



When survey science met online tracking:

Presenting an error framework for metered data

Oriol J. Bosch | THE LONDON SCHOOL OF ECONOMICS / RECSM-UPF

Melanie Revilla | RECSM-UPF

 o.bosch-jover@lse.ac.uk

 orioljbosch

 <https://orioljbosch.com/>



Tracking online behaviours using a meter

Definition

Metered data is obtained from a meter willingly installed or configured by a sample of participants on their devices (PCs, tablets and/or smartphones).

A ***meter*** refers to a heterogeneous group of tracking technologies that allow sharing with the researchers, at least, ***information about the URLs of the web pages visited by the participants***.

Tracking online behaviours using a meter

Definition

Metered data is obtained from a meter willingly installed or configured by a sample of participants on their devices (PCs, tablets and/or smartphones).

A ***meter*** refers to a heterogeneous group of tracking technologies that allow sharing with the researchers, at least, ***information about the URLs of the web pages visited by the participants***.

Sample of participants

Collected from a designed sample of individuals

Nonreactive

Collected by tracking the traces left by individuals when interacting with their devices online.

Tracking online behaviours using a meter

Benefits of metered data

- Objective and free of recall errors
- Continuously collected in real time
- Pre-designed sample of participants



Metered data in past research

33 papers identified, 26 since 2019

Metered data in past research

33 papers identified, 26 since 2019



The sources and correlates of exposure to vaccine-related (mis)information online ☆

Andrew M. Guess^{a,*}, Brendan Nyhan^b, Zachary O’Keeffe^c, Jason Reifler^d

^a Department of Politics, Princeton University, United States

^b Department of Government, Dartmouth College, United States

^c Department of Political Science, University of Michigan, United States

^d Department of Politics, University of Exeter, United Kingdom

ARTICLE INFO

Article history:
Received 11 June 2020
Received in revised form 1 October 2020
Accepted 7 October 2020
Available online 22 October 2020

Keywords:
Vaccine hesitancy
Vaccine skepticism
Online
Information
Social media
Search

ABSTRACT

Objectives: To assess the quantity and type of vaccine-related information Americans consume online and its relationship to social media use and attitudes toward vaccines.
Methods: Analysis of individual-level web browsing data linked with survey responses from representative samples of Americans collected between October 2016 and February 2019.
Results: We estimate that approximately 84% of Americans visit a vaccine-related webpage each year. Encounters with vaccine-skeptical content are less frequent; they make up only 7.5% of vaccine-related pageviews and are encountered by only 18.5% of people annually. However, these pages are more likely to be published by untrustworthy sources. Moreover, skeptical content exposure is more common among people with less favorable vaccine attitudes. Finally, usage of online intermediaries is frequently linked to vaccine-related information exposure. Google use is differentially associated with subsequent exposure to non-skeptical content, whereas exposure to vaccine-skeptical webpages is associated with usage of webmail and, to a lesser extent, Facebook.
Conclusions: Online exposure to vaccine-skeptical content is relatively rare, but vigilance is required given the potential for exposure among vulnerable audiences.

© 2020 Elsevier Ltd. All rights reserved.

Published in final edited form as:

Nat Hum Behav. 2020 March 02; 4(5): 472–480. doi:10.1038/s41562-020-0833-x.

Exposure to untrustworthy websites in the 2016 U.S. election

Andrew M. Guess¹, Brendan Nyhan^{2,*}, Jason Reifler³

¹Department of Politics and Woodrow Wilson School, Princeton University, Princeton, NJ, USA

²Department of Government, Dartmouth College, Hanover, NH, USA

³Department of Politics, University of Exeter, Exeter, UK

Abstract

Though commentators frequently warn about “echo chambers,” little is known about the volume or slant of political misinformation people consume online, the effects of social media and fact-checking on exposure, or its effects on behavior. We evaluate these questions for the websites publishing factually dubious content often described as “fake news.” Survey and web traffic data from the 2016 U.S. presidential campaign show that Trump supporters were most likely to visit these websites, which often spread via Facebook. However, these sites 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 widespread speculation about the prevalence of exposure to untrustworthy websites has been overstated.

Predicting Voting Behavior Using Digital Trace Data

Ruben L. Bach¹, Christoph Kern¹, Ashley Amaya², Florian Keusch¹, Frauke Kreuter^{1,3,4}, Jan Hecht⁵, and Jonathan Heinemann⁶

Social Science Computer Review
1-22

© The Author(s) 2019



Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0894439319882896
journals.sagepub.com/home/ssc



International Journal of Public Opinion Research Vol. 31 No. 4 2019
© The Author(s) 2018. Published by Oxford University Press on behalf of The World Association for Public Opinion Research. All rights reserved.
doi:10.1093/ijpor/edy025 Advance Access publication 15 December 2018

Is Facebook Eroding the Public Agenda? Evidence From Survey and Web-Tracking Data

Ana S. Cardenal¹, Carol Galais², and Silvia Majó-Vázquez³

¹School of Law and Political Science, Universitat Oberta de Catalunya, Spain;
²Political Science and Public Law Department, Universitat Autònoma de Barcelona, Spain;
³Reuters Institute for the Study of Journalism, University of Oxford, UK

Abstract

A major concern arising from ubiquitous tracking of individuals' online activity is that algorithms may be trained to predict personal sensitive information. Although previous research on sociodemographic characteristics, little sensitive outcomes. Against this background, we predict voting behavior, which is considered to strict privacy regulations. Using records of online users eligible to vote in the 2017 German federal election, we find that online activity predicts voting behavior in the same individuals. These findings add to the debate on the information flows.

¹Department of Political Science, Princeton University, Princeton, NJ 08544;
²Department of Political Science, Princeton University, Princeton, NJ 08544;
³Department of Political Science, Princeton University, Princeton, NJ 08544

ARTICLE INFO

Article history:
Received 11 June 2020
Received in revised form 1 October 2020
Accepted 7 October 2020
Available online 22 October 2020

Keywords:
Vaccine hesitancy
Vaccine skepticism
Online information
Social media
Search

ABSTRACT

Objective: To investigate the effect of online activity on voting behavior.
Method: We use a randomized controlled experiment to measure the effect of online activity on voting behavior.
Results: We find that online activity predicts voting behavior in the same individuals.
Conclusion: These findings add to the debate on the information flows.

The consequences of online partisan media

Andrew M. Guess^{a,b,1,2}, Pablo Barberá^{c,1}, Simon Munzert^{d,1}, and JungHwan Yang (양정환)^{e,1}

^aDepartment of Politics, Princeton University, Princeton, NJ 08544; ^bSchool of Public and International Affairs, Princeton University, Princeton, NJ 08544;
^cDepartment of Political Science and International Relations, University of Southern California, Los Angeles, CA 90089; ^dData Science Lab, Hertie School, 10117 Berlin, Germany; and ^eDepartment of Communication, University of Illinois at Urbana-Champaign, Urbana, IL 61801

Edited by Christopher Andrew Bail, Duke University, Durham, NC, and accepted by Editorial Board Member Margaret Levi February 17, 2021 (received for review June 29, 2020)

What role do ideologically extreme media play in the polarization of society? Here we report results from a randomized longitudinal field experiment embedded in a nationally representative online panel survey ($N = 1,037$) in which participants were incentivized to change their browser default settings and social media following patterns, boosting the likelihood of encountering news with either a left-leaning (HuffPost) or right-leaning (Fox News) slant during the 2018 US midterm election campaign. Data on ≈ 19 million web visits by respondents indicate that resulting changes in news consumption persisted for at least 8 wk. Greater exposure to partisan news can cause immediate but short-lived increases in website visits and knowledge of recent events. After adjusting for multiple comparisons, however, we find little evidence of a direct impact on opinions or affect. Still, results from later survey waves suggest that both treatments produce a lasting and meaningful decrease in trust in the mainstream media up to 1 y later. Consistent with the minimal-effects tradition, direct consequences of online partisan media are limited, although our findings raise questions about the possibility of subtle, cumulative dynamics. The combination of experimentation and computational social science techniques illustrates a powerful approach for studying the long-term consequences of exposure to partisan news.

argues that media primarily reinforce existing predispositions (16). At the same time, more recent research strongly implies that newspapers and especially cable news can change people's voting behavior, especially those without strong partisan attachments (17–20). We propose an internet-age synthesis that views people's information environments through the lens of choice architecture (21): frictions, subtle design features, and default settings that structure people's online experience. In this view, small changes (or nudges) could disproportionately affect information consumption habits that have downstream consequences.

To that end, we designed a large, longitudinal online field experiment that subtly but naturalistically increased people's exposure to partisan news websites. Our choice of treatment is ecologically valid: Despite the importance of social media for agenda-setting (22) and public expression (23), more Americans continue to say that they get news from news websites or apps than social media sites (24). The intervention thus served as a nudge, boosting the likelihood that subjects encountered news framed with a partisan slant during their day-to-day web browsing experience, even if inadvertently. The powerful, sustained nature of the intervention and our ability to track participants with survey and behavioral data for months provided the opportunity to test a range of hypotheses about the long-term impact

Abstract

ion, minimizing for fragmenting their effect on the rough Facebook important problems combines survey assumption influence a relevant news itative sample of s for the public

the volume ia and fact-websites traffic data ely to visit all share of ericans with ead rsted.

How Much Time Do You Spend Online? Understanding and Improving the Accuracy of Self-Reported Measures of Internet Use

Theo Araujo, Anke Wonneberger, Peter Neijens, and Claes de Vreese

Amsterdam School of Communication Research (ASCoR), University of Amsterdam, Amsterdam, The Netherlands

ABSTRACT

Given the importance of survey measures of online media use for communication research, it is crucial to assess and improve their quality, in particular because the increasingly fragmented and ubiquitous usage of internet complicates the accuracy of self-reports. This article discusses the need to the discussion regarding the importance of self-reports in presenting relevant factors for improving the accuracy of self-reports. Testing survey design strategies for improving the accuracy of self-reports tracking data and survey data have found low levels of accuracy. This article reveals biases due to a range of factors, including (actual) internet usage, proper usage of mobile devices. An effort to reduce inaccuracies of reported internet usage in research practice follow from

COMMUNICATION METHODS AND MEASURES
2016, VOL. 10, NO. 1, 13–27
<http://dx.doi.org/10.1080/19312458.2015.1118446>

The Accuracy of Self-Reported Internet Use—A Validation Study Using Client Log Data

Michael Scharkow

University of Hohenheim

ABSTRACT

The vast majority of empirical research on online communication, or media use in general, relies on self-report measures instead of behavioral data. Previous research has shown that the accuracy of these self-report measures can be quite low, and both over- and underreporting of media use are commonplace. This study compares self-reports of Internet use with client log files from a large household sample. Results show that the accuracy of self-reported frequency and duration of Internet use is quite low, and that survey data are only moderately correlated with log file data. Moreover, there are systematic patterns of misreporting, especially overreporting, rather than random deviations from the log files. Self-reports for specific content such as social network sites or video platforms seem to be more accurate and less consistently biased than self-reports of generic frequency or duration of Internet use. The article closes by demonstrating the consequences of biased self-reports and discussing possible solutions to the problem.

Information flows

^aDepartment of Political Science
^bDepartment of Political Science
^cDepartment of Political Science, University of Amsterdam

ARTICLE INFO

Article history:
Received 11 June 2020
Received in revised form 1 October 2020
Accepted 7 October 2020
Available online 22 October 2020

Keywords:
Vaccine hesitancy
Vaccine skepticism
Online
Information
Social media
Search

Article

Two Half-Truths Make a Whole? On Bias in Self-Reports and Tracking Data

Pascal Jürgens¹, Birgit Stark¹, and Melanie Magin²

 **Routledge**
Taylor & Francis Group

Social Science Computer Review
2020, Vol. 38(5) 600–615
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0894439319831643
journals.sagepub.com/home/ssc



patterns of media usage, but also to, for example, understand the need to precisely attribute data to show that survey-overreporting, little effort has been made. Using data from a multitude of systematic distortions in data along with potential solutions,

known about the volume of social media and fact-checking for the websites surveyed and web traffic data were most likely to visit made up a small share of the subset of Americans with that widespread bias has been overstated.

Inferences for finite populations

Metered data can potentially suffer from different types of errors

Shared devices and observation of only part of the activity

- 60% of desktops, 40% of laptops and tablets, and 9% of smartphones shared to some degree (Revilla et al., 2017)
- 28% with the meter installed in all devices (Pew Research Center, 2020)

Technical issues and reactivity / social desirability bias (Jurgens et al., 2020; Toth and Trifonova, 2020)

Substantive conclusions vary depending on what is considered as a visit (3 seconds / 30 seconds / 120 seconds) (Mangold et al., 2021)

Inferences for finite populations

Metered data can potentially suffer from different types of errors

Shared devices and observation of only part of the activity

- 60% of desktops, 40% of laptops and tablets, and 9% of smartphones shared to some degree (Revilla et al., 2017)
- 28% with the meter installed in all devices (Pew Research Center, 2020)

Technical issues and reactivity / social desirability bias (Jurgens et al., 2020; Toth and Trifonova, 2020)

Substantive conclusions vary depending on what is considered as a visit (3 seconds / 30 seconds / 120 seconds) (Mangold et al., 2021)

Inferences for finite populations

Metered data can potentially suffer from different types of errors

Shared devices and observation of only part of the activity

- 60% of desktops, 40% of smartphones are shared to some degree (Revilla et al., 2020)

- 28% with the meter installed (Revilla et al., 2020)

Technical issues and reactivity (Revilla et al., 2020; Toth and Trifonova, 2020)

Substantive conclusions vary depending on what is considered as a visit (3 seconds / 30 seconds / 120 seconds) (Mangold et al., 2021)

A systematic **categorization** and
conceptualization of metered data errors

Not available

Main goals and contribution

Total Error Framework for metered data

- #1 **Summarize** the data collection and analysis process for metered data.
- #2 **Conceptualize and categorize** all errors components (e.g. measurement errors) and causes (e.g. social desirability) that can occur when using metered data.

Main goals and contribution

Total Error Framework for metered data

- #1 **Summarize** the data collection and analysis process for metered data.
- #2 **Conceptualize and categorize** all errors components (e.g. measurement errors) and causes (e.g. social desirability) that can occur when using metered data.

- 1) Choose the best design **options** for metered data.
- 2) Make **better** informed decisions **while planning** when and how to supplement **or** replace **survey data** with metered data.
- 3) **Help assess** research using metered data.

Main goals and contribution

Total Error Framework for metered data

- #1 **Summarize** the data collection and analysis process for metered data.
- #2 **Conceptualize and categorize** all errors components (e.g. measurement errors) and causes (e.g. social desirability) that can occur when using metered data.

Bosch, O.J., and M. Revilla (2021). “**When survey science met online tracking: presenting an error framework for metered data.**” RECSM Working Papers Series, 62

Adapting instead of reinventing

- Follow approach by Amaya et al (2020) with their **Total Error Framework for Big Data**
- 7 error components of the TSE (Groves et al., 2009) as starting point:
 - Coverage errors, sampling errors, *missing data errors*, adjustment errors, *specification errors*, measurement errors and processing errors

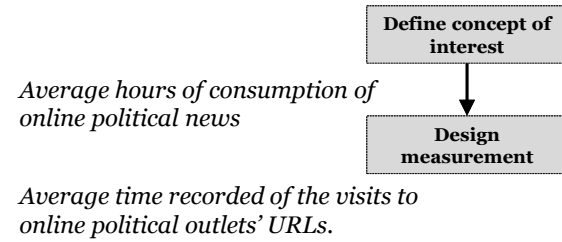
Data collection and analysis process

Data collection and analysis process

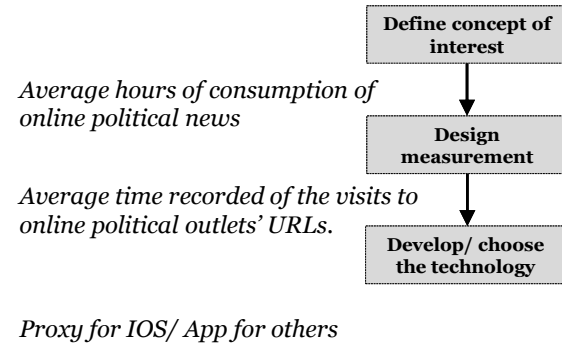
Define concept of
interest

*Average hours of consumption of
online political news*

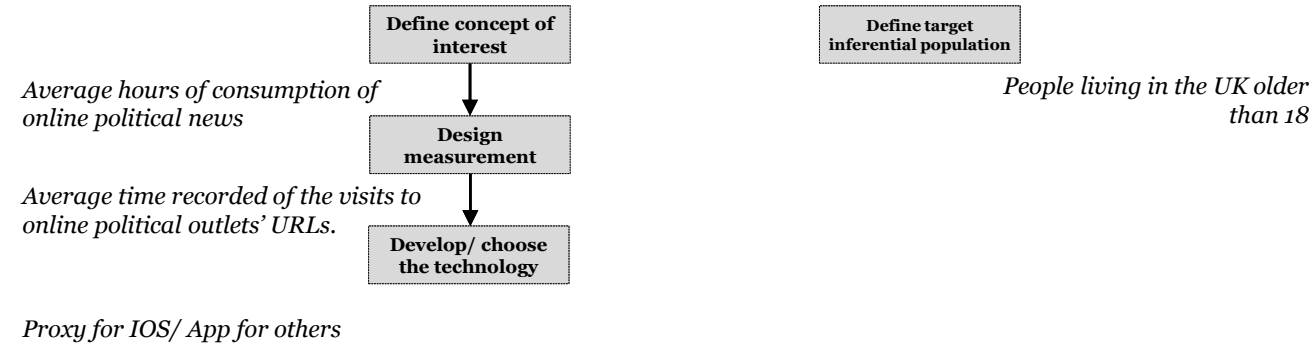
Data collection and analysis process



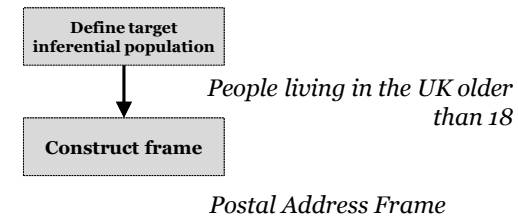
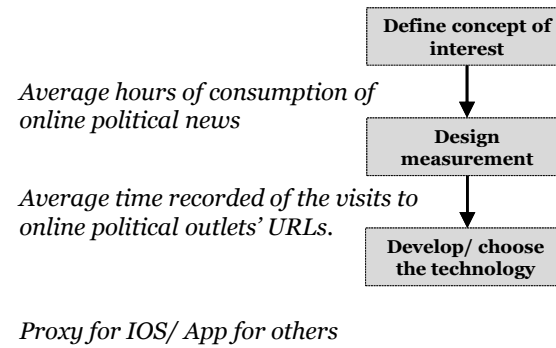
Data collection and analysis process



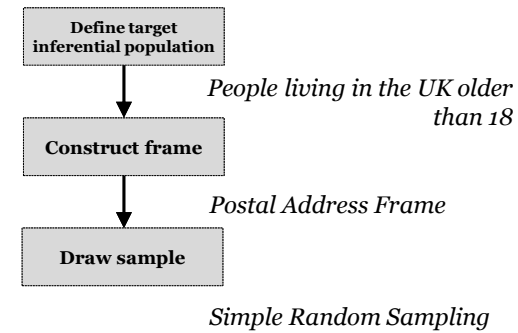
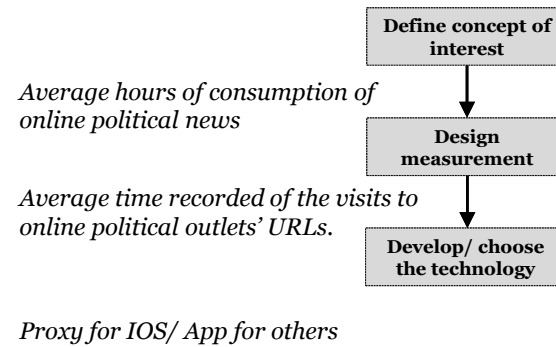
Data collection and analysis process



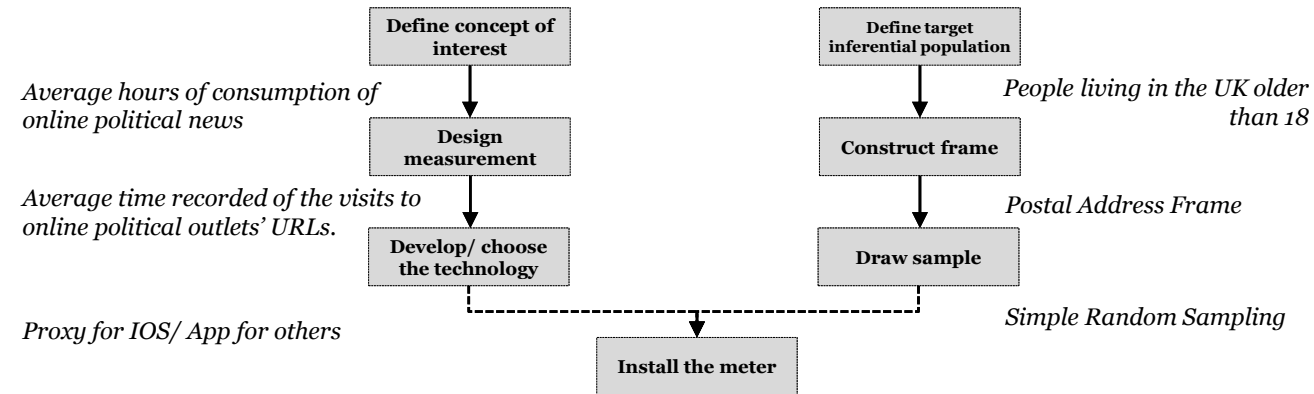
Data collection and analysis process



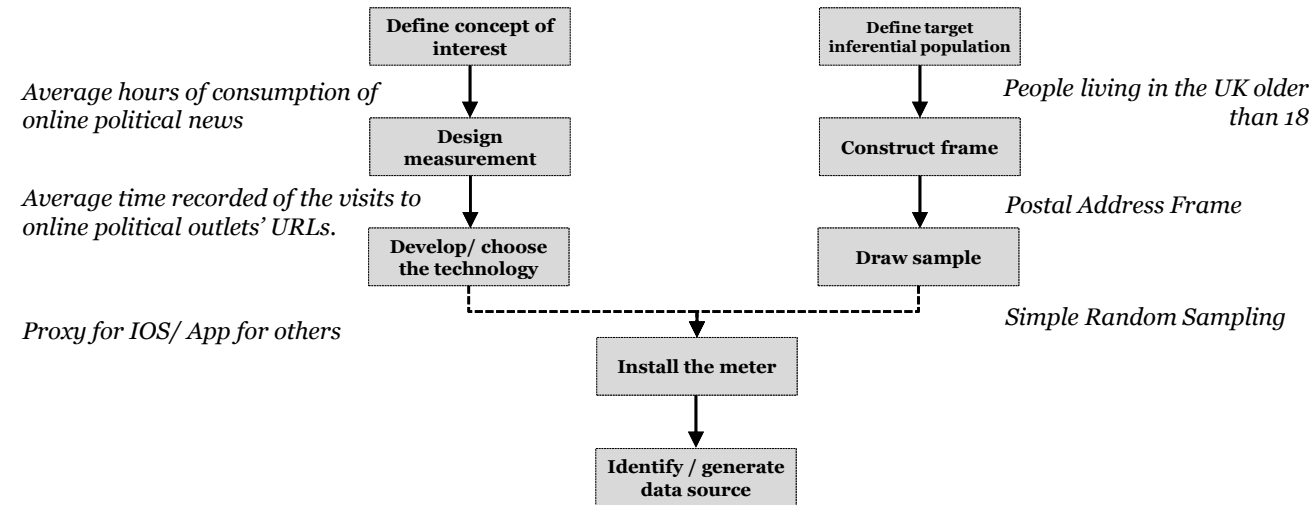
Data collection and analysis process



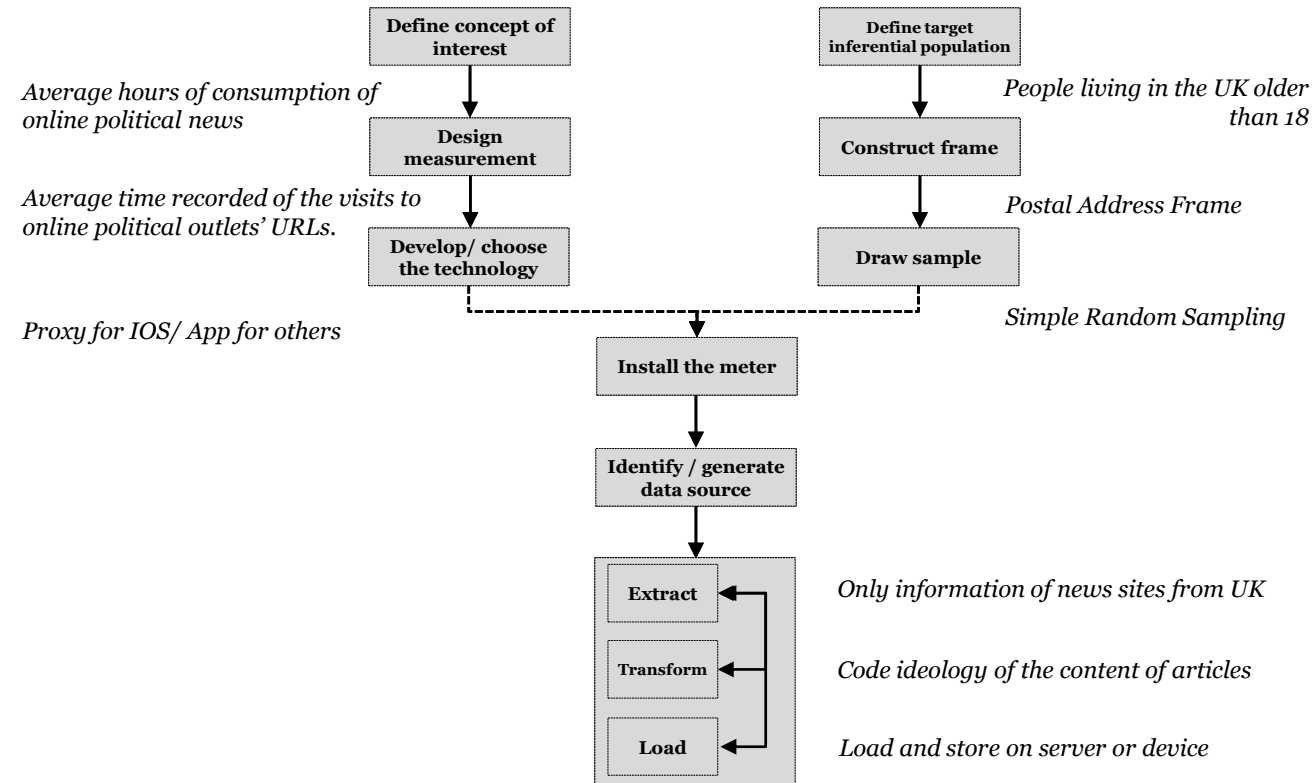
Data collection and analysis process



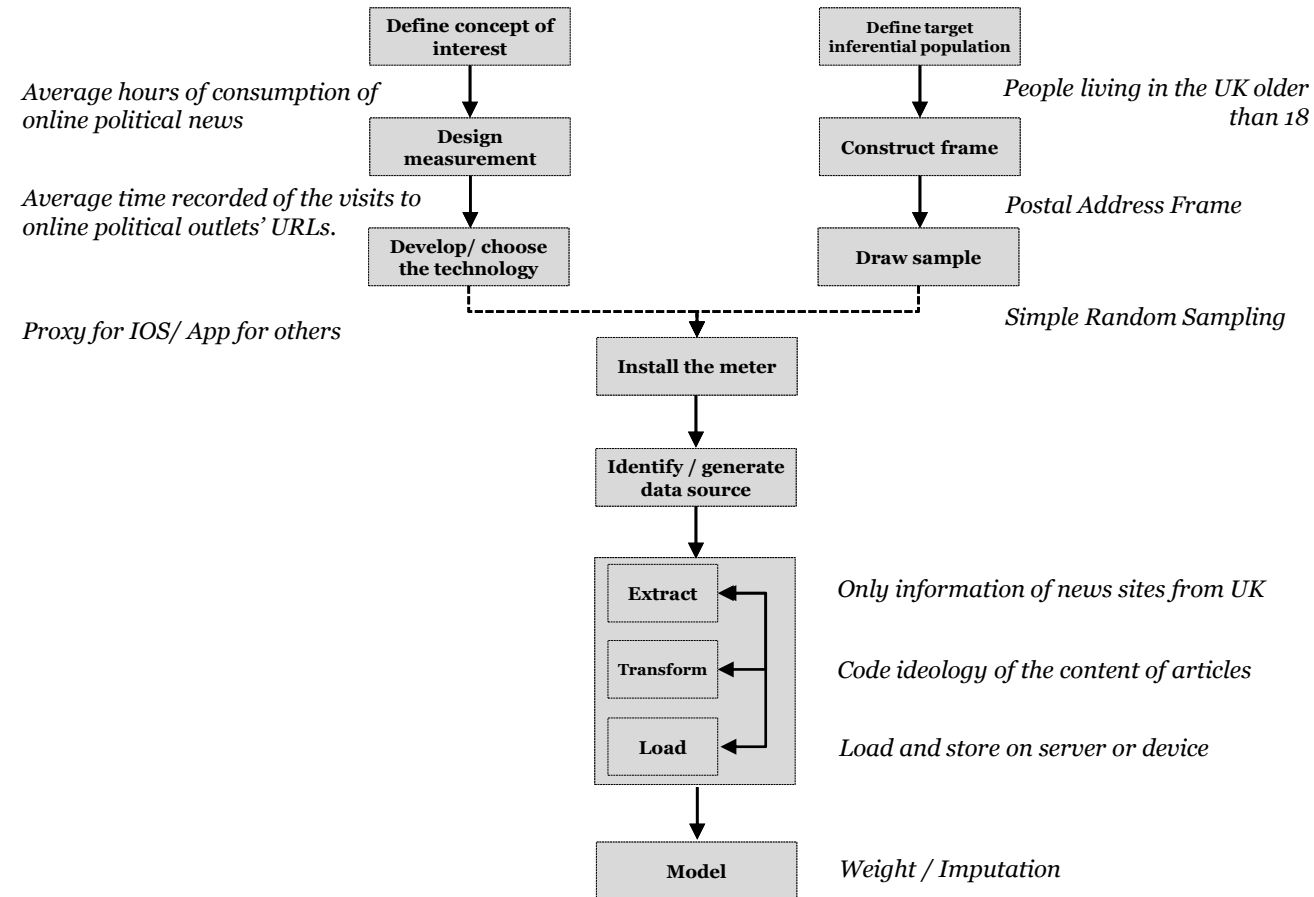
Data collection and analysis process



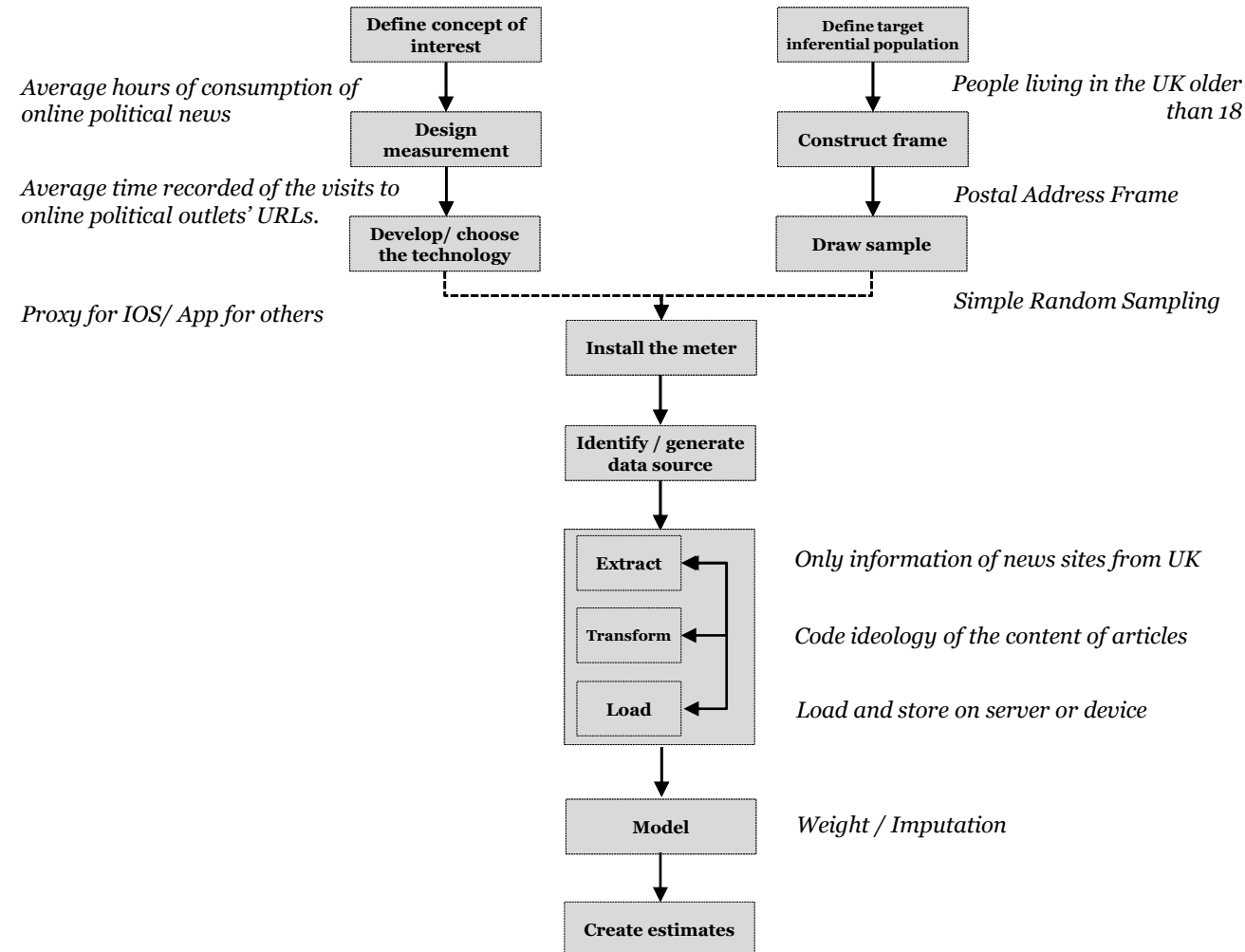
Data collection and analysis process



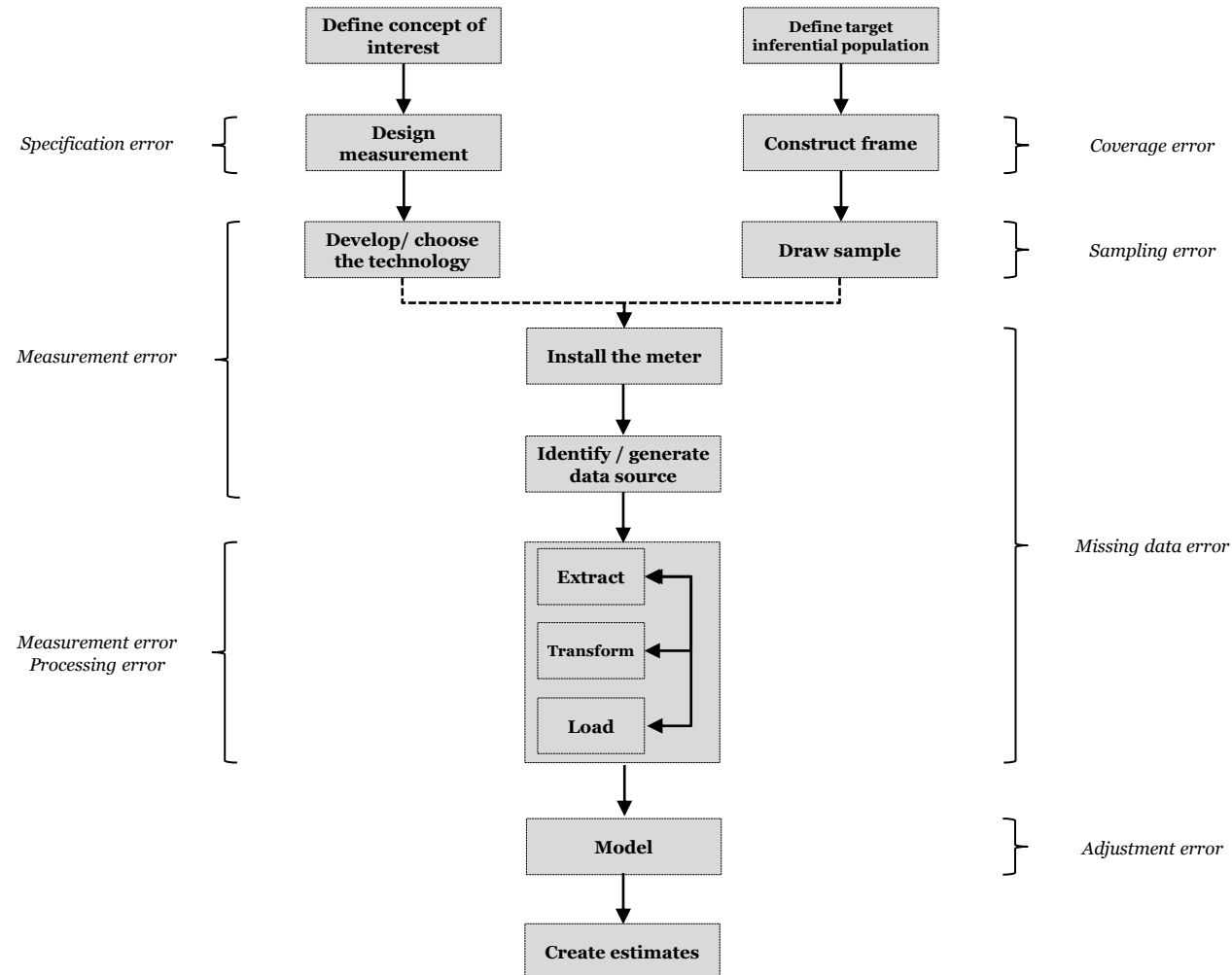
Data collection and analysis process



Data collection and analysis process



Data collection and analysis process



Error components and their causes

Error components	Specific error causes
Specification error	<ul style="list-style-type: none"> – Measuring concepts from which not enough data is available – Inferring attitudes – Defining valid information
Measurement error	<ul style="list-style-type: none"> – Non-trackable target – Meter not installed – Uninstalling the meter – New non-tracked device – Technology limitations – Technology errors – Hidden behaviours – Shared device – Social desirability – Extraction error
Processing error	<ul style="list-style-type: none"> – Coding error – Aggregation at the domain level – Data anonymization
Coverage error	<ul style="list-style-type: none"> – Non-trackable individuals
Sampling error	<ul style="list-style-type: none"> – Same error causes than for surveys
Missing data error	<ul style="list-style-type: none"> – Noncontact – Non-consent – Non-trackable target – Meter not installed – Uninstalling the meter – New non-tracked device – Technology limitations – Technology error – Hidden behaviour – Social desirability – Extraction error
Adjustment error	<ul style="list-style-type: none"> – Same error causes than for surveys

Error components and their causes

Error components	Specific error causes
Specification error	<ul style="list-style-type: none"> – Measuring concepts from which not enough data is available – Inferring attitudes – Defining valid information
Measurement error	<ul style="list-style-type: none"> – Non-trackable target – Meter not installed – Uninstalling the meter – New non-tracked device – Technology limitations – Technology errors – Hidden behaviours – Shared device – Social desirability – Extraction error
Processing error	<ul style="list-style-type: none"> – Coding error – Aggregation at the domain level – Data anonymization
Coverage error	<ul style="list-style-type: none"> – Non-trackable individuals
Sampling error	<ul style="list-style-type: none"> – Same error causes than for surveys
Missing data error	<ul style="list-style-type: none"> – Noncontact – Non-consent – Non-trackable target – Meter not installed – Uninstalling the meter – New non-tracked device – Technology limitations – Technology error – Hidden behaviour – Social desirability – Extraction error
Adjustment error	<ul style="list-style-type: none"> – Same error causes than for surveys

Most specific error causes on the side of measurement

Error components and their causes

Error components	Specific error causes
Specification error	<ul style="list-style-type: none"> – Measuring concepts from which not enough data is available – Inferring attitudes – Defining valid information
Measurement error	<ul style="list-style-type: none"> – Non-trackable target – Meter not installed – Uninstalling the meter – New non-tracked device – Technology limitations – Technology errors – Hidden behaviours – Shared device – Social desirability – Extraction error
Processing error	<ul style="list-style-type: none"> – Coding error – Aggregation at the domain level – Data anonymization
Coverage error	<ul style="list-style-type: none"> – Non-trackable individuals
Sampling error	– Same error causes than for surveys
Missing data error	<ul style="list-style-type: none"> – Noncontact – Non-consent – Non-trackable target – Meter not installed – Uninstalling the meter – New non-tracked device – Technology limitations – Technology error – Hidden behaviour – Social desirability – Extraction error
Adjustment error	– Same error causes than for surveys

Sampling and adjustment errors have no specific error causes

Practical recommendations

- 1. Clearly define what your tracked data is measuring beforehand**

Practical recommendations

1. Clearly define what your tracked data is measuring beforehand

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

Measure: *average time recorded of the visits to online political outlets' URLs.*

Practical recommendations

1. Clearly define what your tracked data is measuring beforehand

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

Measure: *average time recorded of the **visits** to online political outlets' URLs.*

- What is considered a visit?

Practical recommendations

1. Clearly define what your tracked data is measuring beforehand

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

Measure: *average time recorded of the **visits** to online political **outlets**' URLs.*

- What is considered a visit?
- Which online outlets?

Practical recommendations

1. Clearly define what your tracked data is measuring beforehand

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

Measure: *average time recorded of the visits to online political outlets' URLs.*

- What is considered a visit?
- Which online outlets?
- Which URLs should be considered political?

Practical recommendations

1. Clearly define what your tracked data is measuring beforehand

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

Measure: *average time recorded of the visits to online political outlets' URLs.*

- What is considered a visit?
- Which online outlets?
- Which URLs should be considered political?
- What time frame to use to compute an average?

Practical recommendations

- 2. Consider the impact of the chosen technologies on data quality**

Practical recommendations

2. Consider the impact of the chosen technologies on data quality

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

Practical recommendations

2. Consider the impact of the chosen technologies on data quality

Apps	
Where?	Device
Devices	Not iOS
Continuous?	Yes
Types of data	URLs, Time, Device, Search terms, Incognito



Practical recommendations

2. Consider the impact of the chosen technologies on data quality

Apps

Where?
Device

Devices
Not iOS

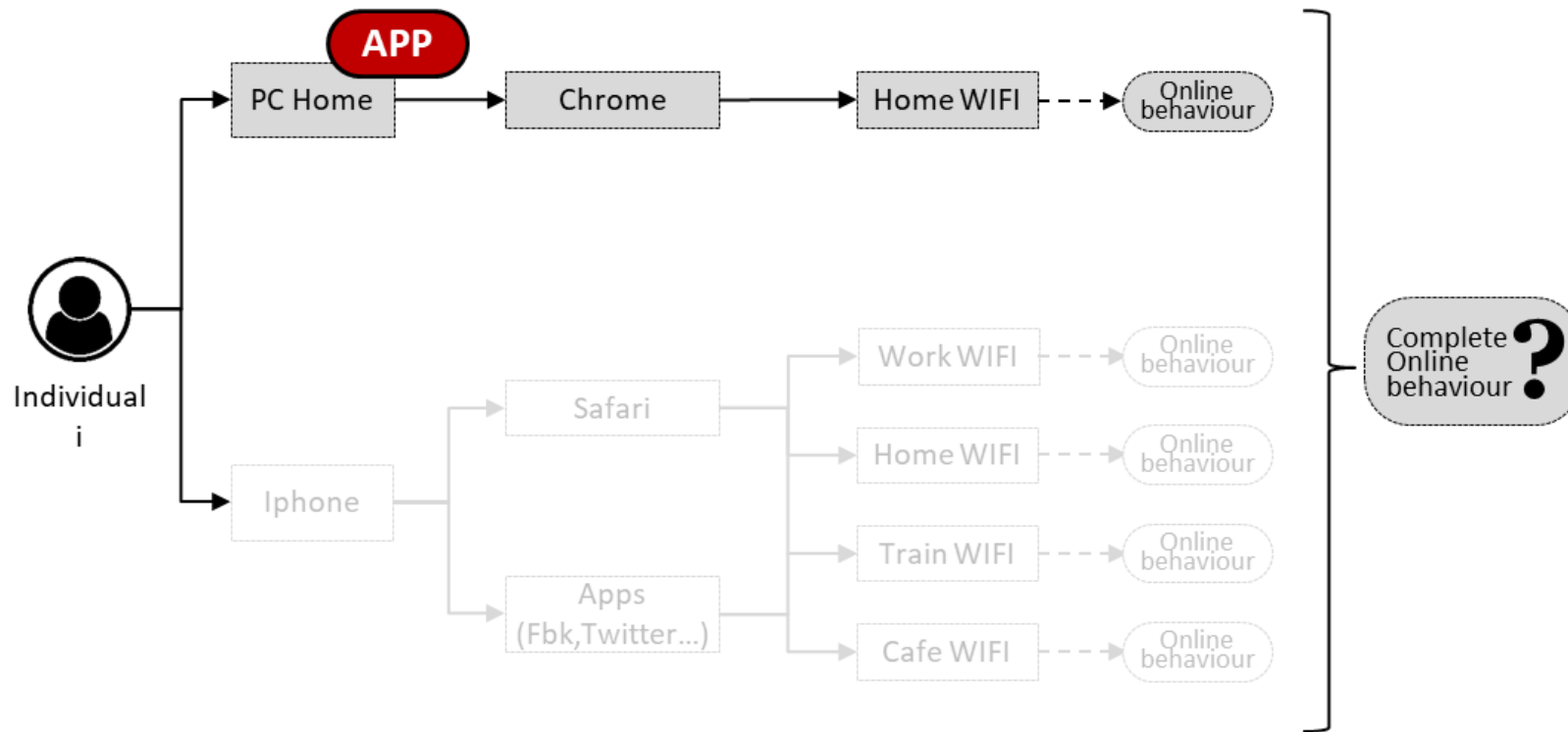
Continuous?
Yes

Types of data
URLs, Time, Device,
Search terms,
Incognito



Practical recommendations

3. Explore strategies to increase the willingness of individuals to install the meter in all targets

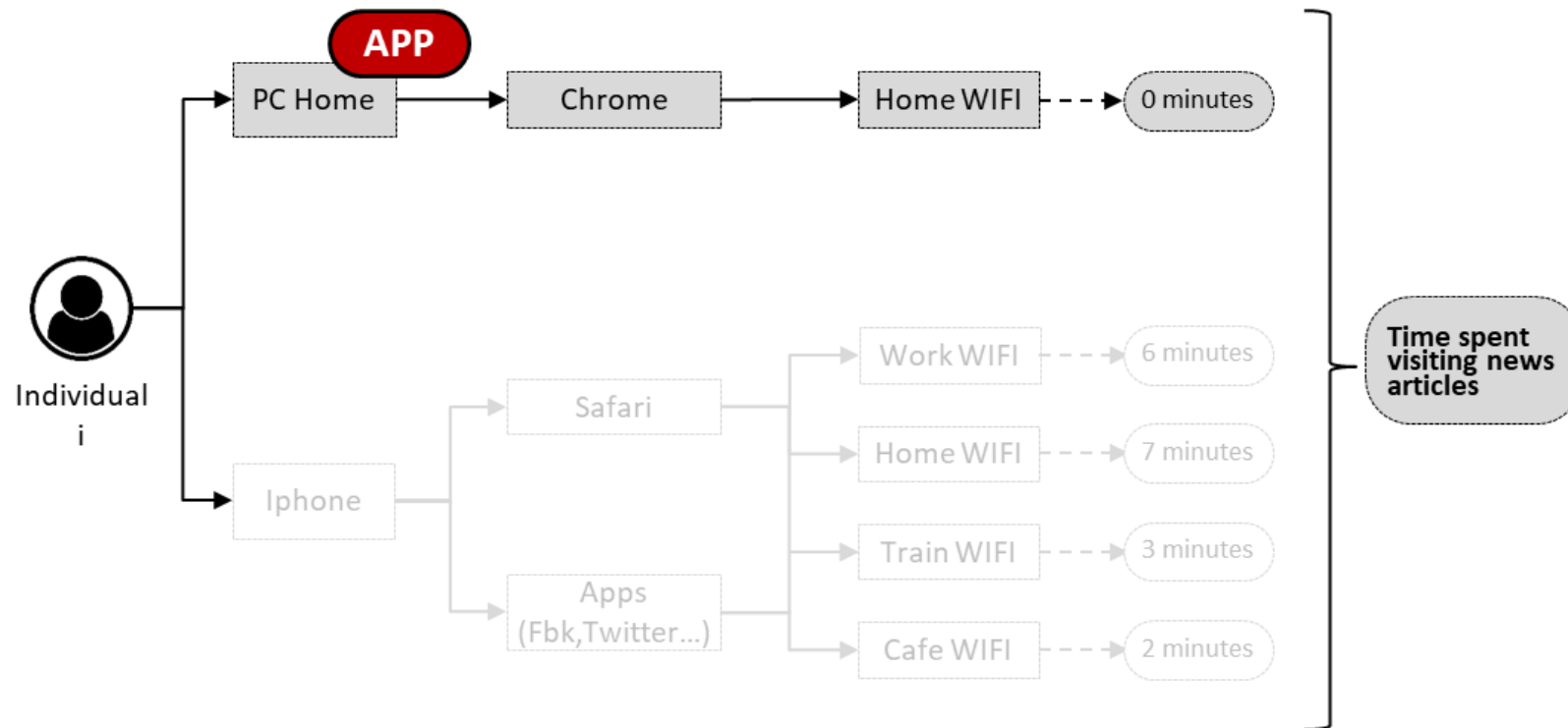


Practical recommendations

- 3. Explore strategies to increase the willingness of individuals to install the meter in all targets**
 - Tracking technologies present different installations processes.
 - Multiple tracking technologies might need to be installed for the same participant.
 - Targets (devices / browsers / networks used) are unknown.

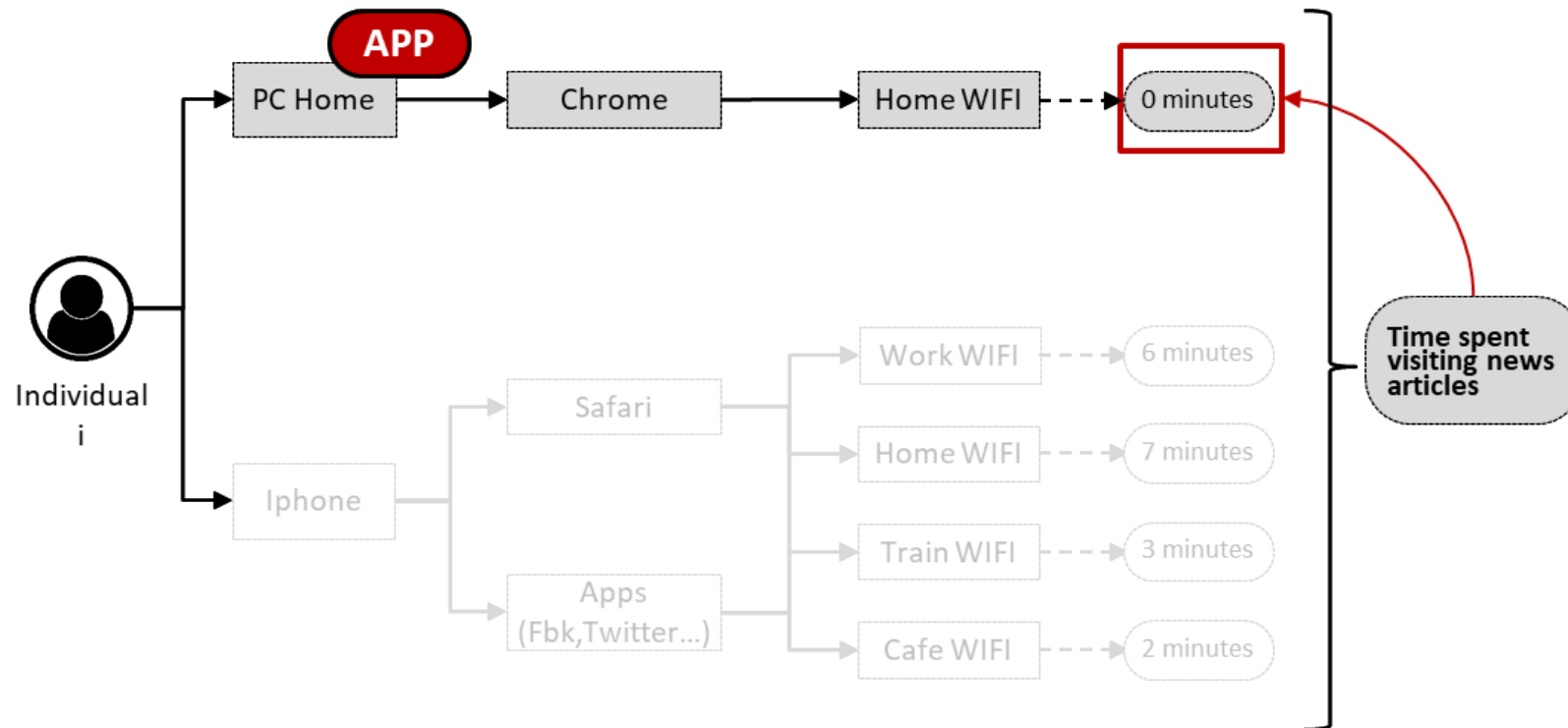
Practical recommendations

4. Define strategies to maximise the information available to identify missing data



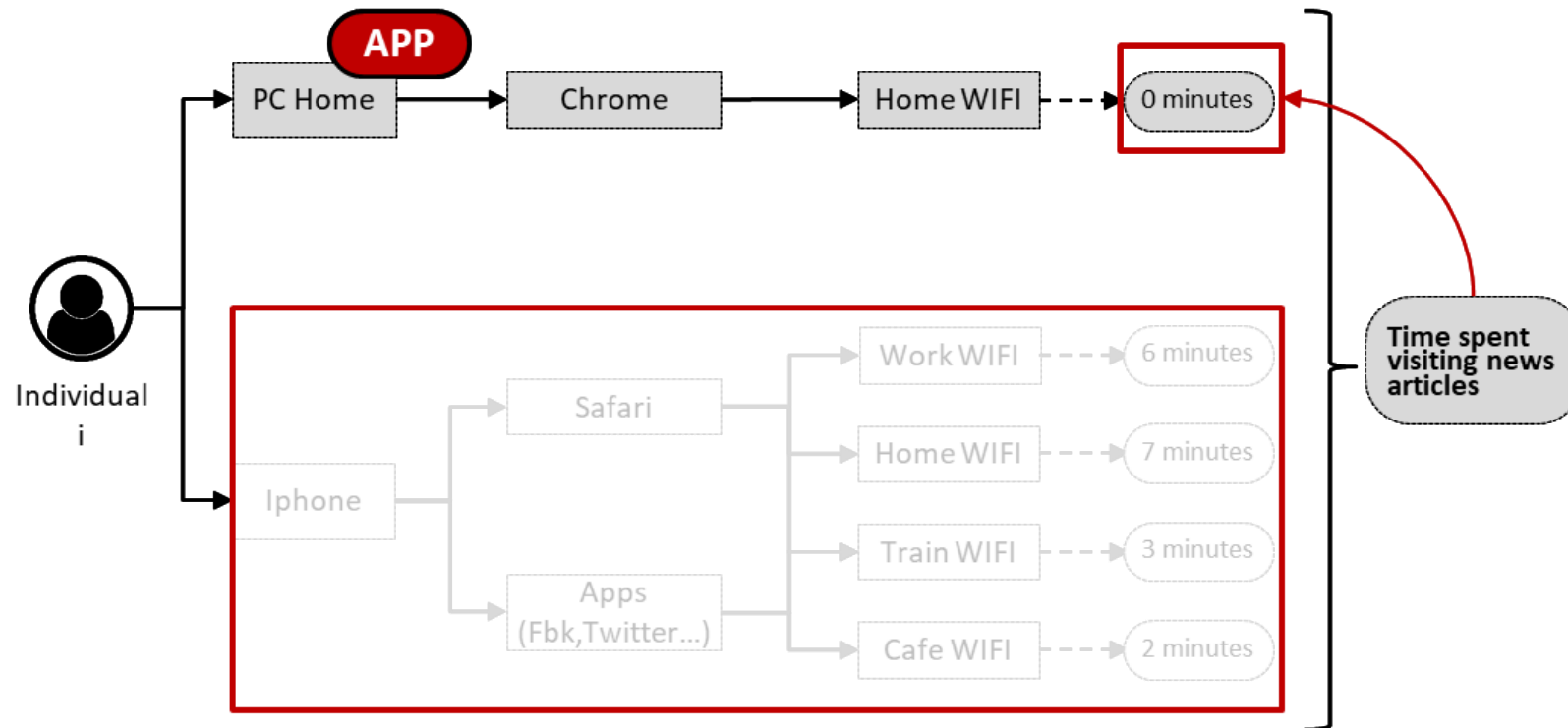
Practical recommendations

4. Define strategies to maximise the information available to identify missing data



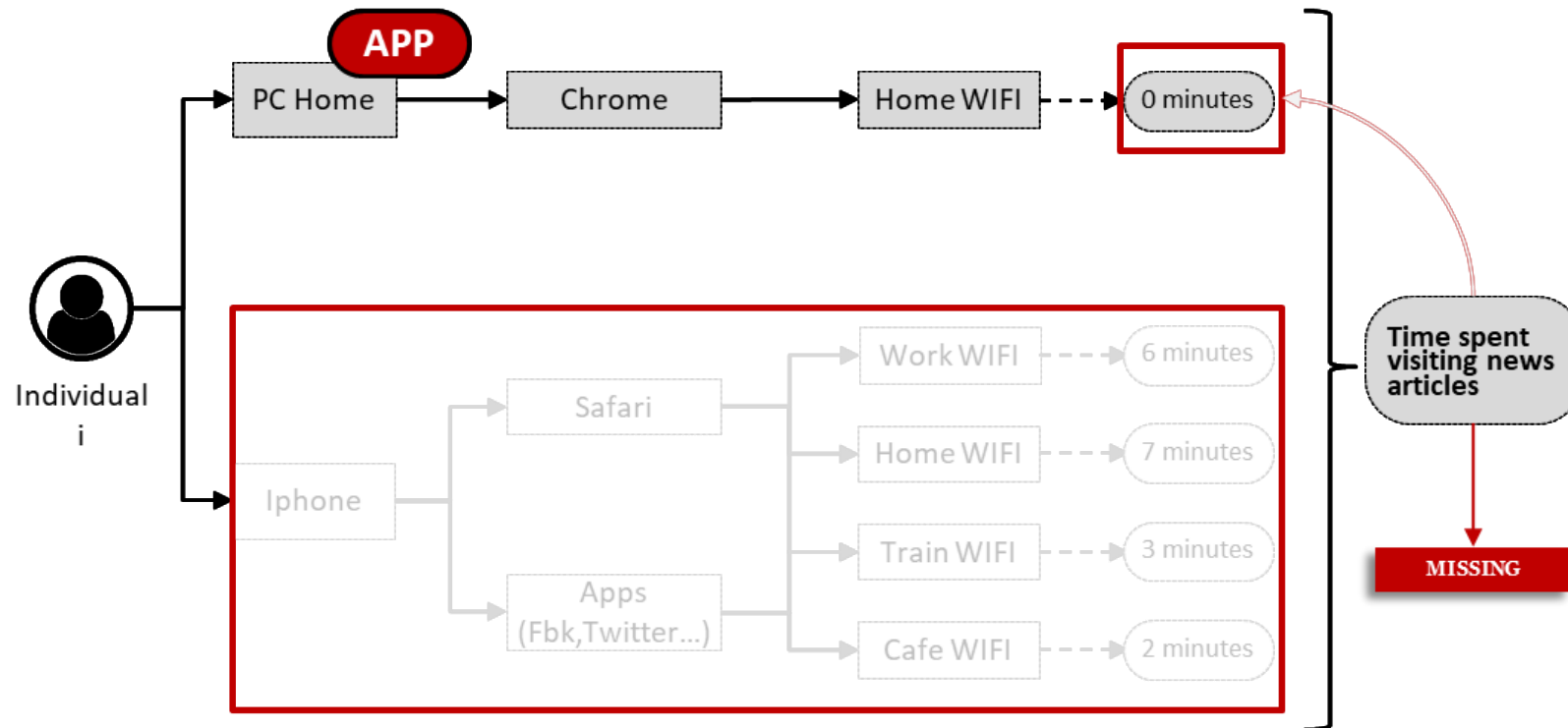
Practical recommendations

4. Define strategies to maximise the information available to identify missing data



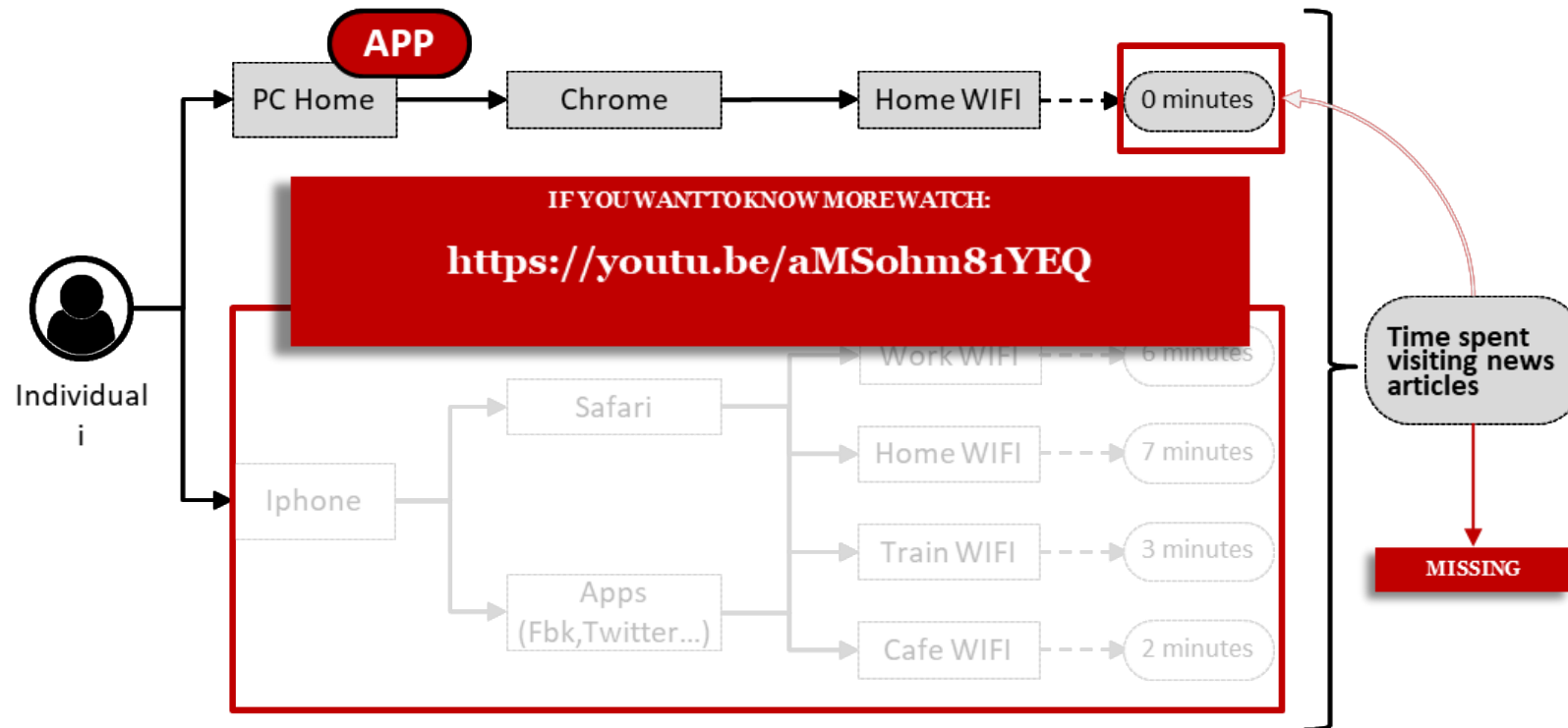
Practical recommendations

4. Define strategies to maximise the information available to identify missing data



Practical recommendations

4. Define strategies to maximise the information available to identify missing data



Limits

1. One specific definition of data quality.
2. Lack of previous empirical research.
3. Tracking technologies are constantly evolving.
4. Metered data errors are considered independently.

Take-home messages

1. Using metered data is complex and many decisions must be taken.
2. Reporting these decisions and conducting robustness checks is necessary.
3. More empirical research is needed.
4. This framework can help on all these aspects.

Thanks!

Questions?

Oriol J. Bosch



o.bosch-jover@lse.ac.uk



orioljbosch



<https://orioljbosch.com/>

Bosch, O.J., and M. Revilla (2021). **“When survey science met online tracking: presenting an error framework for metered data.”** RECSM Working Papers Series, 62



Universitat
Pompeu Fabra
Barcelona



THE LONDON SCHOOL
OF ECONOMICS AND
POLITICAL SCIENCE ■

