Digital Traces via Data Donations

Workshop DGPuK RezFo 2026

Session 1 : Welcome & Intro to Digital Traces

👉 Part of the SPP DFG Project Integrating Data Donations in Survey Infrastructure

UNIVERSITÄT KLAGENFURT

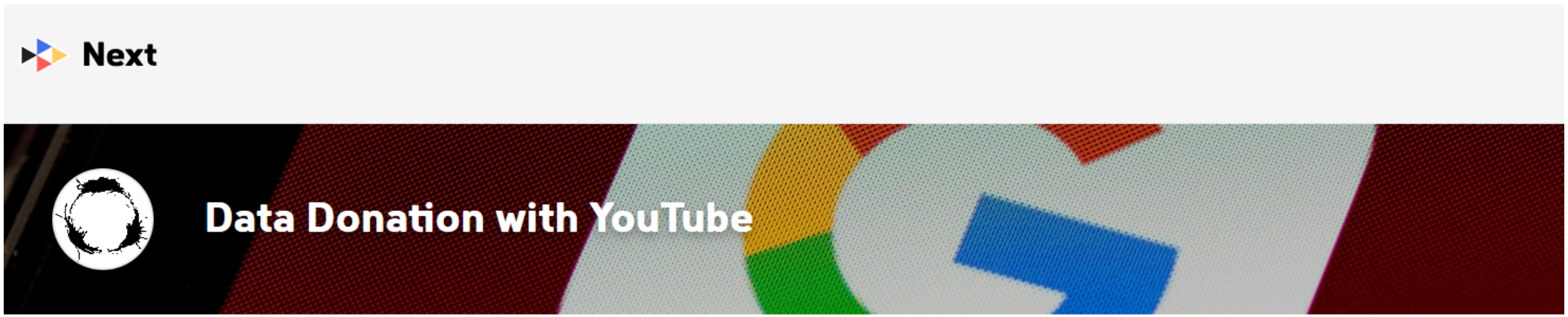
Agenda

1. Intro to the workshop
2. What is digital trace data?
3. How can we collect digital traces?

Image by Hope House Press via Unsplash

Before we start: Have you requested and downloaded your Google Data? 🤔

Otherwise, use this link to request your data now: https://next.eyra.co/a/nWPJC4?p=999 - replace number after *p=* with random number



1. Intro

Source: Image by Markus Winkler via Unsplash

Who are you?

Please raise your hand 👋 if you ....

- are familiar with the term digital trace data
- have worked with APIs
- have worked with data donation
- have worked with automated content analysis
- regularly use programming languages (e.g., R, Python)

Who are you?

In 2-3 sentences, tell us...

- your main research interests
- the methods you mainly use
- related to which theoretical questions/data you are interested in data donation as a method

About me: Valerie Hase

Professor of Digital Media and Communication

- Digital Media and Methods Lab
- University of Klagenfurt

Research interests:

- CSS (automated content analysis, digital traces, bias, data access)
- Digital journalism, crisis communication

👉 More info: github.com/valeriehase & valerie-hase.com

A big thank you 🙌 to the organizers

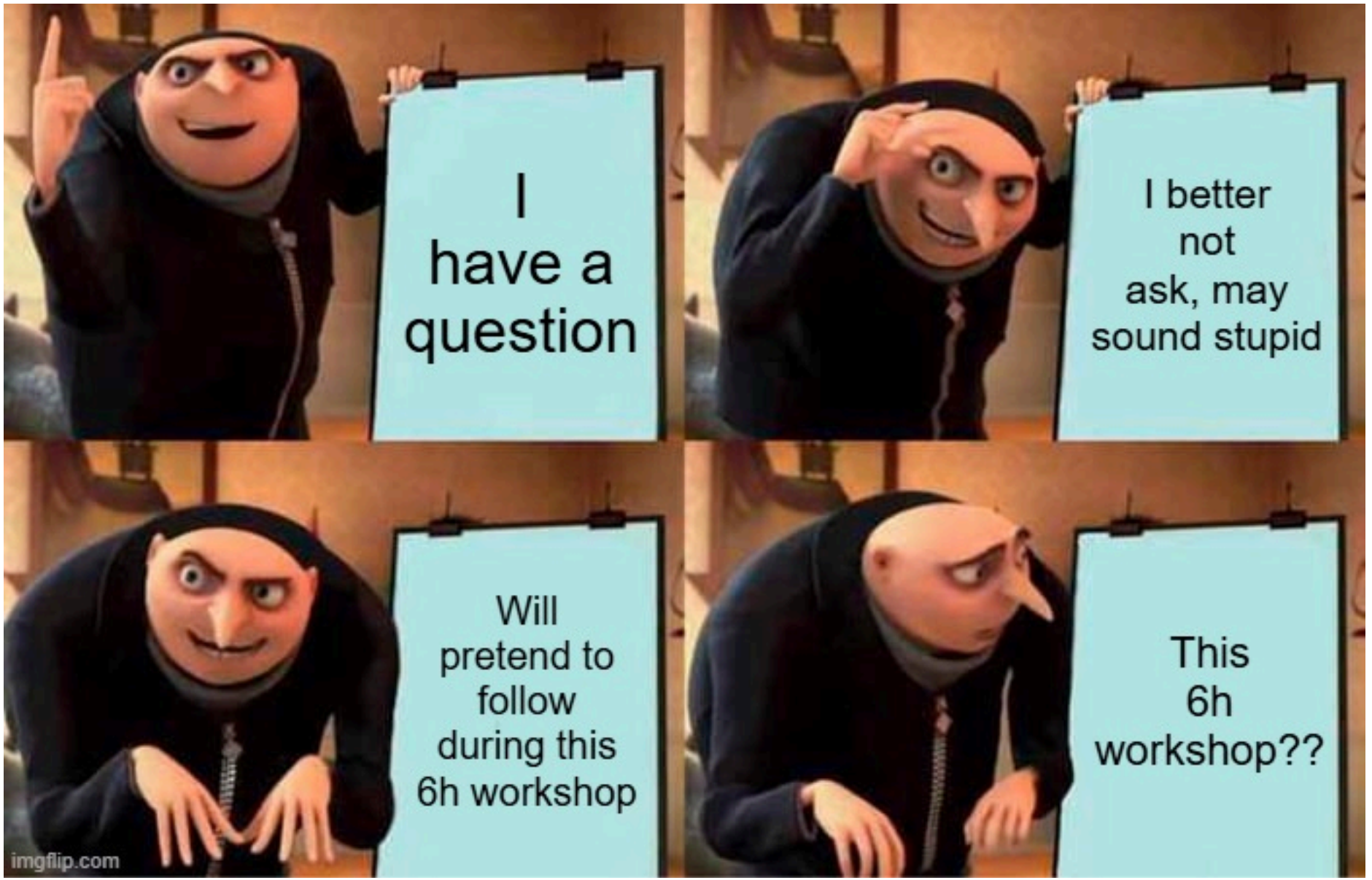
Shoutout to the organizers behind this workshop

- Alicia Ernst & Ulrike Schwertberger (RezFo Young Scholar Network)
- Christina Seeger, Marlis Stubenvoll, Caroline Roth-Ebner, Selina Noetzel & Melanie Hirsch (local organizing team)

What is the goal of this workshop?

- ✅ Understanding digital data traces as a *type* of data
- ✅ Understanding data donation as a *method* of data access
- ✅ Working through key steps of data donation methods (user & researcher view)
- ✅ Discussing when (not) to use data donation studies
- ❌ Detailed implementation (e.g., server set-up, coding data extraction scripts)

How do we communicate in this workshop?



How do we communicate in this workshop?

My goal is that you...

- just **ask right away** if there is something you did not understand
- keep in mind that there **are not stupid questions**
- feel free to ask questions specific to your potential data donation projects!

Timetable

⏰ 11:15-11:45am	Session 1️⃣ : Welcome & Intro to Digital Traces
⏰ 11:45am-1pm	Session 2️⃣ : Data Donation Studies (Participant Perspective)
🥖 1-2pm	Lunch break
⏰ 2-4pm	Session 3️⃣ : Data Donation Studies (Researcher Perspective)
⏰ 4-5pm	Session 4️⃣ : Bias in Digital Trace Data & Outro

2. What is digital trace data?

Source: Image by Markus Winkler via Unsplash

Which examples for digital trace data do you know? 🤔

Example study I

Revisiting the Digital Jukebox in Daily Life: Applying Mood Management Theory to Algorithmically Curated Music Streaming Environments

Alicia Ernst, Felix Dietrich, Benedikt Rohr, Leonard Reinecke, & Michael Scharkow

Department of Communication, Johannes Gutenberg University of Mainz, Mainz, Germany

This manuscript is a preprint and has not been peer-reviewed.

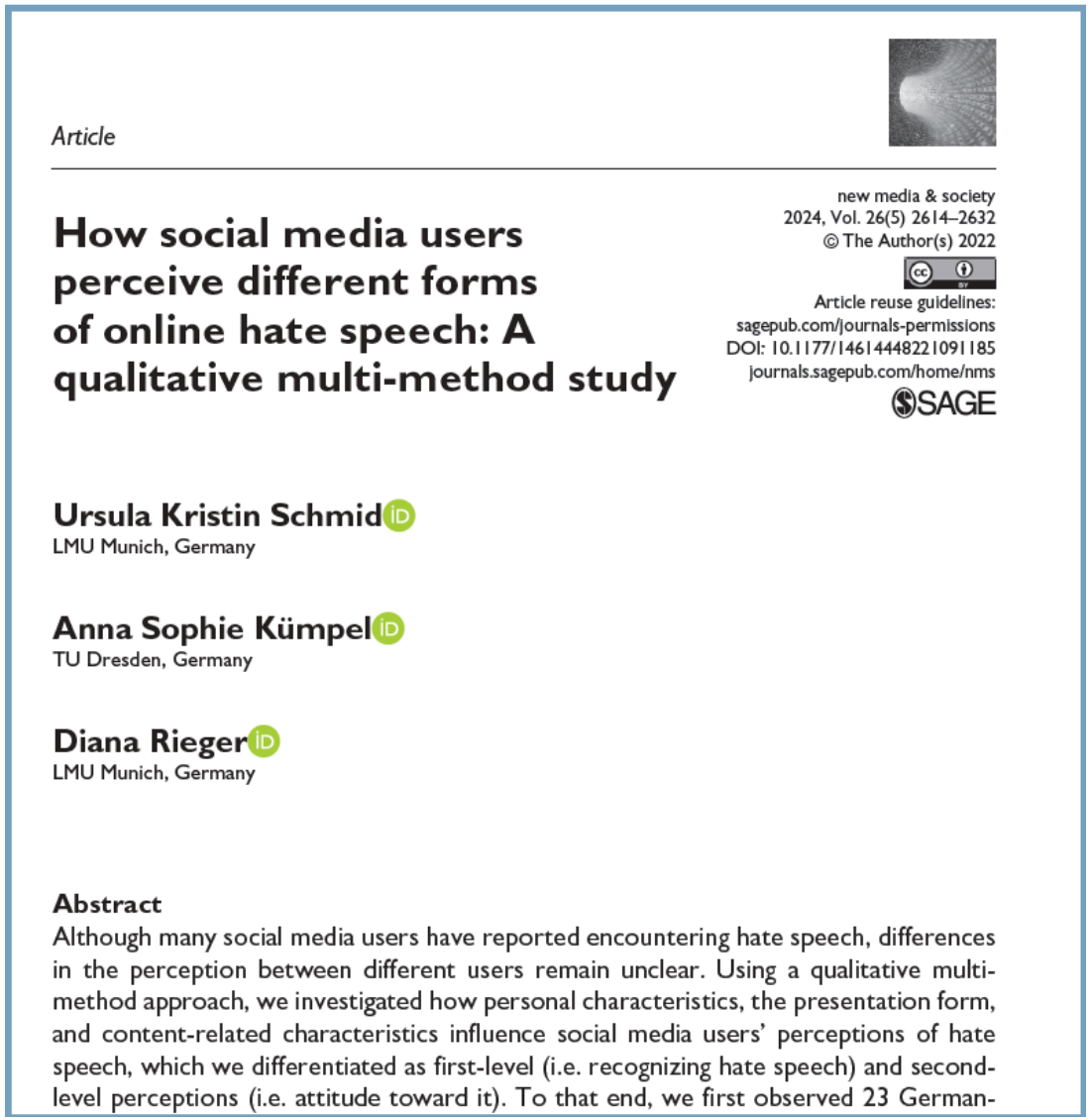
Last updated: January 14th, 2025

Author Note

Correspondence concerning this article should be addressed to: Alicia Ernst, Johannes Gutenberg University Mainz, Department of Communication, Jakob-Welder-Weg 12, 55128 Mainz, Germany. Email: alicia.ernst@uni-mainz.de

Ernst, Dietrich et al. (2025)

Example study II



Article

new media & society
2024, Vol. 26(5) 2614–2632

DOI: 10.1177/14614448221091185
journals.sagepub.com/home/nms
SAGE

How social media users perceive different forms of online hate speech: A qualitative multi-method study

Ursula Kristin Schmid
LMU Munich, Germany

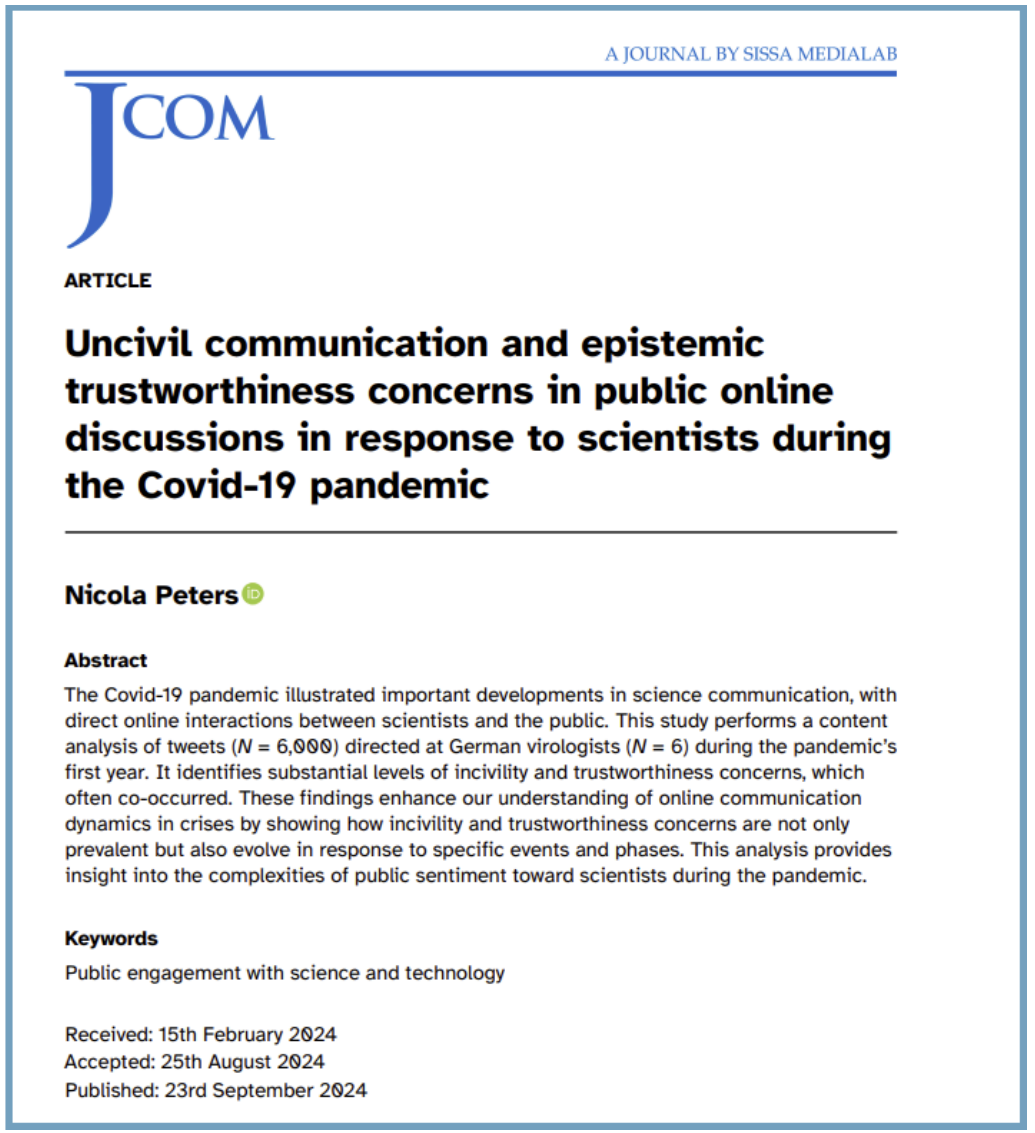
Anna Sophie Kümpel
TU Dresden, Germany

Diana Rieger
LMU Munich, Germany

Abstract
Although many social media users have reported encountering hate speech, differences in the perception between different users remain unclear. Using a qualitative multi-method approach, we investigated how personal characteristics, the presentation form, and content-related characteristics influence social media users' perceptions of hate speech, which we differentiated as first-level (i.e. recognizing hate speech) and second-level perceptions (i.e. attitude toward it). To that end, we first observed 23 German-

Schmid et al. (2024)

Example study III



A JOURNAL BY SISSA MEDIALAB

JCOM

ARTICLE

Uncivil communication and epistemic trustworthiness concerns in public online discussions in response to scientists during the Covid-19 pandemic

Nicola Peters

Abstract

The Covid-19 pandemic illustrated important developments in science communication, with direct online interactions between scientists and the public. This study performs a content analysis of tweets (N = 6,000) directed at German virologists (N = 6) during the pandemic's first year. It identifies substantial levels of incivility and trustworthiness concerns, which often co-occurred. These findings enhance our understanding of online communication dynamics in crises by showing how incivility and trustworthiness concerns are not only prevalent but also evolve in response to specific events and phases. This analysis provides insight into the complexities of public sentiment toward scientists during the pandemic.

Keywords

Public engagement with science and technology

Received: 15th February 2024
Accepted: 25th August 2024
Published: 23rd September 2024

Peters (2024)

What is digital trace data?

Definition 💡 : *The recording and storing of activities on digital platforms to draw conclusions about digital and analog phenomena*

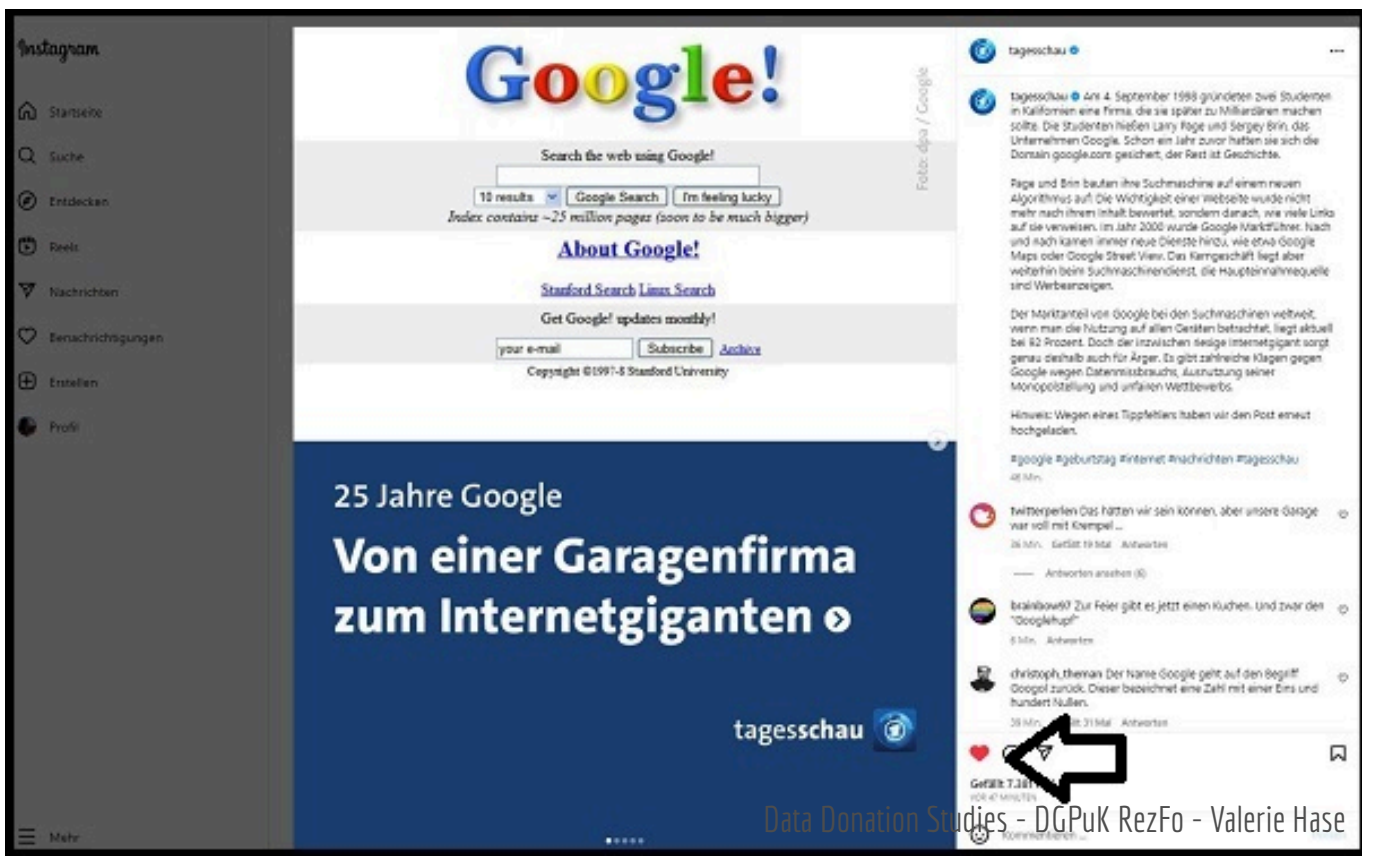
- e.g., tweets, likes, shares on social media
- e.g., geo data (locations, movements)
- e.g., digital payments
- e.g., Spotify playlists

What is digital trace data?

Definition 💡 : *The recording and storing of activities on digital platforms to draw conclusions about digital and analog phenomena*

- e.g., tweets, **likes**, shares on social media

Example: Instagram Like



What is digital trace data?

Definition 💡 : *The recording and storing of activities on digital platforms to draw conclusions about digital and analog phenomena*

- e.g., tweets, **likes**, shares on social media

Example: Instagram Like

Which systems/sensors collect traces?

- Apps & social media platforms (e.g., Instagram)
- Payment systems (e.g., Paypal)
- Wearable devices (e.g., smart watch, fitbit)
- Governmental databases (e.g., health data)

Which types of data does this include?

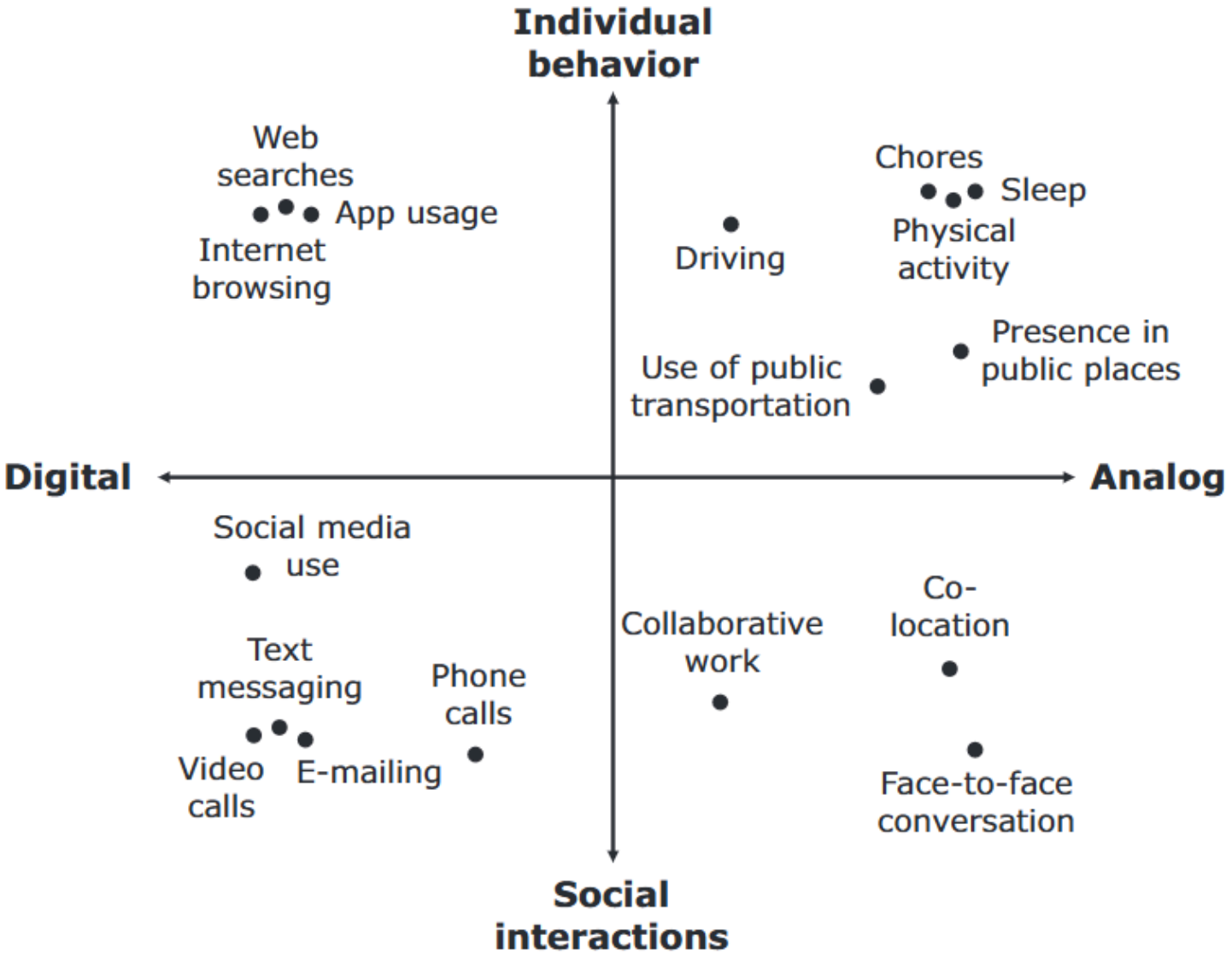
Depending on the data collection method... (Haim & Hase, 2023; Ohme et al., 2024):

- often fine-grained (e.g., time-stamped)
- often longitudinal (e.g., over years, within-individual change)
- often less reactive (e.g., less concerns about social desirability)

Filter

	external_submission_id	engagement_timestamp	day	search_query	donation_platform	donation_type
1	10135	2018-01-03 12:06:02	2018-01-03	robot fail compilation	YouTube	searched
2	10135	2017-01-02 11:53:31	2017-01-02	kuchentv	YouTube	searched
3	6877	2018-10-25 21:35:39	2018-10-25	full house	YouTube	searched
4	6648	2015-11-25 23:06:58	2015-11-25	messias händel halleluja	YouTube	searched
5	10135	2013-04-23 08:45:48	2013-04-23	barlow	YouTube	searched
6	6877	2019-11-01 22:24:05	2019-11-01	csi safri duo	YouTube	searched
7	6877	2013-12-07 19:47:04	2013-12-07	coca cola christmas commercial	YouTube	searched
8	6877	2014-04-13 20:06:51	2014-04-13	dawn of the dead trailer	YouTube	searched
9	6877	2016-05-15 19:42:18	2016-05-15	agnes release me	YouTube	searched
10	6877	2015-06-08 20:25:01	2015-06-08	evanescence rock am ring 2003	YouTube	searched
11	6877	2022-02-15 17:58:46	2022-02-15	missy elliott lyrics	YouTube	searched
12	9126	2021-01-22 18:50:22	2021-01-22	vegan ist ungesund	YouTube	searched
13	10135	2015-06-07 10:51:59	2015-06-07	robert downey jr singing	YouTube	searched
14	10135	2012-08-30 07:22:01	2012-08-30	counter strike	YouTube	searched
15	6877	2014-12-08 21:37:49	2014-12-08	the flash video	YouTube	searched
16	6877	2012-03-27 15:07:56	2012-03-27	ncis mcgee	YouTube	searched
17	9837	2022-01-11 18:14:56	2022-01-11	video in instagram beitrag	YouTube	searched
18	10135	2020-12-23 09:17:48	2020-12-23	unusual memes	YouTube	searched
19	10135	2013-08-14 09:30:16	2013-08-14	all cry	YouTube	searched
20	6877	2012-09-17 20:54:08	2012-09-17	dolph lundgren video	YouTube	searched

Which types of data does this include?



Source: Keusch & Kreuter, 2023, p. 102

What (latent) phenomena are we trying to explain?

- **How (often) people use the internet/apps** (Parry et al., 2021) ...
  - to explain well-being (Ohme et al., 2024) & sleep quality (Siebers et al., 2024)
  - to explain voting behavior (Bach et al., 2021)
- **What content people are exposed to** ...
  - fake news (Lyons et al., 2024)
  - health-related info (Bachl et al., 2024)
- **Where and how people move** to explain ...
  - mobility during pandemics (Li et al., 2021)
  - how deprived neighbourhoods can benefit from green spaces (Pawlowski et al., 2019)
- **Career trajectories / financial behavior** ...
  - how relationships affect womens' earnings (Möhring & Weiland, 2022)
  - whether overqualification can lead to wage loss (Kracke et al., 2018)

Why are digital traces becoming more popular?

- Problems with self-reported data (e.g., via survey)

“How many minutes a day do you use the internet to consume news?”

Source: Image by Scott Graham via Unsplash

- „internet”?
- „news”?
- „how many minutes”?

Why are digital traces becoming more popular?

- **Problems with self-reported data** (e.g., via survey)
  - Inaccurate measurements (recall issues)
  - Bias (Parry et al., 2021; Scharkow, 2016): individual characteristics may predict under- or overreporting
  - Declining response rates in surveys (Luiten et al., 2020)

Why are digital traces becoming more popular?

- **Problems with self-reported data** (e.g., via survey)
- **Availability of digital traces**
  - cheap (e.g., via APIs)
  - large data sets (“big data”)
  - more accurate (“objective data”)

Why are digital traces becoming more popular?

- Problems with self-reported data (e.g., via survey)
- Availability of digital traces

Why are digital traces becoming more popular?

- Problems with self-reported data (e.g., via survey)
- Availability of digital traces

Be careful: These "advantages" of traces are often claimed, but **empirically disputed**.

Digital traces are **neither** necessarily less biased, nor cheaper, or larger (we will discuss this in Session 4️⃣).

(Dis-)advantages of digital trace data

- ✅ More fine-grained, often longitudinal measures due to timestamps
- ✅ Partly measurement of new variables (e.g., algorithmic inference)
- ❌ Still bias due to errors in representation and measurement
- ❌ Implementation can be expensive and cumbersome

👉 More data does not mean better data!

Summary: What is digital trace data? 📚

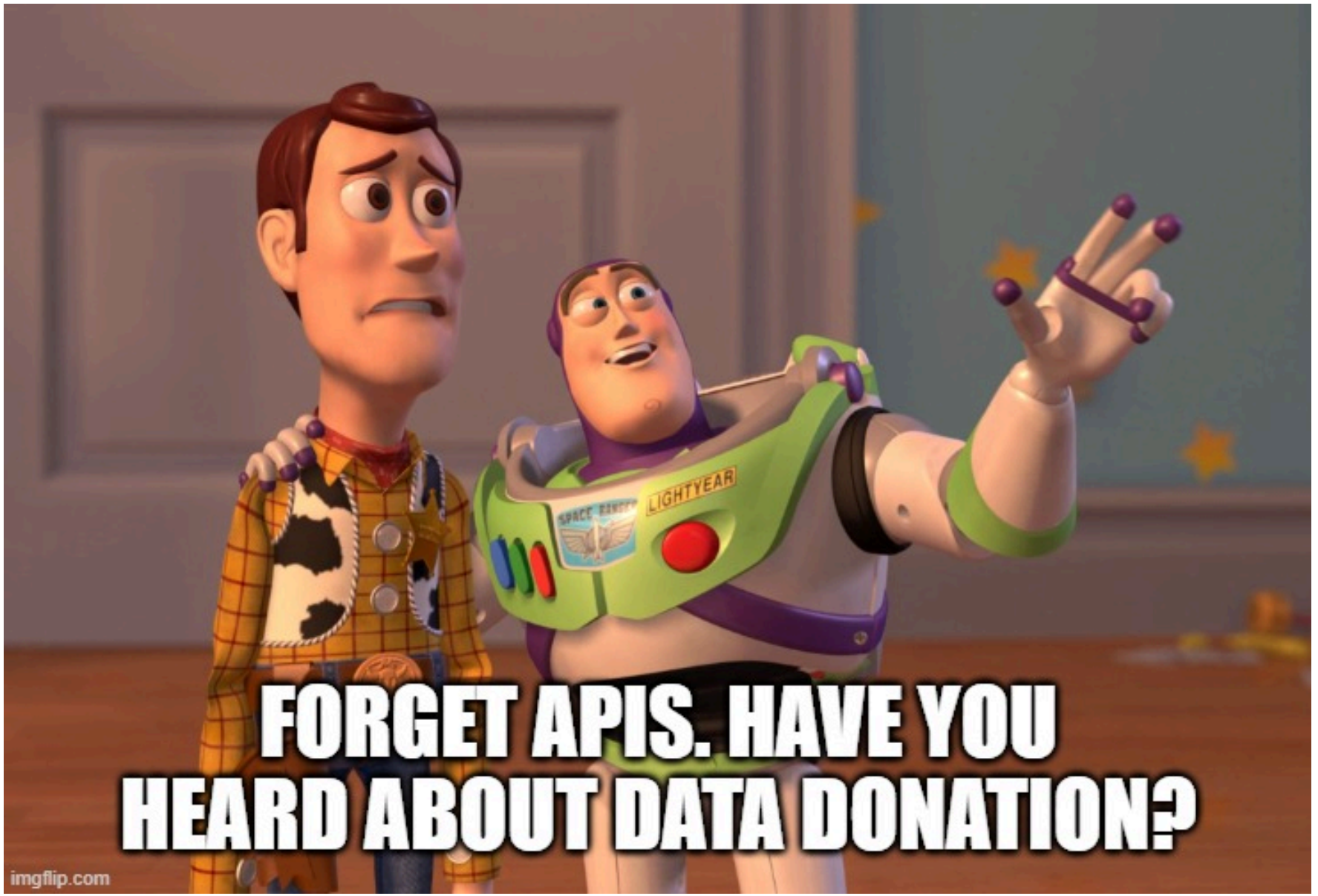
- **Definition**: *The recording and storing of activities on digital platforms to draw conclusions about digital and analog phenomena*
- **Further literature**
  - Keusch & Kreuter (2021)
  - Haim & Hase (2023)
  - Ohme et al. (2024)

3. How can we collect digital traces?

Source: Image by Markus Winkler via Unsplash

Which methods do you know/have you used for collecting digital trace data*?* 🤔

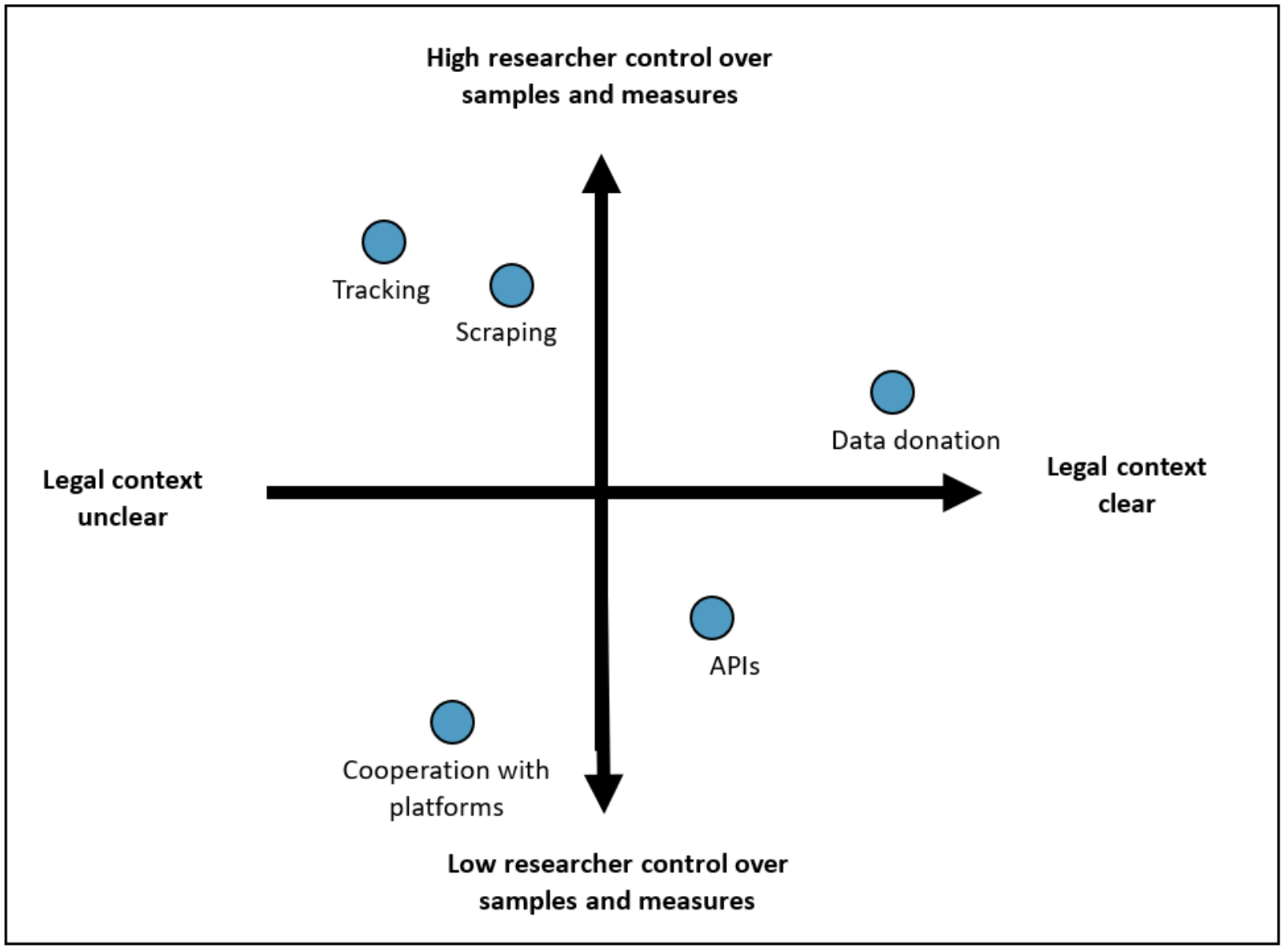
Platform- and user-centric methods



Platform- and user-centric methods

- **Platform-centric** (based on platform cooperation)
  - API (Jünger, 2021)
  - Cooperation with platforms (Wagner, 2023)
- **User-centric** (based on user cooperation and informed consent) or "follow the user" approaches (Caliandro, 2024)
  - Data donation (Carrière et al., 2024)
  - Linkage to existing databases (Sloan et al., 2020)
  - Active sharing via sensors (Struminskaya et al., 2021)
  - Passive sharing via sensors/tracking (Christner et al., 2022)

Platform- and user-centric methods



Platform- and user-centric methods

- **Restrictions of platform-centric methods**
  - Discontinuation of APIs (Freelon, 2018)
  - Concerns about bias (Schatto-Eckrodt, 2022; Ulloa et al., 2025)
- **User-centric methods become more popular, given …**
  - Legal frameworks enabling such studies (GDPR, DSA)
  - Presumably (!) more researcher control
  - Ethical considerations (informed consent)

Summary: How can we collect digital traces? 📚

- **Summary**
  - Platform-centric methods (e.g., APIs) and user-centric methods (e.g., data donation)
  - Key differences: control over samples & measurements, legal & ethical contexts
- **Further literature**
  - Haim & Hase (2023)
  - Ohme et al. (2024)

Questions? 

References

Bach, R. L., Kern, C., Amaya, A., Keusch, F., Kreuter, F., Hecht, J., & Heinemann, J. (2021). Predicting Voting Behavior Using Digital Trace Data. *Social Science Computer Review*, *39*(5), 862–883. https://doi.org/10.1177/0894439319882896

Bachl, M., Link, E., Mangold, F., & Stier, S. (2024). Search Engine Use for Health-Related Purposes: Behavioral Data on Online Health Information-Seeking in Germany. *Health Communication*, *39*(8), 1651–1664. https://doi.org/10.1080/10410236.2024.2309810

Caliandro, A. (2024). Follow the user: Taking advantage of Internet users as methodological resources. *Convergence: The International Journal of Research into New Media Technologies*, 13548565241307569. https://doi.org/10.1177/13548565241307569

Carrière, T. C., Boeschoten, L., Struminskaya, B., Janssen, H. L., De Schipper, N. C., & Araujo, T. (2024). Best practices for studies using digital data donation. *Quality & Quantity*. https://doi.org/10.1007/s11135-024-01983-x

Christner, C., Urman, A., Adam, S., & Maier, M. (2022). Automated Tracking Approaches for Studying Online Media Use: A Critical Review and Recommendations. *Communication Methods and Measures*, *16*(2), 79–95. https://doi.org/10.1080/19312458.2021.1907841

Freelon, D. (2018). Computational research in the post-API age. *Political Communication*, *35*(4), 665–668. https://doi.org/10.1080/10584609.2018.1477506

Haim, M., & Hase, V. (2023). Computational Methods und Tools für die Erhebung und Auswertung von Social-Media-Daten. In S. Stollfuß, L. Niebling, & F. Raczkowski (Eds.), *Handbuch Digitale Medien und Methoden* (pp. 1–20). Springer Fachmedien Wiesbaden. https://link.springer.com/10.1007/978-3-658-36629-2_41-1

Jünger, J. (2021). A brief history of APIs. In *Handbook of Computational Social Science, Volume 2* (1st ed., pp. 17–32). Routledge. https://www.taylorfrancis.com/books/9781003025245/chapters/10.4324/9781003025245-3

Keusch, F., & Kreuter, F. (2021). Digital trace data. In *Handbook of Computational Social Science, Volume 1* (1st ed., pp. 100–118). Routledge. https://www.taylorfrancis.com/books/9781003024583/chapters/10.4324/9781003024583-8

Kracke, N., Reichelt, M., & Vicari, B. (2018). Wage Losses Due to Overqualification: The Role of Formal Degrees and Occupational Skills. *Social Indicators Research*, *139*(3), 1085–1108. https://doi.org/10.1007/s11205-017-1744-8

Li, X., Xu, H., Huang, X., Guo, C., Kang, Y., & Ye, X. (2021). Emerging geo-data sources to reveal human mobility dynamics during COVID-19 pandemic: Opportunities and challenges. *Computational Urban Science*, *1*(1), 22. https://doi.org/10.1007/s43762-021-00022-x

Luiten, A., Hox, J., & Leeuw, E. de. (2020). Survey Nonresponse Trends and Fieldwork Effort in the 21st Century: Results of an International Study across Countries and Surveys. *Journal of Official Statistics*, *36*(3), 469–487. https://doi.org/10.2478/jos-2020-0025

Lyons, B., Montgomery, J. M., & Reifler, J. (2024). Partisanship and Older Americans' Engagement with Dubious Political News. *Public Opinion Quarterly*, *88*(3), 962–990. https://doi.org/10.1093/poq/nfae044

Möhring, K., & Weiland, A. P. (2022). Couples' Life Courses and Women's Income in Later Life: A Multichannel Sequence Analysis of Linked Lives in Germany. *European Sociological Review*, *38*(3), 371–388. https://doi.org/10.1093/esr/jcab048

Ohme, J., Araujo, T., Boeschoten, L., Freelon, D., Ram, N., Reeves, B. B., & Robinson, T. N. (2024). Digital Trace Data Collection for Social Media Effects Research: APIs, Data Donation, and (Screen) Tracking. *Communication Methods and Measures*, *18*(2), 124–141. https://doi.org/10.1080/19312458.2023.2181319

Parry, D. A., Davidson, B. I., Sewall, C. J. R., Fisher, J. T., Mieczkowski, H., & Quintana, D. S. (2021). A systematic review and meta-analysis of discrepancies between logged and self-reported digital media use. *Nature Human Behaviour*, *5*(11), 1535–1547. https://doi.org/10.1038/s41562-021-01117-5

Pawlowski, C. S., Schmidt, T., Nielsen, J. V., Troelsen, J., & Schipperijn, J. (2019). Will the children use it?—A RE-AIM evaluation of a local public open space intervention involving children from a deprived neighbourhood. *Evaluation and Program Planning*, *77*, 101706. https://doi.org/10.1016/j.evalprogplan.2019.101706

Scharkow, M. (2016). The Accuracy of Self-Reported Internet Use—A Validation Study Using Client Log Data. *Communication Methods and Measures*, *10*(1), 13–27. https://doi.org/10.1080/19312458.2015.1118446

Schatto-Eckrodt, T. (2022). Hidden biases – The effects of unavailable content on Twitter on sampling quality. In *Grenzen, Probleme und Lösungen bei der Stichprobenziehung* (pp. 178–195). Halem.

Siebers, T., Beyens, I., Baumgartner, S. E., & Valkenburg, P. M. (2024). Adolescents' Digital Nightlife: The Comparative Effects of Day- and Nighttime Smartphone Use on Sleep Quality. *Communication Research*, 00936502241276793. https://doi.org/10.1177/00936502241276793

Sloan, L., Jessop, C., Al Baghal, T., & Williams, M. (2020). Linking Survey and Twitter Data: Informed Consent, Disclosure, Security, and Archiving. *Journal of Empirical Research on Human Research Ethics*, *15*(1-2), 63–76.