# Data Empathy: A Call for Human Subjectivity in Data Science

**Anissa Tanweer**
University of Washington
Seattle, WA 98195, USA
tanweer@uw.edu

**Brittany Fiore-Gartland**
University of Washington
Seattle, WA 98195, USA
fioreb@uw.edu

**Gina Neff**
University of Washington
Seattle, WA 98195, USA
gneff@uw.edu

**Cecilia Aragon**
University of Washington
Seattle, WA 98195, USA
aragon@uw.edu

## Abstract

Data science is often noted for its apparent ability to displace human subjectivity. We argue that part of the human-centered data science agenda should be to embrace and leverage subjectivity by cultivating data empathy. Data empathy is the ability to share and understand different data valences, or the values, intentions, and expectations around data.

## Author Keywords

Data science, human-centered data science, data valences, data empathy.

## ACM Classification Keywords

H.5.3. Group and organization interfaces: Theory and models.

## Introduction

Data science is often noted for its apparent ability to displace human subjectivity. There is the notion, for example, that “big data” and machine learning algorithms can overturn traditional theory development rooted in the testing of hypotheses constructed from human reasoning [1,8]. And there is the idea that pervasive sensing instruments can subvert the need for self-reported data, illuminating what people “actually” do, as opposed to what they say they do [9,10]. Developments in our ability to computationally collect, store, curate, parse, and analyze data undoubtedly shift the site of human subjectivity in the production of knowledge. But we argue that human subjectivity remains central to the data science endeavor, that this subjectivity should be embraced and leveraged through practices of data empathy, and that data empathy is an integral part of a vision for human-centered data science.

This team of authors has qualitatively studied practices around the collection, sharing, manipulation, analysis, and communication of data in a variety of settings. In this position paper, we garner our collective insights to put forward a vision for human-centered data science based on a principle of data empathy. Empathy means the ability to share in, understand, and identify with the experience of others; to take on another's perspective in an affective or cognitive sense. Data empathy refers to developing this ability for sharing and understanding different data valences, or the values, intentions, and expectations around data.  Data empathy is an ethical and epistemological approach that recognizes manifold, co-existing data valences.

## Human-Centered Data Science vs. Data-Centered Data Science

First, let us consider what an alternative to human-centered data science looks like. Not long ago, one of our subjects was talking about data-driven decision-making in the public sector and remarked, "if you just throw a little data at it, it's got to be better than the way they were doing things before." In the context of the conversation, he was alluding to the undue influence of backroom deals, greasy palms and lobbyists. This concern with unscrupulous politics is valid, but the assumption that simply inserting data into the process will make decision-making more objective or fair is misguided. That is a data-centric version of data science – one that extols the objective purity of data and views human subjectivity as a contaminant. Quite to the contrary, much work has demonstrated that data and algorithms are not inherently objective or fair, and actually have the potential to shape or magnify our biases [2,3,4,6,12,13].

Human judgment is not a contaminant that can be removed from data, but an inherent ingredient in the construction of data sets, which can be thought of as "epistemic achievements that involve categorical judgments" [11:246] and establish patterns of inclusion and exclusion. We argue that although subjective judgment cannot be eliminated, it can and should be acknowledged and accounted for through the consideration of multiple valences in data science. This is not only an ethical position, but an epistemological one: data empathy makes for better, smarter, more powerful data science. Human subjectivity is not poison in the well of data science. It is the antidote.

## Data Valences

Data mean different things to different people at different times and in different contexts. Brittany Fiore-Gartland and Gina Neff call the varied expectations for what data does, "data valences" [5]. Their study surfaces different valences for health and wellness data collected through self-tracking practices: technology companies promoted self-tracking data as a tool for generating actionable insights for the customer; caseworkers used self-tracking data to prompt and organize conversations with patients; clinical practitioners viewed self-tracking data as a liability that requires extra layers of interpretive and managerial work; some patients used self-tracking data for generating self-knowledge through discovery; whereas some patients used self-tracking data to tell their story and validate their experience of illness to themselves and others. Understanding the phenomenon of self-tracking devices, therefore, requires a thoughtful examination of the multiple ways they are being used and made sense of in practice and within different social contexts. This is what we mean by data empathy.

In our observations of data science collaborations, we've seen how valences make a difference in people's work, and how data empathy can strengthen their efforts. In the following example, we discuss how two of the data valences identified by Fiore-Gartland and Neff [5] – transparency and self-evidence – were approached by different players in a data science collaboration, and how data empathy helped participants in the collaboration build trust and make progress on their project.

## Data Empathy in a Data Science Collaboration

The valence of transparency refers to "the benefits of making data accessible, open, shareable, or comparable across cases or contexts" [5:1475], and the valence of self-evidence refers to a view of data as "requiring neither work nor interpretation" [5:1473], such that the data speak for themselves. We saw these valences in tension with one another in a data science collaboration comprised of a team of students and research scientists at a major university, several government agencies, and two non-profit entities. The university team was analyzing information about a suite of public assistance programs in order to determine which programs had the best outcomes.

An academic institute called the Data Science Center (DSC) sponsored the project by paying the salaries of the students and research scientists on the university team, and physically hosting the team's work in their offices. As an organization, part of the DSC's mission is to actively promote open science and reproducible research. In this context, best practices in open science and reproducibility celebrate transparency as a central value and expectation. In practice this translated into a general expectation that any project the institute is involved with will abide by that ethos unless there is a compelling privacy issue over sensitive data. The DSC also encouraged participants in their programs to publicly communicate about and publish aspects of their projects through blogs and other media.

In this case, however, the DSC agreed that before publishing, sharing, or publicly discussing anything about the project, the university team would get input and approval from the external stakeholders, primary among them being the government agencies that generated the data. At first, the university team seemed to view this stipulation as a barrier to transparency and a requirement designed to protect the reputations of the other stakeholders involved. But as the project advanced and the participants were in nearly daily communication with one another, the university team started to understand the concerns of the agencies. A student on the university team talked about her growing recognition that the government agencies felt responsibility for the data, and that relinquishing control over its interpretation was a liability for them:

*I realize they actually love all these cool things we're doing. What they're really concerned about is when people ask them about the things that were presented about their constituency, they need to be able to know where the data came from. [....] So the subset of the data that I took, what went into the filtering? How many people were left out? Hopefully they were left out only because of that filtering that we did, that we were not systematically discriminating against a set of people. I think it's those kind of things that they care about. From our end, we haven't been too careful about that, because we have been pushing ourselves to get to some presentable result.*

Over time, the team realized that the agencies were concerned the data could easily be taken as self-evident, and they insisted on close communication because their knowledge of the data's context of production would be essential to interpreting it. As one of the research scientists on the university team put it:

*They know what's going on in the data. We can't interpret. If we find some weird thing in the data, we have no idea what it is. We don't know whether it's real or not real. Then, you take that back to them and you say, "Okay, what's going on here?" That process was so powerful. I mean, the students just loved it. The [agencies] loved it. [...] I think that's where they started really having more trust, because it felt like we were caring about whether what we were doing made sense.*

In essence, the university team realized the agencies were not trying to obstruct transparency, but were genuinely concerned that the team would be unable to correctly interpret the data because they were far removed from the context of its provenance. As the parties continued to communicate with one another, the university team gained valuable insight into the data through the process put in place for sharing and approval, and the stakeholders gained confidence that the team was treating their data soundly.

In this situation, the university team developed data empathy that allowed them to first understand, and eventually share, the concerns and expectations of other stakeholders in the project. This ultimately strengthened their collaboration and allowed for a richer, more nuanced, and more meaningful engagement with the data. This case underscores the challenges to interpretation posed by the decoupling of data from its context. The agencies were concerned less about releasing the data itself – which was actually publicly available data anyway – and more concerned about losing control over interpretation of the data.

## Conclusion

As it becomes more commonplace for data science to rely on data that is collected from heterogeneous sources or transported across multiple contexts, it is necessary to consider the varied expectations and sense-making practices in which data is enmeshed at every turn. Data empathy is an invitation to explore what these contexts are for different communities and uses, and to unpack the relationships, interpretive structures, values and norms, that shape what data mean and what data do. Unlike simplistic portrayals of data science as a mode of discovery that provides a distant, objective, god's eye view of the world, data empathy puts human subjectivity at the core of a human-centered data science research agenda.

## References

1. Anderson, C. The end of theory: The data deluge makes the scientific method obsolete. *Wired*, 2008. http://www.wired.com/2008/06/pb-theory/.
2. Barocas, S. and Selbst, A. Big Data's disparate impact. *California Law Review 104*, (2016).
3. boyd, d. and Crawford, K. Critical questions for big data. *Information, Communication & Society 15*, 5 (2012), 662–679.
4. Crawford, K. The hidden biases in big data. *HBR Blog Network*, 2013.
5. Fiore-Gartland, B. and Neff, G. Communication, mediation, and the expectations of data: Data valences across health and wellness communities. *International Journal of Communication 9*, (2015), 1466–1484.
6. Gillespie, T., Boczkowski, P.J., and Foot, K.A. Introduction. In T. Gillespie, P.J. Boczkowski and K.A. Foot, eds., *Media technologies: Essays on communication, materiality, and society*. MIT Press, Cambridge, MA, 2014, 1–17.
7. Gitelman, L. and Jackson, V. Introduction. In *"Raw Data" is an Oxymoron*. MIT Press, Cambridge, MA, 2013, 1–14.
8. Hey, T., Tansley, S., and Tolle, K. *The fourth paradigm: Data-intensive scientific discovery.* Microsoft Research, Redmond, WA, 2009.
9. King, G. Ensuring the data-rich future of the social sciences. *Science 331*, 6018 (2011), 719–721.
10. Lazer, D., Pentland, A., Adamic, L., et al. Life in the network: the coming age of computational social science. *Science 323*, 5915 (2009), 721–723.
11. Martin, A. and Lynch, M. Counting things and people: The practices and politics of counting. *Social Problems 56*, 2 (2009), 243–266.
12. The Leadership Conference on Civil and Human Rights. Civil rights principles for the era of big data. 2014. http://www.civilrights.org/press/2014/civil-rights-principles-big-data.html?referrer=https://www.google.com/.
13. [13]United States Executive Office of the President. *Big data: Seizing opportunities, preserving values.* Washington, DC, 2014.