INSTITUT
BARCELONA
ESTUDIS
INTERNACIONALS

European Research Council
Established by the European Commission

web
data
opp

Complementing conventional web survey data with new measurement opportunities to achieve better or new insights

Córdoba, 30 March 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.

web
data
opp

Which new opportunities?

Growing use of (mobile) 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

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

83% of the world population have smartphones[2]

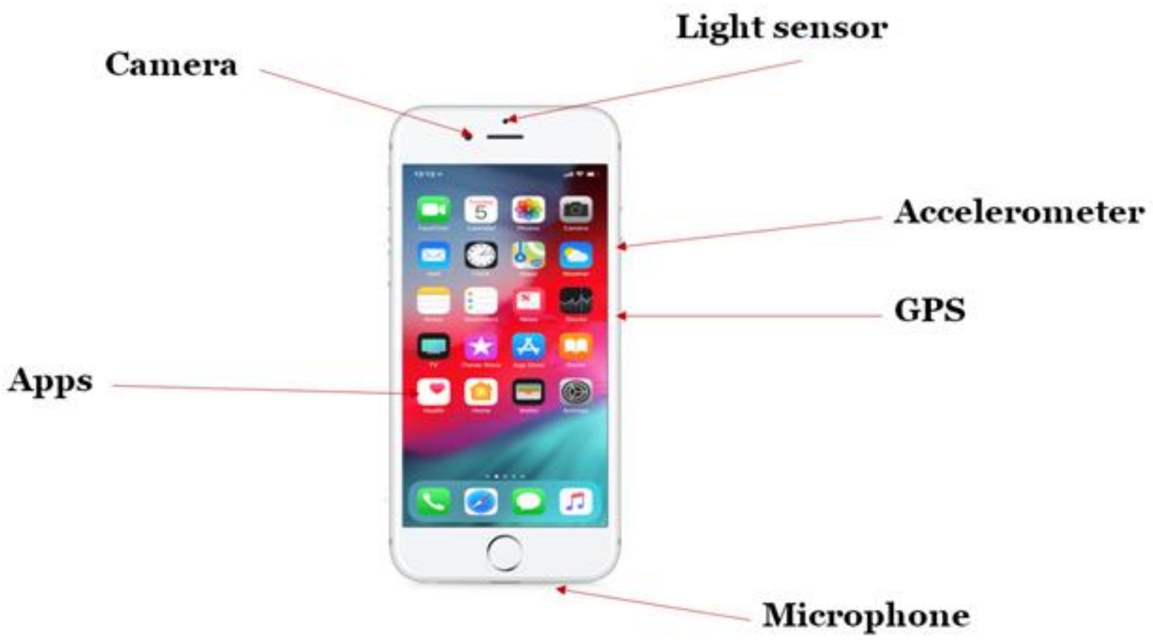
92% of Internet users worldwide access the Internet through smartphones[1]

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

WHICH NEW OPPORTUNITIES

Possible to collect many different types of data

- 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



New data types considered

VISUAL DATA

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

VOICE DATA

Dictation
Voice recording

Most of those data can also be collected for PCs

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. Usually collected in metered panels.

GEOLOCATION DATA

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

IN-THE-MOMENT SURVEYS triggered by such data

How could they help?

HOW COULD THE NEW DATA TYPES HELP?

Main expected benefits (Revilla, 2022)

Researchers

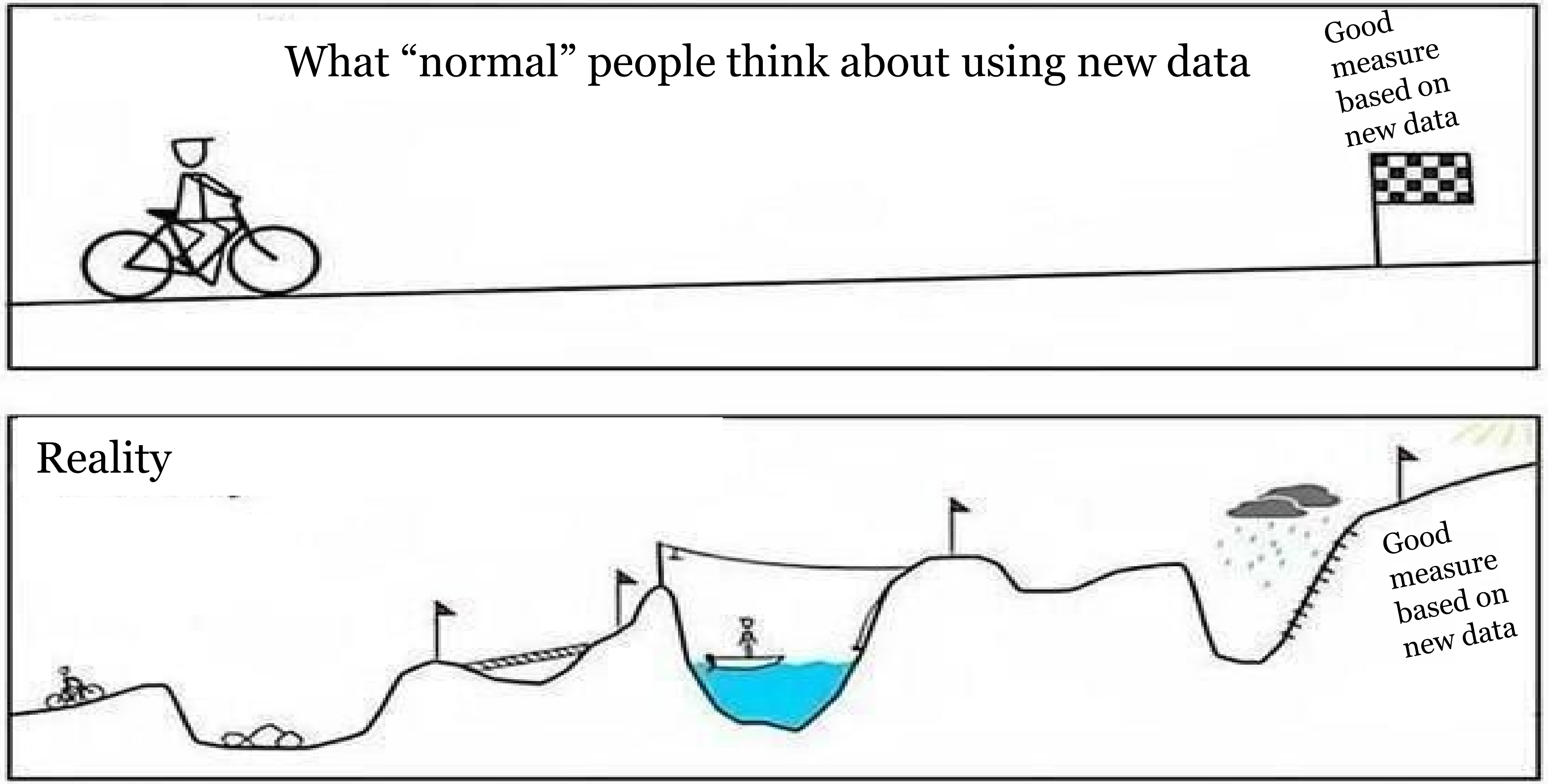
- Reduce some of the issues related to measurement errors
- Massive amount of data
- Real time / continuous (passive data)
- Provide data for new concepts (not measured so far)
- Answer new research questions

Participants

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

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

But this is not that easy…


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

Our goal = get more knowledge that will help better use such data

Example 1

Studying migrants' changes in housing conditions:

how could we use visual data for this?

THIS IS NOT THAT EASY: EXAMPLE VISUAL DATA

Example 1: Studying changes in migrants’ housing conditions

Examples of research questions that could be answered with visual data

- How migrants’ housing conditions change after migrating?
  - For which kind of migrants do they improve?
  - For which kind of migrants do they get worse?
  - For which kind of migrants do they stay the same?

- Which aspects are the ones that change most within the housing conditions?
  - Comfort?
  - Size?

Example 1: Studying changes in migrants’ housing conditions

What could we do to answer such research questions?

Step 1

Identify migrants who recently moved to a new country

– Could be done through surveys

Step 2

Collect information about housing conditions → Ask them to share photos

– Of the place where they lived just before migrating (already saved photos)
– Of the place where they are living now (photos taken during the survey)
– Specific instructions depending on exact aspects of interest within the housing conditions

Step 3

Extract the information from the photos to answer the research questions

Example 1: Studying changes in migrants’ housing conditions

What could we do to answer such research questions?

Step 1

Identify migrants who recently moved to a new country

– Could be done through surveys

Step 2

Collect information about housing conditions → Ask them to share photos

– Of the place where they lived just before migrating (already saved photos)
– Of the place where they are living now (photos taken during the survey)
– Specific instructions depending on exact aspects of interest within the housing conditions

Step 3

Extract the information from the photos to answer the research questions

THIS IS NOT THAT EASY: EXAMPLE VISUAL DATA

Example 1: Studying changes in migrants’ housing conditions

What could we do to answer such research questions?

Step 1

Identify migrants who recently moved to a new country

- Could be done through surveys

Step 2

Collect information about housing conditions → Ask them to share photos

- Of the place where they lived just before migrating (already saved photos)
- Of the place where they are living now (photos taken during the survey)
- Specific instructions depending on exact aspects of interest within the housing conditions

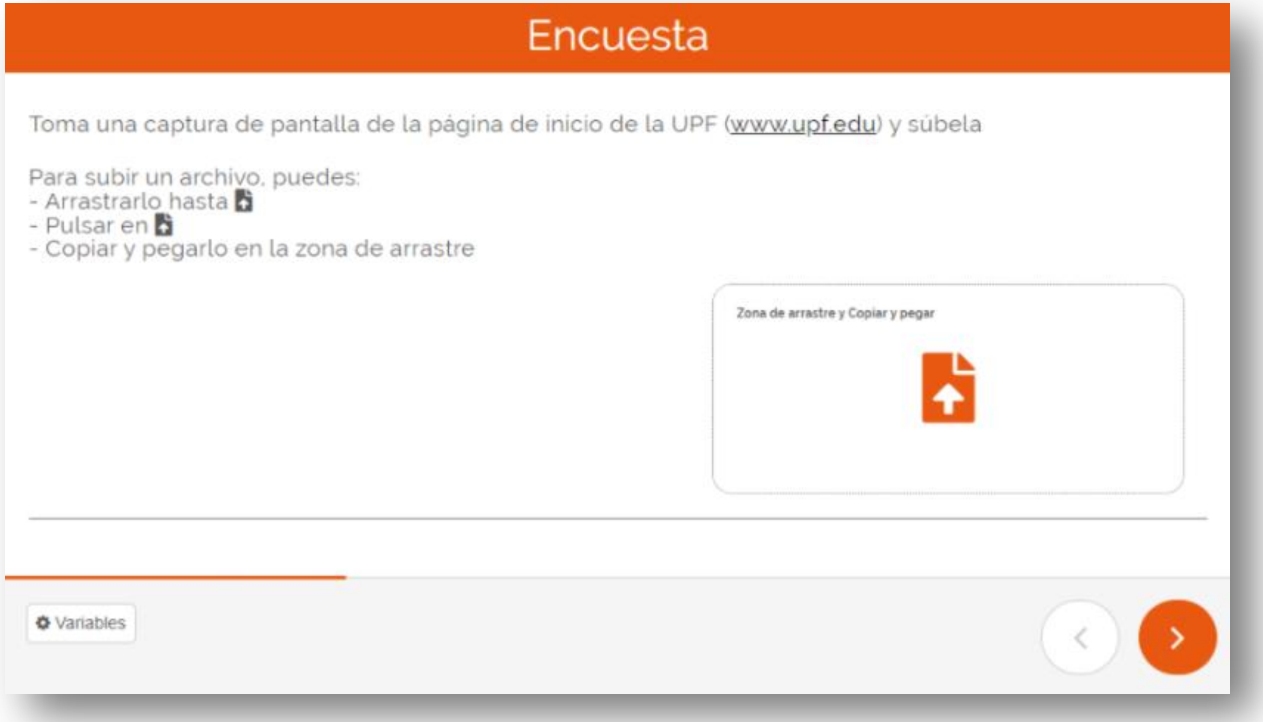
Step 3

Extract the information from the photos to answer the research questions

THIS IS NOT THAT EASY: EXAMPLE VISUAL DATA

Problem 1: Specific tool needed to collect visual data

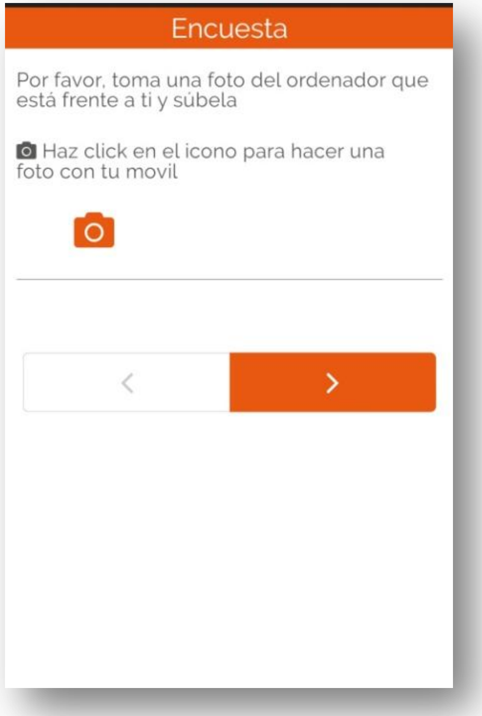
- We developed **WebdataVisual** (Revilla et al., 2022)
  - Can collect visual data already saved on the device
  - Or produced during the survey (through camera or screenshots)



Encuesta

Toma una captura de pantalla de la página de inicio de la UPF (www.upf.edu) y súbela

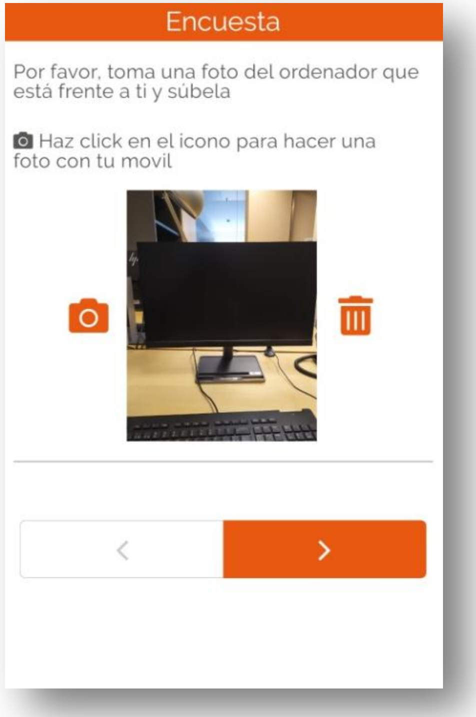
Para subir un archivo, puedes:
- Arrastrarlo hasta
- Pulsar en
- Copiar y pegarlo en la zona de arrastre



Encuesta

Por favor, toma una foto del ordenador que está frente a ti y súbela

Haz click en el icono para hacer una foto con tu movil



Encuesta

Por favor, toma una foto del ordenador que está frente a ti y súbela

Haz click en el icono para hacer una foto con tu movil

More information at: https://www.upf.edu/web/webdataopp/tools

THIS IS NOT THAT EASY: EXAMPLE VISUAL DATA

Problem 2: Respondents should send the images…

- Previous research: **≈ 50%** of the respondents **share images** when asked to do so in web surveys (Bosch et al., 2018)

- Why the others do not?
  - To disentangle the mechanisms behind this non-response
  - We asked about the **skills** + **availability** + **willingness** + **burden** (Iglesias & Revilla, in press)
  - Considering PCs + smartphones & videos + images
  - Data from an opt-in online panel in Spain

➢**Availability** seems to be the most limiting factor for participation

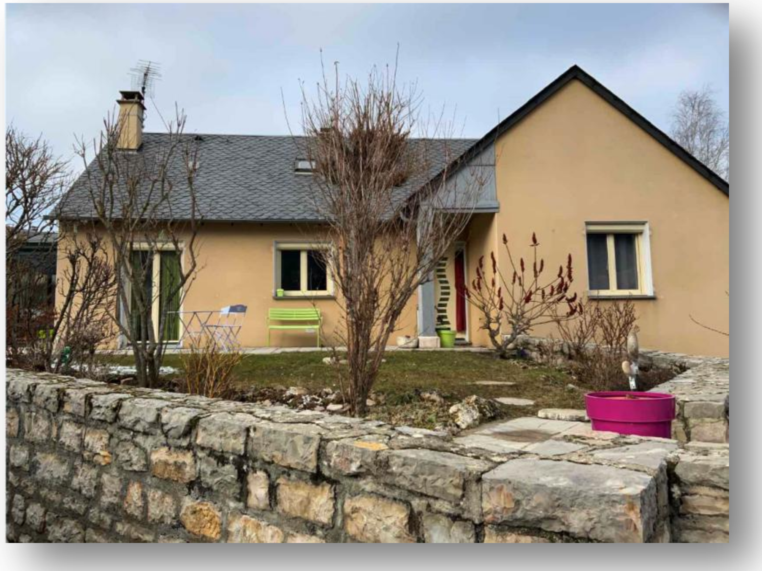
THIS IS NOT THAT EASY: EXAMPLE VISUAL DATA

Problem 3: Extract relevant information from the images

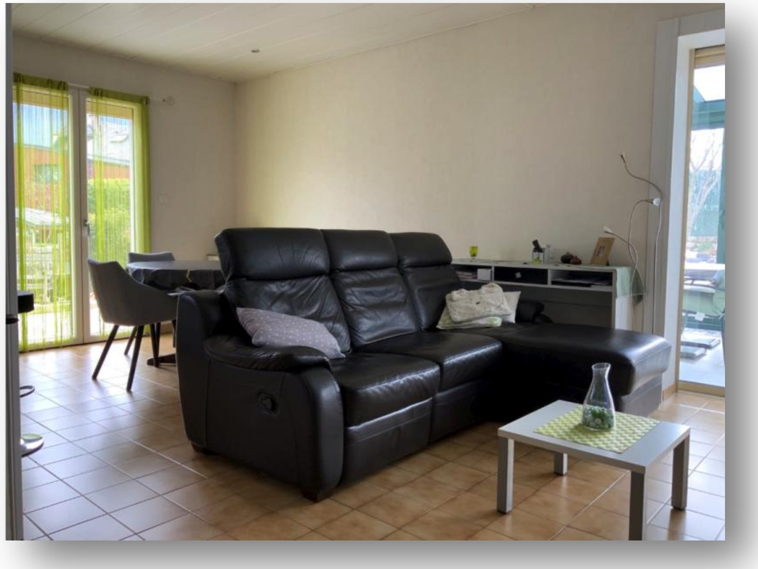
- Extracting information from images = process of "**classification**"
- Quality of the data obtained with images depends on classification
- Key problems:
  - Define properly what we should extract, and which labels we should use
  - Choose the best classification method
    - Can be done manually or automatically (machine learning algorithms)
    - Lot of aspects to balance (features of the tasks, resources available, data quality)
- Practical guide to help researchers interested in using images with these issues (Iglesias et al., 2022)

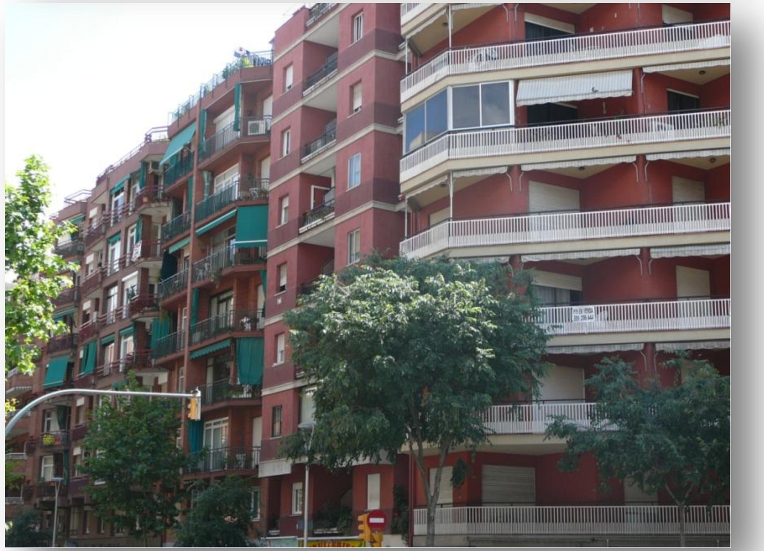
THIS IS NOT THAT EASY: EXAMPLE VISUAL DATA

Problem 3: Extract relevant information from the images

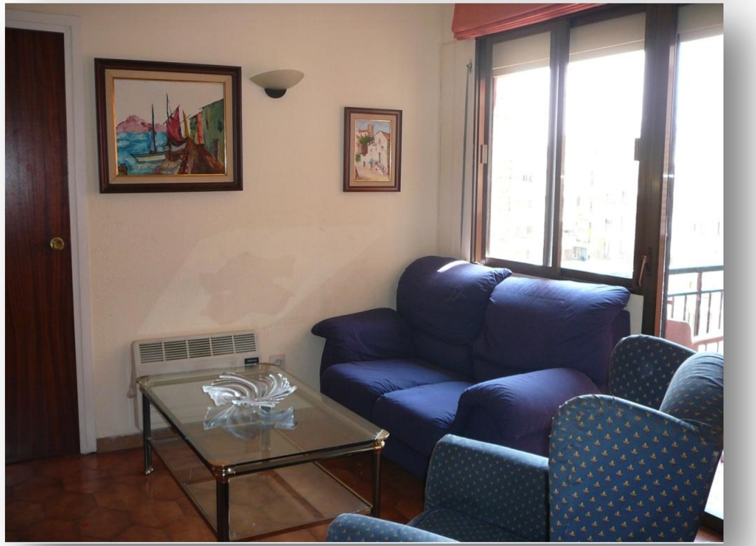


Country of origin





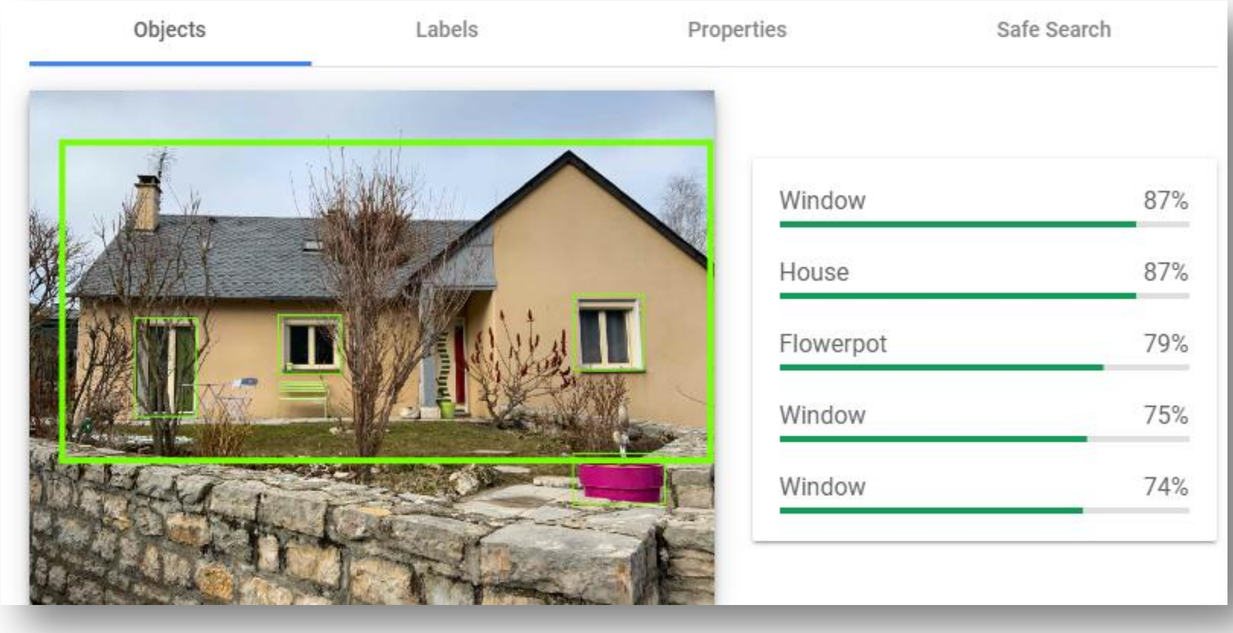
New country



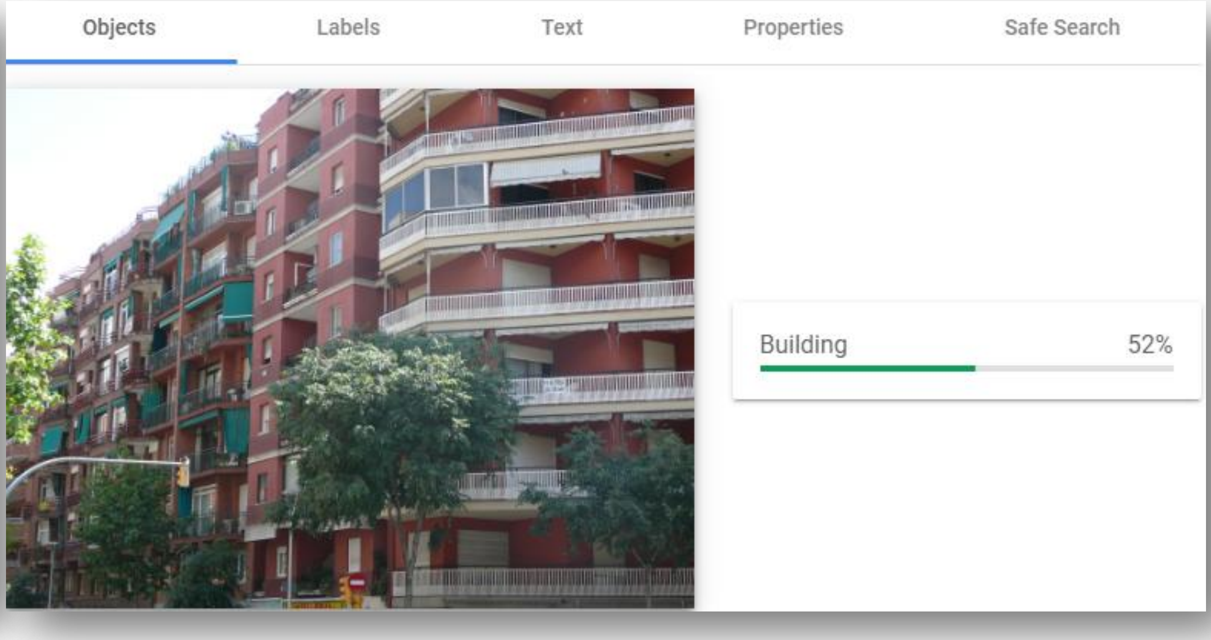
THIS IS NOT THAT EASY: EXAMPLE VISUAL DATA

Problem 3: Extract relevant information from the images

Results from Google Vision API, asking to classify “objects”



Window	87%
House	87%
Flowerpot	79%
Window	75%
Window	74%



Building	52%

More problems

- **Selection** bias?
  - Individuals who send visual data in web surveys: ≠ those who do not? ≠ target population?
  - Depends a lot on the target population
- Data protection and **ethical** issues
  - How to make sure that the consent is really informed?
  - Images might contain personal data → how to deal with such data?
- **Loss of control** for the participants?

Example 2

Studying migrants' online news media exposure:

How could we use metered data for this?

Example 2: Migrants' online news media exposure

Examples of research questions that could be answered using metered data

- Which kind of online news do migrants consume?
  - Do the news mainly come from the country of origin?
  - Or from the country they are now living in?
  - Or still from other countries?

- Which factors influence the kind of news consumed?
  - Are migrants from some specific origins more prone to only read news from their country of origin?
  - Does it depend on the number of years they have been in the new country?
  - Does it depend on migrants' levels of education?
  - Are there differences between men and women?

THIS IS NOT THAT EASY: EXAMPLE METERED DATA

Example 2: Migrants' online news media exposure

What could we do to answer such research questions?

Step 1

Identify migrants

- Could be done through survey
- Could be done using metered data

Step 2

Measure their online news media exposure → metered data

- Distinguishing the news depending on the country publishing them

Step 3

Measure other factors of interest

- Education, gender, number of years in country, etc.
- Could be done through survey

THIS IS NOT THAT EASY: EXAMPLE METERED DATA

Example 2: Migrants’ online news media exposure

web data opp

What could we do to answer such research questions?

Step 1

Identify migrants

- Could be done through survey
- Could be done using metered data

Step 2

Measure their online news media exposure → metered data

- Distinguishing the news depending on the country publishing them

Step 3

Measure other factors of interest

- Education, gender, number of years in country, etc.
- Could be done through survey

THIS IS NOT THAT EASY: EXAMPLE METERED DATA

Example 2: Migrants' online news media exposure

What could we do to answer such research questions?

Step 1

Identify migrants

- Could be done through survey
- Could be done using metered data

Step 2

Measure their online news media exposure → metered data

- Distinguishing the news depending on the country publishing them

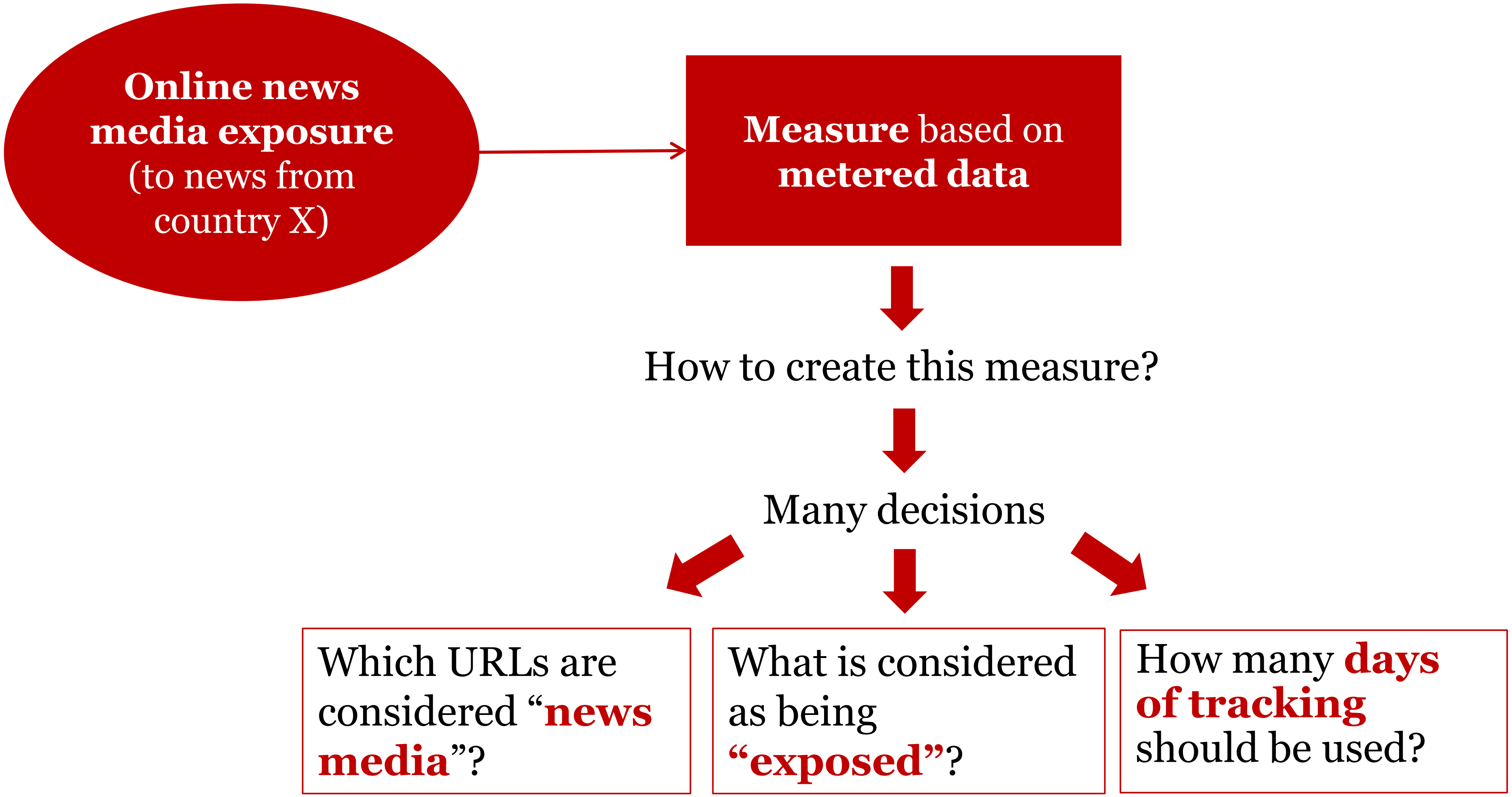
Step 3

Measure other factors of interest

- Education, gender, number of years in country, etc.
- Could be done through survey

THIS IS NOT THAT EASY: EXAMPLE METERED DATA

Problem 1: Operationalizing the concept of interest



Online news media exposure (to news from country X)

Measure based on **metered data**

How to create this measure?

Many decisions

Which URLs are considered "**news media**"?

What is considered as being **"exposed"**?

How many **days of tracking** should be used?

THIS IS NOT THAT EASY: EXAMPLE METERED DATA

Problem 1: Operationalizing the concept of interest

- We studied different ways to operationalize the concept "*online news media exposure*" using metered data (Bosch & Revilla, 2022a)
  - No focus on migrants in our case
  - Only consider written + national news (i.e., news from the country of living)

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

Combining all these decisions

↓

We could create **>8,000** variables that should all measure *"online news media exposure"*

THIS IS NOT THAT EASY: EXAMPLE METERED DATA

Problem 1: Operationalizing the concept of interest

- We studied how these decisions affect the **validity** of the measures (Bosch & Revilla, 2022a)
  - **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[1]

Netquest metered panels in Spain, Portugal, Italy

3 survey waves + metered data 2 weeks before/after each survey

Main results

Low to average convergent validity

High fluctuations in predictive validity depending on the choices

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

THIS IS NOT THAT EASY: EXAMPLE METERED DATA

Problem 2: Identifying all possible types of errors

- Metered data considered as the gold standard in several studies
- But metered data can suffer from different types of errors
- It is crucial to:

1. **Identify** the potential errors
2. Estimate their **size**
3. Find ways to **minimize** them
4. And/or to **correct** for them

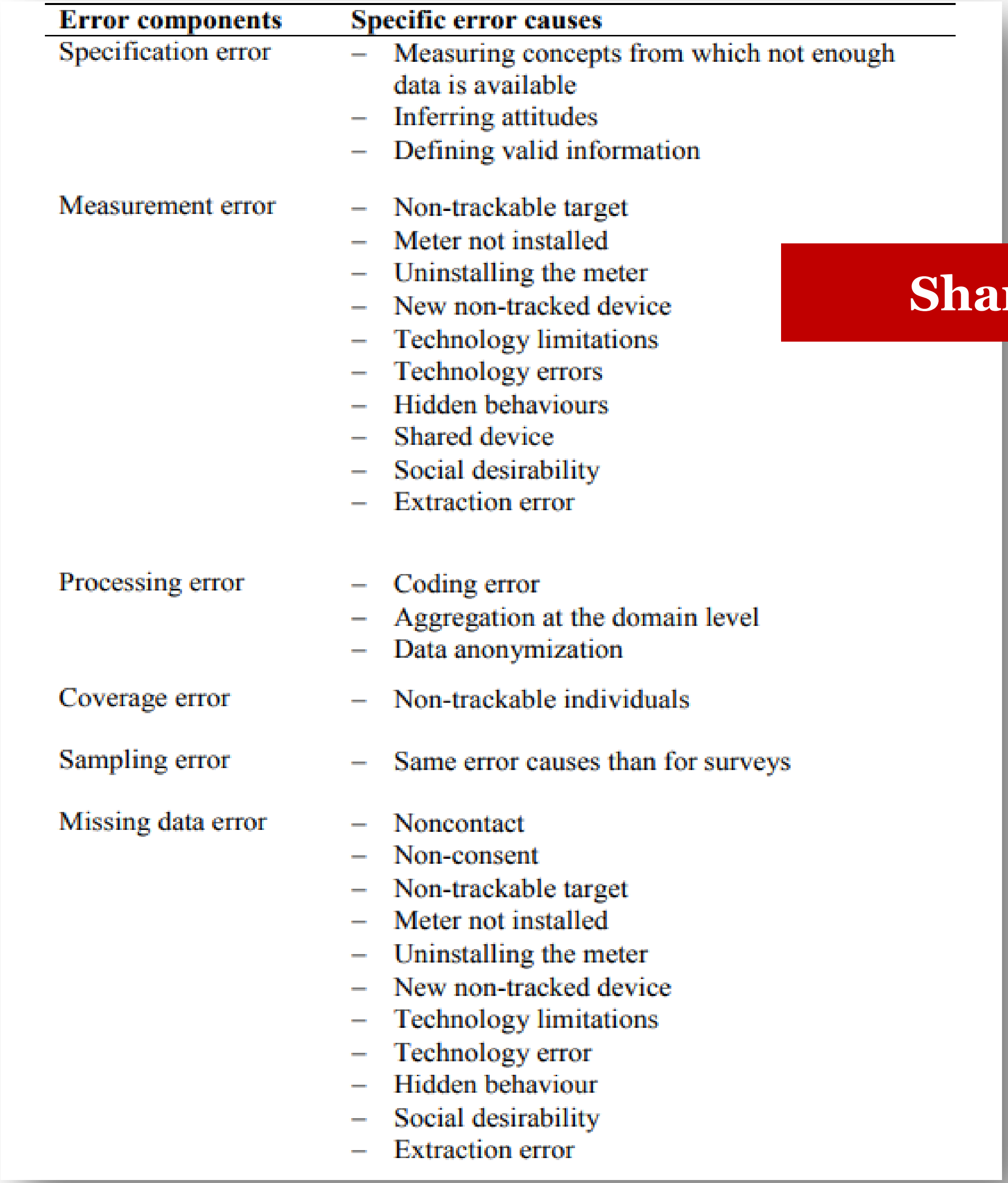
TEM = **Total error framework for metered data** (Bosch & Revilla, 2022b)

Adaptation of the TSE

Overview of all possible errors and their causes

THIS IS NOT THAT EASY: EXAMPLE METERED DATA

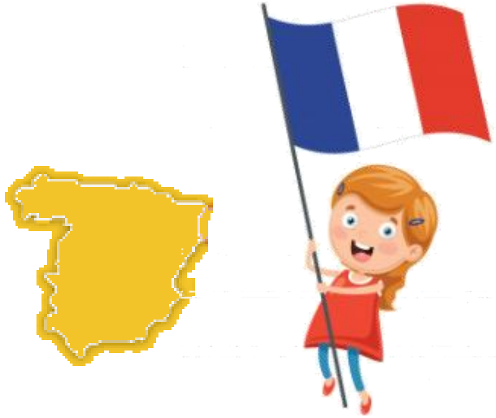
Problem 2: Identifying all possible types of errors



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

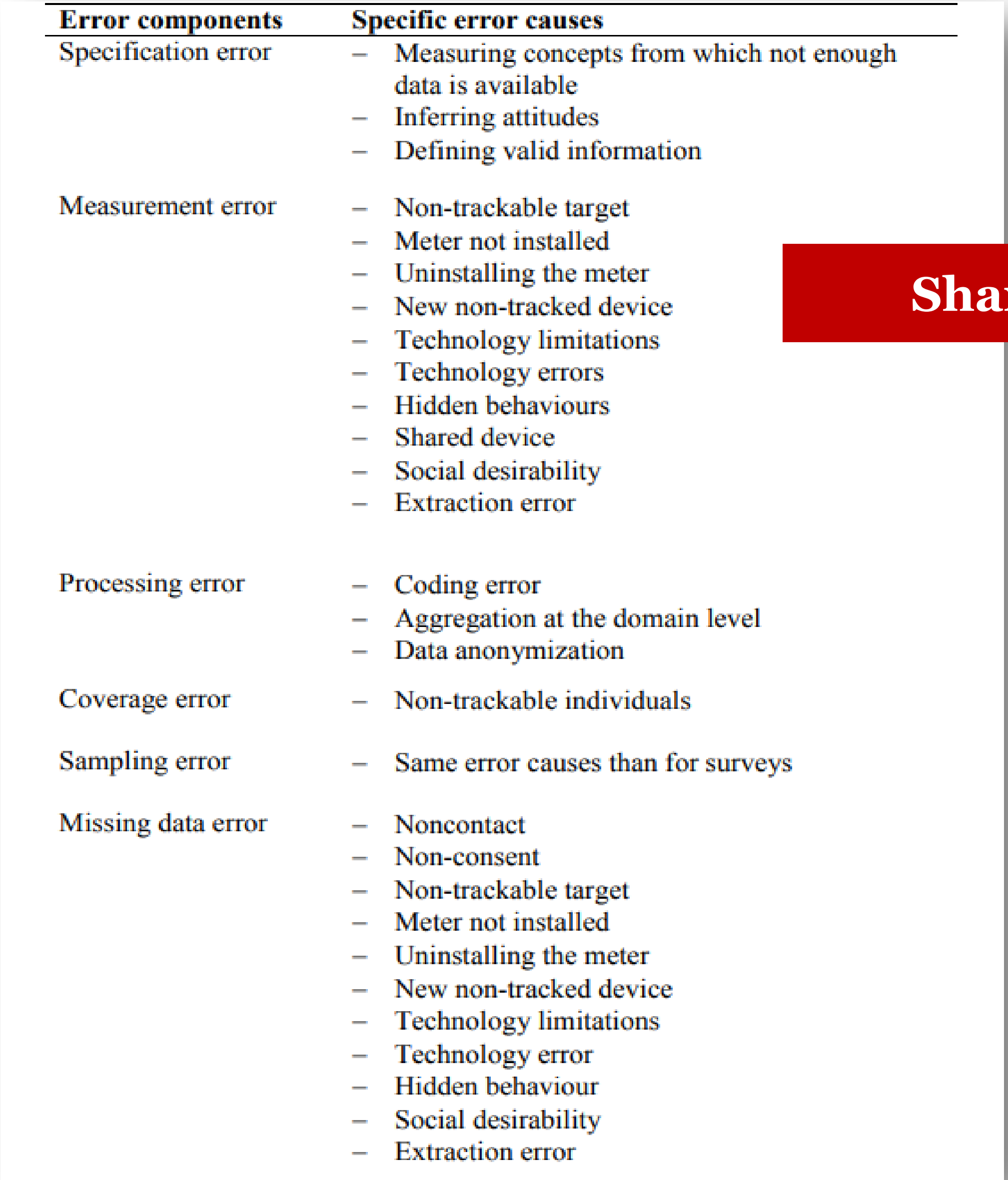
Shared devices

Le Monde
La Russie fait état, pour la première fois, de combats sur son sol



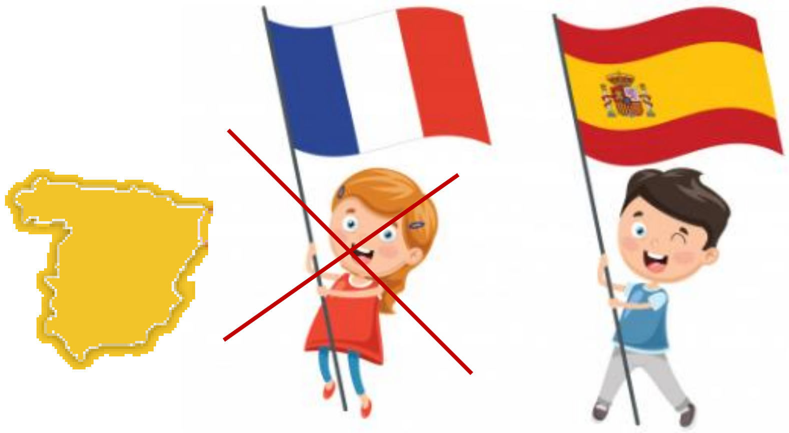
THIS IS NOT THAT EASY: EXAMPLE METERED DATA

Problem 2: Identifying all possible types of errors



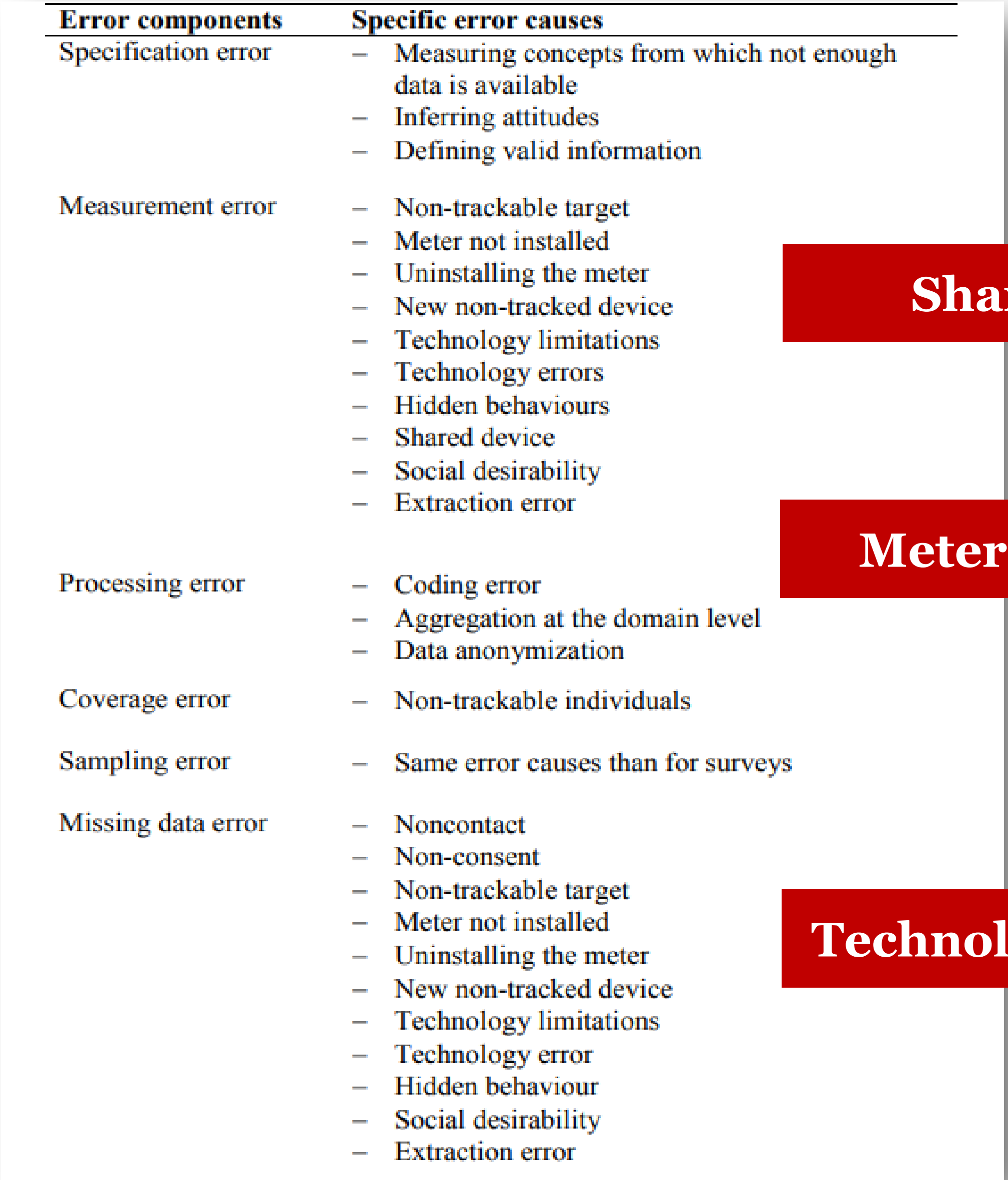
Error components	Specific error causes
Specification error	– Measuring concepts from which not enough data is available
	– Inferring attitudes
	– Defining valid information
Measurement error	– Non-trackable target
	– Meter not installed
	– Uninstalling the meter
	– New non-tracked device
	– Technology 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



THIS IS NOT THAT EASY: EXAMPLE METERED DATA

Problem 2: Identifying all possible types of errors



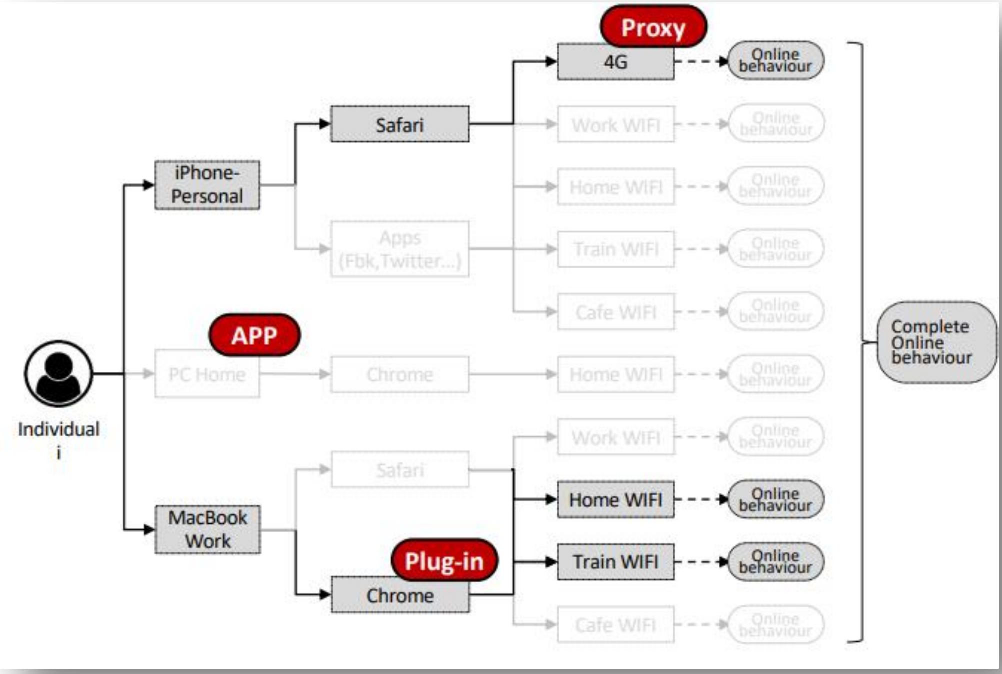
Error components	Specific error causes
Specification error	– Measuring concepts from which not enough data is available – Inferring attitudes – Defining valid information
Measurement error	– Non-trackable target – Meter not installed – Uninstalling the meter – New non-tracked device – Technology limitations – 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

Meter not installed

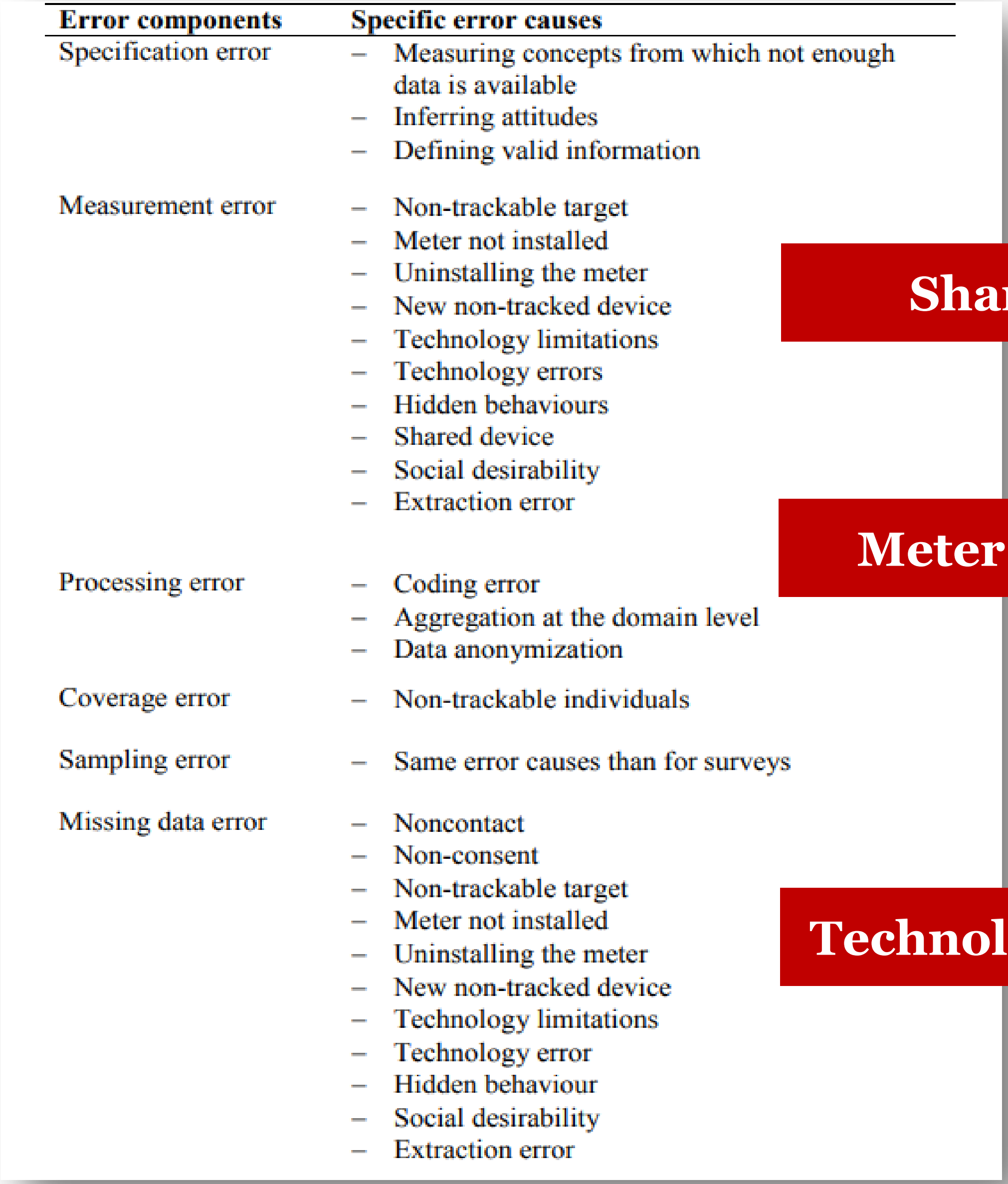
Technology limitations

Tracking undercoverage affected **80-85%** of the participants in the TRI-POL data (Bosch et al., 2022)



THIS IS NOT THAT EASY: EXAMPLE METERED DATA

Problem 2: Identifying all possible types of errors

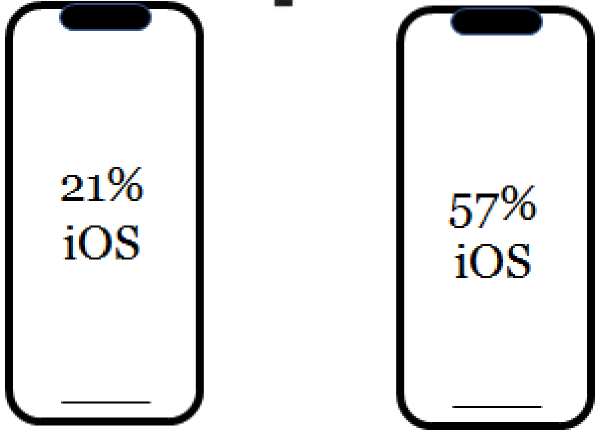


Error components	Specific error causes
Specification error	– Measuring concepts from which not enough data is available – Inferring attitudes – Defining valid information
Measurement error	– Non-trackable target – Meter not installed – Uninstalling the meter – New non-tracked device – Technology limitations – 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

Meter not installed

Technology limitations



https://gs.statcounter.com/os-market-share/mobile/spain; https://gs.statcounter.com/os-market-share/mobile/united-states-of-america

THIS IS NOT THAT EASY: EXAMPLE METERED DATA

More problems

- **Selection bias?**
  - Metered panelists: ≠ non-metered panelists? ≠ target population?
  - Depends a lot on the target population

- Data protection and **ethical** issues
  - How to make sure that the consent is really informed?
  - URLs might contain personal data → need to find ways to pseudonymize

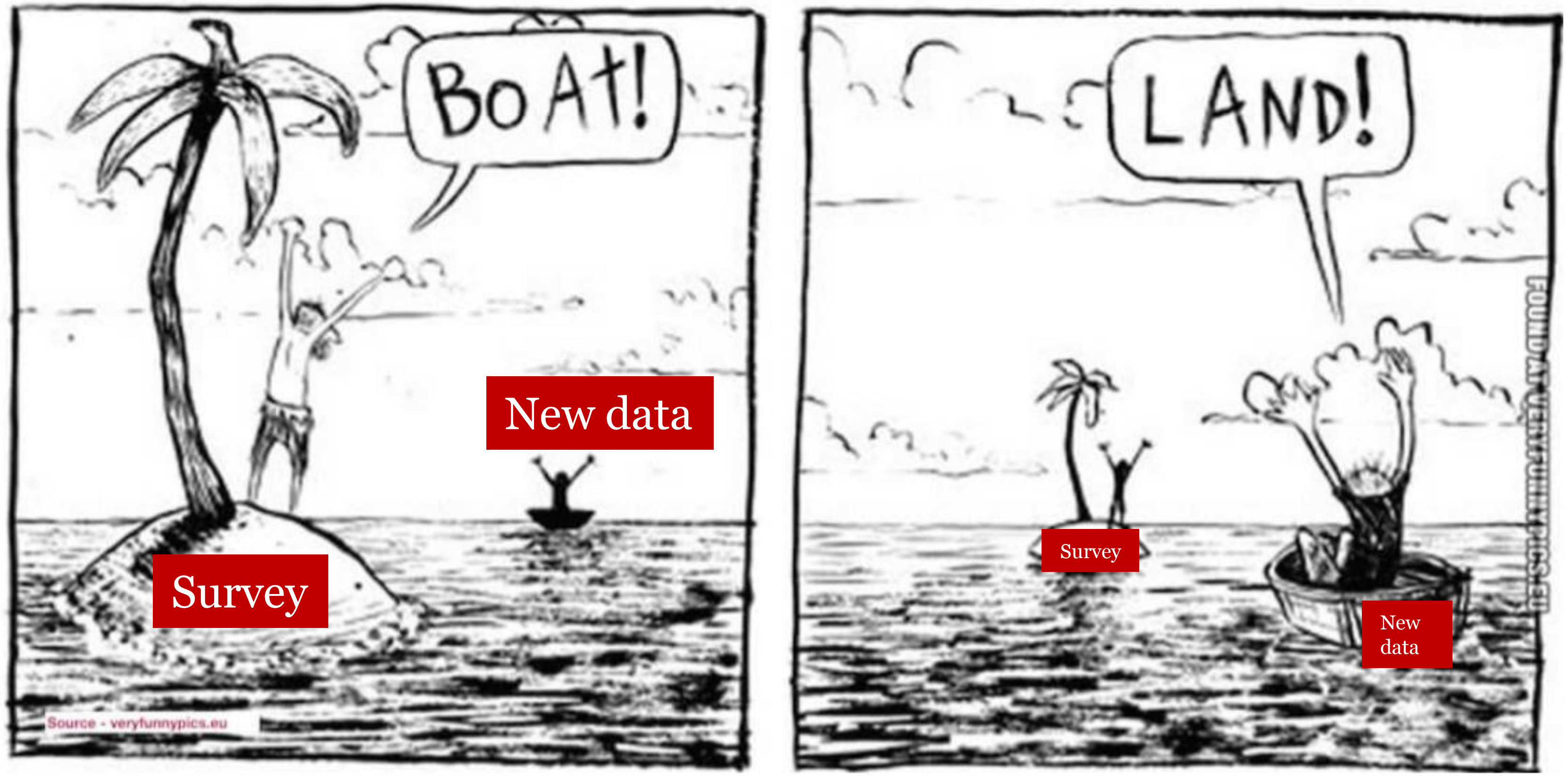
- **Dependence** on private companies

- More **expensive**

Conclusions

CONCLUSIONS

We are not saved yet...



Source - veryfunnypics.eu

CONCLUSIONS

Still a lot to be done…

More research needed for all new 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

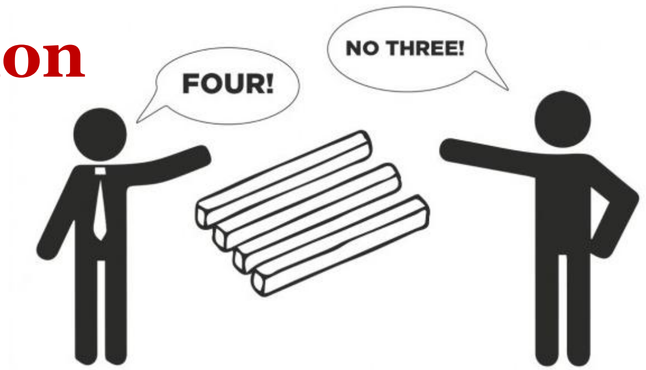
CONCLUSIONS

Still a lot to be done…

More research needed for all new 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., & E. Paura (2018). Answering mobile surveys with images: an exploration using a computer vision API. *Social Science Computer Review*, 37(5): 669-683. https://doi.org/10.1177/0894439318791515
- Bosch, O.J., & Revilla, M. (2022a). 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., & Revilla, M. (2022b). 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., Sturgis, P., & Kuha, J. (2022). Track me but not really: Tracking undercoverage in metered data collection. AAPOR Annual Conference, 11th-13th May 2022.
- Iglesias, P., & Revilla., M. (in press). Skills, availability, willingness, expected participation and burden of sharing visual data within the frame of web surveys. *Quality and Quantity*.
- Iglesias, P., Ochoa, C., and Revilla, M. (2022). What do I do with these images? A practical guide to the classification of images sent by survey participants, RECSM Webinar, 31st March 2022. Available at: https://www.youtube.com/watch?v=IQ0KbO4XsZI
- Revilla, M. (2022). How to enhance web survey data using metered, geolocation, visual andvoice data?. *Survey Research Methods*, 16(1): 1-12. https://doi.org/10.18148/srm/2022.v16i1.8013.
- Revilla, M., Iglesias, P., Ochoa, C., & Antón, D. (2022). WebdataVisual: a tool to gathervisual data within the frame of web surveys. OSF. http://doi.org/10.17605/OSF.IO/R7CAX