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

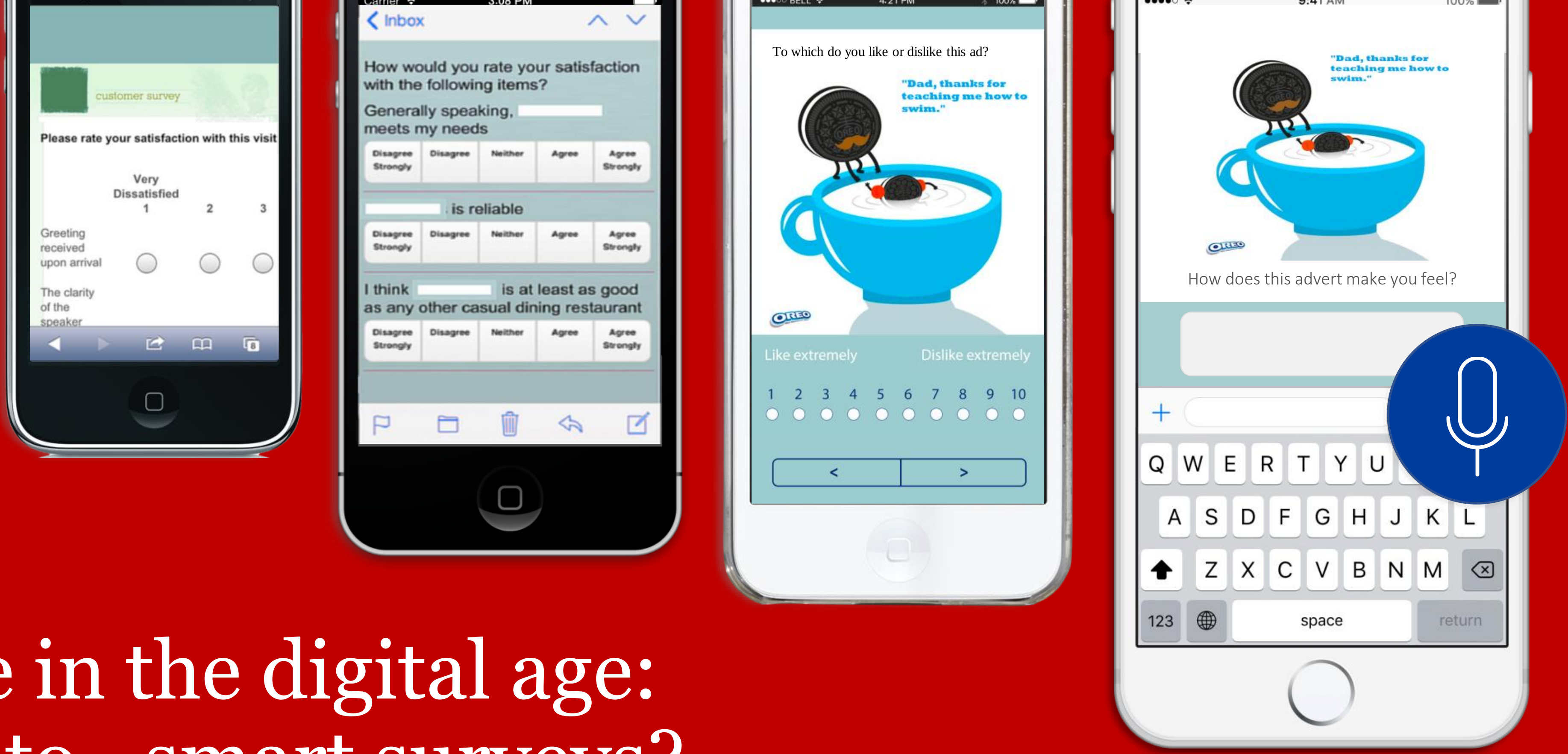
Oriol J. Bosch | Department of Methodology, LSE & RECSM

o.bosch-jover@lse.ac.uk
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).

Who am I?

- PhD Candidate at the **Methodology Department, LSE**
- Upcoming postdoc at the **Leverhulme Centre for Demographic Science, University of Oxford**
- Non-resident research fellow at the **Research and Expertise Centre for Survey Methodology, UPF**
- MSc in Survey Methods for Social Research from the **University of Essex**
- Worked for the **The Alan Turing Institute , University of Southampton, Institute for Social and Economic Research, ESS and Netquest**
- Consultant for the **Wellcome Trust, Social Care Institute for Excellence and MoneyHelper**



Social science in the digital age: from surveys to…smart surveys?

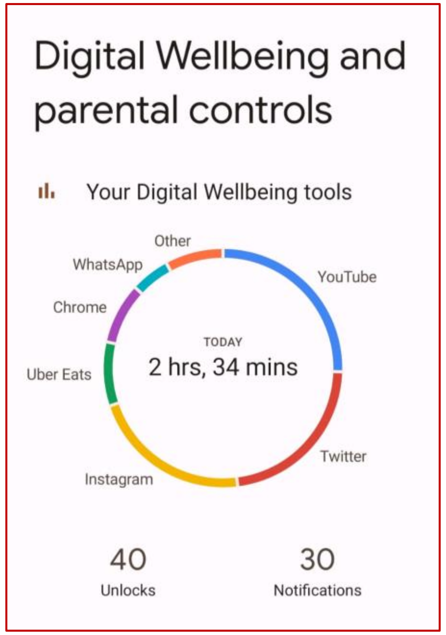
Things have changed…

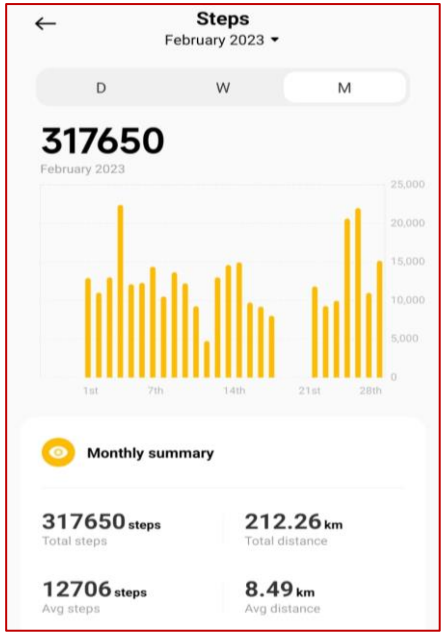
1. What people do on the digital realm can impact both online and offline phenomena.

Things have changed…

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





Things have changed…

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

Mandate: social scientists must find ways to measure the digital behaviour of people

Opportunity: we can directly "observe" what people do in the digital realm

…but surveys are (still) relevant

web data opp

- Some have disregarded surveys and crowned big data as the new gold standard in town.
- But surveys are not only still relevant, but potentially even more important than ever.

Table 3. Different types of quantitative data by discipline, 2014–2015.

Discipline	Survey	Admin	Census	Big data	n
Sociology	51%	42%	16%	4%	277
Political Sciences	41%	58%	9%	4%	308
Economics	32%	74%	19%	3%	374
Social Psychology	69%	5%	0%	2%	235
Public Opinion	86%	16%	3%	5%	81
TOTAL	**49%**	**47%**	**11%**	**3%**	**1275**

Sturgis, P., & Luff, R. (2021). **The demise of the survey? A research note on trends in the use of survey data in the social sciences**, 1939 to 2015. *International Journal of Social Research Methodology*, *24*(6), 691-696.

They might be smart, but they are still surveys

- In this course we will discuss how to properly measure citizens' digital behaviours in the context of smart surveys

They might be smart, but they are still surveys

- In this course we will discuss how to properly measure citizens' digital behaviours in the context of smart surveys

- What is a smart survey?

They might be smart, but they are still surveys

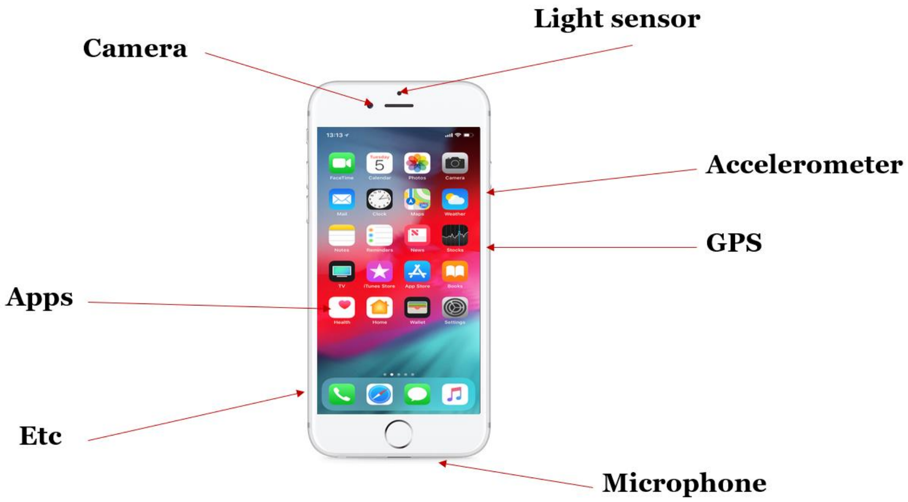
- In this course we will discuss how to properly measure citizens' digital behaviours in the context of smart surveys

- What is a smart survey?
    1. You guessed it, **it is a survey**!

They might be smart, but they are still surveys

- In this course we will discuss how to properly measure citizens' digital behaviours in the context of smart surveys

- What is a smart survey?
  1. You guessed it, **it is a survey**!
  2. It is **linked to data coming from smart devices, or digital systems**. These can be, for instance, mobile device sensors, data donations, tracking apps, linkage to external sensor systems (e.g., Fitbit).



They might be smart, but they are still surveys

webdataopp

- In this course we will discuss how to properly measure citizens' digital behaviours in the context of smart surveys

- What is a smart survey?
  1. You guessed it, **it is a survey**!
  2. It is **linked to data coming from smart devices, or digital systems**. These can be, for instance, mobile device sensors, data donations, tracking apps, linkage to external sensor systems (e.g., Fitbit).

Much of what we know about surveys still apply! We need to sample participants, convince them to participate, and make sure our measurements are valid and reliable

They might be smart, but they are still surveys

web data opp

- In this course we will discuss how to properly measure citizens' digital behaviours in the context of smart surveys

- What is a smart survey?
  1. You guessed it, **it is a survey**!
  2. It is **linked to data coming from smart devices, or digital systems**. These can be, for instance, mobile device sensors, data donations, tracking apps, linkage to external sensor systems (e.g., Fitbit).

Much of what we know about surveys still apply! We need to sample participants, convince them to participate, and make sure our measurements are valid and reliable

But there are new challenges. We need new knowledge, and new practices…that's why we are here!

Why smart surveys?

Researchers

- Reduce measurement issues (e.g., objective)
- Provide new data
- Massive and granular
- Real time

Participants

- Reduce time
- Reduce efforts
- More enjoyable

Measuring what people do online with smart surveys

web data opp

- **Two main approaches** have been used to enhance survey data with digital trace data about people's online behaviours: **web trackers and data donations**

ARTICLES
https://doi.org/10.1038/s41562-020-0833-x
nature human behaviour

Check for updates

Exposure to untrustworthy websites in the 2016 US election

Andrew M. Guess[1], Brendan Nyhan[2] and Jason Reifler[3]

Although commentators frequently warn about echo chambers, little is known about the volume or slant of political misinformation that people consume online, the effects of social media and fact checking on exposure, or the effects of political misinformation on behaviour. Here, we evaluate these questions for websites that publish factually dubious content, which is often described as fake news. Survey and web-traffic data from the 2016 US presidential campaign show that supporters of Donald Trump were most likely to visit these websites, which often spread through Facebook. However, these websites made up a small share of people's information diets on average and were largely consumed by a subset of Americans with strong preferences for pro-attitudinal information. These results suggest that the widespread speculation about the prevalence of exposure to untrustworthy websites has been overstated.

Leveraging Rights of Data Subjects for Social Media Analysis: Studying TikTok via Data Donations

Savvas Zannettou[1], Olivia-Nemes Nemeth[2], Oshrat Ayalon[2], Angelica Goetzen[2], Krishna P. Gummadi[2], Elissa M. Redmiles[2], and Franziska Roesner[3]
[1]TU Delft, [2]Max Planck Institute for Software Systems, [3]University of Washington
s.zannettou@tudelft.nl, {onemes, oayalon, agoetzen, gummadi, eredmiles}@mpi-sws.org, franzi@cs.washington.edu

Abstract

TikTok is a relatively novel and widely popular media platform. In response to its expanding user base and cultural impact, researchers are turning to study the platform; however, TikTok, like many social media platforms, restricts external access to data. Prior works have acquired data from scraping the platform, user self-reports, and from accounts created by researchers for the study's purpose. Existing techniques, while yielding important insights, contain limitations for gathering large-scale quantitative insights on how real TikTok users be-

must be conducted with external data. Prior work has gathered data by scraping the platform (e.g., [2, 19, 20]), an approach that can only collect a few thousands of videos, relies on publicly available information that are included on the web page's source, and is usually biased towards popular videos; from self-reports (e.g., [18, 22, 26]), which suffer from known biases in social media research [27, 42, 9, 12]; or from researcher-created accounts [6, 35], which is a promising technique, but may yield data that ultimately lacks the authenticity, diversity, and account history that real user accounts would contain.

Measuring what people do online with smart surveys

- **Two main approaches** have been used to enhance survey data with digital trace data about people's online behaviours: **web trackers and data donations**

- Very **similar and extremely different approaches**. On many levels:
    - Type of data collectable
    - Ethical concerns
    - Method of collecting data
    - Specific errors for each of them

Measuring what people do online with smart surveys

- **Two main approaches** have been used to enhance survey data with digital trace data about people's online behaviours: **web trackers and data donations**

- Very **similar and extremely different approaches**. On many levels:
  - Type of data collectable
  - Ethical concerns
  - Method of collecting data
  - Specific errors for each of them

- I will teach based on my practical experience and methodological work. That's why **I will focus more on web tracking data.**

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

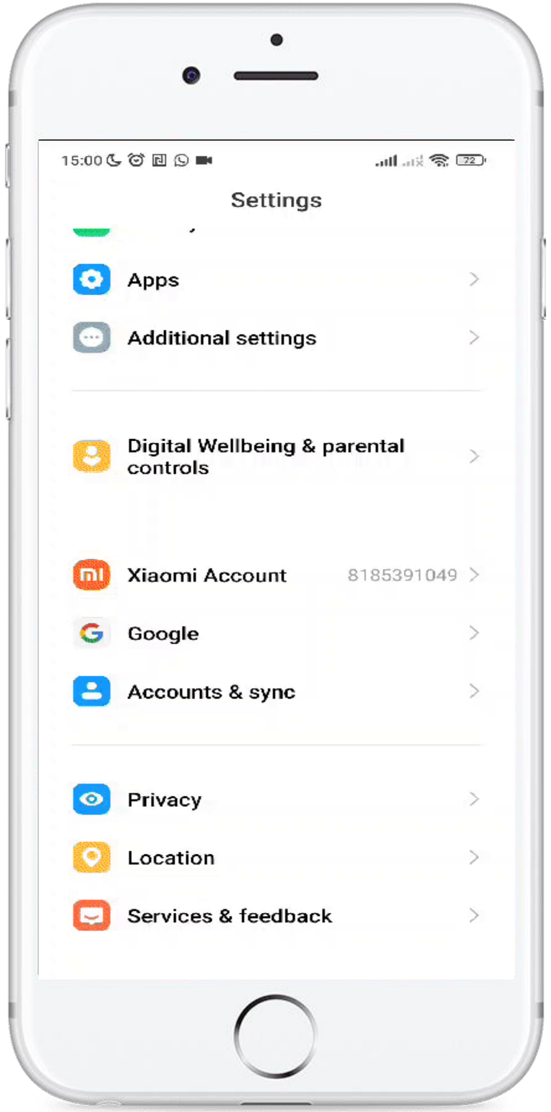
Users directly provide researchers with data that already has been collected by their devices or platforms

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**



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

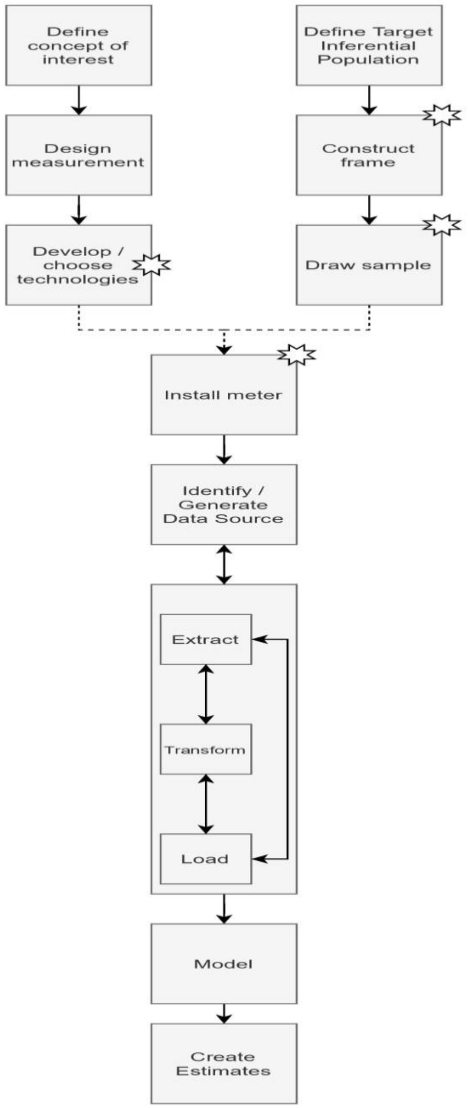
SECTIONS PDF TOOLS SHARE

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)

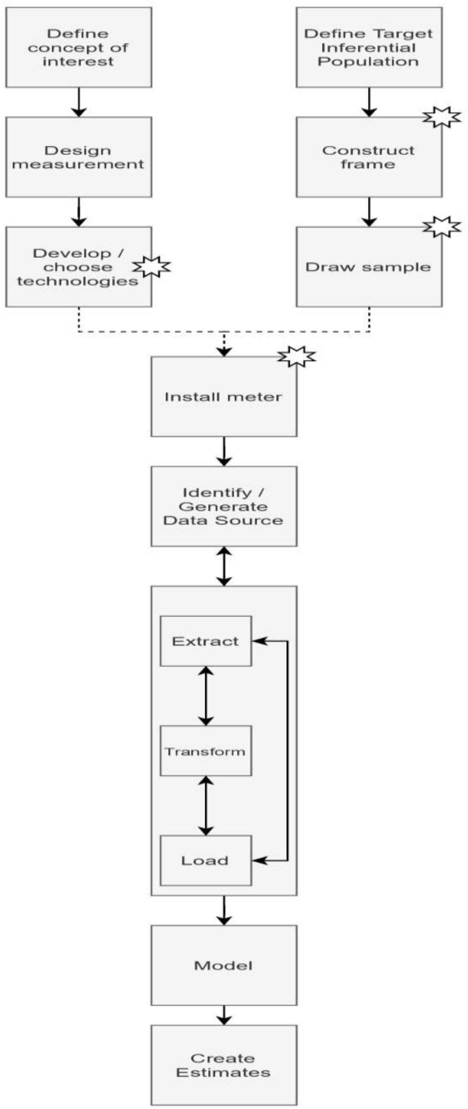
- In general, web tracking data is used to **make inferences** about a **concept of interest** for a given **population**



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

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

- Two parallel processes: **measurement** and **representation**

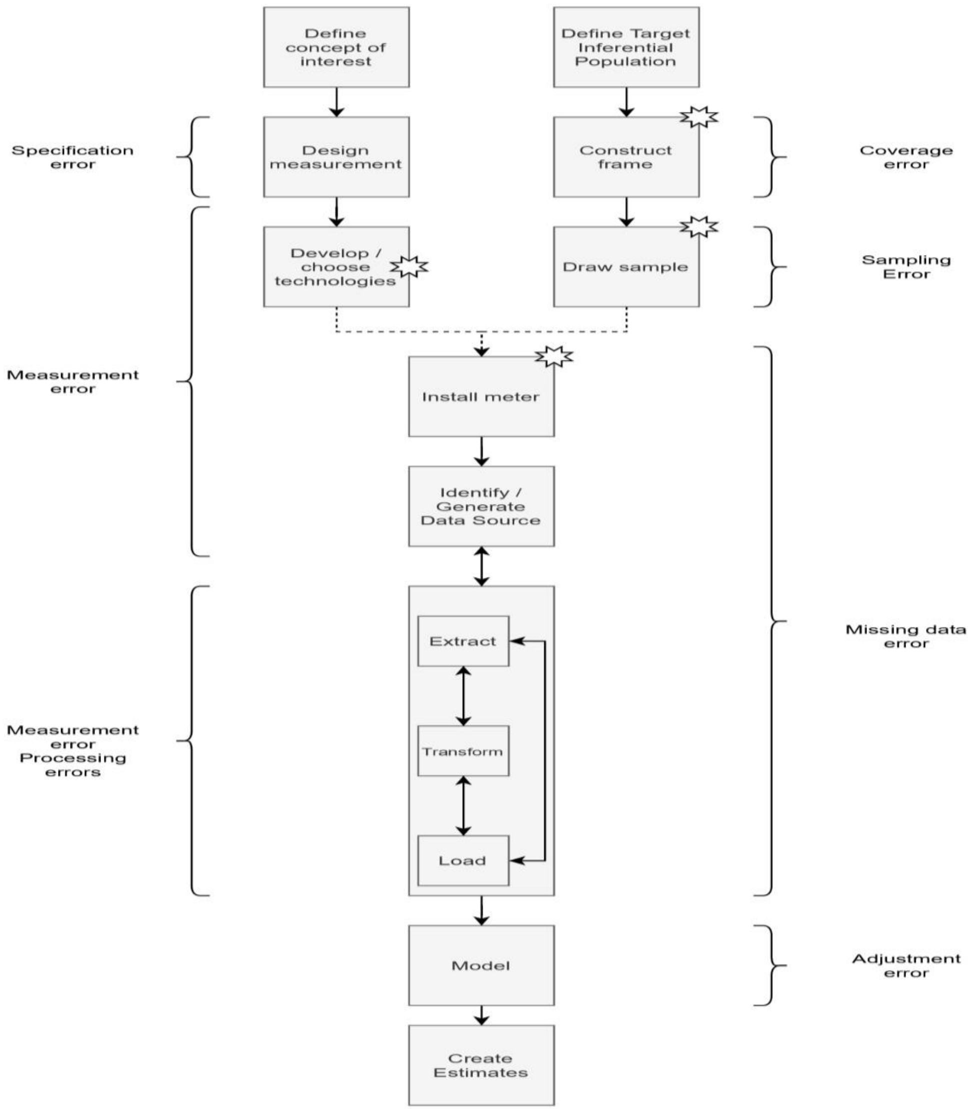


Total Error framework for digital traces collected w/ Meters (TEM)

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

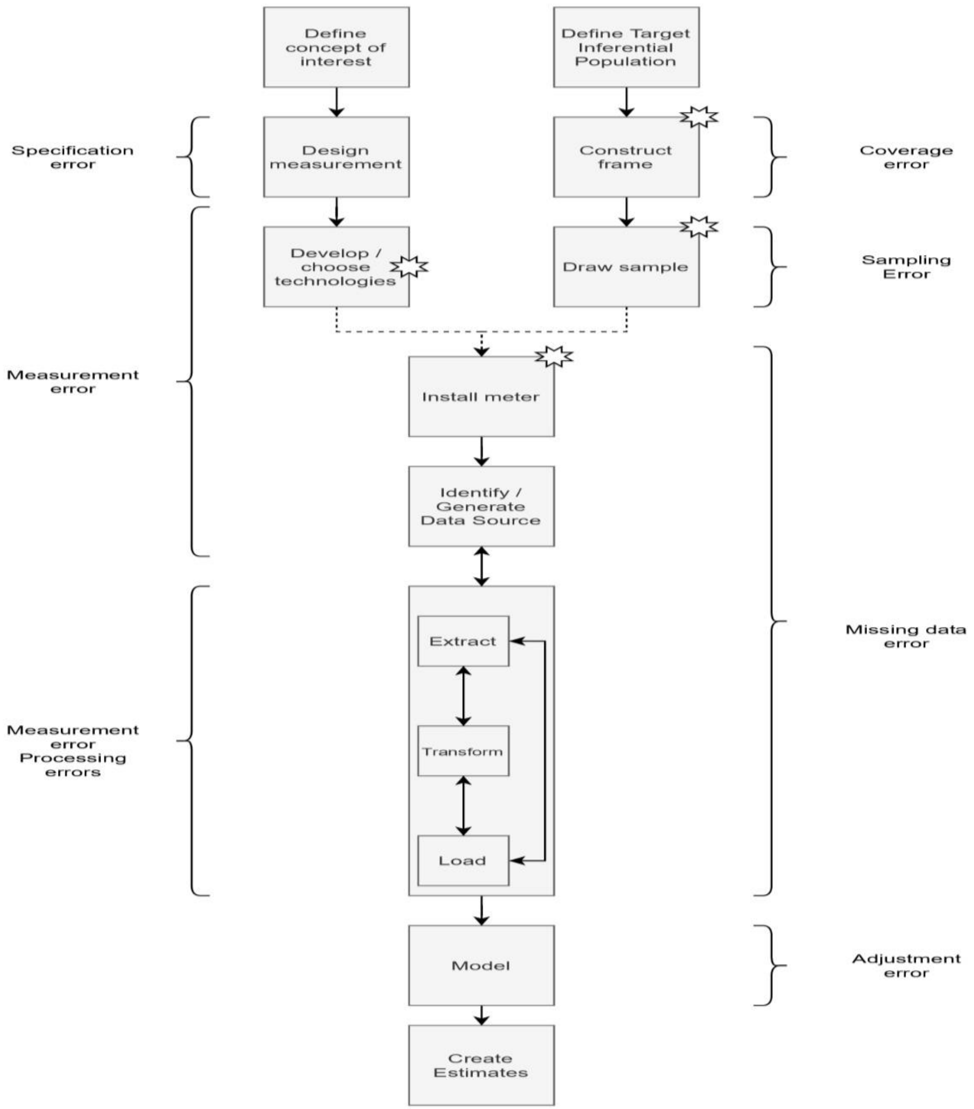
- Two parallel processes: **measurement** and **representation**

- Errors can happen in both sides

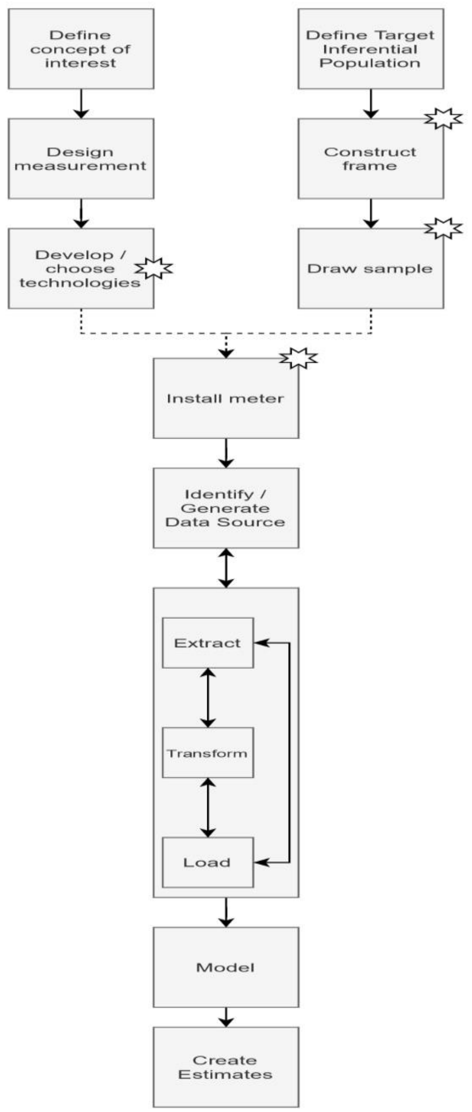


Total Error framework for digital traces collected w/ Meters (TEM)

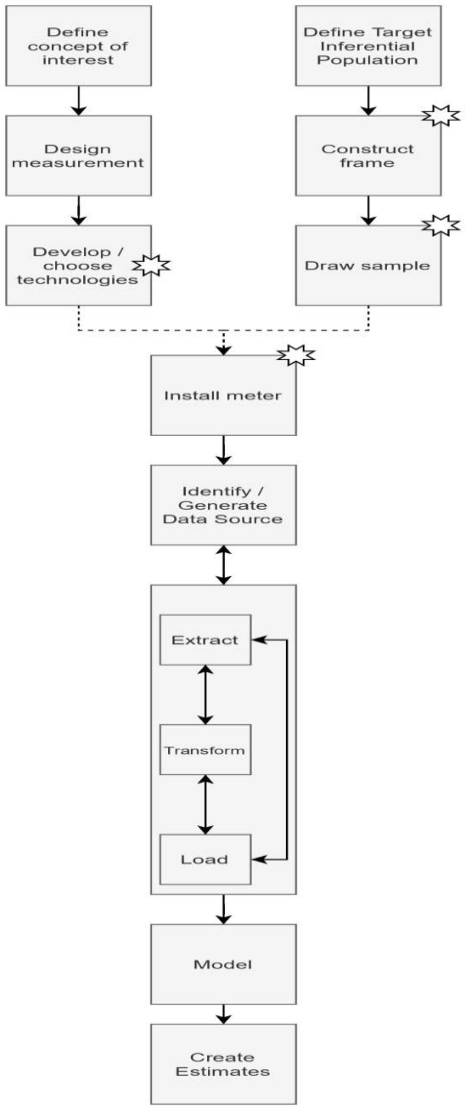
- In general, web tracking data is used to **make inferences** about a **concept of interest** for a given **population**
- Two parallel processes: **measurement** and **representation**
- Errors can happen in both sides
- The goal is to, within the project's **time** and **budget** constraints, **reduce as much as possible** the errors



A step-by-step guide



A step-by-step guide

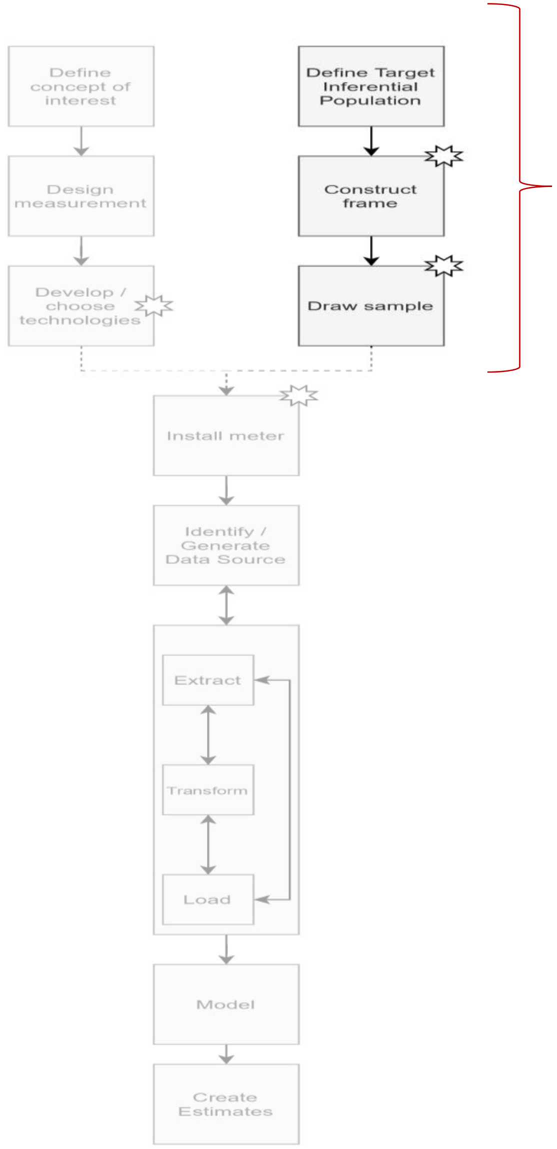


There are many steps to follow when collecting web tracking data.

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

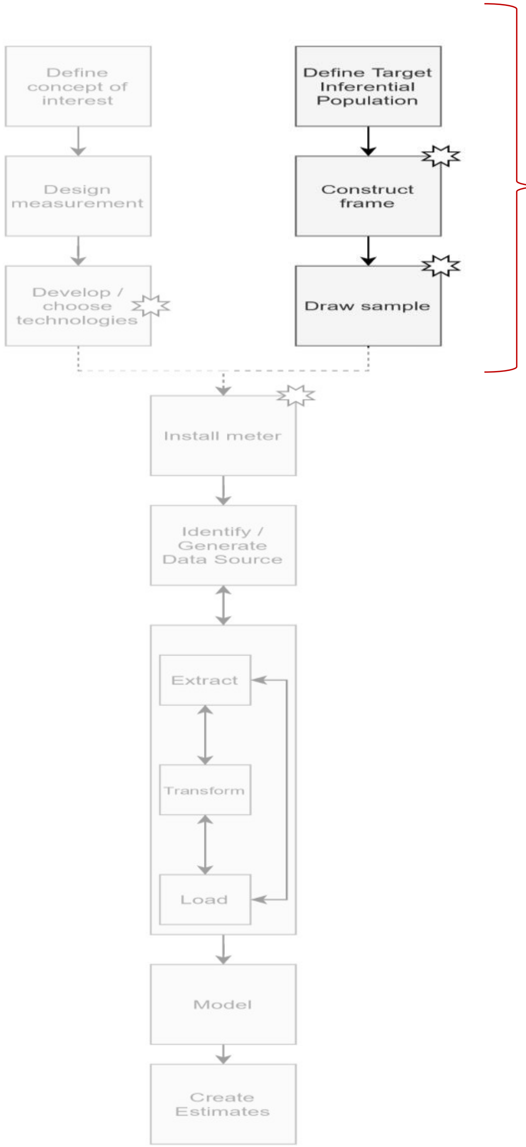
This is rarely acknowledged and understood, we can do better!

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



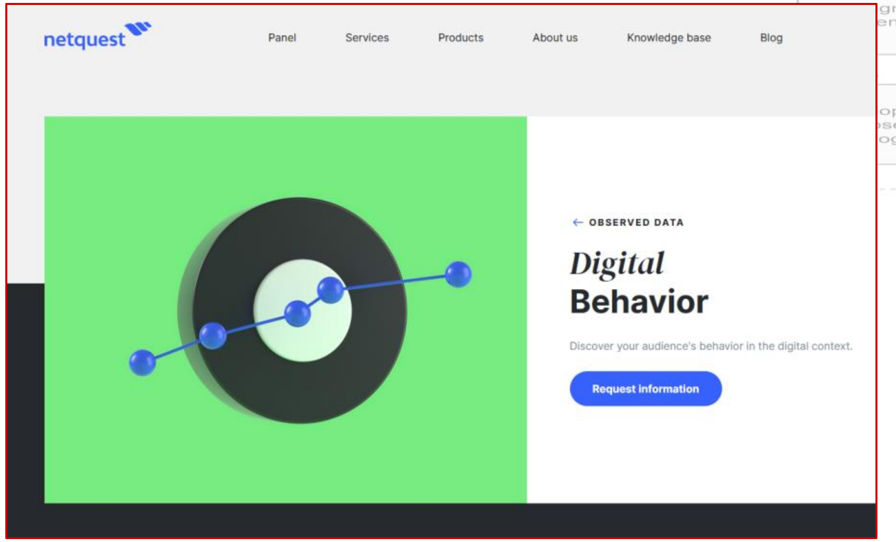
Identical steps as for surveys

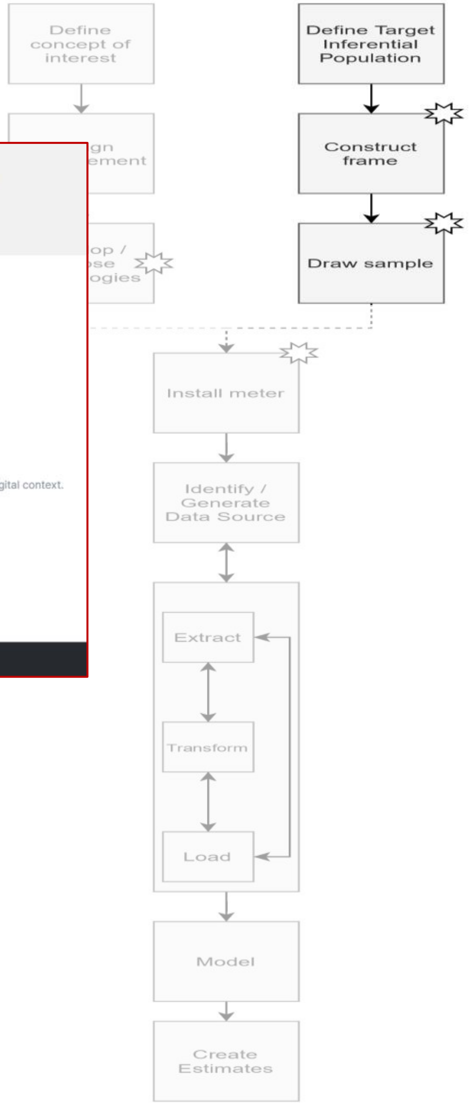
Target population: People living in the UK older than 17

Frame: Postal Address Frame

Sample: Simple Random Sampling

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





Identical steps as for surveys

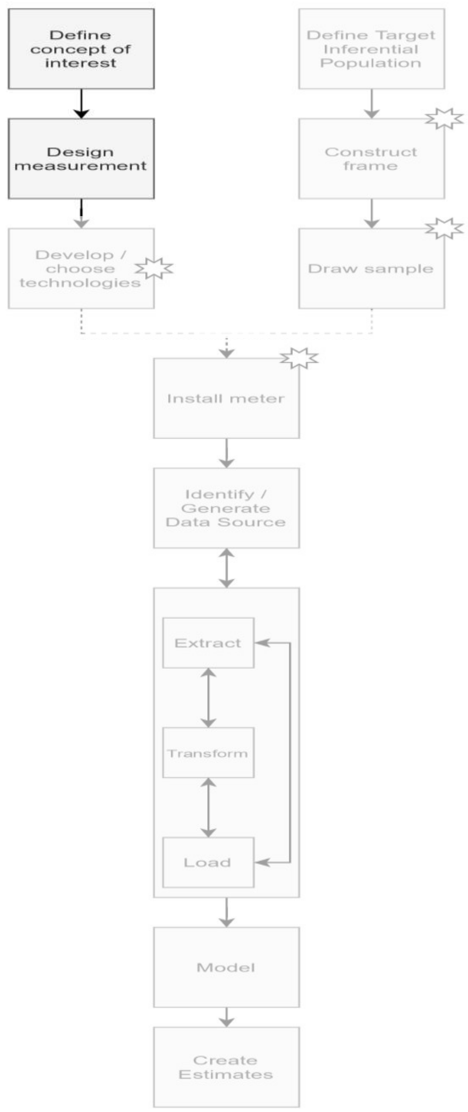
Target population: People living in the UK older than 17

Frame: Postal Address Frame

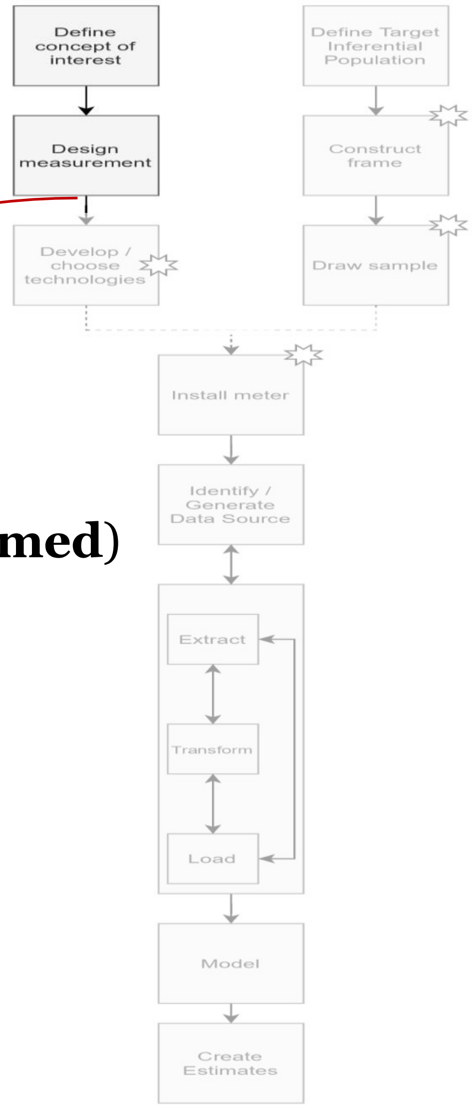
Sample: Simple Random Sampling

Most commonly: non-probability online panels

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*

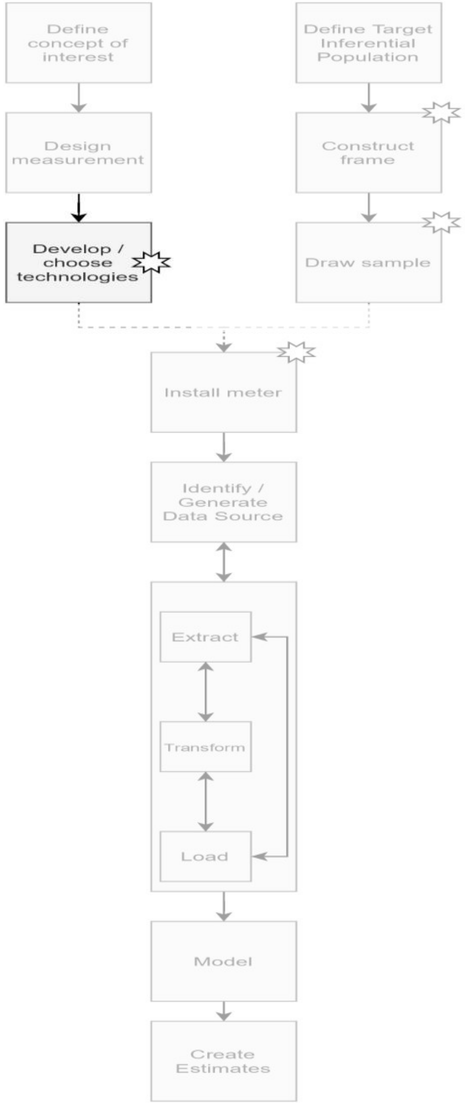
Measure: *average time recorded of the visits to URLs defined as showing written news*

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

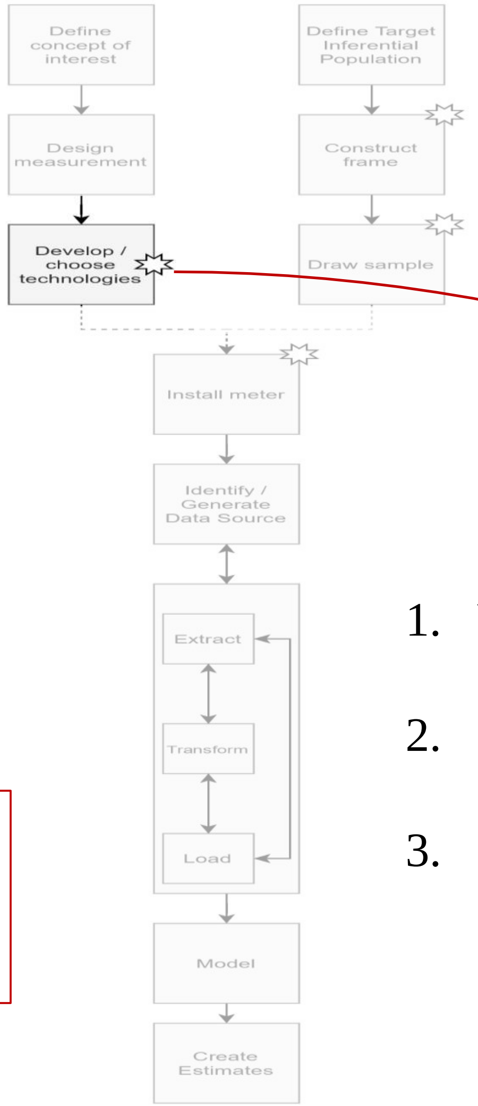
These and other decisions will **determine the measurement used**.

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



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

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]

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.**

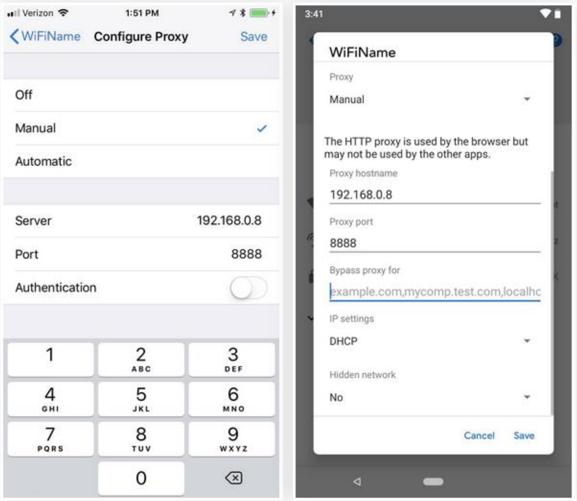
- **These can be**: Proxies, VPNs, Screen-scrapers, Screen recorders, Smartphone-log trackers (and maybe more that I am not aware of).

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

- **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

- **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

Table 1. Overview of existing tools for tracking of online media use on desktop devices.

Available tools	Approach	Types of information	Technical complexity	Privacy features	User experience	Availability
Roxy (Menchen-Trevino & Karr, 2012)	Proxy	actual content, but not from encrypted websites (HTTPS)	high	user-specific and system-wide blacklist; log-in option	relatively complex installation; relatively high level of intrusiveness	code made available open-source
Newstracker (Kleppe & Otte, 2017)	Proxy	content, but no personalization	high	whitelist; log-in option	relatively complex installation; relatively high level of intrusiveness	not open-source
Robin (Bodo et al., 2017)	Proxy	actual content & usage	high	whitelist	relatively easy installation; relatively high level of intrusiveness	not open-source
Eule (Haim & Nienierza, 2019)	Screen-Scraping	actual content & usage of publicly available Facebook posts	medium/high	whitelist; log-in option	relatively easy installation; relatively low level of intrusiveness	code made available open-source
WebTrack (Adam et al., 2019)	Screen-Scraping	actual content & usage	medium/high	blacklist & private-mode option; log-in option	relatively easy installation; relatively low level of intrusiveness	under development

Table 2. Overview of different approaches for tracking online media use for mobile devices.

Approach	Types of information	Technical complexity	User experience	Availability	Available tools
Smartphone log	visited URLs only, no content; can get other behavior data, e.g., calls log.	high	medium (can be highly intrusive depending on the implementation of a specific tool)	yes, but no browsing tracking functionality support (e.g., MobileDNA, not open source)	MobileDNA (Van Damme et al., 2020), the tool is not open source and does not track which URLs were visited
Proxy	URLs + some content (including limited in-app browsing)	high	low (difficult to set up, potentially intrusive)	yes, not academic (e.g., Charles Proxy),	None
Standalone browser/ news app	Content, but only that accessed through this app/browser	medium	medium (highly intrusive)	no (outdated)	None
Browser extension	content, but only that accessed through the browser where the extension is installed	medium	medium (highly intrusive)	no (prototype only)	None
Screen-capturing	All the content including in-app browsing	medium; high for data processing	medium (can be highly intrusive depending on the implementation of a specific tool)	yes, for Android (including open source)	Screenomics (Reeves et al., 2021); unnamed screen recorder (Krieter, 2019b)

Christner, C., Urman, A., Adam, S., & Maier, M. (2022). Automated tracking approaches for studying online media use: A critical review and recommendations. *Communication methods and measures*, *16*(2), 79-95.

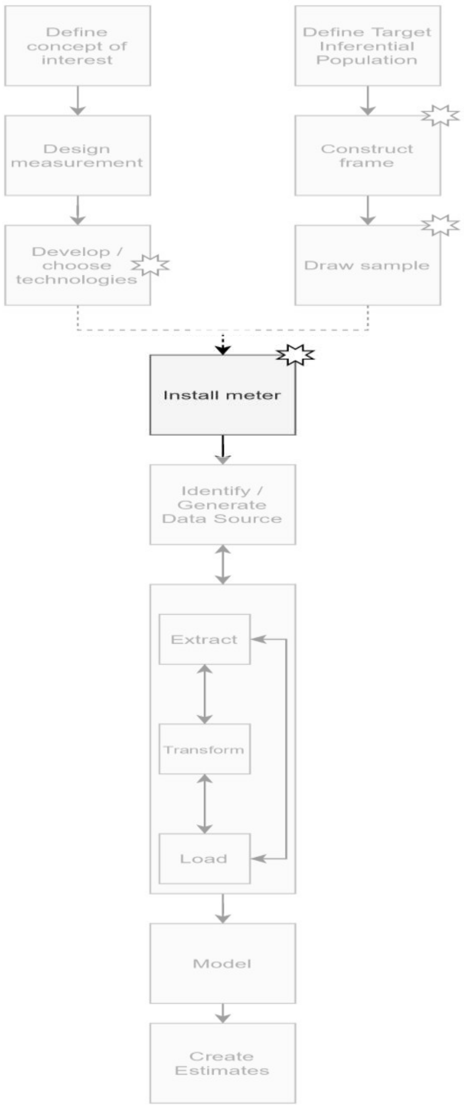
A heterogeneous group of tracking solutions

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

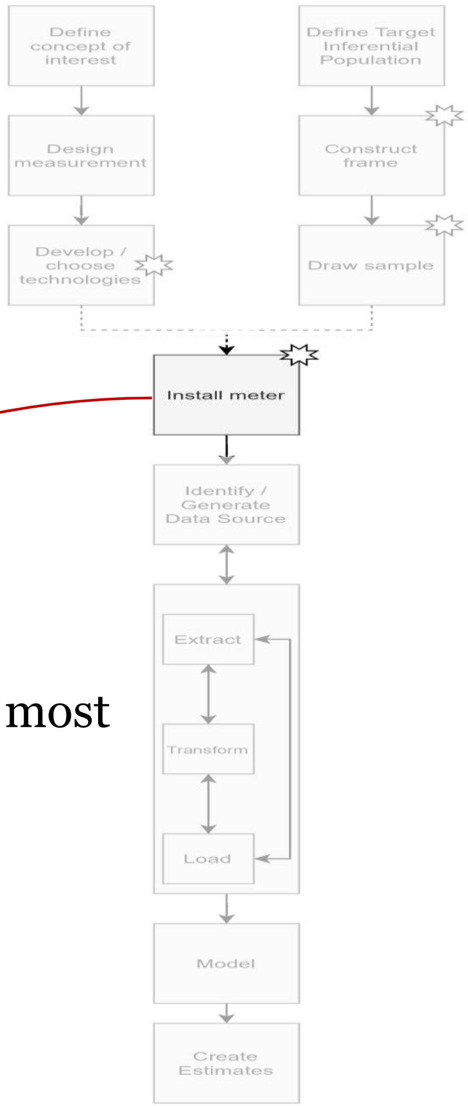
		PC app	PC plug-ins			Android SDK	iOS proxy
			Chrome	Firefox	Safari		
Online tracking							
URLs	Http traffic	Yes	Yes	Yes	Yes	Yes	Yes
	Https traffic	No	Yes	Yes	Yes	Yes	No
	Incognito sessions	No	Yes	Yes	Yes	Yes	No
	HTML	No	Yes	Yes	Yes	No	No
	Time stamps	Yes	Yes	Yes	Yes	Yes	Yes
Apps	App name	-	-	-	-	Yes	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?



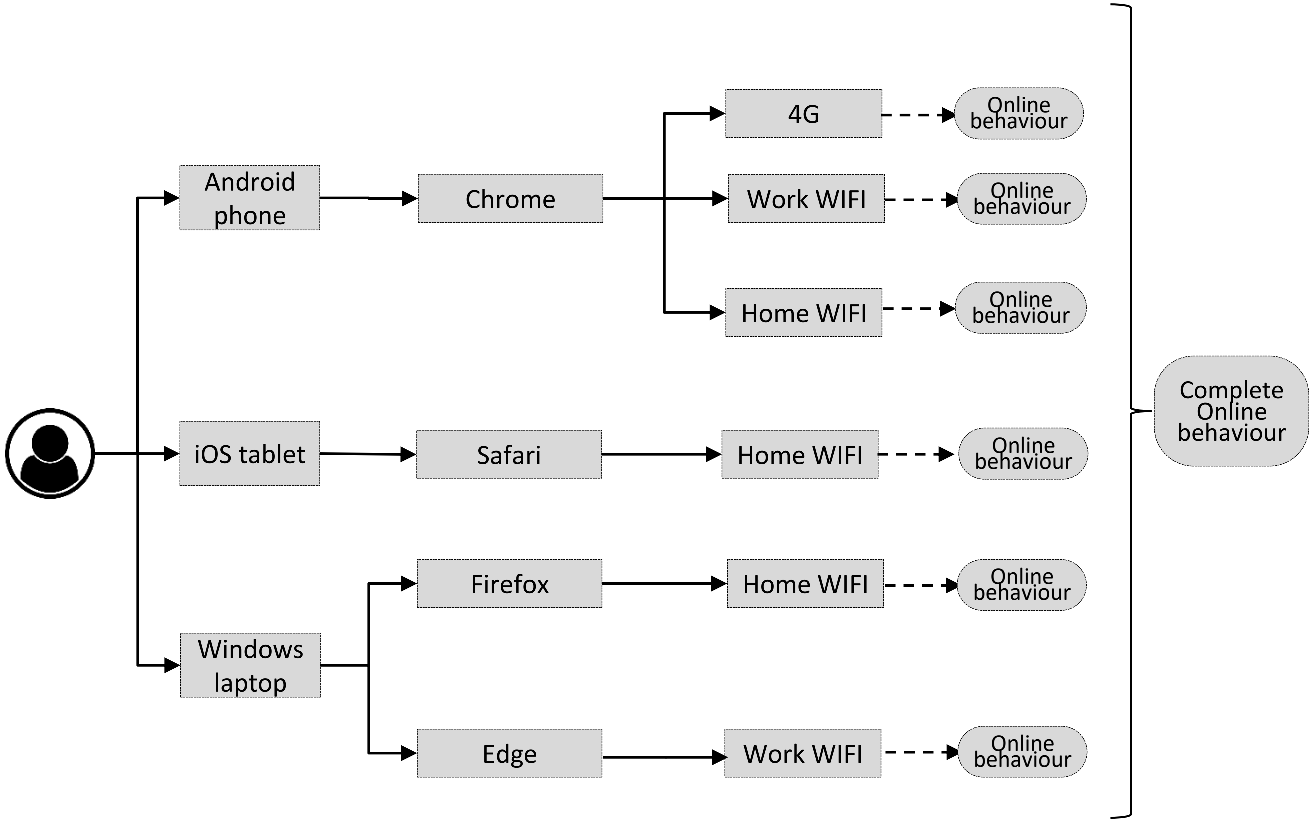
Could you please, maybe, install this meter?



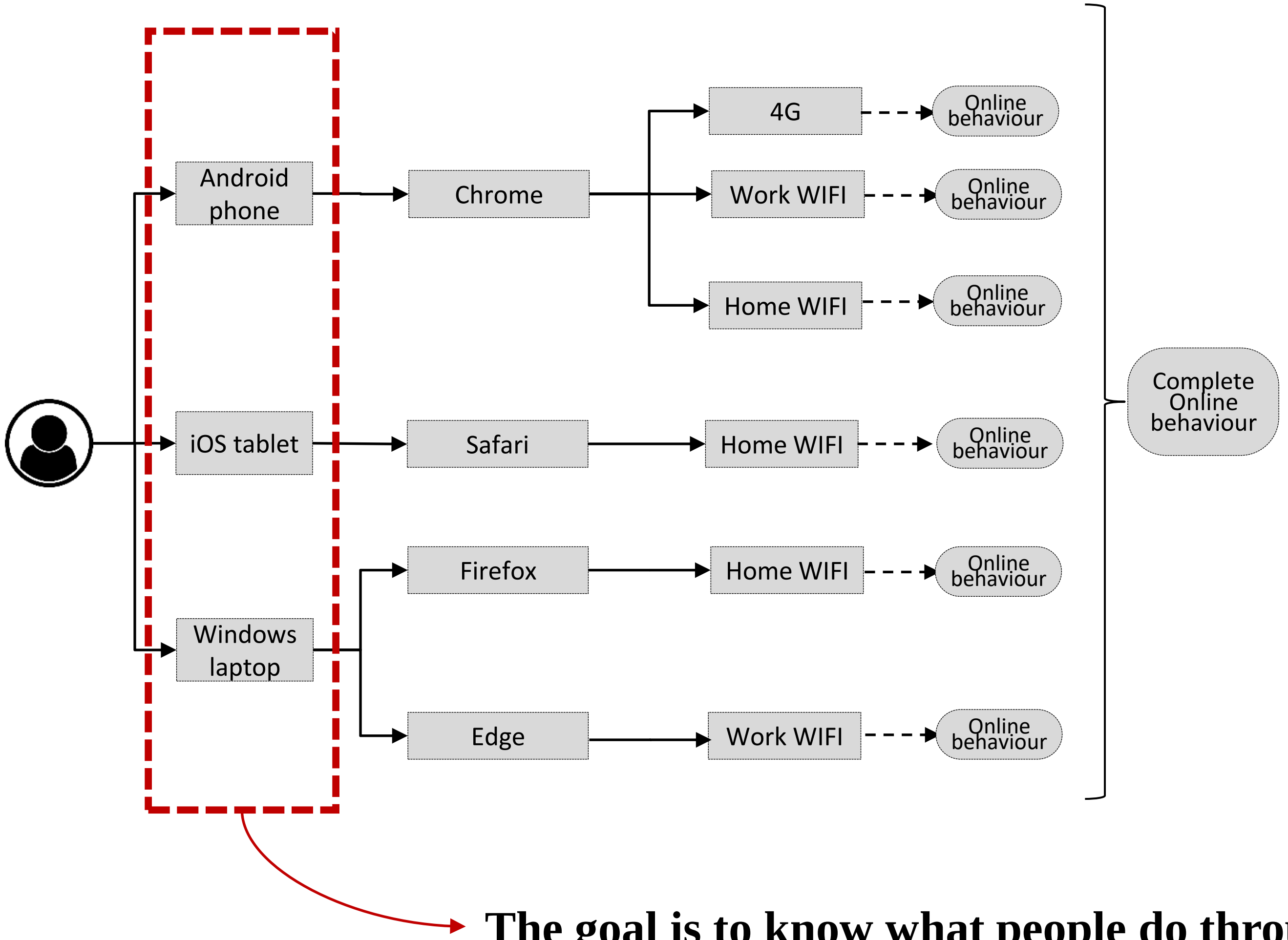
This process is, potentially, one of the most consequential ones for web tracking research. It determines:

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

Could you please, maybe, install this meter?

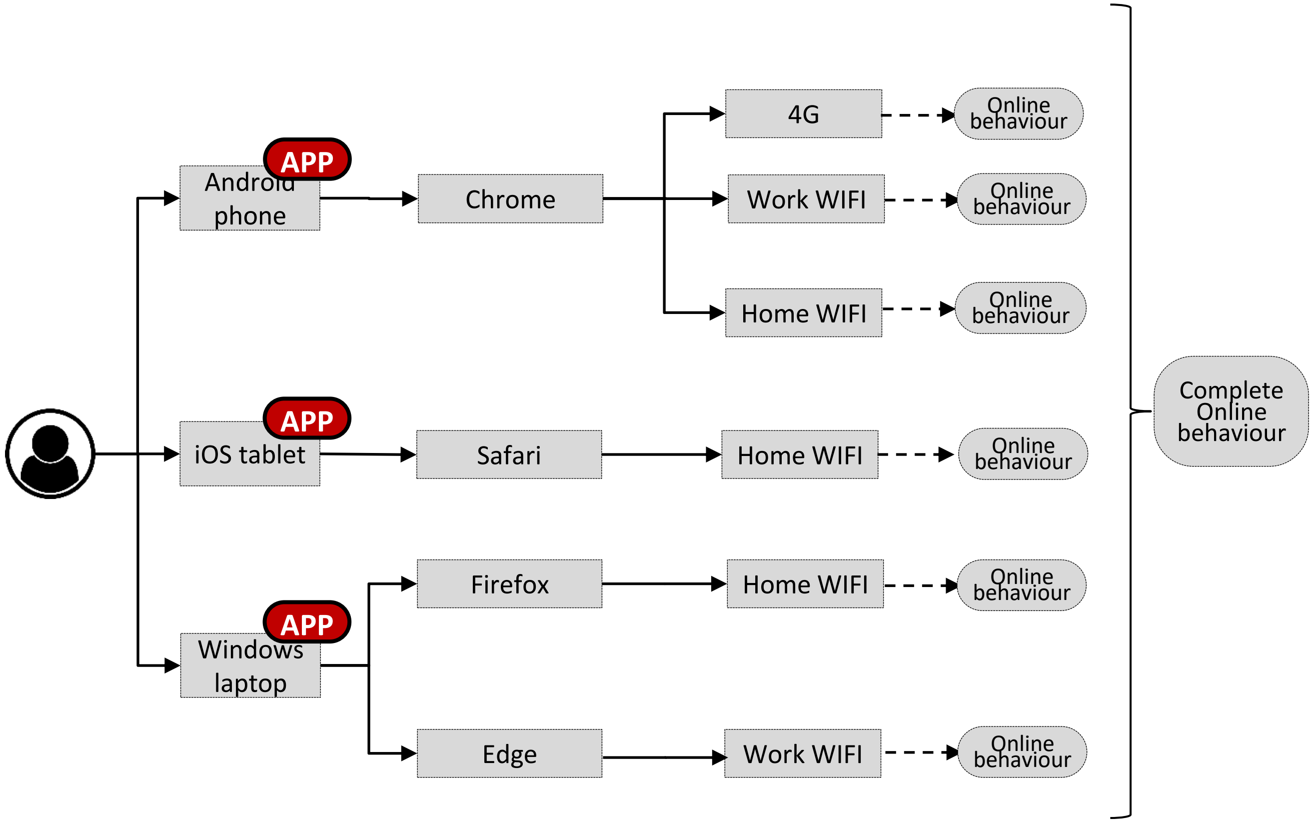


Could you please, maybe, install this meter?



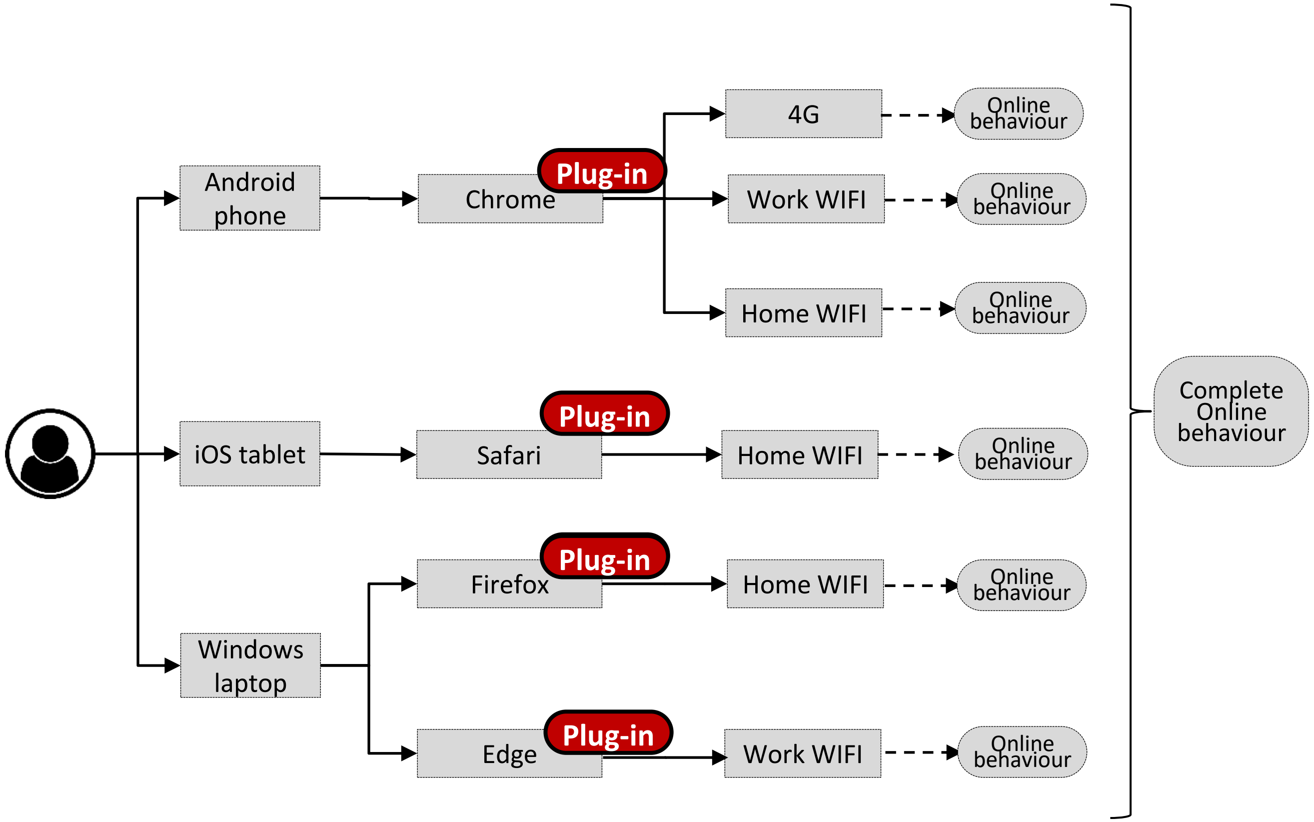
The goal is to know what people do through all their devices

Could you please, maybe, install this meter?



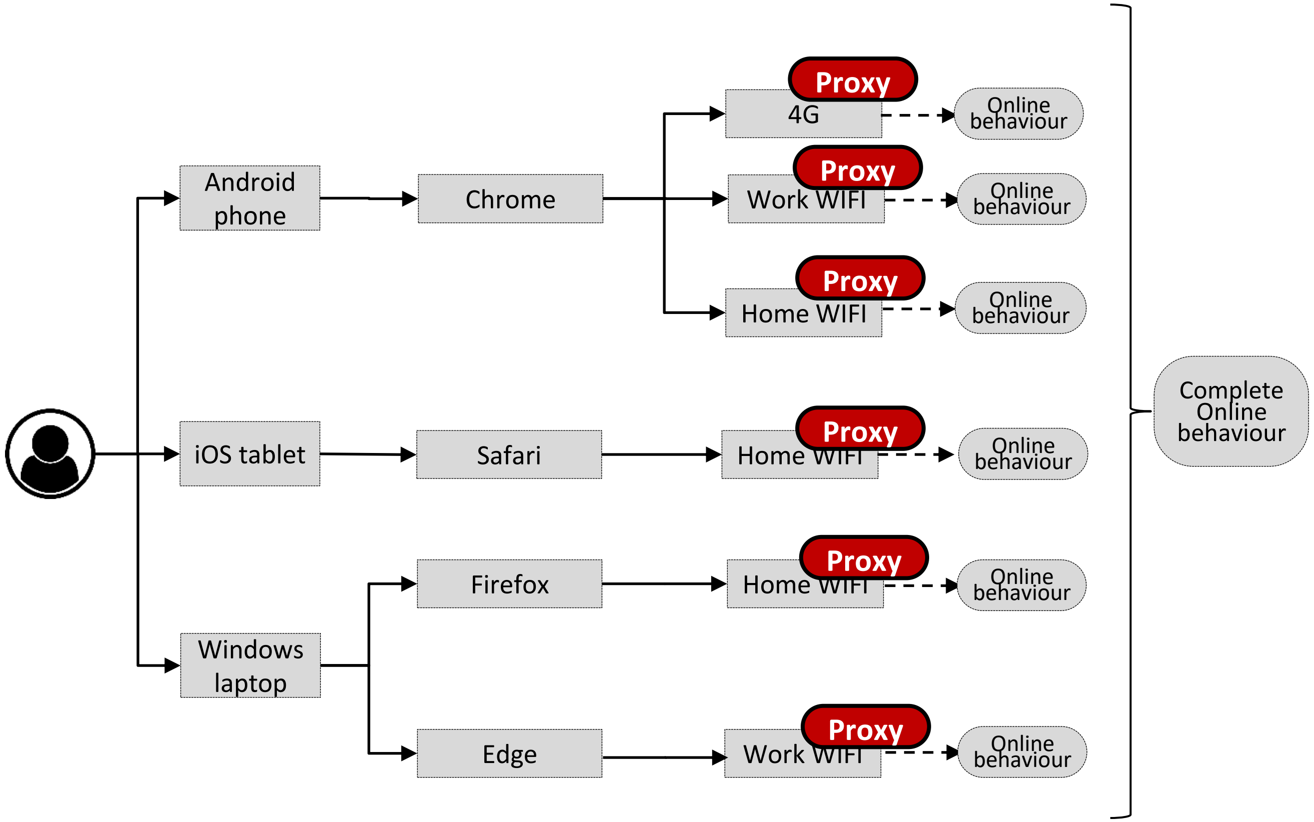
This can be achieved by tracking all devices that someone uses

Could you please, maybe, install this meter?



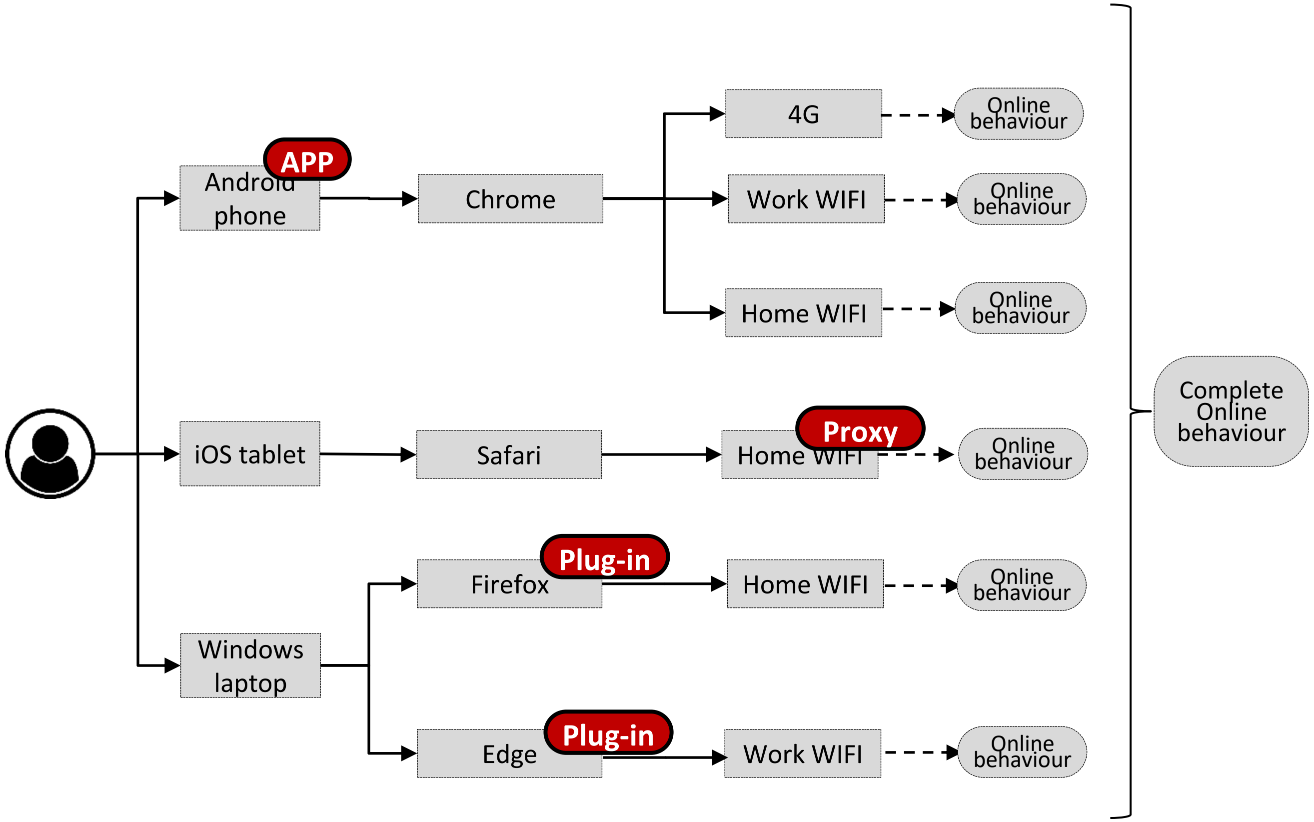
Or all their browsers

Could you please, maybe, install this meter?



Or all their networks

Could you please, maybe, install this meter?



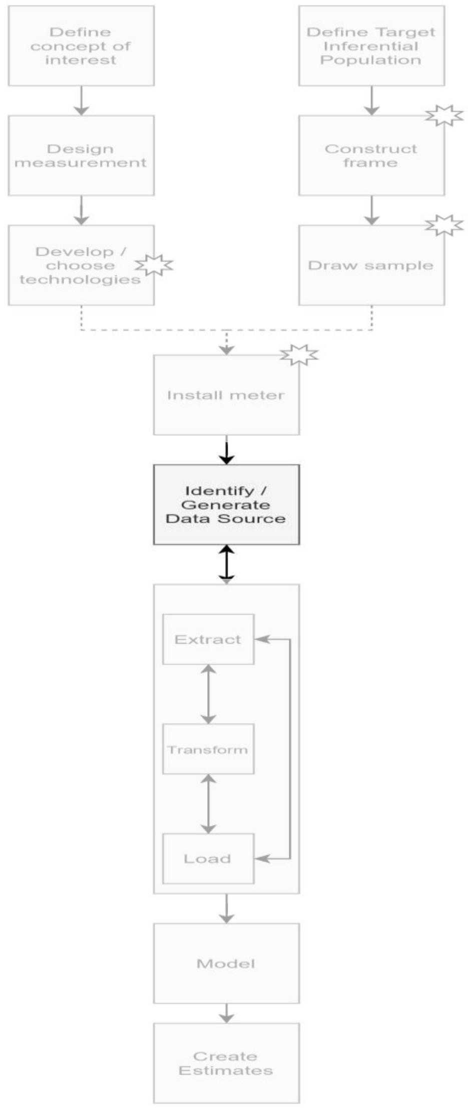
Or a combination of these (most common)

The challenging reality of trying to track people

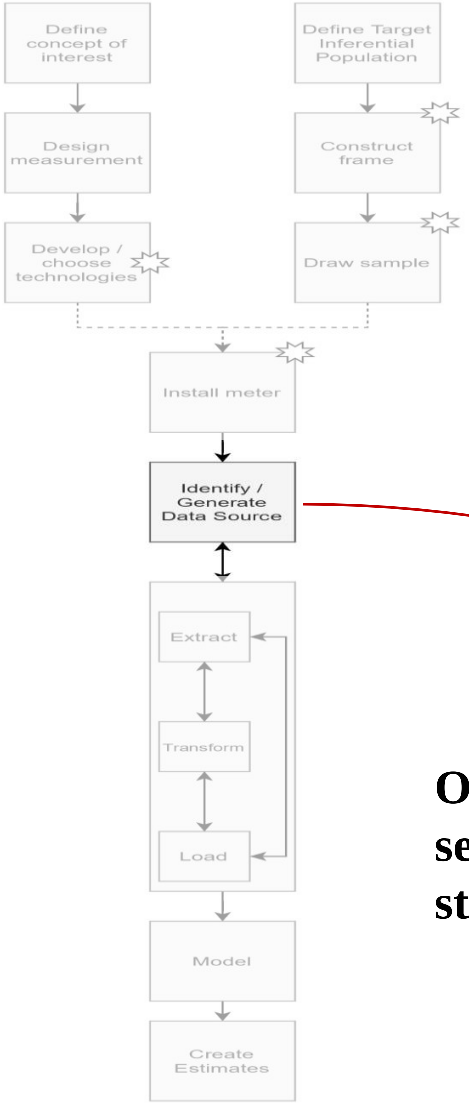
webdataopp

- There is not a one-size-fits-all approach that can track everything people do online.
- For most people, we might have to ask them to install / configure meters in more than one device.
- These trackers can be different for every devices they use (as we have seen before).
- The information about what Devices / Operating Systems / Browsers they use is not available beforehand, needs to be collected from them.
- The devices and browsers that people use, and the versions of their OS, can change over time.
- If we use an already available panel, this is mostly out of our control!

Generate the messy dataset

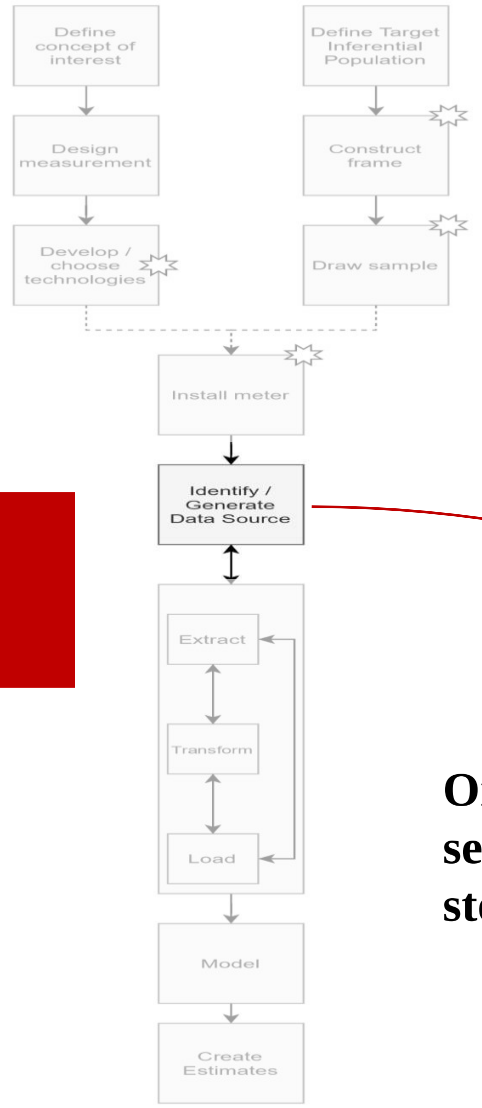


Generate the messy dataset



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

Generate the messy dataset



Sometimes, not all information is tracked! Whitelists and blacklists can be configured (ethically recommended?)

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.

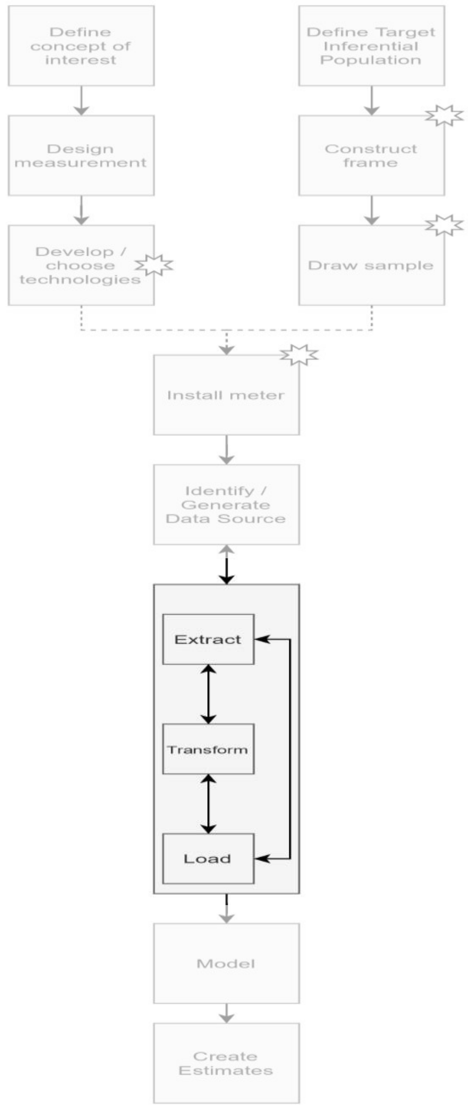
Generate the messy dataset

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

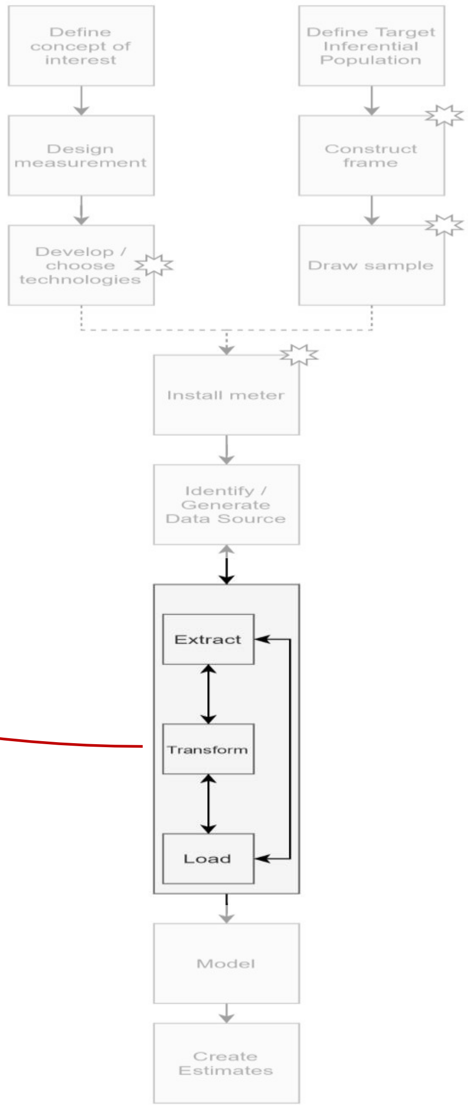
- This is one of the **most basic versions** of what information might be recorder (ID, time stamp, and full URL)
- Other information can be captured, such as **HTML information**. For instance, the **text** each Facebook post seen by a participant, the **number of likes**, the **comments**, why the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

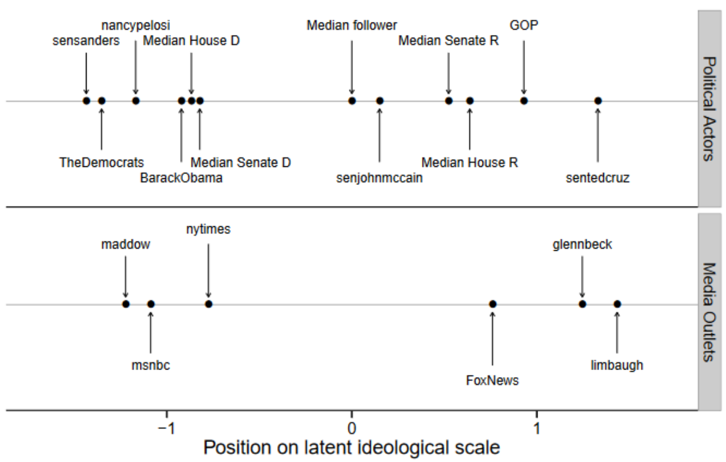
https://www.theguardian.com/environment/2023/jul/11/nuclear-bomb-fallout-site-chosen-to-define-start-of-anthropoce

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.

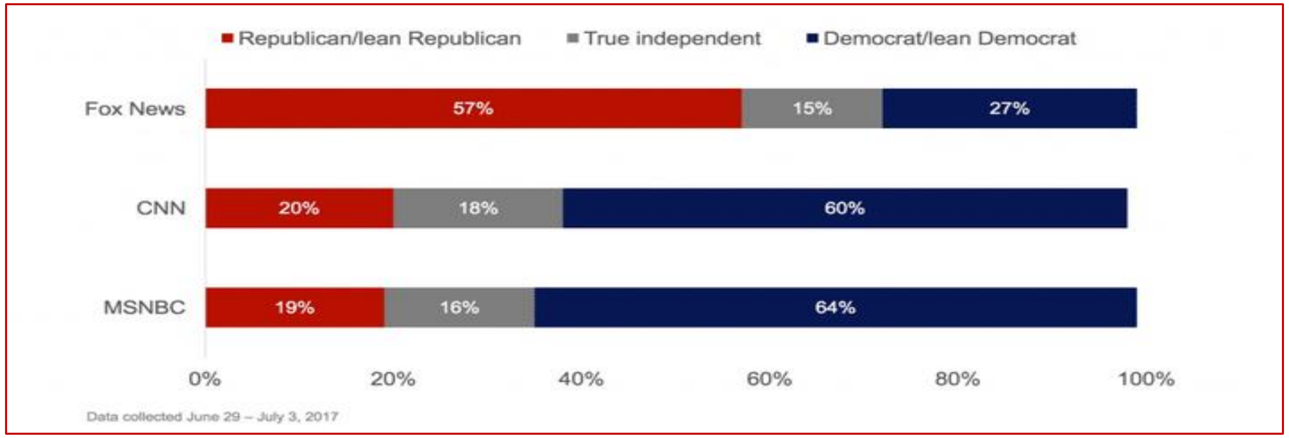
- Most interesting transformation: enriching the information that URLs bring to research.
  1. The content of the URL can be manually identified, and added to the dataset
  2. The webpages can be classified using external information
  3. Machine learning to codify the content exposed to (text / images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

Let's create the dataset to work with

web data opp

- 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).

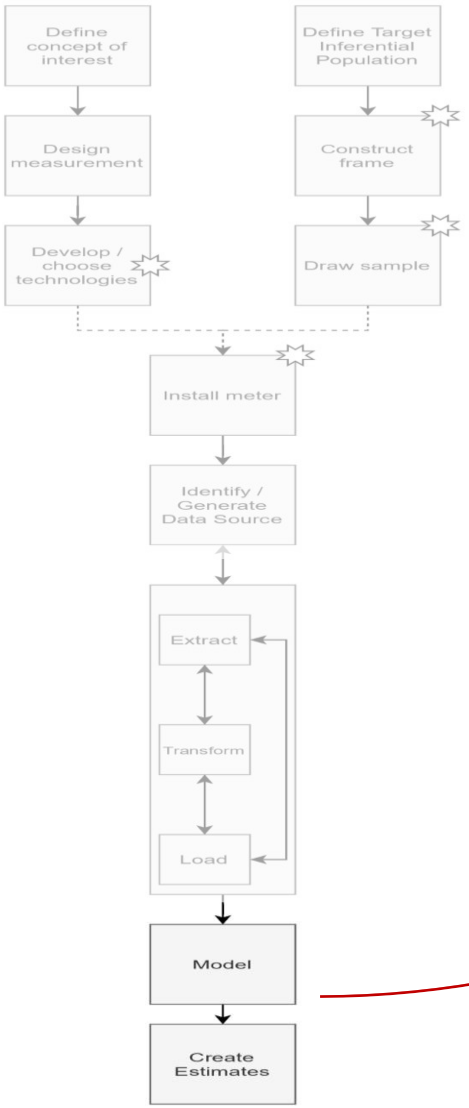
Let’s create the dataset to work with

- 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).

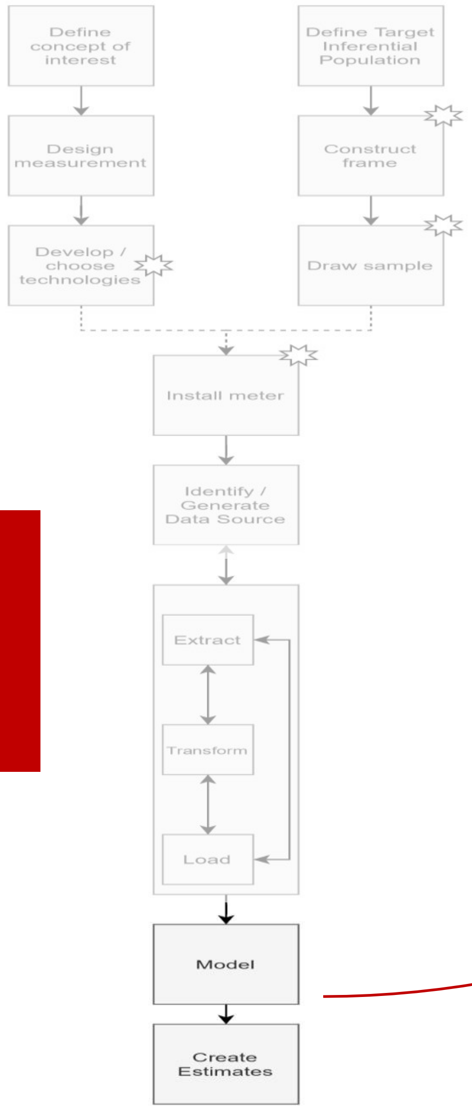
- This is a big difference compared with surveys, that:
  1. Makes the **pre-processing** stage of the research **harder and longer**
  2. But allows for **immense flexibility**, which can be exploited for good (we will talk about this later)

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).

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

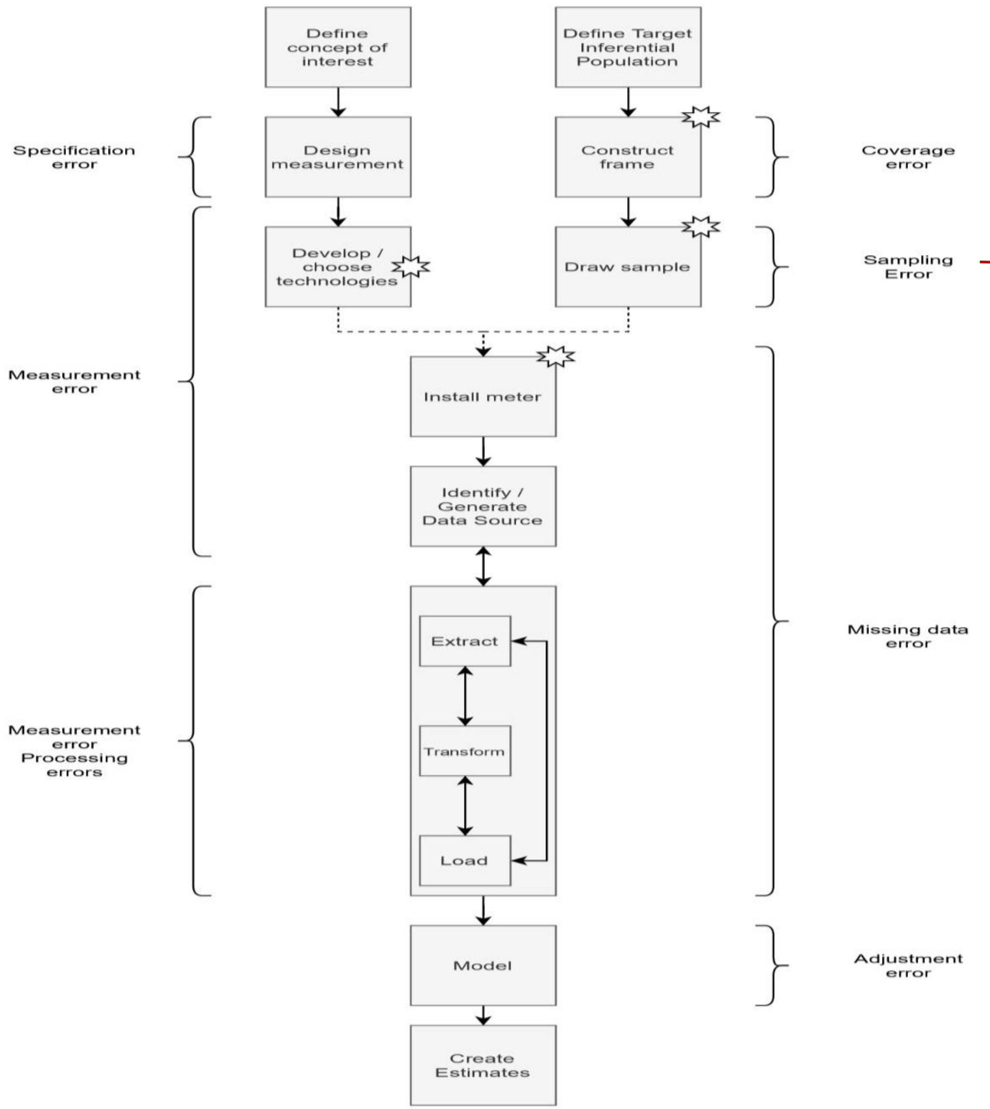


New missingness challenges might require innovative modelling strategies that are not common in surveys. We can discuss this later!

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).

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?

webdataopp

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

Most specific error causes on the side of measurement

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

Sampling and adjustment errors have no specific error causes

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.

Specification errors

Validity

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

Validity

- Does the measurement **reflect the underlying concept** that we intend to measure?

Concept of interest → **Measurement**

Validity

The extent to which an individual encounters written news media

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

Validity

- Does the measurement **reflect the underlying concept** that we intend to measure?

Concept of interest → **Measurement**

Validity

The extent to which an individual encounters written news media

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

<u>Specification error</u>
Traces are not accurately and comprehensively reflecting the concept

Validity

- Does the measurement **reflect the underlying concept** that we intend to measure?

Concept of interest → **Measurement**

Validity

The extent to which an individual encounters written news media

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

What can cause this?

Specification error

Traces are not accurately and comprehensively reflecting the concept

Defining what qualifies as valid information

Concept: The extent to which an individual encounters **written news media**

Defining what qualifies as valid information

Characteristics	Potential choices
List of traces	

web data opp

Defining what qualifies as valid information

webdataopp

1. **Metric:** what can best express variation in the "extent"?

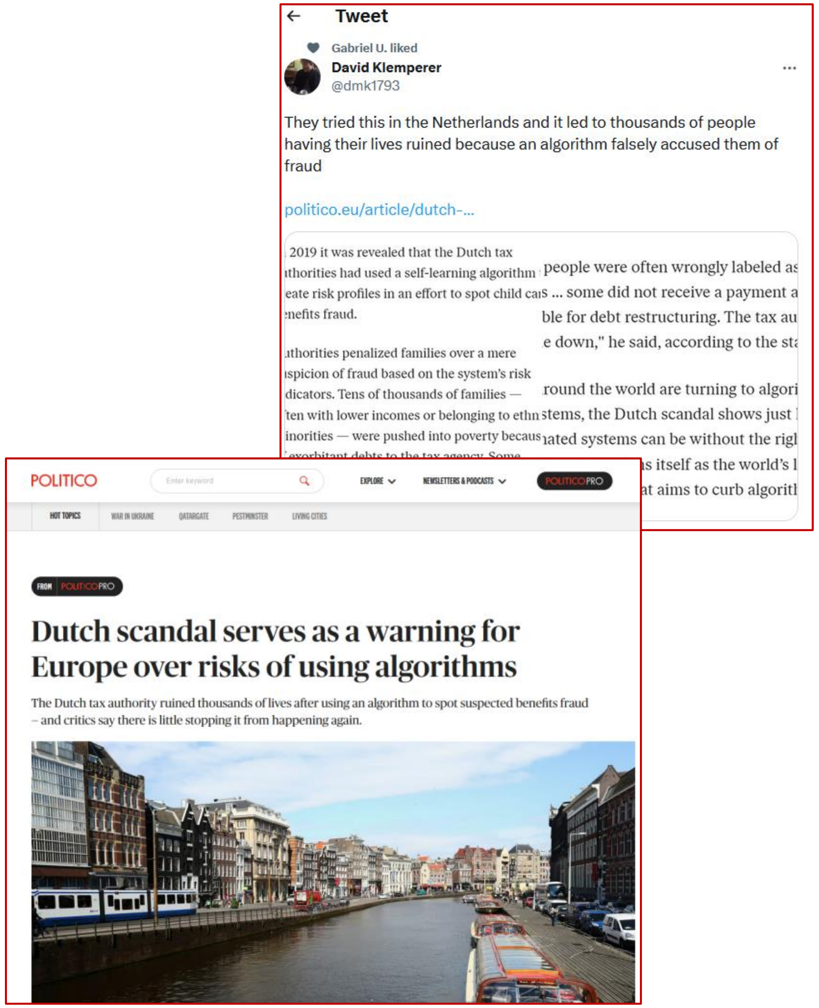
Characteristics	Potential choices
Metric	Visits, Seconds, Days, Media
List of traces	

In surveys, the validity is higher for days or media, is it the same for web tracking?

Defining what qualifies as valid information

2. **List of traces:** what is defined as "written news media"?

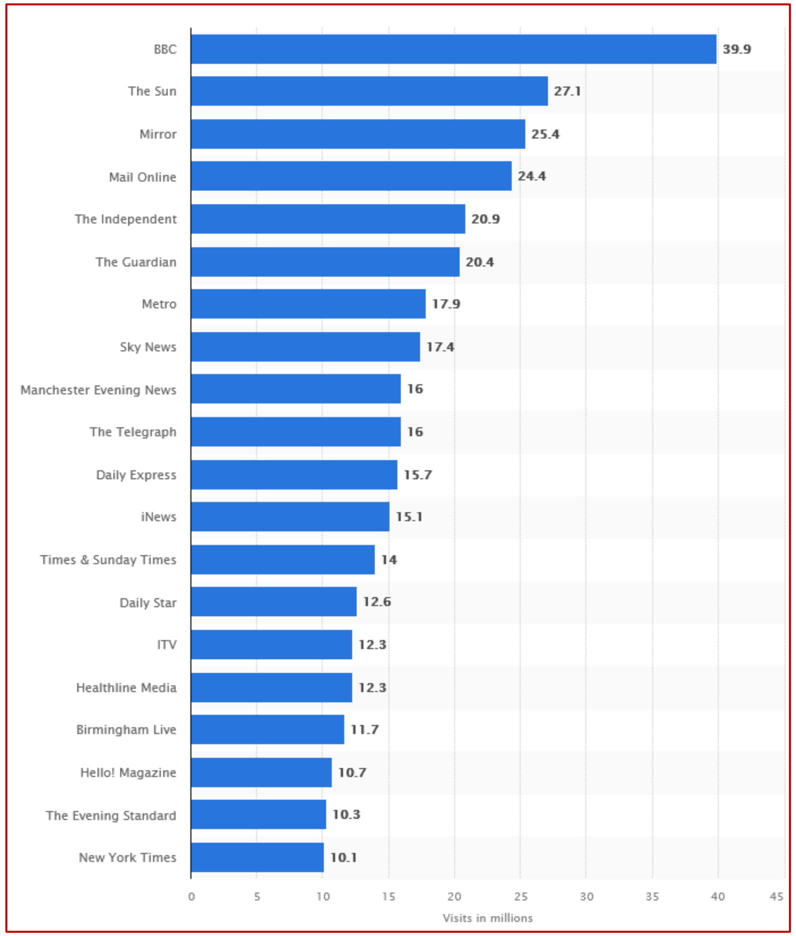
Characteristics	Potential choices
Metric	Visits, Seconds, Days, Media
List of traces	
What is news?	Published by news media, published by any person/media



Defining what qualifies as valid information

2. **List of traces:** what is defined as “written news media”?

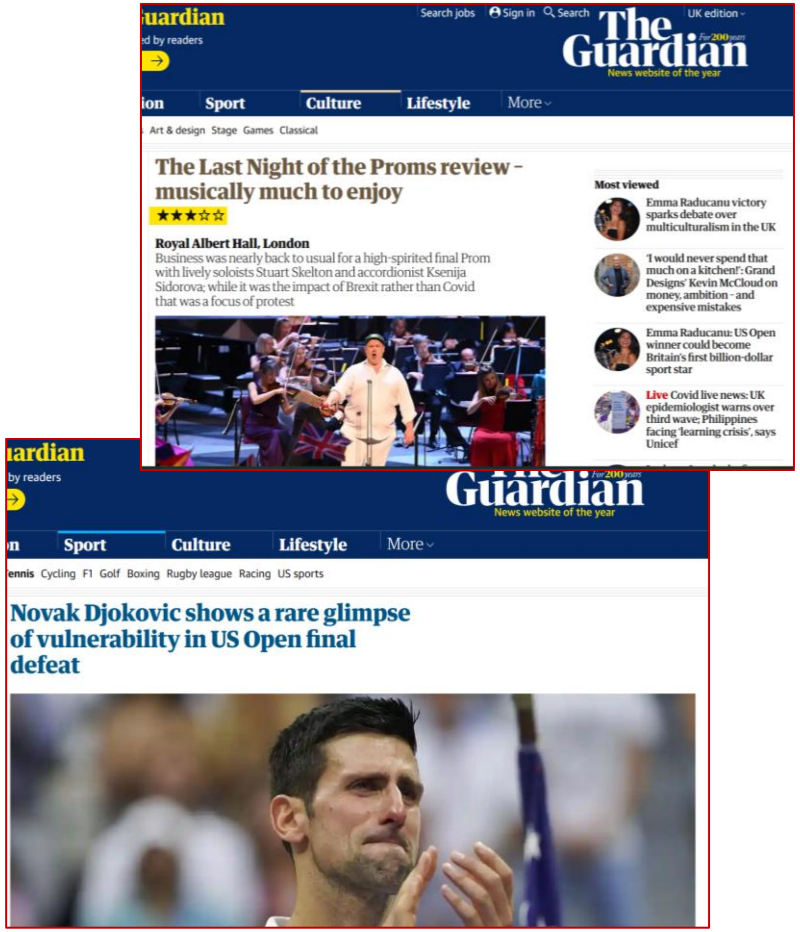
Characteristics	Potential choices
Metric	Visits, Seconds, Days, Media
List of traces	
What is news?	Published by news media, published by any person/media
List of media	Tranco, Alexa, Cisco, Majestic
Top media	10, 20, 50, 100, 200, All



Defining what qualifies as valid information

2. **List of traces:** what is defined as "written news media"?

Characteristics	Potential choices
Metric	Visits, Seconds, Days, Media
List of traces	
What is news?	Published by news media, published by any person/media
List of media	Tranco, Alexa, Cisco, Majestic
Top media	10, 20, 50, 100, 200, All
Information	Broad definition of news, only those identified as "political" news



Defining what qualifies as valid information

3. **Exposure:** what events can be considered as “exposed”?

Characteristics	Potential choices
Metric	Visits, Seconds, Days, Media
List of traces	
What is news?	Published by news media, published by any person/media
List of media	Tranco, Alexa, Cisco, Majestic
Top media	10, 20, 50, 100, 200, All
Information	Broad definition of news, only those identified as “political” news
Exposure	
Time threshold	1 second, 30 seconds, 120 seconds

Exposure might mean just seeing something, or reading part / all of the article

Defining what qualifies as valid information

3. **Exposure:** what events can be considered as "exposed"?

Characteristics	Potential choices
Metric	Visits, Seconds, Days, Media
List of traces	
What is news?	Published by news media, published by any person/media
List of media	Tranco, Alexa, Cisco, Majestic
Top media	10, 20, 50, 100, 200, All
Information	Broad definition of news, only those identified as "political" news
Exposure	
Time threshold	1 second, 30 seconds, 120 seconds
Devices	PC only, Mobile only, All, All without apps

Most research has focused only on behaviours through PCs, is this right?

Defining what qualifies as valid information

webdataopp

4. **Tracking period:** what time period allows to measure "normality"?

Characteristics	Potential choices
Metric	Visits, Seconds, Days, Media
List of traces	
What is news?	Published by news media, published by any person/media
List of media	Tranco, Alexa, Cisco, Majestic
Top media	10, 20, 50, 100, 200, All
Information	Broad definition of news, only those identified as "political" news
Exposure	
Time threshold	1 second, 30 seconds, 120 seconds
Devices	PC only, Mobile only, All, All without apps
Tracking period	2, 5, 10, 15, 31 days

Longer tracking periods might be better, but also more expensive

Defining what qualifies as valid information

Characteristics	Potential choices
Metric	Visits, Seconds, Days, Media
List of traces	
What is news?	Published by news media, published by any person/media
List of media	Tranco, Alexa, Cisco, Majestic
Top media	10, 20, 50, 100, 200, All
Information	Broad definition of news, only those identified as "political" news
Exposure	
Time threshold	1 second, 30 seconds, 120 seconds
Devices	PC only, Mobile only, All, All without apps
Tracking period	2, 5, 10, 15, 31 days

Number of visits, lasting 1 second or more, to the political articles in the top 50 most popular news websites according to Alexa, through PCs, during the last 15 days

SPECIFICATION ERRORS

Defining what qualifies as valid information

Characteristics	Potential choices
Metric	Visits, Seconds, Days, Media
List of traces	
What is news?	Published by news media, published by any person/media
List of media	Tranco, Alexa, Cisco, Majestic
Top media	10, 20, 50, 100, 200, All
Information	Broad definition of news, only those identified as "political" news
Exposure	
Time threshold	1 second, 30 seconds, 120 seconds
Devices	PC only, Mobile only, All, All without apps
Tracking period	2, 5, 10, 15, 31 days

>12k potential combinations

Defining what qualifies as valid information

Characteristics	Potential choices
Metric	Visits, Seconds, Days, Media
List of traces	
What is news?	Published by news media, published by any person/media
List of media	Tranco, Alexa, Cisco, Majestic
Top media	10, 20, 50, 100, 200, All
Information	Broad definition of news, only those identified as "political" news
Exposure	
Time threshold	1 second, 30 seconds, 120 seconds
Devices	PC only, Mobile only, All, All without apps
Tracking period	2, 5, 10, 15, 31 days

>12k potential combinations

Are all these measurements **valid measurements** of the concept of interest?

Defining what qualifies as valid information

Characteristics	Potential choices
Metric	Visits, Seconds, Days, Media
List of traces	
What is news?	Published by news media, published by any person/media
List of media	Tranco, Alexa, Cisco, Majestic
Top media	10, 20, 50, 100, 200, All
Information	Broad definition of news, only those identified as "political" news
Exposure	
Time threshold	1 second, 30 seconds, 120 seconds
Devices	PC only, Mobile only, All, All without apps
Tracking period	2, 5, 10, 15, 31 days

>12k potential combinations

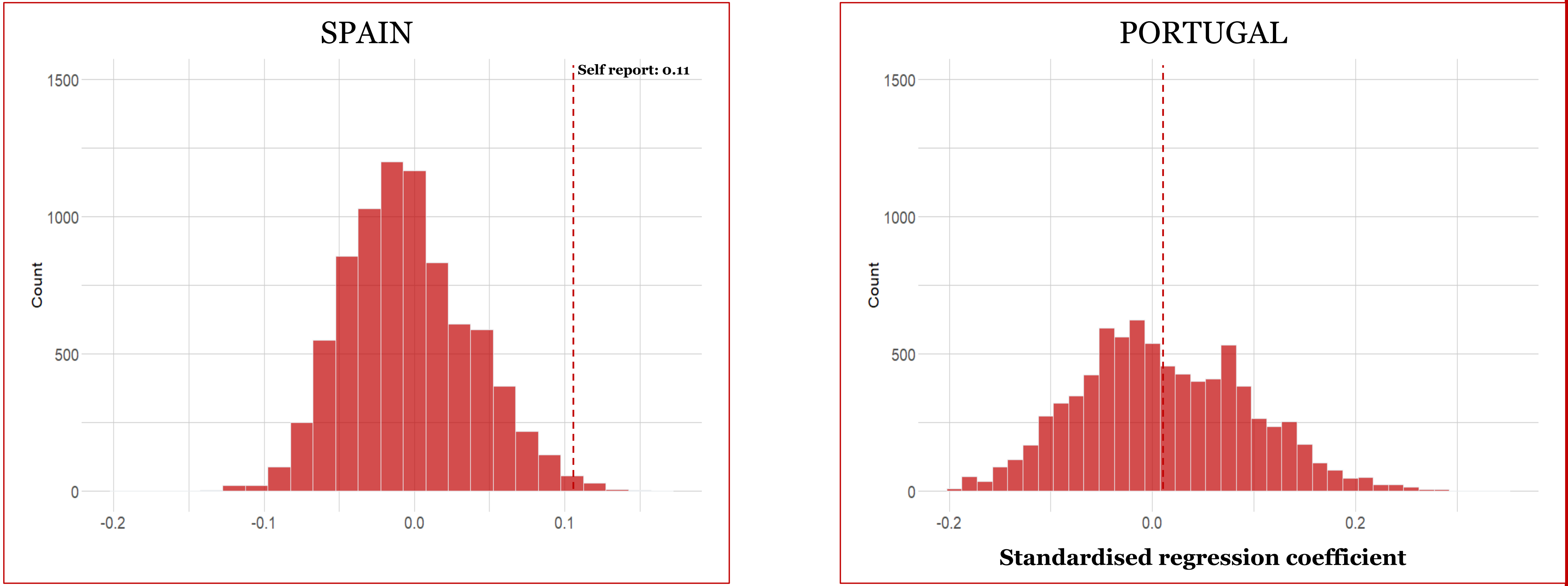
Are all these measurements **valid measurements** of the concept of interest?

If one of these **choices deviates the measurement** from the concept, **specification errors will be introduced**

SPECIFICATION ERRORS

How big of a problem is this?

Association with political knowledge across different specifications



Bosch, O. J., Sturgis, P., Kuha, J., Revilla, M. (2023). Uncovering biases in digital trace data: an assessment of the prevalence and implications of tracking undercoverage when using web tracking data

Measurement errors

Reliability

- Regardless of how valid it is, does the observed values **reflects the hypothetical true value** of our measurement?

True value → **Observed value**

Reliability

- Regardless of how valid it is, does the observed values **reflects the hypothetical true value** of our measurement?

True value → **Observed value**

Reliability

Hypothetical value: 5 minutes exposed to written news

Observed value: 3 minutes exposed to written news

MEASUREMENT ERROR

Reliability

- Regardless of how valid it is, does the observed values **reflects the hypothetical true value** of our measurement?

True value → **Observed value**

Reliability

Hypothetical value: 5 minutes exposed to written news

Observed value: 3 minutes exposed to written news

Measurement error

2 minutes of underestimation

Reliability

- Regardless of how valid it is, does the observed values **reflects the hypothetical true value** of our measurement?

True value $\xrightarrow{\text{Reliability}}$ **Observed value**

Hypothetical value: 5 minutes exposed to written news

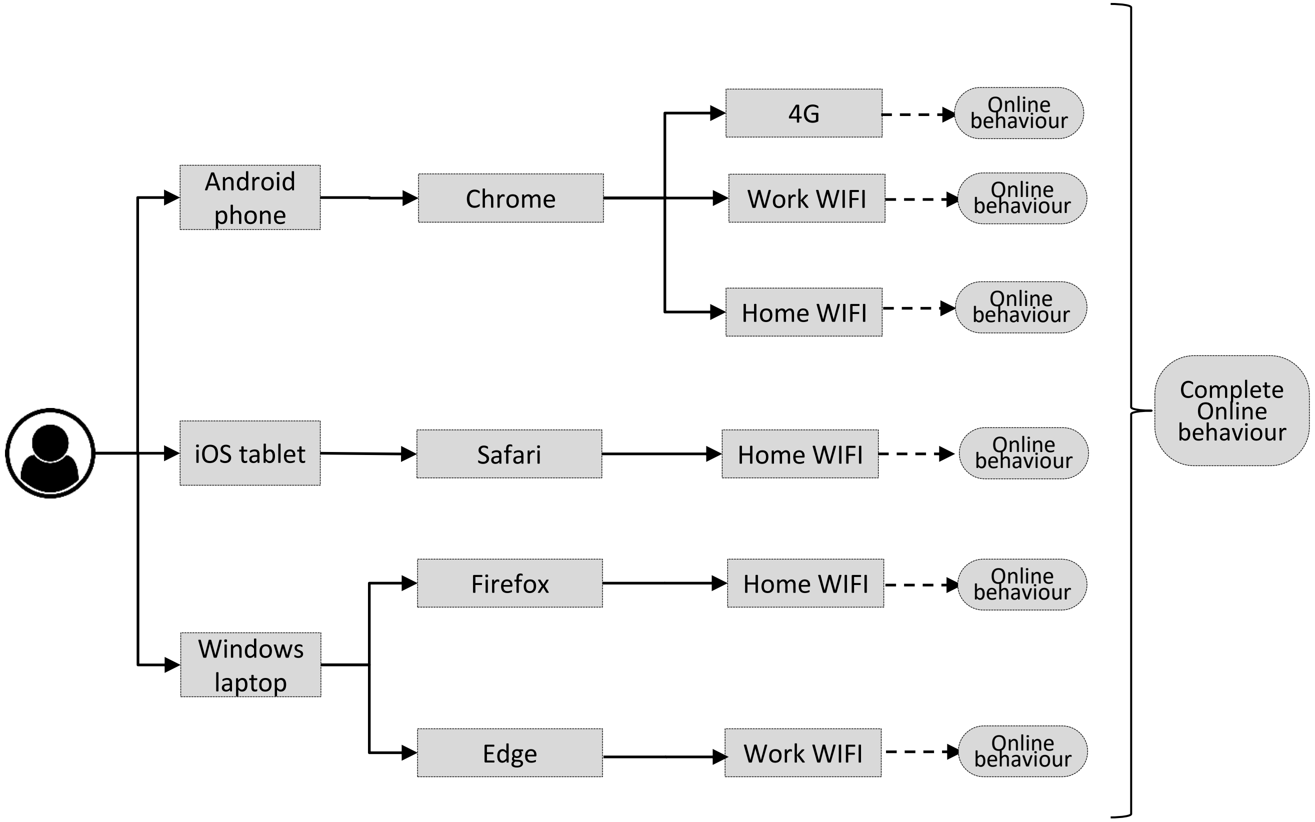
Observed value: 3 minutes exposed to written news

What can cause this?

Measurement error
2 minutes of underestimation

Cause #1: Tracking undercoverage

Cause #1: Tracking undercoverage

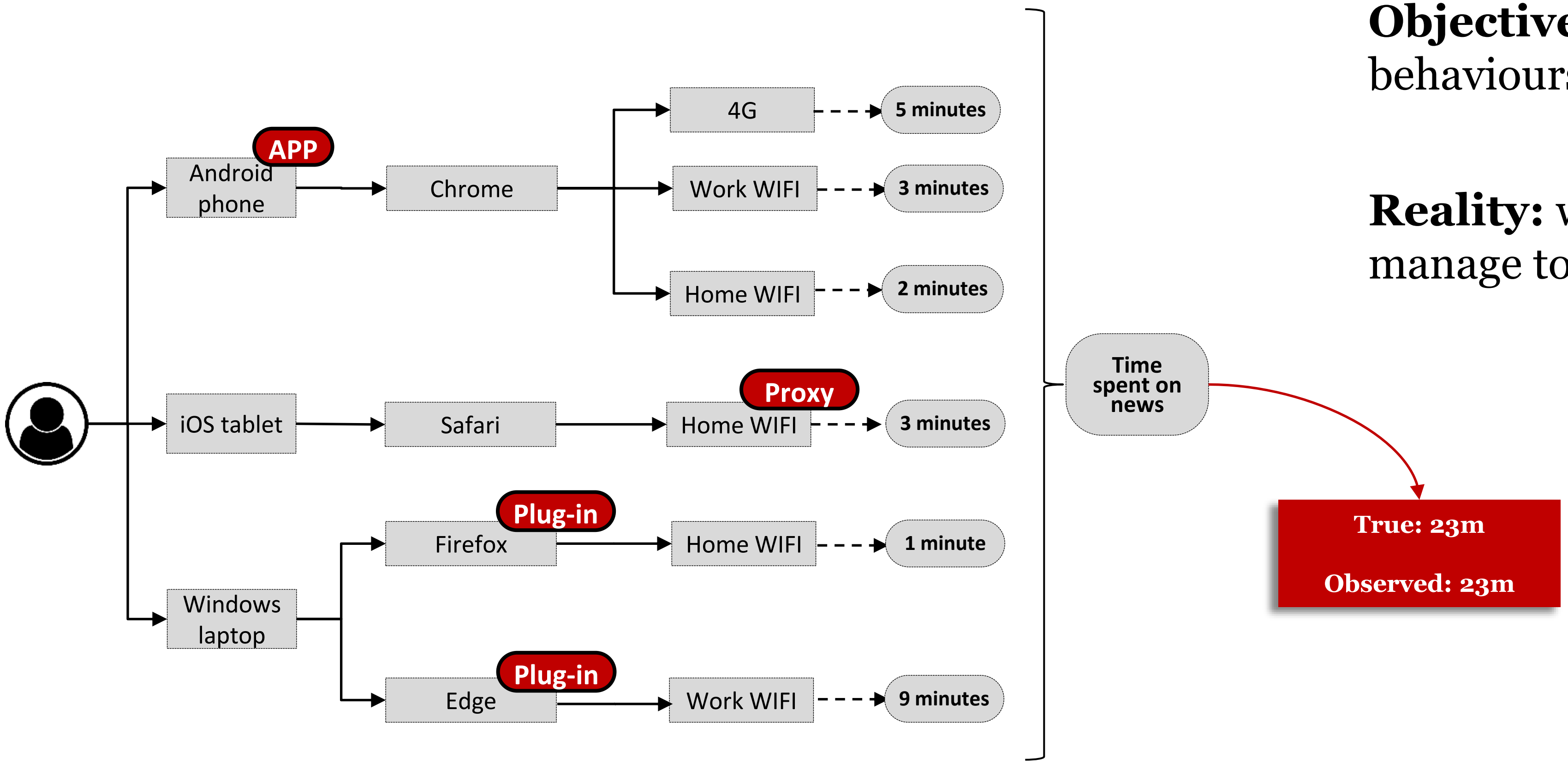


Objective: measuring individuals' behaviours.

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

MEASUREMENT ERROR

Cause #1: Tracking undercoverage

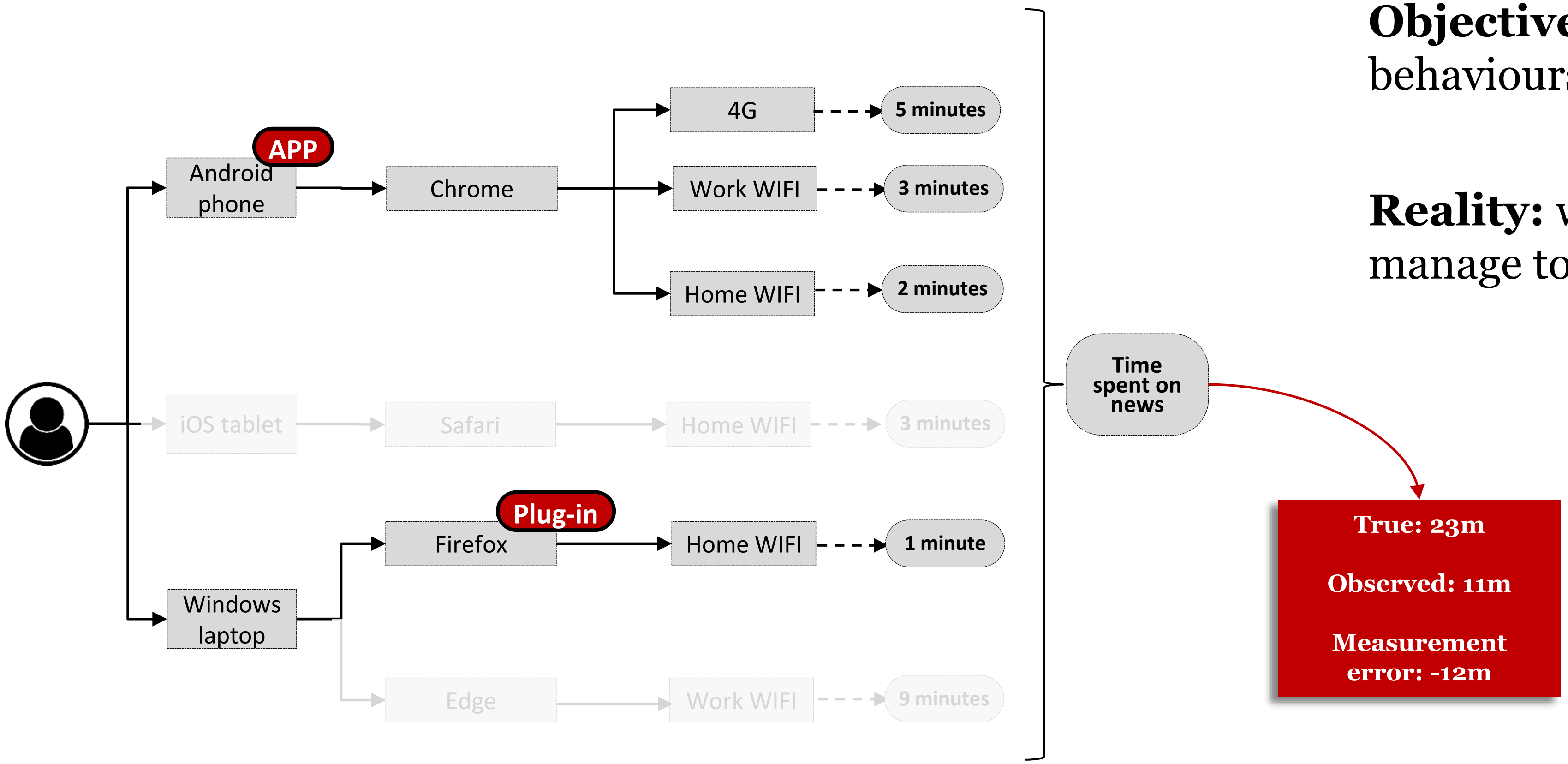


Objective: measuring individuals' behaviours.

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

MEASUREMENT ERROR

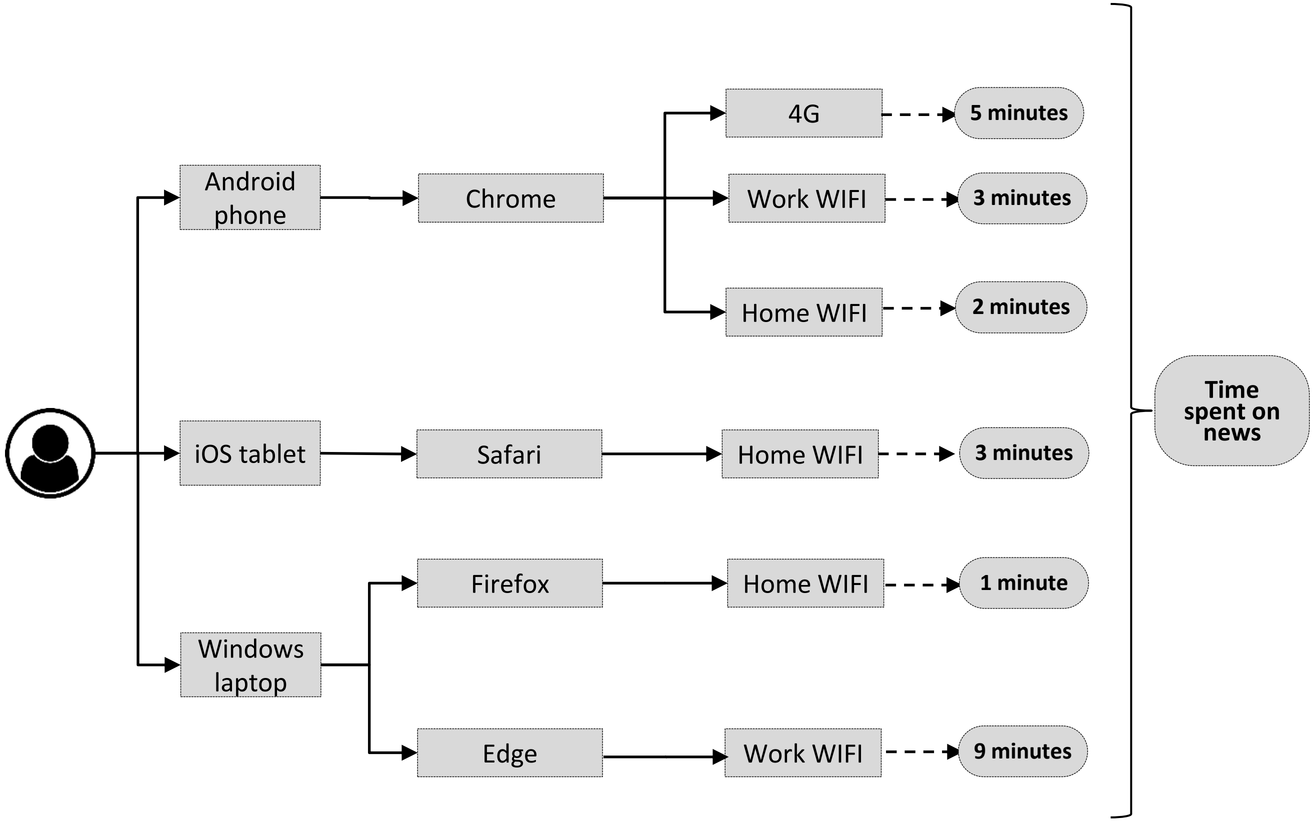
Cause #1: Tracking undercoverage



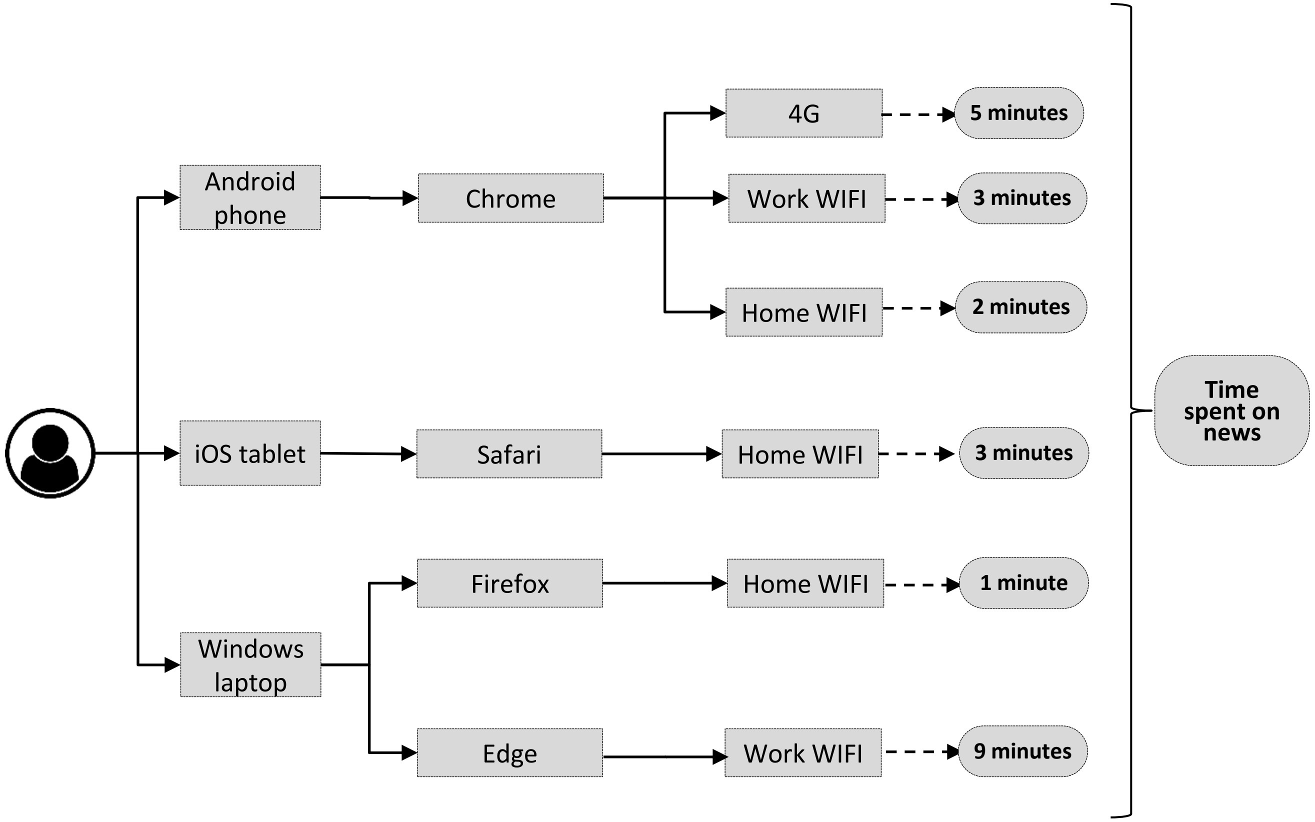
Objective: measuring individuals' behaviours.

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

Why is this happening?

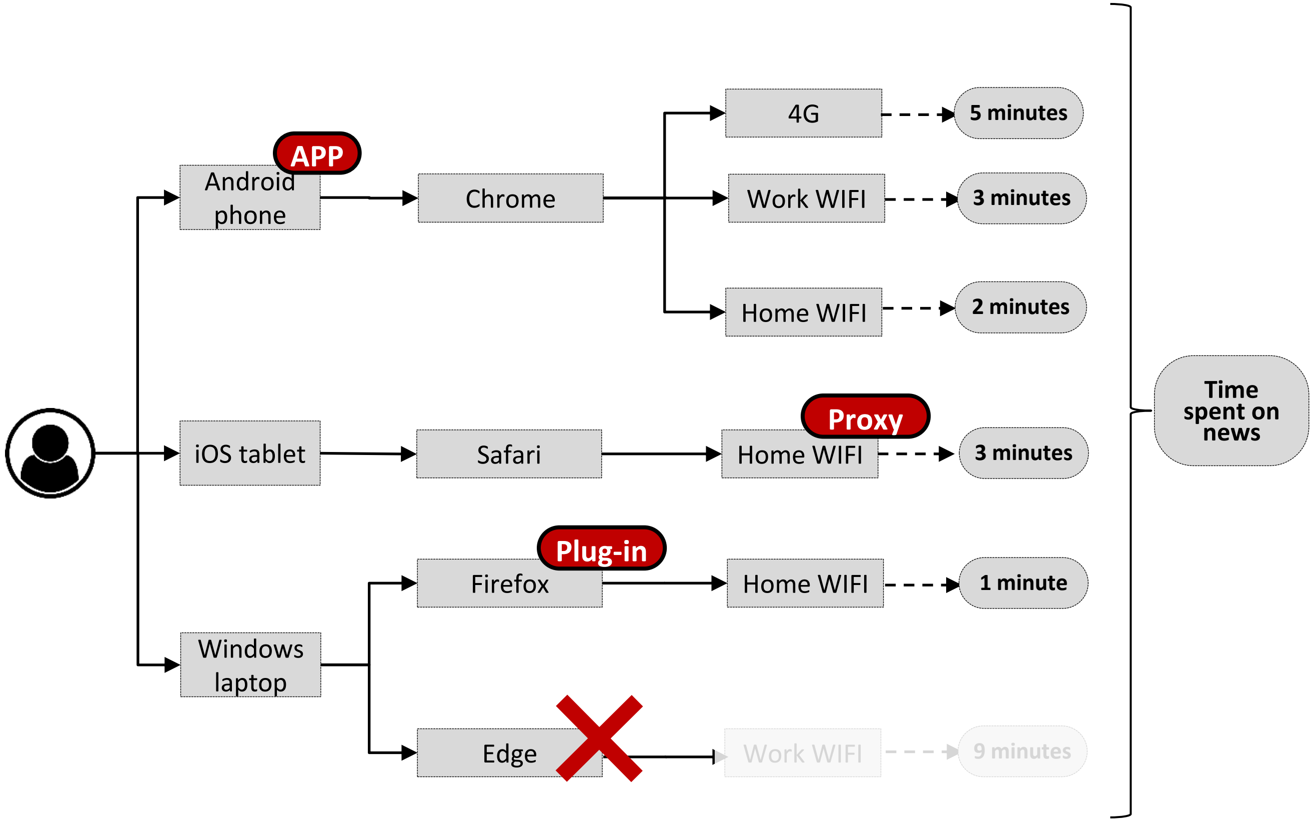


Why is this happening?



Different reasons:

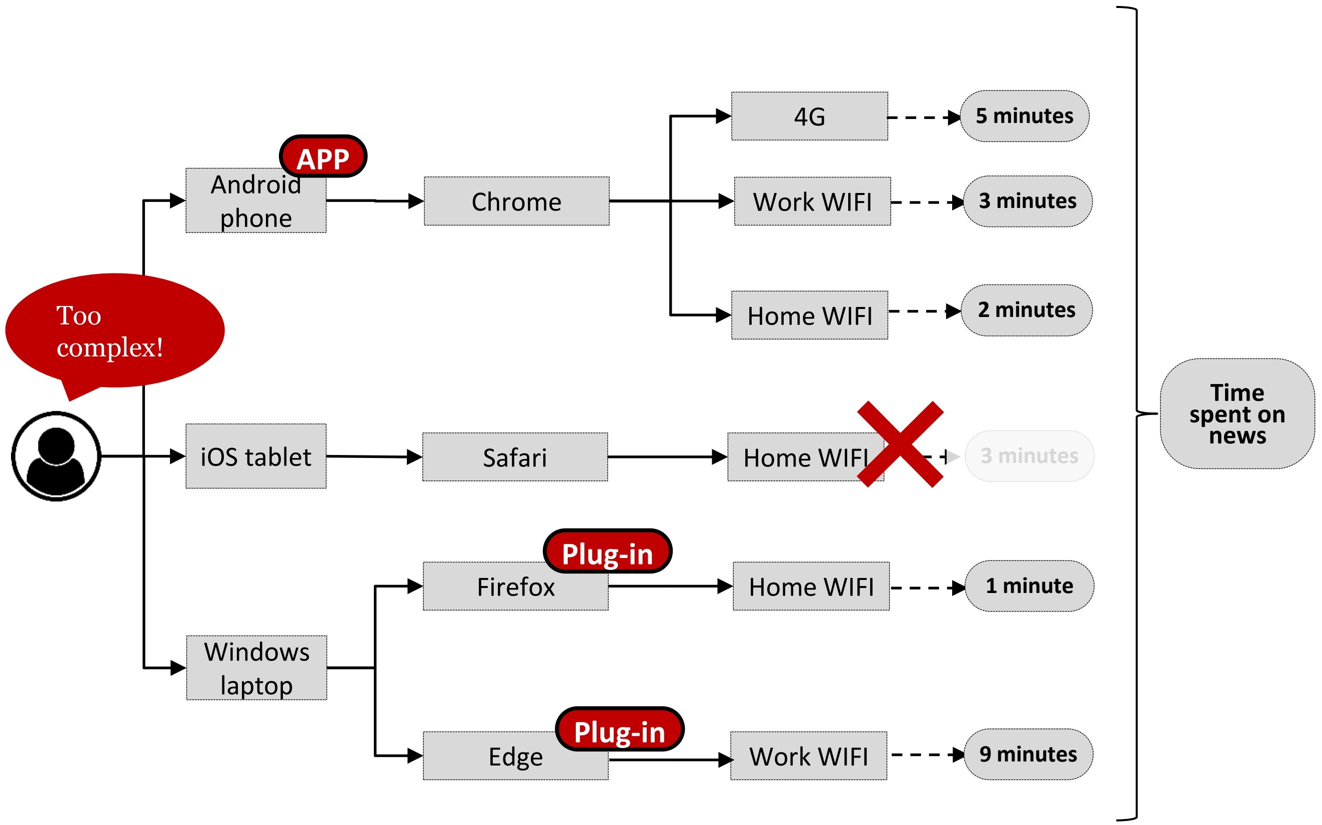
Why is this happening?



Different reasons:

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

Why is this happening?

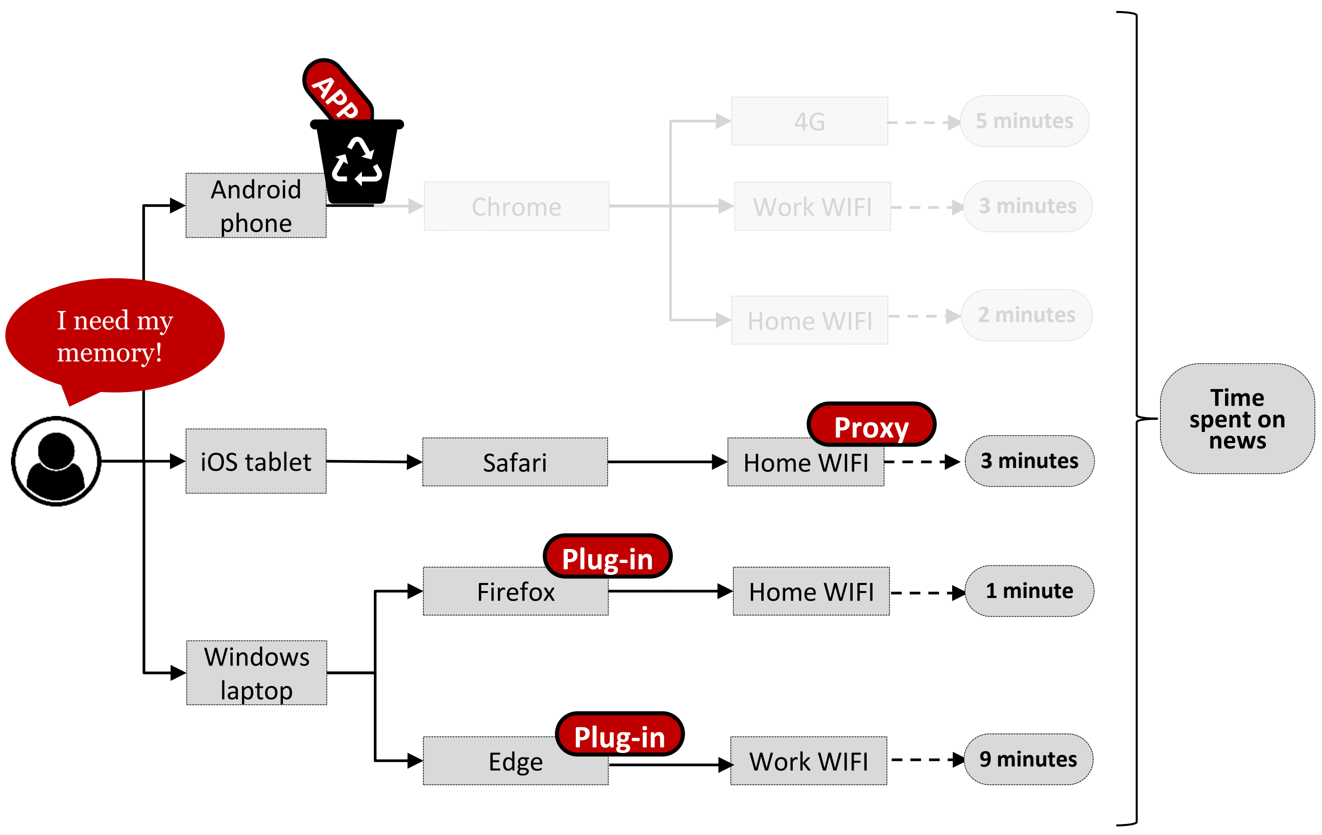


Different reasons:

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

MEASUREMENT ERROR

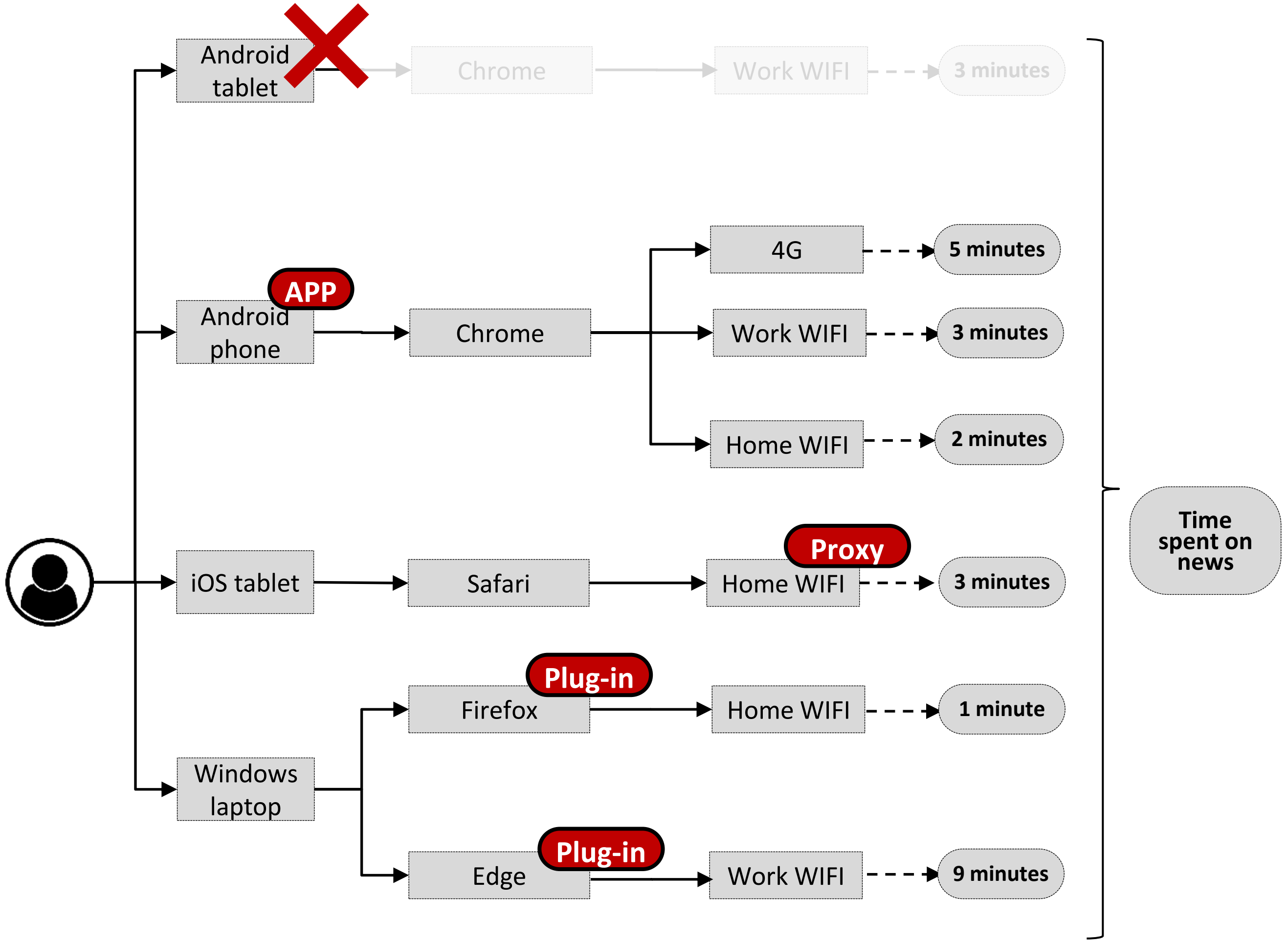
Why is this happening?



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 biases in digital trace data: an assessment of the prevalence and implications of tracking undercoverage when using web tracking data

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 biases in digital trace data: an assessment of the prevalence and implications of tracking undercoverage when using web tracking data

How big of a problem is this?

webdataopp

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 biases in digital trace data: an assessment of the prevalence and implications of tracking undercoverage when using web tracking data

How big of a problem is this?

webdataopp

Proportion of participants with all their devices tracked

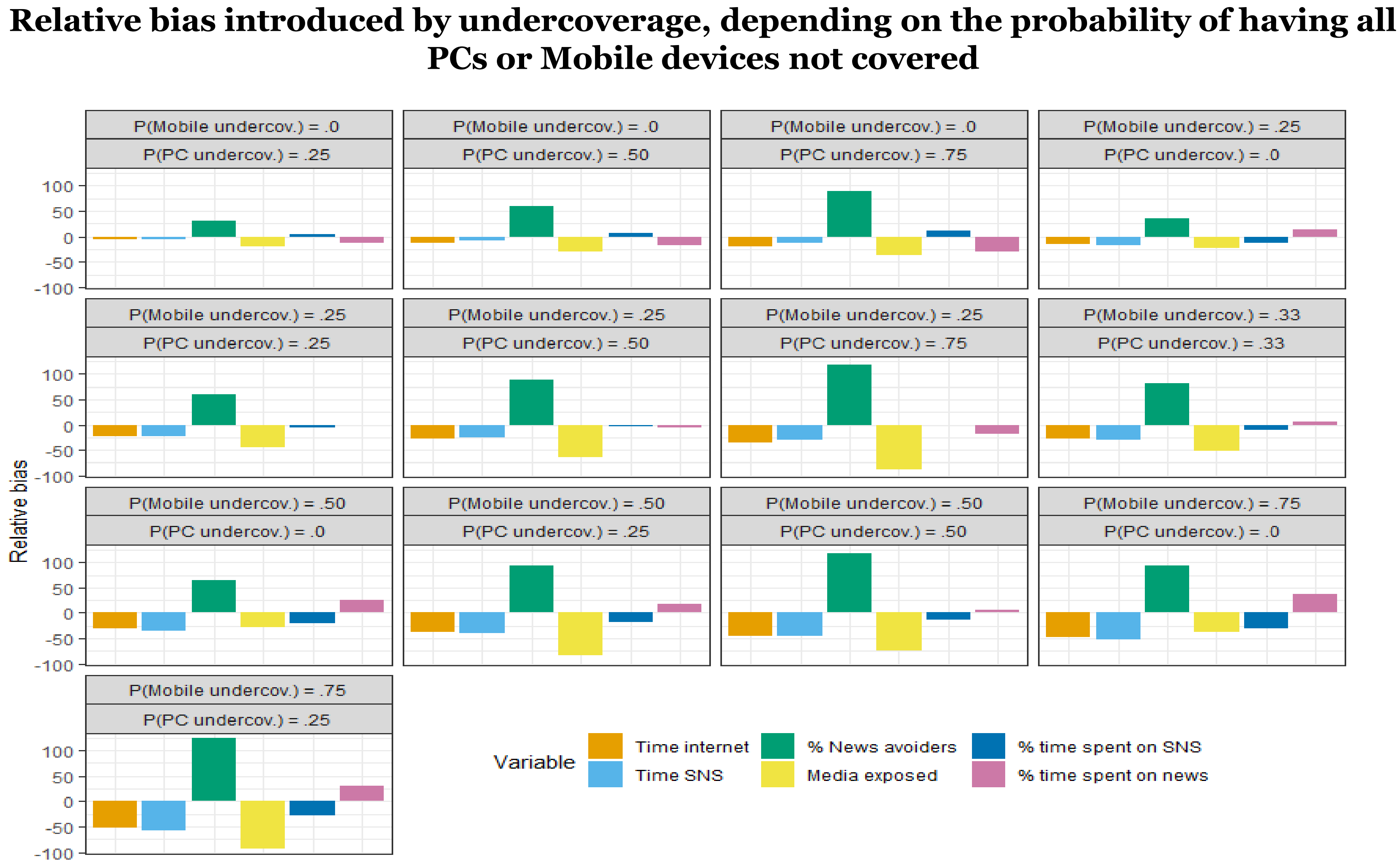
	% fully covered
All participants	26
Participants who reported using…	
… 1 device	100
… 2 devices	34
… 3 devices	13
… 4 devices	1
… +5 devices	0

	% 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 biases in digital trace data: an assessment of the prevalence and implications of tracking undercoverage when using web tracking data

How big of a problem is this?

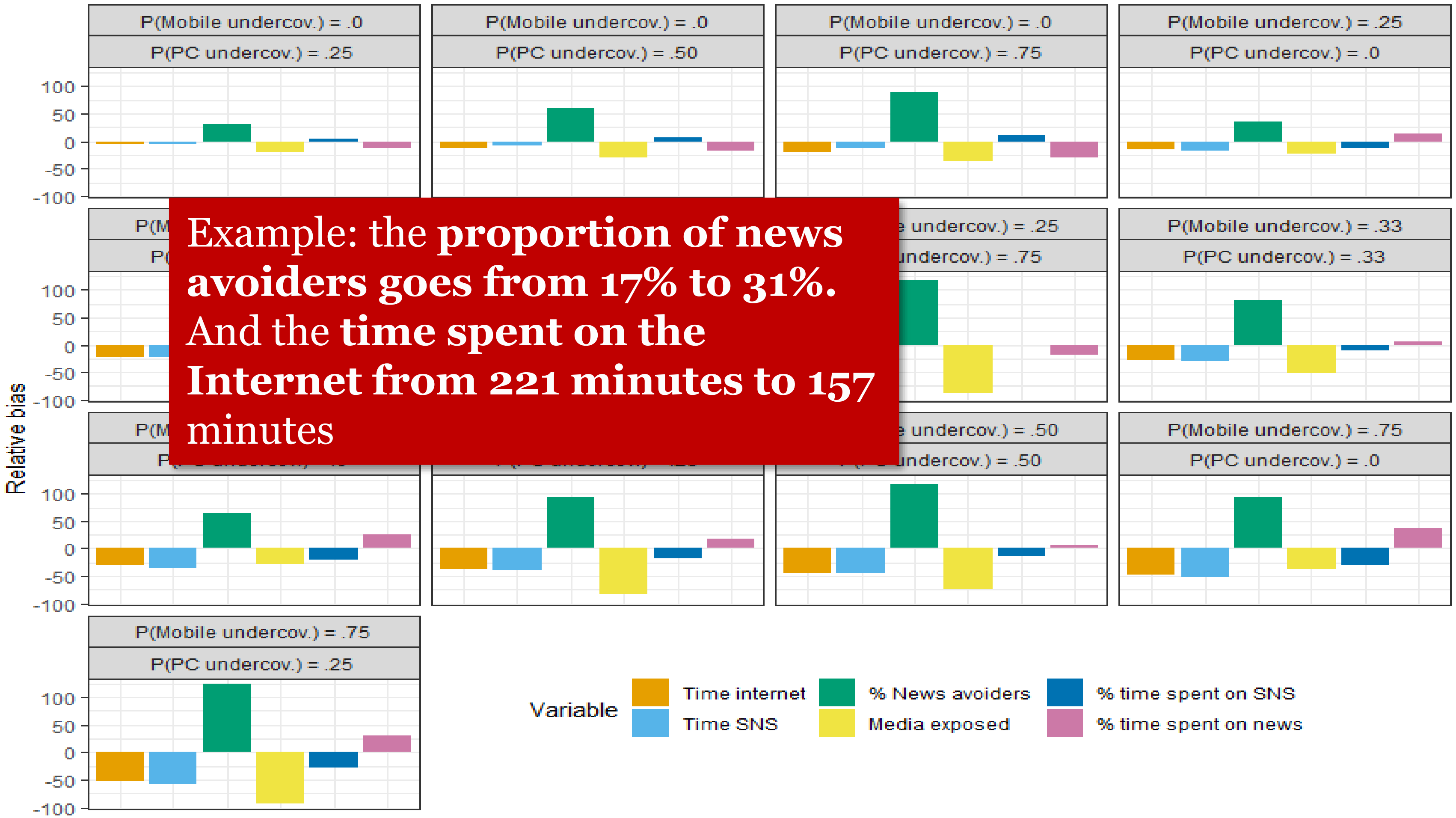


Bosch, O. J., Sturgis, P., Kuha, J., Revilla, M. (2023). Uncovering biases in digital trace data: an assessment of the prevalence and implications of tracking undercoverage when using web tracking data

MEASUREMENT ERROR

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

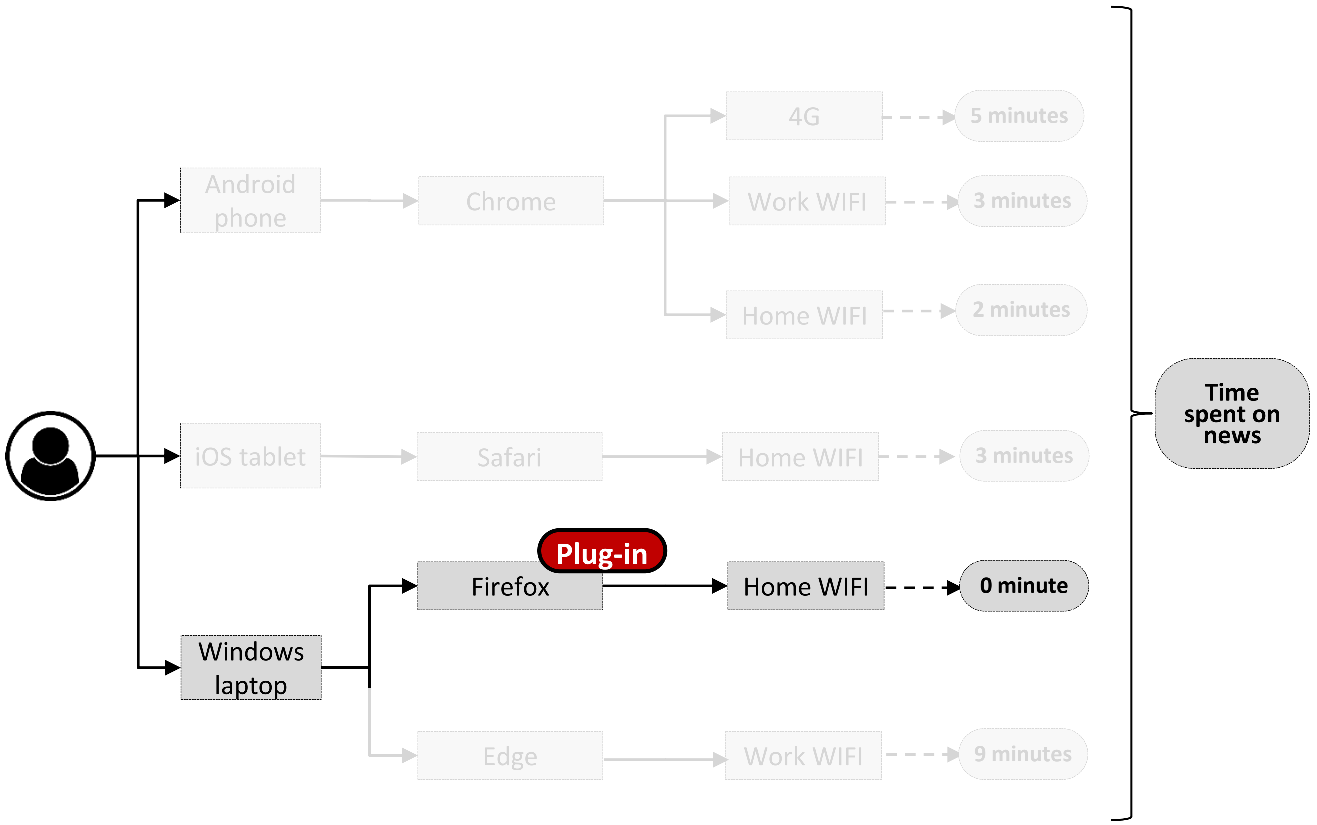


Example: the **proportion of news avoiders goes from 17% to 31%.** And the **time spent on the Internet from 221 minutes to 157** minutes

Bosch, O. J., Sturgis, P., Kuha, J., Revilla, M. (2023). Uncovering biases in digital trace data: an assessment of the prevalence and implications of tracking undercoverage when using web tracking data

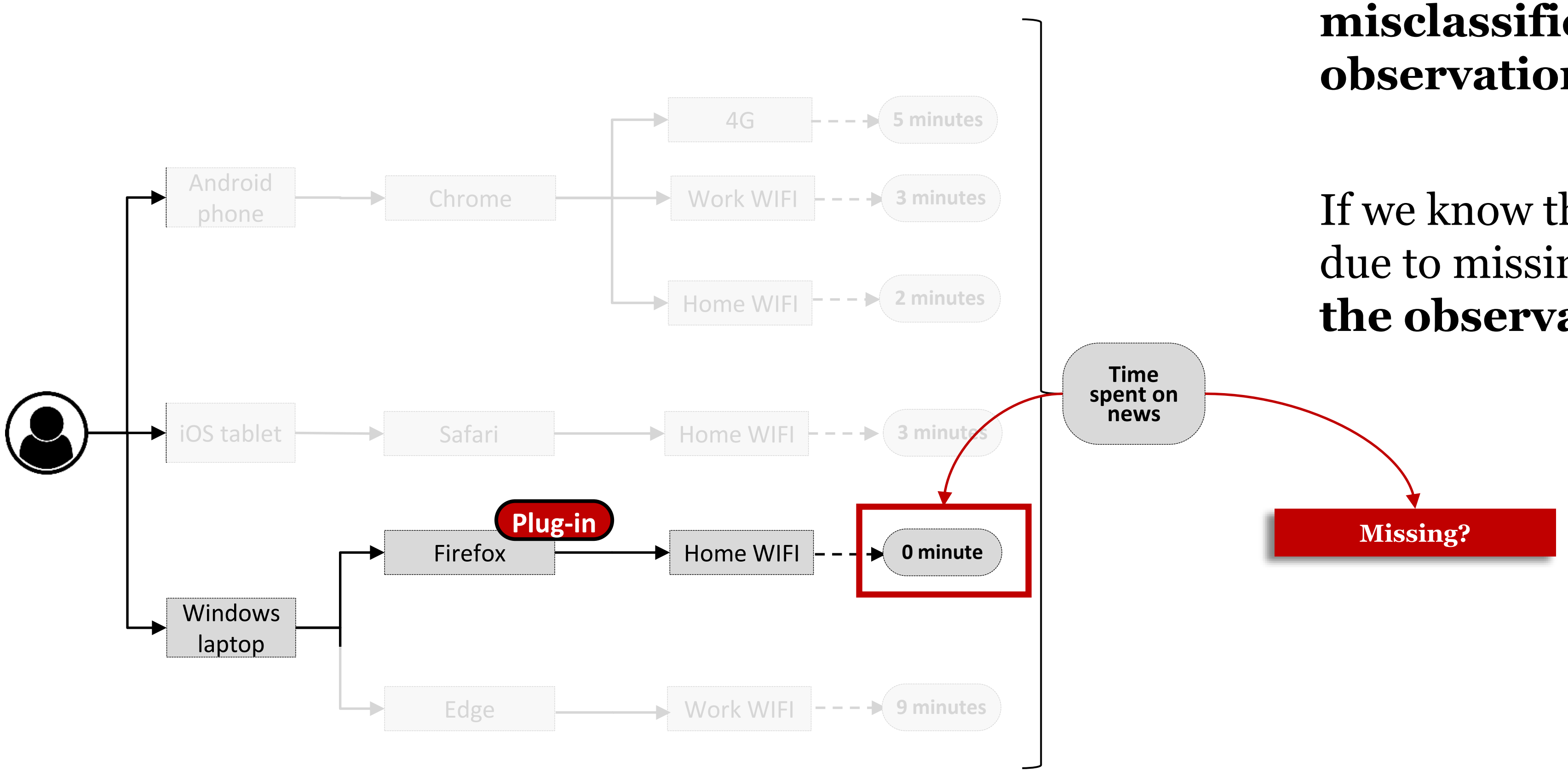
Cause #2: Misclassifying non-observations

Cause #2: Misclassifying non-observations



Sometimes, tracking undercoverage can lead to another error: **a misclassification of non-observations**

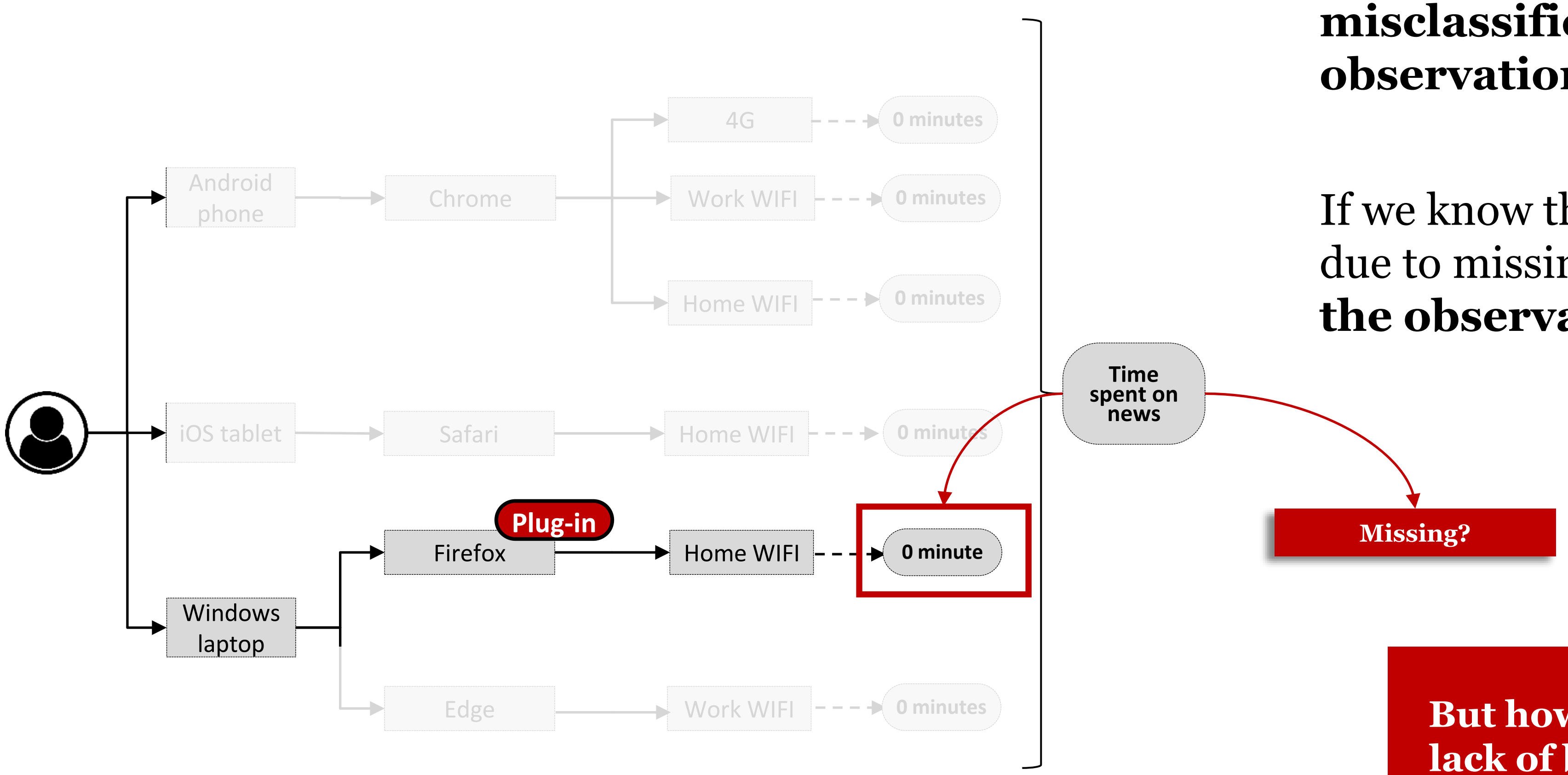
Cause #2: Misclassifying non-observations



Sometimes, tracking undercoverage can lead to another error: **a misclassification of non-observations**

If we know that a lack of information is due to missingness, **should we not treat the observation as a missing?**

Cause #2: Misclassifying non-observations



Sometimes, tracking undercoverage can lead to another error: **a misclassification of non-observations**

If we know that a lack of information is due to missingness, **should we not treat the observation as a missing?**

But how do we know that the lack of behaviour is not real?

How big of a problem is this?

webdataopp

Proportion of participants with error-induced non-observations

	Italy	Portugal	Spain	Argentina	Chile
Facebook	10.5	10.6	11.1	9.8	10.9
Twitter	23.0	17.7	14.7	16.1	21.1
Avg. news outlets	9.0	18.8	11.8	10.0	17.5

Torcal, M., Carty, E., Comellas, J. M., Bosch, O. J., Thomson, Z., & Serani, D. (2023). The dynamics of political and affective polarisation: Datasets for Spain, Portugal, Italy, Argentina, and Chile (2019-2022). *Data in Brief*, *48*, 109219.

How big of a problem is this?

web data opp

Proportion of participants with error-induced non-observations

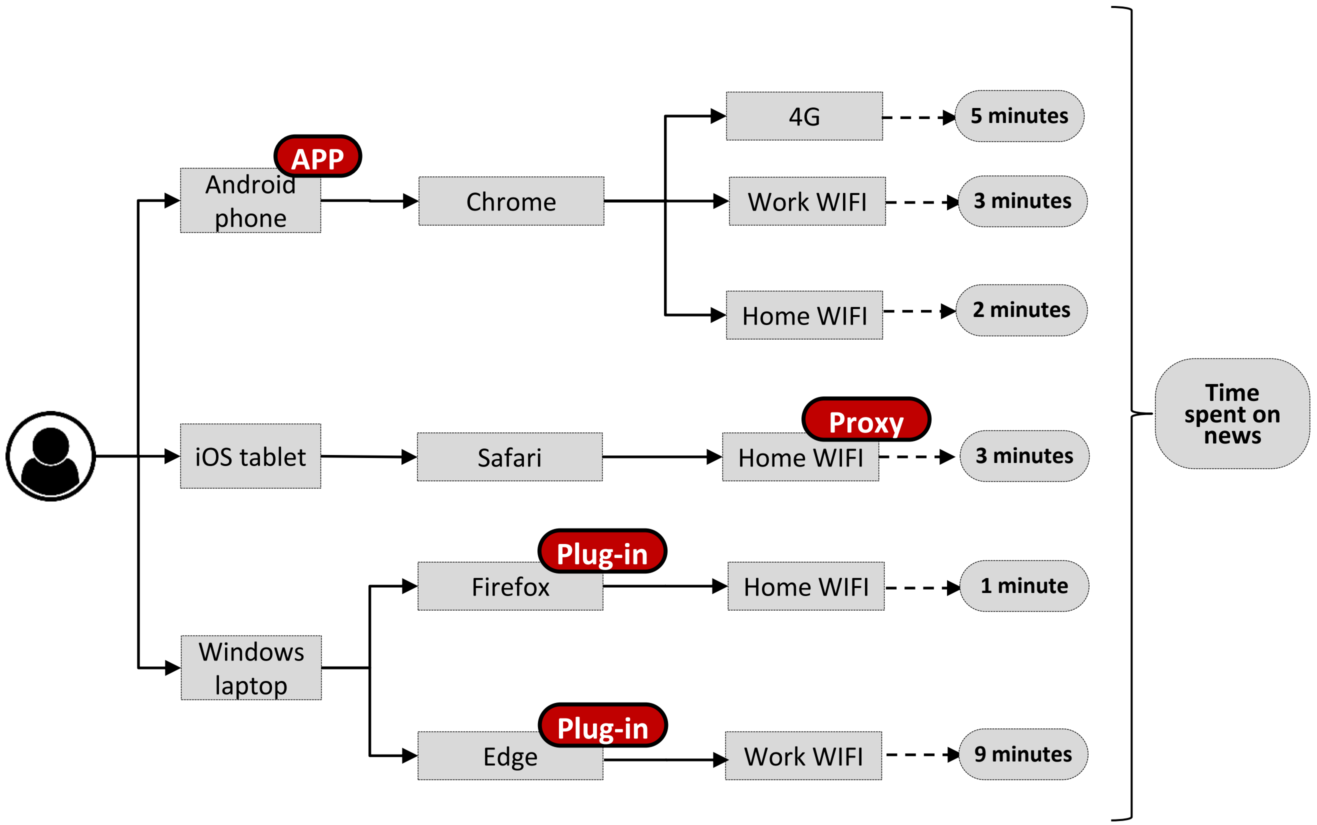
	Italy	Portugal	Spain	Argentina	Chile
Facebook	10.5	10.6	11.1	9.8	10.9
Twitter	23.0	17.7	14.7	16.1	21.1
Avg. news outlets	9.0	18.8	11.8	10.0	17.5

There is a non-negligible risk of **increasing the size of the estimate's measurement errors** if these participants are **not excluded** from the analyses

Torcal, M., Carty, E., Comellas, J. M., Bosch, O. J., Thomson, Z., & Serani, D. (2023). The dynamics of political and affective polarisation: Datasets for Spain, Portugal, Italy, Argentina, and Chile (2019-2022). *Data in Brief*, *48*, 109219.

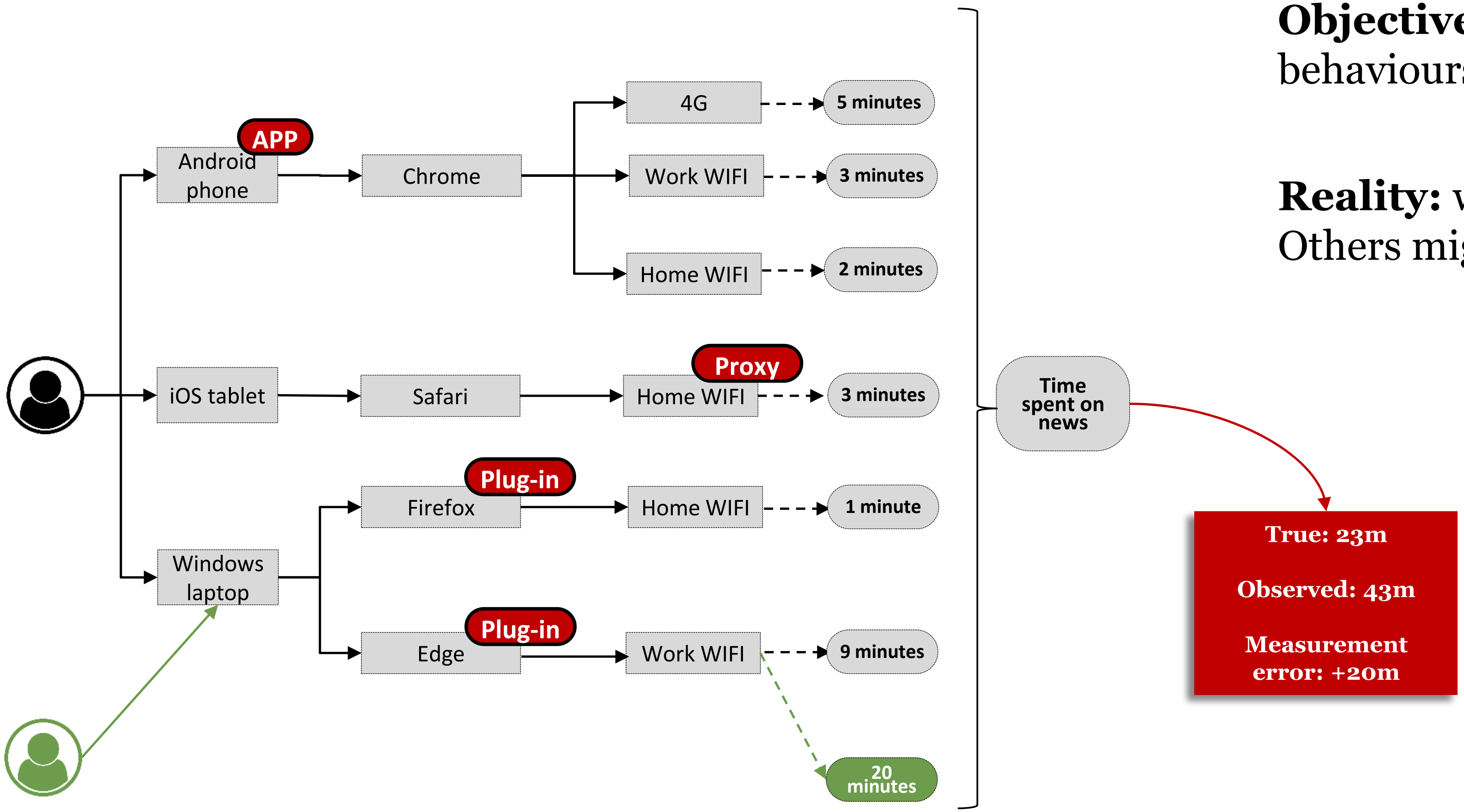
Cause #3: Shared devices

Cause #3: Shared devices



Objective: measuring individuals' behaviours.

Cause #3: Shared devices



Objective: measuring individuals' behaviours.

Reality: we measure devices, not people. Others might use the devices that we track

How big of a problem is this?

webdataopp

60% Desktops are shared

40% Laptops and tablets

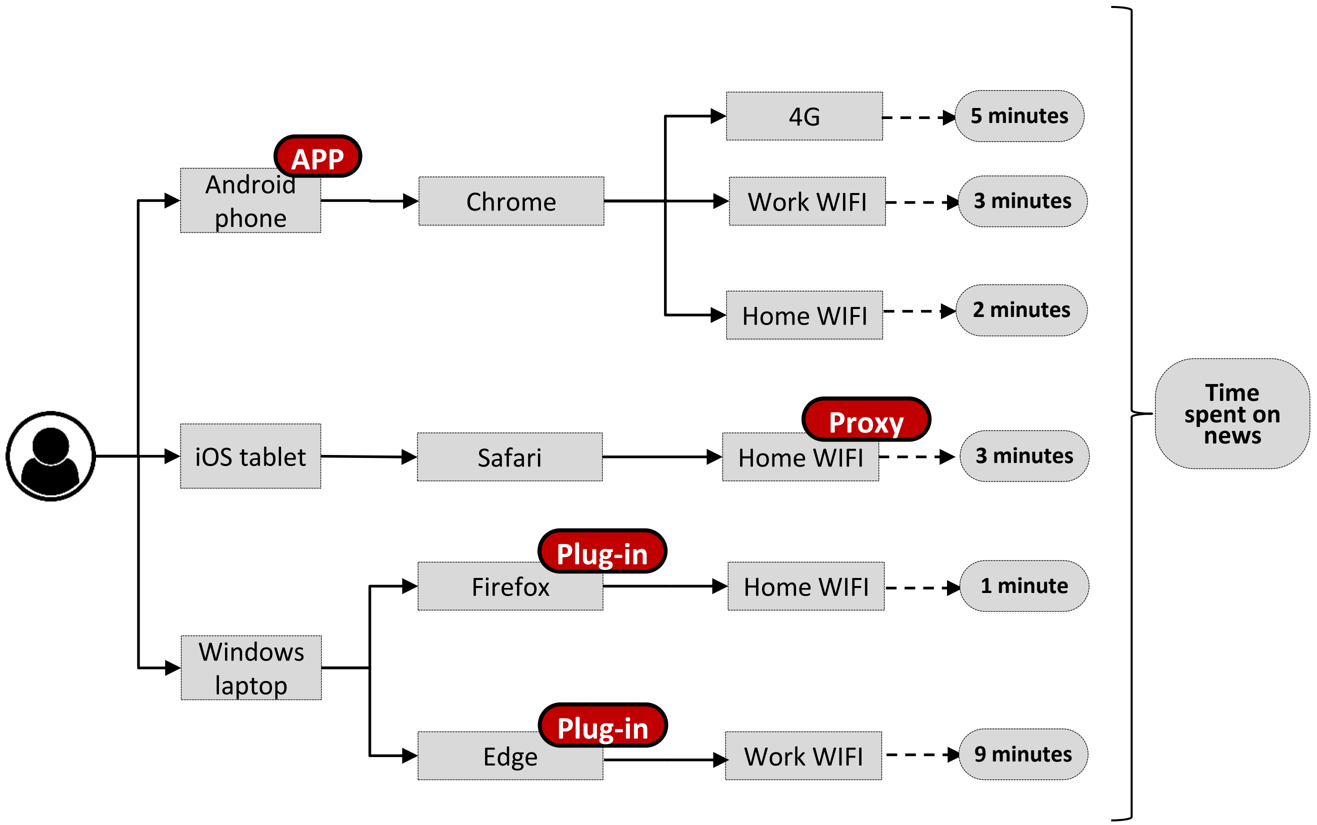
9% Smartphones

Netquest (Spain)

Revilla, M., Ochoa, C., & Loewe, G. (2017). Using passive data from a meter to complement survey data in order to study online behavior. *Social Science Computer Review*, *35*(4), 521-536.

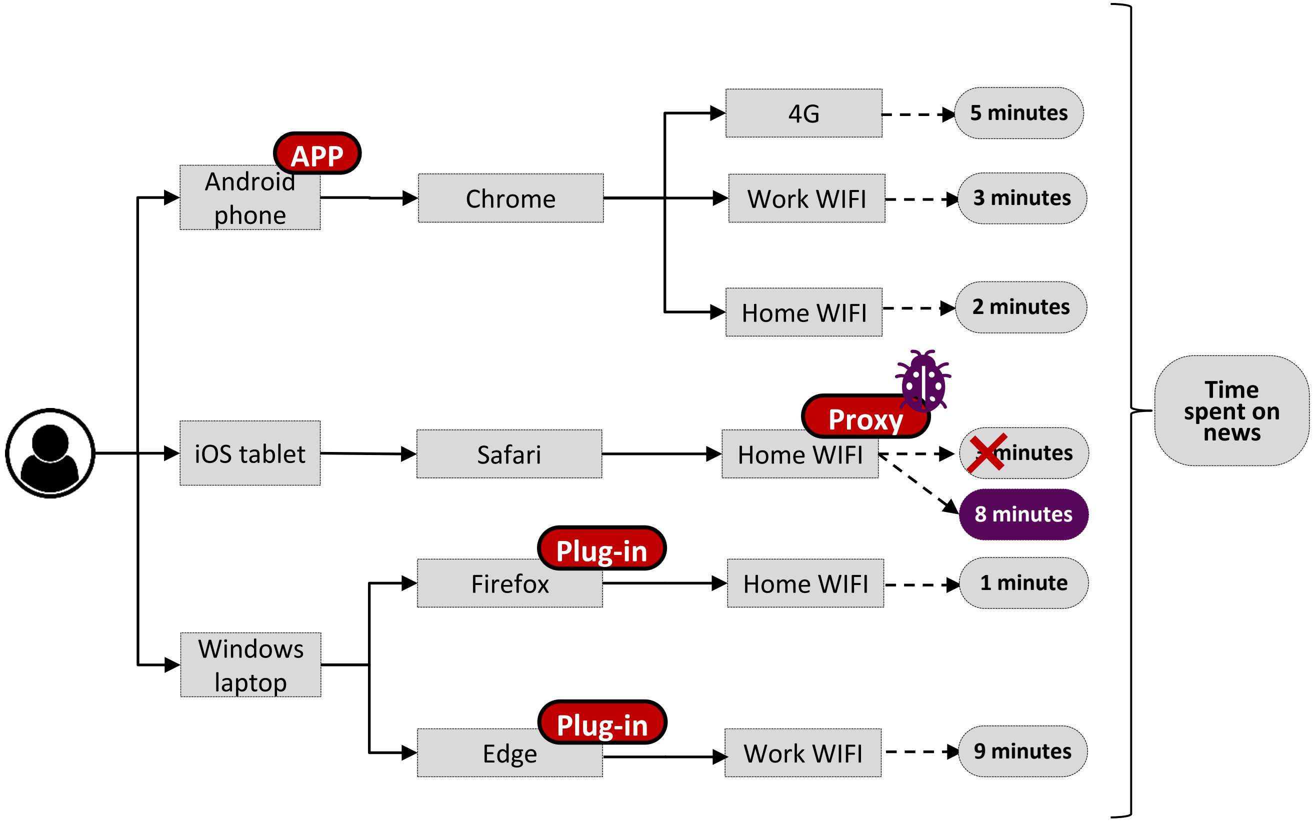
Cause #4: Technology errors

Cause #4: Technology errors



Objective: measuring individuals' behaviours.

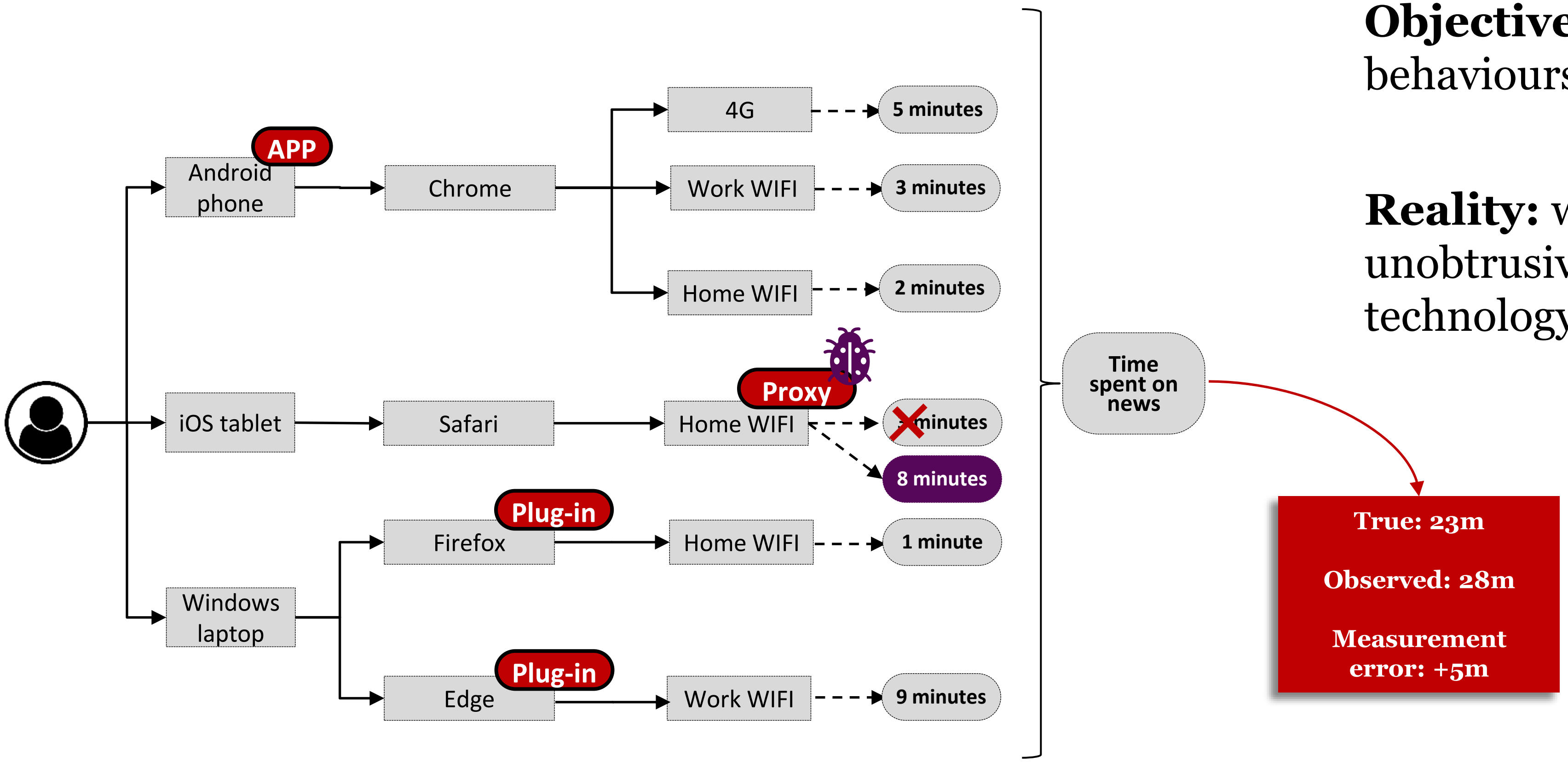
Cause #4: Technology errors



Objective: measuring individuals' behaviours.

Reality: we do not observe behaviours unobtrusively. We measure what the technology captures…which might be wrong

Cause #4: Technology errors



Objective: measuring individuals' behaviours.

Reality: we do not observe behaviours unobtrusively. We measure what the technology captures…which might be wrong

Why is this happening?

webdataopp

1. The devices or third-party apps might **shut down the ability to collect data** when devices are **low on battery**

2. Proxies generates **raw data that must be processed to identify** which part of the tracked traffic was done **passively** by the device (e.g., downloading Facebook information) or **actively** by the participant. This is normally done by trained algorithms. However, this is **not completely accurate**.

3. Since tracking technologies are built on top of OSs and browsers when **new versions of the software** are released, they **can prevent the technologies from working**, causing a loss of information until the technology is adapted to the new version

How big of a problem is this?

Determinant of absolute difference between self-report and web tracking data

	Italy	Portugal	Spain
Tracked on iOS	**57.6****	**35.1***	**56.8***
Internet use	.4**	.2**	.2**
Mobile use	-45.4	-21.5	17.5
Tracking undercovered	12.6	7.1	9.9
Months as panellist	-.1	-.1	.0
Gender	-12.6	5.8	9.3
Age	-1.0*	-.4	-.6*
Educational level	-.8	.0	-1.0**
Constant	189.11*	129.0**	84.6**
Adjusted R^2	.22	.08	.10
N	751	774	908

Absolute error: $|Self - reported\ time\ on\ the\ Internet - Tracked\ time\ on\ the\ Internet|$

Being tracked on an iOS device is associated with having and **absolute difference 35.1 - 57.6** min larger than for those not tracked on an iOS

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.

How big of a problem is this?

webdataopp

Determinant of absolute difference between self-report and web tracking data

	Italy	Portugal	Spain
Tracked on iOS	**57.6****	**35.1***	**56.8***
Internet use	.4**	.2**	.2**
Mobile use	-45.4	-21.5	17.5
Tracking undercovered	12.6	7.1	9.9
Months as panellist	-.1	-.1	.0
Gender	-12.6	5.8	9.3
Age	-1.0*	-.4	-.6*
Educational level	-.8	.0	-1.0**
Constant	189.11*	129.0**	84.6**
Adjusted R^2	.22	.08	.10
N	751	774	908

Absolute error: $|Self - reported\ time\ on\ the\ Internet - Tracked\ time\ on\ the\ Internet|$

Being tracked on an iOS device is associated with having and **absolute difference 35.1 - 57.6** min larger than for those not tracked on an iOS…**do measures from iOS have different measurement properties?**

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.

Cause #5: Technology limitations

Cause #5: Technology limitations

		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

Cause #5: Technology limitations

web data opp

		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

If behaviours happen inside apps, we miss them

Cause #5: Technology limitations

		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

If behaviours happen inside apps, we miss them

If the measurement requires HTML data, only desktops will be trackable

Cause #5: Technology limitations

web data opp

		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

If behaviours happen inside apps, we miss them

If the measurement requires HTML data, only desktops will be trackable.

If behaviours happen in HTTPs webpages, some meter will miss that

Cause #5: Technology limitations

		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

If behaviours happen inside apps, we miss them

If the measurement requires HTML data, only desktops will be trackable.

If behaviours happen in HTTPs webpages, some meter will miss that

Etc…

Other causes

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

Other causes

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

Participants can stop sending their data

Other causes

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

Participants might **change their behaviours once tracked**…but there is **no evidence** of this!

Keusch, F., Bach, R., & Cernat, A. (2023). Reactivity in measuring sensitive online behavior. *Internet Research*, *33*(3), 1031-1052.

Other causes

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

Errors can happen **when extracting the raw data, and transforming it** into variables…in my experience, **very common with panel providers**

Overall

- The few evidence available (nothing published yet) has found that the reliability of measures created with web tracking data is not perfect.

- My own research has shown a reliability on average of around .70. But for some measures other researchers have found values even lower than .50!

- Self-promotion: you can come check my presentation tomorrow to see some preliminary MTMM results

Coverage errors

Non-trackable individuals

Apps	Plug-in A	Plug-in B	Proxy
Where? Device	**Where?** Browser	**Where?** Browser	**Where?** Network
Devices Not iOS	**Devices** Only PC & MAC	**Devices** Only PC & MAC	**Devices** All
Continuous? Yes	**Continuous?** Yes	**Continuous?** No	**Continuous?** Yes
Types of data URLs, Time, Device, Search terms, Incognito	**Types of data** URLs, Time, Device, Search terms, Incognito, HTML	**Types of data** URLs, Time, Device	**Types of data** URLs, Time, Device

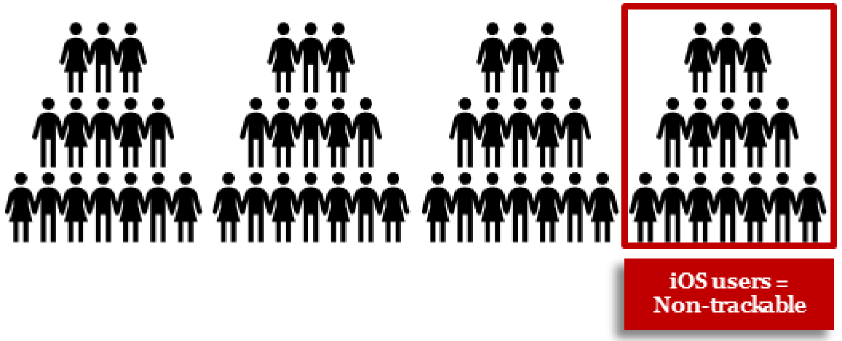
Non-trackable individuals

Apps	Plug-in A
Where? Device	**Where?** Browser
Devices Not iOS	**Devices** Only PC & MAC
Continuous? Yes	**Continuous?** Yes
Types of data URLs, Time, Device, Search terms, Incognito	**Types of data** URLs, Time, Device, Search terms, Incognito, HTML

Depending on the technologies we use, and their capabilities, some participants might not be trackable at all.

Non-trackable individuals

Apps	Plug-in A
Where? Device	**Where?** Browser
Devices Not iOS	**Devices** Only PC & MAC
Continuous? Yes	**Continuous?** Yes
Types of data URLs, Time, Device, Search terms, Incognito	**Types of data** URLs, Time, Device, Search terms, Incognito, HTML



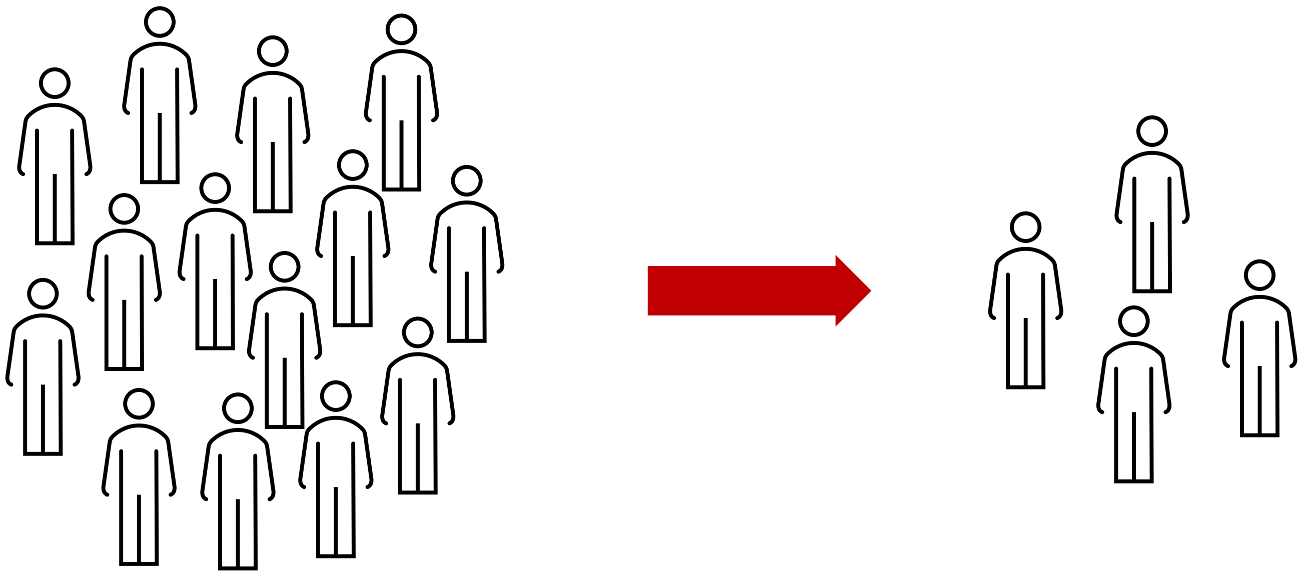
Depending on the technologies we use, and their capabilities, some participants might not be trackable at all.

For instance: people only using iOS devices

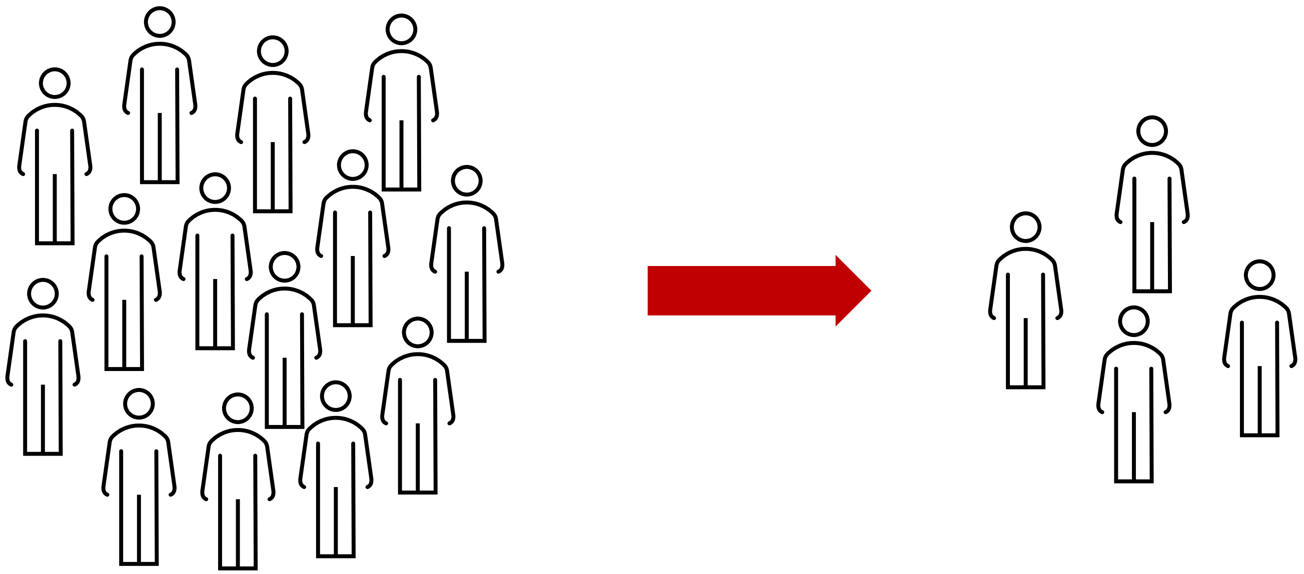
Missing data errors

MISSING DATA ERRORS

From a sample…to a sample of tracked participants



From a sample…to a sample of tracked participants



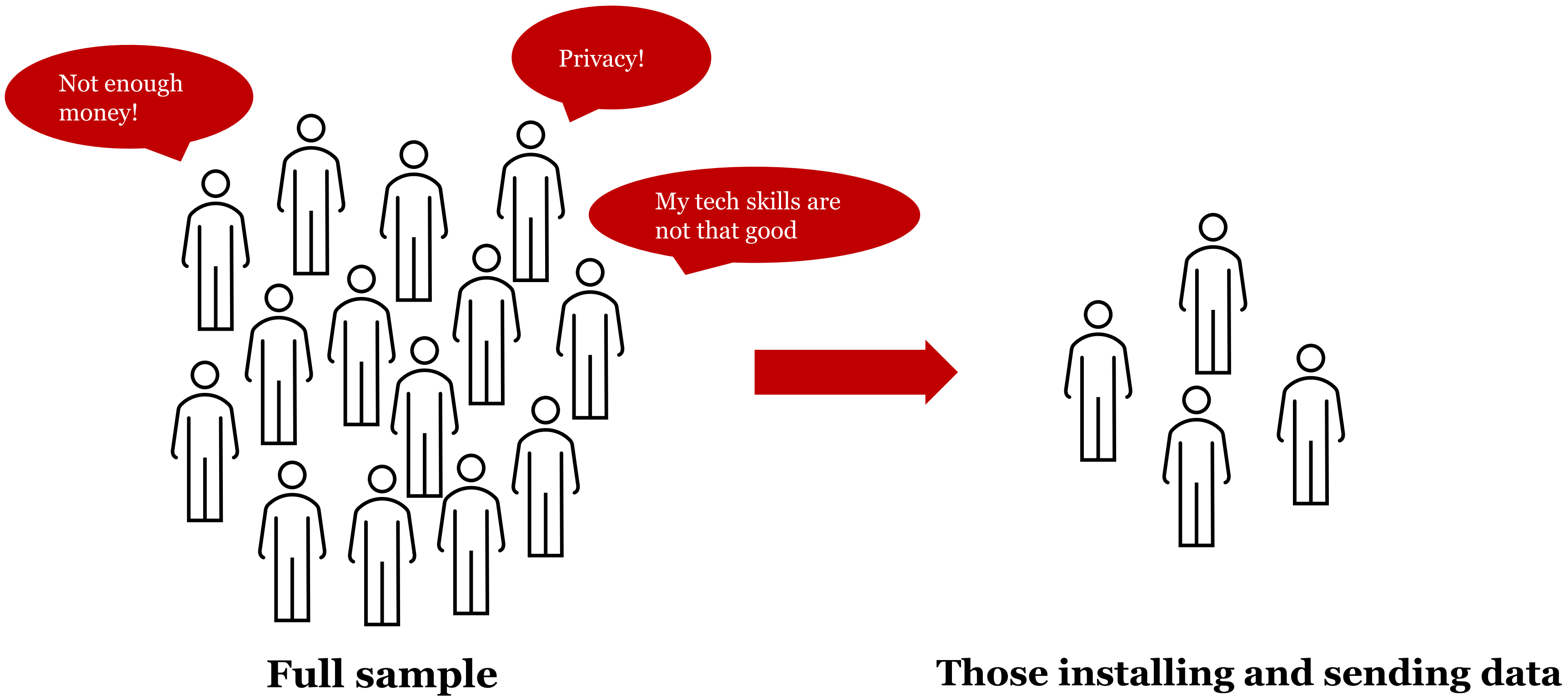
Full sample **Those installing and sending data**

If missing data differ systematically from the available data, biases are introduced

Main cause: non-consent

MISSING DATA ERRORS

Main cause: non-consent



Main cause: non-consent

Table 2 Main reasons* why panelists would accept or not accept the invitation to install a tracking application on their PC

Main reasons for accepting	% (based on N= 171 respondents)
I don't mind/not confidential	37.4
Incentive	25.1
Altruism	14.0
Trust	9.9
Main reasons for not accepting	**% (based on N= 829 respondents)**
Privacy	72.6
No trust	7.0
No reason	5.4
I do not own the PC I use	5.8

Note: * We present all reasons that are mentioned by at least 5% of the respondents. When a respondent provided several reasons, we take them all into account.

Table 3. Reasons for and against participation in passive mobile data collection (*n* = 1,947)

Reasons for not participating		Reasons for participating	
Privacy, data security concerns	44%	Interest, curiosity	39%
No incentive; incentive too low	17%	Incentive	26%
Not enough information/control of what happens with data	12%	Help research, researcher	18%
Do not download apps	7%	Trust, seems legitimate, safe	11%
Not interested, no benefit	6%	Will help create better products & services	7%
Not enough time, study too long	5%	No additional burden	6%
Do not use smartphone enough; not right person for this study	5%	Like surveys & research	4%
Not enough storage	1%	Fun	3%
Other reasons	6%	Other reasons	4%
NA	3%	NA	2%

NOTE.—Percentages do not add up to 100 because respondents could mention multiple reasons.

Revilla, M., Couper, M. P., & Ochoa, C. (2019). Willingness of online panelists to perform additional tasks. *Methods, data, analyses: a journal for quantitative methods and survey methodology (mda)*, *13*(2), 223-252.
Keusch, F., Struminskaya, B., Antoun, C., Couper, M. P., & Kreuter, F. (2019). Willingness to participate in passive mobile data collection. *Public opinion quarterly*, *83*(S1), 210-235.

What can we do about these problems?

Strategies to prevent, identify, correct, and report ***specification errors***

#1: Better defining what qualifies as valid information

#1: Better defining what qualifies as valid information

- Before defining any measurement, list the different design decisions that you will have to make in order to operationalise the concept of interest

Characteristics	Potential choices
Metric	
List of traces	
What is news?	
List of media	
Top media	
Information	
Exposure	
Time threshold	
Devices	
Tracking period	

#1: Better defining what qualifies as valid information

- Before defining any measurement, list the different design decisions that you will have to make in order to operationalise the concept of interest
- List the potential choices that you could make within each design decision

Characteristics	Potential choices
Metric	Visits, Seconds, Days, Media
List of traces	
What is news?	Published by news media, published by any person/media
List of media	Tranco, Alexa, Cisco, Majestic
Top media	10, 20, 50, 100, 200, All
Information	Broad definition of news, only those identified as “political” news
Exposure	
Time threshold	1 second, 30 seconds, 120 seconds
Devices	PC only, Mobile only, All, All without apps
Tracking period	2, 5, 10, 15, 31 days

#1: Better defining what qualifies as valid information

webdataopp

- Before defining any measurement, list the different design decisions that you will have to make in order to operationalise the concept of interest
- List the potential choices that you could make within each design decision

Characteristics	Potential choices
Metric	Visits, Seconds, Days, Media
List of traces	
What is news?	Published by news media, published by any person/media
List of media	Tranco, Alexa, Cisco, Majestic
Top media	10, 20, 50, 100, 200, All
Information	Broad definition of news, only those identified as "political" news
Exposure	
Time threshold	1 second, 30 seconds, 120 seconds
Devices	PC only, Mobile only, **All**, All without apps
Tracking period	2, 5, 10, 15, 31 days

Using this, you can check whether the **literature**, or your **expertise**, favours **specific choices**

#1: Better defining what qualifies as valid information

- Before defining any measurement, list the different design decisions that you will have to make in order to operationalise the concept of interest
- List the potential choices that you could make within each design decision

Characteristics	Potential choices
Metric	Visits, Seconds, Days, Media
List of traces	
What is news?	Published by news media, published by any person/media
List of media	Tranco, Alexa, Cisco, Majestic
Top media	10, 20, 50, 100, 200, All
Information	Broad definition of news, only those identified as "political" news
Exposure	
Time threshold	1 second, 30 seconds, 120 seconds
Devices	PC only, Mobile only, **All**, All without apps
Tracking period	2, 5, 10, 15, 31 days

Using this, you can check whether the **literature**, or your **expertise**, favours **specific choices**

By listing the potential options, and your reasoning for a choice, now you can **report it and be transparent** about what lead you to create a specific measurement

#1: Better defining what qualifies as valid information

- Before defining any measurement, list the different design decisions that you will have to make in order to operationalise the concept of interest
- List the potential choices that you could make within each design decision

When you are defining many different concepts, you can create ad-hoc step-wise procedures for groups of concepts

WHAT CAN WE DO ABOUT SPECIFICATION ERRORS?

#2: Embrace uncertainty

#2: Embrace uncertainty

- Many times, it will not be clear what potential choice is better…which is normal!

Characteristics	Potential choices
Metric	Visits, Seconds, Days, Media
List of traces	
What is news?	Published by news media, published by any person/media
List of media	Tranco, Alexa, Cisco, Majestic
Top media	10, 20, 50, 100, 200, All
Information	Broad definition of news, only those identified as “political” news
Exposure	
Time threshold	1 second, 30 seconds, 120 seconds
Devices	PC only, Mobile only, All, All without apps
Tracking period	2, 5, 10, 15, 31 days

#2: Embrace uncertainty

- Many times, it will not be clear what potential choice is better…which is normal!

- We are dealing with Big Data. We are **not constrained to using one variable because we cannot ask the same question several times** in a questionnaire.

#2: Embrace uncertainty

web data opp

- Many times, it will not be clear what potential choice is better…which is normal!

- We are dealing with Big Data. We are **not constrained to using one variable because we cannot ask the same question several times** in a questionnaire.

- **We can create as many variables as we want, to:**
    1. **Conduct the analyses of interest** with all the potential variables
    2. **Test the quality** of all the potential variables

#2: Embrace uncertainty

- It is not a crazy idea, we have examples: **multiverse analysis**

Perspectives on Psychological Science
2016, Vol. 11(5) 702–712

DOI: 10.1177/1745691616658637
pps.sagepub.com
SAGE

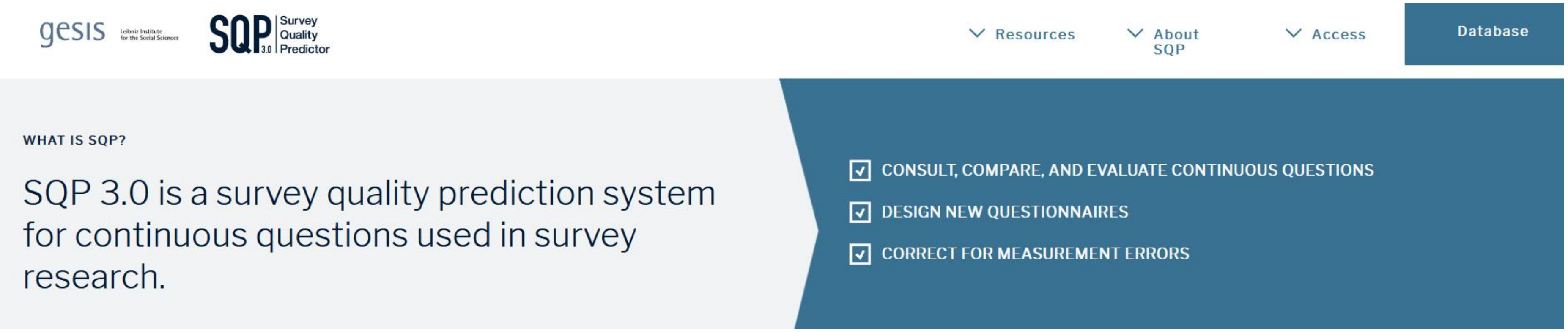
Increasing Transparency Through a Multiverse Analysis

Sara Steegen[1], Francis Tuerlinckx[1], Andrew Gelman[2], and Wolf Vanpaemel[1]
[1]KU Leuven, University of Leuven and [2]Columbia University

Abstract
Empirical research inevitably includes constructing a data set by processing raw data into a form ready for statistical analysis. Data processing often involves choices among several reasonable options for excluding, transforming, and coding data. We suggest that instead of performing only one analysis, researchers could perform a multiverse analysis, which involves performing all analyses across the whole set of alternatively processed data sets corresponding to a large set of reasonable scenarios. Using an example focusing on the effect of fertility on religiosity and political attitudes, we show that analyzing a single data set can be misleading and propose a multiverse analysis as an alternative practice. A multiverse analysis offers an idea of how much the conclusions change because of arbitrary choices in data construction and gives pointers as to which choices are most consequential in the fragility of the result.

#2: Embrace uncertainty

- It is not a crazy idea, we have examples: **Survey Quality Predictor (SQP)**



How can we embrace uncertainty? A practical example

How can we embrace uncertainty? A practical example

web data opp

- **The TRI-POL dataset**

- **Three wave survey** combined with **web tracking data** at the individual level (both PC and mobile data)

- Netquest metered panels
  - **Cross-quotas:** gender, age, education and region
  - **Sample size:** 1,289 (Spain)

- **Spain,** Portugal, Italy, Argentina and Chile

ELSEVIER

Data in Brief

Available online 9 May 2023, 109219

In Press, Journal Pre-proof What's this?

Data Article

The dynamics of political and affective polarisation: Datasets for Spain, Portugal, Italy, Argentina, and Chile (2019-2022)

Mariano Torcal [1], Emily Carty [2], Josep Maria Comellas [3], Oriol J. Bosch [4], Zoe Thomson [1], Danilo Serani [2]

GOBIERNO DE ESPAÑA | MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES | Fundación BBVA | THE UNIVERSITY of MISSISSIPPI | upf. Universitat Pompeu Fabra Barcelona | RECSM Research and Expertise Centre for Survey Methodology | Georgia State University | VANDERBILT UNIVERSITY

How can we embrace uncertainty? A practical example

Concept: The extent to which an individual encounters **written news media**

How can we embrace uncertainty? A practical example

Concept: The extent to which an individual encounters **written news media**

Characteristics	My choices
Metric	Visits, Seconds, Days, Media
List of traces	
List of media	Tranco, Alexa, Cisco, Majestic
Top media	10, 20, 50, 100, 200, All
Information	All URLs, only those identified as political
Exposure	
Time threshold	1 second, 30 seconds, 120 seconds
Devices	PC only, Mobile only, All, All without apps
Tracking period	2, 5, 10, 15, 31 days

+11k potential variables*

- I created **all** the potential variables
- Analyses are computed for each of the +11k variables
- This would take **years and innumerable resources** to be replicated for surveys!

** Not 100% fully crossed. The time metric is not crossed with the 30 seconds and 120 seconds thresholds.*

Assessing the validity of these measures, and their fluctuation

Assessing the validity of these measures, and their fluctuation

We can study the predictive validity of the variables

Assessing the validity of these measures, and their fluctuation

We can study the predictive validity of the variables

- Is the variable a good predictor of a theoretically related measure?

Assessing the validity of these measures, and their fluctuation

Gold standard: how well does media exposure predict political knowledge *gains**

* Dilliplane, S., Goldman, S. K., & Mutz, D. C. (2013). **Televised exposure to politics: New measures for a fragmented media environment**. *American Journal of Political Science*, *57*(1), 236-248.
Prior, M. (2009). **Improving media effects research through better measurement of news exposure.** *The Journal of Politics*, *71*(3), 893-908.

Assessing the validity of these measures, and their fluctuation

Gold standard: how well does media exposure predict political knowledge *gains**

- *Assumption:* exposure to news should impart political information

* Dilliplane, S., Goldman, S. K., & Mutz, D. C. (2013). **Televised exposure to politics: New measures for a fragmented media environment**. *American Journal of Political Science*, *57*(1), 236-248.
Prior, M. (2009). **Improving media effects research through better measurement of news exposure.** *The Journal of Politics*, *71*(3), 893-908.

Assessing the validity of these measures, and their fluctuation

Gold standard: how well does media exposure predict political knowledge *gains**

- *Assumption:* exposure to news should impart political information
- *Analytical approach:* fixed effect regression model of within person change of political knowledge across waves 1-3

* Dilliplane, S., Goldman, S. K., & Mutz, D. C. (2013). **Televised exposure to politics: New measures for a fragmented media environment**. *American Journal of Political Science*, *57*(1), 236-248.
Prior, M. (2009). **Improving media effects research through better measurement of news exposure.** *The Journal of Politics*, *71*(3), 893-908.

Assessing the validity of these measures, and their fluctuation

Gold standard: how well does media exposure predict political knowledge *gains**

- *Assumption:* exposure to news should impart political information

- *Analytical approach:* fixed effect regression model of within person change of political knowledge across waves 1-3

  - **Political knowledge: 1) 5 questions** about politics.

    2) **Basic knowledge** about the **political system**, and the **current cabinet**.

    3) **Sum of correct answers**, hence, it ranges from **0 to 5**

* Dilliplane, S., Goldman, S. K., & Mutz, D. C. (2013**). Televised exposure to politics: New measures for a fragmented media environment**. *American Journal of Political Science*, *57*(1), 236-248.
Prior, M. (2009). **Improving media effects research through better measurement of news exposure.** *The Journal of Politics*, *71*(3), 893-908.

Assessing the validity of these measures, and their fluctuation

web data opp

Gold standard: how well does media exposure predict political knowledge *gains**

- *Assumption:* exposure to news should impart political information
- *Analytical approach:* fixed effect regression model of within person change of political knowledge across waves 1-3

+11k unique coefficients

* Dilliplane, S., Goldman, S. K., & Mutz, D. C. (2013**). Televised exposure to politics: New measures for a fragmented media environment**. *American Journal of Political Science*, *57*(1), 236-248.
Prior, M. (2009). **Improving media effects research through better measurement of news exposure.** *The Journal of Politics*, *71*(3), 893-908.

Assessing the validity of these measures, and their fluctuation

Gold standard: how well does media exposure predict political knowledge *gains**

- *Assumption:* exposure to news should impart political information
- *Analytical approach:* fixed effect regression model of within person change of political knowledge across waves 1-3

+11k unique coefficients

What is the average?

Does it fluctuate?

* Dilliplane, S., Goldman, S. K., & Mutz, D. C. (2013). **Televised exposure to politics: New measures for a fragmented media environment**. *American Journal of Political Science*, *57*(1), 236-248.
Prior, M. (2009). **Improving media effects research through better measurement of news exposure.** *The Journal of Politics*, *71*(3), 893-908.

The impact of design choices on predictive validity

The impact of design choices on predictive validity

- After running the reliability and validity analyses, I created a new dataset, with the following:
  - **Name** of the variables
  - Associated **reliability coefficient** and **standardised reg. coefficient** (predictive validity)
  - **Design choices** of the specific variable, for each **design characteristic**

corr_ALL

Filter

	variable	Coefficient	List	Top	Metric	Time_threshold	Tracking_period	Domain_Subdomain	Device
1163	PRE10_D_News_MobilePC_webapp_10A_120s	-0.1954861	Alexa	10	Day	120	PRE10	Domain	All devices
913	PRE10_D_News_MobilePC_webapp_ALL_1s	-0.1944916	ALL	222	Day	1	PRE10	Domain	All devices
828	PRE10_D_News_MobilePC_webapp_100T_1s	-0.1942604	Tranco	100	Day	1	PRE10	Domain	All devices
868	PRE10_D_News_MobilePC_webapp_200T_1s	-0.1942604	Tranco	200	Day	1	PRE10	Domain	All devices
908	PRE10_D_News_MobilePC_webapp_50T_1s	-0.1932236	Tranco	50	Day	1	PRE10	Domain	All devices
813	PRE10_D_News_MobilePC_webapp_100A_1s	-0.1911152	Alexa	100	Day	1	PRE10	Domain	All devices
853	PRE10_D_News_MobilePC_webapp_200A_1s	-0.1911152	Alexa	200	Day	1	PRE10	Domain	All devices
893	PRE10_D_News_MobilePC_webapp_50A_1s	-0.1911152	Alexa	50	Day	1	PRE10	Domain	All devices
832	PRE15_D_News_MobilePC_webapp_10A_1s	-0.1880830	Alexa	10	Day	1	PRE15	Domain	All devices
827	PRE15_D_News_MobilePC_webapp_100T_1s	-0.1856270	Tranco	100	Day	1	PRE15	Domain	All devices
867	PRE15_D_News_MobilePC_webapp_200T_1s	-0.1856270	Tranco	200	Day	1	PRE15	Domain	All devices
912	PRE15_D_News_MobilePC_webapp_ALL_1s	-0.1841421	ALL	222	Day	1	PRE15	Domain	All devices

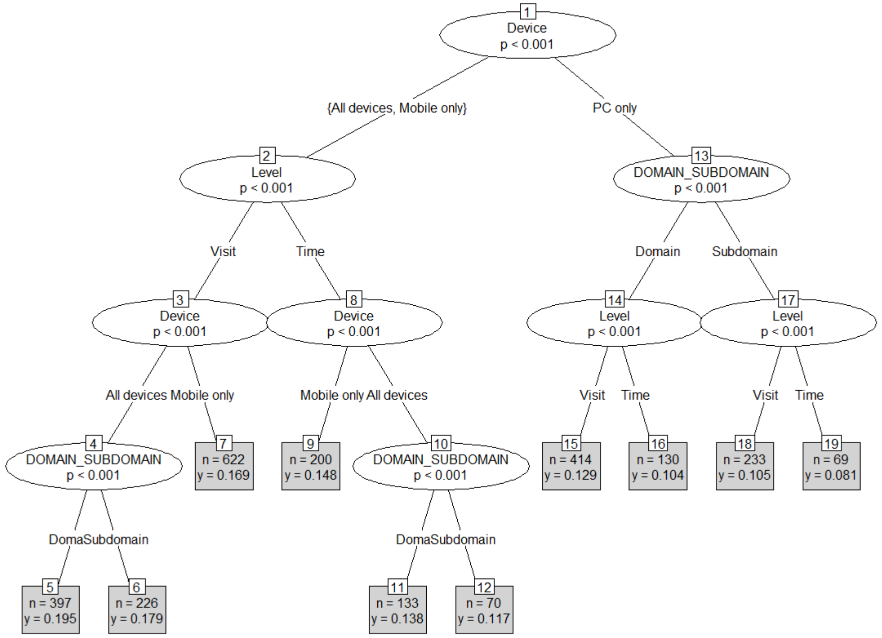
The impact of design choices on predictive validity

- After running the reliability and validity analyses, I created a new dataset, with the following:
  - **Name** of the variables
  - Associated **reliability coefficient** and **standardised reg. coefficient** (predictive validity)
  - **Design choices** of the specific variable, for each **design characteristic**

With this dataset it is possible to **model the effect** of each **design choice** on the estimated **reliability and (predictive) validity,** using the **+11k variables as observations**

The impact of design choices on predictive validity

- To predict the impact of each design choice, we can use a random forests of regression trees*

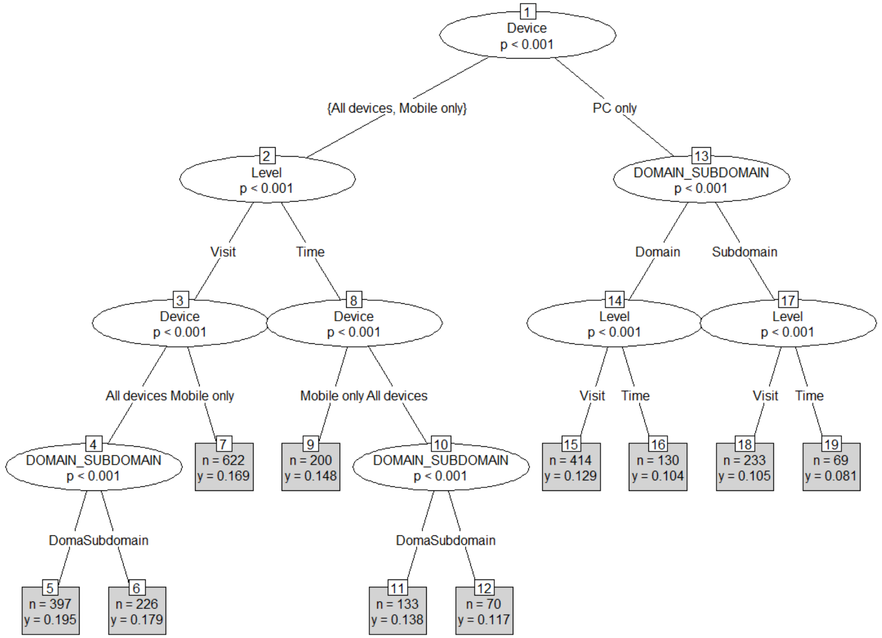


* R package *randomForest: Ntree*: 500 | *Mtry*: 4 | *Node size*: 3 | *Sample fraction*: 80%

The impact of design choices on predictive validity

- To predict the impact of each design choice, we can use a random forests of regression trees*

- We can extract the following information:
    - The variable importance: % increase of MSE
    - And the marginal effect of each choice



* R package *randomForest: Ntree*: 500 | *Mtry*: 4 | *Node size*: 3 | *Sample fraction*: 80%

Your turn to test some of this stuff!

- If you have not, download the files:
  - **reg_all_cross_spain.csv**
  - **multiverse_prediction_code.R**

- We will go together step by step.

- The goal is to show you how, after creating all the variables and running the analyses with all of them, we can make sense of the results.

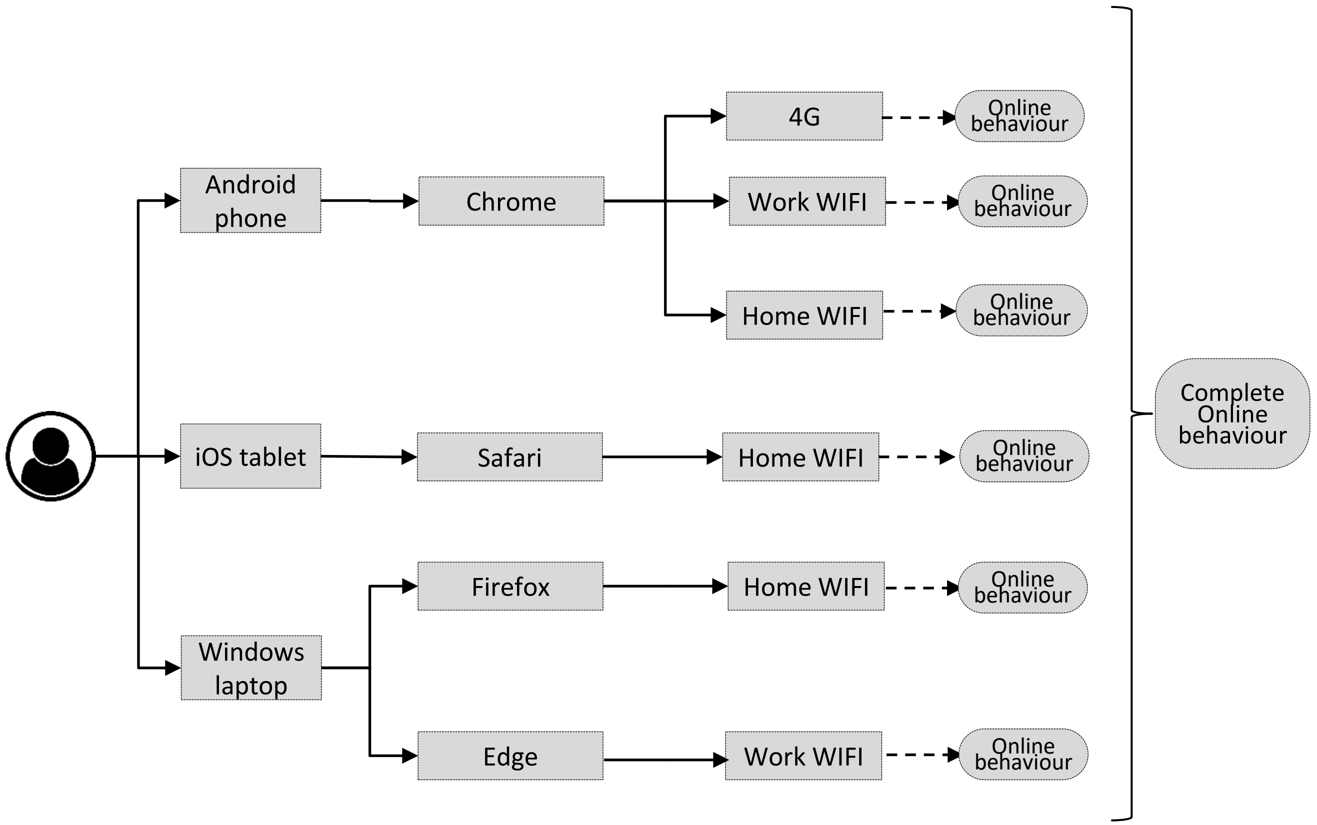
What can we do about these problems?

Strategies to prevent, identify, correct, and report ***measurement errors***

#1: Apply strategies to maximise the coverage of devices/browsers

WHAT CAN WE DO ABOUT MEASUREMENT ERRORS?

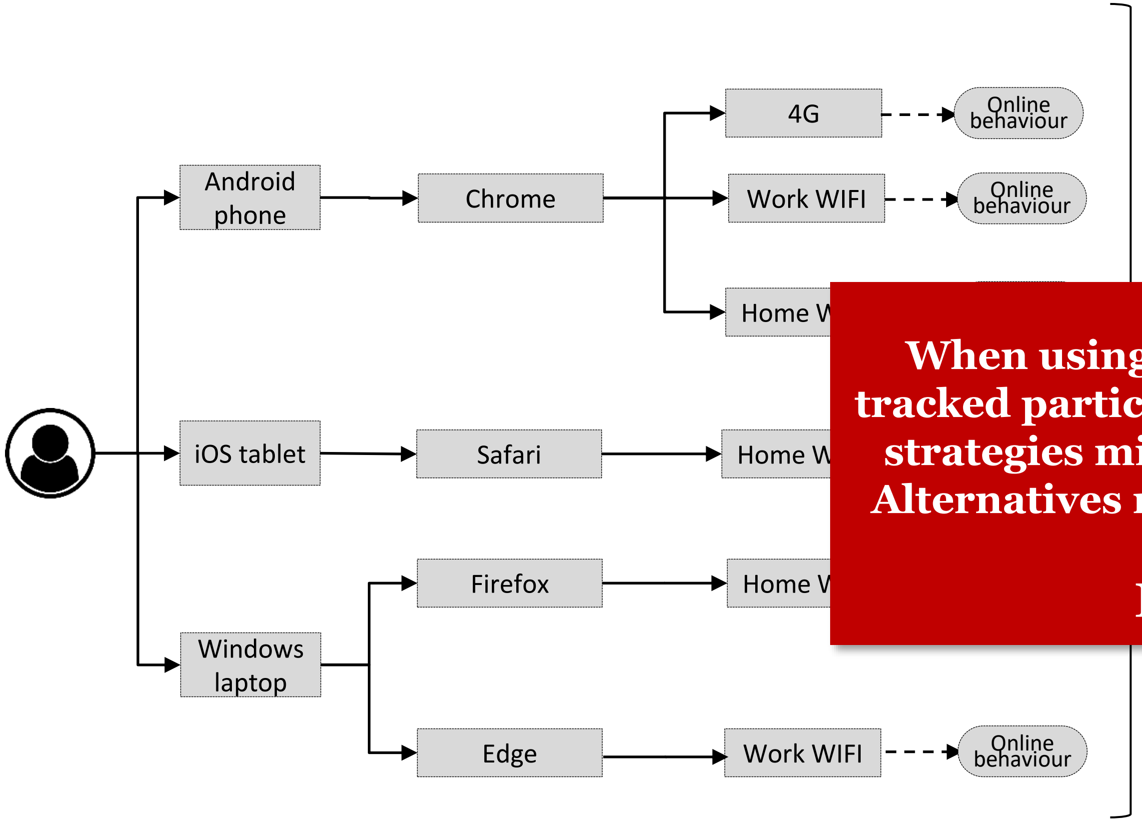
#1: Apply strategies to maximise the coverage of devices/browsers



1. Make sure that your **tracking toolkit can track most devices / browsers**
2. **Collect information** about the devices / browsers they use to go online **to tailor the invitation** to install technologies
3. **Simplify** as much as possible the installation / configuration process
4. **Incentivise** the installation of the technologies in more than one device

#1: Apply strategies to maximise the coverage of devices/browsers

web data opp



When using a panel of already tracked participants, some of these strategies might not be possible. Alternatives must be put in place.

Ideas?

1. Make sure that your **tracking toolkit can track most devices / browsers**

2. **Collect information** about the devices / browsers they use to go online **to** **ilor the invitation** to install chnologies

mplify as much as possible the stallation / configuration process

4. **Incentivise** the installation of the technologies in more than one device

#2: Identify and report undercoverage

#2: Identify and report undercoverage

An approach: combining survey and paradata

During the last 15 days, from how many of these different types of devices have you accessed the Internet (including using apps like Facebook, Twitter or YouTube)? Please, type the number of devices in the respective boxes.

Computer with Windows operating system: **[NUMERIC OPEN BOX]**

Apple computer(s) (MAC): **[NUMERIC OPEN BOX]**

Smartphone or tablet with Android operating system: **[NUMERIC OPEN BOX]**

Apple smartphone or tablet (iPhone or iPad): **[NUMERIC OPEN BOX]**

Others: **[NUMERIC OPEN BOX] (IF >0: "Please, specify: [OPEN TEXT BOX]")**

During the last 15 days, have you used any of the following web browsers to access the Internet through a computer with Windows operating system?

Internet Explorer	
Chrome	
Firefox	
Edge, Opera or others	

During the last 15 days, have you used any of the following web browsers to access the Internet through an Apple computer (MAC)?

	Yes
Internet Explorer	○
Safari	○
Chrome	○
Firefox	○
Edge, Opera or others	○

During the last 15 days, have you used any of the following web browsers to access the Internet through smartphone or tablet with Android operating system?

	Yes	No
Chrome	○	○
Samsung browser	○	○
Firefox	○	○
Edge, Opera or others	○	○

*Compare this information with device **paradata:** Information about **all** the devices and browsers in which they are tracked .*

#2: Identify and report undercoverage

An approach: combining survey and paradata

$$N^o\ of\ devices\ reported\ - N^o of\ devices\ tracked = N^o of\ uncovered\ devices$$

$$N^o of\ uncovered\ devices > 0 = Participant\ is\ undercovered$$

#2: Identify and report undercoverage

An approach: combining survey and paradata

$$N^o\ of\ devices\ reported\ - N^o of\ devices\ tracked = N^o of\ uncovered\ devices$$

$$N^o of\ uncovered\ devices > 0 = Participant\ is\ undercovered$$

We can identify who is undercovered, the extent of this undercoverage, and the type of undercoverage

#2: Identify and report undercoverage

An approach: combining survey and paradata

$$N^o\ of\ devices\ reported\ - N^o of\ devices\ tracked = N^o of\ uncovered\ devices$$

$$N^o of\ uncovered\ devices > 0 = Participant\ is\ undercovered$$

We can identify who is undercovered, the extent of this undercoverage, and the type of undercoverage

We MUST report this information!

#2: Identify and report undercoverage

An approach: combining survey and paradata

$N^o\ of\ de$

$N^o of\ un$

Table 6
Proportion of participants undercovered in terms of device, in all countries for waves 1 and 3.

	Italy		Portugal		Spain		Argentina		Chile	
Device	W1	W3	W1	W3	W1	W3	W1	W3	W1	W3
	76.1	76.7	76.5	75.3	70.3	66.9	70.0	67.9	73.7	72.7
N	842	688	818	675	992	844	1,127	848	958	693

Unweighted proportions.

We can identify who is undercovered, the extent of this undercoverage, and the type of undercoverage

We MUST report this information!

Torcal, M., Carty, E., Comellas, J. M., Bosch, O. J., Thomson, Z., & Serani, D. (2023). The dynamics of political and affective polarisation: Datasets for Spain, Portugal, Italy, Argentina, and Chile (2019-2022). *Data in Brief*, *48*, 109219.

#3: Simulate undercoverage bias

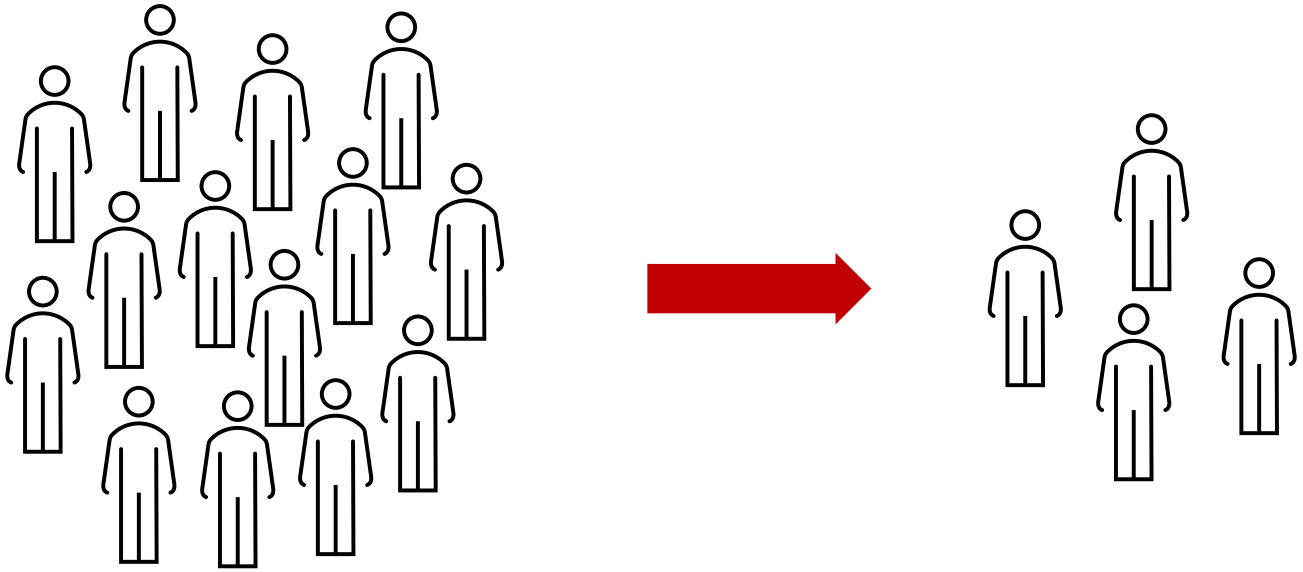
#3: Simulate undercoverage bias

Knowing who is fully covered allows also to simulate bias for them

#3: Simulate undercoverage bias

Knowing who is fully covered allows also to simulate bias for them

- We can treat those subsamples as our “population” of fully covered participants*



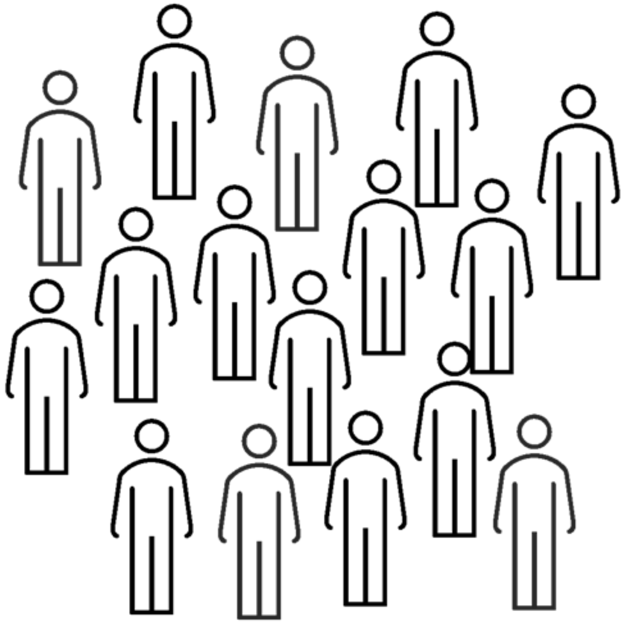
Full sample **Fully covered sample**

* Inverse probability weights computed using the random forest relative frequency method by Buskirk and Kolenikov (2015)

#3: Simulate undercoverage bias

Simulation approach

We can estimate the true estimates of this fully covered subsamples…



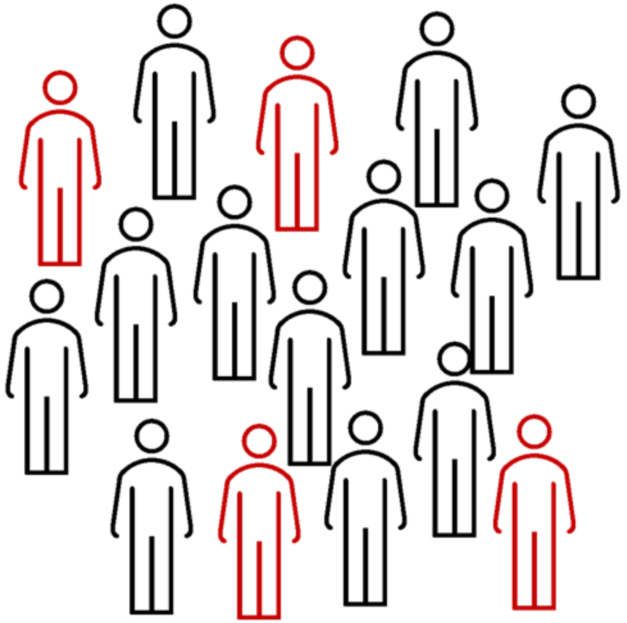
Under	Minutes mobile	Minutes PC	Total
Yes	20	4	24
No	10	6	16
Yes	5	14	19
Yes	26	9	35
No	3	32	35
Yes	14	3	17
No	17	6	23

Complete coverage → **True value:** 40 minutes

#3: Simulate undercoverage bias

Simulation approach

…to then simulate how their estimates would change if some of their information was lost



Under	Minutes mobile	Minutes PC	Total
Yes	0	4	4
No	10	6	16
Yes	0	14	14
Yes	0	9	9
No	3	32	35
Yes	0	3	3
No	17	6	23

Simulated undercoverage → **Biased value:** 18 minutes

→ Difference: 18 minutes = *bias*

#3: Simulate undercoverage bias

Simulating scenarios

- The **key is determining the probability** of being undercovered. But this probability can differ:

    1. We can modify the probability of missing any device / browser
    2. We can modify the probability of missing specific devices / browsers
    3. We can give independent and equal probabilities, or test more complex undercoverage scenarios

- **We can test how this undercoverage will affect the analyses of interest for our project**

#3: Simulate undercoverage bias

Simulating scenarios

- The **key is determi** his probability can differ:

1. We can modify
2. We can modify rs
3. We can give in plex undercoverage scenarios

Table 1. Scenarios for the simulations.

Scenario	*P*(PC undercoverage)	*P*(mobile undercoverage)
1	.25	.0
2	.50	.0
3	.75	.0
4	.0	.25
5	.0	.50
6	.0	.75
7	.25	.25
8	.25	.50
9	.25	.75
10	.50	.25
11	.50	.50
12	.75	.25
*13**	.33	.33

Note: *Scenario 13 represents the actual undercoverage in the sample

- **We can test how this undercoverage will affect the analyses of interest for our project**

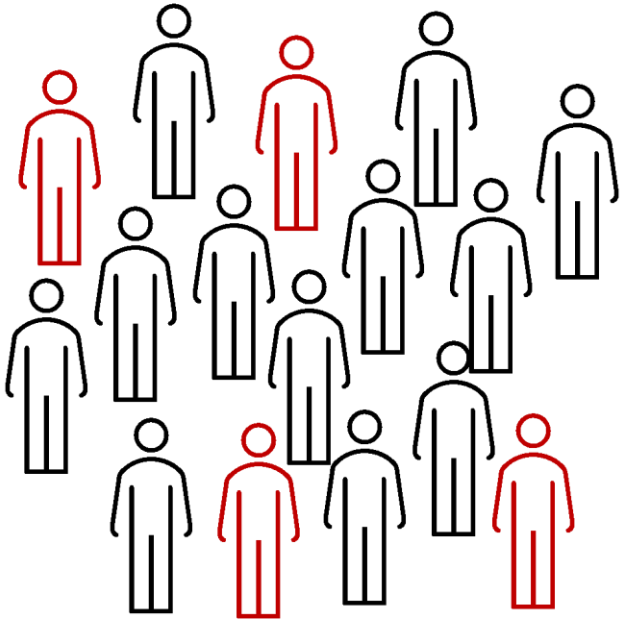
Bosch, O. J., Sturgis, P., Kuha, J., Revilla, M. (2023). Uncovering biases in digital trace data: an assessment of the prevalence and implications of tracking undercoverage when using web tracking data

#3: Simulate undercoverage bias

Monte Carlo simulations

For each scenario, we ran 1,000 random simulations.

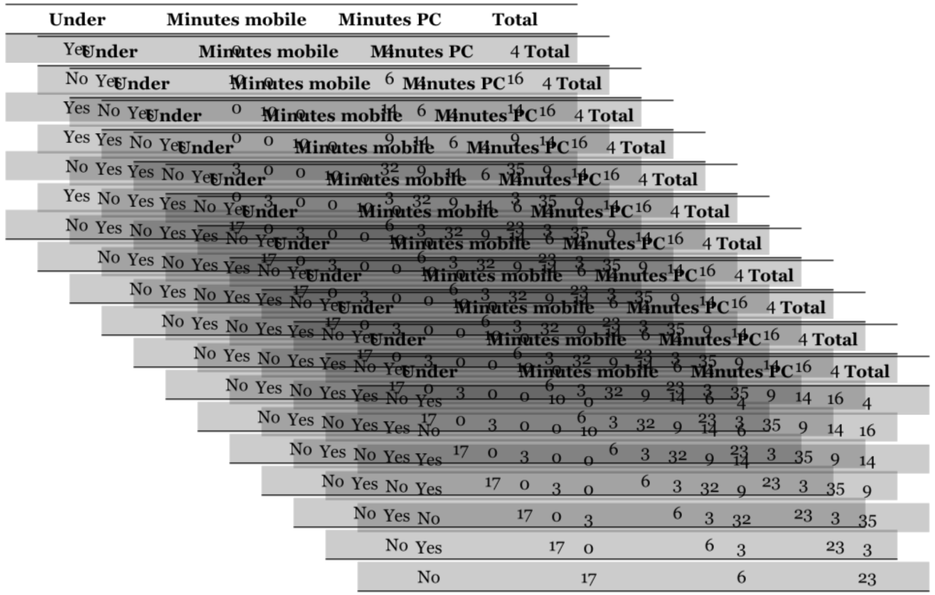
e.g., 25% with no ***computer*** covered → 0.25 probability of being undercovered



#3: Simulate undercoverage bias

Computing the bias

We then computed the *average estimate* of all 1,000 simulations.



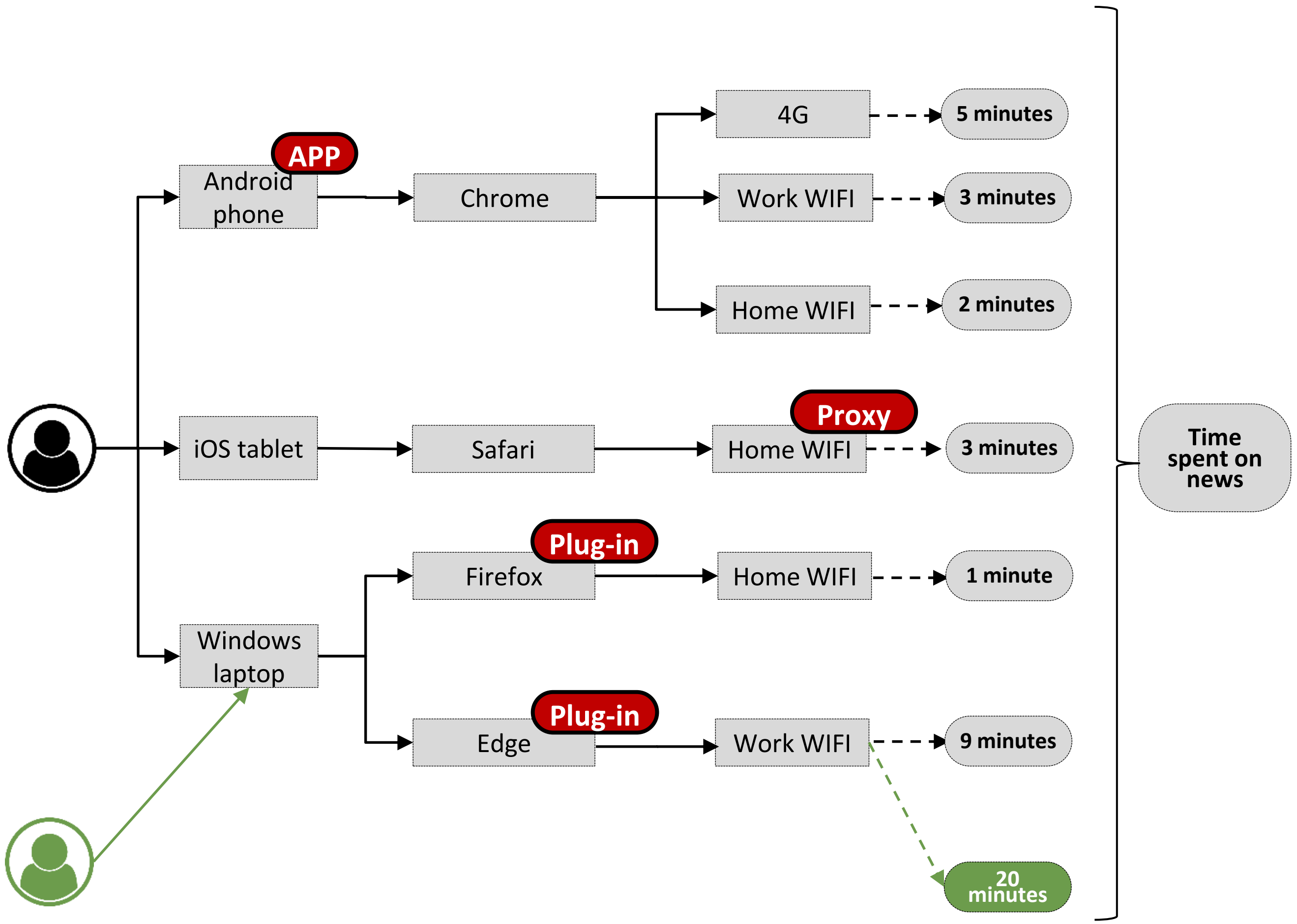
Avg. undercovered estimate: 22 minutes
True estimate: 40 minutes
Difference: 18 minutes → *bias*

Your turn to test some of this stuff!

- If you have not, download the files:
  - **simulation_dataset.csv**
  - **Simulation_code.R**
- We will go together step by step.
- The goal is to show you how, when you get information about whether people are undercovered, you can easily simulate the effect that this might have in your statistics of interest

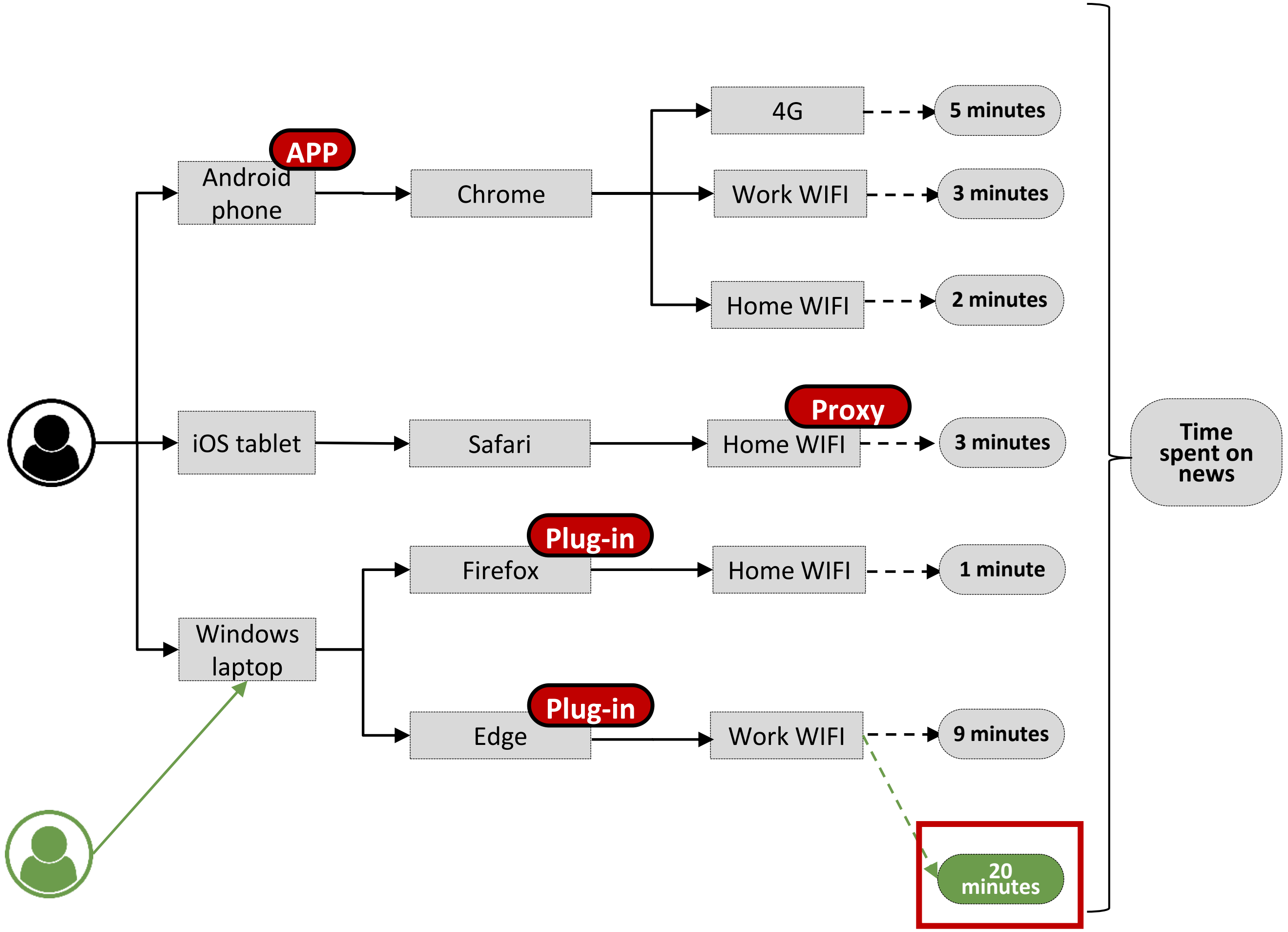
#4: Identify and report those sharing devices

#4: Identify and report those sharing devices



We need **to identify the devices that other people use**, and gather information on what they do there

#4: Identify and report those sharing devices



We need **to identify the devices that other people use**, and gather information on what they do there

We must try to **assess** how much of a problem this is, and potentially **account** for it

#4: Identify and report those sharing devices

An approach: combining survey and paradata

#4: Identify and report those sharing devices

An approach: combining survey and paradata

Use **paradata to identify the devices** that people are tracked on

Ask participant to **self-report whether other people use those devices**, and the extent of this use for the concepts that you want to measure

#4: Identify and report those sharing devices

An approach: combining survey and paradata

Use **paradata to identify the devices** that people are tracked on

Ask participant to **self-report whether other people use those devices**, and the extent of this use for the concepts that you want to measure

This information MUST be reported

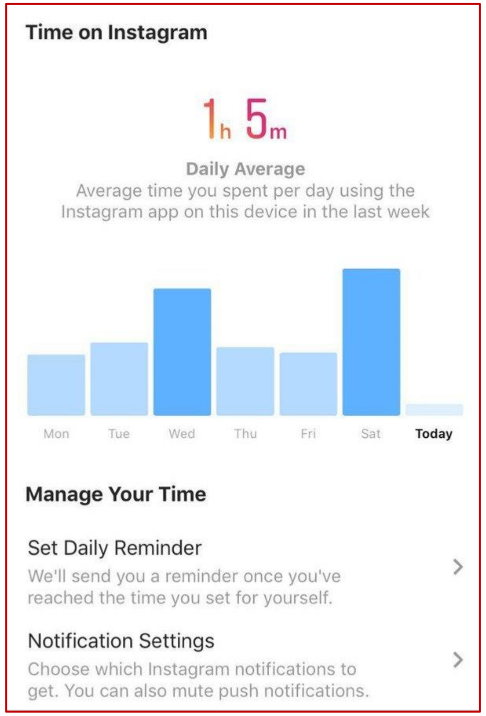
#5: Ideas on how to measure and account for the bias of shared devices?

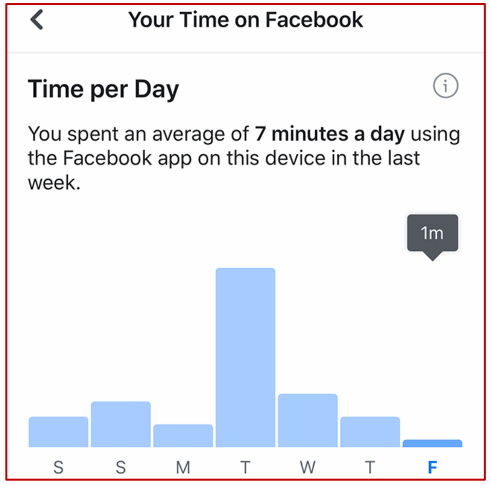
There is no clear established approach yet, currently working on it!

How to use data donations, and what to consider?

One step back: what to consider before collecting any data

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





One step back: what to consider before collecting any data

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

- In order to construct a measurement for 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.

One step back: what to consider before collecting any data

webdataopp

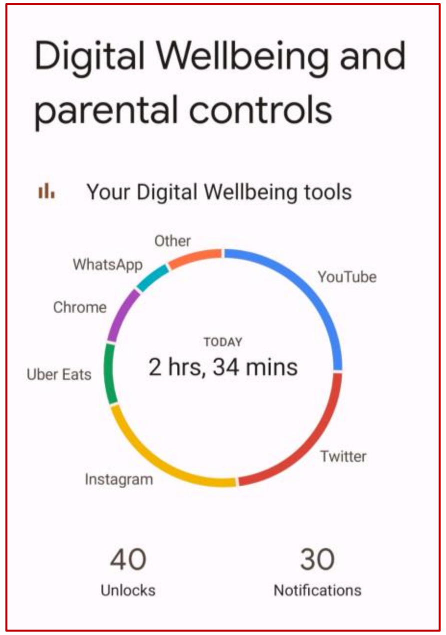
- With data donations, we are asking participants to **donate data that has already been produced by third-parties**, and that they have access to

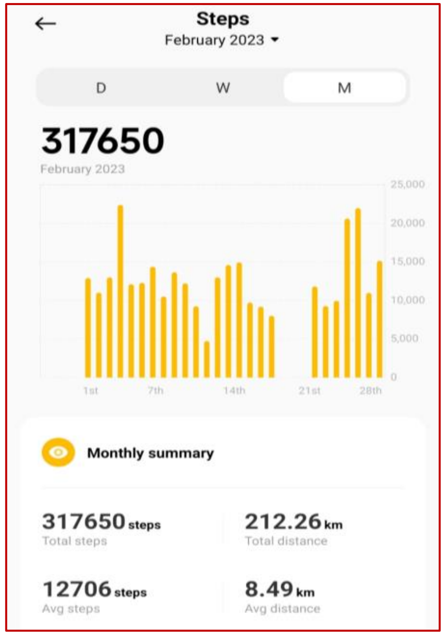
- In order to construct a measurement for a specific concept with data donations, we first need to identify whether there is any **available data** source that participants can access, capture, and share.

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.**

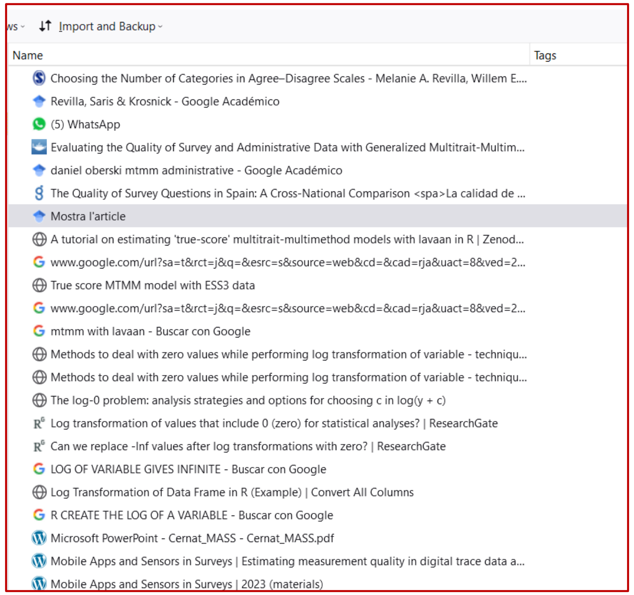


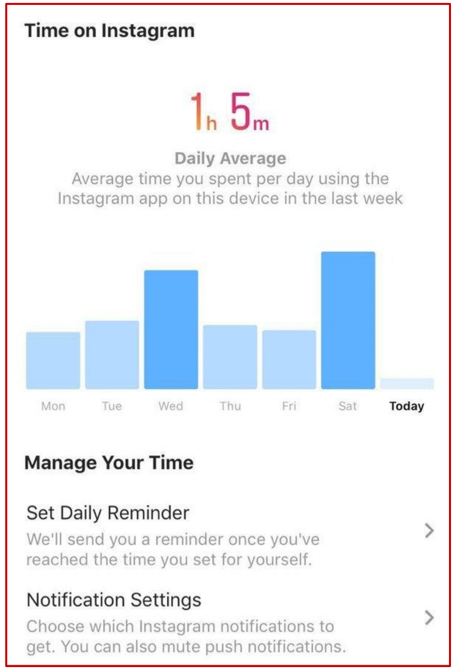


Examples of available data (related to digital behaviours)

- 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.**





The basics of data donations

- Let me be clear: there are many methods to collect data donations, not only one!

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.

The basics of data donations

- Let me be clear: there are many methods to collect data donations, not only one!

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

The basics of data donations

- Let me be clear: there are many methods to collect data donations, not only one!
- A data donation is any instance in which a person accesses some of their personal data, captures it, and shares it with researchers.
- This can be made in many ways! Hence, data donation projects will vary always in three dimensions:

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.

The basics of data donations

- Let me be clear: there are many methods to collect data donations, not only one!

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

- This can be made in many ways! Hence, data donation projects will vary always in three dimensions:
  - How participants **access** the traces of interest
  - *How they **capture** them*
  - *How they **share** the captured information with 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.

The basics of data donations

- Let me be clear: there are many methods to collect data donations, not only one!

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

- This can be made in many ways! Hence, data donation projects will vary always in three dimensions:
    - How participants **access** the traces of interest
    - *How they **capture** them*
    - *How they **share** the captured information with researchers*

Goal: make design decisions across these three dimensions that minimises the required effort of participants to share 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.

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 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.
  - The more the participant must do to capture and share the data, the less data we might be able to collect. But maybe the higher the sense of control?

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
  - The more the participant must do to capture and share the data, the less data we might be able to collect. But maybe the higher the sense of control?

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

Q & A

Thanks!

Oriol J. Bosch|PhD Candidate, The London School of Economics

o.bosch-jover@lse.ac.uk

orioljbosch

https://orioljbosch.com/

European Research Council
Established by the European Commission