DS Connector Assistant Onboarding Guide

Fall 2016

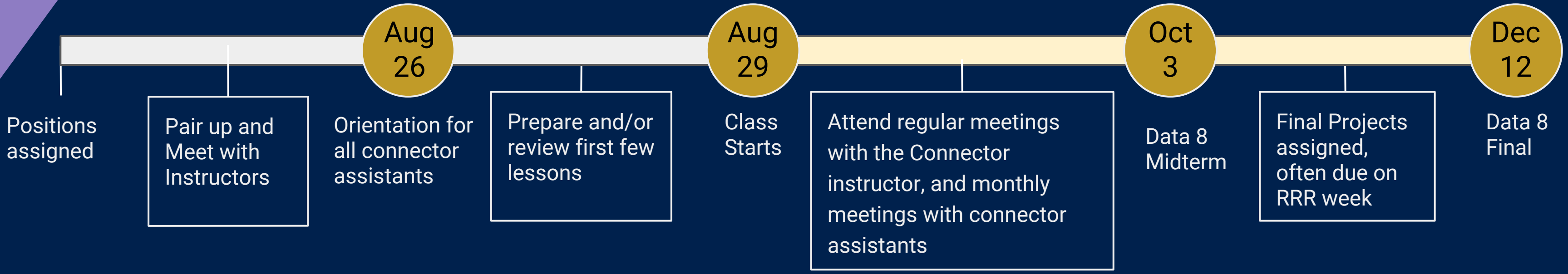
Welcome to the Data Science Education Program!
This short informational handout is designed to provide incoming Data Science Connector assistants with a brief orientation and reference support for specific issues. This is not designed to be prescriptive, since pedagogy varies widely by subject, so it will also highlight aspects of the student experience that are important to understand. As a Connector Assistant you serve as an advocate for the students and will work with them and the instructor to create a supportive learning environment. Since the program is still relatively new and most content is original, don't worry if you encounter frequent issues as some growing pains and issues should be expected. The support team is always available to help.

Table of Contents

(by page #)

Developed by the DSEP Student team

Connector Assistant Program Overview



Positions assigned

Pair up and Meet with Instructors

Aug 26

Orientation for all connector assistants

Prepare and/or review first few lessons

Aug 29

Class Starts

Attend regular meetings with the Connector instructor, and monthly meetings with connector assistants

Oct 3

Data 8 Midterm

Final Projects assigned, often due on RRR week

Dec 12

Data 8 Final

Requirements for Connector Assistants

General Requirements

- Assisting students with difficulties in labs
- Attend monthly Connector Assistant meetings to discuss difficulties and successes
- Review course content before it is covered in class
- Regular meetings with the Instructor regularly to share insights and stay up-to-date on class assignments and projects

Advanced, connector-specific help

- Hosting office hours to help students with material when needed
- Generating LaTex/PDF solution manuals
- Building or contributing to modules
- Finding areas in the class that could be improved from the student’s perspective
- Leading mini in-class review sessions that cover key concepts covered in class or how to write functions for a regression line etc.
- Whatever you and the Instructor deem necessary

Class Structure

Lecture

Instructors usually start with a lecture beginning with a motivating example that describes a key issue to the connector's domain and how data science can be used to help resolve it. To emphasize new techniques used and the data analysis process, Connector Assistants have often sought out material that may be especially engaging to them.

Lab

Most instructors build labs in iPython notebooks, but some prefer other methods that are better-suited for their course (e.g. TensorFlow). Labs teach material through a hands-on exploration that can be any part of the data analysis process from scraping data to cleaning datasets, creating visualizations, or even defending conclusions related to the domain. Collaboration is encouraged because there are many diverse abilities, backgrounds and interests in the class.

During lab time, Connector Assistants can help by **leading demos** on the board if enough students are struggling with a common issue, or by **individually helping** students to understand and resolve their error messages. CAs can also answer questions about the material. CAs explain the material to make sure they understand their work on both a technical (code-based) and ideological (content related to the domain) level instead of providing the answer.

Developing and Preparing Content

1. Cleaning Data Sets

Cleaning data sets can be done by you and the instructor so they can be read with a .read_table() function allowing students to spend more time on the analysis. Some useful tools for this are the pandas .read_csv documentation that underlies the .read_table() function. It is also useful in some domains to teach students to read in complex tables with APIs, so just be sure to include proper instructions and documentation for those to assist students.

*Resources to find datasets are covered in additional information.

3. Autograding vs Grading for Content

NBGrader is the most commonly used software for autograding, and tutorials can be found at the link. Other Instructors prefer OK tests that run similarly. Both can be built into labs and homeworks. Other connectors prefer to grade based on output by having students turn in a visualization or write-up and the code used to lead to it.

2. Tools used in the analysis

iPython: Developed by Berkeley's Fernando Perez, and is the work environment for data 8. Most help for iPython can be found under the help tab in the jupyter environment, but there are additional resources that cover interactable widgets, viewing objects in different forms and even grading help at Fernando's tutorial

New methods: If the course introduces students to new analytic tools (MatPlotLib, Pandas, etc.), you should help provide the students with documentation for them. Some modules can also cover concepts before they are taught in Data 8, which can be very helpful to students if they are taught clearly.

Assignments

Homework

Connector homework give students a chance to review course material covered, while emphasizing important concepts. Connector homework varies from iPython notebooks, data visualizations to responses to questions. Homework can be submitted via **bCourses,** email, etc.

LaTex is used to create solutions manuals for students to look over after they get homework back. LaTex can be difficult to learn, so it's best to get a head start if you and the instructor intend to co-develop them.

Quiz/Test

Connectors often use quizzes to assess how well students are doing and track progress. They are designed to be brief checkins to see if students are understanding the material and are usually 3-4 questions long. Final exams covers materials from throughout the semester and can take the whole class period. These can be highly stressful for students out if implemented during finals week.

Final Project

The final project serves as a way for students to apply their domain knowledge, analytics and creativity to a project that gives them something to show after the class. Roughly 3-4 weeks and can be in partners, groups or individually and serves as a way to demonstrate understanding of several aspects of data analysis in a domain. Some students express great interest in working in groups, but there have been difficulties matching students who will work best together. Students have an opportunity to present their projects in BIDS during RRR week.

Resources and Tips

Piazza

Some courses find it best to set up a piazza page for students to talk on. However, piazza pages also need to be monitored frequently by the instructor and assistant to answer questions.   Here is a quick start guide

Work Environment

Python: Though covering the basics of the python language, many students who are new to coding still have trouble with syntax or other concepts throughout the semester that we are creating guides to address, so some growing pains should be expected.
This has been the most prominent issue encountered by students in the connectors, and simply requires a reference guide or quick explanation to clear up the difficulty. You don't need to understand every error message shown, but essential errors (null pointer exceptions, neverending iterations etc.) should be explained.

Jupyter File Management

Upload: Every individual file must be uploaded into the same folder as the ipython notebook or be accessible in some other way (deep linking), but students do not always realize this. Also the file must be a table alone and not have other scripts saved in it.
Deep-linking files: Files must be linked by consecutive folders with slashes (/)

Additional online resources

Language Guides

- **Matplotlib:** matplotlib.org/api/pyplot_api.html#matplotlib.pyplot.plot
- **Pandas:** pandas.pydata.org/pandas-docs/stable
- **SciPy:** http://www.scipy-lectures.org/
- **Regex:** regexr.com
- **Scipy:** docs.scipy.org/doc
- **Tensorflow:** tensorflow.org/versions/r0.9/api_docs/index.html

Databases

- Berkeley Library Dataset purchasing program
- Google's dataset search engine
- Amazon public datasets (requires free account)
- https://github.com/ali-ce/datasets
- https://github.com/caesar0301/awesome-public-datasets

Data 8 Syllabus Fall 2016

WEEK 1: Intro, Cause and Effect

WEEK 2: Expressions, Sequences, Data types
Begin using tables

WEEK 3: Mon - HOLIDAY
Tables 2, Arrays and Histograms

WEEK 4: Defining Functions and Histograms
Applying functions to tables

WEEK 5: Functions and tables 3
-Project 1 assigned
Sampling, Iteration

WEEK 6: Empirical distribution of sample and statistic
Randomness

WEEK 7: Total Variation distance -
Project 1 due
Hypothesis Testing (P-value and Error probabilities)

WEEK 8: Guest Lecture/Exploration -Midterm
Midterm Review

WEEK 9: Bootstrap Confidence Interval, and CI applications

WEEK 10: Avg, SD, Normal Curve, Central limit theorem -Project 2 assigned
Adjustments between Normals

WEEK 11: Scatter plots and correlation -Proj 2 checkpoint
R, regression, Least squares line

WEEK 12: Regression Inference -Proj 2 due

WEEK 13: Classification and comparing numerical samples -Proj 3 assigned

WEEK 14: Pivot; Comparing two categorical samples

WEEK 15: Causality, Bayes' Rule, Decisions -Proj 3 due