Surveillance & Society

Editorial | Big Data Surveillance: Introduction

Mark Andrejevic

University of Queensland, Australia.
m.andrejevic@uq.edu.au

Kelly Gates

University of California, San Diego, US.
kagates@ucsd.edu

One of the most highly publicized avatars of high-tech surveillance in the networked era is the drone, with its ever-expanding range and field of vision (and sensing more generally), combined with its ominous military capabilities. One of the less publicized facts about the deployment of surveillance and military drones is that in addition to weapons, cameras, and other sensors, they are equipped with a device called an "Air Handler" that can capture all available wireless data traffic in the area. As one of the rare news accounts about this device put it, when a drone goes out on a mission, "the NSA [National Security Agency] has put a device on it that is not actually under the control of the CIA or the military; it is just sucking up data for the NSA" (Goodman 2014). The drone then comes to represent a double-image of surveillance: both the familiar "legacy" version of targeted, purposeful spying and the emerging model of ubiquitous, opportunistic data capture. As one of the reporters interviewed about his research on the "Air Handler" put it, "the NSA just wants all the data. They want to suck it up on an industrial scale. So they're just piggybacking on these targeted operations in an effort to just suck up data throughout the world" (Goodman 2014). This description of the NSA's approach to data collection parallels the widely publicized comments of the CIA's Chief Technology Officer about contemporary strategies of surveillance: "The value of any piece of information is only known when you can connect it with something else that arrives at a future point in time…Since you can't connect dots you don't have, it drives us into a mode of, we fundamentally try to collect everything and hang on to it forever" (Sledge 2013). For drones, the signal-saturated sky is a sea of electromagnetically encoded data that can be captured, processed, refined, and perhaps put to use.

The collect-everything approach to monitoring and intelligence—embodied in the Air Handler, PRISM (the CIA's secret mass electronic surveillance and data mining initiative), and a litany of other programs both public and private—is our starting point for exploring the connection between surveillance and so-called "big data." If conventional definitions of surveillance emphasize its systematic and targeted character (the notion that there is a specific "object" of surveillance), both aspects undergo some significant modifications when the goal is, generally speaking, to capture as much data as possible about everything, all the time, and hold on to it forever. Moreover, the ambitious scope of such surveillance raises a host of important issues associated with the infrastructure for collecting and storing huge amounts of data as well as the techniques and technologies for putting it to use. Even if the underlying goal of capturing information for the pursuit of some form of advantage, leverage, or control remains constant (see, for example, the contributions of both Reigeluth and van Otterlo to this issue), conventional understandings of the operation of surveillance and its social consequences are being reconfigured by the "big data" paradigm.

Andrejevic, M. and K. Gates. 2014. Editorial. Big Data Surveillance: Introduction. *Surveillance & Society* 12(2): 185-196. http://www.surveillance-and-society.org | ISSN: 1477-7487

Some definitions will help specify the character of the changes associated with big data-driven forms of surveillance. For our purposes, the notion of "big data" refers to both the unprecedented size of contemporary databases and the emerging techniques for making sense of them. This understanding of big data will have consequences for our reconfigured definition of data-mining enhanced surveillance. What is significant about the big data moment is not simply that it has become possible to store quantities of data that are impossible for any individual to comprehend (The Library of Alexandria did that, as does the night sky, and the human brain), but the fact that this data can be put to use in novel ways—for assessing disease distributions, tracking business trends, mapping crime patterns, analysing web traffic, and predicting everything from the weather to the behavior of the financial markets, to name but a few examples. (For more on the logic of prediction and pre-emption, see Thomas's contribution to this issue.) Humans are fated to live in environments that contain more information than they can ever fully register and comprehend. The advent of big data marks the moment when new forms of sense-making can be applied to the accumulated data troves (and, correspondingly, the moment when these troves can be amassed, stored, and shared in forms that are amenable to such techniques). So we take the term big data to refer to a combination of size, storage medium, and analytic capability. To refer to big data is not simply to claim that databases contain more information than ever before (although they do), but also to consider the new uses to which that data is put—the novel forms of "actionable intelligence" that emerge from the analysis of ever-expanding data sets. The Library of Congress, for example, has been around for a while, but as its contents are digitized, algorithms can search for patterns and correlations that have been hitherto impossible to detect. If, in the past, there were practical limitations on the ability to track the simultaneous movements of tens or hundreds of thousands of people through a major city, for example (it would be prohibitively expensive to hire enough people to tail everyone and take notes), today the ability to discern useful but non-obvious patterns from the data depends on complex technical systems. Humans simply cannot do that kind of data analysis unassisted.

The formation of big data systems, understood in these terms, has direct consequences for associated forms of surveillance, which avail themselves of both the burgeoning databases and the techniques for making sense of them. Perhaps the most obvious of these is that big data surveillance necessarily relies on automated data analytics. The emerging, massively data-intensive paradigm of knowledge production relies more than ever on highly complex automated systems that operate beyond the reach of human analytical capacities. The CIA can only aspire to "collect everything" if it has at least the hope of putting to use the world redoubled in digital form—something the agency could not hope to do with its small army of human spies alone, or even with the computing capacity it possessed merely a decade or so ago. The reliance on automated data analytics—or data mining—has its own consequences derived from the fact that the goal of such processes is to unearth indiscernible and un-anticipatable patterns from the data (see, for example, Chakrabarti 2009). That is, the data analytic process and its results are systemically, structurally opaque. The legal theorist Tal Zarsky (2013) describes the decisions based on such automated data-mining processes as "non-interpretable" (and thus non-transparent) because of their inherent complexity: "A non-interpretable process might follow from a data-mining analysis which is not explainable in human language. Here, the software makes its selection decisions based upon multiple variables (even thousands)" (2013: 1519). In this regard, the advent of big data surveillance augurs an era in which determinations of risk and suspicion result from complex data interactions that are both un-anticipatable and inexplicable. The database can generate patterns that have predictive power but not necessarily explanatory power. According to this logic, we need not and cannot know how the correlations were derived, or what causal explanations might explain them; we must simply accept that the data science knows best.

Relatively early examples of data mining retained some connection to intuition and explanation. In the 1970s, for example, political operatives discovered that people who drove Mercurys (a model of car) were more likely to vote Republican. There might not be a clear explanation for this, even if it is, perhaps, an intuitable finding, in the sense that the car in both appearance and reputation carried with it a certain set of

associations. We might not be particularly surprised, for example, to learn that, more recently, the music streaming service Pandora discovered that Bob Marley fans are more likely to vote Democratic than Republican (Dwoskin 2014). However, the real prize, from a data-mining perspective, is the generation of completely un-intuitable correlations that nonetheless have predictive (or sorting) power. Let's imagine, for example, that data analysis indicated viewers of a certain age who bike to work and wear glasses are more likely to respond to a toothpaste ad that emphasizes the product's cavity-fighting power than its brightness-inducing qualities. The lure here would be the attempt to find some kind of underlying connection between the stated attributes and the prediction. However, the real goal of data mining is to move beyond this lure in order to arrive at correlations generated by thousands of variables over domains of millions of data points in ways that are untranslatable into any intuitively available pattern.

Zizek, following Lacan, describes this kind of functional non-knowledge as the domain of the symbolic Real: "There is a symbolic Real, which is simply meaningless scientific formulae…you cannot understand quantum physics [for example], you cannot translate it into our horizon of meaning; it consists of formulae that simply function" (Zizek and Daly 2004: 97). We might say something similar of algorithmic correlations and predictions: they do not provide us with underlying, common sense explanations, but offer findings based on a level of complexity that makes them, in some cases, utterly inexplicable. The algorithms do not integrate the findings into our "horizon of meaning" (we may never really understand *why* a particular set of variables is more or less likely to yield a desired outcome); they simply *function*.

The very opacity of the data-mining process suggests that the potential uses of any data set cannot be defined in advance: it may become useful in conjunction with yet-to-be collected data, and it may illuminate activities or outcomes to which it seems entirely unrelated. That is, the specific justification for collecting the data may come only after the fact, thus demanding that all data be collected and stored for its future use-value—its correlational and predictive potential—even if there are no envisioned uses for it at present. Big data surveillance, in this regard, is structurally speculative: data that is seemingly entirely unrelated to a particular strategic objective may well yield the most useful unforeseen correlations.

There is a second sense in which data collection in the context of big data surveillance is speculative: that of attempting to amass an archive that can be searched and sorted retrospectively. The goal is to collect data about everyone, because one never knows who might end up doing something that needs to be excavated and reconstructed (for example, breaking a law). If the archive is complete, according to this logic, then no matter who the suspect is, a relevant data trail can be reconstructed. The data load generated by mobile phones is a case in point. Police have already used mobile phone data to catch thieves by placing them at the scene of the crime and then confirming that their movements coincided with a subsequent car chase (Perez and Gorman 2013). One of the fantasies related to the post-911 goal of "total information awareness" involves the generation of a complete archive that would supplement (or displace) the vagaries of reported actions and memories by externalizing them in the form of machine-readable databases. Hence the recent spate of proposed and actual data retention laws in Europe, Australia, and elsewhere: typically these laws require that telephony and internet providers store data for a minimum period of time (from six months to a year or more) so that they can be accessed by authorities if necessary. Such laws rely on the oligopolistic or monopolistic character of service provision, offloading the surveillance responsibility onto large private sector operators.

Data retention initiatives are often criticized for treating everyone as a potential suspect, but this is not quite right: the assumption is that the vast majority of those tracked will not be considered suspects, and even further (as we are likely to be reminded) that data retention can help exonerate the innocent. Thus, the corollary to the repeated (and questionable) refrain that we need not worry about new forms of data collection as long as we are not doing anything wrong is that the database can come to serve as our alibi. If we are falsely accused or merely the target of suspicion, we have recourse to the monitoring archive that holds records of our whereabouts, our activities, our interactions, and our communications. Alternatively,

for those who run afoul of the law, the archive can be used to link them to the scene of a crime, to reconstruct their movements, to identify and eventually capture them, as was ostensibly demonstrated by the pursuit of the Boston Marathon bombing suspects, originally identified by searching through the archive of security camera video of the event.

However, the Boston Marathon bombing case in fact underscores some holes in the logics of big data that introduce important complications in how this conjuncture of technologies and practices is conceptualized and studied as a problem in Surveillance Studies. There is a case to be made that identifying and capturing the bombing suspects had more to do with human grunt work than automated, high-tech forms of surveillance, big-data or otherwise (see, for example, Sheth and Prasch 2013). Ultimately, the correct suspects were identified by a team of human investigators manually combing through surveillance video gathered on the ground from local businesses. The young men were then apprehended after they shot a campus police officer at MIT, carjacked someone who managed to escape, and then had a shootout with the police. (The surviving suspect, Dzhokhar Tsarnaev, was captured when a Boston resident went outside to smoke a cigarette and saw blood on the side of his boat.) More to the point, if predictive analytics were indeed so sophisticated, why was the Tsarnaev brothers' plot not detected in advance? Given what we now know about the NSA's massive data-gathering machine, how did these two prolific users of email, web browsers, cell phones and social media slip under the predictive radar?

The failure of predictive surveillance in Boston Marathon bombing case has been used (predictably) to provide justification for the need to gather *more* data and develop greater predictive-analytic capacities. But we might also detect here a sense in which the promises of big data and predictive analytics carry with them what William Bogard calls "the imaginary of surveillant control," with the emphasis on imaginary—"a fantastic dream of seeing everything capable of being seen, recording every fact capable of being recorded, and accomplishing these things whenever and wherever possible, prior to the event itself" (1996: 4-5). The fact that this visionary prophetic omniscience remains a "fantastic dream" does not mean that we should be unconcerned about the implications of the new totalizing practices of data collection and mining. Instead it means that we should be careful not to inadvertently allow our thinking to reinforce the flip side of the determinist logic that underpins big data boosterism.

Another important implication of big data-driven surveillance, as we are conceptualizing it here, concerns the importance of attending to the physical and logistical infrastructure that enables it. If the advent of big data is inseparable from the ability to assemble, store, and mine tremendous amounts of data, and if the processing of these data troves is necessarily automated, then infrastructure remains central to emerging forms of surveillance. In fact, the size of the infrastructural capacity is central to understanding what distinguishes something called "big data" from earlier forms of data collection, data-mining and database management. As engineers are keenly aware, increasing the *scale* of technical systems does not involve simply making them bigger in any straightforward sense; instead it often requires their complete reinvention. A similar implication follows for the *effects* of scaled-up systems: they do not simply affect more people or reach more territory, but instead can lead to radical social and material transformations. The mass production of standardized shipping containers, for example, required a massive build-out of infrastructure that in turn radically transformed the terms of global trade and the political-economic geography of the planet (Levinson 2008).

Infrastructures are challenging to describe and virtually impossible to study in their entirety, as Lisa Parks (2012) has noted. There is, on the one hand, a political economy of big data surveillance that explores the implications of ownership and control over the surveillant resources, including the platforms and the networks, the server farms, the algorithms and the cultivation and allocation of data-mining expertise. Such an approach also necessarily considers the relationship between political control and economic resources—the data processing capacity of multinational internet corporations like Google and Facebook, for example, or the ability of the state to access data troves accumulated by these commercial entities.

These companies already have the data collection infrastructure in place to transform communication systems into high-resolution surveillance systems on a global scale. If at one time the accumulation of the world's stored data was the province of the academic sectors and the state (and their various libraries and databases), the accelerated monetization of information has contributed to a dramatically expanding role for the private sector. One recent roundup of the world's largest databases includes three commercial companies in the top five (Anonyzious 2012).

There are also epistemological dimensions to the central role of infrastructure in big data surveillance. Access to the big data resources housed by large corporate entities structures both access to useful "knowledge" and the character of this knowledge itself: not necessarily comprehended content, but excavated patterns. As Google likes to put it: no human reads your email—rather, machines turn it into metadata so as to correlate patterns of communication with patterns of advertising exposure and subsequent purchasing behavior. As Jeremy Packer (2013) argues, the form of knowledge on offer is tied to the infrastructure that generates it. The excavation of purely correlational findings that nonetheless have pragmatic value relies on access to databases and sense-making infrastructures. Packer uses the example of Google (as does the CIA's Gus Hunt, in describing the inspiration of his agency's data-mining practices), whose "computations are not content-oriented in the manner that advertising agencies or critical scholars are. Rather, the effect is the content. The only thing that matters are effects: did someone initiate financial data flows, spend time, consume, click, or conform? Further, the only measurable quantity is digital data" (2013: 297). For Packer, a critique of this shift to effect-*as*-content relies upon an engagement at the level of infrastructure: "Understanding media not merely as transmitters—the old 'mass media' function—but rather as data collectors, storage houses, and processing centers, reorients critical attention toward the epistemological power of media" (2013: 296).

The shift from content to metadata has implications for the convergent character of data mining: even though marketing and national security have received the lion's share of media attention, data analytics play an increasingly important role in a growing number of spheres of social practice, from policing to financial speculation, transport and logistics, health care, employment, consumption, political participation, and education. Data collected by a particular application can often be repurposed for a variety of uses. The music streaming service Pandora, for example, gathers data about user preferences in order to provide customized listening recommendations, but also correlates listening habits with geography, and geography with voting patterns, in order to infer information about listeners' political leanings. Another example is an application developed by Microsoft Research that apparently is able to predict users' moods based on their patterns of smart phone use.

The point is that so-called "function creep" is not ancillary to the data collection process, it is built into it—the function *is* the creep. Continuous repurposing of information initially gathered for other purposes is greatly facilitated by digitization, which makes storage, sharing, and processing easier. But function creep is also made enabled by the new "save everything" logic of automated data analysis (Morozov 2013), where the relevance of any piece of data to future correlations and predictions can never be ruled out in advance.

All of these characteristics of the deployment of big data and associated forms of data mining have implications for how we think about the changing character of surveillance in data-driven contexts. (Klauser and Albrechtslund's contribution to this special issue proposes a framework for exploring these changes.) To approach the developing character of big data surveillance, we might start by considering some recent influential definitions of surveillance. In their report on "The Surveillance Society" for the UK information commissioner, David Murakami Wood and colleagues define surveillance as, "purposeful, routine, systematic and focused attention paid to personal details, for the sake of control, entitlement, management, influence, or protection" (Murakami Wood et al. 2006: 4). They further emphasize that, "surveillance is also systematic; it is planned and carried out according to a schedule that

is rational, not merely random" (Murakami Wood et al. 2006: 4). Similarly, in his influential formulation of "dataveillance" in the digital era, Roger Clarke refers to "the systematic monitoring of people or groups, by means of personal data systems, in order to regulate or govern their behaviour" (Clarke 1987). He subsequently distinguishes between targeted personal dataveillance and "mass dataveillance, which involves monitoring large groups of people" (Clarke 2003). A third example is Haggerty and Ericson's (2000) concept of the "surveillant assemblage," often cited in recent work to theorize the shape and character of distributed, networked data monitoring on a mass scale (Cohen 2013; Klauser 2013; Murakami Wood 2013; Vukov and Sheller 2013; Salter 2013; etc.). The surveillant assemblage, to remind readers, "operates by abstracting human bodies from their territorial settings and separating them into a series of discrete flows. These flows are then reassembled into distinct 'data doubles' which can be scrutinized and targeted for intervention" (Haggerty and Ericson 2000: 606).

Such definitions remain productive for analysing many forms of contemporary surveillance, but they require some qualification in the context of big data. In particular, the speculative and totalizing aspects of big data collection transform the systematic and targeted character of surveillance practices. The notion that surveillance is systematic and targeted takes on a somewhat different dimension when the goal is to capture any and all available data. Drone data collection via the "Air Handler," for example, is more opportunistic than systematic (as was the capture of information from local WiFi networks by Google's Street View cars). By the same token, the uses for such data are often more speculative than defined.

The very notion of a surveillance target takes on a somewhat different meaning when surveillance relies upon mining large-scale databases: the target becomes the hidden patterns in the data, rather than particular individuals or events. Data about the latter are the pieces of the puzzle that need to be collected and assembled in order for the pattern to emerge. In this regard, the target for data collection becomes the entire population and its environment: "all of reality" is, as Packer puts it, "now translatable. The world is being turned into digital data and thus transformable via digital manipulation" (2013: 298). This, of course, is the wager of big data surveillance: that those with access to the data have gained some power over the informated world, and that the patterns which emerge will give those with access to them an advantage of some kind. Big data surveillance, then, relies upon control over collection, storage, and processing infrastructures in order to accumulate and mine spectacularly large amounts of data for useful patterns. Big data surveillance is not about understanding the data, nor is it typically about explaining or understanding the world captured by that data—it is about intervening in that world based on patterns available only to those with access to the data and the processing power. Big data surveillance is not selective: it relies on scooping up as much information as possible and sorting out its usefulness later. In the big data world there is no functional distinction between targets and non-targets (suspects and non-suspects) when it comes to data collection: information about both groups is needed in order to excavate the salient differences between them.

Big data surveillance looks much less parsimonious than the panoptic model that has played such an important role in conceptualizing and critiquing the relationship between surveillance and power. Bentham's model was meant to leverage uncertainty in the name of efficiency—the properly functioning panopticon allowed just one supervisor to impose discipline upon the entire prison population. Bentham speculated that once the system had been implemented it might continue to function even in the absence of a supervisor—with just an empty, opaque tower looming nearby to warn inmates that they might be monitored at any time. This is the logic of "dummy" speed cameras or surveillance cameras: that the spectacle of surveillance carries with it its own power. Compared to this alleged model of efficiency, the big data model looks somewhat extravagant: rather than mobilizing uncertainty (as to whether one is being watched or not), it mobilizes the promise of data surfeit: that the technology is emerging to track everything about everyone at all times—and to store this data in machine-readable form. Of course, the danger of such a model is that it must necessarily fall short of its goal; the only way to ensure that nothing is overlooked is to reproduce the world in its entirety in data-recorded form, and therefore to record data

about the recording process itself, and so on, indefinitely. The result is that one of the hallmarks of big data surveillance is its structural incompleteness.

This incompleteness has its own consequences, given that the big data model attempts to move beyond the sampling procedures of other forms of partial data collection to encompass the entire population. Put somewhat differently, there is no guarantee that the data collected is either comprehensive or representative. The sheer size of the database is meant to compensate for these lacks, but does not prevent systematic forms of bias from working their way into the data trove (or the algorithms that sort it). Shoshana Magnet (2011) and Kelly Gates (2011), for example, have each demonstrated the ways in which debunked conceptions of racial identity work their way into biometric identification technologies. The database carries with it associations of objectivity that can "launder" the forms of bias that are baked into the data collection and sorting processes. Thus, a closer look at the labor that goes into shaping these processes is a crucial component of critical approaches to big data surveillance (see French's contribution to this issue).

One rejoinder to such critiques, however, is that they rely upon an outmoded approach to the data: an over-emphasis on its content rather than its functional efficacy. This is Packer's Kittler-inspired point: when the effect is the content, all other questions of referentiality (is the data representative, complete, etc.?) fall by the wayside. Presumably if there is a problem somewhere along the line, then the ability to attain the desired effect is impaired, but relative to what? There is no standard of truth, or even correctness in such a model (as is implied, for example, by the standard of "the population" in probability sampling). There is merely the bar set by existing forms of practice. If a particular recommendation algorithm achieves a higher rate of success in getting people to watch movies, purchase books, or click on links, then it has succeeded. Questions about the representativeness of the data or biases in the algorithm are subordinated to this measure of success. In this regard, it would be somewhat misleading to say that the top priority of big data surveillance is to obtain as accurate and complete a view of the world as possible. Rather its goal is to intervene in the world as effectively as possible, which may well entail lower standards of comprehensiveness and accuracy.

The complications that totalizing forms of data capture and analysis introduce for established ways of defining surveillance also suggest the need to re-examine existing legal frameworks. The idea that the institutions engaged in totalizing data capture are focused on distributions and patterns emerging in populations, and presumably uninterested in targeting particular individuals, allows them to circumvent established rights-based principles. In the United States, for example, the Fourth Amendment to the Constitution prohibits search and seizure of "persons, houses, papers, and effects" without a legally issued warrant that "particularly describ[es] the place to be searched, and the persons or things to be seized." This legal right is systematically subverted, not only by big data's "no content, just metadata" assertion, as José van Dijck refers to it in this issue, but also by a whole parallel extra-legal domain of industry "self-regulation." The self-regulatory approach to privacy protection relies on so-called voluntary disclosure of personal data, written into the incomprehensible, small-type "privacy policies" that people agree to daily as a condition of participation in the online economy (Turow 2006). Such policies rely upon "an intrinsically asymmetrical relationship," as Sara Degli Esposti notes in her contribution to this issue. The point here is that the big data paradigm, in its applications that rely on data about human beings, is built on the foundation of what is, in reality, a regime of compulsory self-disclosure. And this regime is supported by and commensurate with a normalized and permanent "state of exception," in which individual legal rights are always suspended in any case, for all who might choose to opt out (because after all, opting out itself looks suspicious).

In short, if the totalizing and massively scaled-up data paradigm requires new ways of conceptualizing surveillance, it also requires renewed efforts at rescuing and reinventing the legal arguments and interventions that can be used to address and curb these practices. Legal scholar Julie Cohen (2013) offers

one such effort at rethinking privacy in light of big data in her recent *Harvard Law Review* piece, "What privacy is for"; among her recommendations is a reemphasis on the legal and moral necessity of due process. From another angle, Jay Stanley (2013) of the ACLU offers a short but persuasive argument against the effectiveness of applying big data to predict terrorist attacks, using an analogy from physics called Brownian motion: a water molecule's path through water is easy to understand in retrospect but impossible to predict in advance. If strategies for fighting terrorism continue to take the approach of amassing greater and greater quantities of data about everybody, Stanley explains, the outcome is indeed predictable: "many more incredulous senators, amazed that our security agencies failed to thwart attacks when all the signs seemed so 'clear' in advance." Of course, the challenge of reframing the policy debate returns anew with each terrible tragedy of this kind; terrorist attacks in particular seem uniquely suited to technological-solutionist responses.

Certainly another way of challenging the ethical and legal underpinnings of big-data surveillance is by pointing to the glaring contradiction in the demand on the part of state and corporate actors that their operations remain strictly confidential, even as those same operations require individuals to relinquish all rights to privacy and lay their lives bare. Among the many deafening warning alarms raised by Edward Snowden's NSA revelations is the need to hold the state accountable to the same if not greater scrutiny than what it is now demanding of citizens en masse, starting with all claims that state agencies make to so-called state-secrets privileges. And an equal measure of scrutiny should be levelled at the trade-secrets claims of all non-state institutions with big-data capabilities.

Obviously, more systematic and targeted forms of surveillance have not dropped by the wayside, and definitions of surveillance that account for such practices remain relevant. We are not arguing that the character of surveillance has changed in all contexts. Rather we seek to identify salient aspects of emerging forms of data-driven surveillance and thereby to gesture toward their societal consequences. The shift toward totalizing data capture was more or less apparent well before *The Guardian* published its first NSA spying revelation in June 2013. But the steady stream of revelations trickling out from Snowden's files gives us cause to consider how well our established theoretical and political approaches to surveillance measure up to the challenges of big data, real or imagined. Given its heavy reliance on infrastructure, big data surveillance is available only to those with access to the databases and the processing power: it is structurally asymmetrical. Likewise, the forms of knowledge it generates are necessarily opaque; they are not shareable understandings of the world, but actionable intelligence crafted to serve the imperatives and answer the questions of those who control the databases. These forms of knowledge push against any attempt to delimit either the collection of data or the purposes to which that data are turned.

Contributions to this issue

While the range of topics and issues addressed here is by no means exhaustive, the contributors to this special issue cover considerable ground in their analyses of big data and predictive analytics as issues of concern to Surveillance Studies. Some of the contributors offer speculative theoretical reflections, while others offer empirical case studies grounded in specific domains. They address a selection of key areas: pandemics and disease surveillance systems, public health surveillance, self-tracking and the "Quantitative Self" movement, urban digital infrastructures, the big business of big data, crime prediction programs, and more. The authors also elaborate on a number of critical concepts, some borrowed from other sources, that should prove analytically productive for further research: datafication, dataism, dataveillance, analytical competitors, algorithmic governmentality, and others. There are many other areas and topics that might have fit within the scope of this issue. Promising areas for further research include the role of big data surveillance in the finance industry and the analysis of the financial markets; the accelerated monetization of data at the heart of big data as a business model; the impact and role of big data in political polling and public opinion analysis; market research and sentiment analysis; policing, crime mapping and crime

prediction; military applications of big data; and many more. It also is important to note that authors originally submitted work for consideration in this theme issue before the first Snowden NSA story broke in June 2013, with the reviewing and revising process extending through the initial revelation and the steady stream of exposures that followed over the latter half of 2013. Thus, while some of the authors incorporated the newly visible NSA activities into their articles during the revision process, none of the pieces published here are principally focused on PRISM or other NSA programs revealed by Edward Snowden's files.

In her contribution, José van Dijck examines the logic of "datafication" that underpins the big data paradigm—a belief in the value- and truth-producing potential inherent in the mass collection and analysis of data and metadata (Mayer-Schoenberger and Cukier 2013). Among the problems van Dijck identifies with the datafication view is the assumption of a direct and self-evident relationship between data and people, as if the digital traces users leave behind online reveal their essential traits and tendencies, from their spiritual beliefs and intelligence levels to their earning potential and predisposition to diseases. van Dijck sees an ideology of dataism in the celebratory view of datafication espoused by big data proponents—the resurgence of a flawed faith in the objectivity of quantification, and a misguided view of metadata as the "raw material" of social life. The opacity of data science supports the ideology of dataism, obscuring the priorities that shape both the range of data collected and the ways that data gets analysed. van Dijck also explores the problem of trust at the heart of dataism, insisting that what is at stake is not just our trust in specific government agencies or corporations, but in the integrity of the entire networked ecosystem.

In her essay, "When Big Data Meets Dataveillance," Sara Degli Esposti draws substantially on business literature to identify some of the specific ways that business priorities shape the practices of big data analytics, as well as the kinds of answers that are generated. Information technology companies, she explains, are now differentiated based on their capacity for analyzing big data—"analytical competitors" are so named for their ability to outperform their peers in terms of their ability to apply analytics to their own data, drawing on internal expertise (Davenport and Harris 2007). Of course, not all companies have such high-skilled in-house labor, creating a market for "analytical deputies," or companies building their business models on conducting large-scale analyses of data for less data-savvy corporate customers. Degli Esposti employs Roger Clarke's concept of dataveillance to parse out the surveillant aspects of big-data business practices along four dimensions: recorded observation, identification and tracking, analytical intervention, and behavioral manipulation. While these dimensions of dataveillance operate in a feedback loop, it is forms of *analytical intervention* in particular that enable businesses operating in different sectors to achieve key objectives, "from customer acquisition and retention to supply chain optimization; from scheduling or routing optimization to financial risk management; from product innovation to workforce recruitment and retention." Analytical intervention also allows companies to engage more effectively in profit-maximizing price discrimination strategies. Moreover, while businesses depend increasingly on data-mining agents—computer programs designed to automatically identify patterns in data—what is clear from Degli Esposti's discussion is that *human* expertise remains central to the knowledge- and profit-producing potential of big data. Big data companies employ "tech savvy CEOs" as well as highly trained data scientists with the sophisticated statistical and mathematical skills necessary to create advanced analytical applications. Big data surveillance is a high-tech, high-skill operation.

Martin French's contribution also zeroes in on the production processes that constitute big data systems, focusing on the activities of Canadian health care professionals. Specifically, he presents a case study that examines the process challenges associated with the implementation of a large-scale, regionally interconnected public health information system in Ontario, Canada. Countering the "informatic ethos" that conceals the labor that makes IT systems work, French focuses on what he calls "informatics practice," or the combined human and non-human labor activity that materializes information, the practical activities that contribute to "making information a material reality in quotidian life." Viewed

from this on-the-ground perspective, the foundations of new forms of data-driven monitoring appear more subject to the vicissitudes of individual practices, convenience, and happenstance than suggested by monolithic portrayals of big data as a black-box category. The work of public health professionals is often at odds with IT system operations, for example, and such professionals are often very much concerned with patient privacy. In French's case study, "the everyday operation of health IT not only complicated the work of public health, but also blurred public health's surveillant gaze." French's article shifts the lens from the impact of surveillance on those subjected to it, to its impact on the work practices of those tasked with operating and otherwise making the monitoring systems function.

In his contribution, Tyler Reigeluth conceptualizes big data practices as forms of "algorithmic governmentality," (Berns and Rouvroy, in Reigeluth, this issue), considering how these practices take shape as forms of subjectivation in the Foucauldian sense—interventions that bring human subjectivities into being. He expands on the notion of "digital traces," variously understood as the marks, prints, infinitesimal pieces and intersection points that are gathered together and analysed to make sense of individuals and collectives. Even in academic thought, Reigeluth notes, there seems to be some agreement that identities are "the collection or the sum of digital traces." Similar to van Dijck, he asks what is at stake in understanding digital traces as the raw material of human identities, and the basis for discovering fundamental truths about them. Rather than seeing big data analytics as a radical departure from existing forms of subjectivation, Reigeluth suggests that the concept of digital trace can shed light on the ways digital technology is continuous with long-standing institutional and technological arrangements for shaping human subjectivities by structuring the environments they inhabit. Doing so requires opening up the black box of data capture and analysis, following the traces as they enter into and exit algorithmic systems and their physical infrastructures. Focusing on two manifestations of the big data paradigm—a crime prevention program called PredPol (short for "Predictive Policing") and the Quantified Self movement—Reigeluth considers the ironies of algorithmic rationality. "If the ideal individual is perfectly correlated and immanent to his environment," he asks, and "if her singularity can be reduced to the degree to which she fulfills these correlations," then is it actually possible for an ethical and political subject to exist as such?

Martijn van Otterlo explores the ways in which the flexibility of digital environments can double as laboratory and virtual Skinner box, enabling an ongoing process of experimentation in social control. He explores the ways in which digital environments are constantly modulated for the purposes of determining how best to influence the behavior of those whose actions can be captured by interactive forms of data collection. If, as Otterlo suggests, paranoia is a "step in the right direction" to the transformation of cyberspace into myriad cybernetic laboratories, this is a consequence of the capitalist imaginary at play in the fields of big data. He cites the example of a Microsoft patent for an automated system that monitors and analyses employee behavior in order to ensure "that desired behaviors are encouraged and undesired behaviors are discouraged." It turns out that, from the perspective of managers, marketers, and other associated authorities, futuristic digital dreams recapitulate familiar fantasies of control and manipulation with deep roots in the history of media technologies. Otterlo's contribution helpfully opens up a range of social spheres to a consideration of the relationship between big data mining and experimentation beyond the familiar one of advertising. By drawing on examples from a range of recent appropriations of data-mining technology, he argues that we need to consider the implications of big data-driven forms of monitoring and surveillance in the realms of politics, education, policing, and the workforce, among others.

One of the attributes of "big data" that emerge from the contributions to this issue is the increasing reach of data mining, and the unfolding of new registers of monitoring and surveillance. Francisco Klauser and Anders Albrechtslund's article proposes a framework for research on big data based on the four axes of "agency, temporality, spatiality and normativity." The virtue of such an approach is that it draws upon the wide-ranging uses of data-driven monitoring to broaden the reach of the study of surveillance. The article

invokes the seemingly disparate practices of self-monitoring on the one hand, and urban surveillance associated with "smart cities" on the other, to trace commonalities in data-mining techniques and their relation to forms of social control. The authors argue for approaches to data-driven forms of monitoring that supplement critiques of discipline at the individual level with those of regulation at the population level. In other words, when decisions are made at the aggregate level, drawing on probability levels generated by data mining, the focus is not on particular individuals but on aggregate outcomes. The authors also argue for expanding the reach of Surveillance Studies beyond the monitoring of humans to consider the wide array of objects and contexts about which information is collected. The fantasy of "big data" is that it might become powerful enough to create a comprehensive data double of both the social world and the object world (and their interactions).

Lindsay Thomas's exploration of the logic of disease monitoring systems designed to track and anticipate the spread of contagious illnesses doubles as a meditation on the temporality of pre-emption more generally. Disease monitoring partakes of the logic of predictive analytics: the mobilization of data collected in the past to model possible futures in the present. As she puts it, "The future is anticipated and surveilled using past data," a formulation that could apply equally to financial modeling, crime prediction, climate modeling, and much more. The paradox of such forms of modeling, she notes, is not simply that, when it comes to catastrophes like pandemics, they attempt to forestall future events by interjecting them into the present, but that the very attempt to collapse temporality pushes in the direction of a catastrophic stasis: "their continual construction of soon-to-arrive pandemics normalizes catastrophe. They build 'future' catastrophes all around us, teaching us to accept them, and, by extension, the measures we all must take to prepare for them, as given."

Taken together, the contributions to the issue develop some of the emerging themes in explorations of big data. One is the tension between the promise of predictive control embodied in the big data paradigm, and the realities of biased, incomplete data. Closely related to this tension is the promotional hype associated with new forms of data collection and mining, the misplaced faith in big data that Morozov (2013) aptly refers to as "the folly of technological solutionism." The contributors also explore the epistemological claims associated with the forms of "knowledge" that can be extracted from sorting and analyzing increasingly enormous, merged datasets. Perhaps most importantly, they provide a starting point for studying the social, political, and cultural consequences of a burgeoning domain of automated surveillance. In this regard, big data is not simply a matter of the size of the database, but of the claims made on its behalf, and on its application to an ever-expanding range of social practices.

References

Anonyzious. 2012. 10 Largest Databases of the World. *Realitypod.com*, March 24, 2012. http://realitypod.com/2012/03/10-largest-databases-of-the-world/

Bogard, William. 1996. *The Simulation of Surveillance: Hypercontrol in Telematic Societies*. Cambridge: Cambridge University Press.

Chakrabarti, Soumen. 2009. *Data-mining: Know it All*. New York: Morgan Kaufmann.

Clarke, Roger. 1987. Information Technology and Dataveillance. http://www.rogerclarke.com/DV/CACM88.html

Clarke, Roger. 2003. Dataveillance – 15 Years On. http://www.rogerclarke.com/DV/DVNZ03.html

Cohen, Julie. 2013. What Privacy is for. *Harvard Law Review* 126: 1904-1933.

Davenport, Thomas H. and Jeanne G. Harris. 2007. *Competing on Analytics: The New Science of Winning*. Boston: Harvard Business School Press.

Dwoskin, Elizabeth. 2014. Pandora Thinks It Knows if You Are a Republican. *The Wall Street Journal*. February 13. http://online.wsj.com/news/articles/SB10001424052702304315004579381393567130078

Gates, Kelly. 2011. *Our Biometric Future: Facial Recognition Technology and the Culture of Surveillance*. New York: New York University Press.

Goodman, Amy. 2014. Death by Metadata: Jeremy Scahill and Glenn Greenwald Reveal NSA Role in Assassinations Overseas. *Democracy Now!* http://www.democracynow.org/2014/2/10/death_by_metadata_jeremy_scahill_glenn

Haggerty, Kevin D. and Richard V. Ericson. 2000. The Surveillant Assemblage. *British Journal of Sociology* 51 (4): 605-622.

Klauser, Francisco. 2013. Political Geographies of Surveillance. *Geoforum* 49 (October 2013): 275-278.

Levinson, Marc. 2008. *The Box: How the Shipping Container Made the World Smaller and the World Economy Bigger*. Princeton: Princeton University Press.

Magnet, Shoshana. 2011. *When Biometrics Fail: Gender, Race, and the Technology of Identity*. Durham: Duke University Press.

Mayer-Schoenberger, Viktor and Kenneth Cukier. 2013. *Big Data. A Revolution that Will Transform How We Live, Work, and Think*. London: John Murray Publishers.

Morozov, Evgeny. 2013. *To Save Everything, Click Here: The Folly of Technological Solutionism.* New York: PublicAffairs.

Murakami Wood, David, Kirstie Ball, David Lyon, Clive Norris, and Charles Raab. 2006. A Report on the Surveillance Society. Report for the UK Information Commissioner's Office. *Surveillance Studies Network, UK*. http://ico.org.uk/~/media/documents/library/Data_Protection/Practical_application/SURVEILLANCE_SOCIETY_FULL_REPORT_2006.ashx

Murakami Wood, David. 2013. What is Global Surveillance? Towards a Relational Political Economy of the Global Surveillant Assemblage. *Geoforum* 49: 317-326.

Packer, Jeremy. 2013. Epistemology Not Ideology OR Why We Need New Germans. *Communication and Critical/Cultural Studies* 10(2-3): 295-300.

Parks, Lisa. 2012. Things You Can Kick: Conceptualizing Media Infrastructures. Paper presented at the annual meeting of the American Studies Association Annual Meeting, Puerto Rico Convention Center and the Caribe Hilton, San Juan, Puerto Rico, Nov. 15-18, 2012.

Perez, Evan and Siobhan Gorman. 2013. Phones Leave a Telltale Trail. *The Wall Street Journal*, June 15. http://online.wsj.com/news/articles/SB10001424127887324049504578545352803220058

Salter, Mark B. 2013. To Make Move and Let Stop: Mobility and the Assemblage of Circulation. *Mobilities* 8 (1): 7-19.

Sheth, Falguni A. and Rober E. Prasch. 2013. In Boston, Our Bloated Surveillance State Didn't Work. *Salon.com*, April 22, http://www.salon.com/2013/04/22/in_boston_our_bloated_surveillance_state_didnt_work/

Sledge, Matt. 2013. CIA's Gus Hunt On Big Data: We 'Try To Collect Everything And Hang On To It Forever.' *Huffington Post*, March 20. http://www.huffingtonpost.com/2013/03/20/cia-gus-hunt-big-data_n_2917842.html

Stanley, Jay. 2013. The Asymmetry Between Past and Future, and Why It Means Mass Surveillance Won't Work. *ACLU.org,* May 13. https://www.aclu.org/blog/national-security/fundamental-asymmetry-between-past-and-future-and-why-it-means-mass

Turow, Joseph. 2006. *Niche Envy: Marketing Discrimination in the Digital Age*. Cambridge, MA: MIT Press.

Vukov, Tamara and Mimi Sheller 2013. Border Work: Surveillant Assemblages, Virtual Fences, and Tactical Counter-Media. *Social Semiotics* 23 (2): 225-241.

Zarsky, Tal. 2013. Transparent Predictions. *University of Illinois Law Review* 4: 1503-1570.

Zizek, Slavoj and Glynn Daly. 2004. *Conversations with Zizek*. Cambridge: Polity.