**A Toolbox and Typology of Skills for Data Analysis Experts**

Dr. Alon Hasegal, Ramat Gan Academy

Prof. Niv Ahituv, Tel Aviv University and Peres Academic Center

**Introduction: Data Science**

Data Science (DS) is an emerging profession and field of research, and is now producing some of the most important products for modern industries and businesses. DS has developed, in part, due to the need for high-level computing tools that are able to handle vast amounts of data and recognize trends, patterns, and insights that can be used for to make decisions and for research purposes. 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. While communication capacity is increasing by 30% per year, the number of exabytes being produced exceeds communication and storage capabilities (Gupta et al, 2018).

This immense amount of data has led to the rapid development of DS, which includes a number of topics (defined below in the “data cycle” model) related to the conversion of big data into clear and useful information from which relevant meanings can be derived and used for making logical and informed decisions. The results of these data analyses can be used by managers, decision-makers, researchers, advertisers, and policy-makers a wide range of organizational, social, and economic fields 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 a Data Analyst (DA) is to impart all the pertinent outcomes of the data analysis that exploits the tools of DS to the end “client”, namely, the decision maker, be it a manager, a researcher, an expert in a field, a marketing professional, and the like.

The 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 interface, and related fields. Clearly, DS is of growing importance for all sectors of society and governance. The decision-making systems used by individuals, organizations, and governments are constantly being updated with vast amounts of data collected from multiple, varied sources.

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 vast amounts of data, and enable the recognition of patterns and meanings within the data. That is, by effectively combining and comparing data from internal and external sources, DS transforms huge amounts of information 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 already considered successful and prestigious, and has significant potential for further development (McAfee & Brynjolfsson, 2012).

This article presents the toolbox of techniques used by data analysts along the data cycle. The article first delineates the “data cycle” model, which illustrates how various data analysis tools can be used and adapted in the series of stages. This serves as the generic basis for a wide variety of research and decision-making processes. The article then describes the “clients” of data analysts (e.g., senior and junior managers; professionals such as engineers, government employees; scientific researchers) and assesses which tools are most appropriate for each type of data, 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.

**The Data Cycle**

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 and distance 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 (i.e., experience), 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) are essentially the same at any degree of complexity, and for every sector and field. Figure 1 portrays the Data Cycle, followed by an explanation of each stage and its potential tools.

Fig. 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. Identifying pertinent 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 storing (including cleansing and backup): retrieval of data from various sources and storing it in an accessible format and location; validation and cleansing of data. Potential tools: data transfer technology – communications, clouds, database management software, data validation tools, etc.
4. Data integration: This essential stage allows the user to incorporate data from various sources whose data definition and format were not initially compatible or synchronized. 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, 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 maker(s). 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, the decision maker can improve the usefulness and the effectiveness of future cycles by saving comments and changes and forwarding 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, we classify the tools and associate them with various needs, users, and types of data (Ahituv, 2019).

**Impacts of Data Analysis on the Economy**

Data analysis can have a particularly significant impact for organizations that have adopted DS methods in order to transform 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 securities (stocks, bonds, etc.), discerning anomalies in financial reports, and early identification of trends in public opinion.

Although it has been shown that using advanced data analysis tools and big data applications can help organizations improve their business models and increase their profits, many organizations still do not make the 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 analysis experts who have knowledge that is 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 for data analysis professionals who can locate and optimize information for an organization, perform intelligent data analysis and predictive analytics, adapt operating models according to up-to-date and measurable processes, 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 the corporations in the sharing economy such as Fintech, Insuretech, Cyber Security, Life Science and more, whose core models are based on machine learning and big data methods. 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.