ROAD MAP IN DEVELOPING DATA SCIENCE COMPETENCES

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

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

Abstract

The advent of modern technologies has led to the expansion of the use of databases and data services. This phenomenon has led to the rise of Data Science, and data scientists have become professionals with a particular importance in terms of extracting new knowledge from data.

Achieving a comprehensive competence for learning through “big data” is a task of re-engineering the existing way of learning. It is appropriate to understand what knowledge and skill base data scientists must have in order to meet the current technology capabilities and the expectations of business organizations.

The research addresses the following questions:

1 What skills base should a data scientist have?

2 What is the good practice in data science training?

3 What are the potential students of data science training course and what will be their input knowledge and skills base?

The aim of this study is to present the road map of the progress for a student in developing the identified competences through Data Science Curriculum. In the context of the data science area, the curriculum as an endpoint in the map refers to the building of certain skills that allow extracting knowledge from the available data and is aimed at training students of the University of Library Science and Information Technology in Bulgaria

Keywords: Data science, data science skills, data science competences.

1 INTRODUCTION

In the last few years, digital competence has become a Key requirement for what skills and knowledge people should have in the contemporary knowledge society. In practice, competence is a proven ability to apply knowledge, skills, and attitudes to achieve visible results. In the contemporary world Big Data phenomenon becomes the major factor in data driven decision making [9].

It’s been proved [1] that building up skills for working with big data can respond to key challenges and help make more evidence-based solutions and provide the ability to convert complex, often unstructured data into applicable information as a strategic response to changing global trends.

Volume of captured and recorded has led to significant challenges provoked by the need for accumulated data to be processed and knowledge that comes with them brings benefits. The educational industry faces two specific challenges to respond the market need: on the one hand the demand for well-trained capable professional is growing exponentially, but on the other, the educational institutions are challenged to develop training in this emerging area.

In practice, a large number of training organizations have to catch up with the pace of technological development and dynamics of the Data Science area to offer training to meet the current needs of business. As a result, in a previous survey, we have identified a gap between business requirements for data scientists in job advertisements and the knowledge and skills base that is set in the curriculum and syllabus of Data Science in training organizations [5].

Achieving the necessary competence for learning through “big data” is a task of reengineering existing curricula and programs. It is appropriate to understand what knowledge and skills the data specialists should have in order to respond both to the current technological capabilities and expectations of business organizations, but also to develop an appropriate educational model to help build the necessary competences.

The main purpose of this paper is to contribute to the development of knowledge by developing a roadmap for building the competencies to work with big data Flexible and scalable, to allow adopting the emerging requirements... This research study is based on the following research questions:

1 What are the base knowledge and skills a data scientist should possess?

2 Who are the potential students for a data science training MA program and what are their expected input knowledge and skills?

3 What training road-map may allow for gradual building the identified competences?

4 How to develop and strengthen data scientist competency trained by a Bulgarian university?

In the preparation of this study, we have set the following research tasks:

a) Analyze the framework of knowledge and skills that a modern data specialist should have (by reviewing the literature and existing courses, and analyzing the knowledge and skills required by business organizations in job postings for a Data scientist);

b) To analyze who are the potential students of data science training course and what will be their input knowledge and skills base;

c) To come up with an approach to developing data scientists competence.

The results of each of these tasks are presented in a separate section so that: The first section presents a framework of skills that build the profile of today’s data scientist. The second section presents an analysis of the input skills of a potential group of applicants for training in the Data Science program. In the third section is presented an approach for development of data science competence.

2 BASIC SKILLS AND COMPETENCIES BUILDING UP THE PROFILE OF DATA PROFESSIONAL

From a theoretical point of view, competence is a layered structure that shows the knowledge, skills and components that build it. The development of competence in data science is currently a difficult task mainly due to the lack of a comprehensive framework that can be referred to as a skill guideline required by the IT sector.

The framework defining the knowledge, skills, and competencies of data professionals is becoming an object of constant research. The researchers concentrate their efforts on defining the scope of competencies that build the profile of the modern data professional.

Defining the profile of the contemporary data specialist requires a two-way analysis: (1) analysis of literary sources and (2) review of the required skills by business organizations.

In the process of exploring the available and accessible sources, more than 40 skills have been revealed required from the contemporary data specialist. In order to optimize the workflow, the initial list of the required skills of the data scientists was grouped into a total of three categories – hard skills, soft skills and analytical skills [7]. Each of these categories were referred to particular skills (Table 1.)

Table 1. Data science skills framework [7]

	SOFT SKILLS	HARD SKILLS	ANALYTIC SKILLS (Technical and non-technical)
Tasks	Create and sell stories based on data, verbally and visually	Assure data quality; Build statistical models; Compute similarity Create data products/platforms; Create data visualizations; Integrate data from multiple sources, regardless of its structure and volume	Identify rich data sources; Analyze expected value; Engineer effective solutions; Find answers to important business questions; Improve decision-making; Suggest new business directions; Think data analytically; Use and analyze data; Draw causal conclusions

Skills	Intellectual curiosity business acumen communication skills Communication Entrepreneurship Curiosity	Computer science; Artificial intelligence; Automated analysis of data; Statistics; Big data; Databases; Machine Learning; Mathematics; Networking; Programming; Internet of Things; Cloud computing; Distributed computing; Data processing; Data ingestion; Data mining; Data preparation; Data tools	Academic research; Formulation; Interdisciplinarity; Scientific method; Data analysis design and interpretation; Data visualization; Data warehouses
Competencies	Understanding the basic business objectives and strategies, as this will allow maximum compaction of the knowledge gained from the data; Being able to understand stakeholders and support decision-making; Being able to communicate and disseminate the findings	To have the technical skills for statistical processing to apply in designing and interpreting experiments, modeling and forecasting. Being able to create data artifacts or optimize existing ones.	To know methods of data analysis that automate the construction of analytical models; To improve business management and achievements by enhancing decision-making.

Clearly, a Data Scientist is a professional with diverse competences qualifications [7]. But we need to recognize that the profile of this new job and his or her obligations have emerged from data-driven practice. Generally, the knowledge produced by a data scientists represents a new type of company asset.

Data need to be collected and used, but also protected by professional ethical standards, such as the new GDPR. The Data Scientists have to possess also skills to meet the data protection requirements [8] and to be able to apply appropriate data safety techniques. Or the new category of competences – ethics – has to be included into the scope of Data Scientist' profile. The Data Scientist's ethical skills belong to both hard and soft skills sets. The hard skills are compulsory for Data Scientist to accomplish their activities for collecting and processing of personal data. These include technical skills, such as programming, databases, data handling, etc. to ensure needed infrastructure to guarantee data protection. On the other hand, Data Scientist has to communicate with data subjects to comply with their right to the protection of personal data according to GDPR. These skills are from the set of their soft skills.

In addition to basic skills, data professionals must also have a specific set of competences. We summed up those that match the aims of our research as [2]:

- Ability to extract useful data from huge and diverse repositories, including public and private, and also well and poorly structured sources.
- Ability to verify the obtained data and to judge about their quality
- Ability to interpret (map) obtained data to the context (problem) and to applied appropriate analytic techniques to extract useful patterns, relationships or simply to increase understanding regarding the circumstances associated with the problem.

Based on the hard-soft-analytical-ethical skills model, the distribution of tasks in the field of each of the data science competencies can be represented in the following way (table 2):

Table 2. Structure of Data Science competence [2]

Category	Tasks	Skills
Extract	Choose, Classify, Collect, Compare, Configure, Contrast, Define, Demonstrate, Describe, Execute, Explain, Find, Identify, Illustrate, Label, List, Match, Name, Omit, Operate, Outline, Recall, Rephrase, Show, Summarize, Tell, Translate	Hard skills
Verify	Apply, Analyze, Build, Construct, Develop, Examine, Experiment with, Identify, Infer, Inspect, Model, Motivate, Organize, Select, Simplify, Solve, Survey, Test for, Visualize.	Hard skills Analytical skills
Interpret	Adapt, Assess, Change, Combine, Compile, Compose, Conclude, Criticize, Create, Decide, Deduct, Defend, Design, Discuss, Determine, Disprove, Evaluate, Imagine, Improve, Influence, Invent, Judge, Justify, Optimize, Plan, Predict, Prioritize, Prove, Rate, Recommend, Solve.	Hard skills Soft skills Analytical skills

The information included in table 2 gives us a reason to believe that information providers (so-called information brokers [4]) cover some of the required skills and competencies sought in the profile of the modern data specialist.

3 ANALYSIS OF POTENTIAL “DATA SCIENCE” TRAINING APPLICANTS

Exploring "the Big Data" available today to acquire knowledge regarding the processes, motivation and cause-and-effect, in a needed pace, effectiveness and efficiency is the major challenges faced by different businesses. This requires understanding human cognitive abilities, limitations, and inclinations in seeking information and acquiring knowledge. Fundamental understanding of why and how a given individual value information and knowledge is essential to maximizing chances to meet their needs successfully [4]. The information mediators (info brokers) plays a crucial role in the mediation between information resources and the users of information. This kind of expertise requires special training and professionally oriented education [3].

The increasing complexity of the external environment imposes upon the organization a greater demand for processing information and making quick and rational decisions. Information brokerage is the profession of information mediators. Professionals are entitled to assist clients in surviving in today's world and facing the challenges of the information era. The major role of an IB is to serve their clients by presenting in a "nut shell" the essence of information relevant to the client's problems [4].

The main similarities between the information broker and the data specialist are related to the fact that the specialists from both areas serve society by exploring data and informing their clients, and the knowledge generated by their work helps in making strategic decisions. In practice, each of the two professions requires the use of a wide range of skills that we previously limited to 4 categories. The basic difference is the focus of these two professions. Information brokers are trained to identify and assess different sources of information, not necessary stored on searchable computer repositories, such as obtained in a face-to-face communication, and not necessary exploring Big Data associated techniques. Data scientists are engaged in exploring Big Data stored and accessible via computer technology. Despite those differences, the two professions have a lot in common – both provides information service by communicating the discovered knowledge to their clients.

We have to ask the question: “Is it possible for data science to be a logical continuation of Information brokerage in the course of starting an MA program after the BA one? To find the answer, we need to define the output skills of those finishing the BA program Information Brokers. We have therefore analysed the knowledge and skills in the course of Information Brokerage at the University of Library Studies and Information Technologies (ULSIT-Bulgaria). The analysis included an overview of the curriculum and distribution of each of the course subjects in the above-defined 4 categories of skills.

The group distribution of all disciplines showed that the largest share belongs to the development of hard skills (60%), followed by soft skills (31%), analytical skills (6%) and the smallest share in the curriculum goes to ethical skills (3%). The breakdown of the disciplines forming the curriculum of the specialty by category is presented in table 3.

Table 3. Breakdown of the disciplines forming the curriculum of "Information brokers"

HARD SKILLS			SOFT SKILLS		ANALYTICAL SKILLS	ETHICAL SKILLS
Statistics and Mathematics	Programming	IT basics	Informing processes	Business and Communications	Analysis	Ethics
Discrete Mathematics	Basics of Programming	Computer Architectures	Introduction to Information Brokerage	English as a foreign language	Mathematical Analysis	Legal Aspects of IT
Linear Algebra and Analytical Geometry	Object-oriented Programming	Mobile Technologies	Information Brokerage	Basics of e-business	System Analysis	
Theory Basics of Informatics	Visual Programming Environments	Web Technologies	Information Management	Basics of e-government		
Probability Theory and Math Statistics	SQL	Operational Systems	Knowledge Organization and Management	IT career Development Systems and Competences		
Math Modelling	Data and Algorithm Structures	Databases	Informing Processes and Systems			
SAS	ASP Part I	Computer Networks and Communications	Theory and practice of the counselling			
		Information Systems				

Based on the competencies presented in Table 1, and the analysis of the skills required to build the profile of the data specialist in the first section of this paper, we can claim that in order to move up the profile of the Information Broker towards a Data Specialist it is necessary to provide training with emphasis on some essential for the Big-Data-related hard skills: Statistics, Machine Learning, Cloud computing, Data processing; Data ingestion; Data mining; Data preparation; Data tools, as well as Data analysis design and interpretation; Data visualization; Data warehouses. These disciplines are not part of the BA degree program “Information Brokerage”, but are an essential element of the data specialist profile.

In order to meet the pre-defined conditions, the curriculum of the Master's program "Data Science" must be designed to complement the knowledge and skills of students completing the bachelor's program of "Information Brokerage". Furthermore, the master degree of Data Science should be seen as a logical continuation of the knowledge and skills of students who graduate from bachelor's programs addressing the different aspects of information.

4 MAP FOR DEVELOPING DATA SCIENCE COMPETENCE

Information brokers are trained to handle information but are not prepared to work with data, especially with big data. In the course of their Bachelor's program, they study a number of soft skills disciplines, but disciplines with an emphasis on hard skills are extremely inadequate. In order to be able to realize them as data scientists, it is necessary for the educators to undergo training adapted to their input skills. In this connection, it is necessary to implement several basic steps: *(1) analysis of the input skills of the potential candidates for the specialty; (2) analysis of the data skills required by business organizations; (3) defining a list of mandatory skills in the data scientist’s profile that the information broker does not own; (4) Creating a curriculum and training aimed at building data science skills* (Fig. 1).

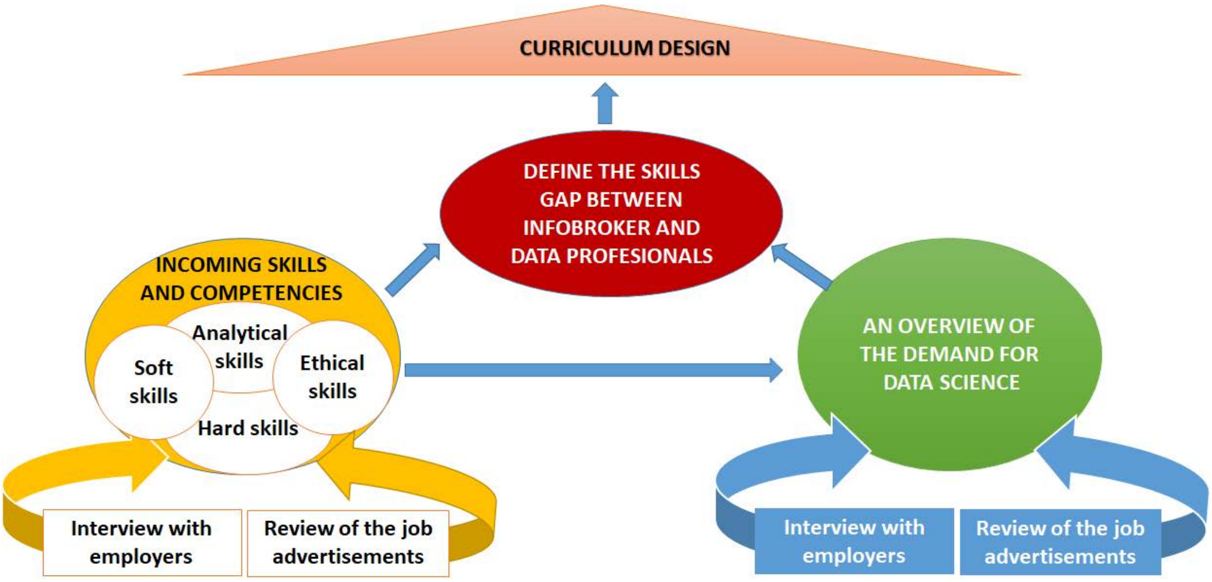
CURRICULUM DESIGN
DEFINE THE SKILLS GAP BETWEEN INFOBROKER AND DATA PROFESIONALS
INCOMING SKILLS AND COMPETENCIES
Analytical skills
Soft skills
Ethical skills
Hard skills
Interview with employers
Review of the job advertisements
AN OVERVIEW OF THE DEMAND FOR DATA SCIENCE
Interview with employers
Review of the job advertisements

Figure 1. A map for developing data science competence

Based on the results of the preliminary analyzes [2,5,6,7,8], we have developed a curriculum model of the MA program. The disciplines in it are presented as a breakdown of the previously used skill categories that can be found in Table 4.

Table 4. Disciplines Included in Data Science Master Program

HARD SKILLS			SOFT SKILLS	ANALYTICAL SKILLS	ETHICAL SKILLS
Statistics and Mathematics	Programming	IT basics	Informing, business and communications	Analysis	Ethics Threat and Fraud Exposure
Statistics	Data warehousing	Data Visualisation Ergonomics	Introduction to Data Science	Big Data Analysis	
Probability Theory and Statistics*	Data mining	Application Architecture	Behavioral Economics	Requirements Development and Management	
Mathematical analysis*	Databases*	Introduction to Cloud Technologies	Customer Relations Management (CRM)	Data Analysis and Management	
	Object-oriented Programming*	Data Technologies	ERP Systems	Modelling and Simulating in a Graphic Environment	
		Internet of Things	Project Development and Management		

The emphasis is on building hard skills (55%), soft skills (23%) and analytical skills (18%). The program includes some of those disciplines that the analysis in the first two sections identified as a minimum requirement to build the profile of the data specialist. There are subjects concerning the processes of storing, retrieving, managing and analyzing data.

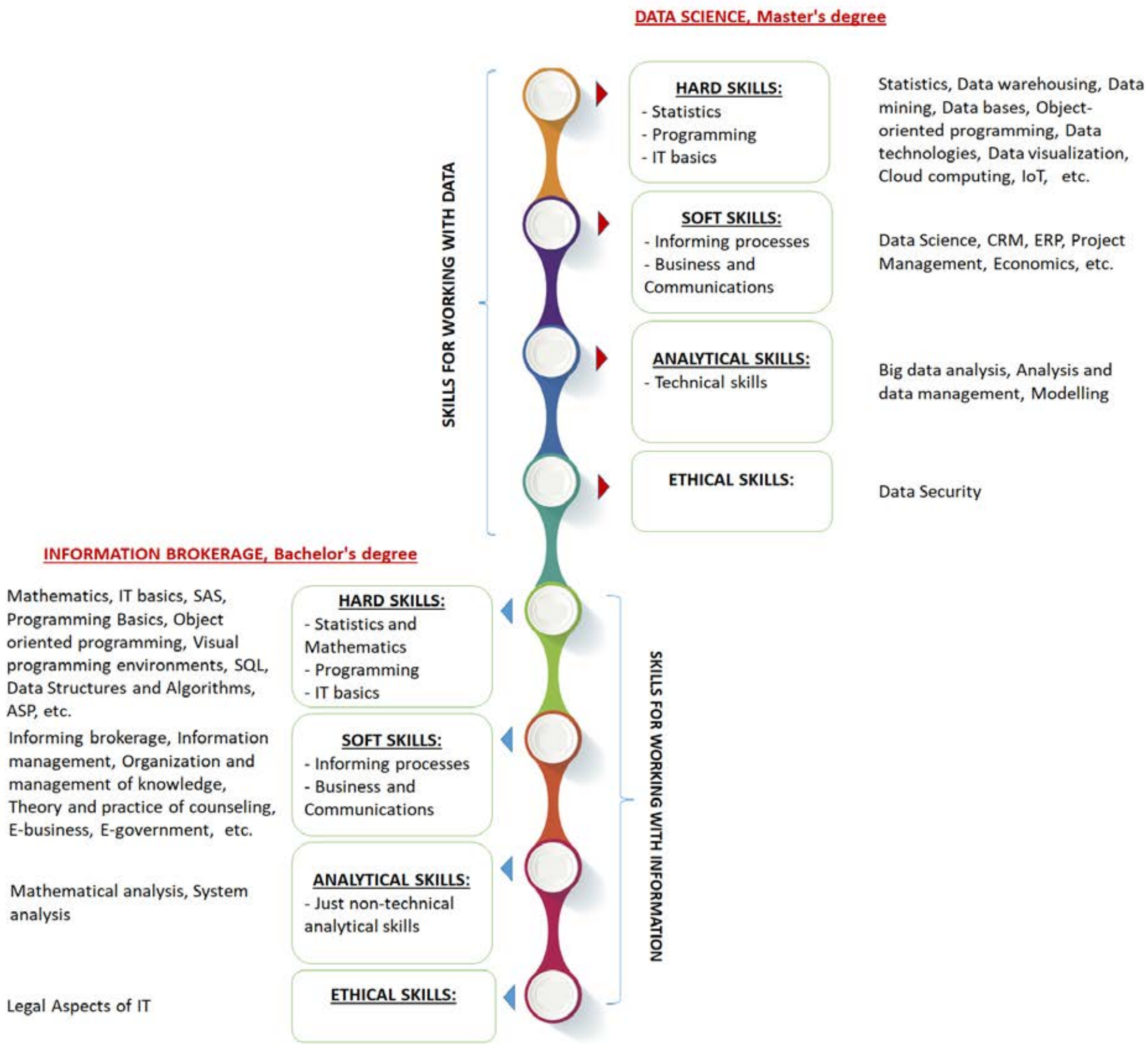
DATA SCIENCE, Master's degree
SKILLS FOR WORKING WITH DATA
HARD SKILLS:
- Statistics
- Programming
- IT basics
Statistics, Data warehousing, Data mining, Data bases, Object-oriented programming, Data technologies, Data visualization, Cloud computing, IoT, etc.
SOFT SKILLS:
- Informing processes
- Business and Communications
Data Science, CRM, ERP, Project Management, Economics, etc.
ANALYTICAL SKILLS:
- Technical skills
Big data analysis, Analysis and data management, Modelling
ETHICAL SKILLS:
Data Security
INFORMATION BROKERAGE, Bachelor's degree
Mathematics, IT basics, SAS, Programming Basics, Object oriented programming, Visual programming environments, SQL, Data Structures and Algorithms, ASP, etc.
HARD SKILLS:
- Statistics and Mathematics
- Programming
- IT basics
Informing brokerage, Information management, Organization and management of knowledge, Theory and practice of counseling, E-business, E-government, etc.
SOFT SKILLS:
- Informing processes
- Business and Communications
Mathematical analysis, System analysis
ANALYTICAL SKILLS:
- Just non-technical analytical skills
Legal Aspects of IT
ETHICAL SKILLS:
SKILLS FOR WORKING WITH INFORMATION

Figure 2. Upgrading the skills of the information broker

In summary, the two trans-disciplinary specialties – “Information Brokerage” and “Data Science” offered in ULSIT – Bulgaria are built to complement each other. In their unity, they provide the skills needed to work with both information and data.

5 CONCLUSIONS

This paper presented a map for building data science competency under specific conditions and the necessity to upgrade the skills of BA information brokers to MA data professionals. The main milestones required for a curriculum, following the route set out on the map, are related on the one hand to a review of input knowledge and skills and analysis of the needs of business organizations and, on the other, an analysis of the necessary and insufficient input skills. Developing the specific skills needed by the information broker to work with large data is possible by adapting the curriculum to the identified needs.

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] B. Daniel, Big data and analytics in higher education: Opportunities and challenges, Brit. J. Educ. Technol, 46, 904–920, 2015.

[2] D. Christozov, K. Rasheva-Yordanova, S. Toleva-Stoimenova. RISKS MANAGEMENT IN DATA SCIENCE TRAINING. Proceeding of Regional international conference on APPLIED PROTECTION AND ITS TRENDS, Zlatibor, 07-10 October, 2018

[3] D. Christozov, S. Denchev, S. Toleva-Stoimenova, K. Rasheva-Yordanova, Training information brokers: A curriculum model. Journal of Issues in Informing Science and Information Technology, 5, 87-94, 2008, Retrieved from http://proceedings.informingscience.org/InSITE2008/IISITv5p087-094Chris441.pdf

[4] D. Christozov, S. Toleva-Stoimenova, The Role Of Information Brokers In Knowledge Management, Online Journal of Applied Knowledge Management A Publication of the International Institute for Applied Knowledge Management, Volume 2, Issue 2, 2014, pp. 109 – 119. Retrieved from https://pdfs.semanticscholar.org/7940/7848512eb0864224c2dd475645aa40b524f9.pdf

[5] K. Rasheva-Yordanova, E. Iliev, V. Chantov, Analysis Of Missing Data Science Competence In It Sector". Proceedings of EDULEARN18 Conference 2nd-4th July 2018, Palma, Mallorca, Spain IATED, ISBN: 978-84-09-02709-5, pp. 7399 – 7403, 2018.

[6] K. Rasheva-Yordanova, E. Iliev, B. Nikolova. "Analytical Thinking As A Key Competence For Overcoming The Data Science Divide". 10th annual International Conference on Education and New Learning Technologies. 2nd-4th of July, 2018, Palma de Mallorca (Spain). IATED, 2018.

[7] K. Rasheva-Yordanova, V. Chantov, I. Kostadinova, E. Iliev, P. Petrova, B. Nikolova, "Forming of Data Science Competence for Bridging the Digital Divide,", 8th edition of the "The Future of Education" conference, PIXEL, Retrieved from https://conference.pixel-online.net/FOE/virtual_presentation_scheda.php?id_abs=3236

[8] 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, ISBN: 978-84-09-05948-5, 2018

[9] M. van Rijmenam, Why All Governments Should Hire a Chief Data Scientist 92 Just Like the US Did, 2015. Retrieved from https://datafloq.com/read/governmentsshould-hire-chief-data-scientist/879