GESIS Leibniz Institute for the Social Sciences

# Linking surveys, web tracking and social media data

Leibniz Association

**Workshop: “Linking Digital Footprint and Survey Data for Open Research”**

Manchester, February 14 2025

Sebastian Stier

# Agenda

1. User-centered collections of social media data
    - Application: Political interest and Facebook news exposure
2. Major challenges in this space and infrastructure work at GESIS

GESIS Leibniz Institute for the Social Sciences

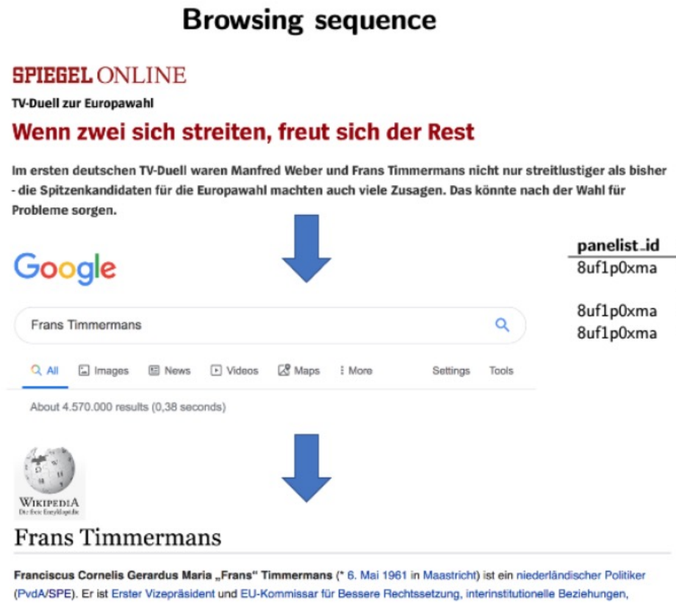
# User-centered collections of social media data

## Application: Political interest and Facebook news exposure

Leibniz Association

# User-centered collections of social media data

- User-centered vs. platform-centered collections of digital behavioral data (Breuer et al., 2023; Stier et al., 2020)
- Web tracking via browser plugins. Limitations so far:
  - data collected by commercial market research panels
  - only URLs delivered –> ex-post web scraping of content necessary
  - no social media content
  - no data archiving and secondary use




**Dataset**

| panelist_id | url | duration |
|---|---|---|
| 8uf1p0xma | https://www.spiegel.de/politik/ausland/europawahl-so-lief-das-tv-duell-zwischen-manfred-weber-und-frans-timmermans-a-1266281.html | 43 |
| 8uf1p0xma | https://www.google.com/search?q=frans+timmermans | 3 |
| 8uf1p0xma | https://de.wikipedia.org/wiki/Frans_Timmermans | 56 |

# Case study on German federal election 2021

- Academic web tracking tool (Adam et al., 2024)
  - Web tracking on desktop computers and laptops
  - Block list of web domains related to porn, banking, illegal content
  - Direct “in situ” scraping of HTML
  - Scraping of public posts seen by participants on Facebook
- 3-month data collection before and after election day
- Quota sample (N=739 persons) recruited from an online access panel

| **Sample** | **Website visits** | | | **Public Facebook posts** | |
|---|---|---|---|---|---|
| N persons | Total visits | News visits | **Facebook visits** | **Total public posts** | Public news posts |
| 739 | 8,358,879 | 104,068 | **137,636** | **370,466** | 68,545 |

# Does Facebook foster inequalities or is it facilitating access to news?

**Related research**

- Facebook as a pathway to news (Fletcher et al., 2018; Scharkow et al., 2021; Stier et al., 2022; Wojcieszak et al., 2021)
- Individual-level inequalities in news exposure within Facebook (Kümpel, 2019; Thorson et al., 2021)
- Political news made up 6-8% of Facebook content during the 2020 US election (Guess et al., 2023)

**Research question**

How is political interest associated with news exposure on websites and on Facebook?

# Measures

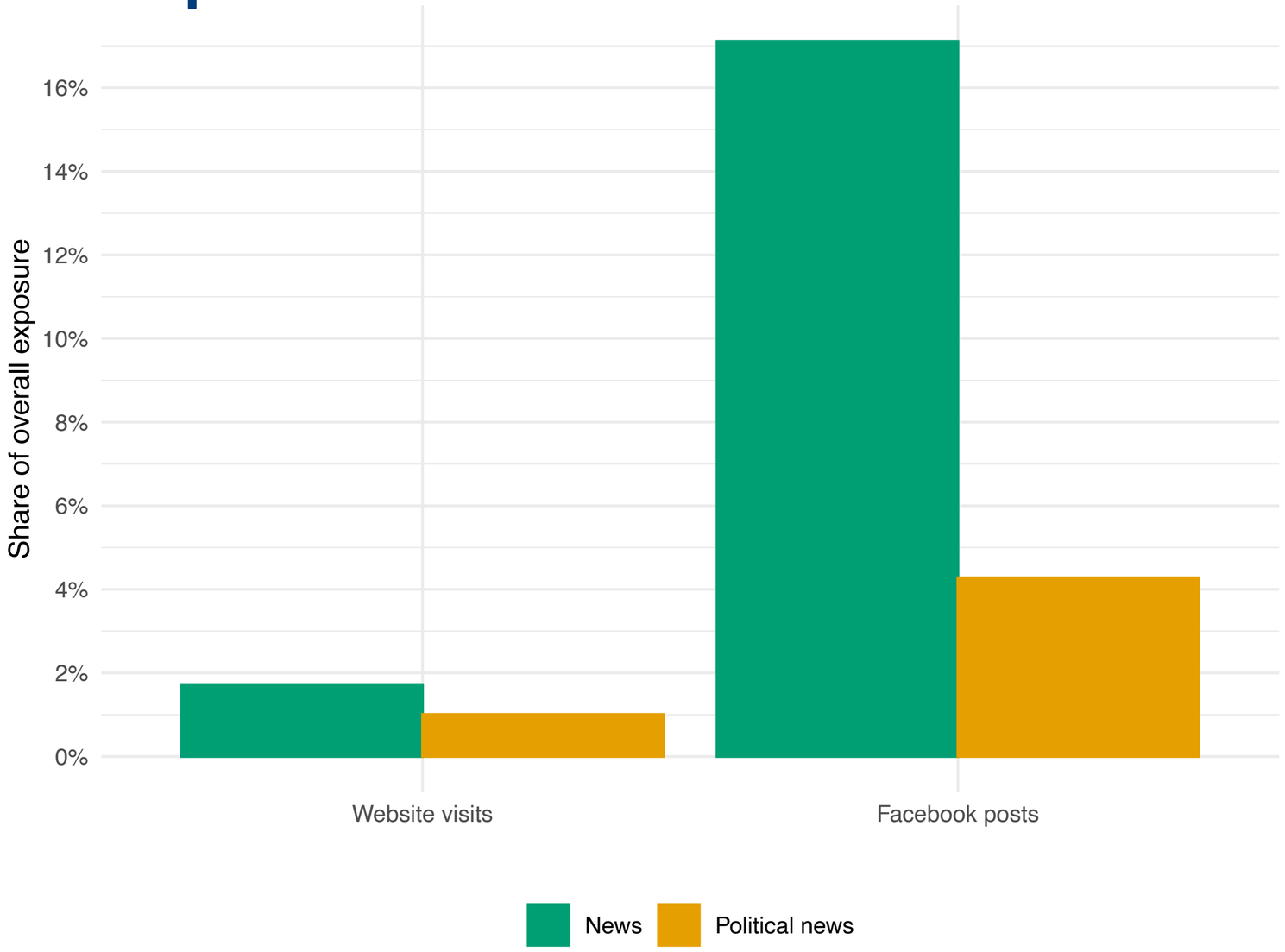
## Behavioral measures

- 266 news domains and their Facebook accounts
- Dictionary to identify political content ($F1_{Facebook} = 0.87$, $F1_{Websites} = 0.85$)
- Dependent variables:
  1. (Political) News website visits
  2. (Political) Facebook news posts seen

## Survey-based measures

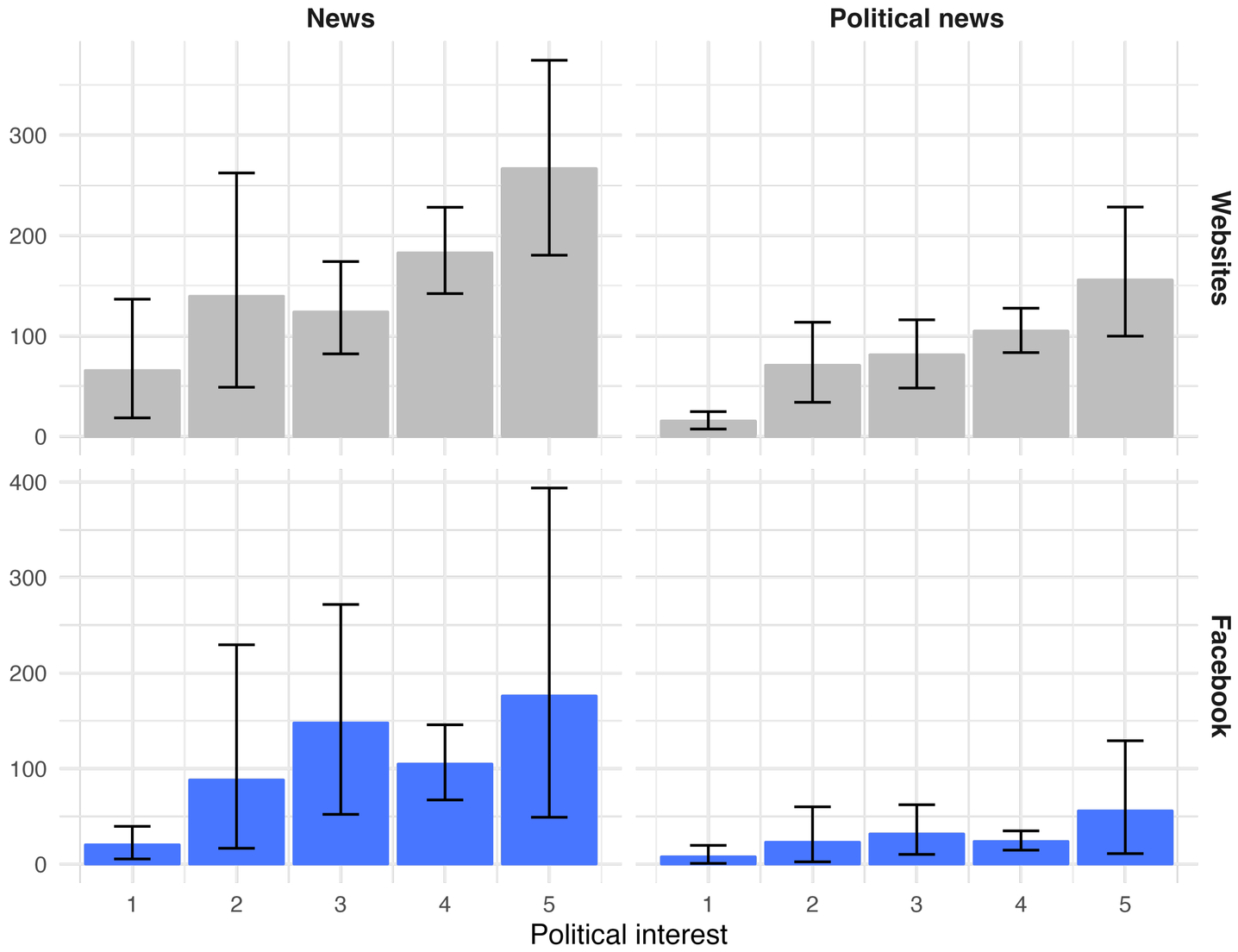
- Independent variable: political interest
- Control variables: gender, age, education, East/West German, political ideology (left/right), political extremism

# Descriptive results



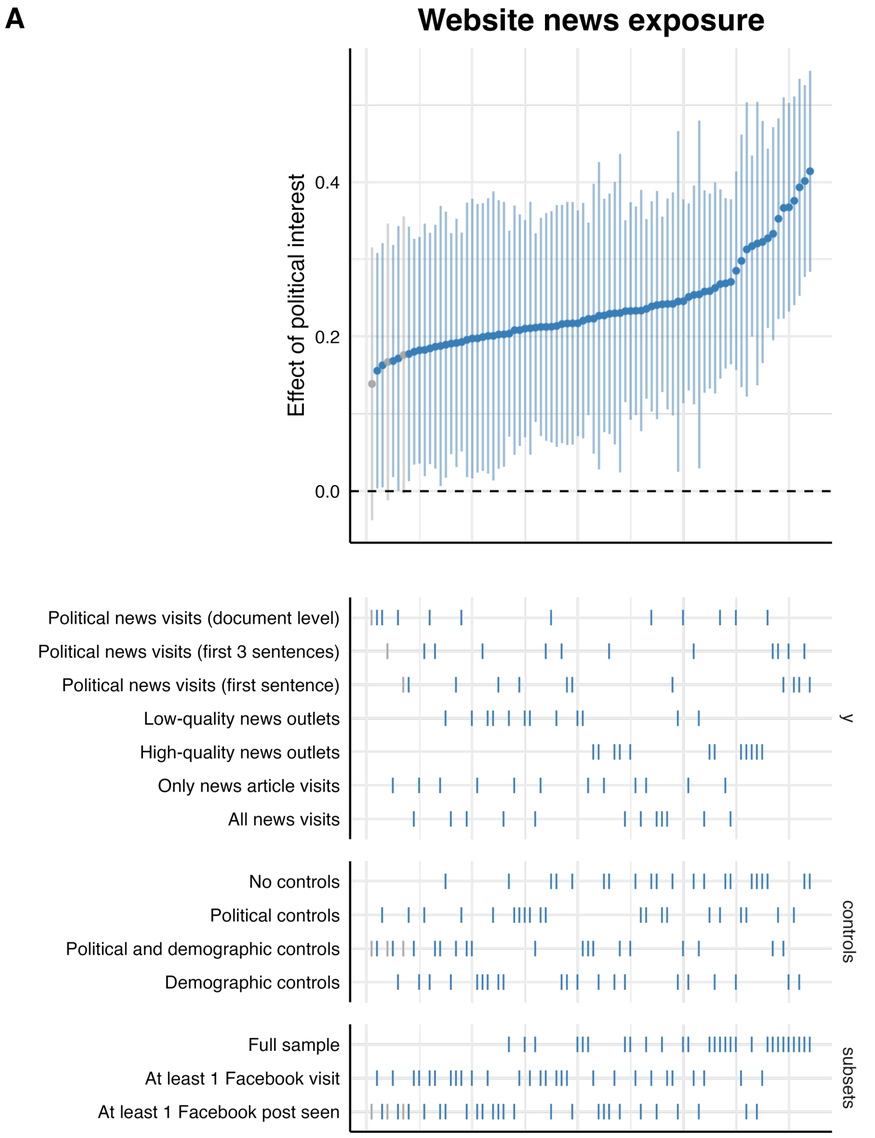

n=490 participants with at least one Facebook visit.

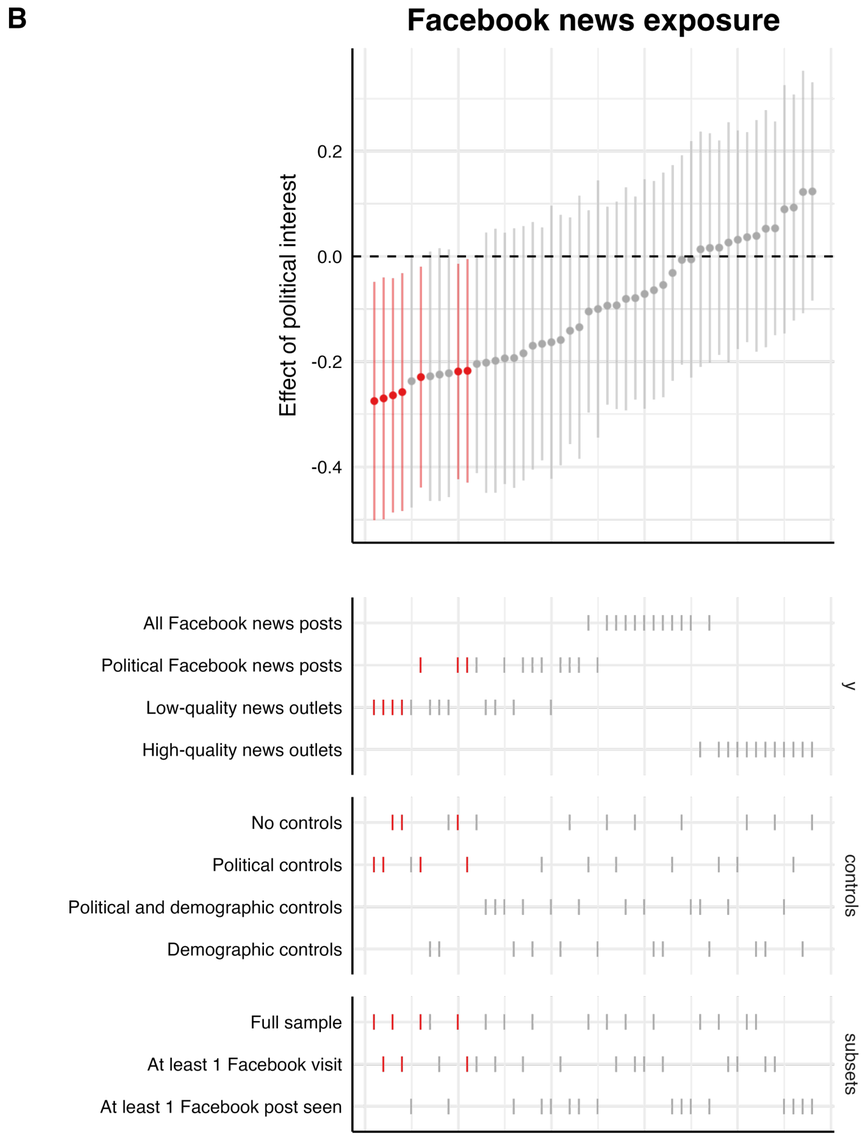
# Bivariate results for political interest




n=488 participants with at least one Facebook visit.

# Multiverse results for political interest




| Dimension | Specifications |
|---|---|
| News quality | • All news<br>• Low-quality news outlets<br>• High-quality news outlets: legacy press & public broadcasting |
| Political content | • Political content (document level)<br>• Political content first sentence<br>• Political content first 3 sentences |
| News main pages | • Main pages included<br>• Main pages not included (only news article visits) |
| Control variables | • Demographic controls: age, male, education, Eastern German<br>• Political controls: political ideology (left-right), political extremism<br>• Combinations of control variables |
| Sample | • Full sample (N=739)<br>• At least one Facebook visit (N=490)<br>• At least one public Facebook post seen (N=327) |

GESIS Leibniz Institute for the Social Sciences

# Major challenges in this space and infrastructure work at GESIS

Leibniz Association

# Challenge 1: Participant recruitment and management

GESIS Panel.dbd

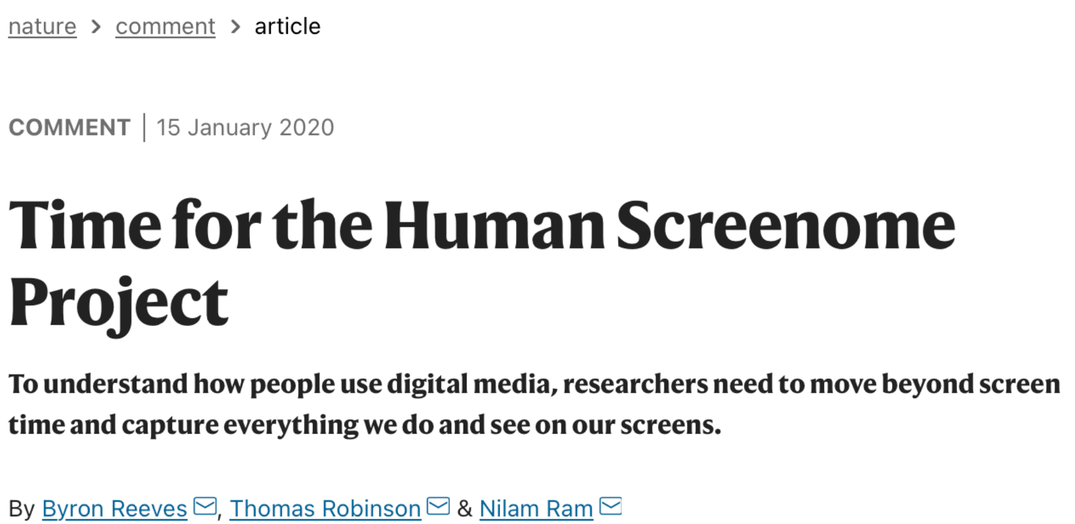
**Design**

- Probability-based and nonprobability-based recruitment arms
- ~6,000 active panelists
- CAWI
- 4-5 survey waves per year
- Digital behavioral data is collected through
  - GESIS Web Tracking (continuous)
  - GESIS AppKit (project specific)
  - Data donations (first pilot studies)

**Recruited via**

# Challenge 2: Tracking mobile content

- GESIS AppKit as a mobile research app
  - Mobile experience sampling
  - Sensor data coming soon
- Tracking content on smartphones is difficult
- Most promising approach: screen capturing/scraping
- Progress by Human Screenome Project, NIO, commercial players like Murmuras



nature › comment › article

COMMENT | 15 January 2020

## Time for the Human Screenome Project

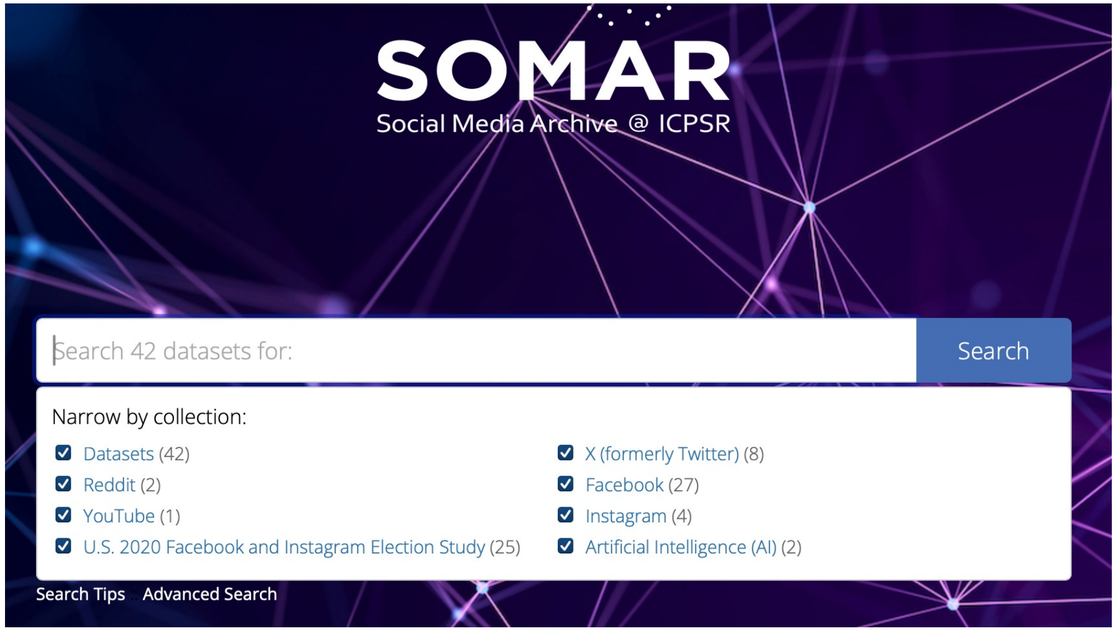
**To understand how people use digital media, researchers need to move beyond screen time and capture everything we do and see on our screens.**

By Byron Reeves, Thomas Robinson & Nilam Ram

# Challenge 3: Archiving and secondary use

**Digital behavioral data poses challenges for archiving**

- (Oftentimes) proprietary
- Contains sensitive personal information: compliance with GDPR?
- Copyright
- Technical challenges (virtual secure access / trusted virtual research environments)




**Upsides**

- We can get consent for reuse of data in user-centered data collections
- Digital Services Act might bring further legal clarity

# Thank you for your attention!

https://sebastianstier.com

@SebStier

Leibniz
Association