# Digital Trace Data

## Part 1

Ridhi Kashyap

University of Oxford

SICSS-Oxford
June 18, 2019

# The Digital Age

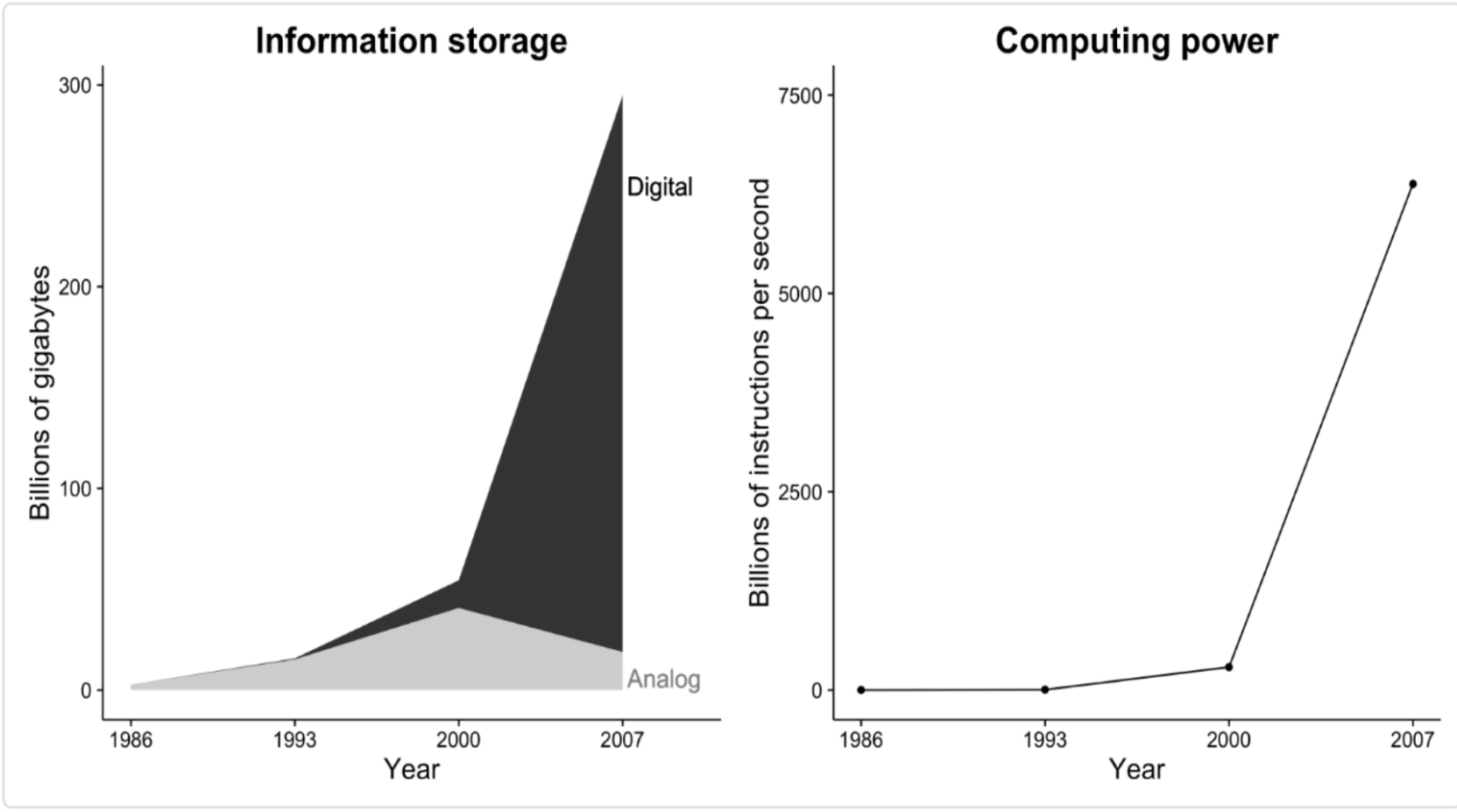



Figure 1.1: Information storage capacity and computing power are increasing dramatically. Further, information storage is now almost exclusively digital (Hilbert and López 2011). These changes create incredible opportunities for social researchers.

# The Digital Age

## The "Big Data" Era: Features

- Explosion in **volume** of data
- **Velocity** with which data are produced
- **Producers of data**: admin, private corporations, individuals. Centralised v. decentralised.
- **Variety** of data, including from technologies such as mobile phones and the "internet of things" (embedded processors in everyday use devices like cars, household devices or thermostats)
  - Not quite **custom-made** for research but **ready-made** (Salganik 2018)

# The Digital Age


# Digital Trace Data: Definition

- The data by-products of the digitalisation of our lives and the adoption of digital technologies and platforms (e.g. social media).
  - Expression of digital interactions and/or phenomena (e.g. tweeting)
  - Capture of "offline life" in digital forms (e.g. accelerometers)
  - Measures of behaviour or activity (contrast with self-reported)

# Digital Trace Data: Examples

- Social media sites

# Twitter

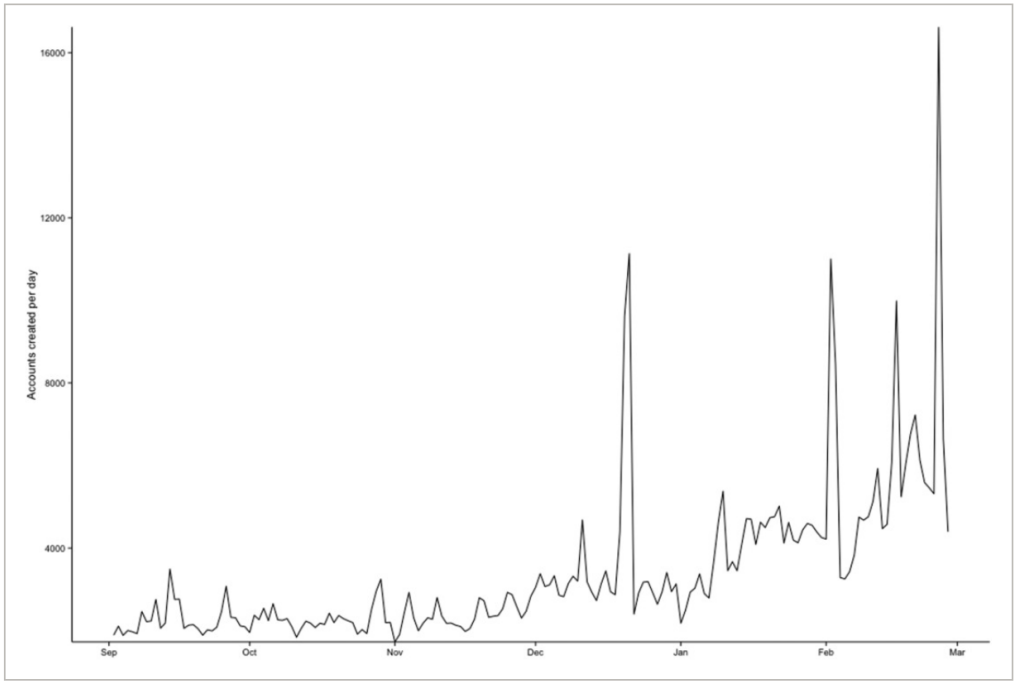


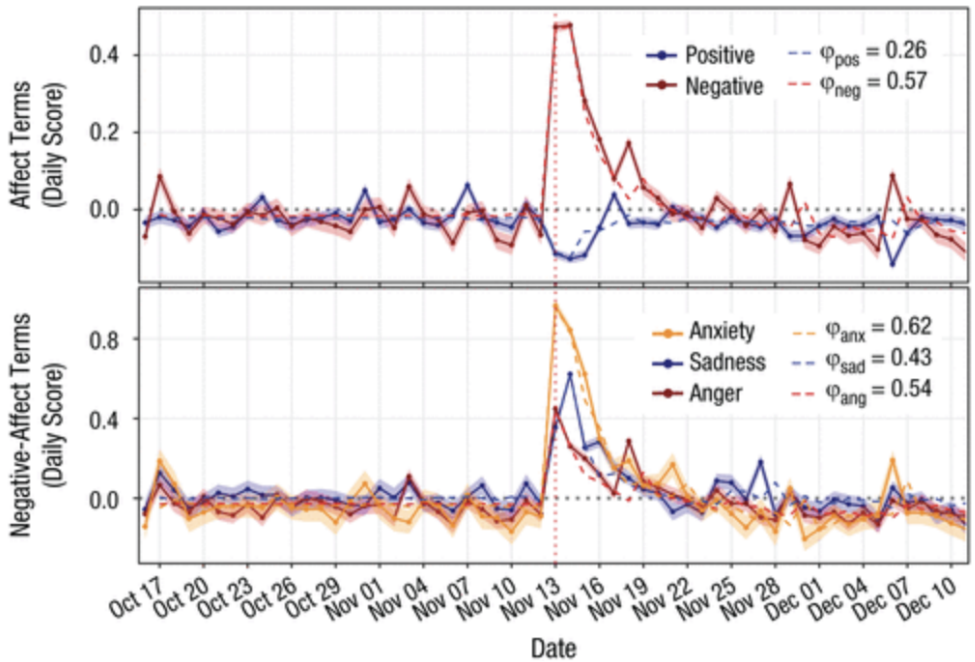

**Figure 3** Open in figure viewer | PowerPoint

Number of new Twitter accounts created daily in Ukraine before and after the protests (2014–15). This figure shows the distribution of account creation dates for users in our dataset (i.e., for those who tweeted at least once about the Euromaidan protests using the keywords and hashtags we designated for collection). There was a relatively stable number of new accounts created prior to the protests, but—beginning in late November—there was a dramatic increase in new account creation, suggesting that citizens joined Twitter in part to follow the protests. This figure is adapted from Metzger and Tucker (2017).

---

[1] Jost, John T., Pablo Barbera, Richard Bonneau, Melanie Langer, Megan Metzger, Jonathan Nagler, Joanna Sterling, and Joshua A. Tucker. "How social media facilitates political protest: Information, motivation, and social networks." *Political psychology* 39 (2018): 85-118.

# 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

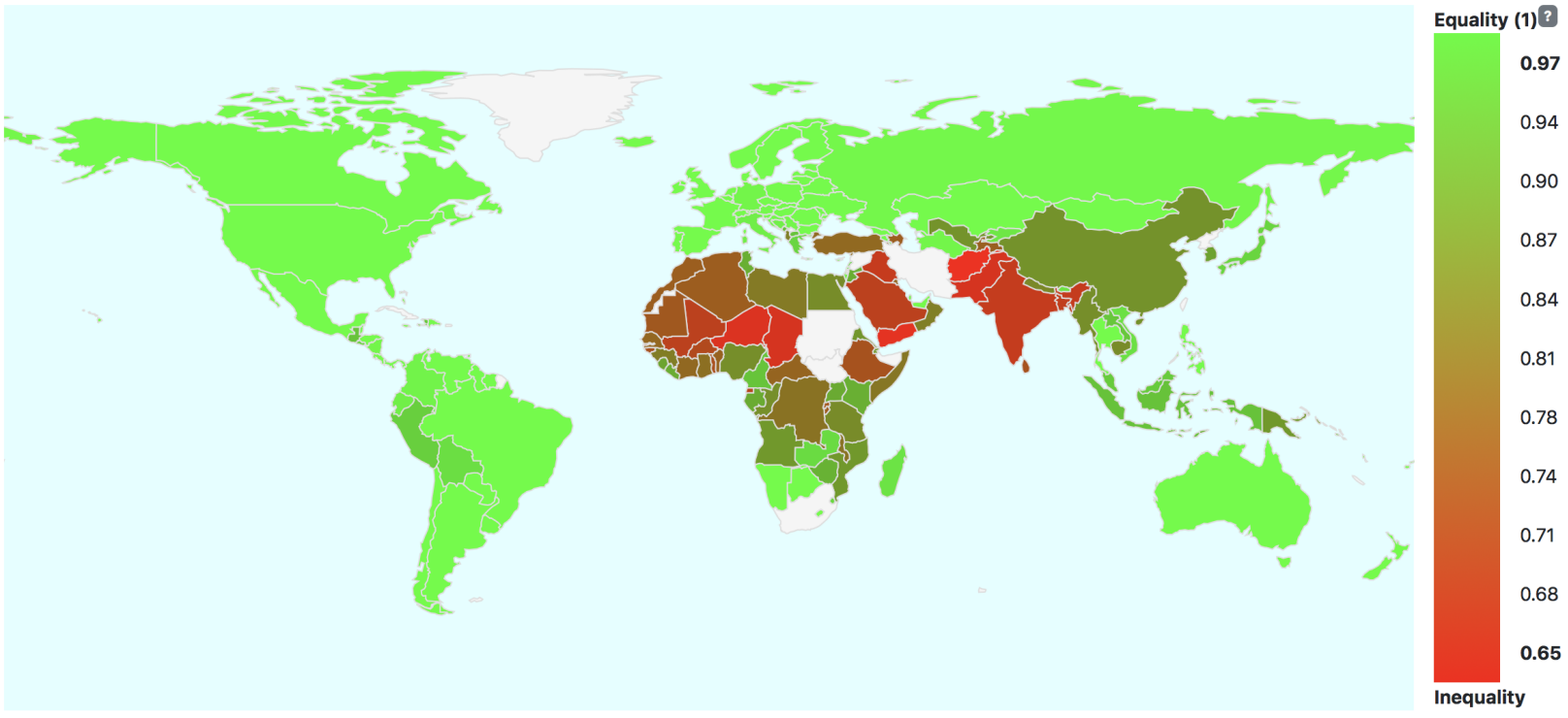



Figure: Gender gaps in internet use computed using data from Facebook (online model) available at `www.digitalgendergaps.org`

[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.

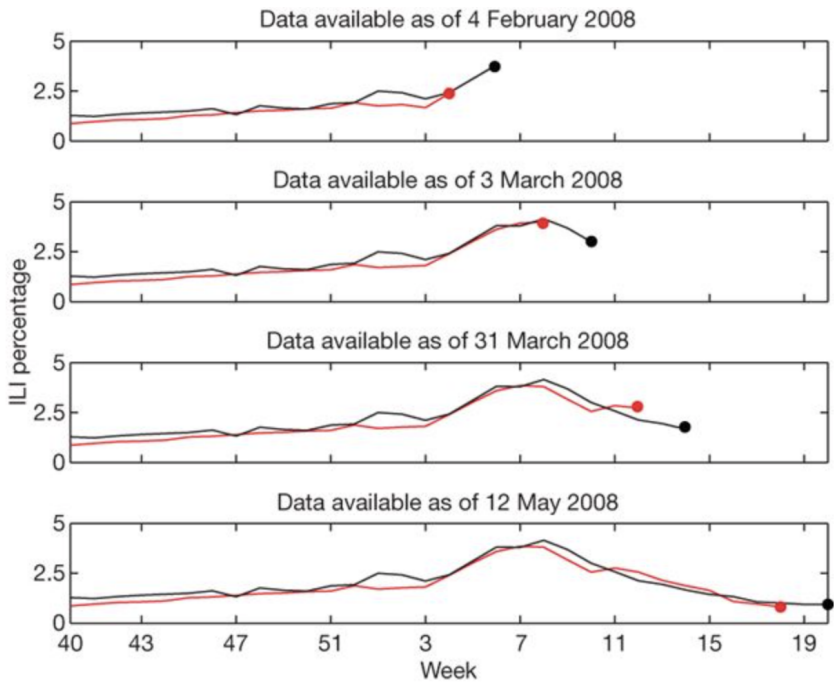
# Digital Trace Data: Examples

- 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 Abortion

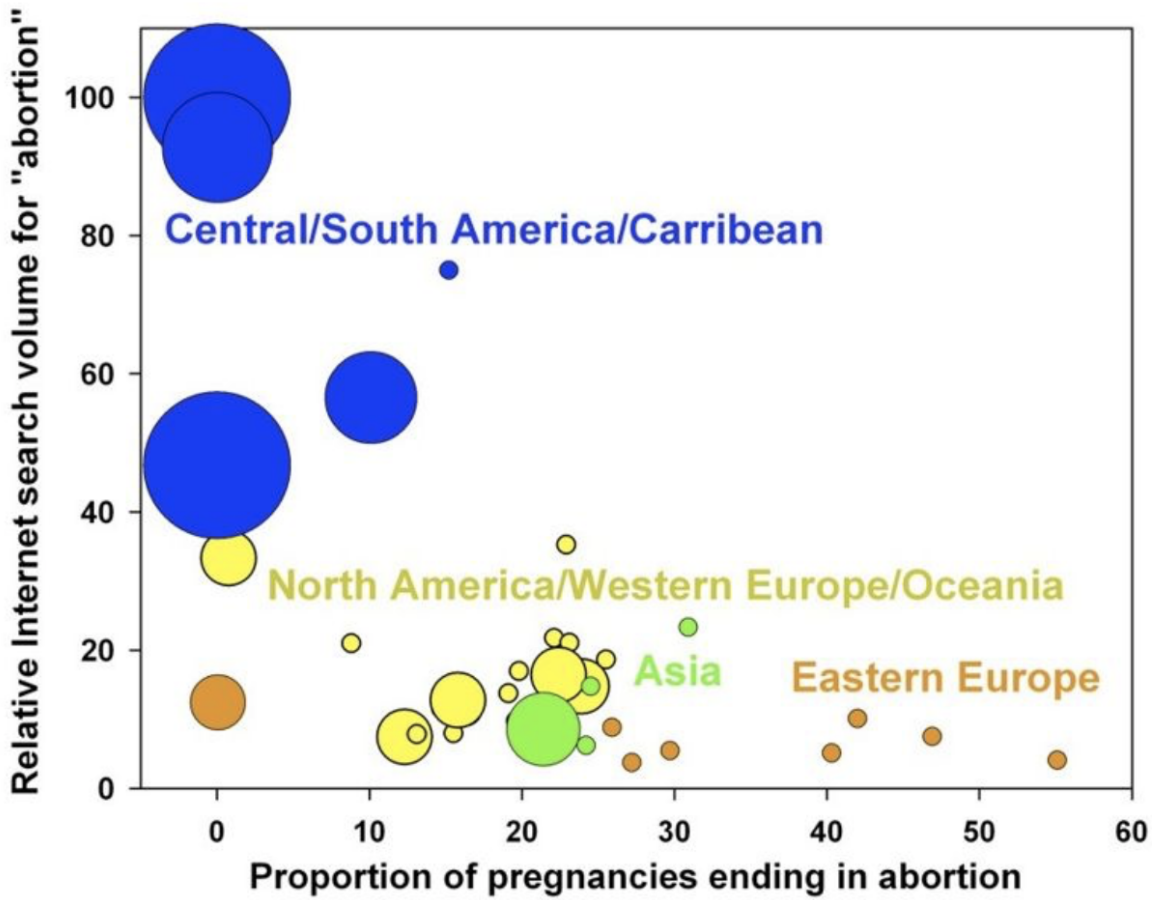



Figure: Relationship between abortion search volume and abortion rates across 37 countries. Marker size indicates the number of restrictions (from 0 to 7) that exist on abortion in each country.

---

[1] Reis and Brownstein (2010), Measuring the impact of health policies using Internet search patterns: the case of abortion, *BMC Public Health*

# Digital Trace Data: Examples

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

# Reddit

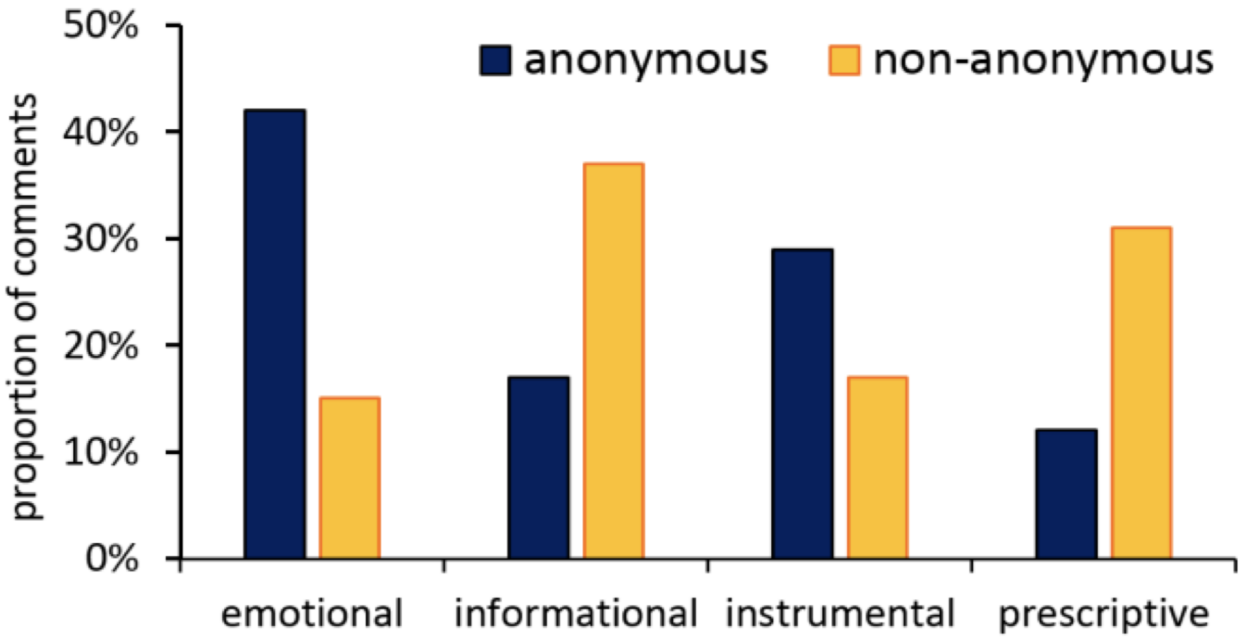



**Figure 4. Posts from anonymous and non-anonymous accounts in the light of social support types.**

[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

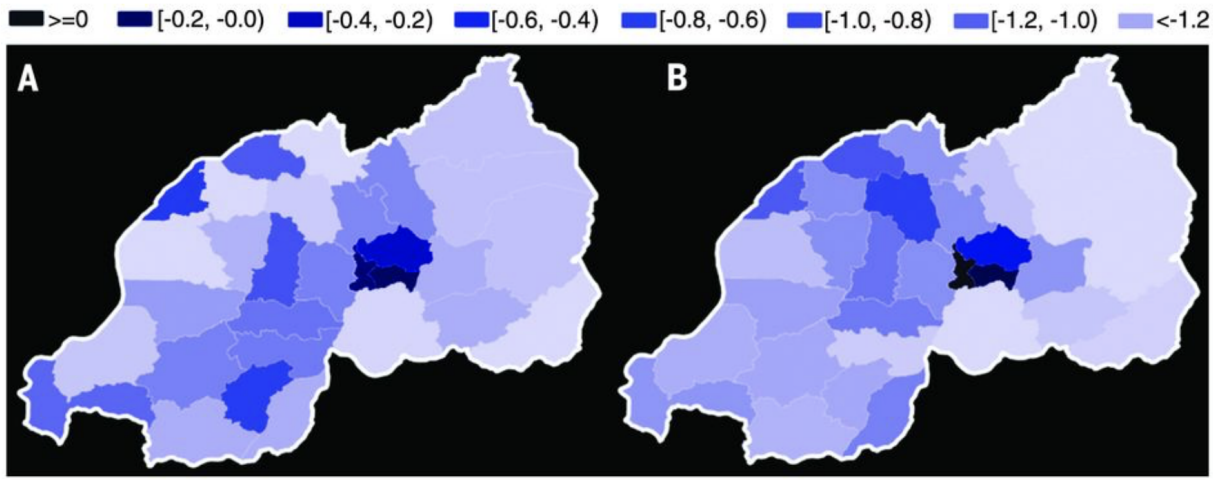



Figure: (A) Predicted composite wealth index (district average), computed from 2009 call data and aggregated by administrative district. (B) Actual composite wealth index (district average), as computed from a 2010 government DHS of 12,792 households.

[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

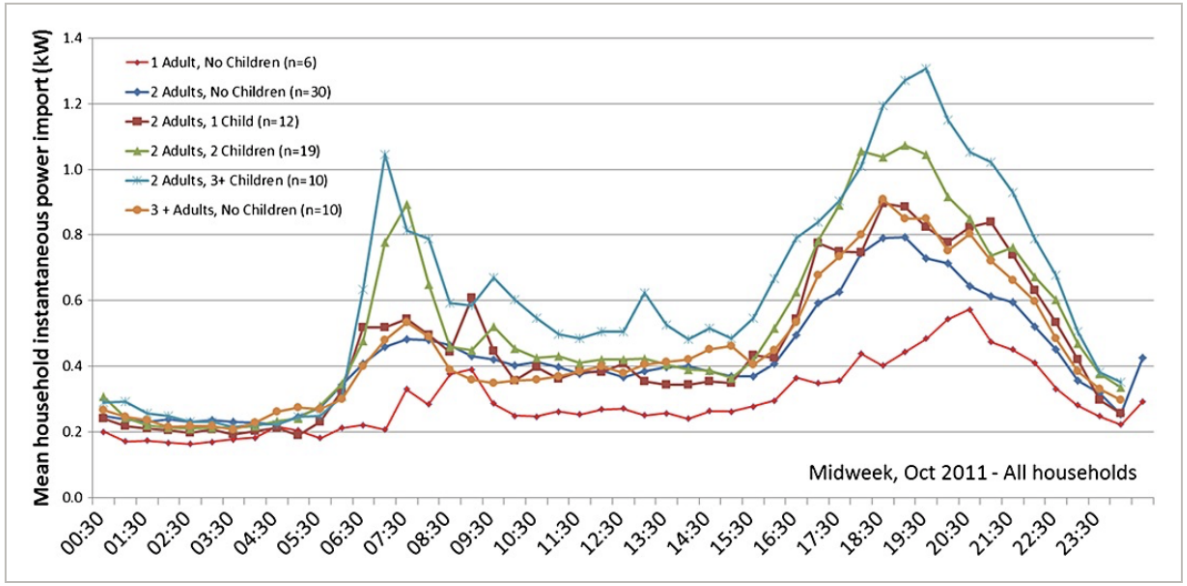



**Figure 2** Open in figure viewer | PowerPoint

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.

# Taxi Meters in NYC

## Why you Can't Find a Taxi in the Rain and Other Labor Supply Lessons from Cab Drivers*

Henry S. Farber

*The Quarterly Journal of Economics*, Volume 130, Issue 4, November 2015, Pages 1975–2026, https://doi.org/10.1093/qje/qjv026

**Published:** 13 July 2015

[1]Farber, Henry S. "Why you can't find a taxi in the rain and other labor supply lessons from cab drivers." *The Quarterly Journal of Economics* 130, no. 4 (2015): 1975-2026.

# Taxi Meters in NYC

"The New York City Taxi and Limousine Commission (TLC), the agency charged with regulating the industry, now requires all taxis to be equipped with electronic devices that record all trip information, including fares, times, and locations. The (currently two) companies that supply these devices report all this information to the TLC on a regular basis, and I have obtained full information for all trips taken in NYC taxi cabs for the five years from 2009 to 2013."

[1] Farber, Henry S. "Why you can't find a taxi in the rain and other labor supply lessons from cab drivers." *The Quarterly Journal of Economics* 130, no. 4 (2015): 1975-2026.

# Promises of Digital Trace Data

- Volume (Big)
  - Useful for heterogeneity, disaggregation, and rare occurrences.
- Velocity: higher frequency or real-time measurement, “always on”

# Twitter

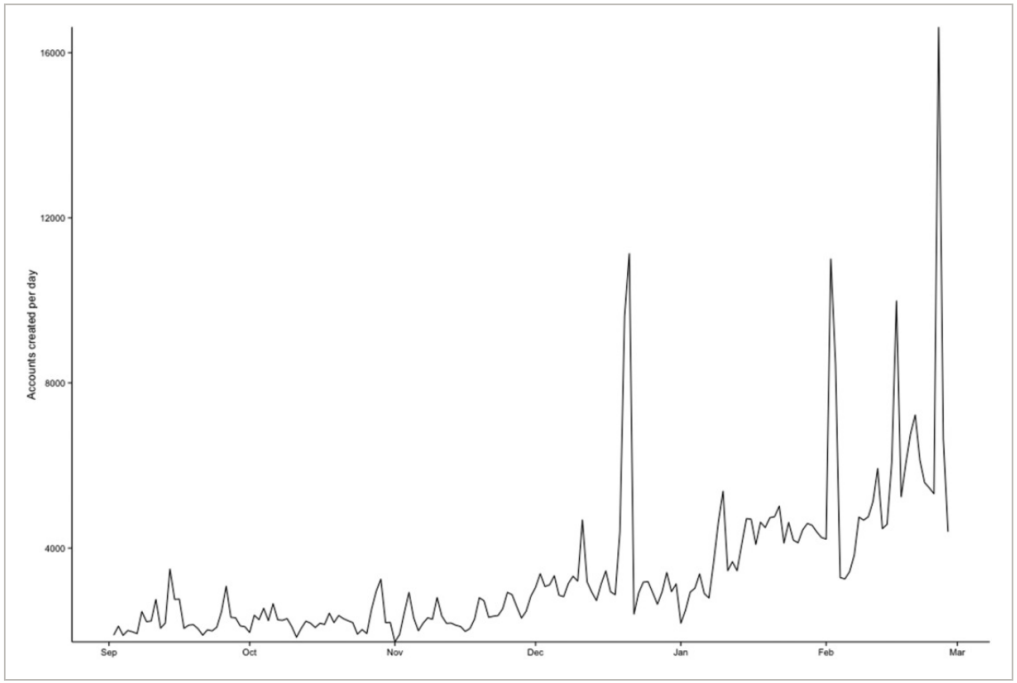



**Figure 3** Open in figure viewer | PowerPoint

Number of new Twitter accounts created daily in Ukraine before and after the protests (2014–15). This figure shows the distribution of account creation dates for users in our dataset (i.e., for those who tweeted at least once about the Euromaidan protests using the keywords and hashtags we designated for collection). There was a relatively stable number of new accounts created prior to the protests, but—beginning in late November—there was a dramatic increase in new account creation, suggesting that citizens joined Twitter in part to follow the protests. This figure is adapted from Metzger and Tucker (2017).

---

[1] Jost, John T., Pablo Barbera, Richard Bonneau, Melanie Langer, Megan Metzger, Jonathan Nagler, Joanna Sterling, and Joshua A. Tucker. "How social media facilitates political protest: Information, motivation, and social networks." *Political psychology* 39 (2018): 85-118.

# Promises of Digital Trace Data

- Volume (Big)
  - Useful for heterogeneity, disaggregation, and rare occurrences.
- Velocity: higher frequency or real-time measurement, "always on"
- Coverage: topics or geographies that might not be covered in other data sources.
- Non-reactive: potential to capture behaviour that might be difficult to measure.

# Facebook and Gender Gaps

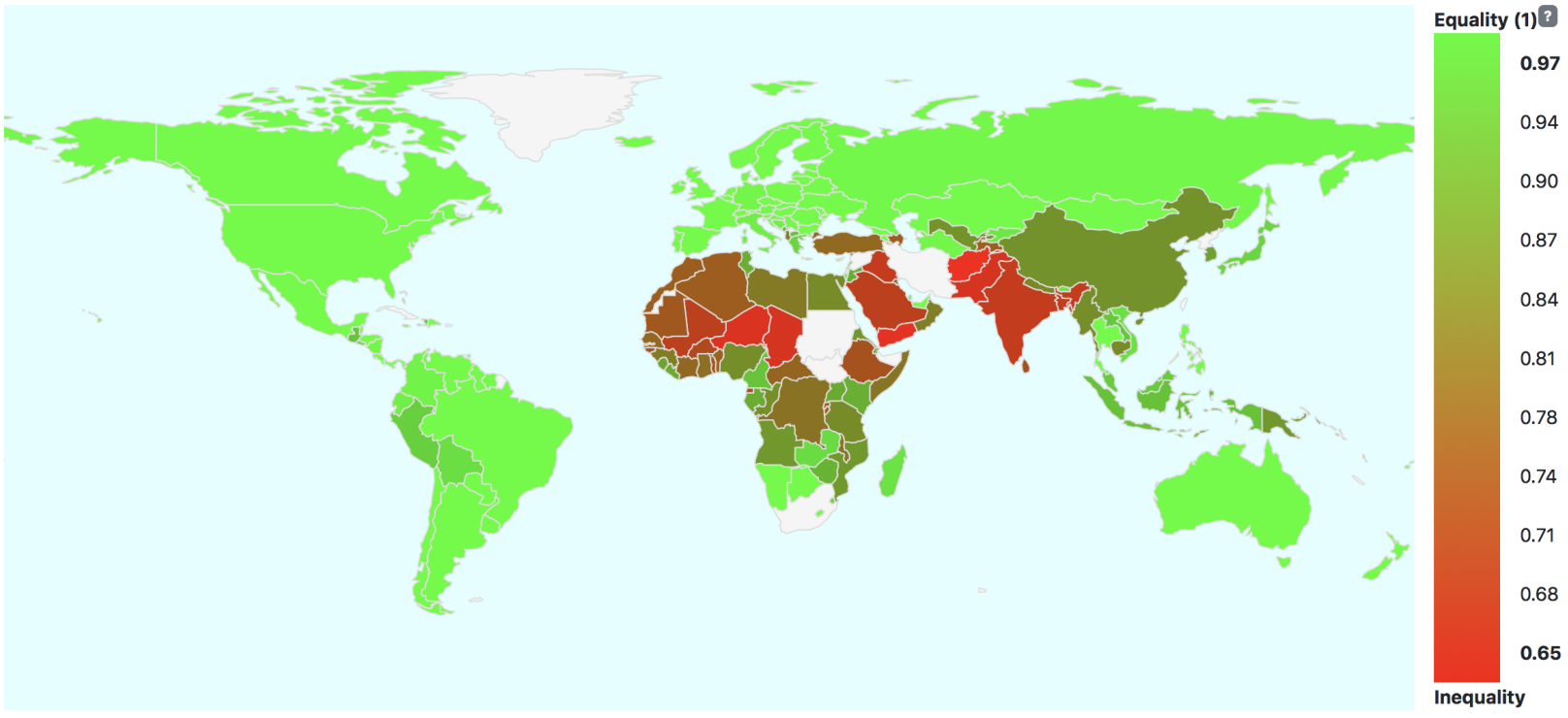



Figure: Gender gaps in internet use computed using data from Facebook (online model) available at `www.digitalgendergaps.org`

[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.

# Measuring Migration using Facebook

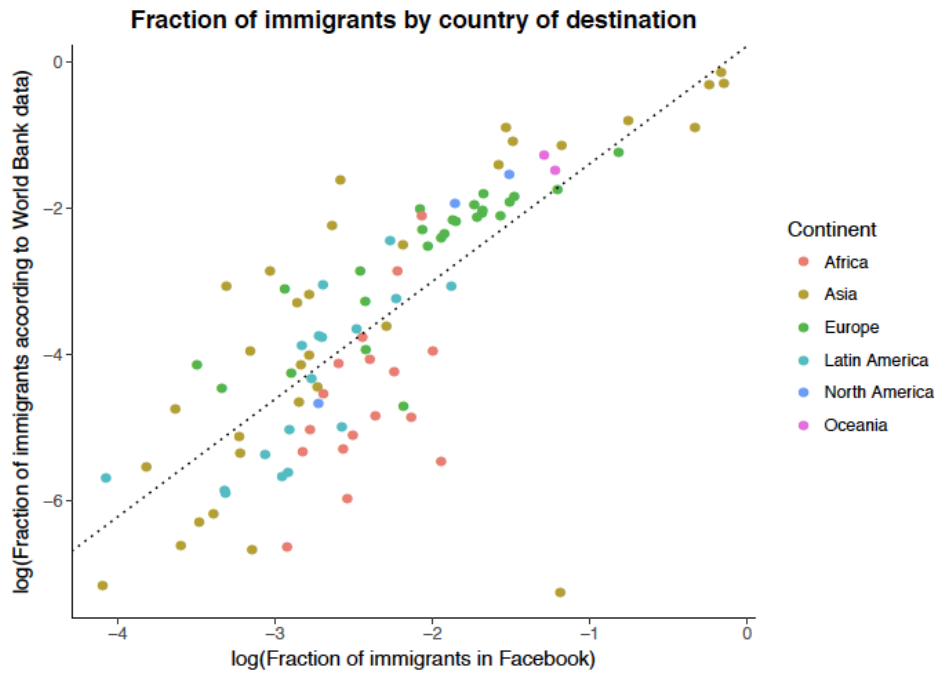


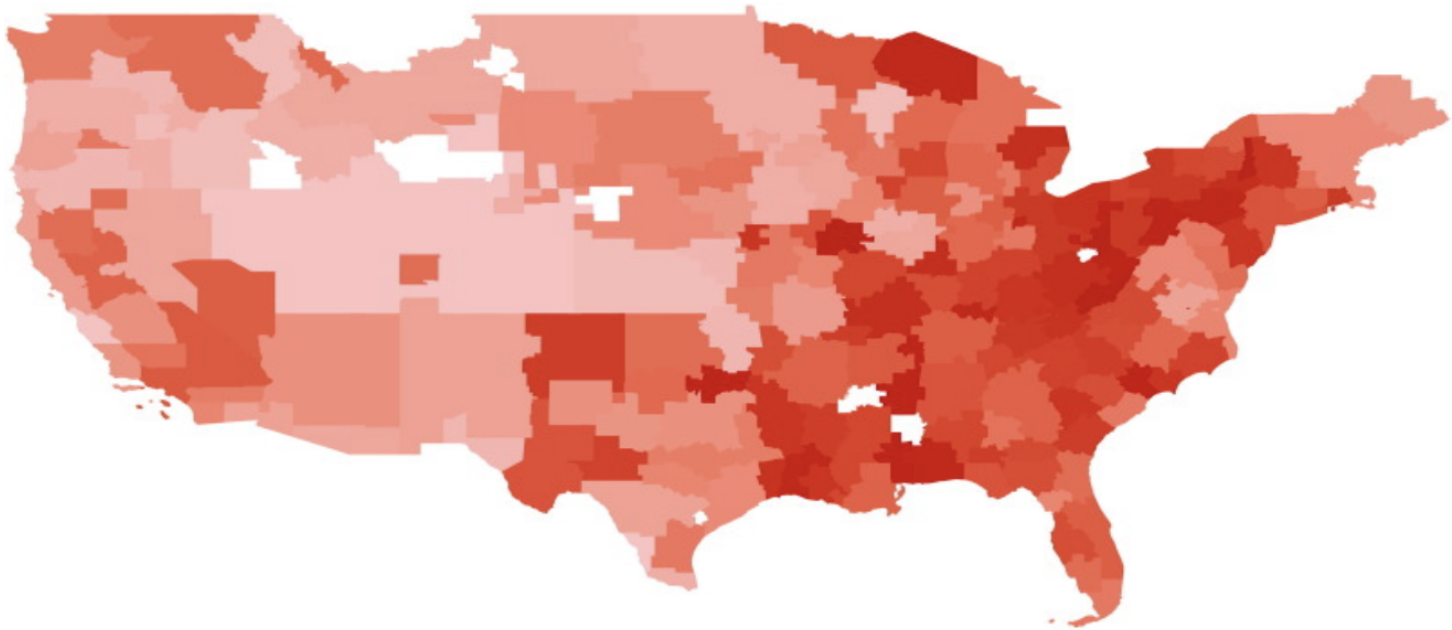

Figure 2: Relationship between stocks of migrants from the Facebook data set, for countries with at least one million Facebook users as of 2016, and the respective estimates from the World Bank (2015). The data points indicate the fraction of immigrants in the population, on a log scale, by country of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

# Promises of Digital Trace Data

- Volume (Big)
  - Useful for heterogeneity, disaggregation, and rare occurrences.
- Velocity: higher frequency or real-time measurement, "always on"
- Coverage: topics or geographies that might not be covered in other data sources.
- Non-reactive: potential to capture behaviour that might be difficult to measure.

# Racial Bias and Google Search



Download full-size image

Fig. 2. Racially charged search rate, media market.

*Notes*: This maps search volume for "[Word 1](s)," from 2004 to 2007, at the media market level. Darker areas signify higher search volume. White areas signify media markets without data. Alaska and Hawaii, for which data are available, are not shown.

[1] Stephens-Davidowitz, Seth. "The cost of racial animus on a black candidate: Evidence using Google search data." *Journal of Public Economics* 118 (2014): 26-40.

# Pitfalls of Digital Trace Data

- Dirty
- Inaccessible: often data from companies with different incentives and goals
- Sensitive
- Incomplete
- Non-representative
- Algorithmically confounded, black-box algorithms
- Drift: population drift, usage drift, system drift

# Measuring Migration using Facebook

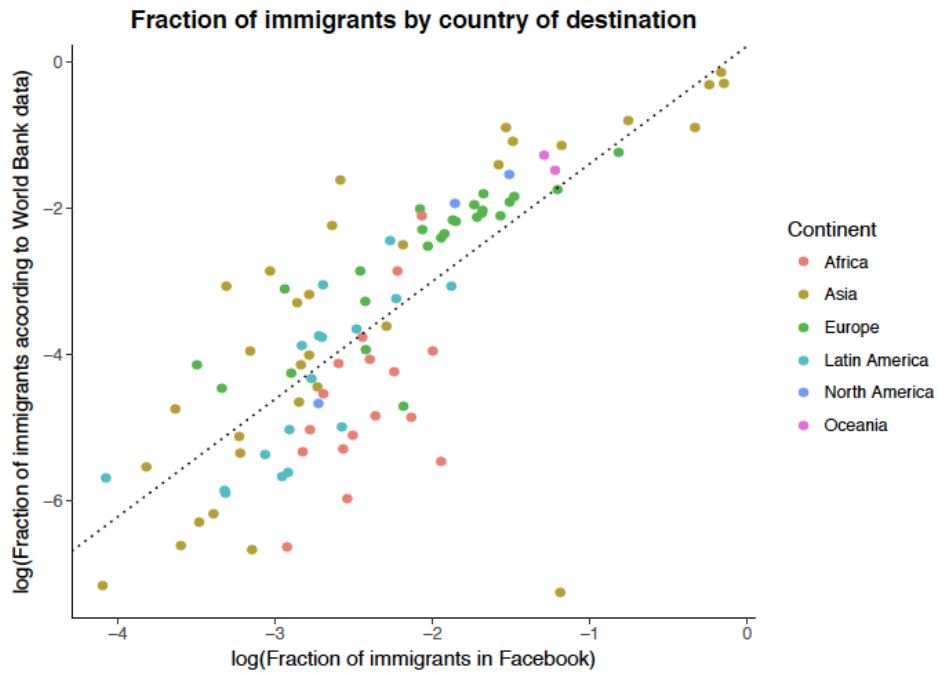
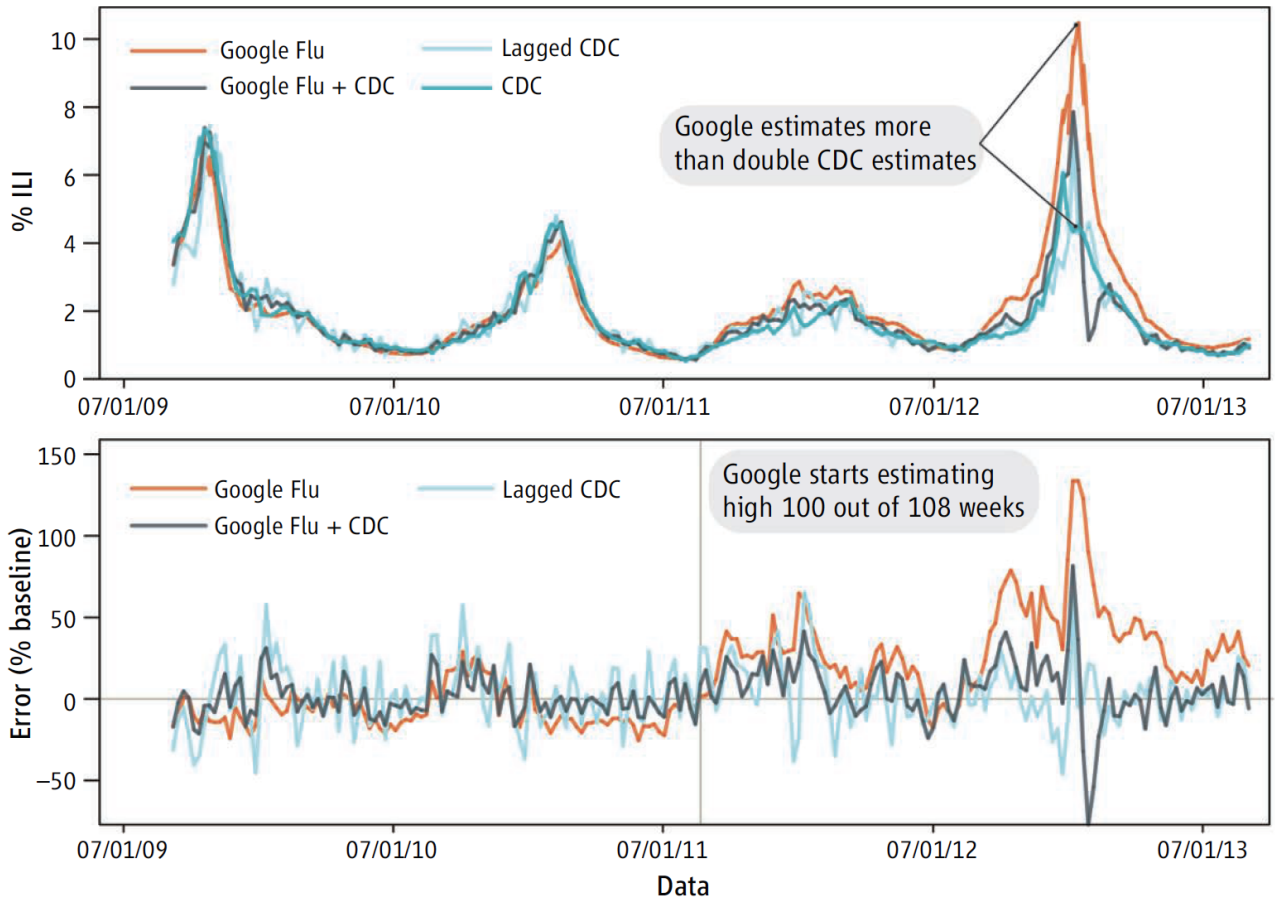



Figure 2: Relationship between stocks of migrants from the Facebook data set, for countries with at least one million Facebook users as of 2016, and the respective estimates from the World Bank (2015). The data points indicate the fraction of immigrants in the population, on a log scale, by country of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

# 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

- Measurement
  1. Operationalising constructs at the macro-level
  2. Nowcasting and filling data gaps. Compare with "offline" benchmarks and models.
- Digital platforms as microcosms of society: testing theories
- Implications of digital technologies for social processes

# Hybrid Research Designs

- Combining digital trace data with conventional data sources like surveys
- Apps for data generation and extraction
- Combining digital traces with experiments (e.g. bots)
- Ethics