Alisa Bilal Zorić
Pollytecnic Baltazar Zaprešić, Croatia
abilal@bak.hr

Data science: fundamental principles

Abstract

We live in a world where we collect huge amounts of data. Traditional methods and techniques are no longer sufficient to process them. In addition to the sophisticated development of computers, new ways of processing data are evolving. Data Science is a new emerging multidisciplinary field that combines classical disciplines like statistics and mathematics with computer science. The main goal of Data Science is to turn large sets of both unstructured and structured data into useful information that can help organisations to make powerful data-driven decisions. At a high level, data science can be described as a set of fundamental principles necessary for successful extraction of information from data. Since we collect data all the time and about anything, its application is diverse. The most common application is in healthcare, travel, e - commerce, sports, government, social media, etc. The goal of this paper is to introduce data science and to present its benefits and application in various fields.

Keywords: data science, data analytics, big data

1 Introduction

The amount of stored data increases rapidly, and today, the main problem is not how to collect data, but how to extract useful information from them. There are many powerful tools for data scientist that can help them in this process, but in order to use them wisely, data scientists must have much pre-knowledge from statistics, math and computer sciences, and they also need to be able to see business problems from a data perspective. There are a different definitions of data science ("data science is a set of fundamental principles that support and guide the principled extraction of information and knowledge from data" (Provost and Fawcett, 2013), "data science is a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data" (Dhar, 2013), "data science is an amalgamation of classical disciplines like statistics, data mining, databases, and distributed systems"(Van der Aalst, 2016)), but they all have in common that Data Science is a multidisciplinary field that deals with data processing (both unstructured and structured data), analysis, and extraction of useful information from data using various statistical methods and computer algorithms. The goals of the paper are: (i) to present a short summary of the history and definition of data science; (ii) to elaborate similarities and differences between Business Intelligence and Data Science, (iii) to overview the life cycle of data science, and (iv) to outline the benefits and various applications of data.

2 History of data science

In recent years, especially with growth of the Internet and cheaper IT equipment, the amount of data that is available has grown tremendously. We start to collect data from various sources like web server logs, online transaction records, tweet streams, social media, data from all kind of sensors and enormous datasets present computational problems. Now, the problem is not finding the data, but figuring out how to use it effectively. Traditional statistics was not enough because it required a lot of time. New discipline called Data Science that combines the knowledge of traditional statistics with computer science, has emerged.

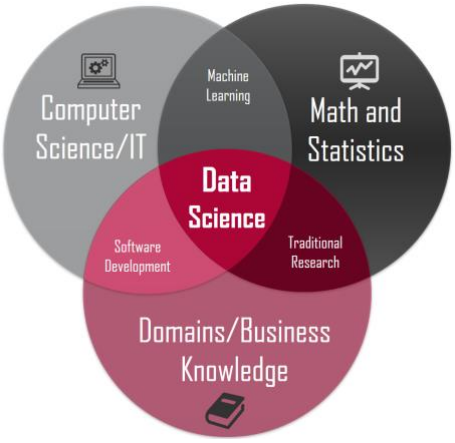

Figure 1: Data science
(https://towardsdatascience.com/introduction-to-statistics-e9d72d818745)

In 1977 the International Association for Statistical Computing was founded and their mission was to integrate traditional statistical methodology, modern computer technology and domain experts' knowledge to transform data into information. In Mining Data for Nuggets of Knowledge (Zahavi, 1999), author pointed out that traditional statistical methods work well with small data sets, but to handle the massive amounts of data, new tools have to be developed. Data Science Journal began to publish in 2002 and in 2008 "Data Scientist" became a buzzword, mostly by DJ Patil and Jeff Hammerbacher (Foote, 2016). In 2011, new concept of Data Lakes was introduced. Data Lakes store information using a non-relational database (NoSQL) and does not pre-categorize the data like Data Warehouse. Data science works much more with data lakes than with data warehouse.

In the last decade, data science has become an important part of business and academic research. It has spread to all areas of human activity and now influences economics, healthcare, governments, genetics, business, finance, social sciences, medical informatics, etc.(Provost and Fawcett, 2013).

In the book *Data Science for Business: What you need to know about data mining and data-analytic thinking,* authors outline several fundamental concepts of data science. These concepts arose from different fields that study data analytics and refer to the links between data science and the business problems or technical solutions. First fundamental concept refers to the systematic approach and importance of each step in the process of extracting useful knowledge from data. The second concept highlights the importance of the context in which the results of the analysis will be used. One concept emphasizes the importance of decomposing the initial problem into smaller units so that they can be easier to solve using known data mining tools. The next concept points out the importance of information technology in finding unknown patterns and correlations. The last concept relates to the problem of overfitting a dataset. Overfitting is a modelling error, which occurs when a model is too closely fit to a limited set of data. It works perfectly on its training data, and poorly on a new, unseen data. This concept is most important to understand when applying data-mining tools to real problems (Provost and Fawcett, 2013).

3 Business Intelligence (BI) and Data Science

In this section, we will explain similarities and differences between Business Intelligence and Data Science, since these terms are often mixed or used as synonyms. They both work with large amount of data, and they both have the capability for interpreting data into useful information for better decision making, but the approaches are different. While data has

become bigger and more complex, the traditional BI platforms have become inadequate to handle such data. The major difference between Business Intelligence and Data Science is in data. Business Intelligence works with highly structured data, while Data Science can work with all types of data, structured, semi-structured and unstructured data gathered from different sources.

Table 1: Comparison of business intelligence and data science
(Own resource)

Business intelligence	**Data science**
Analyses past data	Past data is analysed for future predictions
Works with static structured data	Works with dynamic unstructured data
Statistics and Visualization are the two skills required for business intelligence.	Statistics, Visualization, and Machine learning are the required skills for data science.
Analytical method	Scientific method
Data stored in a Data Warehouse	Data stored in a Data Lakes

Business Intelligence analyses previous data to find insight to describe current state and business trends. It helps interpret historical data and it is mainly used for reporting. Data Science can also analyse the past data but in order to make future predictions, and it is mostly used for Predictive Analytics or Prescriptive Analytics.

Merritt-Holmes (2016) highlights that BI systems are designed to analyse real events based on real data, while Data Science is focused on a future. BI make detailed reports to understand the current trends, it does not interpret the information to predict what might happen.

We can conclude that Business Intelligence is a part of Data Science. While Business Intelligence is limited to the area of business operations, generating dashboards and reports based on the internal structured data, Data Science focuses on generating insights out of the all kinds of data (https://data-flair.training/blogs/what-is-data-science/). Data Science uses a wide variety of complex statistical algorithms and predictive models and is much more complex compared with Business Intelligence. Result of complex predictive analytics is a data model used for future predictions and forecast growth of the business.

4 Data science life cycle

Data Science Life Cycle consists of five main phases. First, there is Data Discovery, which involves retrieving data from various sources. This data can be in structured or unstructured format. In structured format, data is organized into a formatted repository, usually a database, and this type of data is much easier to analyse. Unstructured format of data is format that is not uniform and standardised. The most common type of unstructured data is text from documents, presentations, transcripts, blogs, social media sites, etc. Other types of unstructured data are video files, images, audio files, sensor data, etc. In this phase, initial hypotheses to test are formulated.

Second phase is Data Preparation in which collected data is transformed into a prespecified format. It includes complex methods for data cleaning, data reduction and data transformation. Afther that, there is a Model Building in which various mathematical models are build and applied to generate a satisfactory result. After measuring the models, parameters are calibrated to optimal values. In this phase various statistical formula and visualization tools are used to understand the relations between variables and various

learning techniques (clustering, classification, association, etc.) are analysed to build the model. Once the data is prepared and the models are built, fourth phase called Operationalize begins. In this phase, information is gathered and outcomes are obtained based on initial requirements. In the last phase, results and findings are presented to decision makers.

5 Application of data science

Application of data science is numerous. The most common use is in finance, genetics, banking, medicine, business and transportation for problems like financial trading, credit scoring, fraud detection, online advertising, direct marketing, internet search, recommendations for cross-selling, etc. Many companies have focused their business on data. They use data to find hidden patterns that will help them find appropriate solutions and improve decision-making process. This can help organizations understand their customers, markets, and the business as a whole by anticipating growth, trends and business insights based on huge amounts of data (https://intellipaat.com/blog/what-is-data-science/). Principles and techniques of data science are also applied to general customer relationship management to analyse customer behavior in order to reduce attrition and to increase expected customer value. In the finance industry, data science is used for credit scoring, fraud detection and trading. In the healthcare, classification algorithms can be used to detect cancer and tumors at an early stage using Image Recognition software. In Genetic Industries, data science is used for analysing and classifying patterns of genomic sequences. Using Machine Learning, Data Scientists have developed recommendation systems that recommend different products to customers based on their historical habits. In manufacturing, industrial robots use Data Science technologies such as Reinforcement Learning and Image Recognition to take over repetitive jobs. In transport, Self-Driving Cars are developing based on Reinforcement Learning and Detection algorithms. Another application of data science is in conversational agents (Siri by Apple). They use Speech Recognition system to understand users, to convert human speech into textual data and to provide an appropriate response.

6 Conclusion

With the enormous increase in data, there is a constant need for analysing such a large amount of data. Data Science can manage this data and develop beneficial machine learning models that predict future results.

We can conclude that Data Science is emerging multidisciplinary field with roots in mathematics, statistics, and computer science. As it engages in extracting, analysing, visualizing, managing and storing large amounts of data, it has a very wide range of application from business and finance to healthcare and transportation. The main goal of Data Scientists is to recognize and use meaningful insights from data in order to help organisations in taking smarter decisions. During that process, they use different tools and methods to identify redundant patterns and hidden knowledge within the data. They also use the most powerful hardware, most efficient algorithms and programming systems to solve the data related problems.

In this paper, we wanted to introduce data science as a new, powerful field with various applications that can provide a competitive advantage and long-term stability.

7 Bibliography

1. Barber, M. (2018). *Data science concepts you need to know! Part 1*. Retrieved 15. 9. 2019 from https://towardsdatascience.com/introduction-to-statistics-e9d72d818745.

2. Dataflair Team. (2019). What is data science?: a complete data science tutorial for beginners [Blog]. Retrieved 8. 10. 2019 from https://data-flair.training/blogs/what-is-data-science/.

3. Dhar, V. (2013). Data science and prediction. *Communications of the ACM*, 56(12), 64–73.

4. Foote, K. D. (2016). *A brief history of data science*. Retrieved 17. 10. 2019 from https://www.dataversity.net/brief-history-data-science/#.

5. Grossmann, W. and Rinderle-Ma, S. (2015). *Fundamentals of business intelligence*. Berlin; Heidelberg: Springer.

6. Merritt-Holmes, M. (2016). *10 differences between data science and business intelligence*. Retrieved 15. 10. 2019 from https://www.itproportal.com/2016/08/18/10-differences-between-data-science-and-business-intelligence/.

7. Provost, F. and Fawcett, T. (2013). *Data science for business: what you need to know about data mining and data-analytic thinking* (1 st ed.). Sebastopol: O'Reilly.

8. Provost, F. and Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. *Big data*, 1(1), 51–59.

9. Van der Aalst, W. (2016). Data science in action. In W. van der Aalst, Process mining (pp. 3–23). Berlin; Heidelberg: Springer.

10. What is data science? [Blog]. (2019). Retrieved 12. 10. 2019 from https://intellipaat.com/blog/what-is-data-science/.

11. Zahavi, J. (1999). *Mining data for nuggets of knowledge.* Retrieved 7. 10. 2019 from https://knowledge.wharton.upenn.edu/article/mining-data-for-nuggets-of-knowledge/.