Commentary

Datatrust: Or, the political quest for numerical evidence and the epistemologies of Big Data

Gernot Rieder[1,2] and Judith Simon[1,2]

Abstract

Recently, there has been renewed interest in so-called evidence-based policy making. Enticed by the grand promises of Big Data, public officials seem increasingly inclined to experiment with more data-driven forms of governance. But while the rise of Big Data and related consequences has been a major issue of concern across different disciplines, attempts to develop a better understanding of the phenomenon's historical foundations have been rare. This short commentary addresses this gap by situating the current push for numerical evidence within a broader socio-political context, demonstrating how the epistemological claims of Big Data science intersect with specific forms of trust, truth, and objectivity. We conclude by arguing that regulators' faith in numbers can be attributed to a distinct political culture, a representative democracy undermined by pervasive public distrust and uncertainty.

Over the past few years, there has been growing interest in so-called 'evidence-based policy making'. While the concept is not new (Solesbury, 2002), the latest push for more data-driven modes of governance has been considerable (Haskins, 2014). Against the backdrop of multiple crises, policymakers seem ever more inclined to legitimize specific ways of action by referring to 'hard' scientific evidence suggesting that a particular initiative will eventually yield the desired outcomes (Urahn, 2015). Across many areas of public service—be it healthcare, education, or law enforcement—a steady influx of "data for policy" (European Commission (EC), 2015a) is meant to offer guidance in a moment marked by high levels of complexity and uncertainty (Nowotny et al., 2001).

Legislators' current emphasis on evidence and results correlates with a recent techno-scientific development—the advent of Big Data (Mayer-Schönberger and Cukier, 2013). While state bureaucracies have relied on statistics and numerical information for centuries (Cohen, 2005), new analytical techniques promise to improve upon former methods in several ways: Whereas data analysis has traditionally been costly and time-consuming, it is now fast and cheap; whereas previously one had to settle for samples, the ongoing computerization of society makes it possible to glean data from entire populations; whereas once there was need for theory, through sheer volume the data now speak for themselves; whereas in the past measurements were tainted by human bias, agnostic algorithms now guarantee an impartial view from nowhere. Together, the alleged qualities of Big Data technologies feed into what Rob Kitchin (2014a) has described as the "articulation of a new empiricism", which operates as a "discursive rhetorical device" designed to promote the utility and value of new analytical services.

Policymakers on either side of the Atlantic have bought into the hype, usually without much regard

[1]IT University of Copenhagen, Denmark
[2]University of Vienna, Austria

Corresponding author:
Gernot Rieder, IT University of Copenhagen, Rued Langgaards Vej 7, 2300 Copenhagen S, Denmark.
Email: gernot.rieder@univie.ac.at

for nuance or subtlety. In official documents and speeches, Big Data is referred to as the "new oil of the digital age" (EC, 2012), the next "industrial revolution" (EC, 2014b), "gold" (EC, 2014a), a game-changing "key asset" (EC, 2015b) for creating value, increasing productivity, and boosting growth. The technology is not only expected to improve public administration by "advanc[ing] government efficiency" (Executive Office of the President (EOP), 2014) and enabling "better services" (EC, 2013), but also to support "evidence-informed decision making" (EC, 2015a) by providing real-time feedback, generating solutions, and predicting outcomes, always ensuring that "regulation is empirically justified in advance" (Sunstein, 2012). Although this focus on technology-driven benefits has in some cases expanded to include consideration of potential risks and pitfalls, political leaders remain firmly committed to "harness[ing] the power of Big Data" (Kalil and Zhao, 2013).

Much effort has already gone into challenging the buzz-laden assumptions of modern-day "data-ism" (Brooks, 2013). Investigating both the politics and power of contemporary data practices, scholars from different disciplinary backgrounds have identified a range of social, ethical, and legal issues—from privacy and security (Ohm, 2010) to transparency and accountability (Pasquale, 2015), bias and discrimination (Barocas and Selbst, 2015)—emphasizing that Big Data's presumed benefits may come at a cost. But while there has been a steady stream of critical reactions across academia and the media, attempts to gain a better understanding of the socio-historical foundations of policymakers' push for numerical evidence have been rare. Put differently, even though the rise of Big Data and related consequences has been a major issue of concern, its significance for and embeddedness in a long-standing culture of measurement and quantification has not. As Barnes (2013) poignantly states: "Big Data, little history."

One reason for this lack of historical contextualization can be attributed to the dynamics of Big Data discourse: Presented as a rupture and revolution with no ties to the past, discussions about Big Data have focused on the modalities of change rather than forms of continuity. The 'now' is said to be fundamentally different from what came before, the 'new' supersedes the 'old'. This narrative of novelty and disruption, exemplified in notions such as Anderson's (2009) "Petabyte Age", is both powerful and convenient, but discourages appreciation of Big Data as a specific amalgamation, a "conjuncture of different elements, each with their own history, coming together at this our present moment" (Barnes, 2013). Yet it is precisely the recognition of Big Data's diverse roots, its connection to prior epistemic practices, that may provide greater insight into the current excitement's underlying norms and values.

Such exploratory analysis requires some conceptual rethinking: Instead of narrowly defining Big Data in mere technical terms—e.g., Laney's (2001) popular 'three Vs', which reductively characterize Big Data as an increase in (data) volume, velocity, and variety—it seems more productive to think of it as the terminologically contingent manifestation of a complex socio-technical phenomenon that rests on an interplay of technological, scientific, and cultural factors (cf. boyd and Crawford, 2012). While the *technological dimension* alludes to advances not only in hardware, software, but also infrastructure and the *scientific dimension* comprises both mining techniques and analytical skills, the *cultural dimension* refers to (a) the pervasive use of ICTs in contemporary society and (b) the growing significance and authority of quantified information in many areas of everyday life, including public administration and decision making. Ultimately, this broader interpretative approach may assist in "deconstructing the black boxes of Big Data" (Pasquale, 2015) by paying attention not only to the mechanical, but also to the mental workings of an otherwise opaque phenomenon.

Investigations into the roots and antecedents of Big Data may take different paths: Barnes and Wilson (2014), for instance, examine the origins of the social physics movement, whose monistic urge—that is, the assumption that the laws of physics apply to both natural and social worlds—was later incorporated into spatial analysis, shaping the use of Big Data in present-day geography. Morozov (2014), drawing on Medina's (2011) *Cybernetic Revolutionaries*, details the Allende administration's Project Cybersyn to highlight the intellectual affinities between socialism, cybernetics, and Big Data, while Grandin (2014), citing Dingens (2005), reports on the Pinochet regime's Condor data bank to locate the "anti-socialist origins of Big Data", a juxtaposition of historical events illustrating that the idea of data-facilitated control may in fact appeal to different ends of the political spectrum. Last but not least, Mackenzie (2013) provides an empirical account of how recent shifts in programming practice relate to what Adams et al. (2009) have labeled "regimes of anticipation", demonstrating how the current emphasis on machine learning and predictive modelling is entangled with a concerted cultural effort to reduce uncertainty by fostering the continuous assessment of the 'not yet'.

While these examples offer unique perspectives, each focusing on particular cases and ideas, they are also similar in that they seek to situate Big Data discourse within a larger historical context, attributing meaning to what all too often takes the form of pure marketing. We suggest that such attempts to historicize and contextualize are crucial as they may (a) provide better

insight into the epistemological foundations of contemporary data science, (b) deepen our understanding of the norms, values, and expectations driving the current climate of hope and hype, and (c) indicate potential social and ethical ramifications, serving as a guiding compass at a time when technical innovation continues to outpace government regulation (Rubinstein, 2013). We would like to contribute to this research agenda by suggesting what may prove another fruitful avenue of investigation: The data hype's reliance on specific forms of trust, truth, and objectivity.

As boyd and Crawford (2012) have argued, Big Data is not just about technological progress, but about a "widespread belief that large data sets offer a higher form of intelligence and knowledge that can generate insights that were previously impossible". Leonelli (2014) makes a similar argument, stressing that the novelty of Big Data science does not lie in the sheer quantity of data involved, but in the "prominence and status acquired by data as commodity and recognized output." But where does this prominence and status come from and what exactly are the roots of the belief that more data equals better insight?

An initial answer would be that data are often perceived as raw, objective, and neutral—the "stuff of truth itself" (Gitelman, 2013). But, as historians of science and technology have repeatedly shown, conceptions of objectivity, truth and truthfulness, trust and trustworthiness may vary, they are "situated and historically specific" (Gitelman, 2013). Therefore, it is important to clarify which particular version of these concepts manifests within Big Data discourse. One way to identify such differences is through comparison, which may involve tracing conceptual shifts and changes over time.

In his book *The Social History of Truth*, Shapin (1994) emphasizes the central role of trust in building and maintaining social order. Societies are made through acts of trust—without trust, they may falter and collapse. The allocation of trust and trustworthiness can thus be understood as the "great civility", granting the conditions in which people can colonize each others' mind. Although often rendered invisible, trust as the "cement of society" is also essential to the construction and establishment of epistemic systems. The production of scientific knowledge, for instance, rests on myriads of social and material interactions, which take for granted the reliability of numerous stabilized norms and relationships. As a result, scientific distrust and skepticism only takes place "on the margins of trusting systems."

But such systems of trust are not fixed—conceptions of whom to trust, what to trust, and in what circumstances, are subject to change: While in premodern society it was the politically and economically independent gentleman who was generally conceived as a credible truth-teller, modern society accorded trust to the "abstract capacities" (Giddens, 1990) of "faceless institutions" (Shapin, 1994). The veracity of testimony was no longer underwritten by personal virtue, but by an elaborate system of institutionalized norms and standards, rigorously policed in a great "panopticon of truth" (Shapin, 1994). A different form of trust first accompanied and then superseded the premoderns' faith in the integrity of the solitary knower and the moderns' confidence in the rigor of institutionalized expertise, a type of trust that has gained considerable traction with the arrival of Big Data: people's trust in numbers.

While the general history of quantification can be traced back much further, Desrosières (1998) identifies 17th-century English political arithmetic as the "basic act of all statistical work (in the modern sense of the term), implying definite, identified, and stable unities." Whereas early records of baptisms, marriages, and burials were meant to attest to the existence of individuals and their family relations, later statistical surveys such as the one underlying the 18th-century French "adunation" were intended to support the unification of national territory in order to establish a "politico-cognitive construction of a space of common measurement". Examples such as these highlight the close relationship between statistics and state-making: Numbers allowed for coherence and generality, enabling central governments to exercise administrative control over matters of taxation and economic development at a time when the familiarity of face-to-face interactions gradually gave way to the anonymity and complexity of expanding trade and business networks.

But behind those numbers still stood individual experts and prestigious institutions—numbers did not speak for themselves. Quite to the contrary, it was the cultivated judgment of an administrative elite that guaranteed the trustworthiness of numerical information; deployed by outsiders, statistics counted for little. As Porter (1995) explains, numbers could only "provide a modest supplement to institutional power." Their credibility rested on the authority and integrity of a bureaucracy whose members believed that measurements only became useful when subject to expert interpretation. For them, nothing could be reduced to inflexible laws, abstract formulas, or technical routines. Agreements were reached through informal discussion rather than formal procedures. In general, decisions were rarely entrusted to the numbers.

The demand for quantitative rigor increased during the first half of the 20th century: Instead of expert judgment, the pursuit of technical discipline required an "ideal of self-sacrifice"; instead of professional autonomy, the desire for precision imposed adherence to a strict "regime of calculation"; instead of elite

discretion, it became necessary to "manage by the numbers" (Porter, 1995). The result was what Porter refers to as the "cult of impersonality", a specific culture of quantification that seeks to reduce the human element as much as possible, preferring formalized principles to subjective interpretation, uniform standards to methodological tinkering, the rule of law to the rule of men. The goal was to attain "mechanical objectivity" (Daston and Galison, 1992), a disinterested science that "eradicates all that is personal, idiosyncratic, perspectival." In this brave new world, trust no longer resides in the integrity of individual truth-tellers or the veracity of prestigious institutions, but is placed in highly formalized procedures enacted through disciplined self-restraint. Numbers cease to be supplements. They are couched in a rhetoric of factuality, imbued with an ethos of neutrality, and presented with an aura of certainty. They step out of the shadows of their human creators, enter center stage, and, in the arguments and claims of countless profiteers, start to speak for themselves.

What are the reasons for this shift toward mechanical objectivity? On the one hand, technological progress played a significant role. The growing availability of ever more capable machinery changed the face of the accounting profession. The idea was powerful: The more mechanized a process, the more automated a procedure, the less the need for—and danger of—subjective human intervention (Venturini et al., 2014). In the words of Daston and Galison (2010), "instead of freedom of will, machines offered freedom from will". The virtuous machine was conceived as the "ultimate outsider", and it was not long until "it became the greatest in the kingdom of quantification" (Porter, 1995). Consequently, the "honest instrument" with its "glow of veracity" both served as a means to and symbol of mechanical objectivity (Daston and Galison, 1992, 2010).

On the other hand, there was a social dimension: The pursuit of quantitative rigor was seen as a strategy to adapt to new external pressures in a rapidly changing political environment. War and economic crisis had left their marks, and the dynamics of democracy increased the need for hard evidence and professional accountability. Confronted with public distrust, invasive auditing, and competing political demands, bureaucratic agencies and scientific communities sought to withstand scrutiny and minimize responsibility by adhering to rigid protocols and explicit decision criteria. This willingness for personal restraint is a sign of professional weakness rather than strength: The more permeable the boundaries of a discipline, the higher its vulnerability to outside criticism, the more tempting the language of mechanical objectivity becomes. Consequently, the appeal of standardized methods is especially great in cultures where the faith in other forms of trust has been shattered. As Porter (1995) notes, methodological strictness and objective rules may serve as an alternative to trust and shared beliefs. Where trust is missing and suspicion prevails, numbers are meant to fill the gap: Regarded as carefully measured matters of fact, they are expected to offer a sense of fairness and justice, a way of making decisions without having to decide, a chance to de-politicize legislation. This push for impersonal numerical evidence is however not so much rooted in the inner workings of quantitative professions, but in the needs and demands of a specific socio-political culture, a democratic system undermined by pervasive distrust and uncertainty. It is on these grounds that the Big Data phenomenon continues to blossom.

The epistemic promises of Big Data connect to the ideal of mechanical objectivity in several ways, not only fortifying but also expanding the appeal of the doctrine:

First, a child of new analytical techniques and the progressing computerization of society, Big Data pledges to extend the reach of automation, from data collection to matters of storage, curation, and analysis. The virtuous machine emerges as ever more powerful as it covers increasingly large parts of the analytical and decision-making process.

Second, by capturing massive amounts of data and focusing on correlations rather than causes, Big Data claims to reduce the need for theory, models, and, in extension, human expertise. In addition, modern data analysis software is often thoroughly opaque, with a phenomenology that emphasizes both uniformity and impersonality.

Third, Big Data promises to expand the realm of what can be measured. Trackers, social media, and the Internet of Things allow to trace and gauge movements, actions, and behaviors in ways that were previously unfeasible. Fully quantified and free from bias, Big Data pushes the tenets of mechanical objectivity into ever more areas of application.

Fourth and finally, settling for neither the present nor the past, Big Data aspires to calculate what is yet to come. Smart, fast, and cheap predictive techniques are meant to support decision making and optimize resource allocation across many government sectors, applying a mechanical mindset to the colonization of the future.

The limitations of these "sociotechnical imaginaries" (Jasanoff and Kim, 2009) have been discussed elsewhere (e.g., Kitchin, 2014b), but the point here is to develop a better understanding of how the current language of Big Data-related hope and hype intersects with and relies on particular forms of trust and objectivity, which, in turn, can be conceived as products of a specific socio-political culture. In a climate of distrust, crisis, and uncertainty, officials' adherence to supposedly impartial numbers may be regarded as a strategy of defense, an attempt

to shield themselves from increased public and judicial scrutiny. It is not by coincidence that the European Commission, whose authority continues to be challenged by citizens and national governments alike, has emerged as one of the most zealous political quantifiers.

Big Data has been repeatedly criticized for its positivist epistemology and its support of techno-capitalism, and while such criticism has its merits, it pays little attention to the circumstances and dynamics that contribute to the creation and internalization of corresponding norms and values. Our proposition is simple: Instead of focusing exclusively on the potential consequences of the Big Data phenomenon, we can gain additional insight from examining its social and political, but also its technical and epistemic roots. Such an approach may foster more, not less, critical engagement as it shifts the perspective and situates Big Data discourse within a broader historical narrative. As Barnes and Wilson (2014) argue:

> By showing that Big Data is historical, we show the assumptions that were built into it, as well as the contestations around them. Big Data becomes no longer a black box, self-contained, sealed and impregnable, but is opened up, available for verbalist discussion and contestation.

We wholeheartedly agree.

Declaration of conflicting interests

Funding

References

Adams V, Murphy M and Clarke AE (2009) Anticipation: Technoscience, life, affect, temporality. *Subjectivity* 28(1): 246–265.

Anderson C (2009) The end of theory: The big data deluge makes the scientific method obsolete. *Wired*. Available at: http://www.wired.com/2008/06/pb-theory/ (accessed 2 May 2016).

Barnes TJ (2013) Big Data, little history. *Dialogues in Human Geography* 3(3): 297–302.

Barnes TJ and Wilson MW (2014) Big Data, social physics, and spatial analysis: The early years. *Big Data & Society* 1(1): 1–14.

Barocas S and Selbst AD (2015) Big Data's disparate impact. *California Law Review* (forthcoming 2016). Available at: http://ssrn.com/abstract = 2477899 (accessed 2 May 2016).

boyd D and Crawford K (2012) Critical question for Big Data. Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society* 15(5): 662–679.

Brooks D (2013) The philosophy of data. *New York Times*. Available at: http://www.nytimes.com/2013/02/05/opinion/brooks-the-philosophy-of-data.html (accessed 2 May 2016).

Cohen BI (2005) *Triumph of Numbers: How Counting Shaped Modern Life*. New York, NY: W. W. Norton & Company.

Daston L and Galison P (2010) *Objectivity*. New York, NY: Zone Books.

Daston L and Galison P (1992) The image of objectivity. *Representations* 0(40): 81–128.

Desrosières A (1998) *The Politics of Large Numbers. A History of Statistical Reasoning*. Cambridge, MA: Harvard University Press.

Dingens J (2005) *The Condor Years: How Pinochet and his Allies Brought Terrorism to Three Continents*. New York, NY: The New Press.

European Commission (EC) (2012) *From Crisis of Trust to Open Governing*. Available at: http://europa.eu/rapid/press-release_SPEECH-12-149_en.htm (accessed 2 May 2016).

European Commission (EC) (2013) *EU Leaders Call for Action on Digital Economy, Innovation and Services*. Avaialble at: http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=3369 (accessed 2 May 2016).

European Commission (EC) (2014a) *The Data Gold Rush*. Available at: http://europa.eu/rapid/press-release_SPEECH-14-229_en.htm (accessed 2 May 2016).

European Commission (EC) (2014b) *Towards a Thriving Data-Driven Economy*. Available at: http://ec.europa.eu/information_society/newsroom/cf/dae/document.cfm?doc_id = 6210 (accessed 2 May 2016).

European Commission (EC) (2015a) *Data for Policy: When the Haystack Is Made of Needles. A Call for Contributions*. Available at: http://ec.europa.eu/digital-agenda/en/news/data-policy-when-haystack-made-needles-call-contributions (accessed 2 May 2016).

European Commission (EC) (2015b) *Making Big Data Work for Europe*. Available at: https://ec.europa.eu/digital-agenda/en/big-data (accessed 2 May 2016).

Executive Office of the President (EOP) (2014) *Big Data: Seizing Opportunities, Preserving Values*. Available at: http://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_5.1.14_final_print.pdf (accessed 2 May 2016).

Giddens A (1990) *The Consequences of Modernity*. Cambridge: Polity Press.

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

Grandin G (2014) The anti-socialist origins of big data. *The Nation*. Available at: http://www.thenation.com/article/anti-socialist-origins-big-data/ (accessed 2 May 2016).

Haskins R (2014) *Show Me the Evidence: Obama's Fight for Rigor and Results in Social Policy*. Washington, DC: The Brookings Institution.

Jasanoff S and Kim S-H (2009) Containing the atom: Sociotechnical imaginaries and nuclear power in the United States and South Korea. *Minerva* 47(2): 119–146.

Kalil T and Zhao F (2013) Unleashing the Power of Big Data. *The White House Blog*. Available at: https://www.white-house.gov/blog/2013/04/18/unleashing-power-big-data (accessed 2 May 2016).

Kitchin R (2014a) Big Data, new epistemologies and paradigm shifts. *Big Data & Society* 1(1): 1–12.

Kitchin R (2014b) *The Data Revolution. Big Data, Open Data, Data Infrastructures & Their Limitations*. London: Sage.

Laney D (2001) 3D management: Controlling data volume, velocity and variety. Available at: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf (accessed 2 May 2016).

Leonelli S (2014) What difference does quantity make? On the epistemology of big data in biology. *Big Data & Society* 1(1): 1–11.

Mackenzie A (2013) Programming subjects in the regime of anticipation: Software studies and subjectivity. *Subjectivity* 6(4): 391–405.

Mayer-Schönberger V and Cukier K (2013) *Big Data. A Revolution That Will Transform How We Live, Work, And Think*. New York, NY: Houghton Mifflin Harcourt.

Medina E (2011) *Cybernetic Revolutonaries: Technology and Politics in Allende's Chile*. Cambridge, MA: MIT Press.

Morozov E (2014) The planning machine. *New Yorker*. Available at: http://www.newyorker.com/magazine/2014/10/13/planning-machine (accessed 2 May 2016).

Nowotny H, Scott P and Gibbons M (2001) *Re-thinking Science: Knowledge and the Public in an Age of Uncertainty*. Cambridge: Polity Press.

Ohm P (2010) Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review* 57: 1701–1777.

Pasquale F (2015) *The Black Box Society: The Secret Algorithms that Control Money and Information*. Cambridge, MA: Harvard University Press.

Porter TM (1995) *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton, NJ: Princeton University Press.

Rubinstein IS (2013) Big Data: The end of privacy or a new beginning? *International Data Privacy Law* 3(2): 74–87.

Shapin S (1994) *A Social History of Truth: Civility and Science in Seventeenth-Century England*. Chicago: The University of Chicago Press.

Solesbury W (2002) The ascendancy of evidence. *Planning Theory & Practice* 3(1): 90–96.

Sunstein CR (2012) *Regulation in an Uncertain World*. National Academy of Sciences. Available at: https://www.whitehouse.gov/sites/default/files/omb/inforeg/speeches/regulation-in-an-uncertain-world-06202012.pdf (accessed 2 May 2016).

Urahn SK (2015) A tipping point on evidence-based policymaking. *Governing*. Available at: http://www.governing.com/columns/smart-mgmt/col-state-local-government-tipping-point-evidence-based-policymaking.html (accessed 2 May 2016).

Venturini T, Laffite NB, Cointet J-P, et al. (2014) Three maps and three misunderstandings: A digital mapping of climate diplomacy. *Big Data & Society* 1(1): 1–19.

This article is a part of Special theme on Critical Data Studies. To see a full list of all articles in this special theme, please click here: http://bds.sagepub.com/content/critical-data-studies.