Skill Typology and Toolboxes of Data Experts

Dr. Alon Hasgal†  
 Department Name  
 Ramat Gan Academy  
Ramat Gan, Israel   
 alon.h@iac.ac.il

Prof. Niv Ahituv  
 Department Name  
 Tel Aviv University   
Tel Aviv, Israel ahituv@tauex.tau.ac.il

ABSTRACT

CCS CONCEPTS

• Insert CCS text here • Insert CCS text here   • Insert CCS text here

KEYWORDS

Insert keyword text, Insert keyword text, Insert keyword text, Insert keyword text

1 Introduction: Data Science

Data Science (DS) is a rapidly growing profession and field of research, now producing some of the most important products and outputs for modern industries and businesses. DS covers multiple fields and issues (defined in the “data cycle” model below). DS emerged in response to the need for high-level computing tools capable of converting big data into clear and useful information by identifying trends, patterns, and insights that can then be used to make intelligent and informed decisions. DS is valuable to managers, decision-makers, researchers, advertisers, and policy-makers in a wide range of organizational, social, and economic fields. DS has multiple applications such as: monitoring and predicting social and physical trends (i.e., human behavior, infectious diseases, or global warming), building innovative business models, customizing services and e-commerce, and in multidimensional research fields such as brain sciences (Brock & Khan, 2017). The responsibility of DS experts is to impart all the pertinent outcomes of data analyses conducted using DS tools to their clients or decision-makers, who may be managers, researchers, marketing professionals, or experts in other various fields.

This rapid development of DS is the result of many factors, including the widespread use of social media and the world-wide web, and advances in big data, artificial intelligence, human/computer interfaces, and related fields. DS is clearly of growing importance for all sectors of society and governance because the decision-making systems used by individuals, organizations, and governments are continually being updated with vast amounts of data collected from multiple, varied sources. The amount of data produced in the world exceeds 2.5 exabytes per day, and the amount of information being collected is increasing by about 20% per year. Although communication capacity is increasing by 30% per year, the number of exabytes of data being produced exceeds communication and storage capabilities (Gupta et al., 2018).

At the same time, there has been tremendous progress in the development of sophisticated technologies based on machine learning and inference techniques that can sift through the vast mass of data and enable the recognition of patterns and meanings within it. That is, by effectively integrating and comparing data from internal and external sources, DS transforms huge amounts of data into clearly understandable knowledge. This knowledge can then be used by experts in various fields to improve the quality of their decision-making processes. Studies show that organizations using advanced data analysis techniques such as artificial intelligence, machine learning, information management, and big data applications can improve their business performance (Davenport, 2017). Thus, the field of DS is considered successful and prestigious, and has significant potential for further development (McAfee & Brynjolfsson, 2012).

This article presents the tools and skills used by various types of data experts. It first presents a “data cycle” model illustrating how various data analysis tools can be used and adapted in a series of stages in the decision-making process. This serves as the generic basis for a wide range of research questions or business decisions. The article then describes the clients of data experts (i.e., senior and junior executives and managers, professionals such as engineers, government employees, scientific researchers) and assesses which tools are most appropriate for each type and at each stage of the cycle. This assessment takes into account the types of data that are needed and available: financial, scientific, human resources, statistics, data from external and internal sources, language-based data, visual data (images), graphic data, and the like. It also considers the type of problem to be solved: managerial, marketing, research, medical, and more.

**1.1 The Data Cycle Model**

Every decision-making process is based on a data cycle, which culminates in a decision being made. The cycle can be short and based on few data items, such as when we decide whether it is safe to cross the street. In such a simple case, we first identify the problem or the mission (crossing the street safely). We collect data (number of cars passing by, width of the street), and estimate our walking speed. We integrate this data, operate an algorithm based on our past experiences, analyze the results, make a decision, then store and communicate feedback for future similar activities.

Obviously, most decisions made by organizational bodies and research teams are far more complicated. However, the stages of the Data Cycle (DC) model (Fig. 1) are essentially the same at any degree of complexity, and for every sector and field.



**Figure 1: The data cycle.**

1. Problem definition: Initial definition of the problem to be solved, mission, or purpose for which data is required. Potential tools: formulation methods, quantitative models, qualitative approaches, mathematical tools, etc.
2. Identification of data sources: Determining which types of data are most pertinent to solving the problem, and where they can be located. Potential tools: internet browsers, indices, search engines, international organizations, statistics bureaus, etc.
3. Data collection and storage: Data are retrieved from various sources, validated and cleaned, then stored in an accessible format and location, and backed up. Potential tools: data transfer technology, communications, online clouds, database management software, data validation tools, etc.
4. Data integration: This essential stage allows the user to incorporate and integrate data from various sources, even if the data definitions and formats were initially incompatible or unsynchronized. Potential tools: conversion programs, indices, metadata tools, etc.
5. Data mining: Selection of relevant data from the big data. Potential tools: filters, data retrieval techniques, identification tools, artificial intelligence (AI) tools, heuristics, etc.
6. Processing and analysis: Selected data are screened, processed, and analyzed. Potential tools: algorithms, AI tools, machine learning, data processing programs, heuristics, etc.
7. Visualization: Presentation of the results to decision-makers. Potential tools: dashboard software, graphical tools, reporting systems, interactive systems, voice, UX programs, etc.
8. Learning and decision-making: This stage represents the purpose of the data cycle – making the decision based on the results of the data analysis. Potential tools: decision support tools, what-if software, simulation tools, visualization tools.
9. Feedback: This stage is not always necessary, but since certain decisions are often repeated, decision-makers can improve the usefulness and effectiveness of future data cycles by saving comments and changes and transmitting them to others as needed. This stage includes also knowledge management. Potential tools: reporting systems, interactive reactions, fine tuning tools, DevOps tools, agile design tools, machine learning, knowledge management tools, etc.

Since the needs involved at each stage are different, a wide and varied range of tools are used to support the DC. In the following sections, the tools are classified and associated with various needs, users, and types of data (Ahituv, 2019).

**1.2 Impacts of Data Analysis on the Economy**

Data analysis and DS methods can have a particularly significant impact for organizations in terms of transforming the vast amount of information collected from sources within and outside the organization into valuable knowledge, such as predicting consumer behavior, improving digital (online) visibility, personalized advertising, predicting needed or desired products or services, improving customer experiences, algorithmic trading in financial securities, discerning anomalies in financial reports, and early identification of trends in public opinion.

Although it has been proven that using advanced data analysis tools and big data applications can help organizations improve their business models and increase profits, many organizations still do not make effective use of the information they have (Brock & Khan, 2017). A study commissioned by KPMG International Data and Analytics (Thomas et al., 2016) found that only 35% of the surveyed managers use data analytics to improve their services, organizational processes, and business models. This is due, at least in part, to a lack of data experts with knowledge relevant to the needs of the organization. Therefore, the field of data analysis has great potential for major expansion and development. There is increasing demand by organizations for data analysis professionals who can locate and optimize information, perform intelligence analysis and predictive analytics, adapt and update operating models, and offer guidance and advice as needed. Data experts are being hired to work with companies of all sizes and in many industry sectors that wish to increase their market share, improve their performance, and customize their products and services to provide a better customer experience (Davenport & Patil, 2012). Recently, there has been intense activity among startup companies, new media corporations (who may need such analytics performed on a daily basis), and corporations in the sharing economy such as FinTech, InsureTech, cyber security corporations, the life sciences industry and more, whose core models are based on machine learning and methods for analysis of big data. A longitudinal survey found that graduates with degrees in data analysis are in demand to fill a broad range of positions and roles in every organizational field (Vossen, 2014). Moreover, there is a great need for advanced data analysis in many scientific research fields, such as: genetics, climate, sociology and economics, market research, geophysics, and more.

**1.3 Objective**

Over the years, the need has arisen for a more accurate typology of the professional skills that data experts need to work in the various fields mentioned. The purpose of the current article is to offer a clearly organized typology defining the skills and tools that the market requires of data experts. There has been little previous research on this topic, and most of it was based on information from professional websites pertaining to data analysis.

**2. Methods**

This meta-study presents analysis of information gathered from resources gathered by the authors and previous research by others. The proposed typology is based on a preliminary study of surveys pertaining to the skills and tools of various specializations of data experts, which had been published in academic journals and on professional websites (Ahituv & Hasgall, 2017). The study also presents a comparative analysis based on the findings of a survey of some 20 data experts working in various specialization of the field.

To understand the needs of organizations, an analysis was conducted on employment advertisements for data analysts that were placed on major employment website sites in Israel (Alljobs,[[1]](#footnote-2) SQLink,[[2]](#footnote-3) John Bryce[[3]](#footnote-4)) and a survey conducted by KDNuggets (2018). The typology reflects classification according to four main factors: market needs, complexity of data, characteristics of data, and required tools.

**3. Levels of Expertise in Data Analysis**

The preliminary study (Ahituv, 2019) identified three essential types of data experts: 1) Web analysts who analyze data collected from various digital sources and online social networks; 2) Data analysts who work with data from businesses and organizations to help make decisions, create business models, and in marketing; and 3) Data scientists, who work with vast amounts of scientific data and observations regarding unstable or chaotic conditions, with the goal of predicting trends and developing computerized models (Ahituv & Hasgall, 2017).

***Level 1: Web Analysts***

Web analysts primarily use data collected from online social activity, such as databases of companies that maintain online social networks, e-commerce sites, and other websites. Analysis of this type of data enables identification of trends and development of processes to promote engagement between users and various communities, products, and services. Web analysts generally focus on questions and problems relevant to customer services and online marketing, such as customization of advertisements for services and products. The primary tools they use are surveys, descriptive statistics, and statistical inference.

Previous research on this segment of the industry has found the clients are comprised of businesses and corporations seeking to personalize and tailor the services they offer based on customer behavior through the commercialization of data pertaining to individuals. The primary methods used are data collection, data integration, and completing the database with data from other online and offline sources (Barutçu, 2017). An interesting and innovative case study conducted in the United States found that the majority of data analysis for the trade market is conducted in order to improve segmentation of products, services, companies, and customers using open-source tools that are already available on the market, and do not necessitate development of new algorithm codes (Evgeniou & Niessing, 2017).

An analysis by the GSM Association indicates that decision-makers utilize the insights gained through data analysis to understand and improve the organizational value chain to best meet consumers’ interests, in terms of offering low prices and tailored services. Data experts can conduct these analyses without the need to develop new algorithms for each case (Kearney, 2018), so web analysists must be familiar with existing tools and how to utilize them effectively. Previous research and surveys indicate that most of the demand for data analysis in the marketing field pertains to establishing the correlations between events, business models, consumers’ current social behavior, and creation of models for predicting their future behavior. There is a need for web analysts who specialize in fields such as medicine, business and commerce, government, the financial sector, and more.

***Level 2: Data Analysts***

Data analysts develop basic algorithms for the analysis of vast amounts of data. The skills they need include advanced knowledge of mathematics, statistics, programming, algorithmic studies, models of complexity, machine learning, and artificial intelligence. Data analysts address complex questions and problems, develop models to analyze the existing situation, and perform predictive analytics. They often specialize in data analysis for various sectors of business and industry. They are less focused on deep learning or DS research (Ahituv & Hasgall, 2017).

The main role of this type of data expert, as shown in the DC model, is data processing and analysis. Professionals working at this level, including economists and statisticians, have extensive learning and experience in various types of information systems and data analysis, with a focus on database management using pre-defined and open-source (free of charge) algorithms to analyze data related to business, social, or organizational issues.

Data analysts use multidisciplinary, systemic, and creative thinking, which that are “essentially human”. That is, their insights are derived from human intellectual abilities, not exclusively from mathematical analysis of data (Knowles-Cutler, 2018). Data analysts develop competitive intelligence, analyze trends, and assess the potential of new products and services, taking in to account the social processes, available innovative technologies, market trends, and economic issues that are relevant for the organizations they work with.

Data analysts address a wide range of phenomena: consumer behaviors, how people intend to vote in elections, habits in television watching and internet surfing, crime statistics, population trends, companies’ financial performance, competition for customers between large supermarkets and discount chain stores – the list is virtually endless.

Given the vast amount of information currently being collected by organizations, there is a growing need for data analysts with the ability to work in a complex technological environment.

Overall, a data analyst is responsible for retrieving, collecting, and organizing data, then using the data to reach meaningful conclusions. Analysts examine how data can be used to answer questions and solve problems. They are “digital detectives” striving to understand how various processes work and in order to help others understand them; thus, being a data analyst has the potential to be a challenging, creative, and rewarding career.

A comprehensive survey of the skills and capabilities of data analysts, conducted by the McKinsey Corporation found that main role of professional data analysts in today’s market involves identifying innovative business opportunities and optimally tailoring products and services to customers (Court, 2015). This is accomplished through: data mining from primary and secondary sources; cleansing and organizing raw data; utilizing statistical, mathematical, and/or technological tools and analytic processes to get a broad view of the data; and identifying and analyzing interesting trends in the data. Data analysts identify opportunities for improving an organization’s operational processes. They present technical analyses of the results to business clients or internal teams by creating reports, visualizations, and dashboards to help them interpret and understand the data, make informed decisions, and set goals.

Based on the above, the skills on which data analysis is based may be referred to as “creation of quantitative mental models” (as opposed to DS, which focused on creating mathematical models per se, as discussed below). Thus, a highly skilled data analyst must have the skills and ability to do the following:

* Fully and deeply understand the needs of the specific client
* Make a work plan: become familiar with the organizational environment and market; formulate questions and sub-questions for investigation that will lead to evidence-based and relevant recommendations; convert a qualitative question (for example regarding a business practice) into quantitative data that is validated and integrated with qualitative information, which taken together describe the existing situation; optimize the data in order to find an answer or response to the research question(s)
* Carry out a professional work process: identify a relevant database, organize it, select data from it in accordance with the needs of the question to be researched; create and present clear PivotTables and graphs that confirm a reality of the market (this differs from academic research questions, which are designed to uncover new knowledge)
* Design a detailed and professional methodology of the work process; clearly articulate the relevance of the data to be analyzed and what can be proven by it
* Present and visualize insights in a manner that will have an immediate and convincing influence on the customer/decision-maker

***Level 3: Data Scientists***

This is a high-level specialization that involves the design of complex models and development of the necessary original algorithms for the modeling, computer learning and machine learning that form the basis for artificial intelligence. Only those who studied computer science, high-level statistics, and DS are qualified to work at this level. Data scientists must be able to cope with a large mass of chaotic, unsynchronized, and apparently incomprehensible data and to create data matrices (clustering) from them. They develop and implement advanced statistical methods to create a new structure for the data that show the changing situation and can be used in forecasting and predictive analytics.

**3.1 Typology of Data Experts**

Table 1 summarizes the typology of data experts and the skills required by each, thus illustrating the differences between them.

**Table 1**

*Characterizations of the Three Levels of Data Experts*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Data Expert (level)** | **Characteristics of the Data** | **Complexity** | **Work Processes** | **Demand in the Market, Clients** |
| **Web analysts** | Information on websites and the internet | * Marketing * Tailoring products and services to customers * Customer service | * Posts on social networks * User profiles * History and characteristics of online activity | * Marketing managers * E-commerce managers * Customer service managers |
| **Data analysts** | * Medium-sized data mass, arranged in tables * Economic, business and budgetary data (sales, income, expenses, etc.) * Data from online social activity | * Business models * Databases on competitive business practices and social trends * Realizing the potential for immediate application of micro-data * Application of algorithmic or mechanical processes to existing knowledge | * Statistical analysis on existing data related to a business strategy * Design and maintenance of data systems and databases * Demonstrating patterns in data * Predictive analytics * Preparing reports * Visualization of results | * Senior executives * Department heads * Consulting firms * Startups and developing businesses * Production managers |
| **Data scientists** | * Large mass of unstructured, unsynchronized data * Digital signals * Internet of Things (IoT) | * Level of strategizing * New trends * Making predictions for business and social issues | * Management of large-scale structured and unstructured data * Production of databases * Modeling databases using algorithms and scientific observations * Creating new models * AI-based analysis methods, deep learning machine learning * Developing search engines and content-based recommendation systems | * Managers * Senior executives Research directors * Consulting and forecasting firms * Academic and research institutes |

The first type, web analysts, focus on data from online social activity. In contrast, data analysts focus on the emerging needs of businesses in various sectors and content worlds. They specialize in dealing with a medium-sized data mass relevant to a specific research question, on which they can perform BI analysis, trend analysis, or analysis of the relationships between specific business models and organizational activities. The third type, data scientists, use intelligent algorithms and machine learning to deal with a large mass of data, including unstructured data that covers a broad range of activities and contents and is not necessarily linked to a specific research question. Data scientists often alert their clients to new issues to be examined or suggest they change existing processes, based on their analysis.

**Core Skills and Tools of Types of Data Experts**

This section shows the connections between the demands and needs of organizations, the challenges facing the various data experts, and the tools each type uses, as given in Table 2.

**Table 2**

*Data Experts’ Tools and Skills, Stages in the Data Cycle, and Types of Clients*

***Skills and Tools of Web Analysts***

The main tasks of web analysts are: identifying trends, usage patterns, mapping ongoing processes (patterns of online interaction or shopping), analyzing, categorizing, and ordering quantitative data and continuous variables. They must understand how to utilize statistical indices (means, standard deviations, medians, percentages), data distributions and histograms, correlations and causalities. They detect exceptions to expected patterns and attempt to explain them. They use website management tools such as change monitors, aggregators from social media, content websites, and news sites. They must have strong ability for analyzing qualitative data such as observations and interviews, and for integrating quantitative and qualitative data. They must also be able to use Google tools such as Google Trends and Google Analytics.

***Skills and Tools of Data Analysts***

Data analysts’ work is focused on specific questions. They must be proficient in retrieving data using search engines and targeted search tools for locating online information, databases, profiles, companies, images, and videos. They use organizational diagnostic tools for strategic diagnostics and business models. “Essentially human” skills such as systemic, comprehensive, and multidisciplinary thinking. They must have strong ability for managing databases using Excel, and experience with Structured Query Language (SQL), Python / R, and JavaScript.

For dealing with business issues, data analysts should be able to operate Business Intelligence (BI) systems. They use open-source artificial intelligence algorithms to sort and filter data and present it in the appropriate context. Data analysts present their integrated insights to clients using display systems, infographics, symbols, charts, and BI visualization tools such as Tableau, Power BI, Qlik and microservices tools and technologies. Each BI tool has its own strengths, though there are high-quality, freely available tools such as Power BI, which can create end products that are equivalent in quality to those produced by Tableau. BI tools are useful in evaluating the quality and features of data, including identifying trends and generating insights. These products are designed for technical and non-technical users, and include “drag-and-drop” interfaces to enable users to easily and seamlessly create accessible visualizations.

At each stage of the DC, data analysts must consider the needs of the project’s clients. For example, data visualization gives clients a view of the data that would otherwise be impossible. Therefore, effective data analysts know how to appropriately leverage technologies such as dashboards and other platforms for visually integrating insights from the data. In comparison to traditional flat files, dashboards make data more tangible and easily understandable for decision-makers and other end users. Another important aspect of data analysts’ work involves merging and integrating data from various sources based on shared keys, group aggregations, and statistical calculations. Data analysts must be able to program, write, and use various scripts. They manage databases, address challenges in storing and processing multiple databases, including classical relational databases, and conduct analysis of qualitative data.

***Skills and Tools of Data Scientists***

Data scientists are expected to have a deep understanding of the field in which they are working so they can quickly understand the data and the relevant issues. For example, a data scientist who works for a retailer would know that December is their busiest month, and thus will be able to foresee logistical problems or irregular sales behavior. Data scientists must communicate with team members and stakeholders. Data scientists need strong technical knowledge, problem-solving skills, and be proficient in using code languages. By understanding the mathematics (especially statistics) on which machine learning models are based, data scientists are able to make adjustments and improve model performance.

Prospective employers want candidates who are skilled in data wrangling and data preparation, since the datasets they work with are often unruly or chaotic and require pre-processing. Skills in machine learning and modeling are core skills of data scientists, enabling them to offer high-quality solutions to organizations. As datasets grow ever-larger, so does the demand for data scientists who can interpret and gain insights from big data. Since traditional analysis tools are limited, a specific set of scientific tools is needed for the analysis and manipulation of very large data sets, such as: BI data simulation tools, NoSQL languages (which may sometimes be effective replacements for SQL), and frameworks for processing big datasuch as Apache Hadoop and Apache Spark. Since the infrastructures of Google Cloud, Microsoft Azure, and Amazon Web Services are similar to that of the open-source Hadoop system, learning Hadoop is an advantage. Spark utilizes up-to-date and high-power graphic processing units (GPUs) and includes a machine learning library and built-in capabilities for streaming data such as web data or stock prices, allowing for analysis in real time. Hadoop is also known for YARN (Yet Another Resource Negotiator).

To address these topics, data scientists must know how to develop machine learning models, ranging from traditional statistical models such as linear models or support vector machines (SVMs) through the latest deep networks. To be successful and influential in their field, data scientists must understand machine learning models, when each can be applied, and their limitations. Thus, data scientists should strive to continually develop their skills in the area of predictive modeling, since organizations around the world are interested in predicting trends, classifying customers, or developing innovative technical solutions. Skill in predictive analytics is essential. The StrataScratch platform offers practice problems for testing and improving skills and knowledge on topics such as the benefits of specific models, how to adjust and improve model performance, and classification of missing values.

A survey conducted on the Kaggle website (Kaggle, 2017; KDNuggets, 2018) indicates the main tools and areas of interest for data scientists: supervised learning models, unsupervised learning models, deep learning frameworks, deep neural networks, algorithms for recommender systems, methods for the use and troubleshooting with classification, and advanced regression analysis techniques.

Data scientists identify behavioral patterns and construct behavioral models. Figure 2 shows an analysis of the findings of a survey examining the main challenges faced by data scientists.

Graphical user interface

Description automatically generated with low confidence

**Figure 2: Challenges faced by data scientists at work.**

Figure 3 shows the common methods used by data scientists in their work, from the same survey.

Chart, bar chart

Description automatically generated with medium confidence

**Figure 3: Methods used by data scientists at work***.*

The results of the survey show that data scientists face a wide range of professional challenges, and which are not necessarily connected to each other. Data scientists need to fully understand data analysis methods and technologies such as: order management, group templates for data, artificial intelligence, neural networks and deep learning (sub-fields of computer science and information science), as well as statistical methodologies.

Adapting data to a business strategy requires an education in business administration and information systems; adapting data to the prevailing social culture requires an understanding of the social, sociological, and cultural spheres. The data must be presented to clients and managers in appropriate language. This requires understanding of the customer experience, human-machine interfaces, and the usability, accessibility, and user-friendliness of the data. Data scientists often specialize in a chosen field such as medicine, business and commerce, government, or finances, among others.

Key skills and fields of interest that distinguish data scientists from the other types of data experts (web analyst/data analyst) is the data scientists’ ability to work with advanced processes such as forecasting and building statistical models. Data scientists predict the future using data from the past. They use algorithms in recommendation systems, supervised and unsupervised learning models, deep learning frameworks, and deep neural networks.

תמונה שמכילה שולחן

התיאור נוצר באופן אוטומטי

**Figure 4: Data scientists’ skills.**

As shown in Figure 4, the tools needed by data scientists are diverse and sophisticated. As datasets grow ever-larger, so does the demand for data scientists who can interpret and gain insights from big data. Since traditional analysis tools are limited, a specific set of scientific tools is needed for the analysis and manipulation of very large data sets, such as BI data simulation tools. Data scientists work with technologies for distributed storage and parallel processing (Hadoop, Spark) for big data. Their work includes data cleaning and processing, using models for supervised and unsupervised learning, creating new classifications from data chaos, forecasting, building statistical models, and analyzing visual and audio data. They integrate IoT with machine learning and artificial intelligence to create deep learning. Following these processes, the data scientist presents data to decision-makers, using visualization technologies.

Therefore, data scientists must be proficient in programming, scripting, and code development in relevant languages such as: R, Python, SQL, Java / Scala, and C / C ++. Their work processes include data aggregation, variable reduction, data cleaning, ordering, and integration (using KNIME or Pentaho for example), as well as object-oriented programming languages that arrange data in multidimensional arrays and non-relational systems (also known as NoSQL), problem-solving using advanced classification/regression, and especially document-management systems.

The survey conducted by KDNuggets (2018) and posted on the Kaggle website shows the main tools recommended for data scientists: supervised and unsupervised learning models, deep learning frameworks, deep neural networks, algorithms in recommendation systems, methods and problem-solving with advanced classification / regression.

A picture containing timeline

Description automatically generated

**Figure 5: Code languages used by data scientists.**

Figure 5 shows the code languages required by data scientists, especially NoSQL languages, which are sometimes a useful replacement for SQL. The growing size of databases in many industries necessitate proficiency in the use of platforms for big Data, especially Apache Hadoop and Apache Spark. Since the infrastructures of Google Cloud, Microsoft Azure, and Amazon Web Services are similar to that of the open-source Hadoop system, learning Hadoop is an advantage. To address these topics, data scientists must know how to develop machine learning models, ranging from traditional statistical models such as linear models or support vector machines (SVMs) through the latest deep networks. Familiarity and expertise with available machine learning models is one of the areas where data scientists can be most influential. As a result, data scientists should strive to continually develop their capabilities in predictive modeling, which is needed by organizations around the world who are interested in predicting trends, classifying customers, or building new technical solutions. Skill in predictive analytics is essential in DS, therefore, data scientists need understand machine learning models, their uses, and limitations. The StrataScratch platform offers practice problems for testing and improving skills and knowledge on topics such as the benefits of specific models, how to adjust and improve model performance, and classification of missing values.

**4. Conclusions**

In order to make informed decisions in today’s world, business organizations, public and governmental agencies, and individuals all rely on data experts who are able to understand and make sense of vast amounts of data. Professional data experts have a wide range of skills, tools, and resources that use in each stage of the data cycle, including data collection, cleaning, integration, and analysis, which enable them to offer recommendations. They also assist in long-term storage of data and management of organizational knowledge. The data analyses are used to draw conclusions and make predictions. Data experts (analysts and scientists) who have abilities in collecting, sorting, storing and managing data, and creating useful knowledge from them, are key in every organization in the business, public, or academic sectors.

This field has experienced major development in recent years in throughout all sectors of industry in general and the high-tech sector in particular. There is also a high demand in the public sector for skilled data experts. There is a perceived lack of such experts, which is expected to become even greater in coming years.

An in-depth investigation of the skills, methods, and tools needed for this work led the authors of this article to conclude that these are different types of data experts, differentiated by the level of data management and analysis, the types of data they work with, the tools they use, and the products and outputs they provide: web analysts, data analysts, and data scientists.

Web analysts enable informed and focused customer-supplier interactions. They have extensive knowledge the integration and management of data pertaining to social interactions, social networks, and use of other online data and applications that enable interactions between people or between people and services/products. They must understand how to use models for marketing and social psychology, and be familiar with specific products and services. This type of professional is in demand by online corporations and in the fields of advertising, marketing, customer services, and sales of products and services available in physical or digital spaces.

Data analysts use existing data to generate insights that can be used to guide strategic and operational decisions. They integrate formal and informal information and human intelligence (Humint). They must have broad knowledge of the target market, and the organization’s strategy and needs. They know how to find and integrate relevant data, primarily using existing applications and algorithms, open-source and pre-programmed modules, and digital tools for locating and analyzing data.

Data scientists create an understandable and intelligent order out of chaotic big data that are seemingly uncorrelated. They must know how to use AI, machine learning, deep learning, and deep neural networks, in order to recognize patterns and correlations, make forecasts and predictions, and develop innovative business methods and models, based on chaotic databases.

In addition to these three basic types of data experts, there are many other professionals engaged in information retrieval (information managers or database managers), creating architectures of data structures (formerly called database administrators, now referred to as data architects or data engineers). Together, they create a professional array of experts who know how to advance a critically important resource: data. Organizations that work effectively with these professionals will succeed in today’s competitive market over those that do not.

REFERENCES

Ahituv, N. What should be taught in an academic program of data sciences? *Proceedings of the Digital Presentation and Preservation of Cultural and Scientific Heritage, vol. 9* 2019. Sofia, Bulgaria: Institute of Mathematics and Informatics.

Ahituv, N., Hasgall, A. Adapting undergraduate programs of data science to the professional requirements of the industry. *Research in Geoinformatics: Proceedings of the Geophysical Center of the Russian Academy of Sciences 5*, 1 (2017), 157.

Brock, V., Khan, H. U. Big data analytics: Does organizational factor matters impact technology acceptance? *Journal of Big Data 4,* 1 (2017), 1-28. <https://doi.org/10.1186/s40537-017-0081-8>.

Barutçu, T. M. Big data analytics for marketing revolution. *Journal of Media Critiques 3,* 11 (2017), 163-171.

Chapman, P., Clinton, J., Kerber, R., Khabza, T., Reinartz, T., Shearer, C. Wilrth, R. CRISP-DM 1.0: Step-by-step data mining guide. *SPSS Inc.* *9*, 13 2000.

Court, D. *Marketing & sales: Big data, analytics****,*** *and the future of marketing & sales*. McKinsey & Company, 2015; <https://www.mckinsey.com/~/media/McKinsey/Business%20Functions/Marketing%20and%20Sales/Our%20Insights/EBook%20Big%20data%20analytics%20and%20the%20future%20of%20marketing%20sales/Big-Data-eBook.ashx>

Davenport, T. H. & Patil, D. J. Data scientist: The sexiest job of the 21st century. *Harvard Business Review 90*, 5 (2012), 70-76.

Evgeniou, T. & Niessing, J. *Data analytics****:*** *A marketing segmentation case study*. Insead, 2017; <https://inseaddataanalytics.github.io/INSEADAnalytics/BoatsSegmentationCaseSlides.pdf>

Gupta, P. K., Tuncer Ö., & Mayank, S *Predictive intelligence using big data and the internet of things*. IGI Global. 2018.

Kaggle. *Machine learning & data science survey*. Kaggle, 2017; <https://www.kaggle.com/kaggle/kaggle-survey-2017>

KDNuggets. *The 13th annual KDNuggets software poll*, 2018; <https://www.kdnuggets.com/polls/2012/analytics-data-mining-big-data-software.html>

Kearney, A. T. *The data value chain*. GSM Association, 2018; <https://www.gsma.com/publicpolicy/wp-content/uploads/2018/06/GSMA_Data_Value_Chain_June_2018.pdf>

McAfee, A. & Brynjolfsson, E. Big data: The management revolution. *Harvard Business Review 90*, 10 (2012), 60-68.

Thomas E., Murray R., Brad F., & Anthony C. Building trust in analytics: Breaking the cycle of mistrust in D&A. *KPMG*, 2016.

1. Vossen, G. Big data as the new enabler in business and other intelligence. *Vietnam Journal of Computer Science 1*, 1 (2014), 3-14.

Conference Name:ACM Woodstock conference

Conference Short Name:WOODSTOCK’18

Conference Location:El Paso, Texas USA

ISBN:978-1-4503-0000-0/18/06

Year:2018

Date:June

Copyright Year:2018

Copyright Statement:rightsretained

DOI:10.1145/1234567890

RRH: F. Surname et al.

Price:$15.00

1. <https://www.alljobs.co.il/SearchResultsGuest.aspx?page=1&position=1732&type=&city=&region=&gclid=CjwKCAjw4qCKBhAVEiwAkTYsPAbe-iU7zfiIvDoULQxwSz0xdDltmpS-LLvXAj24qQ6COmUWumPHqBoCwfEQAvD_BwE> [↑](#footnote-ref-2)
2. <https://www.sqlink.com/career/bidbabig-data/data-analyst/?utm_source=google&utm_campaign=DSA_IT&AgId=%D7%93%D7%90%D7%98%D7%94_%D7%90%D7%A0%D7%9C%D7%99%D7%A1%D7%98&utm_term=&AdPos=&utm_content=520118266557&device=c&GeoLoc=1007991&content_site=&utm_medium=cpc&gclid=CjwKCAjw4qCKBhAVEiwAkTYsPAAxXGwHPyJKdFnZRIBSICEDtRFGGVpJivtr6Gsj83WQjClFVqwXCRoC6gEQAvD_BwE> [↑](#footnote-ref-3)
3. <https://www.johnbryce.co.il/jobs> [↑](#footnote-ref-4)