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

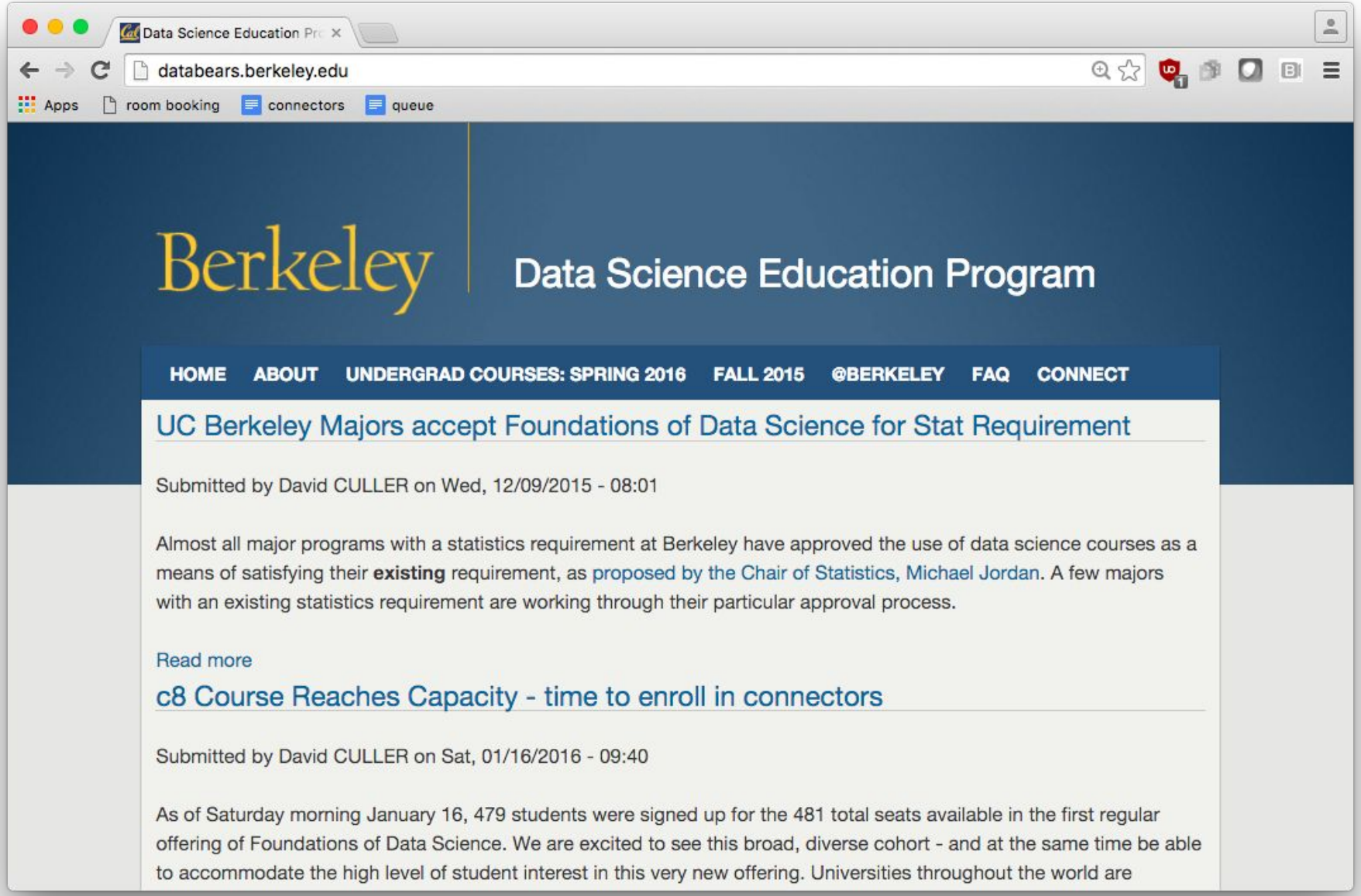
DATA 8: First day of class, Spring 2016

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



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

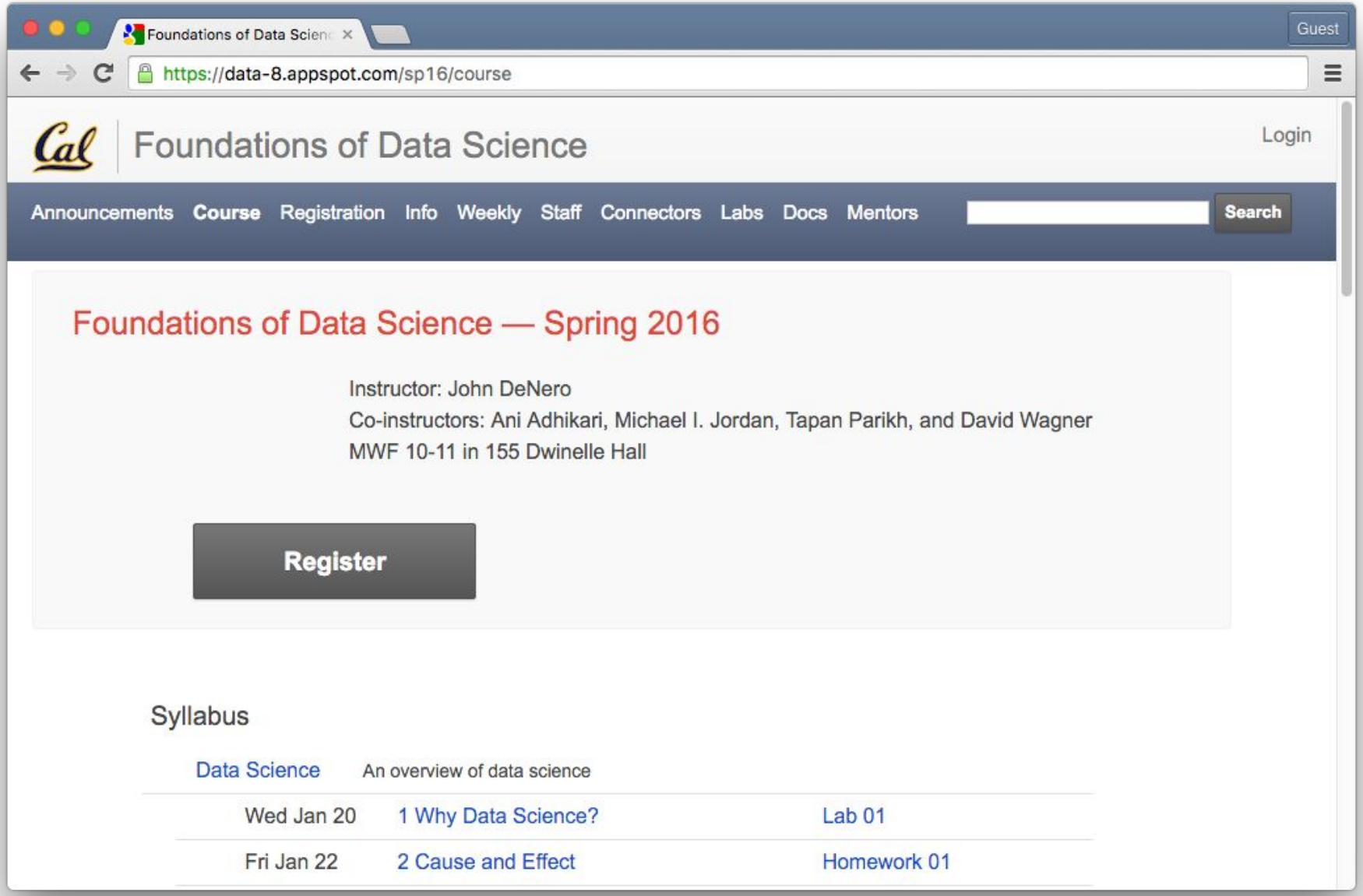
- Why Jupyter notebooks?
  - Provide a natural environment for introducing data science skills to students
  - Let students develop an explicit computational narrative with data
  - Interactive substrate for distributing course content

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



Course website: **data8.org**

Syllabus and links to lecture videos

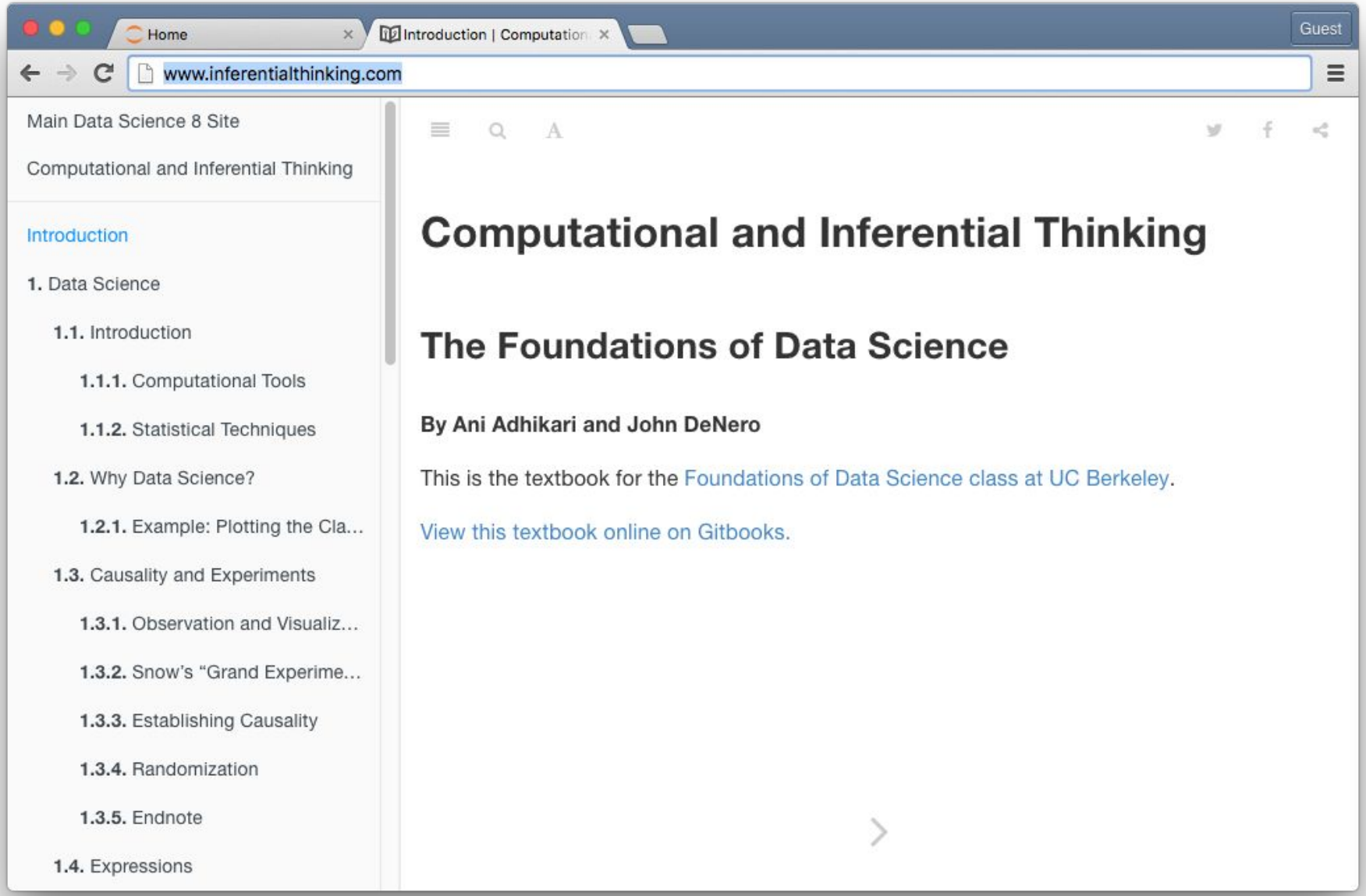
- An overview of data science
- Using Python to manipulate info in table data structures
- Interpreting and exploring data through visualizations
- Sampling: Understanding the behavior of random selection
- Making predictions from data
- Inference: Reasoning about populations by computing over samples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

Course website: data8.org

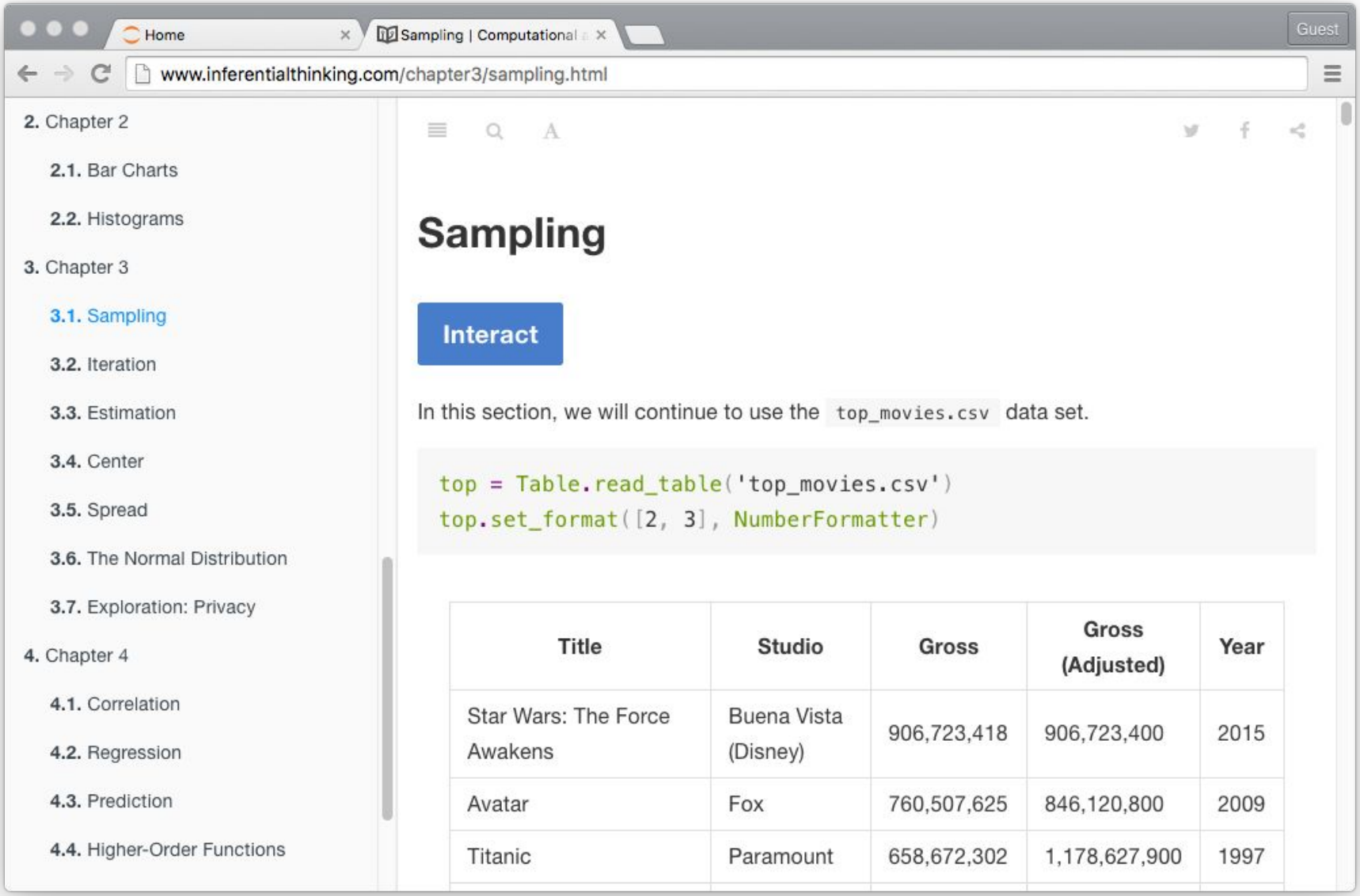
- data8.org is primarily a student-facing website and its links to content such as computer lab assignments will **not** work for anyone who doesn't have a course account

- 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

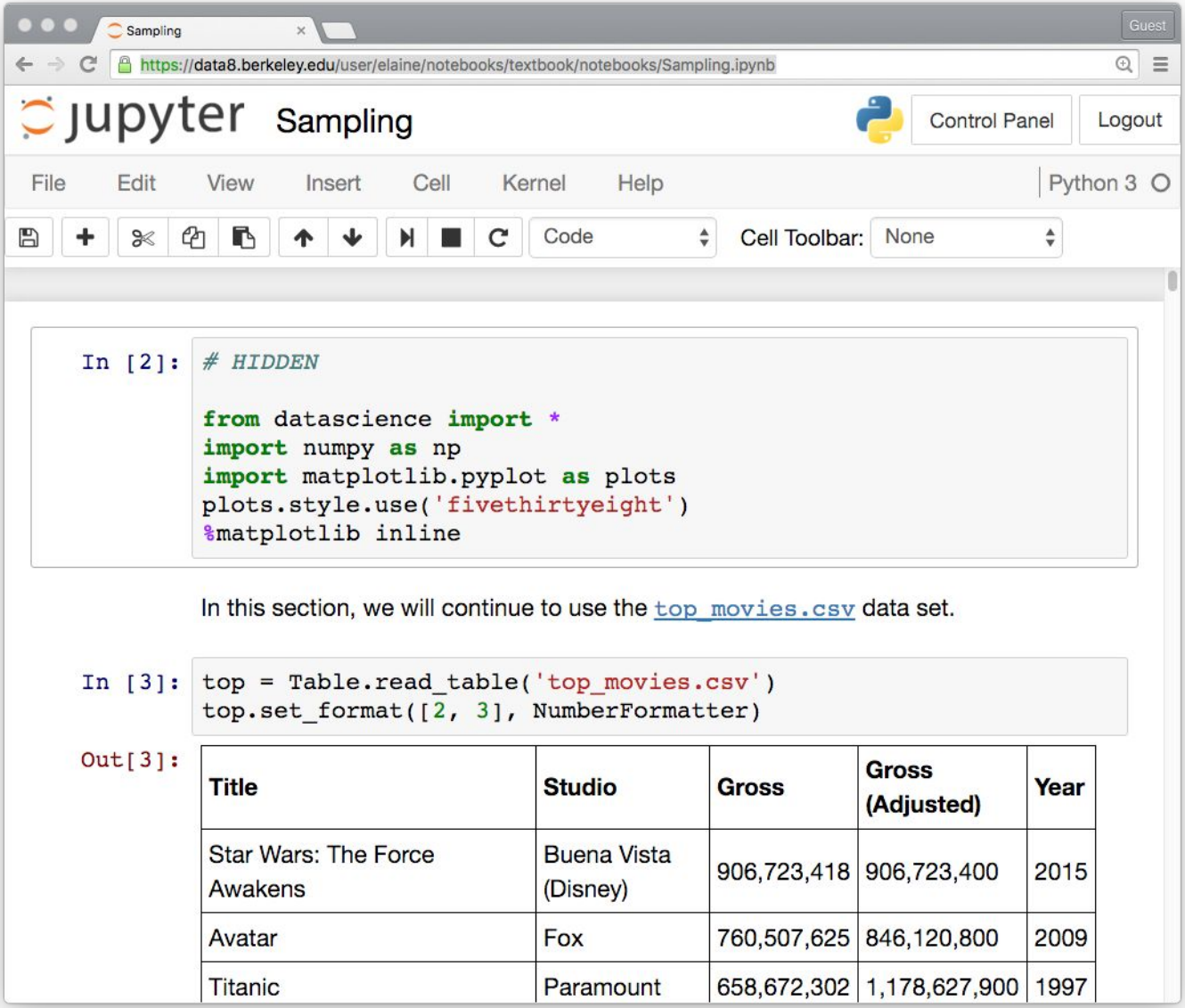
Online textbook: www.inferentialthinking.com



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



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



Link to the same notebook in Binder:

http://bit.ly/sampl-nb

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 JuypterHub 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

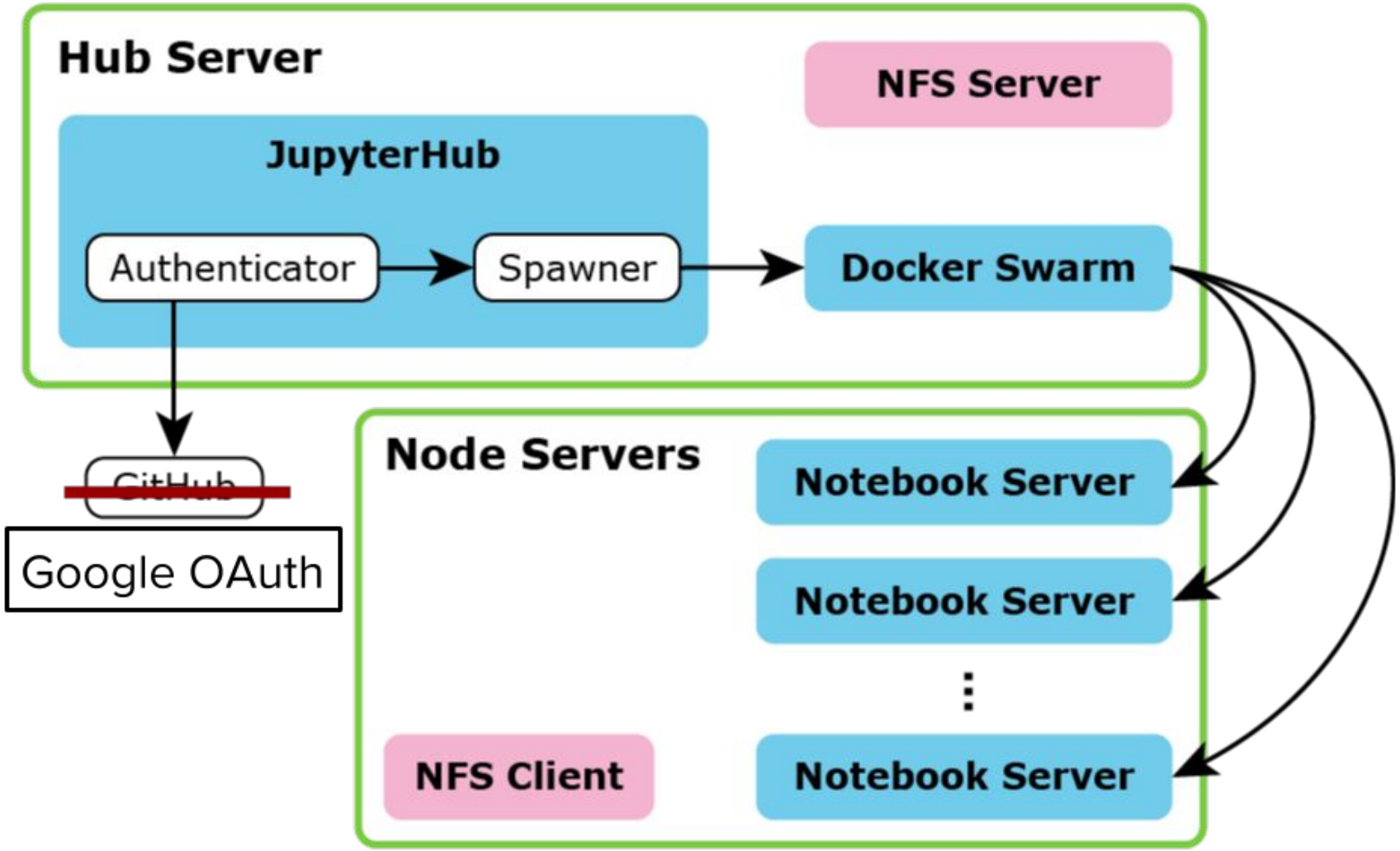
- We were about to level out a new basement to give students computers with Jupyter installed...
- ...until we discovered that **Jessica Hamrick** had deployed JupyterHub to the cloud for a UC Berkeley cognitive science class of 220 students
- We thought we could do it too!

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

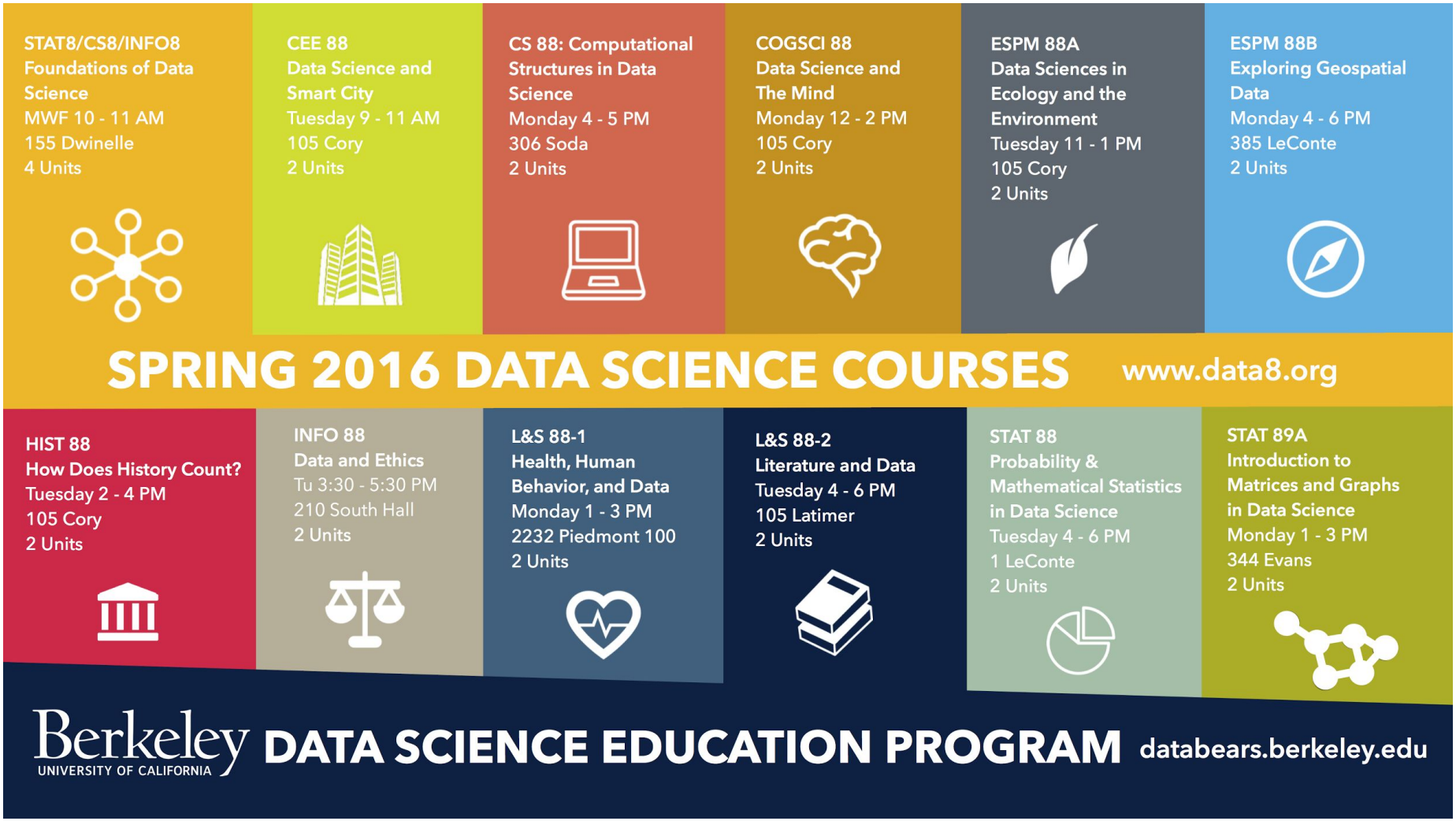


Modified from a diagram by Jessica Hamrick

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



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 made a forum thread for these issues, though reports of trouble have greatly declined since we upgraded to JupyterHub 0.5.0.

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 any courses at UC Berkeley that want to use Jupyter notebooks
- Adding JupyterHub for a course 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**

Some 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
Jupyter(Hub) Developers