Day 1: Introduction to Working with Open Data

Raymond Yee

January 21, 2014 (`http://is.gd/wwod1401`)

Goals Today

- Learn the purpose, structure, expectations for the course
- Begin thinking together about open data and Python
- Dive into some programming

Course Overview

INFO 290T- Working with Open Data
`http://www.ischool.berkeley.edu/courses/290t-wod`
Spring 2014 / CCN: 41620
T,Th 2:00-3:30pm 202 South Hall
Office Hours: T, Th 3:30-4:30pm, 302 South Hall (along with *possible virtual office hours*)
Instructor: Raymond Yee, Ph.D.
Contact info: yee@berkeley.edu
Twitter: @WorkingOpenData / @rdhyee
Tutor: AJ Renold (arenold@ISchool.Berkeley.EDU)

bcourses site to be unveiled soon…

Course Description

Open data – data that is free for use, reuse, and redistribution – is an intellectual treasure-trove that has given rise to many unexpected and often fruitful applications. In this course, students will

1. learn how to **access, visualize, clean, interpret, and share data**, especially open data, using **Python**, Python-based libraries, and supplementary computational frameworks
2. understand the theoretical underpinnings of open data and their connections to implementations in the physical and life sciences, government, social sciences, and journalism.

Working with Open Data (WwOD) is a *technical* course with a strong focus on the social-political context and domains of application of open data.

Prerequisite

Info 206 Distributed Computing Applications and Infrastructure or equivalent background with Python.

Grading Scheme

Grading Scheme:

Subject to Change

1. problem sets (25%)
2. mid-term exam (30%)
3. project proposal (5%)
4. final project (25%)
5. participation (15%)

Coverage vs Discovery

- “It doesn’t matter what we cover. It matters what you discover.” Victor Weisskopf –> via Noam Chomsky, YouTube

Coverage vs Discovery

- "It doesn't matter what we cover. It matters what you discover." Victor Weisskopf –> via Noam Chomsky, YouTube
- Learning by Doing: together and by yourself

Main Textbook

Wes McKinney. `Python for Data Analysis`. (O'Reilly Media, 2012). I strongly recommend getting a paper copy as well as accessing any electronic versions

- oreilly.com
- Proquest.safaribooksonline at UCB

IPython Notebook Integral to Course

Working through IPython Notebooks created by the instructor is the primary vehicle for learning at the beginning of the course.

See A gallery of interesting IPython Notebooks · ipython/ipython Wiki

Supplementary Materials

I plan to supplement the book with materials covering the following topics:

- open data, open content in various fields
- using JavaScript, HTML5, CSS together with Python for data presentation, analysis, and visualization, (e.g., d3.js)

In addition to survey materials on the public domain, creative commons, and open data movements, I'll focus us on

- Wikipedia, dbpedia, Freebase data
- census data

and other data sets still to be determined, probably large open scientific data sets

What we did in 2013

A narrative about last year's course co-written by Fernando Perez and Raymond Yee: Exploring Open Data with Pandas and IPython at the Berkeley I School – includes abstracts of last year's projects.

- WwOD 2013 github repo
- List of class resources: homework and midterm exam
- schedule from 2013

Flow of Logic for Course

- Working on exercises -> Working on Projects

Flow of Logic for Course

- Working on exercises -> Working on Projects
- The US Census + the Wikipedia is an integrating framework

Flow of Logic for Course

- Working on exercises -> Working on Projects
- The US Census + the Wikipedia is an integrating framework
- standard Python -> Python in the context of the IPython Notebook -> integrating JavaScript

Flow of Logic for Course

- Working on exercises -> Working on Projects
- The US Census + the Wikipedia is an integrating framework
- standard Python -> Python in the context of the IPython Notebook -> integrating JavaScript
- computing on Wakari -> computing on notebook -> computing on a cluster (and in the cloud)

Course Outline

Last revised: 2014.03.06

1. Introduction to Working with Open Data
2. Setting Up for Python & IPython
3. Setting Up Cont'd: Environments and Contexts
4. Numpy & Pandas Intro
5. Geographical Hierarchies in the Census
6. Generators for Geographic Entities
7. Calculating Diversity I
8. Calculating Diversity II
9. Creating Projects
10. Calculating Diversity III
11. Looking Ahead to Projects, Plotting, and Baby Names
12. Baby Names and Plotting
13. Baby Names II and mpld3
14. Baby Names III
15.
16.
17. MIDTERM (Day 17, 2014-03-18)

Major Deadlines

- Preliminary Project Proposal (Friday March 21, 2014)
- Mid-term exam: Tuesday, March 18, 2014 (Day 17)
- Project Proposal: March 15, 2014 11:59pm Wed, April 2, 2014
- In-class presentations: April 24, April 29, and May 1, 2014 (Days 26, 27, & Day 28)
- Submission of Final Report and Open House/ Poster Session: Tuesday, May 6, 2014 (Day 29)

It is the student's responsibility to notify the instructor(s) in writing by the second week of the semester of any potential conflict(s) and to recommend a solution, with the understanding that an earlier deadline or date of examination may be the most practicable solution.

It is the student's responsibility to inform him/herself about material missed because of an absence, whether or not he/she has been formally excused.

Stay at home if you are sick

Campus Flu Guideline:

> *Students that they should not come to class if they become ill.The University has adopted the CDC recommendation that members of the campus community who develop flu-like illness should self-isolate until at least 24 hours after they are free of fever or signs of fever without the use of medication. Let your students know that they should follow this recommendation in deciding whether or not to come to class*

In return: there will be flexibility and good judgment in how course requirments will be handled.

Projects

- Participants will work on tangible **projects related to the overall theme**.
- The projects need to be thorough analyses of some open dataset or datasets.
- Students can select from a list of projects or they can propose other projects of comparable scope and intent.
- Students will have opportunities to brainstorm ideas, choose a specific focal point (drawing from structured feedback from other students and the instructor), craft a proposal for their projects, and then present their work at the end of the course.

Laptops in classroom

I would like everyone to bring a notebook computer to class so that we can work together in class on programming assignments. **If you are not able to do so, check in with me.**

Why Python?

see McKinney's narration: `http://proquest.safaribooksonline.com/book/programming/python/9781449323592/1dot-preliminaries/id2700570`

Working definition of open data

From `http://en.wikipedia.org/w/index.php?title=Special:Cite&page=Open_data&id=532390265`:

> *Open data is the idea that certain data should be freely available to everyone to use and republish as they wish, without restrictions from copyright, patents or other mechanisms of control.*

`http://opendefinition.org/`:

> *A piece of content or data is open if anyone is free to use, reuse, and redistribute it — subject only, at most, to the requirement to attribute and/or share-alike.*

OKFestival /OKCon as indicator of vibrancy of the international open data community

- http://okfestival.org/
- http://okfestival.org/after/
- video streams
- start with opening plenary
- final report
- OKCon - Open Knowledge Conference (Sept 2013 in Geneva)

list of working groups

http://okfn.org/wg/ includes:

- Open Government Data
- OpenGLAM (galleries, libraries, archives, and museums)
- Open Economics
- Open Science
- Open Development
- Open Sustainability

A Motivating Example: Racial Dot Map

`http://bit.ly/rdotmap`

and

`http://bit.ly/rdotmapintro`

With any luck, we will not only understand how the map works, we'll also be able to reproduce it and enhance it by the end of the semester. That is how to turn **Census 2010** data into a map.

Activity 1

Group activity – discuss and enter answers at `http://bit.ly/wwod1401Q`

- What you hope to learn and accomplish in the course?
- Name 2 to 4 types of data (or datasets) that interest or intrigue you. Bonus: explain why
- What's one of the more complicated example of Python programming you've done so far?
- What questions do you have for the instructor?

Activity 2: Setting up for Programming

We'll study the **population of countries** before we dive into the US Census.

- For homework and in the next class, we'll focus on getting your own laptop ready for programming.
- For today, I want you to sign up for the free account on Wakari.io.
- Then, I'll show you how to load up today's notebook into Wakari: `https://raw.github.com/rdhyee/working-open-data-2014/master/notebooks/Day_01_A_World_Population.ipynb`

Upcoming events

- (Jan 21) Twitter / sfopendrinks: Lovers of openness! Join us . . .
- (Feb 1) Wikipedia:Meetup/ArtAndFeminism - Wikipedia, the free encyclopedia
- (Apr 3-4, 5-6) Twitter / hypothes_is: Save the date: I Annotate 2014. . . .
- (July 15-18, 2014 in Berlin)OKFestival 2014

Homework

Try to get IPython installed on your computer before class

AJ wrote a nice set of notes on how to do so: `https://github.com/rdhyee/working-open-data-2014/wiki/IPython-Installation-Options`

Readings

- read Python for Data Analysis Chap 1. Preliminaries : Safari Books Online The instructions for using Enthought Python Distribution are out of date. If you are looking for a distribution, follow the installation instructions for Anaconda for your computer platform.
- read `PfDA`, Chap 3 Python for Data Analysis > 3. IPython: An Interactive Computing and Development Environment
- skim `PfDA`, Appendix: Python Language Essentials – to help remind yourself of key elements of standard Python
- skim `PfDA`, Chap 2 Introductory Examples

APPENDICES

Random and not-so random questions for me that open data can help answer

Reading the news, world news, local news, tech news, understanding new contexts, deepening old interests, controlled serendipity.

- What music to listen to
- What book to read?
- When can I see the next episode of White Collar?
- When was BWV 156 (Ich steh mit einem Fuß im Grabe) first performed?
- How to invest our money?
- What programming language to learn next?
- What charity to give money to?
- What restaurant to try?
- What to cook?
- How should we take care of our physical health?

Some Big Questions for the Course

- What are the essential characteristics of open data?
- What are the costs and benefits of open data?
- How to map the universe of open data? What's out there? What data is not available in open form?
- What can we learn from open data?
- What business models?
- What are people doing with open data?
- What are the different common formats used to represent open data? (e.g., CSV, XML, KML, SHP in data.gov) – and how can we use Python to process those formats?
- What are the issues that we face in combining open data with closed data

data.gov as a good example

`http://www.data.gov/`

`http://www.data.gov/about`:

> *The purpose of Data.gov is to increase public access to high value, machine readable datasets generated by the Executive Branch of the Federal Government.”*
>
> *A primary goal of Data.gov is to improve access to Federal data and expand creative use of those data beyond the walls of government by encouraging innovative ideas (e.g., web applications). Data.gov strives to make government more transparent and is committed to creating an unprecedented level of openness in Government. The openness derived from Data.gov will strengthen our Nation’s democracy and promote efficiency and effectiveness in Government.*

datasets in data.gov

Motivation: why I care about open data and why you might care

Traditional motivations given for open government data:

- transparency
- accountability
- efficiency
- innovation

My personal interests in the area:

- Open data useful testbed for working on data of all sorts, because of zero financial costs and minimal restrictions on use, reuse, redistribution
- Growing community around open data because of these low barriers. . . democratization of data. . . many more of us can participate in working with open data and attract a wide range of people I love to learn and to think and to understand, a big believer of computational and information systems as mind augmenters/extenders and open data (as

Examples of Open Data

- http://www.data.gov/
- https://data.sfgov.org/
- https://data.acgov.org/
- http://data.openoakland.org/
- http://www.socrata.com/discover/video-case-study-somerville-ma/
- http://sunlightfoundation.com/projects/
- http://oad.simmons.edu/oadwiki/Data_repositories#Astronomy
- http://opencontext.org/
- http://courtlistener.com/
- http://aws.amazon.com/publicdatasets/
- http://openmetadata.lib.harvard.edu/bibdata
- http://www.bart.gov/schedules/developers/api.aspx / http://www.bart.gov/schedules/developers/index.aspx
- http://www.ncdc.noaa.gov/