Toward a Hermeneutics of Data

Amelia Acker
University of Pittsburgh

Editor: Bradley Fidler

Recently I watched an all women panel on careers in data science hosted by the University of California, Berkeley's iSchool.[1] The panel members had a range of backgrounds and training, from advertising to educational research, statistics, and topic modeling. Some of the roundtable's experts had PhDs, and a few had MBAs. Each of the panelists worked at Bay Area startups and commerce sites in northern California (think Airbnb, Eventbrite, and Jawbone). These corporate data scientists represent a promising—and fast paced—new field of commerce, analytics, knowledge, and perhaps most importantly, technical change in the present world of networked computing. I was struck by the variety of different ways these information professionals approached the idea of "data" as they were speaking about the nature of their work. The engaging discussion on data science illustrated how data is not just a byproduct of computing technologies but an engine for dynamic change that drives society in different, fascinating directions.

Data science is the systematic process of creating, building, and organizing knowledge with data. It has recently become a "new" area of interest in computing sciences, bioinformatics (including public health), learning sciences, business and marketing, and the information sciences. Higher education institutions have begun to offer master's degrees in data science—few programs exist at the undergraduate or doctoral level, but many are soon to come.[2] The "newness" of data science has become all the rage of late, but for some, it's just a fresh coat of paint. As others have taken pains to point out, the discipline of data science simply appears to consolidate and leverage principles and techniques from a number of fields that already exist, such as statistics, machine learning, knowledge management, and information retrieval.[3,4] What's new is that data science aims to confront the massive volumes of data created and collected today. Looking closely at data now that it is big can inspire us to ask questions about how it has been handled, modified, managed, and circulated since people started leveraging data with information systems and computing machines.

New academic programs aren't the only place where we are seeing the impact of the "data deluge." Increasingly, we are seeing a public consciousness around personal data generation and collection by states and corporations. Data collection (telephony metadata, in particular) has come under intense, international political debate since the Snowden leaks in 2013. Earlier this spring, the US Circuit Court of Appeals for the Second Circuit found that the bulk collection of telephony metadata by the US National Security Agency (NSA) is not authorized by the USA PATRIOT Act, saying that the collection "exceeds the scope of what Congress has authorized."[5] Since the Snowden leaks, media coverage, online activism, and political pressure from around the world brought the normally banal term "metadata" to center stage despite the fact the collection of data about citizens is far from a recent development in surveillance states.

Consumers are increasingly aware that the online traces they create generate data that can be aggregated and turned into black gold. We're also seeing consumer backlash against the aggregation, collection, and data protection that has resulted in numerous security breaches to information systems that regularly put consumers, workers, and citizens at risk. Ethnographers, legal theorists, and communication scholars have suggested that new cryptocurrencies and data-obfuscation techniques in email encryption[6,7] have, in part, stemmed from this new consumer consciousness about how user data is aggregated and applied into new commercial products. From Home Depot and Target to the Office of Personnel Management hacks, social media users and ordinary citizens are facing security breaches that increasingly reveal the staggering amount of information that is collected through networked infrastructures about their behavior, preferences, relationships, and activities. Although citizens' concerns about data collection have existed for many decades,[8] and metadata and surveillance programs that leverage data into commercial applications and state governance are not new issues, I'm interested in asking how historians of computing are confronting new conceptions of data that circulate in society—in academic, commercial, and civic spheres—and what we might have to contribute to new scholarship about data.

In the last Think Piece article in the *Annals,* William Aspray presented information domains as a promising area for computing historians to consider as way of getting at the "larger meaning of information and information technology in society."[9] I want to take Aspray's argument a little further, building off of his notion of information domains such as data curation or archival science, and suggest that data—how it is being created, packaged, deployed, and understood—is fruitful for computing historians to consider as part of a larger trend over the last 50 years toward networked information systems. For information scholars, the difference between information and data is context. A piece of data without context is without meaning, but when

Published by the IEEE Computer Society

data is put into context through practices such as aggregation, description, classification, organization, or application, it becomes meaningful information to people and machines. Data that has become information may also have multiple layers of context or acquire more contextual information over time. For example, in the US, birth records are connected to social security numbers that can be aggregated in the Social Security Death Index (SSDI) database of death records. A database of death dates also carries lifetimes of information, including legacy information systems (such as analog birth and death records) as well as evidence for other kinds of data (such as population statistics). The data aggregated and classified in the SSDI database acquire multiple layers of context and carry multiple ontologies about the categories of "life" and "death" depending on how the information is accessed and interpreted. This is just one example of how data can become information with different layers of context.

Here, I want to bring data to the fore, the ways that it figures into different realms, as a phenomena that increasingly seems to penetrate all information domains, including fields of scholarship and areas of society. I argue that historians of computing committed to documenting, charting, understanding, and explaining technological change can expand and shape a growing area of scholarly research, which is increasingly being called "data studies" and which has strong and direct links to the history of computing research agenda.

Historians of computing committed to documenting, charting, understanding, and explaining technological change can expand and shape a growing area of scholarly research.

Data as Computing History

Data, as traces of transmission, are becoming the fundamental organizing principle of emerging cultural records that represent vast swaths of data created as part of networked computing infrastructures.[5] In my work on new information objects created with mobile computing infrastructures, I am particularly drawn to the origins of how data, or traces of data, come to be in information systems. In my earlier work on the Short Message Service (SMS) text messaging protocol,[10] I was concerned with the metadata that encapsulated text messages as part of their transmission across wireless networks. The metadata of text messages (not the message content itself) are used to route network traffic information and to locate senders and recipients of text transmissions in wireless networks. The NSA surveillance programs that were uncovered in 2013 showed us that these context traces, the data about text messages, are useful and collected in all sorts of ways by network operators, handset manufactures, standards organizations, and surveillance programs.[11] These routing data are metadata, which represent their own kind of documentation, records of transactions between people, institutions, machines and cultures through time. The existence of new kinds of data, such as telephony metadata from text messages, points to a shift in the history of recorded information and the ways we communicate with mobile networks that is different from earlier, analog communication networks. We need metadata about transactions for the networked information infrastructure to work. Histories of data help us understand how these layers of context and meanings are acquired through their development, stabilization, and circulation.

Scholars have studied the emergence of the data collection, privacy, and the surveillance society as social constructions since the 1960s, and they can help us begin to make sense of this current data deluge. For example, JoAnne Yates,[12] Geoffrey Bowker and Susan Leigh Star,[13] and Christine Borgman[14] have examined the origins of new kinds of documents, formats, and information objects in information infrastructures in distinct eras and expert domains. Other information infrastructure scholars such as Michael Buckland,[15] David Ribes,[16] and Matthew Mayernik[17] have analyzed the stabilization of formats, systems, and standards and their influence on computing in cultures of information and documentation work. There has also been a spate of work that focuses on how these traces of transactions in the histories of networks has shifted away from organizational and expert cultures to see how new data subjects are developing.[18–20] Still other scholars, influenced by Michel Foucault, come to the study of data traces in information infrastructures by way of privacy and data collection techniques under legislation, network architecture, and technical politics.[21–24] A final body of scholarship points to the ways that the Internet supports new modes of being, such as the "algorithmic self," where users create corpora of personal data traces across social media platforms.[25–27]

This is certainly not an exhaustive list of the work being done in data studies, but it is evidence that a growing number of social scientists, media scholars, and organizational theorists are engaging with data in the recent history of technology. Still, few examine the origins and stabilization of data as a focal point. If the rise of data science, legislation around data collection, and consumer consciousness toward data generation is part of everyday life, how can historians of computing help apprehend data and its growing centrality to the information domain of data scholarship in particular?

One way might be to analyze and describe how network infrastructures are created by the generation and implementation of data, which can provide a way for us to examine the development and design of networking architecture and technologies, in what Andrew Russell calls, "histories of networking."[28] Yet, data and infrastructures have always had an intertwined existence, and the entanglement has become tighter and harder to distinguish, describe, and interpret with new Internet technologies and next-generation wireless networks.[29] In a relatively short time (less than 25 years), mobile computing with handsets has become the primary way of communicating information in terms of volume, frequency, and penetration for much of the developed world.[30] Clearly, historians of computing must account for digital traces and new formats such as telephony metadata, but it is uncertain whether existing approaches that describe data can account for the complexities of today's networked infrastructures. To my mind, data's impact on society and studies of data have reached a point for which it is now time for historians of computing to historicize data directly. And there needs to be an equitable balance between studying the effect of data and studying context—the processes of its creation, stabilization, and transmission in information infrastructures. Given the possibilities of emerging data in contemporary society, it is time for us to consider that data may become a central part of the history of computing, and it will need to be for the foreseeable future.

It is time for us to consider that data may become a central part of the history of computing, and it will need to be for the foreseeable future.

I propose that one way to do this might be to look at data within different scales of infrastructure, as Paul N. Edwards has suggested.[31] Studying data at different scales produces different views of how technology develops as well as how specific technologies affect individual practices (such as recordkeeping or evidence building) and in organizational practices (like business communications), as James Cortada has extensively documented.[32,33] Building upon Thomas Misa's framework for scalar analysis,[34] Edwards describes the micro-, meso-, and macro-scales of society and how infrastructure can be approached in different ways at each scale. Micro refers to the individual or personal level, the day-to-day practices that make up our lives. The meso-scale is the organizational or institutional change that we see with groups of people across weeks and years. Finally, the macro-scale refers to infrastructure over long periods of time, decades or even centuries (what some have called the "long now"[35]).

The beauty of approaching data as it moves through information infrastructures at different scales of analysis is that scales of inquiry are adaptable, like a pocket telescope, extensible and collapsible with quick gestures. Although many studies of data in computing history are at the macro-scale, I'm particularly drawn to the meso-levels of infrastructures, where people create and rely upon new forms of data as information. It is at the meso-level where ethnographers who examine information systems (such as Peter Botticelli,[36] Kalpana Shankar,[37] and Susan Leigh

What will the future of data studies be? How can historians be faithful to particular information ages' data and distinguish them?

Star[38]) all find rapid change—in this messy, in-between area where groups of people are communicating with documents and where the stuff of data creation, stabilization, reception, and circulation actually happens.

The Future of Data Studies

We have benefited from the applications of information domains, theories of infrastructure, and histories of computer networks, but I believe that now we should turn toward studies of data at different scales of information infrastructure. Careful studies of data, and their interpretation and development in histories of networking can tell us more about context, change, and continuities over time when it comes to computing more broadly. There is a role for historians of computing to tell us more about how this moment of data science came to be by looking at information domains through data and the ways data acquires layers of context to become information. Aspray, and others have argued that histories of computing, and the Internet in particular, have been too focused and limiting.[28,39,40] A hermeneutics of data is needed at the micro-, meso-, and macro-scales of networked infrastructures. In asking the field to turn toward the "newness" of data in this moment, I am not arguing that we should jump on the "big data" bandwagon. Instead, I am asking for us to consider how and why data came to be viewed as new again, arguably one of the major cultural developments of the past decade across national boundaries and across fields of expertise and practice.

Young investigators at the nexus of computing history, information infrastructure studies, and communications have begun to examine data at different scales of infrastructure in some interesting and fascinating ways. For example, Brian Beaton has recently written about software that promotes new types of everyday data gathering with mobile devices, and he calls for specific groups of social actors to rework their social relations around continuous data exchange and to form themselves into new types of networked subjects (what he calls "crowdsourced selves").[41] Kevin Driscoll has written about the social history of database technologies, finding that data structures within collections can heavily influence the flavors of database populism that may arise and recordkeeping possibilities with systems such as in social media.[42] Megan Finn has examined analog data artifacts during the 1857 Tejon earthquake, showing that historical information infrastructures of the period before standardized timekeeping shaped popular understandings of natural disasters.[43]

Social histories such as these, of data moving through scales of information infrastructures, represent a valuable intervention into histories of networking but also into studies of information, system design, and communication technologies in different realms of society. What will the future of data studies be? How can historians be faithful to particular information ages' data and distinguish them? Some readers may find this call too concerned with the present, but with a turn toward the study of data in context, we can disinter the ways in which infrastructures of transmission shape recorded information, in this moment and over time. Computing historians are uniquely positioned to probe the entanglement of networked infrastructures, data, and cultures of computing in the recent past and near future. But the key requirement is to first elevate data, making it central to computing history, as it already is within society.

References and Notes

1. Berkeley iSchool, "Technology for the Greater Good: Careers in Data Science," video, 2014; www.ischool.berkeley.edu/events/20140903careersnindatascience/video2.
2. M. O'Neil, "As Data Proliferate, So Do Data-Related Graduate Programs," *Chronicle of Higher Education*, 3 Feb. 2014.
3. B. Darrow, "Data Science Is Still White Hot, But Nothing Lasts Forever," Future of Work blog, *Fortune*, 21 May 2015; http://fortune.com/2015/05/21/data-science-white-hot/.
4. J. Furner, "Information Science Is Neither," *Library Trends*, vol. 63, no. 377, 2015, pp. 362–363.

5. United States Court of Appeals for the Second Circuit, *ACLU v. Clapper*, 7 May 2015, p. 97.
6. B. Maurer, T.C. Nelms, and L. Swartz, "'When Perhaps the Real Problem Is Money Itself!' The Practical Materiality of Bitcoin," *Social Semiotics*, vol. 23, no. 2, 2013, pp. 261–277.
7. F. Brunton and H. Nissenbaum, "Vernacular Resistance to Data Collection and Analysis: A Political Theory of Obfuscation," *First Monday*, vol. 16, no. 5, May 2011; http://firstmonday.org/article/view/3493/2955.
8. S.E. Igo, "The Beginnings of the End of Privacy," *The Hedgehog Rev.*, vol. 17, no. 1, Spring 2015; www.iasc-culture.org/THR/THR_article_2015_Spring_Igo.php.
9. W. Aspray, "Information Society, Domains, and Culture," *IEEE Annals of the History of Computing*, vol. 37, no. 2, 2015, pp. 2–4.
10. A. Acker, "The Short Message Service: Standards, Infrastructure and Innovation," *Telematics and Informatics*, vol. 31, no. 4, 2014, pp. 559–568.
11. S. Landau, "Making Sense from Snowden: What's Significant in the NSA Surveillance Revelations," *IEEE Security & Privacy*, vol. 11, no. 4, 2013, pp. 54–63.
12. J. Yates, *Control Through Communication: The Rise of System in American Management*, JHU Press, 1993.
13. G.C. Bowker and S.L. Star, *Sorting Things Out: Classification and Its Consequences*, 1st ed., MIT Press, 1999.
14. C.L. Borgman, *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*, MIT Press, 2007.
15. M.K. Buckland and R.R. Larson, "Metadata as Infrastructure: What, Where, When and Who," *Proc. Am. Soc. Information Sciences and Technology*, vol. 42, no. 1, Jan. 2005.
16. D. Ribes, "Ethnography of Scaling, or, How to a Fit a National Research Infrastructure in the Room," *Proc. 17th ACM conf. Computer-Supported Cooperative Work & Social Computing*, 2014, pp. 158–170.
17. M.S. Mayernik, "Research Data and Metadata Curation as Institutional Issues," *J. Assoc. Information Science and Technology*, preprint, 30 Jan. 2015; http://onlinelibrary.wiley.com/doi/10.1002/asi.23425/abstract.
18. W.H.K. Chun, *Programmed Visions: Software and Memory*, MIT Press, 2011.
19. Rita Raley, "Dataveillance and Countervailance," *Raw Data Is an Oxymoron*, L. Gitelman, ed., MIT Press, 2013, pp. 121–146.
20. A.R. Galloway, *Protocol: How Control Exists after Decentralization*, MIT Press, 2006.
21. S.E. Landau, *Surveillance Or Security?: The Risks Posed by New Wiretapping Technologies*, MIT Press, 2010.
22. K. Shilton, "Participatory Personal Data: An Emerging Research Challenge for the Information Sciences," *J. Am. Soc. Information Science and Technology*, vol. 63, no. 10, 2012, pp. 1905–1915.
23. L. DeNardis, "Hidden Levers of Internet Control," *Information, Communication & Society*, vol. 15, no. 5, 2012, pp. 720–738.
24. K. Crawford, "When Big Data Marketing Becomes Stalking: Can Data Brokers Be Trusted to Regulate Themselves?" *Scientific Am.*, 28 Jan. 2004; www.scientificamerican.com/article/when-big-data-marketing-becomes-stalking/.
25. I. Gershon, *The Breakup 2.0: Disconnecting over New Media*, Cornell Univ. Press, 2011.
26. A.E. Marwick, *Status Update: Celebrity, Publicity, and Branding in the Social Media Age*, Yale Univ. Press, 2013.
27. A.N. Markham et al., "Algorithmic Identity: Networks, Data, and the Terrible Beauty of the Black Box," Selected Papers of Internet Research, 2014; http://spir.aoir.org/index.php/spir/article/view/891/466.
28. A.L. Russell, *Open Standards and the Digital Age*, Cambridge Univ. Press, 2014.
29. B. Fidler and A. Acker, "Metadata and Infrastructure in Internet History: Sockets in the Arpanet Host-Host Protocol," *Proc. 77th ASIS&T Annual Meeting*, vol. 51, no. 1, 2014.
30. Int'l Telecommunications Union, "World Telecommunication/ICT Indicators database," 18th ed., Dec. 2014; www.itu.int/en/ITU-D/Statistics/Pages/publications/wtid.aspx.
31. P.N. Edwards, "Infrastructure and Modernity: Force, Time, and Social Organization in the History of Sociotechnical Systems," *Modernity and Technology*, T.J. Misa, P. Brey, and A. Feenberg, eds., MIT Press, 2004, pp. 185–225.
32. J.W. Cortada, *The Digital Hand: Volume II: How Computers Changed the Work of American Financial, Telecommunications, Media, and Entertainment Industries*, Oxford Univ. Press, 2005.
33. J.W. Cortada, *Information and the Modern Corporation*, MIT Press, 2011.
34. T.J. Misa, "How Machines Make History, and How Historians (and Others) Help Them to Do So," *Science, Technology, & Human Values*, vol. 13, nos. 3–4, 1988, pp. 308–331.
35. D. Ribes and T.A. Finholt, "The Long Now of Technology Infrastructure: Articulating Tensions in Development," J. *Assoc. Information Systems*, vol. 10, no. 5, 2009, pp. 375–398; http://search.proquest.com/openview/12a0d1f3490378e6817f3fc4ed200341/1?pq-origsite=gscholar.
36. P. Botticelli, "Records Appraisal in Network Organizations," *Archivaria*, vol. 1, no. 49, Jan.

2000; http://journals.sfu.ca/archivar/index.php/archivaria/article/view/12743/13929.

37. K. Shankar, "Ambiguity and Legitimate Peripheral Participation in the Creation of Scientific Documents," *J. Documentation*, vol. 65, no. 1, 2009, pp. 151–165.
38. S.L. Star, "The Politics of Formal Representations: Wizards, Gurus, and Organizational Complexity," *Ecologies of Knowledge: Work and Politics in Science and Technology*, S.L. Star, ed., SUNY Press, 1995, pp. 88–118.
39. W. Aspray and B.M. Hayes, *Everyday Information: The Evolution of Information Seeking in America*, MIT Press, 2011.
40. T. Haigh, A.L. Russell, and W.H. Dutton, "Histories of the Internet: Introducing a special issue of Information and Culture," *Information & Culture*, vol. 50, no. 2, 2015, pp. 143–159.
41. B. Beaton, "Safety as Net Work: 'Apps Against Abuse' and the Digital Labour of Sexual Assault Prevention," *MediaTropes*, vol. 5, no. 1, 2015, pp. 105–124.
42. K. Driscoll, "From Punched Cards to 'Big Data': A Social History of Database Populism," *communication +1*, vol. 1, no. 1, 2012, article no. 4.
43. M. Finn, "Information Infrastructure and Descriptions of the 1857 Fort Tejon Earthquake," *Information & Culture*, vol. 48, no. 2, 2013, pp. 194–221.

***Amelia Acker** is an assistant professor in the School of Information Sciences at the University of Pittsburgh. Her research interests include information infrastructure studies, archival science, and data studies, specifically the material production and transmission of information objects in networked recordkeeping systems over time. Acker has a PhD in information studies from the University of California, Los Angeles. Contact her at aacker@pitt.edu.*

cn *Selected CS articles and columns are also available for free at http://ComputingNow.computer.org.*