Abstract

In this talk we will discuss the 3-year, NSF funded *Data Science for All* seminar series; including the motivation behind extra-curricular seminars, the targeted audience of students and faculty, our experiences developing and delivering the seminars, what’s worked (and what hasn’t), as well as the materials available for faculty at any institution to download and use. Two of the main goals of the seminar series are: (1) increase and diversify the number of undergraduate and community college students aware of data science, and (2) increase and diversify the population of undergraduate and community college students possessing data gathering, wrangling, and cleansing skills.

There is currently a great need to grow our Nation’s data science capabilities. One way to meet this need is by creating data scientists through advanced degree programs – a costly and time-consuming approach. An alternative is to augment the data science workforce with graduates possessing basic data science skills. According to the National Academies of Science, Engineering, and Medicine (2017), many data science roles can be filled by undergraduate students, including data wrangling; the acquisition, profiling, and transforming of data prior to analysis that constitutes 80% of a data scientist’s work. Shifting data wrangling to lesser-trained employees allows data scientists to focus on complex tasks; extending existing data science resources and providing opportunities for a diverse population of undergraduate students.

These seminars seek to provide this much-needed data wrangling workforce by introducing data science concepts to undergraduate students early in their academic careers through low-risk extracurricular seminars. Each seminar includes optional pre-seminar materials that provide a gentle introduction to the topic, a hands-on seminar without prerequisites, and an optional post-seminar assignment they can complete to earn a digital badge attesting to their new skills. Although some seminars (e.g., Python foundations) provide skills that are useful in other seminars, each seminar is designed as a stand-alone unit to minimize the commitment both by students attending and faculty wanting to implement it.

To reach a diverse population and provide them skills in the discipline, the seminars focus on entry-level concepts in data science, using an experiential learning approach and current tools in the field. In 2019, the initial seminar series included foundational seminars in both statistical concepts and Python, data wrangling using Apache Spark and Jupyter notebooks, exploring graph databases, and an introduction to concepts in machine learning. The teaching materials available for download include: slides, Jupyter notebooks, exercises, teaching notes, datasets, Canvas course packages, test materials, and guidelines for setting up digital badges.

Acknowledgement:

This work is funded in part by NSF Grant #1829622

References