Data Sharing Practices in Big Social Data: A Review of the ASONAM Community

RESEARCH DATA ALLIANCE
Co-chair Ethics and Social Aspects of Data IG

Candice L. Lanius
Department of Communication and Media

Rensselaer Polytechnic Institute

My Dissertation Research

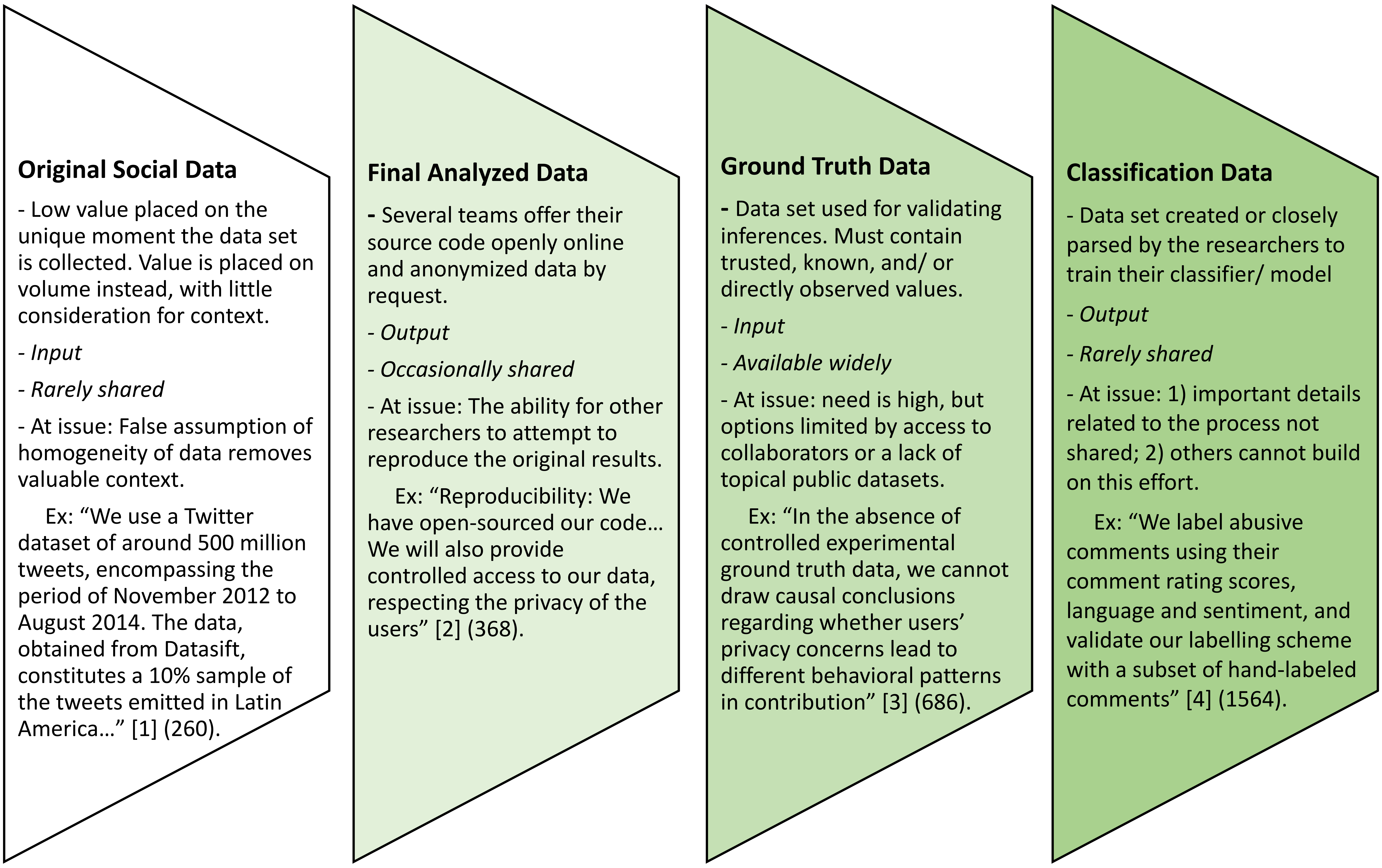
- My research is a sociological study of data scientists, researchers, and engineers who use large-scale social media data to answer social, behavioral, and physical questions about society and our larger environment: i.e. big social data.
- I focus on the research design of big social data analysts to discover moments of interpretation and argumentation in the research process.

Major Outcome: Discovery of Data Sharing Practices

- Research Design is the strategic plan built by the researcher to coherently and logically organize the research, including literature review, question formulation, data selection, collection, processing, analysis, and conclusions.
- One key aspect is the input and output of research data, and my study reveals data sharing, soliciting, and obfuscation practices at different points throughout the process across the community.
- These data sharing practices will be of interest to the Research Data Alliance, especially for data handling, education, big data practitioners, and those supporting the expansion of small scale social and behavioral data sharing.

Method

- Data: A sociological survey sent to 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining participants who used large-scale social media datasets as part of their analysis. A thematic analysis of associated papers in the conference proceeding.
- 34 projects selected and 90 authors solicited with 25 responding to the survey.



Original Social Data

- Low value placed on the unique moment the data set is collected. Value is placed on volume instead, with little consideration for context.

- *Input*

- *Rarely shared*

- At issue: False assumption of homogeneity of data removes valuable context.

Ex: “We use a Twitter dataset of around 500 million tweets, encompassing the period of November 2012 to August 2014. The data, obtained from Datasift, constitutes a 10% sample of the tweets emitted in Latin America…” [1] (260).

Final Analyzed Data

- Several teams offer their source code openly online and anonymized data by request.

- *Output*

- *Occasionally shared*

- At issue: The ability for other researchers to attempt to reproduce the original results.

Ex: “Reproducibility: We have open-sourced our code… We will also provide controlled access to our data, respecting the privacy of the users” [2] (368).

Ground Truth Data

- Data set used for validating inferences. Must contain trusted, known, and/ or directly observed values.

- *Input*

- *Available widely*

- At issue: need is high, but options limited by access to collaborators or a lack of topical public datasets.

Ex: “In the absence of controlled experimental ground truth data, we cannot draw causal conclusions regarding whether users’ privacy concerns lead to different behavioral patterns in contribution” [3] (686).

Classification Data

- Data set created or closely parsed by the researchers to train their classifier/ model

- *Output*

- *Rarely shared*

- At issue: 1) important details related to the process not shared; 2) others cannot build on this effort.

Ex: “We label abusive comments using their comment rating scores, language and sentiment, and validate our labelling scheme with a subset of hand-labeled comments” [4] (1564).

Conclusion

- Current barriers to sharing research data amongst the big social data community include cost, culture, and outdated regulations surrounding human-centered data. These barriers are specific to the different types of data and their uses within the research design. My research helps to understand these barriers in specific settings to work to remove them.

References

[1] Korkmaz, G., J. Cadena, C. J. Kuhlman, A. Marathe, et. al. (2015, August). Combining heterogeneous data sources for civil unrest forecasting. pp. 258-265.
[2] Devineni, P., D. Koutra, M. Faloutsos, & C. Faloutsos. (2015, August). If walls could talk: Patterns and anomalies in Facebook wallposts. pp. 367-374.
[3] Kayes, I., N. Kourtellis, F. Bonchi, & A. Iamnitchi. (2015, August). Privacy concerns vs. user behavior in community question answering. pp. 681-8.
[4] Liao, C., A. Squicciarini, C. Griffin, & S. Rajtmajer. (2015, August). A hybrid epidemic model for antinormative behavior in online social networks. pp. 1563-4.

Research Data Alliance 8th Plenary Meeting, September 15-17, 2016, Denver, Colorado.