Title

Making big data, in theory

Permalink

https://escholarship.org/uc/item/0ds9b7pr

Journal

First Monday, 18(10)

ISSN

13960466

Author

Boellstorff, Tom

Publication Date

2013-10-07

DOI

10.5210/fm.v18i10.4869

Peer reviewed

First Monday, Volume 18, Number 10 - 7 October 2013

PEER-REVIEWED JOURNAL ON THE INTERNET

Making big data, in theory

by Tom Boellstorff

Abstract

In this paper, I explore four conceptual interventions that can contribute to the "big theory" sorely needed in regard to big data. This includes temporality and the possibilities of "dated theory," the implicit histories of the *meta-* prefix shaping notions of metadata, "the dialectic of surveillance and recognition," and questions of interpretation understood in terms of "rotted data" and "thick data." In developing these concepts, I seek to expand frameworks for addressing issues of time, context, and power. It is vital that a vibrant theoretical discussion shape emerging regimes of "big data," as these regimes are poised to play an important role regarding the mutual constitution of technology and society.

Contents

1. Introduction

> [P]ioneer Victorian histories of the British press, written in 1850–87... genuflected before the altar of technology by capitalizing the first letters of Newspaper Press... Wiser and more skeptical, [later writers] stopped capitalizing the "Newspaper Press." Perhaps we should do the same in relation to the "Internet." [1]

We live in a time when big data will transform society. Or so the hype goes.

Like any myth, the current hullabaloo regarding big data is overblown but contains grains of truth. There are the relatively easy responses: there is no unitary phenomenon "big data." There is no singular form of "society." And so on. Yet the impact of big data is real and worthy of sustained attention.

You may have noted that I am not capitalizing "big data." This is my first theoretical intervention in what will be a theoretical essay deeply informed by history. As James Curran notes above, there is a background to capitalization practices with regard to technology—and value in treating big data as a common noun, vulnerable to reconfiguration. My goal is to explore four possibilities for reconfiguration that can contribute to the "big theory" sorely needed in regard to big data. I ground these explorations in my primary disciplinary background of anthropology but draw from other fields: etymology as much as ethnography, philosophy as much as science studies.

This essay originates in conceptual interests that have haunted me throughout my work on digital culture and also my earlier research on sexuality in Indonesia (e.g., Boellstorff, 2005; 2007; 2008; 2012). There is also a proximate motivation: the Edward Snowden affair of 2013. His disclosure that the National Security Agency was monitoring domestic "metadata" launched a vociferous debate about big data, surveillance, and the public good—a debate that at the time I write has not died down (as I composed the first draft of this essay Snowden was still trapped in the Moscow airport). At various points I will draw on aspects of the Snowden affair and the broader discussions linked to it.

This essay covers a great deal of conceptual ground: I am less interested in offering closure than opening conversations. Concepts I will develop include "dated theory," "metastasizing data," "the dialectic of surveillance and recognition," and "rotted data." The title "Making big data, in theory" flags the themes of "making" and "theory" that appear throughout. One analysis of over 27,000 social science articles published between 2000 and 2009 found that "only about 30% of Internet studies cite one or more theoretical references, suggesting that Internet studies in the past decade were modestly theorized." [2] There is a great need for theorization precisely when emerging configurations of data might seem to make concepts superfluous—to underscore that there is no Archimedean point of pure data outside conceptual worlds. Data always has theoretical enframings that are its condition of making: those who actually work with big data know that although it "can be illuminating, it is not unproblematic. Any dataset offers a limited representation of the world" (Loudon et al., 2013).

The stakes are high. Algorithmic living is displacing artificial intelligence as the modality by which computing is seen to shape society: a paradigm of semantics, of understanding, is becoming a paradigm of pragmatics, of search. [3] Contemporary computational language translation, for instance, does not work by trying to get a computer to intelligently understand language: systems like Google Translate work by matching texts from a vast corpus, without the computer ever "knowing" what is said. Historically this lack of knowing was a problem to debate, as in "Chinese room" thought experiments questioning if a person locked in a room and given English instructions for using Chinese characters could be said to understand Chinese. [4] But while the possibility of artificial intelligence is certainly still debated, what is most striking is the degree to which such questions have simply been set aside. Moreover, emergent paradigms of algorithmic living pose the possibility that pragmatics and semantics might converge, that "use" will be the "meaning" that matters in what is claimed to be a new age of big data.

Big data—this vast and changing corpus—is thus at the heart of the idea that a shift to algorithmic living is nigh, despite the concept's relative novelty. The term "big data" likely dates informally to the 1990s, first appearing in an academic publication in 2003 (Lohr, 2013) but gaining wider legitimacy only around 2008 (Lohr, 2012; see Bryant et al., 2008). Nonetheless, in less than a decade big data has risen to a dominant position in many quarters of the technology sector, academia, and beyond. Massive amounts of grant money, private- and public-sector labor, and capital—corporate, state, and military—now flow into the generation, capture, and analysis of big data. The humanities and social sciences face threats and opportunities, not least because "ethnography" is often presented as the Other to big data, raising fascinating issues regarding their recombination (Manovich, 2011). It is vital that a vibrant theoretical discussion shape these emerging paradigms, for "big data" is poised to play an important role regarding the mutual constitution of technology and society in the twenty-first century.

2. Dated theory

Spatial metaphors of mobility and omnipresence are salient in discussions of big data, but big data is also a profoundly temporal phenomenon—caught up in debates over time, technology, and theory. Consider how I speak of the Snowden affair, even though by the time this essay is published those events will have changed. What does it mean to say that by the time you read this article it will be dated? Have I thereby limited its utility: will it be interesting in 2014, in 2024? Could there be any relevance to being behind?

I want to pause on this issue of the timely, an issue that seems ever to nip at the heels of discussions regarding big data and the digital more generally, threatening analytical purchase. I want to consider the value of arguments whose time is out of joint, that are untimely (Grosz, 2004)—that are *dated*. Developing what I will term "dated theory" is important for addressing relationships between big data, representation, surveillance, and recognition.

My notion of "dated theory" is informed by the history of the data concept. In tracing its rise in the seventeenth and eighteenth centuries, Daniel Rosenberg noted it "is the plural of the Latin word *datum*, which itself is the neuter past participle of the verb *dare*, to give. A 'datum' in English, then, is something given in an argument, something taken for granted." [5] But my linking "data" to "dated" does not invoke a false cognate: "date" also comes from *datum*, a shared etymology that is a literal effect of correspondence:

[6]

> In classical Latin, the date of a letter was expressed by a phrase such as *data xiiii K. Maias de Tarentino* "(letter) sent from Tarentum on 18th April"... Hence *data*, the first word of the formula, came to be used as a term for the time and place stated therein. (*OED*, 2013a)

"Data," then, has long been connected to the "time and place" of a letter's sending—what is now termed "metadata," a central topic of the Snowden affair. By "dated theory," I mean to foreground how data is always a temporal formation; "data" always has a "date" that shapes claims to truth made on its behalf.

The notion of dated theory is useful due to fears that "in the time it takes to formulate, fund, conduct, revise, and publish a significant research question, we all are left to worry that changes in the media environment will render our work obsolete." [7] While I share Karpf's skepticism regarding this view, one thing that clearly contributes to such unease is the idea that analytical value hinges on anticipation. This is shaped by positivist traditions that equate scientific value with predictive laws, as well as the hype-filled rhetoric of Silicon Valley showmanship, where "trending" is at a premium. This rhetoric also shapes scholarship: "the dominant tense... is that of the proximate future. That is, motivations and frames... portray a *proximate future*, one just around the corner." [8] This recalls prolepsis, the literary device of the "flashforward" that appears in a phrase like "you're a dead man" when spoken by a villain to the person about to be murdered (Boellstorff, 2011).

To respond to the proleptic temporal imaginaries that often accompany big data, it can be helpful to recall that knowledge production is never separate from the knowledge producer. A discussion of dated theory is a discussion of dated theorists. Consider the well-documented temporal politics of anthropology. Originating largely in the colonial encounter, anthropology was dominated by an "denial of coevalness... a persistent and systematic tendency to place the referent(s) of anthropology in a Time other than the present of the producer of anthropological discourse." [9] Within anthropology, this tendency has been deeply critiqued: the notion of "salvage anthropology" dates to 1970 (Gruber, 1970), and there have been many calls to transcend the "savage slot" to which anthropology traditionally consigned its object of study, to "find better anchor for an anthropology of the present." [10] These reflexive debates have shaped a critique of the dominant temporal imaginary of its practitioners. This is the trope, perhaps best known through the mythical image of the pioneering Bronisław Malinowski, where a lone anthropologist arrives at a tropical island and "discovers" a "remote" tribe whose members have ostensibly lived unchanged for centuries (an image I have played on and destabilized). [11] The famous "Far Side" cartoon by Gary Larson where a group of panicked "natives" quickly hide their televisions and other technology as a pair of anthropologists arrive in a canoe is but one example of this trope's pervasiveness.

This temporal imaginary still motivates some ethnographers. But the critique, however incomplete, has had consequences: a viable ethnographic career can now involve studying immigration, scientists, or queer Indonesians. Most importantly, there is a widespread understanding that addressing researcher subjectivity makes research more scientific, robust, and ethical.

In contrast, there has been little discussion of the temporal imaginary of big data researchers. How does time shape their subjectivities and the making of big data? It may be that the temporal imaginary is not one of digging into the past but looking into a future more than proximate—a *distal future* that can be predicted and even proleptically anticipated.

And the paradigmatic figure of this researcher? One candidate might be Hari Seldon, the protagonist of Isaac Asimov's classic 1951 science fiction novel *Foundation*. Seldon, the greatest of all "psychohistorians," has been put on trial by the Galactic Empire, twenty thousand years from now. His crime is one of anticipation: to threaten panic by using what we can term big data to predict the Empire's fall "on the basis of the mathematics of psychohistory." [12] Seldon's defense is that he seeks to create an "Encyclopedia Galactica" to shorten the subsequent period of anarchy:

> By saving the knowledge of the race. The sum of human knowing is beyond any one man; any thousand men. With the destruction of our social fabric, science will be broken into a million pieces... *But*, if we now prepare a giant summary of *all* knowledge, it will never be lost. [13]

Asimov's Wikipedia-before-its-time, his vision of what we can anachronistically but accurately term "big data as social engineering," resonates with a contemporary context where "the deployment of algorithmic calculations... signals an important move—from the effort to predict future trends on the basis of fixed statistical data to a means of pre-empting the future." [14] Now, I am not claiming that all those who work with big data have read Asimov or aspire to be Hari Seldon, any more than all ethnographers wish to discover an "untouched" native tribe. My point is that we need to "date" not just big data but the temporal imaginaries shaping those who use it.

For instance, the language of ages, eras, and epochs is near-ubiquitous in scholarship on the digital. There has been disagreement with such "framings that imagine the digital in terms of epochal shifts" [15]—for instance, because "the term Digital Age stratifies media hierarchies for those who are out of power." [16] These concerns are valid (including the concern that a language of "ages" can lead to ignoring history), but it is also valuable to contend that "the era of Big Data has begun." [17] Periodization is not just Silicon Valley hype: for social theorists no less than geologists, it can be analytically useful and empirically accurate. We do not want to throw out the baby with the temporal bathwater, for many scholarly works on the digital employ periodizations in an insightful manner. Of course, phrases like "the digital age" are now commonly used even by corporate executives (e.g., Schmidt and Cohen, 2013). But that periodization can be overused or misused does not mean it is useless. Periodizations are heuristics not immutables. We may argue over the precise end of the Mesozoic Era, just as we may argue over the precise beginning of the Industrial Revolution. But the contested character of periodizations gives them value; they thereby represent one important means for producing "dated theory."

3. Making metadata

The Snowden affair foregrounded an ostensible subset of big data, "metadata"—taken to mean things like the time a cellphone call was placed, its duration, and the caller's location, in contrast to the conversation itself. A fundamental power move by representatives of the United States government was to contend that NSA surveillance practices were benign because at issue was only metadata (see Mayer, 2013). Attempts to depoliticize metadata thus hinged on asserting a self-evident distinction between data and metadata, and critical responses often challenged this claim. [18] This shows how one of the most pressing issues in regard to theorizing big data involves historicizing metadata and mapping out its conceptual implications.

The notion of metadata precedes that of big data, having been coined in 1968 by the computer scientist Philip R. Bagley (1927–2011):

> To any data element... can be associated... certain data elements which represent data "about" the related element. We refer to such data as "metadata"... [19]

In using the *meta-* prefix at the dawn of the Internet age, Bagley drew on layers of history with embedded assumptions. In particular, the prefix contains an *unacknowledged tension between laterality and hierarchy*. This tension has never been resolved, with implications for power, selfhood, and community.

Prior to Bagley's coinage, in the study of language, information, and communication *meta-* dates to the 1950s, when the linguist Roman Jakobson began developing the notion of "metalanguage":

> A discrimination clearly anticipated by the Ancient Greek and Indic tradition... a need to distinguish between two levels of language, namely the "object language" speaking of items extraneous to language as such, and on the other hand a language in which we speak about the verbal code itself. The latter aspect of language is called "metalanguage," a loan-translation of the Polish term launched in the 1930s by Alfred Tarski. [20]

Metalanguage, then, is language about language—for instance, "I dreamed of a unicorn" is language, and "a unicorn is an imaginary horse with a single horn on its head" is metalanguage. A pragmatic rule lets a minister state "I pronounce you husband and wife." Given this, "the statement, 'In our society, when a proper religious or judiciary functionary so empowered sincerely utters to a man and woman, "I pronounce you husband and wife," the latter are married,' is a metapragmatic utterance describing the effective use of this formula." [21]

Jakobson's reference to Tarski reveals a second history of the *meta-* prefix, linked to "metamathematics"—a term developed by David Hilbert in the 1920s, but connecting to scholarship going back to thinkers like Russell, Frege, Gödel, and Whitehead (see Lee, 1997). It is this tradition that seems to have most directly influenced communication studies and thus big data. For instance, the authors of *Pragmatics of human communication* introduced their theory of "metacommunication" by noting that "When we no longer use communication to communicate but to communicate about communication... In *analogy to* *metamathematics* this is called metacommunication." [22]

Throughout the twentieth century use of the *meta-* prefix expanded, including metaknowledge (Watzlawick et al., 1967), metaindexicality (Lee, 1997), even metaculture (Urban, 2001). Yet the prefix has retained a fundamental instability. On one hand it is used hierarchically, so there can be "two levels of language," [23] metaknowledge can be knowledge "of a second order," [24] or metadata can be "transcendent and overarching." [25] This is the framework of a zero-degree referent (e.g., language, knowledge, or data), and then "meta" phenomena that lie above or below. On the other hand (sometimes by the same author) the *meta-* prefix is used laterally, so that metalanguage is "a language in which we speak about the verbal code itself," [26] or metaculture is "culture that is about culture." [27]

The origin of this double sense of the *meta-* prefix is rarely addressed: "meta" is not yet a well-dated theory. Consider (as Jakobson's reference to "Ancient Greek and Indic tradition" suggests) μετα's long history, in which the prefix originally had only a lateral meaning: "[I]n ancient Greek and Hellenistic Greek the prefix is [used] to express notions of sharing, action in common, pursuit, quest, and, above all, change (of place, order, condition, or nature)" (*OED*, 2013b). This original meaning lies at the heart of mass and digital "media": "Aristotle... speaks of two elements, namely air and water, as of two 'betweens.' In other words, he is the first to turn a common Greek preposition—*metaxú*, between—into a philosophical noun or concept: *tò metaxú*, the medium." [28] "Metamedia" is a redundant term given this original meaning of the *meta-* prefix. This laterality is now nearly forgotten, appearing in only a handful of terms. These include "metaphor" (literally, "to carry across"), "metathesis"—and, most interestingly, "metastasis," which by the late Renaissance had become a medical term regarding the transference of function between organs. Its converse was *redux*, the return of a diseased organ to its original state (Maurer, 1997). Metastasis effected a change of state, not a state above; there may be value in theorizing data that "metastasizes."

But if μετα originally referenced laterality—"before" and "after"—how did it come to take on the hierarchical, abstracting meaning contained in phrases like "going meta?" It was classification error, a mistake in book shelving. Andronicus of Rhodes, who in the first century B.C. created what became the definitive edition of Aristotle's works, "grouped a number of Aristotle's writings into a single volume and placed it after (*meta-*) the physical treatises (*physika*). Thus the term *metaphysika*, which became the name of Aristotle's work, did not mean what it subsequently came to mean—a subfield of philosophy." [29] Due to this legacy of "the 12 books that, unfortunately, go under the title of 'Metaphysics'," [30] the term became "used as a name for the branch of study treated in these books, and hence came to be misinterpreted as meaning 'the science of things transcending what is physical or natural'... notwithstanding the fact that μετα does not admit of any such sense as 'beyond' or 'transcending'" (*OED*, 2013c).

This history of misinterpretation is hardly obscure; it can be found on the "meta" entry on Wikipedia. [31] In recounting it I mean to imply neither etymological determinism nor Eurocentrism. As many scholars of Aristotle note, the constructedness of the term "metaphysics," which "does not occur in any work of Aristotle known to us," [32] does not mean that there can be no concept of metaphysics at all. At issue is rather the deafening silence regarding the term's contingency and the tensions built into the *meta-* prefix itself. Specifically, *the notion of "metadata" is derived from this mistranslation of "metaphysics"* away from laterality and toward hierarchy. Particularly from the seventeenth century, "metaphysics" took on a strongly Christian cast: it may be that the only novel use of *meta-* as a prefix prior to the nineteenth century comes from John Donne's 1615 conception of a "meta-theology" above the personal Gods of the Reformation (Aronson, 2002; *OED*, 2013c). That metadata might be shaped by these notions of metaphysics is worthy of attention in a domain where some people describe themselves as "technology evangelists" and use "avatars," sometimes on Apple computers whose logo recalls the bite from the Tree at Knowledge in the Garden of Eden (Halberstam, 1991).

The pivotal point is that the threshold that causes something to move from a zero-degree category to its "meta" analogue is not a priori. It is an act of classification, and as such "valorizes some point of view and silences another." [33] I seek to challenge assumptions of a neat division between data and metadata not just because metadata can be more intrusive than data, but because the very division of the informational world into two domains—the zero-degree and the meta—establishes systems of implicit control. Indeed, once a zero-degree/meta distinction is accepted, it becomes impossible to know when to stop. For instance, if we assume that to talk about language one must use a "metalanguage," then we need "a metametalanguage if we want to speak about this metalanguage, and so forth in theoretically infinite regress." [34] So data about metadata could be termed "metametadata," but the fact that a "theoretically infinite regress" is built into the *meta-* prefix indicates a flaw with the concept—namely, obviating the contestable social practices by which data is constituted as an object of knowledge.

These issues around metadata are not limited to the online. During the Snowden affair many were surprised to learn of a longstanding system of monitoring physical metadata, "the Mail Isolation Control and Tracking program, in which Postal Service computers photograph the exterior of every piece of paper mail that is processed in the United States." (Nixon, 2013) Here, the division between data and metadata might seem beyond debate (as well as recalling the historical link between writing "data" on an envelope and being "dated"). After all, what could be easier to distinguish than the address written on an envelope from the letter inside it?

However, I want to question this and all divisions between the zero-degree and the meta. What if the division was framed not in terms of letters in envelopes with their interiors and exteriors, but the two sides of a postcard? Postcards were controversial during their emergence in the late nineteenth century because their "contents" could be read by anyone (Cure, 2013); they trouble the distinction between form and content (Boellstorff, 2013). How might analogizing the postcard provide one way to rethink this binarism? If I could take a postcard and bend it into a Mobius strip I would be even happier: a vision of form and content as intertwined at the most fundamental level, such that acts of "meta" assignation are clearly the cultural and political acts they are, rather than pregiven characteristics.

It is noteworthy that computer scientists and engineers have long recognized this fact. For example, file systems generally consist of at least two "layers," one storing portions of the "data," the second storing "metadata" (for instance, who owns the file or when it was created). But this is a conceptual distinction, and a file system operation (for example, writing a file) is viewed as a single operation that puts some information in the "data" portion and some information in the "metadata" portion of the file system. The "metadata" are not treated specially because of any "meta" characteristic, but because of their use as data (for instance, different rates of access or different storage properties). [35] These practical insights regarding the constructedness of the zero-degree/meta distinction are echoed in scholarly work on digital technology that highlights the often-hidden "labor of generating metadata." [36] For instance, one senior librarian at the British Library noted:

> [T]he perils of outsourcing some of the labor of generating metadata to India, where even the best English-speaking operators working with the digital copies of newspapers may not recognize common English place names. Thus, even the most supposedly neutral activity of metadata creation... shows that culturally situated knowledge can still be important for meaningful mark-up in the digital age. [37]

These examples underscore the practical and political consequences of theory. It is not just that terming things "data" is an act of classification; terming things "metadata" is no less an act of classification and no less caught up in processes of power and control. It is founded in a long and convoluted history of tensions between hierarchical and lateral thinking that shape everything from file systems to societies. [38] This history undermines any attempt to treat the distinction between zero-degree data and metadata as self-evident.

I have sometimes jokingly said that defining "meta" is like defining another four-letter word, "porn": you know it when you see it. But this parallel is surprisingly accurate, because what is widely understood with regard to the obscene—but almost never acknowledged with regard to the *meta-* prefix—is that both are (like all social phenomena) defined by communities of practice. What counts as obscenity depends on the norms of a particular time and place; what counts as "meta" is similarly contextual. In linguistic terms, it is misleading to seek a parallel between the zero-degree and the meta in terms of a structural distinction like that between speech and grammar. It is more effective to think in terms of "codeswitching" between English and Spanish, the movement between formal and informal registers of a language, or even "tagging"—understood as emergent acts of labeling that become generally accepted "hashtag" categories over time.

It is therefore empirically accurate and politically imperative "to treat meta-languages not merely as systems of analysis, but as practices of communication" (K. Jensen, this issue). Consider how search terms, a prototypical example of "metadata," can become "data" through social practice. This happens when such searches are (often inaccurately) used to track possible influenza outbreaks due to increased searches for phrases like "flu treatment" (Crawford, 2013). Another example of this occurred when LGBT activists responded to the heterosexist stance of former Pennsylvania congressman Rick Santorum by using his name as part of a search string for sexual fluids, temporarily pushing a "spreadingsantorum.com" Web site created by activists to first place in Google's results for the term "santorum" (Gillespie, 2012). In instances like these, phenomena typically classed as metacommunication act as forms of communication.

This is perhaps the most important theoretical issue with regard to the making of metadata, one with implications for social theory more generally. The fact that the act of "meta" assignation is culturally contextual is relevant to any use of the prefix, from metaphysics to metapragmatics, from metacommunication to metamedia, from metaknowledge to metaculture. Indeed, Gregory Bateson—one of the classic anthropologists most cited by scholars of communication and digital culture—theorized that play and fantasy were a kind of metacommunication crucial to the evolution of communication itself: "a very important state in this evolution occurs when the organism... becomes able to recognize... that the individual's and its own signals are only signals." [39] A better understanding of the making of meta will therefore be central to understanding emerging forms of big data and their social implications.

4. The dialectic of surveillance and recognition

The Snowden affair magnified existing debates regarding big data, surveillance, and state power. It was, after all, to challenge this power that Snowden took the risks he did, emphasizing "the greatest fear that I have regarding the outcome for America of these disclosures is that nothing will change." [40] His fear was well-founded; public surveys revealed just how many Americans were ambivalent, indifferent, or even enthusiastic about this state surveillance (Ohlheiser, 2013). Debates over the making of big data clearly draw from a contested cultural logic of monitoring, privacy, and disclosure. My tentative name for this logic is "the dialectic of surveillance and recognition."

Snowden's revelations of big-data-enabled state surveillance left many grasping for precedents, analogues, metaphors. One of the most common was "Orwellian;" it had already been noted that big data "carries a darker connotation, as a linguistic cousin to the likes of Big Brother" (Lohr, 2012). Yet many found this a limited trope (e.g., M. Jensen, 2013), not least because "what Orwell didn't see was where technology... gave enormous power to groups of people to build things as complicated and wonderful as... Wikipedia." [41] In other words, the Orwellian metaphor does not capture how the "big data" concept includes both relatively non-intentional data (like GPS data generated by a moving cellphone) and relatively intentional data (like a Facebook posting).

Perhaps this is why Snowden invoked not George Orwell but Michel Foucault: "if a surveillance program produces information of value, it legitimizes it... . In one step, we've managed to justify the operation of the Panopticon." Snowden here referenced Foucault's discussion in *Discipline and Punish* of the Panopticon, proposed by the utilitarian philosopher Jeremy Bentham as part of prison reform. [42] A prison would be composed of cells facing a central tower, the Panopticon, allowing a single guard to monitor the prison. Furthermore, the Panopticon would be designed so that prisoners could never tell if someone was in the tower. They would internalize the Panopticon's gaze and monitor their own behavior: "Hence the major effect of the Panopticon... to arrange things that the surveillance is permanent in its effects, even if it is discontinuous in its action." [43] The Panopticon provided a visual metaphor that seems prescient when an NSA surveillance program can be code-named "prism": "in order to be exercised, this power had to be given the instrument of permanent, exhaustive, omnipresent surveillance... thousands of eyes posted everywhere, mobile attentions ever on the alert, a long, hierarchized network." [44]

However, from a Foucauldian perspective the master metaphor for making big data should not be the Panopticon, but the confession. While published only one year after *Discipline and Punish* and sharing many themes with that earlier work, in *The History of Sexuality, Volume 1* Foucault turned even greater attention to how power, knowledge, and selfhood come together in specific historical contexts. In a chapter titled "The Incitement to Discourse," he traced the emergence of "a political, economic, and technical incitement to talk about sex... in the form of analysis, stocktaking, classification, and specification, of quantitative or causal studies." [45] Sex was *made into data*, with two crucial implications. First, that data was part of a state project: "Sex was not something one simply judged; it was a thing one administered. It was in the nature of a public potential; it called for management procedures." [46] Second, this data was produced through a "confessional" discourse drawing from Christianity and the psychoanalytic encounter between therapist and patient:

> [I]t is in the confession that truth and sex are joined, through the obligatory and exhaustive expression of an individual secret... it is also a ritual that unfolds within a power relationship, for one does not confess without the presence (or virtual presence) of a partner who is not simply the interlocutor but the authority who requires the confession. [47]

The confession is a modern mode of making data, an incitement to discourse we might now term an *incitement to disclose*. It is profoundly dialogical: one confesses to a powerful Other. This can be technologically mediated: as Foucault noted, it can take place in the "virtual presence" of authority. This is the only occurrence of "virtual" in *The History of Sexuality, Volume 1* and its inclusion is significant. To further address how this incitement to disclose plays out in contemporary contexts, it will be helpful to consider Charles Taylor's discussion of "the politics of recognition" that he saw as central to modernity:

> The thesis is that our identity is partly shaped by recognition or its absence, often by the *mis*recognition of others, and so a

> person or group of people can suffer real damage, real distortion, if the people or society around them mirror back to them a confining or demeaning or contemptible picture of themselves. [48]

In speaking of the "dialectic of surveillance and recognition," I seek to link the notion of confessional discourse to the politics of recognition. [49] A thesis for further research is that the rise of big data is accompanied by a discourse that links surveillance to recognition, that *frames surveillance as a form of belonging*. No discourse is singular and there are certainly reverse discourses, counterdiscourses, and alternate discourses. At issue is not that the kind of state surveillance highlighted by the Snowden affair is uncontested (because it obviously is contested), but understanding how so many find surveillance acceptable and even pleasurable: "play is crucial in understanding the new social data." [50] One of the most important political lessons of Foucault's work was the insight that resistance often emerges from within a discourse in a complex fashion poorly represented by purist notions of oppositionality. With regard to queer politics, Foucault noted that the mid-nineteenth century pathological understanding of homosexuality:

> [M]ade possible the formation of a "reverse" discourse: homosexuality began to speak in its own behalf, to demand that its legitimacy or "naturality" be acknowledged, often in the same vocabulary, using the same categories by which it was medically disqualified. There is not, on the one side, a discourse of power, and opposite it, another discourse that runs counter to it. [51]

It is not yet clear what kind of reverse discourses will emerge with regard to big data and its dialectic of surveillance and recognition. However, one clue can be seen in the fact that many responses to the making of big data are implicitly calls not for its abolition, but its extension. In a critique of big data, Kate Crawford noted how "data are assumed to accurately reflect the social world, but there are significant gaps, with little or no signal coming from particular communities," so that "with every big data set, we need to ask which people are excluded. Which places are less visible? What happens if you live in the shadow of big datasets?" (Crawford, 2013). Many other scholars echo this concern that we "be aware of... doubts over data representativeness when generalizing from search engine users to an entire population." [52] I share this concern that more people be included. The point is that in an almost homeopathic fashion, the remedy lies within the conceptual horizon of the illness it is to mitigate—within the dialectic of surveillance and recognition.

5. Rotted data, thick data

Snowden justified his revelations of NSA surveillance by arguing such data collection always takes place in an interpretive frame—even one applied after the fact, so that a government could "go back in time and scrutinize every decision you've ever made." [53] He thus linked claims about data and temporality to "scrutiny"—to the culturally contextual work of interpretation. This echoes an emerging literature challenging the notion of "raw" data. In their introduction to *"Raw data" is an oxymoron*, Gitelman and Jackson noted the edited volume's title references a statement by Geoffrey Bowker. [54] In full, that statement is "raw data is both an oxymoron and a bad idea; to the contrary, data should be cooked with care." [55] This of course references Claude Lévi-Strauss's (1969) *The raw and the cooked*:

> The concept "raw data" can be aligned with Lévi-Strauss's use of the term "raw"... [to describe] a vast mythological set... His argument was that a series of binaries characterized this set, many of which were a variant of what we would call the nature/society divide. The natural was the raw (honey) and the social was the cooked (ashes). [56]

Strikingly, this binarism of raw/cooked is both etic (an outsider framework) and emic (used in everyday practice). For instance, in their study of the Swedish intelligence community, Räsänen and Nyce found that "intelligence practitioners use the term 'raw data' as a common sense category," and sought to "challenge these practical understandings of central categories like the raw and the cooked." [57] These categories are incredibly important with regard to big data. One reason is the implication that the "bigness" of data means it must be collected prior to interpretation—"raw." This is revealed by metaphors like data "scraping" that suggest scraping flesh from bone, removing something taken as a self-evidently surface phenomenon. Another implication is that in a brave new world of big data, the interpretation of that data, its "cooking," will increasingly be performed by computers themselves.

Yet as the authors above (and others) have noted, this is another instance where classic anthropological work provides insight. Lévi-Strauss opened *The raw and the cooked* by emphasizing movement between emic and etic: "the aim of this book is to show how empirical categories... which can only be accurately defined... by adopting the standpoint of a particular culture—can nonetheless be used as conceptual tools." [58] In this book Lévi-Strauss often treated the raw and cooked in a dichotomous fashion. However, in "The culinary triangle," published one year after *The raw and the cooked*, he placed these categories in a triadic relationship with the "rotted." [59] In this full theorization, "raw" and "cooked" are not set in a binary where raw = nature and cooked = culture. Instead, they are framed as elements of a "culinary triangle" shaped by the *intersection* of the binarisms of nature/culture and normal/transformed (Figure 1). Here, "the raw constitutes the unmarked pole, while the other two poles are strongly marked, but in different directions: indeed, the cooked is a cultural transformation of the raw, whereas the rotted is a natural transformation." [60]

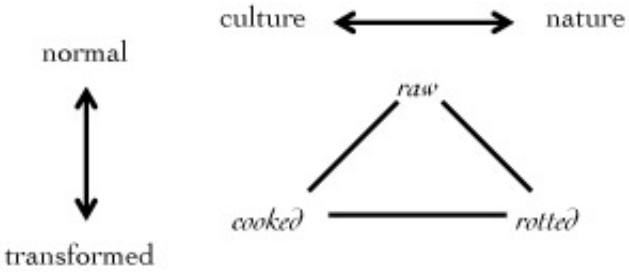

Figure 1: Lévi-Strauss's culinary triangle (image by author).

This trichotomy holds stimulating potential for theorizing the making of big data. Like "raw" and "cooked," the category of the "rotted" is emic as well as etic, as seen in the notion of "bit rot." This can refer to the materiality of data—the way that decaying computer tape and damaged hard drives cause data loss. However, it has long been observed that what is often at stake in "bit rot" is not the loss of data as much as the ability to interpret: "Long before the disk wears out or succumbs to bit rot, the machine that reads the disk has become a museum piece. So the immediate challenge is not preserving the information but preserving the means to get at it." [61]

In the context of the raw and cooked, the "rotted" allows for transformations outside typical constructions of the human agent as cook—the unplanned, unexpected, and accidental. Bit rot, for instance, emerges from the assemblage of storage and processing technologies as they move through time. But "rotting" moves between nature and society as well as between intentional and unintentional. Rotting can be "either spontaneous or controlled;" [62] in the latter case it is often termed "fermentation" or "distillation" and produces everything from bread and cheese to beer and wine.

Given longstanding notions of "distilling" meaning from big data, [63] the notion of rotted data might thereby provide a conceptual lens to consider imbrications of the intentional and random in the generation, interpretation, and application of big data. A "culinary data triangle" places raw and cooked in a logical rather than temporal relation. With three terms in a triangle rather than two terms in a row it is easier to avoid imputing a timeline. It is easier to avoid assuming that the raw comes before the cooked, and thus easier to challenge the claim to power embedded in the temporal argument that data comes before interpretation. A notion of "rotted data" reflects how data can be transformed in parahuman, complexly material, and temporally emergent ways that do not always follow a preordained, algorithmic "recipe."

It is possible to build on classic anthropological insights beyond those of Lévi-Strauss. In challenging the notion of "raw data," Gitelman and Jackson further echo Snowden, Bowker, and many others when emphasizing that "data need to be imagined as data to exist and function as such, and the imagination of data entails an interpretive base." [64] This reference to an "interpretive base" recalls Clifford Geertz's interventions into another debate over making data nearly a half-century earlier. In "Thick description: Toward an interpretive theory of culture," Geertz (an anthropologist often cited by those discussing big data [65]) responded to cognitive anthropologists like Ward Goodenough and Charles Frake, whose scholarship in turn contributed to the algorithmic frameworks central to contemporary big data. [66] Geertz first summarized the philosopher Gilbert Ryle's thought experiment regarding what Ryle termed "thick description”:

> Consider, he says, two boys rapidly contracting the eyelids of their right eyes. In one, this is an involuntary twitch; in the other, a conspiratorial signal to a friend. The two movements are, as movements, identical.... Yet the difference... is vast... the winker has not done two things, contracted his eyelids and winked, while the twitcher has done only one, contracted his eyelids. Contracting your eyelids on purpose when there exists a public code in which so doing counts as a conspiratorial signal *is* winking. [67]

While Geertz’s argument is more complex than I can present here (including a third boy parodying the wink of the second), of relevance is his emphasis that material gesture and semiotic framing are on the same ontological plane—a plane of public, shared meaning. This is what makes data “thick”: “what we call our data are really our own constructions of other people’s constructions of what they and their compatriots are up to.” [68] Geertz was discussing ethnographic data, but “thickness” is relevant to any form of data. Indeed, Geertz noted that an approach like that of Goodenough “is to claim that [culture] consists in the *brute pattern of behavioral events* we observe... from this view of what culture is follows a view, equally assured, of what describing it is—the writing out of systematic rules, *an ethnographic algorithm*.” [69] It was against this earlier iteration of data patterns and algorithmic living that Geertz emphasized the value of an interpretive base: “It is not against a body of uninterpreted data, radically thinned descriptions, that we must measure the cogency of our explications.” [70]

I am being purposely anachronistic: Geertz’s phrase “ethnographic algorithm” is not identical to contemporary uses of algorithms in big data analysis. But nor are they entirely separable, for a historical legacy links them. This includes critiques of the structuralism of Lévi-Strauss and others like Lacan. In contrast, what makes data “thick” is recognizing its irreducible contextuality: “what we inscribe (or try to) is not *raw* social discourse.” [71] For Geertz, “raw” data was already oxymoronic in the early 1970s: whether cooked or rotted, data emerges from regimes of interpretation: “Nor... have I been impressed with claims that structural linguistics, computer engineering, or some other advanced form of thought is going to enable us to understand men without knowing them.” [72]

In placing Lévi-Strauss and Geertz into conversation with contemporary debates regarding “raw data,” I mean neither to reduce them to each other nor claim that they provide a solution. Instead, like the notion of “metadata,” the notion of “raw data” demands an ongoing theoretical response. A valuable part of that response can be rethinking rhetorics of the unprecedented and accelerated that imply we have nothing to learn from the history of social theory—that big data could mean “the end of theory” (Anderson, 2008). Against the “thin” notion of raw data, we can think of data not just as cooked or rotted but “thick.” This highlights how big data is never ontologically prior to interpretation—“interpretation is at the center of data analysis” [73]—and interpretation takes place within horizons of culture that are embedded in contexts of power.

6. Conclusion: Making up big data

In this essay, I have sought to open up multiple lines of inquiry with regard to the making of big data. Building on a range of scholarly conversations, I have explored temporality and the possibilities of “dated theory,” the implicit histories shaping metadata, “the dialectic of surveillance and recognition,” and questions of interpretation understood in terms of “rotted data” and “thick data.” The goal has been to expand frameworks for addressing issues of time, context, and power. As an ethnographer I appreciate the value of focused and localized explanation, but we cannot cede more generalized theorizing to only some disciplines and methodological approaches. There is a need for what I term “platform-agnostic theory”—for theories that make claims about patterns and dynamics beyond the case study and the individual field site, even as those specificities shape the building of theory as well as its contextual modification. In taking big data down a notch (beginning by not capitalizing the term), we can understand it as a conceptual rubric, but also as a field site amenable to cultural critique and ethnographic interpretation.

It is fitting and disturbing that as I write the first draft of this conclusion Edward Snowden is still in limbo at the Moscow airport—in a place of transit now become a site of temporal non-belonging. Placed outside resolution by the state power he challenged, the very form of his plight is of a piece with the regime of making big data revealed by his disclosures. He and many others have worked to show that the perils and promises of big data hinge on recognizing that big data is not just “made” but “made up” in the sense that Ian Hacking has spoken of “making up people,” that classifications “affect the people classified, and... the affects on the people in turn change the classifications.” [74]

Like ethnographies, makings up of big data “are, thus, fictions; fictions, in the sense that they are ‘something made,’ ‘something fashioned’—the original meaning of *fictiō*—not that they are false, unfactual, or merely ‘as if’ thought experiments.” [75] They are more than “scrapes” of reality—they are part and parcel of that reality, immanent to the human condition. Big data is always already “big theory” as well, acknowledged or not. How these informational regimes shape societies into the emerging future depends in no small measure on our ability to understand and respond to the making up of big data itself.

About the author

Tom Boellstorff is Professor of Anthropology at the University of California, Irvine. His publications include *Coming of age in Second Life: An anthropologist explores the virtually human* (Princeton University Press, 2008).
E-mail: tboellst [at] uci [dot] edu

Acknowledgments

Many persons have helped me think through the topics I address in this essay. I particularly thank Ken Anderson, Geoffrey Bowker, Mic Bowman, Axel Bruns, Tarleton Gillespie, Klaus Bruhn Jensen, Elizabeth Losh, Annette Markham, Bill Maurer, and Nick Seaver. The Intel Science and Technology Center for Social Computing has served as a generative intellectual space for this work.

Notes

[1] Curran, 2012, pp. 34, 60.

[2] Peng et al., 2012, p. 655.

[3] For discussion of the notion of “algorithmic living,” see Mainwaring and Dourish, 2012.

[4] Searle, 1980; see, e.g., Collins, 1990.

[5] Rosenberg, 2013, p. 18.

[6] In some Indo-European languages like German, *datum* still means “date.” My thanks to Axel Bruns for reminding me of this point.

[7] Karpf, 2012, p. 640.

[8] Dourish and Bell, 2011, p. 23, emphasis added.

[9] Fabian, 1983, p. 31.

[10] Trouillot, 1991, p. 40.

[11] Boellstorff, 2008, pp. 3–4.

[12] Asimov, 1951, p. 26.

[13] Asimov, 1951, p. 28.

[14] Amoore, 2009, p. 53.

[15] Ruppert et al., 2013, p. 22.

[16] Ginsburg, 2008, p. 139.

[17] boyd and Crawford, 2012, p. 662.

[18] I am not arguing that government surveillance does not include both data and metadata, but that there have been attempts to treat the latter as safer to monitor, and that the distinction between the two is not a priori. For one example of a popular response to the attempted depoliticization of metadata, see the cartoon “Nothing to worry about, it’s just metadata” by Jeff Parker, first posted on 12 August 2013; see

http://www.truthdig.com/cartoon/item/nsa_its_just_metadata_20130812/, accessed 19 September 2013.

19. Bagley, 1968, p. 91.

20. Jakobson, 1980, p. 86.

21. Silverstein, 2001, p. 383.

22. Watzlawick et al., 1967, p. 40, emphasis added.

23. Jakobson, 1980, p. 86.

24. Watzlawick et al., 1967, p. 260.

25. Beer and Burrows, 2013, p. 51.

26. Jakobson, 1980, p. 86.

27. Urban, 2001, p. 3.

28. Kittler, 2009, p. 26.

29. Anagnostopoulos, 2009, p. 18.

30. Kittler, 2009, p. 24.

31. See http://en.wikipedia.org/wiki/Meta, accessed 27 July 2013.

32. Merlan, 1968, p. 175.

33. Bowker and Star, 1999, p. 5.

34. Watzlawick et al., 1967, p. 193.

35. I thank Mic Bowman for these insights regarding file systems.

36. Losh, 2009, p. 266.

37. *Ibid.*

38. Ironically, markings on the outside of ancient Mesopotamian clay vessels to indicate their contents may "represent the precise beginning of the technology of writing" (Schmandt-Besserat, 1980, p. 357). What is now seen as metadata may have come first.

39. Bateson, 1972, p. 151.

40. Edward Snowden, in Rodriguez, 2013.

41. Cory Doctorow, interviewed in Porzucki, 2013.

42. While not mentioning Foucault by name, the reference is probably intentional. As one observer noted, "make no mistake, Snowden has carefully read his Michael Foucault (he also stressed his revulsion facing 'the capabilities of this architecture of oppression')" (Escobar, 2013).

43. Foucault, 1977, p. 201.

44. Foucault, 1977, p. 214.

45. Foucault, 1978, pp. 23–24.

46. Foucault, 1978, p. 24.

47. Foucault, 1978, pp. 61–62.

48. Taylor, 1994, p. 25.

49. Other important discussions of recognition and belonging include Fraser, 2000; Povinelli, 2002.

50. Beer and Burrows, 2013, p. 51.

51. Foucault, 1978, p. 101.

[52] Trevisan, 2013, p. 2.

[53] Edward Snowden, in Rodriguez, 2013.

[54] Bowker, 2013, p. 1.

[55] Bowker, 2005, p. 184.

[56] Bowker, 2013, p. 168.

[57] Räsänen and Nyce, 2013, pp. 656, 660.

[58] Lévi-Strauss, 1969, p. 1.

[59] Lévi-Strauss addressed the category of the rotted only occasionally in *The raw and the cooked* (e.g., pp. 176, 254). *The raw and the cooked* was first published in 1964, and "The Culinary Triangle" in 1965 (see Lévi-Strauss, 1997).

[60] Lévi-Strauss, 1997, p. 29.

[61] Hayes, 1998, p. 410.

[62] Lévi-Strauss, 1997, p. 29.

[63] For instance, Frankel and Reid, 2008.

[64] Gitelman and Jackson, 2013, p. 3.

[65] For instance, Räsänen and Nyce, 2013, p. 659.

[66] These historical linkages are complex and still insufficiently researched. However, it is clear that many contemporary algorithmic methods for analyzing big data originate in mid-twentieth century work in cognition that was highly interdisciplinary, drawing on the research of psychologists like Amos Tversky (Nick Seaver, personal communication). This shaped a generation of cognitive anthropologists for whom it was possible "to regard all culture as information and to view any single culture as an 'information economy'" (Roberts, 1964, p. 438). Such a paradigm led, for instance, to forms of mathematical consensus analysis based on the premise that "Each systemic culture pattern may be thought of as having an associated semantic domain" (Romney et al., 1986, p. 315).

[67] Geertz, 1973, p. 6; emphasis in original.

[68] Geertz, 1973, p. 9.

[69] Geertz, 1973, p. 11, emphasis added.

[70] Geertz, 1973, p. 16.

[71] Geertz, 1973, p. 20, emphasis added.

[72] Geertz, 1973, p. 30.

[73] boyd and Crawford, 2012, p. 668.

[74] Hacking, 2006, p. 23.

[75] Geertz, 1973, p. 15.

References

Georgios Anagnostopoulos, 2009. "Aristotle's works and the development of his thought," In: Georgios Anagnostopoulos (editor). *A companion to Aristotle*. Malden, Mass.: Wiley-Blackwell. pp. 14–27.

Louise Amoore, 2009. "Algorithmic war: Everyday geographies of the War on Terror," *Antipode*, volume 41, number 1, pp. 49–69.
doi: http://dx.doi.org/10.1111/j.1467-8330.2008.00655.x, accessed 20 September 2013.

Chris Anderson, 2008. "The end of theory: The data deluge makes the scientific method obsolete," *Wired*, volume 16, number 7, at http://www.wired.com/science/discoveries/magazine/16-07/pb_theory, accessed 7 July 2013.

Jeff Aronson, 2002. "When I use a word: Meta-," *British Medical Journal*, volume 324, number 7344 (27 April), p. 1022.

Isaac Asimov, 1951. *Foundation*. New York: Gnome Press.

Philip R. Bagley, 1968. *Extension of programming language concepts*. Philadelphia: University City Science Center.

Gregory Bateson, 1972. "A theory of play and fantasy," In: Gregory Bateson. *Steps to an ecology of mind*. New York: Ballantine Books. pp. 150–166.

David Beer and Roger Burrows, 2013. "Popular culture, digital archives, and the new social life of data," *Theory, Culture & Society*, volume 30, number 4, pp. 47–71.
doi: http://dx.doi.org/10.1177/0263276413476542, accessed 20 September 2013.

Tom Boellstorff, 2013. "An afterword in three postcards," In: Dominic Power and Robin Tiegland (editors). *The immersive Internet: Reflections on the entangling of the virtual with society, politics and the economy*. Houndmills, Basingstoke, Hampshire: Palgrave Macmillan, pp. 247–252.

Tom Boellstorff, 2012. "Rethinking digital anthropology," In: Heather A. Horst and Daniel Miller (editors). *Digital anthropology*. London: Berg, pp. 39–60.

Tom Boellstorff, 2011. "But do not identify as gay: A proleptic genealogy of the MSM category," *Cultural Anthropology*, volume 26, number 2, pp. 287–312.
doi: http://dx.doi.org/10.1111/j.1548-1360.2011.01100.x, accessed 20 September 2013.

Tom Boellstorff, 2008. *Coming of age in Second Life: An anthropologist explores the virtually human*. Princeton, N.J.: Princeton University Press.

Tom Boellstorff, 2007. *A coincidence of desires: Anthropology, queer studies, Indonesia*. Durham, N.C.: Duke University Press.

Tom Boellstorff, 2005. *The gay archipelago: Sexuality and nation in Indonesia*. Princeton, N.J.: Princeton University Press.

Harry M. Collins, 1990. *Artificial experts: Social knowledge and intelligent machines*. Cambridge, Mass.: MIT Press.

Monica Cure, 2013. "Tweeting by mail: The postcard's stormy birth," *Los Angeles Times* (22 June), at http://articles.latimes.com/2013/jun/22/opinion/la-oe-cure-postcards-twitter-20130623, accessed 28 July 2013.

Geoffrey C. Bowker, 2013. "Data flakes: An afterword to 'Raw Data' is an oxymoron," In: Lisa Gitelman (editor). *"Raw data" is an oxymoron*. Cambridge, Mass.: MIT Press, pp. 167–171.

Geoffrey C. Bowker, 2005. *Memory practices in the sciences*. Cambridge: Mass.: MIT Press.

Geoffrey C. Bowker and Susan Leigh Star, 1999. *Sorting things out: Classification and its consequences*. Cambridge: Mass.: MIT Press.

danah boyd and Kate Crawford, 2012. "Critical questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon," *Information, Communication & Society*, volume 15, number 5, pp. 662–679.
doi: http://dx.doi.org/10.1080/1369118X.2012.678878, accessed 20 September 2013.

Randall E. Bryant, Randy H. Katz, and Edward D. Lazowska, 2008. "Big-Data computing: Creating revolutionary breakthroughs in commerce, science, and society," *Computing Research Consortium*, at http://www.cra.org/ccc/resources/ccc-led-white-papers/, accessed 20 September 2013.

Kate Crawford, 2013. "The hidden biases in big data," *HBR Blog Network* (1 April), at http://blogs.hbr.org/2013/04/the-hidden-biases-in-big-data/, accessed 5 July 2013.

James Curran, 2012. "Rethinking Internet history," In: James Curran, Natalie Fenton, and Des Freedman. *Misunderstanding the Internet*. London: Routledge, pp. 34–65.

Paul Dourish and Genevieve Bell, 2011. *Divining a digital future: Mess and mythology in ubiquitous* *computing*. Cambridge, Mass.: MIT Press.

Pepe Escobar, 2013. "Digital Blackwater rules," *Asia Times* (11 June), at http://www.atimes.com/atimes/World/WOR-03-110613.html, accessed 11 July 2013.

Johannes Fabian, 1983. *Time and the other: How anthropology makes its object*. New York: Columbia University Press.

Michel Foucault, 1978. *The history of sexuality*. Volume 1: *An introduction*. Translated by Robert Hurley. New York: Vintage Books.

Michel Foucault, 1977. *Discipline and punish: The birth of the prison*. Translated by Alan Sheridan. New York: Vintage Books.

Felice Frankel and Rosalind Reid, 2008. "Big data: Distilling meaning from data," *Nature*, volume 455, number 7209 (4 September), p. 30.
doi: http://dx.doi.org/10.1038/455030a, accessed 20 September 2013.

Nancy Fraser, 2000. "Rethinking recognition," *New Left Review*, volume 3, pp. 107–120, and at http://newleftreview.org/II/3/nancy-fraser-rethinking-recognition, accessed 20 September 2013.

Clifford Geertz, 1973. "Thick description: Toward an interpretive theory of culture," In: Clifford Geertz. *The interpretation of cultures: Selected essays*. New York: Basic Books. pp. 3–32.

Tarleton L. Gillespie, 2012. "The relevance of algorithms: The case of 'spreading santorum'," Internet Research 13.0, Association of Internet Researchers (AoIR), Salford, U.K. (October).

Faye Ginsburg, 2008. "Rethinking the digital age," In: David Hesmondhalgh and Jason Toynbee (editors). *The media and social theory*. New York: Routledge. pp. 127–144.

Lisa Gitelman and Virginia Jackson, 2013. "Introduction," In: Lisa Gitelman (editor). *"Raw data" is an oxymoron*. Cambridge, Mass.: MIT Press, pp. 1–14.

Elizabeth Grosz, 2004. *The nick of time: Politics, evolution, and the untimely*. Durham, N.C.: Duke University Press.

Jacob Gruber, 1970. "Ethnographic salvage and the shaping of anthropology," *American Anthropologist*, volume 72, number 6, pp. 1,289–1,299.
doi: http://dx.doi.org/10.1525/aa.1970.72.6.02a00040, accessed 20 September 2013.

Ian Hacking, 2006. "Making up people," *London Review of Books*, volume 28, number 16 (17 August), pp. 23–26, at http://www.lrb.co.uk/v28/n16/ian-hacking/making-up-people, accessed 20 September 2013.

Judith Halberstam, 1991. "Automating gender: Postmodern feminism in the age of the intelligent machine," *Feminist Studies*, volume 17, number 3, pp. 439–460.

Brian Hayes, 1998. "Bit rot," *American Scientist*, volume 86, number 5, pp. 410–415.
doi: http://dx.doi.org/10.1511/1998.5.410, accessed 20 September 2013.

Roman Jakobson, 1980. "Metalanguage as a linguistic problem," In: Roman Jakobson. *The framework of language*. Ann Arbor: Michigan Studies in the Humanities. pp. 81–92.

Morten Høi Jensen, 2013. "What everybody gets wrong about Orwell," *Salon* (19 June), at http://www.salon.com/2013/06/19/big_brother_is_the_wrong_metaphor_for_our_time/, accessed 12 July 2013.

David Karpf, 2012. "Social science research methods in Internet time," *Information, Communication & Society*, volume 15, number 5, pp. 639–661.
doi: http://dx.doi.org/10.1080/1369118X.2012.665468, accessed 20 September 2013.

Friedrich Kittler, 2009. "Towards an ontology of media," *Theory, Culture & Society*, volume 26, numbers 2–3, pp. 23–31.
doi: http://dx.doi.org/10.1177/0263276409103106, accessed 20 September 2013.

Benjamin Lee, 1997. *Talking heads: Language, metalanguage, and the semiotics of subjectivity*. Durham, N.C.: Duke University Press.

Claude Lévi-Strauss, 1997. "The culinary triangle," In: Carole Counihan and Penny Van Esterik (editors). *Food and culture: A reader*. London: Routledge, pp. 28–35.

Claude Lévi-Strauss, 1969. *The raw and the cooked*. Translated by John and Doreen Weightman. Chicago: University of Chicago Press.

Steve Lohr, 2013. "The origins of 'Big Data': An etymological detective story," *New York Times* (1 February), at http://bits.blogs.nytimes.com/2013/02/01/the-origins-of-big-data-an-etymological-detective-story/, accessed 5 July 2013.

Steve Lohr, 2012. "How big data became so big," *New York Times* (11 August), at http://www.nytimes.com/2012/08/12/business/how-big-data-became-so-big-unboxed.html/, accessed 5 July 2013.

Elizabeth Losh, 2009. *Virtualpolitik: An electronic history of government media-making in a time of war, scandal, disaster, miscommunication, and mistakes*. Cambridge, Mass.: MIT Press.

Melissa Loudon, Bill Maurer, Jessamy Norton-Ford, Martyn Fricker, Joshua Blumenstock, 2013. "Big data in ICT4D: What can we learn from prepaid mobile airtime transactions?" *Proceedings of ICTD 2013: Sixth International Conference on Information and Communication Technologies and Development* (7–10 December, Cape Town, South Africa).

Scott Mainwaring and Paul Dourish, 2012. "Intel Science and Technology Center for Social Computing: White paper," at http://socialcomputing.uci.edu/ISTC-Social-Whitepaper.pdf, accessed 19 August 2013.

Lev Manovich, 2011. "Trending: the promises and the challenges of big social data" (28 April), at http://www.manovich.net/DOCS/Manovich_trending_paper.pdf, accessed 7 July 2013.

Bill Maurer, 1997. "Creolization redux: The plural society thesis and offshore financial services in the British Caribbean," *New West Indian Guide*, volume 71, numbers 3–4, pp. 249–264.

Jane Mayer, 2013. "What's the matter with metadata?" *New Yorker* (6 June), at http://www.newyorker.com/online/blogs/newsdesk/2013/06/verizon-nsa-metadata-surveillance-problem.html, accessed 28 July 2013.

Philip Merlan, 1968. "On the terms 'metaphysics' and 'being-qua-being'," *Monist*, volume 52, number 2, pp. 174–194.
doi: http://dx.doi.org/10.5840/monist196852216, accessed 20 September 2013.

Ron Nixon, 2013. "U.S. Postal Service logging all mail for law enforcement," *New York Times* (3 July), at http://www.nytimes.com/2013/07/04/us/monitoring-of-snail-mail.html, accessed 28 July 2013.

Abby Ohlheiser, 2013. "Twitter's jaded reaction to the NSA's phone records collection program," *Atlantic Wire* (5 June), at http://www.theatlanticwire.com/technology/2013/06/twitters-jaded-reaction-nsas-phone-records-collection-program/65951/, accessed 12 July 2013.

Oxford English Dictionary (*OED*), 2013a. "date, n.2," at http://www.oed.com/, accessed 9 July 2013.

Oxford English Dictionary (*OED*), 2013b. "theory, n.1.," http://www.oed.com/, accessed 11 July 2013.

Oxford English Dictionary (*OED*), 2013c. "metaphysics, n. pl. [1989 edition]," http://www.oed.com/, accessed 15 July 2013.

Tai-Quan Peng, Lun Zhang, Zhi-Jin Zhong, and Jonathan J.H. Zhu, 2012. "Mapping the landscape of Internet studies: Text mining of social science journal articles 2000–2009," *New Media & Society*, volume 15, number 5, pp. 644–664.
doi: http://dx.doi.org/10.1177/0263276409103106, accessed 20 September 2013.

Nina Porzucki, 2013. "NSA leak: Did George Orwell get it right in *1984*?" *PRI's The World* (12 June), at http://www.theworld.org/2013/06/nsa-leak-orwell-1984/, accessed 12 July 2013.

Elizabeth A. Povinelli, 2002. *The cunning of recognition: Indigenous alterities and the making of Australian multiculturalism*. Durham, N.C.: Duke University Press.

Minna Räsänen and James M. Nyce, 2013. "The raw is cooked: Data in intelligence practice," *Science, Technology, & Human Values*, volume 38, number 5, pp. 655–677.
doi: http://dx.doi.org/10.1177/0162243913480049, accessed 20 September 2013.

J.M. Roberts, 1964. "The self-management of cultures," In: Ward Goodenough (editor). *Explorations in cultural anthropology: Essays in honor of George Peter Murdock*. New York: McGraw-Hill. pp. 433–454.

Gabriel Rodriguez, 2013. "Edward Snowden interview transcript full text: Read the *Guardian's* entire interview with the man who leaked PRISM,” *Policymic*, at http://www.policymic.com/articles/47355/edward-snowden-interview-transcript-full-text-read-the-guardian-s-entire-interview-with-the-man-who-leaked-prism, accessed 12 July 2013.

A. Kimball Romney, Susan C. Weller, William H. Batchelder, 1986. “Culture as consensus: A theory of culture and informant accuracy,” *American Anthropologist*, volume 88, number 2, pp. 313–338.
doi: http://dx.doi.org/10.1525/aa.1986.88.2.02a00020, accessed 20 September 2013.

Daniel Rosenberg, 2013. “Data before the fact,” In: Lisa Gitelman (editor). *“Raw data” is an oxymoron*. Cambridge, Mass.: MIT Press, pp. 15–40.

Evelyn Ruppert, John Law, and Mike Savage, 2013. “Reassembling social science methods: The challenge of digital devices,” *Theory, Culture & Society*, volume 30, number 4, pp. 22–46.
doi: http://dx.doi.org/10.1177/0263276413484941, accessed 20 September 2013.

Denise Schmandt-Besserat, 1980. “The envelopes that bear the first writing,” *Technology and Culture*, volume 21, number 3, pp. 357–385.

Eric Schmidt and Jared Cohen, 2013. *The new digital age: Reshaping the future of people, nations and business*. New York: Knopf.

John R. Searle, 1980. “Minds, brains, and programs,” *Behavioral and Brain Sciences*, volume 3, number 3, pp. 417–457.

Michael Silverstein, 2001. “The limits of awareness,” In: Alessandro Duranti (editor). *Linguistic anthropology: A reader*. Oxford: Blackwell, pp. 382–401.

Charles Taylor, 1994. “The politics of recognition,” In: Charles Taylor. *Multiculturalism: Examining the politics of recognition*. Edited and introduced by Amy Gutmann. Princeton, N.J.: Princeton University Press. pp. 25–73.

Filippo Trevisan, 2013. “Social engines and social science: A revolution in the making” (15 May), at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2265348, accessed 6 July 2013.

Michel-Rolph Trouillot, 1991. “Anthropology and the savage slot: The poetics and politics of otherness,” In: Richard G. Fox (editor). *Recapturing anthropology: Working in the present*. Santa Fe, N.M.: School of American Research Press. pp. 17–44.

Greg Urban, 2001. *Metaculture: How culture moves through the world*. Minneapolis: University of Minnesota Press.

Paul Watzlawick, Janet Helmick Beavin, and Don D. Jackson, 1967. *Pragmatics of human communication: A study of interactional patterns, pathologies, and paradoxes*. New York: Norton.

Editorial history

Received 16 September 2013; accepted 17 September 2013.

Making big data, in theory
by Tom Boellstorff.
First Monday, Volume 18, Number 10 - 7 October 2013
http://firstmonday.org/ojs/index.php/fm/rt/printerFriendly/4869/3750
doi:10.5210/fm.v18i10.