# What Should be Taught in an Academic Program of Data Sciences?

Niv Ahituv

[1] Coller School of Management, Tel Aviv University, Tel Aviv, Israel
ahituv@tauex.tau.ac.il

**Abstract.** The new academic discipline of Data Sciences (DS) has been developed in recent years mainly because of the need to make decisions based on huge amounts of data -- Big Data. In parallel, there has been a huge progress in the development of technologies that enable to identify patterns, to filter big data, and to provide relevant meanings to information, due to machine learning and sophisticated inference techniques. The profession of Data Scientist (or Data Analyst) has become highly demanded in recent years. It is required in the business sector where data is the "oxygen" for business survival; it is needed in the governmental sector in order to improve its services to the citizens; and it is very imperative in the scientific world, where large data depositories collected in varied disciplines have to be integrated, mined and analyzed, in order to enable interdisciplinary research. The purpose of this paper is to demonstrate how the scientific discipline of Data Sciences fits into academic programs intended to prepare data analysts for the business, public, government, and academic sectors.

The article first delineates the Data Cycle, which portrays the transformation of data and their derivatives along the route from generation to decision making. The cycle includes the following stages: problem definition → identifying pertinent data sources → data collection, and storing (including cleansing and backup) → data integration → data mining → processing and analysis → visualization → learning and decision-making → feedback for future cycles. Within this cycle, there might be sub cycles, where a number of stages are repeated and reiterated. It should be noted that the data cycle is generic. It might have slight variations under various circumstances, however, there is not much difference between the scientific cycle and all the other cycles.

Each stage within the cycle requires different tools, namely hardware and software technologies that support the stage. This article classifies these tools. The final part of the article suggests a typology for academic DS programs. It outlines an academic program that will be offered to those wishing to practice the Data Analyst profession. An introductory course that should be mandatory to all students campus-wide is sketched.

**Keywords:** Data Sciences, Big Data, Data Mining, Academic Program in Data Sciences, Data Analyst.

Digital Presentation and Preservation of Cultural and Scientific Heritage. Conference Proceedings. Vol. 9, Sofia, Bulgaria: Institute of Mathematics and Informatics – BAS, 2019. ISSN: 1314-4006, eISSN: 2535-0366

## 1 Introduction

Data Sciences (DS) have developed in recent years mainly because of the need to make analysis and decisions based on huge amounts of data – Big Data. In parallel, there has been a huge progress in the development of technologies enabling to identify patterns, to filter big data, and to provide relevant meanings to information, due to machine learning and sophisticated inference techniques. DS deal with handling and transforming of vast amounts of data into clear and useful information for decision-making. The amount of data produced in the world is more than 2.5 Exabyte per day; the communication capacity increases by 30% per year; and the amount of information collected grows by 20% annually (IDC, 2014). Experts in the field can improve the quality of decisions by properly utilize the organization and external data sources. Research shows that organizations that use advanced analytics, including artificial intelligence and machine learning, information management and Big Data applications, generate improvements to their business performance (Davenport, 2017). The Harvard Business Review defined the domain of data analysis as a prestigious field with a wide potential for development (McAfee & Brynjolfsson, 2012). New-media corporations such as Google, Facebook, Twitter, and Waze deal daily with data collection and analysis based on organizational and business models that they maintain. Technology companies such as Intel, companies from classic industries such as pharmaceutics and banking, trading companies such as eBay, many start-up companies, and small and medium-sized companies are also interested in using their internal and external data to increase market share, and to improve product customization. Recently, there has been intensive activity in the areas of Fintech, Insuretech, Cyber Security, Life Science, and others.

Governments and the public sector organizations also rely on huge and integrated depositories of data in order to better the service provided to citizen. This approach enables the provision of a "one stop station" for the citizen using the services (privacy issues will not be discussed here).

Scientific research is moving very fast in the direction of encouraging multi-disciplinary studies such as brain research, global warming, genetic sequencing, global macroeconomics, and many more, where data collected from various sources, under the standards of distinct disciplines whose definitions and formats are disparate, need to be integrated and analyzed.

Therefore, it is obvious that the introduction of DS into the academic arena is inevitable. Scholar and research students will be helpless without some background in DS. The following sections will portray how a DS program should be designed.

## 2 The Data Cycle

Every decision-making process is based on a data cycle culminating in a decision being made. The cycle can be very short and based on a 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), collect visual data on cars passing by, estimate the width of the street and our walking speed, integrate this data, operate an algorithm based on our

past experience (i.e., machine learning), analyse the results, make a decision, store feedback for future similar activities.

Obviously, most of the decisions taken by organizational bodies and by teams are much more complicated. However, the stages of the Data Cycle (DC) are nearly the same for each degree of complexity, in each sector, and for each discipline. Figure 1 portrays the Data Cycle.

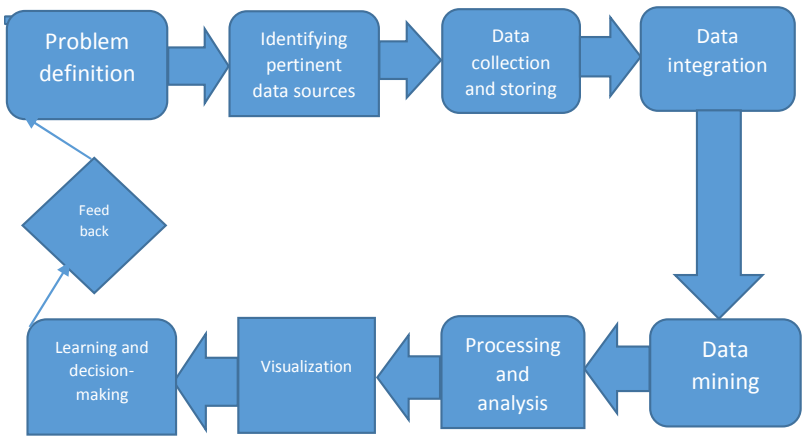


**Fig. 1.** The Data Cycle

We will briefly describe each stage of the DC and list out potential tools that can support each stage:

1. **Problem definition:** An initial definition of the problem, or the mission, or the purpose, for which data is required. ***Potential tools:*** *formulation methods, quantitative models, qualitative approaches, mathematical tools, and the like.*
2. **Identifying pertinent data sources:** Understanding what data are pertinent, and where they can be located. ***Potential tools:*** *browsers, indices, search engines, international organizations, statistics bureaus, and the like.*
3. **Data collection and storing (including cleansing and backup):** retrieval of data from various sources and store them in an accessible location; data validation and cleansing. ***Potential tools:*** *data transfer technology – communications, clouds, database management software, data validation tools, and the like.*
4. **Data integration:** This (very important) stage should allow the user to incorporate data from varied sources whose data definition and format were not initially compatible, nor are they synchronized. ***Potential tools:*** *conversion programs, indices, metadata tools, and the like.*
5. **Data mining:** Selection of relevant data out of the Big Data. ***Potential tools:*** *filters, data retrieval techniques, identification tools, AI tools, heuristics, and the like.*

6. **Processing and analysis:** The data that were selected earlier are now screened, processed, and analyzed. ***Potential tools:*** *algorithms, AI tools, machine learning, data processing programs, heuristics, and the like.*
7. **Visualization:** Presentation of the results to the decision maker(s). ***Potential tools:*** *dashboard software, graphical tools, reporting systems, interactive systems, voice, UX programs, and the like.*
8. **Learning and decision-making:** The final stage that is the purpose of the data cycle. The results are displayed to the decision makers and decisions are taken. ***Potential tools:*** *decision support tools, what-if software, simulation tools, visualization tools.*
9. **Feedback for further cycles:** This stage is not always necessary. However, very often, the need to make a certain decision is repetitive, so the customer (the decision maker) can affect the usefulness and the effectiveness of the cycle by forwarding comments and changes. ***Potential tools:*** *reporting systems, interactive reactions, fine tuning tools, DEVOPS tools, agile design tools, machine learning, and the like.*

As seen in this section, the needs in each stage are different. There is a large variety of tools that support the DC. The purpose of an academic program in DS is to expose the students to the various stages and tools and to train them either to be employed in the industry or in the academia as Data Analysts or Data Scientists, or to proceed in research and teaching by developing an academic career in the DS area.

In addition to an academic program in DS, it is recommended to offer an introductory course in DS that will be mandatory for every student, particularly in "soft" disciplines such as the Humanities, Social Sciences, Law, Management, Arts and the like. This course should be the gate to the contemporary digital world.

The next section describes the typology of DS programs. Then, an example of a whole program will be displayed, and finally, the mandatory introductory course will be presented.

## 3 Data Sciences Programs Typology

After examining and analysing various DS programs[1], it appears that the prevailing programs in the DS field are not identical in neither the content studied nor in the subjects, the depth and the expansion of the learning. In our opinion, the programs can be divided into three types (knowledge levels):

- Program Type "A" – Specialization in digital spatial data analysis - WEB analysis: This program focuses on analysing data retrieved from the digital world. Development of algorithms is not required but only familiarity with those available "on the shelf". This level focuses on the study of models and methods and the use of existing open source systems. The study focuses on the ability to choose software for the analysis of information, and to build queries relevant to

[1] The evolving IT landscape. 2017. What's The Big Data? https://whatsthebigdata.com/2012/08/09/graduate-programs-in-big-data-and-data-science/

managerial functions in the digital world, mainly to marketing and personal customization of advertisements, services, and products.

- Program Type "B" – Data Analysis: In these programs, the graduate learns to develop basic algorithms for data management. The programs require a higher level of mathematics, statistics, and algorithmic studies, as well as models of complexity. They involve machine learning and artificial intelligence. The student learns how to program and to develop data analysis systems, how to master advanced statistics, how to use artificial intelligence, and how to develop complex queries. The student should be able to construct models for analysing an existing situation and predicting future forecast. Graduates of this program have the skills to comprehend all areas of analysis and the huge data currently available in the market. These practical programs emphasize specialization in business data analysis in various industrial sectors. Less emphasis is dedicated to machine learning and deep learning.
- Program Type "C" – Theoretical material and advanced programming: This program focuses on the scientific and theoretical study of data and information: At this level, the graduate learns to develop complex and original algorithms for building system models and predicting future processes. This program requires a great deal of theoretical learning in the fields of algorithms, computer science, machine learning, AI, and related approaches. It is usually offered under the Computer Sciences "umbrella".

Type "C" programs are also taught in advanced degrees. Studying for a Master's degree in this field requires extensive prior knowledge in Mathematics and Computer Sciences. The student learns how to use artificial intelligence and how to study neural networks in order to detect anomalies and create optimal models.

In sum: Type "A" program prepares graduates mainly for digital marketing firms and advertising agencies. The graduates work mainly as technician of the profession and need guidance from experts.

A graduate of the "B" type program can manage the entire data-cycle, and turn it into valuable knowledge: from the data collection process using advanced mining methods, via the employing appropriate algorithms and data management systems, and up to the implementation of operational systems for visualization.

Type "C" program graduates are very sought after by the big New Media corporations. These graduates are sought more in R&D organizations, and rather less in business or governmental bodies. The graduates are very important for the development of the theory, tools and research in the DS field, leading to major scientific breakthroughs and advances in this field. These graduates can lead complex DS projects, train experts, and manage teams of technical personnel.

According to an analysis of about 200 wanted ads in the DS field, published by organizations in different fields and sectors in Israel, 70% of the ads were from business organizations and 30% from government organizations. About 20% of the ads were from large organizations [over 5000 employees] and the rest were published by medium and small organizations. The conclusion from this analysis is that the program Type "B is the most demanded in the industry. Hence, the next sections present the structure of such a program.

# 4 Structure of the Undergraduate Degree Program in Data Science Type "B"

Type "B" program should be composed of several categories of courses:

- **Introductory courses:** Courses that build the student's knowledge base. The courses include introductory courses in programming, mathematics, statistics, databases, and some basic courses in the social sciences, in order to enable the graduate to connect methodologies with academic disciplines needs.
- **Fundamental courses:** These courses enable the student to study advanced statistical methods, probability theory, data structures, and communication. The student should be exposed to the fundamentals of data mining, data analysis and the use of business intelligence as a basis for understanding strategy and forecasting future phenomena.
- **Core courses:** These courses will focus on a deep understanding of the analysis of big data using complex statistical models, distributed computing, machine learning, filtering, and heuristic programming. The core courses should train the student in analysing data in response to queries in a variety of content areas: organizational innovation, marketing requirements, data management and use, and the like.
- **Labs and practical workshops:** Labs and practical workshops enable students to turn theoretical learning into practical learning while actually experimenting with data mining, database management, analysis, data display, and statistical analysis.
- **Focus courses:** courses with a specific orientation toward an industrial sector, or specific business domains. For example:
  - Focus on organizational and business data analysis: financial data, Fintech, insurance, banking, investments management, risk management, capital market data, and predictive maintenance data.
  - Focus on analysing public health data: quality of life data for human communities, building quality life models, preventing future diseases and epidemics, and formulating programs in community health management.
  - Focus on the management of public institutions, hospitals, the defense establishments, government institutions, and more. This specialization enables adapting the operational model to data collected in the field, focusing on services to the citizens, resource allocation and the like.
  - Focus on smart technology: IOT technology expertise, automation, optimization, useful data analysis in smart cities, and cyber-analysis of data.
  - Focus on analysis of the education sector. Adapting personal study programs, adapting classes and methods of study, and analysis of data on a specific student, a class, or an entire school.

It is recommended to offer the courses as in the following distribution (please see Table 1):

**Table 1.** Course Summary Credits of a Type B Program.

| Course Type | Number of Courses | Practice Hours |
|---|---|---|
| Introduction | 11 | 7 |
| Fundamental | 9 | 5 |
| Core | 11 | 1 |
| Focus courses | 3 | |
| Laboratories, practical workshops and Practicum | 8 | |
| Total | 46 | Of which: 13 hours of practice |

Other recommendations:

- Allocate approximately one-third of the hours for exercises in programming, mathematics, and statistics.
- Offer at least 3-4 courses specializing in data analysis in specific market areas such as organizational and financial data analysis, smart technology, IOT and smart cities, industrial technology and medical data analysis.
- Offer at least six practical laboratory courses and practical workshops, in which practicum projects in the field of data analysis in various market areas will be carried out by the students, under the guidance of experts.

Table 2 provides the titles of the courses in a Type "B" program.

**Table 2.** List of courses recommended in an undergraduate DS program.

| **Introductory courses** |
|---|
| Programming: Python 1 |
| Introduction to Data Science |
| Discrete Mathematics |
| Introduction to Statistics – Descriptive Statistics |
| Linear Algebra |
| Programming R |
| Introduction to Social Sciences |
| Data Structures |
| Calculus |
| Programming: Python 2 |
| Introduction to Economics |

| **Fundamental courses** |
|---|
| Statistics 2: Statistical Inference |
| Data Management |

| |
|---|
| Database Administration |
| Introduction to Marketing |
| Introduction to Analytics and Business Intelligence |
| Algorithmics and complexity |
| Data Mining 1 |
| Probability Theory |
| Data Searching and Retrieval |

| **Core courses** |
|---|
| Analysis Methods for Big Data |
| Complex Statistical Models |
| Distributed Architecture and Clouds |
| Statistical Programming – Python + R |
| Machine learning and Deep Learning |
| Advanced Data Mining |
| Artificial Intelligence |
| Research Methodologies |
| Innovation in the Digital Era |
| Legal and Ethical Aspect of Information |
| Visualization |

| **Focusing courses** |
|---|
| A variety of courses related to specific industries, such as Fintech, Health, IOT, Smart Cities, Optimization and more. |

# 5 Mandatory Introductory Course

As mentioned earlier, it is imperative to introduce each student to the digital world. This is especially important for students who will not be exposed to DS programs, or even to a small selection of DS courses along the track of their studies (e.g., Humanities, Arts). Here is a suggested syllabus for an introductory course that should be offered to all the students, regardless of their major area of specialization:

**Introduction to Data Sciences.**

**Duration:** 1 semester

**Load:** 3 frontal hours per week, 2 lab and workshop hours per week

**Prerequisites:** None

**Purpose:** The purpose of the course is to provide the student with initial acquaintance with the tools the digital world provides for learning and research. This knowledge is especially important to student who do not intend to study in technology-oriented tracks.

**Topics:**

1. The data cycle
2. Review of the various tools that support each stage of the data cycle
3. Data types: numerical, alphabetical, analog, non-structured, etc.
4. Tools in more details: data retrieval, data mining, statistical tools, visualization tools, etc.
5. Introduction to programming: algorithms, programming language, procedural programming, flowcharting, problem decomposition
6. A taste of programming in Python.

## 6 To be Continued

There should be at least four tiers in DS education:

The bottom tier is the mandatory introductory course, as described above.

The third tier includes academic programs concentrating in DS, similar to types A, B, C as discussed above.

The highest tier should include graduate programs (Master's and Ph.D.) which will train researchers in the area.

The second tier is skipped in this article. It relates to DS courses or concentration subprograms that are unique to each discipline in the academic sphere. Namely, each school in an academic institution should offer a package of courses tailored to the use of DS in the specific school. For instance: DS in Social Sciences; DS in Management and Business Administration; DS in Law; DS in the Humanities; DS in Arts; DS in Medicine; DS in Engineering, DS in Economics, etc. Each package will train the students how to apply DS in their area of studies. A fast development of such packages is crucial in order to accelerate the entrance of the students and the future researchers to the digital world.

## References

Davenport, T. (2017). *The Analytics Supply Chain*. Retrieved from The Analytics Supply Chain: https://www.linkedin.com/pulse/analytics-supply-chain-tom-davenport/

IDC. (2014). *The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things*. Retrieved from The Digital Universe of

Opportunities: Rich Data and the Increasing Value of the Internet of Things: https://www.emc.com/leadership/digital-universe/2014iview/index.htm

McAfee, A., & Brynjolfsson, E. (2012). *Big Data: The Management Revolution*. Retrieved from Big Data: The Management Revolution: https://hbr.org/2012/10/big-data-the-management-revolution

Received: May 05, 2019
Reviewed: June 25, 2019
Finally Accepted: July 05, 2019