The Journal of Open Source Education

Course Materials for Data Science in Practice

Thomas Donoghue[1], **Bradley Voytek**[1, 2, 3], **and Shannon Ellis**[1, 2]

1 Department of Cognitive Science, UC San Diego **2** Halıcıoğlu Data Science Institute, UC San Diego **3** Neurosciences Graduate Program, UC San Diego

DOI: 10.21105/jose.00121

Software

- Review
- Repository
- Archive

Submitted: 01 July 2020
Published: 23 May 2022

License
Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License (CC BY 4.0).

Summary

Data Science in Practice is a collection of openly available materials, including tutorials and assignments, for learning how to integrate the many skills of data science. The course materials focus on the day-to-day practicalities of hands-on data science, with a particular emphasis on gaining a working familiarity with real-world applications and gaining a 'data intuition.' This collection of materials was originally developed for a course at UC San Diego designed to teach creative and practical data science at scale (Donoghue et al., 2021). The materials for this course have been updated and made publicly available, and are hosted at: https://datascienceinpractice.github.io/.

Topics covered in the data science in practice tutorials include:

- An introduction to relevant tooling, including Python, the scientific Python ecosystem, version control, Github, and Jupyter notebooks.
- Practicalities of working with data, including finding data, evaluating data sources, gathering data, data wrangling and data cleaning.
- Concepts and concerns for using and interpreting data, including data ethics, privacy, and security.
- Introductory statistical concepts, including distributions, statistical tests, and testing statistical properties and assumptions in real datasets.
- Simple data analyses, such as linear models, clustering, dimensionality reduction and classification.
- How to report on data and analysis, through data visualization and narrative text, with a focus on clear explanations.

These topics are further explored in available assignments, which cover:

- Wrangling, cleaning, and combining multiple messy and heterogeneous datasets.
- Collecting web data, applying data protection policies, anonymizing data, and adversarial attacks for deanonymizing data.
- Data analyses, including statistical analyses, applying linear models, and creating visualizations.

The materials are developed in the Python (>= 3.6) programming language, using a standard collection of packages in the scientific Python environment, which can be installed using the Anaconda distribution. All materials are built as Jupyter notebooks, with the assignments being built with the nbgrader extension (Jupyter Project et al., 2019). All the materials are hosted online, using the Jupyter Book tool (Executable Books Community, 2020), from which all the source notebooks can be downloaded to be run locally.

Statement of Need

The field of data science has been rapidly expanding, creating a need for accessible and scalable materials. There is high interest for instruction in data science, and a need in both academia and industry for trained and skilled practitioners. Developing such skills requires hands-on experience and expertise. To address this need, the materials here are focused on practical code-based tutorials, and guided assignments that allow users to practice applying the topics and ideas under study.

There are many available resources for topics within and related to data science, including dedicated tutorials for data science tools and software packages. What can still be difficult, for the novice, is learning how to find and navigate through these materials. A key goal of this course and these materials is to offer a curated introduction to the many topics and available tools, and some initial guided work to make sure users can start to engage with the many aspects of data science. Throughout the course materials, there are many links to other resources. The goal is that these materials be a starting place for the potential user, and a launching off point to the many other more specific resources and tutorials available.

Data science is an interdisciplinary field, requiring expertise from across a range of relevant fields - including technical aspects such as software, computation, statistics, mathematics and machine learning, as well as topics such as research design, contextual understanding of data, ethics, and an understanding of the potential impacts. These materials aim to encompass these multiple elements of data science, focusing not only on the technical aspects of doing data science, but also acknowledging and emphasizing the social impacts and responsibilities of practicing data scientists. These materials are part of an emerging field of integrated data science, as compared to more traditional courses and materials that focus on, for example, detailed machine learning or computation.

Instructional Design

This set of materials was originally created as core materials for a university course, Data Science in Practice, taught at UC San Diego (Donoghue et al., 2021). This course was first taught in the Spring of 2017 and has about 400 students per iteration. The scale of this course originally prompted the development of standalone materials and assignments, that we are now making more generally available.

The full course is supplemented by lectures and lab sections, and is designed as a project-based course. Students work through the materials and assignments presented here, with the goal of building towards doing realistic data science projects. In these projects, students must find openly available datasets, develop a proposal, and then execute analyses to come to an answer. Students must then contextualize the results as a computational notebook that lists their questions and hypotheses, background, ethical considerations, data sources and reliability, results, and conclusion, intermixed with the code and visualizations used to perform the analyses.

In order to encourage users of the public website to also continue to pursue independent data science projects, the website also includes a description of the project outline from the course. This includes guidelines for how to complete data-driven projects using openly available datasets and tools, including listings of available data repositories.

Conclusion

Altogether, these materials offer a general, hands-on introduction to the practice of data science, and its many facets. These materials serve as a complement to many other resources dedicated more specifically to technical skills, and aim to introduce and contextualize the interdisciplinary nature and practical components of working with real world data.

Acknowledgments

We would like to thank the course staff, including the instructional assistants, teaching assistants, and instructors who have contributed time, work, and ideas to this course and these materials, as well as all the students who have taken this course and provided valuable feedback.

References

Donoghue, T., Voytek, B., & Ellis, S. E. (2021). Teaching Creative and Practical Data Science at Scale. *Journal of Statistics and Data Science Education*, *29*(sup1), S27–S39. https://doi.org/10.1080/10691898.2020.1860725

Executable Books Community. (2020). *Jupyter book* (Version v0.10) [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.4539666

Jupyter Project, Blank, D., Bourgin, D., Brown, A., Bussonnier, M., Frederic, J., Granger, B., Griffiths, T., Hamrick, J., Kelley, K., Pacer, M., Page, L., PÃ©rez, F., Ragan-Kelley, B., Suchow, J., & Willing, C. (2019). Nbgrader: A Tool for Creating and Grading Assignments in the Jupyter Notebook. *Journal of Open Source Education*, *2*(11). https://doi.org/10.21105/jose.00032