Stats 170A/B, Project in Data Science

Michael Carey and Padhraic Smyth
Department of Computer Science
Bren School of Information and Computer Sciences
University of California, Irvine
Outline of Today’s Lecture

• Class organization and schedule

• Discussion about projects

• Python software

• Data science in the real-world
Philosophy behind this Class

• Provide an experience of how data science works in the real-world
  – Defining a problem
  – Identifying, understanding, exploring relevant data
  – Extracting, cleaning, management of data
  – Exploration and analysis of data
  – Building models from data (e.g., via machine learning)
  – Evaluating models: how well do they predict
  – Communicating your results to others

• Tie together ideas from different courses you have taken and give you experience in applying these ideas to real-world data
  – Databases, software, algorithms, machine learning, statistics
Components of Data Science

Statistics
(Mathematical and Probabilistic Foundations)

Computing
(Algorithms and Software)

Applications
(Analyzing Real Data)
Course Outline

• Students work individually or in 2-person teams to solve a data science problem using real-world data sets

• Students will define their own projects (with input from instructors)

• Projects will include the full pipeline of data analysis activities
  – Data management
  – Data extraction (e.g., from Web sites)
  – Data cleaning, exploration
  – Data analysis and predictive modeling (machine learning, statistics)
  – Experimental evaluation
  – Reporting
Organizational Items

• Class Website:
  – Class wiki page
  – This is where to find assignments, links to resources such as software, data sets, project guidelines, etc

• Meet weekly in Professor Carey’s office
  – Small class, don’t need a full lecture hall

• No official office hours (given the class size) but if you want to meet with either of us outside of class time please email us to set up a time.
Organizational Items

• 2-quarter class (Winter and Spring)
  – Think of it as one 20-week class
  – Will propose and define your project this quarter and work on it in Spring

• No midterms or final exam
  – But there will be regular reporting and some presentations
  – Also, individual homework assignments during the first six weeks

• Textbook and Reading Materials
  – No official textbook
  – Links to relevant texts (available online via UCI library) on the class wiki page
Textbooks

- *Data Wrangling with Python: Tips and Tools to Make Your Life Easier*
  By Jacqueline Kazil and Katharine Jarmul, O'Reilly Media, 2016.

- *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython* (2nd Edition)
  By William McKinney, O'Reilly Media, 2017.

- *Principles of Data Wrangling: Practical Techniques for Data Preparation*

- *Mining the Social Web* (2nd Edition, Chapters 1 and 9 in particular)
  By Matthew Russell, O'Reilly Media, 2014.

- *Hands-On Machine Learning with Scikit-Learn and TensorFlow* (Chapters 1 through 4 in particular)
  By Aurelien Geron, O'Reilly Media, 2017.

All of these titles are available for free online via the UCI Library's subscription to Safari Books Online ([http://proquest.safaribooksonline.com/](http://proquest.safaribooksonline.com/)).
How this Course will work

• Q1: Weeks 1 to 6: Lectures and Assignments
  – Review general principles of data science
  – Weeks 1 to 3: databases, data extraction, data cleaning
  – Weeks 4 to 6: text analysis, data exploration, machine learning
  – Combination of lectures, assignments, and background reading

• Q1: Weeks 7 to 10: Project Proposals
  – Project proposals from student teams
  – Feedback from instructors, refine proposal, oral presentation at end of quarter

• Q2: Work on Projects
  – Build and use a prototype system/pipeline
  – Develop ideas, implement algorithms, make use of libraries and packages
  – Conduct experiments with real data sets
  – Test and evaluate your system in a systematic manner
  – Communicate your results (presentations and reports)
<table>
<thead>
<tr>
<th>Date</th>
<th>Lecture Topic</th>
<th>Relevant Resources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mon, 1/08</td>
<td>Course overview and plan</td>
<td></td>
</tr>
<tr>
<td>Wed, 1/10</td>
<td>Dusting off your databases</td>
<td>Any CS122a textbook</td>
</tr>
<tr>
<td>Mon, 1/08</td>
<td><em>(No Class)</em></td>
<td></td>
</tr>
<tr>
<td>Wed, 1/17</td>
<td>Data wrangling concepts and issues</td>
<td>Hellerstein <em>et al</em> book</td>
</tr>
<tr>
<td>Mon, 1/22</td>
<td>Python data wrangling basics</td>
<td></td>
</tr>
<tr>
<td>Wed, 1/24</td>
<td>Pandas, Dataframes, and SQL</td>
<td></td>
</tr>
<tr>
<td>Mon, 1/29</td>
<td>XML, JSON, and semistructured data</td>
<td></td>
</tr>
<tr>
<td>Wed, 1/31</td>
<td>Twitter, Tweepy, and SQL vs. NoSQL databases</td>
<td></td>
</tr>
<tr>
<td>Mon, 2/05</td>
<td>Exploratory data analysis and data visualization</td>
<td></td>
</tr>
<tr>
<td>Wed, 2/07</td>
<td>Text analysis techniques</td>
<td></td>
</tr>
<tr>
<td>Mon, 2/12</td>
<td>Machine learning and statistical modeling in Python I</td>
<td></td>
</tr>
<tr>
<td>Wed, 2/14</td>
<td>Machine learning and statistical modeling in Python II</td>
<td></td>
</tr>
<tr>
<td>Date</td>
<td>Lecture Topic</td>
<td>Relevant Resources</td>
</tr>
<tr>
<td>-----------</td>
<td>---------------------------------------------------</td>
<td>--------------------</td>
</tr>
<tr>
<td>Mon, 2/19</td>
<td><em>(No Class)</em></td>
<td></td>
</tr>
<tr>
<td>Wed, 2/21</td>
<td>Project planning, proposals, and guidelines</td>
<td></td>
</tr>
<tr>
<td>Mon, 2/26</td>
<td>Project idea meetings</td>
<td></td>
</tr>
<tr>
<td>Wed, 2/28</td>
<td>Project idea meetings (cont.)</td>
<td></td>
</tr>
<tr>
<td>Mon, 3/05</td>
<td>Project planning meetings</td>
<td></td>
</tr>
<tr>
<td>Wed, 3/07</td>
<td>Project planning meetings (cont.)</td>
<td></td>
</tr>
<tr>
<td>Mon, 3/12</td>
<td>Oral project proposal meetings</td>
<td></td>
</tr>
<tr>
<td>Wed, 3/14</td>
<td>Oral project proposal presentations</td>
<td></td>
</tr>
</tbody>
</table>
Assignment 1

Refer to the Wiki page

Due noon on Wednesday January 18th to EEE dropbox

Will require (e.g., the EDB distribution of) PostgreSQL (https://www.enterprisedb.com/downloads/postgres-postgresql-downloads).
<table>
<thead>
<tr>
<th>Assignment</th>
<th>Due Date (and Time)</th>
<th>Topic</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>HW 1</td>
<td>Wed, 1/17 (11:45 PM)</td>
<td>Schemas and SQL Refresher</td>
<td>HW1</td>
</tr>
<tr>
<td>HW 2</td>
<td>Mon, 1/22 (11:45 PM)</td>
<td>Data Wrangling Principles</td>
<td>HWn</td>
</tr>
<tr>
<td>HW 3</td>
<td>Mon, 1/29 (11:45 PM)</td>
<td>Data Wrangling with Python and Pandas</td>
<td>HWn</td>
</tr>
<tr>
<td>HW 4</td>
<td>Mon, 2/05 (11:45 PM)</td>
<td>Capturing and Wrangling Twitter Data</td>
<td>HWn</td>
</tr>
<tr>
<td>HW 5</td>
<td>Mon, 2/12 (11:45 PM)</td>
<td>Exploratory Data Analysis and Visualization</td>
<td>HWn</td>
</tr>
<tr>
<td>HW 6</td>
<td>Wed, 2/21 (11:45 PM)</td>
<td>Taming Tweets with ML and Statistics</td>
<td>HWn</td>
</tr>
<tr>
<td>Proj 1</td>
<td>Mon, 3/07 (11:45 PM)</td>
<td>Written Project Proposal, Take 1</td>
<td>Projn</td>
</tr>
<tr>
<td>Proj 2</td>
<td>Mon, 3/12 (11:45 PM)</td>
<td>Written Project Proposal, Take 2</td>
<td>Projn</td>
</tr>
<tr>
<td>Proj 3</td>
<td>Wed, 3/14 (due in class)</td>
<td>Oral Project Proposal</td>
<td>Projn</td>
</tr>
</tbody>
</table>
Grading

• Only one grade, assigned at end of Spring quarter

• Winter quarter (50% of total grade)
  – 50% project proposal
  – 40% homeworks
  – 10% class participation

• Spring quarter (50% of total grade)
  – Distributed across project progress reports, final report, class presentations and participation

• Participation = attending class and participating in class discussion

• No grading of late homeworks
Academic Integrity

- Students will be expected to adhere to the UCI and ICS Academic Honesty policies (see [http://www.editor.uci.edu/catalogue/appx/appx.2.htm#academic](http://www.editor.uci.edu/catalogue/appx/appx.2.htm#academic) and [http://www.ics.uci.edu/ugrad/policies/index.php#academic_honesty](http://www.ics.uci.edu/ugrad/policies/index.php#academic_honesty) to read their details).

- Any student found to somehow be involved in cheating or aiding others in doing so will be academically prosecuted to the maximum extent possible: that means that you could fail this course in its entirety. (Ask around - it's happened.) Just say no to cheating!

- This information and associated links are also posted on the class Website
Questions outside Class? use Piazza

- Use Piazza for questions (outside of class time) related to the class
  - Assignments, lectures, projects, data sets, ideas, etc
  - Link = [piazza.com/uci/winter2018/stats170a/home](http://piazza.com/uci/winter2018/stats170a/home)

- Instructors will try to quickly answer questions
  - Students should also feel free to also answer questions
  - If you wish you can use “private mode” to ask questions that only the Professor will see
  - (This way you won’t get lost in our daily faculty e-mail overload)
Introductions

• Instructors
  – Professor Mike Carey
    • Research: database systems, data management, ... → “Big Data” 😊
    • Industry: 13 years in industry, continues to work with startups (e.g., Couchbase), ...
  – Professor Padhraic Smyth
    • Research: machine learning, statistics
    • Industry: consults/works with Google, Samsung, Amazon, eBay, Netflix, ..... 

• Students
  – Introduce yourself
  – What do you hope to get out of the project class?
  – Programming skills?
CLASS PROJECTS
Projects

- Individual or 2-person teams
  - Note that Assignments in weeks 1 to 6 are *not* team-based – these will be worked on and submitted individually
  - For 2-person teams we expect twice as much output and contributions of each individual to be clearly identified in reports

- Each team will propose its own project
  - Suggestions for multiple different projects will be provided
  - Extensive use of libraries (in addition to writing some of your own code)

- Projects will be graded based on
  - Initial proposal
  - Weekly updates
  - Intermediate and final reports
  - In-class presentation

[We will discuss all of this in more detail in future lectures]
What we expect in a Class Project

• Required components
  – Automatically extract a large-scale data set from Twitter
  – Combine Twitter data with at least one other large-scale data set
  – Make use of data management, cleaning, exploration, visualization tools
  – Develop a prediction/forecasting system using the data sets

• Software development
  – You will make use of existing libraries and tools (e.g., PostgreSQL and Python)
  – You are also expected to implement some components of the pipeline yourself

• Evaluation
  – You will need to systematically evaluate your prototype
    • E.g., runtime, predictive accuracy, accuracy as a function of data set size, etc.

• Reporting
  – You will be required to generate reports, graphs, Jupyter notebooks, etc.
Sources of large data sets that could be used for projects

Twitter data: large streams of tweets via Twitter API

Text from 4 million Wikipedia articles
Example of a Class Project

• Data sources
  – Twitter API: tweets mentioning certain keywords, over time, with metadata
  – Census or government maps of population by US county
  – Weather data over time for US locations
  – Historical data on consumer confidence over time

• Create query tool that can compute relative popularity of a keyword
  – over time (time-series plot)
  – Over space (tweets are mapped to location)

• Extension 1
  – Predict popularity of a keyword by week, given historical data

• Extension 2
  – Investigate correlation of keywords with weather data (time and space)

• Extension 3
  – How well can consumer confidence be predicted from tweet sentiment over time?
Average Happiness for Twitter

https://hedonometer.org/index.html
Another Possible Class Project

• Define a set of entities of interest
  – E.g., movie stars listed in the IMDB data set (see Homework 1)
  – E.g., sports stars, musicians, etc, from Wikipedia
  – E.g., products and brands (e.g., cars, shoes, phones, apps)

• Crawl Twitter for historical mentions of these entities
  – E.g., for all of 2014-2017

• Build a system that can answer queries and display results
  – E.g., how many tweets per week did entity A get versus entity B in state X
  – E.g., how many positive versus negative tweets did entity A get over time

• Use machine learning/statistics to forecast
  – Popularity (number of tweets) for any entity for week T, given data to T-1
  – Or predict tweet sentiment (proportion positive/negative) for an entity
Tweets mentioning Coke (green) and Pepsi (red) from chimpler.wordpress.com
Modeling Human Behavior using Social Media

From Lichman and Smyth, ACM SIGKDD 2014
Geolocated Tweets around UC Irvine
https://gis.ncdc.noaa.gov/maps/ncei/summaries/daily
Spatial Data (from US Census)

From https://anaconda.org/jbednar/census/notebook
Spatial Census Data with Color-Coding

From https://anaconda.org/jbednar/census/notebook
Travel Monitoring

Traffic Volume Trends

Traffic Volume Trends is a monthly report based on hourly traffic count data reported by the States. These data are collected at approximately 4,000 continuous traffic counting locations nationwide and are used to estimate the percent change in traffic for the current month compared with the same month in the previous year. Estimates are re-adjusted annually to match the vehicle miles of travel from the Highway Performance Monitoring System and are continually updated with additional data.

See our Frequently Asked Questions (FAQ’s) to find more information on the Traffic Volume Trends report.

To view PDF files, you can use the Acrobat® Reader®.
To view XLS files, you can use the Excel Viewer.

<table>
<thead>
<tr>
<th>Monthly Report</th>
<th>PDF Version</th>
<th>XLS Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>October 2017</td>
<td>PDF (616 KB)</td>
<td>XLS (496 KB)</td>
</tr>
<tr>
<td>September 2017</td>
<td>PDF (704 KB)</td>
<td>XLS (488 KB)</td>
</tr>
<tr>
<td>August 2017</td>
<td>PDF (412 KB)</td>
<td>XLS (488 KB)</td>
</tr>
<tr>
<td>July 2017</td>
<td>PDF (409 KB)</td>
<td>XLS (492 KB)</td>
</tr>
<tr>
<td>June 2017</td>
<td>PDF (408 KB)</td>
<td>XLS (484 KB)</td>
</tr>
<tr>
<td>May 2017</td>
<td>PDF (404 KB)</td>
<td>XLS (487 KB)</td>
</tr>
<tr>
<td>April 2017</td>
<td>PDF (407 KB)</td>
<td>XLS (481 KB)</td>
</tr>
<tr>
<td>March 2017</td>
<td>PDF (428 KB)</td>
<td>XLS (484 KB)</td>
</tr>
<tr>
<td>February 2017</td>
<td>PDF (399 KB)</td>
<td>XLS (458 KB)</td>
</tr>
<tr>
<td>January 2017</td>
<td>PDF (401 KB)</td>
<td>XLS (460 KB)</td>
</tr>
</tbody>
</table>

https://www.fhwa.dot.gov/policyinformation/travel_monitoring/tvt.cfm
OpenStreetMap powers map data on thousands of web sites, mobile apps, and hardware devices.

OpenStreetMap is built by a community of mappers that contribute and maintain data about roads, trails, cafés, railway stations, and much more, all over the world.

Local Knowledge

OpenStreetMap emphasizes local knowledge. Contributors use aerial imagery, GPS devices, and low-tech field maps to verify that OSM is accurate and up to date.

Community Driven

OpenStreetMap’s community is diverse, passionate, and growing every day. Our contributors include enthusiast mappers, GIS professionals, engineers running the OSM servers, humanitarians mapping disaster-affected areas, and many more. To learn more about the community, see the OpenStreetMap Blog, user diaries, community blogs, and the OSM Foundation website.
Inside Airbnb
Adding data to the debate

How is Airbnb really being used in and affecting your neighborhood?

OUT OF MORE THAN 27,000 LISTINGS:

16k are for the entire home (58%)

87% highly available (more than 60 days/year)

29% multi-listings (where the host has other listings)

FILTER by Neighborhood
Chelsea

50+ data points per listing

SEE Airbnb ACTIVITY OVER TIME IN YOUR NEIGHBORHOOD

HOST “JOHN D” 17 listings

VIEW TOP HOSTS’ MULTIPLE LISTINGS

NEXT...

VISIT insideairbnb.com

SHARE it widely
#insideairbnb #illegalhotels #affordablehousing #nyc

DOWNLOAD the data
(open source; 50+ data points per listing)

The data Airbnb doesn’t want you to see!
The Yelp dataset is a subset of our businesses, reviews, and user data for use in personal, educational, and academic purposes. Available in both JSON and SQL files, use it to teach students about databases, to learn NLP, or for sample production data while you learn how to make mobile apps.

The Dataset

4,700,000 reviews
156,000 businesses
200,000 pictures
12 metropolitan areas

1,000,000 tips by 1,100,000 users
Over 1.2 million business attributes like hours, parking, availability, and ambience
Aggregated check-ins over time for each of the 156,000 businesses

https://www.yelp.com/dataset
### Reddit Statistics 2015

<table>
<thead>
<tr>
<th>Stat</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>total # posts</td>
<td>668M</td>
</tr>
<tr>
<td>total # users posting</td>
<td>8.2M</td>
</tr>
<tr>
<td># words per post</td>
<td>30.6</td>
</tr>
<tr>
<td>total # words</td>
<td>&gt;20 billion</td>
</tr>
</tbody>
</table>

---

**Daily # of Reddit Submissions from 2006 - 2015**

Data via Reddit

By Max Woolf — minimaxir.com

Made using R and ggplot2
http://commoncrawl.org/
SOFTWARE ENVIRONMENTS FOR THIS CLASS
Software for Future Assignments and Projects

• **Python**
  – Python will be the primary language we will use in much of this class
  – Assume that all students have a working knowledge of Python 3

• **Packages and Libraries**
  – We will make extensive use of additional packages and libraries in Python, e.g.,
    • Pandas for data manipulation
    • Scikit-learn: machine learning library
    • Scientific computing/graphs/etc: matplotlib, numpy, scipy, etc

You should download and install the Anaconda package: it contains many packages you need for this class
Screenshot of the Spyder IDE

```python
# simple demo of NLTK functions applied to a Web page

# start up NLTK (usually not needed, i.e., already installed)
# import nltk
# nltk.download()

# import the packages that we will need
import nltk, re, print
from nltk import word_tokenize
from urllib import request
from bs4 import BeautifulSoup

# read the text from the Web page for Assignment 1
url = "http://www.ics.uci.edu/~smyth/courses/new175/assignment1.xht"
html = request.urlopen(url).read().decode('utf8')

# strip out most of the HTML stuff...
raw = BeautifulSoup(html).get_text()

# segment the text into sentences (useful for tokenizing - it's not necessary to do
sents = nltk.sent_tokenize(raw)

# extract tokens
tokens = word_tokenize(raw)
tokens[1:20]
tokens[21:50]
```
$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$
scikit-learn

Machine Learning in Python

- Simple and efficient tools for data mining and data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Classification

- Identifying to which set of categories a new observation belong to.
- **Applications**: Spam detection, image recognition.
- **Algorithms**: SVM, nearest neighbors, random forest, ... — Examples

Regression

- Predicting a continuous value for a new example.
- **Applications**: Drug response, Stock prices.
- **Algorithms**: SVR, ridge regression, Lasso, ... — Examples

Clustering

- Automatic grouping of similar objects into sets.
- **Applications**: Customer segmentation, Grouping experiment outcomes
- **Algorithms**: k-Means, spectral clustering, mean-shift, ... — Examples

Dimensionality reduction

- Reducing the number of random variables to consider.
- **Applications**: Visualization, Increased efficiency
- **Algorithms**: PCA, Isomap, non-negative matrix factorization. — Examples

Model selection

- Comparing, validating and choosing parameters and models.
- **Goal**: Improved accuracy via parameter tuning
- **Modules**: grid search, cross validation, metrics. — Examples

Preprocessing

- Feature extraction and normalization.
- **Application**: Transforming input data such as text for use with machine learning algorithms
- **Modules**: preprocessing, feature extraction. — Examples

News

- On-going development: What's new (changelog)

Community

- Questions? See stackoverflow # scikit-learn
- Mailing list: scikit-learn-

Who uses scikit-learn?
Eager Execution

We're announcing eager execution, an imperative, define-by-run interface to TensorFlow. Check out the README to get started today.

LEARN MORE

TensorFlow 1.3 has arrived!

We're excited to announce the release of TensorFlow 1.3! Check out the release notes for all the latest.

UPGRADE NOW

The 2017 TensorFlow Dev Summit

Thousands of people from the TensorFlow community participated in the first flagship event. Watch the keynote and talks.

WATCH VIDEOS
It looks like there is a clear decision boundary between the two classes. Now we need to implement logistic regression so we can train a model to predict the outcome. The equations implemented in the following code samples are detailed in "ex2.pdf" in the "exercises" folder.

Figure from http://nbviewer.jupyter.org/github/jdewitner/ipython-notebooks/blob/master/notebooks/ml/ML-Exercise2.ipynb
DATA SCIENCE IN THE REAL WORLD
Top 5 Market Caps in $B on 03/31/2017

- Apple: 754
- Google: 573.6
- Microsoft: 509.6
- Amazon: 423.6
- Facebook: 410.9

03/31/2017

Same Day, 11 years ago

- ExxonMobil: 368.2
- GE: 361.7
- Microsoft: 278.2
- BP: 237.4
- Citigroup: 234.8

03/31/2006

Graphic from https://medium.com/startup-grind/
What is Data Science?

Data science involves the full lifecycle of data:
from messy unstructured data.....to predictions and decisions

Data science is broader than just databases, statistics, ML, algorithms
.....but these are all critical components

Key aspects of data science include

– Domain knowledge and problem definition
– Data preparation/organization/management
– Understanding of uncertainty (statistics)
– Computing, algorithms, fitting models, machine learning
– Iterative exploration and experimentation
– Human judgement and interpretation
How does Amazon forecast how many items for its warehouses?

From dailymail.co.uk

From www.formaspace.com

From linkedin.com
How does Facebook predict what content to show you?

MONTHLY USERS ON FACEBOOK 2004-2017

Over 30 billion pieces of content shared every month

Over 3 billion photos uploaded each month

The Friendship graph

500M users each connect to an average of 130 other users = ~ 60 Billion Edges

Graphics from Lars Backstrom, ESWC 2011
How do companies decide what ads to show you?
How is Data Science used in each of these Organizations?
Organizations

- facebook
- Google
- Amazon
- Spotify
- Disney
- Kaiser Permanente
- Blizzard
- Honda

Data Science Applications

- Online advertising
- Automated recommendations
- Demand forecasting
- Fraud detection
- Churn prediction
- Automated customer support
Organizations

- facebook
- Google
- Amazon
- Spotify
- Disney
- Kaiser Permanente
- Blizzard
- Honda

Data Science Applications

- Online advertising
- Automated recommendations
- Demand forecasting
- Fraud detection
- Churn prediction
- Automated customer support
Data Pipelines

Unstructured Data → Extracted Data → Transformed Data → Data for Modeling → Predictive Model → Predictions/Decisions
Hidden Technical Debt in Machine Learning Systems

Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

Scullley et al, NIPS 2015 Conference
Example: Predicting Popularity of Music Artists over Time
Example: Predicting Popularity of Music Artists over Time

[Diagram showing data flow from Twitter, Spotify, and Census Bureau to Extracted Data and then to Cleaned Data]
Example: Predicting Popularity of Music Artists over Time
Assignment 1

Refer to the Wiki page

Due 11:45pm on Wednesday January 18th to EEE dropbox

Will require (e.g., the EDB distribution of) PostgreSQL (https://www.enterprisedb.com/downloads/postgres-postgresql-downloads).
<table>
<thead>
<tr>
<th>Date</th>
<th>Lecture Topic</th>
<th>Relevant Resources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mon, 1/08</td>
<td>Course overview and plan</td>
<td></td>
</tr>
<tr>
<td>Wed, 1/10</td>
<td>Dusting off your databases</td>
<td>Any CS122a textbook</td>
</tr>
<tr>
<td>Mon, 1/08</td>
<td><em>(No Class)</em></td>
<td></td>
</tr>
<tr>
<td>Wed, 1/17</td>
<td>Data wrangling concepts and issues</td>
<td>Hellerstein <em>et al</em> book</td>
</tr>
<tr>
<td>Mon, 1/22</td>
<td>Python data wrangling basics</td>
<td></td>
</tr>
<tr>
<td>Wed, 1/24</td>
<td>Pandas, Dataframes, and SQL</td>
<td></td>
</tr>
<tr>
<td>Mon, 1/29</td>
<td>XML, JSON, and semistructured data</td>
<td></td>
</tr>
<tr>
<td>Wed, 1/31</td>
<td>Twitter, Tweepy, and SQL vs. NoSQL databases</td>
<td></td>
</tr>
<tr>
<td>Mon, 2/05</td>
<td>Exploratory data analysis and data visualization</td>
<td></td>
</tr>
<tr>
<td>Wed, 2/07</td>
<td>Text analysis techniques</td>
<td></td>
</tr>
<tr>
<td>Mon, 2/12</td>
<td>Machine learning and statistical modeling in Python I</td>
<td></td>
</tr>
<tr>
<td>Wed, 2/14</td>
<td>Machine learning and statistical modeling in Python II</td>
<td></td>
</tr>
</tbody>
</table>