# The Data Science Revolution

Wil Aalst

# The Data Science Revolution

## How learning machines changed the way we work and do business

Wil M.P. van der Aalst

[1] Process and Data Science, RWTH Aachen University, Aachen, Germany
[2] Fraunhofer FIT, Sankt Augustin, Germany
wvdaalst@rwth-aachen.de
www.vdaalst.com

**Abstract.** Data science technology is rapidly changing the role of information technology in society and all economic sectors. Artificial Intelligence (AI) and Machine Learning (ML) are at the forefront of attention. However, data science is much broader and also includes data extraction, data preparation, data exploration, data transformation, storage and retrieval, computing infrastructures, other types of mining and learning, presentation of explanations and predictions, and the exploitation of results taking into account ethical, social, legal, and business aspects. This paper provides an overview of the field of data science also showing the main developments, thereby focusing on (1) the growing importance of learning from data (rather than modeling or programming), (2) the transfer of tasks from humans to (software) robots, and (3) the risks associated with data science (e.g., privacy problems, unfair or nontransparent decision making, and the market dominance of a few platform providers).

**Keywords:** Data science · Machine learning · Artificial Intelligence · Responsible data science · Big data

## 1 Introduction

The International Federation for Information Processing (IFIP) was established in 1960 under the auspices of UNESCO as a result of the first World Computer Congress held in Paris in 1959. This year we celebrate the 60th anniversary of IFIP. IFIP was created in 1960 because of the anticipated impact and transformative power of information technology. However, the impact of information technology over the past 60 years has been even larger than foreseen. Information technology has dramatically transformed the lives of individuals and businesses. Over the last 60 years, *data science*, i.e., extracting knowledge and insights from structured and unstructured data, has become the main driver of such transformations. In this paper, we reflect on the impact of data science and key developments.

In [2], data science is defined as follows: "*Data science is an interdisciplinary field aiming to turn data into real value. Data may be structured or unstructured,*

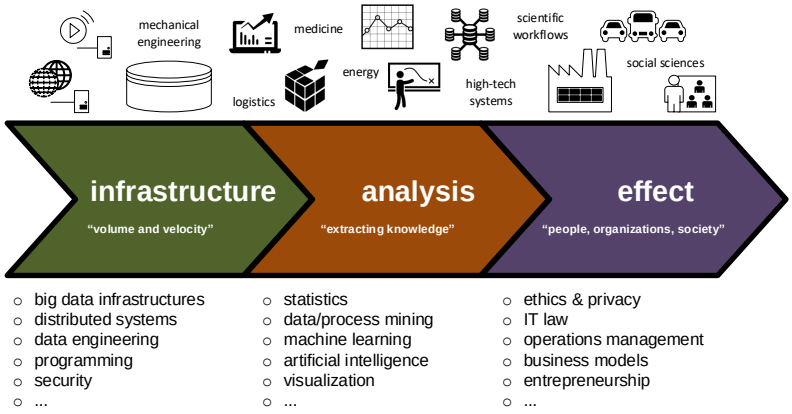


**Fig. 1.** Overview of the key ingredients of data science [4].

*big or small, static or streaming. Value may be provided in the form of predictions, automated decisions, models learned from data, or any type of data visualization delivering insights. Data science includes data extraction, data preparation, data exploration, data transformation, storage and retrieval, computing infrastructures, various types of mining and learning, presentation of explanations and predictions, and the exploitation of results taking into account ethical, social, legal, and business aspects.*" Data science can be seen as an umbrella term for machine learning, artificial intelligence, mining, Big data, visual analytics, etc. The term is not new. Turing award winner Peter Naur (1928-2016) first used the term 'data science' long before it was in vogue. In 1974, Naur wrote [15]: "A basic principle of data science, perhaps the most fundamental that may be formulated, can now be stated: The data representation must be chosen with due regard to the transformation to be achieved and the data processing tools available" . In [15], Naur discusses 'Large Data Systems' referring to data sets stored on magnetic disks having a maximum capacity of a few megabytes. Clearly, the notion of what is large has changed dramatically since the early seventies, and will continue to change.

Figure 1 visualizes the above definition. The top part shows that data science can be applied in many different areas, i.e., most data science approaches are generic. The middle part shows that there are three main ingredients: *infrastructure* (concerned with the huge volume and incredible velocity of data), *analysis* (concerned with extracting knowledge using a variety of techniques), and *effect* (concerned the impact of data science on people, organizations, and society). The diagram shows the *interdisciplinary nature* of data science. As an example, take a self-driving car. To build a self-driving car one needs an *infrastructure* composed of sensors (camera, lidar, radar, etc.), hardware controllers, networking capabilities, powerful processing units, and *analysis* techniques for perception (e.g., convolutional neural networks), localization, prediction, planning, and control using this infrastructure. However, one also needs to consider

the *effect*. The self-driving car has to be economically feasible and may require new business models. While creating such a car, one also needs to consider legal and ethical implications. In July 2016, Tesla reported the first fatality of a driver in a self-driving car triggering heated debates. Who is responsible when the car crashes? What decisions should be taken when a crash is unavoidable (e.g. protect passengers or pedestrians)? Who owns the data collected by the self-driving car? Due to the huge impact of data science on people, organizations, and society, many legal, ethical, and financial aspects come into play.

As Figure 1 shows, the field of data science is broad, building on multiple scientific disciplines. How does data science relate to hyped terms such as Big data, Artificial Intelligence (AI), and Machine Learning (ML)? Big data seems to be out of fashion and AI and ML are the new buzzwords used in the media. AI may mean everything and nothing. On the one hand, the term AI roughly translates to "using data in an intelligent manner" looking at its use in the media. This means that everything in Figure 1 is AI. On the other hand, the term is also used to refer to very specific approaches, such as deep neural networks [13]. The same applies to the term ML. All subfields of data mining (classification, clustering, patterns mining, regression, logistic regression, etc.) can be seen as forms of machine learning. However, ML is also used to refer to only deep learning.

John McCarthy coined the term AI in 1955 as "the science and engineering of making intelligent machines". Today, the field of AI is typically split in symbolic AI and non-symbolic AI. *Symbolic AI*, also known as Good Old Fashioned AI (GOFAI), uses high-level symbolic (i.e., human-readable) representations of problems, logic, and rules. Experts systems tend to use symbolic AI to make deductions and to determine what additional information it needs (i.e., what questions to ask) using human-readable symbols. The main disadvantage of symbolic AI is that the rules and knowledge have to be hand-coded. *Non-symbolic AI* does not aim for human-readable representations and explicit reasoning, and uses techniques imitating evolution and human learning. Example techniques include genetic algorithms, neural networks and deep learning. The two main disadvantages of non-symbolic AI are the need for a lot of training data and the problem of understanding why a particular result is returned. Symbolic AI is still not widely adopted in industry. Although non-symbolic AI performed worse than symbolic AI for many decades, by using back-propagation in multi-layer neural networks, non-symbolic AI started to outperform conventional approaches [17, 13]. As a result, these techniques are now also used in industry for tasks such as speech recognition, automated translation, fraud detection, image recognition, etc.

The successes of non-symbolic AI are amazing. However, AI is only a small part of data science, often tailored towards specific tasks (e.g., speech recognition) and using specific models (e.g., deep convolutional neural networks). The same applies to ML (which is often considered to be a subfield of AI). When the term AI or ML is used in the media, this often refers to data mining, pat-

tern/process mining, statistics, information retrieval, optimization, regression, etc.

This paper aims to 'demystify' data science, present key concepts, discuss important developments. We also reflect on the impact of data science on the way we work and do business. The paper is partly based on a keynote given at the IFIP World Computer Congress (WCC 2018) on 18 September 2018, in Poznan, Poland. It extends the corresponding keynote paper [4].

The remainder is organized as follows. Section 2 metaphorically 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). Section 3 discusses the shift from modeling and programming to data-driven learning enabled by the abundance of data. Due to the uptake of data science, traditional jobs and business models will disappear and new ones will emerge. Section 4 reflects on these changes. Data science can make things cheaper, faster, and better. However, also negative side-effects are possible (see the "fire" element mentioned before). Therefore, Section 5 discusses the topic of responsible data science in the context of the growing dominance of digital platforms. Section 6 concludes this paper.

## 2 The Four Essential Ingredients of Data Science

In this section, we define the four essential elements of data science [4]. As metaphors, we use the classical four elements: "water", "fire", "wind", and "earth" (see Figure 2). According to Empedocles, 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 the nature and complexity of all matter in terms of simpler substances. To explain the essence of data science, we use "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 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. Note that Figure 2 complements Figure 1, allowing us to emphasize specific aspects.

### 2.1 The "Water" of Data Science

The first essential element of data science ("water") is the data itself [4]. The exponential growth of data and data processing capabilities since the establishment of IFIP in 1960 is evident:

- Things are getting exponentially *cheaper*, e.g., the price of storage dropped from one million euros per MB in the 1960-ties to 0.00002 cents per MB today.

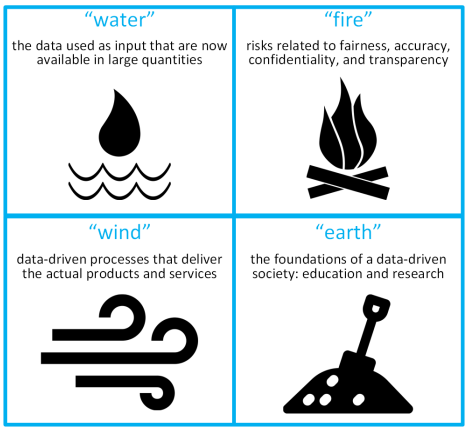


**Fig. 2.** The "water", "fire", "wind", and "earth" of data science [4].

- Things are getting exponentially *faster*, e.g., the number of floating-point operations per second increased from a few kFLOPS ($10^3$ floating-point operations per second) to hundreds of PFLOPS ($10^{15}$ floating-point operations per second).
- Things are getting exponentially *smaller*, e.g., the size of a transistor decreased from a few centimeters ($10^{-2}$ meter) to a few nanometers ($10^{-9}$ meter).

The reductions in price and size and the increase in speed apply to *processing* (i.e., CPU and GPU processors), *storage*, and *communication*. A GPU may have thousands of processors and a company like Google has over one million servers.

The above numbers illustrate our increased capabilities to process large amounts of data. These are used to align the digital world and the physical world. For example, organizations are collecting 'events' at a large scale [2]. Events may take place inside a machine (e.g., an X-ray machine, an ATM, or baggage handling system), inside an enterprise information system (e.g., an order placed by a customer or the submission of a tax declaration), inside a hospital (e.g., the analysis of a blood sample), inside a social network (e.g., exchanging e-mails or twitter messages), inside a transportation system (e.g., checking in, buying a ticket, or passing through a toll booth), etc. The uptake of the so-called Internet of Things (IoT) resulted in many connected devices ranging from light bulbs to wearable heart monitors [18]. Events may be 'life events', 'machine events', or 'organization events'. These may be stored in traditional relational databases (e.g., Microsoft SQL Server, Oracle Database, MySQL, and IBM DB2), NoSQL databases (e.g., CouchDB, MongoDB, Cassandra, and HBase), or distributed ledgers using blockchain technology (e.g., Ethereum, NEO, Cardano). The term *Internet of Events* (IoE), coined in [1], refers to the omnipresence of event data in all application domains.

In 2001, Doug Laney introduced the first three V's describing challenges related to Big data: Volume, Velocity, and Variety [12]. Later, additional V's were added: Veracity, Variability, Visualization, Value, Venue, Validity, etc. The above refers to the first 'V' (Volume) describing the incredible scale of some data sources. The second 'V' (Velocity) refers to the speed of the incoming data that need to be processed. In many applications, it has become impossible to store all data and process it later. Such streaming data needs to be handled immediately. The third 'V' (Variety) refers to the different types of data coming from multiple sources. Structured data may be augmented by unstructured data (e.g. free text, audio, and video).

### 2.2 The "Fire" of Data Science

The second essential element of data science ("fire") refers to the dangers of using data in an irresponsible way [4]. 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-based decisions may be unfair or non-transparent. Confidential data may be shared unintentionally or abused by third parties.

When IFIP was created sixty years ago, one could not foresee the possible risks related to data science. The Facebook-Cambridge Analytica scandal in 2018 and many other scandals involving unacceptable uses of data heavily influenced public opinion. Also books such as "Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy" [16], created increased awareness of the risks associated with data science. The European General Data Protection Regulation (GDPR) [8] is a response to such risks. However, legislation such as GDPR may also inhibit certain applications. Hence, technological solutions involving distribution and encryption are needed. In Section 5, we elaborate further on the topic of responsible data science.

### 2.3 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 typically not a goal in itself. Data are there to support processes. 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.

Data science can be used to *improve tasks within the process*; e.g., certain checks can be automated using a neural network trained on many examples. Data science can also be used to *improve the design or management of the*

*whole process*, e.g., using process mining, one can identify root causes for specific bottlenecks and deviations. Data science can also be used to *create new products and services*, e.g., Spotify is able to recommend music based on the listener's preferences.

Clearly, there is a tradeoff between "water" and "wind". Giving up privacy concerns may lead to better processes, services, and products.

### 2.4 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* [4]. 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 all levels. People need to be aware of the way algorithms make decisions that may influence their lives. Hyped terms such as Big data, Artificial Intelligence (AI), and Machine Learning (ML) are not well understood and may mean anything and nothing (see Section 1). Some examples of phenomena that illustrate the need for education.

- *Correlation is not the same as causality.* It may be that ice cream sales and crime rates strongly correlate. However, this does not imply that one causes the other. The so-called hidden variable is the weather. Higher temperatures lead to higher ice cream sales and higher crime rates. It makes no sense to try to reduce crime by putting a ban on ice cream sales.
- *Simpson's paradox.* It may be that within different subsets of data a variable has a positive influence, whereas it as a negative influence when considering all the data. For example, in each study program, females are more likely to pass. However, when considering all students, males are more likely to pass. Another example is that for both young patients and old patients, exercising has a positive effect on one health. However, when looking at all patients, there is a negative correlation between exercising and health.
- *Hacking deep neural networks.* Given a well-trained neural network that is able to separate cars and horses, it is possible to add a bit of 'noise' to the images (invisible to the human eye) such that horses are classified as cars and cars are classified as horses. The same applies to speech recognition.
- *Homomorphic encryption.* It is possible to do computations on encrypted ciphertexts such that the encrypted result, when decrypted, matches the result of the operations as if they had been performed on the non-encrypted data.
- *Secure multi-party computation.* It is possible to jointly compute a function over multiple parties that all keep their data private. Hence, one can apply data science techniques without sharing the actual data.

The above example phenomena and the oversimplified coverage of AI in the media illustrate that policy and decision makers need to know more about data science. This cannot be left to "the market" or solved through half-hearted legislation like the GDPR [8]. To remain competitive, countries should invest in data science capabilities. This can only be realized through education and research.

## 3 Learning Versus Modeling and Programming

Currently, many fields of science are undergoing a paradigm shift. A new generation of scientists emerges that focuses on the analysis and interpretation of data rather than models detached from data. This shift is caused by the availability of data and easy-to-use data-science tooling.

The fields of science can be roughly split into:

- *Formal sciences* (logic, mathematics, statistics, theoretical computer science, etc.) that are based on a priori knowledge or assumptions that are independent of real/life observations.
- *Natural sciences* (physics, chemistry, biology, etc.) that study natural phenomena (atoms, molecules, gravity, magnetism, cells, planets, etc.).
- *Social sciences* (psychology, sociology, economics, literature, etc.) that study human and societal behavior.
- *Applied sciences* (engineering, medicine, software engineering, etc.) that apply scientific knowledge to practical applications (e.g., creating systems).

Natural, social, and applied sciences heavily depend on observations of natural phenomena, people, and systems. In *inductive research*, the goal of a researcher is to infer models and theories (i.e., theoretical concepts and patterns) from observed data. In *deductive research*, the goal of the researcher is to test models and theories using new empirical data. The importance of data has grown in all of these fields. This is a direct result of our ability to observe natural phenomena, people, and systems much more directly.

Consider, for example, the social sciences with research methods such as surveys (i.e., questionnaires), laboratory experiments, field experiments, interviews, and case studies. Traditional surveys may have low response rates and a sample bias (the set of participants that was invited and accepted may not be representative). Laboratory experiments are often too small and also have a sample bias. Interviews and case studies tend to be subjective. Therefore, most scientific results cannot be reproduced. This is commonly referred to as the "replication crisis" [11]. Therefore, younger social science researchers prefer to use research methods that use objective larger-scale observations. For example, directly recording the activities of participants rather than relying on self-reporting or more field experiments with many subjects rather than lab experiments with a only few subjects.

Another example is the uptake of computational biology and bioinformatics where large collections of biological data, such as genetic sequences, cell populations or protein samples are used to make predictions or discover new models and theories.

Also the field of computer science is changing markedly. There seems to be less emphasis on theoretical computer science due to the desire to relate models and theories to real-world phenomena. It is no longer fashionable to create new modeling languages and to prove properties in self-created artificial settings. Instead, sub-disciplines related to data science are rapidly growing in the

number of students and researchers. Automated learning (e.g., machine learning, different forms of mining, and artificial intelligence) are replacing parts of programming. Rules are no longer programmed but learned from data. This is changing computer science. For example, how to verify the correctness a system that uses neural networks?

The shift from modeling and programming to automated learning is affecting science and also the economies that build upon it. Consider for example the way that marketing changed. Today's marketeer is expected to have data science skills. In fact, many professions have become much more data-driven or are about disappear (see next section).

## 4 Machines Versus People

The uptake of data science will continue to change 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 [4]. As a result, many professions will cease to exist [9, 10, 14]. At the same time, new jobs, products, services, and opportunities emerge.

*The frontier between the tasks performed by humans and those performed by machines and algorithms is continuously moving and changing global labor markets.* In [9], Frey and Osborne provide predictions for the computerization of 702 occupations. They estimate that 47 percent of jobs in the US will be replaced by (software) robots.

In [14] three types of roles are identified: stable roles (work that remains), new roles (new types of work that did not exist before), and redundant roles (work that is taken over by e.g. robots). Examples of redundant roles are clerical work (e.g., data entry), factory work (e.g., assembly), postal service, and cashiers. Of the new roles mentioned in [14], most are related to data science.

In [10] three waves of automation are predicted: (1) *algorithmic wave* (replacing simple computational tasks in data-driven sectors such as banking), (2) *augmentation wave* (replacing more complex clerical work and materials handling closed environments such as warehouses), and (3) *autonomous wave* (replacing physical work in transport, construction, and healthcare). The algorithmic wave is currently in full swing. The augmentation wave has started with the uptake of *Robotic Process Automation* (RPA) and robots in production and warehouse. This wave is likely to come to full maturity in the next decade. The autonomous wave uses technologies that are already under development, but, according to [10], will only come to full maturity on an economy-wide scale in the 2030s.

As a concrete example, consider the uptake of *Robotic Process Automation* (RPA) [5]. RPA software provided by vendors such as UIPath, Automation Anywhere, and Blue Prism provides software robots (bots) replacing humans. In the 1990-ties Workflow Management (WFM) software already aimed to realize Straight Through Processing (STP), i.e., handling cases with no or little human involvement. However, in many cases, this was not cost-effective because existing systems needed to be changed. Moreover, WFM often failed because of a limited understanding of the complexity of the actual processes performed

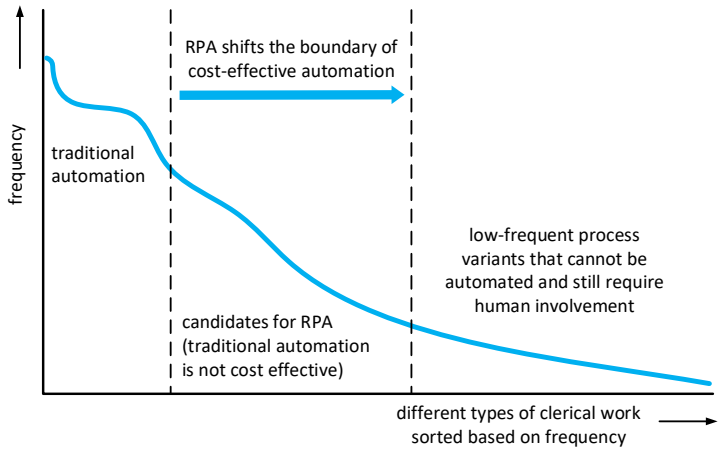


**Fig. 3.** Robotic Process Automation (RPA) has lowered the threshold for process automation [5].

by people. The key difference between RPA and traditional WFM is that RPA does not aim to replace existing (back-end) information systems. Instead, software robots interact with the existing information systems in the same way as humans do. RPA software interacts with the pre-existing applications through the user interface directly replacing humans, i.e., automation is realized by taking over tasks from workers without the back-end systems knowing. A typical RPA scenario is a sequence of copy-and-paste actions normally performed by a human. Compared to WFM, RPA lowers the threshold for automation. As shown in Figure 3, the focus of traditional automation is on high-frequent clerical work. For less frequent work, automation is not cost-effective. RPA shifts the boundary where automation is still cost-effective. Machine learning, artificial intelligence, and process mining are enabling technologies making RPA possible. The software robots need to learn from humans and need to be monitored continuously. Process mining [2] can be used to identify repeating work. Moreover, if a software robot malfunctions due to technical glitches, exceptions, changing user interfaces, or changing contextual factors, then this can be detected using process mining. Machine learning and artificial intelligence can be used to learn specific tasks. In more advanced applications of RPA, work is flexibly distributed over workers and software robots. For example, tasks are initially performed by robots and are escalated to workers the moment there is a complication or exception. Similarly, workers can hand off work to robots using an 'auto-complete' option. Moreover, the RPA solution may adapt due to changes in the underlying process (e.g., concept drift). This illustrates that the border between humans and (software) robots will continue to shift.

Reports such as [9, 10, 14] analyze the impact for specific groups (e.g., based on gender, education, or nationality). Although it is difficult to predict such

phenomena accurately, it is clear that the impact of data science on the work of people is accelerating and will be larger than ever before.

## 5 Responsible Data Science in a Platform Economy

The distribution of work between humans, machines, and algorithms is changing due to the uptake of data science. Moreover, the growing importance of data is also changing the economy and leads to new concerns related to privacy and fairness.

In recent years we have witnessed the rise of the platform economy [6]. The world's most valuable public companies are five American technology firms: Microsoft, Amazon, Apple, Alphabet (Google), and Facebook. These companies are closely followed by Chinese tech giants such as Alibaba, Tencent, Baidu, and Xiaomi, and many more US-based internet companies such as Netflix, eBay, Uber, Salesforce, and Airbnb. These organizations were able to grow extremely fast due to the digital platforms they provide. Some (e.g., Amazon, Alibaba, Airbnb, Uber, and Baidu) provide a transaction platform that matches supply and demand, others provide a technical infrastructure that other people can build upon (e.g., the App stores of Google and Apple), and some provide both (e.g., Amazon also offers cloud services). Successful digital platforms have the characteristic that they tend to grow very fast and, in the end, often one winner remains. For example, Amazon and Alibaba are dominating the way we buy products, Google is controlling the way we search, and Facebook is controlling the way we socialize. Apple, Alphabet, and Microsoft are controlling the platforms we use (iOS, Android, and Windows). After a platform becomes the market leader, it is very difficult to compete for organizations that started later, e.g., for a new company it is difficult (if not impossible) to compete with Google's search engine or with Amazon's marketplace. Large tech companies use profits generated with one platform to create other platforms. See, for example, the current competition to become the leading digital platform for smart homes (e.g., Amazon Alexa, Apple Homekit, Google Assistant, Philips Hue, and Samsung SmartThings). Often "the winner takes it all" due to low marginal transaction costs and so-called network effects [6]. The resulting monopoly may stifle innovation and makes society dependent on a few technology providers. Moreover, the platform providers may use their profits to extend their platforms in other directions. For example, Google is using its profits from search engine marketing to invest in many other services and products (e.g., autonomous driving) and Amazon is using its marketplace to promote its own products (smartphones, televisions, speakers, TV-series, diapers, etc.).

Next to the large-scale economic concerns, there are also smaller-scale concerns impacting individuals. The *Responsible Data Science* (RDS) initiative initiated by the author in 2015, aims to address problems related to *fairness*, *accuracy*, *confidentiality*, and *transparency* [3]. Figure 4 shows the key challenges of RDS:

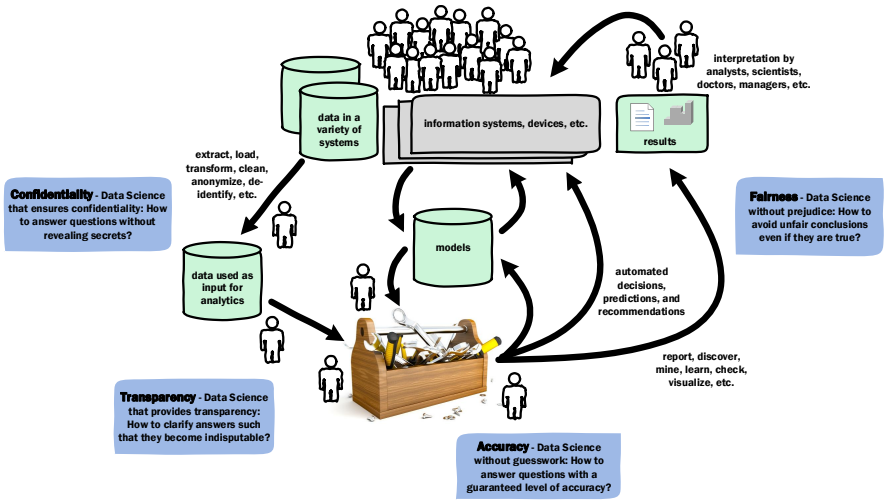


**Fig. 4.** *Fairness*, *Accuracy*, *Confidentiality*, and *Transparency* (FACT) are the main concerns of *Responsible Data Science* (RDS).

- *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?

To sustain the use of data science, it will become increasingly important to address concerns related to fairness, accuracy, confidentiality, and transparency.

Data science techniques need to ensure *fairness*: Automated decisions and insights should not be used to discriminate in ways that are unacceptable from a legal or ethical point of view. This may conflict with the goal of applying data science, e.g., discriminating between students that will be successful in their studies or not, discriminating between customers that will be able to pay back the loan or not, or discriminating between patients that will benefit from treatment or not. These decisions can all be seen as classification problems: The goal is to explain a response variable (e.g., the person will pay back the loan) in terms of predictor variables (e.g., credit history, employment status, age, etc.). Ideally, the learned model explains the response variable in terms of predictor variables. However, these may correlate with sensitive attributes such as gender, nationality, age, etc. As a result, the learned classifier may effectively reject cases from certain groups of persons. This explains the question "How to avoid unfair conclusions even if they are true?".

Data science techniques also need to ensure *accuracy*: Most data science techniques return an answer even when there is not enough evidence in the data. When using many variables relative to the number of instances, classification may result in complex rules overfitting the data [3]. This is often referred to as the curse of dimensionality: As dimensionality increases, the number of combinations grows so fast that the available data become sparse. With a fixed number of instances, the predictive power reduces as the dimensionality increases. Using cross-validation most findings (e.g., classification rules) will get rejected. However, if there are many findings, some may survive cross-validation by sheer luck.

Data science techniques also need to ensure *confidentiality*: The results should not reveal certain types of personal or otherwise sensitive information. The importance of protecting personal data is widely acknowledged and supported by legislation such as the General Data Protection Regulation (GDPR) [7] which states that *"The principles of data protection should apply to any information concerning an identified or identifiable natural person. Personal data which have undergone pseudonymisation, which could be attributed to a natural person by the use of additional information should be considered to be information on an identifiable natural person. To determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly. To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments. The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable."* Organizations that violate the the GDPR can be fined to up to 20 million euro or up to 4 percent of the annual worldwide turnover. The regulations have been criticized for the administrative overhead they generate and the unclear compliance requirements.

There are many techniques to anonymize data, e.g., data masking, generalization, pseudonymization, swapping, perturbation, and sampling. The problem is that also anonymized data may unintentionally reveal sensitive information, e.g., by combining results and background information. There are also more advanced approaches using (homomorphic) encryption or secure multi-party computation. Unfortunately, legislation and policy makers are lagging behind. Therefore, confidentiality concerns may lead to inaccurate, non-optimal, and even unfair decision making [3].

Finally, data science techniques need to ensure *transparency*: It should be clear how the data were processed and that the stakeholders correctly understand the results. Often results highly depend on data cleaning, selection, and parameters of the techniques used. It is easy to change the outcome by altering the analysis-pipeline. Moreover, users that do not understand the phenomena

described in Section 2.4 will have difficulties interpreting data science results accurately. Black-box machine learning techniques such as deep neural nets provide results (e.g., decisions) without explanation. In many application domains this is unacceptable (e.g., sending people to prison, rejecting job applications, or medical decisions). Explainable AI refers to methods and techniques enhancing artificial intelligence such that the results of the solution can be understood by human experts. However, explanations may be very complex and full transparency may not always be possible or required.

The above challenges show that there is a need for data science techniques that are responsible (i.e., "green") by design. This will be one of the main data science challenges of the coming years.

## 6 Conclusion

Data science has become one of the main ingredients of information processing and is changing the way we use and think about IT. Data science includes data extraction, data preparation, data exploration, data transformation, storage and retrieval, computing infrastructures, various types of mining and learning, and the presentation of explanations and predictions. Moreover, given the impact of data science, also ethical, social, legal, and business aspects play a major role. The multidisciplinary nature makes data science particularly challenging.

In this paper, we used the classical four elements ("water", "fire", "wind", and "earth") as metaphors for the essential ingredients of data science. Moreover, we zoomed in on the growing importance of learning from data (rather than modeling or programming) and the transfer of tasks from humans to (software) robots. We also elaborated on the risks associated with data science, e.g., the dominance of digital platforms and concerns related to fairness, accuracy, confidentiality, and transparency.

In 1943, IBM's president, Thomas J Watson, apparently said: "I think there is a world market for about five computers." Up until the 1950-ties many people thought that a handful of computers per country would suffice. This illustrates that, in 1960, the year that International Federation for Information Processing (IFIP) was established, it was impossible to foresee the importance of data science. However, data science has become the main driver of transformation in society and business, and is likely to remain a key topic for IFIP in the next 60 years.

**Acknowledgments**: We thank the Alexander von Humboldt (AvH) Stiftung for supporting our research.

## References

1. W.M.P. van der Aalst. Data Scientist: The Engineer of the Future. In K. Mertins, F. Benaben, R. Poler, and J. Bourrieres, editors, *Proceedings of the I-ESA Conference*, volume 7 of *Enterprise Interoperability*, pages 13–28. Springer-Verlag, Berlin, 2014.

2. W.M.P. van der Aalst. *Process Mining: Data Science in Action*. Springer-Verlag, Berlin, 2016.
3. 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.
4. W.M.P. van der Aalst. Responsible Data Science in a Dynamic World: The Four Essential Elements of Data Science. In L. Strous and V.G. Cerf, editors, *Internet of Things - Information Processing in an Increasingly Connected World*, volume 548 of *IFIP Advances in Information and Communication Technology*, pages 3–10. Springer-Verlag, Berlin, 2018.
5. W.M.P. van der Aalst, M. Bichler, and A. Heinzl. Robotic Process Automation. *Business and Information Systems Engineering*, 60(4):269–272, 2018.
6. W.M.P. van der Aalst, O. Hinz, and C. Weinhardt. Big Digital Platforms - Growth, Impact, and Challenges. *Business and Information Systems Engineering*, 61(6):645–648, 2019.
7. Council of the European Union. General Data Protection Regulation (GDPR). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC, April 2016.
8. 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.
9. 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.
10. 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.
11. J.P.A. Ioannidis. Why Most Published Research Findings Are False. *PLoS Medicine*, 2:e124, 2005.
12. D. Laney. 3D Data Management: Controlling Data Volume, Velocity, and Variety (Research Note 949). Technical report, META Group, February 2001.
13. Y. LeCun, Y. Bengio, and G. Hinton. Deep Learning. *Nature*, 521:436–444, 2015.
14. T.A. Leopold, V. Ratcheva, and S. Zahidi. The Future of Jobs Report. Technical report, Centre for the New Economy and Society, World Economic Forum, 2018.
15. P. Naur. *Concise Survey of Computer Methods*. Studentlitteratur Lund, Akademisk Forlag, Kobenhaven, 1974.
16. C. O'Neil. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing Group, New York, 2016.
17. D.E. Rumelhart, G. Hinton, and R.J. Williams. Learning Representations by Back-Propagating Errors. *Nature*, 323:533–536, 1986.
18. L. Strous and V.G. Cerf, editors. *Internet of Things - Information Processing in an Increasingly Connected World*, volume 548 of *IFIP Advances in Information and Communication Technology*. Springer-Verlag, Berlin, 2018.