ANALYSIS OF ONLINE COURSES TO ACQUIRE DATA SCIENCE LITERACY

P. Petrova, I. Kostadinova, V. Chantov

University of Library Studies and Information Technologies (BULGARIA)

Abstract

Data science skills relate to the competencies needed by everyone who wants to be successfully listed on the labour market. Successful data analysis increases the competitiveness of both the organization and every individual. More and more organizations allocate resources and have units built, even entire departments for processing and analysing Big Data and forecasting. Evidence is the increasing offering of Data Science courses in the web.

This paper aims to analyse available online courses offered by different learning organizations. Platforms for massive open online courses (MOOC) have been studied, some of which are: Coursera, EdX, Cloudera, intellipaat, openHPI. An analysis of the learning methodology has been made. The differences, strengths and weaknesses of the offered courses are shown.

This paper also discusses the possibility of adapting data science courses to a specific group of learners who are not IT professionals to build their data science literacy.

Keywords: analysis, online courses, data science, literacy, e-learning.

1 INTRODUCTION

Data science helps companies to better understand their users, better manage their business, and to be more competitive. Analysis of big data supports decision-making in each area. Banking institutions process large data to improve their success in detecting fraud. Asset management companies use large data to predict the probability that the price of the securities moves up or down at a certain time.

Companies use big data to determine what their consumers are interested in and use this information to make decisions about what products and services they produce and offer. Companies also use algorithms to create customized suggestions to their users based on their behaviour and current preferences.

Data science provides an advantage, but to get it, a company needs to have specialists to understand and analyse data.

2 DATA SCIENCE LITERACY

To acquire Data Science Literacy, we must first understand what data science is. From the definitions in different sources [18, 17, 9] it can be deduced the general assertion that data science is the ability to extract valuable information from different types and volumes of data using scientific techniques, algorithms and systems. The data science field employs mathematics, statistics and computer science disciplines, and incorporates techniques like machine learning, cluster analysis, data mining and visualization. [17] The goal of data science is to gain insights and knowledge from any type of data — both structured and unstructured. Data science is often confused with data mining. However, data mining is a subset of data science. It involves analyzing large amounts of data, such as big data, in order to discover patterns and other useful information. Data science, or data-driven science, combines different fields of work in statistics and computation in order to interpret data for the purpose of decision making. Data science covers the entire scope of data collection and processing. [15]

While a Business Intelligence analyst helps to understand and communicate data patterns from a business perspective, a data scientist is the person who helps gather, process and analyze data. Data scientists should also be skilled in communicating those findings and offering recommendations to others in the business. [13]

Data scientists must possess a combination of analytic, machine learning, data mining and statistical skills, as well as experience with algorithms and coding. Along with managing and interpreting large amounts of data, many data scientists are also tasked with creating data visualization models that help

illustrate the business value of digital information. [14] The data scientists are hired by organizations to help them turn raw data into valuable business information.

Data science is a discipline that has been reinforcing in academia over the past few years. [10] Data scientists are a new breed of analytical data expert who have the technical skills to solve complex problems and the curiosity to explore what problems need to be solved. There is a difference between job description and academic program for data scientist. An effective approach to overcoming the differences between academic programs and business needs is rival mentoring. [11] This approach helps universities to train students in right technologies, techniques, competences and skills.

3 DATA SCIENCE COMPETENCE AND SKILLS

There's not a definitive job description when it comes to a data scientist role. But here are few typical job duties for data scientist according [10, 1, 16]:

Data science competences and skills:

- Collecting large amounts of unruly data and transforming it into a more usable format.
- Solving business-related problems using data-driven techniques.
- Working with a variety of programming languages, including SAS, R and Python.
- Having a solid grasp of statistics, including statistical tests and distributions.
- Staying on top of analytical techniques such as machine learning, deep learning and text analytics.
- Communicating and collaborating with both IT and business.

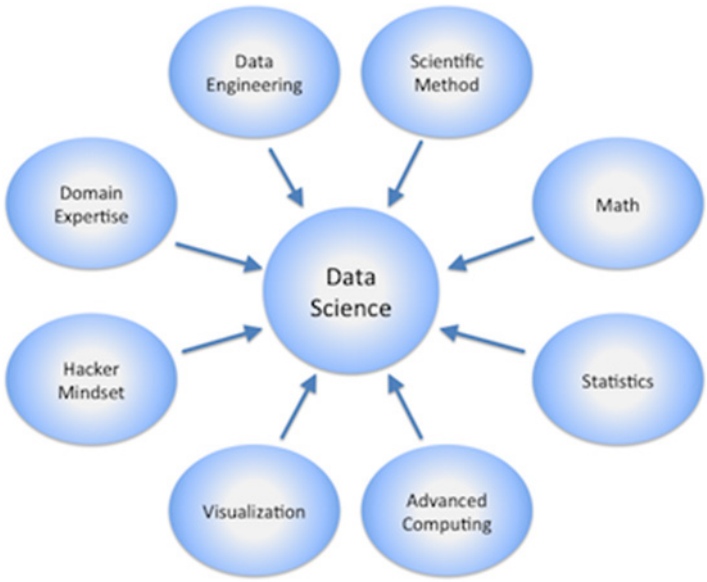

Figure 1. Data science disciplines [1]

Technologies commonly used by data scientists:

- Data visualization: the presentation of data in a pictorial or graphical format so it can be easily analyzed.
- Machine learning: a branch of artificial intelligence based on mathematical algorithms and automation.
- Deep learning: an area of machine learning research that uses data to model complex abstractions.
- Pattern recognition: technology that recognizes patterns in data (often used interchangeably with machine learning).

- Data preparation: the process of converting raw data into another format so it can be more easily consumed.
- Text analytics: the process of examining unstructured data to glean key business insights.
- Prediction: predict a value based on inputs.
- Classification: spam or not spam detection
- Automated processes and decision-making: credit card approval
- Segmentation: demographic-based marketing
- Optimization: risk management

4 DATA SCIENTIST TREND

According [2] job positions as:

- "Management analysts" who conduct research and develop procedures to allow organizations to run more efficiently, and
- "Market research analysts and marketing specialists" who research market conditions and create marketing campaigns and has data science knowledge

Will increase annual earnings by 2026, but job position as

- "Computer and information systems managers" who plan, direct, and coordinate computer systems and does not analyze data will decrease annual earnings by 2026.

According [6] top 25 job position in UK for 2017, Data scientist is on 6th position but only after HR, Audit, Design, Tax and Finance manager. In this list Data scientist is the most wanted and well-paid IT job in Great Britain for 2017. For 2018 data scientist keeps the salary rate (average base salary 45000 pounds) and currently number of job openings are 578.

Data scientist is a modern job position which according to Robert Half Technology's 2018 IT salary report is the best paid job position. The survey also reveals the average salaries for each role based off experience, shown in Table 1.

Table 1. *The 7 most in-demand tech jobs for 2018 [13]*

Job	25th percentile	50th percentile	75th percentile	95th percentile
Business intelligence analyst	$83,750	$104,000	$130,250	$175,750
Data scientist	$100,000	$119,000	$142,750	$168,000
Database developer	$97,750	$104,000	$130,250	$175,750
Data security administrator	$100,000	$117,500	$135,750	$168,750
System administrator	$64,500	$78,750	$95,750	$102,500

- 25th percentile: entry-level workers or those in industries with less competition
- 95th percentile: significant experience, certifications, specializations, high level of expertise, work in a strategic and highly complex role or in a highly competitive industry for talent.

5 DATA SCIENCE ONLINE TRAINING COURSES

The trend of data scientist job position predetermines the interest in setting up training courses in this direction. Here is an overview of online courses and specializations offered by widespread online learning platforms. According the definition of data scientist and job description following skills and experience have to look for:

- Programming languages (specifically Python or Java)
- Strong analytical skills
- Strong mathematical skills

- A masters or Ph.D.

5.1 Cloudera Data Scientist Training

Cloudera offers two 4-days training-led-courses. Data scientist, oriented to developers and Data Analyst, oriented to analysts.

Using scenarios and datasets from a fictional technology company, students discover insights to support critical business decisions and develop data products to transform the business. The material is presented through a sequence of brief lectures, interactive demonstrations, extensive hands-on exercises, and discussions. The Apache Spark™ demonstrations and exercises are conducted in Python and R using the Cloudera Data Science Workbench (CDSW) environment.

This four-day workshop covers data science and machine learning workflows at scale using Apache Spark 2 and other key components of the Hadoop ecosystem. The workshop emphasizes the use of data science and machine learning methods to address real-world business challenges.

The course is designed for data scientists who currently use Python or R to work with smaller datasets on a single machine and who need to scale up their analyses and machine learning models to large datasets on distributed clusters. Data engineers and developers with some knowledge of data science and machine learning may also find this workshop useful.

Workshop participants should have a basic understanding of Python or R and some experience exploring and analysing data and developing statistical or machine learning models. Knowledge of Hadoop or Spark is not required. [4]

Covered topics in this training including some of mentioned in part 2 technologies:

- Inspecting data quality
- Exploring data
- Extracting, transforming, and transforming data
- Regression, Classification, Clustering models
- Cross-validating models and tuning hyperparameters
- Machine learning
- Hadoop ecosystem
- Apache Spark 2
- Spark, Spark SQL, and Spark MLlib
- PySpark and sparklyr
- Cloudera Data Science Workbench (CDSW)
- HDFS data and Hive tables using Hue

5.2 Coursera Data Science Specialization [5]

A nine-course introduction to data science, developed and taught by leading professors, created by John Hopkins University.

This Specialization covers the concepts and tools throughout the entire data science pipeline, from asking the right kinds of questions to making inferences and publishing results. There is a final Capstone Project where every student applies the skills learned by building a data product using real-world data. At completion, students will have a portfolio demonstrating their mastery of the material.

Time to completion can vary, but most learners are able to complete the Specialization in 3-6 months.

Tools that is used in the program to ask the right questions, manipulate data sets, and create visualizations to communicate results are version control, markdown, git, GitHub, R, and Rstudio.

Covered topics in this training including some of mentioned in part 2 technologies:

- Finding Answers
- R Programming (Practical R Exercises in swirl)
- Getting Data
- Exploratory Data Analysis
- Reproducible Research
- Statistical Inference
- Regression Models
- Practical Machine Learning
- Building Data Products
- Types of Questions
- What is Data?
- Big Data
- Textual Data formats
- Getting and Cleaning Data
- Raw and Processed Data
- Summarizing Data

This specialization contains courses for beginners also for advanced learning.

5.3 EdX Foundations of Data Science

This course conducted by Berkeley University of California teaches you basic programming skills for manipulating data. You will learn how to use Python to organize and manipulate data in tables, and to visualize data effectively. No prior experience with programming or Python is needed. Not all data is numerical – you will work with textual data and with maps. Throughout, the underlying thread is that data science is a way of thinking, not just an assortment of methods. The course also emphasizes interpretation and communication, which are essential skills for all data scientists.

Length of the course is 5 week and teach how to use computation to help data tell their story. Basics of Python 3 and how to use it as a tool for data analysis. Fundamental principles and methods of visualization.

Covered topics in this training including some of mentioned in part 2 technologies:

- Correlation and the phenomenon of regression to the mean
- Linear regression
- Quantifying uncertainty and generating 95% confidence intervals using the bootstrap method
- Classification using the k-nearest neighbours algorithm
- How to evaluate the accuracy of a classifier
- Machine learning

5.4 Intellipaat Data Science Training Course [8]

A complete Data Science bootcamp specialization training course from Intellipaat that provides detailed learning in data science, data analytics, project life cycle, data acquisition, analysis, statistical methods and machine learning. Learners gain expertise to deploy Recommenders using Apache Mahout, data analysis, data transformation, experimentation and evaluation. There are no particular prerequisites for this Training Course.

Covered topics in this training including some of mentioned in part 2 technologies:

- Introduction to Data Science and Statistical Analytics
- Introduction to R and R-Studio

- Data Exploration, Data Wrangling and R Data Structure
- Data Visualization - Bar Graph (Simple, Grouped, Stacked), Histogram, Pi Chart, Line Chart, Box (Whisker) Plot, Scatter Plot, Correlogram
- Statistics - Probability, Normal Distribution, Binary Distribution, Hypothesis Testing, Chi Square Test, ANOVA
- Predictive Modelling - Linear and Logistic Regression
- Classification - Decision Trees
- Time Series

6 COURSES ANALYSIS AND COMPARISON

Based on previous research data science skills are defined [12]. According this research there are soft, hard and analytical skills for competences in data science. In Table 2 are shown the online training platforms described in section 4.

Table 2. *Data science courses in widespread online training platforms*

	Soft skills	Hard skills	Analytical skills
Competencies	1. Understanding the basic business objectives and strategies, as this will allow maximum compaction of the knowledge gained from the data; 2. Being able to understand stakeholders and support decision-making; 3. Being able to communicate and disseminate the findings.	1. To have the technical skills for statistical processing to apply in designing and interpreting experiments, modelling and forecasting; 2. Being able to create data artifacts or optimize existing ones.	1. To know methods of data analysis that automate the construction of analytical models; 2. To improve business management and achievements by enhancing decision-making.

According to the program that offers the training courses, the diagram presented in Fig. 2 is derived. It can be seen that all the skills mentioned in Table 2 are covered in the Coursera specialization.

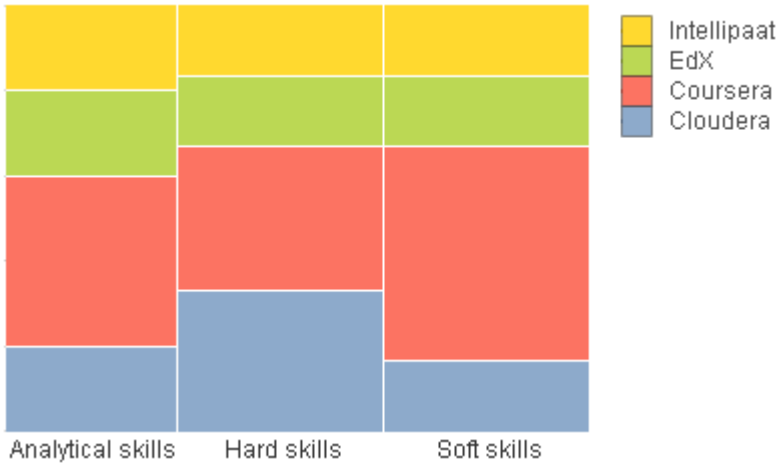

Fig.2 Data science skills covered from different platforms

In the courses studied, the hard skills defined in Table 2 are mainly covered. These are programming, working with mathematical and statistical models, visualization of data. The hard skills are mainly focused on mastering data processing techniques and tools. From Fig. 3, it can be seen that the courses focus mainly on the development of hard skills.

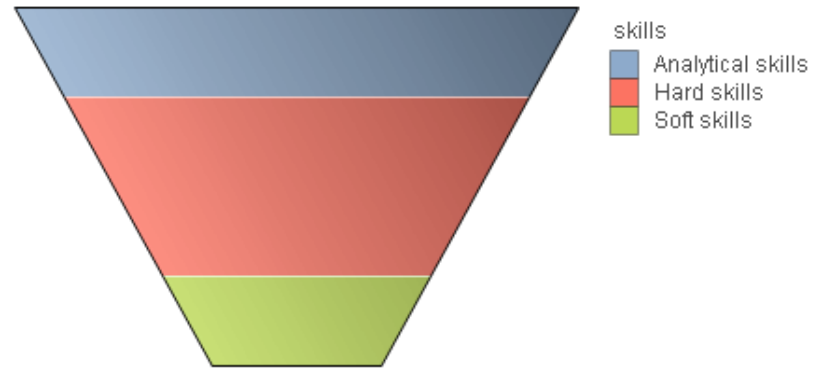

Fig.3 Distribution of the types of skills in the online training platforms

Fig. 4 provides a comparative feature of the online courses reviewed based on the skills they offer.

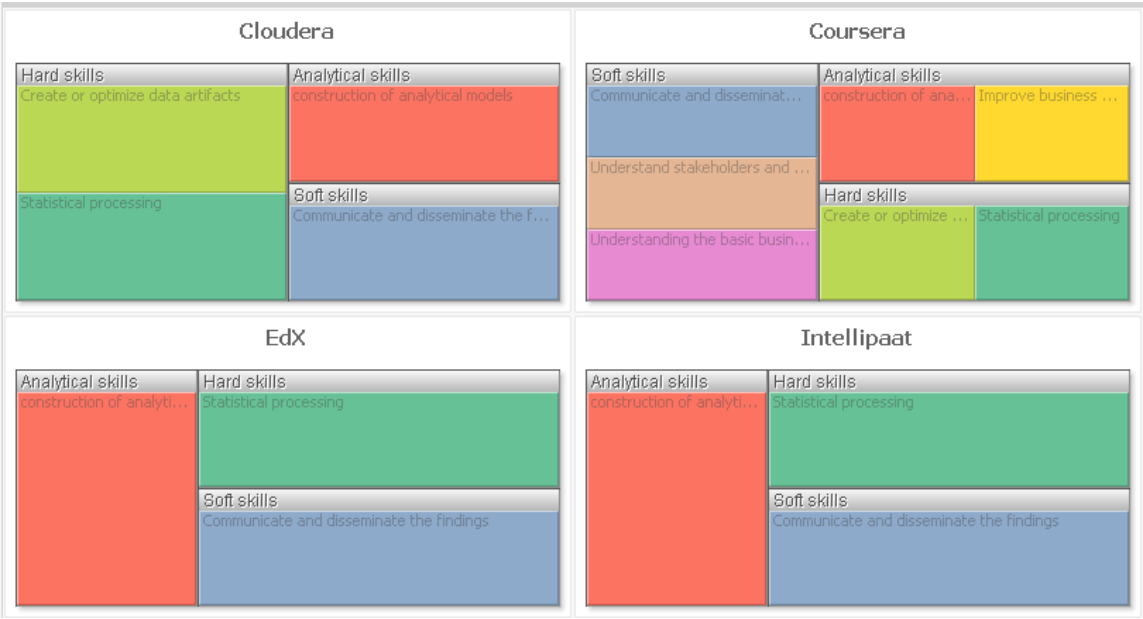

Fig. 4 Comparative characteristic of online learning platforms according to the skills they offer

It can be seen from Fig. 4 that all the skills in Table 2 are covered in the data science of the coursera. Not all courses are a study in this paper. Online platforms offer training for a wide range of learners - analysts who want to improve their skills, IT specialists who want to retrain as data scientists or even beginners who do not know data-processing technology. This study aimed to analyse and compare the existing on-line courses offered by leading online learning platforms. The main skills that these courses focus on are hard skills like programming, data modelling techniques, statistics and forecasting. This shows that the focus of the training is primarily on a group of users defined in [11, 3] as specialists. Interesting is the case of training "users" of data science or those identified as non-specialists.

7 CONCLUSION

Today, companies gather tons of data that are often overlooked or lack understanding of how to use them. This data, through meaningful information extraction and discovery of actionable insights, can be used to make critical business decisions and drive significant business change. It can also be used to optimize customer success and subsequent acquisition, retention, and growth. Data scientists can have a major positive impact on a business' success, increasing profit and reduce financial loss, which are some of reasons why it is so important to train people in data science.

Data scientists is an important and in high-demand role that can have significant impact on a business' ability to achieve its goals, whether they are financial, operational or even strategic.

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] Alex Castrounis, "What Is Data Science, and What Does a Data Scientist Do?", Mar 7, 2017, https://www.innoarchitech.com/what-is-data-science-does-data-scientist-do/

[2] Andy Kiersz and Rachel Gillett, "21 High-Paying Jobs of the Future", OCT 27, 2017, http://www.businessinsider.com/best-jobs-future-growth/#21-farmers-ranchers-and-other-agricultural-managers-1

[3] Christozov, D., Toleva-Stoimenova S., Rasheva-Yordanova K., Vukarski I. Developing Big Data Competences in the Digital Era. Big data, Knowledge and Control Systems Engineering, BdKCSE'2016. pp. 97-104. ISSN – 2367-6350.

[4] Cloudera, Inc., "Cloudera Data Analyst Training: Using Pig, Hive, And Impala With Hadoop", Training Sheet, 1-888-789-1488 or 1-650-362-0488 , 2014, https://university.cloudera.com/static/pdfs/Cloudera-Data-Analyst-Training.pdf

[5] Coursera Inc., "Data Science Specialization", 2018, https://www.coursera.org/specializations/jhu-data-science.

[6] Edith Hancock, "The 25 best jobs in the UK right now ", Business Insider UK, Jan. 30, 2017, http://uk.businessinsider.com/best-jobs-in-the-uk-in-2017-2017-1/#7-supply-chain-manager-19

[7] EdX, "Foundations of Data Science", 2012–2018 edX Inc., https://www.edx.org/professional-certificate/berkeleyx-foundations-of-data-science

[8] Intellipaat, "Data Science Training Course", 2011-2018 https://intellipaat.com/data-scientist-course-training/

[9] Investopedia, Data Science, 2018, Investopedia, LLC, https://www.investopedia.com/terms/d/data-science.asp

[10] Kirk Borne, "What is a Data Scientist?", SAS Institute Inc. 2018, https://www.sas.com/en_us/insights/analytics/what-is-a-data-scientist.html

[11] Petrova P., Boyadzhiev D., "Training young lecturers", XIV-th International Conference "Challenges in Higher Education and Research in the 21st Century", May 31 - June 3, 2016, Sozopol, Bulgaria, ISBN: 978-954-580-356-9, Heron Press Ltd., Vol.14, 2016, p.23-26

[12] 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

[13] Sarah K. White, "The 7 most in-demand tech jobs for 2018 — and how to hire for them', Today's top stories, CIO, APR 13, 2018 - https://www.cio.com/article/3235944/hiring-and-staffing/hiring-the-most-in-demand-tech-jobs-for-2018.html

[14] TechTarget network of technology, "Guide to big data analytics tools, trends and best practices", October 2017, https://searchenterpriseai.techtarget.com/definition/data-science

[15] Techterms, Data Science Definition, Updated: August 17, 2017, https://techterms.com/definition/data_science

[16] Vincent Granville, "40 Techniques Used by Data Scientists", July 4, 2016, https://www.datasciencecentral.com/profiles/blogs/40-techniques-used-by-data-scientists

[17] WhatIs? - https://whatis.techtarget.com/

[18] Wikipedia - https://en.wikipedia.org/wiki/Data_science