International Society of the Learning Sciences

What Data is in K-12 Data Science? An Analytic Approach to Understanding the Data Used in K-12 Data Science Courses

Rotem Israel-Fishelson, Peter F. Moon, Rachel Tabak, David Weintrop
rotemisf@umd.edu, pmoon@umd.edu, retabak@umd.edu, weintrop@umd.edu
University of Maryland

Abstract: To help students understand the role of data in their lives, it is important to introduce them to foundational data science. This paper presents an analysis of the data used in YouCubed's Explorations in Data Science curriculum. This work provides an analytic understanding of how data science is situated in students' lives and provides insights into ways curricula are preparing them for a data-rich world.

Introduction

As the role that data play in society grows, it is increasingly important to introduce foundational concepts and practices of data science to K-12 students (LaMar & Boaler, 2021). The last five years have seen the emergence of several year-long data science curricula (e.g., IDSSP Curriculum Team, 2019) that are playing an important role in defining what constitutes data science at the high school level. Given the ubiquity of data in the lived experiences of today's high school students, there is a tremendous opportunity to situate data science instruction in ways that draw on learners' interests, experiences, and cultures. Understanding the types of datasets being used in data science curricula and their alignment with the experiences of learners is an open question. In this work, we propose an approach for answering the overarching question: *What data is being used in K-12 data science curricula?* The proposed approach attends to what data is used, when it is from, its size, and how it relates to learners. To demonstrate this approach, we analyzed the YouCubed Explorations in Data Science curriculum (YouCubed, 2022), a year-long, freely available, and widely used data science curriculum. This work seeks to shed light on the nature of the datasets that are at the heart of a high school data science curriculum. In doing so, it contributes to our understanding of how best to introduce K-12 students to data science.

Analytic approach

To understand the data in K-12 data science curricula, we developed an analytic approach that attends to four interrelated measures (below). These measures were then applied to the 56 datasets identified in YouCubed.

- Topic - the ideas and topics represented, based on the Bootstrap:Data Science curriculum (Schanzer et al., 2022): *Sports, Politics, Media & Entertainment, Environment & Health, Education, Nutrition,* and *Other*.
- Recency - when the data is from, ranked by the following coding scheme: *Fresh* (just-created data), *Recent* (data from the last 3 years), *Past decade*, *Older than a decade*, and *Not relevant* (fictitious data/ no date).
- Size - the number of observations or entries in the dataset. We classified the datasets into five sizes: very small (<25), small (25-100), medium (100-1,000), large (1,000-10,000), and very large (>10,000).
- Proximity – how the datasets relate to learners, adopted from Lee and Delaney (2022). We used a slightly modified 5-point scale, ranging from 0-4, with 0 describing content-agnostic data, 1 describing fictional data, 2 describing data about a topic that might be familiar to some but not all students, 3 describing data on niche topics, and 4 capturing data that students collected about themselves and their peers.

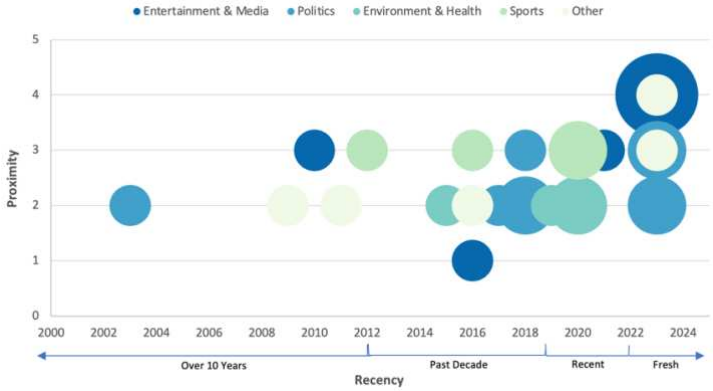
Findings

Our analysis revealed that the most common topic, comprising 16 of the 56 datasets, was Entertainment & Media. The next most common were Politics (12 datasets), with topics such as demographics and economics, followed by Environment & Health (9 datasets), and Sports (5 datasets). Regarding recency, the results show that most datasets are relatively recent, with 25 of 56 being from the last 3 years, with 13 of those datasets coded as *fresh,* meaning the data was/will be collected by the students. In analyzing datasets' sizes, we found that most of them were relatively small, with 13 Very Small datasets and 18 Small. In addition, the curriculum includes 10 Medium, 2 Large, and 4 Very Large datasets. Our analysis reveals that the curriculum includes datasets that fall across the full range of proximity levels, with a majoring of datasets rated as 2 or 3, meaning they are using real data. Specifically, 28 datasets are not related directly to the lives of students (level 2), and 18 datasets cover topics rated as more directly relating to students (level 3). In investigating the intersection of multiple categories, we found that datasets dealing with Entertainment & Media and Environment & Health tend to be relatively small. At the same time, topics related to Politics include very large datasets. In looking at the interaction of Proximity and

Topic, we can see that some topics like Politics and Environment & Health had a high level of datasets ranked as Level 2, which means they used real data but were not closely associated with the students' knowledge or interests. On the other hand, most Entertainment datasets were categorized as Levels 3 and 4 representing data that was either collected by the students or directly related to their personal lives. This analysis shows the types of topics that are close to students and identifies opportunities for potentially revising datasets within a curriculum to help make datasets and materials of greater interest to students. The final intersection we highlight is the relationship between Proximity and Recency, which sheds light on the question of whether more recent datasets are more *proximate* to students. Figure 1 illustrates that the answer to this question is generally yes. This can be seen in the trend where the higher the proximity level, the more recent the data is. This trend exists across all analyzed topics.

Figure 1

The Proximity of datasets in the YouCubed Explorations in Data Science curriculum, organized on a timeline. The bubble size represents the number of datasets with that Proximity/creation date.



Discussion and conclusion

This work presents a set of analytic lenses to deepen our understanding of the data used to introduce K-12 students to the field of data science. In attending to the topic, recency, size, and proximity of the data, we get a picture of what data are used in the YouCubed curriculum and see how it succeeds in incorporating datasets that are recent, *proximate*, and span various topics. At the same time, the analysis reveals opportunities for improvement, such as possibly revisiting the dataset coded as Politics to either make them more proximate to the learner, more recent, or replace them with datasets from other categories. This proposed analytic framework can be useful to curriculum developers to ensure that their materials incorporate various topics that are relevant to learners. Similarly, an educator could use this coding scheme to identify potentially problematic or less interesting datasets to make a more engaging curriculum. For researchers, this framework may be useful to understand the characteristics of a curriculum or as a way to compare different curricula. As the presence and importance of data in society continue to grow, it is important that all students develop a basic understanding of data science. A key aspect of this is developing effective, engaging, and equitable data science learning opportunities, which includes creating curricula with datasets that engage students. This paper builds on prior research to further develop an analytic framework that researchers, designers, and educators can use to ensure that K-12 data science curricula prepare students to be data-literate citizens.

References

IDSSP Curriculum Team. (2019). *Curriculum frameworks for introductory data science*.

LaMar, T., & Boaler, J. (2021). The importance and emergence of K-12 data science. *Phi Delta Kappan*, *103*(1), 49–53.

Lee, V. R., & Delaney, V. (2022). Identifying the content, lesson structure, and data use within pre-collegiate data science curricula. *Journal of Science Education and Technology*, *31*(1), 81–98.

Schanzer, E., Pfenning, N., Denny, F., Dooman, S., Politz, J. G., , … Krishnamurthi, S. (2022). Integrated data science for secondary schools: Design and assessment of a curriculum. *Proc. of SIGCSE 2022*, 22–28.

YouCubed. (2022). Explorations in data science. YouCubed High School Data Science Course.

Acknowledgments

This research is supported by the National Science Foundation (Award # 2141655).