Selected Papers of Internet Research 16:
The 16th Annual Meeting of the
Association of Internet Researchers
Phoenix, AZ, USA / 21-24 October 2015

THE ROLE OF BREAKDOWN IN IMAGINING BIG DATA: IMPEDIMENT TO INSIGHT TO INNOVATION

Anissa Tanweer
University of Washington

Brittany Fiore-Gartland
University of Washington

Cecilia Aragon
University of Washington

Introduction

As the era of "big data" unfolds, researchers across myriad disciplines are increasingly engaging with large, complex datasets distributed across networked technologies. This emerging data-intensive mode of inquiry has been called the "fourth paradigm" of scientific exploration (Hey et al. 2009), inspiring research areas termed "e-science" and more recently "data science," that call for the development of new knowledge infrastructures that support the development of new types of software ecologies (e.g. Borgman 2007; 2015; Edwards et al. 2013). The pursuit of data-intensive scientific discovery in academia means that a wide array of digital technologies and internet-enabled infrastructures that produce, process, manage, store, and analyze the increasing volume, variety, and velocity of data have become inextricable from the everyday work practices of a growing number of scholars. Difficult to envision or engage with in their entirety, these vast and distributed datasets demand new strategies for knowing, seeing, and communicating with data. One such strategy involves breakdown.

We argue that encounters with breakdown – conceived of as points at which progress is stopped due to a material obstacle – represent essential sites of knowledge production for data science and any other big data analysis.

Description of Study

This research is part of an ongoing ethnographic study of data science communities and collaborations in academia. For this paper, we embedded ourselves within a recurring Data Science Collaboration (DSC) program that takes place over the course of a few

Suggested Citation (APA): Tanweer, A., Fiore-Gartland, B., Aragon, C. (2015, October 21-24). *The role of breakdown in imagining big data: Impediment to insight to innovation*. Paper presented at Internet Research 16: The 16th Annual Meeting of the Association of Internet Researchers. Phoenix, AZ, USA: AoIR. Retrieved from http://spir.aoir.org.

months at a large public university. This program brings together data science methodology experts with domain researchers across a range of disciplines, from astronomy to oceanography to political science, to collaborate on data science projects throughout the academic term. A central feature of the program is the co-location of project collaborators two days per week in an effort to advance collaboration and productivity in a short period of time. Our research is based on participant-observation within the space of co-location, semi-structured interviews with all participating data science methodology experts and domain researchers, and archival analysis of project documentation and communication that occurred online.

Breakdowns in Big Data

Participants in the DSC were conducting research on massive data sets. For example, an astronomer in the group was analyzing pixels from a telescopic sky survey contained in a database with a trillion rows that was stored in the cloud and analyzed using distributed virtual machines. Another participant was a political scientist conducting textual analysis on the entire 90-terabyte corpus of web pages published in a single web domain over the course of several years. Working with data at this scale and volume obscures its materiality and renders it invisible and unknowable in its entirety. Someone manually entering information into an Excel spreadsheet can see the entire data set and be intimately familiar with its contents. This is clearly not the case for someone working with 90 terabytes of html code or a table with a trillion rows - these big data sets are partially obscured from human comprehension by their sheer volume.

Susan Leigh Star has demonstrated the ways in which obscured information infrastructure “becomes visible upon breakdown” (1999, p. 382), while Graham and Thrift have highlighted the way breakdown brings materiality to the fore: “Things only come into visible focus as things when they become inoperable – they break or stutter and they then become the object of attention” (2007, p. 2).

In an essay on the value of “broken world thinking” as a lens for furthering media and technology studies, Steve Jackson refers to breakdown in two senses: as the inevitable decay of systems under the inescapable law of entropy, and as points of breakage resulting from “bumping up against the limits of existing protocols and practices” (2013, p. 228). In the context of the DSC, we understand breakdown in the latter sense, as a point of stoppage forced by a material obstacle. We observed numerous instances in which encounters with such breakdown illuminated the materiality of the data and provided an occasion for the researchers to envision its content and structure.

In one case, Louis was surprised that a dataset didn’t take up very much space on a disk, yet when he opened his data in the statistical software package R, it unexpectedly exceeded 60 gigabytes of RAM and maxed out his computer’s processing capabilities, making it impractical to manipulate without a distributed computing system. In trying to figure out why the data seemed to be “growing” to an unmanageable size, Louis and his data science mentor realized that R, by default, was processing the data as a dense matrix, in which every coordinate contained a value. The mentor realized that in actuality, Louis had a sparse matrix, with most of the coordinates containing zeroes. In this example, the encounter with breakdown led to important insights about the data’s

content (it consisted mostly of zeroes) and its structure (it could be organized as a sparse matrix).

Not only does breakdown serve to foreground the otherwise invisible, but a number of scholars have also noted that breakdown often leads directly to productivity (e.g., Petroski, 1985; Graham & Thrift, 2007), and that "[innovation's] engine is breakdown and repair" (Jackson, 2013, p. 228). In keeping with this vein of thinking, we observed how encounters with breakdown in the DSC not only surfaced the materiality of big data, but also how they generated important insights that the researchers and data scientists then leveraged for innovation and discovery. For example, after Louis' encounter with breakdown generated insight into the nature of his data, he and his mentor developed a strategy for writing code that would allow them to work with a sparse matrix in R. This was a novel innovation for the software package and reduced RAM consumption by compressing the data to a manageable level.

By focusing on breakdown in this example and others from the DSC, we were able to trace the iterative process of envisioning the data, generating insights into the data, translating insights into strategies for managing the data, and encoding those strategies into reparative communication with the data.

Conclusion

Our ethnographic study of data-intensive research in an academic setting highlights the importance of the negotiations between the partial images of data and the material interactions with data. We find that encounters with breakdown became sites that inspired new ways of knowing, seeing, and communicating with one's data. Often dismissed as impediments that slow or derail a typical process of scientific inquiry, we argue that these encounters are underappreciated resources for innovation and productivity, and that they represent essential sites of knowledge production for data science and any other big data analysis. Thought of in this way, productive encounters with breakdown have significant design implications, in that the data science community may benefit from tools and platforms that facilitate communication between data and researcher by more explicitly calling attention to points of breakdown.

References

Borgman, C. L. (2007). *Scholarship in the digital age*. Cambridge, MA: MIT Press.

Borgman, C. L. (2015). *Big data, little data, no data: Scholarship in the networked world*. Cambridge, MA: MIT Press.

Edwards, P. N., Jackson, S. J., Chalmers, M. K., Bowker, G. C., Borgman, C. L., Ribes, D., Burton, M., & Calvert, S. (2013) Knowledge infrastructures: Intellectual frameworks and research challenges. Ann Arbor: Deep Blue. http://hdl.handle.net/2027.42/97552.

Graham, S. & Thrift, N. (2007). Out of order: Understanding repair and maintenance. *Theory, Culture & Society*, *24*(3), 1–25. doi:10.1177/0263276407075954

Hey, A., Tansley, S. & Tolle, K. (Eds.) (2009). The Fourth Paradigm: Data-intensive scientific discovery. Microsoft Research. http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_complete_lr.pdf

Jackson, S. J. (2013). Rethinking repair. In T. Gillespie, P. J. Boczkowski, & K. A. Foot (Eds.), *Media technologies: Essays on communication, materiality, and society* (pp. 221–239). Cambridge, MA: MIT Press.

Petroski, H. (1985). To engineer is human: The role of failure in successful design. London: Macmillan.
Star, S. L. (1999). The ethnography of infrastructure. *American Behavioral Scientist*, *43*(3), 377-391.