ANALYSIS OF DATA SCIENCE COURSES THROUGH THE PRISM OF THE DIGITAL DIVIDE

I. Kostadinova, P. Petrova, E. Iliev

University of Library Studies and Information Technologies (BULGARIA)

Abstract

Data Science is the science of pooling, processing, and analysing vast streams of (non-structured) data in order to understand and analyse events related to them. In a particular situation, large volumes of data have to be processed in order to derive certain results. Regardless of the field of activity and the type of specialists, working with data is a valuable ability to acquire knowledge. Data Science offers advanced approaches to discovering such knowledge and forecasting.

This article examines various data science trainings worldwide. The scope of the courses, the teachers, the duration and the way of assessment are part of the surveyed indicators.

Data training is typical for IT professionals, but it is also needed by professionals from all other areas. The needs of different user groups, potential learners in these courses have been defined for the purpose of overcoming digital divide. On the basis of this analysis, appropriate learning content is proposed for the different user groups.

Keywords: Analysis, digital divide, data science, learner groups, training.

1 INTRODUCTION

Over the last few years, several new concepts have entered the IT world, behind which are significant technologies that give a new look to the modern world. Massive digitization and the use of modern ICT have led to a different lifestyle and handling of information. Modern digital devices (personal and industrial) have become a convenient way to collect huge amounts of large data (big data), and the skills to process and analyse have proved to be key in the desire for successful development of any organization. Large arrays of structured and unstructured data are so voluminous and complexed that they can not be collected, selected, processed, or managed through widely used software tools within acceptable time frames. Data science creates mechanisms to analyse this data and capture the core patterns, and then programs these models (or algorithms) into computer applications [6] [7].

Over the last few years there has been a huge interest to the organizations in the area of big data and data science. There is a huge need for data science specialists who can not only extract the useful knowledge from the data but also be able to judge how to use this knowledge. Needed is a complex of skills to get the data specialist to get the knowledge from the processed data. The framework defining the knowledge and competencies that data professionals must possess continues to grow. There is a Big Data Divide, which can be considered as a particular case of digital divide, which has so far been seen as a skill and inability to handle ICT [3].

This paper aims to analyse offered courses of data science in the world. The first part examines the world-wide courses, including the learning material, the lecturers, the duration and the way of assessment. The second part examines the potential users of the courses studied according to the topics set in the training material. The needs of different user groups, potential learners in these courses are defined for the purpose of overcoming Big Data Divide. In the third part, on the basis of the analysis, we offer appropriate learning content for the different user groups.

2 DATA SCIENCE LEARNING COURSES AROUND THE WORLD

Worldwide, a variety of data science trainings are available. There are a large number of both paid and free courses and lessons that a motivated person could use as a springboard for a rewarding and lucrative career. They are organized both traditional at the place and remotely, or fully on-line. There are also training organizations offering training materials for self-study against registration, and they also provide an opportunity for certification of the learners.

As K.Rasheva - Yordanova says in his paper Forming of Data Science Competence for Bridging the Digital Divide [16], the construction of an IT specialist, Information brokerage and Data scientist is carried out by constructing differently for each of them the mastery of Hard skills, soft skills and analytical skills (hard and soft). According to fig.1 Skills base in data scientist profile [16], the data science learning and the formation of Data Scientist should include all three skills.

In this paper, we will look at the current data science learning courses in the world in terms of their duration, the scope of the learning material, the lecturer's qualifications and experience, and the evaluation methods used. As a result of this review, it is expected to build up an insight into what specialists are expected to be produced for the labor market and what they will be able to do.

The review includes educational programs from several universities offering data sciences learning [14] which, according to Olivia Krauth, are among the top 10 of University for this specialty, as well as several other smaller universities.

Because of the volume limit of the paper, Table 1 contains a review of the data of only 7 training programs and 7 qualification courses.

Looking at the indicators of the duration of the learning course, the scope of the learning material, the qualification and the experience of the lecturers and the evaluation methods used, we aim to build a complete picture of the current data science learning courses.

A review of the scope of the curriculum aims to show how well the offered learning courses and learning programs at the University are able to provide training that shapes these skills. Also, whether there is a difference between the teaching material offered by the courses and teaching material that is offered by the Universities.

The review of lecturer qualifications will give an insight into the extent to which the teaching material is tied to practical examples. The transfer of the personal practical experience of the teachers reflects the way in which the learning content and the degree of learn by the learners are presented.

Also is considered which training courses have the minimum requirements for course enrollment and what are these requirements. Minimum-requiring courses imply experience in the field before the course starts.

Consideration of the duration of training courses is not a significant factor but rather an orientation. Typically, the education programs last 3-4 years for bachelor degrees or 1-2 years for a master's degree. In training courses, the form and distribution of lessons is an important part of determining the duration of the course. Here, the duration can be as much as 1-2 months for accelerated courses and up to 1 year for Saturday-Sunday courses.

There are the Universities like CMU [2], which takes an interdisciplinary approach to data science, offering multiple master's degrees with different focuses and tracks. Top options for big data include computational data science and information systems management, with a focus on business intelligence and data analytics. With most its programs lasting two years, tuition varies by program and by semester. Other universities allow only one or several courses to be chosen for learning [11].

Table 1. *Comparative analysis of active training of data science around the world.*

	COURSES	PREREQUISITES:	SCOPE OF THE LEARNING MATERIAL	Qualification and experience of the lecturers	LEARNING DURATION	WAY OF EVALUATION
QUALIFICATION COURSES	Principles of Data Science in University of Cambridge [20]	Mathematical knowledge of linear algebra, calculus, optimization, probability and statistics, some experience with at least one language or package to handle data analysis.	Statistical Learning; Linear Regression; Classification; Resampling Methods; Linear Model Selection and Regularization	Dr XXXXXXXX from Computer Laboratory, University of Cambridge	16 hours	Literature survey (50% of final mark) Practical project (50% of final mark)
	Intro to Data Science - Udacity[19]	Python programming experience or understanding of concepts such as variables, functions, loops, and basic python data structures like lists and dictionaries	Data Manipulation; Data Analysis with Statistics and Machine Learning; Data Communication with Information Visualization; Data at Scale -- Working with Big Data	Instructor	2 months	Quizzes
	Approximation, sampling and compression in data science - Isaac Newton Institute for Mathematical Sciences of Cambridge [12]	No	Challenges in optimal recovery and hyperbolic cross approximation; Mathematics of data: structured representations; Approximation, sampling, and compression in high dimensional problems	Several University lecturers	5 months	----------
	Data Science Specialization - Jhon Hopkins University [13]	No	The Data Scientist's Toolbox; R Programming; Getting and Cleaning Data; Exploratory Data Analysis; Reproducible Research; Statistical Inference; Regression Models; Practical Machine Learning; Developing Data Products; Data Science Capstone	Professor and Associate professor, Biostatistics	10 months	Project
	Data science Software University – Bulgaria [17]	previous experience in Python programming, required is to take the course math concepts for developers	Data collection; Data cleaning and preparation for analysis; Applying the scientific method to real issues and problems; Analyzing and visualizing data; Basics of data modeling; Build a complete application: from raw data to decision making;	programmer and lecturer	2 months	Project

QUALIFICATION COURSES	Social Data Science (master) Oxford Internet Institute [15]	English language Higher level	Foundations and Frontiers of Social Data Science; Applied Analytical Statistics; Research Methods for Social Data Science; Foundations of Visualization; Special topics in Research Design; Python for Social Data Science; Data Analytics at Scale; Machine Learning; Experiments for Data Science; Human and Data Intelligence; Statistical Analysis of Networks; Sociological Analysis; Survival Analysis	Several University lecturers; several data scientists and researchers	10 months	dissertation of up to 15,000 word
	Microsoft Professional Program for Data Science [11]	No	T-SQL;Analyzing and Visualizing Data with Excel; Power BI; Codding with Python and R Azure Machine Learning; Use Code to manipulate and Model Data; Implementing Predictive Analytics with Spark in Azure HDInsight or Spark	Lecturer s are Associate Professor, Biostatistics; Program Manager in Microsoft; Lead Instructor; Senior Content Developer;	10 course/ 16-32 hours per course for 3 month Important: *you can choose even one of 10 courses*	Practical project and certificates
DEGREES AND SPECIALIZATIONS IN HIGHER EDUCATION	Analytics (Master) in American University [1]	Bachelor degree/not specified in what /; Work experience	Designed for Working Professionals; sharing experience of professionals	Professor and Academic Director, MS Analytics	5 - semester	project
	Data Science (Master) Stanford University [18]	Bachelor Degree /not specified in what /	Foundational; Data Science Electives; Advanced Scientific Programming and High-Performance Computing Core; Specialized Electives;	Research Scientists & Lecturers	2 years	Practical Component
	Information and Data Science University of California Berkeley [21]	Bachelor degree /not specified in what /	Research Design; Data Engineering; Machine Learning; Mining and Exploring; Data Visualization; Ethics and Privacy; Statistical Analysis; Communicating Results	Professors	3 semesters : three paths: full time, accelerated, or part time	---------
	Data Science- Varna Free University of Bulgaria [5]	Bachelor degree /not specified in what /	Probability and Statistics; Computer Architecture And Operating Systems; Data Structures And Algorithms; Programming; Databases; Data Analytics; Methods; Programming For Data Science **(Python);** Distributed And Cloud Computing; Data Mining; Algorithms; Machine Learning; Social Network; Analysis; Data **Quality;** Cryptography;	Professors, Assoc. Prof.; Project Manager; IT Business Consultant; Senior Consultant with hands-on experience in BIS and ETL on international projects	1 year	Dissertation of 80 pages

DEGREES AND SPECIALIZATIONS IN HIGHER EDUCATION	Massachusetts Institute of Technology/ Business Analytics (Master) [10]	Bachelor degree /not specified in what /	The Analytics Edge; Applied Probability and Stochastic Models; Analytics Capstone; Optimization Methods Analytics Lab Machine Learning From Analytics to Action and 3 from list of 48 approved electives	Professors ,	1 year	Project
	Northwestern University/ Analytics (MSIA) and Predictive Analytics (MSPA)- [8]		Predictive analytics;Java & python programming; business communication and analytics consulting; data visualization; analytics for big data; deep learning; Healthcare Analytics, Predictive Models for Credit Risk Management, Optimization & Heuristics, or Social Networks Analysis.	Professors	15-month professional master's degree	Professional Practicum + Project
	New York University / Data Science (Master and Ph.D.) [9]	strong knowledge in mathematics, computer science, and applied statistics; bachelor and Standardized Tests	Development of new methods for data science; Machine Learning; Big Data; Regression & Multivariate Data Analysis; Fundamental Algorithms; Inference and Representation	----------	2 years	Project

Analysing the data in Table 1, several conclusions can be drawn:

- Most qualification courses do not demand a learning start.
- The distribution and duration of the learning courses is tailored to the employed people. Courses are organized more concisely or on days off so as to be convenient for those employed.
- Educational programs at Universities require prior knowledge and education, although only a few of them contain specific initial knowledge as a requirement.
- The learning courses included in the qualification courses are of a mandatory nature. Some universities offer a list of elective courses until the necessary credits have been accumulated.
- The training in the syllabuses is conducted exclusively by scientists – mainly professors, while in the training courses there is a combination of scientists and practitioners.
- Organized courses and training programs generally do not check with what initial skills and knowledge will be start the learners. Only one of the listed learning courses is conducting an incoming test to check the student's incoming knowledge.

3 WHO NEEDS TO BE LEARNED IN DATA SCIENCE OR POTENTIAL USER GROUPS

The preparation of each training course or program aims to teach specialists certain needs of the economic market. On this basis, we can state that the training courses and programs currently offered are intended to meet the needs of the data scientist and the training material in them aims to train such experts. The availability of multiple data science training courses in any training format and duration speaks of a 'hunger' in the labor market for such professionals.

In the article Forming of Data Science Competence for Bridging the Digital Divide, Rashev - Yordanova et al. [16] talks about a new digital divide, namely data science divide, forming three cases of data science divide according to this criterion. (1) a division between firms with better human capital and those with no analytical skills; (2) IT specialists who can learn from big data and others who can

only handle 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 big data.

The review of current global data science training courses in our view, also creates an idea of who really needs to learn data science on the basis of the offered courses. Based on this, two more data science divide cases can be distinguished according to the skills of the potential learners, namely:

- active data science specialists who have knowledge and are involved in such an activity, but need additional qualifications, to refresh knowledge or just need certification to maintain their current position (Group A).
- Those willing to study data science – users who do not have any basic skills, data science non-specialists but want to develop in this area (Group B).

Reviewing at Table 1 from the point of the data science divide, we can say that for newly formed groups it is advisable to target specific types of courses or training programs in view of the goals of each learner. According to Table 2, for current Data Science specialists (Group A), it is advisable to enroll on a specific course rather than starting training from the beginning. In this way, they will improve or refine their skills in a relatively short time. All the same, users in this group can choose a course or a program that is for beginners. They will have the opportunity to go through the whole data science training, which will give them a refresh of the old knowledge and learning the new in the field. As a negative, however, starting this kind of training wants a considerable time.

On the other hand, those who want to start and develop in the field of data science (Group B) are advised to begin full-time training in a program that will give them a complete insight into data science and its specifics.

Table 2. *Appropriate learning to the potential user groups.*

Learners Group	DEGREES AND SPECIALIZATIONS IN HIGHER EDUCATION	QUALIFICATION COURSES
ACTIVE DATA SCIENCE SPECIALISTS (GROUP A)	• Appropriate, but time consuming	• Appropriate, recommended
THOSE WHO WILL TRAIN IN DATA SCIENCE (GROUP B)	• Appropriate, recommended	• Inappropriate, • limited scope

4 CONTENT ANALYSIS AND PROPOSAL FOR SELECTING APPROPRIATE LEARNING CONTENT FOR DIFFERENT USER GROUPS

Based on the review and analysis of the previously described training programs and courses, there are some courses that are found in every form of data science learning. These are Databases system and SQL, Programming with Python or R, Statistics and Machine Learning, Big Data, Data Analytics Methods; Regression Models, Developing Data Products and Data Analysis. Various additional courses, such as Human and Data Intelligence, Data Structures and Algorithms, Analysing and Visualizing Data with Excel; Power BI, Use Code to manipulate and Model Data, and others that complement and target learning. There is the ambition of the various educational institutions to shape programs and courses that are attractive to future learners so as to attract more of them. However, this does not change the essence of data science training. On the basis of the formed groups of learners, Table 3 proposes training courses and programs with appropriate content to the learners' skills and knowledge.

Table 3. *Proposal for selecting appropriate learning content for different user groups.*

Learners Group	DEGREES AND SPECIALIZATIONS IN HIGHER EDUCATION	QUALIFICATION COURSES
ACTIVE DATA SCIENCE SPECIALISTS (GROUP A)	----------	✓ Contemporary business models with big data ✓ Development of new methods for data science; ✓ Big Data; Regression & Multivariate Data Analysis; ✓ Business communication and analytics consulting; ✓ Data Mining ✓ Sociological Analysis ✓ Developing Data Products ✓ Mathematics of data structured representations ✓ Approximation, sampling, and compression in high dimensional problems; ✓ Applying the scientific method to real issues and problems ✓ Data Analytics Methods ✓ Human and Data Intelligence
THOSE WHO WILL TRAIN IN DATA SCIENCE (GROUP B)	✓ Basics of data science (including disciplines Data at Scale -- Working with Big Data; Fundamental algorithms, databases; machine learning; data collection; getting and cleaning data; statistics; Basics of data modeling, Research And Development Internship	✓ Basic in programming with Python(R); ✓ Fundamental algorithms; ✓ Machine Learning; ✓ Analysing and visualizing data ✓ Data collection; ✓ Research Methods for Social Data Science ✓ Developing Data Products ✓ Getting and Cleaning Data ✓ Statistics and analysing ✓ Data Analytics Methods ✓ Basics of data modelling ✓ Research and Development Internship

According to the distribution in Table 3, active data science specialist will be able to refresh their knowledge and retrain through training courses in a field they believe they need. This is largely possible now because many Universities offer courses from their learning programs as stand-alone for study through various MOOC platforms like Coursera, eDX and others.

For those who will train in DATA SCIENCE (group B), it will be given the opportunity to learn about the possibilities of data science if its enrol in the offered "Basic of data science" curriculum. Individual courses may be offered in parallel with this program (see Table 3).

However, in order to determine the right set of training courses and to build a data science training program, it is necessary to research what search the companies as a "data scientist specialist" and, on that basis, to place the learning disciplines.

5 CONCLUSIONS

The need for data science specialists has spawned a wide variety of courses and trainings in the world. Compared to traditional forms of training, online courses are also available, organized in

various time periods and intervals. Some Universities, alongside their data science learning programs, also offer self-study in a specific course in this program.

The market for data science education is extremely rich and diverse. Organized courses and training programs generally do not check with what initial skills and knowledge will be start the learners. Different levels of learners' initial knowledge lead to demotivation and stopped the learning of learners with low knowledge. Different levels of learners' initial knowledge have an impact on the rate of transmission of learning content and the number of graduates in data science.

In order for the content of a learning course to be correctly defined, it is necessary to consider what initial knowledge and skills the learners start to take. Accordingly, proposals have been prepared for the selection of appropriate learning content for the different user groups and are shaped according to the learners' initial skills. Having basic skills and knowing what the industry is looking for as a "data scientist", courses and programs can be developed that not only attract more learners but also produce more and more skilled professionals.

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] American University/Analytics (Master) – [Online resource] 8.05.2018- https://www.american.edu/kogod/graduate/analytics/

[2] Carnegie Mellon University - [Online resourse] 8.05.2018-https://www.cmu.edu/graduate/data-science/index.html

[3] 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.

[4] Data Science and Online Education By:Fox, G (Fox, Geoffrey)[1] ; Maini, S (Maini, Sidd)[1] ; Rosenbaum, H (Rosenbaum, Howard)[1] ; Wild, D (Wild, David)[1] Book Group Author (IEEE 2015 IEEE 7th International conference on cloud computing technology and science (CLOUDCOM) Book Series: International Conference on Cloud Computing Technology and Science Pages: 582-587 DOI: 10.1109/CloudCom.2015.82 Published: 2015.

[5] Data Scence(Master) in Varna Free University - [Online resource] 6.05.2018- http://datascience.vfu.bg/

[6] Delaney Connie W., Charlotte A. Weaver, Judith J. Warren, Thomas R. Clancy, Roy L. Simpson. Big Data-Enabled Nursing: Education, Research and Practice//Springer, 2017, pp.488 DOI 10.1007/978-3-319-533000-1, ISBN 978-3-319-53299-8.

[7] Hayashi, Chikio (1998-01-01)."What is Data Science? Fundamental Concepts and a Heuristic Example". In Hayashi, Chikio; Yajima, Keiji; Bock, Hans-Hermann; Ohsumi, Noboru; Tanaka, Yutaka; Baba, Yasumasa. *Data Science, Classification, and Related Methods*. Studies in Classification, Data Analysis, and Knowledge Organization. Springer Japan. pp. 40–51. doi:10.1007/978-4-431-65950-1_3. ISBN 9784431702085.

[8] Northwestern University/ Analytics (MSIA) and Predictive Analytics (MSPA) - [Online resource] 7.05.2018- http://www.mccormick.northwestern.edu/analytics/curriculum/

[9] New York University / Data Science (Master and Ph.D.) - [Online resource] 6.05.2018- https://cds.nyu.edu/academics/ms-in-data-science/

[10] Massachusetts Institute of Technology/ Business Analytics (Master) - [Online resource] 8.05.2018- http://mitsloan.mit.edu/master-of-business-analytics/

[11] Microsoft Professional Program for Data Science - [Online resource] 7.05.2018- https://academy.microsoft.com/en-us/professional-program/tracks/data-science/]

[12] Isaac Newton Institute for Mathematical Sciences of Cambridge/ Approximation, sampling and compression in data science - [Online resource] 5.05.2018 - https://eu.udacity.com/course/intro-to-data-science--ud359

[13] Jhon Hopkins University/ Data Science Specialization - [Online resourse] 5.05.2018- https://www.coursera.org/specializations/jhu-data-science

[14] Krauth, O. Photos: The top 10 universities for data science. // TechRepublic/Big data – online resource] 05.05.2018 https://www.techrepublic.com/pictures/photos-the-top-10-universities-for-data-science/1/

[15] Oxford Internet Institute/Social Data Science (master) - [Online resource] 8.05.2018 https://www.ox.ac.uk/admissions/graduate/courses/msc-social-data-science?wssl=1

[16] Rasheva-Yordanova, K., Chantov V., Kostadinova I., Iliev E., Petrova P., Nikolova B. Forming of Data Science Competence for Bridging the Digital Divide. 8th edition of the "The Future of Education" conference, PIXEL, Florence, 2018.

[17] Software University Bulgaria/Data science - [Online resource] 7.05.2018 - https://softuni.bg/trainings/1919/data-science-june-2018

[18] Stanford University/ Data Science (Master) - [Online resource] 7.05.2018 - https://statistics.stanford.edu/academics/ms-statistics-data-science

[19] Udacity/ Intro to Data Science [Online resource] - 8.05.2018 - https://eu.udacity.com/course/intro-to-data-science--ud359

[20] University of Cambridge/ Principles of Data Science - [Online resource] 9.05.2018 https://www.cl.cam.ac.uk/teaching/1516/L120/

[21] University of California Berkeley/Information and Data Science - [Online resource] 8.05.2018 - https://www.ischool.berkeley.edu/programs/mids