WIP: Bridging the Data Gap – Introducing Unplugged Data Science

1st Mariana Arboleda
Department of Education
Universidad del Norte
Barranquilla, Colombia
marianaaarboleda@uninorte.edu.co

2nd Camilo Vieira
Department of Education
Universidad del Norte
Barranquilla, Colombia
cvieira@uninorte.edu.co

3rd Jennifer L. Chiu
Department of Curriculum, Instruction, and Special Education
University of Virginia
Charlottesville, USA
jlc4dz@virginia.edu

Abstract

This innovative practice paper outlines a design-based research study to enhance data science literacy in students using an unplugged approach. With the rapid growth of data and its impact on daily life, there's a need for new educational methods to help students understand and use data effectively. Data science involves math, statistics, computation, AI, and machine learning, but teaching these concepts can be challenging without access to computers or the internet, potentially widening educational inequities. Inspired by unplugged computer science education, this study supports data science education in unplugged classrooms. Initial research focused on creating and testing two learning units on data visualization and summarization using innovative activities and reflection prompts. Analyzing students' work from a pilot with ten eleventh-graders in a low-income public school in a Global South country, the study aims to understand how these activities support data visualization skills and identify limitations. Future work will develop and implement lesson plans to promote data and AI literacy, aiming to make data science education more inclusive across diverse backgrounds.

Keywords—data science education, k-12 education

I. Introduction

We live in an era where data powers everything. Large amounts of data are being generated every day, and the ability to extract meaningful insights from this data has become essential. Data Science has gained importance as the area of knowledge responsible for turning data into information. It is often described as the multidisciplinary intersection of mathematics, computer science, and subject matter expertise that allows for data collection, preparation, analysis, and interpretation to make informed decisions [1].

In the last decade, DS has become a wide professional field. Data roles are now present across all corporate levels and areas of responsibility in companies [2]. And that growth has led education to adapt and integrate DS across educational levels. Universities were among the first to introduce DS courses. A "consensus curriculum" has started to emerge with a significant focus on computing and using specialized programming languages (such as R and Python) [3]. However, there is uncertainty about how this can be translated for K-12 settings [4].

DS is still at the center of the debate of whether it should be considered an extension of Statistics as it is the discipline that concerns the analysis of data [5]. As such, the first attempts at introducing DS education in schools have been aimed at the mathematics curriculum [6]. However, we know that while mathematical and statistical knowledge is essential to data analysis, DS requires multiple skills from Computer Science. Students should also learn about collecting data from multiple sources, storing structured and unstructured data, cleaning and processing large datasets, creating visualizations, and analyzing data to make data-informed decisions. Data Scientists often need to build physical systems, understand and manage databases, and automate different solutions through coding to collect, store, and analyze data. Also, they need computational thinking (CT) skills (e.g., decomposition, abstraction, pattern recognition, depuration, and algorithm thinking) to develop comprehensive solutions and ultimately transform isolated data into knowledge. It is this set of abilities related to computing that makes DS a science, or "the science of learning from data" [7, p. 5].

Even though attaching DS education to the mathematics curriculum might give students opportunities to see the relevance of mathematics in their lives [8], such integration presents multiple challenges. First, teachers do not necessarily have the technological knowledge to integrate data science concepts into their lessons. Second, incorporating technological tools may distract students from learning the key DS concepts, focusing on understanding the tool instead. Third, access to technology in schools is often restricted. Most opportunities for utilizing computers are limited to computer science classes, especially in contexts with restricted resources. These limitations can hinder students' exposure to practical data science applications and limit their ability to develop essential skills.

In this work-in-progress paper, we propose using unplugged activities to foster data science in the classroom. By focusing on unplugged activities, we can reduce the cognitive load on students learning how to use specialized technologies, thereby allowing them to concentrate on the fundamental concepts of data science. Simultaneously, this approach addresses teachers' time limitations and lack of knowledge by providing a more accessible and versatile way to develop essential knowledge about data.

In the rest of the article, we present our theoretical framework, building on the Computer Science use of Unplugged Activities and Cognitive Loadings Theory to introduce our innovative practice of designing unplugged DS activities. We then present a sample of the work we used during the pilot testing and finally discuss future work and limitations.

II. Theoretical Framework

A. Unplugged Activities

Unplugged activities have been a matter of interest in Computational Thinking (CT) education as a method to promote learning without technology as a mediator or "doing computing without computers" [9]. The unplugged activities range from games and puzzles to outdoor projects that illustrate key ideas in computer science. A recent systematic review and meta-analysis found that unplugged activities foster CT education at K-12 levels, increasing students' algorithmic thinking, decomposition, abstraction, and pattern recognition skills [10].

Unplugged activities can also reduce the challenges of teaching computing concepts by making them more accessible. Teachers report seeing value in unplugged activities as they help engage in topics they are not experts on and gain confidence in integrating them into their practice [11].

Despite its popularity in CT education, very few initiatives exist to integrate unplugged activities in DS education. The activities focus on understanding machine learning concepts, such as presenting linear classification through a pinboard activity [12] and introducing facial recognition with cartoon characters, photos, and descriptions [13]. However, they lack continuity and are limited in presenting concepts, so there is a need for more activities that also foster essential DS skills.

B. Cognitive Load Theory

The Cognitive Load Theory (CLT) describes how students process information as they learn new concepts and skills, given a cognitive architecture [14]. The cognitive architecture includes a working memory (WM), which is limited in time and space, and a long-term memory (LTM), which is vast. When students explore new learning materials, they use what they know from the LTM to make sense of the new information in the WM, managing the cognitive load.

Learning computer programming is complex, as students need to process multiple interacting elements simultaneously, including the problem, the algorithm, the syntax, and how the computer works. Hence, identifying effective ways to reduce such cognitive loads can contribute to supporting student learning. The CLT suggests that reducing extraneous elements from the instructional design may enable the learners to maximize the intrinsic and germane loads, contributing to an effective learning process. A recent study showed that using unplugged activities and providing examples for students to develop early schemata helped students develop basic computational thinking skills [15].

This study proposes using unplugged activities to integrate DS concepts and practices into K-12 settings. Such an approach will allow teachers and students to focus on learning specific DS concepts and skills without the extraneous load generated by trying to understand how to use a visualization tool or a programming language. Moreover, these activities will address the challenges of limited teachers' technological knowledge and limited technological infrastructure in schools.

III. Methods

This work-in-progress is part of an ongoing project of creating and integrating unplugged Data Science lessons in K-12 schools. We will design, test, and publish multiple educational materials for teachers to include in their classes. This paper focuses on the pilot activities tested during the second semester of 2023 and presents the progress and future work. The pilot was conducted in a low-income public school during the senior year of ten students (five women and five men). The pilots took part in their mandatory mathematics lessons but were facilitated by an outside tutor, not their usual teacher. The classes were divided into one-and-a-half-hour lessons, which the students considered as their preparation for the nationwide standard test.

IV. Procedures

A. The context

The traditional approach to teaching data science relies heavily on technology and computer-based tools [16]. However, in low-resource contexts where technology is not readily available, educators face unique challenges in delivering data science education. Also, while teachers might be experts in their subject area, in this case, Mathematics, they might lack the Technological Knowledge necessary to introduce DS into their practice.

The pilot study was conducted in a small school on the Caribbean coast. The school building has seven classrooms, one for each year, for a total population of about 180 students. The classrooms have the essential equipment of a blackboard and chairs, but no tables, computers, or internet access is available for students. Most students have a personal or shared cell phone as their only way to access the internet.

B. The practice

We designed two sets of lesson plans with an approximate duration of 10 hours each. The lessons were designed to present DS in real-world situations and in a way that students could see the relevance of the new concepts they were learning.

In this project's scope, we define DS as a set of skills that allow students to solve problems using data, including:

- Reading graphs: Interpreting different visualizations, recognizing patterns, and explaining trends beyond focusing on specific data points.
- Working with “messy” data: Recognizing datasets are rarely perfect and analysis-ready. Identifying common problems different datasets can have, such as missing data, duplication, and erroneous information.
- Working with multiple data sources: Comparing different data sources and recognizing the differences in data collected through surveys, sensors, or software. Reflecting on the data gathering process and selecting appropriate methods to solve specific problems.
- Selecting and evaluating data visualizations: Understanding the basic concepts of good data visualization techniques, avoiding misleading practices. Creating charts and communicating results.

Data is the heart of DS. Multiple authors have found that DS education is more meaningful for students when the data they study refers to their interests or personal experiences [17]. A recent study that systematically analyzed the 296 datasets used in four major DS Curriculums found that most datasets were less than ideal for K-12 settings. The authors found that only about 36% of the datasets overlapped with students’ interests. Also, most of them were outdated, and did not reflect the complexity of real data [18].

In our study, we worked with different datasets from real-world sources connected to students’ interests. For example, we worked with growth data from a Facebook group, indices of retention and usage of multiple social networks, Google search data from important events, and a topic of their choosing (via Google trends). We also relied on data from their own context, as the school’s past performance in a national standardized test, and self-generated data from measuring their reaction time.

Because the lesson plans were tailored to their experiences, students placed a high value on the data we chose. However, the activities were designed to be easily adaptable and represent a situation within different contexts and interests.

1) Introduction to DS and data visualization

The first lesson plan introduces students to DS and data visualization. The first lesson, presented in Figure 1, starts with a narrative discussing a real life scenario using Facebook data.

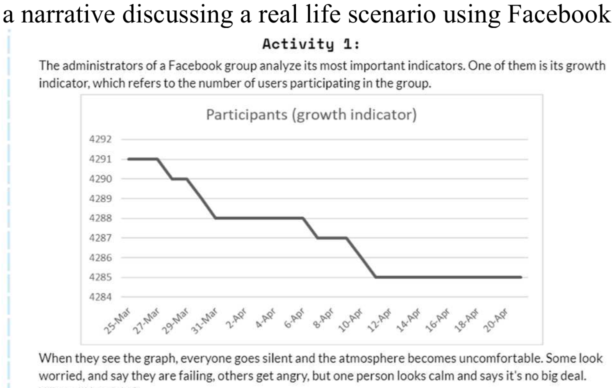

Figure 1 Session 1: First introduction to reading charts. The activity presents a situation using real data from a ***Facebook*** *group.*

Then, the students discussed what they thought about the problem, identified the axis distortion, and wrote why they believed the data was presented in a graph instead of plain numbers. Figure 2 presents another example of a data visualization activity. In this activity, students analyzed the graph but had no context for the phenomena it illustrated. Students only knew that the data source was Google searches, which came from their official Google Trends website.

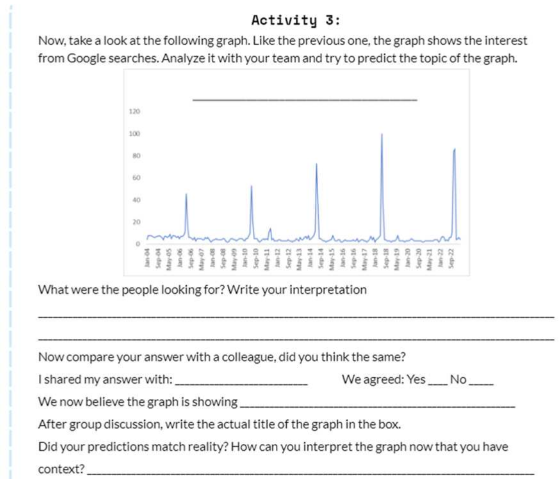

Figure 2 Session 4: Students work in pairs and use their pattern recognition skills to predict the phenomena behind the presented graph.

Students were challenged to predict the phenomena and defend their answers. In this activity, students were invited to look at the trends and patterns (i.e., identifying peaks every four years) and connect those findings with their knowledge (i.e., important events that happen every four years). This set of activities also allowed them to discuss how datasets “do not say much by themselves.” Therefore, we consider that content matter expertise is fundamental in DS.

2) Introduction to data analysis

The second lesson plan introduced students to basic data analysis concepts and reporting through a real-life narrative. First, we presented students with a familiar interface of a messaging app and a situation, as shown in Figure 3: Their “boss” or school coordinator asked them to analyze their school’s performance in a standardized test, using data from the previous years. It is important to note that, even though the dataset was anonymized, it was real data from their own context.

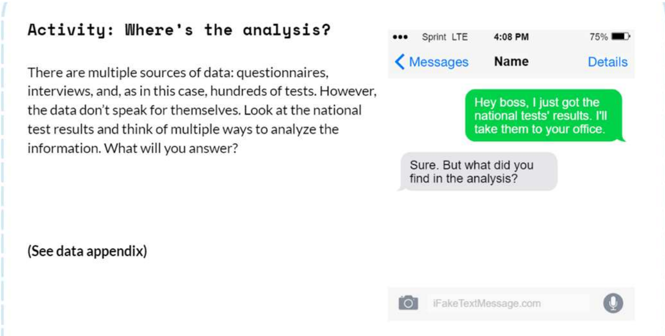

Figure 3 Session 10: The activity presents a narrative of the school's coordinator asking for a written report analyzing the students' performance on the nationwide standardized test

The lesson progressed to introduce the first steps of Exploratory Data Analysis and data storytelling. Given the limitation of not having a computer, students worked in groups to describe the dataset. This led to creative problem-solving, such as working with subsets of the data, performing a global calculation (of the average), comparing the results vs. processing the complete dataset, and answering questions such as "What would happen if the dataset had more than 100 records? How would you approach it? And if it had more than 1000 records," "Would you be able to find the highest score? How long would it take?" The students reflected on the need for computation in DS.

The activities required students to write, reflect, and communicate their results, which are essential skills in DS. After analyzing the data, students were asked to draft a report summarizing their findings and insights, as shown in Figure 4. This task encouraged them to articulate their thought processes and practice making data-informed conclusions. Additionally, each group prepared a scaffolded report in which they interpreted the data and shared their analysis with the rest of the class. These activities helped students develop data interpretation skills and emphasized the importance of effective communication and reflective practices in DS.

By the end of the intervention, students could accurately describe different types of data visualizations, extract meaningful information, and suggest data-based conclusions. And, as one of the students mentioned, "Now we see data and graphs everywhere."

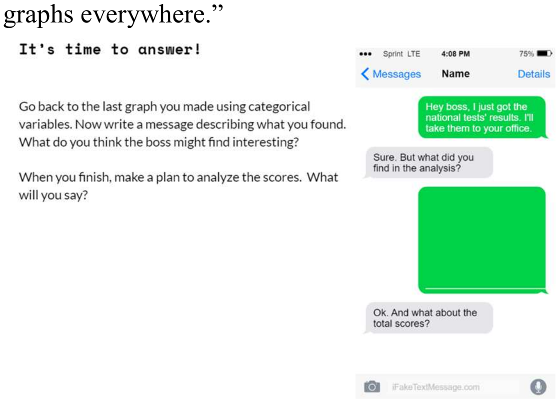

Figure 4 Session 14: The students apply their knowledge to summarize their findings in a familiar format such as a chat interface

V. Machine Learning and Artificial Intelligence

Our current efforts are focused on designing unplugged machine learning (ML) and artificial intelligence (AI) activities. Given the increasing popularity and influence of AI and ML in our daily lives, our aim is to create accessible learning materials to educate students on their core concepts to better understand and navigate the world around them.
For example, when introducing students on how machines learn, we present concepts such as *Training data, Prediction,* and *Error.* Through a drawing game, we guide students to make their personal "training dataset" visible. Following the game, students are asked "How do you know that a house is a house?", and connect their thought process of guessing a drawing, with the predictions of a ML algorithm.
Additionally, we are preparing to address ethical issues related to ML and AI, such as data and algorithmic bias, incomplete datasets, and the reliability of these systems. By incorporating these topics, we aim to empower students so they can responsibly engage with these technologies.

VI. Future Work

We will continue designing and improving a collection of lesson plans to be included not only in mathematics classes but also in diverse subject areas. This interdisciplinary approach aims to demonstrate the versatility and applicability of data science concepts across the curriculum, fostering a broader understanding among students.
We recognize the limitation of our current work, which lacks empirical results. Therefore, a crucial next step is to conduct research with larger groups to evaluate the efficacy of unplugged activities compared to computer-based data science teaching methods. This comparative study will provide valuable insights into the strengths and weaknesses of each approach, guiding future iterations of our lesson plans and teaching strategies.

VII. Discussion

Our approach of creating lesson plans that are easily adaptable to match students' interests and contexts aligns with the recommendations by Israel-Fishelson and colleagues in [18]. The authors invite curriculum designers to think of ways to allow students to choose the datasets themselves or to empower educators to adapt the content based on their knowledge about their students. This strategy is intended to help engage students as their interest in learning impacts their motivation and overall success in class.

Moreover, our observations from the pilot corroborate findings from previous research, which indicate that many secondary school students struggle with data analysis, even in simple formats. Students often focus on individual data points, missing broader patterns and trends and failing to detect online sources' bias [8]. To address these challenges, we shifted from the traditional method, where students created charts from very small, simulated datasets that did not require summarization, and students had to plot dot-by-dot. Instead, we presented data visualizations first and invited students to focus on the "whole picture." This approach helped students develop a more comprehensive understanding of data, allowing them to identify patterns and trends more effectively.

Integrating unplugged DS will impact the broad context of data science education, as it will promote the inclusion of students regardless of their background or socioeconomic status.

Acknowledgment

This work was supported by the Center for Global Inquiry and Innovation at the University of Virginia.

VIII. References

[1] W. Finzer, "The Data Science Education Dilemma," *Technol. Innov. Stat. Educ.*, vol. 7, no. 2, 2013, doi: 10.5070/T572013891.

[2] L. Cao, "Data Science: Profession and Education," *IEEE Intell. Syst.*, vol. 34, no. 5, pp. 35–44, Sep. 2019, doi: 10.1109/MIS.2019.2936705.

[3] A. Schwab-McCoy, C. M. Baker, and R. E. Gasper, "Data Science in 2020: Computing, Curricula, and Challenges for the Next 10 Years," *J. Stat. Data Sci. Educ.*, vol. 29, no. sup1, pp. S40–S50, Jan. 2021, doi: 10.1080/10691898.2020.1851159.

[4] S. Schreiter *et al.*, "Teaching for statistical and data literacy in K-12 STEM education: a systematic review on teacher variables, teacher education, and impacts on classroom practice," *ZDM – Math. Educ.*, vol. 56, no. 1, pp. 31–45, Feb. 2024, doi: 10.1007/s11858-023-01531-1.

[5] I. Carmichael and J. S. Marron, "Data science vs. statistics: two cultures?" *Jpn. J. Stat. Data Sci.*, vol. 1, no. 1, pp. 117–138, Jun. 2018, doi: 10.1007/s42081-018-0009-3.

[6] T. Weiland and C. Engledowl, "Transforming Curriculum and Building Capacity in K–12 Data Science Education," *Harv. Data Sci. Rev.*, vol. 4, no. 4, Oct. 2022, doi: 10.1162/99608f92.7fea779a.

[7] D. Donoho, "50 Years of Data Science," *J. Comput. Graph. Stat.*, vol. 26, no. 4, pp. 745–766, Oct. 2017, doi: 10.1080/10618600.2017.1384734.

[8] T. LaMar and J. Boaler, "The importance and emergence of K-12 data science," *Phi Delta Kappan*, Aug. 2021, doi: 10.1177/00317217211043627.

[9] T. Bell, J. Alexander, I. Freeman, and M. Grimley, "Computer science unplugged: school students doing real computing without computers," *N. Z. J. Appl. Comput. Inf. Technol.*, vol. 13, no. 1, pp. 20–29, 2009.

[10] P. Chen, D. Yang, A. H. S. Metwally, J. Lavonen, and X. Wang, "Fostering computational thinking through unplugged activities: A systematic literature review and meta-analysis," *Int. J. STEM Educ.*, vol. 10, no. 1, p. 47, Jul. 2023, doi: 10.1186/s40594-023-00434-7.

[11] T. Bell and J. Vahrenhold, "CS Unplugged—How Is It Used, and Does It Work?," in *Adventures Between Lower Bounds and Higher Altitudes: Essays Dedicated to Juraj Hromkovič on the Occasion of His 60th Birthday*, H.-J. Böckenhauer, D. Komm, and W. Unger, Eds., Cham: Springer International Publishing, 2018, pp. 497–521. doi: 10.1007/978-3-319-98355-4_29.

[12] E. Ossovski and M. Brinkmeier, "Machine Learning Unplugged - Development and Evaluation of a Workshop About Machine Learning," in *Informatics in Schools. New Ideas in School Informatics*, S. N. Pozdniakov and V. Dagienė, Eds., Cham: Springer International Publishing, 2019, pp. 136–146. doi: 10.1007/978-3-030-33759-9_11.

[13] J. Ho, "AI classroom activity: Facial recognition," Teacher Magazine. Accessed: May 19, 2024. [Online]. Available: https://www.teachermagazine.com/au_en/articles/ai-classroom-activity-facial-recognition

[14] J. Sweller, J. J. G. van Merriënboer, and F. Paas, "Cognitive Architecture and Instructional Design: 20 Years Later," *Educ. Psychol. Rev.*, vol. 31, no. 2, pp. 261–292, Jun. 2019, doi: 10.1007/s10648-019-09465-5.

[15] C. Vieira, R. Gómez, M. Gómez, M. Canu, and Mauricio Duque, "Implementing Unplugged CS and Use-Modify-Create to Develop Student Computational Thinking Skills: – A Nationwide Implementation in Colombia," *Educ. Technol. Soc.*, 2023, doi: https://doi.org/10.30191/ETS.202307_26(3).0012.

[16] M. Çetinkaya-Rundel and C. Rundel, "Infrastructure and Tools for Teaching Computing Throughout the Statistical Curriculum," *Am. Stat.*, vol. 72, no. 1, pp. 58–65, Jan. 2018, doi: 10.1080/00031305.2017.1397549.

[17] V. R. Lee, M. H. Wilkerson, and K. Lanouette, "A Call for a Humanistic Stance Toward K–12 Data Science Education," *Educ. Res.*, vol. 50, no. 9, pp. 664–672, Dec. 2021, doi: 10.3102/0013189X211048810.

[18] R. Israel-Fishelson, P. Moon, R. Tabak, and D. Weintrop, "Understanding the Data in K-12 Data Science," *Harv. Data Sci. Rev.*, vol. 6, no. 2, Apr. 2024, doi: 10.1162/99608f92.4f3ac3da.