JAMIA Open, 1(2), 2018, 159–165
doi: 10.1093/jamiaopen/ooy040
Advance Access Publication Date: 1 October 2018
Application Notes

Application Notes

An overview of two open interactive computing environments useful for data science education

Robert Hoyt[1] and Victoria Wangia-Anderson[2]

[1]Pensacola, Virginia Commonwealth University Health System, Florida, USA and [2]University of Cincinnati, Cincinnati, Ohio, USA

Corresponding Author: Robert Hoyt, MD FACP, 304 Port Royal Way, Pensacola, FL 32502, USA (rehoyt@gmail.com)

Received 2 January 2018; Revised 2 August 2018; Editorial Decision 22 August 2018; Accepted 31 August 2018

ABSTRACT

Objective: To discuss and illustrate the utility of two open collaborative data science platforms, and how they would benefit data science and informatics education.
Methods and Materials: The features of two online data science platforms are outlined. Both are useful for new data projects and both are integrated with common programming languages used for data analysis. One platform focuses more on data exploration and the other focuses on containerizing, visualization, and sharing code repositories.
Results: Both data science platforms are open, free, and allow for collaboration. Both are capable of visual, descriptive, and predictive analytics
Discussion: Data science education benefits by having affordable open and collaborative platforms to conduct a variety of data analyses.
Conclusion: Open collaborative data science platforms are particularly useful for teaching data science skills to clinical and nonclinical informatics students. Commercial data science platforms exist but are cost-prohibitive and generally limited to specific programming languages.

Key words: data science, data mining, data interpretation statistical

INTRODUCTION

Data science has been defined as "the scientific study of the creation, validation and transformation of data to create meaning."[1] While there is not a clear consensus regarding the exact definition of data science, one could argue that data science is a useful umbrella term, encompassing other terms such as data analytics, data mining, and machine learning. Data science includes all data processes from start to finish, in contrast to data mining and data analytics that focus primarily on the final analytical step. The first published use of the term "data science" was reported in a paper by William Cleveland in 2001, in which he called for expansion of the scope of statistics.[2]

Data science has experienced recent popularity which has resulted in multiple new courses and the creation of data science centers across the nation. Data scientists are in great demand today and sought after in all industries. The fundamental skill sets required for data scientists are: mathematics, statistics, domain expertise, programming in multiple languages, database management, data visualization, predictive modeling, machine learning, and big data expertise.[3] According to McKinsey Global Institute, there will be a shortage of 140 000 to 180 000 data scientists by 2018.[4]

As healthcare organizations seek to utilize the growing amounts of electronic health data in a strategic and meaningful manner, there will be realization from health informatics academic programs to integrate data science principles and applications into curricula. This will be necessary if academic programs aim to train graduates prepared for the new job opportunities data science affords in healthcare. To support biomedical data science education, new tools are needed capable of producing descriptive, visual, and predictive analytics rapidly and easily in a collaborative environment. Many healthcare workers can perform simple descriptive and visual analytics using spreadsheet tools but lack the expertise to query databases using programming languages or perform

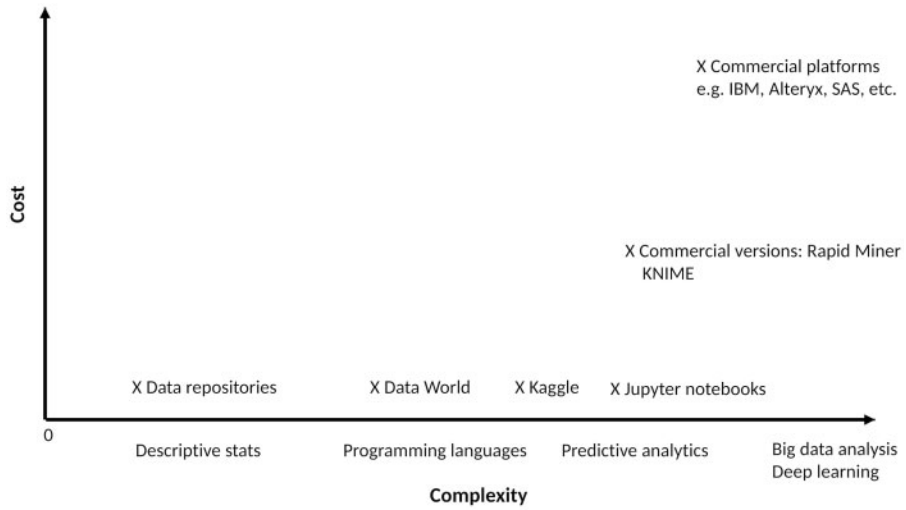

Figure 1. Data science platforms cost versus complexity.

predictive analytics using either advanced statistics or machine learning algorithms.[5]

A wide variety of web-based data science platforms are currently available. In their simplest form, they may be open data repositories, such as the NIH Data Sharing Repository, Data.Medicare.gov, and the Harvard Data Explorer.[6–8] They are "open" to the public but are not truly "collaborative" (there is no sharing function) and lack analytical tools and tutorials. Open, in the context of data use means users "can freely access, use, modify and share for any purpose."[9] Data collaboratives are where "participants from different sectors, in particular companies, exchange their data to create public value."[10]

At the other end of the spectrum are many commercial data science platforms that are closed to the public, can be collaborative only with a subscription, have extensive analytical tools and tutorials, but are associated with a substantial subscription cost, which is rarely disclosed.[11–15] Most of the commercial platforms can handle "deep learning" and "big data" which requires specific software, more processing speed, and robust storage. Data science platforms are popular, as evidenced by the addition of five new commercial platforms in 2017.[16] The commercial platforms tend to be geared towards experienced data scientists.

In the middle, are several data science platforms worth mentioning. RapidMiner data science platform offers a free and a fee-based comprehensive version and KNIME is a similar program that is based on open-source software. Each of these are code free platforms that utilize "visual operators" that are arranged in a sequence to execute a process, such as data preparation or data analytics. Both offer machine learning functionality, as well as integration with the programming languages R and Python. However, enterprise commercial versions are needed for true collaboration.[17,18] Kaggle is a data science platform traditionally used for data competitions. It does however offer "open datasets", programming languages and a collaborative space for a team approach. Creating a data project requires knowledge of either R or Python programming language.[19] Figure 1 shows a diagram of the spectrum of data science platforms with cost on the y axis and complexity on the x axis.

In this Application Note, we will highlight two affordable open collaborative data science platforms that support data science education for clinical and nonclinical faculty and students.

Objectives

The primary objective is to demonstrate and discuss two open collaborative data science platforms, capable of analyzing healthcare data using descriptive, visual, and predictive analytical tools and how they might be used in informatics education.

METHODS

General

Data World (DW) was selected as a data science platform for education because it is affordable (free for academic use), open and collaborative and has commonly used analytical tools. Students can select from a wide variety of health-related DW datasets or upload their own. They can begin with exploratory data analysis (EDA) using the internal spreadsheet tools to look for missing data, visualize distributions and identify the type of data (numerical, categorical, etc.). The next logical step would be to use the internal SQL tool to combine and refine tables to answer a question or hypothesis. This step is supported by SQL tutorials. Finally, as the student's data science education and experience progress they can connect their datasets externally with programming languages such as R and Python to conduct more advanced exercises such as predictive modeling using machine learning.

Jupyter Notebooks were selected as a recommended option for educators and students because it is both free and an effective environment to leverage data science projects, teaching, and learning. It is easy to install and use, minimizing the learning curve for students and faculty. It is web-based and offers interactive computing. This is a feature that many other programming editors do not have. Jupyter Notebook was initially revamped to support three main data science programs: Julia, Python, and R. However, over time it has continued to expand to support many more programming languages. Jupyter Notebooks can be easily shared and offer an environment that allows for results of executed programming code to be displayed in the same notebook as other outputs such as images, visualization, equations, and video. Comments and narrative can be added to the notebook allowing users to gain insight from the notebook. HTML and markdown (using markup language) can be added alongside executable code. This allows users of the notebook to share

Table 1. Initial project page and user interface functions

Feature	Function
Initial page when launched	
New project/file creation and file upload	+Add icon: create new data project and/or upload data files. Files can be csv, xlxs, json, xml, zip, or pdf. File can be added from a URL as well. Data can be shared from Box, Dropbox, and Google Drive
Overview	State the objectives of the data project
Insights	Ability to postdata visualizations on main project page with comments
People/Access	Project participants can be selected with credentials to view only, view and edit, or view, edit, and manage
Discussion	Section for comments by internal and external individuals
Settings	Input name, objectives, open vs private status of data
Like and share button	Share on Facebook and Twitter
Data sources	Ability to add files or link to a dataset. Also includes documents such as data dictionaries to explain data. SQL and SPARQL query results are also posted
Icons	Access to help portal, API documentation, connectors, intro datasets, video tutorials and tutorials of SQL, SPARQL, and Markdown languages
	Ability to send invitations to collaborators
User interface functions	
Workspace	Another option to upload files, write SQL queries and read data dictionaries
Data sources	Same as explained in data sources above
File specifics	Lists the author, when created and modified, file size, labels and description, % missing and data alerts (such as those values > 4 standard deviations above normal) and the ability to download to edit and fix file
Column options	Sort ascending or descending
Data	Where the user would view raw data in a spreadsheet format. "i" icon lists missing data, distinct values, mean, minimum, maximum, standard deviation, skewness, and kurtosis
Chart	Creates charts automatically or manually. Results can be published as insights. Charts can be bar, line, pie, stacked bar multiline, scatter plot, grouped bar, area, or bubble. The chart axis can be flipped
Source	Raw data can be viewed, for example, comma separated
Connect to 3rd party tools	Connect to Python, R, Tableau, KNIME, Chart Builder, Excel, Google Data Studio, Jupyter, Plotly, Power BI, Keshif, MicroStrategy, SPSS Modeler, and SPSS Statistics. Integrated with Canvas learning management system. Future addition: integration with Blackboard learning management system (LMSs)
Other	Query the file (SQL or SPARQL), delete or report an issue

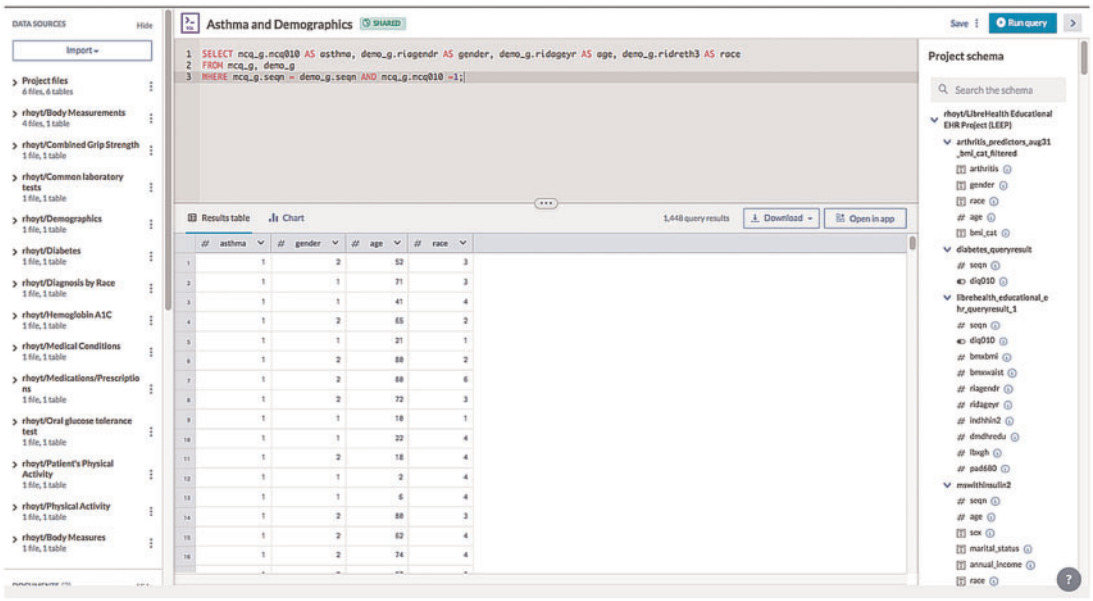

Figure 2. Data query illustration using data world.

meaningful documentation and organize their work well while enabling users to track their work overtime efficiently. Educators can gain insight from the detailed documentation and the outputs. Students can learn from each other once they share their work. Educators seeking to teach a variety of languages, for example R and Python are able to do it all using one type of editor.

Table 2. Main page and user interface functions

Feature	Function
Initial Page when Jupyter Notebook Launches—Dashboard	
New project/file creation and file upload	Able to create new notebook and/or upload data files. Files can be csv, Excel, SAS, text files, or other files. The name of the Jupyter Notebook along with the file extension is listed within a subdirectory. Under the "Running" tab all the running Jupyter Notebooks are shown and any or all can be Shut down from that interface
Overview	State the title of the notebook. A title or headings can also be added to cells where code is written
Insights	Ability to insert files or write code to present data visualizations with comments into the Notebook Editor
People/Access	Through Jupyterhub users and authentication can be managed using PAM, OAuth or integrated with directory service systems. Files can be shared externally on GitHub and viewed using Jupyter Notebook Viewer or accessed through a server. Token-based authentication is an available option since the more recent version of Jupyter Notebook was released
Discussion	Comments can be added anywhere in the code and are typically added by the author or anyone with permission to make edits. This creates narrative that can be included alongside the code as discussion or instructions. Markdown cells are a feature of Jupyter Notebooks and are instrumental for adding markup to Jupyter Notebooks. For example, horizontal rules, headings, equations, and lists are examples of what can be added in the markdown cell as markup language
Settings	
Like and share button	There is no Like or Share button, but you can render Jupyter Notebook in GitHub. The Notebook is also available in html format allowing those with the link to have access. The notebook can also be e-mailed, shared via Dropbox, or made available on the Jupyter Hub server
Data sources	Ability to load actual data files or links to the data. Jupyter Notebook can connect to relational databases using ODBC. Data can also be loaded into Jupyter Notebook as a dataframe. Programming code is executed to create the dataframe
Icons	Featured in the user interface not this first page
User interface	
Jupyter Notebook Editor	It is user-friendly interactive computing environment through which live code is written and executed. Jupyter Notebook is launched through a terminal. In this case, the workspace is the Jupyter Notebook Editor. New Notebooks can be created here and saved. Copies can be created as well. File can be downloaded from this Editor as the Notebook file (.ipynb), Python (.py), HTML (.html), rest (.RST), and PDF (.pdf), for example. The notebooks can also be exported as PDF, HTML, slide shows etc.
Data sources	Same as explained in data sources above
File specifics	A file can be opened in the Jupyter Notebook Editor. The file includes a menubar with options such as File, Edit, View, Insert, Cell, Kernel, and Help. It includes a Toolbar and cells. The Kernel being used is also displayed in the file. From the editor menubar, the Jupyter Notebook file can be saved
Column options	Jupyter Notebook supports a large number of programming languages such as Python, R, Julia, Ruby etc. Code is written and executed against the loaded data which would be in the form of a dataframe in Python, for example. Indexes can be assigned to columns, subsets of columns in the data can be presented, methods can be applied to columns, new columns can be added to the data
Data	Data are viewed by executing code to display it. Statistics are generated based on the code that is written. Programming code has to be written and executed against the dataframe to present missing data, distinct values, mean, minimum, maximum, standard deviation, skewness, and kurtosis
Chart	Charts are created by executing written programming code. A successful execution results in the visualization/s. Results can be published as insights. Charts can be bar, line, pie, stacked bar multiline, scatter plot, grouped bar, area, or bubble. Programming code can be written to change the formatting of the axis
Source	Raw data can be viewed by executing code that shows the dataframe within a cell in the Notebook
Connect to 3rd party tools	Accessing 3rd party tools is often done through already installed packages or by writing script to install the needed package. Widgets can also be added such as buttons and text input
Other	Help in the menu bar offers a user interface tour; help can also be sought through the Github help repository; output in Jupyter Notebook can contain HTIML, images, videos, equations, plots in addition to the statistical output

Materials

Data World (DW) is a Public Benefit Corporation, established in 2016 and free for academic use. As of mid-2018, there were about 16 000 projects and datasets listed on DW covering multiple industries. Approximately, 3600 health-related projects and datasets have been uploaded and new datasets are added weekly.[20] The main DW page consists of the following sections demonstrated in Table 1. Once a new project is created, the user has the option to "launch workspace" with the following features also summarized in the table.

SQL and SPARQL queries can be performed on the datasets. SPARQL is a newer programming language for querying data stored in the resource description framework format.

A new DW project was created by author R.H. based on uploading data from approximately 9600 NHANES patients from the 2011 to 2012 collection period.[21,22] These data were selected because: (1) they are freely available and in the health domain. One hundred twenty-five NHANES data tables (csv files) and data dictionaries are available on the website to users via an external link.

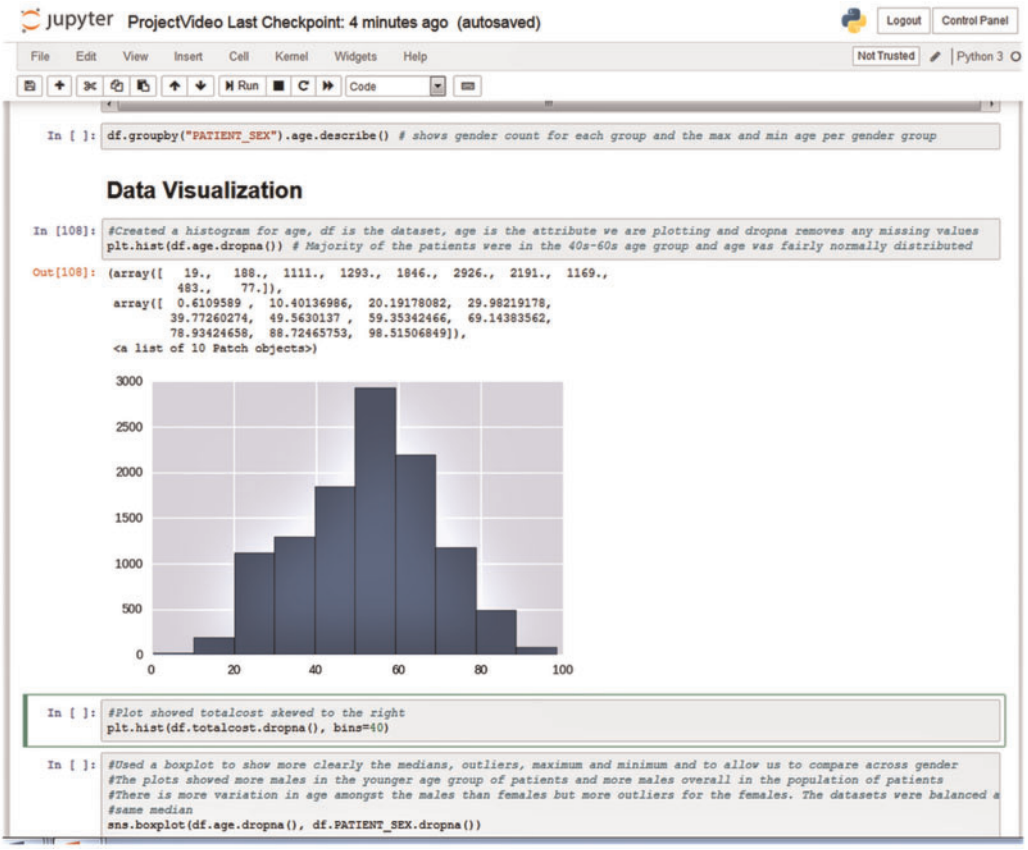

Figure 3. Data visualization using Jupyter Notebook.

(2) Some of these same data were used to populate an open-source electronic health record, known as LibreHealth, being developed for education.[23] Figure 2 displays a SQL query at the top that combines demographic and medical condition tables with the output noted below in a spreadsheet. After a SQL query is run, the user can post the results as an "insight" on the main project page or export it as a csv or xlsx file, copy the URL or connect to a third-party tool, as listed in Table 1. Author (R.H.) has posted 15 insights on the project website with methodology, results, and comments. Insights were based on questions such as "how many medications are taken per NHANES patient?" (descriptive statistics) or "does hand grip strength correlate with pulmonary function testing"? (linear regression) or "do demographics predict insulin resistance?" (logistic regression). Other insights compare NHANES data as a US population benchmark with other datasets reported in the literature. Completed SQL queries are retained and can be shared for educational purposes. This platform was not designed to handle big data or deep learning at this time.

Jupyter Notebooks

The name Jupyter is derived from the programming languages **Ju**lia, **Py**thon and **R** which were supported through Project Jupyter.[24] Jupyter Notebook began as IPython. The project develops open-source software and open standards for computing across multiple programming languages. The notebook is an open-source application and client–server architecture that facilitates the creation and sharing of documents that contain live code, equations, visualizations, and text. The client is what we see, and the server runs the application and makes it available in a browser. Notebooks are commonly used for data preparation, statistical modeling, data visualization, and machine learning. Notebooks consist of three basic elements: (1) documents that contain both code and text, so they are machine and human readable. (2) Jupyter Notebook App is an application that permits the editing and operation of notebooks in a web browser. The App displays the Notebook, as well as a dashboard that serves as a control panel for local files. The App can function offline or installed on a remote server (Jupyter Hub). (3) Kernel is the computational engine that executes the code in the Notebook.[24]

Given the ability to include text and code, this platform has been used by faculty to teach common programming languages, such as R and Python and to collaborate on data science projects.[25] Jupyter notebooks function as both lab and presentation notebooks. Because they contain text and code they can function as "lab notebooks" so data insights can be generated and shared with others working on similar data challenges. General features and functions are listed in Table 2.

Data Science includes processes such as identifying data and acquiring it, cleaning and preprocessing the data, visualizing, modeling and analyzing the data, and communicating findings. The visualization process was utilized for data exploratory purposes and resulted in the graphic presented in Figure 3. Jupyter Notebook was selected because it allowed for user-friendly documentation, built-in graphic presentation, ability to be embedded into a web browser, line-by-line code execution, and the complete integrated development environment (IDE). Python code was written to generate and display the visual plot illustrated in Figure 3. The executed code installed and

imported a Python data visualization library called Matplotlib. In Figure 3, the code transformed the data into a graphic in the form of a plot to effectively represent demographic information about the patient population. Documentation was maintained within Jupyter Notebook by including comments illustrated in Figure 3. Once Jupyter Notebook was installed, the server ran locally and was accessible from the localhost. The notebook server was therefore accessible from the browser.

RESULTS

Figures 2 and 3 demonstrate common platform functionality that students would likely use in data science exercises.

DISCUSSION

This Application Note discusses a variety of new data science platforms currently available, from simple to complex. Two open and collaborative platforms are emphasized because they are free for use by faculty and students and offer novice to intermediate data science experiences. Examples of data manipulation, visualization, and analysis are possible using the two platforms. Because these platforms are integrated with the most common programming languages for data analysis (Python and R), the platforms could be used by informatics faculty and students to augment data science education. Both platforms also provide an opportunity to utilize machine learning, an increasingly common approach to healthcare predictive analytics. DW is a platform most appropriate for those new to data science because it emphasizes EDA with visualization and programming with SQL. Jupyter Notebooks assume programming knowledge and are best suited for users with moderate data science knowledge and experience.

Open collaborative data science platforms that are populated with real or secondary patient data, operational data, or population health data would likely appeal to a wide audience of clinical and nonclinical faculty and students. A new data project would be created for a class with group assignments to perform exercises or investigate a hypothesis. Clinical students could generate descriptive statistics on epidemiological data or other open biomedical datasets. Nonclinical students, such as health informatics and health information management students could benefit from an entire range of data analytics from simple descriptive analytics and visualization to advanced predictive modeling. This would be particularly useful, when analyzing real world healthcare data, such as those found in NHANES data. Information security and ethics are both relevant in the context of these tools. Access should be limited to authenticated users. Credentials should not be sent without encryption, and end-users should not trust a notebook from a source that is not trusted. This would be useful information to include in training developed by faculty who use these tools to educate students.

CONCLUSION

Data science is a new and broad information science field that demands multiple skill sets. Until there are adequate data scientists, it is likely that many healthcare organizations and academic programs will want to utilize the "Wisdom of Crowds"[26] with a wide variety of internal and external experts such as statisticians, computer scientists, and data analysts. Having a spectrum of data science platforms is important for users with variable data science expertise. Affordable open collaborative data platforms associated with a variety of analytical tools will encourage team work and shared expertise.

In this Application Note, we presented two "sandbox" platforms appropriate for a variety of clinical and nonclinical faculty and students involved with data science and informatics education.

FUNDING

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of interest statement. None declared.

CONTRIBUTOR

Dr Hoyt was responsible for the conception and design of the paper as well as the writing and revision of the majority of the paper and final approval. Dr Wangia-Anderson contributed the content for the section on Jupyter Notebooks, helped revise the entire paper and gave approval of the final version. The authors agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

REFERENCES

1. Data Science Association. http://www.datascienceassn.org. Accessed August 10, 2017.
2. Cleveland DS. Data science: an action plan for expanding the technical areas of the field of statistics. *Int Stat Rev* 2001; 69 (1): 21–6.
3. Donoho D. 50 Years of Data Science. R Software. 2015;41. Technical Report, Department of Statistics, Stanford University, 2015. Presented at the Tukey Centennial Workshop, Princeton U. http://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf.
4. Davis P. McKinsey Report Highlights the Impending Data Scientist Shortage Pivotal Blog. 2013. https://blog.pivotal.io/data-science-pivotal/news/mckinsey-report-highlights-the-impending-data-scientist-shortage. Accessed July 6, 2017.
5. McNeill D. *Analytics in Healthcare and the Life Sciences. Strategies, Implementation Methods, and Best Practices*. Upper Saddle River, NJ: Pearson; 2014.
6. National Library of Medicine. NIH Data Sharing Repositories. https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html. Accessed November 2, 2017.
7. Data.Medicare.gov. https://data.medicare.gov. Accessed November 4, 2017.
8. Dataset Explorer. Harvard University. https://nhanes.hms.harvard.edu/transmart/datasetExplorer/index. Accessed February 2, 2018.
9. Open Definition. www.opendefinition.org. Accessed November 4, 2017.
10. Data Collaboratives. www.datacollaboratives.org. Accessed November 4, 2017.
11. IBM Data Science Experience. https://datascience.ibm.com. Accessed November 4, 2017.
12. Dataiku. www.dataiku.com. Accessed November 4, 2017.
13. Domino Data Science Platform. www.dominodatalab.com. Accessed November 4, 2017.
14. DataScience. www.datascience.com. Accessed November 4, 2017.
15. Alteryx. https://www.alteryx.com. Accessed November 4, 2017.
16. Piatetsky G. KD Nuggets. Gartner 2017 Magic Quadrant for Data Science Platforms: gainers and losers. https://www.kdnuggets.com/2017/02/gartner-2017-mq-data-science-platforms-gainers-losers.html. Accessed November 8, 2017.
17. RapidMiner Data Science Platform. https://rapidminer.com. Accessed November 8, 2017.

18. KNIME. https://www.knime.com. Accessed November 8, 2017.
19. Kaggle. https://www.kaggle.com. Accessed November 8, 2017.
20. Data World. https://data.world. Accessed May 25, 2017.
21. LibreHealth Educational EHR Project. https://data.world/rhoyt/libre-health-educational-ehr (sign in may be required). Accessed June 15, 2017.
22. National Health and Nutrition Examination Survey (NHANES). https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2011. Accessed June 6, 2017.
23. LibreHealth EHR Project. https://librehealth.io/teams/education/. Accessed June 15, 2018.
24. Project Jupyter. https://jupyter.org. Accessed May 20, 2018.
25. Berkeley Division of Data Sciences. Data 8: Foundations of Data Science. https://data.berkeley.edu/education/courses/data-8. Accessed May 22, 2018.
26. Surowiecki J. *The Wisdom of Crowds: Why the Many are Smarter than the Few*. Boston, MA: Little Brown; 2004.