6.
A closer look at the data – data quality and representativeness

Verification

- Verified accounts
  - Real name policy
- Verified content?

Facebook now lets you choose from more than 50 gender options

- … but still deletes accounts of drag queens because they don‘t follow the „real name policy“

Heklina (left) said: 'I have been Heklina for 20 years, and I have Facebook telling me Heklina does not exist. So they're basically wiping you out of existence'

http://www.dailymail.co.uk/news/article-2776944/Facebook-apologizes-drag-queens-policy.html

Comparability?

Little or no standards related to:

- Data collection
- Data cleaning
- Analysis
- Tools
- Collection period

Data loss

	A	B	C	D	E
1	**Hashtag**	**Tweets**	**Hydrated**	**Deleted**	**% Deleted**
2	JeSuisJuif	96,518	89,584	6,934	7.18%
3	JeSuisCharlie	6,503,425	5,955,278	548,147	8.43%
4	JeSuisAhmed	264,097	237,674	26,423	10.01%
5	CharlieHebdo	7,104,253	6,554,231	550,022	7.74%
6	**Total**	**13,968,293**	**12,836,767**	**1,131,526**	**8.10%**

Charlie Hebdo Tweets and Deletes

Summers, E. (2015), "Tweets and Deletes: Silences in the Social Media Archive", available at: https://medium.com/on-archivy/tweets-and-deletes-727ed74f84ed (accessed 6 September 2015).

Reproducibility

- Is is possible to reproduce an exact dataset?
- How to work with data collected by others?

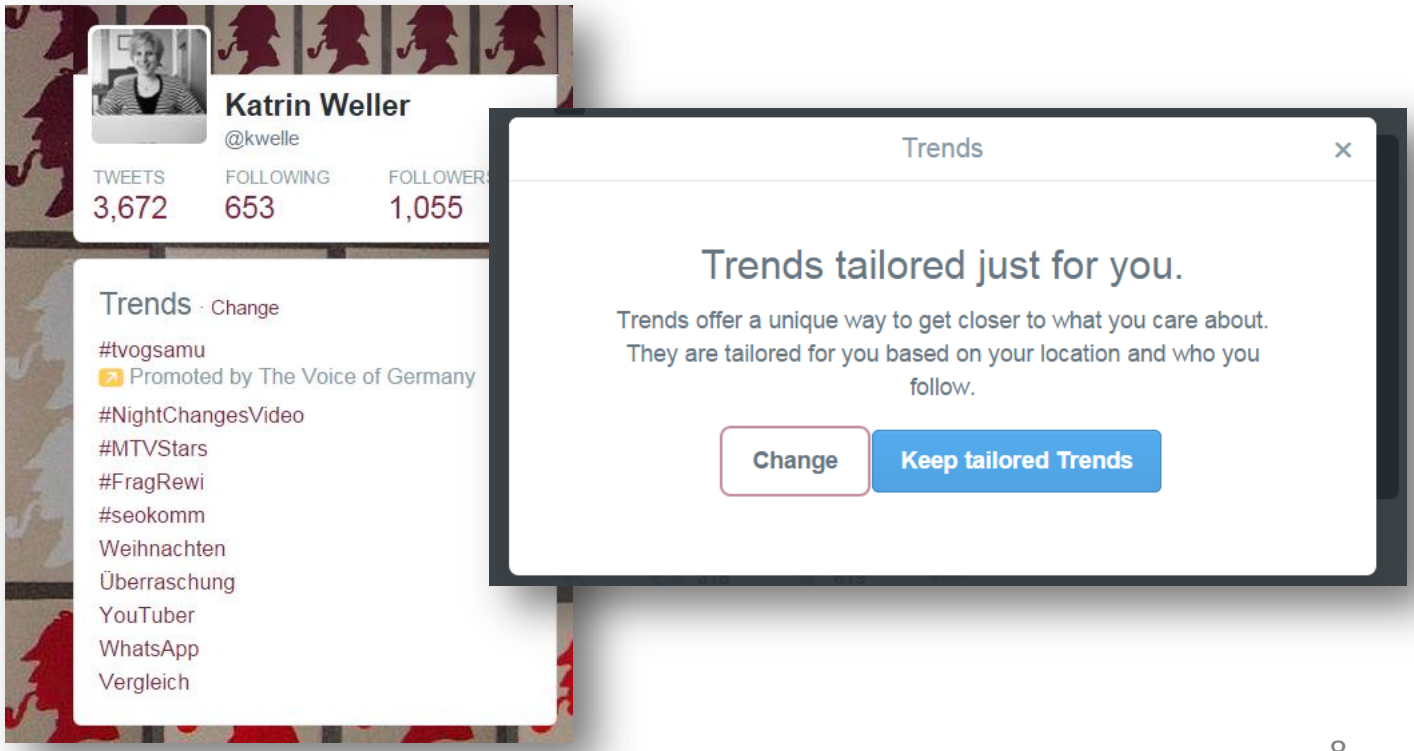
Data Quality

- E.g. comparison of Twitter API and Reseller data

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. Retrieved from http://arxiv.org/abs/1306.5204

Personalization

- „More like me“
- Personalized resutls and interfaces



Representativeness

- Who is using a platform
- Who is using specific features within a platform
- Time and duration of data collection
- Sampling
- Languages

Bruns, A., & Stieglitz, S. (2014). Twitter data: What do they represent? It - Information Technology, 56(5), 240–245. http://doi.org/10.1515/itit-2014-1049

Representativeness

"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.

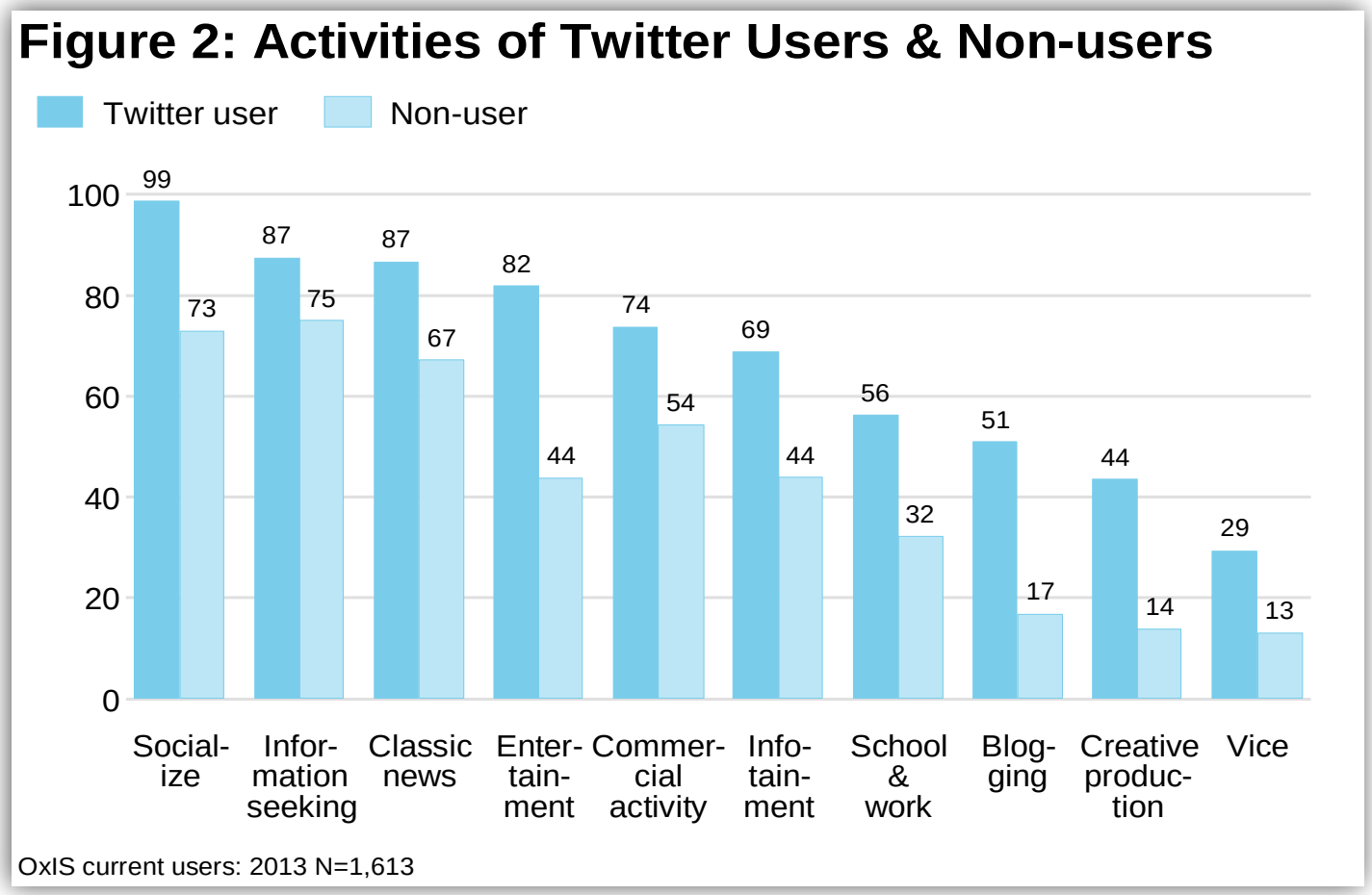
Representativeness

"About a third of all UK Internet users have a twitter profile; a subset of that group are the active tweeters who produce the bulk of content; and then a tiny subset of that group (about 1%) geocode their tweets (essential information if you want to know about where your information is coming from)."

Graham M. (2012). Big data and the end of theory?". The Guardian. Retrieved from: http://www.theguardian.com/news/datablog/2012/mar/09/big-data-theory

gesis Leibniz Institute for the Social Sciences

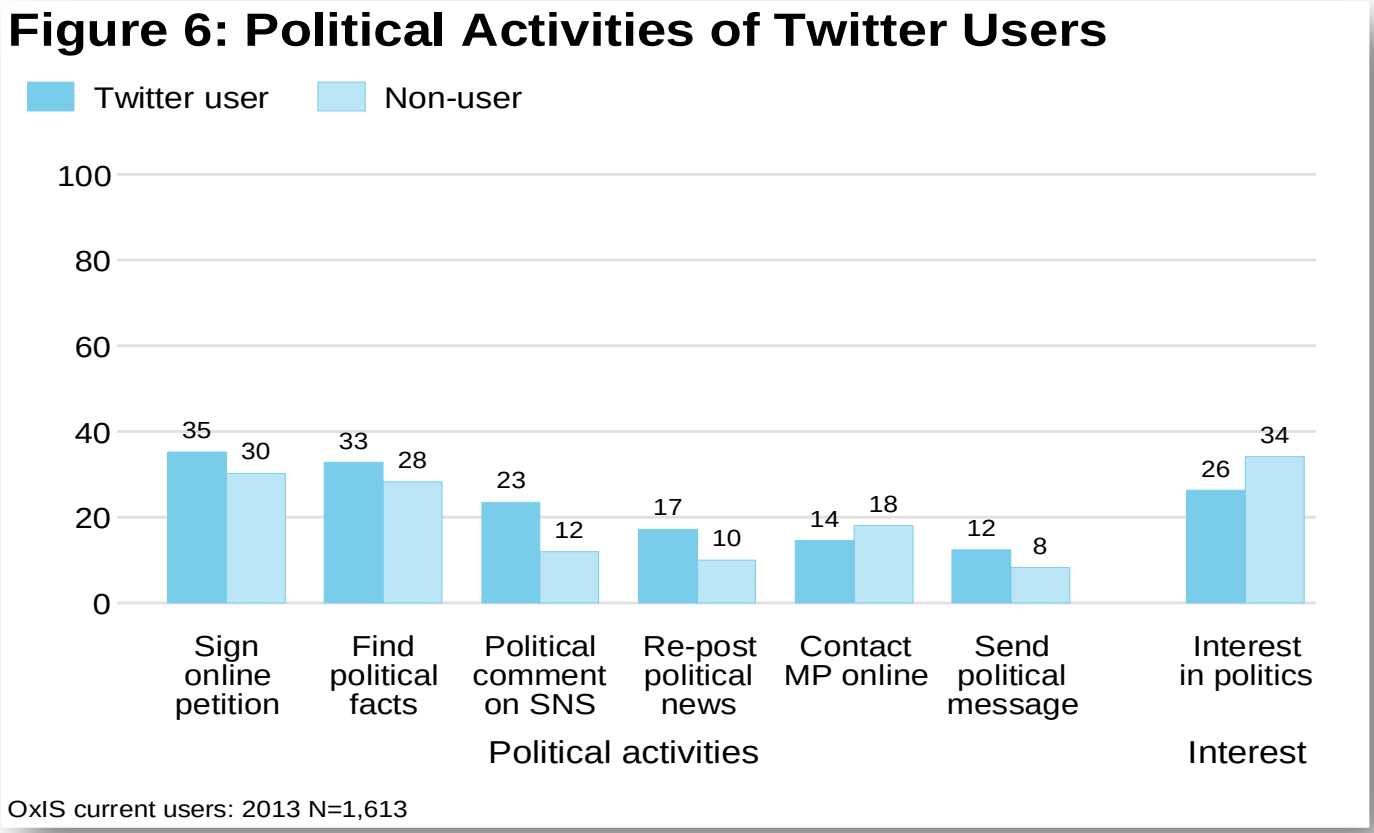
Representativeness



Blank, G. (2014). Who uses Twitter? Representativeness of Twitter Users. Presentation at General Online Research GOR 14. Retrieved from: http://conftool.gor.de/conftool14/index.php?page=downloadPaper&filename=Blank-Who_uses_Twitter_Representativeness-119.pptx&form_id=119&form_version=final

gesis Leibniz Institute for the Social Sciences

Repräsentativität



Blank, G. (2014). Who uses Twitter? Representativeness of Twitter Users. Presentation at General Online Research GOR 14. Retrieved from: http://conftool.gor.de/conftool14/index.php?page=downloadPaper&filename=Blank-Who_uses_Twitter_Representativeness-119.pptx&form_id=119&form_version=final

Beginning 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

Beginning of Theory?

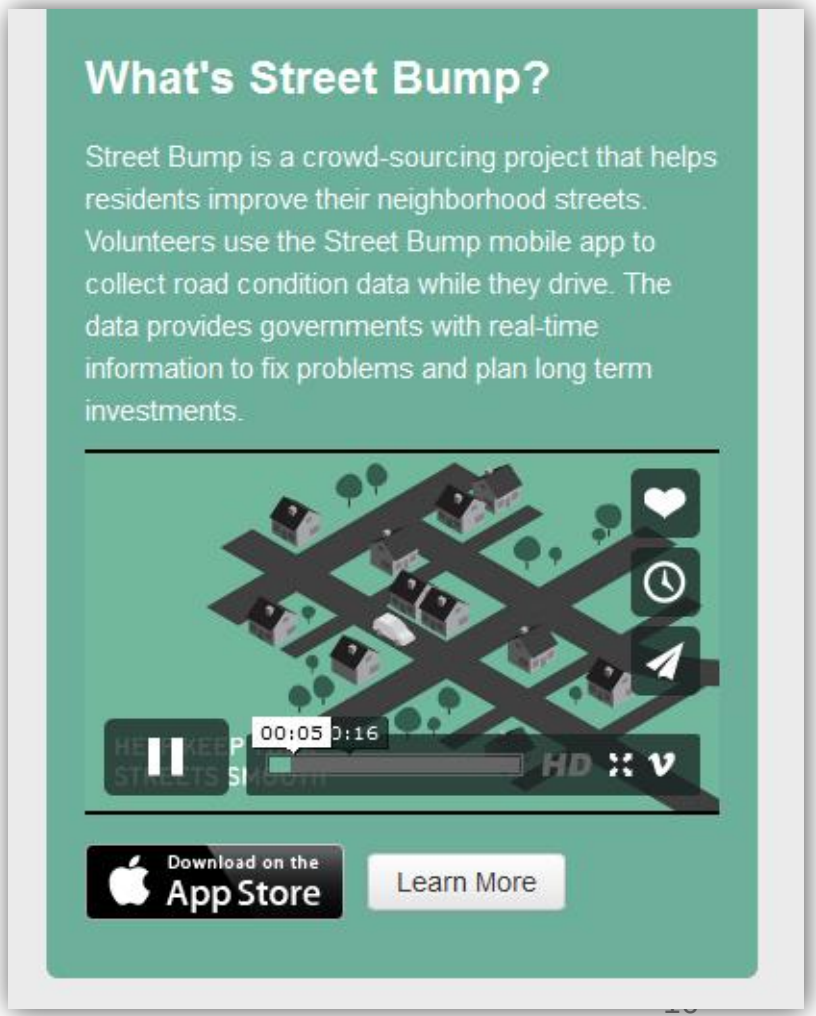
"There are a lot of small data problems that occur in big data," says Spiegelhalter. "They don't disappear because you've got lots of the stuff. They get worse."

Tim Harford (2014): Big data: are we making a big mistake? FT Magazine, retrieved from: http://www.ft.com/cms/s/2/21a6e7d8-b479-11e3-a09a-00144feabdc0.html#axzz2xGqAnW8a?utm_source=pocket&utm_medium=email&utm_campaign=pockethits

gesis Leibniz Institute for the Social Sciences

Representativeness

- Who is excluded?



http://streetbump.org

http://www.wired.com/2014/03/potholes-big-data-crowdsourcing-way-better-government/

The forgotten features of social media platforms

Forgotten features?

Negative actions, e.g. unfriending on Facebook

Bevan, J.L., Ang, P.-C. and Fearns, J.B. (2014), "Being unfriended on Facebook: An application of Expectancy Violation Theory", Computers in Human Behavior, Vol. 33, pp. 171-178, DOI:10.1016/j.chb.2014.01.029.

Quercia, D., Bodaghi, M. and Crowcroft, J. (2012), "Loosing 'friends' on Facebook", in Proceedings of the 4th Annual ACM Web Science Conference (WebSci12), Evanston, IL, ACM Press, New York, pp. 251–254, DOI:10.1145/2380718.2380751.

Xu, B., Huang, Y., Kwak, H. and Contractor, N. (2013), "Structures of broken ties: exploring unfollow behavior on Twitter", in Proceedings of the 2013 Conference on Computer Supported Cooperative Work, San Francisco, CA, ACM Press, New York, pp. 871-876, DOI:10.1145/2441776.2441875.

Forgotten features?

- (a) functionalities that are used less over the course of time (e.g. the blogroll feature in blogs) and eventually even drop out of researchers' conscience (or even of the platform's interface).
- (b) functionalities that are introduced at some point when social media research has already moved on from general descriptions of the platform to practical use cases and are not added to the researchers' already established set of key metrics to analyze for studying user activities. E.g. Favs on Twitter, Hashtags on Facebook.

Katrin Weller , (2016) "Trying to understand social media users and usage: The forgotten features of social media platforms", Online Information Review, Vol. 40 Iss: 2, pp.256 – 264.

Practice

How well do you know the platforms you use?

- Of which platforms do you think you know (or use) all features?
- Which features do you think are not so well known at the platforms you are frequently using?

Have you ever done the following on Facebook?

- Saved a link to read later?
- Tagged others in photos?
- Sorted friends to different lists (and managed access rights accordingly)?
- Followed someone (instead of friending)?
- Hidden (or deleted) one of your posts?

Getting to know new platforms

- Pick one of the following platforms that you have not used before:
  - Pinterest
  - Tumblr
  - Instagram
  - Vine
  - Twitter
- (You may have to create an account)
- Try out the platforms for *10 minutes*. What do you think are the main features and functionalities? How important are the following aspects:
  - Content creation
  - Content sharing
  - User networks (e.g. following)
  - User interactions / conversations
  - Rating (e.g. likes, recommendations)

Research ethics

Scientists Are Just as Confused About the Ethics of Big-Data Research as You

By **Sarah Zhang**, www.wired.com
Mai 20., 2016

Original anzeigen
ethics

When a rogue researcher last week released 70,000 OkCupid profiles, complete with usernames and sexual preferences, people were pissed. When Facebook researchers manipulated how stories appear in News Feeds for a mood contagion study in 2014[1], people were really pissed. OkCupid filed a copyright claim to take down the dataset; the journal that published Facebook's study issued an "expression of concern." Outrage has a way of shaping ethical boundaries. We learn from mistakes.

http://www.wired.com/2016/05/scientists-just-confused-ethics-big-data-research/

Ethics

- Informed consent?
- Data is already public

- There is a lot of ongoing discussion
  - Is it ok to quote tweets and to mention user names
  - Big data vs. Small data
  - When are you obliged to quote tweets

gesis Leibniz Institute for the Social Sciences

AoIR ASSOCIATION OF INTERNET RESEARCHERS

Member Login | Join AoIR | Mailing Lists | Donate | Contact Us

search here ... Go

Home | About | News | Conferences | Publications | Community | Awards | Ethics

Ethics

AoIR has an ongoing commitment to ensuring that research on and about the Internet is conducted in an ethical and professional manner. The Ethics Working Committee, as composed of ethicists and researchers from various regions and countries, has produced two major reports to assist researchers in making ethical decisions in their research:

- *2012: Ethical decision-making and Internet research 2.0: Recommendations from the AoIR ethics working committee* [PDF]
- *2002: Ethical decision-making and Internet research: Recommendations from the AoIR ethics working committee* [PDF]

Researchers, students, ethicists, and related institutional bodies and academic organizations in the domain of Internet research may turn to these ethics document as a starting point for their inquiries and reflection. Just as these documents were immeasurably enriched by comments and contributions from AoIR members, we hope that readers will continue to call attention to issues and resources in Internet research ethics for debate and deliberation by the ethics working committee.

AoIR @ Twitter

- #AoIR2016 at @HumboldtUni w/ the help of @hiig_berlin. Have you registered? Help out by adding to travel grants https://t.co/HxWOOHRp6t about 4 weeks ago
- Joining #AoIR2016 Berlin? Book hotels EARLY. Our con is close to the Day of German Unity - a big national holiday. https://t.co/IPWF989Qsb about 1 month ago

http://aoir.org/ethics/

Conclusions 6

Lessons learned

- Reflections on theory and methodology are needed in order to continuously improve quality of approaches.
- There is a growing interest in research ethics in social media research.

If you have time to read 3 papers…

- Bruns, A., & Stieglitz, S. (2014). Twitter data: What do they represent? It - Information Technology, 56(5), 240–245. http://doi.org/10.1515/itit-2014-1049
- 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. Retrieved from http://arxiv.org/abs/1306.5204
- Zimmer, M. (2010). "But the data is already public": on the ethics of research in Facebook. Ethics and Information Technology, 12(4), 313–325. http://doi.org/10.1007/s10676-010-9227-5