Follow the User?!

Data Donation Studies for Collecting Digital Trace Data

Session 4 : Bias & Outro

Frieder Rodewald (University of Mannheim) & Valerie Hase (LMU Munich)

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

Agenda

1. Bias in Data Donation Studies
2. What's Next for Data Donation?
3. Outro

Image by Hope House Press via Unsplash

1) Bias in Data Donation Studies (Valerie)

Source: Image by Markus Winkler via Unsplash

What is bias?

Definition 💡: *Deviations from the true value of a theoretical concept introduced by its measurement* (Peytchev, 2013)

- Non-systematic errors: random deviations influence variance of estimates
- Systematic errors (or: **bias**): depend on omitted variables

👉 Bias can influence descriptive results but also attenuate/inflate inferential conclusions.

What is bias?

- Errors in representation: *Who participates in data donation studies?*
- Errors in measurement: *Which latent concepts can we measure with data donation studies?*

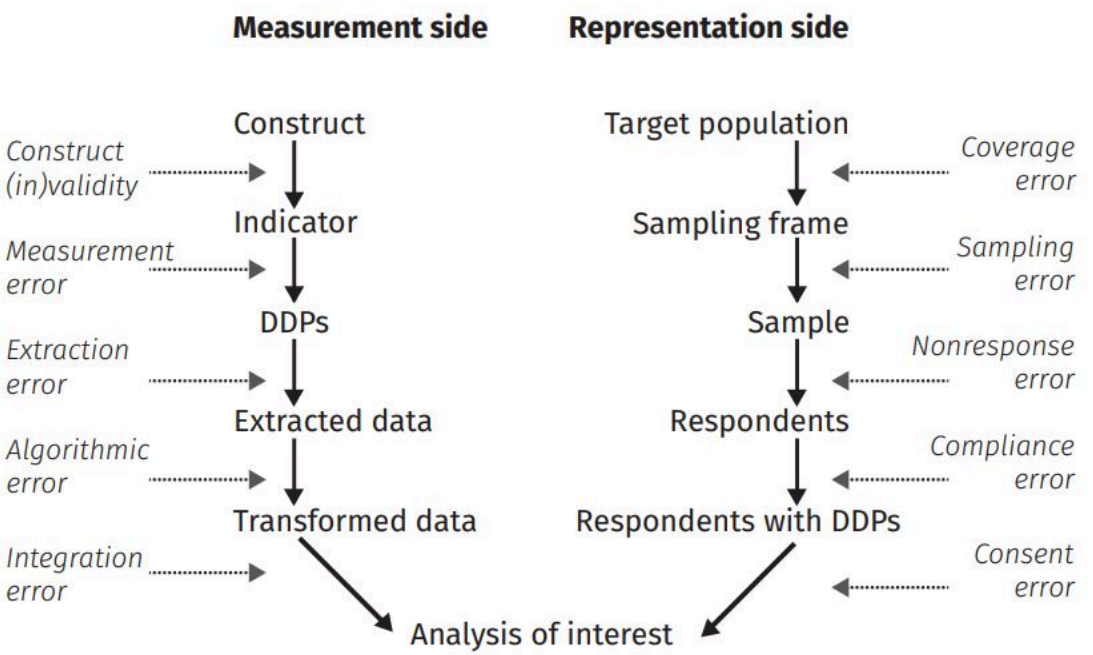


Figure 2. "Total error framework" for social-scientific data collection with DDPs. Each step in the data collection process is shown, together with the errors resulting from this step. Subsequent processing, modeling, and inference steps (Amaya et al., 2020) are omitted.

Source: Image from Boeschoten et al., 2022, p. 396

Errors in representation

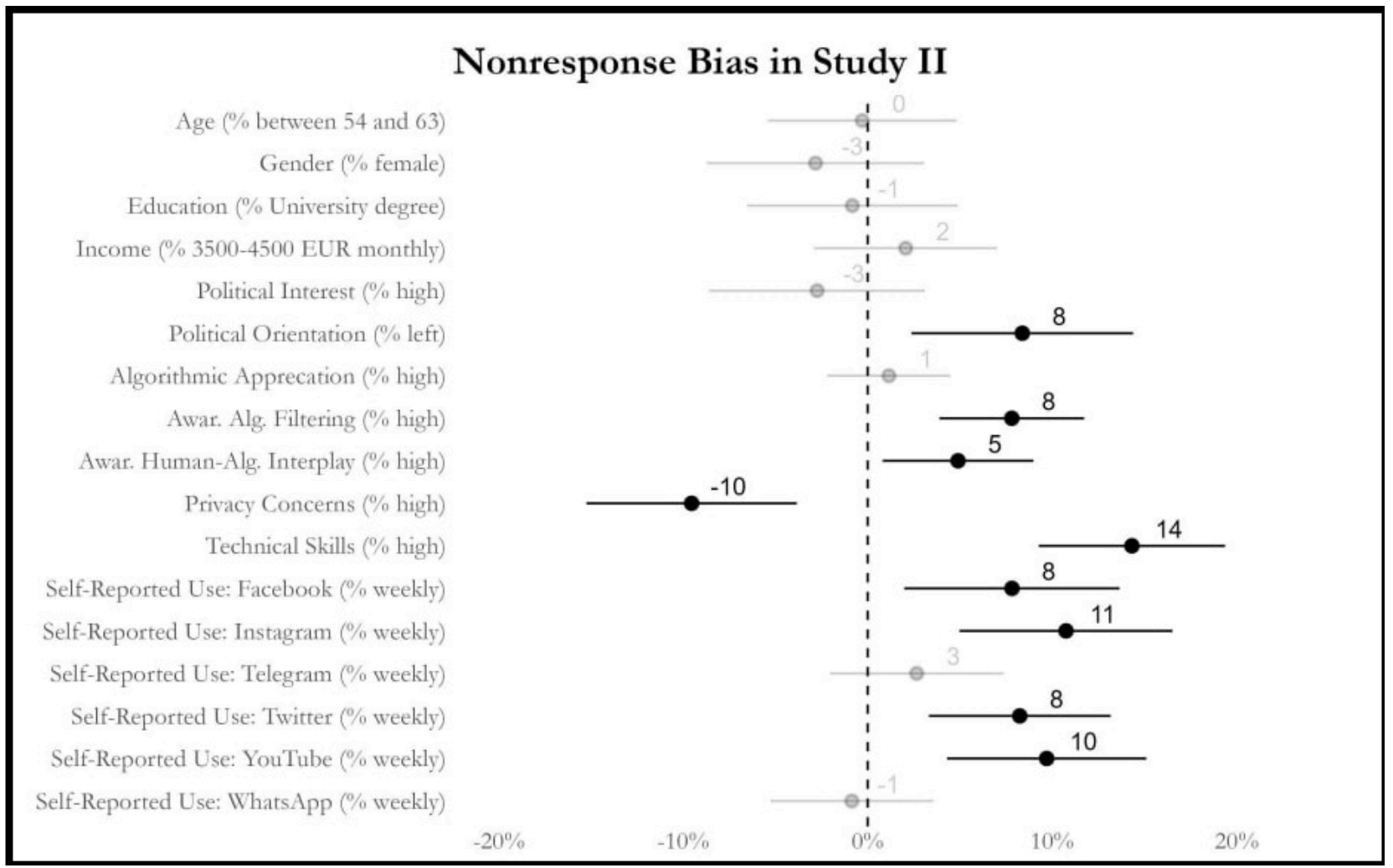
For example ...

- **Coverage error**: Who is (not) represented in the sampling frame? (e.g., social media users vs. YouTube users)
- **Sampling error**: Who is (not) represented in the sample? (e.g., non-probability samples)
- **Non-response error**: Who does (not) want to participate in the data donation?
- **Compliance error**: Who is (not) able to participate in the data donation?

What do you think: Which participant characteristics may correlate with non-response or non-compliance? 🤔

Errors in representation

Example study by Hase & Haim (2024):



Source: Figure from Hase & Haim (2024)

Any ideas (from your discipline): How can we quantify/address errors in representation? 🤔

Errors in representation: Quantification

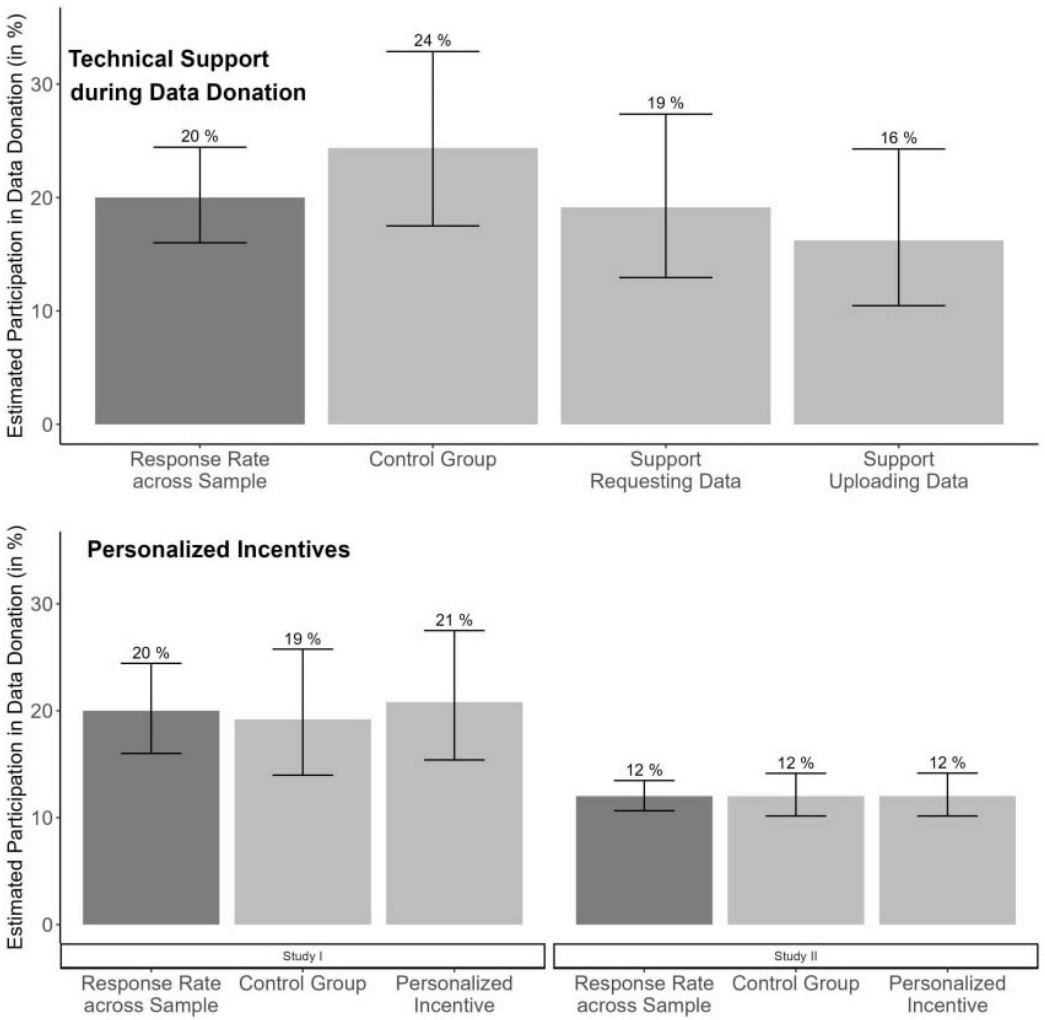
- Response rates across study stages
- Para data as quality indicators (e.g., speeding)
- Non-response bias (e.g. traits of survey vs. donation participants)

Errors in representation: Solutions

- **A posteriori strategies**:
  - Infrastructure: Integration in probability-based panels
  - Survey design strategies (e.g., incentives, study framing)
  - DDT design (e.g. UX-perspective)
- **Post hoc strategies**:
  - Statistical modeling (e.g., weighting, see Pak et al., 2022)

Errors in representation: Solutions

For now: limited studies, limited success of existing solutions



Source: Figure from Hase & Haim (2024)

What do you think: How could errors in measurements sneak into data donation studies? 🤔

Errors in measurement

For example ...

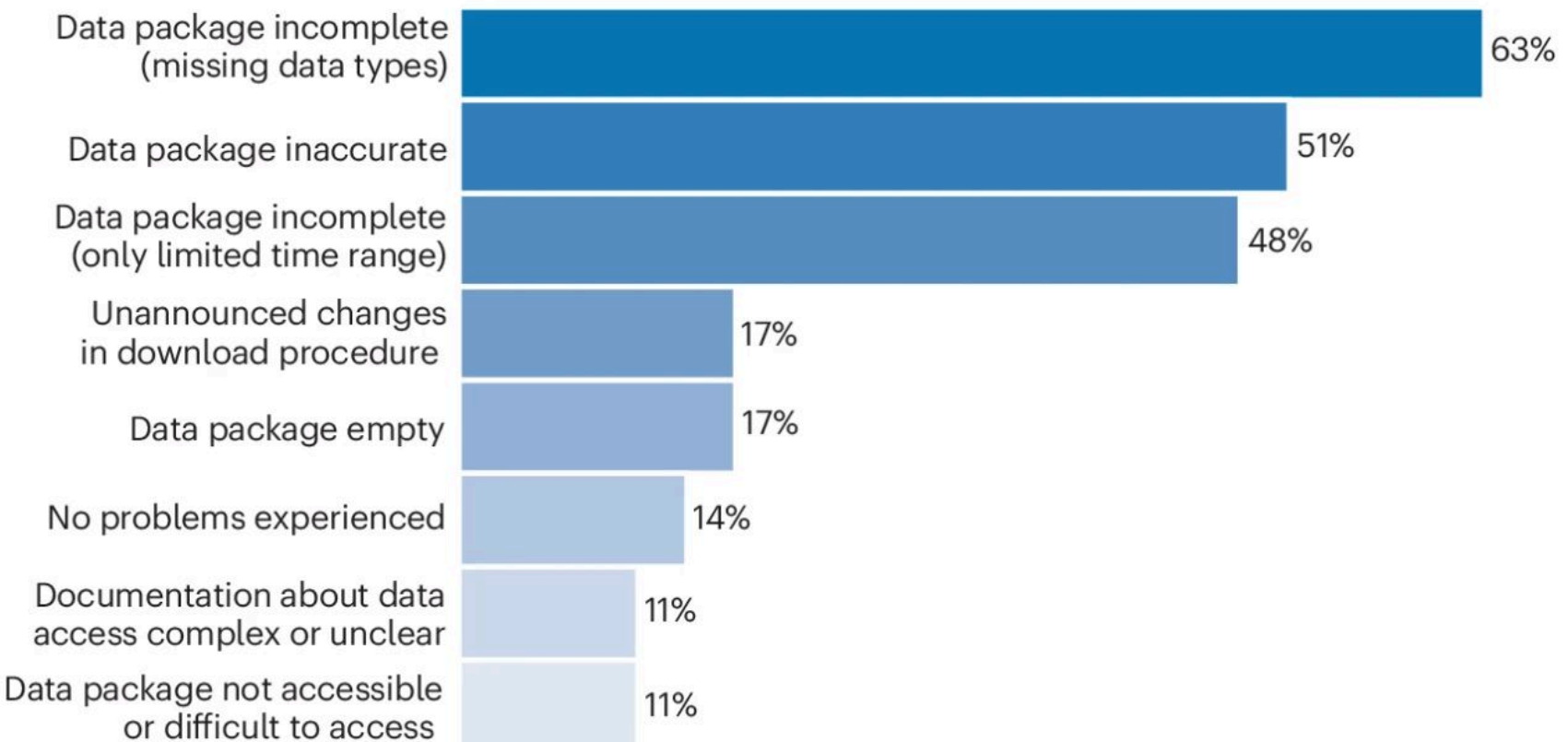
- **Construct (in-)validity**: How do DDP variables relate to latent measurements? (e.g., likes vs. political participation)
- **Measurement error**: How correct is data in our DDP? (e.g., missing data)
- **Extraction error**: Did we extract all relevant files and variables?

Errors in measurements

Example study by Valkenburg et al. (2024):

Fig. 1: Common problems in platform data donations experienced by researchers.

From: It is time to ensure research access to platform data



Data from a June 2024 survey among 51 data donation researchers.

Source: Figure from Valkenburg et al. (2024)

Any ideas (from your discipline): How can we quantify/address errors in measurements? 🤔

Errors in measurement: Quantification

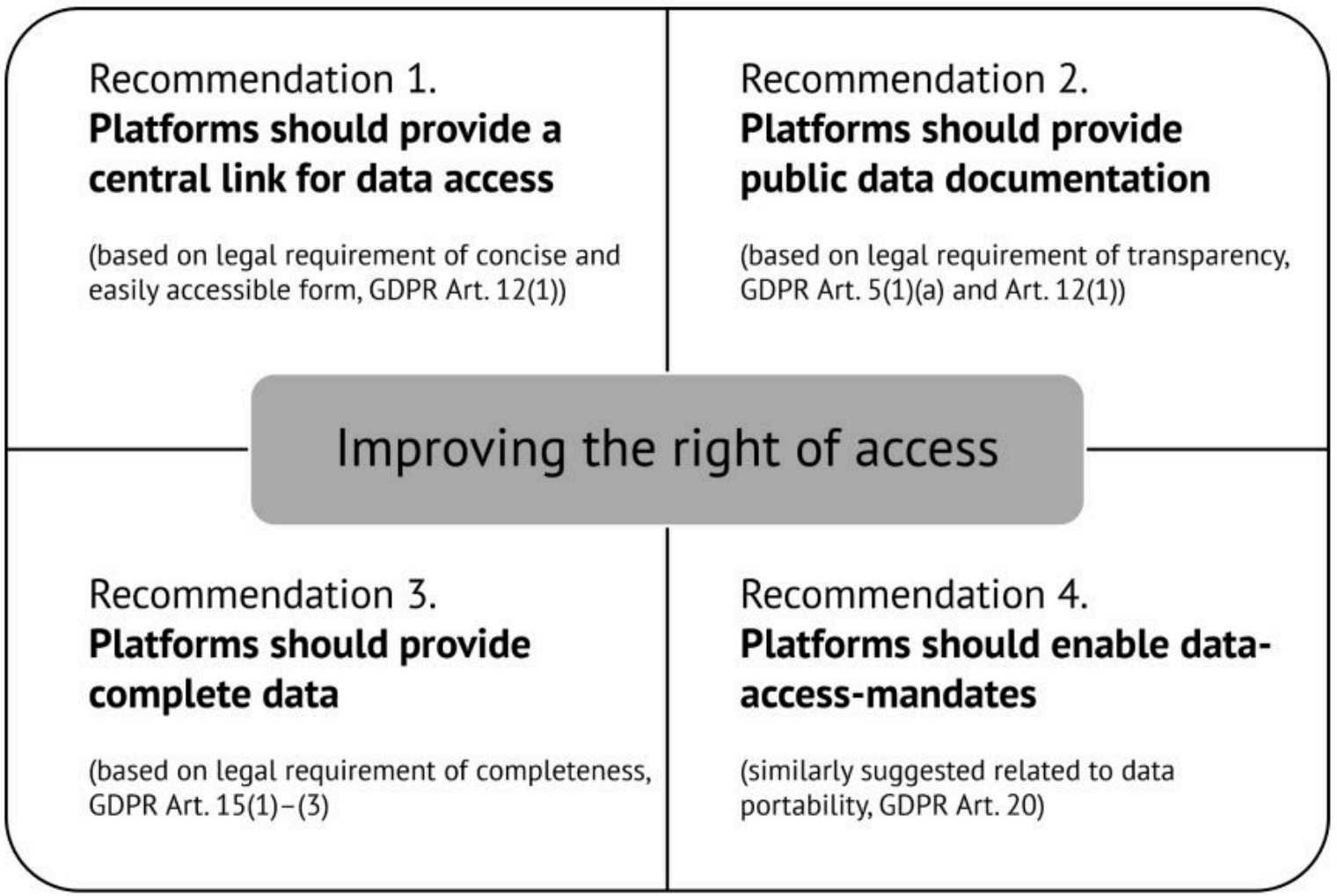
- Para data (e.g., failed uploads)
- Correlation between self-reported and observed behavior
- Multi Trait Multi Method (MTMM) approaches (Cernat et al., 2024)
- Estimation of misclassification effects (TeBlunthuis et al., 2024)

Errors in representation: Solutions

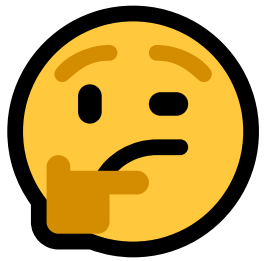
- **A posteriori strategies**:
  - Talk to everyone (e.g., IRB, Data Strward)
  - Repeated testing & DDP download
  - Simulate downstream errors (Bosch et al., 2024)
- **Post hoc strategies**:
  - Multiverse approaches
  - Statistical error correction (TeBlunthuis et al., 2024)
  - Error documentation (Gebru et al., 2021)

Errors in representation: Solutions

Example study by Hase et al. (2024):



Source: Figure from Hase et al. 2024

Questions? 

2) What's next for data donation studies? (Frieder)

Source: Image by Markus Winkler via Unsplash

The road ahead I: Advancing the method

- Multimodal & cross-platform data 📸 (Wedel et al., 2024)
- In-tool, local classification (e.g., local SML/LLMs?)
- Workflow/UX-perspective

Source: Image by DariuszSankowski via Pixabay

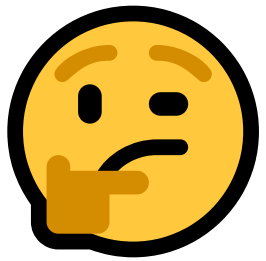
The road ahead II: Data as a political tool

- Platforms do (willingly?) not provide data according to the GDPR/DSA (Hase et al., 2024)
- The EU has started to sanction platforms like X/TikTok
- DSA may become the subject of larger geo-political debates with the USA (Seiling et al., 2025)

Source: Image by WilliamCho via Pixabay

The road ahead III: Can we improve & apply the method?

- Can the method actually be applied for empirical research? (few examples, like Thorson et al., 2021; Wojcieszak et al., 2024)
- Requires interdisciplinary perspectives (e.g., addressing bias, integration in probability-based panels)

Questions? 

3) Outro (Frieder)

Source: Image by Markus Winkler via Unsplash

We want your feedback! 🙃

👉 Please fill out this 3-minute feedback form: https://forms.gle/KLMweywhW7odGyfk8

QR code for survey

Thanks for joining the workshop 

Quellen

Bosch, O. J., Sturgis, P., Kuha, J., & Revilla, M. (2024). Uncovering Digital Trace Data Biases: Tracking Undercoverage in Web Tracking Data. *Communication Methods and Measures*, 1–21. https://doi.org/10.1080/19312458.2024.2393165

Cernat, A., Keusch, F., Bach, R. L., & Pankowska, P. K. (2024). Estimating Measurement Quality in Digital Trace Data and Surveys Using the MultiTrait MultiMethod Model. *Social Science Computer Review*, 08944393241254464. https://doi.org/10.1177/08944393241254464

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, *64*(12), 86–92. https://doi.org/10.1145/3458723

Hase, V., Ausloos, J., Boeschoten, L., Pfiffner, N., Janssen, H., Araujo, T., Carrière, T., De Vreese, C., Haßler, J., Loecherbach, F., Kmetty, Z., Möller, J., Ohme, J., Schmidbauer, E., Struminskaya, B., Trilling, D., Welbers, K., & Haim, M. (2024). Fulfilling Data Access Obligations: How Could (and Should) Platforms Facilitate Data Donation Studies? *Internet Policy Review*, *13*(3). https://doi.org/10.14763/2024.3.1793

Hase, V., & Haim, M. (2024). Can We Get Rid of Bias? Mitigating Systematic Error in Data Donation Studies through Survey Design Strategies. *Computational Communication Research*, *6*(2), 1. https://doi.org/10.5117/CCR2024.2.2.HASE

Pak, C., Cotter, K., & Thorson, K. (2022). Correcting Sample Selection Bias of Historical Digital Trace Data: Inverse Probability Weighting (IPW) and Type II Tobit Model. *Communication Methods and Measures*, *16*(2), 134–155.

https://doi.org/10.1080/19312458.2022.2037537

Peytchev, A. (2013). Consequences of survey nonresponse. *The ANNALS of the American Academy of Political and Social Science*, *645*(1), 88–111. https://doi.org/10.1177/0002716212461748

Seiling, L., Ohme, J., & De Vreese, C. (2025). Wird Europa den DSA in Verhandlungen mit Trump opfern? *Tagesspiegel*. https://background.tagesspiegel.de/digitalisierung-und-ki/briefing/wird-europa-den-dsa-in-verhandlungen-mit-trump-opfern

TeBlunthuis, N., Hase, V., & Chan, C.-H. (2024). Misclassification in Automated Content Analysis Causes Bias in Regression. Can We Fix It? Yes We Can! *Communication Methods and Measures*, *18*(3), 278–299. https://doi.org/10.1080/19312458.2023.2293713

Thorson, K., Cotter, K., Medeiros, M., & Pak, C. (2021). Algorithmic inference, political interest, and exposure to news and politics on Facebook. *Information, Communication & Society*, *24*(2), 183–200. https://doi.org/10.1080/1369118X.2019.1642934

Valkenburg, P. M., Van Der Wal, A., Siebers, T., Beyens, I., Boeschoten, L., & Araujo, T. (2024). It is time to ensure research access to platform data. *Nature Human Behaviour*, *9*(1), 1–2. https://doi.org/10.1038/s41562-024-02066-5

Wedel, L., Ohme, J., & Araujo, T. (2024). Augmenting Data Download Packages – Integrating Data Donations, Video Metadata, and the Multimodal Nature of Audio-visual Content. *Methods, Data, Analyses (Online First)*, 32 Pages. https://doi.org/10.12758/MDA.2024.08

Wojcieszak, M., Menchen-Trevino, E., Clemm Von Hohenberg, B., De Leeuw, S., Gonçalves, J., Davidson, S., & Gonçalves, A. (2024). Non-News Websites Expose People to More Political Content Than News Websites: Evidence from Browsing Data in Three Countries. *Political Communication*, *41*(1), 129–151.