Information Science Students' Background and Data Science Competencies: An Exploratory Study

Ariel Rosenfeld
Department of Information Science, Bar-Ilan University, Ramat-Gan, Israel

Avshalom Elmalech
Department of Information Science, Bar-Ilan University, Ramat-Gan, Israel

Many Library and Information Science (LIS) training programs are gradually expanding their curricula to include computational data science courses such as supervised and unsupervised machine learning. These programs focus on developing both "classic" information science competencies as well as core data science competencies among their students. Since data science competencies are often associated with mathematical and computational thinking, departmental officials and prospective students often raise concerns regarding the appropriate background students should have in order to succeed in this newly introduced computational content of the LIS training programs. In order to address these concerns, we report on an exploratory study through which we examined the 2020 and 2021 student classes of Bar-Ilan University's LIS graduate training, focusing on the computational data science courses (i.e., supervised and unsupervised machine learning). Our study shows that contrary to many of the concerns raised, students from the humanities performed as well (and in some cases significantly better) on data science competencies compared to those from the social sciences and had better success in the training program as a whole. In addition, students' undergraduate GPA acted as an adequate indicator for both their success in the training program and in the data science part thereof. In addition, we find no evidence to support concerns regarding age or sex. Finally, our study suggests that the computational data science part of students' training is very much aligned with the rest of their training program.

Keywords: computational data science training, data science competencies, data science education

Information science (IS) is a discipline concerned with knowledge relating to the origination, collection, organization, storage, retrieval, interpretation, transmission, transformation, and utilization of information (Belkin and Robertson, 1976; Borko, 1968; Williams, 1988). Today, we witness a high volume of data that comes from a variety of sources and in many different forms (e.g., text, video, and audio). In order to derive scientifically valid insights from these data, the LIS community is gradually adopting data science (DS) techniques. While DS can be traced back to the fields of statistics and computer science (Davenport, 2020), DS techniques have also been used in the past decade to address a wide variety of research questions from various disciplines outside the exact sciences, such as music (Burgoyne et al., 2015), literary studies (Rommel, 2004), archaeology (Eiteljorg, 2004; Forte, 2015), linguistics (Hajič, 2004), history (Thomas, 2004; Zaagsma, 2013), philosophy

KEY POINTS:

- The field of library and information science is suitable for students with diverse backgrounds, including those without a computational background.
- The LIS field provides an excellent opportunity for older individuals to pursue a second career in data science.
- Undergraduate GPA can serve as an indicator of potential success in LIS studies.

(Ess, 2004), and many more. For illustrative purposes, let us consider the following two examples, which capture different underlying DS assumptions:

Example 1: Given a set of manuscripts, one may be interested in automatically dating a manuscript, identifying its author, its origin, and so on.

Example 2: Given a set of images, one may be interested in automatically categorizing the images into similar groups (e.g., genres or style), identifying patterns between different properties of the images (e.g., European-based images are associated with greater use of dark colors), signaling out abnormalities in the set (e.g., an image which poorly aligns with the others), and so forth.

In order to provide students with the computational tools needed for tackling real-world challenges such as the two examples outlined above, LIS training programs are gradually expanding their curricula to include not only "classic" IS subjects but also state-of-the-art computational data science courses (Urs & Minhaj, 2022; Wu et al., 2022; Zhang et al., 2022). Unlike "classic" LIS subjects, DS is associated more with so-called "computational thinking" (Wing, 2006). Specifically, DS often requires practitioners to write scripts, apply statistical models, and leverage large amounts of structured and/or unstructured data for extracting mathematically meaningful patterns and insights. For that reason, the introduction of DS topics in LIS programs has led our departmental officials and prospective students to raise concerns regarding the appropriate background students should have in order to succeed in this part of their LIS training. These concerns often revolve around a few key issues: relevant academic background, prior success in academic screening tests or in previous academic degree, age ("too old to comprehend technologically heavy material"), and sex ("computer-related studies are dominated by men").

In order to address these concerns, we report on an exploratory study through which we examined the 2020 and 2021 student classes of Bar-Ilan University's LIS graduate program, focusing on the computational data science courses (i.e., supervised and unsupervised machine learning) given by the two authors. Our study shows that contrary to many of the concerns raised, students from the humanities performed as well as those from the social sciences in data science competencies and had better success in the training program as a whole. In addition, students' undergraduate GPA, upon which acceptance to our program is commonly made, acted as an adequate indicator for their success in both the training program and the data science part thereof. Our results combine to suggest that the computational data science part of the students' training is very much aligned with the rest of their training program, thus mitigating many of the expressed concerns.

Background

Equipping LIS students with DS capabilities

IS training programs worldwide are reinventing themselves and developing new curricula to produce information professionals with the right knowledge and skills to meet changes in societal needs and the labor market (Bronstein, 2015; Hjørland, 2002; Johnson, 1999; Juznic & Badovinac, 2005). A central aspect of this process is the integration and development of DS competencies in LIS training programs (Wang, 2018; Zuo et al., 2017). Unfortunately, DS is a difficult-to-define practice (Garber, 2019; Van Dyk et al., 2015). Its true age, its relationship to previously existing fields like statistics and computer science, and even the profile of its practitioners are widely discussed and debated (Davenport, 2020). Traditionally, DS has been framed as an expansion of statistics (Cleveland, 2001). However, over the years, it became evident that DS competencies span much wider than statistics and encompass various other skills such as analytical skills, "open-mindedness skills," communications skills, mathematical skills, programming skills, and many more (Doyle, 2019). The diverse skills of DS practitioners are also apparent in their extremely diverse academic backgrounds and training (Davenport & Patil, 2012), ranging from exact sciences such as experimental physics to social sciences such as sociology. As a result, an agreed-upon definition of the skills of data science practitioners is currently unavailable (Fayyad & Hamutcu, 2020), despite various attempts to properly define it in various contexts (e.g., Agarwal & Dhar, 2014; Cao, 2017; Costa & Santos, 2017; Dhar, 2013; Van der Aalst, 2014).

Educators and scholars have struggled with defining their DS curriculum (Baumer, 2015; Brunner & Kim, 2016). Since DS methodologies and competencies are required by practitioners of various disciplines, DS courses are available to most students regardless of their backgrounds and major (Dichev & Dicheva, 2017). Previous research suggests that the main focus of DS courses should be on computation, statistics, machine learning, visualization, and ethics (Dichev & Dicheva, 2017). The Park City Math Institute has identified eight main subject areas of DS: data description and curation, mathematical foundations, computational thinking, statistical thinking, data modeling, communication, reproducibility, and ethics (De Veaux et al., 2017). Recent research that surveyed 69 college and university faculty teaching DS courses composed a list of topics most often taught in introductory DS courses for all disciplines. The list includes the following topics: data visualization, data cleaning, ethics, data management, statistical methods, professional practices, data architecture, and machine learning (Schwab-McCoy et al., 2020). The DS topics emerging from these and other lists can be attributed to the four fundamental DS competencies required by almost every DS and IS project: data pre-processing, data exploration, data analysis, and data presentation (Kang et al., 2015). Practically this entire skill set is commonly reflected through the development and deployment of machine learning (ML) (Jordan & Mitchell, 2015) techniques, which are central to the work of most data science practitioners. Specifically, the ability to leverage supervised and unsupervised machine learning algorithms to tackle various types of data is imperative for DS practitioners and encompasses most DS skills discussed above. For instance, Example 1 above may require knowledge of supervised learning techniques, while Example 2 may require knowledge in unsupervised learning techniques.

In this work, we focus on these two types of ML skills as pertaining to the two investigated academic courses given in our training program. We discuss these two types of ML skills in detail next.

Core machine learning approaches

Machine learning (ML) involves computer programs discovering how they can perform tasks without being explicitly programmed to do so (Shalev-Shwartz & Ben-David, 2014). Traditionally, ML techniques are divided into categories which vary in their underlying assumptions, theoretical basis, evaluation metrics, and application settings. The two most fundamental categories are (1) *Supervised learning*: where a program receives example inputs (also known as training data) along with their desired outputs (also known as supervisory signals), given by a "supervisor" (commonly, a human annotator), and the goal of the program is to learn a general rule that maps new, unseen inputs to their correct output; this category may apply to Example 1 above; and (2) *Unsupervised learning*: where the training data are not associated with any outputs, leaving it to the program to figure out what (possibly complex) structure and patterns are "hidden" in its inputs; this category may apply to Example 2.

Next, we formally define supervised and unsupervised machine learning and highlight the principles and techniques taught in our program.

Supervised ML techniques

Supervised ML is concerned with automatically learning a mapping between inputs into outputs based on a training set of examples consisting of input-output pairs (Russell & Norvig, 2002). Specifically, assume that we have a dataset $D = \{(x_1; t_1), \ldots, (\mathrm{x_N}; \mathrm{t_N})\}$, where x_i is a (commonly, vector) representation of the input and t_i is the output signal associated with x_i (commonly a label or a real value). A supervised ML technique is thus concerned with learning a "good" mapping for predicting the outputs for new, possibly unseen, inputs. By a "good" mapping one usually refers to a combination of mathematical properties such as accuracy and robustness.

The two most prominent prediction settings of supervised ML are classification and regression. In classification settings, the task is predicting a discrete label or category (commonly, from an unordered set). Considering Example 1, given a manuscript x, we may be interested in automatically identifying its author from a list of potential authors $t \in T$. In this example, the training data D may consist of a set of manuscripts (x_i represented in some standard form), each associated with the name of its author ($t_i \in T$). The selected classification algorithm thus needs to predict the correct author for a new unseen manuscript. In regression settings, on the other hand, the task is predicting an integer or continuous number (commonly a real value). In Example 1, this may refer to predicting the publication year of a manuscript. As in the classification case, the training data would probably consist of manuscripts (x_i), here each of them associated with its publication year (t_i). The selected regression algorithm would thus need to predict the publication year of a new unseen manuscript. The differences between the two supervised learning settings entail

different algorithms, selection criteria, evaluation metrics, and more (see Shalev-Shwartz & Ben-David, 2014 for further reading).

In our training program, we teach the basic principles of empirical risk minimization and the following classic classification techniques: support vector machines (SVM), naive Bayes, decision trees, K-nearest neighbor algorithm, and neural networks. In addition, we teach the following regression techniques: linear regression and logistic regression. In both settings, special attention is placed on the selection, evaluation, and comparison of supervised ML techniques and the use of feature selection and parameter optimization as part of the learning process.

Unsupervised ML techniques

Unsupervised learning is concerned with automatically learning previously undetected patterns in a dataset with no pre-existing output signals (Russell & Norvig, 2002). In contrast to supervised learning, which makes use of input-output pairs, unsupervised learning techniques receive only inputs and are faced with the challenge of identifying commonalities, patterns, and anomalies based on the inputs alone. Formally, an unsupervised algorithm receives as input a dataset $D = \{x_1, \ldots, x_N\}$, where x_i is a (commonly, vector) representation of the input and outputs some form of description for the hidden structures of the data.

Because there is no "ground truth" supervisory signal associated with the data, it is difficult to capture the appropriateness of an algorithm trained with unsupervised learning in a learning setting. As a result, a wide variety of intriguing challenges, limitations, considerations, and best practices for the use of unsupervised learning have emerged. In the scope of our training program, we focus on the three most prominent lines of approaches: (1) clustering, where the learning algorithm seeks to identify data instances that are similar to each other and groups them together, ideally revealing the internal structure of the input space; considering Example 2 above, this would mean dividing the images into distinct subgroups such that each subgroup is cohesive yet different from the others; (2) anomaly detection, where a learning algorithm looks for unusual patterns in inputs; using Example 2, this could mean identifying the images that are not similar to any other painting in the set or fail to align with the general pattern in the set, and from an applicative perspective, the learning algorithm can be used to flag these images in a dataset for further consideration; and (3) association, where a learning algorithm looks for certain features of a data sample that correlate with other features of that sample; considering Example 2, this could mean that key attributes of an image may be associated with other attributes, such as the origin of the image perhaps being associated with the use of different colors.

In our training program, we teach the fundamental principles and challenges of unsupervised learning and the following classic clustering techniques: K-means and hierarchical clustering; anomaly detection techniques: local outlier factor, cluster analysis-based outlier detection, and deviations from association rules and frequent itemsets; and association rule mining techniques: aprioiri algorithm and FP-growth. In all three settings, special attention is placed on the selection, evaluation, and comparison of the different techniques and the use of feature reduction and visualization as part of the learning process.

Bar-Ilan University's LIS training program

Our LIS graduate training program is offered to students with diverse backgrounds and does not require any prior technological background. Our LIS program, which is the only one in Israel, lasts for two to three years, during which students are trained in both "classic" LIS topics and DS-related topics. The curricula offered in our program are provided in the Appendix. One of our program's missions is to provide our students with in-depth knowledge in DS in addition to a broad knowledge in IS. The main DS courses taught in our program are mathematical foundations for DS, statistics, introduction to programming in python, advanced programming, data visualization, supervised learning, and unsupervised learning. These courses are specifically targeted to develop the four fundamental DS competencies in our students (Kang et al., 2015):

- data pre-processing: the ability to extract usable data from a larger set of raw data (Bartschat et al., 2019; Hand & Adams, 2014; Swamidason, 2019);
- data exploration: the ability to spot trends in data, perform exploratory analysis to make sense of the data, and identify interesting hypotheses (Adèr, 2008; Russo & Zou, 2019; Simmons et al., 2011);
- data analysis: the ability to construct the right model for the data, transfer data to concrete knowledge, and evaluate the model's ability to address the hypotheses of the study (Awan et al., 2019; Gibert et al., 2010; Raschka, 2018); and
- data presentation and communication: the ability to communicate/explain the data to people of different skill sets (i.e., management), explain the importance of patterns in the data, and suggest solutions (Gilpin et al., 2018; Vellido, 2019; Vellido et al., 2011).

At the beginning of their training, students take the introductory DS courses, which provide them with the basic mathematical, programmatic, and technical skills required by the more DS-intensive courses: supervised and unsupervised learning. These courses focus on the core DS competencies discussed above. In this study we focus on these two DS-intensive courses: supervised and unsupervised learning as they best reflect the students' acquired DS competencies.

Methodology

In order to identify students' and departmental officials' concerns regarding student success, we informally interviewed the head of the LIS department and the two administrative figures in the department who are in charge of student recruitment and are in direct contact with prospective students. These short interviews included a single question: "What are the main concerns expressed by prospective students?" In addition, we assembled a focus group consisting of all lecturers in our department (both tenure-track and non-tenure-track) to discuss their concerns regarding the appropriate background that students should have in order to succeed in our program and in the computational part thereof.

The expressed concerns can be generally categorized into the following: concerns regarding future employment, concerns regarding appropriate student background for the training program, and appropriate student background for the computational DS part of the training program. Additional "general" concerns such as which teaching platforms will

be used during the COVID-19 pandemic were also expressed but are not considered further, as those are not the focus of this study.

Starting with student background, we have identified the following key concerns: (1) humanities graduates may face difficulties in coping with DS content due to their limited mathematical background; (2) students' undergraduate GPA (which is the main screening tool for our program) may be a poor predictor for student success in the training program, and in the DS part thereof, due to variability in students' academic backgrounds (i.e., institution, major, etc.); (3) low SAT scores may deter prospective students from DS, since similar undergraduate programs commonly require high SAT scores;[1] (4) older students may be less technologically oriented and thus struggle more with the computational content; and (5) female prospective students often perceive DS to be a male domain.

It is important to note that the above concerns are not unique to our LIS training program but can be found in different forms across different fields. For example, Tariq and Durrani (2012) found that male students, younger students (aged 18-29), and those with previous academic mathematical background tend to present greater confidence in their mathematical and computational skills. Similarly, Guo (2017) has shown that older adults attending programming courses reported higher levels of frustration, including a perceived lack of opportunities to interact with tutors and peers and trouble dealing with constantly changing technologies.

In order to address the above five concerns, we investigated the classes of 2020 and 2021 and explored students' performance in the Supervised Machine Learning (SML) course, the Unsupervised Machine Learning (UML) course, and in the training program as a whole. (The SML and UML courses' contents are outlined above.) The SML course is given by the second author, who was the recipient of the distinguished lecturer award of Bar-Ilan University for 2017, and was attended by 26 students in 2020 and 27 students in 2021. The UML course is given by the first author, who was the recipient of the distinguished lecturer award of Bar-Ilan University for 2018, and was attended by 31 students in 2020 and 28 students in 2021. Overall, in 2020, 24 students attended both courses (9 male, average age of 35 ± 7 years), and in 2021, 20 students attended both courses (8 male, average age of 32 ± 7.5 years). The student classes do not differ significantly in terms of sex and age. The fact that both courses were taught by distinguished lecturers allows us to safely assume that the teaching level in both courses was average or above.

Since the two courses were given by the authors, we had the unique opportunity to investigate the students' competencies first hand. To that end, we devised two final projects for the two courses: In the SML course, students were assigned the classic task of predicting the price of an artifact given a labeled data set. In the UML course, students were given a large set of documents (in our case, academic papers) and were asked to explore the possible "hidden" patterns in the set. The projects were performed individually and checked for plagiarism. Each of the authors manually and independently examined and graded each student's assignment using the following criteria: (1) data pre-processing, (2) exploration, (3) analysis, and (4) presentation and communication. These criteria correspond with the four fundamental DS competencies discussed above.

Specifically, in addition to grading the work on the standard scale, the authors rated the competency level of each student on each of the examined criteria, on a 5-point Likert scale ranging from "very poor" to "very competent." Overall, each student was assigned eight scores (four scores for each course).

In addition to these scores, we extracted the following additional information about the students which corresponded to the raised concerns:

1. undergraduate major;
2. undergraduate GPA;
3. SAT score;
4. age; and
5. sex.

This information was cross-referenced with the students' Master's average grade, which acted as an indicator of their success in the entire LIS training program.

Unlike age and sex, which did not differ between the 2020 and 2021 student classes, our student profile changed with respect to other criteria. Specifically, the students' undergraduate major and GPA were significantly different. The COVID-19 pandemic brought changes that have impacted our potential student pool, including significant changes in the job market and changes in our teaching platforms, leading to a different mixture of student profiles. Specifically, while students' SAT scores were roughly the same across the two classes (583 ± 92 in 2020, 577 ± 86 in 2021), students' undergraduate major and GPA were not. In 2020, out of the 24 students, 12 graduated with a major in the humanities, 11 in the social sciences, and one in the exact sciences. However, in 2021, out of 20 students, only four graduated with a major in the humanities, 11 in the social sciences, and five in other disciplines such as nursing and engineering.

Analysis and results

In order to properly address the raised concerns, we first had to determine what constituted "success" in the DS part of the training. Recall that we focused on two separate courses (SML and UML), in each of which the four DS competencies (discussed in Section 2) were graded. A preliminary question could therefore be posed: Are DS competencies related? While one cannot conclusively determine a possible relation, we considered two correlation-based questions: (1) Are students' scores on DS competencies correlated between the two courses? and (2) Are students' scores on DS competencies correlated with one another?

Starting with the comparison of the two courses, Table 1 summarizes the results. As can be seen in the table, the scores of both courses seem to correlate. To investigate this correlation, we summarized the scores for each course into a single one and examined the Spearman Rank Correlation (r) between them. The results indicate a significant moderately positive correlation of $r = 0.47$; $p < 0.05$ in 2020, and $r = 0.71$; $p < 0.01$ in 2021. We continue by analysing the correlation between the courses on each of the four competencies: for data pre-processing $r = 0.68$; $p < 0.01$ (for 2020) and $r = 0.62$; $p < 0.01$ (for 2021), for data exploration $r = -0.23$; $p = 0.28$ (for 2020) and $r = 0.17$; $p = 0.47$, for data analysis $r = 0.36$;

Table 1: Students' scores on the four examined competencies in 2020 and 2021

	2020		2021	
Competency	SML	UML	SML	UML
Pre-processing	4:25 ± 0:6(4)	4:2 ± 0:8(4)	4:7 ± 0:5(5)	4:55 ± 0:6(4)
Exploration	4:92 ± 0:4(5)	4:33 ± 0:6(4)	4:3 ± 0:6(4)	4:5 ± 0:5(4)
Analysis	3:37 ± 1:3(3)	3:57 ± 0:9(3)	4 ± 0:8(4)	4 ± 0:9(4)
Presentation	4:46 ± 0:8(5)	4 ± 0:7(4)	3:95 ± 0:9(4)	4:1 ± 0:6(4)

Notes. Scores are on a scale of 1-5 (5 being the highest) and are reported along with the rounded standard deviation. Medians are reported in parentheses.

$p < 0.1$ (for 2020) and $r = 0.2$; $p = 0.39$ (for 2021), and for data presentation $r = 0.04$; $p = 0.85$ (for 2020) and $r = 0.07$; $p = 0.77$ (for 2021). Overall, the correlation between the data pre-processing competency (in both classes) and the data analysis competency (in 2020) across the two courses were found to be significant, while the data exploration and data presentation competencies were not.

We now turn to investigate the correlation between each of the competencies in each of the two courses. The correlation matrix is provided in Tables 2 and 3. The results suggest that for both the SML and UML courses, most competencies are strongly correlated.

Success in the DS part of our program

We start by examining whether the students' undergraduate major and GPA, SAT score, age, or sex were indicative of their success in the DS part of the training (as measured by the four DS competencies). Starting with students' undergraduate majors, we compared those who graduated from the social sciences and those who graduated from the humanities in 2020 (a single student from 2020 who graduated from the exact sciences was omitted). Note that we did not consider the 2021 class in this analysis since it consisted of only four humanities graduates. First, we examined the students' scores in the examined competencies across the two investigated courses. Using the Mann-Whitney U test we found that a statistically

Table 2: Correlation matrix between each of the four examined competencies for 2020

Competency (2020)	Exploration	Analysis	Presentation
Pre-processing	0.36*,0.31	0.78***,0.5**	0.6***,0.44**
Exploration	-	0.78***,0.17	0.6***,0.29
Analysis	-	-	0.58***,0.75***

Notes. Each cell reports two correlations: first for the SML course and second for the UML course.
* < 0.1. ** < 0.05. *** < 0.01.

Table 3: Correlation matrix between each of the four examined competencies for 2021

Competency (2021)	Exploration	Analysis	Presentation
Pre-processing	0.54**,0.76***	0.8***,0.66***	0.59***,0.53**
Exploration	-	0.75***,0.67***	0.76***,0.64***
Analysis	-	-	0.76***,0.72***

Notes. Each cell reports two correlations: first for the SML course and second for the UML course.
* < 0.1. ** < 0.05. *** < 0.01.

significant difference is encountered in the data analytic competency for both courses, $p < 0.01$ for the SML course and $p < 0.1$ for the UML course. In both courses, humanities graduates scored higher than those of social sciences.Tables 4 and 5 summarize the results. In addition, in the SML course, humanities graduates scored higher on all the examined criteria (on average, although the difference is not statistically significant), while in UML the same is true for only two of the four competencies.

Considering the students' undergraduate GPA, we find a few positive correlations between grades and competency scores, indicating the predictive value of the undergraduate GPA.Table 6 summarizes the results.

SAT scores and age seem to be poorly associated with students' scores on any of the examined competencies. Specifically, we could not identify any statistically significant correlations between the SAT scores and/or age and any of the competencies examined in this study using correlation analysis. While most correlations were positive, all had relatively high p values.

Considering the students' sex, using the Mann-Whitney U test, we were unable to detect any statistically significant differences between male and female students in the examined competencies. However, a notable difference was detected on the data analytic competency, where female students scored 20% higher than male students on average in 2020 (3.6 compared to 3) and 13.5% higher in 2021 (4.2 compared to 3.7).

Table 4: SML students' scores on the four examined criteria for the classes of 2020 and 2021

	2020		2021	
Competency	Social Sciences	Humanities	Social Sciences	Humanities
Pre-processing	$4 \pm 0.7(4)$	$4.4 \pm 0.5(4)$	$4.82 \pm 0.4(5)$	$4.5 \pm 0.6(4)$
Exploration	$4.8 \pm 0.6(5)$	5(5)	$4.2 \pm 0.6(4)$	$4.25 \pm 0.5(4)$
Analysis	$2.7 \pm 1.6(3)$	**4 ± 0.85 (4)**	$4.2 \pm 0.6(4)$	$3.75 \pm 0.95(3)$
Presentation	$4.2 \pm 0.75(4)$	$4.7 \pm 0.64(5)$	$3.7 \pm 0.9(3)$	$3.75 \pm 0.95(3)$

Notes. Scores are on a scale of 1-5 (5 being the highest) and are reported along with the rounded standard deviation. Medians are reported in parentheses. The result in bold is significant with $p < 0.01$.

Table 5: UML Students' scores on the four examined criteria for the classes of 2020 and 2021

	2020		2021	
Competency	Social sciences	Humanities	Social sciences	Humanities
Pre-processing	4 ± 0.77(4)	4.45 ± 0.49(4)	4.4 ± 0.67(4)	4.5 ± 0.58(4)
Exploration	4:.4 ± 0.66(4)	4.2 ± 0.6(4)	4.5 ± 0.52(4)	4.25 ± 0.5(4)
Analysis	3.14 ± 1.06(3)	**3.73 ± 0.64(4)**	3.9 ± 0.95(3)	3.75 ± 0.95(3)
Presentation	4.2 ± 0.4(4)	4.1 ± 0.63(4)	4.1 ± 0.83(4)	4.25 ± 0.5(4)

Notes. Scores are on a scale of 1-5 (5 being the highest) and are reported along with the rounded standard deviation. Medians are reported in parentheses. The result in bold is significant with $p < 0.01$.

Success in our LIS program

We now turn to examine whether students' undergraduate major and GPA, SAT score, age, and sex are indicative of students' success in our LIS program (as measured by the students' Master's average grade). For the 2020 class, we found a statistically significant difference between students based on their undergraduate majors: Humanities graduates' average Master's final grade was significantly higher than that of the social sciences' graduates (89.3 ± 3.04 compared to 85.6 ± 4.74, $p < 0{:}05$). It is, however, important to note that no significant differences were detected when examining these students' average SAT scores (595.3 ± 92 in the humanities compared to 573.7 ± 100 in the social sciences, $p = 0{:}36$). As before, we did not consider the 2021 class in this analysis due to very low number of humanities graduates (4).

The students' undergraduate GPAs were found to be strongly associated with their Master's final grade, $r = 0.67$, $p < 0.01$ (for 2020) and $r = 0.56$; $p < 0.05$ (for 2021). Specifically, the average Master's final grade of students with below 85 in the undergraduate GPA was 6 points lower (in 2020) and 2.5 points lower (in 2021) than students with above 85 GPA, $p < 0.01$.

Table 6: Correlation between the students' undergraduate GPA and the examined competencies across the two courses

	2020		2021	
Competency	SML	UML	SML	UML
Pre-processing	0.59***	0.43**	0.37	0.24
Exploration	0.002	0.13	0.25	0.4
Analysis	0.62***	0.44	0.37	0.59**
Presentation	0.55***	0.34*	0.13	0.39

* < 0.1. ** < 0.05. *** < 0.01.

Table 7: Correlation between the examined competencies across the two courses and the students' success in the rest of the training program

	2020		2021	
Competency	SML	UML	SML	UML
Pre-processing	0.66***	0.53***	0.54**	0.65***
Exploration	0.05	0.03	0.46**	0.47**
Analysis	0.77***	0.39*	0.65***	0.63***
Presentation	0.67***	0.25	0.58***	0.43*

* < 0.1. ** < 0.05. *** < 0.01.

In addition to the weak correlation between SAT scores and the examined competencies discussed above, we found some correlation between SAT scores and the students' success in our training program, of $r = 0.55$, $p < 0.05$ (in 2020) and $r = 0.36$, $p = 0.12$. Specifically, using the Mann-Whitney U test, we found that the average Master's final grade of students with below average SAT scores (<555) was lower by 4 points (in 2020) and 1 point (in 2021) compared to students with above average SAT scores, $p < 0.05$.

As was the case for DS competencies, using the Mann-Whitney U test, we were unable to detect any statistically significant differences between male and female students. In our case, male students averaged 86.4 ± 5 (in 2020) and 91.2 ± 4 (in 2021), while female students averaged 88.2 ± 3.5 (in 2020) and 92.3 ± 4 (in 2021).

Similarly, age seems to be weakly associated with the students' Master's final grades. A non-significant slight positive correlation of $r = 0.16$ (in 2020) and $r = 0.3$ (in 2021) were found.

Data science competencies and success in our LIS program

Before we conclude our analysis, we further look at whether the examined DS competencies correlate with student success in the rest of their IS training. This examination will help situate the DS part within the LIS program in terms of student success.

As can be seen in Table 7, for both courses and examined years, most competencies are found to be significantly correlated with the students' success in the rest of their LIS training.

Discussion

We start by discussing the results of the preliminary examination of the DS competencies.

The results indicate that the DS competencies are correlated between the two courses. This could suggest that DS competencies are manifested in a similar way in both the supervised and unsupervised learning contexts. For example, we speculate that pre-processing and analysis are highly programmatic and technologically oriented and, as a result, students tend to score similarly on these competencies regardless of the learning setting. Turning to investigate the possible relation between the competencies themselves, we found that for both courses, most competencies are strongly correlated.

We now turn to address the key concerns for students' success in the DS part of their training. Interestingly, contrary to the concerns raised by prospective students and departmental officials, social sciences graduates did not score significantly higher on any of the examined competencies. In fact, in 2020, humanities graduates scored significantly higher on the data analytic competency. Moreover, female students tended to score higher than male students in both years. It is important to note that, as of today, the data science profession is heavily dominated by male non-humanities majors. The results could suggest that students from diverse backgrounds can succeed in leading LIS programs and acquire the necessary DS competencies that are required of LIS professionals. In addition, no meaningful correlation was detected between age and the examined DS competencies. These results could help in mitigating some of the associated concerns of humanities graduates, female prospective students, and older prospective students. In addition, the results show that the students' undergraduate GPA adequately correlates with the examined DS competencies, while students' SAT scores do not. These results suggest that prior success in academic studies (at the undergraduate level) can indicate potential success in the DS part of our LIS training better than SAT scores can. SAT scores, which our focus group had initially believed could act as good indicators for student success, only weakly correlated with the examined DS competencies. Note, however, that both students' undergraduate GPA and SAT scores do correlate with the students' success in our Master's program, as we will discuss next.

We now turn to address the concerns regarding students' success in our LIS training program as a whole. Similarly to the analysis of student success in the DS part of their training, humanities graduates' average Master's final grade is higher than that of social sciences graduates. In addition, female students are slightly more successful than male students (female students average 1−2 points higher than male students, yet the difference is not statistically significant), and a weak positive correlation between age and success in the program was found, meaning that older students were slightly more successful than younger students. As before, these results could help in mitigating some of the associated concerns of humanities graduates, female prospective students, and older prospective students. As was the case before, students' undergraduate GPA was strongly correlated with the students' Master's final grade. Unlike the above analysis, SAT scores were found to better correlate with student success in the LIS training program compared to their correlation with the DS competencies. We believe that this is the result of the integrative nature of the SAT score as well as the students' Master's final grade. Both scores reflect a complex combination of sub-scores which, individually, may or may not align with each other.

Table 8 summarizes the examined criteria, comparing their respective indication strength for success in the DS part of the training as well as the program as a whole.

As for the possible relation between students' performance in the DS courses and their success in the rest of their LIS training, we found that most DS competencies were well correlated with students' success.

The data analysis competency is of special interest. Specifically, it was almost consistently the lowest-rated competency out of the four examined in both the SML and UML courses (Table 1), it was correlated with most other competencies (Table 2 and 3), and it was correlated both with students' undergraduate GPA (Table 6) and with their Master's

Table 8: Summary of possible indicators for student success in the DS part and the LIS training as a whole

Indicator	DS	LIS
Sex	Weak	Weak
Age	Weak	Weak
SAT	Weak	Moderate
Undergraduate major	Moderate	Moderate
Undergraduate GPA	Moderate	Strong

Note. Each cell represents the indication's strength.

final grade (Table 7). This result leads us to conjecture that the data analysis competency is the pivotal competency in our DS training. As such, we believe that this competency should play a greater role in the training and evaluation of LIS students, possibly in the scope of additional courses. With greater emphasis on this competency, student data analysis performance should improve, which, in turn, could lead to the training of better LIS professionals.

Conclusions

In this study, we have identified key concerns expressed by prospective students and departmental officials regarding student success in our LIS training program. We focused on concerns regarding the computational DS part of our training. In order to address these concerns, we examined the 2020 and 2021 student classes of Bar-Ilan University's LIS graduate training. We investigated their background, acquired DS competencies, and success in our LIS training program.

We found that humanities graduates perform as well as social sciences graduates in both the DS part of the training and the training program as a whole (and sometimes, even better). In addition, we further found that the students' undergraduate GPA was a consistent predictor of student success. Moreover, we found that students' SAT scores did not indicate success in the DS part of the training but did indicate success in the training program as a whole. Female students tended to outperform male students in the measured performances, and no meaningful relation was found between students' age and their measured performances. We believe that these results combine to relieve some of the concerns expressed by prospective students and departmental officials. Specifically, we find no evidence that supports most of the raised concerns, while in many cases, we were able to present evidence to the contrary.

We recognize that the current study is limited by the amount, quality, and diversity of the data used. In the context of this work, our sample consists of two classes (2020, 2021) where the number of examined students (24 and 20, respectively) was relatively low. In that respect, the entire studied population is the set of students who attended our program in its current version (approximately 150 students). Since the acceptance criteria

and students' demographics did not change significantly over these years, our sample is reasonably representative of our program. However, all examined students are from a single university (Bar-Ilan University) and all are from a single country (Israel), which may limit the generalization of our results to additional programs. In addition, our work focuses on student academic success and does not consider the employment status of our graduates, which could help in mitigating additional concerns expressed by prospective students. In future work we plan to investigate this issue with our graduates. This phase may be very challenging, since the very definition of LIS practitioner in the industry is not well defined, some students who attend our program are already employed in similar professions, and the employment possibilities in the LIS field vary over time. Last, our study focused on the DS competencies within the LIS training program. Future work will examine other components of the LIS training program using the same methodology of this study.

Ariel Rosenfeld is a senior researcher in the Information Science Department at Bar-Ilan University, Israel. Before joining Bar-Ilan, he was Koshland Postdoctoral Fellow at the Computer Science & Applied Mathematics Department, Weizmann Institute of Science, Israel. He is the recipient of the Victor Lesser Distinguished Dissertation Award for 2017. He obtained a PhD in computer science from Bar-Ilan University following a BSc in computer science and economics, graduated "magna cum laude," from Tel-Aviv University, Israel. His research focus is AI, information science, and scientometrics. Email: Ariel.Rosenfeld@biu.ac.il

Avshalom Elmalech is a distinguished faculty member in the Information Science Department at Bar-Ilan University, Israel. He has a profound passion for applied data science, and his research endeavors revolve around harnessing the power of data science to enhance various disciplines. His innovative work at the intersection of data science and diverse fields has contributed significantly to bridging gaps and unlocking new possibilities. Furthermore, he is an ardent advocate for effective pedagogical methods in teaching data science at the higher education level. His enthusiasm for cultivating the next generation of data scientists is evident in his commitment to developing innovative and accessible approaches to teaching this vital field. Through his research and dedication to education, he continues to make meaningful contributions to both the advancement of knowledge and the preparation of future data science leaders. Email: Avshalom.Elmalech@biu.ac.il

Acknowledgments

We would like to express our deep gratitude to department officials and especially Mrs. Ronit Barak for their help in assembling the data and insightful discussions.

Note

1 SAT scores are intended to measure a high school student's readiness for college (not graduate school) (Manhattan Review, 2022).

References

Adèr, H. J. (2008). *Advising on research methods: A consultant's companion*. Johannes van Kessel Publishing.

Agarwal, R., & Dhar, V. (2014). Big data, data science, and analytics: The opportunity and challenge for is research. *Information Systems Research*, *25*(3), 443–448. https://doi.org/10.1287/isre.2014.0546

Awan, S. E., Bennamoun, M., Sohel, F., Sanfilippo, F. M., & Dwivedi, G. (2019). Machine learning-based prediction of heart failure readmission or death: Implications of choosing the right model and the right metrics. *ESC Heart Failure*, *6*(2), 428–435. https://doi.org/10.1002/ehf2.12419

Bartschat, A., Reischl, M., & Mikut, R. (2019). Data mining tools. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *9*(4), e1309. https://doi.org/10.1002/widm.1309

Baumer, B. (2015). A data science course for undergraduates: Thinking with data. *The American Statistician*, *69*(4), 334–342. https://doi.org/10.48550/arXiv.1503.05570
Belkin, N. J., & Robertson, S. E. (1976). Information science and the phenomenon of information. *Journal of the American Society for Information Science*, *27*(4), 197–204. https://doi.org/10.1002/asi.4630270402
Borko, H. (1968). Information science: What is it? *American Documentation*, *19*(1), 3–5.
Bronstein, J. (2015). An exploration of the library and information science professional skills and personal competencies: An Israeli perspective. *Library & Information Science Research*, *37*(2), 130–138. https://doi.org/10.1016/j.lisr.2015.02.003
Brunner, R. J., & Kim, E. J. (2016). Teaching data science. *Procedia Computer Science*, *80*, 1947–1956. https://doi.org/10.1016/j.procs.2016.05.513
Burgoyne, J. A., Fujinaga, I., & Downie, J. S. (2015). Music information retrieval. In S. Schreibman, R. Siemens, & J. Unsworth (Eds.), *A new companion to digital humanities* (pp. 213–228). Wiley. https://doi.org/10.1002/9781118680605.ch15
Cao, L. (2017). Data science: Challenges and directions. *Communications of the ACM*, *60*(8), 59–68. https://doi.org/10.1145/3015456
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. https://doi.org/10.1111/j.1751-5823.2001.tb00477.x
Costa, C., & Santos, M. Y. (2017). The data scientist profile and its representativeness in the European e-Competence framework and the skills framework for the information age. *International Journal of Information Management*, *37*(6), 726–734. https://doi.org/10.1016/j.ijinfomgt.2017.07.010
Davenport, T. (2020). Beyond unicorns: Educating, classifying, and certifying business data scientists. *Harvard Data Science Review*, *2*(2). https://doi.org/10.1162/99608f92.55546b4a
Davenport, T. H., & Patil, D. (2012, October). 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*, 15–30. https://doi.org/10.1146/annurev-statistics-060116-053930
Dhar, V. (2013). Data science and prediction. *Communications of the ACM*, *56*(12), 64–73.
Dichev, C., & Dicheva, D. (2017). Towards data science literacy. *Procedia Computer Science*, *108*, 2151–2160. https://doi.org/10.1016/j.procs.2017.05.240
Doyle, A. (2019). Important job skills for data scientists. Liveabout dotcom. https://www.thebalancecareers.com/list-of-data-scientist-skills-2062381.
Eiteljorg, H., II (2004). Computing for archaeologists. In S. Schreibman, R. Siemens, & J. Unsworth (Eds.), *A companion to digital humanities* (pp. 20—30). Wiley. https://doi.org/10.1002/9780470999875.ch2
Ess, C. (2004). Revolution? What revolution? Successes and limits of computing technologies in philosophy and religion. In S. Schreibman, R. Siemens, & J. Unsworth (Eds.), *A companion to digital humanities* (pp. 132—144). Wiley. https://doi.org/10.1002/9780470999875.ch12
Fayyad, U., & Hamutcu, H. (2020). Toward foundations for data science and analytics: A knowledge framework for professional standards. *Harvard Data Science Review*, *2*(2). https://doi.org/10.1162/99608F92.1A99E67A
Forte, M. (2015). Cyber archaeology: A post-virtual perspective. In P. Svensson& D. T. Goldberg (Eds.), *Between humanities and the digital*. MIT Press. https://doi.org/10.7551/mitpress/9465.003.0027
Garber, A. M. (2019). Data science: What the educated citizen needs to know. *Harvard Data Science Review*, *1*(1). https://doi.org/10.1162/99608f92.88ba42cb
Gibert, K., Sànchez-Marrè, M., & Codina, V. (2010, July). Choosing the right data mining technique: Classification of methods and intelligent recommendation. International Congress of Environmental Modelling and Software. Ottawa, ON. https://scholarsarchive.byu.edu/iemssconference/2010/all/147
Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. In *IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)* (pp. 80–89). IEEE.
Guo, P. J. (2017). Older adults learning computer programming: Motivations, frustrations, and design opportunities. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (pp. 7070–7083). http://dx.doi.org/10.1145/3025453.3025945
Hajič, J. (2004). Linguistics meets exact sciences. In S. Schreibman, R. Siemens, & J. Unsworth (Eds.), *A companion to digital humanities* (pp. 79—87). Wiley. https://doi.org/10.1002/9780470999875.ch7
Hand, D. J.& Adams, N. M. (2014). Data mining. Wiley StatsRef: Statistics Reference Online. https://doi.org/10.1002/9781118445112.stat06466.pub2

Hjørland, B. (2002). Domain analysis in information science: Eleven approaches—Traditional as well as innovative. *Journal of Documentation*, *58*(4), 422–462. https://doi.org/10.1108/00220410210431136
Johnson, I. M. (1999). Librarians and the informed user: Reorienting library and information science education for the "information society". *Librarian Career Development*, *7*(4), 29–42. https://doi.org/10.1108/09680819910276941
Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260. https://doi.org/10.1126/science.aaa8415
Juznic, P., & Badovinac, B. (2005). Toward library and information science education in the European union: A comparative analysis of library and information science programmes of study for new members and other applicant countries to the European Union. *New Library World*, *106*(3/4), 173–186. https://doi.org/10.1108/03074800510587372
Kang, J. W., Holden, E. P., & Yu, Q. (2015). Pillars of analytics applied in MS degree in information sciences and technologies. In *Proceedings of the 16th Annual Conference on Information Technology Education* (pp. 83–88). https://doi.org/10.1145/2808006.2808028
Manhattan Review. (2022). The SAT as predictor of success in college. https://www.manhattanreview.com/sat-predictor-college-success/
Raschka, S. (2018). Model evaluation, model selection, and algorithm selection in machine learning. arXiv preprint arXiv:1811.12808. https://arxiv.org/abs/1811.12808
Rommel, T. (2004). Literary studies. In S. Schreibman, R. Siemens, & J. Unsworth (Eds.), *A companion to digital humanities* (pp. 88–96). Wiley. https://doi.org/10.1002/9780470999875.ch8
Russell, S. J., & Norvig, P. (2002). *Artificial intelligence: A modern approach*. Pearson.
Russo, D., & Zou, J. (2019). How much does your data exploration overfit? Controlling bias via information usage. *IEEE Transactions on Information Theory*, *66*(1), 302–323. https://doi.org/10.1109/TIT.2019.2945779
Schwab-McCoy, A., Baker, C. M., & Gasper, R. E. (2020). Data science in 2020: Computing, curricula, and challenges for the next 10 years. *Journal of Statistics Education*, *29*, 1–17. https://doi.org/10.1080/10691898.2020.1851159
Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge University Press.
Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632
Swamidason, I. T. J. (2019). Survey of data mining algorithm's for intelligent computing system. *Journal of Trends in Computer Science and Smart Technology*, *1*(01), 14–24. https://doi.org/10.36548/jtcsst.2019.1.002
Tariq, V. N., & Durrani, N. (2012). Factors influencing undergraduates' self-evaluation of numerical competence. *International Journal of Mathematical Education in Science and Technology*, *43*(3), 337–356. https://doi.org/10.1080/0020739X.2011.618552
Thomas, W. G. (2004). Computing and the historical imagination. In S. Schreibman, R. Siemens, & J. Unsworth (Eds.), *A companion to digital humanities* (pp. 56–68). Wiley. https://doi.org/10.1002/9780470999875.ch5
Urs, S. R., & Minhaj, M. (2022). Evolution of data science and its education in iSchools: An impressionistic study using curriculum analysis. *Journal of the Association for Information Science and Technology*, *74*(6), 606–622. https://doi.org/10.1002/asi.24649
Van der Aalst, W. M. (2014). Data scientist: The engineer of the future. In *Enterprise interoperability VI* (pp. 13–26). Springer.
Van Dyk, D., Fuentes, M., Jordan, M. I., Newton, M., Ray, B. K., Lang, D. T., & Wickham, H. (2015). ASA statement on the role of statistics in data science. *Amstat News*, *460*(9), 24.
Vellido, A. (2019). The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural Computing and Applications*, *32*(6), 1–15. https://doi.org/10.1007/s00521-019-04051-w
Vellido, A., Martín-Guerrero, J. D., Rossi, F., & Lisboa, P. J. G. (2011). Seeing is believing: The importance of visualization in real-world machine learning applications. In *Proceedings: 19th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN 2011: Bruges, Belgium* (pp. 219–226).
Wang, L. (2018). Twinning data science with information science in schools of library and information science. *Journal of Documentation*, *74*(2). https://doi.org/10.1108/JD-02-2018-0036
Williams, M. E. (1988). Defining information-science and the role of ASIS. *Bulletin of the American Society for Information Science*, *14*(2), 17–19.
Wing, J. M. (2006). Computational thinking. *Communications of the ACM*, *49*(3), 33–35.

Wu, D., Xu, H., Sun, Y., & Lv, S. (2022). What should we teach? A human-centered data science graduate curriculum model design for iField schools. *Journal of the Association for Information Science and Technology*, *74*(6), 623−640. https://doi.org/10.1002/asi.24644

Zaagsma, G. (2013). On digital history. *BMGN-Low Countries Historical Review*, *128*(4), 3–29.

Zhang, Y., Wu, D., Hagen, L., Song, I.-Y., Mostafa, J., Oh, S., Anderson, T. D., Shah, C. K., Bishop, B. W., Hopfgartner, F., Eckert, K., Federer, L., & Saltz, J. S. (2022). Data science curriculum in the iField. *Journal of the American Society for Information Science*. https://doi.org/10.1002/asi.24701

Zuo, Z., Zhao, K., & Eichmann, D. (2017). The state and evolution of US iSchools: From talent acquisitions to research outcome. *Journal of the Association for Information Science and Technology*, *68*(5). 1266–1277. https://doi.org/10.1002/asi.23751

Appendix: Our training program

Below is the list of all courses offered in Bar-Ilan University's LIS graduate program. All lecturers of the program were asked to provide a concise description of their course. We present the provided descriptions *unaltered*, in alphabetic order, as we believe they best reflect how the lecturers perceive their courses' contents.

- Advanced programming in Python: The aims of the course are to impart advanced knowledge and understanding of object-oriented languages, in addition to imparting knowledge in harvesting and presenting digital information.
- Data visualization: The course presents a theoretical framework to design and assess the effectiveness of data visualizations, letting the students practice visualization creation using several tools including Tableau, Python, and more.
- Digital humanities: This course introduces students to the discipline of digital humanities, demonstrating the use of technological tools and different methods for the study of the humanities. The course focus on different computer applications that are in use for: Text research, Tools for data visualization, Distant reading, Annotation tools, Digital Libraries and Humanities.
- Digital humanities seminar: The seminar deals with the historical development of digital humanities and various studies done in the field. The seminar provides tools and knowledge for research in the field.
- Geographic information: The purpose of the course is to bring the student into the world of geographical information science both theoretically and in practice. The course introduces GIS concepts and GIS tools used to visualize real-world features, discover patterns, and communicate information.
- Introduction to digitization of textual and non-textual information: This course serves as an introduction to the principles and functions that govern archive management. The course presents an historical perspective on the development of archives to the management of records in the modern state. Students learn different approaches to classifying archival functions and the actions required to support them. The course describes the concepts and technological issues related to the processing of government information and focuses on policies and practices related to the national State Archives.
- Introduction to programming in Python: The aim of the course is to impart a basic knowledge of programming principles as well as the ability to design and develop simple computer software in the Python language.

- Semantic web: The goal of this course is to introduce the main concepts of the semantic web (web 3.0), ontologies and basics of various semantic technologies developed in the last two decades such as XML, RDF/S, SPARQL. In addition, students learn about large semantic web projects and initiatives, such as schema.org, DBPedia, and Wikidata.
- Statistics: The aim of the course is to provide students with advanced knowledge in statistics and research methods, to develop abilities for advanced statistical analyzes that are suitable for scientific research work, including quantitative thesis work, development of critical thinking ability on quantitative research. The course presents the statistical software SPSS, focusing on regression and variance analysis (ANOVA), reading outputs and writing appropriate conclusions.
- Supervised learning: This course is intended to give student deep understanding of the main algorithms used in the field of supervised machine learning. The course is very practical with students using python to train and test models on various datasets.
- The evolution of text from manuscripts to digitization: The course prepares students to deal with questions that arise with the transition of text to digital format and the internet. The course includes the following topics: basic concepts in paleography and codicology, chapters in the history of the book, in handwriting, print and in the age of the internet and digitization, acquaintance with databases and projects that include various texts in the humanities, the importance of scientific editions and their creation, including digital scientific editions.
- Unsupervised learning: This course introduces students to principles underlying the development and implementation of big data solutions. The course provides extensive opportunities for hands-on application of big data principles and practices in the development, implementation and evaluation of data-driven solutions while focusing on unsupervised learning techniques.
- Web environment-standards and technologies: This course introduces students to the main standards governing content representation on the web such as HTML, XML, CSS, Unicode, and JavaScript. The students learn to combine these technologies to code and style HTML pages and augment them with JavaScript code to get a dynamic page behavior.