gesis Leibniz Institute for the Social Sciences

Potentials and Pitfalls of Social Media Data

Indira Sen & Katrin Weller

Leibniz Association

gesis Leibniz Institute for the Social Sciences

4: Potential Pitfalls of Social Media Data

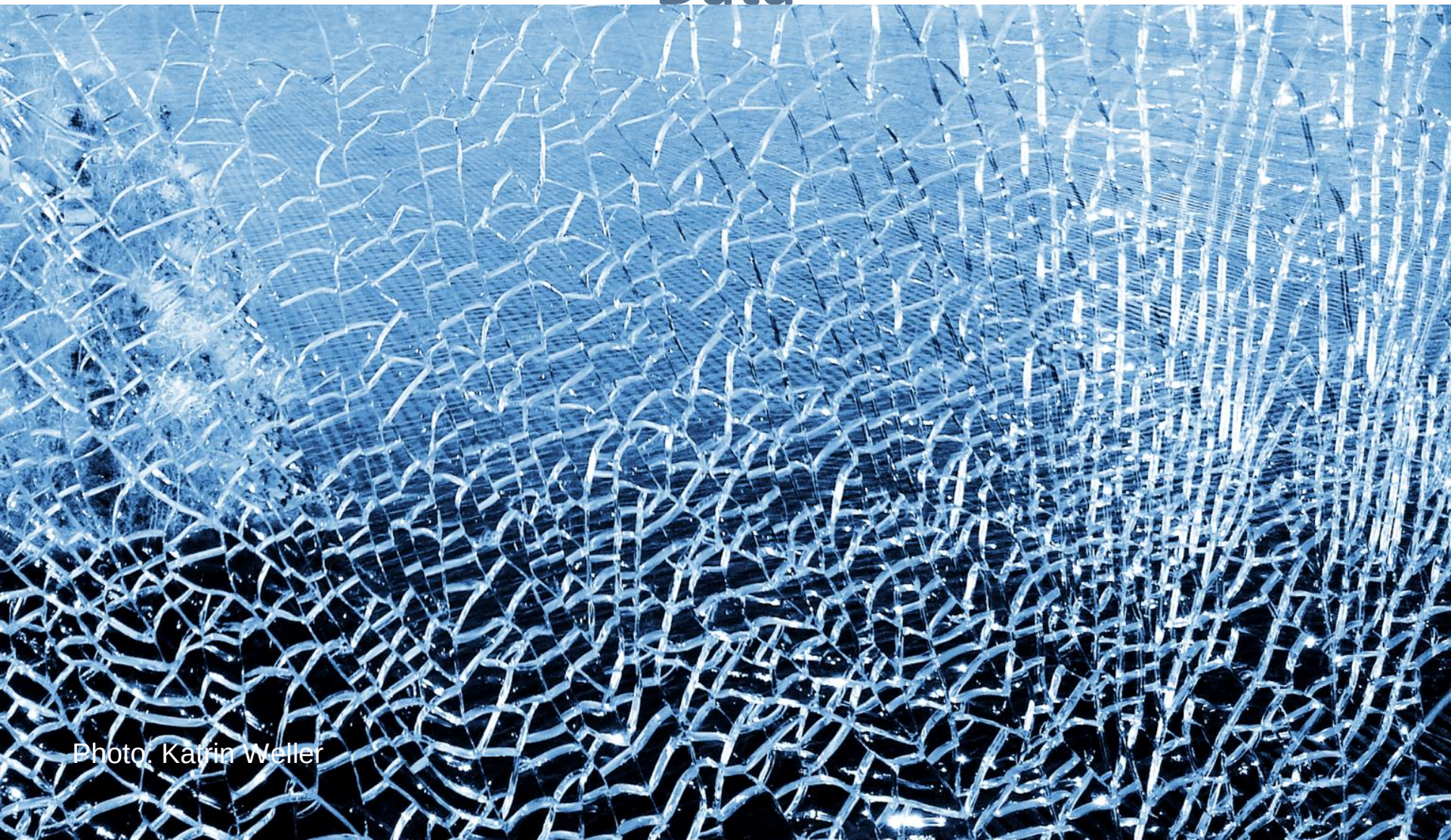


Photo: Katrin Weller

Potential Pitfalls in Studying Social Media Data

- Diverse levels of pitfalls
- Practical challenges induced by the social media landscape, data accessibility, nature of the data (found, not designed), lack of standard methods, ethical conflicts, …
- This has already lead to critique in the past

Critical concerns and practical challenges including data access

- Boyd, Danah, & Crawford, K. (2012). Critical questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon.
- Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013). Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose.
- Bruns, A. (2019). After the 'APIcalypse': Social media platforms and their fight against critical scholarly research.

End of Theory?

"The core challenge is that most big data that have received popular attention are not the output of instruments designed to produce valid and reliable data amenable for scientific analysis."

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google Flu: Traps in big data analysis. Science, 343(6176), 1203-1205.

End of Theory?

"The interesting point is that these limitations can (and have to) be addressed by theory guided research that is typically conducted by social scientists. Accordingly, opportunities emerge for those social and behavioral scientists who are willing to collaborate with the Big Data researchers in the natural, engineering, and computer sciences."

Snijders, C., Matzat, U., & Reips, U.-D. (2012). 'Big Data': Big gaps of knowledge in the field of Internet. International Journal of Internet Science, 7, 1-5. Retrieved from http://www.ijis.net/ijis7_1/ijis7_1_editorial.html

Inferences from Digital Traces: There are *certainly* challenges

- Issues with Twitter data
  - [Tufekci] on the representative and methodological issues of using Twitter
  - Total Twitter Error [Hsieh and Murphy]

- Social Data Biases and Pitfalls [Olteanu et al]
  - Biases and errors linked to each stage of the data handling pipeline
  - Normalizing Digital Trace Data [Jungherr]

Examples

Predicting elections with twitter: What 140 characters reveal about political sentiment, Tumasjan et al., 2010

Detecting influenza epidemics using search engine query data. Ginsberg et al., 2009

gesis Leibniz Institute for the Social Sciences

But also pitfalls

Predicting elections with twitter: What 140 characters reveal about political sentiment,
Tumasjan et al., 2010

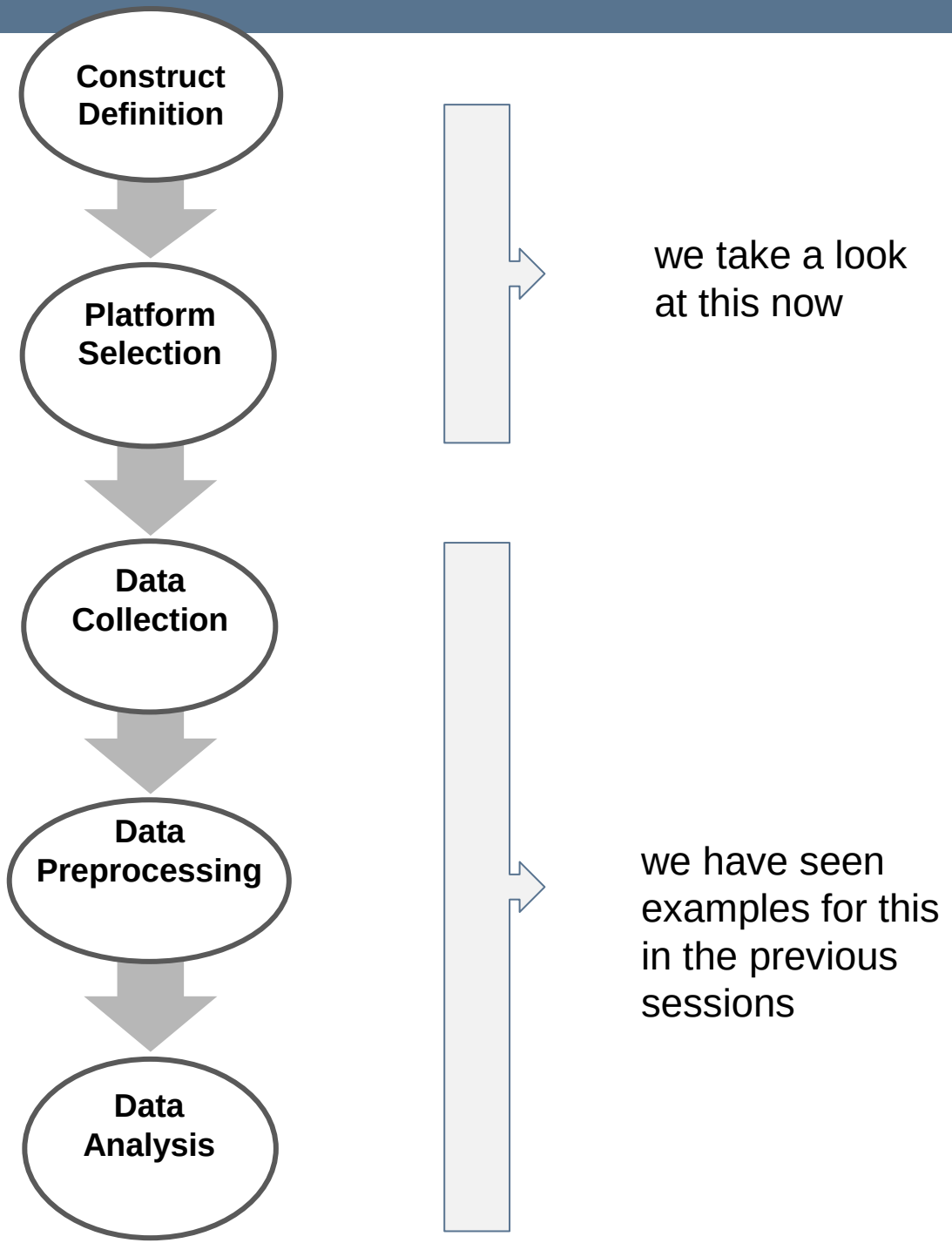
Why the pirate party won the german election of 2009 or the trouble with predictions: A response to Tumasjan et al. […]”.
Jungherr et al., 2012

Detecting influenza epidemics using search engine query data
Ginsberg et al., 2009

The parable of Google Flu: traps in big data analysis,
Lazer et al., 2014

prototypical workflow in social media research

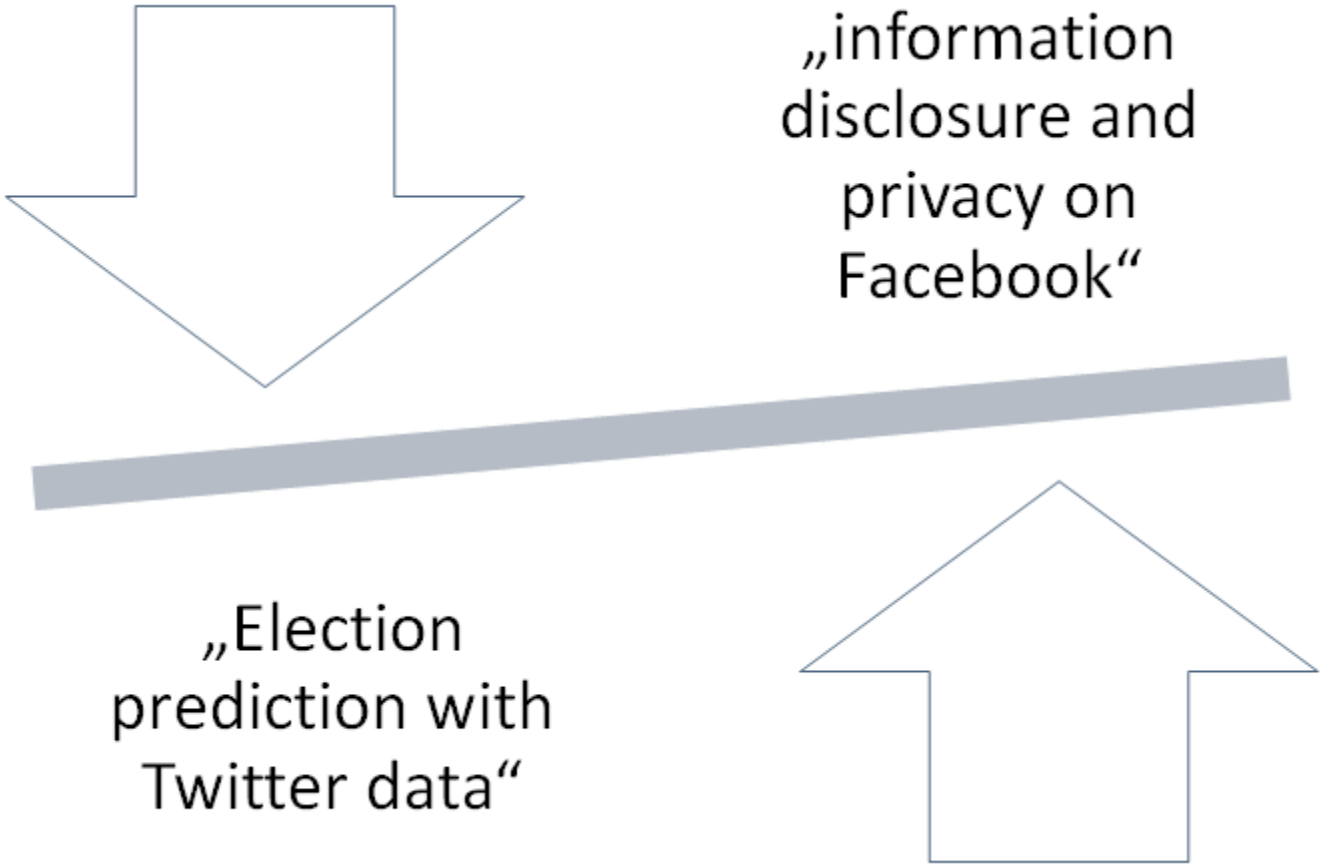
- in practice this is less linear and more iterative
- design choices might need to be revised



Construct definition

- What do we want to measure?
  - Concept / topic that is going to be studied.
  - Constructs can/should be conceptually defined and have a theoretical foundation.
  - Different levels of abstraction. Some constructs cannot be observed "directly".
  - Examples: poverty, education, life satisfaction, urban mobility, social exclusion.
- SMD often act as proxies for more general constructs.

Different streams of social media research



Construct definition

Examples:

Measuring health factors based on social media posts, e.g. obesity, alcohol consumption, insomnia.

- estimating obesity rates (e.g. per geographical region) based on tweets

vs.

- how do people share body images and react to those in online environments?

Construct definition

Examples:

Mobility and location.

- Geo-located Twitter as proxy for global mobility patterns [Havelka et al., 2013].

vs.

- Approaches for measuring accuracy of geo-information derived from Twitter.

Platform selection

#1: Model organisms in social media research?

https://en.wikipedia.org/wiki/Model_organism#/media/File:Drosophila_melanogaster_-_side_(aka).jpg

Tufekci, Z. 2014. Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls. In Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media (ICWSM).

Platforms and traces

- Platforms are shaped by their **affordances**.
- Mostly affordances are used to describe the action possibilities / platform functionalities made available to users by means of technology.
- Platform affordances are affected by technology, platform strategies (marketing), users.

Bucher T and Helmond A (2018) The Affordances of Social Media Platforms. In: Burgess J, Poell T, and Marwick A (eds), *The SAGE Handbook of Social Media*, London: SAGE Publications, pp. 233–253.

Platforms and traces

- Platform effects are important to consider when studying social media [Pfeffer and Malik].
- Affordances shape, *what kind of* ***traces*** users may leave behind.
- *Meaning of traces* may also evolve over time.

Platforms and traces

Examples of affordances that affect traces

- Endorsement? Twitter changing from ★ to ❤. Facebook adding different emotional states to the like button.
- Location? Adding geocodes, plain text fields (“on earth”), choosing from lists of places...

Platforms and traces

Digital traces

- We view the traces people leave in digital platforms as potential **signals for attitudes and behaviour.**
- traces can be actions, content, interactions, connections, metadata.

From construct to collection?

How to operationalize my construct of interest and translate it into a specific data collection approach?

- What will be the selection criteria? Keywords? Persons? Groups? Interactions with specific content?
- What data collection period?
- What would I be missing?
- Is it technically possible and ethically ok?
- How would I want to handle the data in the end (manually, programmatically?)

Potential Pitfalls of SMD

From lots of challenges to a structured reflection on potential errors/pitfalls/limitations?

- How can we make research with digital traces
  - more valid and reliable?
  - more transparent?

- Additionally, how do we improve communication between different disciplines working with Digital Traces?

Our approach: learning from survey research

- Survey research and social media research share some objectives.
- Survey research has constantly evolved as a field - also thanks to to critical reflects on survey methodology.

I. Surveys

Sampling Frame

Survey

Responses

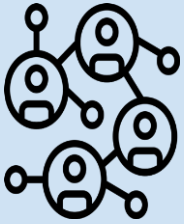
RQ: What do Germans think of the Chancellor?

Estimate

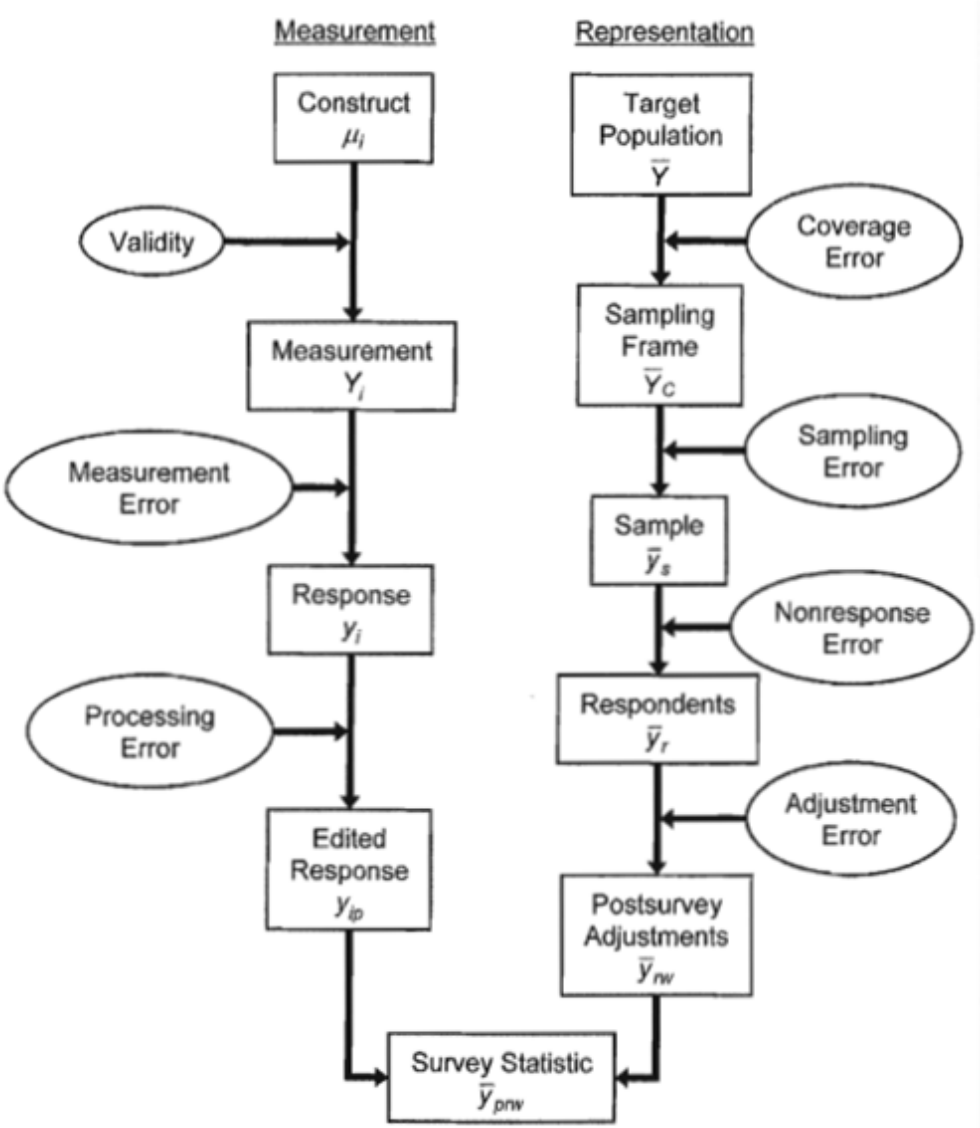
II. DVD

Digital Platform

Query



Users and traces



Survey Lifecycle from a Quality perspective. Adapted from Groves et al. (2011)

gesis Leibniz Institute for the Social Sciences

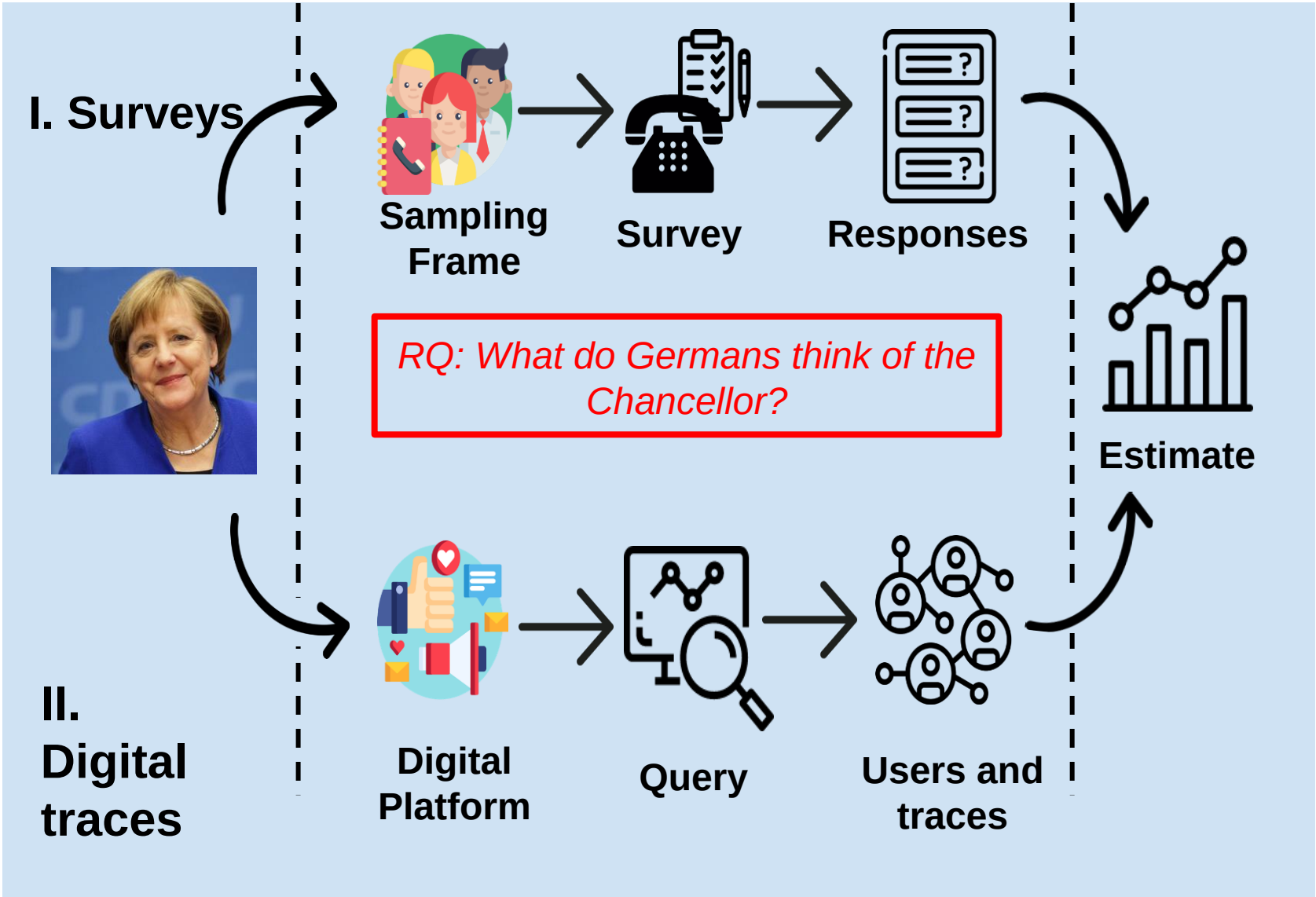
4: Identifying Pitfalls

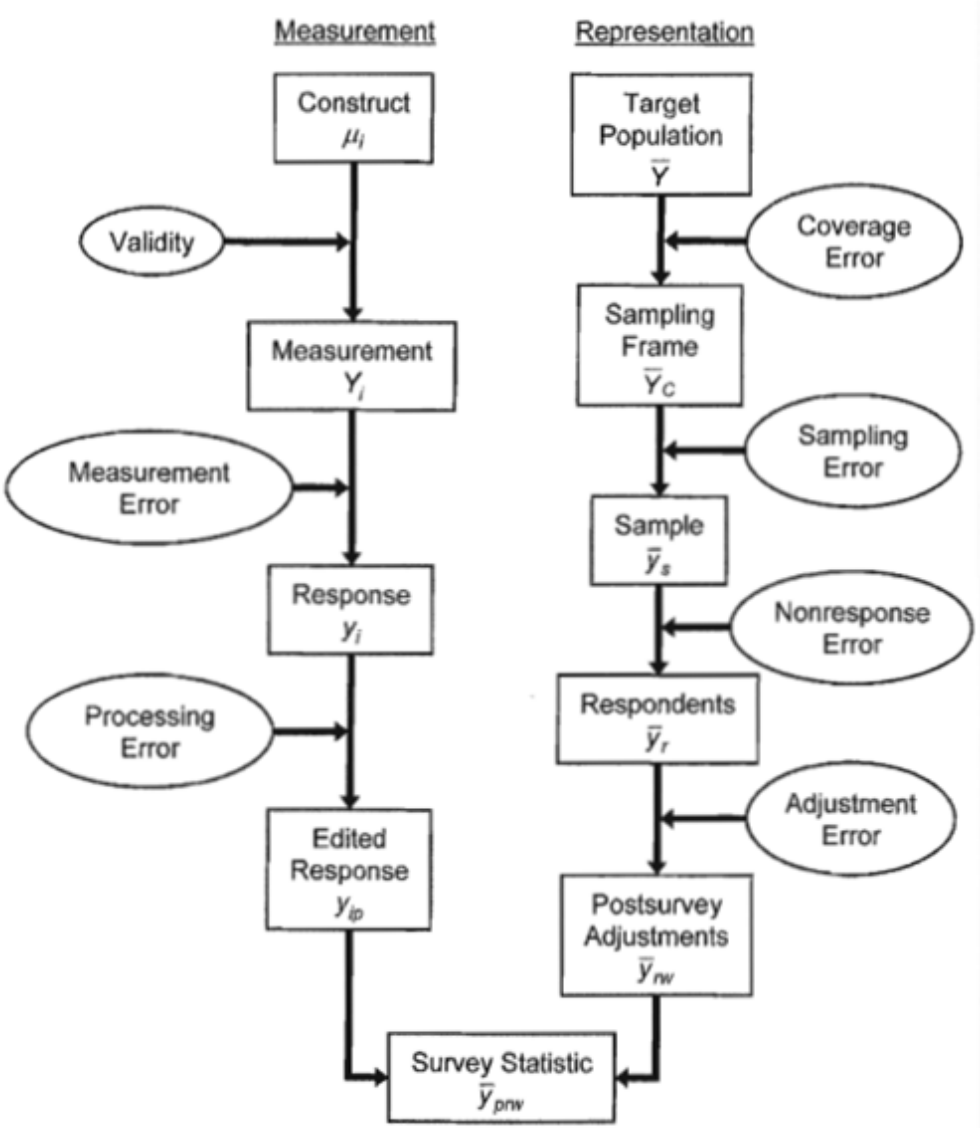
Photo: Katrin Weller

Total Survey Error (TSE)

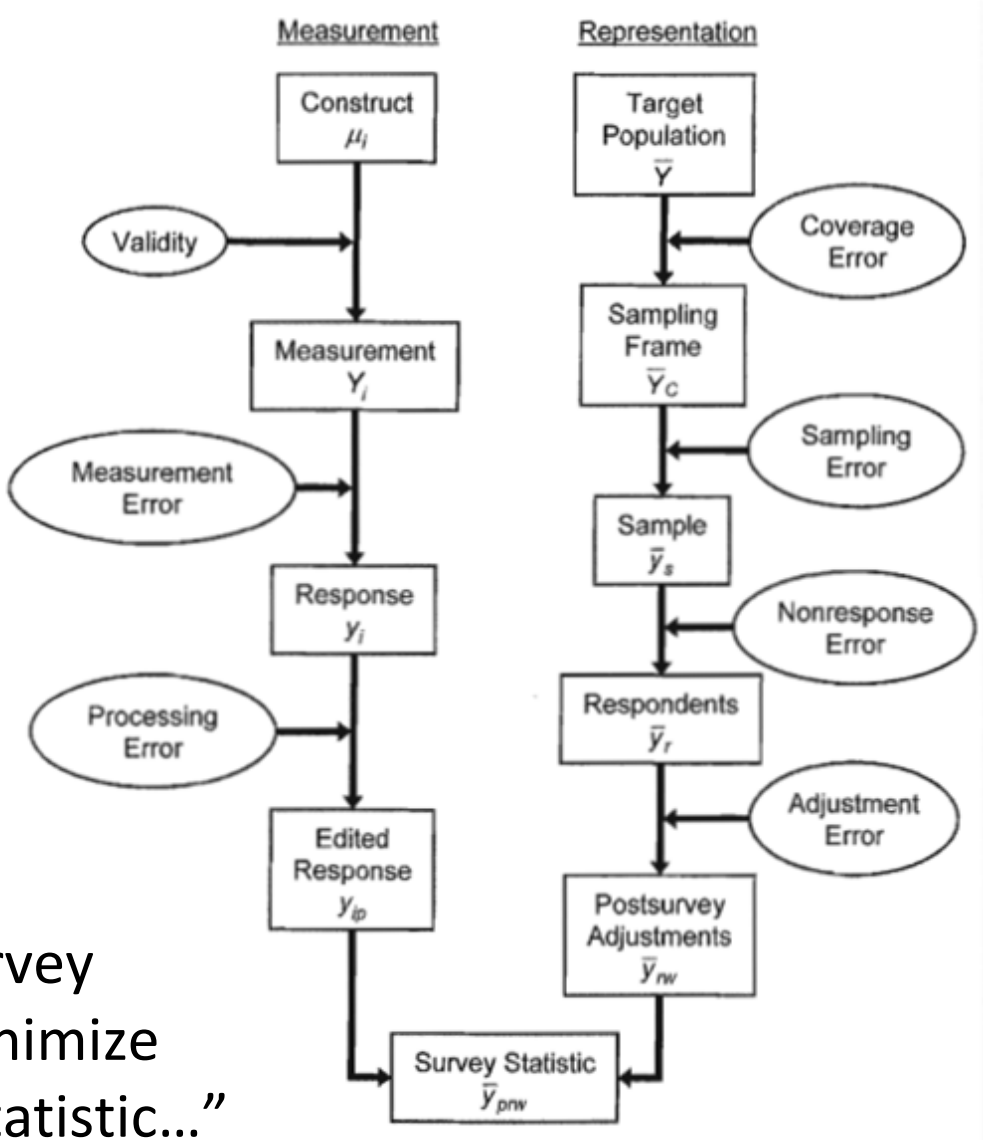
Total Survey Error (TSE) Framework

- Different approaches to create frameworks for identifying potential errors in survey research.
- Most prominent approach by Groves et al.
- Distinguishes between Measurement and Representation Erros.
- Based on the survey lifecycle (typical workflow).


I. Surveys
Sampling Frame
Survey
Responses
RQ: What do Germans think of the Chancellor?
Estimate
II. Digital traces
Digital Platform
Query
Users and traces

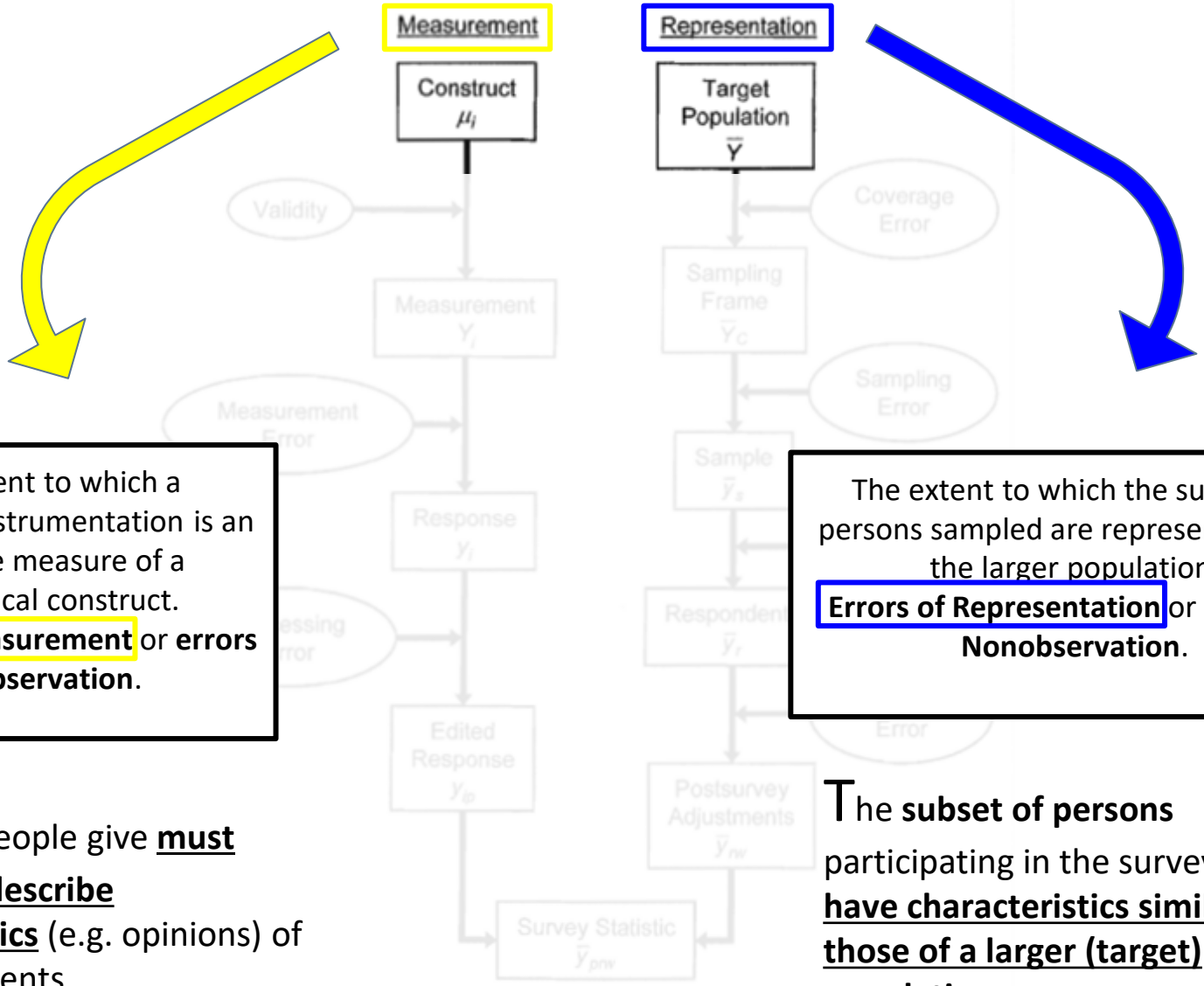


Survey Lifecycle from a Quality perspective. Adapted from Groves et al. (2011)



"The job of a survey designer is to minimize error in survey statistic…"

Survey Lifecycle from a Quality perspective. Adapted from Groves et al. (2011)



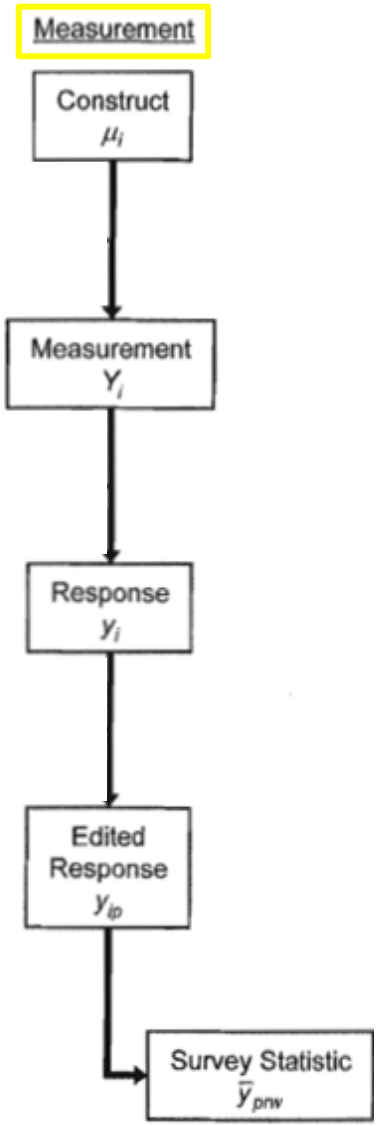
The extent to which a given test/instrumentation is an effective measure of a theoretical construct.
Errors of Measurement or **errors of Observation**.

Answers people give **must accurately describe characteristics** (e.g. opinions) of the respondents.

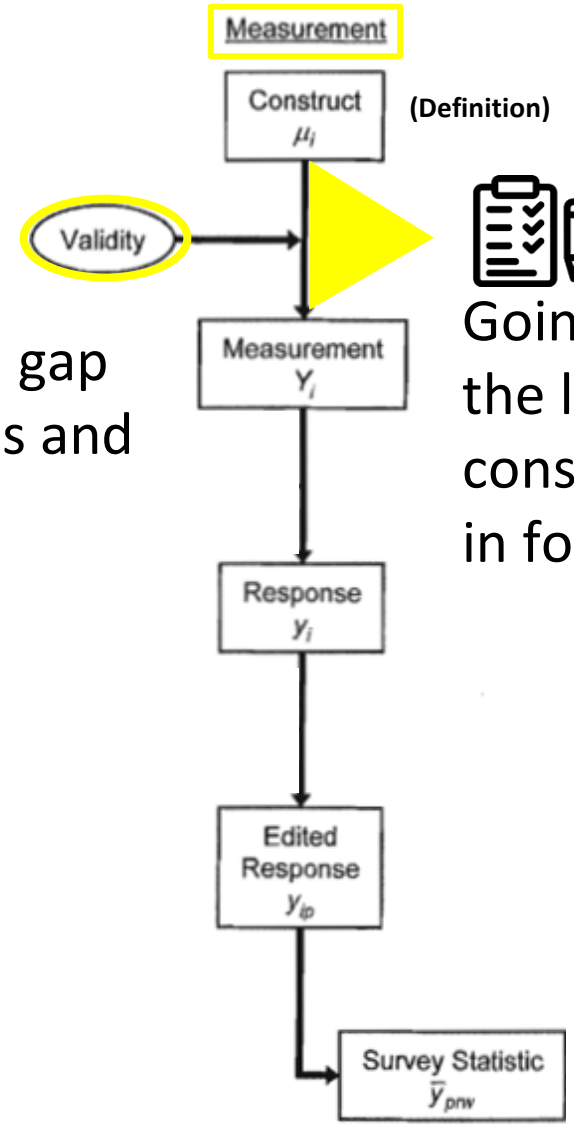
The extent to which the subset of persons sampled are representative of the larger population.
Errors of Representation or **errors of Nonobservation**.

The **subset of persons** participating in the survey **must have characteristics similar to those of a larger (target) population**.

Survey Lifecycle from a Quality perspective. Adapted from Groves et al. (2011)



Survey Lifecycle from a Quality perspective. Adapted from Groves et al. (2011)

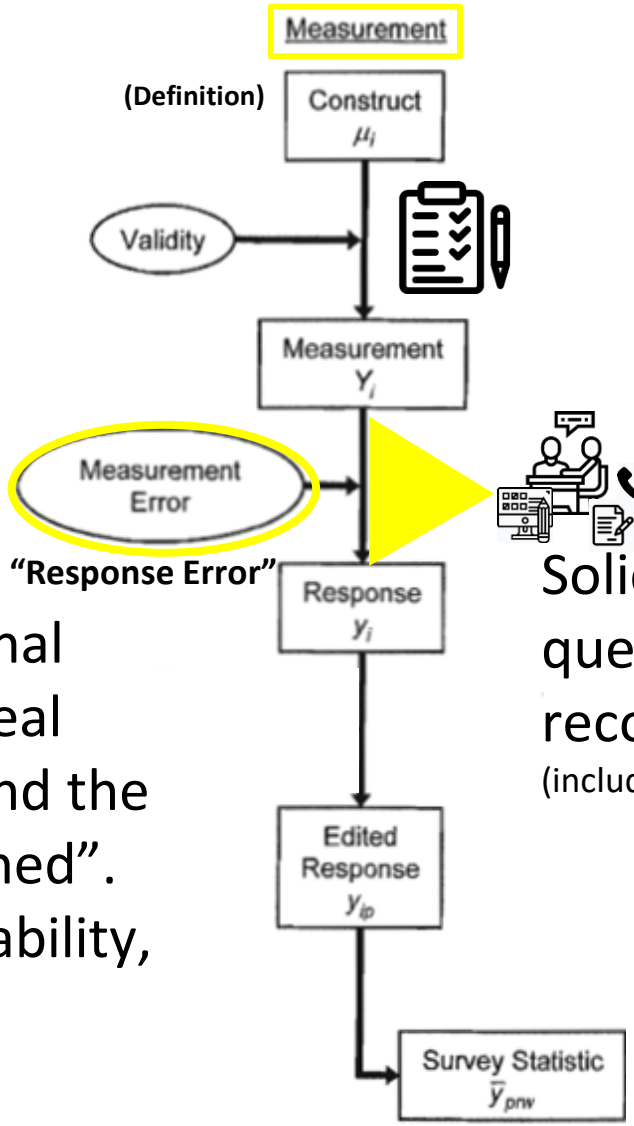


"The observational gap between constructs and measures"

Questionnaire design

Going from a *clear definition* of the latent, unobservable construct to a tangible stimulus in form of questions and items

Survey Lifecycle from a Quality perspective. Adapted from Groves et al. (2011)

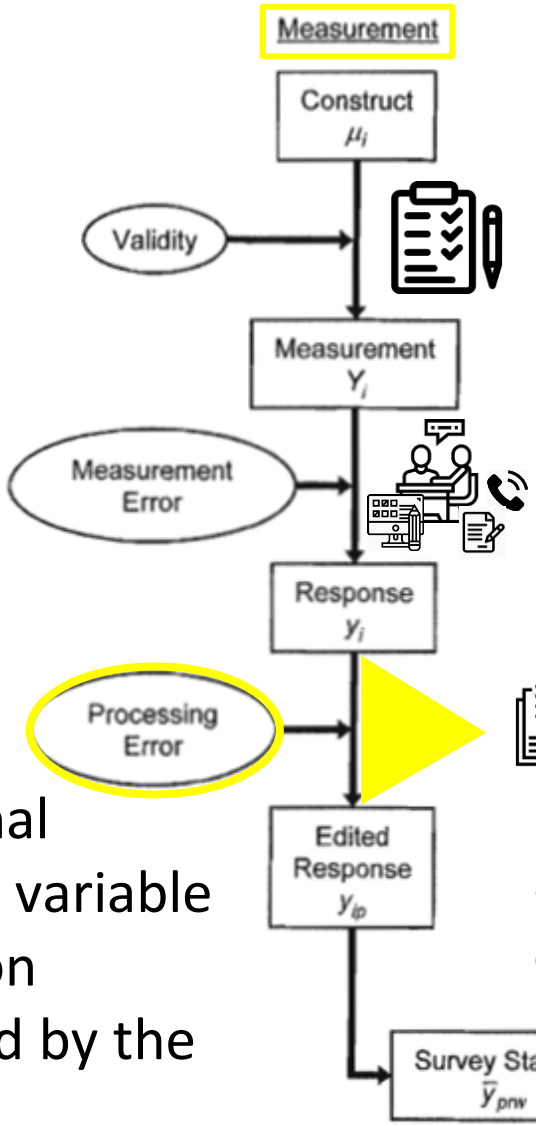


Response collection

Soliciting responses to survey questions from respondents, recording them
(includes choosing collection mode)

"The observational Gap between Ideal Measurement and the Response Obtained". E.g., social desirability, memory gaps

Survey Lifecycle from a Quality perspective. Adapted from Groves et al. (2011)

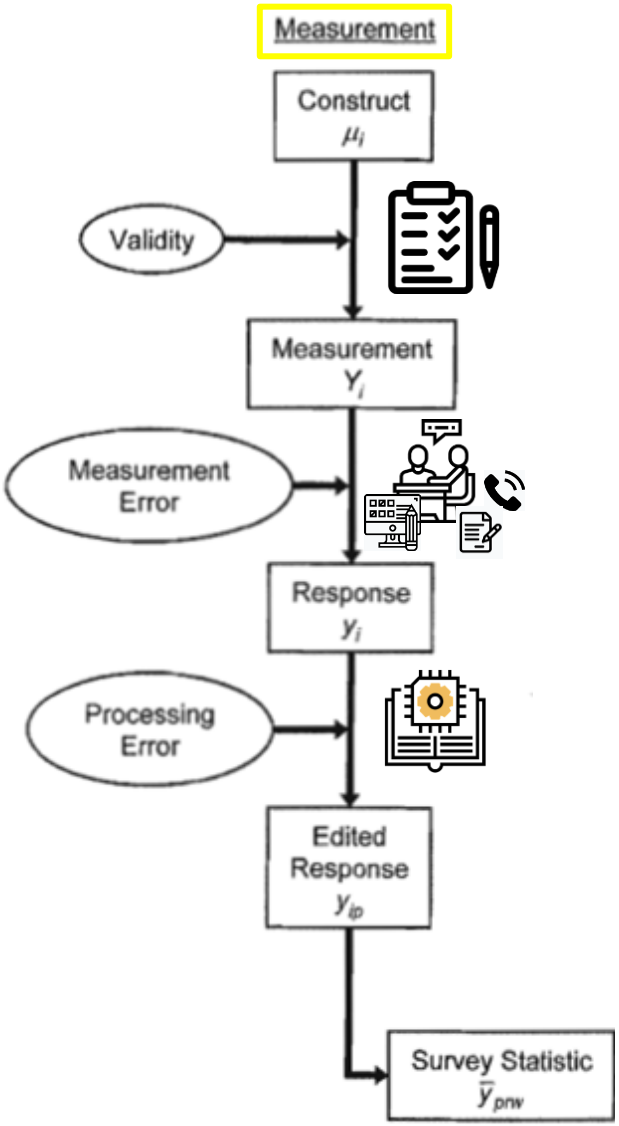


Preprocessing

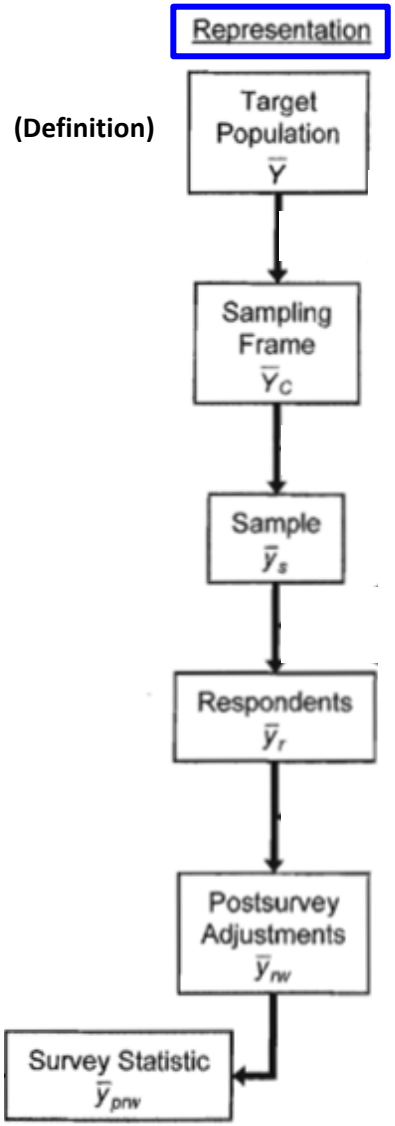
Responses are cleaned, annotated, merged, categorized, etc.

"The observational gap between the variable used in estimation and that provided by the Respondent"

Survey Lifecycle from a Quality perspective. Adapted from Groves et al. (2011)



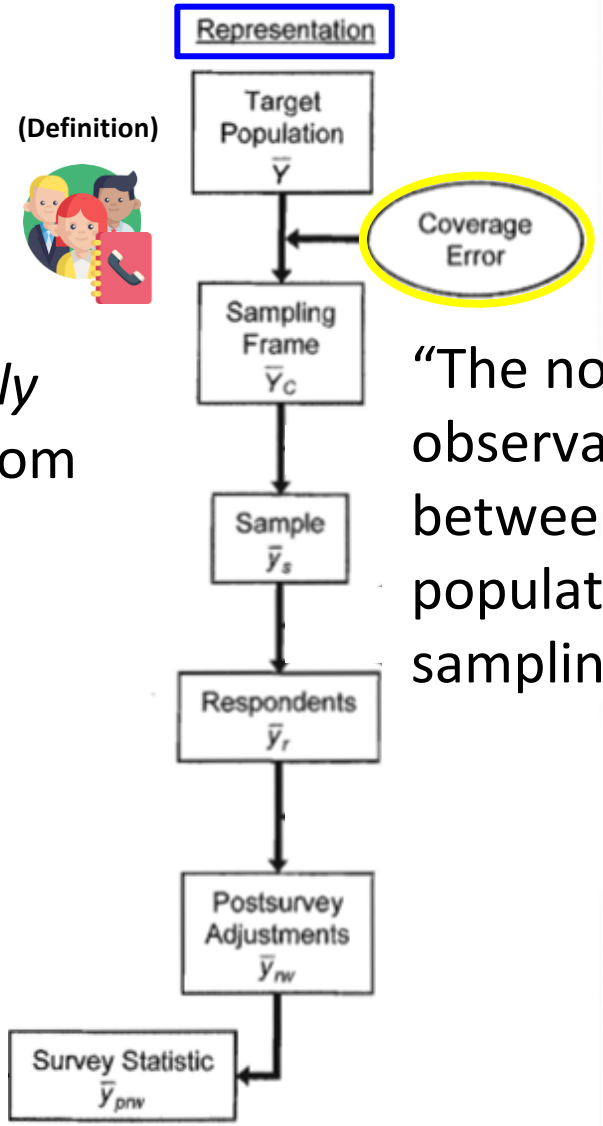
Survey Lifecycle from a Quality perspective. Adapted from Groves et al. (2011)



Survey Lifecycle from a Quality perspective. Adapted from Groves et al. (2011)

Sampling Frame Selection

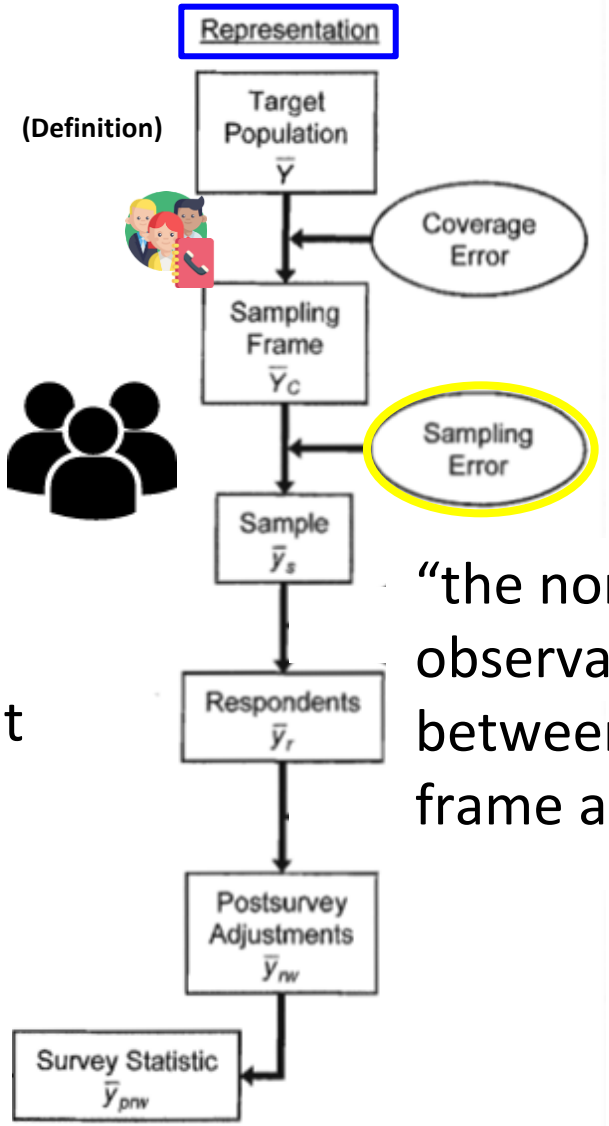
Chose a frame (register, list) best approximating the *clearly defined* target population. From this frame, samples will be drawn.



"The non-observational gap between the target population and the sampling frame "

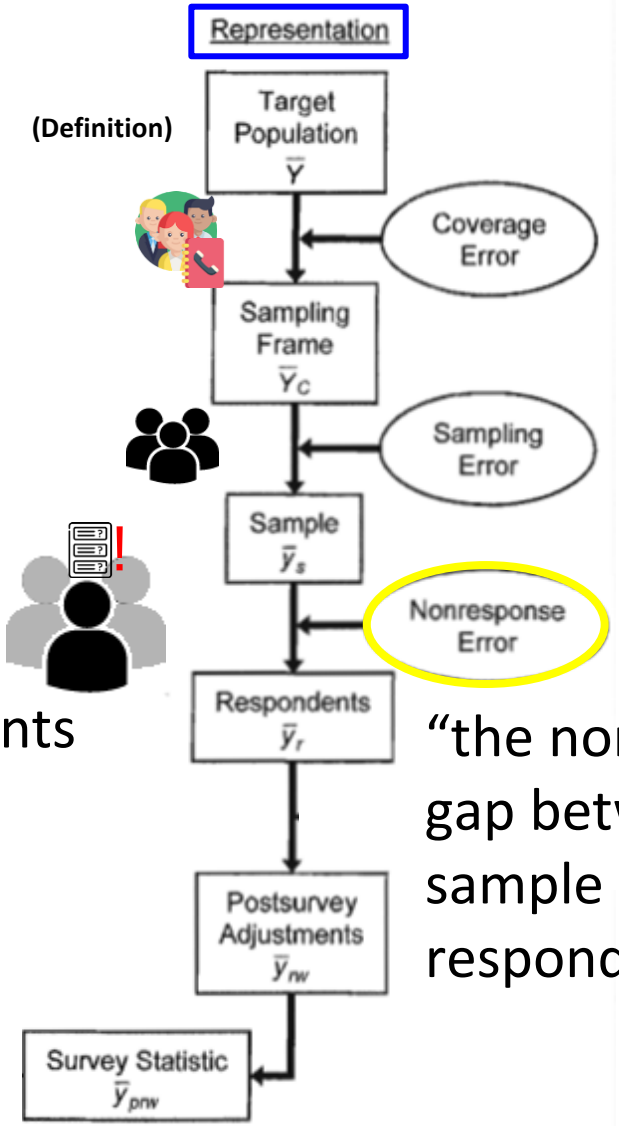
Sampling

Selecting elements from the sampling frame to be surveyed, so that their characteristics best represent the frame → the TP.



"the non-observational gap between the sampling frame and the sample"

Survey Lifecycle from a Quality perspective. Adapted from Groves et al. (2011)

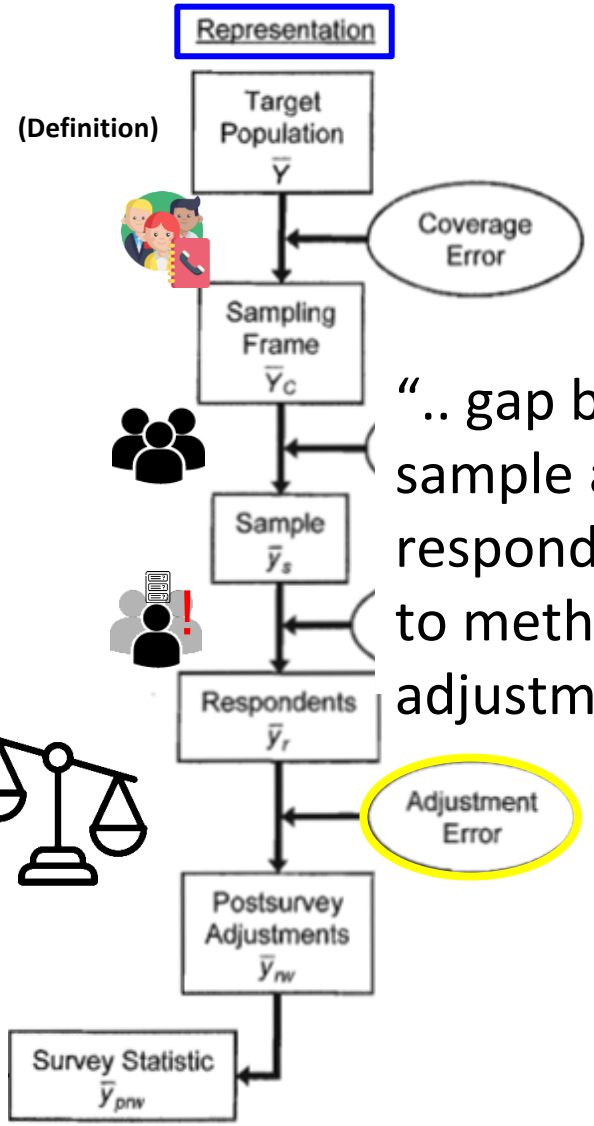


Recruit respondents

Contact sample elements to elicit responses.

"the non observational gap between the sample and the respondent pool"

Survey Lifecycle from a Quality perspective. Adapted from Groves et al. (2011)

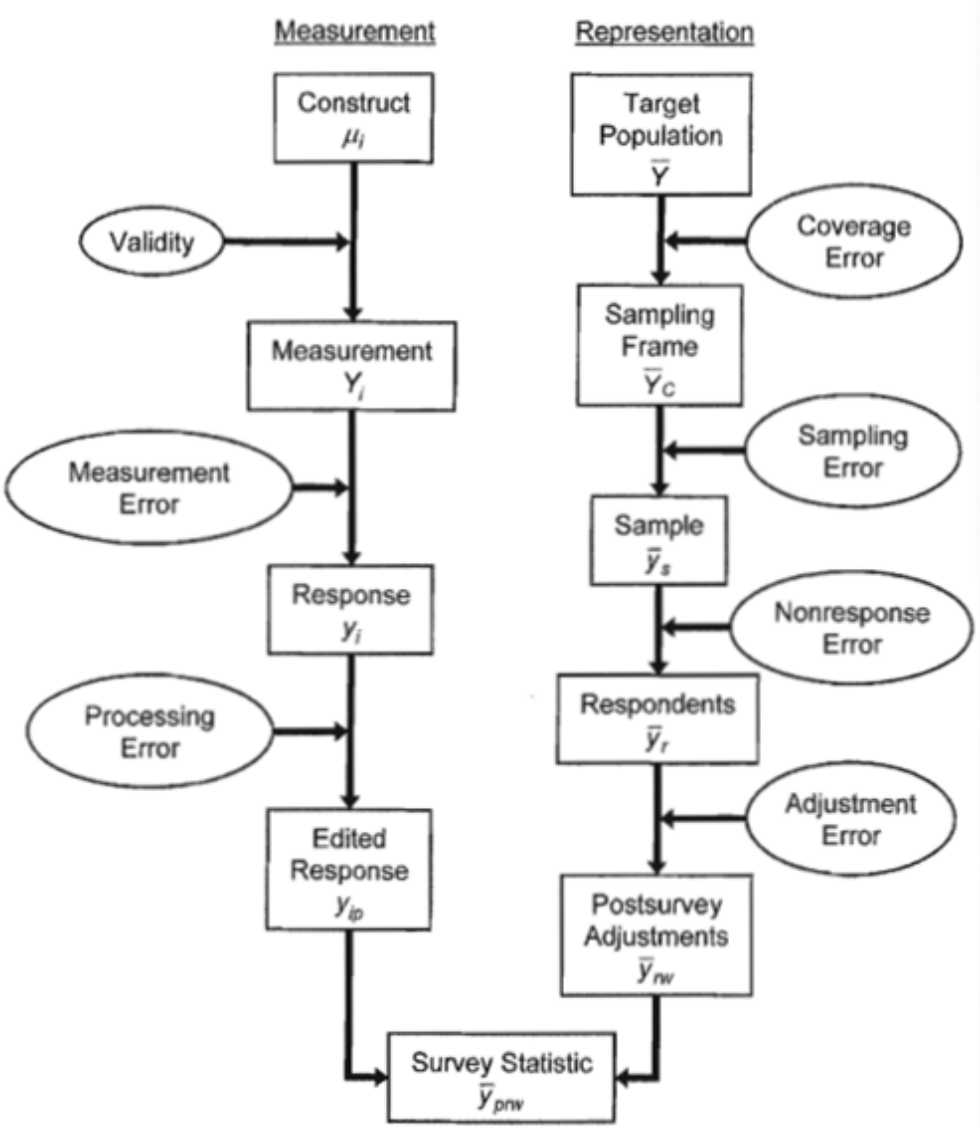


"‥ gap between the sample and the respondent pool… due to method of adjustment "

Posts. Adjustment

Correct the weight individual cases have for the survey statistic

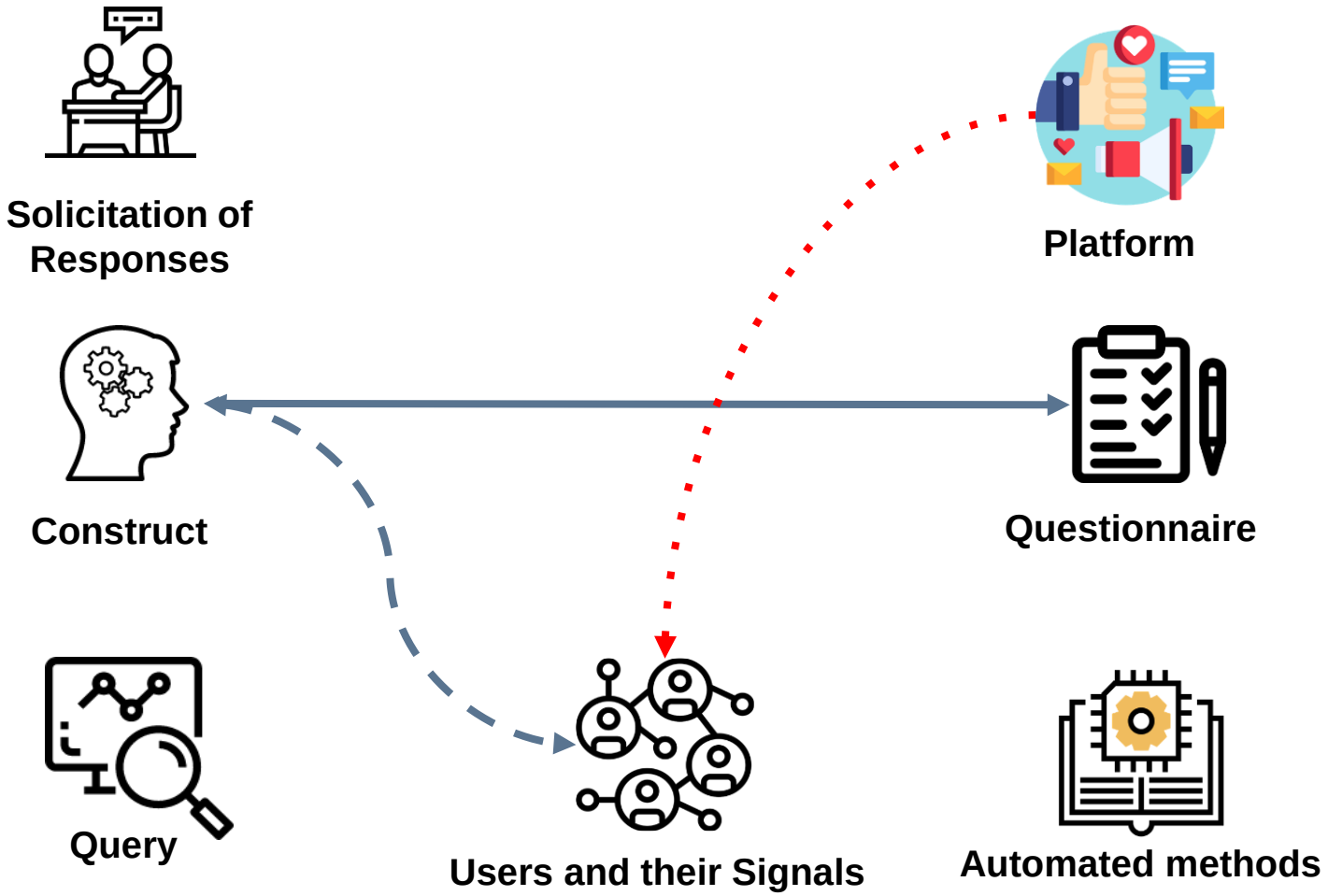
Survey Lifecycle from a Quality perspective. Adapted from Groves et al. (2011)



Survey Lifecycle from a Quality perspective. Adapted from Groves et al. (2011)

Can we apply the TSE to research which uses *digital traces* instead of surveys?

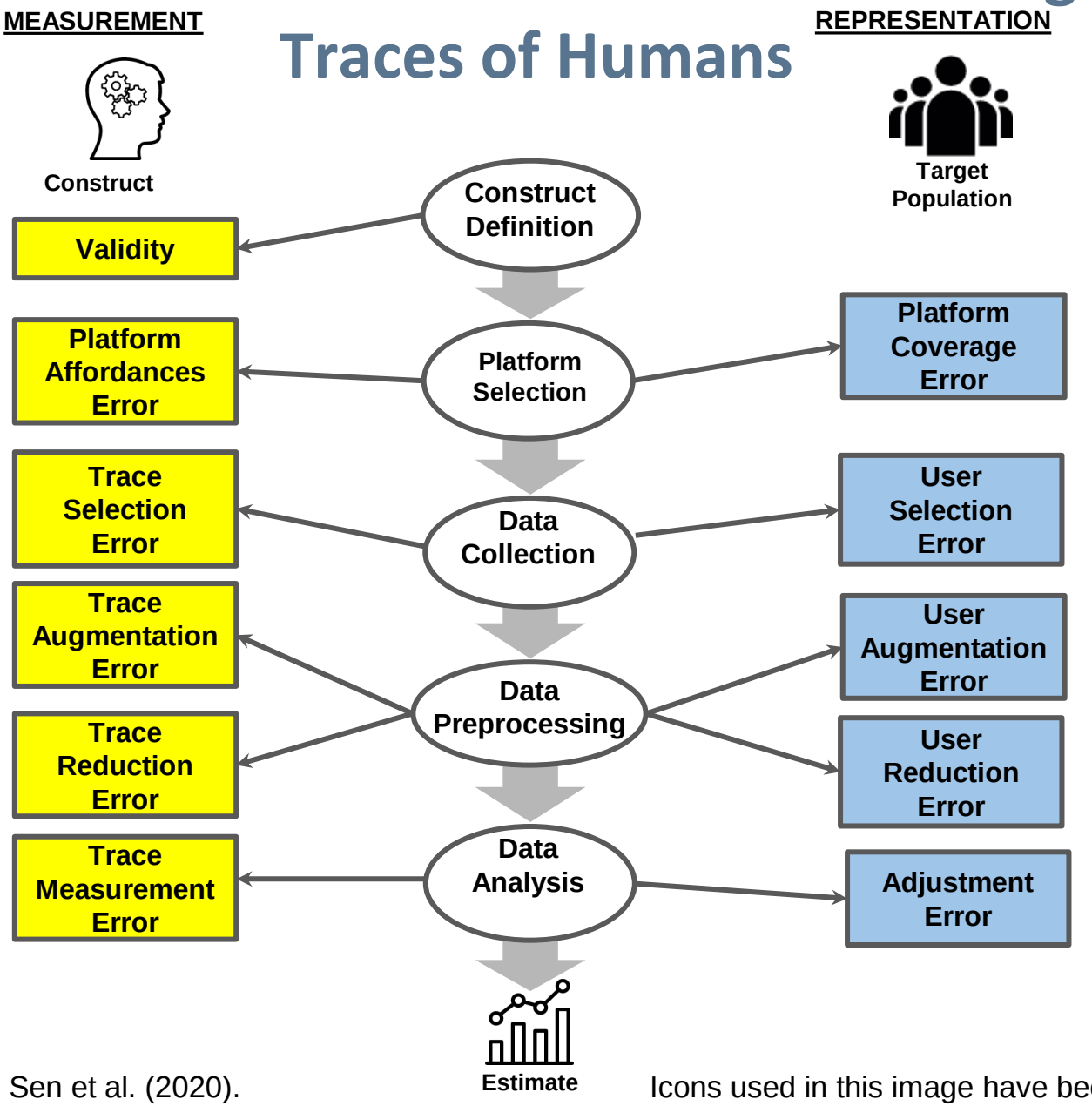
Can we apply the TSE to social science research which uses *digital traces* instead of surveys?



Next Week: Can we apply the TSE to social science research which uses *digital traces* instead of surveys?

- The link between the construct and digital **traces** is much less straightforward
- The **Platform norms** and **affordances** affect how users generate traces on that platform
- Queries select **both traces and users =** representation and measurement errors harder to disentangle
- Due to large-scale and high-resolution data, **automated or semi-automated methods** are used which come with their pitfalls

TED-On: A Total Error Framework for Digital Traces of Humans



Sen et al. (2020).

Icons used in this image have been designed by Becris, EliasBikbulatov and Pixel perfect from www.flaticon.com