Equalizing Data Science Curriculum for Computer Science Pupils

Koby Mike
Faculty of Education in Science and Technology
Technion - Israel Institute of Technology
Haifa, Israel
kobymike@gmail.com

Tamir Hazan
Faculty of Industrial Engineering and Management
Technion - Israel Institute of Technology
Haifa, Israel
tamir.hazan@technion.ac.il

Orit Hazzan
Faculty of Education in Science and Technology
Technion - Israel Institute of Technology
Haifa, Israel
oritha@technion.ac.il

ABSTRACT

Data science is a new interdisciplinary field of research that focuses on extracting value from data. As an interdisciplinary science it integrates knowledge and methods from computer science, mathematics and statistics, and the domain knowledge of the data. As data science is still forming as a domain, several points of view exist on how to teach data science. The curriculum of many undergraduate data science programs includes advanced knowledge and skills in mathematics, statistics, computer science, and one or more data domains.

Several initiatives for designing high school data science curricula have emerged recently. Since it is unrealistic to teach all the above-mentioned advanced topics at the high school level, the high school data science curriculum focuses on a broad understanding of the data science workflow rather than on mathematical and algorithmic details that characterize undergraduate programs.

High school computer science pupils, however, are expected to have a deep understanding of algorithms, as algorithms are the heart of computer science. In this paper, we present our attempt to adapt a data science course to computer science high school pupils that incorporates both a broad view on data science and data workflow, as well as deep understanding of data processing algorithms and specifically, machine learning. This course is taught for 10th grade computer science pupils in an Israeli public school, and was also taught in a summer workshop for computer science teachers.

CCS CONCEPTS

• **Social and professional topics → Computing education**.

KEYWORDS

Data science education, computer science education, interdisciplinary, high school, secondary education, K12

Koli Calling '20, November 19–22, 2020, Koli, Finland

ACM ISBN 978-1-4503-8921-1/20/11…$15.00
https://doi.org/10.1145/3428029.3428045

ACM Reference Format:
Koby Mike, Tamir Hazan, and Orit Hazzan. 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 Calling '20), November 19–22, 2020, Koli, Finland*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3428029.3428045

1 INTRODUCTION

Data science is a new field of research that focuses on extracting insights, knowledge, and value from data. It is an interdisciplinary field that integrates knowledge and methods from computer science, mathematics and statistics, and the domain knowledge of the data. The radical growth in recent years in the availability of both data and 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, most of which are intended for undergraduate students, as either a major or a minor.

As data science is inherently interdisciplinary, there are many different ways to assemble a data science faculty or school [2]. Accordingly, many views on how to teach data science exist and recommendations for undergraduate curricula in data science have been published in recent years by several committees [5–7, 18].

Following the universities and the industrial demand, several initiatives to design a data science curriculum for high schools have emerged as well [9, 12, 15]. Data science education requires advanced knowledge and skills in mathematics, statistics, and computer science, as well as expertise in the data domain. Since it is unrealistic to teach all of these advanced topics at the high school level, data science curricula for high school focus on understanding what data is, how it is collected, and how it can be processed to solve real life questions and problems, rather than delving into the depth of mathematics and algorithms that characterizes undergraduate programs in data science.

At the same time, since algorithms are the heart of computer science [10], computer science high school pupils are expected to be familiar with the algorithmic facet of data science. Gal-Ezer et al. [10], who designed the foundation of the Israeli high computer science curriculum about three decades ago, state that "Its emphasis is on the basics of algorithmics, and it teaches programming as a way to get a computer to execute an algorithm" (p. 73).

In this spirit, we designed a data science curriculum that balances the need to teach both the breadth of data understanding, processing, and workflow, and the depth of the algorithmic facet of data science, specifically, machine learning algorithms. This curriculum was piloted with 10th grade computer science pupils in a 3-year pilot that involved teaching computer science classes in a public school, and a summer workshop in which computer science teachers were exposed to the program and learned it.

The rest of the paper is organized as follow: Section 2 reviews relevant works that have done so far on data science education in secondary education. Section 3 describes the curriculum and Section 4 describes the pilot and the research tools. In section 5 we describe the experience gained during the pilot implementation and in Section 6 we conclude.

2 BACKGROUND

The radical growth, in recent years, in the availability of both data and the computational resources required to process them, has led to a corresponding increase in the demand for data scientists [2]. As a result, new data science education programs have been opening at a growing rate [18]. Most of these programs are intended for undergraduate students, and vigorous discussions have been taking place regarding the appropriate details of these curricula [5–7, 18]. Undergraduate data science programs are usually 3 to 4 years long and include teaching a large body of knowledge from the fields of computer science, mathematics, and statistics. Some programs are designed as interdisciplinary programs and include the domain knowledge in one or more disciplines [1, 14, 16, 22].

There have also been several initiatives to teach data science to elementary, middle and high school pupils. The International Data Science in Schools Project (IDSSP) published a curricular framework for introductory data science developed for high schools [9]. IDSSP curriculum is planned for 240-360 hours spanned over two years. The first unit of the program (for 11th grade pupils) is aimed to develop pupil's enthusiasm for data science. It includes seven topics organized around the data cycle including data gathering, exploration, handling and communicating. The second unit (for 12th grade pupils) is aimed to introduce different data types, to introduce different ways of analyzing and to reinforce the different phases of the cycle of learning from data. As such, it includes 10 units offering introduction to different data type (e.g. text, images), different analysis types (e.g. machine learning), and different applications (e.g. recommendation systems). The IDSSP curriculum is planned for various populations of pupils, considering pupils who do not have programming skills, and thus, it uses tools that do not require programming.

Heinemann et al. [15] published a draft curriculum for a high school data science. The curriculum is planned for 10th to 12th grade pupils and is divided into four modules: (a) from data to information: an introduction to data and data science (b) big data and artificial intelligence: introduction to machine learning as a way to analyze data. (c) data projects: practical experience by working on realistic data sets. (d) data science and society: societal aspects of data projects. While Heinemann et al. [15] focus on the role of data and social issues, we focus on the algorithmic aspects of data science.

In Los Angeles, the Mobilize Introduction to Data Science program, currently being implemented in 45 schools, aims "to develop computational and statistical thinking skills so that students can access and analyze data from a variety of traditional and non-traditional sources" [12].

Other initiatives expose school pupils to the power of data science in the form of extra-curricular programs. Srikant and Aggarwal developed a half-day long data science tutorial for children in grades 5 through 9 [20], which includes developing a friend predictor and completing the full cycle of data application development using a spreadsheet software. Dryer, Walia and Chattopadhyay developed a data mining workshop based on seven modules of data mining, big data, ethics, and privacy [8] in which pupils used RapidMiner, a graphical environment, to analyze data. Bryant et al. [4] presented a one-week long programming camp for middle school pupils that emphasizes data science ideas and includes Python programming and data analysis tasks. Haqqi et al. [13] 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, project proposals, mentorship, field trips, and final posters and presentations. Mariescu-Istodor and Jormanainen designed a two-hour lesson on machine learning for pupils aged 13 to 19 [17] that uses only mathematical and computational knowledge that the pupils already acquired in other areas.

3 DATA SCIENCE CURRICULUM FOR COMPUTER SCIENCE PUPILS

The program described in this paper is a data science and machine learning program for 10th grade computer science pupils that is integrated into the current, official high school computer science curriculum in Israel. Since 1998, at which time the Israeli high school computer science curriculum was first implemented [11], the curriculum has been updated majorly only once, in 2010, in light of new developments in the field of computer science. In addition, on-going, minor updates are incorporated into the curriculum on a regular basis. These updates include both new content as well as adaptations for new populations, targeting younger audiences, mainly elementary and middle school pupils. Like the general computer science high school curriculum taught in Israel, the data science program is designed to accommodate two levels. The basic level is taught in the 10th grade and includes developing a project in Python as part of the lab-based learning unit in which students are exposed to other programming paradigms in addition to the main one – the object-oriented paradigm. In this paper we describe the basic level that is taught in the 10th grade. The extended level is taught in the 11th and 12th grades and is currently under development. The extended level elaborates on both the data science process and machine learning algorithms, with emphasis on deep learning.

There are several approaches to teaching machine learning, depending on the target audience [21]. While some courses focus on the mathematical and algorithmic aspects of machine learning, others focus on understanding how to use it. In either case, machine learning can be taught as a stand-alone course or as a module within a data science curriculum.

The data science curriculum, as derived by the above-mentioned curriculum guidelines, is based on the data life cycle. Pupils learn to ask questions about a specific topic, to collect, clean, explore and

Table 1: Data science for high school curriculum - Topics and number of hours

Topic	Hours
Introduction to data science	3
Image as data	9
Introduction to machine learning	3
The K-nearest neighbors (KNN) algorithm	9
Data and tables	9
The perceptron algorithm	9
The support vector machine (SVM) algorithm	9
Artificial neural networks	9
Final project development	30
TOTAL	90

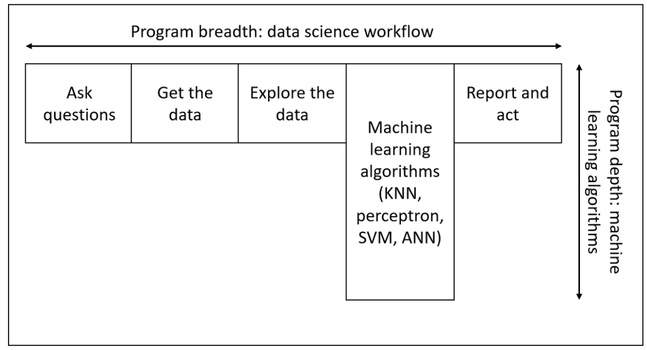

Figure 1: Data science curriculum with emphasis on machine learning

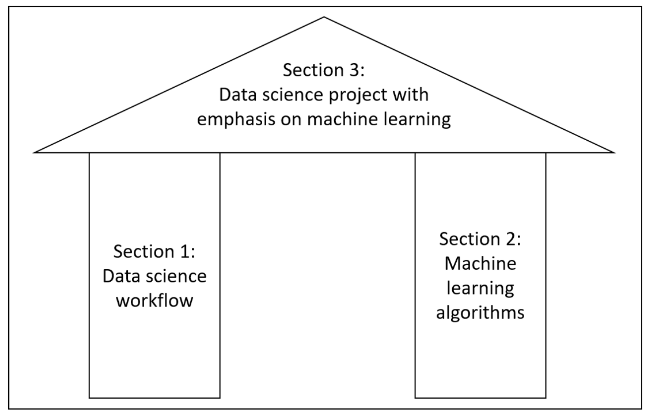

Figure 2: Data science projects with emphasis on machine learning

model relevant data, and to use these models to make decisions and to evaluate the appropriateness and quality of the predictions of the machine learning algorithms. As the mathematical and computational backgrounds of students are limited, deep mathematical or algorithmic aspects are skipped, and in some cases, pupils just learn to use the algorithms.

The rationale for the design of the data science curriculum for computer science pupils is to teach the data life cycle process while extending the level of algorithmics and programming in the machine learning unit (see Fig. 1). The curriculum includes four machine learning algorithms: K-Nearest neighbors (KNN), perceptron, support vector machine (SVM), and neural networks (NN). The first two algorithms are simple to understand and are taught mostly for pedagogical reasons (as explained below); the last two algorithms are useful algorithms in many academic and industrial applications. Since the curriculum is based on constructivism and active learning, the first two units of the curriculum – data science workflow and machine learning – are interlaced (see Table 1). The pupils first learn one data type (images) and one basic machine learning algorithm (KNN), thus enabling them to program and run their first machine learning model on data they collect themselves. The two units continue in parallel: additional data types and data processing methods are taught in parallel to more sophisticated machine learning algorithms.

The third section, project development, is based on the first two units (see Fig. 2). While the project focuses on the implementation of machine learning algorithms, the pupils experience all steps of the data science workflow[1], including asking questions, looking for and collecting data, data exploration, modeling, and reporting.

4 PILOT AND RESEARCH TOOLS

The pilot was conducted over the last three academic years (2017/8 to 2019/20), during which three different classes learned the data science program (one class per year). The learning environment is Google Colab[2] and code is written in Python.

In the design process of the curriculum, data was collected from pupils and computer science teachers using the following research tools:

(1) Observations in the pilot classes.
(2) Pupils' class work, final projects and reflections.
(3) Observations in a summer workshop for computer science teachers.
(4) Weekly questionnaires completed by the above-mentioned computer science teachers.
(5) Resarchers' diaries and reflections were documented and analyzed on a regular basis.

5 LESSONS FROM THE PILOT STAGE

In this section we highlight the lessons learned during the implementation phase with respect to the three units of the curriculum: data science workflow, machine learning and projects.

5.1 Unit 1 – Data Science Workflow

The first steps of the data science workflow include gathering the data and understanding its meaning. Many types of data exist: numbers, tables, text, images, sound, maps, geographic data, time

[1] We use the agile data science workflow as a reference. See https://academy.vertabelo.com/blog/agile-data-science-improve-workflow-with-scrum/
[2] https://colab.research.google.com/

series, and many more. Of this rich and diverse variety, we decided to start with images since:

(1) Images are naturally the most visual data type and thus, easier to understand.
(2) Images are represented in the computer as two-dimensional matrices of pixels. Accordingly, basic image processing operations, such as cropping or embedding an image inside another image, are good ways to practice many computational thinking and programing concepts, including arrays, multidimensional arrays, loops, conditions, and files. This unit is, therefore, useful for strengthening pupils' programming skills as well.
(3) The visual output of the image-processing exercises enables the pupils to demonstrate both their programming skills as well as their creativity.

We acknowledge that there will be data science students who are vision impaired and an image-based curriculum may not accommodate them. Thus, in addition, we plan to develop a module that focus on audio signals.

The final project was introduced in the second year of the pilot. This year we formally taught only the images data type and so all the pupils that year selected projects with images as data. To broaden the variety of possible projects, in the third year of the pilot, we added formal lessons on data exploration and processing of tables and, as a result, the projects in the third year were more diverse and included images, tables, and textual datasets.

5.2 5.2 Unit 2 - Machine Learning

Common types of machine learning algorithm tasks include classification, regression, and clustering. The curriculum for 10th grade focuses only on classification. Since 10th grade pupils learn about the different types and applications of machine leaning only on a basic introductory level, the other types of machine learning tasks are planned to be taught in the extended data science program intended for 11th to 12th grades.

The 10th grade curriculum includes several theoretical and practical topics regarding classification, such as underfitting and overfitting, model complexity, model evaluation, the confusion matrix, and other classification error metrics. To mitigate potential cognitive overload, these topics are introduced gradually during the school year. As mentioned, the curriculum covers four machine learning classification algorithms:

(1) The K-nearest neighbors (KNN) algorithm
(2) The perceptron algorithm
(3) The support vector machine (SVM) algorithm
(4) Artificial neural networks

The KNN and perceptron are simple algorithms to teach and learn and are part of the curriculum for pedagogical reasons. These algorithms are not commonly used for industrial or academic applications: KNN due to its high computational complexity, and perceptron due to convergence limitations. Thus, we also included in the curriculum two more powerful algorithms: SVM and ANN. While pupils wrote their own implementation of KNN and perceptron, we used the scikit-learn[3] Python library to implement SVM and ANN. We asked the teachers who participated in the summer workshop what order they think those algorithms should be taught in. All of twelve teachers who answered this question suggested that KNN should be taught first, 11 suggested that perceptron should be taught second, 8 suggested that SVM should be taught third, and 9 suggested that ANN should be taught last.

In more detail, the KNN algorithm is a simple and intuitive classifier whose implementation requires calculation of the distance between the instance to be classified and its neighbors, a procedure that requires the computation of vector distances. Even though the concept of vector distance is not part of high school mathematics, it is easy to show that it is just an extension of the two-dimensional Pythagoras theorem. Pupils can fully understand the algorithmic, mathematic, and programmatic aspects of the KNN algorithm and can implement it by themselves. The other algorithms included in the curriculum are more complex, due to either mathematical or algorithmic complexity. Thus, KNN is a good algorithm to start with, as it does not require mathematical knowledge that exceeds high school level knowledge and pupils can fully implement this algorithm programmatically.

The perceptron algorithm is a classification algorithm in and of itself and is also used as the basic building block of neural networks. Even though the learning algorithm of the perceptron is simple for implementation by a computer program, its mathematical proof requires linear algebra knowledge, and specifically, the dot product operator and the representation of a hyperplane as a vector, topics that are beyond the scope of high school mathematics. Moreover, since the mathematical proof is technical, it does not help pupils understand why this algorithm converges to a correct separation line (or plane in higher dimensions). Based on the feedback received from the computer science teachers who participated in our summer workshop, we replaced the mathematical proof with a visual animation of the convergence process. Many online resources exist for such animations [4] and pupils can also write their own animation program.

5.3 Unit 3 - Projects

Project-based learning is a teaching method in which learners solve problems taken from real-life situations [3] and which offers many advantages, such as active learning and increased motivation [19]. The project development work process includes most of the steps included in real data science projects: asking questions, data collection, feature extraction and data exploration, training and testing of models, and writing a final report.

We began the project phase by asking the pupils to propose their own ideas for project topics. This phase was not trivial for the pupils, as they had to propose (a) a topic for which they had access to data, and (b) a classification problem they could solve. Most pupils required two or more iterations of proposing ideas, formulating classification problems, and looking for appropriate data before the project topic was finalized. We found that class discussions about the topic selection process of each pupil were beneficial for the other pupils as well, as they helped them refine the formulation of their own project topics.

[3]https://scikit-learn.org/

[4]i.e. https://www.cs.utexas.edu/~teammco/misc/perceptron/

Eventually, several pupils selected projects that were related to their hobbies, such as basketball, robotics, and computer games. It was evident that these pupils had more domain knowledge than pupils who selected topics based on general interest. The average classification performance the pupils attained, as well as the final grades for the projects were not significantly different for pupils who selected topics based on their hobbies and pupils who selected topics based on their general interests. This phenomenon can be explained by the fact that even though pupils who selected topics based on their hobbies had more domain knowledge than the other pupils, their knowledge was not sufficient to significantly improve the classification performance. In a real-life data science classification project, classification performance is the crucial parameter for project success. In our case, however, since each pupil worked on a different problem with a different dataset, it was difficult to compare and grade the projects based on their classification performance. Thus, grades were given based on other parameters, mainly understanding of the problem, understanding of the steps required to process the data, and code quality.

6 CONCLUSION

Since data science is a new field, the curriculum for data science is still in the making and since many varied audiences are interested in data science, it is reasonable that a tailored curriculum be developed for each such audience or group of audiences. In this paper, we presented one such possible curriculum for computer science high school pupils.

In this paper, we reported on the pilot implementation and focused on the successful delivery of the curriculum. We plan to continue examining the curriculum implementation in the coming years as this program is taught by more teachers in more classes. Also, research is required to define and measure success parameters of data science curriculum.

REFERENCES

[1] Paul Anderson, James Bowring, Renée McCauley, George Pothering, and Christopher Starr. 2014. An undergraduate degree in data science: curriculum and a decade of implementation experience. In *Proceedings of the 45th ACM technical symposium on Computer science education*. ACM New York, NY, USA, 145–150.

[2] Francine Berman, Rob Rutenbar, Brent Hailpern, Henrik Christensen, Susan Davidson, Deborah Estrin, Michael Franklin, Margaret Martonosi, Padma Raghavan, Victoria Stodden, et al. 2018. Realizing the potential of data science. *Commun. ACM* 61, 4 (2018), 67–72.

[3] Phyllis C Blumenfeld, Elliot Soloway, Ronald W Marx, Joseph S Krajcik, Mark Guzdial, and Annemarie Palincsar. 1991. Motivating project-based learning: Sustaining the doing, supporting the learning. *Educational psychologist* 26, 3-4 (1991), 369–398.

[4] Caelin Bryant, Yesheng Chen, Zhen Chen, Jonathan Gilmour, Shyamala Gumidyala, Beatriz Herce-Hagiwara, Annabella Koures, Seoyeon Lee, James Msekela, Anh Thu Pham, et al. 2019. A Middle-School Camp Emphasizing Data Science and Computing for Social Good. In *Proceedings of the 50th ACM Technical Symposium on Computer Science Education*. ACM New York, NY, USA, 358–364.

[5] Andrea Danyluk, Paul Leidig, Lillian Cassel, and Christian Servin. 2019. ACM Task Force on Data Science Education: Draft Report and Opportunity for Feedback. In *Proceedings of the 50th ACM Technical Symposium on Computer Science Education*. ACM New York, NY, USA, 496–497.

[6] Richard D De Veaux, Mahesh Agarwal, Maia Averett, Benjamin S Baumer, Andrew Bray, Thomas C Bressoud, Lance Bryant, Lei Z Cheng, Amanda Francis, Robert Gould, et al. 2017. Curriculum guidelines for undergraduate programs in data science. *Annual Review of Statistics and Its Application* 4 (2017), 15–30.

[7] Yuri Demchenko, Adam Belloum, Wouter Los, Tomasz Wiktorski, Andrea Manieri, Holger Brocks, Jana Becker, Dominic Heutelbeck, Matthias Hemmje, and Steve Brewer. 2016. EDISON data science framework: a foundation for building data science profession for research and industry. In *2016 IEEE International Conference on Cloud Computing Technology and Science (CloudCom)*. IEEE.

[8] Amber Dryer, Nicole Walia, and Ankur Chattopadhyay. 2018. A Middle-School Module for Introducing Data-Mining, Big-Data, Ethics and Privacy Using RapidMiner and a Hollywood Theme. In *Proceedings of the 49th ACM Technical Symposium on Computer Science Education*. ACM New York, NY, USA, 753–758.

[9] Nicholas Fisher, Ajay Anand, Robert Gould, John Bailer ans Tim Hesterberg, James Bailey, Raymond Ng, Wesley Burr, James Rosenberger, Alan Fekete, Neil Sheldon, Alison Gibbs, and Chris Wild. 2019. *Curriculum Frameworks for Introductory Data Science*. http://www.idssp.org/files/IDSSP_Data_Science_Curriculum_Frameworks_for_Schools_Edition_1.0.pdf

[10] Judith Gal-Ezer, Catriel Beeri, David Harel, and Amiram Yehudai. 1995. A high school program in computer science. *Computer* 28, 10 (1995), 73–80.

[11] Judith Gal-Ezer and David Harel. 1999. Curriculum and course syllabi for a high-school CS program. *Computer Science Education* 9, 2 (1999), 114–147.

[12] Robert Gould, Moncado-Machado Suyen, Molyneux James, Johnson Terri, and Trusela LeeAnn. 2018. Mobilize: a Data Science Curriculum for 16-year-old students. *Iase-Web. Org* (2018), 1–4. https://iase-web.org/icots/10/proceedings/pdfs/ICOTS10_9B1.pdf?1531364299

[13] Saman Haqqi, Raja Sooriamurthi, Brian Macdonald, Cheryl Begandy, Judy Cameron, Berni Pirollo, Evan Becker, Jacqueline Choffo, Christopher Davis, Margaret Farrell, et al. 2018. Data jam: introducing high school students to data science. In *Proceedings of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education*. ACM New York, NY, USA, 387–387.

[14] Jessen Havill. 2019. Embracing the Liberal Arts in an Interdisciplinary Data Analytics Program. In *Proceedings of the 50th ACM Technical Symposium on Computer Science Education*. ACM New York, NY, USA, 9–14.

[15] Birte Heinemann, Simone Opel, Lea Budde, Carsten Schulte, Daniel Frischemeier, Rolf Biehler, Susanne Podworny, and Thomas Wassong. 2018. Drafting a data science curriculum for secondary schools. In *Proceedings of the 18th Koli Calling International Conference on Computing Education Research*. ACM New York, NY, USA, 1–5.

[16] Sami Khuri, Miri VanHoven, and Natalia Khuri. 2017. Increasing the Capacity of STEM Workforce: Minor in Bioinformatics. In *Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education*. ACM New York, NY, USA, 315–320.

[17] Radu Mariescu-Istodor and Ilkka Jormanainen. 2019. Machine Learning for High School Students. In *Proceedings of the 19th Koli Calling International Conference on Computing Education Research*. ACM New York, NY, USA, 1–9.

[18] Rajendra K Raj, Allen Parrish, John Impagliazzo, Carol J Romanowski, Sherif Aly Ahmed, Casey C Bennett, Karen C Davis, Andrew McGettrick, Teresa Susana Mendes Pereira, and Lovisa Sundin. 2019. Data Science Education: Global Perspectives and Convergence. In *Proceedings of the 2019 ACM Conference on Innovation and Technology in Computer Science Education*. ACM New York, NY, USA, 265–266.

[19] Bina Ramamurthy. 2016. A practical and sustainable model for learning and teaching data science. In *Proceedings of the 47th ACM Technical Symposium on Computing Science Education*. ACM New York, NY, USA, 169–174.

[20] Shashank Srikant and Varun Aggarwal. 2017. Introducing data science to school kids. In *Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education*. ACM New York, NY, USA, 561–566.

[21] Elisabeth Sulmont, Elizabeth Patitsas, and Jeremy R Cooperstock. 2019. What is hard about teaching machine learning to non-majors? Insights from classifying instructors' learning goals. *ACM Transactions on Computing Education (TOCE)* 19, 4 (2019), 1–16.

[22] Andrea Tartaro and Renee J Chosed. 2015. Computer scientists at the biology lab bench. In *Proceedings of the 46th ACM Technical Symposium on Computer Science Education*. ACM New York, NY, USA, 120–125.