gesis

Leibniz Institute
for the Social Sciences

GDPR & Ethical Issues with Social Media Data

SSHOC / DARIAH Bootcamp

Katrin Weller & Oliver Watteler

Feb. 2021

Leibniz
Association

Syllabus and references:

tinyurl.com/1e99am7a
Password: gdpr+sme2021

Research ethics, data protection and the specific contexts for studying social media

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’?

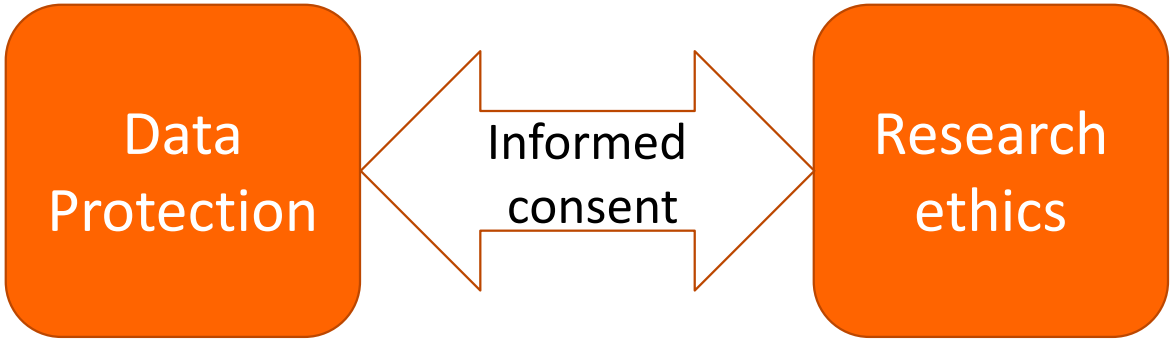
- Data protection
  - part of personality right to privacy
- “Privacy is a personal condition of life characterised by seclusion from, and therefore absence of acquaintance by, the public (Neethling 2005).”
- Prevention of unwanted disclosure of personal information or the misuse of such information
  - 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

Link between data protection and research ethics: informed consent

Informed consent means for example:

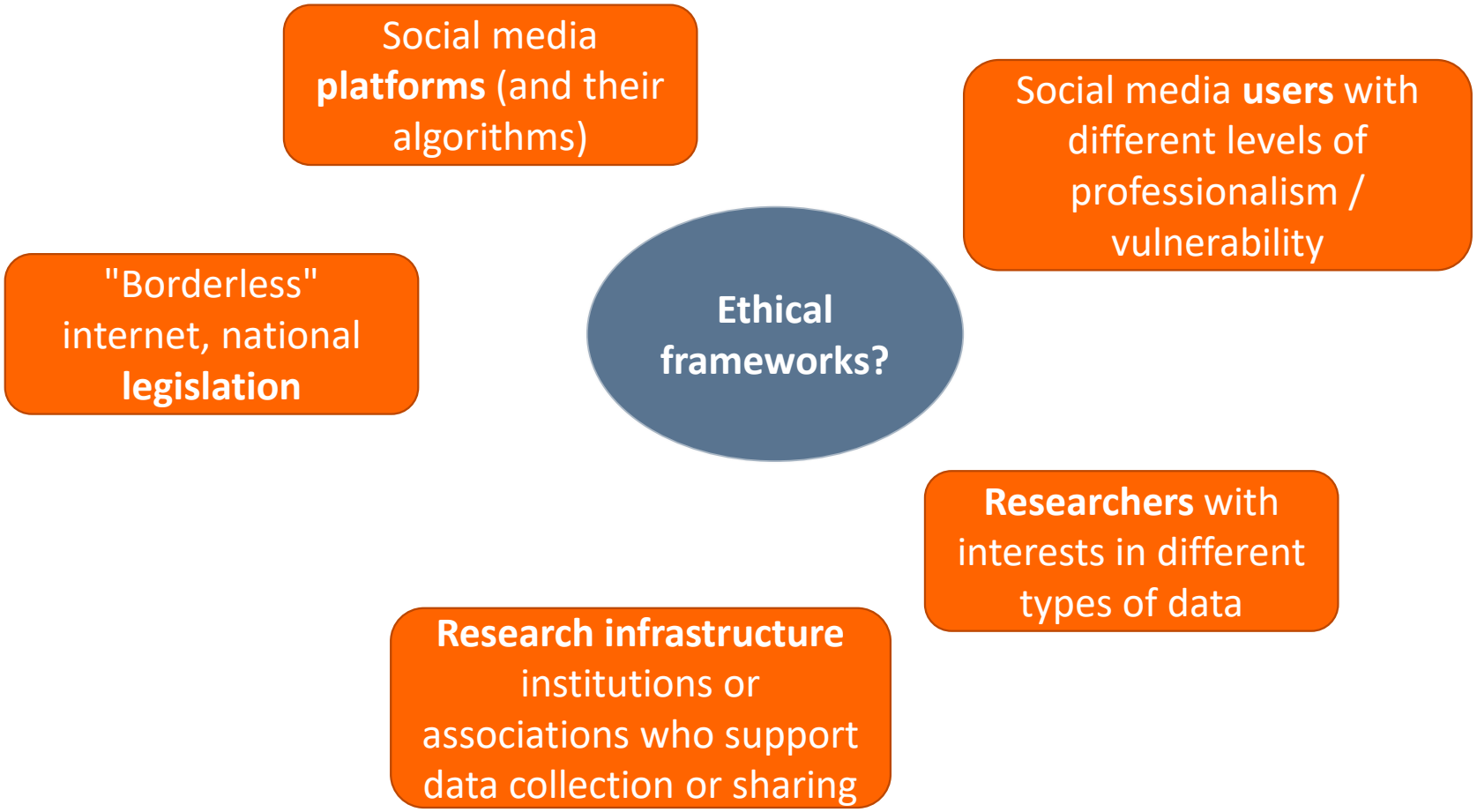
- information
- transparency
- chance to disagree

Regularly in social media research: lack thereof

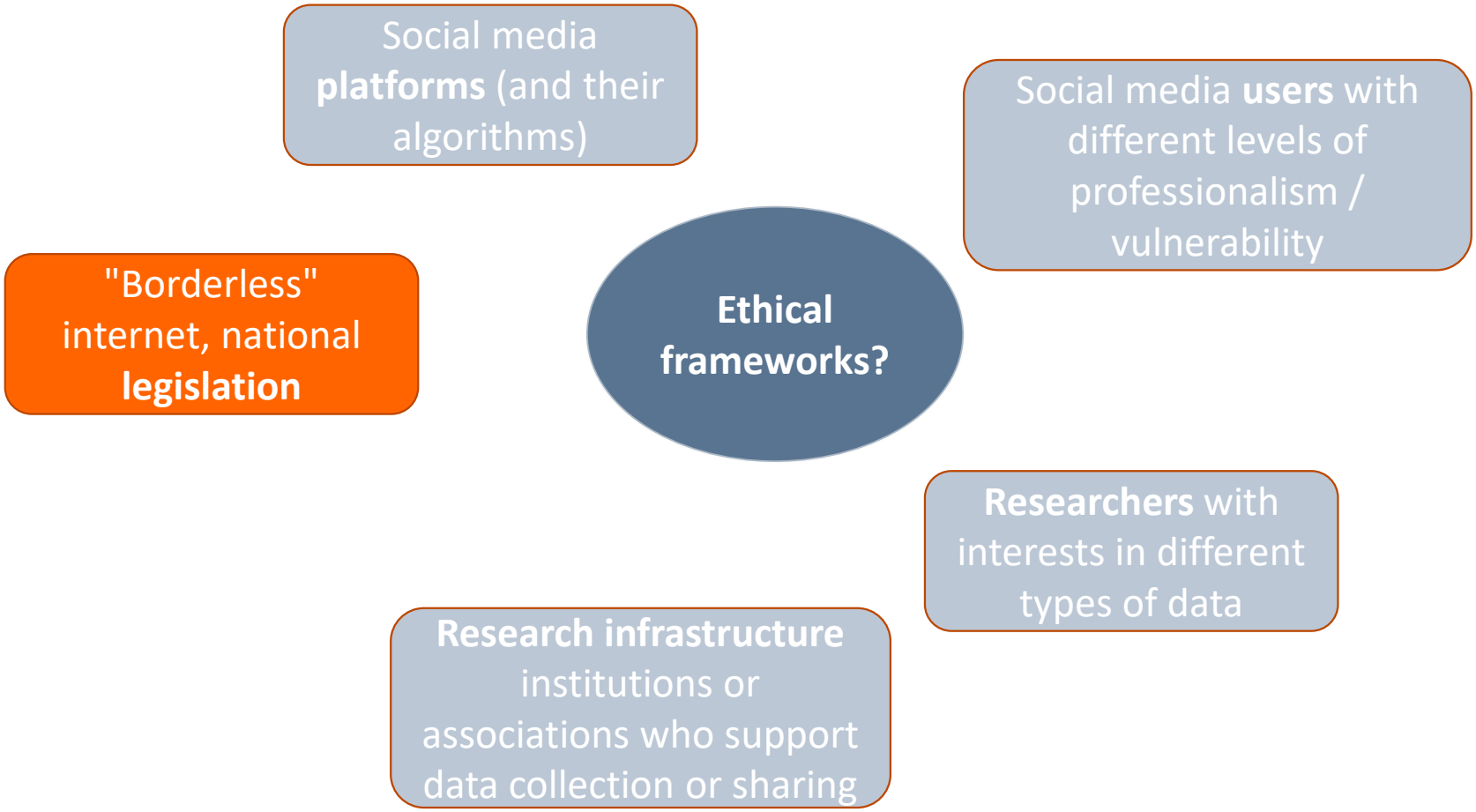


gesis Leibniz Institute for the Social Sciences

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

Terms of service

Ethical guidelines

Data protection legislation - overview

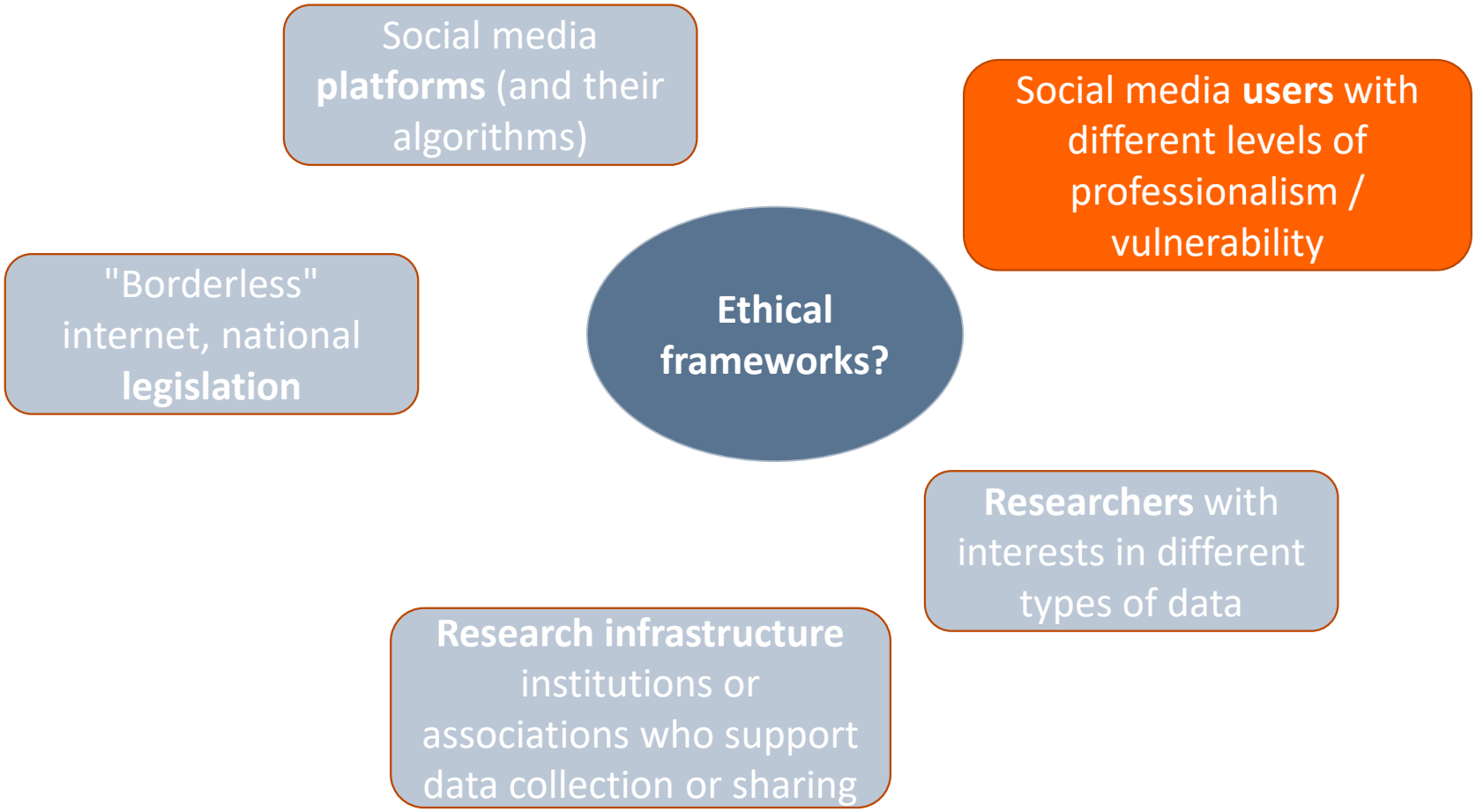
- Since 25 May 2018, the EU General Data Protection Regulation (GDPR) applies:
  - 99 articles and 173 recitals.
  - Applies directly.
  - Intended to harmonize data protection law EU-wide.
  - BUT about 150 “opening clauses” or exemptions.
- GDPR (factually) integrated into hierarchy of norms:
  - Legislation 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

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%

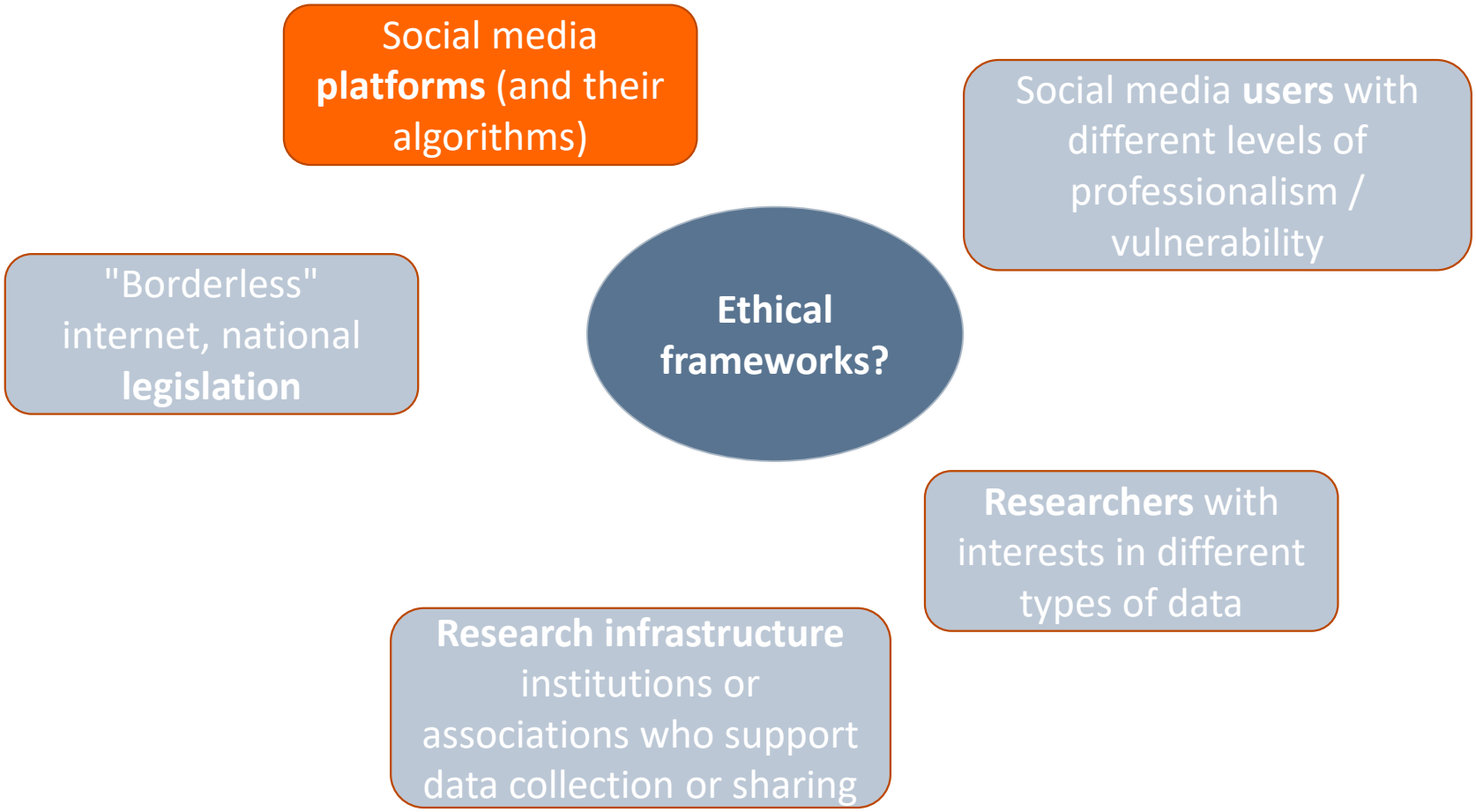
Note. The shading was used to provide a visual cue about higher percentages.

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)

Different platform approaches



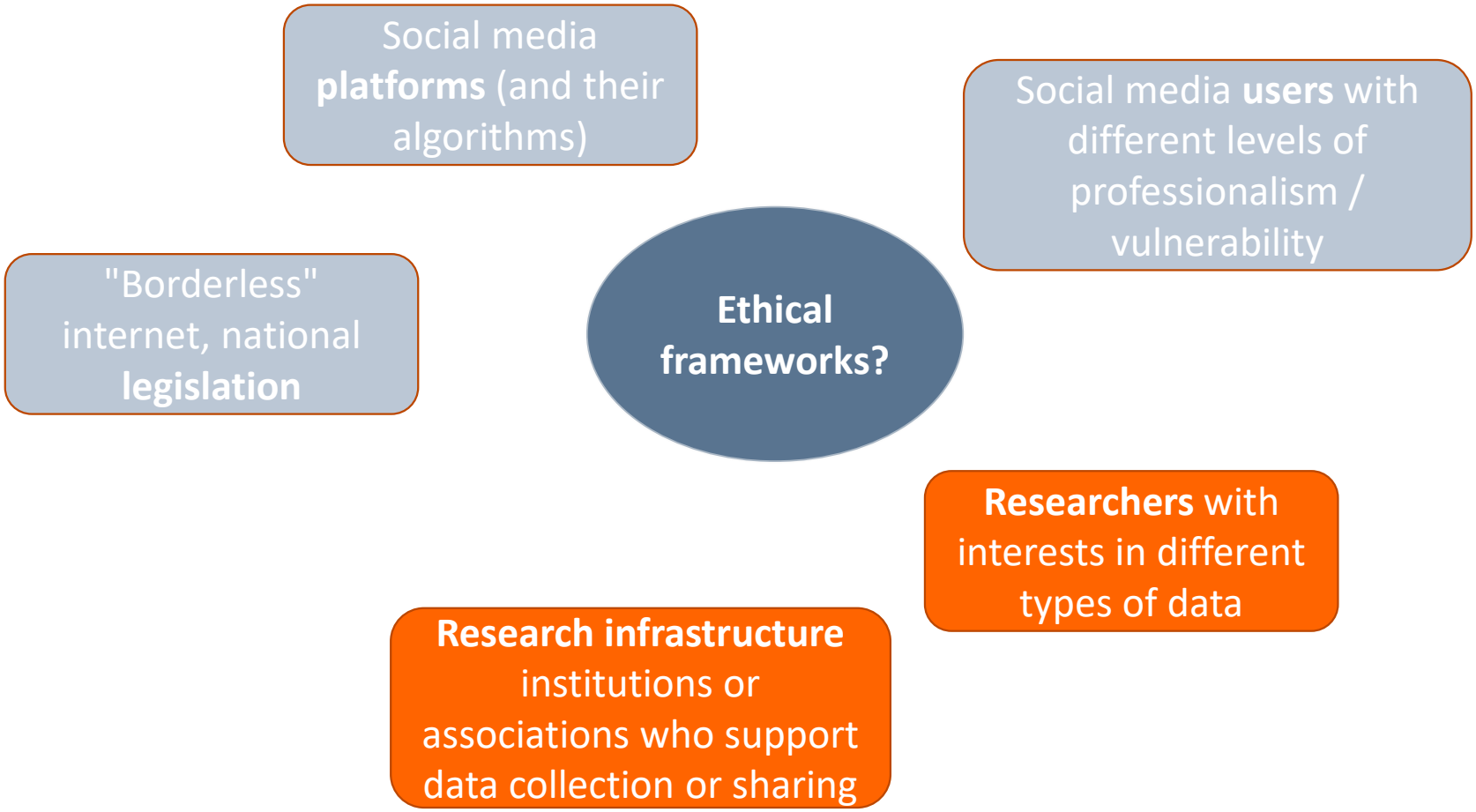
Not all platforms are equal

- 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 at the moment.

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.

Different steps in the research process

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.

Research Data Lifecycle

Guiding questions RDM 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?
- Does an informed consent exist 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?

Study design and data collection

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

Example

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

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
- 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

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.

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

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.

Working groups
(results to be presented on Thursday)

Working groups

- In **teams of about 2-3 persons** please select one of the following 3 example cases taken from existing social media research.
- Alternatively, you may choose to work on **your own choice** of a case scenario.

Example cases

Case 1: Collecting data from vulnerable groups

Sensitive information and interacting with user groups

This case is focusing mainly on the study design and data collection phase. It generally asks for reflecting on the role of social media users in the research process, and places a specific focus on vulnerable group. We have chosen examples from the medical domain to illustrate challenges with vulnerable groups.

Example cases

Case 1: Collecting data from vulnerable groups

Literature for this case:

- The following chapters from Michael Zimmer and Katarina Kinder-Kurlanda (Eds.), Internet Research Ethics for the Social Age. Peter Lang. Full book in Scribd, https://de.scribd.com/document/360717441/Internet-Research-Ethics-for-the-Social-Age-New-Challenges-Cases-and-Contexts-Full
  - Eskisabel-Azpiazu, Amaia; Cerezo-Menéndez, Rebeca; Gayo-Avello, Daniel (2017). An Ethical Inquiry into Youth Suicide Prevention Using Social Media Mining.
  - Ferguson, Robert Douglas (2017). Negotiating Consent, Compensation, and Privacy in Internet Research: PatientsLikeMe.com as a Case Study.
- 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/

Example cases

Case 2: Automated analyses and inferences

Ethical responsibilities in algorithmic inferences

This case has a focus on the data analysis perspective, where often algorithmic approaches are trained for automated analyses. Gender detection algorithms are often trained on image data – but do not perform equally for all cases. And algorithms have an impact within academia, but also well beyond it.

Example cases

Case 2: Automated analyses and inferences

Ethical responsibilities in algorithmic inferences

Literature for this case:

- 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

Example cases

Case 3: Data releases or "THE DATA IS ALREADY PUBLIC"

The "Tastes, Ties, and Time" Dataset and the "OK Cupid Dataset"

This case has a focus on the data sharing perspective. Two different examples from the past should be considered. 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.

Example cases

Case 3: Data releases or "THE DATA IS ALREADY PUBLIC"

The "Tastes, Ties, and Time" Dataset and the "OK Cupid Dataset"

Literature for this case:

- 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/

Example cases

Case 4: Your Choice

Alternatively, you can decide to present a case of your choice or based on your own work or experience

(e.g., a paper you have recently read or your own research design / work in progress etc.)

If you select this option, please also **share useful references** with the course.

Working groups

- In **teams of about 2-3 persons** please select one of the example research cases.
- **Please work through the documents and present your case in our next session using the following three slides as orientation.**

1. Short summary

- *Shortly present the case you selected: what was the research objective, what data has been used?*

2. Challenging areas for research ethics and data protection

- *Where in the research design do you see the main challenges for data protection and research ethics?*
- *Do you see room for improvement? What could have been done differently?*

3. Your comments and Questions

- *What about the case would you like to discuss in the group?*
- *Do you have any open questions?*
- *What lessons learnt would you like to share?*

Questions welcome:

katrin.weller@gesis.org

oliver.watteler@gesis.org

Thanks for participating

gesis Leibniz Institute for the Social Sciences

Leibniz Association