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: Unstructred and Semi-structured data

Data in Databses (relational and SQL)

Tuesday, May 24th : 5 to 7pm

Data tells a Story: QA/QC

EDA and Interactive Visaulzation (Buidling 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”)

jupyter

INSTALL PROJECT COMMUNITY DOCUMENTATION NBVIEWER BLOG DONATE

Open source, interactive data science and scientific computing across over 40 programming languages.

https://jupyter.org

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;

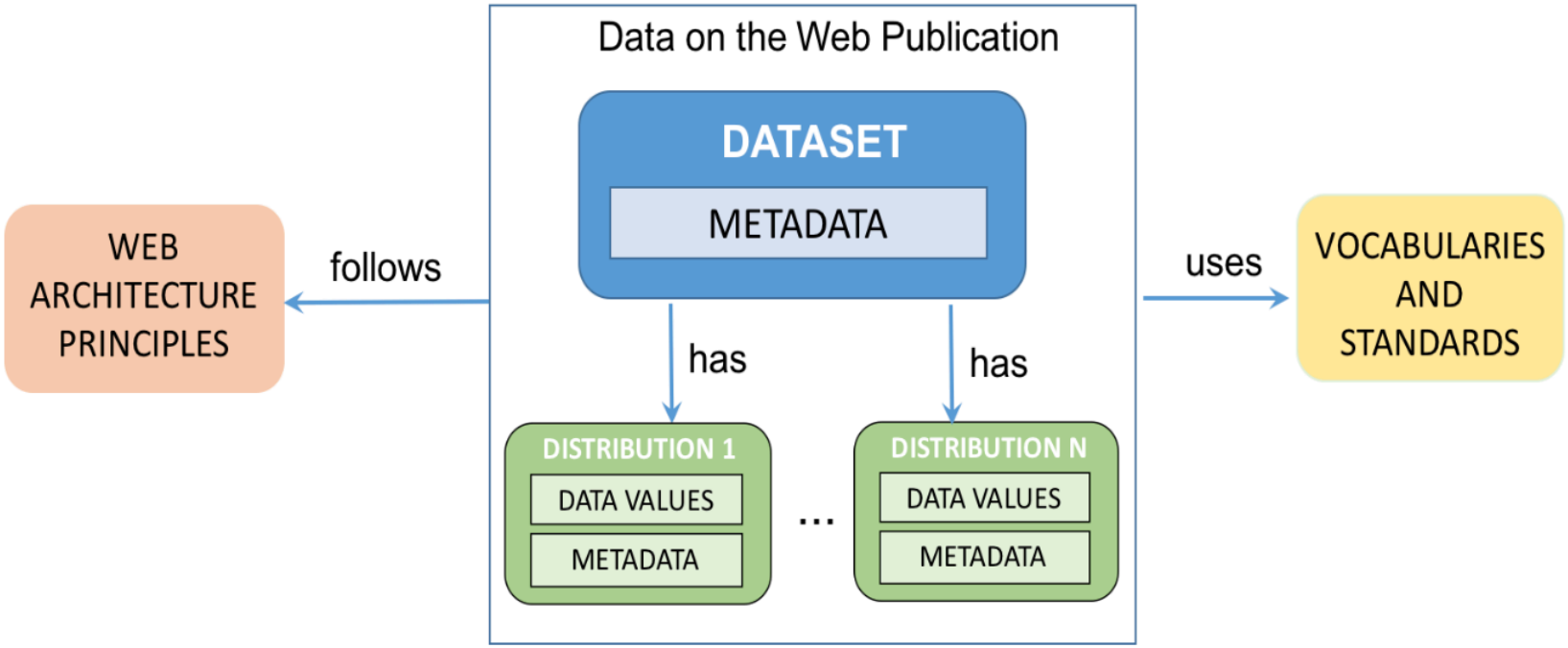
Code to Parse

```
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

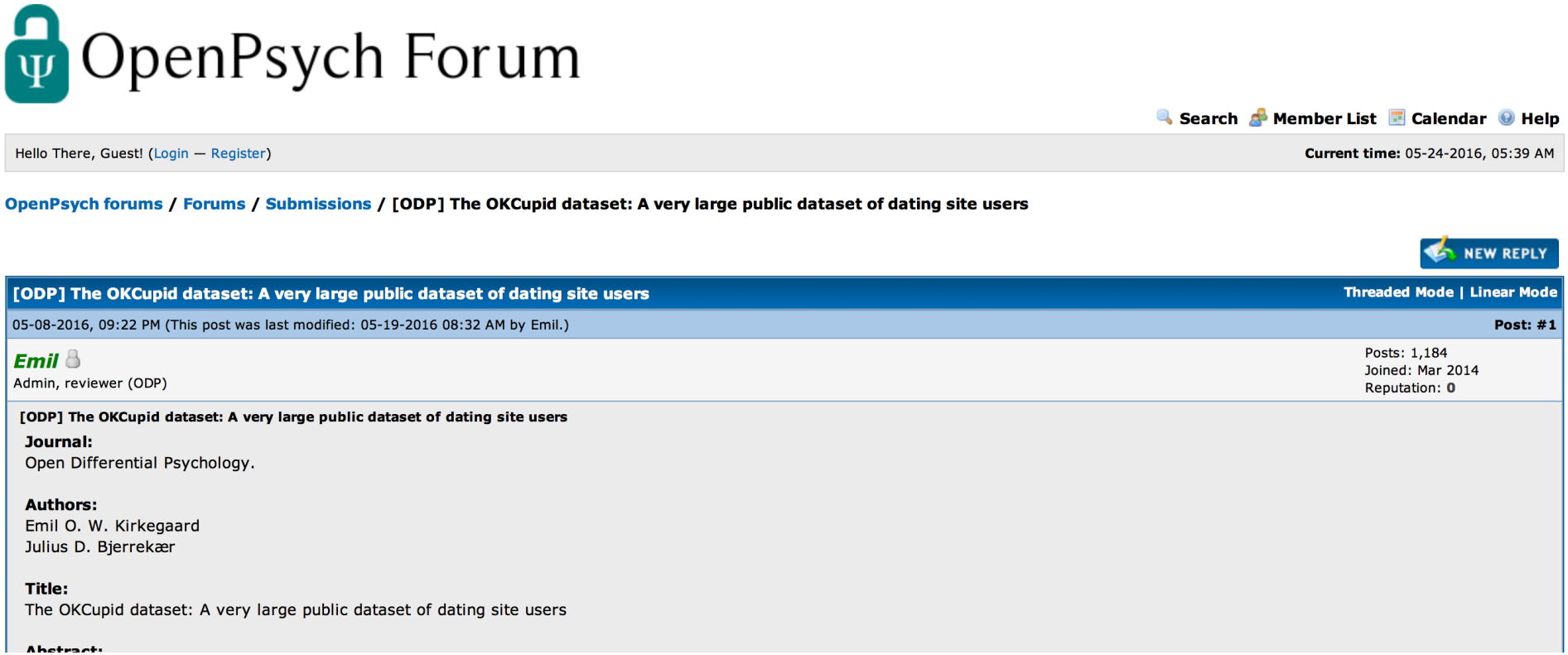


http://www.w3.org/TR/dwbp/

What data is okay to scrape?

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

Public does not equal consent



OpenPsych Forum

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

a fact-based worldview

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

Elements Console Sources Network Timeline Profiles

```
<!DOCTYPE html>
<html lang="en-US">
▶<head>…</head>
▼<body class="wordpress ltr en_US child-theme y2016 m05 d23 h20 monday logged-out home singular singular-page singular-page-1331 page-template-default page-1331 gecko primary-inactive secondary-inactive subsidiary-active">
  ▶<div id="body-container">…</div>
    <!-- .content .hfeed -->
    <!-- <div class="push"></div> -->
    <!-- #container -->
    <!-- #body-container -->
  ▶<div id="footer-container">…</div>
    <!-- #footer-container -->
    <div class="licence-agreementbase" style="display:none;"> </div>
  ▶<div class="licence-agreement" style="display:none;">…</div>
    <script type="text/javascript" src="https://checkout.stripe.com/checkout.js"></script>
    <script type="text/javascript" src="http://www.gapminder.org/wp-content/plugins/stripe-checkout-pro/assets/js/vendor/parsley.min.js?ver=2.3.7"></script>
    <script type="text/javascript" src="http://www.gapminder.org/wp-content/plugins/stripe-checkout-pro/assets/js/vendor/moment.min.js?ver=2.3.7"></script>
    <script type="text/javascript" src="http://www.gapminder.org/wp-content/plugins/stripe-checkout-pro/assets/js/vendor/pikaday.min.js?ver=2.3.7"></script>
    <script type="text/javascript" src="http://www.gapminder.org/wp-content/
```

… body

Styles Event Listeners DOM Breakpoints Properties

Filter .cls

```
element.style {
}
html, body {                                style.css?ver=1.1:133
  height: 100%;
}
body {                                      style.css?ver=1.1:110
  background-color: #fff;
  font-family: 'Open Sans', sans-serif;
  color: #231f20;
}
body {                                      21px.css:29
  line-height: 24px;
  font-family: Georgia, Times, "Times New Roman", serif;
  color: #333;
  background: #fff;
```

margin – border – padding – 867 × 753

Filter Show all

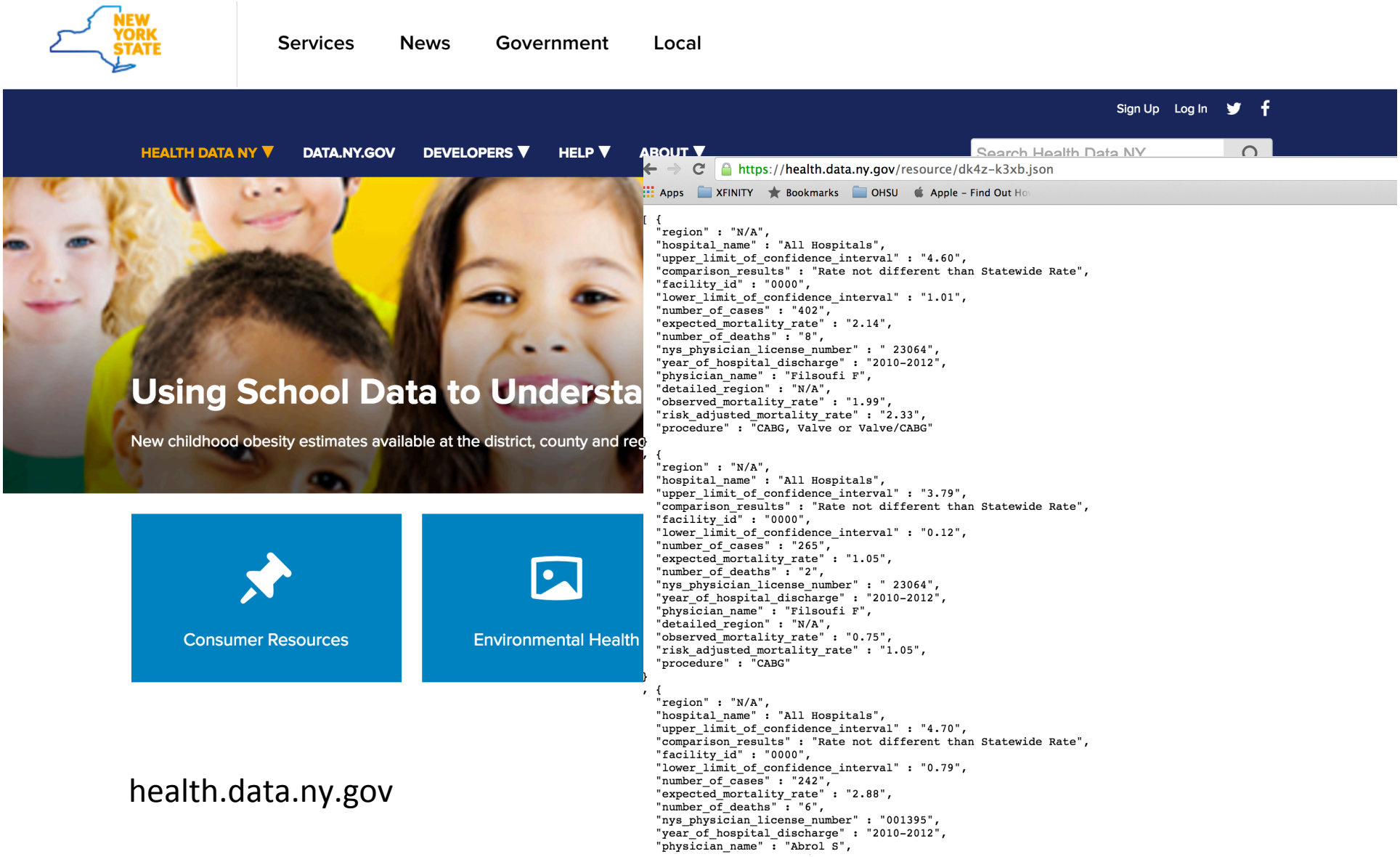
background-attachment scroll

background-clip 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?

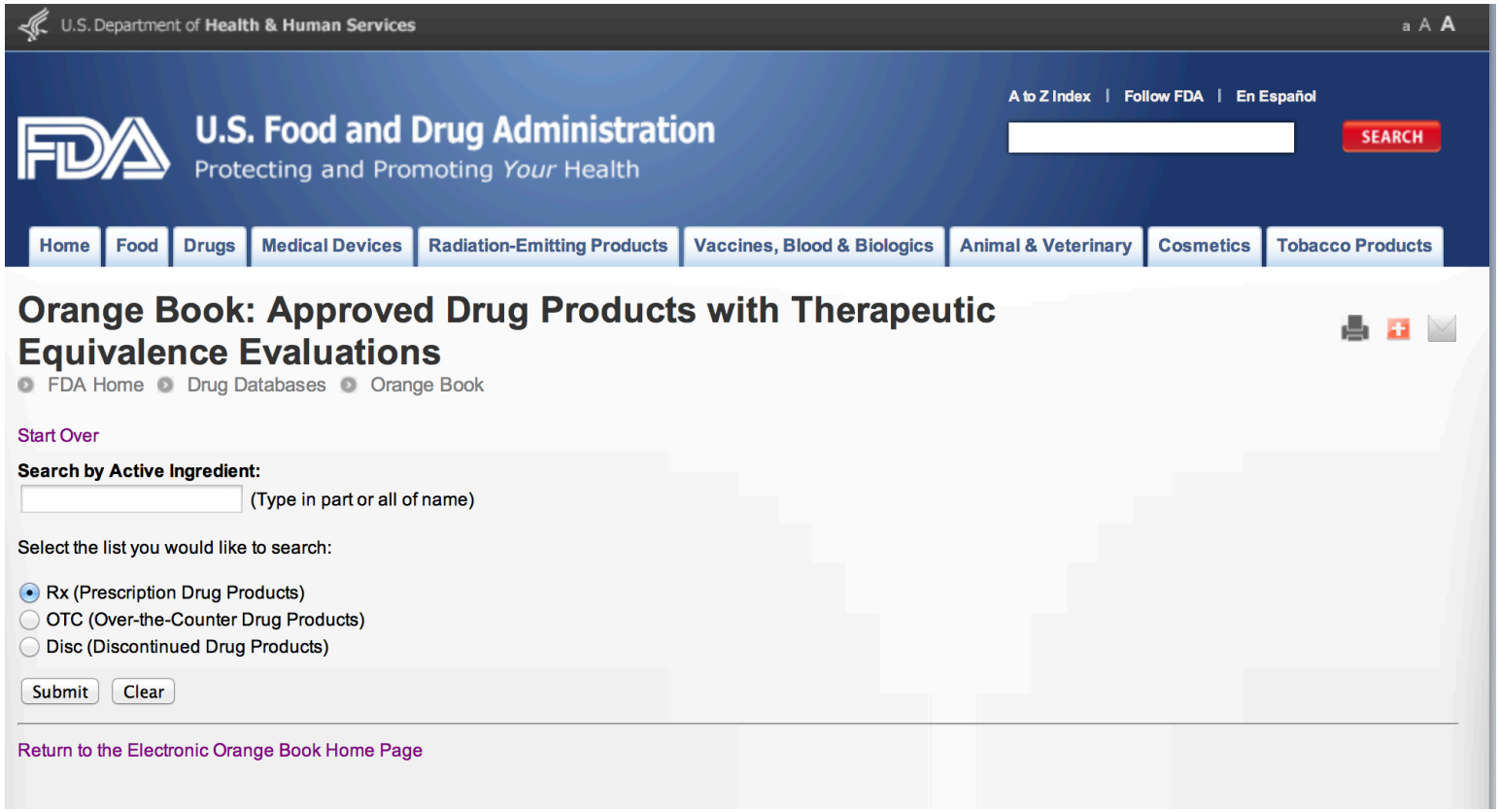
Where do we find the data to answer this?



health.data.ny.gov

Question #2

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



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