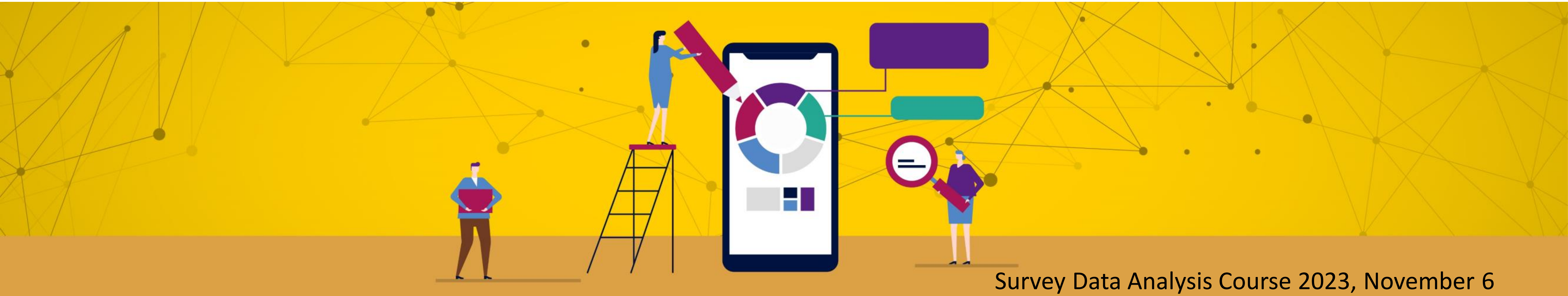
Utrecht University

Designed Big Data
Digital trace data collection using apps, wearables, and data donation

Bella Struminskaya



Survey Data Analysis Course 2023, November 6

Can anonymized data from mobile phone networks predict poverty and wealth?

- Anonymized call records (1.5 mil)
- Telephone survey (n=856)

ECONOMICS

Predicting poverty and wealth from mobile phone metadata

Joshua Blumenstock,[1*] Gabriel Cadamuro,[2] Robert On[3]

Accurate and timely estimates of population characteristics are a critical input to social and economic research and policy. In industrialized economies, novel sources of data are enabling new approaches to demographic profiling, but in developing countries, fewer sources of big data exist. We show that an individual's past history of mobile phone use can be used to infer his or her socioeconomic status. Furthermore, we demonstrate that the predicted attributes of millions of individuals can, in turn, accurately reconstruct the distribution of wealth of an entire nation or to infer the asset distribution of microregions composed of just a few households. In resource-constrained environments where censuses and household surveys are rare, this approach creates an option for gathering localized and timely information at a fraction of the cost of traditional methods.

Reliable, quantitative data on the economic characteristics of a country's population are essential for sound economic policy and research. The geographic distribution of poverty and wealth is used to make decisions about resource allocation and provides a foundation for the study of inequality and the determinants of economic growth (*1*, *2*). In developing countries, however, the scarcity of reliable quantitative data represents a major challenge to policy-makers and researchers. In much of Africa, for instance, national statistics on economic production may be off by as much as 50% (*3*). Spatially disaggregated data, which are necessary for small-area statistics and which are used by both the private and public sector, often do not exist (*4*, *5*).

In wealthy nations, novel sources of passively collected data are enabling new approaches to demographic modeling and measurement (*6*–*8*). Data from social media and the "Internet of Things," for instance, have been used to measure unemployment (*9*), electoral outcomes (*10*), and economic development (*8*). Although most comparable sources of big data are scarce in the world's poorest nations, mobile phones are a notable exception: They are used by 3.4 billion individuals worldwide and are becoming increasingly ubiquitous in developing regions (*11*).

Here we examine the extent to which anonymized data from mobile phone networks can be used to predict the poverty and wealth of individual subscribers, as well as to create high-resolution maps of the geographic distribution of wealth. That this may prove fruitful is motivated by the fact that mobile phone data capture rich information, not only on the frequency and timing of communication events (*12*) but also reflecting the intricate structure of an individual's social network (*13*, *14*), patterns of travel and location choice (*15*–*17*), and histories of consumption and expenditure. Regionally aggregated measures of phone penetration and use have also been shown to correlate with regionally aggregated population statistics from censuses and household surveys (*8*, *18*, *19*).

Our approach is different from prior work that has examined the relation between regional wealth and regional phone use, as we focus on understanding how the digital footprints of a single individual can be used to accurately predict that same individual's socioeconomic characteristics. This distinction is a scientific one, which also has several important implications: First, it allows for the method to be used in contexts for which recent census or household survey data are unavailable. Second, when an authoritative source of data does exist, it can be used to more objectively validate or refute the model's predictions. This limits the likelihood that the model is overfit on data from a single source, which is otherwise difficult to control, even with careful cross-validation (*20*). Third, our approach allows for a broad class of potential applications that require inferences about specific individuals instead of census tracts. As we discuss in the supplementary materials (section 6), future iterations of this approach could help to improve the targeting of humanitarian aid and social welfare, disseminate information to vulnerable populations, and measure the effects of policy interventions.

For this study, we used an anonymized database containing records of billions of interactions on Rwanda's largest mobile phone network and supplemented this with follow-up phone surveys of a geographically stratified random sample of 856 individual subscribers. Upon contacting and surveying each of these individuals, we received informed consent to merge their survey responses with the mobile phone transaction database. The surveys solicited no personally identifying information but contained questions on asset ownership, housing characteristics, and several other basic welfare indicators. From these data, we constructed a composite wealth index using the first principal component of several survey responses related to wealth (*21*, *22*) (supplementary materials section 1D). For each of the 856 respondents, we thus have ~75 survey responses, as well as the historical records of thousands of phone-based interactions such as calls and text messages (Table 1).

We use the merged data from this sample of 856 phone survey respondents to show that a mobile phone subscriber's wealth can be predicted from his or her historical patterns of phone use (Fig. 1A) (cross-validated correlation coefficient r = 0.68). Our approach to modeling combines feature engineering with feature selection by first transforming each person's mobile phone transaction logs into a large set of quantitative metrics and then winnowing out metrics

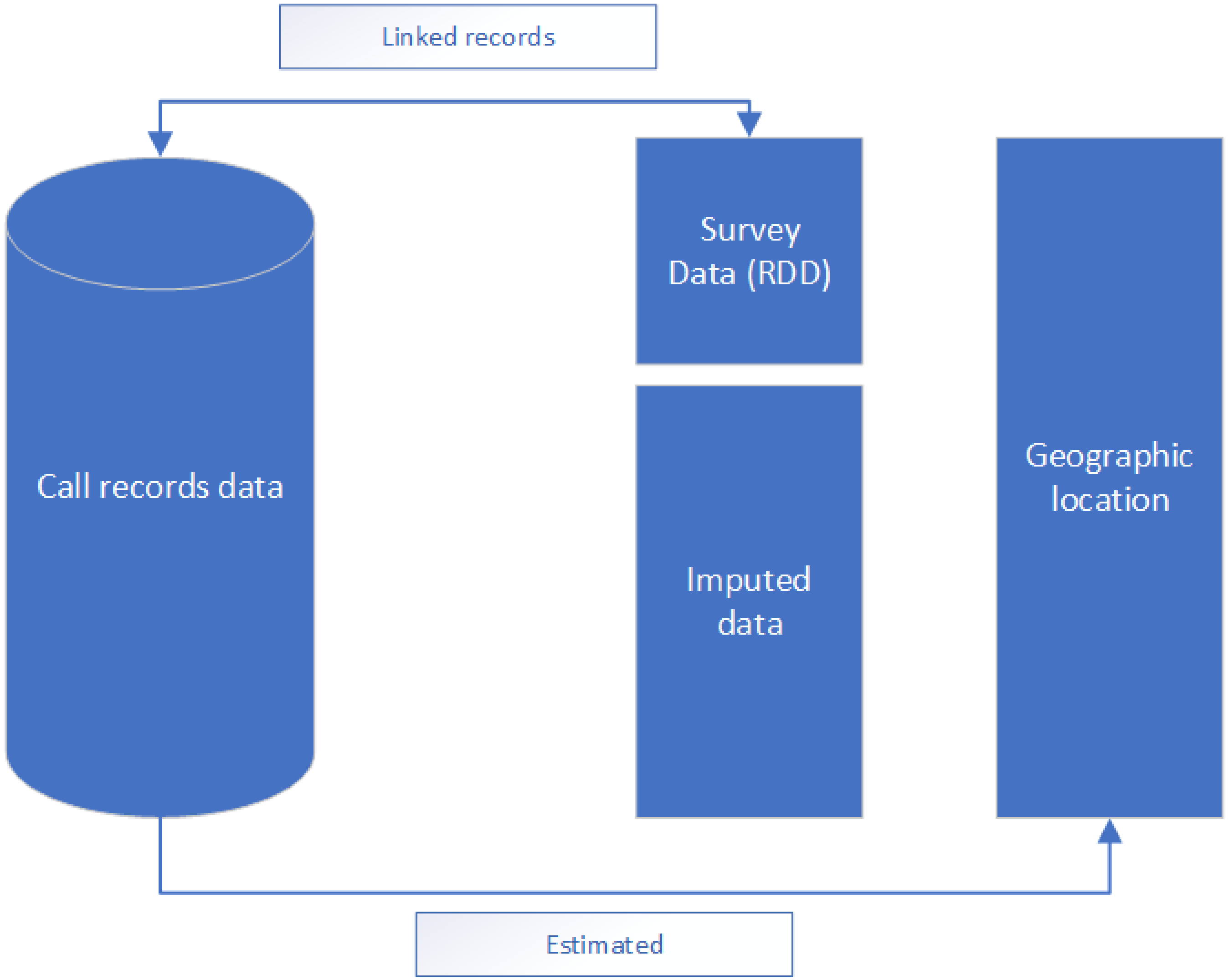
[1]Information School, University of Washington, Seattle, WA 98195, USA. [2]Department of Computer Science and Engineering, University of Washington, Seattle, WA 98195, USA. [3]School of Information, University of California, Berkeley, Berkeley, CA 94720, USA.
*Corresponding author. E-mail: joshblum@uw.edu

Table 1. Summary statistics for primary data sets. Phone survey data were collected by the authors in Kigali, in collaboration with the Kigali Institute of Science and Technology. Call detail records were collected by the primary mobile phone operator in Rwanda at the time of the phone survey. Demographic and Health Survey (DHS) data were collected by the Rwandan National Institute of Statistics. N/A, not applicable.

Summary statistic	Phone survey	Call detail records	DHS (2007)	DHS (2010)
Number of unique individuals	856	1.5 million	7377	12,792
Data collection period	July 2009	May 2008–May 2009	Dec. 2007–Apr. 2008	Sept. 2010–Mar. 2011
Number of questions in survey	75	N/A	1615	3396
Primary geographic units	30 districts	30 districts	30 districts	30 districts
Secondary geographic units	300 cell towers	300 cell towers	247 clusters	492 clusters

(Blumenstock et al. 2015)

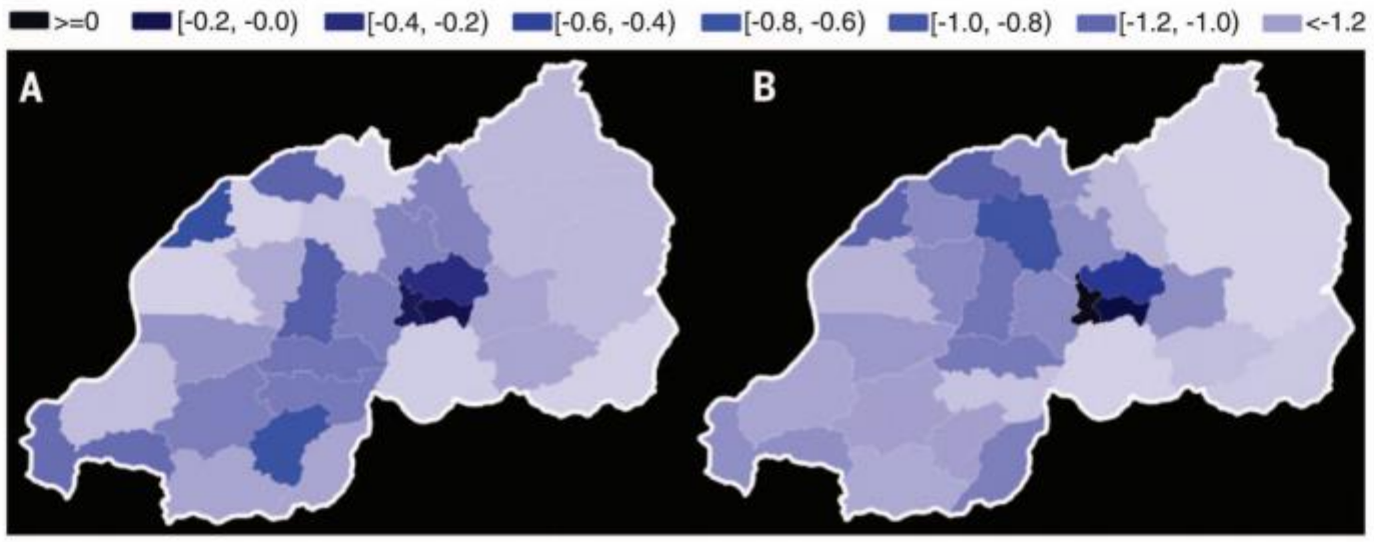
Can anonymized data from mobile phone networks predict poverty and wealth?



- Call activity
- SMS activity
- International communications
- Network structure
- Movement
- etc.

Can anonymized data from mobile phone networks predict poverty and wealth?

- Anonymized call records (1.5 mil)
- Telephone survey (n=856)
- ‘Gold standard’ f2f Demographic and Health Survey (n=12792)

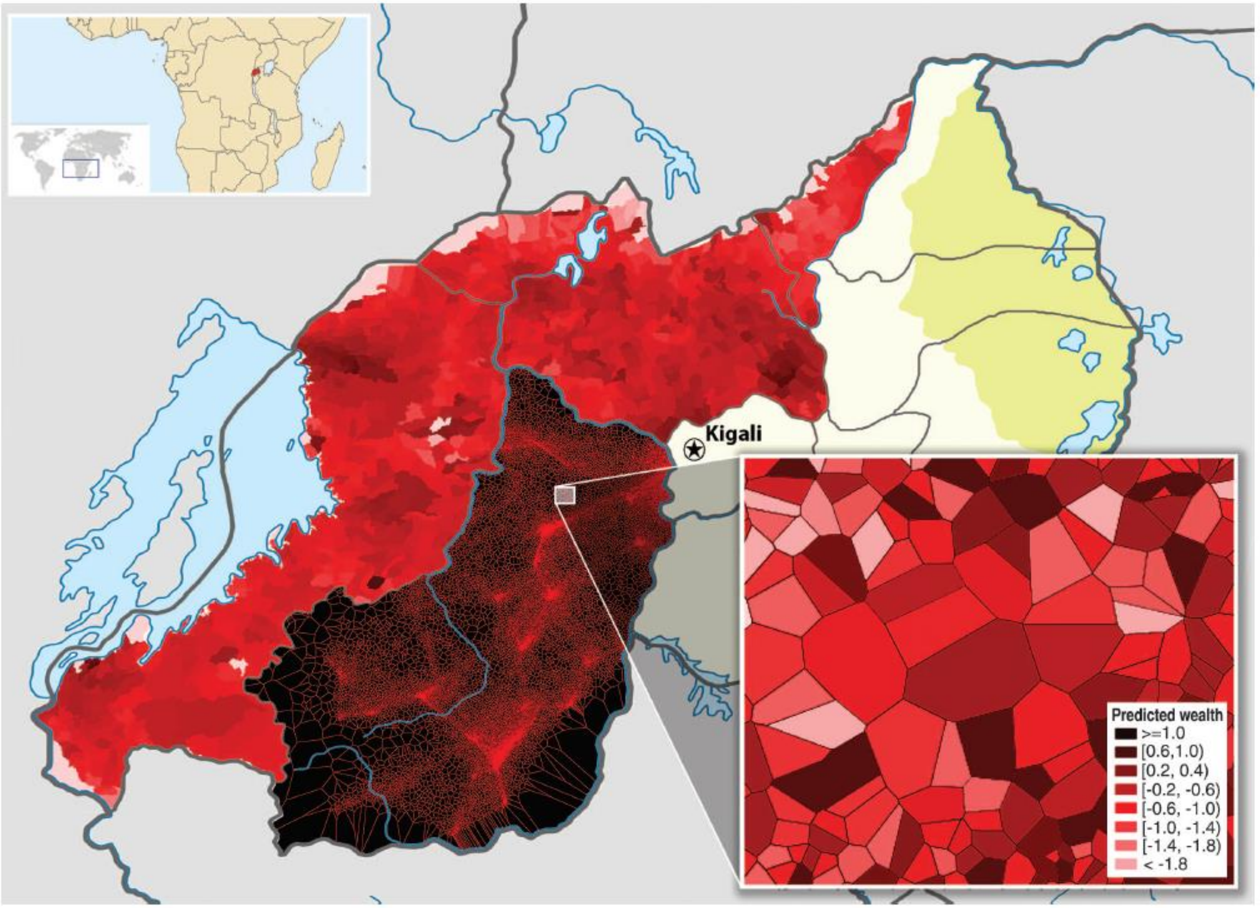


Composite wealth index: A – predicted from call data, B – actual from DHS, *r*=0.79

(Blumenstock et al. 2015)

Added value

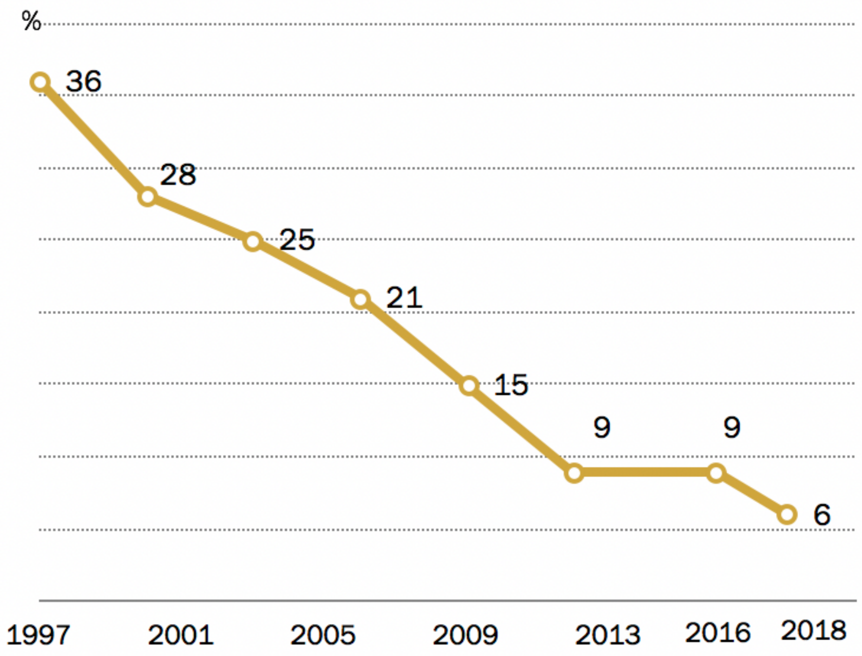
- High-resolution maps of poverty and wealth
- Small area estimation: survey provided estimates on cluster level, call records much richer
- Timely data
- Costs (12,000 vs. 1 Mil)



Decreasing response rates

After brief plateau, telephone survey response rates have fallen again

Response rate by year (%)



Note: Response rate is AAPOR RR3. Only landlines sampled 1997-2006. Rates are typical for surveys conducted in each year.
Source: Pew Research Center telephone surveys conducted 1997-2018.

PEW RESEARCH CENTER

The New York Times

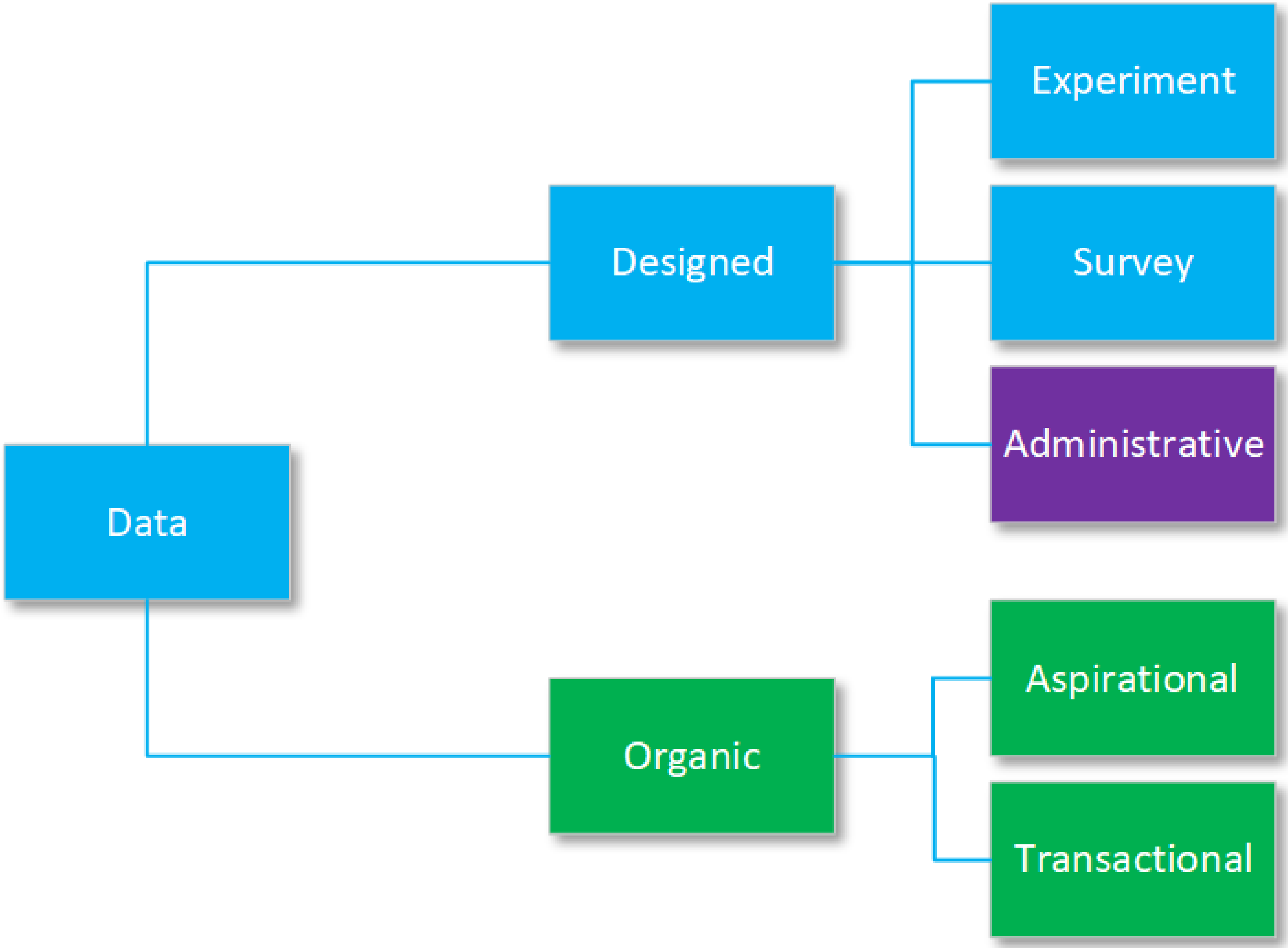
Get Ho

Surprising Poll Results: People Are Now Happy to Pick Up the Phone

Pollsters are used to having their calls screened. But when everyone is stuck at home, a stranger with some survey questions can be a lifeline.

https://www.pewresearch.org/fact-tank/2019/02/27/response-rates-in-telephone-surveys-have-resumed-their-decline/

https://www.nytimes.com/2020/04/17/us/politics/polling-coronavirus.html



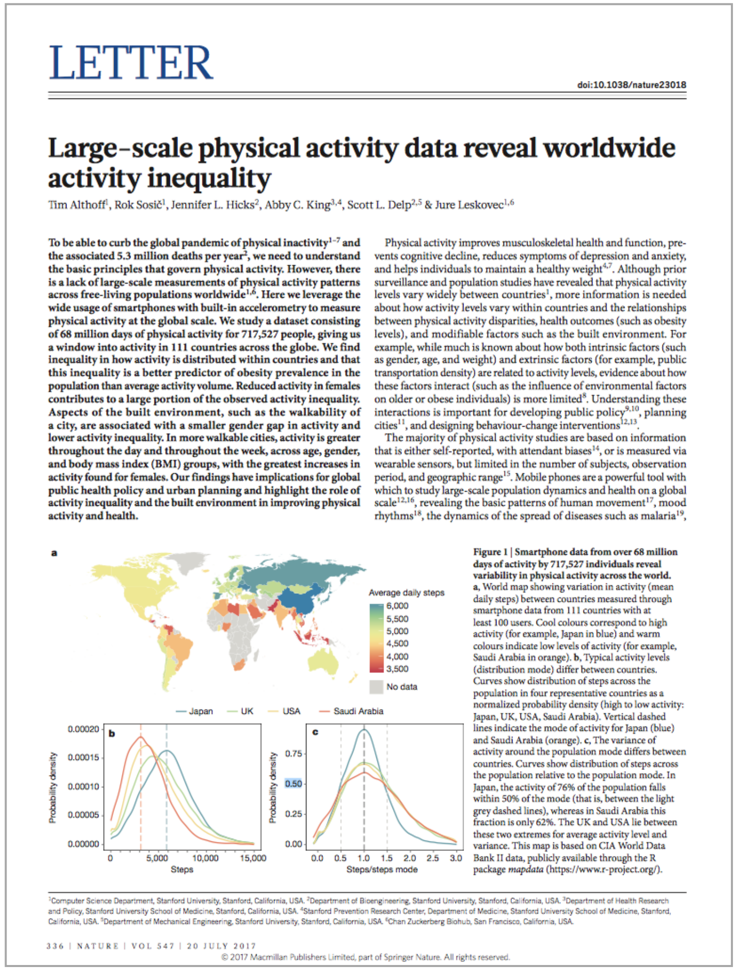
(Source: Kreuter 2018)

Surveys & Big data

Surveys	Big data
• “Designed” data: Collected for the research purposes • Researcher control over content • Large number of covariates • Detailed documentation of the data generating process	• “Organic” data: Collected for purposes other than research • No control over content • Limited number of covariates • No / little documentation • Access issues • (Missingness & coverage)
• High nonresponse • Small N • Measurement error (recall, social desirability)	• Large N • No measurement error due to self-report

(based on Baker 2018, Groves 2011, Sakshaug 2015, Salganik 2018)

Example: Althoff et al. (2017)



LETTER

doi:10.1038/nature23018

Large-scale physical activity data reveal worldwide activity inequality

Tim Althoff[1], Rok Sosič[1], Jennifer L. Hicks[2], Abby C. King[3,4], Scott L. Delp[2,5] & Jure Leskovec[1,6]

To be able to curb the global pandemic of physical inactivity[1–7] and the associated 5.3 million deaths per year[2], we need to understand the basic principles that govern physical activity. However, there is a lack of large-scale measurements of physical activity patterns across free-living populations worldwide[1,6]. Here we leverage the wide usage of smartphones with built-in accelerometry to measure physical activity at the global scale. We study a dataset consisting of 68 million days of physical activity for 717,527 people, giving us a window into activity in 111 countries across the globe. We find inequality in how activity is distributed within countries and that this inequality is a better predictor of obesity prevalence in the population than average activity volume. Reduced activity in females contributes to a large portion of the observed activity inequality. Aspects of the built environment, such as the walkability of a city, are associated with a smaller gender gap in activity and lower activity inequality. In more walkable cities, activity is greater throughout the day and throughout the week, across age, gender, and body mass index (BMI) groups, with the greatest increases in activity found for females. Our findings have implications for global public health policy and urban planning and highlight the role of activity inequality and the built environment in improving physical activity and health.

Physical activity improves musculoskeletal health and function, prevents cognitive decline, reduces symptoms of depression and anxiety, and helps individuals to maintain a healthy weight[4,7]. Although prior surveillance and population studies have revealed that physical activity levels vary widely between countries[1], more information is needed about how activity levels vary within countries and the relationships between physical activity disparities, health outcomes (such as obesity levels), and modifiable factors such as the built environment. For example, while much is known about how both intrinsic factors (such as gender, age, and weight) and extrinsic factors (for example, public transportation density) are related to activity levels, evidence about how these factors interact (such as the influence of environmental factors on older or obese individuals) is more limited[8]. Understanding these interactions is important for developing public policy[9,10], planning cities[11], and designing behaviour-change interventions[12,13].

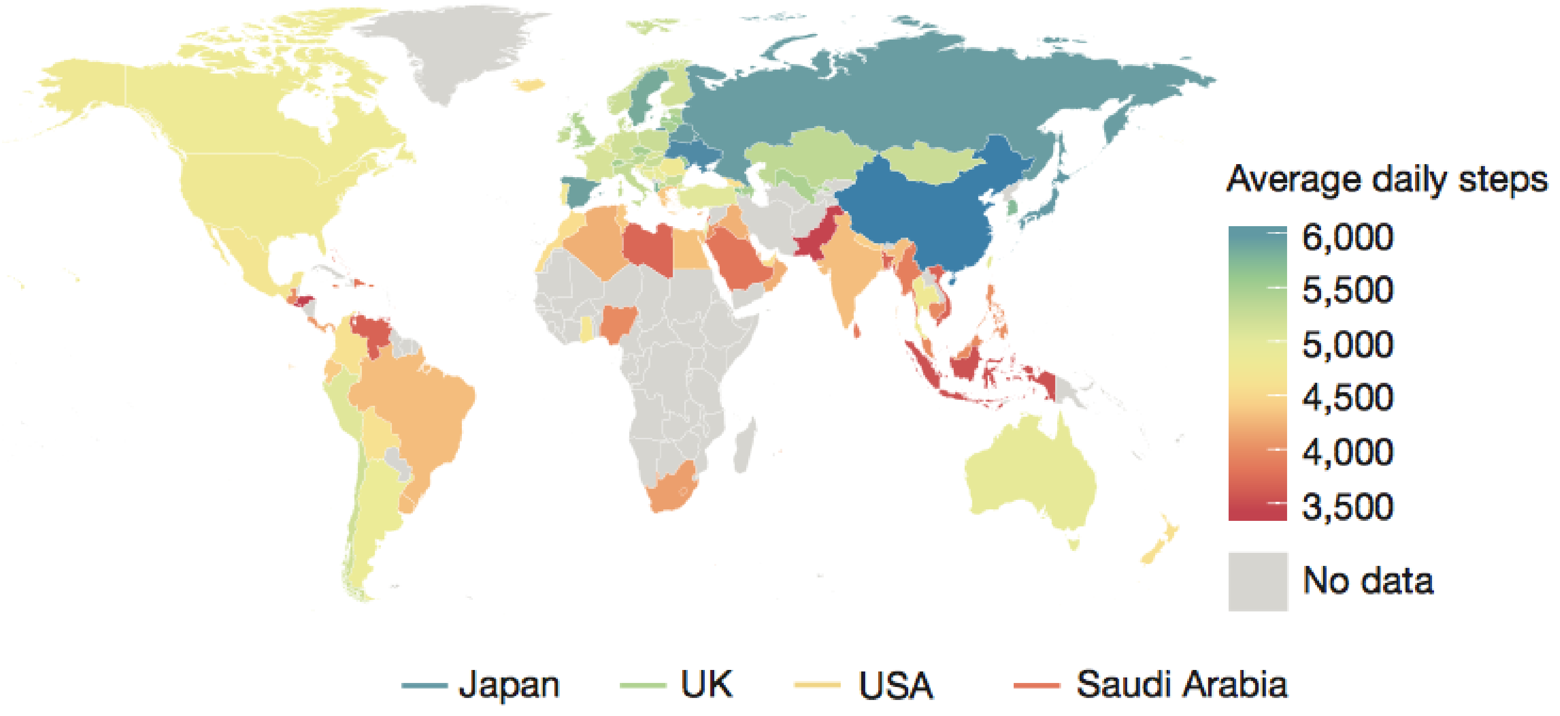
The majority of physical activity studies are based on information that is either self-reported, with attendant biases[14], or is measured via wearable sensors, but limited in the number of subjects, observation period, and geographic range[15]. Mobile phones are a powerful tool with which to study large-scale population dynamics and health on a global scale[12,16], revealing the basic patterns of human movement[17], mood rhythms[18], the dynamics of the spread of diseases such as malaria[19],

Figure 1 | Smartphone data from over 68 million days of activity by 717,527 individuals reveal variability in physical activity across the world. **a,** World map showing variation in activity (mean daily steps) between countries measured through smartphone data from 111 countries with at least 100 users. Cool colours correspond to high activity (for example, Japan in blue) and warm colours indicate low levels of activity (for example, Saudi Arabia in orange). **b,** Typical activity levels (distribution mode) differ between countries. Curves show distribution of steps across the population in four representative countries as a normalized probability density (high to low activity: Japan, UK, USA, Saudi Arabia). Vertical dashed lines indicate the mode of activity for Japan (blue) and Saudi Arabia (orange). **c,** The variance of activity around the population mode differs between countries. Curves show distribution of steps across the population relative to the population mode. In Japan, the activity of 76% of the population falls within 50% of the mode (that is, between the light grey dashed lines), whereas in Saudi Arabia this fraction is only 62%. The UK and USA lie between these two extremes for average activity level and variance. This map is based on CIA World Data Bank II data, publicly available through the R package *mapdata* (https://www.r-project.org/).

[1]Computer Science Department, Stanford University, Stanford, California, USA. [2]Department of Bioengineering, Stanford University, Stanford, California, USA. [3]Department of Health Research and Policy, Stanford University School of Medicine, Stanford, California, USA. [4]Stanford Prevention Research Center, Department of Medicine, Stanford University School of Medicine, Stanford, California, USA. [5]Department of Mechanical Engineering, Stanford University, Stanford, California, USA. [6]Chan Zuckerberg Biohub, San Francisco, California, USA.

336 | NATURE | VOL 547 | 20 JULY 2017

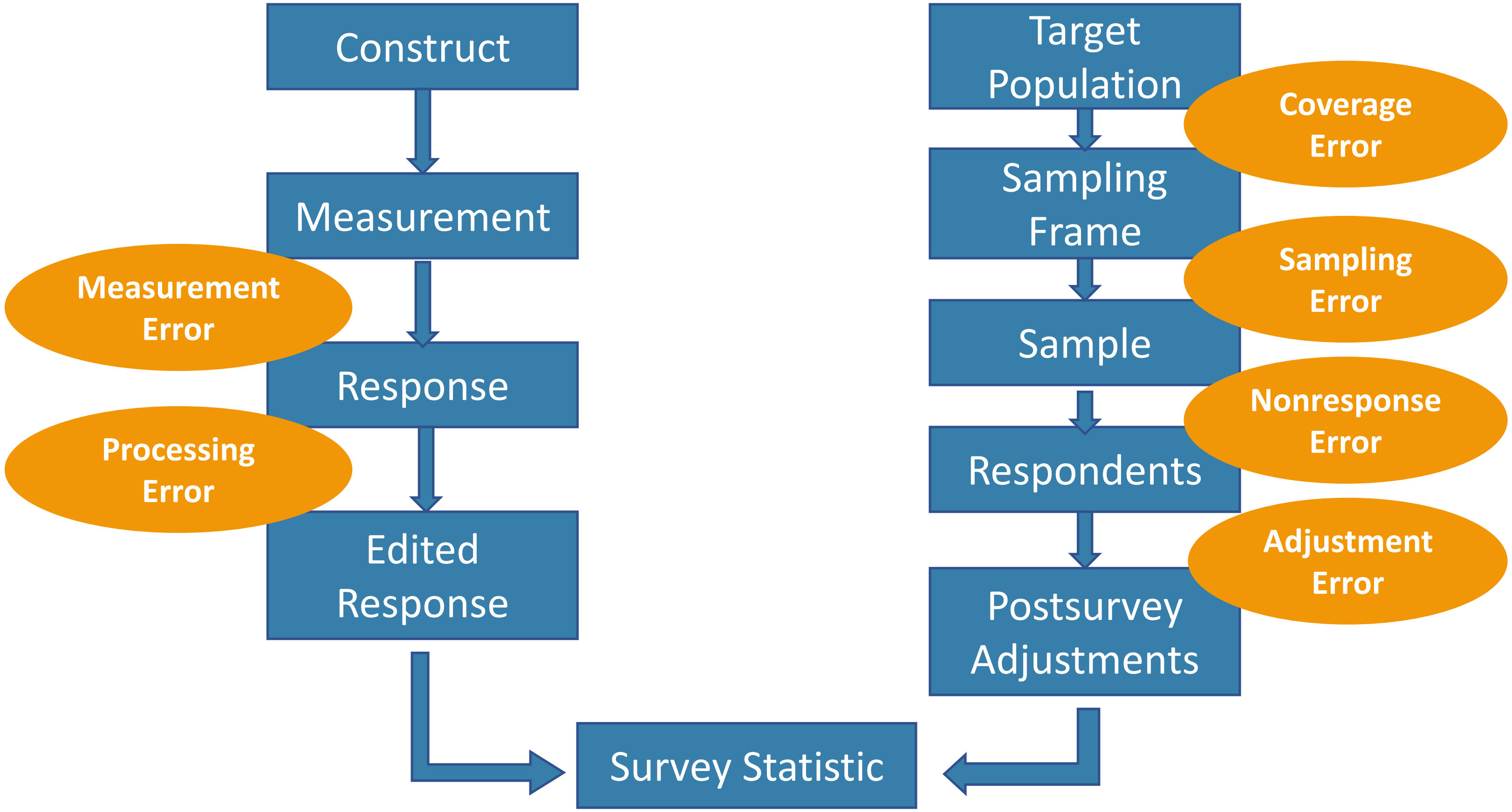
© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Althoff, T., Hicks, J. L., King, A. C., Delp, S. L., & Leskovec, J. (2017). Large-scale physical activity data reveal worldwide activity inequality. *Nature*, *547* (7663), *336-339*

Total Survey Error & Althoff et al.

device malfunction/
out of power



only iPhone users

only volunteers

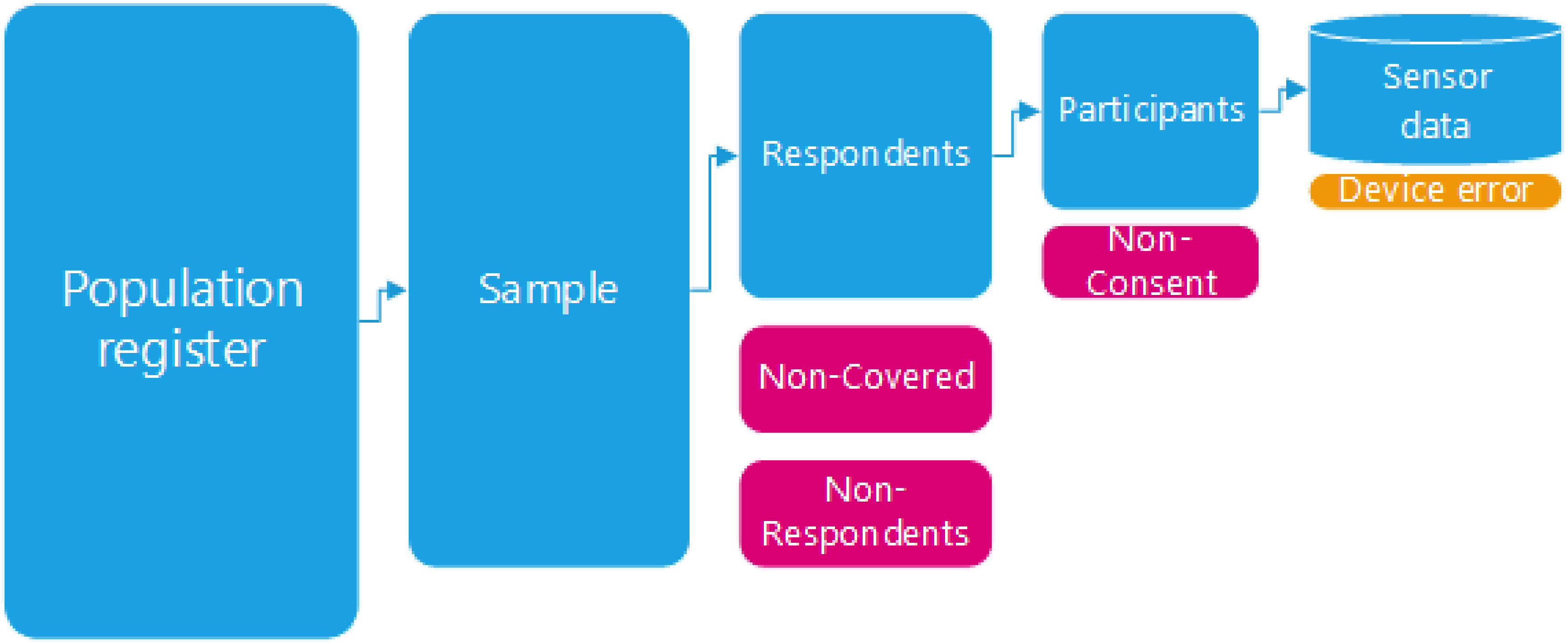
only those who
downloaded the app

population
characteristics known?

(Groves et al. 2004: 48)

Introducing “design” to Big Data

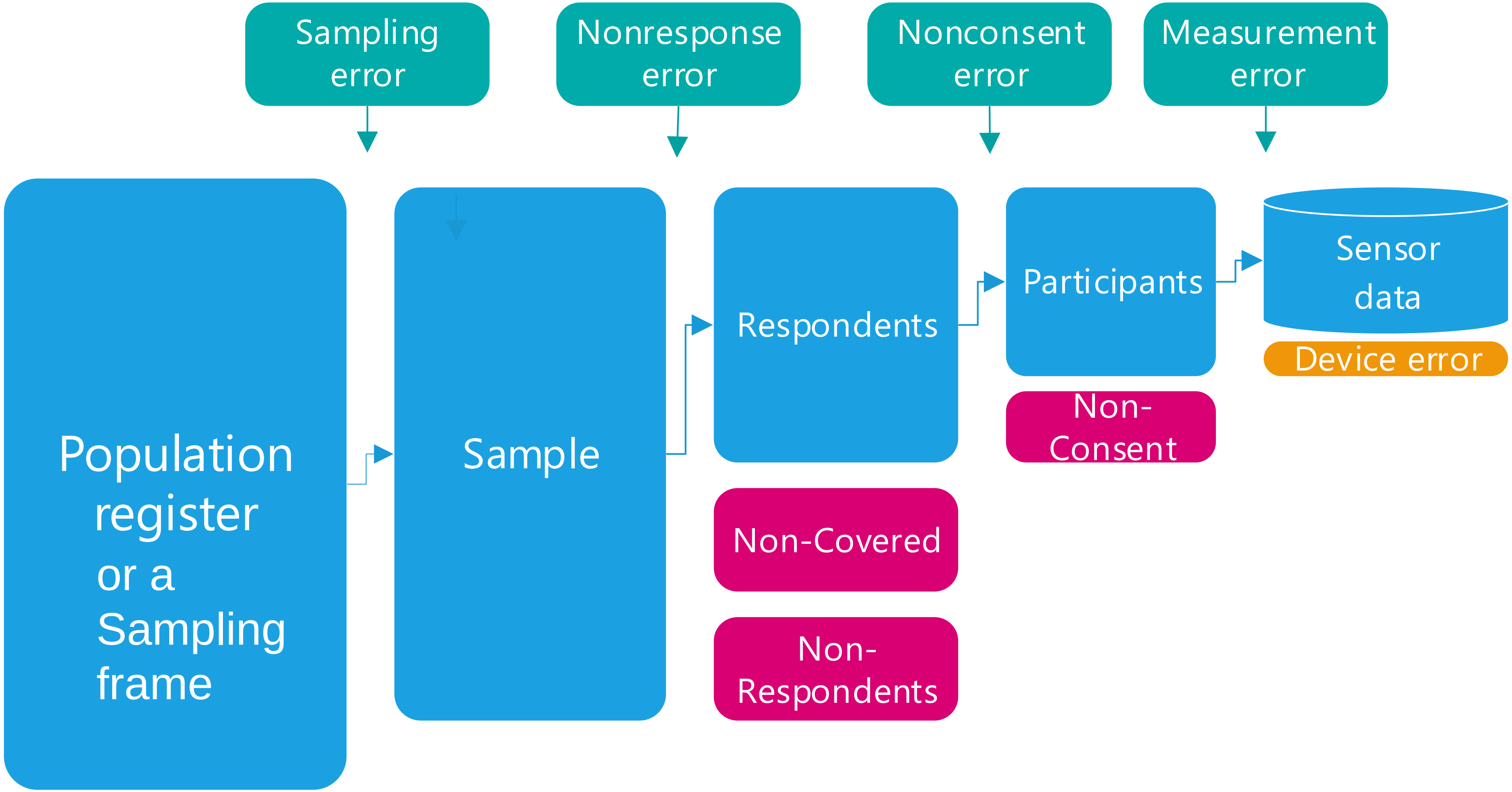
- Smartphone sensor data have many characteristics of Big Data
  - Large volume, high velocity, variety of data formats
- Combining passive smartphone data collection with self-reports through surveys introduces “design” to Big Data



Big data particularly useful for

- Replace surveys/most survey questions
    - Travel
    - Budget
    - User groups/online communities
    - administration
- Increase survey data quality
    - Adding administrative data
    - Adding sensor data
    - Using social media data as a qualitative/pilot study
    - Transaction data? As an explanatory variable?

“Designed Big Data”



Definition of digital traces

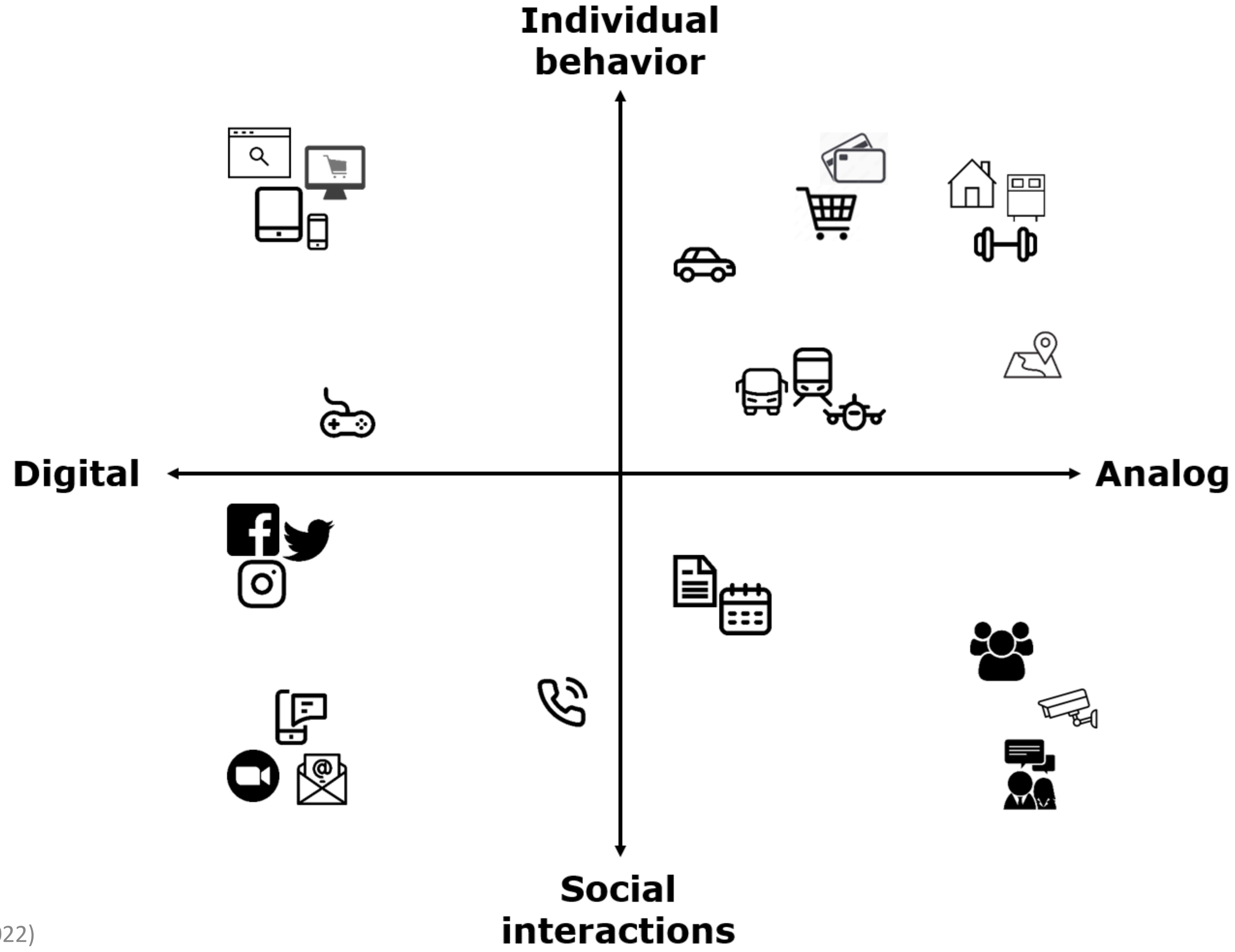
"Records of activity (trace data) undertaken through an online information system (thus, digital)"

(Howison et al. 2011:769)

"Behavioral residue [individuals leave] when they interact online"

(Hinds & Joinson 2018:2)

Exercise: Where and when do you leave digital traces?


Individual behavior
Digital
Analog
Social interactions

Passive data collection using smartphone sensors

Bella Struminskaya

with thanks to Florian Keusch

b.struminskaya@uu.nl
http://bellastrum.com/

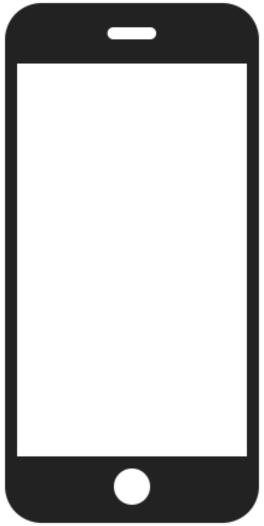
Native smartphone sensors

NFC

Bluetooth

Air humidity sensor

Thermometer



Proximity sensor

Wi-Fi

Microphone

GPS

Light sensor

Cellular network

Camera

Fingerprint sensor

Barometer

Compass

Accelerometer

Pedometer

Gyroscope

What can you measure with these sensors?

Native smartphone sensors

Proximity

NFC

Bluetooth

Air humidity sensor

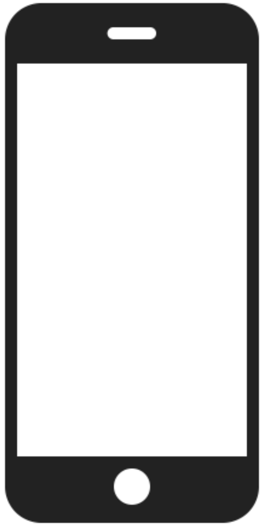
Thermometer

Location

Wi-Fi

GPS

Cellular network



Proximity sensor

Microphone

Light sensor

Ambience

Fingerprint sensor

Camera

Barometer

Compass

Accelerometer

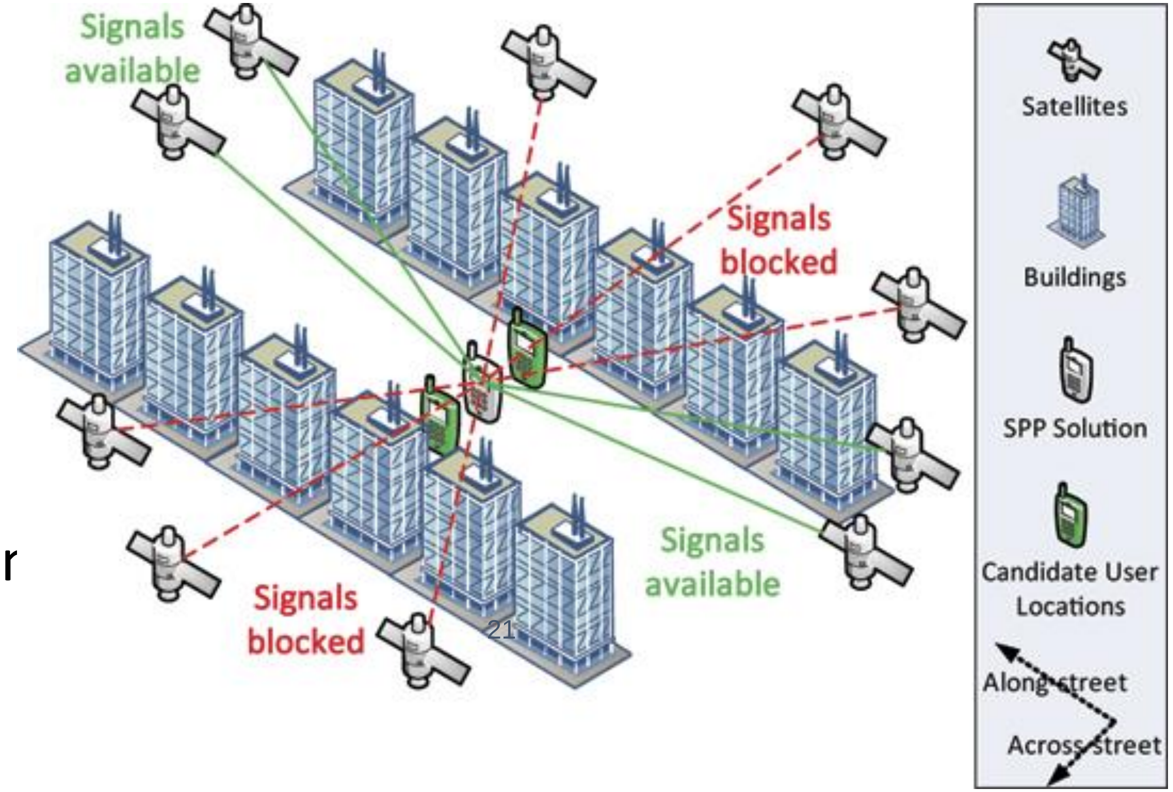
Pedometer

Gyroscope

Physical activity

Geolocation

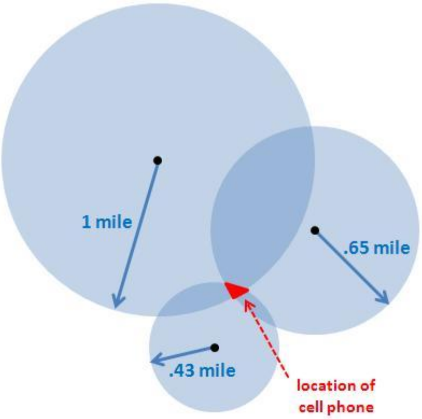
- GPS
  - Provides coordinates in longitude & Latitude
  - Based on distance (= rate x time) to at least 4 satellites
  - Newest generation has accuracy within 30 centimeters
  - Works without cell/Internet connection
  - Performs worse in 'urban canyons', indoors, & underground
  - Constant tracking is very battery-draining



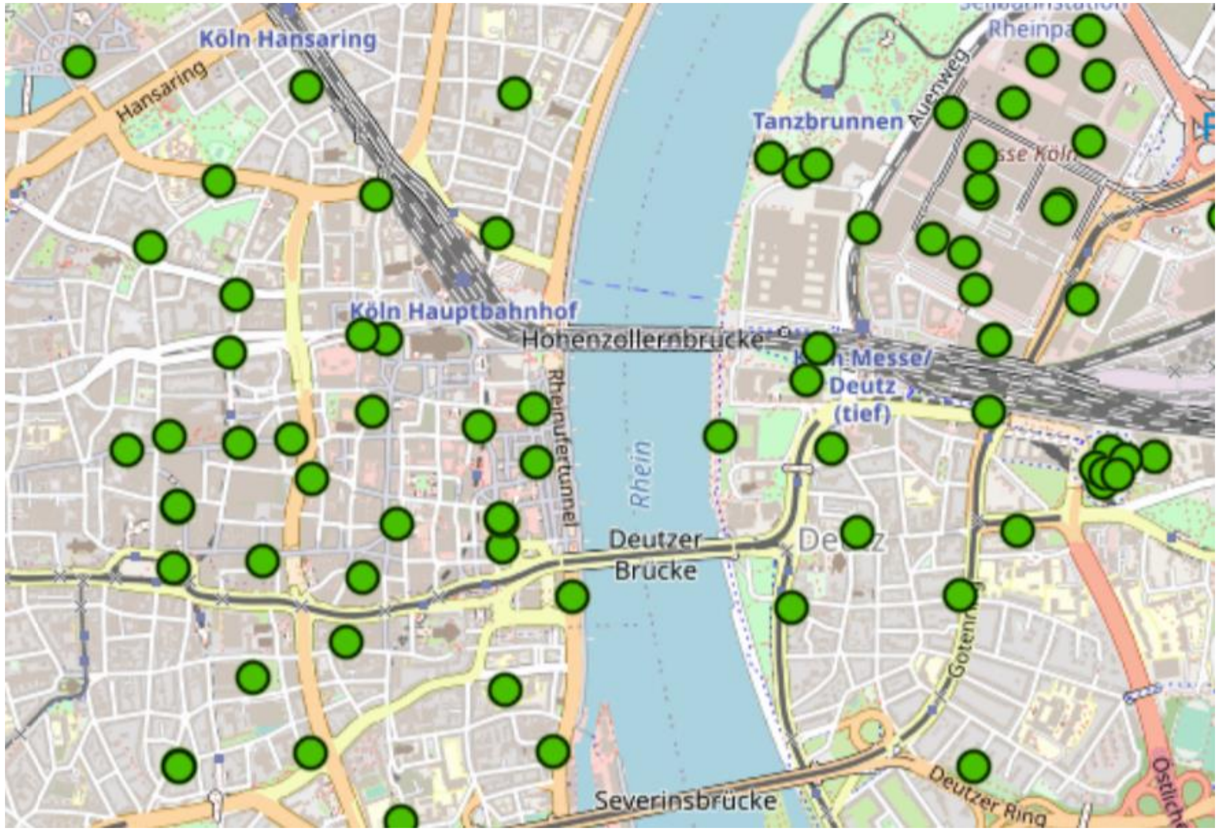
Source: https://www.gpsworld.com/wirelesspersonal-navigationshadow-matching-12550/

Geolocation

- GPS
- Cellular network
  - Multilateration of radio signals between (several) cell towers
  - Works even if GPS is turned off
  - If there is no signal then location information will be missing



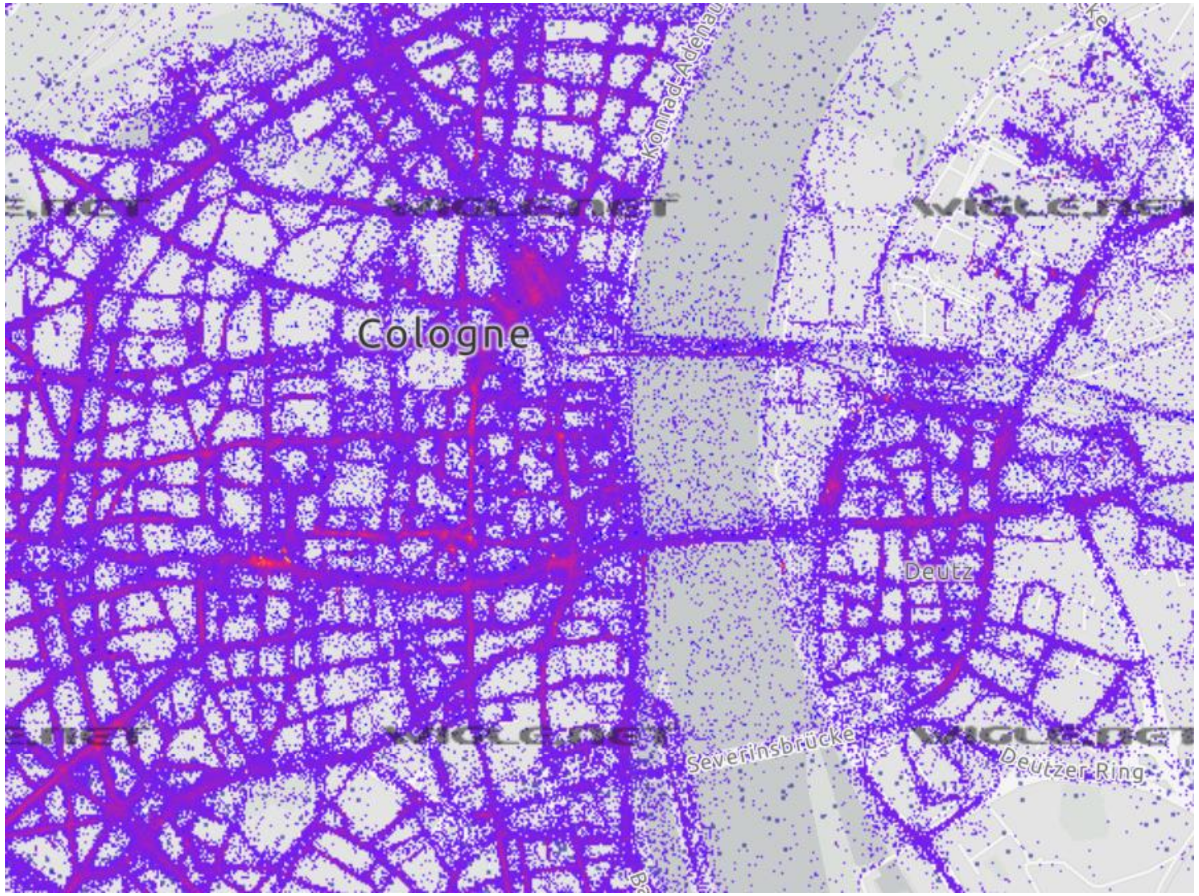
Source: https://searchengineland.com/cell-phone-triangulation-accuracy-is-all-over-the-map-14790



Source: https://www.cellmapper.net

Geolocation

- GPS
- Cellular network
- Wi-Fi
    - Inferring location from Wi-Fi access points (AP)
    - Can overcome problem of ‘urban canyons’ and indoor tracing



Source: https://www.wigle.net

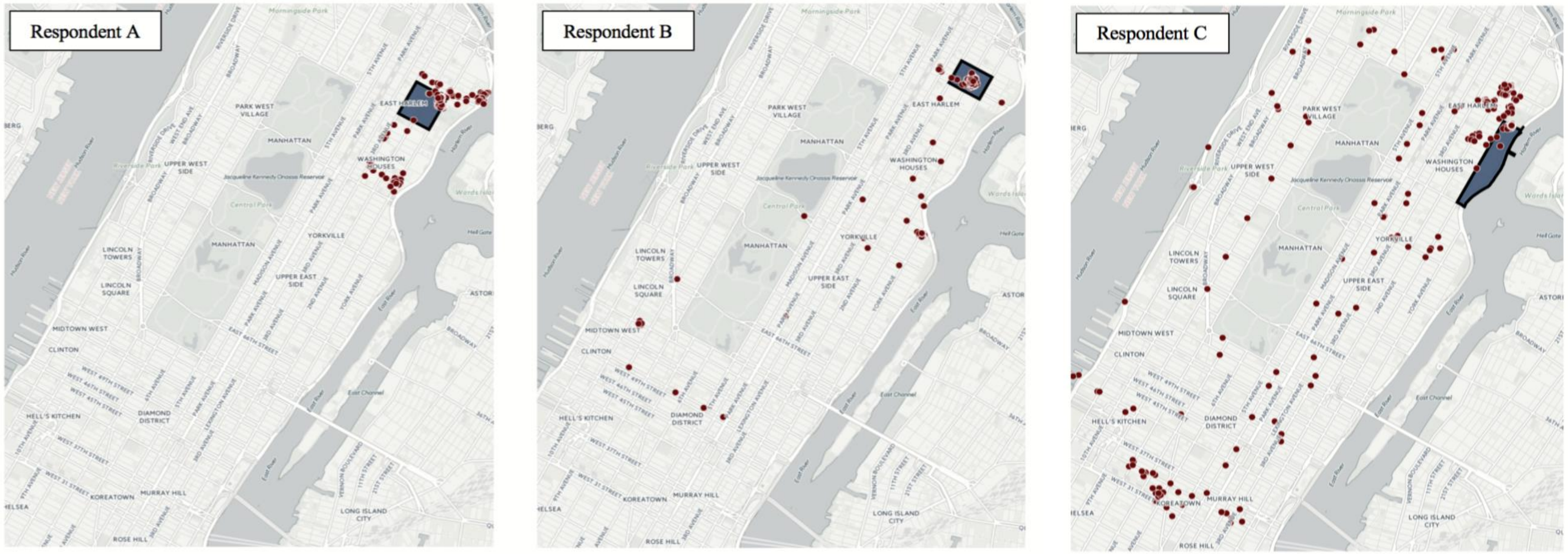
Geolocation

- GPS
- Cellular network
- Wi-Fi
- Hybrid positioning systems
  - Uses combination of systems to make location more accurate (assisted GPS - AGPS)
  - E.g., fall-back on X if Y is not available

Example: Aging in activity space

(York Cornwell & Cagney 2017, 2020)

- *Real-time Neighborhoods and Social Life Study* (RNSL)
- 60 participants aged 55+ in NYC provided with iPhones to carry for 7 days
- GPS-tracking (every 5 min) from 9 a.m. to 9 p.m and four EMAs per day



Example: Aging in activity space

(York Cornwell & Cagney 2017, 2020)

- Activity spaces vary considerably in size
- Participants spent ~40% of their time outside their residential tracts
  - On average >10 min in 9+ tracts
- Activity spaces larger among younger and more advantaged social groups (i.e., whites, those with college degree, car owners)
- Participants with less education and lower incomes spend more time outside
  of their residential tracts
- Four main activities outside of residential tracts
  - Shopping, exercising, socializing, participating in social groups or activities
- Poverty rates in nonresidential tracts lower than in residential tracts
- Higher concentrated disadvantage in an area associated with higher odds of self-reporting pain

Example: How do people find work after prison?

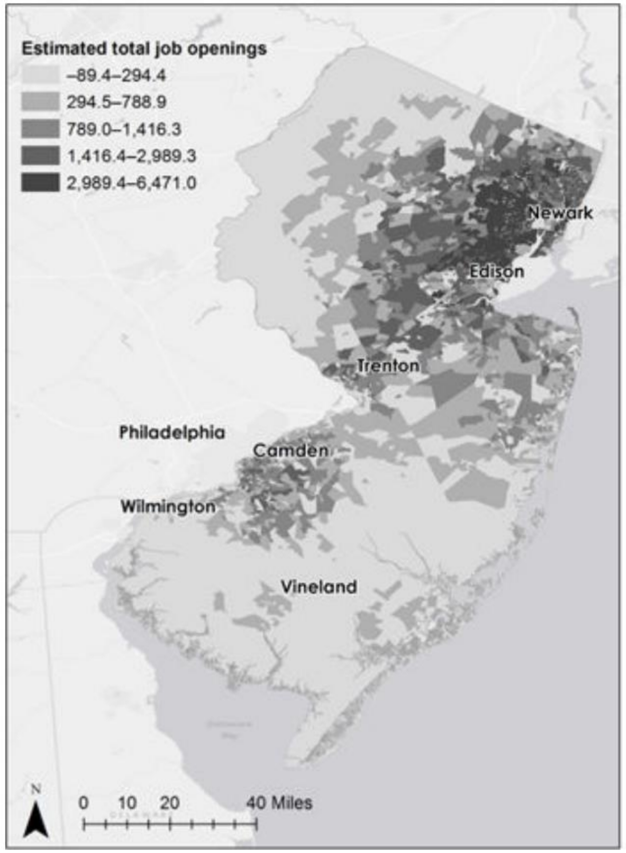
(Sugie 2018; Sugie and Lens 2017)

- Newark Smartphone Reentry Project (NSRP) 2012-2013
  - N = 133 with 8,000 daily observations (89% response, 1.5% noncompliance)
  - 3 months of data collection
- Men recently released from prison
  - Difficult group to follow due to unstable circumstances
- Loaner smartphones (Android)
- Surveys twice a day (EMA) about social interaction, job search & work, and emotional well-being
- Sensing
  - GPS location
  - Calls and messaging (encrypted)
- Survey triggered by calls/messages from new telephone numbers

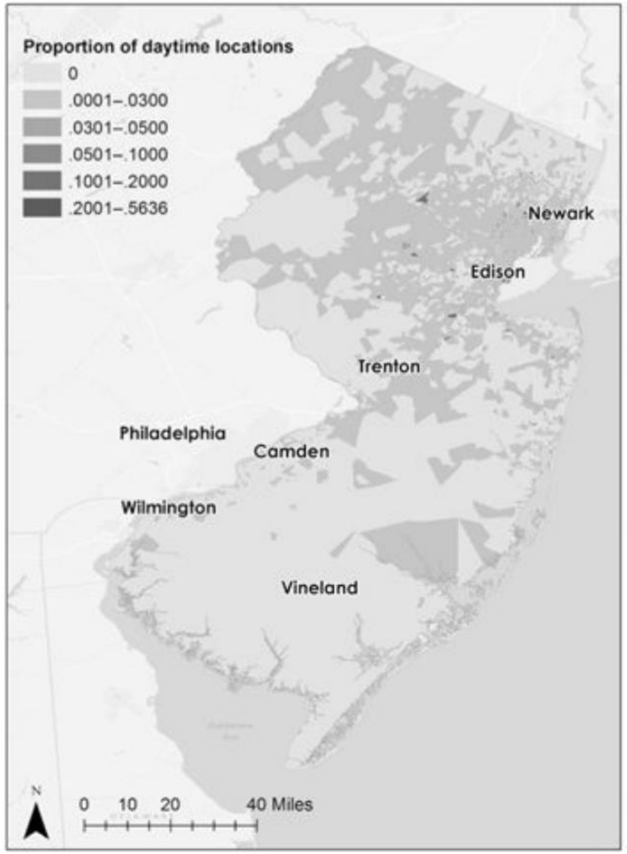
Example: How do people find work after prison?

(Sugie 2018; Sugie and Lens 2017)

- Spatial mismatch
  - Low-skilled, nonwhite job seekers within central cities, job opportunities in outlying areas
- Hypothesis
  - Parolees lack info on job openings are geographically restricted, unable to travel to find work
- Findings
  - Residential mismatch lengthens time to employment
  - But mobility can compensate for residential deficits



Job openings



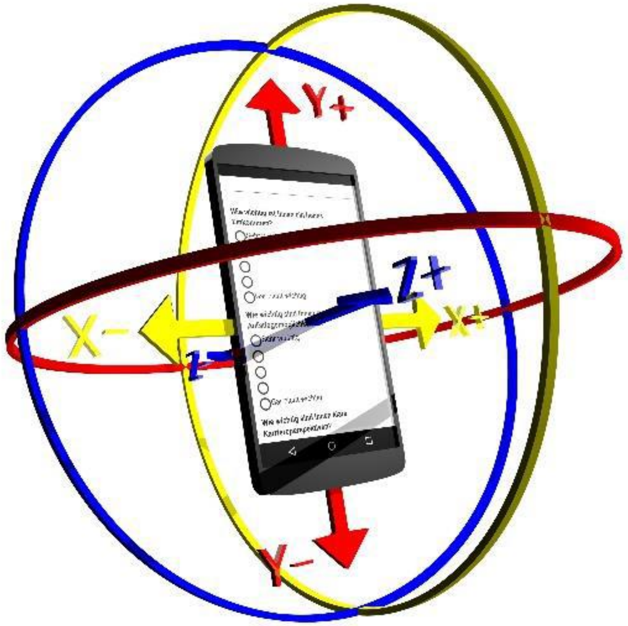
Daytime locations of parolees

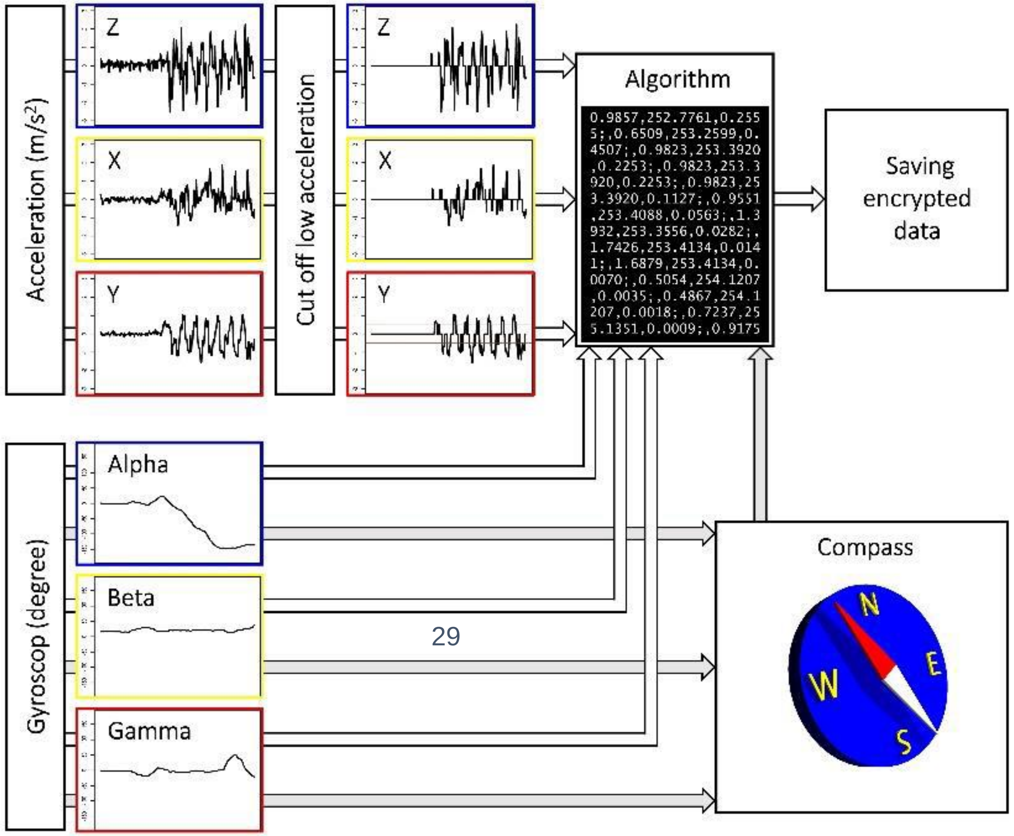
Physical activity

- Accelerometer
- Gyroscope

Source: https://www.techradar.com/news/wearables/10-best-fitness-trackers-1277905

Sources: https://www.actigraphcorp.com/actigraph-wgt3x-bt/, https://www.activinsights.com/products/geneactiv/





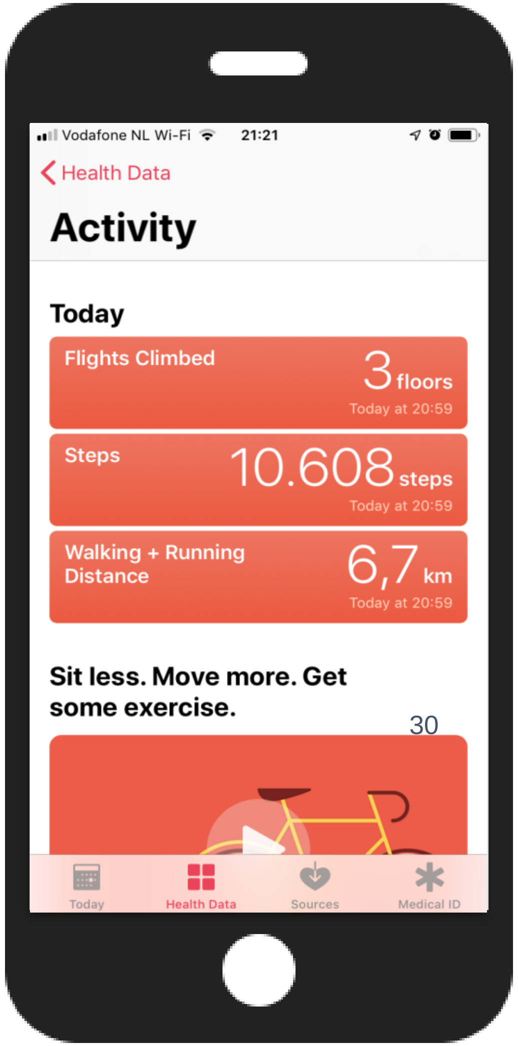
Schlosser et al. (2019)

Physical activity

- Accelerometer
- Gyroscope

  and

- Magnetometer
  - Serves as compass
- Barometer
  - Allows to track changes in elevation



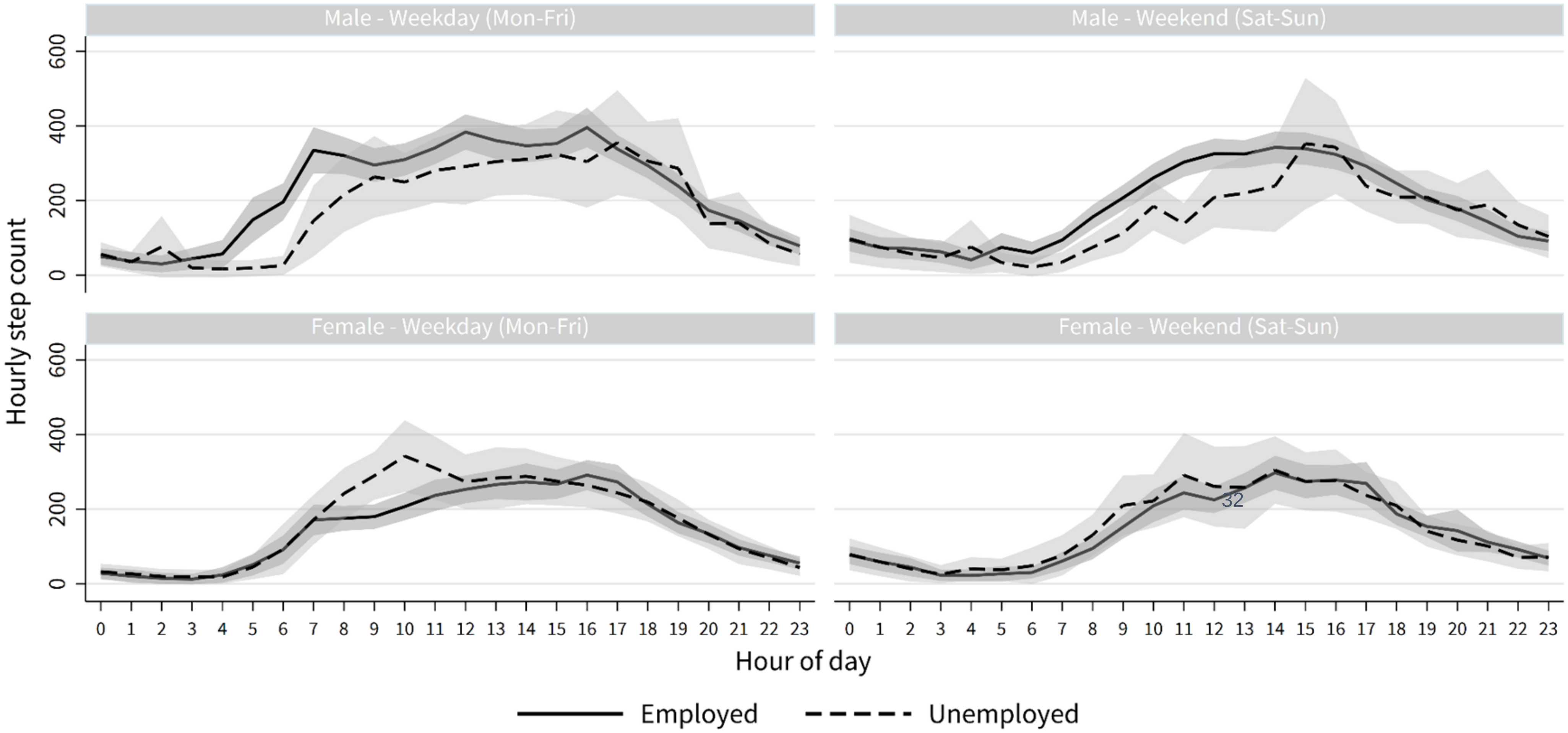
Example: What are the effects of unemployment?

(Kreuter et al. 2020)

- ~650 Android smartphone owners from German panel study "Labour Market and Social Security" (PASS) downloaded *IAB-SMART* app for 6 months
- Survey questions triggered by…
  - Schedule: *Q*s about affective impact of daily smartphone use, Big 5 personality, employment and job search activities, use of smartphones in everyday life, etc.
  - Geolocation: 400 job centers - *Q*s about visit to job center
- Five passive data collection modules:
  - Location using GPS, Wi-Fi, and cellular sensors every 30 min
  - Activity and means of transportation (e.g., walking, biking, riding in/on a motorized vehicle) using accelerometer and pedometer every 2 min
  - Call and texting behavior using phone and SMS logs
  - Use of apps installed on smartphone
  - Social network characteristics from contact lists

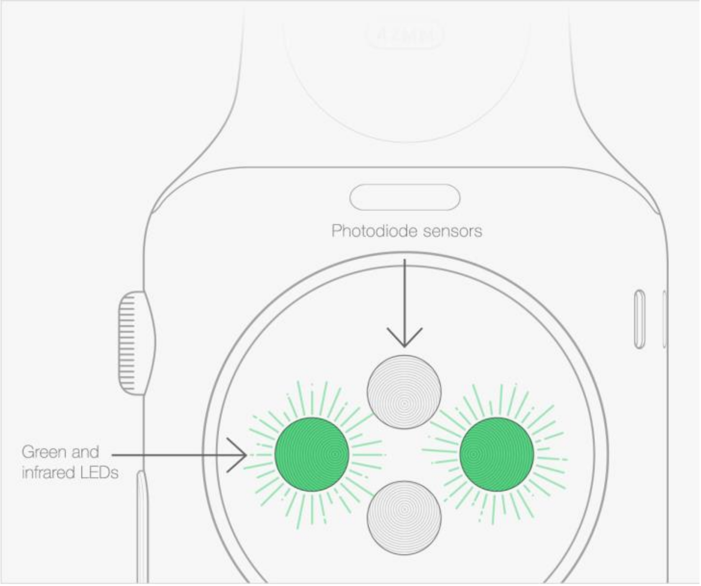
Example: What are the effects of unemployment?

(Bähr et al. in preparation)

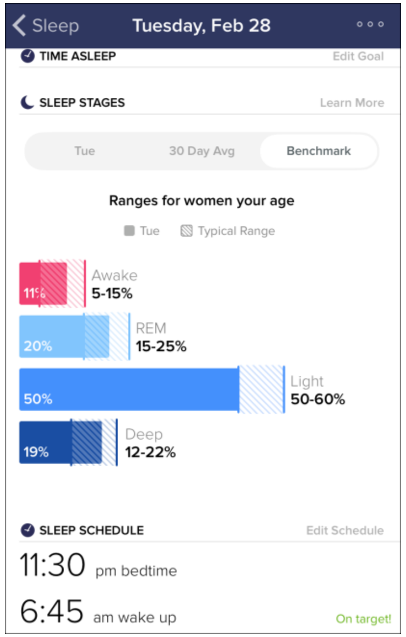


Heart-rate

- Most wristbands use LED-based system
  - Light “shines” onto skin, sensor detects blood
    volume changes
  - “... finely-tuned algorithms are applied to measure heart rate automatically and continuously...” (https://help.fitbit.com/articles/en_US/Help_article/1565)
  - Samsung Galaxy S uses similar system
- Used in combination with accelerometer to determine sleep phases (e.g., on Fitbit)



Source: https://exist.io/blog/fitness-trackers-heart-rate/



Source: https://help.fitbit.com/articles/en_US/Help_article/2163

Sound & light

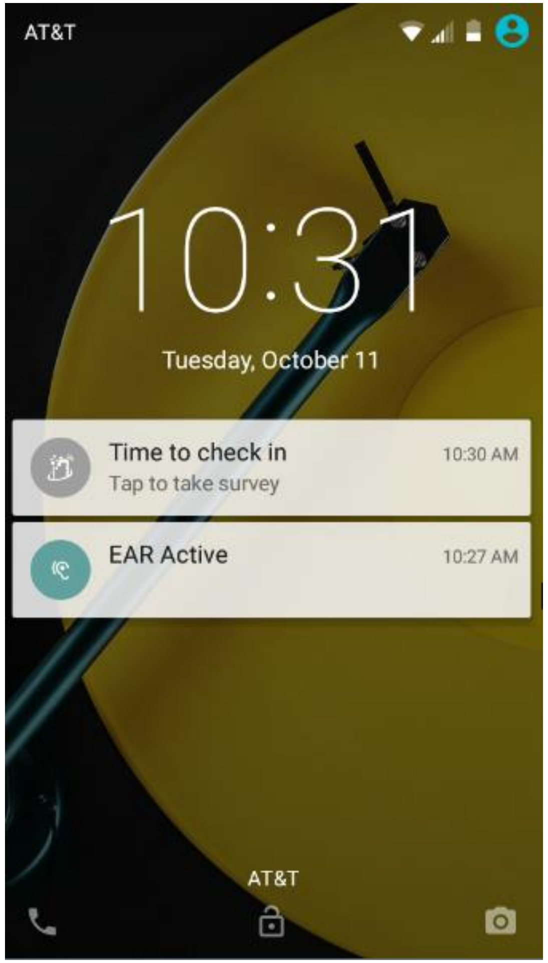
- Microphone
  - “Actively” records answers to survey questions
  - “Passively” measures ambient noise (e.g., clutter), music, and conversations
  - To preserve privacy, classifiers determine that participant is, for example, “around conversation” but not able to reconstruct content or to identify individual speakers
- Light sensor
  - Used to adjust display brightness
  - In combination with other sensors (e.g., accelerometer, microphone) infers idle state of phone/user & sleep

Source: https://www.theverge.com/circuitbreaker/2017/9/15/16307802/apple-iphone-x-features-specs-best-worst

Example: *Daily Experiences and Well-being Study*

(Fingerman et al. 2020)

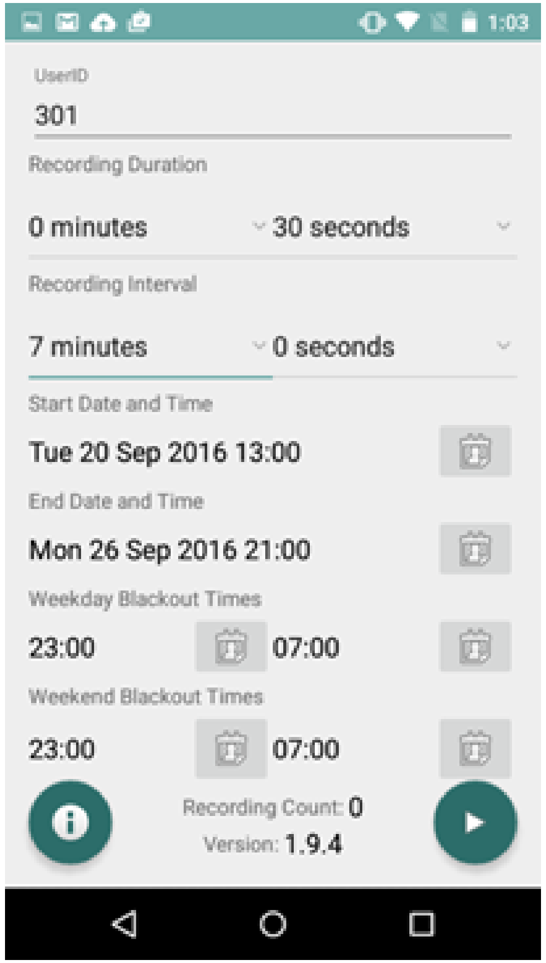
- Telephone screening to identify home-dwelling aged 65+ in Austin, TX (n=333)
    - Oversample of Blacks and Hispanics
    - Without cognitive impairment, not working full-time
- Goal: Study influence of social engagement on physical activity, health, and cognitive status
- In-home interview followed by 5 days of:
    - Actigraphy
    - Loaner Android device with apps to <u>record sound</u> and prompt for ecological momentary assessment (EMA) - no other smartphone functionality
- Daily reminder phone calls & in-home assistance



Example: Electronically Activated Recorder (EAR)

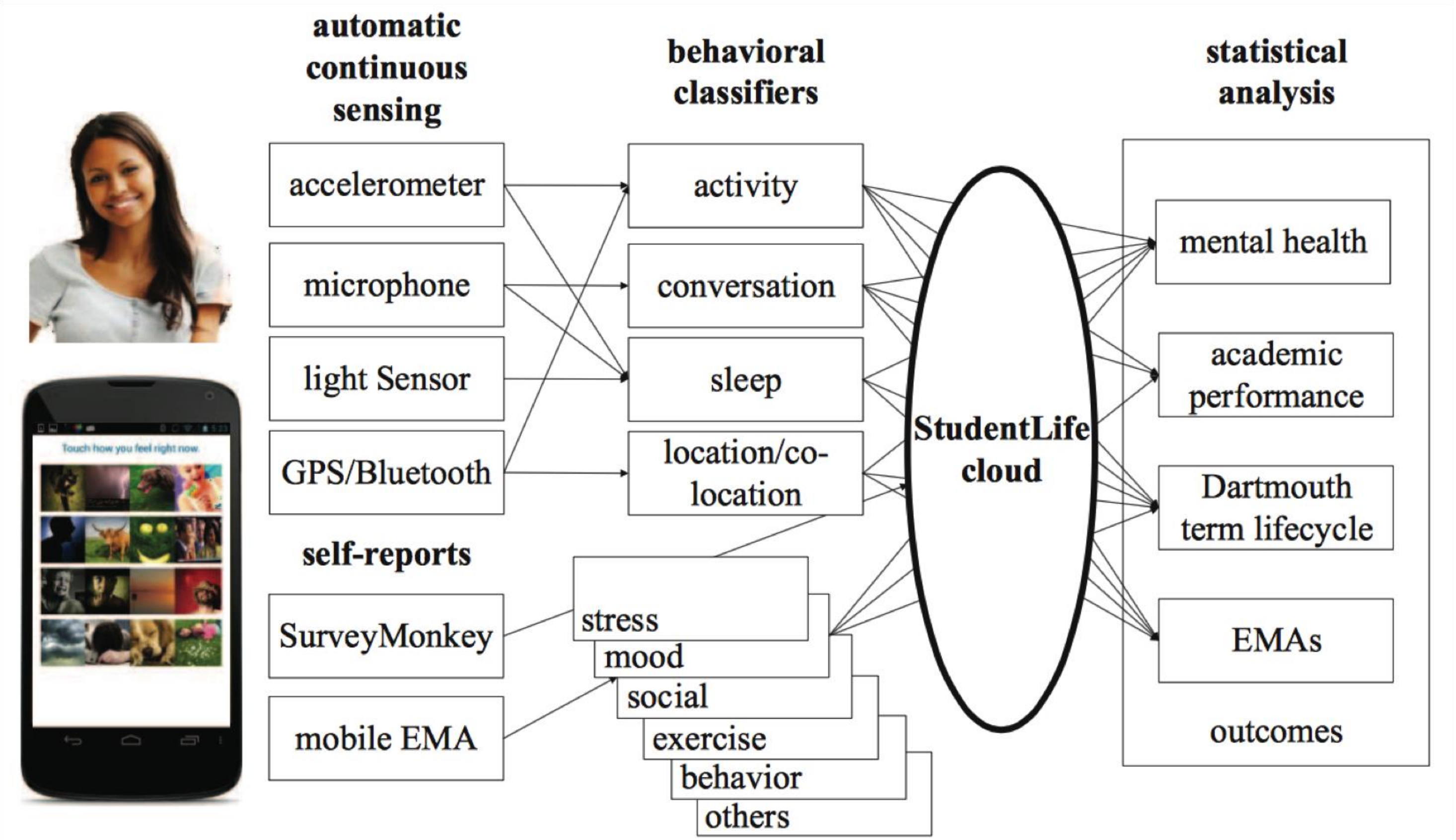
(Fingerman et al. 2022)

- During in-home interview, interviewers entered settings in *EAR* app on phone
  - 30s of recordings every 7 min during waking hours
  - Total of 135,078 audio files
- Devices obtained by interviewer on day 5
- Interviewers responsible for upload and transfer of all data from various devices
- Coders rated each file containing sound for presence of television
- Findings:
  - More TV watching when alone
  - More loneliness reported during periods of TV watching



Does mental health of students change over the course of a term? (Wang et al. 2014)

- 48 students (U.S. college)
- 10 weeks
- Android phones
  (37 provided, 11 own)
- EMA 8 times a day
- Pre- and post-survey



Does mental health of students change over the course of a term? (Wang et al. 2014)

- Students who sleep less, interact less with other students, have fewer co-locations with others more likely to be depressed
- Students around more conversation and students who move around less while on campus do better academically

Correlation with depression

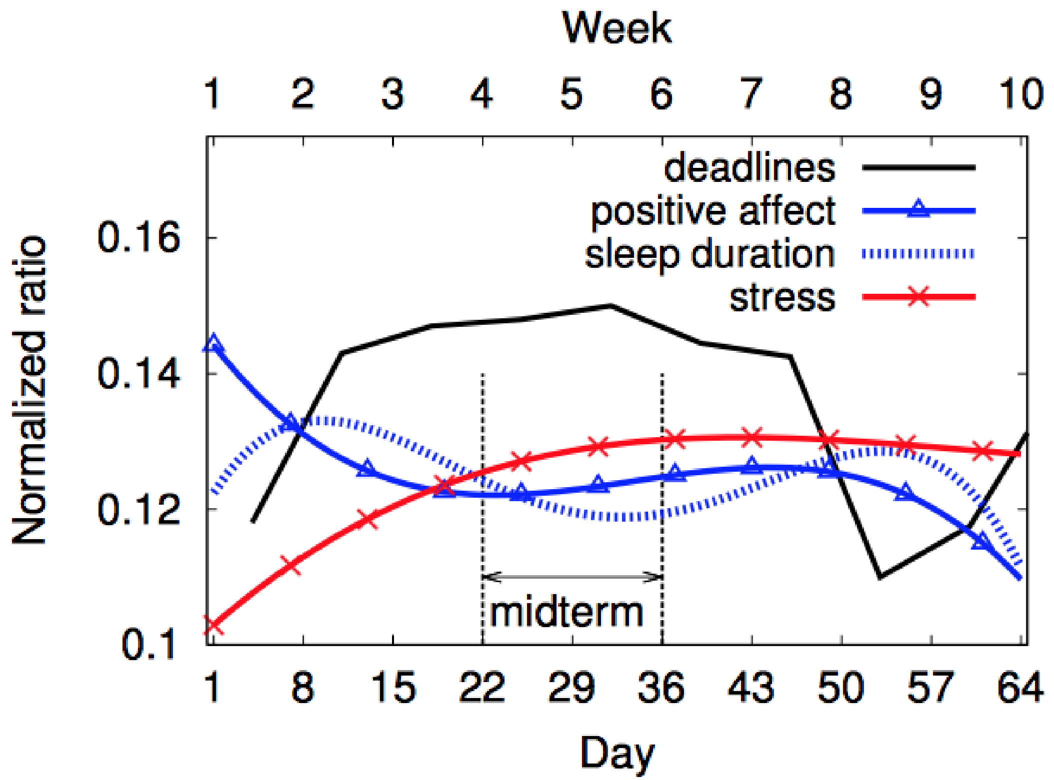
automatic sensing data	r	p-value
sleep duration (pre)	-0.360	0.025
sleep duration (post)	-0.382	0.020
conversation frequency during day (pre)	-0.403	0.010
conversation frequency during day (post)	-0.387	0.016
conversation frequency during evening (post)	-0.345	0.034
conversation duration during day (post)	-0.328	0.044
number of co-locations (post)	-0.362	0.025

Correlation with academic performance

academic performance	Sensing Data	r	p-value
spring GPA	conversation duration (day)	0.356	0.033
spring GPA	conversation frequency (day)	0.334	0.046
spring GPA	indoor mobility	-0.361	0.031
spring GPA	indoor mobility during (day)	-0.352	0.036
spring GPA	indoor mobility during (night)	-0.359	0.032
overall GPA	activity duration	-0.360	0.030
overall GPA	activity duration std deviation	-0.479	0.004
overall GPA	indoor mobility	-0.413	0.014
overall GPA	indoor mobility during (day)	-0.376	0.026
overall GPA	indoor mobility during (night)	-0.508	0.002
overall GPA	number of co-locations	0.447	0.013

Does mental health of students change over the course of a term? (Wang et al. 2014)

- Start of term: high positive affect and conversation levels, low health, healthy sleep, and daily activity
  patterns
- As term progresses: stress rises; activity, sleep, conversation, and positive affect, visits to the gym and attendance drop



(a) EMA and sleep data

Proximity - Bluetooth

- Short-range communication between devices up to 30
  - e.g., hands-free devices, audio speakers, printers
- Enabled healthcare devices can connect to smartphones or other hubs to transmit data
  - e.g., weight, blood pressure, temperature, heart rate, etc.
- Beacons = small Bluetooth transmitters
  - Need to be dispatched by researcher
  - Bluetooth needs to be activated on receiving device
  - Great for indoor tracking

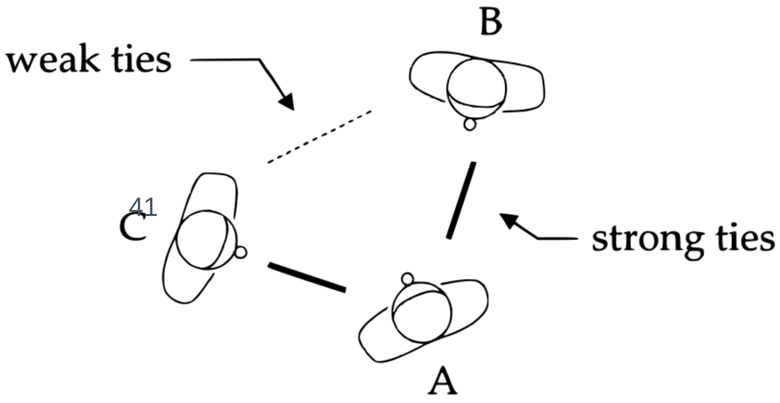
Source: https://www.renesas.com/jp/en/solutions/proposal/bluetooth-low-energy.html

Source http://www.fenc.com/dynafeed/

Source: Silvana Jud

Proximity - RFID & NFC

- Radio-frequency identification (RFID): electromagnetic fields to automatically identify and track tags attached to objects ~1 meter (3 feet)
  - e.g., assembly lines, merchandise in warehouses, livestock
- Near-field communication (NFC): communication between devices by bringing them within 4 cm (1.6 in) of each other
  - More secure than RFID
  - e.g., contactless payment, data transfer, key cards
- All of them (incl. Bluetooth) can be used to track "social ties"



Source: https://upload.wikimedia.org/wikipedia/commons/2/2a/Weak-strong-ties.svg

Example: How do people interact in large social networks?

(Stopczynski et al. 2014)

- *Copenhagen Networks Study*: 1,000 smartphones handed out to Danish students
- Extensive questionnaire upon enrollment: 310 questions on topics from public health, psychology, anthropology, and economics
- Combination of Bluetooth and Wi-Fi networks to collect information about absolute location and relative location to each other
  - Additional data sources: call and text logs, social media data

Example: How do people interact in large social networks?

(Stopczynski et al. 2014)

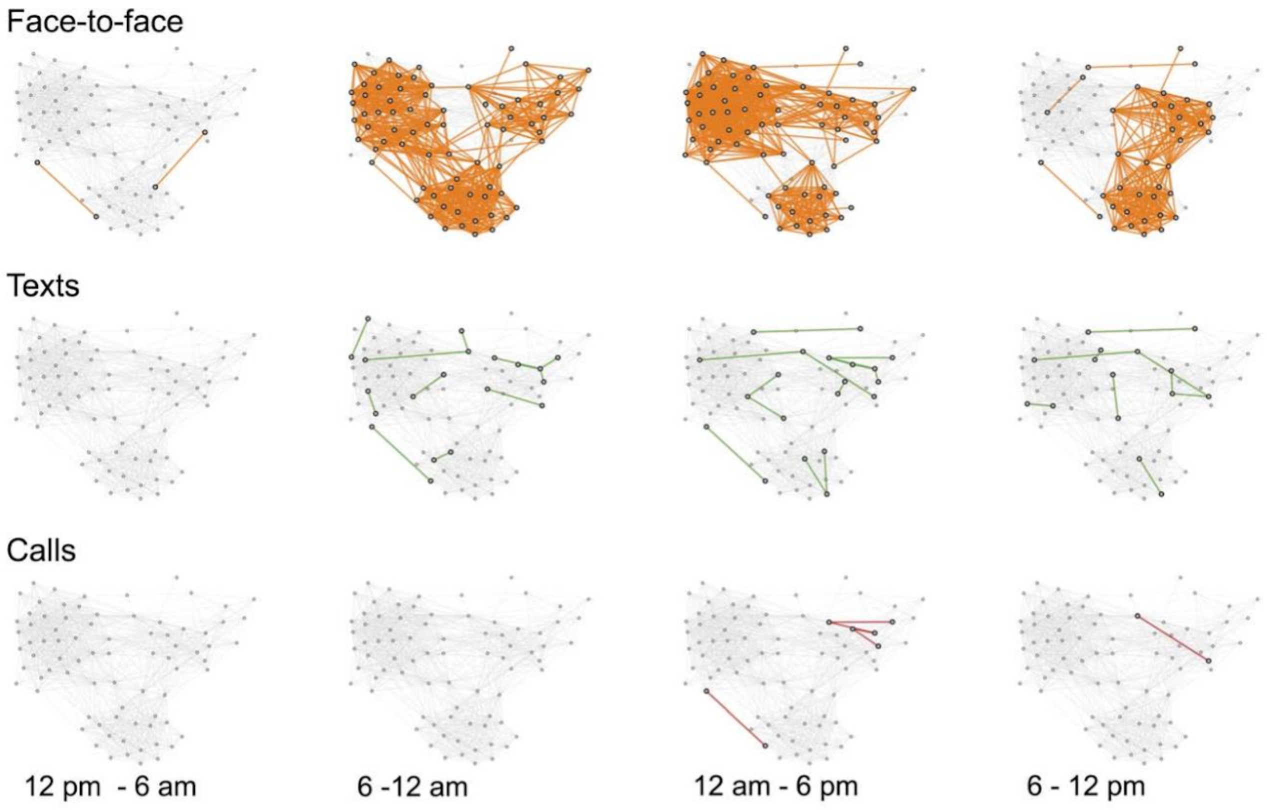
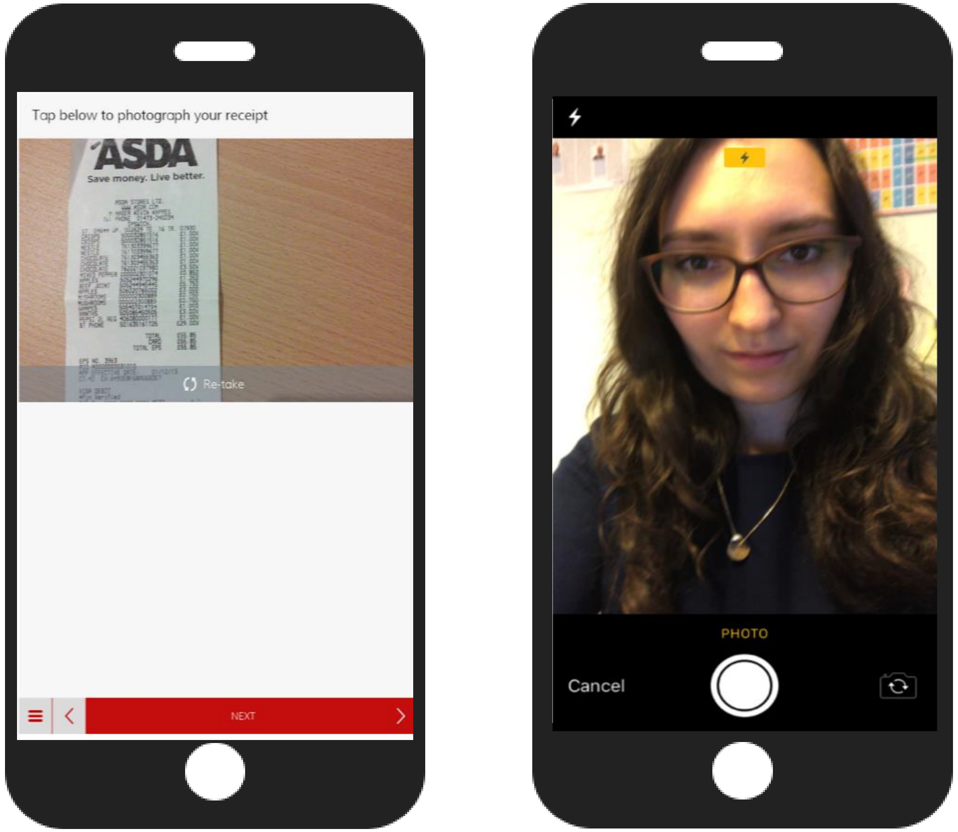


Figure 11. Daily activations in three networks. One day (Friday) in a network showing how different views are produced by observing different channels.

Images

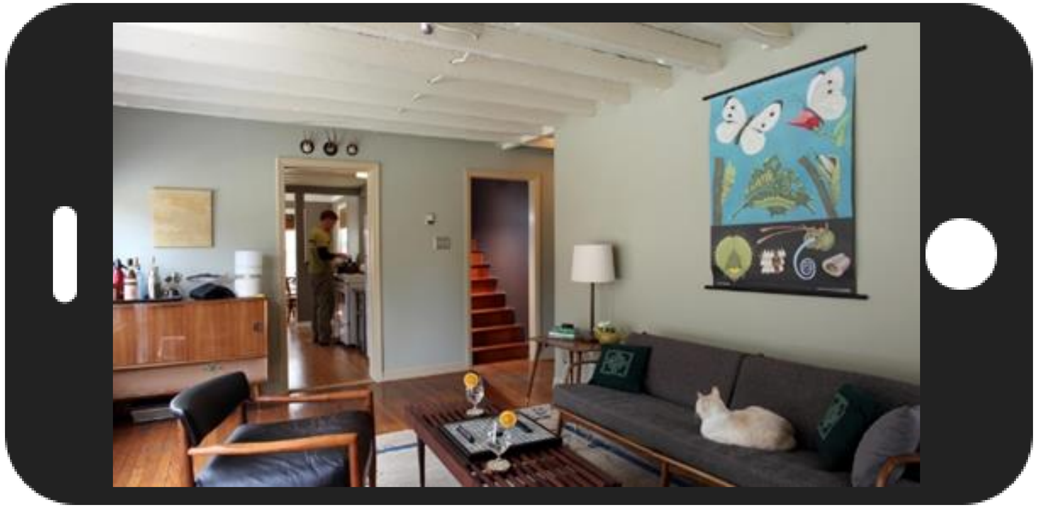
- Photos
  - Food, receipts, physical surroundings, etc.
- Video
- Barcodes
- Linear distance (iPhone Measure app)



Jäckle et al. (2018)

Example: Taking pictures of surrounding

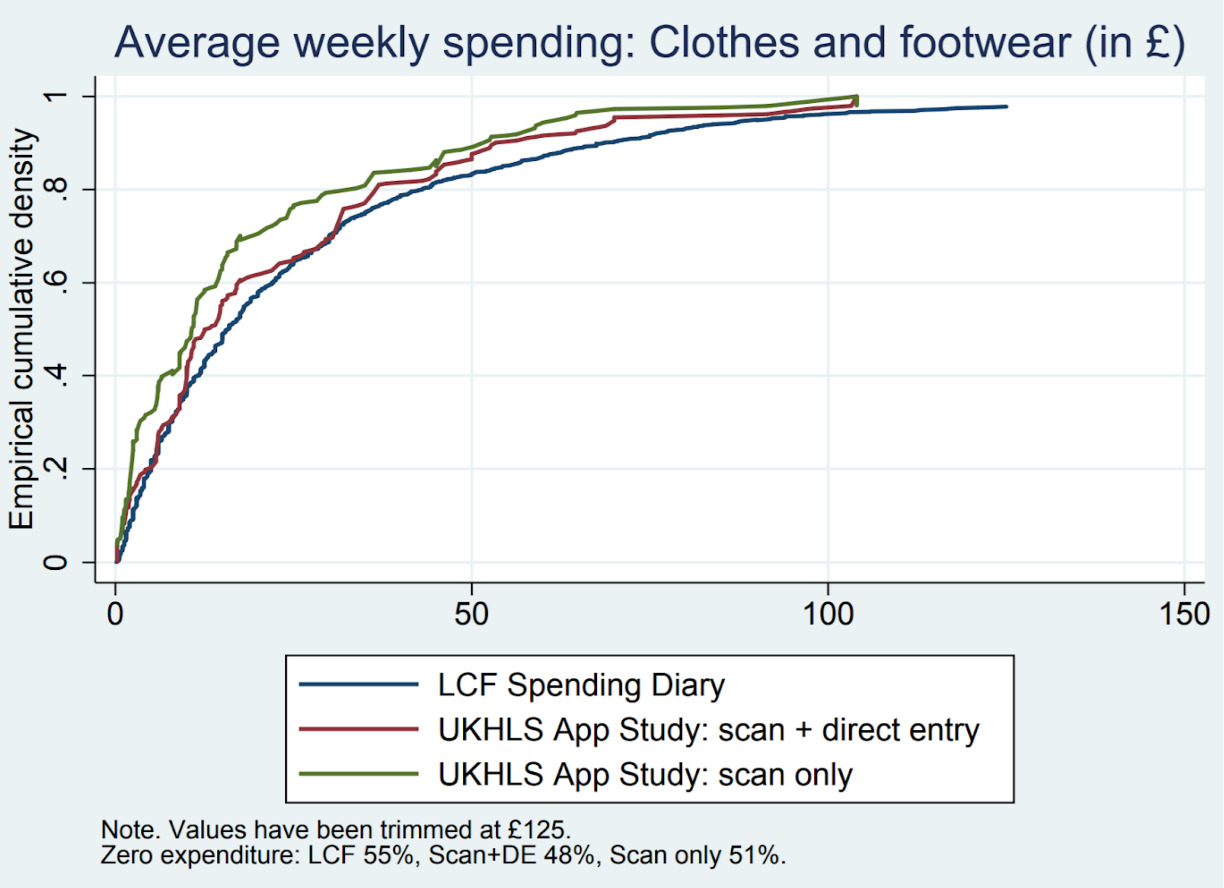
- *Daily Experiences and Well-being Study* (Fingerman et al. 2020)
- Interviewers used phone app when returning to pick device up (day 5)
- After completing all other activities, asked participant for consent to take picture of room they spend most time in
  - Up to 3 photos
  - Careful selection of motive to avoid recording any PII
- Environmental conditions of room hand-coded
  - Lighting, conditions, etc.

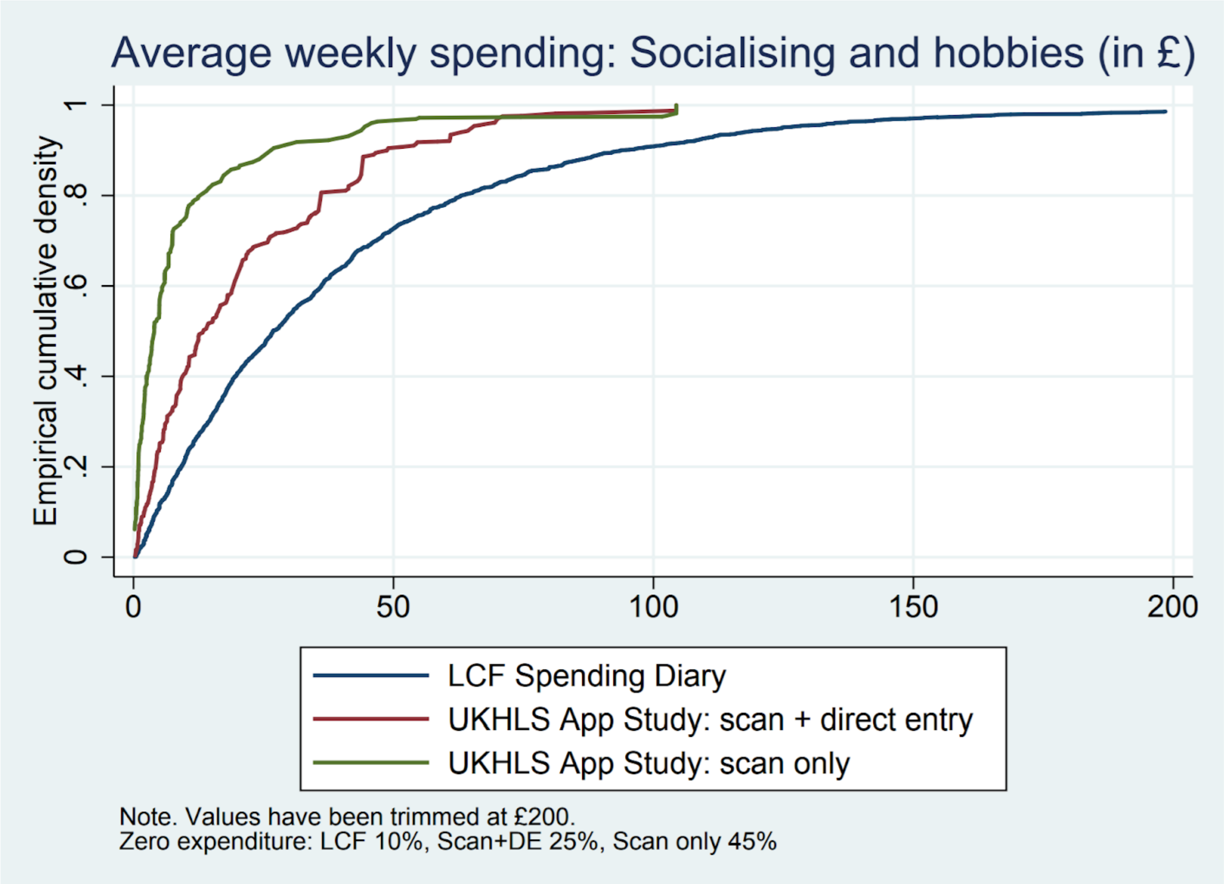


Example: How much do households spend on goods and services? (Jäckle et al. 2019; Wenz et al. 2018)

- UK Innovation Panel Wave 9 participants invited to download app (iOS & Android) to smartphone or tablet and use it to report purchases of goods and services for 1 month
  - 270 participants (13%) used app at least once
- Participant could scan and upload receipts, record purchase without receipt, report day without purchases
  - App sent push notifications once a day
- Scanned receipts hand coded
- Total expenditure (scan + direct entry) comparable to benchmark (LCF)
  - Expenditure more comparable for some categories than for others

Example: How much do households spend on goods and services? (Jäckle et al. 2019; Wenz et al. 2018)





Self-reports on smartphones

- Diary studies
  - e.g., time use, expenditure, food consumption via app or web browser

Daily overview

Adding activities

Adding activity information

Elevelt et al. (2019)

Self-reports on smartphones

- Diary studies (e.g., time use, food consumption) via app or web browser
- Ecological Momentary Assessment (EMA)/Experience Sampling Method (ESM) via app
  - Collecting data several times a day on several days per week allows tracking of change within individuals in much detail
  - Immediate reporting increases ecological validity
  - Participants “pinged” to report about current circumstances
    - Objective situation: e.g., “What are you doing?”
    - Subjective state: e.g., “How anxious are you right now?”
  - Time-based vs. geolocation-based vs. event-based

Example time-based EMA: How do environmental factors affect happiness?

(MacKerron & Mourato 2013)

- *Mappiness* app installed by ~22,000 self-selected iPhone users and used up to 6 months
- EMA questions: how happy, relaxed, and awake users feel and whom they were with at two or more random points during the day
- Physical setting measured by GPS, appended with information from objective spatial data (broad habitat and land cover type, weather conditions, and daylight status)
- On average, participants significantly and substantially happier outdoors in all green or natural habitat types than in urban environments

Example time-based EMA: How does physical activity affect happiness? (Lathia et al. 2017)

- *Mood-Tracking Application* on smartphones of 12,000 volunteer Android users for up to 17 months
- EMA questions: affect two or more times during the day
- Physical activity for immediately preceding fifteen minute period measured both by self-report (EMA) and passively by accelerometer

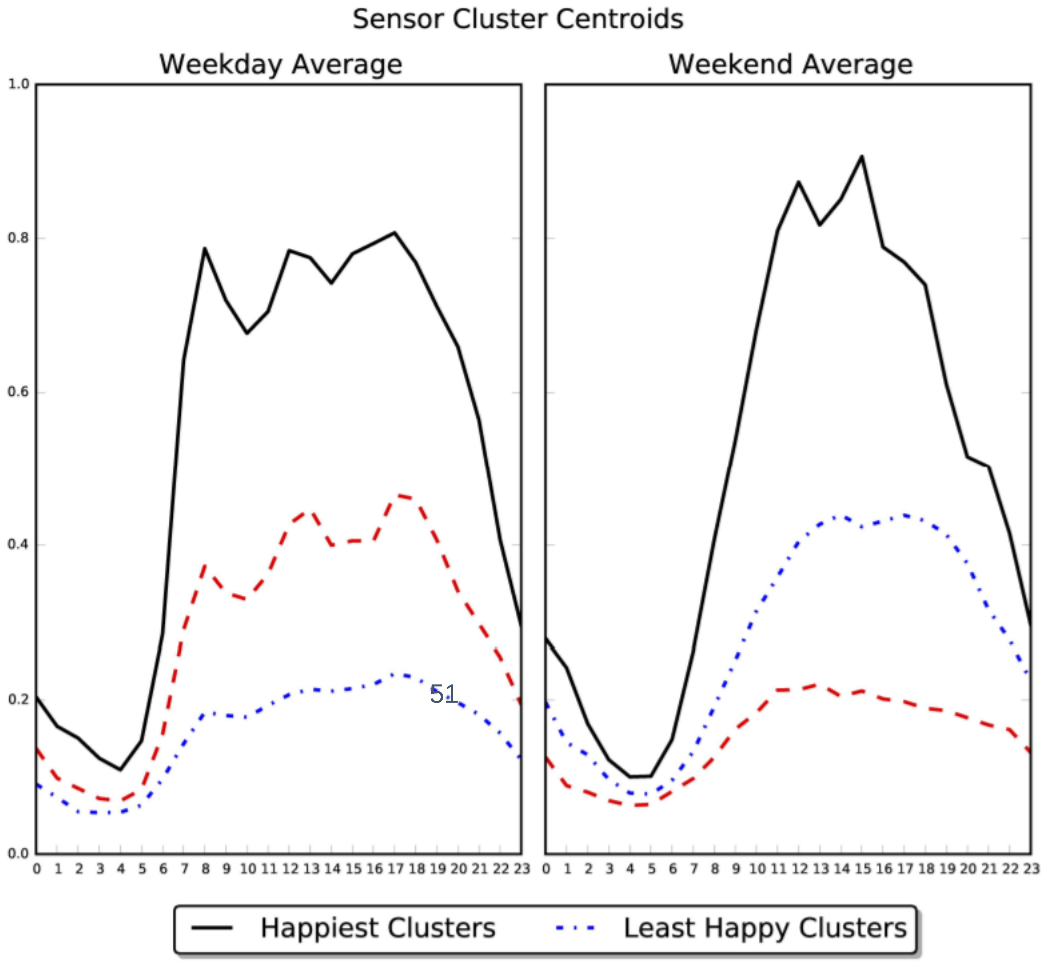
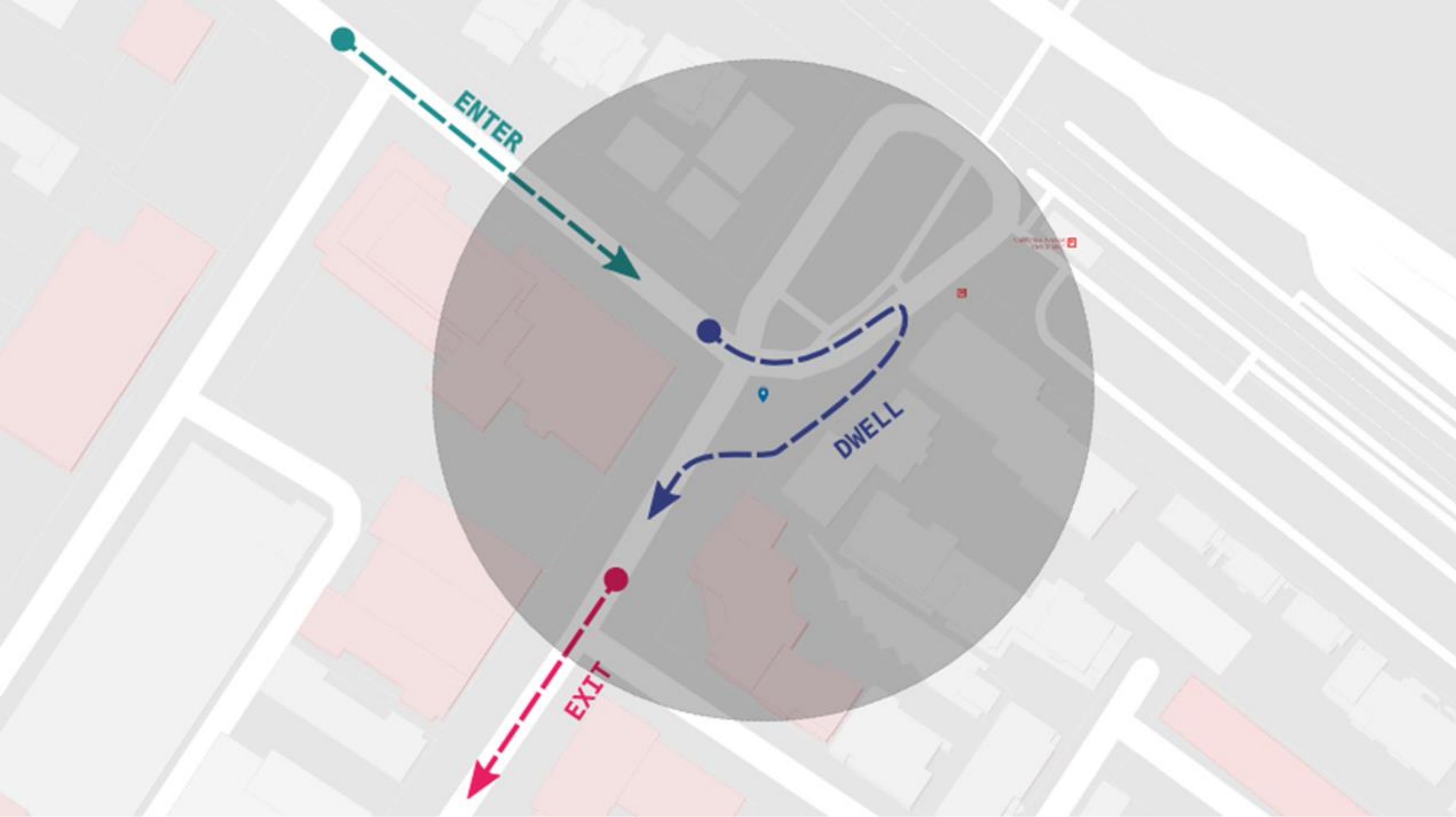


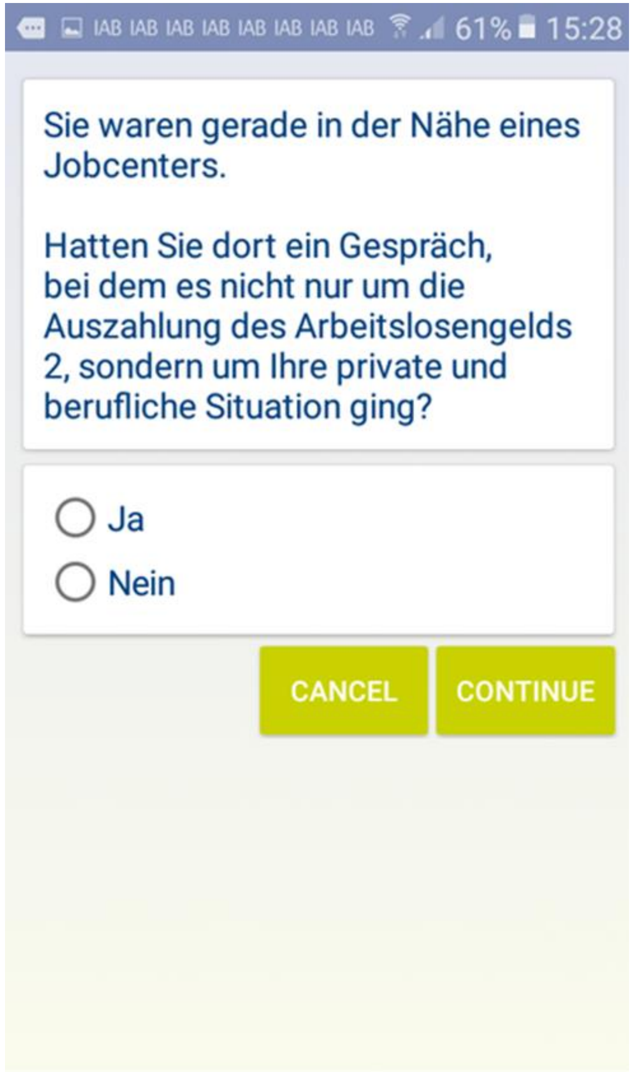
Fig 3. Centroids for the clusters generated from (left) weekday and (right) weekend activity profiles.

doi:10.1371/journal.pone.0160589.g003

Example geolocation-based EMA ("Geofencing"): Visits to job centers (Haas et al. 2020)

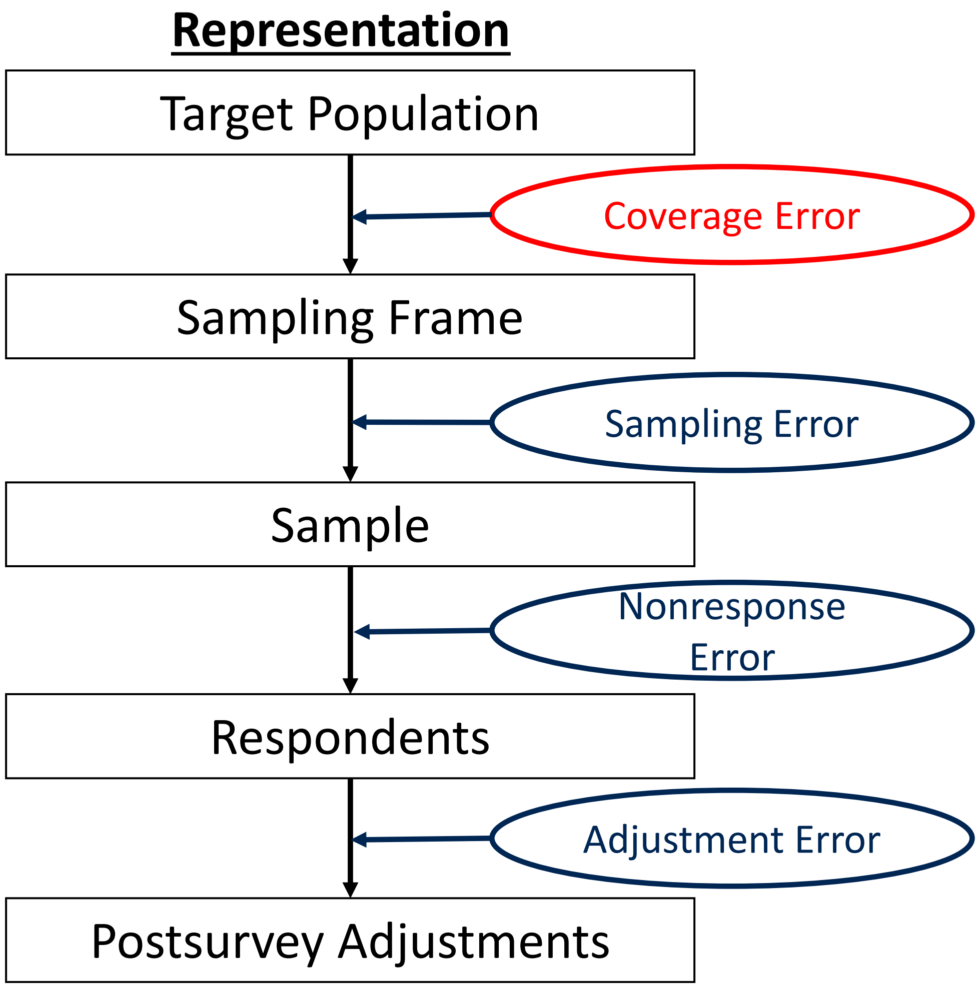


Source: https://developers.google.com/location-context/geofencing/



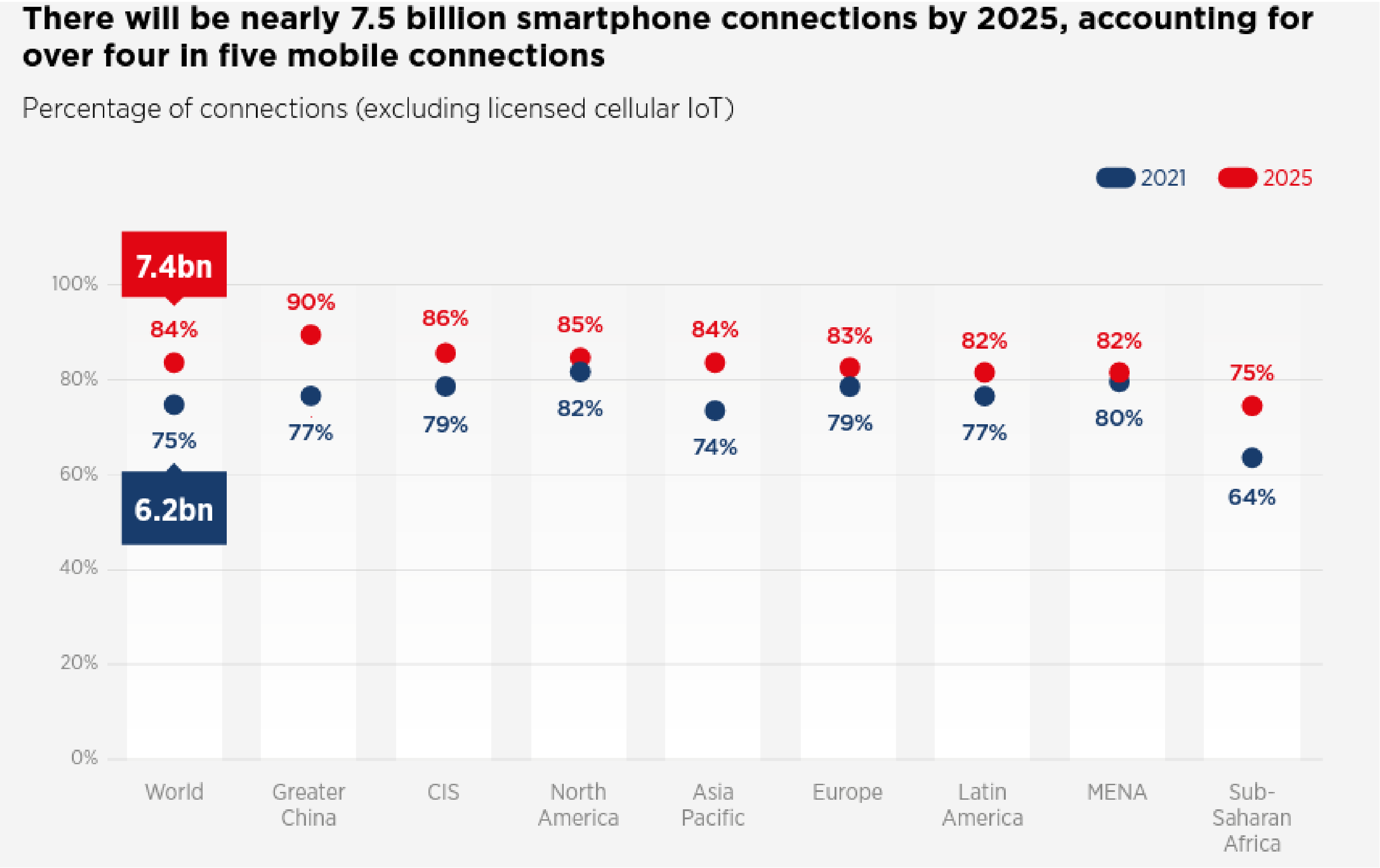
Practical implementation

Representation error in app, sensor & wearables data collection



- **Coverage error:** A study relies on participants to share data from their fitness wristbands to analyze weekend vs. weekday activity by race & ethnicity. The rate of ownership of these devices is lower in the study population than in the general population.

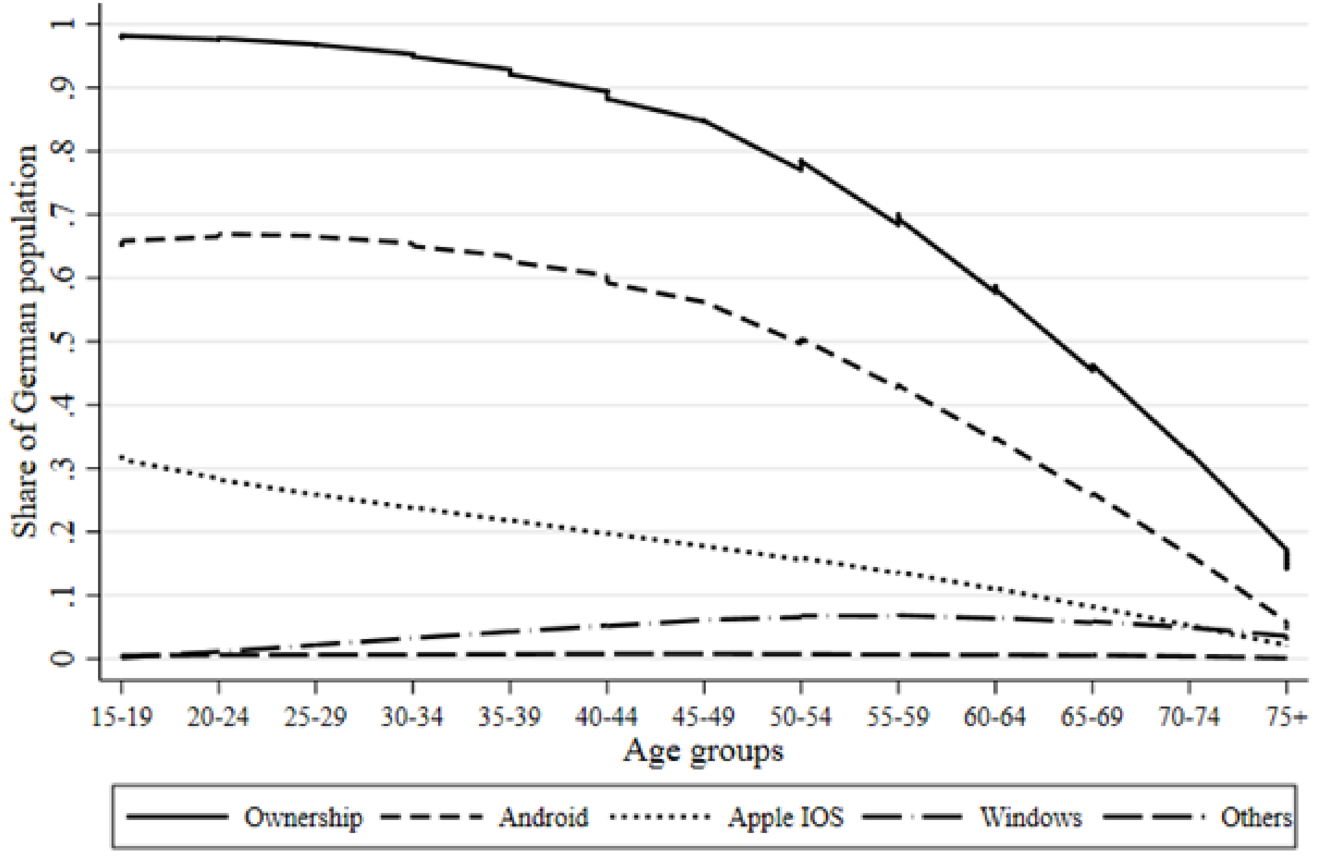
BYOD: Coverage smartphones



Source: https://www.gsma.com/mobileeconomy/wp-content/uploads/2022/02/280222-The-Mobile-Economy-2022.pdf

BYOD: Smartphone coverage bias in Germany

(Keusch et al. 2020)



- Smartphone ownership higher among…
  - …younger
  - …male
  - …higher educated
  - …people in New States
  - …people living in larger communities
- ➢ Digital Divide

Source: PASS Wave 11; n = 13,703; Locally weighted scatter-plot smoother (LOWESS) regression

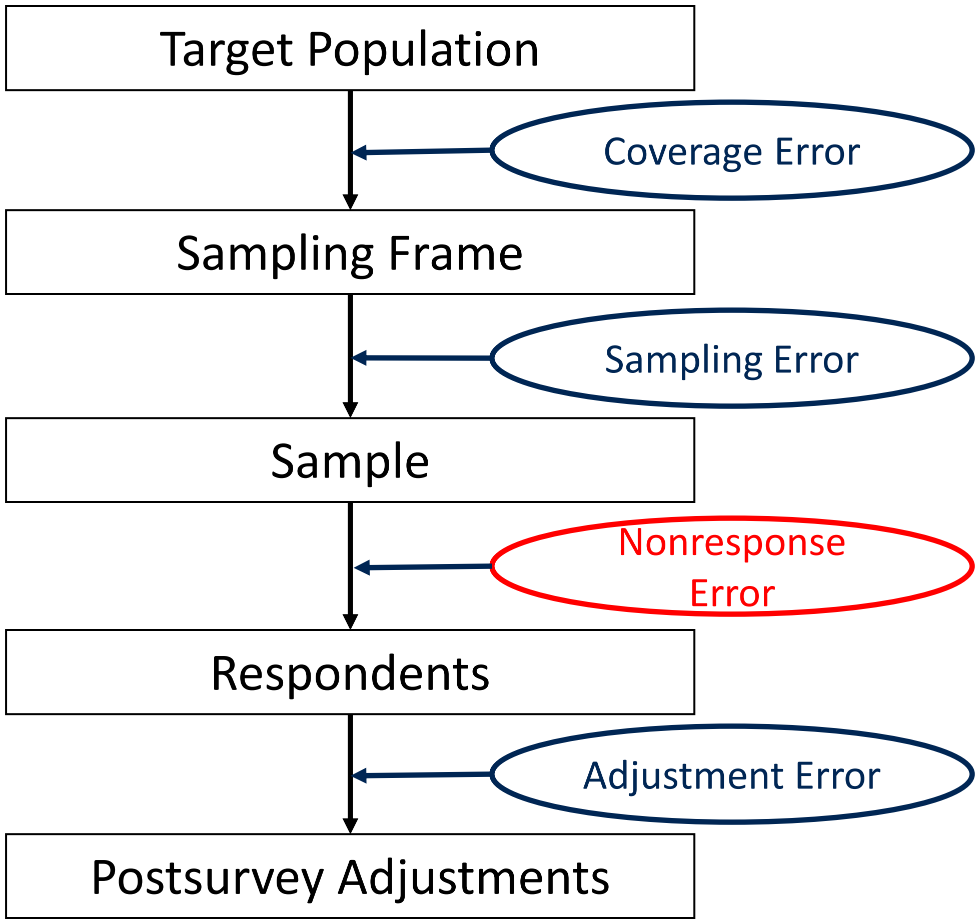
BYOD: Smartphone coverage bias in Germany

(Keusch et al. 2020)

- *Overall smartphone coverage bias* in many substantive estimates relatively small; especially once adjusting for sociodemographic differences between general population and smartphone owners
  - High social inclusion: +2.8 p.p.
  - Size of personal network: n.s.
- Comparable *Android smartphone coverage bias* after sociodemographic adjustment
  - High social inclusion: +1.6 p.p.
  - Size of personal network: n.s.
- Much larger *iPhone coverage bias*, even after adjusting for sociodemographics (up to 11 p.p.)

Representation error in app, sensor & wearables data collection

Representation



- **Nonparticipation error:** Participants are provided with actigraphs to measure sleep patterns for a week. Those who do not sleep well remove the device at night because it disturbs their sleep.

Nonparticipation: Willingness to participate (WTP) & actual participation

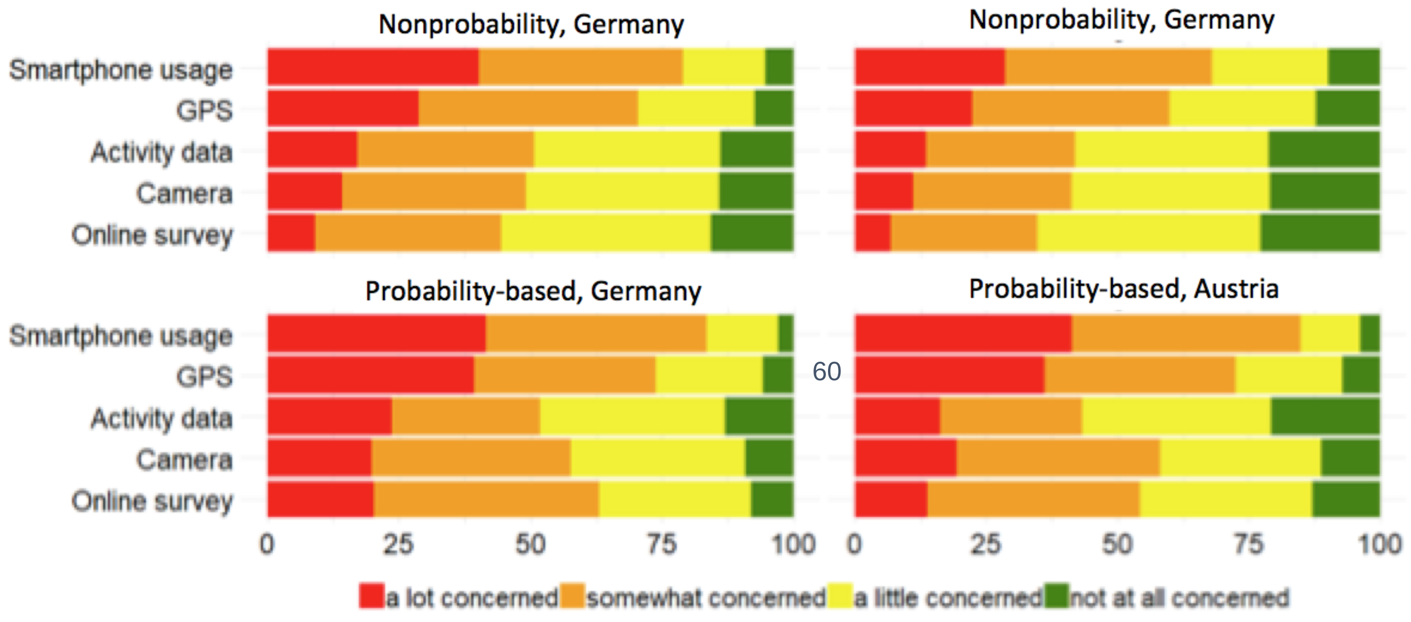
- Many studies work with self-selected volunteer samples
- General willingness varies by sensor and task
  - 29% (Spain) - 52% (Mexico) for taking pictures, 19 (Portugal) - 37% (Chile) for sharing GPS location (Revilla et al. 2016)
  - Mobility (GPS, accelerometer): 37% willing (81% participate) (Scherpenzeel 2017)
  - Physical activity (wearables): 57% willing (90% participate) (Scherpenzeel 2017)
- Willingness/Downloading research app in general population usually rel. low
  - 35% downloaded & registraterd CBS Travel app (McCool et al. 2021)
  - 24% registered in the Household Budget app (Rodenburg et al. 2022)
  - 18% would install app to track URLs of visited websites (Revilla et al. 2019)
  - 17% downloaded UK Understanding Society IP budget app (Jäckle et al. 2019)
  - 15% downloaded IAB-SMART app (Keusch et al. 2022a)

Mechanisms of (non-)participation: Privacy concern

- Participants might have concerns about potential risks related to sensor data
  - Data streams could be intercepted by unauthorized party
  - Connecting multiple streams of data could re-identify previously anonymous users
  - Information could be used to impact credit, employment, or insurability

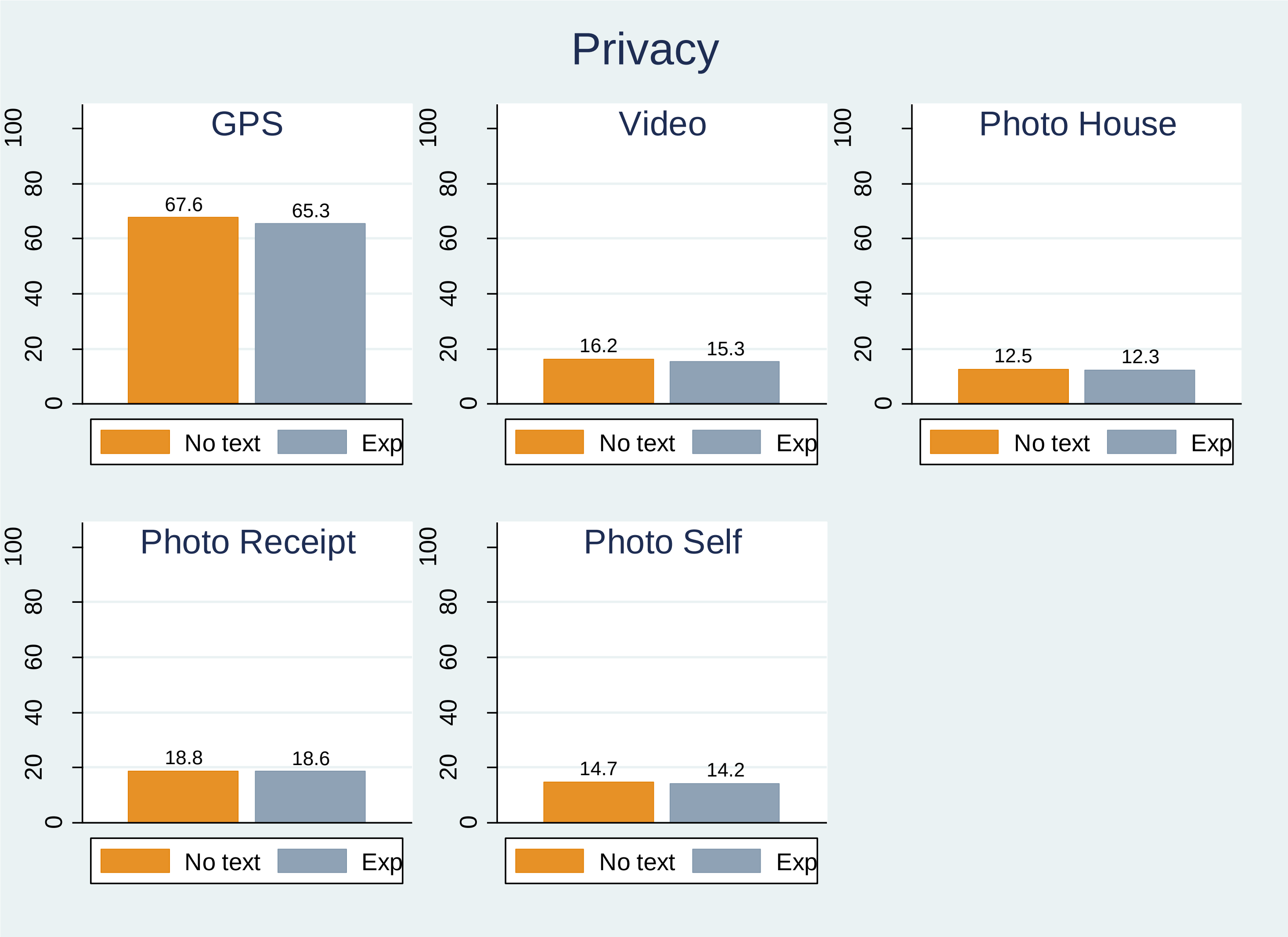
- Higher privacy & security concerns correlate with lower WTP

(Keusch, et al. 2019; Revilla et al. 2019; Struminskaya et al. 2020; 2021; Wenz et al. 2019; Wenz & Keusch in press)



Keusch et al. (2021)

No Effect of Emphasizing Privacy



n=1883, Dutch smartphone & tablet users

"The data you provide will be treated confidentially. It will only be available to researchers conducting this study and your personal information will not be shared with third parties. The results of the survey will only be made available in the anonymized form. Your data is safe in all of our surveys. From the statistical information by CBS personal information can never be inferred."

- No significant differences

Struminskaya et al. (2021)

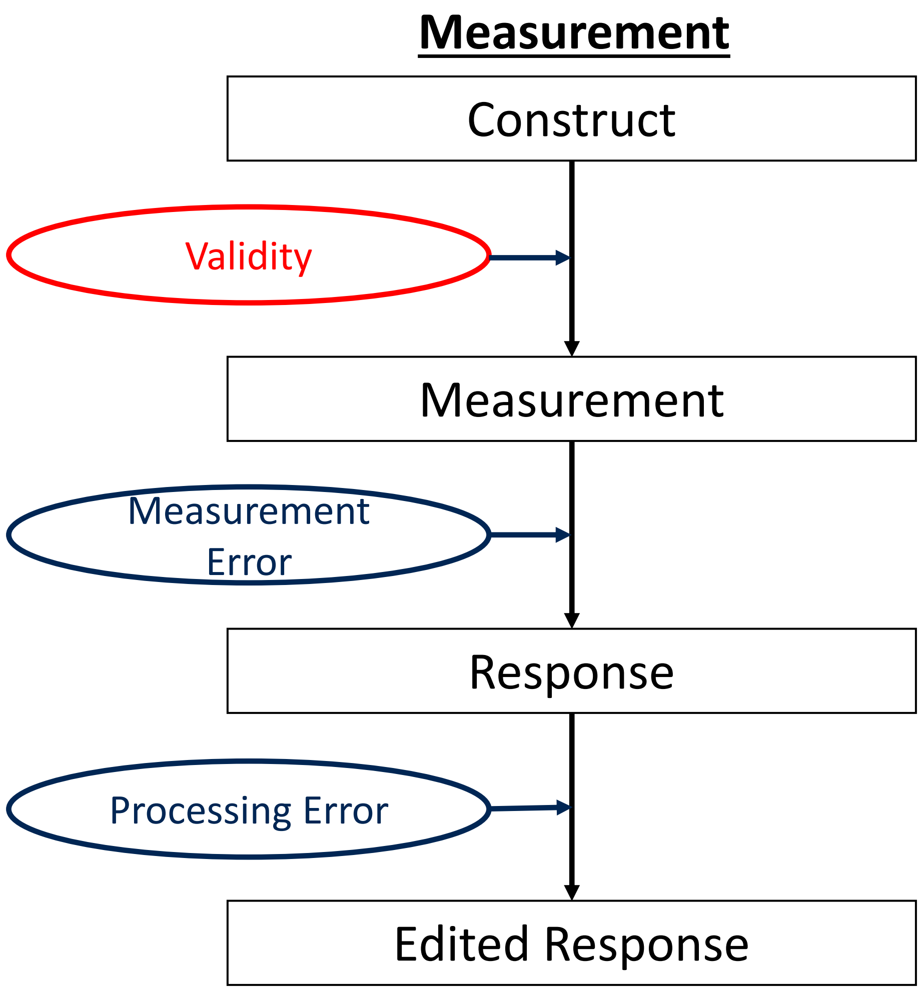
Mechanisms of (non-)participation: Incentives

- Inconsistent findings of effect of incentives on participation
- Hypothetical WTP increases with incentives for downloading app and staying until end of study (Keusch et al. 2019; Wenz & Keusch in press)
- IAB-SMART, Germany (Haas et al. 2021)
  - 20€ for installation increase installation rate over 10€ (16% vs. 13%)
  - Bonus incentive for consenting to all 5 data collection functions no effect
- Statistics Netherlands Travel App (McCool et al. 2021)
  - 5€ unconditional + 5€ Registration + 5€ after 7 days: 30%
  - 5€ unconditional + 10€ after 7 days: 36%
  - 5€ unconditional + 20€ after 7 days: 40%
- UK IP Spending Study (Jäckle et al. 2019)
  - £6 incentive for installation does not increase installation rate over £2

Other Mechanisms on (non-)participation

- Agency: WTP higher for tasks where participants have agency over data collection (Revilla et al. 2019; Keusch et al. 2019; Struminskaya et al. 2020; 2021; Wenz & Keusch in press)
- Sponsor: WTP higher for university sponsor vs. market research and statistical office (Keusch et al. 2019; Struminskaya et al. 2020)
- Framing: emphasizing benefits does not influence WTP (Struminskaya et al. 2020; 2021)
- Smartphone skills: more activities on smartphone (e.g., using GPS, taking pictures, online banking, etc.) correlates with higher WTP (Keusch et al. 2019; Struminskaya et al. 2020; 2021; Wenz et al. 2019; Wenz & Keusch in press)
- Experience: prior research app download increases WTP (Keusch et al. 2019; Struminskaya et al. 2020; 2021)
- Sociodemographics: educational attainment (Jäckle et al. 2019; Keusch et al. 2021, 2022; McCool et al. 2021; Wenz & Keusch in press) and age (Jäckle et al. 2019; McCool et al. 2021; Keusch et al. 2022; Wenz & Keusch in press) correlated with WTP

Measurement error in app, sensor & wearables data collection



- **Validity:** You are using actigraphy to detect intensity of physical activity in a sample of older adults. Your study population is very sedentary and it is difficult to identify physical activity versus usual activity.

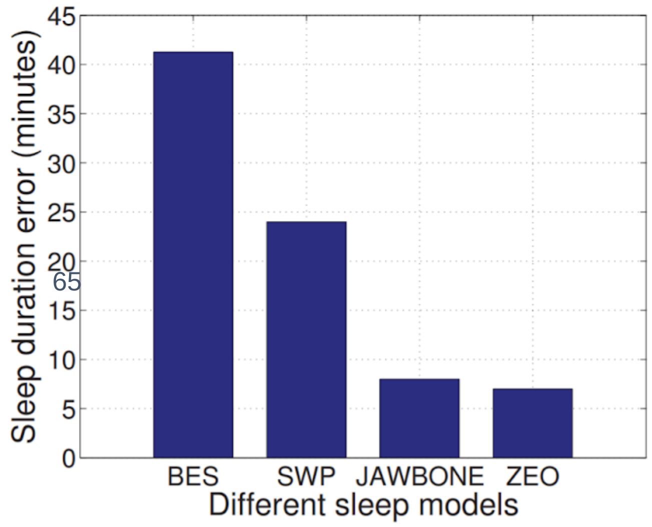
Can you really measure sleep using sensors?

- Does physical inactivity and low heart-rate equal sleep?

Percent of time that self report differs from accelerometer	
Self_report: asleep; Accelerometer: not in bed	6.8%
Self_report: not asleep; Accelerometer: in bed	14.9%
Self_report: asleep; Accelerometer: not asleep	14.0%
Self_report: not asleep; Accelerometer: asleep	10.5%

Kapteyn et al. (2019)

- Does absence of light, sound, and activity measure by a smartphone equal sleep?
- But for <u>some phenomena</u>, sensors seem to be provide highly valid data

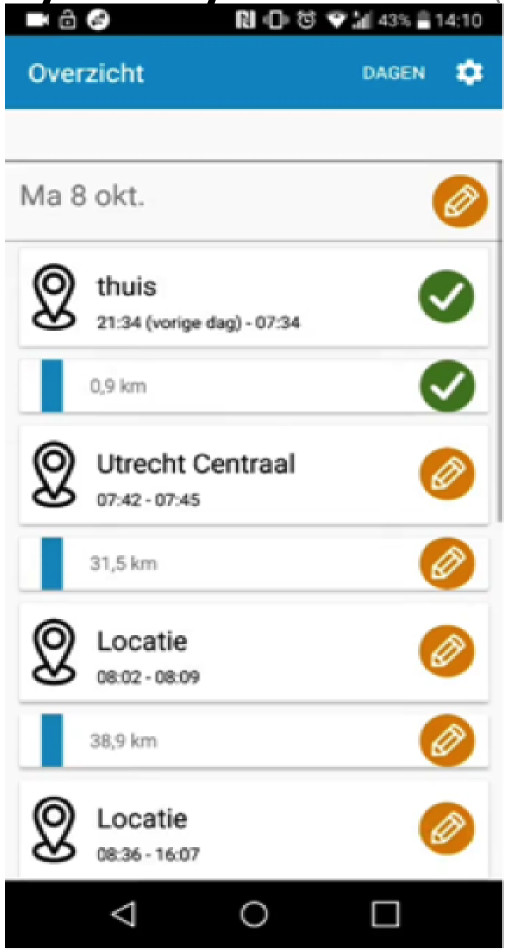


Chen et al. (2013)

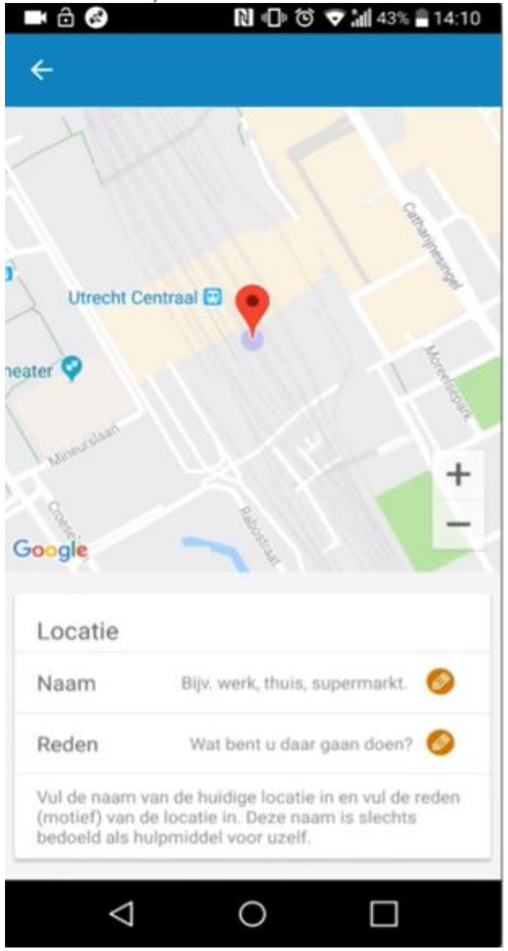
Example: How do people move around in everyday life? (McCool et al. 2021)

- Everyday mobility field test in Dutch general population (Nov-Dec 2018)
  - Statistics Netherlands Travel App (Android & iPhone) [Code for back end & front end]
  - Data collection for 7 days per participant
  - N = 1,902 invited, 674 registered (35.4%)
- Sensing location per second (when moving) & per minute (when still):
  - OS-specific location API implementation (iOS = Core Location, Android = Google Location Services API, fallback: native Android location API)
  - Global Navigation Satellite System (GNSS), local Wi-Fi, Cell towers (for Android ‘high-accuracy mode’ available: including Bluetooth as well)
  - Raw data & processed data stored locally on mobile device
- User annotation: participants provide additional information that helps understand travel behavior (label stops and motives for travel)

Example: How do people move around in everyday life? (McCool et al. 2021)



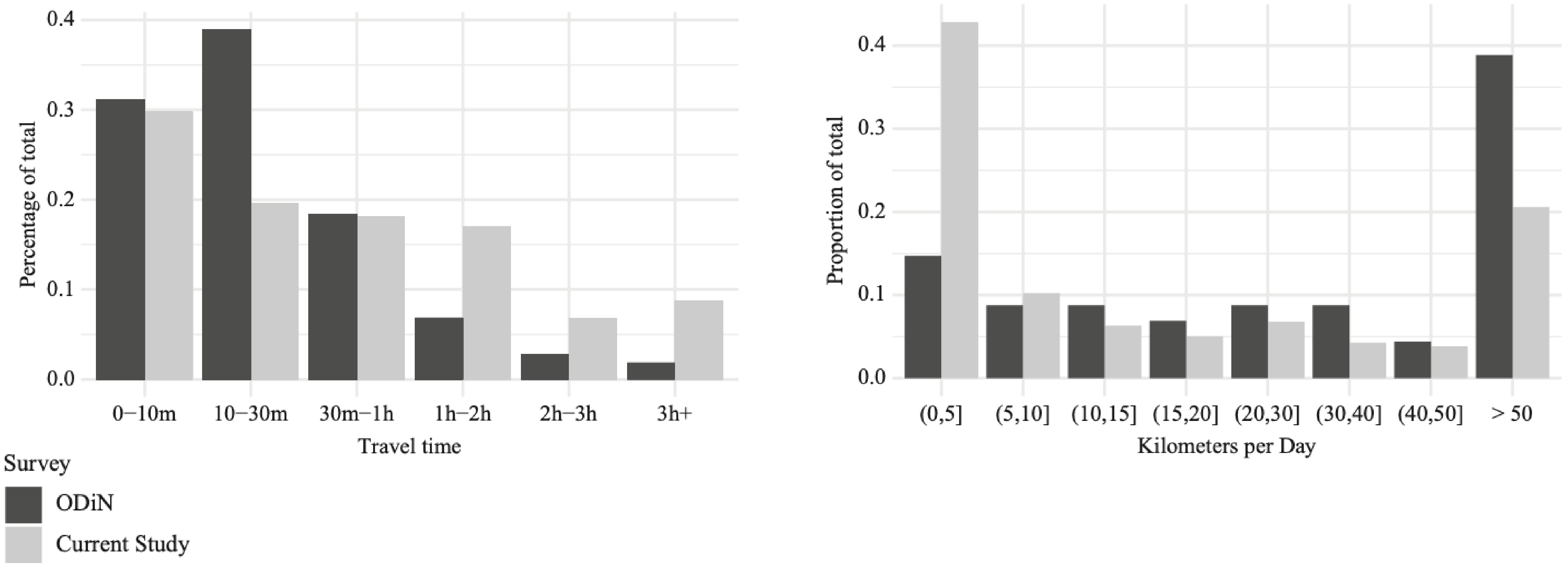
Daily overview



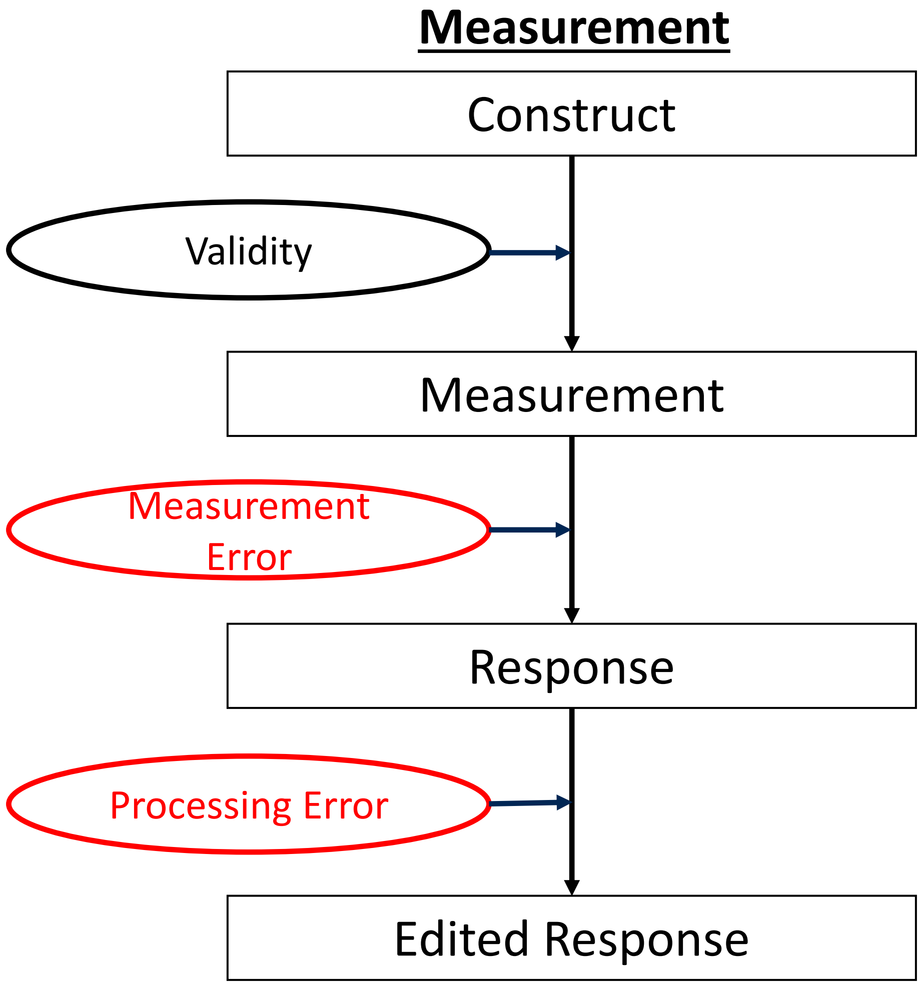
Questions about stops

Questions about trips

Example: How do people move around in everyday life? (McCool et al. 2021)



Measurement error in app, sensor & wearables data collection

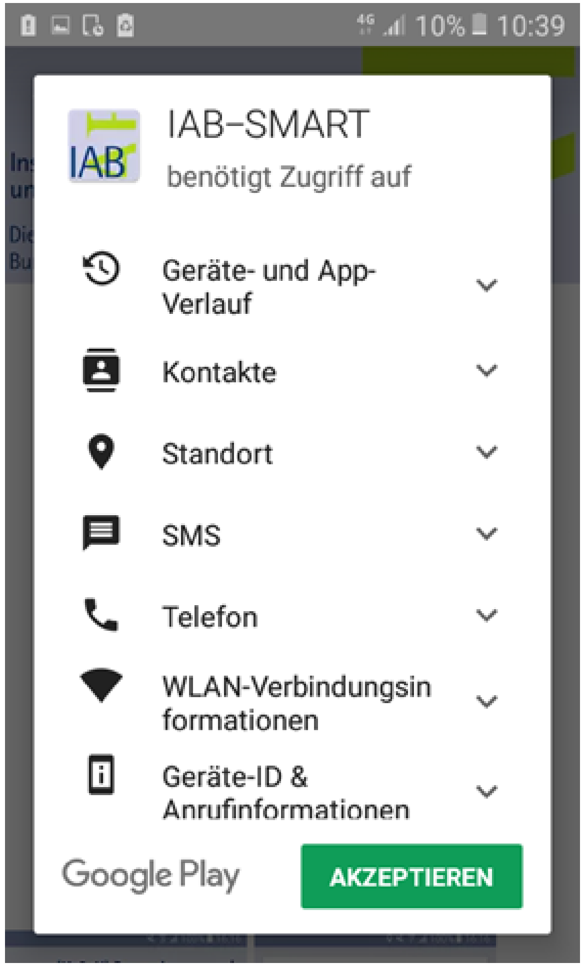


- **Measurement error:** GPS is less precise in urban areas where there are many large buildings.
- **Processing error:** Raw accelerometer data is classified as different types of activity based on where sensor/phone is located (e.g., pocket vs. purse).

Practical implementations and operational considerations

Example consent: IAB-SMART (Kreuter et al. 2020)

Google Play Store

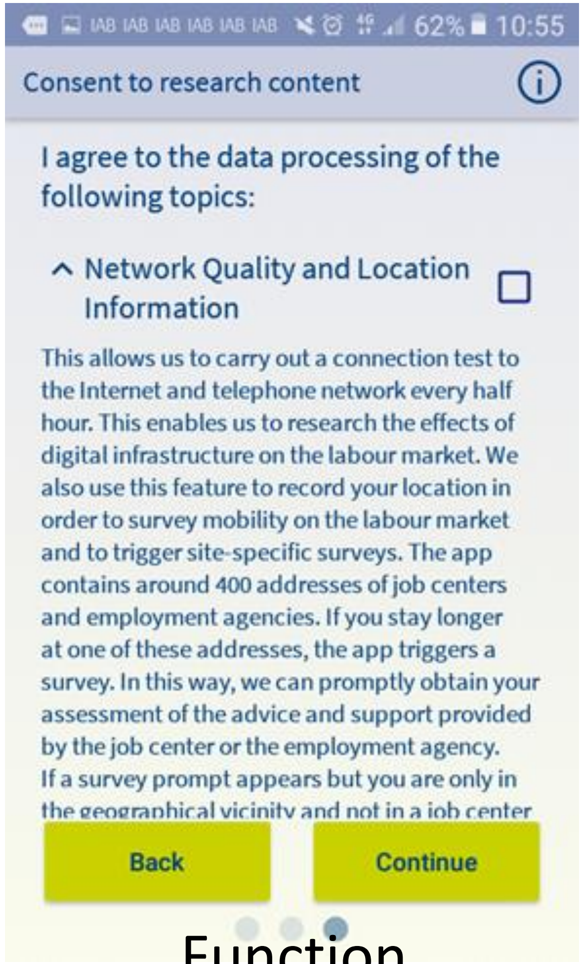


Google Permissions

App installed

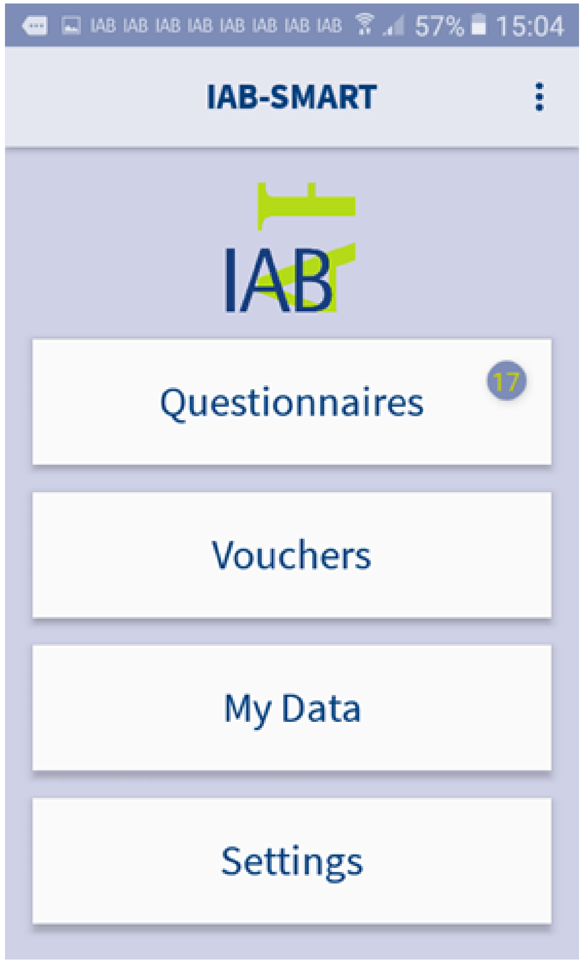
Example for consent: IAB-SMART (Kreuter et al. 2020)

Individual consent screen



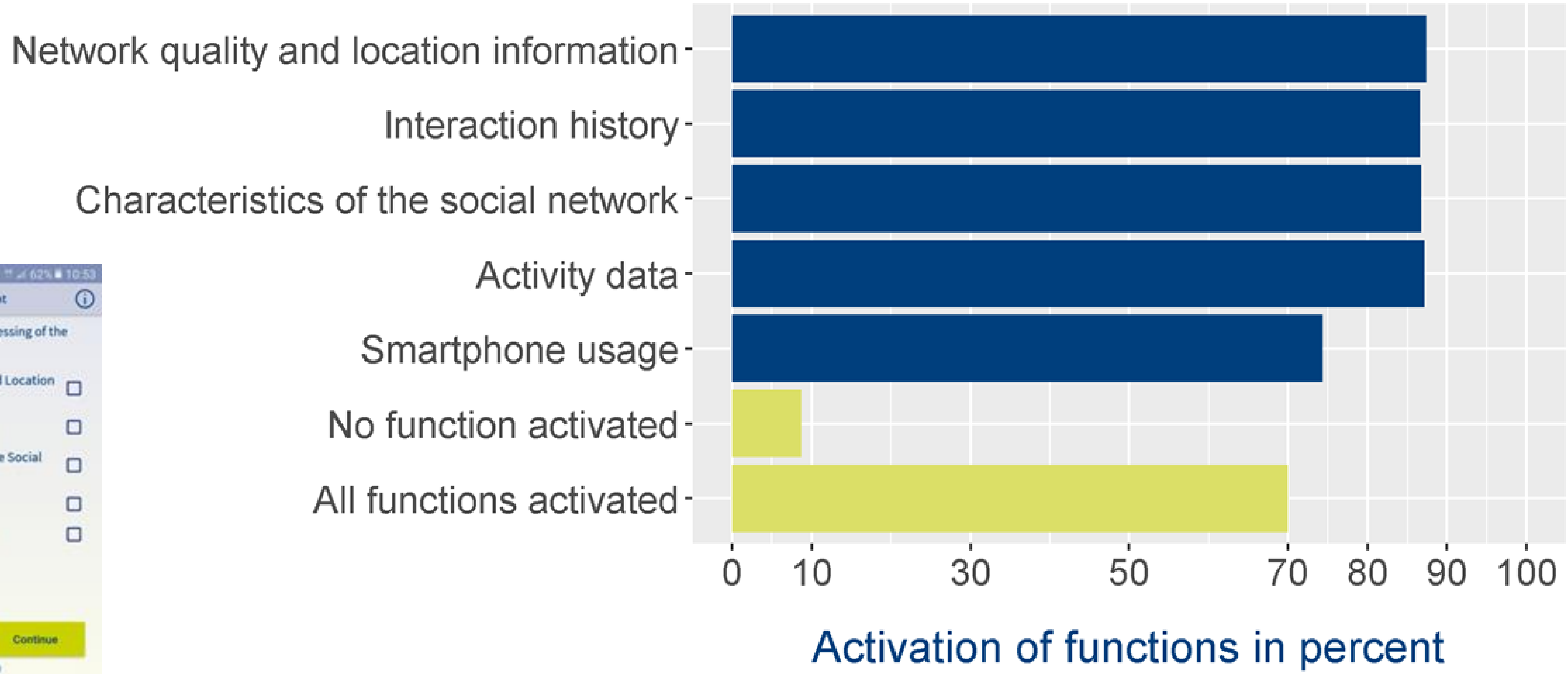
Function explanation

Full consent

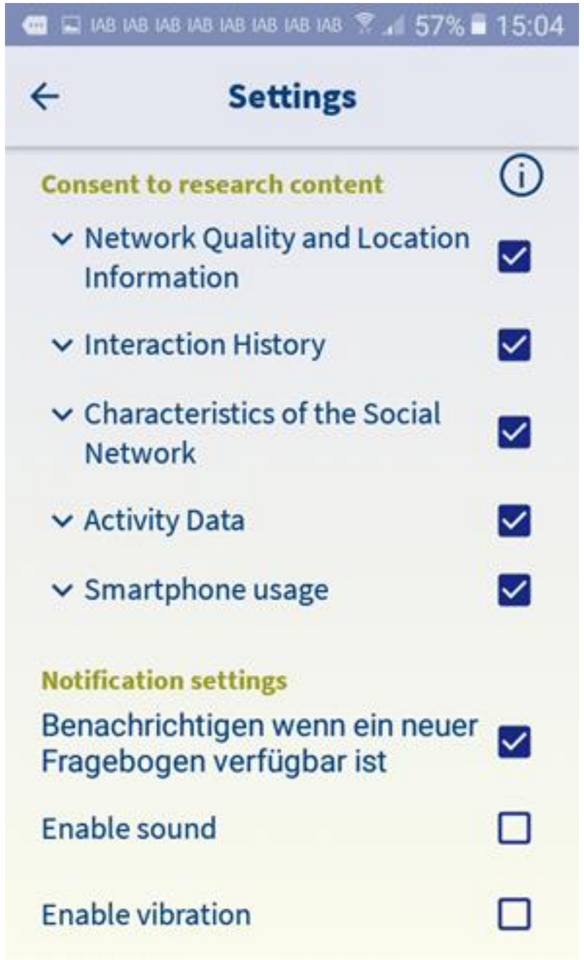


App home screen

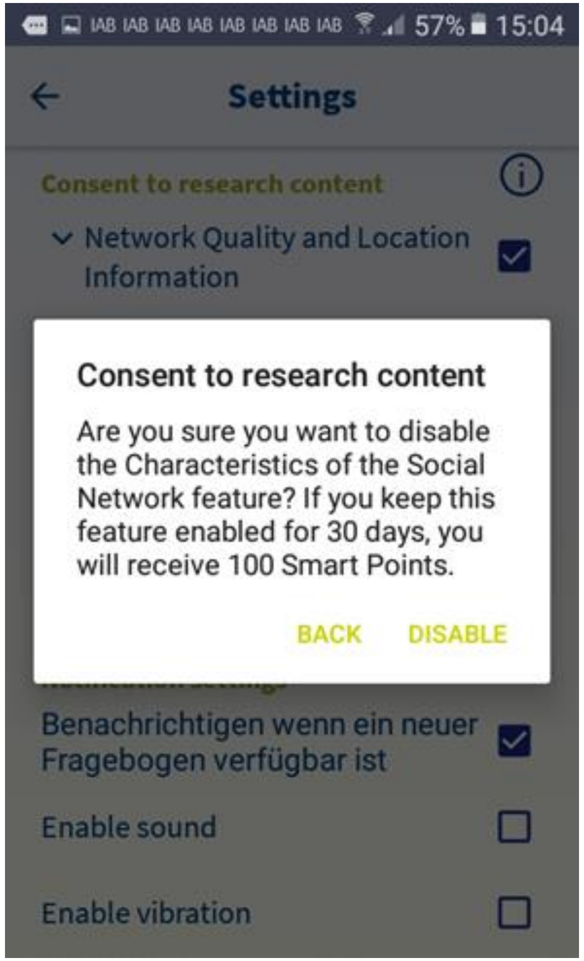
Example for consent: IAB-SMART (Kreuter et al. 2020)



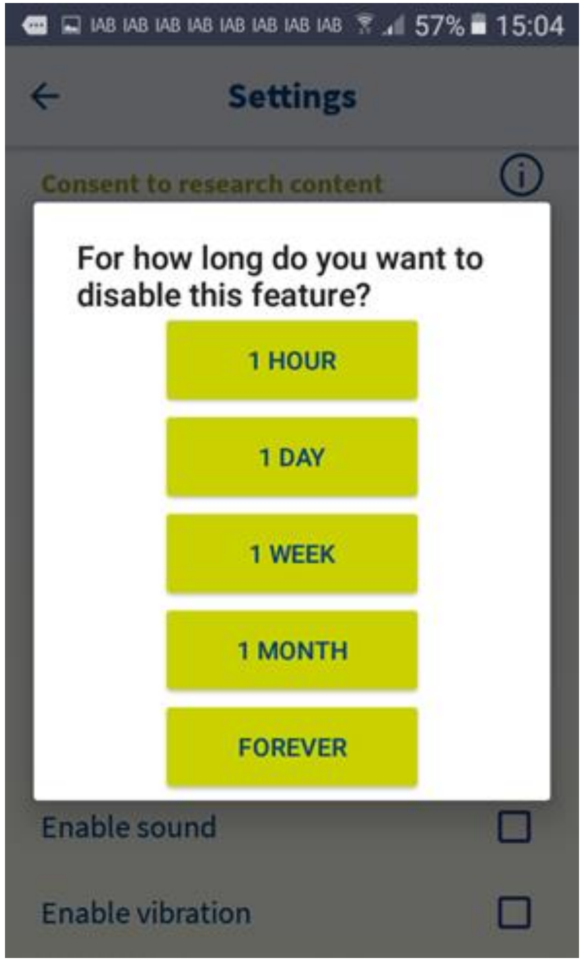
Example for consent: IAB-SMART (Kreuter et al. 2020)



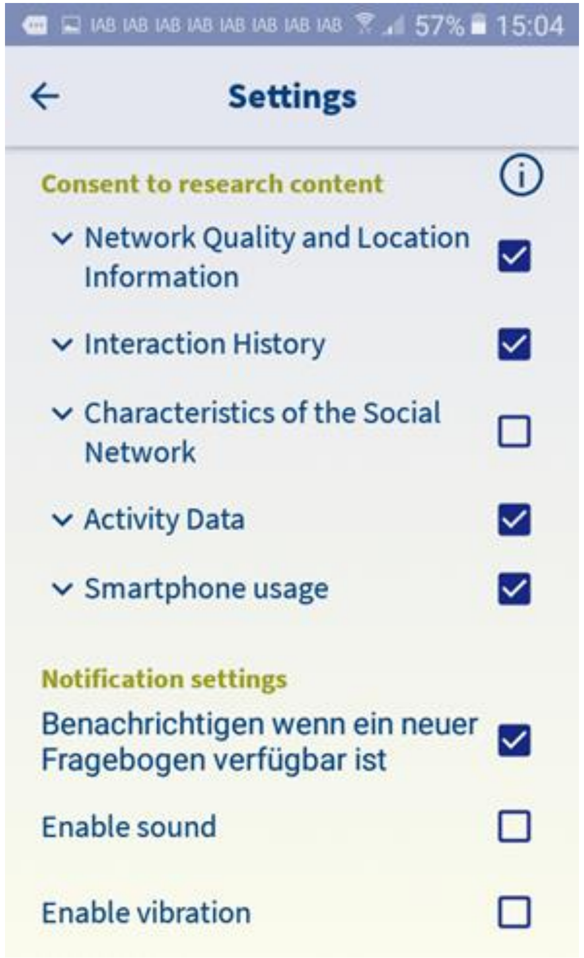
App settings



Withdrawing consent



Withdrawing consent



App settings

Example consent: One-time GPS (Struminskaya et al. 2021)

General consent

Framing, agency, privacy explanation

GPS measurement

Frequency of sensor measurement

- *Sampling rate* defines frequency of measurement (usually in Hz)
- Realized frequency depends on various factors
    - Sensor’s technical capabilities (max. sampling rate)
    - Outside factors (e.g., sleep/battery saving mode, technical failure)
    - Design decisions by researcher
- In practice, there seem to be four groups of measurement frequencies
    - Contingent
    - Discrete
    - Continuous
    - Combination

Frequency of sensor measurement

- Contingent
  - Measurement only at specific times, adding individual data points to survey, as if additional question was asked
  - e.g., GPS location whenever EMA is answered (MacKerron & Mourato 2013)

Frequency of sensor measurement

- Contingent
- Discrete
  - Usually to conserve battery and storage and/or to protect privacy
    - In case of GPS, allows to calculate activity radius but not specific traces
  - e.g., GPS every 5 min from 9 am to 9 pm (York Cornwell & Cagney 2017), every 15 minutes (Sugie 2018), every 30 minutes (Kreuter et al. 2020)
  - e.g., audio recordings for 30s every 7 min during waking hours (Fingerman et al. 2020, 2022)

Frequency of sensor measurement

- Contingent
- Discrete
- Continuous
  - Tracking of smartphone-mediated behavior usually always on (e.g., Kreuter et al. 2020; Stachl et al. 2020; Sugie 2018)
  - Accelerometer & gyroscope usually always on to detect movement
    - Actual frequency depends on device, model, etc.
  - GPS collected at high frequency allows measurement of exact route
    - But may have negative effect on phone performance
  - Microphone always on (e.g., Wang et al. 2014)
    - But data pre-processed on device to save storage and preserve privacy

Frequency of sensor measurement

- Contingent
- Discrete
- Continuous
- Combination
  - Combining fine grained tracking with saving battery and reducing invasiveness
  - e.g., measure activity (accelerometer) only at specific times during the day - 15 min twice a day (Lathia et al. 2017)
  - e.g., reduce sampling rate of GPS if idle, based on accelerometer measures - once every second when in motion but only every minute when still (McCool et al. 2021)

Battery life

- High sampling rates might reduce battery
  - Especially for GPS tracking

- Newer models might go to sleep mode and/or turn off data collection when reaching low battery level

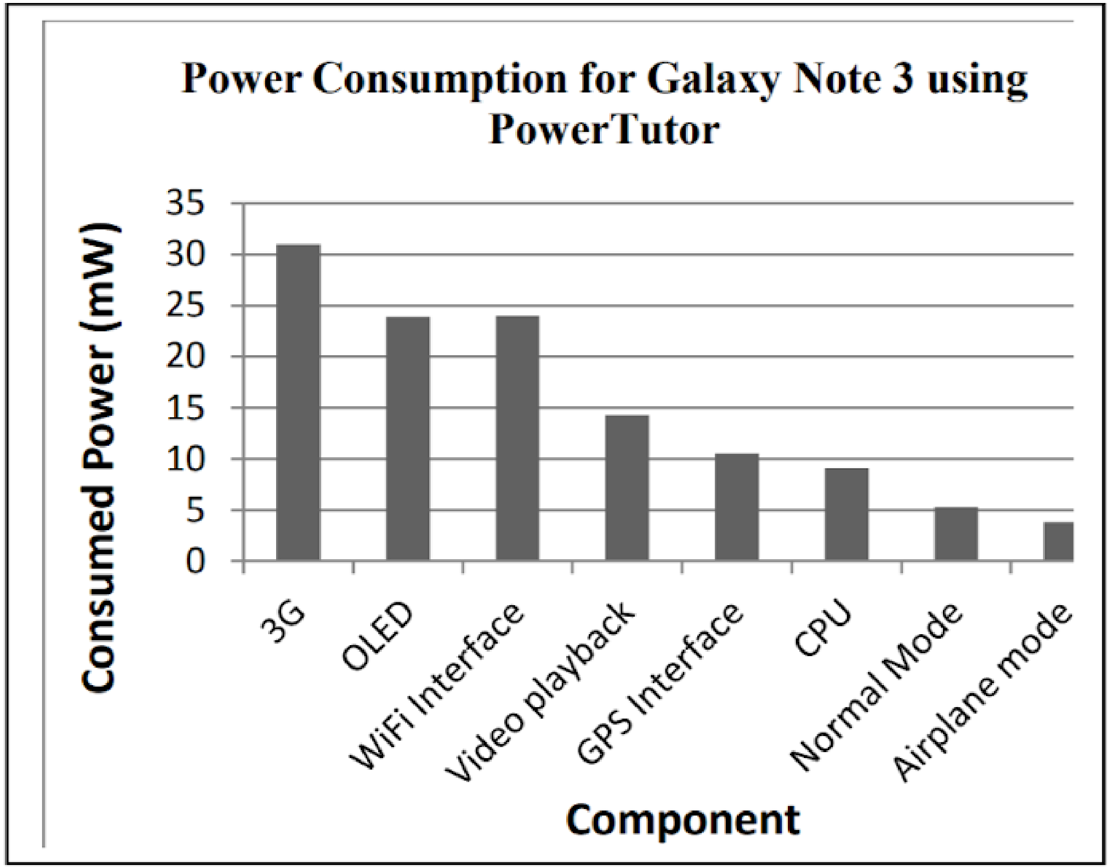
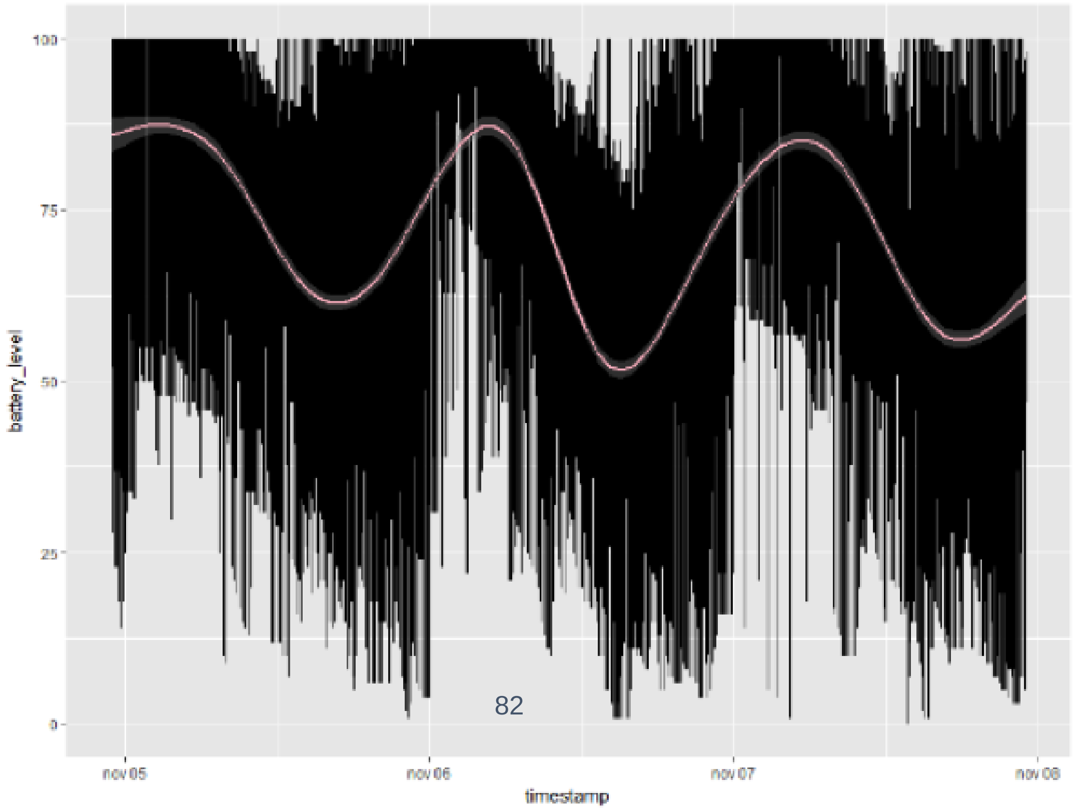


Fig. 5. Power measurements for Galaxy Note3 using PowerTutor.

Tawalbeh et al. (2016)

Battery life example: Statistics Netherlands Travel App (McCool et al. 2019)

- Battery levels for all participants Nov 5-8, 2018
    - Battery levels follow circadian pattern
    - Very few batteries run empty over course of four days



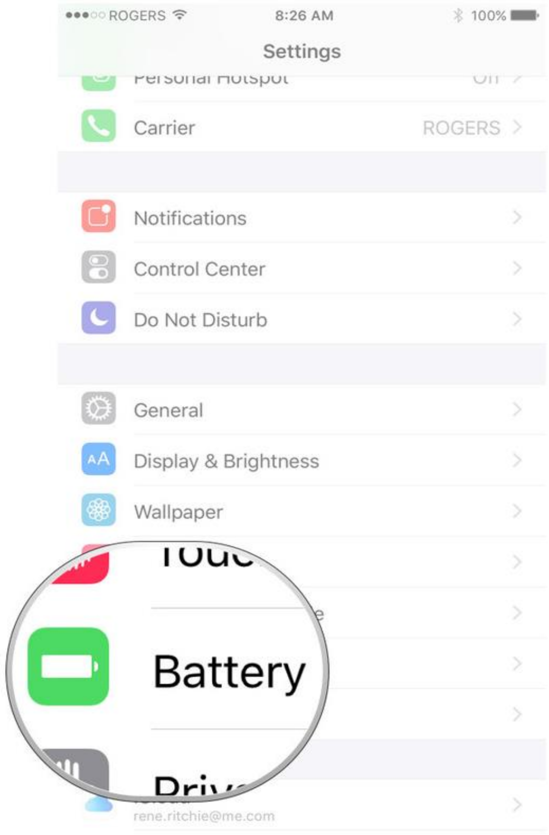
- For loander devices, participants need to be reminded to charge battery

Exercise

Find out what drains your battery!

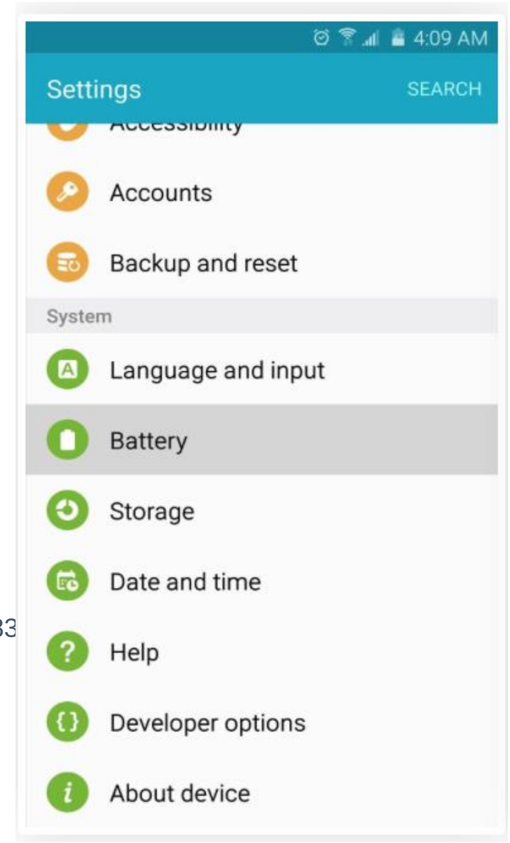
- Settings → Battery

iPhone



Source: https://www.imore.com/how-see-whats-using- battery-life-your-iphone-or-ipad

Android



Source: https://android.gadgethacks.com/how-to/whats- draining-your-androids-battery-find-out-fix-for-good-0162267/

Data transfer

- Some systems store data first on device and then transmit them to server at predefined intervals or once connected to Wi-Fi
  - e.g., for smartphones, if no Wi-Fi connection available for longer time, more expensive cell connection is used

- Other systems require researcher to collect device and download data manually
  - e.g., research-grade accelerometers

Storage

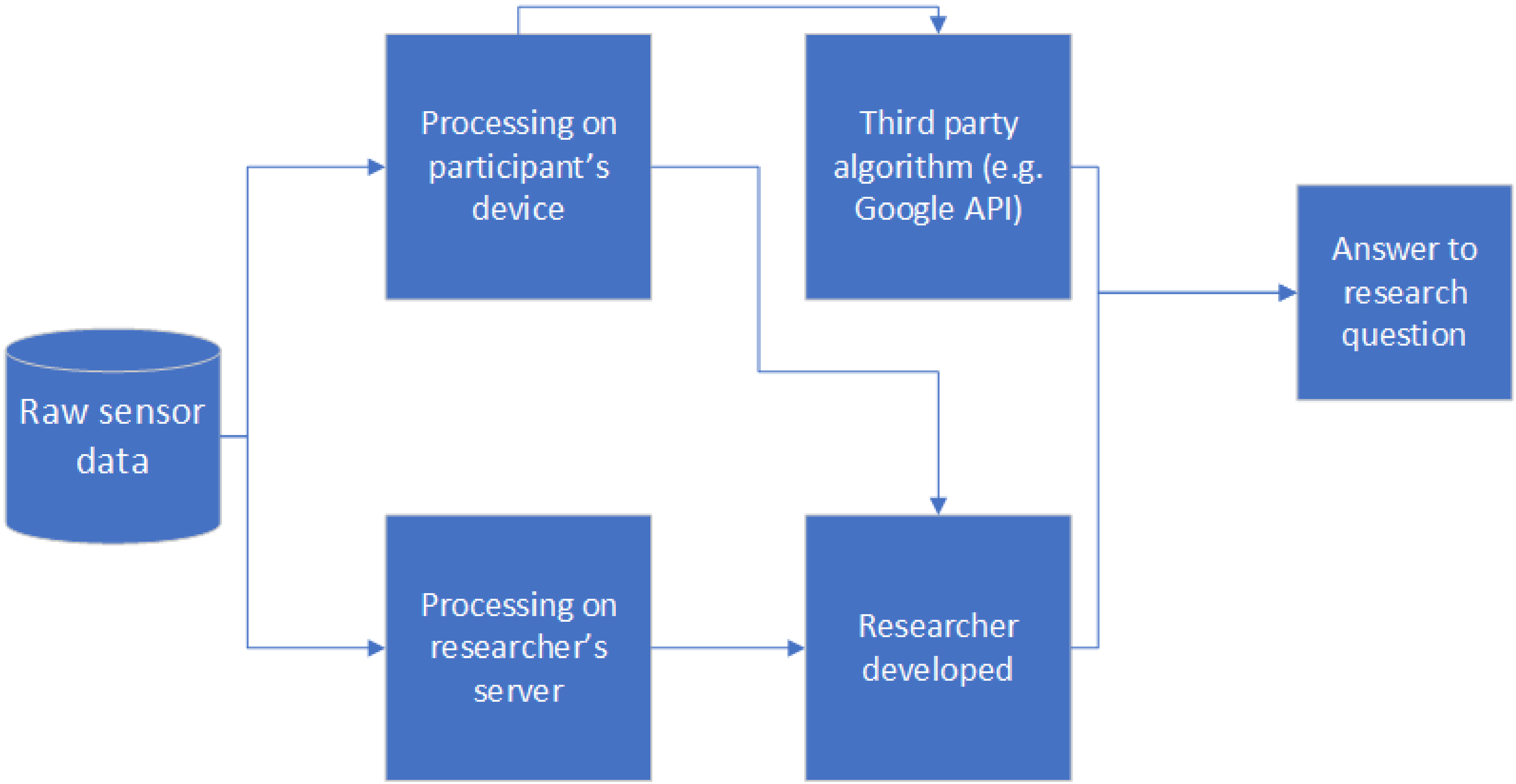
- Size of data = Sampling rate * Field period
  - Even small samples, might produce *“Big Data”*
  - e.g., accelerometer data (three coordinates) collected at 60 Hz (i.e., 60 measurements per sec) for 10 min creates 108,000 data points per participant
- ...think whether you really need this amount of data and if so, have the appropriate infrastructure ready
- Processing data on device and only transmitting processed/aggregated data saves storage for researchers
  - e.g., using Google API to automatically classify accelerometer data into transportation modes
  - e.g., OCR for receipt scanning

Costs

- Sensor measurement might seem relatively inexpensive because of small/no marginal costs of additional data
- Costs come from various sources
  - App development & maintenance (potentially for multiple OS)
  - Loaner devices
  - Incentives for participants
  - Technical support (e.g., hotline)
  - Storage infrastructure
  - Data handling & analytical skills (e.g., data wrangling, working with machine learning algorithms, data linkage, etc.)
  - Data visualization skills (e.g., GIS expertise)

Data from sensors, apps, and wearables

From raw data to insights



Examples of sensor data: Accelerometer


```
SMotion_Time_Stamp;SMotion_without_Gravity;SM_with_Gravity
323,337,354,370,386,404,421,437,454,471,487,504,520,538,564,570,587,604,721,737,758,770,786,804,821,837,854,871,887,904,920,937,954,970,987,1004,1021,1
,3954,3971,3987,4004,4021,4037,4054,4071,4087,4104,4121,4137,4154,4171,4187,4204,4221,4238,4254,4271,4287,4304,4321,4337,4354,4371,4387,4404,4421,4437,
2,0.52,0.30,0.12,0.18,0.12,0.10,0.10,0.17,0.34,0.29,0.22,0.16,0.18,0.26,0.19,0.23,0.16,0.17,0.13,0.11,0.05,0.05,0.11,0.13,0.10,0.10,0.37,0.12,0.07,0.08,
06,0.08,0.12,0.22,0.22,0.13,0.37,0.81,0.54,0.28,0.29,0.44,0.28,0.23,0.26,0.32,0.23,0.32,0.36,0.18,0.19,0.24,0.31,0.20,0.13,0.21,0.23,0.21,0.16,0.13,0.1
.15,9.80,9.88,10.04,10.47,9.85,10.02,10.02,10.12,10.08,9.61,9.58,9.81,10.12,10.08,9.96,9.89,9.43,9.37,9.51,9.98,10.07,10.00,9.83,9.84,9.76,9.71,9.83,9.
.85,9.84,9.87,9.91,9.90,9.89,9.80,9.75,9.75,9.75,9.89,9.94,9.95,9.95,9.84,9.95,10.07,9.92,9.97,9.71,9.76,9.61,9.53,9.55,9.44,9.41,9.62,9.78,10.00,10.10,
222,237,254,274,288,304,321,337,354,372,388,404,422,438,455,472,488,504,522,538,554,572,588,604,621,638,673,688,710,721,738,754,772,791,805,821,838,854,
3838,3854,3871,3887,3904,3921,3938,3954,3971,3988,4006,4021,4038,4054,4071,4088,4104,4121,4138,4155,4172,4188,4204,4222,4238,4255,4271,4288,4304,4322,4
,0.08,0.07,0.10,0.11,0.10,0.15,0.18,0.12,0.19,0.07,0.15,0.08,0.14,0.30,0.16,0.12,0.18,0.10,0.13,0.06,0.04,0.04,0.12,0.35,0.48,0.28,0.31,0.31,0.18,0.11,
5,0.53,0.14,0.23,0.16,0.15,0.19,0.25,0.22,0.23,0.09,0.26,0.22,0.19,0.12,0.10,0.13,0.11,0.12,0.12,0.19,0.17,0.08,0.06,0.08,0.08,0.02,0.01,0.04,0.04,0.11,
.91,9.78,9.82,9.82,9.90,9.86,9.76,9.82,9.82,9.90,9.88,9.86,9.93,9.93,9.87,9.82,9.67,9.77,9.84,10.00,10.02,9.90,9.84,9.77,9.76,9.77,9.83,9.85,9.97,10.01,
111,128,144,452,569,587,619,637,717,734,751,769,786,802,819,836,853,871,888,906,923,940,958,974,990,1008,1025,1042,1059,1076,1093,1110,1127,1144,1161,1
,4254,4271,4288,4306,4324,4340,4358,4374,4391,4408,4425,4442,4459,4476,4493,4511,4529,4547,4565,4582,4599,4616,4633,4651,4668,4684,4702,4719,4736,4753,
6,0.25,0.15,0.15,0.37,0.29,0.13,0.30,0.29,0.24,0.46,0.52,0.42,0.41,0.18,0.10,0.26,0.31,0.29,0.36,0.43,0.31,0.15,0.15,0.14,0.13,0.23,0.35,0.34,0.30,0.31,
36,0.25,0.57,0.36,0.37,0.36,0.31,0.36,0.30,0.26,0.20,0.20,0.10,0.10,0.09,0.19,0.18,0.11,0.38,0.23,0.23,0.51,0.56,0.01,0.60,0.68,0.66,0.35,0.24,0.43,0.2
9.67,9.59,9.57,9.53,9.58,9.68,9.68,9.70,9.67,9.59,9.62,9.58,9.65,9.56,9.46,9.40,9.42,9.42,9.52,9.64,9.63,9.67,9.67,9.68,9.77,9.87,9.70,9.70,9.51,9.32,9.
,9.56,9.58,9.58,9.55,9.53,9.59,9.63,9.65,9.61,9.60,9.60,9.54,9.53,9.57,9.58,9.58,9.57,9.56,9.59,9.68,9.63,9.63,9.63,9.67,9.66,9.60,9.60,9.57,9.55,9.54,
183,199,216,236,249,266,282,299,316,333,348,366,382,399,416,433,449,465,482,500,515,532,549,566,582,599,616,632,649,666,682,699,716,733,749,766,783,799,
,0.07,0.09,0.10,0.11,0.08,0.09,0.07,0.07,0.07,0.06,0.04,0.07,0.12,0.11,0.06,0.10,0.15,0.14,0.21,0.08,0.08,0.08,0.13,0.50,0.38,0.29,0.34,0.36,0.47,0.46,
,9.93,9.85,9.65,9.53,9.44,10.19,10.04,9.91
137,608,725,847,852,852,853,855,856,856,860,863,863,864,865,866,866,866,867,867,867,868,870,917,918,918,926,927,927,947,989,990,990,991,991,992,1001,10
2797,2807,2821,2831,2840,2855,2865,2875,2896,2906,2927,2936,2946,2956,2966,2976,2997,3007,3017,3027,3052,3069,3077,3090,3105,3115,3125,3135,3146,3155,3
,5192,5202,5213,5222,5232,5243,5252,5263,5273,5293,5304,5314,5324,5335,5345,5366,5376,5386,5406,5417,5429,5480,5480,5481,5482,5482,5532,5533,5533,5533,
1,7451,7485,7486,7486,7511,7511,7512,7518,7529,7538,7549,7559,7569,7579,7589,7600,7610,7621,7631,7641,7651,7661,7671,7681,7691,7701,7711,7722,7731,7741,
61,9861,9862,9864,9865,9865,9865,9866,9866,9866,9867,9867,9867,9868,9868,9868,9869,9869,9869,9922,9922,9922,9922,9923,9927,9964,9965,9968,9978,10022,10
,11612,11621,11632,11641,11651,11661,11671,11681,11691,11701,11712,11722,11732,11742,11752,11762,11772,11782,11793,11803,11814,11827,11837,11847,11857,
06,13516,13537,13547,13556,13569,13578,13590,13611,13641,13665,13681,13691,13722,13732,13745,13775,13784,13795,13805,13836,13846,13856,13866,13877,1388
5821,15832,15842,15859,15863,15875,15883,15896,15913,15922,15932,15943,15953,15963,15973,15983,15993,16004,16015,16025,16037,16047,16056,16067,16078,16
```

Examples of sensor data: GPS and trip motives

	device_id	latitude	longitude	accuracy	speed	altitude	timestamp
1:	23	52.09460	5.134593	15.204	0	54.7	2018-10-31 22:53:22
2:	23	52.09460	5.134593	15.204	0	54.7	2018-10-31 22:54:22
3:	23	52.09460	5.134593	15.204	0	54.7	2018-10-31 22:56:40
4:	23	52.09460	5.134593	15.204	0	54.7	2018-10-31 22:57:40
5:	23	52.09460	5.134593	15.204	0	54.7	2018-10-31 22:59:05

38524:	23	52.09464	5.134572	15.175	0	54.7	2018-11-12 00:00:33
38525:	23	52.09464	5.134572	15.175	0	54.7	2018-11-12 00:00:33

	device_id	local_stop_id	begin_timestamp	end_timestamp
1:	23	7	2018-10-31 17:05:47	2018-10-31 17:11:51
2:	23	11	2018-10-31 17:26:56	2018-10-31 17:31:39
3:	23	5	2018-10-31 17:32:51	2018-10-31 17:40:09
4:	23	4	2018-10-31 17:45:13	2018-10-31 19:03:58
5:	23	8	2018-10-31 19:04:08	2018-11-01 12:53:08
6:	23	9	2018-11-01 13:00:52	2018-11-01 15:47:21

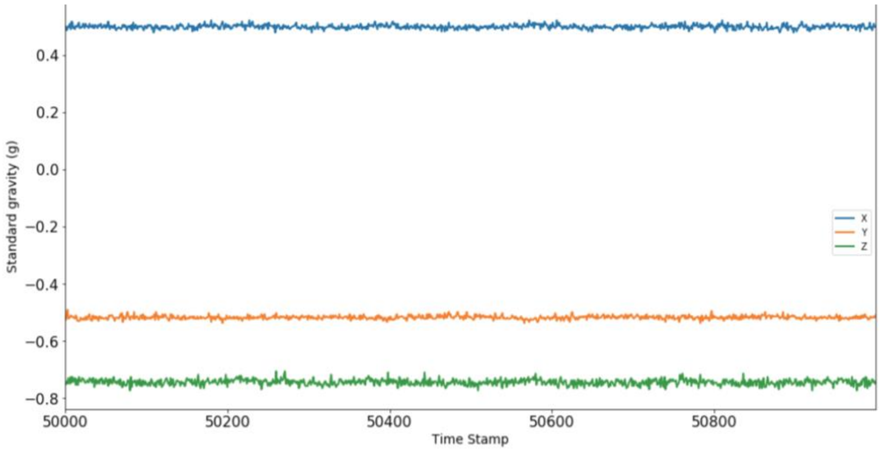
	device_id	local_stop_visit_id	motive
1:	23	3	Home
2:	23	2	PaidWork
3:	23	1	Home
4:	23	7	Transfer
5:	23	6	Transfer
6:	23	4	Home

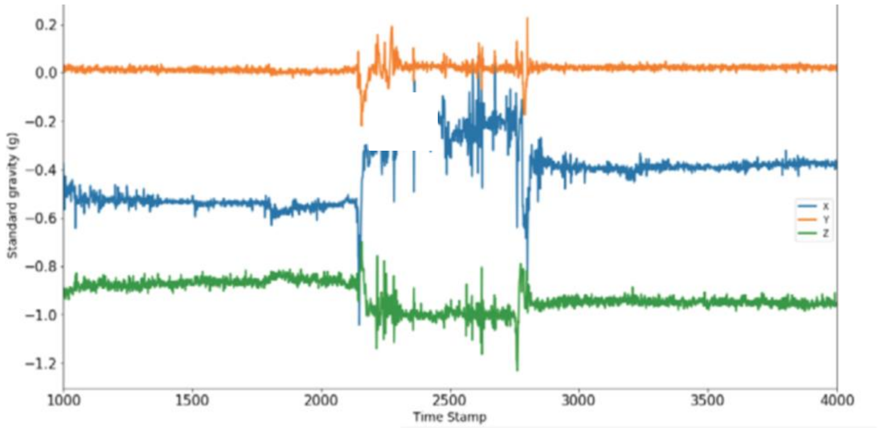
Processing raw data

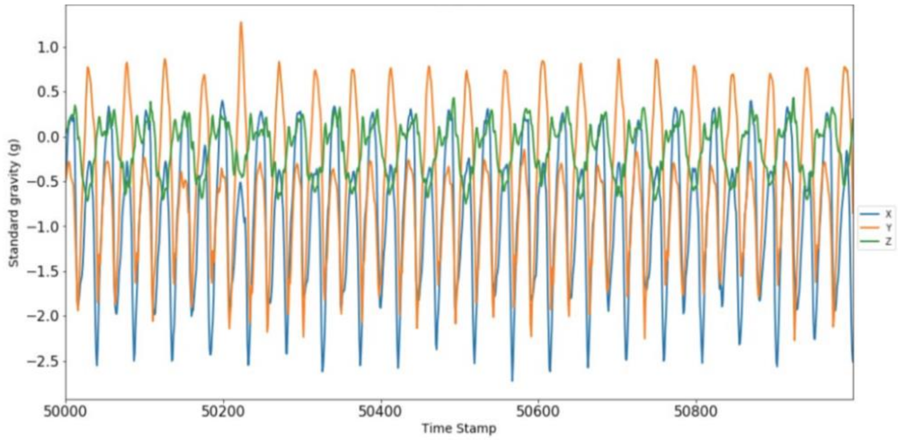
- Data needs to be cleaned and processed before analysis (*Data wrangling/munging*)
    - This usually takes much longer than data analysis (80/20 rule)
- Aggregation of raw data to meaningful data point level
    - What is “meaningful” depends on research and use of data
- Processing of raw data can happen on
    - User’s device using (built-in) third party or researcher-developed algorithm
        - Preserves storage and protects privacy
        - No access to raw data
    - Researcher’s server
        - Full control over data processing
        - All data needs to be transfered

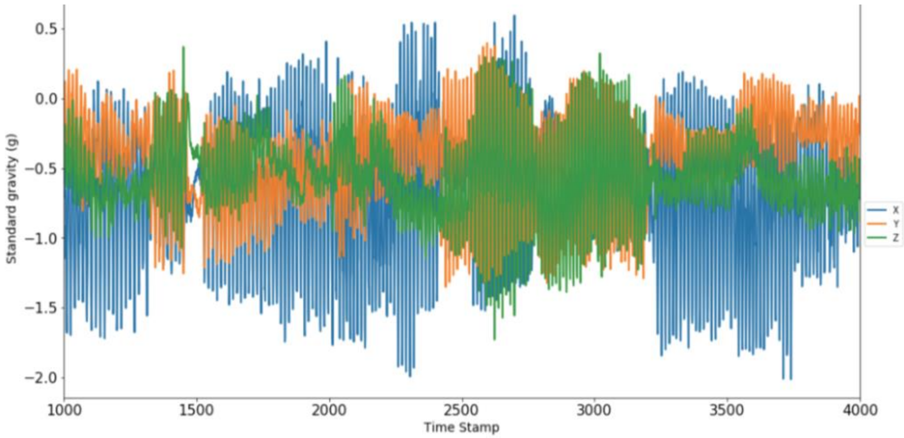
Exercise: Detect types of activity

Match:
Brushing teeth
Jogging
Sleep
Sitting









Mulder et al. (2019)

Model building pipeline

Data Cleaning & pre-processing

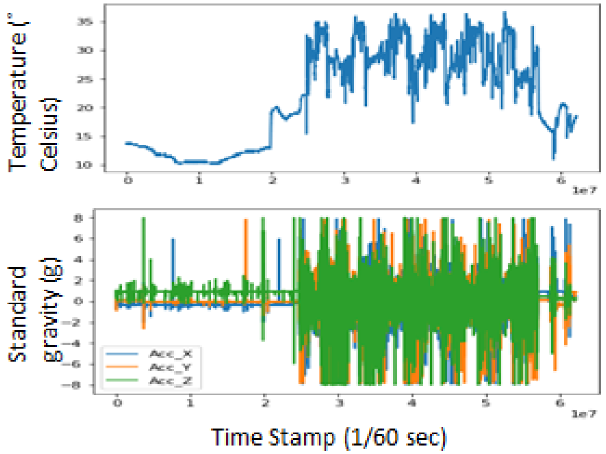
- Removal non-wear time
- Removal of high frequency (frequency higher than 15 Hz)
- Data with wear time less than 7 days discarded

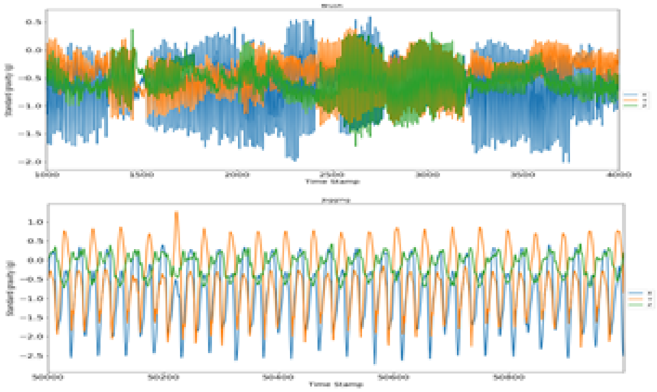
Feature Engineering

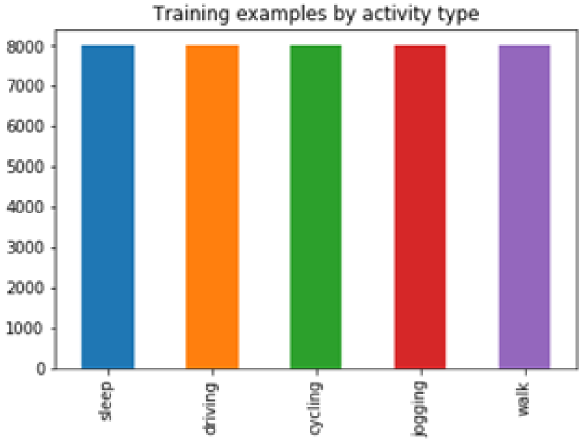
- **Time domain:** X, Y, Z, temperature, mean, median, standard deviation, RMS, percentile distribution
- **Frequency domain:** FT, dominant frequency selection, power of signal

Model building & validation

- Optimizing the epoch time
- Preparing balanced dataset
- Train/test splitting of 80%/20%
- Training and validation of the model (SVM, RF, and LR model)



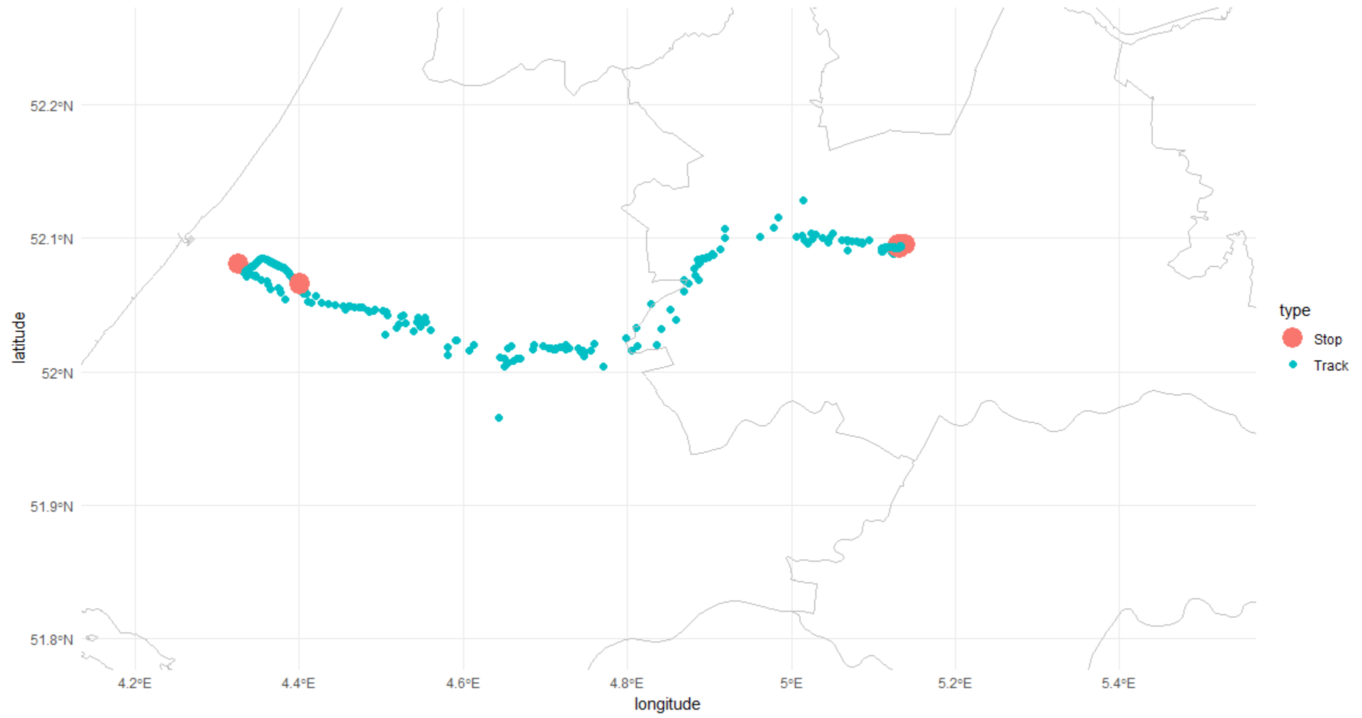


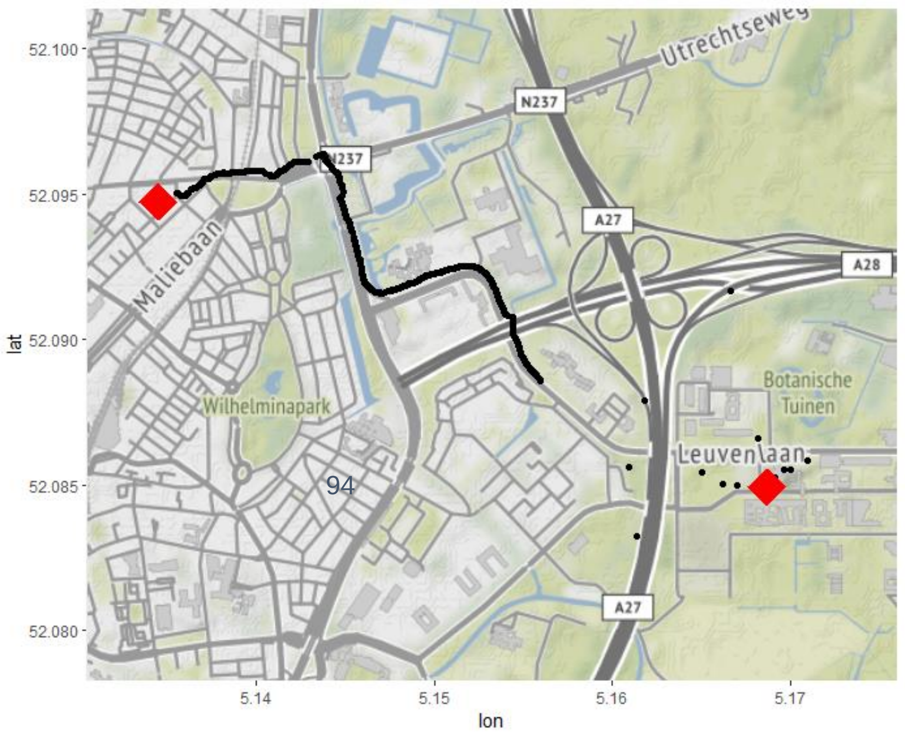


Mulder et al. (2019)

Example: GPS tracks and stop detection (McCool et al. 2019)

- Stops defined based on "static" location: radius has to be (pre)defined by researcher

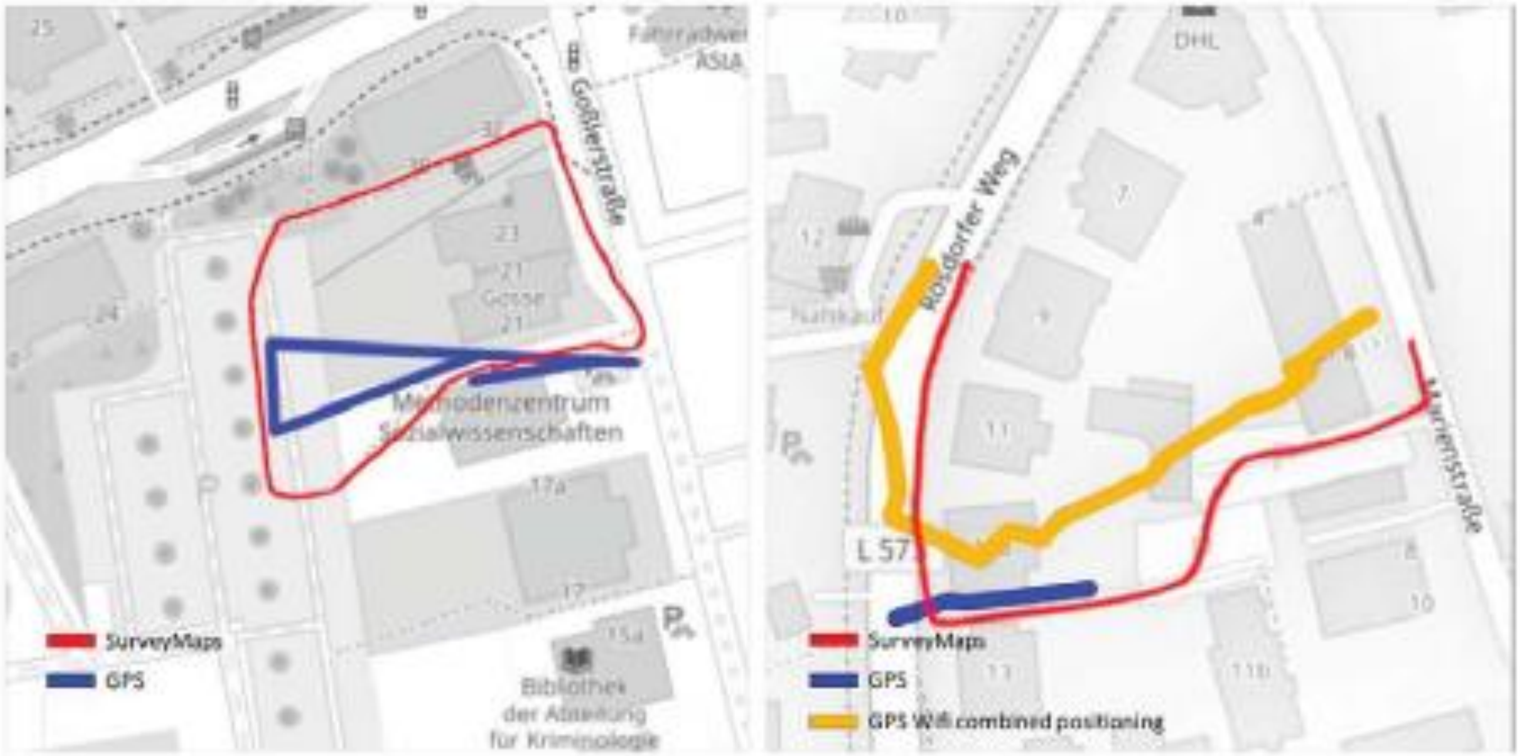




Errors when collecting, processing, and interpreting sensor data

Errors during data collection

- Sensor-based errors/differences
  - Differences between types of sensors as well as brands and models of devices
  - Not one sensor/device per se better than others, depends on what should be measured under what circumstances



Schlosser et al. (2019)

Errors during data collection

- Sensor-based errors/differences
- Device handling
    - Measurement might differ depending on where/
      how sensor/device is worn
        - e.g., differences in how men and women carry around smartphones

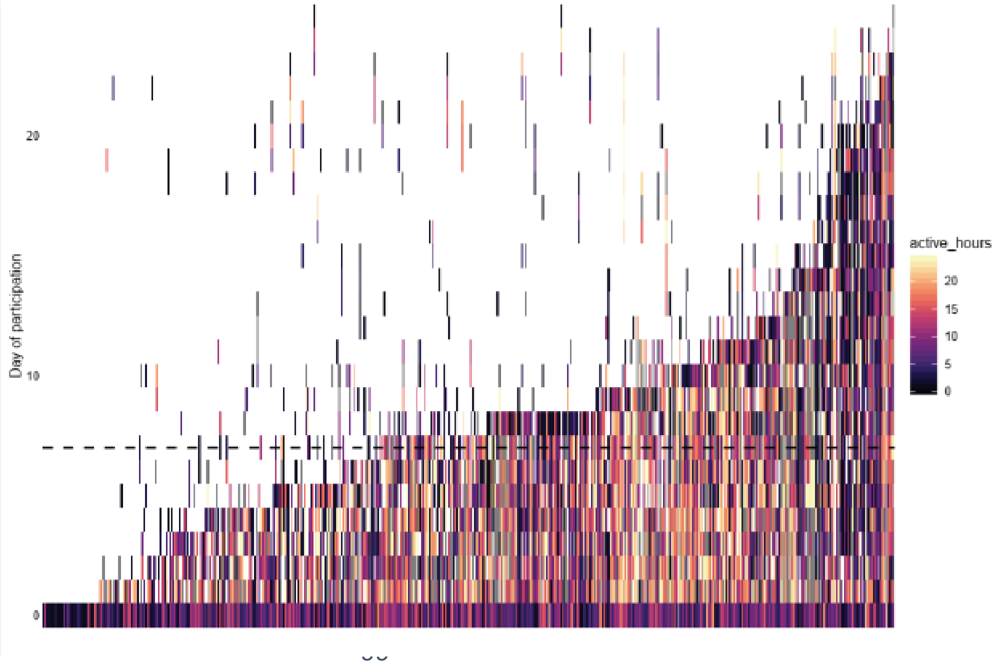
Sztyler et al. (2017)

Errors during data collection (Keusch et al. 2022)

Behavioral barriers – Smartphone...	**Sample 1**	**Sample 2**
...shared with another person	2%	1%
...not always on	32%	44%
...left at home	17%	14%
...carried in purse/backpack/bag when not at home	46%	30%
...left stationary when at home and not asleep	66%	47%
...turned off or in other room at night	49%	34%
n	3,956	2,525

Errors during data collection

- Sensor-based errors
- Device handling
- Missing data
    - Technical issues:
        - Urban canyons, underground, etc. when collecting GPS
        - Device out of power or sleep mode
        - iOS blocks collection of location in background
        - …
    - Noncompliance:
        - Leaving device at home
        - Deliberately turning device off at certain locations or times
        - Forgetting to turn device back on again
        - Missing permissions
        - …



Active hours per user

McCool et al. (2019)

Errors during data collection

- Sensor-based errors
- Device handling
- Missing data
- Erroneous/Invalid data
    - e.g., fake GPS apps, VPN

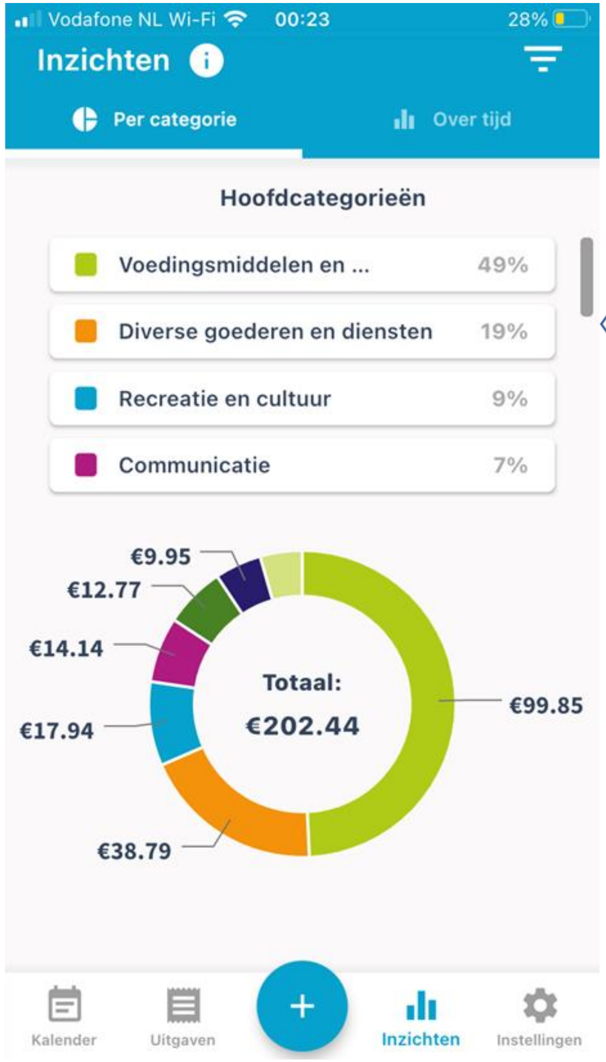
Source: Sebastian Bähr

Errors during data collection

- Sensor-based errors
- Device handling
- Missing data
- Erroneous data
- Providing feedback & measurement reactivity
  - e.g., participants show 7% more physical activity when wearing Fitbit (with feedback) compared to when wearing GENEActive (no feedback) (Darling et al. 2021)

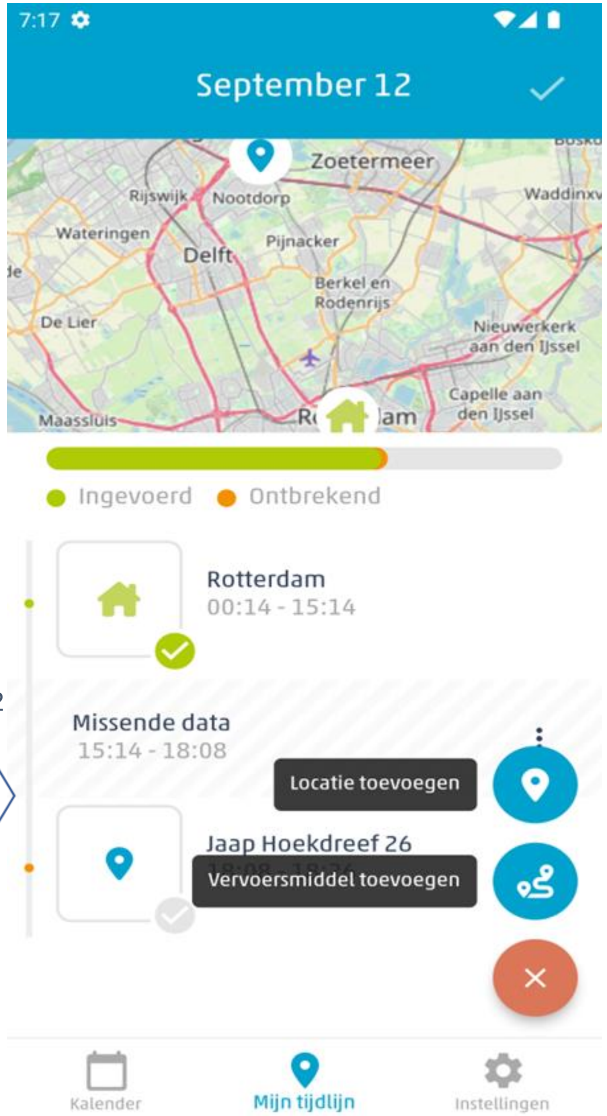
Source: https://twitter.com/mbrennanchina/status/1128201958962032641

Providing feedback to participants



Household Budget Survey (HBS)
fall 2021,
NL, ES, LU
N=3916,
Completion = 16%
No influence of feedback on representativeness, data quality

Travel app possibility to provide context to passive data, add data

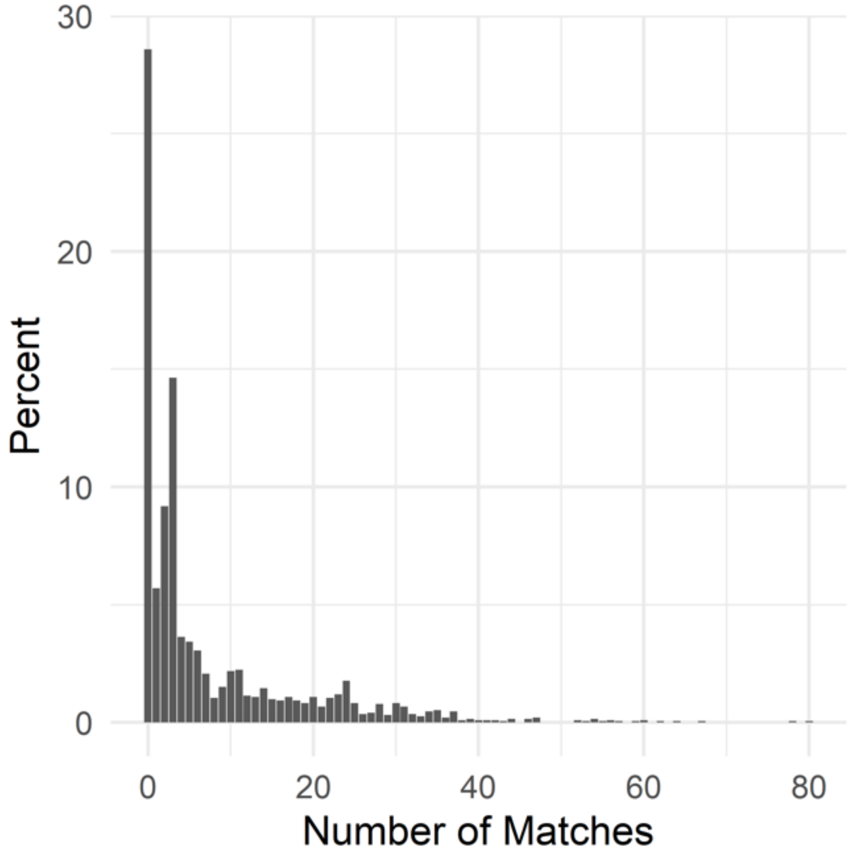


Errors during inference & interpretation

- Raw sensor data must be processed and classified to infer behavior
- “Black box” approach when using third-party algorithm to classify data on device
  - What looks like raw data to researcher is actually (heavily) pre-processed
  - e.g., activity classification was trained based on data from young adults (“WEIRDOS” ©Mick P. Couper) → used to classify behavior of general population
  - e.g., smartphone forgotten at home in a bag → respondent is asleep
- Self-report still needed for validation

Errors during inference & interpretation

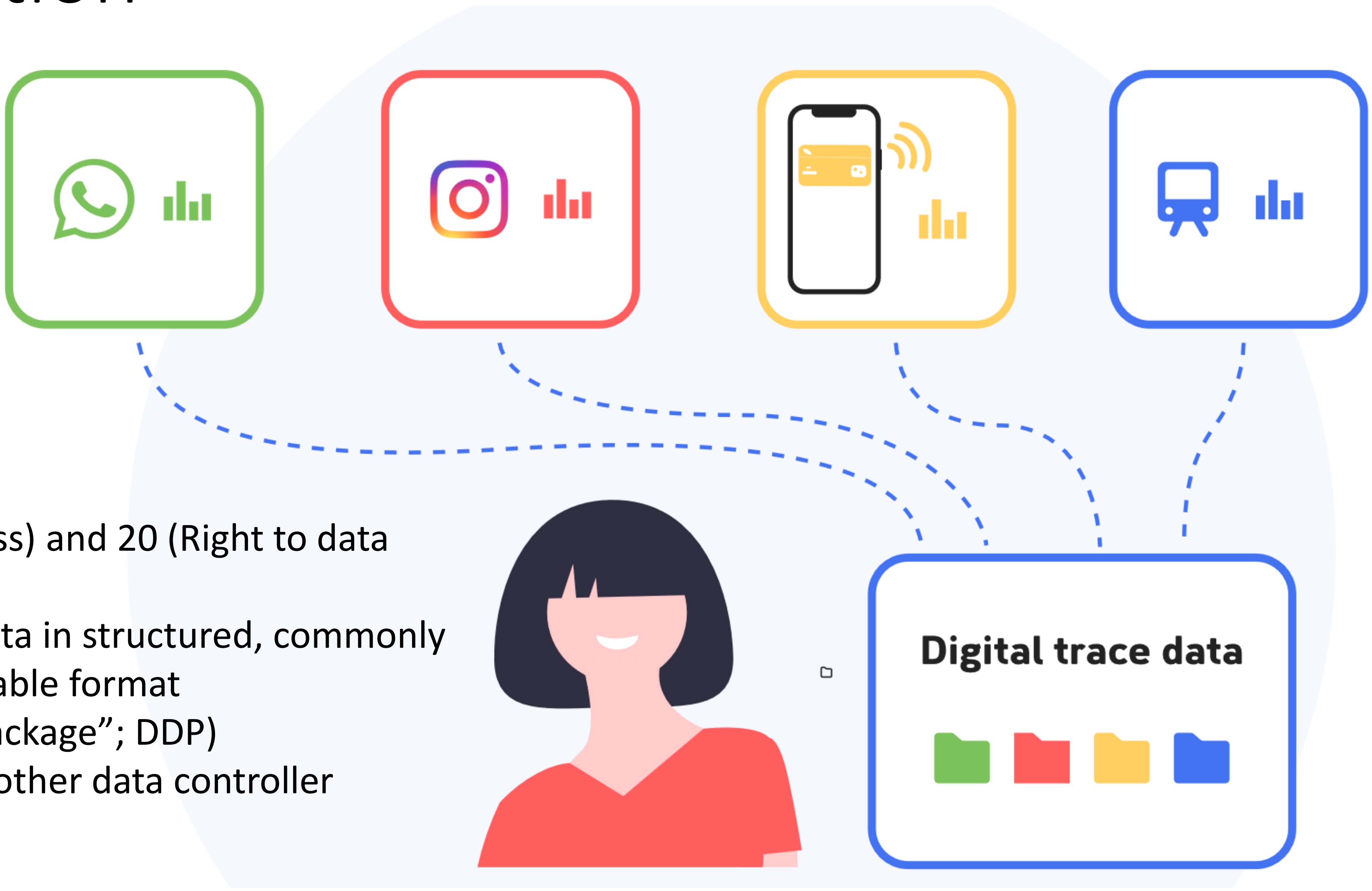
- Errors can also arise when using third-party reference databases
- Challenges in matching places to points of interest (Eckman et al. 2020)
  - Data from multiple bases agreed only on 53 of 1,928 collected coordinates
  - Disagreement between data sources
    - "Strip mall problem": grocery store on Google, liquor store on Foursquare, dentist's office on Yelp
    - Agreement for very big stores, airports, etc.
  - Sometimes, too many matches
  - Sometimes, no matches at all



Eckman et al. (2020)

Data donation

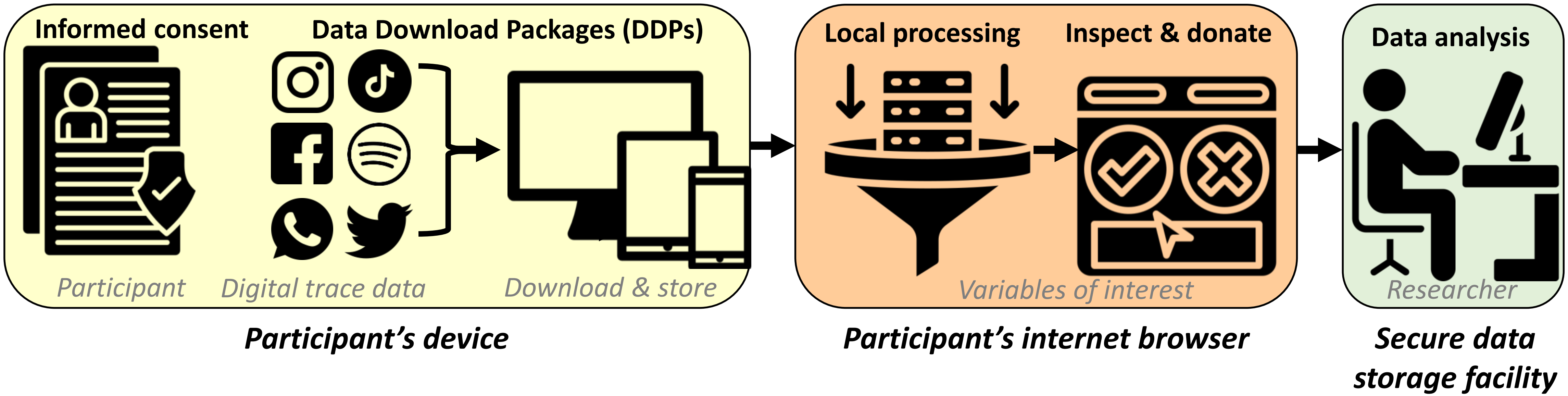
Data donation



GDPR Art. 15 (Right of access) and 20 (Right to data portability)

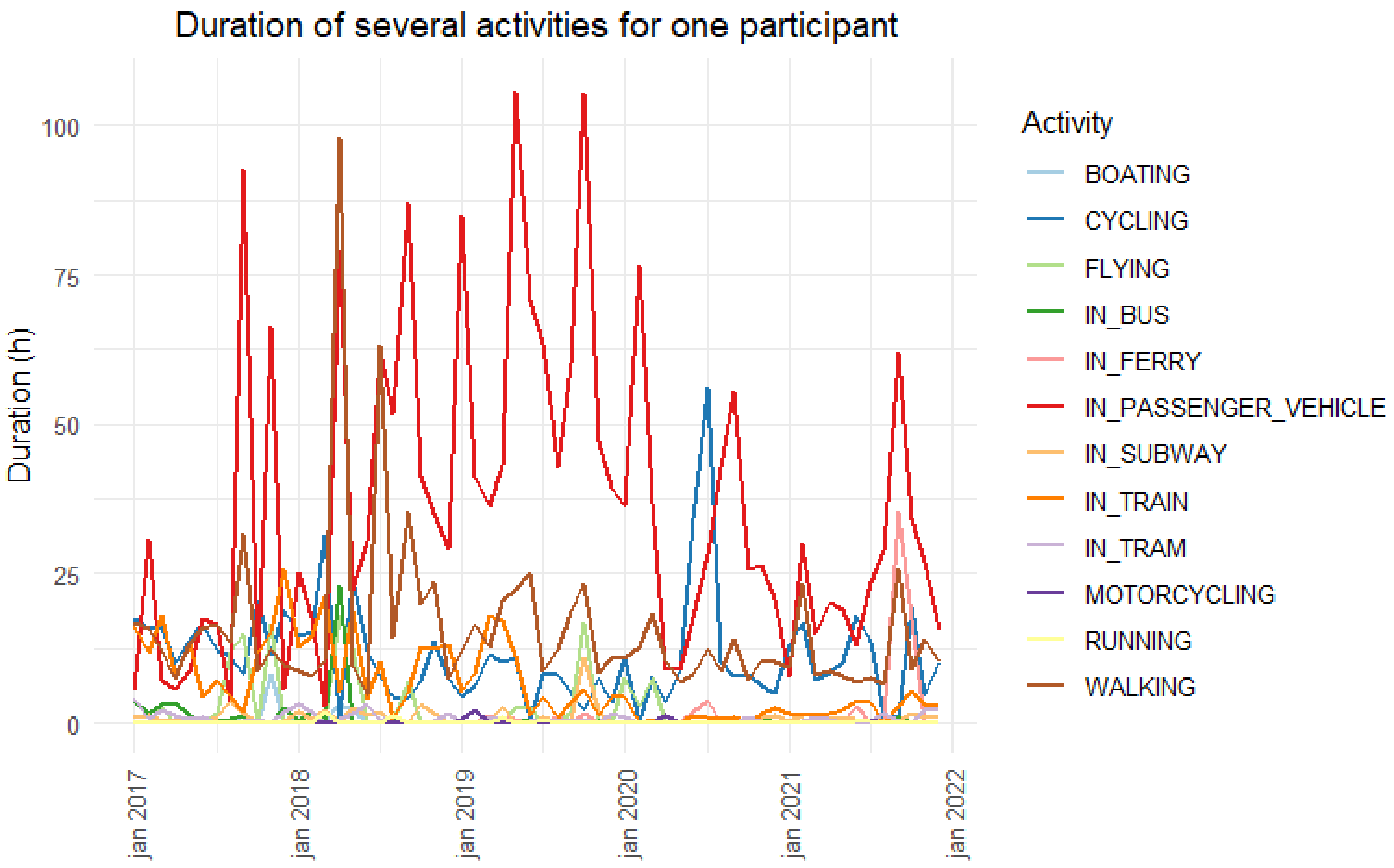
- Receive personal data in structured, commonly used machine-readable format ("Data Download Package"; DDP)
- Transmit data to another data controller

Local processing



[Boeschoten et al., 2023]

Google Location History Data Donation



Data donation

Advantages

- Allows access to data from digital platforms that cannot be collected otherwise
- Works for many platforms
  - Facebook, Instagram, WhatsApp, Google, YouTube, Netflix, Apple Health, Fitbit, …
- User retains control over what data are donated

Challenges

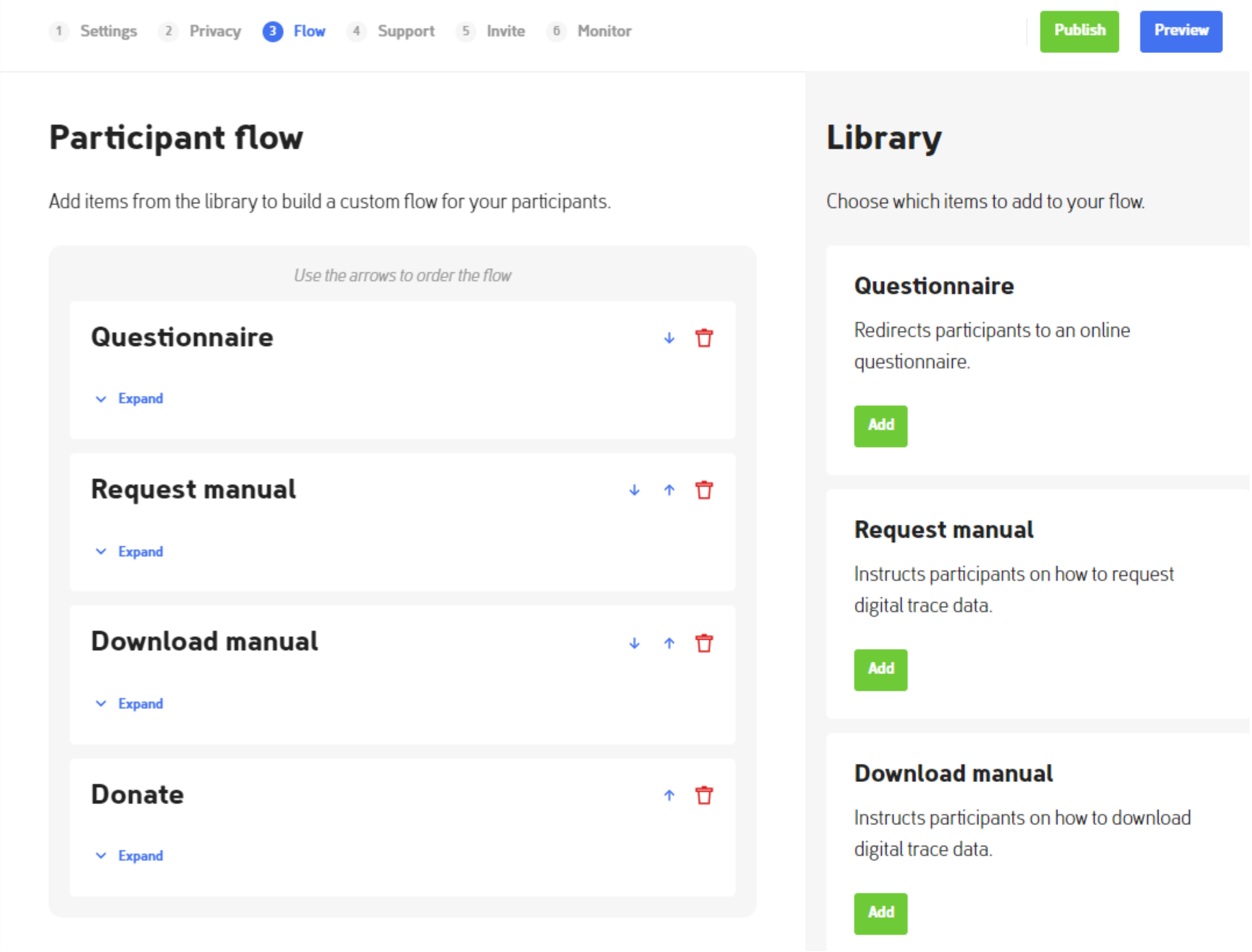
- Data donation process rather cumbersome for users (willingness/participation)
- Linking between donated data and other data (e.g., from survey) not well implemented yet
- Technical know-how needed

Software (Port) to build a complete study flow

- Consent form
- Privacy policy
- Support

- Send out invites
- Monitor progress

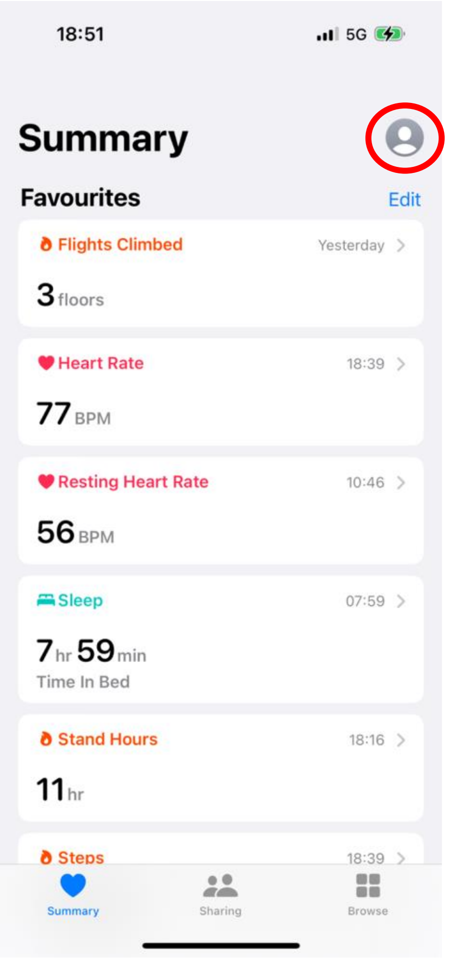
Researcher perspective:

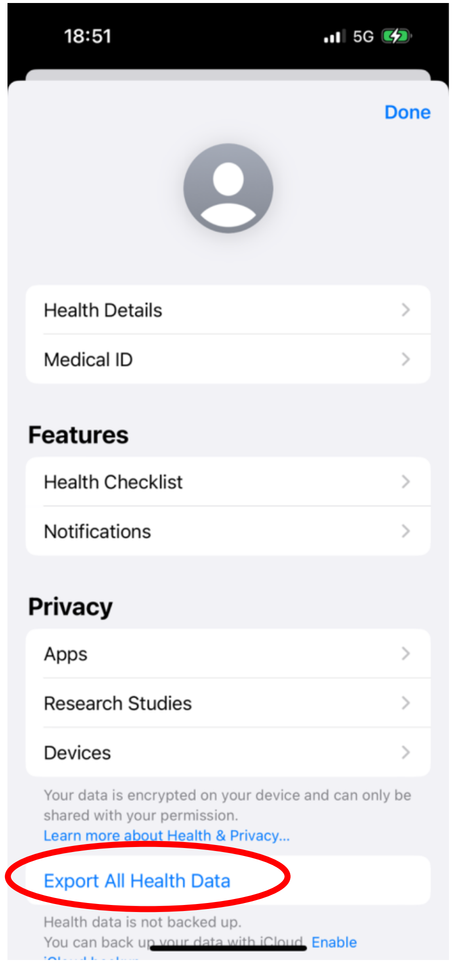


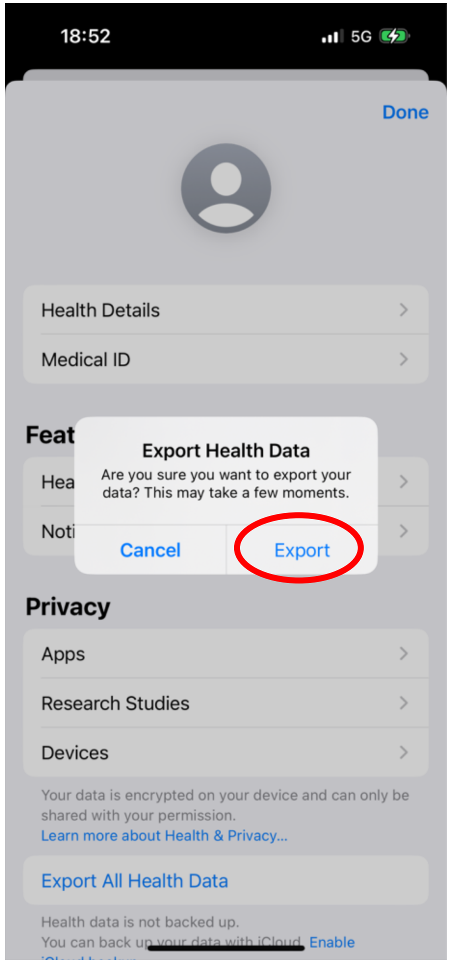
See Boeschoten et al. 2023 article for description of the software

Exercise: Request your DDP

Download DDP from iHealth







Request DDP from Google (with location history)

Walk-through as shown to study participants on the following slides

- Step 1: Navigate to https://takeout.google.com and log in
- Step 2: Click "Deselect all"
- Step 3: Find "Location History"
- Step 4: Check "Your locations and settings from Location History"
- Step 5: Scroll to bottom of page
- Step 6: Click "Next step"
- Step 7: Scroll to bottom of page
- Step 8: Click "Create export" button
- Step 9: Check email for message from Google

Google Account

Google Takeout

Your account, your data.
Export a copy of content in your Google Account to back it up or use it with a service outside of Google.

CREATE A NEW EXPORT

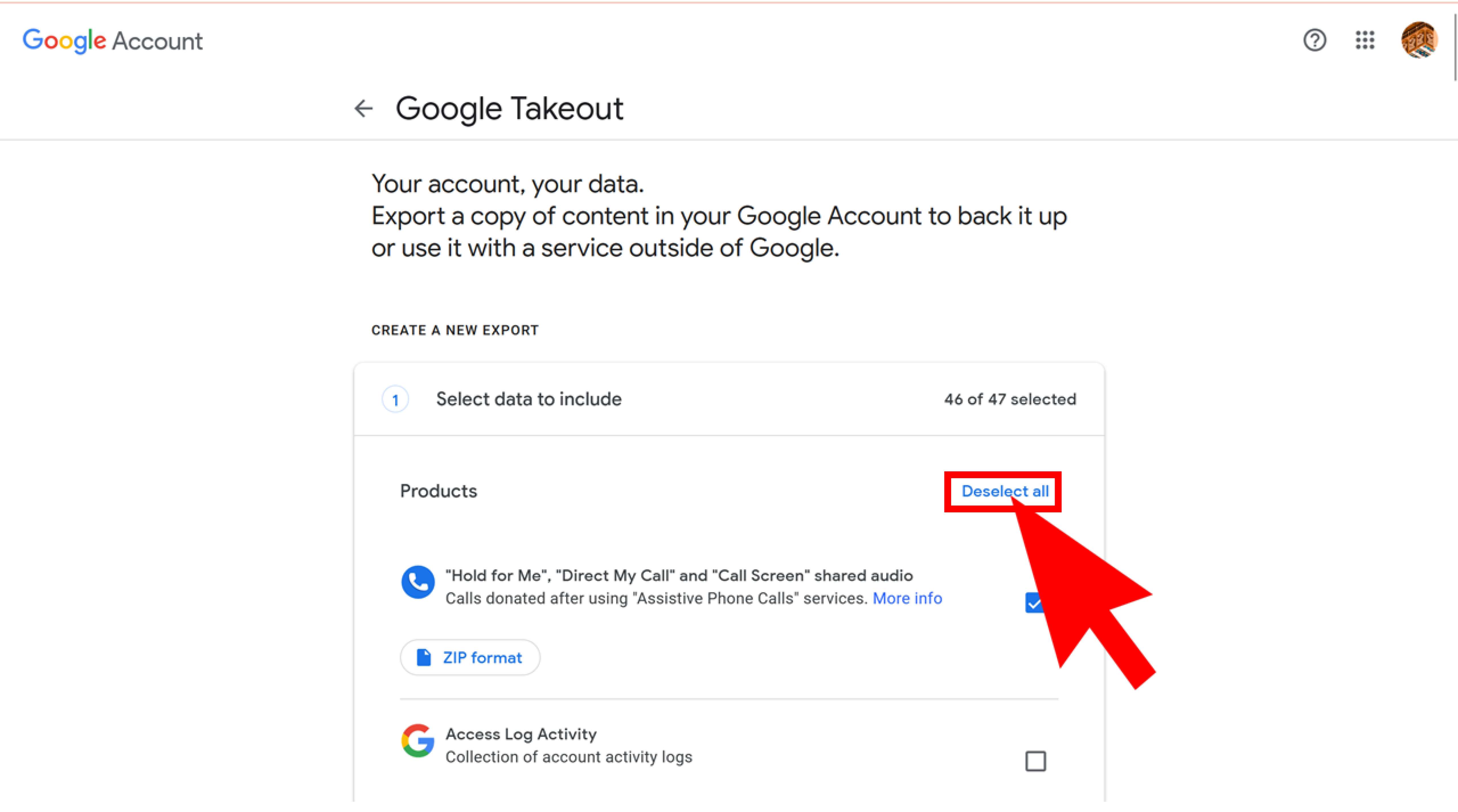
1 Select data to include 46 of 47 selected

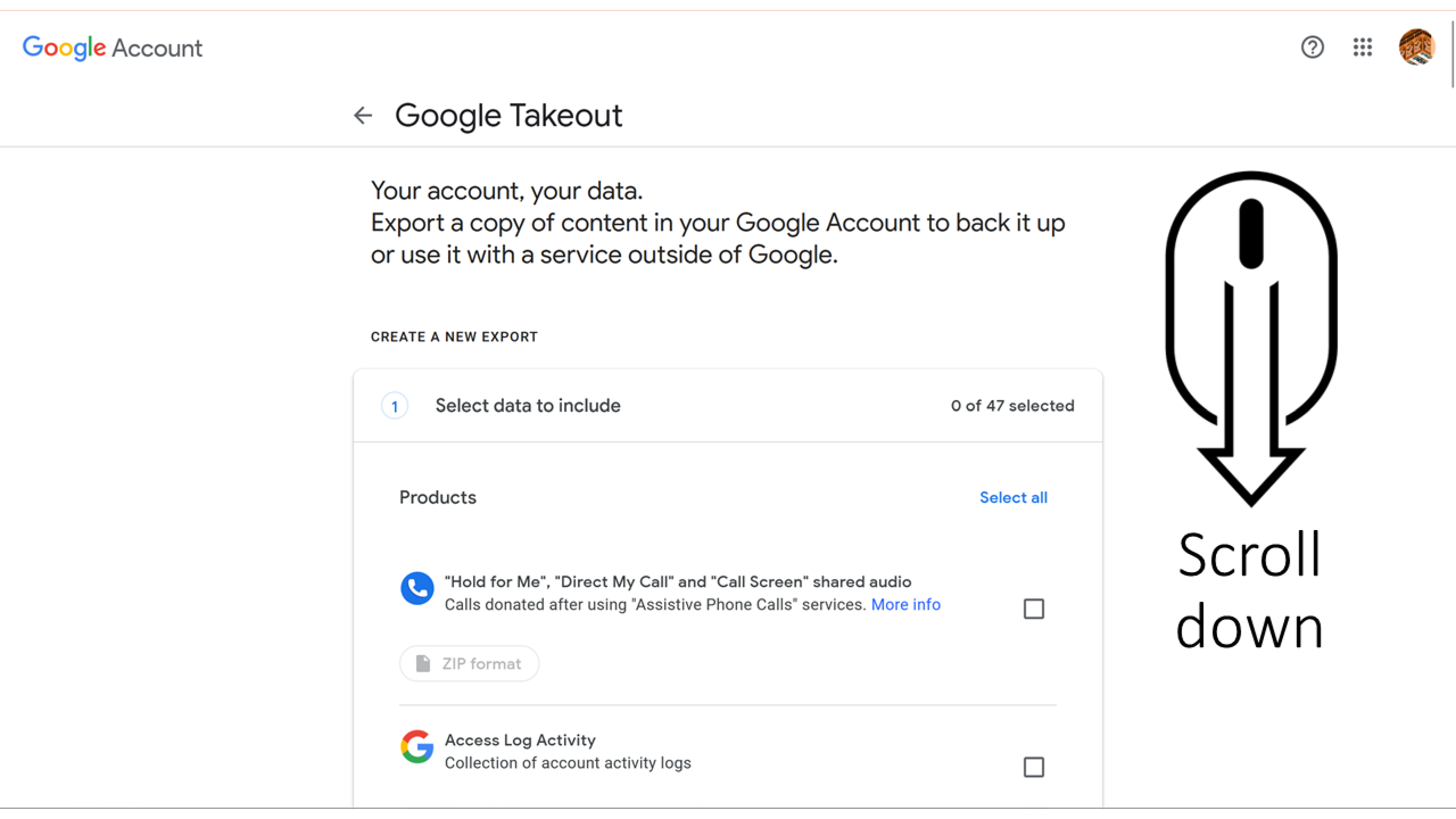
Products Deselect all

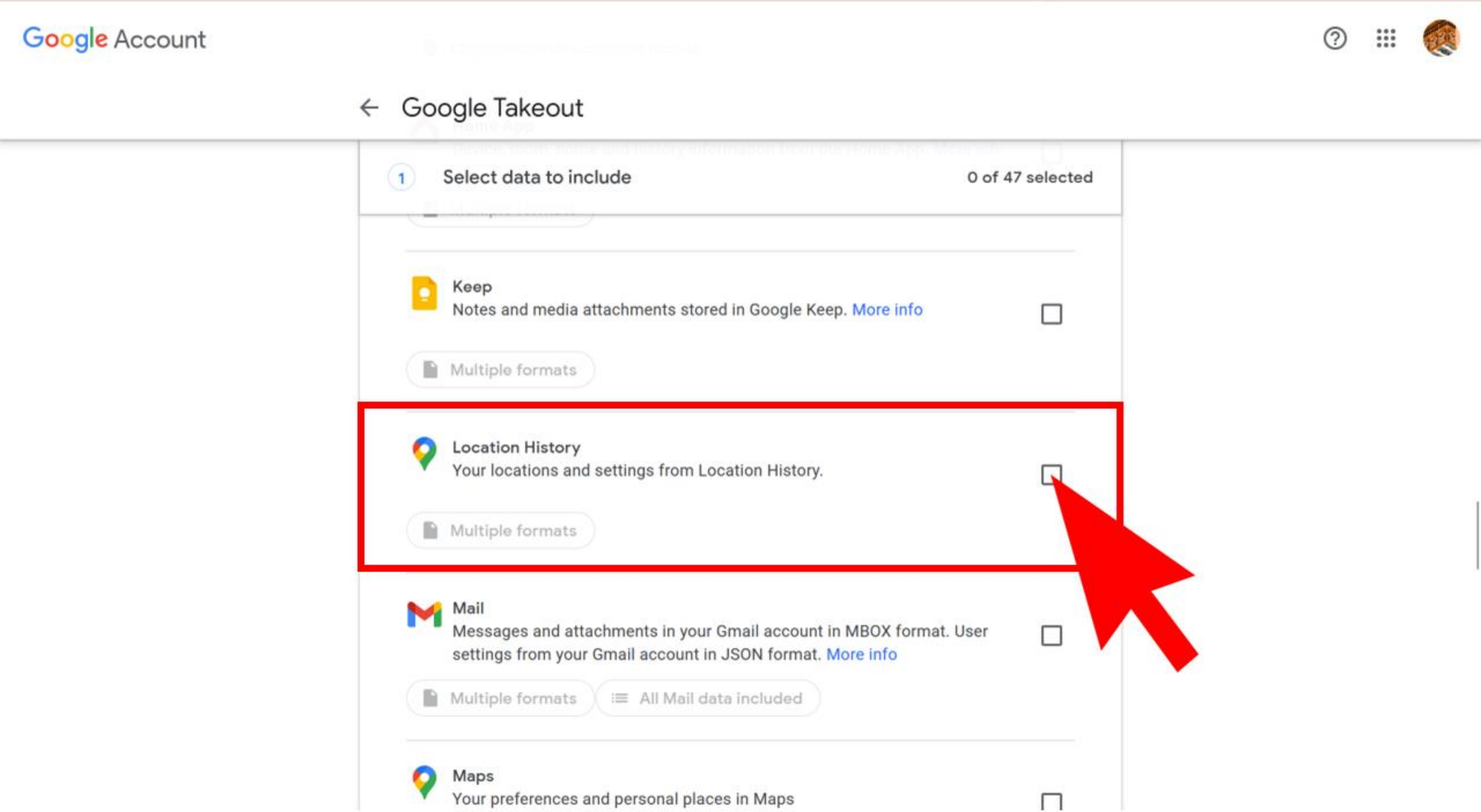
"Hold for Me", "Direct My Call" and "Call Screen" shared audio
Calls donated after using "Assistive Phone Calls" services. More info

ZIP format

Access Log Activity
Collection of account activity logs


Google Account
Google Takeout
Your account, your data.
Export a copy of content in your Google Account to back it up or use it with a service outside of Google.
CREATE A NEW EXPORT
1
Select data to include
46 of 47 selected
Products
Deselect all
"Hold for Me", "Direct My Call" and "Call Screen" shared audio
Calls donated after using "Assistive Phone Calls" services. More info
ZIP format
Access Log Activity
Collection of account activity logs


Google Account
Google Takeout
Your account, your data.
Export a copy of content in your Google Account to back it up or use it with a service outside of Google.
CREATE A NEW EXPORT
1 Select data to include
0 of 47 selected
Products
Select all
"Hold for Me", "Direct My Call" and "Call Screen" shared audio
Calls donated after using "Assistive Phone Calls" services. More info
ZIP format
Access Log Activity
Collection of account activity logs
Scroll down


Google Account
Google Takeout
1
Select data to include
0 of 47 selected
Keep
Notes and media attachments stored in Google Keep. More info
Multiple formats
Location History
Your locations and settings from Location History.
Multiple formats
Mail
Messages and attachments in your Gmail account in MBOX format. User settings from your Gmail account in JSON format. More info
Multiple formats
All Mail data included
Maps
Your preferences and personal places in Maps

Google Account

Google Takeout

1 Select data to include 1 of 47 selected

Keep
Notes and media attachments stored in Google Keep. More info

Multiple formats

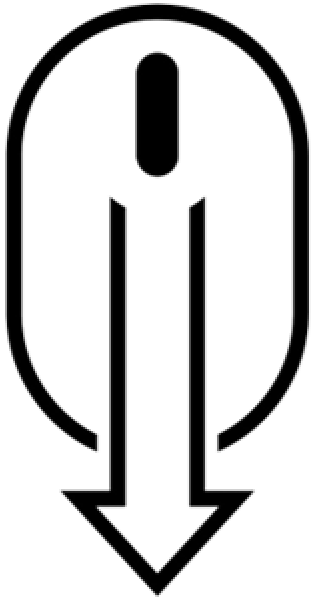
Location History
Your locations and settings from Location History.

Multiple formats

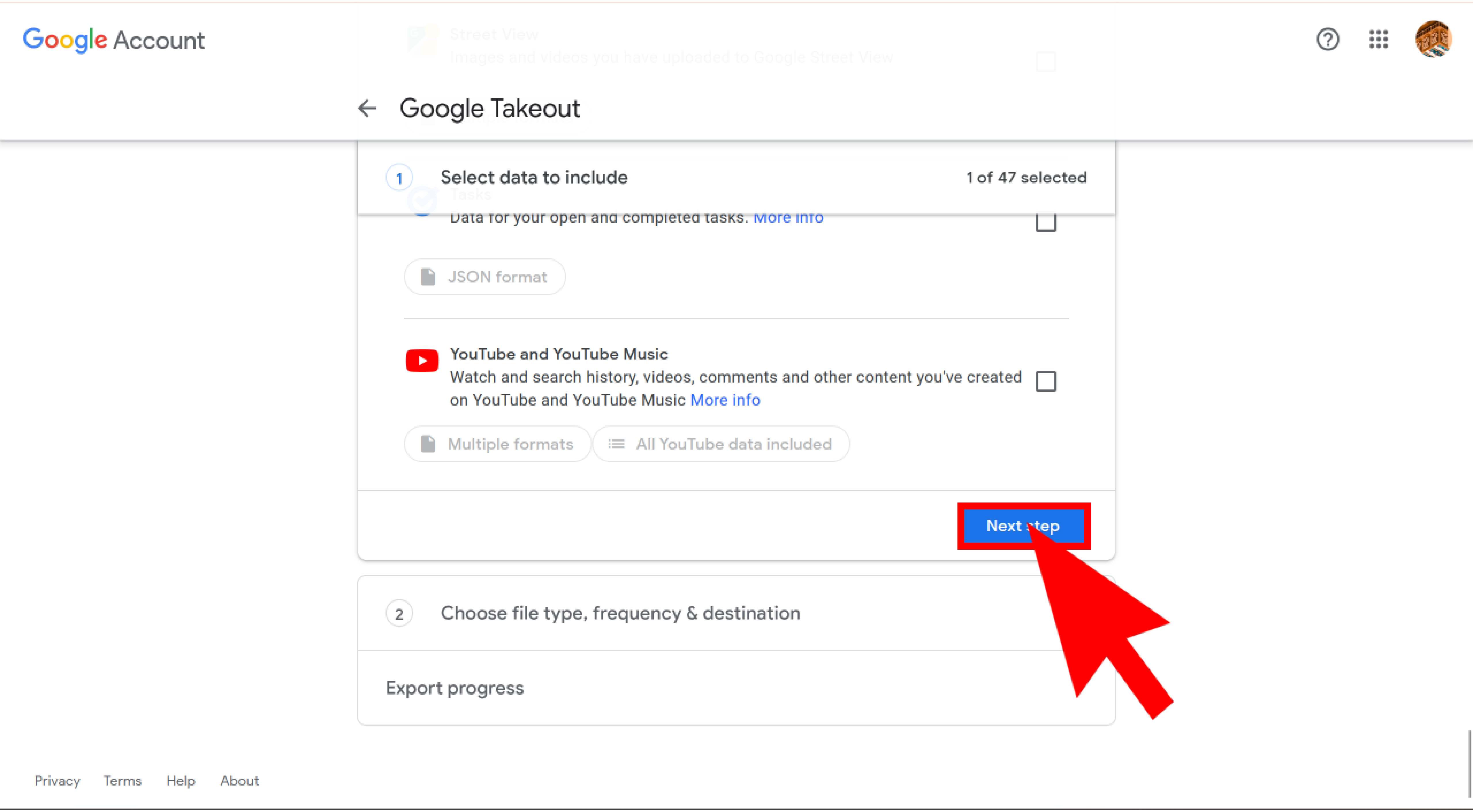
Mail
Messages and attachments in your Gmail account in MBOX format. User settings from your Gmail account in JSON format. More info

Multiple formats | All Mail data included

Maps
Your preferences and personal places in Maps



Scroll down


Google Account
Google Takeout
1
Select data to include
1 of 47 selected
Data for your open and completed tasks. More info
JSON format
YouTube and YouTube Music
Watch and search history, videos, comments and other content you've created on YouTube and YouTube Music More info
Multiple formats
All YouTube data included
Next step
2
Choose file type, frequency & destination
Export progress
Privacy
Terms
Help
About

Google Account

← Google Takeout

Your account, your data.
Export a copy of content in your Google Account to back it up or use it with a service outside of Google.

CREATE A NEW EXPORT

✓ Select data to include — 1 of 47 selected

2 Choose file type, frequency & destination

Destination

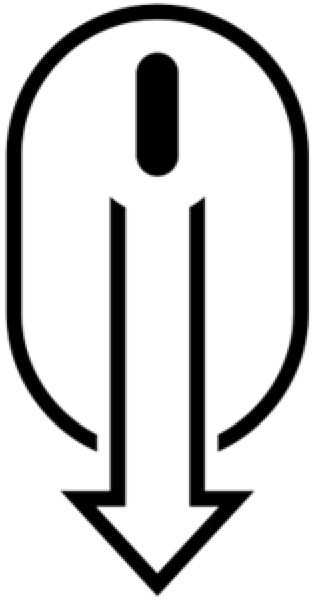
Transfer to:

Send download link via email

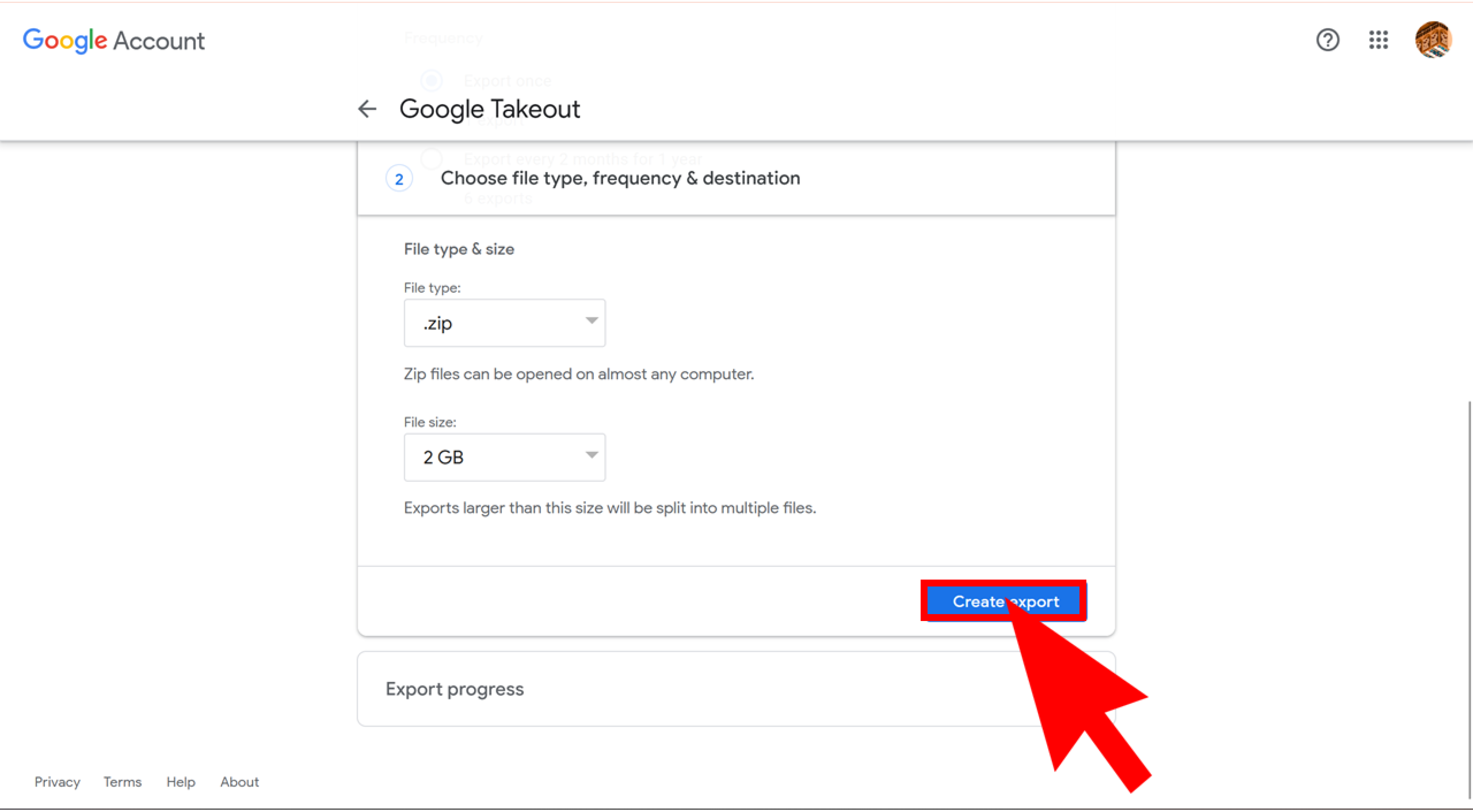
When your files are ready, you'll get an email with a download link. You'll have one week to download your files.

Frequency

Export once



Scroll down


Google Account
Google Takeout
2
Choose file type, frequency & destination
File type & size
File type:
.zip
Zip files can be opened on almost any computer.
File size:
2 GB
Exports larger than this size will be split into multiple files.
Create export
Export progress
Privacy
Terms
Help
About

Your account, your data.
Export a copy of content in your Google Account to back it up or use it with a service outside of Google.

Google Takeout

CREATE A NEW EXPORT

Select data to include — 1 of 47 selected

Choose file type, frequency & destination

Export progress

Google is creating a copy of files from Location History
This process can take a long time (possibly hours or days) to complete. You'll receive an email when your export is done.
Created: April 28, 2023, 4:34 PM

Cancel export

+ Create another export

Demo: Data Donation

Try out the Port software

- If you have requested a package, try out the software
- **No data** will be stored or sent anywhere, this is just a demo!

iHealth (step extraction):
https://eyra.github.io/port-activity-pilot-ihealth/

Google Location History (walking and biking extraction):
https://eyra.github.io/port-activity-google-location-history/

Additional resources

Selected resources for app development

- Commercial/Off-the-shelf existing platforms
  - Movisens: https://www.movisens.com/en/
  - MOTUS: https://www.motusresearch.io/en
  - Murmuras: https://murmuras.com/

- Commercial app builders (usually no special knowledge required)
  - Appypie https://www.appypie.com
  - Ethica Data: https://ethicadata.com/

Selected resources for app development

- App builders for specific OSs (require some programming knowledge)
  - Apple Research Kit: http://researchkit.org/
  - ResearchStack for Android: http://researchstack.org/

- Open source platforms/frameworks (require programming knowledge)
  - AWARE: https://awareframework.com/
  - Beiwe Research Platform: https://www.beiwe.org/
  - PACO: https://pacoapp.com/

Selected resources for EMA/ESM

- Specific EMA/ESM software
  - mEMA: https://ilumivu.com
  - ExpiWell: https://www.expiwell.com/
  - LifeData: https://www.lifedatacorp.com/ecological-momentary-assessment-app-2/
  - SEMA3: https://sema3.com/
  - Other online survey software, such as Blaise5 (https://blaise.com/products/blaise-5), can be used as sample management system that can send surveys at specific time
- Myin-Germeys, Inez, and Peter Kuppens. (Eds.). 2022. *The open handbook of experience sampling methodology: A step-by-step guide to designing, conducting, and analyzing ESM studies.* (2nd ed.) Leuven: Center for Research on Experience Sampling and Ambulatory Methods Leuven

Other resources

- For visualization of location data:
  - Shiny app Utrecht University (R code): https://github.com/sobradob/shinyapp
- For data processing:
  - R package for log data analysis (Stachl): https://osf.io/ut42y/

Our book…

Keusch, Florian, Bella Struminskaya, Stephanie Eckman, and Heidi Guyer. forthcoming. *Data Collection with Wearables, Apps, and Sensors.* https://bookdown.org/wasbook_feedback/was/

Exercise:
Apple Health Data

- If you have an iPhone or an Apple watch (or your new course-friend with an iPhone graciously shares data with you)
- Download your data, prepare it for analysis and find out something about yourself (see Blackboard for code)
  - Go to Health on your iPhone
  - Click on the icon ‘Personalize’
  - Click on Export All Health Data
- Hand in the finished exercise per email to b.struminskaya@uu.nl