The Birth of a New Discipline:

Data Science Education

Jacob Mike

The Birth of a New Discipline:

Data Science Education

Research Thesis

In Partial Fulfilment of the Requirements for the Degree

of Doctor of Philosophy

Jacob Mike

Submitted to the Senate of the

Technion- Israel Institute of Technology

Haifa

Iyar 5782, May 2022

The research thesis was done Under the Supervision of Professor Orit Hazzan in the Faculty of Education in Science and Technology at the Technion – Israel Institute of Technology, Haifa.

The generous financial help of:

- Bernard M. Gordon Centre for System Engineering
- The Technion Machine Learning & Intelligent Systems (MLIS) Center
- Mr. Eyal Kaplan
- Stella and Yoel Carasso - the Beyachad Foundation

is gratefully acknowledged.

This study received the approval of:

- Technion Behavioral Sciences Research Ethics Committee: 2018-073, 2019-001, and 2019-059
- Chief Scientist, Ministry of Education: 11550.

List of Publications

Mike, K., Hazan, T. and Hazzan, O. (2020). Equalizing data science curriculum for computer science pupils. In Koli Calling '20: Proceedings of the 20th Koli Calling International Conference on Computing Education Research, Koli, Finland. ACM, New York, NY, USA.

Mike, K., Nemirovsky-Rotman, S. and Hazzan, O. (2020). Interdisciplinary education - The case of biomedical signal processing, *IEEE EDUCON 2020 Conference*, Porto, Portugal. pp 339-343.

Mike, K., Hartal, G. and Hazzan, O. (2021). Widening the shrinking pipeline: The case of data science, *IEEE EDUCON 2021 Conference*, Vienna, Austria.

Hazzan, O., & Mike, K. (2022). Teaching core principles of machine learning with a simple machine learning algorithm: The case of the KNN algorithm in a high school introduction to data science course, *ACM Inroads*, 13(1), 18-25.

Mike, K. and Hazzan, O. (2022). Interdisciplinary computer science education - The case of graduate psychology students, *IEEE EDUCON 2022 Conference*, Tunis, Tunisia.

Mike, K. and Hazzan, O. (2022, in press). Teaching machine learning: A white box approach, *Statistical Education Research Journal (SERJ*).

Mike, K., Ragonis, N., Rosenberg-Kima, R. B. and Hazzan, O. (2022, in press). Computational thinking in the era of data science, *Communications of the ACM.*

Mike, K. and Hazzan, O. (under revision). What is common to transportation and health in machine learning education: The domain neglect bias, *IEEE Transaction on Education.*

Mike, K. and Hazzan, O. (in preparation). The birth of new discipline: Data science education, *Harvard Data Science Review*.

List of Abbreviations

AI – Artificial Intelligence

CK – Content Knowledge

LG – Learning Goal

ML – Machine Learning[1]

PBL – Project Based Learning

PCK – Pedagogical Content Knowledge

PK – Pedagogical Knowledge

TPACK – Technological, Pedagogical, and Content Knowledge

[1] In titles, quotes, and references, I use the term “machine learning” and not ML

Table of Content

Abstract **1**
1. Introduction **3**
1.1. Research problem 3
1.2. Research goal 3
1.3. Research question 3
1.4. Research importance 4
1.5. Research contribution 4
2. Background **5**
2.1. Data Science 6
2.1.1. The interdisciplinary development of data science 6
2.1.2. Data science as a science 10
2.1.3. Data science as a research method 11
2.1.4. Data science as a discipline 13
2.1.5. Data science as a workflow 16
2.1.6. Data science as a profession 17
2.1.7. Three main characteristics of data science 18
2.2. Data Science Education 20
2.2.1. Data science for undergraduates 21
2.2.2. Data science in school 29
2.2.3. Machine learning education 30
3. Research Design **35**
3.1. Research problem 35
3.2. Research goal 35
3.3. Research question 35
3.4. Research method 35
3.5. Research population 39
3.6. Research tools 40
3.7. Research timeline 41
3.8. Ethics 41
3.9. Validity and reliability 42
3.10. Research limitations 43
4. Findings **45**
4.1. Case study 1 - The domain neglect bias 47
4.1.1. Case study – Biomedical signal processing projects 47
4.1.2. The domain neglect experiment 53
4.1.3. Conclusion 64
4.2. Case study 2 - The base rate neglect cognitive bias 66
4.2.1. Machine learning classifiers performance indicators 66
4.2.2. The base rate neglect cognitive bias 67
4.2.3. The base rate neglect experiment 69
4.2.4. Results 71
4.2.5. Discussion 77
4.2.6. Conclusion 78

4.3. Case study 3 - Bridging gender gaps in STEM education 79
4.3.1. Relevancy of data science for social science and digital humanities researchers... 80
4.3.2. The data science workshop for social sciences and digital humanities researchers 82
4.3.3. Workshop applicants and participants 85
4.3.4. Data collection and analysis..... 86
4.3.5. Discussion 91
4.3.6. Conclusion 93
4.4. Case study 4 - Interdisciplinary introduction to computer science course 95
4.4.1. Data science for psychological science specialization..... 96
4.4.2. The computer science for psychological science course 97
4.4.3. Course participants..... 99
4.4.4. Data collection and analysis..... 101
4.4.5. Discussion 105
4.4.6. Conclusion 106
4.5. Case study 5 - Teaching core concepts of machine learning to non-major data science students 108
4.5.1. The data science for psychology science course..... 109
4.5.2. The object-process duality theory 110
4.5.3. White box understanding of machine learning algorithms 111
4.5.4. Hands-on activity for learning the mathematics underlying machine learning 113
4.5.5. The KNN comprehension experiment 118
4.5.6. The KNN comprehension questionnaire..... 119
4.5.7. Data analysis methods..... 123
4.5.8. Results..... 125
4.5.9. Discussion 127
4.5.10. Conclusion 130
4.6. Case study 6 - The pedagogical chasm 131
4.6.1. Diffusion of Innovation theory 131
4.6.2. Crossing the Chasm theory 133
4.6.3. The data science curriculum case study from the diffusion of innovation perspective 133
4.6.4. Teachers' perspective..... 136
4.6.5. The pedagogical chasm..... 138
4.6.6. Discussion 140
4.6.7. Conclusion 141
4.7. Conclusion 142
5. Discussion..... 145
5.1. Data science education challenges and opportunities..... 146
5.2. Recommendations..... 148
5.2.1. Interdisciplinarity challenges and opportunities 148
5.2.2. Learners' diversity challenges and opportunities 151
5.2.3. Multifaced discipline challenges and opportunities..... 152
5.3. Future research directions 153
References 154

List of Tables

Table 2.1 – Theoretical background of each case study 5
Table 3.1 - Research Field 38
Table 3.2 - Research population 39
Table 3.3 - Research tools 40
Table 3.4 - Research timeline 41
Table 3.5 - Validity and reliability in different disciplines and research paradigms 43
Table 4.1 – Case studies 46
Table 4.2 – Biomedical signal processing projects - intervention program 49
Table 4.3 - Students' perception of required knowledge to meet the project goal 51
Table 4.4 - Students' perception of the research project success factors 52
Table 4.5 - The Traffic Light Classification Question 53
Table 4.6 - The Carcinoma Classification Question 54
Table 4.7 - Number of answers to the traffic light classification question and to the carcinoma classification question 55
Table 4.8 - Relationship between explanation type and decision-making type 63
Table 4.9 - The lion classification question 70
Table 4.10 - The tomato's disease classification question 70
Table 4.11 - Research population 71
Table 4.12 - Comparison of answer distribution by type of question 76
Table 4.13 - Comparison of answer distribution by population 77
Table 4.14 - Workshop Topics and Hours 84
Table 4.15 - Computer science for psychological science - Topics and number of hours 99
Table 4.16 - Data Science for Psychological Science course - Topics and number of hours 110
Table 4.17 - The classification algorithm of the KNN algorithm 114
Table 4.18 - Students' self-evaluation of prior knowledge (1=low, 5=high) (n=17) 119
Table 4.19 - Mapping the questions of the comprehension questionnaire according to the KNN process steps and object properties each question elicits 123
Table 4.20 - Coding scheme of answers to the KNN comprehension questionnaire 124
Table 4.21 - Coded answers 125
Table 4.22 - Data science for 10th grade - Topics and number of hours 134
Table 4.23 - The adoption process of the 10th grade data science curriculum 136
Table 4.24 - Data science education challenges with respect to the education fields of the disciplines generating data science 142
Table 5.1 – Data science education challenges and opportunities 147

List of Figures

Figure 2.1 - Map by John Snow showing the clusters of cholera cases in the London epidemic of 1854 (Snow, 1855) 12
Figure 2.2 - The data science Venn diagram 14
Figure 2.3 - Simplified data science workflow (Mike and Hazzan, in preparation) 16
Figure 2.4 - The data life cycle (Berman et al., 2016) 17
Figure 3.1 – Grounded theory and action research in the context of the data science education research 36
Figure 4.1 - Distribution of answers to the traffic light classification question (N=222) 56
Figure 4.2 - Distribution of explanation types to the traffic light classification question (N=197) 57
Figure 4.3 - Traffic light classification: Number of answers (X-axis) by explanation types (n=197) 58
Figure 4.4 - Distribution of answers to the carcinoma classification question (N=205) 58
Figure 4.5 - Distribution of explanations to the carcinoma classification question (N=170) 59
Figure 4.6 - Carcinoma classification question: Number of answers (X-axis) by explanation types (n=170) 60
Figure 4.7 - Explanation types: Traffic light classification question vs. carcinoma classification question (n=170) 61
Figure 4.8 - Confusion matrix 67
Figure 4.9 - The medical diagnosis confusion matrix 69
Figure 4.10 - Lion classification (phase 1): Distribution of answers (n=98). 72
Figure 4.11 - Lion classification (left) vs. tomato disease classification (right): Distribution of answers (n=153) 73
Figure 4.12 - Lion classification vs. tomato disease classification: distribution by phase and population 74
Figure 4.13 - Computer science and statistics knowledge vs. the domain knowledge dimensions of data science education 81
Figure 4.14 - Participants' pre-workshop knowledge in programming, statistics and machine learning (on scale 1-5, n=38) 83
Figure 4.15 - Workshop Participants' intrest in studying the different workshop topics (on scale 1-5, n=38) 84
Figure 4.16 - Applicants' research disciplines by gender 85
Figure 4.17 - Workshop applicants' academic rank by gender 86
Figure 4.18 - Preferred workshop format by gender 91
Figure 4.19 - Students' self-evaluation of their prior knowledge (n=36) 100
Figure 4.20 - Students' perception of the contribution of educational activities to their learning (n=25) 102
Figure 4.21 - Students' preferred course format (n=10) 103
Figure 4.22 - Students' perception of course pace (N=165 collected by 15 students in 11 questionnaires at the end of each lesson) 104
Figure 4.23 - Proposed range of possibilities for data integration in computer science education 106
Figure 4.24 - The KNN algorithm worksheet (Part 1) 116
Figure 4.25 - The KNN algorithm worksheet (Part 2) 117

Figure 4.26 - KNN comprehension questionnaire 122
Figure 4.27 - Diffusion of Innovation Timeline (Rogers, 1962) 132
Figure 5.1 – Data science education 145

Abstract

Data science is a new field of research that focuses on turning raw data into knowledge and value. The radical growth in recent years in the availability of data and of the computational resources required to process them, has led to a corresponding increase in demand for data scientists. As a result, new data science education programs are opening at a growing rate. Over the past few years, several manuscripts on data science programs were published, focusing on their curricular aspects. The pedagogy of those programs, however, has received only little attention. To partially close this gap, my PhD research focuses on the pedagogical aspects of data science.

The objectives of the research are a) to map gaps, if any, between the desired curriculum and pedagogy of data science and the existing ones, and b) to outline guidelines for the development of curricula and pedagogy for data science in a way that mitigate the challenges and leverages the opportunities of the field. From these objectives the following research question is derived: What are the challenges and opportunities of data science education?

The research integrates grounded theory research and action research methodologies. Data were collected by qualitive and quantitative tools. The researcher participated in the teaching and learning of 22 data science courses and workshops as a student, a teacher or a lecturer, in high schools, IDF, and academia. Research participants included high school pupils, soldiers, undergraduate students, computer science high school teachers, graduate students, and faculty members at research universities. Data were collected by participant observations, questionnaires, interviews, documents, and research diary.

Based on the data analysis, a theoretical framework has been developed that organizes the challenges and opportunities of data science education. Based on this framework, the research thesis has been formulated. The thesis claims that *it is essential to develop data science education as an interdisciplinary field of its own, relying on the principles of interdisciplinary education, and the integration of knowledge from the disciplines of mathematics education, computer science education, statistics education, and the application domain education.* This thesis is substantiated in this thesis by six case studies that illustrate that the challenges of data science education are *not* mitigated, and the opportunities are *not* leveraged within the disciplines of mathematics education, statistics education, computer science education, and application domains education, but can be addressed (mitigated or leveraged respectively) in the context of data science education.

In the discussion, I propose that the three main characteristics of data science education are similar to the three main characteristics of data science: interdisciplinary, diversity of learners and multifaced discipline.

The theoretical contributions of the research are a) Development of a theoretical framework of challenges and opportunities of data science education; b) Development of analysis frameworks for the comprehension of data science concepts; and c) Elaboration of the theory of interdisciplinary education, proposing a new interpretation for the term interdisciplinary pedagogy and introducing the concept open interdisciplinarity. The practical contribution of the research is the developments of new curriculum and teaching methods for data science.

1. Introduction

Data science is a new field of research that focuses on turning raw data into knowledge and value. The radical growth in recent years in the availability of data and of the computational resources required to process them, has led to a corresponding increase in demand for data scientists (Cao, 2017). As a result, new data science education programs are opening at a growing rate. Over the past few years, several manuscripts on data science programs were published, focusing on their curricular aspects, for example (Danyluk & Leidig, 2021; De Veaux et al., 2017; *EDISON Data Science Framework (EDSF)*, 2017; Heinemann et al., 2018). The pedagogy of those programs, however, has received only little attention.

Moreover, data science is relevant not only for undergraduate data science students. Methods to extract knowledge and value from data are relevant to researchers in various disciplines including exact sciences, life sciences, social sciences and digital humanities (Berman et al., 2016). Data science programs are therefore taught for learners at a diversity of backgrounds and a variety of levels. The diversity of learners' backgrounds and levels requires the development of both new curriculum and new teaching methods that can mitigate the challenges of teaching data science to such diverse learners and leverage the new opportunities in data science education.

1.1. Research problem

Over the past few years, several manuscripts on data science programs were published, intended primarily of undergraduate data science students and focusing on their curricular aspects. The pedagogy of those programs, however, has received only little attention. To partially close this gap, my PhD research focuses on the pedagogical aspects of data science.

1.2. Research goal

The objectives of the research are a) to map gaps, if any, between the desired curriculum and pedagogy of data science and the existing ones, and b) to outline guidelines for the development of curricula and pedagogy for data science in a way that mitigates the challenges and leverages the opportunities.

1.3. Research question

From these objectives the following research question is derived: What are the challenges and opportunities of data science education?

1.4. Research importance

The importance of the research stems from the increasing importance of *data* in everyday life, in industry and in research and the increasing importance of *data science* as a discipline of its own which leads to growing demands for data science educational frameworks for diverse populations of learners.

1.5. Research contribution

The theoretical contributions of the research are:

1. Development of a theoretical framework of challenges and opportunities of data science education.
2. Development of analysis frameworks for the comprehension of data science concepts.
3. Elaboration of the theory of interdisciplinary education, proposing a new interpretation for the term interdisciplinary pedagogy and introducing the concept open interdisciplinarity.

The practical contribution of the research is the developments of:

1. New data science curricula for diverse populations of learners (e.g., high school pupils, and social science and digital humanities researchers).
2. New teaching methods for data science.

2. Background

The focus of this research is the challenges and opportunities of data science education. It is not simple, however, to define what data science is and no single definition for data science exists. Accordingly, in this section, I first present data science from several perspectives (Section 2.1) and then present the current status of data science education (Section 2.2).

Note: Chapter 4 (Findings) presents six case studies. To ease the reading of this Background chapter, as well as of the case studies, this chapter presents only the general relevant background for all the case studies. Each case study includes its relevant theoretical background. Table 2.1 lists the theories presented in each case study.

Table 2.1 – Theoretical background of each case study

Case study category	Case study	Section	Relevant theory
Cognitive	The domain neglect bias	4.1	Cognitive biases (Kahneman & Tversky, 1973) Description-based decisions and experience-based decisions (Hertwig et al., 2004; Kudryavtsev & Pavlodsky, 2012). The system 1- system 2 theory (Kahneman, 2011)
	The base rate neglect bias	4.2	The base rate neglect (Kahneman and Tversky, 1973)
Social	Balancing gender gap in STEM education	4.3	Expectancy value theory (Eccles & Wigfield, 2020)
Pedagogical	Interdisciplinary CS1 course	4.4	Self-determination theory (SDT) (Deci & Ryan, 2013)
	Teaching core concepts of machine learning to non-major data science students	4.5	The process-object duality theory (Sfard, 1991)
	The pedagogical chasm	4.6	Pedagogical Content Knowledge (Shulman, 1986) Technology, Pedagogy, and Content Knowledge (Mishra & Koehler, 2006) Diffusion of innovation theory (Rogers, 1962) Crossing the chasm theory (Moore, 2002)

2.1. Data Science

Many efforts have been invested in defining what data science is; such a definition, however, has not been reached yet. One reason for the difficulty to reach a one agreed-upon definition for data science is its multifaceted nature: it can be described as a scientific research paradigm, as a research method or a research skill, as a discipline, as a workflow, and as a profession. One single definition just cannot capture this multi-faceted essence of data science.

In this section I first review from an interdisciplinary point of view the background to the development of data science (Section 2.1.1). I then present what data science is from various perspectives: data science as a science (Section 2.1.2), data science as a research method (Section 2.1.3), data science as a discipline (Section 2.1.4), data science as a workflow (Section 2.1.5) and data science as a profession (Section 2.1.6). I conclude with highlighting three main characteristics of data science: interdisciplinarity, learners' diversity and multifaced discipline (Section 2.1.7).

2.1.1. The interdisciplinary development of data science

Data science is emerging from an interdisciplinary integration of mathematics, statistics, computer science and many other application domains such as, business analytics, bioinformatics, and education. On the surface, nothing in data science is new; any data science method or tool used today can be traced back to statistics, computer science, data mining, bioinformatics, and other data intensive disciplines. The innovation, however, is expressed in the integration - the holistic view of data and methods to get knowledge and value, either financial, social, or educational.

The term *data science* emerged in statistics, computer science and application domains independently over the years (Irizarry, 2020). In this section, I review the growth of the focus on data in the disciplines forming data science: statistics (Section 2.1.1.1), computer science (Section 2.1.1.22.1.1.2), and business analytics as an example to an application domain (Section 2.1.1.3).

2.1.1.1. Data science origins in statistics

In the paper "50 years of data science", Donoho (2017) describes the role of statistics in the creation of data science. The origins of data science, Donoho claims, are rooted in Tukey's 1962 paper "The future of data analysis" (Tukey, 1962), in which Tukey presents a broader vision for the field of statistics, whose focus is data analysis rather than statistical inference. In the

following years, additional voices from the statistician community have been heard calling to shift the focus of statistics from mathematical modeling to learning from data.

According to Wikipedia, in 1985, C.F. Jeff Wu used the term Data Science for the first time as an alternative name for statistics in a lecture given to the Chinese Academy of Sciences in Beijing (Wu, 2021). Twelve years later, in 1997, he gave a lecture entitled "Statistics = Data Science?" in his appointment to the H. C. Carver Professorship at the University of Michigan (Wu, 1997). In his lecture, Wu presented his vision for the future directions of statistics, including handling large and complex data, using neural networks and data mining methods, and representation and exploration of knowledge by computational algorithms. Wu also suggested a new statistics curriculum, one of its recommendations was to create a more balanced curriculum with heavier emphasis on data collection, scientific/mathematical basis for modeling, computing for large/complex systems, and an interdisciplinary training in which undergraduate students will learn a minor in cognition and graduate students will learn 30 to 50 percent of the curriculum outside the department of statistics.

Four years later, in 2001, Cleveland (2001) published the paper “Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics” in which he called for the establishment of a new field, based on statistics that focused on data analysis. According to Cleveland, the new field should have the following six area of interest:

- *Multidisciplinary Investigations*: data analysis collaborations in a collection of subject matter areas.
- *Models and Methods for Data*: statistical models; methods of model building; methods of estimation and distribution based on probabilistic inference.
- *Computing with Data*: hardware systems; software systems; computational algorithms.
- *Pedagogy*: curriculum planning and approaches to teaching for elementary school, secondary school, college, graduate school, continuing education, and corporate training.
- *Tool Evaluation*: surveys of tools in use in practice, surveys of perceived needs for new tools, and studies of the processes for development of new tools.
- *Theory*: foundations of data science; general approaches to models and methods, computing with data, teaching, and tool evaluation; mathematical investigations of models and methods, computing with data, teaching and evaluation.

Since then, in parallel to the development of data science, it has been accepted that statistics is an inherently necessary component of data science and that statistical analysis skills are required for professional data scientists.

2.1.1.2. Data science origins in computer science

In 1966, Peter Naur proposed to the editor of the *Communications of the ACM* a new terminology and a new name for computer science, that will stress the centrality of data in computer science (Naur, 1966). The new term Naur proposed was *datalogy* defined as "*the science of the nature and use of data*" that "*might be a suitable replacement for computer science.*" (Naur, 1966).

Data mining is one branch of computer science that today is closely associated with data science, and whose roots can be found in computer science. It has been used first by statisticians, data analysts, and information systems management professionals and was adopted by the computer science community in the 1990s (Fayyad et al., 1996). While the statistician community was ambivalent regarding *data mining* as it was not considered an acceptable research strategy (Lovell, 1983), in the computer science community, the term *data mining* gained popularity. A related term to data mining is *knowledge discovery in databases* (KDD), coined by Piatetsky-Shapiro in 1989 at the first KDD workshop to emphasize that knowledge is the end product of a data-driven discovery (Fayyad et al., 1996). In the last three decades, however, data mining and knowledge discovery are used interchangeably (Piatetsky-Shapiro, 1990).

In 2000, about 10 years after the first KDD workshop, Piatetsky-Shapiro (2000) predicted that a) in parallel to the growth of computational power and data storage capabilities, data mining and KDD will become important research methods; b) data mining and KDD tools will be better integrated into commercial databases; and c) new application will emerge in e-commerce and drugs discovery.

Today, data mining is used to in variety of disciplines for a variety of applications, for example, improving the prediction of heart failure (Ishaq et al., 2021), financial fraud detection (Al-Hashedi & Magalingam, 2021), and exploration of user behavior (Su & Wu, 2021). In these applications, data mining is used to extract meaningful data from datasets using statistical, machine learning, mathematical and artificial intelligence techniques (Al-Hashedi & Magalingam, 2021).

Data mining techniques can also utilize data exploration to integrate large quantities of unrelated data, find useful correlations, and recover valuable information from the data (Su & Wu, 2021).

In parallel to the growth of data mining and KDD, other branch of computer science research faced the challenges of collecting and managing huge amounts of data, that could not be collected, stored and analyzed with the conventional storage techniques (Cox & Ellsworth, 1997). These challenges led to the development of the *Big Data* technology, which consists on the exploration of datasets with high volume, variety, velocity, and/or variability that require a scalable architecture for efficient storage, manipulation, and analysis (Chang et al., 2015). The ability to store and analyze big data is one of the key drivers of the growth of data science (Berman et al., 2018).

About 15 years ago, Wing (2006) wrote her famous CACM viewpoint on computational thinking, which is considered a cornerstone in the leverage of computational thinking as an essential thinking 21st century skill. In this viewpoint, Wing denotes the duality of code and data and writes that computational thinking is *"It is interpreting code as data and data as code".* In practice, today, machine learning consists of algorithms that can learn directly from data, and in these algorithms, indeed data becomes code.

2.1.1.3. Data science origins in application domains: The case of business analytics

In a paper titled "Data Scientist: The Sexiest Job of the 21st Century" published in the Harvard Business Review in October 2012, Davenport and Patil present the origins of data science in business analytics (Davenport & Patil, 2012). According to this paper, the term "data scientist" was first coined in industrial context by Patil and Hammerbacher in 2008, the data analytics leaders in LinkedIn and Facebook at that time. The focus of the new data scientist profession was to extract insights from big data. At that time, Davenport and Patil claimed, no university programs offered degrees in data science and, in addition, only little consensus was reached with respect to the data scientist role in the organization, how data scientists can add the most meaningful value to the organization, and how the performance of data scientists should be measured.

In the book "Data Science for Business", Provost and Fawcett (2013) describe the emergent of data science from the business need to extract knowledge and value from data:

"The past fifteen years have seen extensive investments in business infrastructure, which have improved the ability to collect data throughout the enterprise. Virtually every aspect of business is now open to data collection and often even instrumented for data collection: operations, manufacturing, supply-chain management, customer behavior, marketing campaign performance, workflow procedures, and so on. At the same time, information is now widely available on external events such as market trends, industry news, and competitors' movements. This broad availability of data has led to increasing interest in methods for extracting useful information and knowledge from data—the realm of data science." (Provost & Fawcett, 2013, p. 1)

2.1.2. Data science as a science

Empirical science has been always about data. Kepler used data about the movement of the planets collected by Tycho Brahe and proved Copernicus theory of the solar system. The first steps of evidence-based medicine started in about 1750 when James Lind, a Scottish naval surgeon, found that oranges can cure Scurvy, a disease known today to be caused by vitamin C deficiency by conducting a controlled experiment. Are Kepler and Lind the fathers of data science, looking for patterns and models in raw data? While Kepler used data that was collected manually and Lind performed his own experiment to collect the data, data science today is more than empirical science. Data science views the data itself as a natural resource and deals with methods for extracting value out of this data. While science focuses both on understanding the world and on developing tools and methods to perform research, data science focuses on understanding *data* and developing tools and methods to perform research on *data* (Skiena, 2017).

In 2005, the National Science Board of the National Science Foundation published the "Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century" report (Simberloff et al., 2005). This report presents data as a natural resource that need to be collected and managed to allow science to exhaust its full value. In this report, data scientists are defined as "*information and computer scientists, database and software engineers and programmers, disciplinary experts, curators and expert annotators, librarians, archivists, and others, who are crucial to the successful management of a digital data collection*" (p. 27).

In 2007, Jim Gray, who was the manager of Microsoft Research eScience group at that time, presented to the National Research Council a presentation entitled "eScience – A

Transformed Scientific Method" (J. Gray, 2007). In this presentation, Gray presented eScience as an integration of information technology and science. Gray presented this integration as the fourth paradigm of science. The first scientific paradigm, established thousands of years old, is the *empirical science*, in which scientist described natural phenomena. The second scientific paradigm, applied hundreds of years old, is the *theoretical paradigm*, in which scientists build models of nature. The third scientific paradigm, started only several decades ago, is the *computational paradigm*, in which scientists simulate complex phenomena using algorithms and computers. The fourth scientific paradigm, according to Gray, is *data exploration*, according to which data is captured or simulated, and analyzed by scientist to infer new scientific knowledge. In 2009, Tony Hey, Stewart Tansley, and Kristin Tolle edited a book called "The Fourth paradigm: Data-Intensive Scientific Discovery" (Hey et al., 2009). This book, devoted to Jim Gray, presents the full transcript of Gray's lecture, as well as more than 20 examples of data-centric scientific research and methods to promote scientific discovery using data in various disciplines such as earth science, environment, health and wellbeing.

In 2014, the StatSNSF (National Science Foundation Directorate for Mathematical and Physical Sciences Support for the Statistical Sciences at NSF) committee defined data science as the "science of planning for, acquisition, management, analysis of, and inference from data" (De Veaux et al., 2017; Johnstone & Roberts, 2014)

Following Gray, in the 2018 report "Envisioning the Data Science Discipline, The Undergraduate Perspective", data science is described as a shorten of the term Gray used "data-intensive science", and *"the conduct of data analysis as an empirical science, learning directly from data itself. This can take the form of collecting data followed by open-ended analysis without preconceived hypothesis (sometimes referred to as discovery or data exploration)"* (National Academies of Sciences et al., 2018).

2.1.3. Data science as a research method

Data science integrates research tools and methods from statistics and computer science that can be used in research conducted in various disciplines, e.g., research in social science and digital humanities or research of human aspects of engineering. In this section I present two data science methods out of many: exploratory data analysis (section 2.1.3.1) and machine learning as a research method (section 2.1.3.2).

2.1.3.1. Exploratory data analysis

Exploratory data analysis (EDA), whose source is in statistics, is used to find new patterns and relationship between variables by visualizing raw data and giving meaning to these visualizations (Tukey & others, 1977). In the context of research, exploration and visualization can uncover unknown patterns and relationships between variables that can reveal new knowledge. One well known example of visualization is the revealing of the source of cholera in the London epidemic in 1854 by John Snow, using a map showing the geographical spread of the cholera cases (see Figure 2.1). Another example is Goggle corelate, a tool used to find correlations between temporal search patterns of different terms to find correlations between phenomena (Mohebbi et al., 2011).

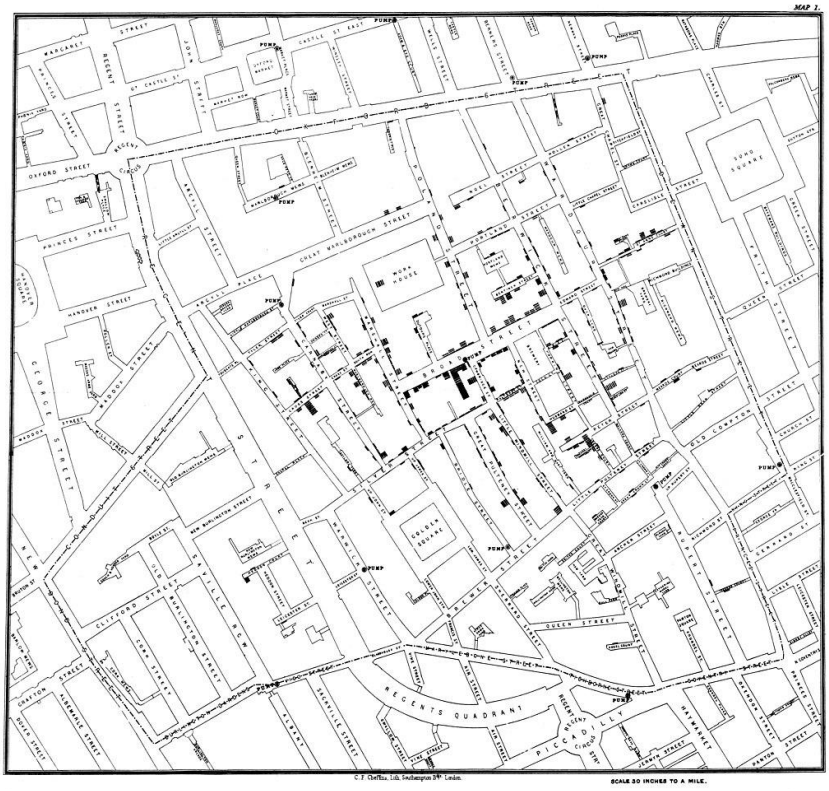


Figure 2.1 - Map by John Snow showing the clusters of cholera cases in the London epidemic of 1854 (Snow, 1855)

2.1.3.2. Machine learning as a research method

ML algorithms, whose source is in computer science, have various research application. For example, ML algorithms can find complex and nonlinear relationship between variables. In

addition, new ML techniques, such as, word embeddings, convert non-numerical complex data (such as text, images, audio, and video) to quantitative data, and thus, enable, among other things, to apply quantitative research methods on complex and non-numerical data.

Drug discovery is one example that illustrates how ML is applied as a research method. In a recent survey on ML applications for drug discovery, Vamathevan et al. (2019) writes "*Opportunities to apply ML occur in all stages of drug discovery. Examples include target validation, identification of prognostic biomarkers and analysis of digital pathology data in clinical trials. Applications have ranged in context and methodology, with some approaches yielding accurate predictions and insights"* (Vamathevan et al., 2019).

Another example is the application of ML methods in social science research. For this research, complex human generated data, such as social networks posts, are used to map social phenomena. For example, Prebor (2021) investigates different kinds of feminism in orthodox groups on social media.

Grimmer et al. (2021) review current use of ML in social science research and state that:

> *"Social scientists are now in an era of data abundance, and machine learning tools are increasingly used to extract meaning from data sets both massive and small… inclusion of machine learning in the social sciences requires us to rethink not only applications of machine learning methods but also best practices in the social sciences… [machine learning] is used to discover new concepts, measure the prevalence of those concepts, assess causal effects, and make predictions. The abundance of data and resources facilitates the move away from a deductive social science to a more sequential, interactive, and ultimately inductive approach to inference"* (Grimmer et al., 2021)

2.1.4. Data science as a discipline

Data science integrates knowledge and skills form several disciplines: computer science, mathematics, statistics, and an application domain. One way to present such relationships is by a Venn diagram which shows logical relationship between different sets. Conway (2010) was the first to propose a Venn diagram for data science as a discipline; following Conway, many other Venn diagrams for the discipline of data science were proposed (Taylor, 2016). Figure 2.2 shows the version of Venn diagram for data science proposed by Mike and Hazzan (2022).

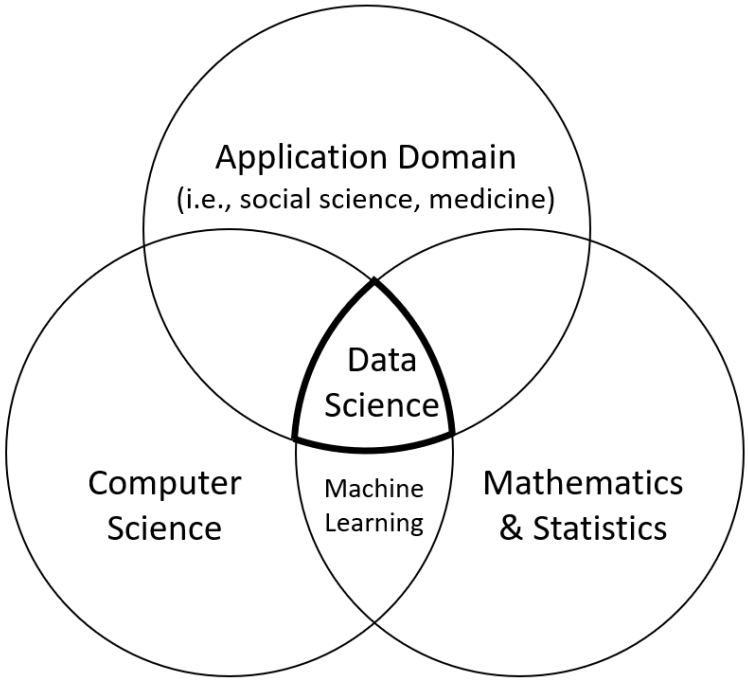


Figure 2.2 - The data science Venn diagram

Researchers recognize three levels of integration between two or more distinguished disciplines: *Multidisciplinary*, *Interdisciplinary and Transdisciplinary* (Alvargonzález, 2011). In what follows, I investigate these levels of integration from an educational perspective.

Multidisciplinary is the lowest level of integration. In multidisciplinary education, learners are expected to gain knowledge and understanding in each discipline separately. In undergraduate education, in multidisciplinary study programs, students select a collection of courses in different scientific topics (in most cases – two) without specific specialization.

Interdisciplinary represents a higher level of integration than multidisciplinary. In interdisciplinary education, after learners gained basic knowledge and understanding in each discipline separately, they are expected to understand the interconnections between the disciplines and to be able to solve problems that require applying knowledge and methods from each discipline. Gender studies, for example, combines knowledge from different disciplines such as humanities, social sciences, life sciences, forming a holistic view on the field.

Transdisciplinary is the highest level of integration. A transdisciplinary discipline is generated when the knowledge, methods and values formed in the intersection of the original

disciplines are detached from the original disciplines and create a separate new discipline on its own.

What level of integration does data science reflect today? What is the desired level of integration for data science in the future? In the 2015 NSF report Strengthening Data Science Education Through Collaboration, that summarized the NSF sponsored workshop on data science education, the debate about the status of data science is clearly phrased:

> *"Data science can either be seen as an integrated combination of contributing disciplines or a separate, still emerging discipline with its own content and expected competencies. Many participants supported the view that data science is gradually emerging as a discipline that is different from the academic disciplines that contribute to it. Data science should be freed from the contributing disciplines so that it can gain its own identity. For this to be possible, data science will need to develop its own independent theory and methods."* (Cassel & Topi, 2015, p. ii)

Today, data science is gradually moving from multi-disciplinarity to interdisciplinarity, but challenges emerged in the integration process of the different bodies of knowledge and traditions originated in the different fields and scholar cultures. The ACM Data Science Task Force final report states explicitly that:

> *"True interdisciplinary work is challenging. If each component remains independent, the relationships remain blurred and the opportunities for cross-fertilization are reduced… Early programs in Data Science will often work with a group of existing courses from the participating disciplines. That is practical and easy to bring a new program into existence... Cross references between courses, projects that call upon topics learned in other courses, and a comprehensive project to bring all the pieces together are essential to turn a mixed set of courses into a cohesive, interdisciplinary program."* (Danyluk & Leidig, 2021, p. 11)

In this spirit, it can be said that the integration of computer science, mathematics, and statistics to form a unified and holistic discipline focused on data will be described eventually as trans-disciplinarity.

To conclude this section, I note that data science is special case of interdisciplinarity, since the component of the application domain can be almost any discipline and therefore, the application domain component is practically infinite. Therefore, I propose to call this special case of interdisciplinarity *open interdisciplinarity*.

2.1.5. Data science as a workflow

Data science is commonly presented as the workflow of generating value and data-driven actions from data. This workflow starts with the collected real-world data (either available data or collected intentionally), proceeds through exploration and modeling, and continues to the publication of new knowledge and/or performance of data-driven actions in the real world (see Figure 2.3). The data science workflow is based on previous data processing workflows such as the Cross-Industry Standard Process for Data Mining (CRISP-DM) (Shearer, 2000). The data science workflow is iterative when each cycle can open a new cycle of exploration based on the finding obtained in the previous one.

The 2015 NSF report that summarized the NSF sponsored workshop on data science education, presents one definition of data science that reflects this perspective at data science as a workflow: *Data science is a process, including all aspects of gathering, cleaning, organizing, analyzing, interpreting, and visualizing the facts represented by the raw data."* (Cassel & Topi, 2015, p. iii)

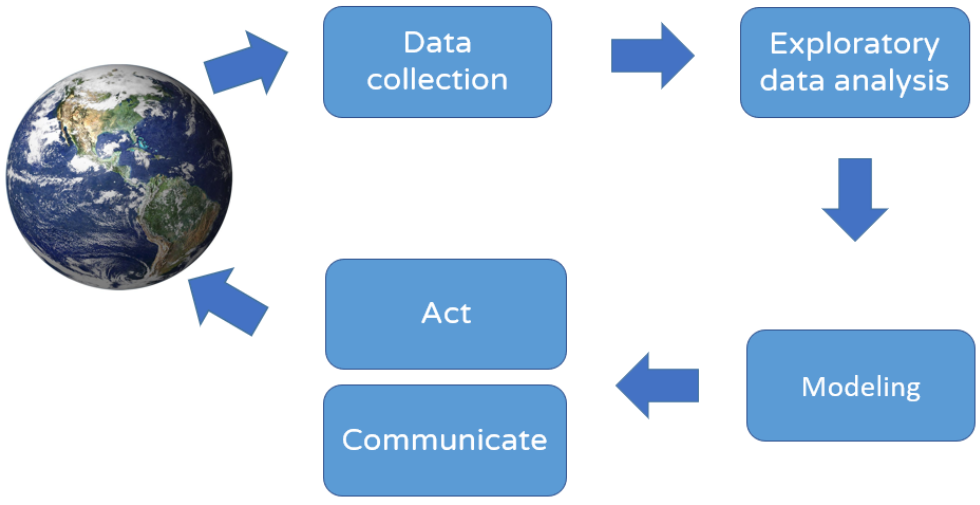


Figure 2.3 - Simplified data science workflow (Mike and Hazzan, in preparation)

The 2016 report of Data Science Working Group of the NSF Computer and Information Science and Engineering Advisory Committee (Berman, Rutenbar, Christensen, Davidson, Estrin, Franklin, Martonosi, et al., 2016) presents a broader perspective of the data science workflow, named the *data life* cycle (see Figure 2.4). The data life cycle includes not only the workflow (from collection to action and publication) but also the environmental and social aspects of data science such as regulations and ethics.

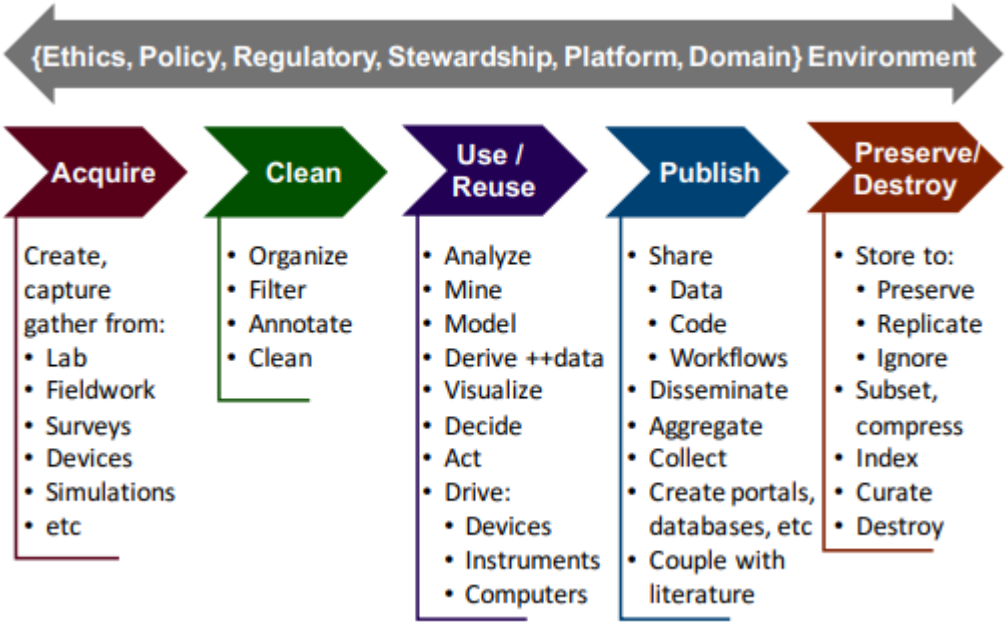


Figure 2.4 - The data life cycle (Berman et al., 2016)

2.1.6. Data science as a profession

In this section, I present the profession of data scientist, the professional who develops and applies the data science methods, belongs to the community of data scientists and adheres its norms. As described in Section 2.1.1.3, the roots of data science are in the industry, and therefore, the definition of data science as a discipline can be derived from the description of the profession of data scientist. Irizarry (2020) proposes that the term data science was coined to improve the communication between human resource recruiters in the industry and work applicants, specifying that:

> *"As the demand in employees capable of completing data-driven projects increased, the term data scientist quickly became particularly prominent because it helped*

recruiters specify the type of employee they wanted. Postgraduate degrees in the two disciplines most associated with data analysis and management, Statistics and Computer Science, did not guarantee the expertise needed to successfully complete these projects. Programming skills and experience organizing and analyzing messy, complex, and large datasets were fundamental." (Irizarry, 2020)

In their paper "Data Scientist: The Sexiest Job of the 21st Century" mentioned above, Davenport & Patil (2012) describe the profession of data scientist as *"a high-ranking professional with the training and curiosity to make discoveries in the world of big data… More than anything, what data scientists do is make discoveries while swimming in data. It's their preferred method of navigating the world around them".*

Harris et al. (2013) surveyed data scientists, working in a business environment, to learn about their perception of the profession of data science. They found 4 types of data scientists: data businessperson, data creative, data developer and data researcher, and 5 clusters of skills these professional need: statistics, programming, mathematics, machine learning and big data, and business.

The NIST Big Data Interoperability Framework report defines data scientist as *"a practitioner who has sufficient knowledge in the overlapping regimes of business needs, domain knowledge, analytical skills, and software and systems engineering to manage the end-to-end data processes in the data life cycle"* (Chang et al., 2015).

2.1.7. Three main characteristics of data science

The various perspectives of data science presented in this section: data science as a science (Section 2.1.2), data science as a research method (Section 2.1.3), data science as a discipline (Section 2.1.4), data science as a workflow (Section 2.1.5) and data science as a profession (Section 2.1.6) highlight three main characteristics of data science: *interdisciplinarity, diversity of learners and multifaced discipline*:

- *Interdisciplinarity* – data science is an interdisciplinary body of knowledge that integrates the disciplines of mathematics, statistics, computer science, and almost infinite possible application domains.

- *Diversity of learners* – due to various disciplines that can be part of data science and the essence of data science as a research method, data science is relevant to diversity of learners from various backgrounds and various educational levels.
- *Multifaced discipline* – in this section I presented multiple facets of data science: data science as a science, as a research method, as a discipline, as a workflow and as a profession. This multifaced perspectives of data science is challenging for data science education. I further elaborate on this in Chapter 5 - Discussion. It is interesting to note that similar to mathematical concepts, that can be perceived as a process or as an object (Sfard, 1991), data science can also be precepted as a process (i.e., research method or workflow) or as an object (i.e., science, a discipline, and a profession).

In this thesis I claim that the challenges and opportunities of data science education drives the need for its development as a new discipline - data science education. As I will demonstrate in Chapter 5, the three main characteristics of data science: *interdisciplinarity, diversity of learners and multifaced discipline*, also characterize the emerging field of data science education.

2.2. Data Science Education

Data science is a young field of research and its associated educational knowledge – data science education – is even younger. As for today, data science education is not yet recognized as a distinct field and is mainly discussed in the context of the education of the disciplines that compose data science, i.e., computer science education, statistics education, mathematics education and the educational fields of the applications domains, such as business education or medical education. There are voices, however, which call to integrate the relevant knowledge from these educational disciplines, and to form a coherent and integrative data science education theory, based on which data science program can be designed. The thesis of this research proposes that due to the challenges and opportunities of data science education, it is essential to develop it as an interdisciplinary field of its own, relying on the principles of interdisciplinary education. In this section, I present the current state of data scient education.

Several initiatives have been launched to integrate the disciplines composing data science to develop a data science curriculum (Cassel & Topi, 2015; Danyluk & Leidig, 2021; De Veaux et al., 2017; *EDISON: Building the Data Science Profession*, 2022; National Academies of Sciences et al., 2018). Such integration is challenging not only due to the inherent difficulty in forming an interdisciplinary program in general, but also due to the unique interdisciplinarity essence of data science, which in addition to computer science, mathematics and statistics, should include the option to add any application domain. It is therefore important to view both data science and data science education as interdisciplinary fields, while the later integrates knowledge, theories and pedagogies taken from computer science education, statistics education, mathematics education and the educational fields of application domains from different disciplines and a variety of cultures (Hazzan & Mike, 2021). In addition, data science education should also build on the theory and practice of interdisciplinary education.

Though the main wave of discussions about data science education have taken place since 2015, it is important to mention that as a respond to the market demand for data scientists, hundreds of new data science programs were launched prior to 2015 and about 200 colleges and universities offered both graduate and undergraduate degrees in data science (Cassel & Topi, 2015). Some of these programs were launched as programs in related topics such as business analytics or Knowledge Discovery is Databases and only later, evolved into a data science program. For

example, the College of Charleston in South Carolina opened an undergraduate degree in Knowledge Discovery is Databases in 2003 as a joined venture of several faculties: mathematics, statistics, computer science, and biology (Anderson et al., 2014). In 2012, when the term data science became much more popular than "Knowledge Discovery is Databases", the program changed its name to data science, to attract more students.

In Section 2.2.1, I present the short history of data science education from the perspective of the several committees that gathered in the past seven years to define the data science curriculum for undergraduates. Following the development of data science programs for undergraduates, several initiatives to fit data science for school pupils were promoted (representative initiatives are presented in Section 2.2.2). As ML is an important component of data science which is both interdisciplinary and challenging for teaching and learning, I dedicate Section 2.2.3 for the presentation of the current state of ML education.

2.2.1. Data science for undergraduates

This section reviews several of the main initiatives whose aim was to formulate guidelines for an undergraduate data science curriculum:

- Strengthening Data Science Education Through Collaboration, 2015 (Section 2.2.1.1)
- Curriculum Guidelines for Undergraduate Programs in Data Science, 2016 (Section 2.2.1.2)
- The EDISON data science framework, 2017 (Section 2.2.1.3)
- Envisioning the Data Science Discipline, 2018 (Section 2.2.1.4)
- The ACM Data Science Task Force, 2021 (Section2.2.1.5)

2.2.1.1. Strengthening Data Science Education Through Collaboration, 2015

The report "Strengthening Data Science Education Through Collaboration" (Cassel & Topi, 2015) is the product of the 2015 workshop on data science education funded by the National Science Foundation (NSF). The workshop was organized by the ACM Education Board and Council and took place in Arlington, VA.

One of the main motivations to organize this workshop was the shortage of data science professionals, which in turns, prevented organizations and societies from extracting the potential benefits from data science. While by 2015, several educational programs have already opened to address this shortage, the workshop participants felt that the quality and the future direction of

these programs, as well as their structures and practices, do not deliver a wide perspective at data science and can be improved. Therefore, the goal of this workshop was *"to start a conversation to address these concerns and develop a deeper integrated understanding of the best ways to offer data science education, ultimately leading to a better prepared workforce." (p.3)*

The workshop started with an attempt to define data science. However, since consensus regarding neither a definition for data science nor whether defining data science is possible within the timeframe of the workshop were achieved, *"[t]he workshop participants made a conscious decision not to spend a significant amount of time defining data science".* This un-agreement on the definition of data science also led to un-agreement regarding the core of data science as well as for the need of such a core, and therefore, the discussion did not attempt to define a curriculum.

Even though data science and its core were not defined in this workshop, the report details the following topics as fundamental for data scientists:

- Machine learning
- Statistical inference and modeling
- Data and database management
- Data integrity, privacy and security
- The [application] domain of activity for which data is important

With respect to the first characteristics of data science mentioned in Section 2.1.7, interdisciplinarity, the report raised several very important questions:

- How much of data science is about the theory and concepts related to handling data, independently of the application domain?
- How much should data science students learn about data related concepts (e.g., integrity, privacy, security, storage and manipulation, sharing, distribution processes, machine learning, statistical methods, algorithms, data structure, etc.) without addressing what purposes these data-related activities serve?
- Is it meaningful to have a data science degree without addressing some application domain?
- How can we balance the various disciplinary perspectives?
- Do the answers to the above questions vary pending on the context?

2.2.1.2. Curriculum Guidelines for Undergraduate Programs in Data Science, 2016

In the 2016 summer, the National Science Foundation (NSF) and the Institute for Advanced Study at Princeton University funded another workshop that focused on formulating curriculum guidelines for a data science undergraduate degree. This workshop was held at the Park City Math Institute of Princeton University at New Jersey.

The workshop product is a report named "Curriculum Guidelines for Undergraduate Programs in Data Science" (De Veaux et al., 2017). The report emphases that that the guidelines are not prescriptive, but rather, should serve to inform and enumerate the core skills that a data science major should have.

The curriculum guidelines are based on six principles: data science as science, the interdisciplinary nature of data science, data at the core, analytical (computational and statistical) thinking, mathematical foundations, and flexibility.

While data science is presented as a science, the report states that 1) data science for undergraduate students can be described as an applied field similar to engineering, with an emphasis on using data to describe the world; and 2) The data life cycle (including obtaining, wrangling, curating, managing and processing data; exploring data; defining questions; performing analyses, and communicating the results), should be the core of the data science curriculum.

It is agreed that data science requires mathematical foundations, computational thinking, and statistical thinking, and that students must be prepared to learn new techniques and methods that may not exist at the time of their studies.

The report lists the following core competence for data science graduates:

- Analytical thinking (Computational thinking and statistical thinking)
- Mathematical foundations
- Model building and assessment
- Algorithms and software foundation
- Data curation
- Knowledge transference—communication and responsibility

This list is similar but not identical to the list of core topics presented by Cassel and Topi (2015) (see Section 2.2.1.1). De Veaux et al. (2017) propose that analytical thinking is composed of both thinking skills (computational thinking and statistical thinking) and suggest that an integrated analytical thinking is the required core competence for undergraduate data science students.

Finally, the report outlies the recommended curriculum that includes the following collections of courses:

- Introduction to data science (2 courses): Introductory courses in which the students will gain first practice in the data science workflow.
- Mathematical foundations of data science (2 courses)
- Computational thinking (2 courses): Algorithms and Software Foundations, Data Curation, Databases and Data Management
- Statistical thinking (2 courses), Introduction to Statistical Models, Statistical and Machine Learning
- Course from an outside discipline, to let student gain expertise in an application domain
- Capstone course

With respect to the application domain, the report states that "*the practical real-world meanings come from interpreting the data in the context of the domain in which the data arose*" (p.3). The application domain, however, which is a fundamental topic in Cassel and Topi (2015) report, is not mentioned as a core idea by De Veaux et al., (2017). Instead, De Veaux et al., (2017) emphasis the importance of a capstone project (the 6th item in the above list) in which the students practice their data science skills in a real-world context.

Yet, the special interdisciplinary essence of data science as well of data science education was acknowledged, as the following quotes indicates:

"Nature of Data Science Data science is inherently interdisciplinary. Working with data requires the mastery of a variety of skills and concepts, including many traditionally associated with the fields of statistics, computer science, and mathematics. Data science blends much of the pedagogical content from all three disciplines, but it is neither the simple intersection, nor the superset of the three. By applying the concepts needed from each discipline in the context of data,

the curriculum can be significantly streamlined and enhanced. The integration of courses, focused on data, is a fundamental feature of an effective data science program and results in a synergistic approach to problem solving" (De Veaux et al., 2017, p. 3).

2.2.1.3. The EDISON data science framework, 2017

In parallel to the efforts invested to define data science and its curriculum in the United States, in Europe, between 2015 and 2017, the EDISON project was funded by the European Commission to define the data science profession to promote the education and training of data scientists (EDISON: Building the Data Science Profession, 2022). The EDISON project produced the EDISON Data Science Framework (EDSF) which is a collection of four documents that define the data science profession and its required competences and knowledge, and presents a model for a data science curriculum (*EDISON Data Science Framework (EDSF)*, 2017). The framework includes the following documents:

- Data Science Competences Framework (CF-DS)
- Data Science Body of Knowledge (DS-BoK)
- Data Science Model Curriculum (MC-DS)
- Data Science Professional Framework (DSPP)

The framework defines five groups of data science competence:

- Data Science Analytics (including Statistical Analysis, Machine Learning, Data Mining, Business Analytics, others)
- Data Science Engineering (including Software and Applications Engineering, Data Warehousing, Big Data Infrastructure and Tools)
- [Application] Domain Knowledge and Expertise
- Data Management and Governance (including data stewardship, curation, and preservation)
- Research Methods for research-related professions and Business Process Management for business-related professions

For each group, the framework details the competencies, body of knowledge, skills, learning goals and proposed courses.

In addition, the framework details the required professional (soft) skills with respect to three groups of professional skills:

- Data Science Professional or Attitude skills
- 21st Century workplace skills
- Data Scientist in the modern agile data driven organization

As a general competence, data science literacy is defined as commonly required competences and skills for data science-related and enabled occupations.

Similarly to the Park City Math Institute recommendations (which published the Curriculum Guidelines for Undergraduate Programs in Data Science, see Section 2.2.1.2), the EDISON project recommendations regarding the integration with application domains knowledge is not defined. For application domain related competencies, the EDISON framework exemplifies the required competencies with the domain of business analytics. In the document defining the body of knowledge (Part 2), additional component is added to represent the body of knowledge of a specific application domain knowledge, but the specific details of the integration are not elaborated in the framework:

> *"The subject domain related knowledge group (scientific or business) is recognized as essential for practical work of Data Scientist what in fact means not professional work in a specific subject domain but understanding the domain related concepts, models and organization and corresponding data analysis methods and models. These knowledge areas will be a subject for future development in tight cooperation with subject domain specialists."* (EDISON Data Science Framework (EDSF), 2022, Part 2, p. 12)

2.2.1.4. Envisioning the Data Science Discipline, 2018

In December 2016, The National Academies of Sciences, Engineering, and Medicine in United States established the Committee on Envisioning the Data Science Discipline: The Undergraduate Perspective to set forth a vision for the emerging discipline of data science at the undergraduate level (National Academies of Sciences et al., 2018). The committee was founded by the National Science Foundation (NSF).

The committee recognize the complexity of forming a data science program from two main reasons: the diversity of the backgrounds from which data science education programs were developed and the complex essence of data science itself. The committee final report states that:

> *"Current data science courses, programs, and degrees are highly variable in part because emerging educational approaches start from different institutional contexts, aim to reach students in different communities, address different challenges, and achieve different goals. This variation makes it challenging to lay out a single vision for data science education in the future that would apply to all institutions of higher learning, but it also allows data science to be customized and to reach broader populations than other similar fields have done in the past."* (National Academies of Sciences et al., 2018, p.16).

The report further describes 'data scientist' not as a single profession, but as a collection of several types of professions, each with a different focus. For example, the report mentions: data scientists who manage the platforms on which data science models are created, data scientists who focus on managing data storage solutions, experts in statistical modeling and machine learning, data visualization experts, and business analytics.

The committee defines data acumen as "*the ability to understand data, to make good judgments about and good decisions with data, and to use data analysis tools responsibly and effectively*" (p. 12). Building on the Park City Math Institute workshop (see Section 2.2.1.2), the committee defines the following concept areas for data science:

- Mathematical foundations
- Computational foundations
- Statistical foundations
- Data management and curation
- Data description and visualization
- Data modeling and assessment
- Workflow and reproducibility
- Communication and teamwork
- [Application] Domain-specific considerations
- Ethical problem solving

The committee also details several models to teach data science. Starting with introductory exposure to data science, summer programs and bootcamps, certificates in data science, minor in data science, and finally major in data science. Massive open online courses (MOOCs) can also

play a major role in proliferation of data science skills. Though the focus of this committee was undergraduates, the report also points out middle school and high school data science curriculums, as undergraduate curriculum has the potential to drive the developments of middle school and high school curricula.

Regarding the application domain, the committee state that "*Effective application of data science to a domain requires knowledge of that domain*". To achieve this goal, the committee recommends a) to grounding data science instruction in substantive contextual examples and b) to reinforce skills and capacities developed in data science courses in the context of a specific application domains. For these purposes, the committee propose several models including a track in the application domain area, specialized connector courses that link data science concepts directly to students' fields of interest, a minor in the application domain area, or a co-major or double major in an application area.

2.2.1.5. Computing Competencies for Undergraduate Data Science Curricula, 2021

The ACM data science task force (DSTF) was formed at 2017 by the ACM Educational Council. The task force mission was to "*explore a process to add to the broad, interdisciplinary conversation on data science, with an articulation of the role of computing discipline-specific contributions to this emerging field*" (Danyluk & Leidig, 2021, p. 6).

In the task force's final report, the special contribution of each of the disciplines constructing data science – computer science, statistics, and the application domain – is recognized. Effective instruction, however, it emphasizes, depends on the integration of the instruction of the different subject matters. The report state that:

> *"Each component of the Data Science environment: the domain that provides the data; statistics and mathematics for analysis, modeling, and inference; and computer science for data access, management, protection, as well as effective processing in modern computer architectures, is essential. However, a random collection of the three elements does not constitute a meaningful Data Science program. Data Science is interdisciplinary and requires the effective integration of the three components to produce meaningful results."* (Danyluk & Leidig, 2021, p. 10).

Since the origin of the task force is computing, it focused on the computer science component of the data science curriculum. In 2018, the task force conducted a survey of ACM members, representing academic institutions and industry organizations. Based on the survey data, the ACM curriculum recommendations for computer science curriculum, and feedback on early drafts of the final report, the task force ensembled the following list of core *computer science* knowledge areas (KA) for data science:

- Analysis and Presentation
- Artificial Intelligence
- Big Data Systems
- Computing and Computer Fundamentals
- Data Acquisition, Management, and Governance
- Data Mining
- Data Privacy, Security, Integrity, and Analysis for Security
- Machine Learning
- Professionalism
- Programming, Data Structures, and Algorithms
- Software Development and Maintenance

The ACM task force acknowledges also the importance of diversity and devotes one of the chapters in the final report to the theme of broadening participation in data science. Its recommendations include a) reporting student and faculty demographics as part of assessment, b) designing data science course content to support a diverse student body, and c) training faculty members in inclusive and diverse teaching pedagogy.

2.2.2. Data science in school

A significant part of this research took place in schools, either with pupils (cases 4-10 in Table 3.1) or with teachers (cases 15-16 in Table 3.1). Therefore, in this section, I elaborate on data science education on the school level.

Following the development of data science courses and data science curriculum for undergraduates, an effort invested to adjust and develop data science courses and data science curriculum for school pupils. A symposium on data science curriculum for schools took place in

Paderborn in 2017 (Biehler & Schulte, 2018), and soon later, a draft data science curriculum for school pupils was published (Heinemann et al., 2018). Following this publication, the International Data Science in Schools Project (IDSSP) published a curricular framework for introductory high school data science curriculum (Fisher et al., 2019). In Los Angles, the Mobilize Introduction to Data Science program, currently implemented in 45 schools, aims to develop computational and statistical thinking skills so that pupils can access and analyze data from a variety of traditional and non-traditional sources (Gould et al., 2018).

Other initiatives aim to expose school pupils to the power of data science in the form of extra-curricular programs. For example, Srikant and Aggarwal (2017) developed a half-day long data science tutorial for children in grades 5 through 9, which included the development of a friend predictor (i.e., a friend recommendation algorithm) and the implementation of a full cycle of data application development using a spreadsheet software. Dryer, Walia and Chattopadhyay (2018) developed a data mining workshop based on seven modules of data mining, big data, ethics, and privacy in which the pupils used RapidMiner, a graphical environment for data analysis. Haqqi et al., (2018) introduced Data Jam, a 4-month long competition aimed at introducing high school pupils to data science concepts. The program included teacher workshops, pupil workshops, homework assignments, projects, mentorship, field trips, and final posters and presentations. Bryant et al. (2019) presented a one-week long programming camp for middle school pupils with emphasis on data science, which included Python programming and data analysis tasks.

2.2.3. Machine learning education

ML is one of the effective methods for modeling huge and complex data. As a research field, it is located in the intersection of statistics and computer science (see Figure 2.2). Unlike traditional parametric modeling, which presumes an underlying statistical behavior model of the researched phenomena and aims to fit the best parameters of this model, ML models learn from experience (Goodfellow et al., 2016) and can learn complex data patterns directly from raw data. In the context of data science, ML modeling is a significant phase of the data science workflow (Berman et al., 2018). ML requires understanding of advanced mathematical and computer science knowledge and therefore, is a complex topic for teaching and learning.

While ML education is significant part of data science education, it is also discussed separately in the contexts of computer science education, statistical education, and application

domains education. In this section I elaborate on this research. it is important to note though, that while on the one hand, ML is usually described as the field created in the intersection of computer science and statistics (see Figure 2.2), on the other hand, ML education builds on the application domain knowledge, and hence should be considered as data science education. However, in this thesis, I will follow the convention related to ML education and data science education in any specific context I discuss.

Researchers identify three types of ML learners: majors, non-majors, and users (Bhatt et al., 2020; Sulmont et al., 2019b): *majors*, *non-majors*, and *users*.

Majors are students who major in disciplines such as electrical engineering, computer science, or data science and have an extensive mathematical background that enables them to learn ML algorithms with all of their mathematical details. Such learning is commonly referred to as learning ML as a white box (Georgiopoulos et al., 2009).

Non-majors are students from other disciplines, such as social sciences or life sciences, who usually lack the mathematical background required for learning ML algorithms with all their mathematical details, and so usually learn ML algorithms as a black box.

Users of ML algorithms need to understand ML as part of their professional or day to day life, i.e., physicians who use ML algorithms as diagnostic tools or business managers who uses ML algorithms for financial decision making. While the goal of ML curricula for majors and non-majors is to teach how the ML algorithms work and how to generate new ML applications, the goal of ML curricula for users is to teach the functionality and limitations of the ML algorithms, including the interpretation of their output and predictions.

In this section I present the educational challenges of ML for majors (Section 2.2.3.1), non-majors (Section 2.2.3.2).

2.2.3.1. Machine learning for data science majors and allied majors

Data science programs require extensive knowledge and skills in mathematics, statistics, and computer science (P. Anderson et al., 2014; Danyluk & Leidig, 2021; De Veaux et al., 2017; Demchenko et al., 2016). In particular, ML requires an understanding of the algorithms themselves, as well as a broader view of the algorithms in the context of the application domain. Such understanding requires knowledge about the role of ML as a component of the data science

process, handling of biases in the data, the role of training and test data, the evaluation of ML methods and models in the context of the application, ethics, and social responsibility.

ML is a complex topic for students (Sulmont et al., 2019a, 2019b). It requires an understanding of the algorithms themselves and of the underlying mathematics, as well as a broader view of the algorithms in the context of the application domain (Biehler & Schulte, 2018). Such understanding requires knowledge about the role of ML as a component of the data science process, handling biases in the data, comprehension of the role of the training and tests data, an evaluation of ML methods and models in the context of the application, as well as ethics and social responsibility (Danyluk & Leidig, 2021; De Veaux et al., 2017).

One of the main challenges of using ML algorithms is ML *explainability* and *interpretability*. Several definitions have been proposed for these terms and no consensus has been arrived at yet (Marcinkevičs & Vogt, 2020). Both terms are concerned with humans' ability to understand the predictions of ML algorithms and the reasons for these predictions (Doshi-Velez & Kim, 2017). Developing both interpretable models and explanation methods is currently one of the main efforts the ML research community is engaged in (Carvalho et al., 2019; Suresh et al., 2020) as interpretability methods are required to help users build trust in ML models and understand their capabilities. Rudin proposes that new ML techniques should be developed that are inherently interpretable (Rudin, 2019).

In the context of education, interpretability is discussed mainly in the context of users and not in the context of students. Duri and Long (Long & Magerko, 2020) define 17 competencies of AI literacy, several of which are connected to understanding the role of humans in developing the model, choosing data for training, and examining the algorithms and their biases. Suresh et al. (Suresh et al., 2020) found that people trust incorrect ML recommendations for tasks that they perform correctly most of the time, even when they have much prior knowledge about ML or are given information indicating that the system is not confident in its prediction. Several methods have been proposed to help users interpret the outcomes of ML. For example, Suresh et al. (Suresh et al., 2021) suggest accompanying the model's prediction output with additional examples, which the users are familiar with and are not necessarily taken from the given data, that generate the same prediction, to help users interpret the algorithms' output. Bhatt et al. (Bhatt et al., 2020) explored whether ML developers consider the users when they investigate their model's explainability.

They found that the majority of ML developers tend to consider explainability with respect to their own ability to debug the model rather than the end users' ability to interpret it.

2.2.3.2. Machine learning for non-majors

ML algorithms are particularly difficult for those *not* majoring in statistics, computer science, or data science. Sulmont et al. (2019a, 2019b) described non-major learners of ML as learners who lack sufficient knowledge in mathematics, statistics, and programming. For example, according to the results of their interviews with instructors of ML courses offered to such learners, Sulmont et al. (2019a) mapped the difficulties such learners encounter while learning ML according to the Structure of the Observed Learning Outcome (SOLO) taxonomy. The SOLO taxonomy, which classifies learning outcomes in terms of their complexity, consists of five levels: pre-structural, unistructural, multi-structural, relational, and extended abstract (Biggs & Collis, 2014). Sulmont et al. (2019a) found that, a) at the unistructural stage of the SOLO taxonomy, students' preconception of human thinking vs. computer processing was a barrier; b) at the relational stage, understanding decision making in ML was a barrier; and c) in the extended abstract stage, students' difficulty perceiving the limits of ML application correctly was a barrier as well. In addition, both mathematics and programming were found to be barriers for this population, and so mathematics was omitted from ML courses in almost all cases, and programming was not included in all cases. It is worthwhile to note that this is also an example to a pedagogical challenge that is not mitigated by the disciplinary pedagogy of ML, which supports my thesis that data science education should be developed as an interdisciplinary field to develop pedagogical tools for the mitigation of such challenges. Sulmont and her colleagues conclude that *"Realizing that higher SOLO learning goals are more difficult to teach is useful for informing course design, public outreach, and the design of educational tools for teaching ML."* (p. 1). In this spirit, one of the case studies presented in this thesis (Section 4.5), presents an educational tool that can support non-major data science students learning ML by mitigating the mathematical barrier.

The term "black box" understanding refers to understanding the relations between the input and the output of an algorithm without understanding how the algorithm itself works. Taking the black box approach, it is possible to learn the principles of ML even without sufficient mathematical and computational knowledge. For example, one can understand *how to use* a

logistic regression as a classifier without understanding *how the algorithm works*, i.e., without understanding the process required to find the model parameters. Biehler and Schulte (2018), however, ask *"what if machine learning is used in a data science course: Would it be appropriate to treat it as a black box…? Probably not"* (p. 9). One reason for this assertion is that ML algorithms require multiple human decisions and tuning to achieve high performance, for which white box understanding is needed (see below).

For example, one challenging task when designing an ML algorithm is hyperparameter tuning (Sulmont et al., 2019b). ML hyperparameters are parameters of algorithms that are not learned from the data, but rather are set by the human developer of the application. Such parameters are, for example, the number of layers in a neural network or K in the KNN algorithm. The hyperparameters are used to control the learning process of the algorithm. Hyperparameter tuning is essential for optimizing the performance of the learning algorithm, and at the same time, requires (at least) a process conception of the mathematical concepts used in the ML algorithm (see Section 4.5.2). Thus, whereas it is possible to teach the principles of ML without teaching the mathematical and computational knowledge required to fully understand them, individuals lacking this knowledge may find it difficult to optimize the performance of ML algorithms.

The term "white box" understanding, on the other hand, refers to understanding the details of the algorithm; that is, *how* it works, which further enables learners to proceed to object conception (see Section 4.5 – Teaching core concepts of machine learning to non-major data science learners). Using this terminology, an individual must have a white box understanding of the algorithm to understand its parameters and hyperparameter tuning and, as a result, is able improve the algorithm's performance. I note that the application of ML has two distinct phases each of which can be understood as either black box or white box: model training and model usage. However, since each phase may require different mathematical knowledge, each of them can be understood, independently, as either a black box or a white box.

3. Research Design

This section describes the research design. The research problem, goal and questions have already been presented in the introduction (Chapter 1) and are repeated here for the readership's convenience (Sections 3.1, 3.2, and 3.3 respectively). Following the research question, I present the research method (Section 3.4), research population (Section 3.5), research tools (Section 3.6), and research timeline (Section 3.7). Finally, the ethical considerations (Sections 3.8), research validity and reliability (Section 3.9), and research limitations (Section 3.10) are discussed.

3.1. Research problem

Over the past few years, several manuscripts on data science programs were published, addressing primarily undergraduate data science programs, focusing on their curricular aspects. The pedagogy of those programs, however, has received only little attention. To partially close this gap, my PhD research focuses on the pedagogical aspects of data science.

3.2. Research goal

The objectives of the research are to map gaps, if any, between the desired status of data science curriculum and pedagogy and the existing one, and to outline guidelines for the development of data science curricula and pedagogy in a way that mitigates the challenges and leverages the opportunities.

3.3. Research question

From these objectives the following research question is derived: What are the challenges and opportunities of data science education?

3.4. Research method

The research integrates *grounded theory* (Glaser & Strauss, 2017) and *action research* (Reason & Bradbury, 2001) methodologies. The need to apply both grounded theory and action research methodologies stemmed from the fact that data science education is a young field: Lack of developed theory for data science education led to the need to adopt the grounded theory approach, and lack in data science established curricula for the various populations who participated in this research led to the need to apply action research.

The relationship between grounded theory and action research in the context of my research is presented in Figure 3.1. The action research cycle starts with the *plan phase*. The basic theory that guided the planning phase is didactic transposition (Chevallard, 1989), in this case adopting data science curricula designed for data science undergraduate students to different populations such as non-major data science students and high school computer science pupils. This approach has been used previously in the context of mathematics (Kang & Kilpatrick, 1992) and computer science (Hazzan et al., 2010) and therefore it is reasonable to use it also in the context of data science. For this purpose, I participated as a student in three data science and machine learning (see Table 3.1, cases 1-3).

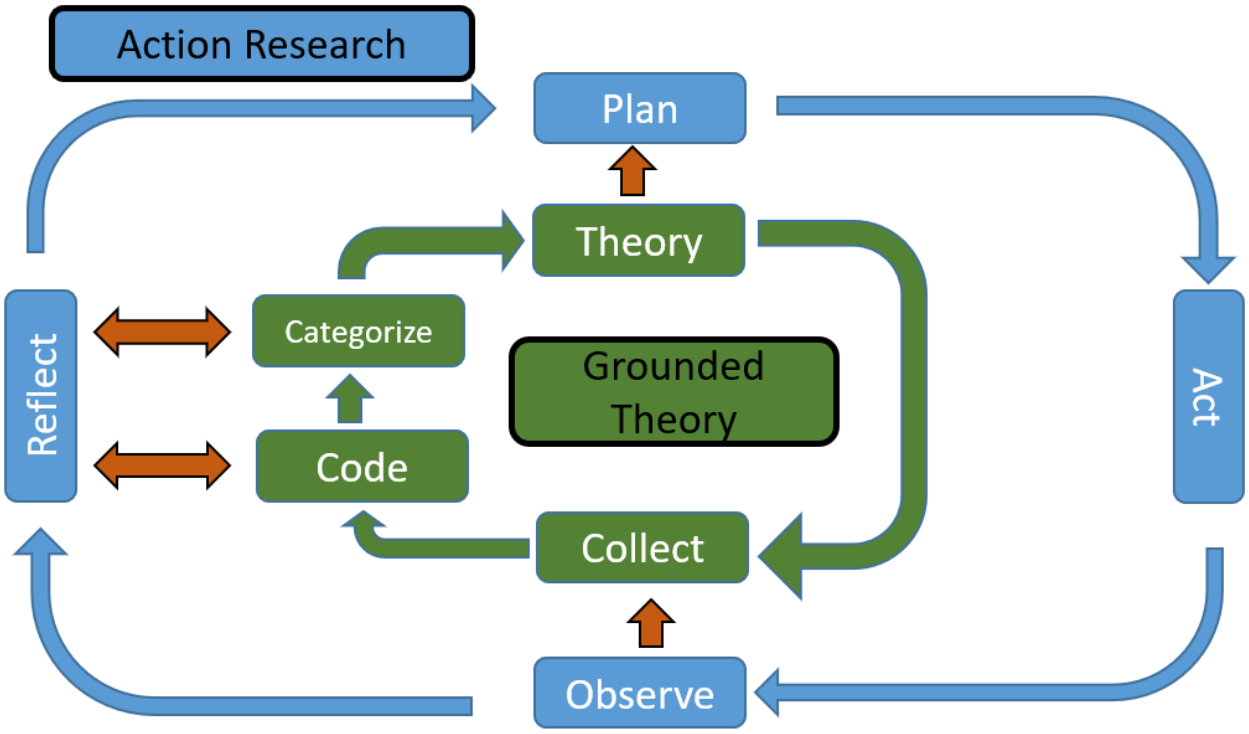


Figure 3.1 – Grounded theory and action research in the context of the data science education research

The *act phase* of the action research included the design of new data science curriculum and teaching methods for data science for different populations and the teaching of these courses (see Table 3.1, cases 4-21).

The *observe phase* of action research and the *collect phase* of grounded theory overlap. The researcher participated in the teaching of 18 data science courses and workshops as a teacher

or a lecturer in a variety of environments: high schools, IDF and academia (see Table 3.1, cases 4-21). The researcher also participated as an observer in the capstone course of fourth year undergraduate electrical engineering students (see Table 3.1, case 22). In all cases, data were collected by qualitive and quantitative tools (see Section 3.6).

The *reflect phase* of action research overlaps the *code and categorize phases* of grounded research. In details, the coding of the collected data and its categorization led to the identification and gradual understanding of the challenges and opportunities of data science education.

Finally, each research cycle is closed, and the next cycle starts by using the knowledge encapsulated in data science education challenges and opportunities framework (see Chapter 5), developed in the previous cycles, for the improvement of existing curricula and for the development of new curricula and teaching methods.

Beyond the actual application of the different research phases in several cycles, the thesis, presented in Chapter 4, was developed gradually. Specifically, based on and in parallel to this gradual understanding of the challenges and opportunities, and the applications of these findings for the improvement of existing curricula and for the development of new curricula and teaching methods in different learning environments, the following thesis was developed and formulated gradually : *It is essential to develop data science education as an interdisciplinary field of its own, relying on the principles of interdisciplinary education, and the integration of knowledge from the disciplines of mathematics education, computer science education, statistics education, and the application domain education*.

Table 3.1 - Research Field

Case	Description	Dates	Researchers' role
1	The course "Introduction to Data Analysis" at the Faculty of Industry and Management at the Technion,	Spring 2018-19	Student
2	The course machine learning at Coursera.	Spring 2019-20	Student
3	The specialization deep learning at Coursera.	Spring 2019-20	Student
4 5 6 7	Data science course for 10^{th} grade pupils	2018-19 2019-20 2020-21 2021-22	Curriculum development, Teacher
8 9 10	Deep learning course for 11^{th}-12^{th} grade pupils	2019-20 2020-21 2021-22	Curriculum development, Teacher
11 12 13	Introduction to machine learning, IDF	August 2019 September 2019 November 2020	Curriculum development, Lecturer
14	Data science for all course, undergraduates, Haifa University	2019-20	Curriculum development, Lecturer
15 16	Data science professional development for computer science teachers	June 2020 July 2021	Curriculum development, Lecturer
17 18	Data science workshop for researchers in social science and digital humanities, Technion	February 2020 September 2020	Curriculum development, Lecturer
19 20	Data science for psychological sciences, graduates, Tel Aviv University	2020-2021 2021-2022	Curriculum development, Lecturer
21	Machine learning applications in economic research, graduates, Tel Aviv University	Winter 2021	Lecturer
22	Fourth year projects for undergraduate electrical engineering in biomedical signal processing	2019-20	Observer

3.5. Research population

Research participants included high school pupils, soldiers, undergraduate students, computer science high school teachers, graduate students, and faculty members at research universities (see Table 3.2).

Table 3.2 - Research population

Educational level	**Population**	**Number of participants**
High school	Pupils	72
Professional development	Soldiers	60
Undergraduates	Data science	70
	Allied to data science (i.e., computer science, electrical engineering)	104
	Other disciplines (i.e., social science)	10
Professional development for graduates	Computer science high school preservice and in-service teachers	141
Graduates	Other disciplines (i.e., social science)	110
Staff members	Data science	10
	Other disciplines (i.e., social science)	14
Total		**591**

3.6. Research tools

Data were collected by participant observations, questionnaires, interviews, documents, and research diary (see Table 3.3). Due to the large number of research tools, a detailed description of the design of the research tools and the contribution of each research tool to the research is presented in the findings section (Section 4), together with the findings they support.

Table 3.3 - Research tools

Educational level	**Population**	**Participant observations**	**Questionnaires**	**Interviews**	**Documents**
High school	Computer Science Pupils	550	9		42 project reports
Professional development	Soldiers	120	202		
Undergraduate students	Data science		130		
	Allied to data science	39	183	5	12 project reports
	Non-majors	26	6		
Graduate	Computer science teachers	54	307	8	28 project reports
Graduate students	Non-majors	212	552		55 project reports
Faculty members	Data science and allied			7	
	Non data science or allied		15	12	
Total		**1001**	**1395**	**32**	**137**

3.7. Research timeline

The research was carried out during 4 years: October 2018 – May 2022 (See Table 3.4 - Research timeline).

Table 3.4 - Research timeline

Year	Phase	Cases	Product
Oct 2018-Aug 2019	Pilot research	1,4	Research proposal Initial results
Sep 2019-Aug 2020	Main research	2,3,5,8, 11,12,14, 15,17	Initial data analysis Data science education challenges framework – Version 1 Papers: Mike 2020; Mike, Nemirovsky-Rotman, and Hazzan 2020
Sep 2020-Aug 2021	Main research	6,9,13,16 18,19	Data analysis and validation Data science education challenges framework – Version 2 Papers: Hazzan and Mike 2021; Mike, Hartal, and Hazzan 2021; Mike, Hazan, and Hazzan 2020
Sep 2021-May 2022	Conclusion	7,10,20,21	Data science education challenges and opportunities framework Papers: Hazzan and Mike 2022; Mike et al. in press; Mike and Hazzan 2022 Final report

3.8. Ethics

This study received the approval of the Technion Behavioral Sciences Research Ethics Committee: 2018-073, 2019-001, and 2019-059 and the approval of the Chief Scientist Ministry of Education: 11550.

Since in some cases the researcher was a teacher / lecturer, all the research questionnaires were filled anonymously to avoid pressure on the participants to fill in these questionnaires. In addition, students in the courses taught by the researcher were not asked to volunteer for interviews. All information from the observations was collected without specifying the names of the participants.

3.9. Validity and reliability

The terms *validity and reliability* refer to level of accuracy and truthfulness of the data collection tools, data analysis, and findings (Brink, 1993). The main concern of validity is whether the research tool indeed measure what they are planned to measure, and whether the data analysis results and research findings indeed represent the real world from which data were collected. The main concern of reliability is whether the research tool measurements, data analysis results and research findings are persistent.

The research presented in this thesis used both qualitative and quantitative methods of data collection and analysis. Research validity is based on the diversity of research tools, the number of and the diversity of the research populations and research fields; research reliability is based on the number of research cycles carried out over four years. In some cases, findings derived from qualitative data analysis are followed by a quantitative examination of the results.

It is interesting to note that the interdisciplinarity of data science highlights the opportunity to examine the terms validity and reliability from three perspectives: statistics, computer science and the application domain (see Table 3.5).

- **Statistics:** In statistical modeling, the terms bias and variance are used to measure the accuracy of an estimator. The bias refers to how much the expected value of the estimator is close to the real value of the estimated parameter; bias is therefore may be seen as the analogous to validity. The variance is a measure of the distribution of the estimator around its expected value, and therefore, the variance can be viewed as the analogous to reliability.
- **Computer science:** In machine learning, developed mainly in the context of computer science, the terms train accuracy and test accuracy refer to the prediction accuracy of a machine learning algorithm. The training error is the prediction error on samples that the machine learning algorithm has trained on. The training error reflects the accuracy in which the model represents the training data; therefore, the training error can be viewed as the validity of the learned model. The test accuracy refers to the accuracy of the predictions of the algorithm on data samples it was not trained on; therefore, the test accuracy is therefore similar to research reliability.
- **Application domain:** Data science is relevant for a variety of application domains and research paradigms, including qualitative research, for which we examine the interpretation of validity and reliability. In qualitative research, the terms validity and reliability are sometimes

avoided and replaced by terms such as credibility, trustworthiness, truth, value, applicability, consistency, and confirmability, when referring to evaluation criteria of the scientific merit of qualitative research (Brink, 1993). As part of the current discussion, I would suggest another angle for the examination of these terms: internal and external validity (Denzin, 2017). In the context of qualitative research, internal validity refers to the extent to which research findings are a true representation of reality, and therefore is similar to validity; external validity refers to the degree to which such representations of reality can be applicable across groups, and therefore similar to reliability

Table 3.5 - Validity and reliability in different disciplines and research paradigms

Discipline	Validity	Reliability
Statistics	Bias	Variance
Machine learning	Training error	Test error
Qualitative research	Internal validity	External validity

3.10. Research limitations

In this section, I list the research limitations:

1) Despite the fact that the research scope is comprehensive in terms of populations (see Table 3.1), it still does not cover all the potential populations of data science learners, e.g., practitioners in the industry, in governmental offices and ministries, and in local authorities.
2) In most of the cases presented in Table 3.1) the researcher was involved either as a student, curriculum developer, teacher, or lecturer, a fact that might decrease the objectivity of the research.
3) Groups 4-7 (data science course for 10^{th} grade pupils) and 8-10 (deep learning course for 11^{th}-12^{th} grade pupils) of Table 3.1 learned in one school, with pupils of high learning skills

and motivation (AMAT[2]). Therefore, these groups of pupils do not represent the high school pupils' population.

4) The pupils in groups cases 4-6 are the same pupils in groups 8-10 in the following year. Therefore, the actual number of high school pupils participated in this research is smaller than 72.
5) The experiments presented in Section 4.1.1 (biomedical signal processing projects) and Section 4.5 (teaching core concepts of machine learning to non-major data science learners) were conducted only once and therefore, could not be validated through several research cycles. In additions, the experiments were conducted with a small number of participants, who may not represent their entire population.

[2] Scientific and Technological Reserve (AMAT in Hebrew) is a prestigious program of excellence of technological education in the Israeli education system. Source: Ministry of Education, https://pop.education.gov.il/scientific-technological-reserve/

4. Findings

In this section, I present my thesis, according to it, in parallel to the development of data science as a profession and a research field, *it is essential to develop data science education as an interdisciplinary field of its own, relying on the principles of interdisciplinary education, and the integration of knowledge from the disciplines of mathematics education, computer science education, statistics education, and the application domain education.* This thesis is derived from the understanding that a) data science raises new educational challenges and opportunities; b) the mitigation of these challenges and the leverage of these opportunities require an interdisciplinary approach; and c) a disciplinary perspective (which is based on the educational disciplines of the disciplines that compose data science) is not sufficient for the mitigation of these challenges and the leverage of these opportunities.

The thesis has emerged gradually, based on iterative process of data collection and analysis conducted during the entire research process (described in Section 3.4). From the many evidence that could illustrate the thesis, I chose six case studies, presented in this chapter. Each case study highlights a challenge and an opportunity of data science education, for which an interdisciplinary approach is required to mitigate the challenge and/or to leverage the opportunity. The case studies are presented by three categories - *cognitive, social*, and *pedagogical* - according to the educational theories on which they are based (see Table 2.1). Table 4.1 details the case studies presented in this chapter and the challenges and opportunities each case study demonstrates.

After the presentation of the six case studies (Sections 4.1-4.6), I conclude this section in Section 4.7 by a) Table 4.24 which summarizes the insights regarding the field of data science education derived from all six case studies and b) strengthening my thesis based on the reports of the *data science education* committees (see Section 2.2.1).

Note: The presented case studies are based on papers co-authored (either submitted or published) in the past three years. Therefore, in the description of the case studies I use a plural form (we).

Table 4.1 – Case studies

Category	Case study	Challenges	Opportunities
Cognitive	The domain neglect bias (Section 4.1)	Integration of the application domain knowledge Project Based Learning (PBL)	Teaching STEM with real-world context
	The base rate neglect bias (Section 4.2)	Lack of statistical knowledge Teachers' gap in statistical Content Knowledge (CK) Teachers' gap in statistical Pedagogical Content Knowledge (PCK)	Teaching statistical literacy
Social	Balancing gender gap in STEM education (Section 4.3)	Lack of computer science knowledge	Balancing gender gaps in STEM education
Pedagogical	Interdisciplinary CS1 course (Section 4.4)	Lack of computer science knowledge	Teaching STEM with real-world data
	Teaching core concepts of machine learning to non-major data science students (Section 4.5)	Lack of mathematical knowledge	Interdisciplinary pedagogy
	The pedagogical chasm (Section 4.6)	Teachers' gap in Content Knowledge (CK) Teachers' gap in Pedagogical Content Knowledge (PCK)	Professional development for teachers

4.1. Case study 1 - The domain neglect bias

Data science is an interdisciplinary integration of mathematics, statistics, computer science and an application domain. The data collected in this case study indicate that when performing data analysis and modeling, significant part of data science learners tends to neglect the application domain knowledge. *Cognitive biases* are phenomena of the human brain that may cause erroneous perceptions and irrational decision-making processes (Kahneman & Tversky, 1973). In the spirit of other cognitive biases, this phenomenon, observed in this research, was titled *the domain neglect bias*.

In this section, I first describe the case study of interdisciplinary biomedical signal processing projects in which the domain neglect bias was first identified (Section 4.1.1). Next, I describe the experiment that was carried out to quantify the extend of the domain neglect bias (Section 4.1.2). Finally, I discuss the implications of these results to data science education (Section 4.1.3).

4.1.1. Case study – Biomedical signal processing projects[3]

One essential element of engineering is solving practical, real-life problems using scientific knowledge and skills. Nowadays, problem solving processes of many real-life problems require an interdisciplinary approach (Alvargonzález, 2011). Project-based learning (PBL) is a pedagogical method through which students gain knowledge and skills by developing solutions to real-life problems (Hadim & Esche, 2002). PBL has been recognized as an efficient method for promoting interdisciplinary learning (Liu & Huang, 2005).

PBL has many advantages as a pedagogical tool. It promotes active learning and increases students' interest in science and technology (Duarte et al., 2012), promotes the acquisition of 21st century skills such as teamwork and interpersonal communication, and can also promote students' entrepreneurship skills (Dragoumanos et al., 2017). In addition, interdisciplinary projects can increase students' motivation and engagement in real-life problems (Yogeshwaran et al., 2019) and promote cooperation between disciplines (Ramamurthy, 2016). To successfully accomplish interdisciplinary learning in a PBL environment, project teams must include specialists in several

[3] This case study is based on Mike, K., Nemirovsky-Rotman, S. and Hazzan, O. (2020). Interdisciplinary education - The case of biomedical signal processing, *IEEE EDUCON 2020 Conference*, Porto, Portugal. pp 339-343.

disciplines. Using cross-disciplinary teams is one way to make sure a project team includes specialists in all required disciplines (Othman et al., 2017).

Working in cross- or interdisciplinary teams, however, is not always possible, and in many cases significant knowledge gaps exist within the project team in one or more of the project domains (Setiawan, 2019).

This section presents a case study of fourth-year electrical engineering students are working in pairs on biomedical signal processing projects (case 22 in Table 3.1, p. 38). Such projects are evidently interdisciplinary. The teams, however, are homogeneous and the students lack essential medical expertise required to achieve solutions applicable by physicians. Furthermore, the students acknowledge this knowledge gap only in the advanced phases of the project, and so critical phases, such as goal setting and planning, are performed without the required knowledge.

In this case study, the domain neglect bias was first identified. Following this case study, an experiment to quantify the domain neglect bias was designed and performed, and its results are presented in the next section (4.1.2).

4.1.1.1. Biomedical Signal Processing

Biomedical signal processing is an interdisciplinary field focused on algorithms and methods whose purposes are to make medical decisions based on biomedical data and to pre-process signals to improve the quality of those decisions. Biomedical signal processing is an interdisciplinary science lying in the intersection of electrical engineering, math and statistics, and biomedical science.

From an image processing perspective, the treatment of medical images is inherently different compared with that of everyday life images (such as for entertainment purposes) in the sense that the diagnostic medical information must be preserved in order to protect patients from potentially harmful situations. Considering this requirement, collaboration with radiologists is crucial when developing and testing the performance of medical data algorithms. For example, in the field of image processing, mathematical fidelity criteria are commonly used to assess processed images with respect to desired results. When dealing with medical images, however, these criteria are insufficient, and images should also undergo subjective evaluation tests by specialist

physicians to assess their quality from the medical perspective. Thus, *medical knowledge and expertise is vital to achieve satisfactory performance of imaging systems*.

4.1.1.2. Biomedical signal processing projects in the Faculty of Electrical Engineering

Dr. Shira Nemirovsky-Rotman was a PhD student in the Faulty of Electrical Engineering in the Technion, and an instructor in the Vision and Image Sciences Laboratory (VISL) at the Andrew and Erna Viterbi Faculty of Electrical Engineering at the Technion – Israel Institute of Technology. I met Shira in the summer of 2019, as she felt that the students she mentored lack knowledge in biomedicine domain and looked for help developing a method to teach this content.

We decided to develop an interdisciplinary intervention program in biomedical signal processing for undergraduate electrical engineering students. The program objectives are: (a) to teach students the basic knowledge of the biomedical data domain; (b) to convey to students the importance of understanding the medical domain while designing and developing image processing algorithms which will be applicable to physicians; (c) to refer students to relevant resources for self-learning material about biomedical signals. The intervention program consists of regular, short key lectures, as well as group meetings with radiologists and scientists from the industry who provide valuable medical knowledge and expertise. The intervention program includes six meetings scheduled to take place during the semester. Table 4.2 presents the program schedule and the topics covered in each meeting. In addition, individual project follow-up meetings are held with the project supervisor.

Table 4.2 – Biomedical signal processing projects - intervention program

Week	Content
1	Introductory tutorial on ultrasound imaging and presentation of project goals
3	First meeting with radiologists to present project results and ask for medicine-related feedback
6	Presentation of self-learning materials on biomedical signals
8	Tutorial on objective fidelity criteria of medical image processing and second meeting with radiologists
10	Third meeting with radiologists
13	Meeting with representatives from the biomedical signal processing industry to receive feedback on the relevance and applicability of the current research results

4.1.1.3. Data analysis

To evaluate the intervention program, we collected data from questionnaires answered by both student who participated the intervention program and students from previous years who did not participate the intervention program. Data from students who participated the intervention program was collected using bi-weekly anonymous questionnaires, and from students from previous years who did not participate the intervention program data was collected by interviews and a single questionnaire. In the following sections, I compare the perception of the importance attributed to the application domain knowledge by students who participate the intervention program and by students who did not participate the intervention program.

Perceptions of students who did not participate the intervention program

We asked students from previous semesters what they think their knowledge gaps were when they were working on their projects. The students told us that they lacked knowledge on the way physicians examine the images and what is considered a good image. Indeed, they discovered this knowledge gap only after they already had a running algorithm. Due to this knowledge gap, they used objective image fidelity evaluation criteria, which are mathematical metrics that do not represent the medical criteria well, nor what physicians are really looking for.

Another student told us that, retrospectively, it was difficult to understand the project goals without the medical knowledge. After running several signal processing algorithms, there was certainly an output, but understanding whether or not this product fulfilled the project goals was impossible without understanding the image from the point of view of medical decision making.

Students from previous semesters thought that this kind of intervention course would be of great benefit to current students. According to these students, in addition to providing a platform for the transfer of medical knowledge, the meetings should be interactive and used to discuss processed images with the physicians. This way, the physicians could serve as medical specialists in cross-disciplinary teams.

Comparison of students' perceptions: Participated in the intervention program vs. did not participate in the intervention program

At the end of the intervention program, we asked both students who participated the intervention program and students from previous years who did not participate the intervention

program to indicate on a scale of 1 to 5 how much knowledge they required to learn to meet the project goals. Students' ratings are presented in Table 4.3. It can be seen that a) the students' perception of the required knowledge in statistics and in biomed is lower than the required knowledge in electrical engineering, and b) the perceived required knowledge in similar for both groups. Nevertheless, while the perception of the two groups with respect to the required knowledge is similar, this perception is explained differently by the students in the two groups (the ones who participated the intervention program and the ones who did not). The students who participated the intervention program perceived the required knowledge in the biomed domain as relatively low due to their understanding that the radiologists are supporting them in this aspect. For example, one student wrote that *"we had continuous guidance from the radiologist"*. On the other hand, the students who did not participate in the intervention program perceived the required knowledge in the biomed domain as relatively low due to their perception of this knowledge as not crucial for their research. For example, one student explained that *"it is required to have* [only] *basic knowledge in medical domain*".

Table 4.3 - Students' perception of required knowledge to meet the project goal

Discipline	**With intervention (n=8, scale 1-5)**	**Without intervention (n=4, scale 1-5)**
Electrical engineering	4.25	4.50
Statistics	2.25	3.00
Biomed	2.50	2.75

We further compared the students' perception of the success factors of the research project (see Table 4.4). While students answers to self-learning as success factor are similar between students who participated in the intervention program and students who did not in the intervention program, a large difference is expressed between the two groups of students with respect to the perceived contribution to the project success of the cooperation with the radiologists. Indeed, the students who participated in the intervention program actually met a radiologist during their project (and accordingly rated this factor as 4.62), while the students who did not participate in the intervention program, worked on the project without any cooperation with radiologists (and accordingly rated this factor as 2.5). The students who participated in the intervention program explained that *"cooperation with the radiologist is necessary"* while the students who did not

participated in the intervention program explained that *"feedback from radiologist is not as important as self-learning".*

Table 4.4 - Students' perception of the research project success factors

Success factor	**With intervention (n=8, scale 1-5)**	**Without intervention (n=4, scale 1-5)**
Self-learning on biomedical diagnosis	3.87	3
Self-learning on image processing	4.12	4.5
Cooperation with radiologists	4.62	2.5

4.1.1.4. Discussion

It is evident that while the students *do* understand the connection between their project and the medical domain, they *do not* necessarily understand, at the beginning of the semester, the importance of having medical knowledge for their project success. However, even at an early stage of the project, the students attribute greater importance to the medical aspect of the project than they do at the very beginning of the semester. They notice the variety of challenges they expect to encounter, including code writing, image processing, and what the images should include, i.e., what is important and what can be ignored. The students acknowledge the importance of cooperating with medical specialists and specify questions they intend to ask the radiologists.

While the previous year's students had no sources of information regarding medical knowledge, the students who participated in the intervention program understood that they need access to such knowledge and accumulate questions for the medical specialists. While students in previous years could use only mathematical fidelity criteria, the students who participated in the intervention program had medical specialists whom they can consult with and learn from.

The findings regarding the electrical engineering students' perceptions with respect to the biomedical domain knowledge led us to hypothesize that students' neglect of the application domain knowledge might be wide phenomenon, common to many learners. We therefore designed an experiment on a larger scale, described in the next section.

4.1.2. The domain neglect experiment

To quantify the domain neglect phenomena, an experiment was designed. As part of the experiment, the participants were asked to compare the performance of two machine learning algorithms in two settings. The participants' answers to these questions allows to infer their awareness to the underlying application domain. In this section I first describe the domain neglect questionnaire (Section 4.1.2.1). Next, I present the data and its analysis (Section 4.1.2.2). Finally, I further discuss the expression of the new defined domain neglect cognitive bias (Section 4.1.2.3)

4.1.2.1. The domain neglect questionnaire

Two questions were developed for this experiment: the traffic light classification question (see Table 4.5) and the carcinoma classification question (see Table 4.6). In these questions, participants are asked to compare the performance of ML classifiers in two different domains: transportation and medical diagnosis.

Table 4.5 - The Traffic Light Classification Question

Researchers developed two machine learning algorithms to classify the red light and the green light of a traffic light. Machine A is always correct when the light is red but is wrong with a probability of 0.001 when the light is green. Machine B is always correct when the light is green, but is wrong with a probability of 0.0001 when the light is red. Which of these machines is better in your opinion? (a) Machine A (b) Machine B (c) Both are the same (d) It is not possible to answer the question since data is missing (e) I do not know Can you explain your choice? _____

Table 4.6 - The Carcinoma Classification Question

Researchers developed two machine learning algorithms to classify two types of carcinomas: SCC and BCC. Machine A is always correct when the carcinoma is of type SCC but is wrong with a probability of 0.001 when the carcinoma is of type BCC. Machine B is always correct when the carcinoma is of type BCC but is wrong with a probability of 0.001 when the carcinoma is of type SCC. Which of these machines is better in your opinion? (a) Machine A (b) Machine B (c) Both are the same (d) It is not possible to answer the question since data is missing (e) I do not know Can you explain your choice? _____

In the traffic light classification question, the participants are requested to compare two ML algorithms that were trained to detect the light in a traffic light. The two machines are different in their performance: while Machine A may classify a green light as red with probability 0.001, Machine B may classify a red light as green with probability 0.0001. Even though Machine B makes only a tenth of the mistakes Machine A makes, the severity of the mistakes Machine B makes, detecting a red light to be green, is much higher than the severity of the mistakes Machine A makes, detecting a green light to be red. The students were asked to choose from five optional answers. Beside the three comparison options: (a) Machine A's performance is better; (b) Machine B's performance is better; and (c) both machines perform equally well, two more options were added: (d) It is not possible to answer the question since data is missing and (e) I do not know the answer. In addition to the multiple-choice question, the students were asked to explain their answer to the multiple-choice question in an open question.

In the carcinoma classification question (see Table 4.6), the participants are requested to compare two ML algorithms that were trained to detect two types of carcinomas: basal cell carcinoma (BCC) and squamous cell carcinoma (SCC). The two machines differ in their performance yet have the same misclassification rate: Machine A may classify BSS as SCC with a probability of 0.001, and Machine B may classify SCC as BCC with a probability of 0.001 as

well. Even though both machines may err with the same probability, 0.001, since SCC is a more severe disease than BSS, the consequences of errors in this case are worse. As in the traffic light question, students were given five optional answers and were asked to explain their multiple-choice question in an open question.

The questions were presented to the research participants using questionnaires. The traffic light classification question was answered by 226 participants. Detailed divide of the participants by population is presented in Table 4.7.

Table 4.7 - Number of answers to the traffic light classification question and to the carcinoma classification question

Population	**Number of answers to the traffic light classification question**	**Number of answers to the carcinoma classification question**
Data science, computer science and electrical engineering undergraduate students	104	101
Computer science teachers	59	52
Social science and digital humanities graduate students and researchers	59	52
Total	**222**	**205**

4.1.2.2. Domain neglect experiment results

The distribution of answers to the traffic light classification question for all populations is presented in Figure 4.1. Answers were classified into three groups: 46% thought that Machine A (which errs on green light) performs better than Machine B; 30% thought Machine B (which errs on the red light) performs better (26%) or the same as Machine A (4%) as Machine A; and 24% indicated that they cannot answer the question either because data is missing (20%) or they do not know (4%).

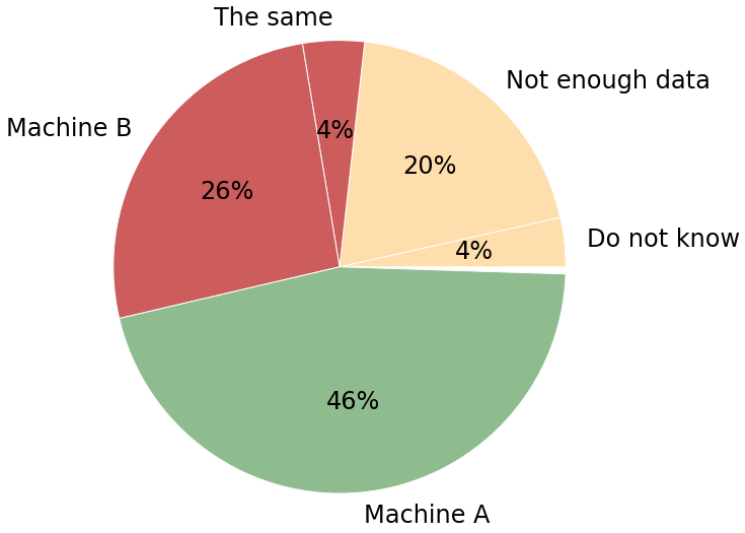


Figure 4.1 - Distribution of answers to the traffic light classification question (N=222)

Analysis of the answers to the open question, in which the students were requested to explain their choice, revealed three types of answers:

- **Comparison based on domain considerations.** For example, one student explained that "it is more important not to make mistakes on a red light to reduce car accidents".
- **Comparison based on mathematical considerations.** For example, one student explained that "Machine B's error probability is lower".
- **Comparison** based on both domain considerations and mathematical considerations, **debating which consideration is more important**. For example, one student explained that "I guess the mistake in red classification is more critical in the context of accidents. So even though B is much less wrong, its mistakes are costlier. It remains to be decided whether the mistake in red will indeed "happen" 10 times more than the mistake in green - to determine which one is a better classifier. And this is information that does not appear in the question (and in any case may be moral)."

I classified participants' explanations based on these categories (see Figure 4.2). While most participants (62%) took the application domain knowledge into consideration, more than one third of the participants (38%) considered only the mathematical consideration (30%) or debated which of the two considerations is more important (8%).

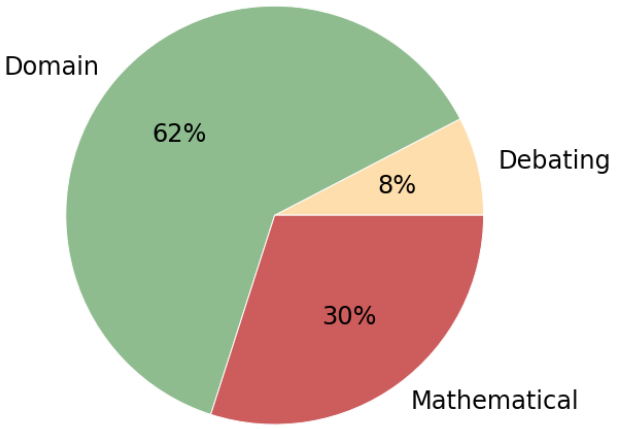


Figure 4.2 - Distribution of explanation types to the traffic light classification question (N=197)

The relationship between the answers to the traffic light classification question and the types of explanations the students used is presented in Figure 4.3. While all participants who indicated Machine A's performance as being better justified their answers with considerations taken from the application domain knowledge, most participants who selected Machine B as the better machine justified their selection with mathematical considerations. It is interesting to note that participants who answered that data is missing used considerations from all three categories: some participants claimed that data explaining that the severity of the mistakes was not given, hence referring to missing application domain knowledge; others explained that the base rate of the events in the dataset was not given, hence referring to mathematical considerations; and the rest of the participants who answered that data is missing debated between the two types of considerations. Very few students choose the answers "Do not know" or "The same". None of them explained his/her answer with a domain consideration.

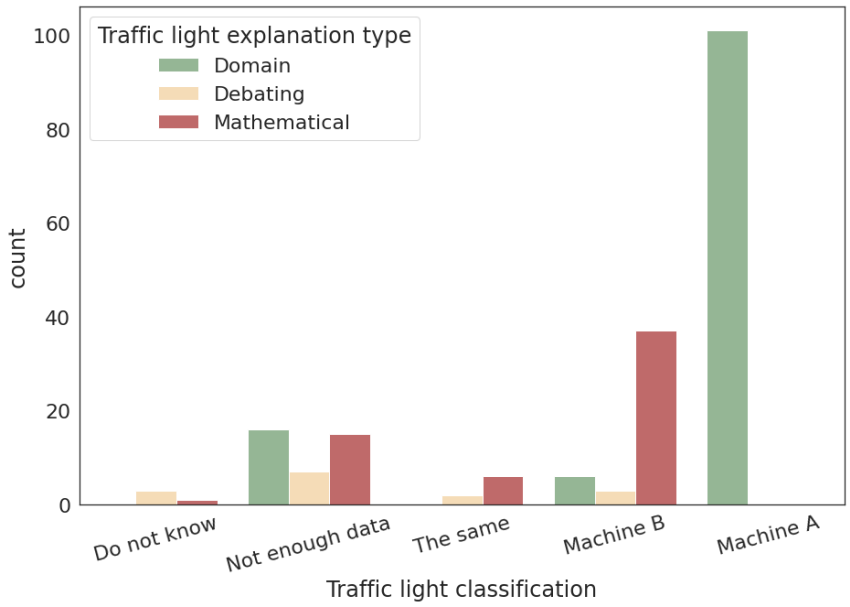


Figure 4.3 - Traffic light classification: Number of answers (X-axis) by explanation types (n=197)

A total of 205 participants answered the carcinoma classification question (see Table 4.6). Most of them (58%) did not select one of the classifiers, but rather stated that there was not enough data to answer (41%) or that they do not know the answer (17%). More than one third of the participants (37%), however, indicated that the two machines performed equally well, ignoring the (possible) real-world difference in severity between the two types of carcinoma. Only a few students choose either Machine A or Machine B.

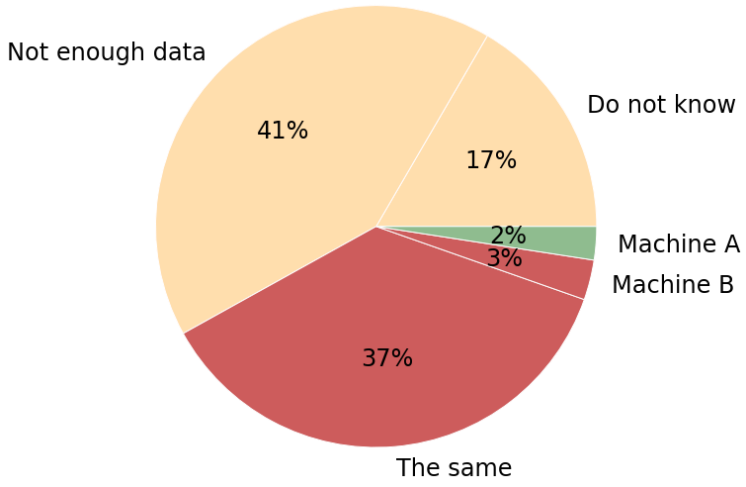


Figure 4.4 - Distribution of answers to the carcinoma classification question (N=205)

As with the traffic light classification question, the participants were requested to explain their answer to the carcinoma classification question in an additional open question. The same three types of explanations that were given for the traffic light classification question answers were given for the carcinoma classification question answers: (a) Comparison based on domain considerations; (b) comparison based on mathematical considerations; and (c) comparison based on both considerations, without deciding which consideration is more important. While some participants explained that they do not know the answer since they do not know about the severity of the carcinomas, hence they considered the application domain in the real world, other participants explained that they do not know the base rate of the different carcinomas, hence considered the mathematical properties of the problem. In general, however, relative to the traffic light question, fewer participants based their explanations of their answer to the carcinoma classification question on the application domain (48% vs. 62%) and more students based their explanations on mathematical considerations (45% vs 30%). It is reasonable to assume that these differences can be explained by the participants' lack of familiarity with the medical diagnosis domain.

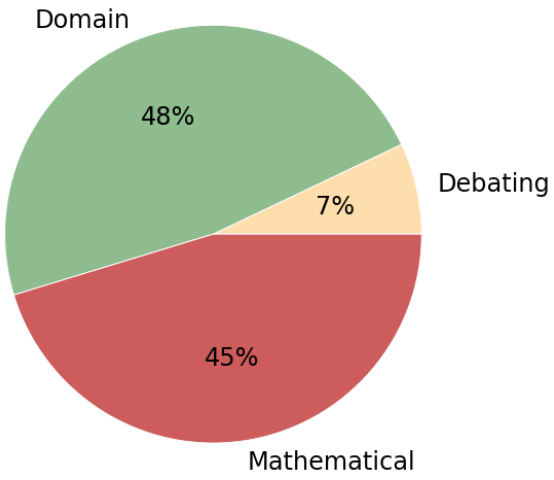


Figure 4.5 - Distribution of explanations to the carcinoma classification question (N=170)

Again, the relationship between the answers to the carcinoma classification question and the types of explanations the participants used was analyzed (see Figure 4.6). While all participants who indicated that they do not know the answer explained that they do not know the severity of the different cancers, hence used domain considerations, all participants who thought that the two machines performance is the same explain it using mathematical considerations. Similar to the

explanations for the traffic light classification question, participants who answered that data is missing were using considerations from all three categories. With respect to the traffic light classification question, significantly less students debated regarding the correct answer.

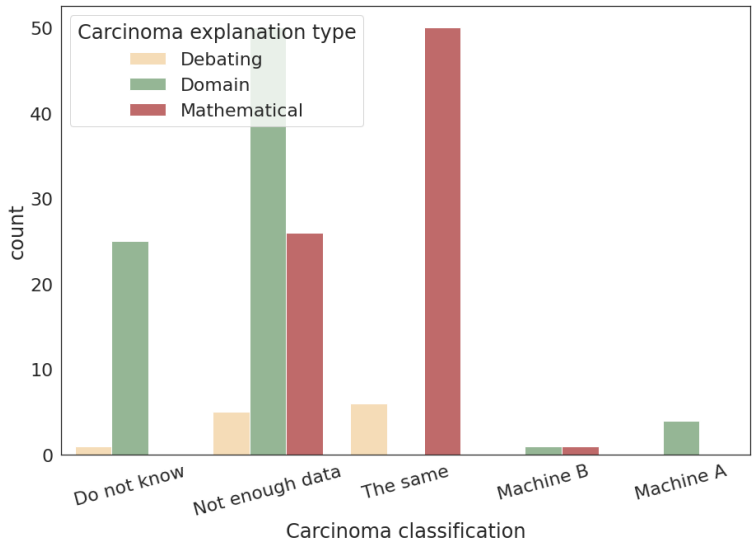


Figure 4.6 - Carcinoma classification question: Number of answers (X-axis) by explanation types (n=170)

Finally, it is interesting to compare the types of explanations participants gave for the traffic light classification with those given for the carcinoma classification (see Figure 4.7, the background color is according to the color of the column and the color of the frame is according to the color of the row). Most of the participants who used mathematical explanations for the traffic light classification question also used mathematical explanations for the carcinoma classification question (41 out of 47). The pattern is different for participants who used domain explanations or debated when explaining the traffic light classification question. While about two thirds of the participants used domain explanations in both cases (73 out of 109), about one quarter (28 out of 109) of the participants w*ho used domain explanations for the traffic light classification question used mathematical explanation when explaining their answer to the carcinoma classification question*; participants who debated when explaining the traffic light classification question answer, gave either mathematical explanations (7 out of 14) or domain explanations (5 out of 14) when explaining their answer to the carcinoma classification question. The explanation of two participants debated in both questions.

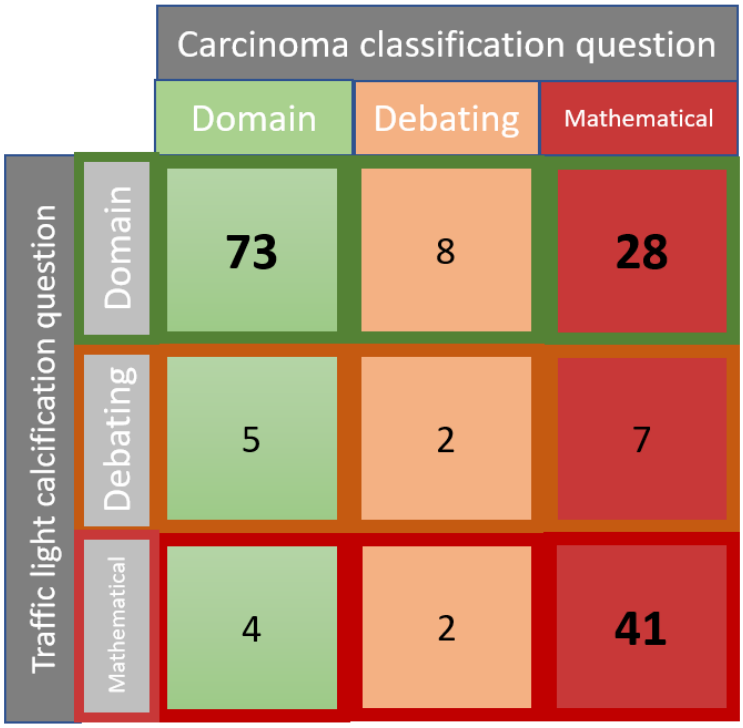


Figure 4.7 - Explanation types: Traffic light classification question vs. carcinoma classification question (n=170)

4.1.2.3. Discussion

From a research perspective, the questions designed for the research: the traffic light classification question and the carcinoma classification question, can be viewed as cognitive experiments. Like other cognitive experiments that, for example, request subjects to select one of two given investments with different interest and risk levels to invest in (Hertwig et al., 2004), our questionnaire presents two ML algorithms with different performances and asks participants to determine which is better (see Table 4.5 and Table 4.6).

Researchers distinguish between two types of such experiments: *description-based decisions* and *experience-based decisions* (Hertwig et al., 2004; Kudryavtsev & Pavlodsky, 2012). These terms refer to how the question is formulated when it is presented to the research participants. In *description-based decision* type questions, subjects receive, for example, a verbal description of two investments and, based on this description, must choose one. In *experience-based decision* type questions, subjects may receive, for example, the opportunity to (virtually) invest in the two investments being compared and are asked to choose their preferred investment after gaining some personal experience and learning how the investments behave. Researchers

found a considerable difference between decision making based on description and decision making based on experience. For example, in the case of description-based decisions, people make choices overweighting the probability of rare events, while in the case of decisions from experience, in contrast, people make choices that underweight the probability of rare events (Hertwig et al., 2004).

In the experiment described in this section, the questions the participants were asked in the questionnaire are hybrid and are each composed of both a description-based decision and an experience-based decision. The performance of the ML algorithms is described verbally (description-based) but the meaning of such performance in the real world is not described verbally and must be interpreted by the participants based on their own experience in the related real-world domain (experiment-based). While the traffic light classification question is based on a domain that is well known to the students based on their daily life, the carcinoma classification question refers to medical diagnosis, which (hopefully) most electrical engineering, computer science, and data science students are not familiar with from their daily life. Therefore, it is expected to see different patterns of answers to questions with similar mathematical properties depending on the participants' experience in different domains. In both cases, however, analysis of the participants explanations found that they explained their answers using three types of explanations but with different distributions: domain considerations, mathematical considerations, and a debate between the two.

The terms experience-based decision and description-based decision were coined to describe different situations in which human beings are required to make decisions. Our research suggests that these terms may also characterize a pattern exhibited by the decision maker. By implementing this idea in the analysis of the students' answers to the two questions, it is possible to characterize the connection between the type of explanation and the familiarity with the domain, using the experience-based decision and the description-based decision terms, to which the familiarity-based decision is add (see Table 4.8).

Experience-based decision:

- Students who used domain-based explanations in both the familiar and unfamiliar domains. *That is, they did not exhibit the domain neglect bias.*

- Students who used domain-based explanation in the familiar domain and debate in the unfamiliar domain, explicitly stated that they were not familiar with the domain and therefore could not decide.

Familiarity-based decision:

- Students who used a mathematical explanation in the unfamiliar domain, when in the familiar domain they either used a domain-based explanation or debated, showing once again that familiarity with the domain matters.

Description-based decision:

- Students who used mathematical explanations in both the familiar and the unfamiliar domains – that is, they exhibited the domain neglect bias.

This pattern suggests that some of the students were consistent in using domain-based explanations or mathematical explanations in both cases, while other students changed their pattern of decision making depending on their familiarity with the domain. The system 1- system 2 theory, proposed by Tversky and Kahneman (1983) may explain this phenomena. According to this theory, humans have two thinking modes: *System 1* represents a fast, instinctive, and emotional thinking mode; *System 2* is slower, more deliberative, and more logical. One function of System 2 is to detect and correct errors of System 1. It seems that in the case of the unfamiliar domain, System 2 fails more often to fix the intuitive answer of System 1. Yet, further research is needed for checking the relevance of the System 1- System 2 theory for the explanation of these results.

Table 4.8 - Relationship between explanation type and decision-making type[4]

		Unfamiliar domain	
		Domain-based explanation	Math-based explanation
Familiar domain	Domain-based explanation	Experience-based decision	Familiarity-based decision
	Math-based explanation	N/A	Description-based decision

[4] The debating-based explanation does not appear in the table as only a small number of students used it.

4.1.3. Conclusion

This case study demonstrates two challenges of data science education. The first part, biomedical signal processing projects (Section 4.1.1), demonstrates the challenge of mentoring learners who are not specialists in the application domain, a project development in an *interdisciplinary project-based learning (PBL)* environment. Based on my teaching experience with other populations (see Table 3.2 - Research population), PBL challenges several populations of data science learners including high school pupils, computer science and electrical engineering undergraduate students, data science majors without a specialization in an application domain and computer science teachers. The second part of this case study, which presents the domain neglect bias (Section 4.1.2), demonstrates that it is challenging to *integrate the application domain knowledge* in the context of data science education.

As part of the thesis justification, I now explain why these challenges are not traditionally addressed by the educational communities of each of the disciplines whose intersection creates data science. My thesis states that *it is essential to develop data science education as an interdisciplinary field of its own, relying on the principles of interdisciplinary education, and the integration of knowledge from the disciplines of mathematics education, computer science education, statistics education, and the application domain education*. The thesis is justified as none of the generating fields of data science education, *alone*, deals with all the challenges. Moreover, some challenges and opportunities are not addressed by any of the generating fields of the proposed discipline of data science education. In this case study, I assert that statistical education, mathematics education, computer science education and application domain education do not deal with the challenges of interdisciplinary project-based learning and integration of the application domain knowledge.

In the context of statistical education, statistical thinking emphasizes the importance of the context of a given problem when performing statistical investigations and drawing conclusions (Ben-Zvi & Garfield, 2004). In the context of computer science education, computational thinking emphasizes problem formulation, problem decomposition, abstraction, and generalization (Wing, 2006). Interdisciplinarity is not in the focus of computer science education (Mike & Hazzan, 2022), and therefore, the challenges of *integrating the application domain knowledge* and *interdisciplinary project-based learning* are not discussed, researched and mitigated in the context

of computer science education. In the context of the application domain education, the focus is on application domain knowledge, not on the integration of this knowledge in the data science workflow, as is demonstrated in this case study in the contexts of PBL and machine learning. Therefore, only statistical education and interdisciplinary education deals with the challenges of interdisciplinary projects based learning and integrating the application domain knowledge.

The interdisciplinary approach of data science education not only mitigates these challenges, as demonstrated in the case of biomedical signal processing projects (Section 4.1.1), but also presents a new opportunity of *teaching STEM with real-world context*. As data science projects always integrate real-world context of some application domain, teaching STEM subjects (for example computer science) in the context of data science leverages the opportunity of giving context to these STEM subjects.

4.2. Case study 2 - The base rate neglect cognitive bias

In section 4.1, I described the domain neglect cognitive bias, a novel cognitive bias found in this research. In this section I describe the expression of a well-known cognitive bias, *the base rate neglect*, first described by Kahneman and Tversky (1973), in the context of data science education

Kliegr et al. (2021) present theoretical difficulties in understanding the outcomes of ML algorithms, i.e., interpret its predictions and performance, duo to human cognitive biases. They do not present, however, empirical evidence for their claims. In this section I present evidence that show that about one third of students in an Introduction to ML course, from varied backgrounds (computer science students and high school teachers, data science, engineering, social science and digital humanities), fail to correctly evaluate ML algorithm performance due to the base rate neglect bias.

In the previous case study, the concept of machine learning classifier has been introduced. In this case study, a deeper understanding of classifiers is required, which refers to their performance indicators. Therefore, in this case study, I first describe performance indicators of machine leaning classifiers (Section 4.2.1) and then describe the base rate neglect bias and its expression in the understanding the performance indicators of machine learning classifiers (Section 4.2.2). I then describe the base rate neglect experiment (Section 4.2.3) and its results (Section 4.2.4). Finally, I discuss the implications of these results to data science education and how they support the thesis of this research (Section 4.2.6).

4.2.1. Machine learning classifiers performance indicators

A classifier is an algorithm whose task is to classify objects to one of several classes. It extracts by itself the classification rules from a labeled training data set. As its name implies, a training data set contains examples that are labeled with the class they belong to. In the training phase, the machine learns the classification rule by iterating over the tagged examples, gradually updating the parameters of the classification model to minimize the classification error. For example, a classifier might be presented with 10,000 photos, of which 5,000 are tagged as cats and 5,000 are tagged as dogs. In the training phase, the classifier learns a model that can distinguish between photos of cats and dogs. After the training phase, the classifier uses this model to classify photos it has never seen before.

Classifiers might also produce false predictions, mislabeling objects of one class as objects of the other class. A common way to present classification correctness is a confusion matrix, whose rows represent the real labels, and the columns represent the predicted labels by the classifier. The number on the diagonal of the confusion matrix represent the numbers of the correctly classified objects; numbers that are not on the diagonal represent the number of incorrectly classified objects, since the predicted label differs from the true label. Figure 4.8 represents a confusion matrix of the classifier presented above, that was train to classify cats and dogs and classifies 4% of the cats as dogs (20 out of 500) and 5% of the dogs as cats (25 out of 500).

		Predicted label	
		Cats	Dogs
Real label	Cats	480	20
	Dogs	25	475

Figure 4.8 - Confusion matrix

4.2.2. The base rate neglect cognitive bias

Cognitive biases are phenomena of the human brain that may cause erroneous perceptions and irrational decision-making processes (Kahneman & Tversky, 1973). For example, consider the following question regarding the classifier presented in Figure 4.8: If an image is classified as an image of a cat, what is the chance that this is indeed a photo of a cat? If the proportion of cats and dogs in the population is balanced, we can tell that the classifier will classify 505 images as cats (the sum of the left column), out of which 480 are cats (the upper left cell), and so the answer is 480/505 ≈ 95%. But what if the base rate of cats and dogs in the population is not balanced? In this case we cannot simply sum up the left column, but rather must consider the base rate of each class. Research indicates, though, that humans tend to neglect the base rate when performing such calculations.

For example, Casscells et al. (Casscells et al., 1978) asked 20 interns, 20 fourth-year medical students, and 20 attending physicians at four Harvard Medical School teaching hospitals the following question, referred to as the medical diagnosis problem:

"If a test to detect a disease whose prevalence is 1/1000 has a false positive rate of 5%, what is the chance that a person found to have a positive result actually has the disease, assuming you know nothing about the person's symptoms or signs?"

The correct answer to this question is about 2%, based on Bayes theorem (see Equation 1). Note that in the original phrasing of the question, the conditional probability of True-Positive is not given, and is assumed to be 1, meaning that the test is positive whenever the patient has the disease.

Eq. 1

$$p(Actually\ has\ the\ disease \mid Found\ positive) =$$

$$= \frac{p(Positive \mid Disease) \cdot p(Disease)}{p(Positive)} =$$

$$= \frac{(Positive \mid Disease) \cdot p(Disease)}{p(Positive \mid Disease) \cdot p(Disease) + p(Positive \mid No\ Disease) \cdot p(Do\ disease)} =$$

$$= \frac{1 \cdot 0.001}{1 \cdot 0.001 + 0.05 \cdot 0.999} =$$

$$= 0.0196 \approx 2\%$$

Only 18% of the participants in Casscells et al. experiment gave the correct answer. Several explanations were suggested for this phenomenon (Koehler, 1996). The base-rate neglect fallacy, identified by Kahneman and Tversky (Kahneman & Tversky, 1973), is one of these explanations, suggests that this mistake results from ignoring the base rate of the disease in the population. Another explanation was suggested by Leron and Hazzan (Leron & Hazzan, 2009), who explained this error using the dual-process theory (Kahneman, 2002).

A known method of mitigating the base-rate neglect cognitive bias is to work with frequencies (natural numbers) instead of probabilities (percentages). For example, Cosmides and Tooby (Cosmides & Tooby, 1996) rephrased the medical diagnosis problem as follows:

"One out of every 1,000 Americans has disease X. A test has been developed to detect when a person has disease X. Every time the test is given to a person who has the disease, the test comes out positive (i.e., the "true positive" rate is 100%). But sometimes the test also comes out positive when it is given to a person who is completely healthy. Specifically, out of every 1,000 people who are perfectly healthy, 50 of them test positive for the disease (i.e., the "false positive" rate is 5%). Imagine that we have assembled a random sample of 1,000 Americans.

They were selected by a lottery. Those who conducted the lottery had no information about the health status of any of these people. Given the information above, on average, how many people who test positive for the disease will actually have the disease? _____out of_____".

Cosmides and Tooby (Cosmides & Tooby, 1996) found that 56% of the participants in their research answered the medical diagnosis problem correctly when it was formulated using frequencies. When presented in this format (with frequencies, not probabilities), it is natural to represent the information in a confusion matrix (see Figure 4.9). In this case, it is easy to compute the precision as 1/(50+1), without using Bayes theorem.

		Predicted label	
		Detected	Not detected
Real label	Disease	1	0
	No disease	~50	~950

Figure 4.9 - The medical diagnosis confusion matrix

Casscells et al. (1978) and Cosmides and Tooby (Cosmides & Tooby, 1996) quantified the base rate neglect bias among medical students and staff in the context of medical diagnosis. In the research presented in this thesis, the base rate neglect bias was quantified among students taking an Introduction to ML course, in the context of ML classifiers, as is described below.

4.2.3. The base rate neglect experiment

Data was collected using a questionnaire. The questionnaire included two questions regarding the base rate neglect – the lion classification question (see Table 4.9 **- The lion classification question**) and the tomato's disease question (see Table 4.10). The lion classification question is the ML analog of the medical diagnosis problem proposed by Casscells (1978). In this question, the students are asked to evaluate the probability of a photo to contain a lion if an ML detected a lion in this photo, that is, the true positive rate of the lion detector. The false positive rate, that is the probability of a photo that do not contains a lion to be detected as a lion photo is given as 5%. As the false negative rate (lion photos that are not detected) is given as 0, all the lion photos will be detected. The question is thus: what is the percentage of non-lion photos in the detected-as-lion-photos group? This percentage depends on the ratio of lion photos in the dataset, that is, the base rate of lion photos, which according to the base rate neglect bias, humans tend to ignore. Similar to the medical diagnosis problem, the base rate of lions is given as 1:1000 and therefore, based of Bayes theorem, the true positive rate is about 2%.

The tomato's disease classification question is a ML analogous to the medical diagnosis problem phrased with frequencies (natural number) instead of probabilities (percentage) as proposed by Cosmides and Tooby (Cosmides & Tooby, 1996). In this case, a ML algorithm is trained to detect diseased tomatoes. Again, 5% of the healthy tomatoes are detected as diseased, but the question phrasing is that 5 out of 100 healthy tomatoes are detected as diseased. The base rate is also given as frequency, and 1:1000 is phrased as the disease attacks about 1 out of 1,000 bushes. The answer, that is the true positive rate, is the same as in the previous question, about 2%.

Table 4.9 - The lion classification question

A machine learning algorithm was trained to detect photos of lions. The algorithm does not err when detecting photos of lions, but 5% of photos of other animals (in which a lion does not appear) are detected as a photo of a lion. The algorithm was executed on a dataset with a lion-photo rate of 1:1000. If a photo was detected as a lion, what is the probability that it is indeed a photo of a lion? (a) About 2% (b) About 5% (c) About 30% (d) About 50% (e) About 80% (f) About 95% (g) Not enough data is provided to answer the question

Table 4.10 - The tomato's disease classification question

A machine learning algorithm was trained to detect a leaf disease in photos of tomato bushes. The algorithm detects the diseased bushes perfectly, but 5 out of 100 healthy bushes are also detected as diseased. The disease affects about 1 out of 1,000 bushes. If a bush is detected as diseased, what is the probability that it is really diseased? (a) About 2% (b) About 5% (c) About 30% (d) About 50% (e) About 80% (f) About 95% (g) Not enough data is provided to answer the question

The research was conducted in two phases: In the first phase, to verify the existence of the base rate neglect bias in the context of interpretability of ML algorithms, only the lion question (Table 4.9) was asked. In the second phase, to study the effect of using frequencies instead of probabilities, the tomato question was phrased and added to the questionnaire. Accordingly, in the second phase, both the lion classification question and the tomato's disease classification question were asked in the questionnaire.

The questionnaire was distributed to learners from a variety of backgrounds who were taking Introduction to ML courses. (Note: The course was not the same for all populations; each population was studying a designated course designed for it.) In the first phase, total of 98 learners answered our first questionnaire; in the second phase, a total of 174 learners answered the second questionnaire. In total, 272 students answered the two questionnaires – none of them answered the two questionnaires (See Table 4.11).

Table 4.11 - Research population

	Number of participants		
Population	**Phase 1**	**Phase 2**	**Total**
Undergraduate computer science electrical engineering and data science students	51	68	119
Computer science high school teachers	26	34	60
Social sciences and digital humanities researchers[5]	21	51	71
Total	**98**	**153**	**251**

4.2.4. Results

Figure 4.10 presents the results of lion's classification question in the first stage. As can be seen, the ML learners exhibited the same patterns found by Casscells et al. (1978): the majority ignored the base rate and answered the lion classification question incorrectly (61%); and, among them, the most common erroneous response was 95% (given by 32% of the participants in the first phase). While Casscells et al. (1978) found that 18% of the participants gave the correct answer,

[5] The social sciences and digital humanities researchers are represented in my research by 4 groups (see Table 3.1): Groups 17-18 (data science workshop for researchers in social science and digital humanities, Technion) and groups 19-20 (data science for psychological sciences, graduates, Tel Aviv University). These groups are internally diverse in participants disciplines, academic level and statistical background.

in our experiment, 39% answered this question correctly, that is, more than double than the rate found by Casscells et al. (1978). This difference can be explained in two ways. First, the question was asked as a multiple-choice question (rather than open one as in Casscells et al's case), providing the participants with several options. Thus, participants could consider the different answers or even guess the correct answer. The second explanation is derived from the participants' answers to the open question asked just after the lion's classification question, in which they were asked to explain their selected answer. More than half of the participants who answered the question correctly explained their answer by Bayes' rule by which they could calculate the answer precisely. It is, therefore, reasonable to assume that our population had more background in probability than did the populations of the Casscells et al.'s experiments.

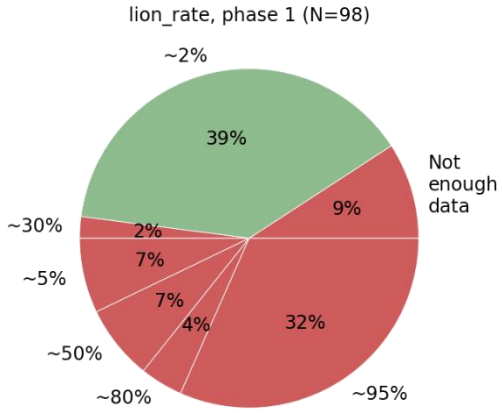


Figure 4.10 - Lion classification (phase 1): Distribution of answers (n=98).

Following Cosmides and Tooby (Cosmides & Tooby, 1996), in the second phase, the two questions were asked: the lion's classification question and the tomato disease question, which is an identical question but phrased with frequencies instead of probabilities, i.e., natural numbers instead of percentage. The answers to these questions exhibited the same pattern found by Cosmides and Tooby (Cosmides & Tooby, 1996) (see Figure 4.11). While in the original study (Cosmides & Tooby, 1996), 56% answered the frequencies question correctly, in our research 58% answered this question correctly.

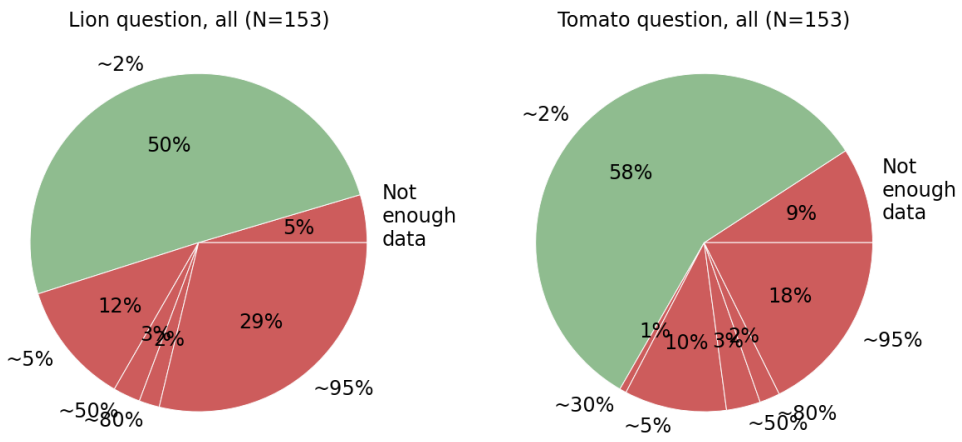


Figure 4.11 - Lion classification (left) vs. tomato disease classification (right): Distribution of answers (n=153)

I also note that the participants in the second phase, who had both questions, exhibited a higher rate of success answering the lion's classification (probabilities) question (50%; see Figure 4.11) compared with participants who answered the first phase questionnaire in which only the probabilities (with percentages) formulation appeared (39%; see Figure 4.10). An explanation to this phenomenon can be find at Ejersbo and Leron (Ejersbo & Leron, 2014), who suggest that the question phrased with frequencies (the tomato classification question in our case) bridged the intuitive and analytical thinking and helped participants reach a correct analytical solution for the probabilities question (the lion photos classification question in our case). It is also suggested that since more students answered the question phrased with probabilities correctly when it was also phrased with frequencies, questions phrased with frequencies should be considered as part of ML courses.

We now delve into the details of the different populations participated in this experiment, Figure 4.12 presents the answer distribution of the three research populations: computer science, electrical engineering, and data science undergraduate students (upper row), social science and digital humanities researchers (middle row), and computer science high school teachers (lower row). For each population, the left column presents the answers to the lion classification question in Phase 1, the middle column presents the answers to the lion classification question in Phase 2 and the right column presents the answers to the tomato's disease classification question presented only in Phase 2. To simplify the presentation of the different distributions, instead of presenting the prevalence of each of the 7 possible answers (about 5%, about 30%, about 50%, about 80%,

about 95%, and not enough data is provided to answer the question), all the wrong answers are unified and are represented by an “incorrect” category, and the 2% correct answer is represented by a “correct” category.

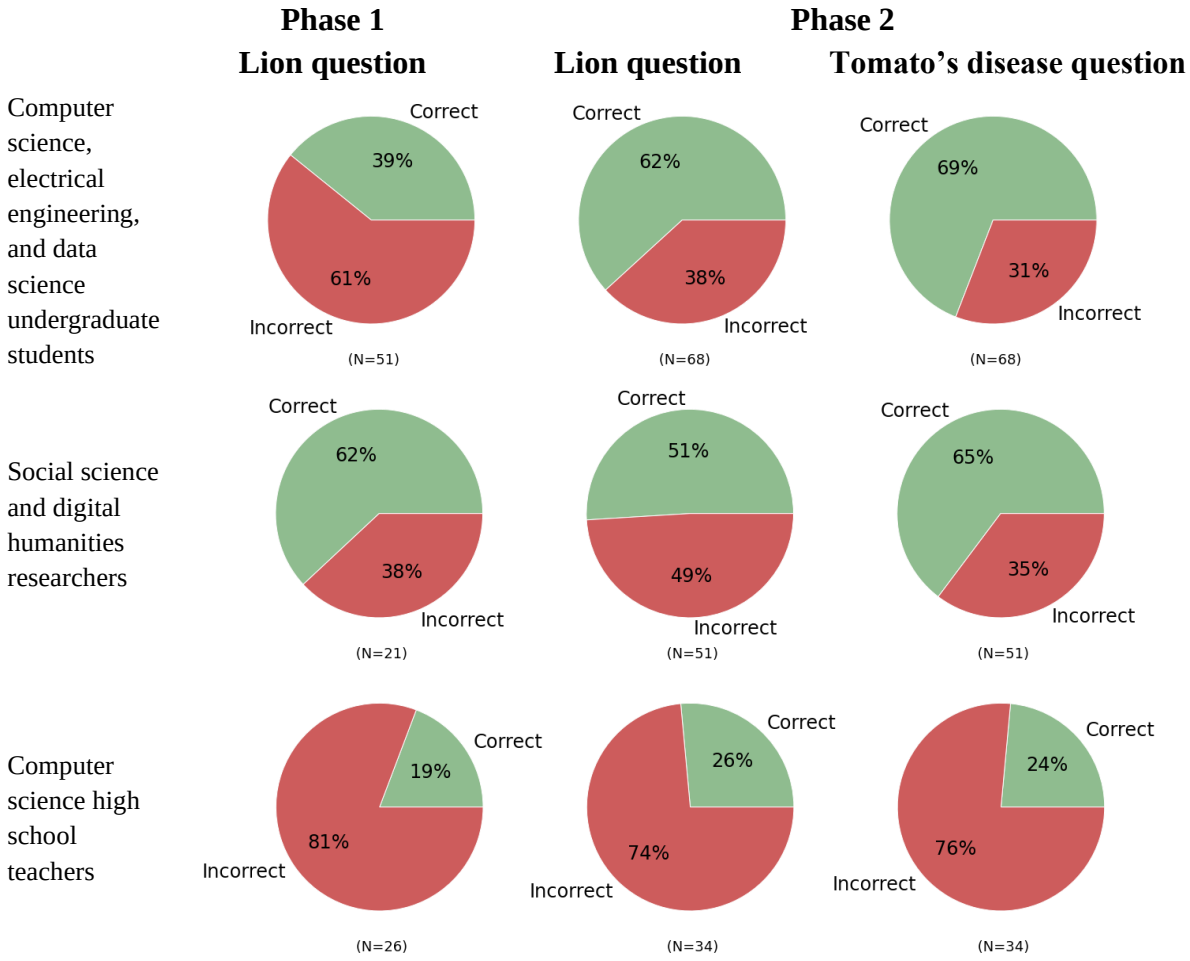


Figure 4.12 - Lion classification vs. tomato disease classification: distribution by phase and population

To compare the success rates between populations and between questions, a statistical test was applied to examine the significance of the differences between the distributions (Dodge et al., 2006). In this case, we are only interested in the success rate (correct / incorrect), and do not need to distinguish between the different types of errors. In other words, the different types of errors are treated as one type of answers (incorrect) and all answers are designated as one of two types: correct (~2%) or incorrect (all other answers). Therefore, two-tailed binomial test is used to test the hypothesis that the distribution of the answers to the two questions is significantly different for the different populations, for each phase (Bower & Richards, 2005; Shah & Kumar, 2002).

Table 4.12 presents the distribution of answers by question and by population and the significance of the difference between the answers for each population (the lion question in both phases, and the line question vs. the tomato question in Phase 2). As can be seen:

- For the undergraduate students, the difference between the answers to the lion classification question in Phase 1 and Phase 2 is significant, while the difference between the answers to the lion classification question in Phase 2 and the answers to the tomato's disease question is not significant. This implies that for this population, the tomato's disease question indeed mitigates the base neglect bias.
- For the social science and digital humanities researchers, the difference is not significant. This might be due to the heterogeneous statistical knowledge of this population.
- For the computer science high school teachers, the difference is not significant as only a low percentage of this population solved the tomato's disease question correctly, and therefore bridging the base neglect bias could not be supported.

Table 4.13 presents the comparison between the distribution of the answers of the different populations. As can be seen:

- The difference between the answers of the computer science, electrical engineering, and data science undergraduate students and the answers of the social science and researchers is significantly different only with respect to their answers to the lion classification question in Phase 1.
- The computer science high school teachers population is significantly different from both computer science, electrical engineering, and data science undergraduate students and social

science and digital humanities researchers in the three comparisons due to their low percentages of correct answers to the three questions.

Table 4.12 - Comparison of answer distribution by type of question

	Computer science, electrical engineering, and data science undergraduate students	**Social science and digital humanities researchers**	**Computer science high school teachers**
Lion classification Phase 1 vs. Phase 2			
Lion classification phase 1	Correct:20 Incorrect: 31	Correct:13 Incorrect: 8	Correct:5 Incorrect: 21
Lion classification phase 2	Correct:42 Incorrect: 26	Correct:26 Incorrect: 25	Correct:9 Incorrect: 25
Two tailed binomial value	p=0.00 (*)	p=0.11	p=0.27
Lion classification Phase 2 vs. tomato disease classification			
Lion classification phase 2	Correct:42 Incorrect: 26	Correct:26 Incorrect: 25	Correct:9 Incorrect: 25
Tomato disease classification	Correct:47 Incorrect: 21	Correct:33 Incorrect: 18	Correct:8 Incorrect: 26
Two tailed binomial value	p=0.26	p=0.051	p=0.84

(*) significant $p<0.05$

Table 4.13 - Comparison of answer distribution by population

Population 1	Computer science, electrical engineering, and data science undergraduate students	Social science and digital humanities researchers	Computer science high school teachers
Population 2	Social science and digital humanities researchers	Computer science high school teachers	Computer science, electrical engineering, and data science undergraduate students
Lion classification phase 1	p=0.04 (*)	p=0.00 (*)	p=0.04 (*)
Lion classification phase 2	p=0.11	p=0.00 (*)	p=0.00 (*)
Tomato disease classification	p=0.54	p=0.00 (*)	p=0.00 (*)

(*) significant $p<0.05$

4.2.5. Discussion

The ubiquity of ML algorithms in decision-making processes and the importance of ML interpretability and explainability is this context prompt ML researchers to develop new methods to better interpret and measure the performance of ML algorithms. There are, however, two side to ML interpretability and explainability: the machine and human beings. Advancing ML interpretability and explainability is, therefore, not the sole responsibility of ML researchers, but also of ML educators.

The questions presented (Table 4.9 and Table 4.10) can be used not only for research but also for teaching. The questionnaire is used as a pre-lesson activity for the topic of ML algorithm performance measurement. This pre-lesson activity reflects to the students their own biases and can, thus, enhance their understanding of the broad considerations required when implementing ML algorithms in real-world applications. There is a need to develop and share additional teaching methods for this important topic.

4.2.6. Conclusion

This case study demonstrates three challenges of data science education. First, it demonstrates the challenges of *lack of statistical knowledge*. Many of the participants in the first phase who answered the lion question correctly explained their answer using Bayes theorem, that is, used their knowledge in statistics to interpret correctly the performance of the ML classifier. Therefore, statistical knowledge can migrate the base rate neglect bias. It is also evident that computer science teacher, who have significantly higher mistake rate than the other populations, lack the (statistical) *Content Knowledge (CK)* and the (statistical) *Pedagogical Content Knowledge (PCK)*, required to practice and teach (respectively) data science. In Case Study 6 (the pedagogical chasm), I will elaborate on these challenges.

It can be seen that in the second phase, the tomato question, phrased with frequencies instead of probabilities, mitigated the base ate neglect for computer science, electrical engineering and data science undergraduate students. This pedagogical strategy, borrowed from statistical education, should be integrated into data science education. In computer science education or application domain education, however, there is no research and pedagogical methods to mitigate the base rate neglect bias. Moreover, computer science teachers do not have the required content knowledge in statistics and do not have the required pedagogical content knowledge in statistical education. Therefore, an interdisciplinary approach is required in data science teacher preparation.

Such an interdisciplinary approach not only mitigates the challenges, but also fosters new opportunities. This case study demonstrates the opportunity of *teaching statistical literacy* in the context of data science.

4.3. Case study 3 - Bridging gender gaps in STEM education[6]

This section describes the data science workshops for social sciences and digital humanities researchers (Cases 17,18 in Table 3.1 - Research Field). A significant majority (44 out of 53, 83%) of the participants of these workshops self-identified as women, although the base rate of woman among social sciences and humanities graduates in Israel is approximately 66%. This gender proportion, the opposite of that prevailing in STEM studies, led to the examination of the workshop from a gender perspective.

Data show that women are underrepresented in STEM subjects in K-12, academia and industry. For example, the statistics of the U.S. Department of Education (*Digest of Education Statistics*, 2020) on earned bachelor's degrees in 2018-19: Women were awarded only 21 of degrees in the computer science.

As data science is an emerging discipline, Berman and Bourne, (2015) propose that it has the potential to narrow the gender gap in the STEM subjects. This case study demonstrates one method to achieve this goal. The findings in this case study indicate that the women participating in the data science workshop perceived it as an opportunity to acquire research tools rather than programming tools. Framing the workshop as a research workshop and not as a programming workshop reduced prevalent gender barriers in STEM, encouraging a majority of women researchers to participate.

In this section I first describe the relevancy of data science for social science and digital humanities researchers (Section 4.3.1) and then describe the data science workshop for social sciences and digital humanities researchers (Section 4.3.2) and the applicants and participants of the workshops (Section 4.3.3). Next, I describe the data collection and analysis of the participants motivations to participate the workshop, based on the expectancy value theory (Section 4.3.4). In the discussion I present how the interdisciplinarity of data science may narrow the gender gap in STEM education (Section 4.3.5). Finally, I discuss the implications of these results to data science education (Section 4.3.6).

[6] This case study is based on Mike, K., Hartal, G. and Hazzan, O. (2021). Widening the shrinking pipeline: The case of data science, *IEEE EDUCON 2021 Conference*, Vienna, Austria.

4.3.1. Relevancy of data science for social science and digital humanities researchers

We live in the age of huge amounts data created by humans and machines. Humans create text, audio, photos, films and other types of data and share them in a variety of social networks. For example, 326 million new users joined social media between April 2021 and April 2022, equating to growth of almost 900,000 new users every day (*Digital around the World*, 2022). Machines generate and collect data about many facets of our lives: health, economy, education, transportation, meteorology, and so on. Wearable healthcare sensors, for example, can collect continuously data about pulse rate, blood pressure and other vital signs. These data can be analyzed to monitor the health status and to predict sickness. Accessible to all, regardless of socioeconomic status, location and nationality, social media collect data on people's interests, preference and other human characteristics that, in turn, can be used to predict changes in the social sphere and political trends.

Whether created by humans or machines, data are relevant for researchers in the social sciences and digital humanities. Text is one key data type in which researchers from both the social sciences and digital humanities are interested. As data, texts attract considerable attention not only in academia but also in the industry since they represent the most prevalent form of data available for both academic research and commercial applications.

As mentioned, the demand for data science education increases, leading to the opening of multiple academic courses, massive open online courses (MOOCs) and other programs that teach a wide range of learners (in terms of ages, occupations and education levels) in different levels of depth (Berman et al., 2018). With regard to education for adults, on one end of the spectrum lie academic programs for data scientists that include advanced mathematics, statistics and computer science, for both undergraduates and graduates(P. Anderson et al., 2014; Danyluk & Leidig, 2021; De Veaux et al., 2017; Demchenko et al., 2016). Undergraduate programs that integrate data science with another disciplines are also available (Havill, 2019; Khuri et al., 2017; Plaue & Cook, 2015). On the other end of the spectrum, we find experts in various disciplines who need data science skills to improve their professional work and to remain relevant in the information age.

This case study focuses on the second population, and specifically on researchers in social sciences and digital humanities who are experts in their disciplines but at the same time, in most cases, novices without background in the other two main components of data science, that is,

computer science and mathematics and statistics. Figure 4.13 - Computer science and statistics knowledge vs. the domain knowledge dimensions of data science education plots computer science and statistics knowledge against the application domain knowledge dimensions of data science education programs.

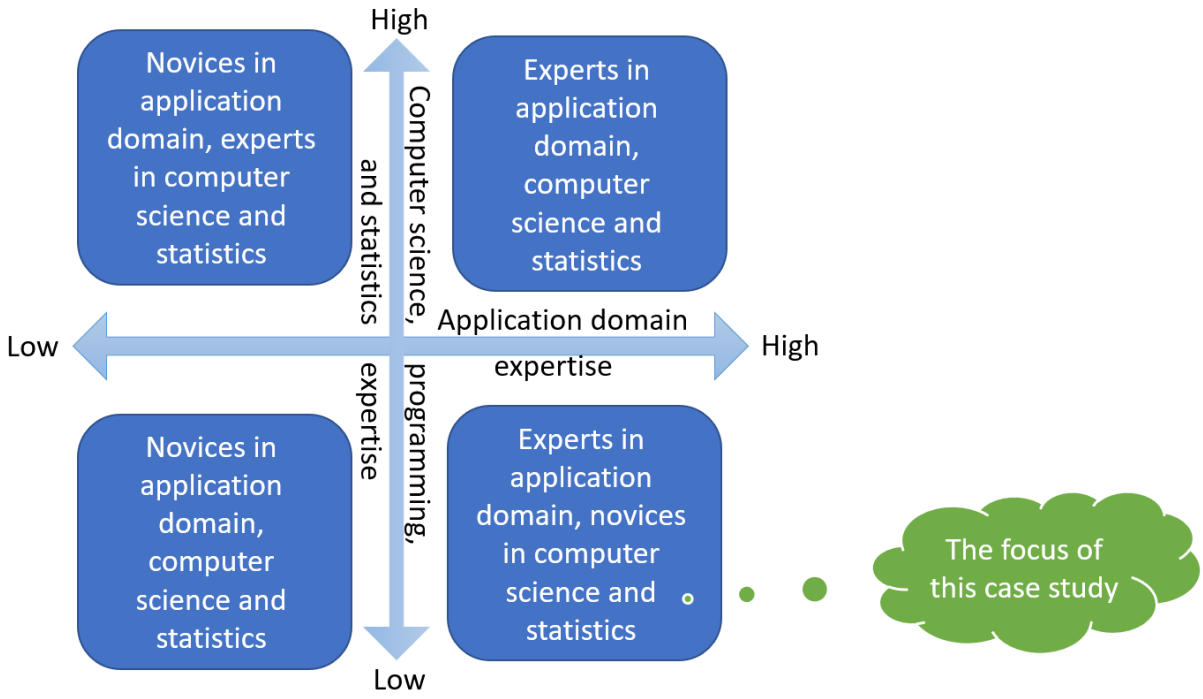


Figure 4.13 - Computer science and statistics knowledge vs. the domain knowledge dimensions of data science education[7]

It is important to educate social sciences and digital humanities researchers to understand data science given their role in decision making processes in a variety of application domains. While data science is clearly essential for the social sciences, the need of data science for digital humanities is less obvious. The Stanford Data Science Initiative explains: *"Data science for humanity – not only for the academic study of the humanities and the social sciences, but also for the betterment of humanity itself – is a deeply interdisciplinary effort… This interdisciplinarity is twofold: data science itself has benefited from the complex, important, and consequential research questions focused both on the rich history of human cultures and societies, and on the present state*

[7] The participants in this case study are internally diverse in their academic level and statistical background.

of humankind – in both its triumphs and its troubles. Data science can help diagnose those troubles, and suggest solutions." (*Data Science for Humanity | Data Science*, 2022)

Another example is course IS417, "Data Science in the Humanities", offered by the University of Illinois School of Information Sciences. The relevance of this course for the humanities stems from the fact that "*cultural materials usually come to analysts as unstructured texts, images, or sound files, forcing explicit decisions about data modelling and feature extraction. Cultural questions also highlight the importance of interpreting statistical models in relation to a social context. Last but not least: songs, poems, and stories confront us with vivid problems.*" (*Data Science in the Humanities*, 2022)

4.3.2. The data science workshop for social sciences and digital humanities researchers

Data science still lacks a commonly accepted definition, and its curricula are still in the process of being shaped (Berman et al., 2018). Thus, in designing our workshop, we had to consider both what the researchers already knew about data science and what they needed to learn in data science that will be valuable for their research. The workshop was designed according to the first principles of instruction (Merrill, 2002):

- Learning is promoted when learners are engaged in solving real-world problems.
- Learning is promoted when existing knowledge is activated as a foundation for new knowledge.
- Learning is promoted when new knowledge is applied by the learner.
- Learning is promoted when new knowledge is integrated into the learner's world.

To tailor the workshop for the researchers, a questionnaire was distributed to the applicants, asking them to rate their current knowledge level in data science related skills, such as programming and statistics, and asking what kind of research they expected to be able to perform at the end of the workshop (see Figure 4.14). Based on the analysis of their answers, the following guidelines were applied in the workshop curriculum design process:

- No previous programming knowledge is required.
- While there are many data analysis tools that do not require programming skills, the researchers did expect to learn to program. Thus, we decided to teach them Python (see Figure 4.15).

- As the learners had no or little programming experience, we also taught computational thinking and the basic principles of algorithmic thinking (Wing, 2006).
- With respect to the data to be analyzed in the workshop, most applicants indicated that they were interested in collecting data from the web. Therefore, we taught Python with emphasis on web scraping (see Figure 4.15).
- With respect to data analysis, most applicants indicated that they were interested in text analysis. Thus, the main theme of data analysis was text analysis with machine learning.

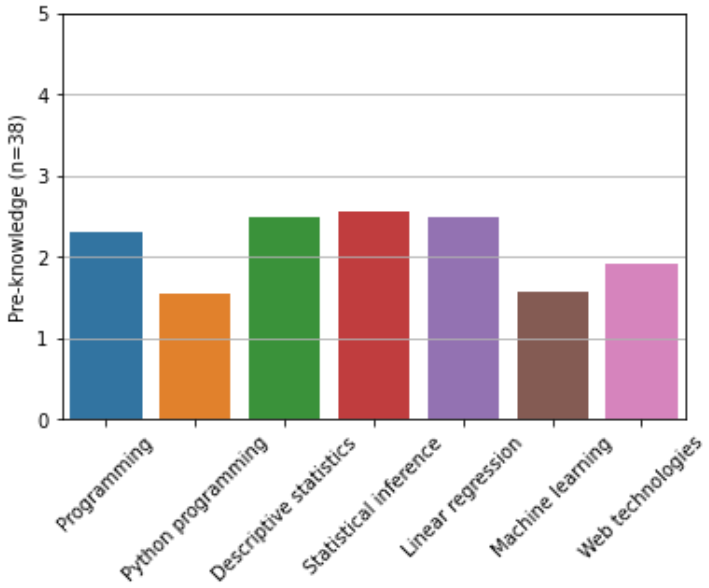


Figure 4.14 - Participants' pre-workshop knowledge in programming, statistics and machine learning (on scale 1-5, n=38)

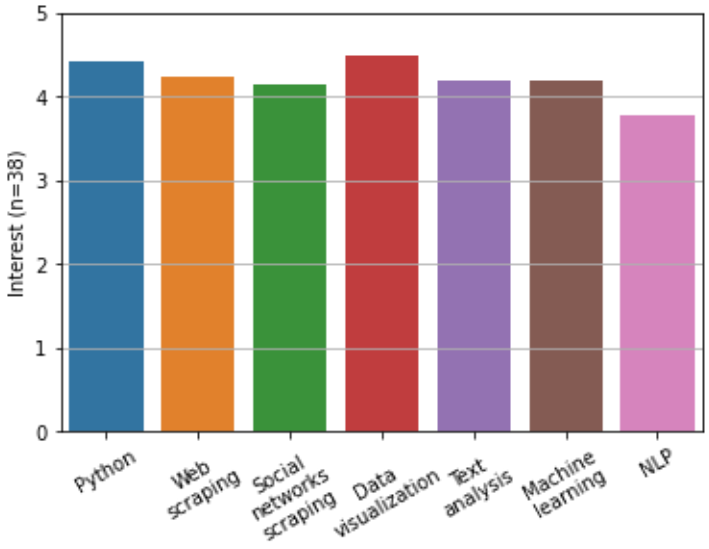


Figure 4.15 - Workshop Participants' intrest in studying the different workshop topics (on scale 1-5, n=38)

Two workshops were conducted, one face-to-face and one online, due to COVID-19 pandemic. The first workshop was based on 40% lectures, 40% exercises and 20% project. The second was an online hybrid workshop with prerecorded lectures and synchronous online ZOOM tutorials. The workshop topics and hours are detailed in Table 4.14. In both cases, the workshop was held in an intensive two-week format.

Table 4.14 - Workshop Topics and Hours

Chapter	Topics	Hours
Introduction to data science	What is data science, data science workflow	4
Introduction to web scraping	HTML, JS, ethics	6
Computational thinking	Computational thinking	4
Python programming	Data types, loops, conditions, file IO	20
Web scraping in Python	Beautiful Soup, requests, API libraries, Selenium	12
Data analysis in Python	Pandas, Matplotlib, Seaborn, SK-learn	12
Machine learning fundamentals	Types of machine learning, machine learning project workflow, machine learning algorithm performance estimation and improvements, hyper-parameter tuning	10
Machine learning algorithms	KNN, SVM, ANN	12
Text analysis	Bag of words, TFIDF, word embedding	10
Total		**80**

4.3.3. Workshop applicants and participants

One-hundred and ninety-five potential participants completed the preliminary questionnaire. Figure 4.16 shows the numerous disciplines of the researchers who completed this questionnaire. Figure 4.16 shows their gender distribution according to academic rank. Out of these applicants, 53 participated in one of two data science workshops: 29 participated in the February 2020 workshop and 24 in the September 2020 workshop.

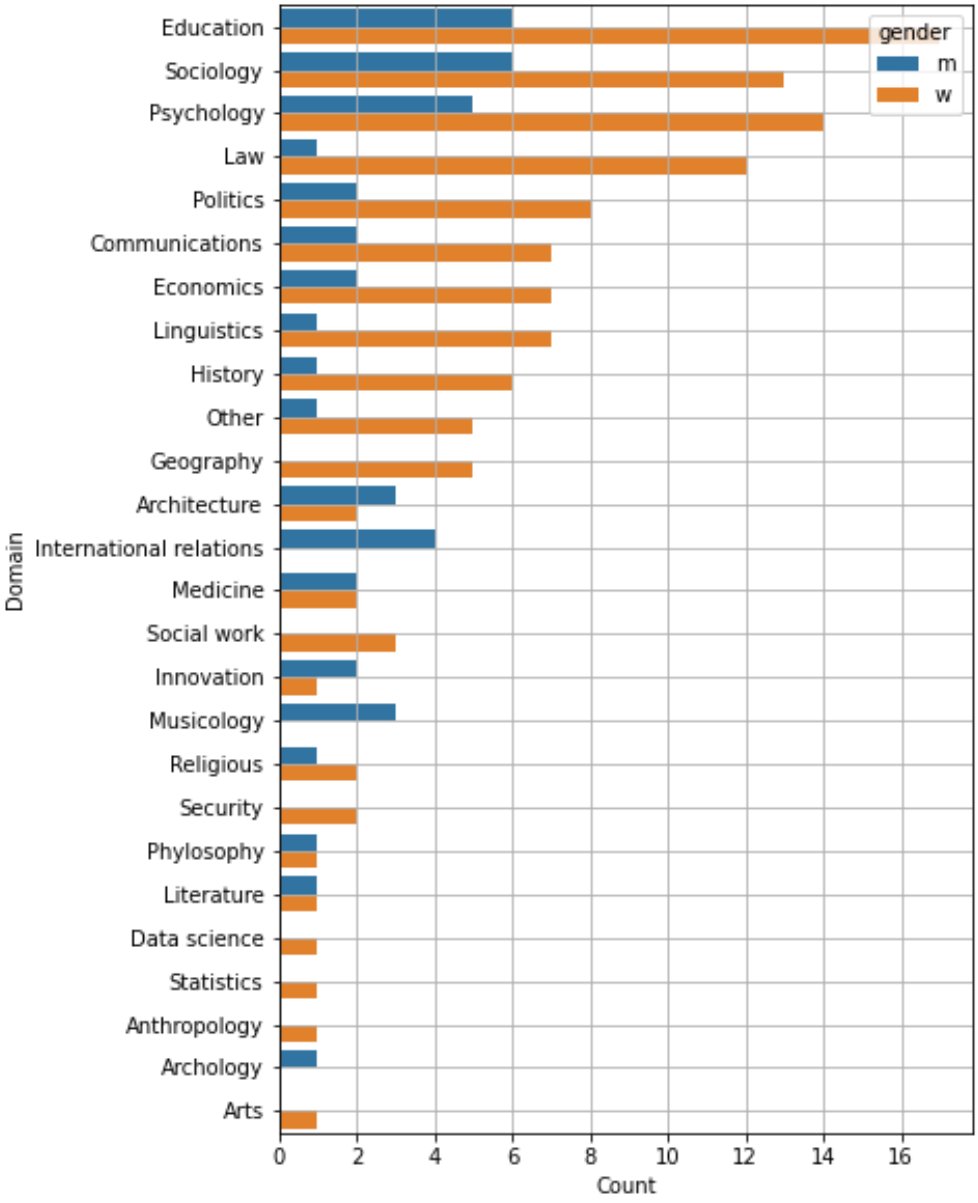


Figure 4.16 - Applicants' research disciplines by gender

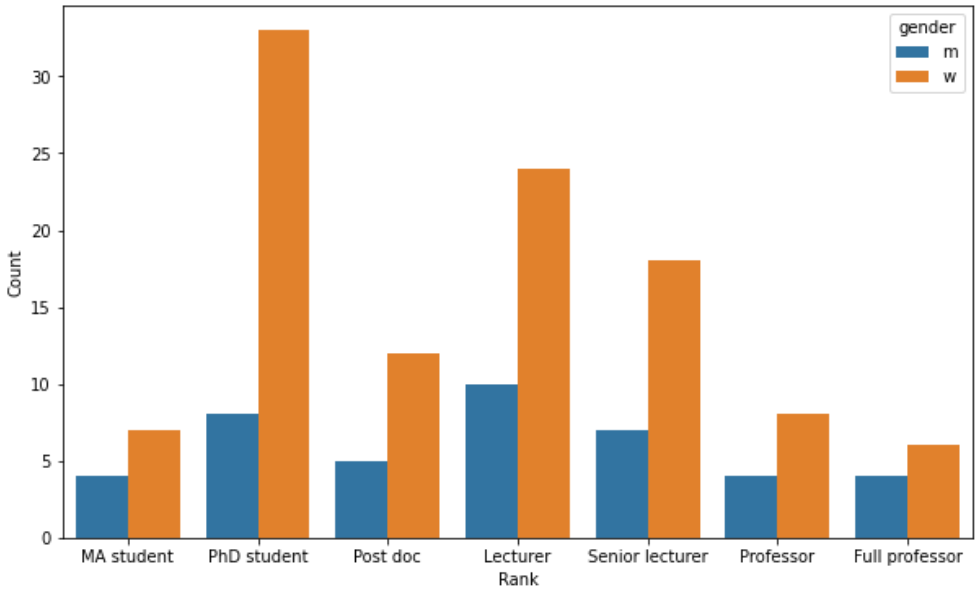


Figure 4.17 - Workshop applicants' academic rank by gender

4.3.4. Data collection and analysis

Data were collected using preliminary questionnaires distributed to social sciences and humanities researchers aimed at assessing their interest in the workshop, questionnaires for workshop participants, and participant observations.

The data were analyzed with respect to the expectancy value theory, a motivational theory that explains the motivation to engage in activities that involve achievements (Eccles & Wigfield, 2020). As the theory was originally developed to explain women's participation in STEM-related studies and occupations, it was found as a suitable theoretical framework for analyzing our data and presenting the results. According to expectancy value theory, the most influential factors in achievement-related choices are expectations for success and subjective task values. Below, our data on workshop participants' expectations are analyzed in terms of these factors.

4.3.4.1. Expectancy for Success

Expectancy for success is defined as individuals' beliefs as to how well they perform on an upcoming task. The name of the workshop, "Data Science for Social Science and Digital Humanities", did not reveal exactly what learners would have to do. The workshop syllabus specified the learning topics (see Table 4.14), but did not detail the tasks. Thus, the participants'

preconceptions of what data science was were crucial for understanding the nature of the workshop.

As mentioned, the data science field is both new and interdisciplinary – two properties found to be important for analyzing these preconceptions. First, as data science is a new field without a commonly accepted definition (Berman et al., 2018), learners may not have previous experience with it. Indeed, out of the 38 participants who completed the pre-workshop questionnaire, only six had experience with data science research, eight had participated in research involving data science, without however engaging in this aspect themselves, and 24 were complete inexperienced with data science.

Second, data science is an interdisciplinary field. While the Venn diagram of data science (see Figure 2.2) is symmetric, assigning equal value to each of its source disciplines, learners may differentially perceive the balance between them. For example, in an unpublished survey of undergraduate data science students at the beginning of an Introduction to Data Science course, we asked the students to list five terms associated with data science and three things they expected to learn in the course. The students mentioned terms associated with computer science about three times more than they did terms associated with statistics and mentioned expectations to learn topics related to computer science twice frequently as topics related to statistics. This small survey provided a preliminary indication to the fact that as suggested by Berman et al. (2018), different perceptions of data science could be reflected in attaching different weights to its components.

To understand how workshop participants perceived it in advance, we asked them in the preliminary questionnaire about their learning expectations. Out of 31 respondents, 18 indicated they expected to acquire data collection and analysis tools, eight indicated they expected to learn both data collection and analysis tools and programming, and five indicated they expected to learn programming. Therefore, we can assume that the participants' expectations for success in the workshop were based on their previous successful experience in data collection and analysis more than in programming.

Next, we asked the participants to rate their programming experience on a 1-5 scale. Out of 38 respondents, 11 indicated they had no programming experience at all and 16 indicated they had very little. When asked what challenges they expected to face during the workshop, 59% indicated programming as a challenge, 41% indicated available learning time and 22% indicated

the use of the tools acquired in their own research. Thus, even though most workshop participants had no or only little programming experience, and most perceived programming as a challenge, they were not discouraged from participation in the data science workshop, as it was perceived as more connected to data collection and analysis rather than to programming. In other words, though the workshop did include a considerable programming component, this was not perceived as a barrier, as it is sometimes perceived in more intensive computer science courses (Klawe, 2015; Kordaki & Berdousis, 2020), and thus, the participants did not have any reason to question their success in the workshop and were motivated to participate in it even without prior programming experience.

4.3.4.2. Subjective Task Value

Subjective task value is divided into four categories: interest value, utility value, attainment value and relative cost. Two of these categories are relevant to our data analysis: utility value and relative cost.

Utility Value

Utility value is the value that the specific task achievement will contribute to the overall goals of the individual. As this was seen as one of the most important motivation factors in our case study, we inductively analyzed our data from this perspective.

Based on a qualitative analysis of all the open-ended questions two major themes emerged as salient for the women researchers. The first theme is *skill acquisition* and specifically the ability to develop *independence in future research* that is methodologically based on data science tools. Only two men (out of 45) respondents addressed this issue. The second theme centered on *data science as a cutting-edge field,* reflecting back on the positionality of the researchers engaging in it and labeling them as more scientific. This perception represented a quality label attached to data science and endorsed by the women researchers. This theme also revealed a highly gendered pattern with 34 references by women (out of 119) researchers and only two from men (out of 45). In what follows, we further elaborate on these two themes.

<u>Skill acquisition and independence in future research</u>

When referring to their purpose in joining the workshop, almost all women researchers described their wish to acquire new tools for research and thereby gain access to a new

methodology. This purpose was similar across academic ranks and across universities. For example, one of the participants, a postdoc in political science, explained that her goal was *"to enrich the tools at my disposal"*. Other participant, a lecturer in a faculty of law, reflected that her objective was *"to gain basic skills for working with data science methods in order to engage in meaningful collaborations with data scientists, and be independent in my research"*. Finally, almost all women respondents indicated a highly instrumental purpose in joining the workshop: to utilize new research tools, and thus become competent in a field outside their discipline.

A specific focus in this skill acquisition process was the level of independence in data science research that may be gained by participating in the workshop. The women researchers hoped not only to attain new tools, but to be as independent as they could in applying them. Bearing in mind that the workshops were an intensive two-week process, the women researchers depicted what the satisfactory outcome for them. One of the participants, a lecturer at an academic center for internet research, said her goal was *"independence in research"*. Other participant elaborated: "*I would like to gain tools I can use to conduct research on my own, independently*". One another participant, a doctoral student from a psychology department, wrote: *"It's important for me not to be dependent on others with expertise in computer science or mechanical engineering to do the analysis for me"*. Another participant, a lecturer at a faculty of law, shared that her goal was *"to learn to do the work that computer science research assistants have done* [for me so far]*"*.

Attaining a cutting-edge and science-oriented position

The answers to the questionnaires reveal that the women researchers find data science as essential for current science. Since the world is perceived to be made of data, data science tools, they contend, would be an appropriate methodology and should be applied by all students as early in their academic trajectory as possible. For example, one of the participants, an associate professor in a department of information science said: *"I think that the knowledge in this field is a basic academic proficiency that nowadays every researcher needs"*. Other participant, a doctoral student added that data science *"is a large part of the world we live in today, and it is hard to stay relevant without at least knowing these tools"*.

When asked to shortly describe the workshop while taking it, one participant, an associate professor in a department of education said that it should be a required course for senior social scientists who did not take programming courses when they were undergraduates.

One of the participants, a historian, said that her goal was to *"learn the tools that are currently seen as the spearhead of humanities research... the future is there"*, and another participant, an information system coordinator added: *"This is a literacy that is taking center stage"*. This narrative linked the newness of data science with its importance, revealing that the participants acknowledged that in order to be significant in their research arena, one needed to also be on the cutting edge.

The women researchers felt that their current toolkit was ineffective and in need of revision. Big data tools, which are recognized innovative and fashionable, were also considered more science-oriented and hence more attractive. One of the participants, an associate professor in a department of education explained that *"data science can add a quantitative aspect to fields that were until now only studied through qualitative lenses"*. Other participant, a graduate student in a program in science, technology and society added that the data science *"way of thinking can influence a whole new generation of future researchers"*. By referring to the "way of thinking", another participant delineated that the social sciences and humanities popular and established way of thinking was not scientific enough, being based on qualitative analysis and small samples of data. Data science, on the other hand, utilized quantitative analysis and extremely large samples, resulting in a research process perceived as more scientific. Consequently, arguments generated through big data analysis were viewed as having greater value, reflecting on both the importance of the knowledge produced and the researcher's reputation.

Relative Cost

The relative cost factor represents the course of action an individual may choose as an alternative to the specific task. In our case, there are multiple alternatives for studying data science, including many online resources and MOOCs. These courses are intended mainly for self-learning. Our first workshop was held in February 2020, before COVID-19, and was face-to-face. Our second workshop was held in September 2020. At the time we prepared for the second workshop, we were still optimistic that another face-to-face workshop could happen. We wanted, however, to prepare for potential COVID-19 restrictions. Thus, the respondents were asked regarding their interested in participating in the workshop whether they preferred a full face-to-face workshop, a hybrid one with online lectures and face-to-face practice, a fully online synchronous workshop, a hybrid with asynchronous lectures and synchronous practice, or a fully asynchronous workshop.

Women preferred the face-to-face format (M=3.23, n=34) more than did men (M=2.84, n=13) while men preferred the online workshop (M=3.27, n=11) more than did women (M=2.6, n=33). A two-way ANOVA was conducted that examined the effect of gender and format (face-to-face, online) on preference. The results F (1, 87) = 2.87, p = .094, however, did not reveal statistical significance.

Due to COVID-19 social distance restrictions, we held an online workshop, and selected a hybrid format with asynchronous lectures and synchronous practice. Even though this format was not initially preferred by women, it was more similar to the preferred face-to-face format than the fully asynchronous option offered by MOOCs. Considering the importance, the participates attached to data science knowledge, we may conclude that the second workshop format was preferable to a MOOC, for example, since it reduced the relative cost of attending a fully a-synchronic course.

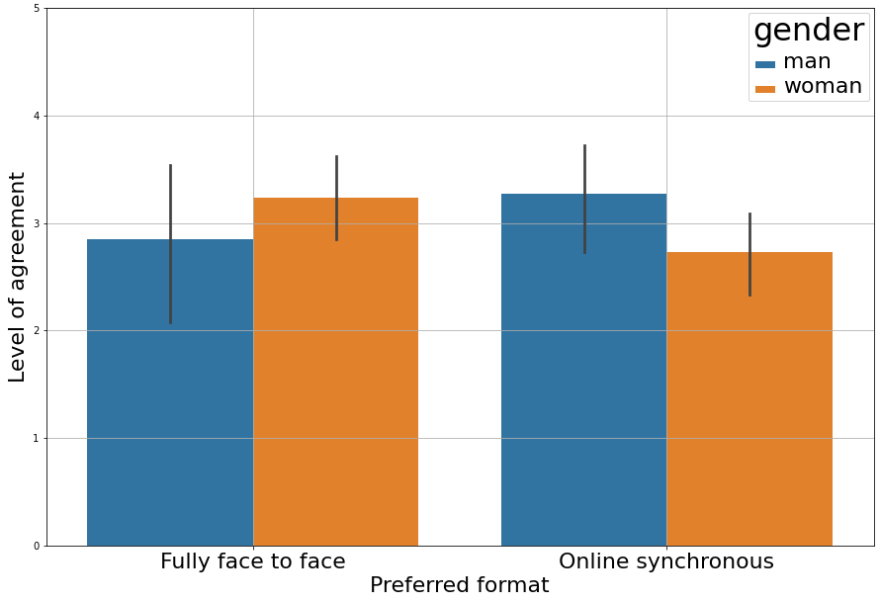


Figure 4.18 - Preferred workshop format by gender

4.3.5. Discussion

A generic data science curriculum can focus on computer science, mathematics and statistics knowledge and skills, with a capstone course integrating a real-life application domain(P. Anderson et al., 2014; Danyluk & Leidig, 2021; De Veaux et al., 2017; Demchenko et al., 2016). When data science is integrated with the studies of specific application domains, a more integrative approach can be taken (Anderson et al., 2014; Khuri et al., 2017). The presented workshop is based

on a third model, in which domain experts, researchers in social sciences and digital humanities, acquire computer science and mathematics and statistics knowledge and skills to enhance their possibilities to investigate data in their disciplines.

Extending their education to acquire interdisciplinary research skills might suit women scientists particularly well. According to Rhoten and Pfirman (2007), interdisciplinary scientific research offers four fundamental categories or mechanisms of activity: cross-fertilization, team-collaboration, field creation, and problem-orientation. Rhoten and Pfirman present evidence and possible explanations for why women researchers are more involved in each of the four interdisciplinary activities. In particular, they suggest that women scientists, especially in the early stages of their careers, spend more time on individual cross-fertilization activities as well as borrow tools, concepts, data, methods, or results from other fields or disciplines at greater rates than their male counterparts. This insight may explain why the women researchers in our study found such great interest in learning data science.

The results of this case study show that one major goal women researchers who participated in the "Data Science for Social Sciences and Digital Humanities Researchers" workshop aimed for was new tools and improved skills to facilitate greater independence in their work with data science tools. Integrating data science within their work was promoted by the interdisciplinary nature of data science, as described in what follows:

- Data science is essentially interdisciplinary. The workshop was designed to use this as an advantage. It did not and could not aim to give the researchers full training in each of the fields data science consists of. Instead, it was a readymade combination of tools that were taught, which provided the participants with an opportunity to gain skills and basic knowledge in the field. This was supported by work on the researchers' personal projects, aimed to allow them to integrate data science tools into their daily research.
- Compared with computer science and other STEM fields, data science was framed as less intimidating. Through the understanding of what data science is (and what it is not), the workshop represented a platform less threatening for the participants. The results, for example, showed that the participants were concerned with the need to have previous programming experience (59%) and were highly concerned with the workshop being time-consuming (40.74%), trying to juggle all their day-to-day tasks, rather than with the skills and knowledge

they needed to master in the workshop, given their lack of pervious knowledge in programming, statistics or mathematics.

- The women researchers were flexible and open to synthetic processes. They took the workshop because they wanted to integrate data science in their current or future work. They realized that in order to produce cutting-edge knowledge and new insights they needed to apply an interdisciplinary approach and integrate data science in their disciplines. They considered data science adequate and productive for this purpose because of its essential interdisciplinarity. Since the researchers' disciplines may or may not be flexible and open to change, the women researchers themselves lead the change. They embody the shared interests of their field and data science and use the workshop to lay the groundwork for an interdisciplinary perspective.

Avnoon (2021) proposes that data scientists are building their professional prestige on several unique skills, of which one is the ability to rapidly learn application domain knowledge. Based on this insight on the fundamental nature of data science as an occupation, we can identify this skill in the participants of the data science for social sciences and digital humanities workshop as well. In this case, however, the participants are acquiring the computational and statistical aspects of data science, as they are already specialists in their original discipline.

It is reasonable to believe that the adequate interdisciplinary perspective to data science is conceptual. Such framing has the potential to enable researchers in the social sciences and digital humanities in general and particularly women researchers to better employ data science as a tool. Therefore, it can help them lead their fields to a more interdisciplinary research paradigm.

Furthermore, the interdisciplinarity approach can also be applied to other STEM subjects and specifically to widen the proverbial shrinking pipeline – the situation where fewer woman than man decide to study computer science, and even less women decide to continue their higher education and peruse an academic career in this field (Camp, 2002).

4.3.6. Conclusion

While the previous case study (Section 4.2) demonstrated the challenge of *lack of statistical knowledge*, this case study demonstrates the challenge of *lack of computer science knowledge*. In the next case study (Section 4.4), interdisciplinary CS1 for non-major data science learners, I suggest that this lack of computer science knowledge can be explained by the fact that computer science courses fitted for non-major data science learners are not available. Hence, this challenge

is not addressed in the context of computer science education as well as by the other two building block of data science – mathematics and statistics, and the application domain.

This case study presents an opportunity to bridge the *lack of computer science knowledge* in the context of data science education. In the next case study, I demonstrate that teaching computer science in the context of data science is beneficial for learners, as it increases their intrinsic motivation. In this case study, I also demonstrate that teaching computer science in the context of data science leverages the opportunity to *bride gender gaps in STEM education*, which has been discussed in the context of computer science education for decades, without meaningful successes (Camp, 2002). As is illustrated in this case study, the interdisciplinary context of data science opens an opportunity to bridge the gender gap in STEM (Berman and Bourne, 2015).

4.4. Case study 4 - Interdisciplinary introduction to computer science course[8]

This section describes the course Computer Science for Psychological Science designed and lectured by me at the School of Psychological Science at the Tel Aviv University in the 2020-21 and 2021-22 fall semesters. This course is the first out of two courses in the Data Science for Psychological Science Specialization (Cases 19, 20 in Table 3.1). The second course in the specialization, titled Data Science for Psychological Science, is the focus of the next Case study: Teaching core concepts of machine learning to non-major data science students.

Data science programs require extensive knowledge and skills in mathematics, statistics, and computer science (Cassel & Topi, 2015; Danyluk & Leidig, 2021; De Veaux et al., 2017; Demchenko et al., 2016; National Academies of Sciences et al., 2018). Such knowledge is crucial for understanding and implementation of advanced data science and machine learning applications. While it is possible to teach and learn the principles of data science without computer science and programming knowledge, using computerized tools such as Orange data science, WEKA, KNIME, and SPSS, such learning is limited.

Graduate psychology students who wish to learn data science have already gained considerable psychology domain knowledge and have an advanced background in statistics (Crooks et al., 2019; Hazan et al., 2018). They still, however, must acquire the required level of computer science knowledge and skills. Learning computer science, however, might be quite challenging (Gorson & O'Rourke, 2020), as such appropriate course that fits their background and needs is not available. On the one hand, the CS1 course for non-majors is adequate for this population, but its learning goals may not meet the data science requirements; on the other hand, the CS1 courses offered to computer science majors are not appropriate for psychology students as these courses are taught on a very high level.

To bridge this gap, a new CS1 course was designed – Computer Science for Psychological Science course – that a) suits social science students and b) meets the required computer science level for data science. The course was evaluated by examining students' learning outcomes,

[8] This case study is based on Mike, K. and Hazzan, O. (2022). Interdisciplinary computer science education - The case of graduate psychology students, IEEE EDUCON 2022 Conference, Tunis, Tunisia.

reflections, and responses on questionnaires. Based on the data analysis according to the self-determination motivation theory (Deci & Ryan, 2013), presented in this case study, it is concluded that the course design enabled to achieve the course's targets.

In this Section, I first describe the data science for psychological science specialization (Section4.4.1), and then describe the computer science for psychological science course, which is the first course in the specialization (Section 4.4.2). Next, I present the course participants (Section 4.4.3) and present the data collection tools and analysis process according to the self-determination theory (Deci & Ryan, 2012) (Section 4.4.4). Next, I discuss the results from the perspective of interdisciplinary computer science courses (Section 4.4.5), and finally, I discuss the implications of these results to data science education (Section 4.4.6).

4.4.1. Data science for psychological science specialization

The Data Science for Psychological Science specialization is a new initiative of the School of Psychological Sciences at the Tel Aviv University. The specialization includes two courses: Computer Science for Psychological Science and Data Science for Psychological Science. The motivation to offer this specialization is three-fold:

- To expose graduate psychology research students to new research methods that are emerging from the new discipline of data science (see Section 4.3.1).
- To expose graduate psychology research students, in general, and cognitive psychology research students, in particular, to state-of-the-art artificial intelligence and machine learning models that are inspired by the human brain and human thinking.
- To address the market demand for data scientists with a confirmed background in social science, in general, and in psychological science, in particular.

The design of the specialization content was inspired by the data science Venn diagram, shown in Figure 2.2. The current psychology curriculum already contains psychology domain knowledge, as well as statistics (Crooks et al., 2019; Dempster & McCorry, 2009; Hazan et al., 2018; Valero-Mora & Ledesma, 2011; Wiberg, 2009), and so the graduate psychology students' understanding of the psychology domain and the statistics components of data science is relatively high compared with their knowledge and understanding of computer science.

The specialization is divided into two courses. The first course, Computer Science for Psychological Science, was designed to fill in the computer science knowledge gap. The second course, Computer Science for Psychological Science, was designed to develop students' ability to complete the data analysis cycle, that is, to explore, analyze, model, and predict using data they collected in the first course. The courses were designed in a flipped classroom format, with a-synchronous pre-recorded lectures and weekly Zoom online meetings devoted to answering students' questions and working on advanced exercises in small groups (Rosenberg-Kima & Mike, 2020).

4.4.2. The computer science for psychological science course

Introduction to computer science and programming courses, often called CS1 (Hertz, 2010), are designed to introduce students to the fundamentals of programming. CS1 courses are currently taught on two distinct levels. On the higher level, CS1 courses are intended for computer science majors or non-majors in allied fields such as data science or other sciences, technology, engineering, and mathematics (STEM) (Bryant et al., 2010). These courses require a deep understanding of computational and algorithmics thinking and set a high bar for students. On the lower level, CS0 courses (Babbitt et al., 2019) or CS1 courses for non-majors (Bryant et al., 2010) aim to expose students to programming and to the essence of computer science; these courses do not expect students to master programming at a high level.

CS1 for computer science majors courses are considered difficult and demanding, and suffer from an overload of learning goals and high failure rates (Becker & Fitzpatrick, 2019). Although many attribute the high failure rates to the difficult nature of programming, Luxton-Reilly (Luxton-Reilly, 2016) asserted that programming is easy and that the high failure rates are the result of non-realistic learning goals and poor pedagogy. CS1 courses can be even more difficult for students who are not majoring in computer science and allied disciplines (i.e., data science) and so several curriculum innovations and pedagogical methods have been proposed to support such students (Cortina, 2007; Forte & Guzdial, 2005; Guzdial, 2003; Sullivan, 2013). Forte and Guzdial (2005) developed a media computation course, a CS1 course tailored for non-majors, in which *"students learn how different media are encoded electronically and how to write basic code for manipulating digital images, audio, video, and text"* (p. 249). A similar initiative was developed by Cortina (Cortina, 2007) for engineering students in the form of an introductory

computer science course for non-majors that "*focuses on the major contributions of computer science from the perspective of the process of computation*" (p. 218). The course does not include programming in any programming language but, rather, focuses on algorithms and on the principles of computing and its history. Sullivan (2013) proposed a data-centric CS1 course, which covers the basics of databases and data mining and offers an introduction to programming and data visualization.

The Computer Science for Psychological Science course was designed to close the computer science knowledge gap. Accordingly, three learning goals (LG) were defined:

- LG1: To develop students' computational thinking, which is considered an important 21st century skill, and specifically, an important thinking skill required for programing (Wing, 2006).
- LG2: To enable students to collect data for social psychological research using web scraping with Python.
- LG3: To enable students to collect data for cognitive psychological research by means of computerized experiments developed with JavaScript.

Following previous research that showed that integrating application domain data can enhance students' engagement with technical content (Wiberg, 2009), the computer science content and programming tasks were linked to the interests of the psychology graduate students. Specifically, two relevant topics for social psychology students emerged in a focus group facilitated by the psychology school staff members: a) social networks analysis and web scraping, and b) experiments in computational cognition. Interviews with psychology school staff members also revealed that, whereas the common programming environment for social networks analysis and web scraping is Python, experiments in computational cognition are programmed in JavaScript. Website design with HTML constitutes a common body of knowledge required for both web scarping and experiments development.

Table 4.15 details the course outline according to the learning goals.

Table 4.15 - Computer science for psychological science - Topics and number of hours

Topic	LG addressed	**Hours**
Computational thinking	LG1	4
Website design and HTML	LG2	8
Website design and JavaScript	LG3	8
Python programming	LG2	16
Web scraping with Python	LG2	8
Computerized experiments with JavaScript	LG3	8
Total		**52**

Although CS1 courses for majors have been the focus of a large body of research (Luxton-Reilly et al., 2018), the scope of research on CS1 for non-majors is rare. In a recent survey, Becker and Quille (2019) mapped only 13 papers on CS1 for non-major courses out of 777 papers on CS1 published by the SIGCES community in the last 50 years. According to the literature review, the new course design is novel in four aspects:

- Whereas interdisciplinary CS1 courses for STEM majors exist (see, e.g., Tartaro and Chosed 2015), it seems that there are no such interdisciplinary courses for non-STEM majors.
- Whereas all existing CS1 courses are intended for undergraduate students, this course is intended for graduate students
- It may be challenging to develop an interdisciplinary course, as an interdisciplinary staff is required (Way & Whidden, 2014). We bypassed this challenge by fostering cooperation between the course staff (the researcher), who contributed the CS specialty, and the graduate students themselves, who contributed the psychology specialty.
- Working with real-life data is known to motivate students in introductory computer science courses. Whereas in the reviewed research works, non-major students learned using pre-prepared datasets, in our course, the students collected their own data, which were also relevant for their thesis research.

4.4.3. Course participants

The research participants were 32 students (20 women and 12 men) who attended the Computer Science for Psychological Science course (the first course in the specialization) during the Fall 2020 semester. Approximately two thirds of the students were MA students and a third were PhD students. Approximately half of the students were studying and researching social psychology and the other half were studying and researching cognitive psychology.

Prior to the onset of the first semester of the specialization, we asked the students to report on their motivation to study the course. For this purpose, we formulated a closed questionnaire based on a previous study that we conducted with social science and digital humanities researchers (see Case study 3 - Bridging gender gaps in STEM education). Eighty-nine percent of the students indicated that they are interested in learning new research tools and eighty-three percent indicated that they are interested in learning Python. We note that while as in Case study 3, most of the participants indicated that they expect to learn research tool, unlike Case study 3, in this case study most of the participants also indicated that they expect to learn programming. This expectation to learn programming can be explained by the framing of this course as an introduction to *computer science* course, as reflected by its name Computer Science for Psychological Science.

We also asked the students to evaluate their pre-course knowledge in the various topics learned in the course, on a 1 to 5 scale (see Figure 4.19). As expected, the students' background in psychological experiments and statistics was much stronger than their background in computer science.

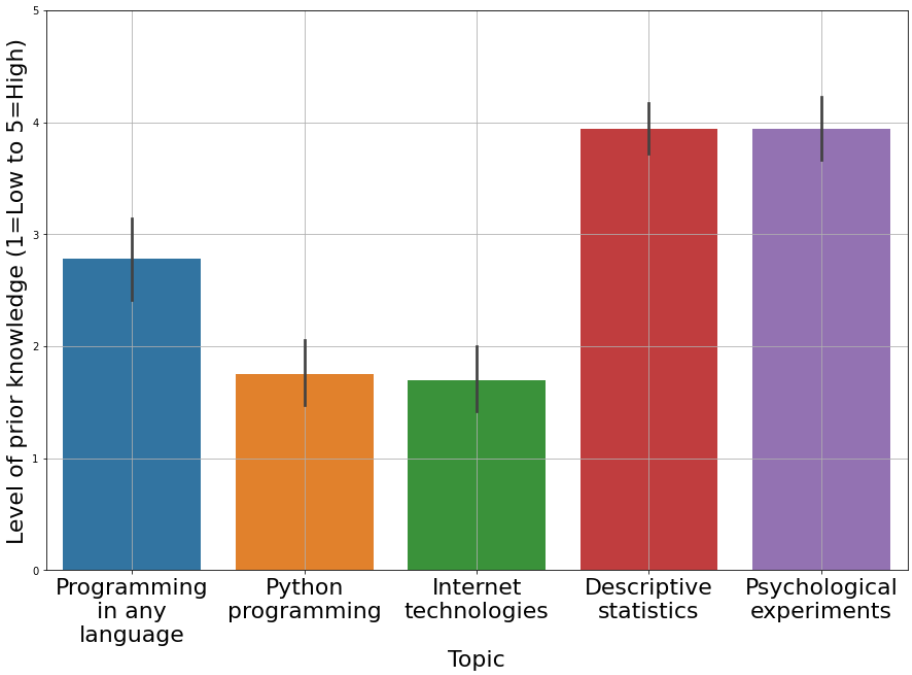


Figure 4.19 - Students' self-evaluation of their prior knowledge (n=36)

4.4.4. Data collection and analysis

Data for this study were collected from students' learning outcomes, reflections, and questionnaires (see Section 3.6)

The data is analyzed from the self-determination theory perspective. The self-determination theory (SDT) is a broad framework for understanding factors that increase intrinsic motivation, autonomy, and psychological wellness (Deci & Ryan, 2013). It emphasizes people's inherent motivational propensities for learning and growing. Research shows that in educational settings, intrinsic motivation results in positive outcomes (Reeve, 2002; Ryan & Deci, 2020).

According to SDT, human beings naturally seek growth and wellness, and three primary needs are particularly fundamental for supporting these processes: autonomy, competence, and relatedness. Autonomy concerns a sense of initiative and ownership of one's actions. It is strengthened by the learner's sense of interest and value and is weakened by external control, whether by means of rewards or punishments. Competence concerns the feeling of mastery. The need for competence can be fulfilled in education settings by well-structured challenges, positive feedback, and opportunities for growth. Relatedness concerns a sense of belonging and connection. It is enabled by the conveyance of indications of respect and caring.

SDT has been already used to evaluate teaching methods in the context of data science and machine learning. For example, Mariescu-Istodor and Jormanainen (2019) used SDT to analyze high school pupils' motivation to learn machine learning.

The data gathered during the course are presented according to the three fundamental needs that support intrinsic motivation: autonomy, competence, and relatedness.

4.4.4.1. Autonomy

Autonomy refers to a sense of initiative and ownership in one's actions. The course was designed to enhance students' autonomy by two means:

- As the course is taught using hybrid asynchronous lectures and synchronous recitations (Rosenberg-Kima & Mike, 2020), students had control over the pace of their learning process, as they could decide when and where to watch the pre-recorded lectures.
- Students work individually on weekly homework assignments, as well as on their final project.

Our data analysis shows that these pedagogical methods meet students' needs for autonomy. In the mid-semester questionnaire, we asked the students to evaluate, on a scale of 1 to 7, the contribution of the different types of educational activities to their learning (see Figure 4.20). It is notable that individual activities, such as learning from recorded lectures or working on tasks individually, received a higher score than group activities, such as discussions or group tasks. When asked to rank their preferred learning formats with respect to the hybrid learning on a 1-5 scale, their responses indicated that they preferred the asynchronous lectures–synchronous recitation format (see Figure 4.21), which, once again, enhances their feeling of autonomy in terms of when and where to study the course.

With respect to the course assignments, in a reflective questionnaire distributed to the students after the lesson in which the students presented their projects, two thirds of the students mentioned the following skills that they can apply autonomically at this stage of the course: development of a software tool required for their research; returning to lecture recordings for learning purposes; searching the web for new required features, new required software libraries, and solutions to problems; understanding code examples; and conducting experiments with code.

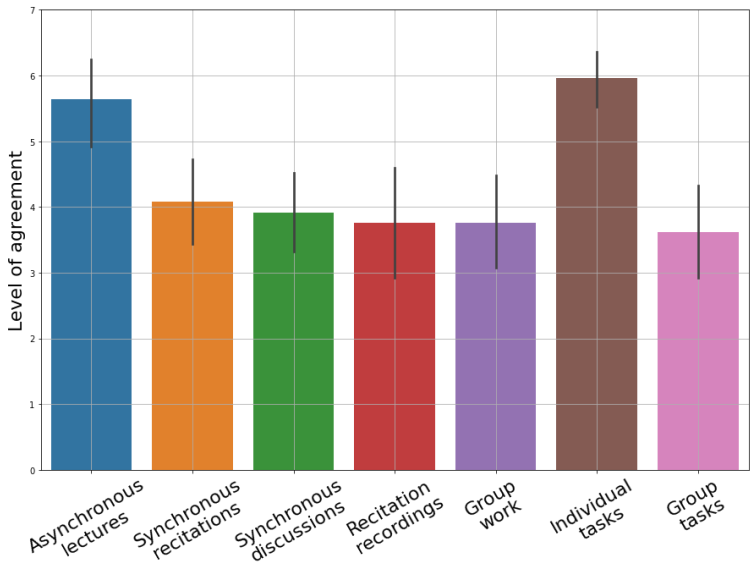


Figure 4.20 - Students' perception of the contribution of educational activities to their learning (n=25)

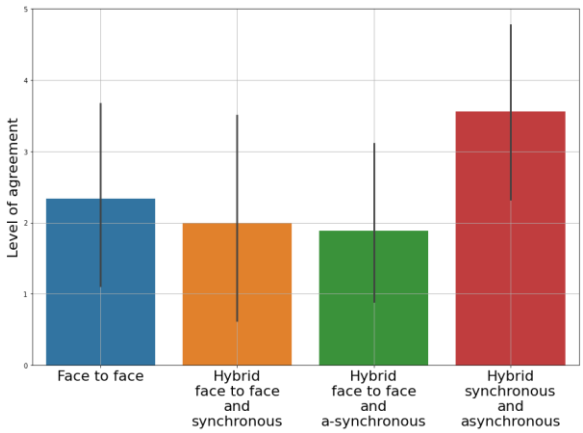


Figure 4.21 - Students' preferred course format (n=10)

4.4.4.2. Competence

Competence refers to a feeling of mastery. The need for competence can be fulfilled in educational settings by well-structured challenges, positive feedback, and opportunities for growth. While the programming level expected at the end of the course is high, the learning curve was designed for novice programmers with no prior programming experience. Accordingly, each computational concept and programming technique was taught in a separate short lecture video to enable the students manage their learning time individually during each week for each topic according to their knowledge level, confidence, and process.

In addition, the students received a task on each specific topic which integrated both computer science and data science concepts. For example, during the course, the students developed gradually a system that collects currency exchange rates from the Bank of Israel website. As part of the first and second topics of the course (computational thinking and introduction to web scraping respectively), the students analyzed the HTML structure of the website and wrote an algorithm in a pseudo code that extracts the required information. As part of the third topic on Python programing, the students analyzed the HTML as a string to extract the required information. As part of the fourth topic, web scraping with Python, the students developed automatic code that retrieves the information from the website and analyze it using library tools. Thus, not only was each new topic accompanied by a relevant task, but in addition, students could construct their knowledge gradually which, in turn, further enhanced their sense of competence and increased their motivation.

This last assertion is supported by students' feedback re the course pace. Each week, we asked the students to state whether the course pace was too slow, exactly right, or too fast for them. It is evident that the students perceived the course pace as adequate (see Figure 4.22), an important component in building students' perception of competence in their process of learning a new skill – programming – and its associated computer science concepts.

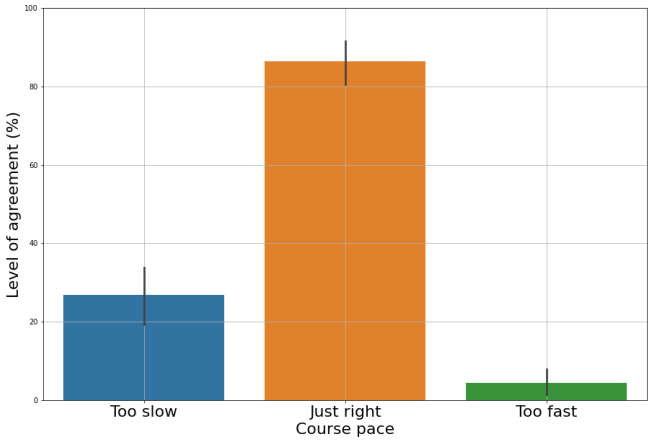


Figure 4.22 - Students' perception of course pace (N=165 collected by 15 students in 11 questionnaires at the end of each lesson)

We also derived the students' programming competence with respect to the programming requirement for data science from the final project, the students developed as the course final assignment. Here are several examples of the cognitive experiments they developed: an experiment

designed to investigate if and to what extent wearing protective masks inhibits human understanding of facial expressed feelings; an experiment designed to investigate the relationship between obsessive compulsive disorder (OCD) and indecisive behavior; an experiment designed to measure visual working memory capacity using the change detection task; and an experiment designed to measure the causal effect of working memory load on outcome-irrelevant reinforcement learning. One student designed a cognitive experiment whose purpose is to *"examine the extent to which working memory, a system that can filter and block the processing of irrelevant information in one's mind, predict credit assignment to outcome-irrelevant task features".* A presentation of the project can be found here[9]. A draft of the research paper written based on this experiment can be found here.

The students worked on their projects individually, and two recitations during the semester were devoted to project-related colleague consultation in small groups. Since the projects the students developed were used in their graduate studies research, that is, they accomplished a meaningful task for their real life, the fact that all students completed their projects successfully confirmed that they reached the required level of programming and fulfilled their need for competence.

In the end-of-course reflection questionnaire, approximately third of the students reflected on their self-perception of their competence, mentioning that they feel competent to develop programs that will meet their future research needs, as well as to learn new techniques when required. For example, one student wrote *"It is great to realize that I can build experiments on my own and I am sure I will use this ability".* Thus, based on students' successful project implementation and on students' reflections, we can conclude that the course learning fulfilled their need for competence both with respect to the required programming competence, the self-feeling of mastery in programming, and self-learning capabilities.

4.4.4.3. Relatedness

Relatedness refers to a sense of belonging and connection. The interdisciplinary design of the course and fact that students develop *experiments for their own research,* based on the material they learn in the course, demonstrate to the students the relatedness of computer science to their

[9] The project, the movie and the paper were created by Ido Ben-Arzi and presented here with his permission

own research and thus enhance their feeling of relatedness to both computer science and data science. Approximately one third of the students mentioned relatedness in their reflections on the final project in the end-of-course reflection questionnaire, stating that the programming tools they learned are powerful for the development of their psychological experiments, and that they can use them to collect data that would not otherwise have been accessible for their research. For example, one student wrote *"I think this course is significant for social science students. The course teaches an important research domain, which is rapidly growing these years"*. Such a statement is similar to those expressed in Case study 3.

4.4.5. Discussion

Integrating real-life data taken from the psychology domain is an important aspect of the interdisciplinary CS1 for psychological science course presented in this thesis. This aspect of the course can be generalized beyond the scope of this specific interdisciplinary course. It has been argued that learning computer science with real data can motivate students and support their learning (R. E. Anderson et al., 2015). It may be difficult, however, to design an interdisciplinary computer science course (Way & Whidden, 2014). One method of designing an interdisciplinary computer science course is to simply integrate data from other disciplines (Tartaro & Chosed, 2015).

We add that there is a hierarchy of possibilities for integrating real-life data in computer science education (see Figure 4.23 - Proposed range of possibilities for data integration in computer science education). We suggest that the two highest levels of the hierarchy, both of which use real-life data collected by the students, not only increase student motivation (as is illustrated in this thesis) but may also be easier to implement than data (either designed or real) provided by the instructor.

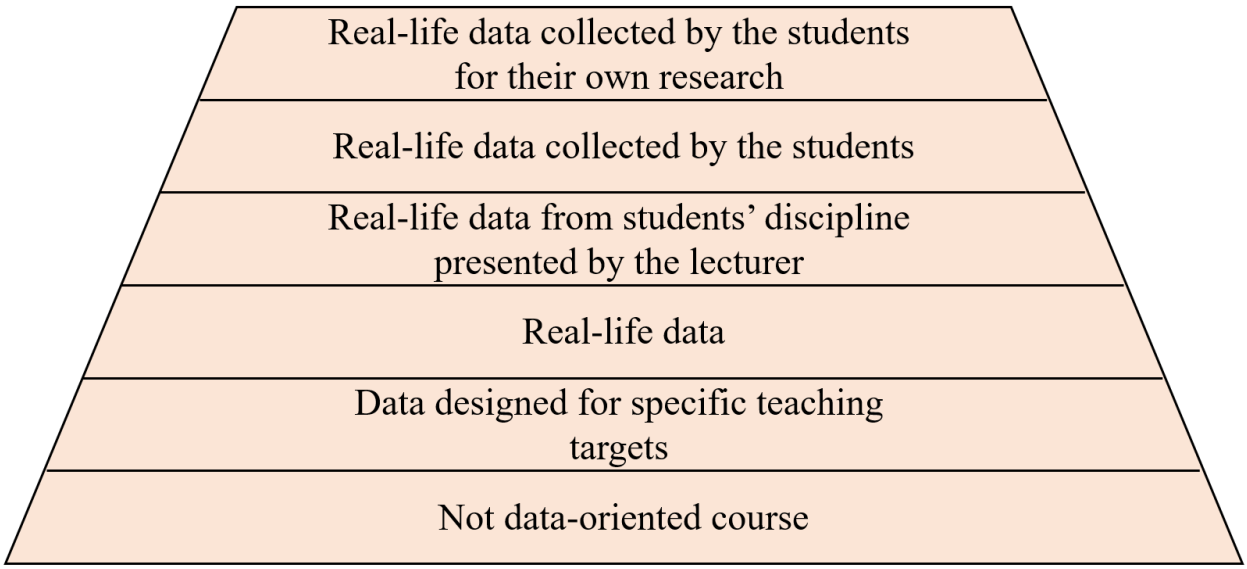


Figure 4.23 - Proposed range of possibilities for data integration in computer science education

4.4.6. Conclusion

This case study demonstrates the challenge of *lack of computer science knowledge*. While this challenge has been demonstrated also in the previous case study – bridging gender gaps in STEM Education, this case study also demonstrates how in the context of data science education this gap can be bridged, and students without computer science background can demonstrate meaningful performances, including a) learning two programming languages and b) developing a computer program to be used in their own research. In other words, the challenge of teaching CS1 to non-majors can be naturally mitigated in the interdisciplinary context of data science education.

This case study also demonstrates the opportunity of *teaching STEM with real-world data*, which supports the mitigation of this challenge. While in the context of computer science education the practice of teaching with real-world data is only currently embraced (Burlinson et al., 2016; Mike et al., in press), in the context of data science education, teaching with real-world data is a core principle.

4.5. Case study 5 - Teaching core concepts of machine learning to non-major data science students[10]

This section is based on the course Data Science for Psychological Science designed and lectured by me at the School of Psychological Science at the Tel Aviv University on 2020-21 and 2021-22 spring semester. This course is the second out of two in the Data Science for Psychological Science Specialization (Cases 19, 20 in Table 3.1). The first course in the specialization, Computer Science for Psychological Science, was presented in the previous Case study (Section 4.4): Interdisciplinary introduction to computer science course.

The focus of this case study is on teaching machine learning (ML) to graduate psychology students. In section 2.2.3 I described the importance and challenge of teaching ML as a white box to non-major data science students. Although the importance of a white box understanding of ML algorithms is clear, an extensive literature survey indicated that pedagogical methods for teaching white box understanding of ML algorithms to learners who lack a mathematical and computer science background, have yet to be proposed. In this case study I describe the pedagogical method I developed to bridge this challenge and its evaluation. Specifically, a hands-on pen-and-paper activity was developed based on theories borrowed from statistics and mathematics education research and students' answers were analyzed using these theories.

In this section, I first describe the Data Science for Psychological Science course, which is the research field of this case study (Section 4.5.1). I then present the object-process duality theory borrowed from mathematical education research, that inspired the design of the hands-on pen-and-paper activity and is also using for as a theoretical background for the data analysis (Section 4.5.2). Next, I explain the concept of white box understanding of ML algorithms and its importance (Section 4.5.3). Next, I illustrate the hands-on activity approach with the KNN algorithm (Section 4.5.4) and then I present the experiment I designed to measure the participants' level of comprehension (Section 4.5.5) and the research tool I designed for this experiment (Section 4.5.6). Then I present the data analysis method (Section 4.5.7) and the results (Section 4.5.8). Finally, I

[10] This case study is based on Mike, K. and Hazzan, O. (2022, in press). Teaching machine learning: A white box approach, *Statistical Education Research Journal (SERJ)*.

discuss the results and its pedagogical implications (Section 4.5.9) and discuss the implications of these results to data science education (Section 4.5.10).

4.5.1. The data science for psychology science course

The Data Science for Psychology Science course is part of a data science specialization track that consists of two courses, each lasting 13 weeks, with four hours of study per week (for a total of 52 hours per course). The first course, Computer Science for Psychology Science, was designed to fill the computer science knowledge gap (see Section 4.4.2).

The second course was designed to develop students' ability to complete the data analysis cycle, that is, to explore, analyze, model, and predict using data (see Table 4.16). The course therefore focused on data analysis including exploratory data analysis, descriptive statistics, statistical inference, and statistics and ML predictive models using Python.

As the psychology curriculum already contains extensive statistics content and the first course teaches the necessary background in programming, the second course elaborates on ML, which is located, in the data science Venn diagram, in the intersection of statistics and computer science (see Figure 2.2). Six ML algorithms were taught in the course:

- The k-nearest neighbors (KNN) and perceptron algorithms, taught mainly for their potential pedagogical value, as they convey the basic principles of ML using simple mathematics and can be taught easily applying the white box approach.
- Linear regression and logistic regression are a little more complex. Since the students are already familiar with the black box implementation of these algorithms from their statistics background, these algorithms are taught in this course with a white box approach including all mathematical details.
- Support vector machines (SVM) and neural networks are taught with partial mathematical details.

Because of the COVID-19 pandemic limitations, the course was designed in a flipped classroom format, with a-synchronous pre-recorded lectures and weekly Zoom online meetings devoted to answering students' questions and working on advanced exercises in small groups (Rosenberg-Kima & Mike, 2020).

Table 4.16 - Data Science for Psychological Science course - Topics and number of hours

Topic	Hours
The data science workflow	4
Table manipulation with Python	4
Visualization with Python	4
Statistical inference with Python	4
Principles of machine learning	8
Supervised machine learning	16
Unsupervised machine learning	4
Text analysis	8
Total	**52**

4.5.2. The object-process duality theory

The different ML algorithms taught in the program require students to understand the mathematical concepts of vectors, vector distances, dot products, derivatives and partial derivatives, and function optimization. None of these mathematical concepts is part of the psychology graduate curriculum (Rabin et al., 2018). The teaching of these algorithms, therefore, requires that the gaps in the students' mathematical knowledge be filled as part of the data science course or, alternatively, that some of the mathematical details be omitted and the gaps filled with alternative intuitive explanations. As our pedagogical goal is to foster students' learning of ML algorithms using a white box approach, our objective was to fill the mathematical knowledge gap. Accordingly, we looked for pedagogical theories and methods that could be borrowed from mathematics education research. From these theories, we found the process-object duality to be suitable.

According to the process-object duality theory, abstract mathematical concepts can be represented in the human mind as either objects or processes (Sfard, 1991). As an object, an abstract mathematical concept is conceived of as a fixed construct, and as a process, an abstract mathematical concept is conceived of as an algorithm or a computation that generates an output from an input. For example, as an object, the concept of a function can be conceived of as a set of ordered pairs $\{(x_i,y_i)\}$, whereas as a process, a function can be represented in the human mind as the steps required to calculate the function output value y_i for a given input value x_i. In the learning processes of most mathematical concepts, the learner passes through three phases. First, the

concept is conceived of as a process. Then, the process is mentally repacked (encapsulated) and an object representation is created in the learner's mind. In the final step, after the concept has been repacked and has become an object, it can be used by the learner as an element of a more complex process. Conceiving of abstract mathematical concepts as objects, therefore, reflects a deeper understanding than conceiving of them as processes.

Based on the process-object theory of mathematics comprehension, it can be argued that understanding the algorithm as a process is an essential step toward understanding it as a procept (explained below) and as an object.

The concept of *procept*, introduced by Gray and Tall (1994), represents the mathematical duality that exists between the understanding of a mathematical concept as a process and as an object. It reflects the idea that an advanced thinker can hold both mental structures of a concept in his or her mind and can move back and forth between the two. Hence, a procept is a hybrid schema, an amalgam of the two representations, as a process and as an object.

In practice, we should ask: How do learners build their mental representations of mathematical objects, and how do they advance their representations along the process-object continuum? A current mathematical education theory suggests that the learning of mathematical concepts is accomplished by participating in mathematical routines, which can be viewed as repeated executions of mathematical tasks (Heyd-Metzuyanim & Graven, 2019; Sfard & Lavie, 2005). There are two types of mathematical routines: rituals and exploration (Lavie et al., 2019). A ritual is the execution of a mathematical routine on a lower level of abstraction, i.e., a simple repetition of the mathematical procedure, based on a representation of the mathematical concept as a process. Exploration is the execution of a mathematical routine on a higher level of abstraction and consists of constructing the mathematical concept as an object. Learners are required to practice rituals to be able to proceed toward an exploration process. Lavie et al. (2019) claimed that *"in initial encounters with a new discourse, the learners can only participate in this discourse in ritualized ways. In further learning, their routines are expected to undergo gradual de-ritualization until they eventually turn into full-fledged explorations*" (p. 1).

4.5.3. White box understanding of machine learning algorithms

The goal of our study was to design a pedagogical approach for white box understanding of ML algorithms. Not only are ML algorithms based on advanced mathematical concepts, but

they themselves can be described as the "*mathematical function mapping [of] some set of input values to output values*" (Goodfellow et al., 2016, p. 5). It is therefore relevant to examine learners' conception of ML algorithms in term of understanding mathematical concepts as processes and objects.

Understanding an ML algorithm as a black box refers to the ability to call a library procedure that executes the algorithm without understanding the internal process of how the algorithm's output is generated. Such understanding, however, is not sufficient, for example, for hyperparameter tuning since students' unfamiliarity with how the algorithm works does not allow them to understand how to tune it. For example, one of the teachers interviewed by Sulmont et al. (2019a) suggested that *"their students cannot understand how tuning works, because they lack the mathematical prerequisite to understand parameters. Therefore, they claim, they think* [ML] *is magic when you tune parameters and get different results."* (p. 11).

Understanding an ML algorithm as a white box refers to knowing how it works. Since ML algorithms are mathematical objects, white box understanding can be further described as understanding ML algorithms as processes, as objects, or as procepts. According to Sfard (1991), understanding an algorithm as an object requires an initial understanding of it as a process. Thus, the more a learner practices the execution of the algorithm as a process, the more chances he or she has of gradually developing an understanding of it as an object. Accordingly, the following question guided our research: What pedagogical methods can enhance students' understanding of an ML algorithm as a process, as the first step toward its encapsulation into an object?

Specifically, we asked ourselves whether it would be effective to implement the algorithms as computer codes and track their execution flow. Tracing problems are a common type of question in computer science education (Hazzan et al., 2020). In most cases, however, such problems focus on software execution tracking rather than on mathematical calculations. Another approach argues that the learning of mathematical concepts can be supported by programming exercises (Leron & Dubinsky, 1995), but it requires an understanding of the mathematical concepts, at least as processes, as well as proficiency in programming. This does not apply, however, to students who are not majoring in data science and who are studying ML; as such learners lack the required mathematical background and programming proficiency, and therefore, implementing an ML

algorithm programmatically might not support their learning of the underlying mathematical concepts.

We thus developed a different approach, building on the mathematical education theory of process-object duality. Accordingly, based on the process-object theory, we first explained the ML algorithms using only mathematical objects the learners are familiar with. For example, in our explanations of the KNN algorithm (see Section 4.5.4), we used only the Euclidean distance, which requires only basic mathematical operations (addition, subtraction, squaring, and taking square roots). Second, building on the theory of mathematical ritual and exploration, we designed learning opportunities based on executing hands-on manual calculations of the algorithms. Similar to first grade pupils learning how to calculate the difference between two numbers by practicing ritual calculation of mathematical exercises until the ritual turns into exploration from which learning is derived, our students learned, for example, to calculate the difference between two vectors, in a hand-on task, as illustrated in the next section. Specifically, we designed pen-and-paper tasks that mimic the hybrid of the mathematical routine and the computer science tracing task, in which the students are required to perform the calculations using pen and paper rather than computer software. These exercises focused solely on the mathematical concepts and were phrased using mathematical equations and symbols rather than code or pseudo-code.

Hands-on calculation, as a pedagogical tool, has been experimentally investigated to determine whether it improves the understanding of various statistical concepts. Although in some experiments, no significant improvement was found (Pfaff & Weinberg, 2009), others showed significant improvement (Hancock & Rummerfield, 2020). In both cases, students' feedback on the activity indicated that it positively influenced their learning and motivation.

In the next section, the hands-on activities designed to impart the mathematical background required for learning ML is illustrated using one of the simplest ML algorithms learned in the course: the KNN algorithm. Similar hands-on tasks were developed for the perceptron, gradient descent, linear regression, logistic regression, and neural network algorithms.

4.5.4. Hands-on activity for learning the mathematics underlying machine learning

To illustrate the idea of using mathematical rituals in the teaching of ML algorithms, we present one example of implementing our hands-on approach for teaching the K-nearest neighbors (KNN) algorithm. KNN is the first algorithm in the curriculum since it is a relatively simple and

intuitive classifier that uses the classification algorithm presented in To generalize the algorithm to the *n*-dimensional case, we present the vector distance between the unknown object, *U*, and the i^{th} sample ($X^{(i)}$) as a generalization of the Pythagoras theorem, given by Equation 3.

(Eq. 3) $$d^{(i)} = \sqrt{(u_1 - x_1^{(i)})^2 + (u_2 - x_2^{(i)})^2 + \cdots + (u_n - x_n^{(i)})^2}$$

Table 4.17 and explained below (Hazzan and Mike, 2022).

The distance between a new object and its neighbors is calculated by vector distances. Several distance measures may be selected; we teach the Euclidian distance measure, as it is more intuitive to explain using its geometrical meaning. We therefore begin with a classification problem of objects with two features, i.e., a two-dimensional case. In this case, the Euclidian distance can be calculated based on the Pythagoras theorem. The distance, $d^{(i)}$, between the unknown object, *U*, and the i^{th} sample ($X^{(i)}$) in the training set is given by Equation 2.

(Eq. 2) $$d^{(i)} = \sqrt{(u_1 - x_1^{(i)})^2 + (u_2 - x_2^{(i)})^2}$$

To generalize the algorithm to the *n*-dimensional case, we present the vector distance between the unknown object, *U*, and the i^{th} sample ($X^{(i)}$) as a generalization of the Pythagoras theorem, given by Equation 3.

(Eq. 3) $$d^{(i)} = \sqrt{(u_1 - x_1^{(i)})^2 + (u_2 - x_2^{(i)})^2 + \cdots + (u_n - x_n^{(i)})^2}$$

Table 4.17 - The classification algorithm of the KNN algorithm

To classify a new object, U, with features u_1, u_2, …, u_n:
(a) Calculate the distance of U from each sample $X^{(i)}$ in the training dataset.
(b) Find the K nearest neighbors of U.
(c) Find the most common label among these K nearest neighbors.
(d) U is classified accordingly to this most common label.

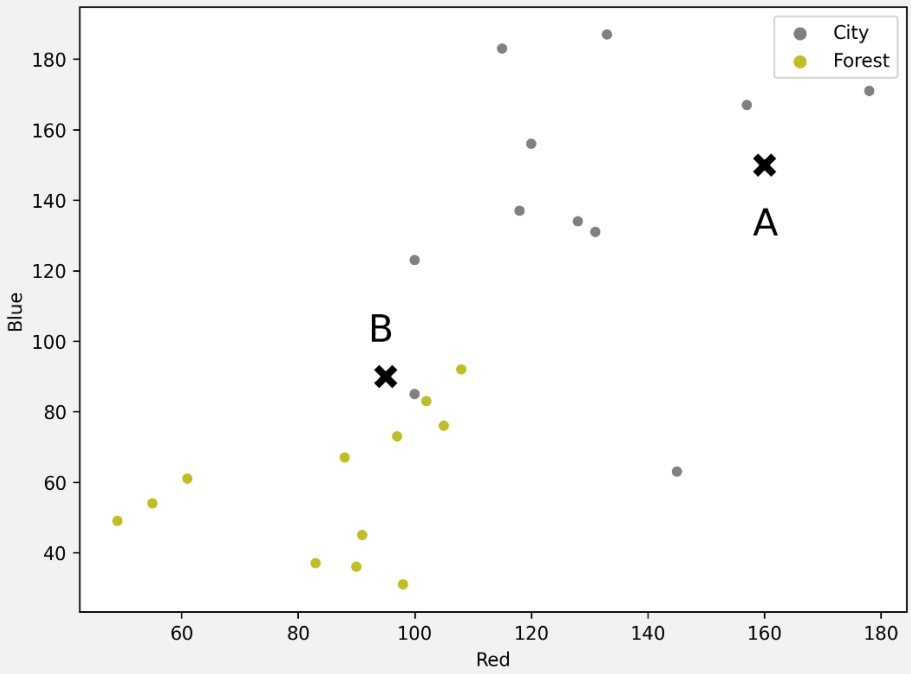
A common task for practicing the KNN algorithm is naturally a classification task of an unknown object based on a training set. One such exercise is the classification of Iris flowers based

on four features of the flower: sepal length (SL), sepal width (SW), petal length (PL), and petal width (PW), using a well-known training dataset containing 150 examples of three types of Iris flowers (Fisher, 1936). But, even with a small dataset of 150 flowers, this task might be too tedious for any learner to execute manually, and the calculation is, therefore, performed by a computer. Simulating the KNN classification process with a computer, however, fail to produce the ritual required to support the learning of mathematical concepts, first as a process and then as an object, and so, we designed a worksheet that guides learners to simulate the KNN algorithm manually and perform the required calculations themselves (see Figure 4.24 and Figure 4.25).

This worksheet guides a ritual execution of the mathematical process of classification using KNN and has two phases:

- The first part (Figure 4.24) requires visual identification of nearest neighbors in a two-dimensional case. It simulates the algorithm's search for the K nearest neighbors of a new instance and its classification according to the majority class. The given dataset was drawn in a way that enables to observe the classification easily without needing to actually calculate the distances.
- In the second part (Figure 4.25), the students are asked to manually calculate the Pythagoras distances in the four-dimensional case for the Iris flower dataset, to find the nearest neighbors, and to classify a new instance according to the majority class. Since the given dataset is small (6 samples of flowers), the students can perform the calculations manually without using a computer.

A KNN algorithm is designed to classify images into two types: urban or forest, based on two features: Red - the mean level of the red color in the image, and Blue - the mean level of the blue color in the image.
The following graph presents the training dataset images (urban images in gray and forest images in green). The graph also shows two unknown images, A and B. Classify images A and B using a KNN algorithm, with K=1 and K=3.



1. For K=1:
   a. The classification of image A is:
      Explain your answer: _____
   b. The classification of image B is:
      Explain your answer: _____
2. For K=3:
   a. The classification of image A is:
      Explain your answer: _____
   b. The classification of image B is:
      Explain your answer: _____

Figure 4.24 - The KNN algorithm worksheet (Part 1)

A KNN algorithm is designed to classify Iris flowers into two classes: setosa and versicolor. Four features are given for each flower: sepal length (SL), sepal width (SW), petal length (PL), and petal width (PW). There are six samples of flowers in the training set.

A researcher found a new flower, U, with the following features: U_{SL}=5, U_{SW}=3, U_{PL}=2, U_{PW}=2.

1. Calculate the distance of the new flower from each sample in the training set:

Sample	**sepal length (SL)**	**sepal width (SW)**	**petal length (PL)**	**petal length (PW)**	**Label**	**$d^{(i)}$**
1	5.1	3.5	1.4	0.2	Setosa	
2	4.9	3	1.4	0.2	Setosa	
3	4.7	3.2	1.3	0.2	Setosa	
4	7	3.2	4.7	1.4	Versicolor	
5	6.4	3.2	4.5	1.5	Versicolor	
6	6.9	3.1	4.9	1.5	Versicolor	

2. Classify this new flower using the KNN algorithm with K=1 and with K=3.
   For K=1:
   a. The indexes of the K closest training examples:
   b. The labels of the K closest training examples:
   c. The final classification is:
   For K=3:
   a. The indexes of the K closest training examples:
   b. The labels of the K closest training examples:
   c. The final classification is:

Figure 4.25 - The KNN algorithm worksheet (Part 2)

4.5.5. The KNN comprehension experiment

To evaluate non-major data science students white box understanding of ML algorithms, we designed an experiment to measure the comprehension of the KNN algorithm. The experiment involved a learning module on the KNN algorithm that included a recorded lecture and the hands-on task. The lecture and the task are based on the following principles, borrowed from mathematical education:

- In the recorded lecture, the students learn all mathematical details of the KNN algorithm, using only mathematical objects the students are familiar with, that is, calculating Euclidian distance using addition, subtraction, squaring, and taking square roots, and finding the minimum of the list of distances, using visual or a table representation, but without using or writing a computer program. The recorded lecture in Hebrew can be found here.
- In the hands-on activity, the students manually simulate the operation of the KNN algorithm and calculate the required operations. This task allows the students to execute the algorithm ritually as a process (see Figure 4.25).

The learning module was integrated into the Data Science for Psychological Science course in the 2021 Spring semester (Case 19 in Table 3.1 - Research Field). To measure the learning outcomes, we designed a KNN comprehension questionnaire (see Figure 4.26). This questionnaire was administered to the students after watching the recorded lecture about the KNN algorithm and completing the KNN hands-on activity. Thus, this questionnaire measured the students' accumulated learning of the KNN algorithm from both the lecture and the hands-on task.

The research population included 23 students. Gender was balanced (12 man and 11 women). Approximately 75% of the students were graduate (MA) students and 25% were PhD students. Approximately half of the students were studying and researching social psychology and the other half were studying and researching cognitive psychology. Two students were studying neural sciences.

An introductory questionnaire was administered to the students before the onset of the semester to collect data on the participants' motivation to study the course and on their previous knowledge. Seventeen students answered this questionnaire. Student motivation was measured by a closed question based on our previous study on the learning of ML by social science and digital humanities graduate students and researchers (see Case study 3 - Bridging gender gaps in STEM

education). Ninety-four percent of the students indicated that they were interested in learning new research tools and 77% said they were interested in learning Python.

With respect to students' previous knowledge, as expected, the students evaluated their background in statistics as being higher than their knowledge in computer science and ML, and their initial knowledge regarding ML as being lower relative to all other topics mentioned in the questionnaire (see Table 4.18).

Table 4.18 - Students' self-evaluation of prior knowledge (1=low, 5=high) (n=17)

Topic	Mean (Variance)
Programing in any language	3.9 (0.6)
Python programing	2.9 (0.9)
Descriptive statistics	4.2 (0.9)
Statistical inference	4.2 (1.0)
Linear regression	4.1 (1.4)
Machine learning	2.3 (1.0)

4.5.6. The KNN comprehension questionnaire

As presented in detail in Section 4.5.2, theories in mathematics education research distinguish between a process and object conception of mathematical concepts (Sfard, 1991). A *process* conception of a mathematical concept is reflected by the ability to follow the calculation of the output of the concept for a specific input. An *object* conception of a mathematical concept, on the other hand, is reflected by the ability to examine it according to its properties, without necessarily checking to see how it works. Similarly, a *ritual* solution of a mathematical task is characterized by meticulously following the entire process of the solution, step by step, while an *explorative* solution of a mathematical task is characterized by the flexible and goal-oriented execution of previously learned rituals that have been conceptualized as objects and can, therefore, be manipulated and examined according to their properties (Lavie et al., 2019).

Accordingly, we characterized students' understanding of the KNN algorithm as a process and as an object as follows:

Students' conception of the KNN algorithm as a *process* in problem-solving processes, whose topic was the KNN algorithm, is exhibited when students specifically address at least one

of the three *steps* required for classification of an unseen sample (P1-P3 below) and/or the step of tuning the K hyperparameter in the performance optimization stage (P4 below).

(P1) Calculate distance from all samples
(P2) Pick the K nearest samples
(P3) Find the label of the majority
(P4) Tune the hyperparameter K to improve performance (to avoid underfitting and overfitting)

Students' conception of the KNN algorithm as an *object* in problem-solving processes, whose topic was the KNN algorithm, is exhibited when students examine the following seven *properties* of the KNN algorithm – the first three (O1-O3) address the classification, the fourth (O4) addresses the algorithm's performance and the last three (O5-O7) address the algorithm's complexity (determined by the number of distance calculations):

(O1) Classification depends on similarity
(O2) Classification is determined by distance
(O3) The classification of a specific unknown example depends on K
(O4) The performance of the KNN algorithm for a specific K depends on the distribution of the data
(O5) The number of distance calculations depends on the number of training samples
(O6) The number of distance calculations depends on the number of features
(O7) The number of distance calculations does not depend on K

To examine students' conception of the KNN algorithm, we used the KNN comprehension questionnaire (see Figure 4.26). Each question enabled us to reveal the students' conception of the KNN algorithm as either a process or an object or both, by examining the entity(ies) the students used in their answers: the four process steps or the seven object properties of the KNN algorithm, respectively.

In Question 1, the students are asked to explain the KNN algorithm to a friend. The aim of the request to explain the algorithm to a friend instead of simply defining it is to guide the students *not* to recall and repeat the KNN definition they learned in the lecture, but rather, to explain it in their own words. The formulation of this question also implies that the explanation should avoid jargon, but rather, should use only simple and understandable language.

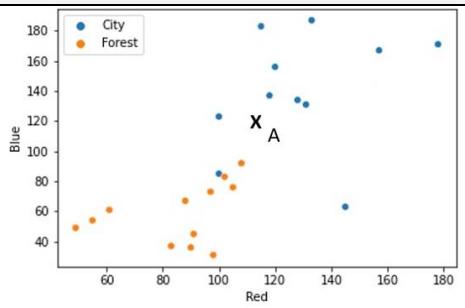
In Questions 2 and 3, the students are asked to indicate whether each of six statements on the prediction of the KNN algorithm, made by six hypothetical students (Alice, Bob, Carol, Dave, Eve, and Frank) is correct or incorrect, and to speculate how the hypothetical student explained his or her answer. These are critical thinking questions, which represent a higher order of thinking based on Bloom's taxonomy (Anderson et al., 2001; Bloom et al., 1956).

In Question 4, the students are asked to examine the computational complexity of the KNN algorithm. While for 4(a) and 4(b), a process conception of the KNN algorithm is sufficient, an object conception is required in order to draw conclusions regarding its complexity in different situations (in 4(c), 4(d) and 4(e)). We note that graduate psychological students are not familiar with the concept of computational complexity and that it is not part of the course curriculum.

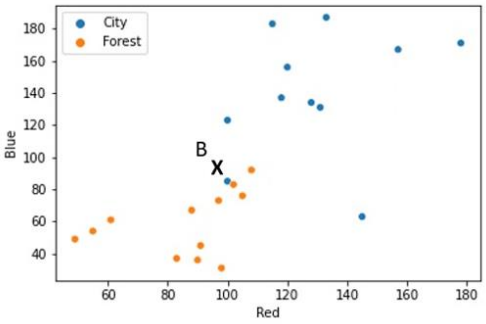
Specifically, in Question 4, the students are asked to indicate, for K=5 and K=11, how many times the square operator has to be calculated in a specific classification problem using a KNN algorithm. Although the K values are different, the square operator must be calculated to find the Euclidian distance between the unknown instance and each instance in the training examples. In other words, 4,000 calculations are required in both cases (4 features *1000 instances in the training set), regardless of the value of K.

1. How would you explain to a friend what the KNN algorithm is?

2. Students were asked to classify Example A in the figure on the right, using the KNN algorithm for K=5.
Alice claims that A's classification is Forest.
a. Is Alice right?
b. In your opinion, how did Alice explain her answer?
Bob claims that A's classification is City.
c. Is Bob right?
d. In your opinion, how did Bob explain his answer?



3. Students were asked to classify Example B in the figure on the right, using the KNN algorithm.
Carol claims that for K=5, B's classification is City.
a. Is Carol right?
b. In your opinion, how did Carol explain her answer?
Dave claims that for K=5, B's classification is Forest.
c. Is Dave right?
d. In your opinion, how did Dave explain his answer?
Eve claims that for any K, B's classification is City.
e. Is Eve right?
f. In your opinion, how did Eve explain her answer?
Frank claims that for any K, B's classification is Forest.
g. Is Frank right?
h. In your opinion, how did Frank explain his answer?



4. In order to classify dogs as Poodle or Labrador, four features were selected: height, weight, tail length, and ear length. The training set included 1,000 dogs, 500 of each kind. Based on this data set, we wish to classify an unknown dog using the KNN classifier.
a. For K=5: How many times is the square operation executed?
b. For K=11: How many times is the square operation executed?
c. What conclusion can you draw from your answers to the above two questions?
d. In your opinion, when are the chances of a correct classification higher?
   - K=5
   - K=11
   - It is impossible to decide
   - I do not know
e. Explain your answer.

Figure 4.26 - KNN comprehension questionnaire

Table 4.19 - Mapping the questions of the comprehension questionnaire according to the KNN process steps and object properties each question elicits

		Question			
Type of conception	**Expression of conception**	**1**	**2**	**3**	**4**
Process conception	(P1) Calculate distance from all samples	X			X
	(P2) Pick the K nearest samples	X	X	X	
	(P3) Find the label of the majority	X	X	X	
	(P4) Tune the hyper parameter K to improve performance				X
Object conception	(O1) Classification depends on similarity	X			
	(O2) Classification is determined by distance	X			
	(O3) The classification of a specific unknown example depends on K	X		X	
	(O4) The performance of the KNN algorithm for a specific K depends on the distribution of the data				X
	(O5) The number of distance calculations depends on the number of training samples				X
	(O6) The number of distance calculations depends on number of features				X
	(O7) The number of distance calculations does not depend on K				X

4.5.7. Data analysis methods

Students' answers to all questions in the KNN comprehension questionnaire (see Figure 4.26) were analyzed to extract the students' conception of the KNN algorithm as either a process or an object or both. Specifically, for each student explanation provided in the comprehension questionnaire, we examined what process steps and/or object properties it includes and determined the student's conception accordingly. The coding scheme we used is presented and illustrated in d Table 4.20.

Table 4.20 - Coding scheme of answers to the KNN comprehension questionnaire

Type of conception	Expression of conception	Coding scheme	Example
Process conception	(P1) Calculate distance from all samples	The text specifies the operation of calculating the distance from all training samples.	*"The algorithm checks the distance to all examples."* (Answer to Q.4)
	(P2) Pick the K nearest samples	The text specifies the operation of picking the K nearest neighbors.	*"If, for example, we defined that the number of neighbors (K) is five - the algorithm will check which five examples are the nearest."* (Answer to Q.1) *"Of the five images closest (most similar) to the new image, most (4) are a city."* (Answer to Q.2)
	(P3) Find the label of the majority	The text specifies the operation of choosing the label of the majority of samples.	*"… and according to the majority, will determine which type it belongs to."* (Answer to Q.1) *"… most (4) are a city."* (Answer to Q.2)
	(P4) Tune the hyperparameter K to improve performance	The text specifies the operation of tuning K to maximize performance.	*"It very much depends on the data; you should check the cross validation and decide accordingly. There is no one correct K."* (Answer to Q.4)
Object conception	(O1) Classification depends on similarity	The text specifies that classification is based on similarity.	*"[the decision] will be determined … by proximity."* (Answer to Q.1)
	(O2) Classification is determined by distance	The text specifies that classification is based on distance.	*"The similarity is expressed by a quantitative distance (squared) from the existing data."* (Answer to Q.1)
	(O3) The classification of a specific unknown example depends on K	The text specifies that different K may lead to different classifications.	*"[the decision] will be determined by the number of nearest items we defined (neighbors)."* (Answer to Q.1) *"He [Frank] missed k=1 and looked only at larger Ks."* (Answer to Q.3)
	(O4) The performance of the KNN algorithm for a specific K depends on the distribution of the data	The text mentions the property that accuracy depends on the properties of the data.	*"You need to know how the data is distributed."* (Answer to Q.4)
	(O5) The number of distance calculations depends on the number of training samples	The text mentions the property that the number of operations depends on the number of training samples.	*"… the number of square operations depends on the number of samples and not on the number of neighbors."* (Answer to Q.4)
	(O6) The number of distance calculations depends on number of features	The text mentions the property that the number of operations depends on the number of features.	*"You need to calculate 1000 distances, each distance has 4 characteristics which is 4000 square operations."* (Answer to Q.4)
	(O7) The number of distance calculations does not depend on K	The text mentions the property that the number of operations does not depend on the number of neighbors.	*"… the number of square operations depends on the number of samples and not on the number of neighbors."* (Answer to Q.4)

4.5.8. Results

In this section, I present the students' conception of the KNN algorithm from the perspective of the process-object duality. The results are analyzed from two perspectives: first from the student point of view (Section 4.5.8.1) and then from the algorithmic perspective (Section 4.5.8.2). Specifically, in Section 4.5.8.1, answers are categorized according to students' conception of the KNN algorithm as either a process, object, or procept. Then, in Section 4.5.8.2, I delve into the details, analyzing what process steps of object properties the students addressed in their answers and what we can learn from this analysis.

4.5.8.1. Students' conception from the process-object duality perspective

Table 4.21 presents the students coded answers to the KNN comprehension questionnaire (Figure 4.26). A cell in the table is colored with green if the participant answers to the KNN comprehension questionnaire indicates the comprehension process steps (P1-P4) or the object properties (O1-O7) according to the coding scheme (Table 4.20). Each process steps and object properties are presented in a different row. Each participant coded answers are presented in a different column. Participants who revealed comprehension of more than half of the process steps and the object properties are colored with dark green. Reviewing the students' coded answers (see Table 4.21) revealed that the students' conception of the KNN algorithm can be classified as follow:

- Process conception: About half of the students (5 out of 12, light green in Table 4.21, participants 1, 2, 5, 6, 8) conceived the KNN algorithm as a process, using in their explanations at least half of its process steps but less than half of its object properties. Clearly, other criteria could be used to determine this conception, but for the sake of simplicity, we set this simple criterion.
- Procept conception: About half of the students (7 out of 12, dark green in Table 4.21, participants 3, 4, 7, 9, 10, 11, 12) conceived the KNN algorithm as a procept, using in their explanations at least half of its process steps and at least half of its object properties. Again, we used this criterion for the sake of simplicity.
- Object conception: None of the students used object properties exclusively in their explanations, and accordingly, none of the students' conception was classified as an object.

Table 4.21 - Coded answers

			Student number											
Type of conception	**Expression of conception**	**Total**	1	2	3	4	5	6	7	8	9	10	11	12
Process conception	(P1) Calculate distance from all samples	6												
	(P2) Pick the K nearest samples	12												
	(P3) Find the label of the majority	12												
	(P4) Tune the hyperparameter K to improve performance	3												
Object conception	(O1) Classification depends on similarity	4												
	(O2) Classification is determined by distance	11												
	(O3) The classification of a specific unknown example depends on K	4												
	(O4) The performance of the KNN algorithm for specific K depends on the distribution of the data	6												
	(O5) The number of distance calculations depends on the number of training samples	6												
	(O6) The number of distance calculations depends on the number of features	6												
	(O7) The number of distance calculations does not depend on K	6												

4.5.8.2. Process steps and object properties conception

We present three observations about the students' conception of each process step and object property, by their prevalence:

- *Similarity*: The students best conceptualized the steps of *picking the K nearest samples* (P2), *finding the label of the majority* (P3), and the property *classification is determined by distance* (O2). This result may be explained by the facts that a) steps (P2) and (P3) are the most practiced steps in the hands-on task (Figure 4.24 and Figure 4.25) and b) this understanding of (P2) and (P3) supports their understanding of (O2). This may explain the fact that all students gave correct answers to all of the closed items of Questions 2 and 3 (See Figure 4.26, Questions 2 a, c and Question 3 a, c, e, g). We note that only one third of the students mentioned that

classification depends on similarity (O1). This can be explained by the fact that students used the terms *similarity* and *distance* interchangeably, due to the graphic representation in which similarity and distance are visually the same. Accordingly, in teaching processes of the KNN algorithm it is important to include also not-graphical questions in which similarity is not identical to distance.

- *Complexity:* Eight students mentioned at least one of the properties: *The number of distance calculations depends on the number of training samples* (O5), *the number of distance calculations depends on the number of features* (O6), and *the number of distance calculations does not depend on K* (O7). Three students consistently mentioned all three properties (students 10, 11 and 12), and five students mentioned one or two of them. This is consistent with the fact that half of the students conceptualize the first process step *calculate the distance from all samples* (P1). Indeed, all students who mentioned this process step also mentioned at least two properties regarding complexity (number of calculations). This result might be due to the technique of solving visual tasks, i.e., Part 1 of the KNN hands-on activity (Figure 4.24) and Questions 2 and 3 on the KNN comprehension questionnaire (Figure 4.26). In these cases, the human mind can find the neighbors instantly, using only its visual processing capabilities, without needing to calculate the distances from all samples in the training set numerically.
- *K hyperparameter:* K as a hyperparameter can reflect either process conception (P4) or an object conception (O3 and O4), pending on its specific role in the description. One third of the students mentioned the property that the *classification of a specific unknown example depends on K* (O3) and one half of the students mentioned the property that the *performance of the KNN algorithm for specific K depends on the distribution of the data* (O4). There is partial overlap (of 2 students: participants 3 and 11) between the students who mentioned these properties regarding K and the students who mentioned the performance optimization phase in which the K hyperparameter is tuned (P4). This might indicate that more students conceptualized the role of K correctly but did not mention the tuning step, as the question (See Figure 4.26, Questions 4) formulation asked them merely to compare the possible performance of two different Ks, and not to describe how to maximize the algorithm performance. This explanation is also supported by the fact that only one quarter of the students mentioned the step of tuning the hyperparameter K to optimize the algorithm performance (P4).

4.5.9. <u>Discussion</u>

From a wider perspective, the process-object duality of the mental representation of mathematical concepts is associated with the phenomenon of reducing the abstraction level when learning abstract mathematical concepts (Hazzan, 1999). In general, students who need to learn mathematical concepts that are too abstract for their current mental representation, i.e., their conception, of the concepts, use several mechanisms to reduce the level of abstraction. As presented in this Case study, one of these is based on the dual process-object representation of mathematical concepts, according to which students conceive abstract concepts that are too abstract as processes or procepts rather than as objects, which are considered a more abstract conception,. This statement has been generalized to include the learning of computer science concepts as well (Hazzan, 2003b, 2003a; Hazzan & Hadar, 2005), and so the theory of process-object representation of mathematical concepts can be used also to describe students' understanding of ML algorithms on different levels of abstraction.

Within this context, the hands-on ritual-based tasks presented in this case study may support students' learning by reducing the level of abstraction (in intermediate learning stages) in several ways:

- Data have specific features that are more concrete (less abstract) than abstract features; images with features "red" and "green" are more concrete than objects with features "x_1" and "x_2".
- Data have specific values that are more concrete than abstract values; an image with features [107, 83, city] is more concrete than an object with features [x_1, x_2, x_3].
- Data are more concrete than just numbers, as they have meaning in the real world; "the level of red in the image is 107" is more concrete than the fact "the value of feature x_1 is 107".
- A small training dataset that learners can iterate manually is more concrete than the huge datasets typically used to demonstrate ML algorithms.
- The manual execution of a task is more concrete than running a simulation, such as a computer program, because attention must be paid to every detail, regardless of its role in the procedure.

Thus, the hands-on activities presented in this case study support learning not only by allowing the students to practice a ritual involving the underlying mathematical concepts, but also by intentionally reducing the level of abstraction of ML algorithms in intermediate learning stages.

Reducing the level of abstraction, however, should be done very carefully, since reducing the level of abstraction too much, may lead learners to conceive the specific as the general case

and this, in turn, may lead to a process or a procept conception. For example, the graphical representation of the KNN, as described in the first part of the KNN hands-on task (see Figure 4.24) in which the students can find the neighbors using visual capabilities without needing to calculate the distance from *all* samples in the training set (since the visual cortex does not really need to calculate each distances), might lead to partial conceptualization of the complexity of the KNN algorithm (see Section 4.5.8). Another example is the small number of training samples in the second part of the KNN hands-on task (see Figure 4.25), in which the students had to calculate the distances of a new example from small a training data set (of only six samples), and so the first step of the KNN algorithm, calculating the distance from *all* the training samples, was not sufficiently clear.

From the abstraction perspective, the three categories of process steps and object properties presented in Section 4.5.8 - *Similarity, Complexity, and K as hyperparameter* - can be organized by the level of abstraction they represent. *Similarity* may be explained visually and, thus, requires the lowest level of abstraction. *Complexity* may need to be explained using large tables and, thus, requires a higher level of abstraction, and *K as a hyperparameter* requires that the classification algorithm be conceived as an object in order to optimize performance, therefore, requiring an even higher level of abstraction.

The above discussion highlights both the practical-pedagogical and theoretical contributions of this case study. Practically, we present a learning module that supports non- major data science students' white box understanding of ML (at least as a process). Theoretically, we introduce a data analysis method to evaluate students' conceptions of ML algorithms.

Combining these two facets of contribution presented in the previous paragraph may guide the teaching process of ML algorithms. For example, a teacher who notices that the majority of his or her class conceives specific ML algorithms as a process, should consider whether more advanced algorithms should be introduced to the class or alternatively, and most probably, additional practice of the specific algorithms is required. An illustrative example for this assertion is the perceptron and neural networks. A teacher should verify that the class conceptualizes the perceptron algorithm as an object before moving forward using it as a building block of a neural network.

4.5.10. Conclusion

This case study demonstrates the challenge *lack of mathematical knowledge*. Ignoring this mathematical gap results in insufficient understanding of core concepts of machine learning and data science. This gap in mathematical knowledge of non-major data science learners in not discussed in the context of mathematics education nor in the context of computer science education. Moreover, this case study presents that an interdisciplinary approach, using theories from mathematical education to solve pedagogical challenges in machine learning education, can be beneficial for the learners. Such an interdisciplinary approach can be attained with data science education.

The term *interdisciplinary pedagogy* is discussed in the literature mostly as a synonym to interdisciplinary education, for example see Cargill, 2005 and Penny, 2009. Others define interdisciplinary pedagogy as an integrative pedagogical approach based on pedagogies of different disciplines (Chesley et al., 2018). In the spirit of interdisciplinary *education*, defined in section 2.1.4 as understanding the interconnections between the disciplines and being able to solve problems that require applying knowledge and methods from each discipline, *interdisciplinary pedagogy* can be defined as *solving pedagogical challenges in interdisciplinary education based on the integration of pedagogical methods from each of the disciplines*. This case study demonstrates the opportunity of *interdisciplinary pedagogy*, using theories and practices borrowed from mathematical education to support learning of ML algorithms.

4.6. Case study 6 - The pedagogical chasm

This case study describes the adoption of a new data science curricula developed for high school computer science pupils by high school teachers (Cases 4-10,15,16 in Table 3.1). As an interdisciplinary profession, data science poses many challenges for teachers. This case study describes our experience with high school computer science teachers and shows that most teachers do not tend to adopt the Israeli data science curriculum.

In this case study I analyze this phenomenon from the perspective of the diffusion of innovation and crossing-the-chasm theories. I first present the diffusion of innovation theory (Section 4.6.1) and crossing-the-chasm theory (Section 0). Next, I present the data science for high school curriculum case study (Section 4.6.3) and the data collected from the teachers (Section 4.6.4). Data analysis reveals that in the case of adoption of new curriculum there is a *pedagogical chasm* (e.g., a pedagogical challenge that reduces the motivation of the majority of the teachers to adopt the curriculum) that slows the adoption of innovation process (Section 4.6.5). Finally, I discuss the results of this case study in the perspective of the islands of innovation in education phenomena (Section 4.6.6) and the implications of the pedagogical chasm to data science education (Section 4.6.7).

4.6.1. Diffusion of Innovation theory

The diffusion of innovation theory was proposed by Rogers based on a study of the diffusion of agriculture technologies in rural farms in the United States about 100 years ago (Rogers, 1962). Nevertheless, this theory is still used today to explain the diffusion of current technologies and innovations. According to this theory, innovations spread in society by flowing from one of five distinct groups of adopters to the next:

Innovators: This group is characterized by enthusiasm for innovation. Its members value the innovativeness of the product and are the first to understand the value of the new technology over existing ones. Innovators are willing to embrace innovation even if a lot of effort is required to use it due to its premature nature.

Early adopters: Early adopters are ready to take risks and embrace innovation at the early stages of innovation maturation. This group is characterized by the ability to quickly identify the added value of the innovation and harness it in favor of a larger vision. The early adopters are not looking for improvement, but rather, promote a vision that is a breakthrough. Innovation for them

is a means of realizing the vision and its importance is derived from its ability to contribute to the fulfillment of the vision it delivers.

Early majority: This group consists of pragmatic people who are interested in adopting mature innovation to improve their performance. This group attempts to reduce costs and is therefore interested in standardization. The early majority is looking to adopt mature and reliable products that include full support.

Late majority: This group consists of conservatives who are fundamentally opposed to innovation and prefer the existing and familiar products.

Laggards: Members of this group oppose innovation, are skeptical about its benefits, and pose difficult questions to the innovators. The laggards are constantly pointing out the gaps between the promises claimed by the new technology and its realization.

It is common to display the five groups on an innovation-adoption timeline (see Figure 4.27). The blue line represents the rate of adoption of each group and the yellow line represents the accumulative adoption of the innovation as time progresses.

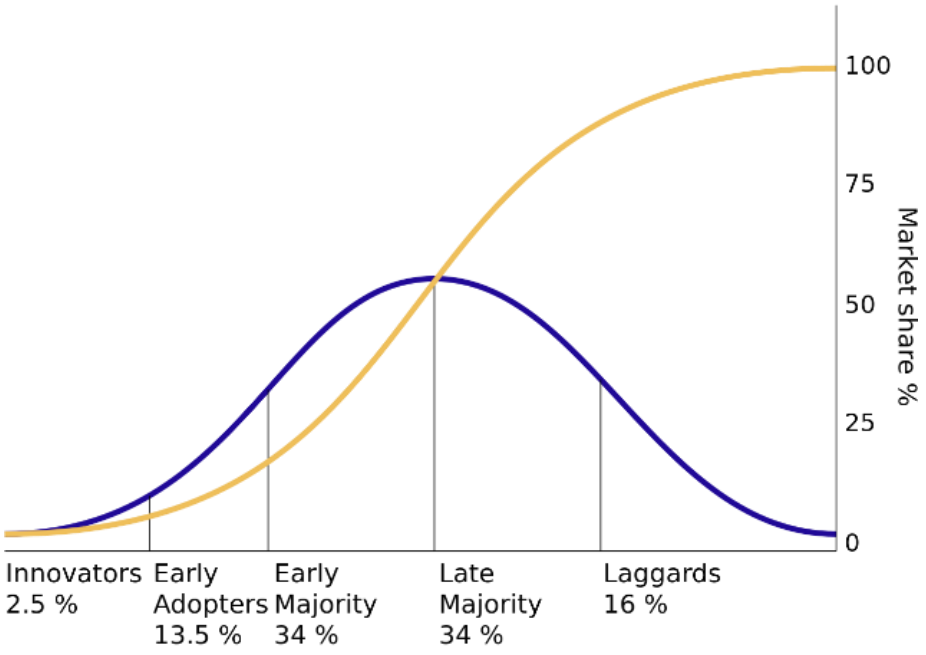


Figure 4.27 - Diffusion of Innovation Timeline (Rogers, 1962)

4.6.2. Crossing the Chasm theory

While some innovations diffuse according to the diffusion of innovation theory, the diffusion rate of other innovations drops before they reach a high level of acceptance. According to the crossing-the-chasm theory, this effect can be caused by discontinuity of the flow of innovation between the *early market* (Innovators and Early Adopters) and the *mainstream market* (Early and Late Majority) (Moore, 2002). Since these two groups (early market and the mainstream market) have different adoption characteristics, there is a *chasm* between them. This chasm stems from the fact that innovation mainstream market adopters rely on recommendations and the adoption experience of other adopters in the same market, and not on the adoption experience of the early market. Mainstream market adopters simply understand that early market adopters are willing to adopt innovations that are not mature enough for them. The chasm causes a reduction in the adoption rate of the innovation after it has been adopted by the early market and leads to its failure to penetrate the central market.

Moore (2002) defines two types of innovation: continuous innovation and disruptive innovation. Continuous innovation is defined as innovation that aims to upgrade existing capabilities of a product without requiring the adopters to change their behavior (e.g., replacing a mobile phone keyboard with a touch screen). Disruptive innovation is defined as innovation that requires the adopters to change their behavior in order to use it (e.g., replacing a mobile phone keyboard with a voice-operated virtual assistant). While early market adopters are willing to adopt disruptive innovations, central market adopters are only willing to adopt continuous innovations. The central market will adopt a disruptive innovation when it reaches a level of maturity, and its compatibility allows the members of the central market to adopt it with no substantial change in the way they operate.

4.6.3. The data science curriculum case study from the diffusion of innovation perspective

A three-year program in data science and machine learning designed for 10th to 12th grade computer science pupils was approved in Israel in 2019 after three years of pilot implementation (Mike, Hazan and Hazzan, 2020). The program was integrated into the current, official Israeli high school computer science curriculum, which has been updated on a regular basis since 1998, when it was first implemented (Gal-Ezer et al., 1995). These updates include both new content as well

as adaptations for new populations, targeting younger audiences, mainly elementary and middle school pupils.

Similar to the general high school computer science programs taught in Israel, the data science program is designed to accommodate two levels. The basic level, which is taught in the 10th grade, focuses on data science workflow and machine learning (see Table 4.22). The extended level is taught in the 11th and 12th grades and focuses on deep learning. Pupils at both levels develop a machine/deep learning project in Python that includes defining the machine learning problem, searching for and collecting data, data exploration, training, evaluation and comparison of several machine learning algorithms, and communication of the results.

The story of the new program started in 2017, when Prof. Tamir Hazan[11], from the Faculty of Industrial Engineering and Management at the Technion, began teaching machine learning as part of the Technion's data science undergraduate program. That year, he also developed and taught a machine learning course to his daughter's 9th grade class. Following his successful teaching experience in school, he presented the program to several groups of teachers, trying to "market" them the program.

Table 4.22 - Data science for 10th grade - Topics and number of hours

Topic	**Hours**
Introduction to data science	3
Introduction to Python	18
Tables with Pandas	6
Exploratory data analysis and visualization	6
Introduction to machine learning	3
The K-nearest neighbors (KNN) algorithm	6
Performance measurements	3
Overfit and underfit	3
Feature engineering	3
Practical development of machine learning project	3
Image as data	3
Introduction to modern machine learning algorithms	3
Final project development	30
Total	**90**

[11] Full name appears with permission.

Even though many teachers heard about this curriculum and started realizing the importance of teaching data science in school, almost none of them adopted it. Clearly, the program was still premature for mass adoption. Nevertheless, the program was adopted by a few teachers, early adopters, who were willing not only to learn the new content knowledge, but also to develop the new pedagogies and teaching methods required to teach it. One of these teachers was Mara Kofler[12], a highly experienced teacher with a strong entrepreneurial sense. Mara assembled a team, recruited new teachers, and initiated implementation of the program in 2018. One of the teachers on her team was Koby Mike, the author of this thesis.

During the first two years of its implementation, the program was taught solely by Mara's team. Based on the experience gained in those two years, the curriculum was redesigned, adequate teaching methods were developed, and it was approved by the Israeli Ministry of Education as a pilot program. At this stage, the program was reintroduced to several groups of teachers. Again, only a few teachers adopted the new curriculum. One of them was Ariel Bar-Itzhak[13], another experienced teacher with an entrepreneurial sense. While teaching the curriculum, Ariel Bar-Itzhak invested many hours developing learning materials and instructional movies for the benefit of both teachers and pupils (see here). At the end of 2020-21 school year (June 2021), after appropriate pedagogy and teaching methods were developed, the curriculum was ready to be adopted by teachers in the mainstream market. Indeed, when the program was introduced for the third time to a group of teachers, about half of the group decided to adopt it and indeed began teaching it the following year. Table 4.23 presents the timeline of this adoption process.

[12] Full name appears with permission.
[13] Full name appears with permission.

Table 4.23 - The adoption process of the 10th grade data science curriculum

Period	Phases of adoption of innovation	Description
September 2017	Innovators	Prof. Hazan developed the program and taught it to his daughter's class.
July 2018	Innovators	Prof. Hazan presents the curriculum to group of leading teachers, including by Mara Kofler
September 2018 – June 2020	Innovators	The program was taught by Mara Kofler's team, including the development of a suitable pedagogy.
July 2020	CHASM	First exposure workshops were held for computer science teachers in which the new program was presented, but the chasm was not crossed. Nevertheless, several early adopters did adopt the innovation.
September 2020	Early adopters	Ariel Bar-Itzhak, one of the early adopters, started teaching the program, developing additional learning material and instructional movies.
July 2021	CHASM crossed	Second exposure workshops were held for computer science teachers in which the new program was presented alongside the teaching pedagogy developed for it.
September 2021	Early majority	The early majority started teaching the program in 20 schools.

4.6.4. Teachers' perspective

In this section I present the teachers' perspective of the challenges involved in the adoption of the new data science curriculum. Data was collected via questionnaires distributed to the teachers at the end of the second exposure workshops (July 2021). In their responses to open questions, the teachers were asked to share what, in their opinion, are the challenges *they* will face in teaching the curriculum, and what challenges, in their opinion, *their pupils* will face while learning it. The teachers' answers are presented according to the three components of teachers' knowledge according to the TPACK framework: pedagogical knowledge (Section 4.6.4.1), content

knowledge (Section 4.6.4.2), and technological knowledge (Section 4.6.4.3) (Mishra & Koehler, 2006). Twenty teachers answered the questionnaire and provided 56 assertions regarding the challenges may be encountered when teaching and learning the new curriculum.

4.6.4.1. Pedagogical Knowledge

About two thirds of the assertions addressed pedagogical challenges, and these were grouped into four categories: a) concepts that are difficult to teach and learn, b) motivation, c) pupils' need to work individually, and d) teaching aids and teachers' support. In the next section I discuss this big portion (two third) of teachers who mentioned pedagogical challenges. Here I present and illustrate the four categories of teachers' assertions that indicate a pedagogical challenge:

Within the 'concepts that are difficult to teach and learn' category, about half of the assertions declared that the curriculum includes many concept that are difficult for teachers to teach and for pupils to learn. Particularly, the teachers mentioned that many concepts require comprehensive mathematics and a solid theoretical background in machine learning and data analysis. For example, one teacher wrote that *"There is a lot of theory here - from my acquaintance with 10th grade pupils, they have no patience for it at this point. That comes later".*

About a quarter of the assertions mentioned the need to enhance pupils' motivation to learn some essential parts of the curriculum that were perceived less interesting. For example, one teacher wrote that *"(Pupils) interest in tabular data"* is a challenge. Another aspect of the curriculum is working on real problems with real data and, so, it can be challenging for the pupils to achieve high performance algorithms. For example, one teacher wrote that *"Unsatisfactory final results"* is a challenge.

About 15% of the assertions regarded the pupils' need to work individually on complex projects. for example, one teacher wrote that pupils need to *"know how to work alone, and look for answers to questions and databases"*.

About 10% of the assertions regarded the lack of learning materials for the new curriculum and the lack of teachers' support. For example, one teacher wrote that *"Arranged lesson plans for teachers with notes, examples, videos beyond advanced training*" is a challenge.

4.6.4.2. Content Knowledge

Approximately one quarter of the assertions addressed the required content knowledge which can be split into two categories: the variety of topics included in the curriculum and the required knowledge in computer science. For example, one teacher wrote that *"The program is difficult to teach, the field is very broad and contains a lot of concepts and a lot of cross-cutting knowledge"*.

4.6.4.3. Technological Knowledge

About 9% of the assertions considered technological challenges. The main technological challenge that the teachers noted is finding suitable datasets for teaching by the teachers and finding suitable datasets for projects to be developed by the pupils. For example, one teacher wrote that *"It is difficult to find good databases on a variety of topics"*. As it turns out, the use of real datasets and the use of dataset search engines is apparently a new activity/skill for most computer science teachers.

4.6.5. <u>The pedagogical chasm</u>

It can be concluded that while some innovative teachers are willing to adopt a new curriculum in its early phases of development, the majority of teachers are willing to adopt a new curriculum only after new pedagogies and teaching methods have been developed for its teaching.

The chasm theory is relevant especially in cases of disruptive innovation, which is, we claim, what the new data science curriculum is for computer science high school teachers. We base our claim on the required change (if the new data science curriculum is adopted) in the three components of teachers' knowledge defined by the TPACK framework: pedagogical knowledge, content knowledge, and technological knowledge (Mishra & Koehler, 2006):

- *The pedagogy component*: From a pedagogical perspective, a new curriculum is considered disruptive if it uses teaching methods that the teacher does not know how to employ. For example, a curriculum whose teaching is based on project-based learning (PBL) is a disruptive innovation, from the pedagogical perspective, for teachers who for many years have used only frontal teaching methods and have never before experienced any active-learning-based pedagogies (such as PBL).

- *The content component:* From a content perspective, a new curriculum can be considered disruptive if it is based on knowledge that teachers lack. For example, a curriculum that teaches the Python programming language is a disruptive innovation, from the content perspective, for teachers who have never before learned Python, as they will have to learn the new content before they can start teaching it.
- *The technology component*: From a technological perspective, a new curriculum is considered disruptive if it uses technology that requires teachers to change their behavior. For example, a curriculum taught in a virtual teaching environment is a disruptive innovation for teachers with no previous experience teaching in such an environment.

As it turns out, the data science curriculum for high school described above is disruptive in terms of all three dimensions:

- The technology component: Teachers must learn new technology, e.g., dataset search engines.
- The pedagogy dimension: Teachers must learn new pedagogies, e.g., using real data that were not generated by the teacher (Mike & Hazzan, 2021);
- The content component: Teachers must learn new content knowledge, e.g., data analysis.

Diffusion of curriculum is similar to diffusion of any other innovation in any other discipline with regard to the content component and the technology component, as any innovation may require users to learn new content or use new technology. The crossing-the-chasm theory can therefore predict the innovation diffusion to the early and mainstream markets based on its properties, such as continuous or disruptive. *Curriculum innovations are, however, different in terms of the pedagogical component*. While any disruptive innovation may require its adopters to learn new knowledge or to use new technology, curriculum innovations may require teachers to adopt also a new pedagogy, hence, crossing a *pedagogical chasm*. With respect to the previous case study (Section 4.5), which presented the opportunity of interdisciplinary pedagogy, it is worthwhile to note the in the case of data science education, the pedagogical chasm in included in the interdisciplinary pedagogical chasm. Therefore, it is not surprising that when asked to share what, in their opinion, are the challenges they and their pupils will face in teaching and learning the curriculum, two third of teachers' assertions indicated pedagogical challenges.

4.6.6. Discussion

Innovation is defined as "*a new idea, device, or method*" (merriam-webster.com). Educational innovation is the invention, implementation, and incorporation of new ideas and methods in educational settings. It is created both within and outside the formal education system by educational entrepreneurs and can manifest in a wide variety of ways, including new curricula, new teaching methods, new management learning methods, new teaching or learning technologies, and so on. While many educational projects are initiated to promote educational innovations, only few succeed and attain systemwide diffusion (Avidov-Ungar & Eshet-Alkakay, 2011; Hazzan & Zelig, 2016; Zelig, 2011). This phenomenon, called *islands of innovation*, describes a situation in which educational innovation fails to exhaust its potential, diffuses only to parts of the relevant educational system, and is implemented only in "islands". Not surprisingly, this phenomenon frustrates many educational entrepreneurs and educators. Fullan (2007) argues that despite many years of research, no satisfactory explanation has yet been given for the failure of many educational reforms to diffuse to the system level, even when they exhibit local success. The chasm theory, specifically with respect to the *pedagogical chasm*, may provide partial explanation for the phenomenon of islands of innovation in education.

This case study demonstrates the pedagogical chasm in the case of a curriculum; however, the pedagogical chasm explanation may be relevant for other types of educational innovation as well. For example, the adoption of a new learning management system may require new pedagogy and new teaching methods rather than conservative face-to-face teaching. In the development of an educational innovation that requires a pedagogical change, educational researchers, innovators, and entrepreneurs should, therefore, first, understand whether a pedagogical chasm exists and if it does, explore its essence; then, if needed, they should develop new pedagogies and teaching methods required to cross the pedagogical chasm so as to facilitate adoption of the innovation by the early majority. This recommendation is correct especially in the case of computer science education in which teachers are asked to adopt new curriculum frequently, and teaching materials and teachers' support are among the core components in successful curriculum implementation. (Hazzan et al., 2008)

4.6.7. Conclusion

In this case study teachers' conceptions about the challenges in teaching and learning of data science curriculum are analyzed according to the TPACK framework. The analysis demonstrates that more than 90% regarded pedagogical and content knowledge challenges while less than 10% of the statements regarded technological challenges. Therefore, this case study demonstrates the challenges of *teachers' gap in Content Knowledge (CK)* and *teachers' gap in Pedagogical Content Knowledge (PCK)*. This gap not only lower the level of teaching, but also slows the rate of adoption of a new data science curriculum. Therefore, to prepare teachers to teach data science, it is not sufficient to teach them separately the relevant content of mathematics, statistics, computer science, and application domains, but an interdisciplinary approach must be taken, both with respect to the content and to the pedagogy (see Case study 5). Clearly, such an approach cannot be developed separately in mathematics education, statistics education, computer science education or application domains education.

This case study also demonstrates the opportunity of *teachers' professional development*. Not only teachers can learn to teach new topics within their own discipline, but they can also learn new interdisciplinary topics and 21st century critical skills.

4.7. Conclusion

Sections 4.1-4.6 present six case studies of data science education that mitigate pedagogical challenges and leverage pedagogical opportunities of data science. As demonstrated, these collection of challenges and opportunities of data science education, presented in Table 4.1, are not mitigated or leveraged (respectively) in the educational context of any of the fields that create data science: mathematics education, statistics education, computer science education, and the education context of the application domains. Accordingly, my thesis claims that *it is essential to develop data science education as an interdisciplinary field of its own, relying on the principles of interdisciplinary education, and the integration of knowledge from the disciplines of mathematics education, computer science education, statistics education, and the application domain education*.

Table 4.24 summarizes the challenges of data science education with respect to the education fields of the disciplines generating data science.

Table 4.24 - Data science education challenges with respect to the education fields of the disciplines generating data science

Challenges	Case study (Section)	Math and Statistical education	Computer science education	Application domain education
Interdisciplinary project-based learning (PBL)	4.1			
Integrate the application domain knowledge	4.1			
Lack of statistical knowledge required for data science	4.2			
Teachers' lack of data science Content Knowledge (CK)	4.2, 4.6			
Teachers' lack of data science Pedagogical Content Knowledge (PCK)	4.2, 4.6			
Lack of computer science knowledge required for data science	4.3, 4.4			
Lack of mathematical knowledge required for data science	4.5			

In Table 4.24, for each challenge, a green shaded box denotes that this challenge *is* mitigated within the education field denoted by the column. A red shaded box denotes that this challenge *is not* mitigated within the education field denoted by the column. For example, the first challenge, *Interdisciplinary project-based learning,* is mitigated in the context of math and

statistical education (by the principles of statistical thinking), but is not mitigated in the contexts of computer science education and application domain education. The cell colors were determined according to the explanations presented in the conclusion sections of each case study. We note thought, that Table 4.24 should be further validated based on a more comprehensive literature review.

The thesis is also supported by the reports of the data science education committees presented in Section 2.2.1:

- In the final report of the NSF workshop on data science education held in 2015 titles "Strengthening Data Science Education Through Collaboration" it is written *"Infrastructure and culture of sharing of materials and experiences among the departments and schools that offer data science programs be supported and encouraged… Forms of collaboration might include web-based tools, conference(s), or a journal focused on data science education."* (Cassel & Topi, 2015, p. iV).
- In the Curriculum Guidelines for Undergraduate Programs in Data Science published in 2016 by the PCMI workshop on data science education it is stated that *"Data science is inherently interdisciplinary. Working with data requires … statistics, computer science, and mathematics. Data science blends much of the pedagogical content from all three disciplines, but it is neither the simple intersection, nor the superset of the three"* (De Veaux et al., 2017, p. 3).
- The EDISON Data Science Framework, a project sponsored by the European committee, states that *"The subject domain related knowledge … is recognized as essential for practical work of Data Scientist … These knowledge areas will be a subject for future development in tight cooperation with subject domain specialists."* (EDISON Data Science Framework (EDSF), 2017, Part 2, p. 12).
- A committee of the National Academies of Sciences wrote in the report Envisioning the Data Science Discipline that "*Effective application of data science to a domain requires knowledge of that domain*" (National Academies of Sciences et al., 2018, p. 29).
- Finally, in 2021 the ACM Data Science Task Force published its final report Computing Competencies for Undergraduate Data Science Curricula in 2021. It stated that *"Each component of the Data Science environment: the domain; statistics and mathematics; and*

computer science, is essential. However, a random collection of the three elements does not constitute a meaningful Data Science program." (Danyluk & Leidig, 2021, p. 10).

Nevertheless, my thesis adds to these reports evidence for their statements derived from an academic perspective. Further, as is summarized in the next chapter (Chapter 5), my thesis presents specific challenges and opportunities of data science education that such an interdisciplinary approach can support their mitigation / leverage respectively.

5. Discussion

Wing (2020) underlined the importance of data science education by asking, and answering: *"What will data science be in 10 or 50 years? The answer to this question is in the hands of the next-generation researchers and educators"*. As for today, however, data science education has not yet been recognized as a distinct field and is discussed mainly in the context of education in the disciplines that compose data science, i.e., computer science education, statistics education, mathematics education, and the educational context of the application domains, such as psychology education and business analytics education.

My thesis claims that a new field is about to be born – data science education – which is based on the educational disciplines associated with the disciplines that compose data science: computer science education, statistics education, mathematics education, and the educational context of the application domains, such as psychology education and business analytics education, as well as theories of interdisciplinary education and pedagogy (see Figure 5.1). In this section I discuss the following aspects of the field to be born of data science education: challenge and opportunities (Section 5.1), recommendations (Section 5.2) and future research topics (Section 5.3).

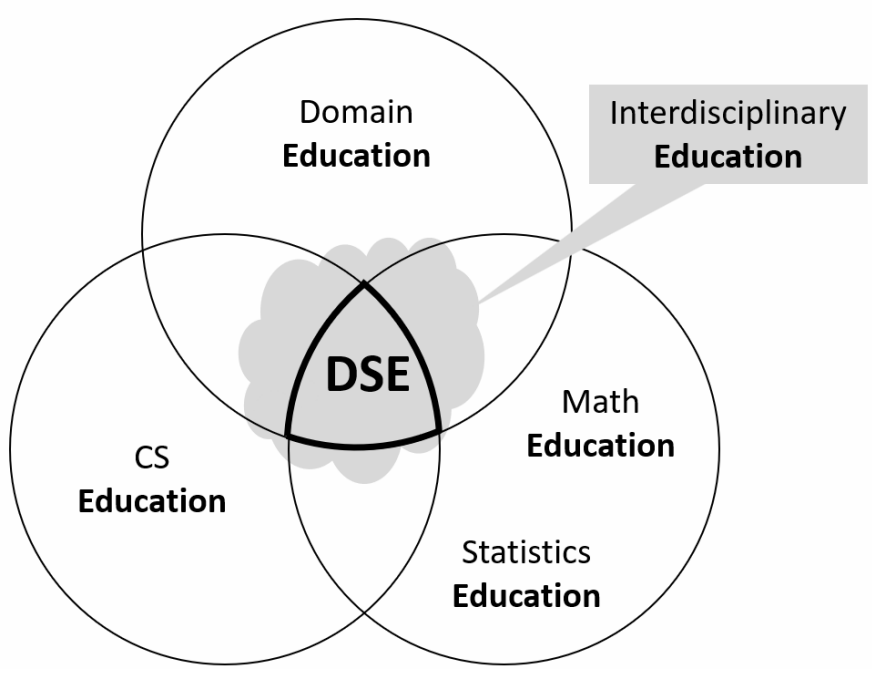


Figure 5.1 – Data science education

5.1. Data science education challenges and opportunities

In Chapter 4, I presented six case studies. Each case study demonstrates (one or more) challenges and opportunities of data science education. The case studies were presented by three categories of challenges: *cognitive, social* and *pedagogical* (see Table 4.1). In this section, I reorganize these challenges and opportunities in a framework according to the three characteristics of data science presented in Section 2.1.7: *interdisciplinarity, variety of learners* and m*ultifaced discipline*.

The challenges and opportunities belonging to the first two categories: *interdisciplinarity* and *variety of learners* are well demonstrated in the presented case studies. With respect to the third category, the fact that data science is a *multifaced discipline*, that can be presented as a science, as a research method, as a discipline, as a workflow and as a profession (see Section 2.1.7), also elevates challenges and opportunities for data science education. In my research, three challenges and one opportunity were discovered:

- Data science can be perceived as a research method, and therefore it requires *research skills*.
- Data science can be perceived as a profession, and therefore requires *professional (soft) skills*.
- Data science can be perceived as discipline, and therefore requires *ethics*.
- Data science is a multifaced disciplined integrating several important 21st century thinking skills and therefore data science education can leverage the opportunity to promote 21st century skills

These challenges and opportunity were discovered in my research, but so far not enough data have been collected to present them as a case study. This fact can be explained by their focus. First, unlike case studies 4.1 – 4.5 which focus on data science content knowledge, these challenges and opportunity focus on skills and social aspects, which can be taught only after content knowledge is mentally constructed (Hazzan & Lis-Hacohen, 2016). Second, unlike case study 4.6 which focuses on data science pedagogical content knowledge, which is in the core this research, these challenges and opportunity focus on other skills, which can be implemented not necessarily in educational contexts. Therefore, additional research is required regarding these challenges and opportunity from the perspective of data science education.

Table 5.1 – Data science education challenges and opportunities

Data science characteristics	Challenges	Opportunities
Interdisciplinary	Integration of the application domain knowledge	Teaching STEM with real-world data
	Teachers' gap in Content Knowledge (CK)	Professional development for teachers
	Teachers' gap in Pedagogical Content Knowledge (PCK)	Interdisciplinary pedagogy
	Project Based Learning (PBL)	Teaching STEM with real-world context
Learners' diversity	Lack of mathematical knowledge	Bridging gender gaps in STEM education
	Lack of computer science knowledge	
	Lack of statistical knowledge	Teaching statistical literacy
Multifaced discipline	Research skills	Teaching 21st century skills
	Professional (soft) skills	
	Ethics	

5.2. Recommendations

In the previous section, I presented a framework for the challenges and opportunities of data science education. In what follows, I present recommendations for the mitigation of the challenges and the leverage of the opportunities. The recommendations are presented by the three characteristics of data science (Sections 5.2.1, 5.2.2 and 5.2.3). For each characteristic, I first present recommendations for its associated challenges, followed by recommendations for its associated opportunities.

5.2.1. Interdisciplinarity challenges and opportunities

5.2.1.1. Challenges

Integration of the application domain knowledge

Educators and learners should be familiar with the domain neglect bias. The traffic light classification question (Table 4.5) and the carcinoma classification question (Table 4.6) can be used as homework or class activity to demonstrate this bias to learners. Educator should note when learners fail to integrate correctly the application domain knowledge in their learning processes in general and project development in particular, increase their awareness to this bias and correct them. Learners should be encouraged to practice data science concepts with data taken from domain they are either specialists in, familiar with, or at least interested in.

Teachers' gap in Content Knowledge (CK)

Data science teachers should be familiar with the Content Knowledge of the disciplines composing data science: mathematics, statistics, computer science and at least one application domain. As teachers are expert in education, the application domain can be education. Data science teachers should also be familiar with Content Knowledge specific to data science, e.g., the data science workflow.

Teachers' gap in Pedagogical Content Knowledge (PCK)

Teachers who intend to teach data science should learn a Methods of Teaching Data Science (MTDS) course. This course should include the following topics:

- Pedagogical approaches taken from mathematical education and from computer science education, as well as their integration with the pedagogy of some application domain into an interdisciplinary education
- Multiple definitions of data science as a multifaced field and what each definition implies for data science teaching
- Challenges and opportunities of data science education
- Cognitive skills required for data science (mathematical thinking, statistical thinking, computational thinking, application domain thinking, interdisciplinary thinking)
- Pedagogical perspective of the data science workflow
- Technologies and learning environments for data science education
- The history of data science
- Social issues of data science (see below)

Project Based Learning (PBL)

Data science educators who ask students to develop a data science project by the project-based learning approach, either as an active learning method or as an evaluation method, should encourage learners to choose data for their project from an application domain they are specialist in, familiar with or at least interested in.

When the application domain of the projects is given, and students cannot choose it, data science educators should check learners' familiarity with application domain, and, if needed, an intervention program, similar to the program presented in Section 4.1.1, should be considered (Mike, Nemirovsky-Rotman and Hazzan, 2020).

5.2.1.2. Opportunities

Teaching STEM with real-world data

It is recommended to teach the fundamentals of mathematics, statistics and computer science required to practice data science, in the context of data science, using real-life data taken from a domain learners are familiar with. Such an interdisciplinary teaching approach promotes learners' motivation (Mike & Hazzan, 2021).

Professional development for teachers

Teachers from all disciplines, both STEM, social sciences and the humanities, are advised, as part of their professional development, to become familiar with the discipline of data science and methods of teaching data science. In such a professional development process, teachers are encouraged to practice data science with data taken from the educational field or from their field of expertise, for example, biology of literature. It is proposed that by such a professional development process, not only the teachers become familiar with data science, but they may also improve their understanding of the application domain they work with, as well as enrich the arsenal of their teaching methods.

Interdisciplinary pedagogy

Interdisciplinary pedagogy has been defined in this thesis as *solving pedagogical challenges in interdisciplinary education based on the integration of pedagogical methods from each of the disciplines* (Section 4.5.10). An example for such an interdisciplinary pedagogy is given in this thesis is teaching machine learning with pedagogical principles of mathematics education (Section 4.5). It is proposed that data science opens the opportunity to further develop this kind of pedagogy which gets recently a lot of attention due to the kind of problems the world faces. Data science educators and data science education researchers should develop and evaluate additional interdisciplinary pedagogical methods to leverage the opportunities of data science education (Mike & Hazzan, in press).

Teaching STEM with real-world context

While the first opportunity presented above, *teaching STEM with real-world data*, is derived from data science's attention to data, this opportunity, *teaching STEM with real-world context*, is derived from integration of a real-would application domain and its importance in the data science workflow. STEM educators from all disciplines are advised to integrate real-word problems in their teaching in general and in project-based learning processes. Such real-world problems may promote their teaching with respect to several aspects:

- Increasing learners' motivation
- Promotion of learners' diversity
- Teaching social issues of STEM

5.2.2. Learners' diversity challenges and opportunities

5.2.2.1. Challenges

Lack of mathematical knowledge

To support learners' understanding of machine learning algorithms (including how to select a machine learning algorithm, how to train it to achieve high performance, and how to use it in real-world context), it is recommended to teach machine learning algorithms as a white box with their details of operation (Mike & Hazzan, in press). To achieve a white box understanding of machine learning algorithms, it is recommended to teach machine learning algorithms using only mathematical objects with which learners are familiar and have already been conceptualized as objects (e.g., Pythagoras theorem for calculating Euclidian distance).

As a first machine learning algorithm, it is recommended to teach the KNN algorithm as its teaching as a white box requires the understanding of only basic mathematical objects commonly known to diverse populations (Hazzan and Mike, 2022).

Lack of computer science knowledge

For data science learners who lack computer science knowledge, it is recommended to teach computer science as an interdisciplinary field in the context of learners' application domain (Mike and Hazzan, 2022). It is further recommended to let learners collect their own data using their own specialty. Such a data collection process not only leverages learners' motivation, but also it reduces computer science educators' load to get data in application domains they are not familiar with.

Lack of statistical knowledge

Statistics is an important component of data science education that should not be neglected (Adams, 2020). In this thesis, the lack of statistical knowledge has been demonstrated with respect to the base rate neglect. Data science educators should be aware to this bias and its implications in data science education. The lion classification question (Table 4.9) and the tomatoes disease classification question (Table 4.10) can be used as homework or class activity to demonstrate the expression of this bias to learners.

5.2.2.2. Opportunities

Bridging gender gaps in STEM education

It is recommended that data science will be developed as a gender-balanced field (Berman & Bourne, 2015). For this purpose, it is recommended to frame data science as a field of its own and not as a subfield of computer science. It is further recommended to frame data science as a data analysis skill or research skill, and not as a programming skill (Mike, Hartal, and Hazzan, 2021).

5.2.3. Multifaced discipline challenges and opportunities

5.2.3.1. Challenges

Research skills

Data science is a research paradigm and a research method. Hence, to apply the full data science workflow, learners should acquire research skills, such as research question formulation, data gathering, data analysis and answering research questions. Learners from various learning levels, i.e., pre-graduates and graduate students, might find these skills challenging. It is recommended, however, to teach these skills as data science content. Yet, further research is needed to develop pedagogical methods to promote the teaching and learning of these skills (see Section 5.3).

Professional (soft) skills

As data science is a profession, it requires professional (soft) skills, both social and cognitive. It is recommended to integrate these skills in the context of data science education. Here are several examples for such skills:

- Social skills:
    - Familiarity with the business environment and strategy (see the domain neglect challenge presented in Section 4.1)
    - Communication and storytelling (see data science workflow in Section 2.1.5)
- Cognitive skills:
    - Avoidance of cognitive biases (see Sections 4.1 and 4.2)
    - Thinking on different levels of abstraction (see Section 4.5.9)
    - Critical thinking (for example for avoiding cognitive biases, see Section 4.1.2)

Ethics

As a profession, data science requires ethics. Understanding ethics requires the understanding of the socio-cultural context of the data. Thus, the domain neglect bias (Section 4.1) not only demonstrates the neglect of the application domain *knowledge*, but also the neglect of the *socio-cultural context* of the problem. In the case described in this thesis, learners neglected application domain knowledge related to human life that depends on the correctness of decisions made by the presented machine learning algorithms. It is therefore recommended that data science educators will emphasis the socio-cultural aspects of data, to increase learners' awareness of ethical issues of data science.

5.2.3.2. Opportunities

Teaching 21st century skills

Several components of data science, such as computational thinking and interdisciplinary thinking, data literacy, and soft skills, are considered as important 21st century skills. Hence, data science teaching is an opportunity to promote these skills in a real-world context as part of the data science learning processes.

5.3. Future research directions

This section presents several research topics that can continue the research described in this thesis and deepen its findings.

(a) The three challenges of data science as a multifaced discipline (see Table 5.1 – Data science education challenges and opportunities): research skills, professional (soft) skills and ethics, were identified in this research but were neither deeply researched nor pedagogical methods to mitigate them were developed. It is recommended to further study these challenges and to develop and evaluate pedagogical methods for their mitigation.

(b) The data science education challenges and opportunities should be studied with additional populations, such as learners in industrial and governmental environments.

(c) The tools developed to support white box understanding of machine learning algorithms by non-major data science learners (see Section 4.5.4) should be further developed and their usages should be further explored.

(d) Table 4.24 should be further validated based on a comprehensive literature review.

References

Adams, J. C. (2020). Creating a balanced data science program. *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*, 185–191. https://doi.org/10.1145/3328778.3366800

Al-Hashedi, K. G., & Magalingam, P. (2021). Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019. *Computer Science Review*, *40*, 100402.

Alvargonzález, D. (2011). Multidisciplinarity, interdisciplinarity, transdisciplinarity, and the sciences. *International Studies in the Philosophy of Science*, *25*(4), 387–403. https://doi.org/10.1080/02698595.2011.623366

Anderson, L. W., Bloom, B. S., & others. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman,.

Anderson, P., Bowring, J., McCauley, R., Pothering, G., & Starr, C. (2014). An undergraduate degree in data science: Curriculum and a decade of implementation experience. *Proceedings of the 45th ACM Technical Symposium on Computer Science Education - SIGCSE '14*, 145–150. https://doi.org/10.1145/2538862.2538936

Anderson, R. E., Ernst, M. D., Ordóñez, R., Pham, P., & Tribelhorn, B. (2015). A data programming CS1 course. *Proceedings of the 46th ACM Technical Symposium on Computer Science Education*, 150–155.

Avidov-Ungar, O., & Eshet-Alkakay, Y. (2011). The islands of innovation model: Opportunities and threats for effective implementation of technological innovation in the education system. *Issues in Informing Science and Information Technology*, *8*, 363–376.

Avnoon, N. (2021). Data scientists' identity work: Omnivorous symbolic boundaries in skills acquisition. *Work, Employment and Society*, 095001702097730. https://doi.org/10.1177/0950017020977306

Babbitt, T., Schooler, C., & King, K. (2019). Punch cards to python: A case study of a cs0 core course. *Proceedings of the 50th ACM Technical Symposium on Computer Science Education*, 811–817.

Becker, B. A., & Fitzpatrick, T. (2019). What do CS1 syllabi reveal about our expectations of introductory programming students? *Proceedings of the 50th ACM Technical Symposium on Computer Science Education*, 1011–1017.

Becker, B. A., & Quille, K. (2019). 50 years of CS1 at SIGCSE: A review of the evolution of introductory programming education research. *Proceedings of the 50th Acm Technical Symposium on Computer Science Education*, 338–344.

Ben-Zvi, D., & Garfield, J. B. (2004). *The challenge of developing statistical literacy, reasoning and thinking*. Springer.

Berman, F., & Bourne, P. E. (2015). Let's make gender diversity in data science a priority right from the start. *PLoS Biology*, *13*(7), e1002206.

Berman, F., Rutenbar, R., Christensen, H., Davidson, S., Estrin, D., Franklin, M., Hailpern, B., Martonosi, M., Raghavan, P., Stodden, V., & others. (2016). Realizing the potential of data science: Final report from the national science foundation computer and information science and engineering advisory committee data science working group. *National Science*

Foundation Computer and Information Science and Engineering Advisory Committee Report.

Berman, F., Rutenbar, R., Christensen, H., Davidson, S., Estrin, D., Franklin, M., Martonosi, M., Raghavan, P., Stodden, V., & Szalay, A. (2016). *Final Report from the National Science Foundation Computer and*. https://doi.org/10.1126/science.1167742

Berman, F., Rutenbar, R., Hailpern, B., Christensen, H., Davidson, S., Estrin, D., Franklin, M., Martonosi, M., Raghavan, P., Stodden, V., & Szalay, A. S. (2018). Realizing the potential of data science. *Communications of the ACM*, *61*(4), 67–72. https://doi.org/10.1145/3188721

Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J. M., & Eckersley, P. (2020). Explainable machine learning in deployment. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 648–657.

Biehler, R., & Schulte, C. (2018). *Paderborn symposium on Data Science education at school level 2017: The collected extended abstracts*. Universitätsbibliothek.

Biggs, J. B., & Collis, K. F. (2014). *Evaluating the quality of learning: The SOLO taxonomy (Structure of the Observed Learning Outcome)*. Academic Press.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). Taxonomy of educational objectives. Vol. 1: Cognitive domain. *New York: McKay*, *20*, 24.

Bower, M., & Richards, D. (2005). The impact of virtual classroom laboratories in CSE. *Proceedings of the 36th SIGCSE Technical Symposium on Computer Science Education*, 292–296.

Brink, H. I. (1993). Validity and reliability in qualitative research. *Curationis*, *16*(2), 35–38.

Bryant, C., Chen, Y., Chen, Z., Gilmour, J., Gumidyala, S., Herce-Hagiwara, B., Koures, A., Lee, S., Msekela, J., Pham, A. T., & others. (2019). A middle-school camp emphasizing data science and computing for social good. *Proceedings of the 50th ACM Technical Symposium on Computer Science Education*, 358–364.

Bryant, R. E., Sutner, K., & Stehlik, M. J. (2010). *Introductory computer science education at carnegie mellon university: A deans' perspective*.

Burlinson, D., Mehedint, M., Grafer, C., Subramanian, K., Payton, J., Goolkasian, P., Youngblood, M., & Kosara, R. (2016). BRIDGES: A system to enable creation of engaging data structures assignments with real-world data and visualizations. *Proceedings of the 47th ACM Technical Symposium on Computing Science Education*, 18–23.

C. F. Jeff Wu. (2021). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=C._F._Jeff_Wu&oldid=1049935836

Camp, T. (2002). The incredible shrinking pipeline. *ACM SIGCSE Bulletin*, *34*(2), 129–134.

Cao, L. (2017). Data Science: A Comprehensive Overview. *ACM Computing Surveys*, *50*(3). https://doi.org/10.1145/3076253

Cargill, K. (2005). Food studies in the curriculum: A model for interdisciplinary pedagogy. *Food, Culture & Society*, *8*(1), 115–123. https://doi.org/10.2752/155280105778055371

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, *8*(8), 832.

Casscells, W., Schoenberger, A., & Graboys, T. B. (1978). Interpretation by physicians of clinical laboratory results. *New England Journal of Medicine*, *299*(18), 999–1001. https://doi.org/10.1056/NEJM197811022991808

Cassel, B., & Topi, H. (2015). *Strengthening data science education through collaboration: Workshop report 7-27-2016*. Arlington, VA.

Chang, W. L., Grady, N., & others. (2015). *Nist big data interoperability framework: Volume 1, big data definitions*.

Chesley, A., Parupudi, T., Holtan, A., Farrington, S., Eden, C., Baniya, S., Mentzer, N., & Laux, D. (2018). *Interdisciplinary pedagogy, integrated curriculum, and professional development*.

Chevallard, Y. (1989). On didactic transposition theory: Some introductory notes. *Proceedings of the International Symposium on Selected Domains of Research and Development in Mathematics Education*, 51–62.

Cleveland, W. S. (2001). Data science: An action plan for expanding the technical areas of the field of statistics. *International Statistical Review*, *69*(1), 21–26.

Conway, D. (2010). The data science venn diagram. *Datist*. http://www.dataists.com/2010/09/the-data-science-venn-diagram/

Cortina, T. J. (2007). An introduction to computer science for non-majors using principles of computation. *Acm Sigcse Bulletin*, *39*(1), 218–222.

Cosmides, L., & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition*, *58*(1), 1–73.

Cox, M., & Ellsworth, D. (1997). Managing big data for scientific visualization. *ACM Siggraph*, *97*(1), 21–38.

Crooks, N. M., Bartel, A. N., & Alibali, M. W. (2019). Conceptual knowledge of confidence intervals in psychology undergraduate and graduate students. *Statistics Education Research Journal*, *18*(1), 46–62.

Danyluk, A., & Leidig, P. (2021). *Computing competencies for undergraduate data science curricula*. Association of Computing Machinery (ACM). https://dstf.acm.org/DSTF_Final_Report.pdf

Data science for humanity | data science. (2022). https://datascience.stanford.edu/research/research-areas/data-science-humanity

Data science in the humanities. (2022). http://ischool.illinois.edu/degrees-programs/courses/is417

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

De Veaux, R. D., Agarwal, M., Averett, M., Baumer, B. S., Bray, A., Bressoud, T. C., Bryant, L., Cheng, L. Z., Francis, A., Gould, R., Kim, A. Y., Kretchmar, M., Lu, Q., Moskol, A., Nolan, D., Pelayo, R., Raleigh, S., Sethi, R. J., Sondjaja, M., … Ye, P. (2017). Curriculum guidelines for undergraduate programs in data science. *Annual Review of Statistics and Its Application*, *4*(1), 15–30. https://doi.org/10.1146/annurev-statistics-060116-053930

Deci, E. L., & Ryan, R. M. (2012). *Self-determination theory.*

Deci, E. L., & Ryan, R. M. (2013). *Intrinsic motivation and self-determination in human behavior*. Springer Science & Business Media.

Demchenko, Y., Belloum, A., Los, W., Wiktorski, T., Manieri, A., Brocks, H., Becker, J., Heutelbeck, D., Hemmje, M., & Brewer, S. (2016). EDISON Data Science Framework: A foundation for building data science profession for research and industry. *2016 IEEE International Conference on Cloud Computing Technology and Science (CloudCom)*, 620–626. https://doi.org/10.1109/CloudCom.2016.0107

Dempster, M., & McCorry, N. K. (2009). The role of previous experience and attitudes toward statistics in statistics assessment outcomes among undergraduate psychology students. *Journal of Statistics Education*, *17*(2).

Denzin, N. K. (2017). *The research act: A theoretical introduction to sociological methods*. Routledge.

Digest of education statistics. (2020). National Center for Education Statistics. https://nces.ed.gov/programs/digest/d20/tables/dt20_322.50.asp

Digital around the world. (2022). DataReportal – Global Digital Insights. https://datareportal.com/global-digital-overview

Dodge, Y., Cox, D., & Commenges, D. (2006). *The Oxford dictionary of statistical terms*. Oxford University Press on Demand.

Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, *26*(4), 745–766.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *ArXiv Preprint ArXiv:1702.08608*.

Dragoumanos, S., Kakarountas, A., & Fourou, T. (2017). Young technology entrepreneurship enhancement based on an alternative approach of project-based learning. *IEEE Global Engineering Education Conference, EDUCON*, *April*, 351–358. https://doi.org/10.1109/EDUCON.2017.7942872

Dryer, A., Walia, N., & Chattopadhyay, A. (2018). A middle-school module for introducing data-mining, big-data, ethics and privacy using rapidminer and a hollywood theme. *Proceedings of the 49th ACM Technical Symposium on Computer Science Education*, 753–758. https://doi.org/10.1145/3159450.3159553

Duarte, C., Oliveira, H. P., Magalhaes, F., Tavares, V. G., Campilho, A. C., & De Oliveira, P. G. (2012). Proactive engineering. *IEEE Global Engineering Education Conference, EDUCON*. https://doi.org/10.1109/EDUCON.2012.6201104

Eccles, J. S., & Wigfield, A. (2020). From expectancy-value theory to situated expectancy-value theory: A developmental, social cognitive, and sociocultural perspective on motivation. *Contemporary Educational Psychology*, 101859.

EDISON: building the data science profession. (2022). https://edison-project.eu/

EDISON Data Science Framework (EDSF). (2017). https://edison-project.eu/edison/edison-data-science-framework-edsf/

Ejersbo, L. R., & Leron, U. (2014). Revisiting the medical diagnosis problem: Reconciling intuitive and analytical thinking. In *Probabilistic Thinking* (pp. 215–237). Springer.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, *17*(3), 37–37.

Fisher, N., Anand, A., Gould, R., Bailer, J., Hesterberg, T., Bailey, J., Ng, R., Burr, W., Rosenberger, J., Fekete, A., Sheldon, N., Gibbs, A., & Wild, C. (2019, September). *Curriculum frameworks for introductory data science*. http://www.idssp.org/files/IDSSP_Data_Science_Curriculum_Frameworks_for_Schools_Edition_1.0.pdf

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, *7*(2), 179–188. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x

Forte, A., & Guzdial, M. (2005). Motivation and nonmajors in computer science: Identifying discrete audiences for introductory courses. *IEEE Transactions on Education*, *48*(2), 248–253.

Fullan, M. (2007). *The new meaning of educational change*. Routledge.

Gal-Ezer, J., Beeri, C., Harel, D., & Yehudai, A. (1995). A high school program in computer science. *Computer*, *28*(10), 73–80.

Georgiopoulos, M., DeMara, R. F., Gonzalez, A. J., Wu, A. S., Mollaghasemi, M., Gelenbe, E., Kysilka, M., Secretan, J., Sharma, C. A., & Alnsour, A. J. (2009). A sustainable model for integrating current topics in machine learning research into the undergraduate curriculum. *IEEE Transactions on Education*, *52*(4), 503–512.

Glaser, B. G., & Strauss, A. L. (2017). *Discovery of grounded theory: Strategies for qualitative research*. Routledge.

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). *Deep learning* (Vol. 1). MIT press Cambridge.

Gorson, J., & O'Rourke, E. (2020). Why do CS1 students think they're bad at programming? Investigating self-efficacy and self-assessments at three universities. *Proceedings of the 2020 ACM Conference on International Computing Education Research*, 170–181.

Gould, R., Suyen, M.-M., James, M., Terri, J., & LeeAnn, T. (2018). Mobilize: A data science curriculum for 16-year-old students. *Iase-Web. Org*, 1–4.

Gray, E. M., & Tall, D. O. (1994). Duality, ambiguity, and flexibility: A "proceptual" view of simple arithmetic. *Journal for Research in Mathematics Education*, *25*(2), 116–140. https://doi.org/10.5951/jresematheduc.25.2.0116

Gray, J. (2007). *EScience – A transformed scientific method*. http://research.microsoft.com/en-us/um/people/gray/talks/NRC-CSTB_eScience.ppt

Grimmer, J., Roberts, M. E., & Stewart, B. M. (2021). Machine learning for social science: An agnostic approach. *Annual Review of Political Science*, *24*, 395–419.

Guzdial, M. (2003). A media computation course for non-majors. *Proceedings of the 8th Annual Conference on Innovation and Technology in Computer Science Education*, 104–108.

Hadim, H. A., & Esche, S. K. (2002). Enhancing the engineering curriculum through project-based learning. *Proceedings - Frontiers in Education Conference*, *2*, 1–6. https://doi.org/10.1109/fie.2002.1158200

Hancock, S. A., & Rummerfield, W. (2020). Simulation methods for teaching sampling distributions: Should hands-on activities precede the computer? *Journal of Statistics Education*, *28*(1), 9–17. https://doi.org/10.1080/10691898.2020.1720551

Haqqi, S., Sooriamurthi, R., Macdonald, B., Begandy, C., Cameron, J., Pirollo, B., Becker, E., Choffo, J., Davis, C., Farrell, M., & others. (2018). Data jam: Introducing high school students to data science. *Proceedings of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education*, 387–387.

Harris, H., Murphy, S., & Vaisman, M. (2013). *Analyzing the analyzers: An introspective survey of data scientists and their work*. O'Reilly Media, Inc.

Havill, J. (2019). Embracing the liberal arts in an interdisciplinary data analytics program. *Proceedings of the 50th ACM Technical Symposium on Computer Science Education*, 9–14.

Hazan, B., Zhang, W., Olcum, E., Bergdoll, R., Grandoit, E., Mandelbaum, F., Wilson-Doenges, G., & Rabin, L. (2018). Gamification of an undergraduate psychology statistics lab: Benefits to perceived competence. *Statistics Education Research Journal*, *17*(2), 255–265.

Hazzan, O. (1999). Reducing abstraction level when learning abstract algebra concepts. *Educational Studies in Mathematics*, *40*(1), 71–90.

Hazzan, O. (2003a). How students attempt to reduce abstraction in the learning of mathematics and in the learning of computer science. *Computer Science Education*, *13*(2), 95–122.

Hazzan, O. (2003b). Reducing abstraction when learning computability theory. *Journal of Computers in Mathematics and Science Teaching*, *22*(2), 95–117.

Hazzan, O., Dubinsky, Y., & Meerbaum-Salant, O. (2010). Didactic transposition in computer science education. *ACM Inroads*, *1*(4), 33–37.

Hazzan, O., Gal-Ezer, J., & Blum, L. (2008). A model for high school computer science education: The four key elements that make it! *Proceedings of the 39th SIGCSE Technical Symposium on Computer Science Education*, 281–285.

Hazzan, O., & Hadar, I. (2005). Reducing abstraction when learning graph theory. *Journal of Computers in Mathematics and Science Teaching*, *24*(3), 255–272.

Hazzan, O., & Lis-Hacohen, R. (2016). *The MERge model for business development: The amalgamation of management, education and research*. Springer.

Hazzan, O., & Mike, K. (2021). A journal for interdisciplinary data science education. *Communications of the ACM*, *64*(8), 10–11. https://doi.org/10.1145/3469281

Hazzan, O., & Mike, K. (2022). Teaching core principles of machine learning with a simple machine learning algorithm: The case of the KNN algorithm in a high school introduction to data science course. *ACM Inroads*, *13*(1), 18–25. https://doi.org/10.1145/3514217

Hazzan, O., Ragonis, N., & Lapidot, T. (2020). *Guide to teaching computer science: An activity-based approach*.

Hazzan, O., & Zelig, D. (2016). Adoption of innovation from the business sector by post-primary education organizations. *Management in Education*, *30*(1), 19–28.

Heinemann, B., Opel, S., Budde, L., Schulte, C., Frischemeier, D., Biehler, R., Podworny, S., & Wassong, T. (2018). Drafting a data science curriculum for secondary schools. *Proceedings of the 18th Koli Calling International Conference on Computing Education Research - Koli Calling '18*, 1–5. https://doi.org/10.1145/3279720.3279737

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science*, *15*(8), 534–539.

Hertz, M. (2010). What do CS1 and CS2 mean? Investigating differences in the early courses. *Proceedings of the 41st ACM Technical Symposium on Computer Science Education*, 199–203.

Hey, A. J., Tansley, S., Tolle, K. M., & others. (2009). *The fourth paradigm: Data-intensive scientific discovery* (Vol. 1). Microsoft research Redmond, WA.

Heyd-Metzuyanim, E., & Graven, M. (2019). Rituals and explorations in mathematical teaching and learning: Introduction to the special issue. *Educational Studies in Mathematics*, *101*(2), 141–151. https://doi.org/10.1007/s10649-019-09890-x

Irizarry, R. A. (2020). *The role of academia in data science education*.

Ishaq, A., Sadiq, S., Umer, M., Ullah, S., Mirjalili, S., Rupapara, V., & Nappi, M. (2021). Improving the prediction of heart failure patients' survival using SMOTE and effective data mining techniques. *IEEE Access*, *9*, 39707–39716.

Johnstone, I., & Roberts, F. (2014). *Data science at NSF*. https://www.nsf.gov/attachments/131215/public/StatSNSF3_31_14.pptx

Kahneman, D. (2002). Maps of bounded rationality: A perspective on intuitive judgment and choice. Nobel prize lecture, december 8. *Retrieved December*, *21*, 2007.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*(4), 237–251. https://doi.org/10.1037/h0034747

Kang, W., & Kilpatrick, J. (1992). Didactic transposition in mathematics textbooks. *For the Learning of Mathematics*, *12*(1), 2–7.

Khuri, S., VanHoven, M., & Khuri, N. (2017). Increasing the capacity of STEM workforce: Minor in bioinformatics. *Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education*, 315–320. https://doi.org/10.1145/3017680.3017721

Klawe, M. (2015). Closing the computer science gender gap: How one woman is making a difference in many lives. *The Conversation*.

Kliegr, T., Bahník, Š., & Fürnkranz, J. (2021). A review of possible effects of cognitive biases on interpretation of rule-based machine learning models. *Artificial Intelligence*, 103458.

Koehler, J. J. (1996). The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behavioral and Brain Sciences*, *19*(1), 1–17. https://doi.org/10.1017/S0140525X00041157

Kordaki, M., & Berdousis, I. (2020). Identifying Barriers For Women Participation In Computer Science. *Pro Edu. International Journal of Educational Sciences*, *2*(2), 5–20.

Kudryavtsev, A., & Pavlodsky, J. (2012). Description-based and experience-based decisions: Individual analysis. *Judgment & Decision Making*, *7*(3).

Lavie, I., Steiner, A., & Sfard, A. (2019). Routines we live by: From ritual to exploration. *Educational Studies in Mathematics*, *101*(2), 153–176. https://doi.org/10.1007/s10649-018-9817-4

Leron, U., & Dubinsky, E. (1995). An abstract algebra story. *The American Mathematical Monthly*, *102*(3), 227–242. https://doi.org/10.1080/00029890.1995.11990563

Leron, U., & Hazzan, O. (2009). Intuitive vs analytical thinking: Four perspectives. *Educational Studies in Mathematics*, *71*(3), 263–278.

Liu, J. S., & Huang, T. K. (2005). A project mediation approach to interdisciplinary learning. *Proceedings - 5th IEEE International Conference on Advanced Learning Technologies, ICALT 2005*, *2005*, 54–58. https://doi.org/10.1109/ICALT.2005.18

Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–16.

Lovell, M. C. (1983). Data mining. *The Review of Economics and Statistics*, *65*(1), 1–12.

Luxton-Reilly, A. (2016). Learning to program is easy. *Proceedings of the 2016 ACM Conference on Innovation and Technology in Computer Science Education*, 284–289.

Luxton-Reilly, A., Albluwi, I., Becker, B. A., Giannakos, M., Kumar, A. N., Ott, L., Paterson, J., Scott, M. J., Sheard, J., & Szabo, C. (2018). Introductory programming: A systematic literature review. *Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education*, 55–106.

Marcinkevičs, R., & Vogt, J. E. (2020). Interpretability and explainability: A machine learning zoo mini-tour. *ArXiv Preprint ArXiv:2012.01805*.

Mariescu-Istodor, R., & Jormanainen, I. (2019). Machine learning for high school students. *Proceedings of the 19th Koli Calling International Conference on Computing Education Research*, 1–9.

Merrill, M. D. (2002). First principles of instruction. *Educational Technology Research and Development*, *50*(3), 43–59. https://doi.org/10.1007/BF02505024

Mike, K. (2020). Data science education: Curriculum and pedagogy. *Proceedings of the 2020 ACM Conference on International Computing Education Research*, 324–325. https://doi.org/10.1145/3372782.3407110

Mike, K., Hartal, G., & Hazzan, O. (2021). Widening the shrinking pipeline: The case of data science. *2021 IEEE Global Engineering Education Conference (EDUCON)*, 252–261.

Mike, K., Hazan, T., & Hazzan, O. (2020). Equalizing data science curriculum for computer science pupils. *Koli Calling'20: Proceedings of the 20th Koli Calling International Conference on Computing Education Research*, 1–5.

Mike, K., & Hazzan, O. (2021). How can computer science educators benefit from data science education? *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*, 1363–1363.

Mike, K., & Hazzan, O. (2022). Interdisciplinary CS1 for non-majors: The case of graduate psychology students. *2022 IEEE Global Engineering Education Conference (EDUCON)*.

Mike, K., & Hazzan, O. (in press). Machine learning education: A white box approch. *Statistics Education Research Journal*.

Mike, K., Nemirovsky-Rotman, S., & Hazzan, O. (2020). Interdisciplinary education—The case of biomedical signal processing. *2020 IEEE Global Engineering Education Conference (EDUCON)*, 339–343. https://doi.org/10.1109/EDUCON45650.2020.9125200

Mike, K., Ragonis, N., Rosenberg-Kima, R. B., & Hazzan, O. (in press). Computational thinking in the era of data science. *Communications of the ACM*.

Mishra, P., & Koehler, M. (2006). Technological pedagogical content knowledge: A framework for teacher knowledge. *The Teachers College Record*, *108*(6), 1017–1054.

Mohebbi, M., Vanderkam, D., Kodysh, J., Schonberger, R., Choi, H., & Kumar, S. (2011). *Google correlate whitepaper*.

Moore, G. A. (2002). *Crossing the chasm: Marketing and selling disruptive products to mainstream customers*. HarperCollins. https://books.google.co.il/books?id=yJXHUDSaJgsC

National Academies of Sciences, Engineering, Medicine, & others. (2018). *Envisioning the data science discipline: The undergraduate perspective: Interim report*. National Academies Press.

Naur, P. (1966). The science of datalogy. *Communications of the ACM*, *9*(7), 485.

Othman, A., Hussin, H., Mustapha, M., & Parman, S. (2017). Cross-disciplinary team learning in engineering project-based: Challenges in collaborative learning. *2017 7th World Engineering Education Forum (WEEF)*, 866–871.

Penny, S. (2009). Rigorous interdisciplinary pedagogy: Five years of ACE. *Convergence*, *15*(1), 31–54.

Pfaff, T. J., & Weinberg, A. (2009). Do hands-on activities increase student understanding?: A case study. *Journal of Statistics Education*, *17*(3), 7. https://doi.org/10.1080/10691898.2009.11889536

Piatetsky-Shapiro, G. (1990). Knowledge discovery in real databases: A report on the IJCAI-89 Workshop. *AI Magazine*, *11*(4), 68–68.

Piatetsky-Shapiro, G. (2000). Knowledge discovery in databases: 10 years after. *Acm Sigkdd Explorations Newsletter*, *1*(2), 59–61.

Plaue, C., & Cook, L. R. (2015). Data journalism: Lessons learned while designing an interdisciplinary service course. *Proceedings of the 46th ACM Technical Symposium on Computer Science Education - SIGCSE '15*, 126–131. https://doi.org/10.1145/2676723.2677263

Prebor, G. (2021). When feminism meets social networks. *Library Hi Tech*.

Provost, F., & Fawcett, T. (2013). *Data science for business: What you need to know about data mining and data-analytic thinking*. O'Reilly Media, Inc.

Rabin, L., Fink, L., Krishnan, A., Fogel, J., Berman, L., & Bergdoll, R. (2018). *A measure of basic math skills for use with undergraduate statistics students: The macs11*.

Ramamurthy, B. (2016). A Practical and Sustainable Model for Learning and. *The 47th ACM Technical Symposium on Computer Science Education, SIGCSE 2016*, 169–174.

Reason, P., & Bradbury, H. (2001). *Handbook of action research: Participative inquiry and practice*. sage.

Reeve, J. (2002). Self-determination theory applied to educational settings. *Handbook of Self-Determination Research*, *2*, 183–204.

Rhoten, D., & Pfirman, S. (2007). Women in interdisciplinary science: Exploring preferences and consequences. *Research Policy*, *36*(1), 56–75.

Rogers, E. M. (1962). *Diffusion of innovations*. Free Press.

Rosenberg-Kima, R. B., & Mike, K. (2020). Teaching online teaching: Using the task-centered instructional design strategy for online computer science teachers' preparation. In *Teaching, Technology, and Teacher Education during the COVID-19 Pandemic: Stories from the Field* (pp. 119–123). Association for the Advancement of Computing in Education (AACE).

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215.

Ryan, R. M., & Deci, E. L. (2020). Intrinsic and extrinsic motivation from a self-determination theory perspective: Definitions, theory, practices, and future directions. *Contemporary Educational Psychology*, *61*, 101860.

Setiawan, A. W. (2019). Detailed comparison of instructor and student-based assessment in project based learning. *IEEE Global Engineering Education Conference, EDUCON*, *April-2019*, 557–560. https://doi.org/10.1109/EDUCON.2019.8725126

Sfard, A. (1991). On the dual nature of mathematical conceptions: Reflections on processes and objects as different sides of the same coin. *Educational Studies in Mathematics*, *22*(1), 1–36. https://doi.org/10.1007/BF00302715

Sfard, A., & Lavie, I. (2005). Why cannot children see as the same what grown-ups cannot see as different?—Early numerical thinking revisited. *Cognition and Instruction*, *23*(2), 237–309. https://doi.org/10.1207/s1532690xci2302_3

Shah, H., & Kumar, A. N. (2002). A tutoring system for parameter passing in programming languages. *ACM SIGCSE Bulletin*, *34*(3), 170–174.

Shearer, C. (2000). The CRISP-DM model: The new blueprint for data mining. *Journal of Data Warehousing*, *5*(4), 13–22.

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, *15*(2), 4–14.

Simberloff, D., Barish, B., Droegemeier, K., Etter, D., Fedoroff, N., Ford, K., Lanzerotti, L., Leshner, A., Lubchenco, J., Rossmann, M., & others. (2005). Long-lived digital data collections: Enabling research and education in the 21st century. *National Science Foundation*.

Skiena, S. S. (2017). *The data science design manual*. Springer.

Snow, J. (1855). *On the mode of communication of cholera (j. Churchill, london)*.

Srikant, S., & Aggarwal, V. (2017). Introducing data science to school kids. *Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education*, 561–566. https://doi.org/10.1145/3017680.3017717

Su, Y.-S., & Wu, S.-Y. (2021). Applying data mining techniques to explore user behaviors and watching video patterns in converged IT environments. *Journal of Ambient Intelligence and Humanized Computing*, 1–8.

Sullivan, D. G. (2013). A data-centric introduction to computer science for non-majors. *Proceeding of the 44th ACM Technical Symposium on Computer Science Education*, 71–76.

Sulmont, E., Patitsas, E., & Cooperstock, J. R. (2019a). Can you teach me to machine learn? *Proceedings of the 50th ACM Technical Symposium on Computer Science Education*, 948–954. https://doi.org/10.1145/3287324.3287392

Sulmont, E., Patitsas, E., & Cooperstock, J. R. (2019b). What is hard about teaching machine learning to non-majors? Insights from classifying instructors' learning goals. *ACM Transactions on Computing Education*, *19*(4), 1–16. https://doi.org/10.1145/3336124

Suresh, H., Lao, N., & Liccardi, I. (2020). Misplaced trust: Measuring the interference of machine learning in human decision-making. *12th ACM Conference on Web Science*, 315–324.

Suresh, H., Lewis, K. M., Guttag, J. V., & Satyanarayan, A. (2021). Intuitively assessing ML model reliability through example-based explanations and editing model inputs. *ArXiv Preprint ArXiv:2102.08540*.

Tartaro, A., & Chosed, R. J. (2015). Computer scientists at the biology lab bench. *Proceedings of the 46th ACM Technical Symposium on Computer Science Education*, 120–125. https://doi.org/10.1145/2676723.2677246

Tukey, J. W. (1962). The future of data analysis. *The Annals of Mathematical Statistics*, *33*(1), 1–67.

Tukey, J. W. & others. (1977). *Exploratory data analysis* (Vol. 2). Reading, MA.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*(4), 293.

Valero-Mora, P. M., & Ledesma, R. D. (2011). Using interactive graphics to teach multivariate data analysis to psychology students. *Journal of Statistics Education*, *19*(1).

Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., & others. (2019). Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery*, *18*(6), 463–477.

Way, T., & Whidden, S. (2014). A loosely-coupled approach to interdisciplinary computer science education. *Proceedings of the International Conference on Frontiers in Education: Computer Science and Computer Engineering (FECS)*, 1.

Wiberg, M. (2009). Teaching statistics in integration with psychology. *Journal of Statistics Education*, *17*(1).

Wing, J. M. (2006). Computational thinking. *Communications of the ACM*, *49*(3), 33–35. https://doi.org/10.1145/1118178.1118215

Wing, J. M. (2020). Ten research challenge areas in data science. *Harvard Data Science Review*. https://doi.org/10.1162/99608f92.c6577b1f

Wu, J. (1997). *Statistics = Data Science?* http://www2.isye.gatech.edu/~jeffwu/presentations/datascience.pdf

Yogeshwaran, S., Kaur, M. J., & Maheshwari, P. (2019). Project based learning: Predicting bitcoin prices using deep learning. *IEEE Global Engineering Education Conference, EDUCON, April-2019*(978), 1449–1454. https://doi.org/10.1109/EDUCON.2019.8725091

Zelig, D. (2011). *Adoption of innovation from the business sector by secondary educational organizations* [PhD Thesis]. Technion.

תקציר

בשנים האחרונות אנו עדים לצמיחתו של תחום מדעי חדש – מדעי הנתונים – העוסק בחקר השיטות להפקת ערך מנתונים. התחום צומח על רקע מהפכת הנתונים המתרחשת בעת הזו הודות לעלייה המשמעותית בכמות הנתונים הזמינים, העלייה בכוח המחשוב הזמין לעיבודם, ופיתוחן של יכולות אלגוריתמיות חדשות.

מדעי הנתונים הפכו לדיסציפלינה עצמאית בעשור האחרון. דיסציפלינה זו מבוססת על מספר תחומים: מתמטיקה, מדעי המחשב, סטטיסטיקה ותחומי הדעת של הנתונים. רוב המורים.ות, חוקרים.ות והמרצים.ות בתחום הגיעו מדיסציפלינות שונות בעלות מגוון מסורות של מחקר והוראה. כתוצאה מכך, לא התפתח עדיין ידע מספק בחינוך למדעי הנתונים, ובפרט, ביחס לתכנון לימודים ודרכי הוראה של התחום. בהתאם, מטרות המחקר הינן למפות פערים, אם קיימים, בין המצב הרצוי של תכנון לימודים ודרכי הוראה לתחום מדעי נתונים לבין המצב הקיים, ולהתוות קווים מנחים לפיתוח תוכניות לימודים ודרכי הוראה למדעי הנתונים באופן שימנף הזדמנויות חברתיות וחינוכיות.

ממטרות אלו נגזרה שאלת המחקר הבאה: מהם האתגרים הפדגוגיים וההזדמנויות החינוכיות בהוראת ובלמידת מדעי הנתונים?

המחקר נערך כשילוב של מחקר מעוגן בשדה ומחקר פעולה, בגישה משולבת איכותית וכמותית. החוקר השתתף בהוראה ולמידה של מדעי הנתונים ב-22 קורסים כמרצה או כסטודנט. משתתפי המחקר כללו תלמידים.ות מכיתות י'-י"ב, חיילים.ות, סטודנטים.ות לתואר ראשון, מורים.ות, סטודנטים.ות לתארים מתקדמים וחברי.ות סגל באוניברסיטאות מחקר. הנתונים נאספו באמצעות תצפיות משתתפות, שאלונים, ראיונות, מסמכים ויומן חוקר. אישורים אתיים לקיום המחקר נתנו על ידי ועדת האתיקה של הטכניון והמדען הראשי במשרד החינוך.

על בסיס ניתוח נתוני המחקר, התגבשה מסגרת תיאורטית מעוגנת בשדה המארגנת אתגרים והזדמנויות בחינוך למדעי הנתונים, על-פי שלוש תכונות מרכזיות של מדעי הנתונים: בינתחומיות, מגוון רחב של לומדים, ומדעי הנתונים כדיסציפלינה רבת פנים. האתגרים שנמצאו במחקר הינם:

- אתגרי בינתחומיות: אתגר בשילוב תחום הדעת של הנתונים, אתגר בהכשרת מורים, ואתגר ביישום למידה מבוססת פרויקטים;
- אתגרי מגוון לומדים: פער בידע קודם של הלומדים.ות, נושאים מורכבים להוראה ולמידה כגון למידת מכונה, והטיות קוגניטיביות;
- אתגרי מדעי הנתונים כדיסציפלינה רבת פנים: קושי בהוראת מיומנויות מחקר, הוראת מגוון רחב של מיומנויות רכות מקצועיות, ואתגר בהוראת אתיקה.

בנוסף נמצאו במחקר הזדמנויות חדשות אותן מאפשר החינוך למדעי הנתונים:

- יצירת איזון מגדרי בתחומי המדע, טכנולוגיה, הנדסה, ומתמטיקה (STEM)
- הוראת תחומי STEM באמצעות נתונים אמיתיים

- הוראת תחומי STEM בקונטקסט של בעיות אמיתיות
- פיתוח מקצועי למורים
- פדגוגיה אינטרדיסציפלינרית
- הוראת אוריינות נתונים
- הוראת מיומנויות המאה ה-21

על בסיס מסגרת מארגנת זו, נוסחה תזת המחקר על פיה *במקביל להתפתחותו של מדעי הנתונים כתחום בפני עצמו יש להקים ולפתח גם את תחום הוראת מדעי הנתונים כתחום בפני עצמו*; תחום הוראת מדעי הנתונים מבוסס על עקרונות הוראה אינטרדיסציפלינרית ומשלב ידע מתחומי הוראת המתמטיקה, הוראת סטטיסטיקה, הוראת מדעי המחשב, והוראת תחומי האפליקציה הרלוונטיים למדעי הנתונים. מסקנות המחקר מתארות את מאפייניו של תחום חדש זה. בנוסף, חיבור זה מציג המלצות לפיתוח תוכניות לימודים במדעי הנתונים ומחקרי המשך.

תרומתו התיאורטית של המחקר באה לידי ביטוי בפיתוח מסגרת תיאורטית לתיאור האתגרים וההזדמנויות בחינוך למדעי הנתונים, בפיתוח מסגרת תיאורטית להערכת רמת הבנה של מושגים במדעי הנתונים ובהרחבה של התיאוריה של חינוך אינטרדיסציפלינרי, כולל הגדרה נוספת למונח פדגוגיה אינטרדיסציפלינרית והגדרת המונח אינטרדיסציפלינריות פתוחה. תרומתו המעשית של המחקר באה לידי ביטוי בפיתוח תוכניות לימודים חדשות למדעי הנתונים, דרכי הוראה חדשות למדעי הנתונים, וכלי מחקר להערכת הבנת מושגים במדעי הנתונים.

המחקר נערך בהנחייתה של פרופ' אורית חזן מהפקולטה לחינוך למדע וטכנולוגיה בטכניון.

אני מודה ל:

1. מרכז להנדסת מערכות ע"ש ב. גורדון בטכניון
2. מרכז למכונות לומדות ובינה מלאכותית (MLIS) בטכניון
3. מר אייל קפלן
4. סטלה ויואל קרסו - קרן ביחד

על תרומתם הנדיבה למחקר.

המחקר קיבל את אישור

1. הועדה לאתיקה במחקר בבני אדם במדעי ההתנהגות בטכניון: 2018-073, 2019-001 ו-2019-059.
2. המדען הראשי, משרד החינוך, 11550.

תחום חדש נולד: חינוך למדעי הנתונים

חיבור על מחקר

לשם מילוי חלקי של הדרישות לקבלת התואר דוקטור לפילוסופיה

יעקב מייק

הוגש לסנט הטכניון – מכון טכנולוגי לישראל

חיפה

אייר תשפ"ב, מאי 2022

תחום חדש נולד: חינוך למדעי הנתונים

יעקב מייק