

Digital Trace Data for Social Research

Ridhi Kashyap

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

SICSS-Oxford
July 6, 2022

The Digital Revolution

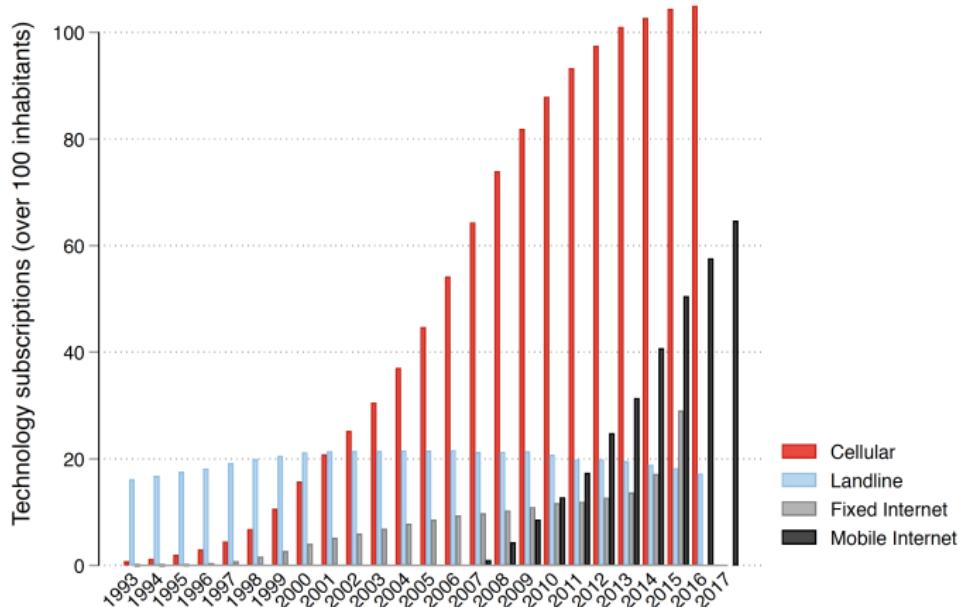


Figure: Technology subscriptions per capita, 1993–2017. Data from International Telecommunications Union (ITU).

The Digital Revolution

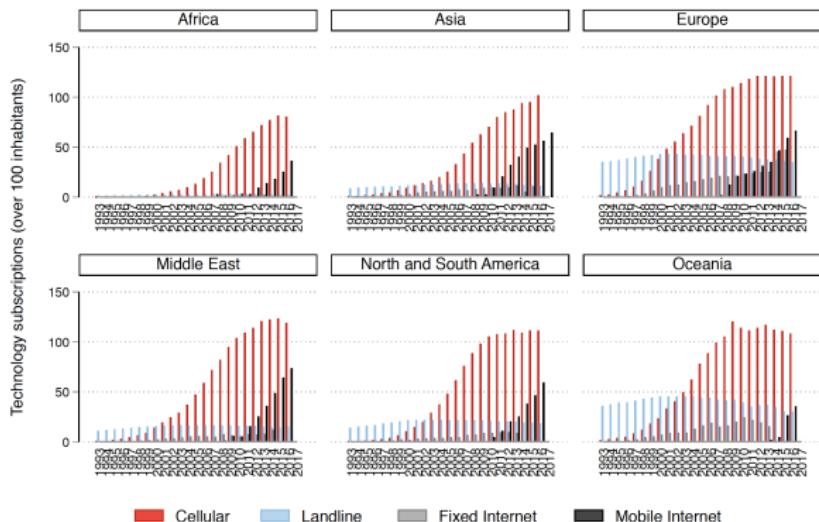


Figure: Technology subscriptions per capita, by global regions, 1993–2017. Data from International Telecommunications Union (ITU).

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

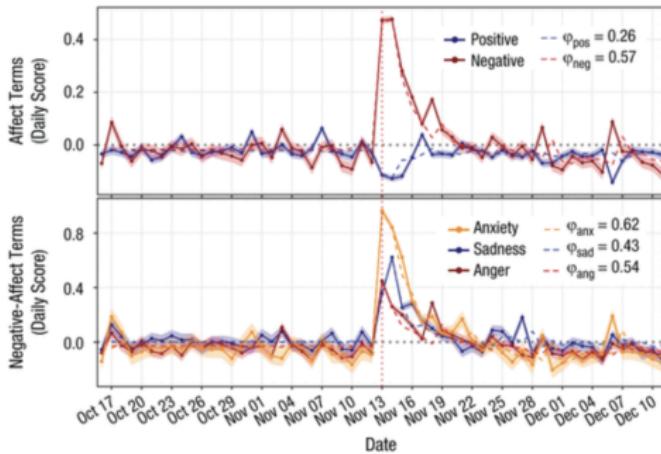
Common Features of Digital Trace Data

- ▶ 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



¹Garcia, 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

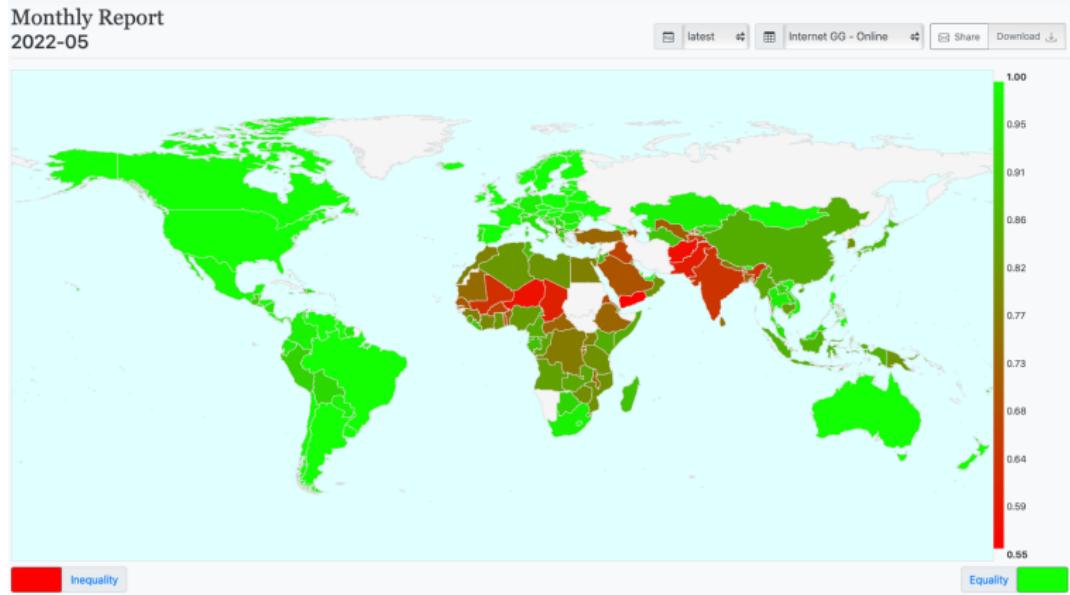


Figure: Gender gaps in internet use computed using data from Facebook (online model) available at www.digitalgendergaps.org

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

LinkedIn Gender Gap Index

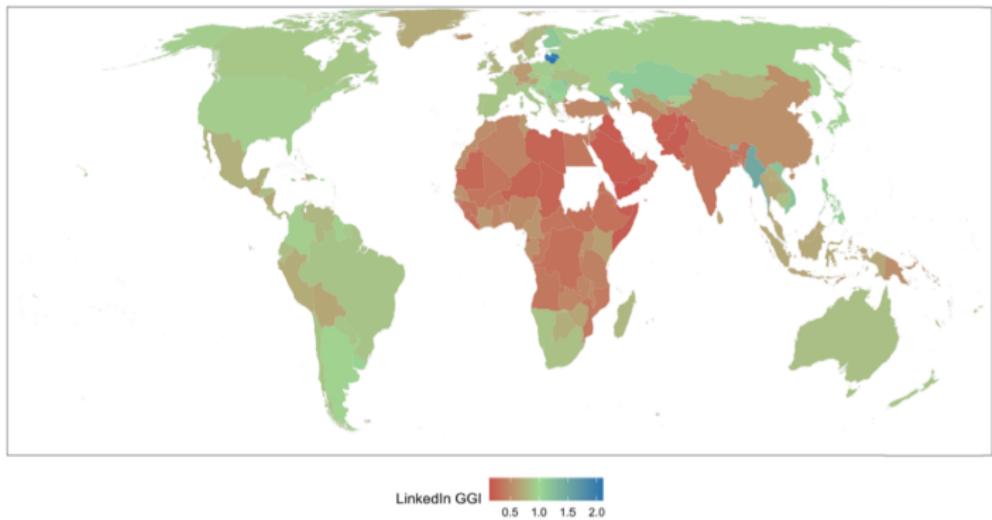


Figure: LinkedIn gender gap index (GGI) (female-to-male ratio of LinkedIn users), across the world

¹Kashyap, Ridhi and Florianne Verkroost "Analysing global professional gender gaps using LinkedIn advertising data." *EPJ Data Science* 10(1): 39

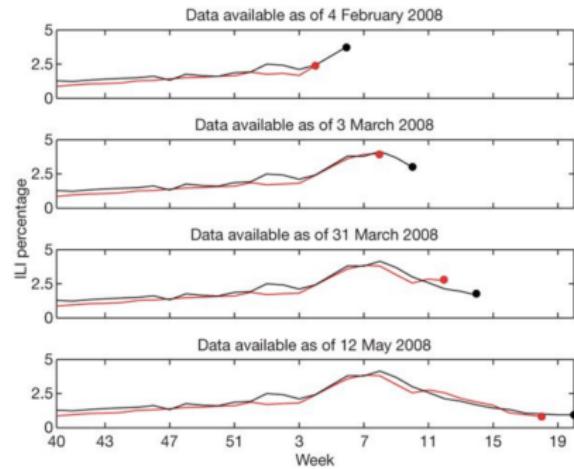
Digital Trace Data: Examples

- ▶ Social media sites
- ▶ Web search queries

Google Flu

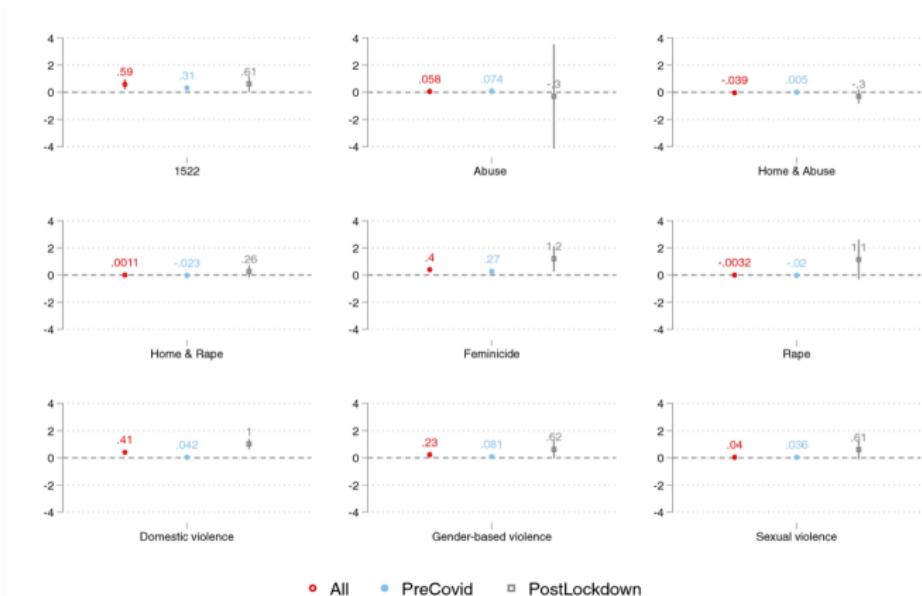
Figure 3 : ILI percentages estimated by our model (black) and provided by the CDC (red) in the mid-Atlantic region, showing data available at four points in the 2007-2008 influenza season.

From: Detecting influenza epidemics using search engine query data



¹Ginsberg, 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



Coefficient plot from regressions of daily 1522 valid calls on Google searches, by selected keywords (whole Italy), lagged predictors

Figure: Coefficient plot from regressions of daily IPV hotline calls on Google searches, by selected keywords (Italy), lagged predictors.

¹Köksal, 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

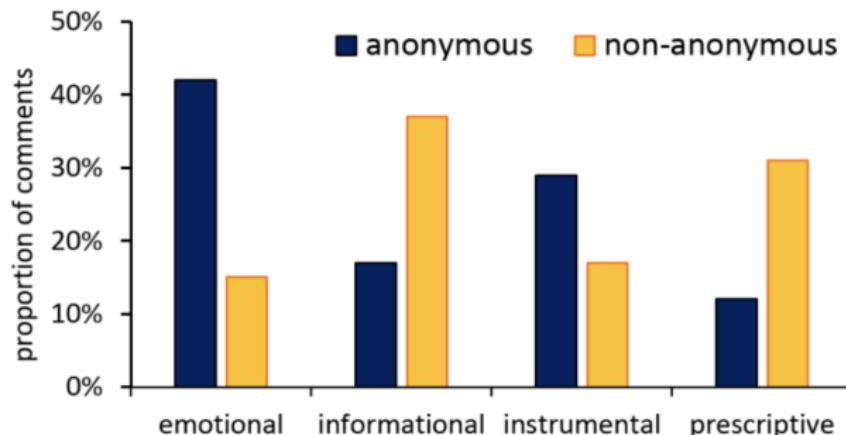


Figure 4. Posts from anonymous and non-anonymous accounts in the light of social support types.

¹De 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

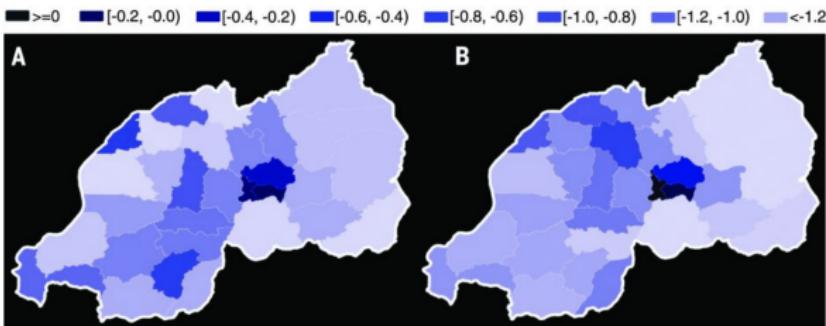


Figure: (A) Predicted composite wealth index (district average), computed from 2009 call data and aggregated by administrative district. (B) Actual composite wealth index (district average), as computed from a 2010 government DHS of 12,792 households.

¹Blumenstock, 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

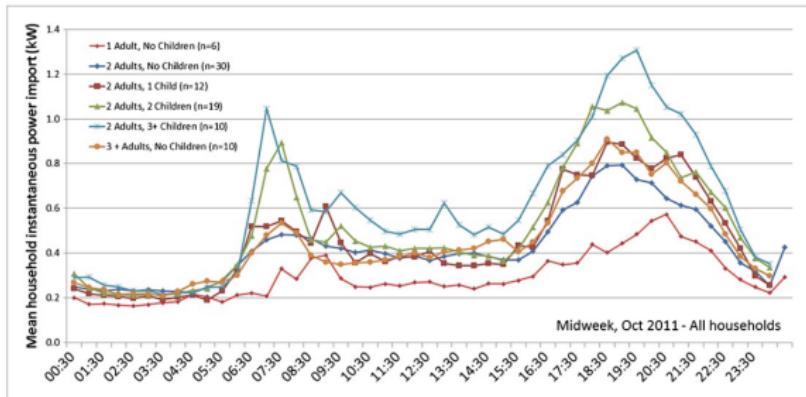


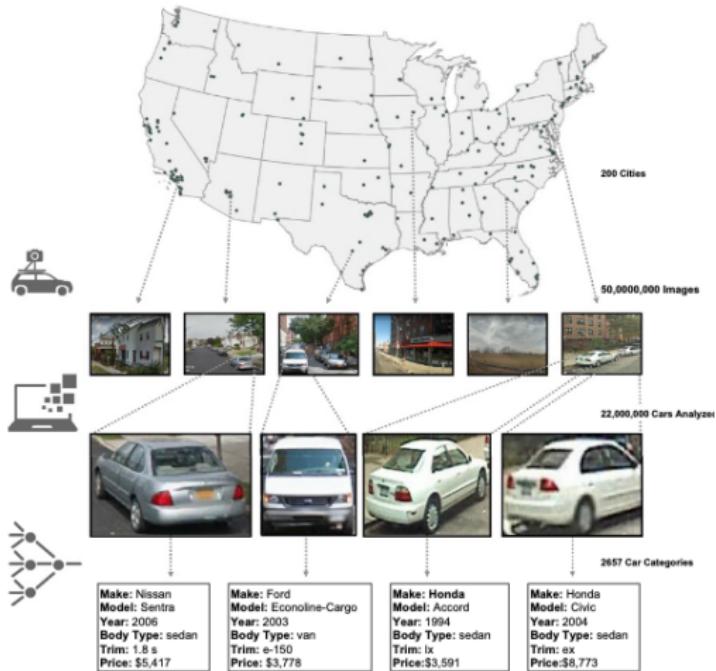
Figure 2

[Open in figure viewer](#) | [PowerPoint](#)

Household load profiles by household composition.

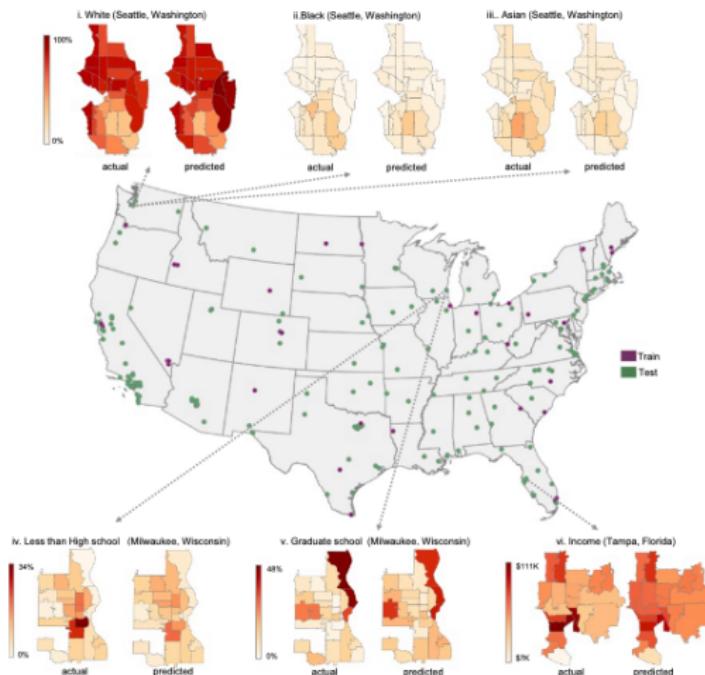
¹Newing, Andy, Ben Anderson, AbuBakr Bahaj, and Patrick James. "The role of digital trace data in supporting the collection of population statistics—The case for smart metered electricity consumption data." *Population, Space and Place* 22, no. 8 (2016): 849-863.

Google Street View



¹Gebru, 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



¹Gebru, 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

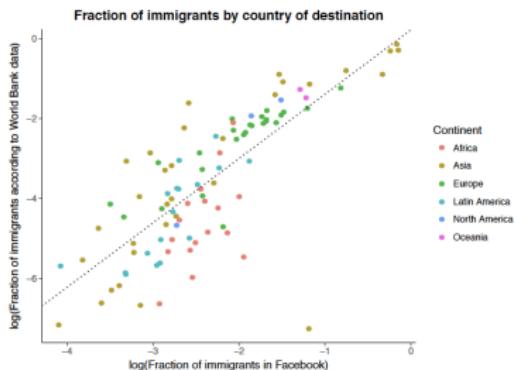
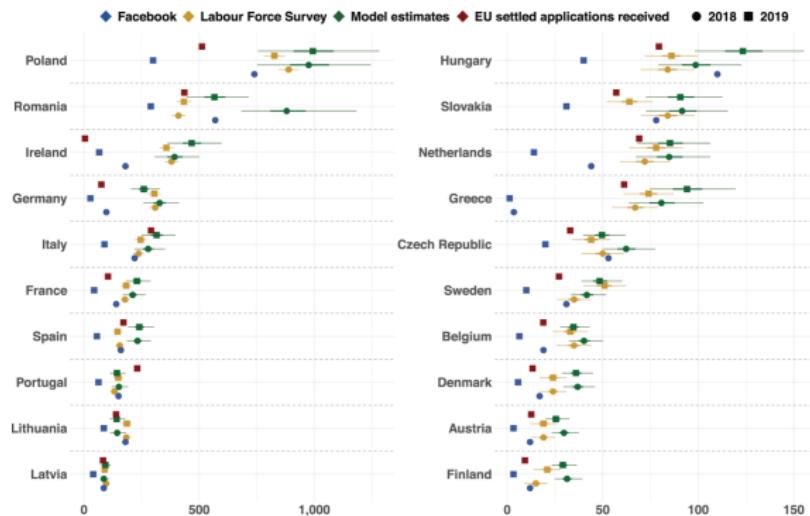


Figure 2: Relationship between stocks of migrants from the Facebook data set, for countries with at least one million Facebook users as of 2016, and the respective estimates from the World Bank (2015). The data points indicate the fraction of immigrants in the population, on a log scale, by country of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

¹Zagheni, 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



¹Rampazzo, 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

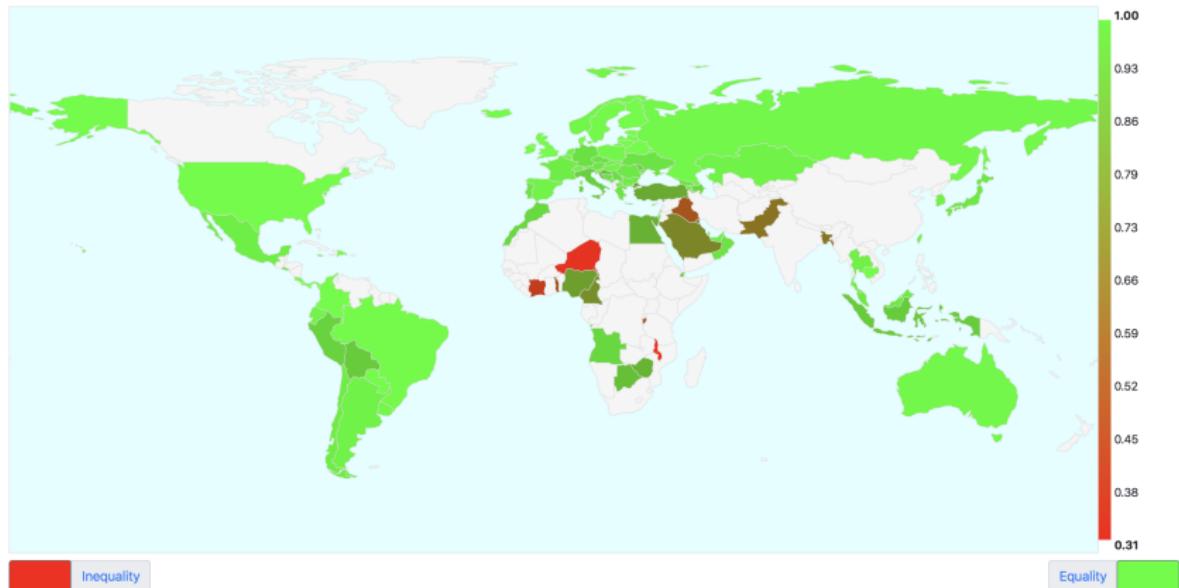


Figure: Gender gaps in internet use computed using data from International Telecommunications Union (ITU) available at www.digitalgendergaps.org

Example (2): Predicting Internet Gender Gaps

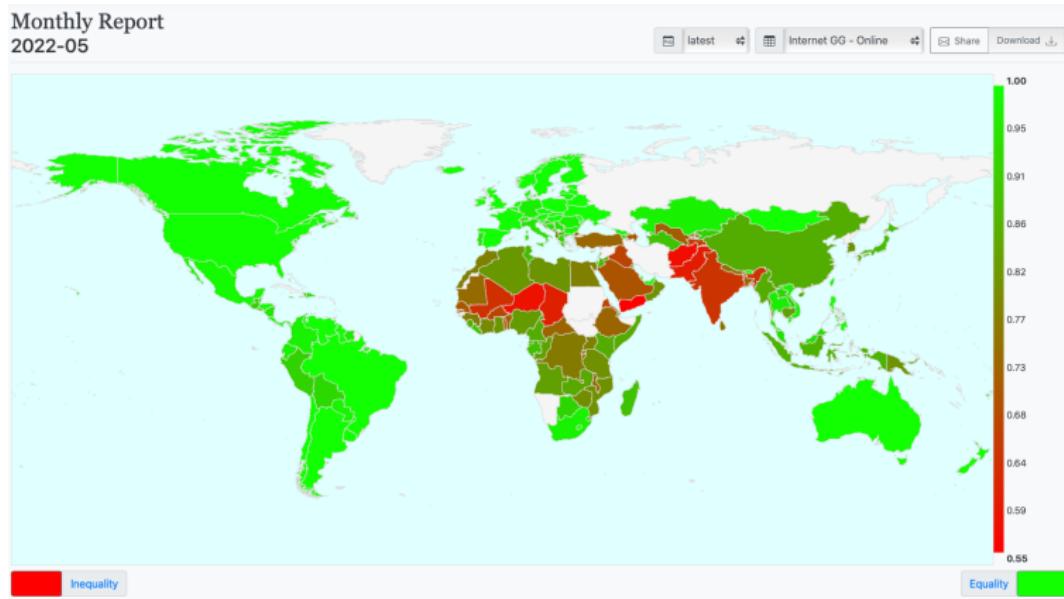


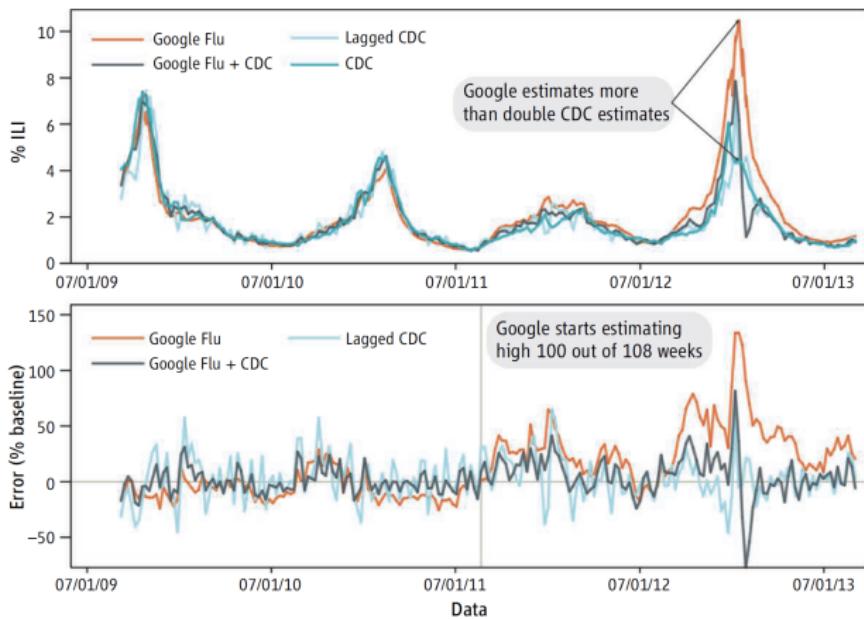
Figure: Gender gaps in internet use computed using data from Facebook (online model) available at www.digitalgendergaps.org

¹Fatehkia, 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



¹Lazer, 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

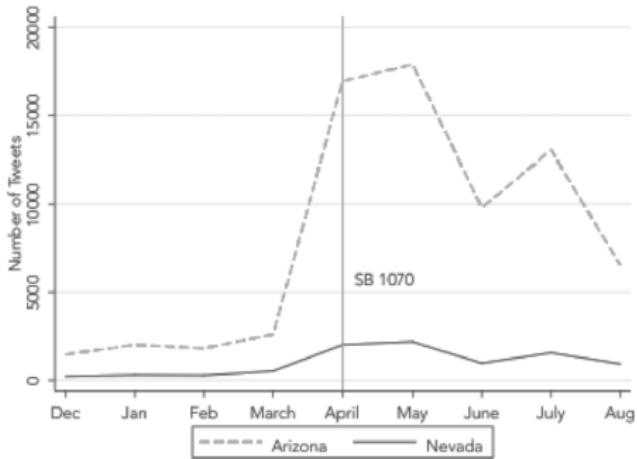


FIG. 2.—Number of Twitter messages related to immigrants per month in Arizona and Nevada (December 2010–August 2011). The vertical line on April 2010 indicates when the Arizona governor approved SB 1070.

¹Flores, 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

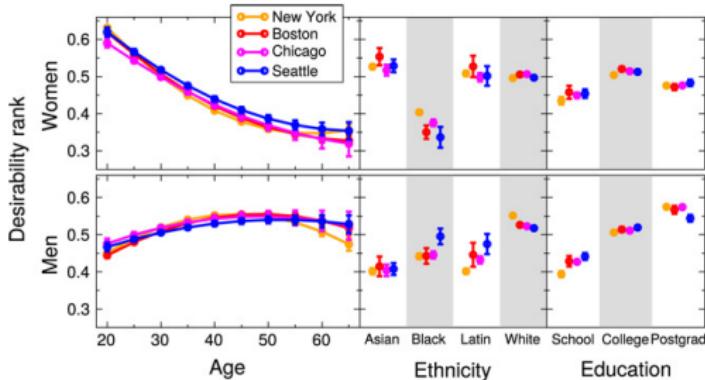


Fig. 2 Desirability, quantified using the measures defined here, as a function of demographic variables of the user population.

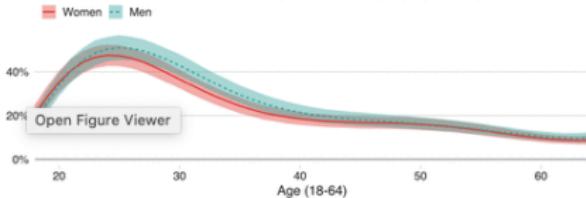
¹Bruch, Elizabeth E., and M. E. J. Newman. "Aspirational pursuit of mates in online dating markets." *Science Advances* 4, no. 8 (2018): eaap9815.

Research Designs using Digital Trace Data

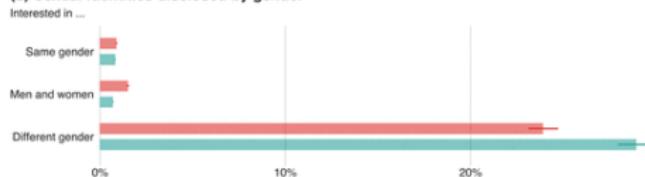
- ▶ Lens for social measurement
 - ▶ Nowcasting social phenomena
 - ▶ Analyzing the impacts of events and shocks in real-time
 - ▶ Measuring tricky, hard-to-capture human behaviours
- ▶ Digitalized lives and the implications of digitalization
 - ▶ Understanding digitalized lives and social dynamics in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

Example: Sexuality Disclosure on Social Media

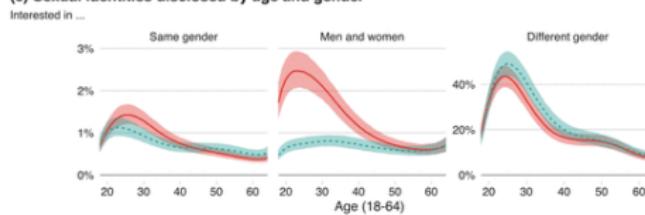
(a) Proportion of US Facebook users disclosing any sexuality by age and gender



(b) Sexual identities disclosed by gender



(c) Sexual identities disclosed by age and gender

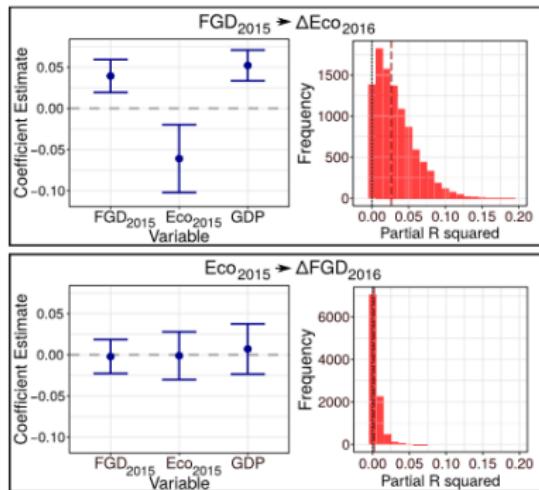


¹Gilroy, Connor, and Ridhi Kashyap. "Digital traces of sexualities: understanding the salience of sexual identity through disclosure on social media." *Socius* 7 (2021): 23780231211029499.

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: Social Media Gender Gaps



¹Garcia, 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.

Accessing Digital Trace Data

- ▶ Screen scraping
- ▶ Querying data from application programming interfaces (APIs) [more on this with Doug and Tobi]
- ▶ Archives or data sharing platforms (e.g. Google Trends).
- ▶ Data sharing agreements with data owners.

Screen Scraping

- ▶ The process of extracting data from web pages in an automated way
- ▶ Lots of data are available on the web – but not all of it is appropriate to use for research.

Considerations when Screen Scraping

- ▶ Read the website's Terms of Service: are you allowed to do this?
- ▶ Larger websites like Facebook, Instagram, NY Times do not allow these practices – but some provide structured access to data through APIs.
- ▶ Even if terms of service does not prohibit this – are the data sensitive? Could the use of the data harm in some way? Is their use ethical?

A Simple Example: Wikipedia Page

World Bank Group (2020) [edit]

Data of the World Bank Group for 2020.^{[3][4][5]} The values in the World Bank Group tables are rounded. All calculations were done on raw data, therefore, due to the nuances of rounding places illusory inconsistencies of indicators arose, with a size of 0.01 year. Since this entry is prone to vandalism and political bias it is recommended to check the original source.

	Countries	all	male	female	gender gap	Δ 2019 all	Δ 2019 male	Δ 2019 female	Δ 2019 gen. gap
		♦	♦	♦	♦	♦	♦	♦	♦
1	Japan	84.62	81.64	87.74	6.10	0.26	0.23	0.29	0.06
2	Singapore	83.74	81.50	86.10	4.60	0.15	0.10	0.20	0.10
3	South Korea	83.43	80.50	86.50	6.00	0.20	0.20	0.20	0.00
4	Norway	83.21	81.60	84.90	3.30	0.25	0.30	0.20	-0.10
5	Australia	83.20	81.20	85.30	4.10	0.30	0.30	0.30	0.00
6	Switzerland	83.10	81.10	85.20	4.10	-0.80	-1.00	-0.60	0.40
7	Iceland	83.07	81.70	84.50	2.80	-0.10	0.00	-0.20	-0.20
8	Israel	82.70	80.70	84.80	4.10	-0.10	-0.30	0.10	0.40
9	Malta	82.65	80.80	84.60	3.80	-0.20	-0.40	0.00	0.40
10	Sweden	82.41	80.70	84.20	3.50	-0.70	-0.80	-0.60	0.20
11	Italy	82.34	80.10	84.70	4.60	-1.15	-1.30	-1.00	0.30

Figure:

https://en.wikipedia.org/wiki/List_of_countries_by_life_expectancy

Wikipedia Page HTML

```
1 <!DOCTYPE html>
2 <html class="client-nojs" lang="en" dir="ltr">
3 <head>
4 <meta charset="UTF-8"/>
5 <title>List of countries by life expectancy - Wikipedia</title>
6 <script>document.documentElement.className="client-js";RLCONF={"wgBreakFrames":false,"wgSeparatorTransformTable":[],"wgDigitTransformTable":[]}
7 "Life expectancy","wgPageContentLanguage":"en","wgPageContentModel":"wikitext","wgRelevantPageName":"List_of_countries_by_life_expectancy","wgRev...
8 RLSTATE={".ext.globalCsJs.user.styles":"ready","site.styles":"ready","user.styles":"ready","ext.globalCsJs.user":"ready","user":"ready","user.optio...
9 "ext.gadget.switcher","ext.centralauth.centralautologin","ext.popups","ext.uls.compactlinks","ext.uls.interface","ext.growthExperiments.SuggestedEdit...
10 <script>(RlQ=window.RlQ||[]).push(function(){mw.loader.implement("user.options#12s51",function($,jQuery,require,module){mw.user.tokens.set({"patrolT...
11 <link rel="stylesheet" href="//w/load.php?lang=en&modules=startup&only=scripts&raw=1&skin=vector"></script>
12 <script async="" src="//w/load.php?lang=en&modules=startup&only=scripts&raw=1&skin=vector"></script>
13 <meta name="ResourceLoaderDynamicStyles" content="">
14 <link rel="stylesheet" href="//w/load.php?lang=en&modules=site.styles&only=styles&skin=vector"/>
15 <link rel="stylesheet" href="//w/load.php?lang=en&modules=site.styles&only=styles&skin=vector"/>
16 <meta name="generator" content="MediaWiki 1.39.0-wmf.16"/>
17 <meta name="referrer" content="origin"/>
18 <meta name="referrer" content="origin-when-crossorigin"/>
19 <meta name="referrer" content="origin-when-cross-origin"/>
20 <meta name="format-detection" content="telephone=no"/>
21 <meta property="og:image" content="https://upload.wikimedia.org/wikipedia/commons/thumb/5/52/Comparison_gender_life_expectancy_WHO.svg/1200px-Compar...
22 <meta property="og:image:height" content="954"/>
23 <meta property="og:image" content="https://upload.wikimedia.org/wikipedia/commons/thumb/5/52/Comparison_gender_life_expectancy_WHO.svg/800px-Compar...
24 <meta property="og:image:width" content="800"/>
25 <meta property="og:image:height" content="636"/>
26 <meta property="og:image" content="https://upload.wikimedia.org/wikipedia/commons/thumb/5/52/Comparison_gender_life_expectancy_WHO.svg/640px-Compar...
27 <meta property="og:image:width" content="640"/>
28 <meta property="og:image:height" content="509"/>
29 <meta property="og:title" content="List of countries by life expectancy - Wikipedia"/>
30 <meta property="og:type" content="website"/>
31 <link rel="preconnect" href="//upload.wikimedia.org"/>
32 <link rel="alternate" media="only screen and (max-width: 720px)" href="//en.m.wikipedia.org/wiki/List_of_countries_by_life_expectancy"/>
33 <link rel="alternate" type="application/x-wiki" title="Edit this page" href="//w/index.php?title=List_of_countries_by_life_expectancy&action=edit...
34 <link rel="apple-touch-icon" href="/static/apple-touch/wikipedia.png"/>
35 <link rel="shortcut icon" href="/static/favicon/wikipedia.ico"/>
36 <link rel="search" type="application/opensearchdescription+xml" href="//w/opensearch_desc.php" title="Wikipedia (en)"/>
37 <link rel="EditURI" type="application/rdf+xml" href="//en.wikipedia.org/w/api.php?action=rdf"/>
38 <link rel="license" href="https://creativecommons.org/licenses/by-sa/3.0/"/>
39 <link rel="canonical" href="https://en.wikipedia.org/wiki/List_of_countries_by_life_expectancy"/>
40 <link rel="dns-prefetch" href="//meta.wikimedia.org" />
41 <link rel="dns-prefetch" href="//login.wikimedia.org" />
42 </head>
43 <body class="mediawiki ltr sitelid-ltr mw-hide-empty-elt ns-0 ns-subject mw-editable page-List_of_countries_by_life_expectancy rootpage-List_of_coun...
44 <div id="mw-head-base" class="noprint"></div>
45 <div id="content" class="mw-body" role="main">
46 <div id="top"></div>
47 <div id="contentSub" style="font-size: small; margin-top: 0.5em;">Content based on the English Wikipedia as of 11:44, 2023
```

48 </div>

A Simple Example: Wikipedia Page

```
library(rvest)
le_wiki <- read_html("https://en.wikipedia.org/wiki/List_of_countries_by_life_expectancy")
le_table <- html_node(le_wiki, xpath = '//*[@id="mw-content-text"]/div/table[1]')
le_rankings <- html_table(le_table1)
head(le_rankings)
```

A tibble: 6 × 9

Countries <chr>	all <dbl>	male <dbl>	female <dbl>	gendergap <dbl>	Δ 2019all <dbl>	Δ 2019male <dbl>	Δ 2019female <dbl>	▶
Japan	84.62	81.64	87.74	6.1	0.26	0.23	0.29	NA
Singapore	83.74	81.50	86.10	4.6	0.15	0.10	0.20	NA
South Korea	83.43	80.50	86.50	6.0	0.20	0.20	0.20	NA
Norway	83.21	81.60	84.90	3.3	0.25	0.30	0.20	NA
Australia	83.20	81.20	85.30	4.1	0.30	0.30	0.30	NA

6 rows | 1–8 of 9 columns

A Simple Example: Wikipedia Page

The screenshot shows a Chrome browser window with the URL <https://en.wikipedia.org/>. The main content is a Wikipedia article about life expectancy, mentioning healthy life expectancy and its reporting issues.

The developer tools sidebar is open, showing the "Developer" tab selected. It displays various developer-related options and keyboard shortcuts.

	Countries	all	male	female	gender gap	Δ 2019 all	Δ 2019 male	Δ 2019 female	Δ 2019 gen. gap
#	♦	♦	♦	♦	♦	♦	♦	♦	♦
1	🇯🇵 Japan	84.62	81.64	87.74	6.10	0.26	0.23	0.29	0.06
2	🇸🇬 Singapore	83.74	81.50	86.10	4.60	0.15	0.10	0.20	0.10
3	🇰🇷 South Korea	83.43	80.50	86.50	6.00	0.20	0.20	0.20	0.00
4	🇳🇴 Norway	83.21	81.60	84.90	3.30	0.25	0.30	0.20	-0.10
5	🇦🇺 Australia	83.20	81.20	85.30	4.10	0.30	0.30	0.30	0.00
6	🇨🇭 Switzerland	83.10	81.10	85.20	4.10	-0.80	-1.00	-0.60	0.40
7	🇮🇸 Iceland	83.07	81.70	84.50	2.80	-0.10	0.00	-0.20	-0.20
8	🇮🇱 Israel	82.70	80.70	84.80	4.10	-0.10	-0.30	0.10	0.40
9	🇲🇹 Malta	82.65	80.80	84.60	3.80	-0.20	-0.40	0.00	0.40

A Simple Example: Wikipedia Page

- the average number of years that a person can expect to live in "full health".^{[1][2]}

Comparing life expectancies across countries can be problematic. For example, due to poor reporting in some countries and various local standards in collecting statistics. This is especially true for Healthy life expectancy, the definition of which criteria may change over time, even within a country.

World Bank Group (2020) [edit]

Data of the World Bank Group for 2020.^{[3][4][5]} The values in the World Bank Group tables are rounded. All

table.wikitable.sortable.static-row-numbers.plainrowheaders.srnr-white-background.jqu...
ue to the nuances of rounding, in some places illusory
0.01 year. Since this entry is prone to vandalism and political
ce.

	Countries	all	male	female	gender gap	A.2019 all	A.2019 male	A.2019 female	A.2019 gen. gap
		♦	♦	♦	♦	♦	♦	♦	♦
1	Japan	84.62	81.64	87.74	6.10	0.26	0.23	0.29	0.06
2	Singapore	83.74	81.50	86.10	4.60	0.15	0.10	0.20	0.10
3	South Korea	83.43	80.50	86.50	6.00	0.20	0.20	0.20	0.00
4	Norway	83.21	81.60	84.90	3.30	0.25	0.30	0.20	-0.10
5	Australia	83.20	81.20	85.30	4.10	0.30	0.30	0.30	0.00
6	Switzerland	83.10	81.10	85.20	4.10	-0.80	-1.00	-0.60	0.40

The screenshot shows the Chrome DevTools context menu open over a table cell. The 'Copy' option is highlighted in blue, indicating it is selected. Other options like 'Cut', 'Paste', and 'Copy XPath' are also visible.

```
> <style data-mw-deduplicate="TemplateStyles:r1094137577" ...>
...   <table class="wikitable sortable static-row-numbers plainrowheaders srnr-white-background jqu...>
...     <tr>
...       <th> Countries </th>
...       <th> all </th>
...       <th> male </th>
...       <th> female </th>
...       <th> gender gap </th>
...       <th> A.2019 all </th>
...       <th> A.2019 male </th>
...       <th> A.2019 female </th>
...       <th> A.2019 gen. gap </th>
...     </tr>
...     <tr>
...       <td> 1 </td>
...       <td> Japan </td>
...       <td> 84.62 </td>
...       <td> 81.64 </td>
...       <td> 87.74 </td>
...       <td> 6.10 </td>
...       <td> 0.26 </td>
...       <td> 0.23 </td>
...       <td> 0.29 </td>
...       <td> 0.06 </td>
...     </tr>
...     <tr>
...       <td> 2 </td>
...       <td> Singapore </td>
...       <td> 83.74 </td>
...       <td> 81.50 </td>
...       <td> 86.10 </td>
...       <td> 4.60 </td>
...       <td> 0.15 </td>
...       <td> 0.10 </td>
...       <td> 0.20 </td>
...       <td> 0.10 </td>
...     </tr>
...     <tr>
...       <td> 3 </td>
...       <td> South Korea </td>
...       <td> 83.43 </td>
...       <td> 80.50 </td>
...       <td> 86.50 </td>
...       <td> 6.00 </td>
...       <td> 0.20 </td>
...       <td> 0.20 </td>
...       <td> 0.20 </td>
...       <td> 0.00 </td>
...     </tr>
...     <tr>
...       <td> 4 </td>
...       <td> Norway </td>
...       <td> 83.21 </td>
...       <td> 81.60 </td>
...       <td> 84.90 </td>
...       <td> 3.30 </td>
...       <td> 0.25 </td>
...       <td> 0.30 </td>
...       <td> 0.20 </td>
...       <td> -0.10 </td>
...     </tr>
...     <tr>
...       <td> 5 </td>
...       <td> Australia </td>
...       <td> 83.20 </td>
...       <td> 81.20 </td>
...       <td> 85.30 </td>
...       <td> 4.10 </td>
...       <td> 0.30 </td>
...       <td> 0.30 </td>
...       <td> 0.30 </td>
...       <td> 0.00 </td>
...     </tr>
...     <tr>
...       <td> 6 </td>
...       <td> Switzerland </td>
...       <td> 83.10 </td>
...       <td> 81.10 </td>
...       <td> 85.20 </td>
...       <td> 4.10 </td>
...       <td> -0.80 </td>
...       <td> -1.00 </td>
...       <td> -0.60 </td>
...       <td> 0.40 </td>
...     </tr>
...   </table>
... </div>
... <div style="background-color: #f0f0f0; padding: 5px; margin-top: 10px; font-size: small;">...
```

Some Limits

- ▶ This was a simple example. Often web pages are complex with many different elements.
- ▶ This yields messy data.
- ▶ Screen scraping can be frustrating or unfeasible. What then?
 - ▶ For complex web pages, crowd worker platforms (e.g. Mechanical Turk) could be an option.
 - ▶ Some web data can be accessed via APIs.

What is an API?

- ▶ Application programming interfaces or APIs are a software intermediary that allows two applications to talk to each other.
- ▶ Web APIs allow one computer (a client) to ask another computer (a server) for some resource over the internet.
- ▶ APIs provide a structured way to access data that are stored in databases that are continuously updated.

Application Programming Interface

- ▶ Modern APIs adhere to standards, that make data exchange programmatically accessible, safe and structured
- ▶ Contrast with web scraping.
- ▶ Two important concepts when using APIs
 - ▶ Credentialling
 - ▶ Rate limiting

API Directory

The screenshot shows the ProgrammableWeb API Directory homepage. At the top, there's a navigation bar with links for 'API DIRECTORY' and 'API NEWS'. Below the navigation is a search bar with the placeholder 'Search over 21,780 APIs and much more'. To the right of the search bar are social media icons for LinkedIn, Facebook, Twitter, and YouTube. A blue button labeled 'ADD API & MORE' is also present.

Below the navigation is a banner for 'MuleSoft Applying DevOps to APIs' with a 'Learn more' button.

The main content area features a large heading 'Search the Largest API Directory on the Web' with a search input field and a 'SEARCH APIs' button. Below this is a 'Filter APIs' section with dropdown menus for 'By Category' and 'Include Deprecated APIs'.

A table lists various APIs with columns for 'API Name', 'Description', 'Category', and 'Submitted'. The listed APIs include:

API Name	Description	Category	Submitted
Google Maps	[This API is no longer available. Google Maps' services have been split into multiple APIs, including the Static Maps API.]	Mapping	12.05.2005
Twitter	[This API is no longer available. It has been split into multiple APIs, including the Twitter Ads API, Twitter Search Tweets, ...]	Social	12.08.2006
YouTube	The Data API allows users to integrate their program with YouTube and allow it to perform many of the operations available on the website. It provides the capability to search for videos, retrieve...	Video	02.08.2006
Flickr	The Flickr API can be used to retrieve photos from the Flickr photo sharing service using a variety of feeds - public photos and videos, favorites, friends, group pools, discussions, and more. The...	Photos	09.04.2005
Facebook	[This API is no longer available. Its functions have been split among the following APIs: Facebook Ads, ...]	Social	08.16.2006

To the right of the main content is a sidebar for 'API UNIVERSITY' with sections for 'FEATURED' and 'LATEST'. It includes links to 'What Are APIs and How Do They Work?', '8 Real World API Strategies and the Keys to Their Success', 'Microservices 101: Understanding and Leveraging Microservices', and 'More for API Providers >'. Another section for 'FOR DEVELOPERS' includes links to 'How to Get Started With Google Actions', 'How to Build a Monitoring Application With the Google Cloud Vision API', 'How to Access Any RESTful API Using the R Language', and 'More for Developers >'. At the bottom of the sidebar is a 'Today in APIs' section with a 'SUBSCRIBE' button.

Figure: <https://www.programmableweb.com/apis/directory>

► Post-API age?

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?