New Opportunities to Enhance or Replace Conventional Web Survey Data

Nuremberg, 20 January 2023

Melanie Revilla | IBEI

***Acknowledgments**:*
This project has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement No. 849165); 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)); and the BBVA foundation under their grant scheme to scientific research teams in economy and digital society, 2019.
I want to thank Oriol Bosch, Patricia Iglesias, and Carlos Ochoa for their feedback on previous drafts of this presentation.

Which new opportunities?

WHICH NEW OPPORTUNITIES

Growing use of Internet

More and more of people's life happens **online**

Average daily time[1] spent online by each internet user

- in 2021 **6h58**
- in 2016 **6h29**

+30 min in 5 years

Affects all domains of life

→ Including aspects related to (un)employment[2]

80% of all **job searches** are done online as of 2022

90% of **recruiters use LinkedIn** to search for candidates to fill job openings

[1] https://datareportal.com/reports/digital-2022-global-overview-report; [2] https://www.zippia.com/advice/job-search-statistics/

Growing use of smartphones

More and more of the online activity is done through **smartphones**

83% of the world population have smartphones[1]

92% of Internet users worldwide access the Internet through smartphones[2]

Smartphones are also used to participate in **web surveys**

Smartphones used in

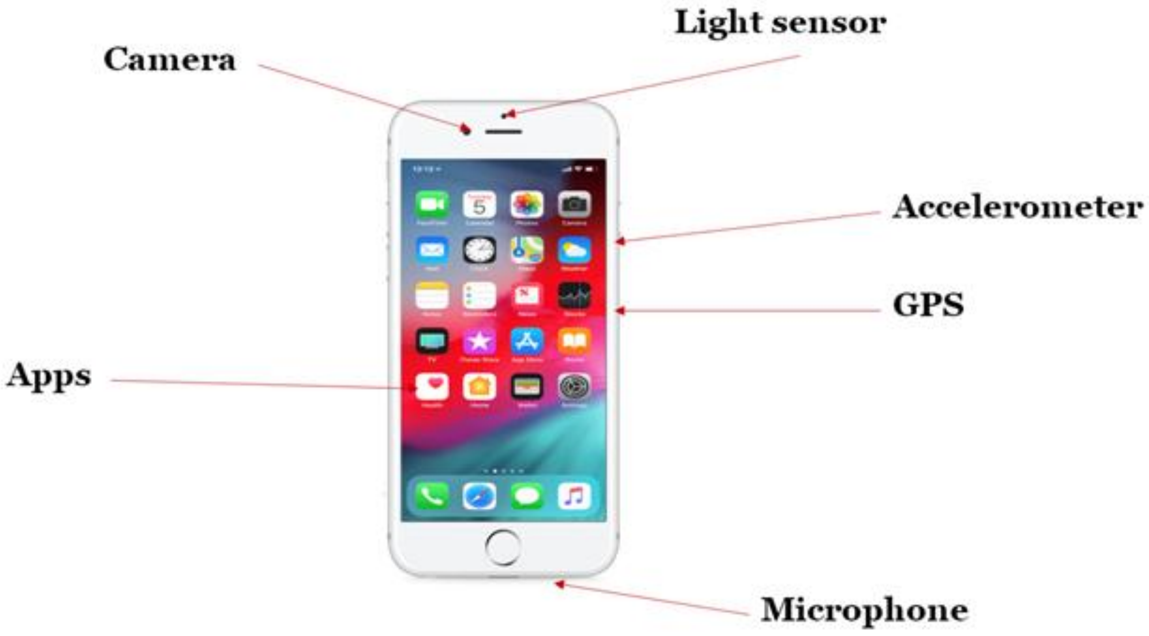
- **79%** of surveys completed by Millennials in US Netquest panel
- **36%** of surveys completed by Boomers in US Netquest panel[3]

➡ Creates both new challenges and opportunities

[1] https://www.bankmycell.com/blog/how-many-phones-are-in-the-world; [2] https://datareportal.com/reports/digital-2022-global-overview-report; [3] In 2017/2018

Different new opportunities

- Focus on possibility to **collect other data types**
  - Lot of different data types
  - Each one has its own potential benefits and risks
  - Important to study them separately
  - But also a lot in common



WHICH NEW OPPORTUNITIES

New data types considered

In-the-moment surveys triggered by such data

METERED DATA

Obtained through a tracking application (“meter”) installed by the participants on their devices to register at least the URLs of the webpages visited

GEOLOCATION DATA

Obtained through a tracking application installed on participants’ mobile devices to register at least the GPS coordinates

Most of those data can also be collected for PCs

VISUAL DATA

Screenshots
Photos/videos taken during the survey
Visual files saved on (or accessible from) the device

VOICE DATA

Dictation
Voice recording

WHICH NEW OPPORTUNITIES

New data types considered

In-the-moment surveys triggered by such data

METERED DATA

Obtained through a tracking application (“meter”) installed by the participants on their devices to register at least the URLs of the webpages visited

GEOLOCATION DATA

Obtained through a tracking application installed on participants’ mobile devices to register at least the GPS coordinates

Benefits not expected for all concepts but enough applications to make the investigation worth it

VISUAL DATA

Screenshots
Photos/videos taken during the survey
Visual files saved on (or accessible from) the device

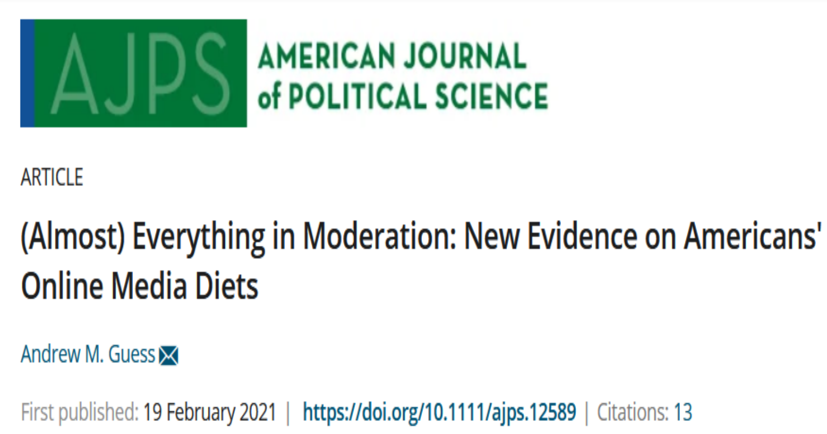
VOICE DATA

Dictation
Voice recording

Using metered data

Metered data are already used in substantive research…

More than **70 papers** published since 2016 using metered data



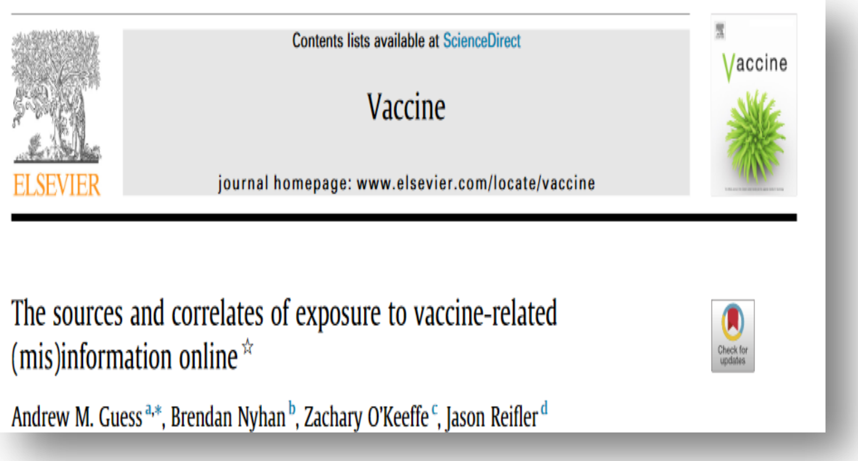
AMERICAN JOURNAL of POLITICAL SCIENCE

ARTICLE

(Almost) Everything in Moderation: New Evidence on Americans' Online Media Diets

Andrew M. Guess

First published: 19 February 2021 | https://doi.org/10.1111/ajps.12589 | Citations: 13



Contents lists available at ScienceDirect

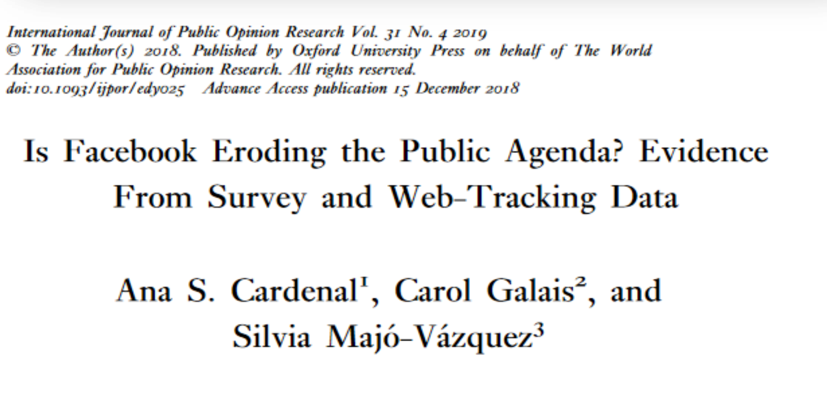
Vaccine

journal homepage: www.elsevier.com/locate/vaccine

ELSEVIER

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

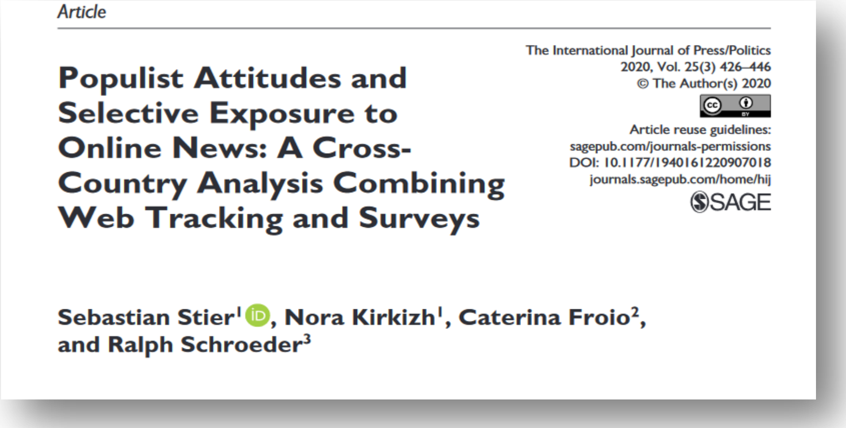
Andrew M. Guess[a,*], Brendan Nyhan[b], Zachary O'Keeffe[c], Jason Reifler[d]



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. Cardenal[1], Carol Galais[2], and Silvia Majó-Vázquez[3]



Article

Populist Attitudes and Selective Exposure to Online News: A Cross-Country Analysis Combining Web Tracking and Surveys

The International Journal of Press/Politics
2020, Vol. 25(3) 426–446
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1940161220907018
journals.sagepub.com/home/hij
SAGE

Sebastian Stier[1], Nora Kirkizh[1], Caterina Froio[2], and Ralph Schroeder[3]

…. because they present many opportunities

Data Characteristics

Passively collected

Data already collected (metered panels)

Continuous data collection

Granular

Massive amount of data

Opportunities

Reduced measurement errors due to recall limitations, to people not knowing, or to social desirability…

Reduced effort for participants

Can decide today to look at what happened months ago

Study before/after

Study the process/journey

Possible applications

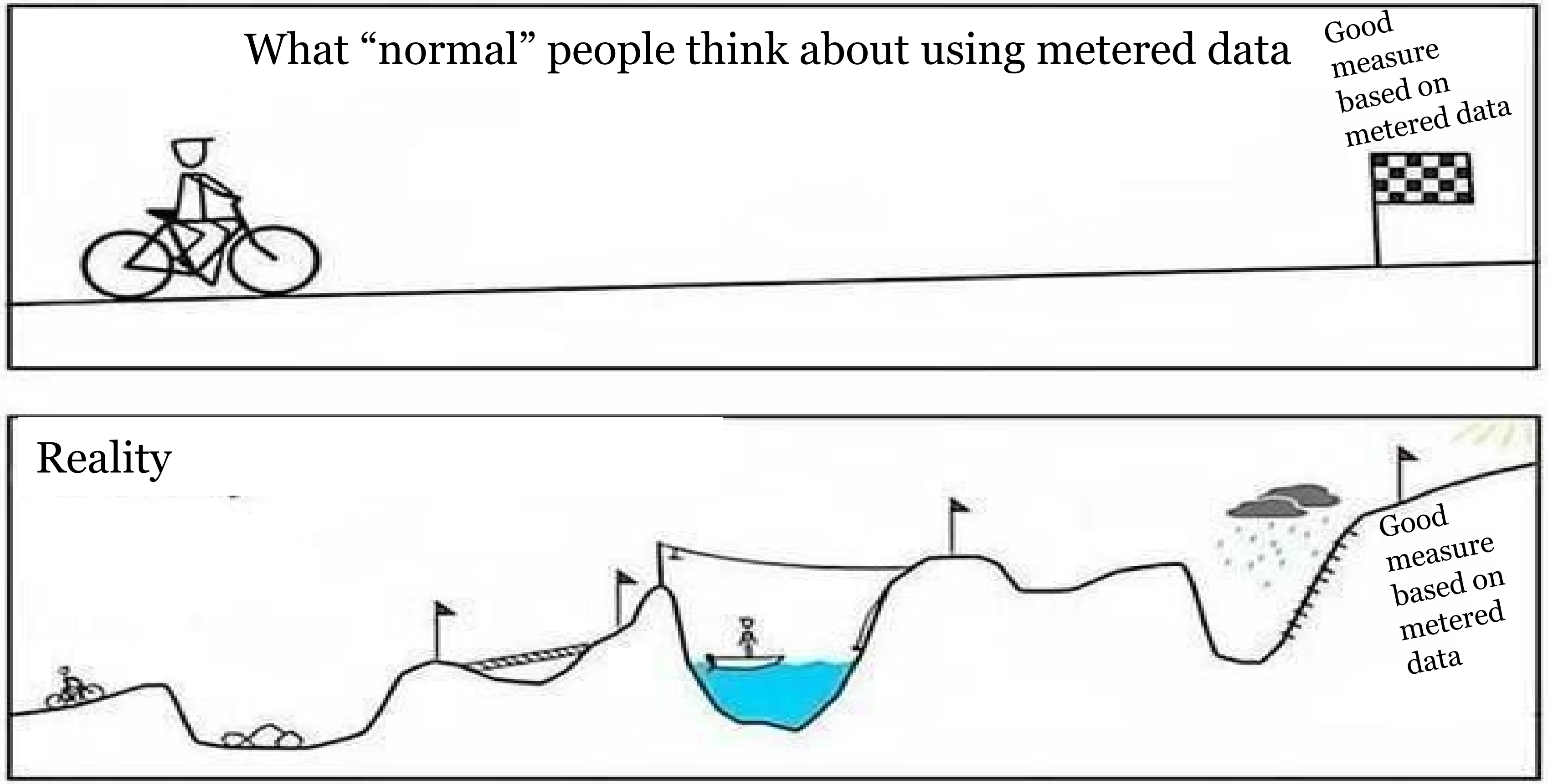
Effect of key events

- ➢ How the pandemic affected online behaviors?
- ➢ How did online behaviors changed across different phases of the pandemic?

How people search for a job online?

- ➢ Which webs do they use?
- ➢ How do they get to the offers they apply to?
- ➢ How many job offers do they visit?

But this is not that easy…


What "normal" people think about using metered data
Good measure based on metered data
Reality
Good measure based on metered data

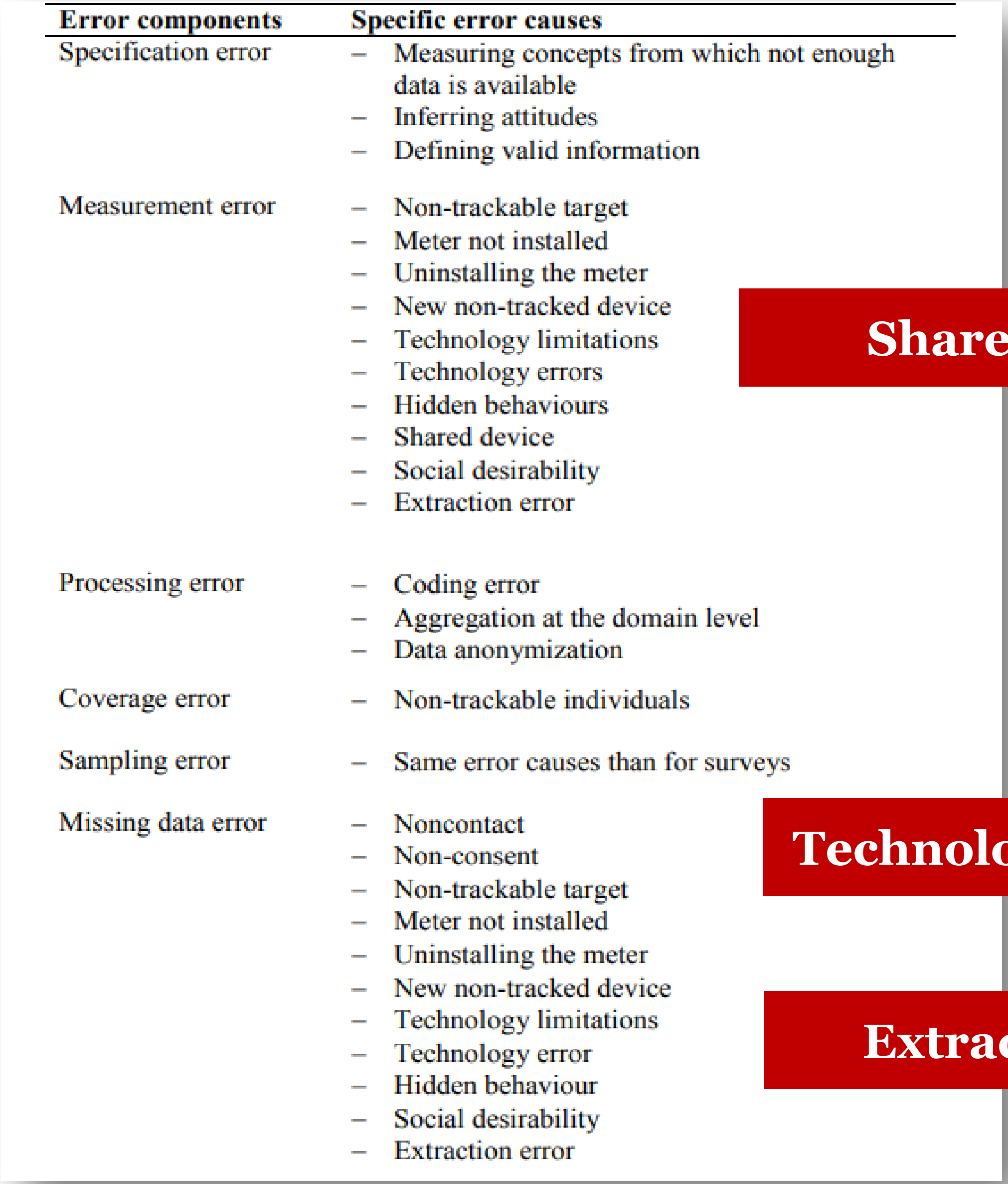
THIS IS NOT THAT EASY

Different types of errors

- Many possible kinds of errors
  - We developed a **Total error framework for metered data** (TEM) = adaptation of the total survey error (TSE) framework to metered data (Bosch & Revilla, 2022a)
  - Provides an overview of all possible errors and their causes

THIS IS NOT THAT EASY

Different types of errors



Error components	Specific error causes
Specification error	– Measuring concepts from which not enough data is available
	– Inferring attitudes
	– Defining valid information
Measurement error	– 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	– Coding error
	– Aggregation at the domain level
	– Data anonymization
Coverage error	– Non-trackable individuals
Sampling error	– Same error causes than for surveys
Missing data error	– 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

Shared devices

Technology limitations

Extraction errors

THIS IS NOT THAT EASY

Size of the errors

- Next, we investigated how large some of these errors are and to what extent they may affect the final estimates (Bosch, Sturgis & Kuha, 2022)

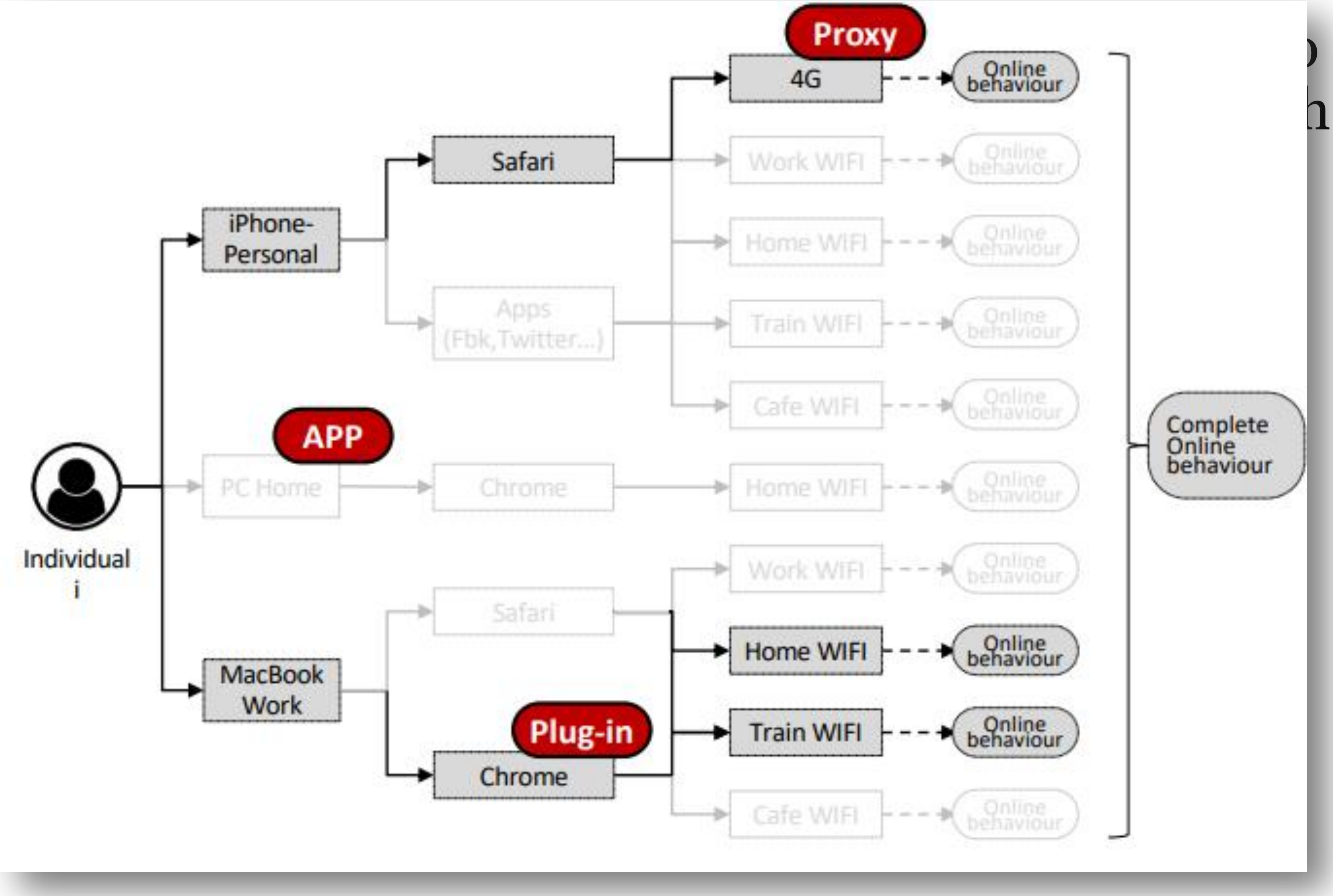
- Focus on **tracking undercoverage**
  - ➢Participants do not install the meter in all devices/browsers

THIS IS NOT THAT EASY

Size of the errors

- Next, we in… what extent they … ha, 2022)

- Focus on tr…
  - ➢Participa…



THIS IS NOT THAT EASY

Size of the errors

- Next, we investigated how large some of these errors are and to what extent they may affect the final estimates (Bosch, Sturgis & Kuha, 2022)

- Focus on **tracking undercoverage**
  - ➢Participants do not install the meter in all devices/browsers

TRI-POL data[1]	Spain, Portugal, Italy 3 survey waves + metered data 2 weeks before/after each survey
Survey+meter	Comparing survey answers to information from the meter We found that **80-85%** of participants are not fully covered
Simulations	Biased univariate and multivariate estimates

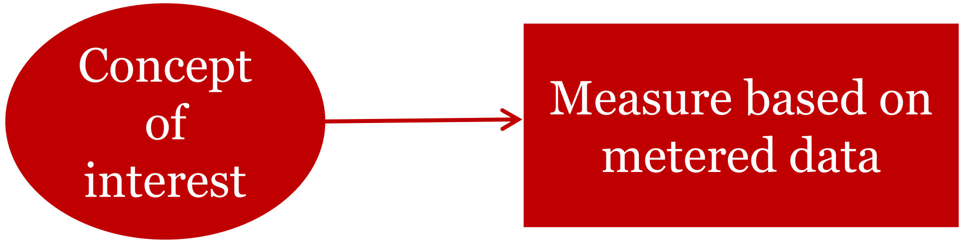
[1] https://www.upf.edu/web/tri-pol

THIS IS NOT THAT EASY

Validity

- We studied the **validity of measures** based on metered data (Bosch & Revilla, 2022b)
  - Focusing on "**online (written) news media exposure**"

- How to create a measure of this concept using metered data?



- Many decisions
  - Which URLs are considered "online written **news media**"?
  - What is considered as **being "exposed"**?
  - How many **days of tracking** should be used?
  - Etc.

THIS IS NOT THAT EASY

Validity

- Combining all these decisions → theoretically we could create **>8,000** variables that should all measure the same concept of interest

Characteristics	Choices
Metric	Visits, Seconds, Days, Media
List of traces	
List of media	Own, Tranco, Alexa, Cisco, Majestic
Top media	10, 20, 50, 100, 200, All
Information	All domain level, subdomains defined 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

THIS IS NOT THAT EASY

Validity

- How do these decisions affect the **convergent** and **predictive validity** of the measures?

Convergent validity All variables measuring the same concept should highly correlate with each other

Predictive validity Measures that correlate more with political knowledge assumed to be better

- TRI-POL data
  - Low to average convergent validity
  - High fluctuations in predictive validity depending on the choices

THIS IS NOT THAT EASY

Summing up

Researchers

More expensive

Dependence on private companies

Selection bias?

Data protection/ethical issues?

Different types of errors

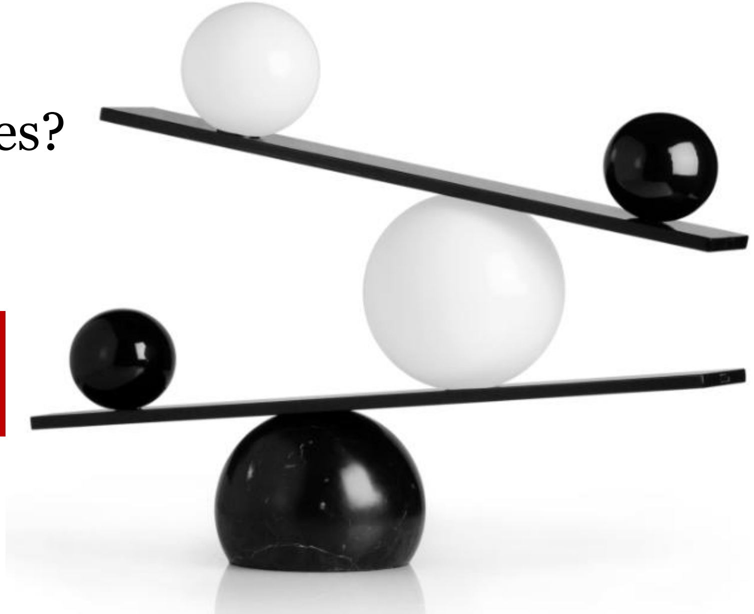
Disadvantages

Participants

Privacy issues?

Loss of control?

New skills needed?



Researchers

Reduce some of the issues related to measurement errors

Massive amount of data
Continuous /real time

➡ New insights

Benefits

Participants

Reduced time dedicated to provide information

Reduced effort

Combining metered data and surveys

Even much more opportunities

Identify and/or compare groups

Identify group of people who suffer a job loss to study the impact of this loss. Differences in how people search for a job online by gender/ethnicity/social class?

Confirm behaviors

Are the behaviors observed with the metered data really done by the sampling unit? Are the behaviors really intentional?

Add information about feelings, opinions, etc.

Did they like specific job search websites?
How did they feel about some job offers?
Did they understand some job offer content properly?

Ask explanations

Why did the participants use this website to search for a job?
Why did they decide to apply to a specific offer?

COMBINING METERED DATA AND SURVEYS

Even much more opportunities

Identify and/or compare groups

Identify group of people who suffer a job loss to study the impact of this loss. Differences in how people search for a job online by gender/ethnicity/social class?

Confirm behaviors

Are the behaviors observed with the metered data really done by the sampling unit? Are the behaviors really intentional?

Add information about feelings, opinions, etc.

Did they like specific job search websites?
How did they feel about some job offers?
Did they understand some job offer content properly?

Ask explanations

Why did the participants use this website to search for a job?
Why did they decide to apply to a specific offer?

COMBINING METERED DATA AND SURVEYS

Even much more opportunities

Identify and/or compare groups

Identify group of people who suffer a job loss to study the impact of this loss.
Differences in how people search for a job online by gender/ethnicity/social class?

Confirm behaviors

Are the behaviors observed with the metered data really done by the sampling unit? Are the behaviors really intentional?

Add information about feelings, opinions, etc.

Did they like specific job search websites?
How did they feel about some job offers?
Did they understand some job offer content properly?

Ask explanations

Why did the participants use this website to search for a job?
Why did they decide to apply to a specific offer?

COMBINING METERED DATA AND SURVEYS

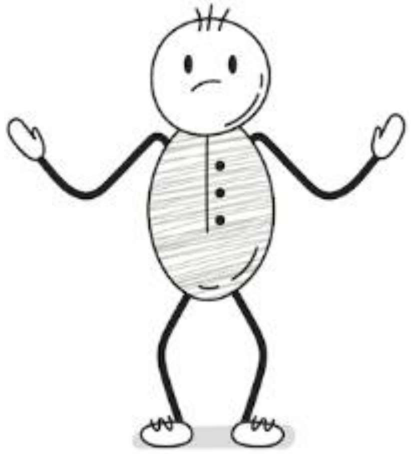
Even much more opportunities

Identify and/or compare groups	Identify group of people who suffer a job loss to study the impact of this loss. Differences in how people search for a job online by gender/ethnicity/social class?
Confirm behaviors	Are the behaviors observed with the metered data really done by the sampling unit? Are the behaviors really intentional?
Add information about feelings, opinions, etc.	Did they like specific job search websites? How did they feel about some job offers? Did they understand some job offer content properly?
Ask explanations	Why did the participants use this website to search for a job? Why did they decide to apply to a specific offer?

Problem due to limitations of human memory

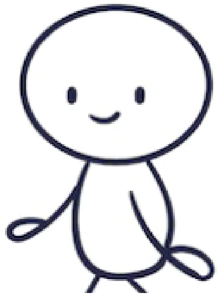
- People **do not recall** all the information we ask in the surveys
  - Human memory is cleaned of irrelevant information when people sleep (Izawa et al., 2019)

- The way we **recall differs from** the way we **experience** things
  - Remembering-self ≠ experiencing-self (Kahneman & Riis, 2005)



How did you feel about the job offer you saw on the 4th of December?

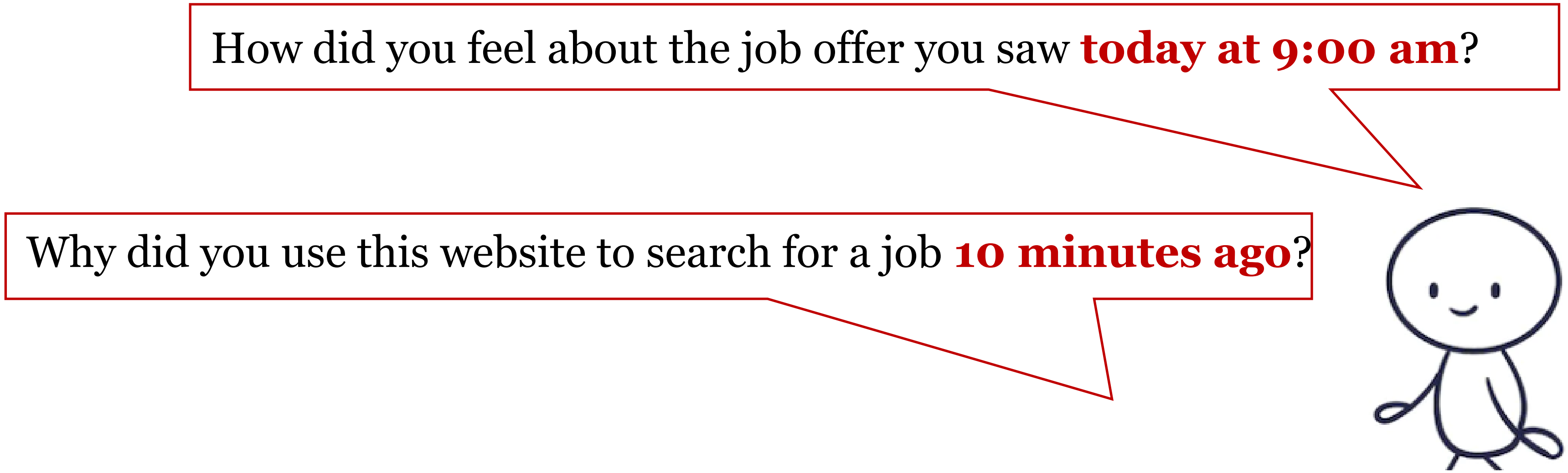
Why did you use this website to search for a job in the last 3 months?



In-the-moment surveys triggered by metered data

- When we detect a behavior of interest using the metered data → send a survey invitation **immediately** to ask more information

- Reduce the time between the event of interest and the survey



How did you feel about the job offer you saw **today at 9:00 am**?

Why did you use this website to search for a job **10 minutes ago**?

Current experiment: content

- Implement in parallel
  - A **conventional** survey
  - An **in-the-moment** survey triggered by metered data
- Event of interest: **online job application**
- Survey asks about:
  - Content of the offer to which the participants apply
  - Fulfillment of the job requirements (and which ones)
  - Reasons for applying
  - How much the offer fits with what they look for
  - Socio-demographic information about the participants
  - Attitudinal questions (e.g., self-confidence, conformity)
  - Evaluation of the survey
- Netquest (metered) panel in Spain (still in programming)

Current experiment: main objectives

Substantive (Maria-José González & Clara Cortina):

- Study differences between men and women in applying online when not meeting all requirements
- Hypothesis: women apply less than men to jobs when not meeting all requirements

Methodological (Carlos Ochoa):

- Study the **feasibility** of using in-the-moment surveys triggered by metered data
- **Compare the samples and quality** of the data obtained in a conventional versus an in-the-moment survey
- Show that we can get **new insights** with the in-the-moment survey

But this is (again) not that easy…

Many challenges

- **Identifying the triggering events**
  - –It is difficult to identify all the job websites where applications can occur
  - –In some job websites, it is not possible to identify if someone applied to an offer based on the URLs (e.g., the URLs do not change when applying)
  - –People can apply to a job online in other ways (e.g., by email or though an app) that cannot be detected with the metered data
  - –The URLs can change so necessary to revise the list very regularly

➢ We are **not** able to detect (and thus invite) all the people applying to a job online

➢ We might also invite people who did not apply to a job, due to shared devices and possible errors

Many challenges

- **Identifying the triggering events**

Example infojobs.net

URL job offer:
https://www.infojobs.net/cornella-de-llobregat/atencion-cliente/of-i7756645bad414cb7e6172261f0587b

When I click "Apply"
*https://www.infojobs.net/candidate/***application***/index.xhtml?id_oferta=7756645bad414cb7e6172261f0587b&searchId=-2147483648&page=1&sortBy=DEFAULT&dgv=7958157879275968900*

I go through an intermediate page:
*https://www.infojobs.net/candidate/***application***/apply.xhtml?dgv=5245958816438271888*

Many challenges

- **Inviting the participants**
  - **How?** Need a tool to send invitations when detecting an event → **WebdataNow** (Revilla et al, 2022) → but also requires a panel software + a survey software + a passive data software
  - **When?** In most cases metered data do not allow to detect the end of an event but only the beginning → which **delay** should we use to send the invitation? How to maximize the chances that participants see the invitation **quickly enough**?
  - **What?** What can we say in the invitation to **motivate people** to participate without revealing information that might not be from the participant (e.g., the job application was made by someone else sharing the participant's device)?
  - **How many times?** Should we invite a participant **several times** if he/she applies to several job offers in a row?

COMBINING METERED DATA AND SURVEYS

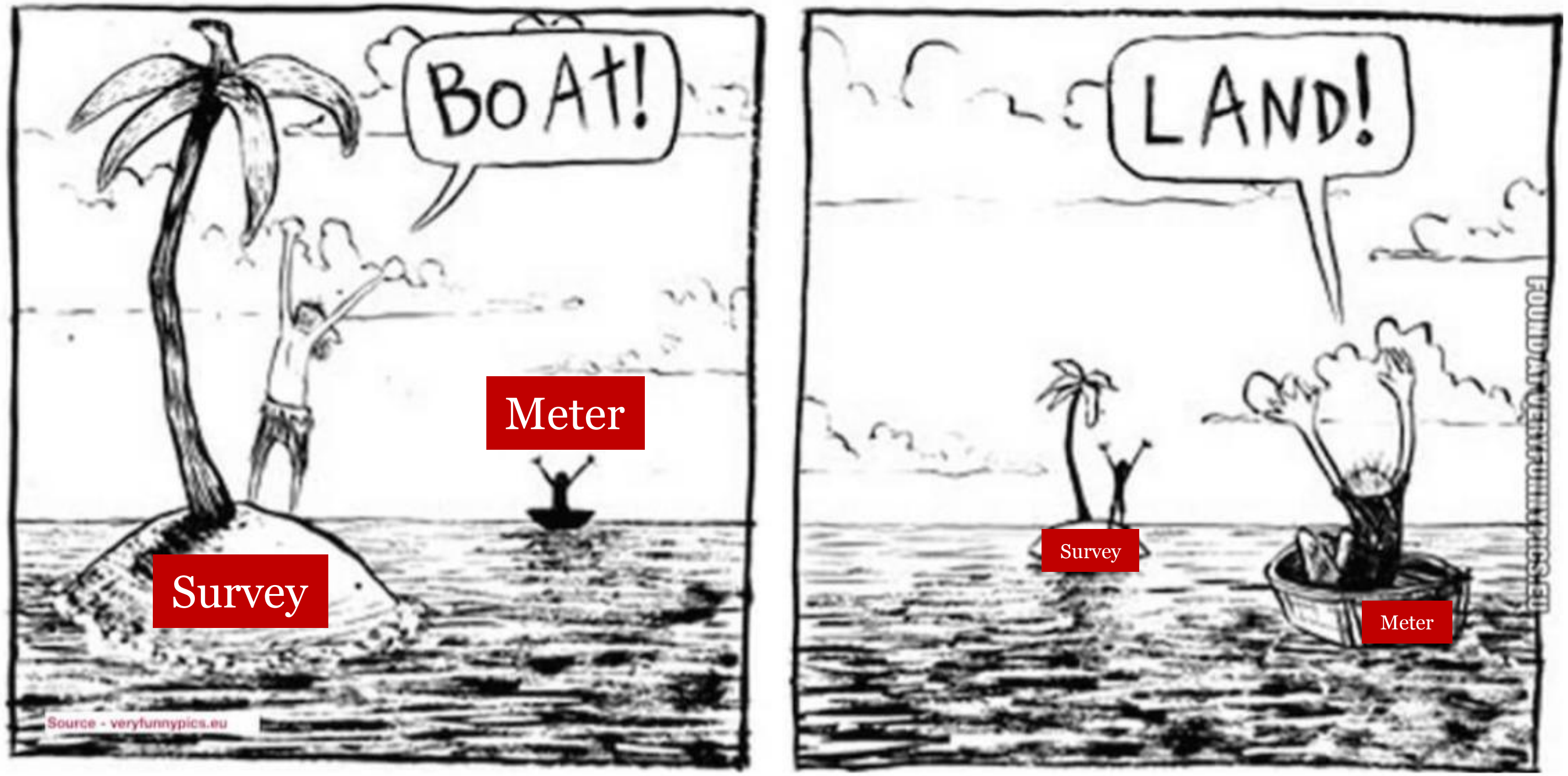
Many challenges

- **Controlling the sample**
  - How to get **sufficient sample size**? If the event is not very common, it can take months to get enough participants
  - How to get a **representative** sample?
- **Getting the information of interest**
  - If we want to combine information from the surveys with information from the metered data, we need to take into account all the possible errors of the metered data
  - If we also need information about the content of the webpages visited, we need to **extract the HTML** and not just the URLs → not all meters allow this + it is difficult to process such information

Conclusions

CONCLUSIONS

We are not saved yet…



Source - veryfunnypics.eu

CONCLUSIONS

Still a lot to be done

More research needed for all 4 types of data

- Learn more about the errors of those data
  - Types of errors, their size and how they affect the results in different contexts

- Better understand **when** to use those data
  - Need to identify when benefits > disadvantages, balancing those for researchers and participants
  - Need to understand better the mechanisms

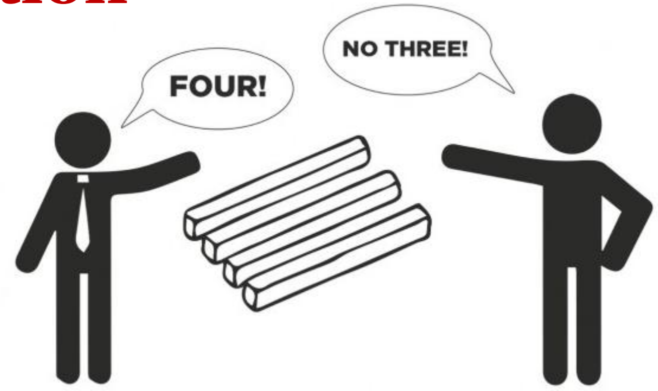
CONCLUSIONS

Still a lot to be done

More research needed for all 4 types of data

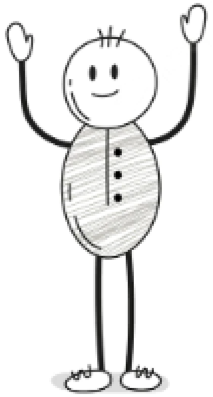
- Better understand **how** to use those data
  - To replace?
    - But errors will always be there → need to **acknowledge them** and think about **their consequences**
  - To combine?
    - Provide **different but complementary information**

➡ Look from different perspectives



Thanks!

Questions?



Melanie Revilla | IBEI

mrevilla@ibei.org

https://www.upf.edu/web/webdataopp

References

- Bosch, O.J., & Revilla, M. (2022a). When survey science met web tracking: presenting an error framework for metered data. Journal of the Royal Statistical Society: Series A (Statistics in Society). https://doi.org/10.1111/rssa.12956
- Bosch, O.J., & Revilla, M. (2022b). Is tracking all that it takes? Exploring the validity of news media exposure measurements created with metered data. AAPOR Annual Conference, 11th-13th May 2022
- Bosch, O.J., Sturgis, P., & Kuha, J. (2022). Track me but not really: Tracking undercoverage in metered data collection. AAPOR Annual Conference, 11th-13th May 2022.
- Izawa, S., Chowdhury, S., Miyazaki, T., Mukai, Y., Ono, D., Inoue, R., Ohmura, Y., Mizoguchi, H., Kimura, K., Yoshioka, M., Terao, A., Kilduff, T. S., & Yamanaka, A. (2019). REM sleep-active MCH neurons are involved in forgetting hippocampus-dependent memories. Science (New York, N.Y.), 365(6459), 1308–1313. https://doi.org/10.1126/science.aax9238
- Kahneman, D., & Riis, J. (2005). Living, and thinking about it: Two perspectives on life. In F. Huppert, B. Keverne, & N. Baylis (Eds.), *The science of well-being*. Oxford, England: Oxford University Press.
- Revilla, M., Ochoa, C., Iglesias, P. Antón, D. (2022). WebdataNow: a tool to send in-the-moment surveys triggered by passive data. OSF. http://doi.org/10.17605/OSF.IO/G3MSC.