Human-Centered Data Science for Collaborative, Interdisciplinary Research

Melissa Bica
Department of Computer Science
University of Colorado Boulder
Boulder, Colorado, USA
melissa.bica@colorado.edu

ABSTRACT

The study of data science practices is necessary with the rise of big data, as well as people wanting to take advantage of it. I come to this workshop from the perspective of a researcher who uses data science methods in collaborative work with scientists across multiple disciplines. This research has enlightened me on how to approach the complex interactions between data science and the domain to which it is applied.

KEYWORDS

Data science, interdisciplinary research

ACM Reference Format:
Melissa Bica. 2019. Human-Centered Data Science for Collaborative, Interdisciplinary Research. In *Proceedings of CHI EA '19: Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '19 Extended Abstracts).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/1122445.1122456

CHI '19 Extended Abstracts, May 4–9, 2019, Glasgow, Scotland, UK

ACM ISBN 978-1-4503-9999-9/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

INTRODUCTION

As a computer and information scientist and researcher in the area of crisis informatics, I conduct muli-method, human-centered empirical research, often utilizing data science techniques. This work is in collaboration with researchers in meteorology, atmospheric science, and social and behavioral sciences at the National Center for Atmospheric Research (NCAR). In particular, my research examines how people both communicate and make sense of forecast and risk information visualizations for hurricanes.

Motivation for Participating in This Workshop

My interest in this workshop is particularly related to the theme of *bridging the gap between the knowledge of data scientists and that of domain experts in various fields of application*, especially given that I identify with both roles and collaborate with others in both roles. In this paper, I will describe some key learnings from my time working in this collaborative data science research space, which I hope can be valuable to those conducting empirical research about such practices.

BRIDGING THE GAP

Translational Work

In my work I collaborate closely with researchers in the same application areas of crisis and severe weather, but who come from different scientific disciplines. Much of my work is translational—I take research goals from my atmospheric and social science collaborators, collect data, and design a research plan utilizing data science methods with the goal to make contributions to each of our collective research domains. My research is human-centered in both the methods of analysis I use as well as the collaborative nature of the research design itself.

To be effective in this kind of interdisciplinary, human-centered work requires the researcher to be an engaged member of multiple research communities—in my case, this includes not only HCI, but also weather and atmospheric science and risk communication. This is one way of instantiating the “human perspective” of data science as described by Blei and Smyth [4], which allows data scientists and domain experts to together “develop computational and statistical tools to explore data, questions, and methods *in the service of the goals of the discipline* [emphasis added].” Some may argue that data scientists do not need domain knowledge to be successful in their work; while this may be true, depending on measures of success, having not only domain knowledge of, but also true engagement with, the discipline to which a data scientist is “in service” can greatly expand the kinds of questions they can answer and contributions they can make, as well as the value of those contributions [5].

Of course, translation goes in both directions. When data scientists are able to translate what they do in ways that make sense and are meaningful to domain experts who are not familiar with the complexities of data science, it supports collaboration and learning for all. As an example, I recently met with a collaborator in the social sciences about a Twitter dataset we were working with to devise a useful coding scheme to help filter the data. One of the first variables was a seemingly (to me) simple binary classification about relevance to a particular disaster event: a tweet was either related to the event or not. It was not until discussing the rest of the seemingly more complex codes and circling back to this first relevance one that we discovered a difference in thinking on one of the most fundamental terms in our discussion: "tweet."

Further discussion revealed that when my collaborator, who had no previous experience working as an analyst of Twitter data, used the term, she referred to the text a user writes that appears on Twitter. When I use the term, I refer to the data object retrieved from an API consisting of hundreds of attributes describing the tweet, with tweet text being just one of them. As a data scientist, determining whether a "tweet" is relevant to an event involves considering all of these attributes to gain as much context as possible around timing, location, and content, as years of work analyzing tweets has taught me that people are not as explicit in what they tweet about as analysts might like [1]. This discrepancy in our ways of thinking about something so fundamental was a strong reminder of the importance of establishing intersubjectivity when working with domain experts on these sorts of projects, especially when they are new to data science, particularly around shared language and techniques.

Data (Science) is Not a Silver Bullet

Many people who first consider Twitter data in relation to their research think of Twitter as being the answer to all, or at least many, of their questions. They want access to the "power of data science" to be able to unlock insights about what people do in an event, how they use a hashtag, or what they think about some phenomenon. While there certainly are many interesting questions that can be answered with *informed, human-centered* analysis of a *thoughtfully collected and curated* social media dataset [6], there are also limitations to what can be gained.

One common misconception about the "power" of both Twitter data and the use of data science methods with such data is along the lines of, "*We can use machine learning to categorize these tweets into these X domain-specific categories!*" Machine learning (ML) can be a great, powerful tool to do certain types of analysis with Twitter data. However, it also has limitations that often need to be actualized for those new to this kind of work and in search of particular results. As mentioned above, people do not use the language researchers might use to describe the same thing: e.g., almost no one other than meteorologists on Twitter uses "funnel cloud" to describe a tornado, and thus filtering Twitter data on that term will not reveal much about how members of the general public reacts to a tornado. Similarly, overly-nuanced questions or categories, for instance differentiating between a *risk* and a *threat*, do not reflect the ways that people actually tweet, and using ML techniques to understand or categorize such themes will not be fruitful.

In my own research, I aim to answer questions about sociotechnical behaviors which are often better described qualitatively, e.g., what it means that different populations share different images of a disaster, or how people threatened by a hurricane make sense of different types of forecast graphics. While I almost always begin analysis with quantitative methods to filter a dataset, I do not rely on such methods alone to answer questions that I and my collaborators find important, but rather tend toward manual data coding and qualitative content analysis (e.g., [2, 3]). In general, research questions must be matched with appropriate methods. Despite the power of ML, data scientists must consider the needs and questions of their research community as well as the affordances of the data and determine whether ML or other methods (or some combination) is the best fit.

CONCLUSION

These examples demonstrate some key principles of data science work I have experienced throughout my interdisciplinary research that ties together HCI, data science, atmospheric science, and social science(s). To bridge the gap between data scientists and domain experts, data scientists can translate the work they do for those in the discipline to which they contribute and establish some intersubjectivity with their collaborators. Data scientists can also immerse themselves in that discipline to gain empathy and be able to contribute meaningfully, and not rely solely on collaborators for all disciplinary knowledge. Finally, across data scientists and domain experts, data science needs to be understood as a set of *tools*, and not a panacea for all research inquiries; like any other approach, it has limitations and requires a thorough understanding of one’s data and one’s domain.

REFERENCES

[1] Jennings Anderson, Gerard Casas Saez, Kenneth M Anderson, Leysia Palen, and Rebecca Morss. 2019. Incorporating Context and Location Into Social Media Analysis: A Scalable, Cloud-Based Approach for More Powerful Data Science. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*. https://hdl.handle.net/10125/59666

[2] Melissa Bica, Julie L Demuth, James E Dykes, and Leysia Palen. 2019. Communicating Hurricane Risks: Multi-Method Examination of Risk Imagery Diffusion. In *Proceedings of the CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019)*. 1–13.

[3] Melissa Bica, Leysia Palen, and Chris Bopp. 2017. Visual Representations of Disaster. In *Proceedings of the ACM 2017 Conference on Computer Supported Cooperative Work*. ACM Press, New York, New York, USA, 1262–1276. https://doi.org/10.1145/2998181.2998212

[4] David M Blei and Padhraic Smyth. 2017. Science and data science. *Proceedings of the National Academy of Sciences* 114, 33 (2017), 8689–8692.

[5] Andrew McAfee, Erik Brynjolfsson, Thomas H Davenport, DJ Patil, and Dominic Barton. 2012. Big data: the management revolution. *Harvard business review* 90, 10 (2012), 60–68.

[6] Leysia Palen and Kenneth M. Anderson. 2016. Crisis informatics–New data for extraordinary times. *Science* 353, June (2016), 224–225. https://doi.org/10.1126/science.aag2579