CHAPTER 1

Introduction

In times of big data and datafication, we should refrain from using the term 'sharing' too lightly. While users want, or need, to communicate online with their family, friends or colleagues, they may not intend their data to be collected, documented, processed and interpreted, let alone traded. Nevertheless, retrieving and interrelating a wide range of digital data points, from, for instance. social networking sites, has become a common strategy for making assumptions about users' behaviour and interests. Multinational technology and internet corporations are at the forefront of these datafication processes. They control, to a large extent, what data are collected about users who embed various digital, commercial platforms into their daily lives.

Tech and internet corporations determine who receives access to the vast digital data sets generated on their platforms, commonly called 'big data'. They define how these data are fed back into algorithms crucial to the content that users subsequently get to see online. Such content ranges from advertising to information posted by peers. This corporate control over data has given rise to considerable business euphoria. At the same time, the power exercised with data has increasingly been the subject of bewilderment, controversies, concern and activism during recent years. It has been questioned at whose cost the Silicon Valley mantra 'Data is the new oil'[1] is being put into practice. It is questioned whether this view on data is indeed such an alluring prospect for societies relying increasingly on digital technology, and for individuals exposed to datafication.

Datafication refers to the quantification of social interactions and their transformation into digital data. It has advanced to an ideologically infused '[…] leading principle, not just amongst technoadepts, but also amongst scholars who see datafication as a revolutionary research opportunity to investigate human conduct' (van Dijk 2014, 198). Datafication points to the widespread ideology of big data's desirability and unquestioned superiority, a tendency termed 'dataism'

How to cite this book chapter:
Richterich, A. 2018. *The Big Data Agenda: Data Ethics and Critical Data Studies*. Pp. 1–14. London: University of Westminster Press. DOI: https://doi.org/10.16997/book14.a. License: CC-BY-NC-ND 4.0

by van Dijk (2014). This book starts from the observation that datafication has left its mark not only on corporate practices, but also on approaches to scientific research. I argue that, as commercial data collection and research become increasingly entangled, interdependencies are emerging which have a bearing on the norms and values relevant to scientific knowledge production.

Big data have not only triggered the emergence of new research approaches and practices, but have also nudged normative changes and sparked controversies regarding how research is ethically justified and conceptualised. Big data and datafication 'drive' research ethics in multiple ways. Those who deem the use of big data morally reasonable have normatively framed and justified their approaches. Those who perceive the use of big data in research as irreconcilable with ethical principles have disputed emerging approaches on normative grounds. What we are currently witnessing is a coexistence of research involving big data and contested data ethics relevant to this field. I explore to what extent these positions unfold in dialogue with (or in isolation from) each other and relevant stakeholders.

This book interrogates entanglements between corporate big data practices, research approaches and ethics: a domain which is symptomatic of broader challenges related to data, power and (in-)justice. These challenges, and the urgent need to reflect on, rethink and recapture the power related to vast and continually growing 'big data' sets have been forcefully stressed in the field of critical data studies (Iliadis and Russo 2016; Dalton, Taylor and Thatcher 2016; Lupton 2015; Kitchin and Lauriault 2014; Dalton and Thatcher 2014). Approaches in this interdisciplinary research field examine practices of digital data collection, utilisation, and meaning-making in corporate, governmental, institutional, academic, and civic contexts.

Research in critical data studies (CDS) deals with the societal embeddedness and constructedness of data. It examines significant economic, political, ethical, and legal issues, as well as matters of social justice concerning data (Taylor 2017; Dencik, Hintz and Cable 2016). While most companies have come to see, use and promote data as a major economic asset, allegedly comparable to oil, CDS emphasises that data are not a mere commodity (see also Thorp 2012). Instead, many types of digital data are matters of civic rights, personal autonomy and dignity. These data may emerge, for example, from individuals' use of social networking sites, their search engine queries or interaction with computational devices. CDS researchers analyse and examine the implications, biases, risks and inequalities, as well as the counter-potential, of such (big) data. In this context, the need for qualitative, empirical approaches to data subjects' daily lives and data practices (Lupton 2016; Metcalf and Crawford 2016) has been increasingly stressed. Such critical work is evolving in parallel with the spreading ideology of datafication's unquestioned superiority: a tendency which is also noticeable in scientific research.

Many scientists have been intrigued by the methodological opportunities opened up by big data (Paul and Dredze 2017; Young, Yu and Wang 2017; Paul et al. 2016; Ireland et al. 2015; Kramer, Guillory and Hancock 2014; Chunara et al. 2013; see also Chapter 5). They have articulated high hopes about the contributions big data could make to scientific endeavours and policy making (Kettl 2017; Salganik 2017; Mayer-Schönberger and Cukier 2013). As I show in this book, data produced and stored in corporate contexts increasingly play a part in scientific research, conducted also by scholars employed at or affiliated with universities. Such data were originally collected and enabled by internet and tech companies owning social networking sites, microblogging services and search engines.

I focus on developments in public health research and surveillance, with specific regard to the ethics of using big data in these fields. This domain has been chosen because data used in this context are highly sensitive. They allow, for example, for insights into individuals' state of health, as well as health-relevant (risk) behaviour. In big data-driven research, the data often stem from commercial platforms, raising ethical questions concerning users' awareness, informed consent, privacy and autonomy (see also Parry and Greenhough 2018, 107–154). At the same time, research in this field has mobilised the argument that big data will make an important contribution to the common good by ultimately improving public health. This is a particularly relevant research field from a CDS perspective, as it is an arena of promises, contradictions and contestation. It facilitates insights into how technological and methodological developments are deeply embedded in and shaped by normative moral discourses.

This study follows up earlier critical work which emphasises that academic research and corporate data sources, as well as tools, are increasingly intertwined (see e.g. Sharon 2016; Harris, Kelly and Wyatt 2016; Van Dijck 2014). As Van Dijck observes, the commercial utilisation of big data has been accompanied by a '[…] gradual normalization of datafication as a new paradigm in science and society' (2014, 198). The author argues that, since researchers have a significant impact on the establishment of social trust (206), academic utilisations of big data also give credibility to their collection in commercial contexts the societal acceptance of big data practices more generally.

This book specifically sheds light on how big data-driven *public health research* has been communicated, justified and institutionally embedded. I examine interdependencies between such research and the data, infrastructures and analytics shaped by multinational internet/tech corporations. The following questions, whose theoretical foundation is detailed in Chapter 2, are crucial for this endeavour: What are the broader *discursive conditions* for big data-driven health research: Who is affected and involved, and how are certain views fostered or discouraged? Which *ethical arguments* have been discussed: How is big data research ethically presented, for example as a relevant, morally right, and societally valuable way to gain scientific insights into public health? What *normativities* are at play in presenting and (potentially) debating big data-driven research on public health surveillance?

I thus emphasise two analytical angles: first, the discursive conditions and power relations influencing and emerging in interaction with big data research; second, the values and moral arguments which have been raised (e.g. in papers, projects descriptions and debates) as well as implicitly articulated in research practices. I highlight that big data research is inherently a ground of normative framing and debate, although this is rarely foregrounded in big data-driven health studies. To investigate the abovementioned issues, I draw on a pragmatist approach to ethics (Keulartz et al. 2004). Special emphasis is placed on Jürgen Habermas' notion of 'discourse ethics' (2001 [1993], 1990). This theory was in turn inspired by Karl-Otto Apel (1984) and American pragmatism. It will be introduced in more detail in Chapter 2.

Already at this point it is important to stress that the term 'ethical' in this context serves as a qualifier for the *kind* of debate at hand – and not as a normative assessment of content. Within a pragmatist framework, something is ethical because values and morals are being negotiated. this means that 'unethical' is not used to disqualify an argument normatively. Instead, it would merely indicate a certain quality of the debate, i.e. that it is not dedicated to norms, values, or moral matters. A moral or immoral decision would be in either case an ethical issue, and '[w]e perform ethics when we put up moral routines for discussion' (Swierstra and Rip 2007, 6).

To further elaborate the perspective taken in this book, the following sections expand on key terms relevant to my analysis: *big data* and *critical data studies*. Subsequently, I sketch main objectives of this book and provide an overview of its six chapters.

Big Data: Notorious but Thriving

In 2018, the benefits and pitfalls of digital data analytics were still largely attributed to a concept which had already become somewhat notorious by then: big data. This vague umbrella term refers to the vast amounts of digital data which are being produced in technologically and algorithmically mediated practices. Such data can be retrieved from various digital-material social activities, ranging from social media use to participation in genomics projects.[2]

Data and their analysis have of course long been a core concern for quantitative social sciences, the natural sciences, and computer science, to name just a few examples. Traditionally though, data have been scarce and their compilation was subject to controlled collection and deliberate analytical processes (Kitchin 2014a; boyd 2010). In contrast, the '[…] challenge of analysing big data is coping with abundance, exhaustivity and variety, timeliness and dynamism, messiness and uncertainty, high relationality, and the fact that much of what is generated has no specific question in mind or is a by-product of another activity.' (Kitchin 2014a, 2)

Already in 2015, The Gartner Group ceased issuing a big data hype cycle and dropped 'big data' from the Emerging technologies hype cycle. A Gartner analyst justified this decision, not on the grounds of the term's irrelevance, but because of big data's ubiquitous pervasion of diverse domains: it '[…] has become prevalent in our lives across many hype cycles.' (Burton 2015) One might say that the '[b]ig data *hype* [emphasis added] is officially dead', but only because '[…] big data is now the new normal' (Douglas 2016). While one may argue that the concept has lost its 'news value' and some of its traction (e.g. for attracting funding and attention more generally), it is still widely used, not least in the field relevant to his book. For these reasons, I likewise still use the term 'big data' when examining developments and cases in public health surveillance. Despite the fact that the hype around big data seems to have passed its peak, much confusion remains about what this term actually means.

In the wake of the big data hype, the interdisciplinary field of *data science* (Mattmann 2013; Cleveland 2001) received particular attention. Already in the 1960s, Peter Naur – himself a computer scientist – suggested the terms 'data science' and 'datalogy' as preferable alternatives to 'computer science' (Naur 1966; see also Sveinsdottir and Frøkjær 1988). While the term 'datology' has not been taken up in international (research) contexts, 'data science' has shown that it has more appeal: As early as 2012, Davenport and Patil even went as far as to call data scientist 'the Sexiest Job of the 21st Century'. Their proposition is indicative of a wider scholarly and societal fascination with new forms of data, ways of retrieval and analytics, thanks to ubiquitous digital technology.

More recently, data science has often been defined in close relation to corporate uses of (big) data. Authors such as Provost and Fawcett state, for instance, that defining '[…] the boundaries of data science precisely is not of the utmost importance' (2013, 51). According to the authors, while this may be of interest in an academic setting, it is more relevant to identify common principles '[…] in order for data science to serve business effectively' (51). In such contexts, big data are indeed predominantly seen as valuable commercial resources, and data science as key to their effective utilisation. The possibilities, hopes, and bold promises put forward for big data have also fostered the interest of political actors, encouraging policymakers such as Neelie Kroes, European Commissioner for the Digital Agenda from 2010 until 2014, to reiterate in one of her speeches on open data: 'That's why I say that data is the new oil for the digital age.' (Kroes 2012)

There are various ways and various reasons to collect big data in corporate contexts: social networking sites such as Facebook document users' digital interactions (Geerlitz and Helmond 2013). Many instant messaging applications and email providers scan users' messages for advertising purposes or security-related keywords (Gibbs 2014; Wilhelm 2014; Godin 2013). Every query entered into the search engine Google is documented (Ippolita 2013; Richterich 2014a). And not only users' digital interactions and communication, but their physical movements and features are turned into digital data. Wearable technology tracks, archives and analyses its owners' steps and heart rate (Lupton 2014a). Enabled by delayed legal interference, companies such as 23andMe sold personal genomic kits which customers returned with saliva samples, i.e. personal, genetic data. By triggering users' interest in health information based on genetic analyses, between 2007 and 2013, the company built a corporately owned genotype database of more than 1,000,000 individuals (see Drabiak 2016; Harris, Kelly, and Wyatt 2013a; 2013b; Annas and Sherman 2014).[3]

One feature common to all of these examples is the emergence of large-scale, continuously expanding databases. Such databases allow for insights into, for example, users' (present or future) physical condition; the frequency and (linguistic) qualities of their social contacts; their search preferences and patterns; and their geographic mobility. Broadly speaking, corporate big data practices are aimed at selling or employing these data in order to provide customised user experiences, and above all to generate profit.[4]

Big data differ from traditional large-scale datasets with regards to their volume, velocity, and variety (Kitchin 2014a, 2014b; boyd and Crawford 2012; Marz and Warren 2012; Zikopoulos et al. 2012). These 'three Vs' are a commonly quoted reference point for big data. Such datasets are comparatively flexible, easily scalable, and have a strong indexical quality, i.e. are used for drawing conclusions about users' (inter-)actions. While volume, velocity, and variety are often used to define big data, critical data scholars such as Deborah Lupton have highlighted that '[t]hese characterisations principally come from the worlds of data science and data analytics. From the perspective of critical data researchers, there are different ways in which big data can be described and conceptualised' (2015, 1). Nevertheless, brief summaries of the 'three Vs' will be provided, since this allows me to place them in relation to the perspectives of critical data studies.

Volume, the immense scope of digital datasets, may appear to be the most evident criterion. Yet, it is often not clear what actual quantities of historic, contemporary, and future big data are implied.[5] For example, in 2014, the corporate service provider and consultancy International Data Corporation predicted that until 2020 'the digital universe will grow by a factor of 10 – from 4.4 trillion gigabytes to 44 trillion. It more than doubles every two years' (EMC, 2014). How these estimations are generated is, however, often not disclosed. When the work on this chapter was started in January 2016, websites such as *internet live stats* claimed that 'Google now processes over 40,000 search queries every second on average (visualize them here), which translates to over 3.5 billion searches per day and 1.2 trillion searches per year worldwide' (Google Search Statistics, 2016). In order to calculate this estimation, the site draws on several sources, such as official Google statements, Gigaom publications and independent search engine consultancies, which are then fed into a proprietary algorithm (licensed by *Worldometers*). Externally, one cannot assess for certain how these numbers have been calculated in detail, and to what extent the provided information, estimations and predictions may be reliable. Nevertheless, the sheer quantity of this new form of data contributes to substantiating related claims regarding its relevance and authority.

As boyd and Crawford argue, the big data phenomenon rests upon the long-standing myth '[…] that large data sets offer a higher form of intelligence and knowledge that can generate insights that were previously impossible, with the aura of truth, objectivity, and accuracy' (2012, 663). This has fostered the emergence of a 'digital positivism' (Mosco 2015) promoting the epistemological assumption that we can technologically control big data's collection and analysis, to the extent that these data may 'speak for themselves' and become inherently meaningful.

This is especially relevant, since these large quantities of data and their interpretation are closely related to promises about profits, efficiency and bright future prospects.[6] Big data – as wider phenomena, and with regards to respective cases – are staged in certain ways. The possibilities and promises associated with the term are used to signify its relevance for businesses (see e.g. Marr 2015; Pries and Dunnigan 2015; Simon 2013; Ohlhorst 2012) and governmental institutions (Kim, Trimi, and Chung 2014; Bertot et al. 2014), and their need to take urgent action. However, despite such claims for its relevance, the collection and analysis of big data is often opaque. This performative aspect of big data, combined with the common blackboxing of data collection, quantitative methods and analysis, is also related to the frequently raised accusation that the term is to a large extent hyped (Gandomi and Haider 2015; Uprichard 2013; Fox and Do 2013).

Apart from the recurring issue that most big data practices take place behind closed curtains and that results are difficult to verify (Driscoll and Walker 2014; Lazer et al. 2014), the problem of assessing actual quantities is also closely related to big data's *velocity*. Their continuous, often real-time production creates an ongoing stream of additional input. Not only does the amount of data produced by existing sources grow continuously, but as new technologies enter the field, new types of data are also created. Moreover, changes in users' behaviour may alter data not only in terms of their quantity, but also their quality and meaningfulness.

Regarding the *variety* or qualitative aspects of big data, they consist in a combination of structured, unstructured and semi-structured data. While structured data (such as demographic information or usage frequencies) can be easily standardised and, for example, numerically or alphabetically defined according to a respective data model, unstructured and semi-structured data are more difficult to classify. Unstructured data refer to visual material such as photos or videos, as well as to text documents which are/were too complex to systematically translate into structured data. Semi-structured data refer to those types of material which combine visual or textual material with metadata that serve as annotated, structured classifiers of the unstructured content.

The possibilities and promises associated with big data have been greeted with notable enthusiasm: as indicated before, this does not only apply to corporations and their financial interests, but has also been noticeable in scientific research (Tonidandel, King, and Cortina 2016; Mayer-Schönberger and Cukier 2013; Hay et al. 2013). This enthusiasm is often grounded in the assumption that data can be useful and beneficial, if we only learn how to collect, store and analyse them appropriately (Finlay 2014; Franks 2012). Related literature mainly addresses big data as practical, methodological and technological challenge, seeing them as assets to research, rather than as a societal challenge. The main concern and aim of this literature is an effective analysis of such data (see e.g. Assunção et al. 2015; Jagadish et al. 2014). Such positions have, however, been called into question and critically extended by authors engaged in critical data studies.

Critical Data Studies

Current corporate or governmental big data practices, and academic research involving such data, are predominantly guided by deliberations regarding their practicability, efficiency and optimisation. In contrast, approaches in critical data studies are not primarily concerned with practical issues of data usability, but scrutinise the conditions for contemporary big data collection, analysis and utilisation. They challenge big data's asserted 'digital positivism' (Mosco 2015), i.e. the assumption that data may 'speak for themselves'.

Critical data studies form an emerging, interdisciplinary field of scholars reflecting on how corporations, institutions and individuals collect and use 'big' data – and what alternatives to existing approaches could look like. Currently, critical data studies predominantly evaluates social practices involving (big) data, rather than operationalising approaches for research using big data. It mainly encompasses research *on* big data, focused on assessments of historical or ongoing big data projects and practices (Mittelstadt and Floridi 2015; Lupton 2013; boyd and Crawford 2012). Such an approach is also taken in this book.

In addition, some researchers have critically engaged and experimented with research *with* big data. For example, this has been done by using data processing software like *Gephi* in order to show how algorithms and visualisation may influence research results. Importantly, research groups such as the *Digital Methods Initiative* explore the possibilities and boundaries of applying and developing quantitative digital tools and methodologies.[7] However, at present, *critical data studies* predominantly refers to the critique of recent big data approaches. As Mosco points out: 'The technical criticisms directed at big data's singular reliance on quantification and correlation, and its neglect of theory, history, and context, can help to improve the approach, and perhaps research in general – certainly more than the all-too-common attempts to fetishize big data.' (Mosco 2015, 205–206) Therefore, in order to rethink how big data are being used (especially in research), it is also desirable that future approaches are informed by critical data studies perspectives, rather than being analysed subsequently.[8]

Also, without using the umbrella term 'critical data studies', various authors have of course nevertheless critically evaluated the collection and analysis of digital user data. These perspectives emerged in parallel with technological developments that allowed for new forms of data collection and analysis. Critical positions also surfaced with regards to the use of big data in research. In 2007, the authors of a *Nature* editorial emphasised the importance of trust in research on electronic interactions, and voiced concern about the lack of legal regulations and ethical guidelines:

> 'For a certain sort of social scientist, the traffic patterns of millions of e-mails look like manna from heaven. [. . .] Any data on human subjects inevitably raise privacy issues (see page 644), and the real risks of abuse of such data are difficult to quantify. [...] Rules are needed to ensure data can be safely and routinely shared among scientists, thus avoiding a Wild West where researchers compete for key data sets no matter what the terms.' (Nature Editorial 2007)

This excerpt refers to familiar scientific tensions and issues that were early on flagged with regards to big data research.[9] Scholars are confronted with methodological possibilities whose risks and ethical appropriateness are not yet clear.

This uncertainty may, however, be 'overpowered' by the fact that these data allow for new research methods and insights, and are advantageous for researchers willing to take the risk. While certain data may be technically accessible, it remains questionable if and how researchers can ensure, for instance, that individuals' privacy is not violated when analysing new forms of digital data. *If* scientists can gain access to certain big data, this does not ensure that using them will be ethically unproblematic. More importantly, the 'if' in this sentence hints at a major constraint of big data research: a majority of such data can only be accessed by technology corporations and their commercial, academic or governmental partners. This issue has been by Andrejevic (2014) the 'big data divide', and has also been addressed by boyd and Crawford, who introduced the categories of 'data rich' and 'data poor' actors (2014, 672ff.; see also Manovich 2011, 5).

Today, globally operating internet and tech companies decide which societal actors may have access to data generated via their respective platforms, and define in what ways they are made available. Therefore, in many cases, scholars cannot even be sure that they have sufficient knowledge about the data collection methods to assess their ethical (in-)appropriateness. This does not merely mean that independent academics cannot use these data for their own research, but it also poses the problem that even selected individuals or institutions may not be able to track, assess and/or communicate publicly how these data have been produced.

The need for critical data studies was initially articulated by critical geography researchers (Dalton and Thatcher 2014; Kitchin and Lauriault 2014) and in digital sociology, with particular regards to public health (Lupton 2014c, 2013). In geographic research this urge was influenced by developments related to the 'geospatial web'. In 2014, Kitchin and Lauriault reinforced the emergence and discussion of critical data studies, drawing on a blog post published by Dalton and Thatcher earlier that year. The authors depict this emerging field as 'research and thinking that applies critical social theory to data to explore the ways in which they are never simply neutral, objective, independent, raw representations of the world, but are situated, contingent, relational, contextual, and do active work in the world' (Kitchin and Lauriault 2014, 5). This perspective corresponds to Mosco's critique that big data 'promotes a very specific way of knowing'; it encourages a 'digital positivism or the specific belief that the data, suitably circumscribed by quantity, correlation, and algorithm, will, in fact, speak to us' (Mosco 2015, 206). It is exactly this digital positivism which is challenged and countered by contributions in critical data studies.

When looking at the roots of critical data studies in different disciplines, one is likely to start wondering which factors may have facilitated the development of this research field. In the aforementioned blog post 'What does a critical data studies look like, and why do we care?' Dalton and Thatcher stress the relevance of geography for current digital media and big data research, by emphasising that most information nowadays is geographically/spatially annotated (with reference to Hahmann and Burghardt 2013). According to the authors, many of the tools and methods used for dealing with and visualising large amounts of digital data are provided by geographers: 'Geographers are intimately involved with this recent rise of data. Most digital information now contains some spatial component and geographers are contributing tools (Haklay and Weber 2008), maps (Zook and Poorthius 2014), and methods (Tsou et al. 2014) to the rising tide of quantification.' (Dalton and Thatcher 2014)

Kitchin and Lauriault explore how critical data studies may be put into practice. They suggest that one way to pursue research in this field is to '[…] unpack the complex assemblages that produce, circulate, share/sell and utilise data in diverse ways; to chart the diverse work they do and their consequences for how the world is known, governed and lived-in' (Kitchin and Lauriault 2014, 6). Already in *The Data Revolution* (2014a), Kitchin suggested the concept of data assemblages. In this publication, he emphasises that big data are not the only crucial development in the contemporary data landscape: at the same time, initiatives such as the digital processing of more traditional datasets, data networks, and the open data movement contribute to changes in how we store, analyse, and perceive data. Taken together, various emerging initiatives, movements, infrastructures, and institutional structures constitute data assemblages that shape how data are perceived, produced and used (Kitchin 2014a, 1)

By drawing on the same idea of digital data assemblages, Lupton outlines a critical sociology of big data (2014b, 93). The author conceptualises big data as knowledge systems which are embedded in and constitute power relations. In a first step, she examines the various fields of their utilisation, such as humanitarian uses, education, policing and security. Moreover, she deconstructs the metaphors which were initially used to describe big data, and how these reflect contemporary criticism. Terms such as 'trails', 'breadcrumbs', 'exhaust', 'smoke signals', and 'shadows' (Lupton 2014b, 108) indicate that big data are commonly seen as signs with a strong indexical quality. The latter part of her analysis also provides an initial overview of themes in the field of critical data studies. However, only in a later online publication (Lupton 2015) does Lupton use the term 'critical data studies'.

A crucial metaphor that Lupton refers to here is the notion of 'raw data' (Boellstorff and Maurer 2015; Gitelman 2013; Boellstorff 2013). The rejection of an idea of data as implicitly 'natural' and 'given', i.e. 'raw', is a crucial tenet in critical data studies. Drawing on Lévi-Strauss's 'culinary triangle' of *raw-cooked-rotten* as well as Geertz' methodological approach and genre of *thick descriptions*, Boellstorff (2013) criticises the nature-culture opposition which is implied in the differentiation between 'raw' (collected) and 'cooked' (processed) data. Rather than being 'pure' expressions of human behaviour or opinions, data in all their manifestations, are always subject to interpretation and normative influences of meaning-making. To frame this fundamental condition of data-driven processes, the author suggests the notion of 'thick data': 'what makes data 'thick' is recognizing its irreducible contextuality: 'what we inscribe (or try to) is not *raw* social discourse.' […] For Geertz, 'raw' data was already oxymoronic in the early 1970s: whether cooked or rotted, data emerges from regimes of interpretation' (Boellstorff 2013).

The idea of rotten data pursues the metaphor of 'raw' and 'cooked' data, but calls attention to the changes in data and their accessibility which go beyond technically or methodologically intended control. Boellstorff (2013) argues that 'the 'rotted' 'allows for transformations outside typical constructions of the human agent as cook—the unplanned, unexpected, and accidental. Bit rot, for instance, emerges from the assemblage of storage and processing technologies as they move through time.'

In a later publication, Boellstorff and Maurer (2015) identified 'relation' and 'recognition' as particularly crucial factors influencing the constant process of data interpretation – which starts with its selection and collection. Data are created and given meaning in interactions between human and non-human actors. Their recognition is socio-culturally and politically defined (Boellstorff and Maurer 2015, 1-6; see also Lupton 2015). In this sense, the term data, derived from the Latin plural of datum, 'that is given', is already misleading, and indicates the term's socially constructed meaning. Strictly speaking, '[o]ne should never speak of 'data'- what is given – but rather of sublata, that is, of 'achievements." (Latour 1999, 42)

It is not surprising that many of the critical approaches to big data are related to fields in which potentially derived information is inherently rather sensitive: in health research and with regards to location-based technology, data critique has emerged as an important general theme. So, the need for critical data studies goes beyond such fields, and should engage with data which have been traditionally seen as sensitive, i.e. allowing for access to information which is commonly treated as private or confidential. One challenge for critical data studies has been (and will be) to demonstrate to what extent seemingly impersonal data are in fact highly sensitive, due to, for example, their corporate, regulatory or technological embedding, and new means for interrelating datasets.

Aims and Chapters

More generally, the aim of this book is to contribute to the emerging field of critical data studies. Specifically, it does so by examining the implications of big data-driven research for ethico-methodological decisions and debates. I analyse how research in public health surveillance that involves big data has been presented, discussed and institutionally embedded. In order to do so, I will examine projects employing and promoting big data for public health research and surveillance.[10] This book realises three main objectives: first, it develops and applies a critical data studies approach which is predominantly grounded in pragmatist ethics as well as Habermasian discourse ethics, and takes cues from (feminist) technoscience criticism (Chapter 2). Second, it identifies broader issues and debates concerning big data-driven biomedical research (Chapter 3). Thirdly, it uses the example of big data-driven studies in public health research and surveillance to examine more specifically the issues and values implicated in the use of big data(Chapters 4 and 5).

This book is divided into six chapters. Chapter 1 introduced the term 'big data' and provided an initial overview of critical data studies. Chapter 2 'Examining data practices and data ethics' focuses on the theoretical foundations of my analysis. The first subchapter 'What it means to "study data" expands on the brief introduction to critical data studies provided above. Adding to the basic principles and historical development outlined in Chapter 1, it offers an overview of themes and issues. The second subchapter, 'Critical perspectives' elucidates why the approach taken in this book should be considered 'critical'. While Habermas' work links this book to critical theory, I also draw on strands in science and technology studies which have explicitly addressed the possibilities and need for normative, engaged analyses; here, I refer mainly to the sociology of scientific knowledge construction, as well as feminist technoscience. The third subchapter on pragmatism and discourse ethics builds upon

Keulartz et al.'s pragmatist approach and Habermas' critical theory notion of discourse ethics.

Chapter 3 'Big data: Ethics and values' describes normative developments which have been discussed with regards to digital data practices, particularly in research. This chapter depicts tensions between values related to personal rights and those linked to the public good, such as the common opposition between privacy and security. Moreover, it shows how transparency and open data relate to (and may conflict with) individuals' privacy and corporate interests in exclusive data access. Based on an overview of the values which have been advanced to justify or critique big data research, I examine how these relate to current negotiations of research methodologies and normativities. This also involves reflections on entanglements between corporate data economies and research analytics. The main purpose of this chapter is to identify broader developments relevant to the case studies, as well as those values which have been comparatively emphasised or neglected.

Chapters 4 and 5 examine the institutional context, methodological choices and justifications of big data-driven research in public health surveillance. In Chapter 4, I show how funding schemes specifically targeted at promoting the use of big data in biomedical research incentivise methodological trends, with ethical implications. Interdependencies between researchers and grant providers need to be seen in the context of funding environments which are partly co-defined by internet/tech corporations. Shedding light on these institutional contexts also facilitates insights into factors co-constructing researchers' decisions to pursue certain topics and approaches.

Chapter 5 goes on to show how such research decisions and developments are translated into research projects. I specifically unpack how the use of big data collected by tech corporations is practically realised as well as discursively presented by researchers. I focus on research projects which have utilised sources that are not traditionally seen as 'biomedical data', but should be seen as such since they allow for insights into users' state of health and health-relevant behaviour. Analyses of specific cases and references to contemporary developments are made throughout the book, but especially in Chapter 5.

While Chapter 4 highlights the institutional conditions for public health surveillance involving digital data by depicting relevant funding schemes, Chapter 5 presents three clusters of cases: 1) Tweeting about illness and risk behaviour; 2) data retrieval through advertising relations; and 3) data mashups. The first cluster examines how Twitter data have been utilised as indicators of health risk behaviour. The second cluster explores researchers' attempts to access, for example, Facebook data via advertising and marketing services. The third cluster focuses on publicly available platforms developed by researchers which draw on data collected by tech corporations such as Google.

These case studies have been chosen because they are not merely clear-cut cases of corporate, commercial data utilisation, but involve more diverse values. More importantly, they are cases in which the analytic possibilities of big data have led to the emergence of 'technosciences', i.e. academic research fields which are substantially grounded in technological changes. It seems important to highlight here that the book's objective is not merely to expose certain projects as 'immoral' (see also Chapter 2). Instead, I want to emphasise the complexities and contradictions, the methodological as well institutional dilemmas, and factors of influence co-constructing current modes of big data research.

The final chapter ties together insights from the analysis, specifically in relation to the critical perspectives and theory introduced in Chapter 2. It emphasises two main issues: first, in the field of big data-driven public health research, one can observe complex (inter-)dependencies between academic research and the commercial interests of internet and tech corporations. This is notably related to two main developments: on the one hand, data access is often controlled by these companies; on the other hand, these companies incentivise research at the intersection of technology and health (e.g. through funding and selective data access).

Second, data practices, norms and the promises of internet/tech corporations are increasingly echoed and endorsed in big data-driven health research and its ethics. These tendencies foster research approaches that inhibit the discursive involvement of affected actors in negotiations of relevant norms. In consequence, I argue that, from a discourse ethics perspective, there is an urgent need to transition from big data-driven to data-discursive research, foregrounding ethical issues. Such research needs to encourage the involvement of potentially affected individuals, as a condition for formative discourse and research ethics grounded in valid social norms.