Winter, Jenifer Sunrise

Conference Paper

Big data analytics, the social graph, and unjust algorithmic discrimination: Tensions between privacy and open data

2015 Regional Conference of the International Telecommunications Society (ITS): "The Intelligent World: Realizing Hopes, Overcoming Challenges", Los Angeles, USA, 25th-28th October, 2015

Provided in Cooperation with:

International Telecommunications Society (ITS)

Suggested Citation: Winter, Jenifer Sunrise (2015) : Big data analytics, the social graph, and unjust algorithmic discrimination: Tensions between privacy and open data, 2015 Regional Conference of the International Telecommunications Society (ITS): "The Intelligent World: Realizing Hopes, Overcoming Challenges", Los Angeles, USA, 25th-28th October, 2015, International Telecommunications Society (ITS), Calgary

This Version is available at:
https://hdl.handle.net/10419/146313

Big data analytics, the social graph, and unjust algorithmic discrimination: Tensions between privacy and open data

Jenifer Sunrise Winter
University of Hawai‘i at Mānoa
jwinter[at]hawaii.edu

Abstract

This paper examines threats to privacy and anonymity accompanying underlying technical changes related to the Internet of Things and big data analytics. Governmental and corporate entities have focused on creating sophisticated graphs of citizens' social media connections and aggregating these data with other sources, such as physical location, biometric data, public records, and online search habits. The increased instrumentation and tracking of natural and social processes and resulting availability of real-time user data has enabled sophisticated user modeling and new algorithms to mine, model, and personalize these data. Even when data seem innocuous or have been anonymized/de-identified, analysis can lead to inferences, re-identification, and subsequent informational harms. The algorithms and machine-to-machine (M2M) communication employed in big data analytics may disadvantage certain individuals or groups. On the other hand, big data analytics has immense potential to enhance public welfare – leading to fairer hiring decisions, government transparency, energy conservation, participatory governance, and substantial advances in medical research and care. This paper first addresses how these developments may create unjust power differentials used by one group to diminish the opportunities of another, threaten to destroy anonymity when engaging in public affairs, and hinder public participation in democratic discourse. Legal and policy barriers to citizen privacy protections and the tension between privacy and open data (for the public good) are then identified and discussed. This paper intends to stimulate a deeper policy debate about how to protect citizens from informational harms and unjust discrimination while opening public access to large data sets required for participatory governance, health advances, and environmental protections.

Keywords: Surveillance, Internet of Things, big data analytics, de-anonymization, privacy, data discrimination, participatory governance, open data

Introduction: The evolving Internet

The Internet has become inextricably meshed with nearly every sector of society, increasing its complexity and creating myriad opportunities and challenges. In its present form as an interconnected network of networks spanning, albeit unevenly, much of the globe, it connects nearly three billion users (International Telecommunication Union, 2014). But while most people think of the Internet as something virtual that is accessed via devices such as a

smartphone or mobile personal computer, it is rapidly becoming part of the natural world itself. As a growing number of ordinary objects are redesigned to include digital sensors, computing power, and communication capabilities – and as new objects (and processes) become part of the Internet, the physical and virtual worlds are merging. This paper outlines key technical changes related to the Internet, including the social semantic web and linked data, increased instrumentation of natural and social processes, and big data analytics. Emerging problems of concern to policymakers are outlined, such as unjust algorithmic discrimination and erosion of anonymity required for democratic discourse. At the same time, exploiting large, open data sets, can substantially benefit the public good. Management of these tensions has emerged as a great challenge for policymakers.

Key technical features of these changes include the increased instrumentation, tracking, and measurement of both natural and social processes. Current research agendas focus on the Internet of Things (IoT), what Gartner (2014) describes as the IoT ecosystem – networks of physical objects embedded with the ability to sense (and sometimes act upon) their environment, as well as related communication, applications, and data analysis. This can also be imagined as a "backbone for ubiquitous computing, enabling smart environments to recognize and identify objects, and retrieve information from the Internet to facilitate their adaptive functionality" (Weber & Weber, 2010, p. 1). While many details in the areas of network standards, semantic specifications, and the 'Things' themselves (i.e., objects embedded in the natural world that are connected to a communication device, such as an RFID chip, a sensor, or an actuator) are still in progress, the IoT is expected to automate the exchange of goods and services and provide a wealth of new data for strategic business and policy decisions. There is no single definition for the IoT – rather, it is a research paradigm encompassing numerous developments intended to

bring everyday objects online and allow them to be uniquely identified over the Internet (Uckelmann, Harrison & Michahelles, 2010). Utilizing a variety of short-range wireless technologies, such as radio frequency identification (RFID), near field communication (NFC), or wireless sensor networks (WSNs), related research by corporations such as Cisco, HP, and IBM – and government ICT strategies, such as those employed by China and the European Union, have focused on fostering this shared vision. By the year 2020, as many as 75 billion "Things" may be connected to the Internet, with room for massive growth (Proffitt, 2013).

> With so much technology and networking available at such low cost, what wouldn't you enhance? What wouldn't you connect? What information wouldn't you mine for insight? What service wouldn't you provide a customer, a citizen, a student or a patient? (IBM, 2008, para. 11)

As tens of billions of everyday objects become equipped with sensors, many new types of data will be collected. Sources are numerous, including biometric data, such as gait or typing pattern, the unique communication signatures of biomedical devices such as pacemakers or insulin pumps, and data from smart homes and appliances, and cars. Sensors could be embedded nearly anywhere, gathering almost any type of data.

Supporting this IoT ecosystem is the growth of machine-readable standards enabling the representation of objects and individuals – the so-called social semantic web (Berners-Lee, 2000; Breslin, Passant & Decker, 2009). Linked data, a method of connecting structured data so it can be connected and referenced via semantic queries that acknowledge contextual data, is enabling a massive, global datasphere that can be searched, and acted upon, by machine-to-machine (M2M) transactions, without human intervention (Heath & Bizer, 2011). O'Reilly and Battelle (2009) note that the Web already includes a "hodgepodge of sensor data," collected from devices such

as mobile phones, that enables machine learning. A goal of instrumentation efforts like the IoT is to provide computers

> with their own means of gathering information, so they can see, hear and smell the world for themselves, in all its random glory. RFID and sensor technology enable computers to observe, identify and understand the world – without the limitations of human-entered data. (Ashton, 2009, para. 5)

With an estimated Zettabyte (1,000,000,000,000,000,000,000 bytes) flowing over the global Internet each year by 2016 (Cisco, 2014), as much data will flow this year as in the entire history of the Internet since its creation in 1969. Increasingly, these are M2M-handled data, such as those used in automobiles and various asset tracking systems. These include data gathered via smart meters, various smart city initiatives, and smart appliances in homes. In addition to information that an average consumer might expect such devices to gather, smart appliances such as televisions or gaming consoles may employ audio recording or facial recognition features that observe when one is watching a particular television program or playing a certain game, along with many other personal details, such as the presence of others in the room. Data from public records are federated with other traces such as location, online search habits, social networks, and affect monitoring via facial recognition (Keller, 2011; Winter, 2016).

New management tools are being created to manage these large and increasingly complex data sets and to provide sophisticated modeling of users (Jaimes, 2010). Over the past decade, there has been growth in parallel and distributed computing systems to handle these data, and we have moved from relational database management systems to graphing databases that enable prediction of human behaviors via awareness of relationships, affiliations, and social influence. Whereas once only a limited number of fields could be cross-referenced, new tools such as Hadoop (an open source version of MapReduce) enable additional data to be included in

analyses. This means that many data that were previously unavailable, or not feasible to analyze due to technical limitations or cost, can now be included in predictive analytics.

Whether these new data sets are gathered illicitly (e.g., via hacking) or via legally-acceptable means, data analytics has clearly moved from a function of an information technology support department to become part of core business and operational functions (Davenport, Barth & Bean, 2012). As data become an increasingly valuable commodity, there is a need for further study of the information markets made possible through these changes and the effect this will have on citizen welfare:

> Policy-makers need to address the question of whether there is need for regulation. Market players want to know the ground on which they compete. Finally, in order to estimate the impact of Ubiquitous Computing technology on the privacy of individuals, it is necessary to understand the motivations and incentives of the parties that operate it. (Bauer, Fabian, Fischmann & Gürses, 2006, p.3)

Data for the public good

There is immense potential for data analytics to serve the public good by fostering government transparency, energy conservation, participatory governance, and substantial advances in medical research and care. Data analytics can also illuminate social injustices in order to correct them. Described as "data for good," Google used its own internal personnel data to identify patterns of discrimination and changed its internal promotion processes to be more inclusive of women (Kang, 2014), and the White House recently appointed its first U.S. Chief Data Scientist, with the aim of harnessing big data for the public good (Patil, 2015). Within the sciences, large data sets are probing new frontiers through exploratory analysis in astronomy and physics (e.g., the Large Hadron Collider). Medical data sets are being mined to create targeted

genomic therapy, increase patient care, and survival rates. The Committee on the Analysis of Massive Data et al. (2013) note that,

> [I]n general, the hope is that if massive data could be exploited effectively, science would extend its reach, and technology would become more adaptive, personalized, and robust. It is appealing to imagine, for example, a health-care system in which increasingly detailed data are maintained for each individual—including genomic, cellular, and environmental data—and in which such data can be combined with data from other individuals and with results from fundamental biological and medical research, so that optimized treatments can be designed for each individual (p. 12)

Open data describes a growing movement to make certain data freely available for citizen use. The goal is to release stored machine-readable data in order to allow third-party developers to create applications and data visualizations that will allow citizens to explore, and test, data. In terms of government data, the goal is to release (or create) valuable data sets in order to build public-facing tools that allow citizens to better understand government processes via increased transparency – building confidence in the process, and encouraging more citizens to take part. Notable examples include the World Bank's release of data to allow research on development and aid spending (http://data.worldbank.org/) and the United States government's open data portal (http://www.data.gov/). Boulton, Rawlins, Vallance and Walport (2011) note increasing pressures to share scientific data so that other scientists (in some cases citizen scientists) can validate findings or re-use data to advance science for public benefit.

There is a blurring boundary between corporate and government data in the open data ecosystem, with governments increasingly relying on data from commercial data aggregators, and aggregators scraping available government data for analysis. In some cases, select corporate data sets are being shared for open data initiatives. For example, energy companies and smart grid research consortia have started to release data to researchers or citizen hackers (St. John, 2014), and there are increasing demands to build application programming interfaces (APIs) to

access energy data that can be used to build apps (Tendril, 2012; Raftery, 2013). This so-called "green button" data can enable energy customers to monitor their usage data in an easy-to-understand format (Chopra, 2011). As discussed later in this paper, the hybrid origin of data being used to make public policy decisions raises many questions about data integrity, context, and the process of its construction.

Are open data personal data? Tensions between privacy and open data

While open data undoubtedly offer a great deal of individual and societal benefit, in most cases, the data in question have been anonymized, stripping them of personally-identifiable information. However, as data are aggregated, re-identification is possible. Certain types of data that are released beyond their original intent are often subject to regulatory requirements that are intended to protect personal data. However, data mining techniques, coupled with an abundance of data, have led to a number of instances of de-anonymization, or re-identification (Schwartz & Solove, 2011). The first well-publicized case was the identification of AOL user Thelma Arnold, also known as "user number 4417749." After AOL released over 20 million anonymous Web queries in 2006, journalists were able to identify Arnold based on her search requests (Barbaro & Zeller, 2006). More recently, researchers at Harvard were able to uniquely identify males who had shared personal DNA sequences on Internet genealogy forums based on publicly available data (Gymrek, McGuire, Golan, Halperin & Erlich, 2013). In another case, data from the Telecom Italia Big Data Challenge, comprised of vehicle location data in the city of Milan, was examined by Manfredi, Mir, Lu, and Sanchez (2014). They concluded that only a few data points were necessary to uniquely identify drivers, and that "there is no known way to anonymize

location data since spatio-temporal data is highly unique to individuals and robust to changes over extended periods of time" (p. 46). Thus, while

> Open data may not seem to be personal data on first glance especially when it is anonymised or aggregated… it may become personal data by combining it with other publicly available data or when it is deanonymised." (Kulk & van Loenen, 2012, p. 196)

While anonymity is not accepted as a universal value, in the United States and other democracies, it is an important aspect of political practice, as it can affect the behavior of those who participate in democratic discourse. Anonymity enables the freedom to express unpopular ideas or be critical of government without risking retaliation (Solove, 2011), and enables citizens to engage in meaningful activities in our everyday lives without inhibition. Its loss may chill valuable opinions and dissent. For example, growing awareness of surveillance may lead citizens to self-censor. The PEN American Center (2013), a group comprised of writers, including journalists, in the United States, found that its members were increasingly engaging in self-censorship after allegations about programs of mass surveillance of the activities of everyday citizens run by the National Security Agency and other security agencies. As concerns about privacy invasions and lack of anonymity mount, citizens' freedom of access to information and their ability to discuss issues relevant to democratic decision-making in their communities is limited.

Due to big data analytics, anonymity on the Web has essentially ceased to exist. Omnibus data aggregators are increasingly able to match real names to online pseudonyms (Angwin & Stecklow, 2010). Even "secure" techniques like Tor, anonymizing proxy software that is often used by journalists or others needing to conceal their identity as they traverse the surface or deep Web, have been revealed to be vulnerable to exploits. Further, by quantifying the minute

differences between users as they touch keys on computer keyboards, researchers are now able to uniquely identify an individual based on his or her typing: "the profiles act as a sort of digital fingerprint that can betray its owner's identity" (Goodin, 2015, para. 2). Via linked data, there has also been a trend towards online identity verification. For example, Google and Facebook attempt to link all online user profiles together via "real name" policies that uniquely identify users. Those who resist may be disadvantaged by being unable to access key services such as email. As boyd notes, "the people who most heavily rely on pseudonyms in online spaces are those who are most marginalized by systems of power" (boyd, 2011, para. 6).

Unjust algorithmic discrimination

A growing body of research addresses emerging privacy and civil liberties concerns related to big data, including unjust discrimination (e.g., Custers, 2013, boyd, Levy & Marwick, 2014). The longstanding practice of gathering (and sharing) customer data for use by advertising and marketing campaigns that increasingly "insert themselves unfiltered into their desired customers' domestic lives in ways that encourage consumers to accept surveillance and relationships tailored to their personal characteristics" (Turow, 2006, p. 295), coupled with social networking analysis, has enabled corporations to assign individuals to social groups. Individuals are then monitored based on these categories. Lyon (2002) describes this as "social sorting":

> Codes, usually processed by computers, sort out transactions, interaction, visits, calls and other activities; they are invisible doors that permit access to or exclude from participation in a multitude of events, experiences, and processes. The resulting classifications are designed to influence and to manage populations and persons thus directly and indirectly affecting the choices and chances of subjects. The gates and barriers that contain, channel, and sort populations have become virtual. (Lyon, 2002, p.13)

The emerging datasphere made possible by these technical and economic developments, and a lack of meaningful regulation to protect abuse of personal data in many contexts, has enabled the creation and sale of large data sets. Mining these reveals patterns that were previously not detectable. Data about a variety of daily tasks that seem trivial is increasingly being federated and used to reveal associations or behaviors, and these analyses and the decisions made based on them pose potential harms. As this happens, sensitive personal information or behaviors (e.g., political or health-related) may be revealed and used to discriminate when individuals seek housing, immigration eligibility, or employment (Winter, 2014). Prior to big data analytics and the ability to capture and store huge volumes of data, many transactions that seemed innocuous – one's movement throughout the day, items purchased at the store, television programs watched, "friends" added or looked at on social networks, or individuals communicated with or who were in close proximity the subject at various times, can all be used to make judgements that affect an individual and his or her life chances. For example, companies might offer different services, products, or prices to individuals based on their data profile (Turow, 2012).

As new forms of price differentiation emerge, insurance will no longer spread risk across a large group. Based on big data analytics, insurers are allocating risk differently (Upturn, 2014):

> A person's future health, like their driving behavior, can also be predicted based on personal tracking to set insurance prices. At an annual conference of actuaries, consultants from Deloitte explained that they can now use thousands of "non-traditional" third party data sources, such as consumer buying history, to predict a life insurance applicant's health status with an accuracy comparable to a medical exam. (p. 6)

This can lead to higher burdens (in the form of denial of coverage or higher costs) for those with certain medical conditions, or even DNA that indicates a higher-than-average probability of certain conditions manifesting. This discrimination can also lead to increased cost burdens for

healthy individuals who are profiled due to late-night driving or living in a low-income area. As Upturn (2014) notes, these groups are already populated by vulnerable social populations.

Another example is the use of big data analytics in the criminal justice system, where data-based risk assessment is now being considered for use *at sentencing*. To cut costs, the State of Pennsylvania, is considering a plan that would allow "some offenders considered low risk to get shorter prison sentences than they would otherwise or avoid incarceration entirely. Those deemed high risk could spend more time behind bars" (Barry-Jester, Casselman & Goldstein, 2015, para. 4).

Looking ahead, there may be instances where the monetary, or discriminatory, value of a data set outweighs an institution's concern with ethical or legal restrictions. For example, the National Institute of Standards and Technology observed that a potential, and unacceptable, use of data gathered via the smart grid might include customer energy usage data that revealed specific lifestyle information that could be used by insurers or a variety of commercial service providers (National Institute of Standards and Technology, 2010). Genetic information is another area of concern. In 2012, large personalized DNA database 23andMe released an API that allows authorized developers to build applications and tools to explore users' genetic data and look for matches among over one million genotyped customers (McMillan, 2012). In addition to aiding customers to learn more about their family origins, 23andMe provides health information, and participants can also opt-in to participate in medical research. On July 20, 2015, a program called Genetic Access Control was posted on GitHub, a high-traffic software repository hosting service (Offensive-computing, 2015). Using 23andMe's API, the creator(s) created software that enables "restrict[ed] access to your site based on traits including sex, ancestry, disease susceptibility

[sic], and arbitrary characteristics associated with single-nucleotide polymorphisms (SNPs) in a person's genotype" (para. 1).

While in the United States the *Genetic Information Nondiscrimination Act* is intended to prevent such abuse, the financial incentives for mining these data are immense, and big data analytics may enable non-protected "proxy" fields (i.e., other patterns or data-based evidence that is not explicitly protected) to be used instead. There are a patchwork of tailored domestic laws protecting collection of personal information and its use, these may be skirted by using other, non-protected information that correlates with the variable of interest. (This may also be possible in jurisdictions where personal data is substantially more protected, such as the European Union). For example, while banks may not deny a woman a home mortgage because she is pregnant, other information may be used to make such a decision. Barocas and Selbst (2016) provide a detailed analysis of how process-oriented civil rights laws (e.g., Title VII) cannot adequately address this disparate impact, since discrimination based on proxies often "discovers" patterns that merely reinforce existing social inequalities. They observe,

> Data mining can go wrong in any number of ways: by choosing a target variable that correlates to protected class more than others would, by injecting current or past prejudice into the decision about what makes a good training example, by choosing too small a feature set, or by not diving deep enough into each feature. Each of these potential errors is marked by two facts: an ex-post determination has been made that the overall outcome is unfair and at least one seemingly nondiscriminatory choice made by a data miner has created a disparate impact. Where data mining goes "right," the data miners could not have been any more accurate given the starting point of the process; it is that very accuracy, exposing an uneven distribution of the attributes that predict the target variable, that gives such a result its disparate impact. (p. 60)

While policymakers may desire to use data for the public good, "the networked nature of modern life can lead to very different outcomes for different groups of people—despite our aspirations to equal opportunity" (boyd, Levy & Marwick, 2014).

Conclusion: Challenges for policy makers related to the emerging IoT ecosystem

This paper described underlying technical changes related to the Internet of Things and big data analytics. The increased instrumentation and tracking of natural and social processes and resulting availability of real-time user data has enabled sophisticated user modeling and new algorithms to mine, model, and personalize these data. Even when data seem innocuous or have been anonymized/de-identified, analysis can lead to inferences, re-identification, and subsequent informational harms. The algorithms and M2M-based communication employed in big data analytics may disadvantage certain individuals or groups. This paper addresses how these developments may create unjust power differentials used by one group to diminish the opportunities of another, threaten to destroy anonymity when engaging in public affairs, and hinder public participation in democratic discourse. On the other hand, big data analytics has immense potential to enhance public welfare – leading to fairer hiring decisions, government transparency, energy conservation, participatory governance, and substantial advances in medical research and care.

There is need for deeper policy debate about how to protect citizens from informational harms and unjust discrimination while opening public access to large data sets required for advances in areas such as participatory governance, healthcare, and environmental protections. This section highlights several areas for further investigation by policymakers and researchers regarding opportunities and challenges of the emerging IoT ecosystem.

First, as big data analytics increasingly moves towards the core of business and governmental decision-making, it is important to thoroughly study both the information markets emerging from these changes and the effects they are having on public welfare (Bauer, Fabian,

Fischmann & Gürses, 2006). As data are collected and shared, we require assessments of the distribution of both benefits and harms resulting from big data analytics, as well as explorations about whether the resulting distributions conflict with general moral and political principles related to community values. For example, in the United States, the concepts of privacy and anonymity – although not explicitly defined or agreed upon – are seen as a basis for protection against informational harms and loss of personal autonomy, and as concepts supporting democratic values such as fairness, justice, and equality. Because regulatory interventions are often devised after harms occur, it is important to catalog these distributions.

A second area relates to data protections. The IoT ecosystem presents marked challenges for the regulation of personal data collection and sharing. First, it involves a wide variety of small, often invisible, objects to collect data. Even where opt-in consent is required, it may be difficult to know if this is being violated (Winter, 2015). Further, even where opt-in consent is required, it would be overwhelming at the interface level to implement such a scheme (e.g., imagine thousands of pop-ups at the interface level). This aspect clearly complicates regulatory or technical schemes that rely on consumer consent. Further, technical approaches in practice today to protect personal information are limited due to the scale and heterogeneity of the IoT. Therefore, encouraging further development and use of privacy enhancing technologies (PETs) that enable users to decide which information they wish to share or erase traces of their activities is an important first step. However, personal data protections must be introduced during the earliest stages of system development and be maintained (in secure fashion) throughout the lifecycle of use. Privacy by design (Cavoukian & Kursawe, 2012) is a design framework with principles that protect personal data while allowing meaningful collection and analysis for other uses. For example, in their case study of smart grid implementation, Cavoukian and Kursawe

conclude that energy utilities should first conduct privacy impact assessments and then only collect data required for primary business purposes. Additional data could be collected with users consent. Using the principles in the privacy by design framework, technologies bridging diverse goals and contexts can be designed with privacy at the core.

This also requires supporting regulation. Weber (2010) points out that existing regulatory approaches fail to account for the IoT ecosystem: "The nature of the IoT asks for a heterogeneous and differentiated legal framework that adequately takes into account the globality, verticality, ubiquity and technicity of the IoT." (p. 30). He acknowledges that neither industry self-regulation nor national laws will be sufficient. As Hildner (2006) observes, self-regulation that is not legally enforceable will not sufficiently limit privacy threats. At the same time, omnibus legal restrictions have proven very difficult to enforce outside of their immediate jurisdiction. Although some (e.g., Weber, 2010) have argued for an international organization to set key principles as a guide to more detailed regulation originating in the private sector, the feasibility of this approach is uncertain. While the United Nations General Assembly has indeed adopted a consensus resolution strongly supporting a right to privacy (United Nations News Centre, 2013), the IoT is a

> complex sociotechnical system comprised of a plurality of actors, networks, institutions, and contexts, and efforts to regulate privacy at an international level may not be sufficient. Instead, we may see global privacy regulations continue at a variety of levels and contexts rather than as a global, coordinated approach. (Winter, 2015)

Where informational harms are identified, it is not clear whether regulatory data protections alone will suffice. The complexity of applying laws to the IoT is highlighted in the recent case of Volkswagen's alleged fraud related to emissions tests. As Rinesi (2015) observes in the context of this story, existing legal frameworks require technical standards to be explicitly defined,

allowing them to be more readily thwarted. Laws "assume a mechanical universe, not one in which objects get their software updated with new lies every time regulatory bodies come up with a new test" (Rinesi, 2015, para. 5). Although this particular case did not involve sharing of personal data, we can imagine many cases where it might. Further complicating legal approaches, Barocas and Selbst (2016) explain that discrimination based on big data analytics is often an

> unintentional emergent property of the algorithm's use rather than a conscious choice by its programmers, [but] it can be unusually hard to identify the source of the problem or to explain it to a court. (p.1)

The question of whether tweaking existing laws (and, if so, which ones) remains a key question, and one that will be aided immensely by further efforts to map out the IoT ecosystem and distribution of benefits and harms.

Third, as policymakers increasingly turn to big data analytics from non-governmental sources (e.g., data aggregators selling Web-based data), it has become more difficult, and costly, to make these decisions transparent (Davies, 2014). Davies argues that the open data environment that users can access requires analysis in order to "to both differentiate datasets to ask about the specific categories of data released - and to treat the collection of datasets as a whole, looking at how datasets interact to construct a range of possibilities." (2014, p. 12). Understanding the context of the data sets and the way they are constructed are both essential.

Further, the statistical tools used to mine data that we are increasingly basing policy decisions on can identify errors in error assessment procedures, but these are less reliable in large data sets. They are built on assumptions that include "various assertions regarding sampling, stationarity, independence, and so on. Unfortunately, massive data sets are often collected in

ways that seem most likely to break these assumptions" (Committee on the Analysis of Massive Data et al., 2013, p. 14).

Finally, because tensions between privacy and open data invariably put some subset(s) of citizens at risk, we have an obligation to make the overall process itself more transparent. There is no way to achieve absolute personal data protection in order to protect against harms for individuals and groups:

> There will necessarily have to be a trade-off, one which is based on an assessment of the relative value of privacy when compared with the possible gains from data analysis. For society to agree on the terms of this trade-off, it will be necessary to understand exactly what are the possible gains from data analysis.

For example, in the case of "green IT" that promises to reduce carbon emissions, save energy, and provide other, positive environmental benefits, it may be that citizens will reach a general consensus that certain data may be used towards this end. However, citizens may not be aware of the risk of de-anonymized data, so the issue of public awareness and means to better guard against re-identification should be further investigated. Further, such a consensus would require trust in the expert systems that promise these improvements.

There is a vital need to enable citizens to think more critically about the construction of databases and the use of data so we can make more informed decisions, both personal and community life. As big data analytics increasingly underlie critical decisions that affect human welfare – and because the access, creation, and distribution of data are key aspects of economic, political, and sociocultural life – the potential for an even greater gap in terms of who can understand and effectively use data is large. This represents yet another dimension of the digital divide. For this reason, big data literacy (e.g., D'Ignazio & Bhargava, 2015) should be fostered through the formal education system, the press, and through other targeted programs.

References

Angwin, J., & Stecklow, S. (2010 October 12). "'Scrapers' dig deep for data on Web." *The Wall Street Journal*. Retrieved from http://online.wsj.com/article/SB10001424052748703358504575544381288117888.html

Ashton, K. (22 Jun 2009). That 'Internet of Things' thing. *RFID Journal*. Retrieved from http://www.rfidjournal.com/article/view/4986

Barbaro, M., & Zeller, T. (2006). "A face is exposed for AOL searcher no. 4417749." Retrieved from http://www.nytimes.com/2006/08/09/technology/09aol.html?pagewanted=all&_r=0

Barocas, S., & Selbst, A.D. (2016). Big data's disparate impact. *California Law Review, 104*.

Barry-Jester, A.M., Casselman, B., & Goldstein, C. (2015, August 4). "Should prison sentences be based on crimes that haven't been committed yet?" Retrieved from http://fivethirtyeight.com/features/prison-reform-risk-assessment/

Bauer, M., Fabian, B., Fischmann, M., and Gürses, S. (2006). Emerging markets for RFID traces. arXiv:cs/0606018 [cs.CY].

Berners-Lee, T. (2000). Weaving the Web: The past, present and future of the World Wide Web by its inventor. London: Texere.

Boulton, Rawlins, Vallance, & Walport. (2011). Science as a public enterprise: The case for open data. *The Lancet, 377*(9778), 14-20.

boyd, d. (2011, August 4). "'Real names' policies are an abuse of power." Apophenia. Retrieved from http://www.zephoria.org/thoughts/archives/2011/08/04/real-names.html

boyd, d., Levy, K., & Marwick, A. (2014). The networked nature of algorithmic discrimination. In S. Gangadharan (Ed.) *Data and discrimination: Collected essays* (pp. 53-57). Washington, DC: Open Technology Institute – New America Foundation.

Breslin, J., Passant, A. & Decker, S. (2009). *The social semantic web*. Heidelberg: Springer-Verlag.

Cavoukian, A., & Kursawe, K. (2012). Implementing privacy by design: The smart meter case. *Proceedings of the 2012 IEEE International Conference on Smart Grid Engineering* (pp. 1-8). Piscataway, NJ: IEEE.

Chopra, A. (2011, September 15). “Modeling a green energy challenge after a blue button.” Retrieved from http://www.whitehouse.gov/blog/2011/09/15/modeling-green-energy-challenge-after-blue-button

Cisco. (2014, June 10). The Zettabyte era: Trends and analysis. Cisco White Paper. San Jose, Ca.: Cisco Systems. Retrieved from http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/VNI_Hyperconnectivity_WP.pdf

Committee on the Analysis of Massive Data, Committee on Applied and Theoretical Statistics, Board on Mathematical Sciences and Their Applications, Division on Engineering and Physical Sciences, & National Research Council. (2013). Frontiers in massive data analysis. Washington, D.C.: National Academies Press.

Custers, B. (2013). Data dilemmas in the information society: Introduction and overview. In B. Custers, T. Calders, B. Schermer & T. Zarsky (Eds.), *Discrimination and privacy in the information society: Data mining and profiling in large databases* (pp. 3-26). New York; Springer.

Davenport, Barth, and Bean. (2012). “How ‘big data’ is different.” Retrieved from http://sloanreview.mit.edu/article/how-big-data-is-different/

Davies, T. (2014). The construction of open government data: decisions points and discriminatory potential. Data and Discrimination. Annual Meeting of the International Communication Association, Thursday, May 22, 2014 - Seattle, Washington.

D’Ignazio, C. & Bhargava, R. (2015). Approaches to building big data literacy. Bloomberg Data for Good Exchange Conference. Retrieved from https://datatherapy.files.wordpress.com/2015/09/edu_dignazio_52.pdf

Gartner. (2014, November 11). “Gartner says 4.9 billion collected ‘Things’ will be in use in 2015.” Retrieved from https://www.gartner.com/newsroom/id/2905717

Goodin, D. (2015, July 28). “How the way you type can shatter anonymity—even on Tor.” Retrieved from http://arstechnica.com/security/2015/07/how-the-way-you-type-can-shatter-anonymity-even-on-tor/

Gymrek, M., McGuire, A.L., Golan, D., Halperin, E., & Erlich, Y. (2013). Identifying personal genomes by surname inference. *Science, 339*(6117), 321-324.

Heath, T., & Bizer, C. (2011). Linked data: Evolving the Web into a global data space. *Synthesis lectures on the Semantic Web: Theory and technology, 1*(1), 1-136.

Hildner, L. (2006). Defusing the threat of RFID: Protecting consumer privacy through technology-specific legislation at the state level. *Harvard Civil Rights-Civil Liberties Law Review, 41*, 133-176.

IBM. (2008, November). A mandate for change is a mandate for smart. Conversations for a smarter planet series, 1. Retrieved from http://www.ibm.com/smarterplanet/global/files/us__en_us__general__smarterplanet_overview.pdf

International Telecommunication Union.(2014). "2014 ICT figures." Retrieved from https://www.itu.int/net/pressoffice/press_releases/2014/23.aspx

Jaimes, A. (2010). Data mining for user modeling and personalization in ubiquitous spaces. In H. Nakashima, H. Aghajan, & J.C. Augusto (Eds.), *Handbook of ambient intelligence and smart environments* (pp. 1015-1038). London: Springer-Verlag.

Kang, C. (2014, April 2). "Google data-mines its approach to promoting women." Retrieved from https://www.washingtonpost.com/news/the-switch/wp/2014/04/02/google-data-mines-its-women-problem/

Keller, J. (2011 September 29). "Cloud-powered facial recognition is terrifying." *The Atlantic Monthly*. Retrieved October 1, 2011, from http://www.theatlantic.com/technology/archive/2011/09/cloud-powered-facial-recognition-is-terrifying/245867/

Kulk, S. & van Loenen, B. (2012). Brave new open data world. *International Journal of Spatial Data Infrastructures Research, 7*, 196-206.

Lyon, D. (2002). Surveillance as social sorting: Computer codes and mobile bodies. In D. Lyon (Ed.) *Surveillance as social sorting: Privacy, risk and automated discrimination* (pp. 14-30). London, UK: Routledge.

Manfredi, N., Mir, D., Lu, S., & Sanchez, D. (2014). Differentially private models of tollgate usage: The Milan tollgate data set. *IEEE International Conference on Big Data, 2014*, 46-48.

McMillan, C. (2012, September 18). An API for your DNA: Genetic discovery made easy. Retrieved from http://www.programmableweb.com/news/api-your-dna-genetic-discovery-made-easy/2012/09/18

National Institute of Standards and Technology. (2010). Introduction to NISTIR 7628: Guidelines for smart grid cyber security. Gaithersburg, MD: NIST.

Offensive-computing. (2015). Genetic Access Control (rbac-23andme-oauth2). Retrieved from https://github.com/offapi/rbac-23andme-oauth2

O'Reilly, T., & Battelle, J. (2009). *Web squared: Web 2.0 five years on*. White paper presented at the Web 2.0 Summit, October 20–22, 2009, San Franciso, CA: O'Reilly. Retrieved from http://www.web2summit.com/web2009/public/schedule/detail/10194

Patil, D.J. (2015). "A memo to the American people from U.S. Chief Data Scientist Dr. DJ Patil." Retrieved from https://www.whitehouse.gov/blog/2015/02/19/memo-american-people-us-chief-data scientist-dr-dj-patil

PEN American Center. (2013). *Chilling effects: NSA surveillance drives U.S. writers to self-censor*. New York. PEN American Center.

Raftery, T. (2013, October 10). "Utilities should open up API's to their smart meter data." Retrieved from: http://greenmonk.net/2013/10/10/utilities-should-open-up-apis-to-their-smart-meter-data/

Rinesi, M. (2015, September 25). "The price of the Internet of Things will be a vague dread of a malicious world." Retrieved from http://ieet.org/index.php/IEET/more/rinesi20150925

Schwartz, P.M., & Solove, D. (2011). The PII problem: Privacy and a new concept of personally identifiable information. *New York University Law Review, 86*, 1814-1894.

St. John, J. (2014, March 13). "Hidden treasure: Two new resources offer up massive amounts of utility data." *Greentechgrid*. Retrieved from: https://www.greentechmedia.com/articles/read/Energy-Data-Treasure-from-Chattanoogas-Smart-Grid-Incubator-and-Pecan-Str

Tendril. (2012, January 20). "NYC Cleanweb Hackathon: Crowdsourcing killer energy apps." Retrieved from http://www.tendrilinc.com/blog/nyc-cleanweb-hackathon-crowdsourcing-killer-energy-apps

Turow, J. (2006). Cracking the consumer code: Advertisers, anxiety and surveillance in the digital age. In K.D. Haggerty and R.V. Ericson (Eds.) *The new politics of surveillance and visibility* (pp. 279-307). Toronto: University of Toronto Press.

Turow, J. (2012). *The daily you: How the new advertising industry is defining your identity and worth*. New Haven, Ct.: Yale University Press.

Uckelmann, D., Harrison, M. & Michahelles, F. (2010). An architectural approach towards the future Internet of Things. In D. Uckelmann et al. (Eds.), Architecting the Internet of Things. Berlin: Springer-Verlag Berlin Heidelberg.

United Nations News Centre (19 December, 2013). "General Assembly backs right to privacy in digital age." Retrieved from http://www.un.org/apps/ news/story.asp? NewsID=46780&Cr=privacy&Cr1=#.UuW_5hB6dD8

Upturn. (2014). *Civil rights, big data, and our algorithmic future*. Retrieved from https://bigdata.fairness.io/

Weber, R.H., & Weber, R. (2010). *Internet of Things: Legal perspectives*. Berlin: Springer-Verlag Berlin Heidelberg.

Weber, R.H. (2010). Internet of Things: New security and privacy challenges. *Computer Law and Security Review, 26*, 23-30.

Winter, J.S. (2014). Surveillance in ubiquitous network societies: Normative conflicts related to the consumer in-store supermarket experience in the context of the Internet of Things. *Ethics and Information Technology, 16*(1), 27-41. doi:10.1007/s10676-013-9332-3.

Winter, J.S. (2015). Privacy challenges for the Internet of Things. In M. Khosrow-Pour (Ed.), *Encyclopedia of Information Science and Technology*, Third edition (pp. 4373-4383). Hershey, PA: IGI Global.

Winter, J.S. (2016). Algorithmic discrimination: Big data analytics and the future of the Internet. In J.S. Winter & R. Ono (Eds.), *The future Internet: Alternative visions*. New York: Springer.