ASEE 2022 ANNUAL CONFERENCE
Excellence Through Diversity
MINNEAPOLIS, MINNESOTA, JUNE 26TH-29TH, 2022 ASEE
Paper ID #38288

Developing a Hands-on Data Science Curriculum for Non-Computing Majors

Xumin Liu

Erik Golen

Rajendra K Raj (Dr)

Developing a Hands-on Data Science Curriculum for Non-Computing Majors

Xumin Liu, Erik Golen, and Rajendra Raj
{xmlics, efgics, rkrics}@rit.edu
Golisano College of Computing and Information Sciences
Rochester Institute of Technology

Abstract

This paper describes the design and development of an appropriate data science curriculum accessible to non-computing majors with little or no programming background. This project took a two-prong approach to address such a curriculum: (1) a Web-based Data Science Learning Platform was developed to offer such students hands-on practice with processing and analyzing data without needing to write code, and (2) a Data Science Curricular Module for teaching data science concepts in both an existing Computer Science Principles course and a follow-on Data Science Principles course. The paper also discusses initial experiences with deploying the curricular module at Rochester Institute Technology.

Introduction

Learning data science has become commonplace in many disciplines and the related curriculum is in great demand[1,2,3]. However, the majority of current data science courses either have prerequisite requirements on programming (such as Python) or are designed with a major focus on programming, which is inappropriate for non-computing majors. First, these students cannot access the traditional data science curriculum due to long prerequisite chains consisting of computing and mathematical topics listed earlier. Second, non-computing majors are usually more interested in learning how to use data science techniques effectively in the context of their disciplines, rather than learn how to write code. Therefore, it is crucial for data science to be brought to non-computing majors in an easy-to-access manner.

In this paper, we describe the design and development of a data science curriculum that is accessible and appropriate for non-computing majors. The curriculum focuses on teaching students fundamental concepts and techniques of data science, introducing them real-world data science problems, and guiding them how to apply data science techniques to tackle those problems in computational way. We first present a Web-based Data Science Learning Platform (DSLP) that demonstrates a computational way of processing and analyzing data and helps students to obtain hands-on experience regardless of their programming background. We then present a Data Science curricular Module that teaches data science concepts in both the existing

Computer Science Principles (CSP) course and a follow-on Data Science Principles (DSP) course. In the end, we briefly describe our experience of teaching the course module at Rochester Institute of Technology in the Fall of 2021 and discuss evaluation results.

Data Science Learning Platform

We have developed a web-based Data Science Learning Platform (DSLP) that works as a middleware between users (i.e., students or instructors) and existing data science tools (e.g., Pandas, Numpy, Scikit-learn) to facilitate users to use the tools without the need for coding. This allows students to focus their learning on the high-level workflow of understanding, exploring, scrubbing, analyzing, and interpreting data instead of how to write lines of code. As shown in

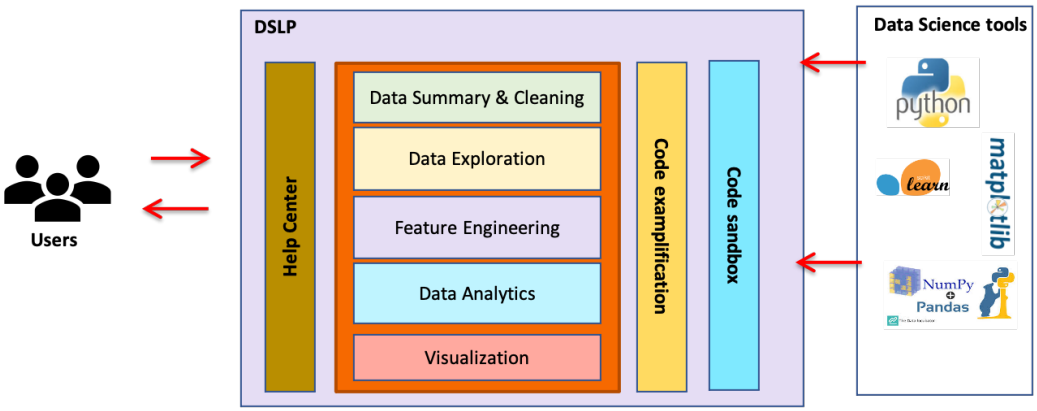

Figure 1: DSLP, a supporting platform for teaching data sciences

Figure 1, the DSLP takes user requests as input and translates them as invocations to the back-end R or Python DS packages. Users can select an in-house dataset or upload their own data to start with and use the DSLP to explore, process, and analyze the dataset. Compared to the existing data science platforms with a graphical user interface (GUI) such as Weka and RapidMiner, the DSLP has the following unique features that make it accessible to non-computing majors and enhance teaching and learning experiences. (1) *Web-based platform:* The DSLP is web-based, which leaves the workload of setting up and maintaining a lab environment to the server side. This lifts the burden from users and allows them to perform hands on practice without the need for downloading, installing, configuring, and upgrading data science frameworks. Meanwhile, as the computationally intensive and high storage demanding tasks such as data storing, processing, and analyzing, are performed on the server side, such a web-based platform takes minimum requirements on the user side, e.g., web browsers. (2) *Help center:* Considering the limited computing and mathematics background of non-computing majors, the DSLP supports a comprehensive and extensible help center to help students better understand data science concepts and techniques. Working with CSP instructors, we will identify the terms that students may have trouble understanding, such as data correlation and heat map, and provide detailed descriptions and examples for comprehension. The help center also covers the explanation of data science methods and helpful tips for using them. Different from the documentation of current data science libraries (such as R and Python data science packages), where the descriptions of methods and parameters are usually brief and technical, our help center is designed for non-computing majors, explaining the terms and methods in a way that they can understand. (3) *Code-exemplification:*

The DSLP contains a code exemplification feature that links each interface-level operations, such as clicking the data visualization button, to the code (in R or Python) invoking the corresponding data science libraries. Users can view the code and learn how to handle data in a computational way. Meanwhile, such code generated from the DSLP will be included in a code repository, which serves as sample code for instructors to teach coding in R or Python. (4) *Code Sandbox:* The DSLP provides a virtual Python IDE environment for students to try their own R or Python code, which allows them to learn programming in these languages without the need for installing, configuring, and maintaining underlying software, such as R Studio and PyCharm.

Data Science Curricular Module

We have developed a Data Science Curricular Module (DSCM) to teach data science to non-computing majors, covering the fundamental concepts, techniques, and application domains of data science. Topics include different types of data, data querying, cleaning, exploration, visualization, feature modeling, and data analytics models. We adopt the format of a curricular module, as it has been common in computing pedagogy as self-contained units of instructions to be incorporated in a specific course[4,5]. The curricular module has two settings, Level I and Level II, targeting the first and second CSP courses for non-computing majors, such as the CS Principles and DS Principles courses for non-majors. Both of these two levels incorporate the usage of the DSLP into the teaching and learning, assisting students to obtain hands on practice of handling, understanding, and analyzing real world data sets, as well as an awareness of data-related ethics and privacy.

The first targeted course is equivalent to AP CSP and is offered to first year college students with little or no computing background, focusing on the principles of computing. Data science is a typical topic in this course, where Level I can be plugged into. Students taking the course are expected to have little or no previous computing experience and mathematics background at the high school level, including high school algebra, foundations of linear functions, composition of functions, and problem solving strategies involving multiple approaches and collaborative efforts. This CSP course is usually designed to introduce students to the central ideas of computing and to instill practices of computational thinking. Typically there are around 10-12 important computing topics included in the course outlines, such as hardware systems, algorithms, computer security, as well as Database/data mining. The proposed curricular module with Level I can be plugged into the course for the topic of database/data mining with a 1-week duration.

The second course is the DS Principles course focusing on managing and analyzing data in a computational way, where Level II can be used. This level of DSCM allows students to delve into the key steps in a data science pipeline, including data formats and querying, cleaning, data exploration and visualization, feature selection and engineering, and data analytics. Case studies, programming assignments, and course projects, are designed to make this course hands-on so students will learn how to apply the learned data science techniques to analyze real world datasets (such as Titanic, COVID, Census datasets) and interpret the analysis results. Data privacy and ethic issues are also covered in the module. It teaches students the basic techniques and provides a high level overview of the theories behind them. Level II is designed to be taught in 12-14 weeks, covering almost a whole semester while also leaving space for instructors to add in their own topics, such as additional prerequisites in computing/mathematics and data science

applications.

Course Module Deployment and Evaluation

We deployed the two levels of DSCM at Rochester Institute of Technology in the Fall of 2021. We taught DSCM level I in the two sections of Computer Science Principle course with the total enrollment of 36 students and DSCM Level II in the one section of Data Science Principle with an enrollment of 6 students. The evaluation focused on how effectively the DSLP and DSCM helped students (1) understand data science principles and practices, and (2) improve their self-efficacy about and interest in data science and computer science. In particular, we designed two sets of hands-on assignments in DSP class, one was to write code using Google Colab, the other one was to use the DSLP to deliver the tasks. The result from end of course student survey showed that students were positive about the course modules and found the DSLP helpful for learning data science better than traditional way, such as writing code. An example of the comments on the DSLP assignments is: *I feel I got most of my learning done through these, and the DSLP platform was extremely useful*. According to the survey results, in CSP classes, 75% students in CSP classes agreed *using the DSLP improved my understanding of data science* and more than 80% students agreed for *using the DSLP improved my confidence in conducting data science inquiries and analytical tasks*. In DSP class, all the students agreed these two statements.

Acknowledgement

This material is based upon work supported by the National Science Foundation under Awards IUSE 2021287. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. The authors thank Dr. Kimberly Fluet for her contribution on designing the survey questions and collecting/analyzing the survey data. The authors also thank the anonymous reviewers for their feedback.

References

[1] Austin Cory Bart, Dennis G. Kafura, Clifford A. Shaffer, and Eli Tilevich. Reconciling the promise and pragmatics of enhancing computing pedagogy with data science. In *Proceedings of the 49th ACM Technical Symposium on Computer Science Education, SIGCSE 2018, Baltimore, MD, USA, February 21-24, 2018*, pages 1029–1034, 2018.

[2] Lillian N. Cassel, Michael Posner, Darina Dicheva, Don Goelman, Heikki Topi, and Christo Dichev. Advancing data science for students of all majors (abstract only). In *Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education, Seattle, WA, USA, March 8-11, 2017*, page 722, 2017.

[3] Jeffrey S. Saltz, Neil I. Dewar, and Robert Heckman. Key concepts for a data science ethics curriculum. In *Proceedings of the 49th ACM Technical Symposium on Computer Science Education, SIGCSE 2018, Baltimore, MD, USA, February 21-24, 2018*, pages 952–957, 2018.

[4] Sushil K. Sharma and Joshua Sefchek. Teaching information systems security courses: A hands-on approach. *Computers & Security*, 26(4):290–299, 2007.

[5] James Walden and Charles E. Frank. Secure software engineering teaching modules. In *In Proceedings of the 3rd annual conference on Information security curriculum development InfoSecCD '06*, page 19–23. ACM, 2006.