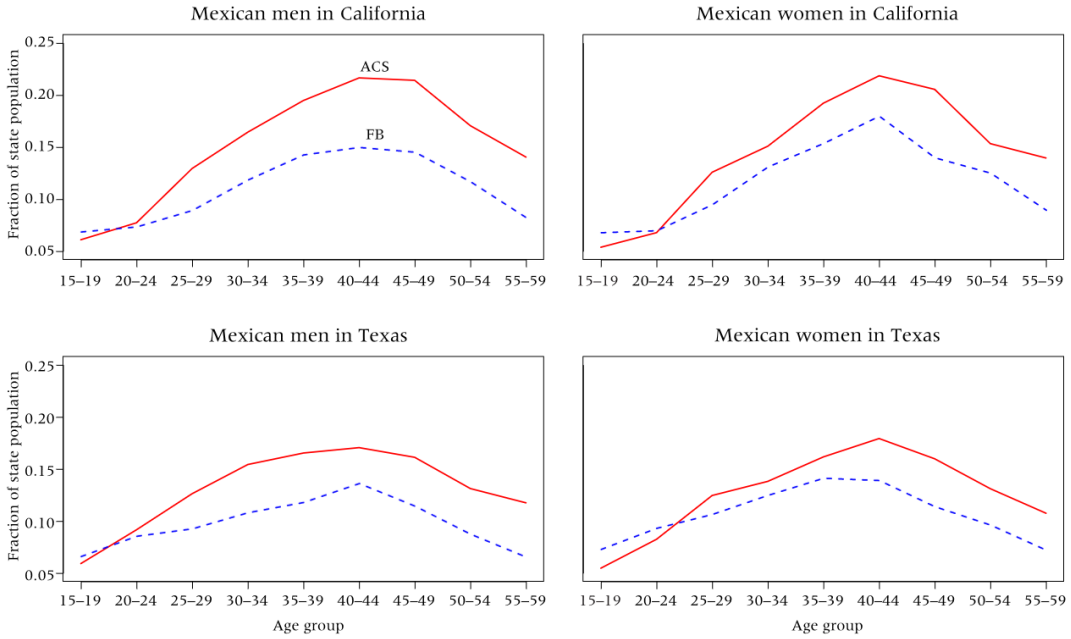


Bastos, M. T., & Mercea, D. (2019). The Brexit botnet and user-generated hyperpartisan news. *Social Science Computer Review*, 37(1), 38–54.

Example: Poverty and Inequality

- Can we use Facebook ad data to estimate stocks of migrants within and across countries?



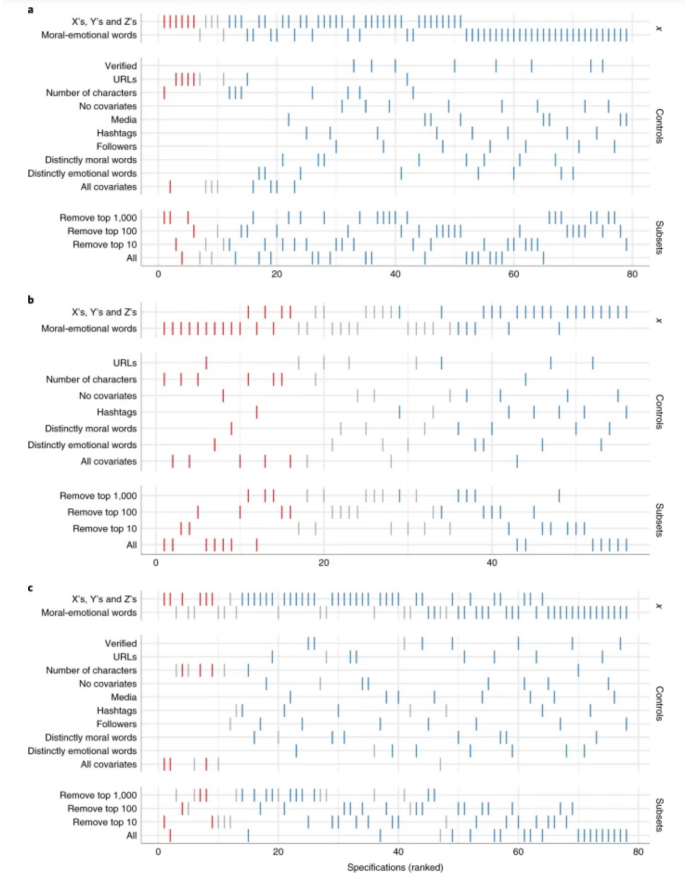


Zagheni, E., Weber, I., & Gummadi, K. (2017). Leveraging Facebook's advertising platform to monitor stocks of migrants. *Population and Development Review*, 721-734.

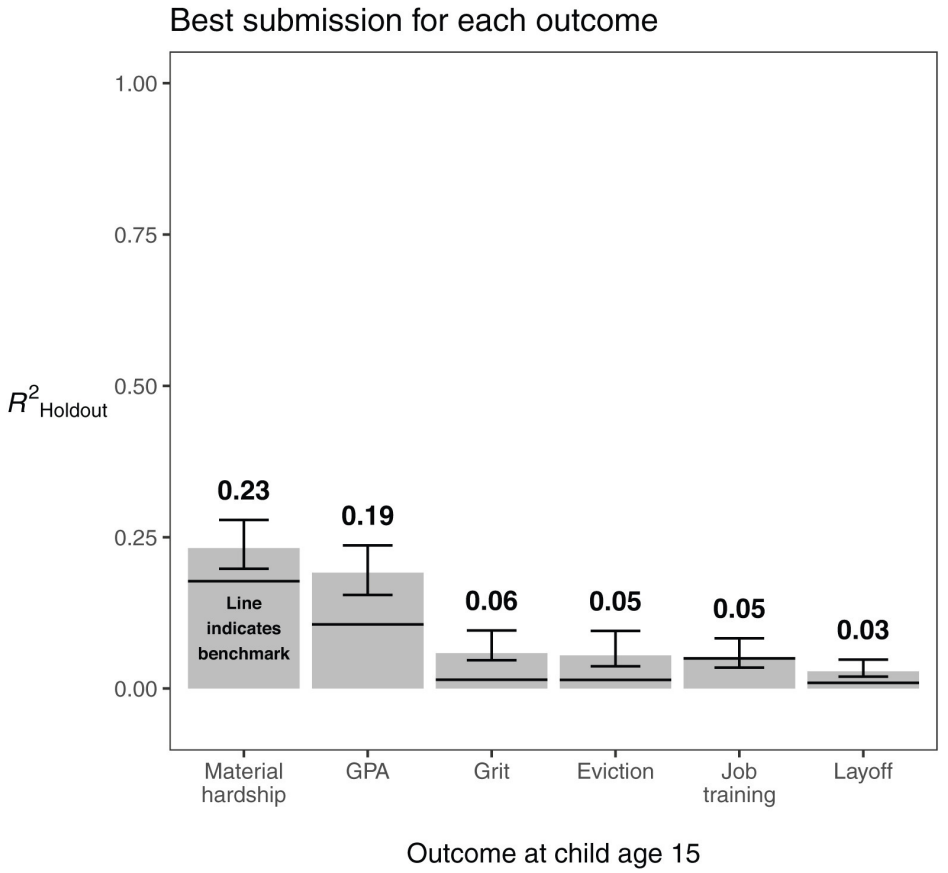
Demonstration

- View/work locally
  - Install python and VS Code on your machine
  - Go to https://github.com/social-research/big-data-workshop and download the repository
  - Open the (unzipped) folder in VS Code
- Alternatively, view/work remotely
  - Go to https://colab.research.google.com/github/social-research/big-data-workshop

Problems and Debates: Reproducibility



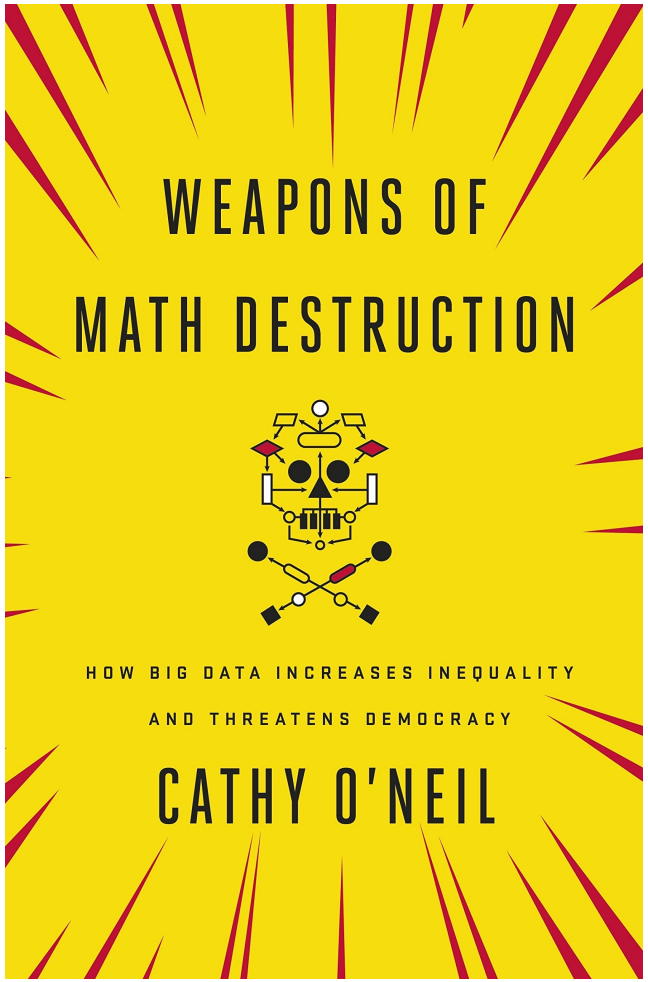
a, COVID-19 corpus. **b**, #MeToo corpus. **c**, #MuellerReport corpus. Each possible model specification (*x* axis) is represented by three vertically aligned points corresponding to the outliers removed and covariates and independent variable included in the negative binomial regression equation (*y* axis). Red indicates a significant ($P < 0.05$) negative regression coefficient, grey indicates a non-significant coefficient and blue indicates a significant positive coefficient.



Burton, J. W., Cruz, N., & Hahn, U. (2021). Reconsidering evidence of moral contagion in online social networks. *Nature Human Behaviour*, 5(12), 1629-1635.

Salganik, M. J., et al. (2020). Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences*, 117(15), 8398-8403.

Problems and Debates: Accountability



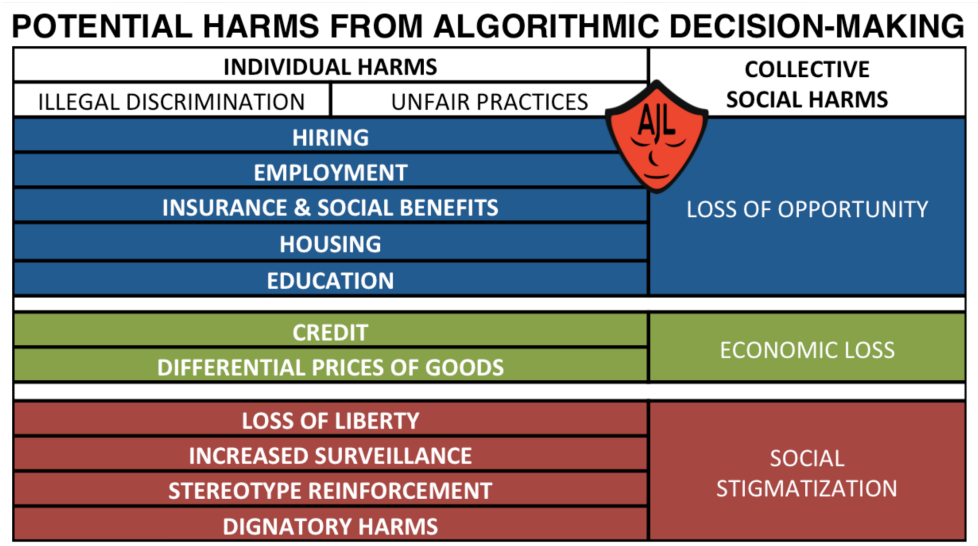
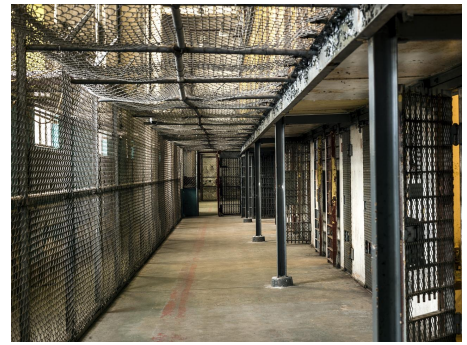
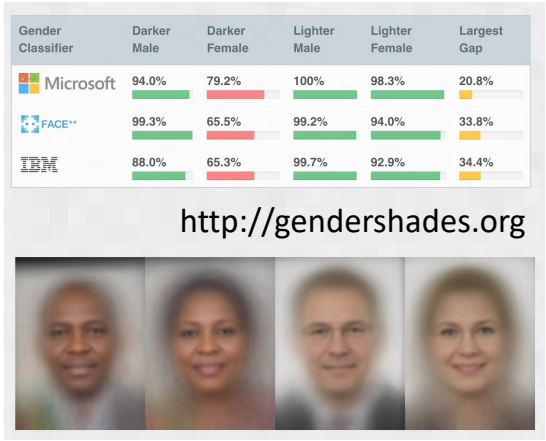


Chart Contents Courtesy of Megan Smith, Former CTO of the United States

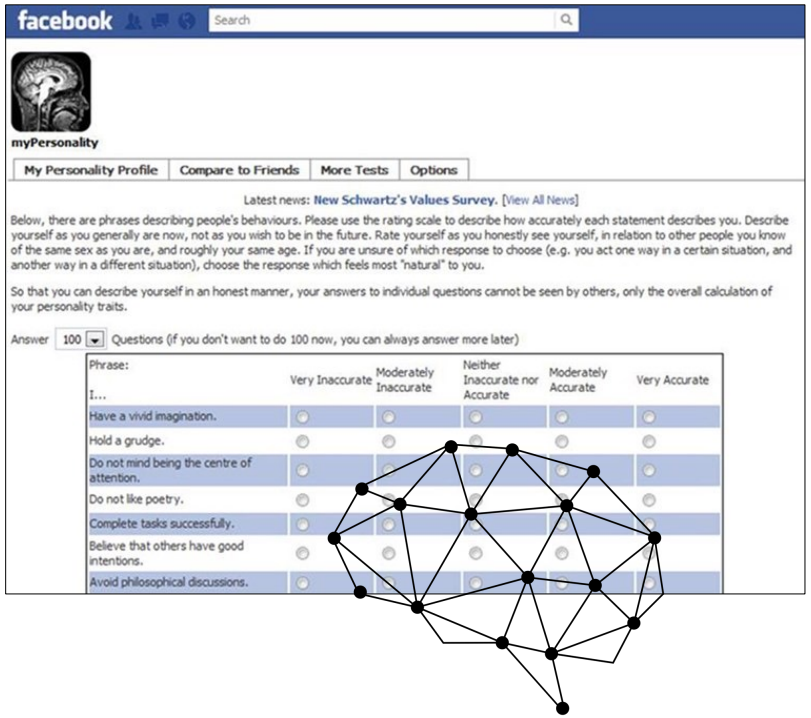




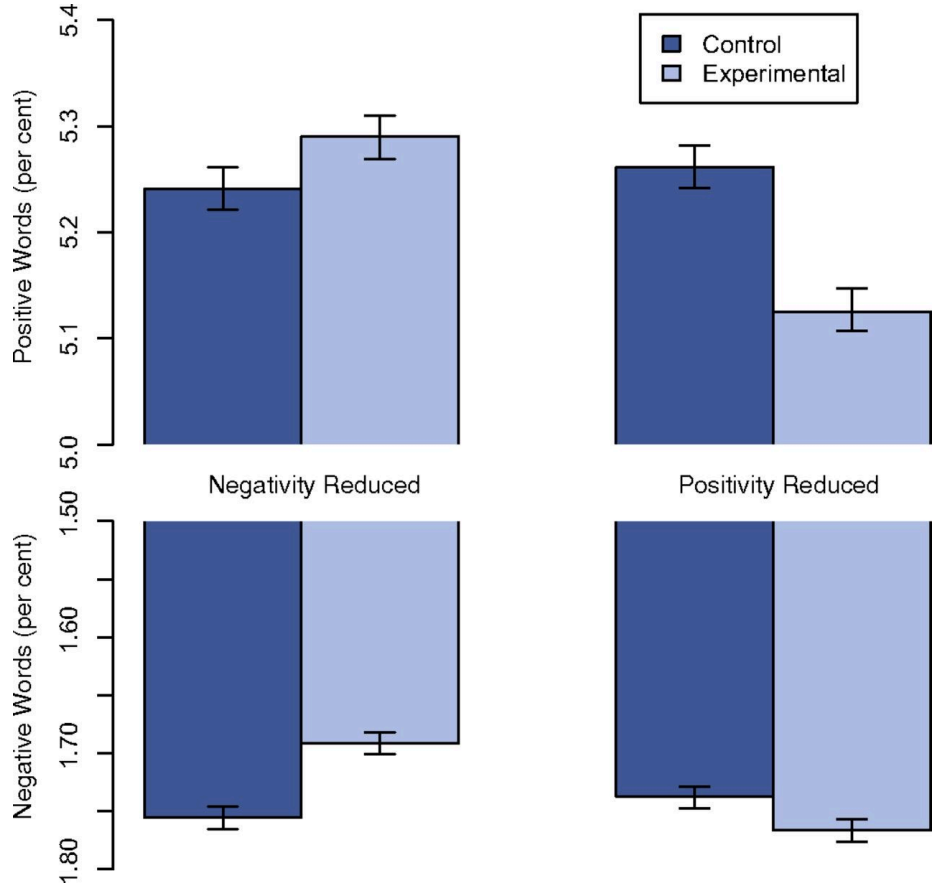
O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Broadway Books.

Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency* (pp. 77-91). PMLR.

Problems and Debates: Research Ethics



Cambridge Analytica



Cadwalladr, C., & Graham-Harrison, E. (2018). Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. *The Guardian*, 17, 22.

Kramer, A. D., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, *111*(24), 8788-8790.

Next Steps

- **Summer course**: LSE ME314: Introduction to Data Science and Machine Learning (taught in R)
- **Online program**: LSE Data Analytics Career Accelerator
- **MSc program**: LSE MSc in Applied Social Data Science
- **Books**
  - Salganik, M. J. (2019). *Bit by Bit: Social Research in the Digital Age*. Princeton University Press.
  - Hogan, B. (2022). *From Social Science to Data Science: Key Data Collection and Analysis Skills in Python*. Sage.
- **Conference**: International Conference on Computational Social Science