gesis Leibniz Institute for the Social Sciences

Acknowledging potential pitfalls in social media research: Between researchers practices and structured documentation approaches

Katrin Weller & Indira Sen
katrin.weller@gesis.org, @kwelle (GESIS & CAIS)
indira.sen@gesis.org, @indiiigosky (GESIS)
November 22, 2022

Member of Leibniz Association

Social Media: A Source of Data for Social Science Research?

Online platforms such as social media platforms have become crucial elements of our lives – and have thus also become the object of academic research.

The "ABC"

Social media data can also help to identify

Attitudes and opinions,

Behavior,

Characteristics

of human users of digital technologies.

Social media platforms as sensors may better recall certain facts than human memory.

Mining communication from existing digital streams can be more timely than creating a survey. They are a valuable source, especially during unforeseeable events.

Sometimes social media data may enable looking into topics for which it would be difficult to recruit study participants otherwise.
They are often created without any stimulus from a researcher.

gesis Leibniz Institute for the Social Sciences

Researchers practices and experiences

Research based on data
from social media platforms is not a
consistent field.

different platforms
different methods
different disciplines
different motivations
different skills

different opportunities to access (restricted) data

similar experiences when interacting with social media platforms and the complexities they are entangled in

gesis Leibniz Institute for the Social Sciences

legal frameworks

research ethics

changing platform access options / ToS

user expectations / privacy

similar experiences when interacting with social media platforms and the complexities they are entangled in

platforms as black boxes

data access

interdisciplinarity

data sharing

methods as black boxes

publishing practices

missing data

Member of Leibniz Association

gesis Leibniz Institute for the Social Sciences

but different conclusions when it comes to addressing specific challenges

E.g., in the context of research ethics:

- ? Big vs. small data
- ? Users as authors vs. users as research subjects
- ? Particularly vulnerable groups (e.g., activists) vs. professional / public accounts (e.g., politicians)
- ? Different practices in quoting from user accounts based on disciplinary requirements

Weller, Katrin, and Katharina E. Kinder-Kurlanda. 2017. "To Share or Not to Share?: Ethical Challenges in Sharing Social Media-based Research Data." In Internet Research Ethics for the Social Age, edited by Michael Zimmer, and Katharina E. Kinder-Kurlanda, 115-129. New York u.a.: Peter Lang.

but different conclusions when it comes to addressing specific challenges

Or in the context of data sharing:

- ? balancing between following principles of good scientific practice and between respecting legal constraints
- ? Perceived ethical obligations *towards the scientific community*
- ? Not sharing data to protect users vs. sharing to include users

Weller, Katrin, and Katharina E. Kinder-Kurlanda. 2015. "Uncovering the Challenges in Collection, Sharing and Documentation: The Hidden Data of Social Media Research?." In *Standards and Practices in Large-Scale Social Media Research: Papers from the 2015 ICWSM Workshop. Proceedings Ninth International AAAI Conference on Web and Social Media Oxford University, May 26, 2015 – May 29, 2015*, 28-37. Ann Arbor, MI: AAAI Press.

social media data aren‘t „ordinary“ research data

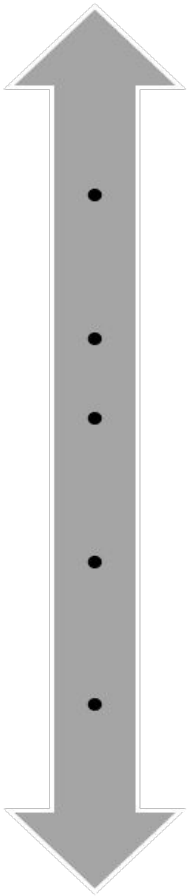
Perceived ethical obligations *towards social media users*

> *“It’s all public, it doesn’t belong to us, we don’t create the data, we don’t evoke it, I mean it’s natural. I don’t think you have the right to really keep other people from it, no.”*

> *“We share datasets with everybody, actually. We don’t feel we own that.”*

how much should I share?

Most reproducibility



What is being shared?

- whole dataset plus additional research information (e.g. scripts)
- whole dataset
- whole dataset, but without direct identifiers (pseudonymization)
- parts of the dataset removed (anonymization)
- changed dataset (e.g. only tweet IDs)

Most privacy

Weller, Katrin, and Katharina E. Kinder-Kurlanda. 2016. "A manifesto for data sharing in social media research."
In Proceedings of the 8th ACM Conference on Web Science (WebSci '16), 166-172. New York: ACM.

gesis Leibniz Institute for the Social Sciences

Researchers want data to be shareable…

“But you can’t make your data available for others to look at, which means both your study can’t really be replicated and it can’t be tested for review. ”

Weller, Katrin, and Katharina E. Kinder-Kurlanda. 2015. "Uncovering the Challenges in Collection, Sharing and Documentation: The Hidden Data of Social Media Research?." In *Standards and Practices in Large-Scale Social Media Research: Papers from the 2015 ICWSM Workshop. Proceedings Ninth International AAAI Conference on Web and Social Media Oxford University, May 26, 2015 – May 29, 2015*, 28-37. Ann Arbor, MI: AAAI Press.

… but are not necessarily keen on reusing existing datasets

"I actually only use [other researchers' datasets] where I'm very sure about where it comes from and how it was processed and analyzed. ***There is too much uncertainty in it****."*

High awareness of potential pitfalls in handling the data –
but lack of good practices for making processes transparent.

“unfortunately, I don’t document much”

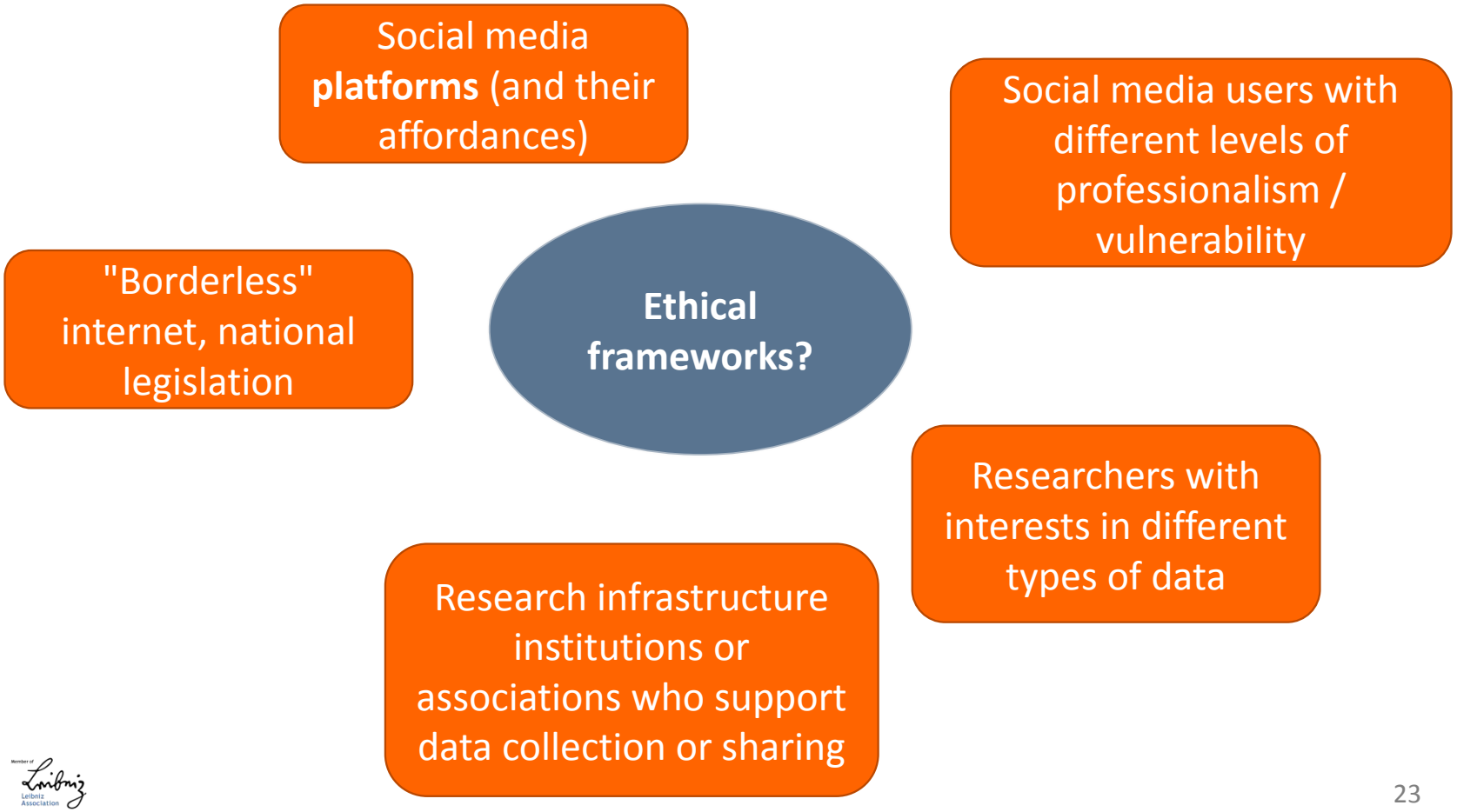
Lots of the decisions that researchers make on a day to day basis are „hidden“. And so are many of their lessons learned and best practices.

Can we support researchers during processes of decision making?

1. Acknowledge structures beyond the individual researchers' influence.

Kinder-Kurlanda, K.E., & Weller, K. (2020). Perspective: Acknowledging Data Work in the Social Media Research Lifecycle. Frontiers in Big Data 3:509954. doi: 10.3389/fdata.2020.509954
https://www.frontiersin.org/articles/10.3389/fdata.2020.509954/full

Different entities that affect potential study design – and research ethics



2. Untangling and documenting choices during the research lifecycle, especially when researchers pursue specific approaches and may have actively decided against others (often due to external factors).

Kinder-Kurlanda, K.E., & Weller, K. (2020). Perspective: Acknowledging Data Work in the Social Media Research Lifecycle. Frontiers in Big Data 3:509954. doi: 10.3389/fdata.2020.509954
https://www.frontiersin.org/articles/10.3389/fdata.2020.509954/full

gesis Leibniz Institute for the Social Sciences

Towards structured reflections on research decisions

Error Frameworks

can

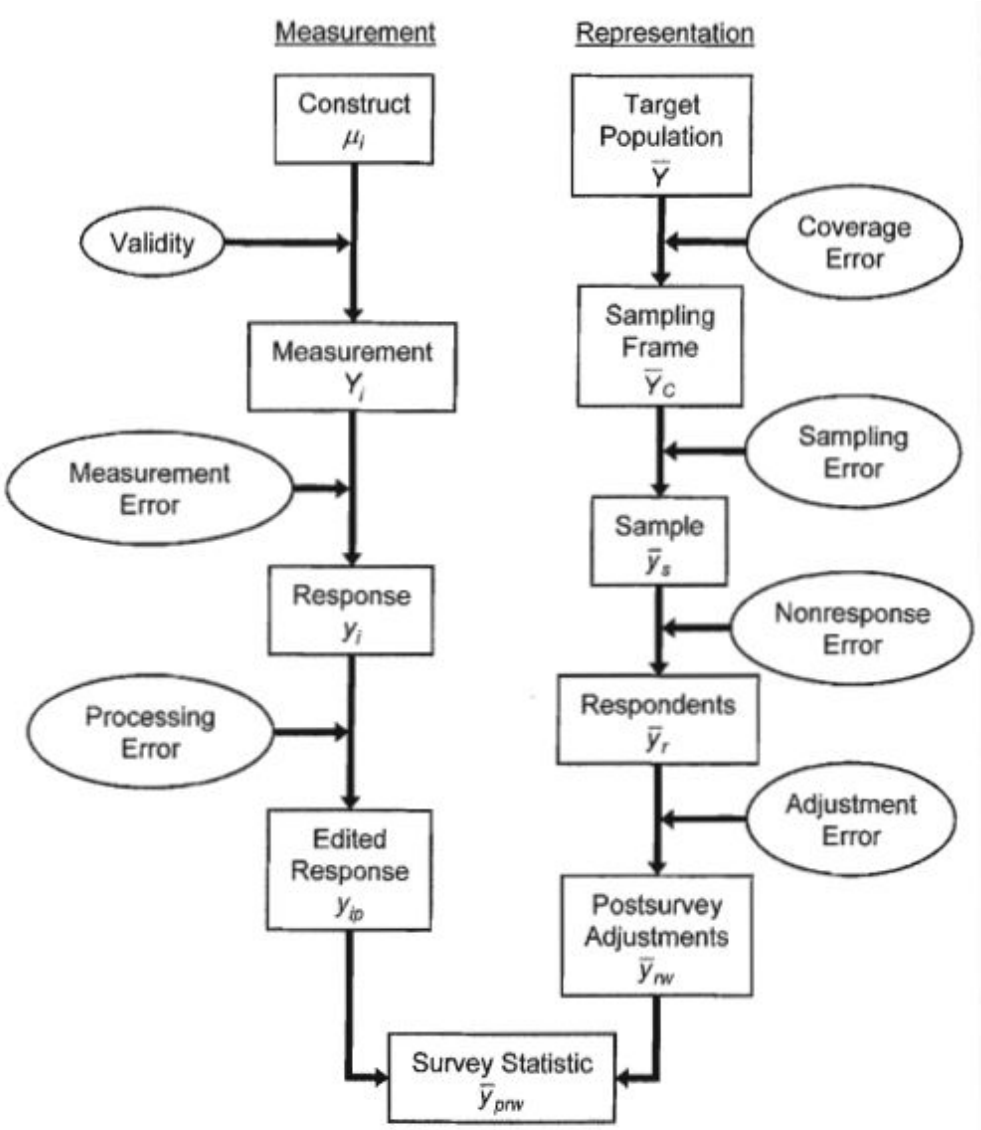
- provide a shared understanding of (and shared vocabulary for) potential challenges and pitfalls
- be a guide for writing methods sections / limitations and for reviewing papers
- encourage and inspire ongoing discussions and reflections on research quality (including methodological CSS research)

Inspiration: *Total Survey Error (TSE) Framework*

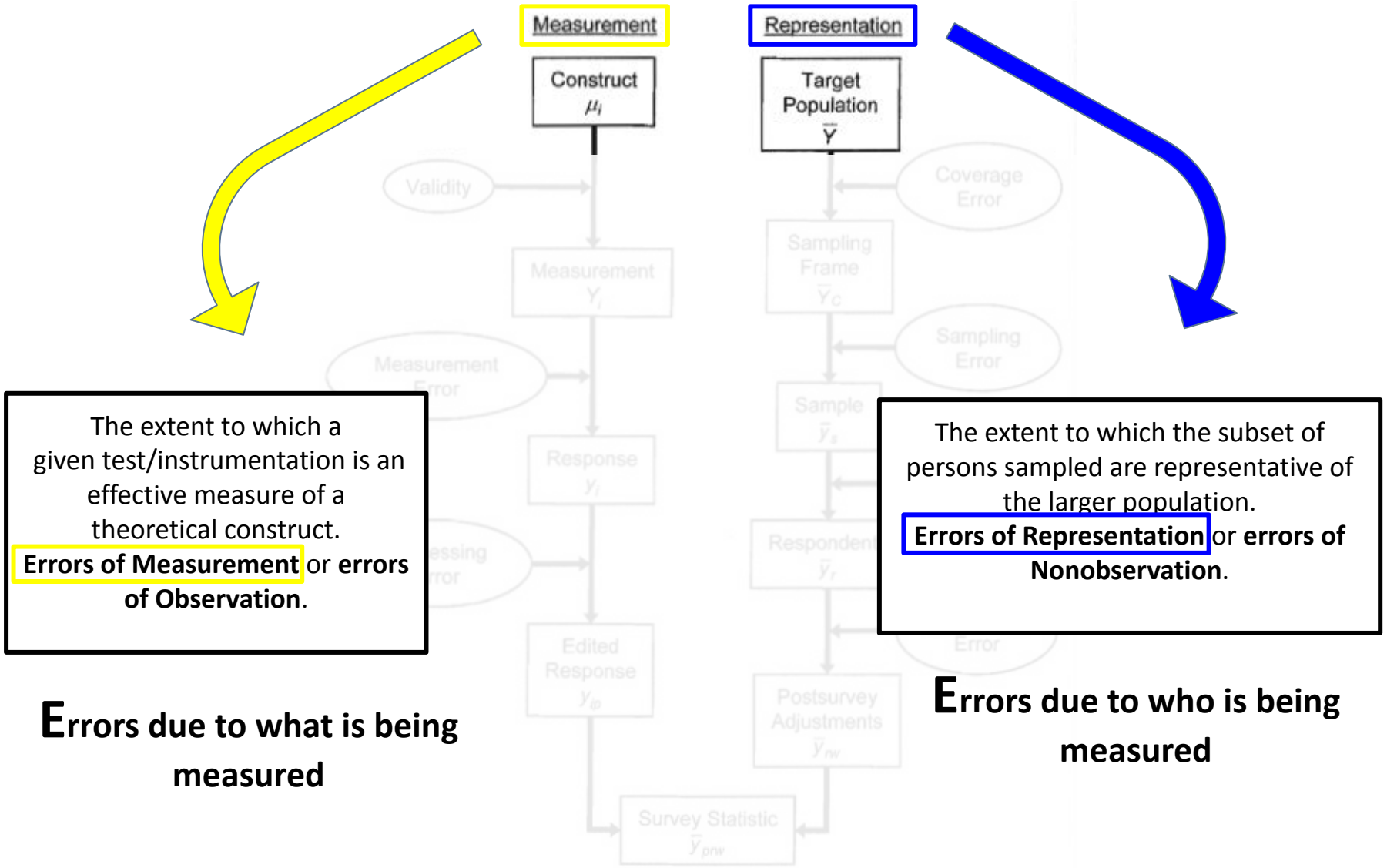
- Different approaches to create frameworks for identifying potential errors in survey research.
- Most prominent approach by Groves et al.
- Based on the survey lifecycle (typical workflow).

Groves Robert M., Fowler Floyd J.Jr., Couper Mick P., Lepkowski James M., Singer Eleanor, Tourangeau Roger. 2011. Survey Methodology, vol. 561. John Wiley and Sons.

Groves Robert M., Lyberg Lars. 2010. “Total Survey Error: Past, Present, and Future.” Public Opinion Quarterly 74(5):849–79.



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**.

Errors due to what is being measured

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

Errors due to who is being measured

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

Other Error Frameworks:

Total Error Framework for Big Data (TEF)

Amaya, Ashley, Paul P. Biemer, and David Kinyon. "Total error in a big data world: Adapting the TSE framework to big data." Journal of Survey Statistics and Methodology 8, no. 1 (2020): 89-119.

Total Twitter Error

Hsieh, Yuli Patrick, and Joe Murphy (2017). "Total twitter error." Total survey error in practice: 23-46.

Total Error Framework for Metered Data

Bosch, Oriol J., and Melanie Revilla. "When survey science met online tracking: presenting an error framework for metered data." (2021).

Adaption to prototypical workflow in social media research?

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

What do we want to study?

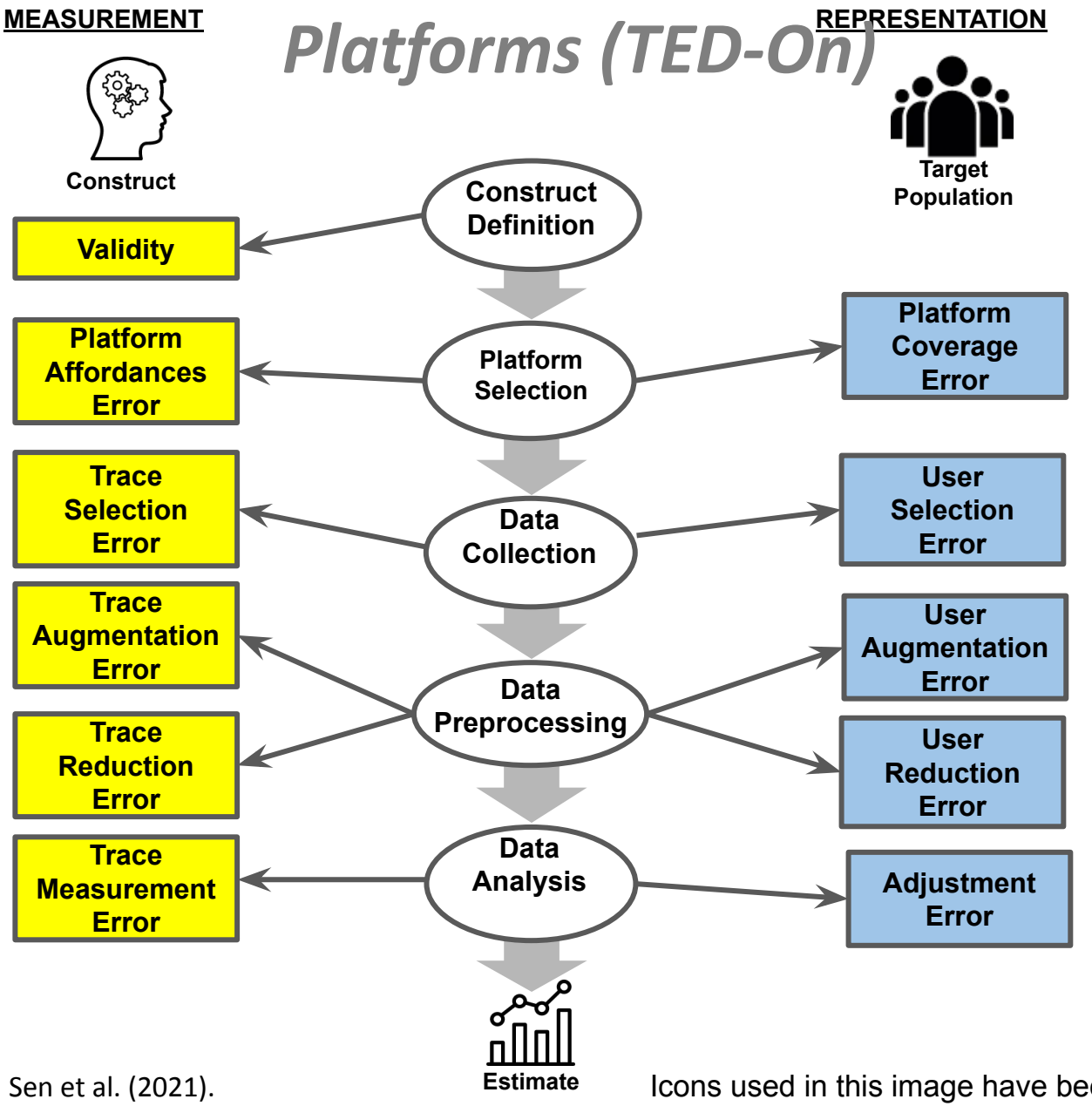
Which platform(s) do we focus on?

How do we collect the data?

How do we clean the data?

How do we analyse the data?

Error Framework for Digital Traces on Online Platforms (TED-On)

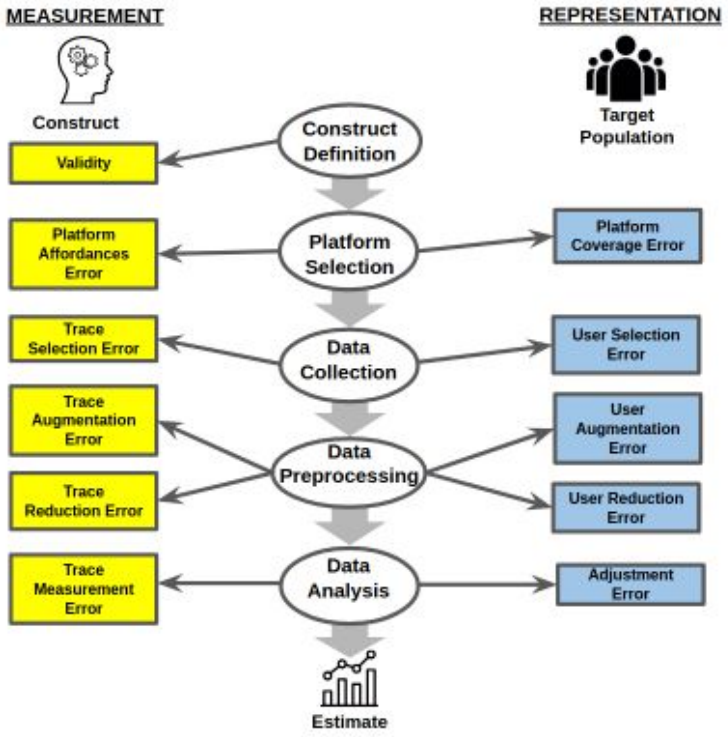


Sen et al. (2021).

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

Error Framework for Digital Traces on Online Platforms (TED-On)

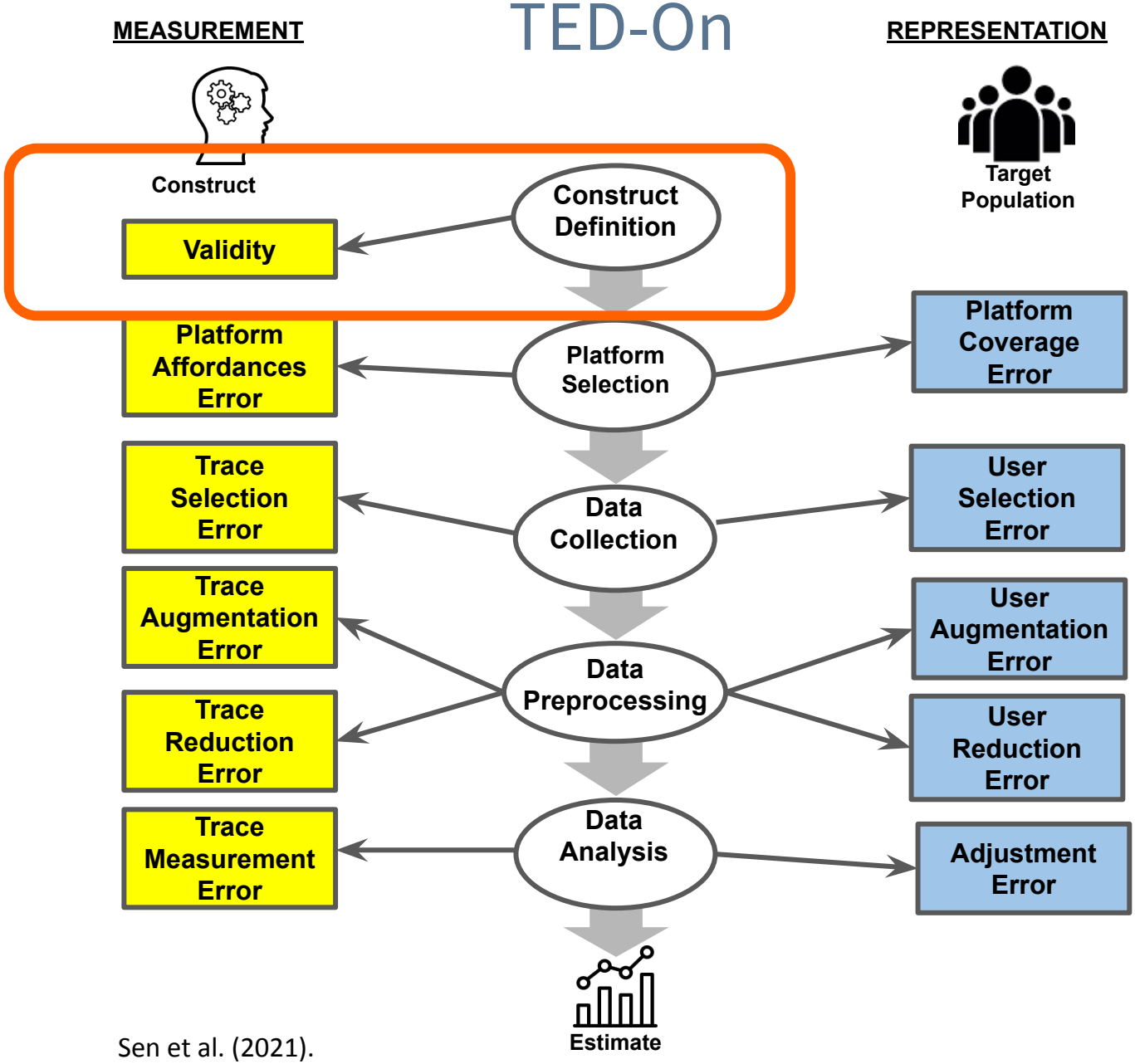
- Distinguishes between:
  - Measurement errors: errors due to <u>what</u> is measured
  - Representation errors: errors due to <u>who</u> is being measured
- Accounts for errors idiosyncratic to digital traces (including web and social media data) such as the *effect of platform recommendation systems*



Sen, I., Flöck, F., Weller, K., Weiß, B., & Wagner, C. (2021). A Total Error Framework for Digital Traces of Human Behavior on Online Platforms. Public Opinion Quarterly. Volume 85, Issue S1, 2021, Pages 399–422, https://doi.org/10.1093/poq/nfab018

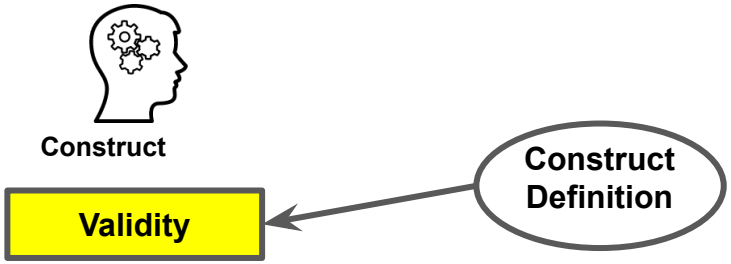
Examples.

TED-On



Sen et al. (2021).

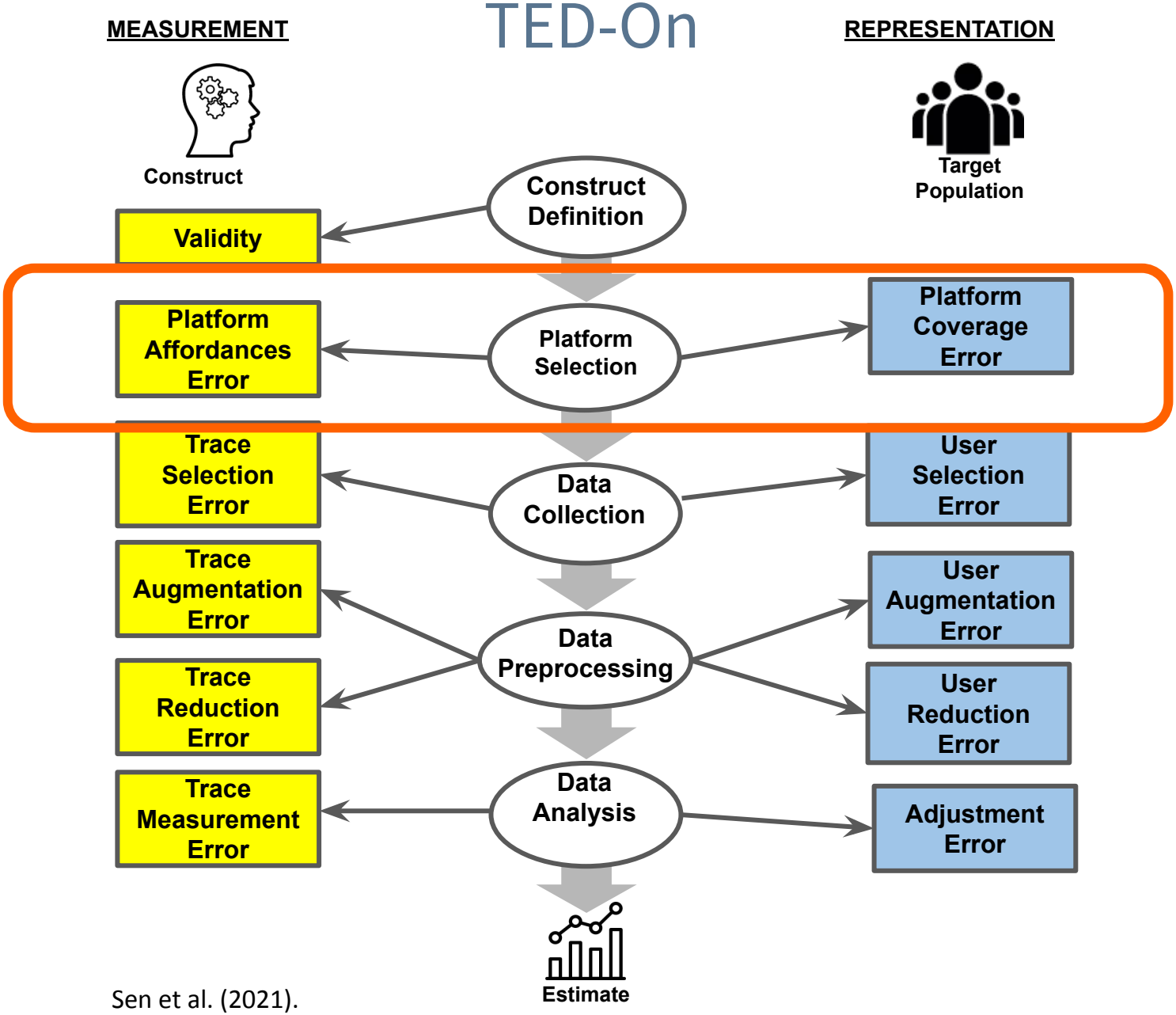
MEASUREMENT



Are we actually measuring what we *think* we're measuring?

For e.g., is someone posts a message with the hashtag #Trump2024, are they supporting Trump or just talking about his presidential bid?

TED-On



Sen et al. (2021).

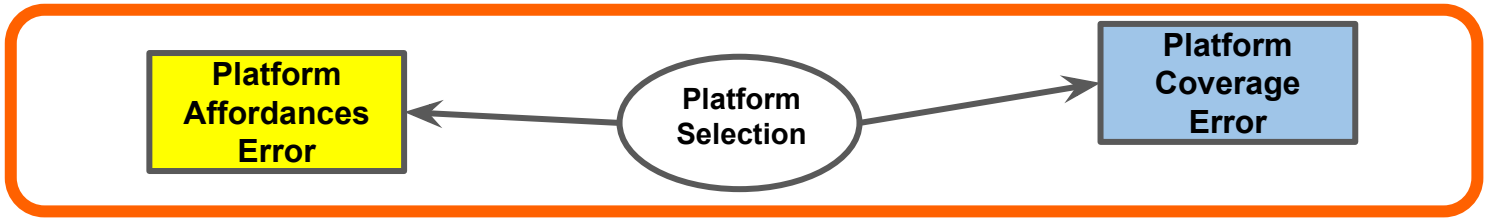
TED-On

MEASUREMENT

Construct

REPRESENTATION

Target Population



Platform has affordances which may distort traces – e.g. on **Reddit:** specific subreddit norms

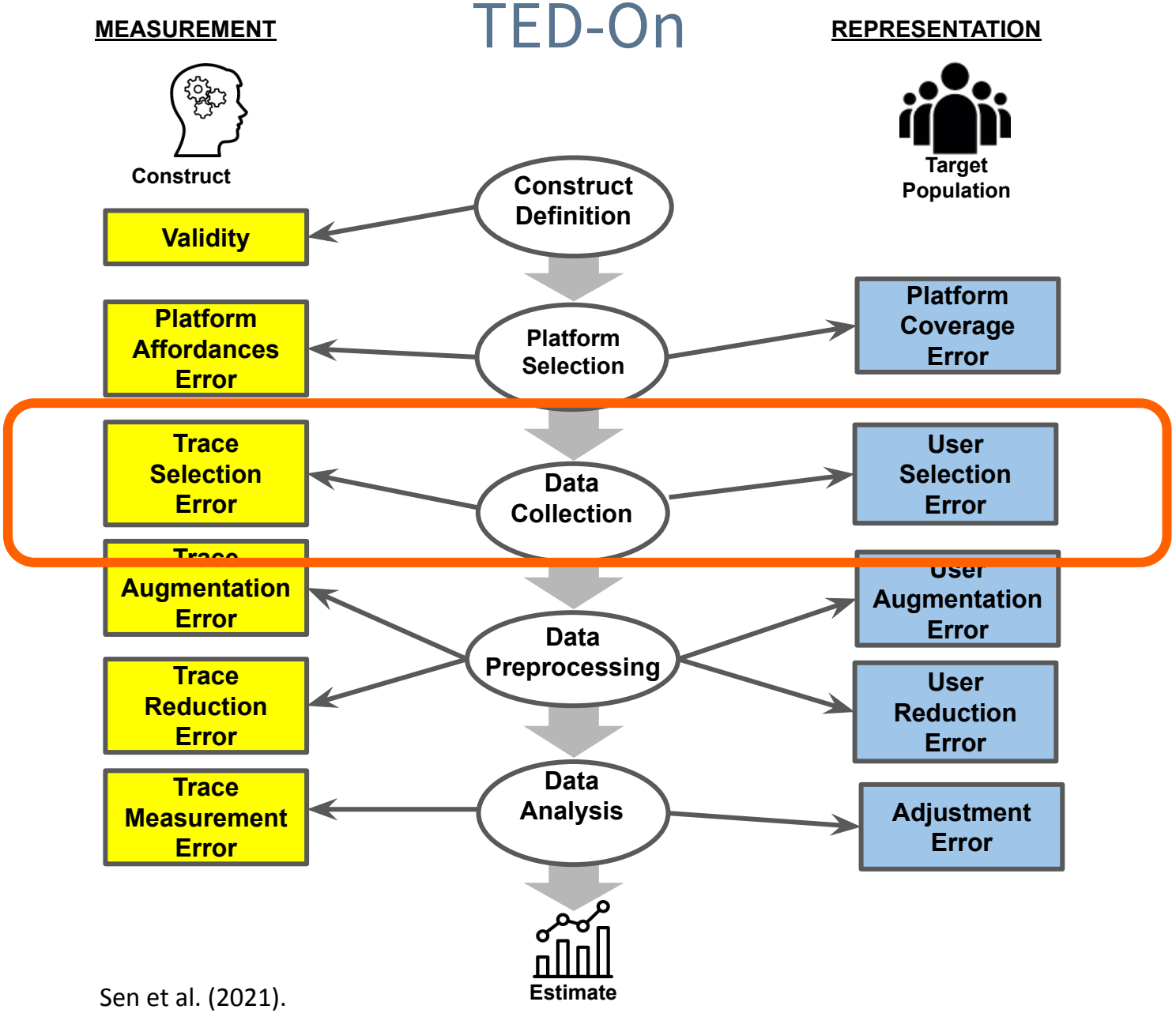
Twitter: trending topics, 280 character limit

Platform is not representative of target population!

And even within the platform: users act differently (who posts about politics?)

Sen et al. (2021).

TED-On



Sen et al. (2021).

TED-On

MEASUREMENT

Construct

REPRESENTATION

Target Population

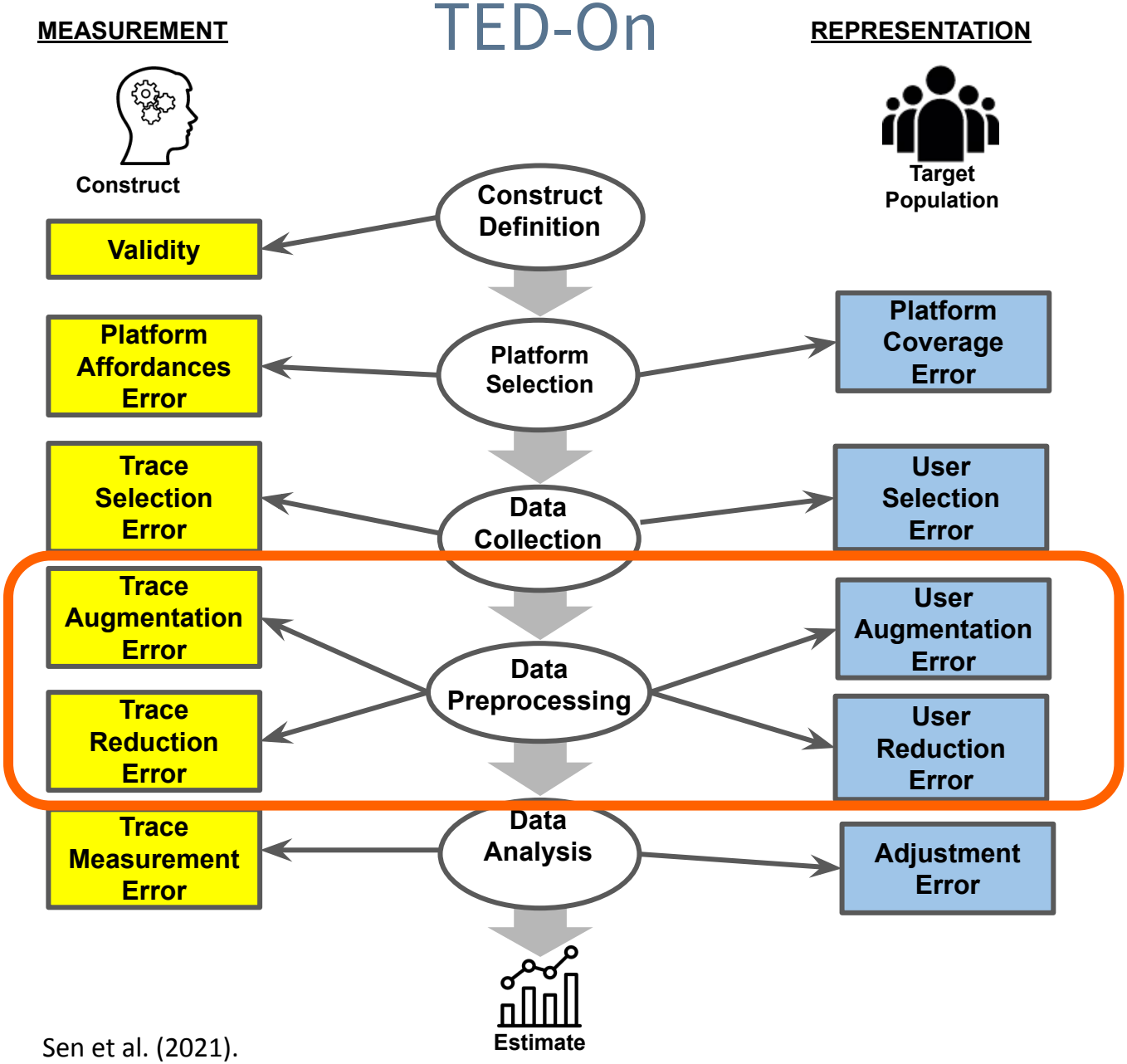
using keywords or search queries and an API

Some posts chosen may not be relevant to the construct

Selection of language for search terms may influence which parts of a user community are left out

Sen et al. (2021).

TED-On



Sen et al. (2021).

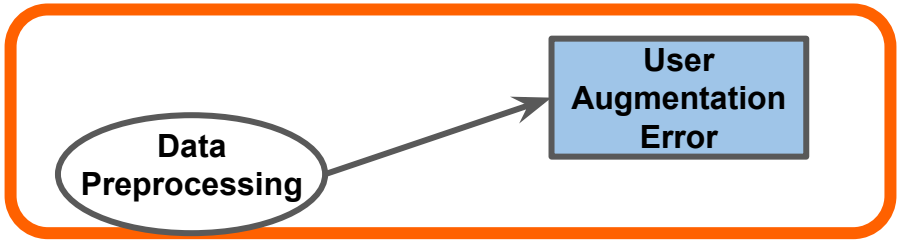
TED-On

MEASUREMENT

Construct

REPRESENTATION

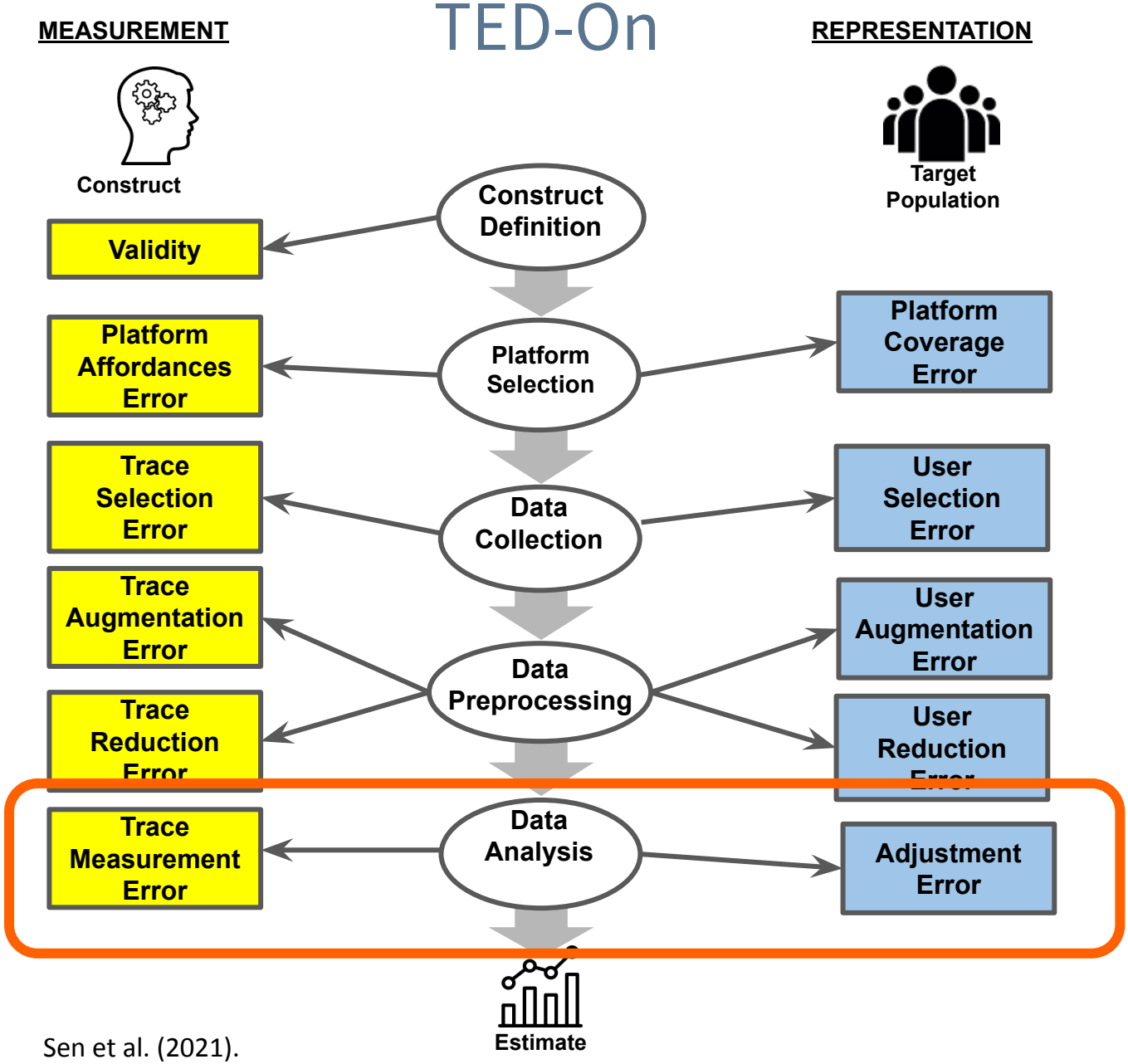
Target Population



demographic data is often important for social science research questions.
But very difficult to obtain from social media data

Sen et al. (2021).

TED-On



Sen et al. (2021).

TED-On

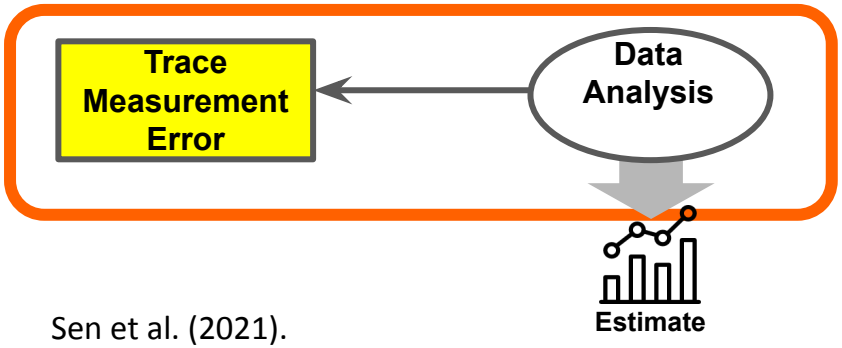
MEASUREMENT

Construct

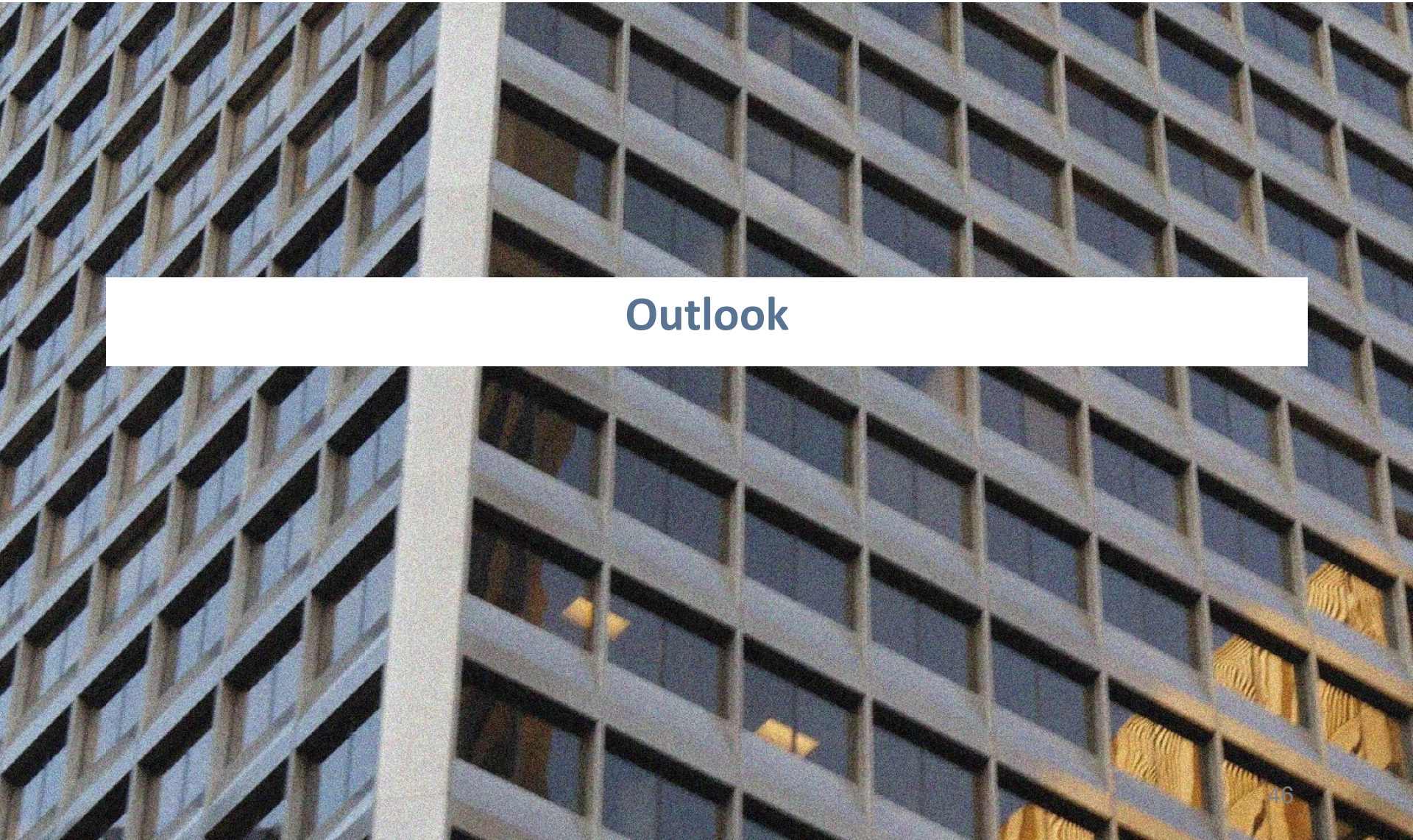
REPRESENTATION

Target Population

Based on the modeling techniques used and the assumptions made, we could have errors dues to data analysis



Sen et al. (2021).



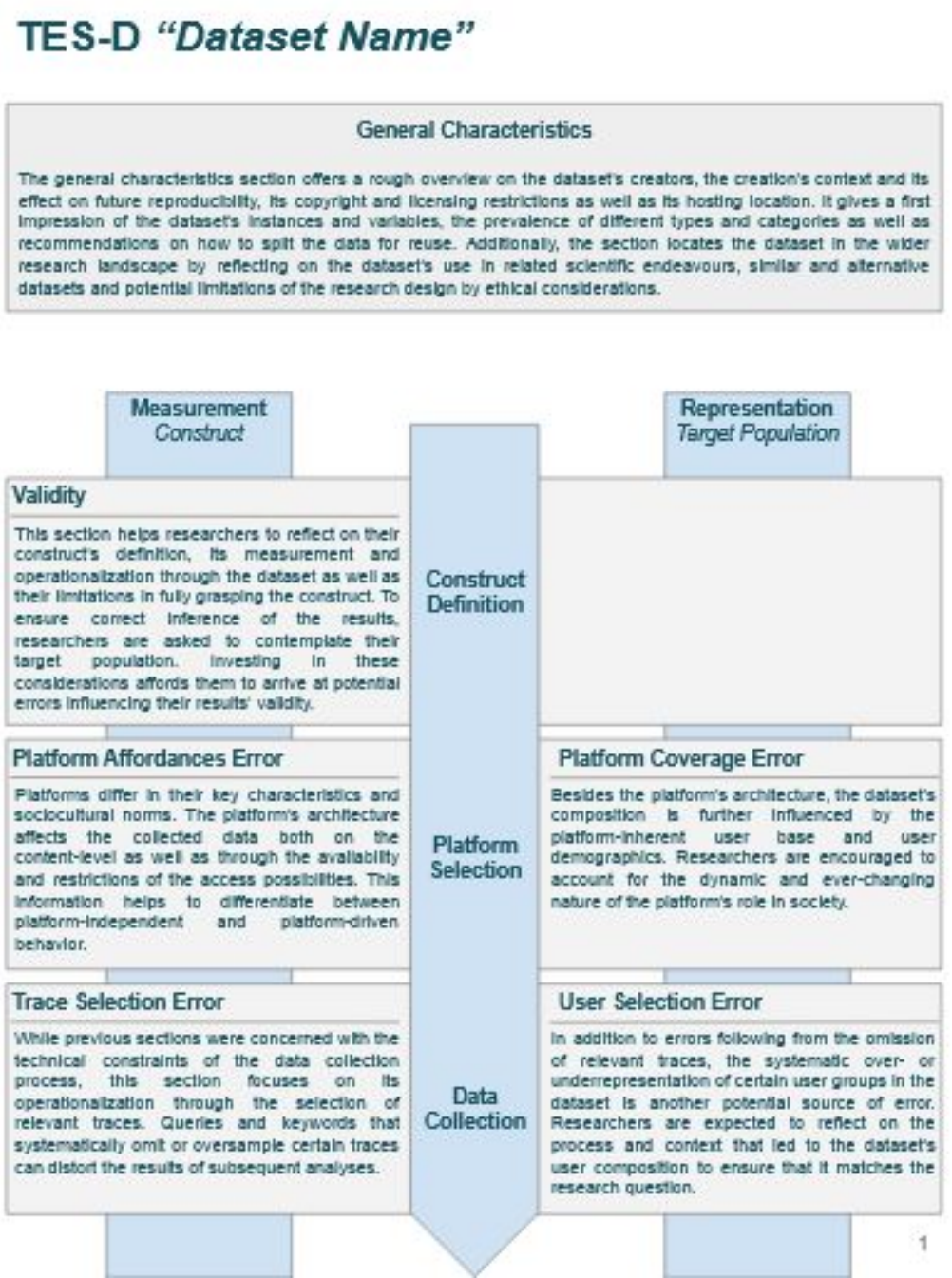
Outlook

Next steps

- Combinations: Bring different error frameworks together, find linking points between frameworks as well as with **existing documentation and data management approaches**
- Specifications: develop new sub-frameworks for specific types of research data or specific application areas.

Template for **Total Error Sheets for Datasets (TES-D):** Documenting errors datasets that use digital traces like social media

(Fröhling et al., in prep)



Error Frameworks

should in the future be complemented by

- Considerations on research ethics for each step in the framework
- Practical guidance for documentation of decision steps and potential error sources
- Standards and best practices for certain steps in the research process

Conclusions

- Research community is aware of challenges and pitfalls in social media research – but lacks initiatives to structurally reflect on them.
- Implementing structured approaches to e.g. review processes, methods sections can raise awareness and guide discussions.

- Accept limitations as natural elements of social media research – focus on making them transparent and understandable.

Further Reading

- Amaya, A., Biemer, P., & Kinyon, D. (2020). Total Error in a Big Data World: Adapting the TSE Framework to Big Data. Journal of Survey Statistics and Methodology, 8(1), 89-119.
- Boyd, D., & Crawford, K. (2012). Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon. Information, Communication & Society, 15(5), 662-679.
- Howison, J., Wiggins, A., & Crowston, K. (2011). Validity Issues in the Use of Social Network Analysis with Digital Trace Data. Journal of the Association for Information Systems, 12(12), 767-797.
- Olteanu, A., Castillo, C., Diaz, F., & Kiciman, E. (2019). Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. Frontiers in Big Data 2(13).
- Sen, I., Flöck, F., Weller, K., Weiß, B., & Wagner, C. (2021). A Total Error Framework for Digital Traces of Human Behavior on Online Platforms. Public Opinion Quarterly 85 (S1): 399-422.
- Tufekci, Z. (2014). Big Questions for Social Media Big Data: Representativeness, Validity and other Methodological Pitfalls. Eighth International AAAI Conference on Weblogs and Social Media

Thanks to our colleagues

Acknowledgements

This presentation is based on joint work with many (former) colleagues at different Depts. at GESIS. Special thanks to:

our collaborators for TED-On Framework:

Fabian Flöck, Bernd Weiß, & Claudia Wagner (GESIS)

- Sen, I., Flöck, F., Weller, K., Weiß, B., & Wagner, C. (2021). A Total Error Framework for Digital Traces of Human Behavior on Online Platforms. Public Opinion Quarterly. Volume 85, Issue S1, 2021, Pages 399–422, https://doi.org/10.1093/poq/nfab018

our collaborators for TES-D Documentation:

Leon Fröhling, Leonie Steinbrinker, Felix Soldner, Maria Zens (GESIS)

Katrin's co-author for studying social media researchers' practices:

Katharina Kinder-Kurlanda (University Klagenfurt)

- Kinder-Kurlanda, K.E., & Weller, K. (2020). Perspective: Acknowledging Data Work in the Social Media Research Lifecycle. Frontiers in Big Data 3:509954. doi: 10.3389/fdata.2020.509954 https://www.frontiersin.org/articles/10.3389/fdata.2020.509954/full

Thank you and
virtual greetings from Köln (Cologne)

Longer form workshop on related themes:

https://training.gesis.org/?site=pDetails&child=full&pID=0x20D86302D1294F288029170F57B3E6D5

gesis training

Search Search GESIS Training

About us ▾ What we offer ▾ Courses & Registration ▾ Other Formats ▾ Covid-19

Introduction to Using Social Media Data for Research: Potentials and Pitfalls

Lecturer(s): Indira Sen, Dr. Katrin Weller

About the lecturer - Indira Sen

About the lecturer - Dr. Katrin Weller

Course description

Please note: There is an additional session on the 12th - 13th December 2022. Please check the schedule for further information.

In this workshop, we provide an introductory overview of the possibilities and limitations of using data collected from social media platforms for research, structured along a theoretical framework and illustrated with practical examples.

register now

About

Date:
05.12 - 06.12.2022

Location:
Online via Zoom

General Topics
Data Analysis

Courselevel
Beginner

Format

Questions welcome!

katrin.weller@gesis.org, indira.sen@gesis.org

gesis Leibniz Institute for the Social Sciences

Member of Leibniz Association

Katrin Weller also acknowledges support from CAIS Center for Advanced Internet Research

CAIS RESEARCH FOR THE DIGITAL AGE