With the surge of interest in all things "data", enrollments in traditional data-oriented courses are at an all time high. The rise of Data Science as a discipline has also led to the creation of new courses whose content significantly overlaps that of an introductory database course. This column presents a series of perspectives on data (base/science) education to help educators think about what we should be teaching in our courses, and what resources we should use to teach them.

This episode contains two contributions. The first discusses a recent effort called OpenDS4All, whose goal is to accelerate the creation of Data Science curricula by providing a "starter set" of open source data science training materials, including PowerPoint slides and Jupyter Notebooks with Python code. The second presents a process for making pedagogical decisions of how to construct a data management class given the wealth of content that can be included.

Susan B. Davidson
University of Pennsylvania

OpenDS4All: Accelerating the Creation of Data Science Curricula at Academic Institutions

Andre de Waal
IBM
andre.dewaal@ibm.com

Ana Echeverri
IBM
ana.echeverri@us.ibm.com

Background and motivation

Over the past decade, the interest in data science careers has been unwavering. In 2001, Linkedin ranked data science specialists as one of its top jobs, and noted that hiring for these roles have grown nearly 46 percent since 2019. Yet there is still a significant data science skills gap in the market and part of the problem is education focused. Research conducted by the University of California, Riverside shows that fewer than a third of the US News & World Report's Top 100 Global Universities offer degrees in Data Science, and most programs are still taught at a graduate or PhD level.

The bottom line is that the growth in demand for data science skills currently outpaces the ability of academic institutions around the world to build data science programs. Although Data Science as a field has existed for several decades, the rapid growth of the last decade, the current skills shortage, and the interdisciplinary nature of the field, have contributed to the difficulty in building new programs.

To help academic institutions overcome these challenges, IBM - working with the University of Pennsylvania and the Linux Foundation – brought to the market a "starter set" of data science training materials launched as an open source project: Open Data Science for All (OpenDS4All). The goal of OpenDS4All is to accelerate the development of data science curricula at academic institutions, making it easier for academic institutions to leverage existing education modules built by professors for professors without the need to build every single course from scratch.

Starting a data science program from scratch is incredibly difficult, as building a curriculum requires significant resources. By making a "starter set" of training materials available containing sets of

PowerPoint slides and Jupyter Notebooks with Python code, we aim to help accelerate the availability of data science skills building programs around the world. By making it open source, we can leverage the open source communities for growth of the available educational content with contributions from other experts in the community.

Current and future content

There are currently 16 educational modules in the OpenDS4All repository on GitHub, and we are constantly striving to add new modules covering topics of immediate relevance and interest. The original content covers the core of data science and includes modules on:

- **Overview** (What is Data Science?)
- **Foundations** (How your computer works and graph theory)
- **Data and Knowledge Modeling** (Representing and codifying knowledge)
- **Data Wrangling and Integration** (Getting Started in Data Science: Data Acquisition and Wrangling)
- **Exploratory Data Analysis** (Information Visualization, aka Visual Analytics)
- **Machine Learning** (Building unsupervised and supervised machine learning models)
- **Model Assessment** (Training, validating and tuning robust models)
- **Scalable Data Processing** (Efficient and cluster-based processing with graph data)
- **Ethics** (Ethics, privacy and fairness in data and algorithms)

The latest module that has been added to the repository covers the basics of deep learning on a well-known image classification data set. It was designed to be a "soft introduction" to image processing using convolutional neural networks (CNNs).

With AI Trust being one of the most pertinent topics currently being discussed in AI and Data Science, two new educational modules based on two of the pillars of AI Trust are in the works. The first module under development is focused on AI Fairness (are certain groups at a systematic disadvantage compared to others) using information from the AI Fairness 360 (AIF360) extensible open source toolkit. The second module is on AI Explainability (the ability to provide a clear and relevant explanation of a model's decision) and is based on the AI Explainability 360 (AIX360) toolkit (also open source). We hope to have these modules incorporated into the repository in the not-too-distant future, and several professors have already expressed their interest in the content.

There are many new developments in Data Science such as the increase in use of natural language processing, federated learning, and automation to name a few. Contributions on any of these topics as well as other relevant topics are welcomed and will be reviewed by OpenDS4All's technical steering committee (TSC) for inclusion into the repository.

Although the Jupyter Notebooks can be run in nearly any local and cloud environment, OpenDS4All provides access to a readily available binder environment so that faculty and students with limited computing resources can experiment with the notebooks. The binder link is available from the GitHub repository. This may be especially useful for students who do not have access to a local installation of Python or a virtual lab where they can execute the code.

Usage

The reception of the developed modules has been overwhelmingly positive, with many academics from all over the world commenting positively on the clear structure, engaging content, and applicability to undergraduate and graduate programs from all disciplines.

One of the most successful ways of promoting the content has been through bootcamps and workshops. These workshops and bootcamps are usually structured as follows:

- Introduction to Data Science and Data Analytics and job opportunities (15 minutes)
- Presentation of one or two of the lectures from the repository (45 – 90 minutes)

- Demonstration and hands-on exercise on how to import a Jupyter notebook and how to run it at least one cloud environment (45 minutes)

Students and professors who attended these workshops and bootcamps loved the hands-on aspects of the modules, experimenting with Python code and running the Jupyter Notebooks, in particular those involving data wrangling, acquisition and web scraping. A good example is the bootcamp we did with the University of Liverpool in the United Kingdom during 2021 where 82 students attended, even though attendance was optional. This is written up as a success story that is available from the GitHub repository.

OpenDS4All has also been adopted as a core curriculum component of the National Student Data Corps. The National Student Data Corps (NSDC) is a community-developed initiative that teaches data science fundamentals to students across the United States, with a special focus on underserved institutions and students.

Lessons learned

During conversations with more than 50 academic institutions on OpenDS4All, we learned that universities are still trying to figure out the best way to teach Data Science. Questions that came up included:

- Should Data Science be taught at the undergraduate level or at the graduate level?
- Should programing in Python or R (as well as mathematics and statistics) be a prerequisite for doing Data Science?
- Should we cater for students preferring low-code environments?

We also underestimated the difficulty of incorporating OpenDS4All content into existing data science programs, due to the somewhat lengthy approval process for new courses. Many academics are also not familiar with open source, and as OpenDS4All is run as an open source software (OSS) project this represents a barrier to adoption.

Although the initial goal of OpenDS4All was to accelerate the development of (new) data science programs worldwide, we also learned that existing programs could benefit greatly from the content by supplementing existing content with some of the modules from the repository. For example, big data is becoming more pervasive, and existing data science programs can augment or enrich their content with novel content from the repository on scalable data processing with Apache Spark.

We also learned that language may be a barrier. To make OpenDS4All more inclusive, we have started translating the hosted educational modules into Spanish and Portuguese. These modules will be incorporated into OpenDS4All during the upcoming months.

How to become involved

There are many ways to interact with this repository:

- Browse and download content (PowerPoint slides, Jupyter notebooks, etc.)
- Contribute content (become a contributor to the project)
- Become involved in the day-to-day management of the project (become a maintainer or committer)
- Provide overall direction and leadership to the project (become a Technical Steering Committee member)

Please visit the GitHub repository at https://github.com/odpi/OpenDS4All, fork it to download content, and star the repository to help us improve the content and keep it relevant.

Acknowledgements

Ana Echeverri, who designed and launched OpenDS4All, reached the final shortlist for the AI Innovator of the Year Awards at AI Summit London 2022 (Solutions Provider award) for OpenDS4All, confirming that OpenDS4All presents a novel and innovative approach to reducing the data science skills gap.

Zachary Ives, who designed much of the original content of OpenDS4All, received the IEEE 2022 IEEE TCDE Education Award for fundamental contributions to Data Science education.