A Neoliberalisation of Social Data?

Big Data and the Future of Official Statistics

Ville Takala

Department of Sociology, Goldsmiths, University of London

Thesis submitted for the degree of Doctor of Philosophy (PhD), 2018

Declaration

I declare that the work presented in this thesis is my own.

Ville Takala

Acknowledgments

First, I would like to thank my supervisors Evelyn Ruppert and Roger Burrows for their continuous support throughout my doctoral studies; I could not have completed this thesis without their guidance. I would also like to extend a special thanks to Evelyn Ruppert for inviting me to participate in the European Research Council funded ARITHMUS research project (CoG 615588). My involvement in the project granted me access to my field-site and gave me the opportunity to work as part of a larger, established research team. Having the chance to share my work with the other researchers on the project ensured that my PhD experience was far less lonely than what it would have been otherwise. I therefore extend my thanks to my ARITHMUS team members: Baki Cakici, Francisca Grommé, Stephan Scheel and Funda Ustek-Spilda.

I would like to thank the staff at Goldsmiths Sociology department, especially Bridget Ward for her support. I thank Monica Greco for providing me with an opportunity to present my work at the IN PROCESS PhD seminar series, and to all other students who participated in them. Sharing one's work with others is surely the best way to garner inspiration and remain motivated throughout a long project. I therefore extend my thanks also to Katharina Hecht and Isabell Loeschner for the writing group sessions that we shared at the London School of Economics and Political Science. I always left those sessions highly motivated and with a strong sense of direction. Thanks also to my upgrade examiners, Aidan Kelly and Dhiraj Murthy, who provided me with important feedback at a critical stage in the thesis. I am enormously grateful to my interviewees at Statistics Finland for taking the time to talk to me despite their busy schedules.

I am also grateful to my close friends and family. I thank Rachael for her friendship and support. And I extend my deepest gratitude to my family,

who have always encouraged me to pursue my dreams.

Finally, I thank the Economic and Social Research Council (ref 1256516) for the financial support without which this thesis would not have been possible.

Abstract

Many of the existing commentaries on enormous, interconnected, dynamic datasets, or Big Data, as they have become commonly known, have highlighted their technical qualities. It has for example been argued that what separates Big Data from previous forms of data are the so called 3Vs; Volume, Variety and Velocity. In contrast, based on a historical, conceptual and empirical analysis, I suggest that what is novel about Big Data is not just its technical composition, but more importantly the changing jurisdictions between the producers of data that its emergence entails. Its technical composition, I argue, is predicated upon how its production is organised.

I suggest that historically, statistics have emerged out of a co-constitutive interaction between methodological and technological developments and changes in the political and administrative world (Desrosières 1998). Further to this, I suggest that the production of data on which statistics have relied has often been monopolised by dominant institutions. By situating Big Data in the context of political economy, I argue that its emergence reflects broader processes of neoliberalisation that have swept over western polities in the past few decades, notably in the sense that it primarily accumulates in the private, rather than the public sector.

By exploring responses to Big Data within National Statistical Institutes, I suggest that it signals not just an increasing privatisation of data production and related infrastructures, but also an increasing pressure to adopt neoliberal rationalities and values in the public sector. I suggest that at Statistics Finland, where my fieldwork was based, these processes are potentially undermining the social welfarist principles upon which the production of official statistics has for long been based. I argue that the social organisation of Big Data must be rethought based on Social Democratic principles and political imaginations and that the

question about the future role of the NSI must form a central component of such considerations.

List of figures

Tables

Table of contents

"Knowledge is power; statistics is democracy" - Olavi Niitamo, former Director General of Statistics Finland

1 Introduction

Empirical sociology has often been placed somewhere between science and literature (Lepenies, 1988) and sociologists have throughout the history of their discipline had to ask themselves whether their accounts of social change have been able to compete with those given usually in a more timely manner by fiction writers (Halsey, 2004). I therefore begin with an anecdote from a recent science fiction novel. When one of the main protagonists in Dave Eggar's novel *The Circle* (2014) suggests that the government might build an online voting system by itself, the other employees at a fictional internet conglomerate, whose social media platform is by this point used by 83 % of the American electorate, burst out in laughter. In the dystopian near future depicted in the novel, therefore, processes of privatisation and neoliberalisation (an ideology that posits that all aspects of society should be organised according to market principles) have gone so far that people simply cannot any longer imagine a role for the state in constructing and maintaining the infrastructure upon which a highly technical society functions.

In this thesis, I explore how the emergence of enormous, interconnected, dynamic datasets – or Big Data, as they have become commonly known, is being interpreted by governments, or more specifically, their statistical agencies. In Official Statistics, Big Data[1] is typically defined by referring

[1] Rather than settling on a technical definition, I understand Big Data as an emerging field of practices "that is not defined by but generative of (sometimes) novel data qualities such as high volume and granularity and complex analytics such as data linking and mining" (Big Data & Society, 2018). I use capitalisation throughout to denote that rather than simply a descriptor of types of data, Big Data is as a concept whose meaning is unsettled and a matter of debate and struggle across numerous fields and settings, of which official statistics is one. I will expand on this theme in chapter four.

to the 3Vs of Volume, Variety and Velocity, which are seen to characterise new data sources such as mobile positioning data, customer club card data, traffic sensor data, electricity meter data and social media data (See for example, UNECE, 2013, 2014b, 2016a). For almost four centuries, states maintained an effective monopoly on data concerning their whole populations, economies and territories. While private corporations also generated data about individuals especially over the last century or so through opinion polling and marketing research and the production of vast archives of customer data, this was limited to more narrow concerns and specific population groups. However, National Statistical Institutes (NSIs) are now rethinking how they produce data and statistics as a result of the production and accumulation of Big Data by private corporations, especially major technology providers that potentially have command over more granular, immediate, varied and detailed data about populations beyond that of any state. What this means for how NSIs conceive of their future role and the very conditions of the production of knowledge of societies is a key question I explore.

In this study, I move beyond proclamations about the revolutionary potentials of Big Data (Mayer-Schönberger and Cukier, 2013) to first undertake a historical review and then empirical exploration of how Big Data is changing and challenging conceptions of data production within a specific domain, that of official statistics. My empirical work focuses on the National Statistical Institute of Finland, a country long considered a stronghold of the Nordic Welfare State model. Furthermore, I situate developments at the Finnish statistical institute in relation to conceptions of Big Data and its implications for the future of official statistics within a transnational field of statistics of which it is a part.

From a rapidly growing corpus of social scientific reflections on Big Data, I highlight two key sources of inspiration. Back in 2007, Savage and Burrows (2007) argued polemically that because sociologists had not

kept up to date with methodological advances in the analysis of Big Data (at this point referred to as “digital transactional data”), they had begun to lose their jurisdiction over social questions. What the authors of the polemic furthermore noted was that increasingly, it was the private sector that produced and had access to Big Data and the skills and expertise required for its analysis. This point is for me an important insight, one that I see as being part of a larger story. Reflecting on developments that were taking place in the US, Starr and Corson (1987) identified already in the 1980s the increasing role of the private sector in selling repackaged public data, privately collected data, and statistical models and analyses of different types. Their analysis of the rise of a “statistical service industry” came with a warning: an increasing production of data in the private sector might in the future jeopardise its free flow, the building block of democratic politics and scientific life.

Although data has undoubtedly been an important resource for capitalist enterprises for centuries (Porter 1995), many have recently argued that as a result of digitisation it has come to occupy a far more critical role than before. Fundamentally, digitisation has opened up massive new expanses of potential data, and new industries and business models have emerged to extract profits from them, most notably in the form of dominant digital platforms (Srnicek, 2017). Whereas the private statistical service industry that Starr and Corson were writing about still relied, for the most part, on the government for the data which it then repackaged into commercial products of various sorts, what has taken place in the years since is that the production, harvesting and analysis of data by private corporations has increased to a point where it has potentially come to outstrip that of states. In this study, I forge an understanding of the circumstances underlying this development and consider its broader social and political implications. Following from this, I argue that what is novel about Big Data is not only its technical composition (Kitchin, 2014a; Kitchin and McArdle, 2016) but perhaps even more importantly the changing jurisdictions between the producers

of data that its emergence entails.

A second key source of inspiration that underpins my work is Rob Kitchin's book *The Data Revolution* (2014b) which, although not an empirical study, is nevertheless tremendously helpful in that it outlines a proposal for how to study Big Data sociologically. Kitchin (2014b: 192) ends his book with the following call:

> *For too long data and the constitution and operation of the assemblages surrounding them have been taken-for-granted, with attention focused on information and knowledge distilled from them. It is time to rectify this neglect.*

Kitchin (2014b: 184) proposes that the way to do this is "First, through philosophical reflection and synoptic, conceptual and critical analysis" and second "through detailed empirical research concerning the genesis, constitution, functioning and evolution of data assemblages." Kitchin (2014b: 188) furthermore explains that "at present, we have little understanding of both the overall construction of data assemblages and their apparatus and individual elements", and that consequently, what we need are case studies that "trace out the sociotechnical arrangements of whole assemblages or document in detail specific aspects of an assemblage such as the sociology and political economy of a community of practice within a sector of big data". For Kitchin (2014b: 17), the production of data is a contested and negotiated process in which vested interests compete over what should be counted and how.

It is this research agenda that I take up in this thesis. Instead of conducting a taxonomy of an entire assemblage, a task that would be cumbersome even for a team of researchers, I explore both conceptually and empirically how the assemblage producing official statistics is currently being reconceived by national statisticians in response to the

increasing production of data by private corporations. By paying attention to specific practices through which Big Data is conceived and experimented with by NSIs, I move beyond statements of data revolutions to exploring how Big Data are or are not influencing the everyday work of statisticians and the reconfiguration of assemblages through which data and official statistics are produced. In the remainder of this introductory chapter, I outline the structure of the thesis and note the occasions where it further addresses some of the research themes and questions prompted by Kitchin.

A good starting point for a study on the pressing issues around data and methods that are emerging in our own time is to look at what others have written about the subject previously. Long before the debate on Big Data, social scientists have not only used statistics in their research, but have also studied them as a topic in their own right. In chapter two, “Early statistics in a transnational context”, I therefore begin the analysis by exploring the literature around the history of statistics and its sociological interpretations. What emerges from this literature review is the conceptual starting point of the thesis. An author that I draw heavily on is Desrosières (1998), who argues that historically, statistics have emerged out of a co-constitutive interaction between, on the one hand, methodological and technological developments, and changes in the political and administrative world, on the other. Law et al. (2011) provide an apt description of this process when they argue that methods have a “double social life” as they are both constituted by the social world of which they are part and active in shaping that same world. Building from the concept of co-constitution, in this introductory chapter I explore four important configurations of the co-constitution of statistics and society in the history of statistics. By demonstrating how advances in statistics have been connected to political contingencies, particularly changing governmental rationalities, historically, I set the analytical framework for the rest of the thesis.

The first configuration I examine concerns how already in early societies rulers maintained some form of statistics on their populations, usually for purposes of taxation and social control. Next, I trace the roots of statistics in their modern form to seventeenth century Germany and England, where university professors and laymen began applying various techniques for the comparison of the relative strengths and weaknesses of states. Thereafter I explore in detail the early parts of the nineteenth century, a period of statistical enthusiasm that resulted in the establishment of statistical societies and institutes across Europe. Finally, I turn my attention to the mathematisation of statistics in the early parts of the twentieth century, before concluding with an analysis of the politics of statistics in the period of welfare states that followed the Second World War. By analysing the interaction between statistics and society in previous time periods, I set the stage for a later analysis of the ways in which Big Data reflects recent trends of advanced liberalism in Western polities.

As Kitchin (2014b: 17) points out, data has both a temporality and a spatiality. What data gets produced and how varies across time and space, depending on, amongst other things, organisational jurisdictions, laws, technologies, statistical methods and techniques. Kitchin (2014b: 17) notes that despite this, to date there have not been many histories and geographies of data assemblages. Kitchin (2014b: 189) therefore suggests that one way to study Big Data is to employ a genealogical method “to trace out the formation and evolution of big data, open data, and data infrastructures more generally, and specific instantiations of them.” Kitchin (2014b: 189) points out that “rather than producing a sanitized, teleological historiography”, such an approach “illustrates how the future is built upon the past, but is not necessarily determined by it in simple cause-effect ways” and that it furthermore “identifies points of confluence when people or ideas come together and give rise to new assemblages and the complex and messy ways in which these then develop”.

Inspired by this research agenda, in chapter three “A brief history of social statistics in Finland”, I consider the historical development of statistics in relation to Finland. My main motivation for this is to situate contemporary changes in the production of official statistics in Finland, which I discuss later on in the thesis, within longer historical reconfigurations. Again, instead of trying to account for the entire assemblage, with all the actors, technologies, legislations and methods etc. involved, I narrow my focus to the main reconfigurations in data production that have occurred in the history of official statistics in Finland. First, I explore statistics in the Kingdom of Sweden, a country that Finland was part of for over half a millennium. Next, I examine statistics in the period beginning at the start of the nineteenth century in which Finland formed an autonomous Grand Duchy within the Russian empire. After that, I explore statistics in the period shortly after Finland gained its independence in 1917 before moving on to a detailed analysis of developments after the Second World War, in particular the period from the 1960s onwards in which a register-based statistical system was put in place as part of the technologies of an emerging welfare state. I then define register-based statistics as a form of welfare state governmentality (Alastalo, 2009b) and explore some of their uses in social scientific research, along with a more general reflection on methodological developments in Finland.

I conclude the chapter by drawing three major analytical conclusions from the review conducted. First, I suggest that historically, statistics have in Finland been produced from data produced by the dominant institution of its time. Whereas the earliest forms of statistics were produced from data compiled by the ruler and the church, their modern form is tied to the centralised authority of the nation state. In light of this, we should perhaps not be surprised by the increasing centrality of private corporations in generating, harvesting and analysing Big Data. Rather, this development reflects broader shifts in political economy, which is a

theme that I address in more detail in chapter four. Second, the increasing centrality of private actors in the production of data suggests a historical break in that the production of data has started to potentially move away from its historical basis in states. Finally, the extensive governmental data collection system in Finland has been made possible by a high level of trust towards the state, and it is not inconceivable that this trust could be lost in the future. These are all fundamentally important points to reflect upon in relation to my empirical material. Without an appreciation of the longer historical trajectory, my ability to interpret current developments would be limited indeed.

Whereas data production in the public sector in Finland particularly in the period of the welfare state has been underpinned by social welfarist aims and objectives, what type of a rationality is driving the increasing production of data in the private sector? Continuing with the understanding that statistics evolve in interaction with political contingencies, in chapter four, “Big Data: A harbinger of utopia, but a utopia for who?”, I take up the challenging task of situating Big Data in the context of contemporary political economy. I suggest that in midst of the recent hype around the socially transformative powers of Big Data it is worth reminding that sociologists have been debating the societal impact of information technologies at least since the 1970s. By situating Big Data in this debate, I argue that instead of a radical rupture from the past, Big Data should in fact be seen as a recent chapter in a much longer development.

This problematisation then points me towards an investigation into the political and economic circumstances in which the computerisation of society kicked into motion in the 1970s. Drawing on a number of scholars, I first argue that information technology should be seen as a key site of ideological battles in our contemporary world, and second, that the digitisation of society has in the past few decades gone hand in hand with a larger ideological shift towards more market oriented

Western societies. Building on from this idea, I then read critically both popular and academic responses to Big Data and suggest that in celebrating the opportunities afforded by it, computational social scientists in particular have tended to fail to account for the fact that much of the data that is opening up exciting new possibilities for social research is simultaneously now also absolutely central to the functioning of modern capitalism.

I therefore suggest that the practical, methodological and ethical issues related to Big Data are tightly connected to the increasingly commercialised context in which much of it is being generated. I end the chapter by taking a critical look at the Big Data industry in its current form. This investigation reveals that behind the many myths surrounding Big Data still exists the hard material reality of a capitalist mode of production ultimately not that different from its industrial predecessor. Far from alleviating the many injustices and inequalities that have come to define the modern world, I suggest that Big Data should in fact be seen as an increasingly important driving factor behind them.

In sum, by building the analysis around the conceptual starting point of the thesis, that data and society are co-constituted, I argue that Big Data reflects a shift in political rationalities towards neoliberalism in Western polities. Although certainly not all Big Data is produced in the private sphere, a significant part is and is therefore underpinned by a market and profit rationality. This central argument is reflected also in the title of the thesis. I have included the question mark to signify that Big Data can be appropriated for other purposes as well, and that in the broader sense, therefore, different Big Data futures are possible (See for example, Ruppert, 2018).

But what does the increasing production of Big Data by private corporations mean for official statistics? Like many existing commentaries on Big Data, the first three chapters are of a conceptual

rather than empirical nature. However, as Kitchin (2014b: 118) points out, what we urgently need at the moment are analyses that move beyond the conceptual to empirical explorations of the workings of Big Data within specific contexts and domains. In chapter five, “Big Data at Statistics Finland: A neoliberalisation of statistical practices?” I situate the concern with the emergence of Big Data to the empirical context of the National Statistical Institute of Finland, Statistics Finland. Drawing on expert interviews with statisticians, I focus my attention on the everyday practices where social transformations and revolutions get enacted. Furthermore, by attending to how Statistics Finland are potentially changing their data infrastructures in response to how they conceive of the “threats” or “opportunities” of Big Data, I provide early empirical evidence of what this means for relations between the public and the private sector in the production of data for official statistics. Although numerous challenges and risks of using Big Data for official statistics remain unsolved, and so far no pre-existing statistics have been replaced by it, NSIs are currently not only changing their ways of thinking but are also taking up entirely new practises in response to the perceived challenge represented it. For the time being, therefore, in official statistics the so called data revolution manifests itself most clearly in changing mentalities and mindsets.

My analysis of my interview material consists of four main parts. First, instead of assuming a predefined definition of Big Data, I start by describing how statisticians understood it and the questions they saw it raising for their professional practice. In contrast to popular debates that locate the newness of Big Data in its technical qualities, for my respondents the biggest change brought about by it was that more and more institutions now have the capacity to produce and analyse data, potentially undermining the role of NSIs as the central producers of data and in turn official statistics. Furthermore, rather than a complete break of past ways of producing and analysing data, my respondents identified many similarities between old and new forms of data, more precisely

between administrative registers and Big Data. Some drew a parallel between the establishment of the register-based statistical system in the post war years, and the period of Big Data statistics emerging at the moment, with the difference between the two being that today, data increasingly accumulates in the private-, rather than the public sector.

Then, by analysing in detail Big Data projects that were ongoing at Statistics Finland at the time of research, I suggest that they are indicative of a neoliberalisation of data infrastructures in at least two ways. First, not only does data increasingly accumulate in the private sector, but amidst public sector cuts Statistics Finland was hard pressed to find resources to tackle Big Data in terms of its processing, analysis, and storage, potentially leading to a more central role for private actors also at the operations at Statistics Finland. Second, success in one Big Data project had been made possible by outsourcing the majority of data handling to the company providing the data. Based on these findings I suggest that in an age of Big Data the production of official statistics is likely to be increasingly influenced, not by the rationalities of government departments, as has been the case in the period of register-based statistics (Alastalo, 2009b), but by those of private corporations.

By interrogating further a set of responses to Big Data at Statistics Finland, I suggest that neoliberalisation can also be identified in the ways that Statistics Finland imagines its future role and relevancy. Younger interviewees in particular felt that in order to remain relevant in the future, Statistics Finland needed to adopt the mentality of an agile and fast paced organisation, often resembling that of a company or a start-up. Drawing on findings such as this I suggest that in a paradigm of Big Data statistics, not only data, but also valuations of professional skills and expertise in its analysis increasingly originate in the private sector. By analysing in detail one effort to build partnerships with companies, I suggest that they too are reflective of growing demands to adopt private sector ways of thinking and acting also in the public sector.

I conclude the analysis in the chapter by exploring companies' attitudes towards data sharing, as perceived by my interviewees. In contrast to the markedly positive picture painted by my respondents, I suggest that there are ways in which the demands of markets are likely to come into conflict with a desire to share data for the advancement of collective goods. I identify here, in its increasingly proprietary nature, a potential contradiction in how the assemblage producing official statistics is currently being reconfigured. In sum, following the central conceptual point of this thesis, that statistics and society are co-constituted, I demonstrate empirically how the production of statistics is in Finland potentially becoming increasingly underpinned by market-, rather than social welfarist-, rationalities.

In laying out his research agenda for Big Data, Kitchin (2014b: 188) notes that ideally, "such studies would also be comparative in nature, contrasting iterations of an assemblage, such as across locales or contexts, or across various types of assemblage" as "comparative research enables generalities and specificities to be identified, and to chart the various contingent and relational ways in which assemblages unfold". This thesis is in a unique position in that it forms part of the European Research Council funded project, ARITHMUS, which, by following the working practices of statisticians at seven different sites, including five European NSIs (Office for National Statistics for England and Wales, Statistics Netherlands, Statistics Estonia, Turkish Statistical Institute, and Statistics Finland), and two international statistical organisations (Eurostat and United Nations Economic Commission for Europe), has sought to examine how digital technologies and Big Data are leading to methodological diversification and innovations in how official statistics are being conceived and produced. By starting from the idea of a transnational field of statistical practices in which the local, the national and the transnational overlap and intersect, one of the project's

major aims has been to move beyond nationally bounded case studies, or what is often termed methodological nationalism (Scheel et al., 2016).

The project's central premise is that the ways in which national statisticians position themselves in relation to phenomenon such as Big Data are "not delineated by national interests and practices" but are instead "part of transnational negotiations, contestations and tensions that cut across numerous NSIs and international statistical organisations" (Grommé et al., Forthcoming: 2). Unsurprisingly, given its sometimes utopian but often practical promises, Big Data has in recent years captured the interest of not just NSIs but also supra-national statistical organisations such as the statistical office of the European Union, Eurostat, the European Statistical System (ESS), the United Nations Economic Commission for Europe (UNECE), and the United Nations Statistical Division (UNSD). In 2013, the Heads of the National Statistical Institutes of the EU signed the Scheveningen Memorandum (2013) to examine the use of Big Data in official statistics. Efforts to engage with Big Data have thereafter often ben led by the supra-national organisations who have provided funding and other forms of support to the NSIs. For example, most of the Big Data experiments that were taking place at Statistics Finland were funded by Eurostat and similar experiments were taking place at other NSIs across Europe. Statisticians from different countries regularly convened to share experiences and to work collaboratively in projects funded by Eurostat. Illustrating the transnational character of Big Data and official statistics, the threats and opportunities of Big Data that my interviewees at Statistics Finland identified closely mirrored international debates on the topic (For a comprehensive overview of the debates around Big Data and Official Statistics, see Kitchin, 2015).

Therefore, in order to avoid the analytical pitfall of treating Statistics Finland as an isolated container of the debates around Big Data and official statistics, in the final chapter before the conclusion I situate the

developments at Statistics Finland within the transnational relations through which much of the practices of modern NSIs are debated and formed. In order to do so, I devise an analytic that draws on a large corpus of data (a collection of over three thousand documents, including fieldwork notes and various policy documents etc.) collected as part of the ARITHMUS project. My approach involves using a number of keywords from my previous chapter on Statistics Finland to conduct keyword searches in the ARITHMUS database. Rather than a comparison, I outline some of the ways in which the major themes of my previous chapter are being discussed at other NSIs and at international forums. Furthermore, although my aim is not to conduct a detailed mapping of the relative positions of authority of the different stakeholders in the field, I highlight some of the tensions and disagreements that are necessarily part of the interactions within a field.

The analysis consists of three parts. First, I explore how the words hype, threat, opportunity and paradigm come up in the ARITHMUS database. In addition to reiterating numerous points that were made also by my interviewees at Statistics Finland, the analysis helps to shed more light on how the increasing involvement of the private sector in the production and management of data is being problematised by official statisticians. The analysis furthermore points to a divergence in how administrative data and Big Data are distinguished from each other at Statistics Finland and at international forums. The analysis also highlights some of the discontents and disagreements that exist between NSIs and the supra-national organisations that increasingly come to influence their work.

In the second part of the analysis, I explore how the words skills, mentality and mindset come up in the broader corpus of data collected as part of the ARITHMUS project. This part of the analysis provides further support to the finding that increasingly, what “modernisation” comes to mean is the adoption of market rationalities also within public sector institutions such as NSIs. The analysis suggests also that cultural

differences towards issues such as data sharing within governments do exist between countries, raising the question whether some governments might currently be better placed than others to take advantage of the affordances of Big Data.

In the final part of the analysis, I explore the ways in which data access and partnerships are debated in the ARITHMUS data. As was the case at Statistics Finland, also at the transnational level the importance of obtaining access to more data sources, especially those produced by the private sector, is identified as a key concern. This includes how actors who might have previously been considered competitors, are seen as possible partners to secure access to new data sources. In marked contrast to the positive outlook of my interviewees at Statistics Finland, however, numerous issues with the increasing need to form partnerships with private actors are identified. Business interests are for example identified as being “transient”, and it is conceded that although modern companies recognise the importance of corporate social responsibility, it is usually not part of their core business models. Some of the material in the ARITHMUS database suggests also that official statisticians are in fact far from comfortable with the increasing involvement of private actors in the production and management of data.

Therefore, in addition to yielding many similar findings to my interviews at Statistics Finland, thus conforming the transnational rather than national character of Big Data, official statistics, and processes of neoliberalisation, the analysis highlights numerous issues in relation to the increasing in involvement of private actors in the production of data for official statistics. In connection to this, the analysis raises questions about the appropriateness of NSIs current responses to Big Data. I conclude by suggesting that due to their role as impartial providers of factual knowledge on societies, NSIs strategic responses to Big Data must in the long run go beyond appropriating the rationalities and practices of actors in the private sector.

I conclude this thesis by reprising its main findings, by suggesting a number of broader analytical themes and problematics that follow from it, and by pointing out directions for future research. I argue that in order to avoid the dystopian future depicted in *The Circle*, the social organisation of Big Data must be rethought based on Social Democratic political imaginations. I suggest that forming such imaginations must begin by establishing a more precise picture of the challenges that Big Data presents for different stakeholders, in addition to the NSIs. I argue that such an understanding is needed for devising an implementing a set of principles and organisational frameworks through which the power of Big Data can be unleashed for the advancement of common, rather than private, goods. I suggest that the question about the future role of the NSI must form an absolutely central component of such considerations.

2 Early statistics in a transnational context

2.1 Introduction. The co-constitution of statistics and society?

A good starting point for a study on the pressing issues around data and methods that are emerging in our own time, is to look at what others have written about the subject previously. I therefore begin the analysis in this thesis by exploring the literature around the history of statistics and its sociological interpretations. What this investigation reveals first and foremost is that long before the debate on Big Data, social scientists have not only used statistics in their research but have also studied them as a topic in their own right.

An idea that comes up repeatedly in this history is that statistics and society are co-constituted. One author that I draw heavily on is Desrosières (1998), who argues that historically, statistics have emerged out of a co-constitutive interaction between, methodological and technological developments, on the one hand, and changes in the political and administrative world, on the other. Resembling Desrosières, Law et al. (2011) suggest that social research methods are part of the "social" in two ways. First, they are social because they are constituted by the social world of which they are part, and second, they are social because they help to constitute that very same world. In other words, in their view methods exist only as long as they have advocates who promote and use them, and by describing social realities, methods help these realities to come into being.

The etymological roots of statistics, the Latin words *statisticum collegium,* which in the late nineteenth century came to designate "the science of the state", implies that the co-constitutive relationship between statistics and society (as state building) is self-evident.

However, as Kullenberg (2011: 64) notes, etymological statements do not describe the processes through which co-constitution happens. Another useful conceptual starting point for this chapter is therefore the work done around Michel Foucault's concept of "governmental rationality" or "governmentality". Foucault, for whom government meant "'the conduct of conduct': that is to say, a form of activity aiming to shape, guide or affect the conduct of some person or persons" (Gordon, 1991: 2) traced the emergence of the notion of "the population" to the end of the eighteenth century. According to Foucault (2007) statistics began in this period to indicate phenomena, movements and regularities that could not be understood as simply the result of the decisions of the sovereign. Thereafter, the population became seen simultaneously as an object in the hands of the governor and as a subject with its own needs, ambitions and possibilities. Like the physical body, it was seen as an entity that could be disciplined through statistical measurements and assessments of things such as birth and mortality rates and life-expectancy tables. In the Foucaultian view, therefore, statistics form part of the processes that "open up society for government" (Hammer, 2011: 82) of which different variations make "some ways of thinking and acting more relevant than their alternatives" (Kullenberg, 2011: 10). By "counting its citizens, territories, resources, problems, and so on […] the State participates in creating both itself, its citizens, and the policies, rights, expectations, services, and so on, that bind them together." (Saetnan et al., 2011: 2). As a result, by critically interrogating the statistics that the government uses and produces it is possible to gauge at how it "thinks" in terms of its priorities, agendas and concerns. Governmental rationalities, in other words, form an integral component of what I mean by the co-constitution of statistics and society.

Drawing on the literature around the history of statistics, in what follows I explore the interaction between statistics and society in four important historical configurations of the co-constitution of statistics and society. In particular, I demonstrate how statistical methods and infrastructures

have often emerged alongside and as part of governmental rationalities concerned with the question of how to best address social issues. Since this question is closely tied to a history of systems of expertise (Barry et al., 1996), in my review I pay specific attention to the changing forms of professional expertise that have accompanied the different historical configurations. The analysis in this chapter is crucial in order to set up the analytical framework for a later analysis of the ways in which Big Data reflects recent developments in “advanced liberalism” in western polities.

I begin the chapter by demonstrating how already in early societies rulers have practiced some form of bookkeeping on the population, usually for purposes of taxation and social control. I suggest that whereas the modern census can be understood as a tool for generating quantitative information on society, its pre-modern form was more clearly an instrument of state power and social control. I suggest furthermore that at the heart of the transformation from pre-modern to modern censuses was a change in the conception of the relation between the state and its subjects.

Next, I trace the roots of statistics in their modern form to seventeenth century Germany and England, where university professors and laymen began applying various techniques for the comparison of the relative strengths and weaknesses of states. Whereas the early German statisticians were connected to the state to an extent where they could not conceive of a civil society distinct from it, the early English statisticians operated largely outside of it, which allowed them to conceive of social phenomenon as existing irrespective of the state. Furthermore, because in England the liberal conception of the state limited its prerogatives, including the organising of large scale surveys, the early English statisticians did not have access to comprehensive data and had to therefore instead use roundabout calculations and

indirect methods. From this context emerged the predecessor of the sample survey, the population multiplier.

Thereafter I explore in detail the early parts of the nineteenth century, a period of statistical enthusiasm that resulted in the establishment of statistical societies and institutes across Europe. In England, where the industrial revolution began, parts of the middle-classes feared a violent revolution from below, and decided to begin systematically collecting data on the living conditions of the working classes in order to alleviate the difficult conditions through policy reform. The twin revolutions in England and France gave rise also to new theorising about society, specifically to the idea that society was governed by social laws and that these laws could be studied. In contrast to common belief, experiments with statistical methods had an important role in the formation of a "sociological imagination".

Finally, I turn my attention to the mathematisation of statistics in the early parts of the twentieth century, before concluding with an analysis of the politics of statistics in the period of welfare states that followed the Second World War. In modern versions of the debate around nature versus nurture it is largely forgotten that sociology emerged out of an institutional battle between the proponents of biological explanations, on the one side, and those who thought that biology and society were best kept separate, on the other. Whereas the latter camp won the institutional battle, the former invented the statistical tools of correlation and regression that are widely used today also outside of the social sciences. I conclude the analysis in the chapter by arguing that in the period of welfare states after the Second World War, the sample survey method became an important tool in assessing Social Democratic concerns with poverty and inequality.

The two histories of statistical infrastructures, on the one hand, and statistical methods, on the other, are necessarily connected to each other in important ways. In what follows, I seek to account for both.

2.2 An instrument of state power and social control: Early examples of official statistics

The first known censuses of agriculture were undertaken already in Babylonian times 3000 years B.C, meaning that the first statistics were compiled relatively soon after the art of writing was invented. In India, a treatise called Arthasastra, probably written between 321–296 B.C., contained a detailed description of the system of data collection relating to agricultural, population, and economic censuses in villages and towns. Also in ancient China the administration counted its subjects in order to determine the income revenues and military strength of different provinces. In Egypt, rulers were compiling statistical overviews long before the birth of Christ and also the Romans took regular censuses of people and property (Eves, 2002; Kuusela, 2011b: 31).

In the middle ages, attempts at conducting censuses were rare. European examples of official statistics from the middle ages include Charlemagne's Survey of the Holy Land and William the Conqueror's Domesday Book, a survey listing all the landholders along with their holdings in England in 1086. Both of these surveys were conducted for purposes of taxation and army recruiting (Higgs, 2004).

An interesting example of the early history of official statistics can be found in the Inca Empire that existed between 1000 A.D and 1500 in South America. There, each Inca tribe had its own statistician, called Quipucamayoc, who kept records of the people, houses, marriages, llamas, and the number of men who could be recruited for the army. At regular intervals, these early statisticians were brought in to the capital

of the kingdom, Cusco, where regional statistics were compiled into a national overview. This system vanished with the fall of the Inca Empire (Kuusela, 2011b: 32).

Claims of the first modern census have been made for Canada, which as a French colony had an enumeration in 1665. The name, sex, marital status and occupation was recorded for the total of 3215 persons. In Europe, the first modern censuses were undertaken in the Nordic countries: Iceland carried one out in 1703, Sweden in 1746 and Denmark in 1769. In 1790, the United States became the first nation to inaugurate a periodic census, publish the results, and to organize its government according to them (Starr, 1987: 12–13). France and Great Britain, where population censuses were for long met with opposition from both citizens and local authorities, conducted their first censuses in 1801 (Diamond, 1999: 9; Hakim, 1985: 39).

John Sinclair's Statistical Account of Scotland, published in twenty-one volumes between 1791 and 1799 was the first book written in English to use the term "statistics" in its title. The book, often considered a cornerstone of modern statistics, was based on a comprehensive social and geographical survey that was sent to local clergy. A pre-planned set of 160 questions was sent to all parishes in Scotland, 40 questions of which covered the geography and topography of the parish, 60 population and related matters, and the rest matters related to agricultural and industrial production. In 1799, Sinclair presented to the General Assembly "a unique survey of the state of the country, locality by locality" (Westergaard, 1932).

Starr (1987: 11) argues that whereas the modern census is a tool for generating quantitative information on society, its pre-modern version was unambiguously an instrument of state power and social control. He suggests that it differed from its modern form in numerous important ways. First, whereas the modern census is an enumeration of an entire

population, the pre-modern censuses were typically limited to males of particular age groups and classes. Second, whereas the modern census provides information at the individual level, its pre-modern version was usually a continuous register. Fourth, in contrast to the modern census whose results are usually made available publically, pre-modern censuses were typically treated as state secrets. Finally, in pre-modern societies statistical agencies were not necessarily distinguished from agencies charged with tax assessments and law enforcement, as is generally the case today. Starr (1987: 12) contends that at the heart of the transformation from pre-modern censuses to the modern census was a change in the conception of the relation between the state and its subjects. Whereas the pre-modern censuses assumed a coercive relationship between the two, the modern census presumes one based on cooperation.

To conclude this section, the early examples of statistics that I have introduced so far can be thought of more as bookkeeping about the population for purposes of surveillance and control, rather than as attempts at statistical reasoning in the more modern sense. When looking for the ancestors of statistics as we understand them today, two strands of development stand out. The first is the tradition of German “Statistik” that developed in German speaking cultural areas in the seventeenth century and the second the tradition of English Political Arithmetic that developed in England at roughly the same time.

2.3 A tool for the comparison of states: Predecessors of modern statistics

2.3.1 German “Statistik”

The etymological roots of the word statistics are in the Latin expression “statisticum collegium”, meaning council of state and in the Italian word

“statista”, meaning statesman or politician. In the 18th century the word “staatenkunde” was used in German speaking cultural areas to describe the comparative description of states (Westergaard, 1932: 2). By the end of the century the term statistics had entered the English language, and the Finnish equivalent, “tilasto”, was introduced in the 1840s (Luther, 1993; Westergaard, 1932: 13).

Westergaard (1932: 4) traces the practice of giving comparative descriptions of states to Aristotle, who in *Politeiai* described the history, present character, public administration, justice system, science and arts, religious life, manners and customs of 158 different states. The parts of Aristotle’s work that survived through the middle-ages provided the model for statistics compiled in 16th century Italy, where, Francesco Sansovino, amongst others, published descriptions of 22 states, including ancient Sparta, Athens and Rome. Interestingly, one of Sansovino’s chapters was devoted to the ideal state, “Utopia”.

During the seventeenth century, similar descriptions started being published in Germany, especially by university professor Hermann Conring (1606-1681), who started the long tradition of German Statistik**.** Conring described the territory, population, administration, aims and armies of Spain, Portugal, France, Japan, Morocco and Abyssinia. Conring did not have any manuscripts for his lectures, but his students published summaries of them after his death (Westergaard, 1932: 7).

A number of German professors continued Conring’s work, including Gottfried Achenwall (1719-1772), who defined statistics as the description of the constitution. According to him, statistics should confine itself to the strictly necessary facts, choosing only those that can throw light on the whole organisation of the state, its powers and weaknesses. Furthermore, statistics should deal with the present time, not the past, and historical descriptions should only be included as introductions. Consequently, Achwell displayed very little interest towards numerical

observations, despite the fact that during his time the first census had already been conducted in countries such as Sweden (Westergaard, 1932: 8–9).

Statistics, as they developed in German universities were literal, rather than numerical descriptions of countries. Drawing on Lazarsfeld (1970), Desrosières (1998: 20), relates this tendency towards literary description to the political situation in Germany during the second half of the seventeenth century. Germany was at that time divided up in to nearly three hundred poor and hostile small cities. Legal disputes over problems of territory, marriage, and succession had to be made by referring to case laws and by examining the archives, giving authority to people who were more inclined toward systematic cataloguing rather than inventing new things.

The comparisons between states were sometimes done by cross-tabulating states in rows and their descriptions in columns. Eventually this required the construction of common referents and criteria, leading to a familiar sounding dispute over whether it was okay to allow objects to lose their singularity by reducing them to numbers. The tabular form invited the comparison of numbers and was therefore an important innovation in the development of quantitative statistics (Desrosières, 1998: 21).

Gradually along the eighteenth and nineteenth centuries, numerical data became more and more important also in the German states. During the eighteenth century statistical activity increased in German-speaking countries probably as a result of courses being offered at many universities. At this point however, the objective of statistics became the establishment of universal social laws and the interest of early statisticians shifted towards the problems created by industrial capitalism (Kuusela, 2011a).

Desrosières (1998: 22–23) argues that by identifying with the viewpoint of the state, the early German statisticians could not conceive of a civil society distinct from it. In England, in contrast, where people were allowed to go about their business in a relatively autonomous way, a different relationship between the state and its subjects was emerging. There, the state was part of society, rather than its totality. From this context emerged another important predecessor of modern statistics, the tradition of English Political Arithmetic.

2.3.2 English Political Arithmetic

Since early on in Europe, population statistics had been gathered by both the government and by the church. By the Late Middle Ages, many churches in Europe had started to keep records of christenings and burials, and in 1538 churches in England were obligated by law to maintain a record of births and deaths. Church rolls were in those days used to monitor the spread of epidemics such as the plague (Luther, 1993).

The seventeenth century was in general a period of rising interest in the possibilities that statistics offered for describing and understanding social phenomenon. In London, the public availability of church rolls made it possible to report births and deaths even on a weekly basis, which provided the material for the early demographic study *The Natural and Political Observations Made Upon the Bills of Mortality*, published in 1662 by merchant John Graunt (1623-1687). The study included an estimation of London's population size and one of the first life expectancy tables (Westergaard, 1932: 16–24).

William Petty (1623-1687), who coined the term "Political arithmetic", systemised and theorised the methods used by Graunt and himself. Petty used ratios, weighted means, and other techniques to estimate population size, agricultural production, trade, and other variables. Being

a follower of Bacon, Political Arithmetic meant for Petty the application of Baconian principles to the art of government. Bacon had drawn a parallel between the natural and the political body and argued that in order to act upon the political body, one had to know it first (Porter, 1986: 19).

Similarly to the early German statistics, Political Arithmetic was concerned with the comparison of the strengths and weaknesses of different countries. Where it differed from the former, however, was in its emphasis of numerical, rather than literal descriptions. The wealth and strength of a country was seen as being strongly connected to the size and character of its population, which prompted an interest in things such as mortality rates. Until the late eighteenth century, the data for these studies were derived solely from the highly unreliable church rolls (Porter, 1986: 19–23).

Desrosières (1998: 24) highlights another important way in which Political Arithmetic differed from the early German statistics. Whereas the German statisticians were academics who wanted to construct an overall logical description of the state in general, the political arithmeticians were practical men that had forged their knowledge in work-related activities. Graunt was for example a merchant and Petty a medical doctor. According to Desrosières (1998: 24) this paved the way for a new social role:"…[that of] the expert with precise field of competence who suggests techniques to those in power while trying to convince them that, in order to realize their intentions, they must first go through him". Whereas the German statisticians had offered a general all-encompassing language, the political arithmeticians offered one that could be precisely articulated through numbers.

In line with the idea that statistics evolve in interaction with political contingencies, Desrosières (1998: 24) identifies a connection between the methods developed by the political arithmeticians and the political

circumstances in which they operated. In England, the liberal concept of the state limited its prerogatives, including the organising of large scale surveys. In 1752 for example, a plan to conduct a population census was opposed by the Whig Party as “utterly ruining the last freedoms of the English people” (Desrosières, 1998: 24). Due to not having access to comprehensive data, the political arithmeticians had to resort to indirect methods and roundabout calculations.

The most famous technique developed by the political arithmeticians was the population multiplier. Since a full census was not possible, the political arithmeticians needed to find a way of inferring the total population of the country from the number of annual births provided by the parish catalogues. Their solution consisted of first taking a census in a few places, then calculating the relationship between the population and the number of births in those places, and finally estimating the total population of the country by multiplying the general sum of births with this number (Desrosières, 1998: 25). This technique can be regarded as an ancestor to the sample survey, which has been the method of choice for data collection for both governments and academics up until very recently (See for example Savage and Burrows, 2007). I will cover sampling in more detail later.

Again tied to changing political circumstances, in the eighteenth century Political Arithmetic went in to decline. Political economists, such as Adam Smith, had doubts not only about the reliability of the available data, but also about the underlying assumptions about state guidance of economic life (Starr, 1987: 14–15). Buck (1982: 28) argues that at this point Political Arithmetic was transformed from “a scientific prospectus for the exercise of state power” into “a program for reversing the growth of government and reducing its influence on English social and economic life.” The decoupling of statistics from state ideology allowed it to re-enter public debate on new terms, which helped to stimulate the development of science.

With innovations such as the population multiplier, Political Arithmetic paved the way for a new social science of statistics that began to emerge from the early nineteenth century onwards. At this point, statistics became connected to the profound social changes set in motion by capitalist industrialisation and urbanisation.

2.4 Accounting for the revolutionary threat: The rise of statistical thinking in the nineteenth century

Due to the diverse set of principles according to which they had been compiled, up until the nineteenth century statistics were usually incoherent, irregularly collected, and difficult to get access to. Another obstacle in the study of human population had for long been the belief that it was too heterogeneous, irregular and unstable to be researched. In the first half of the nineteenth century, however, this belief started to weaken, and research on human populations started to grow rapidly (Kuusela, 2011a).

Porter (1986: 17) argues that in this period statistical writers became increasingly convinced that society was not just a passive recipient of legislative initiatives, but instead, a dynamic and autonomous entity that had to be known before it could be acted upon. For Hacking (1982), on the other hand, the improvements in official statistics in combination with the founding of statistical societies in the early decades of the nineteenth century resulted in an “avalanche of printed numbers”. In an earlier account of the history of statistics, Westergaard (1932: 136) describes the period from 1830 to 1850 as “the Era of Enthusiasm”, in which “statistics attracted public interest to an unusual degree”.

Drawing on Foucault, Barry et al. (1996: 8–9) suggest that “society” emerged in this period as a result of a mutation in the demands of governmental rationalities. Whereas preceding political rationalities had sought to govern “to the minutiae of existence”, a newly forming “liberal” governmental rationality acknowledged that government needed to deal not only with territories and subjects, but also with a complex reality that had its own laws and mechanisms that needed to be known in order to be acted upon. Thereafter government became as much about the technical as the political or the ideological, and the social sciences developed “as a way of representing the autonomous dynamics of society and assessing whether they should or should not be an object of regulation”.

This “liberal” governmental rationality manifested itself as statistical developments on two separate but related fronts. First, in the development of statistical institutions and infrastructures and second, in new theorising about the nature of social phenomenon and their stability. The material conditions underlying both were the tensions brought about by capitalist industrialisation and urbanisation. Due to its position as the first industrialised nation in the world, Britain assumes a central stage also in this period in the history of statistics.

2.4.1 Statistical societies, institutes and congresses

2.4.1.1 Statistical societies

Kent (1981: 5) points out that even though the classical authors of Sociology were born elsewhere, Britain is nevertheless the place where empirical social science first developed. As the first industrialised nation in the world, Britain was also the first to encounter its side effects. In this period, the prospect of violence arising from industrial unrest raised fears in the middle classes, and largely as a result of this, the earliest studies

in to the living conditions of the working classes in Britain trace back to the early parts of the nineteenth century (Kent, 1981: 18).

The interest towards the living conditions of the working classes led to the formation of various statistical societies. The first statistical society was founded in 1833 in Manchester, where rapid urbanisation had brought along fast population growth, acute housing problems, and diseases. The society's objective was to collect social facts that were meant to speak for themselves. By doing so, the society wanted to demonstrate the condition of society without committing itself to party politics. In 1834, the statistical society in Manchester carried out the first house to house social survey in England and perhaps in the world (Kent, 1981: 18–19).

The statistical society that attracted the most public attention, however, was The London Statistical Society, founded in 1834. This society differed from the one in Manchester in that it regarded itself almost as a branch of government. It successfully brought together government officials, party politicians and other distinguished individuals. One of its primary tasks was to provide the parliament with data that it considered necessary for social reform. In 1838, the society started to publish the journal today known as the *Journal of the Royal Statistical Society* and in 1887, the society was granted royal charter. Soon after the society in London was established, similar societies were founded in other parts of the country too (Kent, 1981: 20–22).

Britain was however not the only place where statistical societies were being founded. One famous one was for example founded in the Kingdom of Saxony in 1831, where the society got its mandate directly from the king. The American Statistical Association was founded in Boston in 1839 and in 1888, it started publishing the journal that is today known as *the Journal of the American Statistical Association*. The

statistical societies were generally formed by citizens with very little or no statistical training whatsoever (Kuusela, 2011b: 76–77).

In England, the members of the various statistical societies were mostly middle class men, often professionals, industrialists or members of the establishment. As supporters of free trade policies and economic laissez-faire, they preferred to see the towns and cities, rather than the factories as the cause behind the increasing social problems. The root of statistical activity in the early parts of the nineteenth century was therefore more often that of middle class fear, rather than sympathy towards the working classes. Because the statistical societies focused their attention on the built environment, it is questionable how much, and whether they at all contributed to promoting social legislation in factories (Kent, 1981: 31–33).

The members of the statistical societies believed that large amounts of facts were necessary in order for social reform to be possible, even if they themselves could not necessarily tell what reforms followed from their data. Their goal was not to establish a welfare state, but to create an environment that would foster a healthy and hardworking working class. Improvements in public health and education were seen as key for this goal to be possible. Many believed that if the appropriate policies did not emerge from the data, what was needed was simply more data (Kent, 1981: 31–33).

Kent (1981: 29–30) identifies the near total lack of theoretical tools to interpret data as one of the reasons why activity in the statistical societies soon began to decline. Statistical facts could in the end not speak for themselves, nor could they be separated from opinion. The membership in the statistical societies began to decline already in the late 1840s, at which point statistical activity was already increasingly connected to the government. Albeit short-lived, the statistical societies

laid the foundations for the professional character of statistical work that exists today (Kuusela, 2011a).

By the late nineteenth century, house-to-house surveys, first hand observations, and government statistics were all techniques that had been used by social researchers to gather information about people and their living conditions. Charles Booth, a wealthy businessman from Liverpool, who devoted his life and fortune to surveying poverty in London, combined all these techniques in to one study. In *Life and Labour of the People in London*, published between 1889 and 1903, Booth and his team of researchers mapped out poverty in London by allocating each street in to one of eight categories, from the semi-criminal to the upper middle classes (Kent, 1981: 52–54). Booth's methods were later adopted by Seebohm Rowntree, who along with his team of researchers used them to study the working classes in York by visiting every single working class home in the city (Kent, 1981: 25).

Arthur Bowley (1869 – 1957), an economist and statistician with a background in mathematics, codified and standardised the techniques used by Booth and Rowntree. Desrosières (1998: 166–167) argues that by doing this, Bowley created the scientific and professional norms for the "administrative statistician relying on the knowledge acquired in mathematical statistics". This figure, who became common in official statistics after the 1940s, could not have developed in France or Germany, where academic research was more theoretical and rarely had such strong connections to the administrative world (Desrosières, 1998: 177).

2.4.1.2 Statistical institutes and congresses

In addition to the emergence of statistical societies, the early nineteenth century was also a period when many statistical institutes were being founded. Perhaps the first forerunner of a modern national statistical

institute, Tabellverket, was established in Sweden in 1748, followed by Denmark in 1797 (Luther, 1993: 17–18). The Napoleonic Wars put the development of official statistics to a halt, but the interest quickly reappeared from 1830s onwards with the formation of statistical offices in France, England, Germany, Belgium, Norway, Netherlands and Russia (Luther, 1993: 50).

Following the idea that statistics are the outcome of a co-constitutive interaction between methodological and technological developments, on the one hand, and changes in the political and administrative world, on the other, Desrosières (1998) illuminates ways in which the histories of statistical agencies reflect different roles and rationalities of states in different countries. For example, although both France and England have a long history as unified states, the organisation of the state has assumed very different forms in the two countries, and as a result, so too have their statistical agencies.

In France the state and its statistical apparatus were always more centralised than in England. A general statistical bureau that had been supressed in 1812 was re-established in 1833, at which point it became part of the Ministry of Commerce. Furthermore, in France expert competency was usually internal to the administration, meaning that external experts, such as university people had a smaller influence on statistics than in some other countries. Due to the authority of the administration, statistics also aroused less public scrutiny (Desrosières, 1998: 149).

In England, in contrast, different administrations were more autonomous in relation to each other, and the county and village authorities had more power than in France. Statistics were never centralised into one sole institution, and the national bureaus had to work in collaboration with the local ones. From the outset, therefore, in England official statistics were divided between two branches: the Board of Trade established in 1832

that dealt with economic statistics, and the General Register Office established in 1837 that dealt with social statistics. Furthermore, and as mentioned previously, perhaps due to the general scepticism towards the state, in England attempts at collecting data on the national scale were for long met with opposition (Desrosières, 1998: 149).

In Germany, the early nineteenth century was a period of gradual development towards a unified empire. There, the first statistical bureau was established in Prussia in 1805. The other states, Saxony, Wurtemberg, and Bavaria, also had their own statistical bureaus which remained autonomous until the Nazis gained control in the 1930s. In Germany, the bureaus inherited the three traditions of the eighteenth century: the political, historical, and geographical descriptions developed by university professors, the administrative records kept by officials, and the numerical tables developed by scholarly amateurs. The bureaus were often led by university professors who taught “sciences of the state”, although by now in a more numerical fashion than their seventeenth century predecessors (Desrosières, 1998: 179–180).

Despite the fact that a decennial census provided the basis for political representation in the congress, the US did not establish a permanent bureau of statistics during the nineteenth century. Twenty-one censuses were conducted there between 1790 and 1990. For each one, a superintendent and provisional staff were recruited, only for the group to be disbanded once the work was completed (Desrosières, 1998: 188–189). Desrosières (1998: 196) argues that by recruiting temporary personnel for the census, the American establishment was able to hire experts whose opinions they discreetly wished to influence. This, in combination with a long-standing reluctance to increase the number of federal agencies helps to explain why a permanent Census Bureau was there not created until 1902.

Soon after the statistical institutes were established, international collaboration began to increase. Westergaard (1932) describes the years from 1853 to 1888 as the period of statistical congresses. In these years, the civil servants working in the newly founded statistical institutes began to share experiences with colleagues from abroad. According to Westergaard, the initiative leading to the establishment of the first international statistical congress came from the famous Belgian statistician, Adolphe Quetelet.

The first international statistical congress was held in Brussels in 1853, with about 150 civil servants and scientists attending from all over Europe. The chief objective of the congress was to promote the organisation of statistics and to unify reports from different countries. The congress ended in a declaration that a unified statistical system was necessary. In addition, it recommended that each country should strive towards centralising its statistical activities. After the first congress in Brussels, congresses were held in Paris in 1855, in London in 1860, in Berlin in 1863, in Firenze in 1869, in Hague in 1869, in St Petersburg in 1872, and in Budapest in 1876 (Westergaard, 1932: 175–181).

The aims of the first meetings were more practical and less focused on scientific questions. However, in Firenze in 1867, at the initiative of Quetelet, a resolution was passed that in future congresses there was to be a special section to deal with statistical questions in connection with the theory of probabilities. At the following congress a recommendation was entered that statistical investigations should deal not only with averages, but also with deviations from the mean (Westergaard, 1932: 176–180).

The International Statistical Congresses gradually faded away only to reappear with the formation of the International Statistical Institute (ISI) in 1887 (Westergaard, 1932: 183). The first ISI Session (today called World Statistics Congress) was held in Rome in 1887, and it was

attended by many of the leading statisticians of the time (Westergaard, 1932: 246).

2.4.2 Excursion: statistics, state building and democracy

In line with the Foucaultian view, which I introduced earlier, many have identified a close relationship between state sponsored economic and social development and increases in efforts to collect statistics. Starr (1987: 16–17) proposes as a general hypothesis that the more extensive the scope of state authority over economic and social life, the greater the scope, detail, and volume of statistical inquiry – but notes also a number of objections to such generalisations. First, the interests of the state do not automatically call up systems of thought. Therefore, although the Spanish empire kept more comprehensive accounts of its colonies than did the English, it was the English who came up with the majority of theoretical tools needed to interpret statistical data. Second, although more interventionist states may have a broader interest in statistical enquiry, they may also arouse more public resistance towards data collection. Numerous scholars have related the question of the extent to which people have been willing to tolerate their governments collecting data on them to the question of whether people have trusted their governments, and have therefore proposed that a relationship exists between statistics and democratic political rationality.

Starr (1987: 18–19) highlights a number of hypotheses concerning the historical link between statistics and democracy. First, one of the reasons why democracies have been interested in statistics are their use in distributing representation. In the US, the annual census was originally mandated as a means of allocating seats in the House of Representatives, and the decision to do so helped to change the relation between census takers and respondents. Whereas previously, respondents had often tried to avoid the census due to its association with tax assessment, the decision to make political representation

dependent on it created a strong incentive for participation. Secondly, Starr hypothesises that in democratic states the public might be more willing to give up data also because the aggregate results are usually made available publically. Third, in democracies statistics often function as a means of presenting and evaluating competing claims for legislation, and historical examples suggest that support for public provision of statistics has often been based on political purposes, such as the needs of legislators for information. Finally, democracy might be a particularly hospitable environment for statistics due to their usefulness in representing different interest groups, including minorities.

Analysing the more contemporary couplings between statistics and democracy, Prewitt (1987) identifies two ways in which official statistics, despite their limitations, contribute in his opinion to the functioning of democracy. First, by offering the public information on social conditions for which the government is responsible, statistics enhance democratic accountability. Secondly, by offering especially resource poor social groups a means of drawing recognition, statistics help to ensure that diverse interests are represented in politics. Prewitt sees pros and cons in both uses of statistics but is more optimistic that in the former the benefits clearly outweigh the potential negative effects.

Building from perspectives such as these, Rose (1991) argues that democratic power is *calculated power* in that “numbers are intrinsic to the forms of justification that give legitimacy to political power”, *calculating power* in that “numbers are integral to the technologies that seek to give effect to democracy as a particular set of mechanisms of rule” and finally, democratic power requires citizens who *calculate about power*, making numeracy and numericised spaces of public discourse essential for making up self-controlling democratic citizens. Modern democracy relies, in other words, in numerous ways on the existence of a statistical system and a numerically literate population able to interpret its outputs.

However, the use of statistics is certainly not limited to democratic states only, as non-democratic states also make use of them, albeit potentially in different ways. In general, the interconnections between statistics, state formation and democracy are complex, and no fixed relationship can be assumed. The examples of the co-constitution of statistics and society that I have covered so far in this chapter demonstrate that statistical apparatuses have emerged out of very different social and political contexts, and that all states, democratic or not, rely on statistical systems to know and govern their populations and economies. What is clear however is that official statistics could not have developed far without the theoretical innovations that took place in the nineteenth century, and it is to this history that I turn my attention next.

2.4.3 The discovery of social phenomena and their stability

As I already mentioned previously, the people working in the newly founded statistical societies and institutes were for the most part ordinary citizens with very little or no statistical training at all. The statistical societies in particular were committed to collecting facts that they argued would speak for themselves. However, already at this point many had doubts whether the practice made much sense and suggested instead that data could be interpreted through the use of theory. As argued by both Foucault (2007) and Porter (1986: 40–41) many had at this point come to believe that society, like nature, was governed by laws, and that these laws could be uncovered by the methods of modern social science.

The mathematical foundations of statistics drew heavily on probability theory pioneered in the seventeenth century by the mathematician Blaise Pascal (1623 – 1662). Prompted by a friend interested in gambling problems, Pascal introduced the notion of expected value thereby laying the foundations for probability theory. In the early

nineteenth century, Pierre-Simon Laplace (1749 – 1827), Adrien-Marie Legendre (1752 – 1833) and Carl Friedrich Gauss (1777 – 1855) used the method of least squares to minimize errors in data measurement in the field of astronomy. In 1802, Laplace produced an estimate for the entire population of France based on samples from only three communities. Despite this achievement, it took long before the majority of statistical writers began using probability theory in their work (Hacking, 1990).

According to Desrosières (1998: 67) the key concern for statistical thinkers in the early parts of the nineteenth century was the question of how to make single units out of multiple ones. Soon, the notion of the average value became a key tool in objectifying the social world, and once it became possible to objectify the social world in this way, a debate followed on the nature of the object. What was at stake was the question of whether society could have an autonomous existence in relation to its individual members.

Next, I review a number of answers given to this question. First, I review the answer given by the Belgian statistician Adolphe Quetelet (1796-1874), and second, ones given by some of the classical authors of Sociology. Even though Quetelet is usually not included in the official canon of sociology, his work was similarly to the classics prompted by the two revolutions, the political one in France and the economic one in England. In their work, both Quetelet and the classics tried to come to terms with the rapid breakdown of old social forms and values.

2.4.3.1 Adolphe Quetelet

The Belgian astronomer Quetelet was perhaps the most famous nineteenth-century scientist to believe that a numerical social science could uncover the laws of society. Quetelet contributed to the emergence of modern statistics both as a statistician and as an organiser of

statistical activity in nineteenth century Europe. His interest towards statistics came initially from his work in astronomy, and he was set apart from most of his statistical contemporaries by his commitment to mathematics. On the other hand, not many mathematicians or astronomers were interested in social questions (Stigler, 1986: 161–169).

Quetelet's interest in social phenomena grew especially after a visit to Paris in 1824, where he had familiarised himself with the tradition of Laplacian probability theory. Around the time, the French government had started to publish statistics that indicated that crime remained fairly constant from year to year. From this Quetelet concluded that not only nature but also society was governed by laws. Furthermore, for Quetelet, society should be considered an entity of its own, independent of the actions of its individual members. Poisson's Law of Large Numbers, the theorem that general effects in society are always caused by general causes, became a fundamental principle of Quetelet's "social physics" (Porter, 1986: 43–52).

The relative regularity of births, marriages and deaths from year to year had been discovered already in the eighteenth century, but at that time it had been interpreted as further proof that a divine creature was directing society from above. According to Desrosières (1998: 74–75) Quetelet was able to provide a new answer to the question why there nevertheless was such a remarkable diversity of physical and mental traits between individuals. Quetelet maintained that the law of error (later called the Normal Distribution) was true also for human beings, and that this made it possible to determine the average physical and intellectual features of a population. This divine creation, the average man, had both moral and physical attributes, and his development could be studied over time. Quetelet interpreted the normal distribution as evidence that departures from the mean were like errors in measurement, so that the mean value was the "true value" representing an underlying value or type

(Desrosières, 1998: 52–54). Statistical regularities were for Quetelet an evidence of determinism: an individual might think that marriage was his own choice, but it was in fact pre-determined (Kuusela, 2011b: 83).

Quetelet's major goal was to measure the changes experienced by the average man over time. By doing this, Quetelet hoped to discover the forces that acted on the "social body", and after that predict its future. In practice though, no calculations were included in Quetelet's written work. Porter (1986: 54–55) suggests therefore that Quetelet's lasting contribution to statistics came not from his methodological innovations, which remained modest, but from the fact that he was able to connect French probability theory with the work of the newly founded statistical services.

During his lifetime, Quetelet enjoyed a celebrity status as an organiser of censuses, statistical services and international statistical congresses. He was also important in that he was able to convince his successors that instead of looking for concrete causes of individual phenomena, they should concentrate on the statistical information presented by the larger whole. Quetelet's hypotheses inspired work that later led to the development of many of the statistical principle still in use today (Kuusela, 2011b: 84–87).

Based on Quetelet's idea of the average man, a new type of survey research method, the monograph survey, was for example introduced at the end of the nineteenth century by Frederic LePlay (1806 – 1882). The method was widely used during the nineteenth century especially to explore family budgets and even at the start of the twentieth century the method was officially still accepted by the International Statistical Institute (Kuusela, 2011b: 84).

Probably the most famous of Quetelet's apprentices was Wilhelm Lexis, who is best known for his work on dispersion. Lexis worked mainly on

developing mathematical methods for research on the stability of statistical series, such as the ratio of sexes at birth. According to Porter (1986) Lexis' emphasis on the measurement of dispersion of statistical series was intended as a critique of statistical determinism as advocated by Quetelet. Whereas Quetelet aimed to set every series within a unique "normal" model by assuming their homogeneity and stability, Lexis stressed its fluctuations. It has been suggested that Lexis' analysis of dispersion foreshadowed more modern analysis of variance (Kuusela, 2011b: 86).

2.4.3.2 The birth of modern sociology

The nineteenth century was also the period in which modern Sociology was born. The early classics of the discipline, such as August Comte (1798-1857), Karl Marx (1818-1883), Max Weber (1864 – 1920), Emile Durkheim (1858-1917) were all concerned with objectifying and understanding forces that existed outside of the will of the individual. It is interesting to note that although the story of the birth of sociology is usually presented as a tale of a handful of geniuses inventing theoretical frameworks for interpreting social problems, most of the classics were early on in their careers engaged in empirical research. Perhaps this should not come as such as surprise, especially in light of more recent research that has highlighted the role of the early experiments in statistics in the emergence of the "sociological imagination" (Bayatrizi, 2009). Osborne et al. (2008) go as far as to argue that it is largely thanks to retrospective commentary that the classics are considered theoretical thinkers at all.

August Comte, for example, who coined the term "Sociology", believed that society was governed by social laws but objected strongly to any attempts at quantifying them. Comte's "positive" method was based on observation, experiment and comparison. Although he deemed observation to be key, he never explained how they should be made or

analysed in practice. Comte never engaged in any form of empirical inquiry himself, asserting simply that facts were plentiful (Thompson, 1976).

Max Weber, known usually only as a social theorist, was in fact on numerous occasions engaged in empirical work. Weber's career in empirical research included investigations into agricultural and industrial labour conditions and investigations into workers' attitudes and work histories, utilising both questionnaires and direct observation. In addition, Weber used a statistical approach in a study on the psychological aspects of factory work and in a critique of another study concerning workers' attitudes. In the end, Weber's attitude towards quantitative methods remained ambivalent, and he wondered whether empirical work was best left to psychologists. Like many others, Weber was in the end unable to integrate his empirical studies with his efforts to construct a general sociological theory (Lazarsfeld and Oberschall, 1965).

Also Karl Marx once tested his skills as a quantitative social researcher. In his period of exile in London, Marx used secondary sources, government records in particular, to document the extent to which workers were being exploited by the capitalist system. In 1880, Marx drew up a questionnaire including over one hundred items, and distributed it in France to workers societies, to socialist and social democratic groups, and to newspapers. The survey covered occupation and conditions of work, working hours and leisure activities, terms of employment, wages and the cost of living. In the end, Marx received very few replies and none of the results were ever published (Kent, 1981: 2–3). Important to note also is that that many of the ideas that formed the basis of Marx's theories were already apparent in the empirical enquiries in to life in Manchester's slum districts conducted by Frederick Engels in the 1840s. Marx, who formed a lifelong partnership with Engels, openly admired Engels' empirical accounts of the condition of the working

classes in England (Kent, 1981: 43).

In the study that is generally considered to have founded quantitative sociology, Suicide (1897), Emile Durkheim used statistics to illustrate how the individual was dominated by a moral reality greater than himself. According to Durkheim, statistics revealed a collective impulse that could not be reduced to individual averages as Quetelet had thought. For Durkheim, data on numerous individual actions revealed the social tendencies that were the actual reasons behind individual behaviour (Porter, 1986: 69). Later on, Durkheim attempted a critique of statistical knowledge and distanced himself especially from Quetelet's notion of the average type. For Durkheim, Quetelet's average man was the arithmetic result of selfish individuals, whereas his notion of the collective type was a reflection of a general will that existed before "the vote of the majority" (Desrosières, 1998: 101).

Porter (1995) argues that in the second half of the nineteenth century confidence in the value and reliability of statistical laws had reached a point where the social science of statistics could become a model for certain areas of the physical and biological sciences. Analogies of social science were thereafter used to justify the application of statistical reasoning to problems such as thermodynamics, heredity, and price fluctuations. In light of this it is perhaps surprising that modern sociology has been comparatively slow to pick up a probabilistic mode of reasoning. This has prevented sociologists from participating in interdisciplinary projects to the same extent that for example economists have been able to. Erola (2010) suggests that one of the reasons for this is the fact that that the early classics of sociology, such as Durkheim's Suicide, were written slightly before probability theory matured in the nineteen twenties and thirties.

2.5 How to identify the poor? The mathematisation of statistics in the late nineteenth and early twentieth century

Although the origins of statistical mathematics lie in the advances in probability theory in the eighteenth century, the modern field of statistics only emerged in the late nineteenth and early twentieth century. Desrosières (1998: 103) argues that whereas in the nineteenth century the aim of statistical theory had been to objectify aggregates of individual objects, in the twentieth century its objective became to measure the relationships between objects.

After the Second World War, the hegemonic mode social scientific reasoning has been to regard environmental factors as the key in explaining recursive patterns in social behaviour. In light of this it is perhaps surprising that the origins of mathematical statistics lie not in social scientific work as we typically understand it today, but in an attempt to prove that it is in fact biology, and not the environment, which explains why some people end up affluent and happy, while others remain poor and less happy.

In contemporary versions of the debate around nature versus nurture it is furthermore often forgotten that at least in Britain, modern Sociology emerged out of an institutional battle between the proponents of biological explanations, on the one side, and those who thought that biology and society were best kept separate, on the other (Renwick, 2012). Whereas the latter camp won the institutional battle, and was therefore able to assert a major influence on what was to be deemed as the legitimate mode of reasoning in the social sciences, the former camp invented the statistical tools of correlation and regression widely used today also outside of the domain of the social sciences.

2.5.1 Eugenecists, social reformers, and the question about the nature of poverty

After a period of prosperity in the mid nineteenth century, the Victorian boom, social tensions were again on the rise in the last decades of the century. A depression had decreased prices and profits, and the interests of capitalists and workers were once again diverging. In London, clearances to make way for railways, government offices and warehouses had created an increased pressure on housing, which had led to overcrowding. These conditions resulted in demonstrations, strikes and rioting, raising again the fears of the middle classes that violence might arise from below (Kent, 1981: 64).

In Britain, urban degeneration was seen not just as a domestic problem, but also as something that was threatening the military strength of the empire. While some parts of the middle-classes, sometimes referred to as the social reformers, believed that social tensions could be alleviated by lifting people out of poverty through social reform, others, the eugenicists, believed that the problem had a biological nature, and that it could therefore only be solved through a biological solution. Neither the social reformers, nor the eugenicists believed that the working classes a whole constituted a problem, only its degenerate parts (MacKenzie, 1981: 39–40).

A key methodological concern for the nineteenth century social reformers and eugenicists was therefore the question of how to identify the degenerate parts of the working classes. One solution to the problem was offered by the famous social explorer, Charles Booth, who, along with his team of researchers surveyed the streets of London and allocated each in to one of seven categories, from the wealthy upper-middle and upper classes to the vicious and semi-criminal. Booth concluded that contrary to some popular suggestions according to which a third of London’s population lived in absolute poverty, the actual

number was in fact only eight percent. Even more importantly, only one percent of the entire population of London belonged to the group that constituted the really dangerous parts of the working classes (Kent, 1981: 55, 1985: 56–57).

Booth's suggestions as to what should be done with these people was not much more humane than those put forward by the eugenicists. Booth argued that the very poorest should be allocated in to workhouses to be taught work discipline, while their children would be taken away and raised under strict supervision. Those that failed to work properly in these camps would be sent to poorhouses, and their children would be taken away from them for good. Those that succeeded, however, would be allowed to re-enter society. For Booth, this constituted a form of state socialism where the state was allowed to exercise its power over those that could not help themselves, and that constituted a danger both to their communities and to society as a whole (Kent, 1981: 59, 1985: 55–56).

Another solution to the problem of how to identify the degenerate parts of the working classes came from that part of the middle classes who believed that poverty resulted not from environmental factors, but from biology. Francis Galton (1822 – 1911) used Booth's social classification as an indicator of natural individual aptitude and argued that just like height, genetic worth too was innate and normally distributed. Whereas Quetelet had viewed the normal distribution of human attributes as the result of a large number of variable random causes, Galton wanted to isolate the one cause he saw as crucial, heredity. By seeking an explanation to the paradox that although by knowing the height of the father, one could not determine the height of the son, and yet the dispersion from generation to generation remained constant, Galton was able to formulate the concept of regression (Desrosières, 1998: 113–116).

Galton believed that each individual possessed a fixed quantity of the characteristics that made up his or her civic worth. Whereas the lowest of Booth's categories corresponded to the groups with the smallest quantities of civic worth, the highest categories constituted for Galton the "brains of the nation". Galton therefore, in effect, read the structure of Booth's social classes onto nature. On the basis of this theory, particular social policies were put forward. "Positive eugenics" meant boosting the fertility of the upper classes, while "negative eugenics" meant preventing the lower social groups from breeding. Eugenicists did not regard environmental factors as having much importance, since at the end of the day they would only lead to acquired characteristics that could not be inherited (MacKenzie, 1981: 18–19).

From 1860s to the 1880s, Galton worked on statistical problems with the occasional help of a number of mathematicians. During this period Galton formulated the concepts of regression and correlation, and thereby extended the range of statistics from questions concerning single variables to questions concerning many. Galton was able to demonstrate that genetic combinations were governed by the laws of probability, and that this implied a stability of inherited characteristics (MacKenzie, 1981: 10).

These ideas had a major influence on the mathematician Karl Pearson (1857 – 1936), who while developing and systemising Galton's insights, made many important contributions of his own. The standard formula for the correlation coefficient and the widely-used "chi-square" test of the goodness of fit between observations and theoretical predictions are both named after him. Pearson also became the first head of a statistical department at a university. In addition, he established and edited Biometrika, a journal which from 1901 onwards became the major publication venue for work in statistical theory. Pearson managed his own group of researchers, the biometricians, and taught the first courses in advanced statistical theory (MacKenzie, 1981: 10).

Pearon's student George Udny Yule (1871 – 1951) became interested in applying statistical techniques to social problems. Yule took Pearson's formulation of the correlation coefficient and laid the foundations for partial correlation and linear regression for any number of variables. His interest in the relationship between pauperism and out-relief (assistance given to a poor person who did not live in a workhouse) led him to invent "Yule's Q", a statistic that can be used to measure associations between sets of categorical data (Selvin, 1985: 74–75).

Yule's interest in the relationship between pauperism and out-relief (financial assistance given to poor people not living in a workhouse) was prompted by Booth's studies of poverty in London. Despite this, and in addition the fact that Yule actually wrote a commentary criticising Booth for the way he had interpreted his data, Booth never used correlation or regression in his studies. It is interesting that despite probably being aware of the developments made by Galton, Pearson and Yule through publications such as the Journal of the Royal Statistical Society, none of the late nineteenth century social explorers, Booth, Rowntree, or anyone else, used correlation or regression coefficients even though their accounts of poverty were otherwise highly numerical (Kent, 1981: 97, 1985: 65–66; Selvin, 1985: 70–75).

The credibility of biological explanations to social phenomenon started to weaken from the First World War onwards, only to completely lose momentum as a consequence of the atrocities of the Nazi regime during the Second World War. Some of the research questions of eugenics were integrated in to human genetics, and correlation and regression became widely applied in many fields of research, also outside of the social sciences. At this point, academic sociology became closely tied to an expanding government welfare apparatus (Osborne and Rose, 2008).

MacKenzie's (1981) notes on why the environmental explanations "won" the scientific debate after the Second World War are somewhat haunting. According to Mackenzie, after the war, environmental factors were better suited for the political context of a welfare state that wanted to integrate the majority of the population in to society. He (1981: 50) continues:

> *As that accommodation comes under threat in the 1980s, it would not surprise me if the tactical balance begins to shift back towards eugenics.*

This brings to mind recent statements by the British politician, Boris Johnson, according to whom the recent growth in inequality can partially at least be attributed to the superior intellectual ability of those at the very top of society (Johnson, 2013).

2.5.2 The method of choice of the post-war welfare state: The sample survey

The period after the Second World War was characterised by the prevalence, or indeed dominance, of the sample survey. The Norwegian Anders Kiaer (1838 – 1919) is often attributed as the first person to have suggested that a representative sample should be used instead of a complete enumeration. In 1894, Kiaer conducted a representative survey in Norway that covered occupation, income, expenditure, days missed from work, marriage and the number of children, and soon after, the initiative was widely discussed at the international statistical congress (Kruskal and Mosteller, 1980: 172–175).

Since Kiaer was not mathematically oriented, his presentations did not include a formal description of his method. What Kiaer wanted to demonstrate was that by taking some precautions in the choice of a sample, it was possible to obtain sufficiently good results with a few control variables (already present in exhaustive enumerations) to

suppose that the results were good enough for the other variables too. The technical formulation of the confidence interval was presented by Arthur Bowley in 1906 (Kruskal and Mosteller, 1980: 175–184).

When presenting his method to the International Statistical Institute in 1885, one of the justifications Kiaer gave for it was that in contrast to previous surveys, the representative sample included all classes of society. Kiaer emphasised that in order to properly assess the condition of the working classes, one had to also be familiar with the condition of the other classes. Desrosières (1998: 227) argues that by doing this, Kiaer was among the first to raise the issue of social inequality. Kiaer viewed his survey method as useful in creating funds for retirement and social security, guaranteeing social standardization, and in the statistical treatment of risks.

Desrosières (1998: 221) suggests that the history of the sample survey can therefore be read in parallel with the emergence of the welfare state. Whereas previously statistics had been used for holistic analyses of the social world, by the late nineteenth century they started being used in applying and evaluating policies designed to affect individuals. The establishment of the first laws of social protection, the development of national consumer markets, and the possibility of nationwide electoral campaigns were key factors behind the increasing popularity of the sample survey. All of the aforementioned are indicative of a change whereby local modes of management began being replaced by national ones.

Throughout the nineteenth century, social surveys had especially in Britain been accompanied by a recommendation to improve the morality of the working classes. Both the social reformers and the eugenicists had wanted to identify the morally corrupt parts of the working classes in order to prevent them from causing harm to society. Poverty was in this context seen as fundamentally a spatial phenomenon. The early studies

in to urban poverty put this premise into question. One of Booth's major findings was for example that poverty was barely lower in the whole of London than in the East End. Joseph Rowntree (1836 - 1925), on the other hand, discovered that the percentage of poor people was almost the same in York as in London. As a result, it became increasingly difficult to see poverty as something that resulted from the lack of morality of some individuals in some localities. Instead, poverty started being seen as a structural problem on the national level, one that could not be treated locally, but only with national policies (Desrosières, 1998: 221).

Another contributing factor in the growth in popularity of the sample survey was an increasing debate surrounding the rivalry between large industrial towns. The English Board of Trade received significant public funds to conduct a major survey of working conditions in a large number of different towns in different countries. A by-product of these international comparisons were comparisons between towns within countries. Although the survey did not apply probabilistic methods, it was therefore the first method with an international scope (Desrosières, 1998: 221).

These developments gave Arthur Bowley (1869 – 1957), an economist and statistician with a background in mathematics, the impetus to formulate and standardise the conditions that made representative surveys possible (Hoinville, 1985: 103–104). In contrast to Booth and Rowntree, Bowley did not base his assessments of poverty on visual impressions made during visits to households, but instead, he used quantifiable and constant variables. He was also not interested in the question of whether poverty was caused by a lack of individual morality or by structural forces (Desrosières, 1998: 224). Desrosières (1998: 224, 166–167) argues therefore that by replacing moral judgements with neutral technical assessments, Bowley laid the foundations for a new profession, that of the professional government statistician.

After the Second World War, the sample survey became arguably the most important technical tool of social enquiry, utilised both by the government and by academic researchers. In Britain at least, the background for its growing use in government were the needs of the war time economy. There, the wartime social survey became a means of gathering quantitative information to supplement qualitative data obtained from elsewhere, concerning in particular the question of “public morale”. The early inquiries reflected the immediate needs of the wartime government, and included topics such as nutrition, prevalence of illnesses, and obstacles to service. After the war, the survey started being used as an instrument for planning in the longer term. It was discovered that with the sample survey, it was for the first time possible to for example provide a detailed picture of the housing situation in the country as a whole (Whitehead, 1985: 84–85).

Savage (2010: 201) concludes that after the war, the capacity to conduct a sample survey became the key feature of a modern state. By allowing new ways to track changes in things such as inflation rates, crime rates and poverty rates, it helped to create the concept of a flat, bounded, homogenous national space which did not exist before then. Furthermore, because it allowed for the measurement of “social mobility”, it became a key tool in campaigns for comprehensive, rather than selective education (Savage, 2010: 210–211).

By the 1970s, therefore, the sample survey had become a key tool for assessing social democratic concerns with poverty and inequality, applied often by the government in collaboration with leading social science centres. However, this constellation came in to question in 1980s, as the British government grew increasingly impatient with the time demands of the surveys. As the technique matured and the findings became more and more mundane, the government became less compelled to wait for the results. At this point social scientists had

become increasingly able to analyse data according to their own time schedules and the government became more catholic and eclectic in its choice of research methods (Savage, 2010: 211–212).

2.6 Conclusion

In this chapter, I have reviewed literature around the history of statistics and its sociological interpretations. From this exploration emerged the conceptual starting point of this thesis: that statistics emerge out of a co-constitutive interaction between, methodological and technological developments, on the one hand, and changes in the political and administrative world, on the other. Governmental rationalities, particularly in relation to the question of how to best address social issues, have historically formed an integral component of this co-constitution.

Instead of a conclusive history, I highlighted four important configurations of this co-constitution in the history of statistics. I began by looking at early examples of official statistics when rulers of early civilisations produced some form of aggregate statistics on their populations. In seventeenth century Germany, university professors began using cross-tabulations to make literal comparisons between states, and around the same time in England, the so called political arithmeticians used data obtained from church rolls to calculate the first fertility and mortality tables.

Despite these developments, statistics concerning human populations did not emerge properly until the nineteenth century. Desrosières (1990: 195) argues that historically, statistics have been concerned with the question of “how to make things which hold together”. One influential answer to this question came from the Belgian astronomer and early statistician Quetelet, who introduced the notion of the average value and

normal distribution to the study of human populations. For him, the average physical and intellectual features of a population constituted an average man, a divine creation whose development the social statistician could study.

Quetelet's conception of social forces that existed before the will of the individual informed the classical authors of sociology, such as Comte, Marx, Weber and Durkheim. Although the story of the birth of the discipline is today usually told as a tale of a handful of geniuses inventing theoretical frameworks for interpreting social problems, most of the classics were in fact engaged in empirical social research. More recent commentaries have emphasised the role of the early experiments with statistics, such as those conducted by the political arithmeticians, in the formation of the "sociological imagination" (Bayatrizi, 2009). What Quetelet and the classics furthermore shared was the social context in which they operated, as they were both confronted by the rapid social changes brought about by the twin revolutions in France and England.

The political tensions brought about by industrial capitalism formed the backdrop also for the mathematisation of statistics which occurred in the early parts of the twentieth century. Parts of the working classes were seen as forming a threat to society by the middle classes, and the statistical techniques of correlation and regression emerged out of an attempt to prove that it was in fact biology, and not the environment, which explained why some people ended up poor and dangerous. Poverty was seen as a local problem that could be solved by identifying those who were beyond helping by conventional means.

The expansion of research techniques put this belief in to question. The unprecedented scale of the studies of Booth and Rowntree proved that poverty existed not just locally, but at the national level. In the wake of the Second World War, the nationally representative sample survey became the method of choice of the emerging welfare state. Whereas

the sample survey was the method of choice of the welfare state, it remains to be seen what type of a society Big Data will be a reflection of. It is this intellectual task that I take up in chapter four. Before that however, in the next chapter I explore in more detail the historical development of statistics in the country where my fieldwork was based, Finland.

3 A brief history of social statistics in Finland

3.1 Introduction

Sweden's first national population statistics were compiled in 1749 by the king of Sweden, Fredrik 1. Sweden had for some time been a major military power in Northern Europe, but by 1749 its power was fading. This became ever more apparent with the results of the first population statistics. The government was horrified to discover that instead of a population of twenty million, Sweden only had two million inhabitants. In addition, the population statistics revealed many social problems, such as high mortality and migration rates, and the news about a population enumeration being undertaken raised such widespread interest that the government decided to keep the results secret. Although by now there had been many attempts at counting the population in other countries too, Sweden was the first to start compiling population statistics annually. As a result, Sweden and Finland (a country that was part of Sweden until the early nineteenth century) today have the longest consistent statistics on population and population changes in the world.

The early interest in official statistics in the Nordic countries has continued through the centuries, and today, population statistics in Finland, Sweden, Norway and Denmark are arguably among the most comprehensive in the world. The first statistics were in these countries compiled from records of births and deaths maintained by the Lutheran church. Therefore, when Finland in 1990 became only the second country after Denmark to have its census based entirely on data derived from various administrative registers held within the state, it in fact reverted back to centuries-old data collection methods (Alho, 1999). Due to their perceived advantages, such as low cost and low respondent

burden, register-based statistics have in recent years gained increasing popularity also outside of the Nordic countries (UNECE, 2007).

In this chapter I review the development of official statistics in Finland, from the parish catalogues in the Kingdom of Sweden to the modern administrative registers in Finland as an independent republic. As pointed out by for example Kitchin (2014b: 17), despite data having both a temporality and a spatiality, to date there have not been many histories and geographies of data assemblages. In this chapter, therefore, I seek to ground the ongoing changes in official statistics, which I will attend to in the empirical part of the thesis, in an understanding of historical configurations of the co-constitution of official statistics in Finland. Instead of attempting to conduct a complete taxonomy, I focus my attention on some of the main reconfigurations that have occurred. I highlight in particular shifting jurisdictions between different institutions involved in the production and maintenance of data and the making of statistics. Furthermore, I start each section with a brief historical overview of the period in question. Due to a limited availability of sources, I rely heavily on only a handful of sources especially when covering early configurations of statistics in Finland.

I begin by exploring statistics in the Kingdom of Sweden, a country that Finland was part of for over half a millennium. There, the monarch's administration's initially lukewarm interest towards statistics was later heightened as a result of war. This increasing interest in combination with the arrival of new intellectual currents from central Europe culminated in the formation of the first national population data collection system in the world, Tabellverket (Tables office), in 1756. After that, I examine statistics in the period beginning at the start of the nineteenth century in which Finland formed an autonomous Grand Duchy within the Russian empire. In this period, a temporary weakness of Russia allowed Finland to take a significant step towards independence by establishing its own Central Statistical Office. Then, I explore statistics in the period

shortly after Finland gained her independence in 1917, when, despite the social and political turmoil surrounding it, the work of the Central Statistical Office continued uninterrupted. After that I focus my attention on developments following the Second World War, in particular the period beginning in the 1960s in which a register-based statistical system was established as part of the technologies of an emerging welfare state. Following Alastalo (2009b), I define register-based statistics as a form of welfare state governmentality, after which I explore some of their uses in social scientific research along with a more general reflection on methodological developments in Finland. I conclude the chapter by drawing a number of analytical conclusions from the review conducted.

3.2 Statistics in Finland: From parish catalogues to governmental registers

3.2.1 Statistics in the Kingdom of Sweden (c. 1250 – 1809)

Though interesting to both the Catholic Church in Sweden and to the Greek Orthodox Church in Novgorod (Russia), up until mid-twelfth century, the land area that today constitutes Finland was a political vacuum. In the thirteenth century, however, the struggle for the political and economic control of the coastal region of the Gulf of Finland intensified into a battle of which Sweden eventually came out on top. A peace treaty was signed between Sweden and Novgorod in Nöteborg in 1323, which assigned only eastern parts of Finland to Novgorod, while the western and southern parts of Finland were tied to Sweden and hence to the cultural sphere of Western Europe (Meinander, 2011: 8–10).

Between the fifteenth and the seventeenth centuries, Finland developed into an economically and militarily important part of the Swedish realm.

As a result of Swedish rule, Sweden's social and judicial systems took root also in Finland. Since feudalism was not a part of this system, Finnish peasants never became serfs, but were always able to retain their personal freedom. In that time period, Finland's most important economic and cultural centre was Turku, a town in the east coast of Finland, and a notable trading post already in the Viking period (800 – 1025 AD). The castle of Turku, built between the thirteenth and the sixteenth centuries, is still today Finland's most important religious edifice (Meinander, 2011: 6–17).

By the mid sixteenth century, the Reformation set in motion by Luther had reached Sweden, and eventually, the Catholic Church lost out to the Lutheran faith. As a consequence of the Lutheran conviction that the fellowship with God was personal and direct, and that therefore all Christians should be able to read the Bible for themselves, Finnish-language culture started to rise. Although some historians have estimated that the same tongue has prevailed in Finland for more than ten thousand years, there was no written Finnish language before the sixteenth century. In 1548, The New Testament was translated into Finnish by the Bishop of Turku, Mikael Agricola. Despite this achievement, it was not until the 1880s that Finnish overtook Swedish as the official language in Finland. Until then, the language of the court and the aristocracy was Swedish (Meinander, 2011: 23–25).

Between 1617 and 1721, Sweden was at the height of its power. During this period, often described by popular historians as the Great Power period, Sweden extended its realm around the Baltic, and managed, due to the temporary weakness of Russia, to push the Finnish border further to the east. With consolidation of power in Stockholm, uniform Swedish rule was extended in Finland. The newly established civil service departments were often led by Swedes, a factor that helped to strengthen the relative position of the Swedish language in Finland (Meinander, 2011: 35–39).

The earliest forms of governmental statistics in Finland therefore trace back to the time when Finland was still a province of the Kingdom of Sweden. From the sixteenth century onwards, the Swedish monarch's administration maintained various registers of people that could be taxed. Land registers contained detailed information on individuals practicing agriculture and the customs authorities kept records on foreign trade. Information from various parts of the country were sent to the capital, Stockholm, where a national summary was compiled. By the seventeenth century, the central government had started using this data for investigation and planning (Luther, 1993: 21).

Another important source of early statistics were the parish catalogues maintained by the Lutheran church. These registers included records of births, marriages, deaths, and lists of parishioners, also known as Communion books. Following the example set by countries in central Europe, registers on marriages, christening and burials were ordered compulsory in the 1686 church law (Myllys, 1981: 55). Already at that time the parish catalogues included information on where "a person had come from, how he had behaved and where he had gone" (Nieminen, 1999: 8). Even though the registers were initially collected to serve the church's internal purposes, their usefulness in enumerating social phenomenon was quickly discovered. The first mortality and fertility tables were in Sweden calculated by Lutheran ministers in the early parts of eighteenth century (Nieminen, 1999: 8–9).

However, it was not until the Great Northern War (1700 – 21) that the government truly became interested in the possibilities offered by statistics. The war, which saw Sweden lose its hegemony over the Baltic, resulted in many casualties, raising the government's concern whether a sufficient workforce still existed. In addition, the government believed that outmigration was posing a serious threat to the future existence of the country (Nieminen, 1999: 9). The war had furthermore brought an

end to absolute monarchy, and the resulting power shift from the monarch to a parliament led by the estates resulted in a heightened interest in the wellbeing of the common people (Luther, 1993: 21; Meinander, 2011: 56–57).

Already in the early parts of eighteenth century, some members of the board of trade had suggested that a population count based on parish catalogues should be undertaken. This suggestion was initially objected out of a fear that it would lead to the same dire consequences as to King David in the bible, who having decided to count his kingdom's population had ended up with widespread plague in his kingdom (Luther, 1993: 23). These types of worries were, however, soon set aside by the arrival of new intellectual currents from central Europe.

One sign of the arrival of the Age of Enlightenment in Sweden was the formation of the Royal Swedish Academy of Sciences in 1739. In addition to its central tasks of advancing research in the natural sciences and establishing conditions for economic prosperity, since early on its member also held an interest in demographics (Luther, 1993: 23). As I covered in more detail in the previous chapter, the eighteenth century was in general a period of increasing interest towards quantitative descriptions of social phenomenon. Although Sweden was geographically in the periphery, many of its intellectuals were offered the opportunity to study abroad, which helped to ensure that the Swedish academy was kept up to date on the latest developments in science. Through its connections to science societies in other countries, such as to the Royal Society of London in England, the Swedish academy was aware of the developments in political arithmetic (Luther, 1993: 23; Nieminen, 1999: 9).

In 1746, the Royal Swedish Academy of Sciences handed the parliament an estimation of the population count of Sweden and around the same time, army officials suggested that a population count should be

conducted on a yearly basis. In 1749, the first official population statistics were compiled with the support of the parliament, and seven years later, the first national population data collection system in the world, Tabellverket (Tables Office), was established (Luther, 1993: 23). Although Tabellverket is often regarded as a predecessor to the modern statistical agency, its tasks were modest in comparison.

Three separate steps were used to gather the data in the original population data collection system. First, Lutheran ministers copied the information from the parish catalogues on to forms prepared by the Tables Office. Each church then forwarded its forms to a provost, whose task it then was to prepare a summary for the municipality. Summaries of municipalities were combined in to a summary for the entire province which was then sent to the capital, Stockholm, for a national summary. As I mentioned already in the introduction, the results of the first data collection were treated as a highly sensitive state secret (Myrskylä, 2011).

The first population enumeration was conducted with three different forms. Two forms were used for births, deaths and marriages, and one form was used for the population count divided in to age groups. Initially ministers had to inform the population count on a yearly basis, but later the interval was reduced to three years in order to reduce the work load on the ministers. Births, deaths and marriages were categorised according to gender and month, and each birth was also categorised according to whether it took place within or outside of a marriage (Nieminen, 1999: 12).

Initially only those burials that took place on the churchyard were reported. From early on, children that died before their first birthday were reported separately from others. This practice allowed for the first time the examination of infant mortality. Deaths were classified according to the cause of death and the deceased's marital status. The classification

of diseases caused uncertainty among the ministers, who were later on handed guidelines on the typical signs of different diseases (Nieminen, 1999: 12–13).

Even the earliest data collection forms included a section on a social stratification. In the eighteenth century, the Swedish society was strictly divided in to two halves. The first half was formed of those who belonged to one of the estates, nobility, clergy, burghers or the land owning peasants, and the latter half was formed of the majority of the population, who did not belong to any of the estates (Nieminen, 1999: 12). The early data collection forms are illustrative of how strong social divisions were also in those days. Knights and nobility were included as one group, while burghers, clergy, and land-owning peasants formed a group of their own. Additional groups included those public servants that did not belong to the nobility, and those artisans that were not considered as being part of the burghers. As a consequence of the fixity of the estate system, the estateless population grew fast. A clear division also existed between cities and rural areas. Most trading activities were not allowed outside of the cities, but eventually the government could not prevent it from taking place at the countryside as well (Nieminen, 1999: 14–15).

The reformation of data collection forms in 1802 showed signs of the crumbling of the estate system. From there on, people were no longer divided in to six estates, but instead the new forms included a section on a person’s occupational status. Scholars, store-assistants, journeymen and apprentices were each considered to form a group of their own. With this classificatory scheme, it was for the first time possible to distinguish between the employed and the unemployed population (Nieminen, 1999: 14–15).

3.2.2 Statistics in the Grand Duchy of Finland (1809 – 1917)

The second major turning point in Finland's history occurred as a consequence of Russia's rise to dominance in the Baltic region in the eighteenth century. In this period, Sweden lost its position as a major power in the north and was forced to adapt to a new power balance in Europe. As a result of the 1808 – 1809 war between Sweden and Russia, Sweden lost its eastern part to Russia (Meinander, 2011: 55)

During the Swedish reign, Finland had been a mere group of provinces lacking a sense of national identity, and governed from Stockholm, the capital of the provinces at that time. However, when Finland in 1809 became a part of Russia, it was established as an autonomous Grand Duchy within the Russian empire. The Russian Emperor, Alexander I, became the Grand Duke of Finland, and one of his representatives assumed the position of Governor General (Meinander, 2011: 77). In this period, Finland's highest governing body was a Senate that consisted entirely of Finnish members. Matters concerning Finland were presented directly to the Emperor in St Petersburg without interference from other Russian authorities (Meinander, 2011: 78–79).

By granting Finland extensive autonomy, Alexander I gave Finland the opportunity to develop in to an independent state. In 1812, he decided to move the capital from Turku to Helsinki in order to gain a strategic advantage over the Gulf of Finland. The university which had been founded in Turku in 1640, was also relocated to Helsinki. By making Helsinki the administrative capital of the Grand Duchy, Alexander I hoped to reduced Swedish influences in Finland (Meinander, 2011: 77–81).

The union with Russia gave rise to a nationalist movement in Finland. At the beginning of the nineteenth century, more than 85 percent of the population in Finland spoke various Finnish dialects, but Swedish

remained the language of the administrative elite. In this period, a number of prominent cultural figures, notably Elias Lönnrot (1802 – 1884), J.L. Runeberg (1804 – 1877), and J.V. Snellman (1806 – 1881) became convinced of the intrinsic national spirit of the Finnish language and decided to mould ancient Finnish folk poetry in to a literary whole. The Finnish national epic, *The Kalevala*, written by Elias Lönnrot, was published in 1835 (Meinander, 2011: 87–92).

Russia's defeat in the Crimean War resulted in growing autonomy for the Grand Duchy. In 1863, after a break of more than half a century, the Tsar allowed the Grand Duchy's parliament to assemble. Over the next four decades, the estates met regularly to discuss the language question, the modernisation of society, and the Grand Duchy's constitutional status within the empire. The Conscription Act of 1878 gave the Grand Duchy its own army (Meinander, 2011: 97–100).

The years from 1899 to 1917 are often described as a period of "Russification" and oppression in historical literature. Although it is true that in this period the Russian authorities made many attempts at strengthening their grip on the Grand Duchy, it was also a time period of growing commercial activity and a thriving civic society. Positive social developments freed resources to defend the gradually more outspoken independence within the Russian empire. In 1906, Finland's four estate parliament was replaced by a single-chamber legislature, and the following year its first two hundred members were elected by universal suffrage. Finnish women therefore became first to gain full eligibility to vote in parliamentary elections in Europe (Meinander, 2011: 117–119).

The transition to Russian rule had hardly any noticeable effect in the daily lives of the people, and many governmental practices that had been initiated under the Swedish reign, including the regular compiling of population statistics, were kept in place. In 1812, the area of the so called Old Finland that had previously been part of Russia, was incorporated in

to the Grand Duchy, and as a result, Finland received a considerable Orthodox population. In addition, many Russian officials and soldiers moved to Finland. This population was however never included in the official statistics, since those were based on the Lutheran parish catalogues (Nieminen, n.d.).

As I covered in detail in the previous chapter, the early parts of the nineteenth century was a period of growing enthusiasm towards statistics, eventually leading to the formation of statistical societies and agencies across Europe (Hacking, 1982). The international statistical congress organised every second year provided a forum for debate and collaboration across countries (Westergaard, 1932).

These developments raised an awareness that the Grand Duchy of Finland also needed its own statistical agency. In 1865, during a period of a temporary weakness of Russia, the Russian emperor accepted the senate's request of permission for the founding of a Central Statistical Office (Myllys, 1981: 59). The office's first task was to gather the data scattered around different parts of the government in to one place. As a result, the first Statistical Yearbook of Finland was published in 1870 (Luther, 1993: 55–57).

The late nineteenth century was a period of rapid social change also in peripheries such as Finland. One sign of the changing tides was the gradual crumbling of the estate system. Industrialisation and urbanisation brought with them new occupational groups, and as a result the church found it increasingly difficult to keep a record of its members. Traditionally an individual's occupation had been determined by the occupation of the head of the family, but in the new social circumstances this was no longer necessarily the case. The rapid social changes in the late nineteenth century therefore significantly weakened the reliability of the parish catalogues (Nieminen, 1999: 20).

In a letter sent to the senate in 1869, the Central Statistical Office suggested that due to the unreliability of the parish catalogues, a separate census would need to be conducted in the larger cities (Luther, 1993: 71). Sweden had recently began gathering its population data on individual level forms, and this was in Finland too seen as the solution for the problems that existed with the parish catalogues. Fearing an increase in their work burden, these plans were initially ferociously objected by the Lutheran ministers (Nieminen, 1999: 21)

Despite the objections, population statistics were in 1870 gathered on individual level forms in Helsinki, Turku, Vyborg and Oulu. This was to be the first modern census conducted in Finland. In each city, the census was undertaken by a committee led by the county governor. The committee was formed of people that were familiar with the local conditions, such as the vicar, the mayor and the chief of police. Cities were divided in to separate areas that each had their own calculation office undertaking the actual counting. House owners provided the calculation offices with lists of houses and their inhabitants (Luther, 1993: 71).

One major problem with the parish catalogues was their inability to provide a sufficient level of detail on living conditions. To address this problem, the 1870 censuses gathered information separately in each household and classified them according to neighbourhood and block. Each person was categorised by marital status, mother tongue, literacy rate, religion and occupational status. In addition, information was collected on the blind, the deaf and the mentally disabled. This was also the first time that information was gathered on the number of households, apartments and buildings (Nieminen, 1999: 21).

Preliminary announcements about the purpose of the censuses were unable to alleviate the widespread suspicion that they raised. Especially the poor were eager to leave the city for until the perceived threat was

over. In order to address this issue, the police was commissioned to conduct the information gathering in the poorer neighbourhoods a few weeks after the initial data collection had taken place (Luther, 1993: 71).

Another important step towards national census in the modern sense was the statistics reform in 1877. After this, national population statistics were still compiled from information provided by the parish catalogues, but the accuracy of the information improved somewhat. Births and deaths were for example now reported yearly, which made it possible to calculate accurate life expectancy tables. Statistics on migration were also now reported yearly, unlike previously, where this had been the case only in the larger cities. Whereas previously only those divorces that had resulted from the death of a spouse had been reported, now legal divorces were also included. Classifications on the causes of death were simplified, and only those diseases that had clear symptoms were being reported (Nieminen, 1999: 20).

The nineteenth century was a period of mass migration to America. Initially only people from the western parts of Finland had moved there, but eventually migration started taking place from other parts too. Migration to and from Russia was common as well. The inability to keep statistics on these migration patterns was a big problem since it meant that the actual population was in many municipalities much higher than officially reported (Nieminen, 1999: 23).

Like today, migration was seen as the cause of many social problems. Vaasa and Oulu, the two cities that had experienced the largest amount of out-migration, were the first to start reporting migration numbers. Soon this practice was taken up in other cities and municipalities as well. From 1900 to 1980 migration statistics in Finland were derived from passport catalogues. From early on, migrants were categorised according to sex, age, marital status and occupation (Nieminen, 1999: 24).

In spite of numerous efforts, a national census based on individual level forms was not undertaken in Finland during the nineteenth century. The Russian government had planned to conduct one in mid 1910s, but the plan was postponed when the First World War broke out in 1914 (Nieminen, 1999: 22).

3.2.3 Statistics in Finland after a declaration of independence (1917 – 1945)

In 1917, the Finnish parliament approved a declaration of independence drawn up by the senate. Shortly after, the breach between the left-wing and the right-wing parties became irreconcilable, and the Russian Revolution started to spread to Finland. The revolution in Russia intensified existing social divisions also in Finland and the battle for independence quickly evolved in to a civil war. The short but bitter and bloody war resulted in victory for the centre- and right-wing forces, who had enjoyed the support of German military troops. In 1919, K.J. Ståhlberg was elected president, and Finland became a republic (Meinander, 2011: 125–130).

When Finland gained her independence the work of the Central Statistical Office continued uninterrupted. Soon after the war, the Central Statistical Office returned to its normal routine and was even able to improve on its practice. The war had caused many severe social problems, such as food shortages, that demanded official reports. Research was conducted on, for example, how the civil war had affected the economy and on the living conditions of the war orphans (Luther, 1993: 154–155).

The newly established health and welfare cabinet quickly reinstated research on the living conditions of the working classes. This was seen as a necessary step in order to alleviate the deep social injustices that had led to the civil war. In 1919, the health and welfare cabinet set in

action measures to count all the buildings in the cities. The results were to function as the basis for governmental housing policy. The prohibition law was established in 1919 and brought along with it many new social problems as well. To these the Central Statistical Office responded by conducting research on the assumed links between alcohol consumption and crime (Luther, 1993: 162–163).

As always, also in the early years of independence financial statistics were one of the main areas of concern for the Central Statistical Office. In particular, the Central Statistical Office was commissioned to improve the data on national wealth and personal incomes. This task was made easier in 1921 with the introduction of wealth and income taxes. Data from tax returns were used to calculate how the taxation system could be made more just. Regular economic forecasts started also being conducted as had been requested by the banking authorities (Luther, 1993: 165).

The statistical authorities had long hoped that population statistics would be gathered with individual level forms instead of lists of names collected from the highly unreliable parish catalogues. As I mentioned in the previous section, individual level forms had been used in the larger cities since 1870, but the majority of the population still remained outside of these calculations (Luther, 1993: 177).

In 1923, a committee suggested that each municipality should have its own population register that would be kept by the police chief in the counties and by the register office in the cities. Churches were unenthusiastic about the proposal. On the one hand they did not want to lose their monopoly on population statistics and on the other hand the ministers objected to having their workload increase by having to collect the data on separate forms for each individual (Luther, 1993: 177). Despite these objections, it was gradually possible to increase the number of cities in which a census was conducted. In 1900, a census

was undertaken in four cities, in 1910 in seven cities, in 1920 in ten, and by 1930 in twelve (Luther, 1993: 178).

The Central Statistical Office kept introducing more and more advanced forms of data analysis. In 1920, the first mortality and fertility tables and the first calculation of marital fertility were introduced. In addition to these improvements, the Central Statistical Office published its first population projection in 1932. The projection was only concerned with the capital, Helsinki, but a few years later a projection was conducted for the entire country. It estimated that Finland's population would never exceed four million (Luther, 1993: 178). Finland's population is today close to five and a half million (Statistics Finland, 2017).

Other innovations in official statistics in the years shortly after independence included the marriage law in 1930, which made it possible to obtain individual level data on divorces from court records. Migration statistics had been derived on individual forms from passport catalogues since the beginning of the century, but the problem was that they could not be linked to other population statistics (Nieminen, 1999: 29).

Another important move towards individual level data occurred in 1936, with the introduction of death certificates. Before this, the cause of death had been obtained from parish catalogues, and unsurprisingly the reliability of the statistic improved considerably after the reform. In 1939, birth statistics were also supplied on individual forms by churches (Nieminen, 1999: 29).

There had been many attempts at conducting a national census based on individual forms ever since the formation of the Central Statistical Office in 1865. The need for information on households, living conditions, and buildings became more and more pressing during the early years of independence. International cooperation between national statistical institutes had increased in the early parts of the twentieth century, and

by 1930 Finland and Albania were the only countries in Europe not to have conducted a national census based on individual level data. Although Finland had by then been gathering population statistics for over two hundred years, they had not been based on people filling in their own data (Nieminen, 1999: 31).

In 1938, the Finnish parliament stated a law according to which a national census was to be conducted every decade. This was to be done in direct contact with the people and with one form for each individual. The first national census based on individual level forms was meant to be conducted in 1940, but the plan was set aside when the Second World War broke out in 1939. The census in 1940 was still conducted the highly unreliable parish catalogues (Nieminen, 1999: 31).

3.2.4 Statistics in Finland after the Second World War (1945 – 1960)

Despite conciliatory measures, such as allowing the Social Democrats to participate in parliamentary elections, the wounds sustained in the Civil War would not heal before a national unification in the Second World War. In 1939, the Soviet Union and Germany signed the so called Molotov – Ribbentrop pact, a non-aggression agreement that included a secret protocol relegating Finland to the Soviet sphere of interest. Stalin feared that Hitler would invade northern Soviet Union across Finnish territory, and decided therefore to attack Finland. The "Winter War" (1939 – 1940) ended in a peace treaty in 1940, giving south-eastern parts of Finland to the Soviet Union (Meinander, 2011: 146–149).

When German offensives broke out all over Europe, Finland was left with no other option but to ally itself with Germany. After fighting on the German side against the Soviet, Finland eventually broke its ties with Hitler. The allied leaders had agreed that if Finland committed itself to end its alliance with Germany, it would be allowed to remain a sovereign

nation after the war. The "Continuation War" (1941 – 1944) ended in armistice in 1944. In addition to the areas already lost to the Soviet Union in the Winter War, Finland ceded Petsamo next to the Arctic Ocean, and was forced to lease out Porkkala Peninsula, only thirty kilometres west of Helsinki, as a Soviet military base. In addition, Finland was forced to pay 300 million dollars in war reparations to the Soviet Union (Meinander, 2011: 152–157). Therefore, despite two bloody wars fought against the Soviet Union, Finland was able to maintain its independence through the Second World War.

After many failed attempts at conducting a nationally representative census during the first half of the twentieth century, one was finally successfully conducted in 1950. After the war, the United Nations gave a recommendation on information that should be gathered in national censuses. In addition, Finnish statisticians visited statistical institutes in other Nordic countries to look for examples (Luther, 1993: 242).

The census in 1950 was conducted with the help of accountants who distributed the census forms to the heads of household's three days before the final count. In addition to age, sex and marital status, information was gathered on the place of living, place of birth, language proficiency, religion, nationality, literacy level, education, family structure and occupational status. Whereas previously information had been based on municipalities, this time the information was based on conurbations. The smaller level of granularity of the data made it more usable in town planning (Nieminen, 1999: 32–33).

In connection with the census in 1950, a survey was carried out on the living conditions of the population that had been displaced during the war. The fact that the census now included information on marriages and childbirths made it possible to calculate fertility rates for different age groups (Nieminen, 1999: 33). The war had a big impact on both mortality and fertility, and the population projections that were calculated after the

war differed significantly from the ones conducted in the 1930s (Luther, 1993: 267).

While the results of the first census were still being calculated, a committee was assigned to investigate how official statistics could be improved. In particular, the committee had an interest in improving the accuracy of regional data. The rapid economic and social developments after the war had created a need for municipal level town planning, and data on regional demographic trends were vitally important for its functioning. In 1950, the census had for the first time included information on conurbations, but the committee felt there was still much room for improvement (Nieminen, 1999: 34).

In addition, the committee paid attention to the problems with the existing population statistics. Migration statistics in particular had long been inaccurate and the situation became even worse when the Nordic countries decided to remove the requirement of having a passport when travelling between them. Nevertheless, it was not until the 1960s that the Nordic started to collaborate in keeping statistics on migration (Nieminen, 1999: 34).

The committee had also proposed that population forecasts were to be conducted on a regular basis. At the time the committee's suggestions were considered radical, but the rapid development of computers in the 1960s opened up possibilities that would have been unimaginable ten years earlier. The census in 1960 largely followed the format of previous one. This time however, the results were calculated with a computer, the IBM 1401. From here on, the compilation of most important statistics was no longer dependent on calculations done by hand, as even the most complicated calculations could now be done with computers. By the mid-twentieth century, individual level data was available on all of the most important demographic variables (Nieminen, 1999: 29–35).

3.2.5 Advances in computational techniques, the emergence of the welfare state and the forming of electronic registers (1960 -)

After the Second World War, the sample survey technique had been successfully applied in various fields of research particularly in the US and in India (Luther, 1993: 230). An appreciation of the advantages of sample methods highlighted the need for registers from which samples could be drawn when needed. The development of computers from the late 1950s onwards had made it technically possible to maintain registers on the population, economy and agriculture (Luther, 1993: 260).

The prospect of developing registers was first brought up in the Nordic statistical conference in Helsinki in 1960, where the head of department at the Central Statistical Office in Finland, Olavi Niitamo, suggested that the next step in the development of official statistics would be the production of various kinds of forecasts. According to Niitamo, modern statistics would be able to answer questions such as “what happens if…” and “what should be done in order to…”. A general consensus emerged that statistical offices should place more emphasis on analysis instead of just on compiling the data (Luther, 1993: 260).

During that same conference, the head of the statistical institute in Norway, Svein Nordbotten, suggested a model on how official statistics could be remade in to a statistical archiving system. According to Nordbotten, the old model had involved compiling and publishing statistics only once and then sending them into the archives. Now however, it was technically possible to gather statistics continuously and without necessarily having to form separate entities of data collected at the same period in time. Instead, information on single units (person, family, household, building etc.) could be linked to information on the same unit at a previous period in time. This would be made possible by assigning each unit a unique identification code. Nordbotten’s ideas

quickly gained momentum and soon all Nordic countries were reviewing their practices against the model presented at the conference (Luther, 1993: 260–261).

In addition to the advances in computer technology, another important precondition for the development of population registers was the construction of a comprehensive welfare state. In stark contrast to today, the economic consensus of the post war years posited that income redistribution was in fact beneficial for the overall health of the economy. Between 1950 and 1980, three major social reform packages were passed in Finland. First, a compulsory national insurance scheme was introduced, second, health and social services were expanded, and third, the entire education system was restructured. The period also saw the introduction of income-related pensions for all, the forty-hour work week, and improvements in unemployment benefits (Meinander, 2011: 172–173). In the five decades following the Second World War, Finland turned from a war-ravaged agrarian society into one of the most technologically advanced nations in the world. Between 1948 and 1979, Finland's annual GDP rose faster than that of almost any other West European country, and the average income more than doubled. By the end of the Seventies, Finland had joined the richest third of the European states (Meinander, 2011: 167–168).

The new forms of benefits introduced, such as universal earnings based pension in 1961 and universal sickness insurance in 1963, required a system that would be able to identify individuals unequivocally (Kinnunen, 1998: 118). And since, after intense political battle, the decision was made to extend sickness insurance to all members of society instead of just for workers, also those who resided outside of the workforce, such as students and the elderly, needed a unique identifier (Kinnunen, 1998: 122–123).

The task to develop such a system was handed to the mathematician, Erkki Pale, who had served as a code-breaker in the Second World War. Although no documentary evidence proves that Pale benefitted from his experience of breaking Soviet codes during the war when inventing the logic behind the social security number, his former colleagues have thought this to have been likely. Pale got the initial idea for the number from Sweden, where a personal identity card had been introduced already in the 1940s. In accordance with the Swedish model, the first six numbers in the Finnish social security number consist of a person's birthday, -month, and –year. These numbers are followed by a minus sign for persons born in the 20th century, and a plus sign for those born in the 19th. Following the seven numbers is a three-digit personal number that can be used to distinguish between individuals with the same name born on the same day. For men, this number is odd, and for women, even. In the end, the Finnish model ended up being far more accurate than the Swedish one, as the former was developed at a time when computers could be used to assist in the information processing. Afterwards, Pale's model has been assessed as a nearly perfectly accurate way to identify a person (Kinnunen, 1998: 119–121).

Around the time that the social security number was being introduced, the Ministry of the Interior was planning to introduce a civil registry number for the purpose of identifying individuals in the newly established governmental registers. In order to avoid a situation whereby multiple identification systems would co-exist simultaneously, a decision was made to merge the two numbers. The social security institution KELA was handed the task of assigning an identity number for each citizen and reporting them to other register holders. Between 1964 and 1968, KELA assigned identity numbers to all Finnish citizens, and from there on, it assigned them as people were born or moved in to the country (Kinnunen, 1998: 123–124).

In addition to KELA's register, various other electronic registers were set up by different governmental institutions in the 1960s. For example, the board of education started to keep a register on teachers, the ministry of finance on public servants, and the universities and polytechnics on completed degrees. The Central Statistical Office started a vehicle register in 1965 and the health authorities established one for cancer patients. A central business register was established in 1970 (Luther, 1993: 262).

As a result of the international collaboration between Nordic statistical authorities cited above, the Central Statistical Office was aware of the opportunities offered by governmental registers already before any had been established. Therefore, from very early on, the Central Statistical Office actively sought to gain a foothold in the design of the registers. In the early 1970s, the Central Statistical Office was renamed Statistics Finland and was given a statutory right to both receive data from the various register holders, and to provide them with advice on how the information in the registers should be gathered and maintained. Gradually, as registers became more standardised, Statistics Finland lost its legislative right to influence the design and maintenance of the registers (Alastalo, 2009b: 178–179). According to the most recent statistics act adopted in 2004, state authorities have a statutory obligation to supply data from their possession to Statistics Finland. Furthermore, Statistics Finland can use direct information gathering only when the information is not already available in the registers (Statistics Finland, 2015). This means that in Finland, governmental practices have a major influence on the content of official statistics (Alastalo, 2009b).

As I have documented in the previous sections, population statistics were in Finland for centuries maintained by two separate institutions, the government and the Lutheran church. A solution to the problems that emerged from having information stored in two places was presented in 1969 with the introduction of the Population Register Centre. Since then,

official population statistics have been held in one electronic register, The Population Information System, which includes all individuals living permanently in Finland. In addition to all citizens in Finland, the register includes details on all buildings, their owners, and the people living in them (Population Register Centre, 2015). All Nordic countries established their own Population Register Centre's between 1964 and 1969 (Alastalo, 2009a). In 1974, the task of assigning identity numbers to new citizen's was handed over from KELA to the Population Register Centre (Kinnunen, 1998: 124). The digitisation of government and church records was far from a straightforward task, and required a decade of meticulous work (Alastalo, 2009b: 181). It is interesting to note therefore that in contrast to general perception, the government at least in Finland has in fact been a forerunner in taking advantage of the affordances of digitisation.

The Evangelical Lutheran church was involved in the upkeep of population statistics in Finland all the way until 1999. Since then, the maintenance of the Population Information system has been the responsibility of the Population Register Centre and the Registry Office. The Population Information System is updated through statutory citizen notifications and has been used in elections, taxation, the justice system, administrative planning and in research. Many third- and private sector organisations are also regularly granted permission to access some of the information in the system (Population Register Centre, 2015).

The social security number was first used in the 1970 census. The number of each member in a household was pre-entered in to a form that was then sent to the household to be filled. The social security number was also used to retrieve information from other registers regarding a person's religion and place of birth. Compared to previous census forms, the form in 1970 no longer included a question on occupational status, but instead asked for a person's main source of income, and whether that came from pension or employment. The

census also included a new scale for socio-economic status. In connection with the census, two cohort studies were undertaken, one on fertility and another on the demographic characteristics of some minority groups (Marjomaa, 2000: 275–276; Nieminen, 1999: 41).

The census in 1980 was to a large extent identical with the previous one. However, since the reliability of the Population Information System had increased year by year, by 1980, the census was no longer the only way one could discover the size of the population. Instead, it now functioned more like a large-scale social scientific study that was able to provide a more detailed picture of society than what would have been possible by relying solely on information from the Population Information System (Marjomaa, 2000: 276; Nieminen, 1999: 42).

The main data gathering was still conducted by sending out forms, but many electronic registers were also used. Educational data was retrieved from the degree register maintained by Statistics Finland and information on incomes was retrieved from the tax register. In addition, the social security institution, KELA, provided information on people on benefits (Marjomaa, 2000: 270; Nieminen, 1999: 42).

By 1990, the electronic registers had become so comprehensive and reliable that the census could for the first time be conducted entirely without census forms. Basic information such as age, sex, gender, marital status and place of birth had since 1970 been available in the Population Information System, but forms had been needed to discover people's occupational status (Marjomaa, 2000: 279; Nieminen, 1999: 43).

In addition to the Population Information System, the following registers were used in the 1990 census: the Tax Administration's registers, the Central Pension Security Institute's employment registers, the State Treasury employment register and Municipal Pension Institute's

employment register, the Social Security Institute's register on pensioners, different student registers, the register on job applicants, Statistics Finland's Business register and register of Completed Education and Degrees (Myrskylä, 2011).

Finland was only the second country after Denmark to have its census entirely register based. The total cost of the census in 1990 was one tenth of the fifty million euros (in today's money) paid for the census in 1980. The low cost of the register-based census has made it possible to conduct one every year, and today, the annual cost of data collection for the census is around one million euros (Myrskylä, 2010). Today, a total of 185 registers are maintained by 16 different register holders (Alastalo, 2009b: 173–174). Up to 95 % of official statistics produced by Statistics Finland are derived from information held in the registers (Marjomaa, 2000).

3.2.5.1 A tool of welfare state governmentality?

In the previous chapter, I argued that historically, statistics have emerged out of a co-constitutive interaction between, on the one hand, methodological and technological developments, and changes in the administrative and political world, on the other (Desrosières, 1998). This analytical framework can be fruitfully applied to the emergence of register-based statistics as well. As Alastalo (2009a) explains, register-based statistics emerged as part of the technologies and arrangements of the welfare state, and should therefore be understood as a form of welfare state governmentality. As a result, constructing a similar system elsewhere would constitute far more than a mere technical task, as many countries have recently begun to find out. Whereas technically it might be feasible, the social and political circumstances that make it possible to put in place an extensive governmental data collection system are more difficult to establish.

Kinnunen (1998: 124) describes the social security number on which registers rely as the outcome of “a union between the welfare state, mathematics and technology”. According to her, the development of the social security number was inherently connected to the question of who, in Finland in the 1960s, had the power to determine the structure of social security. Far from being a mere technical tool therefore, the social security number emerged out of a political process whereby a welfare state was established on universalistic principles. After its introduction, all citizens, men and women, children and the elderly, have been registered in governmental systems in Finland first and foremost as individuals. This practice, which seems mundane at first, is not devoid of political consequences.

Kinnunen (1998: 126–127) sees it for example as being tied to the question of how Finns have come to view women’s role in society. For long, Nordic researchers have had to explain to foreign colleagues why it is not sufficient to classify women’s socio-economic status based on their husband’s status. In Finland, where women’s labour market participation rate is equal to men’s, a return to classifications based on households would constitute a step back in time (Kinnunen, 1998: 126–127). According to Kinnunen (1998: 131), the social security number produces a modern notion of individuality, a way to understand oneself and one’s relation to others. As Totto explains (1989: 17–18) the social security number indicates that a person is no longer a nameless member of a family, tribe, profession or caste, but an equal citizen with the same rights and responsibilities as everyone else.

On the other hand, the social security number is also a powerful tool with an element of coercion built in to it. In terms of gender for example, one cannot choose outside of the traditional dichotomy between male and female, but must be one or the other at all times. Ultimately, for Kinnunen (1998: 131), the politics of the social security number boil down to the question of who has the power to determine how and for what purposes

the number is being used, and what rights remain for the individual to contest these uses.

3.2.5.2 An example of register-based research: Mortality

An important part of the welfare state order that emerged after the Second World War was in retrospect the enormous public status enjoyed by social scientists. The writings of leading sociologists, such as Erik Allardt, were carefully read even at the very top of the political establishment (Eskola, 1993). Although perhaps not the most important reason behind the success of social scientists in this period, the possibility of combining data from many different registers opened up previously unimaginable opportunities for quantitative social researchers in Finland.

The following topics have for example been examined with register data in Finland: Socioeconomic differences in cause-specific mortality; the consequences of alcohol tax changes on alcohol-related harm; ageing, long-term care use & end-of-life care; social determinants of crime; partner choice; determinants and outcomes of union dissolution of cohabitations and marriages; determinants of teenage pregnancies and pregnancy outcomes; socio-demographic determinants of fertility behaviour and enterprise demography and job flows (Ilmakunnas et al., 1999; Väisänen, 2013).

Perhaps the most impressive register based dataset so far has been the EKSY (acronym of the Finnish project name, which translates as ‘Living conditions and causes of death’) dataset used mainly by demographers at the University of Helsinki. Since the late seventies, demographers at the University of Helsinki have in collaboration with Statistics Finland and the Institute of Occupational health used the personal identity code to link mortality data to census, and other register data. For example, in the first data linkage in 1978, data on all deaths (approximately 250,000)

during 1971 – 75 was linked with the 1970 census, which covered the total population of 4.6 million at that time. Since the initial dataset, a new one has been compiled every five years. Whereas the initial dataset included information on only 50 variables, the dataset in 1990 covered a total of 245 variables (Valkonen and Martelin, 1999: 213).

The EKSY dataset has allowed both descriptive and explanatory studies on variations in mortality. In terms of descriptive studies, mortality differences have been examined by social group, education, region, marital status and mother tongue. Although methodologically these studies have not offered much new, the availability of data on individuals has allowed for a much greater freedom in forming different combinations of population groups than what has been previously possible. The number of cells in the tables can be massive, sometimes in excess of 100,000. In such a big dataset, all differences that are of relevance to the conclusions are usually statistically significant. By contrast however, not nearly all significant differences have any relevance to the conclusions (Valkonen and Martelin, 1999: 216).

The EKSY dataset has for example been used to test hypotheses concerning factors influencing mortality. Traditionally studies designed to test hypotheses have faced the methodological problem of selection bias. In previous studies, it has been difficult to determine whether it is unemployment itself, and not the personal characteristics of the unemployed, that affects mortality. Because register-based datasets are so large and comprehensive, they can be used to identify even relatively small groups that have been exposed to the factors that the researcher is interested in, without having their personal characteristics influence their exposure. For instance, a study conducted on the relationship between unemployment and mortality in the deep recession years of the early nineties did not lend support to the assumption that unemployment had a clear impact on mortality, at least in the short term. This is an example of how register data has helped to elaborate on an observed

difference between sub-groups in the population (Valkonen and Martelin, 1999: 221).

3.2.5.3 An example of register based research: GIS

Another interesting example of the application of registers comes from the field of Geographic Information Systems, or GIS. In Finland, a national register on buildings and dwellings was established in conjunction with the census in 1980. Data on dwellings and buildings were collected with questionnaires and saved in the Building and Dwelling register maintained by the Population Register Centre. Since then, the register has been updated with statutory notifications from municipal building authorities and local register offices. Register-based statistics on buildings and dwellings have been produced annually since 1987 (Harala and Tammilehto-Luode, 1999: 60–61).

What this means in practice is that nearly all individuals in Finland can be linked into families and household-dwelling units, and to the dwelling or building in which they live. All individuals can also be linked with their employers and to the building in which they work. In addition, all of these units can be located on maps using geo-coordinates. Since 1970, Statistics Finland has provided data in 1 km x 1 km grid squares. These spatially stable units can easily be combined into larger areas (Harala and Tammilehto-Luode, 1999: 62–65).

Since 1992, the GIS-unit at Statistics Finland has provided spatially referenced data to public administrations, the private sector, and academic researchers. Applications of grid square data in social research have included, for example, studies on the regional incidence of diseases, unemployment, changes in rural industrial structure, regional polarisation, and migration (Harala and Tammilehto-Luode, 1999).

3.2.5.4 Advantages and disadvantages of registers

Population registers are often seen as having a variety of advantages in comparison to sample surveys. First, they provide longitudinal data sets with massive sample sizes and many different sampling strategies. In addition, the level of granularity is higher than with samples, as is best demonstrated by small area statistics. Unlike sample surveys, registers do not suffer from low response rates (UNECE, 2007) and with registers, the time spent on data collection can be reduced significantly, and studies can be easily repeated (Gissler, 1999: 245). Furthermore, registers can be used to correct non-response bias in sample-surveys (Lehtonen and Veijanen, 1999). Other standard praises for registers include that they are relatively cheap, they reduce the burden on respondents, they often cover entire populations, and that they are automatically gathered in a machine readable format (Myrskylä, 2010).

On the other hand, the main disadvantage with registers is the researcher's inability to exercise control over content. By relying solely on register data, the academic researcher or official statistician runs the risk of letting governmental practices decide what aspects of the social world can be studied and known (UNECE, 2007).

Overall, in contrast to sample surveys, there have not been many critical examinations of register based statistics. One of the few to have done so, Alastalo (2009b: 174), calls for more critical scrutiny upon the process whereby "information produced in local administrative practices is transformed in to statistical facts". She points out that registers are produced and maintained in local administrative practices, and that changes in them are recorded only when individuals come to interact with the public administration. She highlights four problematic attributes of register based statistics (Alastalo, 2009a: 59–60).

First, registers contain information that has been collected first and foremost for administrative purposes. Hence, when registers are used for official statistics, not nearly all the information is included. The practices and processes according to which information is chosen and then made into official statistics usually remains invisible both for the users and producers of official statistics. For example, the mathematical rules for handling conflicting information in different registers so that they can be combined remain largely black boxed (Alastalo, 2009a: 60).

Second, registers are susceptible to changes in administrative practices and legislation. For example, when the Finnish Tax Administration began using pre-filled tax return forms, it meant that information on job titles would no longer be available, and Statistics Finland had to start looking for alternative sources for the information. Similarly, as a result of the abolishment of the wealth tax in 2006, it became nearly impossible to describe and study changing wealth dynamics within the population (Alastalo, 2009a: 61).

Third, registers are sensitive to local recording practices, which are also poorly documented, and fourth, registers are not necessarily updated when an individual's life circumstances change, unless he or she is in contact with the register holders. As register based statistics have become mundane, their problematic aspects have been largely ignored (Alastalo, 2009a: 61).

3.2.5.5 Privacy concerns

Especially since register data can be mapped so precisely onto space, identification of units has become a crucial issue. Whereas a social researcher might be content with data that is statistically significant at a larger geographical level, other users, such as town planners, sometimes need to know the exact location of a phenomenon. Consequently, data anonymity has presented a considerable challenge

to statistical authorities, in particular to those dealing with geographically referenced data (Harala and Tammilehto-Luode, 1999: 67).

The data protection rules that apply to Finnish official statistics are prescribed in the Statistics Act, the Personal Data Act and the EU Regulation on Community Statistics. According to the Statistics Act, state authorities have a statutory obligation to supply data for statistical production, whereas enterprises, municipal organisations and non-profit institutions are obliged to supply data on matters separately prescribed. Before the data can be used for research purposes, it must to be anonymised. The data must also be made less detailed in order to prevent indirect identification. In 2011, around 200 sets of register data, relating mainly to persons and housing, were supplied to researchers working outside of the statistical office (Statistics Finland, 2013: 3)

Decisions on the provision of statistical data sets for research purposes are made by the Directors of the respective statistics departments or by the Director of the department of Standards and Methods. In special cases, such as in data requests from abroad, the requests are considered by the Ethics Committee of Statistics Finland (Statistics Finland, 2013: 3).

Perhaps surprisingly, the public opinion towards population registers remains positive, and no severe data protection violations have been reported. However, one must question how many Finns are aware of the granularity of the data that their government possesses. Of some significance must also be the fact that register based statistics emerged as part of a governmental system whose function was to distribute benefits to the population. The social state that developed after the Second World War guaranteed a high level of trust towards public administrations in all Nordic countries (Alastalo, 2009b: 182). Alastalo (2009a: 183) points out that because the register-based system is dependent on a high level of trust between citizens and the state, it is

not inconceivable that this trust could be jeopardised as a result of political changes in such a way as to make the use of data from registers for official statistics difficult. In the next section I review methodological developments in social research in Finland.

3.3 A note on methodological developments in social research in Finland

Sociology has been an academic subject in Finland for twice as long as in the other Nordic countries. Already in 1890, Edward Westermarck was named Associate Professor in Sociology at the University of Helsinki. Westermarck was also the first Professor of Sociology at the London School of Economics, a factor which helped Sociology to become accepted as an independent academic subject in Finland as well. Between the two World Wars Finnish social science was dominated by Westermarck's school of Evolutionary Sociology (Allardt et al., 1993: 13–14).

In his studies, Westermarck often looked for the origins of social institutions in human biology, and today, his work might be considered social or cultural anthropology. In 1890, Westermarck published *The History of Human Marriage,* in which he argued against the common belief that promiscuity had been the initial state of human marriage. Westermarck argued that monogamy had been the natural order of things later verified by law. He also fiercely attacked Freud over the idea of the Oedipus complex. According to Westermarck, people who live in close domestic proximity during the first few years of their lives naturally become desensitized to later sexual attractions (Allardt et al., 1993: 44–49).

In his research methodology, Westermarck emphasised the importance of cross-cultural comparisons. Westermarck's source materials were the

descriptions of life in primitive tribes written by travellers and missionaries, which he cross-examined in order to find universal social norms and institutions. According to Westermarck, it was not sufficient to use the comparative method only to categorise social norms and institutions in different societies in different time periods, but one had to search for an answer to the question of why the norms existed in the first place. For Westemarck, the key question was how to best explain the existence of similar norms and institutions across different cultures and the answer was often found in human biology and human emotions. Westermarck and his students advocated strong empiricism and rejected purely theoretical speculations (Allardt et al., 1993: 52–53).

Around the same time with Westermarck's branch of Evolutionary Sociology, another tradition of social research was emerging outside of academia. As I documented in more detail in the previous chapter, in the nineteenth century industrialisation and urbanisation raised an interest towards the living conditions of the working classes especially in England, where the industrial revolution first began (Kent, 1981). In Finland, studies in to the living conditions of the urban and rural poor began to emerge from the 1860s onwards. The founding of The Central Statistical Office in 1865 contributed to the emergence of a tradition of empirical social research in Finland (Alapuro and Alestalo, 1993: 78).

Already in 1848, the philosopher, statesman, and nationalist J.V Snellman had shared his concern about the growing number of poor people in the countryside. Resembling contemporary debates, Snellman had been particularly worried about what would happen to the poorest people in a situation where agriculture would no longer be able to employ as many people as it had done before. Snellman concluded that the eventual disappearing of agriculture would mean that a significant part of the population would have no other choice but to emigrate abroad (Luther, 1993: 91).

For Snellman's relief, wood processing turned out to be an industry that would replace the jobs lost in agriculture. Free trade was established in 1868, and as a result, other industries had started to grow as well. Although industrial capitalism saved many Finns from starvation, it brought along with it many new social problems as well. And irrespective of the new jobs in industrialised cities, the large estateless population in the countryside still formed a major social issue (Waris, 1932).

The first empirical social studies in Finland were often conducted by individuals who were also active members of the newly formed political parties. Inspired by the German *Verein für Socialpolitik*, an organisation that had been established to conduct research and to promote social security legislation, the Economic Society was established in 1891. Its members were pioneers in statistical descriptions of social problems. In 1915, Edvard Gylling published a thorough analysis of population growth in the recently industrialised capital, Helsinki. O.K. Kilpi wrote an analysis of the development of different social classes from 1815 to 1875 based on information derived from the parish catalogues and Eino Kuusi examined seasonal variations in unemployment and considered ways to reduce them (Alapuro and Alestalo, 1993: 81–91; Luther, 1993: 91).

This was also a time period in which many non-governmental organisations started conducting their own research on social conditions in the country. The Worker's Association was established in 1886, and since early on, one of its main tasks was to keep a list of the salaries of its members. In 1888, August Hjelt began a study on the condition of the working classes in Helsinki, which Oskar Groundstroem then led to a successful completion in the 1890s. Another member of the Worker's Association was the medical doctor Wilhelm Sucksdorff, who completed a study on household conditions in the working class districts of Helsinki in 1904. The Worker's Association kept publishing research, data and graphs in its monthly publications throughout the twentieth century. The Women's Association that had been established in 1844 kept a

systematic record on women's salaries and pension benefits (Luther, 1993: 92).

The estates were also interested in the social problems experienced by the urban and rural poor. In 1888, the parliament appointed a committee to gather social statistics in order for them to be used in planning health-, accident- and old age insurance's for the working classes. As a result, various forms of social insurance were established in the early part of the twentieth century. Consequently, the government became increasingly aware of the importance of having accurate social statistics at its disposal (Luther, 1993: 92–94).

The tradition of empirical social research eventually developed in to more modern forms of social policy research, often conducted within the universities. When social research established itself as a serious field within academia after the Second World War, it became difficult to draw clear distinctions between social policy, social history and sociological research (Alapuro and Alestalo, 1993; Haatanen, 1993).

The strong emphasis towards empiricism that was present both in Westermarck's Evolutionary Sociology and in the tradition of empirical social research was a contributing factor when Finnish sociology became heavily influenced by American positivism after the Second World War (Rahkonen, 1995: 13). One reason for this development was the fact that, after the war, American universities had started to offer scholarships to talented young researchers in Finland. After the First World War, Finland had been the only country in Europe to repay its debt to the US and to reward this, the US set up a fund for exchange programs for promising young Finnish scholars. In the decades that followed the Second World War, nearly half of all the professors in the social sciences at the University of Helsinki had at some point in their career spent a year at an American university (Sinnemäki, 2005: 53).

The first social scientist to visit the US was Heikki Waris, who spent the academic year 1934 – 1935 at the University of Chicago, where the methodological currencies had started to shift towards a strong emphasis on the importance quantitative methods (Bulmer, 1986; Mäkelä, 1996: 149). After the Second World War, American sociologists were leading the way in applying statistical techniques to researching social questions. In Finland, sampling techniques had been used in forestry research already in the 1920s, but social researchers were far behind. As a result, when Waris in 1934 used the correlation coefficient in his study on the working classes in Helsinki, he was seen as a forerunner (Luther, 1993: 174). Being a powerful figure in Finnish academia, Waris' emphasis on the importance of rigorous application of statistical techniques had a major impact on the methodological currencies in post-war social sciences in Finland (Mäkelä, 1996: 154).

Hence, after the war, the scientific mode of inquiry became the methodological ideal also for the social sciences. At first, social surveys and quantitative methods were used alongside other source materials and methods, but by 1960, the large majority of doctoral theses at the department of sociology were based on quantitative analysis of survey data. In particular, the factor analysis technique was applied in nearly every thesis (Alastalo, 2005: 73–80).

In the 1960s, Sociology became a popular subject in Finnish academia. The leading figure of Finnish social science at the time, Erik Allardt, enjoyed popularity even at the very top of the political establishment. In a Durkheimian fashion, Allardt applied a variety of research methods in his attempts to demonstrate that sociology did indeed have its own unique research area that set it apart from all the other sciences (Eskola, 1993: 267–273). Allardt utilised both official statistics and other research materials in his studies (Luther, 1993: 255).

The consensus regarding the use of quantitative methods started to

crumble in the 1960s. The Frankfurt school had become influential globally, and through its influence arrived a Marxist paradigm that challenged the credibility of the branch of American sociology that had long been advocated by prominent figures such as Allardt. The 1960s was a time when the epistemological assumptions and the ideological linkages of social science were heavily debated both globally and in Finland. As a result, a strong shift towards an emphasis on qualitative methods occurred in the 1970s and 80s. During the 1990s, only one third of the doctoral theses published at the department of sociology in Helsinki included any form of quantitative analysis whatsoever (Alastalo, 2005: 88–103). Erik Allardt has later noted that when the most heated methodological debate was over, sociological research emerged more pluralistic and richer than what it had been before (Allardt, 1994: 161).

It is indeed fair to say that Finnish sociology is today methodologically pluralistic. The availability of register data since the 1990s has guaranteed that Finnish quantitative research has been leading edge even globally. Collaborations between qualitatively oriented sociologists and quantitatively oriented GIS-experts from the geography department have led to comprehensive studies on large issues such as the increasing regional polarisation within the capital region. Following an international trend, Science and Technology studies have also gained a foothold in the Finnish social sciences.

Despite these examples of a shift towards methodological pluralism old tensions between quantitative and qualitative methods still exist beneath the surface. Some senior figures have for example recently argued that some of the big names within the field of Science and Technology studies advocate such an extreme form of social constructivism that it no longer advances the development of the social sciences. According to these commentators, Sociology should instead re-engage with its past as a subject that tries to find the origin of social institutions in human biology (Roos, 2011). Similar arguments have been made against

feminist criminology (Kivivuori, 2012).

Despite the fact that sociologists and demographers today work in the same department in Helsinki, cooperation has often proved difficult. One senior figure within the department of Sociology explained the situation as a result of the fact that “demographers can count, but they cannot think” and that for demographers “that which cannot be counted, cannot be spoken of” (anonymous, personal communication, 2014). In interesting ways therefore, the methodological debates of today bear a resemblance to debates that have been going on for at least a century. Furthermore, in relation to the history of official statistics it is interesting to note that although the formation of the Central Statistical Office in 1865 played some role in the emergence of empirical social research, methodological currencies in Finnish social science seem to have been far more influenced by internal debates than advances in research technologies or official statistics.

3.4 Conclusion

In this chapter, I have reviewed the development of official statistics in Finland, from the parish catalogues in the Kingdom of Sweden to the modern administrative registers in Finland as an independent republic. The purpose of the review was to ground my understanding of ongoing changes in the production of data for official statistics, which I will attend to in the empirical part of the thesis, within historical reconfigurations of the co-constitution of statistics. With that aim in mind, I draw three major analytical conclusions from the chapter. First, the analysis indicates that historically statistics have been produced from data produced by the dominant institution of its time. Whereas the earliest forms of statistics were compiled by the ruler and the church, their modern form is tied to the centralised authority of the nation state. In light of this we should perhaps not be surprised by the increasing centrality of private

corporations in the generation, harvesting and analysis of Big Data, which is a theme I will attend to in the next chapter. Second, the increasing centrality of private actors in the production of data suggests a historical break from its centralised production by states. Although data production in the private sector has long historical roots, it is not until recently that states have considered this data a viable source for official statistics (NSIs have previously collected data from companies regarding their own operations, but not their customers). Finally, a prerequisite for the extensive governmental production of data in Finland has been the high level of trust of citizens towards the state and it is not inconceivable that this could in the future change in such a way as to put under scrutiny the state's role in collecting and managing information on the population.

Continuing with the understanding that statistics evolve in interaction with political contingencies, in the next chapter I situate the evermore production of data in the private sector in the context of contemporary forms of advanced liberalism. Whereas data production in the public sector in Finland particularly in the period of the welfare state has been underpinned by social welfarist concerns, aims and objectives, what type of a rationality is driving the increasing production of data in the private sector?

4 Big Data: A harbinger of utopia, but a utopia for who?

4.1 Introduction. Myth or revolution?

According to Mosco (2004: 29) myths are neither true or false, but rather living or dead. Understanding them therefore requires more than proving them to be false. Instead, it requires figuring out why they exist, why they are so important to people, what they mean, and what they tell about people's hopes and dreams. For Mosco, myths are stories that lift us out of the banality of everyday life into the possibility of the "sublime", and they exist to help us deal with the inevitable and often unresolvable contradictions in life. Whereas in pre-modern society myths were typically embodied in religion and nature, today we usually find them within the realm of information technology.

Indeed, new technologies have always been surrounded by myths. When first introduced, the telegraph, electrification, the telephone, radio, and television were all accompanied by claims that they would bring about the end of history, geography, and politics. Gradually however, as each of them became normal parts of everyday life, they stopped inspiring grand visions of social change. Ironically, it was at this very point that they became truly influential (Mosco, 2004: 2).

Not so long ago, it was the internet's turn to be celebrated as the technology that would bring about a revolutionary transformation in society. Especially in the late nineties and early two-thousands, many commentators, some of which were popular while others were more academic, were convinced that the internet would quickly bring about a new type of society. In the so called Information Age, communication technology would be available to everyone at a low price. Where

previously people had worked with their hands, in the new society they would work mainly with their heads. Where previously a persons' choice of community had been limited by the accident of birth, in the Information Age it would be entirely open to choice, renewal and change. Since everything that had come before the Information Age was prehistory, there was no need to place it in a larger historical context, or so the argument often went (Mosco, 2004: 35).

Little more than a decade since the biggest hype took place, the internet has already broken most of its utopian promises. Instead of liberating individuals and bringing about true global democracy, we have recently been shocked by revelations that it is being used as a means of mass surveillance by the world's most powerful government (Lyon, 2014). Totalitarian governments in eastern parts of the world, on the other hand, have demonstrated that contrary to often heard claims, the internet can indeed be successfully censored. The internet has not thrown off dictatorships, nor has it brought democracy to centrally controlled states (See for example Murthy, 2013).

Perhaps even more disappointing is the realisation that the internet has not solved the issue of ever-increasing social and economic inequality in our societies. Instead, in the period of the internet's existence, income inequality has accelerated towards pre Second World War levels in many western countries (Piketty, 2014). Particularly in the centres of the digital economy, the labour market has differentiated into a steep polarisation between a small group of people working in the highly paid information industries, and a large pool of people working in poorly paid service jobs (Florida, 2012; Saxenian, 2014).

Therefore, instead of claiming that information technology turns society into a utopia, a better way to approach them would be to say they always map on to, and sometimes even intensify, pre-existing social structures and inequalities (For an analysis along these lines, see Halford and

Savage, 2010). Highly uneven distributions of economic capital create deep wounds between people regardless of whether the disparities are created by industrial or digital capitalism.

And as the internet has become a mundane part of everyday life in Western polities, other mythical technologies have emerged to take its place. A recent all-encompassing technological myth, and one that is already being replaced by other concepts, is Big Data (Couldry, 2013). In the popular discourse around Big Data it is no longer the internet itself, but rather the information deluge made possible by it that is going to bring about "a revolution that will transform how we live, work and think" (Mayer-Schönberger and Cukier, 2013). For some of its advocates, Big Data marks the moment when "the information society finally fulfils its promise" (Mayer-Schönberger and Cukier, 2013: 190), whereas for others it promises to bring about the end of theory, and therefore of science as we know it (Anderson, 2008). As I have already mentioned, these revolutionary proclamations must be understood in the context of a much longer tradition of utopian thinking around societal changes brought about by new information technologies.

In light of the discussion about myths and information technology above, are there any reasons to take the recent hype about Big Data seriously? In public discourse, Big Data has been presented as a solution to the very persistent social and economic problems that exist in societies, and often accompanying such claims has been the notion that the digitisation of society leaves no room for petty politics. For example, the digitalisation of public services is one of the main goals of the right-wing government elected in Finland in 2015. According to its "Vision 2025", "…Finland has by then made a productivity leap in public services and the private sector by grasping the opportunities offered by digitalisation, dismantling unnecessary regulation and cutting red tape" (Ministry of Finance, 2015). Attached to this section of the government program is a report "Let's take a digileap!" by former politician and current Microsoft

advisor Mikael Jungner. In addition to celebrating the many universally shared benefits that digitalisation supposedly will bring about, the report, commissioned by the Confederation of Finnish Industries, warns about the dangers of digitalisation becoming “subject to political passions”. According to Jungner, who curiously enough only recently used to be a key politician in the Social Democratic party in Finland, “A central feature of digitalisation is outsourcing”. He thereby demands that “digitalisation must not be subjected to any kind of political passions”, or otherwise “it might be brought to a halt in Finland and in Europe”. Instead, he encourages us to unanimously embrace the opportunities brought about by digitalisation, a task that will, if carried out successfully “rescue the Finnish welfare state, our way of living and being” (Jungner, 2015). If myths are, as argued by Mosco (2004: 30) “depoliticized speech”, this surely is a great example of them.

However, a sociologist must be wary of simple technological determinism in which a technological development, such as digitalisation, is imagined as if it dropped from the sky fully formed and then exerted effects on society from the outside. Decades of scholarship especially in the field of Science and Technology Studies reminds us that technology is in fact deeply embedded in and shaped by social processes and choices (MacKenzie and Wajcman, 1999). These processes and choices are also far from innocent when the ways in which technology then exerts its impact on society is considered. What I want to do in this chapter therefore is to consider the political and economic purposes that Big Data is currently being mobilized for. By doing so, I ground my understanding of Big Data in an appreciation of not only its technical qualities, but also in a consideration of the social, political and economic circumstances in which much of it is being generated.

In contrast to Jungner’s notion that digitalisation leaves no room for politics, social theory teaches the exact opposite. Barry (2001: 2) for

example argues that political analysis can no longer be confined to the study of political institutions and identities alone. Instead, he suggests that we speak of the government of a technological society, by which he means that technology dominates both the sense of the kinds of problems that government and politics should address, and the possible solutions to them. For him, technology is political not just as an instrument to be used in political battles, but because it is fundamentally tied up with what it means to be human, and with how social institutions function. As a result, ideological battles over the social order often involve efforts to contest the development and deployment of technology as well (Barry, 2001: 8–9). In a world characterised by widespread apathy towards parliamentary politics, technology, in essence, is politics.

In midst of the recent hype around Big Data, it is worth reminding that sociologists have been debating the societal impact of information technologies at least since the 1970s. Computerisation, automation, artificial intelligence, the internet, social media and Big Data all fall under the rubric of information society theory, a theme widely debated over many decades now. I therefore start this chapter by situating Big Data in this debate. Rather than a radical rapture from the past, I argue that Big Data should in fact be seen as a more recent chapter in a much longer development.

This problematisation points me further towards an investigation into the political and economic circumstances in which the computerisation of society was kicked into motion in the 1970s. Although technological progress, and progress more generally, therefore, is today usually associated with the policies of the economic right, this was not always so. In fact, as I will demonstrate, the wide scale introduction of computers, the enablers of Big Data, was strongly associated with a larger ideological shift in what kinds of ideas would dominate societies. The ideologues on the economic right understood from very early on that for free market policies to seem modern, they had to be seen as an ally

of the most modern of technologies, the computers. In sum, I argue that computers sided market liberalism with modernity, which is a crucially important point to keep in mind when I move on to consider the purposes that much of Big Data is currently being mobilized for.

Once that investigation is complete, I turn my focus to Big Data. By critically examining widely circulated popular accounts of Big Data, in particular Mayer-Schönberger and Cukier's, *Big Data: A Revolution That Will Transform How We Live, Work and Think* (2013), I suggest that one of the functions of such work has been to provide a fresh face for the claim that information technology transcends history, geography, and politics as they have come to be known. Next, I explore academic reactions to Big Data. I argue that in celebrating the opportunities offered by Big Data, computational social scientists in particular have tended to fail to account for the fact that much of the data that they use for their research is also now absolutely central to the functioning of modern capitalism. I argue that any reflections on the practical, methodological and ethical issues around Big Data cannot go far without accounting for this fact.

Following from this, I use the final part of the chapter to situate Big Data in the context of political economy. This investigation reveals that behind the many myths surrounding Big Data still exists a material reality of a capitalist mode of production ultimately not that different from its industrial predecessor. In sum, drawing on the conceptual starting point of this thesis, that statistics evolve in interaction with political contingencies, I suggest that increasingly, the production of data comes to be underpinned by a market and profit rationality, rather than the welfarist concerns that underpinned much of the production of data in the period of welfare states.

Popular discourses on Big Data necessarily come to influence also how NSIs conceive of the phenomenon and of its significance for their

practice. For this reason as too, the analysis in this chapter is absolutely central in order to make sense of the empirical findings of my thesis, which I will turn to in the next chapter.

4.2 The making of a Big Data world

4.2.1 Do we live in an information society or in capitalism?

Many influential commentators have recently argued that we now live in an information society. Broadly speaking, the information society is often described as the successor to the industrial society. For its advocates, it is a society in which computer technology has facilitated a new social order, one in which information has replaced labour and fixed machinery as the central organizing principle of society. In these visions, computer technology has had a profound impact on not just the economy, but on all aspects of life.

In midst of current debates around imminent social transformations that will be brought about by the latest wave of information technologies, it is important to remember that the information society debate has been ongoing at least since the 1970s. Already back then, scholars were debating familiar sounding issues, such as “the end of work”, “the advent of the leisure society”, and “fully automated factories”. These themes then went briefly out of fashion only to reappear in the 1990s with the advent of the internet (Webster, 2014: 2–3) . More recently, many have enthusiastically assigned socially transformative powers to concepts such as Social Media and Big Data, and perhaps even more recently Artificial Intelligence and Machine Learning.

Webster (2014: 8–9) places information society theories along a spectrum, where at the one end are those that endorse the view that a significant shift from the past has occurred, and on the other end those

that emphasise continuity between past and present social structures. In addition, he identifies five different criteria by which information society theorists have tried to define the concept: Technological, economic, occupational, spatial and cultural (Webster, 2014: 10).

Technological accounts of the information society focus on the many technical innovations that have taken place since the 1970s. Authors such as Toffler (1980), Negroponte (1995) and more recently Shirky (2008), just to name a few in a very long list, have each on different occasions argued that recent technological innovation has in fact been so profound that it has essentially reconstituted the social world. These accounts are often criticised for being technologically deterministic, that is, they imagine technology as something that comes from the outside of society fully formed and then exert their impact on society in an autonomous way (Webster, 2014: 10–14).

A second way to define the information society has been to chart the growth in economic worth of informational activities in proportion to the GDP. Authors such as Machlup (1962) and Porat (1977) were forerunners in demonstrating the growing economic worth of information industries such as education, law, publishing, media and computer manufacture to the US economy. The rising curve was taken to demonstrate that western societies had indeed become information societies. Many have questioned whether it is possible to infer qualitative changes from purely quantitative indices. The world might in fact today be surrounded by informational activity with very little social and political consequences. For critics, “a nation of pleasure seeking couch potatoes hardly constitutes an information society” (Webster, 2014: 15–16).

Occupational definitions are closely associated with the work of Bell (1973), who coined the term “post-industrial society”. He, along with many others, interpreted the decline of manufacturing employment and the rise of service sector employment as the loss of manual jobs and its

replacement with white-collar work. Later on, influential commentators such as Reich (1991) and Castells (1997) argued similarly that the modern economy was in fact led by people whose main characteristic was the capacity to manage information. Critics have argued that by lumping together qualitatively different occupations in to categories such as “information workers”, these authors have ended up obscuring the power hierarchies that necessarily exist between different occupational groups (Webster, 2014: 17–19).

Spatial definitions of the information society, most notably advocated by Castells (1997) and Urry (2000), highlight the influence that information networks have on the organisation of time and space. By connecting locations, information networks such as the internet, are thought to radically reduce the importance of physical time and distance (Webster, 2014: 19–21). Finally, cultural definitions emphasise the increase in information that ubiquitous media in particular have brought about in to our daily lives (Webster, 2014: 21–23).

One the other side of Webster’s spectrum are theorists such as Herbert Schiller, David Harvey and Jurgen Habermas, who rather than arguing for a radical rupture from the past, see informatisation as a process most distinctively marked by the continuation of pre-existing capitalist relations (Webster, 2014: 7–9). Scholars on this side of the spectrum often view the information society as an ideological concept whose function is to legitimate social conditions in favour of dominant economic and political forces.

Recent research by Piketty (2014) certainly suggests that in the context of a longer history of capitalism, nothing much has changed, but that societies are in fact quickly reverting back to levels of inequality more familiar to the period depicted in Jane Austen’s novels. By fully embracing the notion that we today live in an information society where none of the old rules apply, we risk neglecting the analysis of broader

historical patterns such as those demonstrated by Piketty. Indeed, his work has contributed to a recent growth in scepticism towards “epochal thinking” in the social sciences (Savage, 2014).

What this brief review suggests is that we ought to be cautious of grand proclamations that epochal social changes are just around the corner due to latest advances in information technology, whether that is Big Data or something else. Be that as it may, there is no denying the increasing role that information technology now plays in our lives. Its growing importance is in fact such that we largely take it for granted. Instead of trying to prove or disprove the information society hypothesis therefore, I will instead illustrate some of the ways in which information technology has acted as an important site of ideological and political battle in the past few decades. A key aspect in this history is their connection to the political and economic ideology of Neoliberalism, a European construct which later found its most vocal expression in North America (Mirowski and Plehwe, 2009).

4.2.2 The emergence of computers and the rise of Neoliberalism

4.2.2.1 The demise of Fordism and a new push for automation

One way to understand the rise of information society rhetoric is to look at the political and economic response to the crisis of Fordism in the seventies. Hassan (2008) provides us with a framework from which we can begin to understand the connections between computerisation and Neoliberalism.

After the Second World War, Fordism, a mode of industrial production based on mass production, mass consumption, and the insertion of organisation and planning in to the business cycle had become the standard model of economic production across western countries. Fordism was more than a system of production however, as the effects

of planning and strong partnerships between organised labour, businesses and government, were felt widely across all spheres of life. In stark contrast to today, many countries experienced practically full employment and Keynesian economic policies were the norm rather than the exception. This was also the hay day of Social Democracy, with comprehensive health care and education programs being widely introduced. Importantly, throughout this period the majority businesses were territorially bound to the nation state (Hassan, 2008: 42; Webster, 2014: 75).

This highly regulated economic and social order started to break down in the 1970s. One explanation for the economic crisis that hit western countries particularly hard after the sudden rise in oil prices in 1973, was the problem of over-accumulation, which occurs when the rate of production outweighs the rate of consumption. In the wake of the war's destruction, capital's need for space and flexibility was guaranteed by the fact that many countries had to be rebuilt practically from zero. By the 1970s however, the world market had become crowded and to make matters even worse, many areas of the economy had been rendered off-limit for private investment by a wave of nationalisations that had taken place after the war. The lack of investment and strong union power also meant that research in to new technologies to increase the speed, flexibility and efficiency of the production process was low and heavily regulated (Hassan, 2008: 43–44).

It was in this context that an increasing number of business leaders, economists and politicians started to view regulation as the sole reason for the economic stagnation that had taken place. Arguments that labour unions and left-wing governments alone were to blame for the economic crisis started to gain increasing momentum. And with the crisis of Fordism as a mode of economic production, began the decline of Socialism and Social Democracy as viable frameworks for how to view and organise society (Hassan, 2008: 44–45).

In the context of a perceived need to free up capital, computers were from early on seen as way to radically improve the speed, flexibility and efficiency of the production process. During the cold war, both the US and the USSR had invested heavily in to research in computer technology, and especially the US had believed that in order to respond to a nuclear attack, a system would have to be put in place that was pre-programmed to be executed automatically. These lessons of automated forms of control through computing were not missed by the business community, which had started to automate parts of its production processes. With computer assisted automation, productivity could at least in theory rise endlessly. The important point to take out of this brief history is that despite originating in the sealed off world of military research, computers were in fact from the very beginning developed to a large degree from the perspective of capitalist management (Hassan, 2008: 45–48).

Up until the economic crisis of the 1970s, the then powerful unions had rejected almost any form of automation. Even though businesses were not necessarily happy about this, they had gone along with it because these were the boom years after the war in which profits rose steadily. All of this changed however with the crisis of the seventies. As profits plummeted and unemployment rose rapidly, a relatively tight-knit group of right-wing economists, politicians and media personalities seized the opportunity and began to offer a radically new way of doing things (For a detailed history of Neoliberalism's rise to dominance in the west, see Mirowski, 2013; Mirowski and Plehwe, 2009). A major part of this newly revitalised right-wing intellectual elite’s energies went into pressing for as much automation of industry as possible. Although this was not the first time that big corporations were offering computer applications to the private sector, the collapse of the post-war social contract between labour, businesses and government meant that the way for unrestricted and widespread automation was now clear. Right-wing governments

especially in the US and the UK pushed through the political conditions for automating as much of the production process as possible. In sum, computers contributed to a development whereby from the 1980s onwards, capital's bargaining power far exceeded that of labour's (Hassan, 2008: 48–50). Despite a near total collapse of the capitalist economy in 2007, this development shows no end in sight.

4.2.2.2 A cultural symbol for the superiority of the free market

The economic restructuring that took place from the 1970s onwards could not have taken place without a shakeup in cultural values as well. The move towards a strongly market oriented society required in the words of Harvey (2005) the "construction of consent" or as Hassan (2008: 67) puts it "…the engineering of a shift in what kind of ideas dominate society and to make these ideas so deep-seated that they appear as 'common sense'". Computers played no small part in the ideological battle that ensued.

Streeter (2011) documents these developments from the North American perspective. He points out that although the US is often thought of as a country where market freedom is seen as the archetype of freedom, market individualism has in fact always been only one part of a much larger mix of ideologies. In fact, in the first few decades of the twentieth century, new technology and progress in general were more associated with the American left than the right. Many of the things that made the US the richest and most powerful country in the world, such as the rise of powerful corporations, the effects of the New Deal, and the centrally organised war effort, were accomplished not by entrepreneurs working on their own, but by a close cooperation between the government and large centrally organised businesses. Keynesianism was standard for all political parties, and government regulation of many aspects of society was seen as rational, professional, and forward-looking. Still in the 1960s, calls for rollbacks in government regulation

and praises for free markets had seemed old-fashioned (Streeter, 2011: 72–73).

This was the context in which the newly revitalised group of right-wing intellectuals realised that in order to regain their intellectual authority they had to figure out a way to make market individualism seem modern again. The way that they went about this was to argue that markets were not only generally a more efficient way to organise things, but more specifically, they were superior at handling the most modern of technologies, such as radios, jet planes, and the relatively new invention at the time, the computers. From the late 1960s onwards, therefore, different strands of neoclassical economic thought combined to produce an intellectual framework for the idea that free markets were in fact good for the development of high technology. Right-wing economists, lawyers and intellectuals, many of whom were based at the University of Chicago, began attacking airline regulations, antitrust laws, and the regulation of radio waves. The proponents of the new paradigm argued for the deregulation of the exact same industries that a previous generation of regulators had seen as the prime examples of the limits of free markets. According to the new generation, regulatory bodies could make robust and progressive decisions by thinking in terms of consumer welfare and economic efficiency, instead of focusing on notions of the public good. In the context of a stagnating economy, these thinkers offered not just criticism of previous ways of doing things, but a way out of the crisis that would not mean giving up on established wealth and ways of life (Streeter, 2011: 74).

In the 1980s, neoclassical economics intersected with information society theory. A decade earlier, the notion of an information society had competed with many other concepts, such as “technetronic society”, “telematics society” and “compunications”. The reason information society came out on top was that from the perspective of capital, information had the advantage of being something that could be

imagined as thing-like, and therefore as property. For example, still in the 1970s, the idea that software was something that someone could own was at best controversial. One of the things that the neoclassical law and economics movements did was to revive the language of classical property rights, which then lent itself to the notion that the more rights there is the better. The infusion between the ideas that information was property and that markets were a superior way of organising things laid the foundations for the inflation of the notion of intellectual property. From the 1980s onwards began a process that saw the extension of patents to things like genes and software, the look and feel of computer programs, aspects of a pop stars personality, and eventually to business models such as Amazon's online ordering system (Streeter, 2011: 76–79).

In sum, by the 1980s there had been a general shift in the logic of legal decision making. Whereas previously the burden of proof had been on those that wanted to extend property protections, from here on the burden of proof was on anyone that wanted to prevent legal commodification (Streeter, 2011: 78). After this point:

> *Technology and modernity were no longer on the side of planning or the public good or an example of what democratic government could accomplish; they were on the side of rights, and government was their enemy, just as it was the enemy of rights. (Streeter, 2011: 79)*

There were also more practical ways in which computers became in this period a symbol for the superiority of the free market. For example, the popular myth that the computer industry had emerged solely from the initiative of entrepreneurial individuals such as Steve Jobs and Bill Gates had a big impact on the discourse around big business. Whereas previously the business world had been seen by many as a system dominated by a few large corporations in league with the government, the rise of the computer industry revived the belief that perhaps the

business world was in fact the result of innovative risk-taking individuals in competition with each other, after all (Streeter, 2011: 87).

The idea that computers were made possible by capitalist entrepreneurs and inventors alone also defined the social meaning attached to them. Streeter (2011: 88) argues that in contrast to countries such as France, where the first computers were experienced as part of a telecommunications system provided by the government, in the US:

> *[…] the experience of reading about, buying and using microcomputers created a kind of congruence between an everyday life experience and the neoclassical economic vision – the vision of a world of isolated individuals operating apart, without dependence on each other, individuals in a condition of self-mastery, rationally calculating prices and technology.*

Therefore, in the US, computers provided an everyday experience that made the market-ideology *feel* right (Streeter, 2011: 88). American cultural hegemony being so strong, we can reasonably assume that this has been the way that computers have been experienced in many other parts of the world too.

4.2.2.3 Neoliberal globalisation facilitated by information networks

The claim that computers were from early on developed to a large degree from the perspective of capitalist management makes much sense when we consider another important factor in the downfall of Fordism, namely, the globalisation of production and information and communication services. Indeed, although some businesses have always maintained a global presence, it is only in the last few decades that there has been an exponential growth in their number. In fact, today there are over 50,000 transnational corporations, and whereas in 1950 the majority of North American companies had subsidiaries in fewer than

six countries, today only a few operate on such a small scale (Dicken, 2010). Major corporations are today global to the extent that they no longer necessarily see themselves as being tied to any particular nation state. As is well known, a large company might have its headquarters in the US, design facilities in Europe, and manufacture operations in Asia (Webster, 2014: 79).

The by-products of the globalisation of production have been the globalisation of information services such as advertising, banking, and consultancy, and the globalisation of communications, such as mass media. The development of computer networks, the deregulation of stock markets, and the abolition of exchange controls in particular have led to a vast increase in the volume and velocity of international financial transactions. As a result, the financial sector has become so powerful that when nations lose the confidence of markets, governments must act rapidly to restore “confidence”, or otherwise face the collapse of their currency (Webster, 2014: 80–81).

What is important to take out of this is the point that these processes have been key in undermining the organisational premise of the Fordist system, the nation state. The Fordist order was based fundamentally on the on national government’s capacity to devise and implement policies, on relative immunity from foreign competition, and on distinctively national corporations. Under an ever more global market system, these conditions have become rare. Although the nation state still remains important especially in terms of people’s identities, its economic and political significance has witnessed a steep decline. The rise of transnational corporations has for example obscured what is owned by any given nation. This has then raised the disturbing question of to whom, if anyone, are these transnational corporations accountable to. Most nations are now forced to seek investments from transnational corporations, and the usual precondition for receiving them is the subordination to priorities set by them (Webster, 2014: 83–84).

What does any of the analysis above have to do with Big Data? It is important to remember that transnational corporations could not have risen to global dominance without information technology. As commerce became more and more global, ways of handling information and information flows, an “informational infrastructure” even, was put in place (Webster, 2014: 82). Without the worldwide spread of information technology, transnational corporations could simply not operate at the scale that they do today. Surely this must be at least part of the reason why so much of the hyperbole around information technology comes from the business community. After all, their power in this world depends on their widespread application.

Big Data can therefore be thought of as both the facilitator and result of corporate globalisation. As Herbert Schiller noted already many decades ago, the information explosion of the post-war years is to a large extent a consequence of corporate capitalism’s triumphant expansion. Let us not forget this when we today celebrate Big Data for its potential to be used for good in the world. For in addition to whatever benefits Big Data might end up yielding for society, it is also tightly connected to the growing power that transnational corporations now exert on the world stage.

What the section above has also shown us is that computers, and the information society rhetoric that accompanied them, were far from free of political ideology. Instead, the wide scale introduction of computers was a process deeply embedded in a major transformation in what kind of ideas that would dominate society. The scale of corporate globalisation that has taken place in the last few decades has been made possible by computers, which themselves have become a symbol for the superiority of the free market. The political conclusions that we have drawn from the increasing role that information technology plays in our lives must therefore be seen as a reflection of these broader economic

and social developments. Let us keep this in mind when we today hear political commentators arguing that digitalisation is both inevitable, and ultimately beneficial for everyone. In the next section, I review Big Data as a topic of both popular and academic debate.

4.3 What is Big Data?

A quick analysis using Google Trends global search reveals that it was only towards the end of 2010 that the term “Big Data” had begun to register, just ahead of an explosion of interest from 2011 onwards (Google, 2015). Since then, Big Data has become a buzzword commonly used in business circles, news media and science magazines. In popular accounts, Big Data has been presented as a “problem solver” for almost anything, from breast cancer to low cost governance, from better security to predictive systems, from smart cities and better traffic and water systems, to an end to urban squalor (Uprichard, 2014).

Others have been more critical and argued that the term itself has become so wide-ranging in definition that it no longer constitutes anything meaningful. More recently, the debate on digital data has started to centre around Artificial Intelligence and its related concepts, such as Machine Intelligence and Deep Learning.

To be sure, governments and businesses have been gathering and analysing large datasets for a very long time. In fact, as I pointed out in chapter one, practically all known societies that have existed since the art of writing was invented have practised some form of information gathering (Starr, 1987). In the next section I review some of the ways in which it has been argued that Big Data signals a significant shift from the past.

4.3.1 3Vs?

Although a myriad of definitions of Big Data exist (for a review of twelve different definitions, see Press, 2014), most of them have tended to stress the importance of the so called 3Vs of volume, velocity and variety (Stapleton, 2011).

The first of these, growth in volume, has to a large degree been the result of digitisation. Even if very large datasets, such as the census, are nothing new, there is no denying the major impact that information technology has had on our ability to store and process information on a daily basis. Indeed, the numbers behind such claims are truly staggering. On the internet alone, Google now processes more than 24 petabytes of data, or thousands of times the quantity of all printed material in the U.S Library of Congress, in just one day. Youtube users upload over an hour of video content every second, while Facebook gets over 10 million new photos uploaded every hour. On Twitter, on the other hand, over 400 million new messages are posted every day (Mayer-Schönberger and Cukier, 2013: 8). And these examples are now a few years old.

In addition to volume, Big Data is said to be distinguished by its high velocity. In contrast to traditional forms of data, which tended to present only a snapshot of the world at a particular time, Big Data is generated on a continuous basis, or sometimes even in real time. Digital devices have generated a persistent stream of information unlike previously possible.

Finally, Big Data is said to arrive in much greater variety than previous forms of data. Instead of just numbers on a sheet, Big Data can be pictures, sound or video, or a combination of these. The tendency towards greater variety is reflected in database design, where there has recently been a move from structured designs to ones that can handle a variety of sources and formats.

Kitchin (2014b) elaborates on the standard definition by adding four more characteristics to the list. In addition to the 3Vs, Big Data is according to him exhaustive in scope (covering whole populations instead of just samples), fine-grained in resolution and uniquely indexical (objects have unique identifiers which allow them to be tracked through time and space), relational (made up of common fields that enable linking), and flexible and scalable. For him, together the seven characteristics of Big Data constitute a significant shift, a revolution even, from previous forms of data gathering and analysis (Kitchin, 2014b: 79).

Furthermore, coping with such data is said to require new analytical techniques, from data mining and pattern recognition, to prediction, simulation, optimization, and new forms of data visualization techniques (Kitchin, 2014b: 101). The person with the skills to perform such analyses has been referred to as “the data scientist”, someone who “combines the skills of software programmer, statistician and storyteller/artist to extract the nuggets of gold hidden under mountains of data” (Cukier, 2010).

4.3.2 Correlation over causation and the end of theory?

A heated debate was started a few years ago by the former editor-in-chief at Wired magazine, Chris Anderson, who famously declared that Big Data signalled a new era of knowledge production characterized by “the end of theory” (Anderson, 2008). For Anderson, the ability to automatically analyse large volumes of data is making the scientific method obsolete. With Big Data, correlation becomes more important than causation, and science no longer needs theories and hypotheses to advance. Many similar commentaries have appeared during the past few years often, but not exclusively, from the business community.

Perhaps the best known popular take on Big Data, Mayer-Schönberger and Cukier's "*Big data: a revolution that will transform how we live, work and think"* (2013), reaches the same conclusion as Anderson. The book, which starting from its topic is a good example of modern myth-making around information technology, argues that Big Data is resulting in three shifts in how data is analysed.

First, since Big Data tends to be exhaustive, sampling is no longer required to the same extent as before. Whereas previously zooming in to sub-population usually meant losing statistical significance, Big Data can detect small anomalies and outliers. As a result, inferential statistics will increasingly be replaced by Big Data methods able to mine for anomalies in large datasets (Mayer-Schönberger and Cukier, 2013: 27–31).

Second, because the volume of data is big, statisticians will have to learn to accept its in-exactitude. Big Data increases the likelihood of errors in the data, which however no longer remains a problem because the risk of individual data points biasing the overall analysis is very small (Mayer-Schönberger and Cukier, 2013: 32–34). The focus on exactness is replaced by calculation "with messiness in mind" and where statisticians previously set aside their interest in large samples in favour of more randomness, with Big Data they will have to learn to deal with more imprecision in return for more data (Mayer-Schönberger and Cukier, 2013: 40–41).

These two shifts constitute the preconditions for the most important shift of all, the move from causation to correlation. In contrast to the previous era of hypothesis driven science, datasets are fast becoming too complex to be examined with a hypothesis in mind, but instead, with new computational techniques it is possible to let mathematical algorithms determine the best way to approach the data. As a result, data analysis is moving from a hypothesis driven search for causation to a computer

mediated search for correlations. Increasingly, “knowing what without knowing why is often good enough” (Mayer-Schönberger and Cukier, 2013: 55).

Therefore, although seemingly critical of Anderson’s proclamation (Mayer-Schönberger and Cukier, 2013: 72), Mayer-Schönberger and Cukier end up confirming it. All in all, the book functions as a modern example of myth making around information technology. For its authors “the IT revolution is evident all around us” (Mayer-Schönberger and Cukier, 2013: 78), and we have now arrived at the moment where “where the information society finally fulfils its promise” (Mayer-Schönberger and Cukier, 2013: 190). In contrast to previous myth-makers whose focus was on technology, however, this time we are encouraged to “recast our gaze to focus on the I, the information” (Mayer-Schönberger and Cukier, 2013: 78), which takes on an almost magical power to change the way we know the world. It provides, as Poon (2016: 1091) puts it, “a message that business people are eager to believe: with big data, truth, progress, and the pursuit of profit will finally resonate in perfect harmony.” Ultimately, Mayer-Schönberger and Cukier’s book reminds us of why Big Data has been much more than a technical development. For without a doubt, one of its functions has been to provide a fresh face for the claim that information technology transcends history, geography, and politics as we have come to know them.

4.3.3 Responses from the social science community

Considering that most Big Data, is social data (Uprichard, 2015), it is not surprising to find the topic widely discussed amongst social scientists. As Halavais (2015: 583–584) points out, the challenge that Big Data poses to the social sciences is in fact the very same one that has always been at its core, that of “…connecting the micro-connections between individuals to the vast social structures that shape us (and are shaped by us) as a society.”

As is well known, the social sciences have often found themselves caught somewhere in between the Sciences and Humanities, and so it is also in relation to responses to Big Data that a distinction can be made between social scientists who follow a more strictly scientific approach, and those who identify more strongly with the humanities tradition.

4.3.3.1 Not the end of theory, but the beginning of it?

For leading computational sociologists, Big Data signals, not the end of theory, but the beginning of it. Macy (2015) for example argues that what Big Data really implies is a move from statistical significance to theoretical significance. Speaking metaphorically, for him, Big Data has finally granted the social sciences its very own telescope, but theory is still needed in order to know where the telescope should be pointed at. With so much data around, statistical significance no longer works as a proxy for whether a correlation is meaningful or not, but what is needed instead is some notion of theoretical significance.

A second major shift brought about by Big Data is for Macy a move from individualistic to relational explanations. In contrast to traditional surveys, which treated individuals as isolated objects with various attributes (gender, age, education, occupation etc.), Big Data makes it possible to examine how individual behaviour is influenced by social networks. The big breakthrough with Big Data is therefore for Macy not that it signals the end of theory, but that it grants us access to relational data at a population scale.

Pentland (2014), another founding figure in the field of computational social science, argues similarly that with Big Data, the “old” language of markets and social classes will increasingly be replaced by concepts from what he terms “Social Physics”, such as “exploration”, “engagement” and “social learning”. Similarly to Macy, Pentland (2014:

191–192) argues that in contrast to the traditional concepts and methods, Big Data is able to account both for individual differences and for the relationships between individuals.

In general, computational sociologists such as Macy and Pentland tend to stress the opportunities with Big Data rather than the obstacles and dangers. Optimistically, they believe that practical solutions can be found to the problems associated with Big Data, such as those that exist around ownership and privacy (Pentland, 2012). What they reflect less on however is the central role that the very same data that they use in their research plays in the functioning of modern capitalism. Examined more critically, what this implies is that addressing the problems with Big Data must necessarily go beyond practical solutions to a deeper consideration of what the productive forces and relations of production in contemporary capitalism are. I will return to this theme in the last part of this section.

4.3.3.2 A threat or an opportunity?

More traditionally trained social scientists have identified Big Data both as a threat (Savage and Burrows, 2007) and as an opportunity (Lupton, 2014). Uprichard (2013) takes a strong position in the debate by arguing that now is the time when social scientists must reflect on just what kind of a social science they want to be part of. She argues polemically that left solely in the hands of physical, engineering, computational, and mathematical sciences, Big Data risks creating nothing short of a "methodological genocide" for the social sciences. According to her, we are "slowly but surely, becoming complicit to a deeply positivist, reductionist kind of social science, where variables are the be all and end all, where causality is devoid of meaning, and where non social scientists are the ones ruling the roost in terms of access, collection and analysis - of Big Data, which is *social* data." (Uprichard, 2013: 3–4). For her, Big Data remains very limited in its capacity to address the core task

of sociology, that of identifying and visualising where power networks lie. She therefore calls on sociologists to fight back, not just by training up their skills in data analysis, but also by starting to pose questions such as "…who is doing the counting? Who is making the decisions? Who is deciding what is counted and measured and how these counts and measurements are used and for whom?" (Uprichard, 2013: 4).

Not least due to the fact that funding agencies are increasingly refocusing their resources on projects that involve the analysis of large datasets (Kitchin, 2014b: 143), some social scientists have recently teamed up with computer scientists to explore the opportunities afforded by Big Data to sociological analysis. Reflecting on experiences from one such project, Housley et al. (2014: 4) see in Big Data an opportunity to "'digitally re-master' classic questions about social organization, social change and the derivation of identity from collective life." More specifically, by making it possible to capture "naturally occurring or 'user-generated' data at the level of populations in real or near-real-time", digital social research "…offers the hitherto unrealizable possibility of studying social processes as they unfold…" (Housley et al., 2014: 4). Furthermore, by fostering new relations between researchers and the researched, Big Data offers an opportunity to "motivate a renewed interest in the programme for a 'public sociology', characterized by the co-production of social scientific knowledge involving a broad range of actors and publics" (Housley et al., 2014: 1).

Reflecting similarly on an interdisciplinary project involving both sociologists and computer scientists, Tinati et al. (2014) argue that social science has so far been slow to pick up on the promise of Big Data due to methodological restrictions. For them, the emergence of Big Data mirrors recent theoretical concerns of sociology, such as the move from seeing the social as society bounded by nation-states to understanding it as the mobility and flow of people, objects, images and information (Urry, 2000). With publicly available Big Data it is now possible to start

exploring these issues empirically, but so far, social science methods have been unable account for the key characteristics of Big Data, which are its scale, proportionality, dynamism and relationality. Hence, they call on social scientists to start expanding their methodological repertoire, by amongst other things, forming collaborations with computer scientists, like they have successfully done.

Many have also reflected on the inherent problems that exist in large scale data sets. Lewis (2015) for example warns against treating Big Data as “naturally occurring” and unproblematic. First, he points out that although Big Data is often celebrated for being able to grant access to complete populations, it is in the end still a sample. Certain kinds of people are more likely than others to turn up in certain kinds of datasets, and the information available for each individual may vary hugely. Second, in contrast to the notion that Big Data captures “natural” behaviour, our interactions are in fact constrained by technology in ways which can be difficult to discern. Third, just as Big Data cannot be interpreted independently of the technology through which it is generated, neither can it be interpreted without some understanding of the cultural context in which it has generated. For example, friendship dynamics on social media can operate in entirely different dynamics to real life. Hence, for Lewis, Big Data poses a set of new challenges rather than a simple straightforward solution. Elsewhere, Busch (2014) identifies a total of twelve problems with large scale from distortion to errors, bias, and misinterpretation. Not least because Big Data might increasingly be used to make policy decisions, understanding the manner of its construction is now of key importance. Diesner (2015) points out similarly that conducting data analytics always involves a plethora of small decisions, many of which can have enormous consequences for research outcomes. Hence, she calls for increasing awareness in documenting and accounting for these decisions.

Ruppert (2016), on the other hand, offers a Science and Technology Studies perspective on Big Data. According to her, sociologists should recast their focus from the 3Vs, which for her are qualities, rather than definitions, to the changing data practices which they are an outcome of. She points out that Big Data "are generated and sustained through multiple and selective sociotechnical practices that include not only technologies and people but also norms, values, conventions and rules." (Ruppert, 2016: 2–3). As a result, a focus on what she terms "data practices" can account for the changing relations to data that cut across different contexts, which are of four kinds. First, the digital actions that generate Big Data are also inventive of new forms of sociality. Social networking sites, blogs etc. "instantiate social relations that are part of who we are as individuals and collectives in novel ways." (Ruppert, 2016: 3). Second, Big Data gives rise to new method relations. Various actors are now inventing and experimenting with different methods to represent and enact social worlds with Big Data. A third set of relations comes from the fact that people are now ever more aware of the ways in which they are "being made into 'data subjects', analysed and known" (Ruppert, 2016: 3), as the recent revelations by Edward Snowden have made clear. Finally, Big Data practices are also changing research relations for social scientists, as academia is very much part of both generating Big Data and defining the themes, concepts and concerns that make it up as a field. For Ruppert, the ethical challenge for social science is now "to find ways of being accountable, answerable and responsible to the effects of our methods that take up Big Data and the worlds and ways of being they elevate and promote." (Ruppert, 2016: 4). In sum, for Ruppert, Big Data is not just about its qualities, but about new forms of organising data as well. Following this idea, in the next section I analyse the social context in which much of Big Data is being generated.

4.3.3.3 A corporate takeover of social knowledge production?

A classic study by Graham and Marvin, *Splintering Urbanism* (2001), documents the many ways in which public infrastructure provision of power, transport, communication and water are being "unbundled" and "splintered" as a result of a widespread movement towards privatisation and liberalisation across many countries and continents. There is good reason to argue that this privatisation, or neoliberalisation, of infrastructure now extends to the production of data as well. After all, it is largely private companies that are collecting, organising and owning the data that is generated by our transactions with various, often commercial, platforms both online and offline (Van Dijck, 2013). Many public sector institutions, such as national statistical institutes, which used to hold a monopoly over the production of data for the making of official statistics, now fear that they will increasingly have to negotiate with private companies for access to the best data (Struijs et al., 2014).

boyd and Crawford (2012: 673) have voiced the concern that because Big Data is mostly owned by private companies, it might be creating new hierarchies between researchers with access and those with little or none at all. After all, as things currently stand, only a handful of companies have access to very large social data, which means that they alone set the rules for who can use it and for what purposes. Therefore, in addition to creating hierarchies between researchers, Big Data might also lead to a situation where private companies will increasingly start to impact, if not dictate, the types of questions that are being addressed by social scientific research. The fear is that researchers with access to Big Data might be less likely to choose questions that are contentious to social media companies if they think it may result in their access being cut (boyd and Crawford, 2012: 675).

There is much reason to suggest that the ethical concerns with Big Data go far beyond the question of access. For example, as Thatcher (2014)

points out, currently only a handful of companies get to decide what can be known through and done with Big Data. Since these companies are ultimately driven by the profit motive, they care less about the accuracy of their data than about its potential use in attracting more customers. Hence, when researchers accept the data as inherently meaningful, they also accept "an epistemological framework of knowledge structured through capitalist imperatives." (Thatcher, 2014: 1772). Therefore, even when a researcher gets access to the full volume of existing Twitter data, the quality and characteristics of the data are still beyond her control. In sum, with the move to Big Data comes to the risk that the very limits of knowledge will be "…set through the data infrastructure of private corporations." (Thatcher, 2014: 1766).

Increasingly, Big Data research can be used to not just describe society, but to impact it in very direct ways as well. This became clear with Facebook's 'emotional contagion study' (Kramer et al., 2014) which demonstrated that by choosing what messages to show to its users, Facebook can directly influence their emotional state. Those users who were shown more positive words in their newsfeeds also then posted more positive words, and the same was true for negative words. Reflecting on the study and the public concern it raised, Schroeder (2014) notes that although companies have for quite some time been using market research and advertising to influence people's minds, Big Data represents a leap in how data can be used to manipulate people in powerful ways. Even if Facebook claims that it would never use this type of research for anything else than improving user experience, it is not difficult to imagine how it might one day be used to for example influence people's political behaviour. And were this to be the case, we would probably never even know about it, for as long as Facebook and Google remain commercial services that we use voluntarily, there is no obligation for transparency. Considering the public anger raised by the contagion study, it is in fact likely that this type of research will increasingly be conducted behind the public eye (Schroeder, 2014).

Yet another troubling development are the close links that now exist between academic researchers and private companies. One cannot for example help but to wonder whether boyd and Crawford, who I already introduced earlier in this section, and who are both employed by Microsoft, might in fact exemplify their very own critique (2012). It is after all likely that by aligning themselves closely to Microsoft, they have rendered some modes of critical inquiry in to its workings off limit. As Chan (2015) reminds us, we must be equally worried about what happens when access to Big Data is granted, as we are about the divides that the process might be creating. She argues that collaborations, just as much as divisions between researchers, are the source of the research norms and practices that are currently emerging around Big Data. Cautionary examples of the downsides of public private partnerships come from pharmaceutical, biotechnological and medical research, where an increasing emphasis on corporate funding since the 1980s have led to a series of cases where significant amounts of unfavourable findings on commercial products have been censored from publication. One cannot therefore help but to fear that by owning the most valuable data, and therefore controlling access to it, corporate interests will begin to have an increasing impact on the direction of academic research and higher education more generally.

This impression is further enhanced when one examines the research outputs that have come out of the burgeoning field of computational social science introduced in the beginning of this section. It is notable that in a leading conference in the field, International Conference on Computational Social Science, organised in Helsinki in June 2015, there were in fact not many sociologists among the presenters. Most had a background in computer science or physics, which was clearly reflected in the kinds of questions that they were addressing in their research. It is telling that the one presenter who did identify himself as a sociologist, Macy (2015), already introduced earlier this chapter, defined

computational social science as “computer enabled studies of human behaviour and social interaction”. This is an extremely narrow definition of social science, and one that does not address the fundamental questions of sociology, which have traditionally been about identifying structural roots of inequality and figuring out where power resides. As argued by Crawford et al. (2014: 1667), “Aggregated, individual actions cannot, in and of themselves, illustrate the complicated dynamics that produce social interaction—the whole of society is greater than the sum of its parts.”

It is important to note that certainly not all Big Data is produced in the private sector. Kitchin (2014b: 87–98) for example lists numerous examples of sources of Big Data, of which only part reside in the private sector. Big Data relating to health, public administration, satellite imagery and GPS, traffic control rooms, smart city technologies, university research data, environmental data, are some examples of Big Data that are either entirely publicly produced or produced in conjunction with the public sector who procure the service and dictate the terms of its production. Much of the Big Data that governments for example are hoping to exploit are generated within, rather than outside of them. Furthermore, also civil society is in numerous ways actively involved in both producing and analysing Big Data. Much Big Data is volunteered data, generated for example as part of crowdsourcing (Howe, 2009) or citizen science (Gabrys et al., 2016) initiatives. On the other hand, growing calls for data to be made available publicly and free of charge (Berners-Lee, 2012) has meant that citizens and non-governmental organisations now have access to an increasing amount of data from which to draw their own conclusions and to potentially challenge government and corporate agendas. In sum, Big Data is not only being used to advance corporate agendas, but to challenge them in many ways as well.

Nevertheless, as numerous examples in this chapter have shown, private corporations do seem to have an increasing presence in both producing and analysing Big Data. It is for example noteworthy that the majority of presenters at the conference for computational social science in Helsinki were building their analyses on data obtained not from governments, but from businesses. A major risk with a growing need to collaborate with the private sector in order to access Big Data is that social science might increasingly become reduced to the study of behavioural dynamics. This is exactly the kind of research that companies will want to fund, first, because it yields clear practical applications for their products, and second, because it does not address the difficult questions of political economy that might end up reflecting badly on them. In its current form, what it amounts up to is an apolitical form of social science for an apolitical world. The next Nobel prize in social science might very well go to a computational social scientist, but whether the research leading up to it truly addresses the pressing social and political problems of our time, is a different matter entirely. For little do the computational social scientists reflect on the fact that much of the same Big Data that they use for their research is now at the core of the functioning of modern capitalism. I conclude this chapter by situating Big Data in the context of political economy.

4.4 The political economy of Big Data

4.4.1 Data as a key source of economic value

An important part of the popular discourse around Big Data is the idea that data now constitutes an important form of economic currency. According to Mayer-Schönberger and Cukier (2013: 182), for example, “Data is to the information society what fuel was to the industrial economy: the critical resource powering the innovations that people rely on”. And they are certainly not alone in claiming that although data has

always been important to business, it now occupies a far more critical role than before.

In their popular take on Big Data, Mayer-Schönberger and Cukier (2013: 96) note that in the near future, everything from everyday wearable objects to the built environment will have chips, sensors and other communication modules embedded in them. As a result, language, location and interactions will be recorded in a quantified format for it to be tabulated and analysed. This data is immensely valuable for businesses, who use it to target their own services and to sell it to other third parties, such as advertisers.

Especially those who are keen to promote Big Data would argue that, already today, Amazon knows what we buy, Google knows what we browse online, and Twitter knows what we have on our minds. Facebook, on the other hand, gathers all this information, along with our social networks. Mobile network operators know where we walk, who we talk to, and who is nearby. In a relatively short period of time, these platforms have become “goldmines” of data that can be used to infer various things about people, from their socio-economic statuses to their movements through space and time (Mayer-Schönberger and Cukier, 2013: 96–97).

As Morozov (2015b) explains, this data is now at the core of how many of the world’s most powerful companies operate. In contrast to pre-digital times, when we used cash to pay for services, today we pay for them by surrendering our data. For Morozov, this constitutes a process of double exploitation: first, we are exploited when the data we give up in exchange for relatively trivial services eventually ends up on companies’ balance sheets, and second, when the data is used to customise and structure our world in a non-transparent way through targeted advertising. Whereas cash had no connection to our social lives or life histories, data is in essence a representation of them, to be turned in to money by

internet conglomerates such as Google and Facebook. Hence, these companies are currently promoting free connectivity and digital inclusion not out of their good will, but because they want to own the right to monetise the lives of as many people as possible.

Terranova (2000) identified the growing importance of free labour online at the start of the millennia. Even if the platforms she mentions are different from the ones we use today, the mechanism itself is familiar enough:

> *Simultaneously voluntarily given and unwaged, enjoyed and exploited, free labor on the Net includes the activity of building Web sites, modifying software packages, reading and participating in mailing lists, and building virtual spaces on MUDs and MOOs. Far from being an "unreal," empty space, the Internet is animated by cultural and technical labor through and through, a continuous production of value that is completely immanent to the flows of the network society at large. (Terranova, 2000: 33–34)*

More recently, Ritzer and Jurgenson (2010) have argued that although capitalism has always been characterised by both production and consumption, the recent explosion in user-generated content online has made the two more interlinked than before. In “prosumer” capitalism, value is created from our voluntary use of free services online, but in contrast to previously, this time it is more difficult to determine whether the process is exploitative or not. This is firstly because users generating the content are enjoying themselves while doing it, and secondly, because online platforms also provide a wide base for resistance against commercialisation. Thus, these authors speculate on whether “prosumption” might bring about a new form of capitalism, one characterised by free services and an abundance of content.

Fuchs (2014b: 110–111), on the other hand, explores these issues from a Marxist perspective. Following Marx, he argues the rate of exploitation

is ultimately defined by how much of worker's work time goes unpaid. As users of social media do not receive any salary at all, capitalist "presumption" constitutes for him nothing less than an extreme form of exploitation. For Fuchs, the counter-argument that users of social media are not exploited because in return for their work they receive access to free services would be true only if "free access" could be further converted in to a salary to buy food and housing. Thus, for Fuchs, media in the age of the internet are still fundamentally shaped by capitalist class relations.

Andrejevic (2014) explores the social divides that are being created by Big Data in more detail. According to him, Big Data is bringing about a structural divide between a small minority or people with the capacity to collect, store and mine large quantities of data, and a majority whom the data collection targets. For him, what is most important about Big Data analytics is that they grant the few people with access to the costly infrastructures and technical expertise required by them an unprecedented capability to socially sort, that is, to assign worth to others, who in most cases remain unaware or feel powerless to contest these practices. Although social sorting of this kind is currently used mainly in advertising, it is likely that they will become part of a growing variety of decision-making and forecasting operations. Hence, the big paradox of the emerging Big Data era is that as we overcome the digital divide by granting more and more people access to digital devices, we end up exacerbating a "Big Data divide".

As Qiu (2015) points out, Big Data can be seen as the latest wave of neoliberal expansion. He notes that "While the technological sphere of social media is new, so is the global phenomenon of Big Data worship, the ethical question about 'accessing', privatizing, and commodifying the commons has been a time-honoured concern that goes all the way back to the beginning of the capitalist world-system." (Qiu, 2015: 1092). Similarly to new technologies in previous eras of capitalism, from

shipbuilding to mining to weaponry, media and computing, Big Data represents yet another way to render formerly “inaccessible” regions “accessible” for private wealth accumulation. Hence, for Qiu (2015: 1092), the ethical problem with Big Data is in fact “as old as capitalism itself.”

As a result of such developments, data is now according to many a key, or perhaps the key, economic asset. Many of the major companies that now gather large amounts of data were not originally founded with that purpose in mind, and recently they have realised that as it is practically impossible to know what uses the data might have in the future, the best thing to do is to gather as much of it as possible (Mayer-Schönberger and Cukier, 2013: 98–102). The majority of Facebook’s share value is for example not based on how much money it makes at the moment, but on the financial potential its data holds (Mayer-Schönberger and Cukier, 2013: 118–120).

In addition to transforming traditional industries such as advertising, the economic value that data now holds has given rise to entirely new industries, such as the so called “data brokers” that repackage data into privately held data sources for rent or re-sale purely for profit. Companies such as Epsilon, Acxiom, Alliance Data Systems, eBureau, ChoicePoint, Corelogic, Equifax and Experian have databases with information concerning over 700 million consumers worldwide. Amongst other things, the data analysis products provided by these companies are used to micro-target advertising and marketing services, assess credit worthiness and socially sort consumers. The data concern all aspects of everyday life, from public administration, communications, consumption of goods and media, to travel, leisure, crime, social media actions, and so on. Although data brokers have been around for a long time, it is only recently that their operations have grown in to a multibillion dollar business (Kitchin, 2014b: 42).

According to Mayer-Schönberger and Cukier (2013: 123–149), in a world where data is the main source of economic value, companies will increasingly be divided into those with the data, those with the skills to analyse it, and those with the ability to innovate with it. Large data holders, such as Google, will become ever more powerful due to their scale advantages in storing more and more data, whereas innovative small companies will be able to have a large presence without big investments in physical resources, such as human labour. As a result, the industry will be polarised between a few major players and numerous small start-ups, with little room for mid-sized companies in between (Mayer-Schönberger and Cukier, 2013: 123–149).

Keeping in mind the key economic role that data now holds, a major challenge of the near future is to prevent monopolies on its storage and use. Mayer-Schönberger and Cukier (2013: 183) contrast modern day "data barons" with nineteenth-century robber barons who dominated America's railroads, steel manufacturing and telegraph networks. In the next section, I will explore the possibility that just as the early industrialists were eventually put under control with antitrust legislation, similar measures might be needed to be taken with dominant Big Data companies. So what does the Big Data industry currently look like then?

4.4.2 A global cartel in information resources?

Many now argue that to some extent, a global Big Data cartel already exists. Indeed, few of those who so eagerly celebrate Big Data for its potential to be used for good in the world remember to mention just how unevenly its financial gains are, at least for now, being distributed.

Morozov (2015e), already introduced earlier in this chapter, defines the business models of companies such as Amazon, Facebook, Airbnb and Uber as "platform capitalism". In contrast to the previous economic model, where individual firms competed against each other for

customers, internet firms provide a platform for customers to engage with one another. So whereas taxi companies used to transport customers, Uber claims it is simply connecting drivers with passengers, and where hotels used to offer hospitality services, Airbnb is just connecting hosts with guests.

Two major issues have arisen with the new economic model. First, by clinging on to the platform status, the aforementioned companies have been able to by-bass much of the consumer protection legislation that traditional businesses are subjected to (Morozov, 2015e). This is but one of the many ways in which digitalisation accompanied by free market ideology has started to undermine Social Democracy's most fundamental principle of maintaining a market economy while granting the state a strong mandate to regulate it. Instead, the standard position in Silicon Valley is to argue that the only thing that should regulate the market, is the market itself (Morozov, 2015a).

Second, the major platforms have in effect become monopolies. After all, most of the big companies located in Silicon Valley are there not by accident, but because it is the place where so much of the material resources of data, algorithms and server power demanded by "platform capitalism" today lie. The companies behind these platforms, most of which are also notorious for their ability to evade taxes, do not produce anything on their own, but instead rearrange what others produce in order to make profit for themselves (Morozov, 2015e).

Mosco (2014) documents further the many ways in which the Big Data industry has started to resemble monopoly capitalism. According to him, the Big Data industry has in a relatively short period of time become dominated by only a handful of companies, such as Amazon, Apple, Google, Microsoft and Facebook. In stark contrast to their public image of riding the latest wave of modernity, they have in fact been following some very traditional strategies in economic history.

First, all of the aforementioned companies have recently been cutting prices to the point where smaller companies can no longer compete with them. Historically, once the smaller competitors have been pushed out, prices have begun to rise again (Mosco, 2014: 56–57). Secondly, what they have also started to do is to exercise power up and down the chain of production. Indeed, all of the aforementioned companies are now aiming to build their own computers and thereby challenge old manufacturers like Intel and HP. Even Facebook, which at least so far has not been associated with devices, has recently teamed up with traditional manufacturers to develop its own chip. Google has developed its own semiconductor and Amazon is building a global computer system including its own computers, storage and networking systems, and power stations (Mosco, 2014: 58).

According to Mosco (2014: 58) the major Big Data companies are now "integrating internally to rationalize production from hardware to software, applications, and pricing". The reason they do this is to extend control over markets in order to establish key positions in the development of informational capitalism. With no regulatory framework in sight, the result might soon be a global data cartel not unlike the oil cartel that influenced the energy markets for years. Like the oil cartel before it, the data cartel would "provide for the needs of organizations and individuals, using control over various stages in the production and distribution process that powers global capitalism" (Mosco, 2014: 59).

The new position of power that these handful of companies now find themselves in is well exemplified by a recent statement from Google founder Eric Schmidt, according to whom Google and its competitors should no longer think of themselves as companies, but rather as nation-states. Schmidt put this attitude into practice in 2013, when he travelled to North Korea to meet with its leadership without informing the U.S government (Mosco, 2014: 58–59).

In sum, in a world where data is the new gold, American technology companies are the new imperialists, constantly innovation new ways to suck the world dry of its greatest economic resource. Furthermore, in the data game, Europe is now a mere digital colony (Järvinen, 2014).

Things are of course more complicated than this, as the currently dominant Big Data companies also face a set of challenges from a variety of directions, not least by the fact that by providing better and better cloud services they often end up cannibalizing the sales of their traditional products (Mosco, 2014: 62). Recent developments have nevertheless led many to wonder whether the cloud computing industry could soon face “the problem of monopoly market domination that once led the government to intervene against the power of Standard Oil, IBM, and AT&T” (Mosco, 2014: 60).

Such developments have also raised the questions of whether the enabler of Big Data, cloud computing, should be considered a public utility. After all, cloud computing already shares most of the characteristics of earlier utilities, such as water, gas, and electricity, and as the market is likely to grow into one where a few providers practically everyone, the only thing that is missing for cloud computing to be a utility, is a regulator. Currently the industry is governed by market forces alone, but historically utility markets have tended to become government regulated when one or few producers have come to dominate (Mosco, 2014: 44–45).

For Mosco (2014: 42), for the time being, cloud computing is a singularly market driven project that shows little or no consideration for how it might be used to expand economic or political democracy, to increase worker participation in corporate decision making, or to improve citizen participation in national or community life. Well known political uses, such as its uses in election campaigns, show little consideration for how

it might empower citizens to participate in decision making. Rather, they are examples of population management and control (Mosco, 2014: 25–27).

4.4.3 Governments and Big Data

And by no means is the public sector excluded from cloud computing. The US government for example has recently began outsourcing its cloud computing needs. Increasingly, the scale advantages have started to outweigh the security risks involved when data storage and processing are outsourced to private companies. And as we have recently learnt, contrary to previous claims, the U.S government is in fact working closely with the major tech companies to gather information on both American and foreign citizens (Mosco, 2014: 66–70).

Increasing lobbying efforts from the major technology companies has not been a small factor in this development. In addition to lobbying for deals with government departments, the big technology companies are now also spending millions in efforts to resist the strengthening of privacy laws both in the US and in Europe. For long, the EU was adamant in its resistance to lobbying, but as a result of its weakened economic position it has started to give in. Many European countries are now so desperate for investments that they are willing to give up on privacy for free trade agreements with the US. And as the recent Snowden revelations have revealed, US legislation clearly violates the privacy of EU citizens, a big reason why the EU has been trying to establish its own data privacy regime (Bowden, 2013). In sum, for corporations, and the politicians supporting their agenda, privacy is increasingly seen as a barrier to economic growth (Morozov, 2015d).

Some fear that in addition to taking over data processing and storage from the government, big tech companies might soon take over politics more broadly. Many in Silicon Valley are now promoting a data-driven

approach to governance – “algorithmic regulation” whereby regulations and law enforcement are enacted by algorithms that process the data collected from citizens via smart devices and computers. If all physical objects are embedded with digital technology, companies like Google will increasingly act as gatekeepers between citizens and the government (Morozov, 2014).

Morozov (2014) defines the Silicon Valley approach to politics as “solutionism”: all social problems are to be dealt with apps, sensors and feedback loops. What it amounts up to for him is a technocratic utopia of politics without politics: “Disagreement and conflict under this model, are seen as unfortunate by-products of the analog – to be solved through data collection – and not as inevitable results of economic or ideological conflicts” (Morozov, 2014: 6). Furthermore, if regulation is outsourced to algorithms operated by companies, the role and purpose of the state will become increasingly unclear.

Increasingly, the technology industry believes it can now even solve problems to do with social inequality. According to their reasoning, the rise in income inequality does not matter as long as it is accompanied by a decrease in consumption inequality. So the fact that the gap between the rich and the poor keeps growing and growing matters not as long as the poor have access to the services provided by tech companies, an argument which might be true if the tech companies also provided for food and housing (Morozov, 2015c).

Were this to be the case however, it would raise the question of why bother to have a state in the first place. For Morozov (2015c: 3) the choice we are faced with today is not between the market and the state, but between “a system bereft of any institutional and political imagination – where some permutation of hackers, entrepreneurs and venture capitalists is the default answer to every social problem – and a system, where explicitly political solutions that might question who – citizens,

firms, the state – ought to own what, and on what terms, are still part of the conversation".

4.4.4 Data proletariat of the world unite?

As I have already mentioned earlier in this chapter, automation has generated both utopian and dystopian visions ever since the 1970s. Commentaries on its impact have varied from a world free of work to one characterised by mass unemployment, stark inequality, and widespread social unrest. Recently, there has been a resurgence of the debate. For example, a while ago, a widely reported study by Frey and Osborne (2013) estimated that nearly half of US jobs could be automated within two decades. In another recent popular take on the subject, Brynjolfsson and McAfee (2014: 7–8) argue that we have in fact entered "a second machine age", or a period in which "computers and other digital advances are doing for mental power… what the steam engine and its descendants did for muscle power".

Many fear that the short term effect of automation is likely to be a negative one especially for the middle classes, which for long have acted as the guarantors of social peace and stability. After all, the people who get to work at the major technology companies constitute only a small elite in a global supply chain. These lucky few, sometimes referred to as the "tech aristocracy", get to enjoy high salaries and workplace perks that normal workers can only dream of. At the other end of the supply chain are industrial workers in developing countries where working conditions are not far off from the "satanic mills" of the industrial age. In between these two extremes is a large pool of middle class jobs that often involve the production, processing and distribution of information (Fuchs, 2014a).

Worryingly, it is the latter group of jobs that seems to be most at risk due to automation. Big Data analytics are making it increasingly possible to

subsume into technology much of the labour involved in the information and cultural industries today, and the dystopian future scenario is a labour market steeply polarised between those in low skilled/low-pay service work and those in the upper reaches of organizations (Mosco, 2014: 165–167).

Even the venture capitalists themselves are now actively debating the threat that inequality has started to pose (See for example, Ferenstein, 2014). Lanier (2013) has popularised this debate from within the Silicon Valley bubble. He argues that information on a global scale is currently channelling in to what he calls “Siren Servers” to an extent and at a speed that is soon going to threaten democracy itself. In sum, Big Data is channelling wealth in to the hands of the very few, while at the time destroying middle class jobs and the social democratic political order that accompanied them.

For Lanier, the main problem is currently that people give up their data to companies for free. According to him, the early years of the internet emphasized open access and knowledge-sharing to the extent where it has distracted people from demanding fairness and job security in an economy that is based on information. For him, the threats that Big Data pose could be alleviated by establishing a system whereby companies would compensate people for their data through micro-transactions. In this system, Facebook, Google etc. would have to pay royalty to the user whose data it was selling.

As reasonable as this, essentially market-based solution to Big Data capitalism’s problems, sounds, it is very difficult to imagine how one would get the big technology companies to collaborate in establishing such as system, regardless of the strength of the argument that doing so was in their long term interest. In sum, as things currently stand, and as the numerous examples in this chapter have demonstrated, Big Data is

currently producing fewer winners than we might, and certainly should, hope for.

4.5 Conclusion

In this chapter, I have taken up the challenging task of situating Big Data in the context of political economy. What the analysis has shown is that information technology, Big Data included, are far from free of political rationalities. Importantly, the introduction of computers, the facilitators of Big Data, was an event strongly associated with a larger ideological shift in societies. Amidst all the celebration about the supposed benefits that information technology brings, too little reflection has been paid to the fact that computerisation has coincided with neoliberalism's triumph as a political rationality. This central point has also not been captured in the popular discourse around Big Data, large parts of which has been catered to a business audience. Instead, the narrative has done much to obscure and naturalise developments that are inherently political.

In sum, although not all Big Data is produced by private companies, a significant part is and is therefore underpinned by a market and profit rationality. What the increasing production of data in the private sphere means for the production of official statistics in Finland, which, as I argued in the previous chapter, has for long been underpinned by social welfarist concerns, is a question I explore in the next chapter.

5 Big Data at Statistics Finland: A neoliberalisation of statistical practices?

5.1 Introduction

In the previous chapter I situated Big Data in the context of political economy and argued that its emergence reflects broader processes of neoliberalisation that have swept over western polities in the past few decades. However, my take on Big Data was similar to many others in that it was of a conceptual, rather than empirical nature. But how does the so called data revolution play out at a more practical level? As Kitchin (2014b: 118) points out, what we need at the moment are analyses that move beyond the conceptual to empirical explorations of the workings of Big Data within specific contexts and domains. By drawing on expert interviews with statisticians at the National Statistical Institute of Finland, Statistics Finland, in this chapter I move my focus from conceptual debates to the everyday practices where social transformations and revolutions get enacted.

I draw on three main analytical principles to guide my analysis. First, I follow Ruppert's (2016) suggestion to recast focus from the technical qualities of Big Data, such as the 3vs, to the changing *data practices* that the qualities are the outcomes of. According to her, these practices include "not only technologies and people but also norms, values, conventions and rules." (Ruppert, 2016: 2–3). By analysing the various practices prompted by Big Data at Statistics Finland, I suggest that they are indicative of an increasing "neoliberalisation" of statistical infrastructures (of analysis, processing and storage) and occupational cultures. By neoliberalisation I mean not only the privatisation of public utilities, but also the extension of market rationalities, logics and values to most spheres of life, from the ways governments are organised and

managed to how people organise their social relations (See for example Brown, 2015; Mirowski, 2013; Peck, 2010). I understand neoliberalism therefore as not only an economic doctrine, but a rationality that has sunk so deep in to everyday life that it now passes as common sense (Mirowski, 2013: 28). By referring to neoliberalisation I take up Peck and Tickell's (2002) suggestion to focus on the processes, such as institutional reforms, in which ideologies of neoliberalism are produced and reproduced. Finally, I take up Pelizza's (2016) point that infrastructures are key sites where institutional shifts, and eventually state transformations even, can become visible. By attending to changes in information infrastructures brought about by conceptions of Big Data at Statistics Finland, I provide early evidence of an emerging reconfiguration between the public and the private sector in the production of data and the making of statistics.

The analysis is divided in to four parts. First, instead of assuming a predetermined definition of Big Data such as those I have already cited, I start by describing how statisticians understood it and the questions they saw it raising their professional practice. In contrast to popular debates that locate its newness in its technical qualities, such as the 3vs, my respondents were more inclined to see the newness of Big Data in how the production of data was becoming organised, particularly in the sense that more and more institutions now had the capacity to produce and analyse data, which could end up challenging NSIs in their role as the producers of official statistics. Next, by analysing in detail Big Data projects that were ongoing at Statistics Finland at the time of research I suggest that in an age of Big Data the production of data for official statistics is likely to be increasingly influenced, not by the social welfarist rationalities of government departments, as has been the case in the period of register-based statistics (Alastalo, 2009b), but by neoliberal ones of private corporations. By interrogating further a set of responses to Big Data at Statistics Finland, I suggest that neoliberal rationalities can also be identified in the ways that Statistics Finland imagines its

future role and relevancy. I conclude the analysis by exploring companies' attitudes towards data sharing, as perceived by my interviewees. I suggest that in contrast to a markedly positive picture painted by the interviewees, there are ways in which private sector rationalities are likely to come into conflict with a desire to share data for the advancement of social goods.

It is important to note from the outset that all Big Data projects at Statistics Finland were still at an experimental phase, and that no pre-existing statistics had so far been replaced. In its current form, Big Data did not challenge the register-based statistical system, and as noted in numerous policy reports and opinion pieces, many challenges and risks of using Big Data for official statistics remain unsolved (For an overview, see Kitchin, 2015). What is already clear however is that there is a growing perception in official statistics that Big Data challenges NSIs and their accustomed ways of doing things. For the time being, therefore, the much discussed "data revolution" manifests itself in official statistics most clearly in changing rationalities and mentalities, which, however, are already leading to some very practical outcomes in terms of organisational arrangements etc.

In sum, following the conceptual starting point of this thesis, that statistics and society are co-constituted, and that governmental rationalities form an integral part of this co-constitution, I suggest that Big Data developments indicate that an increasingly neoliberal rationality is becoming part of the making of official statistics. Before the analysis, however, I start by describing my sample and the methods I have applied in its analysis.

5.2 Data and research methods

My sample consists of nineteen semi-structured in-depth interviews conducted with statisticians at Statistics Finland between December 2014 and September 2016. In addition, my analysis is informed by notes taken during three fieldwork visits to Statistics Finland during which I attended the meetings of a recently established Big Data team. Furthermore, I have also studied a number of policy documents, including the minutes from the Big Data team's meetings, which I obtained during the fieldwork.

Roughly half of my interviewees were recruited from the Big Data team that was established at the beginning of 2015. Initially the head of the team, who was also a member of the ARITHMUS Advisory Board (each NSI that participated in the research project had one or two representatives on an Advisory Board that convened annually during the first three years of the project), helped to set up the interviews, but later in the research I opted for setting up the interviews independently via email. Many of the members of the Big Data team that I interviewed were working on projects involving a new data source deemed as Big Data, such as mobile phone data, club card data, web scraping data or social media data, and this was true also for most of the interviewees that were not on the Big Data team. In addition to this, I interviewed two persons from senior management and two statisticians that had worked with population registers over many decades. Three sampling criteria were important to me. First, I wanted to include people with both technical- (i.e. statistics and computer science) and less technical (i.e. sociology and social geography) backgrounds. Second, I wanted to hear the views of both managers and those working in the everyday production of statistics, and finally, I wanted to include interviewees form both older and younger generations of statisticians. In total, my sample included seventeen participants, two of which I interviewed twice. Of the total

sample, 70 per cent were male, 59 per cent had backgrounds in less technical fields, 70 per cent worked in the everyday production of statistics, and 53 per cent had worked at Statistics Finland for ten years or more (Table 1).

Table 1. List of research participants

Interviewee	Gender	E. background	Role	Years at SF
One	Male	Technical	Manager	More than ten
Two	Male	Technical	Statistician	Less than ten
Three	Male	Less technical	Statistician	Less than ten
Four	Male	Less technical	Manager	More than ten
Five	Female	Less technical	Statistician	More than ten
Six	Male	Less technical	Statistician	More than ten
Seven	Male	Less technical	Manager	More than ten
Eight	Male	Technical	Statistician	Less than ten
Nine	Female	Less technical	Statistician	More than ten
Ten	Male	Technical	Statistician	Less than ten
Eleven	Female	Technical	Manager	Less than ten
Twelve	Female	Less technical	Statistician	Less than ten
Thirteen	Female	Less technical	Statistician	More than ten
Fourteen	Male	Technical	Manager	More than ten
Fifteen	Male	Less technical	Manager	More than ten
Sixteen	Male	Technical	Statistician	Less than ten
Seventeen	Male	Less technical	Statistician	Less than ten
Total	70% Male	59% Less technical	65% Stats	53% More than ten

All of the interviews were conducted at the premises of Statistics Finland and they lasted from between forty minutes to one and a half hours. Access to this field site would have been very difficult, if not impossible, without my involvement in the ARITHMUS project. Having the chance to witness from the inside how a government institution addresses a new challenge such as Big Data was a privileged position for a doctoral researcher to be in. My membership on the ARITHMUS team also granted me a certain prestige in the eyes of the participants as I was

usually introduced as “a researcher from London” rather than a PhD student. Especially the earlier interviews were characterised by a nervousness both on the part of the interviewer and the interviewees. In many cases I did not get the feeling that the interviewees necessarily wanted to speak to me that much, this being partially at least a consequence of the relative newness of the topic. Although the interviewees were experts in the field of statistics, Big Data was a new thing for them also. However, as the interviews progressed I became more confident in my ability to conduct them, and I think that this made the interviews more comfortable for the interviewees also.

After I had transcribed the interviews verbatim, I conducted a thematic analysis (Boyatzis, 1998) on the material with the assistance of the qualitative data analysis software NVivo (Bazeley and Jackson, 2013). Since I expected to derive thematic codes from both the research questions and related theoretical framework, as well as the themes that would emerge more spontaneously during coding, I opted for a hybrid approach applying both deductive and inductive codes. When translating the interview quotes from Finnish to English for the chapter, I aimed at translating their meaning rather than literal form.

Rapley (2001), amongst others, advises that when analysing interview data, one should dismiss the idea of gaining access to the intimate interior of a person, and to focus instead on what the interviews contain in terms of performativity and discourse. In an organisational setting a further challenge is to recognize when informants use institutional language and to find ways of moving the discussion beyond pre-prepared statements (DeVault and McCoy, 2001). With these considerations in mind, I tried to construct the interview questions and contexts in such a way as to ensure that participants were interviewed as individuals, rather than as spokespersons for the organisation (Thomas, 1995). I hoped for example that by focusing on educational backgrounds and career trajectories at the beginning of the interviews, I

could encourage the interviewees to reflect on Big Data from the viewpoint of their personal experiences of studying and working in data intensive fields.

Despite these precautions, I suspect that the material might be skewed towards the more positive aspects of Big Data. The interviewees were for example usually keener to emphasise the opportunities with Big Data, rather than to focus on the threats, which I suspect was partially at least a result of them being conscious of the fact that they were also representing their institution in their statements. The increasing role of the private sector and the actions necessitated by it were also painted in a more positive light than what I was expecting. Within the remit of this study it was not possible to account for a potential skew by for example triangulating the interview data with data obtained by other means. My meeting observations were few in number and the findings derived from them did not differ markedly from the interviews.

In sum, I understand the interviews as performed conversations between me, a researcher immersed in the literature around Big Data and the sociology of statistics, and the interviewees, experts immersed in their professional discourses but interested in my research topic. In general, I follow an understanding of qualitative research where its purpose is to describe and illuminate new social phenomenon, rather than to test pre-defined hypothesis or discern casual relationships between factors behind them (Silverman, 2013).

5.3 Findings and analysis

5.3.1 The end of a near monopoly and continuity as well as disruption: How statisticians understand Big Data

5.3.1.1 A meaningful concept or just hype?

Practically all of the interviewees were familiar with the term Big Data, and the vast majority, whilst acknowledging the huge amount of hype and ambiguity surrounding it, saw it as a meaningful and important concept in the context of official statistics. In contrast to the discourse that often focuses on the technical qualities of Big Data, such as the 3vs, for my interviewees the most critical aspect of Big Data was the increasing competition between data producers that its emergence entailed. One manager for example explained to me that the hype around Big Data was here to stay, and that the increasing competition brought about by Big Data meant that also Statistics Finland had to follow it closely:

> *What I always say about the hype surrounding Big Data is that it is hype that is here to stay. While it is of course true that we have always dealt with data that is "big", and that volume is something that grows exponentially, the important difference today is that an increasing number of organisations are obtaining both data and the means to analyse it. It follows from this that also we must monitor the phenomenon closely.* (Interviewee one)

The urgency of this task was amplified by the fact that the use of Big Data was considered to already be the norm in large parts of the private sector. Importantly, the hype surrounding Big Data was viewed not only as something to be approached critically, but also as something potentially beneficial for official statistics. In the current climate where people are more and more reluctant to respond to traditional data gathering methods, such as surveys, the hype around Big Data and the

resulting increase in public awareness around the concept was seen to increase Statistics Finland's chances of obtaining access to new data sources. Furthermore, in terms of the hype surrounding it, Big Data was not considered all that different from previous trendy concepts of bygone years.

Despite agreeing on the importance of the concept, many interviewees raised the issue of its vagueness and questioned whether anyone actually had a clear idea of what it constituted in practical terms. One statistician who had done a long career working with register-based statistics confirmed her familiarity with the term but noted that similarly to Open Data, there was much ambiguity surrounding it and that different people had "different ideas about what it actually constitutes". This ambiguity was thought to be particularly prevalent at international meetings, where concepts such as "The Data Revolution" and "Big Data" were regularly being promoted as "answers to all our problems". Such discussions at international meetings were usually accompanied by a dearth of practical examples, a situation that had however somewhat improved more recently. Some interviewees also considered the biggest hype around Big Data to already have passed, with the term itself having been replaced by ones perceived as more meaningful, such as Machine Learning (the application of artificial intelligence techniques to analysing data) and the Internet of Things (the embedding of the built environment and other physical objects with network connectivity).

5.3.1.2 A threat or an opportunity?

When asked whether Big Data represented a threat or an opportunity for official statistics, interviewees were initially keener to emphasise the opportunities. A statistician with a highly technical background expressed his belief that due to their different nature, both Big Data and official statistics were needed in the future. Whereas the emphasis with Big Data was on timeliness, official statistics had the advantage of

having been validated (for example by having been compared to data collected at a previous point in time). He explained to me that because of this, official statistics would not be replaced by Big Data, but rather, the two would work in a complementary way:

> *I do see it as more of an opportunity. After all, I do not believe that it removes the need for official statistics, because at least in the Big Data research that is currently being conducted the emphasis is usually on timeliness, which means that because the data has not been validated, it cannot be as reliable. You therefore need to supplement Big Data with official statistics, as well as the other way around.* (Interviewee two)

Indeed, Big Data was usually seen as a supplement, rather than a replacement for official statistics. According to some, this was because in contrast to the private sector, where the focus was on analysis and forecasting, in official statistics the focus was on producing data for others to use. Many also saw in Big Data an opportunity to tackle challenges faced by official statistics, such as increasing budget constraints and dropping survey response rates. One interviewee for example explained that for him Big Data reflected a broader change in official statistics where Statistics Finland needed to shift its focus from producing data to exploring what data already existed and finding out ways of accessing it:

> *I see it as an opportunity because for me at least there is a clear trend in official statistics. Whereas previously we could obtain missing data by devising a survey, today this is becoming more and more difficult first, because of increasing budget constraints, and second, because people are less and less willing to respond to them. This means that our focus has got to shift towards exploring what information already exists and figuring out ways of gaining access to it. (*Interviewee three)

Furthermore, by helping to account for new social phenomenon, such as the digital economy, Big Data was also seen as an opportunity to ensure

the continued relevancy of official statistics, something perceived as increasingly important in the current circumstances where National Statistical Institutes were no longer able to take their near monopoly over official statistics for granted. And with the new research methods associated with Big Data came also the opportunity to analyse population registers in novel ways. In addition to guaranteeing an exceptional level of combinability between different registers, the existence of a unique identifier for each Finnish citizen in the form of a personal identity number meant that in theory at least population registers could be treated as Big Data. A statistician with a highly technical background explained:

> *If you know how to link the data it should be perfectly possible to infer similar things from population registers as for example from internet search query data. So for example, although you might not be able to get a direct answer, you could infer that because a person has not been active in this or that register it is very likely that she can be found in this or that one instead. So really there are a lot of opportunities outside of the hype as well.* (Interviewee two)

Despite preferring initially to emphasise the opportunities, respondents were keenly aware that unless attended to, Big Data could end up undermining the role of official statistics. A senior manager explained that were Statistics Finland to fail to attend to Big Data, there were plenty of other actors that would take advantage of it, potentially undermining the role of official statistics:

> *The threat [with Big Data] is that if nothing is done, there are other actors both within the private and the public sector that will take their share of it, and because it offers a practically endless source of data, it could end up substituting the statistics that we produce.* (Interviewee four)

According to many, in the worst-case scenario the emergence of new competitors could lead to the question being raised whether

governmentally produced statistics were needed in the first place. This the interviewees saw as a potentially dangerous development because few other institutions took in their eyes methodological rigour and comparability of data seriously. Ultimately, the development could according to them endanger the continuing need to assess the validity and reliability of the numbers that circulated in the public, as recently demonstrated in the discussion around a “post-fact” era in politics.

Table 2. Opportunities and threats of Big Data for Official Statistics

Opportunities	Threats
• Hype increases awareness of statistics and is therefore useful when negotiating for access to new data sources • Produce more timely outputs • Supplement existing data sources • Cut costs • Compensate for dropping survey response rates • Ensure continued relevancy of statistics by capturing social phenomenon currently not accounted for • Obtain novel insights from population registers with new methods	• Becoming redundant due to increasing competition particularly from the private sector • Losing jurisdiction as validator of official numbers thereby contributing to a “post-fact” era in politics • Losing grasp of how data gets generated in the first instance

Table 2 summarises the opportunities and threats of Big Data for Official Statistics identified by the interviewees. It highlights one of the main arguments of this thesis, that rather than being a settled thing, there are numerous ways of thinking about Big Data and what its implications are, not just between, but also within professional fields such as official statistics. Whereas the aforementioned largely mirror points previously raised in the literature around Big Data and Official Statistics (Kitchin, 2015), thus highlighting the transnational character of official statistics,

much more ambiguity surrounded the question whether Big Data constituted a clear break from past ways of dealing with data.

5.3.1.3 "Just another register" or a new data paradigm entirely?

Much of the initial confusion around the concept of Big Data was related to the question of whether register based statistics constituted a part of it. When asked whether, as is regularly put forward in the hype, Big Data signalled a revolution in how data for statistics were being produced, many interviewees noted their hesitation. A statistician with a long experience of register based statistics for example explained to me that:

> *I am not so sure about it because often when I hear these arguments it makes me think that we here at Statistics Finland have been dealing with Big Data for quite some time now.* (Interviewee five)

Although most interviewees initially argued that register-based statistics did not constitute Big Data, particularly those that had worked in projects involving a Big Data source, such as club card data and mobile phone data, noted the many similarities between the two. Notwithstanding their unprecedented volume, the data sources in question did for example not differ much from governmental registers in terms of their form. According to one interviewee, Big Data was "just another register", this time however owned by a private company:

> *One the one hand there is nothing new in this, because also until now we have produced our statistics largely from registers obtained from elsewhere. One way to see it [the Big Data source in question] is to think of it as just another register, this time however owned by a private company. In this sense the work that we do here in this project is not that different from what we did earlier.* (Interviewee three)

As I covered in more detail in the previous chapter, the popular discourse around Big Data often posits that whereas previously, the production of statistics began with a research question followed by data gathering and

analysis, in an age of Big Data the starting point would be to identify what data already existed followed by a consideration of the questions that could be answered with it. In contrast to this, some interviewees questioned whether the so called traditional model had ever been followed and maintained that their approach had instead always been more of a mixture. A statistician whose career in official statistics spanned many decades reflected on this particularly after having attended the Big Data team's meetings:

> *Especially since I have started participating in the [Big Data] working group it has become clear to me that we have never really followed the traditional model all that consistently. Instead, it has always been the same as here [with Big Data] that we have started with a data source which we have then supplemented with data from somewhere else.* (Interviewee six)

Consequently, many of the challenges with Big Data bore resemblance to the challenges with register-based statistics. The fact that Big Data, like population registers, were originally created for purposes other than official statistics, had at least two important consequences. First, similarly to population registers, the way Big Data was produced and stored might undergo abrupt changes. Whereas a government department might have changed the way it filed information in a register as a result of a legislative change, a Big Data company could change its business model or go bankrupt. In both cases, Statistics Finland would have to look to other sources of data as replacement.

Second, because the data had in both cases originally been produced for purposes other than official statistics, it had to be cleaned, conjoined and modified before it could be used. Alastalo (2009a: 183) defines the production of register-based statistics as a process in which information produced in local administrative practices gets transformed in to statistical "facts" and according to my respondents, the process with Big Data was not all that different. One senior manager for example

explained to me that although with Big Data came the risk that Statistics Finland could lose its grasp on how the data got generated in the first instance, the challenge had existed already with population registers. The manager elaborated on the many similarities between Big Data and population registers:

> *Yes there is a risk of that [losing grip over how the data gets produced] but it does not really differ from what went on previously. It is a very similar process to how we for example ensure that something that the tax office collects is meaningful for the "real world" out there. The data structure of the tax administration is dictated by a well-defined legal framework, and what we do is that we fit that framework to official statistics. It is basically a classification process that has taken place over a very long period of time, so since ancient times really we have started to look at what they have available and began negotiating with them about what they will deliver to us. After that, we have processed the data further, perhaps combining and linking some of its elements. All of this is currently happening in relation Big Data.* (Interviewee seven)

Yet, despite the similarities, Big Data nevertheless signalled for the manager a fundamental shift in the production of data for official statistics. Rather than an abrupt revolution however, Big Data was instead a continuation of processes long in development. The manager, whose career in official statistics spanned many decades and who was now involved in setting up a Big Data infrastructure for Statistics Finland, reminisced that digitisation had been ongoing at Statistics Finland at least since the eighties, and that more recently even the last paper based processes had disappeared. The final remaining survey studies used a wealth register data as their base and systems had become more and more integrated. Still, Big Data was a major change, one resembling the move to the registers, with the difference that this time it centred upon the private sector:

> *If one thinks about the statistics profession in general it is a rather*

> *peculiar profession in the sense that already when I started here at the end of the eighties the house was much further digitised than many other places. Recently even the last paper based processes have disappeared and whereas previously our production consisted of multiple smaller studies, now everything is becoming more integrated and even the few remaining survey studies that we conduct use a wealth of register data as their base. So the process has been moving in that direction [increasingly digitised and integrated] for a long time. Nevertheless, the emergence of Big Data is major change, perhaps a rather similar one to when we started using registers as our main data source. I am inclined to see it as a similar shift as the move to registers, with the difference that this time it centres upon the private sector.* (Interviewee seven)

In sum, and in contrast to the popular discussion on Big Data, which often locates its novelty in its technical qualities, such as the 3vs, my respondents, although by no means denying the importance of the technical aspects, were nevertheless more inclined to see the newness of Big Data in how the production of data was becoming organised. More specifically, in the new circumstances more and more institutions had the capacity to produce and analyse data, which could potentially end up undermining the role of official statistics. In sum, my respondents identified in Big Data the end to the near monopoly on data concerning whole populations, economies and territories long held by statistical agencies.

Second, in contrast to the hype that posits Big Data as an abrupt break from previous ways of producing and analysing data, my respondents identified many similarities between old and new forms. Importantly, many of the challenges with Big Data were similar to ones encountered previously with population registers. Long before the emergence of Big Data sources, Statistics Finland's production had relied on data produced elsewhere, and accessing it had meant similar negotiations and arrangements that were currently necessitated by Big Data.

However, the new negotiations would no longer take place within the public sector, but increasingly, Statistics Finland would have to negotiate with private actors for access to data. As I indicated already in my conceptual analysis of Big Data in the previous chapter, it is here that I identify the most central aspect of the "data revolution" currently under way. In other words, Big Data reflects the broader processes of neoliberalisation that have had a major influence on cultures and economies in the past few decades in that it is increasingly produced and accumulated in the private rather than the public sector. And as I suggested already in chapter three, the increasing centrality of private actors in the production of data suggests a historical break from its centralised production by states.

In the next section I interrogate in detail ongoing Big Data projects at Statistics Finland at the time of the research and suggest that they are indicative of a neoliberalisation of data infrastructures of processing, analysis and storage in at least two ways.

5.3.2 A neoliberalisation of data infrastructures? Big Data projects at Statistics Finland

5.3.2.1 Big Data projects at Statistics Finland at the time of the research

At the time of the research, Statistics Finland was examining the feasibility of multiple Big Data sources, of which perhaps the most promising ones were club card data, mobile phone data, web scraper data and POS terminal data. In the period in which the research was conducted, a heated debate was ongoing in Finland about the size of the public sector amidst a prolonged economic recession. For the first time in its history, Statistics Finland had been forced to undergo layoffs, and its budget had been reduced for some years in a row. Following a relevancy assessment, which had been initiated also as a result of

outside pressure, multiple statistics had been discontinued. This background perhaps helps to explain why the motivation behind all of the Big Data projects initiated was to some extent at least connected to the question of whether costs could be saved.

Important to note is also that developments at Statistics Finland were taking place within a transnational field of statistical practices (Scheel et al., 2016). As I highlighted already in chapter two, international collaboration in official statistics has long historical roots (Westergaard, 1932). However, especially since the founding of the European Union, collaboration between statistical agencies in Europe has become even more central. The majority of the Big Data projects at Statistics Finland were partially at least funded by the statistical office of the European Union, Eurostat, and similar experiments were taking place at other NSIs across Europe. In relation to Big Data, Statistics Finland was in fact a late comer in comparison to some other statistical institutes, most notably CBS in the Netherlands. Statisticians from different countries convened regularly to share experiences and to work collaboratively in projects funded by Eurostat.

The stated motif behind the examination in to club card data at Statistics Finland was the rapidly falling response rates of traditional data gathering methods. Until now, the only way to gain information about people's consumption habits had been to devise a laborious household budget survey which had required respondents to keep a diary of their consumption habits over a period of two weeks. Despite a change introduced at the start of the new millennium whereby instead of keeping diaries, participants only needed to keep the receipts of their shopping, response rates had fallen to a critical level. Especially since retail in Finland was so heavily concentrated (two chains controlled around 80 % of the market), club cards were seen to offer a potentially rich data source concerning people's consumption habits. At the time of the

research, Statistics Finland had made preliminary inquiries to the two biggest retailers about accessing their customer data.

The project examining mobile phone data had been initiated by Eurostat, who, following the example of Estonia, where an arrangement had been put in place whereby a private company obtained the data from mobile phone operators and distributed it forward to other institutions, including the National Statistical Institute, had wanted to examine the feasibility of accessing mobile phone data in other countries too. Statistics Finland approached the data source especially from the viewpoint of tourism statistics, where traditional data gathering methods were proving increasingly inefficient. Whereas before it had been possible to interview people at the borders, the increasingly borderless nature of the EU had made the undertaking much more difficult. Since almost everyone now carried a mobile phone, the data collected by mobile phone operators offered in theory a near full enumeration of travel patterns between and within countries. At the time of the research, Statistics Finland had initiated a dialogue with the largest mobile phone operators in Finland about data access.

Of all the ongoing Big Data projects at Statistics Finland, the joint project examining web scraper data and POS terminal data was the closest to moving in to production. This project too was made possible by Eurostat, who had provided funding for a period of one and a half years with the hope that it would contribute towards automating the production process. The first part of the project explored whether online store data retrieved with a web scraper could be used to produce the producer price indices for services and the second whether POS terminal data could be used for the consumer price index. The construction of the web scraper and associated classification machine were outsourced to a private IT company after a call for bids. Statistics Finland had successfully completed both parts of the project and tests were ongoing whether the statistics produced could be moved in to production. Notably, in relation

to POS terminal data, one company had agreed to act as a permanent data provider for Statistics Finland.

5.3.2.2 Legal, technical, and organisational challenges in accessing data

The challenges in accessing Big Data for Statistics Finland can be divided between the legal, technical and organisational. In practice, the three typically intertwined and interlaced in important ways, which is something I will describe in more detail below. Rather than covering all the problems in each project, I will give examples from each that I think typify one of the different problem areas. I will conclude by suggesting that many of these challenges reflect an increasing neoliberalisation of data infrastructures of processing, analysis and storage.

Perhaps the biggest factor hindering Statistics Finland's access to Big Data was the legal framework in which both it, and the companies providing the data, operated. Although Statistics Finland had extensive legal powers to obtain data from companies about their own operations, this right did not extend to customer data. Nor did the companies usually have the right to pass the data, at least in the rather detailed format usually required by Statistics Finland. Unlike is the case with multinational internet conglomerates, typically the terms of agreement of the Finnish companies in question did not include the right to pass the data on to a third party. A respondent that had participated in the project examining the use of mobile phone data explained that at the heart of the legal conundrum was the question of what constituted personal data:

> *When we requested an anonymised sample for experimental purposes from the [mobile phone] operators they got back to us saying that we needed to get in touch with the Data Protection Ombudsman because they themselves weren't allowed to touch the data. We then got a response from the Ombudsman saying that according to the Personal Data Act, personal data cannot be used*

> *without the person's consent. The question then becomes whether a piece of data should be deemed personal data or not. So for example a travel pattern, where you can see that a person has arrived from that place, gone to that other place, spent two weeks in the country and then left, is that personal data? Even though there is no personal identifier in the data, such as a personal identity number, in the Data Protection Ombudsman's interpretation the possibility of indirect identification could not be ruled out, and therefore it fell within the Personal Data Act. This means that under the current legislation, in order to use the data we would need the consent of everyone who uses a mobile phone.* (Interviewee three)

The prospect of data sharing was, therefore, a new thing for the companies also. For many respondents the biggest challenge with Big Data were therefore not the technical issues related to it, but rather the question of the appropriate legal framework for its utilisation. The respondent cited above explained to me that in Finland concerns around using data of such high precision ran deep, and that he did not consider it at all self evident that Statistics Finland would be given a statutory right to access it:

> *For me, technical issues are not the biggest challenge, but instead the legal framework and its interpretation. Based also on my discussions with the Data Protection Ombudsman the concern around using data of this level of precision run deep. Is society ready for it or does it violate some very basic rights that we have? I do not think it is at all clear that a law will be passed that will give us the right to see just about anyone's location (even though we are only interested in the big picture). At least it will not be done just like that without discussing it first.* (Interviewee three)

He speculated that one of the reasons why Estonia had been successful in putting in place a working model for data sharing was the relatively strong mandate under which its NSI operated. In contrast to Statistics Finland, Statistics Estonia had the legal right to request third party data

from companies. Furthermore, even though the broad legislative framework for data sharing was set by the EU, in practice each member country interpreted it differently, some more loosely and some more strictly.

The question of whether and under what conditions companies were allowed to pass their data was further complicated by the different organisational structures of different companies. This point was well demonstrated in the project examining club card data, where, although only two companies dominated the market in Finland, they did so with very different organisational structures. Whereas one of them operated with a centralised structure, meaning that decisions regarding individual stores were made from the headquarters, the other built its operations around a franchising model, where local storekeepers enjoyed relative autonomy. In relation to the question of who owned the club card data, in the first case the answer was clear, but in the second less so. In theory at least, local storekeepers owned their customer data, meaning that Statistics Finland would have to negotiate with each of them separately for data access. The one instance where a company had agreed to act as a permanent data provider for Statistics Finland had been made possible by the centralised structure of the company and the fact that Statistics Finland had only asked for sales data which did not include information about customers per se.

A further challenge related to the organisational aspect was the multitude of data collection practices that existed between companies, a point also best demonstrated in the project examining club card data. In order for it to be useful for Statistics Finland, data needed to be collected in a format where individual purchases were recorded at the product level. At the time of the research however, only one of the potential data providers collected data at the required level of granularity, and even then there was no guarantee that other technical specifications required by Statistics Finland would be met. In contrast to the establishment of

the register based statistical system, where Statistics Finland had initially at least had a say in the content of the registers (Alastalo, 2009a: 179), in relation to companies' data collection practices Statistics Finland enjoyed no such authority.

Finally, after all the aforementioned challenges remained the question of how the data sharing and processing would be organised in practice. The question was made all the more important by the fact that Big Data did in practice require expertise that Statistics Finland did not possess. In the project examining web scraper and POS terminal data Statistics Finland outsourced both the construction of the web scraper and the system that processed the data at Statistics Finland's end to a private company because, as one respondent put it, "if we talk about daily or even weekly data, our systems are very quickly on their knees." The respondent admitted openly that although part of the budget had from the start been allocated to consulting the private sector, its role had become more important than initially planned due to the increasing budgetary constraints under which Statistics Finland operated. He explained:

> *These acquisitions [from the private sector] have been part of the plan all along, but what has happened since is that as I already mentioned before because we don't have that many resources for development at the moment we have not been able to recruit the manpower needed because even if you are offered the money to hire someone from within [the] house, the basic tasks of a statistical unit are what they are, meaning that everyone have their normal responsibilities to think about as well. In practice therefore we have been running the project with incomplete staffing as a result of which the role of outside actors is very likely to grow.* (Interviewee eight)

In relation to this particular project, one private consultant especially had expressed a desire to offer Statistics Finland Big Data solutions beyond what was needed in the first instance. The respondent interpreted this

as a sign that the company aspired to gain a foothold in the larger statistics market:

> *In contrast to the other service providers, one company has talked specifically about wanting to offer a Big Data solution to us. My impression is that were we to go in that direction [building production around a Big Data infrastructure] this company would be eager to take part in the discussion. It would certainly be an advantage for them if at that point they could refer to already having successfully completed a trial.* (Interviewee eight)

However, were a private company to provide Statistics Finland its Big Data infrastructure, it would raise the question of where the data would be stored. Whereas Statistics Finland's in house server capabilities were insufficient for Big Data, data storage in the cloud was deemed as a serious risk to data protection and confidentiality, the building blocks of Statistics Finland's public mandate. According to one respondent, Statistics Finland's attitude towards data storage was "better safe than sorry":

> *The attitude here could be defined as "better safe than sorry". The traditional perception has been that as long as the data exists in one format or the other within the walls of Statistics Finland it is safe, but as soon as it leaves the house it's jeopardised. More recently this attitude has relaxed perhaps a little bit but at the end of the day it's still pretty close to that.* (Interviewee eight)

At the time of the research the conundrum regarding data storage remained unresolved. For many respondents, the most obvious solution was the construction of in house Big Data capabilities at Statistics Finland. Companies, on the other hand, were aware of the issue and had begun offering Big Data solutions with servers located at the premises of the customer.

As previously mentioned, despite all these challenges, Statistics Finland had successfully obtained access to one Big Data source. A senior manager explained that what had made this possible had been the outsourcing of the data aggregation to the company that produced the data:

> *What we are doing here is that we are not even thinking about obtaining the detailed data, but instead we ask them to deliver daily or monthly averages according to their own product classifications. This we way do not at this point need to worry about the computational problems related to Big Data since we've subcontract the data aggregation to them.* (Interviewee seven)

The manager did admit that this was far from ideal, and that it would bring to the fore the question of whether Statistics Finland had a sufficient grasp on how the data was produced:

> *In relation to mobile phone data we are more interested in having access to the micro level data, but since the current legislation states that only the operators can access it we are forced to specify the request to them. This again is problematic in the sense that we would need to have an understanding of the type of data that is generated to the operators and it's a question in itself what can be digged out from it.* (Interviewee seven)

In addition to the fact that Big Data is primarily produced and accumulated in the private, rather than the public sector, the material above indicates an increasing neoliberalisation of data infrastructures of processing, analysis and storage in at least two ways. First, in midst of public sector cuts Statistics Finland was hard pressed to find resources to adequately tackle Big Data, which was likely to result in a more central role for private actors than originally planned. Perhaps in connection to this, some private actors had begun offering infrastructural solutions beyond mere technical consultancy, raising the question whether Statistics Finland would at some point have to store its data on private

servers that existed outside of its premises. Though the cuts were not historically unprecedented, nor do I want to assess their necessity in this instance, they have taken place in a political moment where economic downturns are also in former strongholds of the Nordic welfare state model, such as Finland, regularly attributed to inefficiencies in the public sector. Just to give one example of this tendency, the current Finnish prime minister has recently stated that Finland has not hundreds, or even thousands, but indeed tens of thousands of civil servants too many (*Helsingin Sanomat*, 2014). Therefore, if current political trends continue, and the funding of the public sector reduced due to it being perceived as primarily a burden for the economy, the role of private actors in managing Big Data will possibly increase. Big Data therefore represents a neoliberalisation of data not just because it accumulates mainly in the private, rather than the public sector, but because the increasing resource demands brought about by it is potentially resulting in a more central role for private actors also at the operations of the governments statistical agency.

Furthermore, I identify neoliberalisation also in the approach taken by Statistics Finland to securing permanent access to a data source. In order to tackle the computational issues of data aggregation and storage, which Statistics Finland was ill equipped to deal with, what had made permanent access possible had been the outsourcing of the majority of data aggregation and management to the company providing the data. Rather than getting 'raw' data, Statistics Finland was content with being handed averages based on the company's own product classification schemes. Alastalo (2009b) argues that because the majority of official statistics in Finland are derived from data produced in registers, the largely black boxed governmental practices according to which they are updated and maintained have had a major influence on the content of official statistics in Finland. Building from this, I suggest that the increasing role of the private sector in the production and infrastructures of data processing and analysis means that in the future

the content of official statistics may increasingly be influenced, not by the rationalities of government departments, but by those of private corporations. What exactly these rationalities are demands further research, but since companies operate in competitive markets, they are likely to be connected to the profit motive. This finding is in line with Thatcher's (2014) argument that with Big Data comes the risk that the making of knowledge will be set through capitalist imperatives.

In order not to exaggerate these findings for the sake of making an argument, it is important to point out that the Big Data projects at Statistics Finland were all still experiments, and that no data produced for traditional statistics had so far been replaced. Despite recognising its potential, interviewees usually saw Big Data as a supplement, rather than a replacement for existing statistics. Therefore, in its current form Big Data did not mount a serious challenge to the register based data used for official statistics. Nevertheless, the findings covered here indicate that the increasing production of data in the private sector is changing ways of thinking within statistical institutes including their roles not only in the production of data for statistics but the infrastructures of its processing, analysis and storage. And as noted previously, these developments seem to indicate a historical break in that data production is moving away from its basis in states.

By exploring further practices and responses prompted by Big Data, in the next section I suggest that neoliberalisation can be identified also in the ways that Statistics Finland imagines its future role and relevancy.

5.3.3 A neoliberalisation of occupational culture? Tackling the challenge of the private

5.3.3.1 A vision for the future: A trusted gatekeeper and public expert?

In spite of the increasing competition and impeding loss of near monopoly on data concerning whole populations, economies and territories, all interviewees expressed a belief in the continued relevancy of NSIs. Trust and neutrality were identified as key competitive factors in the new circumstances. One statistician for example explained to me that Statistics Finland's advantage was in being recognised as an official actor, and that due to their extensive experience of handling sensitive data in the past, NSIs would continue in a similar role in the future:

> *Our advantage is in our status as an official actor. If I think about it from a citizen's point of view, I would not want just anyone to access and manage the Big Data about me. It seems to me quite natural therefore that since NSIs have already until now managed very sensitive data, they will continue to do so also in the era of the new data sources.* (Interviewee four)

Many interviewees suggested that whereas previously the role of the NSI was to gather, produce, analyse and publish statistics, in the future its role would be reduced to publishing data as Open Data for others to use. In this role, neutrality would be key, and NSIs would be well positioned to ensure that everyone, from big banks to news agencies to individual citizens, would gain access to new data at the same time. Many questioned whether a private business such as Google, whose existence depended on making a profit, could be trusted with this role. It was however acknowledged that ultimately it would depend on politicians and voters whether such a responsibility would be handed to a private enterprise.

Some felt that Statistics Finland needed to do more to actively seek a role as a public expert in the use of Big Data. A statistician explained that in her vision for the future Statistics Finland would more courageously comment on misinterpretations of data of which there already circulated numerous examples of:

> *It would be great if our role as experts was emphasised more in the future, that we would have the skills, knowledge and confidence to comment on some of the interpretations [of data] that circulate in the public. There are so many misinterpretations around and we have traditionally been rather shy to comment on them.* (Interviewee nine)

In this new role, a skilful and active use of social media was considered to be paramount. In contrast to previously, where statisticians had had plenty of time to double check with colleagues facts and figures before sending their replies to citizens and other stakeholders, in the fast paced world of social media statisticians would increasingly need to rely solely on their own quick assessments in their communications.

5.3.3.2 Conditions for continued relevancy: New skills and a new mentality towards work

Although Big Data was widely considered to demand new skills from statisticians, many preferred not to speculate about their exact form, since these could only be specified after access to more Big Data sources had been secured. A manager emphasised the primacy of data access but noted that the skill set of a “data scientist” would be increasingly sought after in the future:

> *In principle they can be defined [the skills demanded by Big Data]. A generally accepted definition is that of the data scientist, meaning for example that programming skills will be in high demand. What I would say however is that because a lot of know how already exists at NSIs, the first step is to obtain the data. Only when you have the data can you start asking what kinds of expertise you might need.* (Interviewee

one)

Despite not wanting to speculate about their exact form, many acknowledged that the skill demands were changing. One statistician for example explained that contrary to received wisdom, only a minority of statisticians at Statistics Finland had a background in very technical fields, such as statistics and mathematics, and that already with the current tools available, more technical expertise was needed. Advanced data analysis skills in combination with a deep understanding of statistical theory were perceived as increasingly important as a result of an impeding automation of the more mundane tasks, such as data gathering.

According especially to the younger interviewees, Statistics Finland's future depended on whether its staff was able to adopt a new mentality towards its work. Many saw Statistics Finland as an institution rarely at the frontline of new developments, and with an aging staff very set in its old ways of doing things. The younger interviewees explained that in an age of Big Data, however, where things kept moving faster and faster, no institution or team could afford to stay put and do things the way they had always been done. Instead, when encountering problems, they would have to branch out to other institutions and teams and inquire how they had addressed them. On the other hand, new developments such as Open Data were seen to underscore the importance of actively and successfully campaigning for government funds. Whereas previously Statistics Finland had charged for much of the data it provided, the increasing demands for data to be made available publicly free of charge meant that that revenue would need to be generated from somewhere else, in practice often meaning the government.

A fear expressed by many was that if the public sector failed to present itself as an attractive employer for highly skilled workers, the most innovative work around Big Data would be conducted somewhere else,

often meaning the private sector. And since the private sector was usually able to offer higher wages, the public sector had to figure out other ways to attract employees. One interviewee explained that Big Data skills were highly unevenly distributed in society and that the public sector was left behind already because similarly to before when highly skilled employees would go to work in finance, today they would go to work in the technology companies:

> *One thing that worries me is just how unevenly distributed Big Data skills are in society. And similarly to before, when the best physicists and engineers would go to work on Wall Street, today they go to Google. The public sector is left behind already as a result of this.*
> (Interviewee ten)

Notably, however, when prompted about their reasons for working in the public sector, many interviewees, younger ones very much included, underscored their belief in the value and importance of public service. One statistician with a highly technical background for example explained to me that he believed there were functions in society that the private sector should not be entrusted with. According to him, private actors could for example not guarantee impartiality in the production of statistics in the same manner that a public sector institution could:

> *I do not think I would be working in the public sector, unless I believed that it had an important role to play also in the future. The argument against it often goes that because institutions in the private sector must generate profits in order to exist, they are bound to do things well, whereas the public sector don't have anything to worry about if they do not do such a good job. Personally, I do not believe in this argument, because I think that there are basic functions in society, including the production of official statistics, which cannot be left solely to private actors whose impartiality can easily be questioned.*
> (Interviewee eight)

One way to interpret the material above is to suggest that one response to the increasing competition brought about by Big Data is a perceived need, especially from the viewpoint of younger employees, to adopt private sector ways of thinking and acting also in the public sector. Rather than clinging on to the values of a slow paced bureaucracy, Statistics Finland's employees needed according to many younger interviewees especially to adopt the mentality of agile and fast paced organisations. Although not explicitly spelled out, descriptions of the desired attitude bore resemblance to those often associated with companies or start ups. In sum, whilst emphasising the importance of public service, for many the appropriate response to the increasing competition from the private sector was to become more like it.

What is also worth noting is that the data scientist, identified by one manager as the highly sought after employee of the future, is an occupational category that originates in the private sector. Although what exactly the skills and expertise that constitute a data scientist are is far from self evident, definitions often include business types of skills and attributes. Writing for the Forbes magazine, Hansen (2017) for example explains that data scientists "utilize their knowledge of statistics and modelling to convert data into actionable insights about everything from product development to customer retention to new business opportunities". Similarly Patil and Davenport (2012: 5), in another influential definition, refer to a recruiter at a data science company who explains that data scientists possess not only statistical or analytical capabilities, but also "certain habits of mind" meaning "a feel for business issues and empathy for customers". It is also notable that one of the authors of the piece, DJ Patil, who was appointed as the first U.S. Chief Data Scientist by president Obama, has a background not as a government statistician, but a consultant for private companies such as LinkedIn, PayPal, Ebay and Skype.

In sum, the findings above suggest that in an age of Big Data, not only data, but also valuations of professional skill and expertise in its analysis increasingly originate in the private sector. Data is being "neoliberalised" not just because it increasingly accumulates in the private sector, or because amidst public sector cuts the private sector will come to have an increasing role in the data infrastructures of the public sector, but also because the valuations of professional skill and expertise of the modern data professional originate from a market and profit oriented rationality.

5.3.3.3 Conditions for continued relevancy: A focus on partnerships

In addition to new skills and a new mentality towards work, Big Data was perceived to require increasing efforts in building and maintaining partnerships, not only within the public sector, as had largely been the case before, but now also outside of it. A manager explained to me that although partnerships were not a new thing for Statistics Finland, the fact that they would increasingly have to be formed with actors in the private sector represented a new challenge:

> *One crucial demand brought about by Big Data is the increasing need to build partnerships. This is in itself of course nothing new for us, but because so much of Big Data exists in the private sector, it necessitates completely new conventions and arrangements from us.* (Interviewee four)

The register based statistical system had depended on good relations between different government institutions, and the interviewees hoped that the tradition of openness and trust that had existed over many decades within the public sector in Finland would carry over to arrangements made with the private sector also.

Efforts in building partnerships were closely tied to efforts in lobbying for new legislation. As I pointed out earlier, a major obstacle in gaining

access to Big Data was the legal framework in which both Statistics Finland and the companies providing the data operated. Discussions were ongoing about the renewal of the Statistics Act in Finland, and in these discussions, Statistics Finland had put forward the wish that it's right to request data from companies would be extended to customer data also. One senior manager explained to me that Statistics Finland had lobbied for the new legislation at the Ministry of Finance, and although there had been other reasons for the legal review, such as a new EU data protection initiative that had needed to be harmonised with national legislation, Statistics Finland's lobbying efforts had played an important part. The preparatory work for the new legislation took place in a working committee where different interest groups, such as the Confederation of Industries, some unions, customs, and various state departments and institutions, including Statistics Finland, were represented. The manager explained the process to me in the following way:

> *The first thing that had to be accomplished was to get the Ministry of Finance to start the preparation of the new law. There were other reasons behind it too besides our new data access needs, such as the EU's new data protection directive which had to be harmonised with our national legislation, but we of course propagated it to the ministry and when they were ready to start preparing it we wanted it to include a review of our current data access rights. The way that the process works is that there is a working committee where different interest groups are represented, such as the Confederation of Industries, some unions, customs, other state departments etc. We are represented there by a professional lawyer and a couple of executives.* (Interviewee seven)

Furthermore, the manager revealed to me also that on the side of these efforts, he and the other members of the senior management team had devised, what he termed "the road show", where they visited the different

stakeholders explaining to them Statistics Finland's need for data access:

> *On the side of these efforts we have run a so called "road show" where me and a couple other managers have visited the senior management of the different stakeholders. In addition, we have also gone through some of the biggest companies and some state departments that we know have a mutual interest with us, if I may put it this way. Our point is that we would want an obligation to be created for them to deliver third party data to us. Our current thinking is that in many cases we will have ask people directly for the permission to use the data.* (Interviewee seven)

The manager described the "road show" as a sales operation in which Statistics Finland argued the case that their mission was to make better statistics in new ways, and that old ways were proving increasingly laborious, ineffective and expensive. As part of it, the senior manager and his colleagues explained to the managers of the companies that were the process to go smoothly, in exchange they would be able to offer better statistics also in relation to phenomenon that were of direct interest to them, such as the markets in which they operated. In addition to this, the manager explained that Statistics Finland did have the opportunity to compensate financially for the extra work required from the companies. So far however, Statistics Finland had not paid any meaningful sums as the data requests had been "fairly moderate" in size.

It is interesting to reflect on these findings in light of the historical review in chapter three. As I explained in more detail there, the review suggested that historically statistics have been produced from data produced by the dominant institution of its time. Whereas the earliest forms of statistics were compiled by the ruler and the church, their modern form became tied to the centralised authority of the nation state. I noted furthermore that within the latter arrangement, Statistics Finland's position of authority had witnessed ups and downs through the

years. When for example the register based statistical system was put in place after the Second World War, Statistics Finland had initially had a say in the content and upkeep of the governmental registers, but had lost this power when the registers became more standardised.

As a result of the emergence of Big Data, Statistics Finland increasingly finds itself as one data producer amongst many. In these new circumstances, data access requires the devising of marketing and bartering strategies, such as “the road show”. The extent to which these strategies differ from ones used in the past demands more research, but considering the commercial nature of the new negotiating partners it would be surprising if they did not differ at all. The manager’s framing of the “the road show” as a sales operation speaks volume of the increasingly business type context in which negotiations for data access must now take place. It is yet another example of private sector ways of thinking and doing things becoming more prevalent also in official statistics. And even though access has so far been free or relatively cheap, new dependencies are being created. In light of economic history it is certainly not inconceivable that after having made its production reliant on data produced by a private company, Statistics Finland would find that prices had started to rise.

I conclude the analysis in this chapter by interrogating companies’ attitudes towards data sharing, as perceived by my interviewees.

5.3.4 “Not a question of desire but skill”: Companies’ attitudes towards data sharing

As I explained in more detail in the previous chapter, a key part of the popular discourse on Big Data is the idea that data today constitutes an important form of economic currency. For Mayer-Schönberger and Cukier (2013: 182), for example, “Data is to the information society what fuel was to the industrial economy: the critical resource powering the

innovations that people rely on". Keeping in mind the value that data holds, one might expect companies to be hesitant to share their data unless financially compensated for it.

Perhaps surprisingly therefore, my interviewees did not mention this amongst the concerns that companies had expressed about data sharing. The interviewees were in general markedly more positive about companies' willingness to share their data than what I was expecting, one interviewee describing it as "not so much a question of desire, but having the skill to do so". The interviewee suspected that one of the reasons that had motivated a company to share its data even when not legally bound to do so was the good pre-existing relationship that it had with Statistics Finland. He explained that in Finland companies had many pre-existing legal obligations to supply data about their operations to Statistics Finland, and that in general they did not have a problem with this since in return they got better statistics with which to plan their operations. In his view, therefore, the relationship was not based on bartering, but on cooperation that benefitted both parts. He explained:

> *Firstly, there is the legal obligation to disclose information. So, a company of this size already has many pre-existing obligations to deliver data to us based on cooperation agreements that we have with them. Observed from the outside it looks as if we have a good relationship with them and that the collaboration works. In other words, they feel like they are getting something in return for what they provide to us. To my understanding the relationship is not based on bartering, but instead they regard the information we produce as useful for them, since they constantly use our data to plan their own operations. Other countries have agreed to produce extra reports to companies in exchange for data, but I'm not aware that we would have taken this route.* (Interviewee eight)

Another interviewee explained that although companies had expressed an awareness that they "sat on a goldmine" of data, they would most

likely not see a financial issue in sharing it since Statistics Finland's role was not to provide commercial products, but to describe society "from the high up". So although they recognised the potential financial value of the data they possessed, they would not have a problem sharing it especially once they had extracted the initial value from it by for example selling it to advertisers.

The manager who had participated in the lobbying efforts for data access, on his part, explained to me that modern companies took the notion of corporate social responsibility seriously, and that the CEOs of companies were usually more open to the prospect of data sharing than the interest groups representing them. The interest groups in particular were concerned about the extra work burden that the obligation to provide data would bring on the companies. Furthermore, because similarly to Statistics Finland the companies' existence depended on whether the public trusted them, they had concerns that by sharing their data they might make themselves vulnerable to publicity scandals. The manager cited one instance where a mobile phone operator had shared its data for research purposes, only to find that some of its major customers had ended their contracts having perceived it as a breach of trust.

The markedly positive attitude towards collaborations with the private sector documented in this section suggests that the hope expressed by many, that the tradition of openness and trust that had existed over many decades within the public sector in Finland would carry over to arrangements made with the private sector also, is not entirely unfounded. However, it is here in particular that I identify limitations in my sample. First, although I do not doubt the truthfulness of the views stated, I suspect that they might be skewed towards the more positive aspects of the collaborations with the private sector. Perhaps to some extent at least the positive narratives in these statements are part of the efforts to build partnerships with the new actors. Second, within the

scope of this study it was not possible to interview companies first hand. Therefore, in order to form a more complete picture of the emerging Big Data landscape, in a future study it would be important to identify and interview all the different stakeholders separately. Are for example companies as open to data sharing as suggested by the statisticians?

One factor that could suggest otherwise is the point about the need to maintain customer trust that I mentioned last. For although it might be true as the manager said that modern companies take social responsibility seriously, at the end of the day they operate in competitive markets and exist only as far as they are able to turn in a profit. Unlike is the case for a government department, profits are the very precondition of their existence. Therefore, were the desire to contribute positively to society by sharing their data to come in to conflict with the need to turn in profits, for example by losing customers as a perceived breach of trust as had happened to the mobile phone operator, quite likely the former would give in. As pointed out by for example Schroeder (2014), and as I covered in more detail in the previous chapter, the public anger raised by Facebook's uses of its data for research purposes suggests that companies are in the future more likely to conduct such work behind closed doors.

In contrast to the positive narratives, there is therefore much reason to suggest that the increasing involvement of the private sector in the production of data is far from unproblematic. Regardless of the extent of companies' desire to contribute to the common good by sharing the data they now produce, it is unlikely to take place, at least at the scale that some would like it to, should it come to conflict with the demands of markets. I therefore identify here, in the increasingly proprietary nature of data, a contradiction in how the assemblage producing official statistics is potentially being reconfigured.

5.4 Conclusion

In this chapter, I have explored the uptake of Big Data at Statistics Finland, and suggested that it is in multiple ways indicative of an increasing neoliberalisation of statistical practices. However, as noted previously, what Big Data will mean for the production of official statistics is by no means settled or given. The Big Data projects that I examined were all still in experimental phase, and no pre-existing statistics had so far been replaced by data produced in the private sector. Big Data was in general viewed as a complementary source of data, rather than a replacement for register based statistics. Nevertheless, the analysis provides early evidence of a potential reconfiguration between the public and the private sector in the production of data and the making of official statistics. Examined in the longer historical trajectory, it suggests a shift whereby the production of data for official statistics has started to move away from its historical basis in states. The analysis has furthermore highlighted some of the ways in which the statistical agency is changing its practices in relation to the perceived challenge of Big Data. I argued that they too reflect our current political culture in that private sector ways of thinking and acting are increasingly perceived as the appropriate organisational principle for public sector institutions also. The analysis in this chapter is therefore in line with the central conceptual point of this thesis: That statistics evolve in interaction with political contingencies and that changing governmental rationalities form an integral component of this “co-constitution”.

The analysis has furthermore underscored the importance of understanding Big Data as not just a technical-, but also a political object. How much influence over the data that comes to shape the making of official statistics and in turn the governing of societies should be given to non-state actors? How should the assemblage producing statistics be reconfigured in order to unleash the potential of Big Data for the

advancement of social goods? Some of the findings in this chapter suggest that there are direct ways in which private sector rationalities are in conflict with the desire to share data for the advancement of collective goods.

After I had written this chapter, I shared the results with the other researchers on the ARITHMUS team. Somewhat to my surprise, they informed me that the findings did not differ markedly from their own findings at their respective field sites. In the next and final chapter before the conclusion, therefore, I expand the analysis of Big Data in official statistics by situating my findings at Statistics Finland to the transnational field of statistics of which it is part.

6 Transcending methodological nationalism through an analysis of the ARITHMUS database

6.1 Introduction

Big Data developments at Statistics Finland were not happening in isolation, but rather, much like other work at a modern statistical agency, they took place as part of international relations involving organisations such as Eurostat, UNECE and others through which statisticians from different countries regularly engage and meet to share experiences and work collaboratively. One analytical pitfall would therefore be to treat the developments documented in the previous chapter as being unique to Statistics Finland and not part of these relations and connections.

By starting from the idea of a transnational field of statistical practices in which the local, the national and the transnational overlap and intersect, the ARITHMUS project of which this thesis is part, has sought to move beyond nationally bounded case studies, or what is sometimes termed methodological nationalism (Scheel et al., 2016). Instead of using nation states "as quasi naturally given units of research, analysis and theorisation" the project has approached "the meaning and force of the national" as an empirical question (Scheel et al., 2016: 4). By mobilising a "transversal" method, consisting of numerous research strategies, the project has sought to attend to how practices and discourses travel between and connect sites and scales. Importantly, the transnational field of statistical practices is understood as a field of struggle where "statisticians and other stakeholders (demographers, data scientists, domain specialists etc.) struggle over the devices, truth claims, budgets and methods involved in the production of official statistics in order to advance their relative position[s]" (Scheel et al., 2016: 10). However,

instead of aiming to do a comprehensive mapping of the relative positions of old and new stakeholders in the production of official statistics, the project mobilises the concept of a transnational field of statistical practice as an opening “to examine what kinds of orderings are being done through specific practices that traverse, connect and operate across various sites and scales.” (Scheel et al., 2016: 11).

One example of the strategies through which the project has sought to achieve its goal of moving beyond the national “container” is the approach I took after writing my previous chapter. As part of my method, I shared my findings with the other ARITHMUS team members who emphasised the similarities rather than differences between my findings and the things that they had discovered at their respective field sites. They pointed out to me that for the most part, the narratives that I documented were familiar to them from before. The outcome of the exercise thus underscored the importance of understanding the articulations of Big Data and its impact on official statistics that I discovered at Statistics Finland as very much part of and shaped by forces and dynamics, such as debates and practices, that cut across individual NSIs. As Grommé et al. (Forthcoming: 3) point out, however, this is not to assume that developments such as Big Data are taken up identically across sites. Rather, they may get their specific articulations, depending on for example the different histories and political circumstances in which the different NSIs operate. As I pointed out earlier, however, the ways in which national factors come to play a role in the uptake of developments such as Big Data at NSIs must be treated as an empirical question in need of further analysis.

In this final chapter, therefore, I expand on my analysis by situating the findings of my previous chapter in a transnational context. By mobilising an analytic that draws on a large corpus of data collected as part of the ARITHMUS project, I highlight some of the ways in which the major themes of my previous chapter are being discussed at other NSIs and

at international forums, such as Eurostat and UNECE meetings. Although I cannot claim a comparative study, the analysis is important in order to broaden the scope of my so far nationally bounded understanding of how the assemblage producing official statistics is currently being reconceived by national statisticians in response to Big Data. Furthermore, although my aim is not to conduct a full-fledged analysis of the different actors and their relative positions in the field, I highlight ways in which supra-national organisations often come to play a key role in organising and leading debates and initiatives surrounding Big Data, not only through funding, but also through guidelines, reports, regulations etc. Furthermore, I raise examples of the many struggles and disagreements that are necessarily part of the exchanges and interactions within a transnational field.

I start the chapter by describing in more detail the ARITHMUS dataset and the analytic that I have devised to analyse it. The analysis itself consists of three parts. First, I explore the ways in which the words hype, threat, opportunity and paradigm come up in the ARITHMUS data. In addition to reiterating numerous points that were made also by my interviewees at Statistics Finland, the analysis helps to shed more light on the more problematic aspects of the increasing involvement of the private sector in the production and management of data. The analysis furthermore points to a divergence in how administrative data and Big Data are distinguished from each other at Statistics Finland and at international forums. Importantly, the analysis also highlights some of the disagreements and discontents that exist between NSIs and the supra-national organisations that increasingly come to influence their work.

In the second part of the analysis, I explore how the words skills, mentality and mindset come up in the ARITHMUS database. This part of the analysis also points to a close resemblance between my findings and the ways that these topics are being debated at other NSIs and at

international meetings and conferences. For example, many of the findings provide further evidence to suggest that increasingly, what "modernisation" comes to mean in practice is the adoption of private sector rationalities also within public sector institutions such as NSIs. In terms of new findings, the analysis suggests that cultural differences towards issues such as data sharing within government do exist between countries. I suggest that due to such differences, some governments are potentially better placed than others to take advantage of the affordances of Big Data.

In the final part of the analysis, I explore ways in which data access and partnerships are debated in the documents that make up the ARITHMUS database. As was the case at Statistics Finland, also at international meetings the importance of obtaining access to more data sources is identified as a key concern. Partnerships, also with actors who might have previously be seen solely as competitors, are seen as the appropriate way to securing access to the new data sources. In marked contrast to the positive outlook of my interviewees at Statistics Finland, however, numerous issues with the increasing need to form partnerships with private actors are identified. Business interests are for example identified as being "transient", and it is conceded that although modern companies recognise the importance of corporate social responsibility, it is usually not part of their core business models. Following from this, it is acknowledged that NSIs must in practice be able to offer companies something more than promises of collective contributions to the public good in return for access to their data. Some of the fieldwork findings of other ARITHMUS team members also suggest that statisticians are in fact far from comfortable with the increasing involvement of private actors in the production and management of data. Furthermore, in addition to having to convince external stakeholders, such as companies, about the importance of Big Data for official statistics, the findings in this section shed light on some of the internal politics that taking place in relation to Big Data within institutes. I conclude the

chapter by drawing a number of analytical conclusions from the analysis particularly in relation to the broader themes of this thesis.

6.2 A note on data and method

As I mentioned in the introduction, the aim of moving beyond nationally bounded case studies requires not only conceptual rethinking, but also methodological inventiveness. An attempt to analyse the ARITHMUS database, a collection of over three thousand documents, consisting of for example field work notes, interview transcripts, meeting minutes and various policy documents, could have easily become overwhelming.

To avoid this, I devised an analytic that draws on the large corpus of ARITHMUS data but does so within the confines of the NVivo data analysis software. The ARITHMUS project opted to use an NVivo server environment provided by the project's host university as its data management and analysis tool. The ability to have multiple users working simultaneously in the server environment was a precondition for the collaborative nature of the project. From the very beginning of the project, all researchers shared their data with the other researchers by uploading and storing them on the NVivo server. Therefore, also all the fieldwork material, including the interviews, that I collected, were included in the ARITHMUS database.

My analytic consisted of choosing several keywords from my previous chapter, and running keyword searches on them in the ARITHMUS database. The NVivo software package gives numerous query options for text-mining large corpuses of data (Bazeley and Jackson, 2013: 248–255). Instead of focusing solely on word frequencies, I ran "text search queries", which search for words or phrases, and "compound queries", which search for words in association with each other, measured usually by proximity, in order to locate the documents where the key words were

being discussed in the data. This then gave me lists of documents where the keywords featured. In order to limit the number of documents for analysis, I chose only ones where the key word featured extensively.

The exercise highlighted the technical challenges that are often involved when analysing large collections of data (Edwards et al., 2013). Despite being located in a server environment provided by the university, the software kept crashing repeatedly when running the searches. I therefore often had to limit the number of documents in one way or the other. For example, I regularly filtered documents according to the researcher that had uploaded them, meaning that I usually had to run the same search six times.

After I had chosen the documents for the analysis, I coded them in three separate steps. First, I coded descriptively under a keyword tag the sections in the documents where the keyword appeared. Once I had done this for all of the keywords, I then went through all the keyword tags and coded them further in to more analytical codes. Finally, I merged codes where there was clear overlap between them. As is usually the case with data analysis, the process was far from straightforward, and involved many iterative steps of going back and forth between the different tasks. For example, as my knowledge of the data grew, I discovered more keywords that touched on the themes and issues that I was interested in. Instead of endlessly analysing more and more data, however, I concluded the analysis once I felt that I had covered enough material to substantially expand on my analysis in the previous chapter. In the end, I analysed a total of 109 documents covering, as already mentioned, field work notes, interview transcripts, meeting minutes and a variety of policy documents.

6.3 Findings

6.3.1 Hype, threat/opportunity, paradigm

I started my analysis in the previous chapter by exploring how statisticians at Statistics Finland understood Big Data and the questions they saw it raising for their professional practice. What the analysis showed was that instead of locating its newness in its technical qualities, such as the 3vs, my respondents were more inclined to see the newness of Big Data in how the production of data was becoming organised, particularly in that more and more institutions now had the capacity to produce and analyse data, potentially challenging NSIs in their role as the producers of numerical facts on societies. As the topic of this section indicates, I chose hype, threat, opportunity and paradigm as the keywords through which I explored the occurrence of the aforementioned themes in the larger corpus of ARITHMUS data.

6.3.1.1 Hype

Similarly to my interviews, hype is a theme that features heavily in the ARITHMUS database. Hype is often acknowledged as a permanent feature of the discussion around Big Data, followed by an explanation of why it nevertheless constitutes an important topic for official statistics. These explanations typically underline the increasing digitisation of social life, and the resulting corporate harvesting and commodification of the data. General calls for NSIs to engage with Big Data are often made by individuals higher up in the organisational hierarchies. A Director General of a National Statistical Institute, who has been asked to deliver a keynote speech at an international meeting organised by Eurostat in 2017, for example underlines that irrespective of the hype, NSIs have no choice but to engage with Big Data:

> *Many people are using this term as a kind of hype and a lot of people are saying "big data this and big data that", but I will specify what I*

> *actually mean by that. […] We leave traces all over the place, which means that our lives and our actions and actions of companies are being kind of measured more, detected more than in the past […] that is why big data is not a hype, it is something fundamental and it is something that changes the way we work as well.* (ARITHMUS fieldwork notes 2017)

In contrast, those working closer to the everyday production of statistics often caution NSIs against uncritically following the hype. According to some statisticians, one of the down sides of the Big Data hype is that it raises the expectation in users that any information deficit can now be solved using Big Data. By blindly embracing the poorly understood new data sources, however, NSIs put themselves at risk of jeopardising their main asset, public trust. Diverging attitudes between the users and producers of statistics are therefore identified as a potential source of growing distrust between the two. Discussions such as these typically end with recommendations that NSIs must find ways to balance the high expectations with "realistic" approaches to taking advantage of the opportunities presented by Big Data.

Hype comes furthermore up in relation to increasing fragmentation of tasks within institutions. A statistician who works at Eurostat for example explains that due to the Big Data hype of the past few years, there is a widespread interest towards Big Data in the European Commission. He explains that what this has in practice resulted in is a situation where work on Big Data is scattered around different departments in the Commission, with some departments approaching the subject with more expertise than others:

> *Then you have every DG [Directorate-General] is doing something depending on the theme […] so employment may have an interest in jobs and skills so whatever has to do with big data and jobs and they are behind it. Those who deal with humanitarian aid and external affairs are dealing for example with migration crisis or refugees so*

> *they are impressed by all kinds of anecdotal evidence. That you can do miracles with big data and tracking migrants and so on. Those who are dealing with Energy of course there is Smart technologies. Everywhere you will find an interest in big data because during the last three/four years there was a lot of hype so […]* (ARITHMUS fieldwork notes 2016)

He goes on to explain that to some extent this is unavoidable, and not that different from national governments where themes and issues are also often dispersed across departments and agencies.

6.3.1.2 Threat/opportunity

Similarly to my interviews, Big Data comes up as an opportunity to both improve on existing statistics and to build entirely new ones. Resembling closely my findings at Statistics Finland, Big Data is identified as an opportunity to improve especially the relevancy, speed and cost-efficiency of official statistics. Mobile phone data is regularly taken up as an example of how Big Data can potentially be used to produce more accurate statistics than what can be accomplished with traditional data gathering methods, such as surveys. It is for example argued that whereas the latter generally depend on respondent's recollection and memory of events, Big Data tracks behaviour as it occurs. However, similarly to my interviews, alongside such reflections the continuing need for surveys is usually also underscored. Statistics produced with more traditional methods, which typically go back years, are still needed in order to for example test the "validity" of Big Data models. In building new products, the ability to integrate new (big) data sources with old ones, such as surveys and administrative data, is also often identified as being of key importance, this too being a finding that closely resembles my findings at Statistics Finland.

Policy documents prepared by NSIs often begin by emphasising the need to see the growing interest towards statistics that has resulted also from the debate around Big Data as an opportunity rather than a threat. Nevertheless, numerous threats are identified in relation to the new data sources. A fear that gets expressed by many is that because Big Data statistics can in theory at least be produced faster than traditional statistics, politicians and other data users might increasingly opt to use the former even when made aware of the potential issues in data accuracy. Because “decision makers will make decisions on whatever information they get”, as one keynote speaker at a conference of European statisticians organised by the OECD in 2014 summarises the issue, the prospect of official statistics being replaced by the unreliable Big Data sources is viewed as a threat to the transparency of decision making that a democratic political order is dependent on (ARITHMUS fieldwork notes 2014).

The growing number of data producers, and their increasingly commercial nature, is also problematised on numerous occasions. Already in some of the earlier documents, which date back to 2013, it is noted that data owners are gaining an increasing knowledge of NSIs interest in their data, and that this might eventually lead to increases in the cost of accessing and acquiring that data. It is furthermore often noted that even when permanent data access can be secured, the problem of ensuring the continuing existence of time-series data remains. In contrast to NSIs, businesses and their interests are identified as being “transient”, and this is seen as a serious threat to NSIs desire to ensure the longevity of data. As one national statistician explains at a conference of European statistician organised by OECD in 2016 explains, the crux of the issue with the increasing involvement of the private sector in the production of data can be summarised in the question “what happens when Google goes belly up?” (ARITHMUS fieldwork notes 2016). Statisticians also repeatedly express fears that other public sector institutions might increasingly by-pass NSIs and

instead build their own statistical products with data derived from the private sector.

Two strategic steps are usually advocated as the appropriate means by which to tackle the challenges brought about by Big Data. First, NSIs are encouraged to explore the usability of Big Data in improving official statistics, especially in terms of their timeliness. Although numerous ethical issues with Big Data sources, especially in relation to questions of privacy, are identified, not engaging with them at all is also seen as a problem since even if NSIs did nothing with them, other actors, especially in the private sector, would. NSIs are therefore encouraged to strive towards finding a balance between adhering to their traditionally strict ethical principles on the one hand, and engaging with the opportunities offered by the new data sources on the other. Interestingly, one national statistician makes the point that there would also be an ethical concern with NSIs not engaging with the new data sources, because this would mean that the uses of Big Data would be left to actors who usually care even less about ethics. In response to a presentation on Big Data ethics commissioned by Eurostat in 2016, which focuses on the numerous potential issues with NSIs engaging with Big Data, he observes the following:

> *I am concerned about finding the right balance. In your assignment, you have explored all potential objections to using Big Data in official statistics. But there is also an ethical concern with us not engaging with the data, because even if we did not use them, others still would. For example, we have been experimenting using Twitter data, and our legal experts have been complaining to us about it. But individual social data is already on the market. Individual psychological profiles can be purchased from social media companies. This is the reality, and in this reality we cannot be too strict about ethics.* (ARITHMUS fieldwork notes 2016)

In sum, the statistician seems to be suggesting that due to the increasing involvement of private actors in the production of data NSIs cannot approach Big Data solely from within their traditional conceptions of ethics.

Secondly, numerous discussions highlight the need for NSIs to become more proactive in communicating and marketing their value to society. One strategy document (UNECE, 2014c) for example underscores the need for NSIs to actively campaign for funds. It explains that this need is exacerbated by the growing number of data providers, because "[government] decisions to allocate resources [to different institutions] are taken on the basis of what is perceived, which may or may not accord with reality" (UNECE, 2014c: 3). It therefore encourages other NSIs too to "actively work to close any gap that might exist between perception and reality that have a detrimental effect on the national statistical office's ability to win its 'right' share of business." (UNECE, 2014c: 3). Again, the increasing need to actively market the value of official statistics and to campaign for funds are narratives that featured prominently also in my interviews at Statistics Finland.

6.3.1.3 Paradigm

A keyword search on the word paradigm reveals that, perhaps unsurprisingly, it often appears alongside descriptions of the changing role of NSIs and official statistics. A director of a committee on national statistics in one country summarises the main feature of the change brought about by Big Data as follows:

> *We must move from a paradigm of producing the best estimates possible from a survey to that of producing the best possible estimates to meet user needs from multiple data sources.* (ARITHMUS fieldwork notes 2016)

For others, the paradigm change implies that statistical agencies must shift their focus from producing statistics to a “more service oriented attitude […] to connect, aggregate and tailor” statistical information based on user needs (UNECE, 2015: 4). In the new paradigm, NSIs will be more defined by their value add activities in analysing in interpreting data, rather than by their data collection capabilities.

Methodologically, the new paradigm means moving from sampling to an increasing use of modelling and machine learning techniques. Discussions about appropriate software and Big Data tools feature often in the debates, usually followed by a recognition that the change required goes far beyond the implementation of new technology and methods. Instead, the new paradigm is defined by an abundance of data, where the key concern becomes “what questions to pose and how to draw inference”. Metaphors such as “from farmers to foragers of data” often feature alongside efforts to define the features of the new paradigm. Statisticians also regularly criticise presentations that focus solely on questions of method and software without appreciating the broader change required from NSIs. One statistician for example responds to a lengthy presentation about the impact of Big Data on official statistics at a Eurostat meeting in 2016 with the following comment:

> *The presentation was very much tool oriented. We are very familiar with all these tools and the thing that was missing from the presentation was an acknowledgment of the fact that what is actually changing at the moment is the paradigm around how we conduct research. With Big Data you have data first and then you ask the questions. The issue is therefore not what tools to use but what questions to ask. That’s the crux of the matter, and that is where the skills come in. All too often because the data is so vast and complex one ends up using very simple methods, such as logistic regression.* (ARITHMUS fieldwork notes 2016)

A prominent feature of my interviews at Statistics Finland was that many respondents drew a parallel between Big Data and register based statistics. For some, Big Data was in fact “just another register”, this time however owned by a private company. The question of whether Big Data and administrative registers should be treated together or separately forms a topic of debate also at international meetings. At a Eurostat Big Data task force meeting in 2016, which brings together statisticians that work on Big Data projects at their respective NSIs, a statistician from another task force that focuses on helping NSIs to take advantage of administrative data delivered a presentation on the potential synergies between administrative and Big Data. In the discussion that follows, numerous areas of overlap are identified. According to some, Big Data and administrative data are similar in terms of their data structure, as both involve combining and integrating a multitude of different data sources. It is therefore proposed that quality frameworks from one project could be reused in the other. Questions around data access are identified as another area of potential overlap. Some statisticians however argue that the two are significantly different in terms of their ownership structures, and that different data access strategies are therefore needed.

It is interesting to note that alongside calls to avoid too much overlap between the two projects, it is simultaneously admitted that “there is not always a clear distinction between administrative and big data”. The administrative data project’s aim, it is emphasised, is to focus on “classical” sources of data, such as population and tax registers. In order to achieve this aim, data is considered administrative “if they are collected to comply with regulatory obligations even if they are owned privately”. In distinguishing Big Data and administrative data, therefore, ownership is deemed less important than data structure. Privately owned transport data is taken up as an example of data that can be considered administrative since it is generated as part of a public utility. At another meeting, the issue of public/private ownership of data comes up in

relation to the outsourcing of state services, that is, a recommendation is put forward that contractual requirements for data access and sharing should be put in place as part of any agreements with private contractors (ARITHMUS fieldwork notes 2016). Clearly much conceptual ambiguity over definitions of Big Data exist also beyond Statistics Finland.

In addition to identifying numerous themes that were familiar to me already from my interviews at Statistics Finland, for example the desire of statisticians to see Big Data as an opportunity rather than a threat, or the perceived need for NSIs to shift their focus from producing data to collecting and aggregating already existing data, a look at the broader corpus of ARITHMUS documents has helped to refine the picture on at least two important issues.

First, although my interviewees at Statistics Finland recognised the threat that the increasing involvement of private actors in the production of data posed, they did not go in to too much detail in specifying why it might be an issue. In the ARITHMUS data, in contrast, the growing number of data producers and their often commercial nature is problematised on numerous occasions. The interests of businesses are identified as being “transient” which is seen as a serious threat to ensuring the longevity of data. Google might for example “go belly up”, which would mean that the statistic produced with it’s data would cease to exist along with it. Furthermore, already in the earlier documents the point is made that as businesses gain an increasing knowledge of NSIs interest in their data, prices could start going up. The point is also made that the ethical issues with Big Data would not be solved by NSIs not engaging with it, since other actors, such as companies, are already using it in often dubious ways. In sum, whereas my interviewees often promoted a positive narrative about the increasing involvement of private actors in the production of data, in their internal debates national statisticians appear significantly more concerned about its potential problematic consequences. Partially at least this might be explained by

the different settings of the conversations, but perhaps also differences exist between countries in how the issue of having to increasingly form partnerships with actors in the private sector is viewed by statisticians.

Secondly, it is interesting to compare the way that my interviewees took up the issue of the close proximity between administrative registers and Big Data and the way that the topic is being debated at international meetings. Specifically, to many of my interviewees the distinguishing factor between the two was not so much their different technical quality, but rather the fact that the former was owned by public-, and the latter by private sector institutions. In contrast, in the international debates administrative data and Big Data are not distinguished by who owns them, but rather by their different data structures. The point is made that data can be administrative, such as in the case of transport data, even when it is privately owned, as it might still be "collected to comply with regulatory obligations". It is certainly true that in countries such as the UK, where the privatisation of public services has gone much further than in some other places, the dichotomy introduced by some of my interviewees would be less useful in attempts to distinguish between administrative and Big Data. This finding further highlights the unsettled and often ambiguous nature of the object of Big Data. Whether based on technical, or some other qualities, all definitions of Big Data seem impartial and far from unproblematic. Rather, what comes to constitute Big Data is highly contextual.

Finally, it is also interesting to note that the tensions and disagreements that I highlighted usually took place at meetings commissioned by Eurostat. Whereas Eurostat statisticians often emphasised and underlined the necessity and urgency of the need for NSIs to engage with Big Data, statisticians from NSIs regularly brought up the many issues and challenges associated with the endeavour. Although a general consensus about the importance of Big Data for official statistics exists at these meetings, projects to engage with Big Data are being

promoted mainly by Eurostat, causing many frustrations amongst NSIs who are the ones that must see the projects through in practice, often amidst cuts to their general budgets. In light of the fact that one of the promises of Big Data is the automation of some of the tasks that traditionally went into the making of statistics, perhaps Eurostat's eagerness to push for Big Data in official statistics can partially at least be understood as part of broader neoliberal agendas to reduce the size of the public sector in EU member states? Future research must disentangle in detail the rationalities and agendas of the different stakeholders that are currently participating in the struggle over the meaning and significance of Big Data for official statistics.

6.3.2 Skills, mentality/mindset

Another key finding of my previous chapter was that one response to the increasing competition brought about by Big Data was a perceived need to adopt private sector ways of thinking and acting also in the public sector. From the analysis of my interviews, I chose skills, mentality and mindset as the keywords through which I explored the occurrence of the aforementioned theme in the larger corpus of ARITHMUS data.

6.3.2.1 Skills

Similarly to my interviews, also in the ARITHMUS data the need to adopt new skills features extensively alongside debates about the modernisation of official statistics. In a typical presentation at an international meeting, statisticians define modernisation as a process consisting of new products and services, new data sources and methods, the optimisation of production processes and the enhancement of human knowledge and skills. Investments in staff skills are often identified as one of the key components in the required transformation. Although the idea familiar already from my interviews

that NSIs should first obtain more Big Data before discussing at length the skills that might be needed in its analysis features also in the ARITHMUS documents, many of them go in to much more detail in specifying the types of skills required from statisticians in the age of Big Data.

At one international meeting organised by Eurostat in 2015, for example, a statistician from a European NSI explains that as part of their national Big Data “road map”, they have identified four levels of Big Data skills needed at their institute. At the highest level of hierarchy are “hardcore data scientists” who possess a broad overall knowledge about big data and a deep specialisation in some particular area, and who are usually more directed toward research. At the second level are “developers” who have a knowledge of big data systems and processes and the ability to go from research to statistical production systems. The third level of employees consists of “users” who possess know-how about big data issues in production environments and a good understanding of the specificities of big data sources. At the final level are the remaining staff who mainly need an awareness of the changing world and issues beyond Big Data.

A division between soft and hard Big Data skills is sometimes included in such definitions. Softer skills, such as an awareness of data ethics and governance, it sometimes gets noted, are more difficult to define than “harder” ones, such as those that have to do with methodological knowhow or software. A final report of the results of a survey about “the Skills Necessary for People Working with Big Data in Statistical Organisations” (UNECE, 2014d) published in 2014 defines as the three most important skills for working with Big Data IT skills (noSQL databses, SQL databases and Hadoop), Statistics skills (methodology and standards for processing Big Data, data mining) and Other skills (creative problem solving, data governance and ethics). The survey reports that most of the skills in the “Other skills” category are present at

advanced and intermediate levels at statistical organisations, whereas “IT skills” and “Statistics skills” are often lacking, indicating a shortage of technical skills at NSIs as identified also by my respondents at Statistics Finland.

Yet another familiar theme from my interviews, skills come in the ARITHMUS data also regularly up in relation to calls for the staff at NSIs to adopt a new attitude towards work. To survive in a Big Data world, one presenter at an international conference organised by the Royal Statistical Society in 2015 explains, statisticians need first and foremost a “willingness to learn new things”, which includes learning new technical skills, learning how to find datasets, figuring out who can help with issues and problems, and crossing “boundaries to create a community to make something with big data” (ARITHMUS fieldwork notes 2015). In many documents it is noted that people with Big Data skills are hard to come by, and that NSIs are in competition with big businesses over the best employees. Job satisfaction and the ability to provide experience that can pay off in terms of competitiveness for private sector jobs are identified as advantages that NSIs have in the competition for the best workers. The issue around generational divides that featured prominently in my interviews also comes up in the ARITHMUS data. Specifically, the “stickiness” of personnel is identified as a key challenge in “the management of change”, especially since some of the older employees cannot in practice be brought up to the new skill levels needed to deal with Big Data.

Similarly again to my findings at Statistics Finland, despite a widespread belief that NSIs should learn things by doing as much as possible by themselves, also at other NSIs a lack of Big Data skills has meant bringing in consultants from the private sector. In presentations about the challenge that Big Data poses to official statistics, it furthermore sometimes gets noted that other disciplines, such as computer scientists, have been quicker to adopt the skills needed to deal with Big

Data, thereby allowing them to take up the “high” positions in data analysis that once went to economists. These other professions differ from statisticians especially in that they are far better at “selling” themselves. Official statisticians, whose interests are identified as being markedly different, are criticised for having been far less effective in marketing their expertise.

A final major theme of the keyword search on skills are discussions about the discrepancy between the expectations and reality of working with Big Data sources. A national statistician who has worked in a project examining the affordances of mobile phone data explains that funding for Big Data projects is often gained due to hopes that significant cost savings can be achieved:

> *There is a general misunderstanding that new data sources will be cheaper than existing data-sources and if there is promise that you can get things done with less money this is always a major driver.* (ARITHMUS fieldwork notes 2016)

The statistician explains however that in contrast to popular belief, the costs with Big Data are unknown, and likely to be higher than expected due to the skill demands involved. The IT department at his institute, he furthermore remarks, is aware of the issue but does not speak out on it because it is not in its interest to do so:

> *Only IT-people understand that it [Big Data] will not be cheaper […] IT is not cheap […] but IT-people have no interest in telling this to anybody […] it is very difficult to foresee the real costs of producing statistics with these data-sources […] the people need a completely different skill-set to work with these kinds of data […] and this will be very expensive […] data analysts are expensive.* (ARITHMUS fieldwork notes 2016)

Again, although the issue around the expectations versus reality of working with Big Data was not explicitly stated in my interviews at

Statistics Finland, it is a topic that I recognise from discussions that took place during lunch breaks and other informal occasions during my fieldwork.

6.3.2.2 Mentality/mindset

Keyword searches on the words mentality and mindset reveal again many similar themes to those that featured in my interviews at Statistics Finland, but also new ones. As I mentioned already in the previous section, similarly to my interviews, also at international meetings Big Data is identified as being disruptive of not just methods and techniques, but an entire mentality and mindset. The move from a product to service orientation, which I noted previously, is identified as involving a cultural change at NSIs, one that must begin with the very top level of managers. At a practical level, the shift in mindset involves "a willingness to accept different definitions of quality", since the data sources from which data are derived are becoming increasingly varied.

Importantly, in defining the desired mindset, private companies are often taken up as role models to be followed. At an international meeting organised by Eurostat in 2017 for example, a senior manager explains that not only does Google and Facebook increasingly have the (big) data, they also have the mindset of a big data company, which NSIs in contrast do not. In explaining what he means by this, he takes up the example of register based statistics, which, when introduced, were "violently opposed" for reasons of principle. He explains that the same thing is currently happening in relation to Big Data, and that "the first thing to do", therefore, "is to get the mindset right":

> *The big advantage they [Facebook and Google] have is that they have the big data to do a maximum effect. They also have the mindset of a big data company, which the statistical community does not. When we started using administrative data at [our NSI] statisticians were violently opposed to them with fundamental principle reasons. The*

> *same thing is happening with big data. "This is not statistics, this is not quality", they say. The first thing to do, is therefore to get the mindset right.* (ARITHMUS fieldwork notes 2016)

Elsewhere, the correct mindset in the new circumstances is identified as involving, amongst other things, being more "experimental and aggressive". Here again, the close resemblance to business discourses is clear.

Interesting to note is also that one activity through which NSIs hope to achieve the desired change in attitude and mentality is by organising events such as "hackathons", where staff members collaborate intensively on projects over a pre-defined time period, sometimes in competition with each other. For many years, companies and venture capitalists have viewed hackathons as a quick way to develop new products and to locate new areas of innovation and funding, and it is interesting to note that such organisational forms are currently being appropriated also by government institutions such as NSIs. As I pointed out already based on my interviews at Statistics Finland, it seems that what "modernisation" often comes to mean in practice is the adoption of private sector ways of thinking and acting also in the public sector.

One of the things that my interviewees at Statistics Finland explained to me on multiple occasions was that the register based statistical system had depended on good relations between different government departments in Finland, and that their hope was that the tradition of openness and trust that had existed within the public sector would carry over to arrangements made with the private sector also. The ongoing early experiments with Big Data that I documented provided some support for this view. Furthermore, interviewees were in general markedly more positive about companies' willingness to share their data than what I was expecting, some highlighting the good pre-existing

relationship that had existed between Statistics Finland and many companies as one of the reasons for their optimism.

Mentality and mindset come up also in the larger corpus of ARITHMUS data in relation to discussions about administrative registers. While the register based census remains the method of choice for only a minority of European countries, many of them are currently investigating the possibilities for conducting future censuses using administrative registers (UNECE Statistics Wiki, 2017). One such country, which aims to “achieve” a register based census by 2020, repeatedly underscores the importance of changes in mentality as one the preconditions of the register based census (Matteus, 2013: 65–66). These changes are needed first and foremost within government offices in charge of registers, who must “take the needs of statistics into account and […] not consider them second-rate in comparison with their administrative duties”. In practice, this requires “a much more serious obligation [for the NSI] to instruct registers [offices] on collecting statistical data, and to co-ordinate their activities, especially in terms of activities and methods that guarantee data quality”. Furthermore, the NSI notes that register based statistics “will not work without citizens who understand the need of submitting data and who perform their duties”. The register based census therefore requires “intentional raising of civic awareness and regular correspondence with the public on the topic”.

Another country identifies “departmental data silos” within government as a major national obstacle to unleashing the potential in Big Data. A statistician from the Royal Statistical Society explains that in contrast to some other countries, in this country a silo mentality exists within its government, where different datasets are not shared between departments. To address the issue, the person proposes that the NSI should be given a stronger statutory right to access held within different parts of the government, a move which would guarantee real-time access to what was happening in the country. The person explains:

> *One of the big problems is that there is a silo mentality within Government, and different datasets are held and not shared across departments. The single biggest opportunity is to move where other countries have gone- Canada, New Zealand and Ireland- in giving the statistical office a broad right to data access across departments. At the moment, the [NSI] cannot easily get hold of [..] data [held within different government departments]. If it could, we would have more real-time access to what is going on around the country […] You would not have the privacy issues, because the [NSI] is interested only in aggregate data […] they do not care about us as individuals.* (House of Commons Science and Technology Committee, 2016: 22)

Breaking down data silos by handing the NSI a stronger mandate is therefore identified as a key component of the country's Big Data strategy.

In sum, much of what I uncovered in this section supports the analytical conclusions that I drew based on my interviews. In particular, many of the findings above provide further evidence to suggest that increasingly, government institutions such as NSIs look to the private sector for ideas on how to organise themselves. Big Data is seen to "disrupt" not just methods and techniques, but an entire mentality and mindset. In explaining what the desired attitude looks like, private businesses are regularly taken up as positives examples to be followed. Hackathons, data camps and other organisational practices usually associated with the private sector are increasingly seen as necessary for government institutions as well.

Another key theme in the section above are the discussions about the mentality required by register based statistics. Even though, as I have previously noted, the context in which NSIs operate is increasingly transnational, such discussions do suggest that countries differ in their governmental cultures towards issues such as data sharing. In stark contrast to the country that identified "departmental data silos" as a major

obstacle to unleashing the power of Big Data within its government, my respondents at Statistics Finland insisted repeatedly that government departments in Finland worked in unison and without major dispute. Due to its long experience of register based statistics, and the associated mentalities and practices of data sharing between sectors and agencies, both public and private, perhaps the Finnish state is better positioned than some other to tap in to the affordances of Big Data? This potentially also helps to explain why my interviewees at Statistics Finland were inclined to see the increasing involvement of the private sector in the production of data in more unproblematic terms than what I was expecting. Future research must explore in depth ways in which cultural differences come to play a role in the public-private partnerships that are now increasingly needed due to Big Data. Based on the findings in this section, it seems likely that despite the global nature of the phenomenon, the opportunities and challenges of Big Data do, to some extent, get their local articulations.

6.3.3 Data access, partnerships

In the previous chapter, by analysing in detail Big Data projects that were ongoing at Statistics Finland at the time of the research, I suggested that Statistics Finland had encountered legal, technical and organisational challenges in trying to gain access to more Big Data sources. I suggested furthermore that the legal and organisational aspects took precedence over the technical questions involved, and that the need to build partnerships with different actors was highlighted as a key concern. Some of my respondents explained to me that although partnerships had been very much part of the previous era of statistics also, the fact that they now had to be built with private sector institutions necessitated completely new conventions and arrangements. More specifically, lobbying efforts were needed on multiple fronts. A senior manager explained to me for example that on the side of the lobbying efforts for new legislation, which would grant Statistics Finland a more extensive

right to access data held by companies, he and some other members of the senior management team had devised, what he called “the road show”, where they visited different stakeholders explaining to them their data access needs. The manager, who defined “the road show” as a sales pitch, was also, along with my other respondents, markedly more positive about the prospect of increasingly having to form partnerships with the private sector, than what I was expecting. I ended the chapter by suggesting that the issue was probably far more complex than what I was able to discern based on my limited sample.

I was therefore very curious to see how the issue of partnerships was taken up in the larger corpus of ARITHMUS data. In addition to partnerships, I conducted key word searches on data access and ownership in order to gauge further at the aforementioned themes. In the end I dropped ownership from the analysis since it did not yield any new analytical themes.

6.3.3.1 Data access

One of the things that my respondents underscored to me repeatedly at Statistics Finland was the importance of first obtaining access to more Big Data sources before discussing at length other issues such as what skills might be needed to analyse them. An understanding of the issues with Big Data could only be obtained through practical work with the novel data sources. Furthermore, in relation to data access, considerations about the appropriate legal framework took for my respondents precedence over any technical issues that might be needed to be overcome.

Mirroring the findings of my interviews, data access is identified as a critical issue also at international meetings. At numerous meetings, representatives of different NSIs reiterate the centrality of obtaining access to more data sources in realising “modernisation” within their

institutes. Similarly to my interviewees, issues around data access are seen as being tightly connected to questions of legislation and public opinion. At one international meeting, for example, a statistician explains that because companies often cite legislative concerns as an obstacle to data sharing, regulators need to be approached first. Regulators, on the other hand, he notes, usually highlight the need for a public debate before more legislative powers to NSIs are given. The statistician suggest therefore that regulators and companies should be approached "carefully" as not to give the impression that the NSIs want to impose further regulations on companies. Similarly also to my interviews, statisticians from other countries cite experiences where CEOs of companies have been markedly more positive about sharing their data than employees further down the company hierarchies. Also mimicking my interview findings, the point is often made that companies tend to be happy to share their data especially after it is no longer economically relevant for them, for example after a certain period of time since it's generation has passed.

Other findings help to elaborate on some of the themes of my interviews. For example, an issue that gets discussed on numerous occasions is whether attempts at gaining access to private sector data should be coordinated at the European level. The point is sometimes made that because data is increasingly global, so too should attempts at gaining access to it be. At one meeting that discusses, amongst other things, an international framework for data access principles, the question gets raised whether NSIs should be willing to pay companies for the data or for the service of providing the data. At this particular meeting the consensus is that NSIs should "not pay for data that has a public interest, but only for the service of its pre-processing and transmission". Expertise in statistical analysis and quality assurance are often mentioned as services that NSIs can offer companies in return for their data.

An entirely new theme, one that was not discussed in my interviews, is the need to convince also colleagues at NSIs of the usefulness of the new data sources. One statistician explains that companies will only release their data if they are convinced of its usefulness for NSIs:

> *We need to prove to them that we can use it, that it is useful for official statistics, only then will they give us the data.* (ARITHMUS fieldwork notes 2016)

However, in addition to actors in the "outside world", such as companies and citizens, also colleagues at statistical agencies must be convinced of the value of Big Data:

> *In this report we have to convince people in our house that this is a viable data-source that we need. Then we have to convince the outside world that we need this data, that it is important for official statistics.* (ARITHMUS fieldwork notes 2016)

Such internal and external politics therefore requires "showing usefulness", which is defined as involving not studying everything that can be produced with a new data source, but focusing instead on developing the appropriate methodologies and procedures for a number of topics so that the results can be reproduced by other NSIs in other countries:

> *The main point of the whole [Big Data] project is to "show usefulness", that is what [the funders] are stressing: don't study all sorts of things and say this could be useful, but decide first what you want to do, then do it and develop methodologies that can be used by others […] this [the final report of the project] should become a reference for other countries to go to their data providers and say we need your data.* (ARITHMUS fieldwork notes 2016)

In "demonstrating usefulness", data visualisation techniques are often identified as being a key component.

6.3.3.2 Partnerships

The word partnerships yielded by far the most results of all the keyword searches that I conducted, which indicates its current prominence in official statistics. As one ARITHMUS researcher remarks in her fieldwork notes, partnerships seems to increasingly form one of the buzzwords in the field.

Numerous policy documents and presentations highlight the strategic importance of partnerships as “statistical offices cannot meet the global data challenge alone.” (UNECE, 2016b: 3). This is especially the case, it often gets noted, since NSIs usually do not possess the (big) data, nor the expertise needed in its analysis. Partnerships are furthermore required because users too are partners, and in a changing world official statistics will only stay up to date if NSIs maintain an awareness of the information needs of the people who use their products. In many presentations the argument is made that Big Data calls for statistical agencies to reflect on their role in society and that an important part of this is to acknowledge the importance of forming partnerships with actors who at one point might have been seen solely as competitors. One presenter at an international conference suggests that in addition to gaining a better understanding of user priorities, by forming partnerships NSIs can gain “knowledge of technology”, “clues for future standards and concepts”, “a higher position in society” and the chance to “achieve goals that could not be realised alone”.

Similarly to my interviews, in many of the documents in the ARITHMUS database the point is made that although the technical issues of data analysis are without a doubt challenging, they are usually secondary to the challenges in establishing access to data sources. In many cases, it is noted, Big Data projects exist only if a working partnership with a data provider can be forged. Hackathons, which I already mentioned in a previous section, and training bootcamps, are sometimes taken up as

examples of organisational arrangements through which partnerships with private actors such as start-ups can be formed. As I mentioned already previously, it is interesting that such practices, long used in the private sector, are currently being appropriated by governments also.

Similarly to the discussions around data access in the previous section, also in relation to partnerships the global nature of the issue is identified. At international meetings the point is regularly made that since the devices and data are not based in countries, but are international, partnerships must also be international and therefore established as part of intergovernmental processes. One of the numerous strategy documents on partnerships for example proposes that for multinational data sources “an intergovernmental group should work to forge partnership agreements that can be used by all statistical organizations”. (UNECE, 2014a: 18).

Some of the fieldwork notes collected at the supra-national statistical organisations suggest also that the global nature of Big Data is rearranging the relationship between the NSIs and the supra-national organisations, such as Eurostat, that increasingly set the broader parameters for their work. If for example partnerships were to be increasingly arranged at the supra-national level, this would inevitably take some autonomy away from the NSIs. In such discussions, however, statisticians at the international organisations often stress the importance of involving the NSIs in any negotiations with the private sector. According to a Eurostat statistician, every country that participates in such negotiations can enhance some aspect of the production of Big Data statistics, not least due to their knowledge of national peculiarities and circumstances. He explains:

> *We do not go to [private company] without for example Statistics Belgium. So they are engaged. But of course whatever we learn now from this exercise we can easily export it to Germany. Maybe in*

> *Germany there are some more peculiarities but so they will see that a lot of the investment they do not need to do. This is already an advantage for them to jump onto a train. But we always learn, they may have some greater ideas to improve certain things. We cannot claim that whatever we are going to produce now is perfect so every country that joins can enhance certain aspects of the production.* (ARITHMUS fieldwork notes 2016)

Best practices in forming partnerships are debated extensively at international meetings. Many statisticians suggest that the starting point in forming them should be to identify what benefits they can yield for different stakeholders, especially since, as one ARITHMUS researcher documents in her fieldwork notes, "recognizing [the] capabilities of different partners can lead to win-win situations". Different stakeholders, it is sometimes noted, have different interests, and the key to success is in effectively communicating what NSIs can offer. On numerous occasions, the argument is made that the best way to approach private businesses is to emphasise the social good aspects of statistics to them. One strategy document (UNECE, 2016b: 4), for example, points out that "increasingly, with globalisation of information industries comes corporate social responsibility – a desire to improve lives and be seen to improve lives beyond the narrow impact of a particular commercial service or product", and that this is something that NSIs must tap into.

In addition, a successful strategy involves emphasising the mutual benefits of partnerships. Besides the social good aspect, NSIs can for example offer companies, amongst other things, expertise in handling and interpreting data, a "treasure chest" of existing data and information sets, a reputation for quality and independence, and a collection point for sensitive data. As one keynote speaker at an international conference frames the issue, NSIs advantages includes being seen as the "good, trustworthy data professionals" who also happen to have good connections to policy makers. Another point that is often made is that

because companies already rely heavily in their operations on the data produced by the NSIs, one approach to partnerships is to agree to offer companies tailored data products in return for access to data. By drawing on experiences from one case study, the point is made that the challenge in this is in getting the companies to realise that NSIs can enrich their data and thereby potentially help them in commercialising it. In such discussions, personal relationships between representatives of NSIs and companies are identified as being key. Statisticians are however keenly aware that in some cases, the benefits of partnerships might be experienced only by the NSIs. In such cases, financial compensation for the work involved is suggested as being appropriate. As I previously mentioned however, money should only be paid for the service of providing the data, and not the data itself.

At one international meeting an interesting debate takes place on the topic of finding the correct balance between establishing partnerships and enforcing legal coercion. A representative of a European NSI sites their experience where half of the companies they wanted to form partnerships with declined because the types of analyses the NSI was offering in return for data were already being provided to the companies in a much more detailed form by private market research agencies. The NSI in question therefore decided to seek to enact legislation in order to gain access to the data held by the companies. The representative of the NSI therefore raises the question whether partnerships are sufficient, or whether NSIs will increasingly have to seek to enact coercive legal frameworks in order to gain access to data. In the debate that follows the point is made that extending the legislative powers of NSIs requires public debate, and that caution should be exercised so as not to jeopardise the public trust upon which the NSIs depend on. “From a trust building point of view”, it is noted, “partnership agreements seem to be the better approach”.

In addition to differing interests and asymmetries in rewards, cultural factors are identified as another potential obstacle to partnerships. One presenter at an international conference reminisces that even though partnerships are in principle nothing new to NSIs, even as recently as a decade ago academic researchers requesting data access were considered “intruders”. Another presenter suggests that NSIs have taken the principle of independence “to an art form”, which has resulted in an isolation where NSIs are unable to recognise the mutual interests that they might have with others. Limited resources of NSIs and inflated expectations are identified as yet another set of challenges that NSIs face when forming partnerships. Furthermore, difficult situations are said to arise in situations where some companies are more open to collaboration than others. In relation to mobile phone data, for example, national markets are usually controlled by a number of players, and full coverage of mobile phone data therefore means forming partnerships with all of them, which has often proved difficult.

As I mentioned at the start of this section, my respondents at Statistics Finland were markedly more positive about the prospect of increasingly having to collaborate with the private sector than what I was expecting. I ended the previous chapter by suggesting that the issue was probably far more complex than what I was able to uncover based on my sample. The ARITHMUS database is particularly useful in shedding more light on this topic.

Different NSIs cite for example numerous issues that they have encountered when trying to form partnerships with different actors. Private gain and public good are recognised as sometimes being in conflict with each other, especially since companies operate in competitive markets where they might at some point be forced to seek financial compensation for their data in order to stay competitive. At one international meeting the question is raised whether “public good” truly is a motivating factor for companies. It is agreed that although many

companies recognise the importance of corporate social responsibility, it is rarely part of their core business models. Partnerships are therefore conceded to often require business cases that go beyond promises of collective contributions to the greater good.

In relation to the issue of offering services in return for data, numerous questions are raised. With their limited resources NSIs "can't do everything" and the prospect of offering to evaluate companies' data in return for accessing it might therefore in practice be unrealistic, some statisticians strongly feel. The issue gets also raised that by offering services in return for data, NSIs might end up interfering with markets. In practice therefore any service that NSIs would be willing to offer to one company, they would have to be willing to offer to all other companies as well. The analytical point that I draw from this is that the prospect of increasingly having to partner with private actors in order to obtain access to Big Data raises some very important questions about the independence and impartiality of NSIs.

Finally, some of the material in the ARITHMUS database raises questions about how comfortable NSIs actually are with the increasing involvement of private actors in the production and management of data. In one country, where a private data contractor has been able to obtain a position as a mediator between companies and public sector institutions wanting to gain access to their data, a statistician admits feeling "threatened" by the company. She notes that this company already provides statistics directly to various public sector institutions, and that recently the company has tried to sell their statistical products even to the NSI. She therefore feels that by subcontracting tasks to the company, NSIs are allowing it to enter their "turf":

> *They are conquering our turf. They already do statistics for the police, for the road traffic office, for the […] central bank, for municipalities […] and now they even want to sell us their statistics. […] there is also*

> *the view that we should not give money to private companies like […] we are making them stronger […] we are getting the money and we should use it but instead we give it to them and then they go to the [companies] and get the data and then they will produce the statistics, not us.* (ARITHMUS fieldwork notes 2016)

Offering money to the private contractor is therefore in her opinion not the correct way forward, because she feels that as the statistical office, they should be the ones producing the statistics and not private companies:

> *This is not the way we should go that we buy statistics from them […] we are the statistical office […] we should produce the statistics.* (ARITHMUS fieldwork notes 2016)

The analysis in this final section has again yielded many similar findings to what I was able to discover based solely on my fieldwork at Statistics Finland. Obtaining access to more Big Data sources is identified as a key concern also at the international level, and the way to secure them is by forming partnerships with actors that previously might have been considered purely as competitors. In forming such partnerships, NSIs are advised to emphasise the public good aspect of official statistics. Numerous mutual benefits of such partnerships are also recognised, such as NSIs long established experience in quality assurance, which they can offer companies in return for their data.

In contrast to the markedly positive outlook of my interviewees, however, numerous issues are identified at other NSIs and at international meetings. It is for example recognised that although modern companies care about corporate social responsibility, it is usually not part of their core business model. Therefore, it is conceded that NSIs might increasingly have to offer services in return for data access, which again raises its own set of issues, not least the risks to the impartiality and independence of NSIs that it poses. Finally, some of the fieldwork

findings of the other ARITHMUS researchers suggest that, in contrast to what my fieldwork material indicated, statisticians are far from comfortable with the increasing involvement of private actors in the production and management of statistics.

6.4 Conclusion

In this chapter, I have situated developments at Statistics Finland in the transnational field of statistics of which it is part. Many of the findings that I have uncovered bear close resemblance to the findings of my previous chapter, thus highlighting the transnational, rather than national, character of official statistics, Big Data and processes of neoliberalisation. The analysis has highlighted numerous examples of ways in which struggles over the meaning and impact of issues such as Big Data in official statistics are "not delineated by national interests and practices […] but part of transnational negotiations, contestations and tensions that cut across numerous NSIs and international statistical organisations" (Grommé et al., Forthcoming: 2). Importantly, the analysis has provided further support to the finding that a prominent way in which NSIs are responding to the perceived challenge of Big Data is by adopting private sector mentalities, rationalities, and practices also within their institutes. Furthermore, often these developments are being pushed by the supra-national organisations, and in ways that raise discontent in the NSIs who must see the projects through in practice often amidst cuts to their general budgets. Following the conceptual starting point of this thesis, these findings can be viewed as examples of processes whereby the production of statistics is undergoing transformations in conjunction with broader political changes, such as neoliberal drives to not only reduce the sizes of state agencies, but to make them operate more like actors in the private sector. Future research much disentangle in detail the different agendas and objectives

of the stakeholders that are currently participating in the struggle over the meaning of Big Data for official statistics.

In addition to pointing out similarities between places, however, the analysis has highlighted potential points of divergence in how the challenge of Big Data is being taken up across countries. Specifically, some of the findings in this chapter suggest that NSIs in countries with established traditions in register based statistics not only approach the issue of defining Big Data differently, but also view the partnerships with the private sector often necessitated by it in a more positive light than others. Due to the many similarities between population registers and Big Data noted also by my interviewees, it is possible that countries whose statistical systems rely mainly on registers are better placed to integrate Big Data sources into their production. As Peck and Tickell (2002: 383–384) point out, processes of neoliberalisation are neither monolithic in form nor universal in effect. Although similar processes can be identified across diverse contexts and circumstances, this does not mean that they necessarily lead to a convergence of outcomes, “a neoliberalised end of history and geography”. As I covered in detail in chapter three, the register based statistical system was made possible not just by technological advances, but also by a set of specific social, political, and historical circumstances. In light of this it would be surprising if local conditions would play no role in how Big Data is being taken up and addressed in different countries. Future research must explore and contrast in more detail local articulations of the significance and meaning of Big Data for official statistics.

Perhaps most importantly, this chapter has problematised many aspects of the increasing involvement of private actors in the production of data. In stark contrast to the narratives of my previous chapter, an examination of the larger corpus of ARITHMUS data has revealed that official statisticians are in fact far from comfortable with the idea of increasingly having to partner with private actors in order to gain access to data.

Instead of viewing them solely as potential partners in the advancement of social goods, private companies are also sometimes seen as direct competitors.

Some of the findings of this chapter furthermore raise questions about the appropriateness of NSIs current responses to Big Data. Due to their role as impartial providers of factual knowledge of societies, not least to ensure democratic accountability, NSIs cannot in the end be just one seller and buyer of data in a market place of a plethora of data producers. As this chapter has demonstrated, there are ways in which functioning like a market actor is likely to come into conflict with the principles of impartiality and independence that NSIs role in society is predicated upon. NSIs response to Big Data must therefore, in the long run, necessarily go beyond appropriating the practices, mentalities and rationalities of the private data conglomerates. As Letouzé and Jütting (2014: 15) put it, due to the inherently political nature of the "data revolution" currently underway, engaging in the debates over data ownership and control is for NSIs "not a technical consideration but a political obligation". In the next and concluding chapter, I will provide some reflections on what this might mean, and propose directions for further research.

7 Conclusion

7.1 Summary

In this thesis, I have explored the meaning and significance of Big Data from the viewpoint of states, which for long held a near monopoly on the production of statistics concerning whole populations, territories, and societies. By doing so, I have provided one response to the call for research on Big Data articulated by Kitchin (2014b) amongst others. I have contributed to both our empirical and conceptual understanding of Big Data and the so called data revolution of which it is part.

Instead of starting from a technical definition, I began the analysis by exploring the literature around the history of statistics and its sociological interpretations. From this analysis emerged the conceptual starting point of this thesis: that historically, statistics have emerged out of a co-constitutive interaction between methodological and technological developments and changes in the political and administrative world (Desrosières, 1998). By exploring different historical configurations of this co-constitution, I demonstrated how advances in statistics have often been closely linked to changing governmental rationalities particularly concerning the question how to best address social issues and to regulate society more broadly.

By exploring historical configurations of this co-constitution in Finland, I suggested that the production of data on which statistics have relied has often been monopolised by dominant institutions, first of the church and later the nation state. I argued furthermore that the increasing production of data by private corporations signals a historical shift whereby the production of data has potentially started to move away from its historical basis in states. Following Alastalo (2009b), I demonstrated examples of

the numerous ways in which data production in the period of welfare states has been predicated upon social welfarist governmental concerns.

Building on the framework of co-constitution, I then situated Big Data in the context of contemporary political economy and argued that its emergence reflects broader processes of and rationalities of neoliberalisation that have gained increasing prominence in western polities in the past few decades, particularly in the sense that much of Big Data is currently being produced and accumulating in the private, rather than the public sector. I suggested that as a result, data production is increasingly becoming underpinned by capitalist objectives and rationales.

By exploring responses to and experiments with Big Data both at Statistics Finland and within the transnational field of statistics of which it is part, I suggested that the increasing production of data in the private sector based on market and profit rationalities is changing and challenging how NSIs conceive of how they produce data for official statistics. I suggested that a prominent way in which NSIs are responding to the perceived challenge of Big Data is by adopting neoliberal rationalities and mentalities in their practices. I argued that there are ways in which these rationalities are likely to come into conflict with the social welfarist concerns that for long have underpinned the production of official statistics, at least in welfare states such as Finland.

I will conclude this thesis by first, proposing a set of broader analytical themes that follow from my findings, and second, by highlighting issues and concerns that I have not been able to attend to within the remits of this study, and that must therefore be addressed in future research.

7.2 Analytical themes raised by the research

7.2.1 Big Data is an outcome of how its production is organised

Conceptually, this thesis has reinforced the importance of understanding and studying Big Data, not only as a technical object, but also a socially produced entity. Numerous accounts, both popular and academic, have taken up Big Data solely as a technical object, locating its newness in its technical qualities such as the 3Vs of Volume, Variety and Velocity. Sometimes accompanying such analyses has been the claim that in contrast to previous forms of data, Big Data occurs "naturally" without intervention from human subjects such as researchers. In contrast to surveys, which reflect what people say they do or think, Big Data is claimed to be based on direct measurements of phenomenon and therefore reflective of "actual" transactions, interactions and behaviour of people, societies, and economies.

My analysis does not support such views. Following Ruppert (2016), I approached Big Data as an emerging field of data practices that include "not only technologies and people but also norms, values, conventions and rules" (Ruppert, 2016: 2–3) and that together are generative of the sometimes novel qualities of Big Data. To move beyond attempts to define Big Data through its technical qualities, I sought a way to analyse some of these different elements and their interactions. In relation to the concerns about the future of empirical sociology raised by Savage and Burrows (2007), it is difficult to imagine many topics where social scientific reflections and approaches are currently more sorely needed.

What my analysis showed was that in marked contrast to the hype, Big Data cannot necessarily be easily defined based on its technical qualities. My respondents identified numerous similarities between old and new forms of data, specifically between administrative registers and Big Data. Importantly, both were originally created for purposes other

than official statistics, and had to therefore be cleaned and conjoined before they could be used. Since both were also generated outside of Statistics Finland, they necessitated arrangements and agreements with a variety of stakeholders. In very concrete ways therefore, what data constituted in both of these cases depended not only on the technologies used to produce them, but also on sets of arrangements made between different stakeholders.

Based on both a conceptual and an empirical analysis, I suggested that in an age of Big Data, the content of official statistics is likely to be increasingly influenced not by the motifs and rationales of government departments, as has been the case in the period of register based statistics (Alastalo, 2009b), but by those of private corporations. Furthermore, I suggested that a prominent way in which NSIs are responding to the perceived challenge of Big Data is by adopting the rationalities and mindsets of the private companies that have gained increasing prominence in terms of its production. Following the conceptual framework that society and statistics are co-constituted, I suggested that this reflects recent trends towards neoliberalism in western polities in that the production of official statistics is potentially becoming increasingly underpinned by neoliberal, rather than social welfarist, principles.

I must highlight the role of human agency in all of this. Not only is Big Data an outcome of how its production is organised, but how its production is organised is a question for humans to decide. Arguably, this question forms one of the most pressing political concerns of our time.

7.2.2 How the production of Big Data is organised is one of the most pressing political concerns of our time

Throughout this thesis I have advocated and sought to practice a critical attitude towards some of the claims that have been made about the socially transformative powers of digital technologies. Nevertheless, it seems safe to assume that digital data is becoming ever more central to the running of capitalist businesses, governments and societies more generally. Mayer-Schönberger and Cukier (2013: 182) are clearly on the right track when they claim that: “Data is to the information society what fuel was to the industrial economy: the critical resource powering the innovations that people rely on.”

What follows from this is that data is an increasingly important source of social power, and that is why we must be concerned about its distribution and control. One of the things that I have argued in this thesis is that at least in Finland, data concerning societies has historically been produced by the dominant institution of its time. Reflecting on some of the key findings of this thesis, how much authority over the production and management of data concerning societies should be given to private companies? On other hand, and connected to this, what should the role of the state be? Based on the historical analysis in chapter three, I suggested that the extensive governmental data collection system in Finland was made possible by a high level of trust towards the state, and that it is not inconceivable that this trust could be lost in the future. Since trust towards the state is historically formed, it is interesting to speculate what will happen to this trust as a result of the processes of neoliberalisation that my analysis too has highlighted. If, following recent political trends, the state is increasingly seen as a burden, rather than an enabler and facilitator of things, will citizens still be happy to trust it with their data? Or should, following neoliberal principles, all of the production and management of Big Data be left to the private sector? What will the legitimacy of a neoliberalised state versus privately owned

multinational companies be in the future? For the past decades many have vocally argued that markets are the best way to organise things, and this surely will have an impact on our ability to imagine alternative Big Data futures as well.

Clearly neither capitalist businesses nor states should be idealised. Nevertheless, I do believe that some of the findings in this thesis suggest that states must continue to have a central role in managing and regulating data also in the future. In marked contrast to businesses, NSIs are at least in principle democratically overseen. They are usually part of state departments which are, at least at the level of principle, run by elected politicians. But who regulates Google and Facebook, two prime examples of Big Data companies that increasingly seem to operate beyond and above national governments? It seems that often this role is left to the market, and the market only. One wonders how feasible this arrangement will be in the long run as these companies continue to amass more data, power and responsibility.

Furthermore, my analysis suggests that there are ways in which the need to turn in profits conflicts with a desire to share data for the greater good. Private businesses exist only as far as they are able turn in a profit, and any other considerations must therefore come second to them. Profits are the very precondition of their existence, and the ability to generate them are increasingly tied to the data that they possess. But due to the data that now accumulates especially in the large technology companies, it seems likely that they will start to obtain roles and responsibilities that have previously been under the jurisdiction of the democratic state. The question about the future role of NSIs therefore cannot be separated from larger questions about the types of polities that we want to live in in the future. What might alternative Big Data futures look like then?

7.2.3 The end of new dawn of Social Democracy?

As Ruppert (2015) points out, Big Data is a collective accomplishment of connected and interdependent peoples and technologies. Big Data is, in other words, made possible not just because of entrepreneurial individuals such as Mark Zuckerberg and Steve Jobs, but because billions of people share their lives on commercial platforms provided by them. If we accept that Big Data is socially produced, then surely it is not unreasonable to demand that so too should its benefits be? Based on my analysis, it seems difficult to claim that these benefits are currently shared very fairly or effectively. Instead, reflecting broader trends in western political economies, Big Data seems to increasingly function as the generator of the unnatural riches of a handful of individuals and their families. A world dominated by a few multinational corporations surely cannot be the best thing that we can imagine of our Big Data futures.

In order to avoid the dystopian scenario depicted in *The Circle* that I started this thesis with, therefore, what we must urgently do is to start to cultivate Social Democratic Big Data imaginations. Although their precise form is difficult to pin down, the question about the future role of the NSI must form a central component of such debates. Morozov (2017) for example points out that although in order to exploit all the insights from Big Data it needs to be accumulated in to one entity, this does not mean that that entity must be a big technology firm. He proposes that instead, all the nation's data could, for example, accrue to a national data fund, co-owned by all citizens (or, in the case of a pan-European fund, by Europeans). Companies that want to build new services on top of data would then have to pay a corresponding share of their profits for using it. It is difficult to imagine anyone better suited to run such a data fund that the NSIs with their centuries of experience of handling sensitive data. Based on the analysis that I have conducted in this thesis, we could claim that they have in fact been running one already until now, and that it would therefore be only natural for them to continue to do so also in

the future. In important ways therefore, the future of Social Democracy and the future of NSIs are intertwined.

I conclude this thesis by highlighting issues and concerns that I have not been able to attend to within the remit of this study, and that therefore must be addressed in future research.

7.3 Directions for future research

A major limitation of this study is that it explores the stakes and challenges with Big Data almost exclusively from the viewpoint of just one stakeholder, official statistics. But as I have argued repeatedly, numerous stakeholders are implicated by Big Data, and they collectively produce what Big Data becomes. I have been able to highlight what the challenges with Big Data are for NSIs, but how do for example companies or legislators view them? Because my fieldwork has centred upon Statistics Finland, my coverage of companies' views on Big Data has been limited to second hand sources. In contrast, an important next step would be to hear their views directly. Therefore, large scale research on the social production of Big Data is urgently needed.

Within such a remit, I would propose a multi-sited ethnography where Big Data analytics are used to map and visualise the different stakeholders involved in the production of Big Data. Ideally, such a study would consist of four or five different units of analysis, for example countries, where one researcher would be assigned to each. The project could start by mapping the historical trajectory of statistics in each country, much like what I have done. In the next stage, key stakeholders in Big Data would be identified, potentially with the assistance of large scale data analysis techniques. Previous studies have for example identified key stakeholders within professional fields using Twitter data (Puhakka, 2014), and I see a potential to do something similar in relation

to the different actors who have a stake in Big Data.

After having identified and mapped the different stakeholders, the next phase of the project would consist of the researchers systematically seeking to interview them. These stakeholders would include, but not be limited to, officials in key government departments, politicians, data protection legislators, CEOs and personnel at various companies, and representatives of other non-governmental organizations. By working with a comparative approach from the very beginning, the project could highlight how Big Data does or does not get local articulations depending on the different histories and political circumstances in different countries.

With such an approach, it would be possible to give a more rounded picture of the ways in which Big Data currently challenges different institutions. And by understanding the stakes and challenges involved for different actors, rather than just one or two of them, we would be better placed to construct frameworks and principles through which we will be able to unleash the power of Big Data for the advancement of common, rather than private, purposes.

Bibliography

Alapuro R and Alestalo M (1993) Konkreettinen sosiaalitutkimus [Concrete social research]. In: Alapuro R, Alestalo M, and Haavio-Mannila E (eds.) *Suomalaisen sosiologian historia [The history of Finnish sociology]*. Juva: WSOY, pp. 77–148.

Alastalo M (2005) *Metodisuhdanteiden mahti. Lomaketutkimus suomalaisessa sosiologiassa 1947-2000 [The power shifting methodological trends: Survey research in Finnish sociology 1947-2000]*. Tampere: Vastapaino.

Alastalo M (2009a) Rekisteriperusteinen tilastointitapa: kyseenalaistamaton käytäntö [Register-based statistics: an unquestioned practice]. *Hyvinvointikatsaus* (4): 58–64.

Alastalo M (2009b) Viranomaistiedosta tilastoksi: rekisteriperusteisen tilastojärjestelmän muodostaminen Suomessa [The development of the register based statistical system in Finland]. *Sosiologia* 46(3): 173–189.

Alho J (ed.) (1999) *Statistics, Registers and Science: Experiences from Finland*. Finland: Statistics Finland.

Allardt E (1994) Vertailevan politiikan tutkimus intellektuaalisen omaelämäkerran keskeisenä osana [Comparative political research as a fundamental part of an autobiography]. In: Ahtiainen P (ed.) *Historia, sosiologia ja Suomi. Yhteiskuntatutkimus itseymmärryksen jäljillä [History, sociology and Finland. Social research in search of self understanding]*. Tampere: Tammer-Paino, pp. 149–168.

Allardt E, Alapuro R and Alestalo M (1993) Suomalaisen sosiologian historiasta [Concerning the history of Finnish sociology]. In: Alapuro Risto, Alestalo M, and Haavio-Mannila E (eds.) *Suomalaisen sosiologian historia [The history of Finnish sociology]*. Juva: WSOY, pp. 13–25.

Anderson C (2008) The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *WIRED*. Available at: http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory.

Andrejevic M (2014) Big Data, Big Questions| The Big Data Divide. *International Journal of Communication* 8(0): 17.

Barry A, Osborne T and Rose N (eds.) (1996) *Foucault and political reason : liberalism, neo-liberalism and rationalities of government*. London: Routledge.

Bayatrizi Z (2009) Counting the dead and regulating the living: early modern statistics and the formation of the sociological imagination (1662–1897). *The British Journal of Sociology* 60(3): 603–621. DOI: 10.1111/j.1468-4446.2009.01260.x.

Bazeley P and Jackson K (eds.) (2013) *Qualitative data analysis with NVivo*. Second edition. London: SAGE.

Bell D (1973) *The coming of post-industrial society: a venture in social*

forecasting. Special anniversary edition. Sociology. New York: Basic Books.
Berners-Lee T (2012) Raw data, now! *WIRED UK*. Available at: http://www.wired.co.uk/article/raw-data (accessed 4 June 2018).
Big Data & Society (2018) Big Data & Society: About the Journal. Available at: http://bigdatasoc.blogspot.com/p/big-data-and-society.html (accessed 30 May 2018).
Bowden C (2013) *The US surveillance programmes and their impact on EU citizens' fundamental rights.* European Parliament's Committee on Civil Liberties, Justice and Home Affairs. Available at: http://www.europarl.europa.eu/meetdocs/2009_2014/documents/libe/dv/briefingnote_/briefingnote_en.pdf.
Boyatzis RE (1998) *Transforming qualitative information: thematic analysis and code development*. London: SAGE.
boyd d and Crawford K (2012) Critical Questions for Big Data. *Information, Communication & Society* 15(5): 662–679. DOI: 10.1080/1369118X.2012.678878.
Brown W (2015) *Undoing the demos: neoliberalism's stealth revolution*. New York: Zone Books.
Brynjolfsson E and McAfee A (2014) *The second machine age: work, progress, and prosperity in a time of brilliant technologies*. First Edition. New York: WWNorton & Company.
Buck P (1982) People Who Counted: Political Arithmetic in the Eighteenth Century. *Isis* 73(1): 28–45.
Bulmer M (1986) *The Chicago school of sociology: Institutionalization, diversity, and the rise of sociological research*. Chicago: University of Chicago Press.
Busch L (2014) Big Data, Big Questions| A Dozen Ways to Get Lost in Translation: Inherent Challenges in Large Scale Data Sets. *International Journal of Communication* 8(0): 18.
Castells M (1997) *The rise of the network society*. Cambridge: Blackwell.
Chan A (2015) Big data interfaces and the problem of inclusion. *Media, Culture & Society* 37(7): 1078–1083. DOI: 10.1177/0163443715594106.
Couldry N (2013) A Necessary Disenchantment: Myth, Agency and Injustice in the Digital Age. London School of Economics and Political Science. Available at: http://www.lse.ac.uk/newsAndMedia/videoAndAudio/channels/publicLecturesAndEvents/player.aspx?id=2120 (accessed 17 February 2015).
Crawford K, Gray ML and Miltner K (2014) Big Data| Critiquing Big Data: Politics, Ethics, Epistemology | Special Section Introduction. *International Journal of Communication* 8(0): 10.
Cukier K (2010) Data, data everywhere. *The Economist*, 25 February. Available at: http://www.economist.com/node/15557443 (accessed 26 February 2016).
Desrosières A (1990) How to Make Things Which Hold Together:

Social Science, Statistics and the State. In: Wagner P, Wittrock B, and Whitley R (eds.) *Discourses on Society*. Sociology of the Sciences Yearbook 15. Springer Netherlands, pp. 195–218.
Desrosières A (1998) *The politics of large numbers: a history of statistical reasoning*. Cambridge Massachusetts: Harvard University Press.
DeVault ML and McCoy L (2001) Institutional Ethnography: Using Interviews to Investigate Ruling Relations. In: Gudbrium JF and Holstein JA (eds.) *Handbook of Interview Research: Context and Method*. Thousand Oaks, CA: SAGE, pp. 751–76.
Diamond I (1999) The census. In: Dorling D and Simpson S (eds.) *Statistics in society: the arithmetic of politics*. New York, pp. 9–19.
Dicken P (2010) *Global Shift: Mapping the Changing Contours of the World Economy / Peter Dicken.* Sixth Edition. London: SAGE Publications Ltd.
Diesner J (2015) Small decisions with big impact on data analytics. *Big Data & Society* 2(2): 2053951715617185. DOI: 10.1177/2053951715617185.
Durkheim E (1897) *Suicide: a study in sociology*. Glencoe, Illinois: The Free Press.
Edwards PN, Jackson SJ, Chalmers MK, et al. (2013) *Knowledge Infrastructures: Intellectual Frameworks and Research Challenges*. Working Paper. Available at: http://deepblue.lib.umich.edu/handle/2027.42/97552 (accessed 15 April 2014).
Eggers D (2014) *The Circle*. San Francisco: McSweeny's Books.
Erola J (2010) Why Probability Has Not Succeeded in Sociology. *Sociology* 44(1): 121–138. DOI: 10.1177/0038038509351626.
Eskola A (1993) Sosiologian uudistuminen 1950-luvulla [The renewal of sociology in the 1950s]. In: Alapuro R, Alestalo M, and Haavio-Mannila E (eds.) *Suomalaisen sosiologian historia [The history of Finnish sociology]*. Juva: WSOY, pp. 241 – 285.
Eves H. (2002) A very brief history of statistics. *The College Mathematics Journal; Washington* 33(4): 306–308.
Ferenstein G (2014) Google's Schmidt Says Inequality Will Be Number One Issue For Democracies. In: *TechCrunch*. Available at: http://social.techcrunch.com/2014/03/07/googles-schmidt-says-inequality-will-be-number-one-issue-for-democracies/ (accessed 16 October 2015).
Florida RL (2012) *The rise of the creative class: revisited*. New York: Basic Books.
Foucault M (2007) *Security, territory, population: lectures at the Collège de France, 1977-78*. New York: Palgrave Macmillan.
Frey C. and Osborne M. (2013) *The Future of Employment: How Susceptible Are Jobs to Computerisation?* Oxford Martin School: Oxford.
Fuchs C (2014a) *Digital labor and Karl Marx*. New York: Routledge.
Fuchs C (2014b) *Social Media: A Critical Introduction*. London: SAGE.

Gabrys J, Pritchard H and Barratt B (2016) Just good enough data: Figuring data citizenships through air pollution sensing and data stories. *Big Data & Society* 3(2): 2053951716679677. DOI: 10.1177/2053951716679677.
Gissler M (1999) Routinely Collected Registers in Finnish Health Research. In: Alho J (ed.) *Statistics, Registers and Science: Experiences from Finland*. Helsinki: Statistics Finland, pp. 241 – 254.
Google (2015) *Big Data on Google Trends*. Available at: https://www.google.co.uk/trends/explore#q=Big%20Data (accessed 30 January 2015).
Gordon C (1991) Governmental rationality: an introduction. In: Burchell G, Gordon C, and Miller PM (eds.) *The Foucault effect: studies in governmentality*. London: Harvester Wheatsheaf, pp. 1–51.
Graham S and Marvin S (2001) *Splintering urbanism: networked infrastructures, technological mobilities and the urban condition*. London: Routledge.
Grommé F, Ruppert E and Cakici B (Forthcoming) Data Scientists: A New Faction of the Transnational Field of Statistics. In: Knox H and Nafus D (eds.) *Ethnography for a Data Saturated World*. Manchester: Manchester University Press. Available at: http://research.gold.ac.uk/20522/ (accessed 7 June 2018).
Haatanen P (1993) Sosiaalihistoria [Social history]. In: Alapuro R, Alestalo M, and Haavio-Mannila E (eds.) *Suomalaisen sosiologian historia [The history of Finnish Sociology]*. Juva: WSOY, pp. 13–25.
Hacking I (1982) Biopower and the avalanche of printed numbers. *Humanities in Society* 5(3–4): 279–295.
Hacking I (1990) *The taming of chance*. Cambridge University Press.
Hakim C (1985) Social monitors: population censuses as social surveys. In: Bulmer M (ed.) *Essays on the History of British Sociological Research*. University Press, Cambridge, pp. 39–51.
Halavais A (2015) Bigger sociological imaginations: framing big social data theory and methods. *Information, Communication & Society* 18(5): 583–594. DOI: 10.1080/1369118X.2015.1008543.
Halford S and Savage M (2010) Reconceptualizing Digital Social Inequality. *Information, Communication & Society* 13(7): 937–955. DOI: 10.1080/1369118X.2010.499956.
Halsey AH (2004) *A history of sociology in Britain*. Oxford University Press.
Hammer S (2011) Governing by Indicators and Outcomes: A Neo-Liberal Governmentality? In: Rudinow Saetnan A, Lomell HM, and Hammer S (eds.) *The Mutual Construction of Statistics and Society*. Routledge Advances in Research Methods 2. New York and London: Routledge, pp. 79–95.
Hansen D (2017) How To Become A Data Scientist In 2017. *Forbes*. Available at: http://www.forbes.com/sites/drewhansen/2016/10/21/become-

data-scientist/ (accessed 13 August 2017).
Harala R and Tammilehto-Luode M (1999) GIS and Register-based Population Census. In: Alho, J., Statistics, Registers and Science. In: Alho J (ed.) *Statistics, Registers and Science: Experiences from Finland*. Keuruu: Statistics Finland, Otava Book Printing, pp. 55–72.
Harvey D (2005) *A brief history of neoliberalism*. Oxford: Oxford University Press.
Hassan R (2008) *The information society*. Digital media and society series. Cambridge: Polity.
Helsingin Sanomat (2014) Keskustan Juha Sipilä: Suomessa on kymmeniätuhansia ylimääräisiä virkamiehiä. [The center party's Juha Sipilä argues: Finland has tens of thousands of public servants too many]. Available at: http://www.hs.fi/politiikka/a1414209732254?jako=6966f1a17f673ca5497878e2c46912af&ref=og-url.
Higgs E (2004) *The information state in England: the central collection of information on citizens since 1500*. Basingstoke: Palgrave Macmillan.
Hoinville G (1985) Methodological research on sample surveys. In: Bulmer M (ed.) *Essays on the History of British Sociological Research*. University Press, Cambridge, pp. 101–120.
House of Commons Science and Technology Committee (2016) *The big data dilemma*. London.
Housley W, Procter R, Edwards A, et al. (2014) Big and broad social data and the sociological imagination: A collaborative response. *Big Data & Society*. DOI: 10.1177/2053951714545135.
Howe J (2009) *Crowdsourcing: how the power of the crowd is driving the future of business*. London: Random House Business Books.
Ilmakunnas P, Laaksonen S and Maliranta M (1999) Enterprise demography and job flows. In: Alho J (ed.) *Statistics, Registers and Science: Experiences from Finland*. Helsinki: Statistics Finland.
Järvinen P (2014) Suomi, digitaalinen siirtomaa [Finland, a digital colony]. *HS.fi*, 28 January. Available at: http://www.hs.fi/tekniikka/a1390878024013?jako=60a5bfd602d6c9b385bedf86e490d2d1&ref=og-url (accessed 5 June 2015).
Johnson B (2013) We should be humbly thanking the super-rich, not bashing them. *The Telegraph*. Available at: http://www.telegraph.co.uk/comment/columnists/borisjohnson/10456202/We-should-be-humbly-thanking-the-super-rich-not-bashing-them.html (accessed 29 June 2017).
Jungner M (2015) *Otetaan digiloikka! Suomi digikehityksen kärkeen [Let's take a digi-leap! Finland to the top of digital progress]*. Confederation of Finnish Industries. Available at: http://ek.fi/wp-content/uploads/Otetaan_digiloikka_net.pdf (accessed 27 October 2015).
Kent R (1981) *A History of British Empirical Sociology*. Aldershot:

Gower.
Kent R (1985) The emergence of the sociological survey, 1887-1939. In: Bulmer M (ed.) *Essays on the History of British Sociological Research*. University Press, Cambridge, pp. 52–69.
Kinnunen M (1998) Numeroidut ihmiset. Työeläkekortin numerosta henkilötunnukseksi [Numbered people. From an employment pension card number to an identity number]. In: Paananen S, Juntto A, and Sauli H (eds.) *Faktajuttu [Factual issues]*. Tampere: Vastapaino, pp. 117–134.
Kitchin R (2014a) Big Data, new epistemologies and paradigm shifts. *Big Data & Society* 1(1): 2053951714528481. DOI: 10.1177/2053951714528481.
Kitchin R (2014b) *The data revolution: big data, open data, data infrastructures & their consequences*. Los Angeles: SAGE.
Kitchin R (2015) Big Data and Official Statistics: Opportunities, Challenges and Risks. *Statistical Journal of the International Association of Official Statistics* 31(3): 471–481.
Kitchin R and McArdle G (2016) What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. *Big Data & Society*. DOI: 10.1177/2053951716631130.
Kivivuori J (2012) Feministinen väkivaltatutkimus: akateeminen liike ja ongelmallinen teoria [Feminist crime research: An academic movement and its problematic theoretical basis]. *Tieteessä tapahtuu [Developments in science]* 5(30).
Kramer ADI, Guillory JE and Hancock JT (2014) Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences* 111(24): 8788–8790. DOI: 10.1073/pnas.1320040111.
Kruskal W and Mosteller F (1980) Representative Sampling, IV: The History of the Concept in Statistics, 1895-1939. *International Statistical Review / Revue Internationale de Statistique* 48(2): 169–195. DOI: 10.2307/1403151.
Kullenberg C (2011) Sociology in the Making: Statistics as a Mediator between the Social Sciences, Practice, and the State. In: Saetnan AR, Lomell HM, and Hammer S (eds.) *The Mutual Construction of Statistics and Society*. Routledge Advances in Research Methods 2. New York: Routledge, pp. 64–78.
Kuusela V (2011a) Ensimmäiset merkittävät tilasto-organisaatiot syntyivät 1800-luvulla [The first important statistical organisations were founded in the 19th century]. Available at: http://www.stat.fi/artikkelit/2010/art_2010-12-13_005.html?s=0 (accessed 28 November 2015).
Kuusela V (2011b) Paradigms in Statistical Inference for Finite Populations; Up to the 1950s. Available at: https://helda.helsinki.fi/handle/10138/27416 (accessed 7 March 2015).
Lanier J (2013) *Who owns the future?* London: Allen Lane.
Law J, Ruppert E and Savage M (2011) *The double social life of methods*. Working Paper, March. CRESC, Open University.

Available at: http://research.gold.ac.uk/7987/1/The%20Double%20Social%20Life%20of%20Methods%20CRESC%20Working%20Paper%2095.pdf.
Lazarsfeld P (1970) Notes sur l'histoire de la quantification en sociologie: les sources, les tendances, les grands probl`emes. In: *Philosophie des sciences sociales*. Paris: Gallimard, pp. 75–162.
Lazarsfeld P and Oberschall R (1965) Max Weber and empirical social research. *American sociological review* 30(2): 185–199.
Lehtonen R and Veijanen A (1999) Use of Register Data to Improve the Estimation in a Sample Survey. In: Alho J (ed.) *Statistics, Registers and Science: Experiences from Finland*. Helsinki: Statistics Finland, pp. 197 – 210.
Lepenies W (1988) *Between literature and science: the rise of sociology*. Cambridge University Press.
Letouzé E and Jütting J (2014) *"Official Statistics, Big Data, and Human Development."* Data-Pop Alliance White Paper Series. Paris. Available at: http://datapopalliance.org/item/white-paper-official-statistics-big-data-and-human-development/ (accessed 7 June 2018).
Lewis K (2015) Three fallacies of digital footprints. *Big Data & Society* 2(2): 2053951715602496. DOI: 10.1177/2053951715602496.
Lupton D (2014) *Digital Sociology*. Routledge.
Luther G (1993) *Suomen tilastotoimen historia vuoteen 1970 [The history of the statistical service in Finland up until 1970]*. Helsinki: Tilastokeskus.
Lyon D (2014) Surveillance, Snowden, and Big Data: Capacities, consequences, critique. *Big Data & Society* 1(2). DOI: 10.1177/2053951714541861.
Machlup F (1962) *The production and distribution of knowledge in the United States*. Princeton, N.J., Princeton University Press. Available at: http://archive.org/details/productiondistri00mach (accessed 9 October 2015).
MacKenzie DA (1981) *Statistics in Britain: 1865-1930: The social construction of scientific knowledge*. Edinburgh University Press.
MacKenzie DA and Wajcman J (eds.) (1999) *The social shaping of technology*. 2nd ed. Buckingham: Open University Press.
Macy M (2015) Opportunities and challenges for computational social science. Helsinki. Available at: https://www.youtube.com/watch?v=37QvponcEDc.
Mäkelä J (1996) *Menneisyyden nykyisyys: Kvalitatiivisen ja kvantitatiivisen tuolla puolen [The present of the past: Beyond qualitative and quantitative research]*. Rovaniemi: Acta Universitatis Lapponiensis.
Marjomaa P (ed.) (2000) *Tilastokeskus 1970–2000 [Statistics Finland: 1970–2000]*. Helsinki: Tilastokeskus.
Matteus D (2013) Roadmap to a register-based census. *Quarterly Bulletin of Statistics Estonia* 4: 64–69.

Mayer-Schönberger V and Cukier K (2013) *Big data: a revolution that will transform how we live, work and think*. London: John Murray.
Meinander H (2011) *A history of Finland*. New York: Columbia University Press.
Ministry of Finance (2015) Digitalisation. Available at: http://vm.fi/digitalisaatio?p_p_id=56_INSTANCE_SSKDNE5ODInk&p_p_lifecycle=0&p_p_state=normal&p_p_mode=view&p_p_col_id=column-2&p_p_col_count=1&_56_INSTANCE_SSKDNE5ODInk_languageId=en_US (accessed 27 October 2015).
Mirowski P (2013) *Never let a serious crisis go to waste: how neoliberalism survived the financial meltdown*. London: Verso.
Mirowski P and Plehwe D (eds.) (2009) *The road from Mont Pèlerin: the making of the neoliberal thought collective*. Cambridge, Mass: Harvard University Press.
Morozov E (2014) The rise of data and the death of politics. *The Guardian*, 20 July. Available at: http://www.theguardian.com/technology/2014/jul/20/rise-of-data-death-of-politics-evgeny-morozov-algorithmic-regulation (accessed 9 October 2015).
Morozov E (2015a) Does Silicon Valley's reign herald the end of social democracy? *The Guardian*, 20 September. Available at: http://www.theguardian.com/commentisfree/2015/sep/20/silicon-valley-end-of-social-democracy (accessed 9 October 2015).
Morozov E (2015b) Facebook isn't a charity. The poor will pay by surrendering their data. *The Guardian*, 26 April. Available at: http://www.theguardian.com/commentisfree/2015/apr/26/facebook-isnt-charity-poor-pay-by-surrending-their-data (accessed 9 October 2015).
Morozov E (2015c) Silicon Valley likes to promise 'digital socialism' – but it is selling a fairytale. *The Guardian*, 1 March. Available at: http://www.theguardian.com/commentisfree/2015/mar/01/silicon-valley-promises-digital-socialism-but-is-selling-a-fairy-tale (accessed 9 October 2015).
Morozov E (2015d) What happens when policy is made by corporations? Your privacy is seen as a barrier to economic growth. *The Guardian*, 12 July. Available at: http://www.theguardian.com/commentisfree/2015/jul/12/ttip-your-data-privacy-is-a-barrier-to-economic-growth (accessed 9 October 2015).
Morozov E (2015e) Where Uber and Amazon rule: welcome to the world of the platform. *The Guardian*, 7 June. Available at: http://www.theguardian.com/technology/2015/jun/07/facebook-uber-amazon-platform-economy (accessed 9 October 2015).
Morozov E (2017) To tackle Google's power, regulators have to go after its ownership of data. *The Observer*, 1 July. Available at: http://www.theguardian.com/technology/2017/jul/01/google-european-commission-fine-search-engines (accessed 24

September 2017).
Mosco V (2004) *The digital sublime: myth, power, and cyberspace*. Cambridge Massachusetts: MIT Press.
Mosco V (2014) *To the cloud: big data in a turbulent world*. Boulder: Paradigm Publishers.
Murthy D (2013) *Twitter: social communication in the Twitter age*. Digital media and society. Cambridge: Polity.
Myllys K (1981) Tilastotoimen kehitys eri maissa [The development of official statistics in different countries]. In: Hietala M and Myllys K (eds.) *Tutkija tilastolliset tiedonlähteet [Statistics as a resource for researchers]*. Helsinki: Gaudeamus, pp. 49–77.
Myrskylä P (2010) Rekisterien tilastokäyttö lisääntyy talouspaineissa [The use of registers increases amidst financial pressures]. *Hyvinvointikatsaus* (4): 58–62.
Myrskylä P (2011) More than 250 years of population censuses. Available at: http://www.stat.fi/tup/vl2010/art_2011-03-18_001_en.html (accessed 6 July 2015).
Negroponte N (1995) *Being digital*. London: Hodder & Stoughton.
Nieminen M (1999) *Väestotilastoja 250 vuotta: katsaus väestotilaston historiaan vuosina 1749-1999 [250 Years of Population Statistics in Finland: an examination in to the history of social statistics between 1749-1999]*. Helsinki: Tilastokeskus.
Nieminen M (n.d.) *250 Years of Population Statistics in Finland*. Available at: https://www.stat.fi/isi99/proceedings/arkisto/varasto/niem1020.pdf (accessed 17 November 2015).
Osborne T and Rose N (2008) Populating sociology: Carr-Saunders and the problem of population. *The Sociological Review* 56(4): 552–578. DOI: 10.1111/j.1467-954X.2008.00805.x.
Osborne T, Rose N and Savage M (2008) Editors' Introduction Reinscribing British sociology: some critical reflections. *The Sociological Review* 56(4): 519–534. DOI: 10.1111/j.1467-954X.2008.00803.x.
Patil TH and Davenport DJ (2012) Data Scientist: The Sexiest Job of the 21st Century. *Harvard Business Review*. Available at: https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century (accessed 23 September 2016).
Peck J (2010) *Constructions of neoliberal reason*. Oxford: Oxford University Press.
Peck J and Tickell A (2002) Neoliberalizing Space. *Antipode* 34(3): 380–404. DOI: 10.1111/1467-8330.00247.
Pelizza A (2016) Developing the Vectorial Glance: Infrastructural Inversion for the New Agenda on Government Information Systems. *Science, Technology, & Human Values* 41(2): 298–321. DOI: 10.1177/0162243915597478.
Pentland A (2014) *Social Physics: How Good Ideas Spread-The Lessons from a New Science*. New York: Penguin Press.
Pentland A (Sandy) (2012) REINVENTING SOCIETY IN THE WAKE OF BIG DATA - A Conversation with Alex (Sandy) Pentland.

Edge, 30 August. Available at: https://edge.org/conversation/alex_sandy_pentland-reinventing-society-in-the-wake-of-big-data (accessed 10 March 2015).
Piketty T (2014) *Capital in the twenty-first century*. Cambridge Massachusetts: The Belknap Press of Harvard University Press.
Poon M (2016) Corporate Capitalism and the Growing Power of Big Data: Review Essay. *Science, Technology, & Human Values* 41(6): 1088–1108. DOI: 10.1177/0162243916650491.
Population Register Centre (2015) History. Available at: http://www.vaestorekisterikeskus.fi/default.aspx?id=42 (accessed 17 November 2015).
Porat M (1977) *The information economy: definition and measurement / Marc Uri Porat.* OT special publication; 77-12(1). Washington, DC: USDepartment of Commerce, Office of Telecommunications.
Porter TM (1986) *The rise of statistical thinking: 1820-1900*. Princeton: Princeton University Press.
Porter TM (1995) *Trust in numbers: the pursuit of objectivity in science and public life*. Princeton, N.J: Princeton University Press.
Press G (2014) 12 Big Data Definitions: What's Yours? *Forbes*. Available at: http://www.forbes.com/sites/gilpress/2014/09/03/12-big-data-definitions-whats-yours/ (accessed 29 September 2015).
Prewitt K (1987) Public statistics and democratic politics. In: *The politics of numbers*. Russell Sage Foundation, pp. 261–274.
Puhakka I (2014) *Toimitusten pomot Twitterissä – Verkostoanalyyttinen katsaus jorunalististen päättäjien Twitteryhteisöön [Newsroom bosses on Twitter - A social network analysis approach to identifying the community of news editors on Twitter]*. MA thesis. University of Jyväskylä.
Qiu JL (2015) Reflections on Big Data: 'Just because it is accessible does not make it ethical.' *Media, Culture & Society* 37(7): 1089–1094. DOI: 10.1177/0163443715594104.
Rahkonen K (1995) Teorian vastaanotosta suomalaisessa sosiologiassa [About the reception of social theory in Finnish sociology]. In: Rahkonen Keijo (ed.) *Sosiologisen teorian uusimmat virtaukset [Latest currants in Finnish sociology]*. Helsinki: Gaudeamus, pp. 9–20.
Rapley TJ (2001) The art(fulness) of open-ended interviewing: some considerations on analysing interviews. *Qualitative Research* 1(3): 303–323. DOI: 10.1177/146879410100100303.
Reich R (1991) *The work of nations: preparing ourselves for 21st century capitalism.* 1st ed.. New York: AAKnopf.
Renwick C (2012) *British sociology's lost biological roots: a history of futures past*. Basingstoke: Palgrave Macmillan.
Ritzer G and Jurgenson N (2010) Production, Consumption, Prosumption The nature of capitalism in the age of the digital 'prosumer.' *Journal of Consumer Culture* 10(1): 13–36. DOI: 10.1177/1469540509354673.

Roos J-P (2011) Yhteiskuntatieteiden kurjuus: esimerkkinä kaksi laudaturväitöskirjaa [The miserable state of the social sciences: two laudatur PhDs as examples]. *Yhteiskuntapolitiikka* 3(76).
Rose N (1991) Governing by numbers: Figuring out democracy. *Accounting, Organizations and Society* 16(7): 673–692. DOI: 10.1016/0361-3682(91)90019-B.
Ruppert E (2015) Who owns Big Data? *Discover Society* (23). Available at: http://discoversociety.org/2015/07/30/who-owns-big-data/ (accessed 20 January 2016).
Ruppert E (2016) Big Data economies and ecologies. In: *An End to the Crisis of Empirical Sociology? Trends and Challenges in Social Science Research*. London: SAGE, pp. 13–28.
Ruppert E (2018) Sociotechnical Imaginaries of Different Data Futures. Erasmus University Rotterdam. Available at: https://www.eur.nl/en/essb/3rd-van-doorn-lecture.
Saetnan AR, Lomell HM and Hammer S (2011) Introduction: By the very act of counting: The mutual construction of statistics and society. In: Saetnan AR, Lomell HM, and Hammer S (eds.) *The Mutual Construction of Statistics and Society*. Routledge Advances in Research Methods 2. New York: Routledge, pp. 1–17.
Savage M (2010) *Identities and social change in Britain since 1940: the politics of method*. Oxford: Oxford University Press.
Savage M (2014) Piketty's challenge for sociology. *The British Journal of Sociology* 65(4): 591–606. DOI: 10.1111/1468-4446.12106.
Savage M and Burrows R (2007) The Coming Crisis of Empirical Sociology. *Sociology* 41(5): 885–899. DOI: 10.1177/0038038507080443.
Saxenian A (2014) The Silicon Valley Model: Economic Dynamism, Social Exclusion. In: Castells M and Himanen P (eds.) *Reconceptualizing Development in the Global Information Age*. Oxford University Press.
Scheel S, Cakici B, Grommé F, et al. (2016) Transcending Methodological Nationalism through a Transversal Method? On the Stakes and Challenges of Collaboration. *ARITHMUS Working Paper Series, Paper No. 1.* DOI: 10.13140/RG.2.2.33901.79842.
Scheveningen Memorandum (2013) *Big Data and Official Statistics*. Available at: http://ec.europa.eu/eurostat/documents/42577/43315/Scheveningen-memorandum-27-09-13 (accessed 4 February 2018).
Schroeder R (2014) Big Data and the brave new world of social media research. *Big Data & Society* 1(2): 2053951714563194. DOI: 10.1177/2053951714563194.
Selvin HC (1985) Durkheim, Booth and Yule: non-diffusion of an innovation. In: Bulmer M (ed.) *Essays on the History of British Sociological Research*. University Press, Cambridge, pp. 52–69.
Shirky C (2008) *Here comes everybody: the power of organisation without organisations*. London: Allen Lane an imprint of Penguin

Books.
Silverman D (2013) *Doing qualitative research*. Los Angeles: SAGE.
Sinnemäki A (2005) *Merkintöjä menneestä. Helsingin yliopiston sosiologian laitoksen monisteita*.
Srnicek N (2017) *Platform capitalism*. Cambridge: Polity Press.
Stapleton L (2011) *Taming big data*. CT316. IBM. Available at: http://www.ibm.com/developerworks/data/library/dmmag/DMMag_2011_Issue2/BigData/ (accessed 29 September 2015).
Starr P (1987) The Sociology of Official Statistics. In: *The politics of numbers*. Russell Sage Foundation.
Starr P and Corson R (1987) Who will have the numbers? The rise of the statistical services industry and the politics of public data. In: Alonso W and Starr P (eds.) *The politics of numbers*, pp. 415–447.
Statistics Finland (2013) *Development and challenges of on-line micro-data usage*. Geneva: United Nations Economic and Social Council, Conference of European Statisticians.
Statistics Finland (2015) The Statistics Act. Available at: http://www.stat.fi/meta/lait/tilastolaki_en.html (accessed 17 November 2015).
Statistics Finland (2017) *Preliminary population statistics*. Available at: http://www.stat.fi/til/vamuu/index_en.html (accessed 21 July 2017).
Stigler S (1986) *The history of statistics: the measurement of uncertainty before 1900*. Cambridge, Mass. ; London: Belknap Press of Harvard University Press.
Streeter T (2011) *The net effect: romanticism, capitalism, and the internet*. Critical cultural communication. New York: New York University Press.
Struijs P, Braaksma B and Daas PJ (2014) Official statistics and Big Data. *Big Data & Society* 1(1): 2053951714538417. DOI: 10.1177/2053951714538417.
Terranova T (2000) Free Labor: Producing Culture for the Digital Economy. *Social Text* 18(2): 33–58.
Thatcher J (2014) Big Data, Big Questions| Living on Fumes: Digital Footprints, Data Fumes, and the Limitations of Spatial Big Data. *International Journal of Communication* 8(0): 19.
Thomas RJ (1995) Interviewing Important People in Big Companies. In: Hertz R and Imber JB (eds.) *Studying Elites Using Qualitative Methods*. Thousand Oaks, CA: SAGE, pp. 3–17.
Thompson K (1976) *Auguste Comte: the foundation of sociology*. London: Nelson.
Tinati R, Halford S, Carr L, et al. (2014) Big Data: Methodological Challenges and Approaches for Sociological Analysis. *Sociology*: 0038038513511561. DOI: 10.1177/0038038513511561.
Toffler A (1980) *Third wave*. London: Collins.
Töttö P (1989) *Sosiologia teoriana modernista yhteiskunnasta [Sociology as a theory about modern society]*. A 58. Tampere:

Tampereen yliopisto, yhteiskuntatieteiden tutkimuslaitos.
UNECE (2007) *Register-based statistics in the Nordic countries*. New York and Geneva: United Nations.
UNECE (2013) *What Does "Big Data" Mean for Official Statistics?* 18 March. Geneva: United Nations.
UNECE (2014a) *Guidelines for the establishment and use of partnerships in Big Data Projects for Official Statistics*.
UNECE (2014b) *How big is Big Data? Exploring the role of Big Data in Official Statistics. Draft paper. UNECE Statistics Wikis*.
UNECE (2014c) *How do we define the value and benefits of official statistics in an increasingly competitive data industry?* Geneva: United Nations Economic and Social Council, Conference of European Statisticians.
UNECE (2014d) *Survey about the Skills Necessary for People Working with Big Data in Statistical Organisations*. The Role of Big Data in the Modernisation of Statistical Production. Available at: https://statswiki.unece.org/display/bigdata/2014+Project (accessed 24 September 2017).
UNECE (2015) *Innovative products and services for European statistics: challenges and opportunities*. Geneva: United Nations Economic and Social Council, Conference of European Statisticians.
UNECE (2016a) *Interim Report of the Task Force on the Value of Official Statistics. Conference of European Statisticians. Sixty - fourth plenary session.* Paris, 27 - 29 April.
UNECE (2016b) *Partnerships in data production*. Geneva: United Nations Economic and Social Council, Conference of European Statisticians.
UNECE Statistics Wiki (2017) Collection of papers on transitions to new census methods. Available at: https://statswiki.unece.org/display/censuses/Collection+of+papers+on+transitions+to+new+census+methods.
Uprichard E (2013) Focus: Big Data, Little Questions? *Discover Society* (1). Available at: http://discoversociety.org/2013/10/01/focus-big-data-little-questions/ (accessed 3 July 2015).
Uprichard E (2014) Big-Data Doubts. *The Chronicle of Higher Education*, 13 October. Available at: http://chronicle.com/article/Big-Doubts-About-Big-Data-/149267/ (accessed 3 July 2015).
Uprichard E (2015) Most big data is social data – the analytics need serious interrogation. In: *Impact of Social Sciences*. Available at: http://blogs.lse.ac.uk/impactofsocialsciences/2015/02/12/philosophy-of-data-science-emma-uprichard/ (accessed 20 January 2016).
Urry J (2000) *Sociology beyond societies: mobilities for the twenty-first century*. International library of sociology. London: Routledge.
Väisänen H (2013) The wonderland of statistics? History, access and uses of Finnish register data. Lecture. Goldsmiths, University of London.

Valkonen T and Martelin T (1999) Social Inequality in the Face of Death – Linked Registers in Mortality Research. In: Alho J (ed.) *Statistics, Registers and Science: Experiences from Finland*. Helsinki: Statistics Finland, pp. 211 – 224.
Van Dijck J (2013) *The culture of connectivity: a critical history of social media*. Oxford: Oxford University Press.
Waris H (1932) *Työläisyhteiskunnan syntyminen Helsingin pitkänsillan pohjoispuolelle*. Helsinki: Weilin + Göös.
Webster F (2014) *Theories of the information society*. Fourth edition. International library of sociology. Abingdon, Oxon: Routledge.
Westergaard H (1932) *Contributions to the history of statistics*. London: King.
Whitehead F (1985) The Government Social Survey. In: Bulmer M (ed.) *Essays on the History of British Sociological Research*. University Press, Cambridge, pp. 83–100.