Forming of Data Science Competence for Bridging the Digital Divide

Katia Rasheva-Yordanova[1], Veselin Chantov[2], Iva Kostadinova[3], Evtim Iliev[4], Pepa Petrova[5], Boriana Nikolova[6]

Abstract

Today is important to have a knowledge how to storage, processing, and to searching in data, but more important is to have skills and to know how to extract useful knowledge from the big data and how to use that knowledge. More and more tangible becomes the need to carry out adequate training aimed at acquiring the necessary competencies for evaluation, verification and correct interpretation of statistical measures. The understanding the capabilities of information technology to save all facts and events occurring inside and outside an organization, as well as the detection and causal links explaining behavior, form the mandatory competencies in the age of the big data. The phenomenon "Big Data" opening up a new stage of "digital divide" affecting both organizations and individuals and is primarily the result of the complexity of processing and interpreting of the available data. There is a divide between the people who "haves" and "have-nots" skills and competencies to gain new knowledge from existing data. This article discusses the specifics of digital divide caused by the availability of big data. Based on research have been determined the existing barriers to overcome the problem. The article focuses on formulating the basic set of skills and competencies that must have every data science specialist.

1. Introduction

In the last few years we have witnessed an unprecedented explosion in the interest of organizations in big data and data science. Today, this topic is one of the most-discussed in research and practices as many organizations are striving to use the data they possess or control aiming to improve effectiveness and efficiency of their operations [14].

Modern information technologies allow the registering and storing of all facts related to emerging events. This naturally leads to the accumulation of Big Data. In many cases, searchable data repositories are designed in such a way that they do not allow learning to be done in an easy way. In order to gain knowledge of the accumulated complex and complicated data, computer applications are needed. This in turn limits the number of people who have the necessary experience to take advantage of the data sources [3,4,5]. Thus, the Big Data phenomenon imposes a new social divide between those capable of self-studying from Big Data (members of the s-called “Big Data Elite”) and those relying on intermediaries in order to “study” data. This new divide adds new aspects to the already existing in society digital divide [4].

Today, it is important to understand the available storage, processing and searching capabilities of large data sets, but more importantly, there is the ability to extract the useful knowledge from the data and how to use that knowledge. This phenomenon has allowed the term Data Science to attract considerable attention in recent years, turning professionals defined as Data Scientist into experts of particular importance. The role of these specialists is widely discussed and academically recognized, and the framework defining the knowledge and competencies that data professionals have keeps expanding. The answer to questions is increasingly sought [4] referring to basic attributes of literacy on working with big data, as well as the necessary competencies to learn from them.

The main purpose of this article is to define the necessary competencies that each data specialist must possess in order to overcome digital divide. The article is organized in 3 sections as follows:

[1] University of Library Studies and Information Technologies, Bulgaria, E-mail: k.rasheva@unibit.bg
[2] University of Library Studies and Information Technologies, Bulgaria, E-mail: v.chantov@unibit.bg
[3] University of Library Studies and Information Technologies, Bulgaria, E-mail: i.kostadinova@unibit.bg
[4] University of Library Studies and Information Technologies, Bulgaria, E-mail: e.iliev@unibit.bg
[5] University of Library Studies and Information Technologies, Bulgaria, E-mail: p.petrova@unibit.bg
[6] University of Library Studies and Information Technologies, Bulgaria, E-mail: b.nikolova@unibit.bg

The first section looks at the big data divide phenomenon; the second section reviews the compulsory competencies held by data professionals. The aim of the third section is to present a competency framework of a data specialist in conditions of big data divide.

2. The Phenomenon Big Data Divide

Critical learning of data, knowledge transfer and building a deep understanding of processes are essential for the formation of rational behavior. Retaining knowledge is another essential aspect that affects rational behavior. Today we are faced with a new round of this evolution: literacy is needed to acquire knowledge from the so-called "big data"

The framework of knowledge and skills that a modern data specialist must have continues to expand. There is constant growth in the demand of the business for qualified specialists with the necessary expertise. A new form of division in society is emerging, largely due to the lack of skills to handle available big data.

According to Christozov &Toleva [4] big data represent a new challenge which "addresses the human ability to learn from an amount of data significantly beyond the human cognitive capacity." It can be argued that it is difficult to achieve the necessary competencies to acquire literacy for working with large data. Trends in educated countries show that the young generation is withdrawing from studying topics related to data analysis such as mathematics and statistics. These generations rely on mediators – either human information brokers or computer applications as data mining tools – in coping with Big Data, usually without the necessary understanding of the limitations of application of the tools and the level of relevance of the results to the essence of the problem. This way of researching big data does not generate adequate knowledge of objects and events described by the data. Only a certain elite will be able to take full advantage of the accumulated data, understand the cause-effect relationships in the processes, which will allow the prediction of the results of the activities carried out.

According to Bharadwaj [22], IT staff skills are a critical resource for building measurable business value. The notion that companies have to combine technology with human skills traces its roots in the socio-technical framework. This enhances the role of human skills and states that maximum technological performance requires both management and nurturing of the human skills and knowledge of organizations [23]. This idea is particularly important in the context of big data because skills are not only in the exploitation of technical resources (such as software and infrastructure.) More important is the generation of a vision that guides organizational decision [24].

This gives a reason to believe in the discovery of a new form of digital divide, manifested at different levels within and outside organizations. We disclose three cases of the data science divide: (1) a division between firms that have human capital with better analytical skills and those that do not have it; (2) IT specialists who can learn from big data and others who can only manage, modify, and read them; (3) citizens who apply analytical skills and can handle big data drawing useful information and those who need a mediator to take advantage of the large data.

Studying big data can respond to key challenges and help suggest more effective evidence-based solutions and provide the ability to convert complex, often unstructured data into relevant information as a strategic response to changing global trends [9].

The role of Data Scientist in this situation is related to the creation of products from the available data that acquire their value from the data themselves [15], and the final product in turn generates even more value. Data Scientist is a multidisciplinary profile that seeks knowledge in several learning areas. This specialist relies heavily on the scientific way of doing things, so its research experience is of great significance.

Mikalef, et al. [17] note that data skills are perhaps the most sought-after resource in companies that have big data, as the skills captured by the scientists' profile allow companies to ask the right questions and convert data into practical insights. They conclude that software, infrastructure and data are insufficient to provide any value if personal skills and knowledge are not available to implement them. Such findings have been noted in a number of studies, some of which will be covered in the next section.

3. A review of compulsory competencies in the big data era

In scientific literature, there is an increasingly active participation in the presentation of the Data Scientist's professional profile and the necessary skills that this specialist must possess. The framework of knowledge, skills and competences that shape the data scientist's profile has evolved over the years. For example:

- Cleveland [6], looking at the field of competence of a data specialist broadened the field of statistics to cover data science.
- Press [19] considers Data Science as an area linking statistics and computer science.
- Carlos Costa and Maribel Yasmina Santos [7], add that “the knowledge base expected from a Data Scientist goes beyond the skills of a computer scientist, or a statistician, or even the coupling between these two.”
- The authors Costa and Santos [8] share that “in order to communicate Data Scientist findings and integrate the results into data artifacts deployed in business environments, Data Scientists must have strong social and personal capabilities, like communication, business acumen and curiosity.
- According to Suhailis [21] the skills required to hold the position of data scientist consist of the skills to: model and analyze, data processing, statistic, business domain, soft skills and technical skills.
- In her research Linda Burtch [2] emphasizes that the data scientist should possess technical (incl. Analytics, SAS, R, Python, Coding, Hadoop, SQL, and Database), as wel as non-technical skills (such as Intellectual curiosity, business acumen, and communication skills).
- Ayankoya [1] define that the data scientist is a combination of three basic areas: computer science, statistics and domain knowledge.
- Gehl [11] maintains that the best skill possessed by the data scientist is awareness of the business strategy and the function of the organization.
- Ismail [13] reduces the data scientist skills to five basic ones: business, statistics, machine studies, communications and analysis.
- According to Christozov [3] the three basic competency categories when working with Big Data are: (1) the possibility to retrieve useful data from huge and versatile repositories; (2) the possibility to check retrieved data; (3) the possibility to interpret (map) retrieved data in the context of the issue and to draw useful models, links or just raise awareness regarding circumstances related to the issue
- According to Sicular [20] data scientist is a widely-applied specialist within a variety of organizations, therefore it’s difficult to provide a complete and consistent list of required skills, but points to mandatory data storage, data analysis, data conversion and communication skills.
- According to Christozov and Toleva [4] the ability to address critical information, as well as verifying sources and considering constraints of applied technologies is a factor in generating useful knowledge from the acquired information.
- According to Manieri [16] the data scientist is an expert with the ability tomanipulate and retrieve knowledge and turn it into significant value.
- According to Ismail [13] visualization and communication skills are important because they allow those who are not professional data analysts to interpret the data.
- According to Gupta and George [12] the availability of large data skills is a potential source of competitive advantage.

In order to uncover the main barriers to data science divide at a company level, a profile of the data specialist will be presented in the next section.

4. Data science competence in digital divide conditions

Based on the existing frameworks, a summary profile of the data specialist can be created, combining all the skills discussed above. In this model, we focused on 3 generic skill categories, namely: (1) Hard skills; (2) Soft skills and (3) analytical skills.

Peculiarities of the knowledge and skills possessed in solving specific tasks are presented in table.1.

Table 1. Framework for data science – tasks, skills and competencies

	Soft skills	Hard skills	Analytic skills (hard and soft)

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; 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.

It has been proved that the Big Data Specialist should be able to write in programing languages like Python, R, Java, Ruby, Clojure, Matlab, Pig and SQL [18]. Besides, the data scientist should be familiar with the NLP, machine training, conceptual modeling, statistical analysis, predictive modeling and testing of hypotheses, working with databases. All these skills will be part of the hard skills group.
The category soft skills comprises a great deal of non-technical communication skills, organizational business strategy and understanding of the architecture of the system [13].
Alongside soft and hard skills [1], the data scientist needs to be able to use sophisticated analyses such as forex analysis, visualization and data modeling and machine training to predict what will happen in the future and make recommendations for improving the existing business process. In turn, it can be argued that analytical skills are based on hard skills as the decision making, the elaboration of strategies and the implementation of experimental research are handled using data obtained from other data based on some preliminary processing.
Most IT professionals today have the skills to handle small data. Working with big data, however, requires more than technical literacy, statistics, mathematics, programming and working with a database. The lack of soft skills as well as analytical thinking can be interpreted as the main barrier to the formation of knowledge from big data.
Based on the information presented above, we believe that technical skills can be seen as typical of IT specialists, soft skills – as typical of information mediators, and analytical thinking and analytical skills as typical of data analysts and researchers (see Fig. 1)

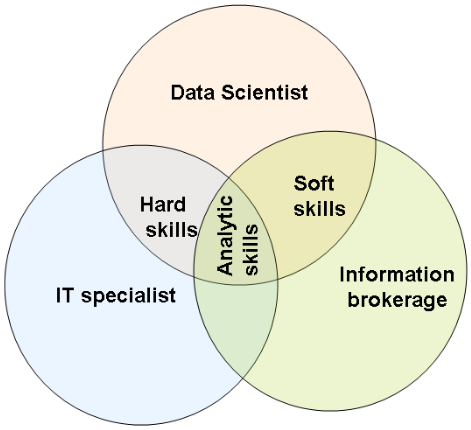

Fig. 1. Skills base in data scientist profile

Technical skills combined with soft skills in the presence of analytical thinking and analysis skills form the general framework of competencies owned by the modern data specialist.

The following dependencies are noticed: (1) The presence of hard skills is a basis for data retrieval and formation of new knowledge. (2) Without the availability of analytic skills, the retrieved data is merely converted data that has no informational value and cannot acquire knowledge useful for developing business strategy. (3) Knowing the business plan and the main directions of the company's development are an important prerequisite for implementing the right algorithm for data research and for achieving the desired results.

The presented competency model also gives us reason to assert that big data divide at company level is further increased due to the wrong selection of human capital when appointing data specialists. Database management skills, statistics and programming languages are just some of the skills you need to work with big data. The data specialist must have expertise in three categories: hard skills, soft skills and analytic skills. The non-coverage of any of the listed categories of data-handling professionals places the company at a disadvantage, and managers are in a losing position.

The Data science divide can be overcome in making the right choice of human capital. The serious problem here is the lack of trained specialists holding this expertise. This opens up new questions related to the training of data professionals. Data Science training needs to be business-oriented. This will increase the quality of the staff on the one hand, and on the other – increase the company's productivity.

5. Conclusion

The emergence of big data has opened up a new digital divide based on the shortage of data professionals with the necessary experience, knowledge and expertise. This article has reviewed the basic competencies that the data specialist must have. On the basis of the existing frameworks, a complex competency model has been deployed, dividing the skills into three groups: soft, hard and analytical skills and the profile of the data specialist has been considered as three professions: IT specialist, information manager and analyst. The training oriented to the needs of the companies and in line with the professional profile of the specialist is the first measure to overcome the data science divide.

Acknowledgment

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] Ayankoya, K., Box, P. O., Calitz, A., Box, P. O., Greyling, J., & Box, P. O. (2014). Intrinsic Relations between Data Science , Big Data , Business Analytics and Datafication, 192–198. doi:10.1145/2664591.2664619

[2] Burtch, L. (2014). 9 Must-Have Skills You Need to Become a Data Scientist. Retrieved from http://www.kdnuggets.com/2014/11/9-must-have-skills-data-scientist.html [Accessed: 24-Marth-2018].

[3] Christozov, D., Rasheva-Yordanova K… Data Literacy: Developing Skills on Exploring Big Data Applications. International Journal of Digital Literacy and Digital Competence. Volume 8 • Issue 2 • April-June 2017. DOI: 10.4018/ IJDLDC.2017040102

[4] Christozov, D., Toleva-Stoimenova S., Big Data Literacy - a New Dimension of Digital Divide: Barriers in learning via exploring Big Data, in Strategic Data Based Wisdom in the Big Data Era,

editors Girard J., Berg K., Klein D., IGI Global, 2015, ISBN13: 9781466681224, ISBN10: 1466681225, EISBN13:9781466681231.
[5] Christozov, D., Toleva-Stoimenova S., Rasheva-Yordanova K., Vukarski I. Developing Big Data Competences in the Digital Era. Big data, Knowledge and Control Systems Engineering, BdKCSE'2016. pp. 97-104. ISSN – 2367-6350.
[6] Cleveland, W. S. "Data Science: an Action Plan for Expanding the Technical Areas of the Field of Statistics. International Statistical Review, Volume 69, Issue 1, April 2001, pp 21-26. ISSN: 03067734, DOI: 10.1111/j.1751-5823.2001.tb00477.x
[7] Costa, C., M. Y. Santos. A Conceptual Model for the Professional Profile of a Data Scientist. Recent Advances in Information Systems and Technologies, Volume 2, ISBN978-3-319-56538-5. DOI 10.1007/978-3-319-56538-5_46
[8] Costa, C., M. Y. Santos, "The data scientist profile and its representativeness in the European eCompetence framework and the skills framework for the information age," International Journal of Information Management, vol. 37, no. 6, pp. 726-734, 2017.
[9] Daniel, B. Big data and analytics in higher education: Opportunities and challenges, Brit. J. Educ. Technol, 46 (2015), 904–920.
[10] Field of Statistics", Int. Stat. Rev., vol. 69, no. 1, pp. 21–26, Apr. 2001.
[11] Gehl, R. W. Sharing , knowledge management and big data : A partial genealogy of the data scientist. 2015, doi:10.1177/1367549415577385
[12] Gupta M., J. F. George, "Toward the development of a big data analytics capability," Information & Management, vol. 53, no. 8, pp. 1049-1064, 2016
[13] Ismail, N. W. Abidin. Data Scientist Skills. IOSR Journal of Mobile Computing & Application (IOSR -JMCA) e -ISSN: 2394 - 0050, P-ISSN: 2394-0042.Volume 3, Issue 4 (Jul. -Aug. 2016), PP 52-61, DOI: 10.9790/0050-03045261
[14] 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.
[15] Loukides, M. What is data science, June-2010. [Online]. Available: https://www.oreilly.com/ideas/what-is-data-science [Accessed: 25-Marth-2018].
[16] Manieri, A., Demchenko, Y., Brewer, S., Hemmje, M., Riestra, R., & Frey, J. Data Science Professional uncovered How the EDISON Project will contribute to a widely accepted profile for Data Scientists. Proseedings of 2015 IEEE 7th International Conference on Cloud Computing Technology and Science Data, p.588-593, doi:10.1109/CloudCom.2015.57
[17] Mikalef, P., I. O. Pappas, J. Krogstie, and M. Giannakos, "Big data analytics capabilities: a systematic literature review and research agenda," Information Systems and e-Business Management, pp. 1-32, 2017.
[18] Mohanty, S., M. Jagadeesh, H. Srivatsa. Big Data Imperatives Enterprise Big Data Warehouse, BI Implementations and Analytics. 1st ed., XXII, Apress, 320 p. ISBN 978-1-4302-4872-9.
[19] Press, G. A Very Short History Of Data Science, Forbes, 28-May-2013. [Online]. Available: https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/#165ec1db55cf. [Accessed: 24-Marth-2018].
[20] Sicular, S. Big Data Analytics Failures and How to Prevent Them, 1(August).
[21] Suhailis, A., Garis Panduan Data Raya Sektor Awam, 2016.
[22] Bharadwaj, A. "A resource-based perspective on information technology capability and firm performance: an empirical investigation," MIS quarterly, vol. 24, no. 1, pp. 169-196, 2000.
[23] Wamba, S. F., A. Gunasekaran, S. Akter, S. J.-f. Ren, R. Dubey, and S. J. Childe, "Big data analytics and firm performance: Effects of dynamic capabilities," Journal of Business Research, vol. 70, pp. 356-365, 2017.
[24] Schoenherr, T., C. Speier - Pero, "Data science, predictive analytics, and big data in supply chain management: Current state and future potential," Journal of Business Logistics, vol. 36, no. 1, pp. 120-132, 2015.