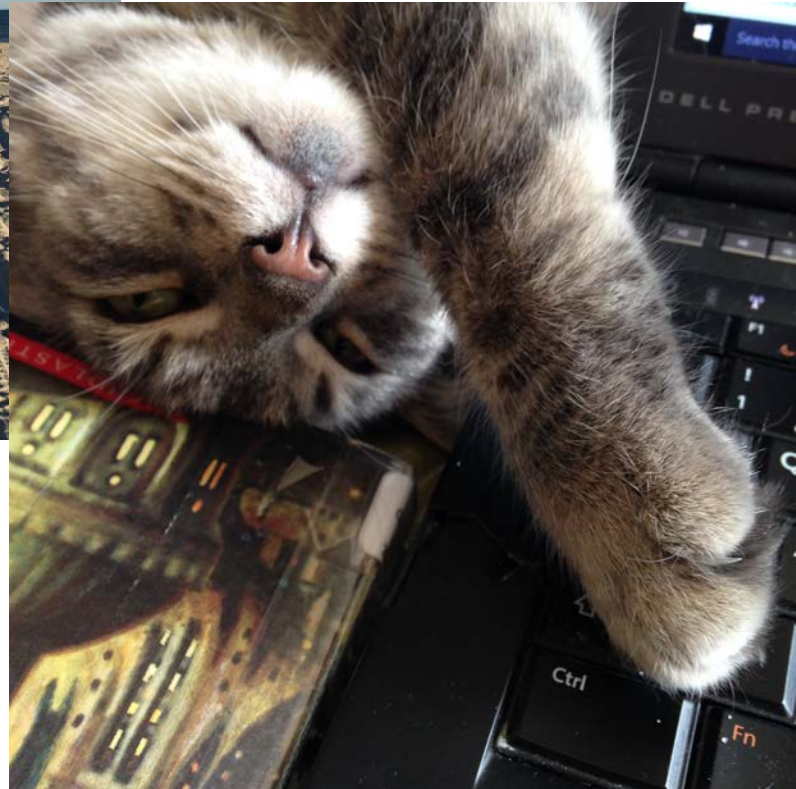


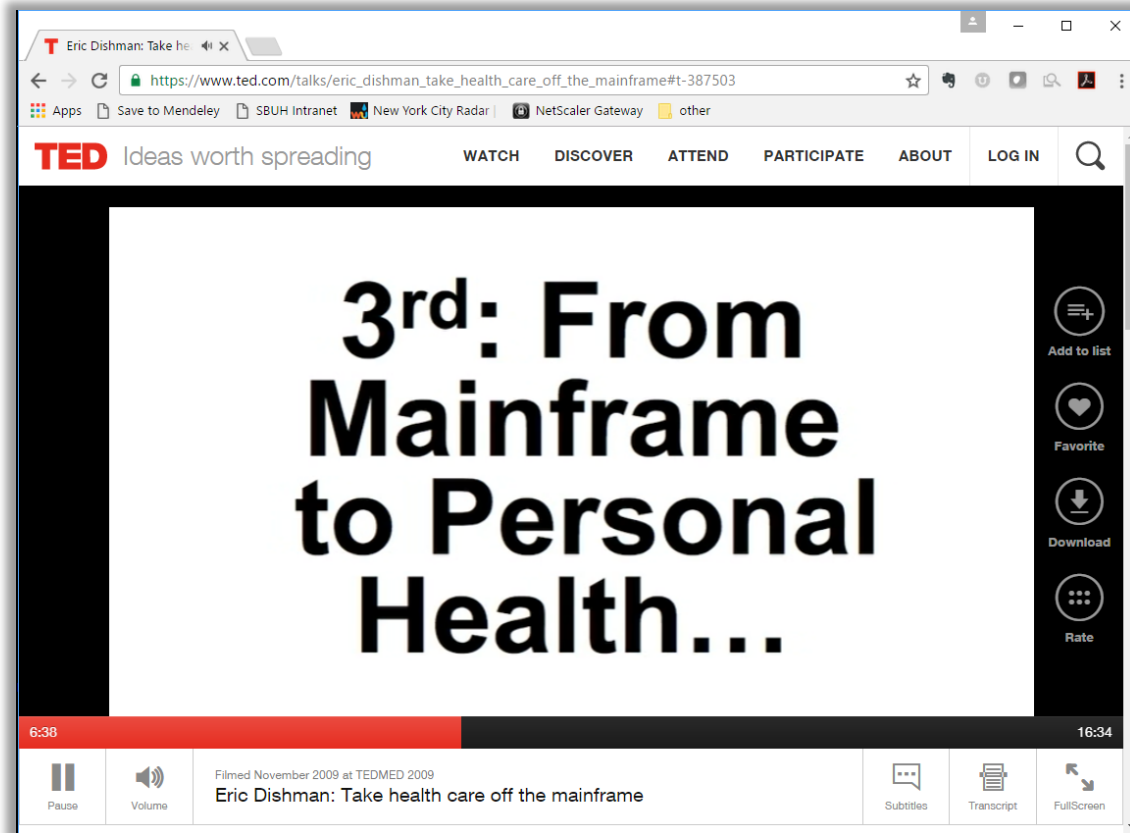
Open Healthcare Data and Tools in Practice

Janos G. Hajagos, PhD
Chief of Data Analytics
Research Assistant Professor
Dept. Biomedical Informatics
Stony Brook University
@jhajagos

AHIMA Data Institute: Making Information Meaningful
Las Vegas, Nevada
12/9/2016



Still Literally True in 2016!



https://www.ted.com/talks/eric_dishman_take_health_care_off_the_mainframe

The examples shown here build on open health data and open source tools



<https://www.flickr.com/photos/122127718@N08/2040238088>

4



https://www.flickr.com/photos/arbre_evolution/8286785236/

The legacy way: analytic software

The screenshot shows a web browser window with the Quora page "How much does SAS cost?". The browser's address bar shows the URL "https://www.quora.com/How-much-does-SAS-cost". The Quora header includes the logo, a search bar, and navigation links like "Ask Question", "Read", "Answer", "Notifications", and a user profile for "Janos".

The main content is an answer to the question. It starts with a paragraph: "if you were using their product for just internal uses. And they charge you a big up front cost followed by about 28% of that amount as an ongoing license fee every subsequent year (it used to be 50%).".

The next paragraph states: "Their basic windows Analytics package costs \$8,700 for the first year. It includes BASE, STAT, and GRAPH products for basic data processing, advanced statistics, and automated production graphics. You will need to pay a chunk extra for modules to access databases directly or using ODBC. See [Order SAS® Software](#)".

The third paragraph says: "For more than one user or versions that run on Linux, Unix (or mainframes) expect to pay more. License fees can easily reach past **\$100,000/year**. Even a single user license for tools like Enterprise Miner can cost something like \$140,000 for the first year."

The final paragraph reads: "Generally you need to speak to someone from SAS to get a quote for more complicated licenses. The SAS sales folks are exceedingly arrogant in many cases and you might find yourself lucky to even get them to spend the time to give you a quote. As a consultant, I tried to get a quote so my client at Morgan Stanley could order the product, and it was a multi-day challenge to get a response. At other times with less impressive sounding clients, I had a hard time even getting a response."

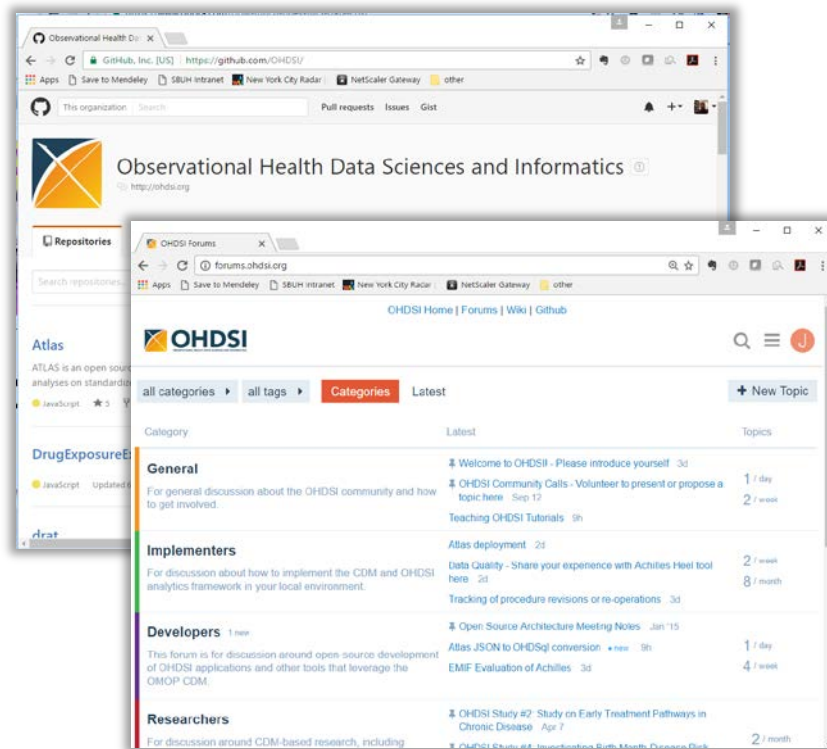
Below the answer is an "EDIT:" label and a row of buttons: "Upvote | 19", "Downvote", and "Comment". To the right of these buttons are social media sharing icons for Facebook, Twitter, and a generic share icon.

On the right side of the page, there is a sidebar with "More Related Questions" including "How much does Cloudera cost?" and "How much does Palantir cost?". Below that is a "Question Stats" section showing "25 Followers", "52,964 Views", "Last Asked Jul 5, 2013", and an "Edits" link.

The open way: software

- Builds on long term investment of open source tools
 - BSD, GNU, Linux kernel, R, Python, LAPACK
- International community of developers from commercial, academic, and government stakeholders
- Collaborative internet tools are used to coordinate development
 - Git and Github
- Source code is made available

Example: OHDSI Software Stack



<http://www.ohdsi.org/>





The legacy way: healthcare data

The screenshot shows a web browser window with the Quora website. The address bar displays the URL <https://www.quora.com/How-much-does-IMS-data-cost>. The Quora header includes the logo, a search bar, and navigation links for 'Ask Question', 'Read', 'Answer', 'Notifications', and a user profile for 'Janos'. The main content area features two answers to the question 'How much does IMS data cost?'. The first answer is by Mike Hamilton, a former chemist turned biotech consultant, dated March 15. It states that the cost varies based on the level of detail and the number of physicians, with a specific example of exceeding \$500,000 for a therapeutic area. The second answer is by Siddharth Mishra, who used to work at IMS Health, dated July 4. He explains that the cost is not fixed and depends on the type of data (e.g., Sell-in, MIDAS, PADS) and the company requesting it, ranging from 10,000 dollars to 4 million dollars. Both answers include upvote/downvote buttons and social media sharing options. A 'Question Stats' sidebar on the right shows 39 followers, 6,705 views, and the question was last asked on April 29.

Quora Ask or Search Quora Ask Question Read Answer Notifications Janos

Mike Hamilton, Former chemist turned biotech consultant
Written Mar 15

It's not cheap, but it all depends on what you want. If you just want high level sales data, it might only cost a few tens of thousands if it's for a single product. If you want detailed prescribing data for all available physicians, i've seen the cost exceed **\$500,000** for a therapeutic area.

1.4k Views

Upvote Downvote Comment

Siddharth Mishra, used to work at IMS Health
Written Jul 4

thanks for A2A.

Its not fixed, depends on the data you want like you are looking for Sell- in data, sell out data, MIDAS, PADS or data view and also on the company requesting. For eg. a large company like Novartis may be able to leverage its position better and get good rates from IMS for a selected data type while a smaller comapany may have to pay what IMS quotes to them. It also depends on the quotations sent by IMS and rates are not standardized as far as I know.

But the range can be (which i have experineced) anywhere from 10000 dollars for a simple data shared every month to 4 million dollars for a complete dedicated data centre.

650 Views · Answer requested by Yuki Ishii

Upvote Downvote Comment

Question Stats

- 39 Followers
- 6,705 Views
- Last Asked Apr 29
- Edits

The open way: Data portal

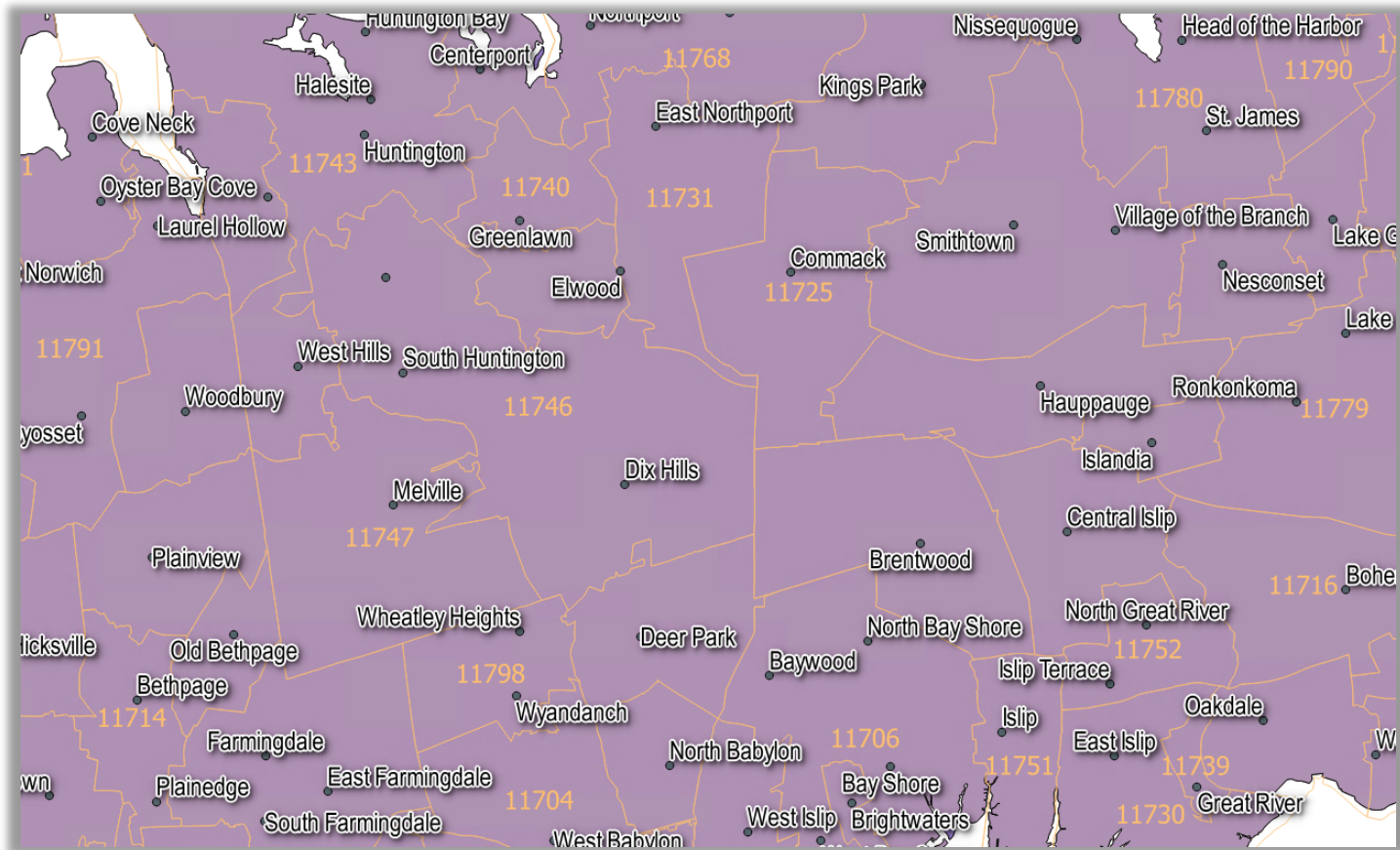


Example 1: Enriching your local patient data

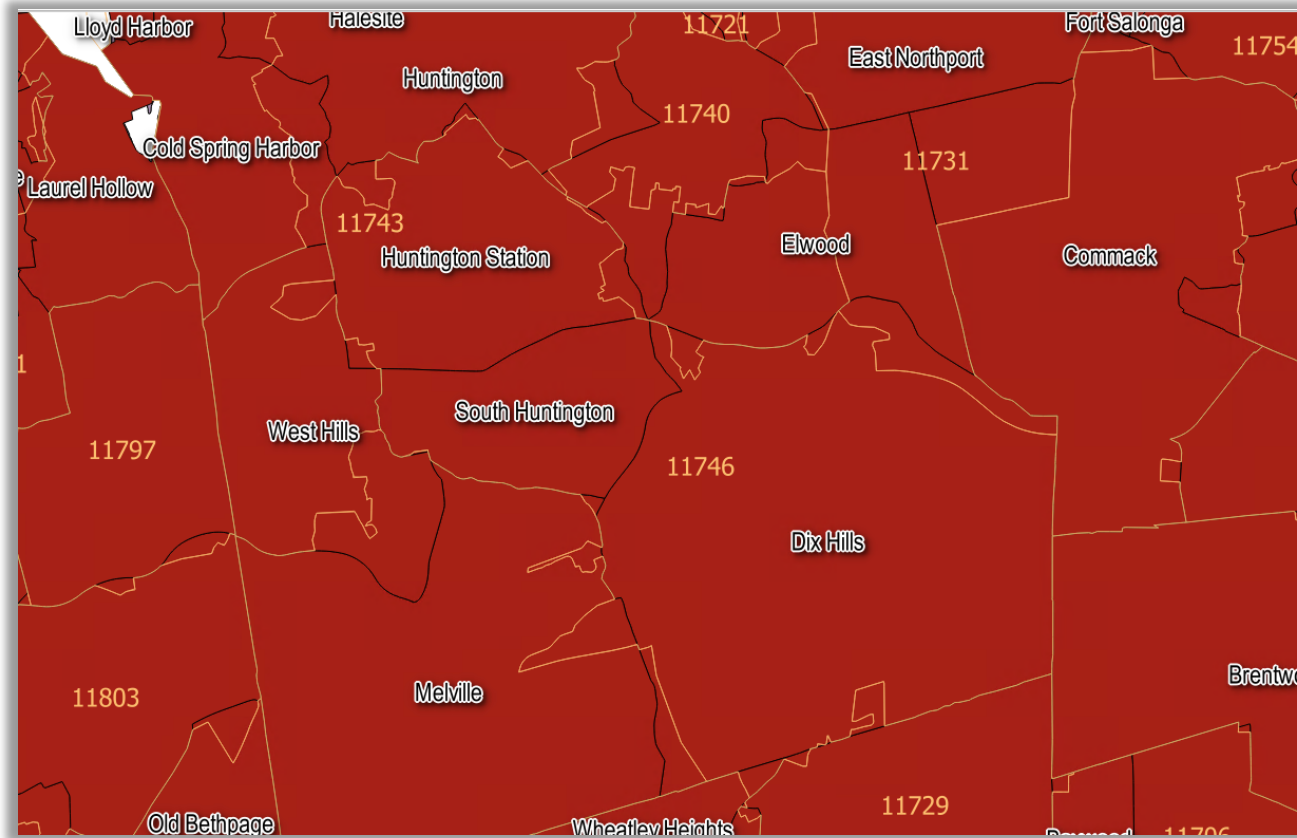
| patient_id | Address |
|------------|--|
| 1001 | 100 Mains Street, Springfield, MA 01103 |
| 1022 | 12 Oak Drive, Springfield, MA 01105 |
| 3033 | 1001 East Main St., Greenfield, MA 01301 |
| 4010 | 101 Route 9a, Deerfield, MA 01342 |

Going beyond the zip code

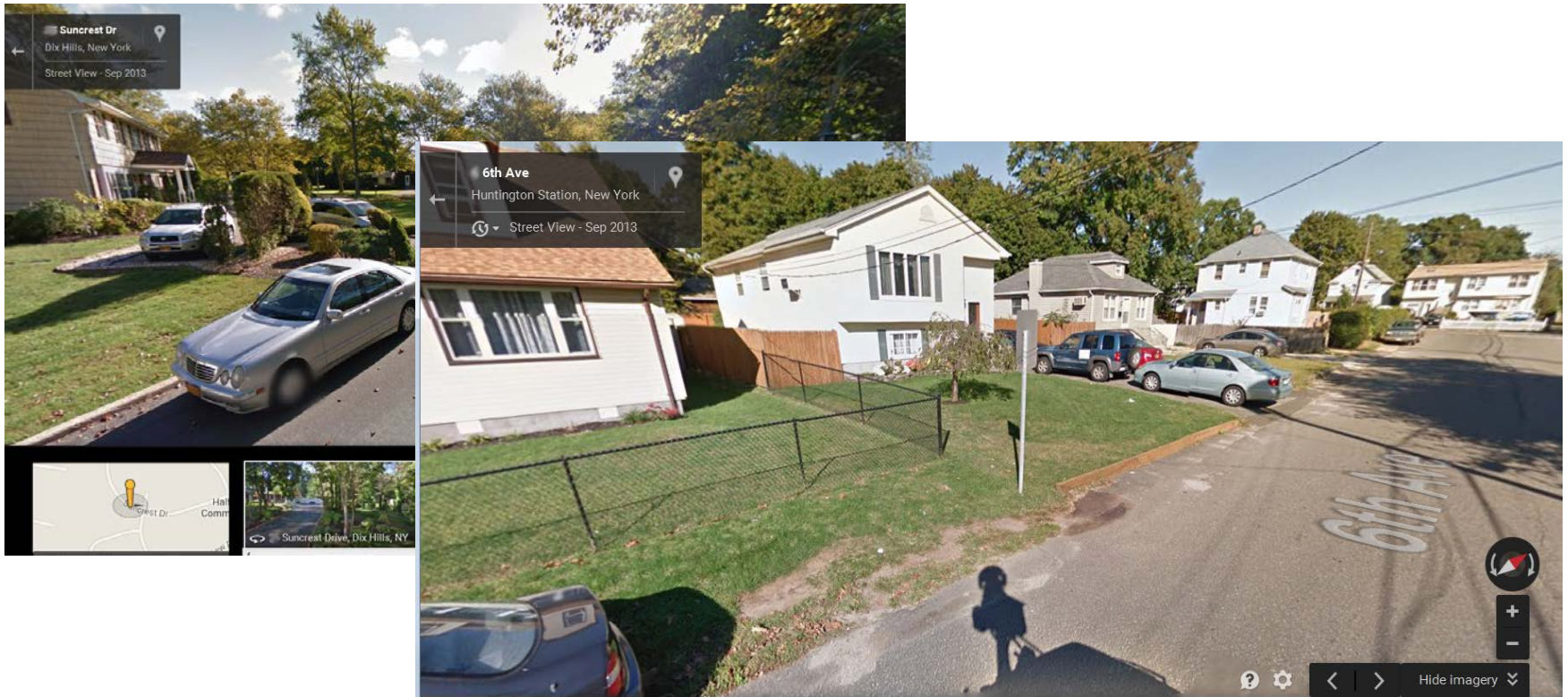
11746 – Huntington Station and Dix Hills



Two Census-Designated Places



Understanding socio-economic determinants health for your patients



Google Map's Street View

American FactFinder - Search Results

https://factfinder.census.gov/faces/nav/jsf/pages/searchresults.xhtml?refresh=t

United States Census Bureau

AMERICAN FactFinder

Feedback FA

MAIN COMMUNITY FACTS GUIDED SEARCH **ADVANCED SEARCH** DOWNLOAD CENTER

Search - Use the options on the left (topics, geographies, ...) to narrow your search results

Your Selections

'Your Selections' is empty

load search | save search

Search using the options below:

Topics
(age, income, year, dataset, ...)

Geographies
(states, counties, places, ...)

Race and Ethnic Groups
(race, ancestry, tribe)

Industry Codes
(NAICS industry, ...)

EEO Occupation Codes
(executives, analysts, ...)

To search for tables and other files in American FactFinder:

Enter search terms and an optional geography and click **GO**

Select Topics

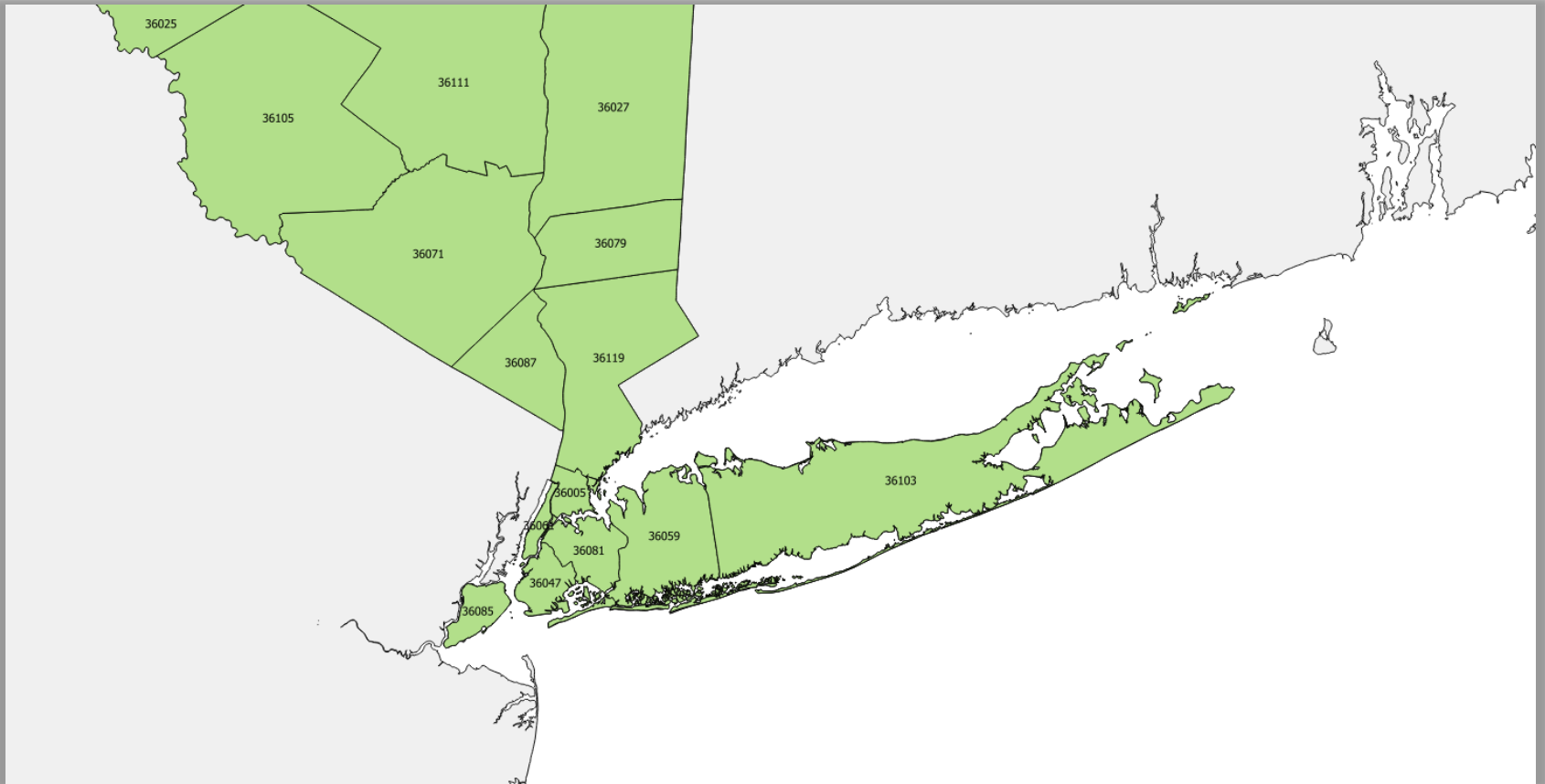
Select Topics to add to 'Your Selections'

- People
 - Basic Count/Estimate
 - Age & Sex
 - Age Group
 - Disability
 - Education
 - Employment
 - Income & Earnings
 - Insurance Coverage
 - Language
 - Marital & Fertility Status
 - Origins
 - Population Change
- Poverty
 - Food Stamps/SNAP (759)
 - Heating and Cooling Assistance (320)
 - Poverty (5,326)
- Relationship
- Veterans

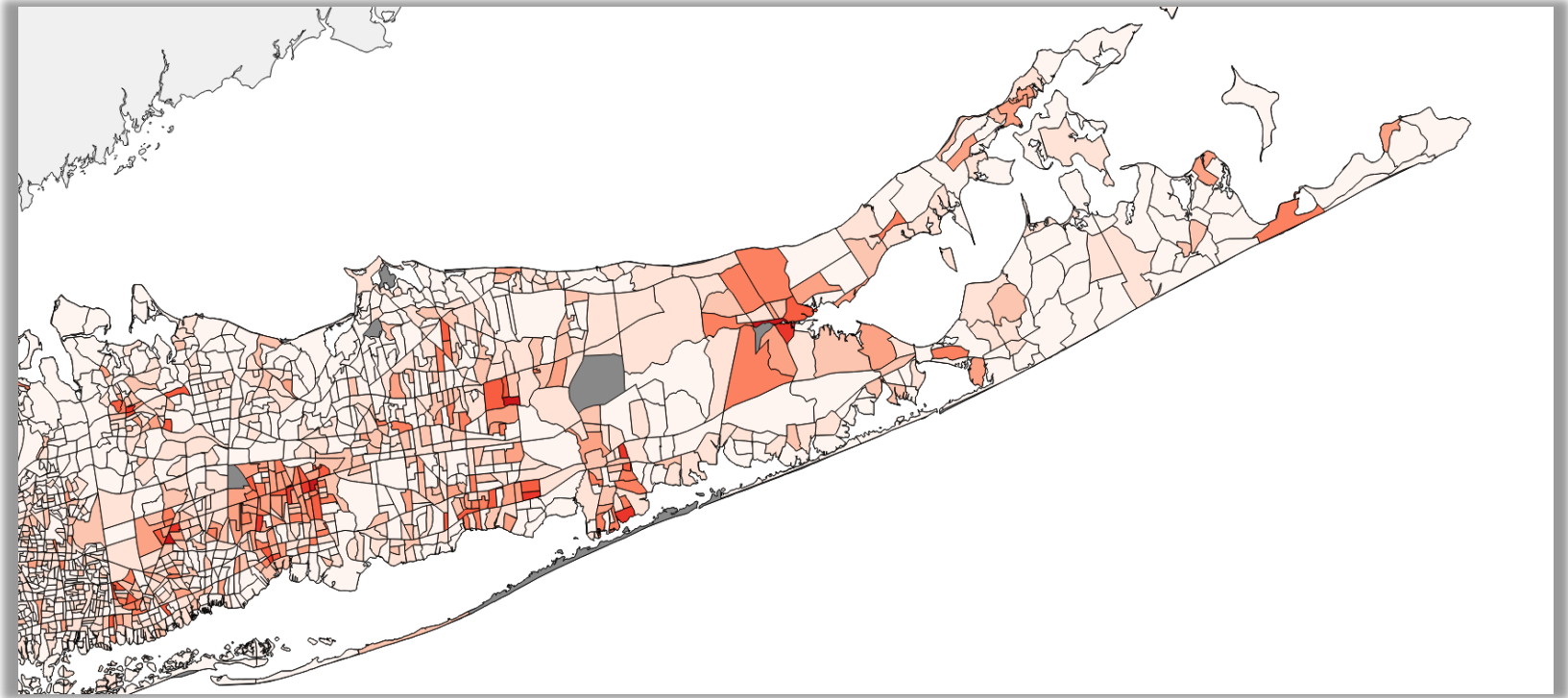
Note: The **Race & Ethnicity** topic is available under the **Race and Ethnic**

<https://factfinder.census.gov/>

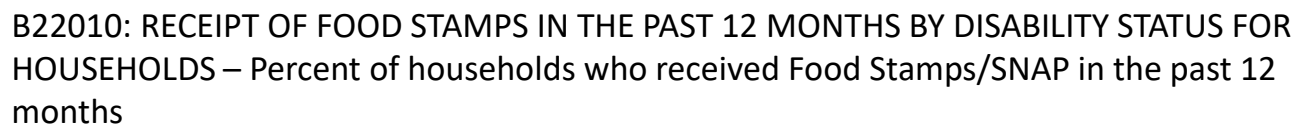
AT THE BLOCK GROUP LEVEL



SNAP By Geographic Region



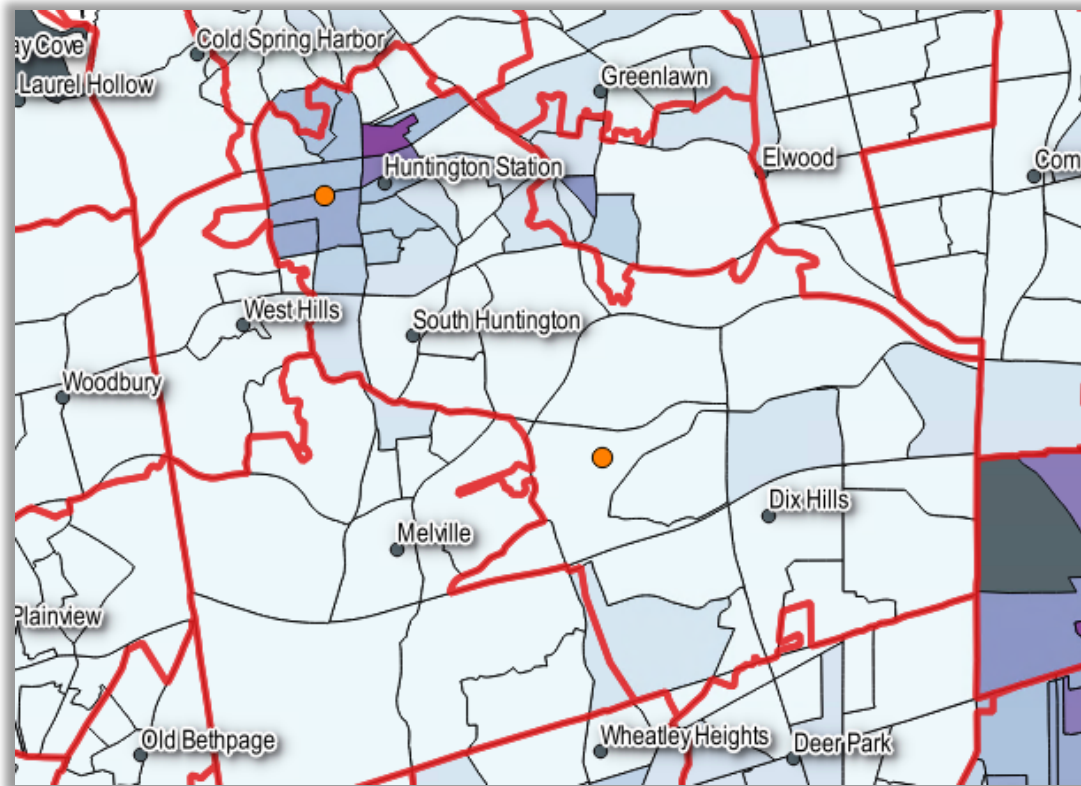
B22010: RECEIPT OF FOOD STAMPS IN THE PAST 12 MONTHS BY DISABILITY STATUS FOR HOUSEHOLDS



Using the PostGis Tiger Geocoder in PostgreSQL

```
SELECT (tt.geo).geomout, (tt.geo).rating,  
       ST_Y((tt.geo).geomout) as latitude,  
       ST_X((tt.geo).geomout) as longitude,  
       tiger.pprint_addy((tt.geo).addy) as  
matched_address, (tt.geo).addy.zip as  
matched_zip5  
  
FROM (select tiger.Geocode('?? Suncrest  
Dr., Dix Hills, NY 11746', 1) as geo) tt;
```

Geocoding Results



| | geomout | rat... | latitude | longitude | matched_address | matched_zip5 |
|---|--|--------|----------|-----------|---------------------------------------|--------------|
| 0 | 0101000020AD100000E5744739715752C06804F93EFA664440 | 0 | 40.80451 | -73.366 | Suncrest Dr, Dix Hills, NY 11746 | 11746 |
| 0 | 0101000020AD1000007F87CC2FCC5A52C0A2F984B9E46B4440 | 0 | 40.84292 | -73.418 | 6th Ave, Huntington Station, NY 11746 | 11746 |

Example 1: Data sources and tools

- PostgreSQL - <https://www.postgresql.org/>
- PostGIS and Tiger Geocoder - <http://www.postgis.net/>
- Shape files - <https://www.census.gov/geo/maps-data/data/tiger-line.html>
- Raw ACS files - <http://www.census.gov/programs-surveys/acs/data/data-via-ftp.html>
- Python Code for extracting ACS variables - <https://github.com/jhajagos/CensusGeographyTools>
- QGIS – open source full featured GIS - <http://www.qgis.org/>

Example 2 - Hospital market analysis – Bakersfield, CA



<https://www.flickr.com/photos/27326512@N00/329820320/>



<https://www.flickr.com/photos/rheinitz/8668769226/>

Medicare Teaming Data

The screenshot shows the CMS.gov website with the URL <https://questions.cms.gov/faq.php?faqId=7977>. The page is titled "Frequently Asked Questions" and includes a search bar, navigation links, and a list of frequently asked questions. The selected question is "What physician shared patient data sets are available?".

What physician shared patient data sets are available?

The physician referral data linked below was provided as a response to a Freedom of Information Act (FOIA) request. These files represent the number of encounters a single beneficiary has had across physicians at intervals of 30, 60, 90 and 180 days. For more details about the file contents for years 2009 - 2015, please see the Technical Requirements document:

http://downloads.cms.gov/foia/Physician_Shared_Patient_Patterns_Technical_Requirements.pdf.

Note: The files range in size from 1-7 gigabytes. Special statistical software is needed for analysis. These files will not fully open in a text editor or Microsoft Excel.

| | | | | |
|--|---------------------------------|---------------------------------|---------------------------------|----------------------------------|
| Physician Shared Patient Patterns - 2009 | 30 day interval | 60 day interval | 90 day interval | 180 day interval |
| Physician Shared Patient Patterns - 2010 | 30 day interval | 60 day interval | 90 day interval | 180 day interval |
| Physician Shared Patient Patterns - 2011 | 30 day interval | 60 day interval | 90 day interval | 180 day interval |
| Physician Shared Patient Patterns - 2012 | 30 day interval | 60 day interval | 90 day interval | 180 day interval |
| Physician Shared Patient Patterns - 2013 | 30 day interval | 60 day interval | 90 day interval | 180 day interval |
| Physician Shared Patient Patterns - 2014 | 30 day interval | 60 day interval | 90 day interval | 180 day interval |
| Physician Shared Patient Patterns - 2015 | 30 day interval | 60 day interval | 90 day interval | 180 day interval |

(FAQ7977)

<https://questions.cms.gov/faq.php?faqId=7977>

For “Big Open Data”* we don’t need to invest in an expensive HIPAA compliant environment

The image shows a screenshot of an email from Amazon Web Services and a terminal window. The email is titled "Amazon Web Services Billing Statement Available" and is addressed to "Inbox". It contains the following text:

Amazon Web Services <no-reply-aw... 4:23 PM (3 hours ago) to me

Greetings from Amazon Web Services,

This e-mail confirms that your latest billing statement, for the account ****2689, is available on the AWS web site. Your account will be charged the following:

Total: \$17.71

You can see a complete break down of all charges on the Billing & Management page located here:

<https://console.aws.amazon.com/billing/home#/bill?statementTimePeriod=1477958400>

To protect your privacy, we can only communicate account information to the e-mail address on file for your account.

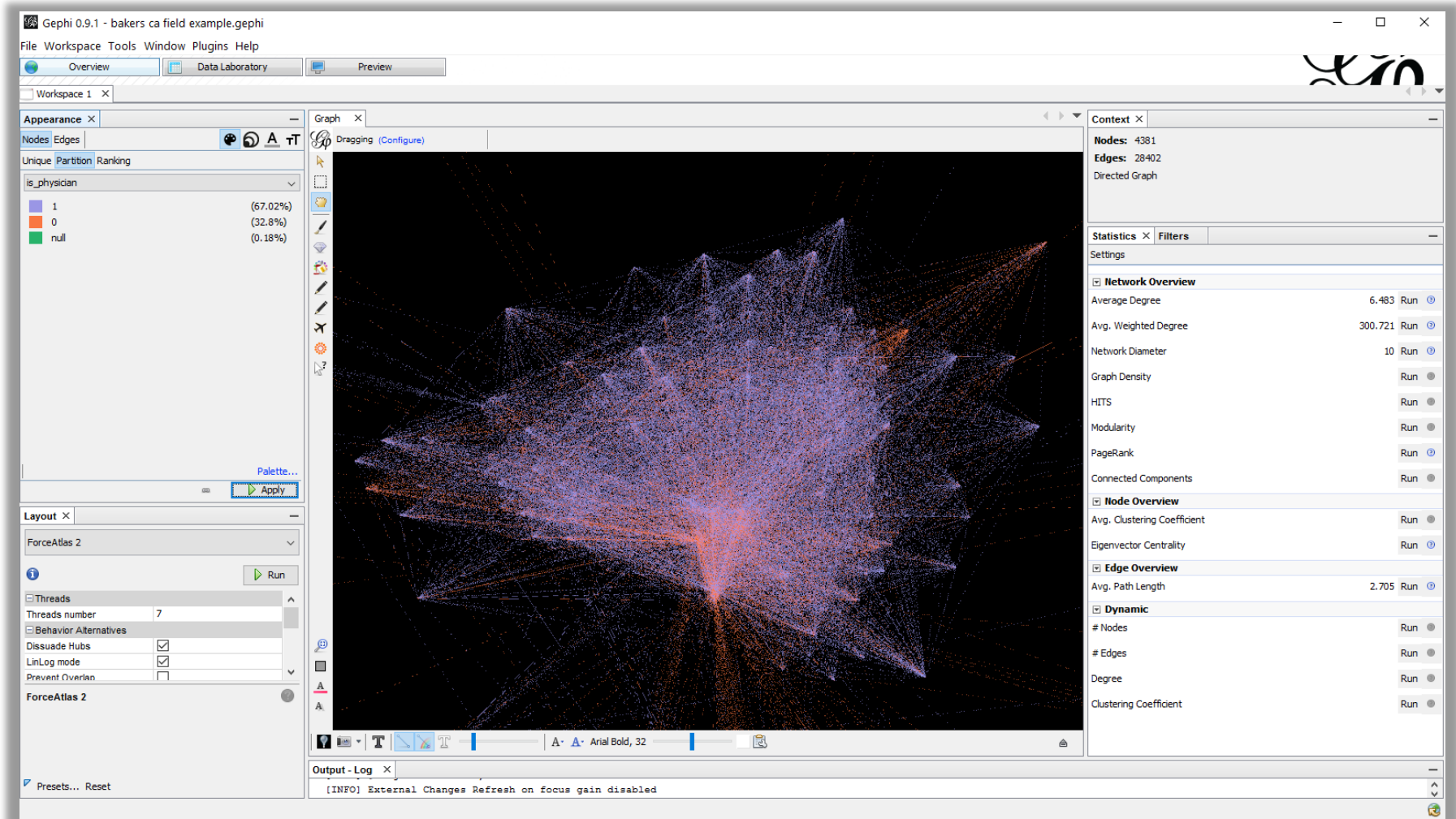
Thank you for using Amazon Web Services.

The terminal window shows the following commands and output:

```
janos — ubuntu@ip-172-30-4-217: /data/flat — ssh — 80x24
ubuntu@ip-172-30-4-217:/data/flat$ ls -sh *.txt *.csv
5.7G npidata_20050523-20161009.csv
12K npidata_20050523-20161009FileHeader.csv
432K nucc_taxonomy_161.csv
2.9G physician-shared-patient-patterns-2015-days180.txt
1.6G physician-shared-patient-patterns-2015-days30.txt
2.2G physician-shared-patient-patterns-2015-days60.txt
2.5G physician-shared-patient-patterns-2015-days90.txt
ubuntu@ip-172-30-4-217:/data/flat$ df -h
Filesystem      Size  Used Avail Use% Mounted on
udev            7.9G   12K   7.9G   1% /dev
tmpfs           1.6G  344K   1.6G   1% /run
/dev/xvda1      7.8G   1.4G   6.0G  19% /
none            4.0K     0   4.0K   0% /sys/fs/cgroup
none           5.0M     0   5.0M   0% /run/lock
none           7.9G     0   7.9G   0% /run/shm
none          100M     0  100M   0% /run/user
/dev/xvdb       148G   76G   65G  54% /data
ubuntu@ip-172-30-4-217:/data/flat$ uname -a
Linux ip-172-30-4-217 3.13.0-91-generic #138-Ubuntu SMP Fri Jun 24 17:00:34 UTC 2016 x86_64 x86_64 x86_64 GNU/Linux
ubuntu@ip-172-30-4-217:/data/flat$
```

*Bigger than supported by a standard business class laptop or desktop or Excel

Gephi Visualization of the teaming data



Eigenvector centrality

- Centrality measures the importance of the node in the network
- Ranks importance of a node (provider) in the network
- Google's PageRank is a variant of this metric
- A higher page rank in search indicates a more relevant search but in a teaming network does not imply that the physician is of high clinical quality

Gephi 0.9.1 - Project 1

FileWorkspaceToolsWindowHelp

OverviewData LaboratoryPreview

Workspace 1Workspace 2

Data Table

NodesEdgesConfigurationAdd nodeAdd edgeSearch/ReplaceImport SpreadsheetExport tableMore actions

Filter:is_family_medicine

| Id | Label | credential | taxonomy_code | taxonomy_name | node_type | is_physician | is_family_medicine | is_internal_medicine | city | specialization | Eigenvector Centrality |
|------------|------------------------|------------|---------------|---|-----------|--------------|--------------------|----------------------|-------------|---------------------------|------------------------|
| 1558372896 | NICHOLS, EDWARD MD | MD | 207R00000X | Allopathic & Osteopathic Physicians - Internal Medicine | leaf | 1 | 0 | 1 | BAKERSFIELD | | 0.662349 |
| 1043290778 | AGGARWAL, ATUL MD | MD | 207R10011X | Allopathic & Osteopathic Physicians - Internal Medicine - Interv... | leaf | 1 | 0 | 1 | BAKERSFIELD | Interventional Cardiol... | 0.568149 |
| 1588619464 | ALIM, MUHAMMAD MD | MD | 207R00000X | Allopathic & Osteopathic Physicians - Internal Medicine | leaf | 1 | 0 | 1 | BAKERSFIELD | | 0.566924 |
| 1134195647 | SINGH, SARABJIT MD | MD | 207RC0000X | Allopathic & Osteopathic Physicians - Internal Medicine - Card... | leaf | 1 | 0 | 1 | BAKERSFIELD | Cardiovascular Disease | 0.563766 |
| 1649382920 | BAER, HAROLD MD FA... | MD FACP | 207RN0300X | Allopathic & Osteopathic Physicians - Internal Medicine - Neph... | leaf | 1 | 0 | 1 | BAKERSFIELD | Nephrology | 0.53562 |
| 1700088036 | SINGH, MATAB MD | MD | 207R00000X | Allopathic & Osteopathic Physicians - Internal Medicine | leaf | 1 | 0 | 1 | BAKERSFIELD | | 0.534222 |
| 1073541512 | ALAM, SYED MD | MD | 207RP1001X | Allopathic & Osteopathic Physicians - Internal Medicine - Pulmo... | leaf | 1 | 0 | 1 | BAKERSFIELD | Pulmonary Disease | 0.518949 |
| 1104941467 | WAHL, DIPPAK MD | MD | 207R00000X | Allopathic & Osteopathic Physicians - Internal Medicine | leaf | 1 | 0 | 1 | BAKERSFIELD | | 0.512479 |
| 1285 | | | | | | | | | | Cardiovascular Disease | 0.511712 |
| 1902 | | | | | | | | | | | 0.49374 |
| 1720 | | | | | | | | | | Nephrology | 0.481389 |
| 1033 | | | | | | | | | | | 0.480749 |
| 1015 | | | | | | | | | | Infectious Disease | 0.480135 |
| 1730 | | | | | | | | | | Interventional Cardiol... | 0.475924 |
| 1720 | | | | | | | | | | | 0.471769 |
| 1952 | | | | | | | | | | Nephrology | 0.470468 |
| 1538 | | | | | | | | | | | 0.470364 |
| 1942 | | | | | | | | | | Nuclear Radiology | 0.465012 |
| 1114 | | | | | | | | | | Critical Care Medicine | 0.451344 |
| 1356304539 | NAIR, SHYAM MD | MD | 174400000X | Other Service Providers - Specialist | leaf | 1 | 0 | 1 | BAKERSFIELD | | 0.4312 |
| 1629077755 | VARANASI, UMAHAHE... | MD | 207RN0300X | Allopathic & Osteopathic Physicians - Internal Medicine - Neph... | leaf | 1 | 0 | 1 | BAKERSFIELD | Nephrology | 0.428627 |
| 1174592547 | KHAN, NASSER MD | MD | 207RC0000X | Allopathic & Osteopathic Physicians - Internal Medicine - Card... | leaf | 1 | 0 | 1 | BAKERSFIELD | Cardiovascular Disease | 0.428519 |
| 1295919371 | SINGH, SARABJEET MD | MD | 207R10011X | Allopathic & Osteopathic Physicians - Internal Medicine - Interv... | leaf | 1 | 0 | 1 | BAKERSFIELD | Interventional Cardiol... | 0.426309 |
| 1548209711 | LAUGHLIN, ROBERT MD | MD | 207R00000X | Allopathic & Osteopathic Physicians - Internal Medicine | leaf | 1 | 0 | 1 | BAKERSFIELD | | 0.418111 |
| 1275580367 | MUBIN, TARIQ MD | MD | 207RN0300X | Allopathic & Osteopathic Physicians - Internal Medicine - Neph... | leaf | 1 | 0 | 1 | BAKERSFIELD | Nephrology | 0.417172 |
| 1891805610 | ALVAREZ, CARLOS MD | MD | 207R00000X | Allopathic & Osteopathic Physicians - Internal Medicine | leaf | 1 | 0 | 1 | BAKERSFIELD | | 0.411133 |
| 1316974694 | HASTA, FAKH-RUDDIN ... | MD | 207R00000X | Allopathic & Osteopathic Physicians - Internal Medicine | leaf | 1 | 0 | 1 | BAKERSFIELD | | 0.408892 |
| 1114937190 | SHUKLA, PANKAJ MD | MD | 207R00000X | Allopathic & Osteopathic Physicians - Internal Medicine | leaf | 1 | 0 | 1 | BAKERSFIELD | | 0.400868 |
| 1790911352 | DESAI, NIRAV MD | MD | 207RC0000X | Allopathic & Osteopathic Physicians - Internal Medicine - Card... | leaf | 1 | 0 | 1 | BAKERSFIELD | Cardiovascular Disease | 0.381854 |
| 1487690830 | NARANG, YADVINDER... | MD FACC | 207RC0000X | Allopathic & Osteopathic Physicians - Internal Medicine - Card... | leaf | 1 | 0 | 1 | BAKERSFIELD | Cardiovascular Disease | 0.379136 |
| 1659340339 | AZIZ, HANY MD | MD | 207RC0200X | Allopathic & Osteopathic Physicians - Internal Medicine - Critic... | leaf | 1 | 0 | 1 | BAKERSFIELD | Critical Care Medicine | 0.370644 |
| 1518290493 | SHAH, RUSHABH MD | MD | 208M00000X | Allopathic & Osteopathic Physicians - Hospitalist | leaf | 1 | 0 | 1 | BAKERSFIELD | | 0.368376 |
| 1245320563 | MATUK, ROBIN MD | MD | 207RG0100X | Allopathic & Osteopathic Physicians - Internal Medicine - Gastr... | leaf | 1 | 0 | 1 | BAKERSFIELD | Gastroenterology | 0.367203 |
| 1790855195 | BHAMBHI, BRIJESH MD | MD | 207RC0000X | Allopathic & Osteopathic Physicians - Internal Medicine - Card... | leaf | 1 | 0 | 1 | BAKERSFIELD | Cardiovascular Disease | 0.3588 |
| 1841448859 | SANDHU, SUK-DEEP MD | MD | 207R00000X | Allopathic & Osteopathic Physicians - Internal Medicine | leaf | 1 | 0 | 1 | BAKERSFIELD | | 0.355995 |
| 1053363663 | BRAR, HARJEET MD | MD | 282N00000X | Hospitals - General Acute Care Hospital | leaf | 1 | 0 | 1 | BAKERSFIELD | | 0.35204 |

Add columnMerge columnsDelete columnClear columnCopy data to other columnFill column with a valueDuplicate column

Create a boolean column from regex matchCreate column with list of regex matching groups

Negate boolean valuesConvert column to dynamic

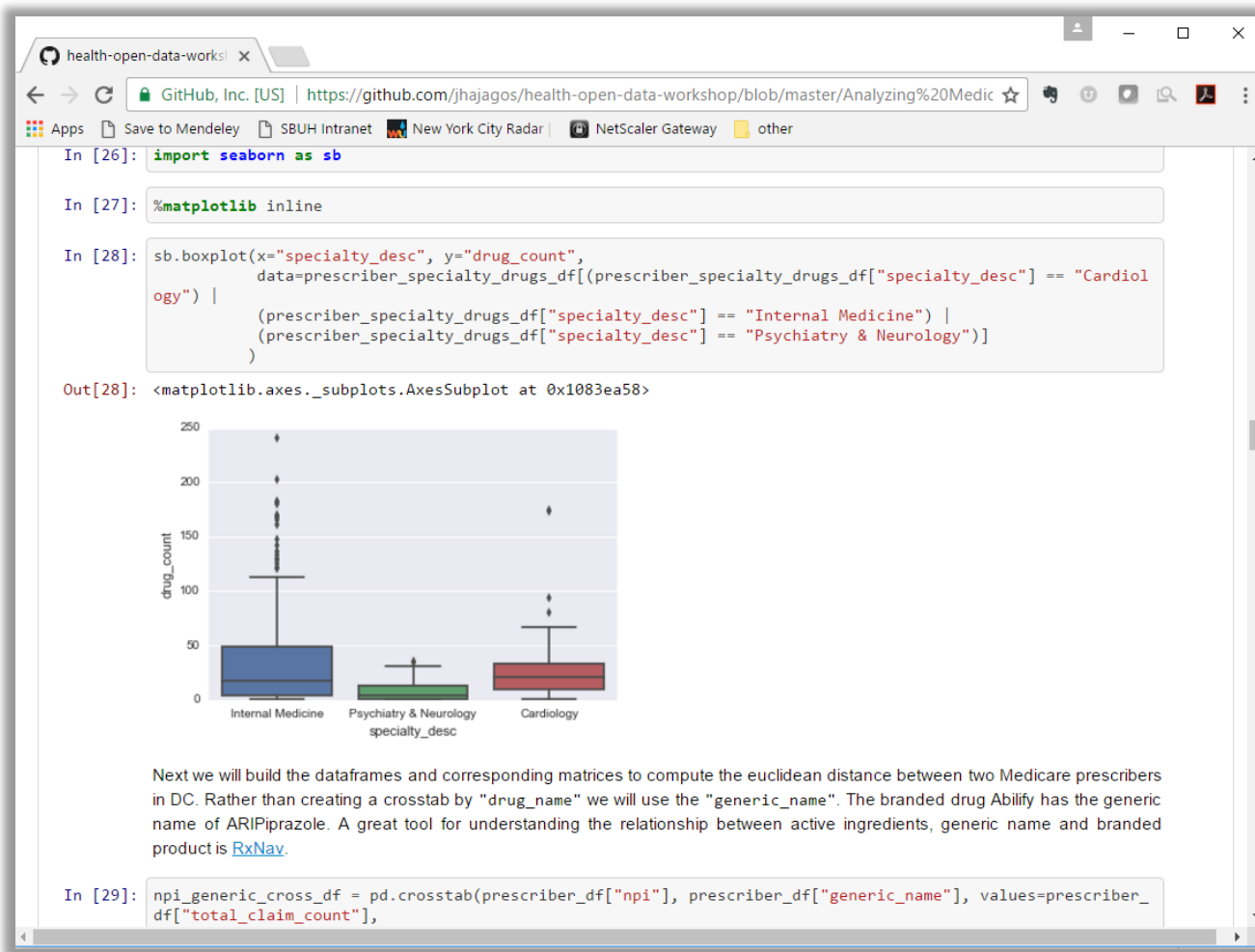
Example 2: Data sources and tools

- NetworkX
- SQLAlchemy
- MySQL
- Gephi 0.9.1
- ETL – load scripts for NPPES transformation script and teaming table load
 - <https://github.com/jhajagos/HealthcareAnalyticTools/>
- NPPES data:
http://download.cms.gov/nppes/NPI_Files.html
- Teaming data:
<https://questions.cms.gov/faq.php?faqId=7977>

Educating Data Scientists to work with healthcare data



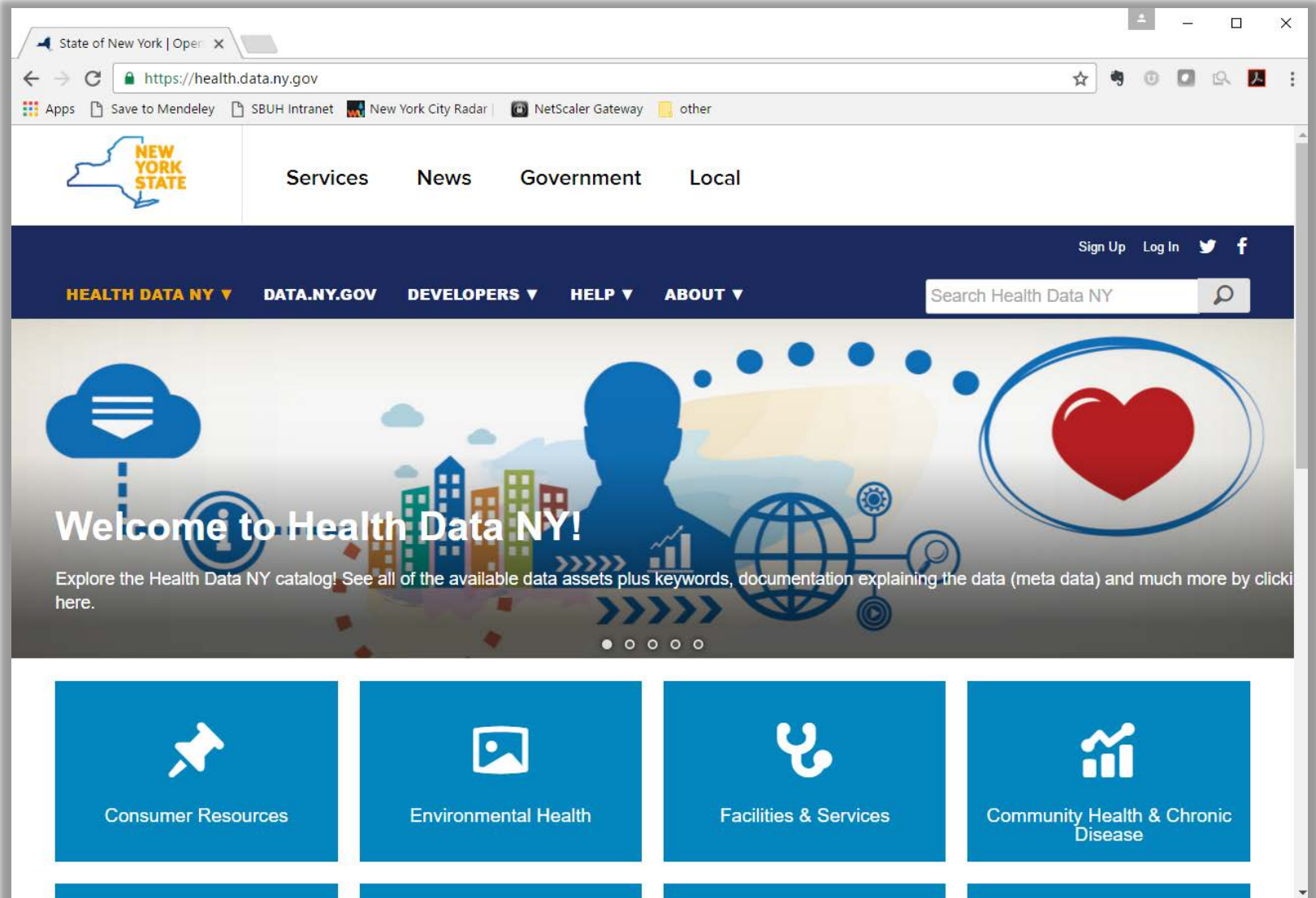
Jupyter Notebooks for Analytic Reproducible Analysis



Example 3: Kidney transplants in NY State



<https://www.flickr.com/photos/tareqsalahuddin/7272346858/>



Hospital Inpatient Discharges (SPARCS De-Identified): 2014

The Statewide Planning and Research Cooperative System (SPARCS) Inpatient De-Identified File contains

Manage More Views Filter Visualize Export Embed About More

| | Patient Disposition | Discharge Year | Discharge Day of Week | CCS Diagnosis Code | CCS Diagnosis Description |
|----|---------------------------------------|----------------|-----------------------|--------------------|---------------------------|
| 1 | Home or Self Care | 2014 | WED | 197 | SKIN/SUBCUT TISS INFECT |
| 2 | Home or Self Care | 2014 | WED | 146 | DIVERTICUL-OSIS/ITIS |
| 3 | Short-term Hospital | 2014 | SUN | 050 | DIABETES W/COMPL |
| 4 | Short-term Hospital | 2014 | SAT | 154 | GASTROENTRTS NONINFCT |
| 5 | Home or Self Care | 2014 | WED | 124 | TONSILLITIS |
| 6 | Home or Self Care | 2014 | MON | 123 | INFLUENZA |
| 7 | Skilled Nursing Home | 2014 | WED | 122 | PNEUMONIA |
| 8 | Home or Self Care | 2014 | MON | 123 | INFLUENZA |
| 9 | Home or Self Care | 2014 | WED | 123 | INFLUENZA |
| 10 | Hosp Basd Medicare Approved Swing Bed | 2014 | FRI | 231 | OTHER FRACTURE |
| 11 | Home or Self Care | 2014 | TUE | 123 | INFLUENZA |
| 12 | Hosp Basd Medicare Approved Swing Bed | 2014 | MON | 122 | PNEUMONIA |
| 13 | Home or Self Care | 2014 | SAT | 122 | PNEUMONIA |
| 14 | Home or Self Care | 2014 | FRI | 155 | OTHER GI DISORDER |
| 15 | Home or Self Care | 2014 | THU | 122 | PNEUMONIA |
| 16 | Home or Self Care | 2014 | MON | 127 | COPD |

Terms of Service Privacy Accessibility Give Feedback

NEW YORK STATE OF OPPORTUNITY. Department of Health

SOCRATA API with pandas library

```
In [16]: kt_url = 'https://health.data.ny.gov/resource/rmwa-zns4.json?ccs_procedure_code=105&$limit=10000'  
         print(kt_url)
```

```
https://health.data.ny.gov/resource/rmwa-zns4.json?ccs_procedure_code=105&$limit=10000
```

Once a URL is constructed to a data source a GET request over HTTP (HyperText Transfer Protocol) can be executed. The HTTP protocol is how most data is transferred from a host/server to the client. Here the client is not a web browser but the Python kernel running on your computer.

Each data source in the Socrata environment is identified by a short string or data tag. The SPARCS 2014 data tag is "rmwa-zns4". In the above request we are asking for a JSON document. JSON or Javascript Object Notation is a text based format for exchanging data in a machine readable format. One can think of JSON as a CSV format for the Internet Era. The Pandas' library function "read_json()" can take a URL and makes a remote call to the Socrata server and read the response. If the URL is misspecified than an error will occur. It converts the JSON response into a dataframe. Pandas' dataframes are powerful constructs for working with table based data.

```
In [17]: kidney_transplants_df = pd.read_json(kt_url)
```

```
In [18]: len(kidney_transplants_df.length_of_stay)
```

```
Out[18]: 1169
```



```
In [31]: kidney_transplants_df.groupby(["facility_name_with_id"])["length_of_stay"].count()
```

```
Out[31]: facility_name_with_id
0001 - Albany Medical Center Hospital      55
0210 - Erie County Medical Center          74
0245 - University Hospital                 59
0413 - Strong Memorial Hospital            60
0541 - North Shore University Hospital     29
0635 - University Hospital SUNY Health Science Center 64
1139 - Westchester Medical Center          27
1169 - Montefiore Medical Center - Henry & Lucy Moses Div 155
1320 - University Hospital of Brooklyn     23
1456 - Mount Sinai Hospital               162
1458 - New York Presbyterian Hospital - New York Weill Cornell Center 211
1463 - NYU Hospitals Center               25
1464 - New York Presbyterian Hospital - Columbia Presbyterian Center 225
Name: length_of_stay, dtype: int64
```

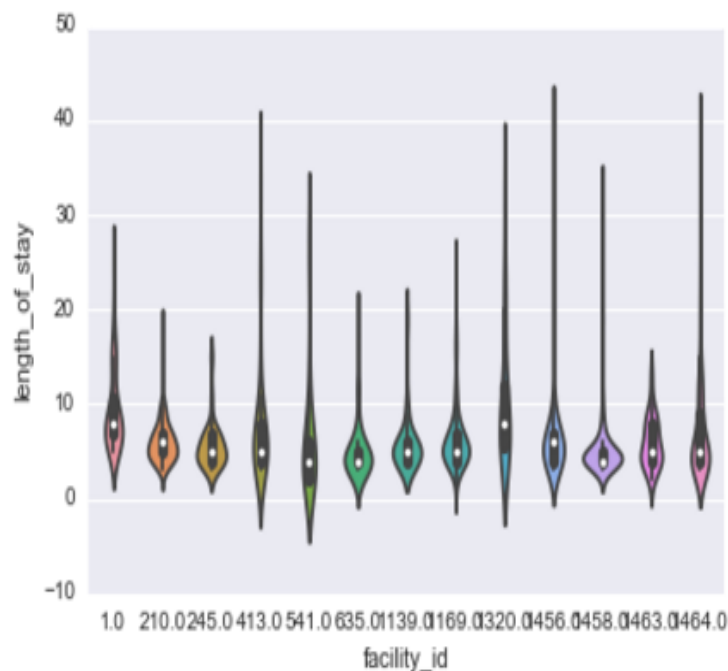
```
In [30]: kidney_transplants_df.groupby(["facility_name_with_id"])["length_of_stay"].mean()
```

```
Out[30]: facility_name_with_id
0001 - Albany Medical Center Hospital      9.327273
0210 - Erie County Medical Center          6.148649
0245 - University Hospital                 6.762712
0413 - Strong Memorial Hospital            9.250000
0541 - North Shore University Hospital     6.137931
0635 - University Hospital SUNY Health Science Center 5.265625
1139 - Westchester Medical Center          8.703704
1169 - Montefiore Medical Center - Henry & Lucy Moses Div 6.477419
1320 - University Hospital of Brooklyn    10.521739
1456 - Mount Sinai Hospital               7.228395
1458 - New York Presbyterian Hospital - New York Weill Cornell Center 5.246445
1463 - NYU Hospitals Center               5.960000
1464 - New York Presbyterian Hospital - Columbia Presbyterian Center 7.235556
Name: length_of_stay, dtype: float64
```

```
In [27]: kidney_transplants_outliers_removed_df = kidney_transplants_df.where(kidney_transplants_df["length_of_stay"] <= 40)
```

```
In [28]: sb.violinplot(x="facility_id", y="length_of_stay", data=kidney_transplants_outliers_removed_df)
```

```
Out[28]: <matplotlib.axes._subplots.AxesSubplot at 0xcd74940>
```



Example 3: Data sources and tools

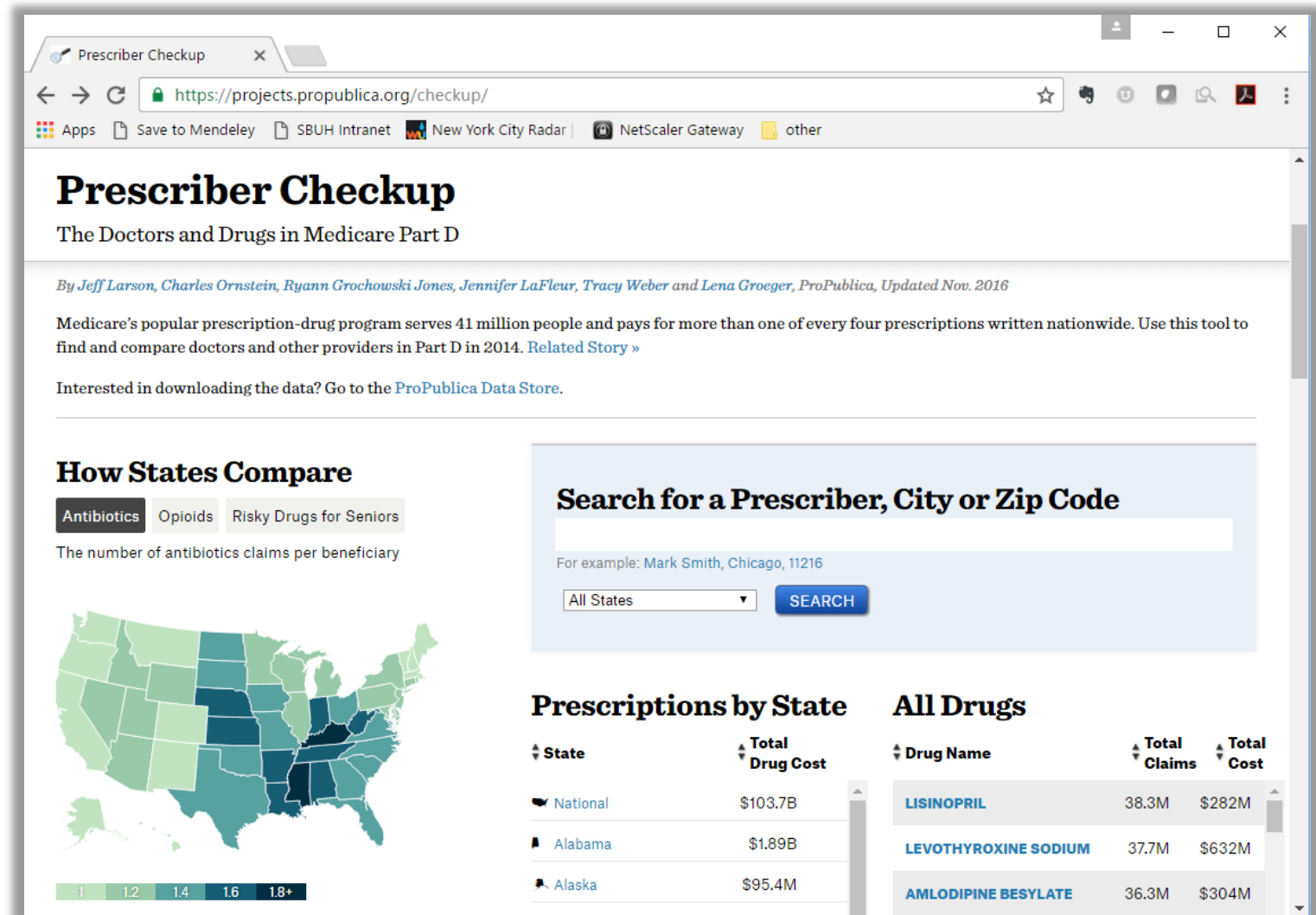
- Notebook - <https://github.com/jhajagos/health-open-data-workshop/blob/master/SPARCS%20Kidney%20Transplants%20in%20NY%20CY%202014.ipynb>
- Anaconda Python distribution - <https://www.continuum.io/downloads>
- Seaborn library - <http://seaborn.pydata.org/>
- pandas - <http://pandas.pydata.org/>
- Socrata API - <https://dev.socrata.com/consumers/getting-started.html>
- SPARCS 2014 discharge data - <https://health.data.ny.gov/resource/rmwa-zns4>

Example 4: Psychiatric drug prescribers in D.C.



<https://www.flickr.com/photos/51274664@N06/6930338021/>

Medicare prescribing data



Can we build a distance metric to find similar prescribers

Prescriber 1 = (0,0,0,0,1,0,0,1,0,0,0,1,1)

Prescriber 2 = (0,0,0,0,1,0,0,1,0,0,0,1,0)

Prescriber 3 = (1,1,1,0,0,0,0,0,0,0,0,0,1)

Where i^{th} entry indicates whether the prescriber prescribes Drug i

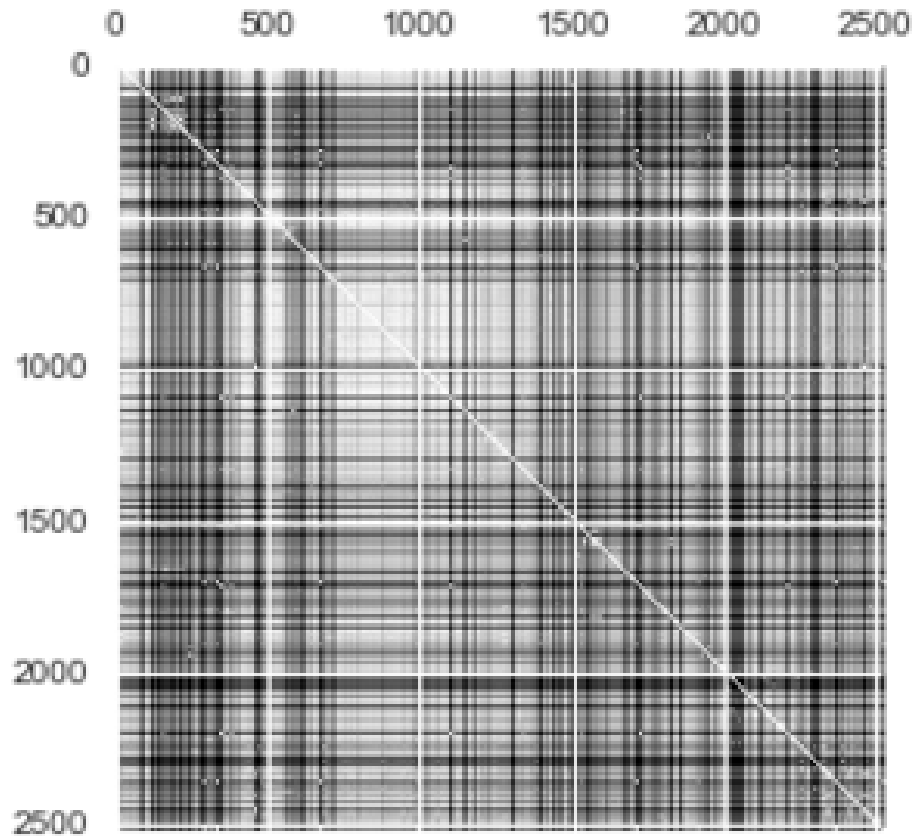
Euclidean distance between providers:

Prescriber 1 and 2 is $\text{Sqrt}(1) = 1$

Prescriber 1 and 3 is $\text{Sqrt}(6) = 2.44$

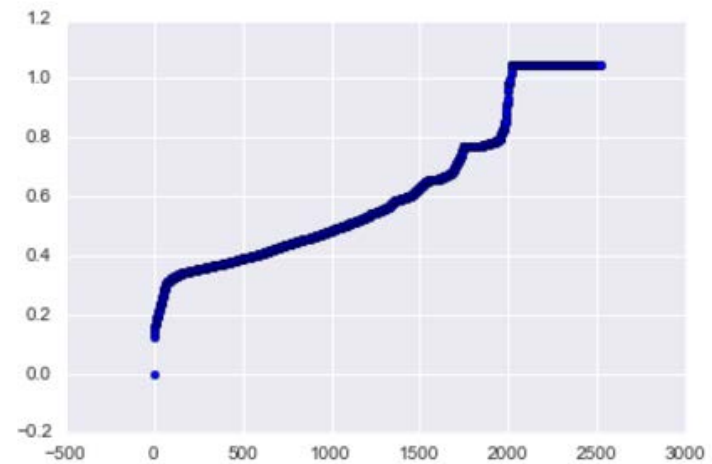
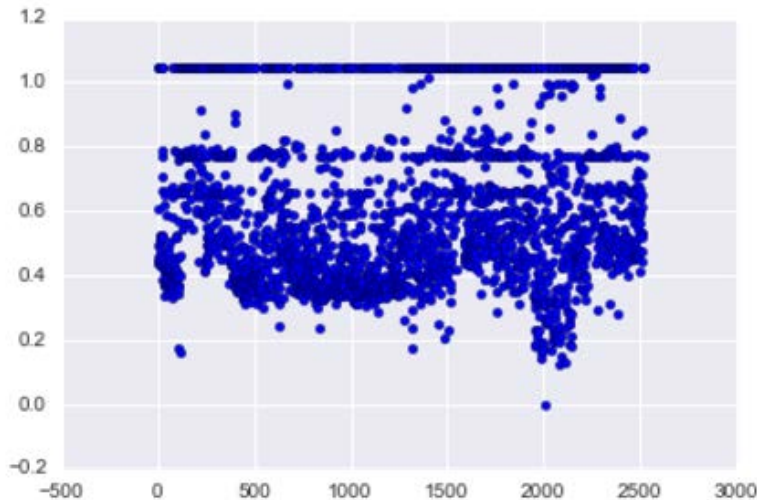
Prescriber 2 and 3 is $\text{Sqrt}(7) = 2.65$

Prescriber Distance Matrix



White to Black – Small distance to big distance

A slice of the distance matrix



Sorted by increasing distance

Pie image: <https://www.flickr.com/photos/aloha75/5953100136/>

Sorted list of NPIs with increasing Rx distance

```
In [51]: providers_sorted = np.lexsort((prescriber_dist[:,2010].tolist(),))
```

```
In [52]: prescriber_specialty_generic_df.iloc[:,0:2].as_matrix()[providers_sorted[0:40],:]
```

```
Out[52]: array([[1487818670L, u'Psychiatry'],
 [1114970076L, u'Psychiatry & Neurology'],
 [1588810162L, u'Psychiatry & Neurology'],
 [1265692115L, u'Psychiatry'],
 [1366766263L, u'Psychiatry & Neurology'],
 [1285815878L, u'Psychiatry'],
 [1992965537L, u'Certified Clinical Nurse Specialist'],
 [1750616645L, u'Psychiatry'],
 [1366618746L, u'Psychiatry'],
 [1801919659L, u'Certified Clinical Nurse Specialist'],
 [1720241011L, u'Neuropsychiatry'],
 [1326086125L, u'Psychiatry'],
 [1790964948L, u'Psychiatry & Neurology'],
 [1033322730L, u'Psychiatry'],
 [1023267606L, u'Psychiatry'],
 [1780766881L, u'Psychiatry & Neurology'],
 [1316089642L, u'Psychiatry'],
 [1164621363L, u'Psychiatry & Neurology'],
 [1184876948L, u'Psychiatry'],
 [1821259581L, u'Psychiatry'],
 [1992746515L, u'Psychiatry'],
 [1821073776L, u'Psychiatry'],
 [1770868796L, u'Nurse Practitioner'],
```

The NPI is the key to rich provider data

[Doc NPI](#) [1992965537](#) [Identifiers](#) [Taxonomies](#) [Contact](#) [Connections](#) [Change Search Type](#) [FAQ](#)

DR. BRITT BALL ALICE APRN, PHD

You can also download the this 1992965537 data report as [csv](#) | [excel](#) | [json](#) | [xml](#)

| Provider Prefix | Provider Last Name | Provider Middle Name | Provider First Name | Provider Credential |
|-----------------|--------------------|----------------------|---------------------|---------------------|
| DR. | BRITT | BALL | ALICE | APRN, PHD |

Gender:
Female

Last update date:
2013-06-21

Is sole proprietor?:
No

In PECOS?:
Yes

Enumeration date:
2008-06-16

Current as of:

Is organization subpart?:
No

Search for ALICE, BRITT in
[PubMed](#)
[Propublica Dollars for Docs](#)

Identifiers

| identifier | description | issuer | state |
|------------|-------------|--------|-------|
|------------|-------------|--------|-------|

Taxonomies

| Taxonomy | State | License Number | Primary? |
|---|-------|----------------|----------|
| 364SP0809X Psychiatric/Mental Health, Adult | DC | RN44959 | |

Example 4: Data sources and tools

- Notebook - <https://github.com/jhajagos/health-open-data-workshop/blob/master/Analyzing%20Medicare%20Part%20D%20Prescriber%20Data.ipynb>
- Medicare Prescriber data - <https://data.cms.gov/Public-Use-Files/Medicare-Provider-Utilization-and-Payment-Data-201/4uvc-gbfz>

