Professional training for data analysts: current trends, required skills and competences*

Ekaterina Terbusheva[1]
terbushevae@gmail. ru

[1] Herzen State Pedagogical University of Russia
St. Petersburg, Russian Federation

Abstract

The article discusses the requirements for different professionals in the field of data analysis and views on the training of such specialists. The key competencies for data analysis are revealed based on a literature review, training programs review and labor market analysis. The paper describes a study of the Russian labor market, which compares the required skills in the vacancies selected by the keywords “Analyst”, “Data analysis”, “Data mining”, “Machine learning”, “Big data” and “Data Science”. The obtained results can be used in the development of data analysis curricula for students of various specialties.

Keywords: *data analysts training, skills for analytics, data analysis curricula, labor market analytics, data science, education.*

1 Introduction

Data analysis is becoming an important competence for many professionals. The analysis of the data accumulated in connection with the global digitalization allows us to extract useful information and make more effective decisions on its basis. For example, in the field of education, information obtained from the learning process data due to analysis can be used to the personalization of learning by creating intelligent learning systems (adaptive learning systems, systems with individual instructions) [Scherzinger et al., 2018] [Chen et al., 2020] [Qi, 2018].

In this regard, training of future specialists for data analysis will be increasingly in demand. According to a joint study by hh.ru and the big data academy from Mail.ru Group, in 2019 there were 9.6 times more vacancies in the field of data analysis compared to 2015 [DS specialists]. Recognizing the importance of qualified human resources, the study of necessary skills, as well as the labor supply and demand, is one of the major focus in economics, sociology, and education. Previously, the skills in demand were easy to assess and adapt to, because changes in a certain area were quite slow and gradual, but now, development in many areas is rapid and the marketable skills can change quickly and not always in a predictable way.

There are different specialists in the field of data analysis. Depending on the application area and the analysis methods used, the following professions are distinguished: business analyst, data analyst, data scientist, big data analytics, marketing analyst, statistician, etc. The rapid development of the data analysis field requires regular research of these and new vacancies in this field. It is important to understand the differences between existing vacancies, as well as the key skills expected from professionals. Such knowledge allows us to timely respond to changes and train qualified specialists.

This article is devoted to the study of vacancies in the field of data analysis in the Russian labor market and the identification of the necessary skills expected from applicants, as well as a review of training programs for such specialists. The study is organized as follows. Section 2 provides an overview of the key competencies and skills of data scientists based on the literature. Section 3 discusses training programs for students in the field of data science, both in Russian universities and abroad. Section 4 describes the study of the Russian labor market, which are compared the required skills in the vacancies selected by the keywords "Analyst", "Data analysis", "Data mining", "Machine learning", "Big data" and "Data Science".

2 An overview of key competencies in data analysis

A data analyst can interact with IT teams, management, and data scientists to set goals, collect data from various sources, clean and reduce data, analyze and interpret results using standard statistical tools and methods, identify trends, correlations and patterns in data sets, make summary reports and visualize data for management, design, create and maintain relational databases and data systems, determine whether the problems are caused by code or data. The term "Data Scientist" appeared recently but has already entrenched itself in the labor market. A data scientist is distinguished from a data analyst by stronger technical skills in mathematics (linear algebra, calculus, probability theory, etc.), statistics (hypothesis testing, summary statistics, etc.), machine learning, data mining, software engineering (including distributed computing, algorithms and data structures), data clean and integration, data visualization, unstructured data processing, programming (R and/or SAS, Python, $C/C++$, Java, Perl), big data processing (Hadoop, Hive, Pig), cloud services (like Amazon S3) [Misnevs et al., 2016].

In the article [Valencia, 2016] the competencies for a data science specialist are divided into specific and general (interdisciplinary) ones. In [Toleva-Stoimenova et al., 2019], this division is defined as hard (technical) and soft (non-technical) skills, and analytical competencies are considered as the intersection point of hard and soft data processing skills. Various specialists include the following skills as soft skills: analytical type of thinking, the ability to work with a large amount of information, attention, systematic thinking, the ability to business thinking and interact with clients, communication and presentation skills, business acumen, enterprise, curiosity, interdisciplinary orientation, associative thinking, creativity. In [Bonesso et al., 2020], based on a study of flexible competencies in data analysis and data science, they were divided into 6 groups:

- awareness (self-awareness, empathy, organizational awareness as an organization's understanding of current capabilities, abilities, potential, and results);
- action (achievement orientation, initiative, focus on performance, self-control, risk management, collection of information, flexibility);
- social (customer focus, persuasion, conflict management, teamwork, developing others, leadership);
- cognitive (diagnostic thinking, pattern recognition, system thinking, lateral thinking);
- exploratory (questioning, observing, experimenting);
- organizational action (visionary thinking, strategic thinking, opportunity recognition.

It is also appropriate to add ethical principles regarding questions of citation, data ownership, data security and privacy to the group of soft social skills.

A comparative analysis of the skills expected from a data analyst in Russia and the United States [Skhvediani et al., 2019] showed that soft skills are most often found in American vacancies, and hard skills in Russian ones. For example, in the Russian Federation programming skills are important, while in the USA these skills are not included in the five most frequent requirements for candidates. Among the 5 main categories of skills, only 3 ones are the same: statistical packages, structured data management, decision making.

3 Trends in training data analysts

In the early 2010s, a shortage of specialists in the fields of big data processing and analysis was identified in the world, and the growth of further demand for such specialists was predicted. To meet the demand, universities have begun to actively develop and offer relevant curriculum.

Currently, the website of the keystone company, which provides recruiting and marketing services in the field of higher education and allows you to find information about degrees and career paths from around the world, presents 312 master's programs, 63 undergraduate programs and 6 PhD programs in the field of data analysis [DegreePrograms]. The top 5 countries in terms of the number of master's degree programs presented on the website are the USA, Great Britain, France, Spain and Germany (Fig. 1). From Russian universities here you can find 8 master's programs: practical data analysis (Ural Federal University, campus and online), data and network analytics (HSE University, online), Data Science (HSE University, campus and online), Data Science (The Skolkovo Institute of Science and Technology, Skoltech), Data Science (Southern Federal University), Data Management (North-Caucasus Federal University, campus). These statistics do not show the total number of programs in the field of data analysis in the countries, since many programs may not be posted on the site or may not be available to foreign students, but it shows the relative development in the training of the specialists in question.

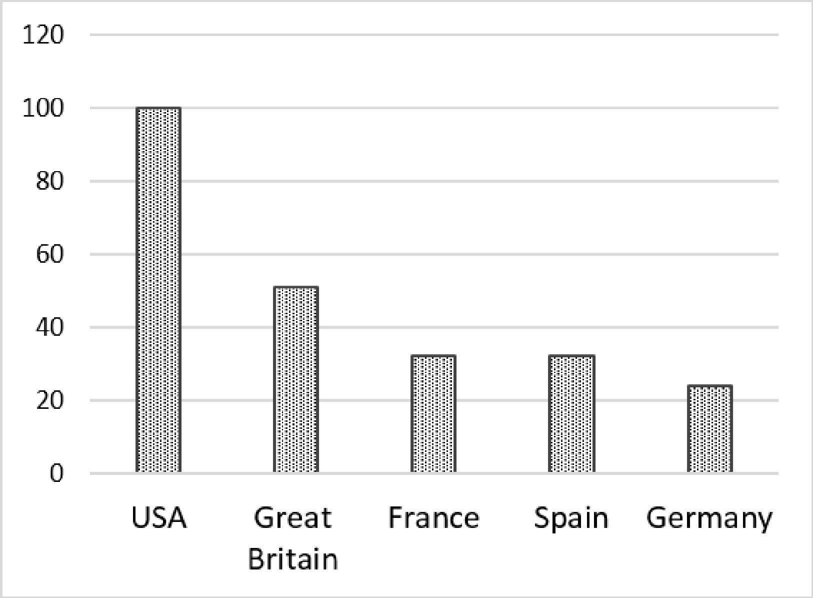

Figure 1: Number of master's degree programs in data analysis available to international students on the keystone website

A group of researchers in mathematics, statistics, and computer science [De Veaux et al., 2017] developed curriculum guidelines for undergraduate programs in data science. To successfully cope with all stages of data processing, from data collection to the presentation of results, it is recommended to develop the following key competencies: analytical thinking, mathematical foundations, model construction and evaluating, fundamentals of algorithms and programming, data preparation and data management, knowledge transfer. The proposed content of the undergraduate program is presented in Table 1.

Currently, there is a tendency in Russia to training of specialists in the field of data analysis under the master's programs. At the Department of Theoretical Informatics of the Faculty of Mechanics and Mathematics of the Moscow State University, it is planned to teach a specific specialization in methods and algorithms of the representing, modeling, and analyzing large data based on the courses "Data Models and Basics of Database Systems", "Databases. Additional chapters", "Big data analytics. Basic algorithms", "Big data analytics. Additional chapters", as well as a number of special workshops on selected areas in data science. It is assumed that students already know the material on linear algebra and its applications, on probability theory and statistics, and on programming [Glavatsky et al., 2016].

An example of a program aimed at developing both technical and flexible skills is the master's program in Machine Learning and Big Data Technologies of Southern Federal University, which has the following structure:

- general professional disciplines (methodology of scientific activity, psychology of personal resource management, professional and academic communication in the field of computer science, research project);
- professional disciplines (algorithms and data structures, mathematical foundations for big data analysis, IS design and management, data mining methods, methods of artificial intelligence, software and hardware of information systems, modern problems and methods of applied computer science, technology of big data analysis, IT projects management, IT security, software and hardware information security and IS administration);
- electives (intelligent Internet technologies, information, and psychological security of the individual).

Not many universities in Russia which provide education for data science professionals. Nevertheless, following the leading universities of the country (St. Petersburg State University, St. Petersburg State University of Economics, Moscow Institute of Physics and Technology, Moscow State University, HSE), training programs related to data analysis and data science are beginning to be developed at other universities. For example, for the direction of training "Applied Mathematics and Informatics" at Petrozavodsk State University in the 2021/2022 academic year, a new master's program "Intelligent Internet Technologies" is opening. Intelligent Internet technologies are based on data collection and analysis solutions in the Internet of Things. The program includes the following academic disciplines: technologies of intelligent spaces, intelligent Internet technologies, cyber-physical systems with artificial intelligence, technologies for programming data processing services, technologies for organizing computing in the Internet environment, planning the capacity of network infrastructures, distributed systems of the Internet of things, management of the software development process, documentation and analytics in IT projects [Korzun et al., 2021].

The existing master's programs require certain technical skills in the fields of mathematics, statistics, programming, and designed for training of students who have an undergraduate degree in information technology and computer science. Such students already have a number of competencies shown in Table 1.

In addition to the incoming requirements focused on technical knowledge and skills, the article [Toleva-Stoimenova et al., 2019] also proposes to evaluate the analytical thinking of students in general as a necessary indicator of the student's readiness to successfully completion of the program. Analytical thinking abilities refer to the highest forms of cognitive activity and include the analysis

Table 1: Content of the undergraduate program in data science

Competence group	
Objectives	**Content**
Mathematical foundations	
- to form values of mathematical methods with understanding of their limitations; - to develop geometric, intuitive, and visual thinking.	- mathematical structures (functions, sets, relations) and logic; - linear modeling and matrix computation (matrix algebra and factorization, eigenvalues/eigenvectors, projection/least-squares); - optimization (calculus concepts related to differentiation); - multivariate thinking (concepts and numerical computation of multivariate derivatives and integrals); - probabilistic thinking and modeling (counting principles, univariate and multivariate distributions, and independence).
Algorithms and Software Foundations	
- to acquire the ability to build algorithmic solutions; - to provide the readiness to implement them by programming in a high-level language.	- development of algorithms (problem decomposition, evaluating alternative solutions, choosing an effective algorithm); - programming concepts (procedural and functional programming); - data structures (lists, vectors, date frames, dictionaries, trees, and graphs); - tools and environments (for input, output and transfer of data, transformation and exploration of data, visualization and analysis, version control tools); - work with big data (parallel programming, distributed data storages, work with streaming data).
Data management	
- develop the ability to effectively apply the principles of data management.	- databases and systems that support big data; - query languages, both for relational databases (sql) and NoSQL systems; - collection of weakly structured data via the Internet, web services, access to streaming data; - cleaning and converting data into structured forms.
Statistical modeling	
- introduce students to statistical data analysis; - prepare for further comparison of linear models and nonlinear approaches.	- concepts of statistical analysis, statistical inference; - exploratory data analysis and graphical data analysis methods; - point and interval estimation and testing of hypotheses: statistical (as the central limit theorem, the law of large numbers) and algorithmic (as bootstrapping) approaches; likelihood theory, Bayesian methods; - Monte Carlo simulation of stochastic systems, inference based on resampling (bootstrap, jackknife, permutations); - introduction to models: simple linear, multivariate and generalized linear models, algorithmic models (like decision trees and nearest neighbor method), unsupervised learning (clustering); - introduction to model estimation and selection (regularization, bias/variance tradeoff, cross-validation, penalized regression, ridge regression).
Statistical and machine learning	
- develop practical skills in applying widespread machine learning methods to solve problems in various fields.	- alternatives to classical regression and classification - algorithmic analysis of models, solving scalability and implementation issues - performance metrics and prediction quality, cross-validation - data transformations (creation of new features, dimension reduction methods, such as principal component analysis, smoothing, and aggregation) - supervised vs. unsupervised learning - ensemble methods (such as boosting, bagging, and model averaging)

of statements or evidence, the ability to draw conclusions using inductive or deductive reasoning, the ability to evaluate, make decisions, or develop solutions to problems. To evaluate analytical competencies, the authors propose a questionnaire with five types of questions:

> [to identify similarities, differences, patterns in sequence, contradictions; to answers following the provided definition instead of using intuitive properties of this entity; to distinguish between facts and opinions; to compare and apply basic mathematical facts in everyday situations; to self-evaluation of their analytical thinking skills.

In the next section we will look at the competencies that should be included in data analysis curricula based on labor market research.

4 Analysis of the Russian labor market

Let's analyze the most popular job site in the Russian Federation - hh.ru in order to determine the current development of analytics in the Russian Federation, to identify the need for data analysis experts and the requirements for them. Vacancies were collected separately for the keywords "Analyst", "Data analysis", "Data mining", "Machine Learning", "Big Data" and "Data Science" using an open project "Automation of search and analysis of vacancies hh.ru" [GitHubProject] and data were analysed using Python libraries. Data mining (DM) is a direction in the field of information technology that studies methods for detecting hidden and useful patterns based on data. Machine Learning and Big Data are related to Data Mining areas. The number of vacancies found on the hh.ru website for each request for one month of 2021 in the Moscow region is shown in Table 2.

Table 2: Number of vacancies on hh.ru

Request on hh.ru	**Number of vacancies**	**Request on hh.ru**	**Number of vacancies**
Data analysis	2000	Machine learning	1506
Analyst	2000	Big Data	1058
Data mining	372	Data Science	596

372 vacancies were found for the keywords "Data mining". The number of different companies that are looking for such specialists is 176. In the description of each job on the site, there is a separate field with key skills expected from the candidate. The analysis of these key skills helps to discover the actual knowledge and technologies used to solve the real problems of data analysis at the present stage. The top 10 key skills that are most prevalent in the "Data mining" vacancies are given in Figure 2.

The most popular skills can be divided into several categories:

1. Data mining

2. Programming languages (often Python, less often C++, Java, VBA)

3. Databases (SQL, Oracle, MS SQL Server)

4. Mathematical statistics

5. Operating systems (Linux)

6. Foreign languages (English)

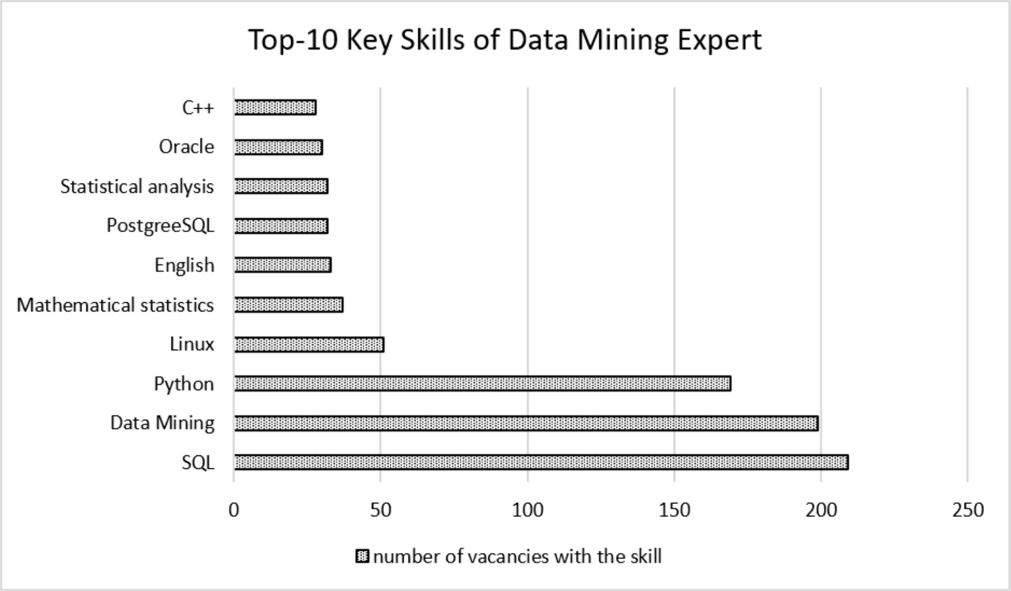

Figure 2: Key skills of a Data mining expert

To show in more detail what is hidden under the first skill “data mining”, let’s look at the identified tasks in the field of data mining based on the ontology of this area [OntoDM], which includes the ontology of data types, the ontology of basic entities, and the ontology of research in the field of DM. The concept “data mining task” refers to the main entities of the DM. The taxonomy of tasks based on data types is given in Figure 3.

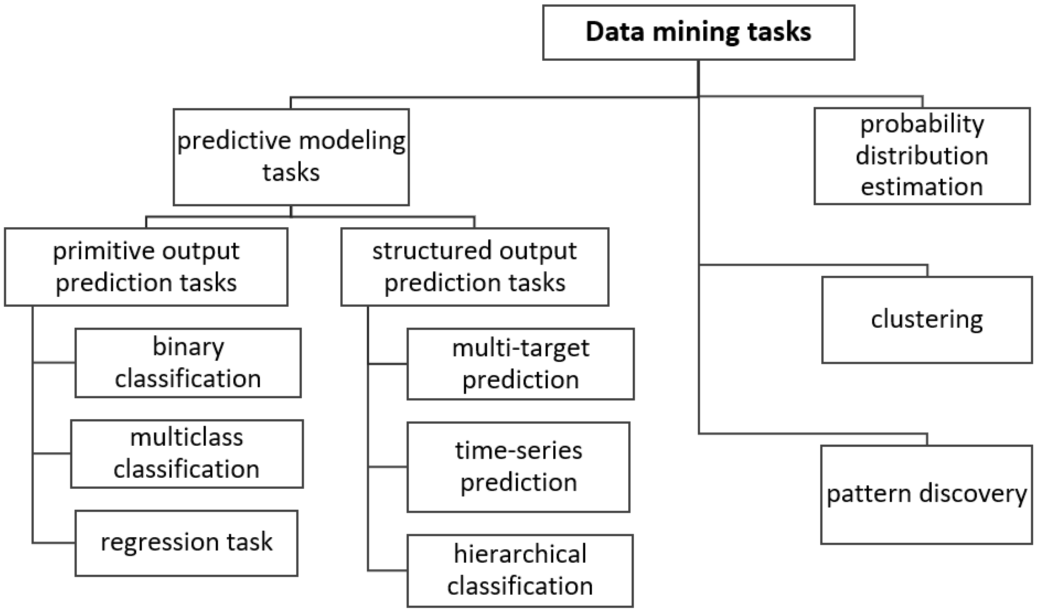
Data mining tasks
predictive modeling tasks
probability distribution estimation
primitive output prediction tasks
structured output prediction tasks
binary classification
multiclass classification
regression task
multi-target prediction
time-series prediction
hierarchical classification
clustering
pattern discovery

Figure 3: Taxonomy of data mining tasks

The key skills for the jobs found on the vacancies website in the other search queries listed at the beginning of this section are similar. The visualization of the key skills in the different groups of vacancies using the Euler-Venn diagram (Fig. 4), which shows the intersections by skills for the studied vacancies, allows to conduct a comparative analysis.

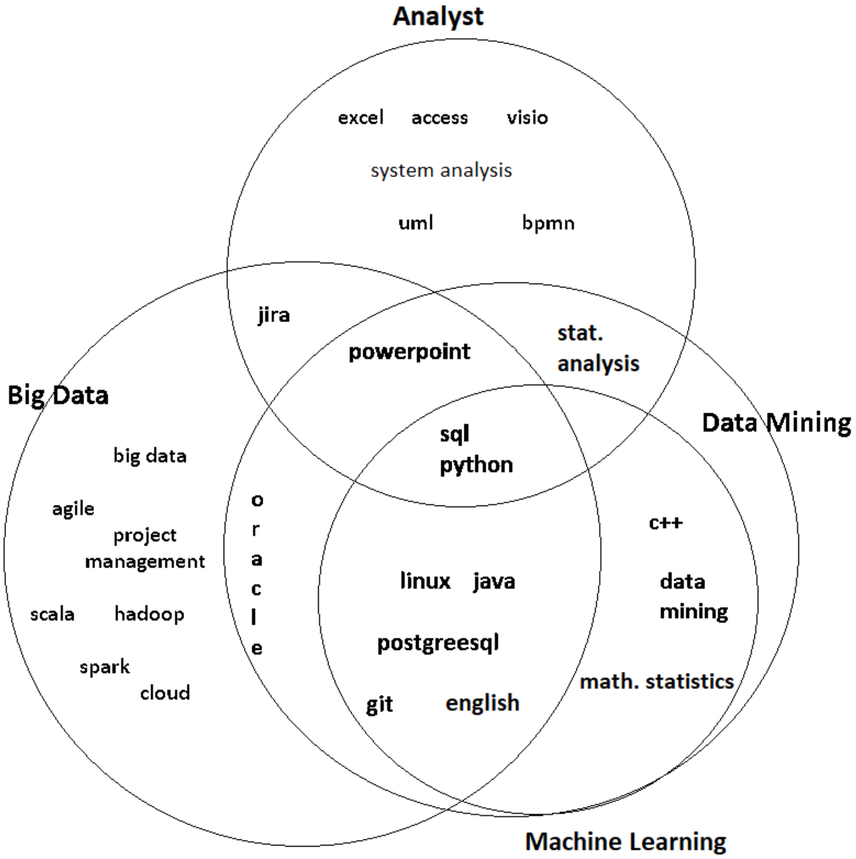

Figure 4: Frequent skills in data analysis jobs

The main differences are:

- vacancies for the query “Data Analysis” and “Analyst” involve a wider range of activities than vacancies for the query “Data mining” (for example, they include specialists who know statistical analysis and data visualization in Excel, as well as professionals who can model business processes using specialized tools);
- vacancies on request “Bid data” also involves knowledge of technologies for working with big data, that cannot be processed using a single PC (such as Hadoop, Spark, Cloud technologies) and project management skills;
- The top skills of a data scientist are identical to the top skills of a machine learning specialist, with the exception of PostgreSQL and the inclusion of project management and teamwork skills.

Conclusion

The demand for data analysis specialists exists and remains unmet, both in Russia and abroad. The rapid development of data science leads to the emergence of new specializations in this field and requires regular updating of existing and the development of new curricula aimed at training qualified specialists. As the overview of the key competencies of the data science specialist showed, it is necessary to develop both the technical rigid and flexible skills described in the article. The study of the labor market in the Russian Federation allowed in addition to the skills in the fields of mathematics, statistics, databases, management and big data processing, software development, which are defined by many authors, to identify additional competencies expected from Russian specialists: skills in working with the Linux operating system, knowledge of English, knowledge of specialized tools for modeling and visualization of business processes.

When training data analysts, we need to consider these key skills that are relevant for solving real-world practical problems. We plan to apply the research results to adapt the educational data mining training program [Piotrowska et al., 2019] to prepare specialists for the organization of personalized learning based on the data of the digital educational environment.

Acknowledgement

The research was carried out within the state assignment of Ministry of Science and Higher Education of the Russian Federation (project № FSZN-2020-0027)

References

[Scherzinger et al., 2018] Scherzinger F., Singla A., Wolf V., Backenköhler M. (2018) Data-Driven Approach Towards a Personalized Curriculum. In: Proceedings of The 11th International Conference on Educational Data Mining (EDM2018).

[Chen et al., 2020] Chen Z., Demmans C. (2020) CSCLRec: Personalized Recommendation of Forum Posts to Support Socio-collaborative Learning. In: Proceedings of The 13th International Conference on Educational Data Mining (EDM). 2020. Pp. 364 – 373.

[Qi, 2018] Qi Z. (2018) Personalized Distance Education System Based on Data Mining. // International Journal of Emerging Technologies in Learning (iJET). 2018. vol. 13, No 07.

[OntoDM] The OntoDM ontology. URL:http://ontodm.com/doku.php

[GitHubProject] Automation of search and analysis of vacancies hh.ru. = Avtomatizaciya poiska i analiza vakansij hh.ru. (In Russ.). URL: https://github.com/capitanov/hh_research

[DS specialists] Data Science specialists: the main skills and demand of employers. 2020. = Specialisty po Data Science: osnovnye navyki i spros rabotodatelej. 2020. (In Russ.). URL: https://hh.ru/article/27128

[Misnevs et al., 2016] Misnevs B., Yatskiv I. (2016). Data Science: Professional Requirements and Competence Evaluation // Baltic Journal of Modern Computing. vol. 4. Pp. 441-453.

[Valencia, 2016] Valencia J. (2016) Data Science: needed competences and applications // Clusters. Research and development. 2016. No 3 (4). pp. 45-49. = Nauka o dannyh: trebuemye kompetencii i ih primenenie // ZHurnal Klastery. Issledovaniya i razrabotki. 2016. No 3 (4). Pp. 45-49. (In Russ.)

[Glavatsky et al., 2016] Glavatsky S., Burykin I. (2016) About courses cycle “Data science and data mining for mathematicians” // CEUR Workshop Proceedings (CEUR-WS.org): Selected Papers of the XI International Scientific-Practical Conference Modern Information Technologies and IT-Education (SITITO 2016), Moscow, Russia, November 25-26, 2016. vol. 1761. Pp. 58–63. = O cikle kursov «Analitika bol'shih dannyh dlya matematikov» // Trudy XI Mezhdunarodnoj nauchno-prakticheskoj konferencii «Sovremennye informacionnye tekhnologii i IT-obrazovanie»

(SITITO'2016), Moskva, Rossiya, 25-26 noyabrya, 2016 (In Russ.). URL: http://ceur-ws.org/Vol-1761/paper07.pdf

[Toleva-Stoimenova et al., 2019] Toleva-Stoimenova S., Christozov D., Rasheva-Yordanova K. (2019). Entry competences assessment of data science potential students // Proceedings of the The 13th annual International Technology, Education and Development Conference (INTED2019), Valencia. Pp. 4248-4256. DOI: 10.21125/inted.2019.1066.

[Bonesso et al., 2020] Bonesso S., Bruni E., Gerli F. (2020) When Hard Skills Are Not Enough: Behavioral Competencies of Data Scientists and Data Analysts. In: Behavioral Competencies of Digital Professionals. Palgrave Pivot, Cham. https://doi.org/10.1007/978-3-030-33578-6_4

[Skhvediani et al., 2019] Skhvediani A., Arteeva V., Sviridenko M. (2019). Comparative Analysis of the Framework of Skills of a Data Analyst Job in Russia and the USA. In: Proceedings The 34th IBIMA conference, Madrid, Spain, November 13-14, 2019.

[Korzun et al., 2021] Korzun D., Bogoyavlensky Yu., Dimitrov V.M., Bogoyavlenskaya O., Petrina O.B., Ponomarev A.V., Marchenkov S.A.(2021) About Master's Degree in Intelligent Internet Technologies at Petrozavodsk State University // In: Proceedings of the conference "Teaching Information Technologies in the Russian Federation-2021", online, May 19-20, 2021. = O magistrature po intellektual'nym internet-tekhnologiyam v Petrozavodskom gosudarstvennom universitete // Sbornik trudov konferencii: Prepodavanie informacionnyh tekhnologij v Rossijskoj Federacii - 2021, onlajn, Maj 19-20, 2021 (In Russ.). URL: https://it-education.ru/conf2021/thesis/4716/

[DegreePrograms] Master's degree Programs. URL: https://www.masterstudies.com/Masters-Degree/

[De Veaux et al., 2017] De Veaux R., Agarwal M., Averett M., Baumer B., Bray A., Bressoud T., Bryant L., Cheng L., Francis A., Gould R., Kim A., Kretchmar R., Lu Q., Moskol A., Nolan D., Pelayo R., Raleigh S., Sethi R., Sondjaja M., Ye P. (2017). Curriculum Guidelines for Undergraduate Programs in Data Science. Annual Review of Statistics and Its Application. 4. DOI: 10.1146/annurev-statistics-060116-053930.

[Piotrowska et al., 2019] Piotrowska X., Terbusheva E. (2019) Educational Data Mining for future educational employees. In: CEUR Workshop Proceedings. NESinMIS-2019 - Proceedings of the 14th International Conference "New Educational Strategies in Modern Information Space". 2019. C. 38-49. URL: http://ceur-ws.org/Vol-2401/PAPER_4.PDF