New Opportunities to Enhance or Replace Conventional Web Survey Data

21 November 2022

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?

Main idea

web data opp

Smartphones are everywhere…

- More people have smartphones than toilets worldwide[1]

… including in **web surveys**

- Smartphones used in
  - **79%** of surveys completed by Millennials
  - **36%** of surveys completed by Boomers[2]

➡ Creates both new challenges and new opportunities

[1] https://www.globalcitizen.org/en/content/access-denied-toilets-Harpic-Waterorg-RB/
[2] Average for the US Netquest panel in 2017/2018 (Bosch et al., 2018)

Main idea

- 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’ 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

These new data are already used in substantive research

- A few examples
  - Metered data
    - Fake news consumption (e.g., Guess et al. 2020)
    - Time spent online (e.g., Festic et al. 2021)
  - In-the-moment surveys
    - Of people leaving polling stations, to predict an election outcome (e.g., Frankovic 2012)
    - To evaluate consumers' exposure to advertisement campaigns or access to health services (e.g., Clemens & Ginnis 2017)
  - Visual data
    - Mosquitoes presence (e.g., Mosquito Alert project[1])
    - Plants diseases (e.g., Kaur et al. 2019)
  - Voice data
    - Level of literacy (ask respondents to read loud some text)
    - Survey panelists' children

[1] http://www.mosquitoalert.com/en/

How could these data help?

Expected benefits

- It is clear that these new data types cannot enhance or replace all conventional survey questions
- However, there are many different questions, that cover concepts from many different disciplines, where we can expect benefits
  - Both on the researchers' and participants' sides

HOW COULD THE NEW DATA TYPES HELP?

Expected benefits

Researchers

- Reduce some of the issues related to measurement errors

Expected benefits

Researchers

- Reduce some of the issues related to measurement errors
- Massive amount of data

Expected benefits

Researchers

- Reduce some of the issues related to measurement errors
- Massive amount of data
- Real time / continuous (passively collected data)

Participants

- Reduce time dedicated to provide information
- Reduce efforts

→ Potential to **answer new research questions**

Expected benefits

Participants

- Reduce time dedicated to provide information
- Reduce efforts

Expected benefits

Researchers

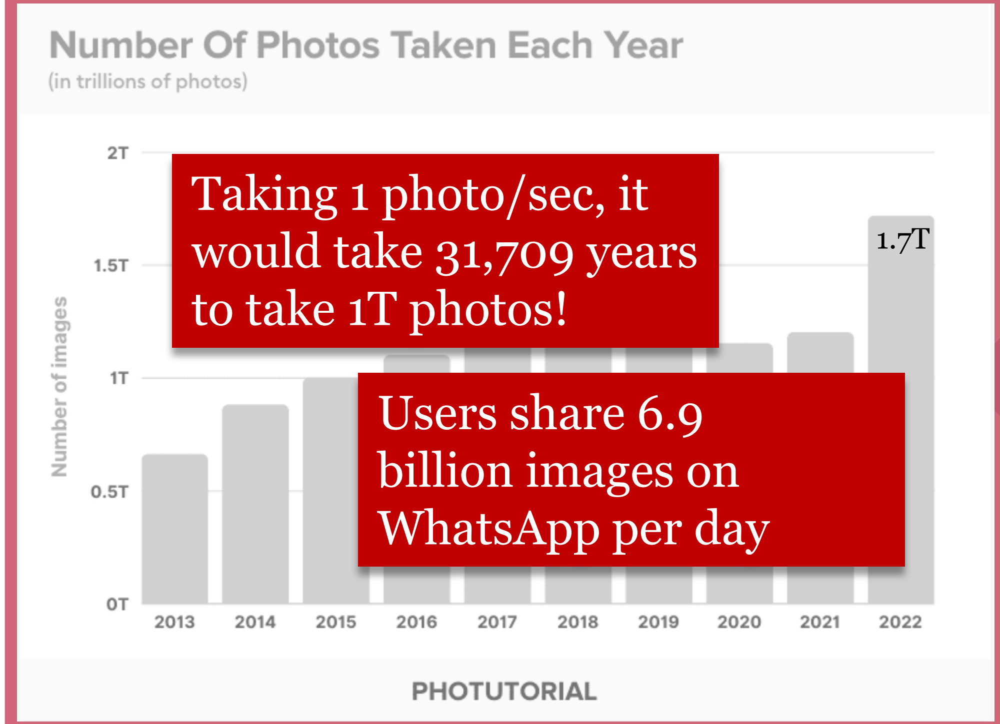
- Reduce some of the issues related to measurement errors
- Massive amount of data
- Real time / continuous (passive data)

Participants

- Reduce time dedicated to provide information
- Reduce efforts
- More enjoyable

HOW COULD THE NEW DATA TYPES HELP?

Expected benefits



Participants

- Reduce time dedicated to provide information
- Reduce efforts
- More enjoyable

Source: https://photutorial.com/photos-statistics/

But this is not that easy...

There are also (new) challenges

Researchers

- Need to adapt tools for data collection
- New skills needed for analyses
- Often more expensive
- Dependence on private companies
- Selection bias in who participates
- New types of errors (e.g., technological errors)
- Ethical / data protection issues

Participants

- Privacy issues
- Loss of control
- New skills needed (e.g., install an app)

Still lot of unknowns

- Do we really observe the benefits in practice?
- Are these benefits higher than the potential disadvantages?
- Not enough research yet to answer these questions
- Besides, it certainly depends on
  - the data types
  - the concepts of interest
  - how we use these data exactly
  - the target population
  - and more!
- Further research needed

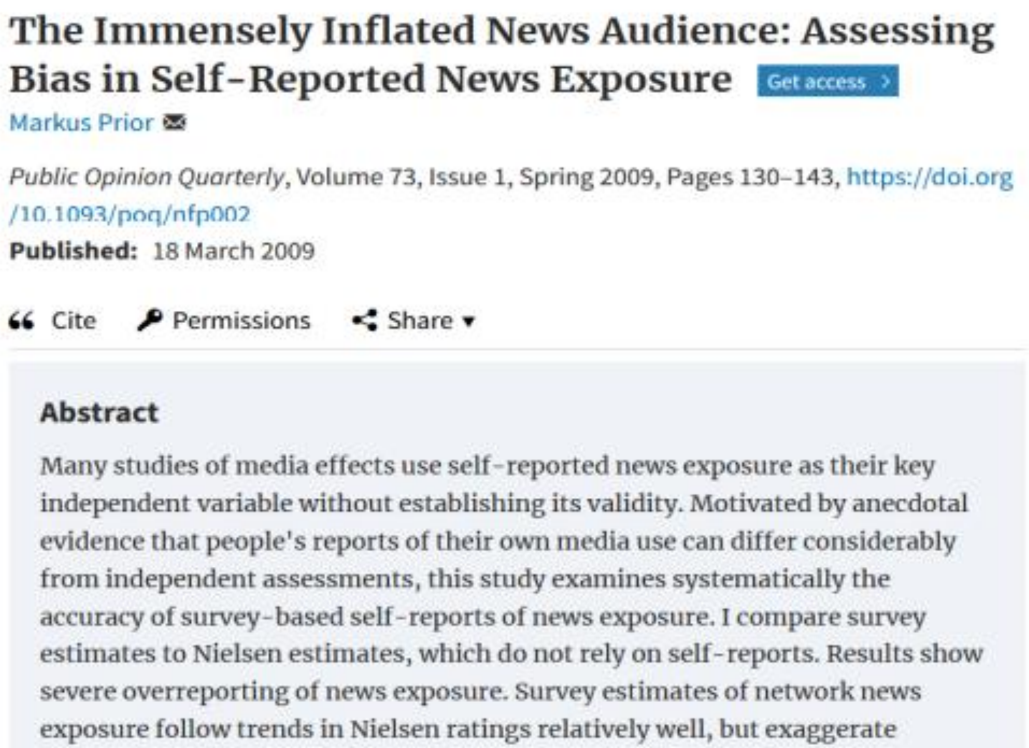
Example of metered data

Already commonly used

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

- Researchers usually assume that measures based on metered data are **perfect**

- Many even use them as the **gold standard**, to which they compare self-reported measures to assess their bias



The Immensely Inflated News Audience: Assessing Bias in Self-Reported News Exposure Get access

Markus Prior

Public Opinion Quarterly, Volume 73, Issue 1, Spring 2009, Pages 130–143, https://doi.org/10.1093/poq/nfp002

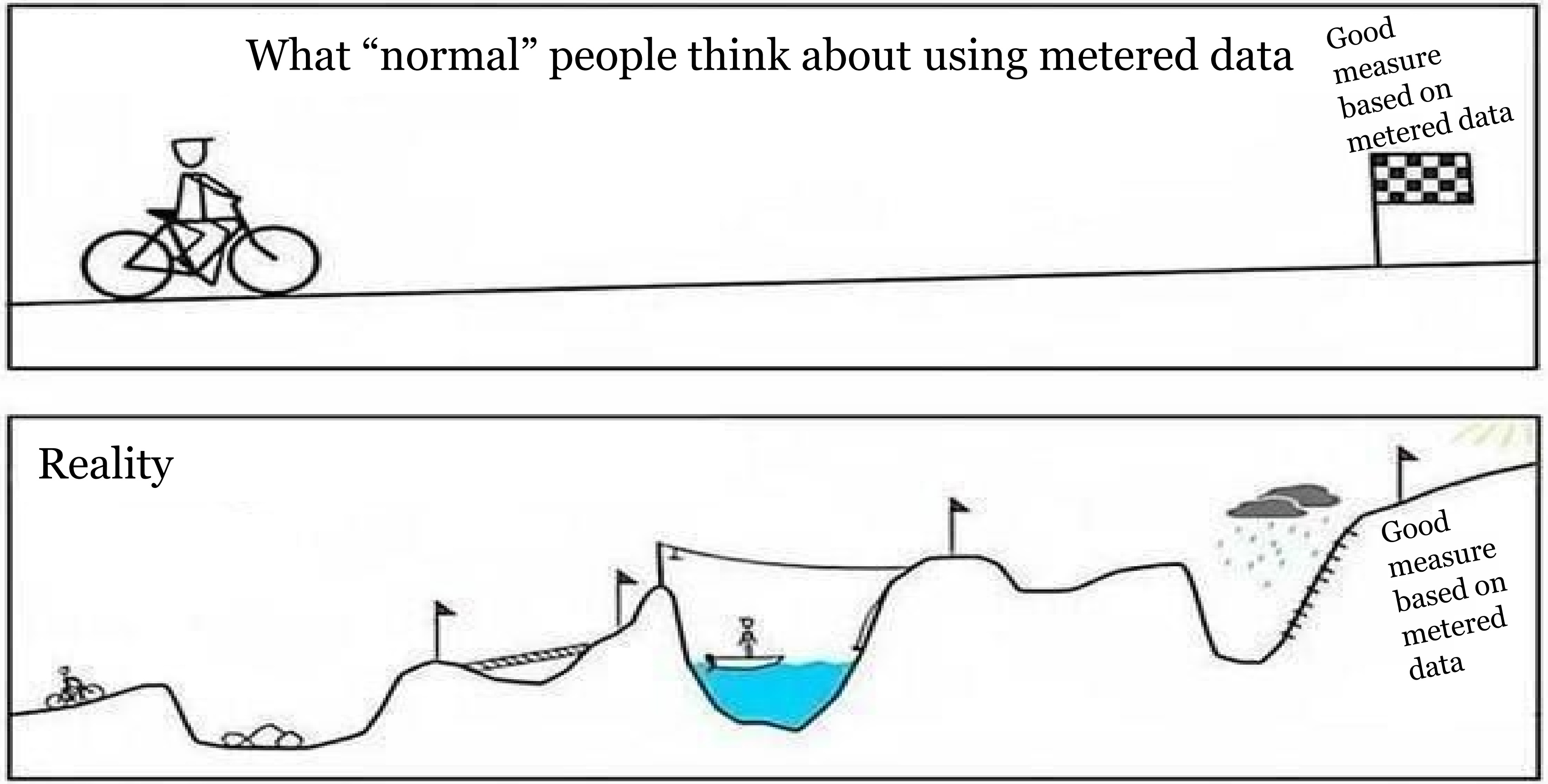
Published: 18 March 2009

Cite Permissions Share

Abstract

Many studies of media effects use self-reported news exposure as their key independent variable without establishing its validity. Motivated by anecdotal evidence that people's reports of their own media use can differ considerably from independent assessments, this study examines systematically the accuracy of survey-based self-reports of news exposure. I compare survey estimates to Nielsen estimates, which do not rely on self-reports. Results show severe overreporting of news exposure. Survey estimates of network news exposure follow trends in Nielsen ratings relatively well, but exaggerate

But this is not so obvious…



An error framework

- Metered data can suffer from lot 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

METERED DATA

An error framework

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

Meter not installed

Shared devices

Technology limitations

Extraction error

Size of the errors

- Next, we investigated how large some of these errors are and to what extent they may affect the estimates (Bosch et al., 2022)

- Focus on **tracking undercoverage**

  – Participants do not install the meter in all the devices/browsers they use

  – Data from the TRI-POL project[1] (Spain, Portugal, Italy): 3 survey waves + metered data 2 weeks before/after each survey

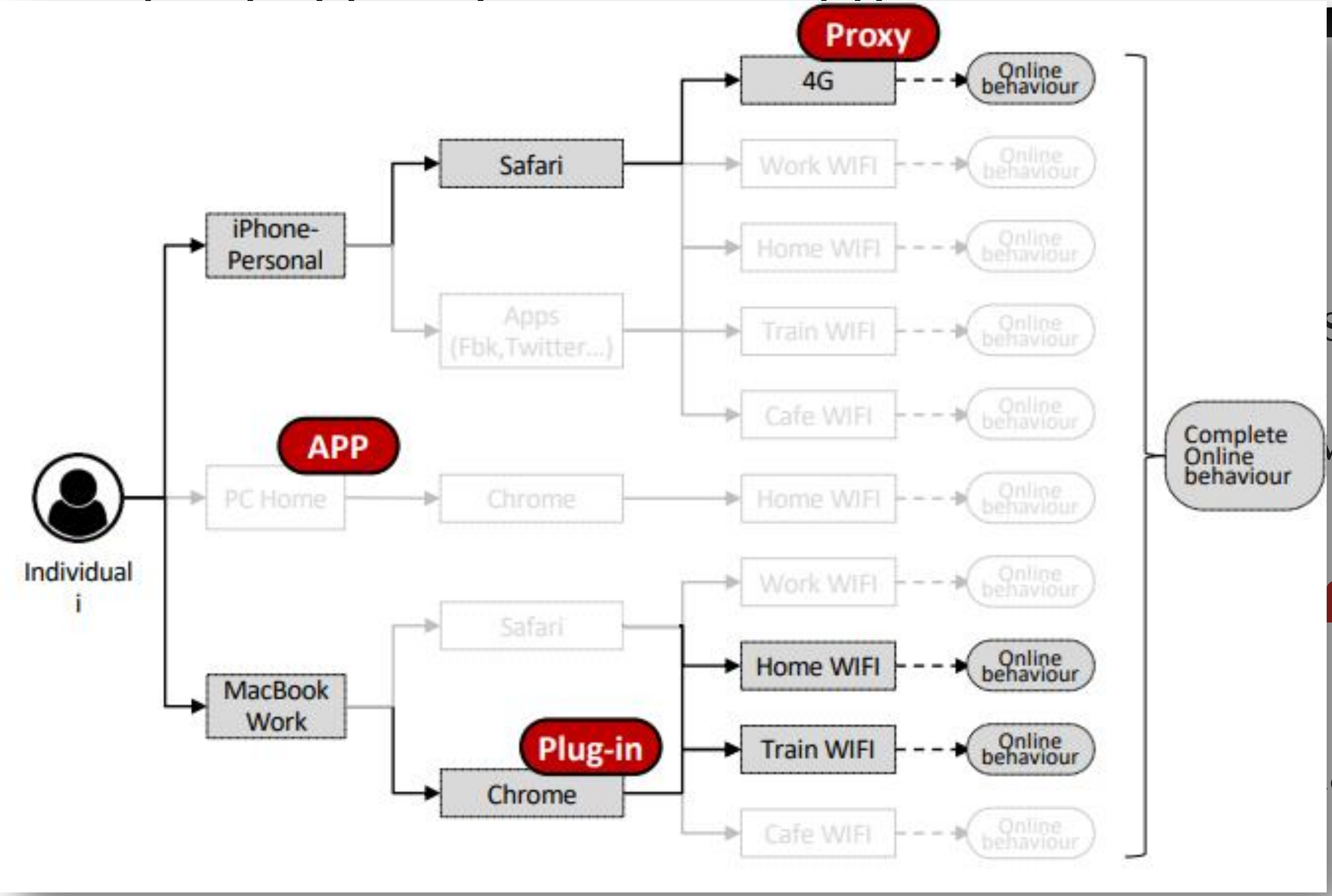
  – Combining survey+metered data, we found that **80-85% of participants are undercovered**

  – Using simulations, we found that tracking undercoverage biased both univariate and multivariate estimates

[1] https://www.upf.edu/web/tri-pol; see also Bosch & Revilla, 2022b

Size of the errors

- Next, we i[...] nd to what extent the[...]
- Focus on[...]
  - Partici[...] sers they use
  - Data fr[...] vey waves + meterec[...]
  - Combi[...] f participants **are ur[...]**
  - Using s[...] ased both univari[...]



[1] https://www.upf.edu/web/tri-pol; see also Bosch & Revilla, 2022b

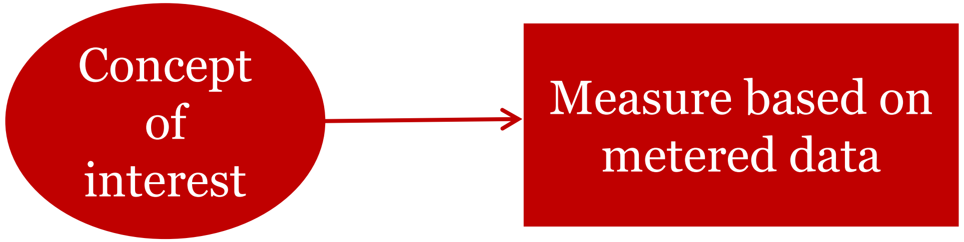
METERED DATA

Validity

- We also studied the **validity of measures** based on metered data depending how we operationalize the concepts of interest

- Focus on "**online news media exposure**" (Bosch & Revilla, 2022c)

- How to create a measure of online news media exposure using metered data?



- Many decisions

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

Validity

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

Validity

- How do these decisions affect the **convergent** and **predictive validity** of online news media exposure measured with metered data?
  - *Convergent validity*: if all variables were measuring the same concept, they should highly correlate with each other
  - *Predictive validity*: measures that correlate more with political knowledge are assumed to be better
- Data from TRI-POL
  - Average to low convergent validity
  - High fluctuations in predictive validity depending on the choices

METERED DATA

Summing up

Disadvantages

Researchers

More expensive

Dependence on private companies

Selection bias?

New types of errors

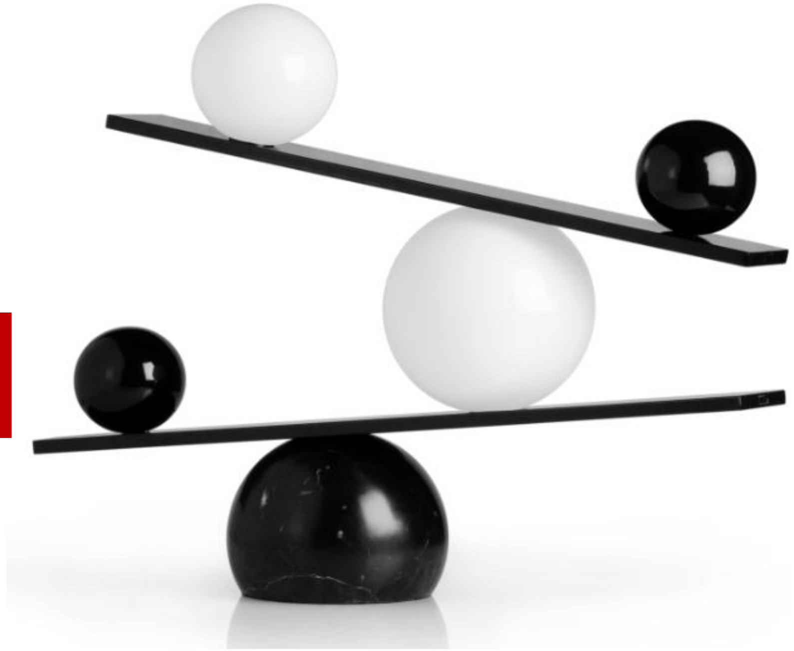
Data protection/ethical issues?

Participants

Privacy issues?

Loss of control?

New skills needed?



Benefits

Researchers

Massive amount of data

Continuous/real time

Reduce some of the issues related to measurement errors

Participants

Reduced time

Reduced effort

Conclusions

Still a lot to be done

- We have been working in different directions but still a lot to do!

- 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

- Better understand **how** to use those data
  - To replace? To combine? How?

Still a lot to be done

- More research needed
  - Both methodological research
  - And applications to key practical issues

- Potentially **broad applications** and **new insights**!
  - Across different disciplines

- But there will **always be errors**…

Do not conclude too much…

- Not realistic to aim to perfect measures
  - Try to minimize errors / correct for them → but still there will be errors

- So… what we can do?
  - Be aware of the errors, **acknowledge them** and think about **their consequences**
  - Look from different perspectives to get **different but complementary information**

CONCLUSIONS

Look from different perspectives

THE BLIND MEN AND THE ELEPHANT

"And so these men of research
Disputed loud and long,
Each in his own opinion
Exceeding stiff and strong,
Though each was partly in the right,
And all were in the wrong!"

John Godfrey Saxe (1816-1887)

References

- Bosch, O.J., & M. Revilla (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., & Revilla, M. (2022b). The challenges of using digital trace data to measure online behaviors: lessons from a study combining surveys and metered data to investigate affective polarization. In SAGE Research Methods Cases. https://dx.doi.org/10.4135/9781529603644.
- Bosch, O. J., Revilla, M. (2022c). 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.
- Bosch, O.J., Revilla, M., & E. Paura (2018). Do Millennials differ in terms of survey participation? *International Journal of Market Research,* 61(4): 359-365.
- Clemens, S., & Ginnis, S. (2017). Mobile-based geo-triggered surveys: Experiences from the field. *Paper presented at the CLOSER "New Technologies to Measure Non-Health Topics in Longitudinal Studies" workshop*. London, U.K.
- Festic, N., Büchi, M. & M. Latzer (2021). How Long and What For? Tracking a Nationally Representative Sample to Quantify Internet Use. *Journal of Quantitative Description: Digital Media* 1(2021), 1–23.
- Frankovic, K. A. (2012). Opinion Polls and the Media in the United States. In: Holtz-Bacha C., Strömbäck J. (eds) Opinion Polls and the Media. Palgrave Macmillan, London. https://doi.org/10.1057/9780230374959_6.
- Guess, A.M., Nyhan, B., & J. Reifler (2020). Exposure to untrustworthy websites in the 2016 US election. *Nature human behavior*, 4(5): 472-480.
- Kaur, S., Pandey, S., & Goel, S. (2019). Plants disease identification and classification through leaf images: A survey. *Archives of Computational Methods in Engineering*, *26*(2), 507–530

Thanks!

Questions?

Melanie Revilla | IBEI

mrevilla@ibei.org

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