data-8.github.io

Data 8: The Foundations of Data Science

The UC Berkeley Foundations of Data Science course combines three perspectives: inferential thinking, computational thinking, and real-world relevance. Given data arising from some real-world phenomenon, how does one analyze that data so as to understand that phenomenon? The course teaches critical concepts and skills in computer programming and statistical inference, in conjunction with hands-on analysis of real-world datasets, including economic data, document collections, geographical data, and social networks. It delves into social issues surrounding data analysis such as privacy and design.

The course is offered in partnership with the UC Berkeley Division of Computing, Data Science, and Society.

Offerings

Each offering site includes links to assignments, slides, and readings. You are welcome to use any of the materials you find.

- Summer 2023
- Spring 2023
- Fall 2022
- Summer 2022
- Spring 2022
- Fall 2021
- Summer 2021
- Spring 2021
- Fall 2020
- Summer 2020
- Spring 2020
- Fall 2019
- Summer 2019
- Spring 2019

- Fall 2018
- Summer 2018
- Spring 2018
- Fall 2017
- Summer 2017
- Spring 2017
- Fall 2016
- Spring 2016
- Fall 2015

Materials

All materials for the course, including the textbook and assignments, are available for free online under a Creative Commons license.

Textbook: Computational and Inferential Thinking: The Foundations of Data Science is a free online textbook that includes interactive Jupyter notebooks and public data sets for all examples. The textbook source is maintained as an open source project.

Assignments: All assignments from Fall 2015 to current iterations of the course are available in this repository as Jupyter notebooks. The notebooks assume a Python 3 installation with the standard modules from an Anaconda installation such as Numpy and Matplotlib, as well as the datascience and, depending on the year, okpy or otter-grader modules.

Lecture Materials: All lecture videos, slides and demonstration notebooks from Fall 2016 to current iterations of the course are available via links on the respective course calendars. To request access to the source of the slides for instructional purposes, please fill out our Data 8 Instructor Interest form.

Infrastructure

All of the software components of the course are maintained as open-source projects. We encourage you to contact us if you want any help using them. We also have prepared a guide on how to set up course infrastructure.

The `datascience` module: The course uses a module for table manipulation, charts, and maps that provides an interface appropriate for an introductory course. The `Table` class is similar to a `DataFrame` in Pandas, but explicitly does not support row indexes, hierarchical indexes, time series data, missing values, slicing, and many other advanced features that can complicate table manipulation for novices. The charting features use Matplotlib, but customize the output to match the pedagogical goals of the course. The mapping features are implemented by Folium, but aim to simplify working with tables and geojson files. While the `datascience` module can certainly be used outside the context of the course, it was specifically designed to support the Data 8 curriculum, while setting up students to transition to more standard tools such as Pandas.

The otter-grader automatic grading software: All notebooks are created using the otter-grader notebook creation format. This software generates two notebooks from a parent notebook. The first contains only "public" tests that are used to help students evaluate whether or not solutions to questions are correct – a type of client-side validation for the student. The second notebook contains solutions as well as "private" tests that students are not able to see. These tests are usually used to evaluate correctness and edge cases as well as assign points. This system is used in conjuction with GradeScope at Berkeley to grade and assign points to student work but an instructor is also able grade notebooks on their own machines, see the documentation at otter-grader, as well as use a free service that we deployed called otter-service-standalone.

Hosted Computing Environment: We provide a hosted environment for our students to edit and execute their Notebooks. It includes two components, a Kubernetes-based deployment of JupyterHub that we have specifically designed for courses, and an assignment server that loads assignments into the students' environment.

If you want more information about any of these tools, please fill out our Data 8 Instructor Interest form or email `ds-help@berkeley.edu` .