Commentary

BIG DATA & SOCIETY

Big Data & Society
July–December 2015: 1–4

DOI: 10.1177/2053951715608877
bds.sagepub.com
SAGE

Big Data and reality

Ryan Shaw

Abstract

DNA sequencers, Twitter, MRIs, Facebook, particle accelerators, Google Books, radio telescopes, Tumblr: what do these things have in common? According to the evangelists of "data science," all of these are instruments for observing reality at unprecedentedly large scales and fine granularities. This perspective ignores the social reality of these very different technological systems, ignoring how they are made, how they work, and what they mean in favor of an exclusive focus on what they generate: Big Data. But no data, big or small, can be interpreted without an understanding of the process that generated them. Statistical data science is applicable to systems that have been designed as scientific instruments, but is likely to lead to confusion when applied to systems that have not. In those cases, a historical inquiry is preferable.

Keywords

Big Data, software architecture, data modeling, design, historical methods, ontology

From 2005 to 2007 I was a Social Media Researcher at Yahoo Research Berkeley (YRB), part of a group of graduate students, UC Berkeley faculty, and professional researchers with a remit to study the vast stores of data generated by Yahoo's social media "properties" like Flickr and Delicious, and to generate and experiment with ideas for new properties (Qi, 2005; Shamma et al., 2007). I recall one afternoon listening to another researcher present some slides describing the current state of the Flickr dataset. He was projecting on the screen a plot of the number of photographs uploaded per user account. The plot had the typical power law shape characteristic of many variables in social media datasets (a few people uploading many photos, many people uploading a few or no photos) but it also had spikes at regular intervals—every multiple of six. After letting the room speculate for a while about why Flickr users seemed to prefer uploading in batches of six, the presenter showed the next slide: a screenshot of Flickr's upload page, a grid of six forms, each of which could be used to select a JPEG from the user's file system and add some textual description. Flickr users hadn't been freely choosing to upload in batches of six, but neither had they been forced to; it simply had been strongly suggested that they should.

Software interfaces are suggestive, sometimes literally so. One of the phenomena we were particularly interested in at YRB was "tagging." Sites like Flickr and Delicious allowed users to add uncontrolled keywords or "tags" to items such as the photos they uploaded or the URLs they saved, and we were interested in understanding how these tags were chosen and used. One complication of studying datasets of tag assignments was that some sites had interfaces that, after a user had typed a few letters of a tag, would suggest possible completions, one of which could then be chosen with a keystroke. How this autocompletion influenced users' tag assignments would depend on how autocompletion was implemented. One implementation might suggest tags most often assigned by other users to the same item, while a different implementation might suggest tags most often assigned by that user to any item in the past. The choice between these two alternatives (which are far from exhaustive) is constrained by the underlying data model. The first implementation assumes that it is meaningful for two

School of Information and Library Science, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Corresponding author:
Ryan Shaw, School of Information and Library Science, University of North Carolina at Chapel Hill, CB #3360, 100 Manning Hall, Chapel Hill, NC 27599-3360, USA.
Email: ryanshaw@unc.edu

different user accounts to assign the same tag to the same item (otherwise why bother suggesting a tag that has already been assigned to that item by another account?). This assumption in turn requires a data model in which a tag is modeled as a relation between a user account and an item. If instead a tag was modeled as a property of an item, then a given tag could only be assigned to a particular item once.

Decisions about how to model user accounts, items, tags, and the relations among them influence, and are influenced by, decisions about how the tagging interface should look and work and decisions about the algorithm used to implement autocompletion. These decisions in turn reflect decisions about the nature and purpose of the system. An image-sharing system intended to be used primarily for re-sharing images originating elsewhere (for example news photos, memes, or screenshots) might choose to model tags as relations between user accounts and images, implying that different accounts' tags for an image express different interpretations of it. On the other hand an image-sharing system intended for sharing one's own photographs might choose to model tags as properties of an image, implying that the tags express an authorized description or categorization. Of course tools are not always used as intended, but an understanding of how a system was intended to be used, and how those intentions were hypostatized in the system's design, can help one interpret users' conduct: to what extent are they working "with" or "against" the system? Taken as a whole, the data model and other design decisions made during the development of a software system constitute the system's *architecture* (Taylor et al., 2010: 1).

Understanding a software system's architecture is part of recognizing what Kirschenbaum (2008) has called the *formal materiality* of the system, the effects of a set of choices that determine what will be easy to do—"frictionless," to use the industry lingo (Wikipedia, 2014)—and what will not. Even the simplest piece of software has embedded within it a series of architectural decisions about what "works" with respect to the purposes for which it was created. This architecture is layered and constantly evolving. Software engineer Jean-Baptiste Quéru (2011) wrote elegantly of the "dizzying but invisible depth" of layered complexity that characterizes contemporary software architecture. Engineers try to manage this complexity by treating systems as consisting of layers, abstractions which help one avoid having to think too much about the dizzying complexity of the overall system. In theory this means that the architectural decisions made at a given layer only have ramifications for adjoining layers. The ability to make new decisions at a given layer without re-engineering the entire stack is what allows software systems to evolve. Evolvability is a critical requirement for any real-world system, but it further complicates the problem of mapping the ontological terrain of software, as that terrain is constantly shifting: interfaces are redesigned, algorithms tweaked, data remodeled. Sometimes these changes reflect deeper changes in the designers' conception of the system, which in turn can result from their observations of how the system is being used. Flickr, for example, started life as a chat network. Only later, after a critical mass of users began using it for sharing their own original images, did the Flickr designers reconceive it as a photo-sharing site (Hoopes, 2004).

To study Big Data is to study the traces left behind by the use of a large, complex, and constantly evolving software system. These traces excite many social scientists, as they seem to provide fine-grained documentation of social or cultural "transactions." The dominant form of studying these traces is "data science," which treats the large software systems that generated them as measuring instruments (Loukides, 2010). *Wired* editor Chris Anderson (2008) famously proclaimed that the sheer quantity of data produced by such systems made them scientific instruments, even if they lacked any coherent model informing their design.

But it is not heaps of transactional data that make an inquiry scientific. Being scientific is an effect of work done to establish stable, quantifiable concepts, and the aim of science is to establish resilient statistical relationships among those concepts (Oakeshott, 2002: 175–178). The statistical relationships emerge from the data, but the stable, measurable concepts do not: the concepts are a prerequisite for the existence of the data. Thus data scientists must design and engineer measuring instruments that will produce data usable within their conceptual framework; the architecture of those instruments must cohere with that framework. Researchers at Facebook and the other corporate owners of Big Data-generating systems recognize this, and in these organizations data scientists work with the engineers designing the systems to establish their usefulness as measuring instruments (Fiore, 2015). This is one possible path for social and cultural inquiry: scientists closely cooperating with engineers to simultaneously build massive software systems and study the behavior of people using them. It is a path that leads social research outside the academy, into a few massively resourced private or government-run research labs (Williamson, 2014).

In 1978 a research programmer at IBM named William Kent wrote *Data and Reality*, a meditation upon the complexities of data modeling. The book surveyed the problem of how the data models that form part of the architectures of information systems relate to our shared reality. Kent's examples were mundane: books in a library, parts in warehouses, and players on

sports teams. He demonstrated that, even for such simple applications, deciding how to store data involves answering a number of interrelated questions. What is one thing? How many things are there? What kinds of things are there? How real are they? How long do they last? Kent emphasized that there are no right answers to such questions: different people in different contexts with different goals will choose different answers as they construct their data models. Data models are practical tools; like maps, they are "correct" to the extent that they get you where you want to go. Furthermore tools are evaluated according to a host of criteria that have little to do with correctness (Kent, 1978: 194). How much do they cost? How fast do they run? How often do they break? How often do they need to be updated? How much training do they require to use? How quickly do they become obsolete? What guarantees do their makers provide? As a result, the data stored and emitted by software do not reflect a coherent theory of the world, but "an amalgam of fragments of theories" (Kent, 1978: 194), small pieces pragmatically selected and loosely joined.

Kent's reflections cast doubt on the possibility of data science. A scientific community requires measuring instruments that make manifest a conceptual framework widely shared by that community. Kent, after 200 pages of reflecting on his experience designing systems for data processing, concluded by highlighting how difficult it is to achieve such a shared view through the mediation of an information system. If the scope of such a system is sufficiently local and limited, he wrote:

> we can share a common enough view of [reality] for most of our working purposes, so that reality does appear to be objective and stable. But the chances of achieving such a shared view become poorer when we try to encompass broader purposes, and to involve more people. This is precisely why the question is becoming more relevant today: the thrust of technology is to foster interaction among greater numbers of people, and to integrate processes into monoliths serving wider and wider purposes. It is in this environment that discrepancies in fundamental assumptions will become increasingly exposed. (Kent, 1978: 203)

The broader the array of uses of a system, the less conceptual coherence its architecture will exhibit, and the less useful it will be as a scientific instrument. If Kent was right, we need not data science but data history.

A historical approach would treat tweets and posts and comments and links and all the rest not as scientific observations but as

> exploits, human doings which have been performed, utterances which have been pronounced, artefacts which have been made, fragments of the bygone purposive engagements of their perhaps unknown authors which have survived (although sometimes recognizably damaged) and are themselves now present. (Oakeshott, 2002: 51)

This shift, from viewing Big Data as scientific measurements toward viewing them as traces left by past engagements, changes the character of Big Data-driven inquiry. Treated as the subject of a scientific inquiry, 100 million tweets are a series of observations generated by the same implicit and unchanging mechanism, the nature of which is to be discerned via statistical generalization from that series. Treated as the subject of a historical inquiry, 100 million tweets are an assembly of individual utterances, the circumstantial relations among which must be discerned through a process of mutual criticism and interpretation. From these circumstantial relations one may be able to infer something about the practices and conventions of Twitter users and designers. Twitter users participate in various complexes of purposive activity—fandom, recruitment, hashtag activism, bots, spam, "weird Twitter," and so on. These practices leave traces that are interpreted by Twitter designers—not only the designers employed by Twitter, but anyone who designs software that interoperates with Twitter.[1] The designers in turn discourage certain practices and encourage others via architectural decisions, decisions that are influenced not only by their interpretations of user practices but also by available technologies, competing products, and prevailing fashions of software development. This process does not work to stabilize a set of assumptions about social reality; but perhaps careful interpretation of the big and small data this process leaves behind can tell us something about the social reality of which it is a part.

Louis Mink, arguing for the independence of historical understanding from scientific explanation, wrote that:

> ... the more I know about the facts of the case, the more necessary it becomes to use something like empathy in order to convert an indigestible heap of data into a synoptic judgment by which I can "see together" all these facts in a single act of understanding. Otherwise, if I am asked what I have learned, I can only point mutely to my filing cabinet. (1966: 42)

The study of Big Data could lead to a more comprehensive understanding of social reality. But achieving that understanding will require developing a sense of the complex materiality of our Big Data-producing information systems, and empathy for the people who fund, design, build, use, and exploit them. Without that

sense and empathy, when we are asked what we have learned from Big Data, we may be left pointing mutely at our data centers.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Note

1. In 2011 Twitter claimed that over 750,000 software developers had created applications interoperating with their "ecosystem."

References

Anderson C (2008) The end of theory: The data deluge makes the scientific method obsolete. *Wired*, 23 June. Available at: https://web.archive.org/web/20150417230631/ http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory (accessed 17 April 2015).

Fiore AT (2015) The limits of data science at scale. 2015 OCLC/Kilgour Lecture, Chapel Hill, NC, 16 March. Available at: https://web.archive.org/web/20150417232718/ https://vimeo.com/123422862 (accessed 17 April 2015).

Hoopes H (2004) Reply to discussion thread "Unfound?". *FlickrCentral*. Available at: https://web.archive.org/web/20150417200653/ https://www.flickr.com/groups/central/discuss/9375/ (accessed 17 April 2015).

Kent W (1978) *Data and Reality: Basic Assumptions in Data Processing Reconsidered*. Amsterdam: North-Holland.

Kirschenbaum M (2008) *Mechanisms: New Media and the Forensic Imagination*. Cambridge, MA: MIT Press.

Loukides M (2010) What is data science? *O'Reilly Radar*, 2 June. Available at: https://web.archive.org/web/20150417230021/http://radar.oreilly.com/2010/06/what-is-data-science.html (accessed 17 April 2015).

Mink L (1966) The autonomy of historical understanding. *History and Theory* 5(1): 24–47.

Oakeshott M (2002) *Experience and its Modes*. Cambridge: Cambridge University Press.

Qi W (2005) Yahoo, UC Berkeley join to create research lab. *The Daily Californian*, 18 July. Available at: https://web.archive.org/web/20150417195644/ http://archive.dailycal.org/article.php?id=19014 (accessed 17 April 2015).

Quéru J-B (2011) Dizzying but invisible depth. 15 October. Available at: https://web.archive.org/web/20150417195936/https://plus.google.com/+JeanBaptisteQueru/posts/dfydM2Cnepe (accessed 17 April 2015).

Shamma DA, Shaw R, Shafton PL, et al. (2007) Watch what I watch: using community activity to understand content. In: *Proceedings of the international workshop on multimedia information retrieval – MIR '07*, 28–29 September, pp. 275–284. DOI: 10.1145/1290082.1290120.

Taylor RN, Medvidović N and Dashofy EM (2010) *Software Architecture: Foundations, Theory, and Practice*. Hoboken: Wiley.

Twitter (2011) One million registered Twitter apps. Available at: https://web.archive.org/web/20150420154319/https://blog.twitter.com/2011/one-million-registered-twitter-apps (accessed 20 April 2015).

Wikipedia (2014) Frictionless sharing. Available at: https://en.wikipedia.org/w/index.php?title=Frictionless_sharing&oldid=605680811 (accessed 17 April 2015).

Williamson B (2014) The death of the theorist and the emergence of data and algorithms in digital social research. Impact of Social Sciences blog, The London School of Economics and Political Science, 10 February. Available at: https://web.archive.org/web/20150322073355/http://blogs.lse.ac.uk/impactofsocialsciences/2014/02/10/the-death-of-the-theorist-in-digital-social-research/ (accessed 22 March 2015).

This article is part of a special theme on *Colloquium: Assumptions of Sociality*. To see a full list of all articles in this special theme, please click here: http://bds.sagepub.com/content/colloquium-assumptions-sociality.