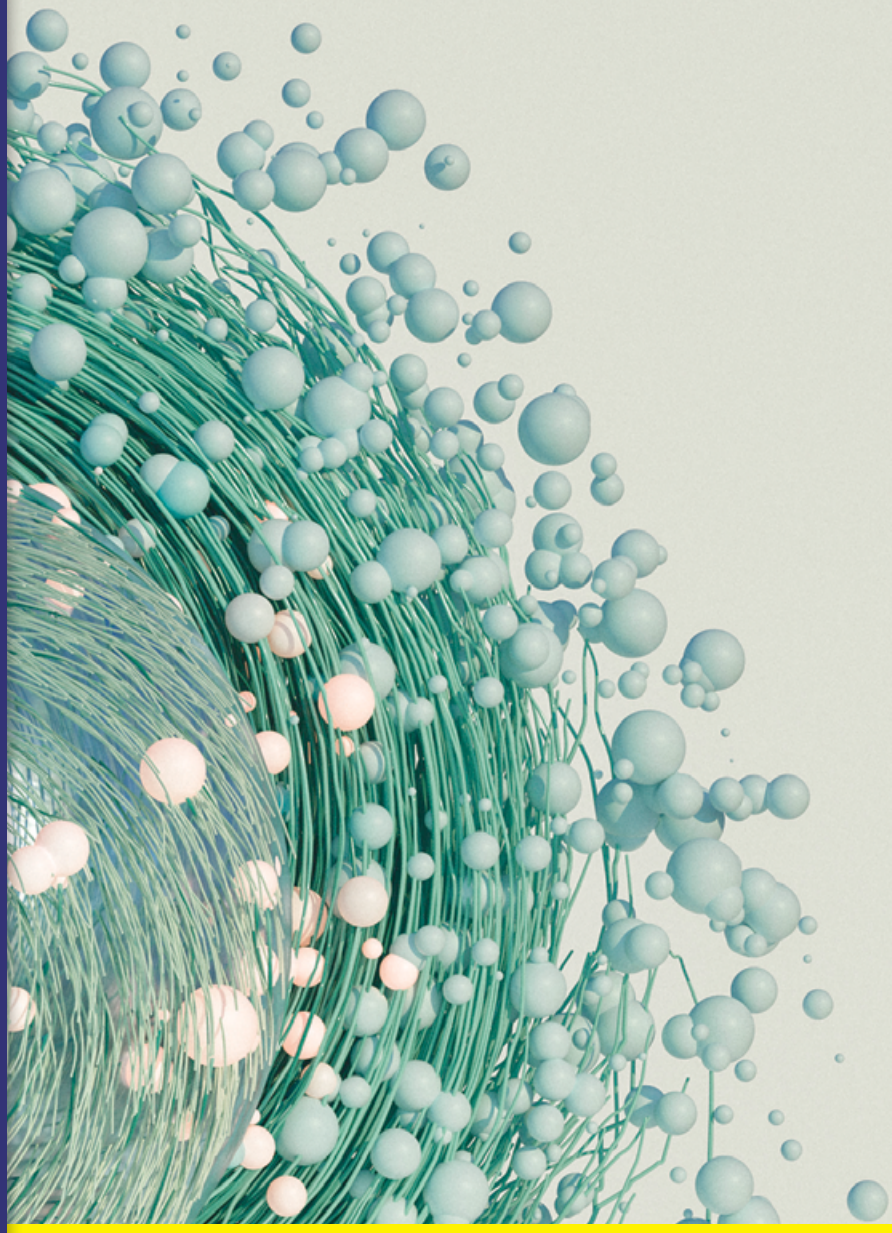
האקדמיה הלאומית הישראלית למדעים
المجمع الوطني الإسرائيلي للعلوم والآداب
THE ISRAEL ACADEMY OF SCIENCES AND HUMANITIES



Data Science Instruction for All Disciplines

Expanded Summary of the Committee Report

Science Policy in Israel

Jerusalem / December 2020

האקדמיה הלאומית הישראלית למדעים
المجمع الوطني الإسرائيلي للعلوم والآداب
THE ISRAEL ACADEMY OF SCIENCES AND HUMANITIES

Data Science Instruction for All Disciplines

Expanded Summary of the Committee Report

Jerusalem / December 2020

Table of Contents

Data Science Instruction for All Disciplines

Introduction

In December 2018, Prof. Nili Cohen, president of the Israel Academy of Sciences, appointed a committee to examine the possibility of incorporating the study of data science in every field of knowledge at universities. Committee members included Prof. Niv Ahituv (Tel Aviv University), Prof. Jonathan Ben-Dov (the University of Haifa and Tel Aviv University – the committee chair), Academy member Prof. Yoav Benjamini (Tel Aviv University), Prof. Yigal Bronner (the Hebrew University of Jerusalem), chair of the Academy's Natural Sciences Division Prof. Yadin Dudai (Weizmann Institute of Science), Prof. Daphne Raban (University of Haifa) and Prof. Roded Sharan (Tel Aviv University).

The committee met five times between January 2019 and January 2020. The meetings included consultations with external experts: Prof. Avigdor Gal (Technion – Israel Institute of Technology), Dr. Renana Kedar (Hebrew University of Jerusalem), Dr. Yossi Ben-Dov (Hebrew Reali School of Haifa), Dr. Avi Cohen (Ministry of Education), and Ms. Ronit Nehemia (Ministry of Education). The committee members, who were divided into subcommittees according to their academic fields, also consulted with other experts, as listed in Appendix A.

The importance of data science

Data science involves the collection, storage, and analysis of data. The data are used to draw conclusions and to predict, generate, and classify knowledge. Data science also entails developing tools that rely on the collected data and its analysis in reference to decision-making issues regarding human and social characteristics. The data come from a broad range of academic disciplines and commercial sources.

The data processing reflects the particular origin of the data and the application of conclusions drawn from it. Many developments in recent years have contributed to the surge in data science, including the widespread use of social and other networks, developments in the field of databases (big data), and the calculations performed on them. Machine learning in various fields, especially machine vision and language analysis, also contributed to the data science surge. In this definition we adopt the generic "data cycle" model which is constant for every process of collection, processing, and presentation of information for decision-making.

Data science is not an interdisciplinary endeavor, in the sense of constituting a particular cross-section of scientific fields; instead, it is a multidisciplinary subject. At this stage, the future of data science is unclear—whether it emerges as an entire discipline in its own right or whether it will remain dispersed among the many fields of knowledge. Given the efforts in Israel and the world during the past five years to develop undergraduate programs in the field, the trend appears to favor the former outcome. For this report, the answer to this question is not essential. Both processes will likely continue in parallel.

The importance of data science in the sciences

There is no need to elaborate on the importance of data science in experimental science. The systematic collection of data, analysis of the collected data, and drawing conclusions from it have been a central part of the scientific method ever since the invention of measuring devices such as the thermometer and pressure gauge. The posing of a scientific question, planning an experiment, documenting implementation, analysis using statistical methods, drawing conclusions, sharing findings with the scientific community, and presenting insights in text and visualization are common to all experimental sciences.

In the humanities and legal scholarship, the use of digital data has undergone a revolution. The so-called "digital humanities" field has generated an abundance of data that can be embedded in learning algorithms to produce new research insights. This data includes a large amount of visual information such as images of manuscripts, works of art, and historical newspapers. Of course, this is in addition to a tremendous amount of textual information containing a wealth of important data that has been digitized in the past decade.

The importance of data science in society and industry

Data science is increasingly important for society and the state. The state's decision-making system, which has always been based on statistical data, is updated with more data gathered from different sources. Macroeconomic data, transaction and business data, health data, and more are used for everyday activity by the state, security forces, police, and large organizations such as banks and local authorities. The interaction between the state and its citizens is becoming digital, generating data that can be used to draw conclusions and make predictions. The data are sometimes essential for defending the state, monitoring crime, and contending with epidemics. As the use of these tools becomes more prevalent, non-scientists will need better to understand their capabilities, advantages, and limitations. In industry, especially in the high-tech industry, there is a high demand for professional personnel with strong data science abilities. The current shortage is expected to grow.

Forming the committee to examine the possibility of incorporating data science instruction into all disciplines

The growing importance of data science for nearly every scientific discipline, and its importance for the fields of science, industry, security, and government, necessitates an examination of how to teach data science to undergraduates across the campus. This examination should evaluate the type and scope of knowledge appropriate to teach in the different disciplines and the students' skills to learn and achieve this.

Israeli academia is committed to imparting necessary data science skills to all of its graduates, just as it is committed to introducing them to various scientific approaches and teaching them to embrace critical thinking. In pursuit of these objectives, the Israel Academy of Sciences and Humanities formed this committee, which operates in conjunction with many active initiatives in the various universities and academic policy institutions in Israel.

The committee worked in conjunction with the Council for Higher Education's committee on undergraduate degree programs in data science and the Planning and Budgeting Committee's steering committee.

Critical points in the proposed approach for teaching data science on campus

Teaching objectives

There are different objectives for teaching data science on campus, corresponding to the students' different characteristics. For our purposes, we will divide the campus population into three groups:

A. Undergraduates in data science

This type of program is generally offered in one of the following departments (or a combination of them): computer science, statistics, industrial engineering and management, electrical engineering, management and information systems, and information science. The current report does not address this group; the Council for Higher Education's committee was formed to address this undergraduate group.

B. Students in core and related departments who are not in a program dedicated to data science

Naturally, some of these students will want to acquire professional data science capabilities in their jobs or graduate studies in data science. Besides, this group may also include students in departments where the analytical and computational requirements

for the degree – in mathematics, computation, statistics, and probability – constitute a basis for future professional work in the field. Some will be from the exact sciences (such as mathematics, physics, and various engineering fields), but some may come from the social sciences (such as economics, management, and psychology) and the life sciences and medicine. For these students, the undergraduate degree serves two purposes:

1. to enable professional employment in the business and public sectors in positions that entail analyzing information from data;
2. to enable engaging in the in-depth study for advanced degrees in data science with relatively little additional investment.

C. Undergraduate students across the rest of the campus

This group includes students in the humanities, law, management, art, and the other social sciences. For these students, the objectives are to:

1. Acquire the ability to identify a need for data and to use data for study and research purposes in their fields of study or their work after completing their degree; to acquire the ability to identify possible sources of data, to generate or retrieve the data, analyze and present it, and share it.
2. Acquire an understanding of the advantages and limitations of using data-intensive methods while maintaining a critical approach to results.
3. Develop sensitivity and a critical perspective for data collection, processing, presentation, and ethical problems concerning society and the individual that arise when analyzing data and its findings.

The content and scope of teaching: The data cycle

The conventional way of acquiring an overall picture of data science is through the "data cycle," a schematic model that characterizes any data-based research process or decision-making, including scientific research, decision-making in an organization, or a personal decision. It is a generic model applicable to any process that starts with defining a need for data and culminates with making a decision, drawing a conclusion or, making a scientific discovery. Therefore, it is valid for any discipline or problem. The data cycle is the central thread in our discussion.

Stages of the data cycle

A. **Defining the problem**: An initial definition of the problem concerning the decision, task, objective, or research in the area of knowledge for which the data was intended. The wording of the problem must explain which types of data are likely to help solve it. The problem may be to describe, predict, understand, categorize, or prove a causal connection between explanatory variables and outcomes.

B. **Collecting the relevant data**: Identifying and tracking down possible sources of data; types of required data, examining the data's suitability for the problem in terms of accessibility, contemporaneity, coverage, possible distortions in the way it was created, and the implications of these distortions on the reliability of the research; generating relevant

data by retrieving it from existing databases or through experiments, computer simulations and surveys; collecting and storing data in a manner that facilitates analysis, or creating a system that enables ongoing data retrieval from a source of information that is continually updated, ensuring the transparency of the process and the ability to reproduce it.

C. **Data integration, cleansing, filtering, and backup**: Removing erroneous data; screening data for its relevance; creating a backup; integration of data from multiple sources that have different definitions; standardization and transformation of indexes (for example, Celsius and Fahrenheit, kilometer and mile, income, number of confirmed carriers on a logarithmic scale, mortality per million people); and definition of primary and secondary outcomes according to their relevance to the problem.

D. **Processing, analyzing and mining**: Processing and analyzing the relevant data. The analysis depends on the type of data, the number of characteristics described, and the data file size. The data analysis generally proceeds from the simplest (such as comparing numerical extracts of data and examining a connection between characteristics) to advanced tools based on statistical models, machine learning, and more.

E. **Visualization and drawing conclusions from the data analysis**: Drawing conclusions from the analysis, assessing their level of uncertainty and other limitations according to the way the data were collected; presenting the results of the processing, analysis, and conclusions in images, text, and orally to decision-makers or researchers. The presentation can be a graphic or tabular visualization, dynamic or static, in an executive summary or lecture. It is imperative to make a clear presentation pertinent to the target audience's field of interest. This targeting is the realm of the man-machine interface.

F. **Implications of the data analysis for knowledge of the problem and feedback on the process**: Conclusions from analyses, despite their importance, are always incomplete and include an element of uncertainty and irrelevance. The last step includes examining the impact of the new knowledge acquired about the original problem in the knowledge field. This examination is a critical stage, whether it involves scientific conclusions, decisions, mapping directions of action, or follow-up research. Sometimes the importance of the activity is illustrated as an ongoing process, and the conclusion will be in the direction of developing unique tools for implementing data analysis processes – for example, via dynamic tables or a dashboard. This stage also includes managing the knowledge accumulated in the organization or individual.

The data cycle, as its name indicates, is a **process that repeats itself**: The last stage in the data cycle usually sheds new light on the original problem in the field of knowledge. New questions, or new facets of the original problem, arise and require defining a follow-up problem, finding data, and so on – a new round of data analysis activity. In such cases, the activity conducted in previous stages may be enhanced to improve performance in the next rounds of the cycle.

The data cycle is described graphically in the diagram below. All data scientists accept its components, but the number of stages can range from six to nine, according to the user's convenience and the importance assigned to the states or sub-stages in the cycle. A relatively simple schematic structure for teaching content is displayed here:

Diagram 1: The data cycle

The toolbox for each stage

Technological/computational tools are required to carry out each stage of the data cycle. Most of the tools are software-based (in addition to the institution's hardware and communications infrastructure). In this toolbox, we distinguish between the **minimal tools** that each student on campus should be able to use (listed under sections "1") and more advanced tools suitable for students with the appropriate background (listed in sections "2").

A. Tools for defining the problem

1. Qualitative methods for defining and formulating problems (such as brainstorming and focus groups); existing quantitative models (surveying existing knowledge and meta-analysis, for example); defining research questions or research hypotheses and formulating hypotheses in light of the data (exploratory data analysis).
2. Defining business dilemmas.

B. Tools for collecting the relevant data

1. Search engines; national and international databases and bureaus of statistics via indexes, databases of local and global organizations (for example, the UN, World Health Organization, UNESCO); technologies for transferring large data files; controlled experiments; surveys.
2. Cloud-based work methods.

C. Tools for cleansing, filtering, backing up and integrating data

1. Visualization of data; logical analysis and probability analyses for data quality control.
2. Filters, ontologies, and metadata; conversion and backup.

D. Tools for Processing, analyzing and mining data

1. Menu-based software for analyzing data without programming and using the software to describe, compare and examine a connection, based on statistical models; decision trees and random forests for predictions.
2. Programming-based software for using existing algorithms such as R and Python; using the software to apply statistical methods and for machine learning, including neural networks and deep learning; tools for processing images, language and voice, morbidity and survival data, and genetic data.

E. Tools for visualization and drawing conclusions from the data analysis

1. Visualization of results (e.g., dashboard), reports in presentations and text; assessing statistical uncertainty; human-machine connection.
2. Simulation; systems for an interactive approach and questioning.

F. Tools for implications of data analysis and feedback on the process

1. If-then algorithms; the tools defined for stage 1.
2. Specification of a decision tool for development; dashboard software.

Ethics and critical thinking in data analysis

In the process of data analysis, ethical questions arise concerning the individual and society. Some of these questions also have legal implications. Data scientists' training requires awareness of these questions and familiarity with the different ways of addressing them. The ethical questions are substantive, and they should be taken into consideration when defining the research questions and methodology.

The question of privacy should be addressed when the data includes personal information, such as medical or legal procedures. The committee calls for maintaining a balance between the societal advantages of full disclosure of data and potential individual harm. Also, the advantage that large companies gain should be weighed against society's interests and those of the individual. Creating information in large databases also raises copyright issues, which have many legal repercussions.

Researchers' ethical conduct in data analysis is as important as the computational procedure and requires full disclosure of conflicting interests, no interference in the process, or discrimination based on gender, age, or other characteristics. The transparency of the analytical process and access to the data are ethical obligations intended to enable other scientists to assess the research conclusions' replicability and validity. Many of the committee's ethical and legal issues involve critical thinking about the methodology, which is the core of quality data work. Students should acquire the ability to evaluate the mathematical procedure and identify the ethical implications and social significance of data collection and use and legal questions that may arise.

The committee noted the first academic course in Israel (as far as we know) on this subject, which was offered at the Technion during the 2019-2020 academic year. (Details are included in the Hebrew report.)

The need for different approaches and emphases in different disciplines

The topics cited in the previous section, as well as the toolbox that accompanies those subjects, should not be presented in the same way to all of the students on the academic campus. Clearly, data science knowledge is not the same for humanities students and law students or students of management and the social sciences. There may also be differences between the disciplines included within each faculty.

The committee recommends exposing every student on campus to all of the data cycle topics during their undergraduate studies. Students should study these topics to develop familiarity with data science and a functional capability in this area. This recommendation may be pursued in several ways: One way is to develop an Introduction to Data Science course; another is to enrich existing courses with research methods, statistics, and computer programming to cover all data cycle topics. It is also possible to combine these two alternatives, covering some of the material in existing courses and teaching the rest in a shorter introductory course.

The system should also offer students options to pursue in-depth study in the field– for example, allowing humanities students to take a more in-depth course in data science offered in another faculty.

Recommendations for implementation in different fields of knowledge

Introduction

The committee recommends that all students learn the data cycle model that corresponds to their field of study. This learning could take place in a dedicated Introduction to Data Science course, by adding content to existing courses, or a disciplinary course that incorporates data science. In the case where a decision is made to offer a dedicated course, the recommendation is for a modular course that can be shaped according to different requirements: starting with the fundamental and generic core, based on the data cycle model, and followed by topics appropriate to the needs of the various fields of research and study (by faculty or department).

The core component

The objectives of the core component in the Introduction to Data Science course include:

A. An understanding of the possibilities that data science offers for academic and applied research in the various fields of knowledge
B. Familiarity with basic concepts in data science
C. An ability to communicate with experts in the field of data science and utilize their assistance
D. Training graduates to work in jobs that demand familiarity with data science and collaboration with data experts

Description of the core component

Course duration:

1-2 semesters

Scope:

Two hours of lectures per week, two practice hours per week

Course topics:

1. The data cycle
2. A survey of tools that support the various stages of the data cycle
3. Types of data in the digital space: numerical, textual, structured, and unstructured
4. Basic implementation of data collection, integration of data from different sources, data analysis, algorithmic learning, and visualization of data (adapted to the field of study)
5. Critical thinking throughout the data cycle (adapted to the field of study):
   * Reliability of the data source
   * Reliability and validity of the data
   * Reliability of the algorithms
   * Adequacy and reproducibility of the research method
   * Ethics in the process of data collection, full disclosure
   * Protecting the privacy of individuals and organizations
   * Checking for distortions in the data and the findings
   * Critical reading of research based on data science
6. Sharing data (open data, data repositories) and rules for citing data sources in data-based academic work

After completing the core component, the students will take the second part of the introductory course, which will be adapted to the scientific discipline they are studying.

The full report presents several examples of syllabi in the second part of the introductory course. The disciplines presented in the report include:

* Humanities (section 3.2 in the full report)
* Social sciences and management (section 3.3)
* Law (section 3.4)
* Life sciences and medicine (section 3.5)
* Exact sciences (section 3.6)
* Engineering (section 3.7)

Recommended methods for teaching data science

Learning based on the analysis of real data

Data science, naturally, deals with data. Exposure to real data, including the difficulties in collecting, processing, and analyzing it, is a central component in teaching the subject.

Learning in teams

In the life of research, as in professional life, data analysis is the work of a team, not just one person alone. In such cases, each team member makes his or her main contribution in one of the stages, according to their area of expertise, but will also be an active participant in all of the other stages.

Online courses

There are many online courses in the English language, and plans to develop additional courses are now being promoted. These welcome efforts will succeed in conveying the more technical content of the subject.

Projects

The courses must include considerable practice exercises, independent work opportunities and culminate in a real project.

Collaboration in teaching data science

University research centers for data science

Research centers for data science were established in various research universities. Some of the universities have also launched degree programs (undergraduate or graduate) in data science. These initiatives will bring together faculty members in this field and even increase their number. The centers are interdisciplinary, spanning all fields of research at the university. Naturally, the data science centers will be partners in establishing the courses recommended in the committee's report. The success of the initiative will depend mainly on collaboration with these data science centers.

From the outset, the collaboration will be expressed by consulting with data experts when creating the syllabi. The committee also sees value in having researchers affiliated with a particular discipline be responsible for developing the course for that discipline. Each institution, faculty, or institutional center for data science should decide on the level of collaboration; they should lead the implementation of the proposed program and structure the relevant courses if they are not giving the courses themselves.

Developing and providing access to special data files for each country

The teaching of data science in a broad scope and incorporating it in numerous fields, as recommended in this report, requires sufficient resources for mining data in each of the fields. Creating these resources is a national imperative, as part of establishing other national infrastructure for science. There is usually sufficient data in the social sciences, but other fields require the formulation of policies for exposing and providing data to the public and for investing in finding the needed resources. The report discusses resources for research in

data science in the local languages, data from the social sciences, and legal data (in light of its particular sensitivity). Each of the sections surveys the current situation and proposes directions for developing resources for the future.

Synergy with academic libraries

The academic libraries serve two principal audiences: students and academic faculty. Library services include developing and providing access to physical and digital information infrastructure and support for learning and research processes. Progress in scientific research and academic instruction requires libraries to be deeply involved in academic processes to continue to provide the best service in each field of knowledge. There has been remarkably rapid growth in data science in the past decade and the more applied big data analysis field. The libraries should develop physical infrastructure for data laboratories for research students, startups, and workshops. They should create "makers" areas that include 3D printers, robots, or other equipment according to research and teaching needs. They should establish training centers for the independent study of data science that include high-quality online academic and professional training courses from Israel and the world; and collaborate with an academic faculty to build unique courses that serve local research and instruction (for example: developing research questions, a literature review, academic writing, the use of data analysis tools, textual analysis, and visualization).

Data science in pre-university education

The committee's recommendations do not pertain to the teaching of data science in high schools. The correct way to inculcate these abilities in high school pupils is not only in lessons devoted to computer science or data science but in everyday tasks of data use in each of the subjects taught in schools, from history to mathematics. Schools that have moved to learning through multidisciplinary projects can include data science skills in each of the projects. The use of data should already be taught and practiced in elementary school and certainly in middle school. Like all other fundamental skills, such as reading literacy or written and oral expression, the use of data should not be left for intensive study in high school before the matriculation exams. To genuinely incorporate data science in the education system, teachers need to be trained in this subject, including those who are already part of the system and those who are now being trained in teachers' colleges to enter the system.

The Ministry of Education formed a professional committee on data science, which is now developing a specialization in high schools. Collaboration with the ministry's committee is recommended.

Committee Members

Prof. Niv Ahituv, Tel Aviv University, National representive to CODATA

Prof. Jonathan Ben-Dov, University of Haifa and Tel Aviv University – Committee chair

Prof. Yoav Benjamini, Israel Academy member, Tel Aviv University

Prof. Yigal Bronner, Hebrew University of Jerusalem

Prof. Yadin Dudai, Chair of the Natural Sciences Division, Israel Academy; Weizmann Institute of Science

Prof. Daphne Raban, University of Haifa

Prof. Roded Sharan, Tel Aviv University

For more information, please contact: **Jo Cohen**, jo@academy.ac.il

האקדמיה הלאומית הישראלית למדעים
المجمع الوطني الإسرائيلي للعلوم والآداب
THE ISRAEL ACADEMY OF SCIENCES AND HUMANITIES

Albert Einstein Square, 43 Jabotinsky Street, Jerusalem, Israel

Tel 972-2-5676222, Fax 972-2-5666059, E-mail info@academy.ac.il

www.academy.ac.il