

Jupyter in UC Berkeley's Data Science Education Program

Elaine Angelino and Sam Lau

JupyterDays Boston 2016
March 17-18, Cambridge, MA

A guided tour

- Publicly available information distributed across many web pages and GitHub repositories running a live suite of data science courses
- Designed for and by many UC Berkeley students, instructors, other teaching / support staff

Who is this presentation for?

- Diverse folks thinking about data science education, using Jupyter notebooks in the classroom, and/or deploying and scaling JupyterHub
- Designed for people who don't have accounts on data8.berkeley.edu

What is in this presentation?

- An overview of the UC Berkeley's new data science education program
- Pointers to current course materials distributed as Jupyter notebooks
- An overview of the live JupyterHub-based infrastructure
- Hopefully, lots of materials to explore, fork, and hack on

Some dependencies

- All course content and software is viewable online
- You'll need a [Git](#) if you want to clone or fork this content
- Course content is distributed as Jupyter notebooks that have several Python dependencies
 - [Python 3](#)
 - [Jupyter](#)
 - [datascience](#)

DATA 8: Foundations of Data Science

DATA 8: First day of class, Spring 2016



Course overview

- Teaches computational and inferential (statistical) thinking through interaction with real data
- Pilot run in Fall 2015 with about 80 students
- Current Spring 2016 enrollment at about 470 students
- Three 50 min lectures & 2 hour computer lab every week

Broader context

databears.berkeley.edu

The screenshot shows a web browser window with the address bar displaying `databears.berkeley.edu`. The page has a dark blue header with the Berkeley logo in yellow and the text "Data Science Education Program" in white. Below the header is a navigation bar with links: HOME, ABOUT, UNDERGRAD COURSES: SPRING 2016, FALL 2015, @BERKELEY, FAQ, and CONNECT. The main content area features two news items. The first item is titled "UC Berkeley Majors accept Foundations of Data Science for Stat Requirement" and is dated "Submitted by David CULLER on Wed, 12/09/2015 - 08:01". The text of this item states that almost all major programs with a statistics requirement at Berkeley have approved the use of data science courses as a means of satisfying their existing requirement, as proposed by the Chair of Statistics, Michael Jordan. The second item is titled "c8 Course Reaches Capacity - time to enroll in connectors" and is dated "Submitted by David CULLER on Sat, 01/16/2016 - 09:40". The text of this item states that as of Saturday morning January 16, 479 students were signed up for the 481 total seats available in the first regular offering of Foundations of Data Science.

HOME ABOUT UNDERGRAD COURSES: SPRING 2016 FALL 2015 @BERKELEY FAQ CONNECT

UC Berkeley Majors accept Foundations of Data Science for Stat Requirement

Submitted by David CULLER on Wed, 12/09/2015 - 08:01

Almost all major programs with a statistics requirement at Berkeley have approved the use of data science courses as a means of satisfying their **existing** requirement, as [proposed by the Chair of Statistics, Michael Jordan](#). A few majors with an existing statistics requirement are working through their particular approval process.

[Read more](#)

c8 Course Reaches Capacity - time to enroll in connectors

Submitted by David CULLER on Sat, 01/16/2016 - 09:40

As of Saturday morning January 16, 479 students were signed up for the 481 total seats available in the first regular offering of Foundations of Data Science. We are excited to see this broad, diverse cohort - and at the same time be able to accommodate the high level of student interest in this very new offering. Universities throughout the world are

Broader context

databears.berkeley.edu

- This is all new, fast-moving, growing, & the intention is to keep growing (up to 3000 DATA 8 students / semester)
- Complemented by a suite of connector courses teaching diverse subjects through the lens of data science
- DATA 8 course is meant to be a foundation for advanced courses to be seeded across the university
- See the report on [Data Sciences @ Berkeley: The Undergraduate Experience](#)

Course design requirements

- Must be accessible to all incoming first-year students
- Assume no computer science background and only high school algebra
- Students interact immediately with data programmatically
 - Can't assume all students have personal computers
 - Can't require students to figure out a local installation
- Provide a platform (technical & intellectual) that students can build on throughout their college careers

Implementation highlights

- Jupyter notebooks + JupyterHub support a solution satisfying all design requirements
- Why Jupyter notebooks?
 - Provide a natural environment for introducing data science skills to students
 - Let students develop an explicit computational narrative with data
 - Interactive substrate for distributing course content

Implementation highlights

- Jupyter notebooks + JupyterHub support a solution satisfying all design requirements
- Why JupyterHub?
 - Multi-user server for Jupyter notebooks can support many users (students, instructors, teaching staff)
 - Enables browser-based interface to computation in the cloud
 - Students only need a browser to start programming, interacting with data, creating a visible record of their analytical steps

Course website: data8.org

The screenshot shows a web browser window with the URL <https://data-8.appspot.com/sp16/course>. The page header includes the Cal logo and the text "Foundations of Data Science". A navigation bar contains links for Announcements, Course, Registration, Info, Weekly, Staff, Connectors, Labs, Docs, and Mentors, along with a search bar. The main content area features the title "Foundations of Data Science — Spring 2016" in red, followed by the instructor and co-instructors' names, and a large "Register" button. Below this, a "Syllabus" section is visible, listing topics like "Data Science", "1 Why Data Science?", "2 Cause and Effect", "Lab 01", and "Homework 01".

Foundations of Data Science

Announcements Course Registration Info Weekly Staff Connectors Labs Docs Mentors Search

Foundations of Data Science — Spring 2016

Instructor: John DeNero
Co-instructors: Ani Adhikari, Michael I. Jordan, Tapan Parikh, and David Wagner
MWF 10-11 in 155 Dwinelle Hall

Register

Syllabus

Data Science	An overview of data science	
Wed Jan 20	1 Why Data Science?	Lab 01
Fri Jan 22	2 Cause and Effect	Homework 01

Course website: data8.org

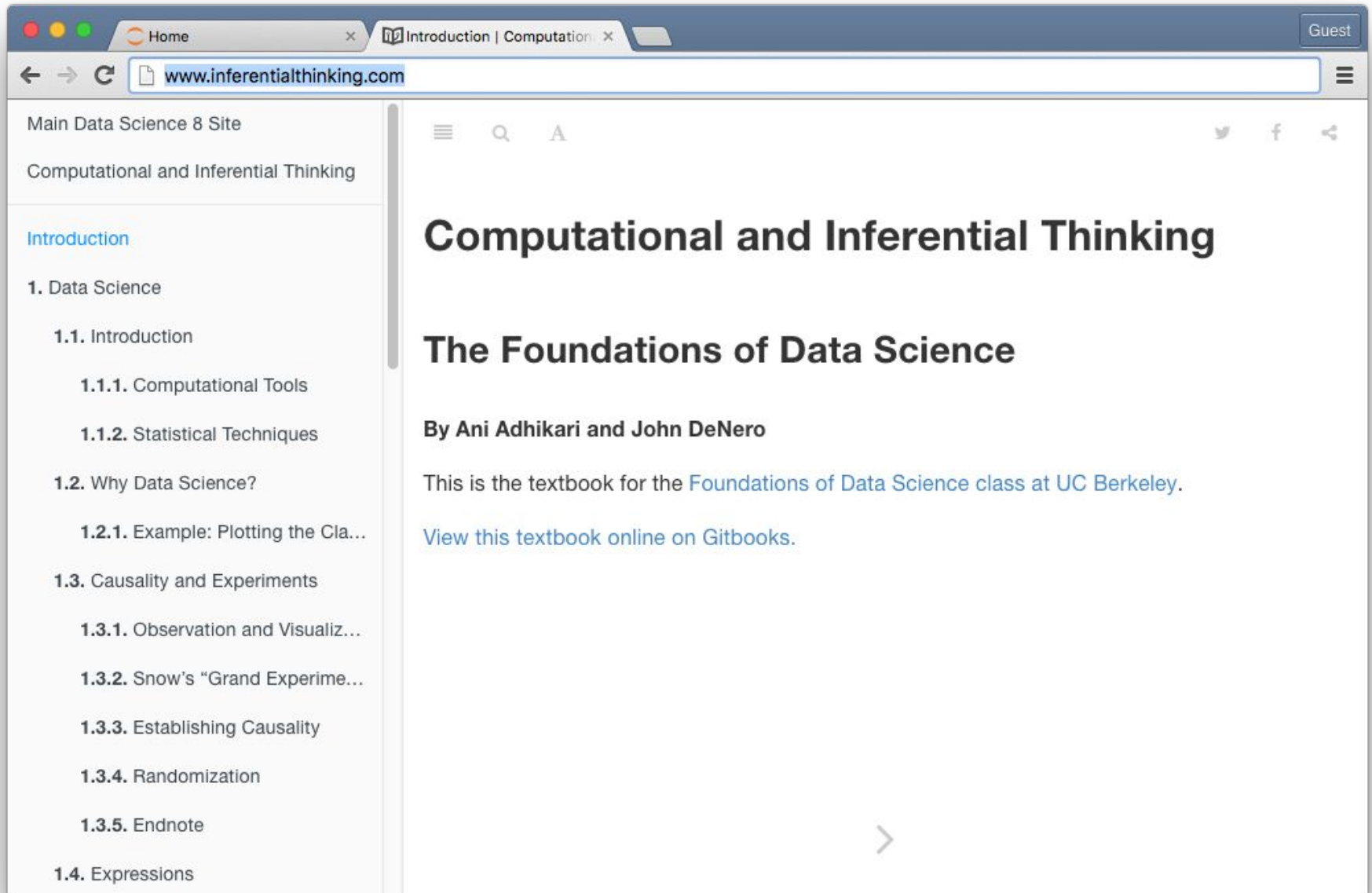
Syllabus and links to lecture videos

- An overview of data science
- Using Python to manipulate info in table data structures
- Interpreting and exploring data through visualizations
- Sampling: Understanding the behavior of random selection
- Making predictions from data
- Inference: Reasoning about populations by computing over samples
- Models: Making assumptions & exploring their consequences

Course website: data8.org

- data8.org is primarily a student-facing website and its links to content such as computer lab assignments will **not** work for anyone who doesn't have a course account
- We'll show you:
 - What students see and interact with
 - How links to interactive materials work for students
 - How to find the source materials hosted across many GitHub repositories at <https://github.com/data-8>

Online textbook: www.inferentialthinking.com



The screenshot shows a web browser window with the address bar displaying www.inferentialthinking.com. The browser has two tabs: 'Home' and 'Introduction | Computational...'. The website layout includes a left sidebar with a table of contents and a main content area. The sidebar lists 'Main Data Science 8 Site', 'Computational and Inferential Thinking', and 'Introduction'. Under 'Introduction', it lists sections 1. Data Science, 1.1. Introduction, 1.1.1. Computational Tools, 1.1.2. Statistical Techniques, 1.2. Why Data Science?, 1.2.1. Example: Plotting the Cla..., 1.3. Causality and Experiments, 1.3.1. Observation and Visualiz..., 1.3.2. Snow's "Grand Experi...", 1.3.3. Establishing Causality, 1.3.4. Randomization, 1.3.5. Endnote, and 1.4. Expressions. The main content area features the title 'Computational and Inferential Thinking' and 'The Foundations of Data Science' by Ani Adhikari and John DeNero. It also includes a paragraph about the textbook being for the 'Foundations of Data Science class at UC Berkeley' and a link to 'View this textbook online on Gitbooks.'.

Home x Introduction | Computational x Guest

← → ↻ www.inferentialthinking.com

Main Data Science 8 Site

Computational and Inferential Thinking

Introduction

1. Data Science

1.1. Introduction

1.1.1. Computational Tools

1.1.2. Statistical Techniques

1.2. Why Data Science?

1.2.1. Example: Plotting the Cla...

1.3. Causality and Experiments

1.3.1. Observation and Visualiz...

1.3.2. Snow's "Grand Experi..."

1.3.3. Establishing Causality

1.3.4. Randomization

1.3.5. Endnote

1.4. Expressions

☰ 🔍 A

🐦 f ↗

Computational and Inferential Thinking

The Foundations of Data Science

By Ani Adhikari and John DeNero

This is the textbook for the [Foundations of Data Science class at UC Berkeley](#).

[View this textbook online on Gitbooks.](#)

➤

Most sections of the online textbook begin with a big blue Interact button ([example section](#))

The screenshot shows a web browser window with the URL `www.inferentialthinking.com/chapter3/sampling.html`. The page title is 'Sampling | Computational'. The left sidebar contains a table of contents with chapters 2 through 4. Chapter 3, 'Sampling', is highlighted. The main content area has a large blue 'Interact' button. Below it, a paragraph states: 'In this section, we will continue to use the `top_movies.csv` data set.' A code block shows the following R code:

```
top = Table.read_table('top_movies.csv')
top.set_format([2, 3], NumberFormatter)
```

 At the bottom, a table displays movie data with columns: Title, Studio, Gross, Gross (Adjusted), and Year. The table lists three movies: Star Wars: The Force Awakens, Avatar, and Titanic.

2. Chapter 2

- 2.1. Bar Charts
- 2.2. Histograms

3. Chapter 3

- 3.1. Sampling**
- 3.2. Iteration
- 3.3. Estimation
- 3.4. Center
- 3.5. Spread
- 3.6. The Normal Distribution
- 3.7. Exploration: Privacy

4. Chapter 4

- 4.1. Correlation
- 4.2. Regression
- 4.3. Prediction
- 4.4. Higher-Order Functions

Sampling


Interact


In this section, we will continue to use the `top_movies.csv` data set.

```
top = Table.read_table('top_movies.csv')
top.set_format([2, 3], NumberFormatter)
```

Title	Studio	Gross	Gross (Adjusted)	Year
Star Wars: The Force Awakens	Buena Vista (Disney)	906,723,418	906,723,400	2015
Avatar	Fox	760,507,625	846,120,800	2009
Titanic	Paramount	658,672,302	1,178,627,900	1997

When a student clicks the Interact button, they're redirected to an interactive Jupyter notebook!



jupyter Sampling  [Control Panel](#) [Logout](#)

File Edit View Insert Cell Kernel Help Python 3

Code Cell Toolbar: None

```
In [2]: # HIDDEN

from datascience import *
import numpy as np
import matplotlib.pyplot as plots
plots.style.use('fivethirtyeight')
%matplotlib inline
```

In this section, we will continue to use the [top_movies.csv](#) data set.

```
In [3]: top = Table.read_table('top_movies.csv')
top.set_format([2, 3], NumberFormatter)
```

Out[3]:

Title	Studio	Gross	Gross (Adjusted)	Year
Star Wars: The Force Awakens	Buena Vista (Disney)	906,723,418	906,723,400	2015
Avatar	Fox	760,507,625	846,120,800	2009
Titanic	Paramount	658,672,302	1,178,627,900	1997

What's going on?

- First, we'll explain where the source material is
- Second, we'll explain the Interact button

The textbook is hosted in a GitHub repo

```
git clone https://github.com/data-8/textbook.git
```

- Most of the underlying source material for the textbook is written in Jupyter notebooks ([example notebook](#))
- [GitBook](#) allows us to [write and organize chapters using Markdown](#)
- Conveniently, [Markdown](#) allows arbitrary HTML inline
- Convert notebook to HTML snippet using [nbconvert](#)
- Include that HTML in the .md file ([example Markdown](#))

The Interact button

- An Interact button in the textbook ([example section](#)) is a link like this:

```
http://data8.berkeley.edu/hub/interact?  
repo=textbook&  
path=notebooks/top_movies.csv&  
path=notebooks/Sampling.ipynb
```

- Uses [DS8-Interact](#), a side server for the DATA 8 JupyterHub deployment that copies remote notebooks and other files into user accounts

```
git clone https://github.com/data-8/DS8-Interact.git
```

Interact links distribute content to students

- Not just for the online textbook
- This link-based system for loading content also works for whole directories and anything that can be checked in
- Used for distributing labs, homeworks, projects, etc.
- Not just for the DATA 8 course
 - Also being used by the connector courses

Our JupyterHub deployment

Origin story

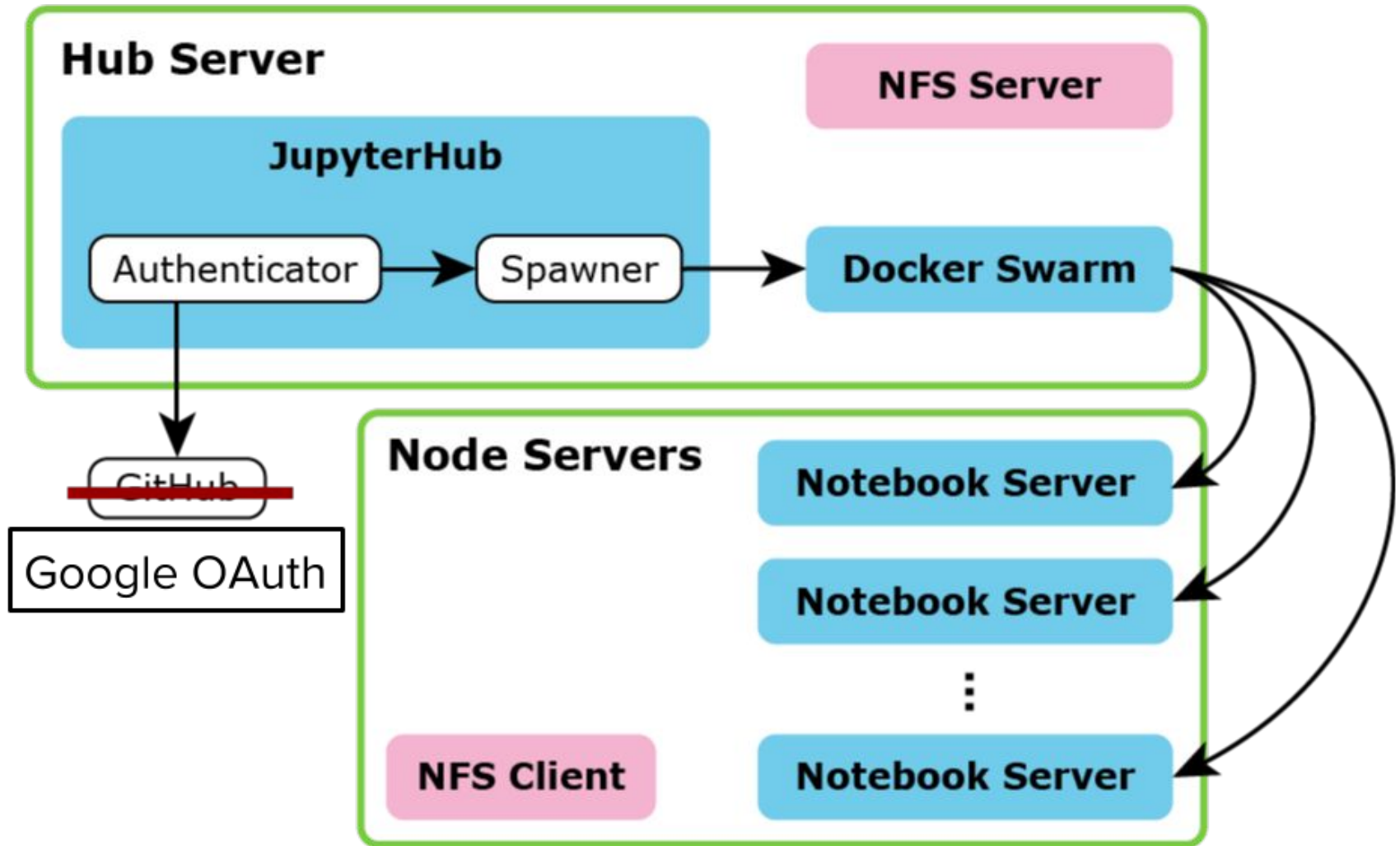
- We were about to level out a new basement to give students computers with Jupyter installed...
- ...until we discovered that [Jessica Hamrick](#) had deployed JupyterHub to the cloud for a UC Berkeley cognitive science class of 220 students
- We thought we could do it too!

Our JupyterHub deployment

```
git clone https://github.com/data-8/jupyterhub-deploy.git
```

- [Our deployment](#) is based on Jessica Hamrick's [jupyterhub-compmodels-deploy](#)
- See Jess's blog post at Rackspace on [Deploying JupyterHub for Education](#) and also the README at [jupyterhub-compmodels-deploy](#) for details

Our JupyterHub deployment



Modified from a diagram by Jessica Hamrick

Technical specs

- Fall 2015 pilot (about 80 students): deployed JupyterHub on bare-metal machines from Berkeley's CS dept
- We gave each student 2GB of RAM
- Expected about 60% of users to be on at any time, so we provisioned 2 machines each with 64 cores + 26GB RAM
- Spring 2016 (about 480 students): we deployed on a donation from [Microsoft Azure](#), using 36 machines each with 8 cores + 14GB RAM

Connector courses

STAT8/CS8/INFO8
Foundations of Data
Science
MWF 10 - 11 AM
155 Dwinelle
4 Units



CEE 88
Data Science and
Smart City
Tuesday 9 - 11 AM
105 Cory
2 Units



CS 88: Computational
Structures in Data
Science
Monday 4 - 5 PM
306 Soda
2 Units



COGSCI 88
Data Science and
The Mind
Monday 12 - 2 PM
105 Cory
2 Units



ESPM 88A
Data Sciences in
Ecology and the
Environment
Tuesday 11 - 1 PM
105 Cory
2 Units



ESPM 88B
Exploring Geospatial
Data
Monday 4 - 6 PM
385 LeConte
2 Units



SPRING 2016 DATA SCIENCE COURSES

www.data8.org

HIST 88
How Does History Count?
Tuesday 2 - 4 PM
105 Cory
2 Units



INFO 88
Data and Ethics
Tu 3:30 - 5:30 PM
210 South Hall
2 Units



L&S 88-1
Health, Human
Behavior, and Data
Monday 1 - 3 PM
2232 Piedmont 100
2 Units



L&S 88-2
Literature and Data
Tuesday 4 - 6 PM
105 Latimer
2 Units



STAT 88
Probability &
Mathematical Statistics
in Data Science
Tuesday 4 - 6 PM
1 LeConte
2 Units



STAT 89A
Introduction to
Matrices and Graphs
in Data Science
Monday 1 - 3 PM
344 Evans
2 Units



Berkeley
UNIVERSITY OF CALIFORNIA

DATA SCIENCE EDUCATION PROGRAM

databears.berkeley.edu

Connector courses

- [Suite of courses](#) in departments across campus introduce diverse subjects through the lens of data science
- Spring 2016 has 11 connector courses: in ethics, cognitive science, geospatial data, statistics (2), ecology, history, computer science, health, smart cities, literature
- Nearly all use Jupyter notebooks and the DATA 8 JupyterHub deployment, Interact links, etc.
- Many connector instructors are new to Python & GitHub!

Technical challenges & possible future directions

- Scaling up to more students
- Scaling out to more courses

Scaling up to more students

- Theoretically, scaling up to more students means we can just add more nodes to the JupyterHub deployment to get the computing power in. However...
- ...we're now discovering bugs that are only discoverable when dealing with scale.
- We were haunted by [a race condition](#) in JupyterHub that resulted in many 503 errors for weeks.
- We've had to make a forum thread for these issues -- students still run into them every day

Scaling up to more students

- Now we have a team of students adding tooling to make deployment more stable
- This includes a development deployment, logging and monitoring, and load testing

Scaling out to more courses

- Goal: eventually have one JupyterHub deployment that can serve all the classes at UC Berkeley that want to use Jupyter notebooks
- Adding JupyterHub for a class should be as easy as creating a class web page
- JupyterHub currently consolidates all users into one system -- we need to split the users into multiple groups

Scaling out to more courses: Some challenges

- Courses have different resources -- some might have AWS credits, others can have Azure credits, etc.
- Instructors need a way to distribute content to students
- Ideally, instructors could also grade assignments easily
- Students should be able to access different hubs for different courses, and build a portfolio of materials over their undergraduate years

Scaling out to more courses: Some proposals

- A JupyterHub Hub, which lists and manages deployments of JupyterHub
- A Dropbox-like interface to GitHub to help instructors with content management
 - See the [design doc](#) for an experiment called [jupyter-synchronized-folders](#)
 - The design doc is structured as (but is not) a [Jupyter Enhancement Proposal](#)
 - We'd love to hear comments via [this pull request](#)

Other resources

- [jupyter-education Google Group](#)
- [JupyterHub Gitter channel](#)

Thanks and acknowledgements!

John DeNero, Ani Adhikari, Michael I. Jordan, Tapan Parikh, David Wagner
[DATA 8 staff](#) + many additional highly motivated amazing undergraduates

[Data Science Connector Course Instructors](#)

[UC Berkeley Data Science Education Program](#)

[Berkeley Institute for Data Science \(BIDS\)](#)

Cathryn Carson

David Culler

Shanaaz Deo

Michele Gleit

Chris Holdgraf

Ryan Lovett

Anthony Suen

Sameera Vemulapalli

Alvin Wan

Steve Yang

Eric Zhao

Jeffrey Anderson-Lee

Matthias Bussonnier

Stacey Dorton

Jessica Hamrick

Frances Hocutt

Kevin Koy

Yuvi Panda

Fernando Perez

Min Ragan-Kelley

Patrick Schmitz

JupyterDays Organizers