gesis Leibniz Institute for the Social Sciences

Social-Media und Text-Mining mit R

Veronika Batzdorfer

Teil I

Herausforderungen mit Social-Web-Daten, 29.09.2022

Member of Leibniz Association

gesis Leibniz Institute for the Social Sciences

Schedule

	Uhrzeit	Inhalt
*	09:00 - 10:30	Konzepte & Herausforderungen bei der Analyse von Social-Web-Daten (Twitter-API)
	10:30 - 11:00	*Kaffeepause*
	11:30 - 12:30	Getting Started mit Twitterdaten: (i) Sampling, (ii) Pre-Processing & (iii) Grundlagen der Textanalyse (Häufigkeiten, Co-Occurences, Netzwerke)
	12:30 - 13:30	*Mittagspause*
	13:30 - 15:00	Twitter Demo & Exkurs Crawling Social-Web-Data
	15:00 - 15:30	*Kaffeepause*
	15:30 - 17:00	Ausblick: Fortgeschrittene NLP-Techniken (z.B. Topic Modelling) & Social-Web-Data-Collection; Bias und Ethik im NLP

https://github.com/nika-akin/-Social-Media-and-Text-Mining-Workshop-2022

Member of Leibniz Association

gesis Leibniz Institute for the Social Sciences

About me

Postdoktorand im

- Team *Digital Society Observatory* (CSS)
- Team *Survey Data Augmentation* (SDC)

Forschungsinteressen

- Verknüpfen von digitale Verhaltensspurdaten mit Surveydaten
- Zeitliche Dynamiken auf Social-Media-Plattformen
- Kausale Identifikation

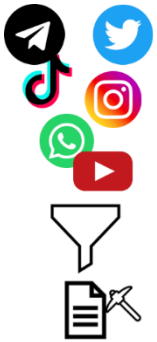
veronika.batzdorfer@gesis.org

About you

- Wie ist Ihr Name?
- Wo arbeiten/studieren Sie?
- Woran arbeiten Sie?

Relevanz von Social-Web-Daten?

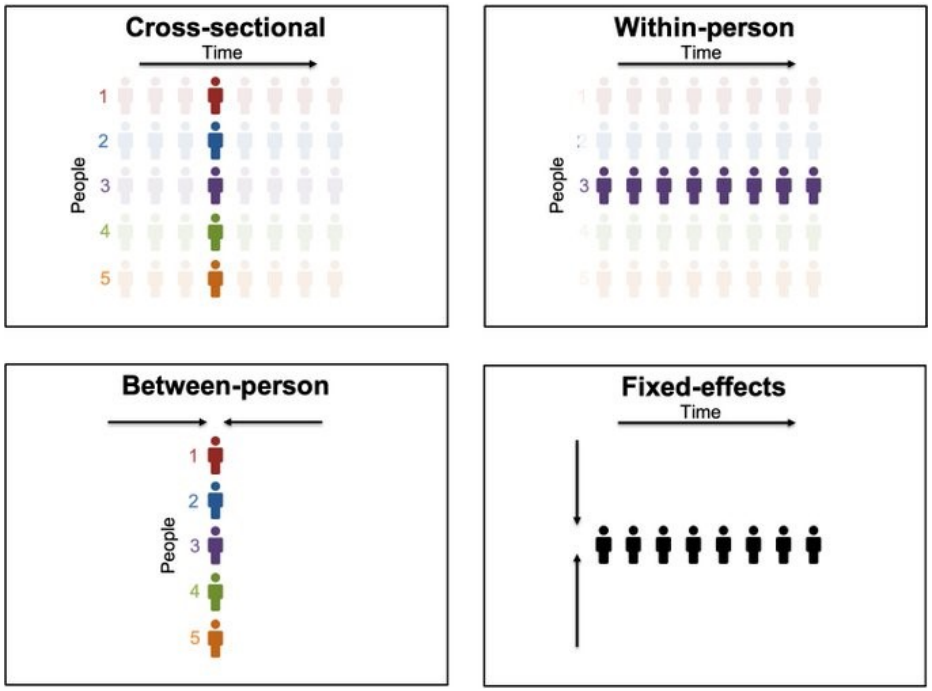
Instrument und **Phänomen**



- psychische Belastung (Guntuku et al., 2017)
- Verbreitung von Fake-News über soziale Medien (Vosoughi et al., 2018)
- Stress (Saha, & De Choudhury, 2017)
- Radikalisierung durch Verschwörungstheorien (Phadke, Samory, & Mitra, 2022)

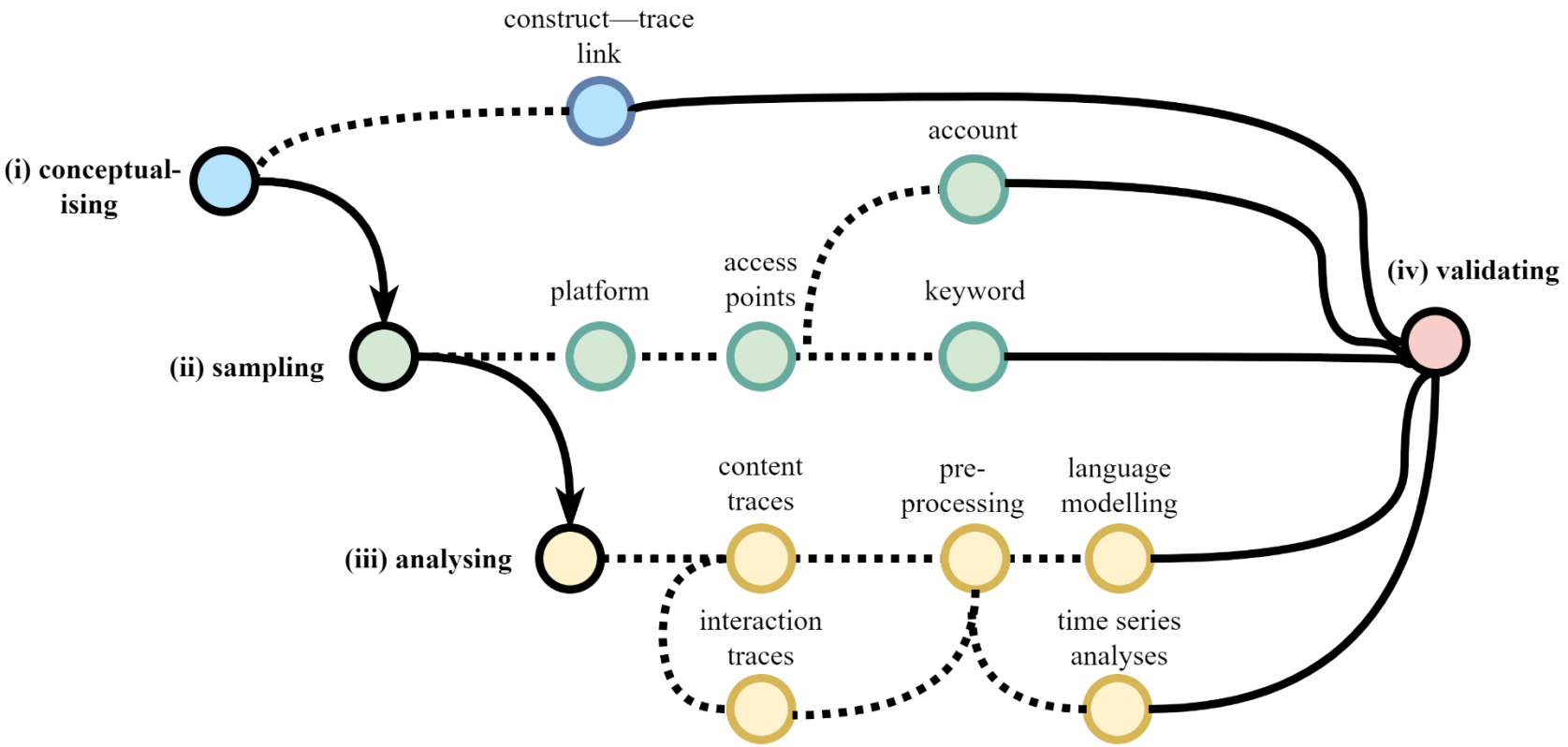
Social-Web-Daten für die Forschung

- 1. Einfach zugänglich (via Access Points wie *Twitter APIs*)
- 2. Beobachtbar
- 3. Skalierbar
- 4. Vielfältige Inhalts- und Interaktionsspurdaten (und Metadaten)

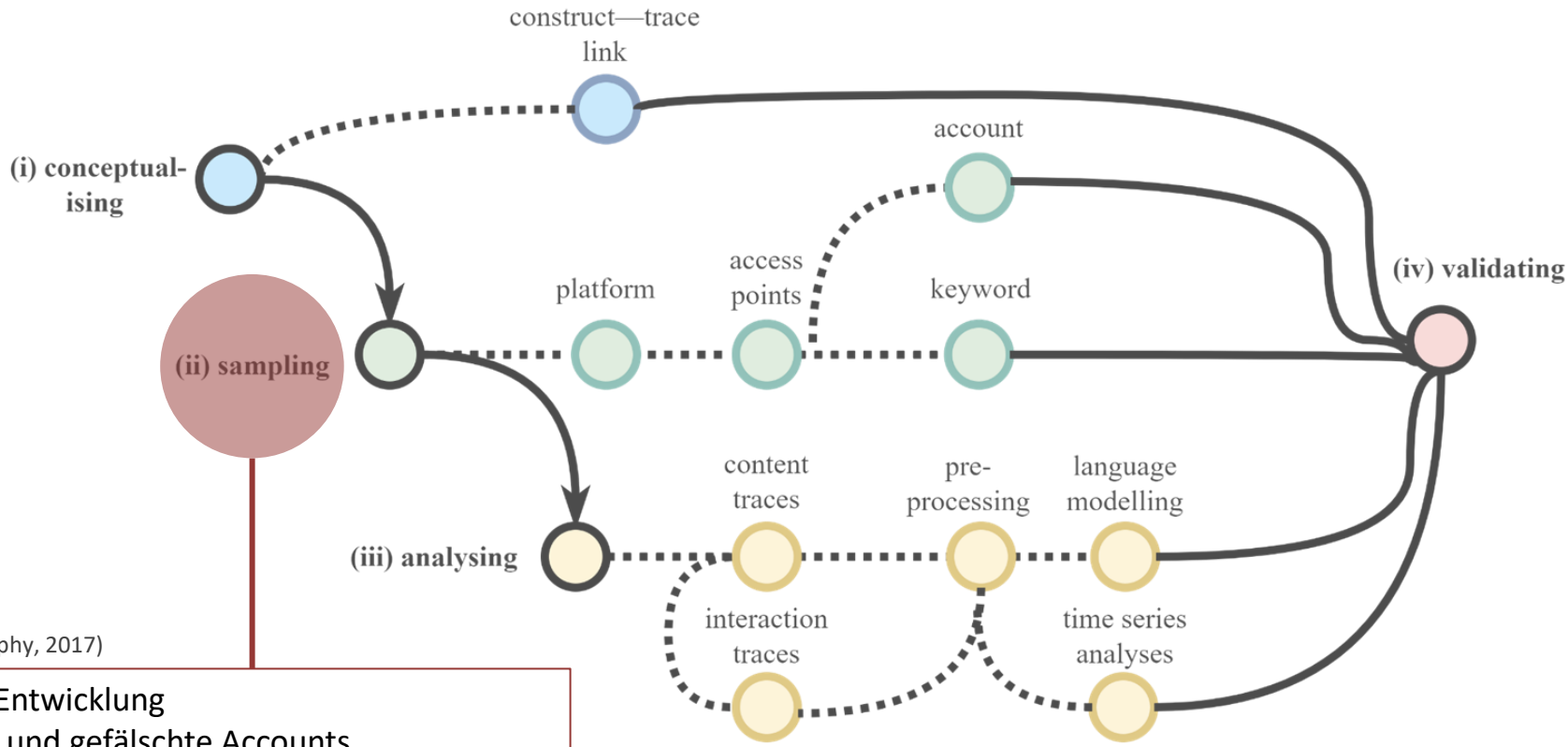


Quelle:
https://twitter.com/SachaEpskamp/status/1376122442599501826/photo/1

Bias im Forschungszyklus



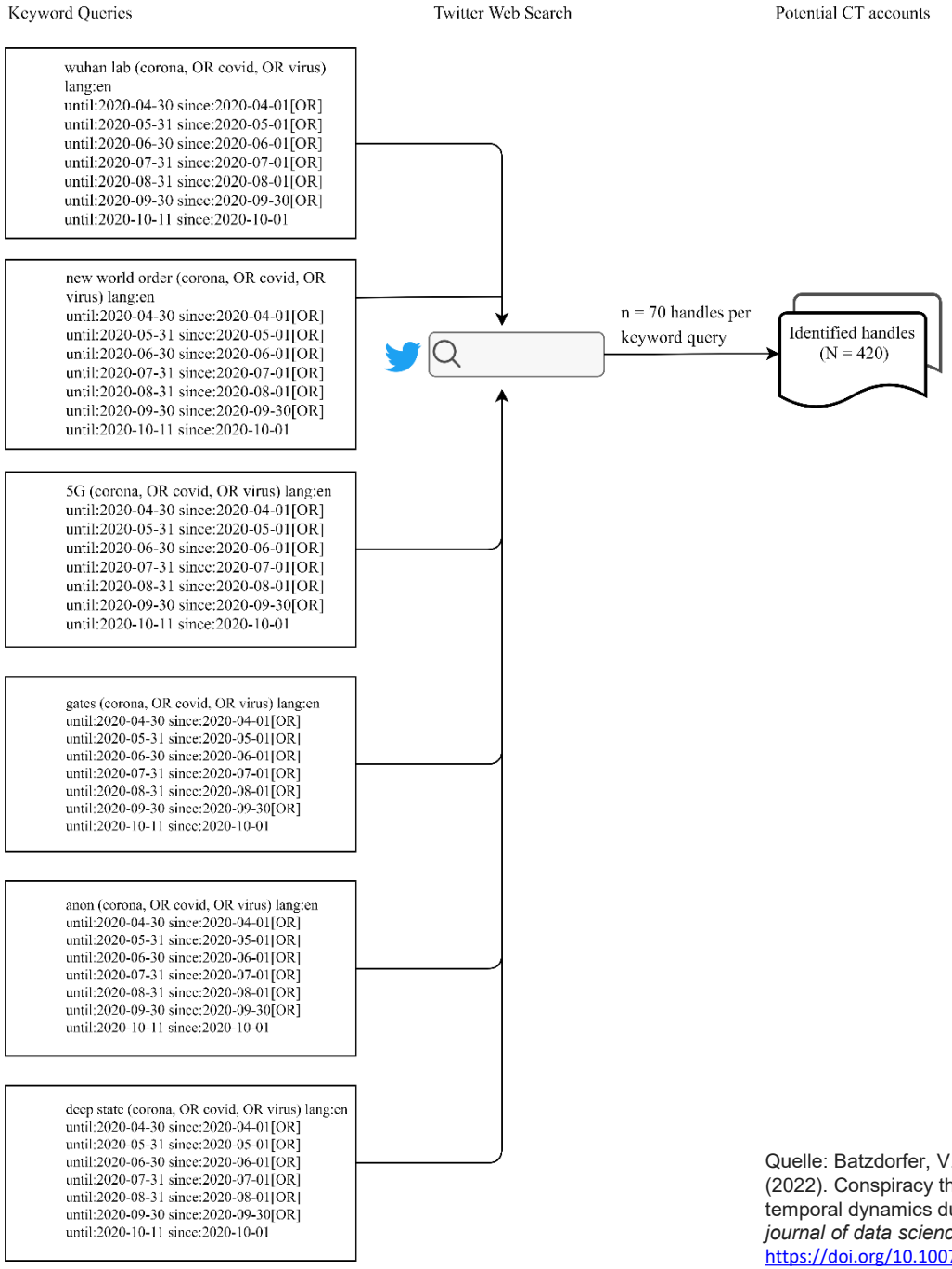
Bias im Forschungszyklus



(z.B., Murphy, 2017)

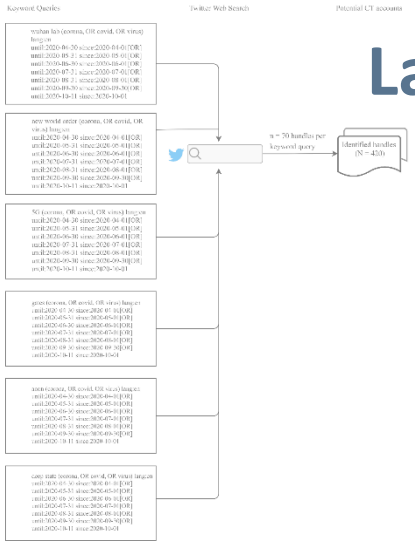
- API-Entwicklung
- Bots und gefälschte Accounts
- Mehrsprachige Daten, nicht-lateinische Schriftzeichen, Bilder
- Datenreliabilität und Repräsentativität
- Abgleich mit externen Daten (z. B. Surveys)
- Inferenz von Geolokalisierung

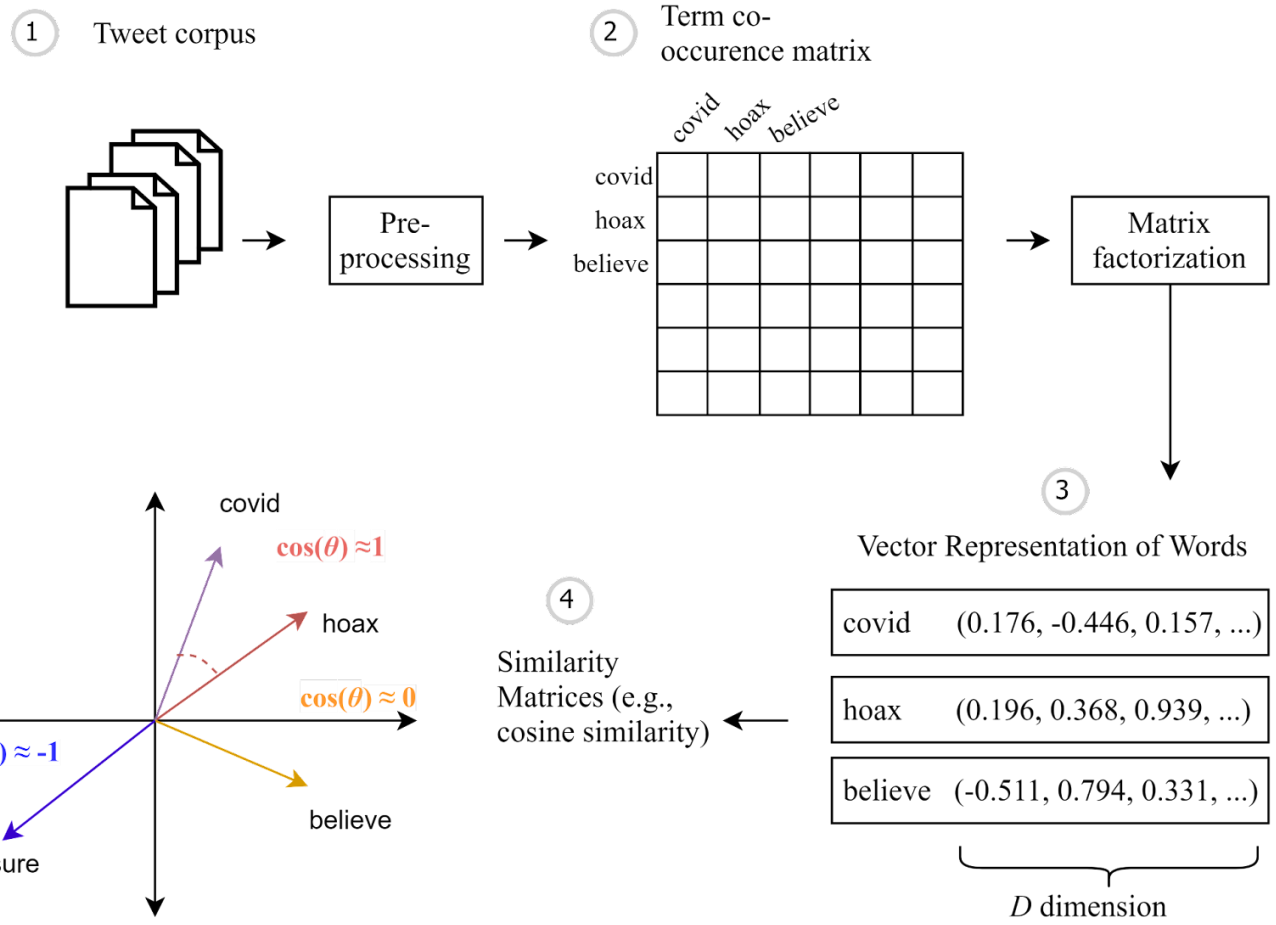
API-Harvesting



Quelle: Batzdorfer, V., Steinmetz, H., Biella, M., & Alizadeh, M. (2022). Conspiracy theories on Twitter: emerging motifs and temporal dynamics during the COVID-19 pandemic. *International journal of data science and analytics*, *13*(4), 315-333. https://doi.org/10.1007/s41060-021-00298-6

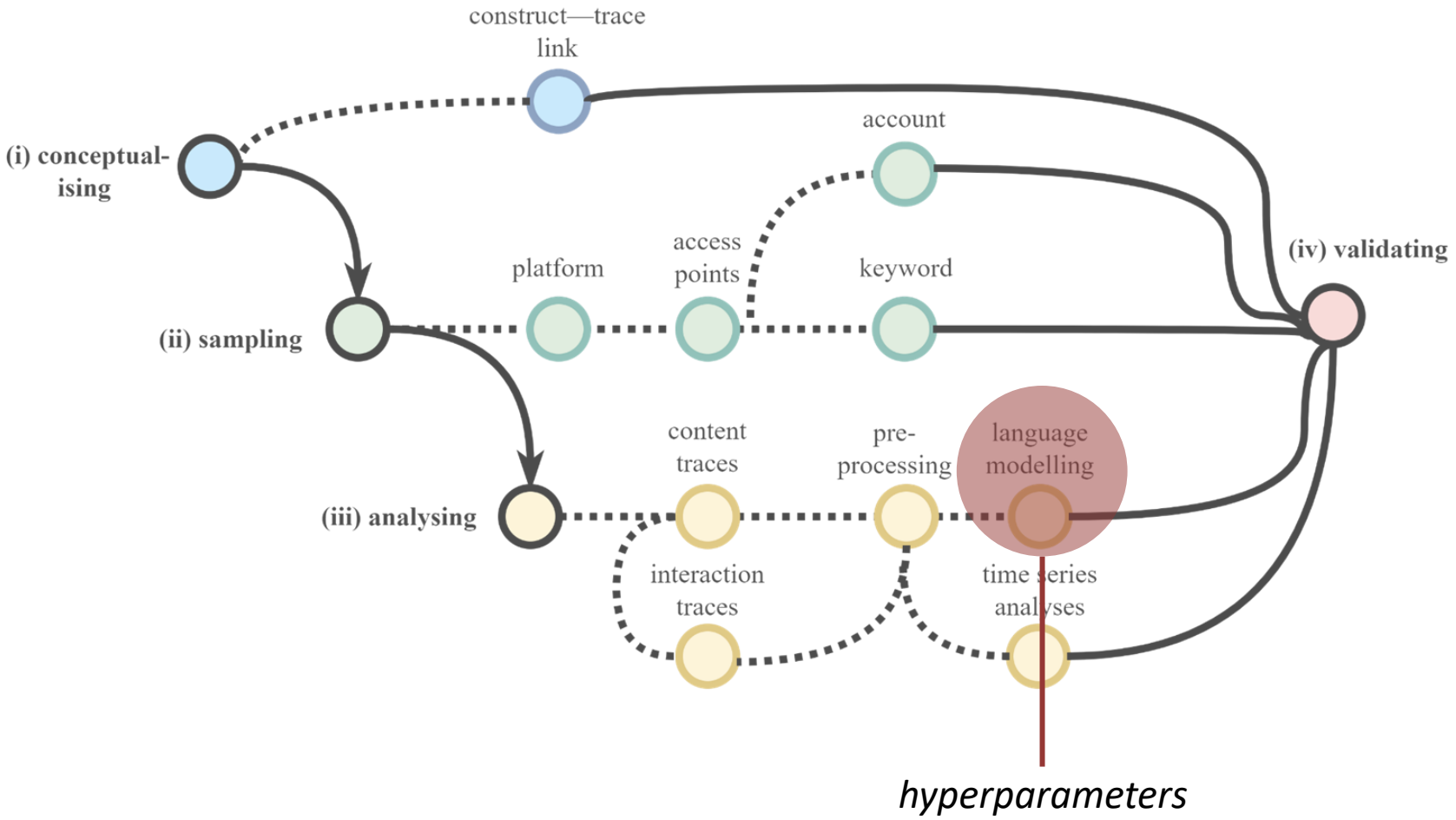
Language Modelling & Verschwörungstheorien



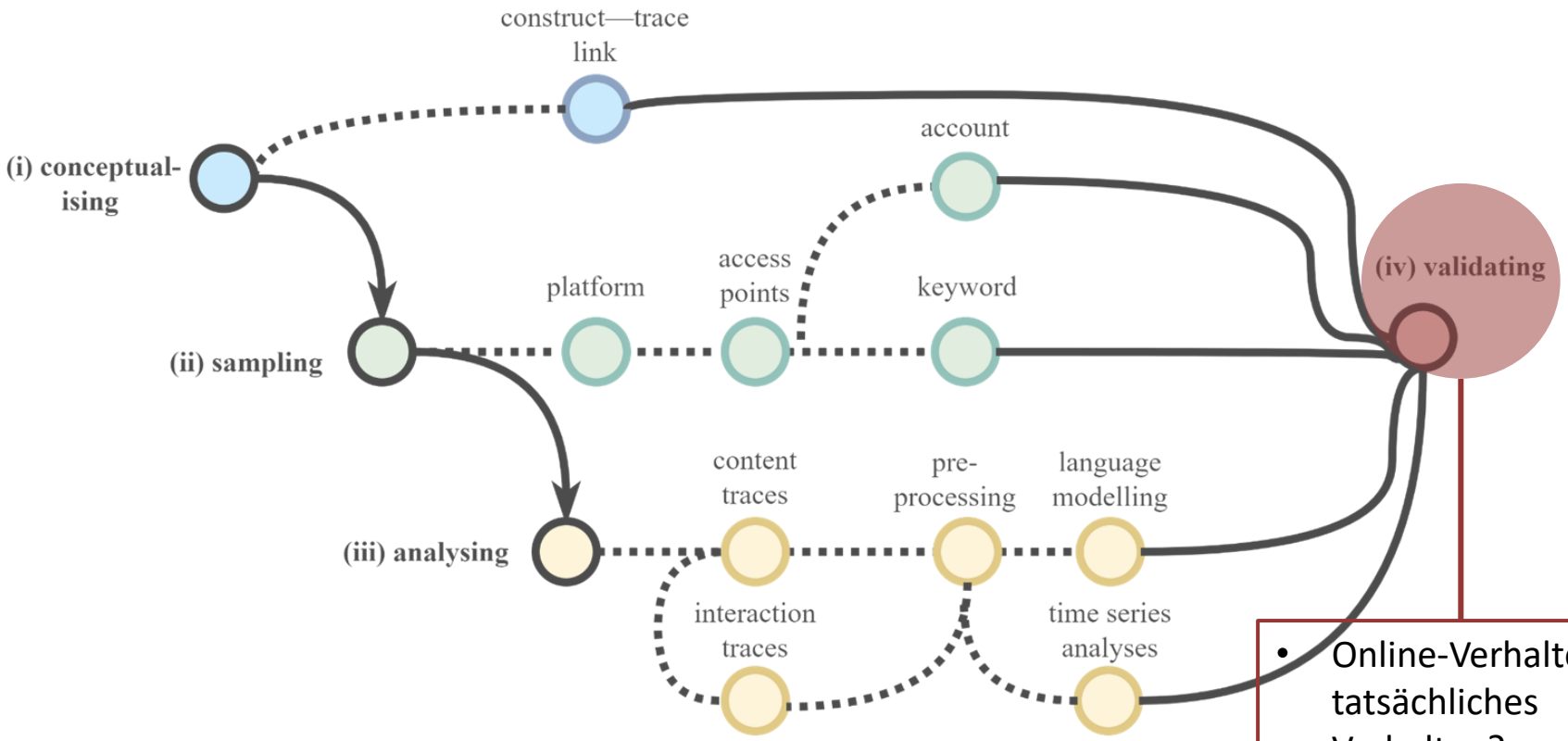


Quelle: Batzdorfer, V., Steinmetz, H., Biella, M., & Alizadeh, M. (2022). Conspiracy theories on Twitter: emerging motifs and temporal dynamics during the COVID-19 pandemic. *International journal of data science and analytics*, *13*(4), 315-333. https://doi.org/10.1007/s41060-021-00298-6

Bias im Forschungszyklus

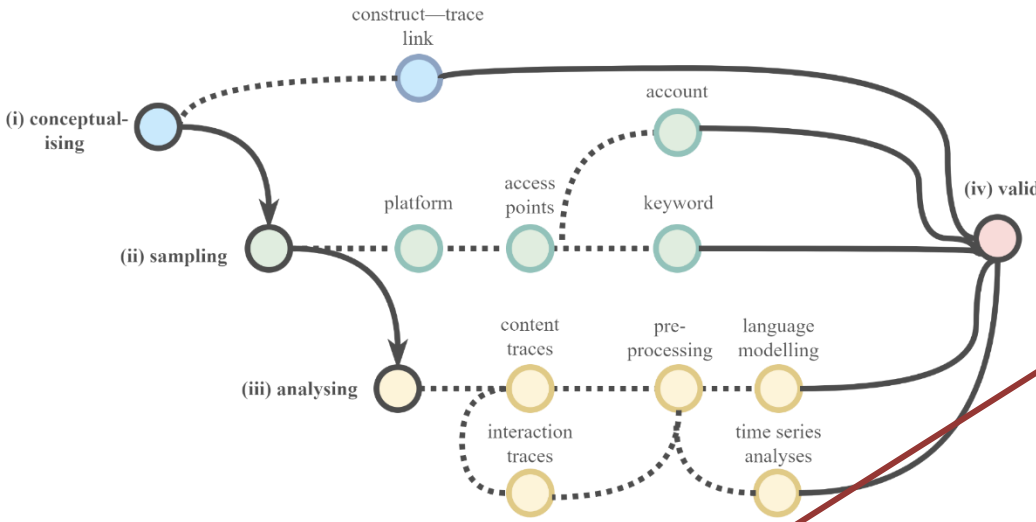


Bias im Forschungszyklus



- Online-Verhalten = tatsächliches Verhalten?
- Mangel Benchmarks zur Überprüfung der Datenqualität

Bias im Forschungszyklus



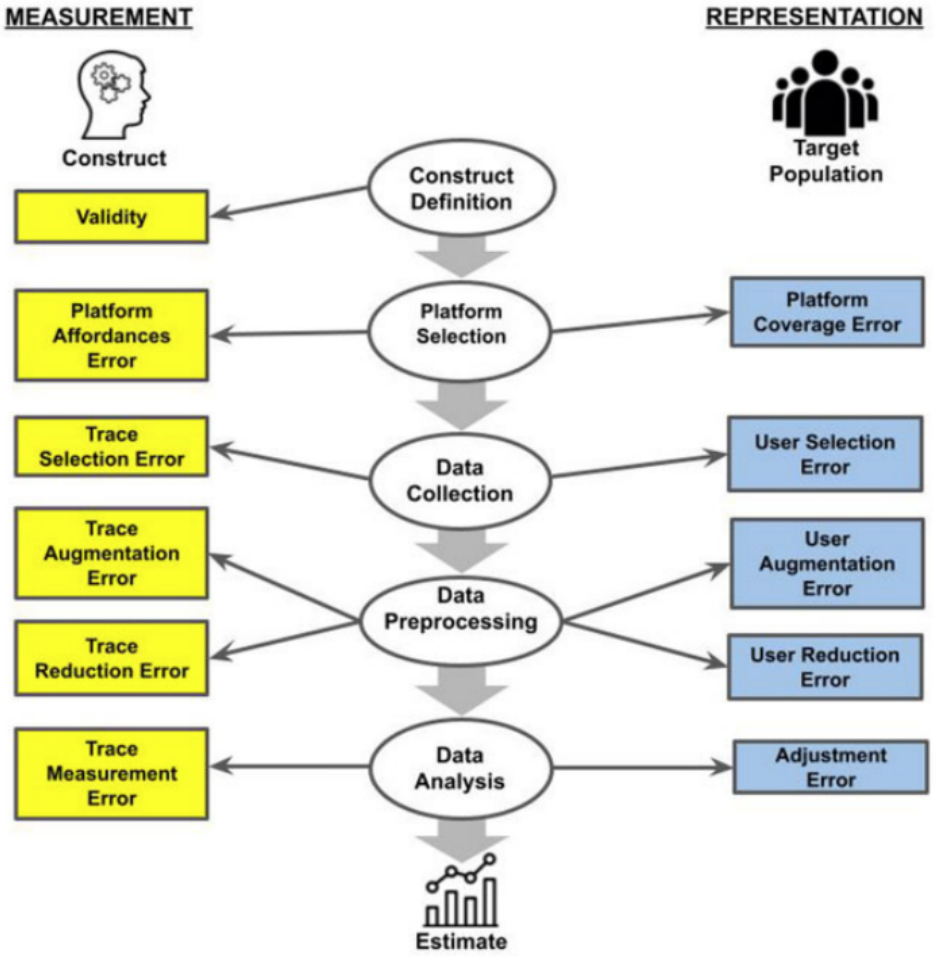
Open-Science

- Bereitstellung von "rehydrierbaren" Datensätzen
- Synthetische Daten

Ethik

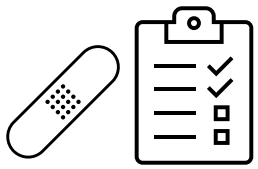
- Kann die Veröffentlichung von Daten die Nutzer gefährden?
- Sind die Nutzer gefährdet und die Inhalte sensibel?
- Tweet/Account zum Zeitpunkt der Erstellung gelöscht?

The Total Error Framework for Digital Traces of Humans (TED)



Quelle: 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.

Bias im Forschungszyklus



Erstellen Sie ein Social-Web-Data-*Fallbeispiel** und versuchen Sie mögliche Fehlerquellen im Forschungszyklus:

- zu benennen
- zu quantifizieren
- zu adressieren

* z.B. Wie würden Forschende die COVID-19-Prävalenz in einer nationalen Bevölkerung anhand digitaler Spuren untersuchen?

Bibliographie

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, *64*(12), 86-92. http://dx.doi.org/10.1145/3458723

Hsieh, Y. P., & Murphy, J. (2017). Total twitter error. *Total survey error in practice*, *74*, 23-46.

Hullman, J., Kapoor, S., Nanayakkara, P., Gelman, A., & Narayanan, A. (2022). The worst of both worlds: A comparative analysis of errors in learning from data in psychology and machine learning. *arXiv preprint arXiv:2203.06498*.

Olteanu, A., Castillo, C., Diaz, F., & Kıcıman, E. (2019). Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, *2*, 13. https://doi.org/10.3389/fdata.2019.00013

Ruths, D., & Pfeffer, J. (2014). Social media for large studies of behavior. *Science*, *346*(6213), 1063-1064. https://doi.org/10.1126/science.346.6213.1063

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. https://doi.org/10.1093/poq/nfab018

Tufekci, Z. (2014, May). Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In *Eighth international AAAI conference on weblogs and social media*.