Digital Trace Data for Social Research

Ridhi Kashyap

Department of Sociology
Leverhulme Centre for Demographic Science
Nuffield College
University of Oxford

SICSS-Duke
June 21, 2022

The Digital Revolution

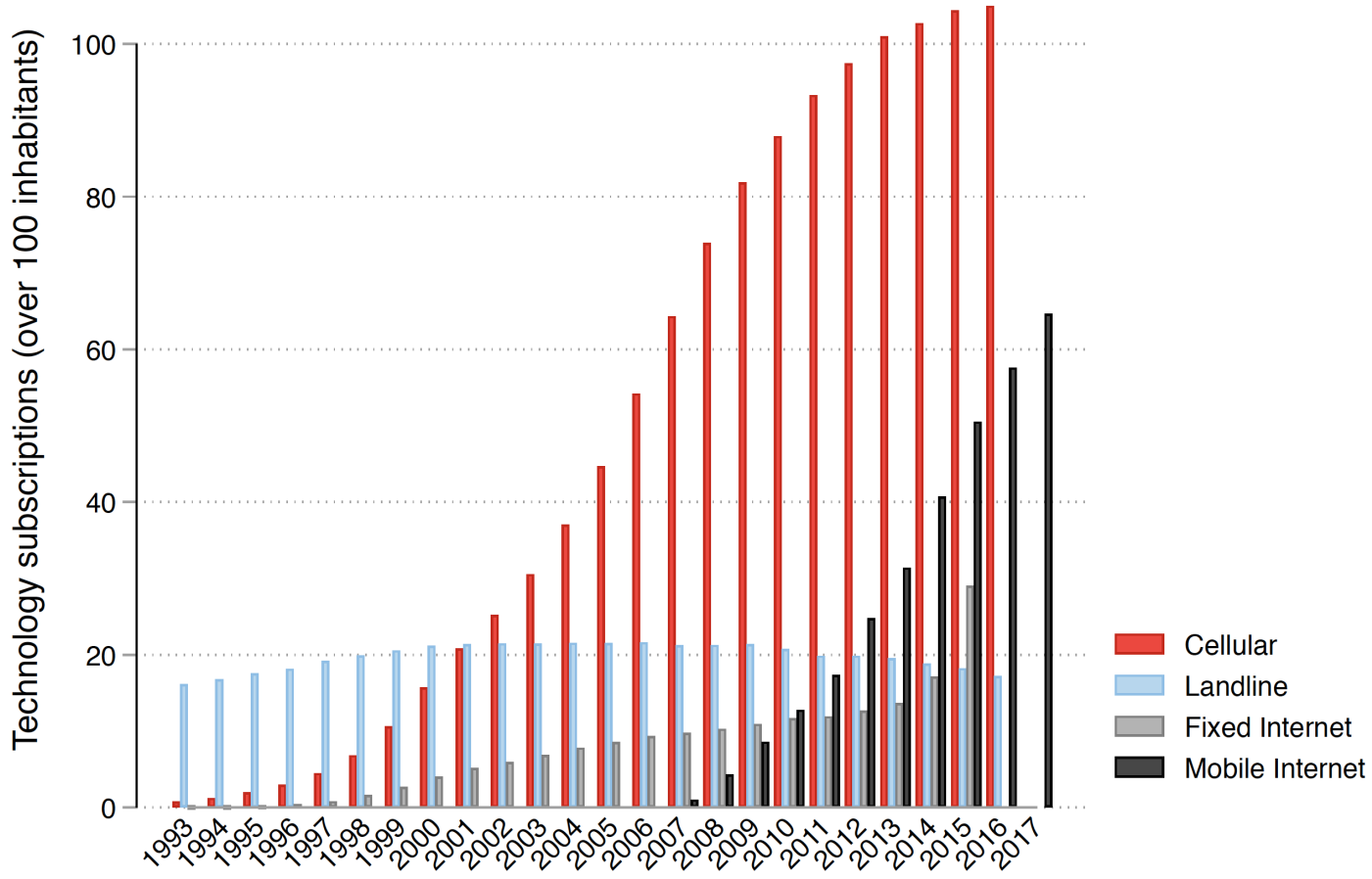


Figure: Technology subscriptions per capita, 1993–2017. Data from International Telecommunications Union (ITU).

The Digital Revolution

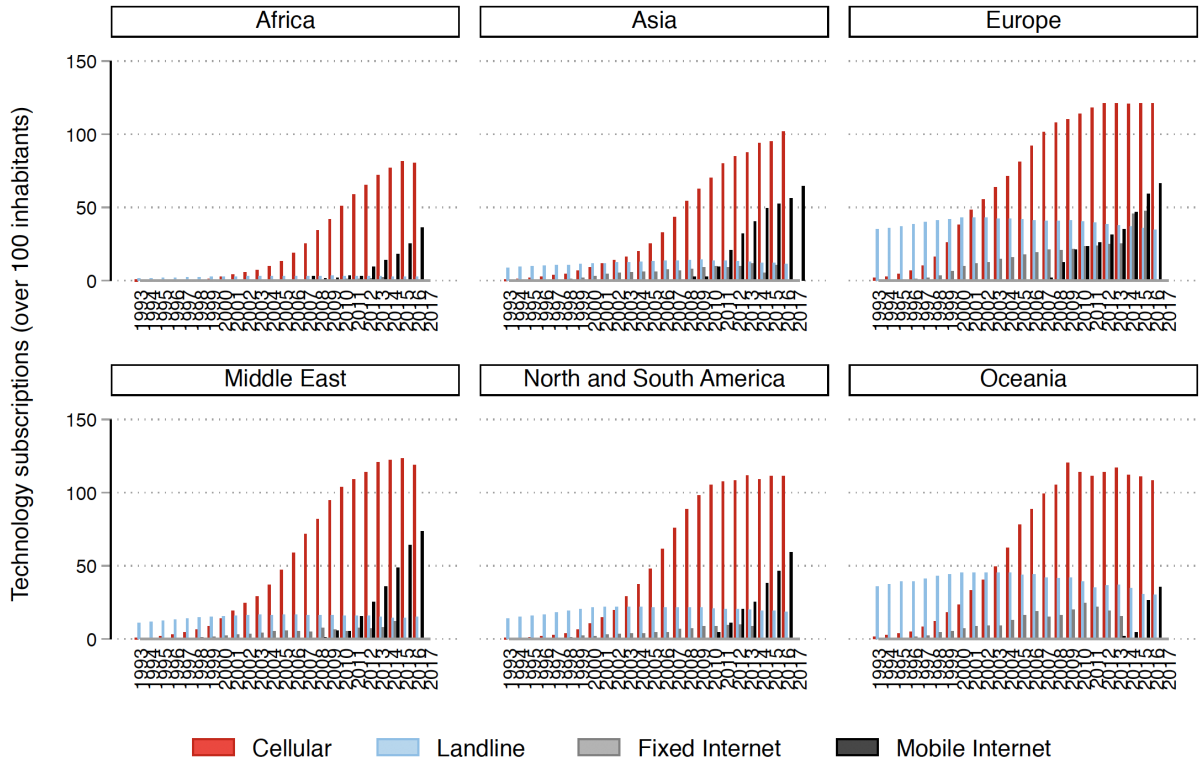


Figure: Technology subscriptions per capita, by global regions, 1993–2017. Data from International Telecommunications Union (ITU).

The Digital Revolution

- **Social transformation:**
  - Digital technologies permeate social, economic, political life.
- **Data revolution:**
  - The digitalisation of our lives creates data by-products: **digital trace data**

A Changing Data Ecosystem

- Digital trace data emerge from two processes –
- Social life is **digitally mediated** and the adoption of digital technologies and platforms (e.g. social media) generates data
  - Companies that provide these services want to capture these data streams because they are intrinsic to business models, e.g. targeted advertising

A Changing Data Ecosystem

- Digital trace data emerge from two processes –
- Social life is **digitally mediated** and the adoption of digital technologies and platforms (e.g. social media) generates data
  - Companies that provide these services want to capture these data streams because they are intrinsic to business models, e.g. targeted advertising
- Digitalisation of data has resulted in the **storage** of diverse types of information – including about **non-digital or offline life**
  - Information linked to everyday activities, e.g. sensors, consumer transactions, video recordings of cities are digitally stored

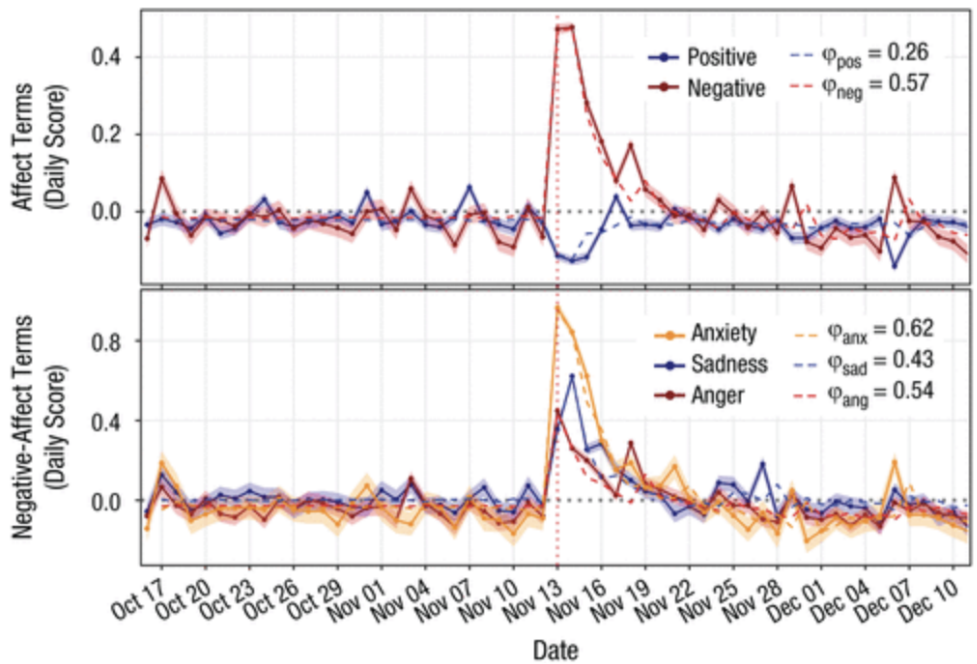
Common Features of Digital Trace Data

- Data that require **repurposing** because they were not intentionally collected for research
- **Non-reactive:** provide opportunities to observe without asking and for more dynamic measurement
- Measures of behaviour or activity (contrast with self-reported)
  - Ethics
  - Accessibility
- Digital trace data come in many forms

Digital Trace Data: Examples

- Social media sites

Twitter



[1] Garcia, David, and Bernard Rime. "Collective emotions and social resilience in the digital traces after a terrorist attack." *Psychological science* (2019): 0956797619831964.

Facebook and Gender Gaps

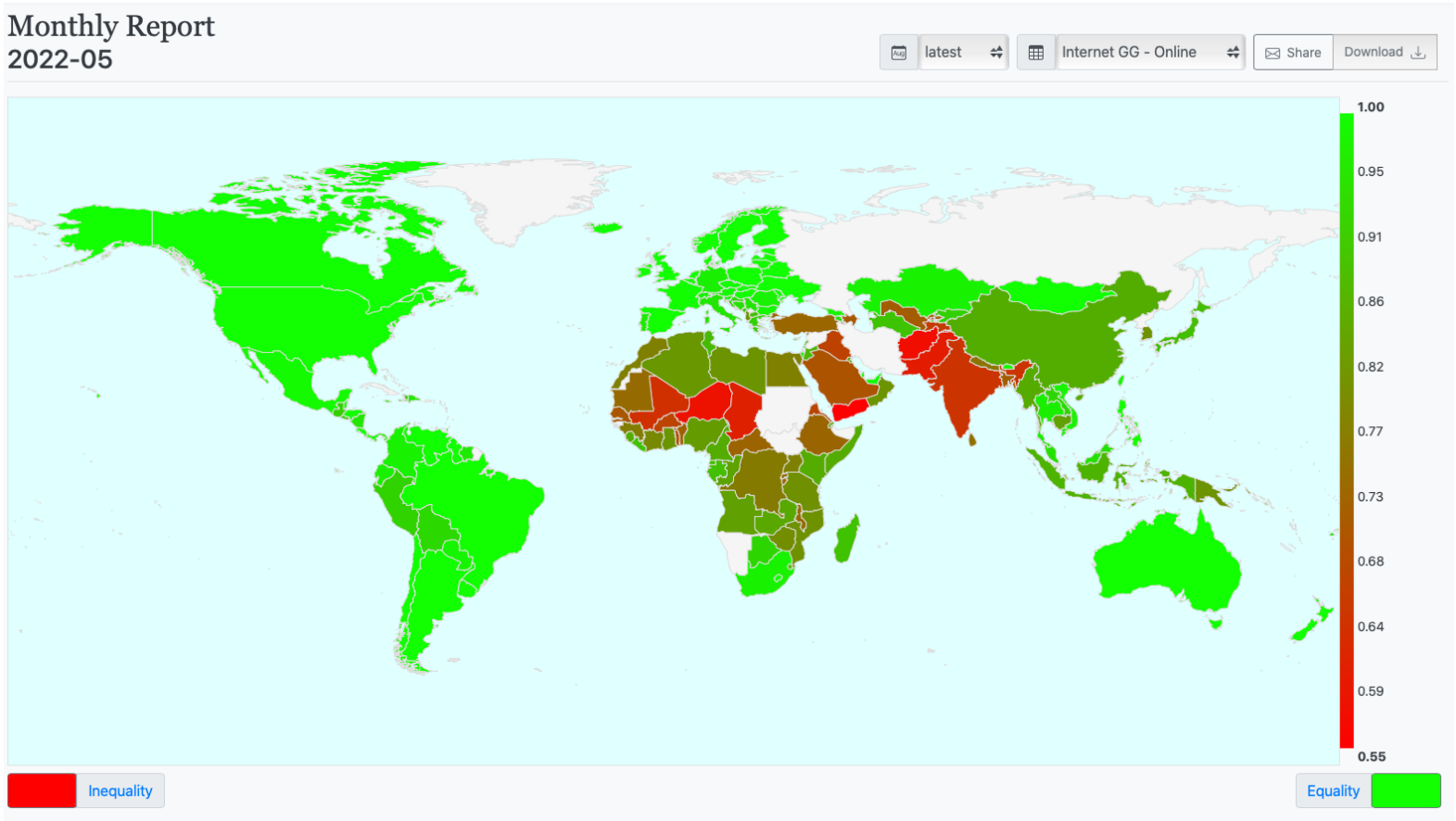


Figure: Gender gaps in internet use computed using data from Facebook (online model) available at `www.digitalgendergaps.org`

[1] Fatehkia, Masoomali, Ridhi Kashyap, and Ingmar Weber. "Using Facebook ad data to track the global digital gender gap." *World Development* 107 (2018): 189-209.

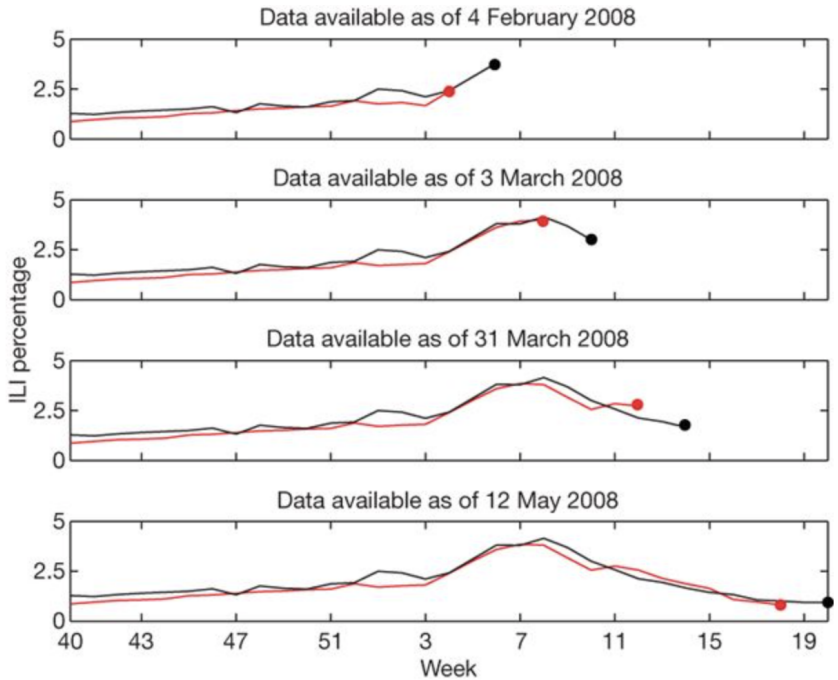
Digital Trace Data: Examples

- Social media sites
- Web search queries

Google Flu

Figure 3 : **ILI percentages estimated by our model (black) and provided by the CDC (red) in the mid-Atlantic region, showing data available at four points in the 2007-2008 influenza season.**

From: Detecting influenza epidemics using search engine query data



[1] Ginsberg, Jeremy, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant. "Detecting influenza epidemics using search engine query data." *Nature* 457, no. 7232 (2009): 1012.

Google Search and IPV

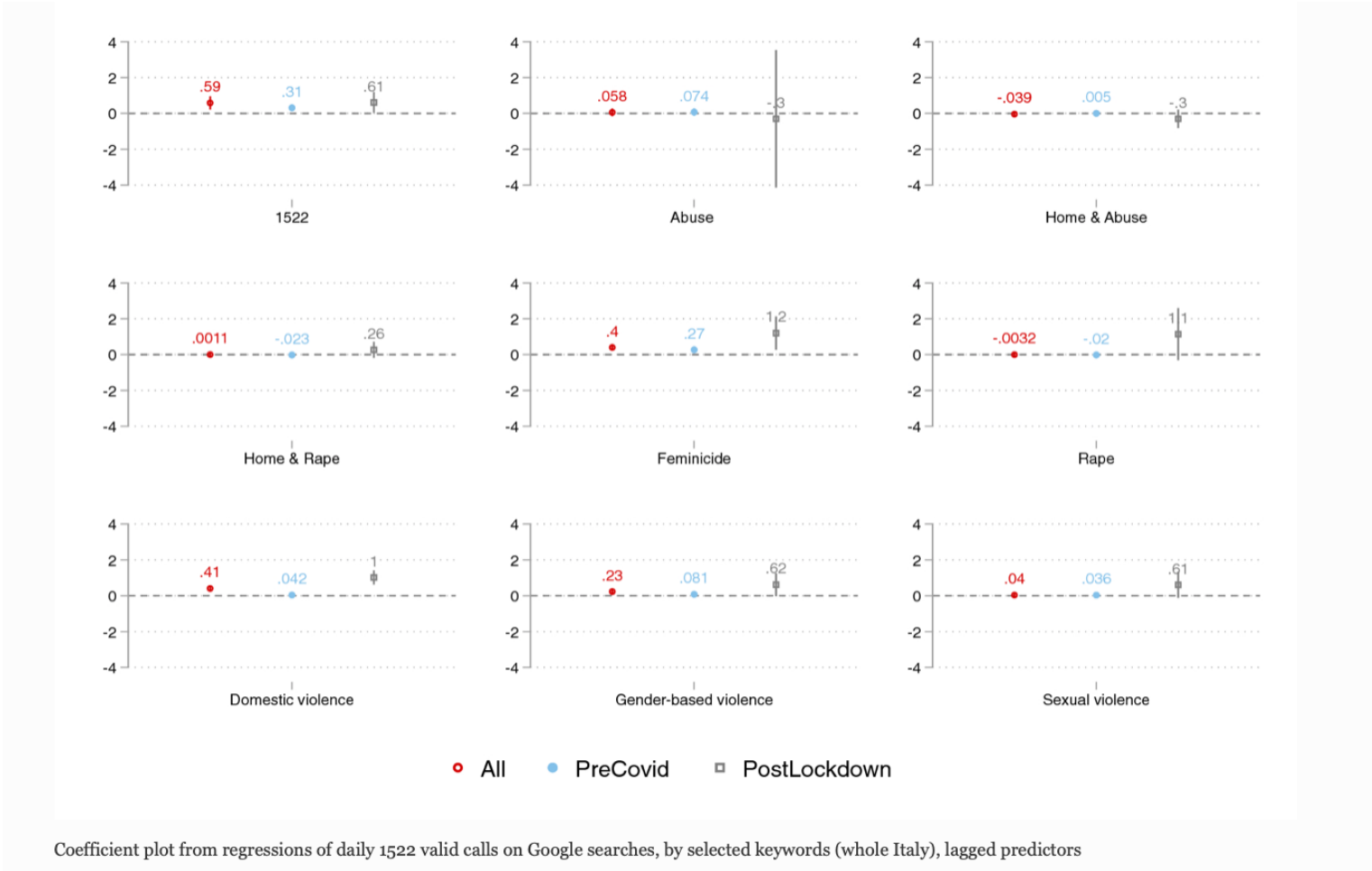


Figure: Relationship between abortion search volume and abortion rates across 37 countries. Marker size indicates the number of restrictions (from 0 to 7) that exist on abortion in each country.

---

[1] Köksal, Selin, Luca Maria Pesando, Valentina Rotondi, and Ebru Şanlıtürk. "Harnessing the Potential of Google Searches for Understanding Dynamics of Intimate Partner Violence Before and After the COVID-19 Outbreak." European Journal of Population (2022): 1-29.

Digital Trace Data: Examples

- Social media sites
- Web search queries
- Blogs and internet forums

Reddit

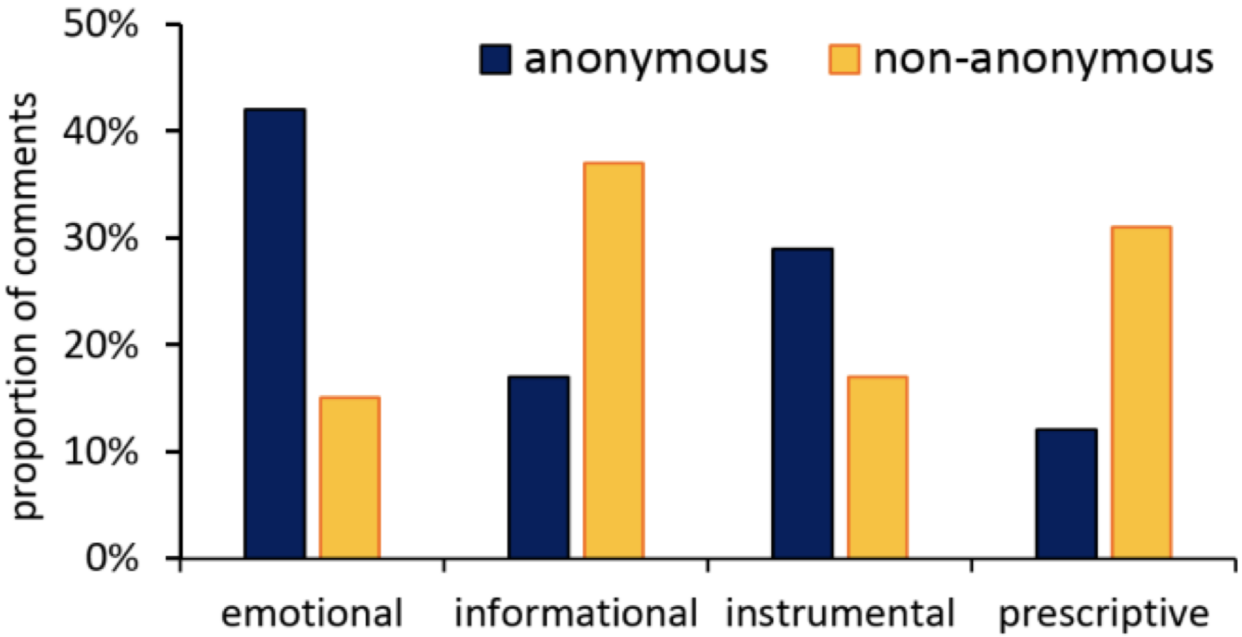


Figure 4. Posts from anonymous and non-anonymous accounts in the light of social support types.

[1] De Choudhury, Munmun, and Sushovan De. "Mental health discourse on reddit: Self-disclosure, social support, and anonymity." In Eighth International AAAI Conference on Weblogs and Social Media. 2014.

Digital Trace Data: Examples

- Social media sites
- Web search queries
- Blogs and internet forums
- Call detail records from mobile phones

Mobile Phones

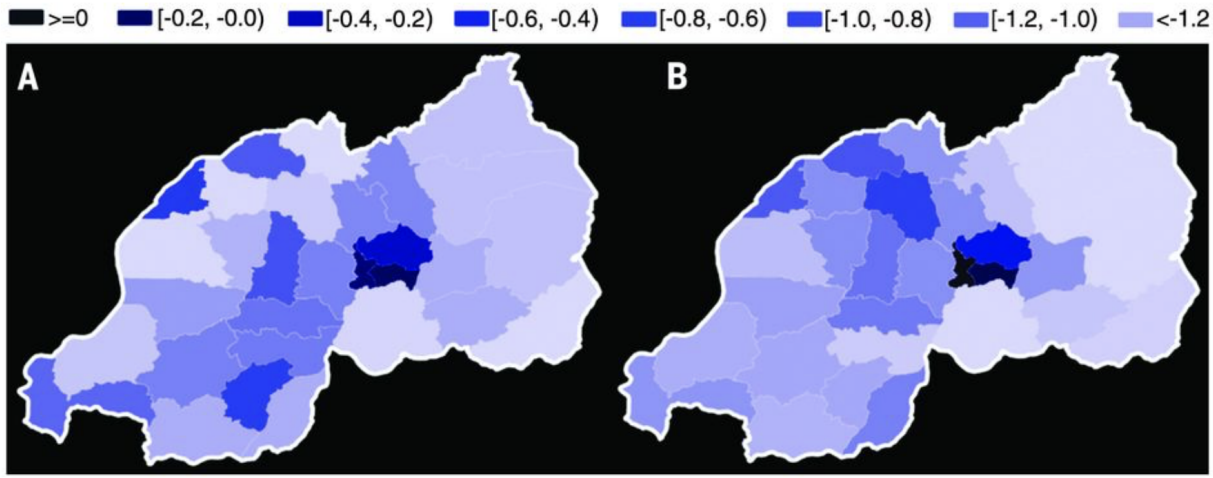


Figure: (A) Predicted composite wealth index (district average), computed from 2009 call data and aggregated by administrative district. (B) Actual composite wealth index (district average), as computed from a 2010 government DHS of 12,792 households.

[1] Blumenstock, Joshua, Gabriel Cadamuro, and Robert On. "Predicting poverty and wealth from mobile phone metadata." *Science* 350, no. 6264 (2015): 1073-1076.

Digital Trace Data: Examples

- Social media sites
- Web search queries
- Blogs and internet forums
- Call detail records from mobile phones
- Sensor data

Electricity Smart Meters

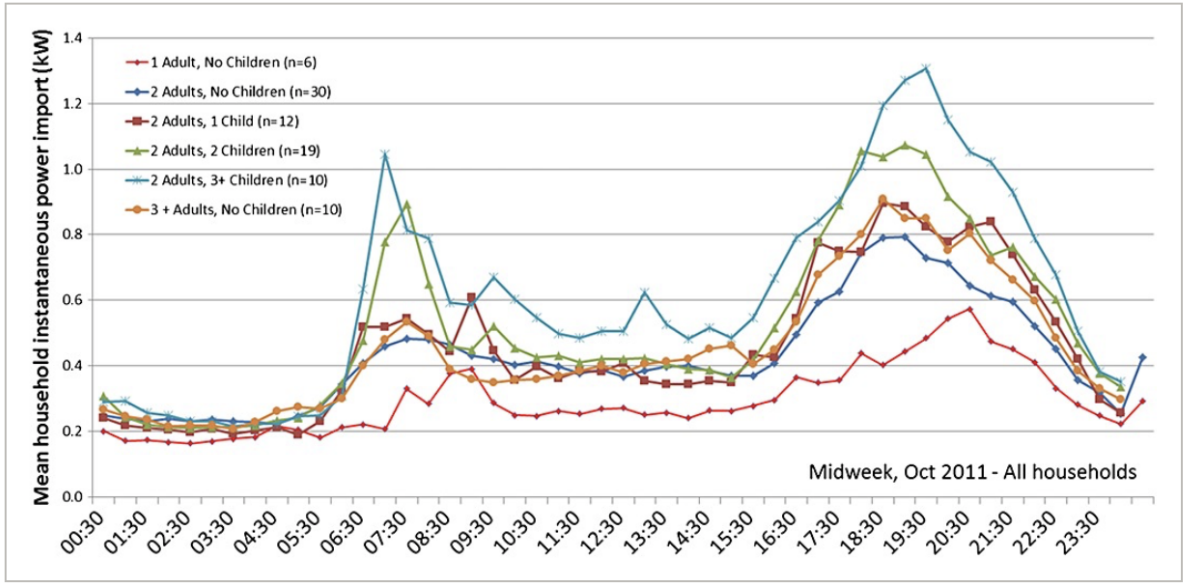
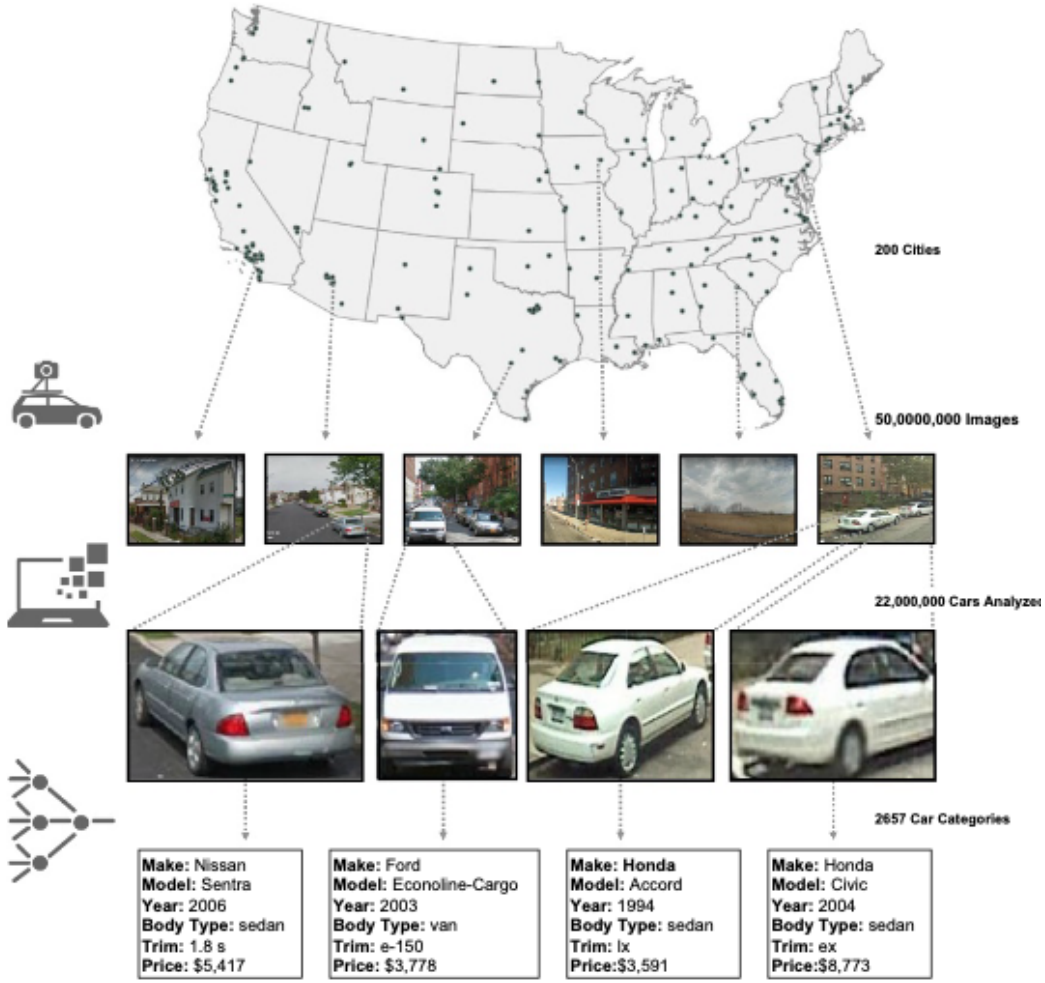


Figure 2 Open in figure viewer | PowerPoint

Household load profiles by household composition.

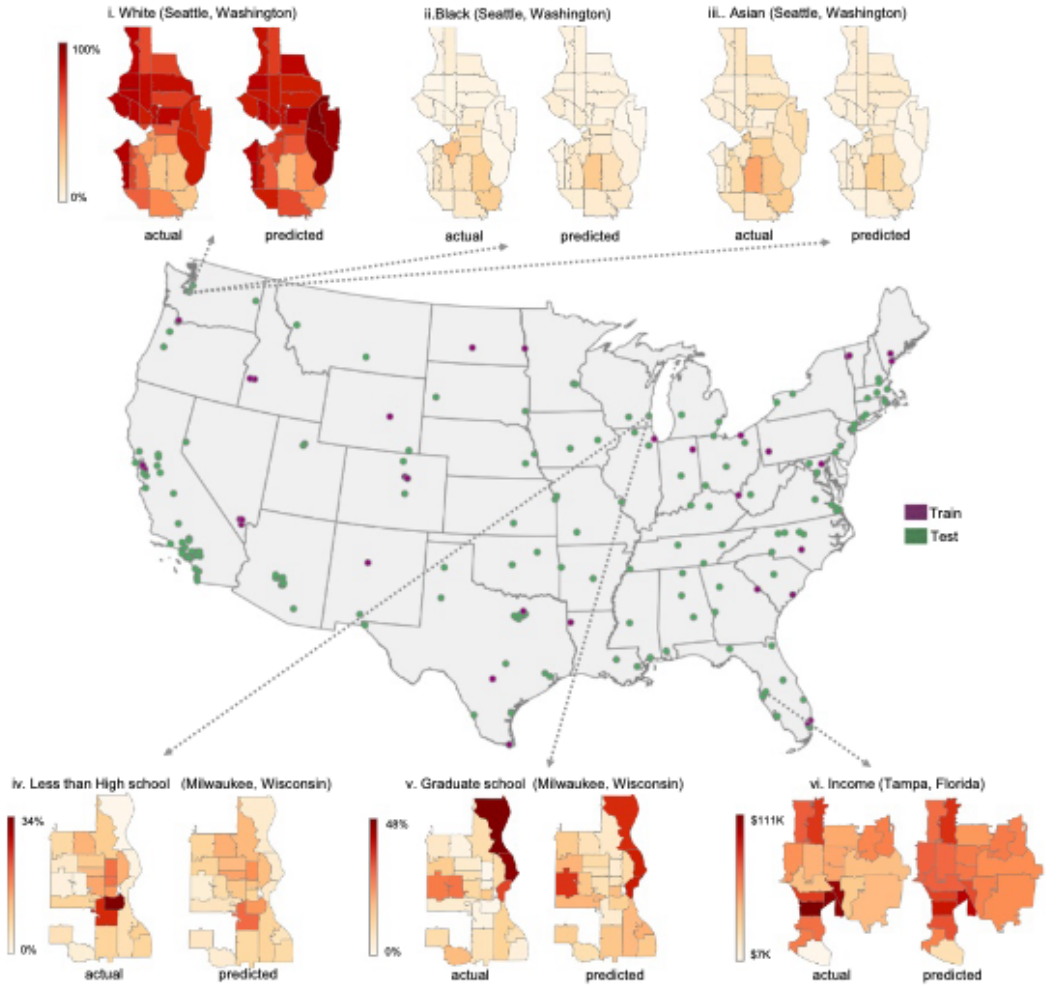
[1] Newing, Andy, Ben Anderson, AbuBakr Bahaj, and Patrick James. "The role of digital trace data in supporting the collection of population statistics–The case for smart metered electricity consumption data." *Population, Space and Place* 22, no. 8 (2016): 849-863.

Google Street View



[1] Gebru, Timnit, Jonathan Krause, Yilun Wang, Duyun Chen, Jia Deng, Erez Lieberman Aiden, and Li Fei-Fei. "Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States." *Proceedings of the National Academy of Sciences* 114, no. 50 (2017): 13108-13113.

Google Street View



[1] Gebru, Timnit, Jonathan Krause, Yilun Wang, Duyun Chen, Jia Deng, Erez Lieberman Aiden, and Li Fei-Fei. "Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States." *Proceedings of the National Academy of Sciences* 114, no. 50 (2017): 13108-13113.

Research Designs using Digital Trace Data

- Lens for social measurement
  - Nowcasting social phenomena
  - Analyzing the impacts of events and shocks in real-time
  - Measuring tricky, hard-to-capture human behaviours
- Digitalized lives and the implications of digitalization
  - Understanding digitalized lives and social dynamics in digital spaces
  - Examining the implications of digital technologies and platforms for social outcomes

Research Designs using Digital Trace Data

- **Nowcasting** social phenomena
  - Better temporal and/or geographical resolution, especially in contexts with data gaps in traditional sources.

Example (1): Predicting Migration using Facebook

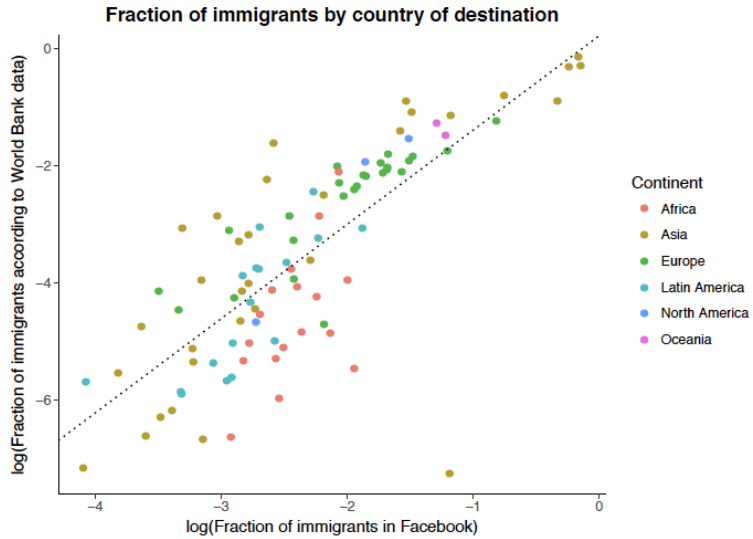
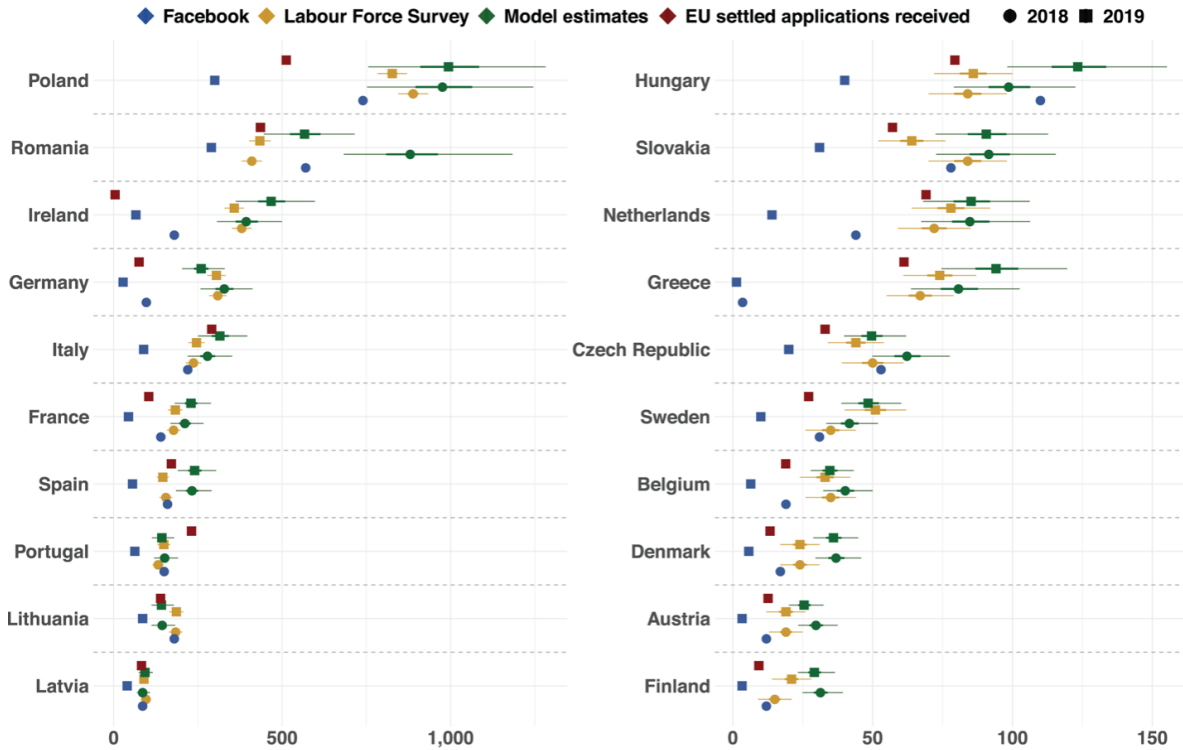


Figure 2: Relationship between stocks of migrants from the Facebook data set, for countries with at least one million Facebook users as of 2016, and the respective estimates from the World Bank (2015). The data points indicate the fraction of immigrants in the population, on a log scale, by country of destination, color-coded by continent. The dashed line is the OLS regression line through the data.

[1] Zagheni, Emilio, Ingmar Weber, and Krishna Gummadi. "Leveraging Facebook's Advertising Platform to Monitor Stocks of Migrants." *Population and Development Review* 43.4 (2017): 721-734.

Example (1): Predicting Migration using Facebook



[1] Rampazzo, Francesco, Jakub Bijak, Agnese Vitali, Ingmar Weber, and Emilio Zagheni. "A framework for estimating migrant stocks using digital traces and survey data: an application in the United Kingdom." *Demography* 58, no. 6 (2021): 2193-2218.

Example (2): Gender Data Gaps

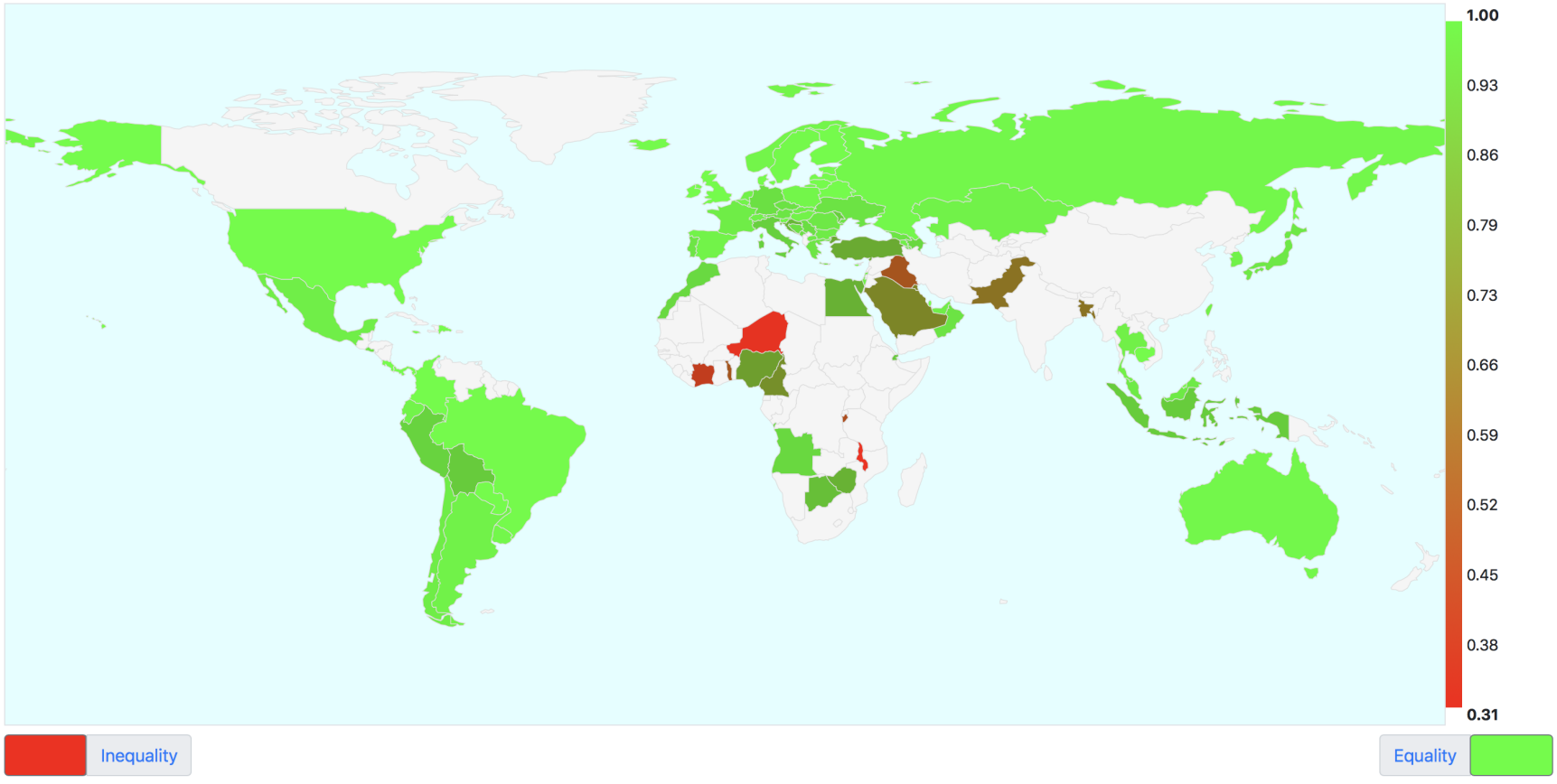


Figure: Gender gaps in internet use computed using data from International Telecommunications Union (ITU) available at `www.digitalgendergaps.org`

Example (2): Predicting Internet Gender Gaps

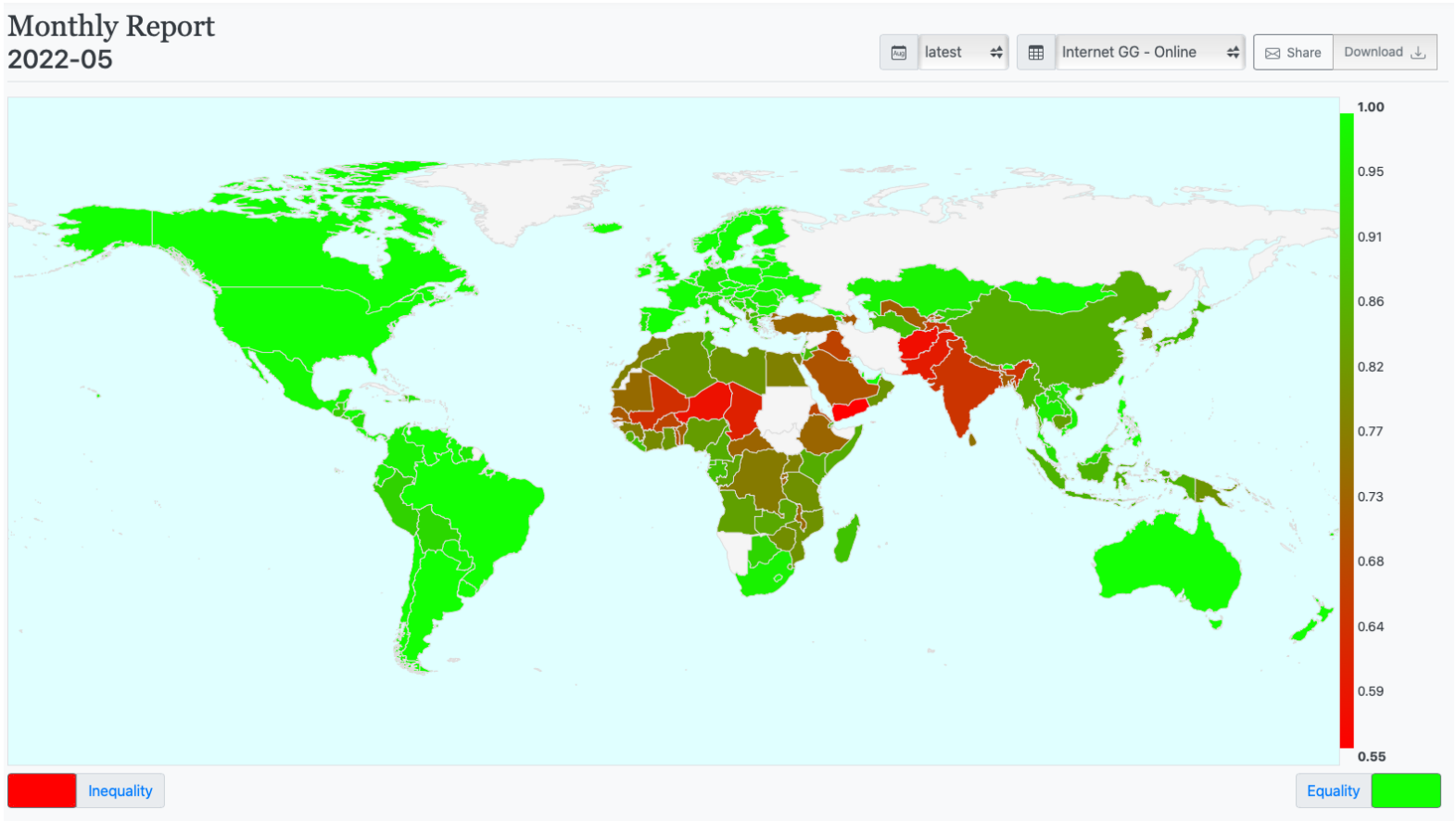


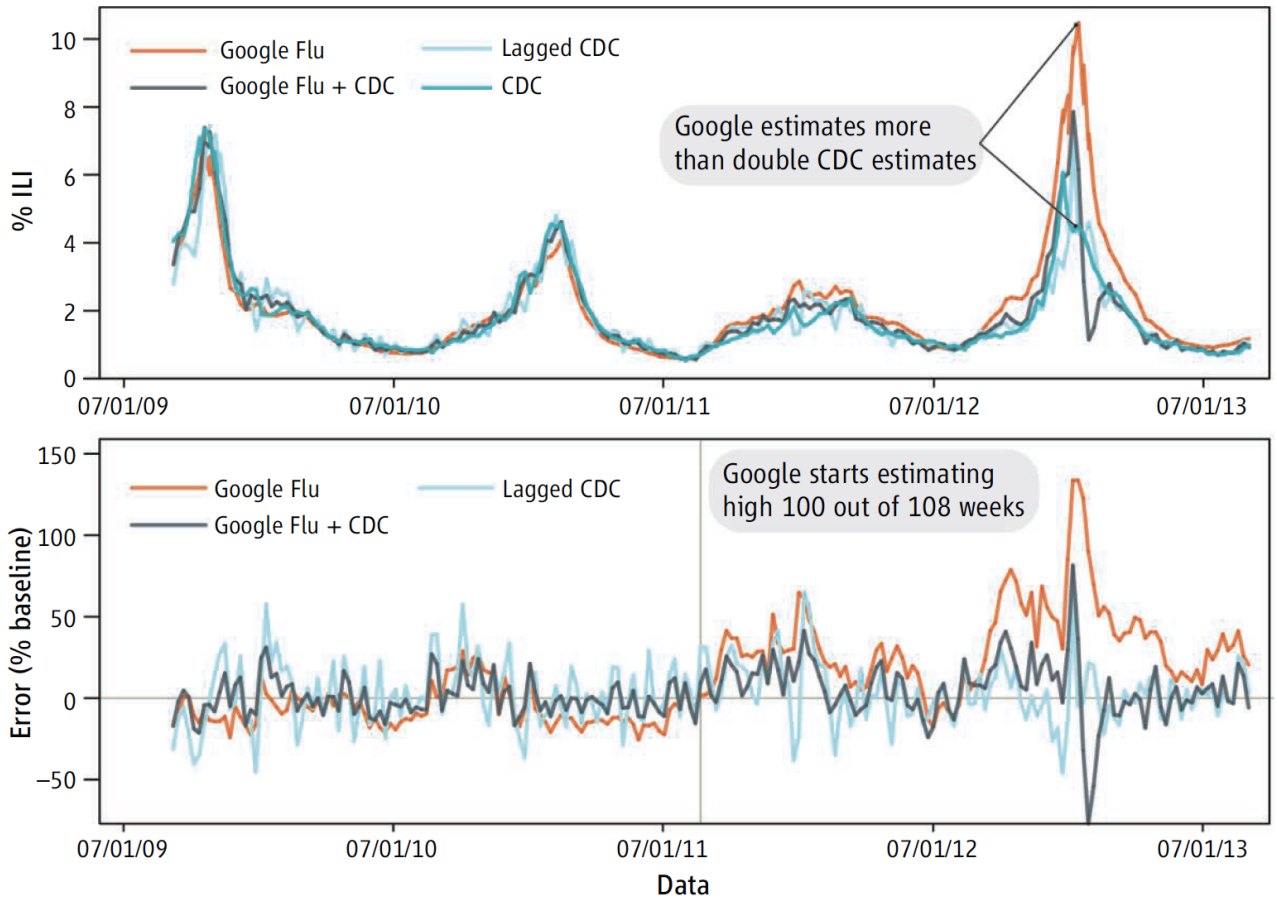
Figure: Gender gaps in internet use computed using data from Facebook (online model) available at `www.digitalgendergaps.org`

[1] Fatehkia, Kashyap, Weber. 2018. "Using Facebook ad data to track the global digital gender gap" *World Development*

Research Designs using Digital Trace Data

- **Nowcasting** social phenomena
  - Digital trace data are **predictors** within a model trained to predict 'ground truth' measure of interest
  - Think carefully about how to operationalize measures from the digital trace source.
    - Algorithmic bias
    - Population drift

Cautionary Tale: Google Flu



[1] Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani. "The parable of Google Flu: traps in big data analysis." *Science* 343, no. 6176 (2014): 1203-1205.

Research Designs using Digital Trace Data

- **Nowcasting** social phenomena
  - Construct generated using digital trace source is a **predictor** within a model trained to predict 'ground truth'
  - Think carefully about how to operationalize measures from the digital trace source.
    - Algorithmic bias
    - Drift: population drift, usage drift, system drift
  - Digital trace often better as complements: combine across multiple sources of data, and consider using multiple platforms.

Research Designs using Digital Trace Data

- Lens for social measurement
  - Nowcasting social phenomena
  - Analyzing the impacts of events and shocks in real-time
  - Measuring tricky, hard-to-capture human behaviours
- Digitalized lives and the implications of digitalization
  - Understanding digitalized lives and social dynamics in digital spaces
  - Examining the implications of digital technologies and platforms for social outcomes

Example: Law and Public Sentiment

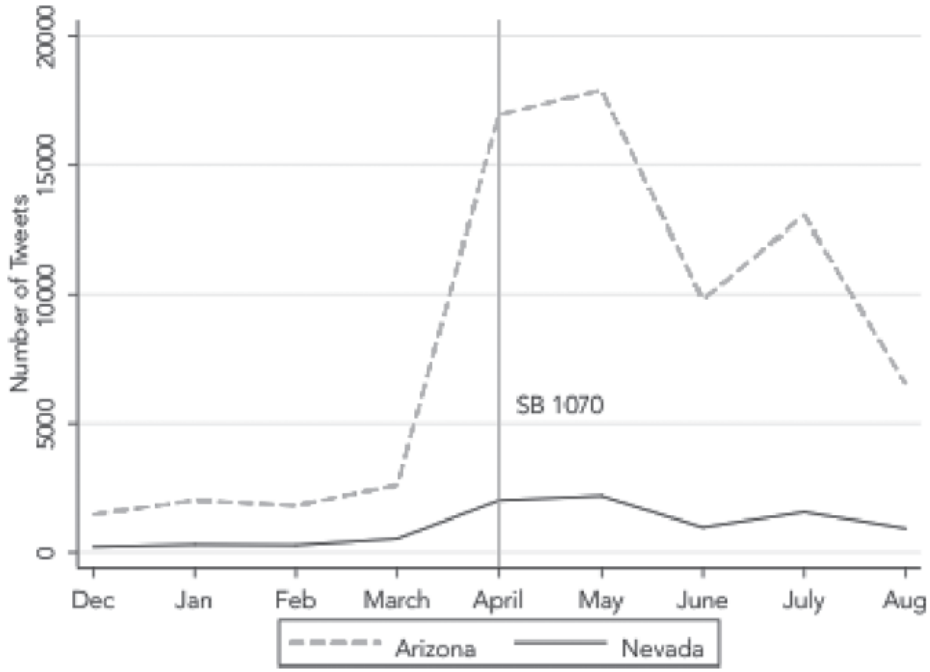


FIG. 2.—Number of Twitter messages related to immigrants per month in Arizona and Nevada (December 2010–August 2011). The vertical line on April 2010 indicates when the Arizona governor approved SB 1070.

[1] Flores, René D. "Do anti-immigrant laws shape public sentiment? A study of Arizona's SB 1070 using Twitter data." *American Journal of Sociology* 123, no. 2 (2017): 333-384.

Research Designs using Digital Trace Data

- Analyzing the impacts of events
- What is the appropriate counterfactual trend?

Research Designs using Digital Trace Data

- Lens for social measurement
  - Nowcasting social phenomena
  - Analyzing the impacts of events and shocks in real-time
  - Measuring tricky, hard-to-capture human behaviours
- Digitalized lives and the implications of digitalization
  - Understanding digitalized lives and social dynamics in digital spaces
  - Examining the implications of digital technologies and platforms for social outcomes

Example: Online Dating

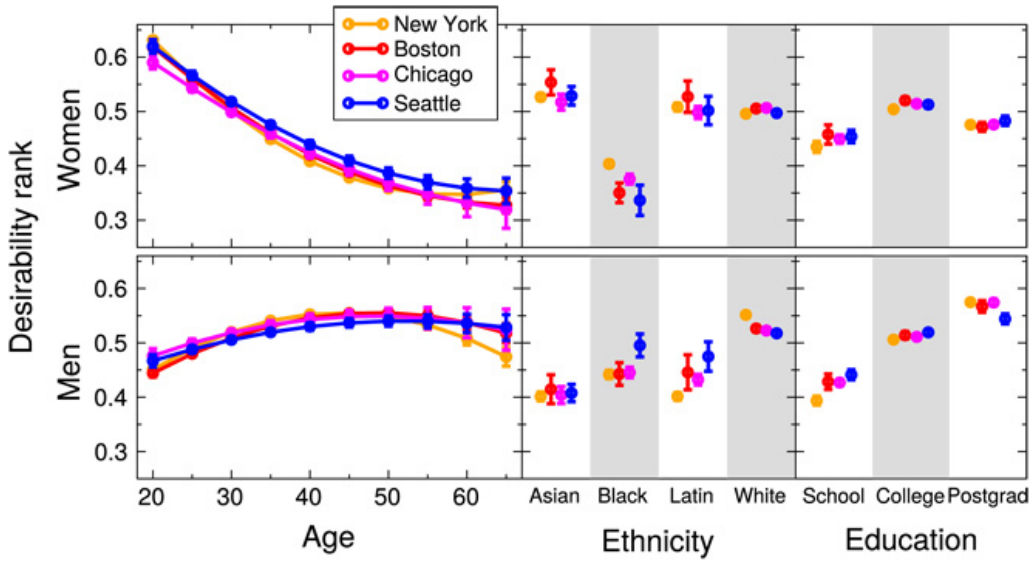


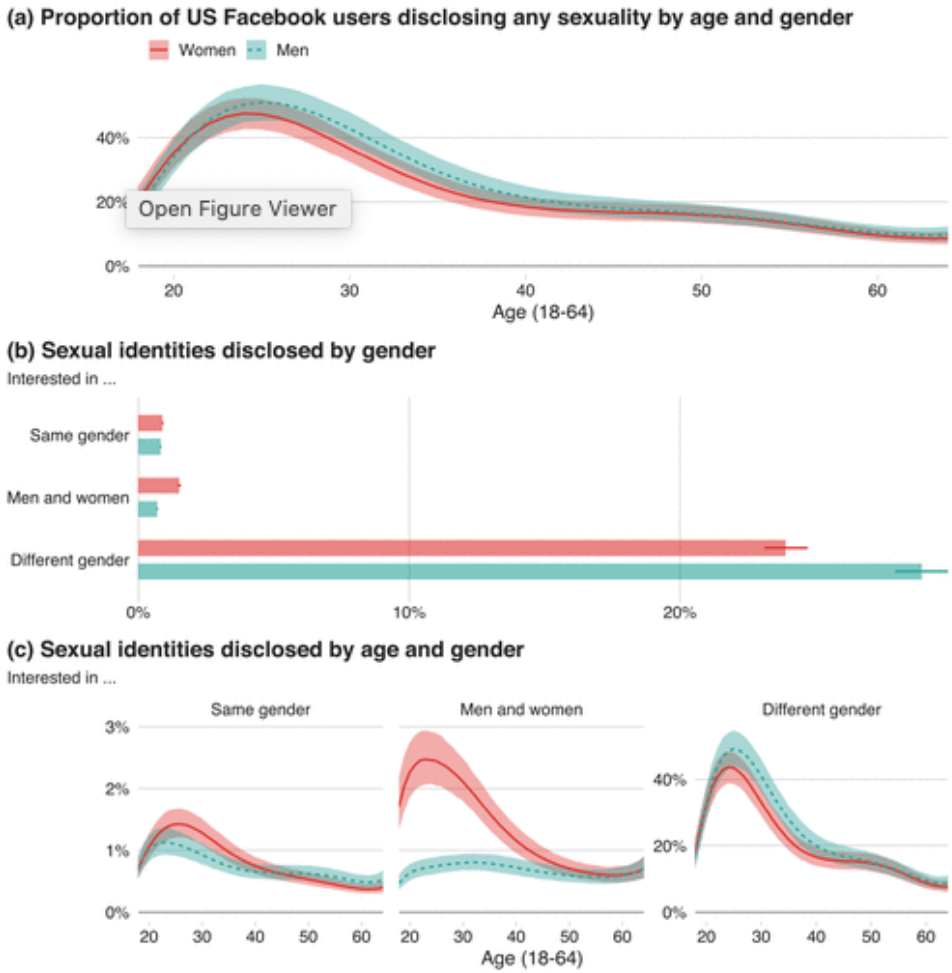
Fig. 2 Desirability, quantified using the measures defined here, as a function of demographic variables of the user population.

[1] Bruch, Elizabeth E., and M. E. J. Newman. "Aspirational pursuit of mates in online dating markets." *Science Advances* 4, no. 8 (2018): eaap9815.

Research Designs using Digital Trace Data

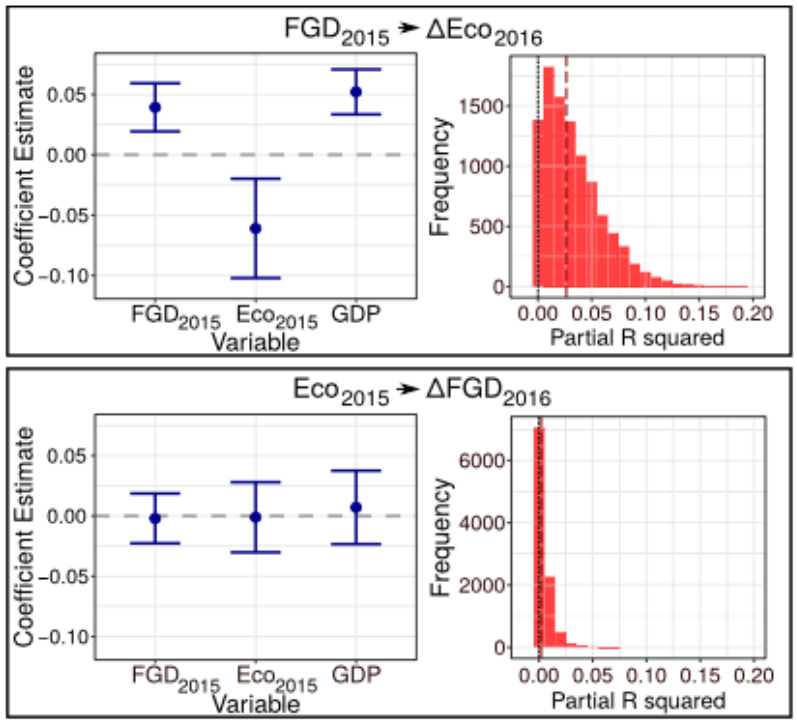
- Lens for social measurement
  - Nowcasting social phenomena
  - Analyzing the impacts of events and shocks in real-time
  - Measuring tricky, hard-to-capture human behaviours
- Digitalized lives and the implications of digitalization
  - Understanding digitalized lives and social dynamics in digital spaces
  - Examining the implications of digital technologies and platforms for social outcomes

Example: Sexuality Disclosure on Social Media



[1] Gilroy, Connor, and Ridhi Kashyap. "Digital traces of sexualities: understanding the salience of sexual identity through disclosure on social media." *Socius* 7 (2021): 23780231211029499.

Example: Social Media Gender Gaps



[1] Garcia, David, Yonas Mitike Kassa, Angel Cuevas, Manuel Cebrian, Esteban Moro, Iyad Rahwan, and Ruben Cuevas. "Analyzing gender inequality through large-scale Facebook advertising data." Proceedings of the National Academy of Sciences 115, no. 27 (2018): 6958-6963.

Considerations when Using Digital Trace Data

- What is your research question? How can digital trace data help you address it?
- Is it possible to use the data in an ethical manner?
- Are the data accessible?
- What is your theoretical or measurement framework?
- Who do your data include and who do they exclude?
- Can you compare or validate your measures against other sources of data?