

# gesis

Leibniz-Institut  
für Sozialwissenschaften



## Research Ethics and Data Protection in Social Media Research

Meet the Experts! – GESIS online talks

*Oliver Watteler & Katrin Weller* · Sep 16, 2021

# Speakers



## Oliver Watteler

- Senior researcher, team Data Acquisitions and Access
- Master in History, Political Science and Philosophy
- Research data management, legal aspects (focus: data protection)
- Contact: [oliver.watteler@gesis.org](mailto:oliver.watteler@gesis.org)



## Dr. Katrin Weller

- Team lead Social Analytics and Services
- PhD in information science
- Social media research methods, altmetrics
- Contact: [katrin.weller@gesis.org](mailto:katrin.weller@gesis.org)

# What is 'research ethics'?

- 'Research ethics'
  - ▶ Moral principles and actions guiding and shaping research
    - from inception to completion,
    - through dissemination and sharing of findings,
    - including archiving and future use.
- Research ethics in the social sciences
  - ▶ Initially 'patient protection' model of medical research
  - ▶ Today broader scope including consideration of benefits, risks and harms to all persons connected with and affected by the research
  - ▶ Including social responsibilities of researchers

# What is ‘data protection’?

- Data protection
  - ▶ part of fundamental right to privacy (or ‘informational freedom’)
- “Privacy is a personal condition of life characterised by seclusion from, and therefore absence of acquaintance by, the public.” (Neethling 2005)
- Prevention of unwanted disclosure of personal information or the misuse of such information
  - ▶ core of data protection
- Legal framework in the European Union
  - ▶ Charter of Fundamental Right of the EU (Art. 8)
  - ▶ GDPR
  - ▶ National and sub-national data protection acts
  - ▶ Specialized laws

# Principles relating to processing of personal data (Art. 5 GDPR)

Art. GDPR	Topic	Meaning
Art. 5 I (a)	Lawfulness	Data must be processed in a legal way (Art. 6) and transparent for ‘data subjects’; no surprises or covert activities.
	Fairness	
	Transparency	
Art. 5 I (b)	Purpose limitation	Data may only be collected “for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes”; research exemption: research seen as in line with initial purposes.
Art. 5 I (c)	Data minimisation	Limit amount of data collected.

# Principles relating to processing of personal data (Art. 5 GDPR)

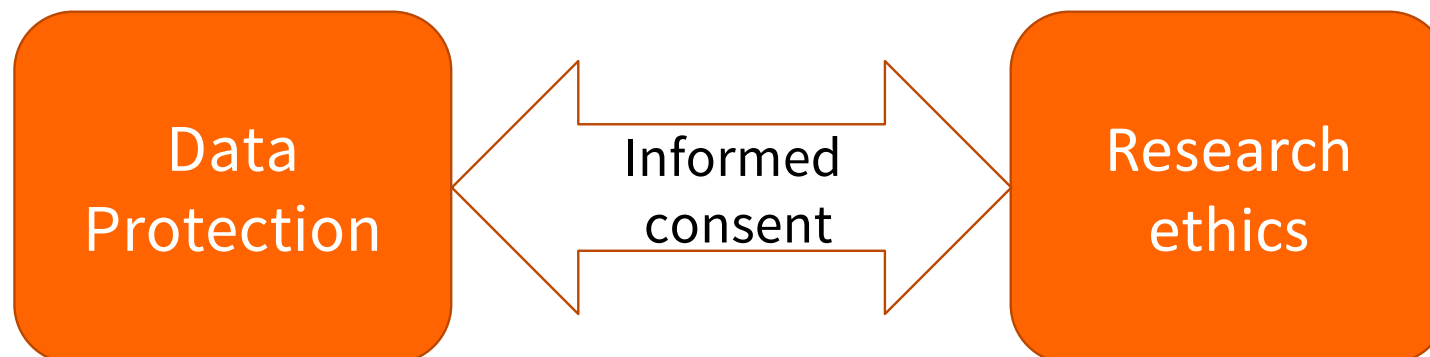
Art.	Topic	Meaning
Art. 5 I (d)	Accuracy	Data collected for a given purpose should be kept correct and deleted or corrected without delay if necessary.
Art. 5 I (e)	Storage limitation	Research exemption: longer period, if “appropriate technical and organizational measures” are implemented.
Art. 5 I (f)	Integrity	Protected against “unauthorized or unlawful processing and against accidental loss, destruction or damage”.
	Confidentiality	
Art. 5 II	Accountability	Controller (or processor) in charge and liable.

# Link between data protection and research ethics: informed consent

Informed consent means for example:

- Information
- Transparency
- Minimal requirement > chance \*not\* to consent

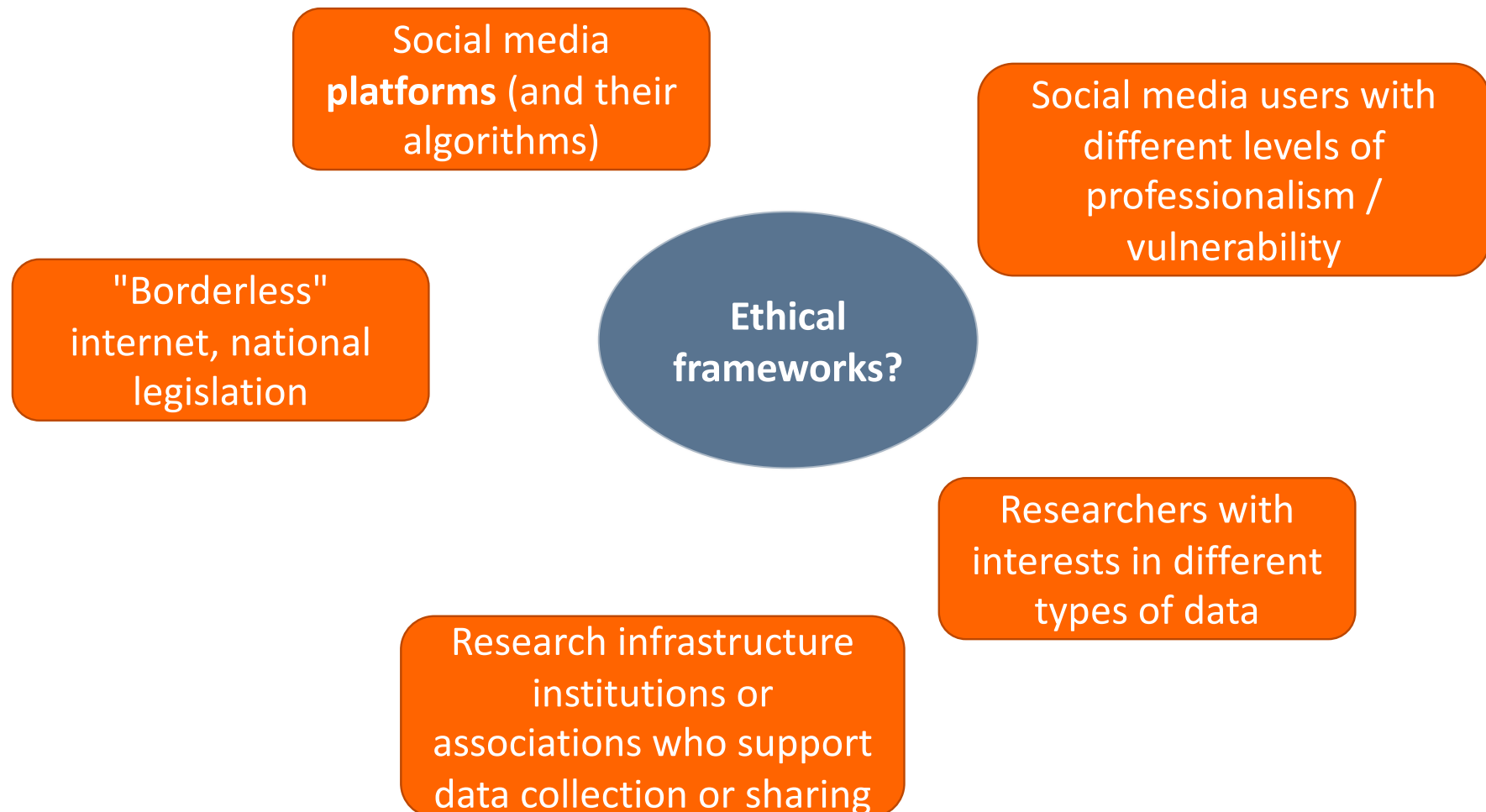
Regularly in social media research: **lack thereof**



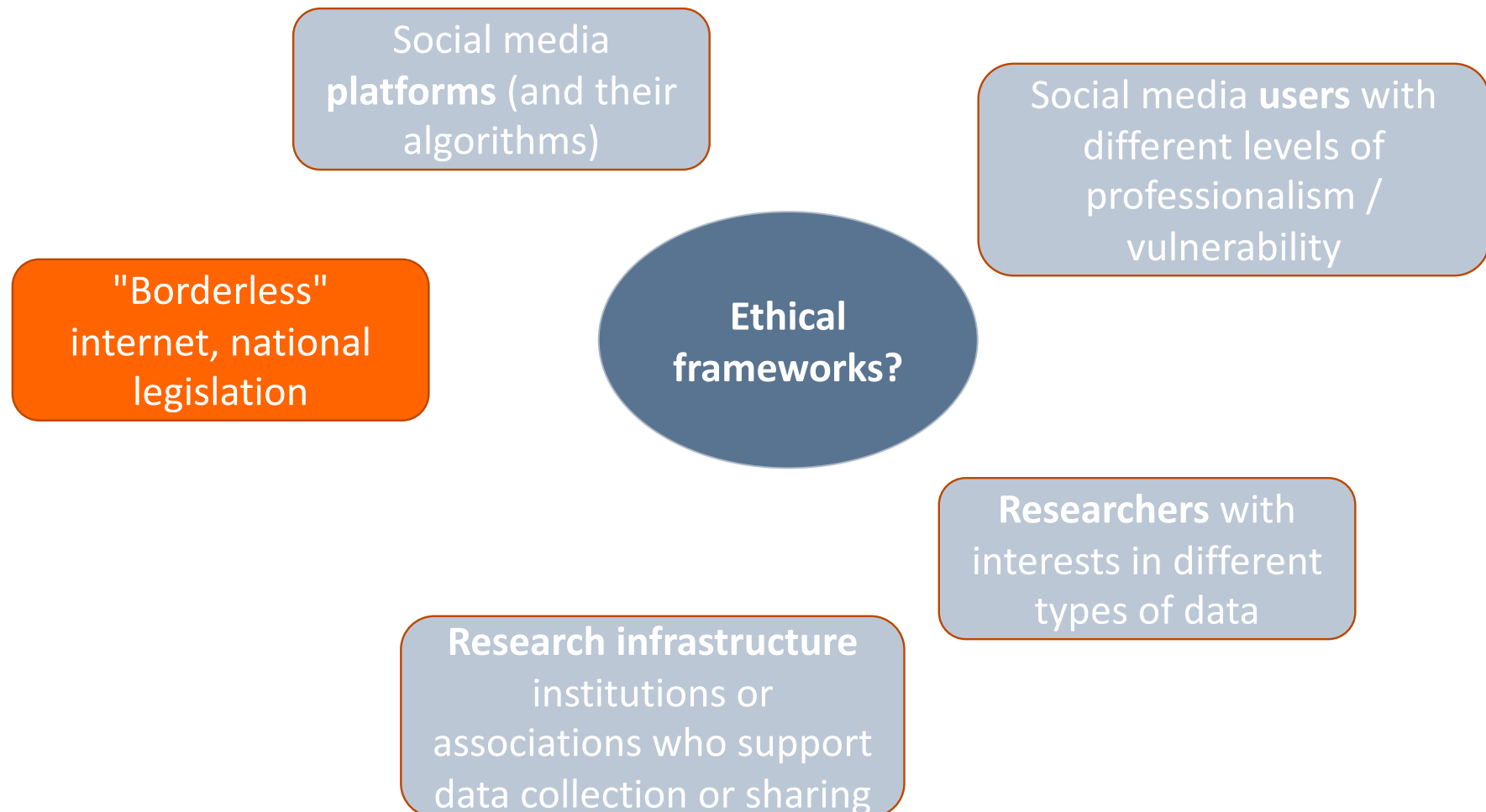
# Actors and Dependencies



# Different entities that effect potential ethical standards in social media research



# Different legal frameworks

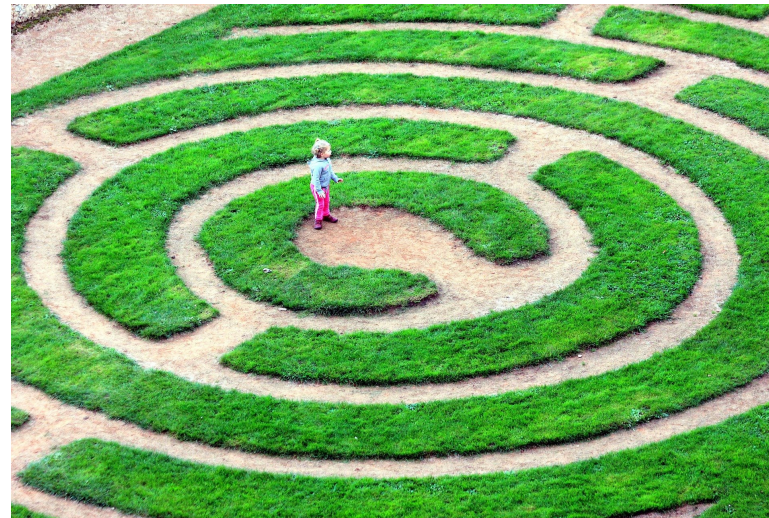


# Facing the maze of ethical and legal challenges

GDPR

Ethical review  
committees

Specialized  
laws  
depending  
on  
research  
purpose



Publishers'  
requirements

Ethical guidelines

Terms of service

# Data protection legislation - overview

- Since 25 May 2018, the EU General Data Protection Regulation (GDPR) applies:
  - ▶ 99 articles and 173 recitals.
  - ▶ Applies directly.
  - ▶ Intended to harmonize data protection law EU-wide.
  - ▶ **BUT** about 150 “opening clauses” or exemptions.
- GDPR (factually) integrated into hierarchy of norms:
  - ▶ Legislation on national (e.g. Federal Data Protection Act) and sub-national level.
  - ▶ Special laws may apply.
  - ▶ Conflict of fundamental rights:  
Freedom of research vs. freedom of personal information.
- **Problem: GDPR catch-all regulation**

# What is ‘personal data’ (Art. 4 (1) GDPR)?

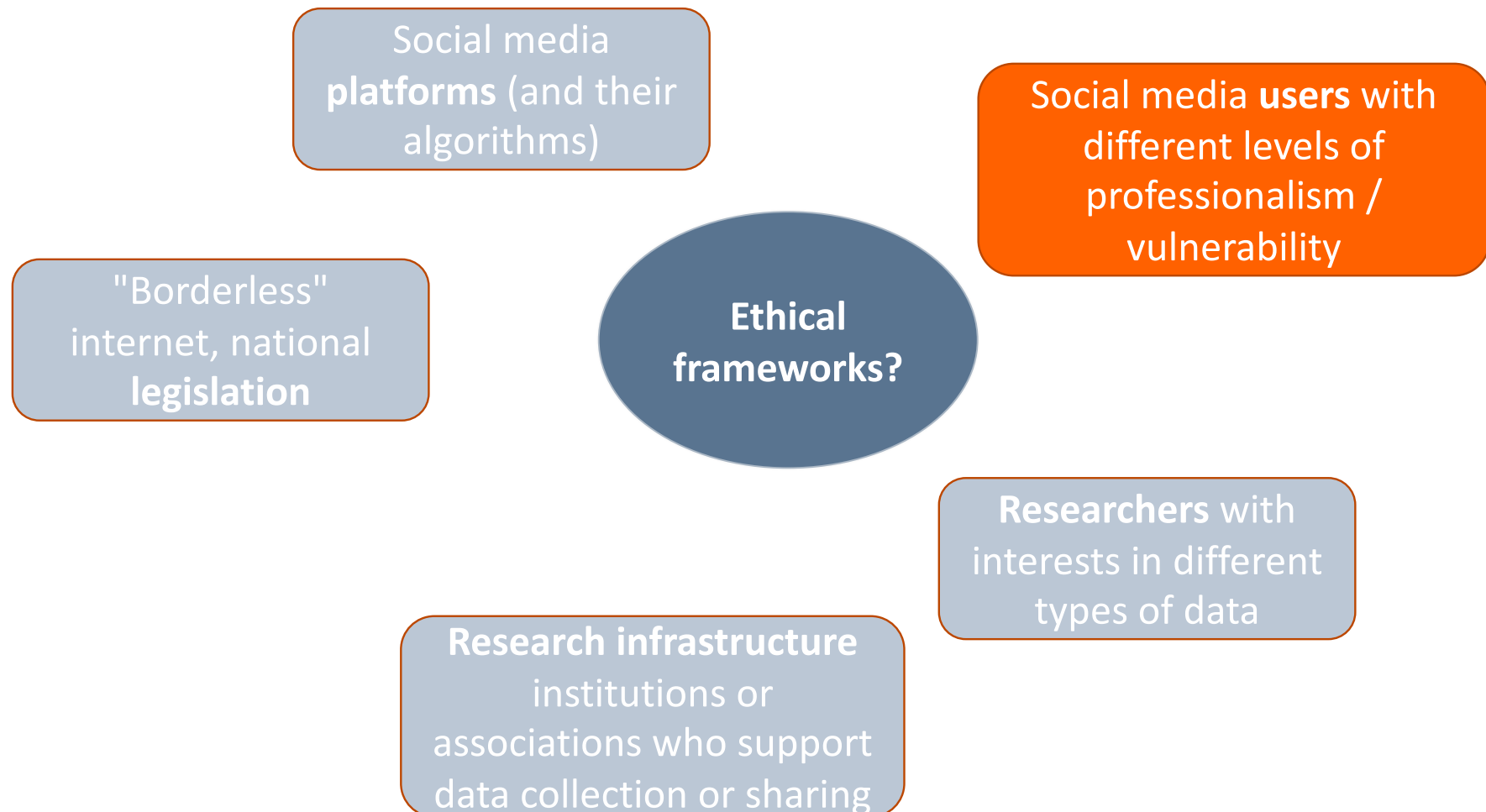
- “(P)ersonal data’ means any information relating to an
  - ▶ identified or
  - ▶ identifiable natural person (‘data subject’);
- an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as
  - ▶ a name,
  - ▶ an identification number,
  - ▶ location data,
  - ▶ an online identifier or
  - ▶ to one or more factors specific to
    - the physical,
    - physiological,
    - genetic,
    - mental,
    - economic,
    - cultural or
    - social identity of that natural person;”

Very broad definition

# Are we talking about ‘personal data’ when it comes to Social Media?

- Examples of ‘personal data’ as identifies in the GDPR and in court decisions:
  - ▶ Facial images
  - ▶ Information on physical or mental health
  - ▶ Geo-locations
  - ▶ Fixed IP addresses (ECJ 2016)
- Problems:
  - ▶ Ubiquity and likability of data > profiling
  - ▶ ‘Entirety’ of data might be revealing
- Answer: most likely “yes”

# Social media users



# Users often unaware of research activities

**Table 2.** Comfort Around Tweets Being Used in Research.

Question	Very uncomfortable	Somewhat uncomfortable	Neither uncomfortable nor comfortable	Somewhat comfortable	Very comfortable
How do you feel about the idea of tweets being used in research? (n = 268)	3.0%	17.5%	29.1%	35.1%	15.3%
How would you feel if a tweet of yours was used in one of these research studies? (n = 267)	4.5%	22.5%	23.6%	33.3%	16.1%
How would you feel if your entire Twitter history was used in one of these research studies? (n = 268)	21.3%	27.2%	18.3%	21.6%	11.6%

Note. The shading was used to provide a visual cue about higher percentages.

Fiesler, C., & Proferes, N. (2018). "Participant" Perceptions of Twitter Research Ethics. *Social Media + Society*, 4(1), 205630511876336. <https://doi.org/10.1177/2056305118763366>



# Not all users are equal

- Celebrities / professional accounts / public figures
- Activists
- Marginalized groups
- Other vulnerable groups (e.g., minors)

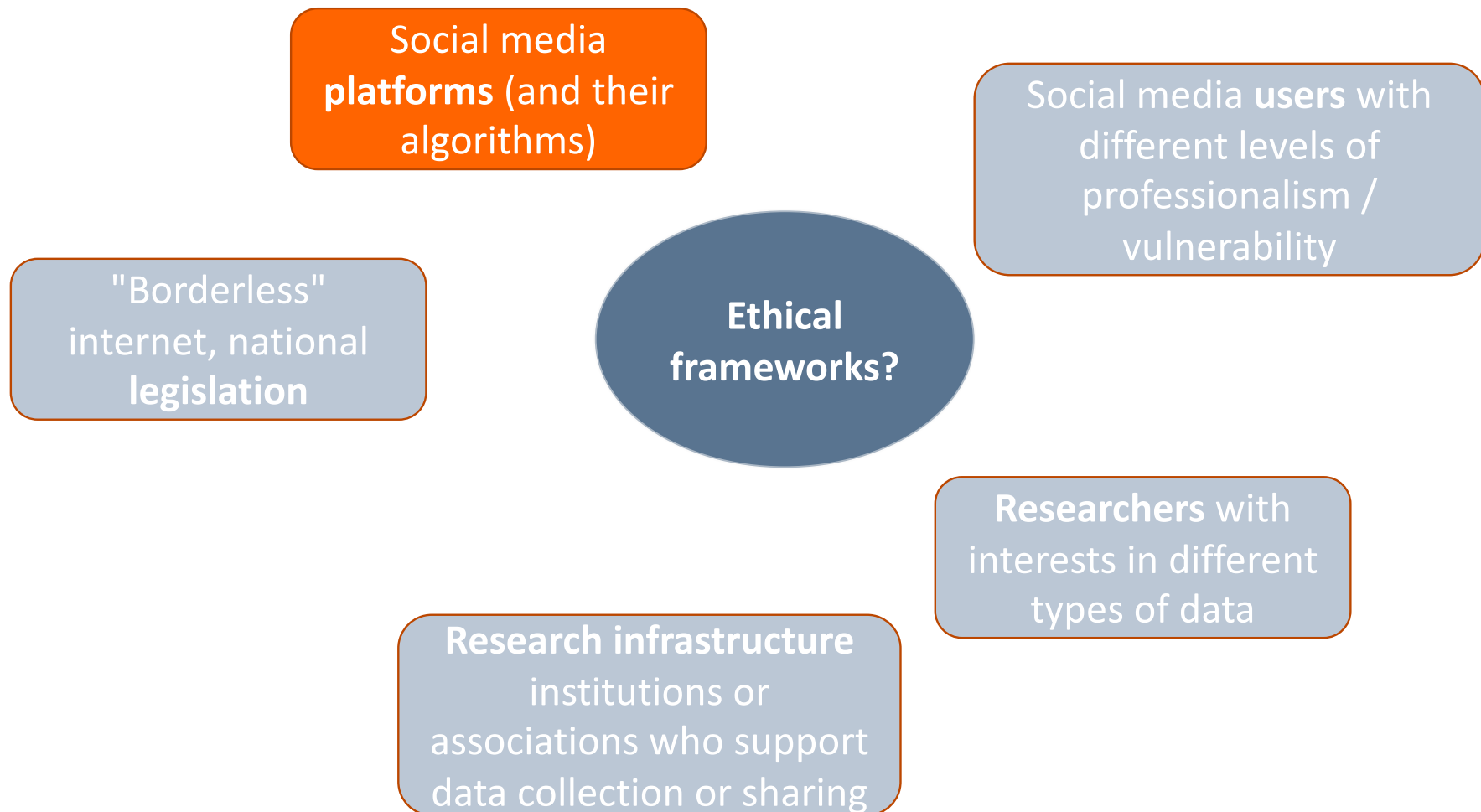
# Collecting data from vulnerable groups

- Research examples from the medical domain illustrate challenges with vulnerable groups, e.g.
  - patient communities
  - related to suicide

Eskisabel-Azpiazu, Amaia; Cerezo-Menéndez, Rebeca; Gayo-Avello, Daniel (2017). An Ethical Inquiry into Youth Suicide Prevention Using Social Media Mining. In: Zimmer & Kinder-Kurlanda (2017).

Ferguson, Robert Douglas (2017). Negotiating Consent, Compensation, and Privacy in Internet Research: PatientsLikeMe.com as a Case Study. In: Zimmer & Kinder-Kurlanda (2017).

# Different platform affordances



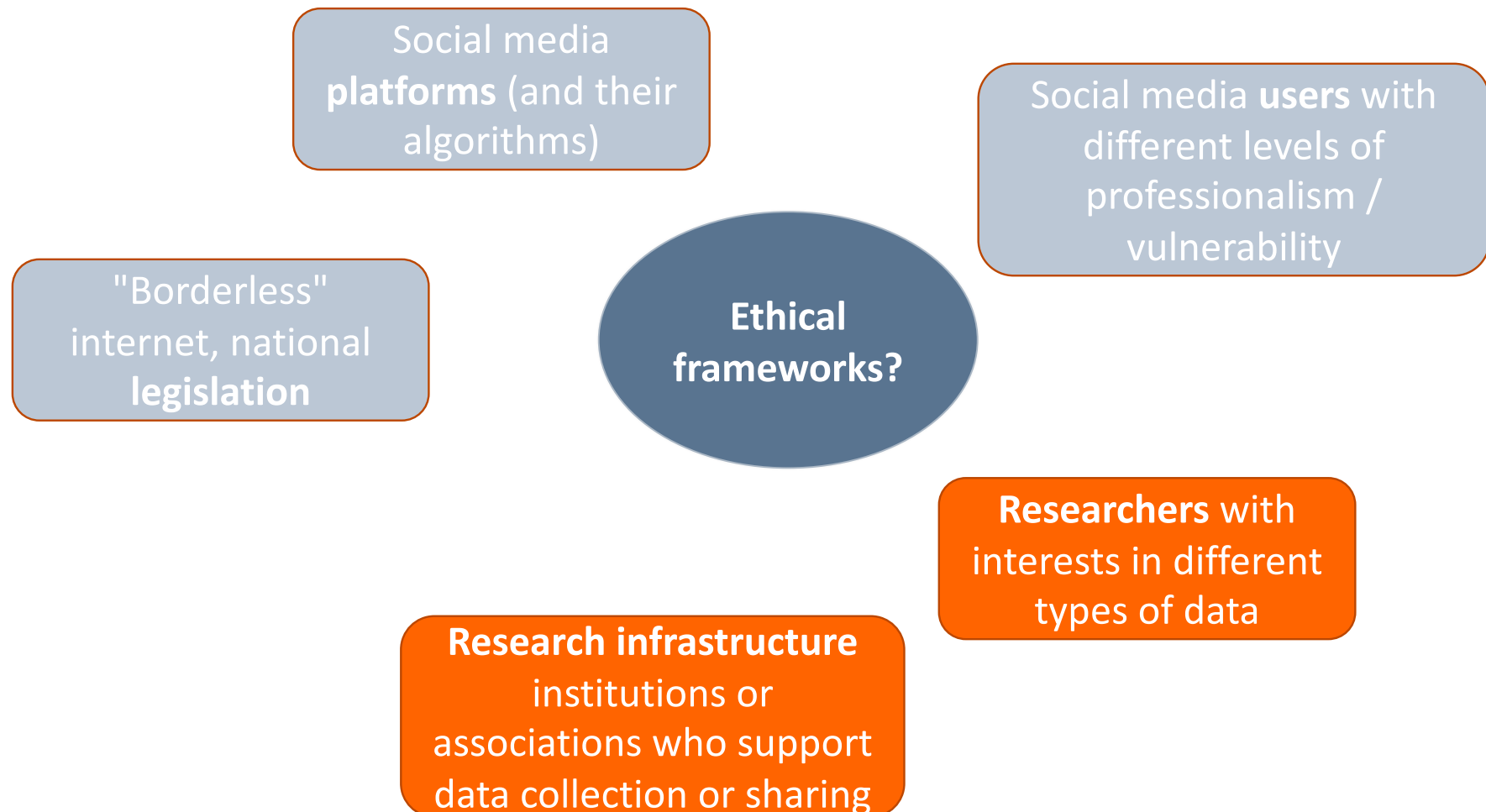
# Not all platforms are equal

- **Users side:** Different options for privacy settings
- **Platform side:** Different ways in which data can be collected from platforms

## **Example: Twitter vs. Facebook**

- Twitter: Simple distinction between either public or protected account. No need for real names. Access options via API.
- Facebook: Complex system of privacy settings that impact visibility of content. Real names requested. (Almost) no access options for researchers.

# Different research approaches



# No formal research field - no standard methods

Diversity of disciplines and approaches.

Lack of standards for

- methods
- documentation / data management
- research ethics

Development of best practices is impacted by the changing nature of social media platforms and their entanglement in broader complexities.

# Guiding Questions & Examples

# Ethical considerations need to be part of the entire research process

- Questions about data protection and research ethics need to be included from the very beginning of a research study.
- Reserve capacities for this during research data management.
- Revisit decisions at later stages of the research process, especially if strategies have changed.



# Guiding questions for researchers

- Will the project collect ‘personal data’?
- What is the legal basis for data processing?
- Who is responsible for data processing in the research project?
- Who has access to the research data?
- What type of personal data is processed? ‘Special categories’ (GDPR) of personal data?
- Has informed consent been obtained from the research participants aka the data subjects (GDPR)?
- Have you made an attempt to get in touch with the the research participants aka the data subjects?
- Can the data be anonymized?

# Research Data Lifecycle



# Study design and data collection



- Which data are suitable to capture a construct of interest?
- Would the data be accessible?
- Which data collection approach?
- What restrictions might be built-in by the platforms?
- What sensitive information might be included?
- How to meaningfully limit data collection and avoid ‘over-collecting’?
- Should data from different platforms/sources be combined?

# Example 1

## Measuring political communication / election debates

- This case represents a very common theme from social media research that exists in several variations.
- Studies on elections exists for different types of social media data, different countries, different countries.
- We focus on election studies based on Twitter data.

# Example 1

## Measuring political communication / election debates

### What to collect?

- All tweets from political candidates for a given election  
→ *public actors*
- Plus the tweets mentioning the candidates  
→ *general public*
- Plus general hashtags related to the election  
→ *potentially including activism*
- Combined with surveys  
→ *data linking challenges*

# Data preprocessing and analysis



- Data collected from social media often needs to be preprocessed or ‘cleaned’?
- Demographic information is often inferred from other available information (names, images)
- Analyses often make use of approaches from network analysis and Natural Language Processing (NLP) - including opinion mining approaches.
- Different/additional challenges when humans are involved in preparing data for analysis (e.g. crowdworkers, research assistants).

# Example 1

## Measuring political communication / election debates

### Preprocessing / analysis:

- Tweet topics and sentiments: popular approaches include mining for opinions on political topics (e.g. presidential approval).
- Filter out specific types of accounts, e.g. bots.
- Identify groups of actors: network structures
- Identify additional characteristics, e.g. gender detection, political affiliation
- Study constructs of interest, e.g. misinformation, sexism.

## Example 2

### Automated analyses and inferences

#### Ethical responsibilities in algorithmic inferences

During data analysis algorithmic approaches are often trained for automated analyses. Gender detection algorithms are often trained on image data – but do not perform equally for all cases.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability and transparency (pp. 77–91).

<http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf>

Wachter, S. and Mittelstadt, B. (2018) "A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI", Columbia Business Law Review. 2 443-493.

<https://academiccommons.columbia.edu/doi/10.7916/d8-g10s-ka92>



# Preserve and publish results and data



- Enhance overall research quality by supporting reproducibility and transparency.
- Publishing datasets can reduce the need to collect the same kind of data for different research projects.
- Several practical challenges often prevent efficient data sharing. Ongoing challenges for research infrastructure institutions.
- Extra need to care for data protection.

# Example 1

## Measuring political communication / election debates

### Data preservation and sharing:

- Twitter data should not be shared in full, as by the Twitter Terms of Services.
- Instead, Tweet IDs may be shared – but need to be “rehydrated” which often implies data loss.
- Deleted tweets can be considered as a withdrawal of consent. Different situation for politicians vs. general public.

## Example 3

### Data releases or “THE DATA IS ALREADY PUBLIC”

### The “Tastes, Ties, and Time” Dataset and the “OK Cupid Dataset”

Two problematic data sharing cases:

- The “Tastes, Ties, and Time” dataset contains Facebook data from university students and was released as anonymized data in 2008.
- In 2016 a dataset that was collected from the dating platform OK Cupid was publicly released.

Zimmer, M. (2010). “But the data is already public”: On the ethics of research in Facebook. *Ethics and Information Technology*, 12(4), 313–325. DOI: 10.1007/s10676-010-9227-5. Author’s copy available at: <http://www.sfu.ca/~palys/Zimmer-2010-EthicsOfResearchFromFacebook.pdf>

Zimmer, M. (2016). OkCupid Study Reveals the Perils of Big-Data Science. <https://www.wired.com/2016/05/okcupid-study-reveals-perils-big-data-science/>

Kirkegaard, EOW, & Bjerrekær, J. (2016). The OKCupid dataset: A very large public dataset of dating site users. *Open Differential Psychology*, Nov. 2, 2016. <https://openpsych.net/paper/46/>

# Conclusions

- Considerations about ethics in social media research are entangled in complex relations of different actors, most prominently platforms and their users.
- Legal regulations may differ across countries – and are often difficult to interpret for specific cases of social media data.
- Some guidance exists, but little standard procedures
  - ▶ projects under way to solve challenges practically.
- Critical reflections need to be built into the entire research process.

# Additional Literature

- Buchanan, Elizabeth A. and Michael Zimmer, "Internet Research Ethics", The Stanford Encyclopedia of Philosophy (Summer 2021 Edition), Edward N. Zalta (ed.)  
<https://plato.stanford.edu/archives/sum2021/entries/ethics-internet-research/> ; last access: 16.09.2021
- NESH (2019): A Guide to Internet Research Ethics. Issued by the The National Committee for Research Ethics in the Social Sciences and the Humanities (NESH) in 2003. Second edition published in Norwegian in 2018 and in English May 2019  
<https://www.forskningsetikk.no/en/guidelines/social-sciences-humanities-law-and-theology/a-guide-to-internet-research-ethics/>; last access: 16.09.2021
- RatSWD [German Data Forum] (2020): Data collection using new information technology. Recommendations on data quality, data management, research ethics, and data protection (RatSWD Output 6 (6)), Berlin  
<https://doi.org/10.17620/02671.51>; last access: 16.09.2021
- RatSWD [German Data Forum] (2020): Data Protection Guide. 2nd fully revised edition. (RatSWD Output 8 (6)), Berlin <https://doi.org/10.17620/02671.57>; last access: 16.09.2021
- Zimmer, Michael T. and Katharina Kinder-Kurlanda (eds.) (2017): Internet Research Ethics for the Social Age. New Challenges, Cases, and Contexts, New York

## Additional Literature cont. 1

- franzke, aline shakti, Bechmann, Anja, Zimmer, Michael, Ess, Charles and the Association of Internet Researchers (2020). Internet Research: Ethical Guidelines 3.0. <https://aoir.org/reports/ethics3.pdf>
- Heider, D., & Massanari, A. (Hrsg.). (2012). Digital ethics: Research & practice. Peter Lang.
- Giglietto, F., & Rossi, L. (2012). Ethics and interdisciplinarity in computational social science. Methodological Innovations Online, 7(1), 25–36. <https://doi.org/10.4256/mio.2012.003>
- Light, B., & McGrath, K. (2010). Ethics and social networking sites: A disclosive analysis of Facebook. Information Technology & People, 23(4), 290–311. <https://doi.org/10.1108/09593841011087770>
- Warfield, K., Hoholuk, J., Vincent, B., & Camargo, A. D. (2019). Pics, Dicks, Tits, and Tats: Negotiating ethics working with images of bodies in social media research. New Media & Society, 146144481983771. <https://doi.org/10.1177/1461444819837715>
- Murphy, J., Link, M. W., Hunter Childs, J., Langer Tesfaye, C., Dean, E., Stern, M., Pasek, J., Cohen, J., Callegaro, M., & Harwood, P. (2014). Social Media in Public Opinion Research: Report of the AAPOR Task Force on Emerging Technologies in Public Opinion Research. <https://www.aapor.org/Education-Resources/Reports/Social-Media-in-Public-Opinion-Research.aspx>
- Steinfeld, N. (2015). Trading with privacy: The price of personal information. Online Information Review, 39(7), 923–938. <https://doi.org/10.1108/OIR-05-2015-0168>

## Additional Literature cont. 2

- ACM (Association of Computing Machinery). (2018, August 22). ACM Code of Ethics and Professional Conduct. <https://www.acm.org/code-of-ethics>
- Vayena, E. a, & Tasioulas, J. b. (2013). Adapting Standards: Ethical Oversight of Participant-Led Health Research. PLoS Medicine, 10(3).  
<http://www.scopus.com/inward/record.url?eid=2-s2.0-84875438870&partnerID=40&md5=31b4854df58aab76ed5693bda8ba0db7>
- Zwitter, A. (2014). Big Data ethics. Big Data & Society, 1(2).  
<https://doi.org/10.1177/2053951714559253>
- Bull, S. S. a, Breslin, L. T. a, Wright, E. E. a, Black, S. R. a, Levine, D. b, & Santelli, J. S. c. (2011). Case study: An ethics case study of HIV prevention research on facebook: The just/us study. Journal of Pediatric Psychology, 36(10), 1082–1092.
- Jang, S. H., & Callingham, R. (2013). Conducting research in social media discourse: Ethical challenges. <http://www.scopus.com/inward/record.url?eid=2-s2.0-84892007811&partnerID=40&md5=f3ae093ab742eaf76e7a715a706890dc>
- Denecke, K. (2014). Ethical aspects of using medical social media in healthcare applications. Studies in Health Technology and Informatics, 198, 55–62.  
<http://www.scopus.com/inward/record.url?eid=2-s2.0-84902284672&partnerID=40&md5=e0c67c3012bd5e45d542f87a4121ab40>

## Additional Literature cont. 3

- Lehavot, K. a b, Ben-Zeev, D. c d, & Neville, R. E. e. (2012). Ethical considerations and social media: A case of suicidal postings on facebook. *Journal of Dual Diagnosis*, 8(4), 341–346.
- McKee, R. (2013). Ethical issues in using social media for health and health care research. *Health Policy*, 110(2–3), 298–301.
- Moreno, M. A. a, Goniou, N. a, Moreno, P. S. b, & Diekema, D. a c. (2013). Ethics of social media research: Common concerns and practical considerations. *Cyberpsychology, Behavior, and Social Networking*, 16(9), 708–713.
- Halavais, A. (2019). Overcoming terms of service: A proposal for ethical distributed research. *Information, Communication & Society*, 22(11), 1567–1581.  
<https://doi.org/10.1080/1369118X.2019.1627386>
- Schmidt, F. A. (2013). The good, the bad and the ugly: Why crowdsourcing needs ethics. *Proceedings - 2013 IEEE 3rd International Conference on Cloud and Green Computing, CGC 2013 and 2013 IEEE 3rd International Conference on Social Computing and Its Applications, SCA 2013*, 531–535.
- O’Neill, B. (2013). Who cares? Practical ethics and the problem of underage users on social networking sites. *Ethics and Information Technology*, 15(4), 253–262.
- Barnett, I., & Torous, J. (2019). Ethics, Transparency, and Public Health at the Intersection of Innovation and Facebook’s Suicide Prevention Efforts. *Annals of Internal Medicine*, 170(8), 565. <https://doi.org/10.7326/M19-0366>



Thank you !

gesis

Leibniz-Institut  
für Sozialwissenschaften

Leibniz  
Leibniz  
Gemeinschaft

## Expert Contact & GESIS Consulting




**Contact:** you can reach the speakers via e-mail:  
[oliver.watteler@gesis.org](mailto:oliver.watteler@gesis.org) & [katrin.weller@gesis.org](mailto:katrin.weller@gesis.org)

**GESIS Consulting:** GESIS offers individual consulting in a number of areas – including survey design & methodology, data archiving, digital behavioral data & computational social science – and across the research data cycle.

Please visit our website [www.gesis.org](http://www.gesis.org) for more [detailed information](#) on available services and terms.

## More Services from GESIS

- Get materials for [capacity building in computational social science](#) and take advantage of our expanding expertise and resources in [digital behavioral data](#).
- Use GESIS data services for [finding data](#) for secondary analysis and [sharing your own data](#).
- Check out the [GESIS blog](#) "Growing Knowledge in the Social Sciences" for topics, methods and discussions from the GESIS cosmos – and beyond.
- Keep up with GESIS activities and subscribe to the monthly [newsletter](#).
-  for publications, tools & services.

\*\*\* Upcoming online workshop; Nov 2-5, 2021 \*\*\*

[Introduction to Social Media as Research Data: Potentials and Pitfalls](#)

# More from CSS Experts in the MTE Series

- June 24 Katrin Weller: **A Short Introduction to Computational Social Science and Digital Behavioral Data**
- July 01 Fabian Flöck, Indira Sen: **Digital Traces of Human Behavior from Online Platforms – Research Designs and Error Sources**
- July 08 Sebastian Stier, Johannes Breuer: **Combining Survey Data and Digital Behavioral Data**
- Sept 16 Oliver Watteler, Katrin Weller: **Research Ethics and Data Protection in Social Media Research**
- Sept 30 Roberto Ulloa: **Introduction to Online Data Acquisition**
- Oct 07 Roberto Ulloa: **Auditing Algorithms: How Platform Technologies Shape our Digital Environment**
- Oct 14 Marius Sältzer, Sebastian Stier: **The German Federal Election: Social Media Data for Scientific (Re-)Use**
- Nov 04 Arnim Bleier: **Introduction to Text Mining**
- Nov 11 Haiko Lietz: **Social Network Analysis with Digital Behavioral Data**
- Dec 2 Olga Zagovora, Katrin Weller: **Altmetrics: Analyzing Academic Communications from Social Media Data**
- Dec 16 Andreas Schmitz: **Online Dating: Data Types and Analytical Approaches**
- Jan 13 Gizem Bacaksizlar: **Political Behavior and Influence in Online Networks**
- Jan 27 David Brodesser: **SocioHub – A Collaboration Platform for the Social Sciences**