Key Aspects of Data Science, Methodologies, Issues, Challenges and its Applications

Richard Luke, Data Scientist, Mzuzu University.

Abstract

This paper gives a summary from three articles in the bibliography on data science. Through the introduction, research methods, data science projects can be improved through a formalized and improve research process. Therefore, this paper focuses on introducing the topics of data science, big data, methodologies of conducting a data science project, challenges and applications. The paper will be interesting to Data Analysts, Data Scientists and Machine Learning Engineers as an introduction.

Index Terms

data science, big data, methodologies

NOMCLATURE

Machine Learning (ML), Return on Investment (RIO), Cross-Industry Standard Process for Data Mining (CRISPDM), Microsoft Team Data Science Process (TDSP)

1. Introduction

In the article, *Data Science: A Comprehensive Overview*, Longbing Cao explores some of the key aspects of the new field of Data Science namely the age of big data, advanced analytics, and Data Science.

He argues that Data Science is a new interdisciplinary field that builds on statistics, informatics, computing, communication, management, and sociology. He highlights topics like datafication and quantification, data initiative by governments, scientific agendas, disciplinary developments, data economy and industry transformation, data professional community, and open model and community.

Mukul Varshney *et al.* defines data science as dealing with large quality of data for the purpose of extracting meaningful and logical results/conclusions/patterns. He reports that Data Science is an emerging area of work concerned with the collection, preparation, analysis, visualization, management, and preservation of large collections of information. Through Data Science, better analysis of the large volumes of data that are becoming available, there is the potential for making faster advances in many scientific disciplines.

Discussing possible shortcomings in Data Science, the authors mention Heterogeneity and Incompleteness, Scale, Timeliness and Privacy of data.

In the article *Data Science Methodologies: Current Challenges and Future Approaches* Inigo Martinez, Elisabeth Viles, Igor G Olaizolaa explores the necessity of developing a more holistic approach for carrying out Data Science projects.

They propose a conceptual framework for managing Data Science projects. According to them a conceptual framework for designing integral methodologies for the management of Data Science projects is built upon a critical review of currently available Data Science methodologies, from which a taxonomy of methodologies can be developed.

2. Aspects of Data Science

2.1. Big Data

This study deals with treatments of large or complex data sets in terms of systematic analysis to extract information from them. It is characterized by volume, variety and velocity

2.2. Advanced Analytics

Data analytics deals with treatments of processed data to provide insights and trends about the future. It is the opposite of analysis which is processing data to get historical insights and trends.

2.3. Data Science

Cao defines data science as the science of dealing with data. Data Science as a multidisciplinary field that lies between computer science, mathematics and statistics, and comprises the use of scientific methods and techniques, to extract knowledge and value from large amounts of structured and/or unstructured data.

3. Methodologies for Data Science

3.1. Ideation and implementation

This is the process of infusing data science into design thinking processes to improve business processes and increase its value. Here the goal is to transform data into information that yields insights or can lead to an impact. Therefore, in this phase business objectives are identified and criteria outlined.

3.2. Project management

Having an effective project management methodology is fundamental for the success of a data science project since the team can monitor and set up goals, stages and outcomes of the project. However, alone it cannot help with the success of a data science project. A methodology that highlights data centric needs of data science while keeping the application focused uses of the models and other artifacts produces is preferred.

3.3. Team management

This entails coordination, collaboration and communication. Most data science projects are completed by teams of a variety of specialized skillsets thus the demand for effective team management. Effective team management demands defining the data science project team by highlighting an intermediary understanding of both language of analytics, domain knowledge thus enabling teams to reduce the gap between team members.

3.4. Data and information management

This is a program that aims to manage teams of people, data, processes and underlying technologies to deliver insights on business intelligence processes. Efficiency here would improve producibility, knowledge management, improved ML, data validations, quality assurance checks, data security and privacy and RIO on IT investments.

Thus, using the taxonomy above, Inigo Martinez, Elisabeth Viles, Igor G Olaizolaa proposes a data science methodology that includes a mixture of statistics, variety of disciplinary knowledge in computing, management and decision making. In this regard, any of the following methodologies would suffice provided they are inclusive: Cross-Industry Standard Process for Data Mining (CRISPDM), Microsoft Team Data Science Process (TDSP), Domino Data Science Lifecyle, RAMSYS, MIDST, Development Workflows for Data Scientists,

Big data ideation, assessment and implementation, Big Data Management Canvas, Agile Delivery Framework, Systematic Research on Big Data, Big Data Managing Framework, The Data Science Edge, Foundational Methodology for Data Science, Analytics Canvas, AI Ops, Data Science Workflow and EMC Data Analytics Lifecycle

4. Issues, Challenges and Applications of Data Science

4.1 Issues

Shortcoming in data science includes heterogeneity and incompleteness, Scale, Timeliness and Privacy of data. To which Cao proposes the following solutions: problem formulation, data step, modelling step and application which can be simplified as business understanding, data acquisition and understanding, and deployment and modelling.

4.2 Challenges

The main challenges are a cross section on the methodologies. For example, team management challenges may include poor coordination, lack of people with analytical skills and lack of transport communication; Project management challenges may include uncertain business objectives, unrealistic expectations and delivery the wrong thing; and Data and information management may include lack of reproducibility, no data validation and data and security challenges.

4.3 Applications

Since Data Science develops from real world application rather than research only therefore its applications include business analytics, prediction, computer vision, natural processing language, bioinformatics, science and research, revenue management, government etc. “The future will be crowded with people trying to apply Data Science in all problems. But it can be sensed that we are going to see some real amazing applications of Data Science for a normal user. The skills needed for visualization, for client engagement, for engineering saleable algorithms are all quite different,” the authors noted.

5. Conclusion

This summary is intended to function as a preliminary point to the data science framework for performing data science projects. The proposed Data Science process follows an iterative structure which has project management, team management and data & information management. Therefore, a good data science project must have some of the following skills: analytical thinking, be methodological, understand statistics, have qualifications like degrees, a background in software engineering, theoretical and domain knowledge. Thus, achieving real world applications is the goal hence the future of data science is bound to include people applying data science to all problems thus the birth of amazing application of data science.

BIBLIOGRAPHY

[1] E. V. I. G. O. Iñigo Martineza, "Data Science Methodologies: Current Challenges and Future Approaches," *Science Direct,* vol. 1, no. 07287, June 15, 2021.

[2] L. CAO, "Data Science: A Comprehensive Overview," *ACM Journal,* vol. 50, no. 3, p. 43, 2017.

[3] S. G. J. ,. A. K. R. Mukul Varshney, "A Study on Issues, Challenges and Application in Data Science," *International Journal of Trend in Scientific Research and Development* , vol. 1, no. 5, pp. 256-533, 2017.