# **The Academic Organisation of Data Science**

# **Introduction**

Around ten years ago so called E-Science institutes starting popping up all over the world, as developing computational methods was high on the research calendar in academia. At the time, these institutes were started to “understand digital devices [and develop] conceptual framings and innovative methods for analysing their effects” (Ruppert, Law & Savage, 2013, p. 29). Today, the buzzwords are arguably ‘Data Science’ and ‘Big Data’ as there are now roughly 40 academic Data Science institutes (DSIs) – sometimes called centres or initiatives. All of these except for one have come into existence between January 2013 and December 2015. In addition, there are now about 120 postgraduate degrees related to terms like Data Science, business intelligence, or advanced analytics around. “Data Science,” and the related “Big Data” surface more and more; academic journals like “Big Data & Society” are now publishing articles about the Big Data phenomenon. But what do Big Data and Data Science refer to? What do these terms mean? And how do these phenomena impact on the organisation of academia?  
 Big Data is generally seen as referring to the huge data sets that are now coming into the possession of companies and universities, sets of data which are so big they are changing the ways in which research is done. Big Data is a phenomenon which is both scholarly, technological and cultural, resting on “the interplay of technology, analysis and mythology” (Boyd & Crawford, p. 663). It is a phrase which “seems to have mysteriously emerged from nowhere over the last couple of years and now it commonly appears in newspaper headlines, the pronouncements of politicians, and the promises made by marketing groups” (Warwick, p. 1). As Fairfield and Shtein argue, “Big Data […] is big news” (2014, p. 38). According to Boyd & Crawford, 2012, “The era of Big Data is underway. Computer scientists, physicists, economists, mathematicians, political scientists, bio-informaticists, sociologists, and other scholars are clamouring for access to the massive quantities of information produced” (p. 663).   
 Data Science is defined by Waller & Fawcett as “the application of quantitative and qualitative methods to solve relevant problems and predict outcomes” (Waller & Fawcett, 2013, p. 78) especially with regards to Big Data. Data Science can thus be seen as the academic response to the sets of Big Data which are available – although one could also argue in favour of simply seeing Data Science as similar to Big Data but happening in a more academic setting. In a leaflet written to promote Imperial College London’s Data Science Institute, it is written that Data Science is “an essential element of all modern interdisciplinary scientific activities” (p. 3) and that “Data Science is not only concerned with the tools and methods to obtain, manage and analyse data, it is also about extracting value from data and translating it from asset to insight” (p. 3).   
 Data Science at the same time make use of methods from many different disciplines all situated “within the broad areas of mathematics, statistics, and information technology” (Diggle, 2015, p. 794). Data Science consists in this view of statistics *and* informatics, which refers to the technology, software and hardware needed (the infrastructure) to convert huge amounts of data into usable formats (Diggle, p. 794). The organisational problem with statistics is that it is a discipline with relations to multiple other disciplines but is usually put together with mathematics in terms of research departments. In Diggle’s opinion universities need to have statistics institutes in order for academics to work together there *and* within a university department. The DSIs which are coming into existence might achieve this goal as well, or so Diggle argues: “Deep involvement of statisticians within the burgeoning number of Data Science institutes might be a more effective tactic to achieve the same goal, at least in the short term” (p. 806). Collaboration is especially important for a discipline like statistics, as the unique selling point of this discipline is that it can be relevant “to the whole of the natural and social sciences” (p. 807). Data Science, and specifically mathematics and statistics should in Diggle’s view work together “when making the case for the fundamental importance of the mathematical sciences to the future health and wealth of UK society” (805).   
 Data Science is different from former ways of analysing data in that the data nowadays is more heterogeneous and unstructured, “emanating from networks with complex relationships among its entities” (Dhar, 2012, p. 2). In order to understand this kind of data, collaboration is needed, as different academic fields all have their own particular skill set which is argued to be not enough in itself to deal with Big Data: “From an engineering standpoint, it turns out that scale matters in that it has rendered the traditional database models somewhat inadequate for knowledge discovery” (Dhar, 2012, p. 3). This is the problem with being a data scientist. There are no individuals “that can possibly have all of what is needed by a data scientist” (Waller & Fawcett, 2013, p. 78). This is also partly why there are so many Data Science Institutes all around the world: to deal with the fact that new opportunities of data collection require new models and methods and collaboration.

# **Research interests**

Since it is estimated that “90% of the data in the world today has been created in the last two years alone and the world’s data will grow by 50 times in the next 10 years” (Technische Universiteit Eindhoven, Web) it is not surprising that DSIs are formed and that social scientists are now writing about Data Science, business analytics and topics like algorithms, as people are realising that (big) data might become *the* power of the 21st century. It is not surprising that Data Science is high on the research agenda, but how do these institutes actually work? And what kinds of data do they work with? Some of the topics that we are interested in discussing by presenting an overview of these institutes are: (1) the different definitions of Data Science as presented by these institutes; (2) the goals of these institutes in terms of output and collaborations; (3) the relations to both existing and emerging academic disciplines and (4) relations to other worlds, like civil society and industrial Big Data practices. By focusing on these four topics, we can look at how new forms of knowledge are (allegedly) being created within the DSIs we discuss and how academic organisation is affected by the Big Data phenomenon. The question we then aim to answer with theses research efforts is the following, as we ask: how does the Big Data phenomenon have an impact on the organisation of academia?

# **Methodology**

One field which has paid attention to the organisation of scientific work since the late 1970s, is STS, a discipline which has “added significantly to our understanding of science” (Roth & McGinn, 1998, p. 216). In particular STSers have been publishing about practice oriented laboratory ethnographies, controversy studies, and historical portraits of science. Nowadays, STSers “take for granted that an understanding of scientific *knowledge* requires a thick engagement with scientific *practice*” (Mody, 2015, p. 1). We propose to study Data Science institutes by looking at scientific practices and infrastructures from an STS perspective. STS literature focusing on knowledge production could prove valuable for discussing today’s scientific infrastructures in terms of how academia is organised, what the goals are of the Data Science institutes, who is involved, and what the relations are between the DSIs and industry and civil society. It could be argued for example that knowledge has moved from the meticulous work in the laboratory to (online) Big Data communities.  
 There is a lot of literature in STS on the training of scientists (Roth & McGinn, 1988; Kohler, 1987). These studies show how “pedagogical institutions can mirror and drive wider cultural change” (Hackett, Amsterdamska, Lynch & Wajcman, 2008, p. 378). With the commercialisation (and arguably de-commercialisation) of science, this topic has never been more relevant than today, as relations between consumers and producers of knowledge are renegotiated with the emergence of DSIs, some of which aim to restructure relationships between academia and industry, and the whole assemblage of how science is organised. One particular STS journal in which many relevant literature is published on both scientific practices and the training of scientists is Big Data & Society. Texts from this journal will especially prove insightful when discussing the nature of the DSIs around the world.   
 In order to expand this project we invite collaboration from researchers from other universities. We do this by making both a Google Map and a Google Spreadsheet containing key information about the DSIs which are publicly accessible and everyone can comment on. This way, we hope to get other academics, not necessarily social scientists, involved in this project as we believe this could lead to more information with regards to the different DSIs worldwide. At this point in time, the map simply presents and overview of where these institutes are, while the spreadsheet contains information, which could provide the basis for a comparison, on the following: the locations of the DSIs; the degrees offered by the universities linked to these institutes; the starting dates of the institutes; information regarding funding and the relationship between the institutes and industry; a list in terms of academic disciplines of who are collaborating in the institutes, how many people are involved and who these people are. In the following we start to compare key themes that surface when studying the information presented by the DSIs by incorporating STS literature and reflecting on our own experiences in working in the DSI at Lancaster University.

# **Some preliminary findings**

As mentioned before, our main goal for this project is to find out how the Big Data phenomenon has an impact on the organisation of academia. We start by discussing the different definitions of Data Science as presented by the DSIs on their respective websites. When looking at these websites there are always short lists with bullet points about certain goals and aspirations, yet there are hardly any DSIs which actually explain what Data Science is (in their view). They all talk about Data Science and say what they aim to achieve, or who is involved in the centre or institute, but they hardly discuss what Data Science is. Why? Is it perhaps because as a term it is as ungraspable or opaque as Big Data? Take this descriptions for example, copied from the websites of the DSIs in Manchester and Michigan:

“Manchester's Data Science Institute acts as an access point to the University’s expertise in data science, facilitates interactions between data science researchers and problem holders, owns the University’s data science strategy, and will deliver sustainable support for the community (Manchester, Web).

“Data science is now widely accepted as the fourth mode of scientific discovery, on par with theory, physical experimentation and computational analysis. Techniques based on Big Data are showing promise not only in scientific research, but also in education, health, policy, and business” (Michigan, Web).

Both of these descriptions illustrate that Data Science is now generally accepted as a way of creating knowledge and that there are researchers ‘doing Data Science’. Yet there is hardly any information on how these institutes work or aim to work, or with which data, or even what Data Science is. This lack of clarity can be explained by the fact that most DSIs are only just beginning to surface and they lack clear definitions and there are hardly any projects within these DSIs which are finished. It seems almost as though these institutes are starting to pop up more and more as reactions to each other, or perhaps to what happens in industry, without the institutes themselves having a clear idea of what needs to be done in the fields related to Data Science, which arguably are all academic fields.   
 Of course, there are exceptions. In Dundee, where the first DSI was founded already in 2011, two years before the second DSI sprang to life, there are academics with a clear idea of how Data Science should be tackled. Even though the following description still does not explain what Data Science is or should be like, it does at least give an indication of what a data scientist is and what kinds of data they work with:

“There has been a recent upsurge of commercial interest in both Business Intelligence (BI) and Data Science (DS). […] BI is about extracting useful information from a mass of raw data.  A large number of systems, techniques and processes can all be involved in doing [this.] DS has much in common with BI but Data Scientists tends to work more with ‘big data’ and have a greater focus on developing custom algorithms and visualizations.  One good definition of a Data Scientist is that they know more statistics than a programmer and more about programming than a statistician.  The term data scientist doesn’t imply that scientific data is involved, although it certainly can be; most data scientists work on commercial data” (Dundee, Web).

The differences between some of the terms used in the DSIs never really become clear though. In Dundee, business intelligence is mentioned, in other parts of the world the word “analytics” is often used in relation to both Big Data, Data Science and Business, but what Data Science entails is never entirely clear. Most DSIs have a description on their website of what their goals are but even these are not completely transparent. As far as we could distinguish, the only description of Data Science is given by Boston University, and this definition still does not say much, as Data Science is defined as “the methodical extraction of knowledge from data” (Web). The question is, with about 40 institutes around the world all working on Data Science, whether they do not define the topic of their research because they (A) do not know themselves what Data Science precisely is, or (B) assume that everyone who reads their websites knows what Data Science is. If the latter is the case, we would like to thank Boston University for being the only university to explain the term.  
 The second question we are interested in is what the goals of the DSIs are in terms of output, like publications, degrees, and collaborations. There are two main goals which come back most often when looking at the different lists of bullet points or explanations presented by the DSIs as to their goals. First of all, as Sabina Leonelli argues, “the novelty of Big Data science lies in (1) the prominence and status acquired by data as scientific commodity and recognised output both within and beyond the sciences and (2) the methods, infrastructures, technologies and skills developed to handle […] data” (2014, p. 2). Most DSIs state that the main purpose of the DSI is to get scientists together from different disciplines to collectively develop frameworks and methods to deal with Big Data. Leonelli is right when she argues that data is seen as a commodity and DSIs are working on methods to delve this commodity. This goal can be described as aiming to do “world-leading research” (Warwick, Web) or “synergis[e] different fields of expertise” (Delft, Web). In Edinburgh, the DSI “acts as a hub for national and international data across the arts and sciences; developing networks of expertise in managing and accessing all forms of data” (Web). The point is clear: collaboration is high on the research agenda, as together academics can work on methods to analyse big amounts of data.   
 Next to this, there are those who write about teaching and training data scientists. In Bournemouth, the first goal listed is to create a platform for “interdisciplinary training of highly skilled and internationally excellent researchers and leaders in the Data Science area covering big data, advanced/predictive analytics, data intensive computing and their innovative business, engineering and science applications, in a cross-disciplinary environment” (Web). The mission in Virginia meanwhile is “to achieve recognized excellence in research and education in the interdisciplinary field of data science” (Web) and at Columbia the DSI is “training the next generation of data scientists and developing innovative technology to serve society” (Web). It is perhaps surprising that there is only one institute which actually mentions (research) ethics with regards to the collection of big amounts of data.   
 There is a third goal which is important, we would argue, but this does not come up as often on the DSIs websites. An exception is the University of Essex’s DSI, where there is research into the “ethical, legal and human rights aspects of data” (Web) which is a very important side to Data Science. We would argue that as the amounts of data which are available are increasing and having an impact on research methods, the question of how we deal with this in terms of values and ethics is an important one. This is thus concerned with studying what the consequences are of the availability of big amounts of data for academia and industry. Related to this, there is an article in Big Data & Society by Rob Kitchin (2014) who writes that because of the availability of big data we are now arguably moving into a fourth paradigm of science. After experimental science, theoretical science, and computational science there is now exploratory science which is “data-intensive; statistical exploration and data mining” (2014, p. 3). The question raised at the University of Essex, about the ethical aspects of data, is therefore a vital one, as the question is how do we make use data as academics and what values do we try to uphold when constructing knowledge?  
 The third question we are interested in discussion is concerned with the relations of DSIs to both existing and emerging academic disciplines, as we wonder: who are collaborating in the DSIs? Thinking about the disciplines which are involved in the different DSIs is very important, as studying which disciplines are represented in DSIs and are doing Data Science can lead to new views on existing and emerging disciplines and changes in epistemologies. As Kitchin (2014) argues, “there is little doubt that the development of Big Data and new data analytics offers the possibility of reframing the epistemology of science, social science and humanities, and such a reframing is already actively taking place across disciplines” (Kitchin, 2014, p. 10). We already gave a few examples of the disciplines involved, and the goals for collaboration, but not explained really in broad lines who is involved in most of the DSIs.   
 As Big Data and Data Science can be about almost any topic, there are many differences in the kinds of disciplines involved, based on the foci of the universities. In general, there are almost always computer scientists, mathematicians, and statisticians working in the DSIs, while engineers are also very common. Of course this can be explained. In Lancaster for example health and environmental issues are high on the research agenda. Therefore, it is logical that Data Science practices in Lancaster focus on developing methods and frameworks for the natural environment. Collaborations in the different DSIs are really about combining the disciplines strongly represented at universities. In Chicago for instance, the DSI is a collaboration between policy experts and computer scientists who are trying to create “computational and data-driven solutions to large-scale social problems in areas such as healthcare, education, sustainability, and community development” (Web).  
 The final question is how academic practices relate to industry. As Beer & Burrows (2013) argue: “digital data inundation is not just a narrow technical methodological matter for the social sciences; it has been argued that it has far broader implications for disciplinary jurisdiction, the relationship between the academy, commerce and the state, and, indeed, for the very nature of the sociological imagination” (p. 47-48). But how does this work out in reality? What are relations like between commerce and academia and the state? There are big dissimilarities between the different DSIs. Some DSIs have for example grown out of collaborations between academia and industry, like Imperial College London’s institute. Other centres do not mention any industry partners, some of which state that they are interested in working with industry and some of which actually already do so but refrain from naming companies.

**What needs to be done after answering these questions?**

* Discuss social science in particular. What is the relation between big data/data science and sociology like? In Lancaster and in other places.

Useful for this is Sophie Mützel (2015) in Big Data & Society:

“projects using Big Data, from data journalism to computational social science, have little engagement with sociology, although many sociological insights could strengthen analyses; in turn, sociologists could benefit from enhanced computing and visualisation skills” (p. 3).

“Big Data and its methods of analysis challenge the praxis of doing sociology. But, to be sure, sociology has much to contribute to the new arenas of social science research: because of its insights and techniques to study meaning and how the social is structured, sociology makes itself very relevant to data science projects mining large data sets” (p. 3).

# **Works Cited**

Agarwal, R., & Dhar, V. Big Data, Data Science, and Analytics: The Opportunity and Challenge for. *Information Systems Research* 25, 443–448.

Beer, D., & Burrows, R. (2013). Popular Culture, Digital Archives and the New Social Life of Data. *Theory, Culture & Society* 30(4), 47-71.

Boston University. Data Science Initiative, Overview. Retrieved from <http://www.bu.edu/datascience/about-dsi/>

Bournemouth University. Data Science Institute. Retrieved from <https://research.bournemouth.ac.uk/centre/data-science-institute/>

Boyd, D., & Crawford, K. (2012). Critical Questions for Big Data. *Information, Communication & Society* 15(5), 662–679.

Columbia University. Data Science Institute, About. Retrieved from <http://datascience.columbia.edu/columbia-data-science>

Diggle, P. (2015). Statistics: A Data Science for the 21st Century. *Journal of the Royal Statistical Society*. 178(4), 793-813.

Fairfield, J., & Shtein, H. (2014). Big Data, Big Problems: Emerging Issues in the Ethics of Data Science and Journalism. *Journal of Mass Media Ethics* 29(1), 38–51.

Hackett, E.J.., Amsterdamska, O., Lynch, M., & Wajcman, J. (2008). *The Handbook of Science and Technology Studies.* Cambridge: MIT Press.

Imperial College, London. Leaflet: Data Science Institute.

Kitchin, R. (2014). Big Data, New Epistemologies and Paradigm Shifts. *Big Data & Society* 1(1), 1-12.

Kohler, R. (1987). Science, Foundations, and American Universities in the 1920s. *Osiris*. 3, 135-164.

Latour, B., & Woolgar, S. (1979). *Laboratory Life: The Social Construction of Scientific Facts*. London: Sage.

Leonelli, S. (2014). What Difference Does Quantity Make? On the Epistemology of Big Data in Biology. *Big Data & Society* 1(1), 1-11.

Lynch, M. (1985). *Art and Artifact in Laboratory Science: A Study of Shop Work and Shop Talk in a Research Laboratory.* London: Routledge & Kegan Paul.

Mody, C. (2015). Scientific Practice and Science Education. *Science Education*. 99, 1026–1032.

Mützel, S. (2015). Facing Big Data: Making Sociology Relevant. *Big Data & Society* 2(1-4).   
Polanyi, M. (1972). *Personal Knowledge: Towards a Post-Critical Philosophy*. New York: Haper Torchbooks.

Roth, W.M., & McGinn, M.K. (1998). Knowing, Researching, and Reporting Science Education: Lessons from Science and Technology Studies. *Journal of Research in Science Teaching*. 35(2), 213-235.

Ruppert, E., Law, J. & Savage, M. (2013). Reassembling Social Science Methods: The Challenges of Digital Devices. *Theory, Culture & Society* 30(4), 22-46.

Technische Universiteit Eindhoven. Data Science Center Eindhoven, Organisation. Retrieved from [https://www.tue.nl/universiteit/faculteiten/wiskunde informatica/onderzoek/onderzoeksinstituten/data-science-center-eindhoven dsce/organization/](https://www.tue.nl/universiteit/faculteiten/wiskunde%09informatica/onderzoek/onderzoeksinstituten/data-science-center-eindhoven%09dsce/organization/)

Technische Universiteit Delft. Delft Data Science. Retrieved from <http://www.delftdatascience.tudelft.nl/>

University of Chicago. Center for Data Science and Public Policy. Retrieved from <http://dsapp.org/>

University of Dundee. Data Science Centre. Retrieved from <https://blog.dundee.ac.uk/datascience/>

University of Edinburgh. Edinburgh Data Science, Aims and Goals. Retrieved from <http://www.ed.ac.uk/data-science/about-us/aim>

University of Essex. Institute for Analytics and Data Science, Our Research. Retrieved from <https://www.essex.ac.uk/iads/research/default.aspx>

University of Manchester. Data Science Institute. Retrieved from <http://www.datascience.manchester.ac.uk/>

University of Michigan. Michigan Institute for Data Science, About MIDAS. Retrieved from <http://midas.umich.edu/about/>

University of Virginia. Data Science Institute, About. Retrieved from <https://dsi.virginia.edu/about>

Waller, M.A., & Fawcett, S.E. (2013). Click Here for a Data Scientist: Big Data, Predictive Analytics, and Theory Development in the Era of a Maker Movement Supply Chain. *Journal of Business Logistics* 34(4), 249–252.

Waller, M.A., & Fawcett, S.E. (2013). Data Science, Predictive Analytics, and Big Data: A Revolution That Will Transform Supply Chain Design and Management. *Journal of Business Logistics* 34(2), 77–84.

Warwick University. Leaflet: Year of Big Data.

Warwick University. Warwick Data Science Institute. Retrieved from <http://www2.warwick.ac.uk/fac/sci/wdsi/>