

Advanced Data After Dark Workshop

Instructors: Shannon McWeeney, PhD | Ted
Laderas, PhD | Melissa Haendel, PhD |
David Dorr, MD, MS | Michael Mooney,
PhD | Jackie Wirz, PhD | Nicole Vasilevsky,
PhD | Bjorn Pederson, MA

Monday, May 23rd : 5 to 7pm

Data Wrangling (Pandas and Python)

Data in Flat Files: Unstructured and Semi-structured data

Data in Databases (relational and SQL)

Tuesday, May 24th : 5 to 7pm

Data tells a Story: QA/QC

EDA and Interactive Visualization (Building R/Shiny Dashboard)

Wednesday, May 25th : 5 to 7pm

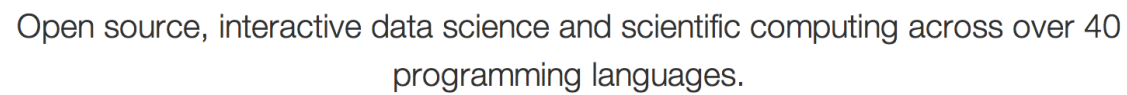
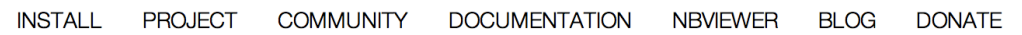
Supervised Learning Algorithms (focused - 2)

Thursday, May 26th : 5 to 7pm

Handle with Care: Caveat and Advice for Machine Learning, Dimensionality reduction, Validation and Evaluation

What is Data Wrangling?

(Also known as “Data Munging”)



Browser test: <https://try.jupyter.org>

Python Packages of Interest

- **Numpy: support for creating and efficiently manipulation large data structures**
- **Matplotlib: visualization of data through graphical plots (requires numpy)**
- Scipy: collection of algorithms for scientific computing
- Scikit-learn: collection of machine-learning algorithms
- **Pandas: data analysis tools + specialized data structures**
- Statistical Packages: Statsmodels + Rpy2(interface between Python and R)

How do we access data?

- Bulk downloads
- API access
- Web scraping

Sources of Error

- Data entry
- Measurement
- Distillation
- Data integration

Data Manipulations

- Filtering, or subsetting: Remove observations based on some condition
- Transforming: add new variables or modify existing variables (e.g., log-transforming)
- Aggregating: collapse multiple values into a single value (e.g., summing or taking means)
- Sorting: change the order of values

Jupyter Python Notebook

Code review

Data Formats

- Delimited values
- Comma Separated Values (CSV)
- Tab Separated Values (TSV)
- Markup languages
 - Hypertext Markup Language (HTML5 / XML)
 - JavaScript Object Notation (JSON)
 - Hierarchical Data Format (HDF5)
- Ad hoc formats

Considerations

- Structured vs Unstructured Data
- Database (SQL vs noSQL)

Caveats re CSV

- CSVs have inherent schemas
 - Schema loss (Db dump)
 - Relationship loss –denormalized dump (a join of two tables)
 - Data Dictionary critical (data type etc)

Example

Date; B1;B2;CS1;CS2;DP;PD;R1;SU

28/01/2016;63;0;;27;16;15;4;22;

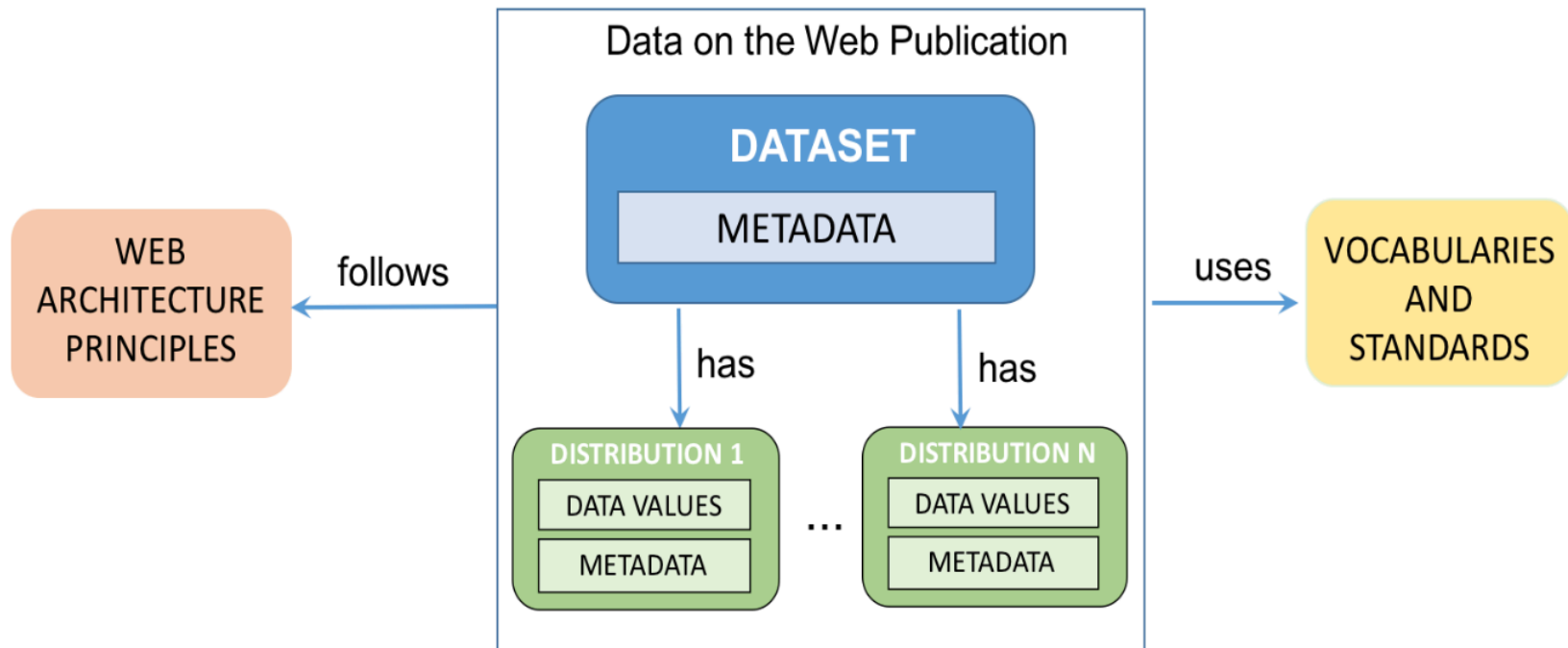
Code to Parse

```
Extracted_data = pandas.read_csv('saddata.csv',  
sep=';', parse_dates=['Date'], dayfirst=True,  
index_col='Date')
```

Web Scraping

- Gives access to data not contained in an API or a file
- Page-reading vs. Screen-reading
- IP (Copyright + Trademark)
 - Legal Notices
 - Review robots file
 - Contact site or Digital media legal organization

Our Goal: Data Standards + Ontologies



<http://www.w3.org/TR/dwbp/>

What data is okay to scrape?

- Public
- Non-sensitive
- Anonymized
- Allowed by License/IP considerations
- Fully referenced -cite source!

Public does not equal consent



[Search](#) [Member List](#) [Calendar](#) [Help](#)

Hello There, Guest! ([Login](#) — [Register](#))

Current time: 05-24-2016, 05:39 AM

[OpenPsych forums](#) / [Forums](#) / [Submissions](#) / [ODP] The OKCupid dataset: A very large public dataset of dating site users

[NEW REPLY](#)

[ODP] The OKCupid dataset: A very large public dataset of dating site users

[Threaded Mode](#) | [Linear Mode](#)

05-08-2016, 09:22 PM (This post was last modified: 05-19-2016 08:32 AM by Emil.)

Post: #1

Emil



Admin, reviewer (ODP)

Posts: 1,184
Joined: Mar 2014
Reputation: 0

[ODP] The OKCupid dataset: A very large public dataset of dating site users

Journal:

Open Differential Psychology.

Authors:

Emil O. W. Kirkegaard
Julius D. Bjerrekær

Title:

The OKCupid dataset: A very large public dataset of dating site users

Abstract:

<http://openpsych.net/forum/showthread.php?tid=279>

Developer Tools (Chrome)



GAPMINDER WORLD VIDEOS DOWNLOADS TEACH IGNORANCE DATA



How not to be ignorant about the World

Ola Rosling & Hans Rosling show simple rules of thumb for dismantling global misconceptions and beating chimpanzees.

Watch TED talk ▶



The screenshot shows the Chrome Developer Tools interface. The **Elements** panel on the left displays the HTML structure of a webpage, including the `<body>` tag and various nested elements like `<div>` and `<script>`. The **Styles** panel on the right shows the CSS rules applied to the selected element. The `body` style is expanded, showing properties like `height: 100%`, `background-color: #fff`, `font-family: 'Open Sans', sans-serif`, and `color: #231f20`. A visual representation of the box model (margin, border, padding) is shown on the right, with dimensions of 867 x 753. The **Properties** panel at the bottom right shows a list of properties, including `background-attachment`, `scroll`, `background-clip`, and `border-box`.

Question #1

- How many incidents have occurred where rate of patient deaths for all cardiac surgical procedures was "significantly higher" than the New York statewide rate?

Where do we find the data to answer this?



Services

News

Government

Local

Sign Up Log In



HEALTH DATA NY

DATA.NY.GOV

DEVELOPERS

HELP

ABOUT

Search Health Data NY

https://health.data.ny.gov/resource/dk4z-k3xb.json

Apps XFINITY Bookmarks OHSU Apple - Find Out Ho

Using School Data to Understa

New childhood obesity estimates available at the district, county and reg



Consumer Resources



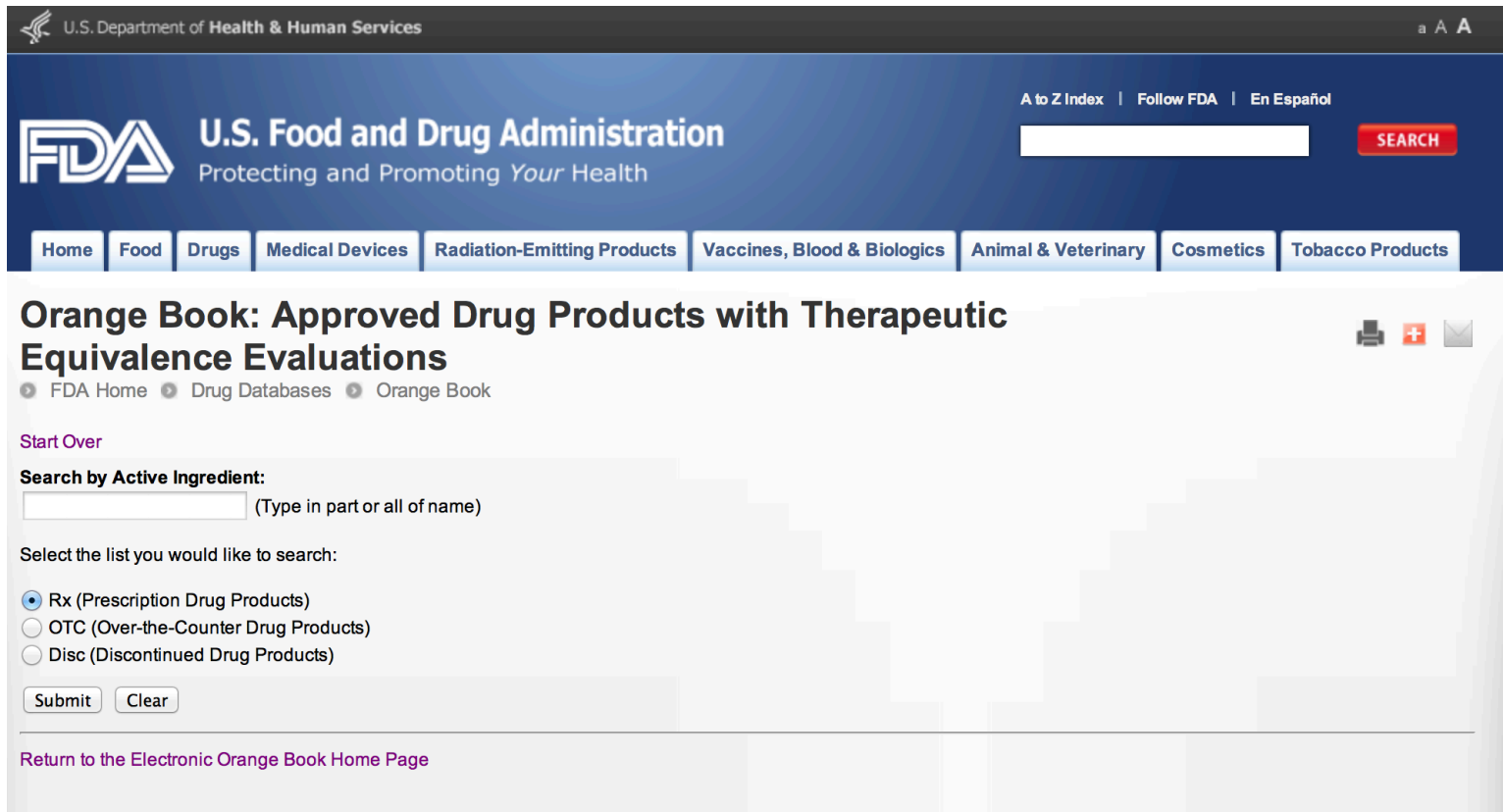
Environmental Health

health.data.ny.gov

```
{
  "region": "N/A",
  "hospital_name": "All Hospitals",
  "upper_limit_of_confidence_interval": "4.60",
  "comparison_results": "Rate not different than Statewide Rate",
  "facility_id": "0000",
  "lower_limit_of_confidence_interval": "1.01",
  "number_of_cases": "402",
  "expected_mortality_rate": "2.14",
  "number_of_deaths": "8",
  "nys_physician_license_number": "23064",
  "year_of_hospital_discharge": "2010-2012",
  "physician_name": "Filsoufi F",
  "detailed_region": "N/A",
  "observed_mortality_rate": "1.99",
  "risk_adjusted_mortality_rate": "2.33",
  "procedure": "CABG, Valve or Valve/CABG"
},
{
  "region": "N/A",
  "hospital_name": "All Hospitals",
  "upper_limit_of_confidence_interval": "3.79",
  "comparison_results": "Rate not different than Statewide Rate",
  "facility_id": "0000",
  "lower_limit_of_confidence_interval": "0.12",
  "number_of_cases": "265",
  "expected_mortality_rate": "1.05",
  "number_of_deaths": "2",
  "nys_physician_license_number": "23064",
  "year_of_hospital_discharge": "2010-2012",
  "physician_name": "Filsoufi F",
  "detailed_region": "N/A",
  "observed_mortality_rate": "0.75",
  "risk_adjusted_mortality_rate": "1.05",
  "procedure": "CABG"
},
{
  "region": "N/A",
  "hospital_name": "All Hospitals",
  "upper_limit_of_confidence_interval": "4.70",
  "comparison_results": "Rate not different than Statewide Rate",
  "facility_id": "0000",
  "lower_limit_of_confidence_interval": "0.79",
  "number_of_cases": "242",
  "expected_mortality_rate": "2.88",
  "number_of_deaths": "6",
  "nys_physician_license_number": "001395",
  "year_of_hospital_discharge": "2010-2012",
  "physician_name": "Abrol S",
  "procedure": "CABG"
}
```

Question #2

- How many FDA-approved drugs were discontinued that had Methadone as an active ingredient?



The screenshot shows the FDA's Orange Book search page. At the top, the U.S. Department of Health & Human Services logo is on the left, and navigation links for 'A to Z Index', 'Follow FDA', and 'En Español' are on the right. Below this is the FDA logo and the text 'U.S. Food and Drug Administration' with the tagline 'Protecting and Promoting Your Health'. A search bar with a red 'SEARCH' button is also present. A horizontal menu contains links for 'Home', 'Food', 'Drugs', 'Medical Devices', 'Radiation-Emitting Products', 'Vaccines, Blood & Biologics', 'Animal & Veterinary', 'Cosmetics', and 'Tobacco Products'. The main heading is 'Orange Book: Approved Drug Products with Therapeutic Equivalence Evaluations'. Below this are breadcrumb links: 'FDA Home', 'Drug Databases', and 'Orange Book'. A 'Start Over' link is in purple. The 'Search by Active Ingredient:' section has a text input field with the placeholder '(Type in part or all of name)'. Below this, it says 'Select the list you would like to search:' followed by three radio button options: 'Rx (Prescription Drug Products)' (which is selected), 'OTC (Over-the-Counter Drug Products)', and 'Disc (Discontinued Drug Products)'. At the bottom of this section are 'Submit' and 'Clear' buttons. A footer link in purple says 'Return to the Electronic Orange Book Home Page'.

U.S. Department of Health & Human Services

FDA U.S. Food and Drug Administration
Protecting and Promoting *Your* Health

A to Z Index | Follow FDA | En Español

SEARCH

Home Food Drugs Medical Devices Radiation-Emitting Products Vaccines, Blood & Biologics Animal & Veterinary Cosmetics Tobacco Products

Orange Book: Approved Drug Products with Therapeutic Equivalence Evaluations

FDA Home Drug Databases Orange Book

[Start Over](#)

Search by Active Ingredient:

(Type in part or all of name)

Select the list you would like to search:

☒ Rx (Prescription Drug Products)
☐ OTC (Over-the-Counter Drug Products)
☐ Disc (Discontinued Drug Products)

Submit Clear

[Return to the Electronic Orange Book Home Page](#)

Validation of Query Results

```
>>> import re
>>> import requests
>>> formurl = 'http://www.accessdata.fda.gov/scripts/cder/ob/docs/tempai.cfm'
>>> post_params = {'Generic_Name': 'Methadone', 'table1': 'OB_Disc'}
>>> resp = requests.post(formurl, data = post_params)
m = re.search('(?!<=Displaying records) *[\d,]+ *to *[\d,]+ *of *([\d,]+)', resp.text)
print(m.groups()[0])
>>> m = re.search('(?!<=Displaying records) *[\d,]+ *to *[\d,]+ *of *([\d,]+)', resp.text)
>>> print(m.groups()[0])
9
```

Orange Book: Approved Drug Products with Therapeutic Equivalence Evaluations

[FDA Home](#) [Drug Databases](#) [Orange Book](#)

[Start Over](#) | [Back to Search Page](#)

Active Ingredient Search Results from "OB_Disc" table for query on "methadone."

Displaying records 1 to 9 of 9

 [Download data](#)

Appl No	Active Ingredient	Dosage Form; Route	Strength	Proprietary Name	Applicant
N006134	METHADONE HYDROCHLORIDE	SYRUP; ORAL	10MG/30ML	DOLOPHINE HYDROCHLORIDE	ROXANE
N017108	METHADONE HYDROCHLORIDE	TABLET, DISPERSIBLE; ORAL	2.5MG	WESTADONE	SANDOZ
N017108	METHADONE HYDROCHLORIDE	TABLET, EFFERVESCENT; ORAL	10MG	WESTADONE	SANDOZ
N017108	METHADONE HYDROCHLORIDE	TABLET, EFFERVESCENT; ORAL	40MG	WESTADONE	SANDOZ
N017108	METHADONE HYDROCHLORIDE	TABLET, EFFERVESCENT; ORAL	5MG	WESTADONE	SANDOZ
A088109	METHADONE HYDROCHLORIDE	TABLET; ORAL	10MG	METHADONE HYDROCHLORIDE	ROXANE
A074081	METHADONE HYDROCHLORIDE	TABLET; ORAL	40MG	METHADONE HYDROCHLORIDE	ROXANE
A088108	METHADONE HYDROCHLORIDE	TABLET; ORAL	5MG	METHADONE HYDROCHLORIDE	ROXANE
A040241	METHADONE HYDROCHLORIDE	TABLET; ORAL	5MG	METHADONE HYDROCHLORIDE	SANDOZ

To Do

- Identify a data set you wish to use for analysis on the web.
- Determine licensing / IP considerations
- Determine if there is a data dictionary
- Assess data access/structure
- Determine best way to extract data
- Determine what manipulations etc will need to be done to data