Utrecht University

Designed big data

Survey Data Analysis 2020

Bella Struminskaya

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

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 (*6*). 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

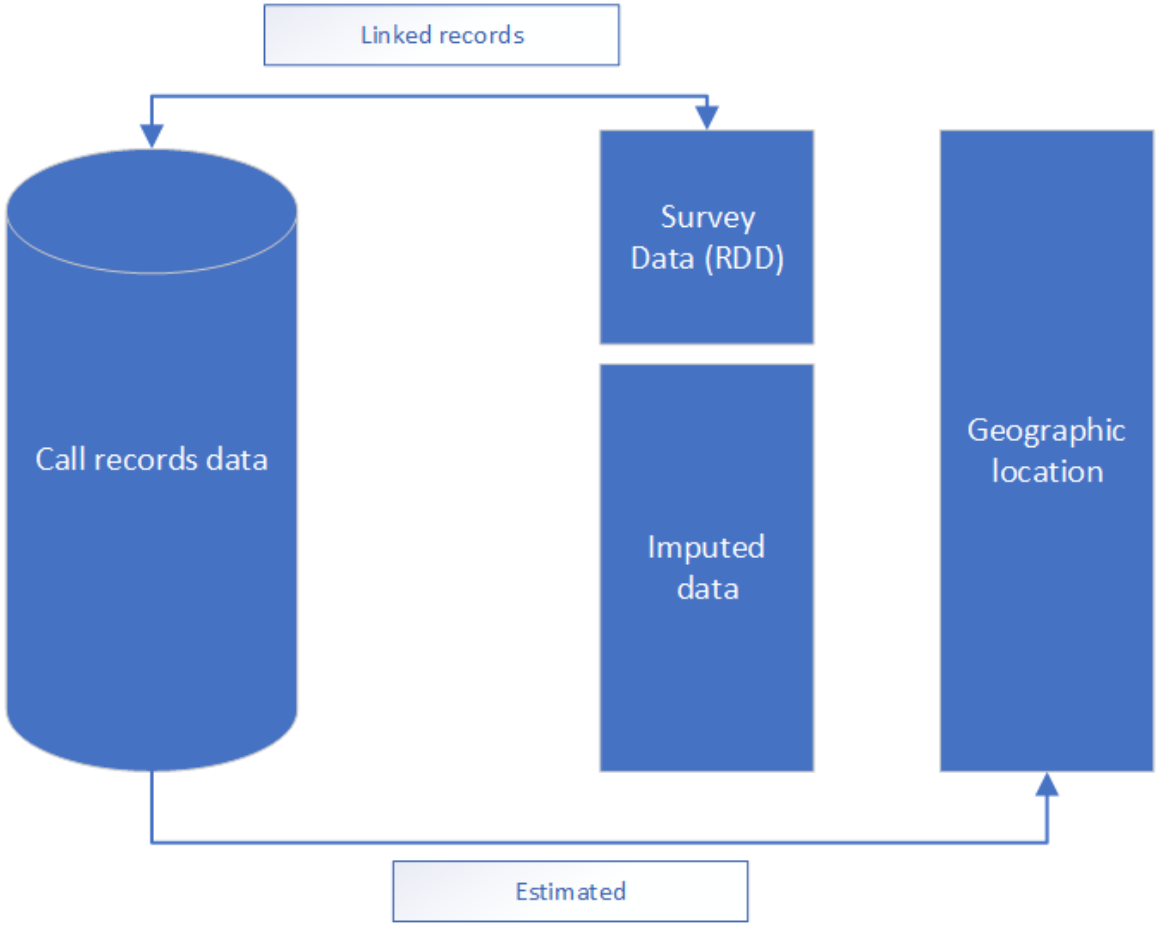
[1]Information School, University of Washington, Seattle, WA 98195, USA. [2]Department of Computer Science and Engineering, University of Washington, Seattle, WA 98195, USA. [3]School of Information, University of California, Berkeley, Berkeley, CA 94720, USA.
*Corresponding author. E-mail: joshblum@uw.edu

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)

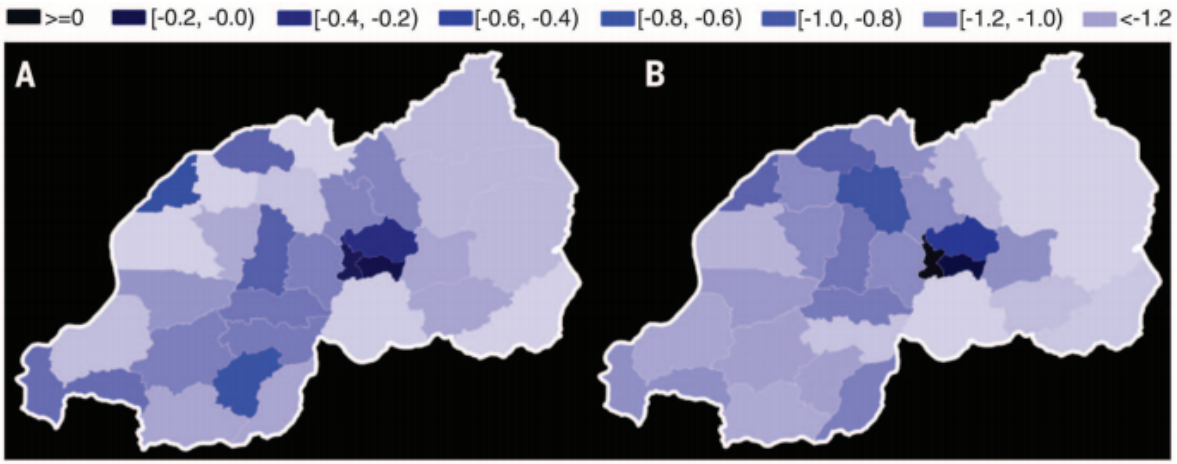
Can anonymized data from mobile phone networks predict poverty and wealth?



- 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)

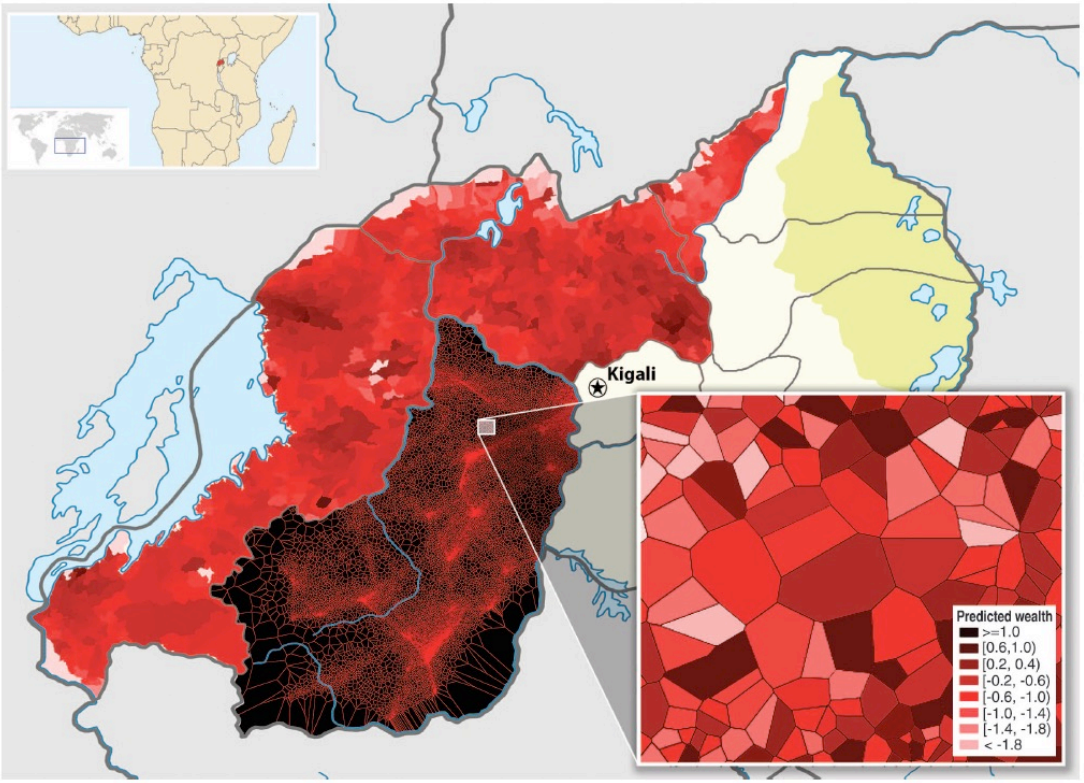


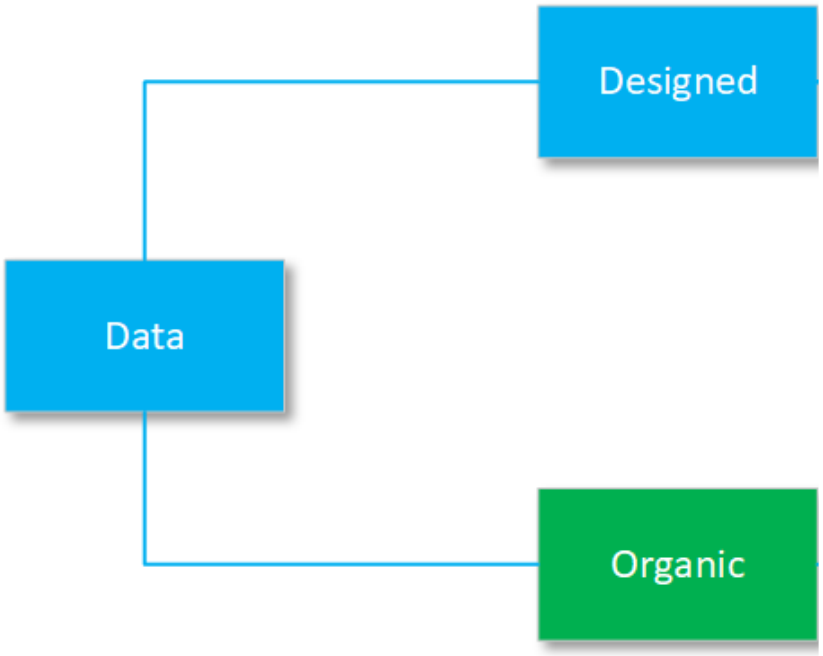
Composite wealth index: A – predicted from call data, B – actual from DHS, *r*=0.79

(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)



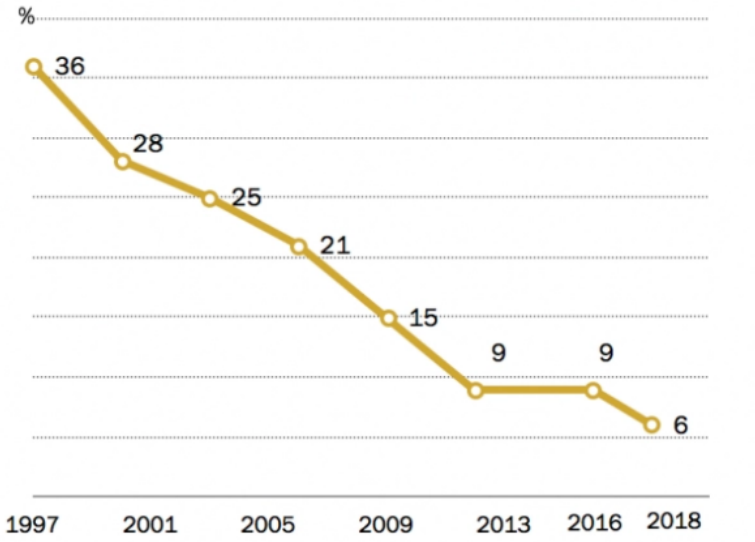


(Source: Kreuter 2018)

Decreasing response rates

After brief plateau, telephone survey response rates have fallen again

Response rate by year (%)

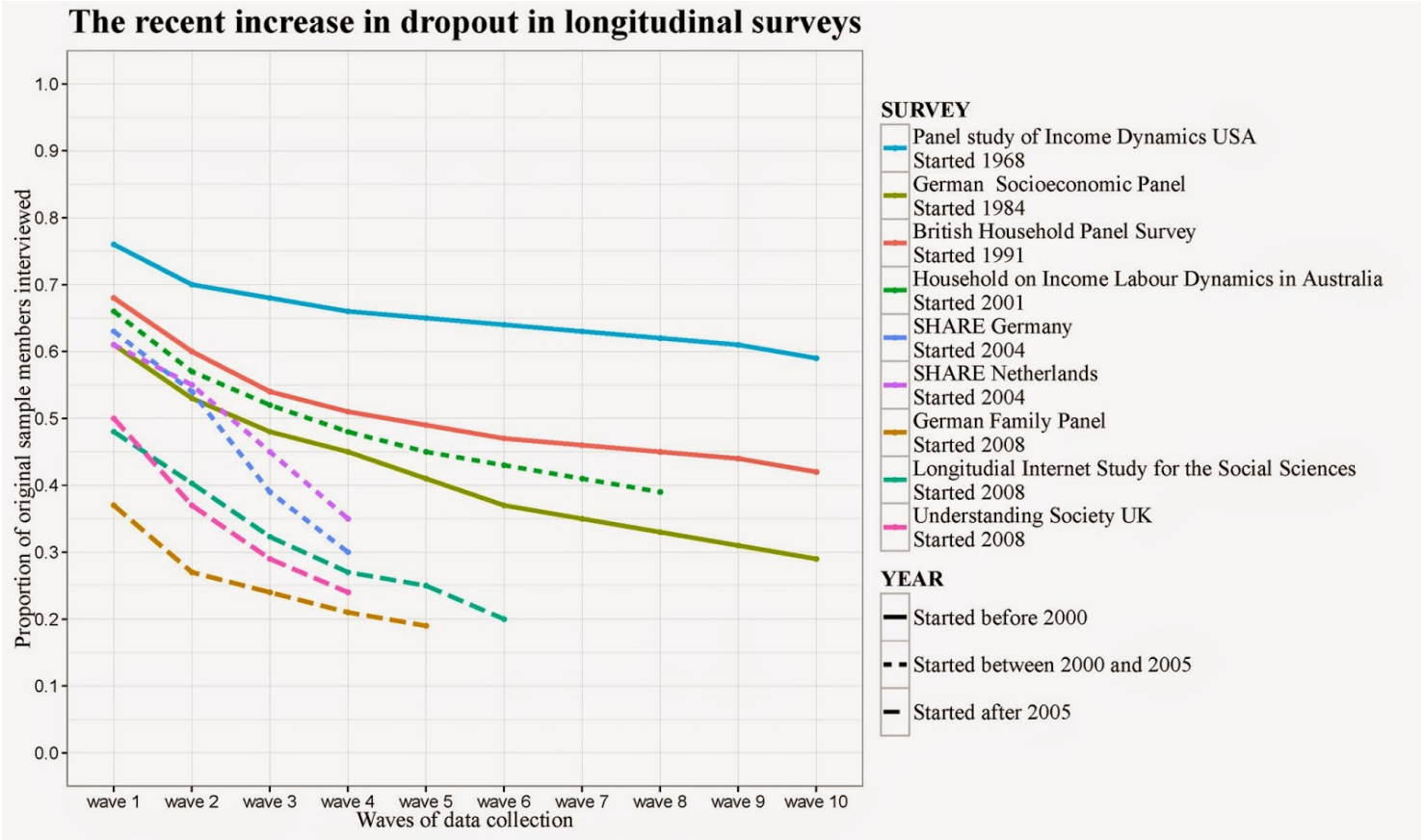


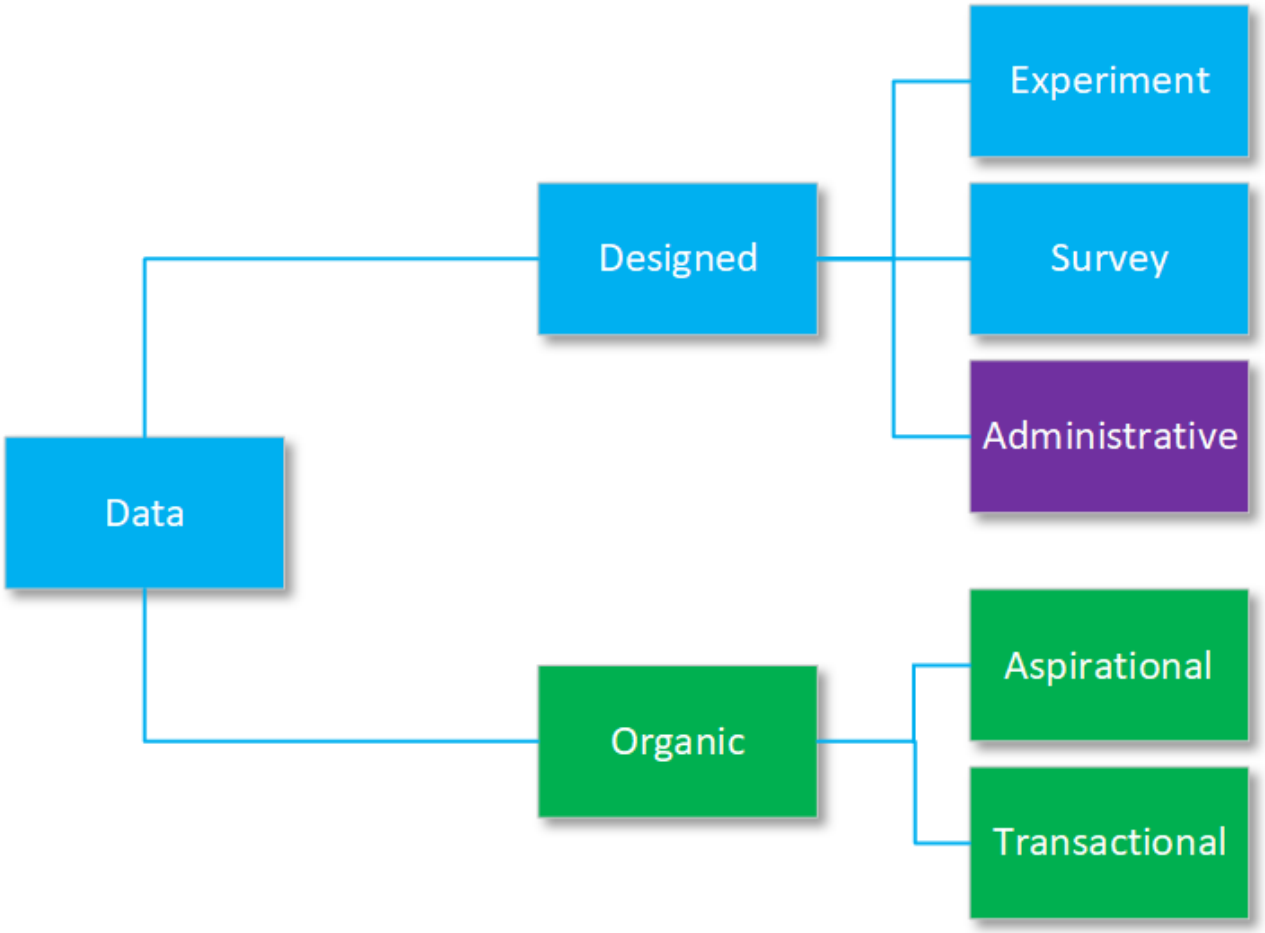
Note: Response rate is AAPOR RR3. Only landlines sampled 1997-2006. Rates are typical for surveys conducted in each year.
Source: Pew Research Center telephone surveys conducted 1997-2018.

PEW RESEARCH CENTER

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

Cumulative nonresponse rates in large panel surveys (peterlugtig.com)





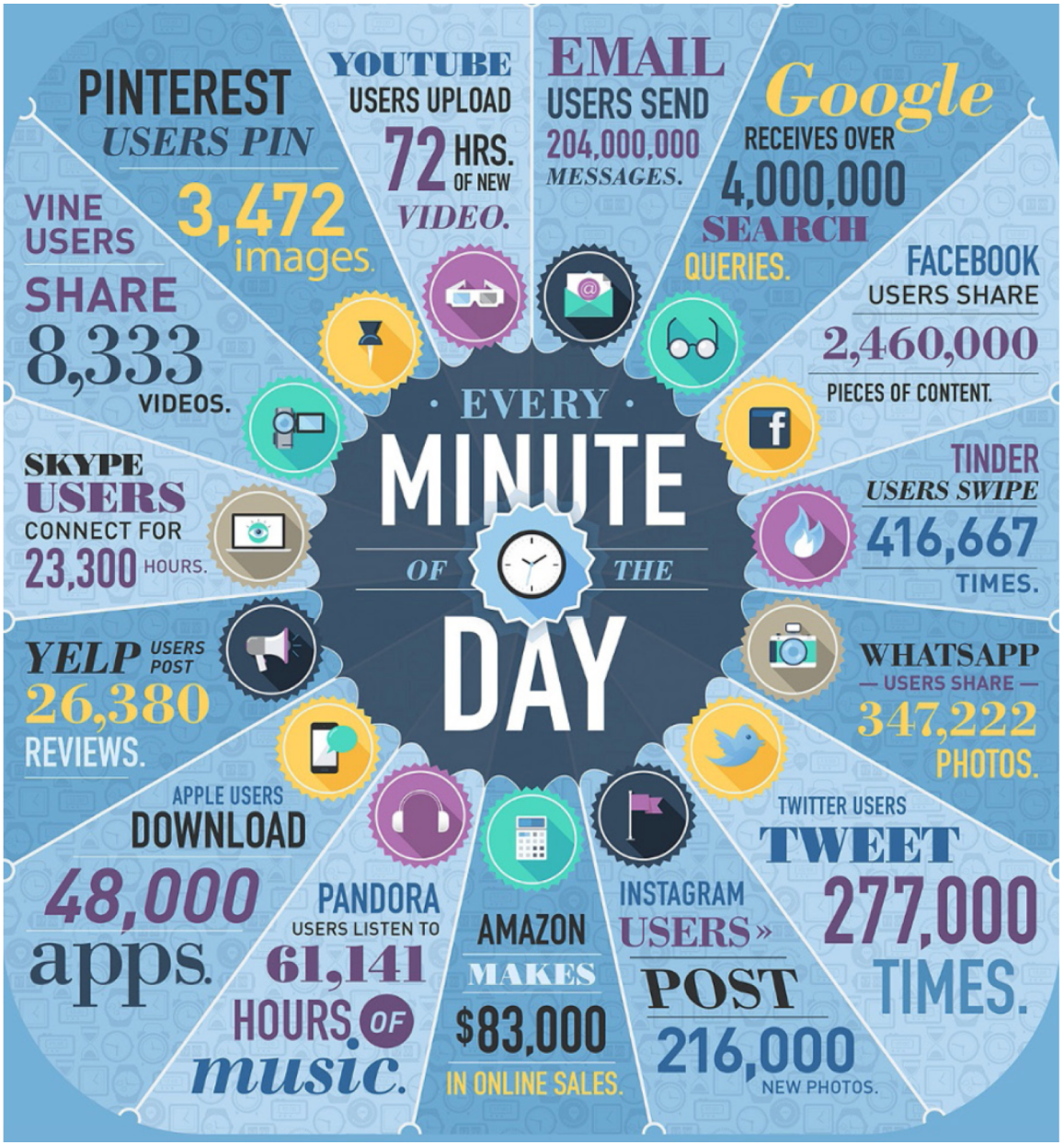
(Source: Kreuter 2018)

Organic / big / found data sources:

1) *Transaction data*: describe an event, e.g., a person interacts with a business or a government entity

2) *Social media data*: scraped from social networks, blogs, web searches etc. E.g., Google Flu Trends

3) *Internet of Things (IoT) data*: data collected from interconnected devices such as autos, household appliances, security cameras, wearable sensors, GPS locators etc. E.g., gathering data on movement of people and things, electricity use (lifestyle & rhythms of daily life)



Characteristics of big data:

1) Volume
2) Variety ((no) structure)
3) Velocity
4) Veracity (accuracy)
5) Variability (differences in meanings across sources)
6) Value
7) Visualization

Source: Baker 2017, Infographic: James 2014

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)

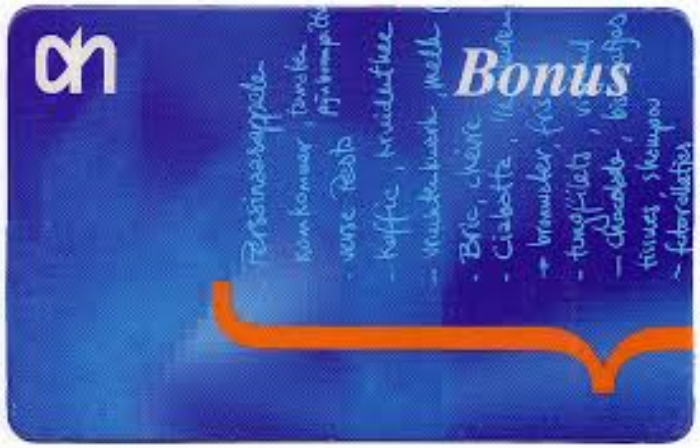
Types of big data

- Types of big data
  - **Administrative data** – provided by persons or organizations for regulatory or other government activities
  - **Transaction data** – generated as an automatic byproduct of transaction and activities (e.g., credit card data, traffic flow data)
  - **Social media data** – created by people with the express purpose of sharing with (some) others
  - **Sensor data** – GPS, accelerometers, heartbeat

Administrative data

- Statistics Netherlands
- Business administration
- Market data
- Pros
    1. Accuracy (?)
    2. Costs (?)
    3. Speed (?)
- Cons
    1. Missing data?
    2. Reliability: data collection and definitions the same?
    3. Validity:
        - Are they measuring what YOU want to measure?
        - Do you KNOW the definitions of variables? Are they yours?

Transaction data



-Data Availability? Often proprietary! A key strength of surveys is public access to data, permitting **replication and reanalysis**
-Not everyone uses cards!
-Knowing what people buy is not the same as understanding WHY they buy!

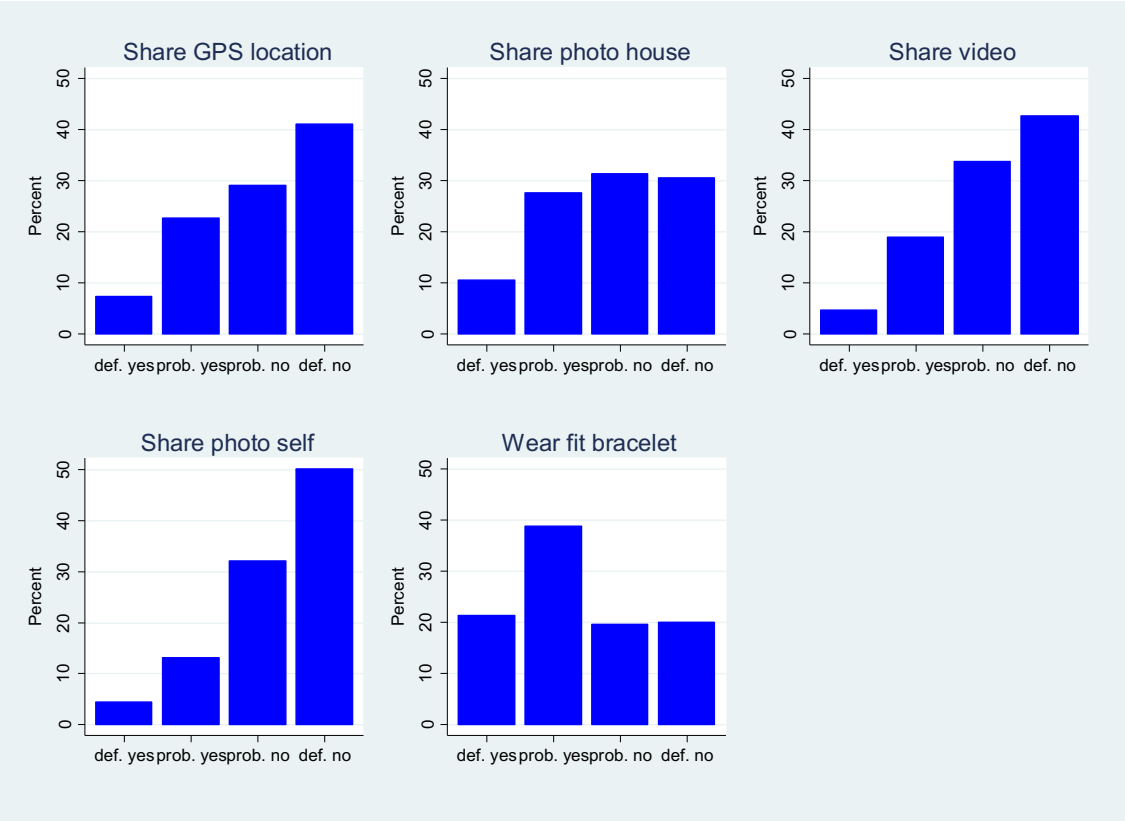
Social Media Data

- Selection bias: “haves” versus “have-nots”
  - Not everyone uses social media!
  - Need to distinguish between producers and users of users of social media – small part of online population actively tweets
- Measurement bias
  - Self-presentation bias: Impression management is a key element of social media
  - The average Facebook user has MANY “friends”

Sensor data

-Not everyone allows you to track them: 47% share their GPS coordinates (Struminskaya et al. 2018)
-some type of activities are difficult to measure
-thresholds for intensity are arbitrary

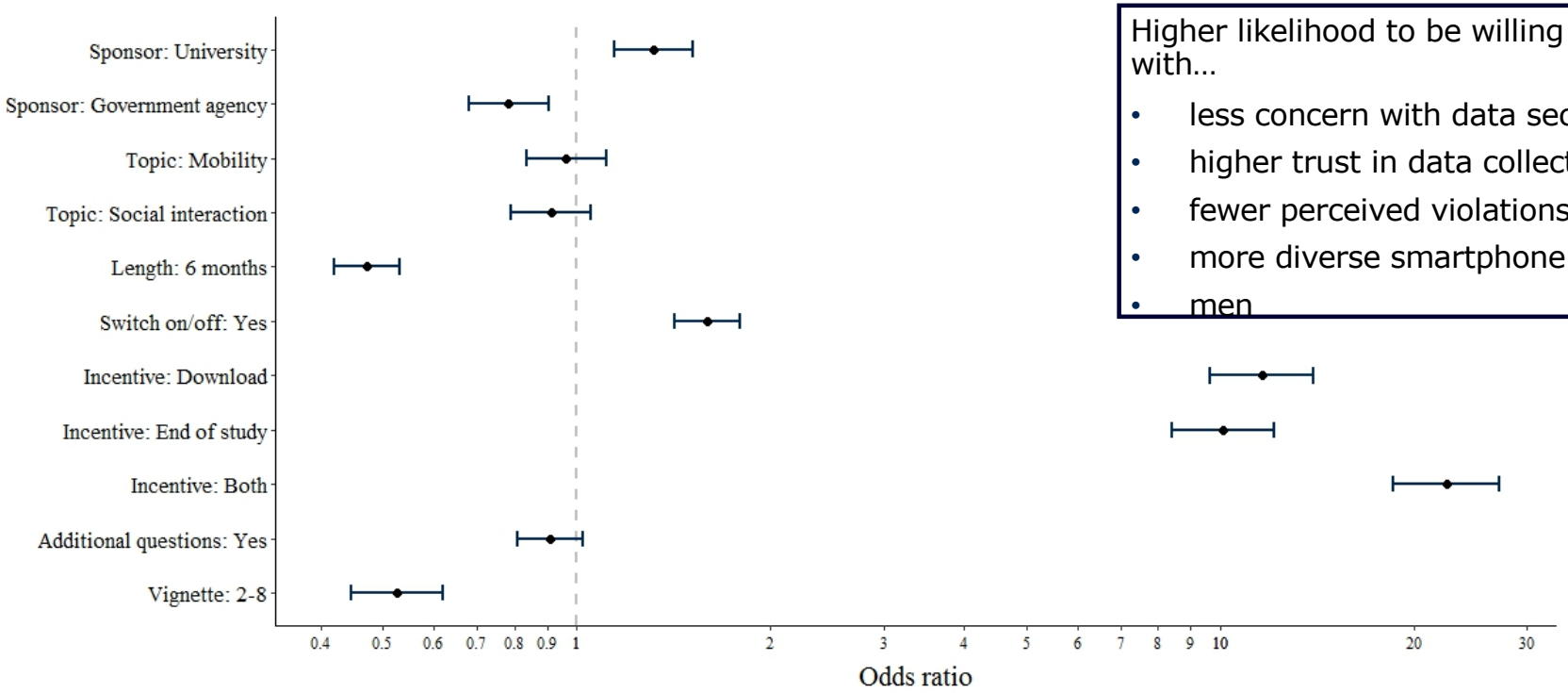
Willingness to collect sensor data



LISS Panel
November 2017
N= 3023
RR = 88.4%

(Struminskaya et al. forthcoming)

Willingness to download research app



Higher likelihood to be willing to participate with…

- less concern with data security
- higher trust in data collecting institutions
- fewer perceived violations of privacy
- more diverse smartphone activities
- men

Odds ratios (points) with 95%-CI (lines) from multilevel logistic regression. DV: Dichotomized willingness to participate in passive mobile data collection. n=1,947 German smartphone users

Keusch et al. (2019)

Reasons (not) to participate

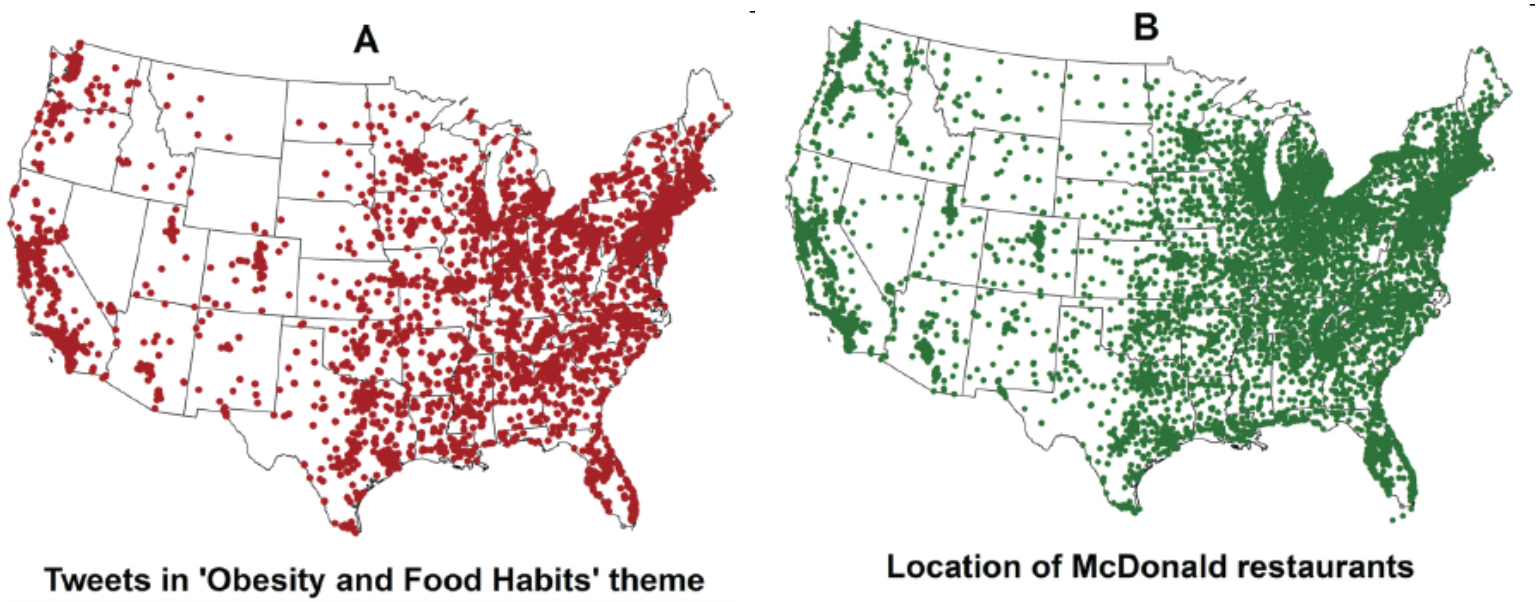
Reasons not to participate		Reasons to participate	
Privacy, data security concerns	44%	Interest, curiosity	39%
No incentive, incentive too low	17%	Incentive	26%
Not enough information provided	12%	Help research, researcher	18%
Generally do not download apps	8%	Seems legitimate, safe	11%
Not interested	6%	Will make products & services better	7%
Not enough time/Study too long	5%	No additional effort	6%
Don't use smartphone enough	5%	Like research & surveys	4%
Not enough storage	1%	Fun	2%
Other reasons	6%	Other reasons	9%
n=1,154		n=900	

N=1947 German smartphone users

Keusch et al. (2019)

Few covariates:

Obesity-Related Tweets and McDonalds Restaurants



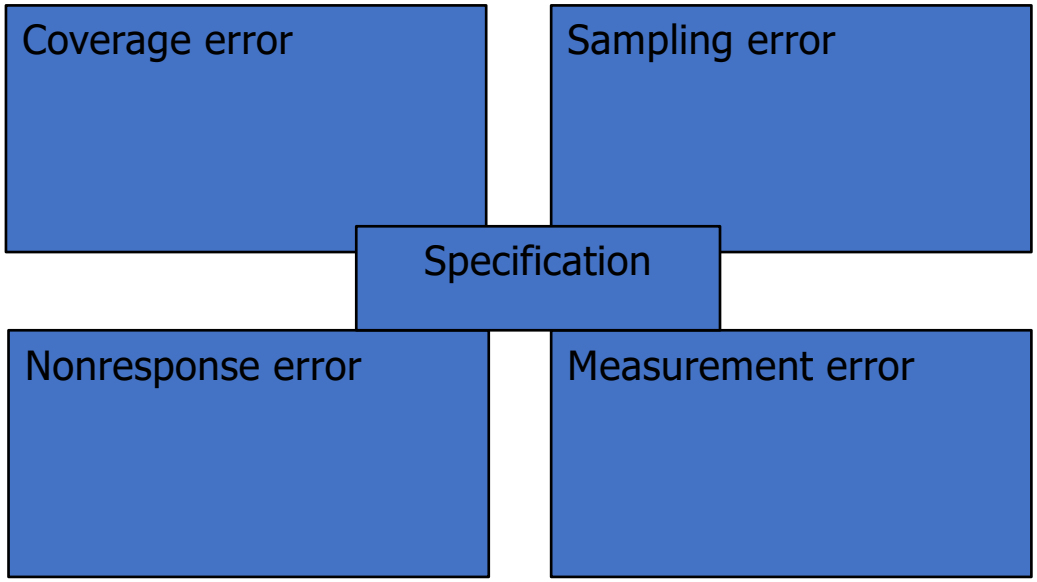
- Ghosh and Guha (2013) report a "strong correlation" between the two
- Any alternative explanation come to mind?

Lack of stability

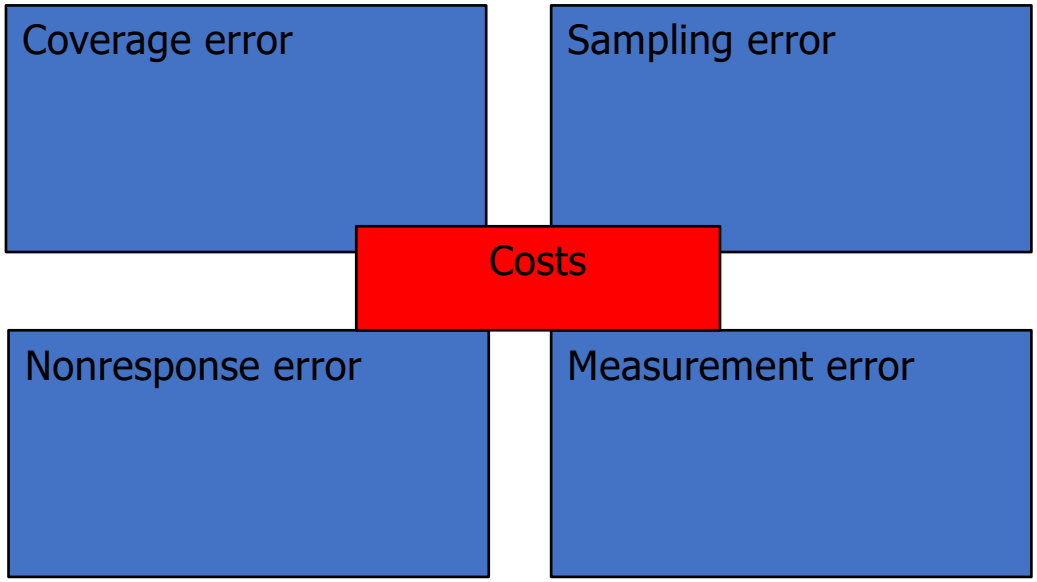
- *Technological speed:*
- What will Facebook or Twitter look like 5 or 10 years from now?
- Mobile phones and sensors look very different now than 10 years ago

- Big data may be good for measuring short term trends, but surveys may be better for longer-run measurement

Total Survey Error/Cornerstones of Survey Research (see Hox et al., 2008)



Total Survey Error/Cornerstones of Survey Research (see Hox et al., 2008)



Coverage in Big Data

- Does the sampling frame include all units of the population?

	High coverage?	Difference frame /population	Adjustment possible
Administrative data	yes, except for people that are not registered, e.g. illegals, homeless	++	Via snowball sample
Transaction data	Only those that pay with cards	+	Cash survey/observation
Social media data	Only those that are on social media	-	General population survey
Sensor data	Only those that wear sensors and allow you to track them	-	Nonresponse survey

Sampling in Big Data

- No differences between big data and survey data
- Is sampling necessary?
  - Often no additional costs for using census instead of sample
- Often the unit of observation is not the individual
  - Transaction with transaction data
  - Verbal comment with social media data
  - Data capture point with sensor data
    - Recode into individual data
    - Dependent observations (many observations from few individuals)

Administrative data	++
Transaction data	+
Social media data	--
Sensor data	--

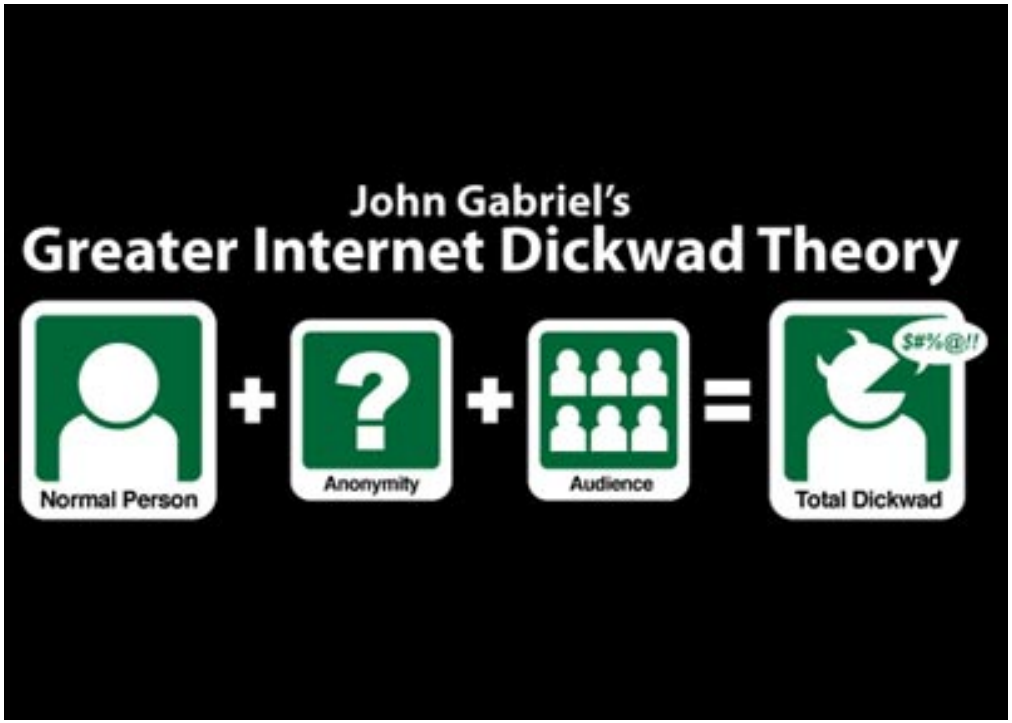
Nonresponse in Big Data

- Sampled but not collected
- In principle, there is no non response in big data
- Nonresponse error/bias is an important issue in surveys
- Check
  - Compare big data to external data
  - Investigate internal variation within the data, e.g. difference in estimates by the number of verbal comments in social media data
  - Examination of adjustment estimates, where each adjustment contains different assumptions about nonresponse

Administrative data	++
Transaction data	++
Social media data	-
Sensor data	--

Measurement in Big Data

- Is the data well-constructed, clear, and not leading or otherwise biasing? (AAPOR report survey quality, 2016)
- Do people provide truthful data?
- Were any respondents removed?



Measurement in Big Data

- Is the data well-constructed, clear, and not leading or otherwise biasing? (AAPOR report survey quality, 2016)
- Administrative data: do you know definitions? Are they the same as yours? Over time?
- Transaction data: accurate?!
- Social media data: Do people provide truthful data?
- Sensor data:
  - Objective weight is about 1 kilo lower than reported weight (Koorenman & Scherpenzeel, 2014)
  - Automatic trip detection with sensors (Geurs et al., 2015): inaccurate with small trips, public transport trips not classified, unsuccessful mode detection in 25% of trips

Administrative data	-
Transaction data	++
Social media data	--
Sensor data	-

Specification in Big Data

- Formulating and answering research questions
  - The construct implied in the data differs from the intended construct that should be measured (validity)
  - Problems of wording, context, concepts
  - Ask what is essential for the research question
- Check with qualitative techniques/interviews

Administrative data	**+/-**
Transaction data	+/-
Social media data	+/-
Sensor data	+/-

Costs in Big Data

- Big data is already out there, so little costs involved!

Administrative data	**++**
Transaction data	++
Social media data	++
Sensor data	+/-

Big data particularly useful for

- Replace surveys/most survey questions
  - Travel
  - Budget
  - User groups/online communities

- 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?

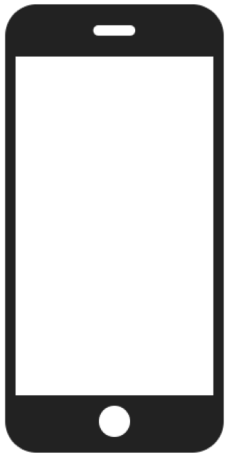
Passive data collection using smartphone sensors

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

Compass

Barometer

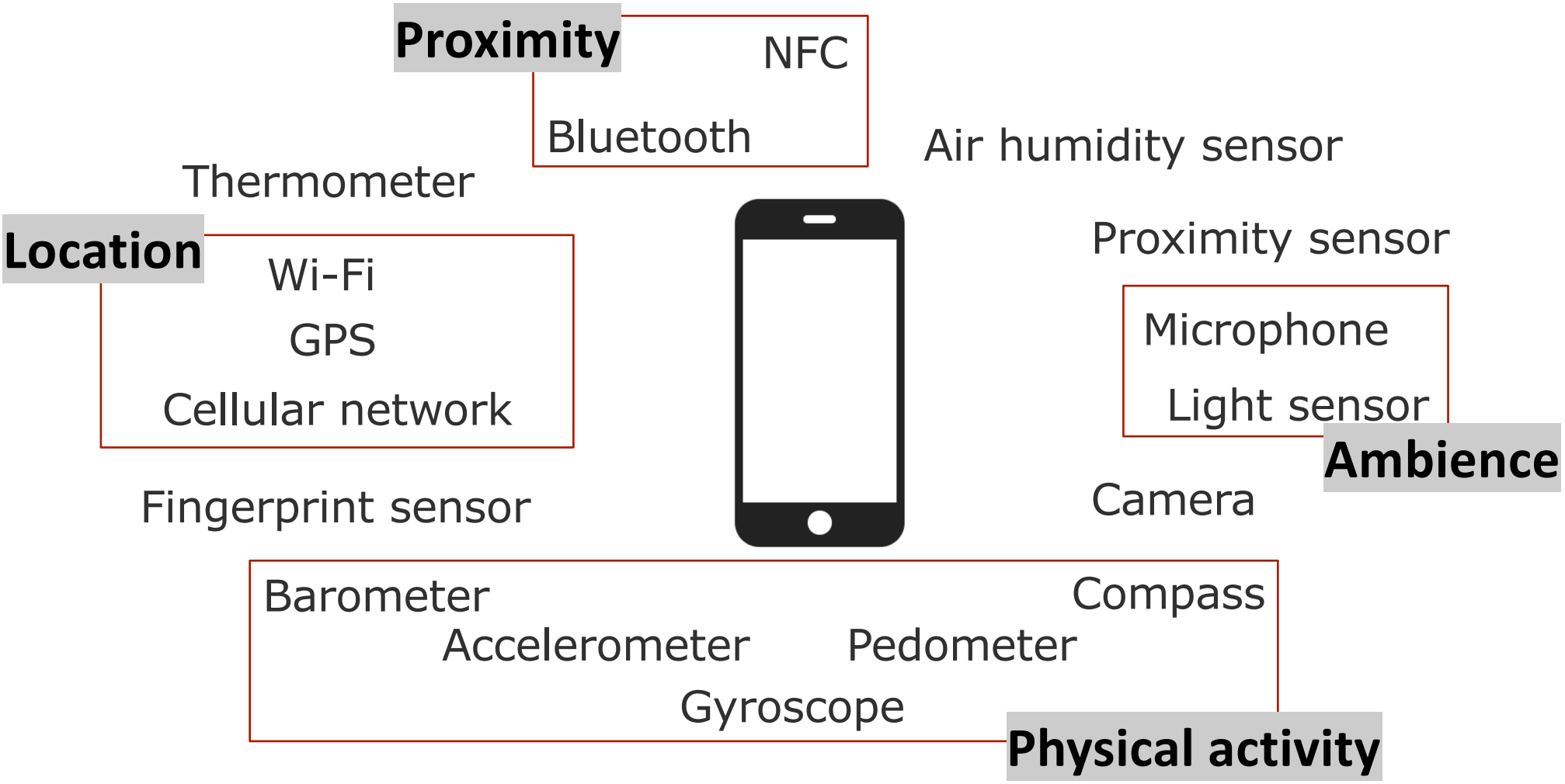
Accelerometer

Pedometer

Gyroscope

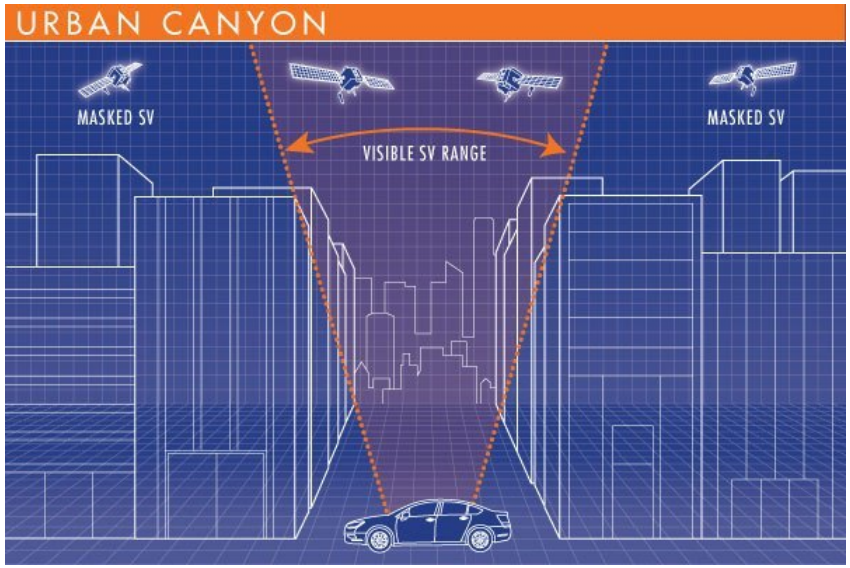
What can you measure with these sensors?

Native smartphone sensors



Location sensors

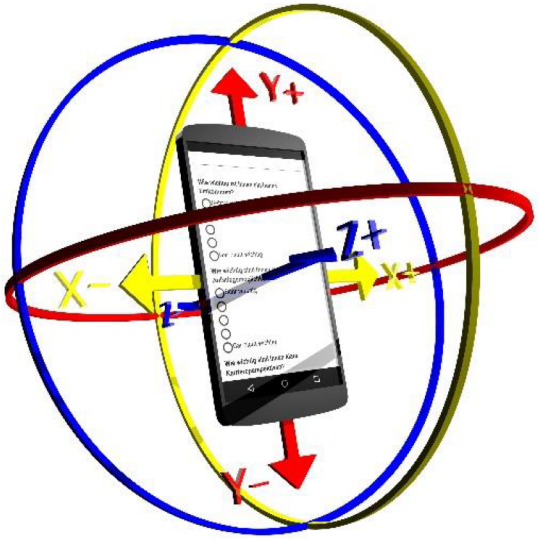
- GPS
  - Coordinates in longitude & latitude
  - High accuracy (newest generation 30 cm)
  - Works without cell/Internet connection
  - Performs worse in ‘urban canyons’, indoors, and underground (pseudo accuracy!)
  - Can be battery-draining
- Cellular network
  - Multilateration of radio signals between (several) cell towers
- WiFi
  - Inferring location from Wi-Fi access points (AP)
- Beacons
  - Bluetooth transmitters for indoors

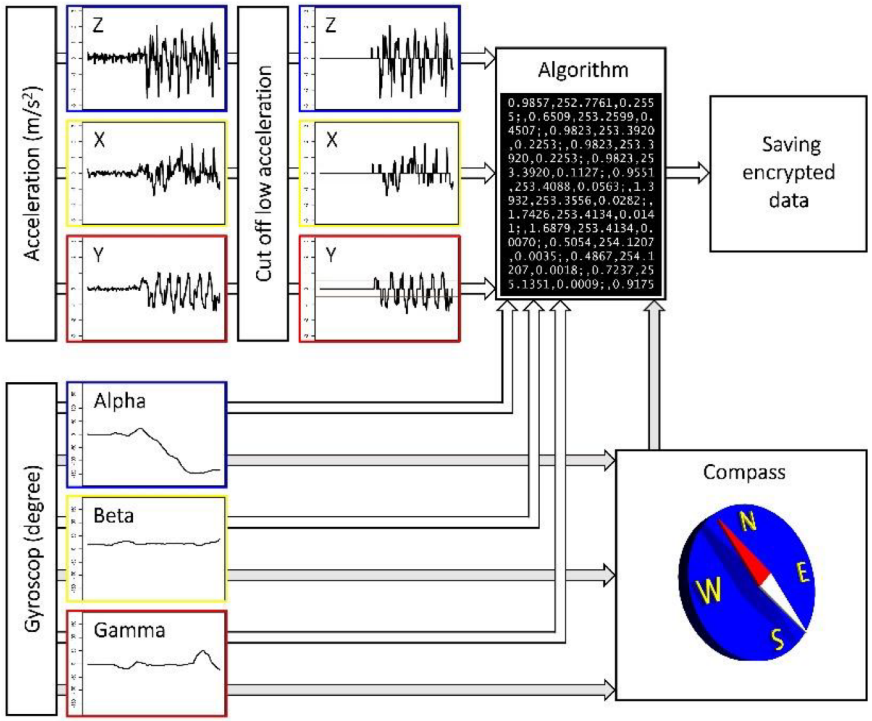


Picture source: https://i2.wp.com/geoawesomeness.com/wp-content/uploads/2014/01/urbancanyon.jpg?fit=600%2C400&ssl=1, https://locatify.com/wp-content/uploads/2015/03/beacon-wall-756x425.jpg

Physical activity sensors

- Accelerometer
- Gyroscope
- Mangnetometer
- Barometer
- Pedometer





Source: Schlosser et al. (2019)

Ambience sensors, proximity sensors

- Camera
  - photos, videos, scanning of bar codes
  - heart rate
  - linear distance
- Microphone
  - active and passive (ambient noise) recording
- Light sensor
  - e.g., identify idle state

- Bluetooth
- RFID
- NFC

Source: Jäckle et al. (2018)

...a selection of research questions

- How do environmental factors affect happiness? (MacKerron & Mourato 2013)
- How do people interact in large social networks? (Stopczynski et al. 2014)
- How much do households spend on goods and services? (Jäckle et al. 2019; Wenz et al. 2018)
- What food and drinks do Americans acquire? (Yan et al. 2019)
- How do people find work after prison? (Sugie 2018)
- Does mental health of students change over the course of a term? (Wang et al. 2014)
- How do people move around in everyday life? (McCool et al. in preparation)
- What are the effects of unemployment? (Kreuter et al. 2018)

StudentLife App: Combining sensor data with self-reports

automatic continuous sensing

- accelerometer
- microphone
- light Sensor
- GPS/Bluetooth

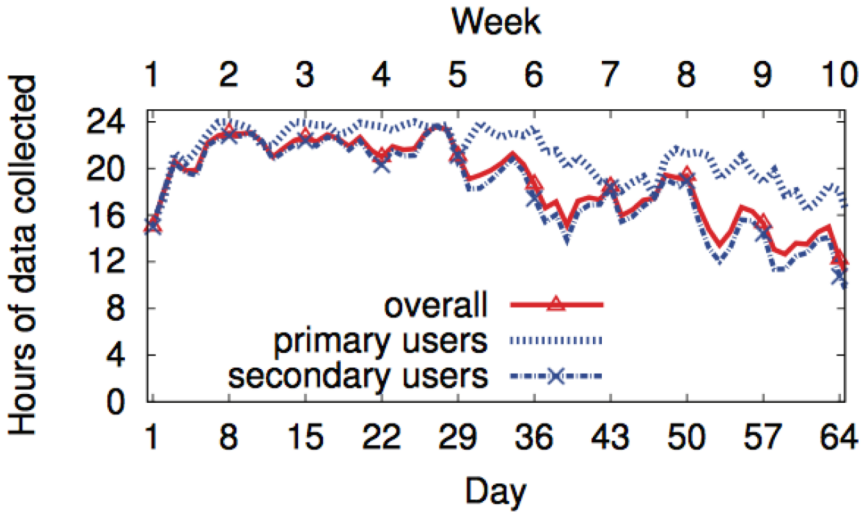
self-reports

- SurveyMonkey
- mobile EMA

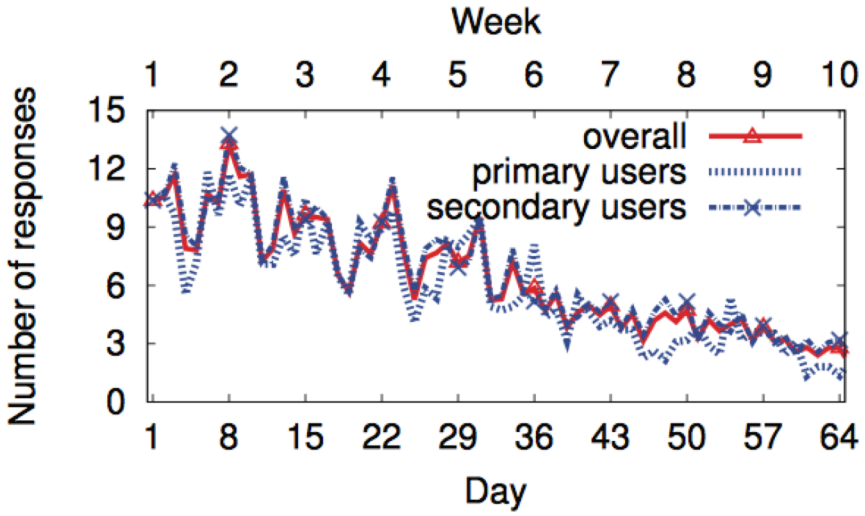
Wang et al. 2014

StudentLife Study Design

- Participants: 48 students of Darmouth College (USA)
- Duration: 10 weeks
- Smartphones: Andriod provided to students (37) or own (11)
- EMAs: about 8 times per day
- Pre- and post-survey
- Incentives: T-shirt and lottery
- Feedback: none

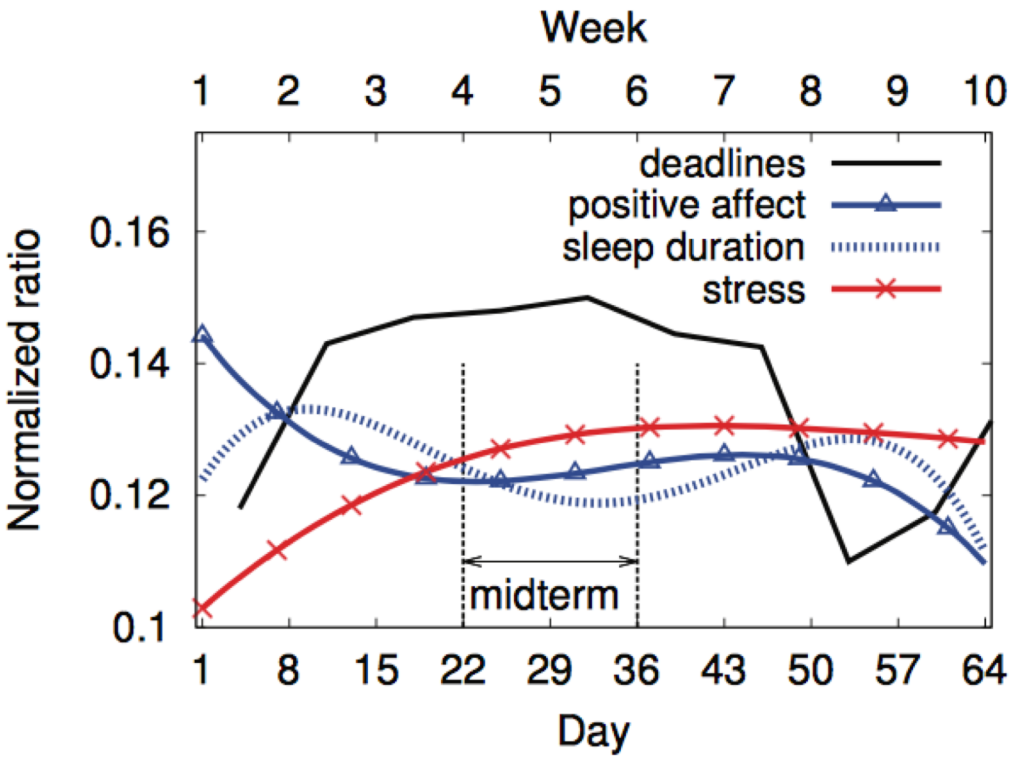


(a) Automatic sensing data quality over the term



(b) EMA data quality over the term

Correlation between self-report & sensor data



EMA & Automatic sensor data (sleep)

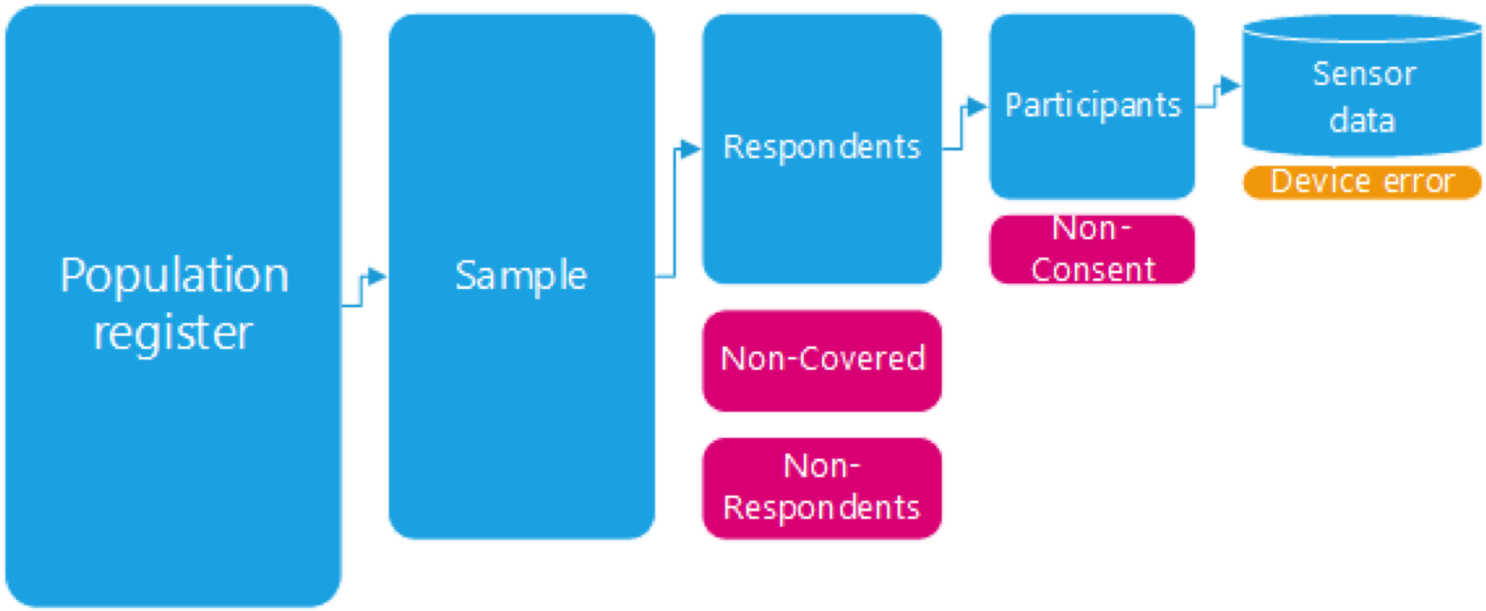
Depression & Automatic sensor data

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

Wang et al. 2014

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



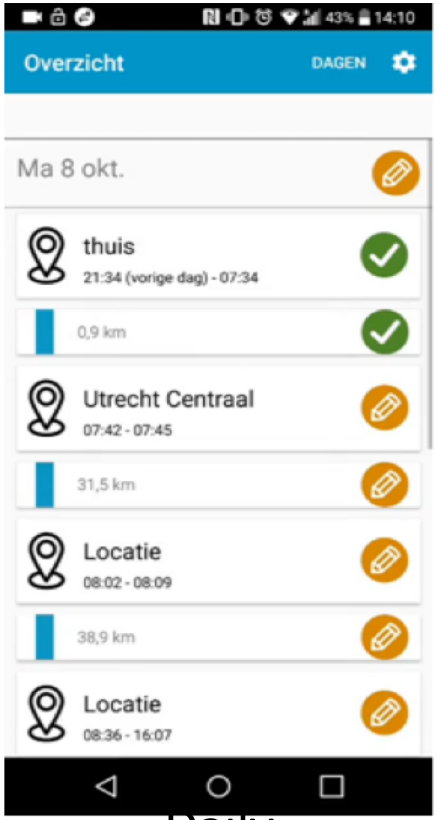
How do people move around in everyday life?

(McCool, Lugtig, and Schouten, under review)

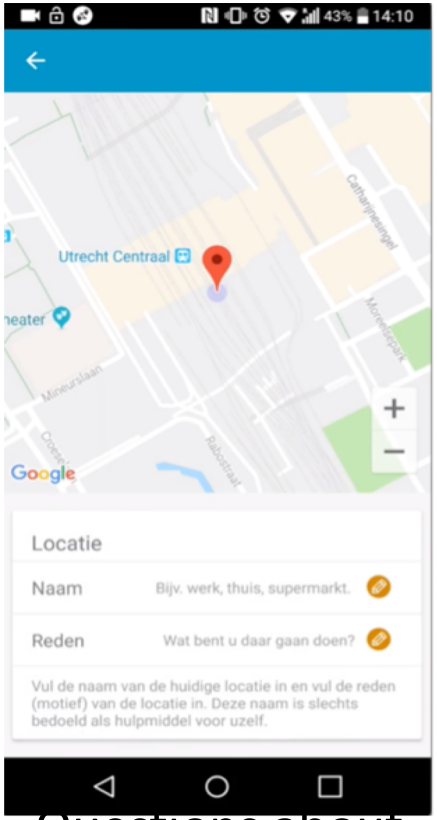
- Everyday mobility field test in the Dutch general population (Nov-Dec 2018)
  - Travel app of Statistics Netherlands (Android & iPhone)
  - Data collection for 7 days
  - N = 1,902
- Sensing location per second (when moving) & per minute (when still):
  - GPS
  - Wi-Fi
- Respondents eagerly provide additional information that helps understand travel behavior (label stops and motives for travel)

How do people move around in everyday life?

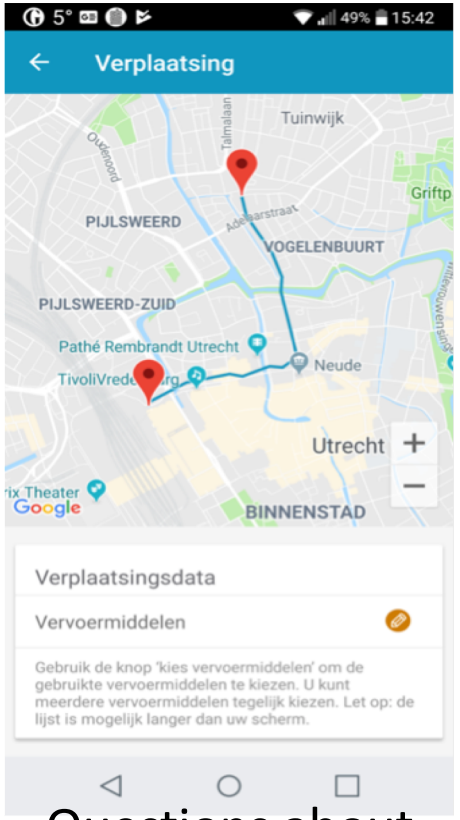
(McCool, Lugtig, and Schouten, in preparation)



Daily overview



Questions about stops

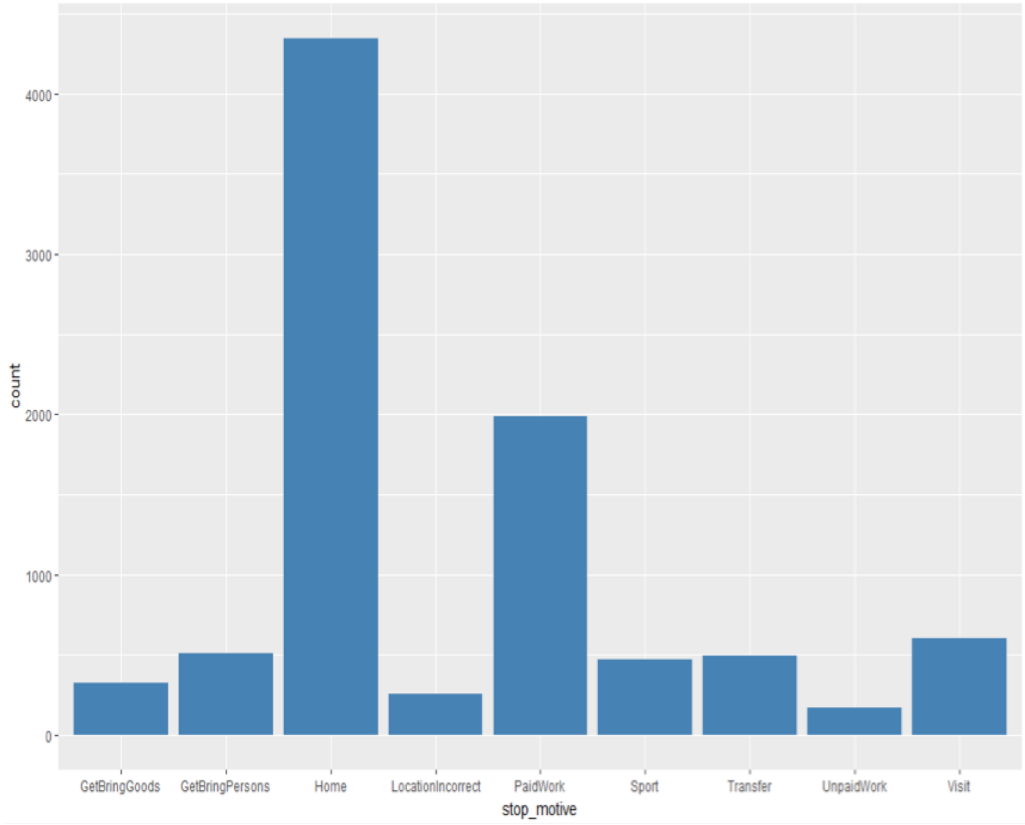


Questions about trips

How do people move around in everyday life?

(McCool, Lugtig, and Schoutenunder review)

- 22,000 stops
  - 13,000 (60%) labeled
  - Overall 50% of respondents give complete/almost complete details



How do people move around in everyday life?

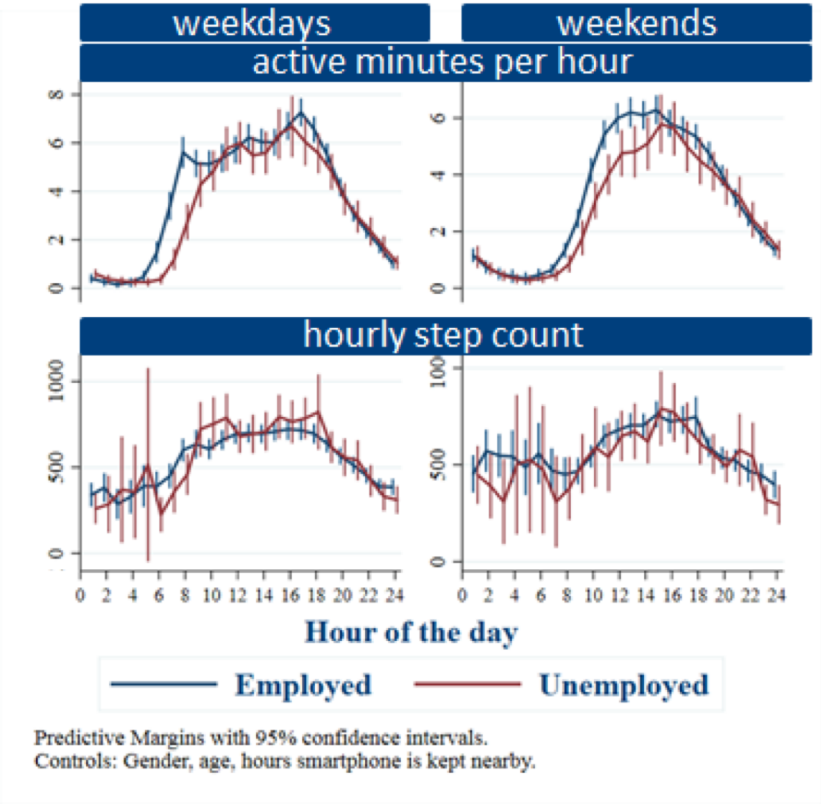
(McCool, Lugtig, and Schouten, under review)

- One week travel

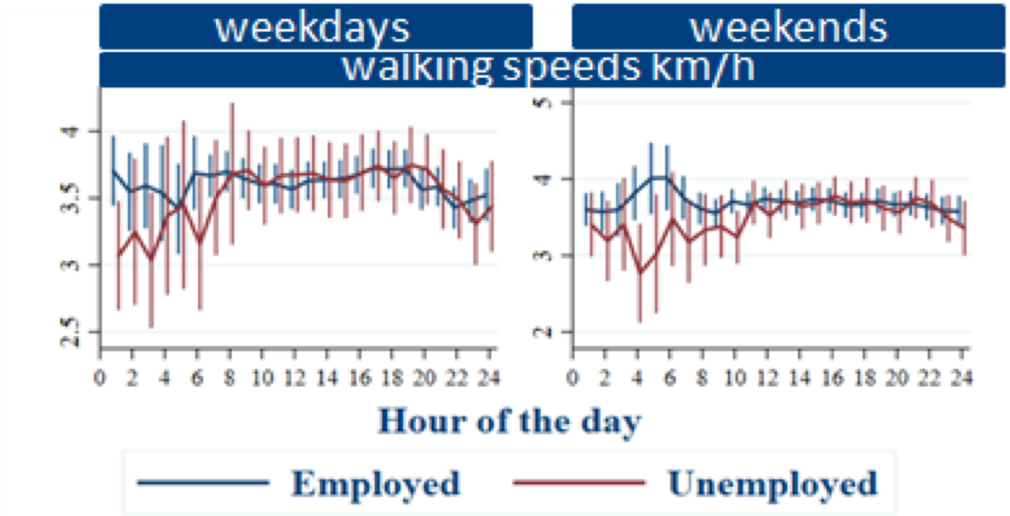
What are the effects of unemployment? (Kreuter et al. 2018)

- 650 Android owners from German panel study "Labour Market and Social Security" (PASS) downloaded *IAB-SMART* app for 6 months
- EMA questions concerning both subjective (e.g., affective impact of daily smartphone use, Big 5 personality) and objective phenomena (e.g., employment and job search activities, use of smartphones in everyday life, memberships in professional and voluntary organizations)
- Five sensing modules:
  - Location using GPS, Wi-Fi, and cellular sensors every 30 minutes
  - Activity and means of transportation (e.g., walking, biking, riding in/on a motorized vehicle) using accelerometer and pedometer
  - Call and texting behavior using phone and SMS logs
  - Use of apps installed on smartphone
  - Social network characteristics from contact lists.

What are the effects of unemployment? (Kreuter et al. 2018)



Predictive Margins with 95% confidence intervals.
Controls: Gender, age, hours smartphone is kept nearby.

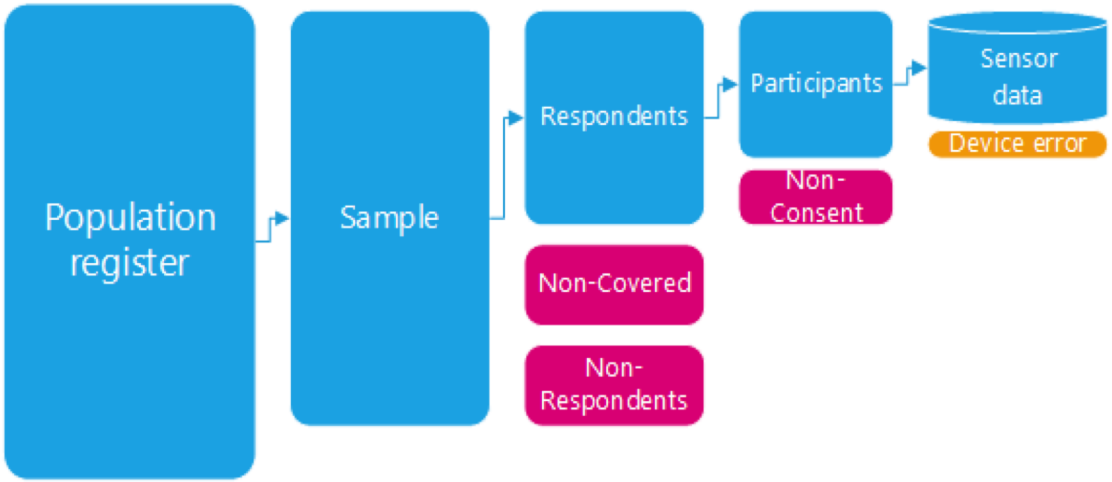


Predictive Margins with 95% confidence intervals.
Controls: Gender, age, hours smartphone is kept nearby.

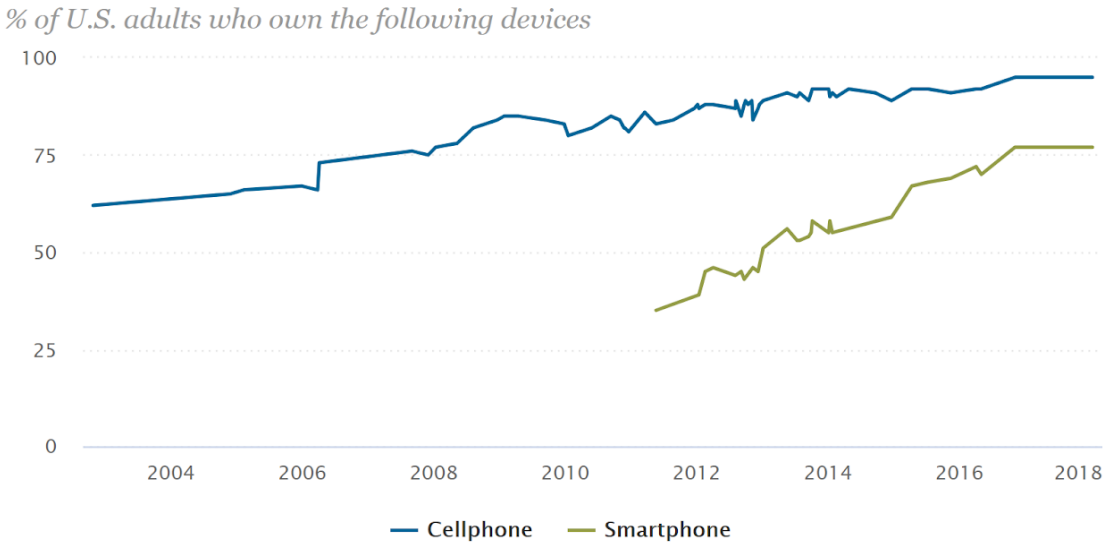
Practical implementation

Practical implementation:

1. Representation: is the population covered?
2. Nonparticipation:
   - Privacy & Consent
3. Measurement:
   - Frequency / sampling rate
   - Errors / missing data / battery life
4. Storage & Costs



1. Representation: Coverage



Picture source:https://www.pewinternet.org/fact-sheet/mobile/

- In Europe, mobile Internet access varies between 31% (Italy) and 84% (Netherlands, Sweden) in 2017 (Eurostat 2018)
- Comparable numbers for Asian-Pacific area (eMarketer Report 2017)
- Levels in Africa substantially lower with much variability across countries (Afrobarometer 2018)

1. Representation: Coverage bias in Germany

(Keusch et al. under review)

- Absolute bias in many substantive measures of PASS for smartphone ownership relatively small (< +/-6 p.p.)
  - Especially once controlling for age (+/-2 p.p.)
- Operating System (OS) Bias:
  - Android smartphone coverage bias generally not much higher than general smartphone coverage bias
  - Large iPhone coverage bias (up to +/-14 p.p.), even when controlling for age (up to +/-12 p.p., especially for measures of satisfaction and deprivation)

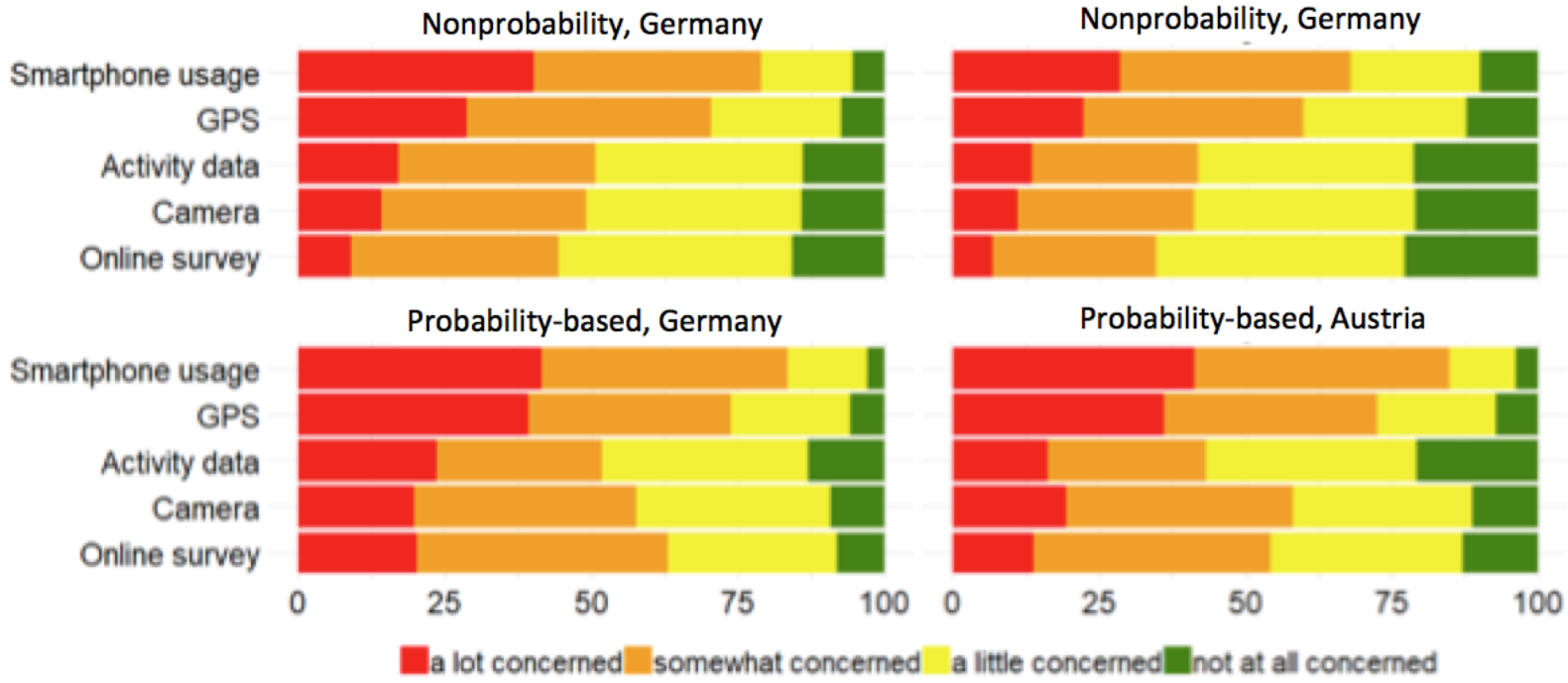
Source: PASS Wave 11; n = 13,703

2. (Non)participation: Willingness to participate (WTP) & actual consent

- Small-scale studies relying on enthusiasts: Willingness = 100%
  - e.g., Wang et al. (2014), Fritz et al. (2017), York Cornwell & Cagney (2017)
- Non-probability online access panels: Willingness between 5% and 56%
  - Varies across countries and by sensor type: 25%-52% for taking pictures, 19-37% for sharing GPS location (Revilla et al. 2016)
- Probability-based panels:
  - LISS Panel: Mobility (GPS, accelerometer) 37% willing, 81% participated; Physical activity (wearables) 57% willing, 90% participated (Scherpenzeel 2017)
  - UK Understanding Society Innovation Panel: Download budget app 17% (Jäckle et al. 2019)
  - German PASS Panel: Download IAB-SMART app 16% (Kreuter et al. 2018)
- Cross-sectional general population studies:
  - CBS Travel App Download & Registration: 35% (McCool et al. 2019)
  - WTP varies from 12% for photo of house to 67% for GPS (Struminskaya et al. 2018)

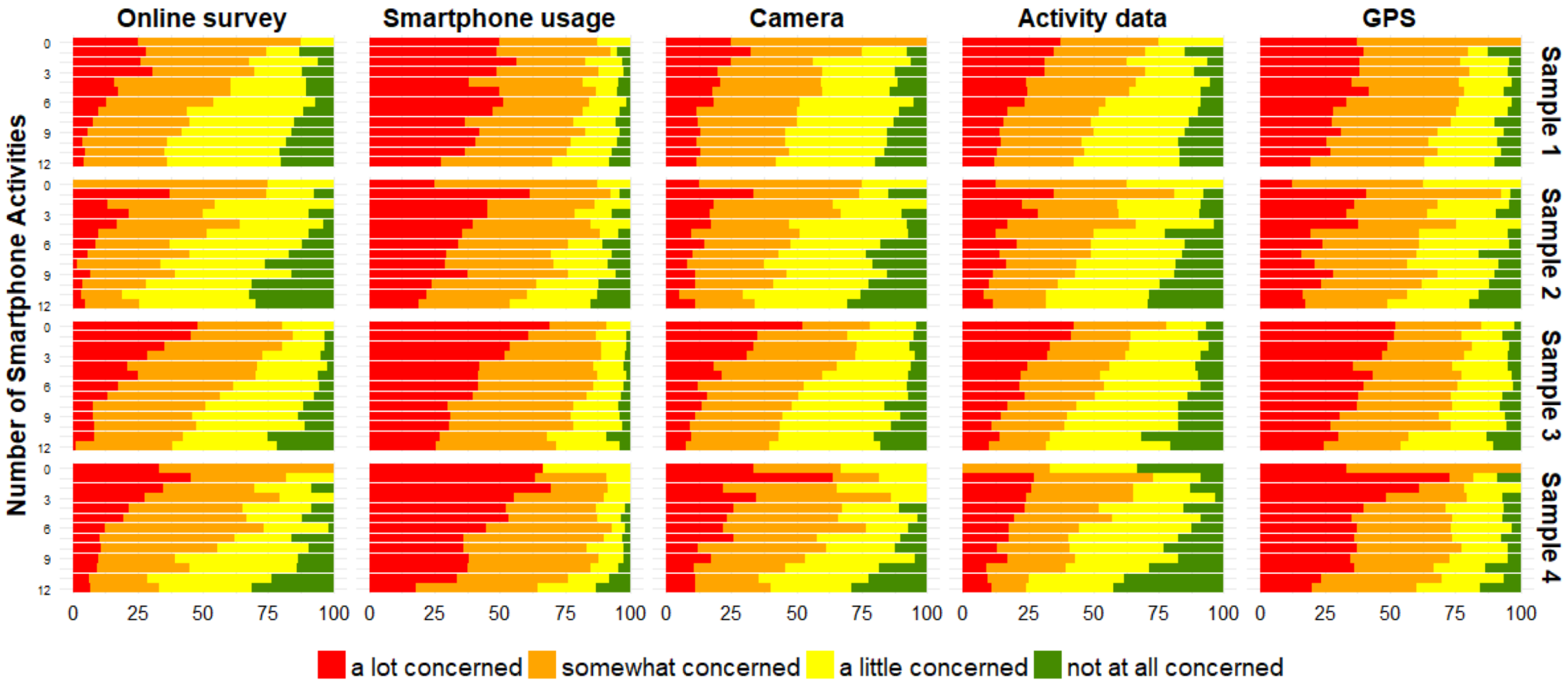
2. Mechanisms of (non-)participation: Concern

- Privacy/security concerns: higher privacy concerns correlate with lower WTP (Keusch et al. 2020; Revilla et al. 2018; Struminskaya et al. 2018; Wenz et al. 2019)



Source: Keusch, Struminskaya, Kreuter, Weichbold (i2020)

Concern by Number of Smartphone Activities



Source: Keusch, Struminskaya, Kreuter, Weichbold (i2020)

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. (2018)

2. Mechanisms of (non-)participation: Incentives

Inconsistent findings

- Keusch et al. (2019):
  - Incentives for app download and staying until end of study increase hypothetical WTP
- Haas et al. (2000): IAB-SMART, Germany
  - 20€ for installation increase installation rate over 10€ (16% vs. 13%)
  - Bonus incentive for consenting to all 5 passive data collection functions has no effect
- McCool et al. (in preparation): CBS Travel App, The Netherlands
  - Bonus incentive for staying until end of field period increase participation: 5€ Prepaid (voucher) +
    - + 5€ Registration + 5€ at the end: 30%
    - + 10€ at the end: 36%
    - + 20€ at the end: 40%
- Jäckle et al. (2019): IP Spending Study, UK
  - 6£ incentive for installation does not increase installation rate over 2£

2. Mechanisms of (non)participation: Privacy & consent

- 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
- Collecting GDPR-conforming consent
- Processing raw data on device
- Collecting data at lowest frequency necessary to answer research question

2. Example for consent: In-browser GPS (Struminskaya et al. 2018)

General consent

Framing, control, and privacy explanation

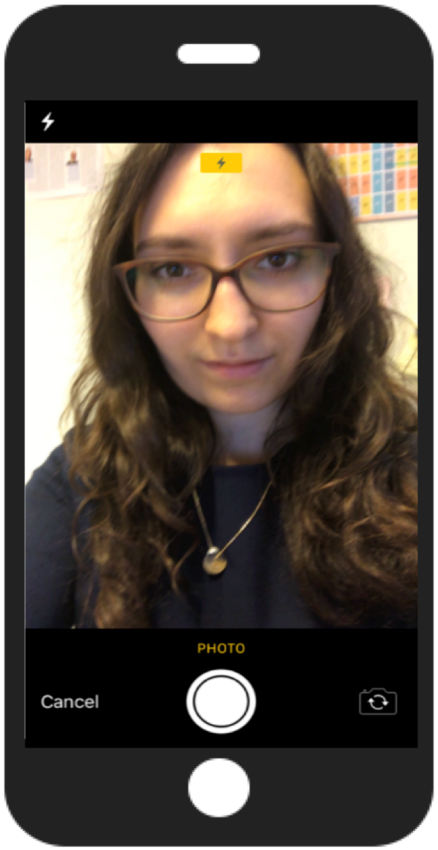
GPS measurement

2. Example for consent: In-browser camera

(Struminskaya et al. 2018)

Video of surroundings

Selfie



Photo/video

2. Example for consent: Tabi App (Lugtig et al. 2019)

- Invitation letter
- Website
- App Download
- Allow tracking

Centraal Bureau voor de Statistiek

<naam>
<adres>
<PC> <plaats>

ons kenmerk
onderwerp CBS-onderzoek
datum

CBS Heerlen
CBS-weg 11
6412 EX Heerlen

<Aanhef>

We zijn met z'n allen veel onderweg. Boodschappen doen met de fiets, wandelen met de hond, met de trein erop uit of met de auto naar het werk. Auto's, fietsen en voetgangers vechten om de beschikbare ruimte. Wat betekent dit voor ons? Kunnen we onze kinderen nog veilig naar school brengen op de fiets? Hebben we meer asfalt nodig? Of juist niet? Om dit soort vragen te beantwoorden voeren het CBS en het ministerie van Infrastructuur en Waterstaat het onderzoek 'Onderweg in Nederland' uit.

Voor dit onderzoek vraagt het CBS een klein aantal personen om met een app korte tijd bij te houden waar ze naar toe gaan. U bent daar één van. U vertegenwoordigt dus veel andere inwoners in Nederland. Voor gemeenten, provincies en voor het Rijk is dit onderzoek de belangrijkste bron van kennis over mobiliteit. Helpt u mee? Zo houden we Nederland samen bereikbaar. Nu en in de toekomst.

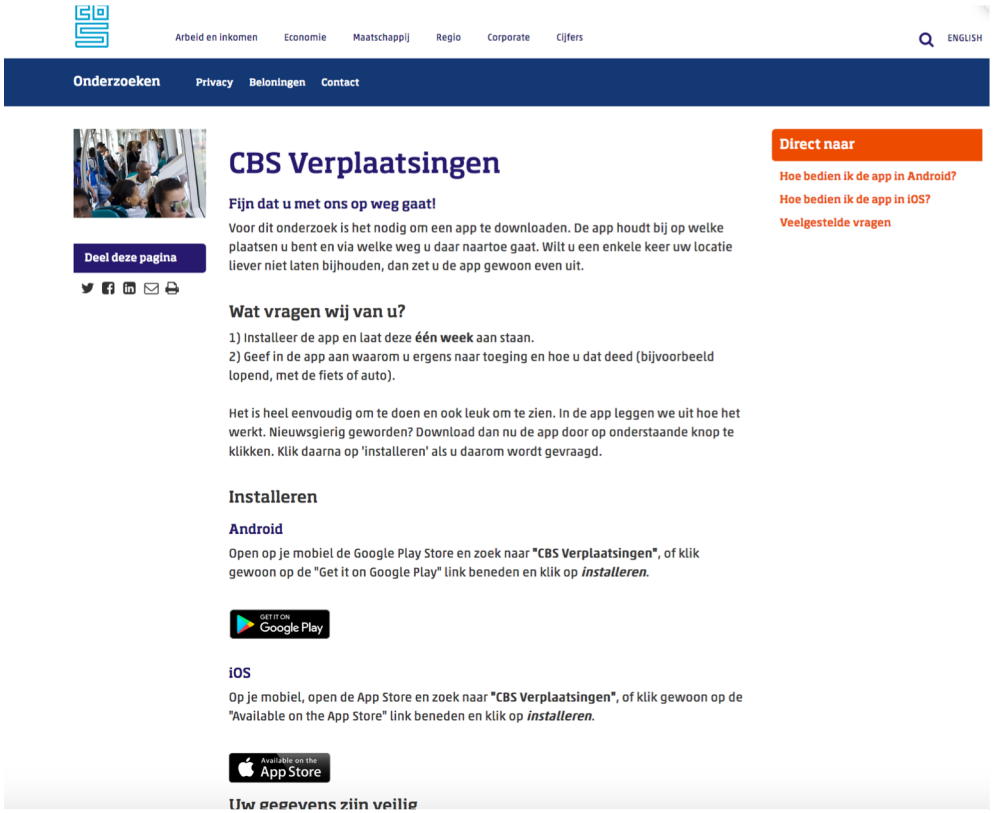
Als dank voor uw hulp krijgt u na afloop van het onderzoek **een cadeaubon van €20.**

Hoe kunt u meedoen?

1. Meedoen kan alleen met een smartphone.
2. Ga met uw smartphone naar de website van het onderzoek: www.tabiapp.eu of gebruik de QR code hiernaast.
3. Op de website kunt u de app downloaden.
4. Na het openen van de app vult u uw gebruikersnaam en wachtwoord in:

Gebruikersnaam: 4035
Wachtwoord: test

5. Het gebruik van de app is heel eenvoudig en wordt in de app zelf uitgelegd.



Arbeid en inkomen Economie Maatschappij Regio Corporate Cijfers ENGLISH

Onderzoeken Privacy Beloningen Contact

Deel deze pagina

CBS Verplaatsingen

Fijn dat u met ons op weg gaat!

Voor dit onderzoek is het nodig om een app te downloaden. De app houdt bij op welke plaatsen u bent en via welke weg u daar naartoe gaat. Wilt u een enkele keer uw locatie liever niet laten bijhouden, dan zet u de app gewoon even uit.

Wat vragen wij van u?

1) Installeer de app en laat deze **één week** aan staan.
2) Geef in de app aan waarom u ergens naar toeging en hoe u dat deed (bijvoorbeeld lopend, met de fiets of auto).

Het is heel eenvoudig om te doen en ook leuk om te zien. In de app leggen we uit hoe het werkt. Nieuwsgierig geworden? Download dan nu de app door op onderstaande knop te klikken. Klik daarna op 'installeren' als u daarom wordt gevraagd.

Installeren

Android

Open op je mobiel de Google Play Store en zoek naar **"CBS Verplaatsingen"**, of klik gewoon op de "Get it on Google Play" link beneden en klik op ***installeren***.

GET IT ON Google Play

iOS

Op je mobiel, open de App Store en zoek naar **"CBS Verplaatsingen"**, of klik gewoon op de "Available on the App Store" link beneden en klik op ***installeren***.

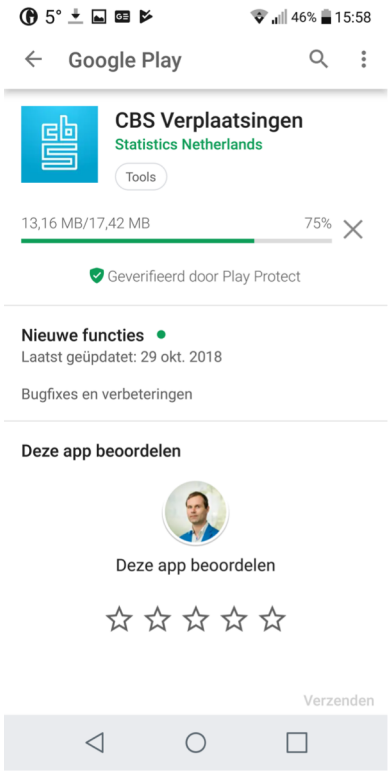
Available on the App Store

Direct naar
Hoe bedien ik de app in Android?
Hoe bedien ik de app in iOS?
Veelgestelde vragen

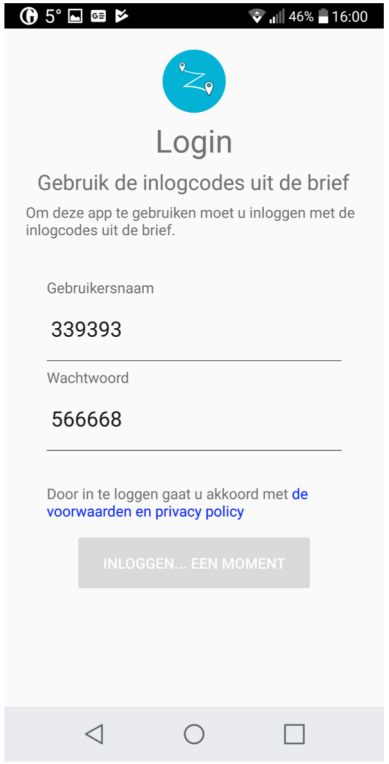
Uw gegevens zijn veilig

2. Example for consent: Tabi App (Lugtig et al. 2019)

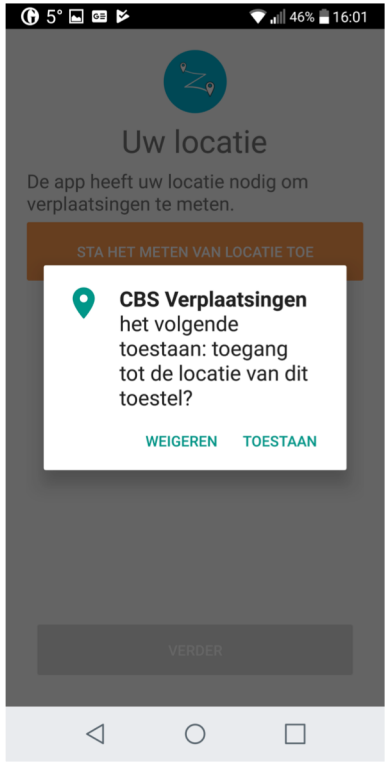
- Invitation letter
- Website
- App Download
- Allow tracking



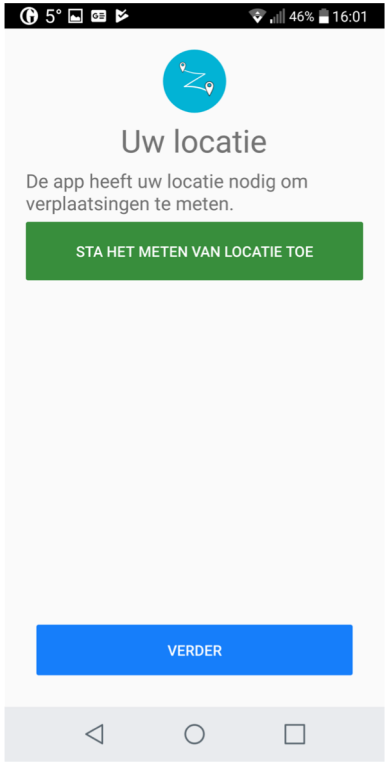
Google play store



Log in with credentials

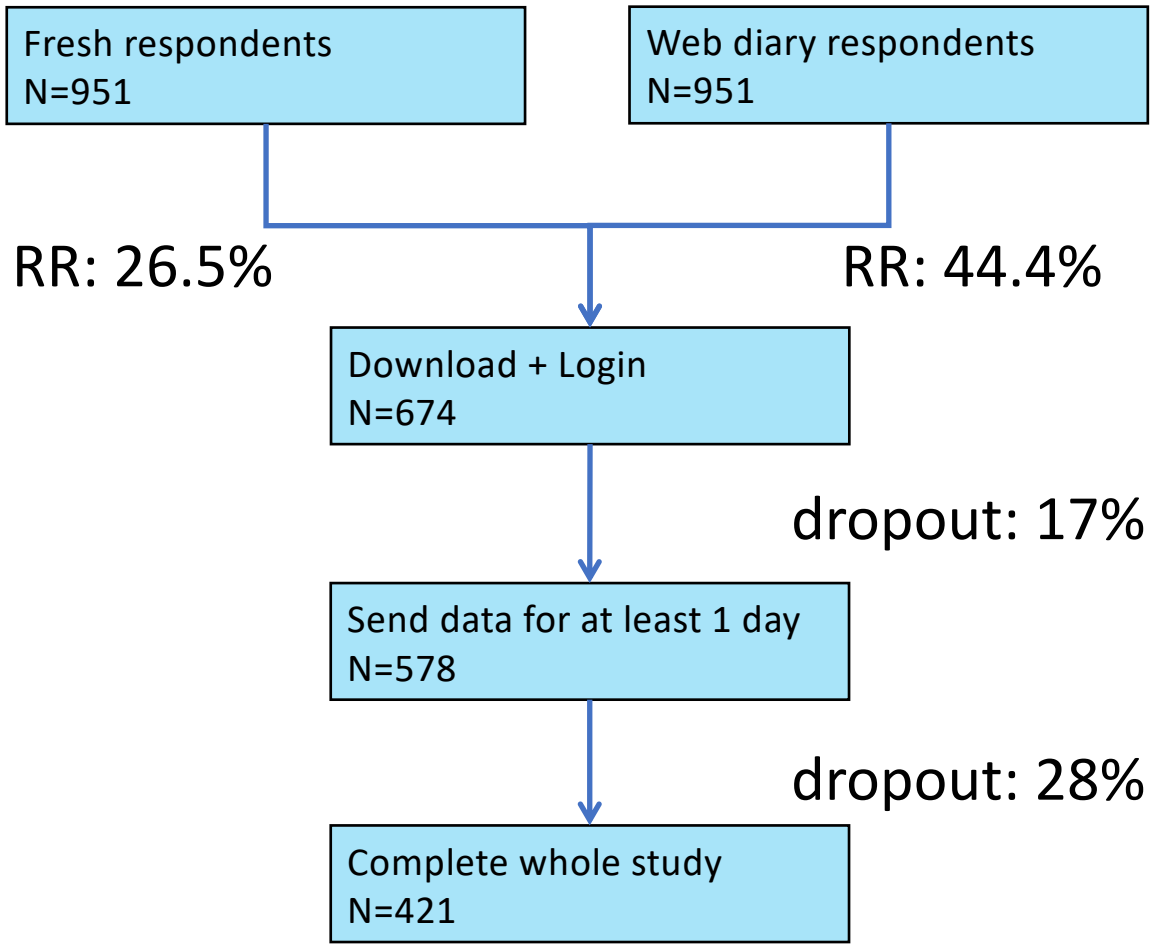


Allow GPS tracking

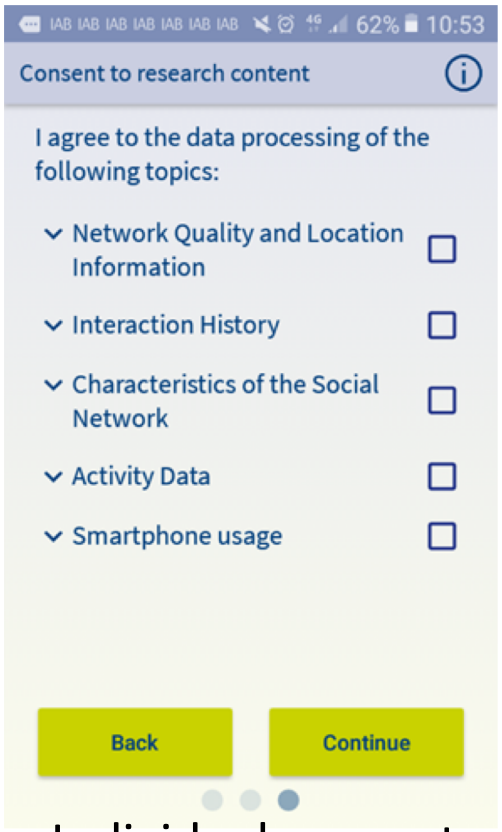


Ready to use

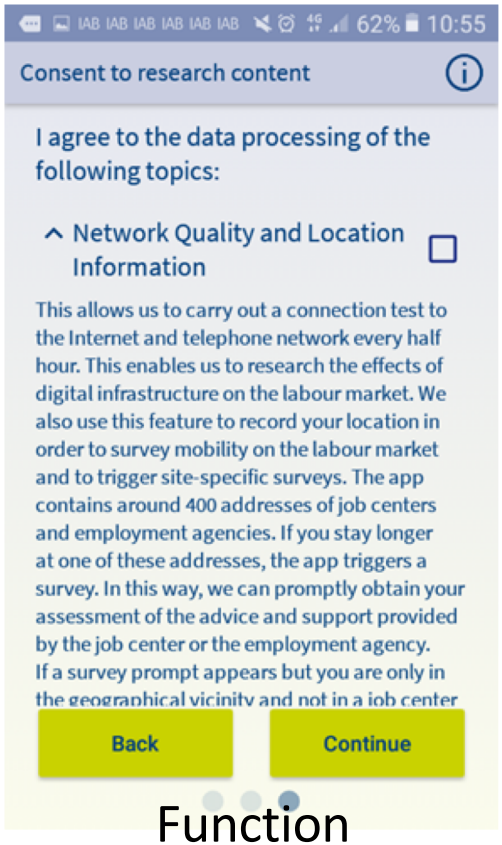
2. Example for consent: Tabi App (Lugtig et al. 2019)



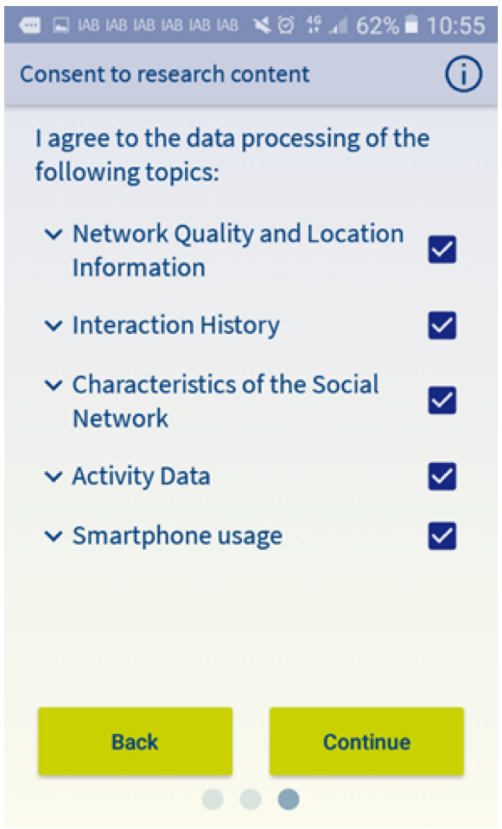
2. Example for consent: IAB-SMART (Kreuter et al. 2018)



Individual consent screen



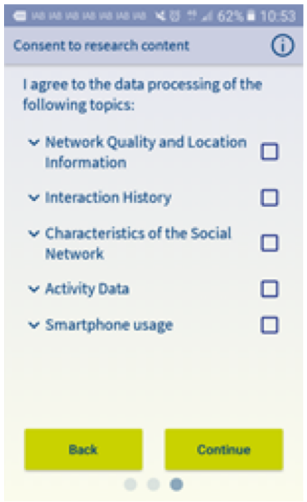
Function explanation

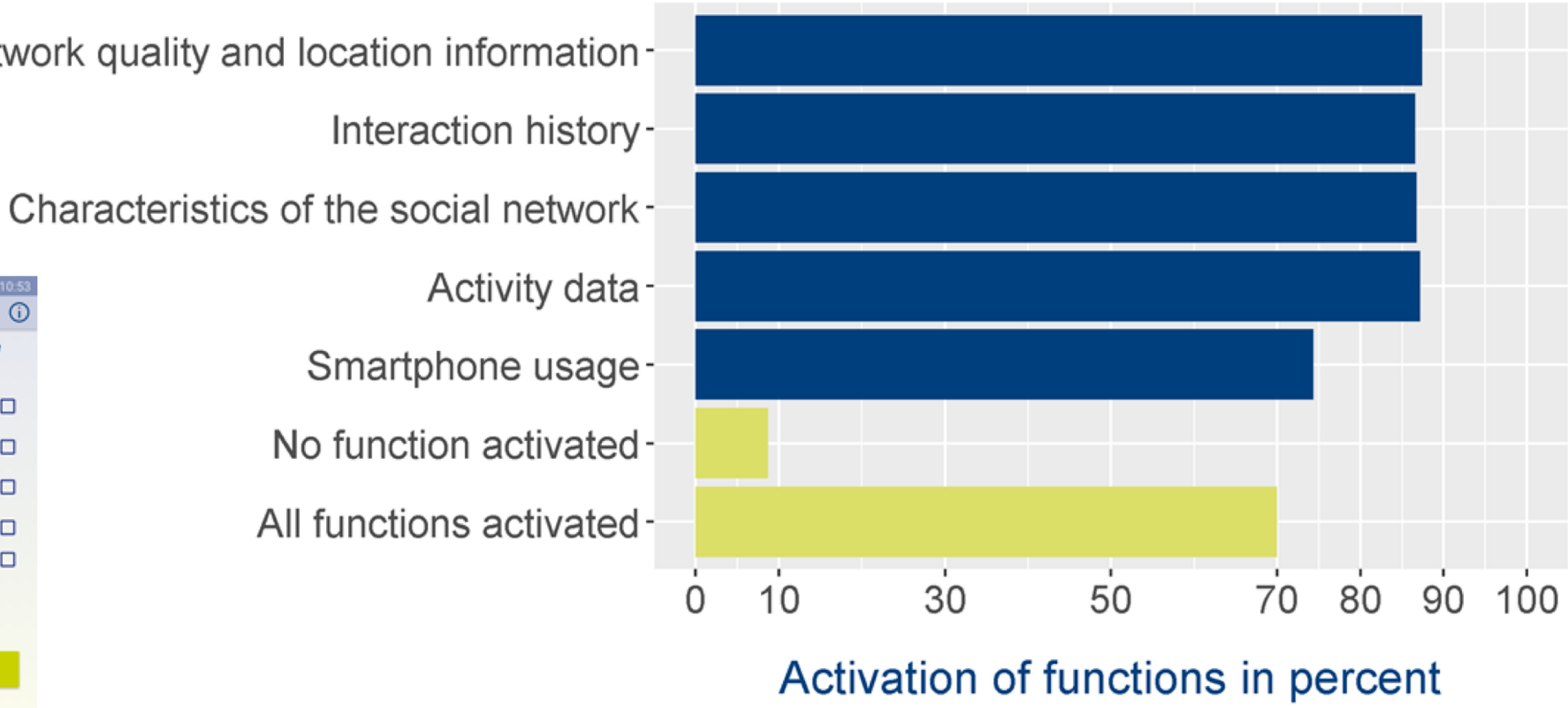


Full consent

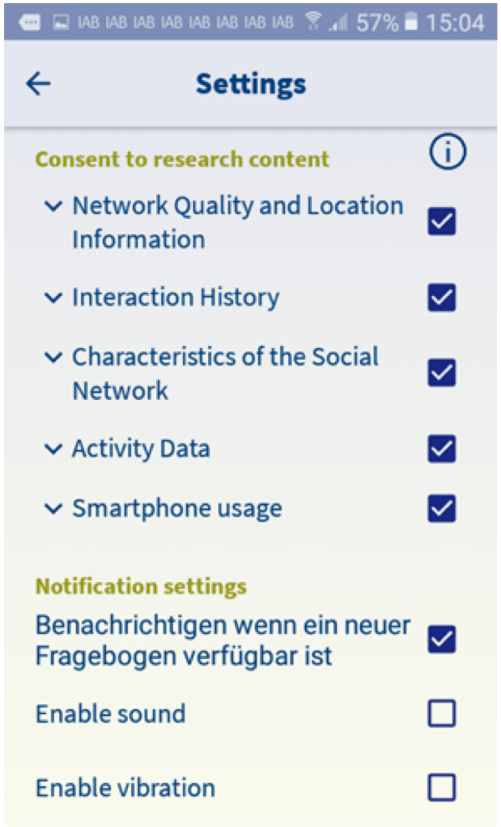
App home screen

2. Example for consent: IAB-SMART (Kreuter et al. 2018)

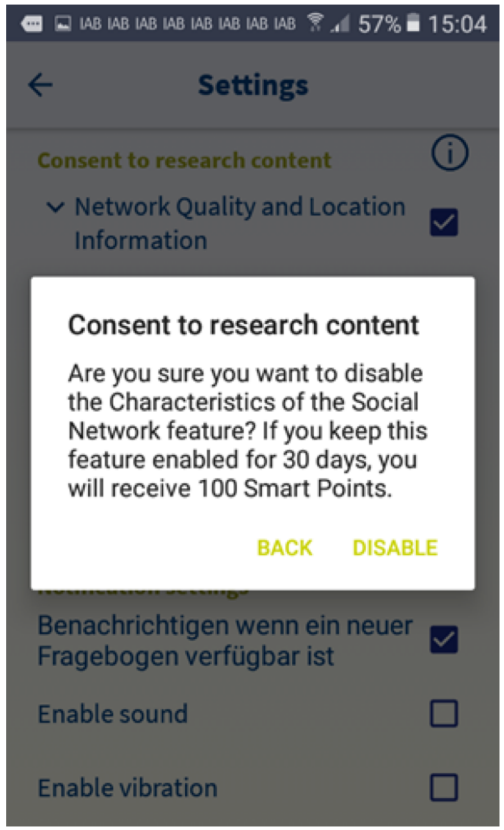




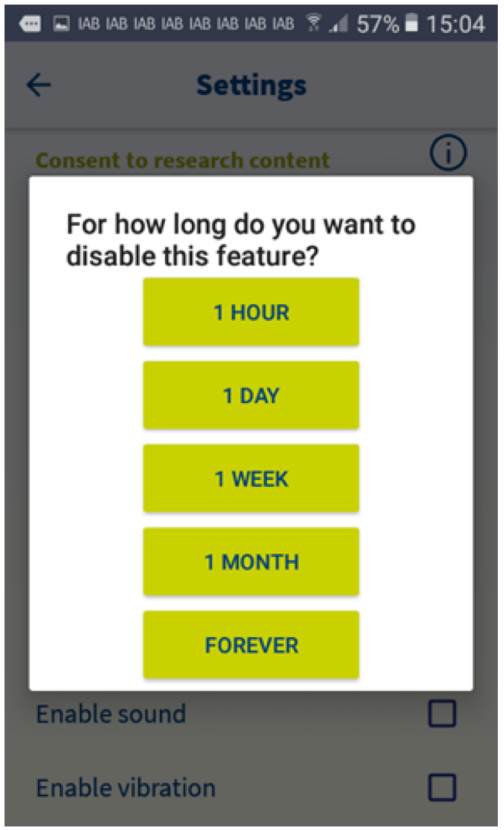
2. Example for consent: IAB-SMART (Kreuter et al. 2018)



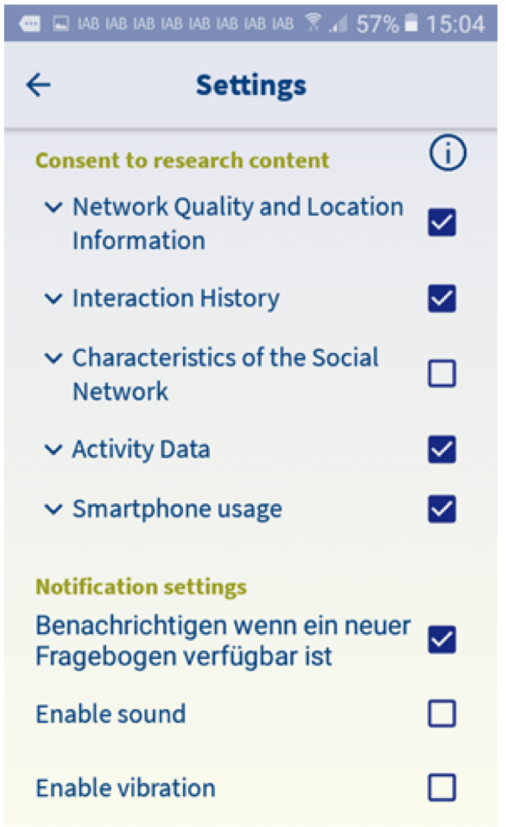
App settings



Withdrawing consent



Withdrawing consent



App settings

3. Frequency of measurement - Sampling rate

- Situational
- Discrete
- Continuous
- Combination

3. Frequency of measurement - Sampling rate

- Situational
  - Measurement only at specific times
  - e.g., GPS location whenever EMA is answered (MacKerron & Mourato 2013)
- 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 (York Cornwell & Cagney 2017) from 9 am to 9 pm, every 30 minutes (Kreuter et al. 2018)

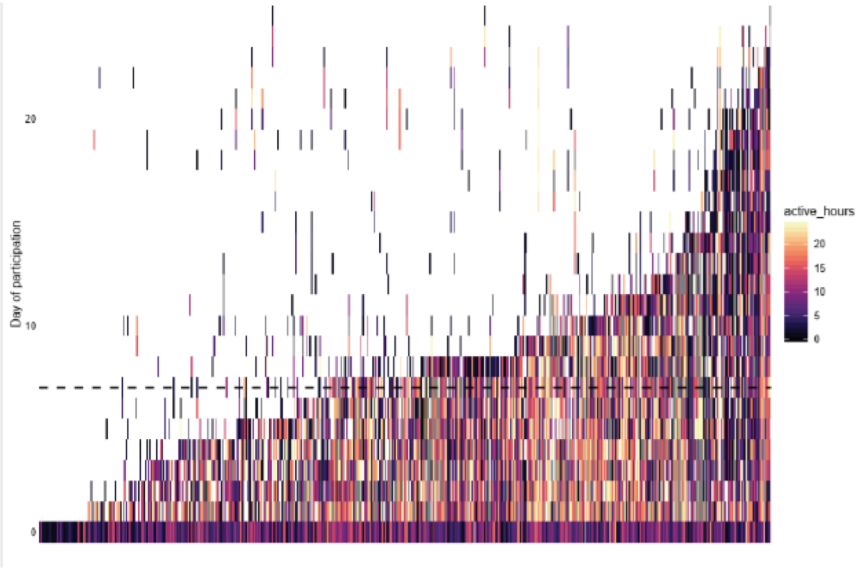
Struminskaya & Keusch 2019

3. Frequency of measurement - Sampling rate

- Continuous
  - Tracking of smartphone-mediated behavior usually always on (e.g., Kreuter et al. 2018; Sugie 2018)
  - GPS collected at high frequency allows measurement of exact route
    - e.g., every sec when moving, every min when still (McCool et al.)
  - Microphone always on (e.g., Wang et al. 2014)
    - But data processed on device to save storage and preserve privacy
- Combination
  - e.g., measure activity (accelerometer) for 15 min twice a day (Lathia et al. 2017)

3. Errors during data collection & Missing data

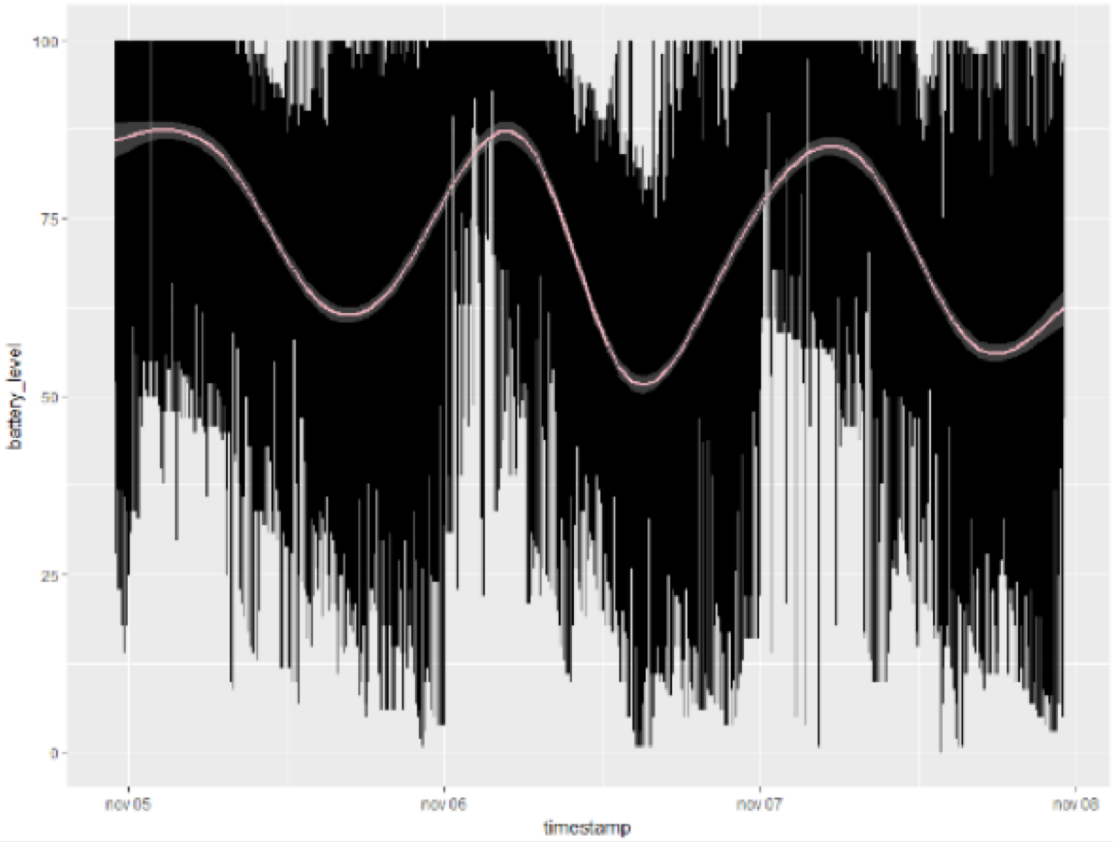
- Sensor-based errors
- Missing data
  - Technical issues: e.g.,
    - Phone out of power or sleep mode
    - iOS blocks collection of location in background
  - Noncompliance: e.g.,
    - Missing permissions
    - Mobile data disabled
    - Leaving phone at home
    - Turning app off at certain locations
- Erroneous data
- Providing feedback & measurement reactivity



Active hours per user (McCool et al. 2019)

3. Battery life example: CBS Travel App (McCool et al.)

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



4. Storage & Costs

- One person’s GPS coordinates collected every five minutes for 7 days would result in 10,080 data points
- Think whether you really need this amount of data!
- Most systems store data first on device and transmit them to server once connected to Wi-Fi
- Processing data on device and only transmitting aggregated data saves storage space
- Besides storage, costs can include:
  - App/In-browser measurement development
  - Incentives
  - Data handling
  - Helpdesk

Group discussion / brainstorm

HEALTH • COVID-19

Contact Tracing Apps Were Big Tech's Best Idea for Fighting COVID-19. Why Haven't They Helped?

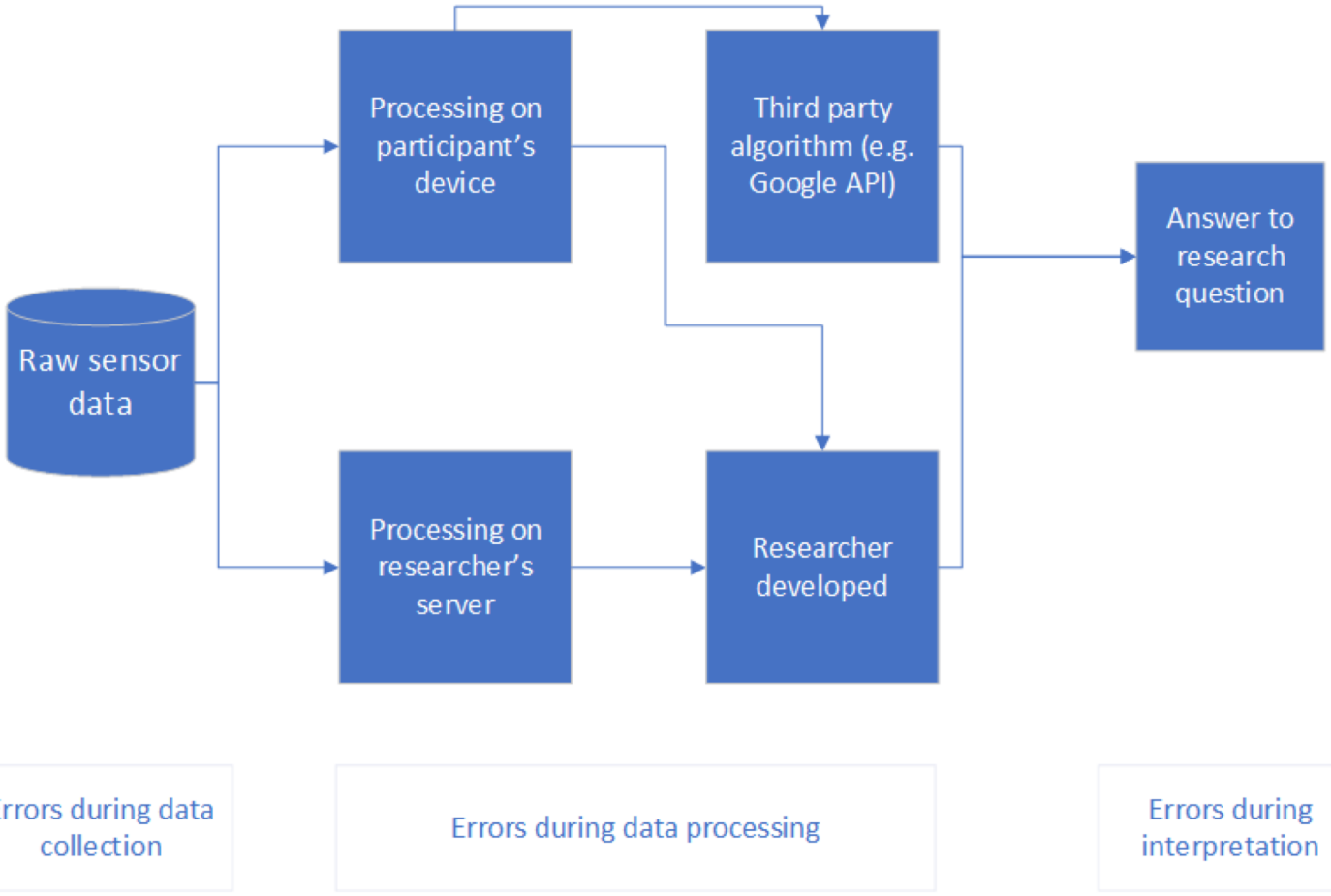
https://time.com/5905772/covid-19-contact-tracing-apps/

Group discussion / brainstorm

- Designed big data application to COVID contact tracing apps
- What might be problematic with contact tracing apps?
- What are the (ideal) conditions when they would work?
- ….
- ….
- ….

When the data is collected…

From raw data to insights



Raw data

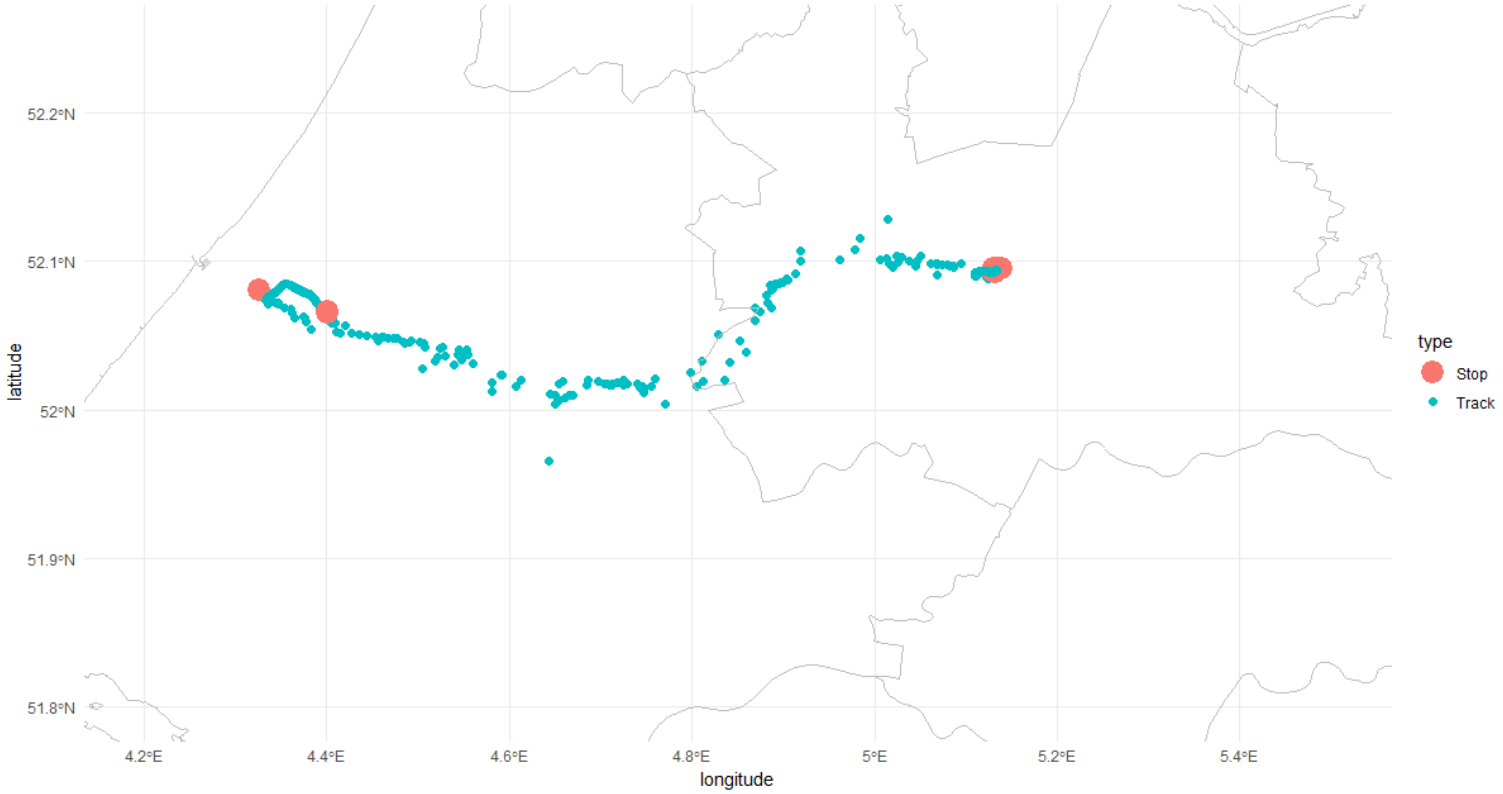
	device_id	latitude	longitude	accuracy	speed	altitude	timestamp
1:	23	52.09460	5.134593	15.204	0	54.7	2018-10-31 22:53:22
2:	23	52.09460	5.134593	15.204	0	54.7	2018-10-31 22:54:22
3:	23	52.09460	5.134593	15.204	0	54.7	2018-10-31 22:56:40
4:	23	52.09460	5.134593	15.204	0	54.7	2018-10-31 22:57:40
5:	23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

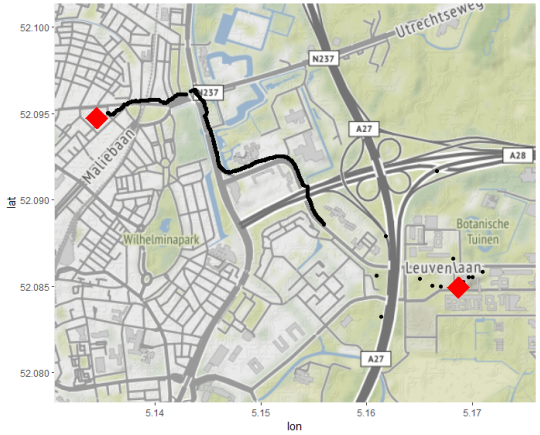
38524:	23	52.09464	5.134572	15.175	0	54.7	2018-11-12 00:00:33
38525:	23	52.09464	5.134572	15.175	0	54.7	2018-11-12 00:00:33

	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

Making inference from raw data





Source: McCool et al. (2019)

RQ: How active people are? (Mulder et al. 2019)

- GENEActive accelerometers
  - Wrist-worn
  - 8 days
  - Sampling rate 60 Hz
- Dutch LISS Panel respondents (N invited=1000, N collected=850)
- Day Reconstruction Method
- Machine Learning to detect physical activity patterns
  - Sedentary: sleeping, sitting, commuting
  - Energetic: walking, jogging, cycling, stair climbing, tooth brushir

Model building pipeline

Data Cleaning & pre-processing

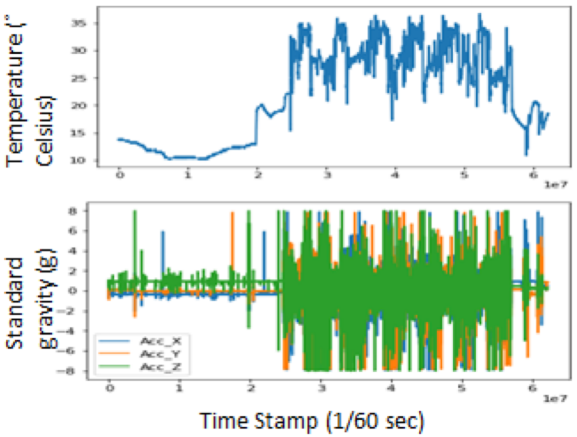
- Removal non-wear time
- Removal of high frequency (frequency higher than 15 Hz)
- Data with wear time less than 7 days discarded

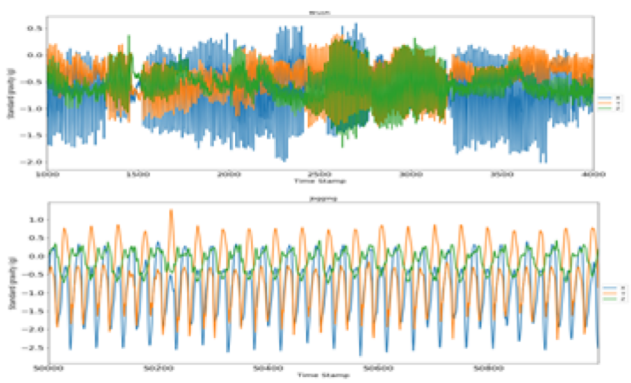
Feature Engineering

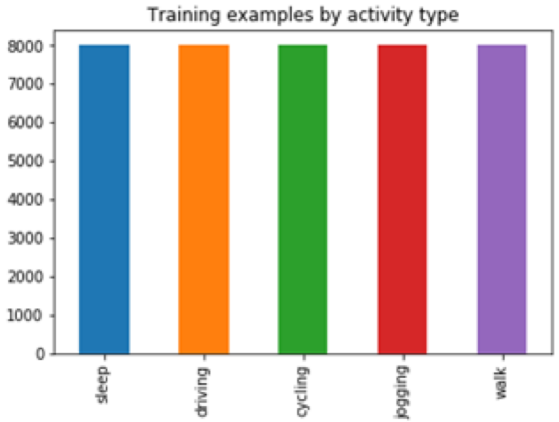
- **Time domain:** X, Y, Z, temperature, mean, median, standard deviation, RMS, percentile distribution
- **Frequency domain:** FT, dominant frequency selection, power of signal

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)

Class exercise: Design a sensor & survey study

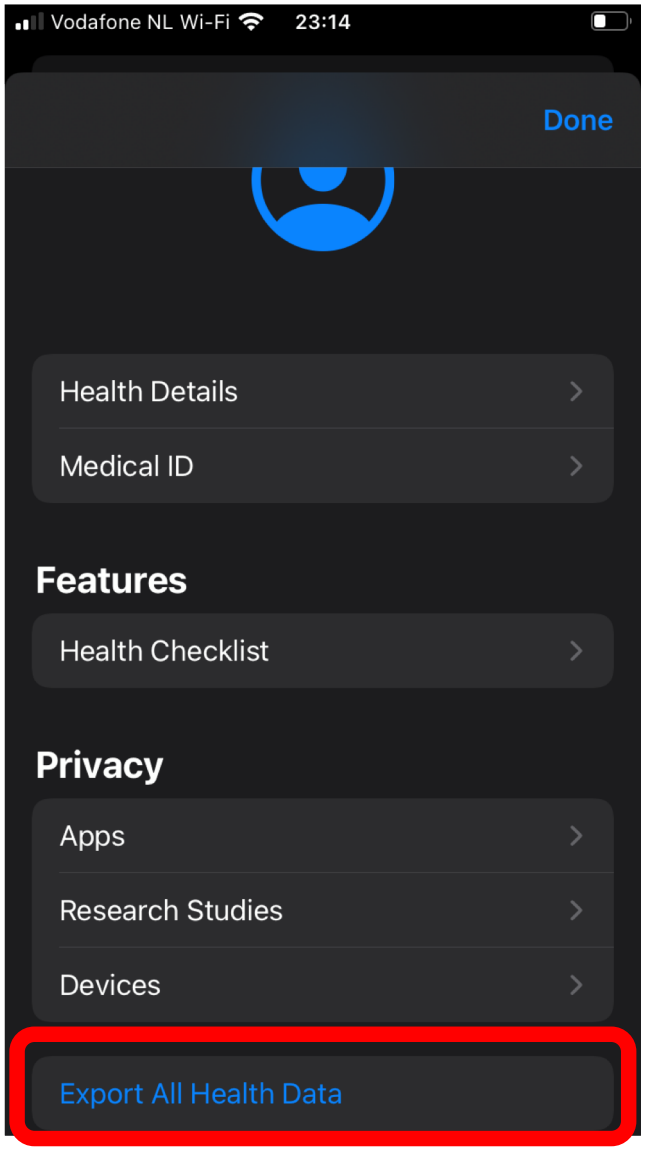
In groups: choose a RQ and suggest study design
20 min (design) + 5 (present)

- RQ1: How do local environmental conditions affect the health of elderly residents in a large megapolis?
- RQ2: How social ties develop in college students over time?
- RQ3: How do people living in households spend money?
- RQ4: How do behaviors of employed and unemployed persons differ?

- What kind of sensor data should be collected and how would you collect these (be specific)
- What kind of survey data should be collected and how would you collect these (be specific)
- What kind of auxiliary data could be used?
- Implementation (sampling rate, loan devices, ESM, etc.)
- Added value (can you answer these questions with survey data?)
- Discuss the design from the TSE perspective (focus on measurement, representation, or both)

Take home exercise

- If you have an iPhone or an Apple watch (or your friend with an iPhone graciously shares data with you or you google how to replicate it for android)
- Download your apple health data, prepare it for analysis and find out something about yourself (see next slide)



Take home exercise

- Go to “Health” -> click on profile -> Download All Health Data
- Follow the steps from here to prepare for analysis: https://taraskaduk.com/posts/2019-03-23-apple-health/
- You will need these packages and use only if you want to change time zones, otherwise comment out #endDate = ymd_hms(endDate,tz="America/New_York"),

```
library(XML)
library(tidyverse)
library(lubridate)
library(scales)
library(ggthemes)
library(dplyr)
library(magrittr)
library(stringr)
library(ggplot2)
```

- Find out about something about yourself! e.g., your average step count, how many steps you go by day of the week and when, or something more interesting
- Send me your code and your result (e.g., a plot) and I will tell Peter ☺

Take home exercise

- It is optional! So no stress…

***German “entspannen” = relax**

Questions? Comments?

Now or contact me at b.struminskaya@uu.nl

References & Additional reading

Baker, Reginald P. 2017. Big Data: A Survey Research Perspective. In „Total Survey Error in Practice“, ed. by P. P. Biemer, E. De Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. Lyberg, C. Tucker, and B. West, Hoboken, NJ: Wiley

Beyer, M. A., and D. Laney. 2012. The Importance of “Big Data”: A Definition. G00235055. Stamford, CT: Gartner.

Callegaro, M., and Yang, Y. 2017. The role of surveys in the era of „big data“. In „The Palgrave Handbook of Survey Research”, ed. by D.L. Vannette and J.A. Krosnick, Palgrave.
doi: 10.1007/978- 3-319-54395-6_23

Groves, R. M. 2011. Three eras of survey research. Public Opinion Quarterly 75(5), 861-871. doi:10.1093/poq/nfr057

Salganik, M. 2018. Bit by bit. Social research in the digital age. Princeton University Press.

Struminskaya, Lugtig, Keusch, Höhne (2020). Using mobile apps and sensors in surveys. Special issue of the Social Science Computer Review (contributions on SSCR website by English et al., Bähr et al., Sepulvado et al., Wenz et al.)