Responsible Data Science in a Dynamic World

Wil Aalst

Responsible Data Science in a Dynamic World
The Four Essential Elements of Data Science

Wil M.P. van der Aalst

Lehrstuhl für Informatik 9, Process and Data Science, RWTH Aachen University,
D-52056 Aachen, Germany
wvdaalst@pads.rwth-aachen.de
http://vdaalst.com

Abstract. Data science is changing our world in many different ways. Data and the associated data science innovations are changing everything: the way we work, the way we move, the way we interact, the way we care, the way we learn, and the way we socialize. As a result, many professions will cease to exist. For example, today's call centers will disappear just like video rental shops disappeared. At the same time, new jobs, products, services, and opportunities emerge. Hence, it is important to understand the essence of data science. This extended abstract discusses the four essential elements of data science: "water" (availability, magnitude, and different forms of data), "fire" (irresponsible uses of data and threats related to fairness, accuracy, confidentiality, and transparency), "wind" (the way data science can be used to improve processes), and "earth" (the need for data science research and education). Next to providing an original view on data science, the abstract also highlights important next steps to ensure that data will not just change, but also improve our world.

Keywords: Data science · Responsible data science · Process mining · Big data.

1 Data Science

This extended abstract is based on a keynote given at the IFIP World Computer Congress (WCC 2018) on 18 September 2018, in Poznan, Poland. The main theme of WCC 2018 was "Information Processing in an Increasingly Connected World: Opportunities and Threats". Data science is the main driver for the changes that create these opportunities and threats. Recent reports [6, 7] indicate that many jobs will cease to exist because of advances in machine learning, artificial intelligence, robotics, and other forms of smart automation. These advances are only possible because of both the availability of data and progress in data science.

It is not easy to define data science. The data science pipeline shown in Figure 1 illustrates the breadth of the discipline. The "infrastructure" part of the pipeline is concerned with the huge volume and incredible velocity of data. Hence,

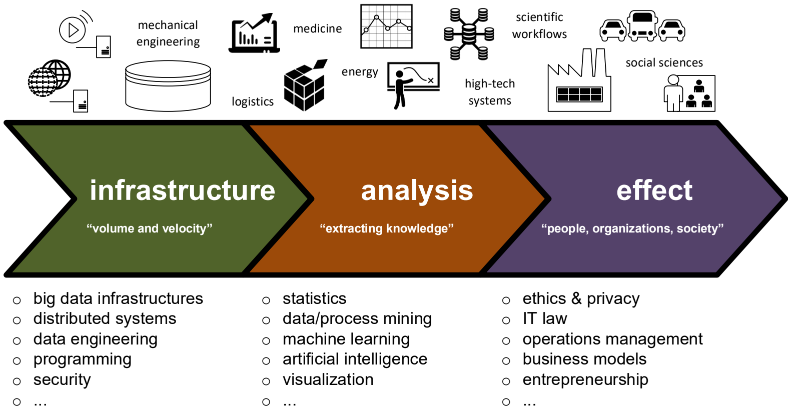

Fig. 1. The data science pipeline showing that different capabilities are needed to turn data into value.

the primary focus is on making things scalable and instant. The "analysis" part of the pipeline is concerned with extracting knowledge. This is about providing answers to known and unknown unknowns.[1] The "effect" part of the pipeline is concerned the impact of data science on people, organizations, and society. Here legal, ethical, and financial aspects come into play.

The uptake of the Internet of Things (IoT) illustrates the pivotal role of data science. More and more devices (light bulbs, clothes, refrigerators, containers, bicycles, etc.) are connected to the internet and produce data. These devices are becoming "smart" by learning from the data collected. The Internet of Things (IoT) depends on the whole data science pipeline shown in Figure 1. We are (or will be) surrounded by smart devices collecting data and the impact of this cannot be overestimated.

In the remainder, we define the four essential elements of data science. As metaphor we use the classical four elements: "water", "fire", "wind", and "earth". According to the Empedocles, a Greek pre-Socratic philosopher who lived in Sicily in the fifth century B.C., all matter is comprised of these four elements. Other ancient cultures had similar lists, sometimes also composed of more elements (e.g., earth, water, air, fire, and aether) that tried to explain nature and complexity of all matter in terms of simpler substances. Today, we know that this is not the case. However, for data science, we are still in the phase where we are looking for the essential elements. This paper uses "water" as a placeholder for the availability of different forms of data, "fire" as a placeholder for irresponsible uses of data (e.g., threats to fairness, accuracy, confidentiality, and transparency), "wind" as a placeholder for the way that data science

[1] "There are known knowns; there are things we know we know. We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns the ones we don't know we don't know." (Donald Rumsfeld, February 12, 2002)

can be used to improve processes, and “earth” as a placeholder for education and research (i.e., the base of data science) underpinning all of this. These four essential elements are discussed in the remaining sections.

2 The “Water” of Data Science

The first essential element of data science (“water”) is the data itself. The exponential growth of data is evident. Figure 2 (inspired by the analysis in [9]) shows the rapid developments in terms of *costs* (things are getting exponentially cheaper), *speed* (things are going exponentially faster), and *miniaturization* (things are getting exponentially smaller). This is not limited to *processing* (i.e., CPU and GPU processors), but also applies to *storage* and *communication.* Consider for example the costs of storage. To store one megabyte (MB) of data in the sixties one would need to pay one million euros. Today, one can buy a 10TB harddisk for less than 300 euro, i.e., 0.00003 cents per MB. Another example is the bandwidth efficiency, also called spectral efficiency, which refers to the information rate that can be transmitted over a given bandwidth. It is the net bitrate (useful information rate excluding error-correcting codes) or maximum throughput divided by the bandwidth in hertz of a communication channel or a data link. The spectacular progress of our data handling capabilities illustrated by Figure 2, explains why data science has become on of the key concerns in any organization. In the sixties, we only had a few “drops of data” whereas today we are facing a “tsunami of data” flooding our society.

Clearly, data science has its roots in statistics, a discipline that developed over four centuries [1]. John Graunt (1620-1674) started to study London’s death records around 1660. Based on this he was able to predict the life expectancy of a person at a particular age. Francis Galton (1822-1911) introduced statistical concepts like regression and correlation at the end of the 19th century. Although data science can be seen as a continuation of statistics, the majority of statisticians did not contribute much to recent progress in data science. Most statisticians focused on theoretical results rather than real-world analysis problems. The computational aspects, which are critical for larger data sets, are typically ignored by statisticians. The focus is on generative modeling rather than prediction and dealing with practical challenges related to data quality and size. When the data mining community realized major breakthroughs in the discovery of patterns and relationships (e.g., efficiently learning decision trees and association rules), most statisticians referred to these discovery practices as “data fishing”, “data snooping”, and “data dredging” to express their dismay [1, 4, 10].

Put differently; most statisticians were focused on techniques to make reliable statements given a few “drops of data”. Such viewpoints turned out to be less effective when dealing with “tsunamis of data”.

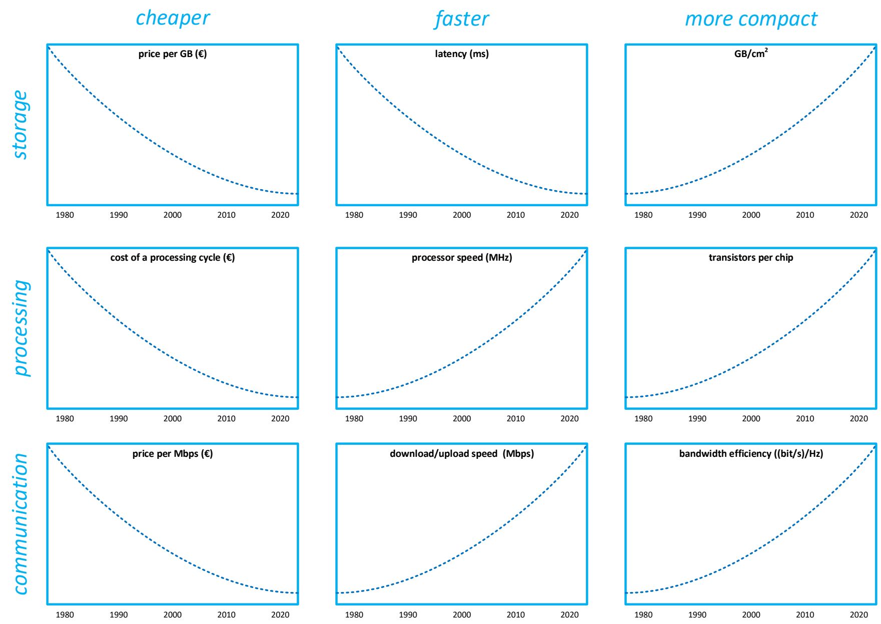

Fig. 2. Moore's law predicts an exponential growth of the number of transistors per chip. This can be generalized to storage and transition and also applies to costs and speed.

3 The "Fire" of Data Science

The second essential element of data science ("fire") refers to the dangers of using data in an irresponsible way. Data abundance combined with powerful data science techniques has the potential to dramatically improve our lives by enabling new services and products, while improving their efficiency and quality. Many of today's scientific discoveries (e.g., in health) are already fueled by developments in statistics, mining, machine learning, artificial intelligence, databases, and visualization. At the same time, there are also great concerns about the use of data. Increasingly, customers, patients, and other stakeholders are concerned about irresponsible data use. Automated data decisions may be unfair or non-transparent. Confidential data may be shared unintentionally or abused by third parties.

From 2015 until 2017, the author led the *Responsible Data Science* (RDS) initiative where the strongest Dutch data science groups joined forces to address problems related to *fairness*, *accuracy*, *confidentiality*, and *transparency* (www.responsibledatascience.org). The goal of RDS is to show that data science techniques, infrastructures and approaches can be made responsible by design. *Responsible Data Science* (RDS) revolves around four main challenges:

- *Data science without prejudice* - How to avoid unfair conclusions even if they are true?
- *Data science without guesswork* - How to answer questions with a guaranteed level of accuracy?
- *Data science that ensures confidentiality* - How to answer questions without revealing secrets?
- *Data science that provides transparency* - How to clarify answers such that they become indisputable?

The term *green data science* was introduced for cutting-edge solutions that enable individuals, organizations and society to benefit from widespread data availability while ensuring *Fairness*, *Accuracy*, *Confidentiality*, and *Transparency* (FACT) [2].

Naïvely one could think that “fire” can be controlled by “water”, however this is not the case. When considering RDS, it is better to consider data as “oil” rather than “water”. It needs to be controlled and stored carefully.

There is a need for new and positive data science techniques that are responsible (i.e., “green”) by design. This cannot be solved by stricter laws. Using the metaphor of “green energy”: We should not be against the use of energy (“data”), but address the pollution caused by traditional engines. Fortunately, there are plenty of ideas to make data science green. For example, discrimination-aware data mining [8] can be used to ensure fairness and polymorphic encryption can be used to ensure confidentiality.

4 The “Wind” of Data Science

The third essential element of data science (“wind”) is concerned with the way data and processes interact. Storing and processing data is not a goal in itself. Data are there to support processes. The campaign “The best run companies run SAP” illustrates that the purpose of information systems is to ensure that processes run well. Data science can help organizations to be more effective, to provide a better service, to deliver faster, and to do all of this at lower costs. This applies to logistics, production, transport, healthcare, banking, insurance, and government. This also applies to individuals. Data science will increasingly support our personal workflows and take over tasks, or at least support them. Data (“water”) can be used to manage and support processes (“wind”) through the use of data science technologies.

An emerging technology linking “water” and “wind” is *process mining* [1]. Process mining bridges the gap between traditional model-based process analysis (e.g., simulation and other business process management techniques) and data-centric analysis techniques such as machine learning and data mining. Process mining seeks the confrontation between event data (i.e., observed behavior) and process models (hand-made or discovered automatically) [1]. The process-mining spectrum is broad and includes techniques for process discovery, conformance checking, prediction, and bottleneck analysis. These techniques tend to be very different from mainstream data mining and machine learning techniques which are typically not process-centric.

Consider for example the topic of *Robotic Process Automation* (RPA). RPA is an umbrella term for tools that operate on the user interface of other computer systems in the way a human would do. RPA aims to replace people by automation done in an “outside-in” manner [3]. This differs from the classical “inside-out” approach to improve information systems. Unlike traditional workflow technology, the information system remains unchanged. The robots are replacing humans while leaving the back-end systems intact. RPA is a way to support processes in a more cost-effective manner. However, this requires learning what humans do by observing them. Data science approaches like process mining can be used to learn the behavior of people doing routine tasks. After the desired behavior has been “played in”, it can be “played out” to handle new cases in an intelligent manner.

RPA illustrates that data science will lead to new trade-offs between what humans do and what robots do [6, 7]. These trade-offs are interesting: How to distribute work between given breakthroughs in data science? Obviously, the question needs to take the “fire” dimension into account.

5 The “Earth” of Data Science

The fourth essential element of data science (“earth”) is concerned with the foundations of a data-driven society: *education* and *research*. Education (in every sense of the word) is one of the fundamental factors in the development of data science. Data science education is needed at any level. People need to be aware of the way algorithms make decisions that may influence their lives. Privacy discussions reveal the ignorance of policy makers and end users. Moreover, to remain competitive, countries should invest in data science capabilities. This can only be realized through education. Data science research plays a similar role. On the one hand, it is key for our education. On the other hand, research is needed to address the many technological and societal challenges (e.g., ensuring fairness, accuracy, confidentiality, and transparency).

Currently, eight of the world’s ten biggest companies, as measured by market capitalization, are American: Apple, Alphabet (incl. Google), Microsoft, Amazon, Berkshire Hathaway, Facebook, JPMorgan Chase, and Bank of America.[2] The two remaining companies are Chinese: Alibaba and Tencent Holdings. This shows the dominance of a few countries due to investments in IT. Most of the companies are relatively new and emerged through the smart use of data. Amazon and Alibaba are dominating the way we buy products. Google is controlling the way we search. Facebook is controlling the way we socialize. Apple, Alphabet, and Microsoft are controlling the platforms we use (iOS, Android, and Windows). Consider for example Facebook. On the one hand, many people are expressing concerns about the use of data. On the other hand, Facebook has

[2] Based on market capitalization data by Bloomberg on 31 March 2018.

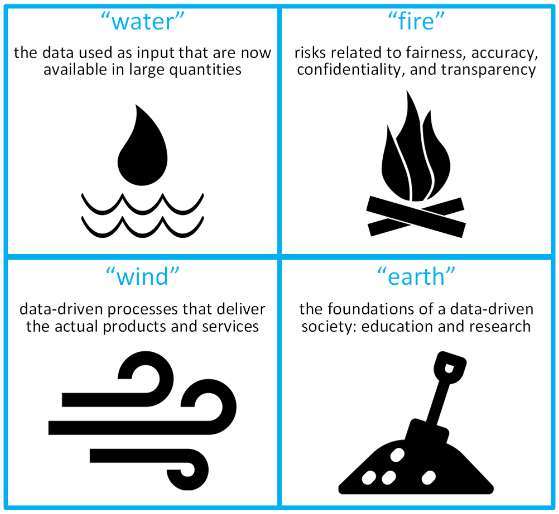
"water"
the data used as input that are now available in large quantities
"fire"
risks related to fairness, accuracy, confidentiality, and transparency
"wind"
data-driven processes that deliver the actual products and services
"earth"
the foundations of a data-driven society: education and research

Fig. 3. The "water", "fire", "wind", and "earth" of data science.

over 2 billion monthly active users that provide personal information in order to use social media. One of the problems of data science is that due to economies of scale "the winner takes it all". This may also apply to education, e.g., on Coursera a few US universities are dominating data science education.

Data science literacy and major public investments are needed to address these concerns. This cannot be left to "the market" or solved through half-hearted legislation like the European General Data Protection Regulation (GDPR) [5].

6 Epilogue

This extended abstract aimed to present some of the key messages of the keynote presentation for the IFIP World Computer Congress (WCC 2018). It stresses the importance of data science for people, organizations, and society. Just like computer science emerged as a new discipline from mathematics in the early eighties, we can now witness that the data science discipline is emerging from computer science, statistics, and social sciences.

In this paper, we discussed the four essential elements of data science: "water" (availability, magnitude, and different forms of data), "fire" (irresponsible uses of data and threats related to fairness, accuracy, confidentiality, and transparency), "wind" (the way data science can be used to improve processes), and "earth" (the need for data science research and education). By presenting data science in this manner, we hope to get more attention for process-centric forms of data science (e.g., process mining), responsible data science, data science education, and data science research. The dominance of a few companies and countries when it comes to data science is undesirable and requires the attention of politicians and policymakers. The IFIP could and should play an active role in this discussion.

References

1. W.M.P. van der Aalst. *Process Mining: Data Science in Action*. Springer-Verlag, Berlin, 2016.
2. W.M.P. van der Aalst. Responsible Data Science: Using Big Data in a "People Friendly" Manner. In S. Hammoudi, L. Maciaszek, M. Missikoff, O. Camp, and J. Cordiero, editors, *Enterprise Information Systems*, volume 291 of *Lecture Notes in Business Information Processing*, pages 3–28. Springer-Verlag, Berlin, 2017.
3. W.M.P. van der Aalst, M. Bichler, and A. Heinzl. Robotic Process Automation. *Business and Information Systems Engineering*, 60(4):269–272, 2018.
4. L. Breiman. Statistical Modeling: The Two Cultures. *Statistical Science*, 16(3):199–231, 2001.
5. European Commission. Proposal for a Regulation of the European Parliament and of the Council on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data (General Data Protection Regulation). 9565/15, 2012/0011 (COD), June 2015.
6. C.B. Frey and M.A. Osborne. The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change*, 114(C):254–280, 2017.
7. J. Hawksworth, R. Berriman, and S. Goel. Will Robots Really Steal Our Jobs? An International Analysis of the Potential Long Term Impact of Automation. Technical report, PricewaterhouseCoopers, 2018.
8. D. Pedreshi, S. Ruggieri, and F. Turini. Discrimination-Aware Data Mining. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 560–568. ACM, 2008.
9. R.Brennenraedts, A. Vankan, R. te Velde, B. Minne, J. Veldkamp, and B. Kaashoek. The Impact of ICT on the Dutch Economy. Technical report, Dialogic, 2014.
10. J.W. Tukey. The Future of Data Analysis. *Annals of Mathematical Statistics*, 33(1):1–67, March 1962.