Digital trace data

Jan Zilinsky (TUM) Summer Institute in Computational Social Science

Munich (2023)

Digital footprints

One perspective: data that is created as a **by-product** of people's interactions with digital systems

A misleading definition? When you post a comment on social media, it's an expression of deliberate action. The post is created intentionally and surely it counts as digital data.

And *other metadata* is generated in connection with your post - users will differ in the extent of their awareness.

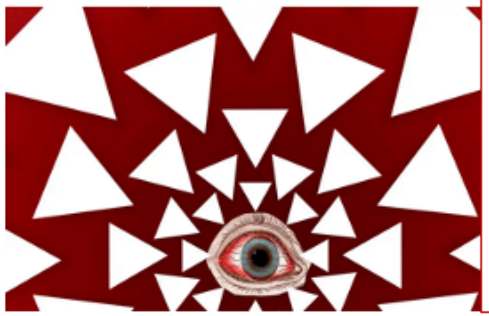
The New York Times

Opinion

YouTube, the Great Radicalizer

By Zeynep Tufekci

March 10, 2018



The New York Times

OPINION

GUEST ESSAY

Does Instagram Harm Girls? No One Actually Knows.

Oct. 10, 2021

UK Parliament

Business | MPs, Lords & offices | About | Get inv

UK Parliament

> Democracy under threat from 'pandemic of misinformation' online – Lords Democracy and Digital Technologies Co

Democracy under threat from 'pandemic of misinformation' online – Lords Democracy and Digital Technologies Committee

The UK Government should act immediately to deal with a 'pandemic of misinformation' that poses an existential threat to our democracy and way of life. The stark warning comes in a report published today by the Lords Committee on Democracy and Digital Technologies.

The report says the Government must take action 'without delay' to ensure tech giants are held responsible for the harm done to individuals, wider society and our democratic processes through misinformation widely spread on their platforms.

The Committee says online platforms are not 'inherently ungovernable' but power has been ceded to a "few unelected and unaccountable digital corporations" including Facebook and Google, and politicians must act now to hold those corporations to account when they are shown to negatively influence public debate and undermine democracy.

The Committee sets out a package of reforms which, if implemented, could help restore public trust and ensure democracy does not 'decline into irrelevance'.

Digital data

- Social life is now digitally mediated
  - Examples: Netflix/YouTube watch histories
  - Terms you will hear: *digital exhaust or passive data; digital footprints/breadcrumbs*
  - Main attribute: it is generated organically as users interact with digital technologies
- Storage of records
  - Consider digitized attendance sheets
  - Are they different from data produced by sensors?

Classic examples

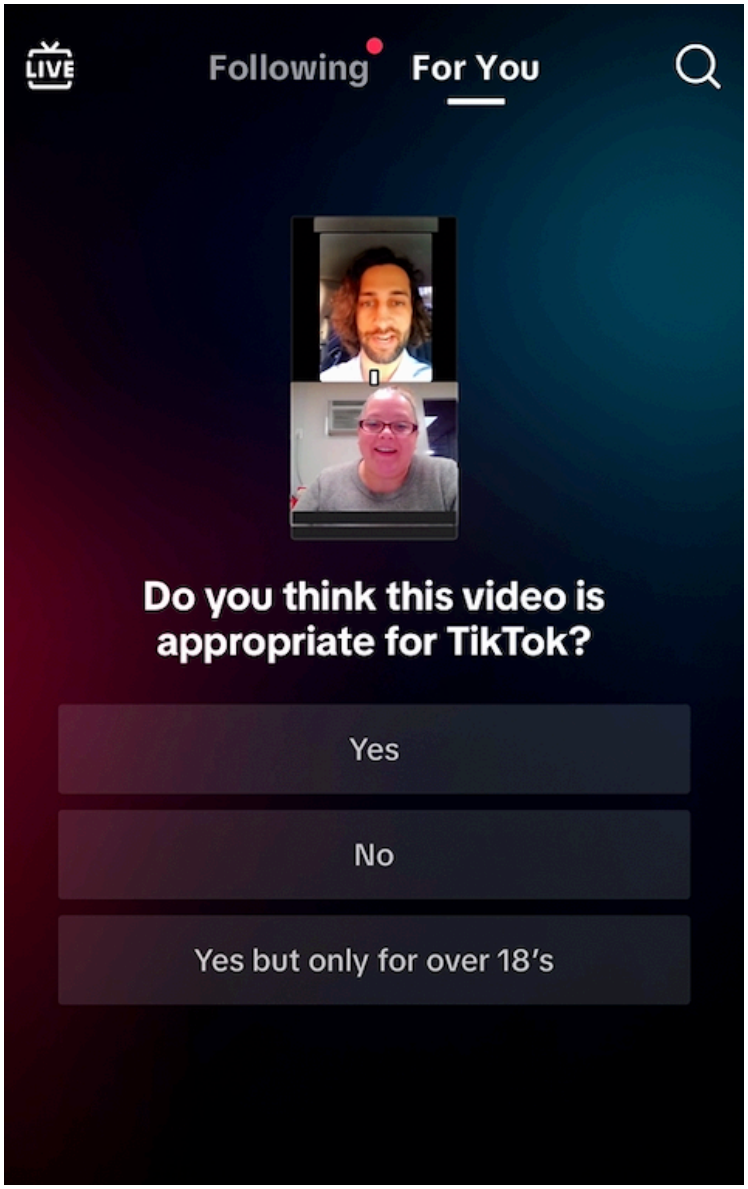
- Web browsing & web searches
- Social media use
- Mobile phone use
- Steps tracking
- Comments on blogs and internet fora
- Online purchases

Storing and digitizing

- Payments in a store with a credit card
- Historical texts/archives and audio-visual data which have been digitized

And also:

- Surveys you fill online
- You decision to report content on TikTok as offensive or inappropriate


LIVE
Following
For You
Do you think this video is appropriate for TikTok?
Yes
No
Yes but only for over 18's

Stylized examples

Web browsing & web searches, social media use, mobile phone use, fitness tracking…

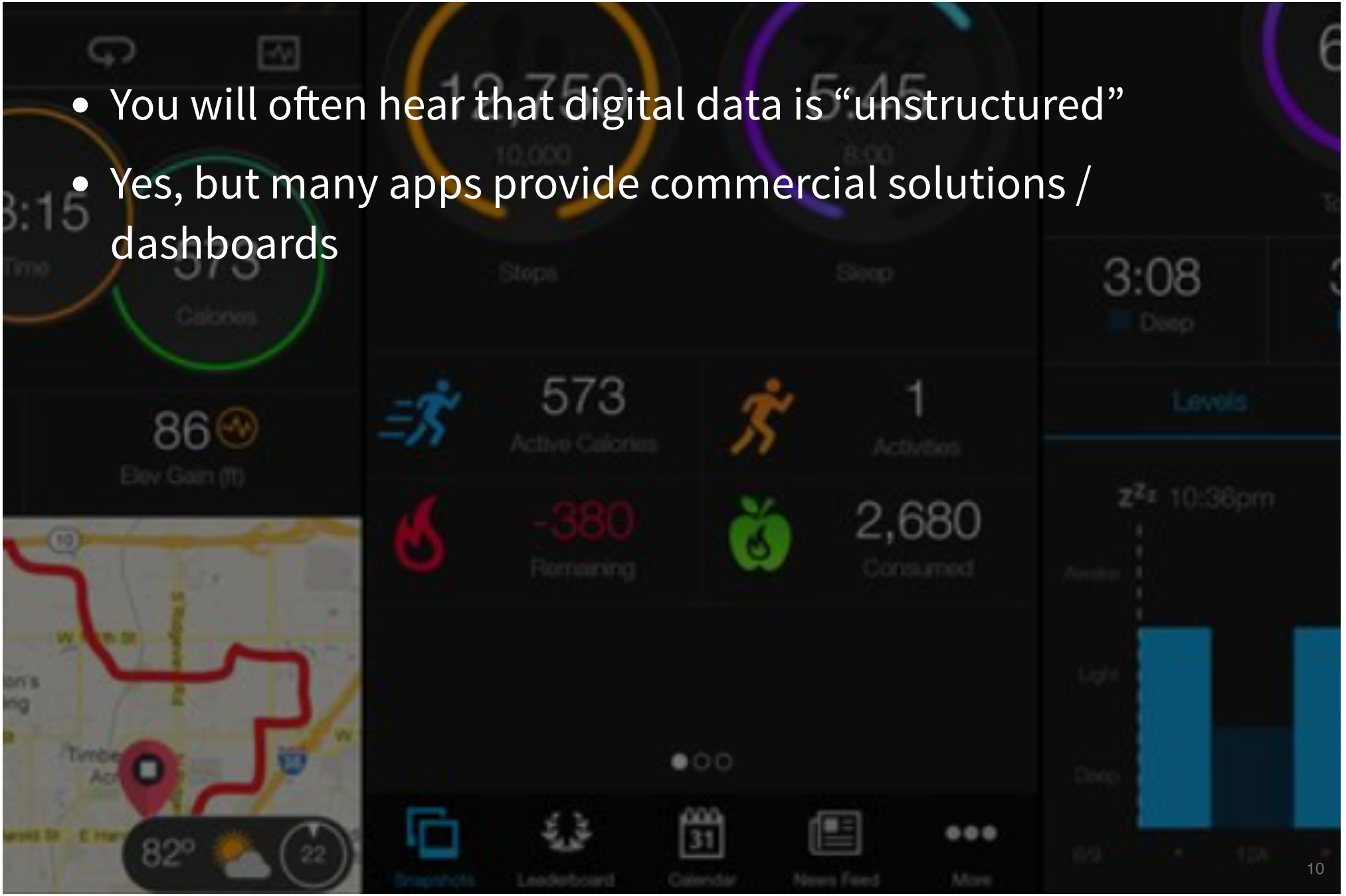
But remember

- There are still time use surveys (AHTUS)
- Gfk MRI asks respondents: “Did you watch movie X in the last 6 months?”
  - Netfilx knows for sure
  - Or does it?
- Inflation tracking involves visits to retail to stores

Characteristics

- Data is (potentially) continuously collected
- Typically unobtrusive
  - Potentially providing a **more authentic snapshot** of behavior, preferences
  - “Non-reactive” aspect emphasizes these is no visible prompting from researchers
- Collection of data may be concealed
  - We will talk about surveillance and ethics
- Often augmented with metadata (location at the time of measurement; social relationships; etc.)

- You will often hear that digital data is “unstructured”
- Yes, but many apps provide commercial solutions / dashboards



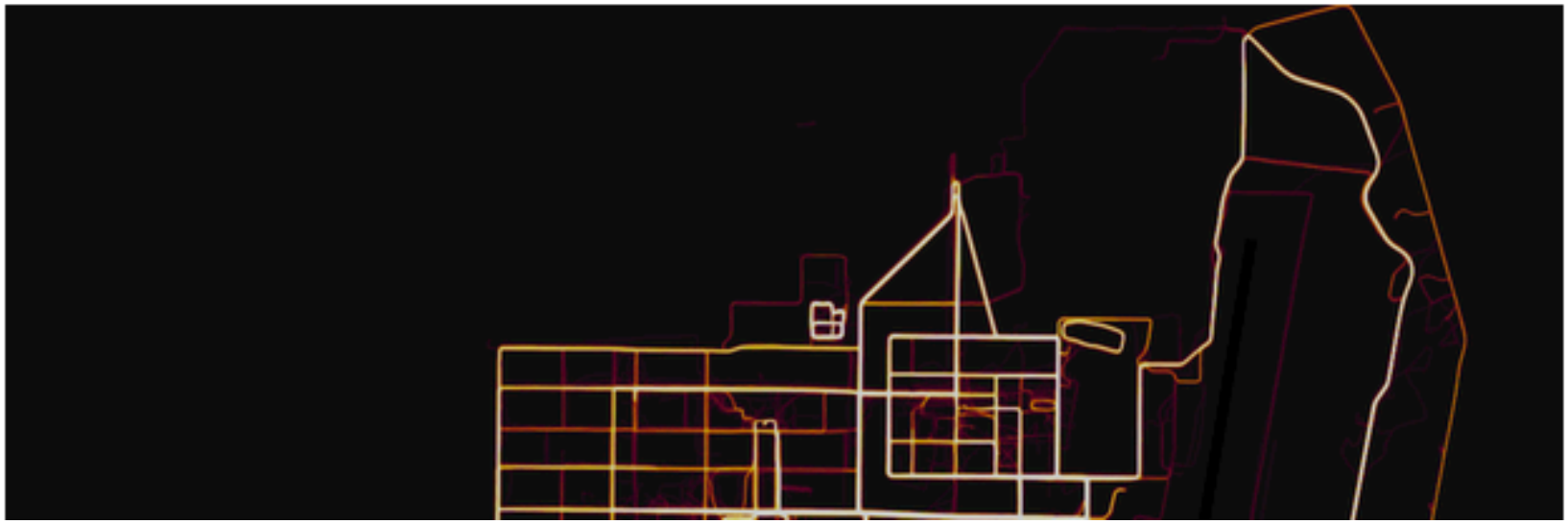
Dynamic responses

- One issue we'll come back to: do we think wearing a smartwatch leads to changes behaviors?
  - If you forget your watch at a friend's place, will you go for a run anyway?
  - This is not just a "data incompleteness" issue
- Distinct issues
  - The Hawthorne Effect: a psychological phenomenon where individuals modify (or improve) their behavior in response to their awareness of being observed
  - Selection bias

Fitness tracking app Strava gives away location of secret US army bases

Data about exercise routes shared online by soldiers can be used to pinpoint overseas facilities

- **Latest: Strava suggests military users 'opt out' of heatmap as row deepens**



Applications across fields

Social science

- Twitter has been used to study how (mis)information spreads; great interest in network structures
- Many attempts to measure public opinion in real time

Public health

- Time spent away from home, time spent at workplaces, etc. (Google Mobility reports)

Business

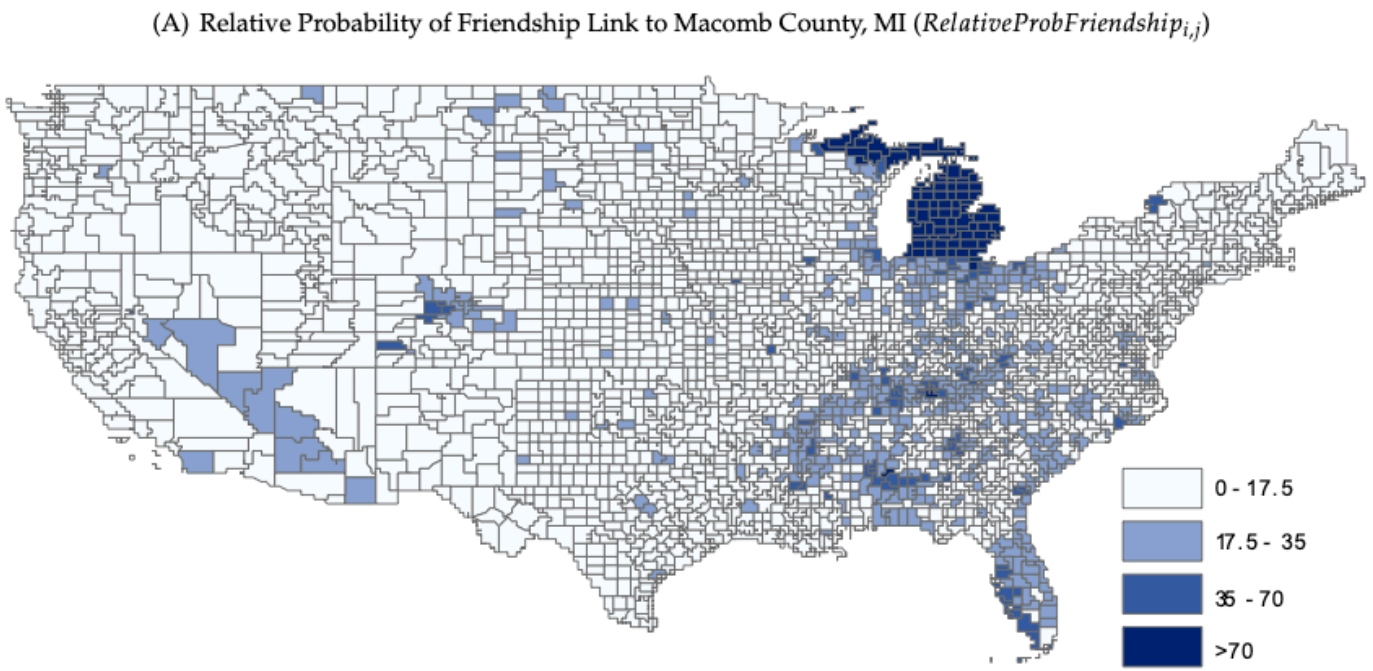
- Foot Traffic Data (Safegraph)

Concrete examples

Social relationships

Facebook Social Connectedness Index

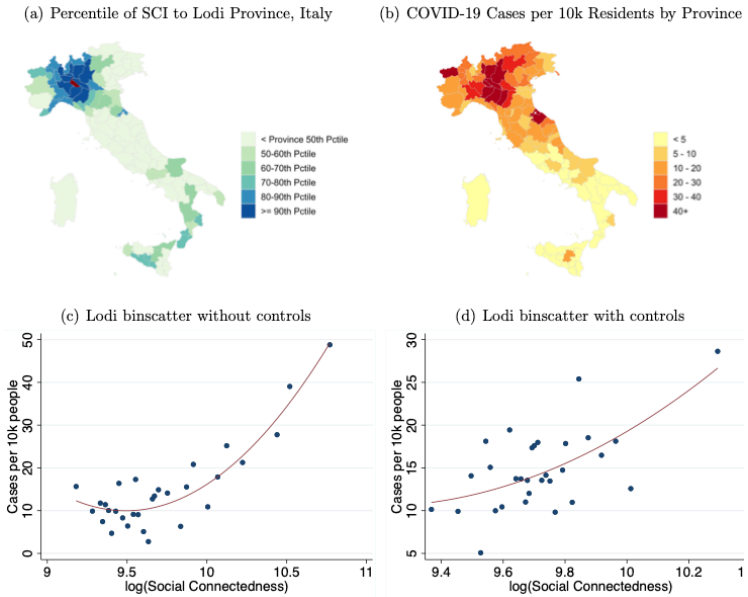
The Social Connectedness Index (SCI) is a measure of the social connectedness between different geographies. Specifically, expressing the relative probability that two individuals across two locations are friends with each other on Facebook.



https://data.humdata.org/dataset/social-connectedness-index

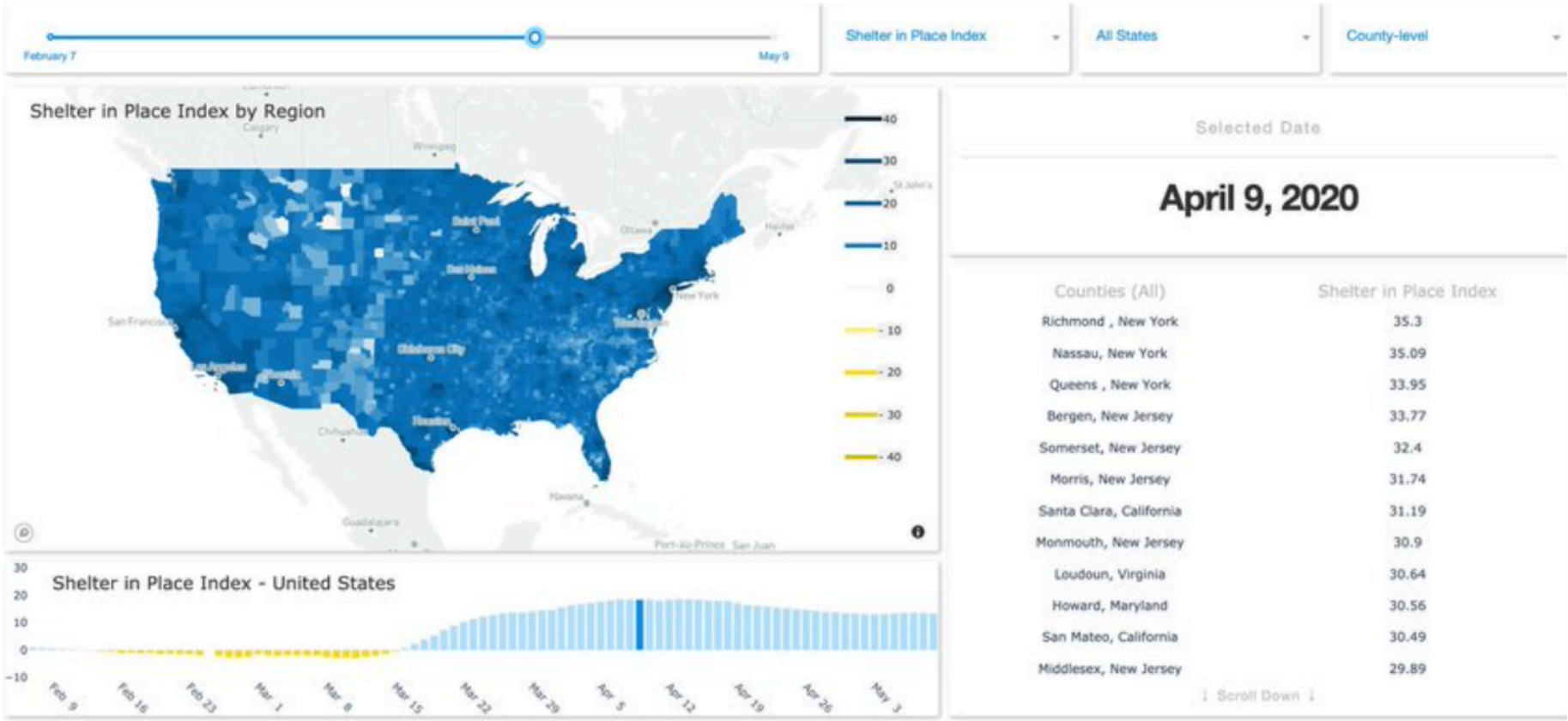
Public health

> Lodi is an Italian province of around 230,000 inhabitants in the heavily impacted region of Lombardy. It contains Codogno, where the earliest cases of COVID-19 in Italy were detected, and was at the center of Italy's outbreak



Paper: Kuchler et al. The geographic spread of COVID-19 correlates with the structure of social networks as measured by Facebook

Public health



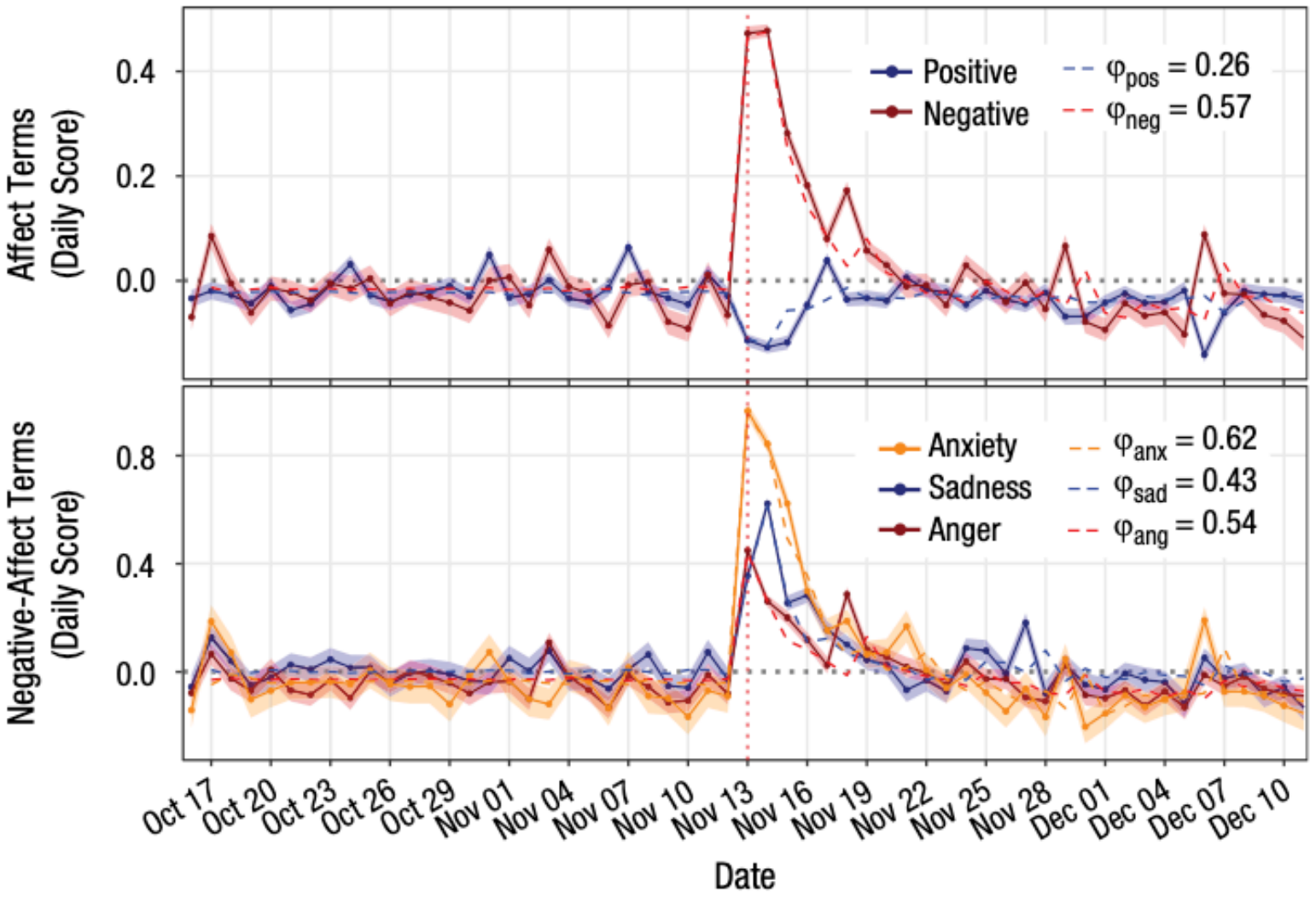
Originally at https://www.safegraph.com/data-examples/covid19-shelter-in-place but as of 2023 no longer available.

Psychology

- Goal: understand how national emotions changed after a tragic event
- Ask yourself: what would the ideal dataset contain?

Garcia and Rime. 2019. Collective Emotions and Social Resilience in the Digital Traces After a Terrorist Attack

- After November 13, 2015, the authors collected tweets with hashtags related to the terrorist attacks in Paris.
- Applied language detection to each tweet in combination with language metadata provided by Twitter (keeping only tweets in French)
- One (bad) option: hashtag tweets as a corpus. Instead: identify a panel of user accounts
- Of the users found in the initial data set 287K (58%) shared location information in their profile
- 62,114 user accounts in France


Affect Terms
(Daily Score)
0.4
0.2
0.0
Positive
Negative
φ_{pos} = 0.26
φ_{neg} = 0.57
Negative-Affect Terms
(Daily Score)
0.8
0.4
0.0
Anxiety
Sadness
Anger
φ_{anx} = 0.62
φ_{sad} = 0.43
φ_{ang} = 0.54
Oct 17
Oct 20
Oct 23
Oct 26
Oct 29
Nov 01
Nov 04
Nov 07
Nov 10
Nov 13
Nov 16
Nov 19
Nov 22
Nov 25
Nov 28
Dec 01
Dec 04
Dec 07
Dec 10
Date

Drawbacks

Weaknesses of Digital Trace Data

- Illusion of completeness (suppose your study participants agree to install a web tracker)
- Often inaccessible
  - Or sudden loss of access
- As monitoring improve, the incentive to manipulate signals increases
- Normalizing constant data collection

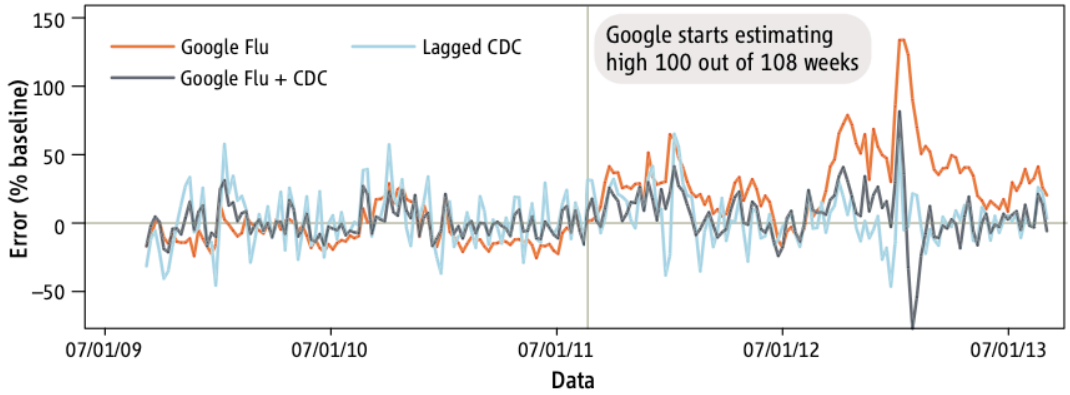
"Campaigns and companies, aware that news media are monitoring Twitter, have used numerous tactics to make sure their candidate or product is trending" (Lazer et al. 2014)

There seem to be unprecedented scientific possibilities thanks to new data collection methods, but…

> “Big data hubris” is the often implicit assumption that big data are a substitute for, rather than a supplement to, traditional data collection and analysis

David Lazer, Ryan Kennedy, Gary King, Alessandro Vespignani. The Parable of Google Flu: Traps in Big Data Analysis

Google Flu Trends



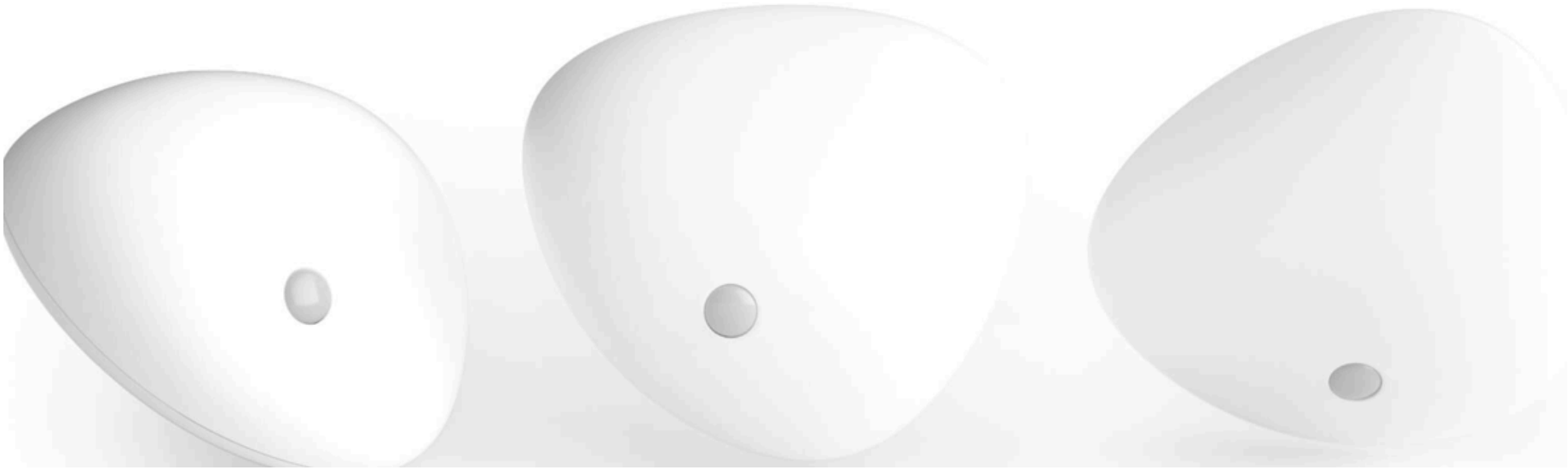
> **Algorithm dynamics** are the changes made by engineers to improve the commercial service and by consum- ers in using that service. Several changes in Google's search algorithm and user behav- ior likely affected GFT's tracking.

> There are multiple challenges to replicating GFT's original algorithm. GFT has never documented the 45 search terms used, and the examples that have been released appear misleading.

WIRELESS

Occupancy Sensor (Workspace)

The Occupancy Sensor knows whether the desk is occupied or free and reports the information to the server. It is used to track workspace vacancies, collecting anonymised data that is later analysed on the backend.



Universities are tracking their students. Is it clever or creepy?

Learning analytics are becoming increasingly popular for improving learning and cutting drop-out rates – but critics question the impact on privacy

The Guardian

Emotional data

The next frontier for learning analytics is feelings. Research is already probing the role of emotions in a student's university experience, and analysts are developing theories about how this "emotional data" can be captured.

Mobile phones can be used to monitor eye movements and body pose, for example, and this data could be used to tailor individual educational programmes based on the teaching methods that elicit the greatest positive emotional response.

AMERICA

Harvard Secretly Photographed Classrooms To Monitor Attendance

November 6, 2014 · 4:31 PM ET

By Sam Sanders

Credits, references, resources

- APIs for social scientists
- Example code for the Social Connectedness Index
  - Michael Bailey, Rachel Cao, Theresa Kuchler, Johannes Stroebel, and Arlene Wong. 2018. “Social Connectedness: Measurement, Determinants, and Effects.” *Journal of Economic Perspectives*, 32(3).
- A crowd-sourced list of datasets from Chris Bail
  - You can also download your viewing history from Netflix
- Thanks also Chris Bail, Carsten Schwemmer, and Oriol J. Bosch