Interest-Driven Data Science Curriculum for High School Students: Empirical Evidence from a Pilot Study

Rotem Israel-Fishelson
University of Maryland
rotemisf@umd.edu

Peter F. Moon
University of Maryland
pmoon@umd.edu

David Weintrop
University of Maryland
weintrop@umd.edu

ABSTRACT

This paper presents a pilot study of an interest-driven data science curriculum for high school students. The curriculum uses authentic and meaningful data exploration activities to situate data science in students' lived experiences. The curriculum aims to lay the computational foundation of data science and equip students with the necessary skills and practices to become informed and active citizens in our data-driven world. The pilot study, conducted in two sections of a computer science class, demonstrates the curriculum's inquiry-based approach, which allows students to formulate questions based on their interests and answer them by manipulating publicly available datasets. The study illustrates how a block-based learning environment and API data retrieval can be harnessed to support data science learning activities that situate the topics in learners' lived experiences and create an engaging learning experience. The study advances our understanding of ways to use novel technologies to introduce learners to data science, emphasizing the practical implications of using authentic data and the inquiry-based approach in curriculum design.

CCS CONCEPTS

• **Social and professional topics** → Professional topics; Computing education; K-12 education.

KEYWORDS

Data Science Education, Interest-Driven Curriculum, High School Students

ACM Reference Format:
Rotem Israel-Fishelson, Peter F. Moon, and David Weintrop. 2024. Interest-Driven Data Science Curriculum for High School Students: Empirical Evidence from a Pilot Study. In *Interaction Design and Children (IDC '24), June 17–20, 2024, Delft, Netherlands*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3628516.3659416

1 INTRODUCTION

Today's youth are increasingly immersed in a data-driven world where information and technology are central to their daily lives. They generate and consume a vast amount of data when they post on social media, choose their streaming content, and use learning environments. As youth navigate this digital landscape, they encounter multiple data sources that constantly influence and shape their experiences. Therefore, understanding the role of data and its impact on their lives is an essential literacy [8]. Youth need the necessary knowledge and skills to analyze, evaluate, and draw meaningful insights from the data [5]. In response to these needs, data science education has gained momentum, and there is a growing call to teach data literacy and foundational data science concepts in K-12 contexts [13, 25].

In recent years, several educational initiatives have emerged to promote data science in primary and secondary schools [21]. One such initiative is the International Data Science in Schools Project (IDSSP). This cross-disciplinary and international project provides training programs and frameworks for developing data science courses [11]. These learning experiences employ a variety of novel technologies and technological tools to make data and data science concepts accessible to younger learners [19]. However, the discipline is still in its early stages, as different stakeholders are working to define what data science entails, what tools and technologies are most effective, and what should be included in the curriculum [24].

Data science education aims to provide students with the core technical skills to analyze datasets, investigate phenomena, pursue questions, and draw conclusions based on the data [25]. Moreover, it aims to equip students with the ability to think critically, make informed decisions, promote fair data practices, and contribute to building a more equitable and inclusive digital world [2]. By harnessing data that align with students' interests and cultural backgrounds, data science education can create captivating learning experiences and equip students with the skills and knowledge to become informed citizens in our data-driven world [10]. The continuous interaction of students with data provides a compelling opportunity to situate data science in their lived experiences [26]. Integrating real-life data that connects to learners' personal experiences, beliefs, and interests can make the learning experience more relevant and meaningful [15] and increase the likelihood of knowledge acquisition [3].

Here, we present our effort to develop and evaluate an introductory interest-driven data science curriculum for high school students. We outline the guidelines for our curriculum aiming to anchor data science in students' lived experiences and present empirical evidence from a pilot conducted in two computer science (CS) classes. We demonstrate this by showing the data exploration activity, which encouraged students to pursue interest-driven data science practices. The work advances our understanding of how to use novel technologies and tools to introduce youth to essential, foundational data science concepts and practices and, in doing so, prepares them to succeed in a data-rich world.

IDC '24, June 17–20, 2024, Delft, Netherlands

ACM ISBN 979-8-4007-0442-0/24/06
https://doi.org/10.1145/3628516.3659416

Table 1: Topics Covered in Each Curricular Unit

Unit 1: Data in Learners' Lives	Unit 2: Computational Foundations of Data Science	Unit 3: Data Science Practices
1.1 Introduction to Data	2.1 What is Data Science	3.1 Intro to Data Visualization
1.2 Data Collection and its Purpose	2.2 Manual Data Processing	3.2 Exploratory Analysis with CODAP
1.3 Using Data: the DIKW Model	2.3 Intro to Programming with EduBlocks: Filtering & Data Transformation	3.3 Graphs and Figures: One Variable
1.4 Sources of Data	2.4 Accessing Data with APIs using RapidAPI	3.4 Graphs and Figures: Two Variables
1.5 Evaluating Datasets	2.5 Preparing Data for Analysis	3.5 Statistical Testing
1.6 Data Collection: Impact and Equity	2.6 Data Analysis in Practice	3.6 Linear Models

2 INTEREST-DRIVEN DATA SCIENCE CURRICULUM

"API Can Code" is an interest-driven introductory data science curriculum that introduces high-school students to the computational foundations of data science through authentic, meaningful data exploration. The project is being carried out in close collaboration with an urban public charter high school as part of a research-practice partnership. The curriculum is grounded in the Interest Development Theory, which highlights that students' interests thrive when they can explore, interact with, and gain meaning from the subject they are interested in [17, 20]. Moreover, when the educational experiences are customized to the learners' backgrounds, they can utilize their prior experience and knowledge, express their viewpoints, and feel empowered to participate in the learning activities [1]. The central idea of this curriculum is anchored in the data science cycle [11]. Students are encouraged to formulate questions based on their interests, identify relevant datasets, programmatically manipulate and analyze the data, and communicate their findings. The interest-driven nature of this inquiry approach provides essential context and meaning for the data [16].

The curriculum is divided into three units, each containing six lessons (Table 1). The first unit aims to help students understand how data impacts their lives. In this unit, students gain an understanding of data and explore the entities that collect it. They are also introduced to the Data-Information-Knowledge-Wisdom (DIKW) model [22]. Moreover, they learn about data sources, ways to evaluate datasets, and how data can influence equity and algorithmic bias. The second unit is designed to help students gain foundational computational skills to programmatically retrieve and manipulate publicly available data from diverse Application Programming Interfaces (APIs). This is done using EduBlocks, a block-based programming tool designed to introduce text-based programming languages, like Python, in a user-friendly and engaging manner [6]. EduBlocks includes blocks that facilitate sending requests to external sources, allowing students to write programs to call APIs and collect data to answer their driving questions. Students are introduced to the field of data science and then practice ways to access and manipulate data from various APIs. The third unit centers on data science practices, including analyzing and visualizing data to extract valuable insights. Students use CODAP to perform data analysis practices, create and interpret a variety of summary plots, and perform statistical tests. CODAP, or Common Online Data Analysis Platform, is a free, user-friendly data visualization, analysis, and exploration tool [4]. Each unit is anchored with data exploration activities that align with students' interests as identified in preliminary research among high school students [12]. The curriculum culminates in a final project where students go through the full DIKW sequence, starting with a driving question, then identifying a relevant API to query, writing a program to manipulate the data, and finally creating a visualization to shed light on their topics.

3 METHODS

3.1 Research Settings and Data Analysis

We conducted a pilot study with two CS classes at a US school to test, refine, and improve our curriculum. The CS teacher underwent weekly professional training with a team of two researchers, where they jointly reviewed the curricular materials, including structured lesson plans, presentations, and worksheets in Google Forms. Each lesson lasted 90 minutes and took place in the same classroom. The first author observed the lessons and took field notes. Additionally, semi-structured interviews were conducted with a sample of five students after each unit. The interviews were audio recorded, transcribed, and analyzed using an open coding approach to identify emerging themes [23]. Furthermore, all artifacts created during the lessons (i.e., EduBlocks programs, CODAP charts, and worksheets' responses) were collected for analysis from students who signed an assent form and provided signed consent from their guardians. The IRB board of the University approved research protocols and data collection methods.

In this paper, we showcase the emerging findings of a lesson in the second unit of the curriculum. It demonstrates the pedagogical approach used to illustrate how the technologies employed can help accomplish the goal of developing an interest-driven data science curriculum. The findings section details the learning activity given to the students, the technological infrastructure used, and the findings derived from the analysis.

3.2 Participants

A total of 41 12th-grade students participated in the curriculum lessons. Parent/guardian consent and child assent were received from 25 students, including 12 males and 13 females. Most students (22 out of 25) identified themselves as Black/African American, one student identified as American Indian or Alaska Native, and two preferred not to specify. The average age of the students was 17. Most students (22 out of 25) had previously taken a computer science course.

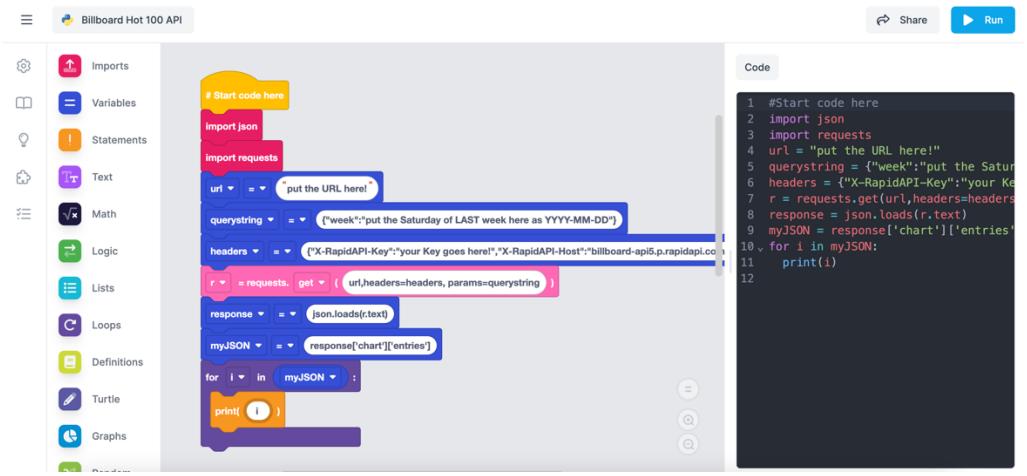

Figure 1: The pre-written EduBlocks program, which returns the most popular songs (in a JSON format) from the Billboard API.

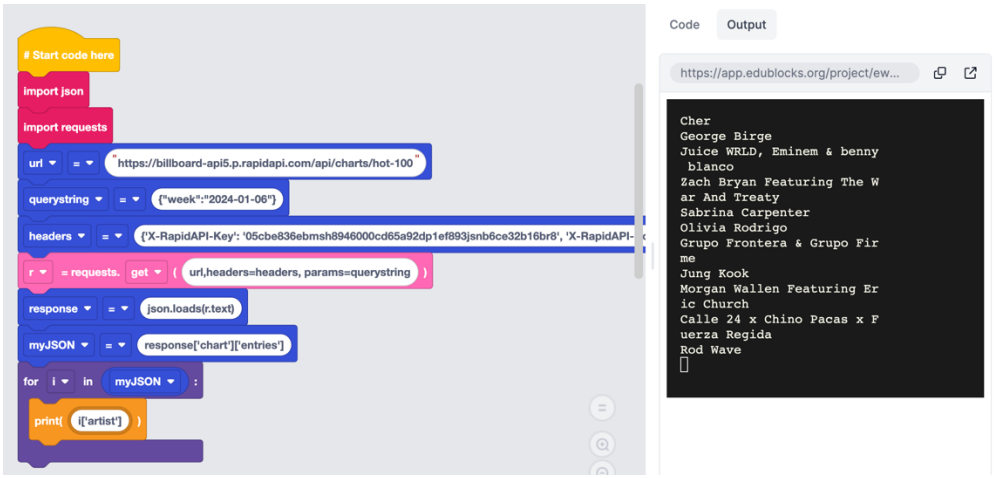

Figure 2: EduBlocks program created by a student to list popular song artists in January 2024.

4 FINDINGS

During the second unit, students learned how to use EduBlocks to programmatically retrieve and manipulate data from RapidAPI's hub, a large repository of publicly available APIs. First, the students were introduced to the concepts of an API and how they can be inquired to retrieve data. They learned what API endpoints are and how to retrieve data from a given API with Python commands inside the EduBlocks environments (Figure 1). For example, one of the data exploration activities in that unit asks students to investigate data from the Billboard API, which returns up-to-date data about the most popular songs, artists, and albums. This activity follows a Use→Modify→Create structure [9]. First, the students receive a pre-written EduBlocks program (Figure 1) that uses the "Billboard Top 100" endpoint to retrieve data in a JSON format about the most popular songs. Students run the provided program and identify the variables and the types of stored data. To do so, they first find and copy an API key and place it in the provided EduBlocks code. Then, the students are asked to modify the EduBlocks program (Figure 2) to return only the lists of artists in the dataset (i.e., all the artists whose songs were among the 100 most popular ones). Next, they are challenged to modify the program and add a loop block to count how many Taylor Swift songs are included in this list. Then, they choose an artist they like and use an `if` statement to check for any songs by their selected artist in the dataset. The last part of the activity asks students to create a new EduBlocks program to answer a question they are interested in that could be answered with data from the Billboard API. Students were asked to log their responses for each phase in a dedicated Google form.

A total of 22 students filled out the Google form that accompanied this data exploration activity. We collected their responses and analyzed them. We found that, except for one student, all students could identify all the variables in the dataset. In their answer to the question "What kind of data is stored?" only nine students referred to the type of variables, i.e., numbers and strings. Admittedly, the other students described the query or the returned data. For

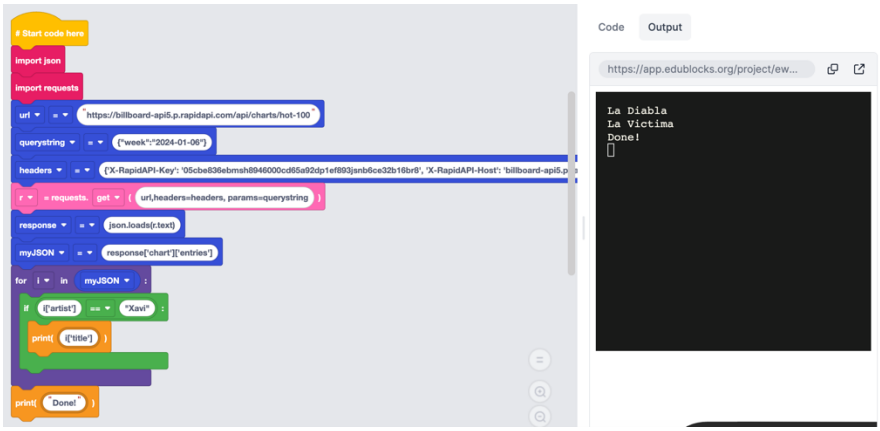

Figure 3: EduBlocks program created by a student to list popular songs by Xavi in January 2024's second week.

instance, one of the students wrote: "The top 100 Billboard songs of last week". This indicates a gap in their understanding and distinction between different data types, even though these were discussed in the first unit.

Three types of responses appeared in the student's reflections on their code modifications. One group described how they added to their `if` statement. For example, one student stated the condition itself: `i[`artist']`, while another expanded the explanation and wrote: *"I modified the code by going from i* to *i['artist']. This allowed me to focus on just the artist data"*. The second group indicated the block they added to the program. For example, one student mentioned: *"I used green if statements"* (Figure 3), while others added an explanation: *"I had to use the green if statement to modify my code to show the specific artist I picked in the dataset. I had to put (i['name'])"*. The third group only focused on the printing operation of the query results, for example: *"I used the print functions to specifically print the names of the artists"*.

The interviews conducted with five students at the end of the unit indicate excitement with the activity. The students stated: *"I enjoy it. I think it was fun"*; *"I think many students engaged in that lesson, I think music was really good [topic] because many kids listen to music"*; *"Honestly, I found it really interesting because I make music myself. And I found it as really good information for me to take in personally"*. In response to the question of what they learned during the activity, some referred to coding in general, for example: *"I learned how to set up the codes in EduBlocks"*, while others described what the activity enabled them to do: *"That activity allows you to define labels. For example, imagine you look for songs, and then it gives you all the songs an artist made this year or last year. You can use EduBlocks to print it using the print function"*.

5 DISCUSSION

As data becomes increasingly ubiquitous, it is crucial to prepare today's youth to be informed, data-literate citizens. Our approach to achieving this goal is to lay the computational foundations for teaching data science by situating it in students' lived experiences, values, and interests. We have anchored our curriculum in data exploration activities supported by rich technology environments that promote valuable practices such as data retrieval, analysis, and visualization. Our approach integrates computing and data science in an authentic way (i.e., using real datasets and live data) while grounding it in students' interests, two instructional strategies found to be important for effective data science instruction [7, 14]. Thus, the learning activities, such as the one explored in the lesson, rely on data from publicly available APIs to support interest-driven inquiry.

The Billboard API exploration activity presented above illustrates the practical application of our pedagogical approach from theory to practice. When students work with data that is relevant to them, they have a better chance to ask thought-provoking questions, analyze data critically, and develop problem-solving skills to gain insights into issues they truly care about [18]. This sense of ownership can enhance motivation and engagement and result in a better understanding of data science practices. Further, the Billboard API activity grounds data science in contemporary popular music in a way that provides both relevance to learners and a mechanism for them to have agency in pursuing topics, in this case, choosing specific music artists, as part of developing foundational data science practices. This work serves as an example of how educators and designers can show youth the prevalence and relevance of data in their world and empower them to engage with it to pursue topics they are passionate about. In doing so, this curriculum demonstrates one approach to grounding data science in the lived experiences of today's youth to better prepare them to be data-literate, informed, and empowered citizens. The work presented here is a first step towards a larger planned analysis of all three units of student work.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation (Award # 2141655). Any opinions, conclusions, and/or recommendations are those of the investigators and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] [1]Azevedo, F.S. 2019. A pedagogy for interest development: The case of amateur astronomy practice. *Learning, Culture and Social Interaction*. 23, (Dec. 2019),

100261. DOI:https://doi.org/10.1016/j.lcsi.2018.11.008.
[2] [2]Biehler, R., Veaux, R.D., Engel, J., Kazak, S. and Frischemeier, D. 2022. Research on data science education. *Statistics Education Research Journal*. 21, 2 (Jul. 2022), 1–1. DOI:https://doi.org/10.52041/serj.v21i2.606.
[3] [3]Brooks, C., Quintana, R.M., Choi, H., Quintana, C., NeCamp, T. and Gardner, J. 2021. Towards culturally relevant personalization at scale: Experiments with data science learners. *International Journal of Artificial Intelligence in Education*. 31, 3 (Sep. 2021), 516–537. DOI:https://doi.org/10.1007/s40593-021-00262-2.
[4] [4]CODAP - Common Online Data Analysis Platform: 2022. https://codap.concord.org/. Accessed: 2022-12-19.
[5] [5]Deahl, E. 2014. *Better the data you know: Developing youth data literacy in schools and informal learning environments*. Massachusetts Institute of Technology.
[6] [6]EduBlocks: 2022. https://edublocks.org/.
[7] [7]English, L.D. and Watson, J. 2018. Modelling with authentic data in sixth grade. *ZDM*. 50, 1–2 (Apr. 2018), 103–115. DOI:https://doi.org/10.1007/s11858-017-0896-y.
[8] [8]Franklin, C. and Bargagliotti, A. 2020. Introducing GAISE II: A guideline for precollege statistics and data science education. *Harvard Data Science Review*. 2, 4 (Oct. 2020). DOI:https://doi.org/10.1162/99608f92.246107bb.
[9] [9]Franklin, D., Coenraad, M., Palmer, J., Eatinger, D., Zipp, A., Anaya, M., White, M., Pham, H., Gökdemir, O. and Weintrop, D. 2020. An analysis of Use-Modify-Create pedagogical approach's success in balancing structure and student agency. *Proceedings of the 2020 ACM Conference on International Computing Education Research* (Virtual Event New Zealand, Aug. 2020), 14–24.
[10] [10]Gould, R. 2021. Toward data-scientific thinking. *Teaching Statistics*. 43, S1 (Jul. 2021). DOI:https://doi.org/10.1111/test.12267.
[11] [11]IDSSP Curriculum Team 2019. *Curriculum frameworks for introductory data science*.
[12] [12]Israel-Fishelson, R., Moon, P.F., Pauw, D. and Weintrop, D. (Accepted). Exploring interest-driven data science through participatory design. *The International Conference for the Learning Sciences – ICLS 2024* (Buffalo, USA, (Accepted)).
[13] [13]LaMar, T. and Boaler, J. 2021. The importance and emergence of K-12 data science. *Phi Delta Kappan*. 103, 1 (Sep. 2021), 49–53. DOI:https://doi.org/10.1177/00317217211043627.
[14] [14]Lee, V. and Wilkerson, M. 2018. Data Use by Middle and Secondary Students in the Digital Age: A Status Report and Future Prospects. *Instructional Technology and Learning Sciences Faculty Publications*. (Jan. 2018), 1–43.
[15] [15]Lee, V.R., Wilkerson, M.H. and Lanouette, K. 2021. A call for a humanistic stance toward K–12 data science education. *Educational Researcher*. 50, 9 (Dec. 2021), 664–672. DOI:https://doi.org/10.3102/0013189X211048810.
[16] [16]Makar, K. and Ben-Zvi, D. 2011. *The role of context in developing reasoning about informal statistical inference*. Taylor & Francis.
[17] [17]Michaelis, J.E. and Weintrop, D. 2022. Interest development theory in computing education: A framework and toolkit for researchers and designers. *ACM Transactions on Computing Education*. 22, 4 (Dec. 2022), 43:1-43:27. DOI:https://doi.org/10.1145/3487054.
[18] [18]Moll, L.C. 2019. Elaborating funds of knowledge: Community-oriented practices in international contexts. *Literacy Research: Theory, Method, and Practice*. 68, 1 (Nov. 2019), 130–138. DOI:https://doi.org/10.1177/2381336919870805.
[19] [19]Moon, P.F., Israel-Fishelson, R., Tabak, R. and Weintrop, D. 2023. The tools being used to introduce youth to data science. *Proceedings of the 22nd Annual ACM Interaction Design and Children Conference* (Chicago IL USA, Jun. 2023), 150–159.
[20] [20]Renninger, K.A. and Hidi, S. 2015. *The power of interest for motivation and engagement*. Routledge.
[21] [21]Rosenberg, J.M., Lawson, M., Anderson, D.J., Jones, R.S. and Rutherford, T. 2020. Making data science count in and for education. *Research Methods in Learning Design and Technology*. Routledge. 94–110.
[22] [22]Rowley, J. 2007. The wisdom hierarchy: Representations of the DIKW hierarchy. *Journal of Information Science*. 33, 2 (Apr. 2007), 163–180. DOI:https://doi.org/10.1177/0165551506070706.
[23] [23]Saldaña, J. 2016. *The coding manual for qualitative researchers*. SAGE.
[24] [24]Schanzer, E., Pfenning, N., Denny, F., Dooman, S., Politz, J.G., Lerner, B.S., Fisler, K. and Krishnamurthi, S. 2022. Integrated data science for secondary schools: Design and assessment of a curriculum. *Proceedings of the 53rd ACM Technical Symposium on Computer Science Education* (Providence RI USA, Feb. 2022), 22–28.
[25] [25]Weiland, T. and Engledowl, C. 2022. Transforming curriculum and building capacity in K–12 data science education. *Harvard Data Science Review*. 4, 4 (Oct. 2022). DOI:https://doi.org/10.1162/99608f92.7fea779a.
[26] [26]Wilkerson, M.H. and Polman, J.L. 2020. Situating data science: Exploring how relationships to data shape learning. *Journal of the Learning Sciences*. 29, 1 (Jan. 2020), 1–10. DOI:https://doi.org/10.1080/10508406.2019.1705664.