Digital Trace Data: Web Scraping

Summer Institute in Computational Social Science @ CU Boulder
August 13th, 2018

Overview

- What is digital trace data?
- What are the strengths and weaknesses of digital trace data?
- Practical tips and tools for web scraping in Python

“[J]ust as the invention of the telescope revolutionized the study of the heavens, so too by rendering the unmeasurable measurable, the technological revolution in mobile, Web, and Internet communications has the potential to revolutionize our understanding of ourselves and how we interact … we have finally found our telescope. Let the revolution begin…”

— Duncan Watts (*Everything is Obvious* 2012, p. 266)

https://cbail.github.io/SICSS_What_is_Digital_Trace.html

Digital Trace Data

Examples:

- Social media sites
- Web search data
- Blogs / internet forums
- Administrative data on websites
- Internet archive
- Digitization of historical texts / archives
- Audio-visual data

https://cbail.github.io/SICSS_What_is_Digital_Trace.html

Digital Trace Data

Digital trace data is a usually *readymade* not *custommade*.

“Big data sources tend to have a number of characteristics in common; some are generally good for social research and some are generally bad.” (Salganik, 2017)

Readymade

Custommade

Figure 1.2: *Fountain* by Marcel Duchamp and *David* by Michaelangelo. *Fountain* is an example of a readymade, where an artist sees something that already exists in the world then creatively repurposes it for art. *David* is an example of art that was intentionally created; it is a custommade. Social research in the digital age will involve both readymades and custommades. Photograph of *Fountain* by Alfred Stiglitz, 1917 (Source: *The Blind Man*, no. 2/Wikimedia Commons). Photograph of *David* by Jörg Bittner Unna, 2008 (Source: _Galleria dell’Accademia, Florence/Wikimedia Commons).

https://www.bitbybitbook.com/en/1st-ed/observing-behavior/
https://cbail.github.io/SICSS_What_is_Digital_Trace.html

Common Characteristics of Big Data

It's big, …

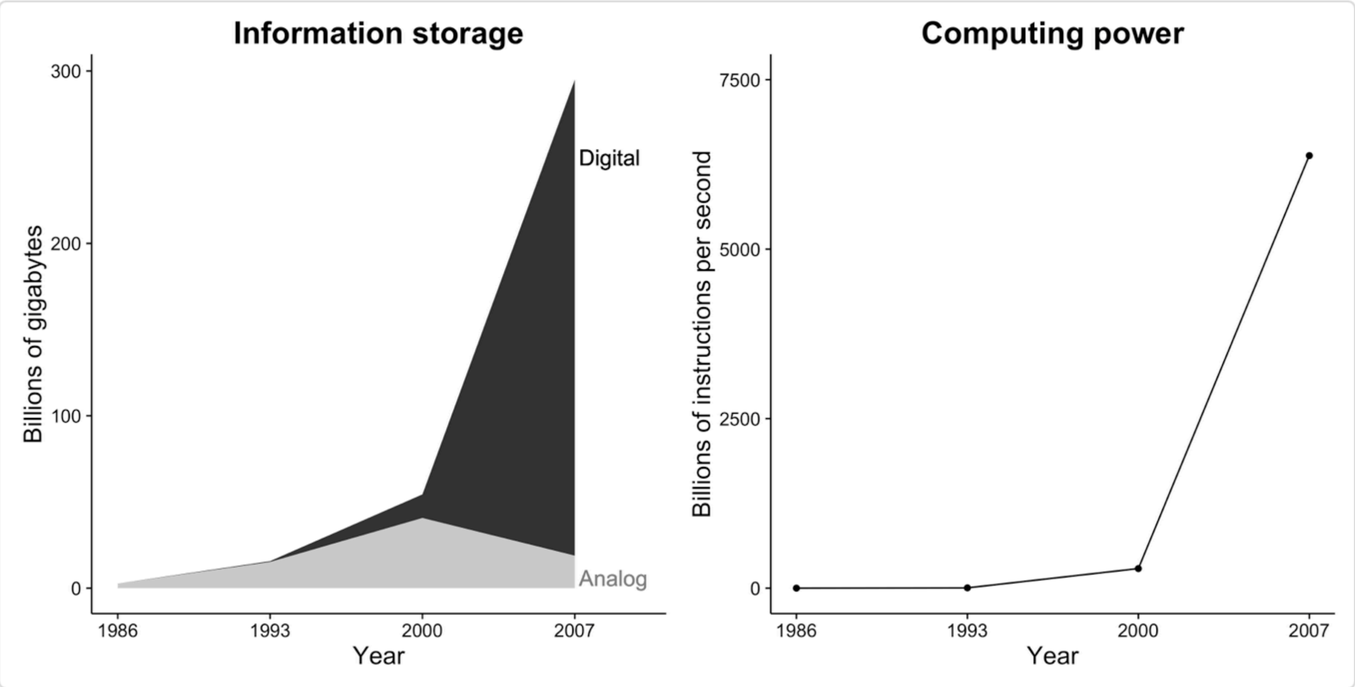


Figure 1.1: Information storage capacity and computing power are increasing dramatically. Further, information storage is now almost exclusively digital. These changes create incredible opportunities for social researchers. Adapted from Hilbert and López (2011), figures 2 and 5.

https://www.bitbybitbook.com/en/1st-ed/introduction/digital-age/
https://cbail.github.io/SICSS_strengths_weaknesses.html

Common Characteristics of Big Data

always on, …

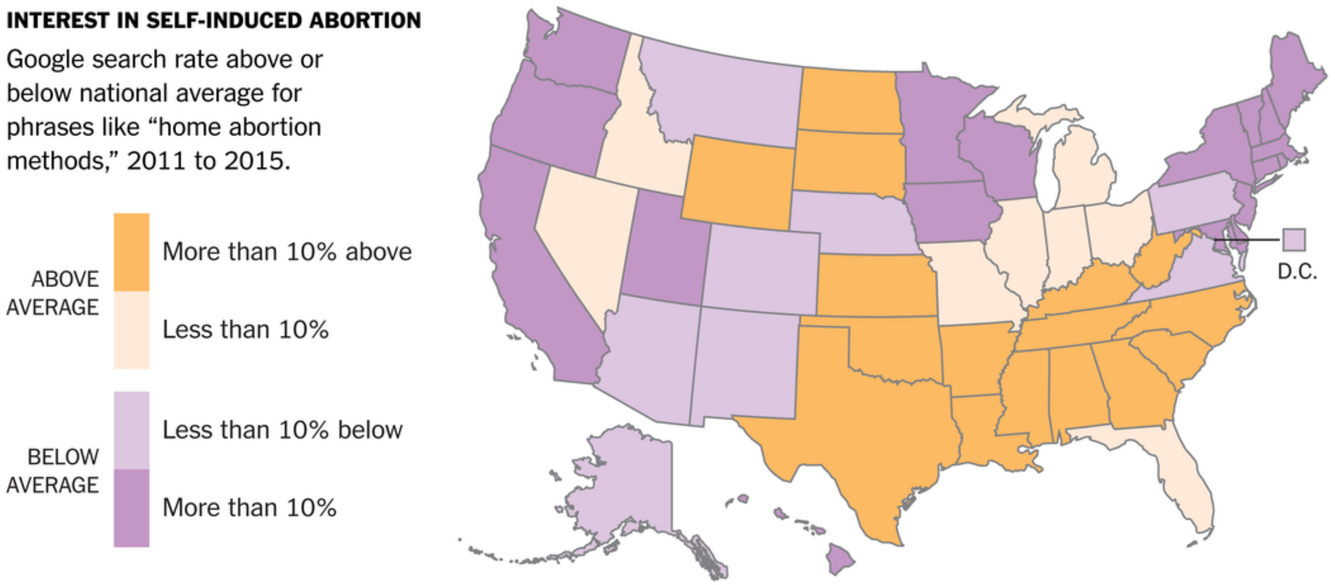
Table 2.1: Studies of unexpected events using always-on big data sources.

Unexpected event	Always-on data source	Citation
Occupy Gezi movement in Turkey	Twitter	Budak and Watts (2015)
Umbrella protests in Hong Kong	Weibo	Zhang (2016)
Shootings of police in New York City	Stop-and-frisk reports	Legewie (2016)
Person joining ISIS	Twitter	Magdy, Darwish, and Weber (2016)
September 11, 2001 attack	livejournal.com	Cohn, Mehl, and Pennebaker (2004)
September 11, 2001 attack	pager messages	Back, Küfner, and Egloff (2010), Pury (2011), Back, Küfner, and Egloff (2011)

https://www.bitbybitbook.com/en/1st-ed/observing-behavior/characteristics/always-on/
http://www.techreviewer.co.uk/we-use-facebook-to-schedule-the-protests-twitter-to-coordinate-and-youtube-to-tell-the-world/
https://cbail.github.io/SICSS_strengths_weaknesses.html

Common Characteristics of Big Data

and nonreactive.



https://www.nytimes.com/2016/03/06/opinion/sunday/the-return-of-the-diy-abortion.html
https://cbail.github.io/SICSS_strengths_weaknesses.html

Common Characteristics of Big Data

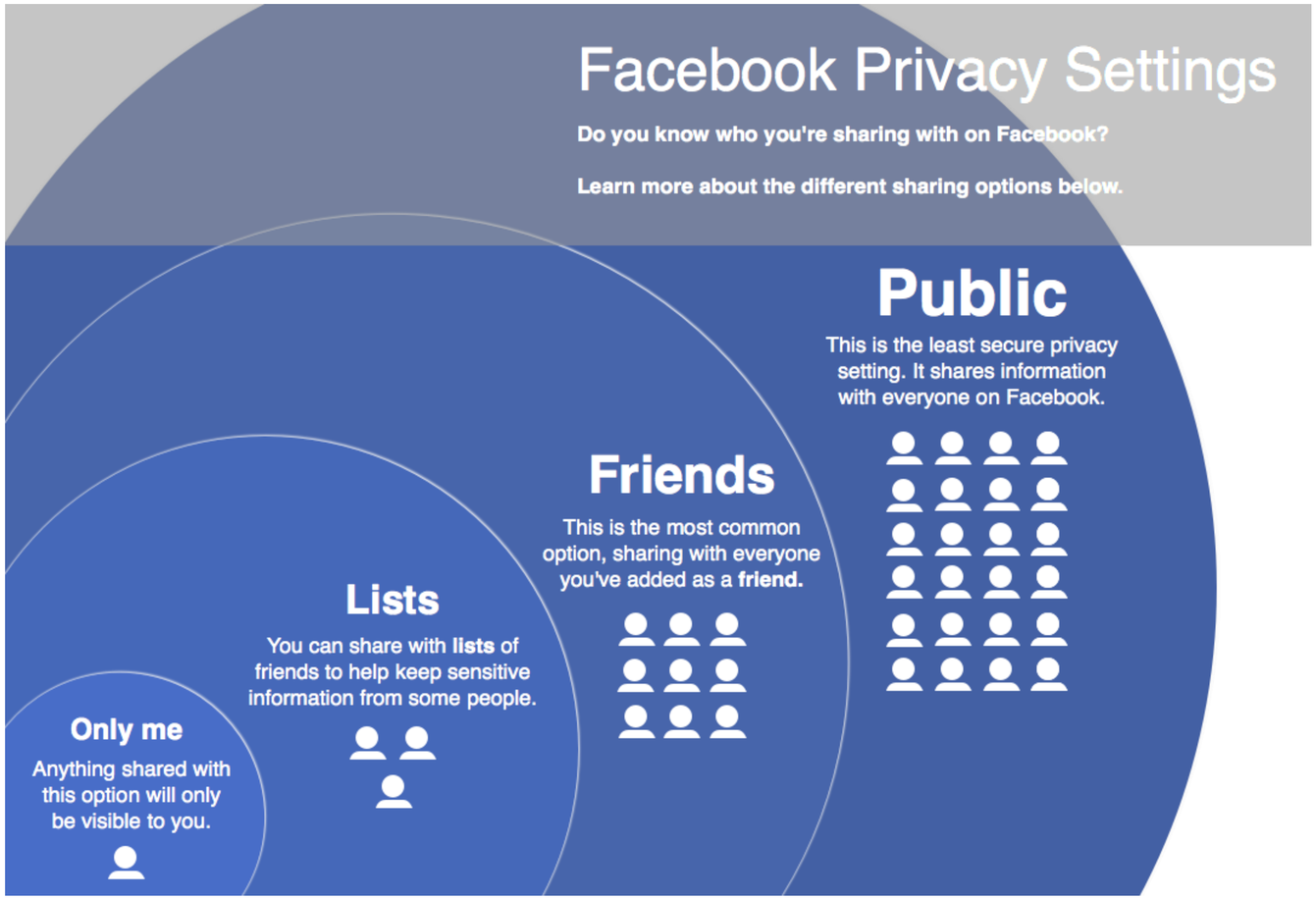
But it's also incomplete, …

"No matter how big your big data, it probably doesn't have the information you want." (Salganik 2017)

https://cbail.github.io/SICSS_strengths_weaknesses.html

Common Characteristics of Big Data

inaccessible, …



https://cbail.github.io/SICSS_strengths_weaknesses.html

Common Characteristics of Big Data

non-representative, …

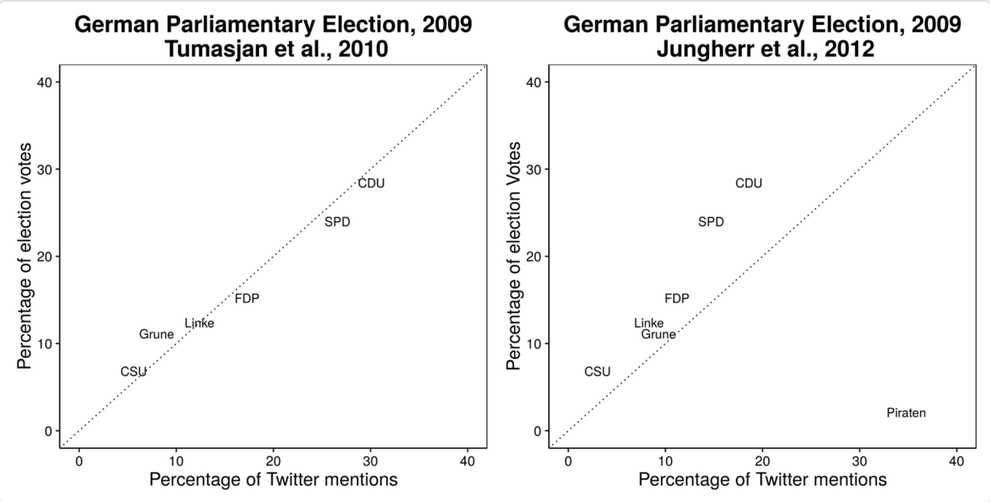
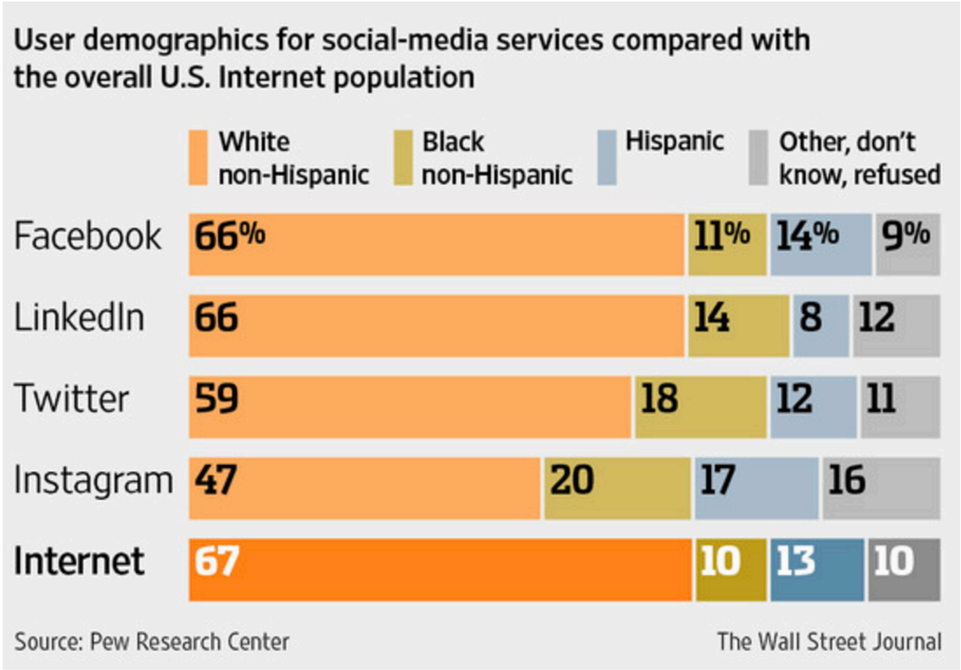


Figure 2.3: Twitter mentions appear to predict the results of the 2009 German election (Tumasjan et al. 2010), but this excludes the party with the most mentions: Pirate Party (Jungherr, Jürgens, and Schoen 2012). See Tumasjan et al. (2012) for an argument in favor of excluding the Pirate Party. Adapted from Tumasjan et al. (2010), table 4 and Jungherr, Jürgens, and Schoen (2012), table 2.



http://www.pewinternet.org
https://www.bitbybitbook.com/en/1st-ed/observing-behavior/characteristics/non-representative/
https://cbail.github.io/SICSS_strengths_weaknesses.html

Common Characteristics of Big Data

drifting, …

List of defunct social networking websites

From Wikipedia, the free encyclopedia

This is a **list of defunct social networking websites**.

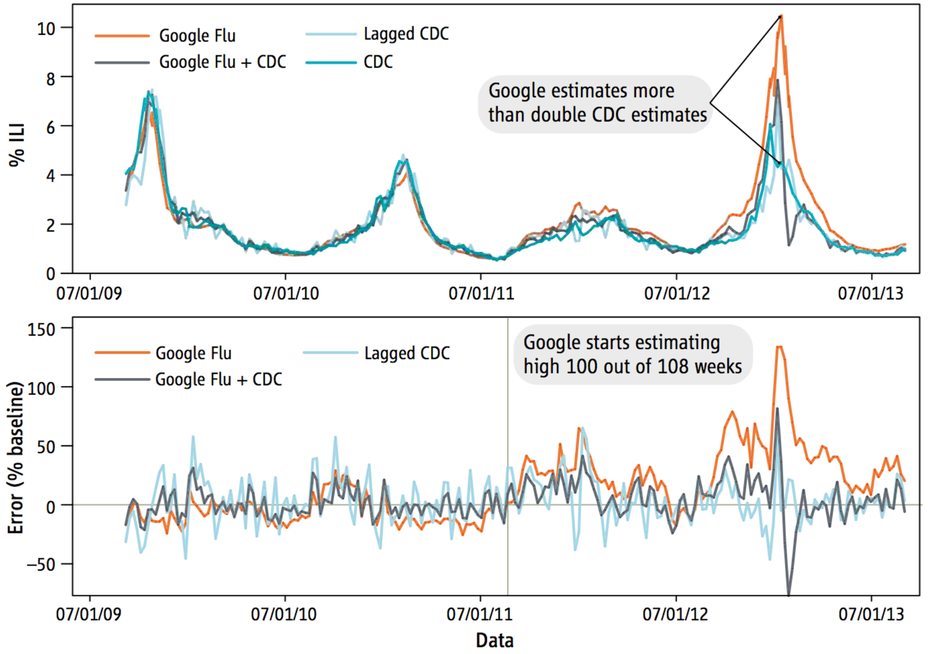
Name	Description/Focus	Registered users	Registration
43 Things	Goal setting and achievement	12,914[1]	Closed on January 1, 2015
App.net	Microblogging		
Avatars United	Online games	22,686,225[2]	Open
Beeso.fr	French black community		Open
Bolt	Teen community website		Open
Boomj.com	Those born 1956-1965 (baby boomers)		Merged with another site
Boredat	Anonymous social network for U.S. college students		Shut down December 31, 2016
Capazoo		230,000	Open
Classical Lounge	Classical music lovers		Sold to investors 2009, closed 2011
eConozco	Bought by Xing		Open
FitFinder	Anonymous UK student microblogging website.	322,113[3]	Open
Formspring	Social Q&A website	290,000,000[4]	Open
FriendFeed	Feed aggregator	1,683[5]	Closed on October 2, 2007
Friends Reunited	Social networking based upon the theme of reunion	1,210 (UK)	Closed on February 26, 2016
Friendster	Social networking website	279,801 (August 2017)[6]	Closed on June 14, 2015
Google Buzz	Social networking, microblogging and messaging tool that was developed by Google		Discontinued December 15, 2011 and replaced by Google+

https://en.wikipedia.org/wiki/List_of_defunct_social_networking_websites
https://cbail.github.io/SICSS_strengths_weaknesses.html

Common Characteristics of Big Data

algorithmically confounded, …

"Behavior in big data systems is not natural; it is driven by the engineering goals of the systems." (Salganik 2017)

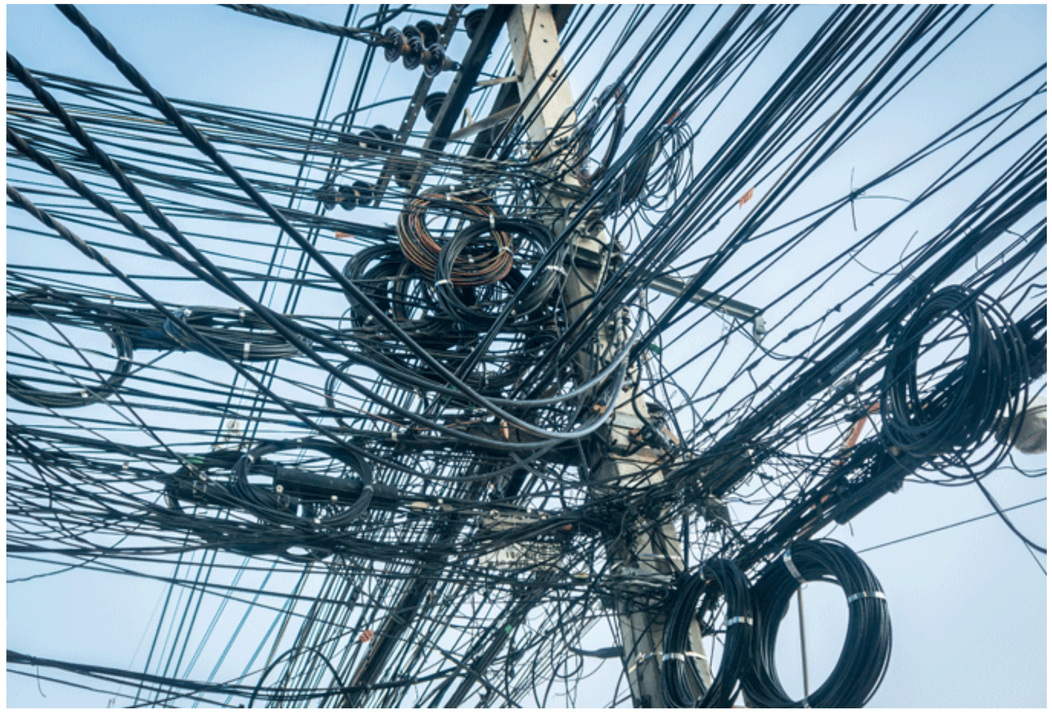


GFT overestimation. GFT overestimated the prevalence of flu in the 2012–2013 season and overshot the actual level in 2011–2012 by more than 50%. From 21 August 2011 to 1 September 2013, GFT reported overly high flu prevalence 100 out of 108 weeks. (**Top**) Estimates of doctor visits for ILI. "Lagged CDC" incorporates 52-week seasonality variables with lagged CDC data. "Google Flu + CDC" combines GFT, lagged CDC estimates, lagged error of GFT estimates, and 52-week seasonality variables. (**Bottom**) Error [as a percentage {[Non-CDC estmate)−(CDC estimate)]/(CDC) estimate)}. Both alternative models have much less error than GFT alone. Mean absolute error (MAE) during the out-of-sample period is 0.486 for GFT, 0.311 for lagged CDC, and 0.232 for combined GFT and CDC. All of these differences are statistically significant at $P < 0.05$. See SM.

http://science.sciencemag.org/content/343/6176/1203
https://cbail.github.io/SICSS_strengths_weaknesses.html

Common Characteristics of Big Data

dirty, …



https://cbail.github.io/SICSS_strengths_weaknesses.html

Common Characteristics of Big Data

and sensitive.

TECH • OKCUPID

Researchers Caused an Uproar By Publishing Data From 70,000 OkCupid Users

TECH • ASHLEY MADISON

What to know about the Ashley Madison hack

Robust De-anonymization of Large Sparse Datasets

Arvind Narayanan and Vitaly Shmatikov

The University of Texas at Austin

"… We apply our de-anonymization methodology to the Netflix Prize dataset, which contains anonymous movie ratings of the 500,000 subscribers of Netflix … We demonstrate that an adversary who knows only a little bit about an individual subscriber can easily identify this subscriber's record in the dataset." (Narayanan & Shmatikov 2008)

http://fortune.com/2016/05/18/okcupid-data-research/; http://fortune.com/2015/08/26/ashley-madison-hack/
https://ieeexplore.ieee.org/document/4531148/
https://cbail.github.io/SICSS_strengths_weaknesses.html

Collecting Digital Trace Data

Popular methods include:

- Web scraping (“screen scraping”) [Now]
- Application Programming Interfaces (APIs) [Later]
- Also, downloading datasets directly from the platform