Data Science at UC Berkeley: 2000 undergraduates, 50 majors, no command line

JUPYTERCON 2017
VINITRA SWAMY, GUNJAN BAID
First day of Data 8, 08/23/17
```python
from datascience import *
import numpy as np
%matplotlib inline
import matplotlib.pyplot as plots
plots.style.use('fivethirtyeight')
import warnings
warnings.simplefilter(action="ignore", category=FutureWarning)

from urlparse.request import urlopen
import re

def read_url(url):
    return re.sub(r'\s+', '', urlopen(url).read().decode())

little_women_url = 'http://data8.org/dataBassets/lec/sp16/little_women.txt'
little_women_text = read_url(little_women_url)
chapters = little_women_text.split('CHAPTER \n\n')[1:]
display = [print(line[:70]) for line in chapters[:10]]

table = Table().with_columns(
    "Jo", np.char.count(chapters, "Jo"),
    "Meg", np.char.count(chapters, "Meg"),
    "Amy", np.char.count(chapters, "Amy"),
    "Beth", np.char.count(chapters, "Beth"),
    "Laurie", np.char.count(chapters, "Laurie")
).cumsum().plot(overlay=True)

Table().with_columns(
    "Characters", [len(c) for c in chapters],
    "Periods", np.char.count(chapters, ".")
).scatter("Periods")
```
The program, in numbers

1500+ undergraduates enrolled in Fall 2017’s Data Science offerings

50/50 gender representation in Data 8

150+ course staff helping manage Data 8

21 Unique data science connector course offerings
The Offerings

1 foundations course, 3 extenders, 20+ connector courses (2 unit seminars), 15+ modules (data science units)

New DS Courses

Fall 2017 Data Science Connector Courses
Diversity

- Nothing but a web browser
- 60+ Majors
- No prerequisites

Year in cumulative undergraduate studies:
- First Year: 16.7%
- Second Year: 28.9%
- Third Year: 25.1%
- Fourth Year: 13.8%
- Fifth Year or Above: 9.9%
- Not an undergrad: 28.8%

How good a programmer do you consider yourself to be?
- No skill at programming: 9.9%
- Terrible programmer: 10.0%
- Bad programmer: 29.8%
- Reasonable programmer: 37.9%
- Good programmer: 10.9%
- Very good programmer: 0.9%
- Don’t know: 0.9%

Graph:
- Fall 2015
- Spring 2016
- Fall 2016
- Spring 2017

- Advanced Data Science Classes (3 new in Sp 17)
- Connector Courses (21 so far)
- Foundations of Data Science (Data 8)
Who are we?

- **Gunjan Baid**
  - Head GSI of Data 8
  - Technical Lead, DSEP Connector Course (CS 88)

- **Vinitra Swamy**
  - DSEP Jupyter Infrastructure Research Team
  - TA of CS Connector Course (CS 88)
Using Jupyter to teach Data Science

THE UC BERKELEY APPROACH
The Start

“... rethinking at a fundamental level what every educated person must know about **quantitative reasoning**: how to effectively understand, process and interpret information, to **inform decisions** in their professional and personal lives and as **citizens of the world** in the 21st century.”

-- from the Chancellor and Provost’s charge to the Data Science Education “Rapid Action Team,” June 2014
Designing a computing course for all

Remove “command line tax”

Make computing relevant for majors that are not CS/Stats

Have course content that appeals to all demographics

Increase computing diversity
JupyterHub

Result: A student with a chromebook can have a fully functional data science environment
Data 8 Course Structure

**Lecture**
- Mon Wed Fri
- 3 hours

**Lab**
- Weekly
- 2 hours

**Lab**
- Weekly
- 2 hours

**Office Hours**
- Tutoring Sessions
- Review Lectures
- Extra Lectures

**Exams**

**Guerrilla Sections**

**Homework Parties**

**Extra Lectures**

**Mandatory**

**Support Resources**
datascience Library

“I have always imagined Paradise will be a kind of library.”

- Jorge Luis Borges
Data Visualization

Project 1 -- Analyzing global population infant mortality rates.
Teaching Classification with Song Lyrics

Project 3 – Classifying song lyrics as hip hop and country

Kaggle Competition
Diverse, Real World Problems

Cancer Screening – what does a positive test result mean?

WWII – RAF predicts number of enemy German warplanes

Estimate the age of the universe using nearby stars

Conditional Probability

Bootstrap Sampling + Confidence Intervals

Regression
How do we enable an abstraction for students?

Infrastructure Architecture
## Challenges

### Technical

- Real time support for students
  - Real consequences
- Beyond Data 8
  - Exporting work for future student use

### Pedagogical

- Balance of computer science/statistics/domain applications
- Class Structure
  - Worksheets vs. Jupyter labs
  - Tests are pencil and paper, real world is not
How we manage 1000+ active users

- Interfacing with Students
- Solving needs as they arrive
- Surrounding Ecosystem
- Autoscaler bot channel
- Connector Courses
The Student Workflow

JUPYTERHUB AND THE SURROUNDING ECOSYSTEM
Data Science Education Stack

- Datascience package, Tables abstraction
- Jupyter Notebooks
- JupyterHub, surrounding ecosystem
- Kubernetes, Docker
The Technology

- Interact Links
- JupyterHub
- OK
The Student Workflow

1. Obtain assignment
2. Complete Assignment
3. Submit Assignment

Interact Links
JupyterHub
OK
The Motivation
The Student Workflow

1. Obtain assignment
2. Complete Assignment
3. Submit Assignment

Interact Links
JupyterHub
OK
Working on Notebooks: JupyterHub
JupyterHub: The Benefits

- No installation or setup
- Easy to troubleshoot
- Equal compute resources
The Student Workflow

1. Obtain assignment
2. Complete Assignment
3. Submit Assignment

Interact Links
JupyterHub
OK
Obtaining Notebooks: Interact Links

https://datahub.berkeley.edu/user-redirect/interact?
account=data-8&repo=social-networks-connector&branch=gh-pages&path=lab01
1. Click assignment "interact link"

### Data 8 Spring 2017

- Complete the Data 8 Final Online Survey by Saturday 5/6.
- Recover a lost point on your final exam by completing the online surveys!
- If 85% of students complete each survey, then survey completers will recover an additional lost point on the final.

- Independent data exploration
  - Instructions are in the Participation section of the course policies.
  - An FAQ is posted on Piazza.
  - Sign up for the exploration showcase 2pm-5pm on Wednesday 5/3 in BIDS (Doe Library).
  - Use the exploration assignment to submit by Tuesday 5/2 @ 11:59pm.

- Final exam review
  - Review lectures by GSIs on Monday 5/1 & Wednesday 5/3 11am in 1 Pimentel
  - Topical review during lab periods
  - Past exams and review resources are on the resources page.
  - This Piazza post describes all review resources, past exams, etc.

### Calendar

**Instructor:** John DeNero

**Lecture:** MWF 11-12 in 1 Pimentel

<table>
<thead>
<tr>
<th>Date</th>
<th>Topic</th>
<th>Lecture</th>
<th>Reading</th>
<th>Assignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wed 1/18</td>
<td>Introduction</td>
<td>Demos, Slides, Video</td>
<td>1.1, 1.2, 1.3</td>
<td>Lab 01: Expressions, Worksheet</td>
</tr>
<tr>
<td>Fri 1/20</td>
<td>Cause and Effect</td>
<td>Slides, Video</td>
<td>Chapter 2</td>
<td>Homework 01 (Due Thu 1/26)</td>
</tr>
</tbody>
</table>
Computational and Inferential Thinking

The Foundations of Data Science

By Ani Adhikari and John DeNero

Contributions by David Wagner and Henry Milner

This is the textbook for the Foundations of Data Science class at UC Berkeley.

View this textbook online on Gitbooks.

The contents of this book are licensed for reuse under Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0)
The Student Workflow

1. Obtain assignment
2. Complete Assignment
3. Submit Assignment

Interact Links
JupyterHub
OK
Submitting Notebooks: OK

Automate Grading & Personalize Feedback.
OK autogrades programming assignments, facilitates submission, composition feedback, and analytics for your class.

Used by courses with millions of submissions
Berkeley CS 61A Data Science 8 CS 168
Then why is the [450, 1500) bar so much shorter than the [400, 450) bar?

Write your answer here, replacing this text.

Submit the assignment

You have now completed lab 1! You can run the first cell below to regrade questions 1 and 2, for which autograder tests were provided.

```python
# For your convenience, you can run this cell to run all the tests at once!
import os
_= [ok.grade(q[:3]) for q in os.listdir("tests") if q.startswith('q')]
```

Once you have checked your solutions, please run the below cell to submit your lab to the OKpy autograder site. Once you run the cell, you will see a URL for the OKpy autograder site. You can click on this URL to verify that your lab was properly submitted.

```python
_= ok.submit()
```

1. Run submit cell
500: Internal Server Error
Redirect loop detected.
Takeaways

• Everyone can learn data science!
• The challenge is removing barriers to entry
• Get started on using these tools:
  
gunjanbaid.github.io/pages/jupytercon/
  
try.datahub.berkeley.edu
Related JupyterCon Events

Upcoming

• Meet the Expert session with Gunjan
• Jupyter User Testing (bit.ly/jupytercon-usertesting)

Past

• Managing a 1000+ student JupyterHub without losing your sanity (Talk)
• Deploying JupyterHub for students and researchers (Tutorial)
Thank You

- Yuvi Panda
- Ryan Lovett
- Professor David Culler
- Professor Cathryn Carson
- Professor John DeNero
- Professor Ani Adhikari
- Professor David Wagner
- Data Science Education Program
- Data 8 Staff
- UC Berkeley Division of Data Science
- Berkeley Institute for Data Science
- ... and many more!
Data Science at UC Berkeley: 2000 undergraduates, 50 majors, no command line

JUPYTERCON 2017
VINITRA SWAMY, GUNJAN BAID