Datafiction, dataism and dataveillance: Big Data between scientific paradigm and secular belief

van Dijck, J.

Publication date
2014

Document Version
Final published version

Published in
Surveillance & Society

Link to publication

Citation for published version (APA):
van Dijck, J. (2014). Datafiction, dataism and dataveillance: Big Data between scientific paradigm and secular belief. *Surveillance & Society*, *12*(2), 197-208.
http://library.queensu.ca/ojs/index.php/surveillance-and-society/article/view/datafication

Surveillance & Society

Article

Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology

José van Dijck

University of Amsterdam, The Netherlands.
j.van.dijck@uva.nl

Abstract

Metadata and data have become a regular *currency* for citizens to pay for their communication services and security—a trade-off that has nestled into the comfort zone of most people. This article deconstructs the ideological grounds of datafication. Datafication is rooted in problematic ontological and epistemological claims. As part of a larger social media logic, it shows characteristics of a widespread secular belief. Dataism, as this conviction is called, is so successful because masses of people—naively or unwittingly—trust their personal information to corporate platforms. The notion of trust becomes more problematic because people's faith is extended to other public institutions (e.g. academic research and law enforcement) that handle their (meta)data. The interlocking of government, business, and academia in the adaptation of this ideology makes us want to look more critically at the entire ecosystem of connective media.

Introduction

When Edward Snowden, on June 10, 2013, made himself known as the whistleblower that had exposed N.S.A. practices of routine surveillance to the news media, he described in detail the "architecture of oppression" that enabled him and many other N.S.A.-contractors to intercept the metadata of three billion phone calls and interactions recorded by Facebook, Google, Apple, and other tech companies. In a videotaped interview, the former CIA-analyst said he could no longer live with the extensive privacy invasion and legal violations he had to perform on behalf of the intelligence community. He also wanted to make people aware of the fact that many agents have complete access to all kinds of communication data, hoping to spark a public debate.

Snowden's disclosures have been more than a wakeup call for citizens who have gradually come to accept the "sharing" of personal information—everything from marital status to colds, and from eating habits to favorite music—via social network sites or apps as the new norm (van Dijck 2013a). Platform owners routinely share users' aggregated metadata with third parties for the purpose of customized marketing in exchange for free services. Many people may not have realized, up until Snowden's leaks, that corporate social networks also—willingly or reluctantly—share their information with intelligence agencies. When Barack Obama defended his administration's policies of mass surveillance saying that there was "no content, just metadata" involved in the PRISM scheme, he added that citizens cannot expect a hundred per cent security and a hundred per cent privacy and no inconvenience. The president's explanation echoed social media companies' argument that users have to give up part of their privacy in exchange for free convenient platform services. In other words, metadata appear to have become a regular *currency* for

van Dijck, José. 2014. Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology. *Surveillance & Society* 12(2): 197-208.
http://www.surveillance-and-society.org | ISSN: 1477-7487

citizens to pay for their communication services and security—a trade-off that has nestled into the comfort zone of most people.

What explains this remarkable tolerance for Big Brother and Big Business routinely accessing citizens' personal information also known as Big Data? Part of the explanation may be found in the gradual normalization of *datafication* as a new paradigm in science and society. Datafication, according to Mayer-Schoenberger and Cukier (2013) is the transformation of social action into online quantified data, thus allowing for real-time tracking and predictive analysis. Businesses and government agencies dig into the exponentially growing piles of metadata collected through social media and communication platforms, such as Facebook, Twitter, LinkedIn, Tumblr, iTunes, Skype, WhatsApp, YouTube, and free e-mail services such as gmail and hotmail, in order to track information on human behavior: "We can now collect information that we couldn't before, be it relationships revealed by phone calls or sentiments unveiled through tweets" (Mayer-Schoenberger and Cukier 2013: 30). Datafication as a legitimate means to *access*, *understand* and *monitor* people's behavior is becoming a leading principle, not just amongst techno-adepts, but also amongst scholars who see datafication as a revolutionary research opportunity to investigate human conduct.

In this article, I would like to deconstruct the ideological grounds of datafication as defined by Mayer-Schoenberger and Cukier and echoed by many proponents of this new scientific paradigm. I will argue that in many respects datafication is rooted in problematic ontological and epistemological claims. However compelling some examples of applied Big Data research, the ideology of *dataism* shows characteristics of a widespread *belief* in the objective quantification and potential tracking of all kinds of human behavior and sociality through online media technologies. Besides, dataism also involves *trust* in the (institutional) agents that collect, interpret, and share (meta)data culled from social media, internet platforms, and other communication technologies.

Notions of "trust" and "belief" are particularly relevant when it comes to understanding *dataveillance*: a form of continuous surveillance through the use of (meta)data (Raley 2013). As Snowden's documents made clear, people have faith in the institutions that handle their (meta)data on the presumption that they comply with the rules set by publicly accountable agents. However, as journalists found out, the N.S.A. regularly defies court rulings on data use, just as corporations are constantly testing legal limits on privacy invasion.[1] More profoundly, the Snowden files have further opened people's eyes to the interlocking practices of government intelligence, businesses, and academia in the adaptation of dataism's ideological premises. Therefore, we need to look into the credibility of the whole ecosystem of connective media. What are the distinctive roles of government, corporations and academia in handling our data? And what kind of critical attitude is required in the face of this complex system of online information flows?

Datafication and "life mining" as a new scientific paradigm

Over the past decade, datafication has grown to become an accepted new paradigm for understanding sociality and social behavior. With the advent of Web 2.0 and its proliferating social network sites, many aspects of social life were coded that had never been quantified before—friendships, interests, casual conversations, information searches, expressions of tastes, emotional responses, and so on. As tech companies started to specialize in one or several aspects of online communication, they convinced many people to move parts of their social interaction to web environments. Facebook turned social activities such as "friending" and "liking" into algorithmic relations (Bucher 2012; Helmond and Gerlitz 2013); Twitter popularized people's online personas and promoted ideas by creating "followers" and "retweet" functions (Kwak et al. 2010); LinkedIn translated professional networks of employees and job seekers into

[1] Advocacy group Consumer Watchdog, in May 2013, filed a suit against Google claiming that Google unlawfully opens up, reads, and acquires the content of people's private email messages.

digital interfaces (van Dijck 2013b); and YouTube datafied the casual exchange of audiovisual content (Ding et al. 2011). Quantified social interactions were subsequently made accessible to third parties, be it fellow users, companies, government agencies, or other platforms. The digital transformation of sociality spawned an industry that builds its prowess on the value of data and metadata—automated logs showing who communicated with whom, from which location, and for how long. Metadata—not too long ago considered worthless byproducts of platform-mediated services—have gradually been turned into treasured resources that can ostensibly be mined, enriched, and repurposed into precious products.

The industry-driven datafication view resonates not only in entrepreneurs' auspicious gold rush metaphors, but also in researchers' claims hailing Big Data as the holy grail of behavioral knowledge. Data and metadata culled from Google, Facebook, and Twitter are generally considered *imprints* or *symptoms* of people's actual behavior or moods, while the platforms themselves are presented merely as neutral facilitators. Twitter supposedly enables the datafication of people's sentiments, thoughts, and gut-feelings as the platform records "spontaneous" reactions; users leave traces unconsciously, so data can be "collected passively without much effort or even awareness on the part of those being recorded" (Mayer-Schoenberger and Cukier 2013: 101). Analysts often describe the large-scale gauging of tweets as using a thermometer to measure feverish symptoms of crowds reacting to social or natural events—an assumption founded on the idea that online social traffic flows through neutral technological channels. In this line of reasoning, neither Twitter's technological mediation by hashtags, retweets, algorithms, and protocols, nor its business model seems relevant (Gillespie 2010).

Researchers endorsing the datafication paradigm tend to echo these claims concerning the nature of social media data as natural traces and of platforms as neutral facilitators. Information scientists have called Twitter a "sensor" of real-time events when processing people's tweets about earthquakes or other disasters (Sakaki, Okazaki and Matsuo 2010); Twitter has also been termed a "sentiment detector" of people's political predilections (O'Connor et al. 2010) and a tool that helps understand the "dynamics of sentiment" by analyzing twitterers' reactions to a specific video fragment (Diakopoulos and Shamma 2010; Bollen, Mao and Pepe 2010. Assessing big data sets collected through social media platforms is increasingly presented as the most scrupulous and comprehensive method to measure quotidian interaction, superior to sampling ("N=all") and more reliable than interviewing or polling. Large amounts of "messy" data replace small amounts of sampled data and, as proponents assert, the sheer size of data sets compensates for their messiness. Some information scientists argue that Twitter is in fact a giant real-time polling tool, ready to become "a substitute and supplement for traditional polling" (O'Connor et al. 2010). There are important parallels between polls and Twitter data, and the correlations found in Twitter results are obviously meaningful. However, caveats about Twitter's alleged representativeness and (technological and commercial) biases are poorly addressed.[2]

Datafication enthusiasts also often assume a self-evident relationship between data and people, subsequently interpreting aggregated data to predict individual behavior. For instance, Quercia et al. (2011) analyzed the relationships between personality and different types of twitterers, finding that

[2] Just a few remarks about Twitter's alleged representativeness and inherent biases. Twitter's user base do not match the demographics of a general public. A Pew Internet and American Life Project, published in February 2012, found that only 15 per cent of online adults use Twitter and only 8 per cent use it on a daily basis (see http://pewinternet.org/Reports/2012/Twitter-Use-2012/Findings.aspx). Furthermore, Twitter deploys several algorithms that favor influential users and allow for manipulation of tweet messages, either by the platform itself or by concerted groups of users (see Cha et al. 2010). Twitter data are often treated as equivalents of polling results, despite explicit disqualifications of the tool's representational value. For instance, the Twitter Political Index or Twindex, launched in January 2012, tracks tweets that mention candidates running for office. Twindex, a partnership between Twitter, search engine Topsy and a bipartisan pair of political pollsters, attempts to measure "the public's shifting moods, as well as to establish Twitter as a platform for civic debate" (see https://election.twitter.com/).

popular and influential users are both “imaginative” and “organized”. On the basis of these patterns, they speculate which users may successfully recommend products or help boost marketing strategies. Along similar lines, a recent study by Kosinski and others (2013) shows how private traits and attributes are predictable from digital records of human behavior; in this case, Facebook Likes were used to “automatically and accurately predict a range of highly sensitive personal attributes including: sexual orientation, ethnicity, religious and political views, personality traits, intelligence, happiness, use of addictive substances, parental separation, age, and gender” (Kosinski, Stillwell and Graepel 2013: 1). The authors conclude that such private information may be used to optimize personalized platform services and offer social psychologists a wealth of data they could never have gained otherwise.

Identifying patterns of conduct or activities out of unconsciously left (meta)data on social network sites increasingly serves to predict future behavior. Information scientists Weerkamp and De Rijke (2012) state it very clearly: “We are not interested in current or past activities of people, but in their future plans. We propose the task of activity prediction, which revolves around trying to establish a set of activities that are likely to be popular at a later time.” They position activity prediction as a special case of “life mining”, a concept defined as “extracting useful knowledge from the combined digital trails left behind by people who live a considerable part of their life online.” The phrase “useful knowledge” begs the question: useful for whom? According to Weerkamp and De Rijke, social media monitoring provides meaningful information for police and intelligence services to forecast nascent terrorist activity or calculate crowd control, and for marketers to predict future stock market prices or potential box office revenues (see also Asur and Huberman 2011. From the viewpoints of surveillance and marketing, predictive analytics—relating (meta)data patterns to individual’s actual or *potential* behavior and vice versa—yields powerful information about who we are and what we do. When it comes to human behavior, though, this logic may also reveal a slippery slope between analysis and projection, between deduction and prediction (Amoore 2011).

A “big data mindset” also seems to favor the paradoxical premise that social media platforms concomitantly *measure*, *manipulate*, and *monetize* online human behavior. Even though metadata culled from social media platforms are believed to reflect human behavior-as-it-is, the algorithms employed by Google, Twitter and other sites are intrinsically selective and manipulative; both users and owners can game the platform. For instance, when Diakopoulos and Shamma (2010) predict political preferences by analyzing debate performance through tweets, they seem to ignore the potential for spin-doctors or partisan twitterers to influence Twitter debates in real time. In marketing circles, the prediction of future customers’ needs is akin to the manipulation of desire: detecting specific patterns in consumer habits often results in simultaneous attempts to create demand—a marketing strategy that is successfully monetized through Amazon’s famed recommendation algorithm (Andrejevic 2011). Social media content, just like internet searches, is subject to personalization and customization, tailoring messages to specific audiences or individuals (Pariser 2011; Bucher 2012). Promoting the idea of metadata as traces of human behavior and of platforms as neutral facilitators seems squarely at odds with the well-known practices of data filtering and algorithmic manipulation for commercial or other reasons.

Datafication and life mining are staked in ideological assumptions, which are, in turn, rooted in prevailing social norms. As said before, users provide personal information to companies and receive services in return—a form of barter. Metadata in exchange for communication services has become the norm; few people appear willing to pay for more privacy.[3] The currency used to pay for online services and for security has turned metadata into a kind of invisible asset, processed mostly separate from its original context and outside of people’s awareness. Social media companies monetize metadata by repackaging

[3] A report by the European Network Information and Security Agency (ENISA 2012), showed that less than one-third of experimental subjects in a study on “privacy-for-data” exchange were willing to pay extra if the service-provider promised not to use their data for marketing purposes.

and selling them to advertisers or data companies. Information scientists often uncritically adopt the assumptions and ideological viewpoints put forward by SNSs and data firms. The datafication paradigm thus performs a profound ideological role at the intersection of sociality, research, and commerce—an inextricable knot of functions that has been conspicuously under-examined.

Dataism: unraveling datafication's ideological underpinnings

The data mining metaphor is grounded in a peculiar rationale that guides entrepreneurs, academics, and state agencies in their pursuit of a new social-scientific paradigm. First and foremost, dataism betrays a belief in the objectivity of quantification and in the potential of tracking all kinds of human behavior and sociality through online data. Secondly, (meta)data are presented as "raw material" that can be analyzed and processed into predictive algorithms about future human behavior—valuable assets in the mining industry. Let me explore in more detail each of these ontological and epistemological assertions underpinning dataism as a belief in a new gold standard of knowledge about human behavior.

A first line of critical inquiry is leveled at the alleged objective nature of data. In a thought-provoking essay, social scientists boyd and Crawford (2012: 2) deconstruct the widespread mythology that "large data sets offer a higher form of intelligence and knowledge that can generate insights that were previously impossible, with the aura of truth, objectivity, and accuracy." Piles of (meta)data are purposefully generated through a number of different online platforms which are anything but objective. Metadata relate to human behavioral acts in the same way as MRI scans relate to body interiors: signs of disease never simply appear on a screen, but are the result of careful interpretation *and intervention* in the imaging process. It took medical technicians decades to learn proper imaging of specific organs; they had to refine protocols for positioning bodies and tweak the machine's performance to enhance the tool's usefulness (van Dijck 2005). Facebook and Twitter are apparatuses that are constantly tweaked to translate friendship or popularity into algorithms while promoting these very computations as *social* values (Manovich 2011; Bucher 2012). "Likes" and "trending topics" buttons may be commonly perceived as icons of spontaneous online sociality, but the algorithms underlying these buttons are systematically fine-tuned to channel user responses (Mahrt and Scharkow 2013).

The idea of (meta)data being "raw" resources waiting to be processed perfectly fits the popular life-mining metaphor. According to Mayer-Schoenberger and Cukier (2013), each single data set is likely to have some intrinsic, hidden, not yet unearthed value, and companies are engaged in a race to discover how to capture and rate this value. But as Lisa Gitelman aptly states, "raw data" is an oxymoron: "Data are not facts, they are 'that which is given prior to argument' given in order to provide a rhetorical basis. Data can be good or bad, better or worse, incomplete and insufficient" (Gitelman 2013: 7). Automated data extraction performed on huge piles of metadata generated by social media platforms reveals no more information about specific human behavior than large quantities of sea water yield information about pollution—unless you interpret these data using specific analytical methods guided by a focused query.

Here is an example to illustrate this point. A team of information scientists collected six months of search results culled at regular intervals from users who first entered the keyword "home mortgage" into a search engine, in order to find how correlations changed over time (Richardson 2008). The data "show" how mortgage seekers, six weeks after their initial query, move from mortgage basics to insurance and taxes; three months later they look for furnishings, and six months after their initial mortgage query they are interested in pools and patio accessories. However, correlations like these do not simply "emerge". They are much rather induced by an implicit question framing the inquiry: what do new homeowners need to buy in the first six months after acquiring their home? Explicating this question reveals that an interpretative frame always prefigures data analysis. Following Gitelman's line of thought, data provide a rhetorical basis for the argument that new homeowners "need" certain stuff at certain moments—a pattern prediction valuable to advertisers.

Making sense of patterns thus requires *critical* interrogation: why do we look for certain patterns in piles of metadata, in whose interests, and for what purposes? Identifying meaningful patterns on the basis of data culled from online platforms is an intrinsically interpretative act, although you may have to spell out implicit prerogatives. Messages from millions of female Facebook users between 25 and 35 years of age posting baby pictures in their Timeline may be endlessly scrutinized for behavioral, medical, or consumerist patterns. Do researchers want to learn about young mothers' dieting habits with the intention of inserting propositions to change lifestyles? Or do they want to discover patterns of consumptive needs in order for companies to sell baby products at exactly the right moment? Or, perhaps far-fetched, are government agencies interested in interpreting these data for signs of postnatal depression or potential future child abuse? Quantitative methods beg for *qualitative* interrogation to disprove the claim that data patterns are "natural" phenomena. Big data research, in other words, always involves an explicit prism (no pun intended).

Raw data do not go in at one end of the digital assembly lines managed by Google or Facebook while processed information comes out at the other end, as Mayer-Schonberger and Cukier (2013: 101) contend. Metadata are value-laden piles of code that are multivalent and should be approached as multi-interpretable texts. According to American scholar John Cheney-Lippold, data are cultural objects "embedded and integrated within a social system whose logic, rules, and explicit functioning work to determine the new conditions of possibilities of users' lives" (Cheney-Lippold 2011: 167). Big Data configured as a rhetorical text which has been generated for specific purposes and which can be probed by various groups of people, offers an alternative to the pervasive mining metaphor. Academics looking at data sets from a social science or humanities perspective may pose very different questions than information scientists; and medical doctors are likely to see different patterns than criminologists (Manovich 2011).

The compelling logic of dataism is often fueled by the rhetoric of new frontiers in research, when large sets of unconsciously left data, never available before, are opening up new vistas. Dataism thrives on the assumption that gathering data happens outside any preset framework—as if Twitter facilitates microblogging just for the sake of generating "life" data—and data analysis happens without a preset purpose—as if data miners analyze those data just for the sake of accumulating knowledge about people's behavior. It may not always be simple to identify in what context (meta)data are generated and for what purposes they are processed. And yet it is crucial to render hidden prerogatives explicit if researchers want to keep up users' trust in the datafication paradigm. Trust is partly grounded in the persuasive logic of a dominant paradigm; for another part, though, faith resides with the institutions that carry the belief in Big Data.

Dataism and the trust in institutions

A second line of critical inquiry is leveled at the institutional structures that scaffold Big Data thinking. Data companies, government agencies and researchers alike underscore the importance of users' trust in societies where growing parts of civilian life—from application procedures to medical records and financial transactions—are moved onto online platforms. Establishing and maintaining the system's integrity is often assigned as a task to "the state"—whereas "the platforms" have to comply with the rules set by government agencies. When Mayer-Schoenberger and Cukier (2013) address the perils of metadata's ubiquitous availability—i.e. profiling based on stereotypes, penalties based on propensities, surveillance based on association, a weakened right to privacy—they hold governments responsible for taking measures to avert these potential risks. The authors of *Big Data* call for a new "caste of big-data auditors we call algorithmists" to "secure a fair governance of information in the big-data era" (Mayer-Schoenberger and Cukier 2013: 184). Academics, too, count on national governments to regulate possibly adverse effects of datafication; but they also turn to data companies when they call for "trust and goodwill" from corporations and ask them to give users "transparency and control" over their information

(Kosinski et al. 2013). In striving for trust and credibility, there is a presumed separation of public, corporate, and state institutions as autonomous bodies that each has a distinctive relationship with users—whether consumers or citizens.

Needless to say, neither "the state" nor "data firms" are monolithic categories. For one thing, various government agencies—besides the N.S.A—each represent a specific relationship with users and thus play a specific role in the maintenance of trust. Agencies like the F.T.C and the N.I.S.T. have the legal means and the political obligation to secure citizens against privacy and exploitation risks propelled by the datafication paradigm.[4] Data companies, for their part, are simultaneously competitors and allies when it comes to winning and keeping users' trust. Users' faith in their data policies may be part of a single company's competitive edge; however, since partnerships in this sector abound it is impossible for users to keep track of who shares data with whom.

And yet, if the Snowden files have taught us anything, it is probably that institutions gathering and processing Big Data are not organized *apart from* the agencies that have the political mandate to regulate them. In fact, all three apparatuses—corporate, academic, and state—are highly staked in getting unrestraint access to metadata as well as in the public's acceptance of datafication as a leading paradigm. Scientists, government agencies and corporations, each for different reasons, have a vested interest in datafied relationships and in the development of methods that allow for prediction as well as manipulation of behavior. The aspirations of all agents to know, predict, and control human behavior overlap to some extent but differ on other accounts. Data firms want their platforms to be acknowledged as objective, standardized aggregators of metadata—better and more precise than the tools government agencies or academics use for measuring consumer sentiment, public health, or social movements.[5] When government agencies and academics adopt commercial social media platforms as the gold standard for measuring social traffic, they in fact transfer the power over data-collection and interpretation from the public to the corporate sector. As boyd and Crawford (2012: 14) argue: "There is a deep government and industrial drive toward gathering and extracting maximal value from data, be it information that will lead to more targeted advertising, product design, traffic planning, or criminal policing."

In this tripartite alignment of forces, government, academia, and data firms are interconnected at the level of personnel as well as through their exchange of innovative technologies, i.e. by co-developing data mining projects. In an article on the Snowden case for *The New York Times*, reporters Risen and Wingfield (2013) bare close connections between Silicon Valley and the N.S.A.: "Both hunt for ways to collect, analyze and exploit large pools of data about millions of Americans. The only difference is that the N.S.A. does it for intelligence, and Silicon Valley does it to make money." Links between data firms and state intelligence agencies show how technical experts rotate jobs between academia and health industries, and move from data firms to financial services or intelligence agencies. The interests of corporations, academics, and state agencies converge in various ways. For instance, Skype and its owner Microsoft readily engaged with the C.I.A. on Project Chess aimed at making Skype calls useable to law enforcement officials. As Timothy Garton-Ash (2013) quipped in an op-ed in *The Guardian*: if Big Brother came back in the 21st century, "he would return as a private-public partnership."

[4] The American Federal trade Commission (F.T.C.) is assigned to protect consumers and to eliminate and prevent anti-competitive business practices; the National Institute of Standards and Technology (N.I.S.T.) is the agency charged with setting federal cybersecurity standards. Both agencies scrambled to restore public confidence after Snowden's revelations about the N.S.A.

[5] Google executives argue that Google search data can reveal trends a week or two earlier than official government statistics (Aspen Institute Report 2010). In addition, it is argued that Google Flu Trends is a better instrument to measure for emerging flu epidemics than national surveillance systems for influenza-like symptoms (Wilson et al. 2009).

What is at issue here is not just an embrace of dataism as a technique of knowing social action—human behavior being measured, analyzed, and predicted on the basis of large sets of metadata—but also as a faith in high-tech companies' and government agencies' intention to protect user data from exploitation. Dataism presumes *trust* in the objectivity of quantified methods as well as in the *independence* and *integrity* of institutions deploying these methods—whether corporate platforms, government agencies, or academic researchers. Trust and independence, however, are embattled notions in an ecosystem of connectivity where all online platforms are inevitably interconnected, both on the level of infrastructure as on the level of operational logic (van Dijck 2013a; van Dijck and Poell 2013). When everything and everyone is connected through the same infrastructure and operates through the same logic—a view theorized by Foucault well before the advent of online technologies.

For instance, the logic of predictive analytics appears to be corroborated by governments, researchers, and corporations alike. Google claims they are much better than state agencies in *forecasting* unemployment statistics or flu epidemics because their web crawlers can determine when an individual is about to start looking for a new job or starts seeking information about influenza. Facebook Likes can potentially *predict* which young mothers may be likely to malnourish their children—information which state health agencies may act upon. And the N.S.A. declares they have *prevented* at least fifty terrorist attacks due to the PRISM scheme, based on data culled from social media platforms and e-mail services. Problematic in these institutional forms of dataism is not only the fact that we lack insight in the algorithmic criteria used to define what counts as job seeking, dysfunctional motherhood, or terrorism. More questionably, the contexts in which the data were generated and processed—whether through commercial platforms or public institutions—all appear to be interchangeable.

What is at stake here is not simply our "trust" in specific government agencies or single corporations, but the credibility of the entire ecosystem—an ecosystem that is fueled by a steady flow of billions of e-mails, video, text, sound, and metadata. The custody over data flows appears to be mired in a fuzzy delineation of territories; access and restrictions to data are fought over both before the public's eye and outside people's realm of knowing. Since Snowden's revelations, users-citizens have increasingly questioned American high-tech companies' cozy relationships with governments, and, in response, some companies have filed court complaints against what they call N.S.A.-bullying tactics. This public struggle over whom to trust with user data may serve to enhance the impression of each institution's independence, and yet, it is obvious that data firms like Google and Facebook do not operate in a vacuum. The ecosystem is typically an infrastructure where no single institution is in command (Brivot and Gendron 2011: 153), but which credibility is disputed in a number of public debates, court struggles, and political skirmishes—including government attempts to curb whistleblowers' leaks.

The interpellation of dataism as a shared belief built on institutional trust seems to be equally important than the interrogation of datafication premises. Embattled notions of "trust" and "belief" are particularly relevant when it comes to understanding *dataveillance* as an increasingly preferred way of monitoring citizens through social media and online communication technologies (Raley 2013). What are the distinctive interests of government, business, and academia in handling our data? Dataveillance raises more questions regarding the credibility of the entire system of online information flows.

Dataveillance and the struggle for credibility

Several months after the N.S.A. revelations started, Google, Facebook, Yahoo and Microsoft struck back at their critics who charged them with government collaboration and betraying users' privacy by suing the Foreign Intelligence Surveillance Agency (F.I.S.A.), which provides the legal framework for N.S.A. operations. Facebook's Mark Zuckerberg claimed in a newspaper interview that the American government had done a "bad job of balancing people's privacy and its duty to protect" and Yahoo's Marissa Mayer conceded they had to fight the N.S.A. in court to maintain her company's trustworthiness towards both users and investors (Rushe 2013). Interestingly, what we saw in the aftermath of the Snowden revelations was that data companies teamed up and rallied against the N.S.A. to regain the public's trust. The disclosure of routine dataveillance tactics threatened to seriously undermine not just people's trust in state agencies or individual corporations, but in dataism's institutional pillars as a whole.

Dataveillance—the monitoring of citizens on the basis of their online data—differs from surveillance on at least one important account: whereas surveillance presumes monitoring for specific purposes, dataveillance entails the continuous tracking of (meta)data for unstated preset purposes. Therefore, dataveillance goes well beyond the proposition of scrutinizing individuals as it penetrates every fiber of the social fabric (Andrejevic 2012: 86). Dataveillance is thus a far-reaching proposition with profound consequences for the social contract between corporate platforms and government agencies on the one hand and citizens-consumers on the other. Let's look more closely at the distinctive role of each actor in this battle for credibility and trust.

From the onset, Facebook and Google superficially anchored their users' expectations of trust in corporate mantras such as "Do no evil" (Google) and "Making the world transparent and connected" (Facebook). To them, the social contract with consumers was staked in making online sociality visible and traceable; part of this call for transparency was requiring authentic and verifiable personal information from their registered customers (van Dijck 2013b). However, platforms offered little transparency in return; from 2007 to this very day, companies like Facebook have engaged in battles with the F.T.C. and courts of law to defend its continuously changing Terms of Use, which keep stretching its privacy policy.[6] Over the past few years, user advocates have taken Facebook and other platforms to court for unlawfully keeping logs of user data. Consumer advocacy groups have tirelessly called for explicating the quid-pro-quos of free online services to help restore public trust in single platforms as well as the ecosystem as a whole. And alternative platforms for search and communication—e.g. Lavabit, DuckDuckGo, Path, Leaf, and Silentcircle—have tried to balance off users' data protection with reliable services. However, it turns out to be very hard to escape from the rules and practices set by the dominant players in the system.

The compliance of high-tech firms with post-Patriot-Act laws, dutifully reported by journalists in the wake of the Snowden affair, certainly contributed to the waning public trust in dataveillance tactics; so it is not surprising to find CEOs from data companies lashing out at the N.S.A. and vocally trying to re-establish their image as neutral facilitators. Platform owners' attitude vis-à-vis administrative bodies are often ambivalent, though. They call upon governments to mend the gaps in laws and policies (Brown, Chui and Minyika 2011: 11), but these same companies warn the government against overregulation and propose to leave "openness" to be regulated by the technology sector itself (Schmidt and Cohen 2013).

A similar ambivalence can be seen coming from the government. Obviously, intelligence agencies have different interests than government regulators. Security and privacy issues often pose contradictory demands, leading to ambivalent legal definitions, such as Obama's validation of metadata ("we're not listening in on your phone conversations") as legitimate means for dataveillance. Citizens' groups rightly call for clear-cut policies that guard privacy and balance it off with security. Bringing legal definitions in tune with advanced technological apparatuses is just one pivotal step in the effort to rebuild trust. As we have seen in the banking crisis, starting in 2008, a loss of trust in the finance sector was caused by a similar murkiness involved in many complex financial schemes and the high-tech based logic of derivatives; after two decades of self-regulation, trust in banking systems has come to an all-time low.

[6] Over the last year, Facebook had to defend its practice of making "shadow profiles" of friends you connect to by copying addresses and phone numbers; the platform also had to defend its automatic assumption that the parents of teenagers using the service have given permission for their names and images to be used in Facebook advertising (Oremus 2013; Goel and Wyatt 2013).

Responsibility for maintaining credibility of the ecosystem as a whole also resides with academics. The unbridled enthusiasm of many researchers for datafication as a neutral paradigm, reflecting a belief in an objective quantified understanding of the social, ought to be scrutinized more rigorously. Uncritical acceptance of datafication's underpinning ideological and commercial premises may well undermine the integrity of academic research in the long run. To keep and maintain trust, Big Data researchers need to identify the partial perspectives from which data are analyzed; rather than maintaining claims to neutrality, they ought to account for the context in which data sets are generated and pair off quantitative methodologies with qualitative questions. Moreover, the viability and verifiability of predictive analytics as a scientific method deserves a lot more interdisciplinary enquiry, combining for instance computational, ethnographic and statistical approaches (Giglietto et al. 2012: 155).

Academics are significant actors in the building of *social trust*: a paradigm resting on the pillars of academic institutions often forms an arbiter of what counts as fact or opinion, as fact or projection. In the world of online sociality, where human behavior is coded into (meta)data and mediated by platforms, the distinctions between facts, opinions, and predictions—between objectivities, subjectivities, and potentialities—are gradually erased. In the words of sociologist Bruno Latour (2007), they are obliterated "in such a way that they are both graduating to the same type of visibility—not a small advantage if we wish to disentangle the mixture of facts and opinions that has become our usual diet of information." If predictive analytics and real-time data analytics become the preferred modes of scientific analysis of human behavior, humanities and social science scholars seriously need to address the fundamental epistemological and ontological questions merely scratched upon in the previous sections.

Meanwhile, as Edward Snowden's unscrupulous actions show, there is an overarching significant actor in the fight for credibility that is often overlooked: users-citizens. When Snowden made the choice to go public with his inside information on N.S.A. dataveillance practices, he not only showed the power of an individual employee to unveil and unsettle a complex state-industrial-academic complex of forces. He also counted on the vigilance of many citizens—researchers, influential bloggers, journalists, lawyers and activists—to take public his concern about the structural flaws in the ecosystem that is currently developing. Over the past decade, the actual power of users-citizens vis-à-vis corporate platforms and the state has triggered substantial debate, albeit mostly in activist and academic circles. Some have found the ability of users to resist platforms' privacy policies and surveillance tactics to be quite limited; individuals are steered by platforms' technologies and business models of single platforms while it is extremely hard to gain insight in the system's interdependence and complexity (Draper 2012; Hartzog and Selinger 2013; Mager 2012). Other researchers have argued in favor of strengthening digital (consumer) literacy particularly at the level of understanding privacy and security in relation to social data (Pierson 2012). And there is a growing mass of critical scholarship stressing the importance of users in baring how connective media are forging a new social contract on societies while refurbishing sociality and democracy in online environments (Langlois 2013; Lovink 2012).

The much wider public debate fueled by Snowden is itself an eminent example of a project to restore the internet's credibility.[7] It is through shocks like these that people become more aware of the institutional and ideological forces involved in an evolving paradigm. The popularization of datafication as a neutral paradigm, carried by a belief in dataism and supported by institutional guardians of trust, gradually yielded a view of dataveillance as a "normal" form of social monitoring. Perhaps it took Snowden to blow

[7] A survey (July 2013) by the Pew Research Center for People and the Press shows that Snowden's revelations have indeed affected public opinion on surveillance and security. The report states that a "majority of Americans – 56% – say that federal courts fail to provide adequate limits on the telephone and internet data the government is collecting as part of its anti-terrorism efforts. An even larger percentage (70%) believes that the government uses this data for purposes other than investigating terrorism." See: http://www.people-press.org/2013/07/26/few-see-adequate-limits-on-nsa-surveillance-program/

a whistle on these increasingly normalized practices, but it certainly takes more than one whistleblower to launch a full-fletched inquiry into the new online pillars of democracy and sociality. The issues put on the agenda by Snowden certainly deserve to remain in the spotlights of public attention until all precarious matters are addressed.

Acknowledgements

I would like to thank two anonymous reviewers of this article for their insightful and constructive comments. I would also like to thank Kelly Gates, Mark Andrejevic, Thomas Poell, Maarten de Rijke, Arnold Smeulders and Ton Brouwers for conversations on this topic and helpful editorial comments.

References

Amoore, L. 2011. Data Derivatives : On the Emergence of a Security Risk Calculus for Our Times. *Theory Culture and Society* 28 (6): 24-43.

Andrejevic, M. 2011. The work that affective economics does. *Cultural Studies* 25(4/5): 604-620.

Andrejevic, M. 2012. Exploitation in the data-mine. In: *Internet and Surveillance: The Challenges of Web 2.0 and Social Media*, eds C. Fuchs, K. Boersma, A. Albrechtslund, and M. Sandoval, 71-88. New York: Routledge.

Aspen Institute report. 2010. The Promises and perils of Big Data. Washington DC: The Aspen Institute. Available at: http://www.aspeninstitute.org/publications/promise-peril-big-data (accessed May 7, 2014).

Asur, S. and B.A. Huberman. 2011. Predicting the Future With Social Media.Paper published by Cornell University Open Library. Available at: http://arxiv.org/abs/1003.5699 (accessed May 7, 2014).

Bollen, J., H. Mao and A. Pepe. 2010. Determining the public mood state by analysis of microblogging posts. *Proceedings of the 12th International Conference on the Synthesis and Simulation of Living Systems*. Odense, Denmark, 2010. Available at: http://pti.iu.edu/pubs/determining-public-mood-state-analysis-microblogging-posts (accessed May 7, 2014).

boyd, d. and Kate Crawford. 2012. Critical questions for Big Data. *Information, Communication & Society* 5(15): 662-679.

Brivot, M. and Y. Gendron. 2011. Beyond panopticism: On the ramifications of surveillance in a contemporary professional setting. *Accounting, Organizations and Society* 36: 135–155.

Brown, B., M. Chui, and J. Manyika. 2011. Are you ready for the era of 'big data'? Trend report McKinsey Global Institute. Available at: http://lonerganpartners.com/sites/lonerganpartners/files/Article%20PDFs/are-you-ready-for-era-of-big-data-10-2011-mckinsey.pdf (accessed May 7, 2014).

Bucher, T. 2012. Want to be on the top? Algorithmic power and the threat of invisibility on Facebook. *New Media & Society* 14 (7): 1164-1180.

Cha, M., H. Haddadi, F. Benevenuto, and K. P. Gummadi. 2010. Measuring user influence in Twitter: The million dollar fallacy. Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media. Available at http://pdfcast.org/pdf/measuring-user-influence-in-twitter-the-million-follower-fallacy Last accessed June 12, 2012.

Cheney-Lippold, J. 2011. A new algorithmic identity. Soft biopolitics and the modulation of control. *Theory, Culture &Society* 28(6): 164-181

Diakopoulos, N. and D. A. Shamma. 2010. Characterizing debate performance via aggregated Twitter sentiment. Paper presented at the CHI Conference, April 10–15, 2010, Atlanta, GA. Available at http://dl.acm.org/citation.cfm?id=1753504 . (Accessed June 13, 2012).

Ding, Y., Y. Du, Y. Hu, Z. Liu, L. Wang, K. W. Ross, and A. Ghose. 2011. Broadcast yourself: Understanding YouTube uploaders. Paper presented at the Internet Measurement Conference, IMC'11, November 2–4, Berlin. Available at http://conferences.sigcomm.org/imc/2011/program.htm (accessed May 7, 2014).

Draper, N. 2012. Group Power: Discourses of Consumer Power and Surveillance in Group Buying Websites. *Surveillance & Society* 9(4): 394-407.

ENISA. 2012. Study on monetizing privacy. An economic model for pricing personal information." Published February 27. Available at: http://www.enisa.europa.eu/activities/identity-and-trust/library/deliverables/monetising-privacy (accessed May 30, 2012).

Garton-Ash, T. 2013. If Big Brother came back, he'd be a public-private partnership. *The Guardian*, 27 June, 2013. Opinion page.

Giglietto, F. L. Rossiand D. Bennato. 2012. The Open Laboratory: Limits and Possibilities of Using Facebook, Twitter, and YouTube as a Research Data Source. *Journal of Technology in Human Services* 30(3-4): 145-159.

Gillespie, T. 2010. The politics of platforms. *New Media & Society* 12(3): 347-64.

Gitelman, L., ed. 2013. *'Raw Data' is an Oxymoron*. Cambridge, MA: MIT Press.

Goel, V. and E. Wyatt. 2013. Facebook Privacy Change Is Subject of F.T.C. Inquiry. *The New York Times*, 11 September 2013. Available at: http://www.nytimes.com/2013/09/12/technology/personaltech/ftc-looking-into-facebook-privacy-policy.html (accessed May 7, 2014).

Hartzog, W. and E. Selinger. 2013. Big Data in Small Hands. *Stanford Law Review Online Perspectives* 66 (81): 3 September. Available at: http://www.stanfordlawreview.org/online/privacy-and-big-data/big-data-small-hands (accessed May 7, 2014).

Helmond, A. and C. Gerlitz. 2013. The Like Economy. Social buttons and the data-intensive web. *New Media and Society*, Available online: http://nms.sagepub.com/content/early/2013/02/03/1461444812472322.abstract

Langlois, G. 2013. Participatory Culture and the New Governance of Communication. The Paradox of Participatory Media. *Television and New Media* 14 (2): 91-105.

Latour, B. 2007. Beware, your imagination leaves digital traces. *Times Higher Literary Supplement* April 6. Available at: http://www.bruno-latour.fr/node/245 (accessed May 7, 2014).

Lovink, G. 2012. *Networks Without a Cause. A critique of social media*. Cambridge: Polity Press.

Kosinski, M., D. Stillwell, and T. Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *PNAS,* online first, March 2013. Available at: http://www.pnas.org/content/early/2013/03/06/1218772110.full.pdf+html (accessed May 7, 2014).

Kwak, H., C. Lee, H. Park, and S. Moon. 2010. What is Twitter, a social network or a news media? *Proceedings of the 19th International World Wide Web (WWW) Conference*, April 26–30, Raleigh NC, 591–600. Available at http://an.kaist.ac.kr/traces/WWW2010.html l (accessed June 12, 2012).

Mager, A. 2012. Algorithmic Ideology. *Information, Communication & Society* 15(5): 769-787.

Mahrt, M. and M. Scharkow. 2013. The Value of Big Data in Digital Media Research. *Journal of Broadcasting & Electronic Media* 57 (1): 20-33.

Manovich, L. 2011. Trending: The Promises and the Challenges of Big Social Data. In: *Debates in the Digital Humanities*, ed. M.K. Gold, 460-475. Minneapolis: University of Minnesota Press.

Mayer-Schoenberger, V. and K. Cukier. 2013. *Big Data. A Revolution that will transform how we live, work, and think*. London: John Murray Publishers.

O'Connor, B., R. Balasubramanyan, B. R. Routledge, N. A. Smith. 2010. From Tweets to Polls: Linking Text Sentiment to Public Opinion. *Association for the Advancement of Artificial Intelligence* (www.aaai.org). 122-129. Available at: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/viewFile/1536/1842 (accessed May 7, 2014).

Oremus, W. 2013. With friends like these. How your friends, family, and co-workers are secretly helping social networks gather intelligence on you. *Slate.com*, June 26. Available at: http://www.slate.com/articles/technology/technology/2013/06/facebook_data_breach_how_social_networks_use_find_friends_to_mine_your_contacts.html (accessed May 7, 2014).

Pariser, E. 2011. *The Filter Bubble: What the Internet Is Hiding from You*. New York: Viking.

Pierson, J. 2012. Online Privacy in Social Media: A Conceptual Exploration of Empowerment and Vulnerability. *Communications and Strategies* 88 (4): 99-120.

Quercia, D., M.. Kosinski, D. Stillwell and J. Crowcroft. 2011. Our Twitter profiles, our selves: Predicting personality with Twitter. *IEEE International Conference on Social Computing*, Boston, 9-11 October. Available at http://ieeexplore.ieee.org/xpls/abs_all.jsp (accessed May 7, 2014).

Raley, R. 2013. Dataveillance and Countervailance. In: *'Raw Data' is an Oxymoron,* ed. L. Gitelman, 121-146. Cambridge, MA: MIT Press.

Richardson, M. 2008. Learning about the world through long-term query behavior. ACM TWeb 2 (4). Available at: ACM Digital Library http://dl.acm.org/citation.cfm?id=1409224 (accessed May 7, 2014).

Risen, J. and N. Wingfield. 2013. Silicon Valley and Spy Agency Bound by Strengthening Web. *The New York Times*, 19 June.

Rushe, D. 2013. Zuckerberg: US Government 'blew it' on NSA surveillance." *The Guardian* 11 September 2013. Available at: http://www.theguardian.com/technology/2013/sep/11/yahoo-ceo-mayer-jail-nsa-surveillance (accessed May 7, 2014).

Sakaki, T., M. Okazaki and Y. Matsuo. 2010. Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. *International World Wide Web Conference Proceedings,* April 26–30, 2010, Raleigh, North Carolina, USA. ACM 978-1-60558-799-8/10/04.

Schmidt, E. and J.Cohen. 2013. *The New Digital Age. Reshaping the Future of People, Nations and Business*. New York: Knopf.

van Dijck, J. 2005. *The Transparent Body. A Cultural Analysis of Medical Imaging*. Seattle: University of Washington Press.

van Dijck, J. 2013a. *The Culture of Connectivity. A Critical History of Social Media.* New York: Oxford University Press.

van Dijck, J. 2013b. 'You have one identity': Performing the self on Facebook and LinkedIn. *Media, Culture & Society* 35(2): 199–215.

van Dijck, J. and T. Poell. 2013. Understanding Social Media Logic. *Media and Communication* 1 (1): 2-14.

Weerkamp, W. and M. de Rijke. 2012. Activity Prediction: A Twitter-based exploration. *SIGIR Workshop on Time-aware Information Access*. August 16, Portland. Available at: http://wouter.weerkamp.com/downloads/taia2012-activity-prediction.pdf (accessed May 7, 2014).

Wilson, N., K. Mason, M. Tobias, M. Peacey, QS Huang, and M. Baker. 2009. Interpreting Google flu trends data for pandemic H1N1 influenza: the New Zealand experience. *European Communicable Disease Bulletin* 14 (44): 429-433.