2018 International Conference on Multidisciplinary Research
Track 1: STEM (Science, Technology, Engineering & Mathematics)
ISBN: 978-99949-0-384-9

Challenges of Big Data from a Philosophical Perspective

Sunil Choenni
Ministry of Justice and Security, WODC
The Hague, The Netherlands

Niels Netten
Rotterdam University of Applied Sciences, Creating 010
Rotterdam, The Netherlands

Mortaza Bargh
Rotterdam University of Applied Sciences, Creating 010
Rotterdam, The Netherlands

Rochelle Choenni
University of Amsterdam
Amsterdam, The Netherlands

Abstract

Due to the many potential applications of Big Data, the expectations are high. However, there are some fundamental objections on the straightforward use of Big Data outcomes. In this paper, we take a philosophical view on the Big Data approach and discuss these objections. Formally, Big Data induces models from very large data sets, which are nevertheless incomplete. In many cases these data sets might be skewed as well. This gives rise to the question to what extent induced models represent the real world adequately, and therefore are sufficiently grounded to base new policies on. We argue that caution is needed in interpreting these models and well thought through strategies are required for using the models in practice in a responsible way. We discuss two strategies that may be used.

Keywords: Big Data, induction, models, incompleteness.

Introduction

Nowadays, our environment has become more complex as it is equipped with many devices, such as camera's and mobile phones that together generate huge amounts of data. These data are of several types. To exploit this data, traditional database techniques are not advanced enough, giving rise to the concept of Big Data. Big Data refers to both structured and unstructured data sets that are so large in volume and complex that traditional systems are not capable of managing, storing and/or processing them within a reasonable time frame. These data sets have proven to encapsulate invaluable and sometimes unexpected information. The data sets have traditionally been used to find some useful insights to enable, for example, cost and time reductions. There are various organizations and companies across a wide range of public and private sectors that are now trying to process raw data in order to derive information useful for specific applications such as public administration, health care, insurance, finance, transportation, logistics and retail (Roger et al., 2012). There are multiple factors that help for effectively retrieving valuable information. Not only does the amount of data available play an important role, the way in which the data is processed has equally important role. Therefore, new intelligent techniques are needed to optimally process the data (Kim et al., 2014).

This need has led to increasing amount of research among data scientists and analysts, regarding questions on how to process and manage the data. However, how to interpret and implement the results in practice is in its childhood. In this paper, we discuss the challenges and pitfalls of implementing the results obtained from data analytics tools.

Results from Big Data are often used on a large scale to make predictions about a wide range of matters (Tien, 2013; Choenni, 2000; Netten et al., 2018). In business settings, for example, predictions can be made about the most optimal business strategy to follow. As the predictions are fully based on the input data, the accuracy of these predictions depends on, among others, the suitability of the data. When for example government institutions use these methods to make predictions about certain future behaviors and potentially impose new regulations based on them, it becomes a necessity to be able to guarantee reasonably accurate results. This entails that the data sets provided to the algorithm are a sufficient reflection of the real world (Choenni et al., 2018).

This gives rise to the question of to what extent Big Data and the models derived from the data are able to represent the real world. More importantly, it gives rise to the question as to whether the use of this approach to gain information on a large scale of complex (social, business, economic, etc.) phenomena is justified. Does this approach provide results that are sufficiently grounded to base new policies on? This paper describes various challenges raised by this Big Data analytics approach. Both theoretical and more practical aspects will be discussed. Firstly, a more thorough explanation will be given on the way in which Big Data models are used. Secondly, the reasoning method behind this approach will be questioned by proposing David Hume's induction problem (Hume, 2003). Then an attempt will be made to provide a solution to this problem through Bayesianism. Lastly, some critical problems to the interpretation and use of the results will be discussed.

Big Data

In some sense Big Data can be regarded as a unique collection of the concepts coming from different fields in computer science and the related fields (Netten et al., 2016). The implementation of these concepts gives rise to many novel applications potentially. Two views may be taken on the exploitation of Big Data. In the first view, referred to as the black box, the implementation of the concepts and the relationships between the concepts are not clear and considered as a black box. We exploit these concepts by offering 'big data sets' and some constraints to the black box and observe the results that are provided by the black box. In case that we are not satisfied by the results, we alter some of the constraints and perhaps also the data and offer them again to the black box. We may repeat this process until we are satisfied by the results provided by the black box.

In the second view, referred to as the open box, the implementation of the concepts and their relationships is fully documented and we are able to track exactly how the results are obtained. In case we are not satisfied by the results obtained from Big Data, we may find out which concepts and relationships contribute to this dissatisfaction and adapt them accordingly to achieve satisfactory results. Due to the complexity of contemporary Big Data systems, questions need to be asked about the feasibility of the open box approach.

We take another view on Big Data, referred to as glass-box, which is in between the two (black box and open box) views. In this view, it is not necessary to know precisely all the ins and outs with regard to the implementation of the used concepts and their relationships. However, the crucial concepts of Big Data and possible relationships between the concepts are specified. Furthermore, it should be transparent how Big Data results are obtained. An example of a glass box view is depicted in Figure 1. In this example, we distinguish a set of algorithms for data processing, data analytics, and data visualization, a Hadoop cluster, and some data storage facilities. We see that the algorithms may exploit the Hadoop cluster for their jobs, and this cluster interacts with the data storage facilities. A Hadoop cluster consists of several machines, which are used to process a (complex) job. Via the so-called "map reduce" paradigm, a job is split into a set of smaller jobs and these jobs are distributed among the several machines for processing. Once a machine has completed its job, the result is collected. The collected results are composed to a final result.

Furthermore we see in Figure 1, the output of the data processing algorithms are used as input by the data analytics algorithms, and so on.

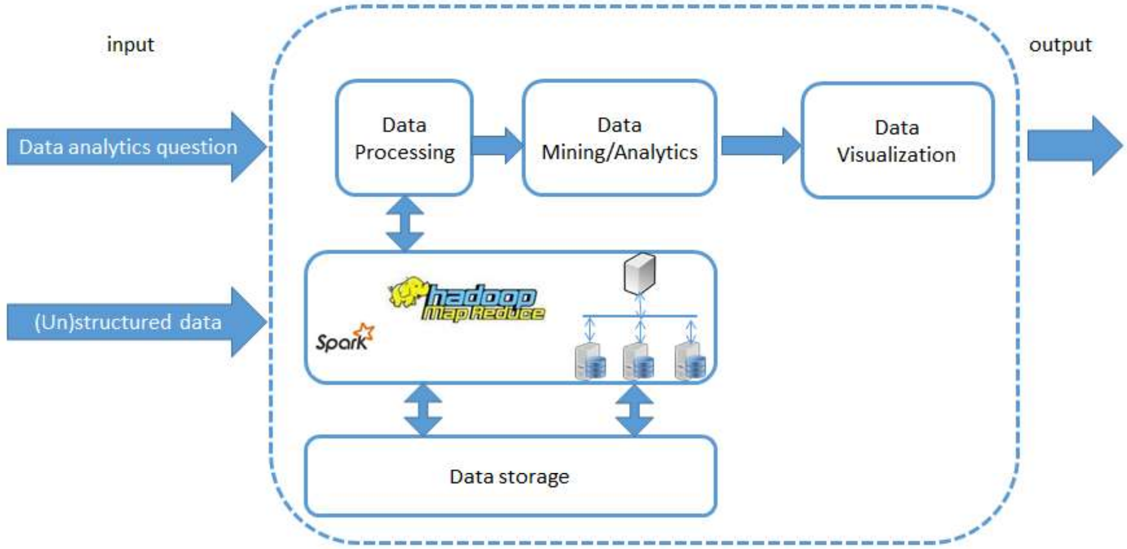

Figure 1 – Illustration of a glass box view

A crucial concept of Big Data is data analytics on which we will focus on in this paper. Data analytics has the goal of learning and drawing conclusions from large amount of integrated data. In order to analyze these large amounts of data, data mining is a key methodology for data analysts. Informally data mining can be regarded as the practice of examining large data sets in order to generate new information. On the other hand, the data set is supposed to represent the real world. If for example a correlation is found between advertisements with luxurious cars and high selling rate, then the result might be a prediction that new commercial campaigns should again involve these cars. This is based on the assumption that if the correlation between luxurious cars and high selling rates has been observed many times already, this correlation will also hold in the future.

Challenges

Formally, data mining is the induction of a model of the environment from large data sets (Choenni et al., 2005). When such models are generated, they are used to retrieve information from (like for prediction of unforeseen outcomes based some observed events). Inductive reasoning has been the subject of philosophical debate for years. Those models that are based on inductive reasoning pose two critical problems as described below.

Firstly, inductive reasoning assumes that when repeating patterns are observed these patterns will always exist and repeat. For example, when we continuously see that a certain event X leads to a certain event Y, we will automatically assume that the next time we observe X it will lead to Y again. Generally, you observe some concrete examples where something is true and from there a universal rule is induced. However, Hume argued that this is in fact a logically incorrect way of reasoning as inductive inferences do not necessarily have to be correct. This will be discussed more thoroughly later on in this paper.

The second problem stems from the fact that the models based on Big Data cannot fully capture the real world. Models are merely reflections of the real world and therefore they will necessarily miss some (relevant) parts of the real world. As the model might miss crucial information it can be an insufficient and thus skewed representation of the world. If this is the case, results derived from the model may be meaningless in the real world. This means that even when we have a perfectly sound reasoning method to generate the model from, we would still never be sure whether the information derived from it is correct since the model might be incomplete.

Furthermore, models are largely used for two reasons. Either as a way of understanding the environment or as an attempt to predict the environment. These reasons are both contradicting as well as reinforcing. Models that make good predictions do not necessarily have to be insightful(being contradictory), but insightful models can result in better predictions (being reinforcing). Each time the results are retrieved and applied to the real world, the model therefore needs to get updated. Essentially, the model continuously needs feedback to adjust its structure and parameters in order to remain as close of an accurate representation of the world as possible, see Figure 2.

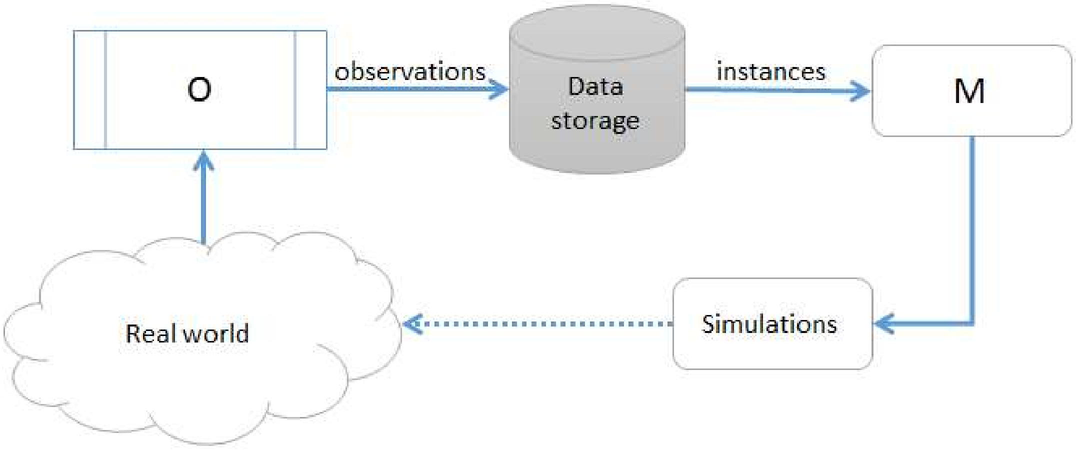

Figure 2 – Relationship between Big Data results and the real world

This new model can then again be used as an up to dated representation of the world to generate information from. Later on, we will explain how this approach in combination with the fact that models are incomplete can pose some ethical problems.

A Philosophical View

The problem of induction concerns the justification of inductive methods. Hume described them as methods that predict or infer that "instances of which we have had no experience resemble those of which we have had experience" (Hume, 2003). He argued that all human knowledge is derived from ideas and the relationship between ideas and impressions. People can recognize structural patterns systematically in these, which gives rise to concepts such as causality. However, it would also be possible to imagine that causality would for example cease to exist tomorrow. Nothing guarantees that patterns that we see today, must necessarily still exist tomorrow. For example, when ice is heated, we expect it to melt. However, it is possible to imagine that from tomorrow on this will no longer be the case. This is what Hume denoted as the induction problem. If nothing guarantees that these relationships will hold, we cannot really gain information from experience. So why do we intuitively reason inductively anyway?

Hume found the answer in habits and associations. When the same sequence of events keeps being observed many times, associations are automatically learned. For example, fire and warmth and ice and cold. Therefore, a certain cause does not have to be necessarily but is just expected by habit. Hume did not find this particularly problematic. If we assume a uniform nature, then it is logical that we would expect certain behaviors based on associations (Glymour, 2015). This does, however, not hold for the luxurious cars and advertisement example proposed earlier. Suppose that the observation has been made many times within the database that a relation between luxurious cars and high selling rates exists. As a result, the prediction is then yielded that a next advertisement will again be most successful if it involves luxurious cars. However, this would not necessarily be logical according to Hume's reasoning about the induction problem. Hume saw a sense of logical reasoning behind induction if we assumed a uniform nature. However, concepts like consumer behavior can be expected to be much less uniform (i.e., more dynamical) than the nature behavior is. It is therefore not completely intuitive to expect all concepts to appear to be uniform.

A well-known answer to Hume's induction problem is Immanual Kant's transcendental argument. Kant found the answer in the transcendental categories of relations. These are synthesized by empirical intuition and create as well as limit all human knowledge by synthetic a priori judgments. These categories form causality under time and space (Kant, 2004). The idea is that we cannot have knowledge of the world in itself, but only of how the world appears to us. However, we know that observations cannot be made without concepts such as causality. Therefore, Kant argued that these concepts precede the experience and can thus escape skepticism. Kant did not prove the existence of these concepts, he merely proved that someone cannot observe anything without adding it to the experience. Again, this argument does not translate well to cases like the luxurious cars and advertisements discussed earlier. In data mining, inductive reasoning is often used to explain abstract concepts like consumer behavior. These are not relations that are experienced.

Many more attempts were made to refute the induction problem that remained unsatisfying. Nowadays, it still seems like there is no solution to the induction problem that can guarantee the correctness of the results. However, in the twentieth century a movement called Bayesianism proposed an insightful alternative. The novel idea was that belief and knowledge do not necessarily have to be either correct or incorrect. Rather, they can come in degrees that conform to certain constraints related to the axioms of probability theory (Grimmett & Stirzaker, 2001). They therefore did not attempt to prove that induction leads to knowledge, but that it can result in a certain degree of knowledge. Probabilistic Bayesianism has now gained a well-established position in philosophy. It is in close correspondence to our intuition. If we see something happen a million times, it is highly likely, but not certain, that it will happen again. The inferences from induction can therefore gain an epistemic status somewhere between the two extremes of right and wrong. This status depends on the quality of the evidence and can be adjusted in the future when new evidence comes to light. This approach to inductive inferences is the only justification to our reasoning with Big Data. Consequently, all results we derive in this manner cannot be perceived as knowledge, but merely as probable conclusions. Whilst this does provide some justification, it is a weakness of the theoretical foundations on which this approach to Big Data is based.

It should be clear that a Big Data approach will not lead to universal truths but at most to some probable conclusions from a philosophical point of view. Furthermore, the probable conclusions are bounded by time and space. Therefore, caution and a good understanding of the probable conclusions are needed in exploiting the results obtained by (Big) Data analytics. To what extent caution should be taken depends on the application at hand. An important parameter that determines this caution is the velocity with which a concept is changing.

Illustrative examples

As explained in the foregoing, Big Data is collected and by means of data mining a model of the world is generated. This prior model will be used to apply Artificial Intelligence (AI) techniques on, which outputs certain results. The information from these results is then processed in the real world. As the results are processed, the world has changed and therefore the model needs to get updated. The prior model is replaced by a second model on which techniques can be applied to retrieve new information. For smaller simple cases this seems like a reasonable approach. However, when this is applied on a large scale by, for example, government institutions, this approach can pose some ethical problems. This will be illustrated by the following example. Suppose a government institution is in charge of policies regarding law reinforcement. The institution now races for example the challenge of fining as many drivers that commit traffic violations as possible. Therefore, a model of the world is induced from Big Data and some AI techniques are applied to them in order to answer the question of which strategy the policemen should follow.

Now suppose that the data include many correlations between drivers with a certain ethnic profile and the committing of traffic violations. It would seem beneficial, but may be unethical or against the law, to instruct all policemen to keep an extra eye on these drivers. This is where the problem arises. As the policy is now made to watch out for these drivers, the perceived world will be altered based on results that were retrieved from the previous model. Naturally, if many policemen are now focusing on these drivers, the same drivers will again get fined more often. This however does not necessarily have to mean that they do indeed commit traffic violations more often than the drivers with another profile. However, with every iteration of this method, the system will keep getting reinforced that it was initially right. With every new model these drivers will get over represented more and therefore the output will remain the same (i.e., the drivers from that certain ethnic profile will be picked up). This would be an ethical problem as policemen are now continuously targeting the same group of people with that certain ethnic profile.

The problem with this approach stems from the fact that the policies that are based on the model will also determine the righteous of the model. Thus, essentially, the model is only reinforcing itself. This illustrates the same problem that Karl Popper already identified. If you are trying to prove the hypothesis that all swans are white, just searching for extra white swans as evidence is not a valid method (Popper, 2005). In this case the hypothesis would be the retrieved result that drivers of the certain ethnicity are committing more violations, which you are then searching evidence for by ordering policemen to target these specific drivers. Therefore, when this method is applied on a large scale, the AI algorithms may actually and eventually influence the real world instead of simply capturing information from it. It could essentially give back that group of drivers accusing them of committing more traffic violations and the outcome will always hold true according to the model. In the context of mortgage provisions, the above discussed issues (may) play a role as well (Berkovec et al., 1994).

In brief, we argue that the real world rules may change. Then, the current rules may not apply in the future, which actually asks for continuous learning. However, even if we want to learn the new rules, the learning is useless if the old rules dictate the scope and space of observations, such that the changes in the real world cannot be observed.

Towards Solutions for Big Data Challenges

Given the fact that the interpretation of data analysis results is far from trivial, we propose to consider these results as a central body of knowledge. From this knowledge, we can derive a hypothesis for an individual case. Then, we search for other evidence(s) that may support or weaken this hypothesis. As an example, suppose that we feed a Big Data tool with a large amount of data of those who were involved in car accidents. After analyzing the data, the tool produces the following profile as result "young men living in areas with zip code 1234 have a higher than average probability to cause car accidents". Now we have a young man, named Mr. Green, who lives in this area (with zip code 1234). A challenging question is how to apply the profile in the case of Mr. Green.

Despite the fact that the profile is a statistical truth and cannot be projected on a specific individual as fact, we search for a way to exploit the knowledge that is captured in the profile. Therefore, we formulate the following hypothesis for Mr. Green: "Mr. Green will cause car accidents". To evaluate this hypothesis, we may use two strategies as described below.

In the *first* strategy, we search for evidences that support the hypothesis, e.g., Mr. Green caused car accidents in the past or an expert agrees with the hypothesis. Note that these evidences should not be based on or derived from the data that are used in the data analysis. Using the same data will shed no new light on the hypothesis, instead it will incorrectly strengthen the hypothesis further. If enough supporting evidence has been collected, the hypothesis can be accepted and the case of Mr. Green should be investigated further. In our example, suppose we examine the drives of Mr. Green to find more evidences that support the hypothesis (that he causes a car accident). Assume that we find that Mr. Green was involved at several car accidents in the past and has filed several insurance claims. However, it was not always clear whether Mr. Green caused the accidents or others who were involved in the accidents. Furthermore, a check on his bank accounts shows that they were practically empty at the time of the accidents. This may be considered as evidence for Mr. Green, who may swindle insurance companies. A disadvantage of the strategy of collecting supporting evidences is that it may strengthen confirmation biases and lead to a *self-fulfilling prophecy*, i.e., a false hypothesis might become true due to this bias.

In the *second* strategy, we search for evidence that weakens the hypothesis "Mr. Green will cause car accidents". If we find a set of evidence that gives rise to rejecting the hypothesis, no action should be taken against Mr. Green. Unlike in the previous strategy, the same data set from which the hypothesis is derived may be used to search for evidences that weaken the hypothesis. The data that are used to induce the hypothesis from the profile can be used to infer other profiles that may be in advantage of Mr. Green. For example, if we are able to derive a new profile like: "Men living in zip code 1233 and have not filed a car insurance claim in the past 5 years and do not drive in leased cars are cautious drivers". If Mr. Green satisfies this new profile, then this will weaken the hypothesis that Mr. Green will cause car accidents frequently. Other data sets can also be used to search for evidences to weaken the hypothesis. A disadvantage of this strategy of collecting weakening evidences is that it may lead to a *self-denying prophecy*, i.e., a true hypothesis might become false due to bias.

Which strategy to use for which application depends on the nature of an application and the impact of possible false positives and false negatives. We note that a false positive refers to an accepted hypothesis while it is false, and a false negative refers to a rejected hypothesis while it is true. The first strategy tends to reduce the false negatives and to increase the false positives, while the reverse is true for the second strategy. The second strategy is applied in contemporary judicial courts. The public prosecutor makes a statement, which is based on police investigations of a suspect. Subsequently, in court the lawyer of the suspect aims to disprove this statement by presenting counter proofs. Such a strategy is chosen to only convict someone if he is indeed guilty, i.e., to avoid false positives. In sensitive applications that have a large impact on someone's life, the second strategy is recommended. The first strategy, on the other hand, focuses on strengthening a hypothesis and avoiding false negatives.

In some application areas that are related to public security (like searching for terrorists), the first strategy might be useful.

Independent of which strategy is used for an application, it makes sense to have an estimate of the impact of false positives and false negatives, and a procedure to anticipate on them. One of these strategies (or perhaps a mix of both) should be tailored to the application at hand.

Conclusion

The use of inductive reasoning to retrieve information from Big Data can, to a certain extent, be justified if we take a Bayesianistic position. This, however, means that we have to give up the idea that the results that we obtain are certainly true. Thus, we acknowledge that we might only be dealing with probable information. But if our model is induced and induction is only probable, the results can be wrong. This, combined with the fact that models can never be an exact representation of the real world anyway, results in much room for error. In practice, the model gets continuously updated to compensate for these errors. However, as explained earlier, when government institutions start applying this method on large scale this can pose some challenges. As illustrated, it can result in ethical problems.

All in all, a lot of caution is needed when using this approach. It is important to be aware of the fact that there are many contributing factors that can be the cause of error. It is clear that induction does not always provide correct results. However, as mentioned before, even if it did, it would still not be possible to completely trust the output. After all, then we would still be dealing with a model which in itself is not an exact representation of the world. Therefore, there is too much uncertainty and the information gained is thus not reliable enough to directly base policies on, especially not when being applied to sensitive cases such as with the example of accident causing drivers. It could possibly have negative consequences as wrong interpretations of the results could impact the real world adversely. Nevertheless, as induction is for a large part a justifiable reasoning method, the results are not entirely useless. The results should be interpreted as rough estimates that can serve as guidelines for further research. Based on this research and with human intervention, these results can then indirectly be used to base policies on. We have discussed two preliminary strategies to benefit from Big Data. However, these strategies need to be further elaborated, which is a topic for our future research.

References

Berkovec, J. A., Canner, G. B., Gabriel, S. A., & Hannan, T. H. (1994). Race, redlining, and residential mortgage loan performance. *The Journal of Real Estate Finance and Economics*, *9*(3), 263-294.

Choenni, S. (2000). Design and implementation of a genetic-based algorithm for data mining. *In VLDB (pp. 33-42), Proceedings of the 26th International Conference on the Very Large Databases*, Morgan Kaufmann.

Choenni, S., Bakker, R., Blok, H. E. & Laat, de, R. (2005). Supporting technologies for knowledge management. *In Knowledge Management and Management Learning* (pp. 89-112). Springer, Boston, MA.

Choenni, S., Bargh, M. S, Netten, N. & Braak, van den, S. (2018). Using data analytics results in practice: Challenges and solution directions. *In ICT-Enabled Social Innovation for the European Social Model (IESI-ESM)*, F. Davide and G. Misuraca (eds), IOS PRESS (to appear).

Glymour, C. (2015): *Thinking Things Through: an introduction to philosophical issues and achievements*. MIT Press: (pp. 171-179).

Grimmett, G., & Stirzaker, D. (2001). *Probability and random processes*. Oxford University Press.

Howson, C. & Urbach, P. (2006) *Scientific reasoning: the Bayesian approach*. Open court Publishing: Chapter 1.

Hume, D. (2003): *A treatise of human nature*. Courier Corporation: section VI.

Kant, I. (2004): *Prolegomena to Any Future Metaphysics: That Will Be Able to Come Forward as Science: With Selections from the Critique of Pure Reason*. Cambridge university press: (pp. 79-122).

Kim, G., Trimi, S. & Chung, J. (2014) Big-Data Applications in the Government Sector. *Communications of the ACM,* 57,3, 78-85.

Netten, N., Braak, van den , S., Choenni, S. & van Someren, M. (2016). A Big Data Approach to Support Information Distribution in Crisis Response. *In Proceedings of the 9th International Conference on Theory and Practice of Electronic Governance* (pp. 266-275). ACM.

Netten, N., Bargh, M. S. & Choenni, S. (2018). Exploiting Data Analytics for Social Services: On Searching for Profiles of Unlawful Use of Social Benefits. *In Proceedings of the 11th International Conference on Theory and Practice of Electronic Governance*. ACM.

Popper, K. (2005). *The logic of scientific discovery*. Routledge.

Roger, C.H, Chiang, H. L. & Storey, V.C. (2012). Business intelligence and analytics: from big data to big impact. *MIS quarterly* (pp. 1165-1188).

Tien, J.M. (2013). Big data: Unleashing information. *Journal of Systems Science and Systems Engineering* (pp. 127-151).