DATA SCIENCE: CHALLENGES AND TRENDS

K. Rasheva-Yordanova[1], S. Toleva-Stoimenova[1], D. Christozov[2]

[1]University of Library Studies and Information Technologies (BULGARIA)
[2]American University in Bulgaria (BULGARIA)

Abstract

With the vast amounts of data available in the world, most companies today are focusing on data usage to identify the strengths, weaknesses, and opportunities in business. An area called "Data Science", closely related to the Data Mining concept. "Data science" is a term that has been in the public perception and imagination only since the first half of the decade, but today it is extremely popular in science, practice, and education. It includes tools, methods, and systems applied to Big Data to leverage knowledge for decision-making. Based on the literature review and initial findings, this research study found that there are differences in the understanding of the term "Data Science" by academics and practitioners. This article aims to outline, on the one hand, a comprehensive framework of Data Science competences and on the other to sum up the inter- and multi-disciplinarity characteristics of the area. In result, a review of the existing definitions and fundamental concept of Data Science is presented. Explored identified directions for development of Data Science competency are discussed.

Keywords: Data Science, Big Data, Data Scientist, Data Science Competences.

1 INTRODUCTION

In the information era “Data Science“, “Big Fata“ or “Information Society” are often used to define important aspects related to the large amount of data and information that are available today. The "data-information-knowledge" sequence characterizes the process of data exploitation. Learning from data becomes the necessary condition for effective behavior.

Organizations today have vast amounts of stored structured and unstructured data from which expect to extract valuable knowledge in a way to enhance their competitiveness. A new field of science and practice is emerging to respond the Big Data challenge; and a new profession – Data Scientist emerged to fulfill the needs for analysts capable to support data driven decision-making processes.

Tukey [11] has made a major contribution to the development of the field of data processing and analysis, defining for the first time the term exploratory data analysis (EDA). In recent years, data analysis has evolved and new scientific fields such as Data Analysis, Data Analytics, Advanced Analytics, Big Data Analytics, Deep Analytics have emerged. Other areas have been established such as Data Mining, Knowledge Discovery and Machine Learning, all of which related to the intelligent data analysis and pattern discovery.

With the advancement of ICT and the computing power of computers, data management, data storage, retrieval and recovery mechanisms, query execution and transaction processing and advanced data analysis have evolved and improved. Integration of all those scientific areas as a new trans-discipline – Data Science – can be seen as the result of natural evolution of all areas linked to data management and exploitation.

"Data Science" is a term appeared in public perception since the first half of current decade, but today it gains popularity in science, practice, and education. The area identified as “Data Science” today includes tools, methods, and systems applied to manage Big Data in a way to leverage knowledge supporting decision-making.

The objective of this paper is to outline the competences associated with the profession of a Data Scientist in a comprehensive framework by summing up the inter- and multi-disciplinarity of the area. The paper is organized in three sections. Section one contrasting the understanding of the term "Data Science" by academics and practitioners, and propose the framework of competences as union of expertise industry is looking for with competences assigned to Data Science in literature. Section two presents Data Science as an inter- and multi-disciplinary field. In section three the available opportunities for competency development in the Data Science area are discussed.

2 DATA SCIENCE DEFINITIONS

According to recent research by Yan and Davis [30], the term “Data Science“ was coined for the first time by Wu [5] during a lecture delivered at Michigan University in 1997. In this lecture, as well as in a lecture from 1998 [6], Wu uses the term Data Science in order to call statistics in a modern way. According to Yan [30], the author of Cleveland [29] “in his publication of 2001, outlines a plan for a ‘new’ discipline, wider than statistics, which he called Data Science, but he did not refer to the term coined by Wu“.

However, a review of the literature shows that for the first time the definition of Data Science appears in the book Concise Survey of Computer Methods [20], where the author defines it as “Data Science, once they Despite the lack of consensus on the origin of the term, it is clear that the importance of data is fully recognized by the academic community and is one of the factors that have led to the emergence of this relatively new scientific field.”[21].

In practice, Data Science has become popular over the last decade with the development of large Internet corporations such as Yahoo, Google, LinkedIn, Facebook and Amazon, and is seen as “potentially one of the most significant advances of the early 21st century“ [17].

Despite the popularity of Big Data and Data Science, not only in research circles but also in public perception, there is currently no consensus on the definition of Data Science. While Wu uses the term "Data Science" to give a modern flavor to traditional statistics, today, most researchers in the field view Data Science as a much broader concept.

The following table provides a chronological representation of the various definitions of Data Science.

Table 1. Data Science definitions and interpretations

Author	*Definition*
Provost & Fawcett, 2013	Includes a number of basic principles that support and manage the extraction of information and knowledge from data [9].
Dhar, 2013	An area that focuses on the collection, retrieval and transformation of data into knowledge [27].
Shum et al., 2013	Data Science is a combination of statistics, computer sciences and information design [28].
Andrejevic, 2014	Offers new ways to use the data needed to make forecasts and make decisions that affect all sectors - from health care to urban planning, financial planning, job screening and admission in education, etc[16]..
Diggle, 2015	Includes not only statistics but also computer science (hardware and software engineering) [19].
Parks, 2017	A combination of techniques, knowledge and skills applied to data to identify hidden knowledge that can be used to guide teams in making important decisions [8].
Cao, 2017	A new interdisciplinary field that includes statistics, computer science, computing, communication, management and sociology to transform data into insights and decisions, following the notion of the data-knowledge-wisdom hierarchy [15].
Turkay et al., 2018	A process of multi-stage knowledge discovery in which useful knowledge is extracted from a raw, often impure collection in a specific context [4].
Donghui & Davis, 2019	Provides principles, methodology, and guidance for data analysis for: (1) an instrument (visualization, data collection, or research tools), (2) value (commercial or scientific), or (3) knowledge (hidden objective practical useful interdependencies) [30].

In search of a uniform definition of the Data Science concept, we found that at least one of the following phrases is present in each of the proposed definitions:

- “includes a combination of statistics, mathematics and programming
- “unites methods, means and technologies“
- “aiming to retrieve meaningful information, knowledge and insights from data”

This makes us suggest a common interpretation of the term, whose subject matter, object and results have been presented in Figure 1.

- The object of Data Science is the available Big Data.

- The subject matter of Data Science is: (1) informed decision-making (by business managers) and (2) developing Data Science competence (by Data Science experts).
- Sought result – as a result of the applied methods, techniques, principles and available resources for working with data is the formation of knowledge and insight from the available data.

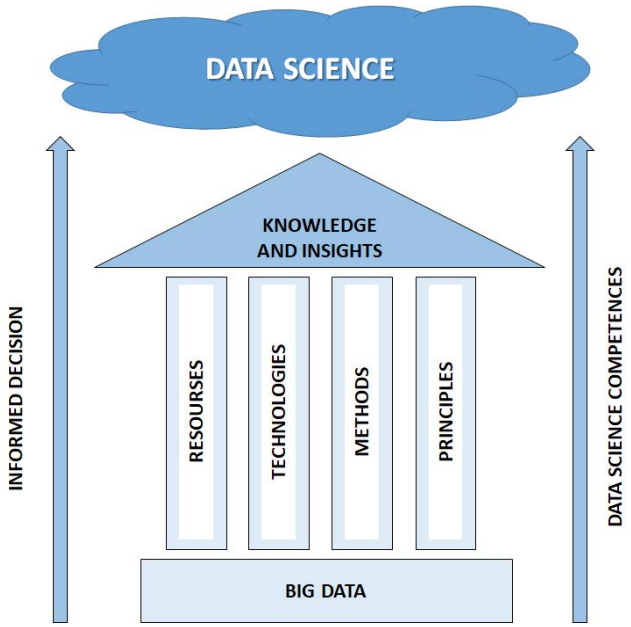

Figure 1. A summary framework of Data Science

In order for insight to be achieved, it is necessary that data experts should have a wide spectrum of knowledge, skills and competence, often uniting various scientific and applied fields. This, in turn, leads us to believe that Data Science is a multi and interdisciplinary field. A review of the literature on this claim will be made in the next section.

The expertise the industry is looking for was investigated by reviewing the requirements of job-offering in “man-power” internet sites as well as by informal interviewing practitioners. In a nut-shell industry is looking for three categories of competences – hard, soft and analytical skills. Job offers address mostly hard skills as mathematics, statistics, and algorithms, which is easier for assessment by reviewing candidates CVs, but soft and analytical skills are defined as critical for selection during the interviews.

3 INTER- AND MULTIDISCIPLINARY OF DATA SCIENCE

Due to the widespread use of ICT and the development of society, most sciences nowadays have to step outside their borders and use new methods and technologies that enable data research and processing. Inter-, multi- and trans-disciplinary studies are becoming ordinary.

It is now common practice for multidisciplinary teams of experts from two or more disciplines to work together to solve a complex problem. Each participant uses theories, methods and techniques from his or her own field and holistic solution is looking as a synergy of applying the approach combining the most relevant from different fields. In this way, multiple perspectives and a variety of tools are included, which contributes to the effectiveness of the work.

Literally, "inter" means inward orientation, interaction, and "multi" means multitude. Therefore, multidisciplinary methods are from different scientific fields which can be combined to retrieve the most useful information in a given scientific problem. In the process of applying specific methods to different scientific fields, they are implicated in multidisciplinary methods. “Trans” refers to building something new on top of multiple backgrounds, creating bridges between disciplines.

Interdisciplinarity refers to the ability to combine several disciplines to extend the benefits of each one. This opens up the possibility of exploring a problem from the perspective of a multitude of scientific

perspectives, so as not only to direct and develop interest, but also to stimulate activity, to present the value orientation of the problem, the result and the application. Interdisciplinarity as an approach is very often associated with interactive learning. The interdisciplinarity of Data Science is determined by the fact that it cannot fit into the traditional classifications of a single discipline, but covers fields such as statistics, computer science, computer technology, communication, management, sociology, and others that provide their criteria to build an emerging new specifics.

Parks [8] examines the interdisciplinarity of Data Science, in terms of the professional profile of data professionals and the "interdisciplinary knowledge" they must possess. According to the author, knowledge and skills in disciplines such as accounting, economics and information systems belong to this category, which allows adequate decision-making in response to specific problems. Bateman [22], considering the interdisciplinary nature of Data Science presents the so-called 'cloud of labels' illustrating the diversity of (sub)disciplines that Data Science encompasses.

The term “transdiscipline” was first coined by Piaget [10] as a „higher stage of succeeding interdisciplinary relationships…which would not cover interactions or reciprocities between specialized research projects, but would place these relationships within a total system without any firm boundaries between disciplines” While multidisciplinary and interdisciplinary teams focus on knowledge transfer across disciplines, transdisciplinary teams use a set of holistic practices that aim to focus on the subject matter of the study, without underestimating the different fields, but without approaching the study from a centralist point of view. Maasen and Lieven [24] characterize transdisciplinary teams as "broadening expertise" and "participatory legitimation" rather than "knowledge legitimation".

The three concepts are similar in common idea of unity, mutual relations, integration of scientific disciplines, transfer of methods among the sciences, etc., which they share. They all relate to the importance of not centralizing thought in one science, but integrating several sciences into research.

A review of the literature revealed differences in the views of different authors regarding the interdisciplinary knowledge that a data specialist should possess. For example:

- Conway [7] claims that Data Science is a field which necessitates hacking skills, knowledge in in Math and Statistics, as well as machine learning as an interception point.
- Tierney [2] looks at Data Science as a common point between statistics, visualisations, pattern recognition, neurocomputing, AI, data mining, databases and data processing.
- Matter [26] presents Data Science as a combination of quantitative methods, computer science and social sciences.
- Ayankoya [12] includes in the building components of Data Science fields like domain knowledge, statistics and analytical techniques, BI, programming and data visualisation.
- van der Aalst [18], 2014 unites Data Science fields like statistics and stochastics, system design, industrial engineering, behavior/social sciences, domain knowledge, visualization, distributed computing, process mining.
- Considering the interdisciplinary nature of Data Science, Eubanks [3] includes areas like hacking and coding, statistics, machine learning, substantive expertise (marketing), data engineer and traditional research.

In a previous study, we presented a competency model based on three types of skills (Figure 2): hard skills, soft skills and analytical skills [13] which we expanded into another group of skills – ethical skills [25].

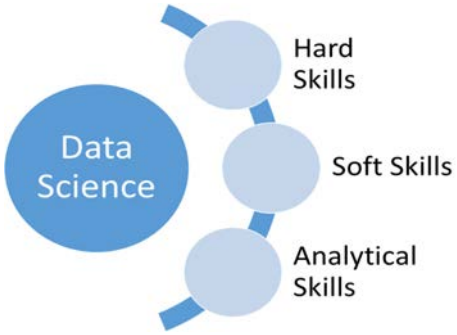

Figure 2. Skills encompassing Data Science competence

The specifics of Big Data and the interdisciplinary nature of Data Science pre-define the existence of challenges that have been explored and analyzed in the next section of the paper.

4 DATA SCIENCE CHALLENGES

Still, the scope, scale and range of Big Data activities are in the early stages of development and Data Scientists face a number of challenges. In the scientific literature, the challenges to Big Data and Data Science are addressed in two aspects: (1) according to the specifics of the tasks performed and the results sought, and (2) the technology, algorithms, methods and analysis techniques.

According to Abualkishik [1] challenges in the field of Big Data can be referred to the Big Data „V's". Handling the rapid growth of data size is problematic and requires efficient scalable storage system, data file formats for several types of data, i.e. columnar data files, compression, and duplication. On the other hand, the velocity of data whereas generating insights from the stored data is needed in a timely manner and efficient cost via new generation ETL and analytical tools.

Wadhwani [14] complements the list of challenges when working with Big Data adding velocity and veracity, data quality, data availability, data discovery, data quality, data extensiveness, personally recognizable information, data assertiveness, quantifiability and data processing to data management, volume, variety and combining multiple data sets. Donahole [23] looks at the real-time challenges of Data Science which are based on perspectives from those experts in the field. She focuses on 5 basic challenges:

Problem identification – One of the major concern in analyzing a problem is to identify it accurately for designing a better solution and defining each aspect of it.

Accessing the right data – The rule is the "right kind of data for the right analysis". It is necessary that Data Scientist receive data in the most appropriate format. Issues, such as hidden data, or insufficient volume and variety of data, are possible and frequent. A challenge can be also the legality of given business to access data.

Data cleansing Vs. Data quality – "Dirty data" create problems with operating costs. Difficulties are noticeable when working with databases that are full of discrepancies and anomalies. In practice, "anomalies as unwanted data leads to unwanted results". Data specialists work with huge amounts of data and spend a lot of time in "sanitizing the data before analyzing". Quality of data is the issue of catching, gathering and recording and analysts cannot influence it.

Lack of professionals – Data specialists are important to bridge the gap between IT department and top management as domain expertise is required for conveying the needs of the business to the IT department. To resolve this, data scientists need to get more insights that are useful from businesses in order to understand the problem and work accordingly by modeling the solutions. They also need to focus on the requirement of the businesses by mastering statistical and technical tools".

Competence and expertise development – The data expert profession requires the application of results using sophisticated data and practical applications. However, a career in Data Science is not only based on momentary expertise, it is based on the response of experts to the needs of industries. There is a need for continuous improvement and upgrading of skills. The knowledge needed today may not be enough tomorrow.

We have limited the list of challenges of Data Science in Big Data context to the following:

Data challenges relate to the characteristics of the data itself (e.g. data volume, variety, velocity, veracity, volatility, quality, discovery and dogmatism).

Technological Challenges relate to technical tools for processing, storage and transfer of large data sets.

Data professionals challenges relate to the competence and expertise of Data Scientists. Data specialists are concerned with analyzing data and extracting hidden meaningful patterns, internal relationships and dependencies that cannot be directly established from the accumulated data. They face many new opportunities and challenges that provide them with Big Data, available methods and technologies for storing, processing and analyzing them.

Data security challenges concerns some ethical aspects of Big Data analysis that affect the privacy and legal issues surrounding copyright and data ownership. Often, users of data services and devices are not informed that they will be included in the large amount of data and/or what that data may subsequently be used for.

With regard to the production and implementation of data specialists, we have included three additional challenges in the field of Data Science:

Educational challenges for training organizations building data specialists - universities sould respond to current labor market needs and offer adequate curricula to enable graduates to pursue the profession as soon as they graduate. This, in turn, requires the involvement of trainers with experience and expertise to meet existing needs. The natural questions that arise are (1) what knowledge, skills and competencies are needed to extract useful value from the accumulated data and (2) what is the frequency that should be followed in auditing and updating the existing curriculum?

Challenges for Big Data business organizations and institutions relate, on the one hand, to the discovery and appointment of a Data Scientist who has the knowledge and experience to handle the data and, on the other, to the storage of the accumulated Big Data asset. The second challenge requires (1) care for the integrity and security of the data and (2) additional resources (technical and financial) that will allow the data to be stored and subsequently analyzed.

Challenges facing Data Scientists - as it has already become clear, upgrading and improving Data Science skills is a prerequisite for a successful data specialist

In this regard, specific dependencies can be identified that affect the pace of development of Data Science:

- lack of Data Scientists with the necessary experience and expertise directly affects the business
- lack of good trainers will affect the quality of Data Science educational programs
- the Data Scientist profession requires a continuous upgrade of skills, which affects the range of competences building the profile of a data specialist on the one hand and the pace of development of Data Science on the other.

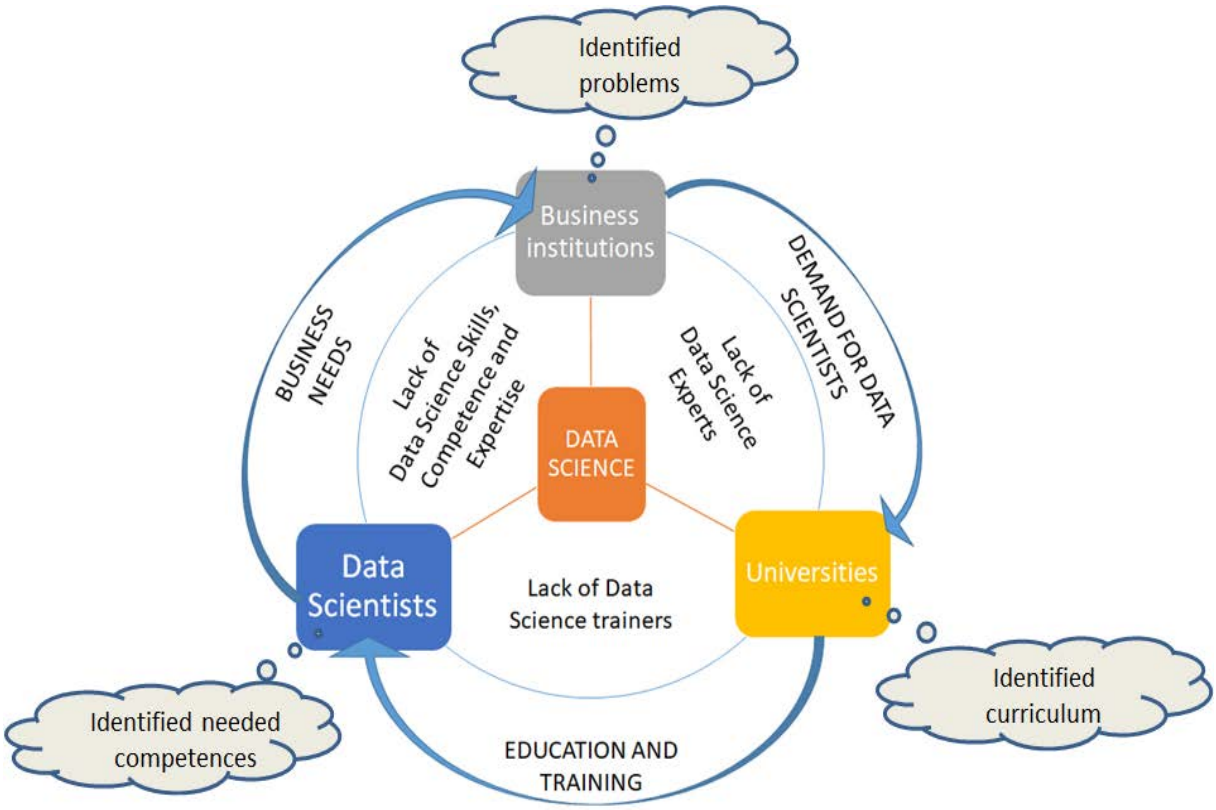

Figure 3. Data Science development loop

Due to increasing demand for obtaining value from data and the high pace of updating the required competences, the universities have to respond by flexible, dynamic, and agile advancement of Data Science curriculum. It depends on communication between three parties: (1) universities that train data specialists; (2) business organizations, that looking for data specialists to solve their emerging problems and (3) data specialist themselves. The lack of a strong link between these three parties will lead to imbalances and a negative impact on the pace of development of Data Science. In fact, universities have the responsibility to train professionals to meet the demand of the business organizations. In order to ensure a high quality learning process in the preparation of students, Data Science trainers with the necessary expertise have to be involved. Developing competences and training professionals in this field represents a significant challenge to educational institutions. The growing job market for Data Scientists defines the demand for constant review and update of existing curriculum to ensure an adequate education. The development of identified skills and responding to continues demand for adding newly emerged competences to the profile of a data specialist, expands the scope of Data Science and leads to its evolution and refinement.

5 CONCLUSIONS

Over the past few years there has been an unprecedented explosion in the organizations interest of Big Data, and the need for people, who are fluent in working with data, has grown exponentially. Many researchers explore the Data Scientist's professional profile and the shortage for Data Science and analytics workers. McKinsey Global Institute claims that the U.S. economy could be short as many as 250,000 Data Scientists by 2024. Nowadays Data Scientist is faced with lots of serious data, process and management challenges, discussed above. The main reason is the trans-disciplinary nature of Data Science. It is built by a synergetic contribution of variety and diverse areas of knowledge – from pure technological achievements to process Big Data via intensive use of data analytical techniques toward different defined by the data domain, way to present and visualize result in a compact and informative forms. The last, but not least is the ability of professional in data Science to work in teams, to respect and adopt others' ideas and to be innovative and creative. These new realities are challenging not only for the business, but also for education, especially when the needs for competences are changing so rapidly.

In conclusion, we argue that the Data Science development depends on the advancements of tools and techniques for collecting, preprocessing and analyzing Big Data, but also from the available human capital. The conservative nature of education, the relatively long period of proving the benefit of made changes, is also a challenge to address the demand in such a dynamic field as Data Science. That is why the solution should be sought in bridging the gap between academic institutions and industry in a way to train students properly.

ACKNOWLEDGEMENTS

This work has been supported by National Science Fund at the Ministry of Education and Science, Republic of Bulgaria, within the Project DM 12/4 - 20/12/2017.

REFERENCES

[1] A. Abualkishik, “Hadoop And Big Data Challenges,” *Journal of Theoretical and Applied Information Technology*, vol. 97, no. 12, pp. 3488- 3500, 2019.

[2] B. Tierney, “Data Science Is Multidisciplinary,” 2012. Retrieved from https://oralytics.com/2012/06/13/data-science-is-multidisciplinary/

[3] C. Eubanks, “Three lessons cross fit taught me about data science,” 2016. Retrieved from https://blogs.gartner.com/christi-eubanks/three-lessons-crossfit-taught-data-science/

[4] C. Turkay, N. Pezzotti, C. Binnig, H. Strobelt, B. Hammer, D. A. Keim, J.-D. Fekete, T. Palpanas, Y. Wang, F. Rusu, “Progressive data science: Potential and challenges,” 2018. Retrieved from https://hal.inria.fr/hal-01961871/document

[5] C. F. J. Wu, “Statistics = Data Science?”, *H. C. Carver Professorship Lecture*, The University of Michigan, Ann Arbor, 1997. Retrieved from http: //www2.isye.gatech.edu/~jeffwu/presentations/ datascience.pdf.

[6] C. F. J. Wu, “Statistics = Data Science?”, *P. C. Mahalanobis Memorial Lecture*, The Indian Statistical Institute,1998.

[7] D. Conway, “Data Science Venn Diagram,” 2010. Retrieved from http://drewconway.com/ zia/2013/3/26/the-data-science-venn-diagram

[8] D.M. Dedge Parks, “Defining Data Science and Data Scientist,” Graduate Theses and Dissertations, 2017. Retrieved from http://scholarcommons.usf.edu/etd/7014

[9] F. Provost, T. Fawcett, “Data Science Its Relationship Data-Driven Decision Making,” vol.1, no.1, pp.51–59, 2013. Retrieved from http://doi.org/10.1089/big.2013.1508

[10] J. Piaget, “The epistemology of interdisciplinary relationships”. In L. Apostel, G. Berger, A. Briggs, & G. Michaud (Eds.) Interdisciplinarity: Problems of teaching and research in universities, pp.127–139, 1972.

[11] J. W. Tukey, “Exploratory Data Analysis,” . Reading, MA: Addison-Wesley, 1977.

[12] K. Ayankoya, A. Calitz, J. Greyling, “Intrinsic Relations between Data Science, Big Data, Business Analytics and Datafication,” Proceedings of the Southern African Institute for Computer Scientist and Information Technologists Annual Conference 2014, pp.192-198, 2014.

[13] K. Rasheva-Yordanova, V. Chantov, I. Kostadinova, E. Iliev, P. Petrova, B. Nikolova, “Forming Of Data Science Competence For Bridging The Digital Divide,” Proceeding of 9th International Conference The Future of Education, pp. 174-179, 2019.

[14] K. Wadhwani, Y. Wang, “Big Data Challenges and Solutions,” 2017. Retrieved from https://www.researchgate.net/publication/313819009_Big_Data_Challenges_and_Solutions

[15] L. Cao, “Data science: A comprehensive overview,” *ACM Comput. Surv.*, vol.50, no.3, Article 43, 2017

[16] M. Andrejevic, “Big Data, Big Questions|The Big Data Divide,” *International Journal of Comm.*, vol.8, no.17. Retrieved from https://ijoc.org/index.php/ijoc/article/view/2161/1163

[17] M. L. Brodie, “On Developing Data Science. Applied Data Science,” 2019. Retrieved from https://www.researchgate.net/publication/333752267_On_Developing_Data_Science

[18] W.M.P. van der Aalst, “Data Scientist: The Engineer of the Future,” Proceedings of the I-ESA. Enterprise Interoperability VI, Springer-Verlag, Berlin, pp.13-28, 2014.

[19] P. J. Diggle, “Statistics: A Data Science for the 21st Century,” Royal Statistical Society, Series A (Statistics in Society), vol.178, part 4,pp. 793-813, 2015.

[20] P. Naur, “Concise Survey of Computer Methods,” Studentlitteratur, Lund, Sweden, 1975.

[21] S. Mantri, “Data Science: Literature Review and State of Art,” 2016. Retrieved from https://www.researchgate.net/publication/323393464_Data_Science_Literature_Review_State_of_Art

[22] S. Bateman, C. Gutwin, M. Nacenta, “Seeing things in the clouds: the effect of visual features on tag cloud selections,” *Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*, pp. 193-202, 2008.

[23] S. Donahole, “5 Real-Time Challenges Faced by Data Science Industry and How to Combat It,” Retrieved from https://towardsdatascience.com/5-real-time-challenges-faced-by-data-science-industry-and-how-to-combat-it-35b9027d4ce9

[24] S. Maasen, O. Lieven, “Transdisciplinarity: a new mode of governing science?” Science & Public Policy, vol.33, no.6, pp.399-410, 2006.

[25] S. Toleva-Stoimenova, K. Rasheva-Yordanova, D. Christozov, “New Dimensions Of Data Science Professional Skills As Emerged By Identified Ethical Issues: GDPR,” *Proceedings of ICERI2018 Conference* 12th-14th November 2018, Seville, Spain, pp. 0488-0497, 2018.

[26] U. Matter, “Data Science in Business/Computational Social Science in Academia?” 2013. Retrieved from http://giventhedata.blogspot.com/2013/03/data-science-in-businesscomputational.html

[27] V. Dhar, “Data Science and Prediction,” *Communication of the ACM*, vol.56, no.12, 2013. http://doi.org/10.1145/2500499

[28] S.Shum, R. Baker, J. Behrens, M. Hawksey, N. Jeffery, R. Pea, “Educational Data Scientists: A Scarce Breed,” 2013. Retrieved from http://simon.buckinghamshum.net/wp-content/uploads/2013/03/LAK13PanelEducDataScientists.pdf

[29] W. S. Cleveland, “Data science: an action plan for expanding the technical areas of the field of statistics”. *International statistical review*, vol.69, no.1, pp.21-26, 2001.

[30] Y. Donghui, E. Davis, “A First Course in Data Science,” 2019. Retrieved from https://arxiv.org/pdf/1905.03121.pdf