Techno-optimism and policy-pessimism in the public sector big data debate

Vydra, Simon; Klievink, Bram

DOI
10.1016/j.giq.2019.05.010

Publication date
2019

Document Version
Final published version

Published in
Government Information Quarterly

Government Information Quarterly xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Government Information Quarterly

journal homepage: www.elsevier.com/locate/govinf

Techno-optimism and policy-pessimism in the public sector big data debate

Simon Vydra[a,b,*], Bram Klievink[b]

[a] Faculty of Technology, Policy and Management, Delft University of Technology, Jaffalaan 5, 2628 BX Delft, Netherlands
[b] Faculty of Governance and Global Affairs, Leiden University, Turfmarkt 99, 2511 DP Den Haag, Leiden, Netherlands

ARTICLE INFO

Keywords:
Big data
Analytics
Government
Public administration
Policy-making
Decision-making
Science-policy interface
Network governance

ABSTRACT

Despite great potential, high hopes and big promises, the actual impact of big data on the public sector is not always as transformative as the literature would suggest. In this paper, we ascribe this predicament to an overly strong emphasis the current literature places on technical-rational factors at the expense of political decision-making factors. We express these two different emphases as two archetypical narratives and use those to illustrate that some political decision-making factors should be taken seriously by critiquing some of the core 'techno-optimist' tenets from a more 'policy-pessimist' angle. In the conclusion we have these two narratives meet 'eye-to-eye', facilitating a more systematized interrogation of big data promises and shortcomings in further research, paying appropriate attention to both technical-rational and political decision-making factors. We finish by offering a realist rejoinder of these two narratives, allowing for more context-specific scrutiny and balancing both technical-rational and political decision-making concerns, resulting in more realistic expectations about using big data for policymaking in practice.

1. Introduction

Despite the elusiveness of the concept of 'big data', its potential to change business and politics is well established in the literature. Some even argue that we are entering a second machine age, implying that computers and big-data-enabled analysis remove mental power constraints much like the invention of the steam engine removed physical power constraints (Brynjolfsson & McAfee, 2014). For social science specifically, the impact of big data can arguably "be compared with the impact of the invention of the telescope for astronomy and the invention of the microscope for biology (providing an unprecedented level of fine-grained detail)" Hilbert, 2016, p. 136). The public sector has not avoided this wave of big data optimism with some authors arguing that not only do "public bodies using big data achieve significantly more positive outcomes and benefits" (Maciejewski, 2016, p. 127), but also that big data "will profoundly change how governments work and alter the nature of politics" (Cukier & Mayer-Schoenberger, 2013, p. 35).

Notwithstanding the high hopes, the adoption of big data appears to be a slow and uneven process that takes different forms and happens at different speeds based on the institutional and policy context (Klievink, Romijn, Cunningham, & de Bruijn, 2017). This is observable globally as certain policy areas see much more big data use than others, but also in regional case-studies that often conclude that "there is still little knowledge of the conditions and determinants for its [big data's] application, especially in public policy domain" (Misuraca, Mureddu, & Osimo, 2014, p. 176), or that "we cannot fully account for the lack of widespread diffusion of the innovative localized [big data] use practices" (Chatfield & Reddick, 2018, p. 346). We ascribe this predicament to a strong emphasis the current literature places on technical-rational factors, which are insufficient to explain the diffusion of big data analytics as adopting IT solutions in public administrations resembles a "mixture of political behaviour, intuition and the exploitation of emerging opportunities, whereas technical rationality plays a minor role" (Nielsen & Pedersen, 2014, p. 419). As a result, the existing literature struggles to explain the uneven adoption of big data analytics for policymaking and the (lack of) change to policymaking practice this entails.

There seem to be two archetypical narratives present in the existing literature. First, a narrative focused on the study of big data analytics as a technological phenomenon, focusing on its comparative (dis)advantages to how 'traditional' data is created, handled, and analysed, often rooted in engineering and computer science disciplines (see for example Dong et al., 2017; Dumbacher & Hutchinson, 2016; Ku & Leroy, 2014; Misuraca et al., 2014). Second, a narrative focusing on decision-making and the study of how quantitative evidence and the advent of big data interacts with political and bureaucratic decision-making, often rooted in public administration and organisational decision-making disciplines (see for example Desouza & Jacob, 2014;

* Corresponding author at: Faculty of Technology, Policy and Management, Delft University of Technology, Jaffalaan 5, 2628 BX Delft, Netherlands
E-mail address: S.Vydra@tudelft.nl (S. Vydra).

https://doi.org/10.1016/j.giq.2019.05.010
Received 11 May 2018; Received in revised form 28 January 2019; Accepted 26 May 2019

Dunleavy, Margetts, Bastow, & Tinkler, 2005; Giest, 2017; Janssen & Kuk, 2016a, 2016b; Klievink et al., 2017). If we would put these two narratives to the extreme – by limiting our focus purely to technology or political decision-making and accepting the underlying assumptions of these narratives as axioms - we could argue that the first narrative is optimist and the latter is pessimist with regards to the impact of big data on policymaking. We attribute this difference to the fact that technology evolves and is adopted very rapidly compared to how slowly political and governance practices change, making the technical narrative optimistic and the policy and decision-making narrative pessimistic about the magnitude of change big data will have on public sector and governance in general. We term these two extremes 'techno-optimism' and 'policy-pessimism'.

Even though these two narratives differ primarily in focus and optimism, this difference translates to important aspects of talking about big data, including something so fundamental as how we define it: The most common big data definition uses a set of 'Vs' – attributes along which big data differs from 'normal' data. Most commonly these V's are volume, variety, velocity, and veracity (IBM, 2012; Ward & Barker, 2013), but sometimes also include variability, visualisation, and value (for review of definitions see Ylijoki & Porras, n.d.). This way of defining big data itself seems to be rather techno-optimist, as the attributes are primarily technical and describe the nature of the data itself (except visualisation and value, which are not commonly used). The policy-pessimist definitions of big data revolve around the social change big data motivates, especially in terms of changes to decision-making processes necessary to make use of big data (Kim, Trimi, & Chung, 2014). These definitions refer to the usage of structured and unstructured data (potentially in combination) from multiple sources both internal and external to an institution, the use of high-frequency data streams, and the use of data for radically different purposes than it was originally intended for (if there was an intent to begin with) (Klievink et al., 2017). Such definitions immediately emphasize the challenges of deriving relevant insight from data and how this insight is used by individuals in making decisions. This makes the two narratives differ not just in their focus but in terms of the fundamental 'unit of analysis': Techno-optimism focuses on data and analytical output whereas policy-pessimism focuses on humans turning data into insight and humans making decisions in bureaucratic structures (with the help of that insight).

Much of the literature on big data in the public sector has ingredients of both narratives, yet whether conscious or not, tends to emphasize or be based on one of them. As alluded to earlier, the emphasis currently seems to be on the techno-optimist side. Yet, we do acknowledge that an unequivocal distinction is very hard to make, as even rather techno-optimist accounts pay lip service to decision-making and politics (Höchtl, Parycek, & Schöllhammer, 2016; Maciejewski, 2016). In fact, even the more policy-pessimist accounts pay lip service to the big data promise and do not dismiss it outright (Iacus, 2015; Lavertu, 2016). Thus, despite our diagnosis of a techno-optimist bias, it is important to note that majority of the existing contributions are not openly and unequivocally techno-optimist and they do address relevant shortcomings, but do not do so systematically or comprehensively (Bertot & Choi, 2013; Chatfield & Reddick, 2018; Einav & Levin, 2013; Katal, Wazid, & Goudar, 2013; Ku & Leroy, 2014; Misuraca et al., 2014; Sagiroglu & Sinanc, 2013). The result of offering only lip service to (as opposed to systematically addressing) the 'opposing' perspective and cherry-picking easy-to-address concerns is that many contributions talk past one another, rendering the existing literature incapable of explaining why is the diffusion of big data analytics in the public sector uneven on so many levels. We aim to help this predicament in two ways: Firstly, we challenge key techno-optimist arguments from a more policy-pessimist lens, thus illustrating its value for interrogating big data use in the public sector. Secondly, we structure the current debate by articulating these two archetypical narratives and making them meet 'eye-to-eye' with the ambition of helping scholars to interrogate their work more systematically.

To do so we need to first disentangle the techno-optimist narrative into key arguments and assumptions, which in itself is a difficult task for two reasons: Firstly, because the benefits and shortcomings of big data articulated in the literature are numerous and how these should be aggregated into 'key arguments and assumptions' is not obvious. Secondly, since existing literature situates itself between the two extremes but not directly on them, it is not possible to directly extract the archetypical techno-optimist narrative from a specific contribution. In other words, we construct techno-optimism as the logical extreme of arguments we identify in the literature, but our construction of techno-optimism and policy-pessimism remains a heuristic fit for the purpose of this paper rather than a robust categorization to sort the current literature by. That said, to provide a structure for his paper we disentangle these two archetypical narratives into four aspects of big data analysis they fundamentally differ on: Firstly, the quality of the data insight and subsequent decision-making. Secondly, the speed of data analysis and subsequent decision-making. Thirdly, the epistemological foundation for the analytics process. Fourthly, overcoming some of the fundamental concerns relevant to big data analytics (in this paper we focus on privacy as an exemplary concern). These four key arguments and assumption are selected because of how foundational they are to the big data in public sector debate (conceptually and in terms of being covered by existing literature), but the two opposing narratives can be constructed using less aggregate and more context-specific set of key arguments and assumptions. These four key arguments and assumptions will be addressed in sections two to five in the order listed above, with each section first briefly outlining the techno-optimist argument for that aspect followed by highlighting shortcomings of that argument. In section six we conclude by summarizing the techno-optimist and policy-pessimist narratives for the four key arguments and assumptions that we deal with in this paper, making the two narratives meet 'eye-to-eye' and highlighting some crucial questions to interrogate research with based on these narratives. In concluding we also offer our take on reconciling the two narratives in a 'best-of-both-worlds' fashion by adopting a more granular approach and focusing on specific big data sources and specific policy questions - a level of analysis at which trade-offs can be meaningfully made.

2. How bigger doesn't always mean better in public decision making

A fundamental argument of a techno-optimist narrative is that big data will provide better information and that this better information will in turn facilitate better decisions. The argument essentially claims that "[t]he more quality and accurate information is available, the better the decisions will be." (Höchtl et al., 2016, p. 152). This notion is based on understanding policy decisions as based largely on empirical input and improving this input then resulting in better regulatory policy (Maciejewski, 2016). This input can of course be (and often is) an estimate, leading some to argue that "[i]f we improve the basis of prior information on which to base our estimates, our uncertainty will be reduced on average. The better the prior, the better the estimate, the better the decision" (Hilbert, 2016, p. 135).

How exactly will data (and subsequently decision making) be "better" is often left unexplained, but some authors provide a bit of elaboration: Maciejewski (2016) argues that using big data methods results in more accurate decision-making due to expansion of databases, more extensive analytics, and better data visualisation and presentation (Maciejewski, 2016). Other authors focus on the overall efficiency gains in the private sector triggered by big data analytics, arguing that it is reasonable to expect similar developments in the public sector (Chen & Hsieh, 2014). In other words, the notion of 'better' can be related to an increase in accuracy (Höchtl et al., 2016), a reduction in uncertainty (Hilbert, 2016), or efficiency gains (Chen & Hsieh, 2014) and is applied to both the insight we can derive from data as well as the decision we make based on this insight.

In this section, we tackle both of the assumptions this argument rests on: That big data provides better insight and that better insight translates into better policy decisions. In Section 2.1 we point to the various important aspects of data quality that make it impossible for a big data source to be 'better' for policymaking in general. In Section 2.2 we point to factors other than data that influence the quality of public decision-making, thus complicating the link between better data and better decisions.

2.1. The myth of 'better' information

Taking accuracy and uncertainty as two aspects of data quality highlighted by the techno-optimist argument, it is important to point out that not all big data sets by default allow for more accurate insights: Firstly, big data sources often struggle with substantial representativeness problems that have been described both empirically and conceptually (Hargittai, 2015; Keith, Ginnis, & Miller, 2016; Liu, Li, Li, & Wu, 2016; Ruths & Pfeffer, 2014; Samarajiva & Lokanathan, 2016), making the resulting insight skewed and thus inaccurate in that sense. Secondly, big data often contain much more 'noise' than 'signal' and this noise has to be removed to arrive at reliable conclusions (Iacus, 2015; Scannapieco, Virgillito, & Zardetto, 2012; Vaccari, 2015), which in itself presents an analytical challenge that introduces inaccuracy (since it is impossible to perfectly distinguish signal and noise). Because of these issues, national statistical institutions are currently primarily focused on creating quality assurance processes for big data sources (Boettcher, 2015; Dumbacher & Hutchinson, 2016; Eurostat Big Data Task Force, 2014; Hackl, 2016), rather than actually using big data for official statistics and policymaking. In fact, when compared to traditional survey-based measures that can be crafted to accurately categorize every individual (Kitchin & Lauriault, 2015), accuracy of big data tends to become more of a concern rather than a demonstrable benefit.

More importantly, accuracy is not the only (and arguably not the most important) attribute of data for policymaking. When it comes to big data, "[Data] [q]uality is composed of several elements, such as accuracy, reliability, relevance or timeliness" (Eurostat Big Data Task Force, 2014, p. 13). When it comes to reliability, arguably the most important metric, big data often perform rather poorly. Reliability in this case refers to the trust policymakers have in a specific indicator, which is established by having a good track record of accuracy and relevance for policy questions (Kitchin & Lauriault, 2015). To amass such a track record, an indicator needs to have a good and long backrun (how far back is the data available for). Given how crucial *data backrun* is for establishing reliability (demonstrated by the Bank of England deciding not to use big data based on insufficient data backrun (McLaren & Shanbhogue, 2011)) and how overlooked the concept is in the current big data debate, it deserves more elaboration: There are broadly speaking four reasons for why data backrun is crucial for data quality. Firstly, better temporal coverage of a data set allows traditional statistical methods to generate better inferential leverage. Secondly, it provides crucial contextualization to any data insight: A 'spike' in an indicator is of little use unless we can compare it to historical data showing how these spikes play out in social reality and how they react to different policies. Thirdly, and perhaps most importantly, as crucial indicators build up a reliable backdrop they get institutionalized into domestic and international policymaking practice. This creates decades of negotiated knowledge between experts, politicians, and institutions on how to measure and adapt these concepts to assure their continuous usefulness (consider the ICLS conferences on Labour market statistics organized by the ILO (Hussmanns, 2007) as an example of such negotiated knowledge and institutionalization). Lastly, this institutionalization also achieves international comparability, which big data sources struggle with as they vary greatly from country to country and cannot really be controlled by a statistical institution since much of big data is privately owned. The example of data backdrop illustrates that the debate about data 'quality' is far more nuanced than a techno-optimist narrative sometimes conveys.

Besides the multiple dimensions of data quality, the argument has a more conceptual (but no less important) dimension: Information can be seen as a contested commodity (Peled, 2014; Ruppert, Isin, & Bigo, 2017) and the use of big data depends on context and governance practices which may be subject to shifting ideals, and the (social) concepts the data ought to represent may be "negotiated, abbreviated and contested" (Robertson & Travaglia, 2015, ¶ 6). In other words, data are objects of knowledge but also power (Ruppert et al., 2017), meaning they cannot be universally "better" in a non-partisan way (Goldston, 2008). This begs the question for whom is big data better, or alternatively for what purpose is it better for. In answering such a question it is vital to understand where and how data are produced, how data are used, and what gets lost along the way: To appreciate what big data analysis tells us, we also need to know what wasn't measured, what data got filtered out, and why (McNeely & Hahm, 2014). This is not just analytical good practice – in some cases politicians are more receptive to big data evidence if they understand how and from which specific group of individuals was the data gathered (Panagiotopoulos, Bowen, & Brooker, 2017). Furthermore, the generation, collection, storage, and processing of big data are done using information systems and algorithms that are perceived to be neutral (McNeely & Hahm, 2014), but are in fact part of bureaucratic systems and structures that are inherently political (Janssen & Kuk, 2016b). All these details make it impossible to assert that big data is somehow objectively 'better'.

2.2. How 'better' information translates to 'better' decisions

The transformation of insight from data into policy is by no means a straight forward process and the emergence of big data influences it as well as it does the data itself. The idea that better information leads to better decisions assumes a rather linear view of policy making, where information only enters at certain places, often represented in terms of a 'policy cycle' (Helbig, Dawes, Dzhusupova, Klievink, & Mkude, 2015). Yet, decisions and policy processes often do not work that way. Rather, they are the product of multiple, interacting actors, that are interdependent and are hard to commit to a common problem, solution or even the value of 'facts' (de Bruijn & Ten Heuvelhof, 2008). The result is a complex policy battle in which decision-making often takes place through small, incremental steps (Lindblom, 1959) and consist of several iterations between processes, making it a plate of spaghetti rather than a cycle (Klijn & Koppenjan, 2015). In the case of big data, the process between information and decisions is subject to politicization in at least two distinct ways.

Firstly, there is the issue described above; transforming big data into information and insights is not a politically neutral process much of which depends on who decides what data is worth, what is included, what is excluded, how data are aggregated, etc. Not to mention, these concerns can often be 'hidden' in complex algorithms and thus extremely difficult to interrogate (Janssen & Kuk, 2016b). Secondly, there are political decisions to be made not only in interpreting the data, but also in gathering it; the algorithms used to capture insights from big data reflect specific conceptions of social phenomena, including preconceptions about factors of importance, expected correlations, or contested assumptions. A telling example of these two points is the debate surrounding the COMPAS risk assessment algorithm meant to predict recidivism. The algorithm, despite not including race as an input, has been argued to label blacks who do not actually re-offend with higher risk scores than whites who do not re-offend and vice versa for those who do re-offend (Angwin, Larson, Mattu, & Kirchner, 2016). The company developing COMPAS as well academics have argued against this critique along technical and methodological lines (Dieterich, Mendoza, & Tim Brennan, 2016; Flores, Bechtel, & Lowenkamp, 2016), with authors of the original critique standing by their conclusions as a response (Angwin & Larson, 2016). The debate is largely technical, but there is an underlying disagreement about the

notion of 'fairness' and whether that refers to accurate calibration between groups (a specific risk score corresponding to the same rate of recidivism across population groups), or to a correct balance between the negative and the positive classes (the average assigned scores to those who reoffend should be identical across population groups) (Kleinberg, Mullainathan, & Raghavan, 2016). Not only is this a clearly political choice, it is also a choice that is difficult to avoid as both notions of fairness cannot be satisfied simultaneously in the vast majority of real-world cases (where we cannot predict perfectly and where base rates differ between groups) (Kleinberg et al., 2016). Hence, parts of what defines the search for evidence in big data and of what we infer from data, are in fact political choices.

Big data arguably adds to this problem, as it is "easy to mistake correlation for causation and to find misleading patterns in the data" (McAfee & Brynjolfsson, 2012, p. 68). There is thus the space for exploiting this 'malleability' of big data insights by policy makers seeking to find evidence for a policy that fits their pre-existing agenda – a behavior well documented in the literature (Kogan, 1999; Marmot, 2004; Nelkin, 1975; Walker, 2000). Given the number of new data sources, methods, and the size of the data itself, it will become increasingly possible to support virtually any policy intervention as 'evidence-based', which greatly expands the room for the 'political game' one can play with data. This renders the concept of 'evidence based policy' less meaningful, but also brings to the forefront some fundamental questions: "To whom do the analytics and findings go to and for which purposes? Who is profiting the most and least from big data?" (Uprichard, 2015, ¶ 2). Given that data are not inherently objective (as addressed above) and that human design and biases affect the methodologies for dealing with data (Crawford, Gray, & Miltner, 2014), questions about actor involvement, agendas, gains, and losses remain crucial.

Given the political nature of collecting and interpreting data, the more data there are, the more political choices will have to be made by those deriving meaning from the data - the analysts. For a policy maker or politician, this presents an uncomfortable situation: Analysts create algorithms to analyze big data, but the algorithms are often very complex and self-adjust, making the (political) choices made along the way difficult to interrogate (as demonstrated by, for example, the COMPAS algorithm) by the very people that (have to) use these statements and insights as a basis for policy (Janssen & Kuk, 2016b; van der Voort, Klievink, Arnaboldi, & Meijer, 2018).

This dynamic introduces more actors into the policymaking process and the necessary public-private partnerships often result in tacit endorsement of security and privacy policies of private sector analytics companies (Bertot & Choi, 2013). Furthermore, this could also change the policy process itself: Entrepreneurial data analysts, scientists and enthusiasts are empowered (by the existence of big data that they can repurpose) to proactively come up with insights and services that may call for a policy response, putting the decision maker into a reactive role (van der Voort et al., 2018). Not only does this provide substantial agenda-setting power to the analysts, it also constitutes a radical decentralization of policymaking: If analysts can provide answers and solutions to problems the decision makers do not know exist yet, the cycle of "goals ➔ gathering information ➔ intervention" that characterizes traditional policymaking is effectively changing to "gathering information ➔ intervention ➔ goals", where the boundary between gathering data, making inferences, and the intervention is increasingly permeable. This is of course not a general trend, but the fact that some interventions can happen in this manner reinforces the observation of Klijn and Koppenjan (2015) that the entire process resembles a plate of spaghetti and happens in a much less structured and less predictable fashion than a techno-optimist narrative assumes.

3. Faster decisions: the unattainable ideal of real-time

Outside of resulting in better information that leads to better decisions, the techno-optimist narrative also maintains that big data analytics produce faster information which in turn leads to faster decisions. Not only is it argued that automation will accelerate some of public administrations' informational tasks (Maciejewski, 2016), but that real-time data streams will reduce the time period between policy coming to effect and being evaluated, as "[d]emographic data, unemployment numbers or migration patterns could be observed in real time, enabling a much faster assessment of whether the implementation of a certain policy was a success or not" (Höchtl et al., 2016, p. 162). In other words, big data will enable policy interventions to happen in real-time or near real-time.

Much like with better data leading to better decisions, this argument rests on two assumptions: That it is possible to generate relevant real-time data to inform policy decisions and that policymaking can adapt to the speed of this data. In this section we question both of these assumptions (in that order), pointing to the fact that many policy decisions are concerned with the long term, that many relevant indicators do not respond to policy interventions "in real-time", and that the speed of policy decision making is constrained by public administration and decision-making dynamics that are not removed by big data (van der Voort et al., 2018).

In policy areas concerned with long-term effects, improvements on how quickly data is available mean very little: What is the impact of education on employment outcomes, of pollution on environment, or of healthcare policy on health outcomes? All of these questions are extremely salient for policy and cannot be answered in real-time, as the effects they are concerned with materialize only years after the policy intervention. The benefits of faster measurement still exist, but a 'data lag' of even a few months is close to insignificant when measuring effects that take years or even decades to materialize, especially if there are other dimensions of quality of the measurement to be considered. Thus, notwithstanding the demonstrable potential of big data to speed up policymaking in multiple policy (Kitchin, 2014b; Lettieri, 2016; Wamba, Edwards, & Sharma, 2012), this potential cannot be extended to policymaking in general.

Furthermore, an effect lag exists even for policies that are meant to have effect as soon as possible and that we often assume can be measured in real-time. Consider employment or unemployment indicators – indicators that many have tried to measure in real time using big data (Antenuccia, Cafarellab, Levensteinc, Red, & Shapiro, 2014; Askitas & Zimmermann, 2009; Choi & Varian, 2009; D'Amuri, 2009; D'Amuri & Marcucci, 2010; Proserpio, Counts, & Jain, 2016; Vicente, Lopez-Menendez, & Perez, 2015) and that are of crucial importance for labour market policy. Both terminating and obtaining employment are not instantaneous (employees have to give notice and job seekers have to be selected, negotiate contracts, etc.) and thus assuming that a labour market policy intervention would yield (un)employment outcomes immediately is misleading. More timely measurement is of course valuable, but even if the indicator we are interested in can be measured in real-time and a policy has immediate effect on individual behaviour, translating that behaviour to a measurable change of an indicator is not instantaneous.

Lastly, much like with the quality of data and decision making, the 'speed' at which data can be generated does not translate directly into the 'speed' of decision making; The 'political game' described in Section 2.2 does not happen instantly as actors have to co-ordinate, negotiate, and often bring in third party companies for their big data analytical expertise (Giest, 2017). This is a lengthy process that gets further extended by disagreements on interpretation, or by a misalignment with a policy window. The 'policy window' concept refers to the fact that for a policy action to be taken, multiple 'streams' have to align (Kingdon & Thurber, 1984), including a 'politics' stream that refers to whether policymakers have the will and opportunity to make the necessary policy (Cohen, March, & Olsen, 1972). In other words, often time it is not enough to identify a problem and conceive a solution, it is also important to implement this solution at the 'right time'. Faster data

analytics are of course helpful in capitalizing on open policy windows, but it is also important to realize that just having a solution to a problem doesn't mean that the corresponding policy action can be taken. Needless to say, big data does not affect the political dynamics that determine when it is the 'right time' to create a specific policy.

4. New epistemological and methodological singularism in the works?

We now move to addressing an assumption we believe to underlie a substantial part of the techno-optimist narrative: The (often implicit) assumption that correlations identified in large datasets (or predictions made by models trained on such data sets) are at least a sufficient replacement for understanding causality of the relationship in question. Needless to say, not all big data analytics are based on this assumption, but a general link between big data analytics and privileging correlation over causation can be observed (Bollier, 2010; Kitchin, 2014a; Kitchin & Lauriault, 2015; Zwitter, 2014), as some argue rather explicitly that "we will need to give up our quest to discover the cause of things, in return for accepting correlations" (Cukier & Mayer-Schoenberger, 2013, p. 29). Perhaps even more importantly, this emphasis on correlation and prediction goes hand in hand with the belief that "[w]ith enough data, the numbers speak for themselves" (Anderson, 2008, ¶ 7). If this logic is applied to public policymaking, it translates into arguments such as "[t]he undeniable truth of facts [provided by big data] cannot be neglected even by the most stubborn politicians" (Höchtl et al., 2016, p. 146).

This leads some to argue that "[b]ig data helps answer what, not why, and often that's good enough" (Cukier & Mayer-Schoenberger, 2013, p. 29). It is important to acknowledge the use of 'often' by Cukier and Mayer-Schoenberger (2013), but in this section we argue that in policymaking more often than not knowing 'what' without the 'why' is not good enough. We first challenge the assumption of 'organic' data and meaningful correlations that can speak for themselves as a fundamental misunderstanding of data in the context of social science (Section 4.1), following which we also illustrate the practical limitations of purely predictive approaches when it comes to answering various policy questions (Section 4.2).

4.1. Death of the scientific method?

One of the most popular and forceful endorsements of the data-driven approach is Anderson's (2008) claim that the scientific method is dead. His argument is that creating models and theories can be useful, but models are never truly correct as reality is too complex to be captured by one (Anderson, 2008). Contrary to models and theories, enormous data sets collected with no specific analytical purpose in mind are argued to 'organically' reflect social reality more so than traditional statistical data (Groves, 2011; Zwitter, 2014, p. 2), making the patterns we find within them meaningful and informative in and of themselves. At face value, at least a part of this argument is true – models are always a simplification of the infinitely complex reality and as such might be useful, but are never truly accurate.

However, following the lines of Anderson's own argument, since reality cannot be captured by a model it cannot be captured by a data set either: Reality is infinitely complex and datasets are inherently finite, making it impossible to capture the 'full domain' of reality within a dataset (Kitchin, 2014a). Secondly, data do not really exist in a vacuum and cannot be meaningful without understanding and interpretation. Whether by design or due to practical limitations, data don't really exist in a "raw" and "organic" form (Gitelman, 2013) and always capture reality from a specific vantage point (Kitchin & Lauriault, 2015). Furthermore, we cannot derive any meaning from data without interpreting them and attaching them to domain-specific knowledge (Clemons & McBeth, 2009; Janssen & Kuk, 2016a; Kitchin, 2014a). Even if the process of translating numbers into data lacks a formal framework, science doesn't circumvent the human perspective (Giere, 2006; Gould, 1981), making it impossible to capture reality in an 'organic' dataset. Because of this, data and the correlations contained in it cannot 'speak for themselves' (Goldston, 2008; Kettl, 2016; Liu et al., 2016) and are in fact crucially dependant on how we make sense of this data and interpret it – a process that is far from organic and objective.

In terms of meaningfulness of correlations, the fact that correlation does not imply causation requires no explanation, but we wish to take this argument even further: In big data sets, as Boyd and Crawford (2012) correctly note, correlation does not really imply much: "[E]normous quantities of data can offer connections that radiate in all directions" (Boyd & Crawford, 2012, p. 668). The reason for this is that with larger sample sizes the criterion of statistical significance is easier to satisfy, as well as the high-dimensionality of big data allowing for more potential correlations. This implies that in large data sets correlations are also less meaningful, for which a mathematical proof can be constructed, showing that there exist "ramsey-style correlations" that exist purely because of the size of a data set and in large data sets these can be the quantifiable majority of statistically significant correlations (Calude & Longo, 2016). This is not to say that we should not look for interesting correlations in big data, but that we should be acutely aware of the fact that some of the 'traditional' risks like spuriousness are even more pronounced in big data sets and that we should opt for statistical rigor over the assumption that correlations are meaningful because of the size and 'organic' nature of the data.

Lastly, even if data could 'speak for itself', using that for policymaking without interpretation could be unlawful. Big data insights often 'hide' a lot of discrimination, as historical exclusion and discrimination of certain groups reflects itself in data (and consequently in models trained on that data). Thus, adhering only to data-driven insight would further reinforce the discriminatory dynamic at play (Barocas & Selbst, 2016), which would be illegal in state-sponsored services (Samarajiva & Lokanathan, 2016).

4.2. Data-driven science in policymaking and public administrations

Despite all the above mentioned problems, there is an argument to be made that public administrations are not academia and could be more open to more inductive approaches, as politicians often revert to 'common sense' in policy decisions (Kettl, 2016, 2018) and the risks of spurious correlations are arguably context dependent. For example, it might be impossible to determine a detailed psychological theory of how work-related frustration translates to job loss, but the absence of such theory does not make this link particularly" risky" and the relationship between the two can arguably still be leveraged to understand labour market policy (United Nations, 2011). In other words, public administrations do not work along clearly demarcated 'inductive' and 'deductive' lines and are often open to 'doing what works' regardless of the epistemological implications. As such, it is important to translate this epistemological dilemma into practical terms.

The main practical limitation of an inductive data-driven approach is that it can be analytically suffocating (Lemire & Petersson, 2017) and even though it is useful for policy questions concerned with prediction, there are many other policy questions it is not useful for: Questions that beg causal proof, questions that beg explanations, or questions that beg comparative judgement (Lemire & Petersson, 2017). This is best illustrated by the difference between causality and prediction policy questions, both of which are extremely important but not answerable by the same methods. Prediction problems essentially require pre-existing knowledge about the causal link between policy intervention and outcomes, including how these outcomes depend on the occurrence of a specific event (Athey, 2017).

For example, consider evacuation policy aimed at minimizing casualties of a natural disaster: The causal link between evacuation and minimizing casualties is self-evident - if people are not in the affected area they will not be hurt by a natural disaster. In this case the

effectiveness of the evacuation policy depends almost entirely on accurately predicting when the natural disaster takes place – evacuate too early or too late and you displace people without preventing casualties. However, for some of the most crucial policy problems the difficulty runs in reverse with the causality question taking precedence over the prediction question: Reducing poverty, increasing employment, or optimizing service delivery are all crucial policy areas where accurate predictions are secondary to understanding the underlying causal mechanisms. In other words, better prediction is extremely valuable for some policy question, but for other policy questions causal explanation (and other approaches and analyses in general) are a more important part of the answer and cannot be replaced solely by predictive methods.

Furthermore, inductive big data analytics are not irreconcilable with the scientific method, provided that they are only used at the stage of hypothesis formulation: Big data can point to interesting novel hypotheses and theories that can further be tested in a rigorous manner (Liu et al., 2016), resulting in more data-driven science, but science nevertheless (Kitchin, 2014a; Kitchin & Lauriault, 2015). Such approaches should allow for leveraging the insight big data can provide without abandoning the scientific method.

5. Unwarranted de-emphasizing of crucial issues: the case of privacy protection

The limitations and challenges to the use of big data in the public sector as outlined above are rarely systematically addressed in scholarly work on the topic. Yes, scholars generally do discuss potential limitations of big data use in the public sector, but these often end at the level of acknowledging the problem and leaving it for future legal or policy solutions. Crucial issues such as privacy are then left with "government is required to pursue this [big data] agenda with strong ethics" (Höchtl et al., 2016, p. 156). At times, these challenges are even omitted entirely. Of course, not every research can address every potential problem with big data use, but in addressing problems it is crucial to engage with how these problems are linked to the process of big data analytics itself in order to avoid assuming that the two are separable.

It is outside of the scope of this paper to provide a comprehensive overview of all the big data challenges that tend to get overlooked, de-emphasized, or addressed selectively in the literature. In light of this paper's objective, we do look at the underlying logic, demonstrating why it is problematic to de-emphasize these issues based on the belief that many of them can be solved down-the-line without altering the fundamental analytics. Using the example of privacy protection we illustrate that the ethical, societal, or other non-technical challenges are inseparable from big data analytics itself. Even though this section focuses on privacy as one of the best known and often referred to problems, an attentive reader can surely apply this more inquisitive approach to a host of other commonly known big data pitfalls.

5.1. Privacy: the big trade-off in using big data

In a way, a techno-optimist view of big data analytics makes it very difficult to engage with the issue of privacy: If we speak of data and the patterns they contain as something that is inherently objective and meaningful (as the techno-optimist narrative suggests), our data sets need to mirror social reality as closely as possible. However, in order to avoid privacy breaches, we need to distort our data. This dilemma is at the root of the trade-off between privacy protection and the validity of empirical inferences one can derive from a dataset (Daries et al., 2014).

This trade-off can be illustrated both conceptually and empirically. Conceptually, achieving a data set that doesn't pose a privacy risk is simple under partition-based privacy standards such as k-anonymity (Sweeney, 2002). We can set k to a large value and distort the data to the point where no two entries can be distinguished from one another, reaching a data set that poses absolutely no risk to privacy, but is also devoid of all meaning. In other words, "[t]o strip data from all elements pertaining to any sort of group belongingness would mean to strip it from its content" (Zwitter, 2014, p. 4). This is because in meeting a specific anonymity requirement the data needs to be manipulated by a combination of suppression of entries and generalization of entire variables (Daries et al., 2014). The problem with those manipulations is that generalizing variables generalizes the data set as a whole and introduces a bias into the correlations and suppressing certain entries introduces a demographic bias (Angiuli, Blitzstein, & Waldo, 2015). More research is needed in this area, but the current research has already shown that reaching k-anonymity (for $k = 5$) can significantly distort conclusions derived from a data set (Angiuli et al., 2015; Waldo, 2016).

The question then becomes whether this trade-off between privacy and accuracy can be reconciled by technological solutions of either improving the popular privacy standards (such as k-anonymity), or creating a different privacy standard altogether. In terms of optimizing k-anonymity, Angiuli et al. (2015) show that the trade-off between distorting the means of variables and distorting the correlations between quasi identifiers is much more acceptable at certain "bin sizes" used for the generalization procedure. Other methods such as introducing "chaff" into the data instead of excessive suppression could also be a (part of the) solution (Waldo, 2016). Another solution would be a different privacy standard altogether, with non-partition-based standards such as differential privacy showing the largest promise by resisting a wider range of privacy attacks (Mohammed, Chen, Fung, & Yu, 2011). Nevertheless, differential privacy still distorts the accuracy of the data and this trade-off is rather explicit in setting the privacy parameter: The more secure this parameter, the more noise is introduced to data with each query and less queries are allowed. Thus, despite its potential to optimize the trade-off (Ghosh, Roughgarden, & Sundararajan, 2012; Mohammed et al., 2011), differential privacy can only be perfect for specific users and single count queries, but not for other types of queries (Brenner & Nissim, 2014). This is not to discredit these technological solutions, but to point out that their potential is merely to optimize rather than completely reconcile the privacy and accuracy trade-off.

Outside of technological solutions to this trade-off, the argument for policy solutions can be made. Here the debate turns even more speculative, since no alternative approaches to de-identification exist in practice. Theoretically, one of the promising concepts has to do with a shift from preventing privacy breaches to punishing them effectively. Such an approach would allow for sharing of de-identified data sets under the condition of tracking how individual users use this data in order to punish re-identification attempts and other misuse (Waldo, 2016). Such developments are extremely speculative, especially because there is no technical solution to enforce such a drastically different system: A scalable and practicable system of enforcement and audit of contracts on data use in the current legal system is difficult to even imagine, let alone implement (Daries et al., 2014). Thus, despite some signs of legislators re-thinking privacy regulation, no significant changes can be expected to happen soon (Angiuli et al., 2015). In sum, the evidence seems to suggest that not distorting data and respecting individual privacy are not (perfectly) reconcilable and that we are far removed from a good technical or policy solution.

6. Discussion and conclusion

Big data is expected to have a profound impact on the public sector. In recent years, a body of literature has emerged highlighting the possibilities of using big data for better insights, better decision making, and for significantly altering policy processes. Yet, the true challenge to these promises lies in where big data meets existing practice in the public sector. Although the literature has not completely neglected these challenges, the current debate on big data in the public sector emphasizes technical-rational factors, focusing much more on data and analytical output rather than on its interaction with the decision-

making process in public administrations. Throughout this paper we have illustrated why political decision-making factors should be taken seriously by critiquing some of the core techno-optimist tenets from a more policy-pessimist angle, constructing these two archetypical narratives in the process.

We have first tackled the claim that big data provide 'better' insights and thus foster better decisions: Not only is big data not always 'better' in terms of accuracy, but there are also multiple dimensions of data quality. Furthermore, translating 'better' evidence into 'better' policy is subject to public administration dynamics much more complex than the techno-optimist narrative assumes. Secondly, we address a similar argument of faster insight resulting in faster policy decisions, which we challenge based on not all policy questions being able to benefit from near real-time measurement because of long-term concerns or natural delays in the causal chain. Furthermore, public decision-making dynamics are not removed by big data and introduce a substantial time lag in and of themselves. Thirdly, we tackled the less clearly articulated but no less important epistemological concerns with big data analytics as both a fundamental misunderstanding of data, but also as a practical limitation in terms of what policy questions can be answered. Lastly, we have argued against how a techno-optimist narrative de-emphasizes certain issues that are in fact crucial and should be an integral part of the debate – an argument that we illustrate on the trade-off between privacy protection and accuracy. In this concluding section, we first summarize the two narratives and have them meet eye-to-eye, and second provide a realist rejoinder.

6.1. Techno-optimism and policy-pessimism: an eye-to-eye comparison

Despite challenging techno-optimist arguments throughout this paper, our goal is not to make a case for policy-pessimism as an alternative. The problem we see in the current literature is not an absence of a critical alternative to techno-optimism, but rather that such an alternative is complex, spans many disciplines, and only seldom makes it into individual research projects and agendas in a systematized and comprehensive way. As a result, even high quality research often subscribes to techno-optimist simplifications in approaching legislation (Bertot & Choi, 2013), privacy (Sagiroglu & Sinanc, 2013), data quality considerations (Ku & Leroy, 2014; Matheus, Janssen, & Maheshwari, 2018), and many other aspects of big data use. Our contribution aims to remedy that by articulating the two archetypical narratives and making them meet 'eye-to-eye', allowing scholars to systematize the way in which they interrogate big data promises and shortcomings, paying sufficient attention to both technical-rational and political decision-making factors.

To provide this eye-to-eye comparison, in Table 1 we summarize both narratives along the four dimensions addressed throughout this paper. In this table we also include a hypothetical set of questions that one of these narratives would interrogate the opposing narrative with. These questions are derived from arguments we have presented in this paper, which also constitutes an important limitation: Since this paper is mainly challenging the techno-optimist narrative from a policy-pessimist lens, the policy pessimist questions are far better anchored in the existing literature. We still derive some key techno-optimist questions from our summary of the narrative, but recognise that a more thorough summary and defence of the techno-optimist narrative would certainly arrive at more informed and grounded techno-optimist questions. Despite this limitation, these questions as presented in Table 1 illustrate the utility of understanding these two narratives as logical extremes that might not provide the best argument, but that are asking important questions.

Despite our diagnosis that the literature as a whole is leaning towards techno-optimism and our subsequent case for the utility of policy-pessimism, systematizing the assumptions and arguments of the literature in this fashion has value even if one disagrees with our diagnosis. The techno-optimist and policy-pessimist systematization

Table 1
An eye-to-eye comparison of techno-optimism and policy-pessimism.

Key issue	The 'techno-optimist' narrative	The 'policy-pessimist' questions	The 'policy-pessimist' narrative	The 'techno-optimist' questions
Quality of big data insight and how that translates into quality of decisions (Section 2)	Big data provides more information which means better insight and better predictive capabilities, which then translates into better informed (and thus generally better) policy decisions.	Is big data better on all data quality dimensions? Can data be universally better? If not, who or what are they better for? How do data get translated to decisions?	On important quality dimensions big data is not better for policymaking than traditional data. Politicians will always cherry-pick data that suits their agenda – more data will diffuse the meaning of 'evidence based' and result in more political strategizing.	How will better estimates and predictions impact decision-making? How can analysts and new data source facilitate better insight? Can certain decisions be automated? How does measuring previously unmeasurable concepts help in policymaking?
Speed of big data analysis and how that translates to speed of decisions (Section 3)	Real-time data streams provide more up-to-date information faster than currently available data, meaning that policy decisions can be made faster, making policy more agile.	Is faster data possible or useful in all policy areas? Can decision making adapt to the speed of data? What gets lost if we remove humans from the equation to allow for faster decisions?	Decision-making will not adapt to the speed of data, as negotiation and interrogation of the data by humans is a crucial part of the process. Faster data is not available for most policy questions.	Does reduced data-lag influence policy-relevant insights? Does better temporal resolution improve insight? Can't certain decisions be reliably automated?
Epistemology of big data analysis (Section 4)	A more inductive approach based on correlation and prediction rather than causation as long as the dataset is of sufficient size.	What is the role of interpretation? How meaningful are correlations in big data? Is this approach appropriate for policy questions not predictive in nature?	No substantial change to scientific method, muting the effect big data analytics will have as they are tailored for inductive exploration and not deductive testing.	Why would public sector not emulate private sector for efficiency gains? Can inductive exploration contribute new and relevant insight? Why not 'do what works' if we can show that it does?
Connection between big data analysis and fundamental concerns related to it (Section 5)	Privacy and other fundamental issues with big data matter, but can be overcome down the line with more advanced technological solutions or policy interventions.	Are these issues related to the type of analytics advocated? What are the trade-offs with these issues? How likely are these issues to be solvable 'down-the-line'?	Privacy and other fundamental big data issues are crucial and cannot be (fully) reconciled with big data analytics. To avoid them we should stop or limit big data analytics.	What technological solutions can provide good results? Should limitations stop progress in terms of big data use? What is the balance of risks and rewards (including the risk of falling behind in data utilization)?

offers a tool that can be fitted to a specific research context: A specific research focus might require these two narratives to emphasize the various important legal or ethical concerns (such as intellectual property rights, data security, liability, accountability, etc.) and de-emphasize some of the points we focus on in this paper. Regardless of the focus, this systematization will still pose important questions and expose where on the axis between the two narratives one is located. That in turn presents two options: Either defend a specific position as the most appropriate trade-off point (argue for one narrative over the other), or find a way to reconcile the two narratives in a 'best-of-both-worlds' fashion. Doing neither results in (unintentional) cherry-picking of the easiest to address problems and not tackling underlying assumptions that can, despite seeming inconsequential, influence research findings.

6.2. A realist rejoinder

To conclude this paper, we offer our take on a rejoinder between techno-optimism and policy-pessimism in the form of a middle-of-the-road realist perspective. To achieve that, we propose a move away from the umbrella terms of 'big data' and 'policymaking' to talking about specific data sources and methods used for specific policy questions. It is difficult to make general conclusions about big data use because there are numerous associated benefits and pitfalls which depend on context. Some of the pitfalls are addressed by this paper, but many are omitted, including the costs and challenges associated with developing skills and infrastructure, representativeness of big data sets, the procurement of data itself and the necessary public-private partnerships, accurately distinguishing 'signal' from 'noise' in big data sets, legal concerns, and many others. On the other hand, there are important and difficult to deny benefits of big data: The speed of data and analysis can be tremendously valuable for time-sensitive policy responses or monitoring systems, the large sample size can mean much more accurate disaggregation of data crucial for group-specific interventions, and analysing novel datasets can provide previously unmeasurable insight. Furthermore, once the infrastructure is in place and skills are developed, the marginal cost of an additional analytical inquiry is miniscule compared to traditional survey based sources (Kitchin & Lauriault, 2015), further reinforced by the fact that response rates to surveys are declining (Bostic, Jarmin, & Moyer, 2016).

Given how many such shortcomings and benefits exist and the absence of a meaningful way to sort them, it might seem that the decision for or against adopting big data is arbitrary or heavily political at best. However, we believe that in looking at specific cases (a data source and a method applied to a policy question) the trade-offs between shortcomings and benefits become meaningful enough to make sound (albeit political) choices on. Consider the example of data backrun mentioned in this paper: Data backrun is of tremendous importance for policy decisions on issues that policy makers have been wrestling with for decades, but of extremely little importance for more recent issues whose emergence coincides with the emergence of big data (such as e-commerce), because for those issues conventional data have no comparative backdrop advantage. This context-specificity applies to all possible pros and cons: Representativeness issues might not be serious in group-specific policy decisions, privacy is almost a non-issue when using aggregated search query data as opposed to individual search query data, and the speed of big data can benefit rapid response policies but does very little for long-term human capital policies. Outside of public policymaking, public administrations also have the task of public service delivery (and optimization), for which data needs can be different and thus also emphasize and de-emphasize various shortcomings and benefits of big data. Not only do these trade-offs become meaningful at the level of individual policy problems and data sources, they also show some space for generalizations: For example, many fundamental economic questions are naturally retrospective, and thus benefit from data accuracy much more than from timeliness (Einav & Levin, 2013), making it unreasonable to expect any shift towards 'relativized exactitude' in solving those policy questions. Through balancing these pitfalls and benefits is how decisions for or against the adoption of big data analytics can be most meaningfully made.

That said, here we draw on policy-pessimism to highlight that making 'meaningful' decisions on big data does not mean making them fully rationally: Public administrations are not purely rational entities and different stakeholders are not only likely to reach different conclusions with regards to whether big data is actually fit for a specific policy question, but also use these conclusions in different ways depending on broader strategic concerns and individual agendas. The process of public administration can resemble a strategic game rather than rational deliberation (Klijn & Koppenjan, 2015) and the adoption of big data is not immune to this dynamic. This means that to understand big data in the public sector, it is important to understand not only the rationality behind balancing the context-specific benefits and pitfalls of big data, but also the actors and institutions that participate in making the decision.

The realist rejoinder we propose can be summarized in three key points: Firstly, big data has multiple aspects of quality (including speed) and the importance of these is crucially dependant on the policy question, data source, and methods. As such, big data will be a 'game changer' for certain policy areas, but will continue to struggle with adoption in other policy areas. Secondly, big data is subject to public administrations and decision making dynamics when used for policy purposes, making the translation from big data insights into policy action rather complex. As such, even 'better' or 'faster' insights could be affected by this process and result in unexpectedly good or bad policy. Finally, as a consequent of these two arguments, big data adoption will remain uneven and will be determined by numerous balancing acts of big data benefits and pitfalls for a specific policy application and data source by networks of actors. These balancing acts will be subject to divergent perspectives, pre-existing agendas, will not be fully rational, and will require time.

We hope that our systematic way of addressing optimist and pessimist arguments and assumptions in the current debate will help scholars and policy makers to interrogate and challenge their own assumptions. This may lead to a better fit between the goals of big data for specific uses and the context in which it will be applied, as well as to more realistic expectations and hence more careful decisions about deploying big data in practice.

Acknowledgements

This work was supported by the European Research Centre for Economic and Financial Governance (EURO-CEFG), a joint research initiative of Leiden University, Delft University of Technology and Erasmus University Rotterdam

References

Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. Retrieved May 16, 2017, from https://www.wired.com/2008/06/pb-theory/.

Angiuli, O., Blitzstein, J., & Waldo, J. (2015). Balancing statistical accuracy and subject privacy in large social-science datasets. *Communications of the ACM, 58*(12)https://doi.org/10.1145/2814340.

Angwin, J., & Larson, J. (2016). ProPublica responds to Company's critique of machine bias story. Retrieved January 25, 2019, from https://www.propublica.org/article/propublica-responds-to-companys-critique-of-machine-bias-story.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. Retrieved January 25, 2019, from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Antenuccia, D., Cafarellab, M., Levensteinc, M., Red, C., & Shapiro, M. (2014). Using social media to measure labor market flows. Retrieved from http://www-personal.umich.edu/~shapiro/papers/LaborFlowsSocialMedia.pdf.

Askitas, N., & Zimmermann, K. (2009). Google econometrics and unemployment forecasting. IZA discussion papers no. 4201. Retrieved May 28, 2019, from https://www.econstor.eu/handle/10419/35733.

Athey, S. (2017). Beyond prediction: Using big data for policy problems. *Science, 355*(6324).

Barocas, S., & Selbst, A. D. (2016). Big Data's disparate impact. *California Law Review, 104*(671).

Bertot, J. C., & Choi, H. (2013). Big data and e-government: Issues, policies, and recommendations. *The proceedings of the 14th annual international conference on digital government research (dg.O 2013)* (pp. 1–10). New York: ACM. https://doi.org/10.1145/2479724.2479730.

Boettcher, I. (2015). Automatic data collection on the internet (web scraping). Retrieved May 28, 2019, from https://ec.europa.eu/eurostat/cros/content/automatic-price-collection-internet-web-scraping-ingolf-boettcher_en.

Bollier, D. (2010). *The promise and peril of big data.* The Aspen Institute. Retrieved May 28, 2019, from https://assets.aspeninstitute.org/content/uploads/files/content/docs/pubs/The_Promise_and_Peril_of_Big_Data.pdf.

Bostic, W. G., Jarmin, R. S., & Moyer, B. (2016). Modernizing federal economic statistics. *American Economic Review, 106*(5), 161–164. https://doi.org/10.1257/aer.p20161061.

Boyd, D., & Crawford, K. (2012). Critical questions for big data. *Information, Communication & Society, 15*(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878.

Brenner, H., & Nissim, K. (2014). Impossibility of differentially private universally optimal mechanisms. *SIAM Journal on Computing, 43*(5), 1513–1540. https://doi.org/10.1137/110846671.

de Bruijn, H., & Ten Heuvelhof, E. (2008). *Management in networks : On multi-actor decision making.* Routledge.

Brynjolfsson, E., & McAfee, A. (2014). *The second machine age : Work, progress, and prosperity in a time of brilliant technologies.* New York: W. W. Norton & Company.

Calude, C. S., & Longo, G. (2016). The deluge of spurious correlations in big data. *Foundations of Science, 22*(3), 595–612. https://doi.org/10.1007/s10699-016-9489-4.

Chatfield, A. T., & Reddick, C. G. (2018). Customer agility and responsiveness through big data analytics for public value creation: A case study of Houston 311 on-demand services. *Government Information Quarterly, 35*(2), 336–347. https://doi.org/10.1016/J.GIQ.2017.11.002.

Chen, Y.-C., & Hsieh, T.-C. (2014). Big data for digital government: Opportunities, challenges, and strategies. *International Journal of Public Administration in the Digital Age, 1*(1), 1–14. https://doi.org/10.4018/ijpada.2014010101.

Choi, H., & Varian, H. (2009). Predicting initial claims for unemployment benefits. Retrieved May 28, 2019, from http://static.googleusercontent.com/media/research.google.com/en//archive/papers/initialclaimsUS.pdf.

Clemons, R. S., & McBeth, M. K. (2009). *Public policy praxis : A case approach for understanding policy and analysis.* New York: Routledge.

Cohen, M. D., March, J. G., & Olsen, J. P. (1972). A garbage can model of organizational choice. *Administrative Science Quarterly, 17*(1), 1. https://doi.org/10.2307/2392088.

Crawford, K., Gray, M., & Miltner, K. (2014). Big data| critiquing big data: Politics, ethics, epistemology| special section introduction. *International Journal of Communication, 8*(10), 1663–1672.

Cukier, K., & Mayer-Schoenberger, V. (2013). The rise of big data: How it's changing the way we think about the world. *Foreign Affairs, 92*(3), 28–40.

D'Amuri, F. (2009). Predicting unemployment in short samples with internet job search query data. MPRA paper 18403. Retrieved May 28, 2019, from https://ideas.repec.org/p/pra/mprapa/18403.html.

D'Amuri, F., & Marcucci, J. (2010). “Google it!” forecasting the US unemployment rate with a Google job search index. *FEEM working paper no. 31*https://doi.org/10.2139/ssrn.1594132.

Daries, J. P., Reich, J., Waldo, J., Young, E. M., Whittinghill, J., Ho, A. D., ... Chuang, I. (2014). Privacy, anonymity, and big data in the social sciences. *Communications of the ACM, 57*(9), 56–63. https://doi.org/10.1145/2643132.

Desouza, K. C., & Jacob, B. (2014). Big data in the public sector: Lessons for practitioners and scholars. *Administration and Society, 49*(7), 1043–1064. https://doi.org/10.1177/0095399714555751.

Dieterich, W., Mendoza, C., & Tim Brennan, M. (2016). COMPAS risk scales: Demonstrating accuracy equity and predictive parity. Retrieved May 28, 2019, from http://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf.

Dong, L., Chen, S., Cheng, Y., Wu, Z., Li, C., & Wu, H. (2017). Measuring economic activity in China with mobile big data. *EPJ Data Science, 6*(1), 6–29. https://doi.org/10.1140/epjds/s13688-017-0125-5.

Dumbacher, B., & Hutchinson, R. (2016). Enhancing the Foundation of Official Economic Statistics with big data. *European conference on quality in official statistics* (Madrid).

Dunleavy, P., Margetts, H., Bastow, S., & Tinkler, J. (2005). New public management is dead–long live digital-era governance. *Journal of Public Administration Research and Theory, 16*(3), 467–494. https://doi.org/10.1093/jopart/mui057.

Einav, L., & Levin, J. D. (2013). The data revolution and economic analysis. *NBER Working Paper, 53*(9), 1689–1699. https://doi.org/10.1017/CBO9781107415324.004.

Eurostat Big Data Task Force (2014). *22 nd meeting of the European statistical system committee.* (Retrieved May 28, 2019, from https://ec.europa.eu/eurostat/cros/system/files/ESSC doc 22_8_2014_EN_Final with ESSC opinion.pdf).

Flores, A., Bechtel, K., & Lowenkamp, C. (2016). False positives, false negatives, and false analyses: A rejoinder to "Machine bias: There's software used across the country to predict future criminals and It's biased against blacks". *Federal Probation,* (2), 80.

Ghosh, A., Roughgarden, T., & Sundararajan, M. (2012). Universally utility-maximizing privacy mechanisms. *SIAM Journal on Computing, 41*(6), 1673–1693. https://doi.org/10.1137/09076828X.

Giere, R. (2006). *Scientific perspectivism.* The University of Chicago Press.

Giest, S. (2017). Big data for policymaking: Fad or fasttrack? *Policy Sciences, 50*(3), https://doi.org/10.1007/s11077-017-9293-1.

Gitelman, L. (Ed.). (2013). *"Raw data" is an oxymoron.* MIT Press.

Goldston, D. (2008). Big data: Data wrangling. *Nature, 455*(7209), 15. https://doi.org/10.1038/455015a.

Gould, P. (1981). Letting the data speak for themselves. *Annals of the Association of American Geographers, 71*, 166–176. https://doi.org/10.2307/2562790.

Groves, R. M. (2011). Three eras of survey research. *Public Opinion Quarterly, 75*(5), 861–871. https://doi.org/10.1093/poq/nfr057.

Hackl, P. (2016). Big data: What can official statistics expect? *Statistical Journal of the IAOS, 32*, 43–52. https://doi.org/10.3233/SJI-160965.

Hargittai, E. (2015). Is bigger always better? Potential biases of big data derived from social network sites. *The Annals of the American Academy of Political and Social Science, 659*(1), 63–76. https://doi.org/10.1177/0002716215570866.

Helbig, N., Dawes, S., Dzhusupova, Z., Klievink, B., & Mkude, C. G. (2015). Stakeholder engagement in policy development: Observations and lessons from international experience. In M. Janssen, M. Wimmer, & A. Deljoo (Eds.). *Policy practice and digital science. Public Administration and Information Technology* (pp. 177–204). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-12784-2_9.

Hilbert, M. (2016). Big data for development: A review of promises and challenges. *Development and Policy Review, 34*(1), 135–174. https://doi.org/10.1111/dpr.12142.

Höchtl, J., Parycek, P., & Schöllhammer, R. (2016). Big data in the policy cycle: Policy decision making in the digital era. *Journal of Organizational Computing and Electronic Commerce, 26*(1–2), 147–169. https://doi.org/10.1080/10919392.2015.1125187.

Hussmanns, R. (2007). *Measurement of employment, unemployment and underemployment – Current international standards and issues in their application.* International Labour Organization. Retrieved May 28, 2019, from http://www.ilo.org/global/statistics-and-databases/WCMS_088394/lang–en/index.htm.

Iacus, S. M. (2015). Big data or big fail? The good, the bad and the ugly and the missing role of statistics. *Electronic Journal of Applied Statistical Analysis, 5*(1), 4–11. https://doi.org/10.1285/I2037-3627V5N1P4.

IBM (2012). *4 Vs.* Retrieved May 28, 2019, from https://www.ibmbigdatahub.com/infographic/four-vs-big-data.

Janssen, M., & Kuk, G. (2016a). Big and open linked data (BOLD) in research, policy, and practice. *Journal of Organizational Computing and Electronic Commerce, 26*(1–2), 3–13.

Janssen, M., & Kuk, G. (2016b). *The challenges and limits of big data algorithms in technocratic governance.* https://doi.org/10.1016/j.giq.2016.08.011.

Katal, A., Wazid, M., & Goudar, R. H. (2013). Big data: Issues, challenges, tools and good practices. *Sixth International Conference on Contemporary Computing,* 404–409. Retrieved May 29, 2019, from https://ieeexplore.ieee.org/document/6612229.

Keith, J., Ginnis, S., & Miller, C. (2016). Addressing quality in social media research: The question of representivity. *Social Research Practice, 2*, 21–30.

Kettl, D. F. (2016). Making data speak: Lessons for using numbers for solving public policy puzzles. *Governance, 29*(4), 573–579. https://doi.org/10.1111/gove.12211.

Kettl, D. F. (2018). *Little bites of big data for public policy.* Thousand Oaks, California: CQ Press.

Kim, G.-H., Trimi, S., & Chung, J.-H. (2014). Big-data applications in the government sector. *Communications of the ACM, 57*(3), 78–85.

Kingdon, J., & Thurber, J. (1984). *Agendas, alternatives, and public policies.* Boston: Little Brown.

Kitchin, R. (2014a). Big data, new epistemologies and paradigm shifts. *Big Data & Society, 1*(1), https://doi.org/10.1177/2053951714528481.

Kitchin, R. (2014b). The real-time city? Big data and smart urbanism. *GeoJournal, 79*(1), 1–14. https://doi.org/10.1007/s10708-013-9516-8.

Kitchin, R., & Lauriault, T. P. (2015). Small data in the era of big data. *GeoJournal, 80*(4), 463–475. https://doi.org/10.1007/s10708-014-9601-7.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. Retrieved from https://arxiv.org/pdf/1609.05807.pdf.

Klievink, B., Romijn, B.-J., Cunningham, S., & de Bruijn, H. (2017). Big data in the public sector: Uncertainties and readiness. *Information Systems Frontiers, 19*(2), 267–283. https://doi.org/10.1007/s10796-016-9686-2.

Klijn, E. H., & Koppenjan, J. (2015). Strategic complexity in governance networks: Strategies, games, rounds, and arenas. In E. H. Klijn, & J. Koppenjan (Eds.). *Governance networks in the public sector* (pp. 66–98). New York: Routledge.

Kogan, M. (1999). The impact of research on policy. In F. Coffield (Ed.). *Speaking truth to power: Research and policy on lifelong learning.* Policy Press.

Ku, C.-H., & Leroy, G. (2014). A decision support system: Automated crime report analysis and classification for e-government. *Government Information Quarterly, 31*(4), 534–544. https://doi.org/10.1016/J.GIQ.2014.08.003.

Lavertu, S. (2016). We all need help: “Big data” and the mismeasure of public administration. *Public Administration Review, 76*(6), 864–872. https://doi.org/10.1111/puar.12436.

Lemire, S., & Petersson, G. J. (2017). Big bang or big bust? The role and implications of big data in evaluation. In G. J. Petersson, & J. D. Breul (Eds.). *Cyber society, big data, and evaluation: Comparative policy evaluation.* New Brunswick: Transaction Publishers.

Lettieri, N. (2016). Computational social science, the evolution of policy design and rule making in smart societies. *Future Internet, 8*(2), 19. https://doi.org/10.3390/fi8020019.

Lindblom, C. (1959). The science of muddling through. *Public Administration Review, 19*(2), 79. https://doi.org/10.2307/973677.

Liu, J., Li, J., Li, W., & Wu, J. (2016). Rethinking big data: A review on the data quality and usage issues. *ISPRS Journal of Photogrammetry and Remote Sensing, 115*, 134–142. https://doi.org/10.1016/j.isprsjprs.2015.11.006.

Maciejewski, M. (2016). To do more, better, faster and more cheaply: Using big data in public administration. *International Review of Administrative Sciences, 83*(1), 120–135. https://doi.org/10.1177/0020852316640058.

Marmot, M. G. (2004). Evidence based policy or policy based evidence? *BMJ (Clinical Research Ed.), 328*(7445), 906–907. https://doi.org/10.1136/bmj.328.7445.906.
Matheus, R., Janssen, M., & Maheshwari, D. (2018). Data science empowering the public: Data-driven dashboards for transparent and accountable decision-making in smart cities. *Government Information Quarterly*. https://doi.org/10.1016/J.GIQ.2018.01.006.
McAfee, A., & Brynjolfsson, E. (2012). Big data: The management revolution. *Harward Business Review, 90*(10), 60–66.
McLaren, N., & Shanbhogue, R. (2011). *Using internet search data as economic indicators. Bank of England quarterly bulletin no. 2011 Q2.* Retrieved May 29, 2019, fromhttps://doi.org/10.2139/ssrn.1865276.
McNeely, C. L., & Hahm, J. (2014). The big (data) bang: Policy, prospects, and challenges. *Review of Policy Research, 31*(4), 304–310. https://doi.org/10.1111/ropr.12082.
Misuraca, G., Mureddu, F., & Osimo, D. (2014). Policy-making 2.0: Unleashing the power of big data for public governance. In M. Gascó-Hernández (Ed.). *Open Government* (pp. 171–188). New York, NY: Springer.
Mohammed, N., Chen, R., Fung, B. C. M., & Yu, P. S. (2011). Differentially private data release for data mining. *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '11 (p. 493). New York*New York, USA: ACM Press. https://doi.org/10.1145/2020408.2020487.
Nelkin, D. (1975). The political impact of technical expertise. *Social Studies of Science, 5*.
Nielsen, J. A., & Pedersen, K. (2014). IT portfolio decision-making in local governments: Rationality, politics, intuition and coincidences. *Government Information Quarterly, 31*(3), 411–420. https://doi.org/10.1016/J.GIQ.2014.04.002.
Panagiotopoulos, P., Bowen, F., & Brooker, P. (2017). The value of social media data: Integrating crowd capabilities in evidence-based policy. *Government Information Quarterly, 34*(4), 601–612. https://doi.org/10.1016/J.GIQ.2017.10.009.
Peled, A. (2014). *Traversing digital babel: Information, E-government, and exchange.* The MIT Press.
Proserpio, D., Counts, S., & Jain, A. (2016). The psychology of job loss: Using social media data to characterize and predict unemployment. *Proceedings of the 8th ACM conference on web science* (pp. 223–232). New York, NY: ACM Press. https://doi.org/10.1145/2908131.2913008.
Robertson, H., & Travaglia, J. (2015). A politics of counting – Putting people Back into big data. Retrieved May 29, 2019, from http://discoversociety.org/2015/07/30/a-politics-of-counting-putting-people-back-into-big-data/.
Ruppert, E., Isin, E., & Bigo, D. (2017). Data politics. *Big Data & Society, 4*(2), https://doi.org/10.1177/2053951717717749.
Ruths, D., & Pfeffer, J. (2014). Social media for large studies of behavior. *Science, 346*(6213), 1063–1064. https://doi.org/10.1371/journal.pmed.0020124.
Sagiroglu, S., & Sinanc, D. (2013). Big data: A review. *In 2013 International conference on collaboration technologies and systems (CTS),* 42–47. https://doi.org/10.1109/CTS.2013.6567202.
Samarajiva, R., & Lokanathan, S. (2016). Using behavioral big data for public purposes: Exploring frontier issues of an emerging policy arena. Retrieved May 29, 2019, from http://lirneasia.net/wp-content/uploads/2013/09/NVF-LIRNEasia-report-v8-160201.pdf.
Scannapieco, M., Virgillito, A., & Zardetto, D. (2012). Placing big data in official statistics: A big challenge? Retrieved May 29, 2019, from https://ec.europa.eu/eurostat/cros/content/placing-big-data-official-statistics-big-challenge-antonino-virgillito-monica-scannapieco_en.
Sweeney, L. (2002). K-anonymity: A model for protecting privacy. *International Journal on Uncertainty, 10*(5), 557–570.
United Nations (2011). Unemployment through the lens of social media. Retrieved May 29, 2019, from http://www.unglobalpulse.org/projects/can-social-media-mining-add-depth-unemployment-statistics.
Uprichard, E. (2015). Most big data is social data – The analytics need serious interrogation. Retrieved February 13, 2018, from http://blogs.lse.ac.uk/impactofsocialsciences/2015/02/12/philosophy-of-data-science-emma-uprichard/.
Vaccari, C. (2015). Report: Twitter data. Retrieved February 23, 2017, from http://www1.unece.org/stat/platform/display/bigdata/Report%3A+Twitter+data.
Vicente, M., Lopez-Menendez, A., & Perez, R. (2015). Forecasting unemployment with internet search data: Does it help to improve predictions when job destruction is skyrocketing? *Technological Forecasting and Social Change, 92*, 132–139.
van der Voort, H., Klievink, B., Arnaboldi, M., & Meijer, A. (2018). Rationality and politics of algorithms. *Will the promise of big data survive the dynamics of public decision making?*Government Information Quarterlyhttps://doi.org/10.1016/J.GIQ.2018.10.011.
Waldo, J. (2016). Can Accuracy and Privacy Co-Exist? Data for policy. Retrieved from https://www.youtube.com/watch?v=Cf9F0i2ZHt8.
Walker, R. (2000). Welfare policy: Tendering for evidence. In S. Nutley, P. Smith, & H. Davies (Eds.). *What works? Evidence-based policy and practice in public services.* The Policy Press.
Wamba, S. F., Edwards, A., & Sharma, R. (2012). Big data as a strategic enabler of superior emergency service management: Lessons from the new south wales state emergency service. *Society for Information Management and MIS quarterly executive pre ICIS workshop* (pp. 1–3). . Retrieved May 29, 2019, from http://ro.uow.edu.au/cgi/viewcontent.cgi?article=1153&context=buspapers.
Ward, J. S., & Barker, A. (2013). Undefined by data: A survey of big data definitions. Retrieved May 29, 2019, from http://arxiv.org/abs/1309.5821.
Ylijoki, O., & Porras, J. Perspectives to definition of big data: A mapping study and discussion. Journal of Innovation Management, 4(1), 69–91. doi: 10.24840/2183-0606_004.001_0006
Zwitter, A. (2014). Big data ethics. *Big Data & Society, 1*(2), 1–6. https://doi.org/10.1177/2053951714559253.

Simon Vydra is a PhD researcher on big data in policymaking at the Faculty of Technology, Policy and Management at Delft University of Technology. His research focuses on how big data is (and can be) used by public administrations to create adaptive and responsive policies. His focus is particularly on social media and search query data in labour market assessment.

Bram Klievink is Full Professor of Public Administration with a focus on Digitalisation and Public Policy, at the Faculty of Governance and Global Affairs, Leiden University. His research is about how digital innovations challenge the incumbent practices and institutions of public governance, and how digitalisation might be used in novel governance arrangements for the digital age.