Quality of Digital Behavioral Data

Florian Keusch
University of Mannheim

✉f.keusch@uni-mannheim.de
@floriankeusch

FORS-SSP
Methods & Research Meeting
Lausanne, Jan 31, 2024

Acknowledgments

- Collaborative research with Ruben Bach (University of Mannheim), Alexandru Cernat (University of Manchester), and Paulina Pankowska (Utrecht University)

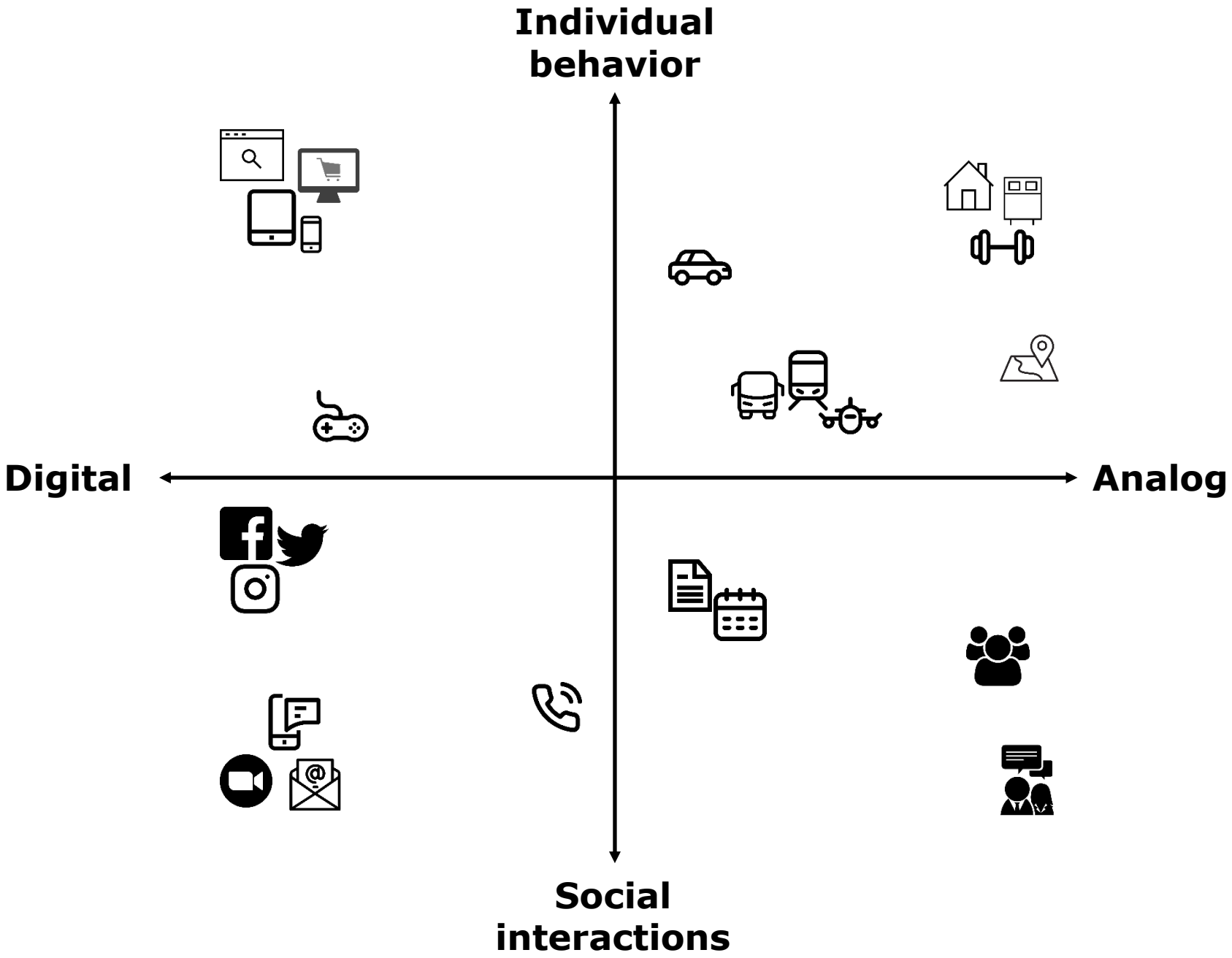
- Funding for this work comes from the *Baden-Württemberg Foundation* through the grant "Filter Bubbles, Alternative News and Political Polarization" (PI: R. Bach) and the *German Science Foundation (DFG)* through the grant "Evaluating Data Sources for Research into Political Reforms: (Non)probability Online Surveys and Big Data" (PIs: A. Blom, F. Keusch, & F. Kreuter).

- We thank Ella Häcker and Leonard Bek for research assistance.

What Are Digital Behavioral Data?

"Records of activity (trace data) undertaken through an online information system (thus, digital)"
(Howison et al. 2011:769)

"Behavioral residue [individuals leave] when they interact online"
(Hinds & Joinson 2018:2)



Keusch, F. & Kreuter, F. (2022). Digital trace data. Modes of data collection, applications, and errors at a glance. In Engel, U. et al. (Eds.) *Handbook of Computational Social Science. Volume 1: Theory, Case Studies and Ethics*, 100-118. Milton Park: Routledge. 10.4324/9781003024583-8

Benefits of Digital Behavioral Data

- Allow measurement of behavior (in-the-moment) at high frequency
  - Evaluation of moment-to-moment changes
  - Without increasing burden on participants
  - Scalable

- Measurement is nonreactive, i.e., without direct solicitation of subject studied
  - Data should be unaffected by measurement itself
  - Reduced measurement error when measuring smartphone use (Kobayashi & Boase 2012; Boase & Ling 2013; Andrews, et al. 2015; De Reuver & Bouwman 2015; Revilla et al. 2017; Jones-Jang et al. 2020), online media consumption (Araujo et al. 2017; Barthel et al. 2020; Haenschen 2020; Junco 2013; Keusch et al. 2022; Revilla, et al. 2017; Scharkow 2016), and mobility (Stopher et al. 2007; Scherpenzeel 2017)

How to Collect Individual-Level Digital Traces

- Use APIs

Meta

An Update on Our Plans to Restrict Data Access on Facebook

April 4, 2018

Chesnot / Getty Images

SCIENCE / TWITTER / TECH

Twitter just closed the book on academic research / Twitter was once an indispensable resource for academic research. That's changed under Elon Musk.

By Justine Calma, a science reporter covering the environment, climate, and energy with a decade of experience. She is also the host of the Hell or High Water podcast.

May 31, 2023, 3:19 PM GMT+2 | 23 Comments / 23 New

https://www.theatlantic.com/technology/archive/2023/02/elon-musk-twitter-ethics-algorithm-biases/673110/

FEATURE

Reddit pricing: API charge explained

Providing a free API was becoming costly for Reddit. Enterprise-scale developers now have to pay for access to Reddit's data.

By Ben Lutkevich, Technical Features Writer Published: 11 Jul 2023

Shutterstock

How to Collect Individual-Level Digital Traces

- Use APIs
- Collaborate with industry

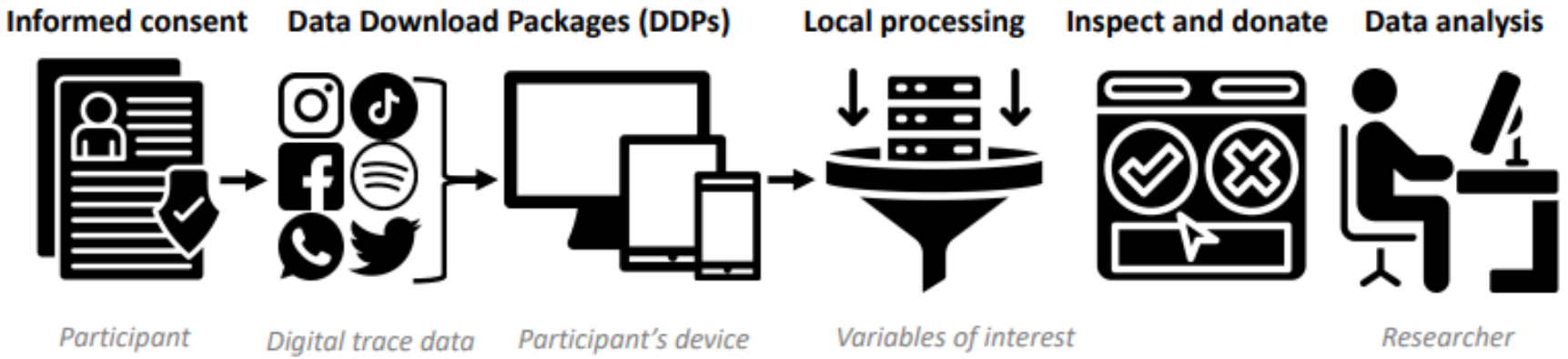
How to Collect Individual-Level Digital Traces

- Use APIs
- Collaborate with industry
- Have users install meters and apps that continuously…
  - …track information on web browsing
  - …logs usage behavior, (native) mobile browsing, and sensor readings

- Allows tracking of individual behavior over longer periods of time
- Various commercial and non-commercial tools available
  - Wakoopa, Movisense, Mumuras, Ethica, Aware, Beiwe, ResearchStack, ResearchKit, umlaut (formerly P3), …
- Some market research companies maintain "metered panels"
  - Netquest, Respondi/Bilendi, Gapfish, Dynata (U.S. only), …

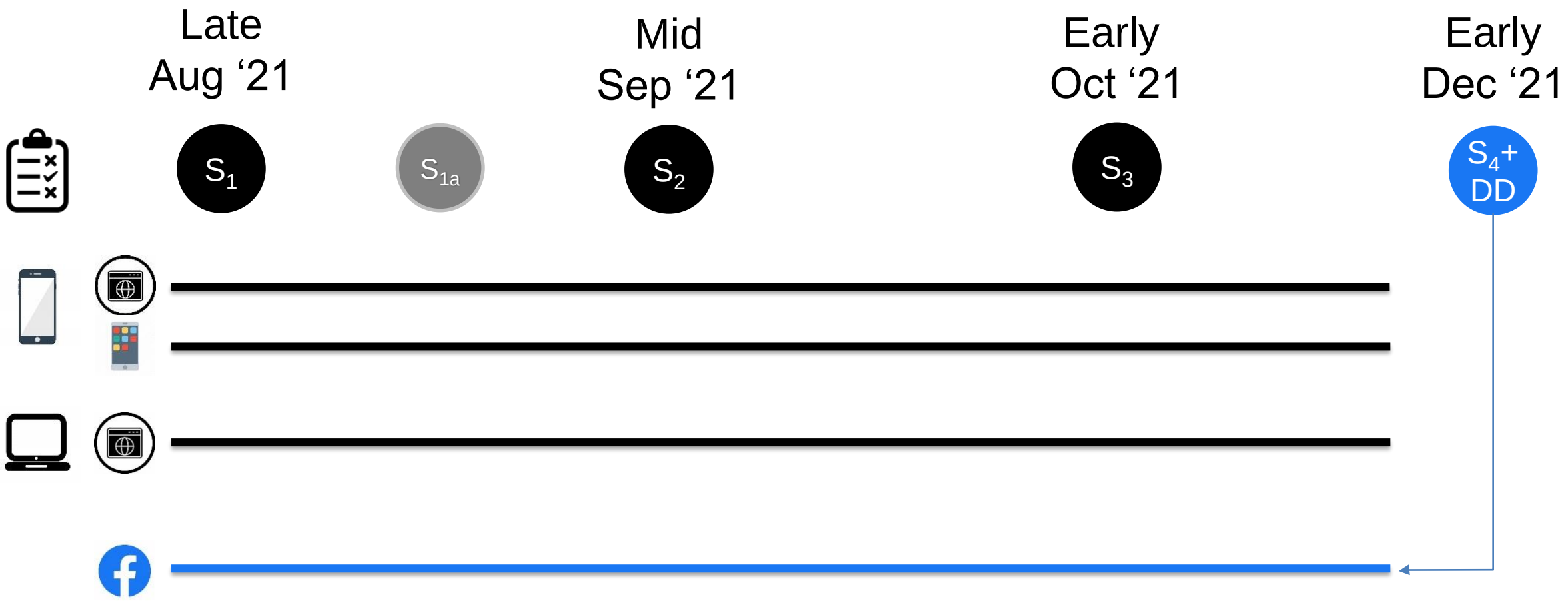
How to Collect Individual-Level Digital Traces

- Use APIs
- Collaborate with industry
- Have users install meters and apps
- Ask users to donate data
  - Takes advantage of GDPR Articles 15 (*Right of access by the data subject*) and 20 (*Right to data portability*)
  - Privacy-preserving data donation platforms



(Boeschoten et al. 2023)

Understand Quality of Digital Behavioral Data from Meters and Data Donation vis-á-vis Self-reports



N=2,100 members of German non-probability online panel

Study 1

Estimating Measurement Quality in Digital Behavioral Data and Surveys Using the MultiTrait MultiMethod Model

Cernat, A., Keusch, F., Bach, R.L., & Pankowska, P.K. (R&R). Estimating measurement quality in digital trace data and surveys using the multitrait multimethod model. *Social Science Computer Review.*

How Best to Measure Phone Usage Behavior?

Survey

- 5-point rating scale
- 7-point rating scale
- Duration (hours and minutes)

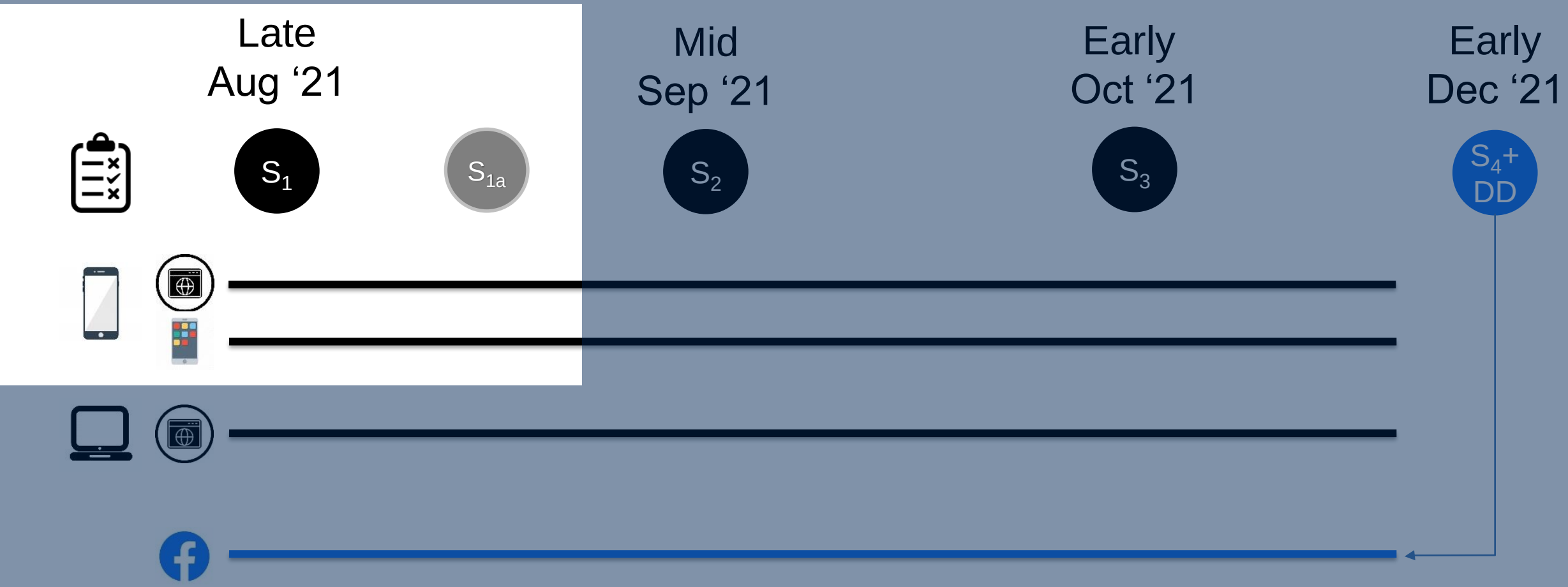
How often do you use your smartphone to [do activity]?
- Once a month or less
- Several times a month
- Several times a week
- Daily
- Several times a day

Tracking Data

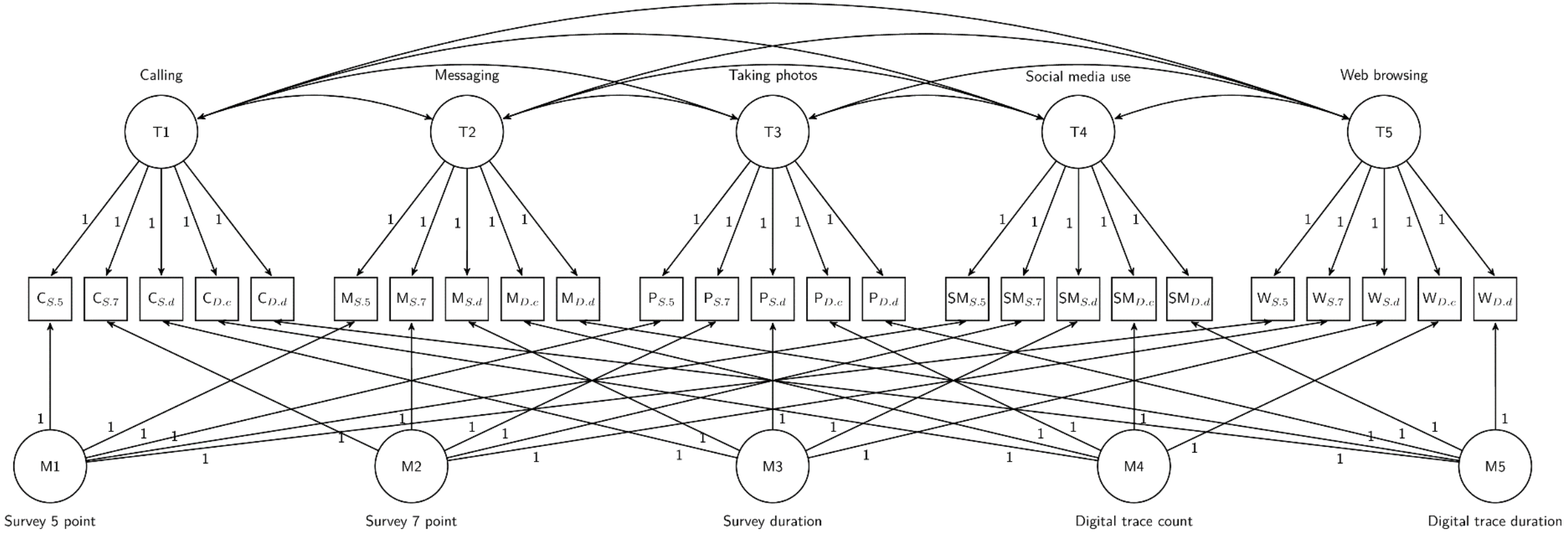
- No. of times activity is recorded
- Time spent on activity

How often do you use your smartphone to [do activity]?
- Less than once a month
- Once or twice a month
- Several times a month
- Once or twice a week
- Several times a week
- Once or twice a day
- Several times a day

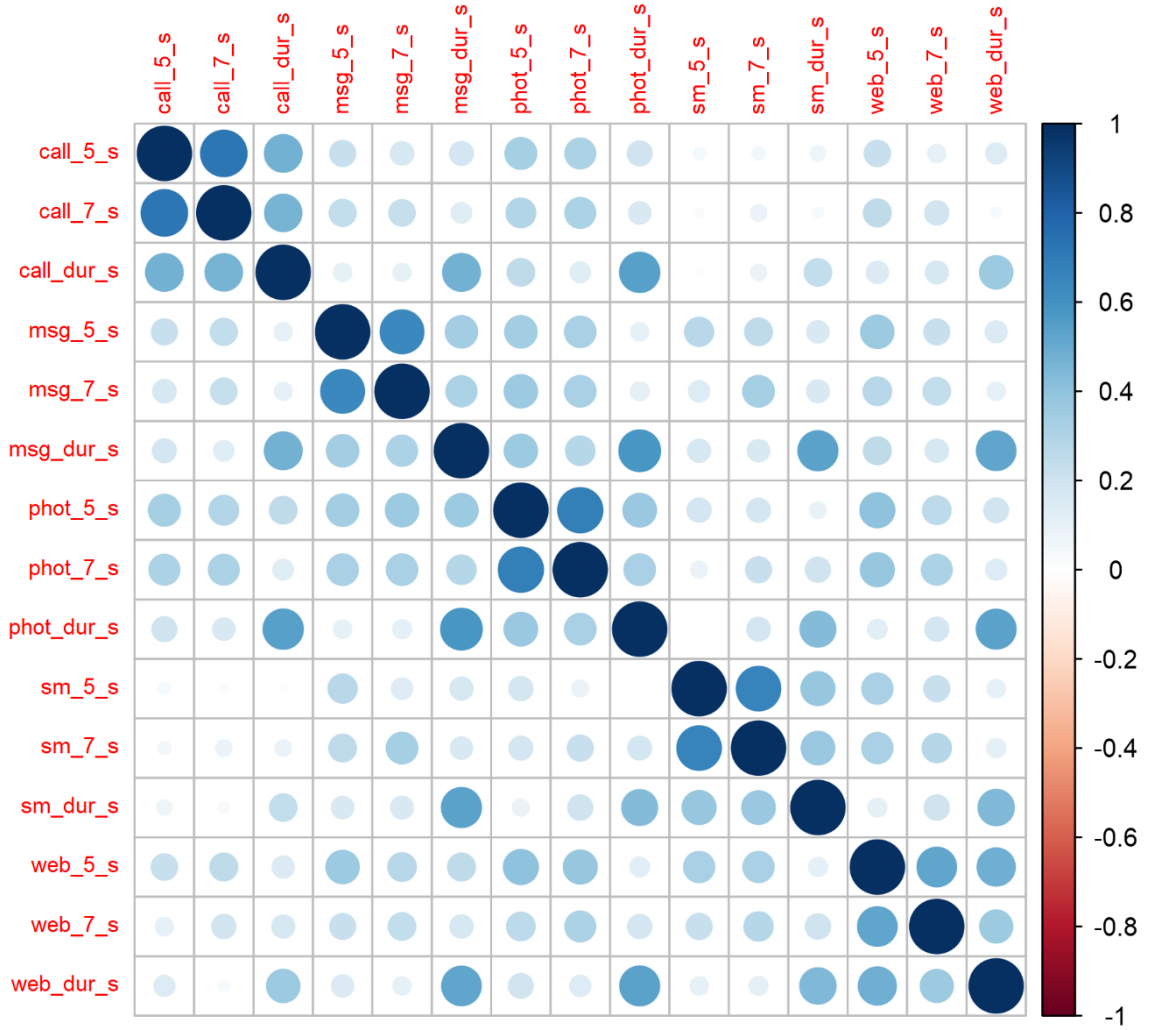
Design Study 1



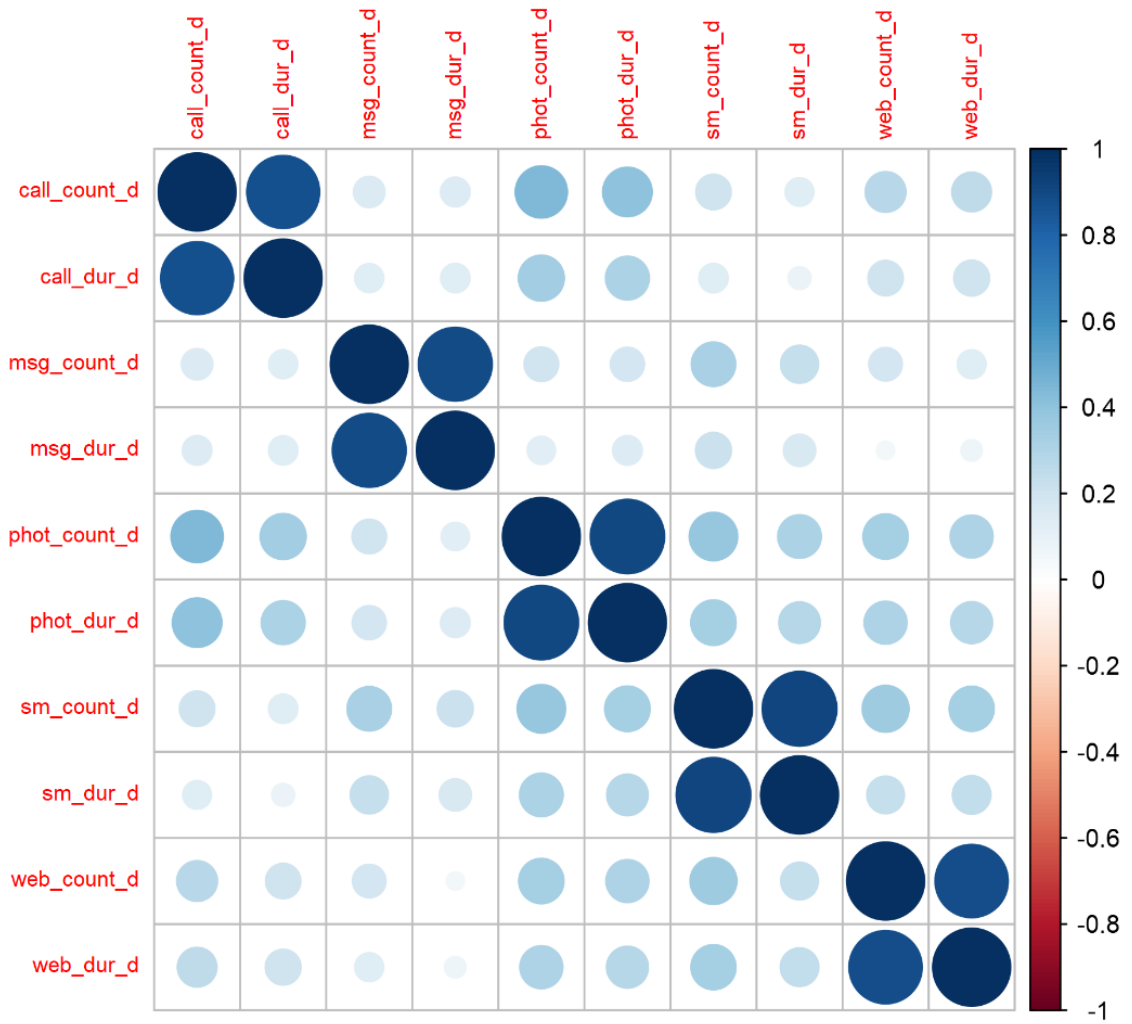
MTMM-Approach



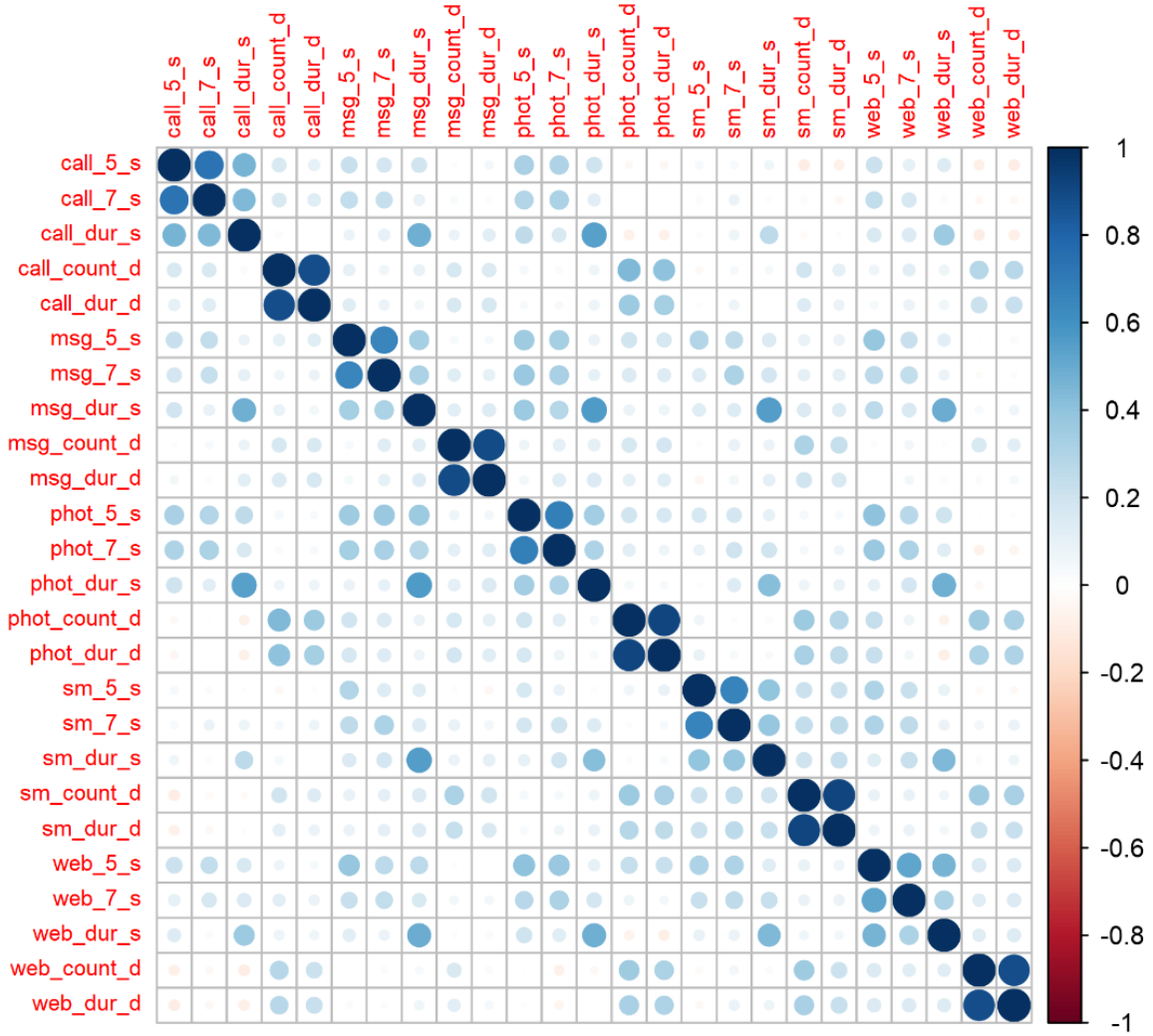
Correlation Matrix Survey



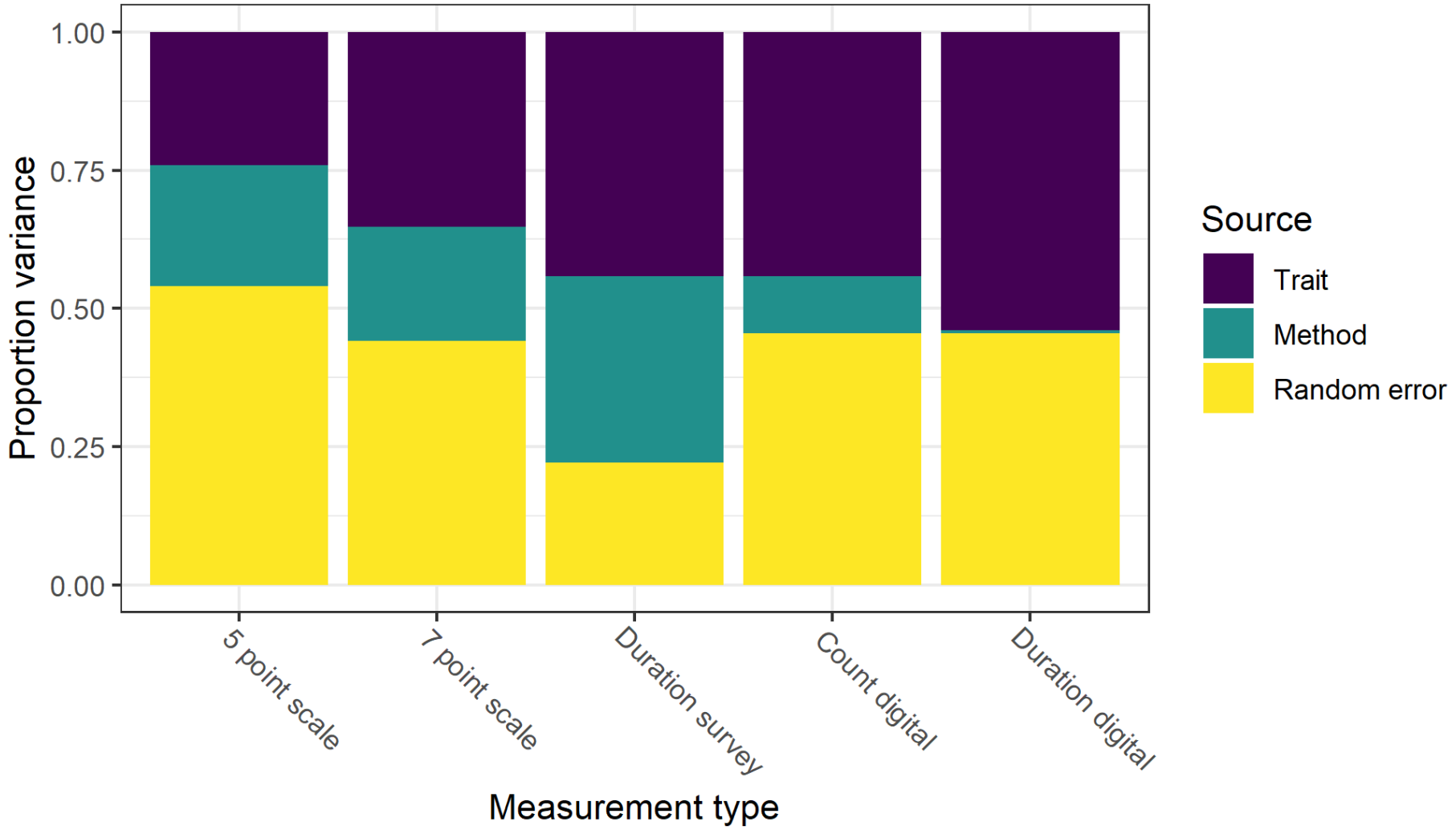
Correlation Matrix Tracking Data



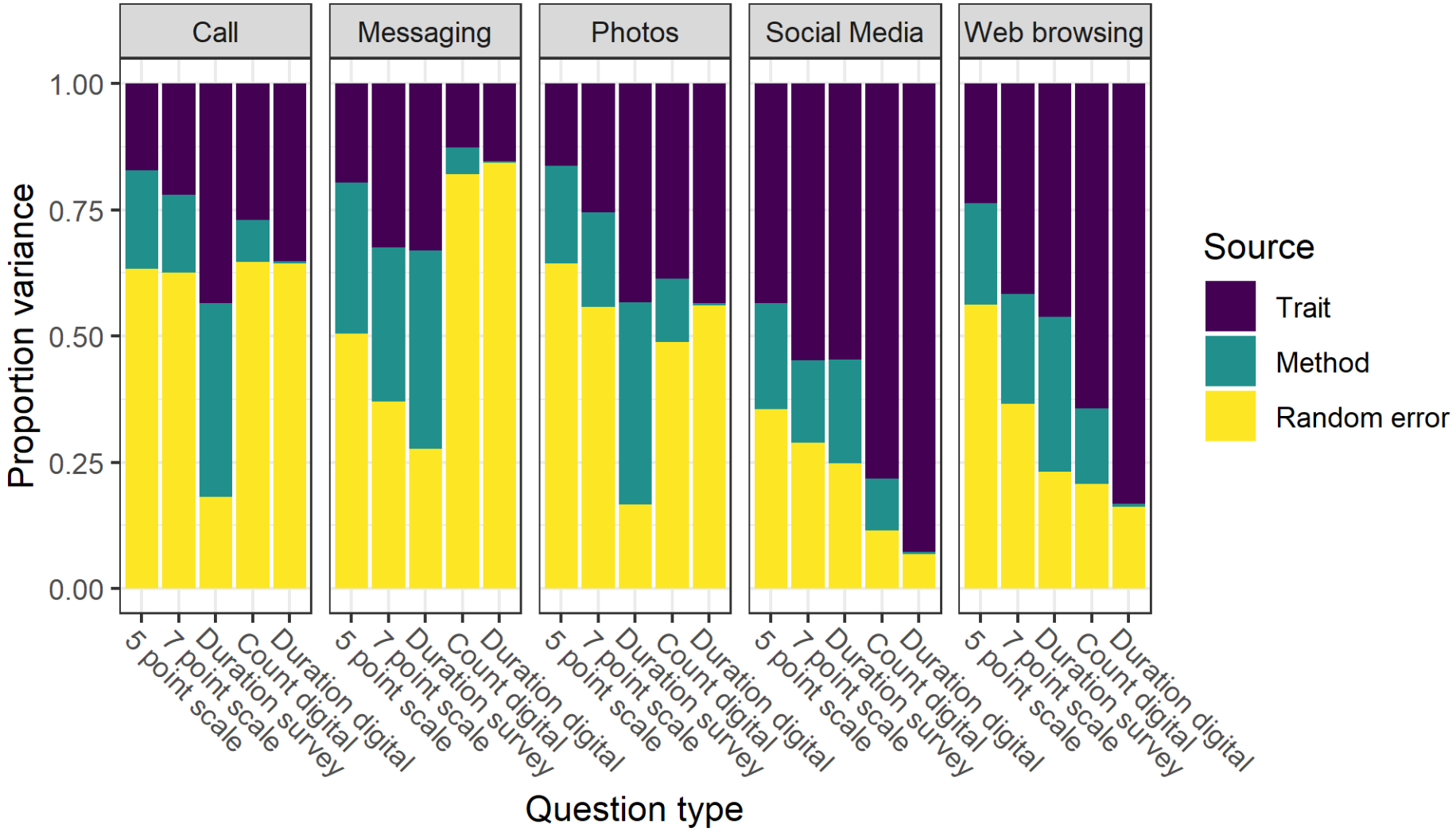
Full Correlation Matrix



MTMM Variance Decomposition – Method



Variance Decomposition – Method x Trait



Conclusion Study 1

- Digital behavioral data seems to measure smartphone activities better than survey
  - But far from perfect

- Problem of text messaging in meter data might stem from how app categories are defined

- Next steps: investigate impact on substantive results and combining measurements

Study 2

Measuring Facebook Use: The Accuracy of Self-reported Data Versus Digital Behavioral Data

Pankowska, P.,K. Cernat, A., Keusch, F., & Bach, R.L. (in preparation). Using hidden Markov models to assess and correct for measurement error in digital trace Data.

How Best to Measure Facebook Usage?

Survey

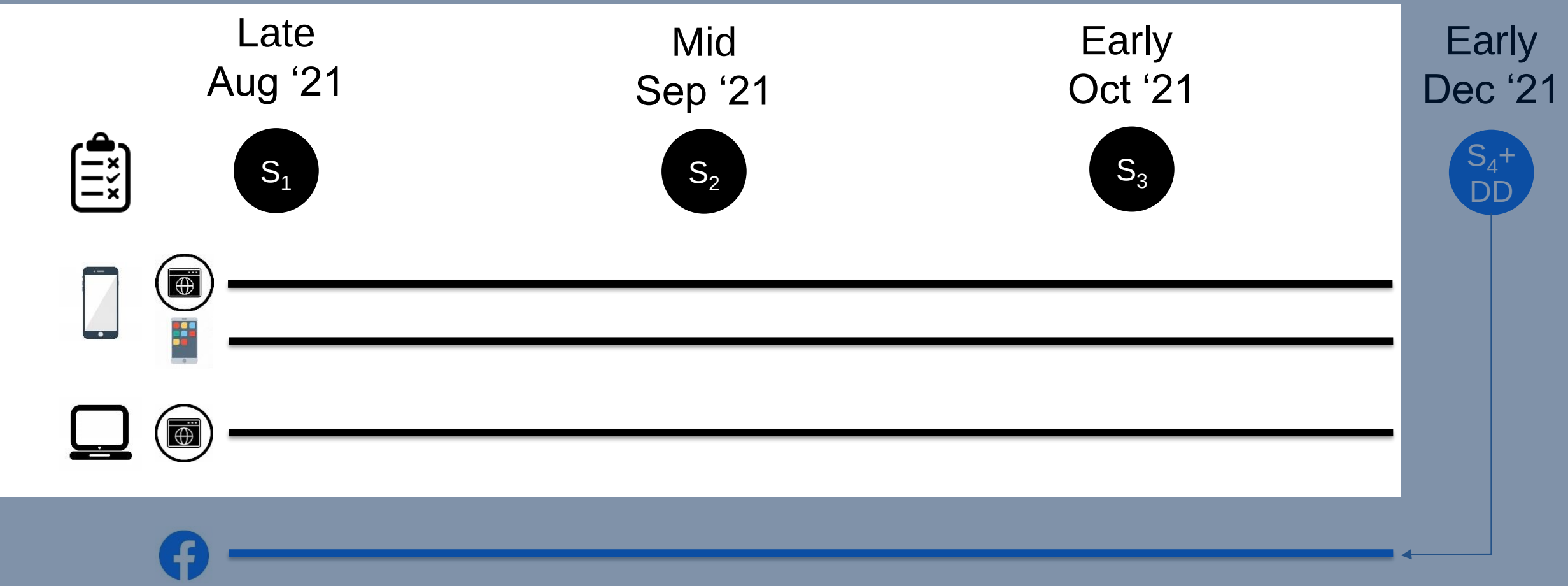
- 5-point rating scale

How often do you use Facebook?
- Once a month or less
- Several times a month
- Several times a week
- Daily
- Several times a day

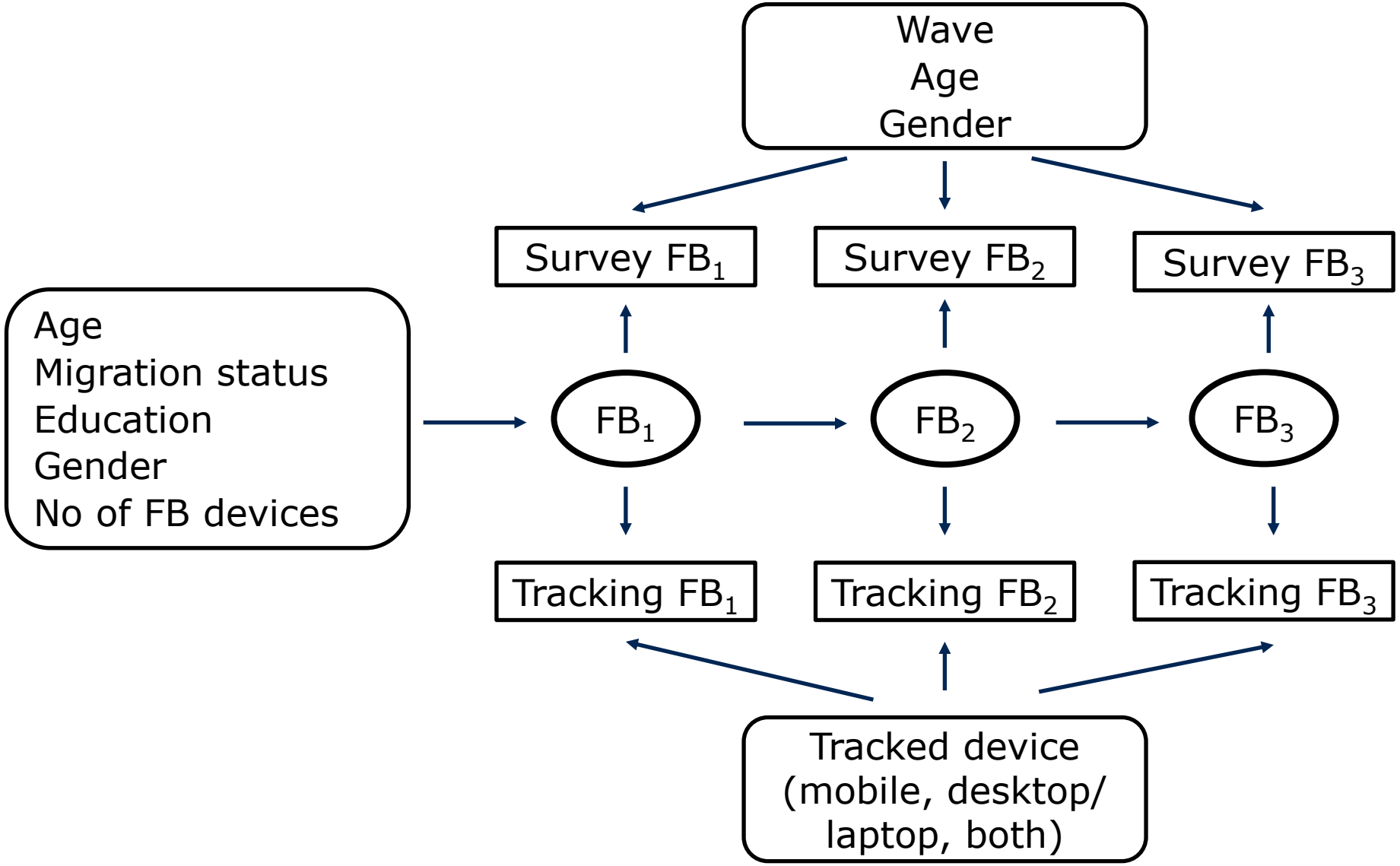
Tracking Data

- Three 10-day periods corresponding to survey waves
- Summing up number of times FB used on mobile and/or desktop/laptop during each period
- Categorizing usage variable in accordance with survey question

Design Study 2



Extended, Two-indicator HMM



Three Classe-Solution

- Both sources measure frequent and infrequent users relatively accurately
- One class characterized by large inconsistencies between the two sources
- For inconsistent class (C1)…
  - …number of devices tracked significant predictor of FB use in digital trace data
  - …probability of being infrequent user much higher if only desktop/laptop (rather than mobile & desktop/laptop) tracked

	Class		
	1	2	3
Size	0.38	0.33	0.29
FB use survey			
Once a month or less	0.00	0.00	0.30
Several times a month	0.01	0.02	0.31
Several times a week	0.10	0.13	0.37
Daily	0.46	0.35	0.02
Several times a day	0.44	0.50	0.00
FB use tracking app			
Once a month or less	0.52	0.01	0.61
Several times a month	0.13	0.02	0.14
Several times a week	0.23	0.17	0.20
Daily	0.10	0.00	0.05
Several times a day	0.01	0.80	0.01

Conclusion Study 2

- For certain groups of people, conclusions about FB usage between self-reports and digital behavioral data very much allign

- For about one third of users, self-reported FB use much higher than digital behavioral data, especially when information from mobile devices missing
  - Might suggest presence of systematic error: when not all devices tracked, FB use underestimated

- Digital behavioral data might reduce error of forgetting and social desirability, but missingness caused by other mechanisms

Study 3

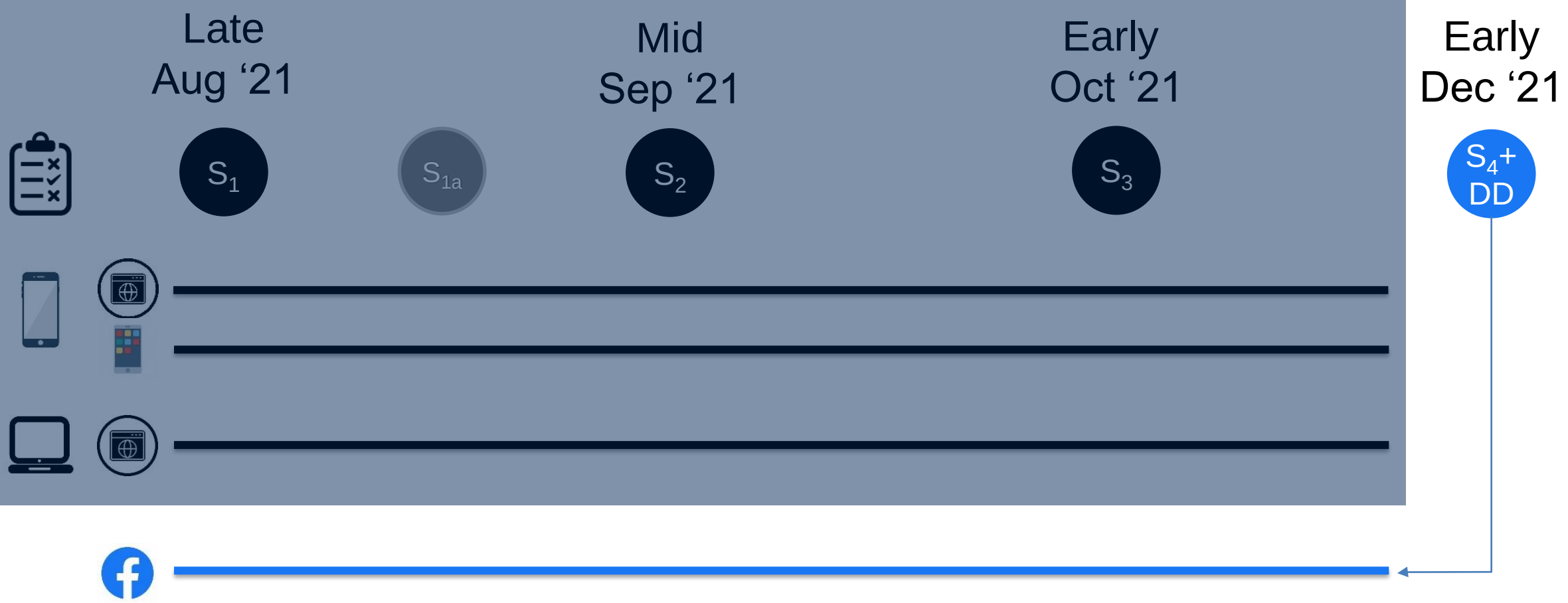
Do you have two minutes to talk about your data? Willingness to participate and nonparticipation bias in Facebook data donation

Keusch, F., Pankowska, P.K., Cernat, A. , & Bach, R.L. (2024). Do you have two minutes to talk about your data? Willingness to participate and nonparticipation bias in Facebook data donation. *Field Methods.* 10.1177/1525822X231225907

Willingness to Donate

- How **willing** are Facebook users to donate their data in a web survey?
  - What effect does the **framing** of the data donation request have on willingness to donate?
- How **successful** are Facebook users donating the data?
- What **bias** does arise from selective willingness to donate and successful donation of Facebook data?
- Can donated data be used to **verify self-reports** on Facebook use?

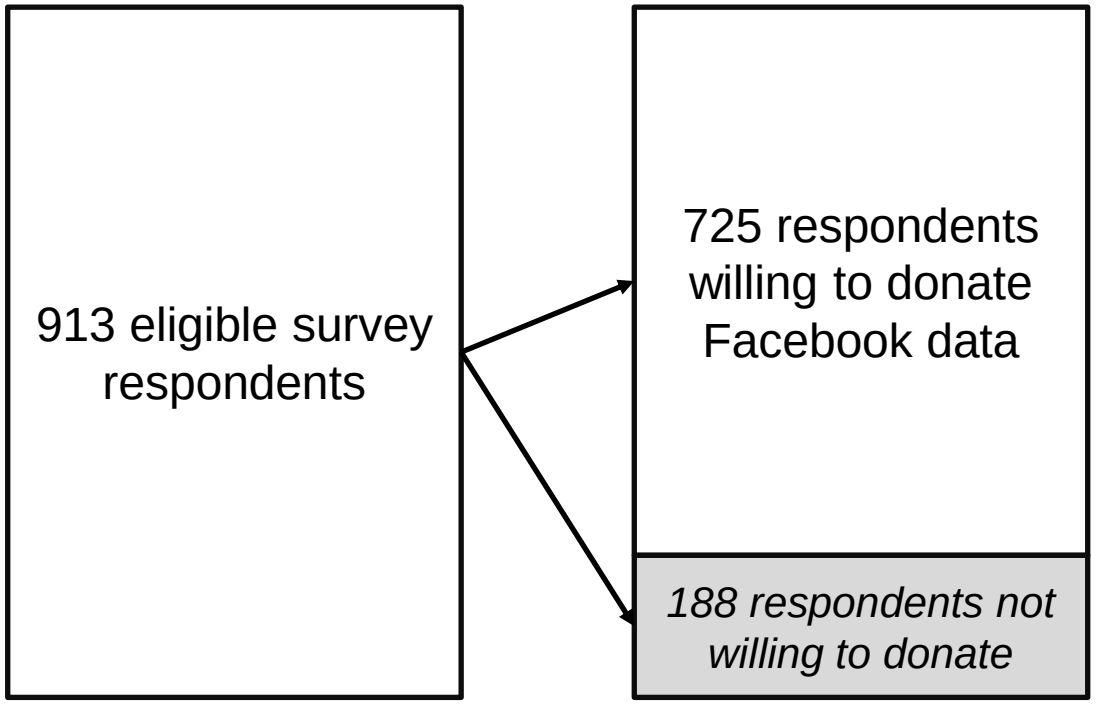
Design Study 3



Data

- 1,083 Facebook users who asked at end of 5-min web survey about willingness to donate FB data
- Request to donate two FB data packages
  - "Account information" and "topics" from past 3 months
  - Separate incentives for survey participation and data donation
- To proceed, participants had to use PC: n = 913
- Experiment: Gain vs. loss framing of data donation

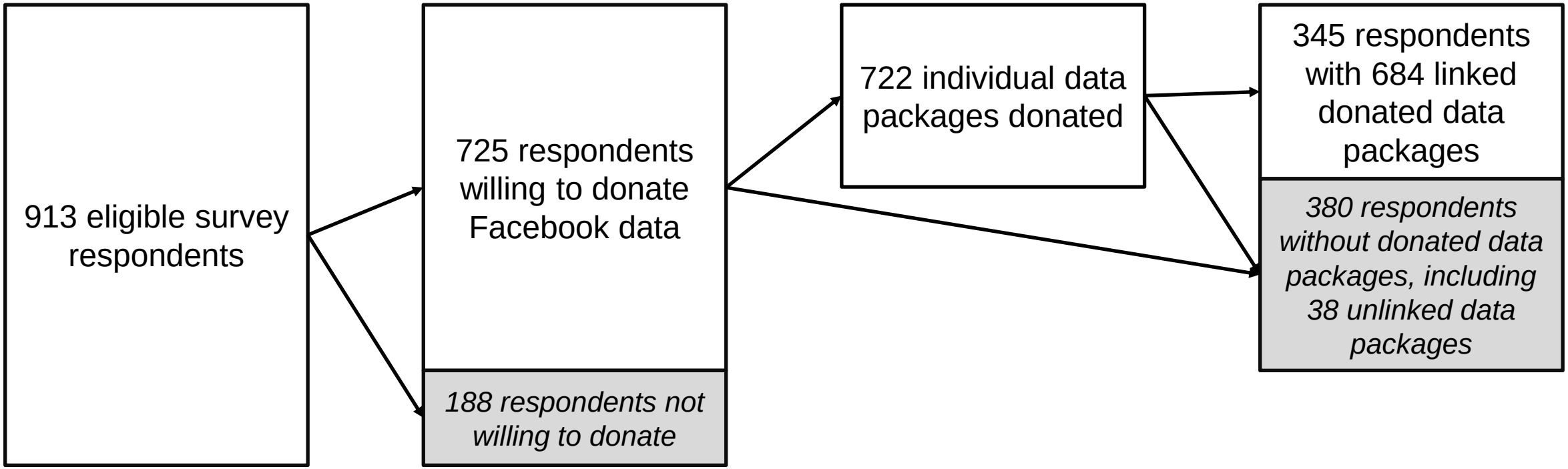
Willingness to Donate FB Data



Willigness to Donate: 79%

- **Men** and those with **higher trust in researchers** sign. more willing to donate
- No sign. effect of general trust, trust in FB, privacy concerns, frequency of FB use, age, education, gain/loss framing
- Main reasons for not being willing to donate (n=140)
  - Wish to protect privacy (24%)
  - Fear of misuse of data (20%)
  - Anticipated technical problems (14%)

Successful Donation



Willigness to Donate: 79%

Succesful Donation: 48%

Successful Donation

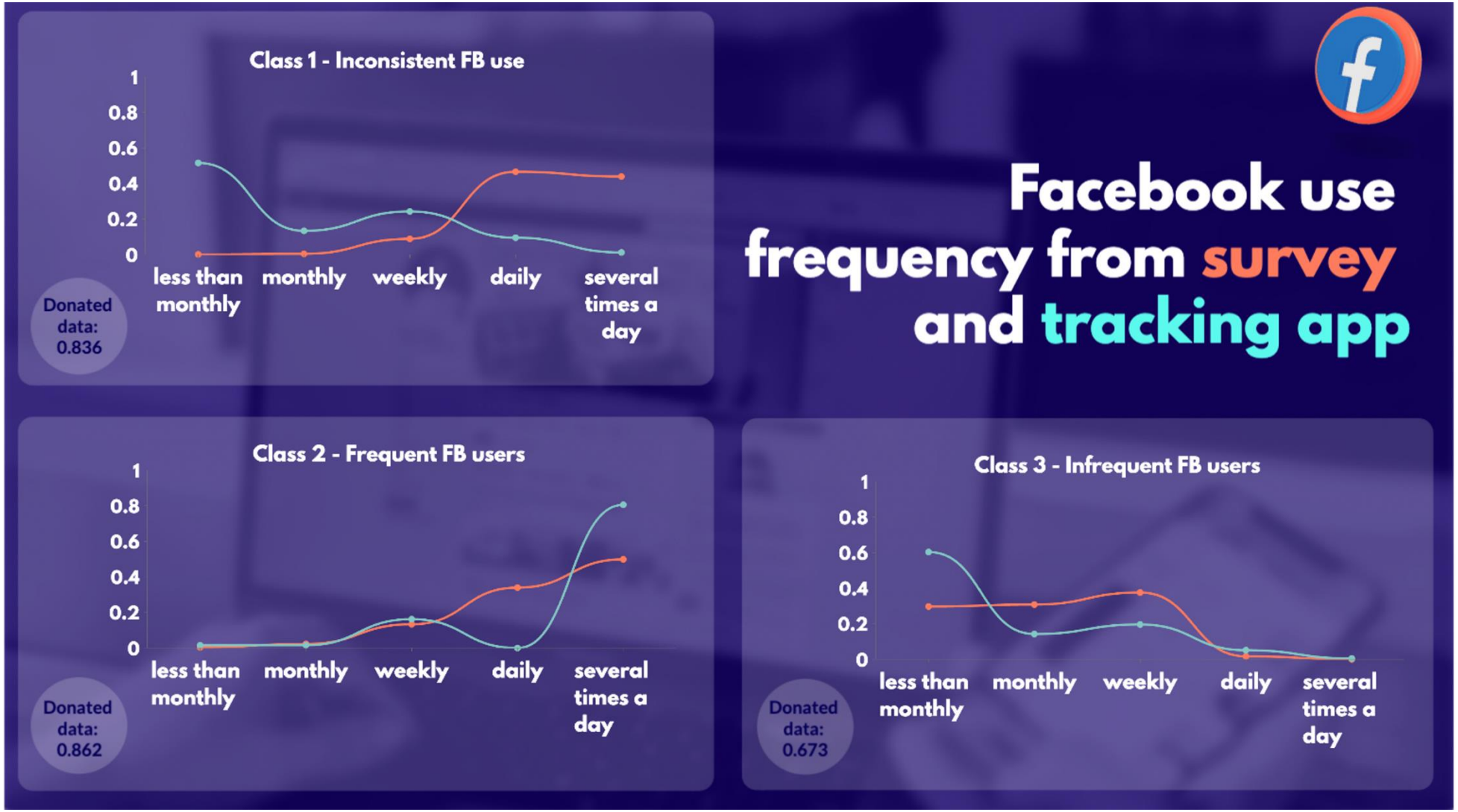
- For 345 (**48%**) survey participants who were willing to donate, we could **link at least one data donation package**
  - Majority of matching as combination of ID + timestamp

- Linkage more likely to be successful for people with **high educational attainment**, **higher trust in researchers**, and **lower trust in FB**

- No sign. effect for privacy concerns, general trust, age, gender, frequency of FB use, and gain/loss framing

- Technical problems stated as main reasons for no successful donation (83%, n=41)

Data Donation to Verify Self-reports



Conclusion Study 3

- While almost 80% of web survey respondents indicated willingness to donate Facebook data, only a little over one third donated
- Reasons for not donating seem to be related to the cumbersome process and linkage problems
  - Implementation will (hopefully soon) become easier
- Donors and nondonors differ in education, trust towards researchers and FB, but no bias in FB use
  - Donated data promising for methodological questions about data quality of trace data and for substantive questions on online media consumption

Summary

	 Surveys	 Meters	 Data Donation
Strengths	• Probability samples • Control of design • Long term comparability	• Direct measurement • Low burden • Detailed/high frequency	• Direct data access • Works with many platforms • High user control • “Gold standard”
Weaknesses	• Fragmented/ discrete information • High burden • Measurement error	• Selective samples • Technology-dependent	• Convoluted process • High drop-out • Linking with other data

Summary & Outlook

- Digital behavioral data have several advantages over self-reports but far from perfect
- Systematic bias when tracking only part of devices does explain some but not all differences observed

- Do different data sources measure the same concepts?
- How to decide which data source to use for what measures?
- Would combining measures from multiple sources improve measurement quality?

Thank You!

Florian Keusch
University of Mannheim
School of Social Sciences
Social Data Science & Methodology
f.keusch@uni-mannheim.de
http://floriankeusch.weebly.com/
@floriankeusch