Berkeley
Division of Computing,
Data Science, and Society

This page is intended as an overall guide to the open-source resources for teaching data science. The development of these resources has been the product of inspiration, innovation, and collaboration by faculty, staff, and students in curriculum design, course development, and untold hours of programming. Many of these resources were developed specifically at Berkeley but others as part of broader open source collaborations (e.g., Project Jupyter). In almost all cases the content and source code are open to use and modify, and throughout there has always been a vision of building tools for expanding data science education to other campuses.

Data 8: Foundations of Data Science

Data 8 (data8.org) is the flagship introductory Data Science class at UC Berkeley and is currently fully enrolled at more than 1,300 students every semester. The vision for the course is to combine teaching computational thinking and statistics, to modernize statistical instruction to center on statistical inference, and to be accessible to a broad range of students, including by eliminating requirements such as calculus, linear algebra, and introductory programming. Both programming and statistics are taught through Jupyter notebooks running Python.

Data 8x - an online MOOC implementation of Data 8 offered through EDx

Outside of Berkeley, this class is available as Data 8x, a popular course online at EDx. Several institutions have deployed Data 8 in a flipped classroom model, using the videos from EDx in combination with in-person lab help.
edx.org/professional-certificate/berkeleyx-foundations-of-data-science

Textbook

The Data 8 textbook is at inferentialthinking.com. The textbook was written by John DeNero and Ani Adhikari and is licensed under a Creative Commons License. One of the key innovations of the textbook is that it has a set of Jupyter notebooks that provide programming illustrations of key concepts, which can be accessed at a local JupyterHub or more generally at mybinder.org. You can access the Jupyter Book repository at github.com/data-8/textbook; the README includes instructions for how to host/adapt the textbook at other sites and how to change the interact links. There is a guide to the Jupyter-based open source textbook publishing at jupyter.org/jupyter-book/. Content and pages in the textbook are written using Jupyter notebooks, generated with Jekyll, and hosted at Github.

Course Webpage

The Data 8 homepage is at data8.org, where you can find the course sites for each iteration of Data 8 since fall 2015. Within the sites for past semester offerings, you can find slide decks and videos for lectures and Jupyter notebooks for in-class demos, as well as labs, homeworks, and projects. Almost all of the resources (e.g., textbook, webpage, teaching materials) used for Data 8 are available the github organization, github.com/data-8.

Guide for Adapting Data 8 to Other Institutions

The *Zero To Data 8 Guide* is a guide to setting up and running Data 8 at other institutions. This guide, at http://data8.org/zero-to-data-8/intro, documents step by step how to manage aspects such as the technical infrastructure, cloud resources, and technical support.

Tables Datascience Package for Python

Data 8 uses a course-specific python package *datascience*, designed for teaching tabular data manipulation and visualization in introductory data science courses. Documentation for the package can be found at data8.org/datascience/. Teaching with this package throughout the Data8 materials allows for a pedagogically clean dataframe concept, without the added complexity of Pandas or R.

Jupyter and JupyterHub

Project Jupyter (jupyter.org) exists to develop open-source software, open-standards, and services for interactive computing across dozens of programming languages. JupyterHub is a tool that allows Berkeley's data science program to utilize cloud computing infrastructure to deploy scalable resources that enable users to interact remotely with a standardized, common computing environment. JupyterHub creates on-demand cloud-based Jupyter notebook servers and are the basis of the technical infrastructure for Data 8 and related classes. The Data 8 JupyterHub deployment is at datahub.berkeley.edu. In addition, the course textbook uses mybinder.org for public interact links. Binder uses a JupyterHub to create temporary user sessions that are open to the public, in this case serving the Data 8 textbook and its environment. *Zero To JupyterHub* is a guide on deployment and maintenance of a JupyterHub, at zero-to-jupyterhub.readthedocs.io. For a smaller class of up to 50 people, *The Littlest JupyterHub* is a simpler version to run Jupyterhub on a single server called the, at the-littlest-jupyterhub.readthedocs.io.

Grading

Data 8, uses an autograding server called Ok.py for grading and other class management purposes. These tests can be seen throughout the Data 8 notebooks. A new project to develop a local autograding solution called Otter Grader, at https://github.com/ucbds-infra/otter-grader. It is designed to be a scalable grader that utilizes temporal docker containers either locally or in a VM. Users can generate Ok tests using a simple test generator at https://oktests.herokuapp.com/

Data Science Modules

The infrastructure developed for Data 8 has also been used to develop teaching modules, often a set of 1-3 notebooks to deploy into an existing class, often outside the realm of computer science. These can range from GIS mapping to neuroscience to text analysis in a humanities course. A showcase of some of the interesting deployments is available at ds-modules.github.io/modules-textbook. The full set of Jupyter notebooks developed for all classes are available at github.com/ds-modules.

Connector Courses

Connector courses are a set of classes complementary to Data 8 that have been developed to expose students to Data Science applications within a subject area. Connectors allow students to apply theoretical concepts from Data 8 to a particular area of interest. A full list of connectors can be found at data.berkeley.edu/education/connectors.

Human Context and Ethics

The Data Science Major and Minor require that all students take a class in Human Contexts and Ethics. HCE education explores how human, social, and institutional structures and practices shape technical work around computing and data, as well as how data, data analytics, machine learning, artificial intelligence, and computing permeate and shape our individual and social lives. Materials are available at https://data.berkeley.edu/degrees/human-contexts-and-ethics. The Human Contexts and Ethics (HCE) Toolkit helps students identify where human power structures and value choices get built into technical work, empowering them to figure out how, when, where they can responsibly and effectively intervene. https://data.berkeley.edu/human-contexts-and-ethics-toolkit

Resources for Instructors - Curriculum Guide

The Data Science Education Program at Berkeley has created a Curriculum Guide at https://ds-modules.github.io/curriculum-guide/intro to help guide instructors with set up, workflow, and pedagogy in teaching data science courses connected to Data 8 and using the same infrastructure. Much of the content in this guide is useful for instructors teaching with Jupyter notebooks and Jupyterhub deployments. Faculty across many departments have been trained to use the data science pedagogy platform in a short summer Data 8 bootcamp intended to get faculty ready to adapt data science teaching tools to their own subject area: sites.google.com/berkeley.edu/ucb-dse-workshop/home.

Contact us at: data.berkeley.edu / dspi@berkeley.edu / @BerkeleyDataSci   v3   March 2020