This is a repository copy of *Big data and data analytics*.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/169086/

Version: Accepted Version

Big Data and Data Analytics

The emergence of social research about big data analytics

In 2012, it was claimed by IBM that 90% of all data in existence had been generated in the previous two years. This was, according to many, the era of "datafication" (van Dijck, 2014); a term used to refer to the processes by which masses of "big data" are generated about aspects of life that were not previously quantifiable. Vast reams of data were being produced by new technologies such as social media platforms (boyd and Crawford, 2012) and smart city infrastructures (Kitchin, 2013). The analytical possibilities of this process of datafication were significant, as became evident for much of the public in 2013 when Edward Snowdon blew the whistle on the data-driven surveillance practices of the USA's National Security Agency and other countries' security services, illuminating how they had intercepted metadata relating to billions of phone calls and online interactions (van Dijck, 2014). By 2014, the term "Big Data" had been added to consultancy firm Gartner's annual hype cycle as industry turned its attention to what this new resource might offer.

It wasn't only the tech sector and consultancy firms that were paying attention to these processes of datafication and the emergence of big data analytics. Even prior to the hype surrounding 'Big Data' the emerged around 2013, social scientists and humanities scholars had begun to question the implications of increasing amounts of 'transactional data' for their disciplines. Savage and Burrows (2007), for example, had pointed to a "coming crisis in empirical sociology". No longer was the social survey – the bread and butter method of quantitative sociologists – necessarily the best way to understand populations. Organisations were increasingly conducting their own in-house sociological research using the transactional data generated as consumers interacted with their systems. They observed how these data, which academic sociologists rarely had access to, offered much more fine-grained insights into, for example, people's social networks, than the resource intensive methods of academia.

Other observers noted both the significant potentials and challenges that "big social data" raised for the humanities (Manovich, 2011) and social sciences (Ruppert, 2013). Manovich (2011) observed that despite the interesting opportunities that such data opened up for the humanities, beyond researchers working in companies such as Facebook and Google, most had limited access to be able to use the data, and the skills required to answer interesting humanities questions using such data were not typically something that humanities scholars have. He observed the emergence of "new kinds of divisions" between "those who create data…those who have the means to collect it, and those who have expertise to analyze it" (p. 10). Ruppert (2013) made similar observations from the perspective of social science, calling for new modes of interdisciplinary collaboration that are able to develop innovations in methods that allow researchers to "innovatively, critically and reflexively engag[e] with new forms of data" (p. 270), and foster a "critique from within" that is able to grapple with issues of privacy, rights and ethics.

Beyond questions about the methodological implications of Big Data on academic social science and humanities, the gaze of scholars began to turn more generally in the direction of the emergent phenomena of Big Data analytics. Early contributions came from those whose work had positioned them close to the newly emerging technologies. danah boyd and Kate Crawford (2012), social scientists based at Microsoft Research, published an essay posing six "critical questions for big data", with a particular focus on social media data. They critiqued the claim of some within the tech sector that, with the advent of big data, numbers had begun to 'speak for themselves'. Instead, they

recognised that the much heralded objectivity and accuracy of big data was misleading. Too often, they observed, patterns could be seen in big data where none actually existed, for example correlations between stock indexes and butter production in Bangladesh. Big data did not always mean better data, and understanding the sample was, in the age of big data, more important than ever. For example, it was vital to recognise that Twitter data did not represent all people – only what Twitter users chose to say on Twitter. Understanding big data in context, they argued, was crucial for interpretation, and just because researchers have access to data does not make it ethical to analyse it. Finally, they made similar observations to Manovich about the emerging divides related to which researchers had access to such data and the skills to analyse it, and which did not.

Urban geographer Rob Kitchin, was also beginning to pose similar questions with a focus on the emergence of Smart Cities. He questioned the claim that big urban data, often generated in 'Smart City' contexts, are neutral and lacking in ideology, noting that such data do not exist independently of, for example, ideas, people and social contexts. He also raised concerns about the surveillance dynamics of emergent Smart City infrastructures. He observed that as urban data systems became more centralised through Smart City developments, they also become more panoptic in design: "rais[ing] the spectre of a Big Brother society based on a combination of surveillance (gazing at the world) and dataveillance (trawling through and inter- connecting datasets), and a world in which all aspects of a citizen's life are captured and potentially never forgotten" (Kitchin, 2013, p. 11).

These early interventions from social scientists also raised the question of what big data actually are. Much of the tech sector had adopted the 3Vs definition: Big Data were data that were so large in volume, variety and velocity that they could not be processed and analysed using standard data technologies. However, as Kitchin and McArdle (2016) observed definitions sometimes also involved other attributes such as exhaustivity, resolution, indexicality, relationality, extensionality and scalability (p. 2). Through examination of a number of supposedly 'big' datasets, Kitchin and McArdle (2016) observed that only a few datasets had all these traits and some were not even significant in volume or variety as the original 3Vs definition implied they ought to be. They concluded that 'Big Data' as a category of data needed to be further unpacked. Van Dijck (2014) also drew attention to the necessity to understand the ways in which metadata (data about data) make up big data sets, particularly in the wake of the Snowdon revelations which identified how national security services apply big data analytics techniques to communications metadata.

Kitchin and McArdle (2016) were not alone in questioning what 'big data' are. boyd and Crawford (2012) had already defined 'big data' not only as a technological and analytical phenomenon, but also a mythological one. The mythology of big data that they embedded in their definition related to the widespread beliefs surrounding 'big data', specifically that big data "offer a higher form of intelligence and knowledge that can generate insights that were previously impossible, with the aura of truth, objectivity, and accuracy" (p. 663). Van Dijck referred to this mythology as "the ideology of dataism" (p. 198). As Beer (2016) argued social scientists and humanities scholars, "need[ed] to take the concept of 'Big Data' seriously…[and] explore the type of work that is being done by that concept" (p. 1), as well as the work being done by applications of so-called big data.

Emerging from these early provocations by social scientists and humanities scholars, a new sub-field of Critical Data Studies, made up of researchers from disciplines including Geography, Media and Communications and Information Studies, began to take shape. Work aiming to define the focus of the 'Critical Data Studies' field was published in Human Geography e.g. Kitchin and Lauriault (2014) and Dalton and Thatcher (2014) and Communications (Iliadis and Russo, 2016). A field specific journal was launched in 2014 - Big Data and Society, edited by sociologist Evelyn Ruppert; and field specific conferences were also launched e.g. 'Data Power' in 2015 (http://datapowerconference.org/) and 'Data Justice' in 2018 (https://datajusticelab.org/data-

justice-conference/). While research interest in Big Data and related fields has expanded across the different disciplines of the social sciences and humanities in recent years, and many disciplinary conferences now have 'data' related tracks, the loosely formed sub-field of critical data studies provides a space in which transdisciplinary insights about 'Big Data' and related analytics are fostered. The following sections will go on to discuss the major claims and developments in this field, as well as its principle contributions, criticisms and possible future developments.

Major claims and developments in the field: philosophical and political

As overviewed above, researchers in the field have advanced a number of arguments in relation to big data and related analytics. Key arguments have been (1) philosophical in orientation, examining the epistemological and ontological claims and implications of big data, and (2) political in orientation, examining issues related to power and governance in the context of big data. While some contributions are more clearly philosophical or political, many researchers in the field draw on both philosophical and political arguments, to varying degrees, in order to advance their claims. This section begins first by considering philosophical critiques of, and insights about, big data, both epistemological and ontological. It then moves on to examine political claims about big data systems, including issues of dataveillance, discrimination, transparency and exploitation. The section finishes by exploring the implications for collective agency in the context of advances in big data analytics.

Philosophical critiques of big data

One of the early critiques that social scientists and humanities scholars advanced in relation to big data was an epistemological critique of the claim that big data had the capacity to generate new, superior forms of knowledge. Epistemology is the branch of philosophy that deals with the nature of knowledge, for example how it is created and what constitutes valid knowledge. The hype surrounding big data had included a number of claims about the far-reaching potential of big data to overturn earlier ways of creating knowledge. These arguments echoed some of the points made by sociologists Savage and Burrows (2007) in their paper on the coming crisis in sociology, however they went further and with less nuance. Leading the way was an audacious claim made in 2008 by Chris Anderson - the editor of tech magazine Wired - that big data analytics made the scientific method obsolete, that we should forget every theory generated by the natural and social sciences, and simply let the numbers "speak for themselves". His ideas caught on in some of the industry hype surrounding 'big data' in the following years, and, unsurprisingly, social scientists and humanities scholars took the bait coming down hard on Anderson's claims.

The epistemological critique of 'big data' centred on a number of key points:

(1) **Data are never "raw", neutral or self-evident representations of the world.** Big data, in particular data generated as a by-product of some other activity (e.g. social media practices) or using sensors (e.g. location or environmental sensors), were often presented by big data advocates as an essentially accurate and truthful representation of phenomena, void of any social or political bias or interference. Early interventions from social scientists and humanities scholars often centred on this claim. Lisa Gitelman's (2013) edited collection, *"Raw Data" is an Oxymoron*, was a key contribution to this debate. As was work by Kitchin (2013; 2014), boyd and Crawford (2012) and van Dijck (2014). All of these critiques advanced, in their own ways, the argument that data are never "raw", socially neutral

representations of the world, rather they are manufactured in complex social contexts shaped by ideologies, beliefs, finance etc.

(2) **The outputs of big data analytics are not neutral representations of the world.** Similar to the argument that data were not "raw", critics also recognised that the analysis of data and interpretation of patterns and relationships in data is also undertaken by people and shaped by their socio-cultural biases, assumptions, contexts (Kitchin, 2014; D'Ignazio and Klein, 2020)

(3) *N ≠ all, data are always partial.* A frequent claim advanced in the hype of big data, was that such data are comprehensive and offer full resolution; that there was no longer any need to sample because data represented the full population (n=all). Critics such as Kitchin (2014) and boyd and Crawford (2012) instead recognised that despite the increasing volume of datasets, big data were still always in some way partial representations of the world, and that it was vital to understand the ways in which any given dataset was incomplete in relation to whatever the analyst was trying to understand. In simple terms, a Twitter dataset only represents (likely a sub-set of) Twitter users and what they share on Twitter, and is not representative of the thoughts of the full population as some big data advocates assumed.

(4) **Big data as a technology for knowledge production is theoretically driven.** While advocates had proclaimed that 'big data' meant the end of theory, critics re-joined with the observation that big data systems are not themselves free of theory and philosophy, rather they were created for a purpose and guided by an agenda, and the analytical techniques and algorithms they use are based upon scientific reasoning and established theories (Kitchin, 2014)

More recently, the epistemological debates around big data have drawn on feminist and decolonial lenses to illuminate some of the ways that historically constituted social injustices become embedded in the 'universalist' knowledge claims emerging from many big data systems. D'Ignazio and Klein's (2020) "data feminism" lens, for example, draws attention to who is engaged in the practice of data analytics, and what this may mean for the knowledge claims that are produced by the field. They observe that it is common for people from "dominant groups" to be the ones who do and benefit from big data analytics, and whose priorities tend to get turned into products. Milan and Treré (2019), take this line of critique further to observe that the critique of big data analytics that has developed in the social sciences also has its own biased epistemology. They argue that research in the field has so far tended to "take as frame of reference the liberal democracies of the West, with their sociocultural substrate and long tradition of representative institutions, rule of law, and citizen involvement in public affairs". Those outside this frame of references, they argue – "the different, the underprivileged, the silenced, the subaltern, and the 'have nots'" who are not necessarily connected to particular geographical locations - have remained in a "blind spot" (p. 320), and their knowledges and experiences need to be centred.

As well as advancing an epistemological critique of big data and the hype and critique surrounding it, social scientists and humanities scholars have also explored the ontological dimensions of big data systems. Ontology refers to the philosophical study of the nature of being e.g. existence, reality etc, and how different entities relate to one another. Big data systems raise a number of ontological questions, for example, Ruppert (2013) asks what the implications of big data might be on the nature of who people are, as individuals and societies. Others have addressed the question of what entities make up a big data system, and what this might mean in relation to the social shaping and implications of big data. Kitchin, for example, advance the idea of a 'data assemblage'; a complex of socio-material entities including cultural (e.g. ideologies, beliefs, norms), social (e.g. regulations, communities) and material (e.g. technologies, infrastructures, investments) factors that all inter-relate to shape how data are produced, processed and used (Kitchin and Lauriault, 2014; Kitchin, 2014). This interest in the materiality of big data systems, both in terms of their constitution and

implications, has been explored by a number of researchers in the field. For example, Bates et al (2016) examine the socio-material constitution of data objects and flows in their work on 'Data Journeys', and Lupton (2018) draws on various theories of materialism to illuminate "the material and embodied dimensions of human–data assemblages as they grow and are enacted, articulated and incorporated into everyday lives".

As Couldry and Powell (2014) assert, the "mutual intertwining of human and material agency is hardly a new insight" (p. 3), but the observation takes on increased significance for research in a context in which big data practices are largely opaque to non-experts. Through drawing on a variety of materialist traditions, such scholars have avoided an idealist critique of big data systems which would understand big data to be driven only by things such as beliefs, ideas and discourses, to illuminate the complex assemblage of interrelated social and material factors that produce 'big data' and the work it does in the world. As articulated by Cote et al (2016):

> "This is not to suggest that Big Data – more specifically processes of datafication -
> – are best or at all understood as socially constructed. Indeed, discursive analysis or
> unreconstructed social theory cannot fully grasp how data re-articulates the social, cultural,
> political and economic in a deeply recursive manner. Thus, any political reckoning must
> equally account for the materiality of data, alongside the logic guiding its processes and the
> practices that deploy its tools." (p. 5)

Political critique of big data

Beyond the philosophical questions around what data are, how they come to be, what work they do, and what kinds of knowledge claims it is possible to make with them, a further – and often related – focus of research about big data analytics has been the politics of data. While there is an underlying politics behind much of the work cited in the above section, two papers published in 2016 and 2017 explicitly layout the case for a focus on "data politics" within the field (Cote et al, 2016; Ruppert et al, 2017).

While stopping short of articulating a singular framework for illuminating data politics, Cote et al (2016) instead argue that to "forensically unpack the value-laden information and knowledge" produced through data analytics, what is required is a "political critique [that] entails questions of data access, technological understanding and capacities, and the ability to critically examine the algorithms of data analytics" (p. 9). Not only was it vital to understand how data and their analysis are constituted in complex socio-material contexts, it was also crucial to pay attention to concentrations of economic power and ownership within the data economy, and how algorithms "enact new procedures of power and knowledge" (p. 8). While Cote et al's (2016) conceptualisation of data politics remains focused on the political constitution of the outputs of big data analytics, Ruppert et al (2017) place more emphasis on the ways in which data has been constituted as a political object with particular "powers, influence and rationalities". They argue that data has become political because it "reconfigures relationships between states and citizens" (p. 1). It has become "a force that is generative of…new forms of power relations and politics at different and inter-connected scales" (p. 2). In the big data era, virtual and actual worlds interact to produce subjects whose political rights become "objects of struggle" in the field of data politics.

While these approaches to defining 'data politics' have their differences, they are similar in their emphasis on issues of power, struggle and relations between different social groups. Within the wider literature on big data analytics, these issues are explored by authors who, on the one hand, frame big data in terms of its capacity for oppression through surveillance, discrimination, exploitation and a lack of transparency, and those that refocus this framing with an emphasis on

agency and empowerment of citizens in the context of big data systems. The following section will outline these debates in more detail.

- **Surveillance, dataveillance and the data gaze.** Researchers of surveillance practices began to pay attention to data-driven systems earlier than many within the social sciences and humanities. For example, Louise Amoore's (2006) research on the expansion of "biometric borders" in the early 2000s which examined the ways in which personal data was increasingly being used to "classify and govern the movement of people across borders" (p. 341).

  Others have also worked to unpack the differences between traditional forms of surveillance and contemporary practices of "dataveillance". Kitchin (2013), for example, notes how while surveillance involves "gazing at the world", the term dataveillance emphasises the practice of "trawling through and interconnecting datasets" (p. 11), and in a piece published post-Snowdon revelations, communications scholar Jose van Dijck (2014) noted that while "surveillance presumes monitoring for specific purposes" the dataveillance enabled by big data analytics "entails the continuous tracking of (meta)data for unstated preset purposes" (p. 205).

  In the Smart City context, Kitchin (2013) pointed to the ways in which big data systems were leading to a centralisation of urban surveillance systems in city control rooms, while in other contexts researchers observed that surveillance systems were becoming more distributed and less panoptic (Galic et al, 2015). These and other studies observed the different ways in which data-driven surveillance systems were shifting the relationship between state and citizens, and the ways that already marginalised citizens could become caught up in these systems, for example because the data and algorithms falsely concluded they were a national security risk. These novel forms of surveillance were not only perceived to impact the power relation between state and citizens, but also consumers and commercial organisations such as platform companies (e.g. Google, Facebook).

- **Discrimination and biased systems.** As big data analytics systems were adopted in more and more areas of society it fast became apparent that the discriminatory and unjust impacts that had been observed in border control and state surveillance systems, were also a feature of other types of big data systems. One area of significant research activity is related to bias in search engines. Using a variety of methodological approaches, researchers have observed how the results of search engines such as Google and Bing have significant racial and gendered biases (Noble, 2018; Otterbacher et al, 2017). What was quickly becoming apparent was that these systems and the data they were ingesting and learning from reflected, and in many cases exaggerated, the significant social injustices related to racism and sexism that are present within the societies that produced the data and systems. While some computer scientists have begun work to try and fix these biased systems (e.g. those engaged in work on 'FATE' (Fairness, Accountability, Transparency and Ethics)), many social scientists and humanities scholars are concerned that the kinds of technical fixes being proposed are unlikely to be sufficient to address such a complex socio-technical problem.

- **Exploitation and colonisation.** Scholars drawing on a Marxist understanding of social relations have long pointed to the exploitative dynamic that sits behind people's online activity, recognising internet users' online activity as a form of 'digital labour' that generates value for capitalist platform owners. More recently, some social theorists have expanded this critique to the wider Big Data context. Thatcher et al (2016), for example, argue that the capturing of 'big data' about people is a form of accumulation by dispossession; an "asymmetric power relationship in which individuals are dispossessed of the data they

generate in their day-to-day lives" (p. 990). A similar argument is put forward by Couldry and Mejias (2019), who also argue that the ways in which data is captured and processed by companies is a form of "appropriation". Both sets of authors argue that these processes of appropriation should be understood as a contemporary form of colonialisation that they term "data colonialism" (Thatcher et al, 2016; Couldry and Mejias, 2019). This understanding of big data as 'colonising' has also been proposed by theorists who adopt a different – non-Marxist – perspective. Ruppert et al (2017), for example, argues that "data colonizes minds…[and] lifeworlds" (p. 5), and Milan et al (2019) adopt a decolonial lens to critique the ways in which research about big data analytics is biased towards a frame of reference centred on the liberal democracies of the West.

- **Black-boxed systems.** These issues of dataveillance, discrimination and exploitation are, Frank Pasquale (2015) argues, further compounded by a further colonisation: a colonisation by "the logic of secrecy" (p. 2). While big data systems have made citizens far more transparent to the watchful eyes of states and companies, these same institutions have in many cases become more secretive, particularly in relation to the functioning of the 'black-boxed' big data systems that increasingly inform their decision making. With a focus specifically on big data systems used in Wall Street (finance) and Silicon Valley (internet), Pasquale explores the different tactics firms use to keep secrets about their black-boxed systems and argues the case for transparency.

While there has been significant focus from social scientists and humanities scholars on the oppressive nature of big data systems, a further body of work has examined the possibilities of agency within this context. As Kennedy et al (2015) assert in the introduction to a special issue on the topic of 'Data Agency', "thinking about agency is fundamental to thinking about the distribution of data power" (p. 2). Early contributions exploring such questions include Couldry and Powell's (2014) essay in which they argue for more attention to be paid to "agency and reflexivity" than theories which only emphasise oppressive forms of algorithmic power. They identify two ways in which this might be done. First, examination of the ways that e.g. community organisations can use data analytics to further their own agendas, for example NGOs such as Mapping for Change. Second, researching people's attempts to create an information economy that is "more open to civic intervention". They conclude by arguing that it might be important for those in the field to highlight not only the risks of big data systems, but also the potential opportunities – "ambiguous as they may be" (p. 4). This line of argument is also presented by Taylor et al (2014) in a paper that reports on ideas emerging from a workshop about how big data might become a "resource for positive social change in low- and middle- income countries (LMICs)" (p. 418). They argue that there are four potential opportunities big data offers in this context: "advocacy; analysis and prediction; facilitating information exchange; and promoting accountability and transparency" (p. 418), yet across these opportunities they recognise that there are challenges relating to issues of privacy and open data. They argue that it is crucial for civil society groups from LMICs to become engaged in the debates around big data practices.

Related to these debates around "data agency" others have explored how societies should respond to emergent data practices in ways that fit with normative values such as ethics and justice. Some of this work is happening in emergent interdisciplinary fields such as Fairness, Accountability and Transparency, Ethics (FATE), which brings together researchers working primarily in the fields of Computer Science, Ethics and Law to develop technical and legal interventions that mitigate against the risks big data analytics poses with regard to privacy and discrimination. However, as some social scientists and humanities scholars have observed, in this field concerns with big data can be framed quite narrowly as a specialist issue focused primarily on technical and legal solutions. Some social scientists have instead argued that responses to emergent data practices should be framed more

broadly through centring the discussion on notions of "data justice" that "relate to long-standing social, political, economic and cultural issues" (p. 873), rather than being specifically technical problems requiring techno-legal solutions (Dencik et al, 2019)

A further line of debate within this area is whether ethics offers a satisfactory framework to address problems identified with big data analytics (Rességuier and Rodrigues, 2020). Many have observed how the discourse around data ethics has in recent years been co-opted by big tech companies and others in the tech sector to 'ethics wash' their products without any deep consideration of their social implications. Social scientists and humanities scholars have therefore questioned whether 'ethics' alone is sufficient to address the problems with big data analytics that researchers and social justice campaigns have observed. Dencik et al (2019), for example, observe that there is a need to "position data in a way that engages more explicitly with questions of power, politics, inclusion and interests, as well as more established notions of ethics, autonomy, trust, accountability, governance and citizenship" (p. 874).

Principal contributions of the field: conceptual, methodological, empirical and community building

The major claims and research developments highlighted in the previous section can be understood as mapping onto a number of key contributions made by the field. Core contributions to knowledge are categorised here as (1) conceptual, (2) methodological and (3) empirical contributions. A fourth contribution is also identified as the creation of research communities interested in critical examination of issues related to Big Data and Analytics from social science and humanities perspectives.

Conceptual contributions

As Kitchin (2014) argued in his book The Data Revolution, "there is a need to develop conceptual and philosophical ways to make sense of data" (p. 25). In 2014, when the book was published, there was a relatively limited conceptual apparatus through which to make sense of the emergence of big data and related analytics, but it was not long before scholars stepped into the space with a variety of conceptual tools that could be used to help illuminate what was unfolding.

Early concepts tended to describe what observers saw. Terms such as "Datafication" for example aimed to describe the unprecedented ways in which more and more aspects of life were being quantified through capture of data via devices and sensors. Similarly, the term "dataveillance" aimed to capture the shift that big data enabled in the context of surveillance allowing watchers to trawl through and interconnect databases (Kitchin, 2013, p. 11).

Researchers also began conceptualising the context that was shaping – and being shaped by – these emergent big data practices. Another contribution from Kitchin and Lauriault (2014) within this area was the concept of the "data assemblage". This concept aims to capture the ways in which big data systems are complexly intertwined with a variety of technical, political, social and economic apparatuses that "frame[] their nature, operation and work" (Kitchin and Lauriult, 2014, p. 7; see also Kitchin, 2014).

A further body of conceptual work has aimed at developing normative concepts that aim to promote a direction of travel for how societies respond to big data and analytics, as well as offering a space for critique of existing practices. The concept of "data agency" (Kennedy et al, 2015) introduced above, for example, emphasises the importance of paying attention to the important ways that data might be used by ordinary people in order to further their own agendas. Similarly, the notion of

"data justice" (Dencik et al, 2019) moves beyond responses that are overly determined by technical and legal responses in order to offer a deeper political critique of what social justice might look like in a datafied society.

While it has been critiqued for its primarily Western orientation (e.g. Milan and Treré, 2019), the above reflects the beginnings of a conceptual framework developed by key scholars in the field of critical data studies that provides an initial tool box for researchers in the field to think with, and to adapt and further develop as they undertake their own investigations into the implications of big data analytics on societies.

Methodological contributions

A further area of contribution has been in relation to advancing methods for conducting empirical research aimed at exploring the interrelation between the relationship between data and society. Lupton (2019; 2020) and colleagues have gathered together a variety of innovative methods that have been developed by researchers in different fields in order to examine the how people make sense of data in the contemporary era, as well as developing innovative methods of their own such as Data Letters and Data Kondo. They list a large number of methods from the field of Human-Computer Interaction including 3D printing of data, data selfies, data comics, among others, as well as highlighting a number of methods developed within the social sciences to explore these issues. One interesting feature of social science methodological innovations in this field has been an interest in mobility, and below we explore some of these methods in more detail.

One interesting methodological contribution that places an emphasis on mobility is that of Data Walks. Powell's (2018) "data walks" or "datawalkshop" method, draws on a tradition of methods that involve walking, including psychogeography. The approach is a participatory methodology, bringing together aspects of both research and public engagement, to engage people in discussion about the data processes that they observe in the urban environment, and exploring them from a particular perspective or matter of concern e.g. surveillance or discrimination. Jarke (2019) builds on this idea of the data walk in her co-design research with older adults in Germany. Through embedding data walks in her research process and evaluating the approach in the context of exploring open government data from the perspective of older citizens, she observes how data walks can be used in complementary ways to help answer questions at different stages of the design process of data-driven systems.

A different methodological approach that draws upon the idea of mobility is Bates et als (2016) Data Journeys methodology. With a focus on the mobility of data, the authors develop a methodological approach to capture the socio-material dynamics that shape the movement of data as they move between different sites from when they are initially created to when they are used in different contexts. Through following the data and using mobile ethnographic methods to capture cultures and contexts of the people that work with data in different places, they are able to shed light on some of the driving forces behind the circulation of data through interconnected big data systems.

Interesting methodological contributions are also starting to emerge through collaborations between social scientists or humanities scholars and computer scientists, some of whom have been inspired by critical debates about big data and analytics in recent years. Some of this work is visible at computer science conferences such as CHI (ACM Conference on Human Factors in Computing Systems) and ACM FAccT (ACM Conference on Fairness, Accountability, and Transparency). While a number of social scientists and humanities scholars have called for such collaboration, and have been actively involved in such work over recent years, there is still more scope for deepening methodological innovation through interdisciplinary engagement between the social and technical disciplines to examine the challenges of datafied societies and shape future developments.

Empirical contributions

While the work cited in this chapter has been primarily of a conceptual or commentary nature, many cited authors based their arguments on their in-depth empirical work that has often been written up in other papers. For example, amongst others, work by Kitchin on Data Dashboards in Dublin's Smart City; Amoore on data analytics in border control; Bates' on the circulation of meteorological and climate data; Kennedy on public perceptions of organisations data practices and governance. These empirical contributions have produced detailed and critical accounts of, for example, the ways in which complex social, cultural, political and economic factors interplay to provide a framework for how big data systems are envisioned and developed, and what the implications of this may be for societies. Much of this work evidences significant concerns and raises many questions about the unjust implications of emergent data practices for already marginalised social groups (e.g. Amore, 2006), or illuminates how patterns of big data sharing and trading might be empowering already powerful agents such as financial companies in relation to societal challenges such as climate change (e.g. Bates et al, 2016).

Emergence of new research and activist communities

A final contribution of the field that can be identified relates to community building. Over the last few years there has been a flourishing of new spaces for people from different backgrounds to share and engage in dialogue about research and ideas related to big data analytics and society. Some of these have emerged from within academia, for example, the research journal Big Data and Society that was launched in 2013, and international conferences such as Data Power and Data Justice (in the Social Sciences and Humanities), and ACM Conference on Fairness, Accountability and Transparency (in the Computer Sciences). Other groups have emerged from activist communities in which academics may or may not be involved. Examples, include Data for Black Lives and the Algorithmic Justice League in the USA, and Indigenous Data Sovereignty Networks in countries including the USA, Canada, Australia and New Zealand.

Key critiques of research in the field: politics, bias and empirical grounding

Critiques of research in this field tends to revolve around disagreements related to the differing political commitments and positionalities of researchers. For example, critiques are advanced that some research is not critical or normative enough, or that it emphasises technical or legal solutions that do not recognise or make explicit the ways in which issues identified reflect longstanding social, political, economic and cultural inequalities (Dencik et al, 2019). These critiques have been most obvious along disciplinary lines. For example, critical social scientists have argued that the techno-legal orientation of research communities such as 'FATE' are limited to the extent that they do not fully grapple with the complex social realities of big data, but instead emphasise the development of technical methods to identify, measure and correct problems within big data systems.

Another critique has been that much work in the field has been strongly western in orientation, with many of the key contributors being based in North American and European universities and drawing upon Western perspectives and theories in their work (Milan and Treré, 2019; Halkort, 2019). Some conferences have tried to address some of these concerns in an effort to decolonise the field, for example in selection of key note speakers, conference themes and financial support for attendees from the 'global south', however there is still much work to be done to broaden and enrich the field beyond predominantly Western perspectives.

A final critique relates to the lack of empirical grounding in some research. For example, some theoretical or speculative work may make assumptions or exaggerated claims about the power of big data analytics, that do not ring true for data scientists and practitioners that have more understanding of the technical and organisational limitations on some of the speculations that have been advanced by some scholars and journalists. These tensions between researchers with different disciplinary backgrounds and underlying philosophical commitments in the way they approach research can often work as a barrier to advancing interdisciplinary research that aims to better understand big data analytics as a socio-technical phenomenon.

Conclusion: the importance of Critical Data Studies and anticipated future developments

Big data and related analytics technologies are likely to become more deeply embedded into different organisational and social contexts in the years to come. While some legal frameworks such as the EU's GDPR aim to curb some of the most significant risks to citizen's privacy and freedom, these laws have their limitations and often are a few steps behind the most recent technological developments. Academic fields such as Critical Data Studies and 'FATE' (Fairness, Accountability, Transparency and Ethics) and related activist communities are therefore crucial to help society keep abreast of big data developments, offer critique of practices that risk unjust social consequences, and develop recommendations for how to resolve issues identified.

With some of the conceptual and normative groundwork now laid for the field of Critical Data Studies, at least in the Western context, we might expect more in depth empirical work to be added to that already produced by scholars in the field in the coming years, and some of this work will add its own conceptual and theoretical insights. Given the Western bias of the existing conceptual toolbox means it is "only partially able to grasp the obscure development, the cultural richness, and the vibrant creativity emerging at the margins of the 'empire'" (Milan and Treré, 2019, p. 321), we might also expect to see a flourishing of concepts, theories and empirical work that advances efforts to decolonise the field. Further, it is anticipated that we will see more interdisciplinary collaboration between those in the social and technical disciplines. Such collaboration would allow for both more forensic unpacking of the work of big data systems and those that develop them, as well as critical design research to advance the development of big data systems that are designed around values of social justice, rather than profit and social control.

Bibliography

- Amoore, L. (2006). Biometric borders: Governing mobilities in the war on terror, Political Geography, 25,336-351
- Bates, J., Lin, Y. W. and Goodale, P. (2016) 'Data journeys: Capturing the socio-material constitution of data objects and flows', Big Data and Society, 3(2), pp. 1–12. doi: 10.1177/2053951716654502.
- Beer, D. (2016) 'How should we do the history of Big Data?', Big Data & Society, 3(1), pp. 181–196. doi: 10.1177/2053951716646135.
- Boyd, D. and Crawford, K. (2012) 'Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon', Information Communication and Society, 15(5), pp. 662–679. doi: 10.1080/1369118X.2012.678878.

- Coté, M., Gerbaudo, P. and Pybus, J. (2016) 'Introduction. Politics of Big Data', Digital Culture & Society, 2(2), pp. 5–16. doi: 10.14361/dcs-2016-0202.
- Couldry, N. and Mejias, U. A. (2019) 'Data Colonialism: Rethinking Big Data's Relation to the Contemporary Subject', Television and New Media. doi: 10.1177/1527476418796632.
- Couldry, N. and Powell, A. (2014) 'Big Data from the bottom up', Big Data & Society, 1(2), p. 205395171453927. doi: 10.1177/2053951714539277.
- D'Ignazio, C and Klein, L. (2020) Data Feminism. MIT Press: Cambridge, Mass.
- Dalton, J. and Thatcher, J. (2014). What does a critical data studies look like, and why do we care? Seven points for a critical approach to 'big data ', Society and Space. Available at: http://societyandspace.com/material/commentaries/craig-dalton-and-jim-thatcher-what-does-a-critical-data-studies-look-like-and-why-do-we-care-seven-points-for-a-critical-approach-to-big-data/ (Accessed: 23 December 2014).
- Dencik, L. et al. (2019) 'Exploring Data Justice: Conceptions, Applications and Directions', Information, Communication & Society. Routledge, 22(7), pp. 873–881. doi: 10.1080/1369118X.2019.1606268.
- Galic et al (2015), Bentham, Deleuze and Beyond: an overview of surveillance theories from the panopticon to participation. Philosophy and Technology, 30, 9-37.
- Gitelman, L. (2013) 'Raw data' is an oxymoron. Cambridge, Mass.: MIT Press. Available at: https://mitpress.mit.edu/books/raw-data-oxymoron (Accessed: 17 August 2020).
- Halkort, M., 2019. Decolonizing Data Relations: On the Moral Economy of Data Sharing in Palestinian Refugee Camps. Canadian Journal of Communication, 44(3).
- Iliadis, A. and Russo, F. (2016) 'Critical data studies : An introduction', Big Data & Society, (December), pp. 1–7. doi: 10.1177/2053951716674238.
- Jarke, J., 2019. Open government for all? Co-creating digital public services for older adults through data walks. Online Information Review.
- Kennedy, H., Poell, T. and van Dijck, J. (2015) 'Data and agency', Big Data and Society, 2(2), pp. 1–7. doi: 10.1177/2053951715621569.
- Kitchin, R. (2013) 'The real-time city? Big data and smart urbanism', GeoJournal, 79(1), pp. 1–14. doi: 10.1007/s10708-013-9516-8.
- Kitchin, R. (2014). *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. Sage: London
- Kitchin, R. and Lauriault, T. P. (2014) 'Towards critical data studies : Charting and unpacking data assemblages and their work (preprint)', Geoweb and Big Data, pp. 3–20. Available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2474112.
- Kitchin, R. and McArdle, G. (2016) 'What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets', Big Data & Society, 3(1).
- Lupton, D. (2018) 'How do data come to matter? Living and becoming with personal data', Big Data & Society. 5(2) doi: 10.1177/2053951718786314.
- Lupton, D. (2019). Data Letters and Data Kondo. Available at: https://livingwithpersonal.data.blog/2019/11/20/data-letters-and-data-kondo/
- Lupton, D. (2020). Innovative and creative methods for researching people's use and understandings of their data – a resource list. Available at: https://livingwithpersonal.data.blog/2020/03/08/innovative-and-creative-methods-for-researching-peoples-use-and-understandings-of-their-data-a-resource-list/
- Manovich, L. (2011) 'Trending: The Promises and the Challenges of Big Social Data', Debates in the Digital Humanities, pp. 460–475. doi: 10.5749/minnesota/9780816677948.003.0047.
- Milan, S. and Treré, E. (2019) 'Big Data from the South(s): Beyond Data Universalism', Television and New Media. SAGE Publications Inc., 20(4), pp. 319–335. doi: 10.1177/1527476419837739
- Noble, S. U. (2018) *Algorithms of oppression : how search engines reinforce racism*.

- Otterbacher, J., Bates, J. and Clough, P. (2017) 'Competent men and warm women: Gender stereotypes and backlash in image search results', *Conference on Human Factors in Computing Systems - Proceedings*, 2017-May, pp. 6620–6631. doi: 10.1145/3025453.3025727.
- Pasquale, F. (2015) The black box society: the secret algorithms that control money and information. Harvard University Press: Cambridge, Mass.
- Powell, Alison. 2018. "Alison Powell on Data Walking". TMG Journal for Media History 21 (2): 146–50. DOI: http://doi.org/10.18146/2213-7653.2018.371
- Rességuier, A. and Rodrigues, R. (2020) 'AI ethics should not remain toothless! A call to bring back the teeth of ethics', Big Data & Society. SAGE Publications Ltd, 7(2), p. 205395172094254. doi: 10.1177/2053951720942541
- Ruppert, E. (2013) 'Rethinking empirical social sciences', Dialogues in Human Geography, 3(3), pp. 268–273. doi: 10.1177/2043820613514321.
- Ruppert, E., Isin, E. and Bigo, D. (2017) 'Data politics', Big Data & Society, 4(2), p. 205395171771774. doi: 10.1177/2053951717717749.
- Savage, M. and Burrows, R. (2007) 'The Coming Crisis of Empirical Sociology', 41(5), pp. 885–899. doi: 10.1177/0038038507080443.
- Taylor, L. *et al.* (2014) 'Big Data and Positive Change in the Developing World', *Policy & Internet*, 6(4), pp. 418–444. doi: 10.1002/1944-2866.POI378.
- Thatcher, J., O'Sullivan, D. and Mahmoudi, D. (2016) 'Data colonialism through accumulation by dispossession: New metaphors for daily data', Environment and Planning D: Society and Space, 34(6), pp. 990–1006. doi: 10.1177/0263775816633195
- van Dijck, J. (2014) 'Datafication, dataism and dataveillance: Big data between scientific paradigm and ideology', Surveillance and Society, 12(2), pp. 197–208. doi: 10.24908/ss.v12i2.4776.