Measuring Citizen's Digital Behaviours Using Web Trackers and Data Donations

Oriol J. Bosch | University of Oxford & RECSM

oriol.bosch-jover@demography.ox.ac

orioljbosch

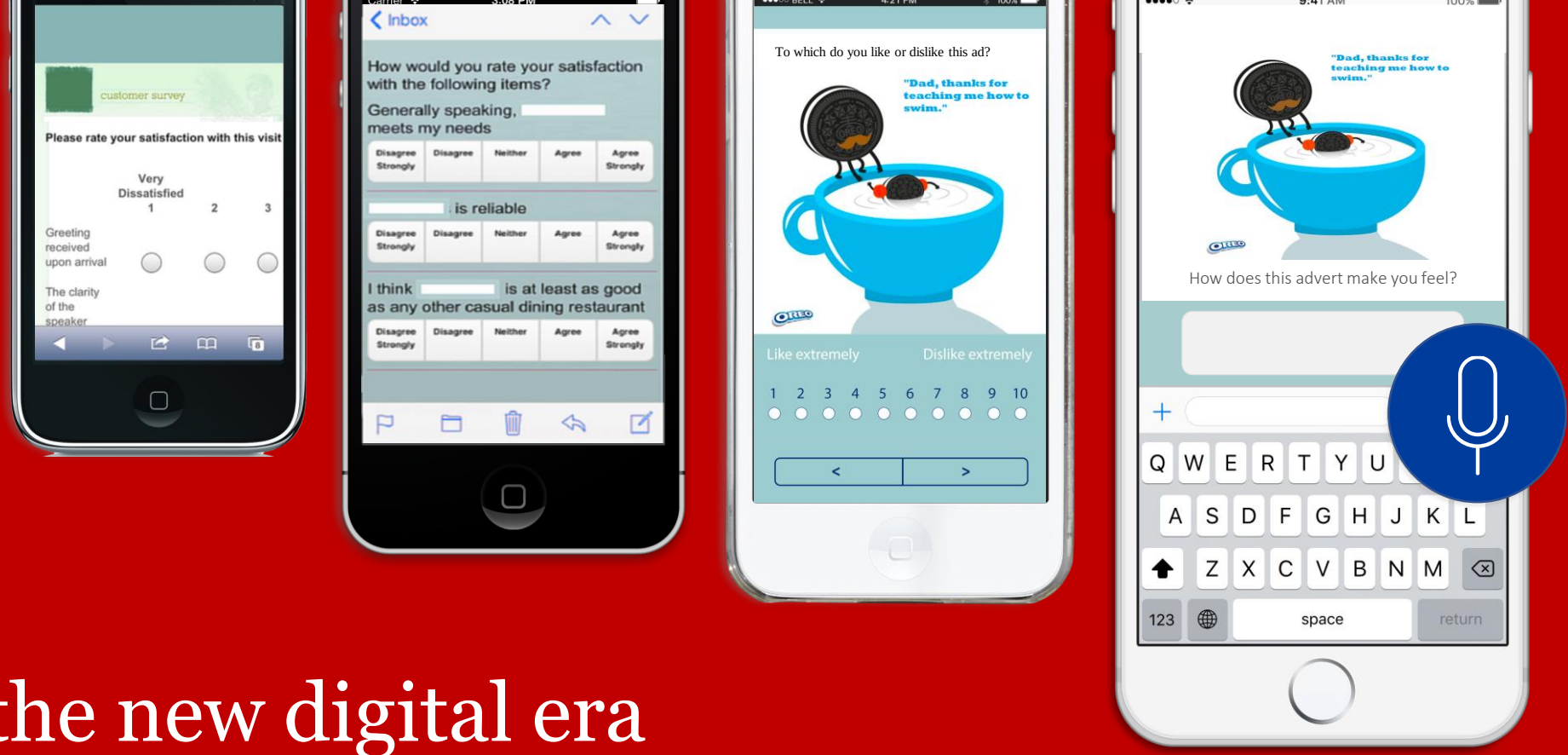
https://orioljbosch.com/

Funding: This project has received funding from the European Research Council (ERC) under the European Unions Horizon 2020 research and innovation programme (grant agreement No 849165; PI: Melanie Revilla); the Spanish Ministry of Science and Innovation under the "R+D+i projects" programme (grant number PID2019-106867RB-I00 /AEI/10.13039/501100011033 (2020-2024), PI: Mariano Torcal); and the BBVA foundation under their grant scheme to scientific research teams in economy and digital society, 2019 (PI: Mariano Torcal).

HOUSE KEEPING

Today's webinar

1. Some context
2. Quick intro to web tracking data and data donations
3. A guide to collecting and using web tracking data
4. Break for questions
5. Some challenges when using web tracking data
6. A guide to collecting and using data donations
7. Q&A



Surveys and the new digital era

The digital era of surveys

1. What people do on the digital realm can impact both online and offline phenomena.

The New York Times

OPINION
GUEST ESSAY

Does Instagram Harm Girls? No One Actually Knows.

Oct. 10, 2021

UK Parliament

Business | MPs, Lords & offices | About | Get inv

UK Parliament

Democracy under threat from 'pandemic of misinformation' online – Lords Democracy and Digital Technologies Co

Democracy under threat from 'pandemic of misinformation' online – Lords Democracy and Digital Technologies Committee

The UK Government should act immediately to deal with a 'pandemic of misinformation' that poses an existential threat to our democracy and way of life. The stark warning comes in a report published today by the Lords Committee on Democracy and Digital Technologies.

The report says the Government must take action 'without delay' to ensure tech giants are held responsible for the harm done to individuals, wider society and our democratic processes through misinformation widely spread on their platforms.

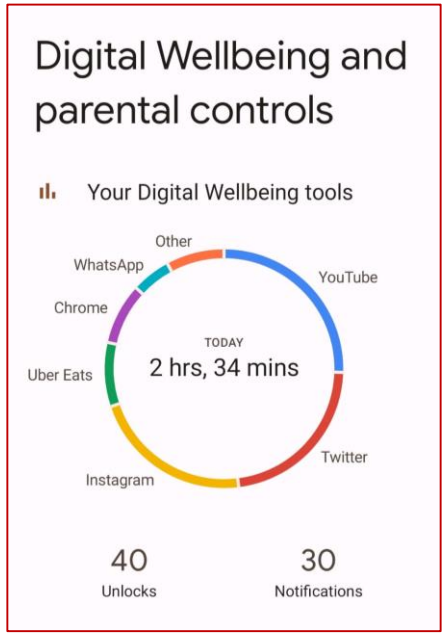
The Committee says online platforms are not 'inherently ungovernable' but power has been ceded to a "few unelected and unaccountable digital corporations" including Facebook and Google, and politicians must act now to hold those corporations to account when they are shown to negatively influence public debate and undermine democracy.

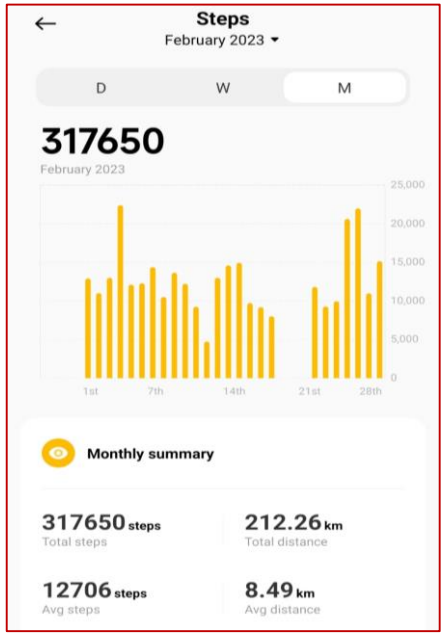
The Committee sets out a package of reforms which, if implemented, could help restore public trust and ensure democracy does not 'decline into irrelevance'.

The digital era of surveys

1. What people do on the digital realm can impact both online and offline phenomena.

2. The digitalisation of our lives is making new types of data available



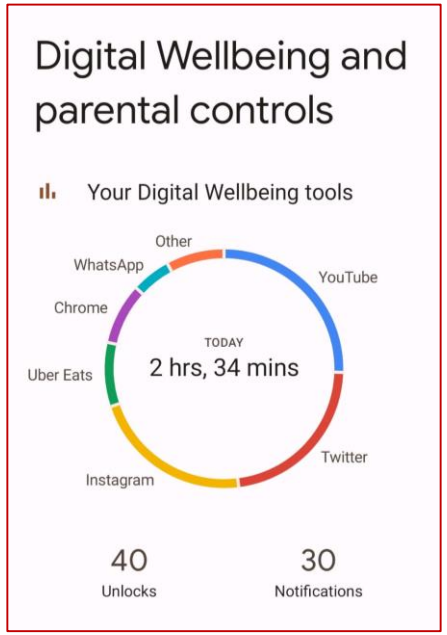


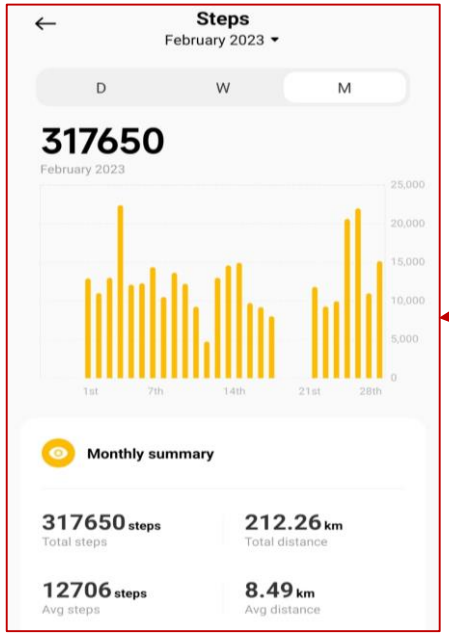
The digital era of surveys

1. What people do on the digital realm can impact both online and offline phenomena.

2. The digitalisation of our lives is making new types of data available

We can ask people to self-report these behaviours





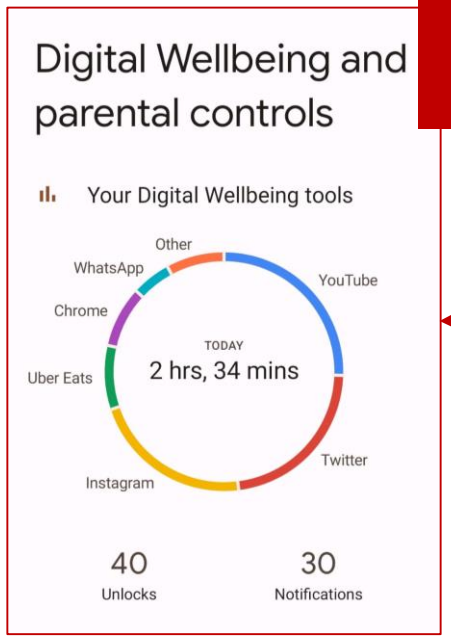
The digital era of surveys

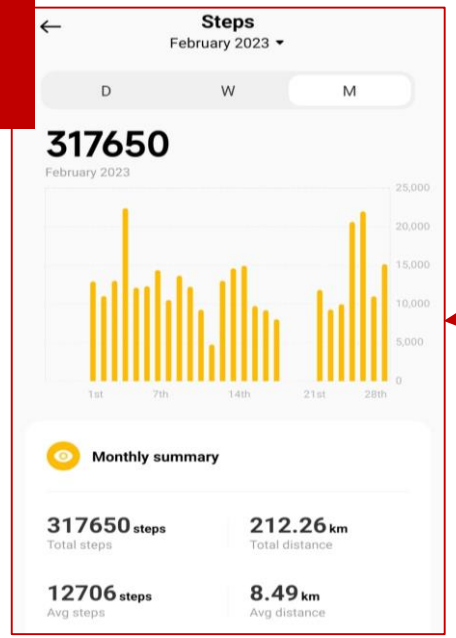
1. What people do on the digital realm can impact both online and offline phenomena.

2. The digitalisation of our lives is making new types of data available

We can ask people to self-report these behaviours

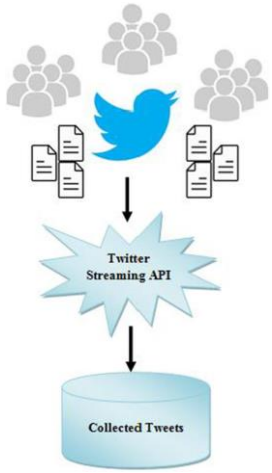
Or we can directly collect these digital traces

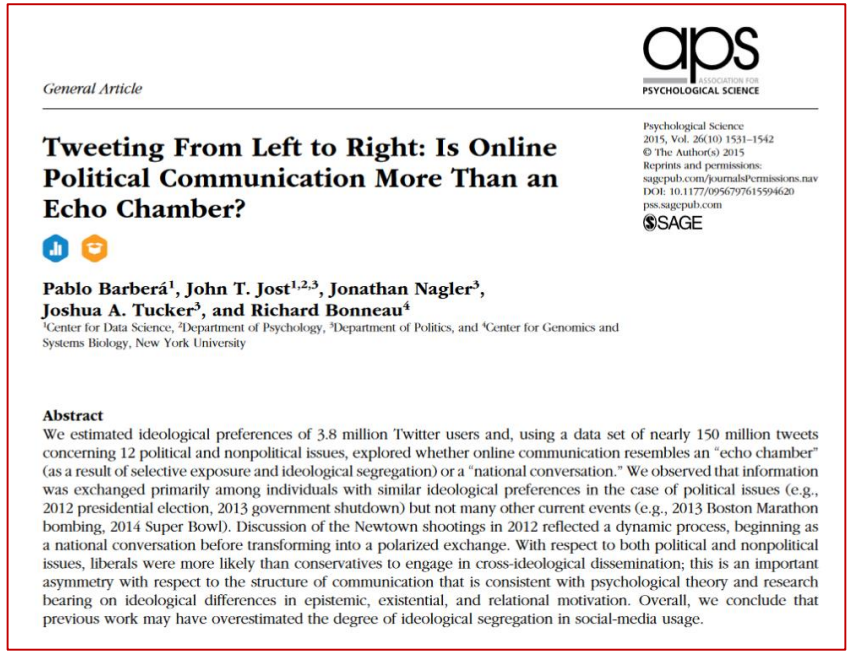




Measuring what people do online with designed digital data

- Most common approach to collect digital traces: collect data directly from the platforms (“Found data”)





General Article

aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Tweeting From Left to Right: Is Online Political Communication More Than an Echo Chamber?

Psychological Science
2015, Vol. 26(10) 1531–1542

DOI: 10.1177/0956797615594620
pss.sagepub.com
SAGE

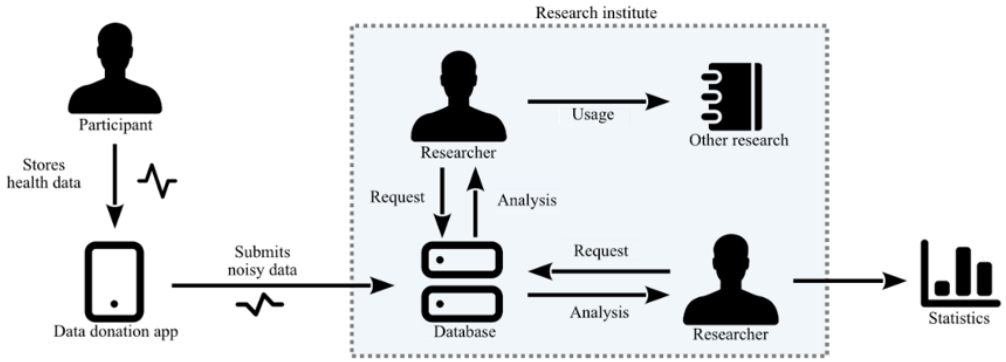
Pablo Barberá[1], John T. Jost[1,2,3], Jonathan Nagler[3], Joshua A. Tucker[3], and Richard Bonneau[4]
[1]Center for Data Science, [2]Department of Psychology, [3]Department of Politics, and [4]Center for Genomics and Systems Biology, New York University

Abstract
We estimated ideological preferences of 3.8 million Twitter users and, using a data set of nearly 150 million tweets concerning 12 political and nonpolitical issues, explored whether online communication resembles an “echo chamber” (as a result of selective exposure and ideological segregation) or a “national conversation.” We observed that information was exchanged primarily among individuals with similar ideological preferences in the case of political issues (e.g., 2012 presidential election, 2013 government shutdown) but not many other current events (e.g., 2013 Boston Marathon bombing, 2014 Super Bowl). Discussion of the Newtown shootings in 2012 reflected a dynamic process, beginning as a national conversation before transforming into a polarized exchange. With respect to both political and nonpolitical issues, liberals were more likely than conservatives to engage in cross-ideological dissemination; this is an important asymmetry with respect to the structure of communication that is consistent with psychological theory and research bearing on ideological differences in epistemic, existential, and relational motivation. Overall, we conclude that previous work may have overestimated the degree of ideological segregation in social-media usage.

Measuring what people do online with designed digital data

- Most common approach to collect digital traces: collect data directly from the platforms (“Found data”)

- These **traces can also be collected in a designed way**: for instance, within a survey



Appenzeller, A., Terzer, N., Krempel, E., & Beyerer, J. (2022, June). Towards private medical data donations by using privacy preserving technologies. In *Proceedings of the 15th International Conference on PErvasive Technologies Related to Assistive Environments* (pp. 446-454).

Measuring what people do online with designed digital data

- Most common approach to collect digital traces: collect data directly from the platforms (“Found data”)
- These **traces can also be collected in a designed way**: for instance, within a survey
- **Two main types** of designed digital data for understanding digital behaviours

Web tracking data

Data donations

A quick intro to web tracking data & data donations

Web tracking data

Direct observations of online behaviours using tracking solutions, or *meters*.

Group of tracking technologies (plug-ins, apps, proxies, etc)

Installed on participants devices

Collect traces left by participants when **interacting with their devices online: URLs, apps visited, cookies…**

Data donations

Users directly provide researchers with data that already has been collected by their devices or platforms

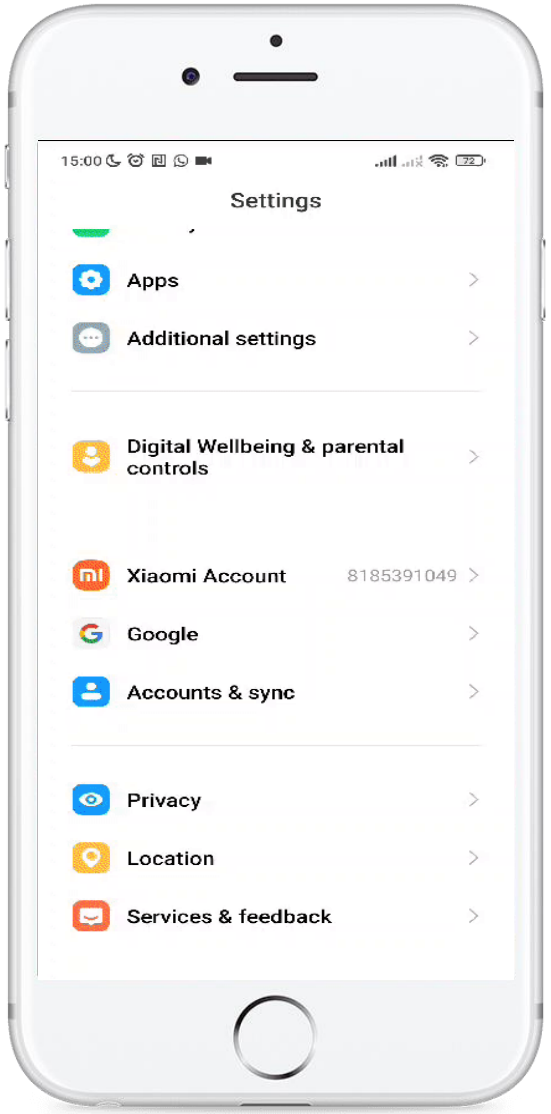
Participants must **access** this data

Capture it in some way

And **share** it with researchers

This **process**, as well as the **traces** collectable, can **vary a lot from project to project**

Platforms that easily allow for this: Meta, Google, Twitter, etc.



A guide to collecting and using web tracking data

Total Error framework for digital traces collected w/ Meters (TEM)

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Royal Statistical Society
Statistics in Society
Series A

ORIGINAL ARTICLE | Open Access

When survey science met web tracking: Presenting an error framework for metered data

Oriol J. Bosch Melanie Revilla

First published: 06 November 2022 | https://doi.org/10.1111/rssa.12956

Funding information: Fundación BBVA, H2020 European Research Council, , Grant/Award Number: 849165; Ministerio de Ciencia e Innovación, , Grant/Award Number: PID2019-106867RB- I00 /AEI/10.13039 /501100011033

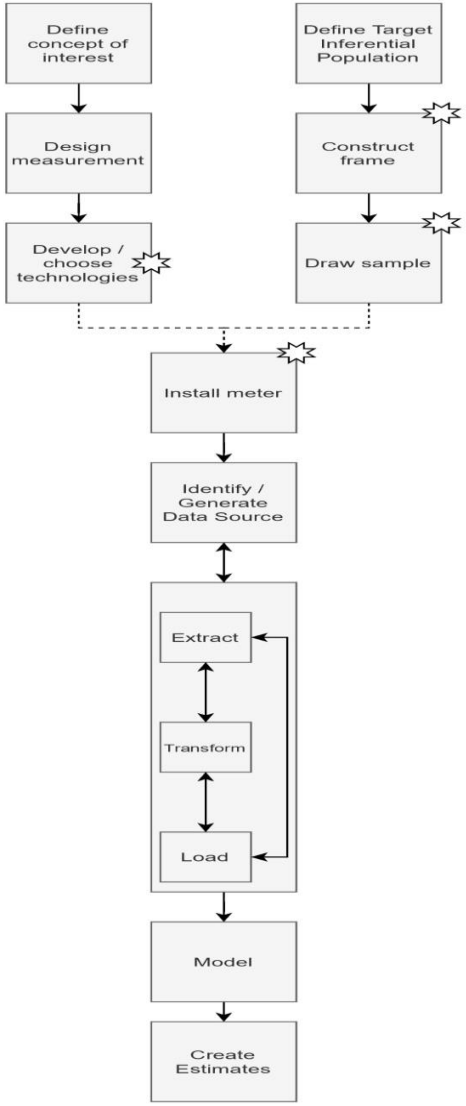
SECTIONS PDF TOOLS SHARE

Abstract

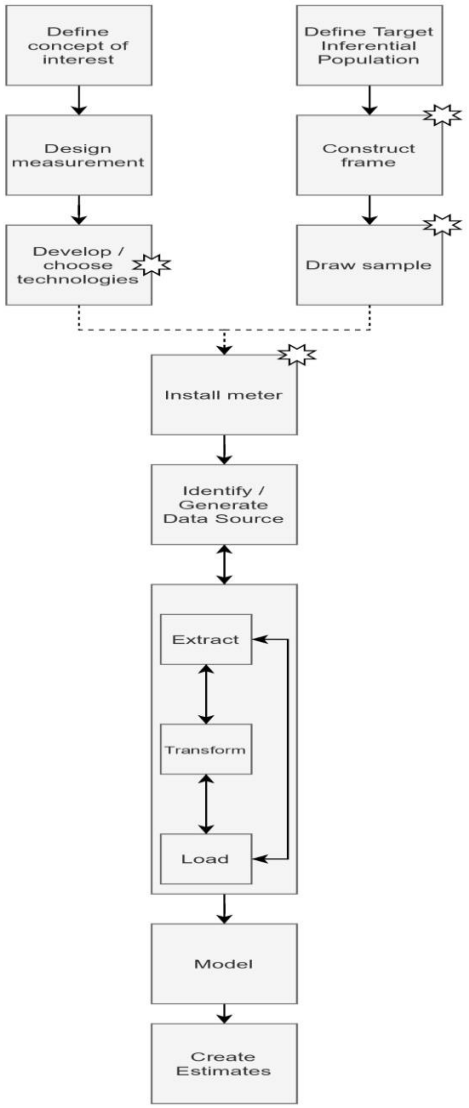
Metered data, also called web-tracking data, are generally collected from a sample of participants who willingly install or configure, onto their devices, technologies that track digital traces left when people go online (e.g., URLs visited). Since metered data allow for the observation of online behaviours unobtrusively, it has been proposed as a useful tool to understand what people do online and what impacts this might have on online and offline phenomena. It is crucial, nevertheless, to understand its limitations. Although some research have explored the potential errors of metered data, a systematic categorisation and conceptualisation of these errors are missing. Inspired by the Total Survey Error, we present a Total Error framework for digital traces collected with Meters (TEM). The TEM framework (1) describes the data generation and the analysis process for metered data and (2) documents the sources of bias and variance that may arise in each step of this process. Using a case study we also show how the TEM can be applied in real life to identify, quantify and reduce metered data errors. Results suggest that metered data might indeed be affected by the error sources identified in our framework and, to some extent, biased. This framework can help improve the quality of both stand-alone metered data research projects, as well as foster the understanding of how and when survey and metered data can be combined.

Total Error framework for digital traces collected w/ Meters (TEM)

- In general, web tracking data is used to **make inferences** about a **concept of interest** for a given **population**



A step-by-step guide

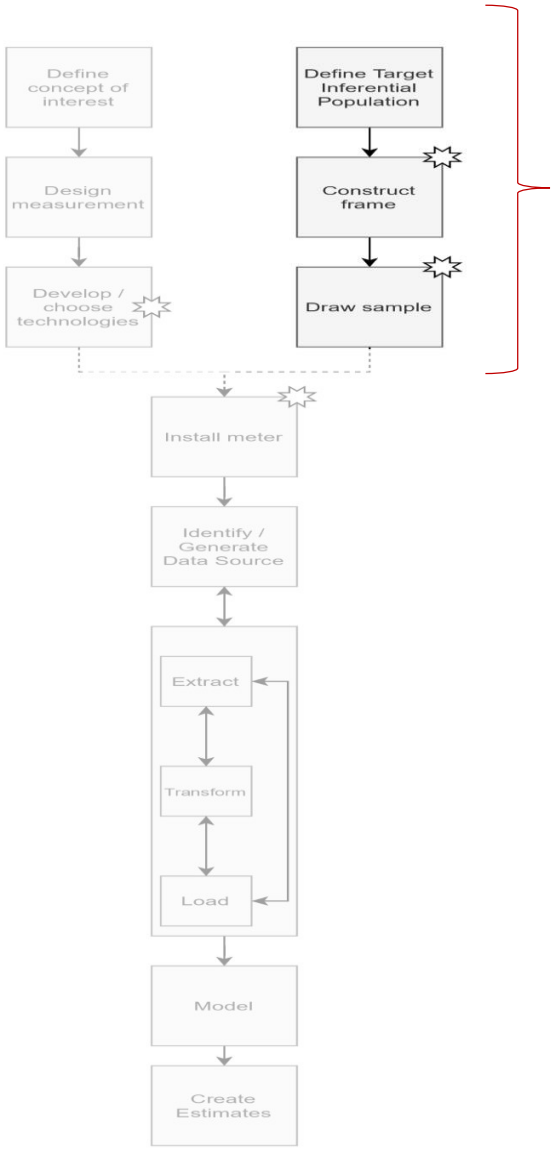


There are many steps to follow when collecting web tracking data.

Many decisions can be made for each step, all with potential impact on data quality

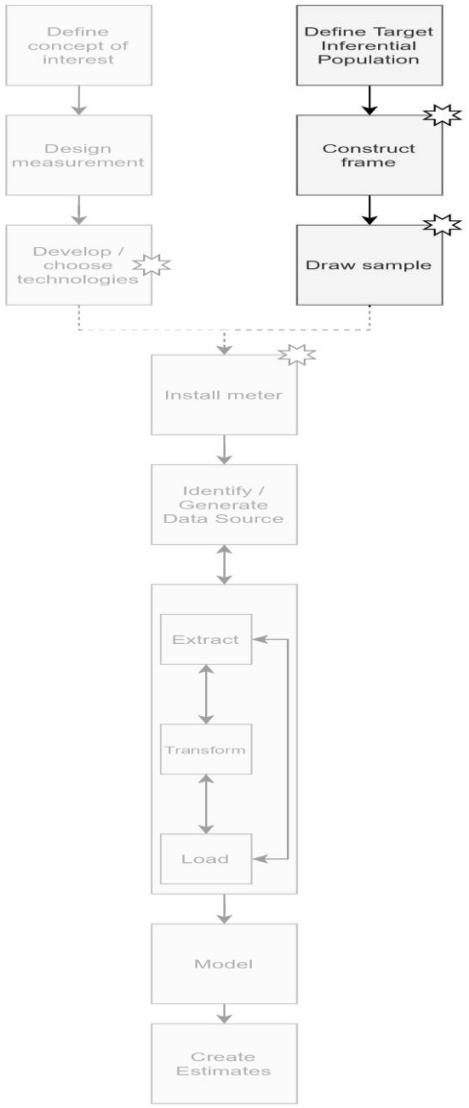
This is rarely acknowledged and understood, we can do better!

First steps on the representation side: same old, same old



Identical steps as for surveys

First steps on the representation side: same old, same old



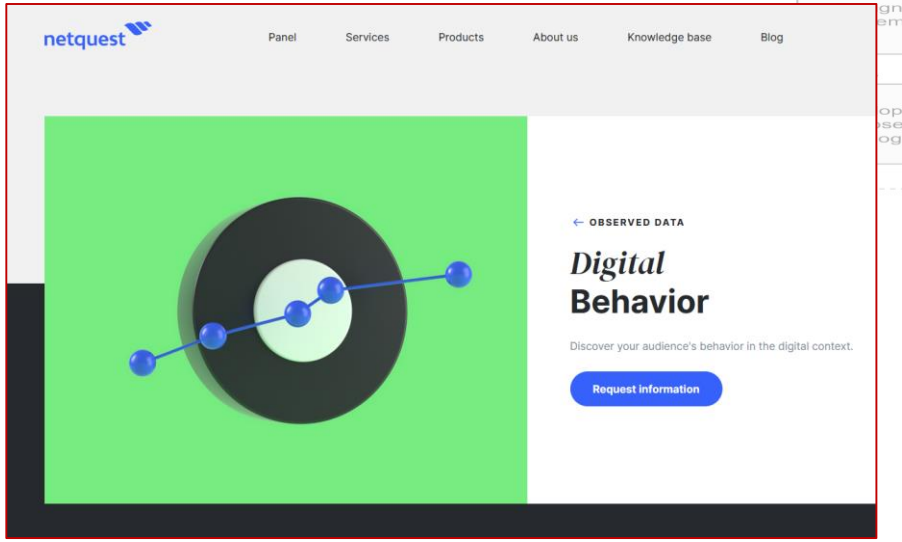
Identical steps as for surveys

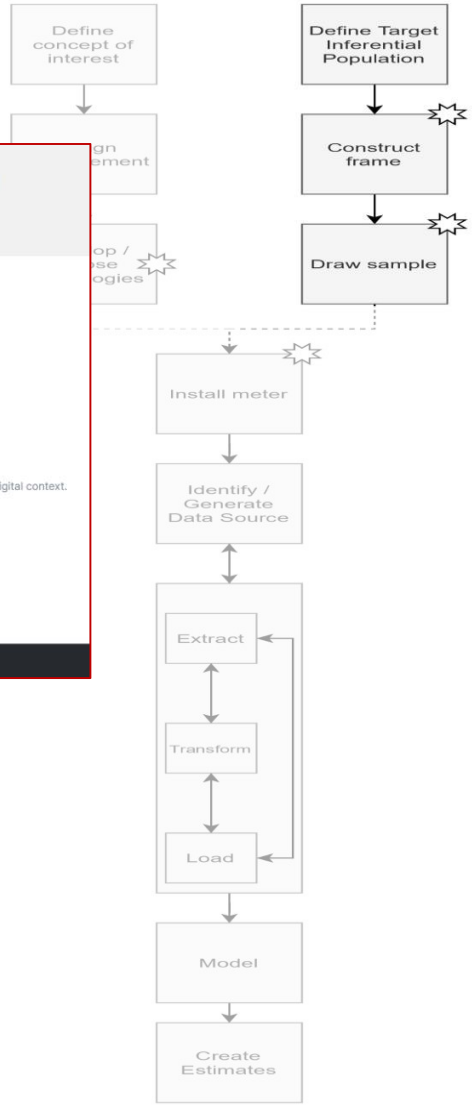
Target population: People living in the UK older than 17

Frame: Postal Address Frame

Sample: Simple Random Sampling

First steps on the representation side: same old, same old





Identical steps as for surveys

Target population: People living in the UK older than 17

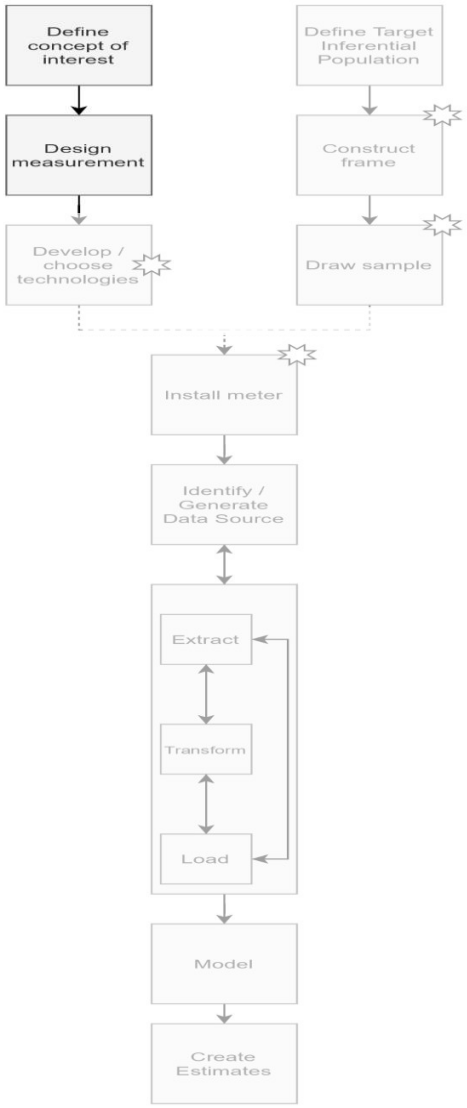
Frame: Postal Address Frame

Sample: Simple Random Sampling

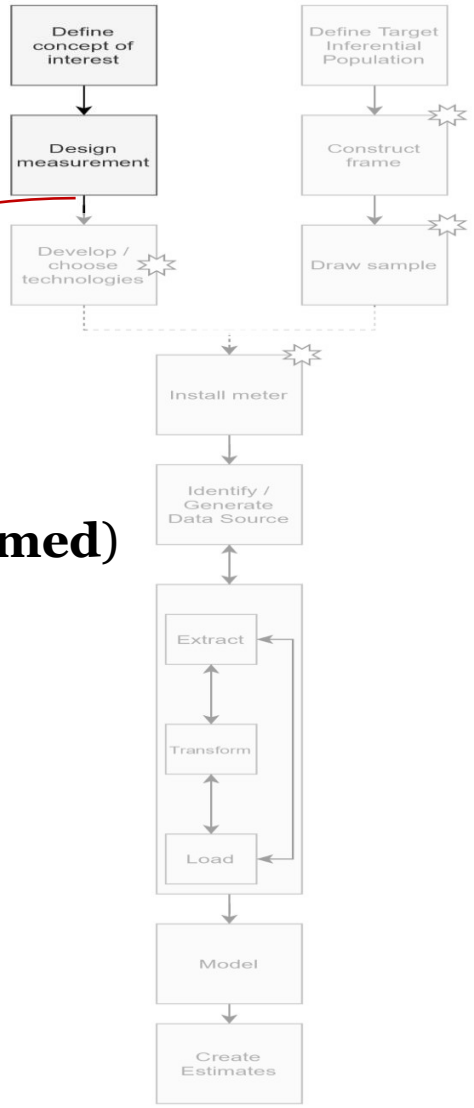
Most commonly: non-probability online panels

YouGovPulse

From concepts to measurements: similar but different



From concepts to measurements: similar but different



Measurements: **Traces** that will be **collected**, **combined** (and **transformed**) to compute a specific variable

From concepts to measurements: similar but different

- Normally not acknowledged: **it is key to clearly define the traces that will be used to measure a specific concept**

Concept of interest → **Measurement**

From concepts to measurements: similar but different

- Normally not acknowledged: **it is key to clearly define the traces that will be used to measure a specific concept**

Concept of interest → **Measurement**

The extent to which an individual encounters written news media

Average time recorded of the visits to URLs defined as showing written news

From concepts to measurements: similar but different

Concept: *average hours of consumption of online political news*

Measure: *average time recorded of the visits to URLs defined as showing written news*

- What traces are considered as a visit?
- Which URLs are considered written news?
- What time frame has been used to compute an average?

From concepts to measurements: similar but different

Concept: *average hours of consumption of online political news*

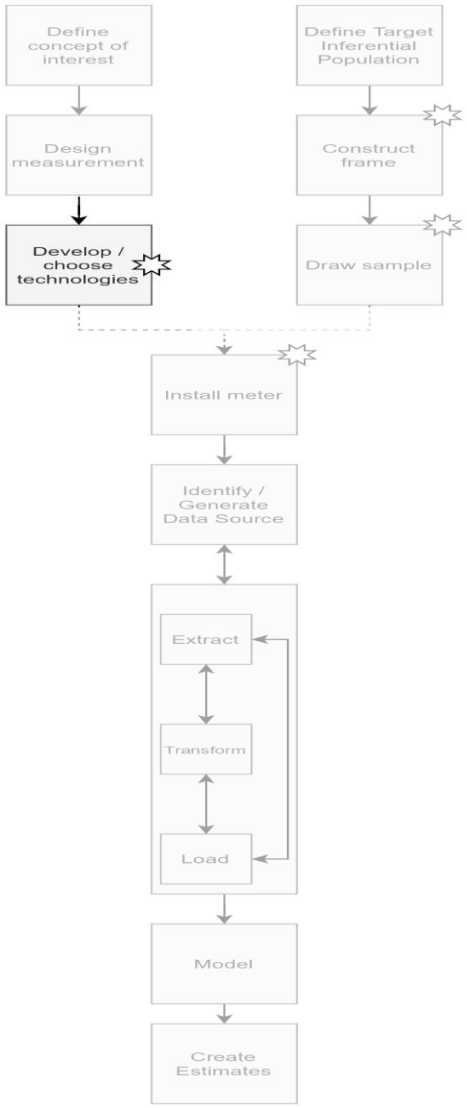
Measure: *average time recorded of the visits to URLs defined as showing written news*

- What traces are considered as a visit?
- Which URLs are considered written news?
- What time frame has been used to compute an average?

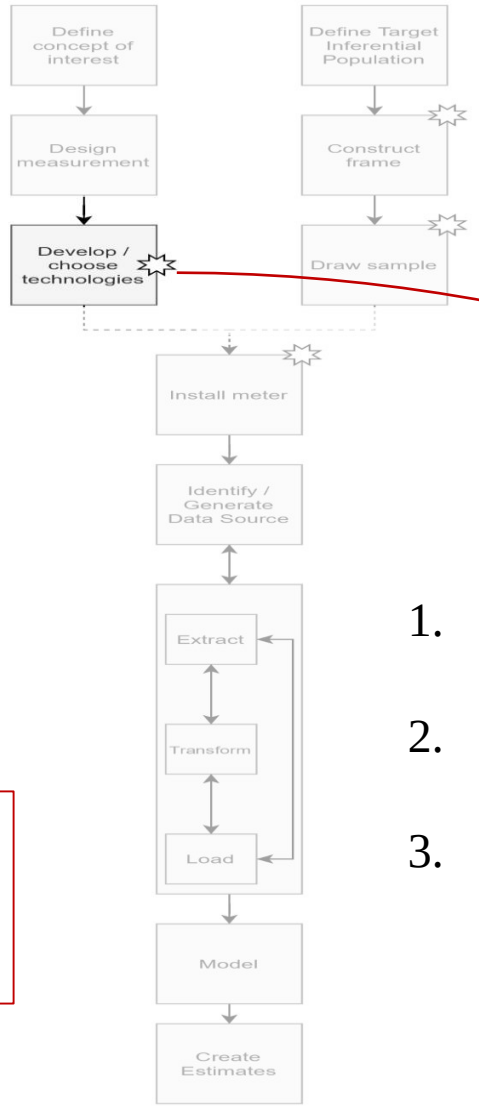
These and other decisions will **determine the measurement used**.

Pretty much as for **surveys** this is determined by the **wording**, **the type of scale**, etc.

Develop or choose the tracking technologies to use



Develop or choose the tracking technologies to use



RealityMine
real life, revealed

wakoopa

COMMUNICATION METHODS AND MEASURES
2022, VOL. 16, NO. 2, 79–95
https://doi.org/10.1080/19312458.2021.1907841
Routledge
Taylor & Francis Group
Check for updates

Automated Tracking Approaches for Studying Online Media Use: A Critical Review and Recommendations

Clara Christner[a], Aleksandra Urman[b], Silke Adam[b], and Michaela Maier[a]

1. We can **develop** tracking technologies from scratch
2. Or use **open-access** technologies already available
3. Or we can use **commercially available** technologies

A heterogeneous group of tracking solutions

- There are **many different types of tracking approaches.**

A heterogeneous group of tracking solutions

- There are **many different types of tracking approaches.**

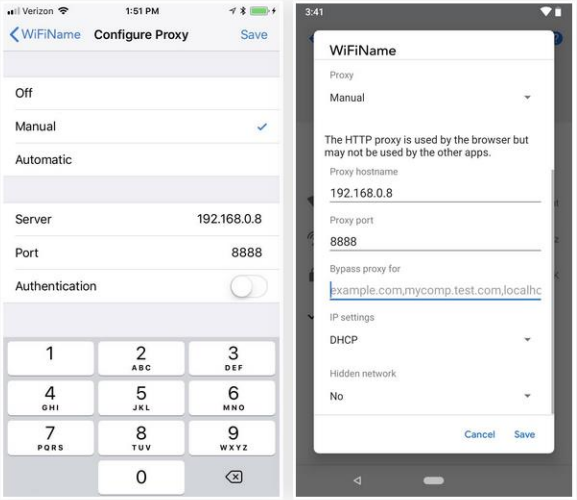
- **These can be**: Proxies, VPNs, Screen-scrapers, Screen recorders, Smartphone-log trackers (and maybe more that I am not aware of).

A heterogeneous group of tracking solutions

- There are **many different types of tracking approaches.**

- **These can be**: Proxies, VPNs, Screen-scrapers, Screen recorders, Smartphone-log trackers (and maybe more that I am not aware of).

- **They can come in different packages for users**: Apps, Browser plug-ins, manual configuration with or without any piece of software required.



https://null-byte.wonderhowto.com/how-to/use-charles-proxy-view-data-your-mobile-apps-send-receive-0185364/

A heterogeneous group of tracking solutions

- There are **many different types of tracking approaches.**
- **These can be**: Proxies, VPNs, Screen-scrapers, Screen recorders, Smartphone-log trackers (and maybe more that I am not aware of).
- **They can come in different packages for users**: Apps, Browser plug-ins, manual configuration with or without any piece of software required.
- **Their capabilities and limitations vary a lot**: not all of them can be installed on all devices. Not all of them can capture the same data. Not all of them have the same level of granularity and accuracy

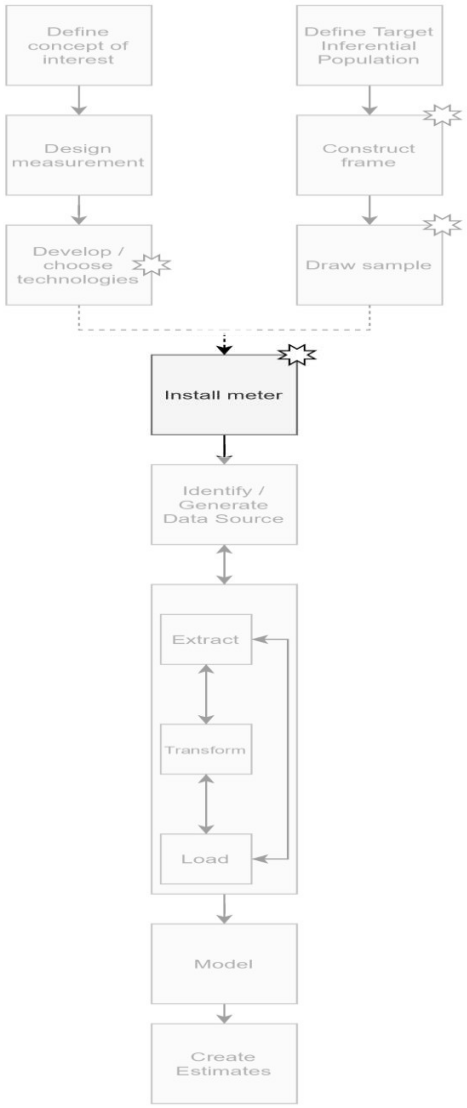
A heterogeneous group of tracking solutions

- Most real-life projects end up using a **combination of approaches**, depending on the devices that people use

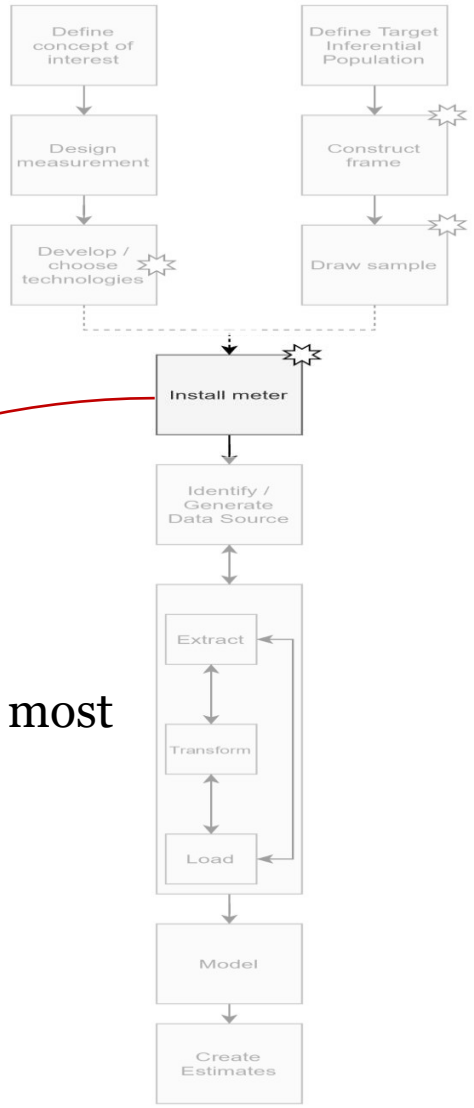
		PC app	PC plug-ins			Android SDK	iOS proxy
			Chrome	Firefox	Safari		
Online tracking							
URLs	Http traffic	Yes	Yes	Yes	Yes	Yes	Yes
	Https traffic	No	Yes	Yes	Yes	Yes	No
	Incognito sessions	No	Yes	Yes	Yes	Yes	No
	HTML	No	Yes	Yes	Yes	No	No
	Time stamps	Yes	Yes	Yes	Yes	Yes	Yes
Apps	App name	-	-	-	-	Yes	Yes
	App usage start time	-	-	-	-	Yes	Yes
	App usage duration	-	-	-	-	Yes	Estimated
	Offline apps	-	-	-	-	Yes	No
	In-app behaviour	-	-	-	-	No	No
Search terms	Search terms	Yes	Yes	Yes	Yes	Yes	No
Device information							
Device type	E.g. desktop	Yes	Yes	Yes	Yes	Yes	Yes
Device brand	E.g. Xiaomi		No	No	No	Yes	Yes
Device model	E.g. S9	No	No	No	No	Yes	Yes
Operating system	E.g. iOS	Yes	Yes	Yes	Yes	Yes	Yes
OS version	E.g. 10.1.2	No	No	No	No	Yes	Yes
Internet provider	E.g. Voxi	No	No	No	No	Yes	Yes

Bosch, O. J., & Revilla, M. (2022). When survey science met web tracking: presenting an error framework for metered data. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *185*(Supplement_2), S408-S436.

Could you please, maybe, install this meter?



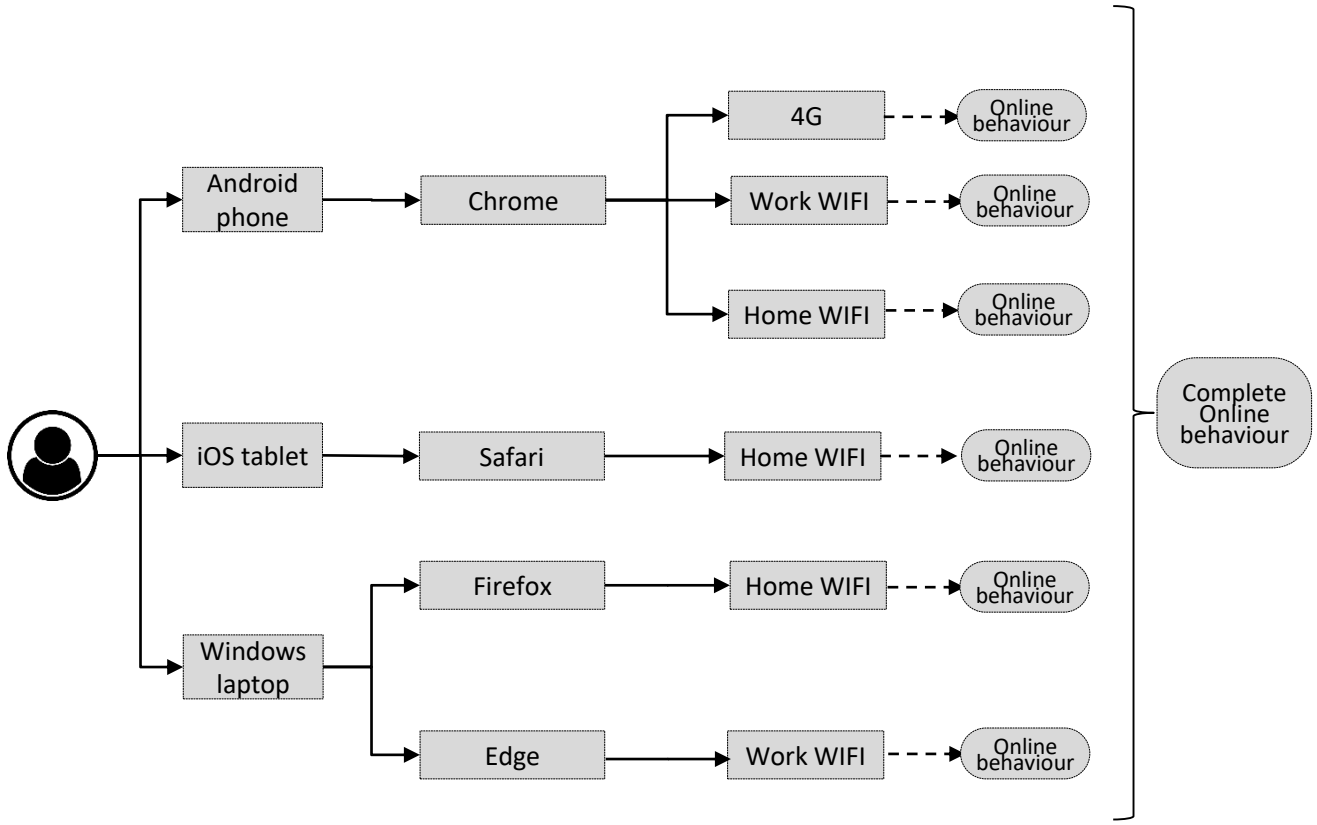
Could you please, maybe, install this meter?



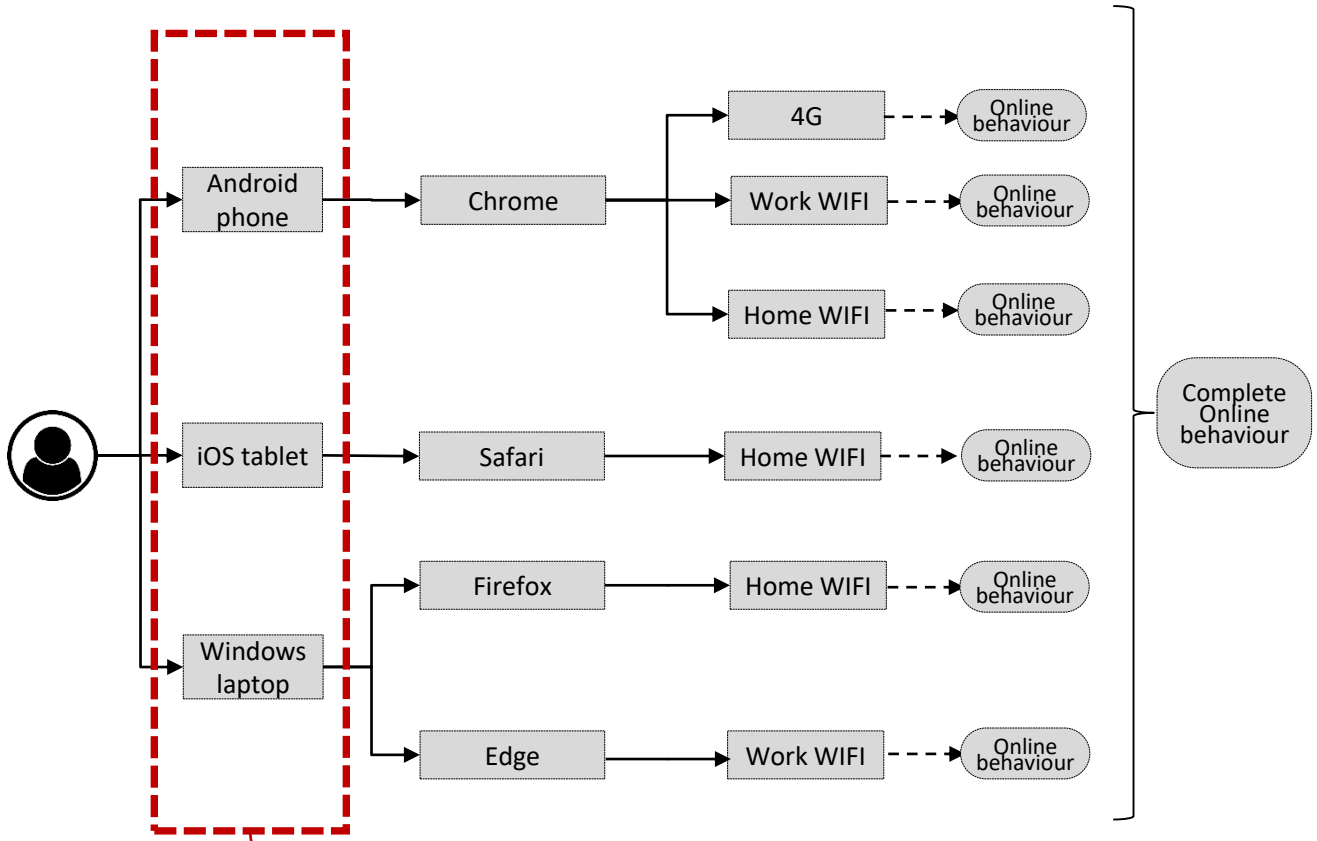
This process is, potentially, one of the most consequential ones for web tracking research. It determines:

1) **Who you track**
2) **And how well you track them**

Could you please, maybe, install this meter?

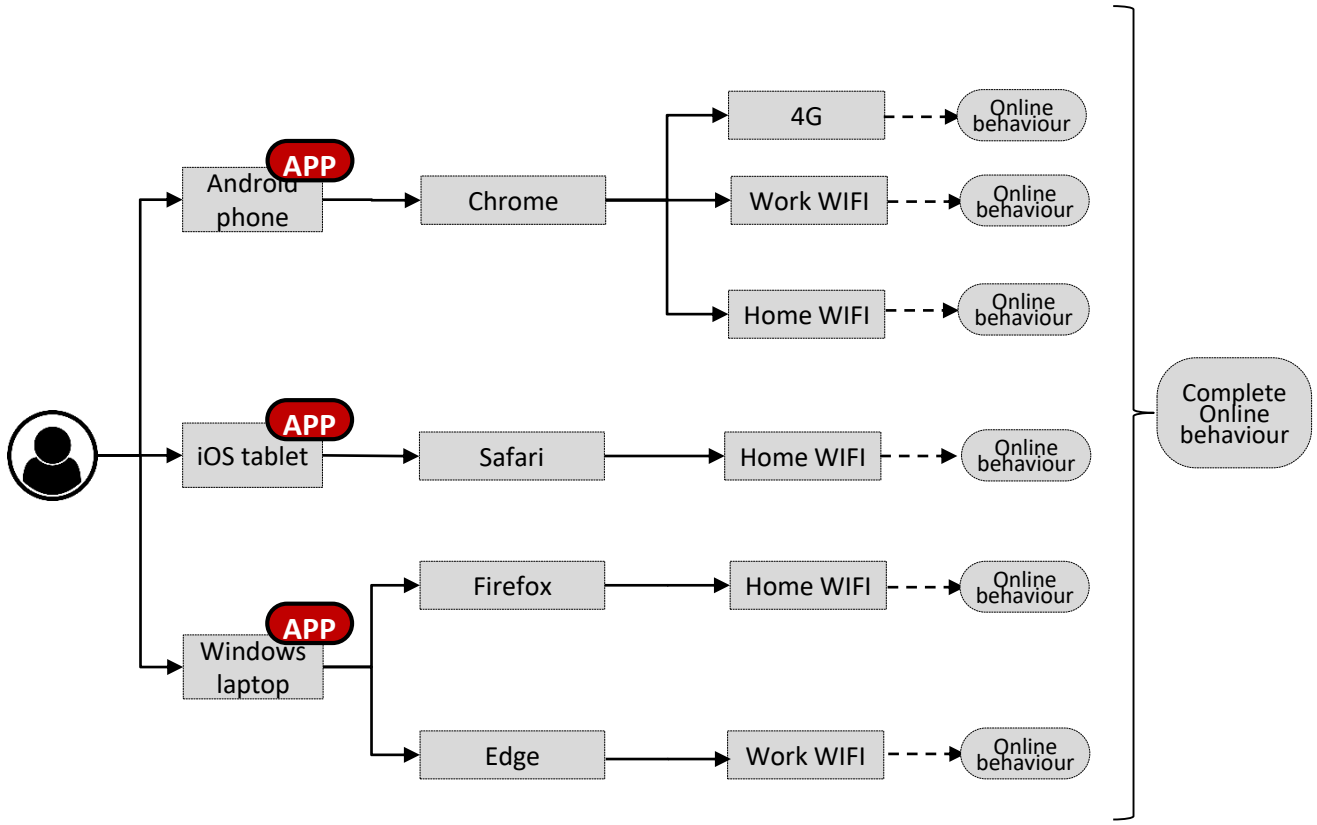


Could you please, maybe, install this meter?



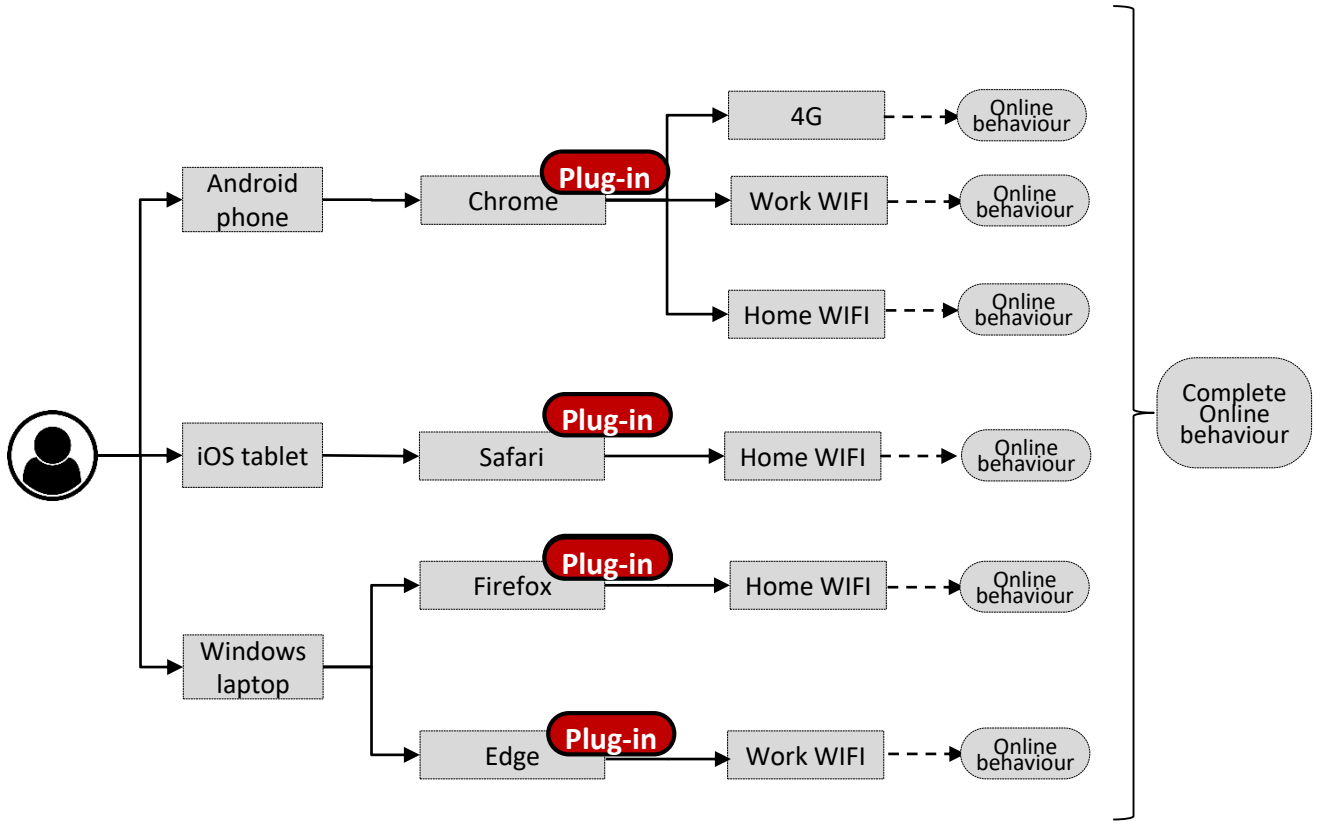
The goal is to know what people do through all their devices

Could you please, maybe, install this meter?



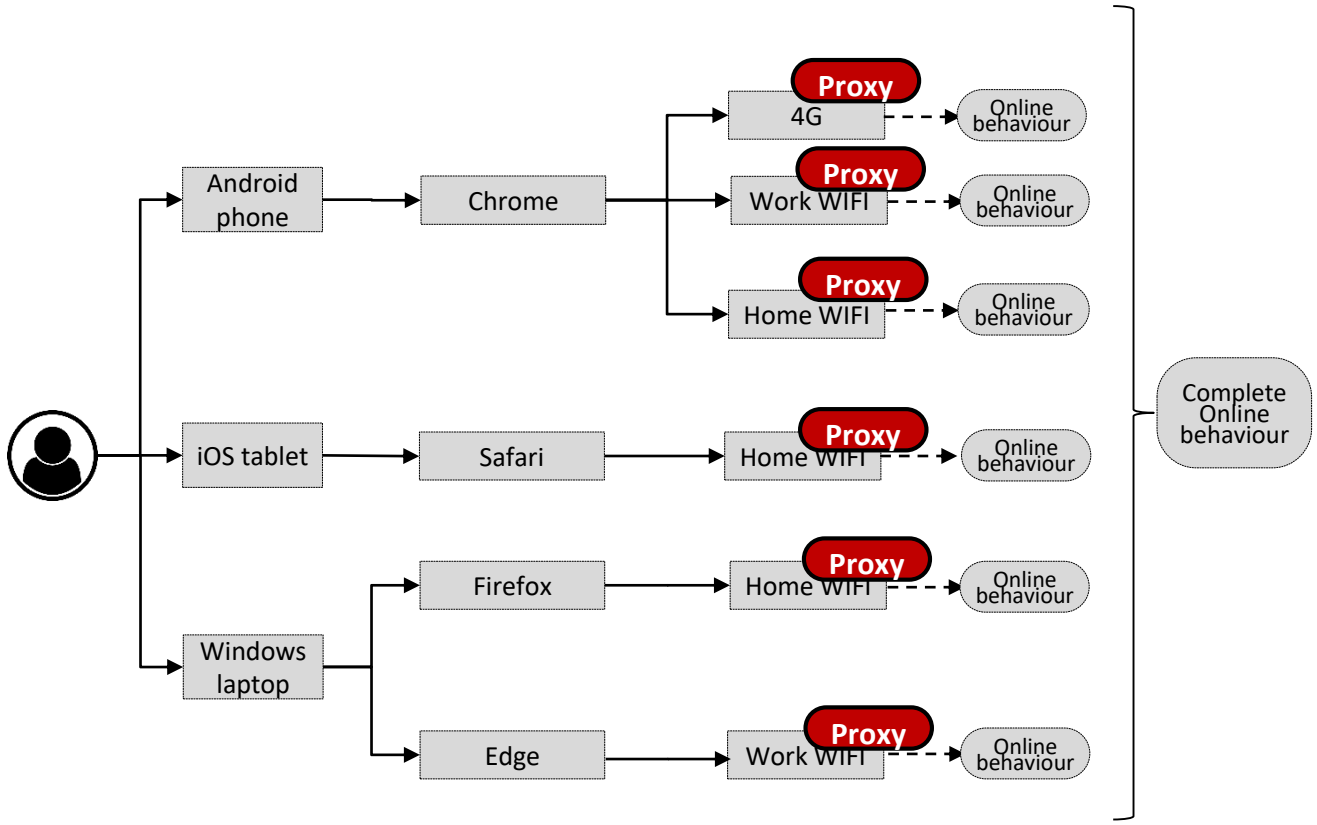
This can be achieved by tracking all devices that someone uses

Could you please, maybe, install this meter?



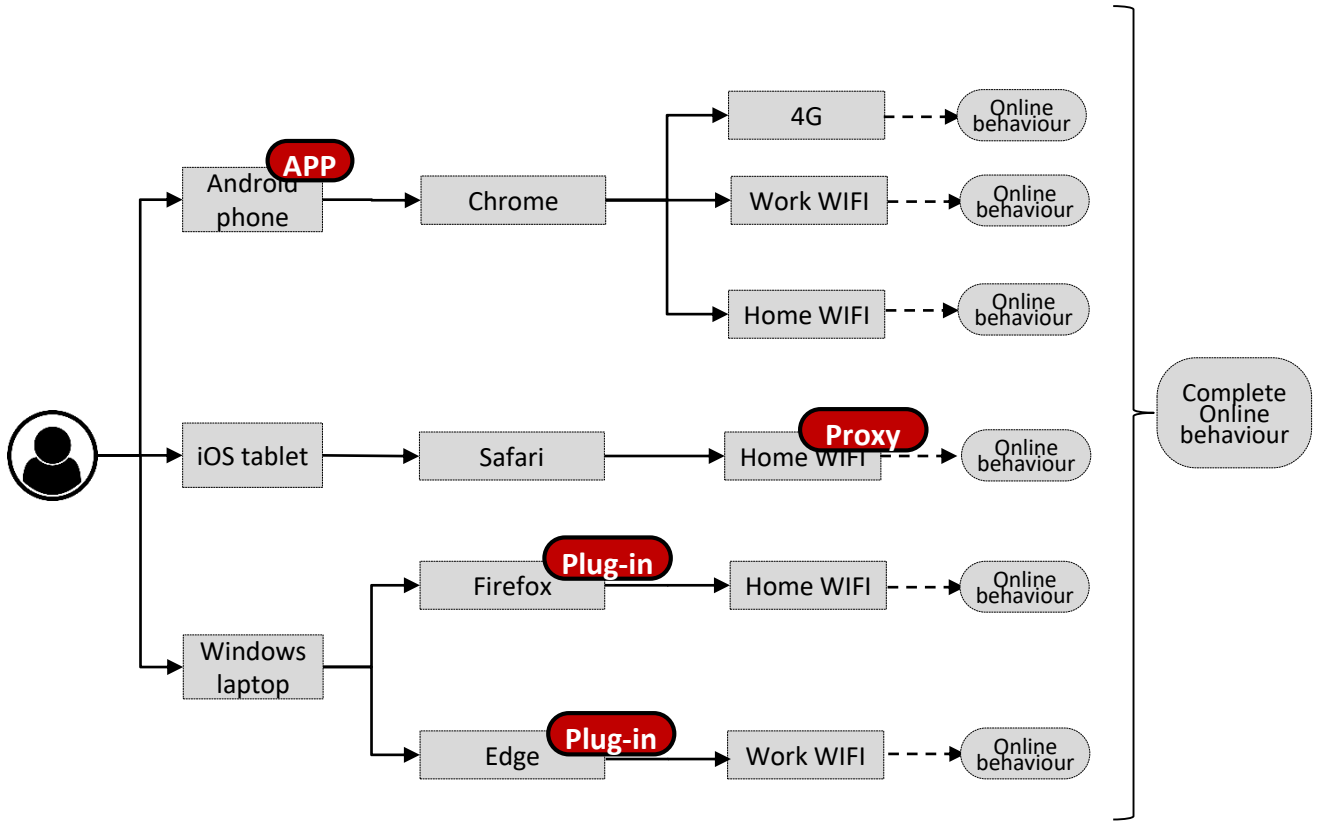
Or all their browsers

Could you please, maybe, install this meter?



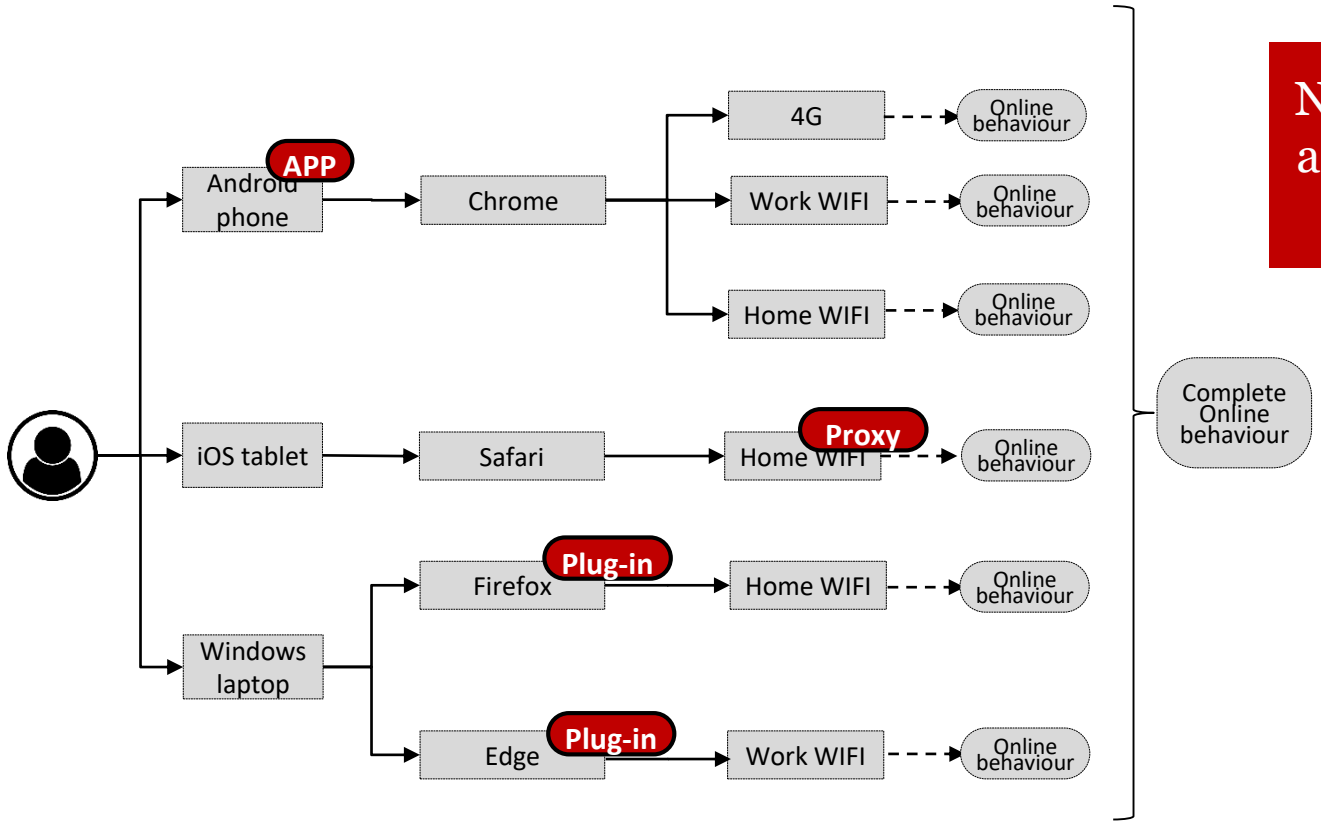
Or all their networks

Could you please, maybe, install this meter?



Or a combination of these (most common)

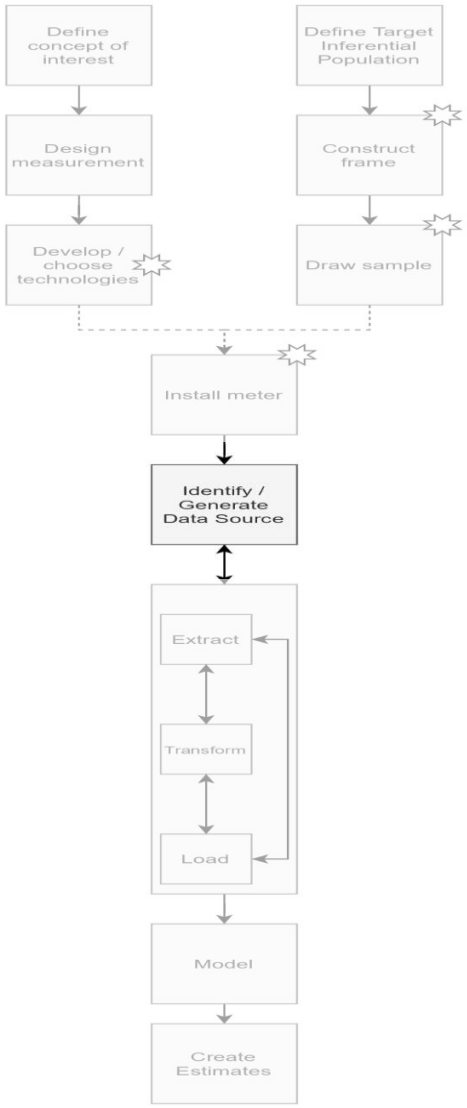
Could you please, maybe, install this meter?



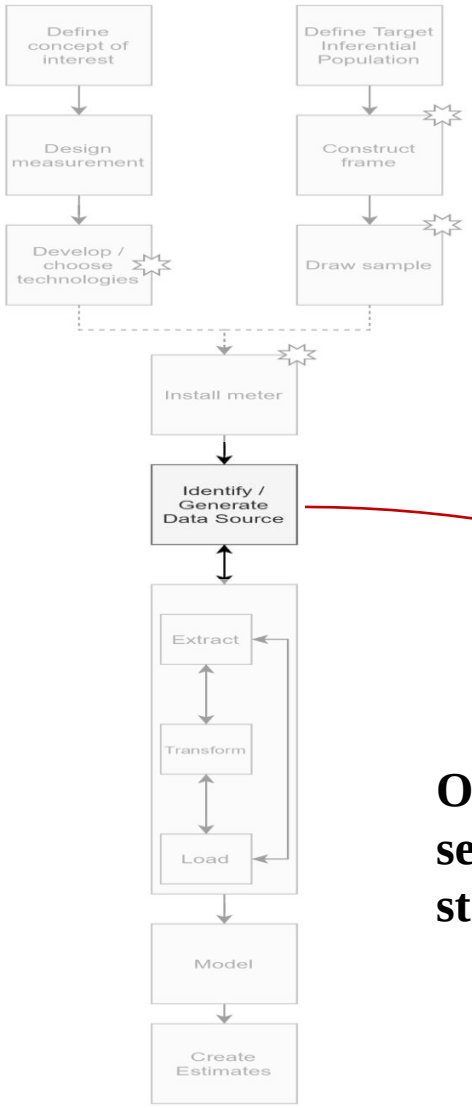
Not always easy to achieve, as we will see later

Or a combination of these (most common)

Generate the messy dataset



Generate the messy dataset



Once the trackers are installed, they start sending information, which is stored in a data storage (e.g., MySQL)

Generate the messy dataset

Figure 1: *Example of web tracking data excerpt*

USERID	STARTTIME	URL
ID:1310	2017-08-13 21:26:45 UTC	HTTPS://WWW.GOOGLE.DE
ID:1310	2017-08-13 21:26:50 UTC	HTTPS://WWW.GOOGLE.DE/SEARCH?Q=BÄCKEREI+GEÖFFNET+IN+DER+NAHE
ID:1310	2017-08-13 21:35:51 UTC	HTTPS://WWW.TWITTER.COM/HOME
ID:2808	2017-08-08 19:28:10 UTC	HTTPS://WWW. YOUGOV.DE/OPI/MYFEED#/ALL
ID:2808	2017-08-08 19:29:10 UTC	HTTPS://WWW.YOUTUBE.COM/WATCH?V=DQW4W9WGXCQ
ID:2808	2017-08-08 19:36:17 UTC	HTTPS://WWW.NETFLIX.COM/WATCH/81441579

- This is one of the **most basic versions** of what information might be recorder (ID, time stamp, and full URL)

Munzert, S., Ramirez-Ruiz, S., Watteler, O., Breuer, J., Batzdorfer, V., Eder, C., ... & Yang, J. (2023). Publishing Combined Web Tracking and Survey Data.

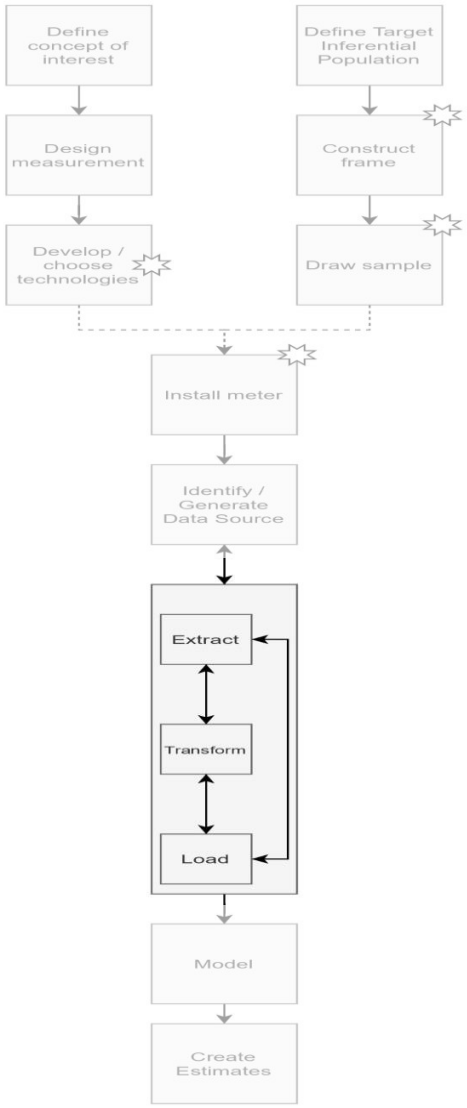
Generate the messy dataset

Figure 1: *Example of web tracking data excerpt*

USERID	STARTTIME	URL
ID:1310	2017-08-13 21:26:45 UTC	HTTPS://WWW.GOOGLE.DE
ID:1310	2017-08-13 21:26:50 UTC	HTTPS://WWW.GOOGLE.DE/SEARCH?Q=BÄCKEREI+GEÖFFNET+IN+DER+NAHE
ID:1310	2017-08-13 21:35:51 UTC	HTTPS://WWW.TWITTER.COM/HOME
ID:2808	2017-08-08 19:28:10 UTC	HTTPS://WWW. YOUGOV.DE/OPI/MYFEED#/ALL
ID:2808	2017-08-08 19:29:10 UTC	HTTPS://WWW.YOUTUBE.COM/WATCH?V=DQW4W9WGXCQ
ID:2808	2017-08-08 19:36:17 UTC	HTTPS://WWW.NETFLIX.COM/WATCH/81441579

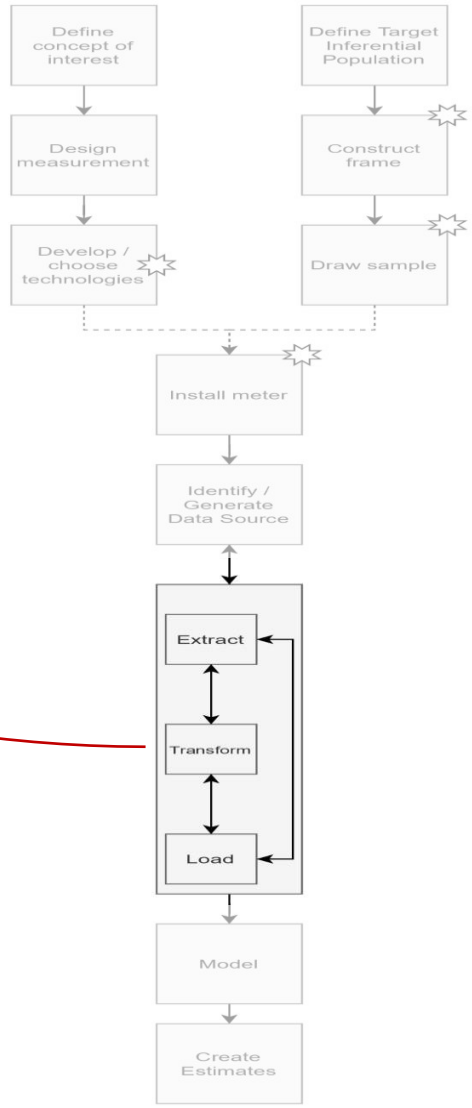
- This is one of the **most basic versions** of what information might be recorder (ID, time stamp, and full URL)
- Other information can be captured, such as **HTML information**. For instance, the **text** each Facebook post seen by a participant, the **number of likes**, the **comments**, why the post was visible, etc.

Let's create the dataset to work with



Let's create the dataset to work with

Most researchers need to process the messy unstructured web tracking data to work with it



Let's create the dataset to work with

- The first step is to **extract the data** of interest. This might mean:

Let's create the dataset to work with

- The first step is to **extract the data** of interest. This might mean:
  - Selecting a **subset of the raw data**. For instance, only full URLs within a given period, or those containing specific values in the URLs

Let's create the dataset to work with

- The first step is to **extract the data** of interest. This might mean:
  - Selecting a **subset of the raw data**. For instance, only full URLs within a given period, or those containing specific values in the URLs
  - Extracting information and **performing calculations to create 'structured' variables** (e.g., counts of visits to specific URLs) ➡ typical SQL queries

Let's create the dataset to work with

- The first step is to **extract the data** of interest. This might mean:
  - Selecting a **subset of the raw data**. For instance, only full URLs within a given period, or those containing specific values in the URLs
  - Extracting information and **performing calculations to create 'structured' variables** (e.g., counts of visits to specific URLs) ➡ typical SQL queries

Figure 1: *Example of web tracking data excerpt*

USERID	STARTTIME	URL
ID:1310	2017-08-13 21:26:45 UTC	HTTPS://WWW.GOOGLE.DE
ID:1310	2017-08-13 21:26:50 UTC	HTTPS://WWW.GOOGLE.DE/SEARCH?Q=BÄCKEREI+GEÖFFNET+IN+DER+NAHE
ID:1310	2017-08-13 21:35:51 UTC	HTTPS://WWW.TWITTER.COM/HOME
ID:2808	2017-08-08 19:28:10 UTC	HTTPS://WWW. YOUGOV.DE/OPI/MYFEED#/ALL
ID:2808	2017-08-08 19:29:10 UTC	HTTPS://WWW.YOUTUBE.COM/WATCH?V=DQW4W9WGXCQ
ID:2808	2017-08-08 19:36:17 UTC	HTTPS://WWW.NETFLIX.COM/WATCH/81441579

Number of visits to google: 2

Number of visits to video platforms: 2

Let's create the dataset to work with

- The second (***optional***) step is to **transform the extracted data**. This might be needed if the defined measurement requires more than simple counts of URLs.

Let's create the dataset to work with

- The second (***optional***) step is to **transform the extracted data**. This might be needed if the defined measurement requires more than simple counts of URLs.

- Most interesting transformation: enriching the information that URLs bring to research.

Let's create the dataset to work with

- The second (***optional***) step is to **transform the extracted data**. This might be needed if the defined measurement requires more than simple counts of URLs.

- Most interesting transformation: enriching the information that URLs bring to research.
  1. The content of the URL can be manually identified, and added to the dataset

https://www.theguardian.com/business/live/2023/jul/12/bank-england-warns-rising-interest-rates-stress-indebted-firm

https://www.theguardian.com/fashion/2023/jul/12/fashion-rental-four-women-on-the-dresses-making-them-a-fortune

https://www.theguardian.com/sport/2023/jul/11/tennis-wimbledon-elina-svitolina-ukraine-war-iga-swiatek

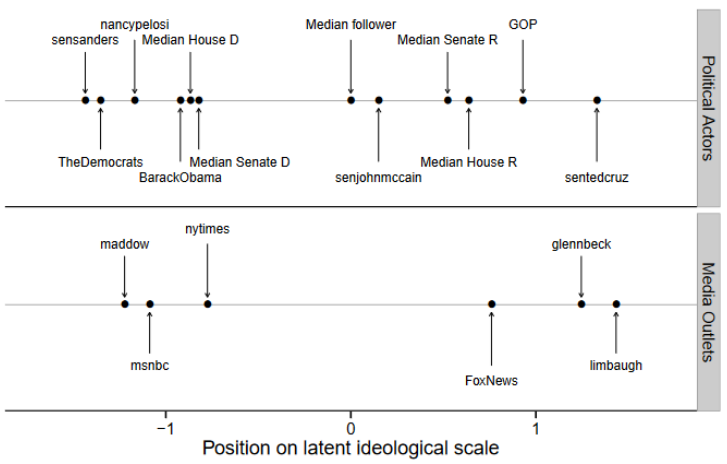
https://www.theguardian.com/environment/2023/jul/11/nuclear-bomb-fallout-site-chosen-to-define-start-of-anthropoce

Let's create the dataset to work with

- The second (***optional***) step is to **transform the extracted data**. This might be needed if the defined measurement requires more than simple counts of URLs.

- Most interesting transformation: enriching the information that URLs bring to research.
  1. The content of the URL can be manually identified, and added to the dataset
  2. The webpages can be classified using external information

Figure S4: Ideology Estimates for Key Political Actors and Media Outlets



Average ideology of participant's media diets

Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., & Bonneau, R. (2015). Tweeting from left to right: Is online political communication more than an echo chamber?. *Psychological science*, *26*(10), 1531-1542.

Let's create the dataset to work with

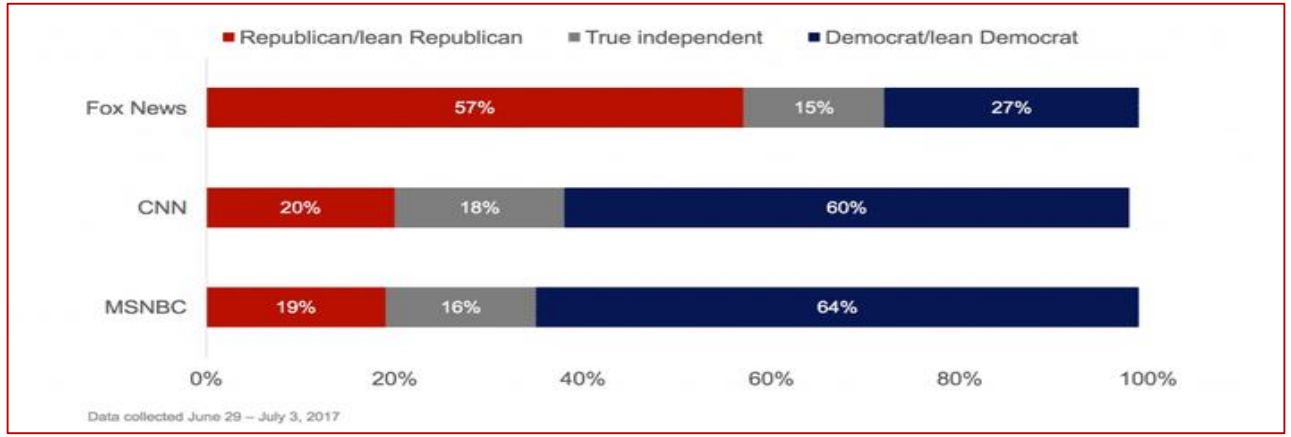
- The second (***optional***) step is to **transform the extracted data**. This might be needed if the defined measurement requires more than simple counts of URLs.

- Most interesting transformation: enriching the information that URLs bring to research.
  1. The content of the URL can be manually identified, and added to the dataset
  2. The webpages can be classified using external information
  3. Machine learning to codify the content exposed to (text / images / video / etc)

Let's create the dataset to work with

- The second (***optional***) step is to **transform the extracted data**. This might be needed if the defined measurement requires more than simple counts of URLs.

- Most interesting transformation: enriching the information that URLs bring to research.
  1. The content of the URL can be manually identified, and added to the dataset
  2. The webpages can be classified using external information
  3. Machine learning to codify the content exposed to (text / images / video / etc)
  4. Measure non-behavioural concepts: e.g., a person's ideology using Correspondence Analysis



Let’s create the dataset to work with

- In the final step the extracted and transformed data sets are ***loaded*** **and stored on the researchers' devices or servers**

Let's create the dataset to work with

- In the final step the extracted and transformed data sets are ***loaded*** **and stored on the researchers' devices or servers**

- All these steps can be done **simultaneously or iteratively** (e.g., extracting information, transforming it, loading it back and extracting it again).

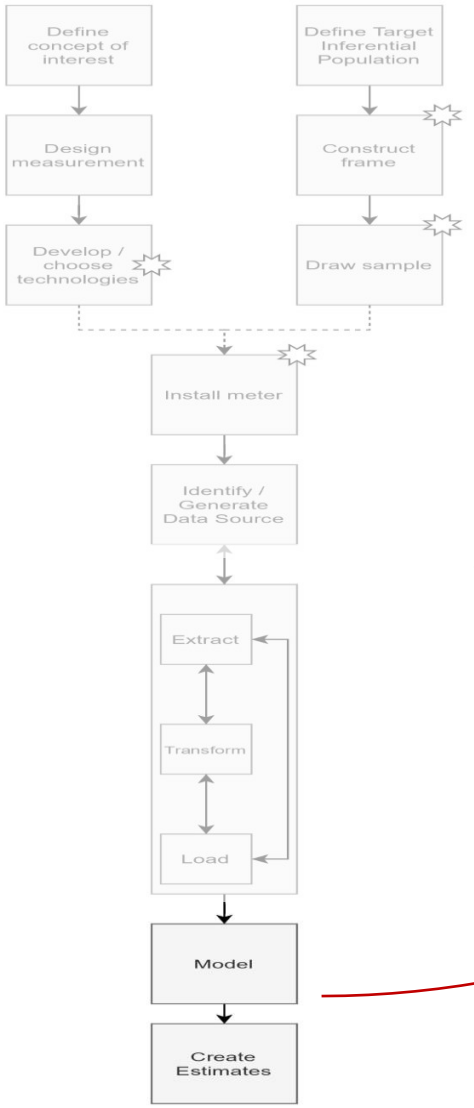
Let's create the dataset to work with

- In the final step the extracted and transformed data sets are ***loaded*** **and stored on the researchers' devices or servers**

- All these steps can be done **simultaneously or iteratively** (e.g., extracting information, transforming it, loading it back and extracting it again).

- This is a big difference compared with surveys, that:
  1. Makes the **pre-processing** stage of the research **harder and longer**
  2. But allows for **immense flexibility**, which can be exploited for good

Modelling and estimating: (for now) same old, same old

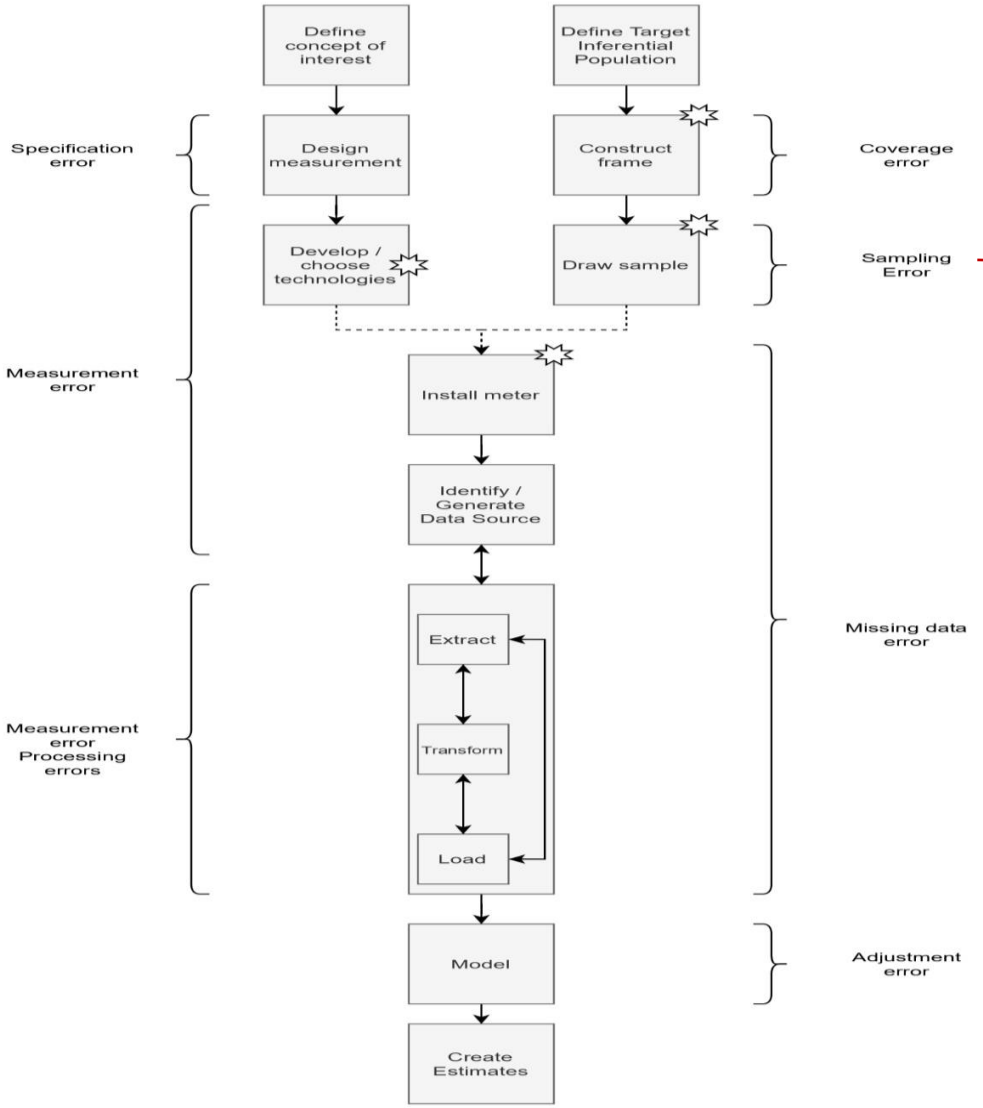


This involves adjusting the data (e.g., weighting and/or imputation). With the adjusted and modelled data, an estimate can be created (e.g., the mean hours of media consumption).

Break for questions

The challenges and errors of web tracking data

Errors can be introduced in every step



Same error components as surveys

What can cause those errors?

Error components	Specific error causes
Specification error	– Defining what qualifies as valid information – Measuring concepts with by-design missing data – Inferring attitudes and opinions from behaviours
Measurement error	– Tracking undercoverage – Technology limitations – Technology errors – Hidden behaviours – Social desirability – Extraction errors – Misclassifying non-observations – Shared devices
Processing error	– Coding error – Aggregation at the domain level – Data anonymization
Coverage error	– Non-trackable individuals
Sampling error	– Same error causes as for surveys
Missing data error	– Non-contact – Non-consent – Tracking undercoverage – Technology limitations – Technology errors – Hidden behaviours – Social desirability – Extraction errors – Misclassifying non-observations
Adjustment error	– Same error causes than for surveys

Bosch, O. J., & Revilla, M. (2022). When survey science met web tracking: presenting an error framework for metered data. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *185*(Supplement_2), S408-S436.

What can cause those errors?

Error components	Specific error causes
Specification error	– Defining what qualifies as valid information – Measuring concepts with by-design missing data – Inferring attitudes and opinions from behaviours
Measurement error	– Tracking undercoverage – Technology limitations – Technology errors – Hidden behaviours – Social desirability – Extraction errors – Misclassifying non-observations – Shared devices
Processing error	– Coding error – Aggregation at the domain level – Data anonymization
Coverage error	– Non-trackable individuals
Sampling error	– Same error causes as for surveys
Missing data error	– Non-contact – Non-consent – Tracking undercoverage – Technology limitations – Technology errors – Hidden behaviours – Social desirability – Extraction errors – Misclassifying non-observations
Adjustment error	– Same error causes than for surveys

Let's deep dive on this error

Bosch, O. J., & Revilla, M. (2022). When survey science met web tracking: presenting an error framework for metered data. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *185*(Supplement_2), S408-S436.

Tracking undercoverage

UNCOVERING DIGITAL TRACE DATA BIASES: TRACKING UNDERCOVERAGE IN WEB TRACKING DATA

Oriol J. Bosch[1, 2, 3], Patrick Sturgis[2], Jouni Kuha[2], Melanie Revilla[4]

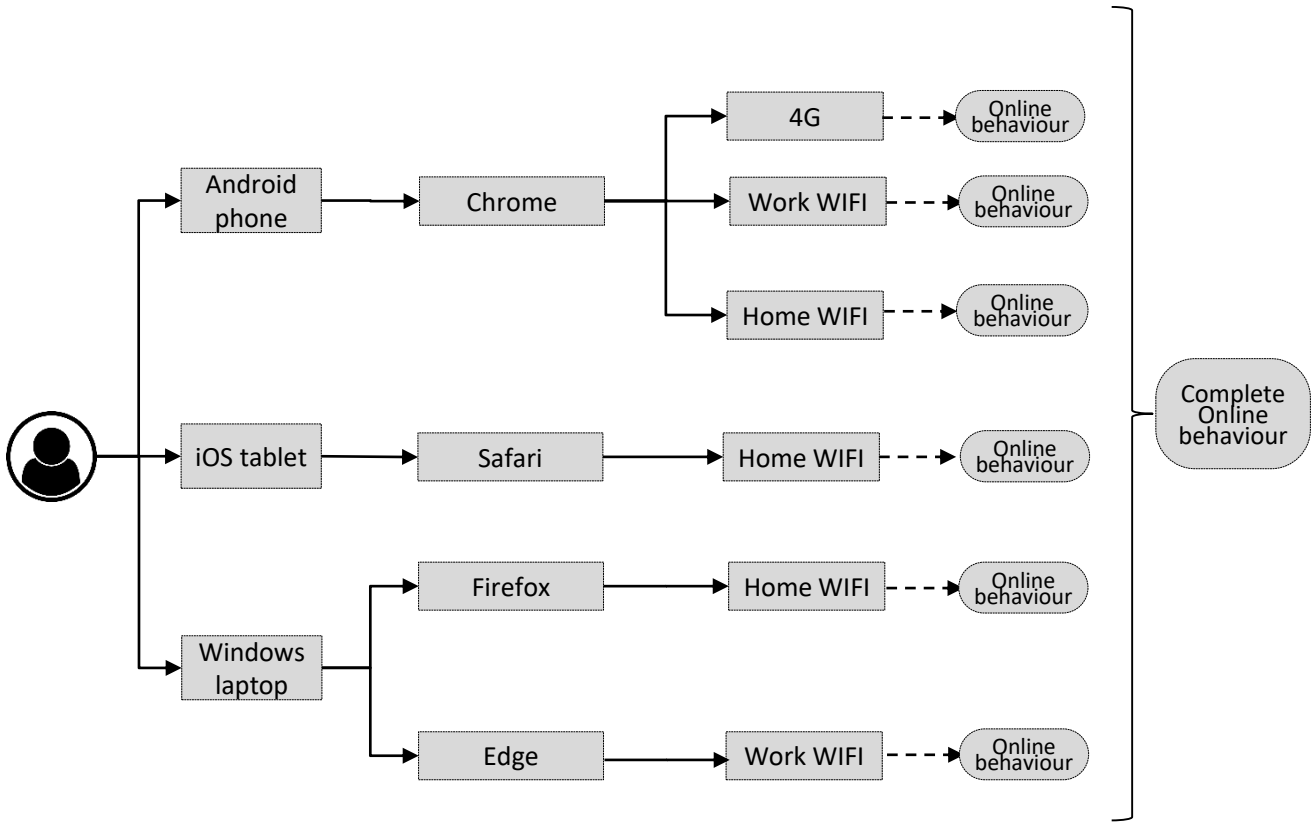
[1] Leverhulme Centre for Demographic Science, University of Oxford
[2] Department of Methodology, The London School of Economics and Political Science
[3] Research and Expertise Centre for Survey Methodology, Universitat Pompeu Fabra
[4] Institut Barcelona Estudis Internacionals (IBEI)

Abstract

In the digital age, understanding people's online behaviours is vital. Digital trace data has emerged as a popular alternative to surveys, many times hailed as the gold standard. This study critically assesses the use of web tracking data to study online media exposure. Specifically, we focus on a critical error source of this type of data, tracking undercoverage: researchers' failure to capture data from all the devices and browsers that individuals utilize to go online. Using data from Spain, Portugal, and Italy, we explore undercoverage in commercial online panels and simulate biases in online media exposure estimates. The paper shows that tracking undercoverage is highly prevalent when using commercial panels, with more than 70% of participants affected. In addition, the primary determinant of undercoverage is the type and number of devices employed for internet access, rather than individual characteristics and attitudes. Additionally, through a simulation study, it demonstrates that web tracking estimates, both univariate and multivariate, are often substantially biased due to tracking undercoverage. This represent the first empirical evidence demonstrating that web tracking data is, effectively, biased. Methodologically, the paper showcases how survey questions can be used as auxiliary information to identify and simulate web tracking errors.

Keywords:
Digital trace data · Web tracking data · Undercoverage · Bias · Media exposure · Monte Carlo simulation

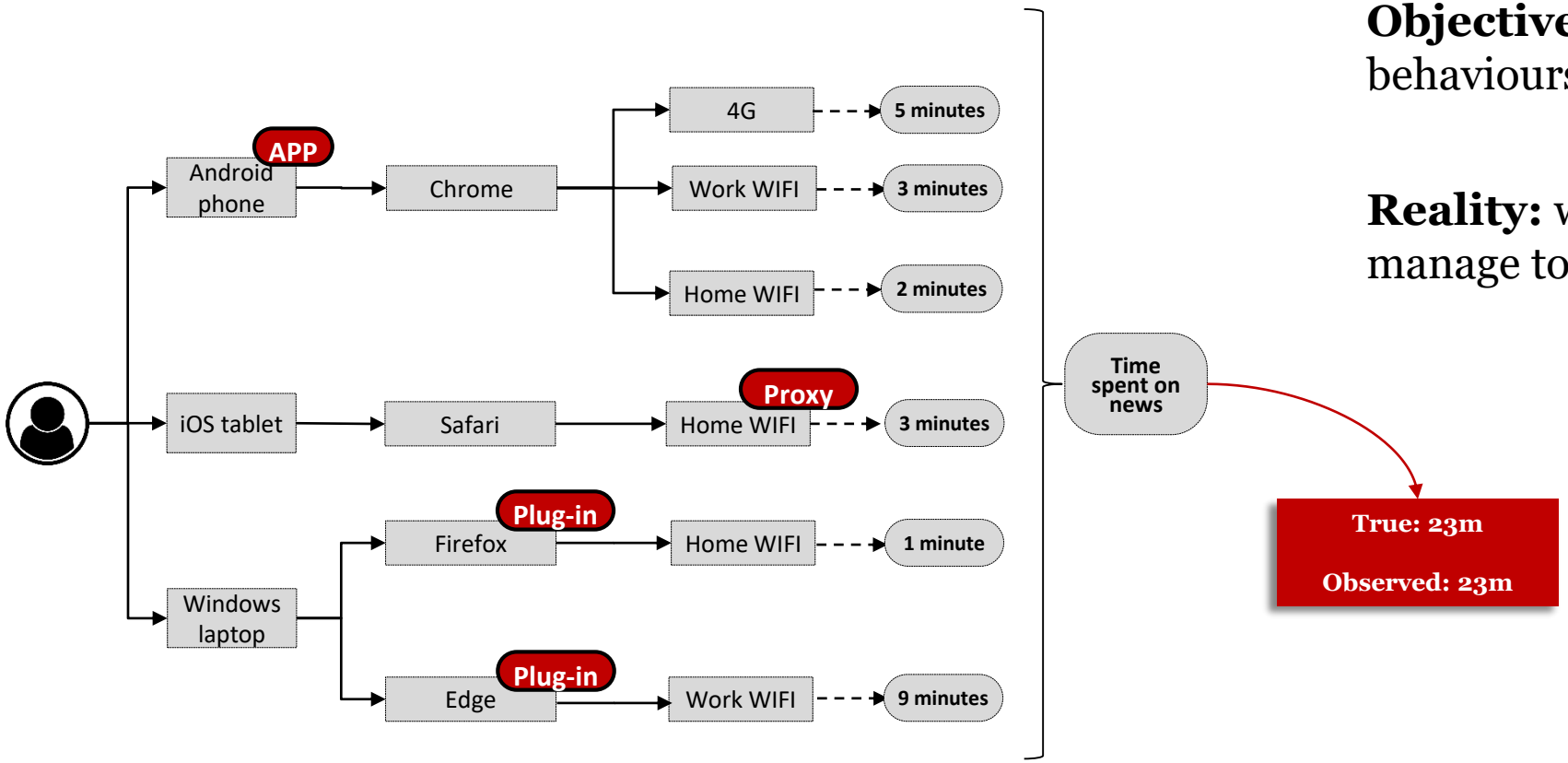
Tracking undercoverage



Objective: measuring individuals' behaviours.

Reality: we only measure what we can manage to track.

Tracking undercoverage



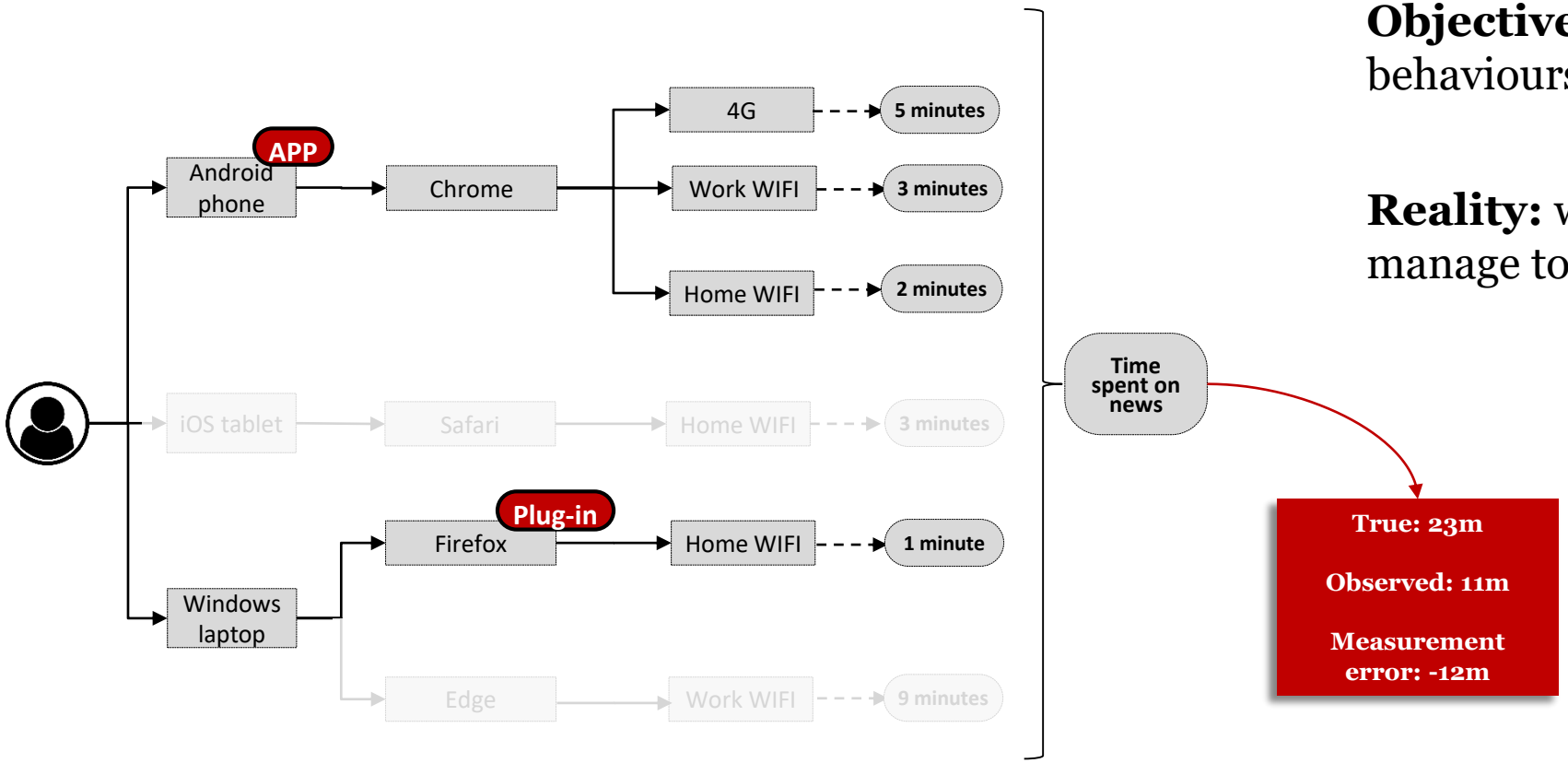
Objective: measuring individuals' behaviours.

Reality: we only measure what we can manage to track. → **All**

True: 23m

Observed: 23m

Tracking undercoverage



Objective: measuring individuals' behaviours.

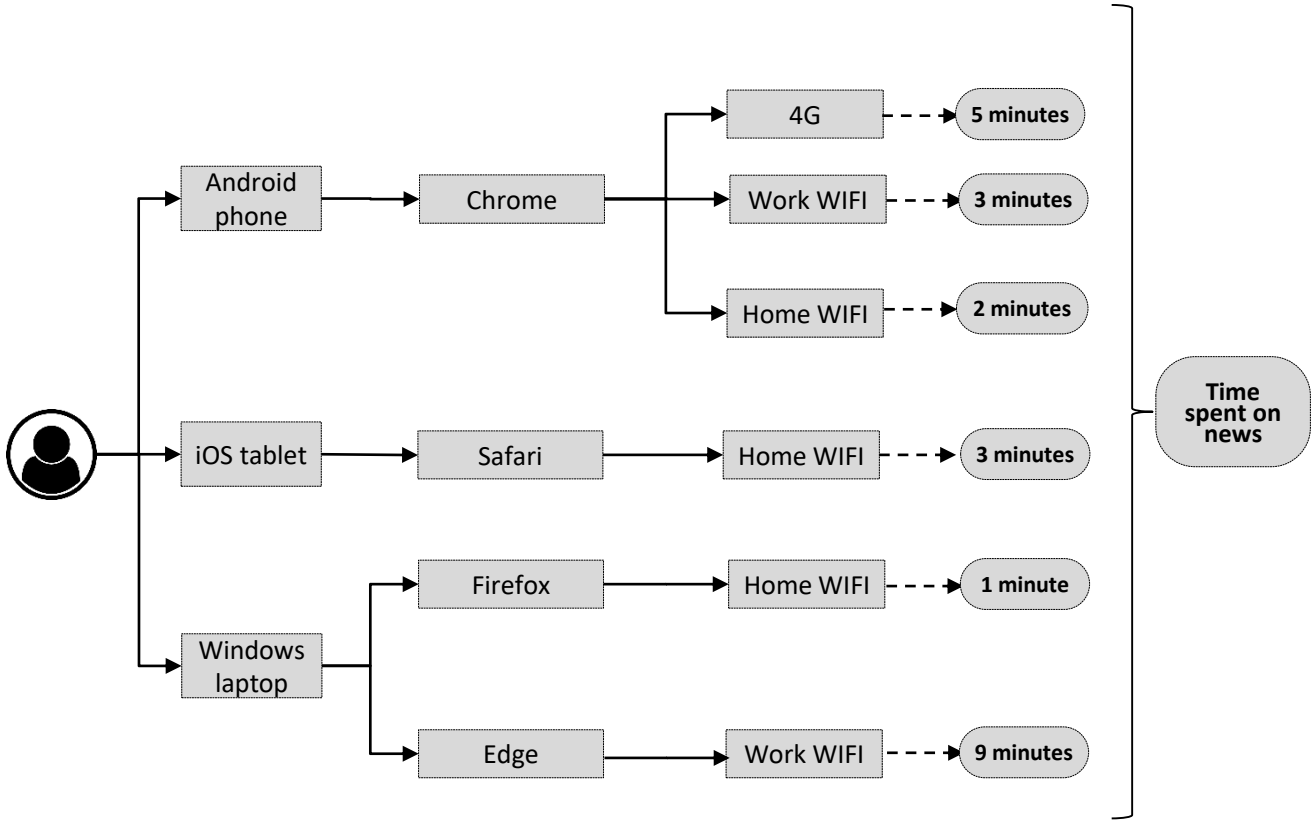
Reality: we only measure what we can manage to track. → **Part**

True: 23m

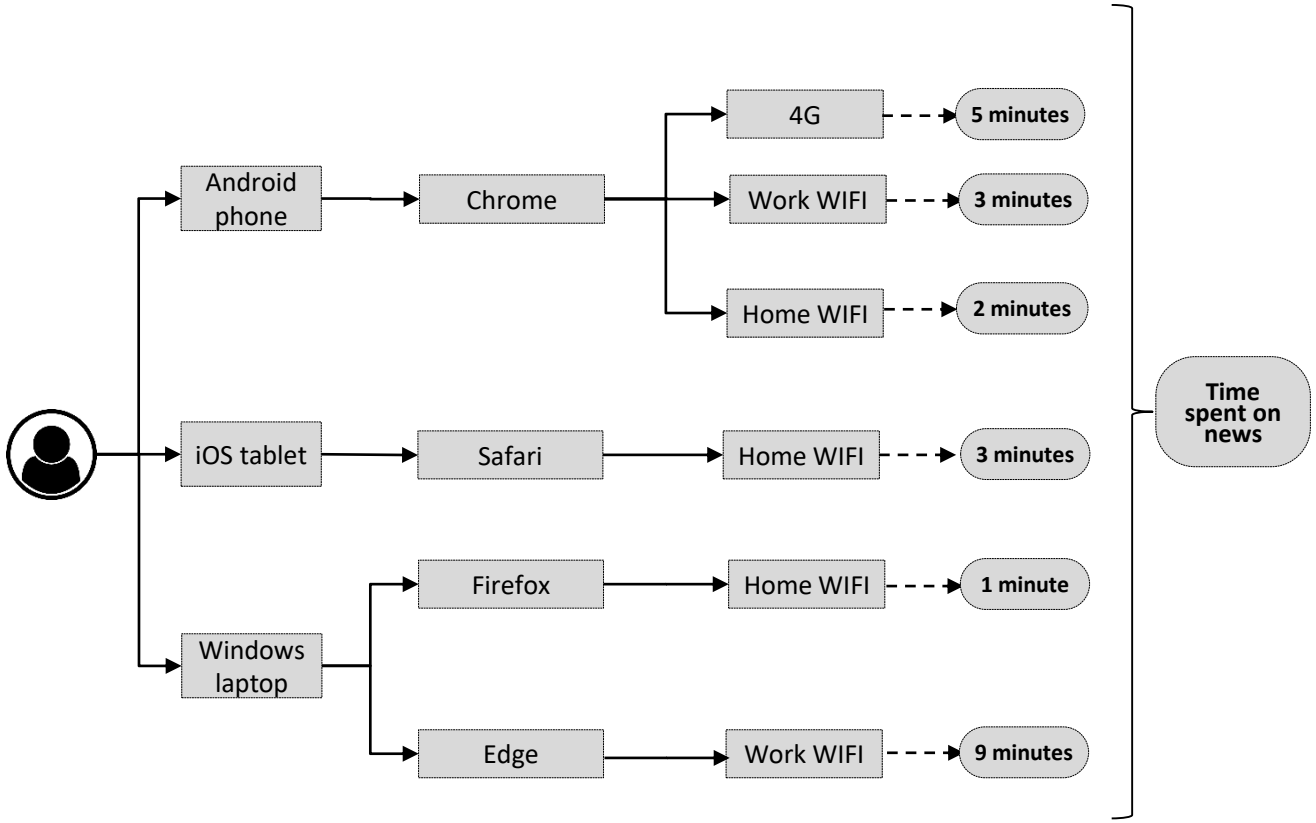
Observed: 11m

Measurement error: -12m

Why is this happening?

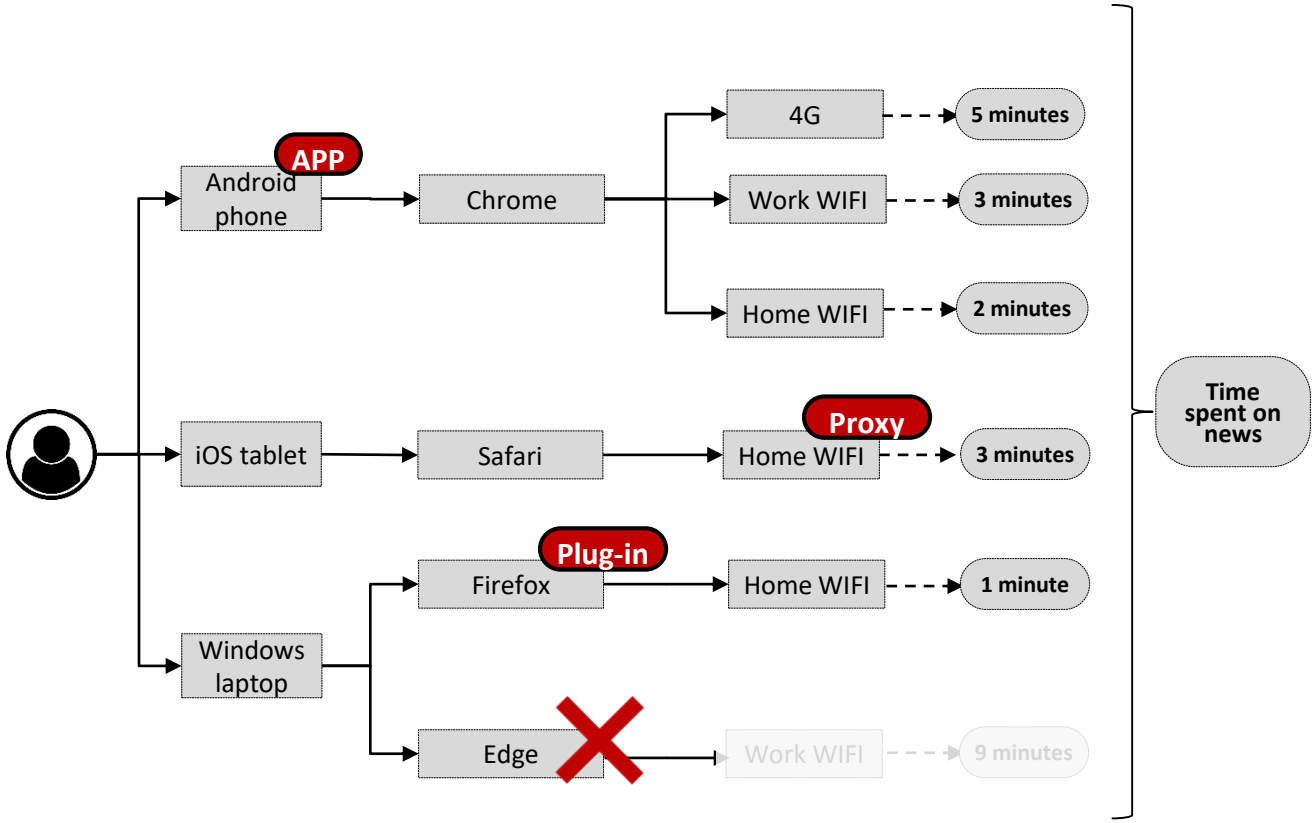


Why is this happening?



Different reasons:

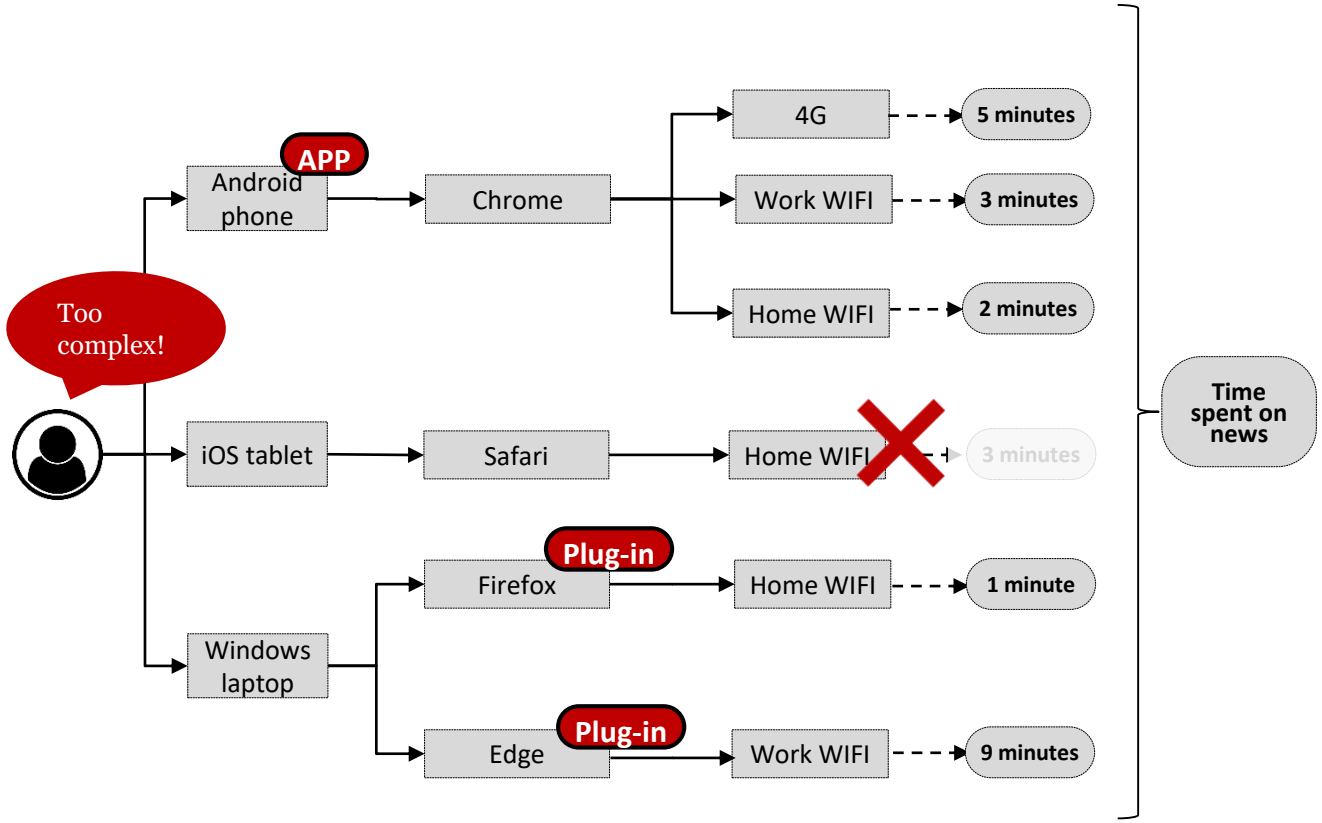
Why is this happening?



Different reasons:

1. Some devices / browsers **cannot be tracked with available technologies**

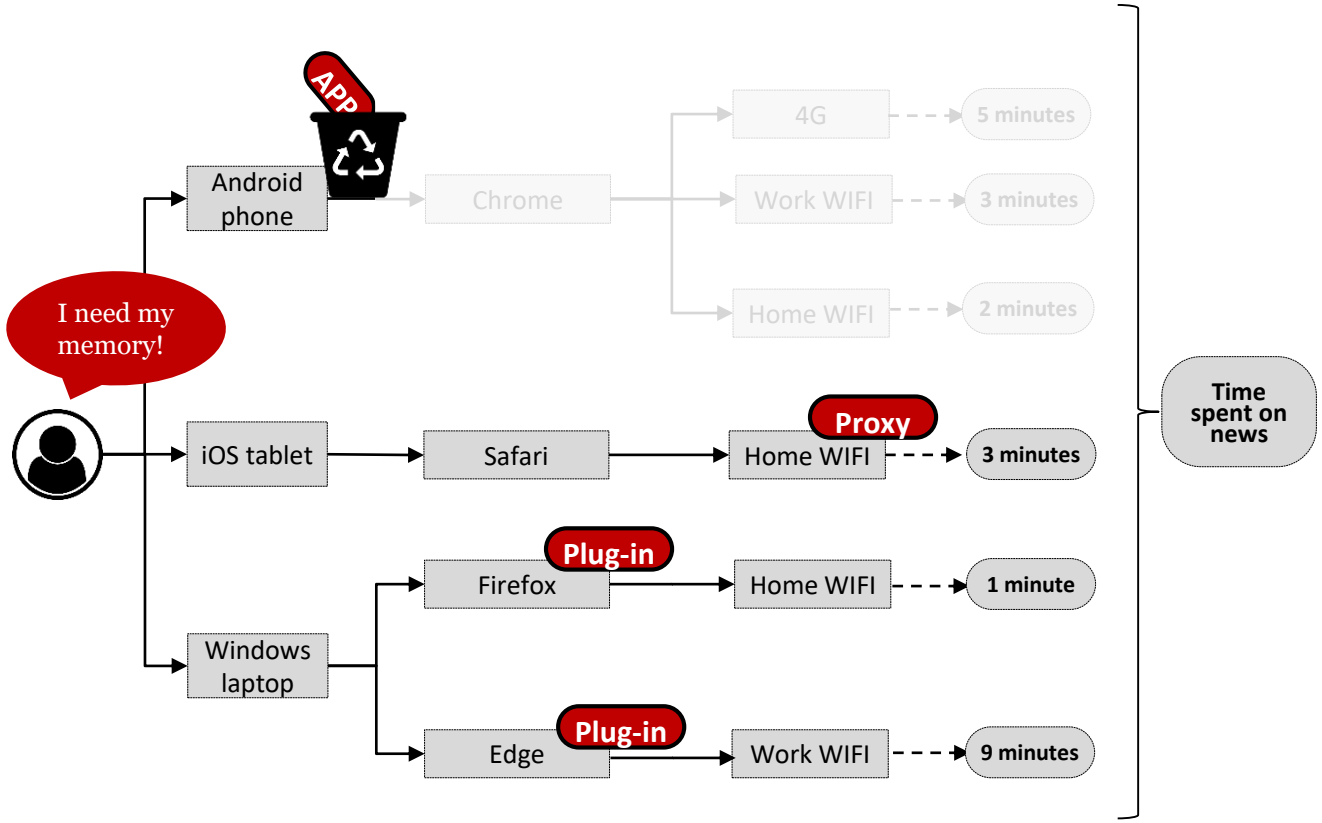
Why is this happening?



Different reasons:

1. Some devices / browsers **cannot be tracked with available technologies**
2. People might **not want to fully comply**

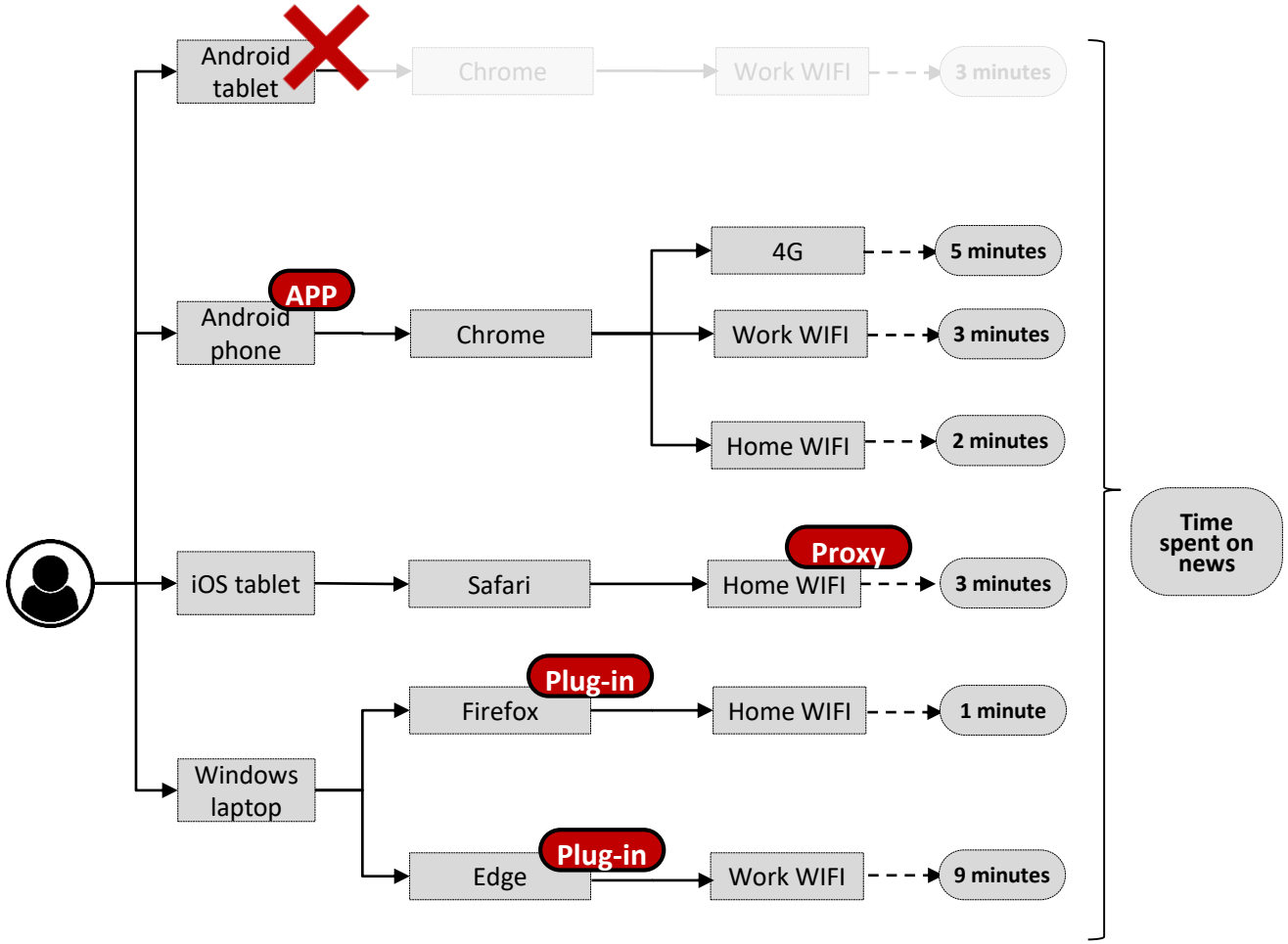
Why is this happening?



Different reasons:

1. Some devices / browsers **cannot be tracked with available technologies**
2. People might **not want to fully comply**
3. People might **uninstall technologies**

Why is this happening?



Different reasons:

1. Some devices / browsers **cannot be tracked with available technologies**
2. People might **not want to fully comply**
3. People might **uninstall technologies**
4. **New device,** we do not even know they have

How big of a problem is this?

Proportion of participants with all their devices tracked

	% fully covered
All participants	26
Participants who reported using…	
… 1 device	100
… 2 devices	34
… 3 devices	13
… 4 devices	1
… +5 devices	0

	% fully covered
Participants who reported using…	
PC	
…Windows	49
…MAC	27
Mobile	
… Android	52
… iOS	10

Bosch, O. J., Sturgis, P., Kuha, J., & Revilla, M. (2023). Uncovering digital trace data biases: tracking undercoverage in web tracking data. https://doi.org/10.31235/osf.io/t2dbj

How big of a problem is this?

Most people do not have all their devices fully tracked

Proportion of participants with all their devices tracked

	% fully covered
All participants	26
Participants who reported using…	
… 1 device	100
… 2 devices	34
… 3 devices	13
… 4 devices	1
… +5 devices	0

	% fully covered
Participants who reported using…	
PC	
…Windows	49
…MAC	27
Mobile	
… Android	52
… iOS	10

Bosch, O. J., Sturgis, P., Kuha, J., & Revilla, M. (2023). Uncovering digital trace data biases: tracking undercoverage in web tracking data. https://doi.org/10.31235/osf.io/t2dbj

How big of a problem is this?

Proportion of participants with all their devices tracked

	% fully covered
All participants	26
Participants who reported using…	
… 1 device	100
… 2 devices	34
… 3 devices	13
… 4 devices	1
… +5 devices	0

	% fully covered
Participants who reported using…	
PC	
…Windows	49
…MAC	27
Mobile	
… Android	52
… iOS	10

The higher the number of devices that people use, the more likely it is that we do not fully track them

Bosch, O. J., Sturgis, P., Kuha, J., & Revilla, M. (2023). Uncovering digital trace data biases: tracking undercoverage in web tracking data. https://doi.org/10.31235/osf.io/t2dbj

How big of a problem is this?

Proportion of participants with all their devices tracked

	% fully covered
All participants	26
Participants who reported using…	
… 1 device	100
… 2 devices	34
… 3 devices	13
… 4 devices	1
… +5 devices	0

	% fully covered
Participants who reported using…	
PC	
…Windows	49
…MAC	27
Mobile	
… Android	52
… iOS	10

We have a problem with Apple devices! (tech reasons)

Bosch, O. J., Sturgis, P., Kuha, J., & Revilla, M. (2023). Uncovering digital trace data biases: tracking undercoverage in web tracking data. https://doi.org/10.31235/osf.io/t2dbj

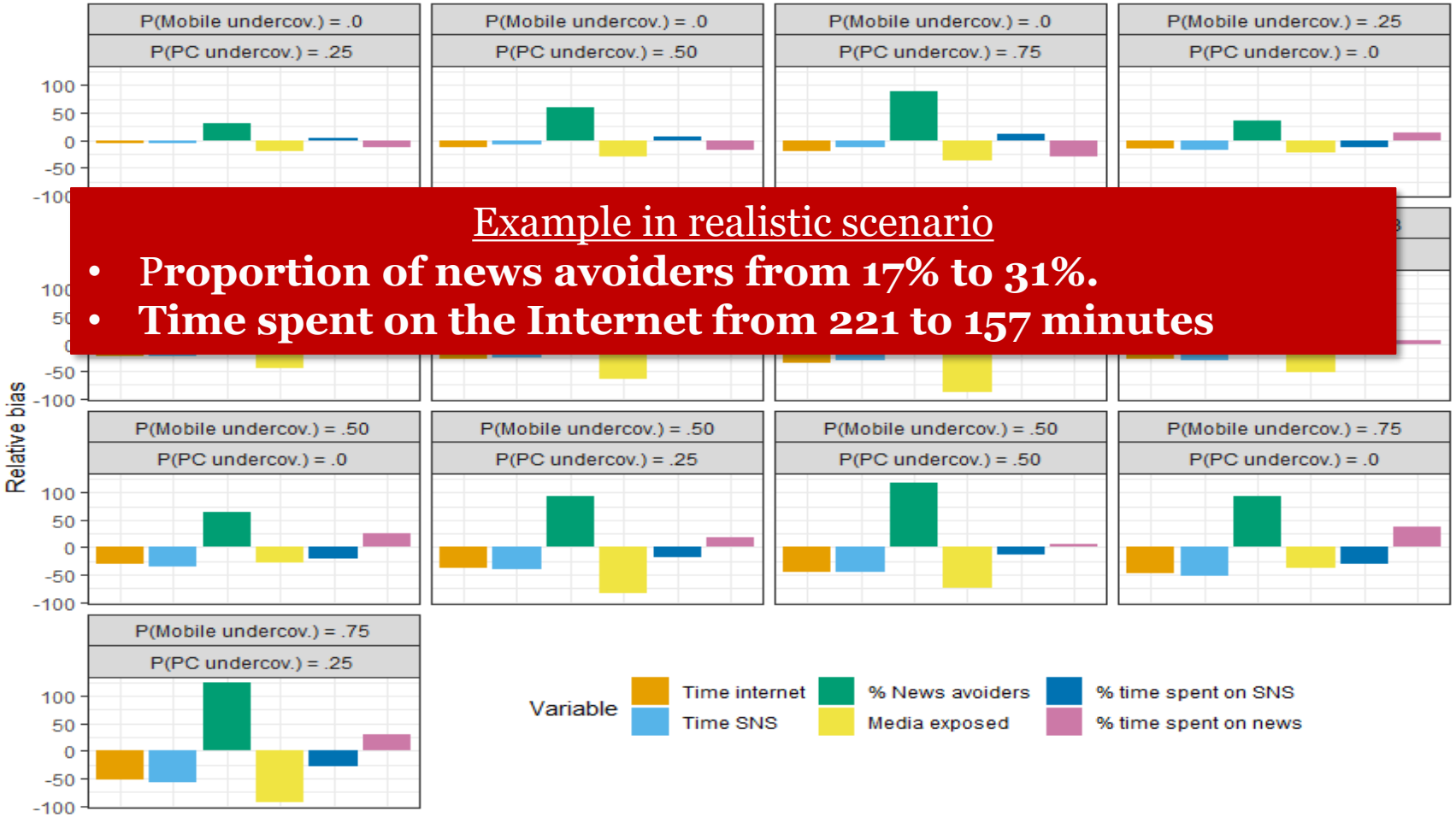
How big of a problem is this?

Relative bias introduced by undercoverage, depending on the probability of having all PCs or Mobile devices not covered

Bosch, O. J., Sturgis, P., Kuha, J., & Revilla, M. (2023). Uncovering digital trace data biases: tracking undercoverage in web tracking data. https://doi.org/10.31235/osf.io/t2dbj

How big of a problem is this?

Relative bias introduced by undercoverage, depending on the probability of having all PCs or Mobile devices not covered



Example in realistic scenario

- **Proportion of news avoiders from 17% to 31%.**
- **Time spent on the Internet from 221 to 157 minutes**

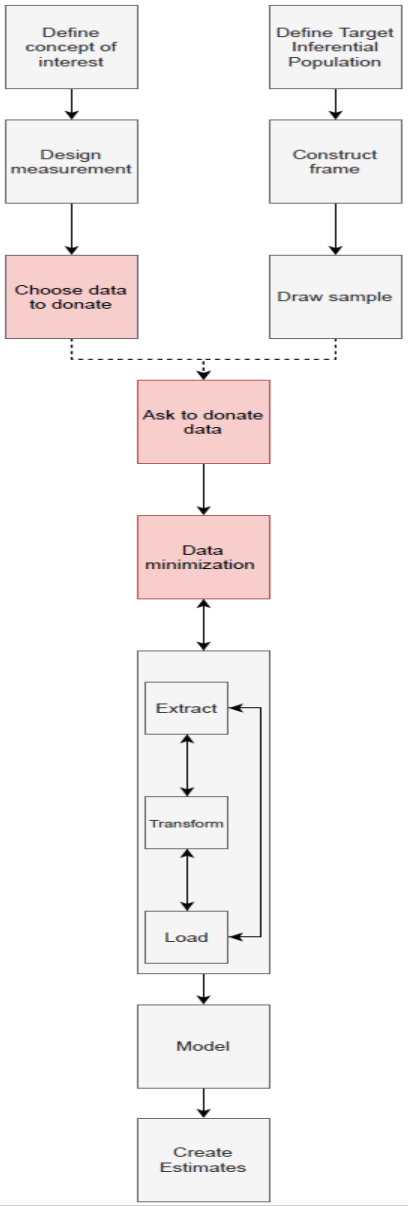
Bosch, O. J., Sturgis, P., Kuha, J., & Revilla, M. (2023). Uncovering digital trace data biases: tracking undercoverage in web tracking data. https://doi.org/10.31235/osf.io/t2dbj

How to use data donations, and what to consider?

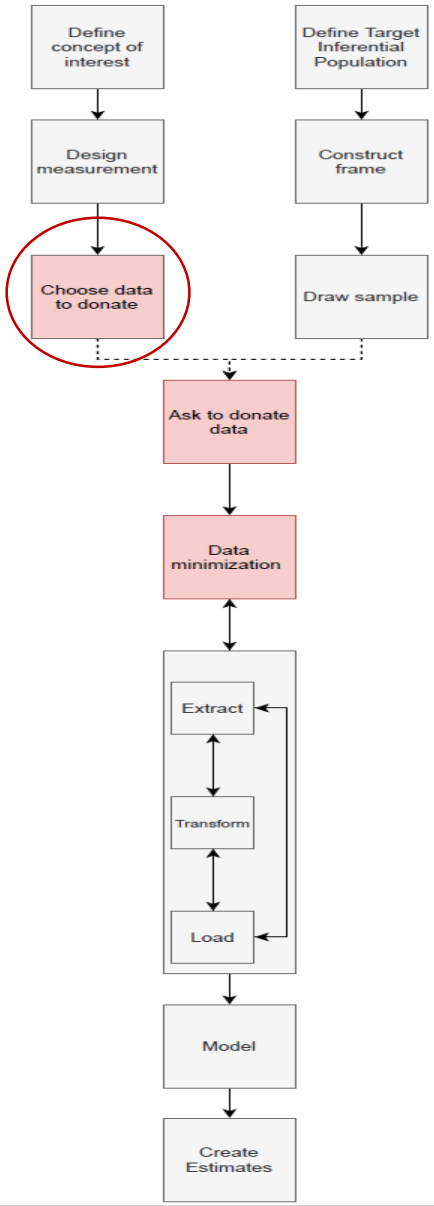
Refreshing our memory

A data donation is any instance in which a person accesses some of their personal data, captures it, and shares it with researchers.

Similar, but different

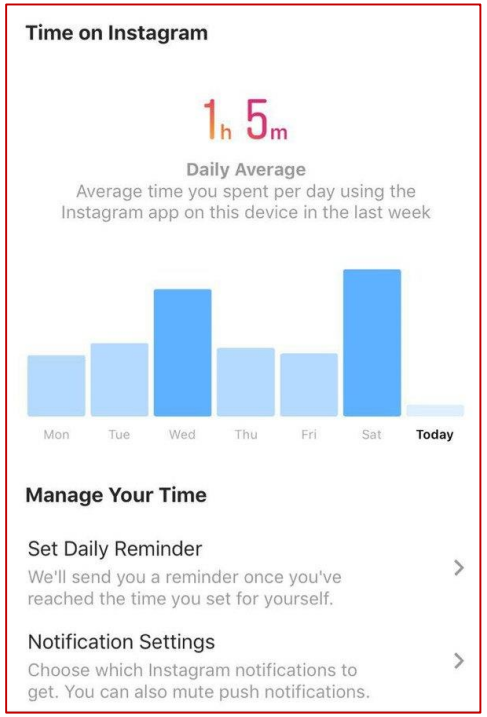


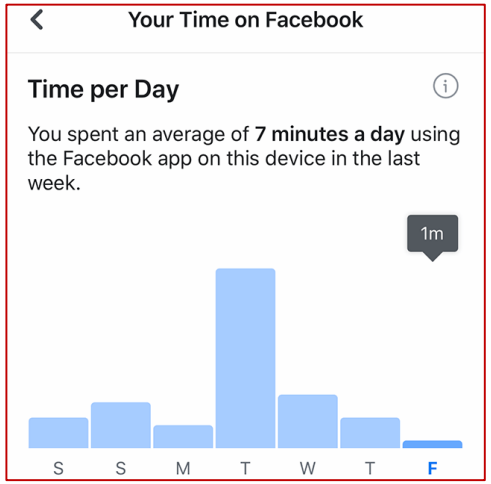
Similar, but different

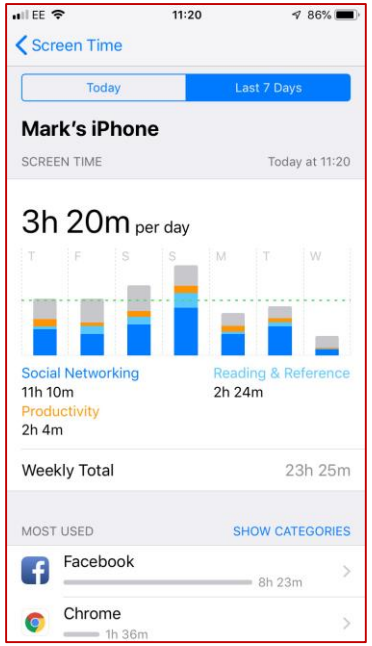


We need to identify what available data might work for us

- With data donations, we are asking participants to **donate data that has already been produced by third-parties**, and that they have access to







We need to identify what available data might work for us

- With data donations, we are asking participants to **donate data that has already been produced by third-parties**, and that they have access to

- To measure a specific concept with data donations, we first need to identify whether there is any **available data** source that participants can access, capture, and share.

We need to identify what available data might work for us

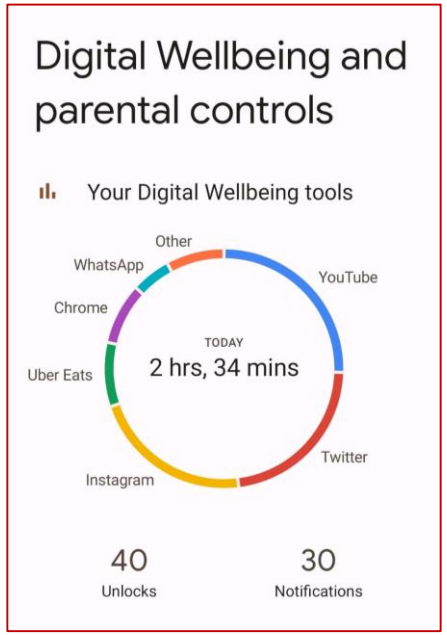
- With data donations, we are asking participants to **donate data that has already been produced by third-parties**, and that they have access to

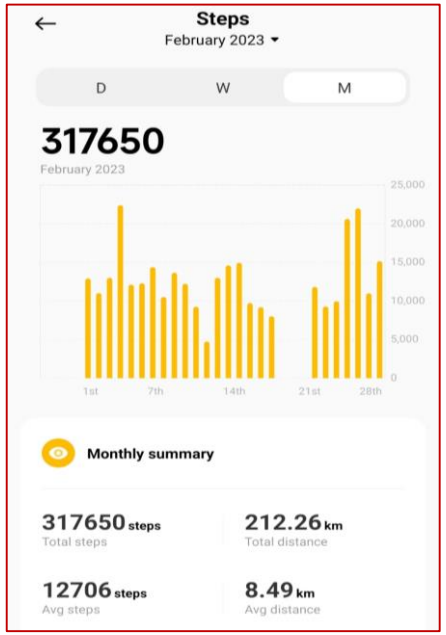
- To measure a specific concept with data donations, we first need to identify whether there is any **available data** source that participants can access, capture, and share.

We are constrained by what other companies have created and collected. We have no control over what data might exist, and the format of it

Examples of available data (related to digital behaviours)

- Information collected and stored by digital devices. Examples could be:
  1. **Device, battery and/or memory usage information.**
  2. **Activity and health data.**

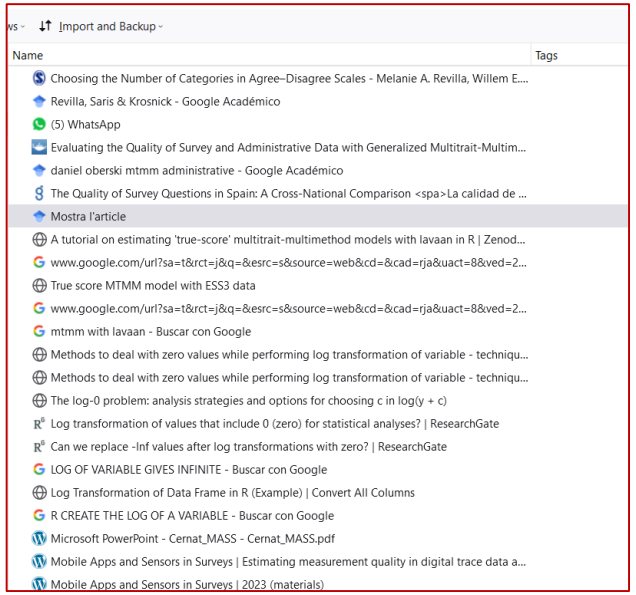


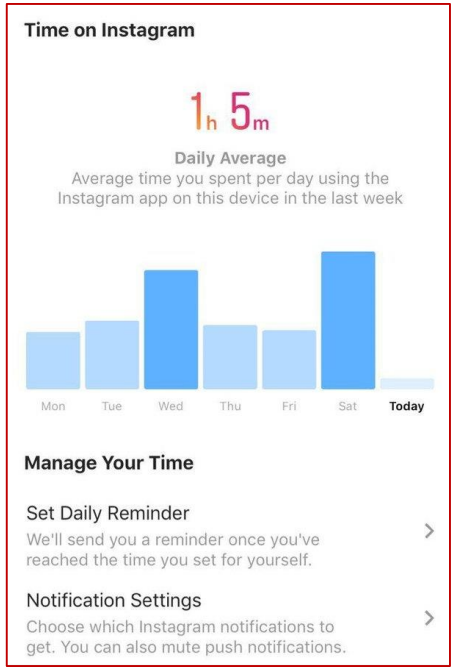


Examples of available data (related to digital behaviours)

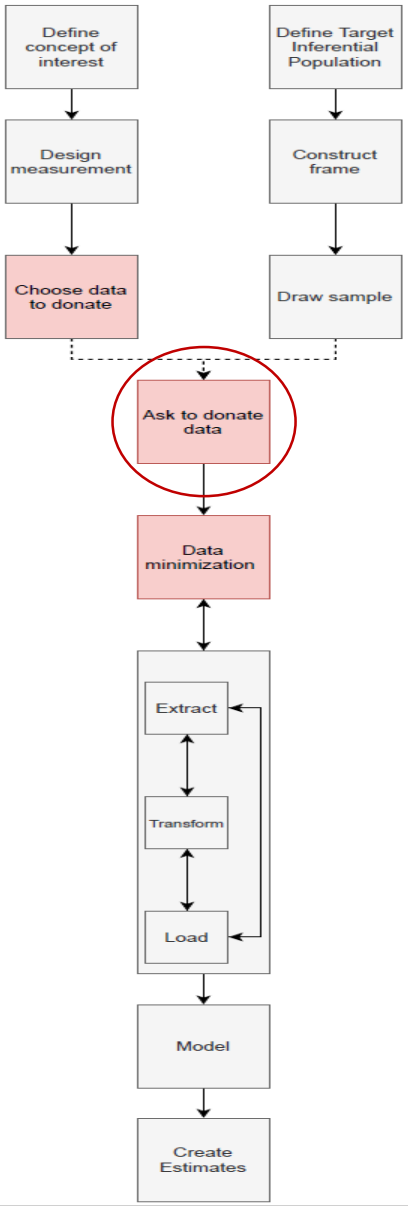
- Information collected and stored by digital devices. Examples could be:
  1. **Device, battery and/or memory usage information.**
  2. **Activity and health data.**

- Information collected and stored by tech companies. Examples could be:
  1. **Browsing history.**
  2. **Social media usage.**
  3. **Location and travel data.**
  4. **Advertisement data.**





Similar, but different



Asking to donate data

- In most cases, when a participant is asked to donate their data, there will always be at least three steps:

  1. Access the data
  2. Capture it
  3. And share it with the researchers

Baumgartner, S. E., Sumter, S. R., Petkevič, V., & Wiradhany, W. (2022). A novel iOS data donation approach: Automatic processing, compliance, and reactivity in a longitudinal study. *Social Science Computer Review*, 08944393211071068.

Asking to donate data

- In most cases, when a participant is asked to donate their data, there will always be at least three steps:

  1. Access the data
  2. Capture it
  3. And share it with the researchers

Goal: make design decisions across these three dimensions that **minimises the required effort** of participants to share data, **while allowing us to collect the necessary data**

How can participants capture and share their data?

Capture

- Take pictures or screenshots
- Take videos or video recordings
- Download the information
- Manually annotate the data / memorize (not ideal).

Share

- Upload within the questionnaire.
- Upload in an outside system.
- Send the data using e-mails or secure sharing systems.
- Manually record the data.

How can participants capture and share their data?

- **The process to capture and share** this data will heavily **vary depending the approaches** selected for the project

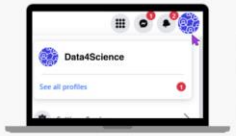
How can participants capture and share their data?

- **The process to capture and share** this data will heavily **vary depending the approaches** selected for the project
  - For instance: downloading a **Data Download Package (DDP)** can be a long and burdensome process. Can take more than one day from the point that the participants asks for the data, and the data is available.

FREE YOUR FACEBOOK DATA

Step 1: Download your data

DATA 4 SCIENCE
Be part of a movement



1 Login to Facebook

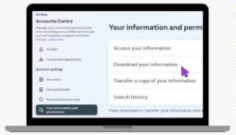
Click on your **profile photo** on the top right of the screen

2 Select 'Settings and Privacy'

Click **Settings** on the dropdown menu

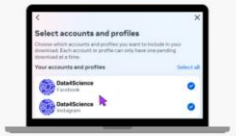
3 Go to 'Accounts Centre'

Click on **Your information and permissions**, then **Download your information**



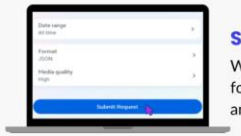
4 Select 'Request a download'

Click on the **Facebook and Instagram** profiles you'd like to download, then **Complete copy**



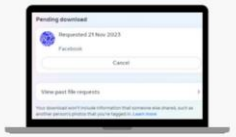
5 Select file options

We suggest you choose **JSON** format, **High** media quality, and **All time** date range



6 Click 'Submit request'

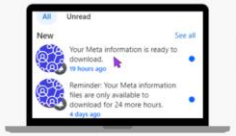
Once you receive a Facebook notification/email, you'll have four days to download your data



FREE YOUR FACEBOOK DATA

Step 2: Free your data!

DATA 4 SCIENCE
Be part of a movement

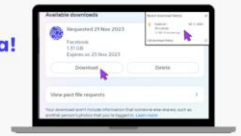


1 Your data is ready

Click on the **Facebook notification or email**

2 Download your data!

Select **Download** and **save the .zip file** to your PC

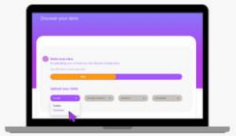


3 Login to Data4Science

Select **Dashboard** on the top menu

4 Go to Facebook

Hover over **Socials** and click on **Facebook**



5 Upload your data!

Drag and drop the .zip file into the box and click on **Submit**

6 Explore your data

Select a chart and start exploring your data!

web data opp

How can participants capture and share their data?

- **The process to capture and share** this data will heavily **vary depending the approaches** selected for the project
  - For instance: downloading a Data Download Package (DDP) can be a long and burdensome process. Can take more than one day from the point that the participants asks for the data, and the data is available.
- Similarly, the **amount of data collectable** and the **perceived privacy concerns** might potentially be affected by the method used.

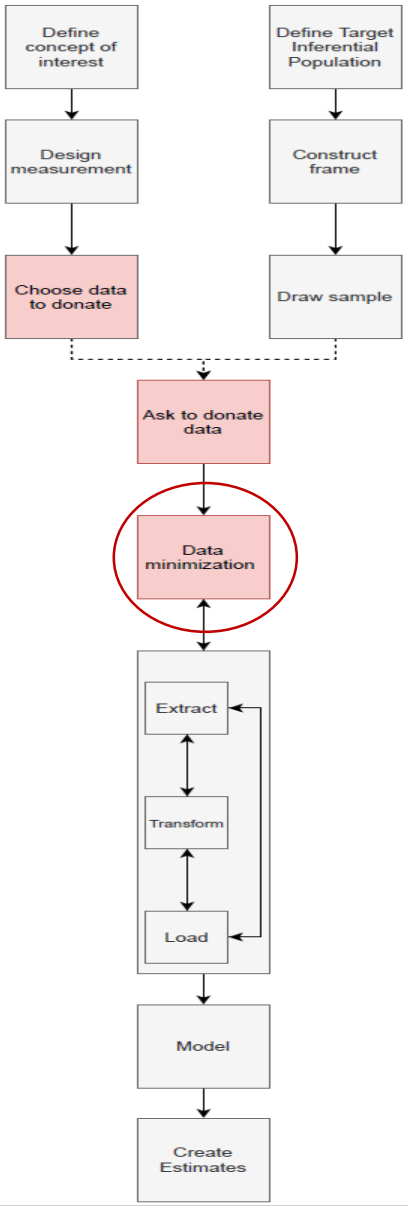
How can participants capture and share their data?

- **The process to capture and share** this data will heavily **vary depending the approaches** selected for the project
  - For instanc ng and burdensome process. Ca s asks for the data, and the dat
- Similarly, the **a oncerns** might potentially be a

Sometimes we might not be able to choose! Some data can only be captured in specific ways.

For example: device usage data cannot (most of the times) be downloaded in any way

Similar, but different



Minimizing the data

- In most cases, the data that people will donate will be plagued with sensitive information
- Data must be minimized, either locally, or before being saved in the servers, to make sure no unintended and sensitive data is collected
- This will normally involve a lot of complex coding, so get ready

Q & A

Thanks!

Oriol J. Bosch | Postdoctoral Researcher, University of Oxford

oriol.bosch-jover@demography.ox.ac

orioljbosch

https://orioljbosch.com/

UNIVERSITY OF OXFORD

LEVERHULME CENTRE FOR DEMOGRAPHIC SCIENCE

RECSM Research and Expertise Centre for Survey Methodology

web data opp