RISKS MANAGEMENT IN DATA SCIENCE TRAINING

Dimitar Christozov, American University in Bulgaria, dgc@aubg.edu

Katia Rasheva-Yordanova, Stefka Toleva-Stoimenova,

State University of Library Studies and Information Technologies

Summary

Contemporary business is challenged by the phenomenon of Big Data. Utilization of opportunities provided by recent advancement of Information Technology reached the point when all data describing what happened in the business are recorded. The phrase "We are drowning in information but starved for knowledge" is the nowadays reality for the majority of business entity. The term Data Science appeared to mark competences needed to explore Big Data in a way to understand better the cause-and-effect that drives the processes and behavior intermediated by computer and communication technologies.

The paper objectives are to open a discussion regarding the risks' management within projects aimed to create the infrastructure allowing the business to benefit from accumulated data – Data Science Business Infrastructure (DSBI). DSBI is composed first of all by professionals with diverse expertise, computer technologies, specialized software, and organizational measures directed to facilitate data collection, data quality, data analytics, and proper use of findings.

The main objectives are to presents our vision on developing training programs directed to build competences needed for Data Scientists, especially in the area of identifying, assessing, evaluating, and managing risks in Big Data processing.

Key words: *Data Science, Big Data, Risk Management, Competences, Training*

INTRODUCTION

Recently, the term "Data Science" became popular to mark the phenomenon of technology mediated learning. Evolution of computer and communication technologies (CCT) and their wide application reach the point when data accumulated and stored on databases in almost every entity represents significant asset, but to gain its value are needed specific competences and technologies. Mikalef, et al. [1] 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.

Today, it is important to understand the available storage, processing and searching capabilities of big 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 term "Big Data" is used to describe the volume, velocity, and variety [2] of data well beyond the human capacity to comprehend without the use of computer technologies [3]. Those

competences are usually developed in different scientific disciplines and often the miscommunication between professionals doesn't allow the business to explore available data in a beneficial way.

Knowledge discoveries from data can lead to more efficient marketing, new revenue opportunities, better customer service, improved operational efficiency, competitive advantages over competing organizations, and other business benefits. Generally, the main purpose of large data analyzes is to help companies make better-informed business decisions [4].

According to [5] the basic concept of data science is the extraction of useful knowledge from data to solve business problems that can be systematically treated following a process with relatively well-defined stages. The results of the scientific data require careful consideration of the context in which they will be used in the relationship between the business problem and the analysis decision. The available means of analysis can be used to find informative data elements within the big data.

Despite the great potential of new technologies, tools, and applications for analysis, the biggest problems faced by practitioners in using these technologies are finding employees with the necessary skills, competent to guarantee that technologies will be used properly in a way to produce correct and reliable results.

The challenges in exploring data, faced by the business, create a challenge to educational industry as well. How to prepare professionals in practically every field to be able to explore Big Data? What are the competences they need? What has to be covered by the curriculum of "Data Science" professional – data scientist? This paper is addressing one of the aspects of those issues - what are the risks a data scientist may face and must consider in executing their profession. How to incorporate the area of risk management into Data Science curriculum? What are the specific risks associated with exploring Big Data?

The paper is dedicated to present the authors view on this issue. What represents the core of Data Science training? What is the place of Risk management in it? What are the specific Data Science risks? Next three sections are answering these questions.

BACKGRUND

The framework defining the knowledge, skills, and competencies of data professionals is becoming an object of constant research. The researchers concentrate their efforts on defining the scope of competencies that build the profile of the modern data specialist. For example, Sicular [5] defines Data scientist as a widely-used specialist within a variety of organizations, making it difficult to provide a complete and consistent list of required skills. The author lays down the skills required for this specialist's profile including data manipulation (warehousing), data analysis, data conversion and communication skills. Ismail [6] limits the skills of data scientist to five main categories: business, statistics, machine training, communication, and analysis. Costa, C.et al. [7] confirm the general assertion that Data Scientist is a multidisciplinary profile that seeks knowledge in several areas of learning. The authors add that this specialist relies heavily on the scientific way of doing things, so research experience is extremely important in shaping his/her profile.

Under the EDISON project [8], the "Knowledge Level for Learning Outcomes approach in the Data Science model curriculum (MC-DS)" includes the limitation to 3 levels: familiarity (as knowledge and comprehension), usage (as application and analysis), creation (as synthesis and evaluation).

In our previews study (see [9] and [10]) we identified three fundamental categories of competences a data scientists need to master (Table 1).

Table 1. Structure of Data Science competences

Category	Tasks	Competence	Skills
Extract	Choose, Classify, Collect, Compare, Configure, Contrast, Define, Demonstrate, Describe, Execute, Explain, Find, Identify, Illustrate, Label, List, Match, Name, Omit, Operate, Outline, Recall, Rephrase, Show, Summarize, Tell, Translate	Ability to extract useful data from huge and diverse repositories, including public and private, and also well and poorly structured sources.	Hard skills
Verify	Apply, Analyze, Build, Construct, Develop, Examine, Experiment with, Identify, Infer, Inspect, Model, Motivate, Organize, Select, Simplify, Solve, Survey, Test for, Visualize.	Ability to verify the obtained data and to judge about their quality	Hard skills Analytical skills
Interpret	Adapt, Assess, Change, Combine, Compile, Compose, Conclude, Criticize, Create, Decide, Deduct, Defend, Design, Discuss, Determine, Disprove, Evaluate, Imagine, Improve, Influence, Invent, Judge, Justify, Optimize, Plan, Predict, Prioritize, Prove, Rate, Recommend, Solve.	Ability to interpret (map) obtained data to the context (problem) and to applied appropriate analytic techniques to extract useful patterns, relationships or simply to increase understanding regarding the circumstances associated with the problem.	Hard skills Soft skills Analytical skills

In the following section share, in brief, the curriculum designed to train data scientists, and further how risk management are addressed in this curriculum

DATA SCIENCE CORE CURRICULUM

Training students to work with Big Data is a complex task. The required competences, as shown in the previous section, are quite diverse, the knowledge and skills needed are typically belong to the scope of different disciplines. They include skills to work efficiently with technology; applying correctly sophisticated mathematical and statistical methods; understanding and exploring a variety of organizational techniques; and possessing deep domain knowledge. The last, but not least is a deep understanding of the way how users perceive information and to find an effective way to visualize results. In designing curriculum two issues needs addressing – current education is based on studying in deep, but narrow disciplines, but Data Science is a much broader area, not covered in full by any specific “classical” educational area; and current trends among students show that the young generation is withdrawing from studying topics related to data analysis such as mathematics and statistics.

Data science curriculum is built on the three fundamental categories of competences defined above:

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

Ability to do all of the data processing via IT, in a highly efficient and effective way, represents an essential part of the curriculum as well.

Critical competence is the ability to understand the properties of accessible and obtained data. "Garbage-in, garbage-out". Understanding data properties include the ability to answer the following questions:

- What represents the data quality?
- What are the relevant criteria to assess data quality of a given data source or problem domain?
- What are the factors and circumstances influencing data collecting and presentation in a given way?
- How will data be used? and
- How to measure whether data availability and data exploration satisfy the above criteria to guarantee meaningful inferences?

Success in the Big Data era also requires competences to obtain meaningful, useful results from data that do not fully satisfy the highest criteria for quality, and competences to make rational inferences under uncertainty.

Those competences are not limited to any particular discipline, major or profession. Nowadays they are an important requirement for every branch and every profession.

The curriculum model of a Master Program, presented here, corresponds to authors understanding regarding contents of a program aiming to train professionals with different backgrounds (see [11]). It is constructed in to cover three or four semesters:

Preliminary requirements:

- Calculus I
- Probability Theory and/or Mathematical Statistics
- Discrete Mathematics
- Software Development (any language)
- Fundamental Data Structures and Algorithms
- Relational databases and SQL

First Semester:

- Introduction to Data Science
- Statistics: parametric and non-parametric methods for inference
- Cloud Computing, including Data Centers, NOSQL DB, Hadoop with Map-reduce.
- Data Analytics
- Visualization.

Second Semester:

- Big Data Analysis: challenges and benefits; Gartner's EIM Maturity models
- Big Data Applications: Architectures
- Data-Driven Management
- Applications:
  - Fraud detection
  - Exploring social networks - behavioral economics - marketing
  - "In-house" data management - ERP, BI

- Data Science in public services - e-Government
- In-memory solutions.

Third semester

- Pre-diploma project
- Diploma Thesis

Further, we share our approach in training risk management across the curriculum. In the next couple of section we elaborate our vision of what represents specific risks faced by data scientists, and in the last section – how this vision is implemented within the curriculum as presented above.

RISK MANAGEMENT AS ESSENTIAL COMPONENT OF TRAINING DATA SCIENTISTS

Risk management, usually, is part of training students in the field of Project Management. Such training is usually dominated by specific domain risks. For example, in IT projects, special attention is given to bugs in the software and risk management stress on training students how to organize development process in a way to identify and fix bugs.

In the field of Data Science hazards are events that lead to wrong decisions, to decisions misled by wrong interpretation of results produced by analytical tools used to process data. In general, risks associated with Data Science are product of human readiness and awareness. Often user considers technology-mediated data processing as a "black box", without clear understanding how given computer application produces the results, how sensitive is such application to data, especially to violation of mathematical conditions regarding the properties of the data to produce reliable results. For example, how sensitive is given statistical technique to a violation of expectation for Gaussian distribution of data.

In addressing Risk Management for Data Science training we consider only those areas of potential hazards which are not covered by other disciplines and represent a specific risk for the success in exploring Big Data. Literally, we identified three areas of hazards – user readiness; data quality, and sensitivity of analytical tools toward violation of required conditions.

Users' Maturity: Level of Awareness and Readiness

The origin of hazards in this category is information asymmetry. Data scientists and end-users often speak different jargons. Miscommunication may result in misinterpretation and misleading and as result wrong decisions and losses. Also, missing assessment of user's awareness and readiness may result in constructing interfaces that doesn't inspire confidence and as a result users are not making "data-driven" decisions.

Data Quality

Data quality is another category of risks. To apply given analytical technique data scientist have to "clean" row data. Often this results in loosing significant information. In Big Data era, verifying data quality and assessing the properties of data according to conditions of given analytical techniques requires specific skills. These skills include, but are not limited to, tracing origin of data, assessing data credibility, accuracy, context of creating and recording, factors influencing data, and many others. In some cases evaluation of properties as precision, repeatability and reproducibility needs understanding the physics, equipment used, competences of people created those data, etc. Inability to verify data quality creates the risk of obtaining misleading result by applying data processing analytical technique as a black-box.

Data processing

Risks in this category are mostly results of exploring analytical applications as “black-boxes”. There are data, there is software – import data into software and obtain results. Lack of competences results in accepting inappropriate for the purpose of analysis technique, applying the technique in a wrong way, or accept results without considering sensitivity effect.

In Tabl.2 we will present how this vision is realized in the curriculum of "Data Science".

Table 2. Curriculum implications

Criterion	Risk	Losses	Guidelines / emphasis in the course	Disciplines overcoming risk
User Readiness	Information Asymmetry	Misinterpretation And Misleading; Wrong Decisions And Losses	how users perceive the domain of data, context, and what are the problems they face and what are acceptable solutions	“Big Data Analysis: challenges and benefits” “Visualization”
Data Quality	Loosing Significant Information	obtaining misleading result by applying data processing analytical technique	tracing origin of data, assessing data credibility, accuracy, context of creating and recording, factors influencing data.	“Data Driven Management”
Sensitivity Of Analytical Tools	Lack of competences	Accepting inappropriate for the purpose of analysis technique, applying the technique in a wrong way, or accept results without considering sensitivity effect	Which case which technique is most appropriate; how to apply it in a proper way, considering objectives, domain, context, and properties of data; and to present results to users, considering their readiness and awareness	The entire program

The course “Big Data Analysis: challenges and benefits” includes training students to assess the maturity level of the business entity, but also how to assess the readiness of specific users. The course “Visualization” teaches students of different techniques applicable to users with a different background – from highly naïve to highly sophisticated. Building skills to assess user’s competences is one of the objectives of these courses. The message to students is to use these techniques also to learn how users perceive the domain of data, context, and what are the problems they face and what are acceptable solutions.

The data quality risk is discussed in the course “Data Driven Management”. The course studies several cases of effects of a violation of data quality properties which lead to wrong decision and significant losses. The two aspects – row data that don’t meet requirements to use given techniques (for example data that are not normally distributed in applying statistics heavily sensitive to this

property of data), and the effect caused by cleaning data via removing incomplete or missing records are included. Specific techniques to mitigate such risks are presented and trained in discussing different applications of data science.

The entire program is oriented to train students to understand in which case which technique is most appropriate; how to apply it in a proper way, considering objectives, domain, context, and properties of data; and to present results to users, considering their readiness and awareness. Often choosing a simple technique, just to visualize data in a proper way is sufficient for the user.

CONCLUSION

Risks are usually the last component addressed in training. Emphasis is given to risks in developing applications, in project management, but quite rare training risk management addresses regular operational business processes. It is assumed that hazards are addressed when the process is designed. From one side data analytics is a regular operational process, but from the other side any particular data analysis is specific, different, and exposed to many hazards. Training students to have in mind potential risks, not directly associated with application development, but how data is analyzed and how results are perceived and used, represents a challenging task in designing curriculum.

The paper shares authors' vision in addressing this problem when the curriculum for a data Science master program was designed.

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] Mikalef, P. et al. (2018) The Human Side of Big Data Understanding the skills of the data scientist in education and industry. IEEE EDUCON 2018 Global Engineering Education Conference, At Tenerife, Canary Islands, Spain.

[2] Laney, D. (2012) The Importance of "Big Data": A Definition, *Gartner,* Retrieved June 21, 2012 from http://www.gartner.com/resId=2057415

[3] Christozov, D., Toleva-Stoimenova S. (2015). 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] Shum, S. B., Hall, W., Keynes, M., Baker, R. S. J., Behrens, J. T., Hawksey, M., & Jeffery, N. (2013). Educational Data Scientists : A Scarce Breed. Retrieved from http://simon.buckinghamshum.net/wp-content/uploads/2013/03/LAK13Panel-Educ Data Scientists. pdf.

[5] Sicular, S. (2015). Big Data Analytics Failures and How to Prevent Them, 1(August).

[6] 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

[7] 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

[8] http://edison-project.eu/sites/edison-project.eu/files/attached_files/node-447/edison-mc-ds-release2-v03.pdf (08.09.2018)

[9] Rasheva-Yordanova et al. (2018) Forming of Data Science Competence for Bridging the Digital Divide. The eight edition of International Conference "Future in education" 2018. New Perspectives in Science Education, Libreria Universitaria Edizioni. ISSN 2384-9509 (in print).

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

[11] Christozov D., Toleva-Stoimenova S., Rasheva-Yordanova K., Vukarski I., (2016) Developing Big Data Competences in the Digital Era, International Conference on Big Data, Knowledge and Control Systems Engineering - BdKCSE'2016, Sofia

Dimitar Christozov, professor in American University in Bulgaria (AUBG), 1 G. Izmirliev square., Blagoevgrad 2700, Bulgaria, Ph.D., (+359 73) 888 443, dgc@aubg.edu

Katia Rasheva-Yordanova, assistant professor in University of Library Studies and Information Technologies, 119 Tsarigradsko shose Blvd., Sofia 1784, Bulgaria, Doctor, (+359) 878 720 961, k.rasheva@unibit.bg

Stefka Toleva-Stoimenova, assistant professor in University of Library Studies and Information Technologies, 119 Tsarigradsko shose Blvd., Sofia 1784, Bulgaria, Doctor, (+359) 887 542 490, s.toleva@unibit.bg