ANALYTICAL COMPETENCES IN BIG DATA ERA: TAXONOMY

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

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

Abstract

Recent advancement of computer technologies, namely Big Data Era, created a huge opportunity in rising human performance, but added a new round in so-called "digital divide" challenge, dividing society into Data elite, who are capable to benefit from those new opportunities, and the rest.

Exploring Big Data is a complex and challenging task, because of not only its 3 V's, but more important because the current state of publicly available sources is flooded with unreliable, difficult to verify information, in addition to free use of languages and jargons, and missing contexts. In this regard, skills to explore and benefit of data availability, and especially competences for performing data analysis are becoming essential for success nowadays. The question motivated our work is: What represents "data analytics competences" in Big Data Era? In previous studies, we have shown that analytical competences represents the cross-point of all other hard and soft skills in data processing, especially in the Big Data context. In this paper, analytical competences are considered in a way to emphasize their importance in almost all stages of data life cycle.

The principal objective of this paper is to develop a classification (taxonomy) of key competences required for data analysis under challenges of the day – today's way of creating, disseminating, and accessing data, as well as data processing, assessing, interpreting, and inferring. The taxonomy particularly concerns areas of competences such as identifying problems and what means relevant information; how to locate, access and verify credibility of sources; ability to analyze consistency of obtained data; to interpret aggregated measures used to represent population of researched objects by clear understanding of their limits, constraints, and how they are applied; and last, but not least, synthesizing solutions in consent with data, preserving professional ethical and moral standards.

As this work explores the essential skills and competences needed to succeed in the area of data analytics, it may serve a large and diverse group of professionals and can help in researching analytical competences within groups. The developed and proposed here analytical competences taxonomy is essential in curriculum design. Such classification can be used as benchmark for designing and executing different kinds of training, qualification, and degree programs in the area of Data Science.

Keywords: Big Data, Data Analytics, Taxonomy, Competences.

1 INTRODUCTION

Data Science, Big Data, Information Society are terms used to mark current state of dependence of people on data and the ways people manage data. The triad "data-information-knowledge" characterizes the process of learning from data. Modern information technologies changed completely the way how data is collected stored, organized, accessed, and processed. But the slogan "drowning in information but starved for knowledge", written more than 35 years ago [9] is becoming more and more valid.

Natural question is what is needed to extract useful value from accumulated data representing the true picture of the observed reality? The easy answer is "data analytics" offer those tools that are designed to explore data. But, what constitute "data analytics"? Great majority of people simply understand a set of software applications and are using them as "black boxes". In many cases this approach works, but the risk of being misled is significant. Data acquired by public sources are polluted in many ways, analytical techniques implemented as those software are valid only when certain conditions are fulfilled. The level of sensitivity toward data of such techniques may result in wrong and misleading inferences.

What competences a user of data analytical tools needs in a way to mitigate the risk of being misled by the results obtained? What represents a comprehensive set of abilities and skills to explore data? How to structure such set of competences? How to train students to increase the probability that they

will become successful in nowadays world? Institutions and people engaged in educating professionals in every area of knowledge nowadays need to find answers to those questions.

This paper is trying to initiate research on structuring and classifying the analytical competences, which can help structuring curriculum in a way to comprehend capabilities of graduates. Taxonomy of competences is designed to follow specific objectives, which provide meaning of priorities and hierarchies. Different objectives may result in different classifications. Next section presents our view regarding the set of competences composing “analytical inquiry” category, stating reducing the risk of misinforming when using computer / software data analytic applications as the leading objective. In our view, this is the most general approach. We follow to great extend the publications of Lumina Foundation [3] regarding Degree Qualification Profile (DQP), adjusted to our own experience. Third section is dedicated to discussion on optional approaches and pro and cons arguments.

2 TAXONOMY OF ANALYTICAL COMPETENCES

Analytical skill is the ability to visualize, articulate, and solve both complex and uncomplicated problems and concepts and make decisions that are sensible and based on available information.

Mastery of this competence involves developing a set of skills such as:

- Identifying, in the unordered description of a situation, the informative elements it contains, classifying them according to the degree of reliability or certainty attached to them.
- Recognising information gaps, for which one must have a prior theoretical model for sorting information elements into categories. Lack of sufficient or sufficiently reliable information can lead to decisions to search for new information or to posit reasonable hypotheses or estimates before attempting to reach conclusions.
- Using analytical tools to organise available information and show the relations between different components. Such tools include visualization to describe data in a meaningful, structured way as figures, outlines, tables, concept maps, graphs, etc.; and inferring from data by using statistical and other techniques.

Data Science supports and encourages shifting between deductive (hypothesis-based) and inductive (pattern-based) reasoning**.** This is a fundamental change from traditional analysis approaches. Inductive reasoning and exploratory data analysis provide a means to form or refine hypotheses and discover new analytic paths.

The paper's primary focus is on developing a classification (taxonomy) of key competences required for data analysis in Big Data context.

The general framework used here includes three categories of variables, described by data collected, subject of analysis:

- Controlled variables,
- Uncontrolled variables,
- Outcomes.

The first two are called also factors, predictors, inputs; and the third one – results. A simplified goal of data analysis is to find the relationship between those three categories of variables. Or, knowledge created is an explicitly defined model describing this relationship.

Having those competences allows an analyst to reach correct understanding about “causes and effect” driving development of particular circumstances, and to present this understanding in a useful form.

According to Microsoft Azure’s blog [12], we typically use Data Science to answer five types of questions:

1 How much or how many? (regression)

2 Which category? (classification)

3 Which group? (clustering)

4 Is this weird? (anomaly detection)

5 Which option should be taken? (recommendation)

In our view the taxonomy, designed to reduce the risk of misinforming, is more generally applicable approach compared with the two other approaches, briefly described in the next section. Following this leading principle, the categories of Analytical Competences can be classified as:

2.1 Identify and Analyze Existence of a Problem

This category of competences addresses ability to analyze particular circumstances, to place facts into given context and identify existence and formulate the problem. These competences usually are defined as "domain" knowledge. Proper structuring or framing of the problem is critical to achieving relevant and actionable outcomes. Sometimes, analytics requires inputs via scenarios. Organizations need to define relevant scenarios that will drive portfolio analyses.

2.1.1 Identify Existence of a Problem

Existence of a problem is defined by comparisons between expected outcomes and data describing the real state. Analytical skills include accepting variability in given boundaries, based on randomness of the uncontrolled factors. Understanding whether the difference between expected and achieved is based on random circumstances or not. The latest indicates existence of a problem and defines the research goal. Data Scientist has to clarify the problem and narrows the scope of the study into understandable, simple, short, and measurable goals, which might be set out in question format.

2.1.2 Identify What Cause the Problem.

This category of competences covers ability to understand the context of appearing the problem and may allow assessing applicability of solution.

The main purpose here is to make sure all the stakeholders understand the who, what, where, when, why and how of the problem [4]:

- **Who**: Who is causing the problem? Who says this is a problem? Who are (not) impacted?
- **What**: What is happen if not solved? What are the symptoms? What are the impacts?
- **Where:** Where does the problem occur? Where does it have an impact?
- **When:** When does this problem occur? When did it first start occurring?
- **Why:** Why does the problem occurring?
- **How:** How should this system work? How is it currently handled?

Then, the analyst has to possess competences to answer questions like the following:

- What are factors (inputs) caused the problem?
- What are the relevant variables?
- What is the acceptable variability of uncontrolled inputs?

A cause and effect diagram, often called a "fishbone" diagram, is used in brainstorming to identify possible causes of a problem and in sorting ideas into useful categories. It demonstrates the relationship between effects and the categories of their causes and to find the root causes.

In this stage, the variables that need to be predicted have been identified, but also the category of problem solution, whether predicting a particular value – regression, or assigning to given category – classification; and the type of learning – supervised or unsupervised; to define the scope of applicable analytical methods.

2.1.3 Identify Whether the Problem is Decomposable

Once the actual problem is identified the next step is to decompose the problem into smaller distinct sub-problems each with their own goals, data, computations, and actions. Often there may be many valid ways to decompose the problem, each leading to a different solution. There may be hidden dependencies or constraints that only emerge in the process of developing a solution.

This category of competences requires deep understanding of problem and interrelations between variables. The aim is to gain additional insights into the problem and also to explore independence or loose relationship which may allow solving multiple, but simpler, problems instead.

2.1.4 Assess Whether Factors with Impact on Problem's Solution are Controllable.

A cause and effect diagram also helps the analysts explore the main variables, understand which are in their control, which are measurable, etc. In addition to the cause-and-effect competences, identifying factors, which are fully under control of the decision maker from factors, which depend on environment, require understanding the limitation of available data, and impact of random forces. The independent (controllable) variables also referred to as predictors are used to predict the values of dependent variables.

2.2 Identify and Assess Information Needs, Data Quality, and whether Data is Relevant to the Problem.

A critical analysis and examination of the data itself needs to be considered. Selecting data that is used in an analysis, proper filtering of irrelevant data, ensuring high data integrity, and data element definition are core in any analytics process. This category of competences addresses ability to analyze credibility of data explored. Additionally, this category includes competences to select analytical tool (technique, method) applicable to the problem and to the properties of explored data. These competences usually are defined as "information competences".

2.2.1 Identify and Assess Information Needed

Understanding problem domain, and the context in which the problem arises, defines how to assess the relevance of data. Defining which variables influence the output is the first step in assessing whether the further analysis may bring reliable result to find a solution of the problem. Identifying which are the input variables, whether they are controllable or not, is of critical importance for the success of data analytics.

2.2.2 Where relevant information can be found (sources)?

Whether available data provides sufficient resource for building the model with acceptable level of certainty? Is it possible to enrich information base to improve the analysis?

Identifying a wider variety of data sources may increase the probability of finding hidden patterns and correlations. For example, to provide insight, it can be beneficial to identify as many types of related internal and external data sources as possible, especially when it is unclear exactly what to look for. Internal data includes policy data, insurance application documents, claim data, claim adjuster notes, incident photographs, call center agent notes and emails. In the case of external datasets, a list of possible third-party data providers, such as data markets and publicly available datasets, are compiled. Some forms of external data may be embedded within blogs or other types of content-based web sites, in which case they may need to be harvested via automated tools. Data providers from third parties, for example social network data or voluntarily contributed data bears the risk of being manipulated. For example, many spurious social media messages could be generated in order to push a statistical index derived from these data in one or another way in case it is known that the index is calculated from such data [1].

2.2.3 Ability to access, to retrieve, and to structure relevant data.

The Data Scientist has to collect sufficiently representative data in order to make sense of the problem at hand. Data can be obtained in several ways - for example by downloading it from a server, by querying a database, or by connecting to a Web API. However, even just text can come in multiple forms, sometimes, the data comes in a compressed form or in a binary format such as Microsoft Excel. This subcategory of Data Scientist's competences covers technical competences needed to obtain data and to present data in a form suitable for further processing. It includes:

- access and query many different databases and data sources (*RDBMS*, *NoSQL*, *NewSQL*),
- integrating the data into an analytics-driven data source (e.g., *OLAP*, *warehouse*, *data lake*, …)
- structuring the data in proper format as the data gathered from different has various shapes and sizes.

One of the major challenges researchers are facing is "How to integrate all the data from different sources to maximize the value of data."

2.2.4 Ability to clean and to explore the acquired data (Transformation stage)

Once Data Scientist obtained the data he/she has to prepare the data from a raw form into a suitable form for data modeling. Transforming stage includes cleaning and removing any noise from the data (remove inconsistencies, dealing with missing data, etc.), normalization, aggregation and generalization. After the transformation and scaling of data duplication, i.e., removing redundancy and efficiently organizing data. This is very important in Big Data projects, which often involve terabytes of data to work with. But the trade-off between cleansing data and striving for quality data needs clear justification. Cleansing data may always result in loss of potentially useful insights [5].

The transforming stage is following by data exploration stage. It is like the brainstorming of data analysis. It includes analyzing the data, finding bias, patterns and dependencies among raw data for deep understanding and validation of its limitations and quality. Feature engineering is an essential part of building any intelligent system. Feature engineering deals with feature selection (cutting down the features that add more noise than information) and construction (creating new features) [16]. It is an important part of cleaning data because it greatly reduces noise and the problem with dimensionality, and can even address the problems with sparse data. That means the model will have an easier time navigating missing values and outliers to more efficiently learn relationships in your data [14].

2.2.5 Ability to verify and to asses credibility of obtained information.

Assessing trustfulness of data is essential to justify output of data analytics. The slogan “garbage in, garbage out” is fully applicable.

Credibility is usually used to evaluate non-numeric data. It refers to the objective and subjective components of the believability of a source or message. Credibility of data has three key factors: reliability of data sources, data normalization, and the time when the data are produced.

The data collected by these authoritative scientific bodies generally have high data authenticity and credibility because researchers in these organizations generally obey research ethics and follow good scientific practices. However sometimes Big data are used from commercial companies or often comes from systems not directly controlled by an organization and may contain inherent biases or outright false values (for example because of bots in social media).

2.2.6 Ability to assess data quality.

There are several measures of data quality as completeness, consistency, uniqueness, integrity, conformity, accuracy, precision, timeliness, reproducibility, etc. This category covers competences to choose the right measures for given case, and skills to measure whether data satisfy quality expectations. A different set of skills and tools are often used for detecting and correcting errors and inconsistencies from a data set in order to improve its quality. A good understanding of architecture, technology, corporate culture, and other factors is often important. Besides a number of essential technical skills are also required when dealing with the data itself including parsing, standardising, record linkage/matching, data scrubbing/cleansing, data profiling and data auditing/monitoring. These skills are often extensively used when conducting projects such as data migrations where data quality improvements need to be achieved in tight timescales.

2.2.7 Competences to Map Obtained Data to Analytical Tool

A critical step in Data Science is to identify an analytic technique that will produce the desired action. Whereas predictive analytics tells what will happen, prescriptive analytics suggests what to do [7]. Prescriptive analytics can identify optimal solutions, often for the allocation of scarce resources. Organizations typically move from descriptive to predictive to prescriptive analytics. Another way of describing this progression is: what happened – why did it happen, what will happen, how can we make it happen [19]. Things can get a little more complicated because the lines between the different types of tools can be a little fuzzy. Some business intelligence tools have data mining and predictive analytics capabilities. Some predictive analytics tools include streaming capabilities.

This is another category of technical competences – competences to apply the right tool in the right way. This category includes competences about the computer application – analytical method implemented and characteristics of data required. In addition, how to transform raw data to suit the tool’s input format. The Data Scientist may encounter a wide variety of implementation constraints. They can be conceptualized, however, in the context of five dimensions that compete for your

attention: analytic complexity, speed, accuracy & precision, data size, and data complexity. Balancing these dimensions is a zero sum game - an analytic solution cannot simultaneously exhibit all five dimensions, but instead must make trades between them. Achieving a desired analytic action often requires combining multiple analytic techniques into a holistic, end-to-end solution [17].

2.3 Synthesizes solution

This category of competences addresses ability to use effectively data analytical tools. It is required specific competences in mathematics, statistics, artificial intellect, and especially efficient use of computer technologies. These competences usually are defined as “machine learning” competences.

2.3.1 Understanding Limitations of Applied Analytical Tool

The truth is there is not a single best algorithm that is universally better in all situations –choosing the best algorithm depends on the problem type, size, available resources, etc. Moreover, machines lack common sense so humans are still needed to supervise.

Increased capacity to process Big Data creates an inherent tendency towards include irrelevant data. On the other hand, used analytical techniques as a “black box” may mislead analysts in interpreting results in the domain context. The black box decision-making is a serious limitation as many policy execution or governance requirements need clear explanations of decisions, e.g. explain to customer why transaction was blocked.

The higher data volume increases the probability that the data files and documents may contain inherently valuable and sensitive information. In this regard, Big data analytics and purely automated processing may use personal data to make decisions affecting individuals. It can relate to privacy and security concern which lead to some ethical challenges. Sometimes during the translation of cultural clichés and stereotypes into empirically verifiable datasets, subjectivity is introduced. In some circumstances even displaying different advertisements can mean that the users of that service are being profiled in a way that perpetuates discrimination, for example on the basis of race, sex, political opinion, religion, etc. Data Scientists must have experience in dealing with such problems [18].

2.3.2 Setting Assumptions

Competences to define assumptions in applying analytical techniques are essential to allow beneficial interpretation and justification of obtained results. Data scientist should be able to generate hypotheses about the ways the concerned system can behave if it is changed in a specific manner. As stated by Ockham's razor the simpler is better so that less assumption is preferable. Clearly, attributes that are not in the model will have no effect on the model’s predictions [6].

Hypothesis generation allows [15]:

- to experiment with theories about the data;
- to take a systems-thinking approach to the problem to be solved;
- to build more sophisticated models based on prior hypotheses and understanding.

2.3.3 Comparing Different Applicable Courses of Action

Choosing the most informative tool for given case, requires ability to compare tools on the base of the properties of available data and data analytics objectives. The Data Scientist have to choose a machine learning algorithm or statistical methodology that suits the data, use case, and available computational resources. Many Data Scientists choose to build and test multiple modelling methodologies to explore outputted predictions.

Modelling is where art meets science. The art comes from the a priori intuition and assumptions on the relationship between each predictor variable—i.e., the variables used to predict the outcome—and the response variable. The science comes from running quantitative analysis on the data and using a large collection of models. Model selection involves limiting the number of parameters and ultimately selecting a model that is understandable and actionable by stakeholders [10]. Question to consider are [13]:

- Does the model appear valid and accurate related to the launch on the dataset?
- Does the model output/behavior make sense to the domain experts?
- Do the parameter values make sense in the context of the domain?

- Is the model sufficiently accurate to meet the goal?
- Are there any tolerable or intolerable errors occurring?
- Are there enough data or we need more inputs? Are there any transformations or adjustments of the data to be performed?
- Will the kind of model chosen support the runtime environment?
- Is a different form of the model required to address the business problem?

2.3.4 Verifying Obtained Model

This category includes competences to apply techniques (cross validation and receiver operating characteristic (ROC) curve analysis, etc.) to verify applicability of the obtained model.

The Data Scientist can estimate how effective it is by applying it to some of the training data and comparing the prediction to the known value. Usually the model is tested in a larger set of cases and it helps in defining the scope of circumstances where it is applicable. This part always precedes deploying or presenting the model [14].

2.3.5 Communicating results

Once Data Scientist derived the intended insights from the model, he/she have to represent them in a user-friendly, re-usable and intelligible format that the different key stakeholders in the project can understand. Presenting information in such a way that people can consume it effectively is a key challenge that requires to be aware of data visualization tools, techniques and technologies used for creating images, diagrams, or animations to communicate, understand, and improve the results of Big Data analyses. Data scientist have to choose from different visualization methods depending on the machine learning model and other Data Science process flow characteristics.

Data visualization combines the fields of communication, psychology, statistics, and art, with an ultimate goal of communicating the data in a simple yet effective and visually pleasing way [16].

3 ALTERNATIVE APPROACHES IN CLASSIFYING ANALYTICAL COMPETENCES

The taxonomy presented in the previous section is only one of the possible ways to classify the competences needed in Data Science. In this section, we present in brief two possible alternative approaches to illustrate from one-side alternatives, but from other the complexity of building such classification and to enforce further discussion.

3.1 Technology Driven Classification

Technology driven classification prioritize the techniques and methods used in data processing. The major categories are listed below in the Table1 [8,11]:

Table 1. Big Data analytics techniques and models

Technique/Method	Description
A/B testing	a control object is compared with many test objects for improvements (split or bucket testing)
Associated Rule Mining	used for relationships among the huge datasets
Classification	used to identify the category of the incoming new data based on the existing set which are already categorized based on some data points
Cluster Analysis	used to group a set of objects into same group based on some commonalities to each other and different to other groups
Crowdsourcing	used to generate the ideas/ innovations from a large group through an online method
Data Fusion and Integration	used to generate more efficient and potential insights of the data using set of models and techniques which integrate multiple data sets and analyze, rather analyzing them independently

Data Mining	used for predictive analytics to identity hidden patterns
Ensemble Learning	a type of supervisory learning technique
Genetic Algorithm	based on the process of Natural evolution
Machine Learning	used to create artificial intelligence by providing some knowledge to the system
Natural Language Processing	a kind of Machine Learning process
Neural Networks	computational techniques, suitable for nonlinear pattern recognition; it uses both supervised learning and unsupervised learning
Network Analysis	used in a network or a graph to illustrate the associations among the discrete nodes
Predictive Modelling	a model built with the help of some of the statistical and mathematical techniques to predict a best possible outcome
Regression	a statistical technique used for prediction which will determine co-variance between dependent and independent variables
Sentiment Analysis	it is a Natural Language Processing which can determine the information of subject from the source of textual data
Spatial Analysis	uses set of statistical models which can be used to explore the geographical data such as Geographical Information Systems (GIS)
Statistics	used for collecting data in the form of surveys or experiments etc., organizing the data in any sorting order or lexical order etc., with an art of interpretation of the data in a form of hypothesis
Supervised Learning	adapts some of the machine learning techniques which can deduce relationships based on prior knowledge (training data)
Simulation	used to mimic the behavior of some complex systems to predict and plan the outcome and measure the results
Unsupervised Learning	uses some of the machine learning models which can identify the hidden patterns from the data without any prior knowledge; an example of Unsupervised learning is Cluster Analysis
Visualization	used to provide a gist of the information in very simple way in terms of a graph, diagram, image or any visual representation to simplify the understanding

3.2 Activities/Application Driven Classification

Application driven classification prioritize the data analytical domain. The particular domain in which data arises determines the types of architecture that is required to store it, process it, and perform analytics on it. There are several ways to characterize data [2]:

a) according to time span in which it needs to be analyzed
   - **Real-time** (financial streams, complex event processing (CEP), intrusion detection, fraud detection)
   - **Near real-time** (ad placement)
   - **Batch** (retail, forensics, bioinformatics, geodata, historical data of various types)

b) according to the degree of structure or organization the Big Data come with.
   - **Structured** (retail, financial, bioinformatics, geodata)
   - **Semi-structured** (web logs, email, documents)
   - **Unstructured** (images, video, sensor data, web pages)

c) according to the types of industries that generate and need to extract information from the data:
   - **Financial services** (high frequency trading)
   - **Retail** (behavioral analysis)
   - **Network security** (intrustion detection, APT's)

- **Large-scale science** (bioinformatics, high energy physics)
- **Social networking** (sentiment analysis, social graphs)
- **Internet of Things**/sensor networks (weather, anomaly detection)
- **Visual media** (scene analysis, image/audio understanding)

4 CONCLUSION

The volume and complexity of captured and stored data continue growing. Dependence on data analytics to understand what is going on and to adjust one's behavior increases. The groups of people needing competences on data analytics also grows as number and as variety and diversity grow. This definitely will result in evolution of understanding of "what represents data analytics competences". The three approaches, presented above, in building taxonomy on data analytics competences simply illustrate this diversity.

Opening a discussion and initiating research regarding defining taxonomy on data analytic competences was the major objectives of this paper. We choose to design it by following relatively abstract approach, which may serve a large and diverse group of professionals.

As this work explores the essential skills and competences needed to succeed in the area of data analytics, it can help in researching analytical competences of DS professionals. Such classification can be used as benchmark for designing and executing different kinds of training, qualification, and degree programs in the area of Data Science – from general education category in universities following liberal arts model, toward narrow professional training.

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] A. Wirthmann, M. Karlberg, B. Kovachev, F. Reis, "Structuring risks and solutions in the use of big data sources for producing official statistics – Analysis based on a risk and quality framework," Working Paper. Retrieved from https://ec.europa.eu/eurostat/cros/system/files/Big%20Data%20Risk%20Paper%20Version1.pdf

[2] Big Data Working Group, "Big Data Taxonomy", Cloud Security Alliance, 2014. Retrieved from https://downloads.cloudsecurityalliance.org/initiatives/bdwg/Big_Data_Taxonomy.pdf

[3] C. Adelman, P. Ewell, P. Gaston, C. G. Schneider, "Degree Qualification Profile," Lumina Foundations Publications, 2014. Retrieved from https://www.luminafoundation.org/resources/dqp

[4] D. Brown, J. Kusiak, "Problem Analysis Techniques," IRM Training, White Paper. Retrieved from https://www.miun.se/siteassets/fakulteter/nmt/summer-university/problemanalysispdf

[5] D. Cielen, A.D. B. Meysman, , M. Ali, "Introducing Data Science: Big data, machine learning, and more, using Python tools," Manning Publications, 2016.

[6] H. Kalechofsky, "A Little Data Science Business Guide," 2016. Retrieved from http://www.msquared.com/wp-content/uploads/2017/01/A-Simple-Framework-for-Building-Predictive-Models.pdf

[7] H.J.Watson, "Big Data: Concepts, Technologies, And Applications," Retrieved from http://www.mercerindustries.com/wp-content/uploads/2015/02/Watson-Tutorial-Big-Data-Business-Analytics-Collaborative.pdf

[8] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, A. Byers, "Big data: The next frontier for innovation, competition, and productivity," McKinsey Global Institute, 2011. Retrieved from https://bigdatawg.nist.gov/pdf/MGI_big_data_full_report.pdf

[9] J. Naisbitt, "Megatrends," New York: Warner Books, 1982.

[10] J.-Y. Rioux, J. Baer, P. Duplessis, J. Quah, S. (Y.) Zhan, "Predictive modelling. Turning Big Data into Big Opportunities," 2018. Retrieved from https://www.cia-ica.ca/docs/default-source/2018/218081e.pdf

[11] K. Venkatram, M. A. Geetha, "Review on Big Data & Analytics–Concepts, Philosophy, Process and Applications," Cybernetics and Information Technologies, Vol.17, No. 2, pp. 3-27, 2017.

[12] Microsoft Azure Blog, Accessed 8 September, 2018. Retrieved from https://azure.microsoft.com/en-us/blog/?v=18.37

[13] Microsoft Azure Blog, Accessed 14 September, 2018. Retrieved from https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/lifecycle-modeling

[14] N. Castle, "Model Building: From Data Cleaning to Deployment," 2017. Retrieved from https://www.datascience.com/blog/predictive-data-models-from-data-cleaning-to-model-deployment=20

[15] Oracle+Datascience.com, Accessed 9 September, 2018. Retrieved from https://rexplorations.wordpress.com/2017/02/17/hypothesis-generation-a-key-data-science-challenge/

[16] S. Agarwal, "Understanding the Data Science Lifecycle," 2018. Retrieved from http://sudeep.co/data-science/Understanding-the-Data-Science-Lifecycle/

[17] "The Field Guide to Data Science", Second Edition, Booz Allen Hamilton Inc., 2015. Retrieved from https://wolfpaulus.com/wp-content/uploads/2017/05/field-guide-to-data-science.pdf

[18] The United Kingdom Information Commissioner's Office, "BD, artificial intelligence, machine learning and data protection," Version 2.2, March 2017.

[19] W. Eckerson, "Gauge Your Data Warehousing Maturity." DM Review, vol. 14, no. 11, pp. 34, 2004.