Individual Paper: DS4ALL: Teaching High-School Students Data Science and AI/ML Using the Texera Workflow Platform as a Service

Jiadong Bai, UC Irvine, jiadongb@uci.edu
Xiaozhen Liu, UC Irvine, xiaozl3@uci.edu
Anthony Cuturrufo, UCLA, anthonycuturrufo1@gmail.com
Alexander Kundu Taylor, UCLA, ataylor2@cs.ucla.edu
Jeehyun Hwang, UCLA, jeehyunhwang@cs.ucla.edu
Mingyu Derek Ma, UCLA, derek.ma@ucla.edu
Xinyuan Lin, UC Irvine, xinyual3@uci.edu
Yanqiao Zhu, UCLA, yzhu@cs.ucla.edu
Yicong Huang, UC Irvine, yicongh1@ics.uci.edu
Yunyan Ding, UC Irvine, yunyad1@uci.edu
Wei Wang, UCLA, weiwang@cs.ucla.edu
Chen Li, UC Irvine, chenli@ics.uci.edu

Session Description

In the summers of 2023 and 2024, we organized an NSF-supported program called "Data Science For All" ("DS4ALL" for short) at the University of California, Irvine to teach data science skills to high-school students. The aim of the program was to provide a learning experience that did not require a programming background so that it could be offered to a diverse range of students. We accomplished the goal by leveraging the open-source Texera system to provide a cloud service for students to learn data science using GUI-based workflows. The platform supports powerful collaboration features like shared editing, execution, and version control. These features not only improved students' experiences when working in teams but also made teaching various topics efficient. Students also learned Python and advanced analytics using artificial intelligence (AI) and machine learning (ML). In this paper, we present the details of the program and the insights we gained.

1. Motivation

Many existing high-school data science curricula require students to have a coding experience in languages such as Python. This prerequisite prevents many students from taking courses due to their limited coding skills. In addition, it is challenging for students to download software packages, set up an integrated development environment (IDE), install libraries, and use services such as GitHub. To address these issues, we want to study whether we can develop a solution to lower students' barriers of learning data science and advanced topics such as AI and ML.

2. Our Method: Summer Program of "Data Science for All"

In the summers of 2023 and 2024, funded by NSF, we, a joint team of UC Irvine and UCLA, organized a program called "Data Science For All" ("DS4ALL" for short) at UC Irvine to teach high-school students skills in data science. We used the open-source Texera system to provide a cloud service for students to do data analytics using GUI-based workflows. A primary advantage of using the platform is that it does not require students to have a programming background, which makes these skills significantly more accessible to students.

Texera is an open-source system that supports collaborative data science using a cloud service. It allows students with no coding skills to collaborate in the data science life cycle. Its primary focus is collaboration as a web service, consistent with the general trend of cloud services for efficient teamwork, as demonstrated by the success of services such as Google Docs. It also has powerful features for advanced programmers to add operators in languages (Python, R, and Java) and debug the execution of a workflow. Fig. 1 shows a screenshot of the system where multiple students collaboratively construct and execute a data science project as a workflow.
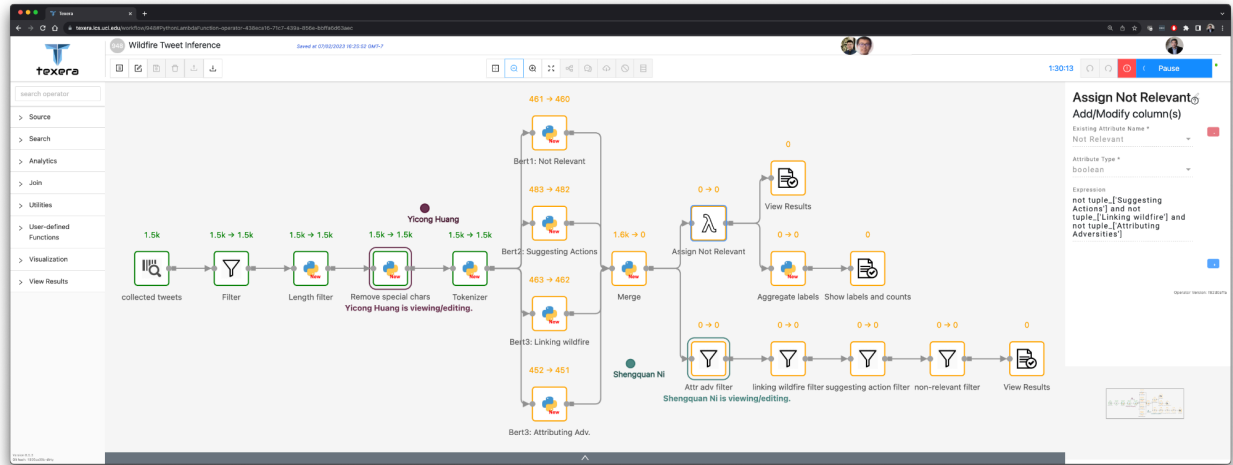
Fig. 1: Two users collaborating on the same Texera workflow.

Recruitments

We reached out to local high schools and had 10 students in summer 2023 and 16 students in summer 2024. We did not consider students' coding backgrounds during the selection process. The program was free because of NSF's support.

Program structure

We organized each summer program to cover Texera usage and data analytics skills over two weeks. In week one, we spent two days teaching students basic concepts about data records and how to use operators to do data wrangling, data preparation, and data visualization. We started teaching Python as a user-defined function (UDF) operator on the third day. Since each operator provides a clean context with a template for the Python script to process incoming data and produce results, students quickly learned how to write code to implement their own logic. In week two, we taught students concepts related to AI/ML. We explained ML models such as linear regression, logistic regression, decision trees, and neural networks. We used sample workflows and operators on Texera to demonstrate how to use these ML models in a data science pipeline and shared them with the students to be included in their workflows. Students formed two-person teams, each choosing their own topic for a capstone project. We gave lectures in the mornings and used the afternoons as lab sessions to work on their capstone projects.

3. Results

As an example, in the 2024 program, 16 students created a total of 132 workflows. 21% of these workflows used ML models, and 64% used Python UDF operators. All the final capstone workflows used Python and ML models.

Advantages of the cloud service

A main benefit of Texera is that it is accessible to students without them needing to download any software. Some students were from low-income families and did not have a laptop. We provided them with laptops, and students were able to quickly access Texera using these laptops. In addition, whenever we identified bugs, we fixed them quickly at night so that students could use the improved service the next day without any disruption.

Easy of use

In week one, the students grasped data-preparation concepts in two days, because of the easy-to-use user interface of Texera. They found the GUI interface intuitive to learn and use. The collaboration features benefited students working in teams. The platform allowed instructors to share sample workflows and ML models with the class. When students encountered issues, instructors could investigate from their laptops after the workflows were shared.

Capstone showcase event

At the end of each program, we organized a showcase event and invited the students' parents, Texera developers, and other guests. The high-school students did a fantastic job showing their final workflows that had many operators including some with ML models. Fig. 2 shows two students presenting their capstone project as a Texera workflow.
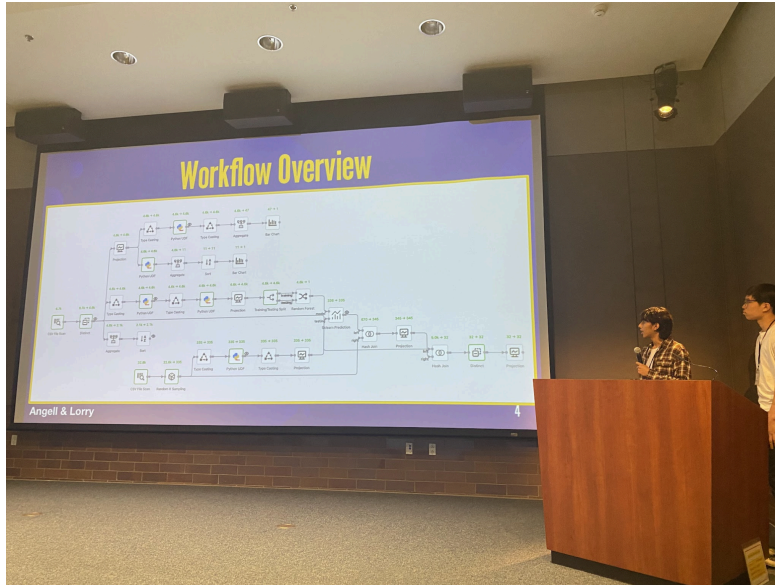
Fig. 2: Two students presenting their workflow

4. Discussion

Many high schools can adopt our successful DS4ALL program to teach students without requiring them to have coding skills. The Texera platform enables a cloud service that provides unique advantages that can reduce students' barriers to learning data science, AI, and ML. It helps students familiarize themselves with these critical topics and attract them to the increasingly important STEM disciplines.

References

"Data Science for All" summer program. (n.d.). https://ds4all.ics.uci.edu/
Texera: Collaborative Machine-Learning-centric data analytics using workflows. GitHub. (n.d.). https://github.com/Texera/texera
Wang, Z., Huang, Y., Ni, S., Kumar, A., Alsudais, S., Liu, X., Lin, X., Ding, Y., & Li, C. (2024). Texera: A system for collaborative and interactive data analytics using workflows. *Proceedings of the VLDB Endowment, 17*(11), 3580-3588. https://doi.org/10.14778/3681954.3682022
Broadening participation in computing. NSF. (n.d.). https://new.nsf.gov/cise/broadening-participation