Utrecht University

# Summer Course Survey Research: Advanced Survey Design

Bella Struminskaya & Peter Lugtig

Department of Methodology & Statistics, Utrecht University


Slides by Toepoel, Struminskaya, Lugtig

Utrecht University

# Big data and TSE

Bella Struminskaya & Peter Lugtig

Department of Methodology & Statistics, Utrecht University

# 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

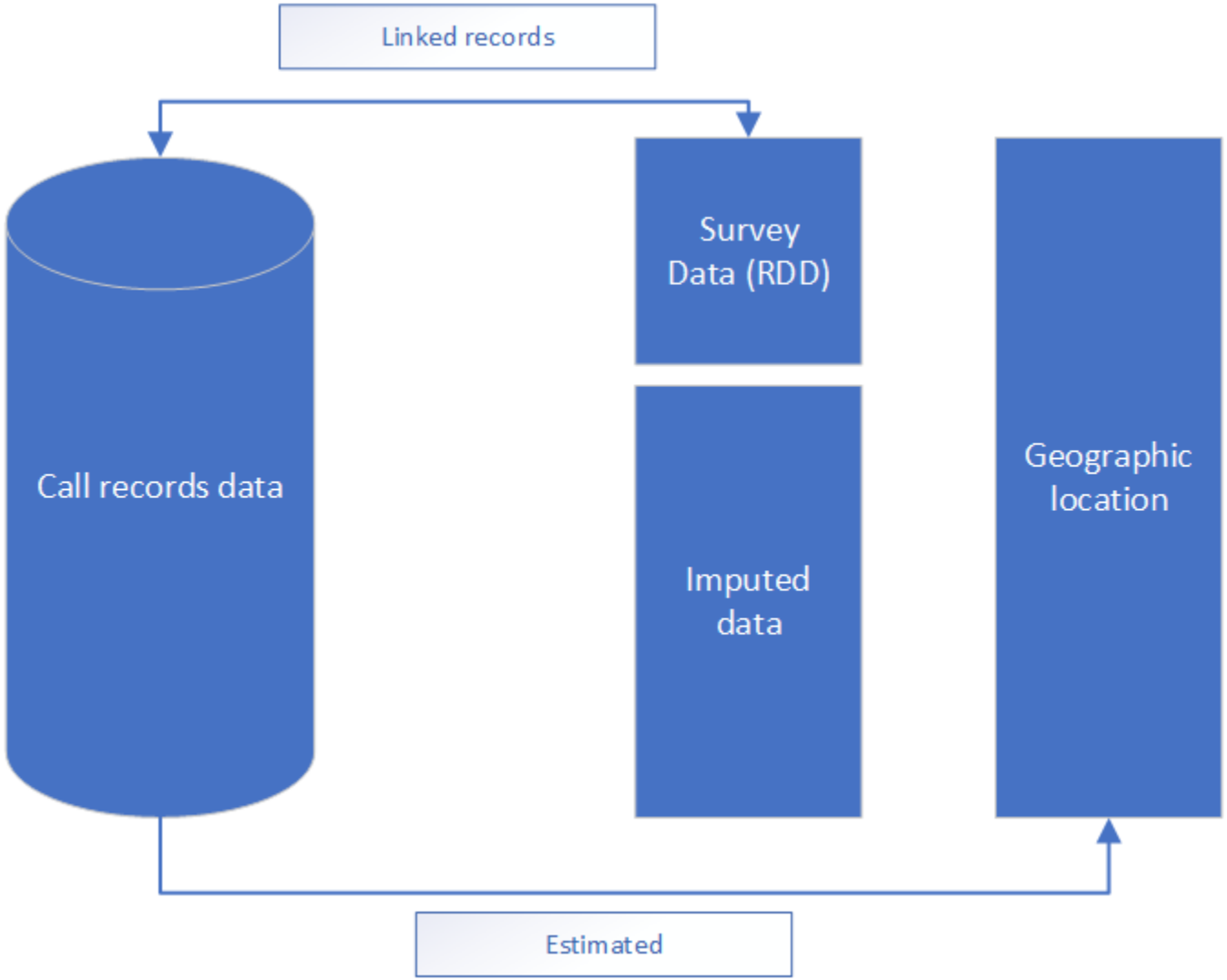
**Table 1. Summary statistics for primary data sets.** Phone survey data were collected by the authors in Kigali, in collaboration with the Kigali Institute of Science and Technology. Call detail records were collected by the primary mobile phone operator in Rwanda at the time of the phone survey. Demographic 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)

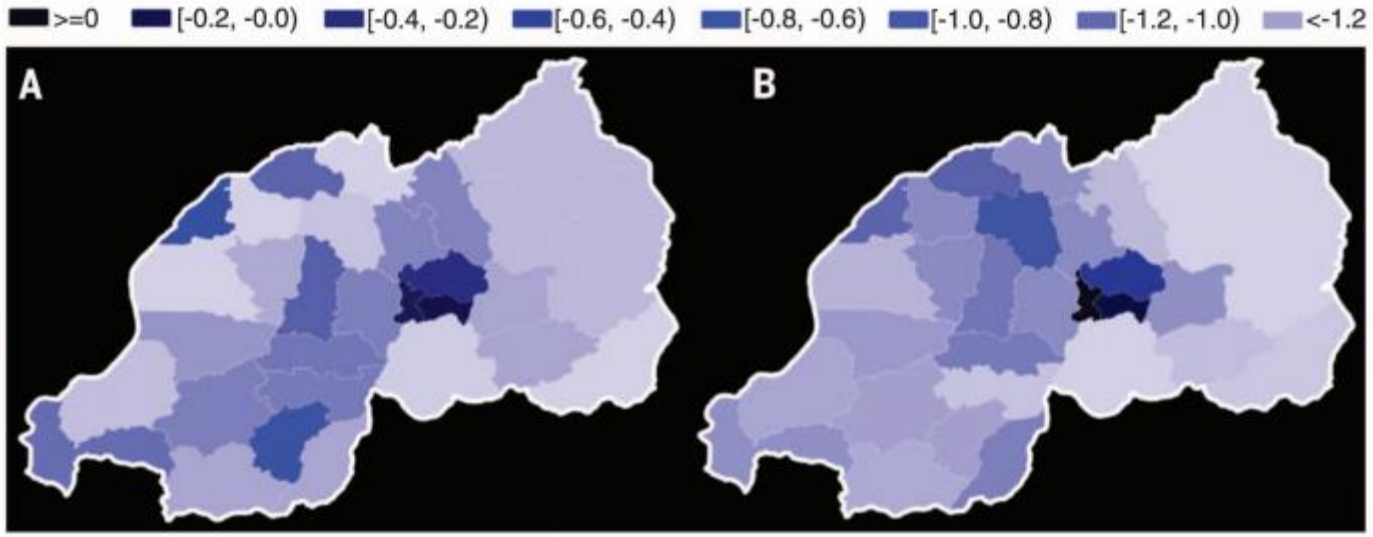
# Can anonymized data from mobile phone networks predict poverty and wealth?




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

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

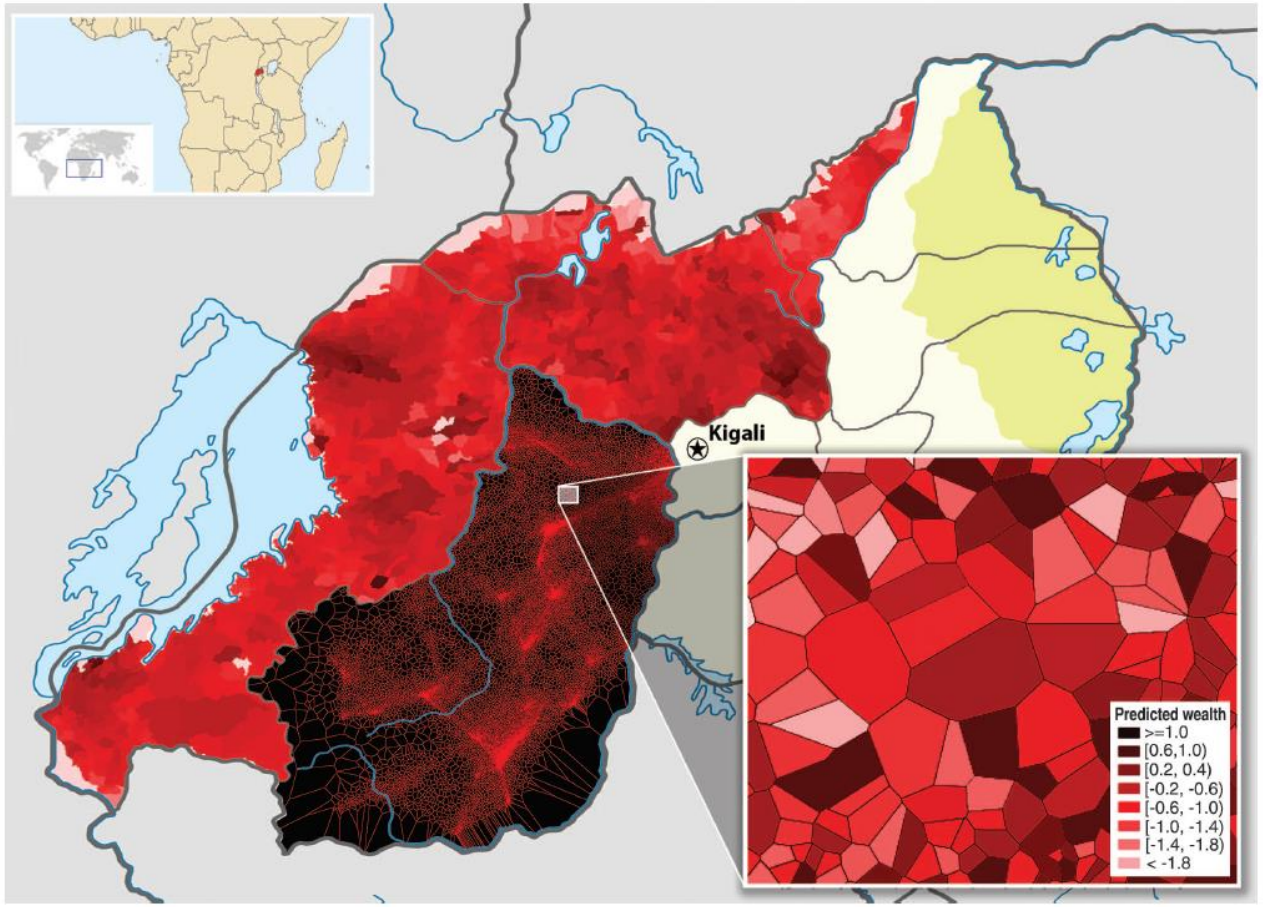
- Anonymized call records (1.5 mil)
- Telephone survey (n=856)
- ‘Gold standard’ f2f Demographic and Health Survey (n=12792)



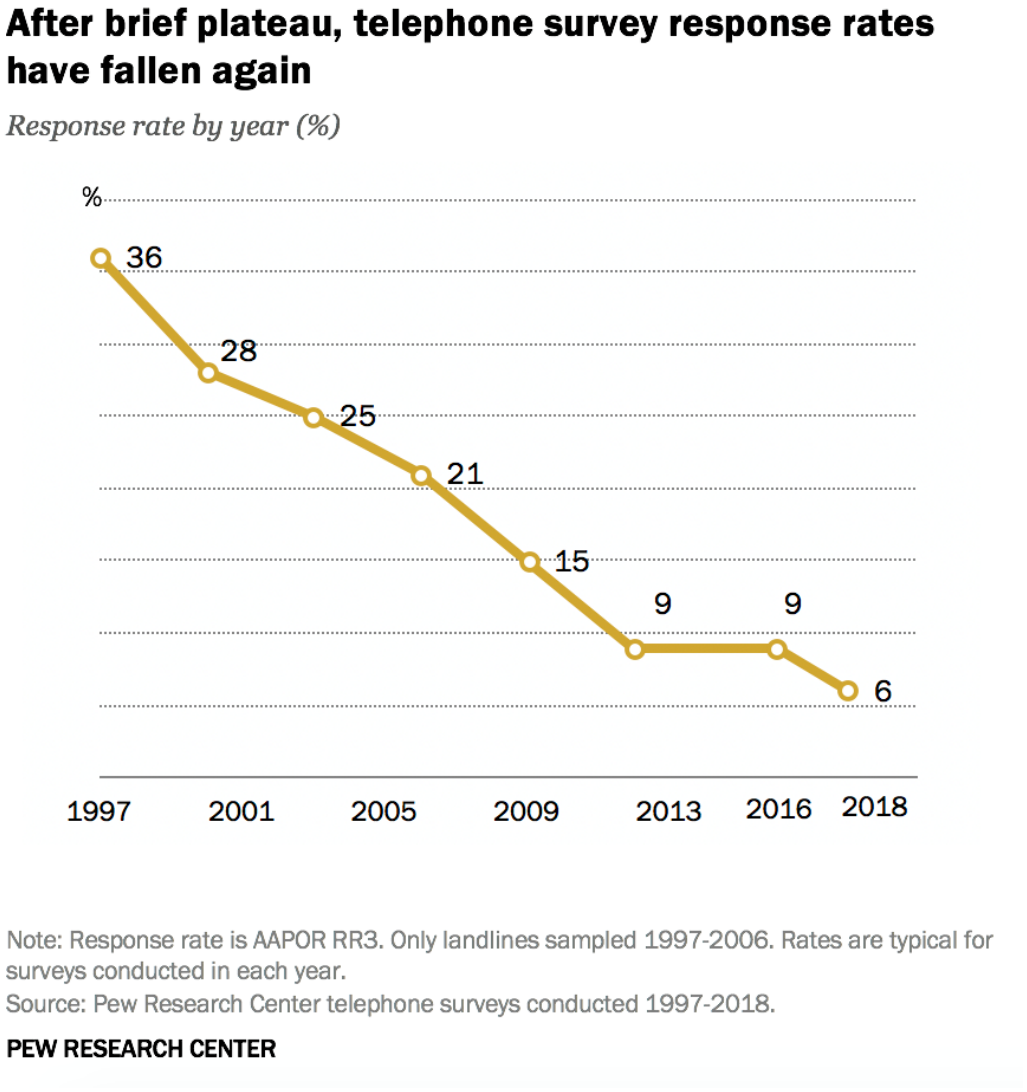

Composite wealth index: A – predicted from call data, B – actual from DHS, *r*=0.79 (Blumenstock et al. 2015)

# 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)



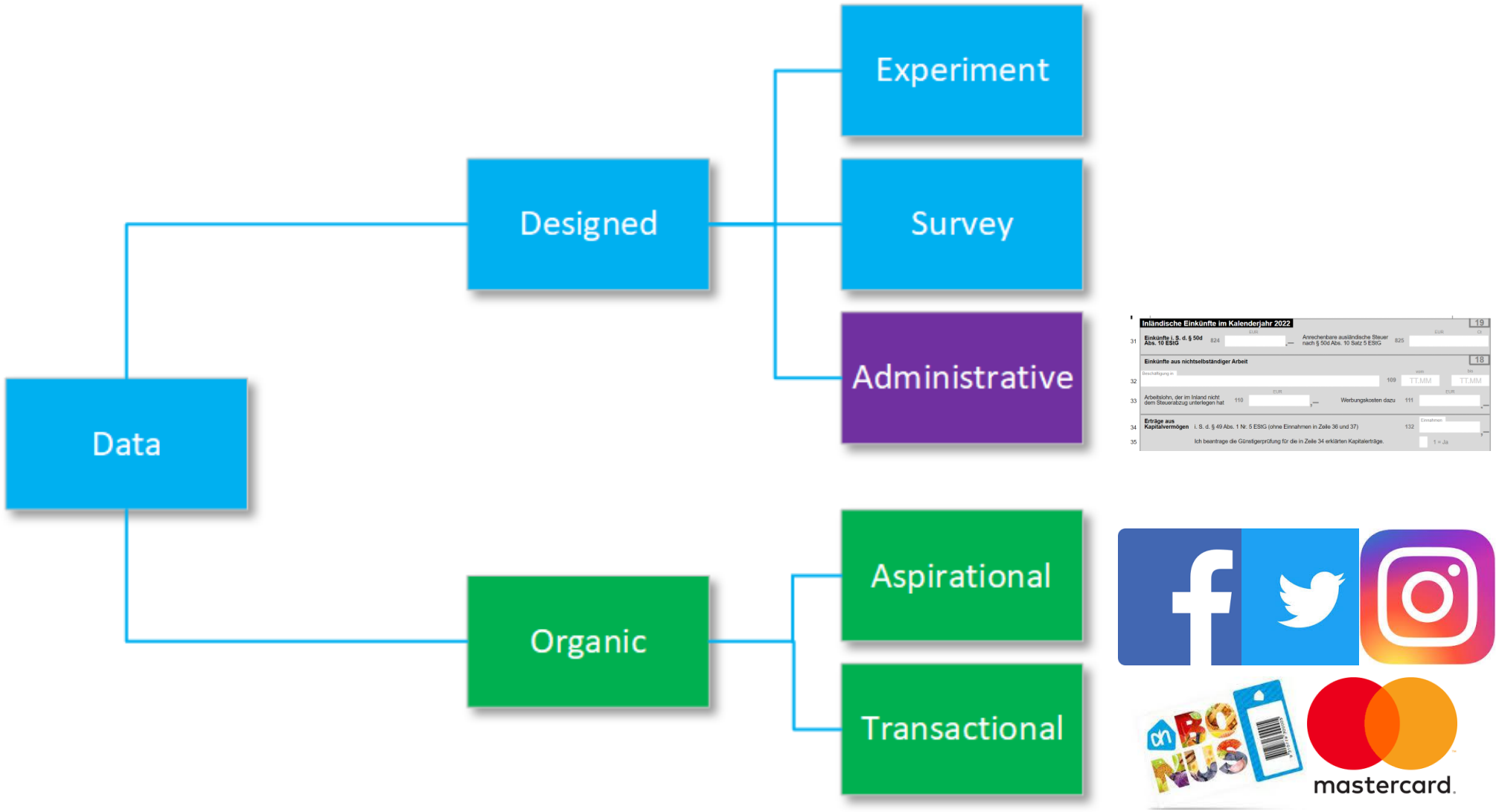
# Decreasing response rates





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



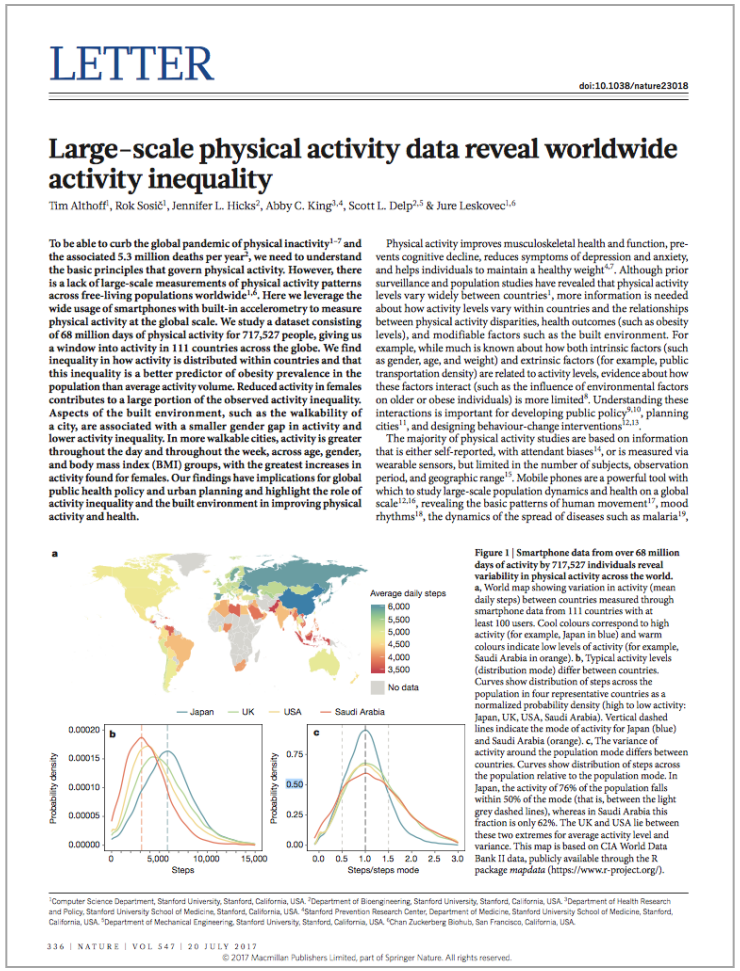

(Source: Kreuter 2018)

# Surveys & Big data

| | |
|---|---|
| • "Designed" data: Collected for the research purposes<br>• Researcher control over content<br>• Large number of covariates<br>• Detailed documentation of the data generating process | • "Organic" data: Collected for purposes other than research<br>• No control over content<br>• Limited number of covariates<br>• No / little documentation<br>• Access issues<br>• (Missingness & coverage) |
| • High nonresponse<br>• Small N<br>• Measurement error (recall, social desirability) | • Large N<br>• 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].

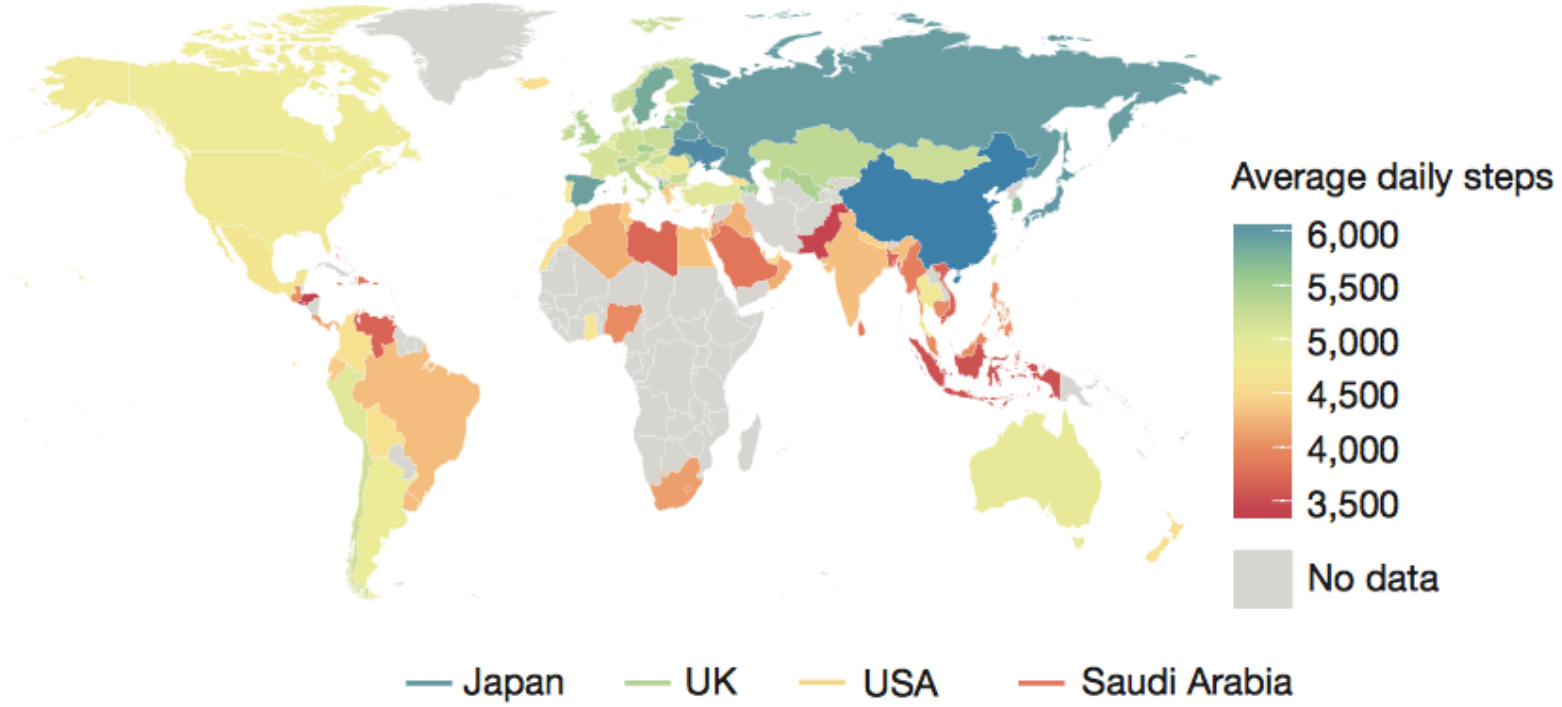
The majority of physical activity studies are based on information that is either self-reported, with attendant biases[14], or is measured via wearable sensors, but limited in the number of subjects, observation period, and geographic range[15]. Mobile phones are a powerful tool with which to study large-scale 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.

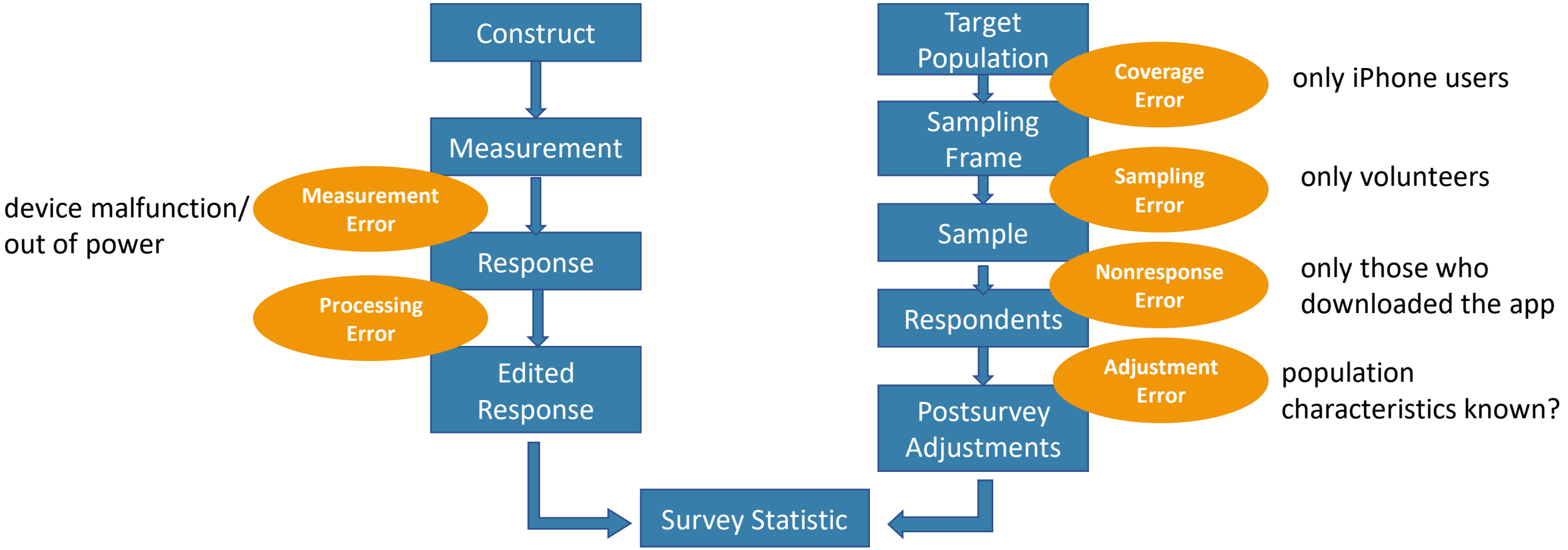







Althoff, T., Hicks, J. L., King, A. C., Delp, S. L., & Leskovec, J. (2017). Large-scale physical activity data reveal worldwide activity inequality. *Nature*, *547* (7663), *336-339*

# Total Survey Error & Althoff et al.



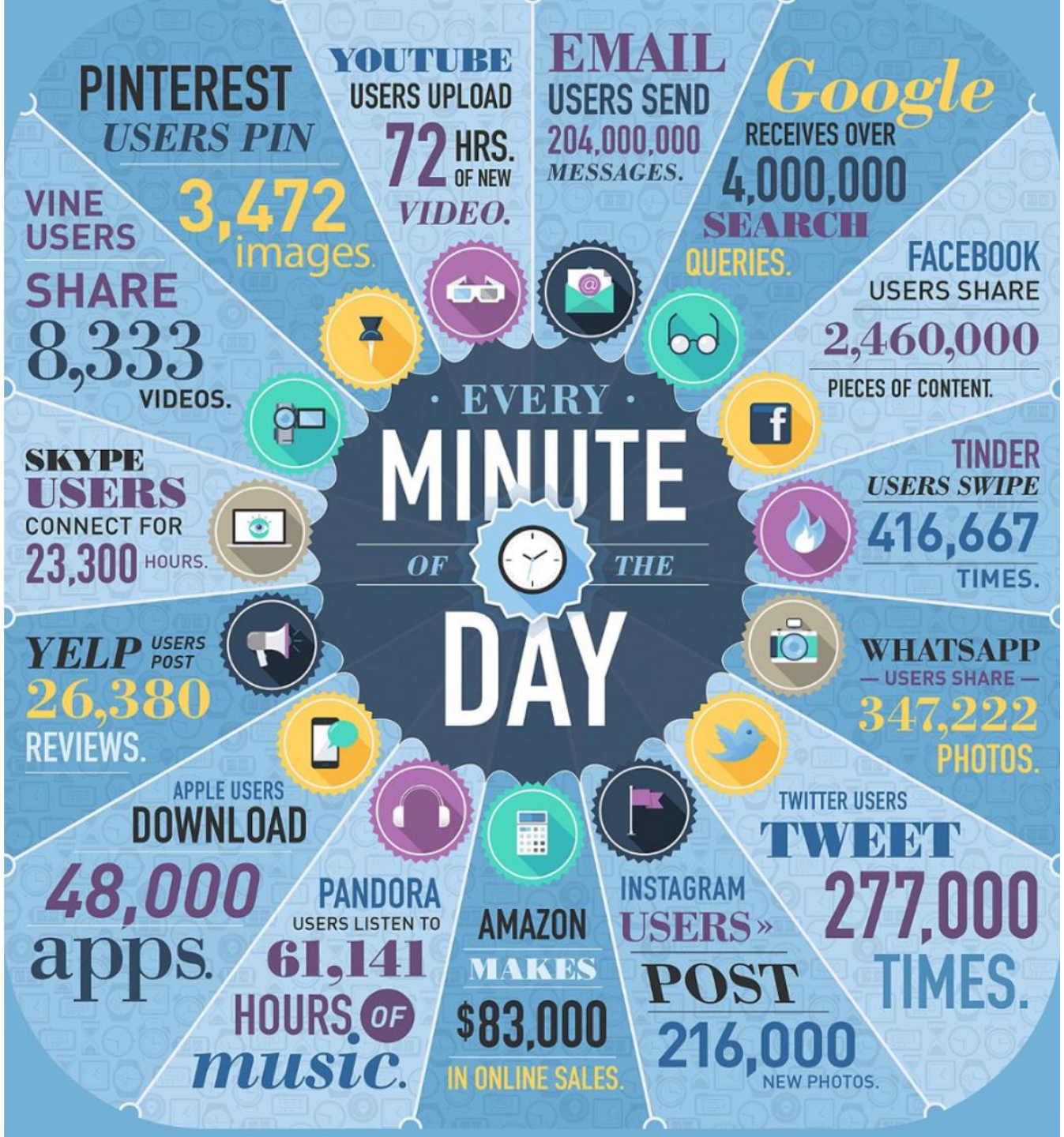

(Groves et al. 2004: 48)

# 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*: data 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:

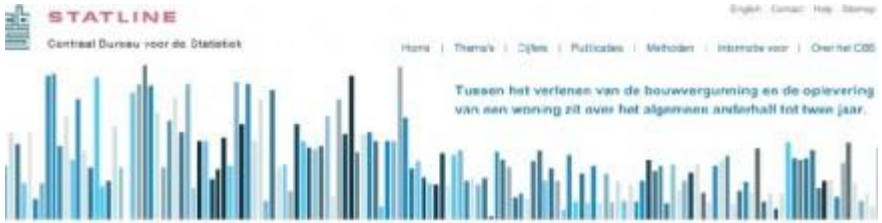
1) Volume
2) Variety ((no) structure)
3) Velocity
4) Veracity (accuracy)
5) Variability (differences in meanings across sources)
6) Value
7) Visualization

Source: Baker 2017, Infographic: James 2014

# 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** – geolocation, accelerometers, heartbeat

# Administrative data





- National Statistical Offices, Tax Office
- 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 *the same as* 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

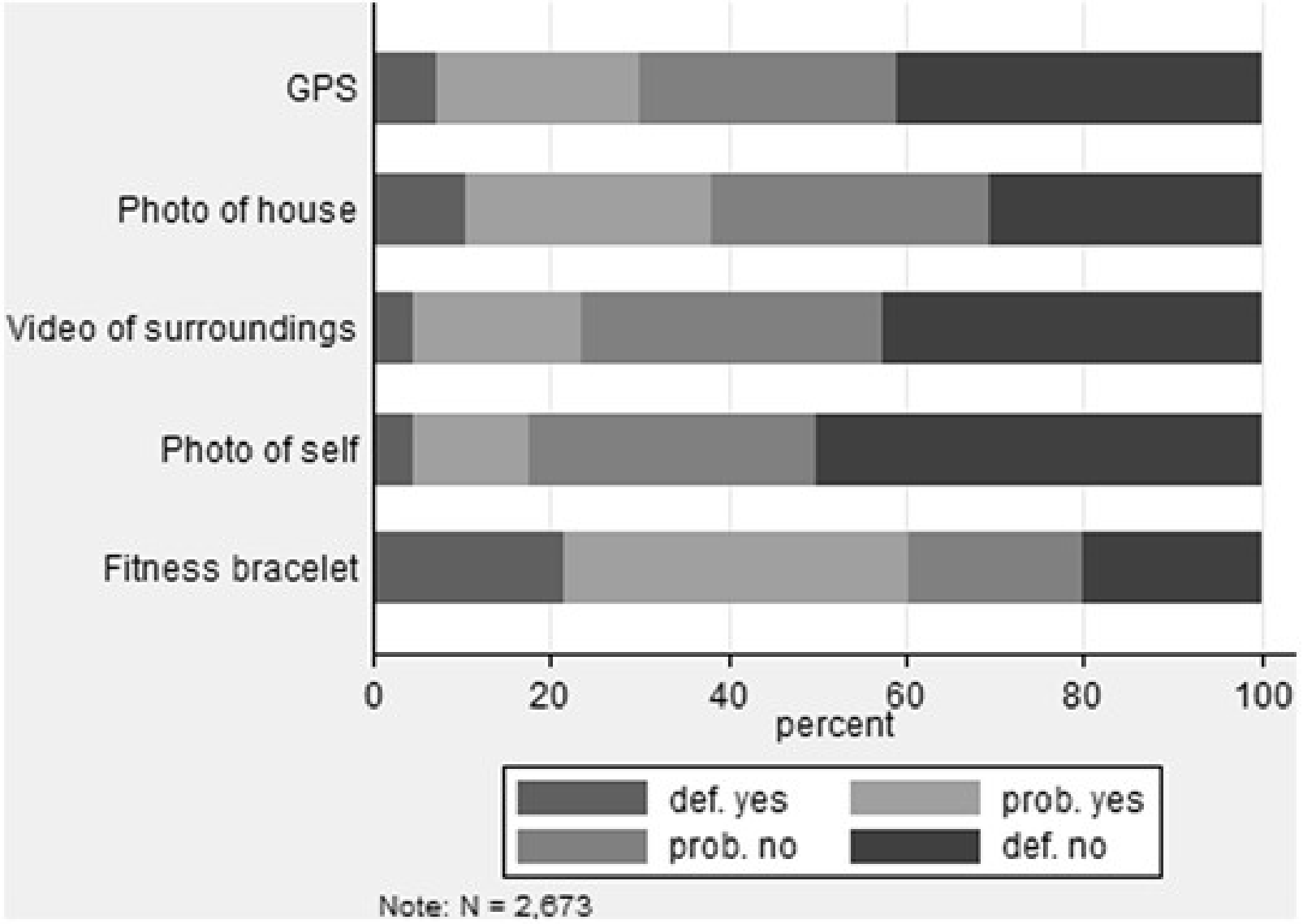
- Selection bias: “haves” versus “have-nots”
  - Not everyone uses social media!
  - Need to distinguish between producers and users of users of social media – small part of online population actively tweets
- Measurement bias
  - Self-presentation bias: Impression management is a key element of social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

# Sensor data



-Not everyone allows you to track them: only 25% is willing to track GPS coordinates (Toepoel & Lugtig, 2014)
-some type of activities are difficult to measure
-thresholds for intensity are arbitrary

# Willingness to collect sensor data

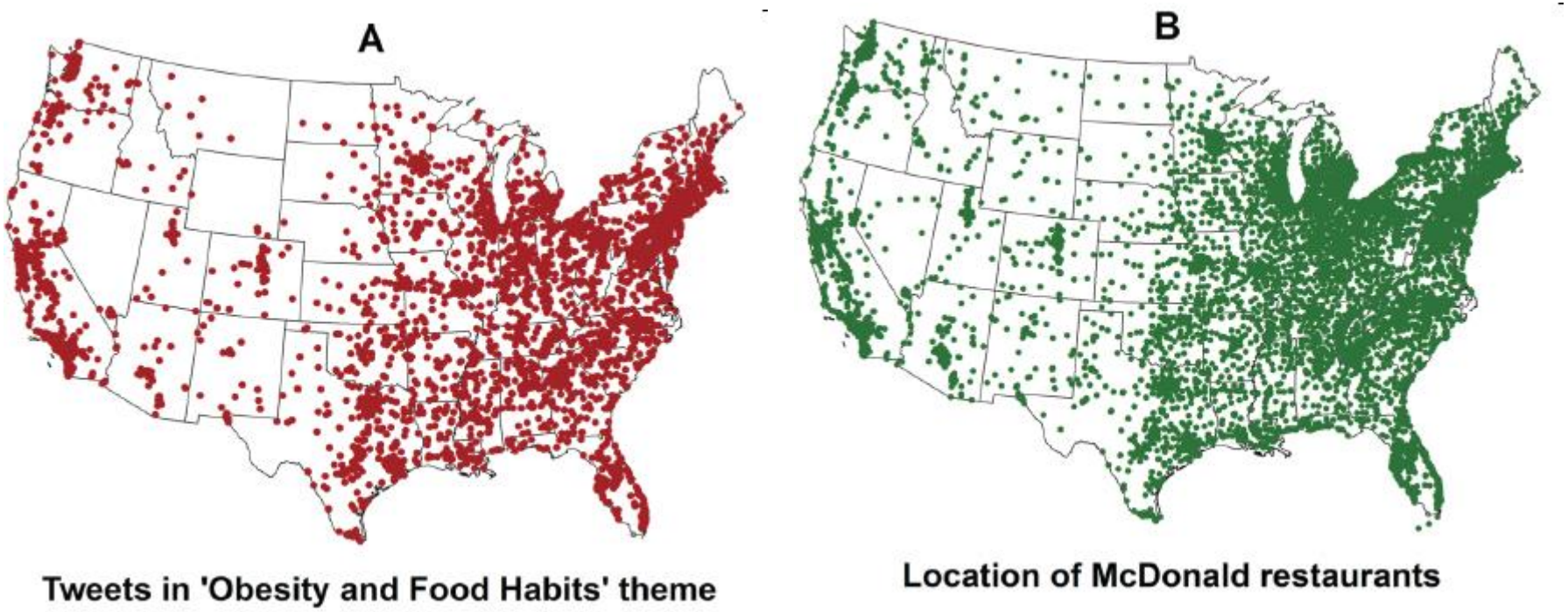



LISS Panel
November 2017

(Struminskaya et al. 2020)

## Few covariates:

Obesity-Related Tweets and McDonalds Restaurants



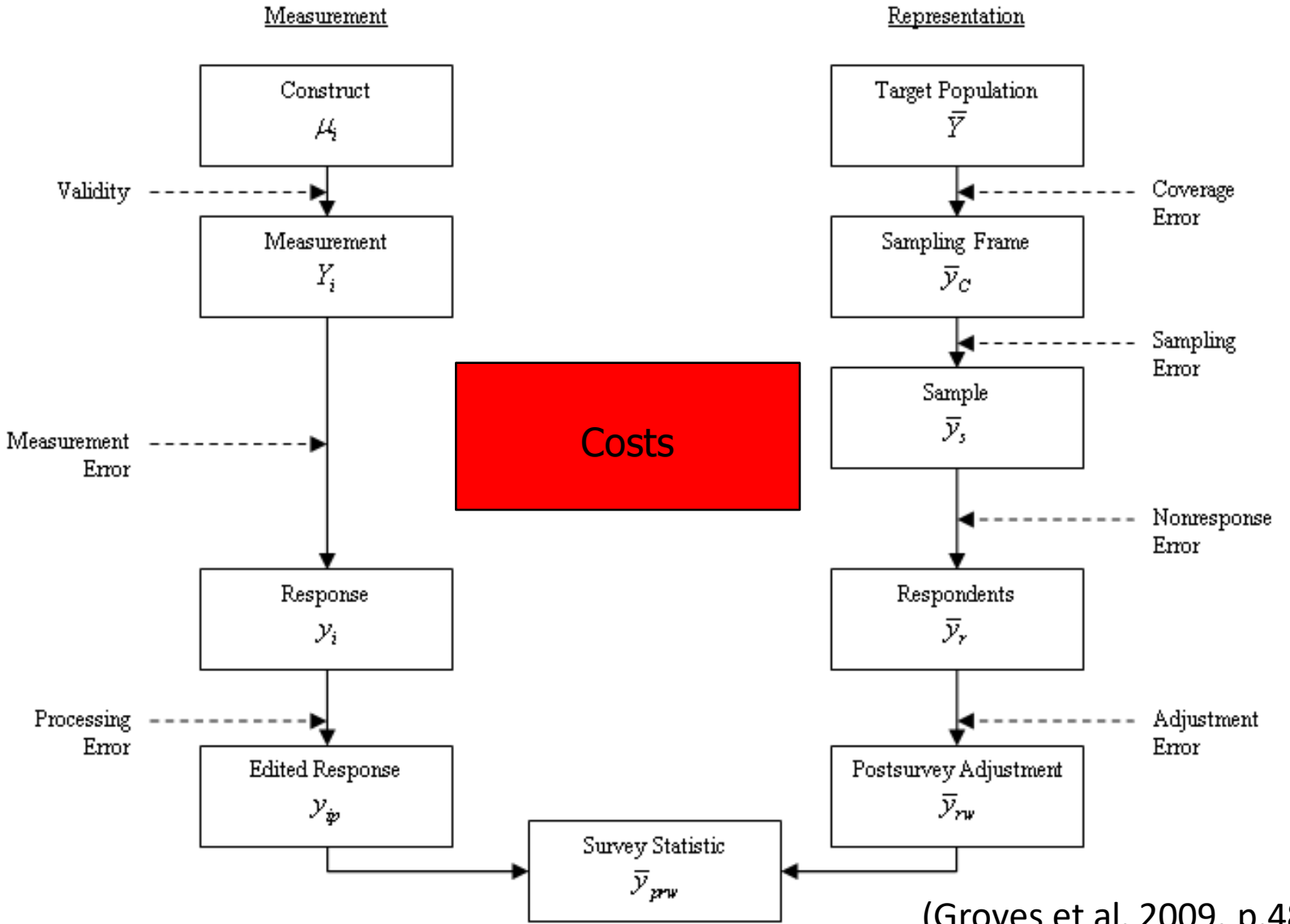

- Ghosh and Guha (2013) report a "strong correlation" between the two
- Any alternative explanation come to mind?

# 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 Framework




(Groves et al. 2009, p.48)

# 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 | -- |

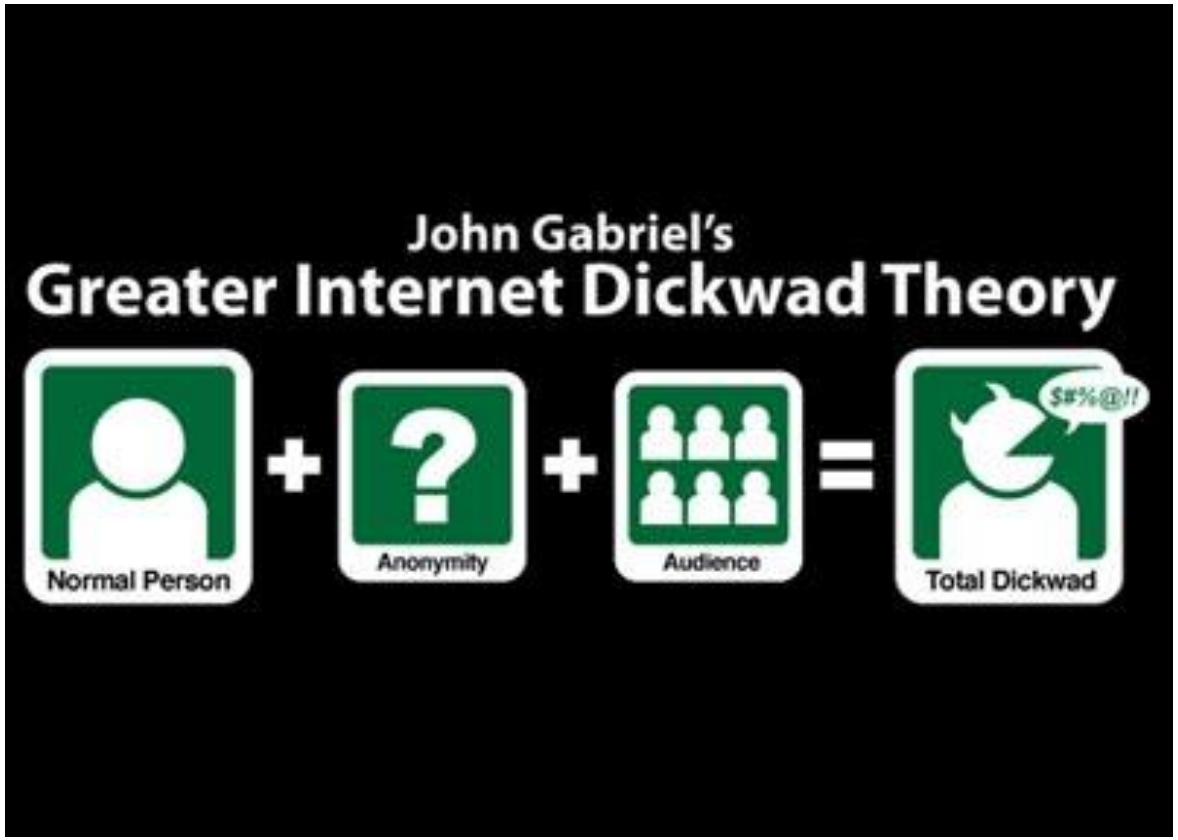
# Nonresponse in Big Data

- 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

A misconception?

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

Utrecht University

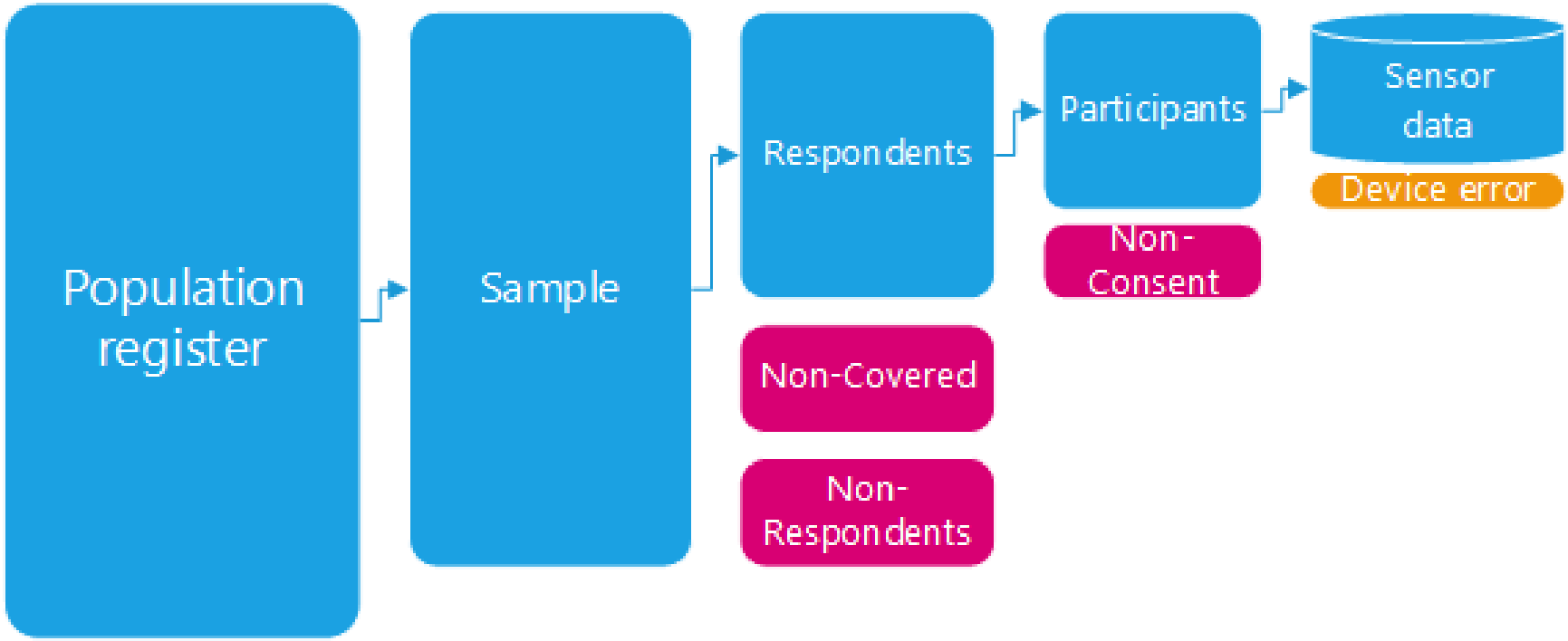
# Designed big data

Bella Struminskaya & Peter Lugtig

Department of Methodology & Statistics, Utrecht University

# Introducing “design” to Big Data

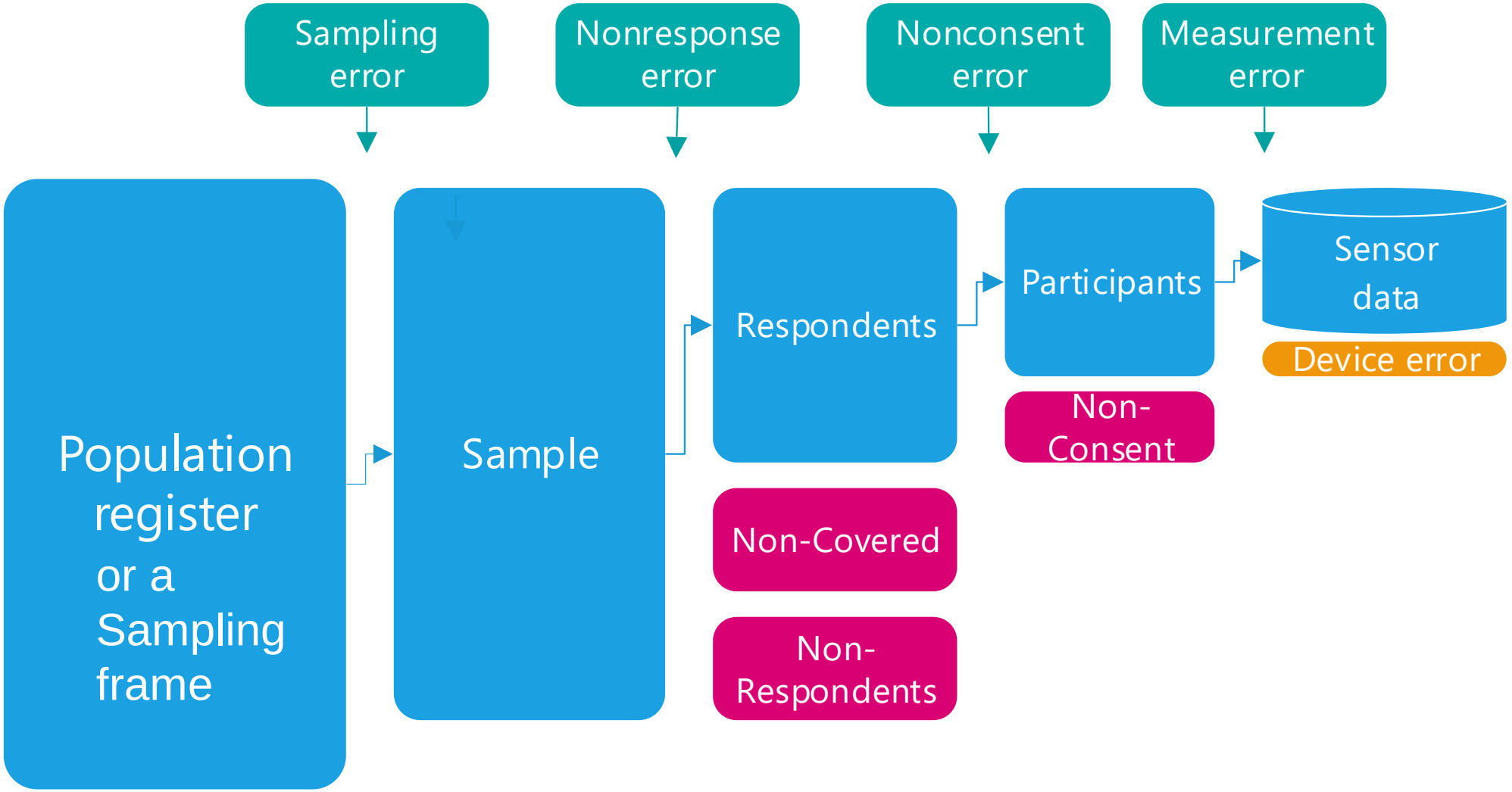
- Smartphone sensor data have many characteristics of Big Data
  - Large volume, high velocity, variety of data formats
- Combining passive smartphone data collection with self-reports through surveys introduces “design” to Big Data



# 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



# Example: Statistics Netherlands' Travel App

## An App-Assisted Travel Survey in Official Statistics: Possibilities and Challenges

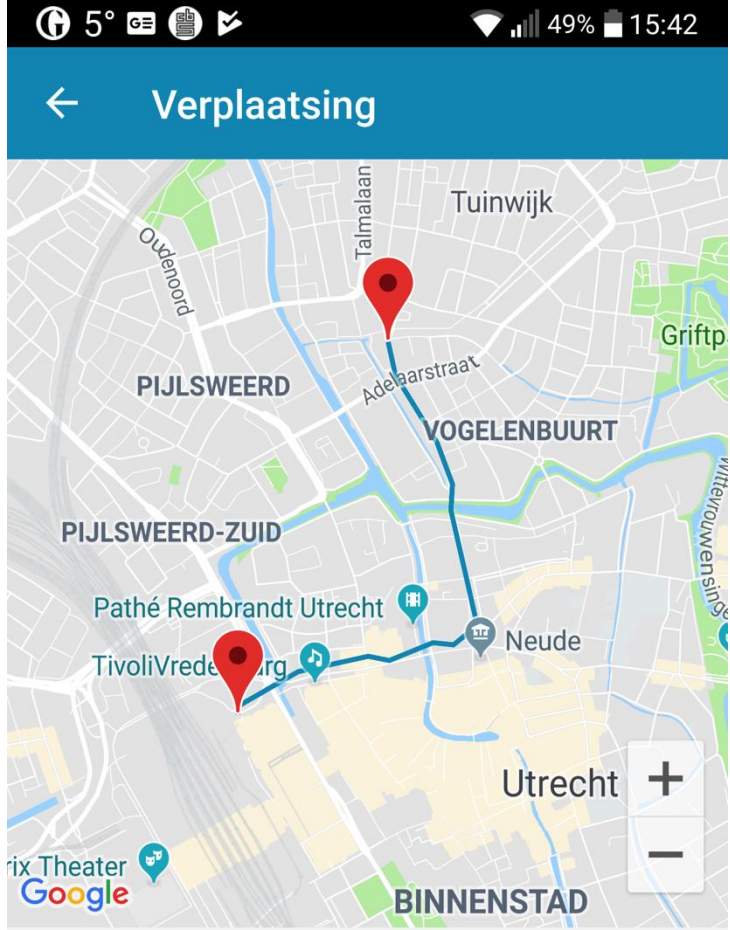
*Danielle McCool*[1], *Peter Lugtig*[1], *Ole Mussmann*[2], *and Barry Schouten*[3]

Advances in smartphone technology have allowed for individuals to have access to near-continuous location tracking at a very precise level. As the backbone of mobility research, the Travel Diary Study, has continued to offer decreasing response rates over the years, researchers are looking to these mobile devices to bridge the gap between self-report recall studies and a person's underlying travel behavior. This article details an open-source application that collects real-time location data which respondents may then annotate to provide a detailed travel diary. Results of the field test involving 674 participants are discussed, including technical performance, data quality and response rate.

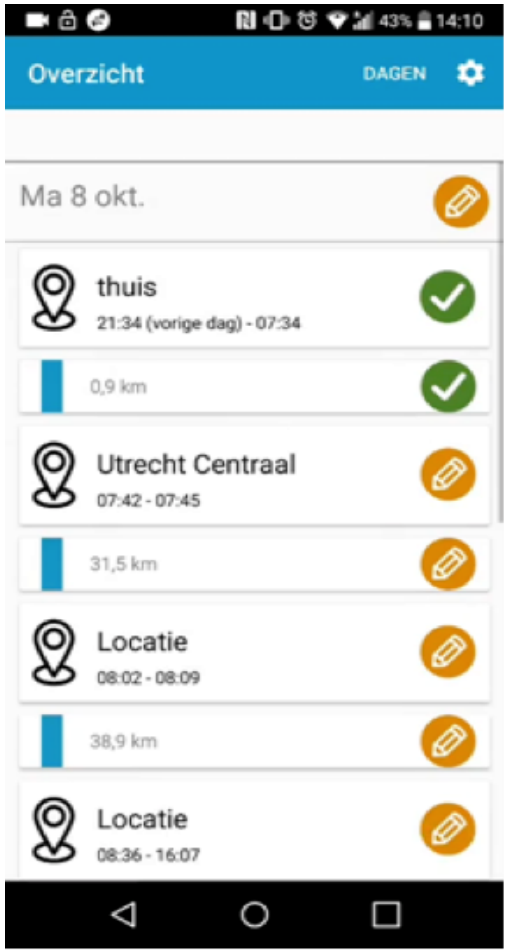
*Key words:* Non-response; travel diary; sensor data for surveys; app design; android background restriction.

### 1. Introduction

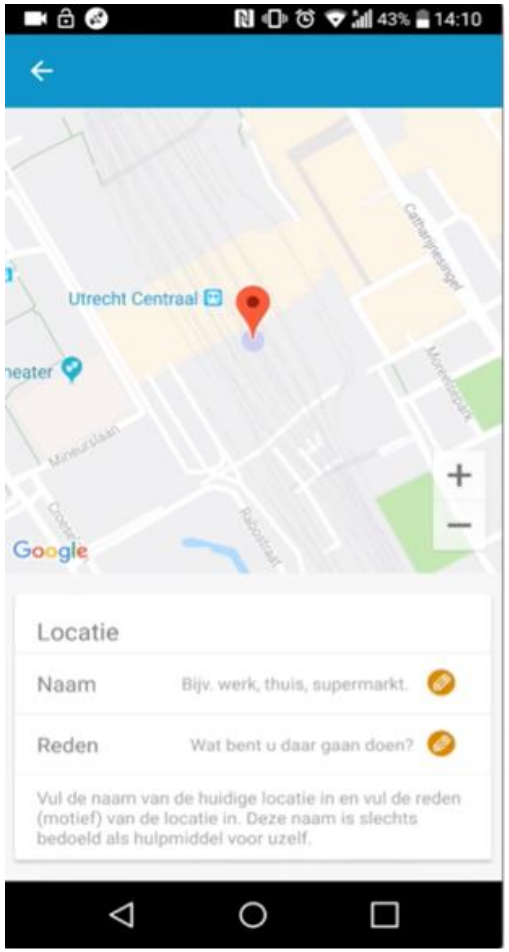
Understanding the true underlying movement behavior of persons in a given geographic area is a key component in the foundation of national infrastructure decisions. Institutions responsible for generating official statistics have designed streamlined instruments to enable the collection of important travel behavior metrics. Most organizations currently implement some form of travel diary survey (TDS), in which participants record a series of trips and stops over a specified time period. When these diaries are completed within probabilistic samples, the aggregate results can be used to model travel demand between



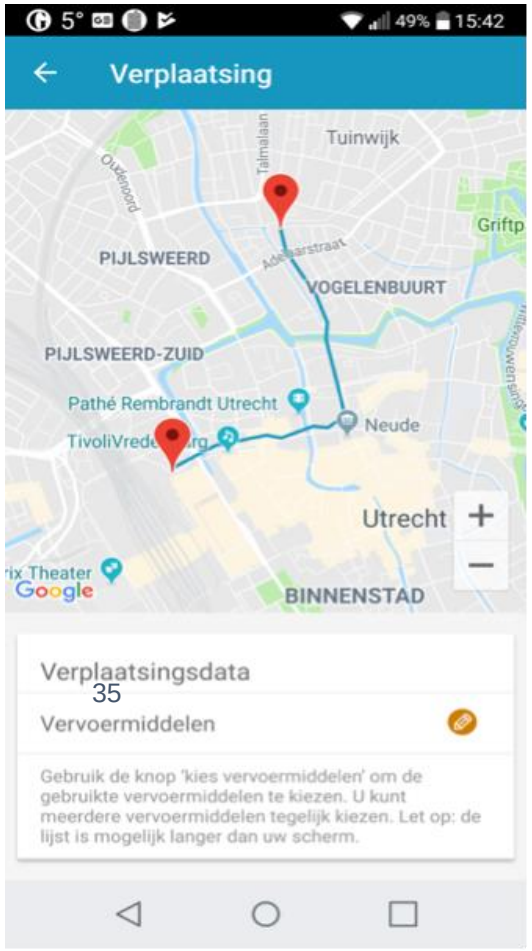
# The app in practice




Daily overview




Questions about stops




Questions about trips

# Fieldwork

## 1. Invitation letter

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.

# Fieldwork

## 1. Invitation letter

- 1b. Website

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

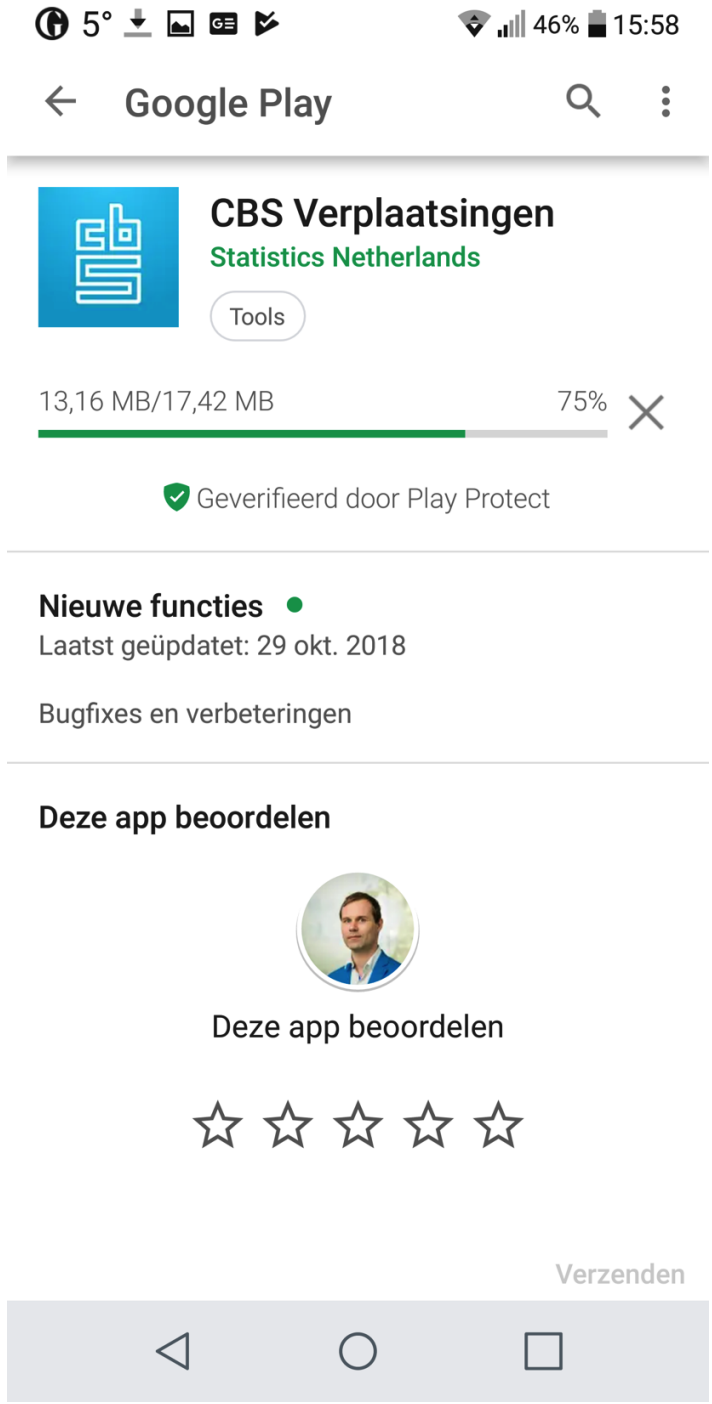
## Uw gegevens zijn veilig

**Direct naar**

- Hoe bedien ik de app in Android?
- Hoe bedien ik de app in iOS?
- Veelgestelde vragen

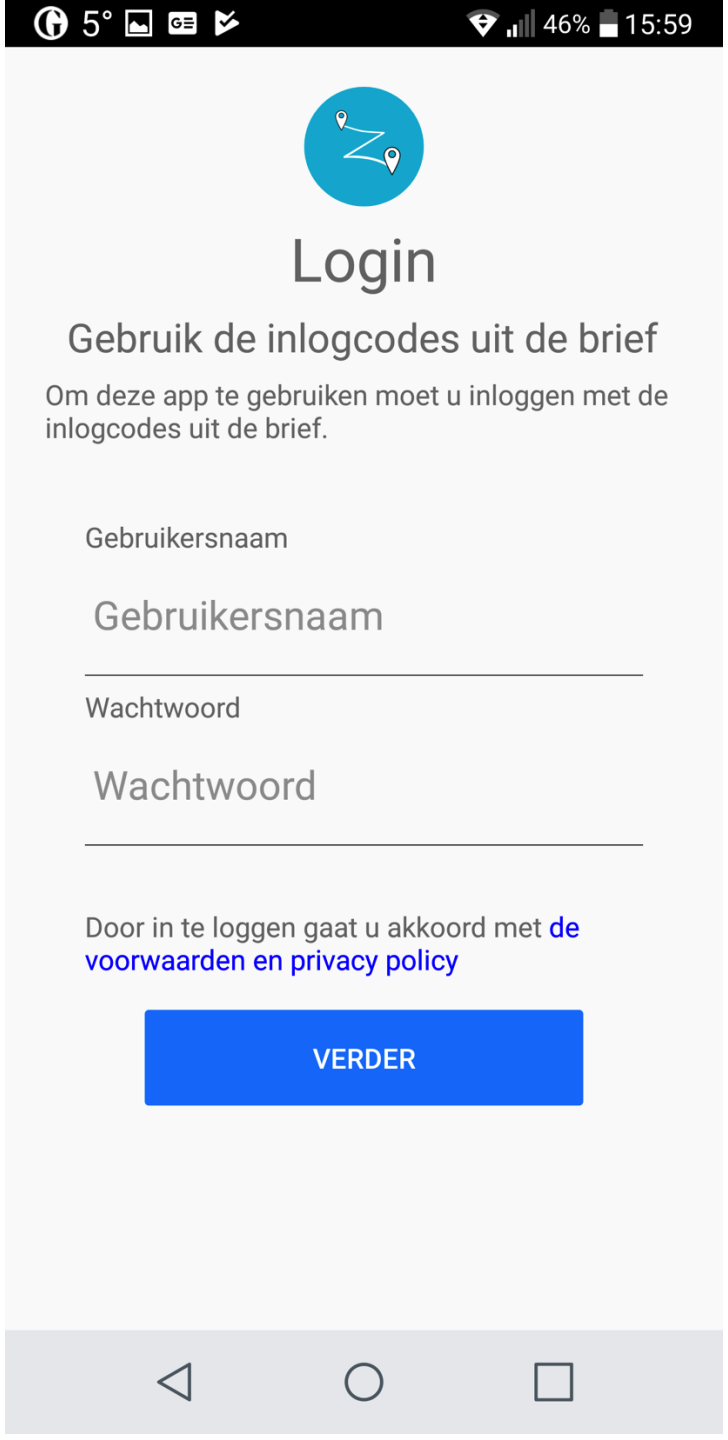
# Fieldwork

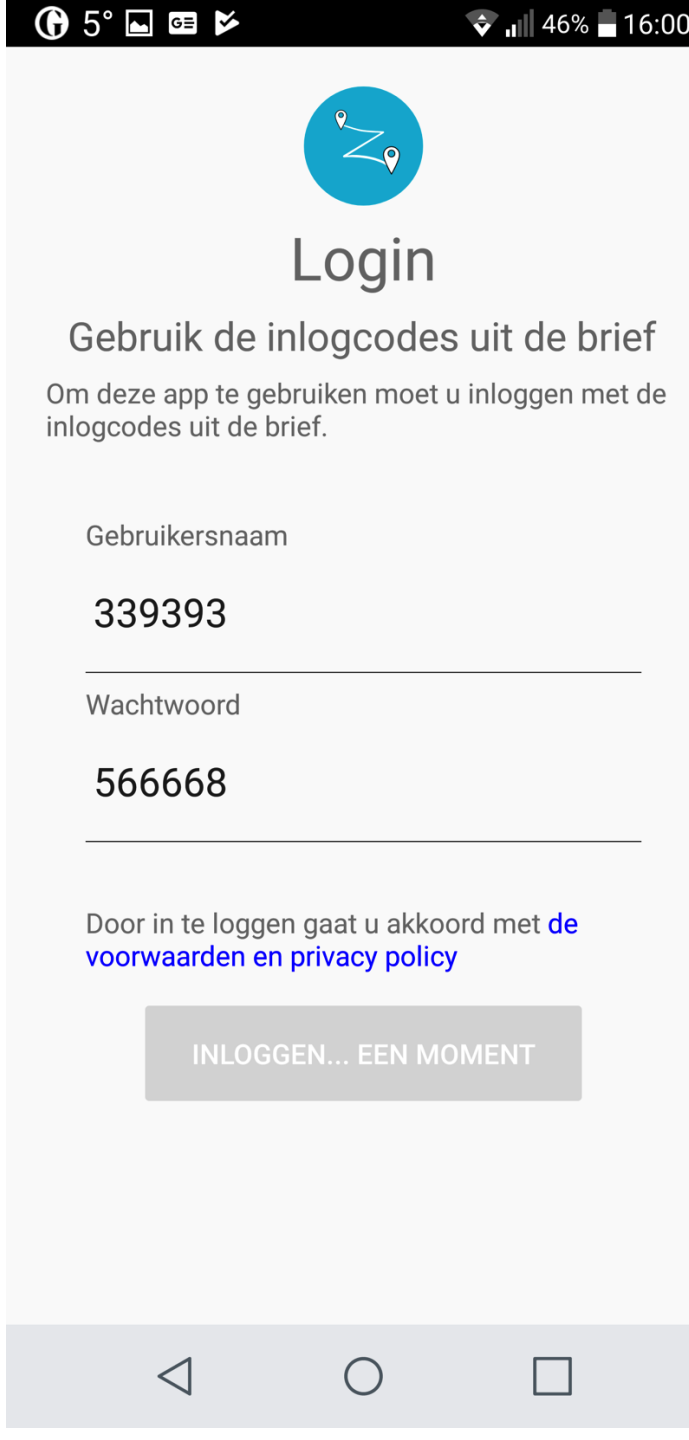
1. Invitation letter
    - 1b. Website
2. Download app



# Fieldwork

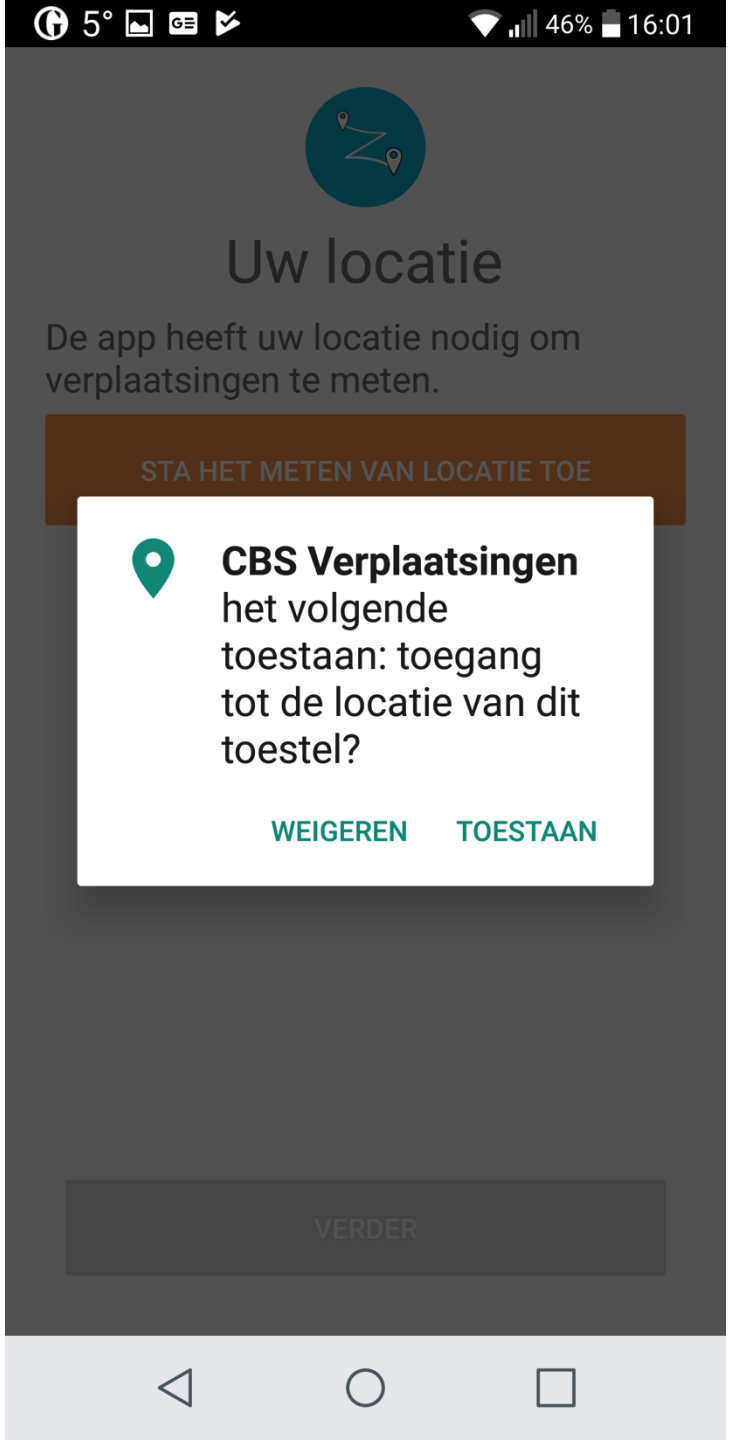
1. Invitation letter
    1b. Website
2. Download app
3. Login

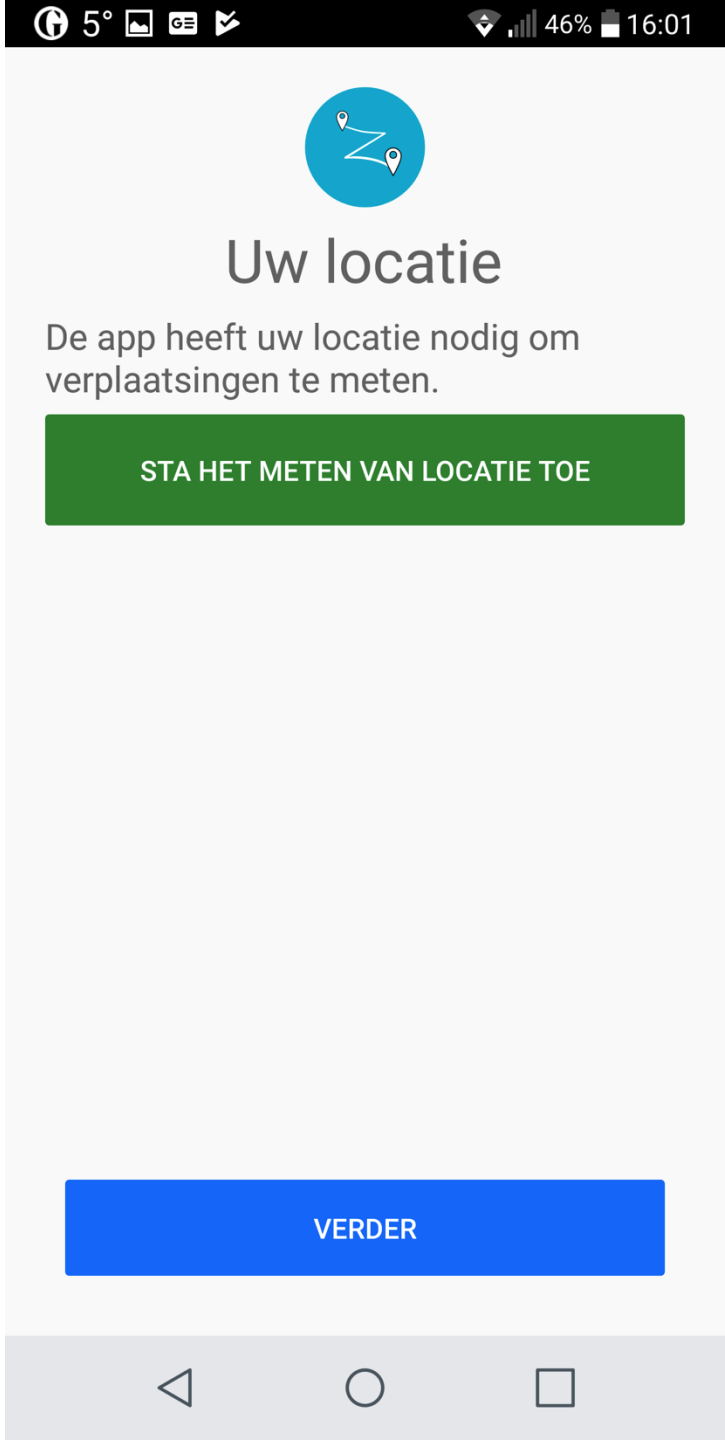


# Fieldwork

1. Invitation letter
   - 1b. Website
2. Download app
3. Login
4. Allow location measurements



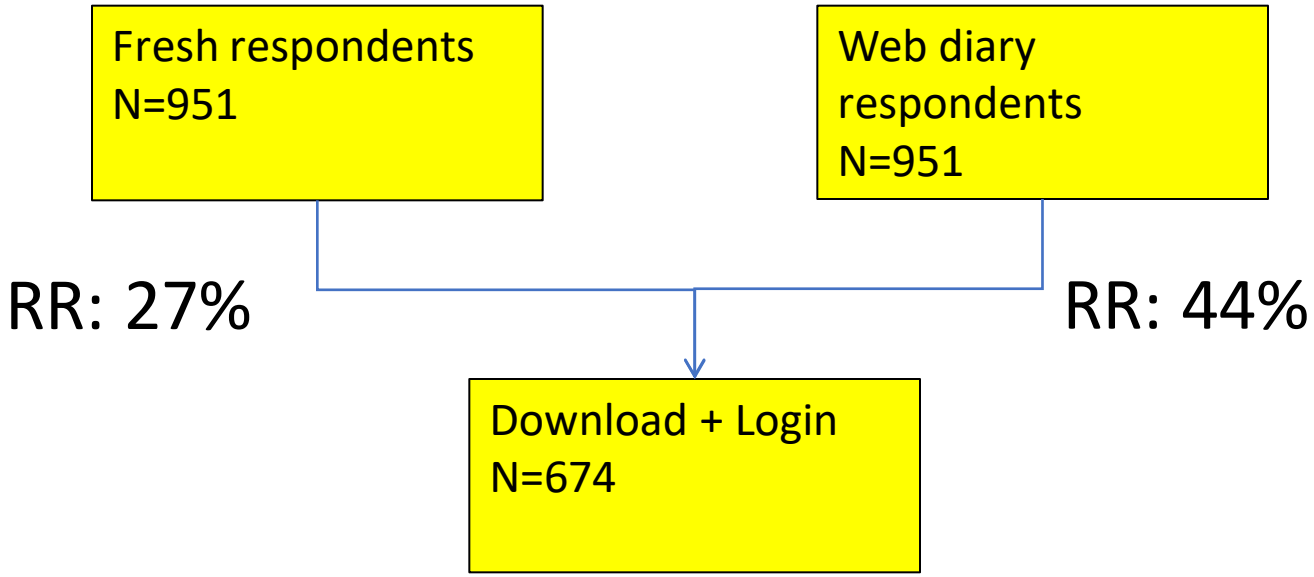



# Fieldwork (overview)

- Start 31 October 2018
- Reminder 14 November
- Encouragement 21 November to those who started, but sent <7 days data

Fresh respondents
N=951

Web diary
respondents
N=951


Fresh respondents
N=951
Web diary respondents
N=951
RR: 27%
RR: 44%
Download + Login
N=674

Fresh respondents
N=951

Web diary respondents
N=951

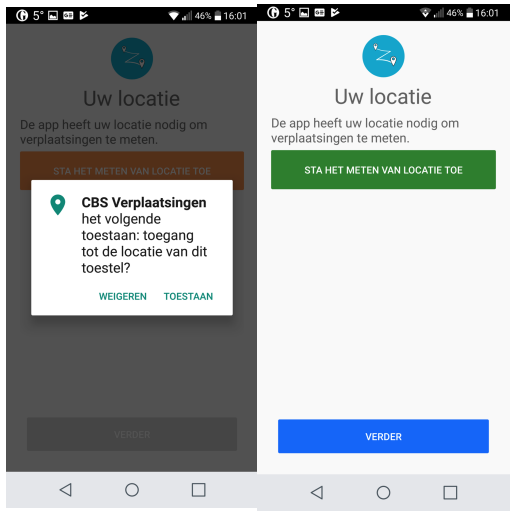
RR: 27%

RR: 44%

Download + Login
N=674

dropout: 17%

Send data for at least 1 day
N=578



Fresh respondents
N=951

Web diary respondents
N=951

RR: 26.5%

RR: 44.4%

Download + Login
N=674

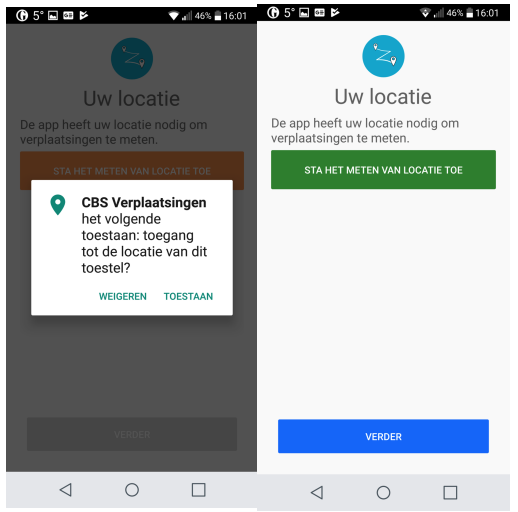
dropout: 17%

Send data for at least 1 day
N=578

dropout: 28%

Complete whole study
N=421

Complete responses:
22% of sample



| | Stage 1 | | | |
|---|---|---|---|---|
| | *AME* | *B* | *se* | *p* |
| **Intercept** | | -0.975 | 0.374 | ** |
| **Sample (ref = New)** | | | | |
| ODiN | 0.158 | 0.769 | 0.104 | *** |
| **Incentive (ref = 5-5-5)** | | | | |
| 5-10 | 0.074 | 0.375 | 0.128 | ** |
| 5-20 | 0.104 | 0.521 | 0.127 | *** |
| **Age (ref = 16-25)** | | | | |
| 26-45 | -0.124 | -0.565 | 0.185 | ** |
| 46-65 | -0.180 | -0.84 | 0.189 | *** |
| >65 | -0.264 | -1.29 | 0.231 | *** |
| **Education (ref = Basisonderwijs)** | | | | |
| Vmbo, avo onderbouw, mbo 1 | -0.012 | -0.071 | 0.332 | |
| Havo, vwo, mbo | 0.100 | 0.513 | 0.31 | |
| Hbo-, wo-bachelor | 0.207 | 0.998 | 0.327 | ** |
| Hbo-, wo-master, doctor | 0.200 | 0.97 | 0.351 | ** |
| Unknown | 0.056 | 0.299 | 0.315 | |
| **Marital status (ref = Married)** | | | | |
| Single | -0.060 | -0.302 | 0.119 | * |
| **Origin (ref = Dutch)** | | | | |
| Not-western | -0.110 | -0.573 | 0.209 | ** |
| Western | -0.049 | -0.244 | 0.179 | |
| **Income (ref = 0-20)** | | | | |
| 21-40 | -0.097 | -0.528 | 0.225 | * |
| 41-60 | 0.025 | 0.125 | 0.208 | |
| 61-80 | 0.072 | 0.347 | 0.202 | |
| 81-100 | 0.064 | 0.31 | 0.203 | |
| Unknown | -0.025 | -0.125 | 0.562 | |
| **Gender (ref = Male)** | | | | |
| Female | | | | |

Covariates from population register:

- Sample (ODIN vs. new)
- Incentive (ref: 5+5+5)
- Age
- Education
- Marital status
- Income quantiles
- Gender
- Having a driver’s license
- Urbanicity
- Ethnicity (ref=Dutch)

Utrecht University

# Digital Trace Data

Bella Struminskaya & Peter Lugtig

Department of Methodology & Statistics, Utrecht University

# What are digital traces and what can they measure?

Slides by Bella Struminskaya and Florian Keusch

# Definition of digital traces

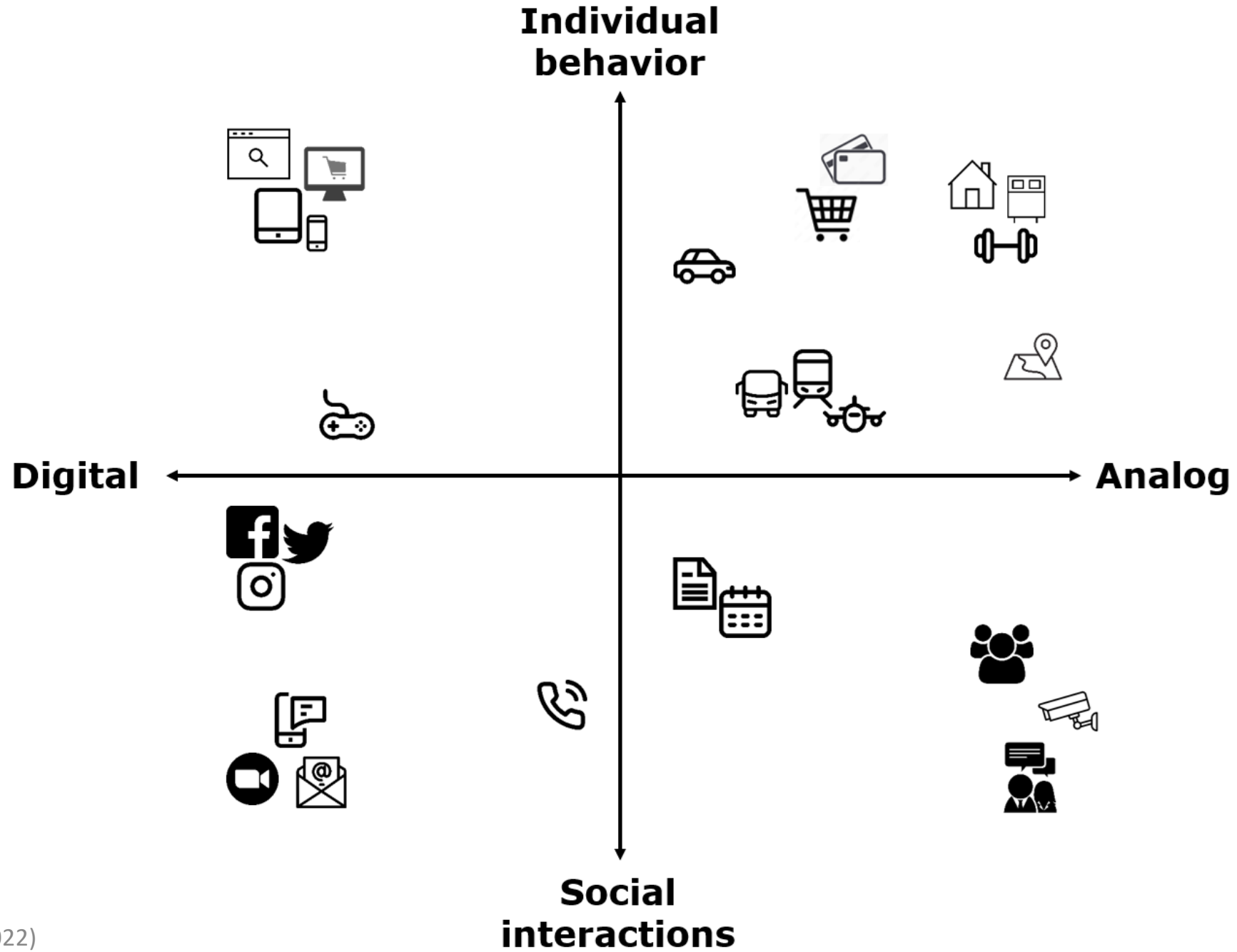
"Records of activity (trace data) undertaken through an online information system (thus, digital)"

(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

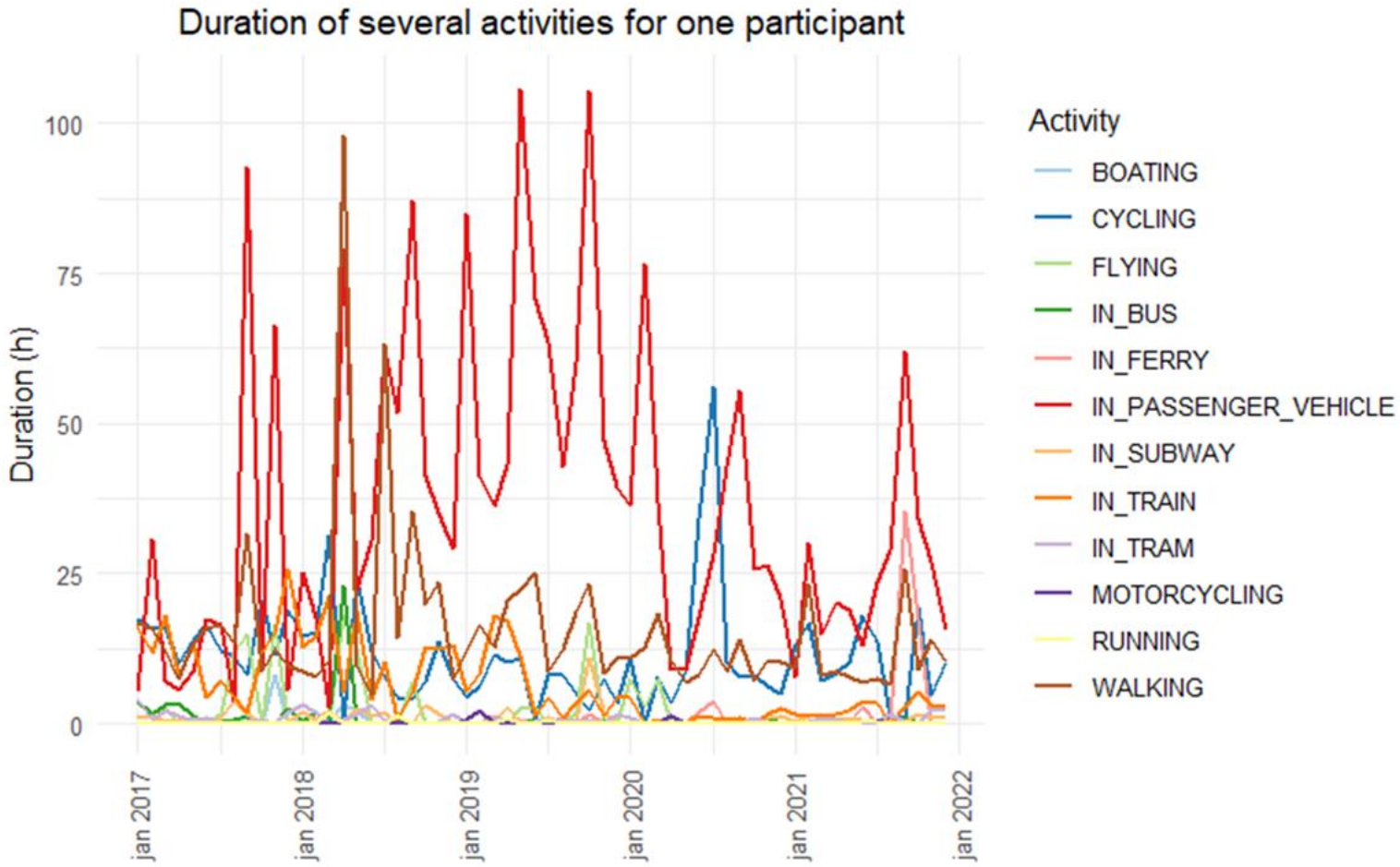
# Advantages of digital trace data

- Allows measurement of behavior at high frequency (in-the-moment)
  - Highly detailed measures
  - Evaluation of moment-to-moment changes
  - Without increasing response burden on participants
  - Scalable

# Example: Internet Usage

| ID | URL | used_at | active_seconds |
|---|---|---|---|
| 1 | tipico.de/de/online-sportwetten | 2017-09-12 10:45:15 | 83 |
| 1 | rib-software.com/de/loesungen/architektur-und-bauplanung/arriba-ava.html | 2017-09-12 10:45:19 | 40 |
| 1 | wetteronline.de/wettertrend/cala-d-or | 2017-09-12 10:45:23 | 182 |
| 1 | facebook.com/messages/t | 2017-09-12 10:45:36 | 2 |
| 1 | facebook.com/messages/t/Peter.Mustermann | 2017-09-12 10:45:44 | 74 |
| 1 | severin-produkttest.de | 2017-09-12 10:45:55 | 8 |
| 1 | severin-produkttest.de/bewerbung | 2017-09-12 10:46:05 | 46 |
| 1 | sparda-b.de | 2017-09-12 10:46:37 | 6 |
| 2 | rp-online.de/nrw/staedte/duesseldorf/airport-duesseldorf-flughafen-nicht-verantwortlich-fuer-das-was-dort-passiert-aid-1.7075717 | 2017-09-12 11:19:48 | 9 |
| 2 | autoscout24.de/ergebnisse?mmvmk0=29&mmvmd0=15537&mmvco=1&body=5&pricefrom=0&priceto=8000&cy=D&zipr=50&fuel=B&powertype=kw&offer=U&offer=J&offer=O&offer=D&lat=49.15148&lon=9.17345&zip=Heilbronn&atype=C&ustate=N%2CU&sort=standard&desc=0&page=1&size=20 | 2017-09-12 12:01:10 | 28 |
| 2 | trauer.weser-kurier.de | 2017-09-12 12:06:45 | 5 |
| 2 | amazon.de/Kee-Hearts/dp/B072Z8DMRY/ref=sr_1_1?ie=UTF8&qid=1505211277&sr=8-1&keywords=kee+of+hearts | 2017-09-12 12:14:44 | 10 |
| 2 | sat1gold.de/tv/zwei-bei-kallwass/video/117-brunos-geheimnis-ganze-folge | 2017-09-12 12:15:10 | 49 |
| 2 | bibeltv.de/livestreams | 2017-09-12 12:28:40 | 1 |
| 2 | arbeitsagentur.de/meine-eservices | 2017-09-12 12:36:46 | 11 |
| 3 | facebook.com | 2017-09-12 12:55:35 | 6 |
| 3 | bild.de/news/ausland/hurrikan-irma/irma-live-ticker-karibik-puerto-rico-domrep-haiti-florida-usa-53130320.bild.html | 2017-09-12 12:55:41 | 2 |
| 3 | spiegel.de | 2017-09-27 09:56:46 | 14 |
| 3 | https://www.spiegel.de/politik/deutschland/wolfgang-schaeuble-ueber-grosse-koalitionen-kann-auf-dauer-nur-schiefgehen-a-1299654.html | 2017-09-27 09:57:00 | 195 |

# Example: Google Semantic Location History



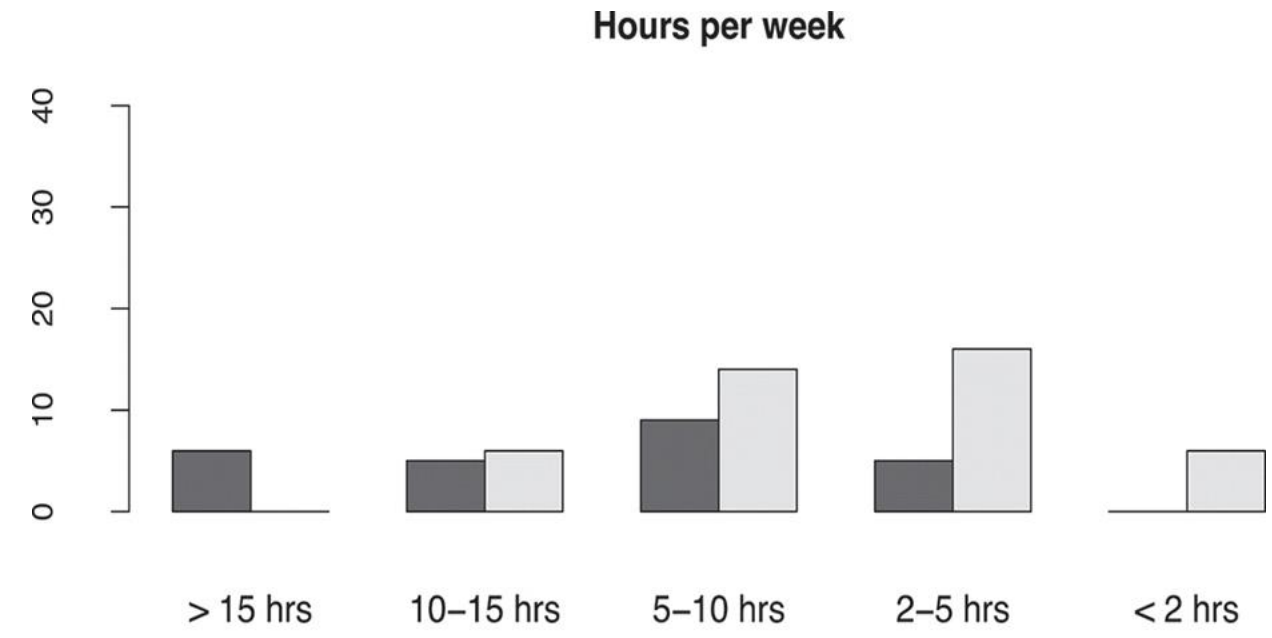
# Advantages of digital trace data

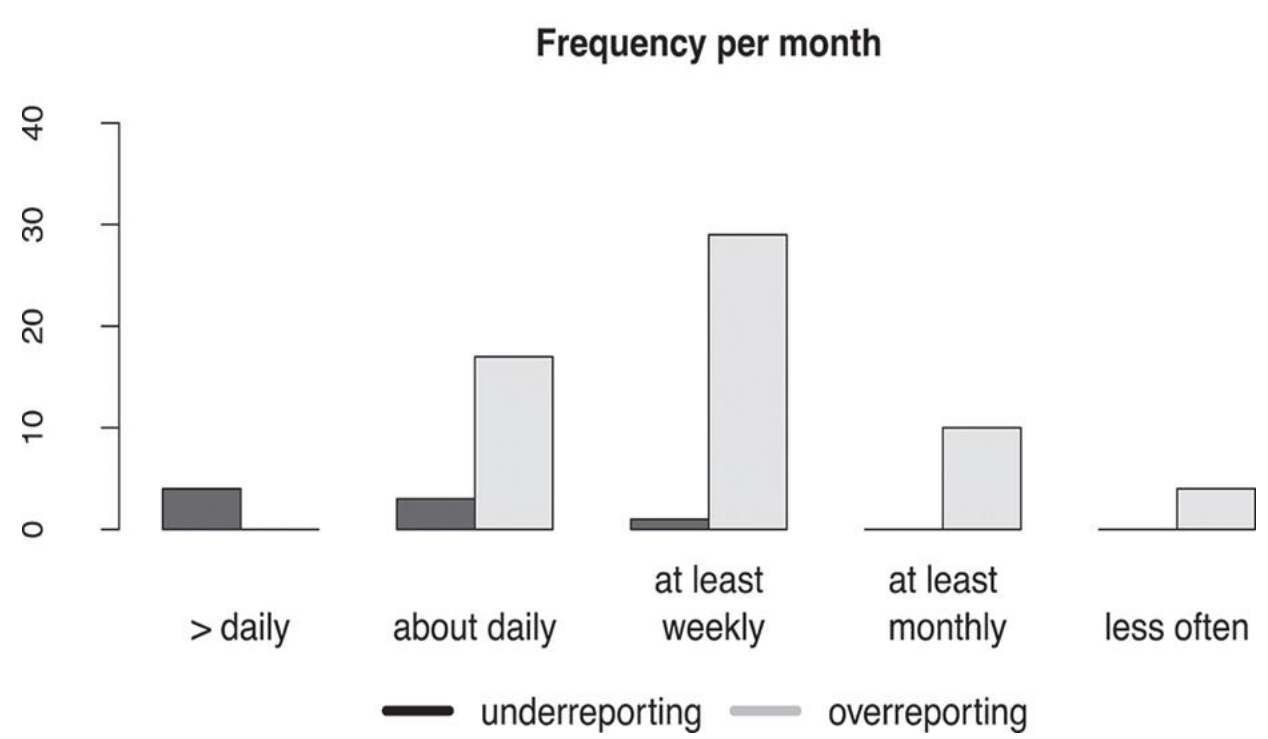
- Allows measurement of behavior at high frequency (in-the-moment)
- Measurement happens “passively”
  - Without direct solicitation of subject studied
  - Digital trace data should be unaffected by measurement itself
  - Reduced measurement error

# Example: Smartphone Use (Jones-Jang et al. 2020)

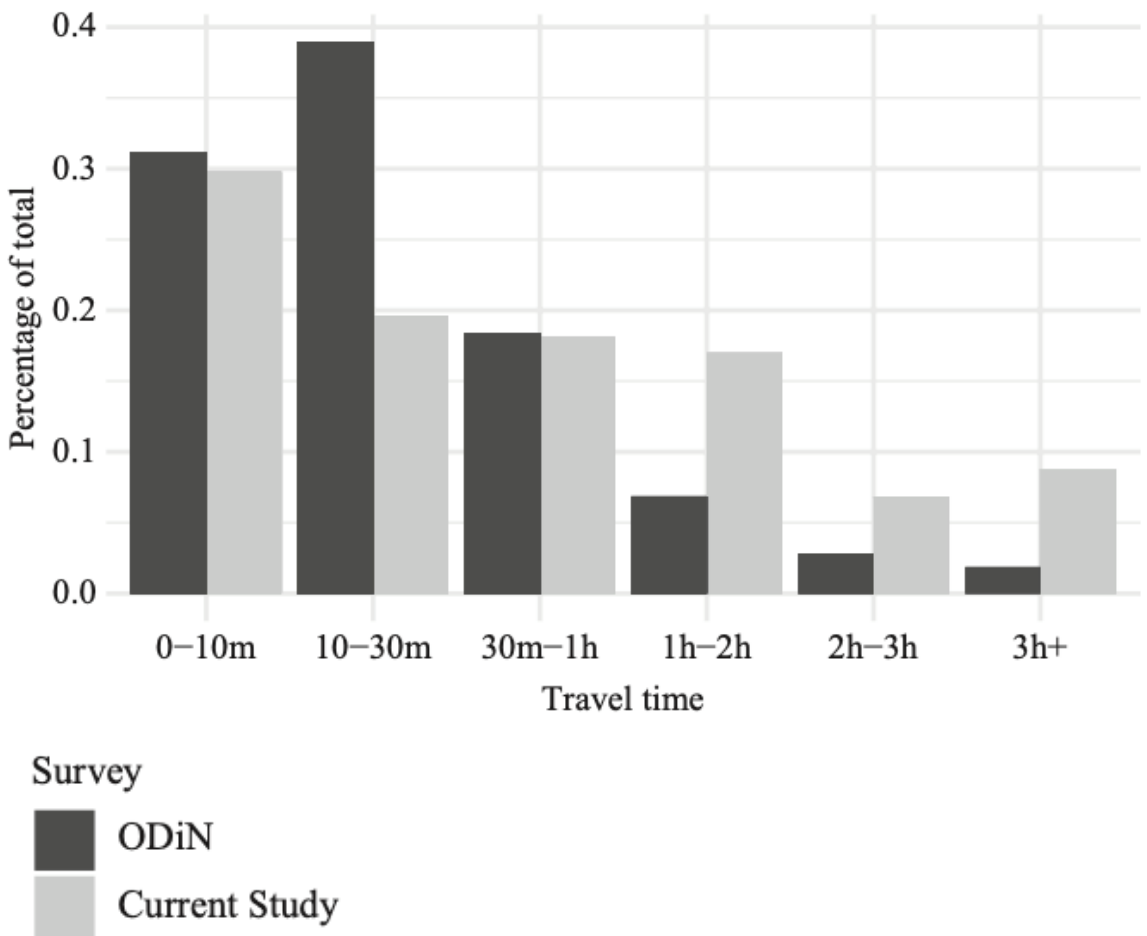
# Example: Internet Use (Sharkow 2016)

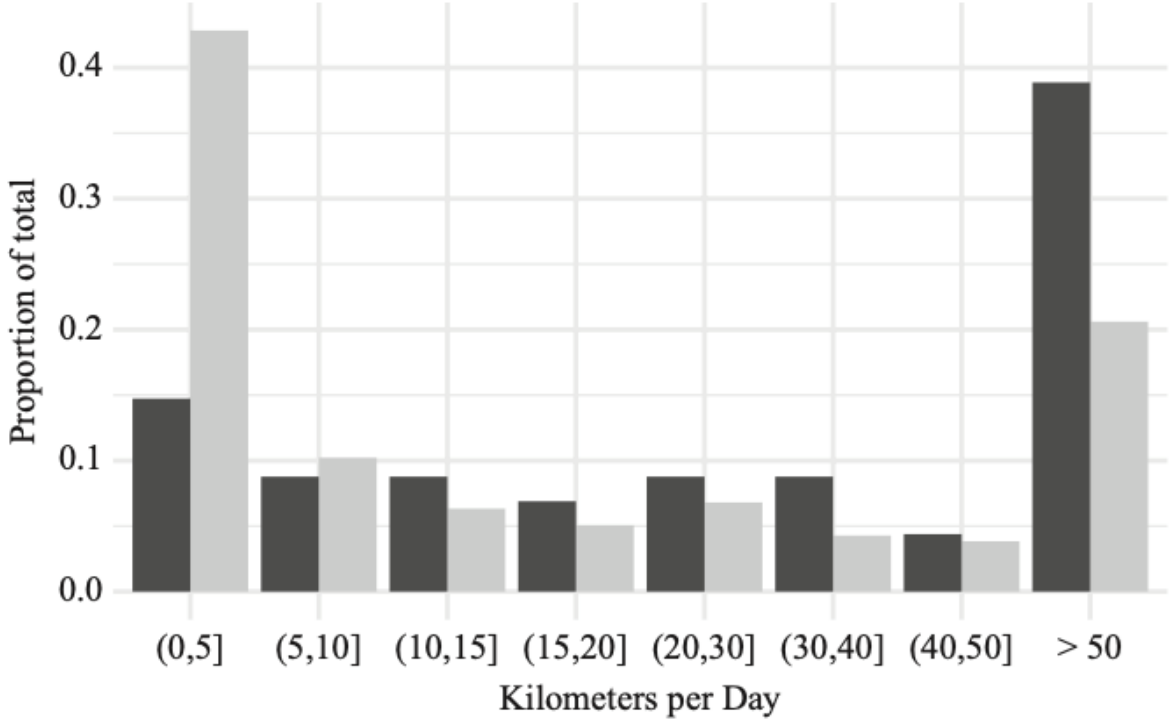
- Under- and overreporting in surveys by “metered” Internet use



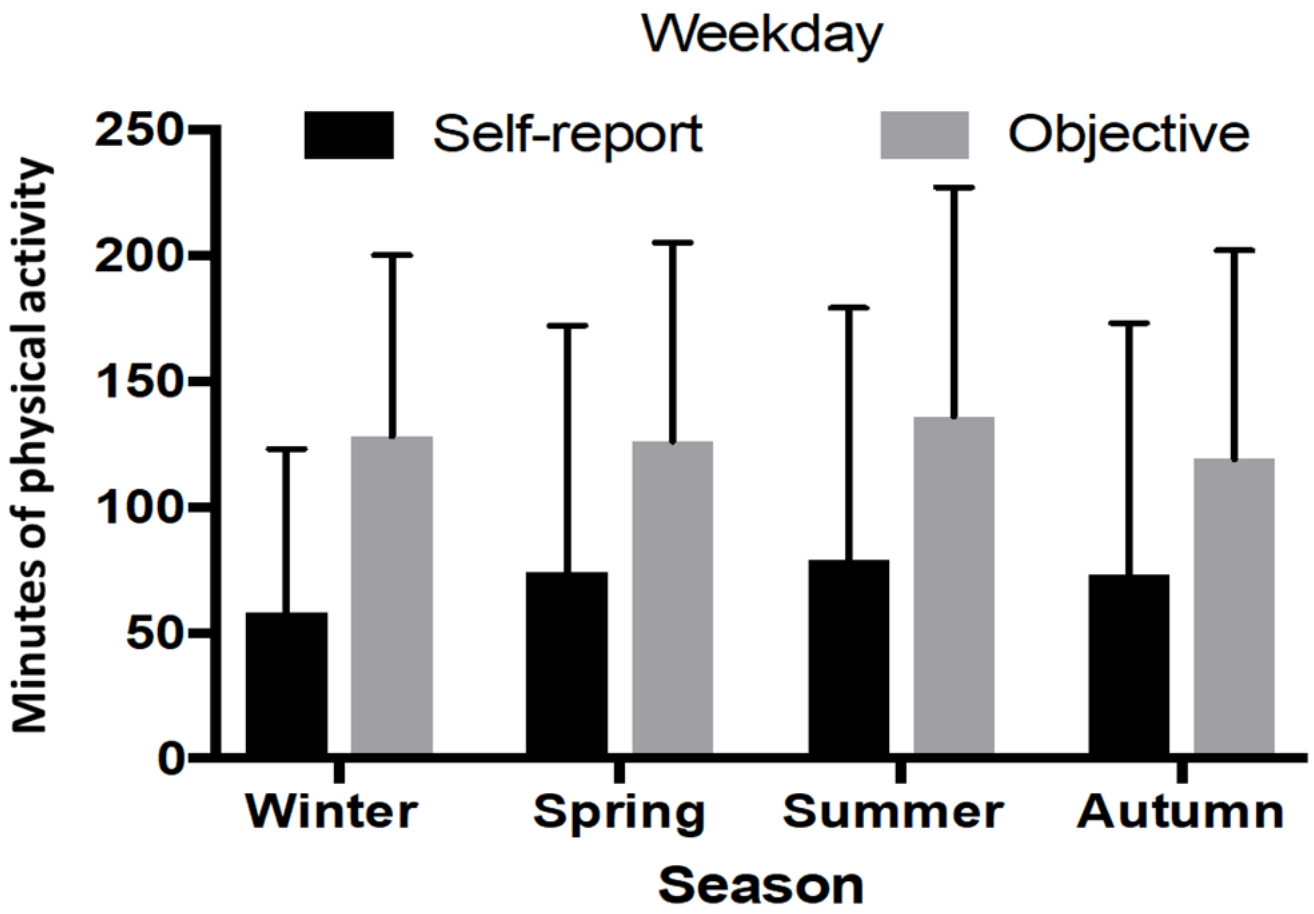



# Example: Mobility (McCool et al. 2021)



# Example: Physical Activity (Gilbert & Calderwood 2018)



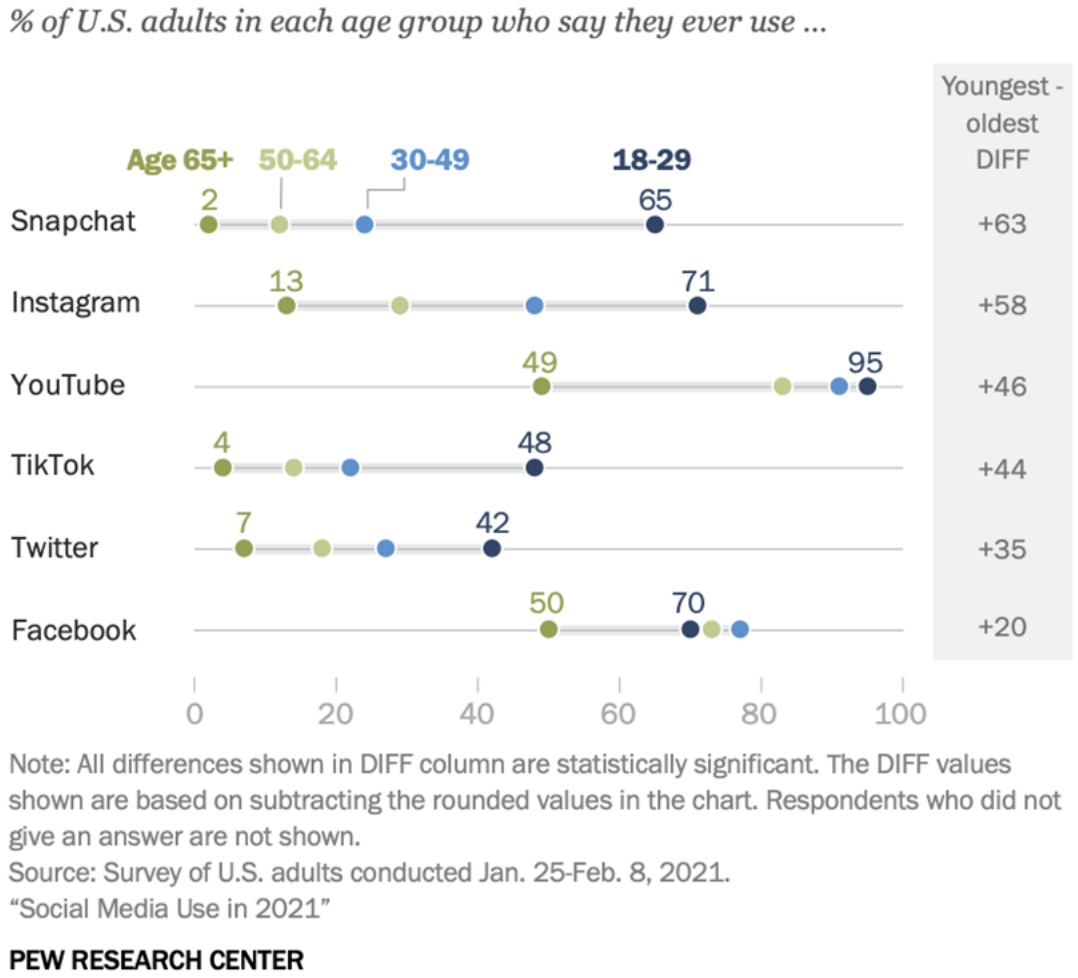

- UK Millennium Cohort Study
  - Children aged 14
  - Accelerometers worn for 2 days (weekday & weekend)

# Challenges of digital trace data

- Data access
  - Often data are proprietary
  - “Post-API age” (Freelon 2018)
    - No Facebook API since April 2018
    - Twitter API about to close (?)
    - Other platforms never had an API
  - Web scraping has legal and ethical limits

# Challenges of digital trace data

- Data access
- Representation
  - Who uses a certain platform?
  - Who is willing to share their data?




Auxier & Anderson (2021)

# Challenges of digital trace data

- Data access
- Representation
- Validity
  - Secondary data, not *designed* for research
  - How well are measures of digital behavior suited for measurement of other phenomena (e.g., attitudes)?

# Challenges of digital trace data

- Data access
- Representation
- Validity
- Unstructured data
  - e.g., classification of URLs and apps

```
com.sec.android.app.clockpackage,2018-06-19 08:20:16.361 +0200,NTP,2018-06-19 08:20:19.305 +0200,NTP,2945,0,0,Unknown,Unknown
com.google.android.music,2018-06-19 08:20:21.202 +0200,NTP,2018-06-19 08:21:18.689 +0200,NTP,57420,35279,63680,Google Play Music,Music & Audio
com.whatsapp,2018-06-19 14:59:07.227 +0200,NTP,2018-06-19 14:59:36.351 +0200,NTP,29076,252113,11862,WhatsApp Messenger,Communication
com.sec.android.app.launcher,2018-06-19 17:02:20.977 +0200,NTP,2018-06-19 17:02:29.893 +0200,NTP,8710,0,0,Unknown,Unknown
com.google.android.apps.maps,2018-06-19 17:02:52.774 +0200,NTP,2018-06-19 17:04:01.406 +0200,GPS,69035,117029,882507,Maps - Navigation & Transit,Travel & Lo
com.google.android.apps.maps,2018-06-19 17:30:26.277 +0200,GPS,2018-06-19 17:30:34.221 +0200,GPS,7941,1156,759,Maps - Navigation & Transit,Travel & Local
org.telegram.messenger,2018-06-19 19:48:37.730 +0200,NTP,2018-06-19 19:48:43.653 +0200,NTP,5920,362,245,Telegram,Communication
org.telegram.messenger,2018-06-19 20:17:42.599 +0200,NTP,2018-06-19 20:18:12.791 +0200,NTP,30191,2952,2446,Telegram,Communication
com.sec.android.app.launcher,2018-06-19 20:22:42.508 +0200,NTP,2018-06-19 20:22:46.876 +0200,NTP,4368,0,0,Unknown,Unknown
com.samsung.android.email.provider,2018-06-19 20:22:46.876 +0200,NTP,2018-06-19 20:22:53.042 +0200,NTP,5547,1310,5078,Unknown,Unknown
com.sec.android.app.launcher,2018-06-19 20:22:53.042 +0200,NTP,2018-06-19 20:22:56.010 +0200,NTP,2667,0,0,Unknown,Unknown
com.android.chrome,2018-06-19 20:22:56.010 +0200,NTP,2018-06-19 20:23:59.152 +0200,NTP,61363,300683,2813090,Chrome Browser - Google,Communication
com.android.chrome,2018-06-19 20:24:31.686 +0200,NTP,2018-06-19 20:26:31.206 +0200,NTP,119514,148480,1220440,Chrome Browser - Google,Communication
com.whatsapp,2018-06-19 22:36:01.409 +0200,NTP,2018-06-19 22:36:07.606 +0200,NTP,6142,323,356,WhatsApp Messenger,Communication
com.sec.android.app.clockpackage,2018-06-20 06:00:03.358 +0200,NTP,2018-06-20 06:05:00.330 +0200,NTP,296954,0,0,Unknown,Unknown
com.sec.android.app.clockpackage,2018-06-20 06:05:02.349 +0200,NTP,2018-06-20 06:06:02.618 +0200,NTP,60235,0,0,Unknown,Unknown
com.whatsapp,2018-06-20 06:06:02.618 +0200,NTP,2018-06-20 06:06:11.920 +0200,NTP,8939,0,0,WhatsApp Messenger,Communication
```

# Challenges of digital trace data

- Data access
- Representation
- Validity
- Unstructured data
- Privacy (Sloan et al. 2020)
  - Consent
  - Disclosure
  - Security
  - Archiving

# How can digital traces be collected in surveys?

# Types of digital trace data collection in surveys

- **Collaborate** with data collecting entity/platform
- Get **aggregate-level data** from platforms and link to your survey
- Ask respondents to **share individual account information** in survey
- Deploy **meters and apps**
- Ask for **data donation**

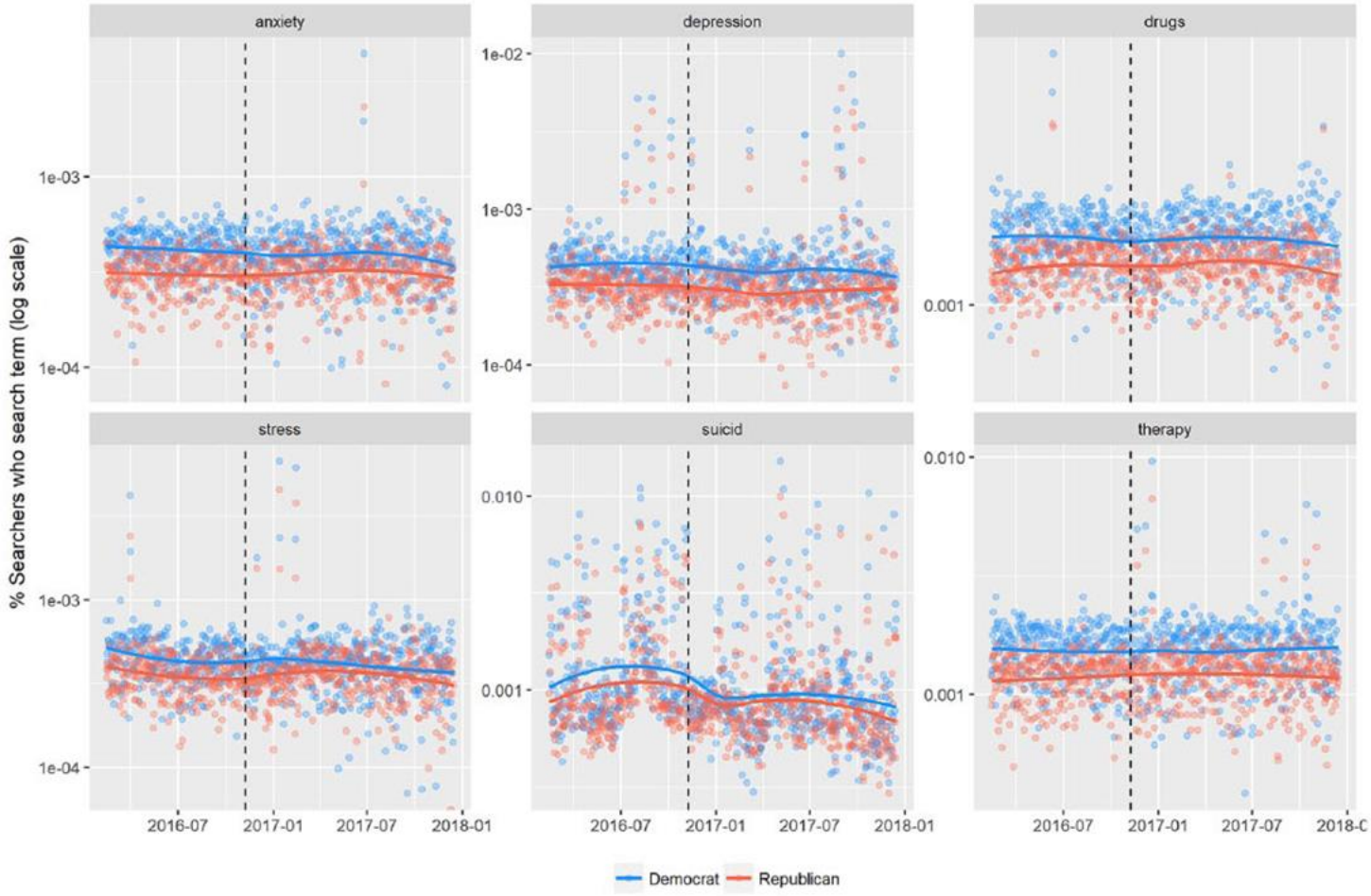
# Collaborate with data collecting entity/platforms

- Direct collaboration with data collecting companies through public-private partnership (Breuer et al. 2020)
  - Contractual agreement between research institution and company
  - “Embedded researcher” (e.g., consultant, research fellow, intern, part-time employee)

# Example: President Trump Stress Disorder

(Krumpenkin et al. 2019)

- Searches from 300,000 Bing users who completed survey on MSN.com
- Changes in mental health related searches
- No difference between searches from Dem and Rep
  - But mental health related searches from Spanish-speaking searchers increased



# Collaborate with data collecting entity

**Advantages**

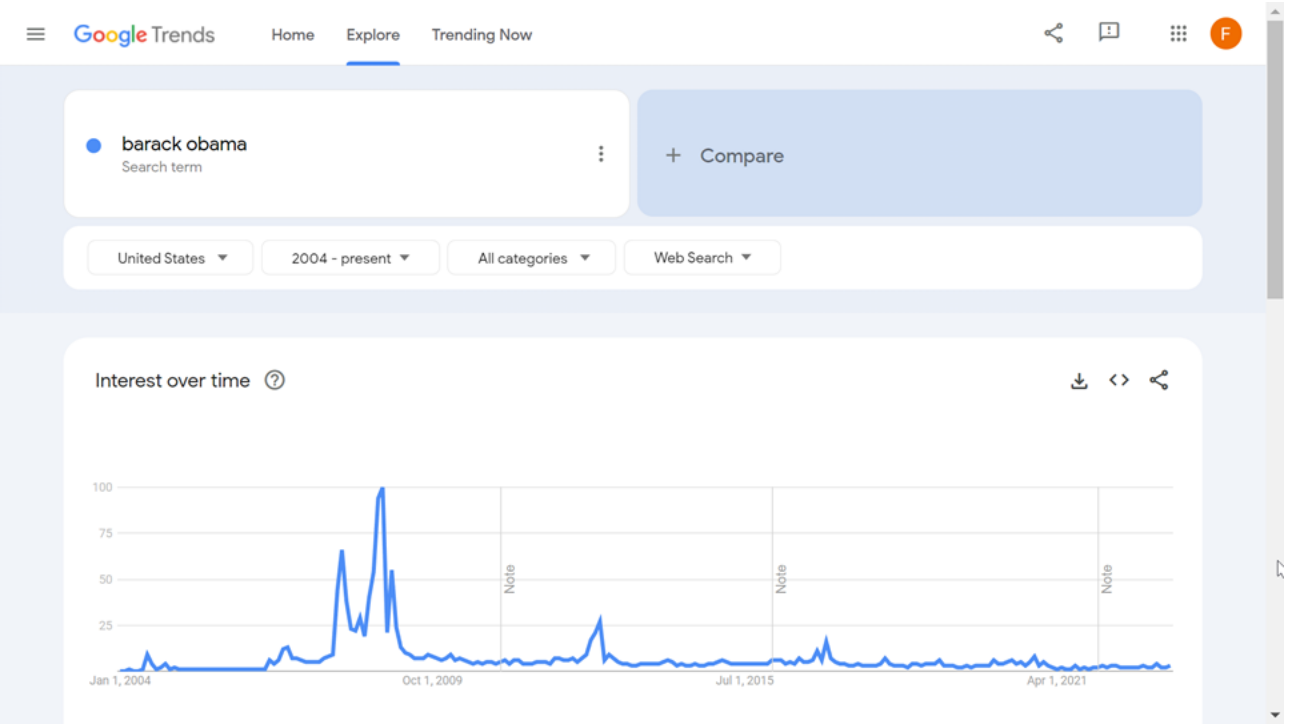
- Full data from one platform
- Combination of data from different sources within platform (e.g., traces, customer survey)

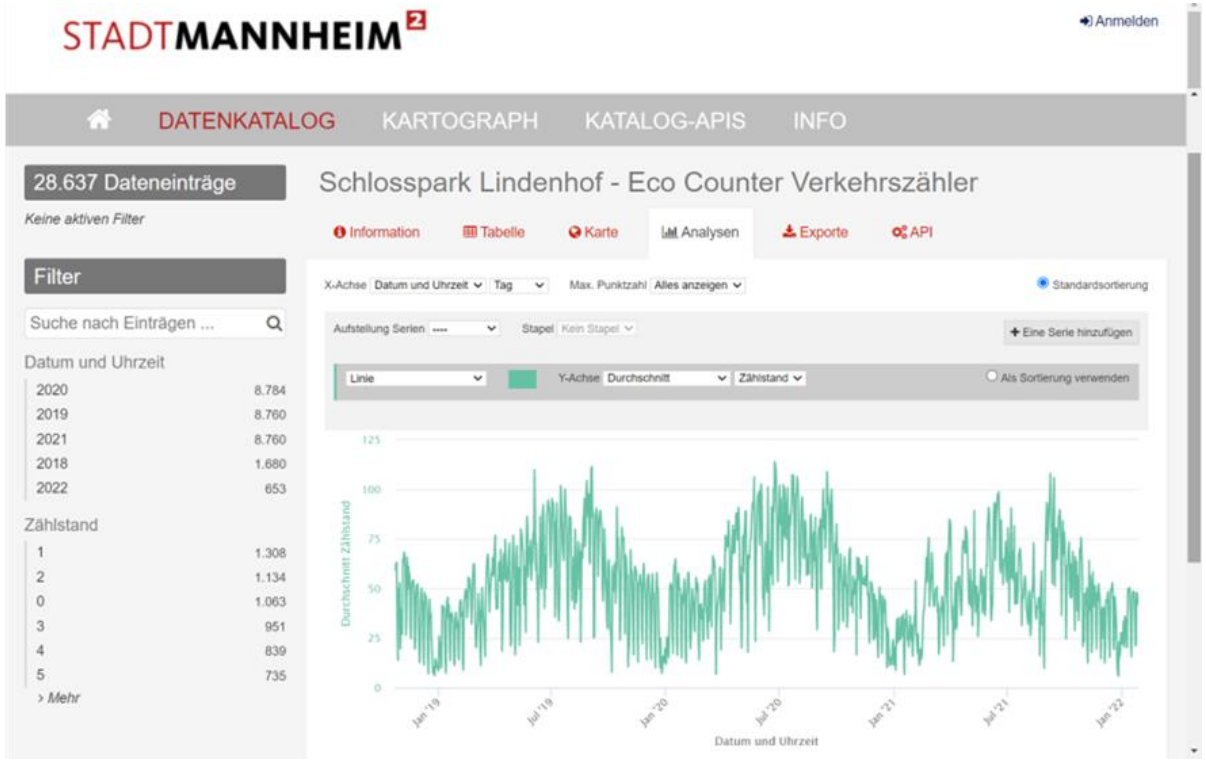
**Challenges**

- Restricted to users of one specific platform (i.e., coverage)
- Restrictions in data use and for publication
- Proliferating disparities in data access

# Aggregated measures of behavior from platforms

- Some platforms provide aggregate-level (e.g., country, state, region, metro) data that can be linked to individual survey data based on geography



# Example: Income inequality and racial bias

(Connor et al. 2019)

- State-level income inequality data used to predict combination of state-level Google Trends of Internet searches containing racial slurs and cross-sectional data from 1.5m U.S. adults on explicit race bias (Race Implicit Association Test, preference for Whites compared with Blacks, feeling thermometers)
- Hierarchical linear modeling (HLM) used to account for nested structure of different data sources
- Significant positive within-state association between income inequality and Whites' explicit racial bias

# Aggregated measures of behavior from platforms

**Advantages**

- “Real-time” data about behaviors (“nowcasting”)
- Easy access
- Freely available

**Challenges**

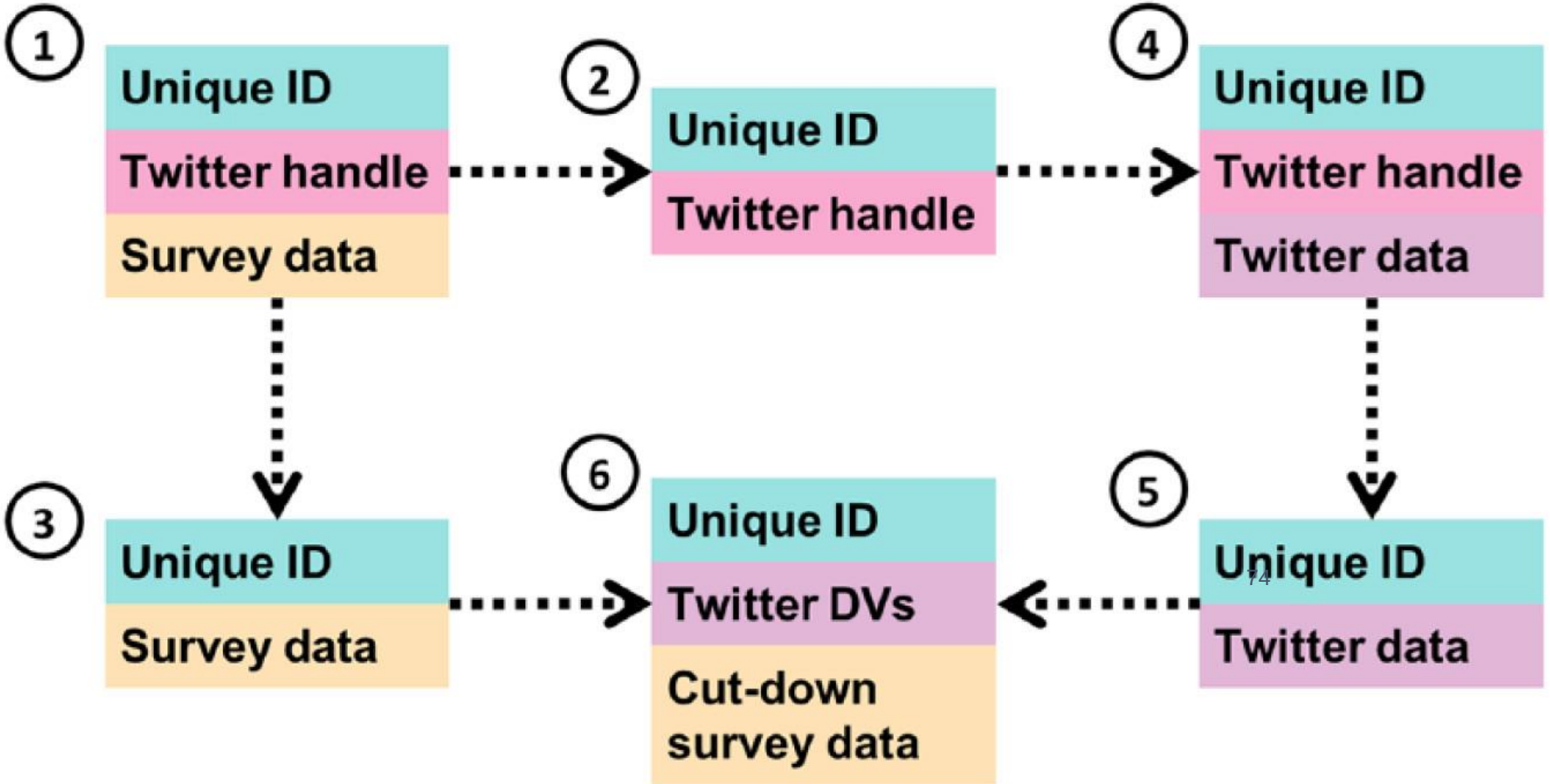
- Only aggregated information
- No covariates

# Ask respondents to share account name

- “Classic” record linkage idea: respondents consent to linkage of survey data with data on a platform via (unique) identifier
  - Twitter handle
  - Facebook friendship request
  - Consent to fitbit data via fitabase
  - …

# Example: Linking Twitter and survey data

(Sloan et al. 2020)



# Example: Linking Facebook to survey data

(Schröder et al. in preparation)

- 2,457 Facebook users in German online access panel asked to message assigned ID to designated study account
- Linking of survey data with publicly available Facebook data

| | Percent of accounts |
|---|---|
| Gender | 87 |
| Date of birth[1] | 13 |
| Current city | 38 |
| State | 9 |
| Hometown | 38 |
| School | 26 |
| University | 12 |
| Job | 29 |
| Relationship status | 24 |
| Email | 0 |
| Friends | 31 |
| Photos | 92 |
| Photo album with profile photos | 82 |
| Photo album with cover photos | 80 |
| Photo album with favorite photos | 21 |
| N | 215 |

[1] Including partial information (only year / only month and day)

# Ask respondents to share account name

**Advantages**

- Individual linkage of survey data and digital traces

**Challenges**

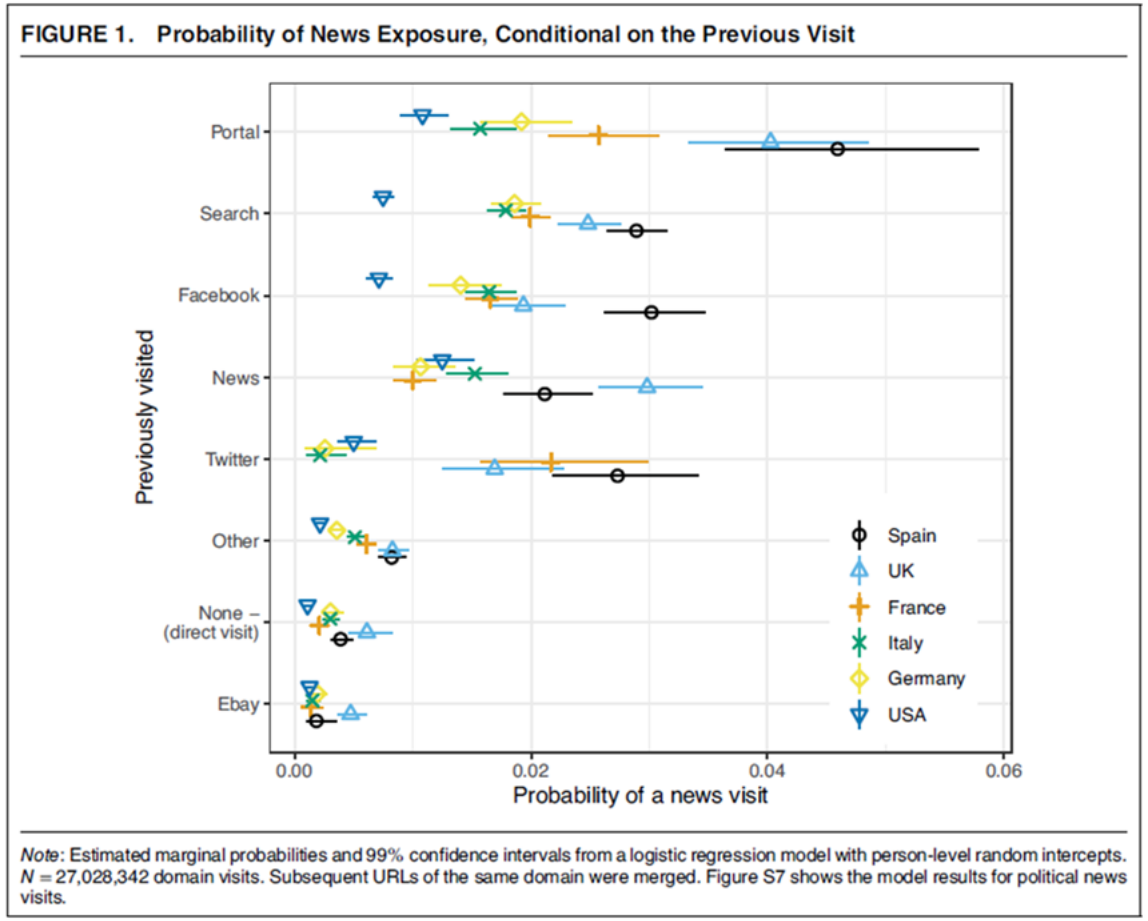
- Ethical data linkage and consent
- Process varies vastly by platform
- Willingness to provide account name information

# Meters and research apps

- Users…
  - …install plug-in to web browser (aka “meter”) that continuously tracks information on web browsing (URLs, HTML code, screen scraping)
  - …download app to smartphone that continuously logs usage behavior, (native) mobile browsing, and sensor readings
- Allows tracking of individual behavior over longer periods of time
- Various commercial and non-commercial tools available
  - See “Resources” slide
- Some market research companies maintain “metered panels”
  - See “Resources” slide

# Example: Online intermediaries and news exposure (Stier et al. 2021)

- Combining web browsing histories and survey responses of >7,000 participants from six countries
- Findings: Substantively large effects of using intermediaries (e.g., Facebook, Twitter, search engine) before visit to news outlet



# Meters and research apps

**Advantages**

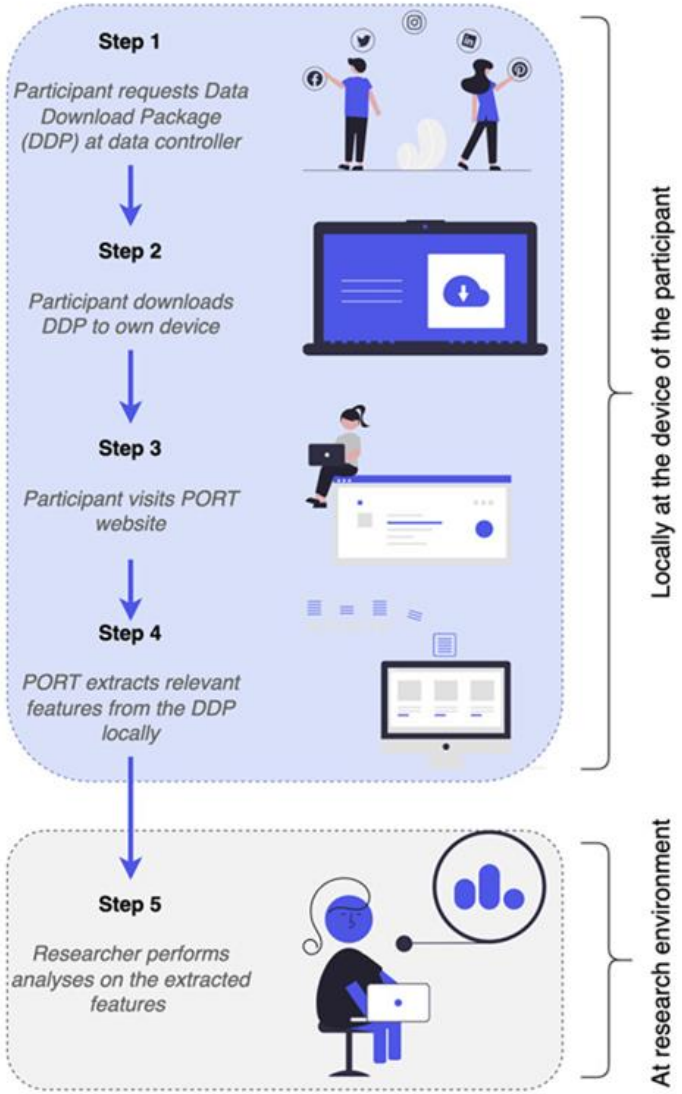
- Direct observation of media consumption
- Unobtrusive, continuous measurement

**Challenges**

- Coverage & nonparticipation
  - Mainly from volunteer panels
- Missing data
  - Due to device settings, turning of meter/app, using multiple devices/browsers
- Construct validity (?)
- Reactivity (?)

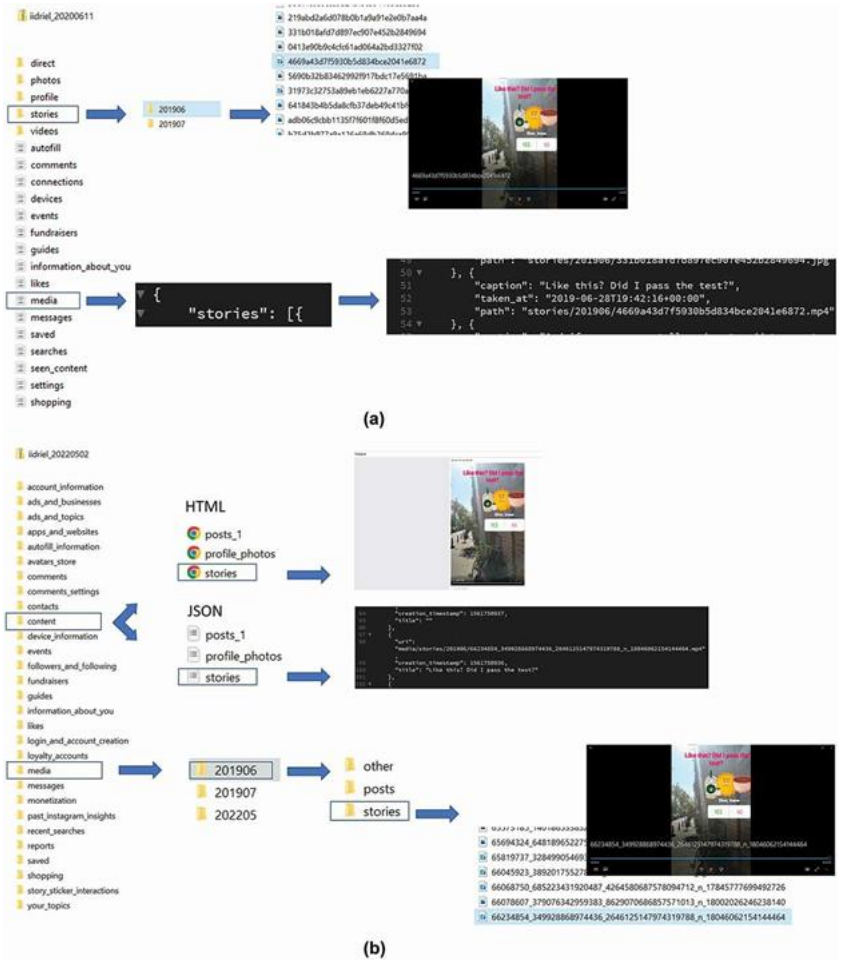
# Data donation (DD)

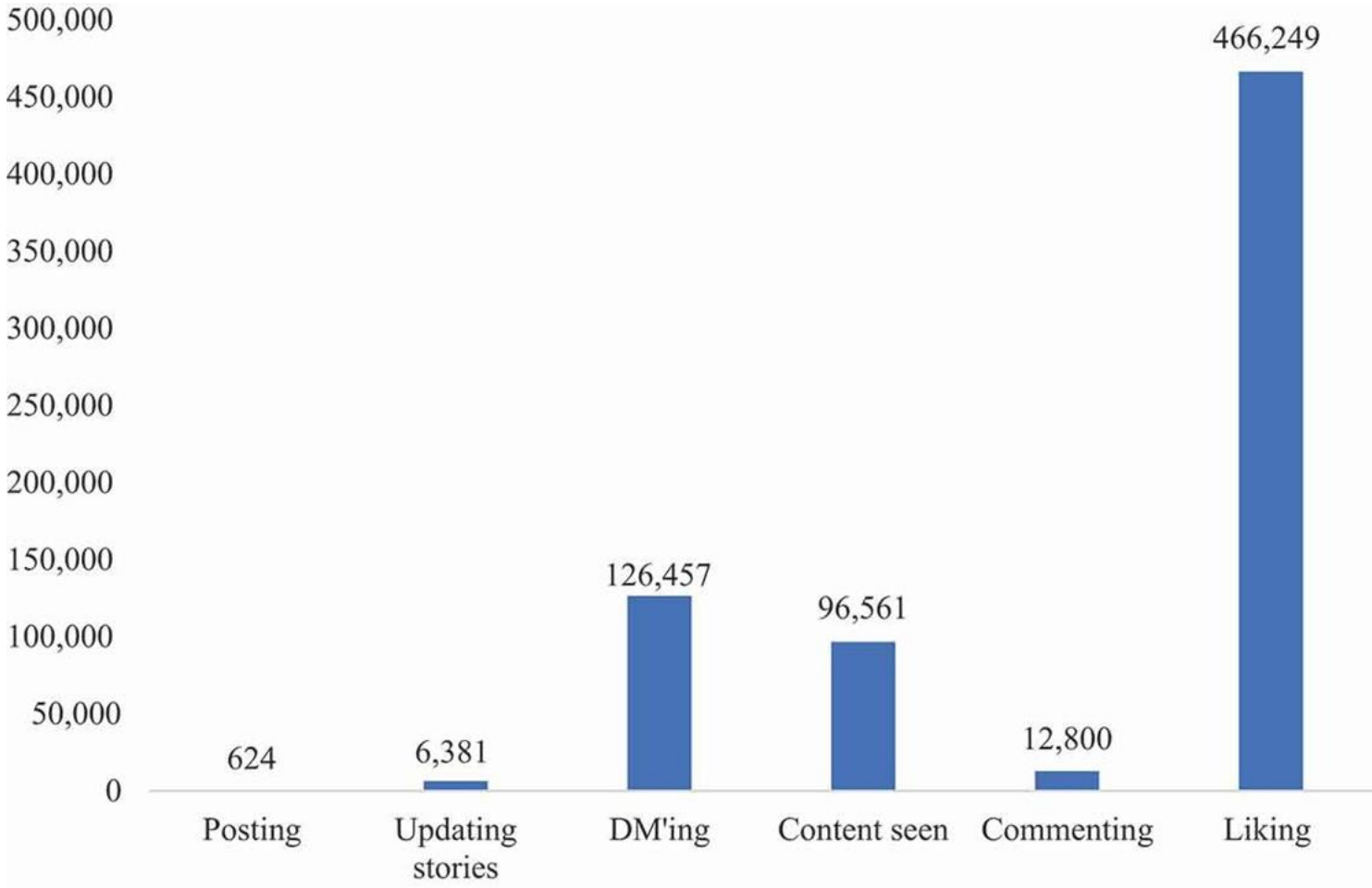
- Takes advantage of GDPR Articles 15 (Right of access by the data subject) and 20 (Right to data portability)
  - Receive personal data in structured, commonly used, and machine-readable format ("Data Download Package"; DDP)
  - Transmit data to another data controller
- Donation of full DDP via direct sharing
  - e.g., ftp server, email
- Privacy-preserving data donation platforms
  - Port (Boeschoten et al. 2022)
  - OSD2F (Araujo et al. 2022)




Boeschoten et al. (2022)

# Example: Instagram DDP donation (van Driel et al. 2022)



# 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

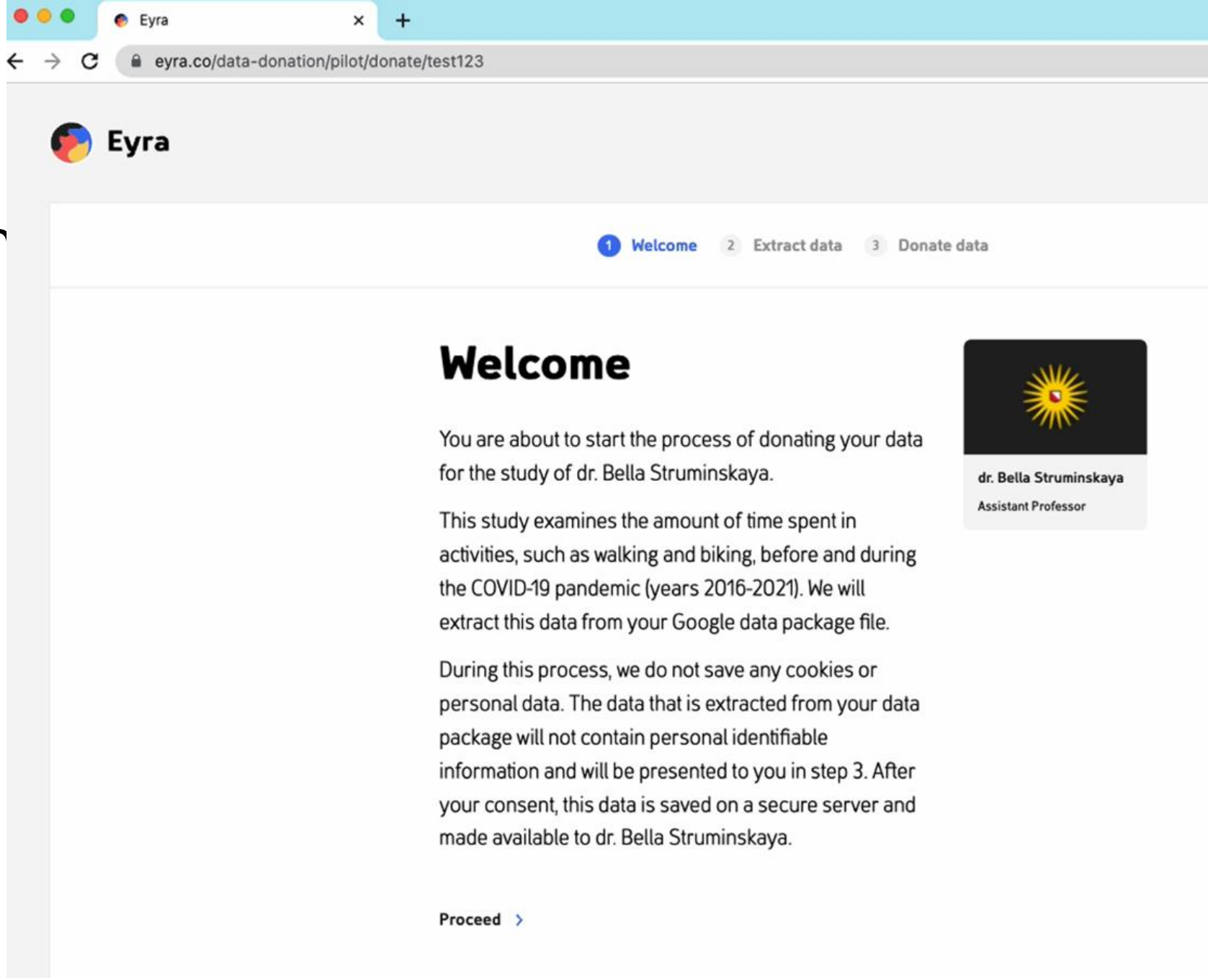
**Challenges**

- 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

# Demo: Data Donation

# Demo: PORT (Google Location History)

[Video]



# Motivation to donate & reasons against DD

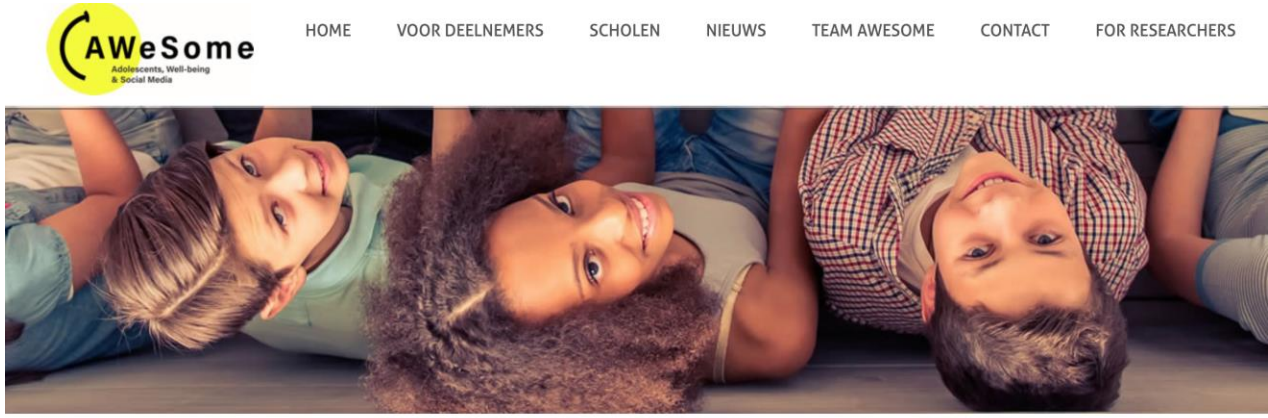
- 54% likely to donate for health research, 31% not likely (Skatova & Goulding 2019)
- 40% of fitness tracker owners willing to donate (Toepoel et al. 2021)
- 30% share Facebook data, 24-40% Twitter data, 60% Spotify (Silber et al. 2021)

+ Prosocial behavior (Skatova & Goulding 2019)
+ Insight into own results / quantified self (Bietz et al. 2019)
- Not gaining direct benefits from data donation (Skatova & Goulding 2019)
- Need to know the consequences of donation (Skatova & Goulding 2019)

- Similar to reasons to share/not share sensor & app data
- Privacy concerns may play a role (e.g., for apps/sensors: Keusch et al. 2019; Struminskaya et al. 2020; Struminskaya et al. 2021)

# Selectivity in donation of social media data

- Project AWeSome (Adolescents, Well-being, and Social Media) by University of Amsterdam
- Topics: social media use, well-being, social relationships, self-regulation
- Teenagers 13-15 yo in NL, recruited f2f at school, parental consent provided (N = 388)
- 80% have Instagram account(s)
- 32% donated Instagram data (raw)


# Privacy and loyalty are key

**Sociability**

- Social comparison
- # good friends
- Friendship quality
- Parental phone rules
- Parental knowledge
- Adolescent dis-closure & secrecy*** (AME=.10)

**Psychological chars**

- Affective well-being
- Cognitive well-being
- Positive affect
- Negative affect
- Self-esteem* (AME=-.08)
- Loneliness
- Self-regulation*** (AME=1.23)

**Social media, SP use**

- # accounts
- Sphone self-monitor.
- Sphone type (iPhone)
- # followers
- Importance followers
- # likes on post
- Eval. of reactions
- Eval. # of reactions
- Importance positive reactions

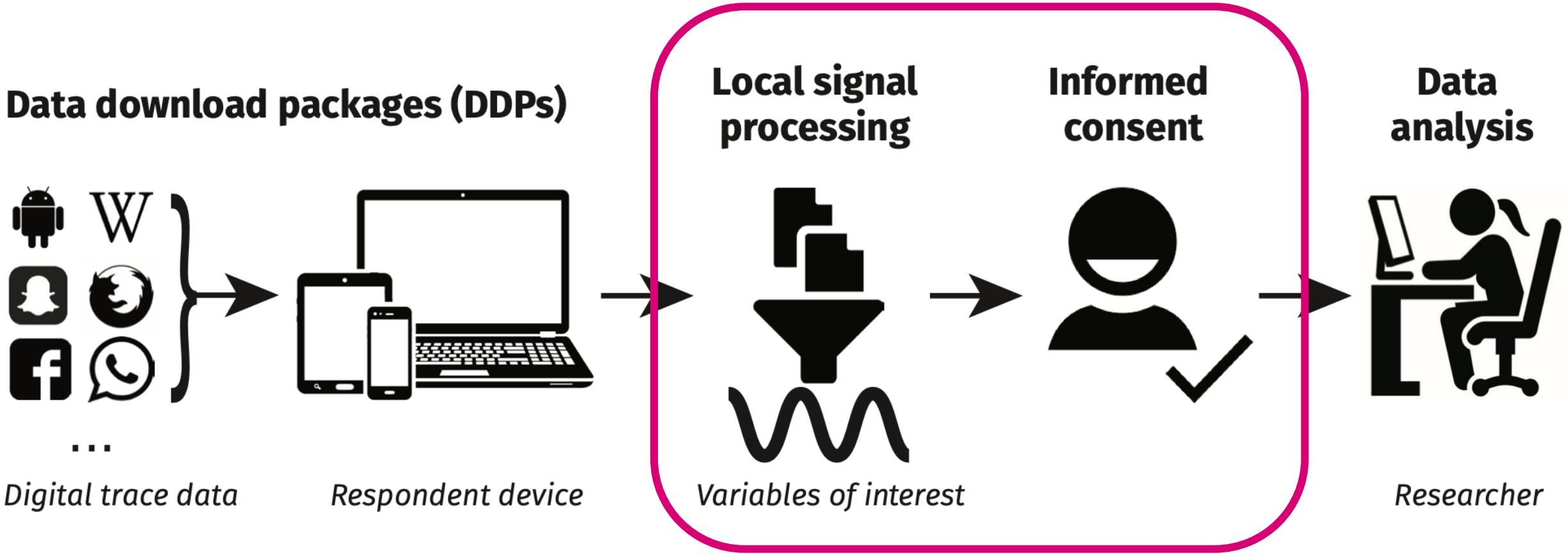
**Study design**

- # completed ESM1
- # completed ESM2*** (AME=.004)
- # completed surveys

Logistic regressions donated (1) vs. did not donate (0), models include sex, grade, education, N varies 146-316.

# Giving agency to participants

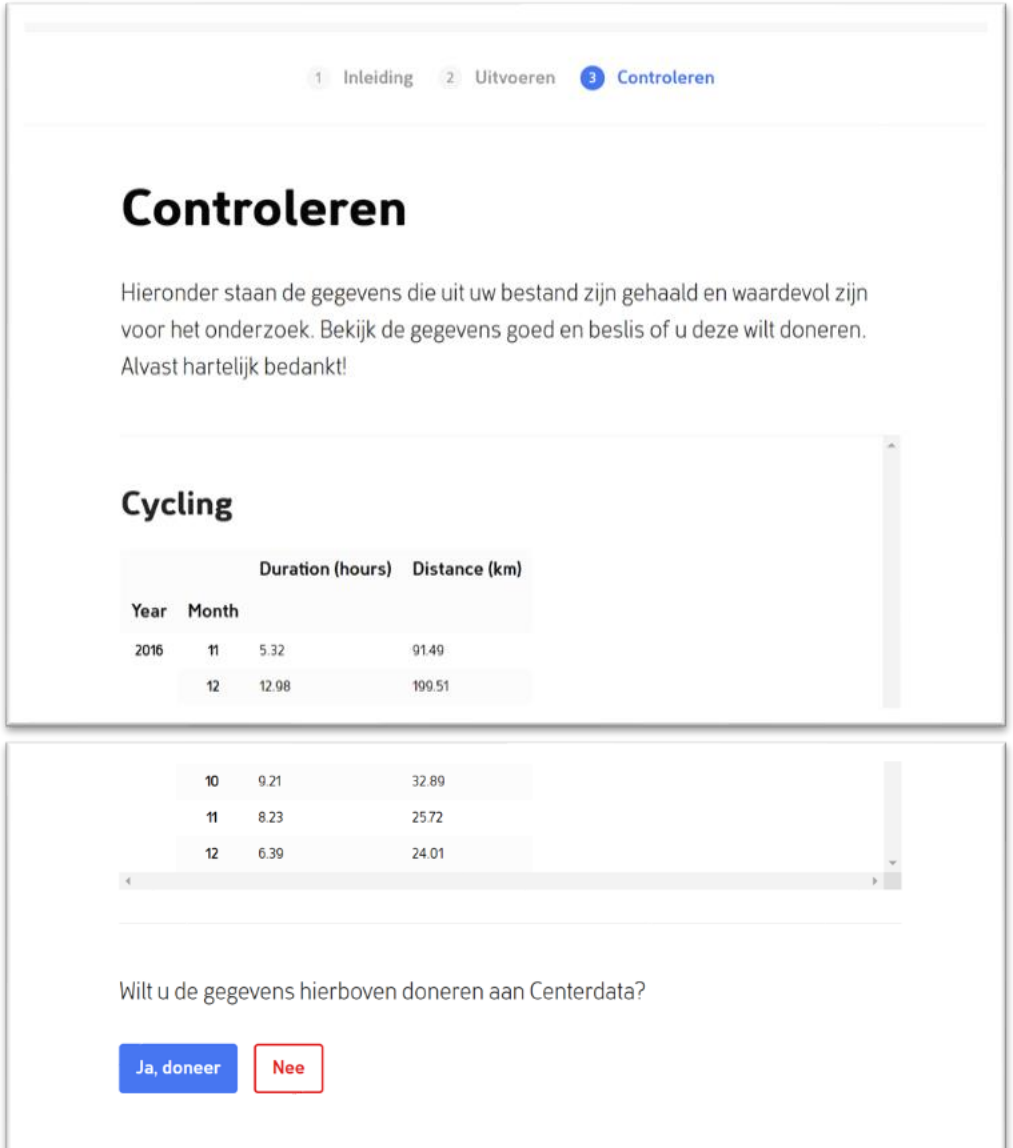
# Privacy-preserving Data Donation Workflow




Boeschoten et al. 2020

# Google Location History Data Donation

- Study in CentERpanel July-August 2022
- N=1035 (75% AAPOR RR1)
- Integration of data donation (PORT)
- Willing = 30%, 144 donated (14%)
- Research questions:
  - Incentive amount (5€ vs. 10€)
  - Visualization prior to request
  - Understanding of consent request (50% asked TRUE/FALSE questions)
  - Nonparticipation bias



# Understanding the consent request

| Statements asked to respondents | Correct % | Incorrect % | Don't know % |
|---|---|---|---|
| You are asked to download information from Google. TRUE | 48.8 | 19.8 | 31.4 |
| The software implemented in the survey will extract the information on the number of hours you cycle, walk, take public transport, travel by car. TRUE | 62.3 | 6.1 | 31.2 |
| Information on all the locations you visited will be shared with Centerdata. FALSE | 39.2 | 31.4 | 29.4 |
| Google collects information on location about everyone. FALSE | 24.8 | 46.6 | 28.5 |
| From the data you will provide, the information can be traced back to you. FALSE | 45.3 | 22.2 | 32.5 |
| You will be able to inspect the data before sending it to Centerdata. TRUE | 59.0 | 7.8 | 33.1 |
| It is impossible to identify you as an individual from the data that you provide. TRUE | 43.4 | 19.6 | 37.0 |

# Incentive & visualization

- No difference in incentives
  - 5€: willing to donate 32% (n=147)
  - 10€: willing to donate 34% (n=159)
  - Chi2(1) = 0.32, p=.574
  - Donated: 48% vs. 46% (Chi2(1) = 0.17, p=.676)

- No difference by showing how data looks like
  - Visualized: willing to donate 34% (n=159)
  - Not visualized: willing to donate 32% (n=147)
  - Chi2(1) = 0.56, p=.456
  - Donated: 46% vs. 48% (Chi2(1) = 0.17, p=.676)

De reisbewegingen kunt u in deze vragenlijst delen met Centerdata. Het is goed om te weten dat locaties die u hebt bezocht niet uit het pakketje worden gehaald en dus ook niet met Centerdata worden gedeeld. Er wordt **alleen** informatie gedeeld **hoe** u zich heeft verplaatst en **hoeveel tijd** u hieraan hebt besteed per maand en jaar.

[*if condition = 1* Een voorbeeld van hoe deze informatie eruitziet ziet u hieronder:

**Cycling**

| Year | Month | Duration (hours) | Distance (km) |
|---|---|---|---|
| 2021 | 8 | 1.14 | 6.32 |

**In Bus**

| Year | Month | Duration (hours) | Distance (km) |
|---|---|---|---|
| 2021 | 8 | 1.97 | 28.23 |

**In Passenger Vehicle**

| Year | Month | Duration (hours) | Distance (km) |
|---|---|---|---|
| 2021 | 8 | 23.31 | 375.84 |

# Understanding the consent request

- 5.5% had everything correct
- Mean correct: 3.23, median = 4
- People with more correct answers more likely to be willing & to donate:
  - 4.54 correct statements for willing
  - 2.56 correct statements for non-willing
  - OR = 1.572, p <.001

  - 5.33 correct statements for donated
  - 3.94 correct statements for not donated
  - OR = 1.795, p <.001

# Who is more willing to donate?

| Characteristic | Odds Ratio | SE | p-value |
|---|---|---|---|
| gender (male) | 1.759 | .293 | .001 |
| age | .992 | .006 | .208 |
| middle education | 1.794 | .412 | .011 |
| high education | 1.786 | .402 | .010 |
| urban | 1.076 | .063 | .212 |
| privacy concern | .966 | .056 | .554 |
| trust Google | 1.221 | .119 | .041 |
| can download | .825 | .129 | .217 |
| smartphone skill | .967 | .097 | .737 |
| no. smartphone activities | 1.128 | .034 | .000 |
| constant | .078 | .064 | .002 |

Logistic regression, n=867, Pseudo-R2=.068

# Who is more likely to donate?

| Characteristic | Odds Ratio | SE | p-value |
|---|---|---|---|
| gender (male) | 1.825 | .512 | .032 |
| age | .971 | .011 | .007 |
| middle education | 1.615 | .660 | .240 |
| high education | 1.509 | .592 | .294 |
| urban | 1.075 | .107 | .465 |
| privacy concern | .969 | .090 | .730 |
| trust Google | .890 | .143 | .469 |
| can download | .724 | .205 | .255 |
| smartphone skill | 1.051 | .172 | .760 |
| no. smartphone activities | .998 | .050 | .978 |
| constant | 4.125 | 5.928 | .324 |

Logistic regression, n=280, Pseudo-R2=.057

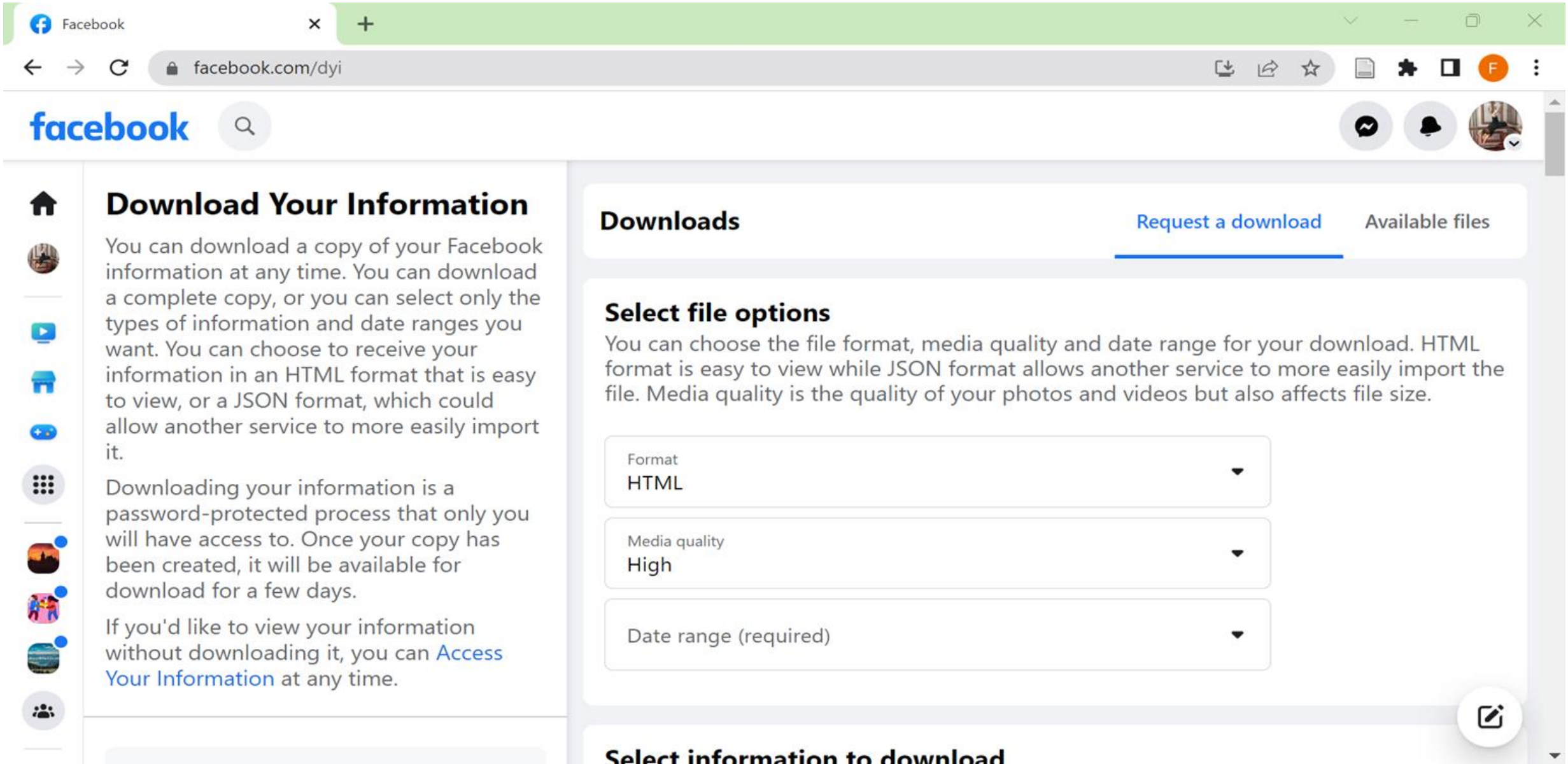
# Summary

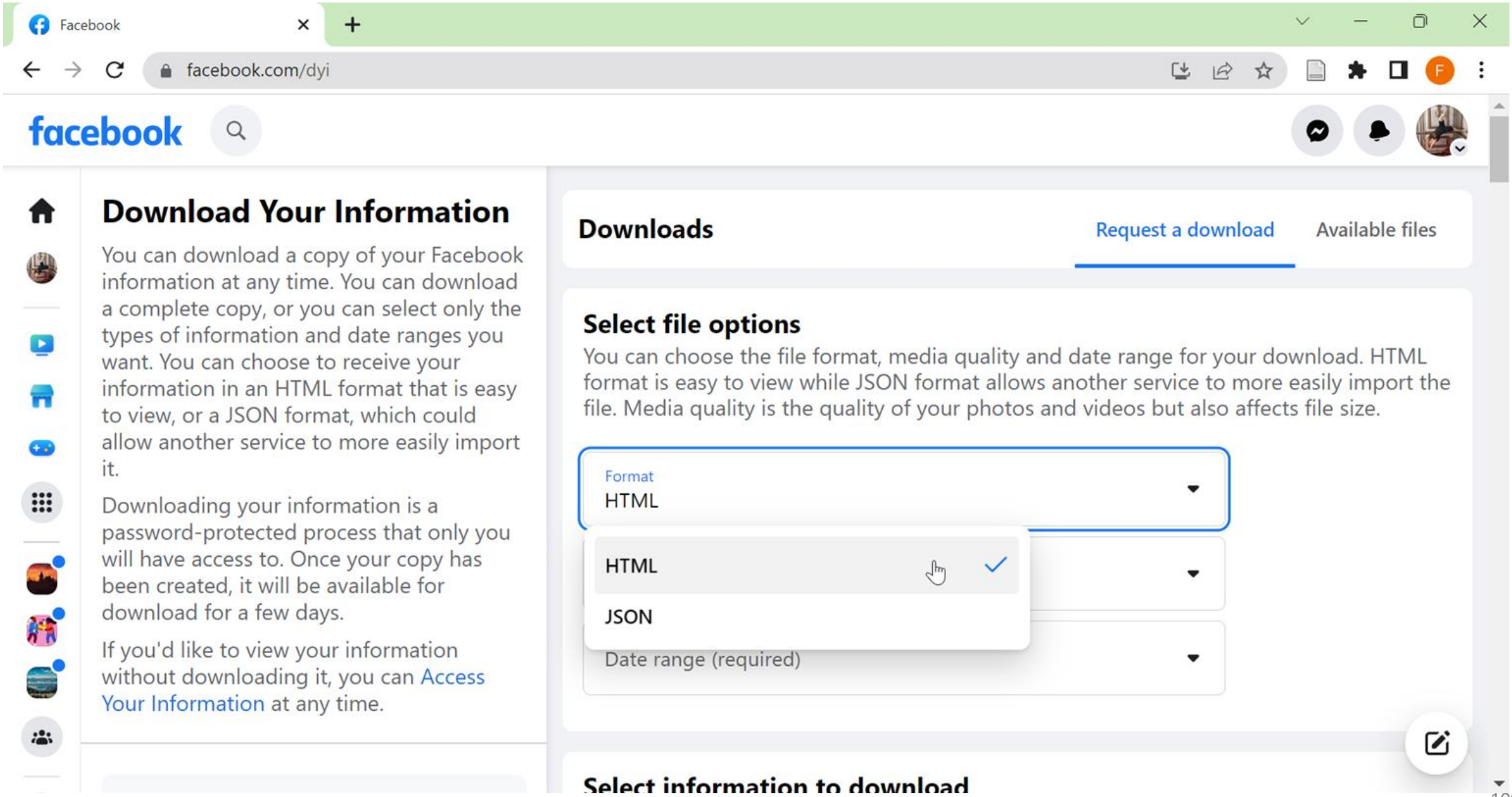
- Mechanisms of willingness still much unknown
- People might have a stable pre-conception (no effect of incentives, visualization), but: panel used to innovations → study different populations, different data types
- Understanding of consent request is not high → further experiments  text length, content, presentation, cognitive interviewing
- About half of the participants willing to donate does not make a donation → dig deeper into the process
- Quality, validity of donations → passive vs. self-report
- Level of aggregation → raw data vs. aggregated

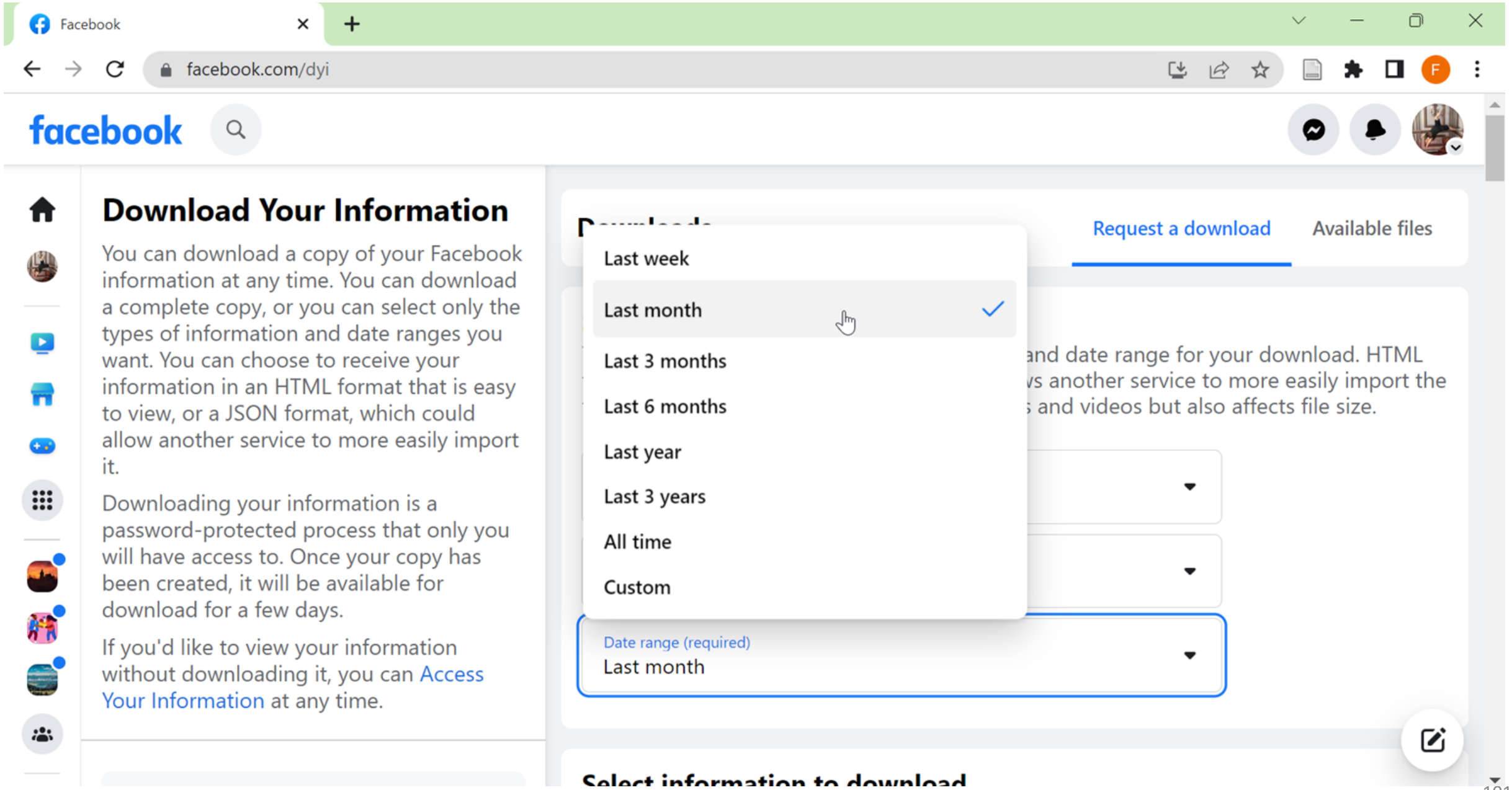
# Exercise: Request your DDP

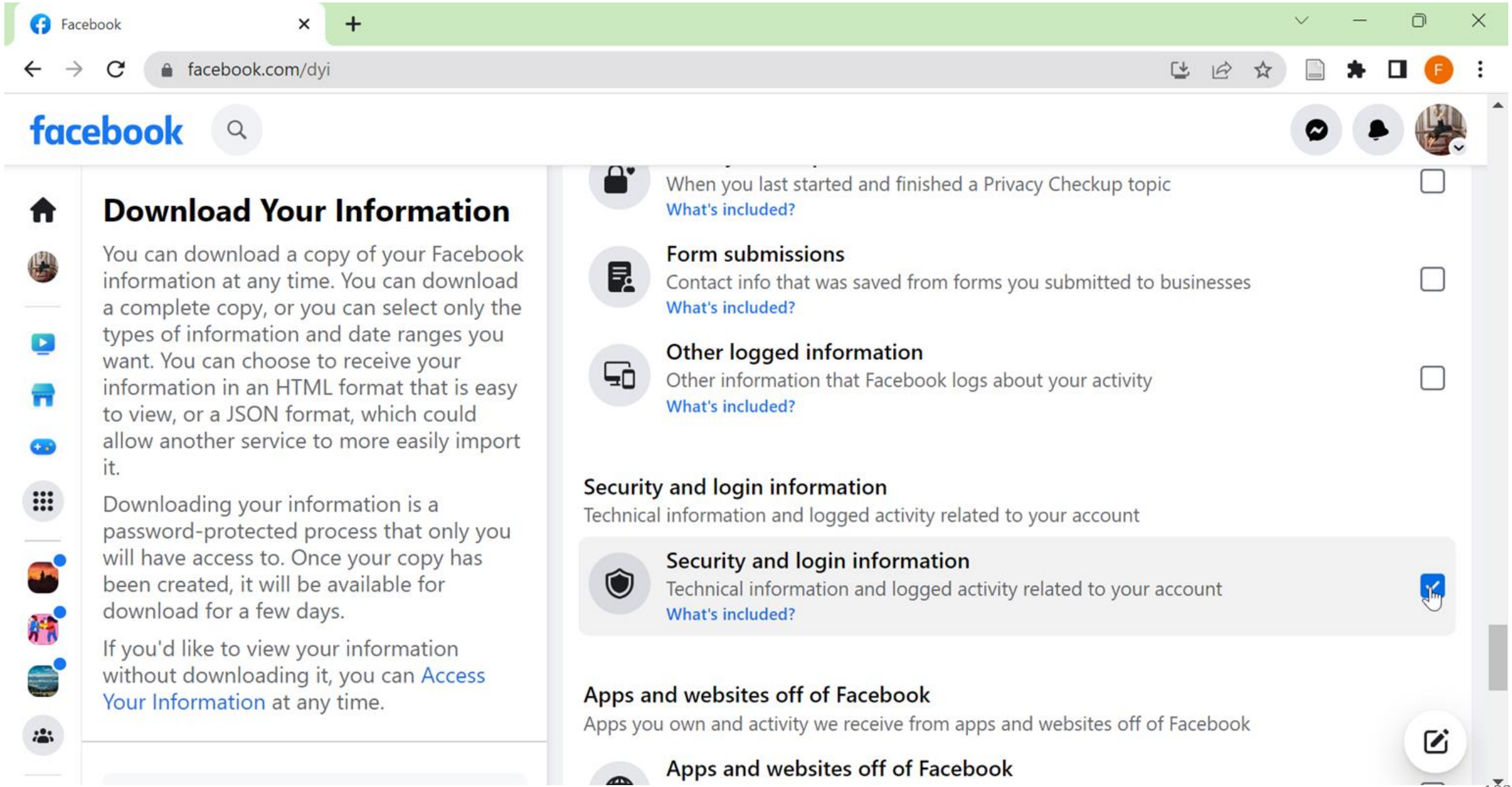
# Download DDP from Facebook

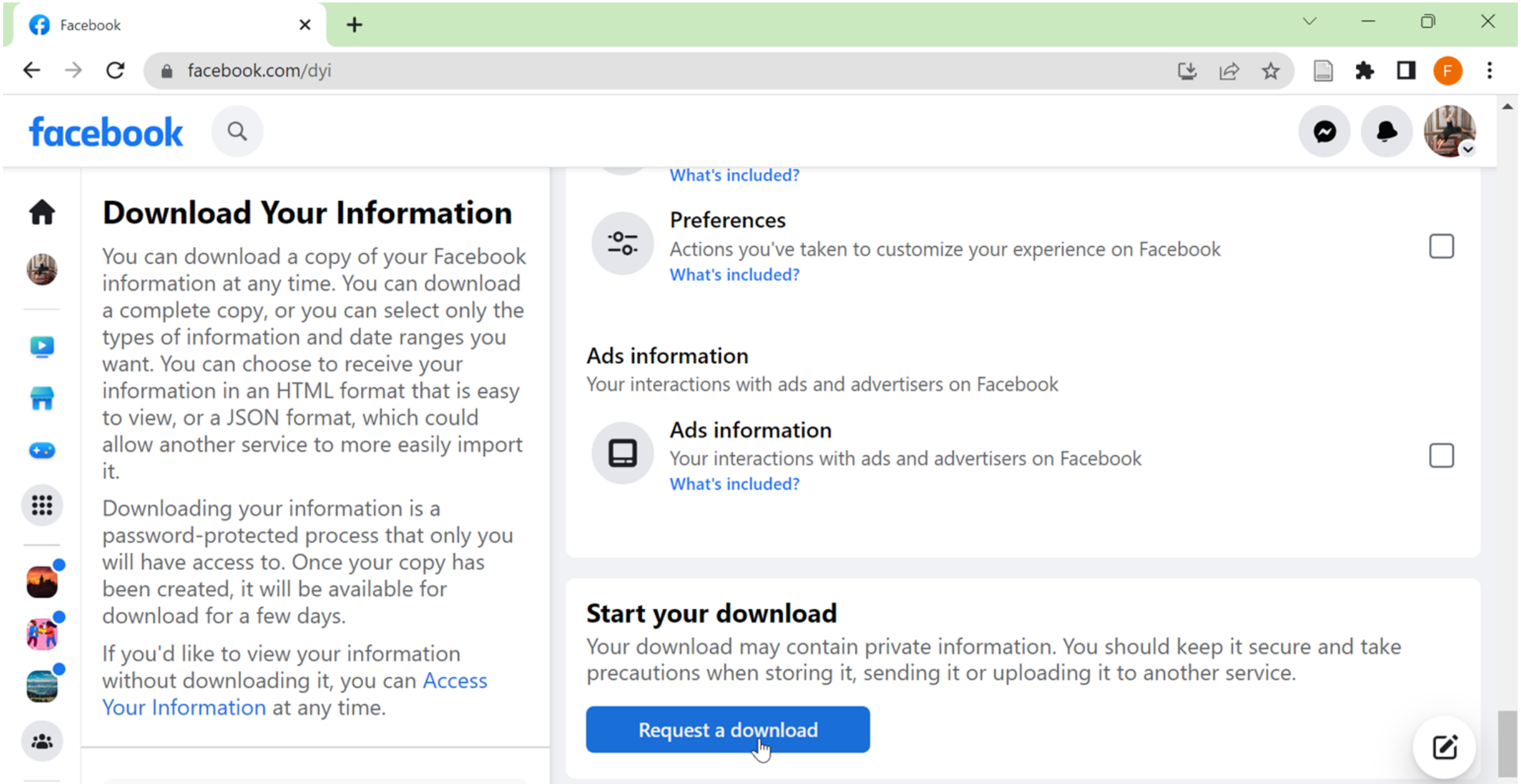
- Step 1: Go to https://fb.com/dyi
- Step 2: Specify data format
- Step 3: Set date
- Step 4: Select data download packages
- Step 5: Request download
- Step 6: Verify with password
- Step 7: Download zip file

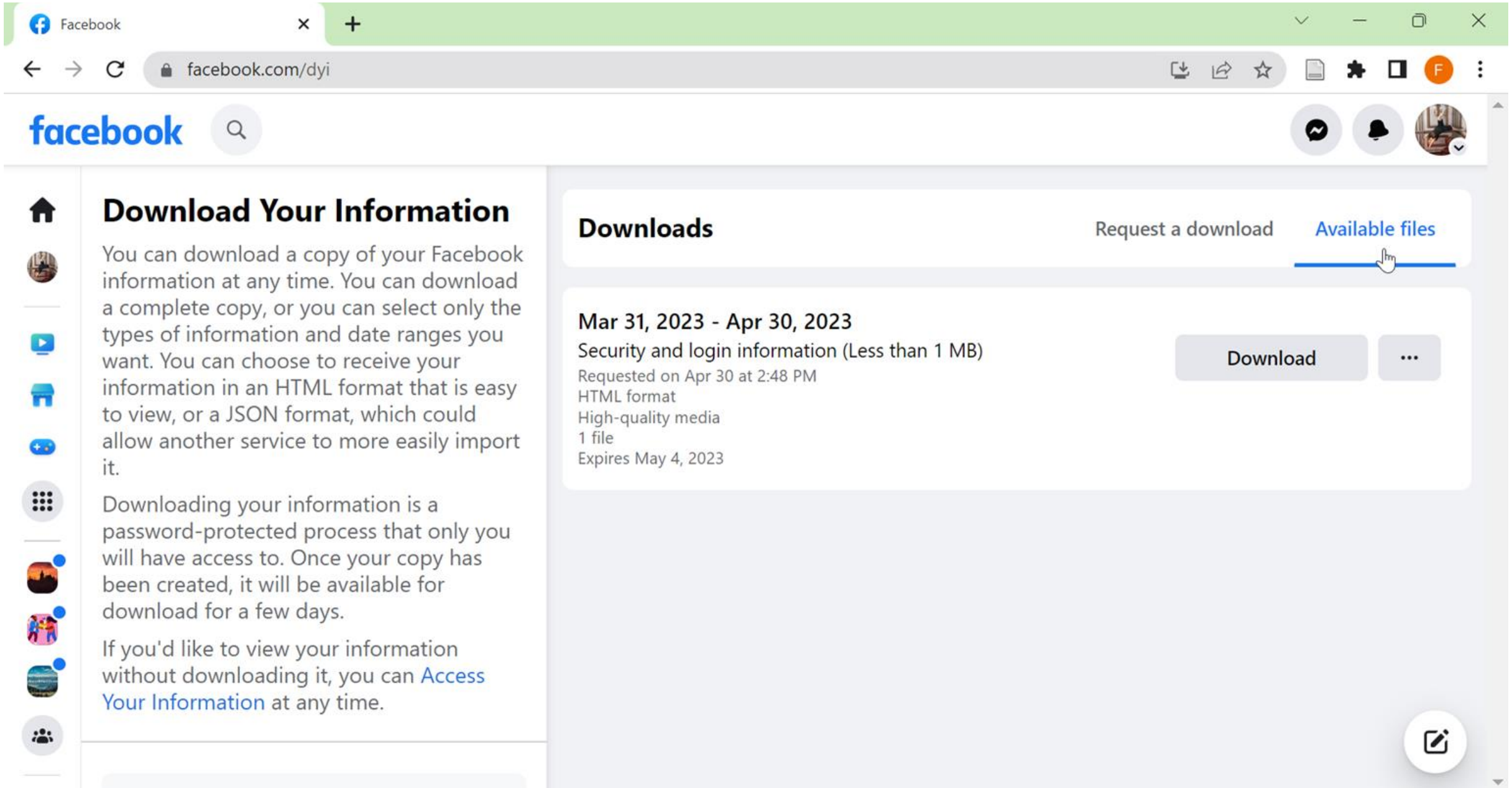

Facebook
facebook.com/dyi
facebook
Download Your Information
You can download a copy of your Facebook information at any time. You can download a complete copy, or you can select only the types of information and date ranges you want. You can choose to receive your information in an HTML format that is easy to view, or a JSON format, which could allow another service to more easily import it.
Downloading your information is a password-protected process that only you will have access to. Once your copy has been created, it will be available for download for a few days.
If you'd like to view your information without downloading it, you can Access Your Information at any time.
Downloads
Request a download
Available files
Select file options
You can choose the file format, media quality and date range for your download. HTML format is easy to view while JSON format allows another service to more easily import the file. Media quality is the quality of your photos and videos but also affects file size.
Format
HTML
Media quality
High
Date range (required)
Select information to download

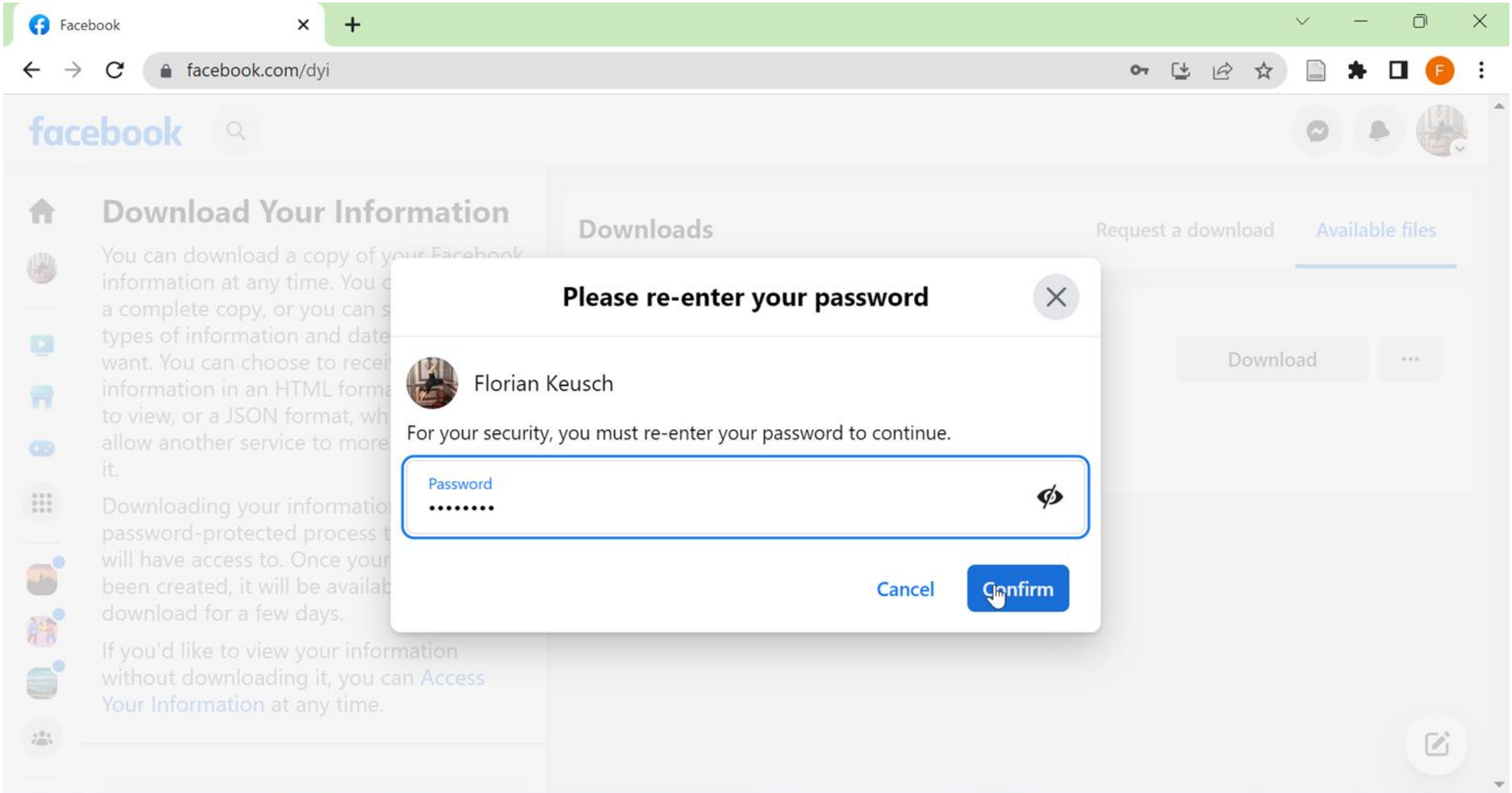

Facebook
facebook.com/dyi
facebook
Download Your Information
You can download a copy of your Facebook information at any time. You can download a complete copy, or you can select only the types of information and date ranges you want. You can choose to receive your information in an HTML format that is easy to view, or a JSON format, which could allow another service to more easily import it.
Downloading your information is a password-protected process that only you will have access to. Once your copy has been created, it will be available for download for a few days.
If you'd like to view your information without downloading it, you can Access Your Information at any time.
Downloads
Request a download
Available files
Select file options
You can choose the file format, media quality and date range for your download. HTML format is easy to view while JSON format allows another service to more easily import the file. Media quality is the quality of your photos and videos but also affects file size.
Format
HTML
HTML
JSON
Date range (required)
Select information to download

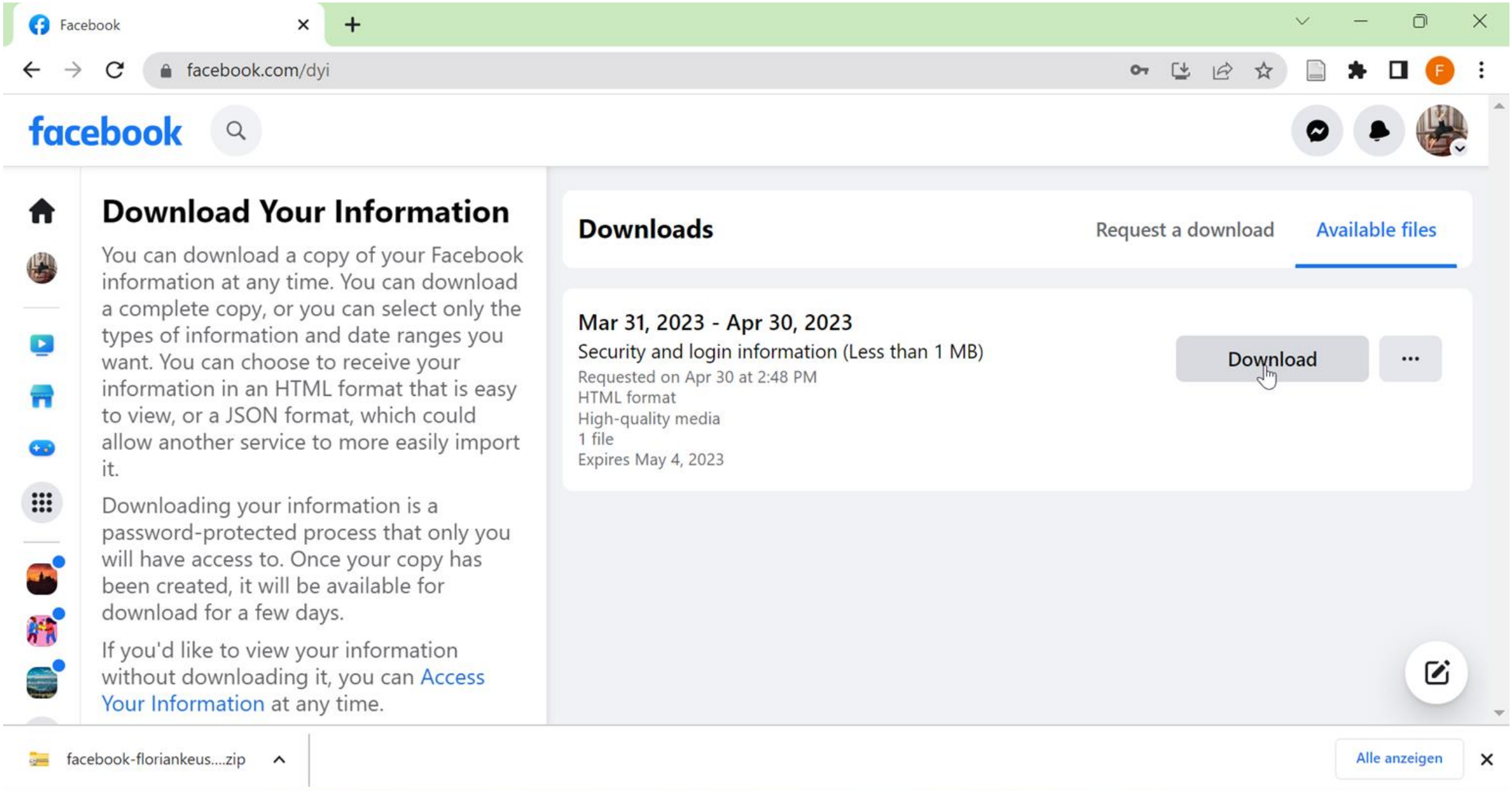

Facebook
facebook.com/dyi
facebook
Download Your Information
You can download a copy of your Facebook information at any time. You can download a complete copy, or you can select only the types of information and date ranges you want. You can choose to receive your information in an HTML format that is easy to view, or a JSON format, which could allow another service to more easily import it.
Downloading your information is a password-protected process that only you will have access to. Once your copy has been created, it will be available for download for a few days.
If you'd like to view your information without downloading it, you can Access Your Information at any time.
Request a download
Available files
Last week
Last month
Last 3 months
Last 6 months
Last year
Last 3 years
All time
Custom
and date range for your download. HTML
vs another service to more easily import the
s and videos but also affects file size.
Date range (required)
Last month
Select information to download


Facebook
facebook.com/dyi
facebook
Download Your Information
You can download a copy of your Facebook information at any time. You can download a complete copy, or you can select only the types of information and date ranges you want. You can choose to receive your information in an HTML format that is easy to view, or a JSON format, which could allow another service to more easily import it.
Downloading your information is a password-protected process that only you will have access to. Once your copy has been created, it will be available for download for a few days.
If you'd like to view your information without downloading it, you can Access Your Information at any time.
When you last started and finished a Privacy Checkup topic
What's included?
Form submissions
Contact info that was saved from forms you submitted to businesses
What's included?
Other logged information
Other information that Facebook logs about your activity
What's included?
Security and login information
Technical information and logged activity related to your account
Security and login information
Technical information and logged activity related to your account
What's included?
Apps and websites off of Facebook
Apps you own and activity we receive from apps and websites off of Facebook
Apps and websites off of Facebook

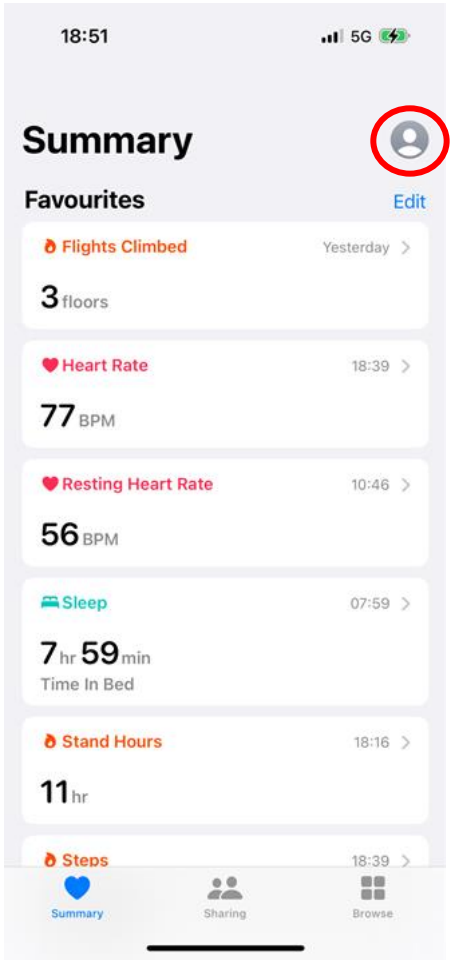

Facebook
facebook.com/dyi
facebook
Download Your Information
You can download a copy of your Facebook information at any time. You can download a complete copy, or you can select only the types of information and date ranges you want. You can choose to receive your information in an HTML format that is easy to view, or a JSON format, which could allow another service to more easily import it.
Downloading your information is a password-protected process that only you will have access to. Once your copy has been created, it will be available for download for a few days.
If you'd like to view your information without downloading it, you can Access Your Information at any time.
What's included?
Preferences
Actions you've taken to customize your experience on Facebook
What's included?
Ads information
Your interactions with ads and advertisers on Facebook
Ads information
Your interactions with ads and advertisers on Facebook
What's included?
Start your download
Your download may contain private information. You should keep it secure and take precautions when storing it, sending it or uploading it to another service.
Request a download

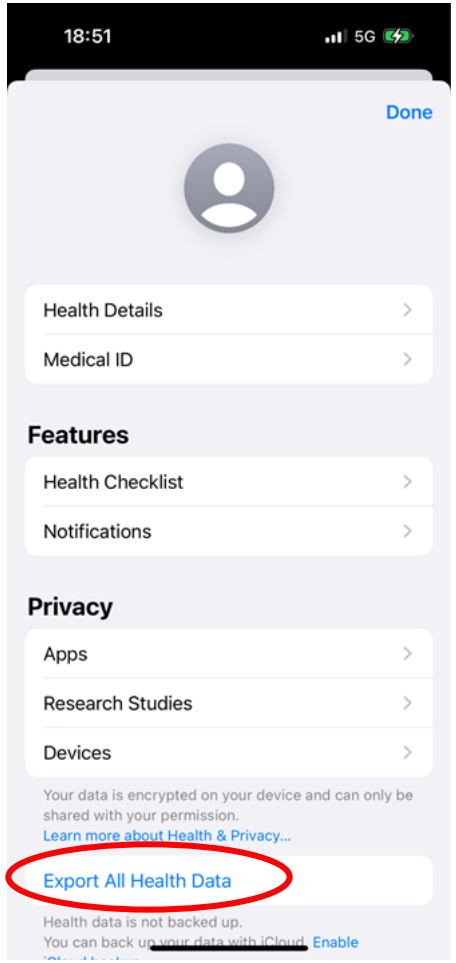

Facebook
facebook.com/dyi
facebook
Download Your Information
You can download a copy of your Facebook information at any time. You can download a complete copy, or you can select only the types of information and date ranges you want. You can choose to receive your information in an HTML format that is easy to view, or a JSON format, which could allow another service to more easily import it.
Downloading your information is a password-protected process that only you will have access to. Once your copy has been created, it will be available for download for a few days.
If you'd like to view your information without downloading it, you can Access Your Information at any time.
Downloads
Request a download
Available files
Mar 31, 2023 - Apr 30, 2023
Security and login information (Less than 1 MB)
Requested on Apr 30 at 2:48 PM
HTML format
High-quality media
1 file
Expires May 4, 2023
Download

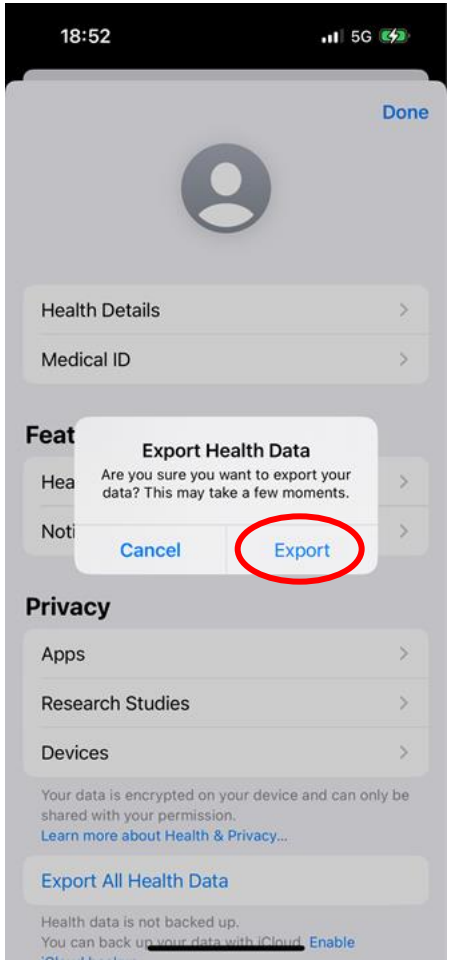

Facebook
facebook.com/dyi
facebook
Download Your Information
You can download a copy of your Facebook information at any time. You c
a complete copy, or you can s
types of information and date
want. You can choose to recei
information in an HTML forma
to view, or a JSON format, wh
allow another service to more
it.
Downloading your informatio
password-protected process t
will have access to. Once your
been created, it will be availab
download for a few days.
If you'd like to view your information without downloading it, you can Access Your Information at any time.
Downloads
Request a download
Available files
Download
Please re-enter your password
Florian Keusch
For your security, you must re-enter your password to continue.
Password
Cancel
Confirm


Facebook
facebook.com/dyi
facebook
Download Your Information
You can download a copy of your Facebook information at any time. You can download a complete copy, or you can select only the types of information and date ranges you want. You can choose to receive your information in an HTML format that is easy to view, or a JSON format, which could allow another service to more easily import it.
Downloading your information is a password-protected process that only you will have access to. Once your copy has been created, it will be available for download for a few days.
If you'd like to view your information without downloading it, you can Access Your Information at any time.
Downloads
Request a download
Available files
Mar 31, 2023 - Apr 30, 2023
Security and login information (Less than 1 MB)
Requested on Apr 30 at 2:48 PM
HTML format
High-quality media
1 file
Expires May 4, 2023
Download
facebook-floriankeus....zip
Alle anzeigen

# 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 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 — 109
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

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 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

# 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

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

# Exercise 2:
# 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 (see slide “Download DDP from iHealth”)
  - Go to Health on your iPhone
  - Click on the icon ‘Personalize’
  - Click on Export All Health Data
- Prepare the data for analysis and find out something about yourself (see Surfdrive for code)
- If you do not have iHealth, use Bella’s data (see file export.zip on Surfdrive)

# Additional reading


Araujo, T., Ausloos, J., van Atteveldt, W., Loecherbach, F., Moeller, J., Ohme, J., Trilling, D., van de Velde, B., de Vreese, C., & Welbers, K. (2022). OSD2F: An Open-Source Data Donation Framework. *Computational Communication Research*, 4(2), 372-387. https://doi.org/10.5117/CCR2022.2.001.ARAU

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.

Boeschoten, L., Ausloos, J., Möller, J.E., Araujo, T., & Oberski, D.L. (2022). A framework for privacy preserving digital trace data collection through data donation. Computational Communication Research, 4(2), 388-423. https://doi.org/10.5117/CCR2022.2.002.BOES

Boeschoten, L., Mendrik, A., van der Veen, E., Vloothuis, J., Hu, H., Voorvaart, R., & Oberski, D.L. (2022). Privacy-preserving local analysis of digital trace data: A proof-of-concept. Patterns, 3(3), 100444. https://doi.org/10.1016/j.patter.2022.100444

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.

Van Driel, I.I., Giachanou, A., Pouwels, J.L., Boeschoten, L., Beyens, I., & Valkenburg, P.M. (2022). Promises and Pitfalls of Social Media Data Donations. *Communication Methods and Measures*, 16(4), 266-282. https://doi.org/10.1080/19312458.2022.2109608