Introduction

Indira Sen

University of Konstanz
Measurement and Representation Biases (MRB) in Digital Trace Data-based Studies

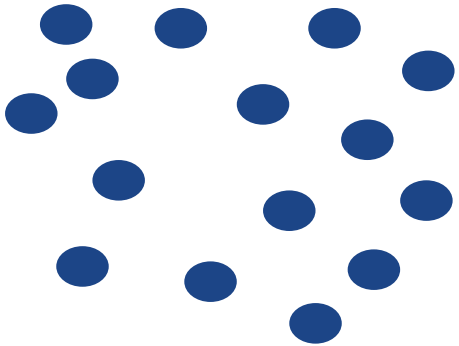
https://github.com/Indiiigo/css_research_projects/tree/main

Agenda

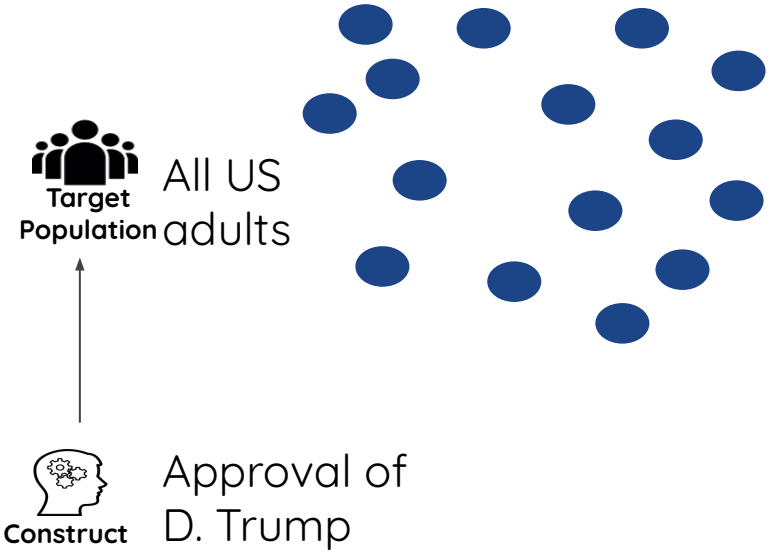
- Background in Research with Digital Traces
- Course logistics

A typical research pipeline with digital trace data

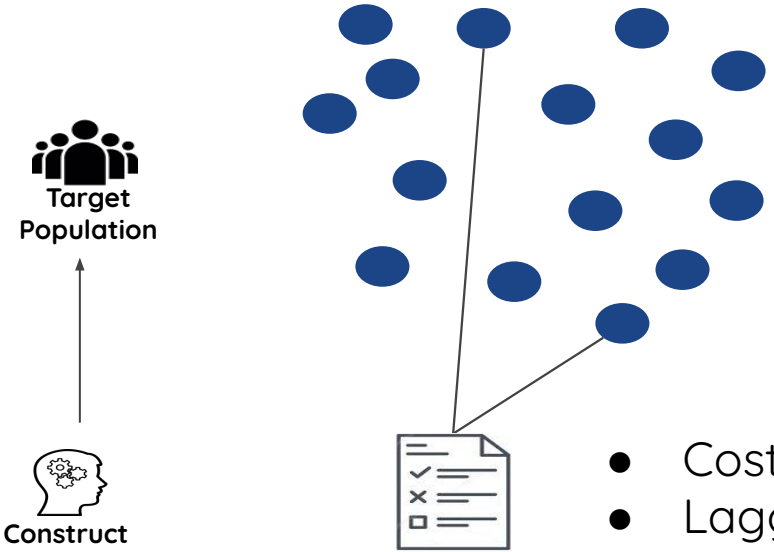
A typical research design with digital traces



A typical research design with digital traces

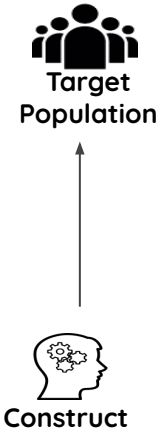


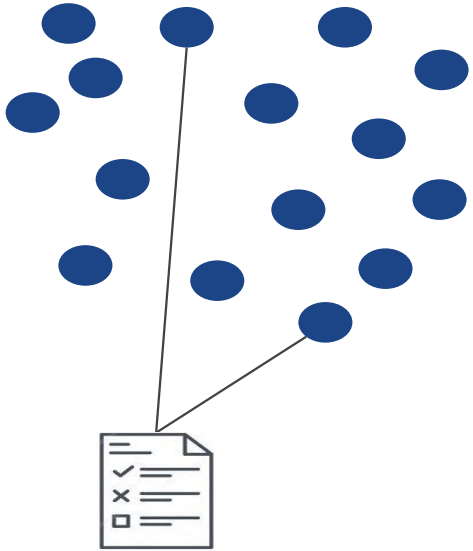
A typical research design with digital traces

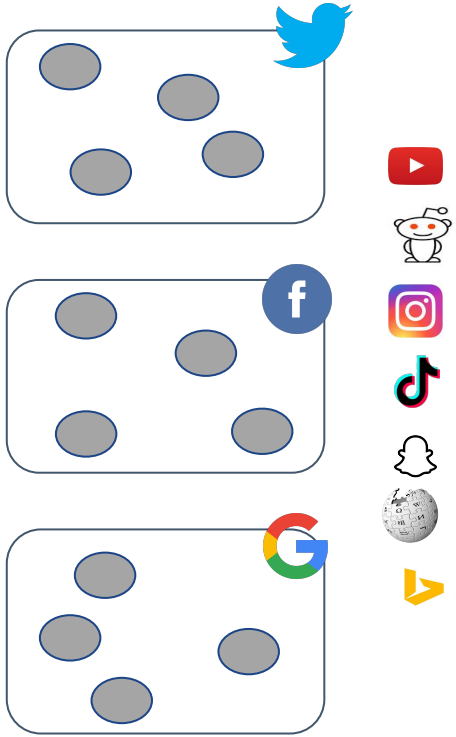


- Costly
- Lagging (after important events)
- Small samples (max. in the thousands)

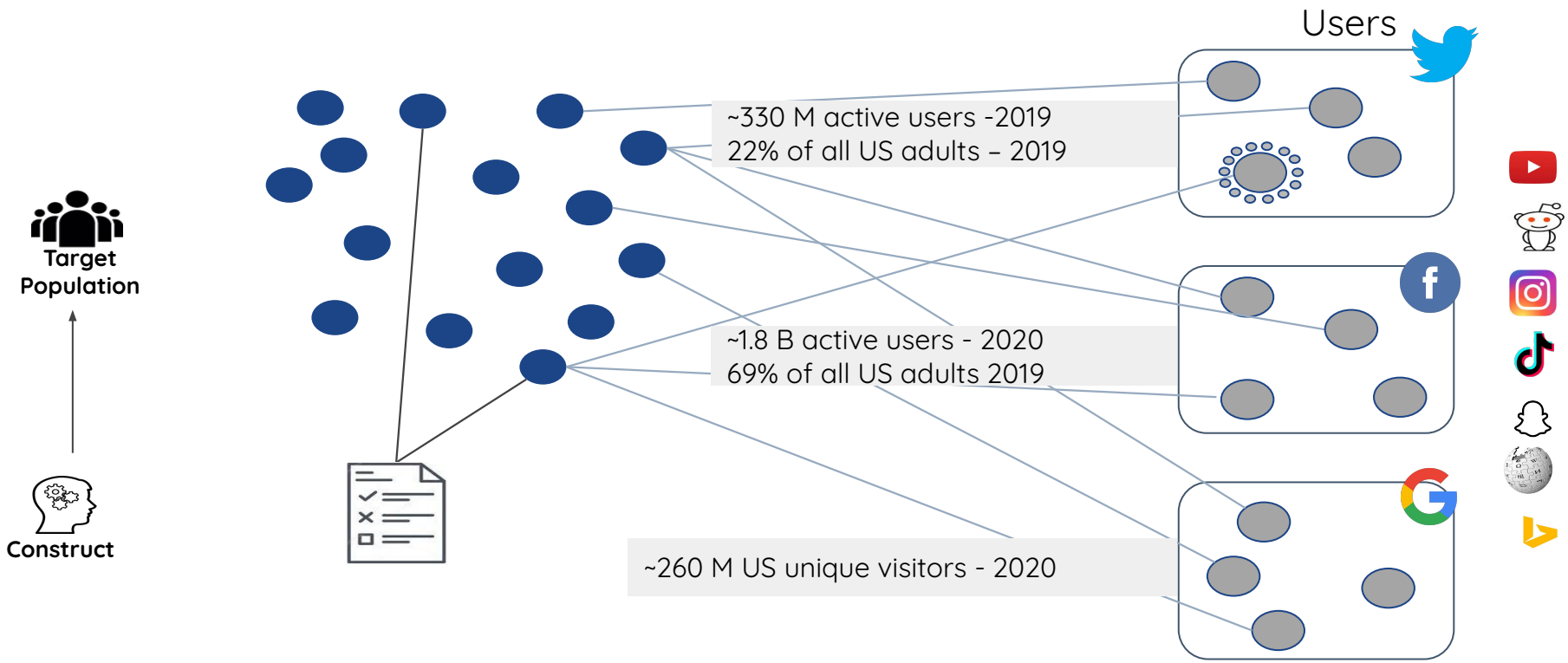
A typical research design with digital traces



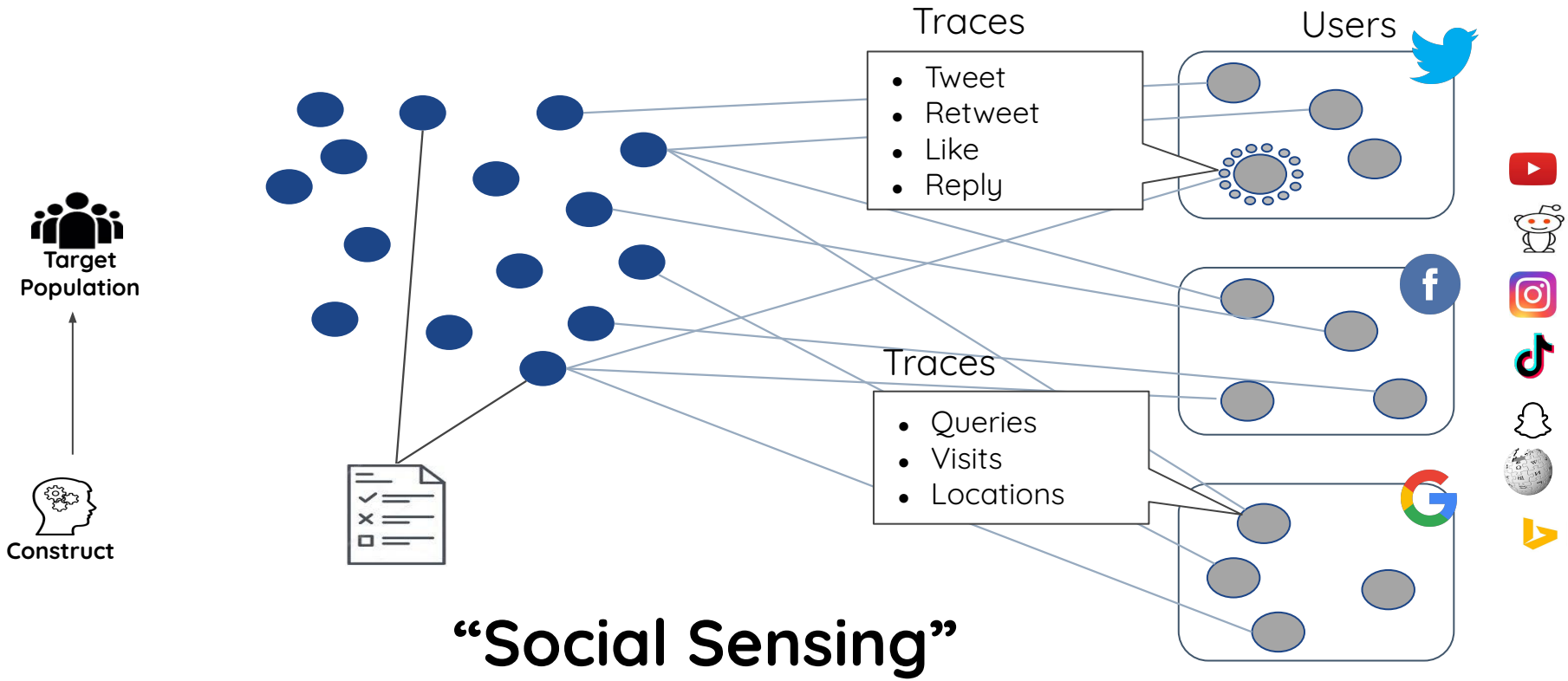




A typical research design with digital traces

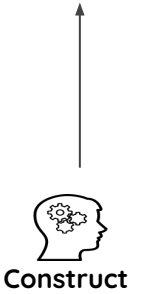


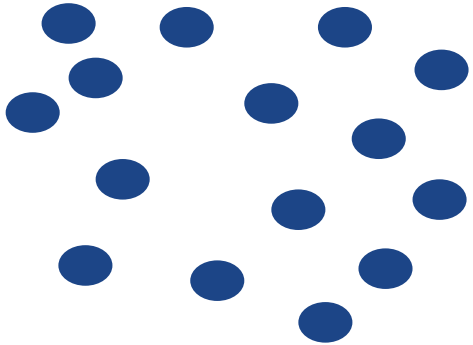
A typical research design with digital traces

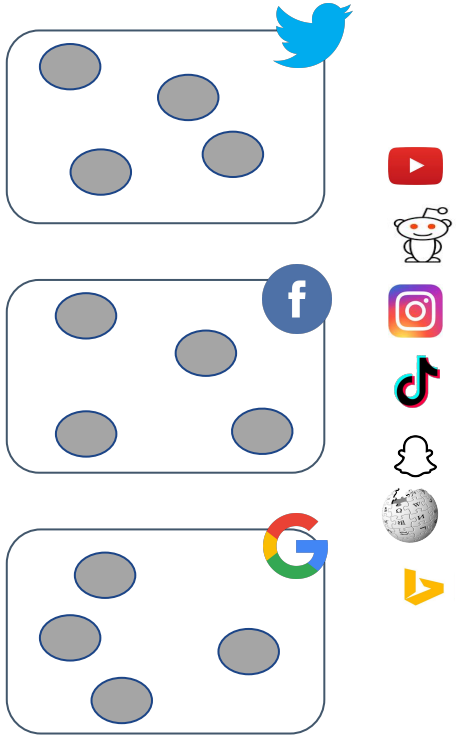


"Social Sensing"

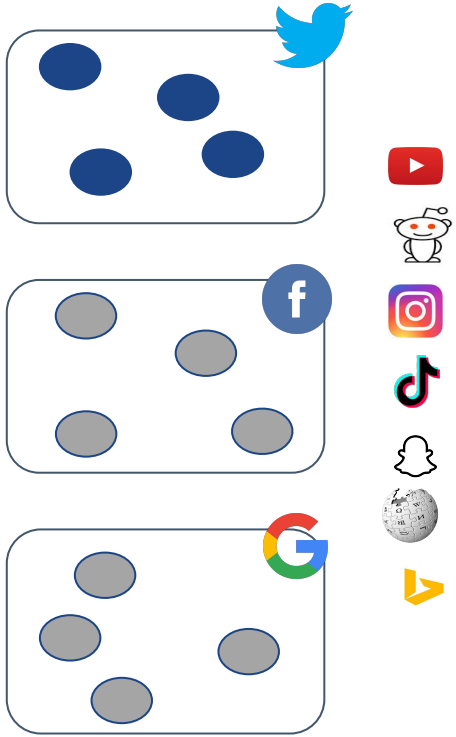
A typical research design with digital traces





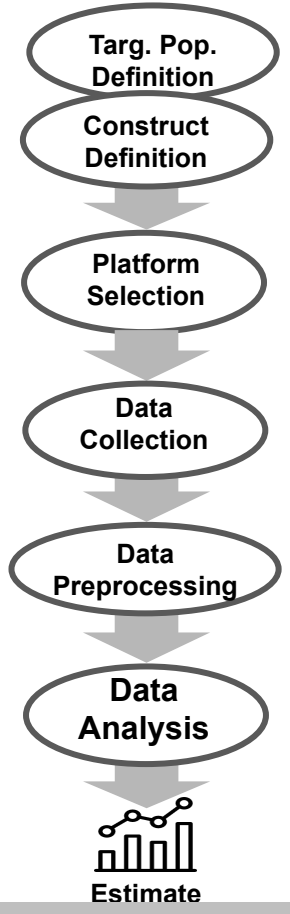


A typical research design with digital traces



“Platform Study”

A typical research pipeline with digital traces



"TED-On: A Total Error Framework for Digital Traces of Human Behavior on Online Platforms" Sen et al., 2021, Public Opinion Quarterly and https://arxiv.org/pdf/1907.08228

Use case: Detecting the flu with digital traces

Google Flu (trends)

nature

Vol 457|19 February 2009|doi:10.1038/nature07634

LETTERS

Detecting influenza epidemics using search engine query data

Jeremy Ginsberg[1], Matthew H. Mohebbi[1], Rajan S. Patel[1], Lynnette Brammer[2], Mark S. Smolinski[1] & Larry Brilliant[1]

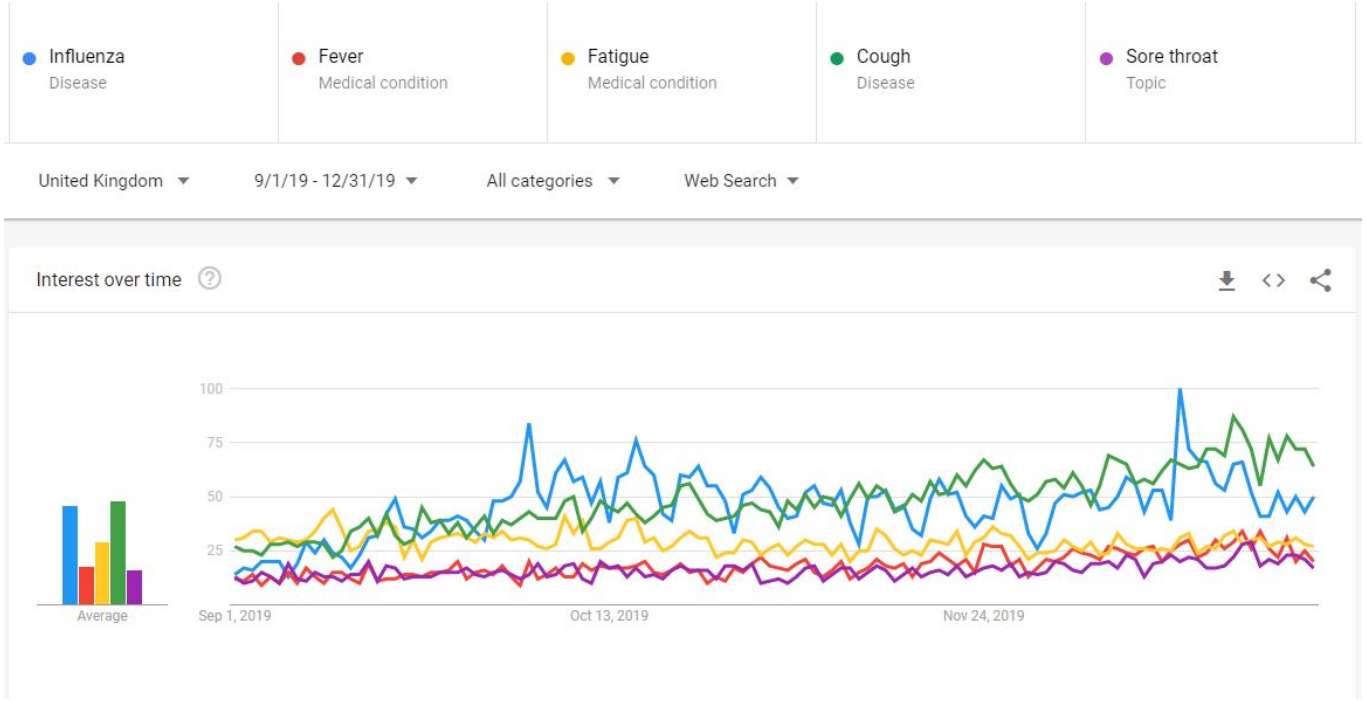
What proportion of US-Americans have the flu?

What is the approval rating of A. Merkel?

Are anti-immigration sentiments on the rise?

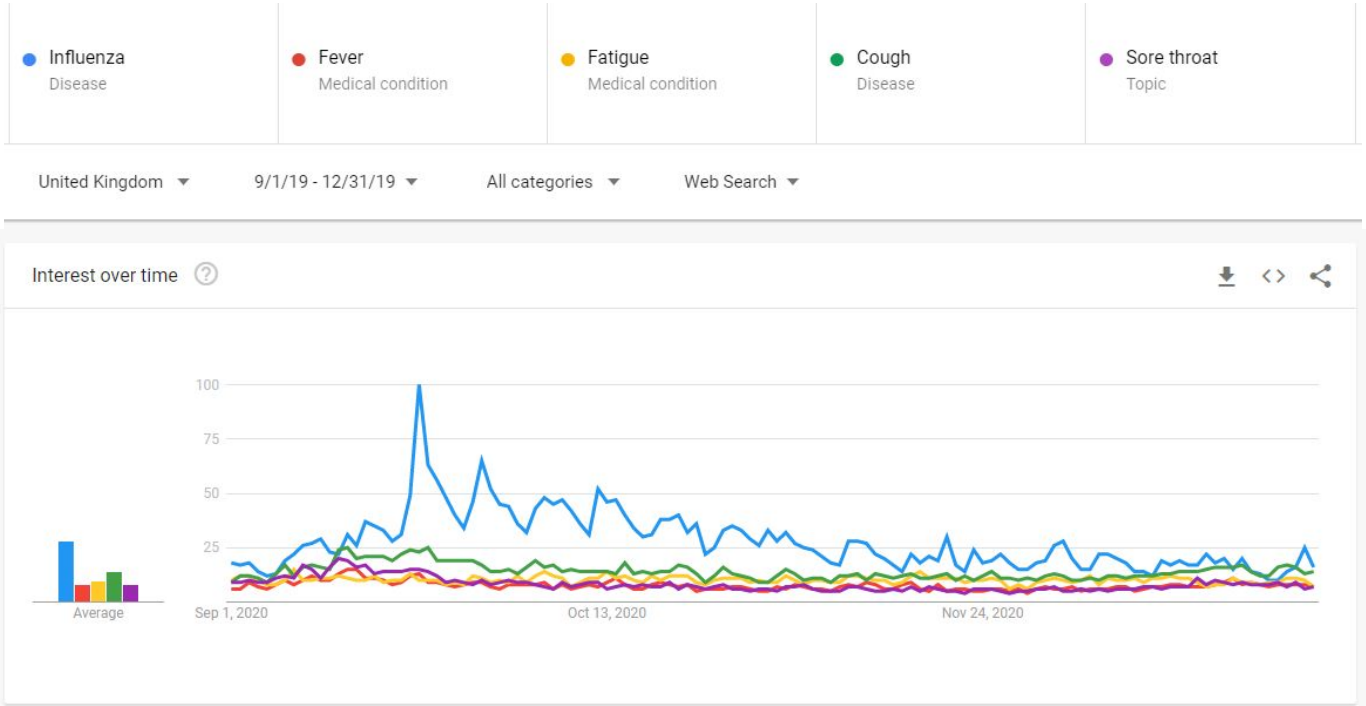
Approach

- Trends in Google searches related to influenza-like illnesses



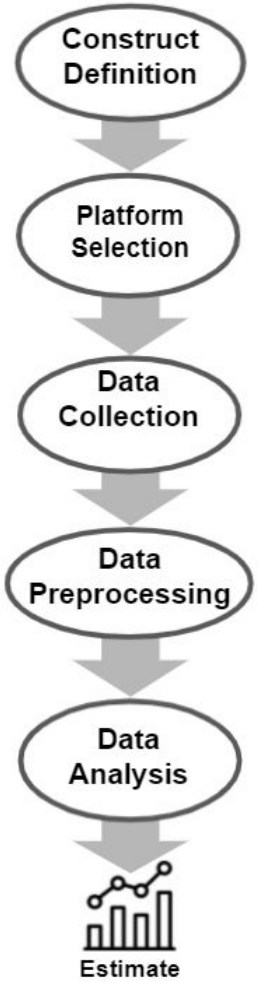
Approach

- Trends in Google searches related to influenza-like illnesses



Approach in a nutshell

- Construct: influenza-like illness (ILI)
- Target population(s): national and regional populations of the US
- Platform: Google web search
- Data collection: Logs 2003-2008, weekly, From 50 mio. search terms, select those that best predict.
- Preprocessing: aggregated per each region, normalized by overall search activity, location via IP.
- Analysis: Regression where DV= ILI doctor visits, IV= ILI-related search fractions, for 1152 data points



Approach

- Result

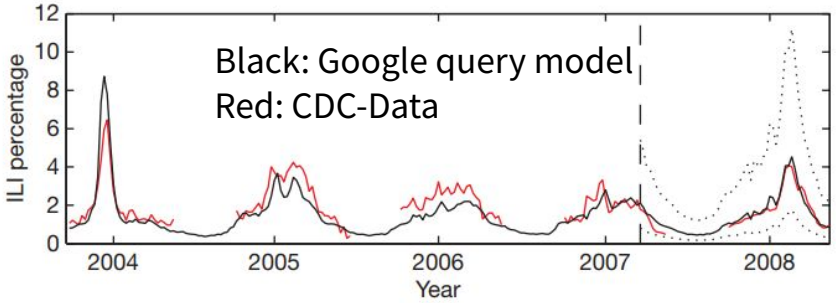


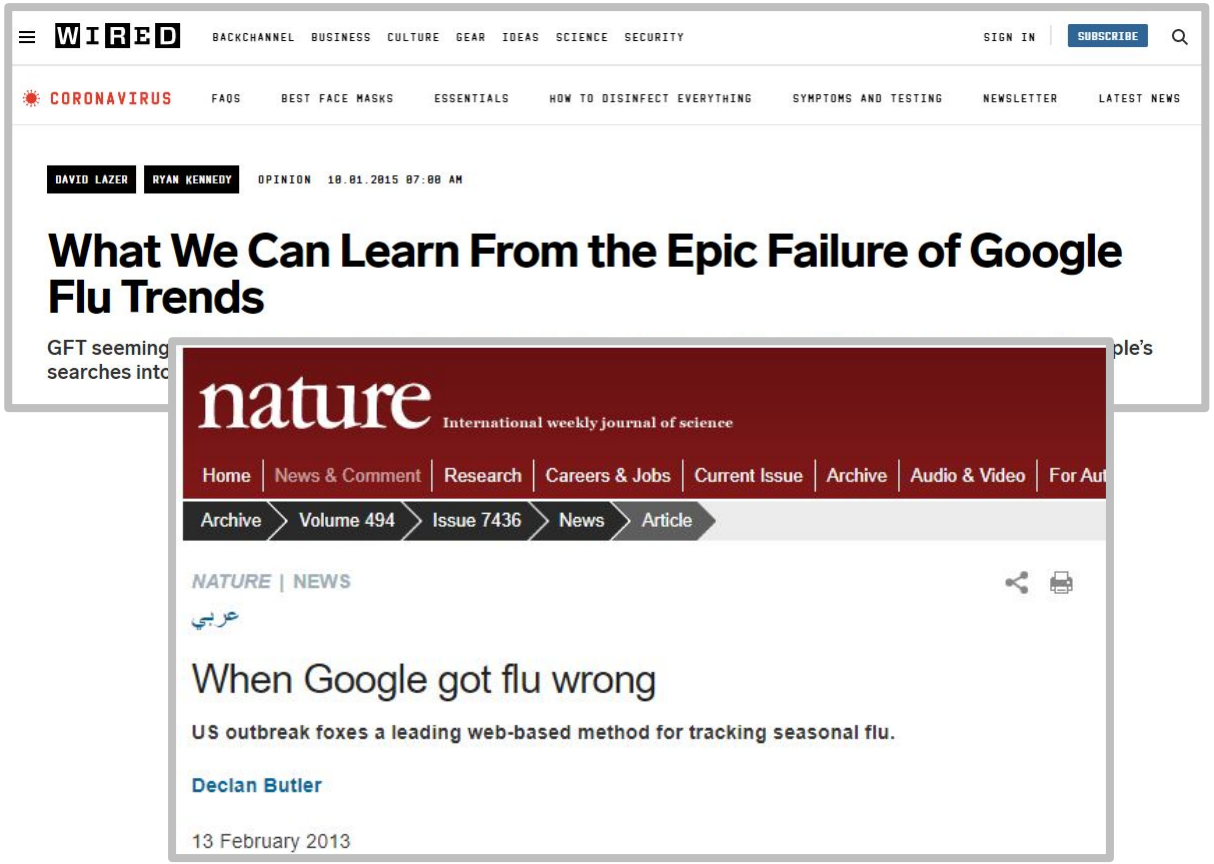
Figure 2 | A comparison of model estimates for the mid-A
(black) against CDC-reported ILI perce
which the model was fi
over 128 points fr
correlation of 0.96
indicate 95% predic
Jersey and Pennsylva
ases New York, New

Problem solved

Good concurrent validity. Content validity of queries seems ok. + other validations.

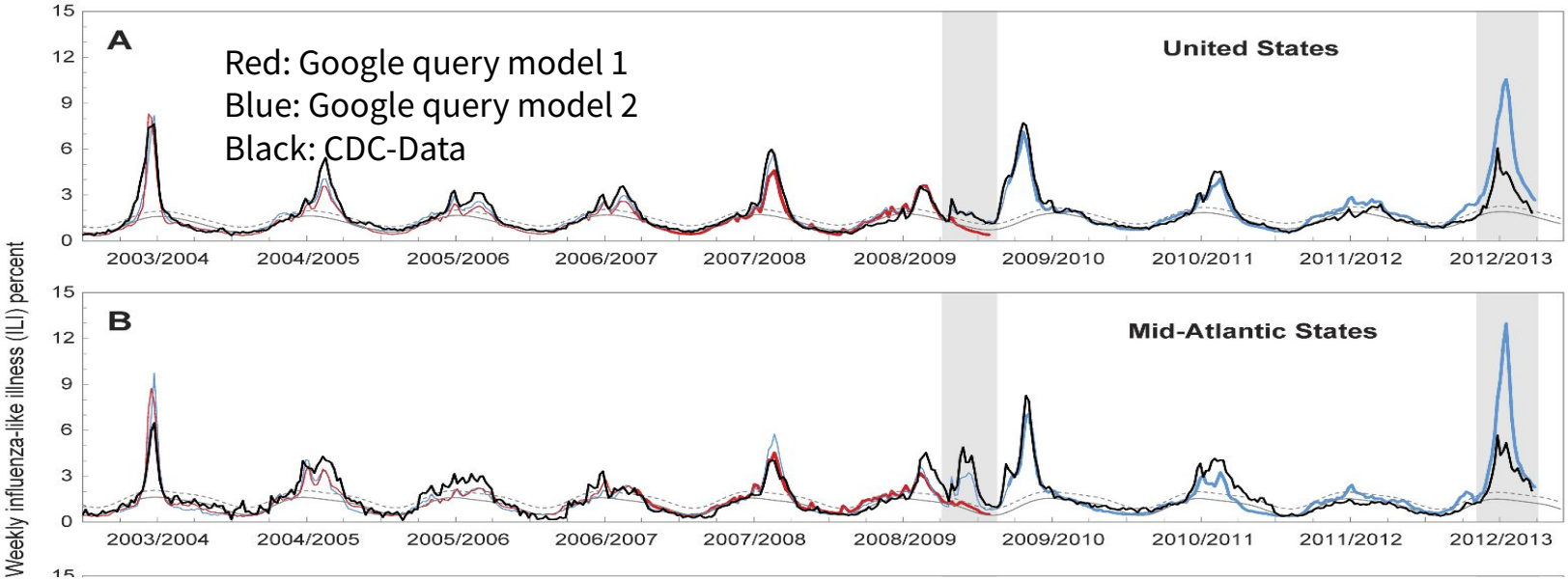
- Now possible: Now- or fore-casting, fine-grained local detection, other countries

The fail



"In February 2013, Google Flu Trends (GFT) made headlines but not for a reason that Google executives or the creators of the flu tracking system would have hoped. Nature reported that GFT was predicting more than double the proportion of doctor visits for influenza-like illness (ILI) than the Centers for Disease Control and Prevention (CDC), which bases its estimates on surveillance reports from laboratories across the United States (1, 2)."

Overestimation



"part flue detector, part winter detector" □ errors are auto-correlated & direction and magnitude change with seasons

Where to be careful

"[…] there are enormous scientific possibilities in big data [but] the core challenge is that most big data that have received popular attention are not the output of instruments designed to produce valid and reliable data amenable for scientific analysis."

"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"

Lazer, David, et al. "The parable of Google Flu: traps in big data analysis." *Science* 343.6176 (2014)

This Course: Objectives

- Critically reflect on the validity and reliability of using digital trace data for social science research

- Systematically assess how to use digital traces and computational models for social science research

- Learn techniques for mitigating errors in studies using digital trace data

Course Logistics

- This is a reading seminar: each session, one person will *lead* the discussion on one (or more) papers.
- The paper(s) will be announced at least one week earlier. All participants are expected to read the paper
- After the lead person has given a presentation on the paper, we will have a discussion
- Each discussant will follow a particular role (more on this later)

Course Logistics (contd…)

- Leaders of the papers are:
  - Me (Indira)
  - Guest speakers (people who have authored papers on this topic)
  - And you!

- Discussants: you

- You will be graded on:
  - presentation of the paper when you are the lead (30%)
  - participation in discussions after each presentation when you are the discussant (40%)
  - report on the paper you led (30%) [to be submitted at the end of the course]

Course Schedule

date	title	who leads
Apr 10	Introduction and kickoff	
Apr 17	no class	
Apr 24	How to read and review a research paper AND overview of research w/ digital traces	
5/1/2024	no class	
May 8	Social data biases (Olteanu)	Indira
May 15	measurement and representation errors (TED-On)	Indira
May 22	guest presentation	guest [TBD]
May 29	no class	
Jun 5	student presentation	
Jun 12	guest presentation [Max Pellert]	Max
Jun 19	student presentation	
Jun 26	guest presentation	guest [TBD]
Jul 3	student presentation	
Jul 10	guest presentation [Giordano de Marzo]	Giordano
Jul 17	student presentation	
Jul 24	guest presentation	guest [TBD]
Jul 31	student presentation	
Aug 7	student presentation	

Which papers are we going read?

- Research on digital trace data for social science
- Particularly errors, biases, and other pitfalls when using digital traces and how to overcome them
- Guest presenters have been chosen keeping with this topic
- You can pick the paper you'd like to discuss, but you have to get it vetted by me first. I also have some suggestions later in the slides
- First presentation slot: June 5th

What you will be graded on

1. When leading and in your report:
    a. Insight into the motivation of the paper. What question does it claim to answer? Are there gaps in motivation? Can it be updated if the paper is a bit old?
    b. Broader context. Reference to other research seen in the lectures, both by the lecturer, guests and by other students: which have similar or opposite aims?
    c. Critical reflection on plausibility, evidence, and insights of the work based on what we have seen in the lectures
2. When discussing others' work:
    a. Situating in relevant related work
    b. Justifying identified strengths and limitations of the work

What you will be graded on

1. When leading and in your report:
    a. Insight into the motivation of the paper. What question does it claim to answer? Are there gaps in motivation? Can it be updated if the paper is a bit old?
    b. Broader context. Reference to other research seen in the lectures, both by the lecturer,
    c. Critical reflection on plausibility, evidence, and insights of the work based on what we have seen in the lectures
2. When discussing others' work:
    a. Situating in relevant related work
    b. Justifying identified strengths and limitations of the work

We will look into how to critically read papers and critique them in the next lecture

Final report on your chosen paper

You can be creative here, but these are recommended subsections:

- Summary: try to be as objective here as possible
- Paper outline: a deeper outline of the main points of the paper, including it's context wrt related work, assumptions made, arguments presented, data analyzed, and conclusions drawn.
- Strengths
- Weaknesses and limitations
- Improvement suggestions and future Work

Send the final report as a PDF document (max. 10 pages, min. font size 11pt) via email to indira.sen@uni-konstanz.de

References do not count towards the page limit.

Discussant Roles

- Based on: https://colinraffel.com/blog/role-playing-seminar.html

Archaeologist. This paper was found buried under ground in the desert. You're an archeologist who must determine where this paper sits in the context of previous and subsequent work. Find and report on one *older* paper cited within the current paper that substantially influenced the current paper and one *newer* paper that cites this current paper.

Academic Researcher. You're a researcher who is working on a new project in this area. Propose an imaginary follow-up project *not just* based on the current but only possible due to the existence and success of the current paper.

Industry Practitioner. You work at a company or organization developing an application or product of your choice (that has not already been suggested in a prior session). Bring a convincing pitch for why you should be paid to implement the method in the paper, and discuss at least one positive and negative impact of this application.

Discussant Roles

- Based on: https://colinraffel.com/blog/role-playing-seminar.html

Social Impact Assessor. Identify how this paper self-assesses its (likely positive) impact on the world. Have any additional positive social impacts left out? What are possible negative social impacts that were overlooked or omitted?

Private Investigator. You are a detective who needs to run a background check on one of the paper's authors. Where have they worked? What did they study? What previous projects might have led to working on this one? What motivated them to work on this project? Feel free to contact the authors, but remember to be courteous, polite, and on-topic.

Discussant Roles [Bonus]

- Based on: https://colinraffel.com/blog/role-playing-seminar.html

>  **Hacker.** You're a hacker who needs a demo of this paper ASAP. Implement a small part or simplified version of the paper on a small dataset or toy problem. Prepare to share the core code of the algorithm to the class and demo your implementation. Do not simply download and run an existing implementation – though you are welcome to use (and give credit to) an existing implementation for "backbone" code.

- If you choose the 'hacker' role and fulfil it for any paper, you get a 10% bonus

Next steps

- Presentations of students start from June 5th. Slots are random, you can pick whichever works best for you.
  a. There's not really much advantage in picking a later slot
  b. 'Register' your slot and paper by May 24th: send me an email with the paper and date you prefer.
  c. Chat with me first if you pick something other than the suggested papers.
  d. If there are conflicts (two or more students pick the same date or paper), I will adjudicate

- Discussions from you start today!
  a. But, you'll be graded on discussions starting from May 8th. Sessions before that are warm-up
  b. Discussant roles apply May 8th onwards. You can pick them (First Come First Serve) when the paper is announced
  c. but I'll try to make sure everyone gets a chance to play all 5 roles

Course Timeline and Deadlines

24.05.24: register *your* paper and timeslot

5.06.24-7.09.24: present your chosen paper

15.09.24: submit final report on your paper

08.05.24: graded discussions start (pick your roles)

08.05.24-7.09.24: participate in discussions of papers by students and guests

Suggested papers (to give you some ideas, but feel free to pick something else…)

Construct definition

Ruths, Derek, and Jürgen Pfeffer. "Social media for large studies of behavior." Science 346.6213 (2014): 1063-1064.

Blodgett, Su Lin, et al. "Language (Technology) is Power: A Critical Survey of “Bias” in NLP." Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020.

Wagner, Claudia, et al. "Measuring algorithmically infused societies." Nature 595.7866 (2021): 197-204.

Platform Effects

Malik, Momin, and Jürgen Pfeffer. "Identifying platform effects in social media data." Proceedings of the International AAAI Conference on Web and Social Media. Vol. 10. No. 1. 2016.

Gligorić, Kristina, Ashton Anderson, and Robert West. "How constraints affect content: The case of Twitter’s switch from 140 to 280 characters." Proceedings of the International AAAI Conference on Web and Social Media. Vol. 12. No. 1. 2018.

Suggested papers (to give you some ideas, but feel free to pick something else…)

Platform Effects (contd…)

Arazy, Ofer, et al. "Information quality in Wikipedia: The effects of group composition and task conflict." Journal of management information systems 27.4 (2011): 71-98.

Data Collection

Zafar, Muhammad Bilal, et al. "Sampling content from online social networks: Comparing random vs. expert sampling of the twitter stream." ACM Transactions on the Web (TWEB) 9.3 (2015): 1-33.

Gaffney, Devin, and J. Nathan Matias. "Caveat emptor, computational social science: Large-scale missing data in a widely-published Reddit corpus." PloS one 13.7 (2018): e0200162.

Pfeffer, Juergen, et al. "This Sample seems to be good enough! Assessing Coverage and Temporal Reliability of Twitter's Academic API." Proceedings of the International AAAI Conference on Web and Social Media. Vol. 17. 2023.

Suggested papers (to give you some ideas, but feel free to pick something else…)

Data Preprocessing and Modeling

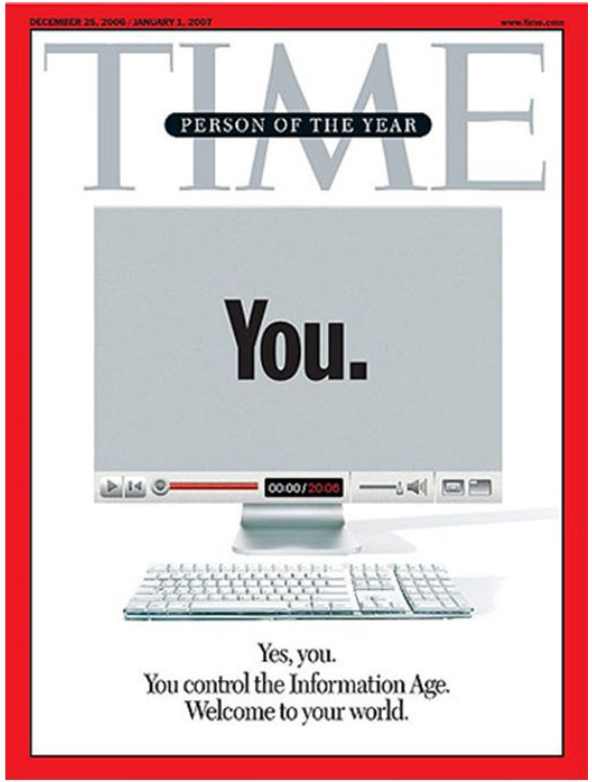
Culotta, Aron. "Reducing sampling bias in social media data for county health inference." Joint Statistical Meetings Proceedings. Citeseer, 2014.

Jurgens, David, et al. "Geolocation prediction in twitter using social networks: A critical analysis and review of current practice." Proceedings of the international AAAI conference on web and social media. Vol. 9. No. 1. 2015.

Cohen, Raviv, and Derek Ruths. "Classifying political orientation on Twitter: It's not easy!." Proceedings of the International AAAI Conference on Web and Social Media. Vol. 7. No. 1. 2013.

Fleisig, Eve, Rediet Abebe, and Dan Klein. "When the majority is wrong: Leveraging annotator disagreement for subjective tasks." arXiv preprint arXiv:2305.06626 (2023).

Lucy, Li, and David Bamman. "Gender and representation bias in GPT-3 generated stories." Proceedings of the third workshop on narrative understanding. 2021.



What brought you to this course?

- Your interest in digital traces (specific types of data or methods you might want to learn more about)
- Prior experience, disciplinary background, etc...

Readings for next lecture (April 24)

1. Keshav, Srinivasan. "How to read a paper." ACM SIGCOMM Computer Communication Review 37.3 (2007): 83-84.
2. Pain, Elisabeth “How to review a paper”

Acknowledgement

Some parts of this lecture are based on the ‘Meet the Experts’ session with Dr. Fabian Flöck and me:
https://www.youtube.com/watch?v=y9mVuQnXWec

Thanks to Fabian for the slides on Google Flu.