Universiteit Leiden
The Netherlands

‘For good measure’: data gaps in a big data world

Giest, S.N.; Samuels, A.

Policy Sciences (2020) 53:559–569
https://doi.org/10.1007/s11077-020-09384-1

DISCUSSION AND COMMENTARY

'For good measure': data gaps in a big data world

Sarah Giest[1] · **Annemarie Samuels**[2]

Published online: 22 April 2020

Abstract
Policy and data scientists have paid ample attention to the amount of data being collected and the challenge for policymakers to use and utilize it. However, far less attention has been paid towards the quality and coverage of this data specifically pertaining to minority groups. The paper makes the argument that while there is seemingly more data to draw on for policymakers, the quality of the data in combination with potential known or unknown data gaps limits government's ability to create inclusive policies. In this context, the paper defines primary, secondary, and unknown data gaps that cover scenarios of knowingly or unknowingly missing data and how that is potentially compensated through alternative measures. Based on the review of the literature from various fields and a variety of examples highlighted throughout the paper, we conclude that the big data movement combined with more sophisticated methods in recent years has opened up new opportunities for government to use existing data in different ways as well as fill data gaps through innovative techniques. Focusing specifically on the representativeness of such data, however, shows that data gaps affect the economic opportunities, social mobility, and democratic participation of marginalized groups. The big data movement in policy may thus create new forms of inequality that are harder to detect and whose impact is more difficult to predict.

Keywords Data gaps · Data quality · Inclusive policymaking · Marginalized groups · Big data

Introduction

Since the amount of data has increased, there is a widespread techno-optimist notion that so-called big data will provide better information and that this better information will in turn facilitate better decisions. Big data is largely referred to as the collection of data so large, varied, and dynamic that it cannot be handled through conventional processing methods and often combines enormous volumes of digital data with advanced data analysis (Klievink et al. 2017; Vydra and Klievink 2019). In this context, some specifically point toward the new forms

✉ Sarah Giest
s.n.giest@fgga.leidenuniv.nl

[1] Institute of Public Administration, Leiden University, Den Haag, The Netherlands

[2] Institute of Cultural Anthropology and Development Sociology, Leiden University, Leiden, The Netherlands

of social data generated by internet users (Mergel et al. 2016). However, marginalized groups often produce less data, 'because they are less involved in the formal economy and its data-generating activities [or because they] have unequal access to and relatively less fluency in the technology necessary to engage online' (Barocas and Selbst 2016, 685). In other words, some people do not engage with activities that advanced analytics is designed to capture (Lerman 2013). Therefore, while there is seemingly more data to draw on for policymakers (Giest 2017), mining data can reproduce existing patterns of discrimination and exclusion by drawing on biased data. At the core of this paper is thus the idea that even though the volume of data has increased in recent years, the quality of the data in combination with potential known or unknown data gaps limits government's ability to create inclusive policies. Simply put, having a lot of data does not necessarily mean that the data are representative and reliable (Desouza and Smith 2014) or that governments are able to utilize them. In this context, Lerman (2013) and Hand (2020) talk about 'big data's exclusions' and 'dark data' respectively. Both conclude that the data used can have hidden data gaps that differ depending on how data was collected and analyzed as well as the kind of questions being asked. In addition, these gaps might contain non-random and systematic omissions, which can lead to data that excludes or underrepresents people at the margins—whether that is due to poverty, geography, or lifestyle (Lerman 2013; Hand 2020).

Beyond this, however, data gaps with a specific focus on marginalized groups and policymaking have received limited attention over the years. The literature on this topic focuses largely on the Global South in the context of data agency and bottom-up data generation as well as defiance (e.g. Milan and Trere 2019). Another stream of the literature highlights potential biases in big data, zooming in on social media data (e.g., Hargittai 2018; Olteanu et al. 2019). For this paper, we are particularly interested in how these data gaps manifest in different areas of government decision-making and how they potentially impact policymaking and public services. *We define data gaps as data for particular elements or social groups that are knowingly or unknowingly missing when policy is made on the basis of large datasets*. We thereby distinguish among three categories that are summarized in Table 1. A data gap may occur either when a part of the necessary data for policymaking is absent or when it is present but underused/of low quality. Importantly, the gap may be either known or unknown. In each case, the data gap may lead to an incomplete picture for policymaking.

First, data may be unavailable, and this gap is known to government. In this scenario, where the gap has been detected, government can compensate with alternative measures, which, as will be discussed below, have their own pitfalls. Policymakers may also decide to not follow-up on collecting missing data. This is what we define as 'primary data gap'. In a second version of this scenario, the data gap might be unknown to government. In this context, hidden data gaps can lead to policymakers relying on datasets that unintentionally underrepresent certain groups, which can potentially have wider repercussions for public decision-making and may overlook smaller, potentially vulnerable groups. In a scenario where awareness of the gap is met with available data, there are additional hurdles that government may encounter. These can originate from the required data being proprietary and in the hands of private companies or government lacking the expertise or resources to utilize them. Finally, the data that

Table 1 Types of data gaps in policymaking with large datasets

	Data unavailable	Data available
Data gap known	Primary data gap	Secondary data gap
Data gap unknown	Hidden data gaps	

are available may also be of poor quality and are unable to be a good 'fix' for the data gap that is being filled. This is what we call a 'secondary data gap'. These aspects are particularly relevant when we turn to 'inclusive policymaking'. The OECD (2019) draws attention to an approach to policymaking that better understands how policies are designed and implemented. This, according to the OECD, builds on reliable and relevant information in order to make informed decisions. If some reliable information is missing or is perceived as complete while experiencing data gaps, this creates an issue for those affected by policies created based on incomplete data.

The following sections will discuss in more detail the primary, secondary and hidden data gaps based on examples. The analysis will also show how flaws in the data have effects on public decision-making and service delivery. The final section is dedicated to raising larger questions around the data input and output in times of big data and how that changes the way governments see and design policies for marginalized groups.

Primary data gap

The primary data gap describes a scenario in which government is aware of the fact that data is missing, but there are limited opportunities to fill this gap due to the lack of appropriate data. In recent years, the technical ability to mine large amounts of data has resulted in ways to replace missing values through, for example, proxy variables. The following section looks at these solutions in order to better understand how they work and whether they are able to indeed provide a more complete 'data picture' for government with a focus on minority groups.

Machine learning systems are increasingly relied upon for many government policies, such as flagging potential welfare fraud recipients or the identification of money laundering schemes. The problem of automating these things is that artificial intelligence learns based on what the human teaches or the data being provided by humans. The way the machines are taught or the data it is trained on can therefore be highly biased (Zhong 2018). There are three main issues that can arise specifically for accurately portraying minority groups in the data. First is the identification and selection of proxies for certain characteristics. Certain features might be less reliable collected from minority groups. This implies that if the reliability of a label is lower for minority groups than it is for the majority group, the system has lower accuracy for the prediction of the minority group due to noise. Second, the sheer amount of data for minority groups is lower, which means that it is harder to model that group in the context of the data. Finally, at times, sensitive attributes, such as race or gender are excluded from the training for a machine learning system. There are however often other features that then become proxies for sensitive attributes, such as neighborhood for race. If such features are included, the bias in the data remains, even if other attributes were actively excluded (Zhong 2018). In other words, in the big data context, 'missing values issues are exacerbated with the amount of available data' (Josse 2016, 62). In order to compensate for this missing data problem, unknown data points are mostly filled in through prediction, imputation and proxies (Williams et al. 2018).

In essence, new techniques and larger datasets give governments the opportunity to make predictions by using variables as proxies for excluded variables. Computer programs, such as machine learning and artificial intelligence algorithms allow imputation, where missing data points can be inferred by looking at, for example, similar people for whom data is available as well as patterns and correlations in recorded data can speak to

information outside of the dataset through proxy variables (Williams et al. 2018). If, however, these data are biased, they show a distorted picture of the population. In essence, the outputs from machine learning and other artificial intelligence analyses are limited to the accuracy of available data (Hashimoto et al. 2018). And this can have real-life effects in decision-making and public service delivery. This is what happened in a recent US example, where Obermeyer et al. (2019) find that a commercial algorithm used for healthcare had a racial bias attributed to label choice. This led to less money being spent on Black patients with the same level of need, because the algorithm had falsely concluded that they were healthier than equally sick White patients. The bias was based on the fact that the algorithm predicted healthcare costs rather than illness, 'but unequal access to care means that we spend less money caring for Black patients than for White patients. Thus, despite healthcare costs appearing to be an effective proxy for health by some measures of predictive accuracy, large racial biases arise' (Obermeyer et al. 2019, 447).

Another example from Global Health statistics is that a region or country's maternal mortality rate is often taken as an indicator for the general functioning of a healthcare system. Yet, in most countries where the maternal mortality rate is calculated this is done in the absence of reliable reporting, through a complex equation of several estimates—often build up from other estimates—and missing numbers (Wendland 2016). Wendland (2016) finds that effects of these uncertainties and missing data get exacerbated in the resulting number, that nevertheless is often treated as fact. But not only do actual problems of maternal mortality, and by extension healthcare systems, become invisible in this process, the political pressure on having successful maternal mortality rates can result in practices underreporting and therefore more missing data.

Secondary data gap

In a scenario where government is aware of a data gap and data is potentially available in different formats, such as social media data, or can be obtained in other ways, we speak of a secondary data gap. The following section highlights a host of issues that arise when government taps into these data sources in order to complete or replace existing datasets.

Statistical Offices are increasingly looking into big data in order to extract additional, relevant and reliable information for the statistical production process. However, this is not an easy task. These datasets, largely stemming from social media sources, are typically not designed by the Statistical Offices themselves, which means their structure and contents need to first be understood (Daas et al. 2015). These data are also more likely to be selective and not representative of the target population of interest. These concerns are paired with the need for specific technical expertise of Statistical Offices, such as advanced high-performance computing and data engineering, which is often not available or only applies to a small number of people (Daas et al. 2015). The development towards citizens sharing more and more data with private, social media platforms also changes the perception and role of Statistical Offices in two ways: First, people are less willing to fill out lengthy surveys—especially if the requested data has already been given to a government body. And second, official statistics are facing more competition. Hence, Statistical Offices have to look for new methodologies and forms of interpretation as well as working with a collaborative network in order to produce timely and relevant statistics to both the public and policymakers (Struijs et al. 2014).

The use of social media data also comes with a host of issues, which include construct, internal and external validity problems (Olteanu et al. 2019). Research has been most concerned with data quality and population biases. For the latter, this means that there are 'systematic distortions in demographics or other user characteristics between a population of users represented in a dataset or on a platform and some target population' (Olteanu et al. 2019, 6). For social media data specifically, it is described that a WEIRD (Western, Educated, Industrialized, Rich and Democratic) group is largely using different social media platforms (Heinrich et al. 2010; Hargittai 2018). There is also evidence that those with a higher socioeconomic status are more likely to be on several social media sites, creating more data points. In other words, 'opinions and behavioral traces of the more privileged are more likely to be represented in data sets that use social media as their sampling frames than the views and actions of the less privileged' (Hargittai 2018, 11). In addition, there is sparsity in social media data of rare elements or phenomena due to many measures following a power-law or heavy-tailed distribution (Baeza-Yates 2013). Further, noise can lead to incomplete, corrupted data or data containing typos/errors or content that is not reliable or credible (Boyd and Crawford 2012; Olteanu et al. 2019). Taken together, these aspects can also create platform-specific phenomena, which means that findings from one platform are hard to generalize to other platforms let alone to an entire population in order to be useful for policymakers (Tufekci 2014).

The idea of digital exclusion can thus be seen as a triple threat. While past research has been focusing on access to ICT and the ability to use it, there is the additional threat of generating unequal amounts and types of data that are then used for policymaking. A variety of research suggests that access to ICT 'is patterned along the lines of socioeconomic status, income, gender, level of education, age, geography and ethnicity' (Selwyn 2002, 5). This would then carry over into the data that is being extracted from the use or non-use of technology more generally and social media platforms specifically. In short, beyond the more simple issue of access/no access to ICT come more complex questions of levels of capability and data generation in that space.

In short, collecting additional data or data acquisition through sources, such as social media is difficult. Research highlights the danger for data collection bias, which is described as 'biases introduced due to the selection of data sources, or by the way in which data from these sources are acquired and prepared' (Olteanu et al. 2019, 13). This has to do with seeming availability of social media data, however many of the platforms are designed to disencourage data collection while also not capturing all relevant data. In times of big data, governments increasingly rely on proxy data to build a narrative from available data that were also often collected for other purposes. Governments rely on social media data specifically, which poses issues linked to the private ownership of the data as well as a lack of accountability as to who is represented in those datasets. For policymaking, as Battersby (2020) highlights for the case of food systems and food security, this can come at a cost to better understand the context of the problem being addressed as well as a limited understanding of who is missing in the data.

While more data are being collected and new ways found to utilize it for application, some emphasize the necessity of qualitative work in order to understand 'who counts and what counts as value added' (Jerven 2013, 112). Collecting data, for example, on HIV-positive patients 'lost to follow up' in ART (antiretroviral therapy) programs is often considered difficult, yet the absence of such data may lead to a biased evaluation of HIV programs. A meta-analysis of 32 studies that attempted to trace 'lost to follow-up' HIV patients in sub-Saharan countries provides significant results for effective policymaking, especially with respect to mortality, which is often underestimated in the absence of data

(Zürcher et al. 2017). Qualitative research on this particular patient group (e.g. Dlamini-Simelane and Moyer 2017) is pivotal to understanding why the gap emerges, as people do not follow up on treatment, in the first place. Targeted research to address data gaps may therefore be both effective for policymaking in understanding the gap and as a source for meta-analysis.

Hidden data gaps

Hidden data gaps describe datasets that are regularly used for policymaking, but contain misrepresentation, bias or missing data without governments being aware. As the following section will show, this has direct effects on public decision-making and service delivery.

These types of gaps are especially an issue when outputs of machine learning and other artificial intelligence analyses are applied to existing databases with a hidden data gap. Due to the underlying incomplete and outdated data, these methods can then result in faulty inferences about underrepresented groups. Systematic biases in existing datasets, such as clinical data, 'can affect the type of patterns AI recognizes or the predictions it may make, and this can especially affect women and racial minorities due to long-standing underrepresentation in clinical trial and patient registry populations' (Hashimoto et al. 2018, 73). In short, because data mining relies on training or existing data as ground truth, when those inputs are biased, the system will produce unreliable or even outright discriminatory results (Barocas and Selbst 2016).

Another dimension to this bias is that computer science needs to engage 'contextually with the relationships between technological interventions and social impact in both the short and long term' (Green 2019, 4). This applies when discrimination occurs due to incomplete or non-representative data, since there is no obvious method to adjust historical data. 'Corrective measures that alter the results of the data mining after it is complete would tread on legally and politically disputed terrain' (Barocas and Selbst 2016, 672). In a qualitative study of AIDS survey data gathering in Malawi, Biruk (2012) finds that it is exactly the standards of 'high-quality' and 'clean' data that make well-trained researchers manage uncertainty. As they use all their social skills to negotiate and elicit exact numbers from their respondents, real-world uncertainties of sometimes bored and annoyed interlocutors become obscured. While this example pertains to the social production of numbers in surveys, as the healthcare algorithm discussed by Obermeyer et al. (2019) shows, similar questions of what and who falls outside of data gathering or gets miscategorized or misrepresented in the process of gathering need to be asked for 'automatically' generated digital datasets.

An example of this is the 2002 finding by the Institute of Medicine (IOM) that there are longstanding and significant disparities in the US healthcare system based on racial or ethnic differences in the quality of the care, which are unrelated to access factors or clinical needs. This is, as Filice and Joynt (2017) lay out in their study, a result of the underlying data, rather than a product of programmers assigning certain factors inappropriate weight. 'Such a possibility has gone unrecognized by most scholars and policymakers, who tend to fear concealed, nefarious intentions or the overlooked effects of human bias or error in hand-coding algorithms' (Barocas and Selbst 2016, 674). Filice and Joynt (2017) find that the US Medicare program does not collect beneficiary race and ethnicity data themselves for its own records, but relies on information they receive from the Social Security Administration (SSA). The SSA receives this data as soon as someone applies for a Social

Security Number (SSN). This makes sense, since most Medicare beneficiaries become eligible as they become eligible for Social Security benefits. However, there are additional channels through which US citizens can receive Medicare where race and ethnicity data is not collected. For example, if someone is eligible through the Railroad Retirement Board (RRB) or if there are non-working spouses who are covered by Medicare, but have never received a SSN. For the latter group specifically, the spouses are classified as the same race/ethnicity as their wage-earning partner and make up about a fifth of Medicare beneficiaries (Filice and Joynt 2017). In addition, SSA data is not routinely updated, so the information collected at the time of application remains in the system unless there is a new application from the same beneficiary. This way of collecting data has been combined with methods to indirectly identify Medicare beneficiaries race and ethnicity.

Discussion and concluding remarks

The big data movement combined with more sophisticated methods in recent years has opened up new opportunities for governments to use existing data in different ways as well as fill data gaps through techniques that make predictions based on lookalike data. However, focusing specifically on the representativeness of such data—existing and newly created datasets—shows that these 'innovative and advanced methodological toolboxes' have trouble accounting for existing biases in the data as well as marginalized developments and cultural factors (Milan and Trere 2019). This applies to the input as well as the output side of such data analyses. For the input, the previous sections show that there is bias in the collection and use of certain datasets. This has to do with hidden data gaps in, for example, administrative data, but also in using social media data to fill known gaps without awareness around their selective nature. This also raises larger questions around sorting and labeling data to be entered into a dataset. Arora (2016) finds that big data architectures are setup in ways that reproduce existing prejudices. One example is the implementation of medical diagnostics software in the Himalayas where the survey on health issues of villagers to feed into the software did not account for illnesses related to social deprivations, such as chronic hunger or long hours in the field. This led to health symptoms having to be entered in the 'other' field.

For the output side, the interpretation and assessment of results are often done by data experts, not by domain experts. 'This is problematic as there are known differences in how non-experts and experts interact with and validate systems outputs' (Olteanu et al. 2019, 20). A similar point is being made by Isoaho et al. (2019) with regards to the use of computational algorithms for text analysis in policy research. They find that such data-driven methods require a genuine understanding of both the techniques applied as well as a 'contextual and semantic understanding' (Ibid, 10). In addition, they find that text modeling is unable to account for what is *not* represented in the frames used. Thus, such computational methods remain complementary to qualitative work and contextual understanding.

To fill some of these data gaps, the data has to further have a certain level of granularity. This means that larger data sets can be broken down by, for example, gender or ethnic group (without re-identifying individuals). However, this is often impossible in the way that data is collected and aggregated. As the UK Office for National Statistics points out, for some seemingly unreported indicators, data already exists, but 'cannot be fully disaggregated' and will be classified as a 'disaggregation gap'. They also pledge to actively collect such data where relevant. For example, for the Sustainable Development

Goals (SDGs), many indicators lack disaggregation by disability, income, ethnicity, age and sex. This has effects on the conclusions that can be drawn from the data as well as how informative it is for policymaking (UK Office for National Statistics 2018). This also identifies gaps that affect specific groups, as highlighted by the 'gender data gap' (Criado Perez 2019). As Criado Perez (2019) points out, 'if there is a data gap for women overall (both because we don't collect the data in the first place and because when we do we usually don't separate by sex), when it comes to women of color, disabled women, working-class women, the data is pratically non-existent'. This is because sex-disaggregated data is missing. A United Nation Women Global Study finds that funding for the implementation of policies related to women in post-conflict contexts remains 'inadequate' (UNW 2015; Criado Perez 2019). This has to do with data not being collected and divided by sex, which results in multiple examples where needs of women were not met. One case is that of Sri Lanka where after the 2004 Boxing Day Tsunami, rebuilding efforts lacked the inclusion of women, and as a result, homes were built without kitchens and the following inability to make food (Criado Perez 2019). Another example from the health sector highlights the use of standardized tests like the electrocardiogram or the physical stress test for a heart attack, which are less conclusive for women. This has to do with the 'standard' level biomarkers which are incorrect for women, resulting in later detection and thus higher death rates following a heart attack (Regitz-Zagrosek et al. 2016; Criado Perez 2019).

In addition, the focus on quantifying policy problems leads to different questions and ultimately different solutions in government. In an in-depth exploration of the role of counting and accounting in Global Health Policy today, Adams (2016, 8) concludes: 'One of the attractions of metrics is their ability to hold status as apolitical or politically neutral forms of evidence.' Of course, she continues to argue, the history of metrics shows how the act of counting is thoroughly political and indeed ironical, as Global Health Policy reproduces some of the exclusionary patterns of colonialism. For example, Livingston (2012) finds that statistical data on cancer in Africa have long been lacking, not because there is no cancer, but because global public health planning has continued a security-driven and racialized imagination of Africa as a place of infectious diseases. Therefore statistical collection 'has focused on disease transmission, vaccination coverage, births, and deaths' (Livingston 2012, 35). Nevertheless, the idea that 'numbers will offer unbiased, apolitical truths about health outcomes or health conditions' is still pervasive (Adams 2016, 8). The production and use of datasets, or metrics, not only in Global Health but in all policymaking is political. Modern governmental efforts to make social processes countable (Scott 1999) inevitably produces gaps. With the abundance of data that is gathered in our present time, and the importance of these 'big data' to policymaking, it has become more vital than ever that we identify and learn to address data gaps in order to move towards inclusive policymaking.

Crucially, for politically marginalized groups being included in data collection efforts might also have adverse political effects (Taylor and Schroeder 2015). For example, 'low-income communities are among the most surveilled communities in America' (Waddell 2016, 1). Public benefits programs, such as child welfare or domestic abuse agendas continuously gather data on their largely poor users. This data, in turn, is being fed back into (predictive) police systems and can cut citizens off from job or loan applications (Waddell 2016). The underlying implication of the issues raised here is that the efforts and resources put into the datafication of policy will affect the economic opportunities, social mobility and democratic participation of marginalized groups. As Lerman (2013) warns, 'these technologies may create a new kind of voicelessness, where certain groups' preferences and behaviors receive little or no consideration when powerful actors decide how to

distribute goods and services and how to reform public and private institutions' (Ibid, 59). Hence, the big data revolution may create new forms of inequality that are harder to detect and whose impact is more difficult to predict.

Taken together, this hints toward how political data can be and how this is often hidden among discussions on methodology and data itself. In order to unravel these politics, Prada-Uribe (2012) highlights the knowledge and the governance effect that datafication has. The knowledge effect describes the ability of data to spread certain knowledge as a universal truth even if the underlying notion—specifically linked to certain groups or countries—is still contested. And at the same time indicators can have a normative effect 'because they produce the standards against which a society's development ought to be measured' (Prada-Uribe 2012, 7). This results in a governance effect where policies strive to accommodate such measures without a check whether they measure relevant dimensions. Battersby (2020) further suggests that the knowledge and the governance effect have a mutually reinforcing relationship, which results in 'politics of measurement'. In order to resolve this, governments need to understand existing gaps in the data as well as what they obscure and why and find solutions for adding additional knowledge through innovative and traditional ways of data collection.

References

Adams, V. (2016). Introduction. In V. Adams (Ed.), *Metrics: What counts in global health* (pp. 1–17). Durham, London: Duke University Press.

Arora, P. (2016). The bottom of the data pyramid: Big data and the global south. *International Journal of Communication, 10,* 1681–1699.

Baeza-Yates, R. (2013). Big data or right data? In *Proceedings of the 7th Alberto Mendelzon international workshop on foundations of data management (Puebla; Cholula)*.

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review, 104,* 671–732.

Battersby, J. (2020). Data gaps and the politics of data: Generating appropriate data for food system assessment in Cape Town, South Africa. In A. Blay-Palmer, D. Conare, K. Meter, A. Di Battista, & C. Johnston (Eds.), *Sustainable food system assessment, lessons from global practice* (pp. 93–110). London, New York: Routledge.

Biruk, C. (2012). Seeing like a research project: Producing "high-quality data" in AIDS research in Malawi. *Medical Anthropology, 31*(4), 347–366.

Boyd, D., & Crawford, K. (2012). Critical questions for big data. *Information, Communication & Society*, *15*(5), 662–679.

Criado Perez, C. (2019). *Invisible women, exposing data bias in a world designed for men*. London: Random House.

Daas, P. J. H., Puts, M. J., Buelens, B., & van den Hurk, P. A. M. (2015). Big data as a source for official statistics. *Journal of Official Statistics, 31*(2), 249–262.

Desouza, K. C., & Smith, K. L. (2014). Big data for social innovation. *Stanford Social Innovation Review, Summer, 2014,* 39–43.

Dlamini-Simelane, T. T., & Moyer, E. (2017). 'Lost to follow up': Rethinking delayed and interrupted HIV treatment among married Swazi women. *Health Policy and Planning, 32*(2), 248–256.

Filice, C. E., & Joynt, K. E. (2017). Examining race and ethnicity information in medicare administrative data. *Medical Care, 55*(12), 170–176.
Giest, S. (2017). Big data for policymaking: Fad or fasttrack? *Policy Sciences, 50,* 367–382.
Green, B. (2019). 'Good' isn't good enough. In *Conference paper, AI for social good workshop at NeurIPS (2019), Vancouver, Canada*.
Hand, D. J. (2020). *Dark data: Why what you don't know matters, a practical guide to making good decisions in a world of missing data*. New Jersey: Princeton University Press.
Hargittai, E. (2018). Potential biases in big data: Omitted voices on social media. *Social Science Computer Review, 38*(1), 10–24.
Hashimoto, D. A., Rosman, G., Rus, D., & Meireles, O. R. (2018). Artificial intelligence in surgery: Promises and perils. *Annals of Surgery, 268*(1), 70–76.
Heinrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*(2), 1–75.
Isoaho, K., Gritsenko, D., & Mäkelä, E. (2019). Topic modeling and text analysis for qualitative policy research. *Policy Studies Journal*. https://doi.org/10.1111/psj.12343.
Jerven, M. (2013). *Poor numbers: How we are misled by African development statistics and what to do about it*. Ithaca: Cornell University Press.
Josse, J. (2016). *Contribution to missing values and principal component methods*. Statistics, HAL Archives ID: tel-015734993. Retrieved April 20, 2020, from https://hal.archives-ouvertes.fr/tel-01573493/document.
Klievink, B., Romijn, B., Cunningham, S., & de Bruijn, H. (2017). Big data in the public sector: Uncertainties and readiness. *Information Systems Frontiers, 19,* 267–283.
Lerman, J. (2013). Big data and its exclusions. *Stanford Law Review Online, 66,* 55–63.
Livingston, J. (2012). *Improvising medicine, an African oncology ward in an emerging cancer epidemic*. Durham: Duke University Press.
Mergel, I., Rethemeyer, R. K., & Isett, K. (2016). Big data in public affairs. *Public Administration Review, 76*(6), 928–937.
Milan, S., & Trere, E. (2019). Big data the south(s): Beyond data universalism. *Television and New Media, 20*(4), 319–335.
Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science, 366*(6464), 447–453.
Olteanu, A., Castillo, C., Diaz, F., & Kiciman, E. (2019). Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data, 2*(13), 1–33.
Prada-Uribe, M. A. (2012). *Development through data? A case study on the World Bank's performance indicators and their impact on development in the Global South*. IRPA Research Paper No. 5. Retrieved April 20, 2020, from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2167366.
Regitz-Zagrosek, V., Oertelt-Prigione, S., Prescott, E., Fanconi, F., Gerdts, E., Foryst-Ludwig, A., et al. (2016). Gender in cardiovascular diseases: Impact on clinical manifestations, management, and outcomes. *European Heart Journal*. https://doi.org/10.1093/eurheartj/ehv598.
Scott, J. C. (1999). *Seeing like a state: How certain schemes to improve the human condition have failed*. New Haven: Yale University Press.
Selwyn, N. (2002). 'E-stablishing' an inclusive society? Technology, social exclusion and UK government policy making. *Journal of Social Policy, 31*(1), 1–20.
Struijs, P., Braaksma, B., & Daas, P. J. (2014). Official statistics and big data. *Big Data and Society, 1*(1), 1–6.
Taylor, L., & Schroeder, R. (2015). Is bigger better? The emergence of big data as a tool for international development policy. *GeoJournal, 80,* 503–518.
Tufekci, Z. (2014). Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In *ICWSM '14: Proceedings of the 8th international AAAI conference on weblogs and social media, 2014*.
UK Office for National Statistics. (2018). *UK data gaps, inclusive data action plan towards the global sustainable development goal indicators*. Retrieved March 18, 2020, from https://www.ons.gov.uk/economy/environmentalaccounts/articles/ukdatagapsinclusivedataactionplantowardstheglobalsustainabledevelopmentgoalindicators/2018-03-19.
United Nation Women (UNW). (2015). *Preventing conflict, transforming justice, securing the peace*. A global study on the Implementation of United Nations Security Council resolution 1325. Retrieved March 18, 2020, from https://reliefweb.int/sites/reliefweb.int/files/resources/UNW-GLOBAL-STUDY-1325-2015.pdf.
Vydra, S., & Klievink, B. (2019). Techno-optimism and policy-pessimism in the public sector big data debate. *Government Information Quarterly, 36*(4), 101383.

Waddell, K. (2016). *How big data harms Poor communities*. The Atlantic, April 8, 2016. Retrieved April 20, 2020, from https://www.theatlantic.com/technology/archive/2016/04/how-big-data-harms-poor-communities/477423/.

Wendland, C. L. (2016). Estimating death: A close reading of maternal mortality metrics in Malawi. In V. Adams (Ed.), *Metrics: What counts in global health*. Durham: Duke University Press.

Williams, B. A., Brooks, C. F., & Shmargad, Y. (2018). How algorithms discriminate based on data they lack: Challenges, solutions, and policy implications. *Journal of Information Policy, 8,* 78–115.

Zhong, Z. (2018). *A tutorial on fairness in machine learning*. Medium, October 22, 2018. Retrieved on April 20, 2020, from https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb.

Zürcher, K., Mooser, A., Anderegg, N., Tzmejczyk, O., Couvillon, M. J., Nash, D., et al. (2017). Outcomes of HIV-positive patients lost to follow-up in African treatment programmes. *Tropical Medicine & International Health, 22*(4), 375–387.