Digital Trace Data

Bamberg Summer Institute in Computational Social Science

Carsten Schwemmer, University of Bamberg

2019-07-30

Many thanks to Chris Bail for providing material for this lecture

What is digital trace data?

What is digital trace data?

"[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 ... . [T]hree hundred years after Alexander Pope argued that the proper study of mankind should lie not in the heavens but in ourselves, we have finally found our telescope. Let the revolution begin."

Duncan Watts (2011, p. 266)

What is digital trace data?

- social media sites
- web search data
- blogs / internet forums
- administrative data on websites
- internet archives
- digitization of historical texts/archives
- audio-visual data

What is digital trace data?

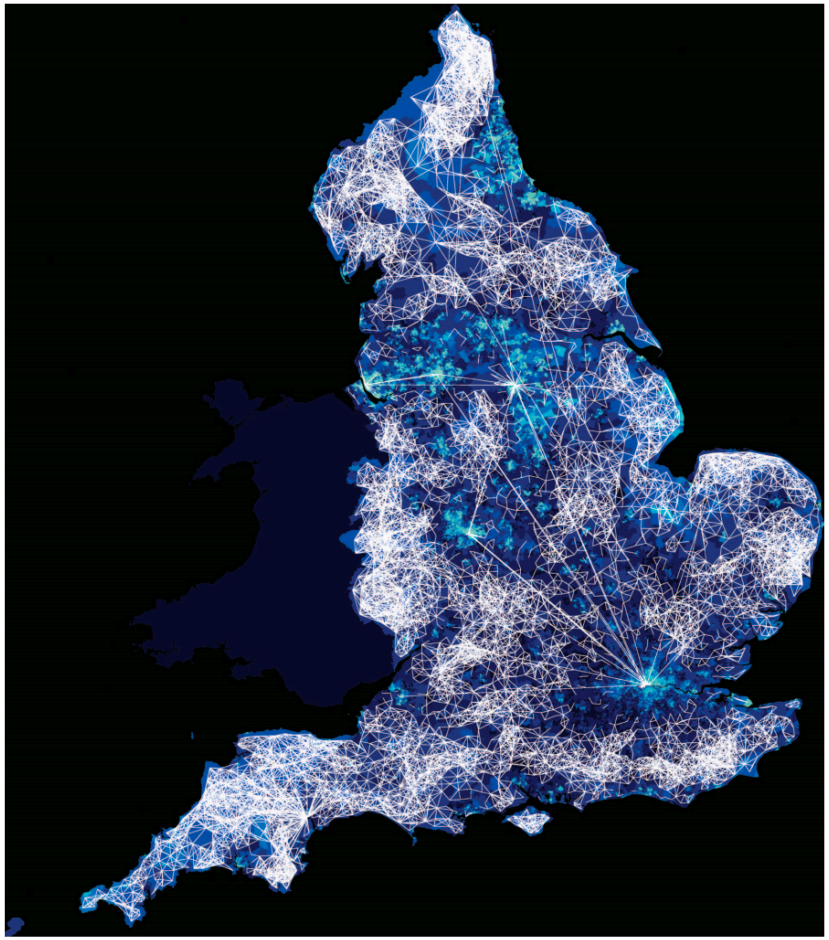


Fig. 1. An image of regional communication diversity and socioeconomic ranking for the UK. We find that communities with diverse communication patterns tend to rank higher (represented from light blue to dark blue) than the regions with more insular communication. This result implies that communication diversity is a key indicator of an economically healthy community. [(*29*) Crown copyright material is reproduced with the permission of the Controller of Her Majesty's Stationery Office]

https://science.sciencemag.org/content/328/5981/1029

What is digital trace data?

https://www.facebook.com/notes/facebook-engineering/visualizing-friendships/469716398919

What is digital trace data?

Strengths of digital trace data

Big

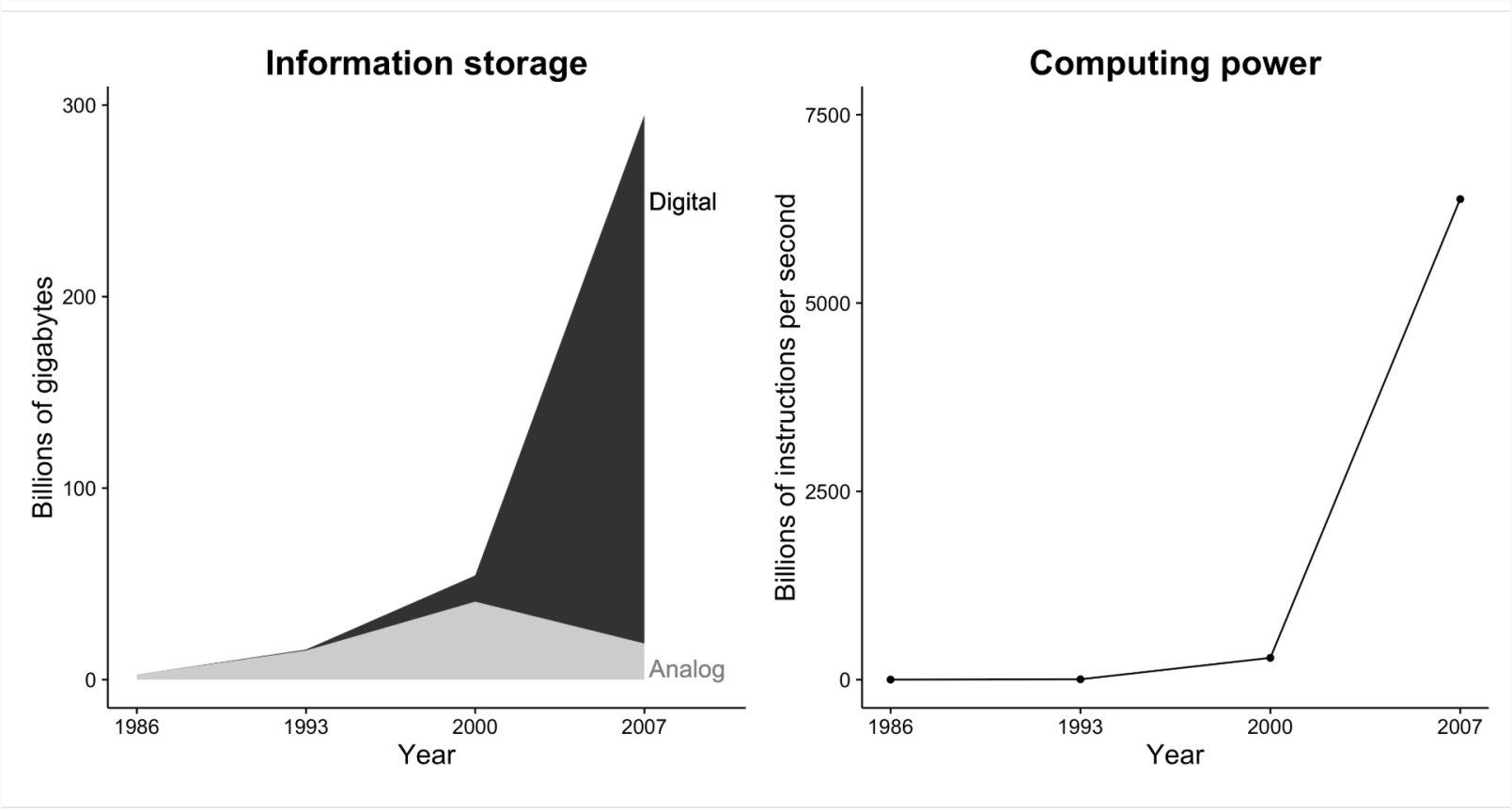


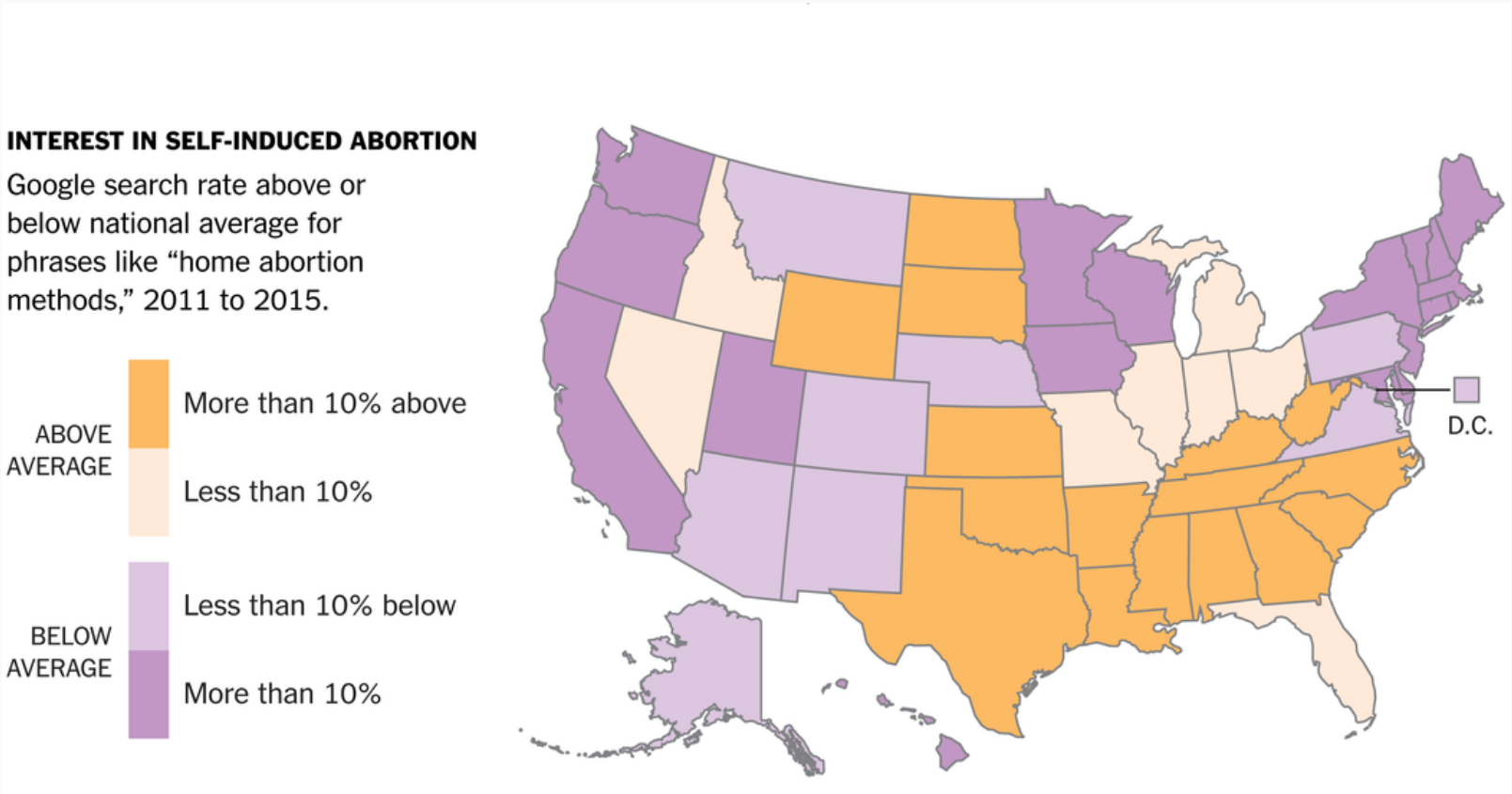
Figure 1.1: Information storage capacity and computing power are increasing dramatically. Further, information storage is now almost exclusively digital. These changes create incredible opportunities for social researchers. Adapted from Hilbert and López (2011), figures 2 and 5.

https://www.bitbybitbook.com/en/1st-ed/introduction/digital-age/

Always on

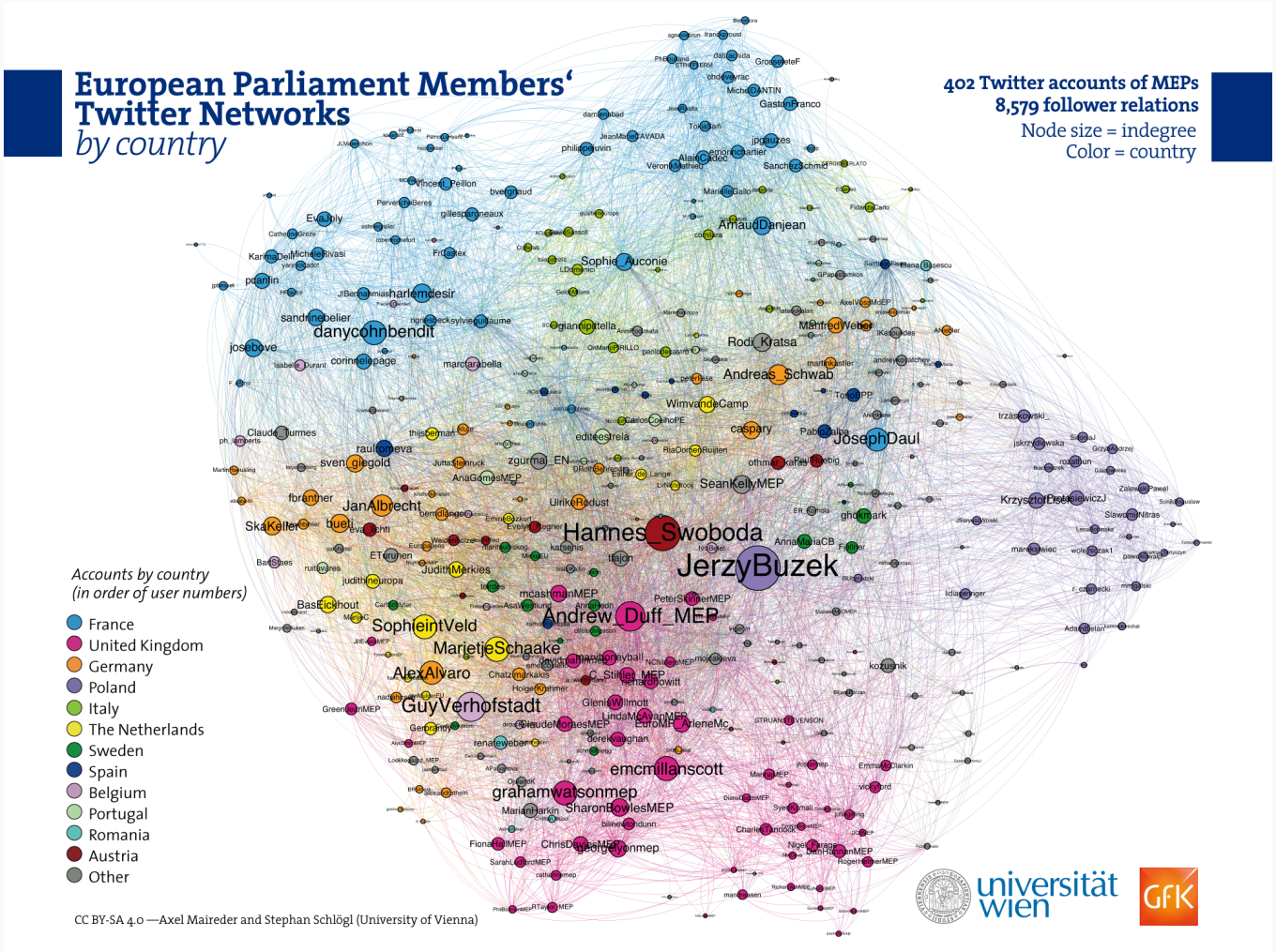
https://en.wikipedia.org/wiki/Egyptian_revolution_of_2011

Non-reactive



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

Captures social relationships



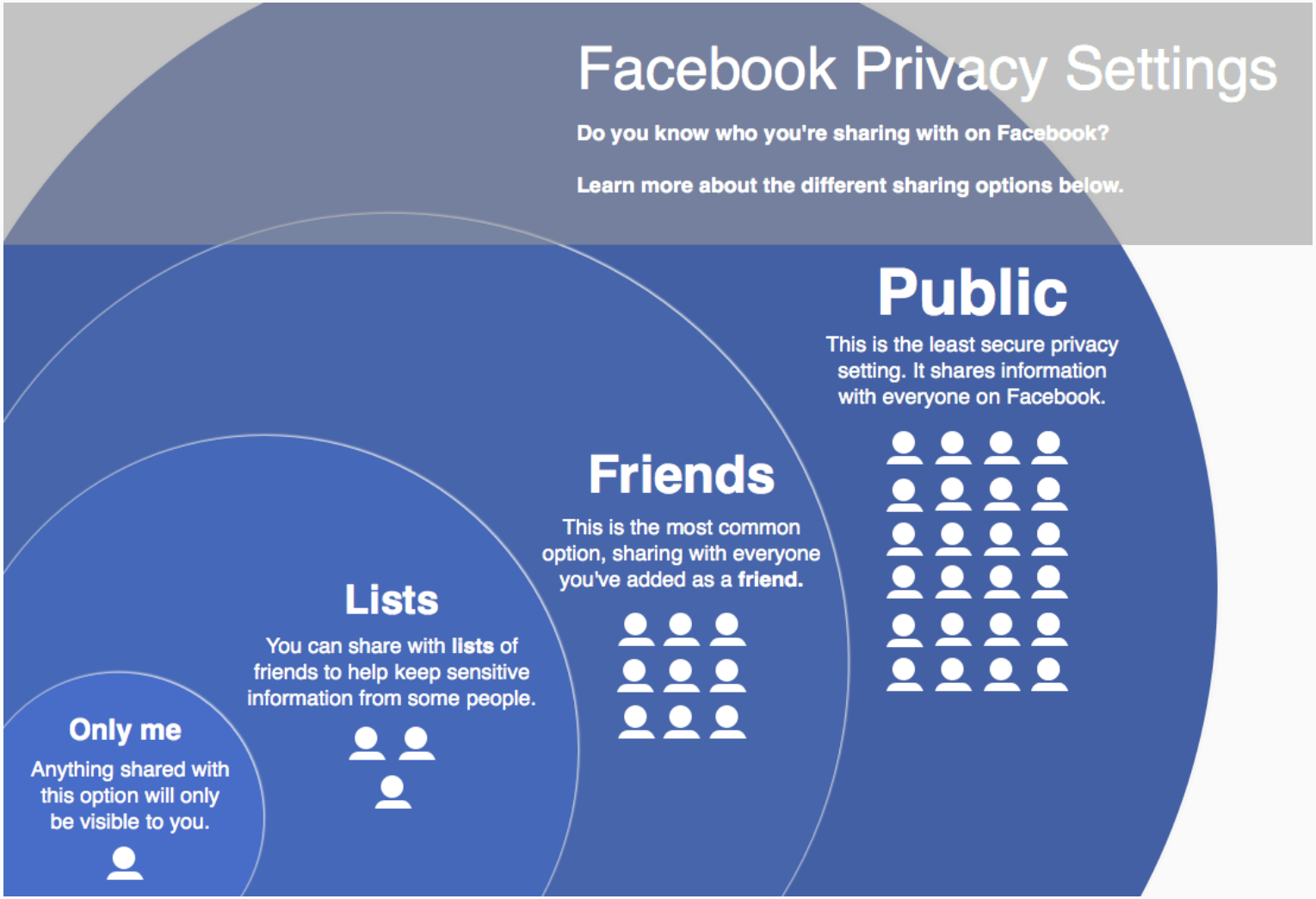
https://homepage.univie.ac.at/axel.maireder/php/wordpress/wp-content/MEPnetwork_country.png

Weaknesses of digital trace data

Incomplete


Like · Comment · Share
7 people like this.
This comment has been removed. You can Undo this, Report it as abusive, or Ban [redacted].
Write a comment...

Inaccessible



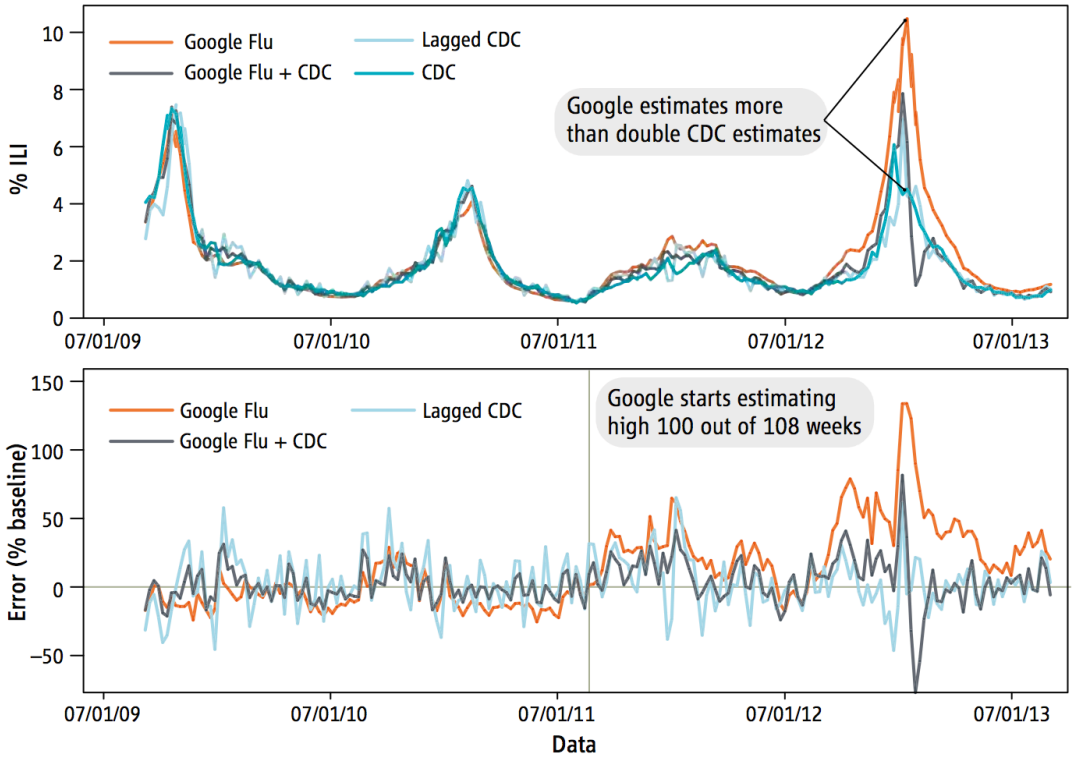
Non-representative

		Age				Gender		Education		
		14-29	30-49	50-64	65+	Men	Women	Low	Mid	High
	WhatsApp	79	71	58	18	56	54	34	59	69
	Facebook	66	52	34	10	39	39	23	41	52
	YouTube	64	50	31	6	41	32	21	36	46
	Instagram	37	15	5	1	13	14	3	9	21
	Twitter	18	9	5	1	10	5	3	6	11
	XING	4	12	5	1	9	3	0,2	4	14
	Snapchat	22	5	1	0	6	6	2	3	7
	LinkedIn	5	9	5	1	8	2	0,3	3	12
	Google+	4	5	3	1	5	1	2	3	5
	Other	10	6	2	1	5	4	1	3	7
	None	3	10	20	29	18	15	19	19	14

https://initiatived21.de/pm-sonderstudie-nrw/

Drifting

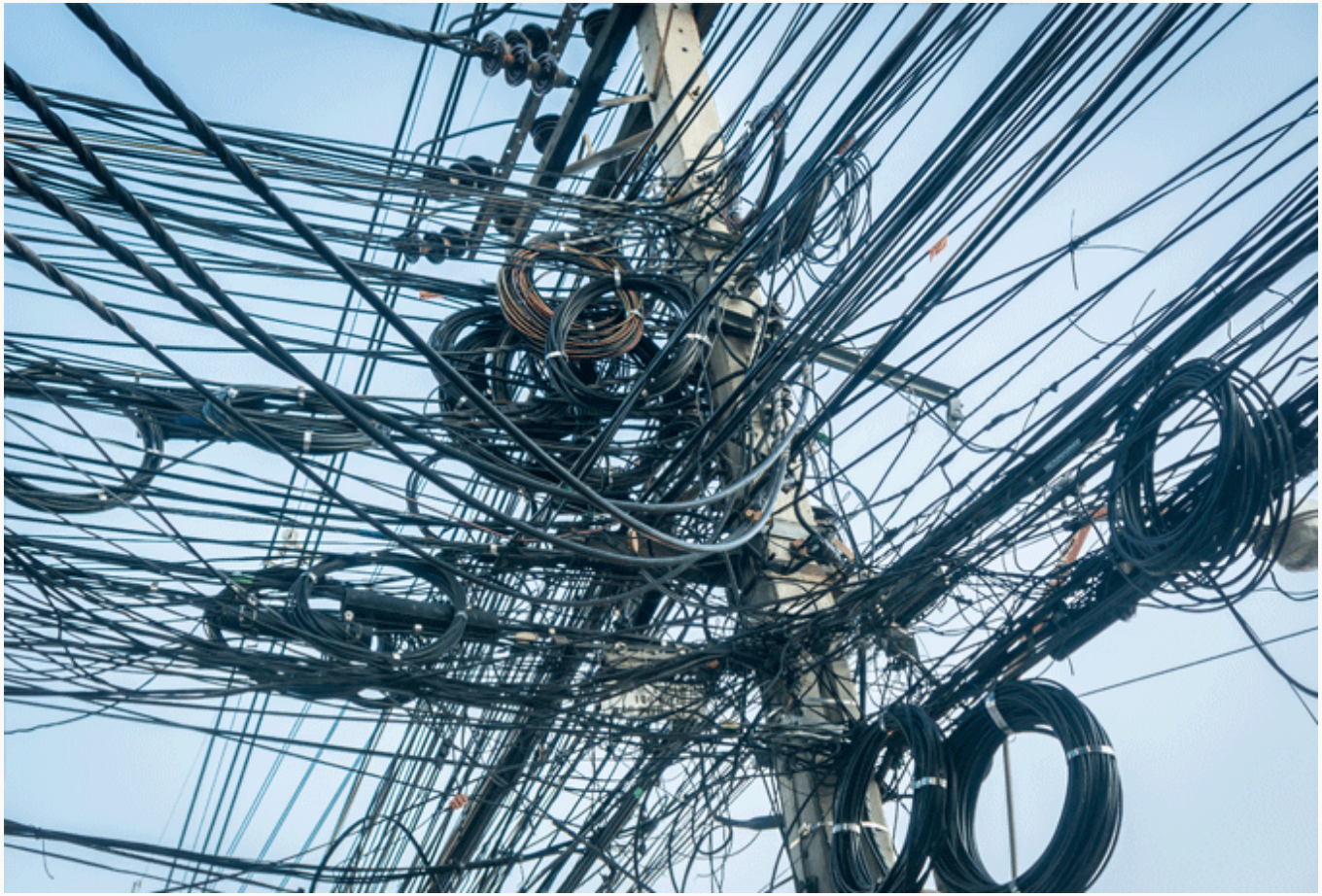
Algorithmically confounded



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.

https://dx.doi.org/10.1126/science.1248506

Unstructured



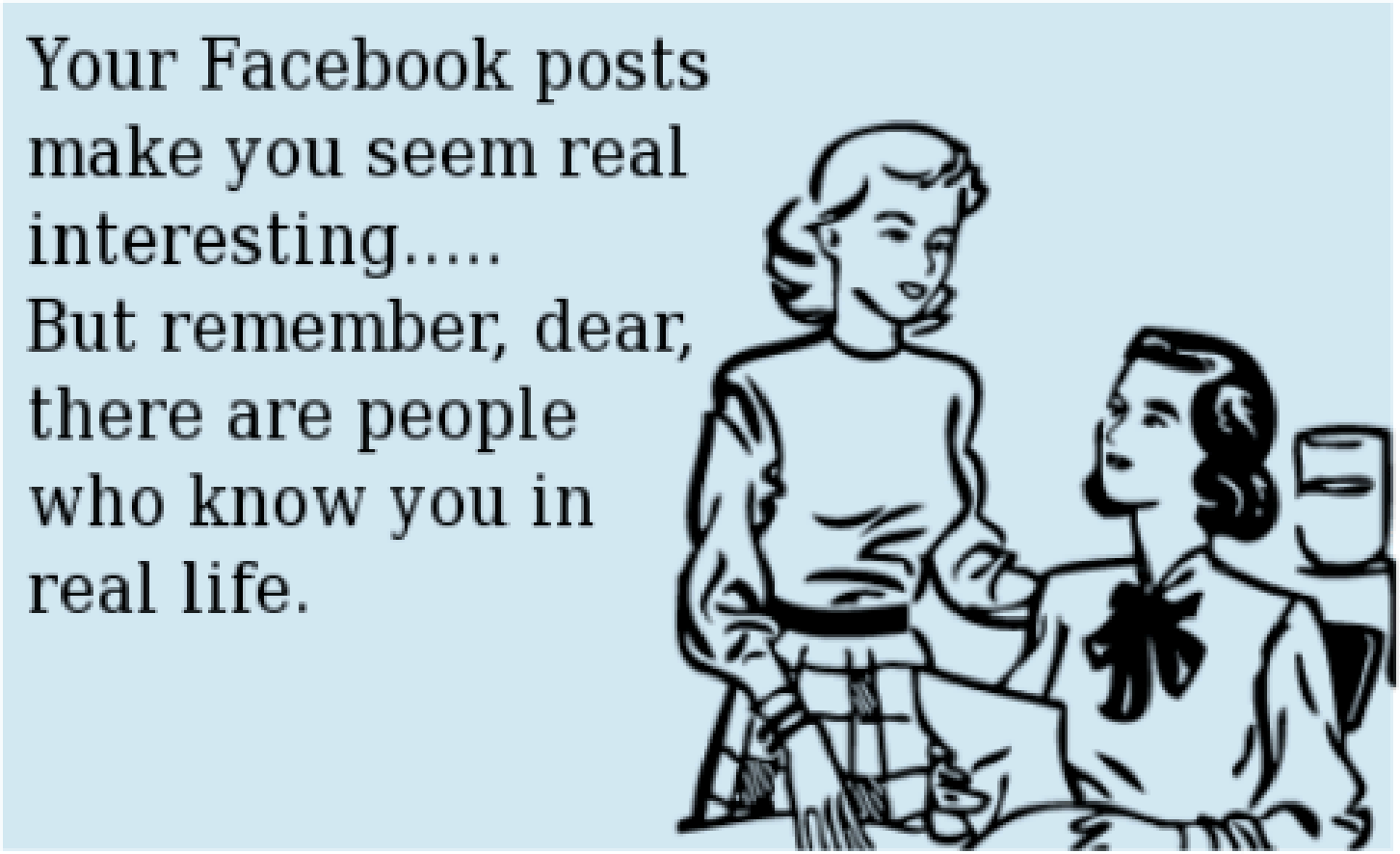
OKCUPID

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

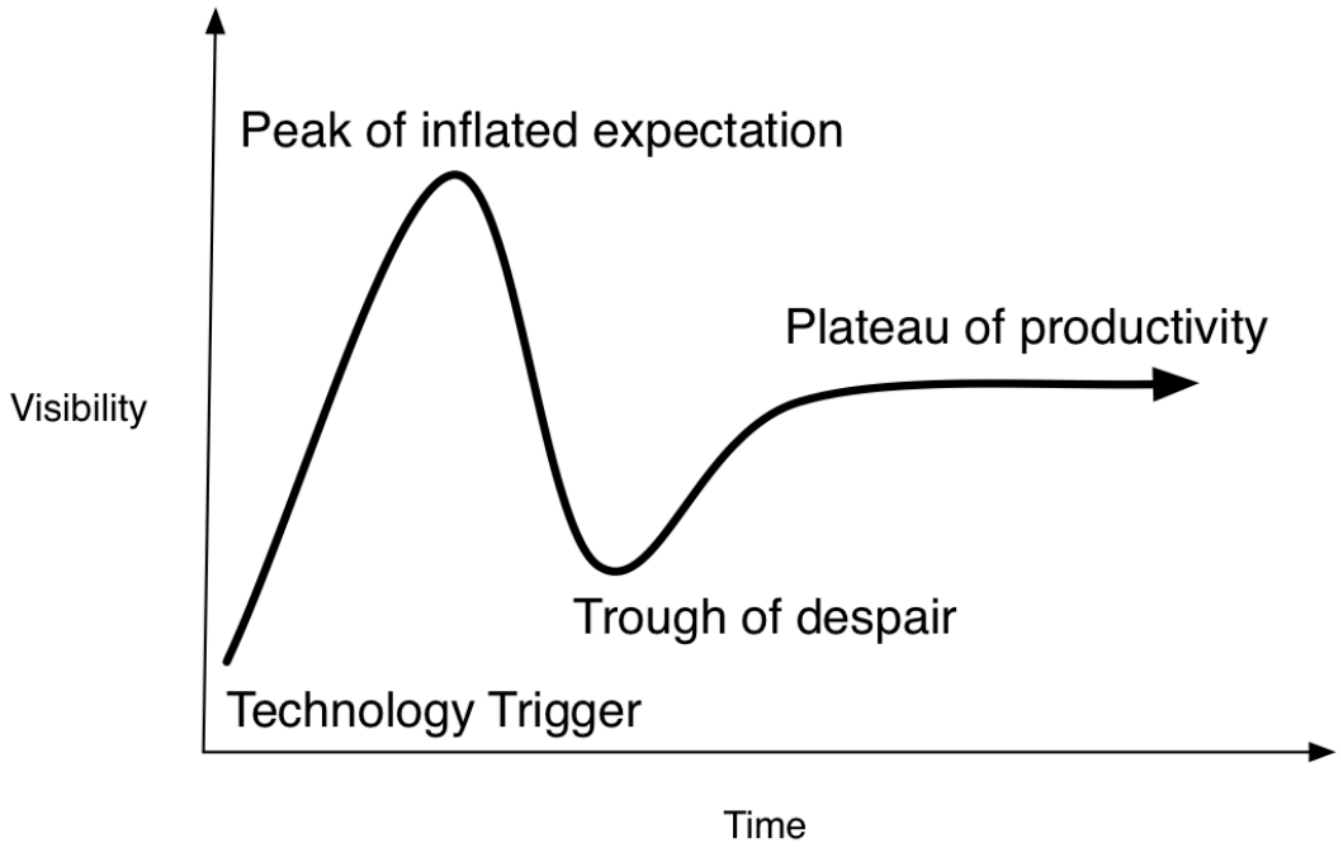
Robert Hackett
May 18, 2016

https://fortune.com/2016/05/18/okcupid-data-research/

Positivity bias



The future of digital trace data



https://commons.wikimedia.org/wiki/File:Gartner_Hype_Cycle.svg

Questions?