Two Introductory Data-driven Activities for Secondary Schools

Giovanna Guerrini[1], Daniele Traversaro[1]

[1]DIBRIS, University of Genoa, Genoa, Italy

Abstract

This paper presents two data-driven activities aimed at secondary school students to promote data thinking skills, the first focusing on data literacy and the second on computer programming with a 'data-centric' pedagogy. We briefly introduce a Python package, called ToyPandas, that we developed to make coding more accessible to novice data science enthusiasts. The paper then discusses the motivations and challenges of providing an introductory data activity for students without a computing or data science background, and presents our experience, discussing some preliminary assessment results.

Keywords

Data science, Data education, Data-driven principles in secondary education, Data-centric introduction to computing

1. Introduction

Data science is rapidly gaining popularity in all fields of application, from science to art, engineering to business, and so on. Even our daily lives depend on decisions based on large amounts of data, which are seen as a resource for discovering valuable knowledge. This has created new business models, services, and jobs, and increased the demand for professional data scientists. As a result, data science curricula have been required in many disciplines in recent years [1, 2]. However, it has also raised a number of concerns, particularly about the ethical and political issues associated with the algorithms used to mine the data. These algorithms often lack transparency, making them "black box" systems that may contain inherent biases in the data, potentially leading to unfair outcomes. Additionally, they can pose a threat to users' privacy. As a result, data literacy has become essential for individuals in both their personal and professional endeavours [3] aligning with the European DigComp framework[1].

Data literacy can be defined as the ability to understand and critically evaluate information derived from authentic data [4]. It lies at the intersection of quantitative reasoning (i.e., the ability to apply mathematical principles to solve real-world problems), authentic context (i.e., the ability to develop an understanding of content knowledge from practice), and data science [4]. The third one is based on three pillars: computing, statistics/mathematics, and domain knowledge. In

SEBD 2023: 31st Symposium on Advanced Database System, July 02–05, 2023, Galzignano Terme, Padua, Italy

†These authors contributed equally.

giovanna.guerrini@unige.it (G. Guerrini); daniele.traversaro@dibris.unige.it (D. Traversaro)

0000-0001-9125-9867 (G. Guerrini); 0000-0002-0696-3671 (D. Traversaro)

[1]https://joint-research-centre.ec.europa.eu/digcomp_en

particular, recent research in data science education has highlighted the importance of cultivating a fresh problem-solving mindset that revolves around a data-driven approach, known as *data thinking* [5]. This has been defined as the integration of computational thinking, statistical thinking, and domain thinking.

In this discussion paper, we present two data-driven introductory learning initiatives that we have carried out over the past year to introduce data literacy and promote data thinking to non-major students. In particular, we describe two activities aimed at secondary school students to introduce them to data visualization and data transformation, introducing Python programming as a means to perform simple data science tasks. The remainder of the paper is organized as follows. In Section 2, we present the motivations for offering introductory data science activities for school students, as well as some challenges in offering such activities. In Section 3 we present some related work. In Sections 4 and 5 we describe the design and setting of the two data-driven initiatives and present some preliminary results. Finally, we conclude the paper in Section 6.

2. Motivations & Challenges

As mentioned above, as we live in the data age, it is fundamental that people have some *basic data skills in order to be active digital citizens*. Students should be aware of the challenges of data ethics and learn how to share, handle, communicate, and interpret data appropriately, regardless of their future careers. In addition, we believe that early exposure to data science topics in school can help to *embed data culture* in students and make them aware that data science has become ubiquitous due to its domain-specific nature. This is also reflected at university, where many non-computing academic programs (e.g. biology, business, medicine, etc.) have started to offer courses that cover data analytics topics, and students need to be aware of this. Furthermore, it is widely recognized that data thinking, like computational thinking, should be introduced to students as early as possible *to spark their interest and create a potential future generation of data scientists*, or at least data literate people.

These are the main motivations for offering data dissemination activities to our students. If we add the fact that it is difficult to find experienced data science teachers or support materials for data science education in (Italian) schools, the issue becomes even more important. Furthermore, *data science can be a means to promote and improve competency-based education with an interdisciplinary approach*. The use of real data can be exploited to design learning units -"Unità di apprendimento" (UdA) in Italian-, creating learning pathways that mix humanities, sciences, coding, and digital skills. In particular, authentic data can be relevant to students' interests and attitudes, which can have a positive impact on their creativity and motivation to learn. In addition, real data sets have been shown to promote data thinking in learning environments designed to develop computational thinking [5]. Therefore, computer science education can also benefit from the potential of data by paving the way for new interdisciplinary learning units where students can integrate data literacy with other soft and hard skills.

On the other hand, early data science activities lead to a number of **challenges** that need to be considered by educators and researchers. Firstly, *achieving the desired learning outcomes*, as data science education often requires some prior knowledge of computing and statistics/mathematics.

Coding and statistical models are important components of data science, and students typically need to learn a programming language and/or some statistical packages (e.g., R, Python, Pandas, NumPy, matplotlib, etc.). *The domain-specific nature of data science* is certainly an opportunity for learning, but could also be a challenge if not addressed in the right way: domain knowledge should be carefully assessed by choosing a dataset that requires a "right" level of knowledge, and with a view to an UdA, teachers from non-computing disciplines should contribute their domain knowledge. Finally, *time and curriculum constraints could be another important barrier to the dissemination of data science in schools*. In our experience, activities have to be embedded in a very short time frame of 10/20 hours maximum, which leads to several challenges in terms of providing the right balance between learning content and *hands-on practical activities*. The latter is a major component of data science education, but often requires a lot of time, as non-computing students often struggle with the technical and coding aspects.

Contribution. Our contribution aims to address the aforementioned challenges by designing two activities aimed at all secondary school students regardless of their background. We present a Python package based on Pandas, called `ToyPandas` (working title), to simplify the syntax and the notional machine (NM) behind Pandas, in order to focus only on the most important aspects for an introductory course in data-centric programming. We will also present a data literacy activity based on a Business Intelligence tool (e.g., Tableau) to introduce students to some relevant data topics, such as data analysis, interpretation, and visualization. Our aim is to give students a first experience in data science and, more generally, to make them aware of the importance of data in the development of a better society.

3. Related Work

Although data education is still in its infancy and rarely integrated into school curricula [6], several teaching units using programming languages or statistical software tools have been proposed in recent years. Bootstrap: Data Science [7] is a curriculum designed for use in integrated contexts at the secondary-school level, based on Pyret, a Python-like programming language for education with an "ad-hoc" NM and native support for tabular and image data. Other examples include IDS[2], a year-long course for high-school students using the R programming language for statistical computing; Passion-Driven Statistics[3], a data-driven CURE used in several US colleges and universities, using several statistical tools (e.g., R, SAS, Python, etc.); Berkeley's Data8 course[4] using Python with a module called `datascience`, designed by them and used to simplify the syntactical knowledge and the use of tabular data. Also, a unit based on survey data using the CODAP[5] online tool [8]; School 21[6] using Tableau; the Tableau's official data literacy courses for schools and academies[7]; and finally, an innovative curriculum where

[2] https://www.idsucla.org
[3] https://passiondrivenstatistics.wescreates.wesleyan.edu
[4] http://data8.org
[5] https://codap.concord.org
[6] https://www.tableau.com/community/blog/2021/8/tableau-brings-data-literacy-classroom-school-21
[7] https://www.tableau.com/it-it/learn/data-literacy

secondary school students learn about big data and its social impact using real multivariate data [9].

Our interventions are less ambitious than the above-mentioned curricula/courses, in fact, they are much shorter/compact in terms of learning content, due to the rather tight time constraints. Our initiatives are mainly designed for school *Pathway for Soft Skills and Educational Guidance* (known in Italy as PCTO). Nevertheless, it is possible to make some reflections/comparisons that will be presented in this paper in the respective sections dedicated to each presented activity.

4. Introduction to Data Literacy using Data Visualization Tools

We designed a teaching format on data literacy aimed at secondary school students that focused on data analysis and visualization. We have experimented with this unit twice, namely in a.y. 2021/2022 and 2022/2023 with grade-11 students of a high school major in humanities (*"Liceo Classico"* in Italian)[8] and was part of a school PCTO activity. In particular, students were enrolled in the *European Classical Lyceum* curriculum, which, compared to the traditional one, includes the study of economics, law, and two foreign European languages, as well as the strengthening of scientific disciplines.

The students had no previous background in computing or data science, so this was their first data-driven experience (taking into account both formal and informal learning). The teaching unit lasted approximately 20 hours, with 5 hours of lecture and laboratory, approximately 12 hours of a homework team project, and 3 hours of final plenary discussion/presentation with classmates and teachers.

Both editions used Tableau[9] software, a BI tool that provides a simple and colorful user interface and enables the creation and sharing of accessible dashboards and visualizations. Although there are many alternative solutions, including educational and open-source options such as Orange[10] and Inzights[11], we chose to use Tableau, primarily because of its intuitive drag-and-drop interface, which is particularly beneficial for our target audience of students with limited digital skills. In addition, Tableau offers a wealth of online resources and support materials that further facilitate the learning process. Finally, we did not consider programming-based solutions such as Python, R, or JavaScript with the D3.js library because the students had no previous experience of computer programming and acquiring such skills in the limited time available would have required a significant cognitive effort.

Unit Overview. The unit consisted of a five-hour lecture (seven for the first edition), consisting of a frontal lesson introducing data and a laboratory lesson using Tableau, a team project that students were expected to spend 12 hours a month on, and a final meeting where student teams could share and discuss their findings. In the lesson, we introduced students to the following topics: *Introduction to Data Science* (e.g. what data science is, big data as a valuable resource,

[8] The Italian *"Liceo Classico"* (literally *"Classical Lyceum"*) is a high school specializing in classical studies, where students study both Latin and Ancient Greek, with advanced curricula in philosophy, literature, and history.

[9] https://www.tableau.com/it-it/products/public We recommend Tableau Public, a freeware solution that only requires an e-mail account and is available both as a computer programme and as a cloud web-app.

[10] https://orangedatamining.com

[11] https://inzight.nz

the 5V model, what a dataset is); *basic Tableau data types* (numerical, textual, geographic) and qualitative/quantitative variables; *data ethics* (basic concepts of data fairness, security, and privacy); *data exploration and visualization* (history of data visualization, summary statistics, basic chart/plot techniques). The first edition also covered *data cleaning*, where students could learn about data quality issues and their causes.

During the hands-on activity, we used a toy dataset and emphasised the importance of effective communication and storytelling techniques when presenting data insights. We also encouraged students to consider ethical issues around data, such as privacy concerns, as Tableau Public shares all results with the wider community.

In the team project, using a collaborative learning approach, *students had to identify a research question based on a real dataset*[12], and from there to create and interpret appropriate visualizations to generate new information. Finally, they had to prepare a *multimedia presentation* to showcase their own data-driven project and discuss the technical expertise and the interpersonal skills they had acquired. The dataset was selected by us, in particular for the first and second edition of the activity we proposed the Olympics dataset[13] and the Economic Freedom Index dataset[14] respectively. Both datasets included various geographical variables, giving students the opportunity to explore and experiment with Tableau's geocoding features such as heat maps (chloropleths). The Olympic dataset required historical-sociological knowledge to explore some aspects (e.g., the evolution of women's participation or the impact of the World Wars on the competitions), while the Economic Index dataset required some economic/financial knowledge to understand the different indicators. This presented an opportunity for students to apply and deepen their understanding of subjects they had previously learned in other disciplines, such as economics, law, and history.

Preliminary Results. At the end of each edition, we collected team projects and responses to a final satisfaction questionnaire. The questionnaire was individual and consisted of the following questions: (DS) *How interested were you in the unit?*; (H1) *How challenging was it?*; (H2) *How much fun was it?*; (H3) *How useful was it?.* All items were rated on a Likert scale from 0 to 3, with options ranging from "not at all" to "very much" (note that H1 is a reverse item).

In particular, the first edition had a population of 28 students, while the second edition had 42 students. The first session was held remotely due to the Covid-19 health emergency, while the second session was held in person in a computer laboratory at the Department of Computer Science of the University of Genoa. The team projects were evaluated according to the following evaluation criteria: presentation quality and time management (multimedia presentation/slides, adherence with assignment deadline); data analysis (numbers of variables considered; types of data operations such as drill-down, filtering, etc.; the depth and structure

[12] In the second edition of the course, the research questions were pre-assigned at the request of the school.

[13] https://www.kaggle.com/datasets/heesoo37/120-years-of-olympic-history-athletes-and-results The Olympics dataset covered all the Games from Athens 1896 to Rio 2016 and contained more than 271 thousand rows and 15 columns. Each row corresponded to an athlete competing in an individual Olympic event and contained basic bio data on athletes and medal results.

[14] https://www.heritage.org/index/ The Economix Freedom Index dataset measures the economic freedom of 184 countries based on trade freedom, business freedom, investment freedom, and property rights. Specifically, we combined five years of observations from 2018 to 2022 into a single dataset to allow students to analyse the global economic prosperity over the past five years.

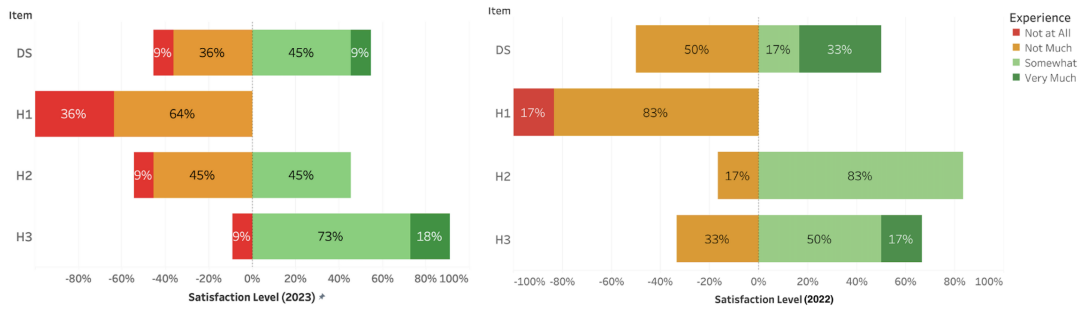

Figure 1: Gantt bar chart of Likert scale survey data (2023 vs. 2022 edition).

of the analysis in relation to the original research question); communication with data (effective use of maps/charts, presence of labels/legends, interactivity in dashboards, etc.), and domain-knowledge (quality of in-depth study, cited sources, etc.). The scores for each dimension were given in decimals, and the final score was calculated as a weighted sum of the different criteria. Presentation, data analysis, and communication had the same weight (0.3), while domain knowledge had a weight of 0.1.

In the first edition, all team projects were of good quality and students seem to appreciate and grasp the potential of a data-driven approach as well as its application in their personal and professional daily lives, as already presented in [10]. In the second edition, the quality of projects has been more erratic. In total, we assessed 11 team projects. The mean was 6.9 and the median was 7. More specifically, we gave three groups a mark of 6, six groups a mark of 7, and two groups a mark of 8.

Figure 1 shows the results of the final questionnaires. Only 6 and 11 students in 2022 and 2023 respectively completed the questionnaire. The median of the item DS is 1.5 [2 in 2023], while the medians of the three items related to the homework (H1, H2, H3) are 1, 2, and 2 respectively [1, 1 and 2 in 2023]. In general, the activity was perceived as difficult in both editions (all H1 responses were negative), but more so in the second, although it was perceived as more useful. The level of interest in the activity is quite similar, although the first edition had a higher percentage of very positive responses. We did not conduct a statistical control between the two sessions as several variables were changed (remote vs. face-to-face, data cleaning phase, dataset change, pre-set research question), some of which were beyond our control.

In an informal way, we can measure the variability between the two editions based on the anonymous free comments from the 2023 questionnaire. Specifically, about half of the participants expressed a preference for working with a different dataset (one that is not economic/legal) and/or having the opportunity to propose their own dataset. Two respondents would also have liked a lesson on multimedia presentations. Finally, one person suggested splitting the theory and practical sessions into separate days. We believe that a data-driven activity will spark curiosity and motivation if the topic is of interest to the student, so in the future, we will propose a number of datasets to choose from and leave the students free to devise their research questions.

5. Introduction to Programming with a Data-Centric Approach

This second contribution stems from a desire to experiment with *data-centric computing* pedagogy in computer programming [11]. In the literature, to our knowledge, the main contributions of this approach are *Berkeley's Data 8* and *A Data-Centric Introduction to Computing* [11]. The latter defines data-centric computing pedagogy as "data science + data structure" (in that order, not commutatively), i.e. a pedagogy that starts with rudimentary data science and flows into more traditional data structures. Both curricula used a property-coding tool developed by the authors, `datascience` and `PyRet` respectively. The former is a Python library that provides native support for tabular data with an easy-to-learn syntax, while PyRet is an educational programming language where image and tabular data types are first-class citizens. Both are courses/curricula that aim to provide students with a good programming foundation that can be applied to data science. Instead, our proposal is to introduce a Python library called ToyPandas (working title), specifically designed for short courses of about 10 hours, divided into two 5-hour lessons. As a result, our learning objectives differ from those of a comprehensive data science curriculum. However, we believe that our library has potential for longer introductory data science activities. It serves as an entry point for beginners to explore data science using Python and a simplified version of the popular Pandas library. Below, we outline the key features of ToyPandas.

ToyPandas. `ToyPandas` is a Python package based on Python Pandas, which *pedagogical aid is to make programming more accessible to beginners thanks to a less complex NM and an easier to learn syntax.* The syntax is simplified but similar to that of the original Pandas, so students can learn the basics and then move on to Pandas, finding a familiar environment and syntax, without having to create a new mental model of the NM from scratch, but rather enriching their existing one. For example, to create a dataframe starting from two series:

```
s1 = series(2, 0, 5, 9)
s2 = series("Bob", "Alice", "Ted", "Carol")
df = concat(s1, s2)
df.show()
```

The syntax and the NM behind are simpler. In original Pandas, you can define series starting from iterables, dictiories, or scalar values (e.g., with `np.arange()`), so the original instruction might be something like `s1 = pd.Series([2, 0, 5, 9])`. This approach assumes prior knowledge of basic Python data types such as lists. However, this can be confusing for beginners who are just starting to learn about series, as they may perceive series as being similar to lists. This can lead to a higher cognitive load as beginners need to grasp multiple concepts at once. We decided to hide an important feature of Pandas, namely the hashable indexes and multi-level/hierarchical indexes. In our library, series are designed to be more similar to Python lists, as they can only have numeric indexes. In addition, when an element is removed from a series, the resulting series still maintains a sequential order of numeric indexes. In contrast, the original library requires manual adjustments to fix the "holes" in the indexes, as they are treated as hashable keys in a map.

Another distinction lies in the data types. In our library, no prior knowledge of Python or NumPy iterable data types is required. Text and heterogeneous series are referred to by different names - "string" and "object" - rather than using a single identifier called "object". Integer and floating-point columns are simply referred to as "int" and "float" respectively, without specifying the bitsize of the type in the column name (e.g., int32, int64, float64 etc.). In general, the syntax of ToyPandas is more expressive and provides clearer semantics. For example, we have simplified the way we access the dataframe elements (`.indexlocate[]` instead of `.iloc[]`). We also made no distinction between `.loc[]` and `.iloc[]`. Students can also learn about iteration, functions, and Boolean conditions by applying the function to an entire series or by querying one or more columns of a dataframe (using the bitwise operators) using the original syntax (e.g. `df[df[(Age > 18) & (Name == "Alice")]]`). The iteration over rows is somehow implicit (a bit like SQL queries)[15].Finally, we have overridden several Pandas methods for series and dataframes, including those for handling missing values (NaN).

From a technical point of view, we created subclasses of Pandas classes (e.g., Series, Dataframe, pd.core.indexing._iLocIndexer etc.). These subclasses simplify tasks such as column renaming, duplicate removal, missing values handling, element access and function calls (arguments, return values, etc.). Furthermore, our classes remain compatible with data visualisation using Matplotlib.

Our package is still a work in progress, but we aim to publish it on PyPI by summer 2023. In April we will test it with 30 secondary school students in a PCTO activity. The unit will consist of two lectures (10 hours in total) and will cover the basics of data and computational thinking: Introduction to data science (what is data science, data ethics); Introduction to algorithms and programming (flowchart, program, executor, etc.); Basic data types (int, float, bool, string, series, dataframe) and their main operations; Starting with data (loading data from .CSV files, understanding data summary); Data exploration and analysis (filtering and retrieving relevant data using conditional expressions with logical operators); Functions (applied to columns, to transform data); Data visualization (using Matplotlib). We plan to evaluate the activity in terms of the pedagogical effectiveness of our personalized, data-centric approach.

6. Conclusions

We presented two data-driven initiatives to promote data literacy and data thinking in secondary schools: the first was based on a data visualization tool, while the second was based on Python programming using a proprietary package to help students perform data science coding techniques more intuitively. We plan to release this package by summer 2023.

As current and future work directions, we mention activities tailored to stress ethical aspects, e.g., in terms of coverage, when applying data transformations and an ongoing project *From botany to Big Data*[16] to develop dashboards analyzing environmental data acquired in various forms observing the plants in school gardens, so to demonstrate how to use digital devices for making the school garden a laboratory for the environment.

[15] More specifically, many Pandas operations are vectorised to improve performance, but this optimization detail is beyond the scope of our introductory intervention.

[16] competenzedigitali.unige.it/pon-green

References

[1] A. C. Bart, D. Kafura, C. A. Shaffer, E. Tilevich, Reconciling the promise and pragmatics of enhancing computing pedagogy with data science, SIGCSE '18, Association for Computing Machinery, New York, NY, USA, 2018, p. 1029–1034. URL: https://doi.org/10.1145/3159450.3159465. doi:10.1145/3159450.3159465.

[2] A. Danyluk, P. Leidig, L. Cassel, C. Servin, Acm task force on data science education: Draft report and opportunity for feedback, SIGCSE '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 496–497. URL: https://doi.org/10.1145/3287324.3287522. doi:10.1145/3287324.3287522.

[3] A. Grillenberger, R. Romeike, Developing a theoretically founded data literacy competency model, in: Proceedings of the 13th Workshop in Primary and Secondary Computing Education, WiPSCE '18, Association for Computing Machinery, New York, NY, USA, 2018. URL: https://doi.org/10.1145/3265757.3265766. doi:10.1145/3265757.3265766.

[4] M. K. Kjelvik, E. H. Schultheis, Getting messy with authentic data: Exploring the potential of using data from scientific research to support student data literacy, CBE—Life Sciences Education 18 (2019) es2. URL: https://doi.org/10.1187/cbe.18-02-0023. doi:10.1187/cbe.18-02-0023. arXiv:https://doi.org/10.1187/cbe.18-02-0023, pMID: 31074698.

[5] K. Mike, N. Ragonis, R. B. Rosenberg-Kima, O. Hazzan, Computational thinking in the era of data science, Commun. ACM 65 (2022) 33–35. URL: https://doi.org/10.1145/3545109. doi:10.1145/3545109.

[6] L. Van Audenhove, W. Van den Broeck, I. Mariën, Data literacy and education. introduction and the challenges for our field 12 (2020) 1–5. doi:10.23860/JMLE-2020-12-3-1.

[7] E. Schanzer, N. Pfenning, F. Denny, S. Dooman, J. G. Politz, B. S. Lerner, K. Fisler, S. Krishnamurthi, Integrated data science for secondary schools: Design and assessment of a curriculum, in: Proceedings of the 53rd ACM Technical Symposium on Computer Science Education V. 1, SIGCSE 2022, Association for Computing Machinery, New York, NY, USA, 2022, p. 22–28. URL: https://doi.org/10.1145/3478431.3499311. doi:10.1145/3478431.3499311.

[8] D. Frischemeier, R. Biehler, S. Podworny, L. Budde, A first introduction to data science education in secondary schools: Teaching and learning about data exploration with codap using survey data, Teaching Statistics 43 (2021) S182–S189. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/test.12283. doi:https://doi.org/10.1111/test.12283. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/test.12283.

[9] E. Gil, A. L. Gibbs, Introducing secondary school students to big data and its social impact: A study within an innovative learning environment, in: Promoting understanding of statistics about society. Proceedings of the Roundtable Conference of the International Association for Statistics Education (IASE), 2016.

[10] G. Guerrini, D. Traversaro, Introduction to data literacy with tableu in high school, La trasformazione digitale nella Scuola, negli ITS, nell' Università e nella formazione professionale., Milan, Italy, 2022, pp. 239–245. URL: https://www.aicanet.it/documents/10776/4555506/ATTI+Didamatica+2022/469726b1-58d5-4f48-8a8f-0fa1d2fb0f34.

[11] S. Krishnamurthi, K. Fisler, Data-centricity: a challenge and opportunity for computing education, Communications of the ACM 63 (2020) 24–26.