ANALYSIS OF MISSING DATA SCIENCE COMPETENCE IN IT SECTOR

K. Rasheva-Yordanova, E. Iliev, V. Chantov

University of Library Studies and Information Technologies (BULGARIA)

Abstract

Over the last few years, there has been an unprecedented growth in the interest of organizations in big data and analysis. The demand for certified data science and big data specialists is becoming more and more intense. For business executives, it is clear that providing human capital with better analytical skills to handle big data will allow for more efficient and effective solutions and will increase overall company productivity. This is the main factor that guides the manager in defining the Skills and Knowledge needed in a good data specialist.

Recent researches have shown that there is a clear mismatch between the needs of the industry for data science specialists and the knowledge and skills offered by higher education institutions. It turns out that most of the universities do not prepare the specialists for the needs of the business. Based on a profile created by the data specialist, this article defines the set of skills that today's graduates do not possess, but should have.

Keywords: data science, big data, data scientist, data science skills.

1 INTRODUCTION

With the advent of big data the search for respective experts has become more intensive. Experts are needed who have the necessary experience, qualifications, knowledge, skills and competences to operate and draw knowledge from the data available. According to a LinkedIn [1] report on the extent of demand for newly emerging job vacancies in the USA in 2017, Data scientist comes second among 20 professions right after Machine Learning Engineer. According to the same survey for the period 2012-2017, there is an unprecedented increase in demand for data scientist. Data shows that interest in data scientist grew nearly 6.5 times within the survey period.

Key factors for the increased demand for data scientist are, on the one hand, the availability of big data that is too large and complex for processes using traditional storage and analysis technologies [3]. On the other hand, managers have an understanding that the knowledge gained from the accumulated data has a strategic importance for the development of the organization they represent.

It is a fact that organizations use their data to improve the efficiency and effectiveness of their operations [4]. Providing human capital with high analytical skills for working with big data allows for more efficient and effective solutions and undoubtedly increases the overall performance of organizations. Despite the great potential of new technologies, tools and applications for analysis, the biggest problems faced by practitioners in using these technologies are finding employees with the necessary skills [5],[6].

Today we speak of the existence of a new digital divide between those who meet the indicators demanded by the business sector and those who do not meet the full range of the requirements. Although many universities and academies offer training courses and master programs for the training of data scientists, there is a shortage of these specialists on the labor market today. Moreover, against the backdrop of the great demand for specialists with the necessary expertise in Data Science and despite the production of new experts from universities, there is now a gap between demand and supply. It turns out that fewer specialists fully meet the demands of the business. As a result, digital divide is at the level of skills, knowledge, qualifications and experience [8].

Tendency indicates a deepening of the problem. Data scientist demand will not only continue to grow but according to Davenport, data scientist is expected to be the most sought-after professional in the industry in the years to come [2].

In order to define the missing competencies between the demand and supply of data scientist, this paper will compare business requirements with the training courses offered in Bulgaria up to 2018. We believe that the lack of data science supply will lead to a crisis in the labor market caused by high

demand and low supply of specialists with the necessary experience and expertise [7]. This will respectively have an adverse effect on the economy at national level: the lack of staff in the internal market will provoke hiring of specialists from abroad.

The aim of the paper is to address the emerging problem of lack of data scientists and to examine the difference between business needs and what education in Bulgaria offers by the middle of 2018. The report is organized in three sections in total. In the first section, an analysis is made of the job advertisements in the IT sector related to Data Science, aiming at building the professional profile of the data scientist. The second section presents an analysis of the Data Science training in Bulgaria with a view in the third section of this paper to defining the correspondences and discrepancies between the demand and supply of the data scientist.

2 AN OVERVIEW OF THE DEMAND FOR DATA SCIENCE

In order to define the sought-after knowledge and experience in the job of a data scientist we have reviewed the job advertisements offered on one of the most popular search and employment sites in Bulgaria - jobs.bg. Nearly 100% of Data Scientist search listings are published by international companies operating on the Bulgarian market and abroad.

The following table summarizes the survey data that will serve as a basis for building a framework for the data scientist's search profile. The analysis of the knowledge, skills and expertise demanded by the business was based on the three-dimensional model "hard skills-soft skills-analytical skills", suggested by Rasheva-Yordanova et al [8]. The table presents the indicators education, hard skills, soft skills, analytical skills, language skills and experience.

Table 1. *Knowledge and experience base in seeking to employ a Data Scientist*

Education	Hard skills	Soft skills	Analytical skills	Experience
BSc in Statistics, Applied mathematics, Computer science or another related field. **Master's degree** in Statistics, Mathematics, Computer Science or related field **Ph.D.** in Statistics, Mathematics, Computer Science or another quantitative field	**Scripting language** (Python, R) **Objective oriented programming language** (Scala, Python, Java, C++) **Databases** **Big Data frameworks** (Hive, Spark, Hadoop) **Statistics** (SPSS, SAS); **Machine learning techniques** (clustering, decision tree learning, artificial neural networks, etc.) **Statistical techniques and concepts** (regression, properties of distributions, statistical tests, and proper usage, etc.)	**Communication skills** - ability to translate technical language to a non-technical audience. **Organizational and leadership skills** - self-motivated, organized	**Creatively thinking** **Analytical and problem solving skills**	Between 2 and 4 years.

Considering the education and qualification factors, businesses require the Data Scientist candidate to hold at least a BA degree (about 24% of job listings). The most common condition for the same indicator was the possession of an MA degree (62%), and the most rarely stated requirement is that the data scientist should have a PhD (only 14%). What brings together all the ads by the "education" indicator is the specialty: all employers require the candidate to be a specialist in Statistics, Mathematics, Computer Science or related field.

As far as the hard skills factor is concerned, demand limits the scope of data scientist most often to scripting language, objective oriented programming language, big data frameworks, statistics, machine learning techniques, statistical techniques and concepts.

In their advertisements, the possible employers seek soft skills like communication, organizational and leadership skills. The required experience varies between 2 and 4 years of relevant work experience.

In order to complement the general framework presented by our available job offers, we believe it is necessary to review the most popular software in the field of Data Science. According to a survey by [9] the popularity of some tools increases proportionally, while others have seen a sharp decline. The

picture presented in this study can tell us what the trends will be in the next few years and what tools should be put in the training courses for data scientists. The survey confirms that the most sought-after hard skills in the past few years are Python, R and SQL, which are currently part of the hard skills list.

3 AN OVERVIEW OF THE SUPPLY IN THE FIELD OF DATA SCIENCE

In order to make a comparison between demand and supply in the field of Data Science, this section of the paper will present an analysis of the knowledge and skills involved in the courses training data scientists.

Based on the analysis presented in the previous section, we know that one of the most frequently asked requirements for candidates is that they have an MA degree. This gives us grounds in our analysis to include both qualification courses and university MA programs that train Data Science specialists. The analysis will again be carried out on the basis of the three-competence model "hard - soft -analytical skills".

Table 2. *Knowledge and skills included in the training programs in Data Science in Bulgaria*

Type of courses	Hard skills	Soft skills	Analytical skills	Other
Qualification courses	**Statistical techniques and concepts** Data acquisition **Data tidying and cleaning** Data visualization, Reading spatial data Databases SQL databases	**Organizational skills** Organizing research	**Analytical and problem solving skills** Analytical graphs; Exploratory data analysis	EDA and processing
Masters programs	**Programming**, **Data Mining Algorithms, Machine Learning**, **Data Quality**, Database Systems,		Data Analytics Methods, Social Network Analysis	Semantic Data, Information Management, Information Management Systems, Cryptography, High Performance Scientific Computing, Security, Neuroscience, Natural Language Processing

The following summary can be made:

- Qualification courses on data science have a total duration of less than 2 months of intensive training. The course includes concepts and tools that are used throughout the data collection and processing process: from asking the right questions to publishing a result. We believe that the form of training as well as the short-term and non-deeper plunge into the problem area, along with the access to the subjective assessment of one or mostly two lecturers are the main shortcomings of the data science qualification courses.
- The length of training in an MA degree program in data science varies between 2 and 3 semesters (about a year, a year and a half). The course covers a number of disciplines, some of which are purely theoretical, and others are beyond the scope of the professional profile of Data scientist.

In order to define which of the two types of training (qualification courses and master programs) is closer to the required skills, knowledge and competencies, in table 3 we have provided a comparative knowledge supply analysis.

Table 3. *Knowledge supply analysis*

	HARD SKILLS				
	Scripting language	Programming language	Databases	Statistics	Machine learning
Qualification courses	✗	✗	✓	✓	✗
Master's programs	✗	✓	✓	✗	✓
	SOFT SKILLS				
	Communication skills	Organizational and leadership skills			
Qualification courses	✗	✓			
Master's programs	✗	✗			

As can be seen from table 3 qualification courses cover 40% of the core hard skills shaping the professional profile of the data scientist. Although MA programs cover a larger percentage of basic hard skills (60%), they cannot fully meet the demand of the business. With regard to the other two types of skills – soft skills and analytical skills, it is clear that neither of the two types of supply meets the criteria laid down by the labor market.

4 AN OVERVIEW OF THE MISSING KNOWLEDGE IN DATA SCIENCE

The preview of the previous section has helped us to discover the missing knowledge in the Data Science training in Bulgaria. It turns out that by completing an MA degree in Data Science in Bulgaria, newly-emerged data scientists do not receive complete training to finalize them as employees. In practice, the knowledge and skills accumulated by the specialists in the course of their training are basic but do not meet the 100% needs of the business. At the moment, even the training aimed at improving the qualification in the field of Data science still does not fully meet market demand.

None of the courses / master programs in Bulgaria use the software that is alleged to be searched by companies (such as tensorflow, weca etc.). Work is done, demonstration and training mostly on well-known paid programs like SAS, while companies rely mostly on open source programs. There is not enough statistics in courses / MA programs. It takes a lot of time and exercises to understand and comprehend the relationships between the different statistical concepts. The same applies to machine self-learning – special attention should be paid to clusters, decision trees and neural networks. The demonstration of individual algorithms from the different data mining approaches will not allow learners / students to make the necessary connections between the different knowledge involved in the training.

Summing up data on the provision of Data Science training programs within Bulgaria, we have defined the skills that Data Science graduates do not possess.

Fig. 1 presents the general framework of the knowledge and skills required in the field of Data Science in Bulgaria.

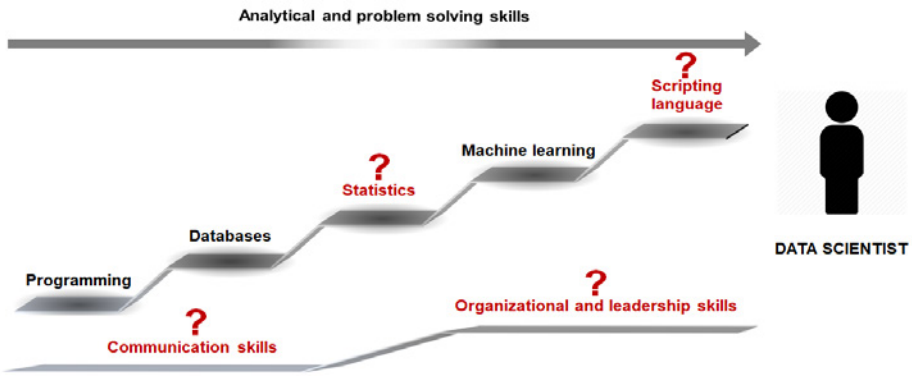

Fig. 1. Data scientist knowledge base

It is clear that modern training programs which prepare some of the most sought-after specialists nowadays manage to provide only 50% of the knowledge and skills required by the profession. This raises a number of questions. It is important to understand what reforms and modifications should be made in curriculum and data science programs so that they can meet the needs of the business and produce specialists that can find realization as soon as they graduate.

5 CONCLUSIONS

This article provides an overview of the demand and supply of data scientists within Bulgaria. In order to define the missing skills among the graduating specialists, the curricula of the Master's degree in Data Science were compared with the required knowledge and skills for recruitment. To accomplish this task, a job review has been made on one of the most popular online search sites. The result of the survey showed that there is a missing link between supply and demand in terms of Data Science. Currently neither the qualification courses nor the MA degree programs meet the needs of the business. The conclusion of the study is that urgent updates are needed on existing curricula and programs to meet the demand. This in turn will only be possible if there is a constant relationship between employers and job seekers.

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] Data Science and Machine Learning Jobs Most In-Demand on LinkedIn. Retrieved from https://www.business2community.com/linkedin/data-science-machine-learning-jobs-demand-linkedin-01986689 on [02.5.2018].

[2] Davenport T., and D. Patil, "Data scientist: The Sexiest Job of the 21st Century," Harvard business review, vol. 90, no. 5, pp. 70-76, 2012.

[3] Debortoli,S., O. Müller, and J. vom Brocke, "Comparing business intelligence and big data skills," Business & Information Systems Engineering, vol. 6, no. 5, pp. 289-300, 2014

[4] Kowalczyk M., P. Buxmann, "Big Data and information processing in organizational decision processes," Business & Information Systems Engineering, vol. 6, no. 5, pp. 267-278, 2014.

[5] LaValle S., E. Lesser, R. Shockley, M. S. Hopkins, and N. Kruschwitz, "Big data, analytics and the path from insights to value," MIT sloan management review, vol. 52, no. 2, p. 21, 2011.

[6] Mikalef, P. et al. (2018) The Human Side of Big Data Understanding the skills of the data scientist in education and industry. IEEE EDUCON 2018 Global Engineering Education Conference, At Tenerife, Canary Islands, Spain.

[7] Petrova P., Boyadzhiev D., “Training young lecturers”, XIV-th International Conference "Challenges in Higher Education and Research in the 21st Century", May 31 - June 3, 2016, Sozopol, Bulgaria, ISBN: 978-954-580-356-9, Heron Press Ltd., Vol.14, 2016, p.23-26]

[8] Rasheva-Yordanova et al. (2018) Forming of Data Science Competence for Bridging the Digital Divide. The eight edition of International Conference “Future in education” 2018. New Perspectives in Science Education, LibreriaUniversitaria Edizioni. ISSN 2384-9509 (in print).

[9] R, Python Duel As Top Analytics, Data Science software – KDnuggets 2016 Software Poll Results Retrieved from https://www.kdnuggets.com/2016/06/r-python-top-analytics-data-mining-data-science-software.html reviewed [03.5.2018]