12-2021

Data Science for All: Apache Spark & Jupyter Notebooks

Scott Jensen

Leslie J. Albert

Esperanza Huerta

Data Science for All: Apache Spark & Jupyter Notebooks

Tutorial

Scott Jensen
San Jose State University
scott.jensen@sjsu.edu

Leslie J. Albert
San Jose State University
leslie.albert@sjsu.edu

Esperanza Huerta
San Jose State University
esperanza.huerta@sjsu.edu

Abstract

The Nation's research enterprise faces a shortage of data scientists. Expanding the pipeline of data science students, particularly from underrepresented populations, requires educational institutions to increase awareness of data science and inspire a passion for data in students as they begin their academic careers. In this tutorial we discuss the development and delivery of a free seminar designed to provide hands-on lessons in the use of both Apache Spark and Jupyter notebooks to students from any academic background in an approachable, no-risk environment. An explanation of the seminar resources, exercises, and implementation guidelines are included, as are lessons learned from several successful seminars held both in-person and virtually at two institutions of high education.

Keywords

Tutorial, Apache Spark, Jupyter, Big Data, Data Science.

Introduction

The seminar discussed in this tutorial was developed under the Data Science for All seminar series, a set of eight hands-on seminars funded by the NSF's CyberTraining program[1] and designed as extra-curricular activities to introduce both four-year and two-year colleges students to topics and tools in data science and Big Data. The goal of the NSF's program is to create innovative and scalable training programs to develop the nation's data science workforce. Most data science programs focus on training data scientists through post-baccalaureate programs. However, that is a time-intensive and costly approach since 80% of a data scientist's workload is the data wrangling that can be done by less skilled data analysts drawn from a wide range of disciplines (National Academies 2017). The seminars developed in this project take a novel approach to augmenting our Nation's data science workforce by introducing a diverse population of students to the possibility of a supporting role in data science as a data analyst, including students without backgrounds in math and science who may have previously found the concept of data science intimidating.

The seminars were developed and piloted at a minority-serving institution, with transfer students representing over 50% of new undergraduate enrollments, and over half of all incoming students identifying as first-generation college students. Some of the obstacles that prevent these students from pursuing a career in data science include: 1) lack of familiarity with the topic as first-generation college students, 2) fear or a lack of self-efficacy in math, science or technology, and 3) as first-generation students they are often balancing work, school, and family, making the time commitment and expense of pursuing additional structured training prohibitive. The seminar is designed to be an engaging and experiential two to three hour hands-on learning experience that shows them that data science is within their abilities and provides

[1] The Data Science for All Seminar Series is supported by NSF grant #1829622

them with skills they can immediately leverage in other courses. The emphasis is on diving in and doing data science, with less focus on theoretical aspects. Although the eight seminars in the series are all on data science topics, each seminar, including this one on Spark and Jupyter notebooks, is designed so it can be taken stand-alone, without any prerequisites. Instead of requiring students to commit to a specific training sequence of seminars, this approach is specifically designed to pique the student's interest in data science by allowing them to start with the seminar they find most interesting and to build confidence in their skills.

The seminar leverages open data and introduces students to Apache Spark and Jupyter notebooks; two in-demand Big Data and data science technologies. The Spark implementation used in the seminar is hosted for free on AWS by Databricks (a firm founded by the developers of Apache Spark) and available to students worldwide. The seminar also includes optional activities if students want to learn more or pursue a digital badge attesting to their new skills. This seminar has been presented both in-person and virtually to 123 students at a four-year university and 68 students at a local community college. One of the main goals in the development of the seminar is to provide a teaching resource that faculty at any two or four-year college could adopt, in whole or in part, either as an extra-curricular activity, or as one or more assignments in a course. The materials are all available on our website, and any verified faculty can request the additional teaching bundle for free.

The remainder of this paper discusses the development and structure of the seminar, lessons learned in presenting it, and the dissemination of the materials.

Seminar Development

The goal of the seminar is to show students how to use Spark SQL to wrangle a large dataset using a Jupyter notebook interface, quickly query and iterate on asking questions of their data, understand the importance of documenting their work using markdown, and quickly create visualizations and publish their notebook to share their results with the world. We use web-based accounts that provide a notebook interface to Spark for free to students through Databricks[2]; a company spun out of Berkeley's AMPLab and founded by some of the original developers of Apache Spark. As Spark is open-source, there are numerous customizations of the Jupyter interface by different software vendors, but if students learn one implementation, switching between vendors is fairly easy. The community accounts provided by Databricks are hosted on AWS for free and students only need a computer with a browser – even a Chromebook will work. Although we have used web-based implementations from other vendors in the past for MIS classes, a few advantages of the Databricks community accounts are: (1) they have offered a consistently available and stable interface, so seminar materials will remain relevant (2) since it was founded by some of the original developers of Apache Spark, the latest stable versions of Spark are always available, and (3) it is extremely easy to sign up for a community account in contrast to accounts with some of the other vendors.

The Datasets Used

Since the seminars are targeting a general student population, one goal is to make the material relatable to students from any major. We use datasets from USASpending.gov[3] and the Social Security Administration (SSA) (Social Security Administration undated). Similarly, in another of the Data Science for All seminars we use an academic dataset available from Yelp (Yelp undated). USASpending is an open data site which provides details on payments related to government contracts dating back to the Obama administration, which initiated the website. Since detailed payment data is not available as a single download file, we used the website's API to build a dataset covering 2014-2018 which includes all of the data for that timeframe. Although the API is not particularly user friendly, we built a python program that allows us to download and zip the data by year, selecting the fields we want to include. The notebook used to download and build the dataset is available to any faculty who wants to build a different dataset. Alternatively, the files we use are also available and include 25 columns of data on contract payments, the paying agencies and vendor information, including their location and information about their ownership (e.g., minority and women owned businesses). The dataset includes nearly 20 million records. Although the size of the dataset is not

[2] https://databricks.com

[3] https://www.usaspending.gov

Big Data in the commercial sense, it is a manageable size for this purpose and is far larger than the datasets most students have worked with previously. Since the seminar is targeted at students with no prior experience in data science, having a larger dataset helps us demonstrate the power of Apache Spark and the approachability of Jupyter notebooks while building students' data science self-efficacy.

To save time when loading the data, we stage the data on AWS using Amazon's S3[4] (Simple Storage Service). Using S3 has a couple advantages. First, when the seminar is delivered in-person, having a classroom of students all uploading the data over the wireless network at the same time could be very slow. Additionally, as we went online due to the pandemic, this approach was more feasible since students would be uploading from home and may not have a fast Internet connection. If using the materials as part of a regular semester course, faculty may want to have the students upload the data using the Databricks GUI, as we do with the SSA data for the digital badge quiz at the end of the seminar. Since the student accounts are hosted on AWS, we include instructions in the faculty materials on how to set up an AWS account and load the data into an S3 bucket configured to only allow access from EC2 instances in the same datacenter. This enables the code in the student notebooks to access the data quickly and ensures there are no data transfer costs incurred by the faculty. The cost to store the data on AWS is less than 5 cents per month. We also provide a notebook that generates the up-to-date configuration file needed to avoid data transfer costs.

The CSV files containing the USASpending data can be loaded directly into Databricks through the GUI and then read into a DataFrame in Spark. However, since an aim of the seminar is to have students working hands-on with the data, we pre-processed the CSV files as a Spark table using the Parquet file format common across Big Data tools, resulting in a file that takes only three minutes to load. Starting with a table in Spark instead of a DataFrame also allows seminar participants to dive straight into using Spark SQL instead of first having to learn PySpark commands to load the data.

Starting With a Question

To emphasize to the students that data science should always start with a question (Shron, M. 2014), we begin the seminar with "Do government agency spending patterns change based on the political party of the presidency?" By examining the period from 2014 through 2018, we are querying the last three years of the Obama administration and the first two years of the Trump administration. We try to avoid any political commentary and point out that individual political views are not relevant when empirically evaluating how the spending by government agencies may change under administrations led by different parties and by presidents with very different views.

Seminar Structure

The seminar has been presented one to two times per semester and although developed and presented by faculty from the business school, the seminar is designed for students from any background and requires no prior programming or data science skills. The seminar is also designed for a two-year college audience. Faculty adopting the material can leverage any tool their school currently uses to enroll students in extracurricular activities, but for this seminar we use the web-based Eventbrite[5] platform which is free for our level of use and allows us to collect registration information plus some limited demographic data for research and reporting purposes. Materials are made available to students through a Canvas shell. The main components we use in Canvas are the homepage, the modules, and the gradebook for awarding the optional digital badge. The homepages for all of the seminars in the series use a consistent format, and each seminar has three modules: an optional pre-seminar module that introduces the topic and includes setup information, a seminar module that includes all of the material students will use in the seminar, and a post-seminar module that includes additional optional materials and a quiz that students can complete to demonstrate their new skills and earn a digital badge.

All of the materials included in Canvas are included in the faculty bundle as discussed below in the section on dissemination, so the focus here will be on the structure of how the seminar is presented, which has

[4] https://aws.amazon.com/s3

[5] https://www.eventbrite.com

evolved over multiple iterations. The agenda for the seminar is composed of the six components listed in Table 1.

Seminar Component	Time (minutes)
Setup • Sign up for Databricks and log into their account • Start (spin) up a Databricks cluster • Import the seminar notebook & attach to the cluster	20
Part 1: Introduction • Import the seminar data • Discuss tabular data (first poll) • Discuss why Spark & Jupyter	20
Part 2: Spark SQL and Basic Charts – Let's Dive In! • SQL Queries • Wrangle the data Spark SQL to visualize as a bar chart	50
15-minute break (time to catch-up any students who have data or notebook issues)	
Part 3: Exploring Spark SQL • More SQL queries • Visualizations: stacked bars, area charts, and more	30
Part 4: Applying Your New Skills, Sharing with the World • Benford's Law – using data for fraud detection • Publishing your results and sharing your notebook	15
Part 5: Load the Notebook for the Digital Badge • Loading data using the Databricks GUI • Doing some initial profiling	20
Questions	10

Table 1. Seminar Components and Timing

Seminar setup

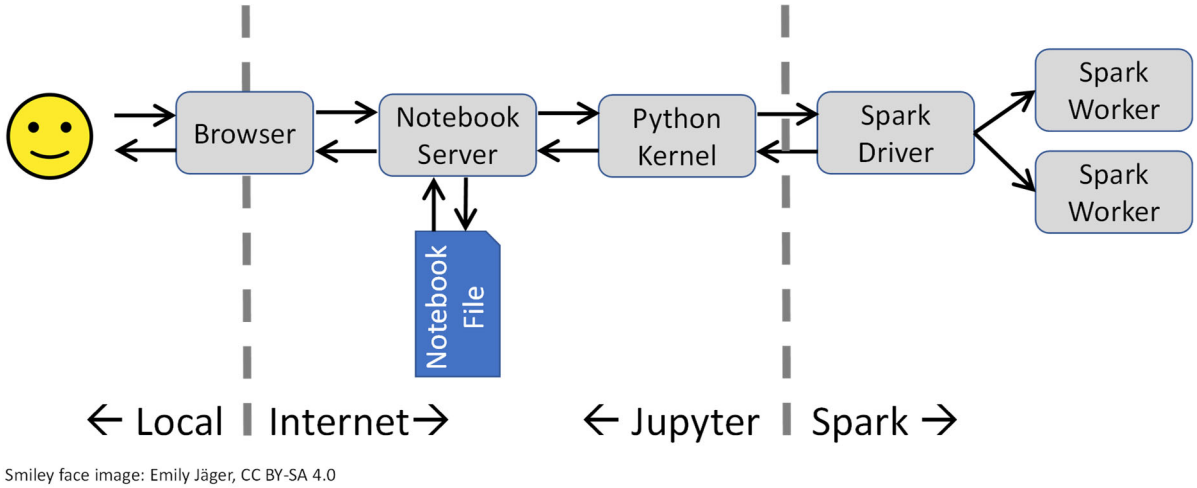
When presenting the seminar virtually, we are online 15-minutes before it starts, and both when presenting online and in-person, we direct the participants who arrive early to sign up for a Databricks account if they have not already done so. During the first 20 minutes, we walk students through the setup process for the seminar, which involves logging into their cloud-based Databricks account, starting a cluster, and importing the notebook. We have created a Jupyter notebook containing partial code which the students then complete during the seminar. The notebook contains extensive documentation as markdown that students can refer back to later. Unlike code comments, markdown compiles into well formatted HTML (similar to any web page). This allows us to describe what the participants are doing and embed graphics that illustrate steps in the process. Notebooks do not contain data, so they are small and easy to share and can be loaded either as a file, or from a URL. Later in the seminar we load the notebook for the digital badge quiz from a file, but initially we load the seminar notebook from our project website. Since clusters take approximately two minutes to start, we walk through importing the notebook while the cluster is starting. After importing the notebook, the cluster is usually running, and students can attach the notebook to the cluster. We describe what is happening behind the scenes at a high level while loading the data in Part 1. These setup steps are also detailed in Canvas with snippets of the slides to avoid losing any participants during the process. Although we allot 20 minutes in the schedule for setup, it is usually completed sooner, which starts us off ahead of schedule.

Part 1: introduction

To guide students through the setup process, we have four numbered steps in Canvas. The first three steps are the setup process as discussed above, and the last step is the loading of the data which we cover in the introduction. In earlier versions of the seminar, we had participants create a DataFrame directly from CSV files compressed using the bzip2 format and then do some initial data wrangling in PySpark. We then created a DataFrame including all five years in a later step of the notebook. This approach had a couple of drawbacks. First, since the seminar's target audience is students who are new to programming and data science, the wrangling in PySpark was often confusing and could be overwhelming as a starting point. Second, creating the DataFrame with the data for all five years later in the seminar added a six-minute step, and if any participant failed to create the DataFrame correctly, it was harder to recover. In the current version of the seminar, it takes three minutes to load the data, and if any participant encounters a problem, we can address it during the break after Part 2 and get them back on track.

To load the data, we create an S3 bucket in our account on AWS, and have code in the notebook that downloads the data from that bucket, generates an MD5 checksum to verify the file has not been corrupted, moves the files to a directory on DBFS in the student's account, and then creates a table in Spark. We use a manifest file written in JSON (that we host on our project's website) to specify what files should be downloaded, where they are staged, and the checksum for each file. This allows us to add or modify the data or where it's staged without changing the notebook itself. The faculty bundle includes directions on setting up an AWS account and how to stage the data and modify the manifest file. We use a "widget" in the notebooks to specify the URL where the manifest file is located. This allows faculty at other schools to post a modified manifest file on the web (such as on their faculty web page) without having to modify code in the notebook itself.

While the data is loading, we discuss at a high level the architecture of Jupyter and Spark, and how they work together as shown in Figure 1. In that diagram, students are the happy user accessing the Internet, loading their notebook into the Jupyter notebook server hosted by Databricks on AWS, and then connecting (attaching) to the Python kernel and Spark session that were started for them when they spun up their cluster. We discuss a bit about the history of both Spark and Jupyter notebooks, that Spark provides them the ability to spread calculations across many computers, and how the introduction of DataFrames in 2013 and particularly Spark SQL in 2016 has enabled the democratization of data – allowing many users within organizations, such as the students themselves, to analyze large volumes of data in a way that would not have been feasible on their laptop directly. When online, we include a poll at this step to keep students engaged while the data is loading.



Smiley face image: Emily Jäger, CC BY-SA 4.0

Figure 1: Architecture of Jupyter and Spark

Part 2: Spark SQL and basic charts

In this part of the seminar, we start with the basic layout of a query and use it to apply data wrangling techniques (Kandel, S. et al. 2011) to profile our data – learning what is in our data, correcting minor errors, and using visualizations to summarize the data. Since one of the goals of the Data Science for All project is to increase the pool of graduates who can perform the 80% of data science that requires only an undergraduate degree, throughout the seminar we emphasize the importance of performing data wrangling to learn about the data. In markdown preceding each code cell we describe what the code is doing and what the student needs to complete, but we also use headers in the markdown to number each step. In PowerPoint, we include that same numbering in the title of each slide. As students can only see a few cells of the notebook on screen, numbering each step makes it easier for them to follow where we are in the notebook and keeps them from getting lost. In every notebook, Jupyter automatically numbers each cell, but this numbering should not be used when identifying steps in the slides. A feature of notebooks is that you can insert a cell anywhere, so if someone asks a question during the seminar, we often respond by having the students insert a cell and write an ad-hoc query. The act of inserting a cell will cause Jupyter to automatically renumber all subsequent cells.

When discussing concepts such as DataFrames and writing SQL queries, we have found it useful to relate these concepts to mental models and everyday practices that may already be familiar to the students. For example, when discussing the tabular layout of data as rows and columns for the SELECT and WHERE clauses, we use Excel spreadsheets as a mental model. Similarly, when introducing the GROUP BY clause, we use sorting M&M candies as an example. At the end of this segment, if any participants are still encountering problems with the data or their initial queries, the 15-minute break allows time to get them back on track while other participants take a break. Although we encourage all students to try writing the queries, we also include a text file in Canvas containing all of the SQL queries and the plot options for each chart.

Part 3: Exploring Spark SQL

In this part of the seminar we incrementally build a more complex query. We start by first listing what we want our query to tell us, and then, using a color-coded query template we introduced in Part 2, we build and run the SQL query. The overall goal of the query is to see a summary by year of the agencies that spent the most. At the end of this segment, students walk through creating a stacked bar chart using the display() function. Although there are Python tools capable of creating more beautiful and complex visualizations, the display() function allows students to visually profile their data very quickly, which is an important skill for data analysts to have (Kozyrkov, C. 2018).

Part 4: Benford's Law and publishing

In this segment of the seminar, students apply a fraud detection technique known as Benford's Law which analyzes the first digit of each payment and generates an interesting visualization that can, at first, seem counterintuitive. When presenting online, we poll the student as to what they expect to see in their results. This SQL query is more complex, and the students are not expected to write the query, but we do introduce the HAVING clause using the same color-coded approach. Here the focus is more on working with the display() function to generate the 100% stacked bar chart shown in Figure 2.

Since this is the end of the notebook (except for optional material on joins), we walk the students through the process of publishing their notebook; which allows them to show their work to others. If they use Spark on future class projects, they can publish and share the results on their LinkedIn profile or other social media sites. If the materials are used in a graded course, the URL for the published notebook can be submitted as the assignment deliverable and it's one of the deliverables for the digital badge quiz.

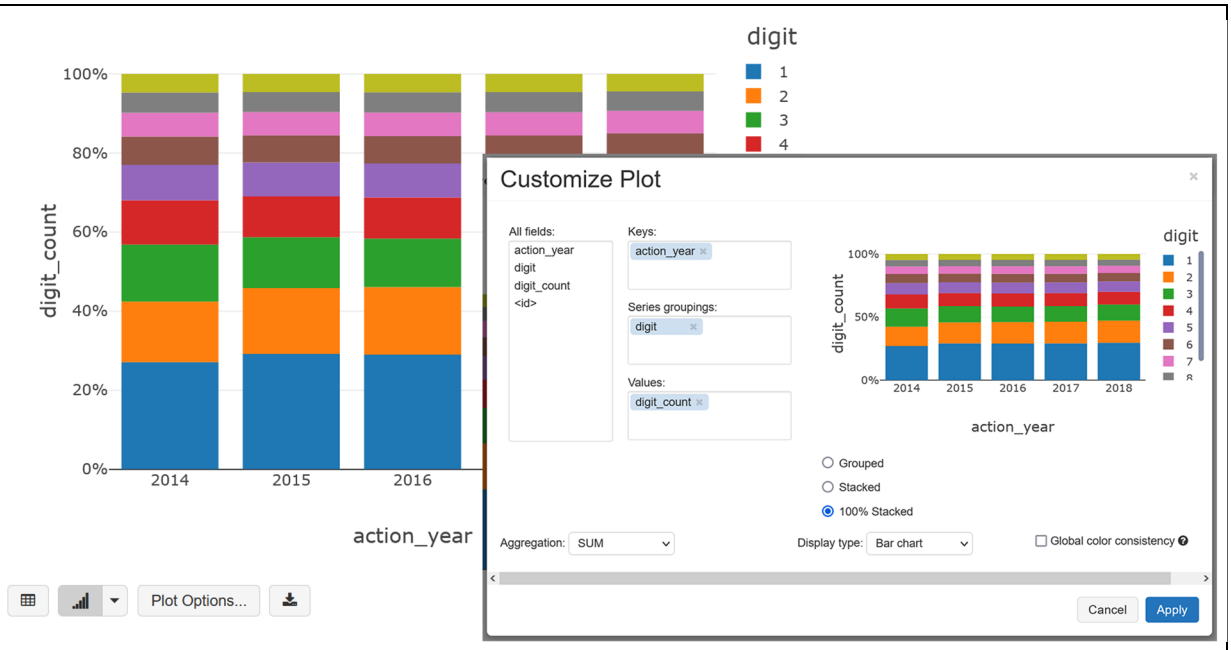


Figure 2: A 100% stacked-bar chart used to illustrate Benford's Law

Part 5: Starting the digital badge quiz

When we initially presented the seminar, digital badges were not yet common, so we found we needed to "sell" students on the idea of posting a digital badge on their LinkedIn profile. Digital badges are now more common, but we have found that it's important to spend time getting students set up to take the digital badge quiz. The quiz includes questions requiring them to write queries in a notebook, run the queries, and use the results to answer the quiz questions. For the quiz, we use a different notebook and datasets from the Social Security Administration (SSA) that contains information on the number of men and women born each year (who applied for a social security card), summarized by their first name. The national data covers 1880-2020, and the state level data covers 1910-2020. We use both datasets in the quiz and load them using the GUI interface. After loading the data, we walk through some basic data wrangling queries. Since students will most likely complete the digital badge quiz at a later date, we include instructions in Canvas on how they can work with their notebooks using a new cluster. Since the Jupyter interface also allows them to clear the memory (state) of the notebook without terminating, we can have students clear and recalculate as they would after starting a new cluster. That shows them how easy it is to pick up the process again later.

Lessons Learned

The Spark and Jupyter notebooks seminar is designed for students with no background in data science or programming and we often see a wide range of both technical knowledge and basic computer skills. One approach that we have found very helpful is to have one to two student teaching assistants (TA's) who have worked through the materials available to walk around and debug any problems students encounter. The TA's need to be able to quickly get students back on track, but also know when to interrupt the presentation if multiple participants are encountering the same issue. Without TA assistance, participants will too often wait until they are thoroughly stuck before asking questions. We have also found it helpful to congratulate students when they encounter an error. We point out that they will learn more from debugging errors than from having code that executes correctly the first time. As long as this is done in a light-hearted and humorous manner, it can help reduce their resistance to asking questions. As we pivoted to presenting online, the student assistants were able to answer questions in the chat window, but for more difficult questions this is challenging. They can open a breakout room in Zoom so the student can share their screen, but the TA and participant are then both "out of the room".

The Canvas course includes a pre-seminar module that provides a gentle introduction to the topic, and in the initial version included steps we wanted participants to complete before attending, such as creating a Databricks account and loading the data. However, a goal of the seminar is to provide an introduction to data science with minimal start-up effort and we found that few participants had completed these preparatory steps beforehand. Fortunately, it's easy (and fast) to create a Databricks account. Additionally, we have changed the process for loading the data so it only takes three minutes and can be performed while covering a couple slides with an overview of the tools we will be using.

We provide a PDF version of the slides during the seminar, but we found that for the initial steps it's also helpful to include snippets of the slides as ordered steps in Canvas (creating a Databricks account, starting a cluster, importing the notebook, and loading the data). If a participant falls slightly behind, but is not willing to ask a question, this allows them to catch up more easily. As we switched to an online format, we also found more participants arrived late. Adding these ordered steps in Canvas provided a path the TA's could use to get late arrivals up to speed.

The aim of the seminars is to introduce students to data science, and not to grade them, so we found it helpful to include a text document in Canvas that contains the code for each of the queries and the plot options for each of the charts. We encourage the participants to work with us through the queries, but point to this file as an additional resource if they are getting stuck. Particularly for participants who are less technical, this can be a useful crutch. Although some may simply copy and paste their queries, they are still able to run the code and see the results in their notebook. We also include a URL in Canvas where they can view the completed notebook online.

We also found it beneficial to focus seminar materials more on specific skills, leaving more advanced skills as optional materials student could explore later if they wanted to dive further into data science. In the initial version of the seminar, we covered both some PySpark methods and Spark SQL queries. However, participants who did not have prior experience programming found the dot syntax used to call the methods of an object could be difficult to grasp in the time available. In subsequent versions of the seminar, we address this by using only Spark SQL, allowing more time to focus on the basic structure of a SQL query. Additionally, in early iterations we included SQL joins, which are an important feature in SQL, but are best left as an optional topic included in the notebook for exploration afterwards.

Dissemination

Student resources and instructor bundles for all eight seminars are freely available under the Creative Commons Attribution-ShareAlike 4.0 International license. This license gives users the right to "copy and redistribute the material in any medium or format," and "remix, transform, and build upon the material for any purpose, even commercially" (Creative Commons undated). The only requirement is that the materials are attributed to the authors of the seminar. Student resources can be found on the Data Science for All website[6] along with an extensive set of data science resources for both students and faculty. Faculty bundles containing additional teaching materials are available on request from the website following verification of faculty status (via a URL to an institutional faculty webpage). Seminar materials are also indexed on Merlot[7], a site developed to host free instructional resources for college faculty.

The faculty bundle includes the PowerPoint slides, the Jupyter notebooks used in the seminar and digital badge quiz, instructions provided to students on creating accounts and signing up for materials, introductory materials, questions (and answers) from the digital badge quiz, teaching notes, the data sets used in the seminar and digital badge, as well as suggestions on timing and presentation of the materials. Also included is a Jupyter notebook used to download data through the USA Spending.gov API and build the seminar dataset. That notebook is configurable to include additional or different fields in the dataset. The dataset used in the digital badge quiz is based on data from the Social Security Administration (SSA). We did some minor data wrangling to enhance the SSA data and combine it into a single compressed file that students can easily load through the Databricks GUI. Since that data source is updated annually, we

[6] https://www.sjsu.edu/datascienceforall

[7] https://www.merlot.org/merlot

also provide the notebook that downloads the data and performs that wrangling. This allows other faculty to either use the data file we provide or create an updated data file in the future.

We use the Canvas learning management system (LMS) to manage the materials and present them to students as modules (with the digital badge quiz as part of the post-seminar module), so for faculty who are using the Canvas LMS at their institution, we also provide a Canvas cartridge that can be imported into an empty Canvas course shell. For those faculty at institutions that do not use Canvas, we also provide Word documents with all of the content from the Canvas pages, as well as the quiz questions, and documentation as to how the materials are structured in Canvas. As mentioned previously, the Spark and Jupyter notebooks seminar is designed to provide a gentle introduction into data science tools and techniques for participants from any academic background; no prior knowledge or experience in data science or programming is needed. As such, faculty may find this seminar to be a valued addition to existing technical courses, such as those in data analytics, data science, or related fields, or non-technical courses, such as courses designed to expose students to STEM career options. This seminar may also be used in conjunction with the other project seminars to create a survey course of data science-related topics. Alternatively, this seminar could be presented as an extracurricular learning opportunity for college students, faculty, and staff, or as part of continuing education or community outreach programs.

Conclusion

Addressing the Nation's shortage of data science capabilities requires novel approaches to expanding the number of undergraduate students interested in data science-related programs and careers, especially those from populations underrepresented in data science. This tutorial provides one such approach – a well-tested, introductory seminar on Apache Spark and Jupyter notebooks that is designed to help participants from a wide range of backgrounds and disciplines feel confident in their ability to learn and use data science tools and techniques, to get excited about playing with data, and to consider careers in data science-related fields. We hope the freely available resources, exercises, implementation guidelines, and lessons learned explained here will streamline faculty efforts and encourage others to offer training on these in-demand tools at their own institutions.

Acknowledgements

The Data Science for All Seminar Series is supported by NSF grant #1829622.

References

Creative Commons, undated. *Attribution-Sharealike 4.0 International*. Last accessed: 9/26/2021, from https://creativecommons.org/licenses/by-sa/4.0

Kandel, S., Paepcke, A., Hellerstein, J. and Heer, J. 2011. "Wrangler: Interactive Visual Specification of Data Transformation Scripts", *CHI 2011*, May 7–12, 2011, Vancouver, BC, Canada.

Kozyrkov, C. 2018. "What Great Data Analysts Do – and Why Every Organization Needs Them", Harvard Business Review Blog. Last accessed: 9/26/2021, from https://hbr.org/2018/12/what-great-data-analysts-do-and-why-every-organization-needs-them

National Academies of Sciences, Engineering, and Medicine, 2017. *Envisioning The Data Science Discipline: The Undergraduate Perspective: Interim Report*, Washington, DC: The National Academies Press (doi 10.17226/24886).

Shron, M. 2014. *Thinking with Data*, Sebastopol, CA; O'Reilly Media, Inc.

Social Security Administration undated. Beyond the Top 1000 Names. Last accessed: 9/26/2021, from https://www.ssa.gov/oact/babynames/limits.html

Yelp undated. Yelp Open Dataset. Last accessed: 9/21/2021, from https://www.yelp.com/dataset