Critiques of Data science

Hanne Keijzer, 6194419

2022

Can data science make the scientific method obsolete?
Master thesis History and Philosophy of Science

Contents

1 Introduction

Once, science was done for the love of it, and those who wanted to benefit from it were dubbed not real scientists. In the words of the legendary Simon Newcomb 'Nature turns a forbidding face to those who pay her court with the hope of gain, and is responsive only to those suitors whose love for her is pure and undefiled.' [26, p.223]

Times have changed. As we will see, contrary to the idea of doing science for the love of knowledge, science's newest hype, data science, is fully aimed at providing immediate applicable products. [42]

This thesis aims to investigate the promise of data science, and will ultimately form a fundamental critique of data science. Thereby trying to start a much needed discussion of the current use of data science, and the claim that the classical scientific method has become obsolete because of data science.

For this purpose, we need to analyse data science and what parts of the scientific enterprise data science might utilise or ignore. We will therefore start with a brief introduction of data science, followed up by a characterisation of data science as illustrated by Chris Anderson's article 'The end of theory.' In this article, data science is described as making the scientific method obsolete where it comes to the production of scientific understanding and knowledge. Another important claim Anderson makes is that through data science, we are able to, and ought to, let the data speak for itself. This thesis will offer fundamental challenges to these two arguments by Anderson.

Although Anderson's article is by no means a scientific article, I believe that Anderson correctly grasps the current practice of data science, and I will underline his ideas and give power to his description by showing its correctness in line with a specific, real data science study. The investigation of his claims will be by no means exhaustive, but this thesis will try to illustrate that data science is in line with both of Anderson's claims. Scientific arguments in favour and against Anderson's article will also be considered. The first part of this thesis will close with a brief look into the history of science, to be able to place the claims made by Anderson in a bigger perspective.

The second part of this thesis will focus on critiques of data science. The relation between the first and the second part of this thesis is that in the second part, Anderson's two main ideas, which were illustrated in the first part, will be challenged. Another relation between the parts is that in the first part critique on Anderson's paper will be discussed, and we will conclude that these arguments do not *fundamentally* challenge data science as Anderson describes it. This critique can be overcome by doing *better* data science. But in the second part, fundamental challenges will be voiced.

The fundamental critique will be of two kinds. First an already existing critique of Dan McQuillan will be investigated and appended. This critique attacks Anderson's idea that we ought to let the data speak for itself. The critique will be appended by forming a foundation of McQuillan's argumentation based on the words of data scientists themselves.

Second, my original, fundamental critique will be voiced. This critique attacks Anderson's idea that data science makes the classical scientific enterprise obsolete

where it comes to the production of scientific knowledge. Because data science does not make use of the classical scientific method, data science loses some of that method's benefits. My concerns can be summed up as follows: First, negative results in data science do not contain the benefits that negative results have in the classical scientific enterprise. Second and last, I believe certain methods of data science are poorly reproducible, if reproducible at all.

Data science is an immensely useful tool and offers great benefits when used correctly. But I believe its fundamental challenges are not properly discussed. This work tries to start the much needed, but not yet held debate about data science's fundamental flaws. And whether you ultimately agree with me or not, we need to have this discussion.

2 Part 1: Analysis of data science

Data science is a new and booming part of science, used abundantly. However, there is a debate going on about what this new practise is exactly, and what its implications are. To be able to discuss the risks of the current data science practice in the latter part of this thesis, it is important to have a shared and correct interpretation of what data science is and how it is being used. We will start with a brief introduction of data science and its methods. It is not possible that we exhaustively discuss these methods, which will become clear shortly.

Next, we quickly move to the central focus of the first part of this thesis: an influential article about data science and its promises, written by former editor in chief of WIRED, Chris Anderson. This article makes several strong claims, the most central are the following two: Data science replaces and ought to replace the classical scientific method where it comes to the creation of scientific knowledge, and; Through data science we can let the data speak for itself. Although Anderson's article is by no means a scientific article, it has been very influential in the scientific debate of data science. Because it is not a scientific article, it is important to make clear that, and why, his article warrants a scientific reaction. For this purpose, this thesis takes multiple steps. We will discuss a case study of data science on cancerous cells to illustrate the correctness of Anderson's claims. Next, scientific reactions to Anderson's article will be discussed. Last, we will investigate a bit of history of science, in which the claims made by Anderson will be put into a larger frame.

2.1 Big data analytics

A few years ago, the enormous amount of data was seen as a serious problem because we lacked the tools to store, analyse, and therefore use it. [35, p.4] Likewise,

for a long time scientists were not able to utilise their data fully. [6, p.8689] Nowadays, the analysis of large amounts of data in scientific context is a fast-growing and influential practise.

There is an abundance of names to describe the practise of analysing large volumes of data, what makes data science different from any of these? In current practise, nothing! Big data analytics, advanced analytics, data mining, data science and large-volume analytics all describe projects and technologies aimed at analysing large amounts of data which connect statistical and computational methods. [9] [18] [42] [8, p.50] [35, p.4,10] They describe the practise of applying advanced analytic techniques to large volumes of data to gather information, facts or predictions of a specific phenomenon, combined with the statistical, technical and domain specific knowledge necessary for the proper execution of this analysis. [35, p.8-9] [6, p.8690-8691] [15, p.2863]

Although it is difficult to pin big data and data science down and give a strict definition of data science, there is a popular trend of vaguely describing data science with the use of the three v's. [35, p.7] [18, p.1] Big data is characterized as the analysis of vast amounts of data - Volume - at a high frequency - Velocity -, in a Variety of ways. [23, p.177] The V of Volume speaks to the large amounts of data handled, and the depth or breath of the data. Velocity speaks to the frequency at which data is generated, shared and analysed. Variety speaks mostly to the different ways in which the data can be combined. It also speaks to the different ways in which different types of data are gathered, structured, warehoused and represented. [42, p.33-34] [32, p.19] [35, p.6-7] [15, p2866] [18, p.1-2]

2.1.1 Methods of data science

Data science is an umbrella term for a variety of ways to analyse big data. Data science makes use of a vast different array of methods to analyse data, which all find their roots in statistics and machine learning. [9] [10] [42] Machine learning algorithms can find anomalies, patterns and correlations in large data sets. [13] It distils patterns from the data and is able to relate these patterns to new data, thereby predicting the characteristics of new data. [34] [18, p.101] also see [42] [14] The variety of machine learning algorithms is so vast that it is by no means possible to discuss them all here. However, to capture the essence of data analysis, we must discuss two aspects which are of crucial importance: the methods of data science and the goals of data science. [42, p.43]

Pattern recognition is a method of data science focussed on distilling correlations from data. The two most important types of pattern recognition are clustering and association rules. see [42, p.43] [9, p.3] [10, p.751]

With clustering, algorithms scan datasets in search of similarities between variables, with the aim of clustering similar variables. Thereby classifying and dividing

the data and mapping the strength of correlations and possible overlap between them. see [42, p.43] [9, p.3] [10, p.751]

Association rules are algorithms that search the data to find patterns of variables that are correlated. Here machine learning algorithms check what the relations are between different variables with the aim of distilling these (cor)relations. see [42, p.43] [9, p.3] [10, p.751]

Next to different methods of analysis there exist different goals of the analysis. The most important goals are those of predictive analysis and prescriptive analysis. [42] A predictive analysis is aimed at making predictions of data with certain characteristics based on data with similar characteristics, or making predictions about characteristics of data based on other characteristics of a specific item in the data. [42] Prescriptive analysis is similar but goes one step further than predictive analysis, and advises a certain course of action based on the predictive analysis. [42] see also [9, p.3] [10, p.751]

Both of these goals can be achieved by machine learning. Big data analysis is new in the combination of pattern recognition in large datasets and the predictive analysis based on this pattern recognition. Although these are two different analytical processes, in data science and big data analysis they are often combined and happen at the same time. When the feedback is done by the algorithm itself, this is called machine learning. [42] [9, p.3] [10, p.751]

2.2 The debate surrounding data science

Let us turn towards the debate surrounding data science. Over 50 years ago, John Tukey wrote an article called 'The Future of Data Analysis.' In it he called for a reformation of academic statistics. He saw an unrecognised scientific discipline interested in learning from data. [10, p.745]

Continuing in that thought, before the turn of the century the debate among philosophers was about whether data science was to be classified as either a scientific discipline or a form of pseudoscience. The attempts to try to show that data science is a scientific discipline or a form of pseudo-science either 1: make use of the demarcation criteria to try and show that data science does or does not satisfy these criteria, or; 2: demonstrate relevant similarities between data science and paradigmatic sciences, and advocate that these similarities warrant an extension of the general concept. [9] However, these attempts clearly fall short because in current practise data science is not seen as or used as a scientific discipline, but as a way of doing science that is used in many different scientific disciplines, see for example [5] [7] and [40].

2.2.1 Anderson, The end of theory: the data deluge makes the scientific method obsolete.

Since the publication of the influential online article [1], called 'The end of theory: the data deluge makes the scientific method obsolete,' the debate surrounding data science has refocussed on the idea that data science is a new way of doing science (a new methodology, if you will).

In his paper, former editor in chief of WIRED Chris Anderson argues that data science makes the scientific method obsolete. Anderson's argument goes as follows: Science makes use of models, but all models are wrong in some form. And the more we learn about biology or physics for instance, the further we find ourselves from a model that can explain it. However, in today's world of the abundance of data, we don't have to settle for wrong models. Through machine learning algorithms we can make use of the abundance of data existing in our digital age, and 'with enough data, the numbers speak for themselves.' [1]

In the classical scientific enterprise, scientists are trained to distinguish correlations from causations, by which they try to understand the underlying mechanism that connects a correlation. This forming of an underlying mechanism is the formation of a theory or a model, which can be tested through hypotheses. However, Anderson claims that in the new age of big data, correlations are enough. We do not need to know *why*, only *what*. He boldly states that 'correlation supersedes causation, and science can advance even without coherent models, unified theories or really any mechanistic explanation at all.'

Anderson's first claim therefore is that the scientific method has become. Moreover, with enough data, we no longer need to know *why*, only *what*. [1] Science no longer needs to be focussed on providing explanations, only providing knowledge. Theories and models have therefore become obsolete. So, we can state that Anderson's first claim is that the scientific method has become obsolete where it comes to the advancement of scientific knowledge.

This first claim results in another claim. Through data science, we can remove the human aspect of science, namely the formation of theories. Humans are flawed, data is not. Anderson's second claim therefore is that we ought to let the data speak for itself, and data science makes this possible.

2.2.2 Arguments agreeing with Anderson

A number of scientific papers have been published in reaction to Anderson's article, either agreeing [18] [31] or not [5] [7] and [40]. It is for example voiced that 'Big data ushers in a new era of empiricism, wherein the volume of data, accompanied by techniques that can reveal their inherent truth, enables data to speak for themselves free from theory.' [18, p.3] And through data science, we are able

to gain insights 'born from the data' [18, p.2], and let the data speak for itself. Moreover, some argue that 'there is no need for a priori theory, models or hypotheses,' [18, p.4] arguing in favour of Anderson's 'theory free science,' and once again letting the data speak for itself.

However, these authors seem to be more careful choosing their words than Anderson seems to be. To a certain extend they place data science somewhat within the current scientific enterprise, arguing it might be interpreted as a forth paradigm, next to the other three paradigms: 1) Experimental science; 2) Theoretical science, and; 3) Computation science (such as models). [18, p.3] Despite their caution, in the end the authors argue that the new scientific paradigm, characterised through data science and 'empiricism reborn,' replaces the need for theory completely. [18, p.4]

Some argue that 'scientists no longer have to make educated guesses, construct hypotheses and models, and test them with data-based experiments and exampled. Instead, they can mine the complete set of data for patterns that reveal effects, producing scientific conclusions *without* further experimentation.' [31, p.6](emphasis in original) For another example of theory free science, see [36]. All these scientific articles argue in favour of Anderson's claim that data science makes the scientific method obsolete where it comes to the generation of scientific knowledge, and that through data science we allow the data to speak for themselves.

2.2.3 Case study

Let us now take a look at a case study, and see how it relates to Anderson's words. A data science study on DNA expressions to detect cancerous cells can be found in [25]. This data science study in which data on known DNA expressions, of which some are known to have cancer, is used to predict whether 'new' DNA expressions contain cancer. This case study makes use of machine learning algorithms illustrated earlier in this thesis.

In this case study, the data scientists are interested in the DNA expression profile of cancerous tissues. They collect DNA expressions of cancerous cells and of non-cancerous cells. For a given microarray[1], through machine learning the scientists compare the DNA expression profile of a cell of interest, which is cancerous, with that of some reference cells which are non-cancerous.

The procedure allows to see which strands of DNA are activated mostly in cancerous cells and which ones in healthy cells. Regardless of the eventual role of these detected strands in the activity of the cells, the data scientists can regard the respective degree of activation as a significant and characteristic feature of these cells, which distinguishes the cancerous ones from the healthy ones, so [25, p.9]

[1] A microarray is a set of DNA sequences representing the entire set of genes, arranged for use in genetic testing.

states. In other words, the study finds correlations between DNA strands and whether the cells are cancerous or not. Moreover, [25] states that the data scientists are able to classify the cancerous cells and their changes in time after addition of potential drugs on the basis of their DNA expression profile. Therefore, the data scientists can detect which drugs are effective, and extrapolate a common pattern that characterizes the development of (pre-)cancerous cells in combination with certain drugs.

In Anderson's words, the data scientists collect large volumes of data of cancerous and non-cancerous cells, throw this in a computing pile, and extract correlations from them. Thereby predicting whether a cell is cancerous, pre-cancerous or non-cancerous based on the given data. This illustrates that the data scientists attempt to let the data speak for themselves.

The paper concludes by stating that the mathematical study of microarrays is a clear example of prediction and inference from unstructured data that is a trademark of modern data analysis. [25, p.10] In this study, knowing *that* certain DNA expressions characterise (pre-/non-)cancerous cells is enough. The formulation of theory or models to understand the correlations found is surpassed. Instead of taking part in the classical scientific enterprise of forming a theory from which a testable hypothesis springs, data science provides us with immediate correlations and predictions. I believe Anderson would argue that in this example the data scientific method make the classical scientific method obsolete where it comes to providing scientific knowledge.

2.2.4 Scientific arguments disagreeing with Anderson

Next to the similarities between the case study and Anderson's words, there is more evidence that can convince us that the picture Anderson sketches needs to be taken seriously. This evidence comes from scientific literature which takes Anderson's words as a starting point, and argues against his statements. Through numerous arguments philosophers wonder whether the acquisition of large data sets 'mean[s], as popular commentators have argued, the end of theory and the end of the scientific method.' 'Unlikely,' they state. And does 'more data at least mean that we can more easily fathom Nature's mysteries? Not necessarily,' they claim. [7, p.2] Moreover, some are 'sceptical that a purely data driven approach - 'blind big data [in the words of Anderson]' - can deliver the high expectations of some of its most passionate proponents.' [7, p.2]

To support these statements, numerous counterarguments are voiced against Anderson's arguments. It is for example voiced that not all data are reliable. [7, p.4] Through an argument like this, philosophers of science argue against Anderson's ideas that through data science we can let data speak for itself. Or they argue against the idea that data science makes the scientific method obsolete and that

through data science a theory free science is possible.

However, I believe these arguments offer no structural, fundamental challenge to data science. I believe these counterarguments can be overcome by doing *better* data science. The argument that not all data are reliable, and that we therefore cannot let the data speak for itself, can for example be countered by advocating for a better and more careful data scientific practice. If the data is collected more careful, and unreliable data is filtered out, it still can be argued that the reliable data can be left to speak for itself. Arguments like this must be take seriously by data scientists to ensure trustworthy data science. However, because an argument like this can be solved by doing better data science, I believe it does not fundamentally challenge Anderson's ideas.

Philosophers have voiced more concerns surrounding data science. Concerns based on quality, ethics – such as privacy, ownership, informed consent, second use – and concerns surrounding inference and its actuation. [39] Next to ethics, the field of data science and the debate of its philosophy mainly focusses on not producing inaccurate, invalid or misleading results. [15, p.2863] [33, p.79]. It goes without speaking that bad input data results in bad output data, this is not specific to data science, but also holds up in the traditional scientific enterprise. And concerns around data ownership and privacy can obviously be solved. Other concerns voiced are the following:

It is argued that correlations observed in different sets of data are not necessarily evidence of dependency. And that data science is full of spurious correlations. [7, p.4] I think the reaction of Anderson would be that it is well know that data sets are full of spurious correlations, and that care must be taken to extrapolate true correlations. Correlations found in different datasets are indeed not necessarily evidence of dependency, and Anderson does not claim they are. When Anderson says that correlation supersedes causation, it seems obvious to me that he means that true correlations supersede true causations. For these reasons this argument can be countered by doing better data science, for example by reducing the risk of false correlations.

It is likewise argued that 'data scientists [must] respect the sensitivity of complex systems to tiny errors in data and the effects of chaos.' [7, p.4] But this does not pose a fundamental challenge to data science, because this can be solved by executing data science more carefully.

Moreover, it is argued that complex systems are strongly correlated, and hence are much more vulnerable to outliers than classical science is. Therefore variance is claimed to be much higher. [40, p.2] I believe these arguments can be countered in a similar way. However, when these arguments are countered with the argument that Anderson is obviously only talking about true correlations, a problem arises. In philosophy of science it is well known that inductive statements can never

be confirmed, simply because new data may contradict old data, no matter the amount of affirmative data. Having seen a million white swans does not necessitate all swans being white, the next observation may be a black swan, contradiction the statement that all swans are white. I believe that this may pose a tougher challenge to data science, maybe even a fundamental challenge. However, I am not aware of any literature connecting data science to these problems of induction, and forming a critique of data science that springs from inductivist literature. This seems promising subject to further research.

Some argue that we need models and theoretical insights to help guide the collection, curation and interpretation of data. [7, p.4] They thereby try to attack Anderson's statement that the scientific method has become obsolete. It is well known that machine learning algorithms and computer science rest on theory. However, this is clearly not the type of theory-use Anderson aims at. The theory used in collection, curation and interpretation of data and used behind the workings of machine learning is not used to form scientific understanding about the correlations found in data. Anderson is clearly aiming at the latter form of using theory in science, i.e. using theory for the production of knowledge.

Some argue that the results of data science do not readily lead to understanding, thereby arguing that the data cannot be left to speak for itself and that theory is needed to transform the products of data science into understanding. [7, p.4] This argument rests on the further specification that machine learning offers no structural explanations of the correlations they reveal, and many correlations may be false-positives. [7, p.4] Some authors claim this is a weak point of data science. However, Anderson claims that correlations are enough for our scientific practise, and we no longer need to understand 'why', only know 'what.'Anderson claims that we do no longer need to settle for incorrect theories and models, and therefore we no longer need to understand the underlying mechanism or be able to generate scientific understanding. We can use correlations and predictions directly, without a need for an explanation of the underlying mechanism.

It seems that [7] and Anderson talk past each other. Anderson claims we no longer need explanations, while [7] claims we cannot generate explanations without theories. I believe therefore that this argumentation does not fundamentally challenge Anderson's ideas.

It is likewise argued that explanations will never arise from a data scientific enterprise that only makes snapshots and does not make use of theory or models. [7, p.7] Data science does not claim to provide explanations of the correlations found. However, data scientists simply argue that in the new age of data science, explanations are no longer necessary. They thereby talk past Anderson.

Some argue that if data science is to be useful, we need to be able to turn data into true predictions. That is, predictions of events in novel circumstances,

or predictions of events before they occur, not post-hoc explanations. [7, p.5] I would counter this argument as follows: Is this - i.e. providing true predictions - not exactly what data science tries to live up to, i.e. immediately applicative information? If the emphasis of the argument is laid on 'true' predictions, it can be countered by advocating that predictions and correlations found in data science must be carefully tested before being asserted in scientific articles or before acting on it. This argument can therefore be countered by advocating for a more careful data science.

It is argued that in a finite-capacity world, too much data is just as bad as no data. Beyond a certain threshold, further data does not add any information because novel data contains less new information. And if new data that contradicts old data are added, it destroys information in the data set, [40, p.10] argues. They even state that this phenomenon is well known in the science of complex systems, a phenomenon called *non-linear saturation*. I believe this argument might partly be answered by taking care when engaging in data science where systems with strong sensitivity to inaccuracies are used. For example, use multiple different but similar datasets and different machine learning algorithms to find and confirm correlations/predictions found through data science, and turn the correlations/predictions into true predictions. However, the argument that too much data is like no data comes the close to forming a fundamental critique, I believe. The authors argue that 'in a finite world, [in data sets] close to capacity, competitive interactions arise which either annihilate the return on investment (information per data unit) or even make it negative, thereby destroying information and productivity.'[p.10] [40] Thereby arguing against the idea Anderson poses that ever-increasing data sets lead to ever-increasing information.

However, the limitation that new data destroys information seems only to pose a problem to data science beyond a certain data-limit. Therefore I wonder whether this poses a fundamental challenge to data science in its entirety or only poses a challenge to the idea that data science is continually increasing due to the enlargement of data sets and information digitalized. In current data scientific practice, this seems to not yet pose a problem.

I furthermore wonder whether the destruction of information plays a role in current data scientific practise. And what is meant with the destruction of data at all? The authors state that 'Eventually, additional data may even contradict previous data, perhaps because of inaccuracy but more devious scenarios are not hard to imagine, thereby destroying information, because the new and the old data annihilate each other. In the latter scenario, information gain turns into information loss: seeing too much starts to be like not seeing enough.' [40, p.10] If contradictions are due to inaccurate data, this can be solved by improving the quality of the data. If the contradictions are not due to inaccurate data, I

wonder whether the destruction of information a bad thing? If a machine learning algorithm distils the correlation that all swans are white from a data set, and new data is added that makes this statement no longer true, the destruction of information seems to be not only not a problem, but even a necessary process in increasing the accuracy of data science methods. However, I am not familiar with the science of complex systems the authors cite. This argumentation line against data science seems fruitful avenue for future investigation.

2.2.5 History of science solely focussed on data

Anderson makes great claims in favour of data science, most notably the claim that data science lets the data speak for itself. Although data science and Anderson's ideas of data science sound very new, it is not the first time such claims were voiced as an improvement of the scientific enterprise. Let us take a brief look at positivism and its alleged strengths, because the strengths of positivism are in line with the strengths Anderson claims data science possesses.

The history of positivism dates back to as early as the 17th century, and in the 19th century positivism became a leading view of how science ought to be practised. Positivism can be described as a social and intellectual movement that tried to do away with ways of knowing other than sensory. Positivism is focussed on the discovery of laws which facilitate explanation and prediction. These laws of nature are derivable solely through empirical data. [28]

Let us focus on two appealing reasons in favour of positivism. The first has to do with the quantitative approach of positivism. Because the only pathway of knowing is sensory, it is claimed that all data can be objectively measured. Much like in data science it is taken for granted that human intuition may be flawed, but it is held that data measured through scientific apparatus contains an objective truth. Therefore quantitative data is seen as more objective, and was even seen as more 'scientific' than qualitative data or for example theories that sprang from human minds. Positivism therefore argued to let the data speak for itself. [29]

Another advantage of the positivist movement is its well defined structure during studies and discussions. Contrary to for example the current diversity of methods between different scientific disciplines, in the positivist approach all sciences can make use of the same method. Moreover, this method was very clearly defined, which was possible because positivism only made use of empirical data. [29] It is beyond the purpose of this thesis to have a detailed look at this method, but it is interesting to note that in positivism all sciences have a similar method. Because of this benefit, positivism can be somewhat related to data science. In data science, it does not matter the type of data or the scientific discipline in which the study takes place. The data can be thrown into a large computer pile, and correlations or predictions can be extracted.

However, there are also differences between positivism and data science. Most notably a difference that has to do with their goals. Both positivism and data science aim at furthering scientific knowledge and understanding. But a goal of positivism is also to provide scientific explanations. As we have seen, data science as Anderson describes it claims it no longer needs to provide explanations of correlations and predictions found in its studies. Although it surpasses the purpose of this thesis to go into further detail regarding positivism and why it was ultimately discarded, this will be a fruitful avenue for further research.

2.3 Conclusion

As we have seen Anderson makes two big claims about data science. First, that data science makes the classical scientific enterprise obsolete where it comes to generating scientific knowledge and understanding. Second, through data science we are able to let the data speak for itself. The first part of this thesis was aimed at providing a foundation for these two claims and thereby arguing that Anderson's ideas of data science are accurate. The second part of this thesis will be aimed at providing fundamental challenges to data science.

3 Part 2: Critique of data science

The aim of this part is to open up a debate of data science. First, an existing critique of data science and its claim that we ought to let data science speak will be investigated. This critique makes use of the idea that data is inherently theory laden. Next, I will voice my own critique of data science, a critique against the idea that data science can replace the classical scientific method where it comes to the generation of scientific knowledge or understanding.

3.1 Fundamental critique based on feminist standpoint theory

Let us take a look at the argument against data science based on the theory-ladenness of observations and data. McQuillan first exposes a preconception of data scientists. Next, he illustrates that this preconception leads to dangers. Finally, he voices his critique of data science, attacking data science and therefore trying to prevent its dangers.

3.1.1 Neoplatonism in data science

In his work *Data Science as Machinic Neoplatonism*, [22], Dan McQuillan exposes a picture of data scientists as scientific realists.[2] McQuillan claims that data scientists hold on to a specific realist believe called neoplatonism.

Neoplatonism is the idea that there is an ideal, mathematical, world hidden behind the world we observe through our senses. They hold that the world we experience is an imperfect imprint of this perfect, inaccessible and ontologically superior world. McQuillan claims that '[a]s a method for revealing a hidden mathematical order in the world, data science strongly echoes this neoplatonic project. For the data scientist, computation plays the role of the intermediary between the imperfect world of data and the pure function that relates the features to the target.' [22, p.261]

There are other sounds relating to realism in data science. Some authors for example cite the Data Science Association's "professional code of conduct," which states that data scientists '[use] the scientific method to *liberate* and *create* meaning from raw data.' [9, p.4] The use of the word 'liberate' seems to fall in line with the idea McQuillan illustrated above, because when one liberates something, the existence of that something is independent of the liberation. The term 'create' seems to be the opposite of McQuillan's idea, because when creating something, the thing created does not exist independently of the creator. The use of the word *create* implies that these products are made by human practises and do not exist independently from these practises. Similar sounds, which are not in line with McQuillan's, can be found in philosophical works on data science, such as [20] or [30]. Because these arguments are voiced by philosophers, and not data scientists themselves, this does not necessarily challenge McQuillan's idea that data scientists themselves hold on to a neoplatonic view of their practise.

Foundation for the argument that data scientists are neoplatonists is lacking in McQuillan's work. Let me therefore try to give it here. I believe a base for this claim can spring from the direct ideas of data scientists. However, these ideas are often left implicit in data science studies. I therefore suspect that a compelling argument in favour of McQuillan's claim might come from viewing data scientists' vocabulary.

In an introductory article of geographic data science studies, Andrienko, a prominent data scientists, states: 'The massive volumes of data contain complex, yet implicit spatial, temporal and semantic interrelations that are waiting to be *uncovered* and made explicit' [2, p.16] (my emphasis). Gennady Andrienko is not just *a* data scientist, he 'is a lead scientist at the Fraunhofer Institute for Intelligent

[2] scientific realism can be characterised by the statement that science aims to give us a literally true story of what the world is like; and acceptance of a scientific theory involves the belief that it is true.

Analysis and Information Systems (IAIS) and professor at City University London. He co-authored monographs 'Exploratory Analysis of Spatial and Temporal Data' (Springer, 2006) and "Visual Analytics of Movement" (Springer, 2013) and 100+ peer-reviewed journal papers.'[3] Therefore, his words carry weight, and the choice to use *uncover* interrelations, instead of, for example, *discover* or even *make*, can be assumed to be not an arbitrary one. Through these words we can argue that at least some data scientists hold on to a neoplatonic view of their practice.

3.1.2 Machinic neoplatonism is problematic

McQuillan argues neoplatonism in data science leads to problems. First, McQuillan claims that the neoplatonic nature of data science makes it problematic because it makes data science hard to constrain. The neoplatonic nature creates structural conditions for specific injustices caused by bad data or false positives, he argues. Moreover, McQuillan claims that correlations found in data are given more weight than testimonies of subjects, which might lead to a situation in which subjects are not able to contest data science because they lack the capacity to express their knowledge in the same way. He states that 'The new paradigm redefines 'the facts on the ground', because, as both Kuhn and Feyerabend pointed out, the very idea of what constitutes facts can change with a shift in the overall pattern of thought (Kuhn 1996).' Against this superior insight 'traditional safeguards and civic protections become ineffective, because the ground they stand on is modified by a new neoplatonism.' [22, p.]

Second, McQuillan claims that if the judgements of machine learning models and data science remain opaque to us, data scientists and others who are faced with the products of data science are released from 'categories of intent or accountability.' This release of accountability could very well result in a lack of concern about the effects of execution of the proposed scheme by an 'abstract authority.' [22, p.263] This line of thinking originates from Hannah Arendt's work on thoughtlessness.

Thoughtlessness is used to explain and comprehend the ability of bureaucrats in Nazi Germany to perform their actions and hide their responsibility behind the bureaucratic process. Because actions were mandated by sources up the chain of command, a Nazi like Eichmann was able to perform his actions thoughtlessly, and hide his responsibility in the source of the command. When we are unable to understand the reason why data science advises a certain course of action, a similar danger hides in data science, McQuillan claims.[4]

[3] see the site of the city university of London

[4] McQuillan does not explain why data science might remain opaque to use. In many machine learning algorithms, the internal workings of a machine learning algorithm, for example the reason why two characteristics are signalled to be correlated, are not known and cannot be known to

But classical science might just as well be seen as superior insight, opaque to many of us. To differentiate the danger of data science from this, McQuillan states that data science differs from science: '[data science] is an apparatus that not only makes possible a certain way of knowing but also acts directly on the knowledge produced. In that sense, it is very different to science, which seeks to distance itself from implementation in order to retain the veil of neutrality.' [22, p.262] The direct application of data scientific knowledge, which McQuillan calls the machinic aspect of data science, can result in an apparent indifference to the consequences of actions mandated by the data science study. When this is combined with the fact that data science produces superior insight, this could lead to an apparent indifference to the consequences of actions mandated by the data science study, McQuillan argues.

One might wonder to what extend the claim that people suffer injustices through data science is actually an argument against data *science*, and not just an argument against big data. As an argument against big data used by governments or companies, I believe this to be a valid argument. However, I wonder to what extend this argument measures up against data science in the fields of geology, marine biology, astronomy or physics for example. But in the specific case study considered earlier in this thesis one can consider how, for example, the results of the study compared between different geological locations or ethnic groups might lead to the injustices McQuillan mentions. Moreover, the overlap between big data and data *science* mentioned earlier in this thesis could also be used as an argument in favour of McQuillan's claim.

3.1.3 Feminist standpoint theory

McQuillan tries to shield us from this machinic neoplatonism through historical critiques of science, more specifically feminist standpoint theory. Although the scientific method is fit for removing personal bias and bad science, some of the sexist and racist prejudices stem from inadequacies in the way scientific methods and norms are conceptualised, standpoint theorists hold. This is the case because the data generated through the scientific method are shaped by the thinking of dominant groups of society and how they think about the natural world, social relations and the way these dictate how society at large understands the world. So, because 'prevailing standards for objectivity are too weak to identify culture-wide assumptions that shape selection of specific scientific procedures as good ones

any human involved, not even the programmer of the algorithm. Anderson characterises this as not problematic, because in the new age of petabyte data knowing *what* is enough, knowing *why* is not necessary any more. However, here McQuillan argues that the opaque nature poses a problem to data science. For more information on opaqueness of machine learning, or as it if often called, black box machine learning, see for example [9, p.13].

in the first place,' data and observations generated through the scientific method are inherently theory laden. [22, p.264] The theory-ladenness of data is a much discussed topic in philosophy of science, and it is well established and accepted in philosophy of science. [20, p.135-136]

The critique McQuillan has on data science is a fundamental critique. Because there is always a perspective through which data are gathered. *All* data are fundamentally biased. Because *all* data is biased, data cannot be let to speak for itself and these data cannot uncover an objective hidden truth behind the world. Using *better* data and being more careful when asserting the products of data science can never overcome the problem that *all* data are theory laden. [22, p.264]

Data scientists might argue that because data science does not make use of theories, it is less susceptible to human biases. However, the data generated and used in data science does not escape these inherent biases. Even when human influence is minimized, feminist standpoint theorists idea that data can never be collected bias free cannot be avoided. Even when data is collected automatically through digital means for example, the digital means cannot escape human and societal influences, and thus the data is framed through human notions and therefore inherently laden with bias.

One might not agree about the dangers McQuillan voiced regarding data science and its possible consequence of thoughtlessness and the danger that data science could and will be seen as superior yet opaque insight. However, feminist standpoint theory and the theory-ladenness of observation are well established and well accepted in the scientific culture. Therefore I believe the extrapolation of feminist standpoint theory to data science must be regarded as a serious critique of data science.

3.1.4 More theory ladenness

McQuillan is not the only one who connects the idea of theory-ladenness in philosophy of science with data science, see for example. [20] and [30]. The first of these takes a deep dive into philosophy of science to debunk the idea of data as an 'indicate basic, incontrovertible facts on a given entity or process,' very briefly connecting this to big data.

The second of these takes a more rigorous approach. This paper called *Aspects of theory-ladenness in data-intensive science* 'takes a detailed look at two algorithms that are widely employed [in data science].' Through theory-ladenness of observationand the argument that the methods of data science are also constructed using theories, it argues that data science is not theory free. Although it correctly argues that data science is theory laden, I have my questions as to whether his argument truly counters Anderson's statement that data science makes the classical scientific method obsolete where it comes to the production of scientific knowl-

edge. When Anderson claims that data science is a rise of theory free science, I believe he argues that science can take place without forming new theories to *explain* correlations found. Anderson claims that correlations are enough, we do no longer need to know *why*, knowing *what* suffices. In that sense, in the age of big data, the formation of theories is no longer necessary where it comes to generating scientific knowledge. Therefore I believe the argument that data and the methods of data science are inherently theory-laden does not accurately attack the claim that because of data science, we do not need theory to explain science any more; and that in the era of big data, knowing *what*, not *why*, is enough. The argument therefore talks past Anderson's arguments.

In the spirit of this argument, one could likewise argue that machine learning algorithms are based on computer science theory, and that data science is therefore not theory free science. This argument talks past Anderson's argument in much the same way.

3.2 My fundamental critique

I wish to add a fundamental critique. A critique based on the method of data science, and a critique of data science's presumption of making the classical scientific method obsolete where it comes to the generation of scientific knowledge. In short, the argument is based on the following: The scientific method is used for a reason, it has great advantages and secure checks and balances. Data science claims to have made the scientific method obsolete where it comes to the generation of scientific knowledge. Thereby possibly surpassing these benefits of the classical scientific method.

Two specific aspects of the classical scientific method will be investigated. First we will take a look at negative results. Second we will take a look at reproducibility.

3.2.1 Negative results

In the classical scientific enterprises, hypotheses spring from theories. Hypotheses can investigate whether a theory can be corroborated or refuted. Thus, hypotheses can be tested by experiments or observations, and either be true or not. If a hypothesis is confirmed it is called a positive result, and the hypothesis can confirm or corroborate the scientific theory from which it sprang. If the hypothesis is found false it is called a negative result, and it may refute a theory. [12, p.892]

Even though they are called negative results, they provide positive benefits and are crucial to scientific progress. They form the collective self-correcting process. [12, p.892] The absence of negative results can 'cause a waste of resources replicating research that has already failed,' wasting time, energy and money. [12, p.892] see also [21, p.172] Moreover, when negative results are found and reported

in scientific research and literature, they can 'serve as a warning to researchers that a particular area or approach is unfruitful.' [38, p.229], see also [38, p.228-229] [11, p.871] [16, p.700] [21, p.172]

A result of this is that negative results are able to delimit the search space. Negative results are able to delimit the search space because they relate back to a hypothesis and a theory. Therefore, when a hypothesis is found to be false, not only the hypothesis itself can be refuted, but the theory from which the hypothesis sprang can likewise be refuted. [38, p.228-229] [11, 871] [16, p.700] [21, p.172]

Negative results help the classical scientific enterprise progresses over time. Because of negative results, over time the search space of a scientific discipline will be delimited more and more. It works as follows: when a hypothesis is refuted, it refutes the theory from which it sprang. When in search of a new theory, the search space from which a new theory is picked is smaller than it once was, because it does not contain the refuted theorie(s) anymore. Therefore classical science and scientific knowledge progresses over time through negative results, because knowing where an explanation or prediction is not to be found is knowledge in itself.[5]

A single negative result can thereby further scientific understanding beyond the refutation of the specific hypothesis refuted. For this reason, negative results form an important tool to further scientific progress. [38, 228-229] [11, 871] [16, p.700] [21, p.172]

Data science produces correlations and predictions. These may turn out to be correct or incorrect. If correct, it gives us a hunch about a relation (maybe causal, but at least correlational) in the real world. The correctness of the prediction does of course not necessitate the truth of the data science study, it could be correct by coincidence. But the same holds for the classical scientific enterprise. If the prediction turns out to be false, it states that a specific correlation or prediction found is not true. This might be called a negative result.

A negative result in data science does not have the same consequence as negative results in the classical scientific enterprise. Because data science does not make use of theory, a negative result can not refute a theory. Therefore a negative result in data science does not further scientific understanding by refuting theories in the same way a negative results in classical science does. But does the negative result in data science not result in the refutation of anything apart of the specific prediction or correlation under consideration?

As we discussed, in classical science, a negative result refutes the theory from which is sprang. I do not believe a refutation of a prediction in data science results

[5]I acknowledge that there is no consensus in philosophy of science over whether scientific progress is even possible. I have taken a practical stance in this debate, i.e. the stance that scientific progress does exist. However, further research might be done to investigate data science in the light of scientific progress, it goes beyond the scope of this thesis to investigate the relation between those ideas and data science.

in the refutation of the machine learning algorithm from which is sprang. Moreover I do not believe that data scientists themselves think that it does. Anderson claims that 'with enough data, anything can be predicted,' implying that if a prediction turns out to be false, or when a false correlation is found, more, or more accurate data is needed. Therefore a negative result in data science does not have the same consequence as a negative result in the classical scientific enterprise. In classical science, negative results refute something bigger than just the hypothesis itself. In data science, I do not believe something bigger than the specific prediction in question can be refuted.

Data science is unable to delimit the search space in the same way. Therefore future data scientific practice will make use of the be exactly the same search space as current data scientific practice. In an article from 1964, called 'the importance of negative results in psychological research,' R. G. Smart uses an analogy to explain the importance of negative results. He states that '[t]he finding that a particular mountain contains no gold fails to move prospectors and spectators. (...) In scientific undertakings, however, the failure to find the gold of positive results has important implications.' [38, p.228] If no gold is found - if hypotheses turn out false - science delimits the search space. However, in this analogy, data scientists are the prospectors and spectators that do not move on to new mountains. They search everywhere for the occasional lucky discovery of a gold nugget, unable to delimit the amount of land in which they search. Where science over time removes numerous mountains from their search space through falsifying theories through hypotheses, data science is unable to delimit its search space. In the far future, data science will be searching the same search space in search of lucky nuggets.

So through negative results, classical science makes a kind of progress that data science by itself is unable to make, namely the progress of delimiting the search space. And through negative results classical science produces scientific knowledge that data science by itself is unable to make, namely the knowledge of where a true theory of a certain scientific discipline can not be found. It can therefore be argued that data science does not make the scientific method obsolete where it comes to scientific progress and the creation of scientific knowledge, because the scientific method contains a progress data science is fundamentally unable to make. This is truly a fundamental challenge to Anderson's picture of data science, because it cannot be overcome by collecting better data or being more careful when asserting correlations or predictions.

A counterargument might be that data science does not make use of the same concept of 'search space' as the classical scientific enterprise does. The reason for this might be that data science does not make use of theories, and that therefore it omits the search space of theories from which a new theory springs in the classical scientific enterprise. Eventhough this might be true, and might be an interesting

argument for future papers, the fact that knowing where answers cannot be found remains a type of scientific knowledge which data science cannot produce.

But what if data scientists counter with the statement that this kind of knowledge is no longer valuable in data science? What if, because data science omits the whole explanatory side of data science, it claims that we can truly rest our scientific understanding on correlations and nothing more than that? If data scientists make these claims, my argument seems to talk past data scientists in much the same way as the arguments discussed in the first part of this thesis. I have difficulty to answer these questions with anything more than the arguments provided above - that negative results contain many positive benefits, and most importantly, that they provide a kind of knowledge and understanding of the world data science can not provide.

3.2.2 Reproducibility

The goal of this section is to raise questions surrounding trustworthiness of correlations found in data science. There is already some discussion surrounding the accuracy of data science and concerns surrounding false correlations. It is for example voiced that: through the use of data scientific methods, data science finds many false correlations because complex systems (like big data sets) are by definition strongly correlated; The strong sensitivity of data to inaccuracies, even in big data, leads to false correlations; An abundance of data leads to many false correlations, telling us nothing in the end, and; Not all data are as reliable. Other aspects mentioned are that many or all datasets are biased, that sensory perception as grounds for knowledge are unreliable, and critics point out that neural nets can throw up false correlations, especially if the datasets they are trained on are small data sets. See for example [40] [20] [7] [15] [41] [5] [39] [23] [37] [4].

These concerns must be taken serious, but, as illustrated in the first part of this thesis, they offer no fundamental challenge to data science. I believe that there are fundamental challenges to data science. One of these reasons was voiced in the previous section, another will be voiced next.

In the classical scientific method, experiments are performed and produce results. To ensure the accuracy of these results, the experiment and results ought to be reproducible. This means that when different scientists perform the same experiment on a similar group of test subject or using similar data, similar results are found. Even though there is talk of a reproducibility crisis in some sciences, in theory results and experiments ought to be reproducible.

Reproducibility is a control mechanism of science. If the original experiment and its results are not reproducible, the outcome of the original scientific study is challenged. The outcome of the original experiment can be concluded as being an accidental correlation, or that execution of the original experiment was flawed. As

a consequence, the truth of the original results can be challenged, and the theory from which the hypothesis sprang might be challenged.

Data science that makes use of certain machine learning algorithms is not reproducible. (The machine learning algorithms in question are deep learning methods.[6]) Because of the use of randomness in these machine learning algorithms, even when the same machine learning algorithm is used on the same set of training data, results may differ; The same exact procedures are by no means bound to have the same results. [4, p.305] Let alone if a different machine learning algorithm is used or a different but very similar set of training data. Moreover, 'many of the other innumerable, and often silent, parameters that control modern deep learning methods plausibly impart similar influence on the final performance, further complicating reproducibility.' [4, p.305] This is a well known fact in computer engineering, but little discussed in data science.

Does this pose a fundamental challenge to data science? I believe deep learning forms a fundamental and unavoidable part of data science. Numerous studies show the abundant presence of deep learning algorithms in data science. Some illustrate the presence of deep learning models and hybrid deep learning models in data science in economics. [27] Others state that deep learning is a 'high focus of data science.' [24, p.1] Moreover, some note the extensive application of deep learning in various fields of science, and state that 'a key benefit of Deep Learning is the analysis and learning of massive amounts of unsupervised data, making it a valuable tool for Big Data Analytics where raw data is largely unlabeled and un-categorized.' [24, p.1] also see [17, p.275] Other studies illustrate that deep learning has a big presence in data science, illustrating the use of data science in Lung Cancer detection, illustrating the presence of deep learning in fluid dynamics and showing the presence of deep learning in fluid dynamics. [19] [3] But further and more technical research ought to be done to determine whether data science can function without irreproducible machine learning algorithms, and how big the presence of randomness in machine learning algorithms really is.

In literature about machine learning there are some solutions offered for the problem of irreproducible machine learning algorithms. One of these is using seeds for random numbers. The idea is that you use the same random numbers every time you run the machine learning algorithm, thereby producing the same result. On first glance the machine learning algorithm is indeed made reproducible, but this turns out to be an illusion. When using the same seed[7] in the learning of an algorithms, you produce the same result. But what differentiates the results found in the algorithm when using one seed and not another? Nothing, the choice of seed is arbitrary. What then differentiates the correlations found with the seed

[6] A form of machine learning often used in data science.

[7] a way to denote that you use the same random numbers every time you run the algorithm

used from the correlations that would have been found from the use of another seed? Nothing. The solution of using seeds therefore does not offer a solution to the irreproducibility of machine learning algorithms.

Even though the machine learning community 'has embraced fairly radical notions of open science, transparency, and reproducibility, [and] many reports are first available as reprints, code is usually available as open source, and most articles rely on data sets available in the public domain,' reproducibility remains a fundamental challenge to many machine learning algorithms. And although 'medical researchers using machine learning would be well served by adopting some of these practices, including open sharing of data, code, and results whenever possible,' this does not seem to solve the problem. [4, p.306]

Can the irreproducibility of machine learning algorithms not be solved by doing better and more accurate data science? And can it not be solved by being more careful with asserting the correlations and predictions found? We could indeed double and triple check the correlations and predictions found through data science, and filter out the false ones. However, if human hands and minds need to differentiate true from false correlations, are we still letting the data speak for itself? Moreover, if predictions and correlations need to be tested (through experiments for example) to assert their truth, is the scientific method made obsolete where it comes to the production of scientific knowledge? The answer might be somewhere in the middle. Further research is needed to determine whether the irreproducibility of machine learning algorithms does fundamentally challenge (all) data scientific studies or not.

4 Conclusion

Data science as Anderson describes it faces us with two propositions. First, through data science we are able to let the data speak for itself. Second, data science makes the scientific method obsolete where it comes to the production of scientific knowledge.

The second part of this thesis was aimed at attacking these two propositions. First, McQuillan's argument, which attacks the idea that we ought to let the data speak for itself, was investigated. His argument rests on established feminist standpoint theory and therefore needs to be taken seriously. Next, my original arguments were voiced against the idea that data science makes the scientific method obsolete where it comes to the production of scientific knowledge.

My critique is twofold. The first argument is based on the idea that data science is unable to benefit from negative results in the same way the classical scientific method does. The implication of this is that data science lacks a kind of structural improvement in scientific knowledge that is present in the classical

scientific method. The second argument is concerned with false correlations. Combined with the fact that data science produces many false correlations, the claim that fundamental methods of data science are not reproducible poses a challenge to confirming correlations found through these machine learning algorithms.

5 Discussion and further research

This thesis begins with an analysis of the data scientific practice. Due to the limited resources of my research, it is not possible to discuss the data scientific practice and its methods exhaustively. Therefore I have used the words of Anderson to characterise data science, and given power to his words through a case study and scientific response to his article. Some might not agree with Anderson's characterisation of data science, and might not be convinced by the arguments I used to support his analysis of data science. I have investigated many data science studies which convinced me of the correctness of Anderson's claims. And in this thesis I have done several things to argue that it is a correct description. However, a broader analysis ought to be done to determine whether Anderson's characterisation is indeed an accurate description of the whole field of data science.

What I can say with certainty is that there are numerous data science studies that are accurately described by Anderson's article, and therefore these studies are fundamentally challenged by the critique voiced in the second part of this thesis. However, it might be the case that the discussion is further than I projected it to be. Data scientists might be more careful in the execution of their practice and be more aware of fundamental flaws than I described. Although I have not encountered either of these in data scientific literature, let us hope this is indeed the case.

McQuillan describes multiple dangers of data science. I have already noted that not everyone might be convinced by the argumentation about the dangers according to McQuillan. However, the argumentation of the theory-ladenness of data is based on established standpoint feminist theory and is able to exist independently of the risks McQuillan voices. I believe McQuillan's argument must therefore be take seriously.

The arguments voiced in the section 'My fundamental critique' are novel arguments and are my original work. The concern about reproducibility in data science based on the use of certain machine learning algorithms is a new argument in the data science discussion. However, this is a well known phenomena in the computer scientific disciplines of Artificial Intelligence and machine learning. To further support the argument that data science is not reproducible, further and more exhaustive research is necessary about the methods of data science.

The argument that negative results in data science lack the benefits they pro-

vide in the classical scientific enterprise is an entirely new argument. It is an argument I have not encountered anywhere, neither in any literature regarding data science nor any literature regarding machine learning. The novelty of this argumentation makes it an important argument, but at the same time makes the foundation of the argumentation not as strong as it could be. At the time of writing this thesis, I am convinced my argument offers a fundamental challenge to the way data science is currently being practised. But there must be arguments or aspects to the debate about negative results that I have missed. Due to the fact that this debate has not yet been had, I have not been faced with any arguments in the defence of data science. I hope this thesis can provoke a response, is thought provocative and adds to the beginning of a debate about the fundamental flaws of the current practice of data science.

This thesis ends with a critique. I challenge the current practice of data science, but I refrain from offering solutions. Ideally I would have offered some revisions to data science, but sadly this surpasses the breath of this thesis and must be subject to further research. I believe the critique of the method of data science and the accompanying conclusion that the scientific method is not made obsolete by data science offers a fruitful start to such a creolisation of data science. Namely that the scientific method is necessary and crucial in structurally furthering of our understanding of the world, and that data science as it is used currently is unable to utilise certain benefits of the scientific method. A possible creolisation therefore could be to use data science as a method in the classical scientific enterprise, for example as a way of forming hypotheses and utilising data science in the 'context of discovery,' and not in the 'context of justification.' See for example [5] for the idea that data science ought to be a fourth paradigm in the classical scientific enterprise.

The debate about false correlations in data science raises challenges related to induction. It for example raises the question whether correlations found in data science can be confirmed at all. Some characterise data science as a new form of radical empiricism, which again raises questions surrounding induction and its confirmation. Induction is a much debated topic in philosophy of science, and the topic is far too broad to say anything about it in the a single thesis. Therefore I have not discussed induction and its relation to data science in this thesis. This is however fertile soil for further research. For more information about data science and induction, see for example [18], which characterises data science as radical empiricism, and [9].

This thesis has attempted to voice fundamental challenges to data science. Whether you agree with me or not about my challenges to data science or the way I analysed data science in the first part, the most important goal of this thesis is to start the discussion. I believe data science originated without its method being

thoroughly investigated or discussed, and I believe there has not been sufficient discussion about the data scientific practice and its fundamental challenges.

Where the strong aspects of data science are obvious and used abundantly, its fundamental weak points are scarcely discussed and ill defined. If the strong points are so visual, but the weak points almost invisible, science runs the risk of over-hyping the data scientific practice. Scientists might believe data science contains only benefits, and a consequence scientists might believe data science really could replace the scientific method, without facing risks or lacking aspects of our classical generation of scientific knowledge. I therefore hope that this thesis adds to the beginning of such a discussion.

6 Acknowledgement

I would like to thank prof. dr. ir. Jan Broersen and dr. Brandt van der Gaast for being my thesis supervisors. I have enjoyed their critique and recommendations greatly.

References

[1] Anderson, Chris. The end of theory: the data deluge makes the scientific method obsolete. *WIRED*. 2008.

[2] Andrienko, Gennandy & Andrienko, Natalia. Geographic data science. *IEEE Computer Graphics and Applications*. 2017 Vol. 37. 15-17.

[3] Baldi, Pierre. Deep learning in biomedical data science. *Annual Review of Biomedical Data Science*. 2018 Vol. 1. 181-205.

[4] Beam, Andrew L.; Manrai, Arjun K.; Ghassemi, Marzyeh. Challenges to the reproducibility of machine learning models in health care. *American Medical Association*. 2020 Vol. 323 No. 4. 305-306.

[5] Bell, Gordon; Hey, Tony; Szalay, Alex. Beyond the data deluge. *Computer Science*. 2009 Vol. 323. 1287-1298.

[6] Blei, David M.; Smyth, Padhraic. Science and data science *Proceedings of the National Academy of Sciences*. 2017 Vol. 114 No. 33. 8689-8690.

[7] Covevney, Peter V.; Dougherty Edward R.; Highfield Roger R. Big data need big theory too. 2016.

[8] Davies, Nigel & Clinch Sarah. Pervasive data science. *IEEE CS,*. 2017 Vol. 16(01). 50-58.

[9] Desai, Jules; Watson, David; Wang, Vincent; Taddeo, Marianrosaria; Floridi, Luciano. The epistemological foundations of data science: a critical analysis. 2022. Available at SSRN: https://ssrn.com/abstract=4008316 or http://dx.doi.org/10.2139/ssrn.4008316.

[10] Donoho, David. 50 years of data science. *Journal of Computational and Graphical Statistics*. 2017 Vol. 26 Iss. 4. 745-766.

[11] Easterbrook, Philippa J.; Berlin, Jesse A.; Gopalan, Ramana; Matthews, David R. Publication bias in clinical research. *The Lancet*. 1991 Vol. 337 No. 8746. 867-872.

[12] Fanelli, Daniele. Negative results are disappearing from most disciplines and countries. *Scientometrics*. 2012 Vol. 90.

[13] Fayyad, Usama; Piatstsky-Shapiro, Gregory; Smyth, Padhraic. From data mining to knowledge discovery in databases *AI Magazine*. 1996 Vol. 17 No. 3.

[14] Han, Jiawei; Kamber, Micheline; Pei, Jian. Data mining: Concepts and techniques *Elsevier*. 2011.

[15] Haug, Frank S. Bad big data science. *IEEE International Conference on Big Data (Big Data)*. 2016. 2863-2871.

[16] Ioannisid, John P.A. Why most published research findings are false. *PLoS Medicine*. 2005 Vol. 2 Iss. 8 e.124. 700-701.

[17] Jan, Bilal; et al. Deep learning in big data analytics: A comparative study. *Computers & Electrical Engineering*. 2019 Vol. 75. 275-287.

[18] Kitchin, Rob. Big data, new epistemologies and paradigm shifts, *Big Data & Society*. 2014 Vol. 1 No. 1. 1-12.

[19] Kuan, Kingsley; et al. Deep learning for lung cancer detection: Tackling the kaggle data science bowl 2017 challenge. *Computer Vision and Pattern Recognition*. 2017.

[20] Leonelli, Sabina;ed: Floridi, Luciano. *The philosophy of data, in:The Routledge Handbook of Philosophy of Information*. 2016. Oxford University Press.

[21] Matosin, Natalie; Frank, Elisabeth; Engel, Martin; Lum, Jeremy S.; Newell, Kelly A. Negativity towards negative results: a discussion of the disconnect between scientific worth and scientific culture. *Disease Models & Mechanisms.* 2014 Vol. 7. 171-173.

[22] McQuillian, Dan. Data science as machinic neoplatonism. *Philosophy & Technology.* 2018 Vol. 31. 1-20.

[23] Mosco, Vincent. *To the Cloud: Big Data in a Turbulent World.* 2014. Routledge.

[24] Najafadabi, Maryam M; et al. Deep learning applications and challenges in big data analytics. *Journal of Big Data.* 2015 Vol. 2 No. 1.

[25] Napoletani, D.; Panza, M.; Struppa, D.C. Agnostic science. towards a philosophy of data analysis. *Found Sci.* 2011 Vol. 16. 1-20.

[26] Newcomb, Simon. Evolution of the scientific investigator *Smithsonian Institution, Annual report.* 1904. 221-233.

[27] Nosratabadi, Saeed; et al. Data science in economics: Comprehensive revies of advanced machine learning and deep learning method. *Mathematics.* 2020 Vol. 8 No. 10.

[28] Park, Yoon; Konge, Lars; Artino; Anthony. The positivism paradigm of research. *Acad Med.* 2020 Vol. 5 Nr. 95.

[29] Pham; Lan Thi Mai. Qualitative approach to research: A review of advantages and disadvantages of three paradigms: positivism, interpretivism and critical inquiry. 2018.

[30] Pietsch. Wolfgang. Aspects of theory-ladenness in data-intensive science. *Philosophy of Science.* 2015 Vol. 82 No. 5. 905-916.

[31] Prensky, Marc. H. sapiens digital: From digital immigrants and digital natives to digital wisdom. *Innovate: Journal of Online Education.* 2017 Vol. 5 Iss. 3. 1-9.

[32] Rathenau Instituut. De datagedreven samenleving. 2016. Den Haag.

[33] Richterich, Annika. *The Big Data Agenda: Data Ethics and Critical Data Studies. in:The Big Data Agenda,.* (2018, University of Westminster Press). 33-51.

[34] Rusitschka, Sebnem; Ramirez, Alejandro. Big data technologies and infrastructures. *EU BYTE*. 2014.

[35] Russom, Philip. Big data analytics. *TDWI Beste Practices Report*. fourth quarter 2011.

[36] Schmidt, Michael; Lipson, Hod. Distilling free-form natural laws from experimental data. *Science*. 2009 Vol. 324. 81-85.

[37] Schwab, Matthias; Karrenbach, Martin; Claerbout, Jon. Making scientific computations reproducible. *Computing in Science and Engineering*. 2000 Vol. 2 No. 6. 61-67.

[38] Smart, Reginald G. The importance of negative results in psycholohical research. *The Canadian Psychologist*. 1964 Vol. 5 No.4. 225-232.

[39] Spinney, Laura. Are we witnessing the dawn of post-theory science? *The Guardian*. 2022.

[40] Succi, Sauro & Covevney, Peter V.;. Big data: the end of the scientific method? *CORR*,. 2018 (1807.09515).

[41] Sullivan, John L. Review of to the cloud: Big data in a turbulent world. *International Journal of Communication*, 2014 Vol. 8. 2343-2347.

[42] Wetenschappelijke Raad voor het Regeringsbeleid. Big data in een vrije en veilige samenleving. 2016. Den Haag.