DESIGNING DATA SCIENCE CURRICULUM IN A WAY TO ADDRESS EXPECTED STUDENTS ENTRY COMPETENCES

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

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

Abstract

Success of a training program depends on how well the program allows students to enter the program in a smooth and easy way. The low percentage of retaining students to great extend depends on the way how the curriculum address difficulties students face when they start the study. Smooth introduction to the field of study and encouragement to accept and benefit of offered challenges, natural to every new endeavor, is essential in designing new curriculum. This is especially important in designing curriculum to train students in a new, immature area of education.

The paper is dedicated to share our experience in designing a Data Science Master Program with emphasis on potential students' portfolio. Designing the curriculum was done in three phases:

1 Defining the mission, objectives and scope of the program, according to market study of the needs of industry.

2 Empirical study of potential students' competences acquired on their bachelor study and related to the objectives of the program as defined on the first stage.

3 Designing curriculum in a way to achieve program mission of training specific for Data Science competences.

The mission, objectives, and scope of the program were defined by obtaining feedback shared by ULSIT alumni, mostly graduated in Information Brokerage Bachelor and Master Programs. The objectives of these programs are to train professional in the information related fields. Students obtained competences serving as mediators between end users and information resources. This includes variety of skills from understanding and comprehends classical librarian issues, via expertise of using information technologies, especially to organize information resources within a social entity, toward analytical and presentational skills. Students are prepared to serve in support of consulting teams, as well as managers of information resources. This background allows alumni to obtain a broader view on the needs of variety of industries in the area currently marked as Data Science.

The target group of potential students in Data Science Master Program was defined as undergraduate students, studying information related fields. The natural target was ULSIT students in majors as Information Brokerage, Computer Science, and Information Technologies. These three disciplines provide the needed diversity of background experience. A survey was designed to highlight analytical skills, identified as the key component of Data Science competences, and implemented in the Fall semester of 2018. About 200 students were approached, and 187 responded. Details of questionnaire, methodology of conducting the survey, and obtained results are presented.

Analysis of the profile of a typical candidate for study in the Data Science Master Program doesn't show needs of adjusting the initial composition of courses to cover the content identified as essential for the program, but emphasize on pedagogical issues and on the way how to organize and submit specific topics, how to order them, and how to monitor students' progress toward desired competences. Special attention is given on students' outcome assessment, which includes both demonstration of theoretical knowledge with close to real life practical problem solving.

Keywords: Curriculum model, Data Science, Competences, Analytical skills.

1 INTRODUCTION

Developing a new degree program is always a challenging project. From point of view of theory of decision making, this is a strategic, long term, highly risky decision made under a set of uncertain issues. This is especially true for a new immature area of knowledge as Data Science. There are numerous studies identified the growing demand for "Data Scientist" in practically every industry (see for example [5]), as well as many studies identifying the competences in data analytics the industries are looking for

([1], [6]). Also many universities are targeting this educational niche ([2], [7]). The area of Data Science emerged recently as a result of fast development of Information Technologies (IT) providing facilities to capture, record, and access data in unknown before "volume", "variety" and "velocity" magnitude according to Gartner's 3V definition [3]. The term Big Data best illustrates the state the industry had reached at the beginning of this millennium. But Big Data also challenged the industry in searching the way to benefit from those valuable resources and the job of "data scientist" emerged when this need was recognized. The role of such professional is to make the potential value of accumulated and continually accumulating data into real, usable value. The needed competences are quite diverse, including how to organize data storages in a way to allow efficient access and summarize (data warehouses), to allow extraction from data useful patterns and relationship (data mining), to develop tools to allow automatic smart reaction (machine learning), and many other knowledge and skills areas. Last, but not least to apply all these techniques to particular domain.

From one side, the competences covered by the field "Data Science" are diverse and the discipline can be explained by terms as multidiscipline or trans-discipline ([4]). From the other side the growing job market for data scientists defines the demand for proper education. Potential students are coming from different domain areas, with different background and a great majority of them assumes this step as a risky project. These both aspects create opportunity for building a successful master program, attracting diverse audience and training students in synergy gaining value through cooperation. But also challenges educators. Overestimating one of the aspects may discourage particular group of students. This may cause two risky effects - retention rate could be high and the beneficial diverse of backgrounds could be lost.

This paper is dedicated to the design of Data Science Master Program in a way to mitigate the risk of discouraging students entering the program. The study includes a survey among undergraduate students who are potential candidates for the master program. Students from majors as Computer Science, Information Technologies, and Information Brokerage, who have passed training emphasizing different aspects of IT, were the primary target, but also students from other non-technical areas as Library Studies and Public Communications were approached.

The paper is organized in the following way: in the next section the leading assumptions are explained and discussed; further we share the approach used to collect data regarding expectations and preparedness of potential students for the Data Science master program. Results of survey are presented in the following section with discussion on the implication on program design.

2 BACKGROUND AND ASSUMPTIONS

The project to design the Master of Data Science program was done in three steps:

1 By studying the literature and discussing the issue with potential employers the mission, major objectives, and the scope of the program were identified. The mission was defined as "educating professionals capable to lead and contribute to organizing and exploring information resources". Objectives include developing competences for analytical data processing; organizing, structuring and implementing Big Data repository; coordinating and directing heterogeneous innovation teams; and efficient use of contemporary IT. The objectives defined the scope of the program.

2 Another objective in designing the program was to allow smooth entrance of students with heterogeneous background obtained in their undergraduate study to allow training tolerance and the soft skills needed. The required competence expected for a successful Data Scientists to possess was analytical and abstract thinking. Profiling potential students from this perspective will allow designing the program in a proper way. Empirical study among undergraduate students majoring different technical and non-technical disciplines was done to identify students ability to analyze data; to look critically on data by distinguishing facts from opinions, understanding and avoiding misleading and misinterpretation; and to make data driven decisions avoiding personal bias.

3 The original design of the program, including identified content and courses were reviewed in the light of obtained results and adjusted in a way to meet the two set of objectives. First to train the identified technical competences to deal with Big Data and second to develop deep understanding about potential risks based on poor interpretation of results obtained via data processing computer applications. The adjustments were than on two levels. On the level of curriculum design redefining the order of offering courses and adjusting the road-map of building the

targeting competences. And on the pedagogy of delivering particular course, having in mind its place in the curriculum.

3 METHODOLOGY

The survey was done in the Fall 2018. About 200 students on their senior year were approached by an on-line questionnaire. Students were originally classified according to their majors as possessing "technical" and "non-technical" background. The essential part of the questionnaire required a student to answer 20 multiple choice questions indicated four categories of competences (five questions per category):

1 Identifying analytical skills - checking ability to identify similarity, difference, progression in a sequence, contradiction (AS);

2 Identifying abstract thinking – answering by following definition regarding the objects instead of intuitive well known attributes (AT);

3 Distinguishing between facts and opinions (FO);

4 Quantitative reasoning – ability to compare quantities and to apply basic mathematical facts as, for example, relationship between the radius and the circle length (QR).

The questions were randomly mixed among different categories.

Data processing includes three steps.

1 In the preliminary processing the inconsistent or invalid responses were removed.

2 Next every student was described according to the number of correct answers in every of this category. For example, if a student X answered correctly on two out five questions in the AS category, two out of five in AT, five of five in FO, and one of five in QR, the student X = {AS {2}, AT (2), FO(5), QR(1)} is represented as the vector X={2,2,5,1}.

3 The profile of an expected student (generalized) for the program was defined as the average of each of these categories. Or the profile of expected students' entry competences is represented by vector SP={AS,AT,FO,QR}.

This is an oversimplified presentation, but for the purpose of the study it serves well in solving the two tasks in designing curriculum – sequencing the courses and what pedagogy to apply in knowledge delivering and training, especially in the first semester courses.

4 RESULTS

4.1 Summary of Survey

About 200 students were asked to fill the online questionnaire and 187 responded:

- 18 don't respond to more than half of questions, and their responses were removed;
- Other 3 have provided invalid data.

The total number of responses used is 166. Answers of students with technical background were 123, and the "non-technical" – 43. An unanswered question is considered as wrong. Table 1 includes the summary of results.

Table 1. Summary of survey

Competence Category	Total	AS	AT	FO	QR
Technical	123	2.85	2.78	2.56	2.92
Non-technical	43	3.12	2.02	3.43	2.24
Total	166	2.92	2.38	2.93	2.53

Some of these results were a bit surprising, but they justified one of the assumptions that both categories of potential students are capable to study in the Data Science Master program and to become successful data scientist.

4.2 Curriculum Design

The initial composition of the program, as defined according to the industry experience and expectations shared in an informal way by alumni, graduated the major of Information brokerage, and serving as information officers in different companies, is presented in Appendix 1. The program is designed for three semesters when the third one is dedicated for the master thesis.

The background of potential students, especially those graduated in the Information Brokerage bachelor program, includes soft-skills competences, especially communication skills, needed to identify problems and to obtain information via encouraging sharing by individuals in a face-to-face conversation. The program emphasizes the technical aspects of designing and implementing computer mediated exploration of data, which corresponds to the profile of expected students.

The results of the survey (Table 1) supported also the initial structure of the program and how that program was scheduled to support gradual mastering of the required knowledge and skills. The major adjustment was done on the micro level – on the level of organizing and delivering given courses. The syllabi of all courses for the first semester were adjusted in a way to support smooth entrance and gradual progress of students with relatively low Analytical Thinking and Quantitative Reasoning. In particular:

- The course “Introduction to Data Science”, starts with several case studies emphasizing the importance and challenges in making data driven decisions by avoiding “common sense” bias. During the second half of the semester further understanding regarding the scope of Data Science is building on top of accumulated intuition by structuring the knowledge addressing different aspects of exploring Big Data.
- The course of “Statistics” was completely revised by moving always from pure mathematical way to present statistical techniques, emphasizing the practice and ability to infer and stressing on interpretation of results obtained by applying given statistical technique having in mind the given properties of data. For example sensitivity to the level of correlation between variables, linearity, etc.
- The course of “Cloud Computing” introduces students not only to this modern way to store data, but mostly to challenges of organizing data centers to store heterogeneous data objects, security and issues related to data protection and ethics in dealing with personal data.
- This “Cloud Computing”, together with the course of “Data Warehouses” stresses on exploring high volume of multidimensional, heterogenic data. The course introduces the importance of the hierarchy of concepts and why generalization is important to have a better view on data. Emphasis is given on building understanding regarding importance of abstract view on given problem and techniques to escape from particularities of given circumstances.
- “Data Mining” course also move always from algorithmic representation of learning techniques, rather addressing interpretation of results and practice in using data miners. The course introduces students to variety of open-source miners, stressing on developing learning skills in mastering the use of specific data analytical technologies.
- The course of “Visualization” includes couple of components addressing cognitive and psychological aspects of learning.

In this way students will step on their soft-skills background to acquire the lacking hard skills competences, which are addressed with more rigor in the following semester.

All courses in the entry semester emphasize on solving practical problems and on accomplishing practical task, where solutions require diverse background, different opinions and there is no single and obvious “correct” solution. The cases are designed to provoke discussion and helps further developing of communication skills and ability to listen and respect opinions of others in a diverse group.

5 DISCUSSION

Such requirements are quite challenging for instructors. Standard textbooks are not suitable for this style of training, as well as most of the available teaching resources. From certain perspective, making a preliminary study regarding expected students' entry competences, exposes difficulties and may generate resistance among instructors, who have to invest significant efforts in designing their courses, without having the opportunity to follow a proven experience.

From program success point of view this approach reduces mitigates the risk of designing the courses in unsuitable for the particular audience way.

6 CONCLUSION

The paper shared experience in addressing the challenges of the day faced by educational institutions in designing curriculum for an emerging scientific field as Data Science. In a nut-shell those challenges are associated with the inevitability of migration from a narrow professional, but rigid training, based on knowledge transfer, toward training competences allowing flexibility in adapting to emerging vulnerable circumstances, life-long learning, and agility. The major problem was to find the right balance between theory and practice to ensure smooth learning progress of students with certain profile in acquiring the needed competences. The Data Science program has to prove the vitality of its design by the performance of students in their professional career.

The Data Science Master Program needs to establish itself into the educational market by breaking the bias of potential students regarding their ability to master demanding technical expertise. Also the bias of the industry regarding expertise acquired in the program may interfere professional career students and to prove potential of selected approach. Inertia of expectations of what higher education has to do is the major stumbling stone in introducing new and nonstandard programs.

ACKNOWLEDGEMENTS

This work has been partially 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] D. Christozov, K. Rasheva-Yordanova, S. Toleva-Stoimenova, Analytical Comptences in Big Data Era: Taxonomy, Proceedings of ICERi2018 Conference, 12th-14th November 2018, Seville, Spain. ISBN: 978-84-09-05948-5, pp 7182-7191, 2018.

[2] D. Christozov, S. Toleva-Stoimenova, K. Rasheva-Yordanova, I. Vukarski Developing Big Data Competences in the Digital Era. Big Data, Knowledge and Control Systems Engineering (BdKCSE'2016), Institute of Information and Communication Technologies, Bulgarian Academy of Sciences, 1 December 2016, pp. 97-104. ISSN – 2367-6350.

[3] D. Laney, 3D Data Management: Controlling Data Volume, Velocity, and Variety, Gartner, file No. 949. 6 February 2001, Retrieved from https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

[4] G. Lotrecchiano, S. Misra. Transdisciplinary Knowledge Producing Teams: Toward a Complex Systems Perspective, Informing Science: The International Journal of Emerging Transdiscipline, Volume 21, pp. 51-74., 2018.

[5] L. Columbus, IBM Predicts Demand For Data Scientists Will Soar 28% By 2020, 2017 Retrieved from https://www.forbes.com/sites/louiscolumbus/2017/05/13/ibm-predicts-demand-for-data-scientists-will-soar-28-by-2020/#4079c5d27e3b

[6] Y. Demchenko; A. Belloum;T. Wiktorski, EDISON Data Science Framework: Part 1. Data Science Competence Framework (CF-DS) Release 2, p. 59, 2017, July 3, (Version Release 2). Zenodo. http://doi.org/10.5281/zenodo.1044346

[7] Y. Demchenko, The Emerging Role of the Data Scientist and the experience of Data Science education at the University of Amsterdam, in: LEARN Toolkit of Best Practice for Research Data Management, pp. 105-115, Retrieved from http://discovery.ucl.ac.uk/1546592/.

APPENDIX 1

Data Science Master Program: Curriculum

Semester	Course
I semester	Introduction to Data Science
	Statistics
	Introduction to Cloud Computing
	Data Warehouses
	Data Mining
	Data Visualization and computer Ergonomics
II Semester	Big Data Analytics
	Architectures of computer application
	Computer forensic
	Behavioral Economics
	Management of Information Resources: ERP systems
	Data Management in Public Entities
III Semester	Internship: Real Data Research Project
	Writing Master Thesis