METTE MY MADSEN *

Data as Monads: How Digital Data can be Understood as the Sum of the Components in the Process of Locating it

Intersections. EEJSP
3(1): 15-30.
DOI: 10.17356/ieejsp.v3i1.284
http://intersections.tk.mta.hu

* [my5madsen@gmail.com] (University of Copenhagen, Denmark)

Abstract

This article concerns epistemology of data, especially digital data or Big Data. It especially problematizes data understood as something fact-like. The empirical object is a case where researchers from an interdisciplinary collaborative research environment working with Big Data, experienced how banal questions about how to find data led to a journey around problems of competing mathematical models, ethical questions and human agency. Based on the empirical material the article unfolds how data itself can be regarded as inherently contextual, fragile and unstable. This is done analytically by shifting focus from the two most dominant understandings of data within social science: data as something 'raw' that can be picked up using the right tools and data as 'shaped' by the tools of collection. Instead the article proposes to regard data as monad-like that is as the unstable constellations of components in the process of locating the data.

Keywords: Big Data; Digital Data; Monads; Interdisciplinary Collaboration; Computational Social Science.

1. Introduction

This article is about how banal questions of how to locate data in a data research project led to the questioning of our understanding of data itself. The empirical material arrives from a research project in which the author participated, namely the Copenhagen based, interdisciplinary, Big Data research project called the Copenhagen Network Study (CNS). This large-scale research project involved scientists from eight institutes who all shared a common data pool comprised of data from various and diverse channels. The data was collected from a population of 800 freshman students at the Danish Technical University (DTU). Within the CNS the involved researchers organized into sub-groups. In the sub-groups they did their own sub-projects both within disciplines and interdisciplinary ones (http://socialfabric.ku.dk/). The article's empirical focus will be on one such interdisciplinary sub-group, a team of sociologists and anthropologists, and this team's research using data from a student party held at DTU campus to investigate different kinds of intensities at this social event. More precisely the article will focus on the problems that arose before the team of researchers could even start what they thought would be their research project, namely the problem of two seemingly banal questions: 'where was the party located?' and 'who participated?'. Initially these questions seemed rhetorical, something that had to be answered only in order to get on with the real analysis, something that would provide the team with the fundamental data material that they were supposed to base their research on. However, it turned out that the path to answering these seemingly simple questions was intertwined with a context of numerable obstacles concerning competing mathematical models, interpretations of calculations, research political questions and human agency. What was interesting for the researchers to notice was how every step taken and problem encountered in finding the data, altered the data, i.e. the data looked slightly differently at every step or problem encountered. Based on this experience, the article argues that if we understand its context not as something outside of data, but as constituting elements of data itself, data can be understood as continuously negotiated constellations of diverse elements of context.

Digital behavioural data, especially of the kind called Big Data, is often assumed to provide analysis with an objectivity that arrives form direct correspondence with the lived world. As such, it is assumed in many different fields of science as well as other social sectors like business or government, that digital data can provide the possibility to measure the social world with an unprecedented precision. Precision here meaning using data to come closer to an 'objective' or 'true' picture of the social world. By suggesting data as constellations of diverse elements the paper seriously questions these assumptions. However, the point here is not to reject them as 'untrue', but to show that they represent merely one perspective on how data can be understood. The advances in the use of Big Data and network analysis do offer the possibility to answer questions, as well as, to question answers arrived at in previous studies. Too often social scientists either reject working with Big Data or attempt to translate the new approaches to digital data in terms of using existing quantitative routines, gaining little other than 'business as usual'. The first part of the article will focus on computational social science and interdisciplinary collaboration using digital

data as scientific environments that have recently inspired many new insights (Knox and Nafus, n.d.). But here it is also shown that the increased attention to digital data research is mainly centred on sharing and developing tools and methods, while 'data' is left unquestioned. As such, the article follows Davies' (2013) point that disciplinary pluralism will help us reveal the cultural and political substrates of disciplines (Davies, 2013), in this case the substrates concerning data. This article will show that we can understand more about the dilemmas and use of Big Data, by putting data itself into the centre of analysis and questioning our understanding of it. The article will demonstrate how there are ways of approaching data that might open up to scientific investigation – not only with digital data, but also *of* digital data, that is to say *in* digital data itself. Thus the primary concern of the article is about the epistemology of data.

To come closer to an analysis of data itself the article will zoom in on two dominant perspectives on data within computational and social science in its second part, namely data as 'raw' and data as 'shaped'. The first represents an understanding of data as something 'out there', as social facts that exist prior to collection and that can be picked up using increasingly refined tools. The latter is an understanding of data as shaped by the very tools of collection and their extended context of biases such as political, ethical, gender or paradigmatic matters. Data can be made more objective or fact-like by accounting for the process and the extended context of collection. Here data does not exist prior to collection, but is in an inherent interrelation with context (Eriksen, 2001). Though distinct perspectives of data, these two do have the understanding that data is, or can at least be transformed into, something objective or fact-like in common. The article will, in its third and last part, follow the understanding of an inherent interrelation between data and context, but will propose a new way of understanding this relation by advancing the concept of *monad* (Latour et al., 2012). Monad here means the duality of something as simultaneously both unit and composition. Using the concept of monad as a perspective from where to look at the epistemology of data the article argues that digital data can be understood as monad-like. The empirical example of the problems encountered at the interdisciplinary research event is used to demonstrate this point: as the context, that is every problem or step taken to find the data, ended up making the data somehow different from before, the article argues that the context can be seen as parts that compose data. Finding data then means settling for one such composition chosen over other possible compositions. This perspective shows data as other than fact-like, as being monad-like, that is to say, being composed of multiple, diverse and changeable elements.

2. Computational social science and the focus on method

Digital data especially of the kind called Big Data has recently and to an ever-increasing degree become a tool to investigate and analyse social interaction and social networks. This is, for example, seen in the variety of themes that are being studied such as; productivity and information diffusion (Aral et al., 2007), collaborative networks (Sonnenberg et al., 2000), interdependence and trust (Tomkins, 2001), prediction of habitual and non-habitual actions (McInerney et al., 2013), behaviour contagion (Centola, 2010) using digital data obtained from 'real life' by channels such

as calls, e-mail and message logs, geo-location, Bluetooth, online social networks, sensory networks, search histories and more. While some find the storing and usage of these large amounts of obtained digital data concerning and call for increased security (Perrig, 2002), others find that the 'real life' aspect of digital data has become superfluous and that computational research of sociality should instead be done in artificial society simulations built from digital data (Gilbert, 2005).

Also within the social sciences Big Data has raised much attention. In 2007 Savage and Burrows declared that the new age of big and available digital data sets would soon cause a major crisis within empirical sociology since the expertise that was formerly the property of scientists would no longer be needed as both digital data and methods of analysis would now be developed in the hands of public and private institutions/organizations.[1] Once, they argue, the sociological methods such as survey and in-depth interview were the link between the empirical world and the data world, a position most important for generating data about the social. Now, these cornerstone sociological methods have become dated as new digital data including methods for collection and analysis have advanced (Savage and Burrows, 2007). Indeed, it seems likely that social scientists will become irrelevant in computational science if the tendency is really that the 'real life' aspect of data has become superfluous!

Other social scientists do not share the 'crisis' view. Lazer et al. (2009) call for an increased attention to a computational social science. Like Savage and Burrows they argue that the emerging field of big digital data research is currently at the hands of large companies and institutions. However, they see a potential for social science in the computational science scene. Even though they list a number of obstacles, they argue that interdisciplinary research between computational scientists and social scientists, or maybe even computationally trained social scientists and vice versa, would be of great value both to society and the sciences. By embracing both computational and social science collaboratively in a 'computational social science' we might enhance our understanding of the social (Lazer et al., 2009).

Ruppert (2013) also sees the emergence of Big Data research as a potential for interdisciplinary collaboration. Indeed understanding Big Data *requires* interdisciplinary collaboration because methods and expertise for understanding Big Data is distributed over a number of different sciences and other sectors like industry, government and business. Her argument, along with others', is that attention should be paid to the tools and methods of digital data collection (Ruppert, 2013; Ruppert et al., 2013).

The main interests of these studies have been the reconfiguration and development of methods and the facilitation of interdisciplinary cooperation and collaboration with which to work with digital data. Studies such as the CNS are indeed inscribed in, if not born out of, this understanding that dealing with large and diverse digital data, any one science or sector is bound to fall short of understanding its complexity and will only gain very limited, that is to say, science- or sector-specific, insights. One of the many things that motivated the CNS was the possibility for a more

[1] In the 2007 paper Savage and Burrows do not actually use the term Big Data. However, in their 2014 follow-up paper they make clear that the focus of both papers was on Big Data and the influence they predict it will come to have on empirical social science (Savage and Burrows, 2014).

multiplex understanding of social networks using diverse data and engaging in interdisciplinary collaborative research. I have shown here that much attention has been paid to the emergence of digital data, collaboration and computation of the social also within the social sciences. This has led to interesting work on the methodologies of different sciences and how collaboration between these can inspire new approaches and insights into the social world. However, there has been less interest in the basic elements of which any computational analysis consists. A computational analysis with large quantities of digital data is always based on mathematical calculation, which means that they basically consist of two things: digital (quantifiable) data and algorithms. Where there have been studies of what algorithms are and do (Blass et al., 2003; Kockelman, 2013; Wilf, 2013) there has not, in the case of digital data, been much attention paid to data in itself and what data is and does. In these times when more and more digital data is being generated and stored and where the Big Data analysis and computational sciences gain more and more ground it is relevant to ask basic questions about what this digital data is. Let me therefore go on by zooming in on data in the next section by introducing two dominant understandings of data. One is data understood as something out there that can be picked up or collected in its 'raw' form, the other is data understood as something that is always 'shaped' by the process and tools of collection. To make the difference between the two understandings of data clearer I will introduce the two as they are put forth in social science interrogations of digital data.

3. Data as 'raw' and data as 'shaped'

To critically interrogate digital data is what Boyd and Crawford do by offering six provocations that problematize widespread assumptions embedded in work with large digital data sets (Boyd and Crawford, 2012). They point out that there is a tendency for computational science working with social science issues to claim an increased objectivity because of the use of digital data. This is grounded in an assumption of digital data and computation as 'the business of facts and not interpretation' (Boyd and Crawford, 2012: 667).

Boellsdorff (2013) is also directing our attention towards the assumption of digital data as fact. He problematizes what he calls 'algorithmic living' (Boellsdorff, 2013: 2). 'Algorithmic living' is understood as a future that Boellsdorf predicts will emerge if the current treatment and understanding of digital data continues. In this future we will rely more and more on computational (algorithmic) analysis and cease asking questions. Even though there is no unitary definition of what Big Data is, it still has real and increasing effects, not only as a research tool, but also in shaping society. This means that before trying to understand what digital data is, it is already in use for computational analysis, policymaking, commerce, etc. (as I have also shown in the multiple works above). Now, Boellsdorf goes on, there is not necessarily anything wrong with using new technologies for all these things, but when its use is its justification we risk a conversion of 'use' into 'meaning' (Boellsdorff, 2013: 2).

Like Boyd and Crawford, Boellsdorf points out that, at present, there exists a widespread social imaginary about digital data as fact-like, or to phrase it differently, that its collection exists prior to any interpretation, that digital data's basic form is

matter-of-fact or 'raw' (Boellsdorff, 2013: 9), an imaginary that shines through in concepts used in computational sciences such as 'scrapings' or 'ground truth'. However, both Boyd & Crawford and Boellsdorff stress that there is always a theoretical and methodological context that any data is inscribed in and developed from. Theory and method is something that in all sciences is constantly developed, scrutinised changed and revisited and as such this context has an aspect of timeliness to it. The timeliness of the context that data is developed in/from must be considered in order to understand the data at hand and the idea that digital data can be considered anything like 'raw' must be challenged and treated as highly problematic. Instead of buying in on the 'rawness' of digital data, Boellsdorff encourages us to think of digital data as 'a field site amenable to cultural critique and ethnographic interpretation' (Boellsdorff, 2013: 11).

In other words what both Boellsdorff and Boyd & Crawford show is that first of all that there exists a social imaginary about digital data as something 'out there' that can be picked up or gathered using the right tools. In this sense data stands as something factual or 'raw' and this imaginary seems to be dominant not only within computational science, but also to a large extent in society at large. However, they argue, the concept of data (no matter how digital or big) calls for an increased attention to context in the form of, for example, what policies, notions, instruments etc. were at play in deciding what is data and how we should credit it. Let us therefore turn our attention to another way of understanding data that incorporates an extended context.

In anthropology, the background from where Boellsdorff draws his anxieties about the rawness of data, there has long been the understanding that no data can be truly 'raw'. Malinowski (1884-1942), whom is by most considered the founding father of modern social anthropology[2], stressed that every phenomenon encountered by a researcher should be studied in its full context (Eriksen, 2001: 15). He set standards for data collection in anthropology because he argued that there was no way to understand scientific findings without understanding the data and there was no way to understand the data without understanding the underlying 'apparatus' for data collection, that is, the context of how and why the data was collected (Malinowski, 1922: 1-6).

The 'hows' and 'whys' of data collection are of course shared among all sciences. Indeed, all serious computational work includes detailed descriptions of how the data was collected; using what tools and for what purposes and what questions we want to answer. However the 'hows' and 'whys' that Malinowski points to might be slightly different. The potential difference is best highlighted in what is now known as the Writhing Culture Debate in anthropology that took onset in the 60's and 70's and fully flourished in the 80's (Eriksen, 2001: 22). Through the Writhing Culture debate the notion of societies as something that could be generalized about was heavily criticized. Instead anthropologists turned their focus to the individual actors and showed how there could be great variation even in very small societies (Barth, 1975; 1993; Eriksen, 2001: 22). But what is most interesting about the debate, in relation to this article, is that it also marked a shift in how data was understood in anthropology.

[2] More precisely; modern British social anthropology.

There was an increased attention to the tools of data collection (in this case the anthropological researcher) and how the tools would always be inscribed in a context of biases that shaped the data they collected. The 'hows' and 'whys' of data collection were now understood not only as accounts of what tools were used and what questions needed to be answered. Indeed both the tools and the questions were themselves realized as inscribed in complex contexts including power relations, human agency, thought regimes, gender, politics and more (Abu-Lughod, 1989; Nader, 1972). In this understanding there can be no collection of data existing prior to interpretation as data is always shaped by the 'hows' and 'whys' of data collection.

The Writhing Culture Debate was initiated as a critique and full break with the assumption that data could be understood as any kind of 'raw'. This article does also, following the here presented researchers, deal with questioning our understanding of data. Like Boellsdorf this article draws its questioning of the social imaginary of digital data as 'raw' from the anthropological notion of data as shaped or inherently contextually constructed. However, the understanding of data as shaped has its own limitations; for once we have accounted for all the layers of context that are implicit in the tools or methods of collection that shaped our data, the data has once again assumed a form that looks critically fact-like. In any case the question remains how we can approach the 'contextuality' of digital data if not only from the accounting-for-the-tools-of-collection angle? How can we start out with data as our object of analysis instead of ending up with it as a product of a shaping process? The article will here turn to look at how researchers work with digital data in order to know more about data itself. Thus, let me now sketch out an empirical example of such work, what I have previously called 'the experiment'. As stated in the introduction, the article's empirical material, the work of one interdisciplinary collaborative team of scientists, was part of a much bigger project: the CNS. To provide a context whereupon the empirical material can be better understood let me start this section with describing in more detail the CNS, then proceed to describe the underlying idea behind 'the experiment' and finally give an empirical, first-hand account of the very initial proceedings and problems of the teams digital data research.

4. 'The experiment': context and problems

The CNS roughly consists of two separate though fully integrated parts. One part is the SensibleDTU project that is based at the DTU. Here computer scientists manage the data collection, the app and storage development and do computational social network studies of unprecedented width and depth. The other half is the SocialFabric project where researchers from the seven involved sciences, physics, public health, psychology, economy, philosophy, sociology and anthropology, from the University of Copenhagen aim their research at answering a range of questions regarding the formation and importance of social networks.

The data was collected from approximately 800 freshman students at the Danish Technical University (DTU) who were given smartphones with a specially designed app that, using several channels for collection, logged their social interactions every five minutes. The channels of data collection were as diverse as Bluetooth signals, geo-location, call and message logs. On top of that data was collected from

questionnaires pushed to the participating students through their smartphones, Twitter and Facebook information, register data and ethnographic records from one year of fieldwork among the students. The ethnographer was herself equipped with a smartphone that collected her data in the same manner as the participating freshman students (for a more detailed account of the data and collection methods see Stopczynski et al., 2014). Domains of science operate primarily on different types of data and different sciences are interested in different questions and use very different tools and methods to engage with data. The idea of the CNS was to collect data from multiple channels on the same population, to enable researchers from disparate domains of science to work together across field boundaries and draw on the different expertise and results generated by such work and thereby achieve a more diverse range of insights (Stopczynski et al., 2014).

The sub-group within the larger CNS project that is the empirical focus of this article was a team comprised of sociologists and anthropologists. The purpose of forming this specific team was to work in an interdisciplinary, tightknit collaborative manner with data analysis, the work I will from here on call 'the experiment'. The overall idea with 'the experiment' was to work in praxis with heterogeneous types of data by investigating a common empirical object through various data types and methodological approaches. The working thesis was to seek out complementarities of the data types and by that hopefully gain not only more complex insights but also new insights about the empirical object (Blok and Pedersen, 2014). 'Interdisciplinary collaboration' was defined by the team members not only as the shared and diverse data, but also as a methodological approach to the data. An approach that would allow the involved sciences to be inspired by each other's data and ways of conducting research. As one way of practicing interdisciplinary collaboration the team wanted to explore the possibility of doing ethnography in the digital data material (for more detailed accounts on how this was done in practice see Blok et al., forthcoming). What would happen, the team asked, if instead of interpreting data by aggregation, it could be interpreted by means of 'walking around' in it? This approach was facilitated by the kind of digital data gathered and stored in the CNS; its fine granularity and sequential density that allowed for the dynamic behavioural traces of single notes to be followed through large timeframes and the network to be seen from that (or other) notes perspective. In combination with the ethnographic field notes it allowed for a simultaneous quali-quantitative view of data sequences. To find new insights the team would therefore not settle for asking and tracing its own questions in the data, as is often the practice in quantitative research (Blok et al., forthcoming). The team wanted to truly experiment by throwing data types and methods into random compositions, compositions that would in turn force forward new questions in the researchers. With this approach the team wanted to test whether it could chase out new understandings by combining different kinds of quantitative and qualitative data and methods of the kind just described. In 'the experiment' this would be done by engaging with the same empirical object, the object of a social event – a student party held at the DTU. The following will describe the very initial engagement of the team with their object of study, or more precisely all the ado that came before the team could engage in the way they thought they would with their object of study.

This specific social event was chosen on the basis that there was both digital and ethnographic data available and that both had a high degree of granularity and thoroughness. It was decided to start out by looking at the ethnographic data (that would be the ethnographic field notes) of the event in question. The field notes were read aloud to the research group. In the ethnographic data material the team found that the party seemed to have happened in waves from 'dead' to 'intense'. The waves were described and characterized by the connection between the number of student participants at the party and the atmosphere – the feeling of intensity. It was decided to narrow the focus from the party in total to the theme of intensity at the party. The team went on by asking how they could identify and describe the intensity of the party through the digital data material. As a starting point the team wanted to look closer at the waves as social rhythms, as the low and high of students throughout the night of the party.

As described a connection between the number of party participants and the intensity/atmosphere at the party was noticed in the ethnographic account. Here the team chose as a theoretical bridge from the qualitative to the quantitative data, Durkheim's theory of the social event concerning how an event gets more intense with more participants (Durkheim, 1915). With its empirical observation and the Durkheimian theory as backdrop the team decided to use Bluetooth data from the participants' smartphones as one possible medium through which to digitally observe the intensities. Bluetooth was chosen especially because it would enable the team to see both the number of participants and their physical proximity. But here the team ran into problems of a very fundamental kind, because how were they supposed to locate the party and in the same breath find the party participants within the much larger sum of data in the total data pool (from 800 students). The problems centred on 'where' and 'who'.

At this stage the ethnographer was the only one known with certainty to have been attending the party. Likewise her smartphone was known to have certainly been at the party. On this basis it was decided to use the ethnographer's Bluetooth data from that night as a starting point and look for the party and participants by registering what other phones the ethnographer's phone had connected to during the time span of the party. Through this manoeuvre the team hoped to be able to digitally reconstruct the party as a social event using the ethnographer's digital data as a calibration point or perspective, from where to look at the relational topography.

But, after more than an hour's work trying to find the ethnographers data in the large data pool, it was discovered that this data had been removed from the data pool and only existed in its own file. This was done on request from some of our fellow researchers from the CNS out of fear that the ethnographer's data would contaminate the data pool because the ethnographer was not an authentic freshman student.

A sociologist, Ben, from the research team did, however, manage to claim a copy of the total data from the day of the party as well as the ethnographer's data from its special file and merge the two, so that he could trace and map the interaction of the Bluetooth signals directly connecting the ethnographer's phone to other phones. After a full day of work Ben discovers that 30 other phones showed relevant Bluetooth connection to the ethnographer's phone.

The ethnographer notices by now that the field notes indicate that the ethnographer, as well as a number of students, had left their bags and phones in a windowsill at the party. This information about basic human agency at the party questioned the credibility of the number of 30 party-participants as it was calculated on the assumption that the ethnographer, party participants and corresponding phones would have been moving around at the party.

Peter, another sociologist from the team, had worked to locate the party in another way, namely by defining the largest clusters of relevant proximity Bluetooth interaction at the DTU, then inserting the ethnographers Bluetooth data and observe what cluster the ethnographer would belong to. Peter had found a group of about 60, including the 30 Ben had found. Finally and by shared effort it now looked like the team had found the party and the participants, in other words the fundamental data of the research project.

Now to recap: After the decision to narrow the investigation down to 'intensity' at the party the search for the fundamental data happened roughly in four phases: 1) the team looked at the ethnographic data and digital data and arrive, with the theoretical help of Durkheim, at the possibility that atmosphere, number of participants and physical proximity combined might tell something about the intensity. 2) The team looked for the ethnographers digital data in the total data pool. This was made impossible by the fact that the ethnographer's digital data had been excluded out of concerns regarding research-politics and contamination. The ethnographer's digital data was located, claimed and reinserted into our copy of the data set from the night of the party. 3) Using a mathematical model the relevant Bluetooth signals were mapped out. However the credibility of the mapping was obstructed by the likelihood that some of the party participants had left their phones on a windowsill. 4) Finally in an interplay between Bluetooth data, inter-human discussion and a competing mathematical model the search was completed and the data needed for the team's further research was located.

By this note the investigation of how to perceive data could have easily ended in the concluding remark 'that is how the team found our data and now the real research could start' – that is, the research *with* the digital data could start. The above-mentioned account of problems would in that way only show the process of how the fundamental data were found or shaped. It would have been a self-contained story about our trouble terminating in endpoint-data. But let me instead use these problems and ask what we might have learned from them instead of dismissing them as merely a troublesome means to an end. Thus this article will instead ask: what can we learn about data when seemingly banal questions of 'where' and 'who' ended up in matters of competing mathematical models, human agency and different research policies? With this question I hope to direct the focus from a narrative of research by the means of data to a narrative concerned with research of or in data itself. In this light the team's quest to find the fundamental data stands out as a point in its own; for what had happened when the questions of 'where' and 'who' were asked was that the understanding of what data is had to be rethought; instead of understanding the search for the fundamental data as a process that shaped our data, we could see how the fundamental data was comprised of the sum of all the things that had happened and were drawn in to find it. In the following section I will make this point clearer.

5. The monadological perspective

We have the ethnographic and digital material from where we choose the parts about Bluetooth and atmosphere. We have the ethnographer's digital data and the research political considerations of CNS colleagues, the considerations that have changed the position of the ethnographers' digital data to be outside of the large data pool. Finally we have the reinsertion of the ethnographer's digital data and the interplay between that, other Bluetooth data and two mathematical models. The point here is that the events described can be understood in three ways, namely 1) as a process that led to data 2) as a process that shaped data, or 3) as parts of data.

The shift seems small, but it still indicates a change in how to understand data. In the first case there is a prescribed static understanding of data as something stable that 'is out there' and that can be found using the right tools. In the second data is understood as shaped by different contextual factors, but again, once shaped, data figures as something that 'is'. In both these understanding it is possible, by accounting thoroughly for the process through which the data has been shaped, to make the data more objective, creditable or fact-like (Boellsdorff, 2013: 3). In the third case there is no processual understanding in play and therefore there is no understanding of data as less or more fact-like. In the third case the data we have and the way we arrived at it is to be understood as one and the same. Let me explain this last point more elaborately by drawing on Latour et al.'s research on new visualization and navigation possibilities in digital data material.

The focus of Latour et al. (2012) is how search-tools and accessibility of large quantities of digital data in combination with new possibilities of visualization leads to new ways of navigating digital data. They argue that this new availability of digital data allows for a re-evaluation of how notions of micro and macro are understood especially in the social sciences. Social science, they claim, has always operated through the understanding that sociality exists on two levels, a micro level that focuses on individuals and a macro level that focuses on the aggregate (the two level standpoint or 2-LS as they name it in the article). Even though there have been multiple attempts to analytically bridge these two levels (Bourdon, 1981; Bourdieu, 1972; Giddens, 1984) they are still the primary foundation for shaping research questions within social science (Latour et al., 2012: 590-591). Latour et al. go on by claiming that the presupposition that there exist two levels will end up biting its own tail, because when we operate with the presupposition that there exist two levels we cease to keep the content of the levels open for enquiry. The really interesting question, they argue, is not how to get from one level to another but to ask: 'What is an element? What is an aggregate?' (Latour et al., 2012: 591). However, they argue, the 2-LS is now being challenged by the accessibility of very large quantities of diverse data. To demonstrate this they use as an example the situation where you look for information on the Internet about a person that you have a business appointment with. They call this imaginary person Hervé C.

The first thing to do is to search on the name Hervé C. on the internet, find the person, where he is employed, CV, publications, projects or the like. All the things we find about Hervé C. are in the language of Latour et al. called *attributes* and it is through all these attributes that we start to form a picture of Hervé C. until we can say

to ourselves 'Who is *this* actor? Answer: *this* network' (Latour et al., 2012: 593). As such Hervé C. the person that at first meant nothing more to us than just a name, is now understood by us or indeed 'is' a large network of attributes we found during our Internet-search. In our understanding this network of attributes is now connected and perceived as one entity, namely Hervé C., the envelope that encapsulates the network of attributes by one name (Latour et al., 2012: 593).

Following the 'monadological principle' (Latour et al., 2012: 600) the categories of element and aggregate versus micro and macro become dissolved and irrelevant because all elements are themselves aggregates and all aggregates are elements in other aggregates. This, I will now argue, might have the potential to help us understand data in new ways.

As I have described above in a previous section, digital data is usually understood as fact-like in one way or the other. Within computational analysis data was, roughly speaking, seen as something 'out there' that could be collected using the right tools. Within the social sciences data was, equally roughly speaking, understood as something that took shape from the tools and extended context in the process of collecting it. The difference between these two understandings is not to be taken as a breaking point where one is rejected for the sake of the other. Rather the social science understanding can be seen as a continuation of the other in the way that the extended context of the tools helps us to see how data takes shape from the tools of collection. Similarly the step taken from the social science perspective to the monadological perspective of data is not a great leap or break. Rather it is an extension of the idea of context shaping data; in a monadological understanding of data the extended context is not simply outside of data but inside. It is an inversion of the understanding of contextually shaped data. Let me explain it in another way:

If we can understand an element in a network as always consisting of more elements or an attributes cloud, can we then also understand data, one of the two basic elements of computational analysis, as always consisting of more data? The question opens the suggestion that if we understand data as inherently unstable constellations of attributes, something that is already always negotiated, fragile and sensible to changes. To make the argument by way of comparison we can say that if the name Hervé C. is the team's initial response to questions of 'where/who' and the things, persons and events during the team's search are equivalent to the Internet search, then Hervé C. after the search is equivalent to the team's data after finding it. For example, we saw in the empirical example how the team brought in Durkheimian theory at an early stage in their search for the fundamental data. By doing this, the team could be seen to have narrowed the search by defining the kind of data they were looking for 'out there' in the data set. However using the monadological perspective 'defining a kind of data' means building up its specific constellation of attributes, and from that perspective the team has started building its data by providing it with an attribute. The shape that the team's data took was due to the specific combination of attributes mobilised to find it. As another example let us recall that when the ethnographer suggests that she as well as several party participants might have left their phones on a windowsill. Here a sudden fragility of data is revealed because this extra peace of contextual information, or indeed the human agency, made the data change from highly factual to highly questionable. In other words the

composition of attributes in the search mirrored a difference in the data. Also the final use of two different mathematical search models, Ben and Peter's, highlight this point. For adding Ben's model gives us a very different perspective on the party than Peter's model. Adding Ben's model gives us 30 participants – a small size party – whereas Peter's model gives us 60 participants – a regular size party. Two different perspectives individualized by their (slightly) different sums of attributes. To put it in yet another way we can depict the difference like this: sum + mathematical model 1 = sum1 and sum + mathematical model 2 = sum2.

So, what does the encapsulation of the context into data itself mean for the understanding of data in this article? It means that there can never be a piece or pieces of data that does or do not itself consist of data, that any data is always a network of things, persons, events, etc. When we change perspective regarding the context that shaped our data from being 'outside' the data itself to being what any data is made of we can escape the fact-like understanding of data. Rather we can understand the data we ended up with as a stance, a specific perspective of the party that was foregrounding other perspectives, because of that exact combination of the different attributes that enabled us to regard something as 'the data'. There is an almost circular movement to this argument: the way we went about the search for data was, so to speak, what we saw, it was the perspective from where we addressed data and it affected the internal properties of our data which in turn affected how the data was understood.

6. Stepping into data

Let me conclude by summing up the point of this article: by applying the monadological perspective on data we might be able to place data itself at the centre of scientific interrogation and thereby gain new insights about one of the basic essentials in any social science research.

Even though theorists within the social sciences (and other disciplines) have been concerned with the epistemology of data and have developed and treated the concept of data with care and reflection, it is of great importance never to stop challenging assumptions or settle with current paradigms about what data is. The questioning and reconfiguration of dominant understandings of data is even more pressing since data is one of the cornerstones of any social science research; leaving data as something fact-like and only concern ourselves with tools and methods we risk getting stuck with an out-dated set of apparatuses for research. In other words, to develop tools and methods for data research it seems a good idea to reflect on what this data is or we might risk that what we use to research data only shows us what we knew already.

The article has demonstrated, through describing 'the experiment' – a collaborate work of different domains of science with heterogeneous types of data, how the emergence of new types, quantities, qualities and combinations of data, for example in certain types of interdisciplinary collaboration and combinations of heterogeneous data types, has potential to review our understanding of data anew. As we saw, the team did not actually set out to investigate digital data as such, but to investigate using digital data. The data was initially just something needed to get on

with the 'real' research, not something that was to be questioned in itself. Perhaps therefore, the problems encountered when approaching the data inspired this article. It was the deconstruction of traditional research frames in 'the experiment' that brought forward self-reflection and the meta-question 'what are the data we are working with?'. This article has been an attempt at answering just that.

Thus, the concern of the article was with the implicit understandings of data that form the basis of the majority of social science research. Two such understandings were drawn out, that of data as 'raw' and data as 'shaped'. This article has demonstrated how we might gain a different perspective on data if we regarded data as 'monadic'. The empirical material illustrated how it was possible to change focus from an understanding that the extended context of data would be something outside of data itself, that the data had a final form after the search process had finished, to regarding data as the sum of the attributes that comprised the search. By this shift of perspective, I argued, we can gain the possibility of analytically stepping into data itself. The leap from outside of data to inside is first of all interesting, because, as I have shown, most research on data is focussed on the methods and tools of data research and not on data itself. However, taking data itself as the point of departure for analysis raises three other points of interest for social scientists: 1) a new field of analysis: data itself as an object of analysis. 2) A new method for analysing data: by identifying the attributes that comprise the perspective that is our data we gain the possibility to open up every single attribute and identify what they consist of (political, mathematical, human, technical etc.). 3) A new epistemological understanding of data: by changing how we see the object of study we might be able to develop even more new tools and methods for analysis.

If we understand data as the relation of attributes we can take the step into our data and investigate it from within. As such it would be possible to move not only between data, but *within* data. Might there be, by adding the monadological perspective to the repertoire of understandings of data, an increased possibility for scientific research not only with digital data but also within data itself?

References

Abu-Lughod, L. (1989) Fieldwork of a Dutiful Daughter. In Altorki, S. and C. F. El-Sohl (Eds.) *Arab Women in the Field: Studying Your Own Society.* Cairo: Syracuse University Press. 139-162.

Aral, S., Brynjolsson, E. and M. Van Alstyne (2007) Productivity Effects of Information Diffusion in E-Mail Networks. *ICIS 2007 Proceedings,* 17. Available at: http://ebusiness.mit.edu/research/papers/234_VanAlstyne_Productivity_Effect_Info_Diffusion.pdf Accessed: 20-03-2017.

Blass, A. and Y. Gurevich (2003) Algorithms: a quest for absolute definitions. *Bulletin of European Association for Theoretical Computer Science 81.* Available at: https://www.microsoft.com/en-us/research/wp-content/uploads/2017/01/164.pdf Accessed: 20-03-2017.

Blok, A., Bornakke, T., Carlsen, J. A. B., Ralund, S., Madsen, M. M., and M. A. Pedersen (forthcoming) The Heterogeneous Party: An Experiment in Big Data from the Bottom Up. *Big Data & Society*.

Blok, A. and M. A. Pedersen (2014) Complementary Social Science?: Quali-Quantitative Experiments in a Big Data World. *Big Data and Scociety,* 1(2): 1-6. DOI: http://dx.doi.org/10.1177%2F2053951714543908

Boellsdorff, T. (2013) Making Big Data, in Theory. *Firstmonday,* 18(10). DOI: https://doi.org/10.5210/fm.v18i10.4869

Boudon, R. (1981) *The Logic of Social Action: an Introduction to Sociological Analysis*. London: Routledge.

Bourdieu, P. (1972) *Outline of a Theory of Practice*. Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/cbo9780511812507

Boyed, D. and K. Crawford (2012) Critical Questions for Big Data. Provocations for a Cultural, Technological and Scholarly Phenomenon. *Information, Communication & Society,* 15(5): 662-679. DOI: https://doi.org/10.1080/1369118x.2012.678878

Centola, D. (2010) The Spread of Behaviour in an Online Social Network Experiment. *Science*, 329(5996): 1194-1197. DOI: https://doi.org/10.1126/science.1185231

Clifford, J. and G. E. Marcus (1986) *Writing Culture. The Poetics and Politics of Ethnography.* Berkeley: University of California Press.

Davies, W. (2013) Empirical Limits. Policy Makers are Mistaken if They Think Legitimacy is Merely a Question of Being Led by Evidence Based Data. *RSA Journal,* (4): 36-39.

Durkheim, E. (1915) *The Elementary Forms of the Religious Life*. London: G. Allen&Unwin.

Giddens, A. (1984) *The Constitution of Society*. Cambridge: Blackwell.

Gilbert, N. (2005) Computational social science: agent-based social simulation. Manuscript submitted for publication.

Eriksen, T. H. (2001) *Small places, Large issues.* Oslo: Universitetsforlaget

Knox, H. and D. Nafus (n.d.) (Eds.) *Big Data and Ethnography.* Manchester: Manchester University Press. Manuscript submitted for publication.

Kockelman, P. (2013) The Anthropology of an Equation. *HAU: Journal of Ethnographic Theory,* 3(3): 33-61.

Latour, B., Jensen, P., Venturini, T., Grauwin, S. and D. Boullier (2012) 'The Whole Is Always Smaller Than Its Parts; A Digital Test of Gabriel Tardes' Monads. *The British Journal of Sociology,* 63(4): 591-615. DOI: https://doi.org/10.1111/j.1468-4446.2012.01428.x

Madsen, M. M., Blok, A. and A. M. Pedersen (forthcomming) Transversal Collaboration: an Ethnography in/of Computational Social Science. In Knox, H. and D. Nafus (Eds.) *Big Data and Ethnography*. Manchester: Manchester University Press.

Malinowski, B. (1964/1922) *Argonauts of the Western Pacific.* London: Routledge & Keegan Paul.

McInerney, J., Stein, S., Rogers, A. and N. R. Jennings (2013) Breaking the Habit: Measuring and Predicting Departures from Rutine in Individual Human Mobility. In: *Pervasive and Mobile Momputing,* 9(6): 808-822. DOI: https://doi.org/10.1016/j.pmcj.2013.07.016

Nader, L. (1972) Up the Anthropologist: Perspectives Gained from Studying Up. In Hymes, D. (Ed.) *Reinventing Anthropology*. New York: Phanteon Books.

Perrig, A., Szewczyk, R., Tygar, J. D., Wen, V. and D. E. Culler (2002) SPINS: Security Protocols for Sensory Networks. *Wireless Networks, (8): 521-534.*

Ruppert, E. (2013) Rethinking Empirical Social Science. *Dialogues in Human Geography,* 3(3): 268-273. DOI: https://doi.org/10.1177/2043820613514321

Ruppert, E., Law, J. and M. Savage (2013) Reassembling Social Science Methods: The Challenge of Digital Devices. *Theory, Culture, Society,* 30(4): 22-46. DOI: https://doi.org/10.1177/0263276413484941

Savage, M. and R. Burrows. 2014. After the crisis? Big Data and the methodological challenges of empirical sociology. Big Data and Society, 1(1) DOI: http://dx.doi.org/10.1177%2F2053951714540280

Savage, M. and R. Burrows (2007) The Coming Crisis of Empirical Sociology. *Sociology,* 41(5): 885-899. DOI: https://doi.org/10.1177/0038038507080443

Sonnenwald, D. H. and L. G. Pierce (2000) Information Behaviour in Dynamic Group Work Contexts: Interwoven Situational Awareness, Dense Social Networks and Contested Collaboration in Command and Control. *Information Processing and Management,* 36(3): 461-479. DOI: https://doi.org/10.1016/s0306-4573(99)00039-4

Stopczynski, A., Sekara, V., Sapiezynski, P., Cuttone, A., Madsen, M., Larsen, J. E., and S. Lehmann (2014) Measuring Large-Scale Social Networks with High Resolution. *PLoS ONE*, 9(4): e95978. DOI: https://doi.org/10.1371/journal.pone.0095978

Tomkins, C. (2001) Interdependence, Trust and Information in Relationships, Alliances and Networks. *Accounting, Organizations and Society,* 26(2): 161-191. DOI: https://doi.org/10.1016/s0361-3682(00)00018-0