International Journal of Communication 8 (2014), 1673–1689 1932–8036/20140005

IJoC Big Data, Big Questions

The Big Data Divide

MARK ANDREJEVIC[1]
Pomona College, USA

This article extends the notion of a "big data divide" to describe the asymmetric relationship between those who collect, store, and mine large quantities of data, and those whom data collection targets. It argues that this key distinction highlights differential access to ways of thinking about and using data that potentially exacerbate power imbalances in the digital era. Drawing on original survey and interview findings about public attitudes toward collection and use of personal information, it maintains that the inability to anticipate the potential uses of such data is a defining attribute of data-mining processes, and thus of the forms of sorting and targeting that result from them.

Keywords: big data, data mining, privacy, digital divide, predictive analytics

Between Me and My Data

Contributing to the buzz around "the personal data revolution," Web founder and knighted new media guru Tim Berners-Lee recently issued a plea for Internet users to be able to access their personal data. All people should have the resources for data-mining themselves, he proclaimed, since "My computer has a great understanding of my state of fitness, of the things I'm eating, of the places I'm at. My phone understands from being in my pocket how much exercise I've been getting and how many stairs I've been walking up and so on" (Katz, 2012, para. 3). Echoing a well-worn set of claims about the power of machines to know ourselves better than we do (e.g., Gates, 1995, on software agents or Negroponte, 1996, on digital butlers), Berners-Lee portrayed the database as a personal-service resource: "If my computer understands all that, then it's in a position to be very valuable to help me run my life, you know, to guess what I need next, to fill in a lot of the context . . . to guess what I want to read in the morning" (Katz, 2012, para. 1, embedded recording).

[1] This research was supported by the Australian Research Council's Discovery Project's funding scheme (DP1092606).

Mark Andrejevic: mark.andrejevic@pomona.edu
Date submitted: 2013–04–06

Of course, Google News and any number of aggregators and services are already hard at work providing these kinds of services, without users needing to get involved or reclaim access to their data trails. Berners-Lee's point, however, is that personal devices might usefully combine data from different social-networking "silos" and other applications and devices, since these form a personal informational nexus where all types of different data rub shoulders (a personal NSA, as it were):

> There are no programmes that I can run on my computer which allow me to use all the data in each of the social networking systems that I use plus all the data in my calendar plus in my running map site, plus the data in my little fitness gadget and so on to really provide an excellent support to me. (Katz, 2012, para. 4)

Berners-Lee is bemoaning a growing separation of people from their data that characterizes the lives of active users of interactive devices and services—a form of *data divide* not simply between those who generate the data and those who collect, store, and sort it, but also between the capabilities available to those two groups. Berners-Lee challenges one aspect of that divide: If we generate data that is potentially useful to us, he reasons, shouldn't we be able to access it and put it to use? Why not overcome this separation between users and their data, and with it the separation between the different data silos we generate on various devices and platforms?

Surely he has a point, but it raises a further one: Even if users had such access, what individuals can do with their data in isolation differs strikingly from what various data collectors can do with this same data in the broader context of everyone else's data. To take a familiar example, Berners-Lee mentions customized news delivery as one possible benefit of self-data-mining: if a computer knows what its users have read in the past, it might be able to predict which news stories will interest them in the future (this is, of course, an echo of Negroponte's [1996] "Daily Me," and perhaps also his "digital butler"). But online news aggregators take into account not only one's own interest patterns (surely not formed in isolation) but also those of everyone else about whom they collect data. This data trove enables them to engage in various forms of collaborative filtering—that is, to consider what the other people who share one's interests are also interested in.

Generalizing this principle from the perspective of data mining, it is potentially much more powerful to situate individual behavior patterns within the context of broader social patterns than to rely solely on the historical data for a particular individual. Put somewhat differently, allowing users access to their own data does not fully address the discrepancies associated with the data divide: that is, differential capacities for putting data to use. Even if users had access to their own data, they would not have the pattern recognition or predictive capabilities of those who can mine aggregated databases. Moreover, even if individuals were provided with everyone else's data (a purely hypothetical conditional), they would lack the storage capacity and processing power to make sense of the data and put it to use. It follows that the structural divide associated with the advent of new forms of data-driven sense making will be increasingly apparent in the era of "big data."

To characterize the differential ability to access and use huge amounts of data, this article proposes the notion of a *big data divide* by first defining the term, then considering why such a divide merits attention, and then exploring how this divide might relate to public concern about the collection and use of personal information. The sense of powerlessness that individuals express about emerging forms of data collection and data mining reflects both the relations of ownership and control that shape access to communication and information resources, and growing awareness of just how little people know about the ways in which their data might be turned back upon them. Although the following research will focus exclusively on personal data—the type of data at the heart of current debates about regulation of data collection online—the notion of a big data divide is meant to invoke the broader issue of access to sense-making resources in the digital era, and the distinct ways of thinking about and using data available to those with access to tremendous databases and the technology and processing power to put them to use.

From a research perspective, boyd and Crawford (2011) have noted the divide between "the Big Data rich" (companies and universities that can generate or purchase and store large datasets) and the "Big Data poor" (those excluded from access to the data, expertise, and processing power), highlighting the fact that a relatively small group with defined interests threatens to control the big data research agenda. This article extends the notion of a big data divide to incorporate a distinction between ways of thinking about data and putting it to use. It argues that "big data mining" privileges correlation and prediction over explanation and comprehension in ways that undermine the democratizing/empowering promise of digital media. Despite the rhetoric of personalization associated with data mining, it yields predictions that are probabilistic in character, privileging decision-making at the aggregate level (over time). Moreover, it ushers in an era of "emergent social sorting," the ability to discern un-anticipatable but persistent patterns that can be used to make decisions that influence the life chances of individuals and groups. In online tracking and other types of digital-era data surveillance, the logic of data mining, which proposes to reveal unanticipated, unpredictable patterns in the data, renders notions such as informed consent largely meaningless. Data miners' claims, discussed in more detail in the following sections, reveal that big data holds promise for much more than targeted advertising: It is about finding new ways to use data to make predictions, and thus decisions, about everything from health care to policing, urban planning, financial planning, job screening, and educational admissions. At a deeper level, the big data paradigm challenges the empowering promise of the Internet by proposing the superiority of a post-explanatory pragmatics (available only to the few) to the forms of comprehension that digital media were supposed to make more accessible to the many. None of these concerns fits comfortably within the standard privacy-oriented framing of issues related to the collection and use of personal information.

A Big Data Divide

In the sense of standing for more information than any individual human or group of humans can comprehend, the notion of big data has existed since the dawn of consciousness. The world and its universe are, to anything or anyone with senses, incomprehensibly big data. The contemporary usage is distinct, however, in that it marks the emergence of the prospect of making sense of an incomprehensibly large trove of recorded data—the promise of being able to put it to meaningful use even though no individual or group of individuals can comprehend it. More prosaically, big data denotes the moment when automated forms of pattern recognition known as *data analytics* can catch up with automated forms of data collection and storage. Such data analytics are distinct from simple searching and querying of large data sources, a practice with a much longer legacy. Thus, for the purposes of this article, the big data moment and the advent of data-mining techniques go hand in hand. The magnitude of what counts as big data, then, will likely continue to increase to keep pace with both data storage and data processing capacities. IBM, which is investing heavily in data mining and predictive analytics, notes that big data is not just about size but also about the speed of data generation and processing and the heterogeneity of data that can be dumped into combined databases. It describes these dimensions in terms of the three "Vs": volume, velocity, and variety (IBM, 2012, para. 2).

Big-data mining is omnivorous, in part because it has embarked on the project of discerning unexpected, unanticipated correlations. As IBM puts it, "Big data is any type of data—structured and unstructured data such as text, sensor data, audio, video, click streams, log files and more. New insights are found when analyzing these data types together" (IBM, 2012, para. 9). Data can be collected, sorted, and correlated on a hitherto unprecedented scale that promises to generate useful patterns far beyond the human mind's ability to detect or even explain. As data-mining consultant Colleen McCue (2007) puts it, "With data mining we can perform exhaustive searches of very large databases using automated methods, searching well beyond the capacity of human analysts or even a team of analysts" (p. 23). In short, data mining promises to generate patterns of actionable information that outstrip the reach of the unaided human brain. In his book *Too Big to Know*, David Weinberger (2011) describes this "new knowledge" as requiring "not just giant computers but a network to connect them, to feed them, and to make their work accessible. It exists at the network level, not in the heads of individual human beings" (p. 130).

Such observations trace the emerging contours of a "big data divide" insofar as putting the data to use requires access to and control over costly technological infrastructures, expensive data sets, and the software, processing power, and expertise for analyzing them. If, as Weinberger puts it, in the era of big data "the smartest person in the room is the room," then much depends on who owns and operates the room. The forms of "knowing" associated with big data mining are available only to those with access to the machines, the databases, and the algorithms. Assuming for the sake of argument that the big data prognosticators (e.g., Mayer-Schönberger & Cukier, 2012) are correct, the era of big data—characterized by the ability to make use of databases too large for any individual or group of individuals to comprehend—ushers in powerful new capabilities for decision making and prediction unavailable to those without access to the databases, storage, and processing power. In manifold spheres of social practice, then, those with access to databases, processing power, and data-mining expertise will find themselves advantageously positioned compared to those without such access. But the divide at issue is not simply between what boyd and Crawford (2011) describe as database "haves" and "have-nots"; it is also about asymmetric sorting processes and different ways of thinking about how data relate to knowledge and its application. The following sections consider each of these issues in turn.

The Big Data Sort

For those with database access, the ability to capture and mine tremendous amounts of data considerably enhances and alters possibilities for engaging in what David Lyon (2002), building on the work of Oscar Gandy (1993), has called "surveillance as social sorting": "a means of verifying identities but also of assessing risks and assigning worth" (p. i). Those with access to data, expertise, and processing power are positioned to engage in increasingly powerful, sophisticated, and opaque forms of sorting that can be "powerful means of creating and reinforcing long-term [or newly generated] social differences" (Lyon, 2002, p. i). The very notion of a panoptic sort is premised on a power imbalance between those positioned to make decisions that affect the life chances of individuals (in Gandy's original work, businesses as both employers and marketers) and those subjected to the sorting process. Subsequently reflecting on the notion of the "panoptic sort," Gandy observed that he had "come to understand that these decisions are not really based on an assessment of who or what people are, but on what they will do in the future. The panoptic sort is not only a discriminatory technology, but it is one that depends upon an actuarial assumption" (Gandy, 2005, p. 2). This observation remains as salient as ever in the era of data mining and predictive analytics, which, while deploying the rhetoric of personalization, also operate at a probabilistic level. In this regard, the assertion that data mining augers a future "in which predictions seem so accurate that people can be arrested for crimes before they are committed," is misleading (Kakutani, 2013, para. 14). Predictive analytics is not, despite the hype, a crystal ball. As one commentator put it,

> When you are doing this kind of analytics, which is called 'big data,' you are looking at hundreds of thousands to millions of people, and you are converging against the mean. I can't tell you what one shopper is going to do, but I can tell you with 90 percent accuracy what one shopper is going to do if he or she looks exactly like one million other shoppers. (Nolan, 2012, p. 15)

But the confusion between fortune telling and forecasting is consequential, for decisions made at a probabilistic, aggregate level produce effects felt at an individual level: the profile and the person intersect. To someone who has been denied health care, employment, or credit, the difference between a probabilistic prediction and a certainty is, for all practical purposes, immaterial.

Social sorting has a long history but comes into its own as a form of automated calculus, as Gandy (1993) suggests, in the era of modern bureaucratic rationality. Thus, it is tempting to note the historical continuity between big data-driven forms of social sorting and earlier forms of data-based decision making, from Taylorist forms of "scientific management" to mid-20th-century forms of redlining in the banking, housing, and insurance industries. Raley (2013), for example, has noted that in an early account of computer-enabled surveillance, David Lyon (1994) "suggests that the difference made by information technologies is one of degree, not of kind, that they simply 'make more efficient more widespread, and simultaneously less visible many processes that already occur'" (Raley, 2013, p. 124). However, a qualitative shift in monitoring-based social sorting results from the "emergent" character of new data-mining processes, which now can generate un-anticipatable and un-intuitable predictive patterns (e.g., Chakrabarti, 2009). That is, their systemic, structural opacity creates a divide between the kinds of useful "knowledge" available to those with and without access to the database.

In the following sections, I argue that emerging awareness of forms of asymmetrical power associated with both the tremendous accumulation of data and new techniques for putting it to work provides a possible explanation for public concern about the collection and use of personal data. Survey after survey, including my own (discussed below), has revealed a high level of concern about the commercial collection and use of personal information online. For example, a 2012 Pew study in the United States revealed that a majority (65%) of people who use search engines did not approve of the use of behavioral data to customize search results, and that more than two-thirds of all Internet users (68%) did not approve of targeted advertising based on behavioral tracking (Purcell, Brenner, & Rainie, 2012). Another nationwide U.S. survey found that 66% of respondents opposed ad targeting based on tracking users' activities (Turow, King, Hoofnagle, Bleakley, & Hennessy, 2009). In a U.S. study of public reaction to proposed "do not track" legislation, 60% of respondents said they would opt out of online tracking, given the choice. My own nationwide survey in Australia revealed strong support for do-not-track legislation (95% in favor). Well over half of the respondents (56%) opposed customized advertising based on tracking, and 59% felt Web sites collect too much information about users. People's continuing use, despite their stated concerns, of services that collect and use their personal information is framed sometimes as a "paradox" (e.g., Norberg, Horne, & Horne, 2007), and sometimes as evidence that people do not really care as much as the research indicates (e.g., Oppmann, 2010). Based on early results of qualitative research on privacy concerns, this article offers an alternative explanation: that people operate within structured power relations that they dislike but feel powerless to contest. On a somewhat more speculative level, I suggest that there is an emerging understanding on the part of users that the asymmetry and opacity of a "big data divide" augurs an era of powerful but undetectable and un-anticipatable forms of data mining, contributing to their concern about potential downsides of the digital surveillance economy. This asymmetry runs deep, insofar as it privileges a form of knowledge available only to those with access to costly resources and technologies over the types of knowledge and information access that underwrite the "empowering" and democratizing promise of the Internet.[2]

Theory's End?

In a much discussed *Wired* magazine article, Chris Anderson (2008) claimed that the era of big data (which he called the "petabyte age") portended the "end of theory"—that is, the coming irrelevance of model-based understandings of the world. As he put it,

> This is a world where massive amounts of data and applied mathematics replace every other tool that might be brought to bear. Out with every theory of human behavior, from linguistics to sociology. Forget taxonomy, ontology, and psychology . . . With enough data, the numbers speak for themselves. (Anderson, 2008, para. 8)

This sweeping, manifesto-like claim invites qualification: Surely, statistical models remain necessary for developing algorithms, and other sorts of models are needed to shape the use of the information generated by increasingly loquacious data. Data scientists have emphasized the importance of domain-specific expertise in assessing the data that gets fed into mining algorithms and shaping the questions that might be put to the data. As McCue (2007) stated in her primer on data mining and predictive analytics, "domain expertise is used to evaluate the inputs, guide the process, and evaluate the end products within the context of value and validity" (p. 22). Indeed, the term *domain expert* emerges

[2] For a good overview of the celebratory, democratizing rhetoric surrounding the reception of the Internet, see Mosco (2004).

against the background of data mining's convergent character to redress both the fact that it is, in a sense, content-agnostic, and the resulting tendency "to treat data analysis as a strictly technical exercise" (Berry & Linoff, 2001, p. 44).

Moreover, the claim that numbers "speak for themselves" overlooks the broader context of the conversation around them (boyd and Crawford met this claim with "a resounding 'no'" [2011, p. 4]). Patterns may emerge from the data, but their relevance or usefulness depends heavily on the questions they address, which in turn depend on who is asking. One thing data cannot do is set the agenda. Anderson's version of big data is an instrumental one abstracted from broader issues of values and goals (questions of social justice, democratic commitments, etc.)—the very issues that the existing bodies of theory sidelined by Anderson are needed to address. Anderson's article is simply a quick-hit magazine piece, but its failure to consider the larger context in which large, for-profit entities (and even governments, which, it turns out, piggyback on these databases) collect, own, and control the data is telling nonetheless. More bluntly, sidelining the broader question of context and values effectively exempts the question of the best uses of the data from the reach of theory and models, leaving it to the imperatives of those with access to the databases. This is the real import of the "end of theory" claim.

With these qualifications in mind, the substance of Anderson's claim is more narrowly interpretable: data mining has the ability to generate actionable information that is both unpredictable and inexplicable (neither needing nor generating an underlying explanatory model). For example, the era of data mining and "micro-targeting" has renewed the salience of a bit of political wisdom discovered early in the 1970s by Republican political consultants in the United States: "Mercury owners were far more likely to vote Republican than owners of any other kind of automobile" (Gertner, 2004, para. 12). As one consultant put it, "We never had the money or the technology to make anything of it . . . but of course they do now" (ibid.).

This kind of inductive correlation, which is relatively easy to generate through data mining, provides predictive power and actionable information but little in the way of explanation. Meanwhile, those interested in using the information for electioneering purposes do not particularly care about any underlying explanation, should there be one. As Anderson (2008, para. 8) pointed out, "Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity."

A defining attribute of this kind of knowledge is the replacement of explanation and causation with correlation and prediction. What is "known" is not an underlying cause or explanation but rather a set of probabilistic predictions. Data mining promises to unearth increasingly unpredictable (in the sense of not being readily anticipatable) and otherwise indiscernible patterns by sorting through much larger data sets—indeed, the goal of data mining is to detect patterns that are not intuitively available to the unaided human eye or mind. That is, the goal is, by definition, to extract non-predictable patterns that emerge only via automated processing of data sets that are too large to make sense of otherwise. As one data-mining textbook observed, "as the world grows in complexity, overwhelming us with the data it generates, data mining becomes our only hope for elucidating the patterns that underlie it" (Chakrabarti, 2009, p. 32). Perhaps unsurprisingly, considering commercial databases' central role in its development, the goals of data mining are often—although not exclusively—portrayed in terms of competitive advantage. "Data mining is defined as the process of discovering patterns in data. The process must be automatic or (more usually) semiautomatic. The patterns discovered must be meaningful in that they lead to some advantage, usually an economic advantage" (Chakrabarti, 2009, p. 27). But numerous other types of advantages are conceivable. McCue's (2007) textbook on predictive policing frames the goal in terms of national security and military advantage: "If knowledge is power, then foreknowledge [via predictive analytics] can be seen as battlespace dominance or supremacy" (p. 48). MIT's big data guru Alex Pentland, who coined the term "reality mining" to describe the breadth and depth of new forms of data capture, anticipates that insights gleaned from the database will help create a more healthy, secure, and efficient world for all:

> For society, the hope is that we can use this new in-depth understanding of individual behaviour to increase the efficiency and responsiveness of industries and governments. For individuals, the attraction is the possibility of a world where everything is arranged for your convenience—your health checkup is magically scheduled just as you begin to get sick, the bus comes just as you get to the bus stop, and there is never a line of waiting people at city hall. (Pentland, 2009, p. 79)

Other benefits could involve new forms of transparency that make various kinds of public records available so as to hold public officials and private entities more accountable.

But even these salutary scenarios belie the "leveling" promise of networked digital technology. The era of big data mining concentrates a particular technique for generating actionable information (to be used for good or ill) in only a few hands, for the specific purpose of gaining some kind of advantage.[3] Tellingly, it posits a form of knowing that allegedly renders obsolete or outdated the very model of Internet empowerment that was supposed to help hold entrenched forms of power accountable by increasing access to forms of knowledge that allowed people to understand the world around them.[4] This is the thrust of Anderson's account of "the end of theory": that understanding the world through the careful, judicious, and informed study of available information is, for a growing range of applications, obsolete in the petabyte era, which promises to unearth powerfully useful patterns from bodies of information that are too large for a single person or group of people to make sense of. At the very moment that the new technology enhances access to traditional forms of understanding and evidence, they are treated as ostensibly outdated.

Even if Anderson is overstating the case and understanding remains an important aspect of knowledge acquisition in the digital era, the point remains: The few will have access to useful forms of "knowledge" that provide an advantage of some kind and that are not just unavailable to the vast majority but *incomprehensible*, in the sense described by Weinberger (2011). This knowledge is unpredictable and inexplicable in the conventional sense (as in the Mercury example: a correlation without an underlying

[3] Big data should not be understood as a static concept, for as more people gain access to data-mining technology, still "bigger" data will remain beyond their reach, available only to those with the resources to support the latest technology and the largest databases.

[4] For a discussion of the promise of Internet empowerment, see Andrejevic (2007, pp. 15–21).

explanation) and therefore opaque to those without access to the database. Thus, individual users have no way to anticipate fully how information about them might prove salient for particular forms of decision making, including, for example, whether they might be considered a security risk, a good or bad job prospect, a credit risk, or more or less likely to drop out of school.. Consider, for instance, the finding that "people who fill out online job applications using browsers that did not come with the computer . . . but had to be deliberately installed (like Firefox or Google's Chrome) perform better and change jobs less often" ("Robot Recruiters," 2013, para. 2). The finding is unexplained and unlikely to be anticipated by the applicants themselves, but it can significantly affect their lives nevertheless. As this example suggests, the forms of social sorting associated with big data mining will range far beyond the marketing realm, feeding into the decision-making processes of those with access to the information it provides and thereby allowing them to affect the life chances of others in increasingly opaque but significant ways.

Whereas it may still be possible to intuitively grasp the link between, for example, a particular brand of car and a political preference, the promise of data mining is to unearth correlations beyond the realm of such imagining. Reverse engineering an algorithmic determination can require as much expertise as generating it in the first place, and the results may have no direct explanatory power. When correlation displaces causality or explanation, the goal is to accumulate as comprehensive and varied a database as possible to generate truly surprising, non-intuitive results. Perhaps a particular combination of eating habits, weather patterns, and geographic location correlates with a tendency to perform poorly in a particular job or susceptibility to a chronic illness that threatens employability. There may not be any underlying explanation beyond the pattern itself.

The basis for the kind of sorting envisioned via big data mining is likely to become increasingly obscure in direct proportion to the size and scope of the available data and the sophistication of the techniques used to mine it. At a recent meeting of the Organisation for Economic Co-operation and Development, one participant observed that data mining entails the loss of "a degree of transparency in why computers make the decisions they do" (Cukier, 2013, para. 6). According to the participant, who is CEO of a data-mining company:

> There are machines that learn, that are able to make connections that are much, much finer than you can see and they can calibrate connections between tons and tons of different facets of information, so that there is no way you as a human can understand fully what is going on there. (J. Haesler, personal communication, February 26, 2013)

To note these characteristics of data mining is not to discount the potential benefits of its anticipated benevolent uses. Yet the shadow of rationalization betokens asymmetrical control: a world in which people are sorted at important life moments according to genetic, demographic, geo-locational, and previously unanticipated types of data in ways that remain opaque and inaccessible to those who are affected. In some instances, this is surely desirable: when, for example, a medical intervention is triggered just in time to avoid more severe complications. At the same time, it is easy to imagine ways in which this type of pre-emptive modelling—what William Bogard (1996, p. 1) has called "the simulation of surveillance"—can be abused. Imagine, for example a world in which private health insurers mine client data in an attempt to cancel coverage just in time to avoid having to cover major medical expenses.

What People Talk About When They Talk About Privacy

The apparent contradictions in public attitudes toward personal-data collection resolve somewhat when viewed against this account of the big data divide and its defining attributes. Those who judge people solely by their actions may conclude, for example, that, "the average American finds a very healthy acceptable balance between privacy and convenience, they give up some privacy and get a lot of convenience" (Oppmann, 2010, para. 11). This framing of the exchange assumes people are aware of the terms of the trade-off and it construes acquiescence to pre-structured terms of access as tantamount to a ready embrace of those terms. On closer examination, such assumptions fall short. The notion of informed consent is a vexed one in the online context, partly because few people read the terms of use they agree to upon joining or signing in. Research indicates that the vast majority of users only skim privacy policies or skip them altogether (e.g., "Regulators Demand," 2009; Turow, Mulligan, & Hoofnagle, 2007), a fact that might be taken as evidence that people do not care about privacy, despite high levels of stated concern and the proliferation of technologies for data capture. A more plausible explanation, based on my research on collection and use of personal information in Australia, is a perceived lack of options combined with lack of knowledge about possible uses of personal information and the absence of any *discernible* negative impact of these uses (e.g., job applicants are likely unaware that their choice of browser might decide whether they are hired).

Particularly striking in my research has been respondents' expressed sense of powerlessness vis-á-vis the arrangements that structure the collection and use of personal information. Despite the persistent focus on privacy issues in both academic research and popular press coverage, privacy arguably takes a backseat to an underlying sense of powerlessness. As one focus group respondent said (eliciting expressions of general assent in the group), "My biggest thing from loss of privacy isn't about other people knowing information about you but kind of being forced or bribed to share your information" (female, 22). In other words, Google may be misapprehending users' concerns when it defends its data scanning practices with the assurance that "no humans read your email or Google account information" (Byers, 2013, para. 6). Users' concerns are over the very fact that it collects this information for allegedly powerful uses that are not fully understood.

The focus group was one of three devoted to discussing the results of a nationwide telephone survey of 1,100 people about Australians' attitudes toward the collection and use of their personal information.[5] The survey results paralleled research in other countries indicating a high level of concern

[5] These survey findings are based on a national telephone survey conducted with *N* = 1,106 adults across Australia between November 17 and December 14, 2011. Managed by the Social Research Centre in Melbourne, the project sourced respondents through random-digit phone number generation for landlines and mobile phones. The final sample consisted of 642 surveys taken via landline numbers and 464 taken via mobile numbers. Reported data were proportionally weighted to adjust for design (chance of selection), contact opportunities (mobile only, landline, or both), and demographics (gender, age, education, and state). A complete summary of the findings and methodology is available online at www.cccs.uq.edu.au/personal-information-project. The survey was followed up by an ongoing series of interviews and focus group discussions. As of this writing, 27 structured interviews were conducted at

about the collection and use of personal information: 59% of respondents said websites collect too much information about people.[6] They also revealed a very high level of support for stricter controls on information collection, including a do-not-track option (92% support), a requirement to delete personal data upon request (96% support), and real-time notification of tracking (95% support).[7] Well over half of the respondents (56%) said they opposed customized advertising based on tracking. The survey results also indicated that people are palpably aware that they know little about how their information is used: 73% of respondents said they needed to know more about the ways websites collect and use their information.[8]

These findings represent a particular type of "big data divide," not between researchers with access to the data and those without, but between sorters and "sortees"—that is, not between those who comprehend the correlations and those who do not, but between those who are able to extract and use un-anticipatable and inexplicable (as described above) findings and those who find their lives affected by the resulting decisions. This formulation can aid consideration of the ways that the post-survey findings from the follow-up focus groups challenge the dominant framing of issues in contemporary discussions of privacy. One repeatedly mobilized frame is perhaps best summed up by Eric Schmidt's notorious observation: "If you have something that you don't want anyone to know, maybe you shouldn't be doing it in the first place" (and his subsequent, related assurance that "if you don't have anything to hide, you have nothing to fear") (Bradley, 2012, para. 3; "Google CEO on Privacy," 2010, para. 1). Gmail's role in

three sites across Australia (Melbourne, Sydney, and Brisbane). Recruited randomly in public spaces for 30- to 45-minute discussions, respondents were screened to include only experienced Internet users. The preliminary interview sample skews young and female, consisting of 19 female respondents and 8 male respondents, all between the ages of 19 and 37. As the project develops, respondents will be selected to counter this skew. Focus group participants were similarly recruited in public spaces at the three research sites and received a $20 iTunes gift card to participate in a 50-minute group discussion. A similar skew applies to the focus group participants: 16 women and 6 men, ages 20–31. The focus group structure was tested on students in an undergraduate seminar, and some of their comments were included.

[6] Actual survey question: "Thinking now about the personal information gathered by ONLINE companies about their consumers, would you say they gather too much, about the right amount or not enough information?"

[7] Survey questions:

Do you think:

1. There should be a law that requires websites and advertising companies to delete all stored information about an individual, if requested to do so?
2. There should be a law requiring Web sites and applications to provide a "do-not-track" option that would prevent them from gathering information about people?
3. There should be a law requiring companies to notify people at the time when they collect data about them online?

[8] Survey question: "How would you describe your understanding of the ways in which companies collect and use the information they gather about people online? Do you feel that you already know as much as you need to know about what companies do in this regard or need to know more about what companies do in this regard?"

the downfall of U.S. General Petraeus lent these remarks a certain salience, but they do not reflect the concerns of most respondents, who emphasize that whereas much of the information they share (and that is collected about them) is mundane, they still dislike being compelled to share it.

One focus group participant, for example, used just one word in response to concerns about the collection and use of personal information: "powerless." Several others in the seven-person discussion group indicated they had also written down "powerless" in their discussion notes. Another participant chimed in,

> You just feel out of control of what people can know about you. It reinforces what the world has come to. I know that in general you share a lot more than you used to, we're used to that. But then I still feel powerless. (male, 21)

The focus group participants repeatedly invoked a feeling of asymmetry that paralleled this sense of powerlessness. As one respondent maintained in a conversation touching on e-mail and social networking: "It's not fair, it's not transparent. It's funny because Facebook is supposed to be all about transparency, and they're the ones who aren't transparent at all" (female, 31). Another respondent explained how this sense of powerlessness influenced her decision not to read privacy policies:

> I just click agree, because what else can I do? I think that frustration sometimes just translates into: "I won't even think about it, because what can I do?" It [Facebook] becomes part of how you connect with people. It's really useful for your career, for your choices in life. It doesn't mean you can't live without it, but living with it becomes important. (female, 29)

Most respondents expressed concern and frustration with the online collection of information about them, but a few said they were unconcerned because there was nothing they could do about it. As one focus group participant put it,

> I don't see it as a threat . . . probably because I don't know much about it. . . . I can't see it affecting me in my everyday life but if you tell me about online privacy . . . then I'll be thinking about it all the time. I'm better off not knowing about it in the first place. (male, 22)

Significantly, even respondents who expressed concern about data collection were vague about actual, perceived, or anticipated harm. When pressed on the concrete content of their concern, respondents tended to fall back, not particularly confidently, on a familiar litany of well-covered privacy concerns: the threat of identity theft or fraud and distaste for data-driven target marketing, which some equated with a limiting form of stereotyping. As one respondent put it,

> It kind of pushes you and says who you are and what you'd like. . . . At the end of the day you do have your right to choose, but this kind of enforces an idea of what you

> should be choosing and limits what it is you can choose from. . . . You either work within that stereotype or they will create another stereotype for you. (female, 25)

Overall, concern about actual harms came across less vociferously than did frustration over a sense of powerlessness in the face of increasingly sophisticated and comprehensive forms of data collection and mining. Focus group participants generally agreed with responses emphasizing that this sense of powerlessness extended to their lack of knowledge over how personal data might be used. As one respondent admitted, “We really don’t know where things collected about us go—we don’t understand how they interact in such a complex environment” (female, 22). Interview respondents and focus group participants alike noted the seemingly endless appetite for personal data: “It’s not just what you want—it’s where you are, what you do. It’s everything. You’re not free any more. You’re just a slave of these companies” (male, 22). This may come across as hyperbolic, but nonetheless noteworthy is the stark contrast between this response and the rhetoric of freedom, empowerment and convenience that has long underpinned the promotion of the online economy. The contrast highlights the challenge posed by the power asymmetries ushered in by big data mining.

Dimensions of the Divide

This article’s analysis suggests that the sense of powerlessness expressed by the focus group respondents operates in at least two dimensions: that of ownership and control over information and communication resources, and that of different approaches to knowledge-based decision making. People are palpably aware that powerful commercial interests shape the terms of access that extract information from them: they must choose either to accept the terms on offer or to go without resources that in many ways are treated as utilities of increasing importance in their personal and professional lives. However—and this is an interpretive, speculative claim—the very vagueness (but vociferousness) of their concerns about information collection may reflect the structural gap in the big data divide: the fact that users of big data rely on the unanticipatable and un-intuitable character of their findings. This vagueness, then, is not necessarily an artifact of laziness or ignorance due to users’ failure to educate themselves about the technologies they use (or to read legalistic, vague privacy policies) but may be a defining characteristic of the data collection strategies to which they are subjected. People can hardly be expected to imagine that, for example, their use of a particular browser might render them more or less desirable to employers, or to envision all the possible patterns generated by the complex interplay of thousands of variables about millions of people, patterns that data-mining strategies have explicitly relegated to the realm of “too big to know or predict.” As one respondent put it, “you end up accepting having no privacy without knowing the consequences” (male, 32).

If, as Helen Nissenbaum (2009) has compellingly argued, privacy is contextual (because of established expectations associated with particular information-collection contexts), then the big data era challenges people to develop “contextual” norms for the use of data whose uses can be radically, unpredictably decontextualized. Thanks to the proliferation of monitoring technologies (license plate readers, smart cameras, drones, RFID scanners, audio sensors, etc.), data scraping continues to extend its reach both online and off, so fewer places and activities are likely to be exempt from the logic of the big data divide, whereby people are separated from their data and excluded from the process of putting it to use. Overcoming the digital divide means exacerbating the big data divide. Greater access to and facility in the use of smartphones and networked laptops, tablets, and computers of one kind or another means more data to store, sort, and mine. More comprehensive forms of data mining promise to serve a growing variety of decision-making, forecasting, and sorting operations. Whereas many of the applications mentioned here are only in their infancy, the pace of change urges the individual to anticipate the social, cultural, and political consequences now. Given the impossibility of adjusting expectations to anticipate correlations that are by definition unpredictable, people face the daunting prospect of finding ways to limit the reach and opacity of emerging forms of social sorting and discrimination. This is the challenge of the big data era.

References

Anderson, C. (2008, June 23). The end of theory: The data deluge makes the scientific method obsolete. *Wired Magazine*, *16*(7). Retrieved from http://www.wired.com/science/discoveries/magazine/16-07/pb_theory

Andrejevic, M. (2007). *iSpy: Surveillance and power in the interactive era*. Lawrence: University of Kansas Press.

Berry, M., & Linoff, G. (2001). *Mastering data mining: The art and science of customer relationship management*. Hoboken, NJ: John Wiley & Sons.

Bogard, W. (1996). *The simulation of surveillance: Hypercontrol in telematic societies*. New York, NY: Cambridge University Press.

boyd, d., & Crawford, K. (2011, September). Six provocations for big data. Presentation at *A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society*, Oxford Internet Institute, Oxford University, Oxford, UK. Available at SSRN http://ssrn.com/abstract=1926431 or http://dx.doi.org/10.2139/ssrn.1926431

Bradley, T. (2012, March 24). Hey Employers—My Facebook Password Is None of Your Business. *PCWorld.* Retrieved from http://www.pcworld.com/article/252514/hey_employers_my_facebook_password_is_none_of_your_business.html

Byers, A. (2013, February 7). Microsoft hits Google email privacy. *Politico*. Retrieved from http://www.politico.com/story/2013/02/microsoft-renews-google-attack-on-email-privacy-87302.html#ixzz2LW6dnEV4

Chakrabarti, S. (2009). *Data mining: Know it all*. Burlington, MA: Morgan Kaufmann.

Cukier, K. N. (2013, February 18). The thing, and not the thing. *The Economist*. Retrieved from http://www.economist.com/blogs/graphicdetail/2013/02/elusive-big-data

Gandy, O. H., Jr. (1993). *The panoptic sort: A political economy of personal information. Critical studies in communication and in the cultural industries*. Boulder, CO: Westview Press.

Gandy, O.H., Jr. (2005, October). If it weren't for bad luck. *14th Annual Walter and Lee Annenberg Distinguished Lecture*. Annenberg School for Communication, University of Pennsylvania, Philadelphia, PA. Retrieved from http://www.asc.upenn.edu/usr/ogandy/Annenberg%20Lecture.pdf

Gates, B. (1995). *The road ahead*. New York, NY: Penguin Books.

Gertner, J. (2004, February 15). The very, very personal is the political. *The New York Times Magazine*. Retrieved from http://www.nytimes.com/2004/02/15/magazine/15VOTERS.html?pagewanted=all

Google CEO on privacy (VIDEO): "If you have something you don't want anyone to know, maybe you shouldn't be doing it." (2010, March 10). *The Huffington Post.* Retrieved from http://www.huffingtonpost.com/2009/12/07/google-ceo-on-privacy-if_n_383105.html

IBM. (2012). Bringing big data to the enterprise*.* Retrieved from http://www-01.ibm.com/software/in/data/bigdata

Improve health care: Win $3 million*.* (2012). *Heritage Provider Network: Health Prize*. Retrieved from http://www.heritagehealthprize.com/c/hhp

Kakutani, M. (2013, June 10). Watched by the Web: Surveillance is reborn. *The New York Times*. Retrieved from http://www.nytimes.com/2013/06/11/books/big-data-by-viktor-mayer-schonberger-and-kenneth-cukier.html

Katz, I. (2012, April 18). Tim Berners-Lee: Demand your data from Google and Facebook. *The Guardian* [London]. Retrieved from http://www.guardian.co.uk/technology/2012/apr/18/tim-berners-lee-google-facebook

Lyon, D. (1994). *The electronic eye: The rise of surveillance society*. Minneapolis: University of Minnesota Press.

Lyon, D. (Ed.). (2002). *Surveillance as social sorting: Privacy, risk and automated discrimination*. New York, NY: Routledge.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Boston, MA and New York, NY: Eamon Dolan/Houghton Mifflin Harcourt.

McCue, C. (2007). *Data mining and predictive analysis: Intelligence gathering and crime analysis*. New York, NY: Butterworth-Heinemann.

Mosco, V. (2004). *The digital sublime: Myth, power, and cyberspace*. Cambridge, MA: MIT Press.

Nissenbaum, H. (2009). *Privacy in context: Technology, policy, and the integrity of social life*. Palo Alto, CA: Stanford Law Books.

Negroponte, N. (1996). *Being digital*. New York, NY: Vintage.

Nolan, R. (2012, February 21). Behind the cover story: How much does Target know? *The New York Times Magazine*. Retrieved from http://6thfloor.blogs.nytimes.com/2012/02/21/behind-the-cover-story-how-much-does-target-know

Norberg, P. A., Horne, D. R., & Horne, D. A. (2007). The privacy paradox: Personal information disclosure intentions versus behaviors. *Journal of Consumer Affairs, 41*(1), 100–126.

Oppmann, P. (2010, April 14). In a digital world we trade privacy for convenience. *CNN.com*. Retrieved from http://www.cnn.com/2010/TECH/04/14/oppmann.off.the.grid/index.html

Pentland, A. (2009). Reality mining of mobile communications: Toward a new deal on data. In S. Dutta & R. Mia (Eds.), *The global information technology information report 2008–2009: Mobility in a networked world* (pp. 75–80). Basingstoke, UK: Palgrave Macmillan.

Purcell, K., Brenner, J., & Rainie, L. (2012, March 9). Search Engine Use, 2012. *Pew Internet and American Life Project, 9*. Retrieved from http://pewinternet.org/Reports/2012/Search-Engine-Use-2012.aspx

Raley, R. (2013). Dataveillance and counterveillance. In L. Gitelman (Ed.), *Raw data is an oxymoron* (pp. 121–146). Cambridge, MA: MIT Press.

Regulators demand clearer privacy policies. (2009, February 16). *Out-Law.com*. Retrieved from http://www.out-law.com/page-9795

Robot recruiters. (2013, April 6). *The Economist.* Retrieved from http://www.economist.com/news/business/21575820-how-software-helps-firms-hire-workers-more-efficiently-robot-recruiters

Turow, J., King, J., Hoofnagle, C. J., Bleakley, A., & Hennessy, M. (2009, September 29). Americans reject tailored advertising and three activities that enable it. Retrieved from http://ssrn.com/abstract=1478214 or http://dx.doi.org/10.2139/ssrn.1478214

Turow, J., Mulligan, D., & Hoofnagle, C. (2007, October 31). Research report: Consumers fundamentally misunderstand the online advertising marketplace. Retrieved from http://www.law.berkeley.edu/files/annenberg_samuelson_advertising.pdf

Weinberger, D. (2011). *Too big to know: Rethinking knowledge now that the facts aren't the facts, experts are everywhere, and the smartest person in the room is the room*. New York, NY: Basic Books.