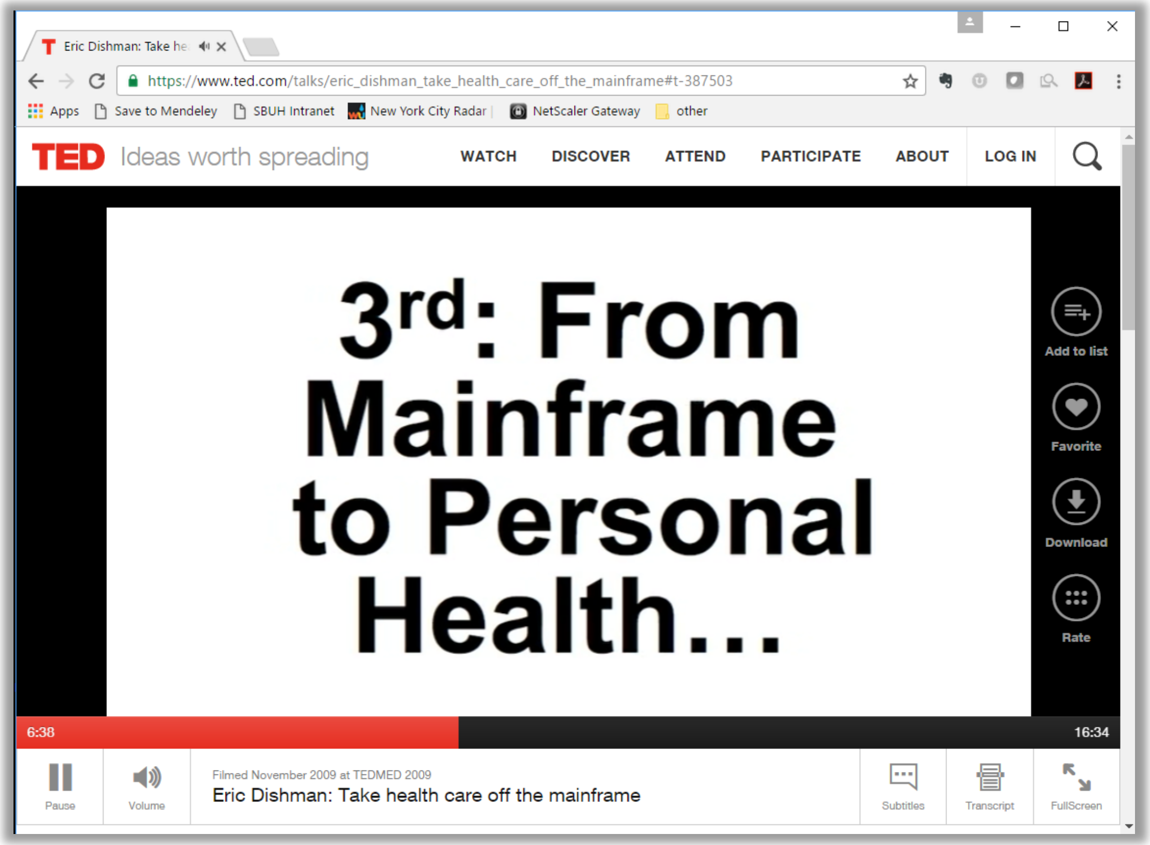
Open Healthcare Data and Tools in Practice

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


Ctrl
Fn

Still Literally True in 2016!



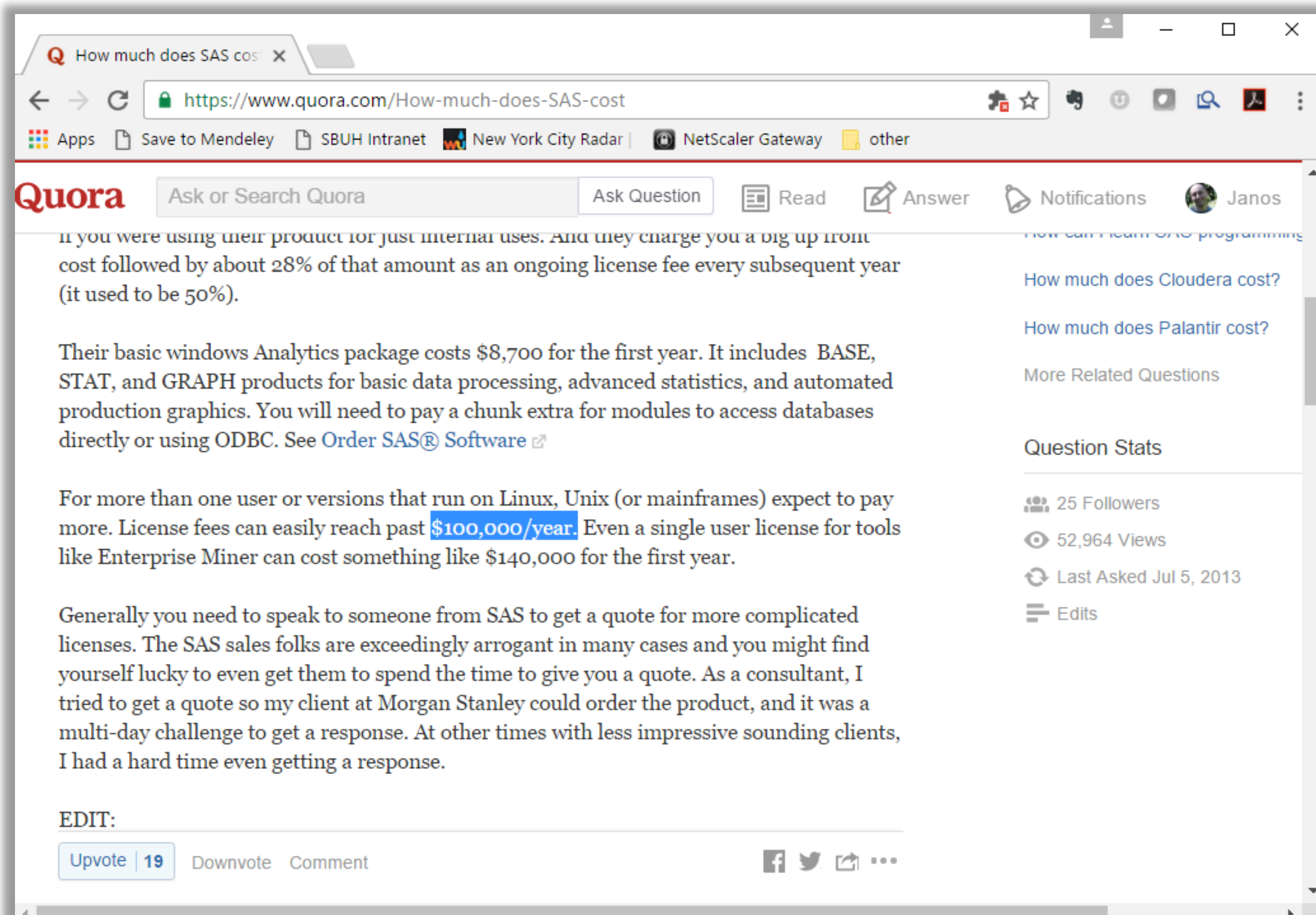
https://www.ted.com/talks/eric_dishman_take_health_care_off_the_mainframe

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/

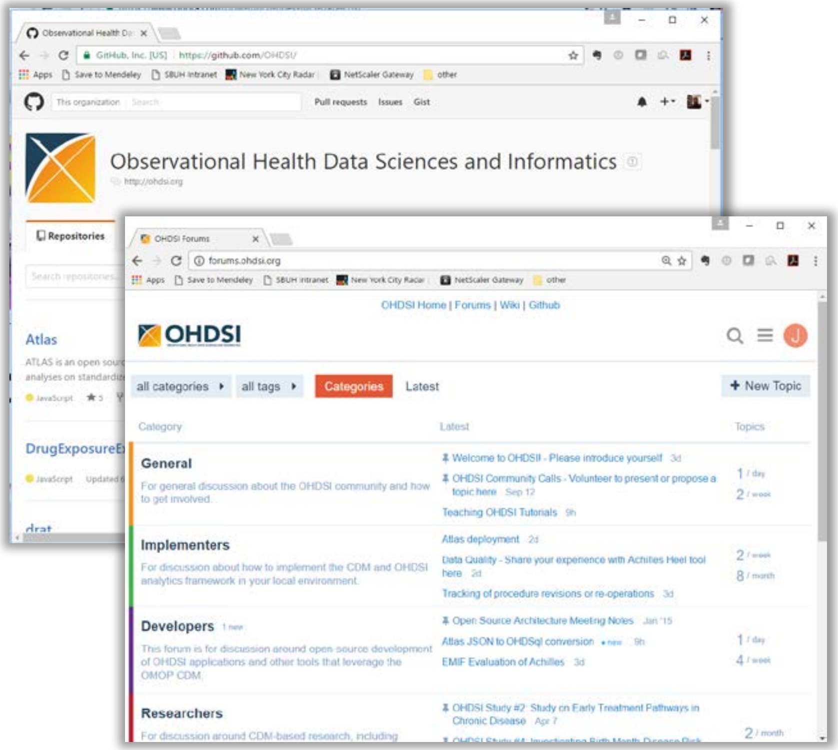
The legacy way: analytic software



The 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

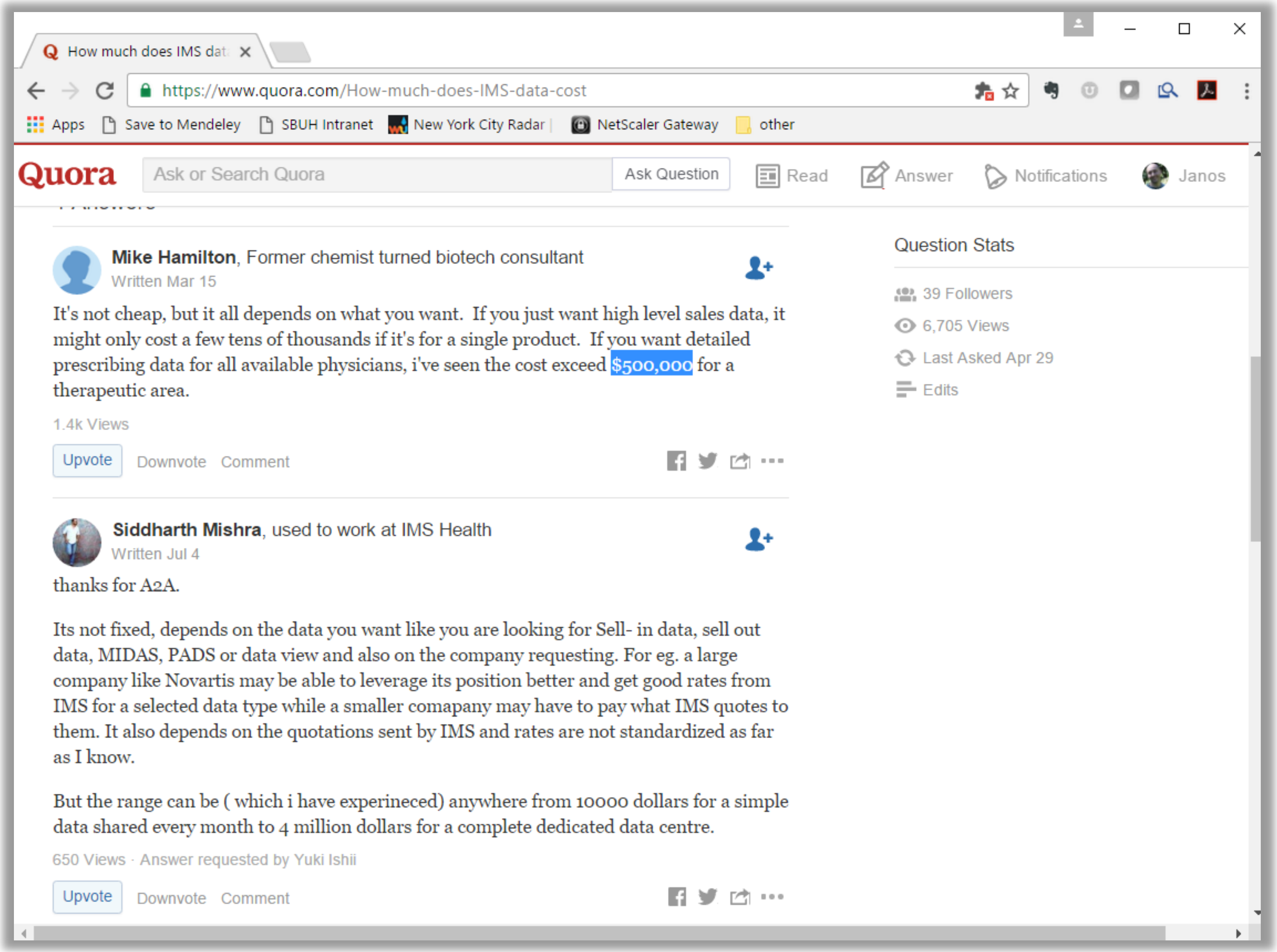
Example: OHDSI Software Stack



http://www.ohdsi.org/


NO TRES
WETLANDS RES
Digging No Shell-Fish
No Entry Within 300 of
espassing Will Result
Plus Cour
Day Nigh

The legacy way: healthcare data



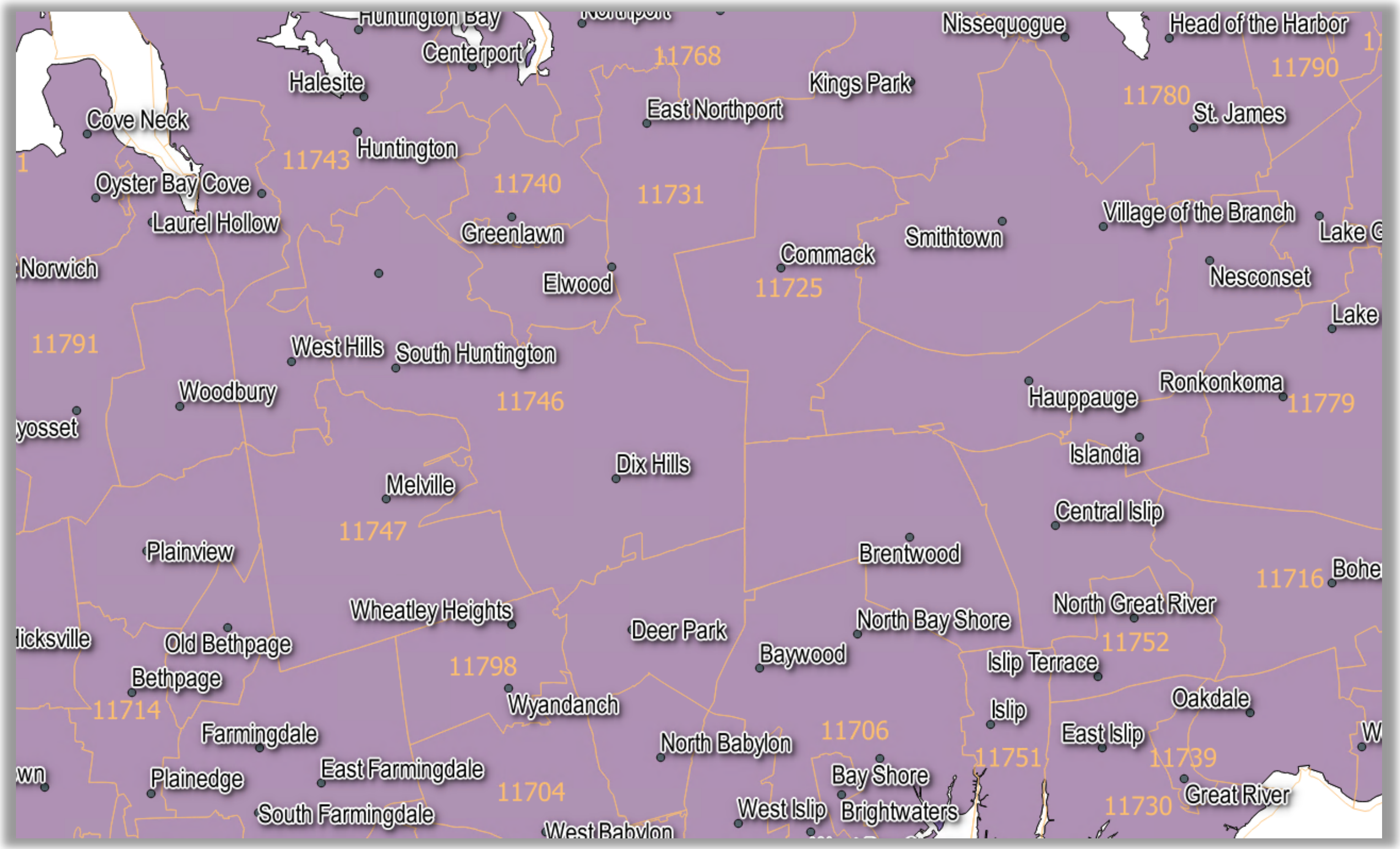
The open way: Data portal

Example 1: Enriching your local patient data

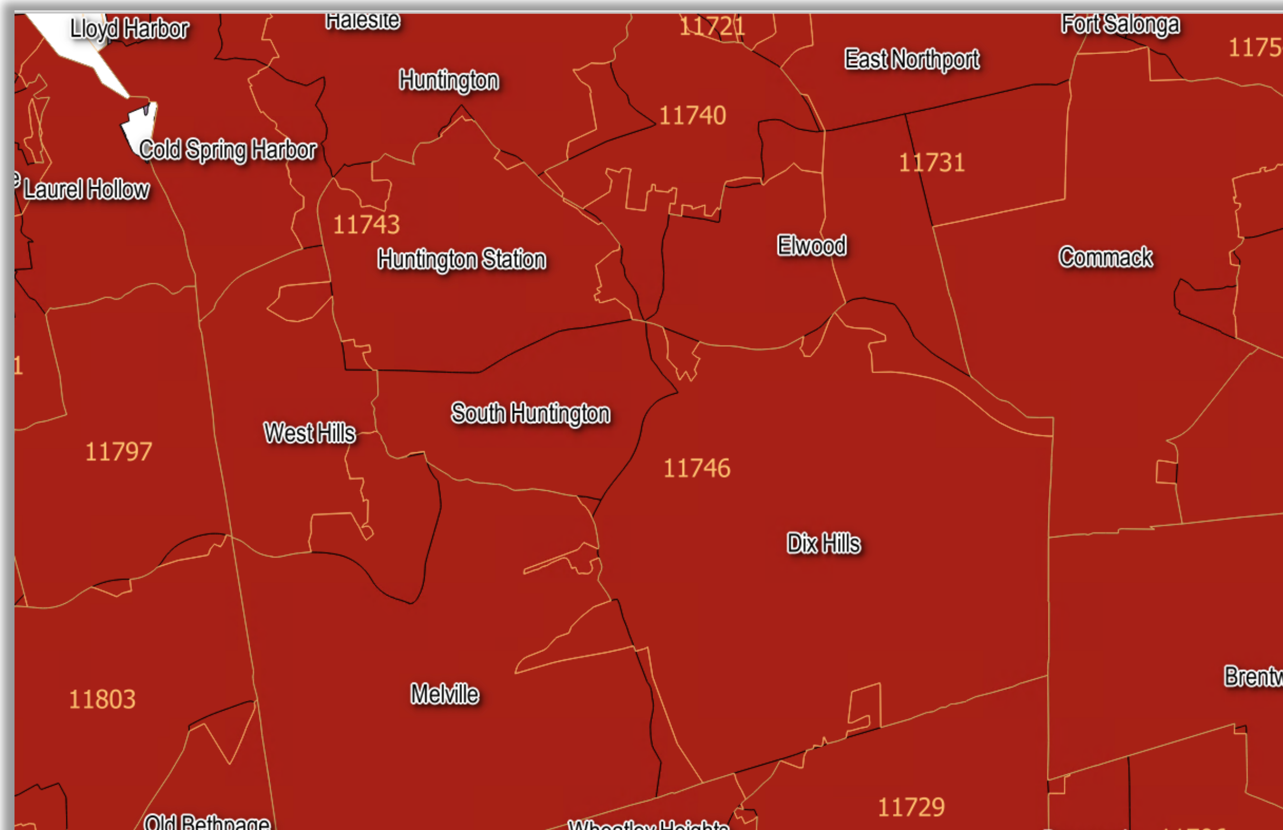
patient_id	Address
1001	100 Mains Street, Springfield, MA 01103
1022	12 Oak Drive, Springfield, MA 01105
3033	1001 East Main St., Greenfield, MA 01301
4010	101 Route 9a, Deerfield, MA 01342

Going beyond the zip code

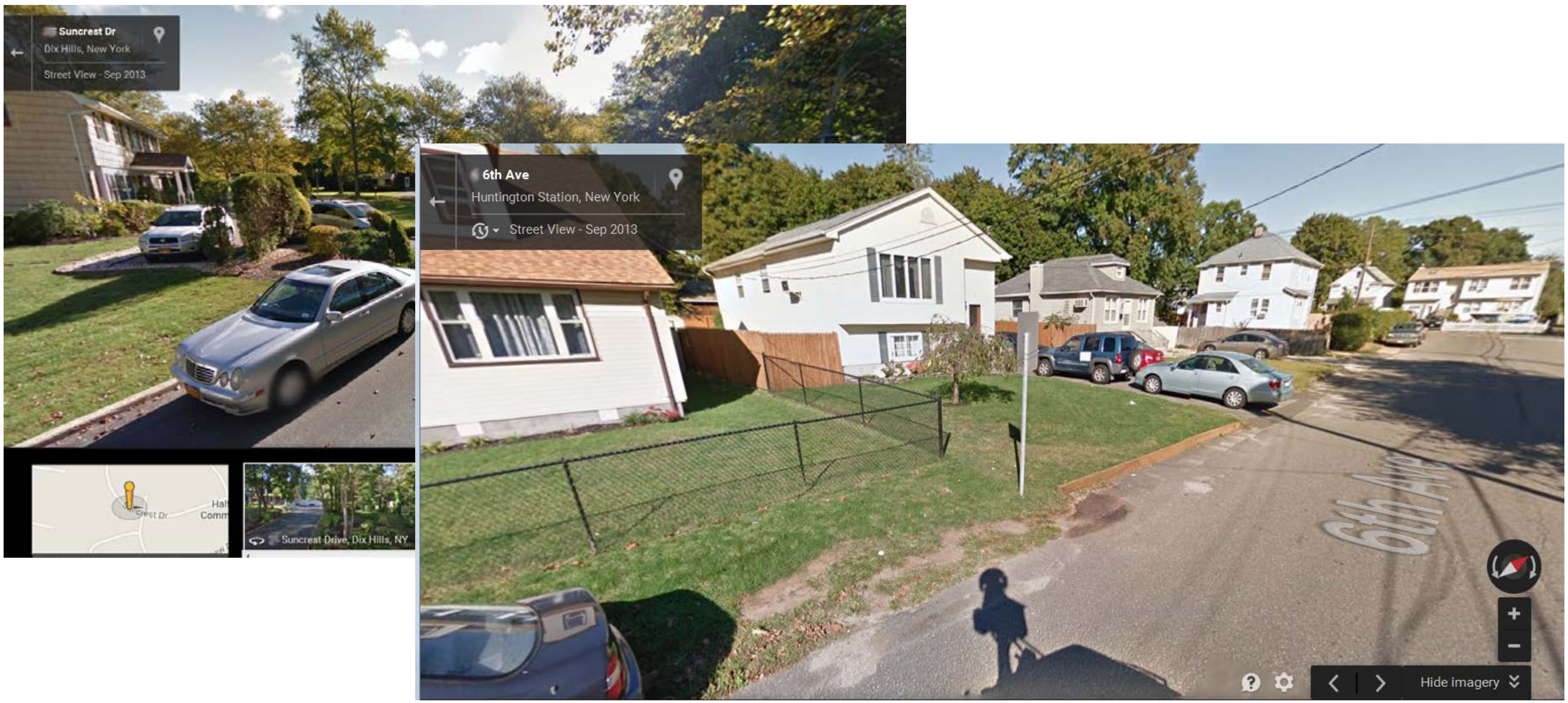
11746 – Huntington Station and Dix Hills



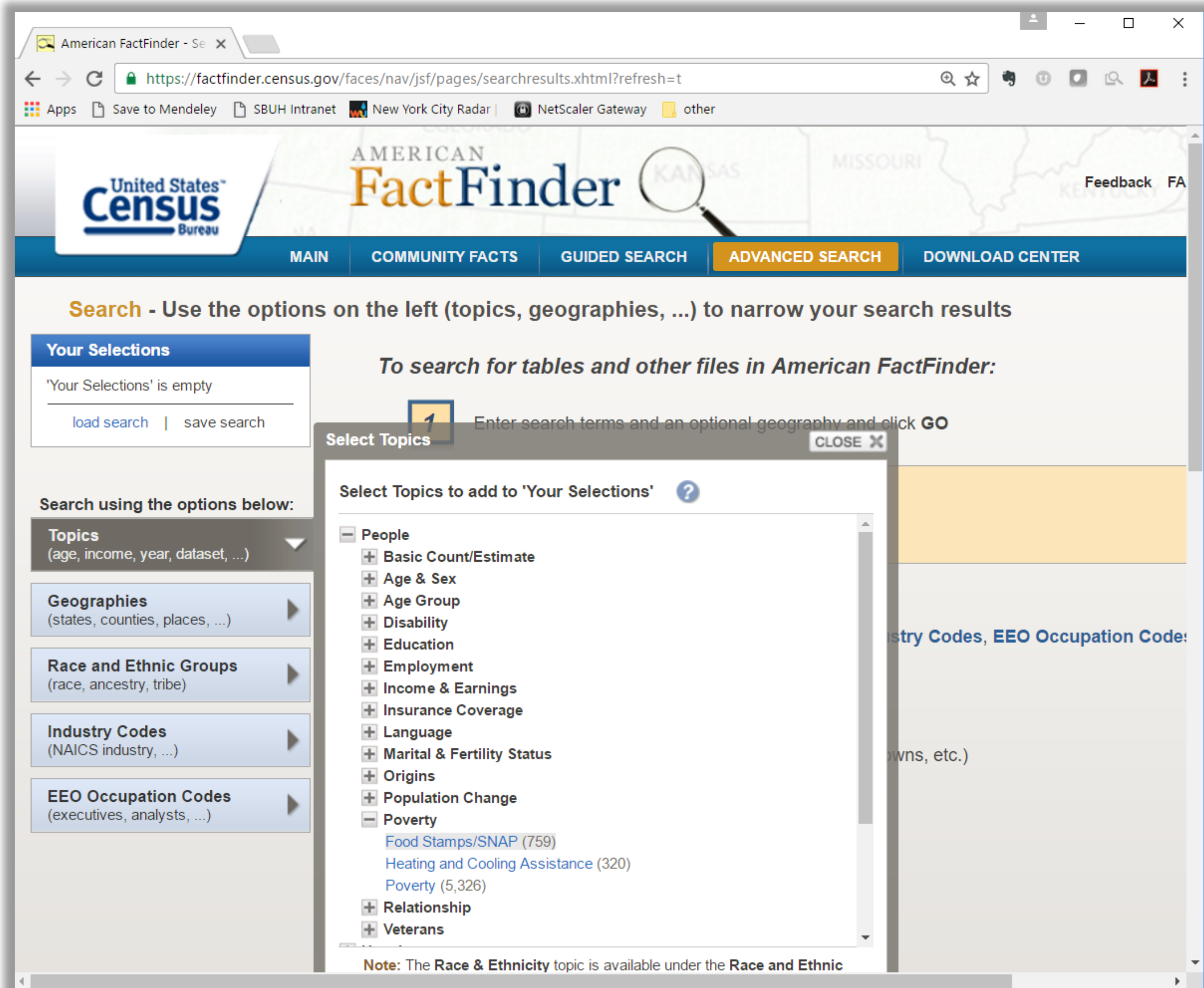
Two Census-Designated Places



Understanding socio-economic determinants health for your patients

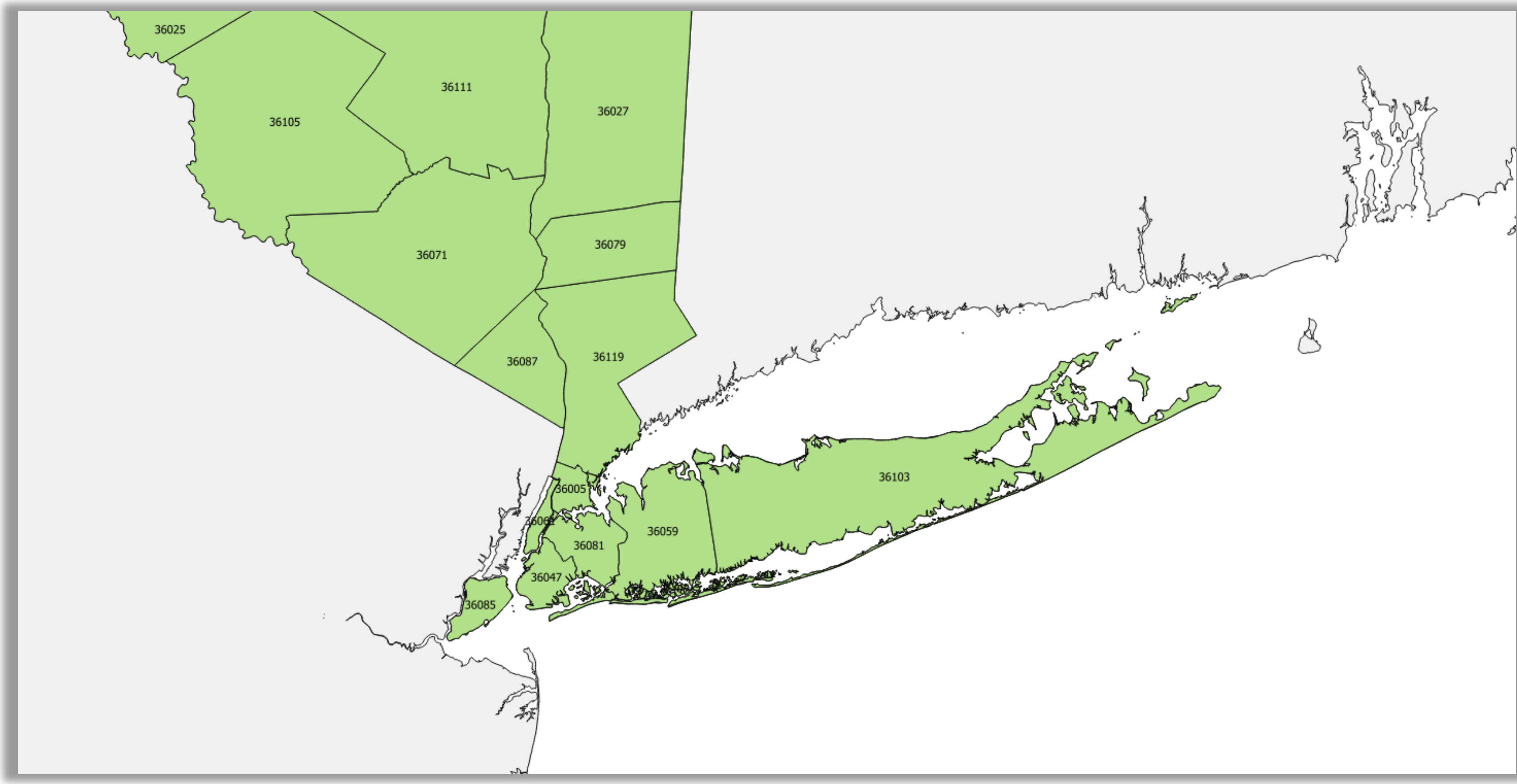


Google Map's Street View

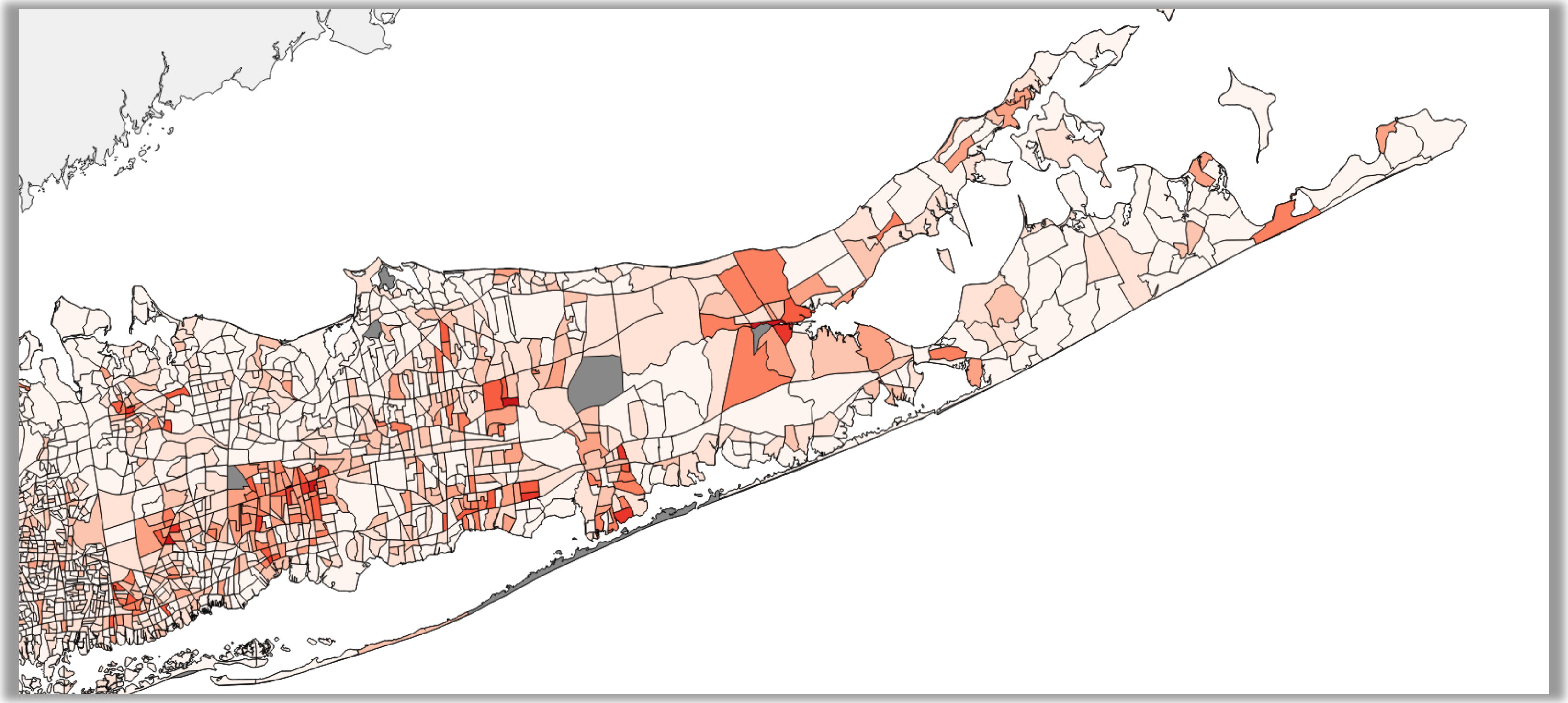


https://factfinder.census.gov/

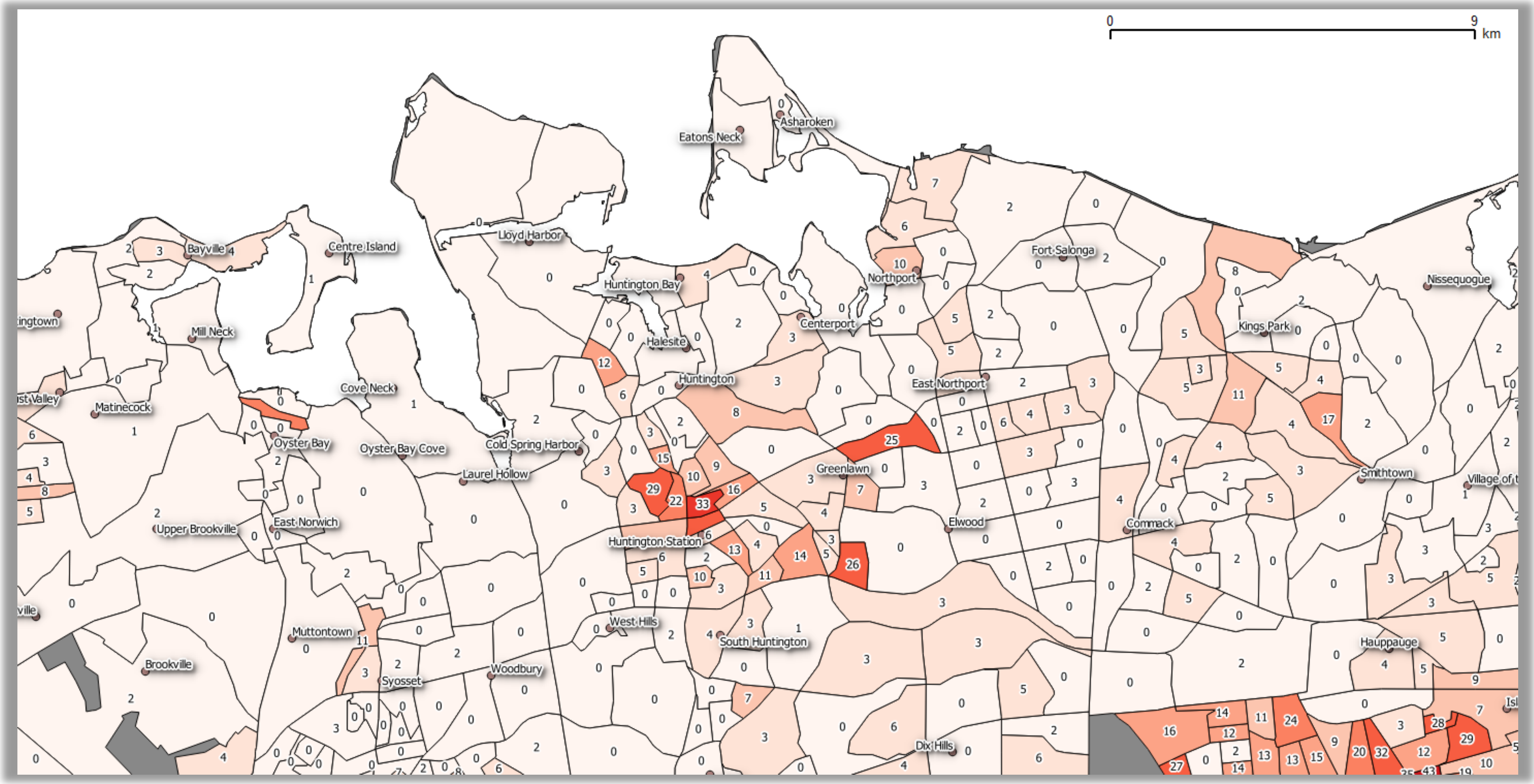
AT THE BLOCK GROUP LEVEL



SNAP By Geographic Region



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

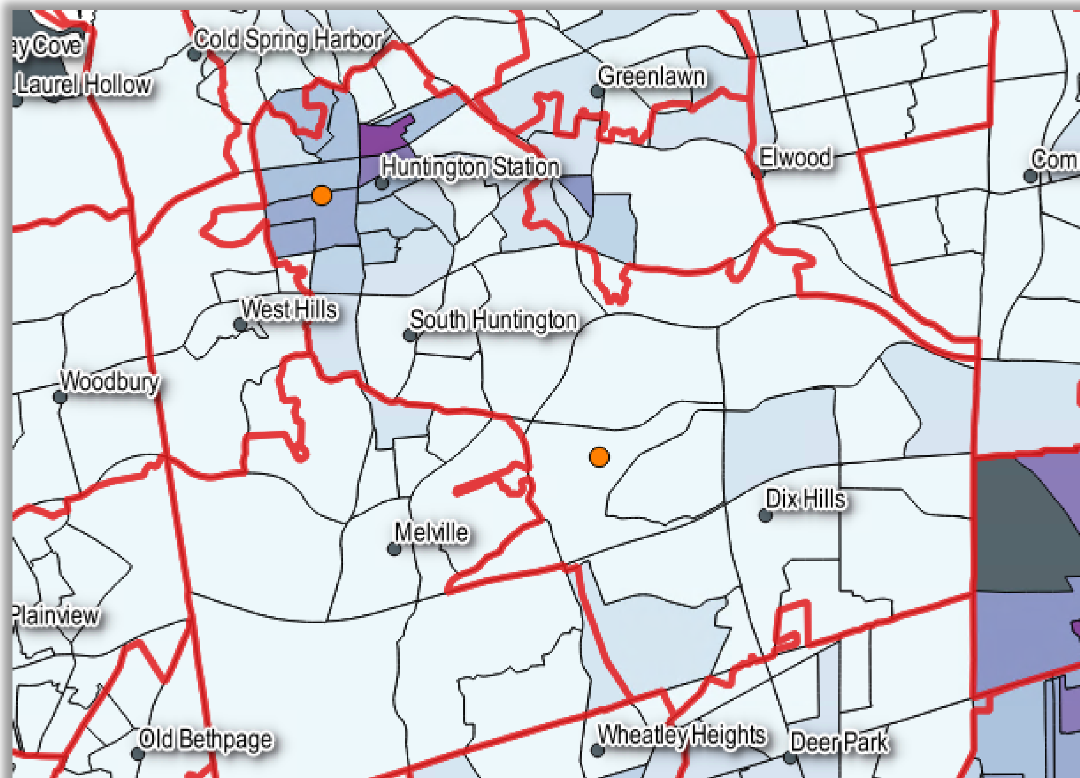


B22010: RECEIPT OF FOOD STAMPS IN THE PAST 12 MONTHS BY DISABILITY STATUS FOR HOUSEHOLDS – Percent of households who received Food Stamps/SNAP in the past 12 months

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
0101000020AD100000E5744739715752C06804F93EFA664440	0	40.80451	-73.366[illegible]	[illegible] Suncrest Dr, Dix Hills, NY 11746	11746
0101000020AD1000007F87CC2FCC5A52C0A2F984B9E46B4440	0	40.84292	-73.418[illegible]	[illegible] 6th Ave, Huntington Station, NY 11746	11746

Example 1:Data sources and tools

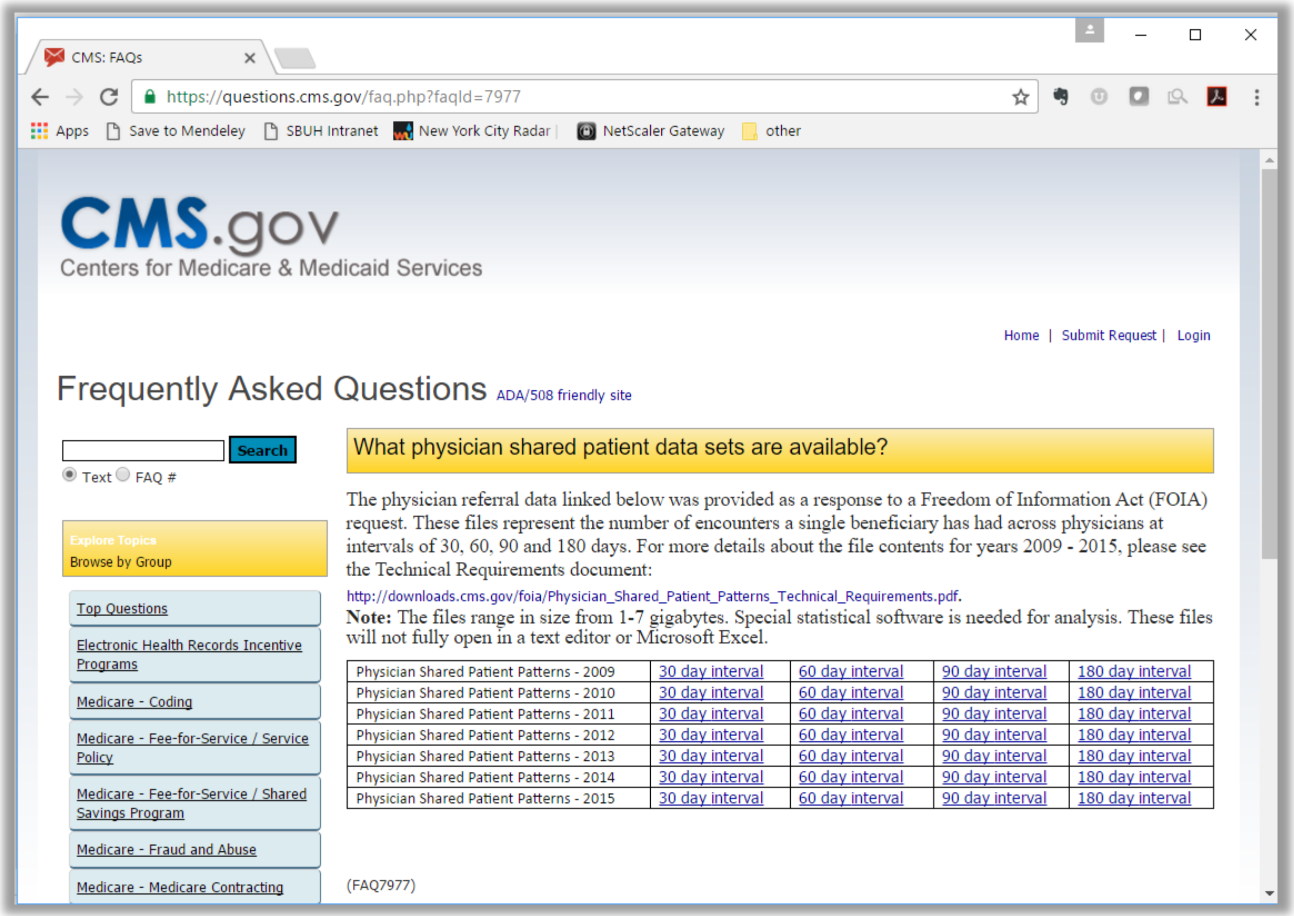
- PostgreSQL - https://www.postgresql.org/
- PostGIS and Tiger Geocoder - http://www.postgis.net/
- Shape files - https://www.census.gov/geo/maps-data/data/tiger-line.html
- Raw ACS files - http://www.census.gov/programs-surveys/acs/data/data-via-ftp.html
- Python Code for extracting ACS variables - 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/

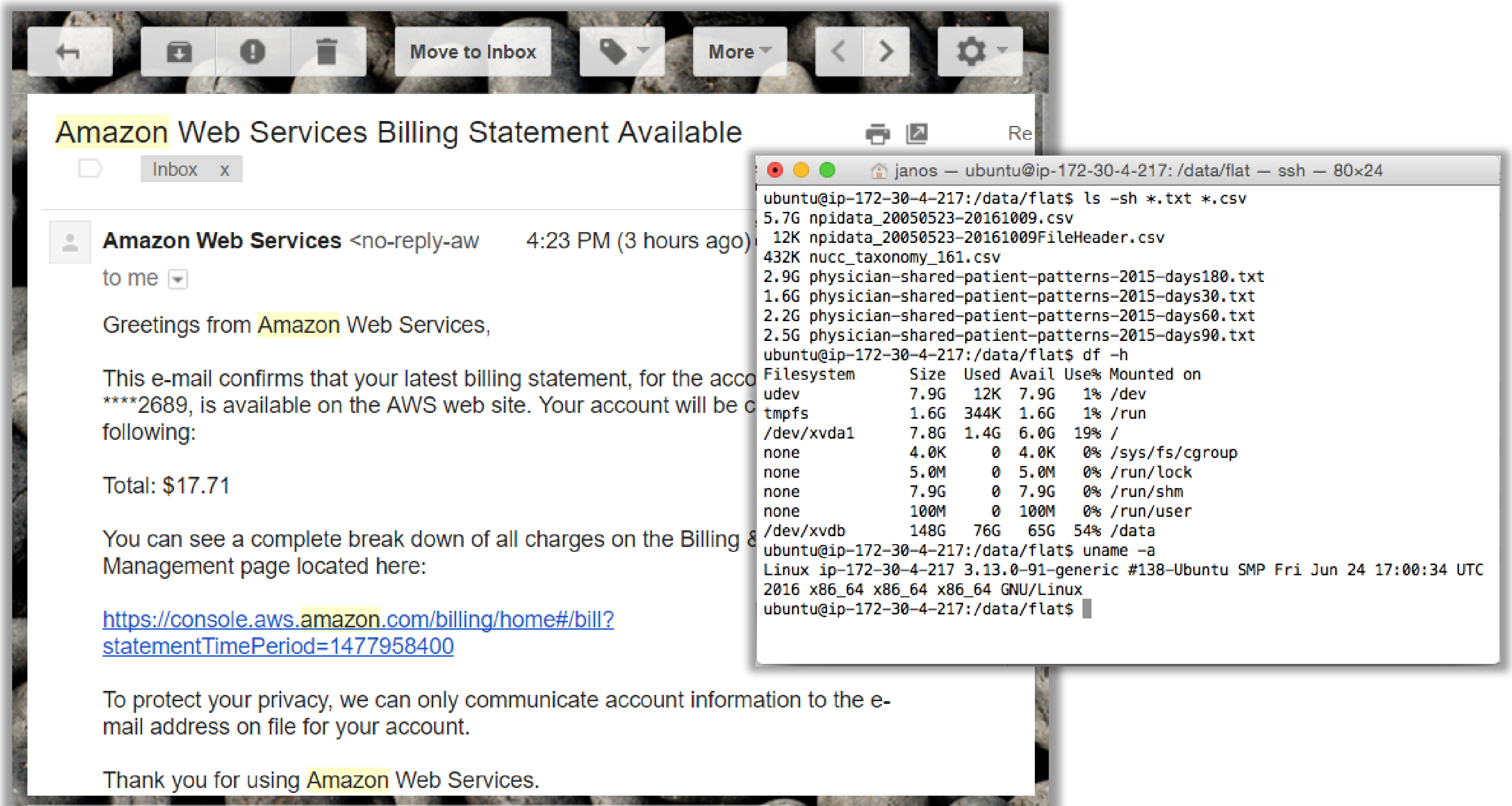
https://www.flickr.com/photos/rheinitz/8668769226/

Medicare Teaming Data



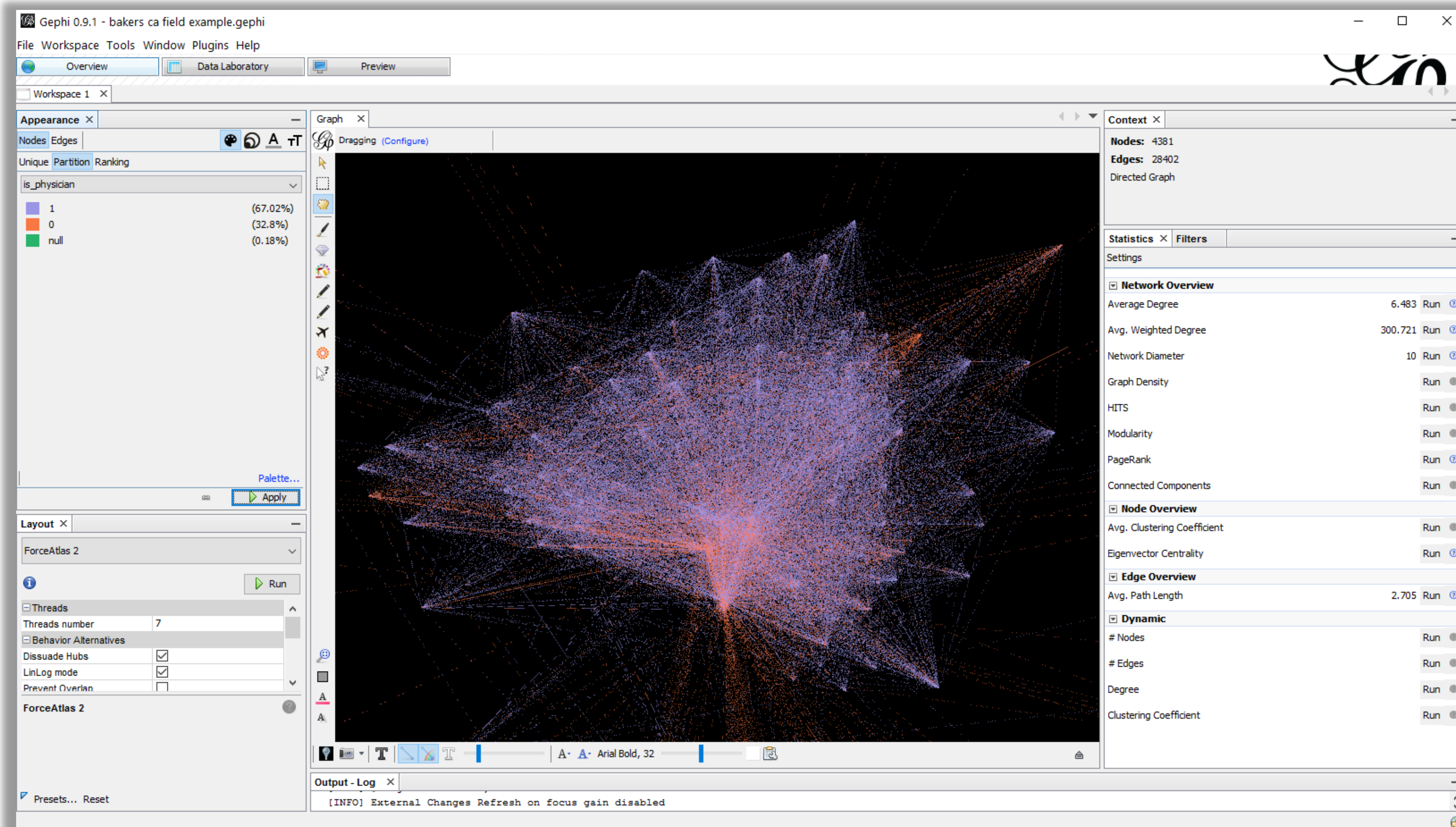
https://questions.cms.gov/faq.php?faqId=7977

For "Big Open Data"* we don't need to invest in an expensive HIPAA compliant environment



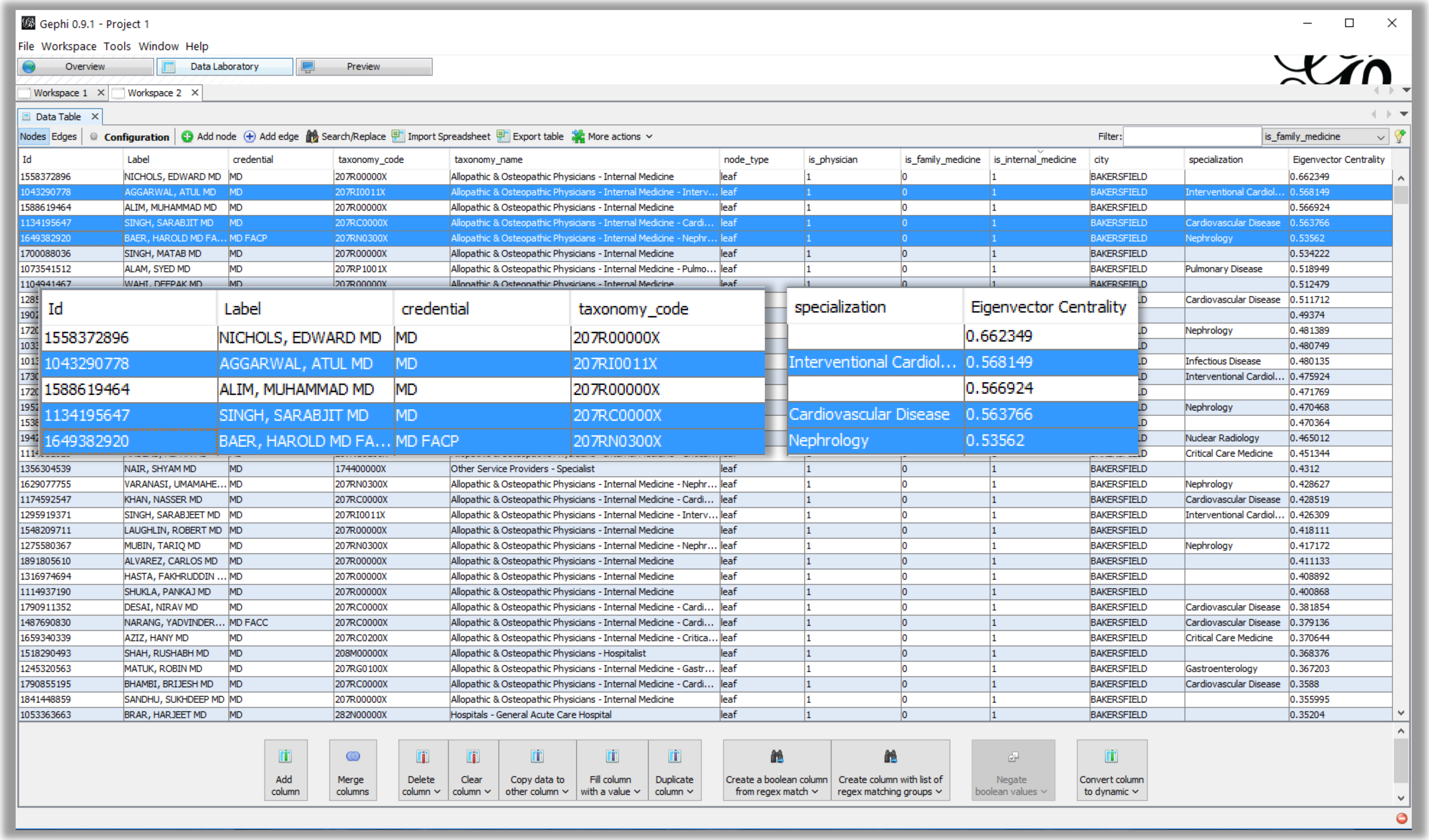
*Bigger than supported by a standard business class laptop or desktop or Excel

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



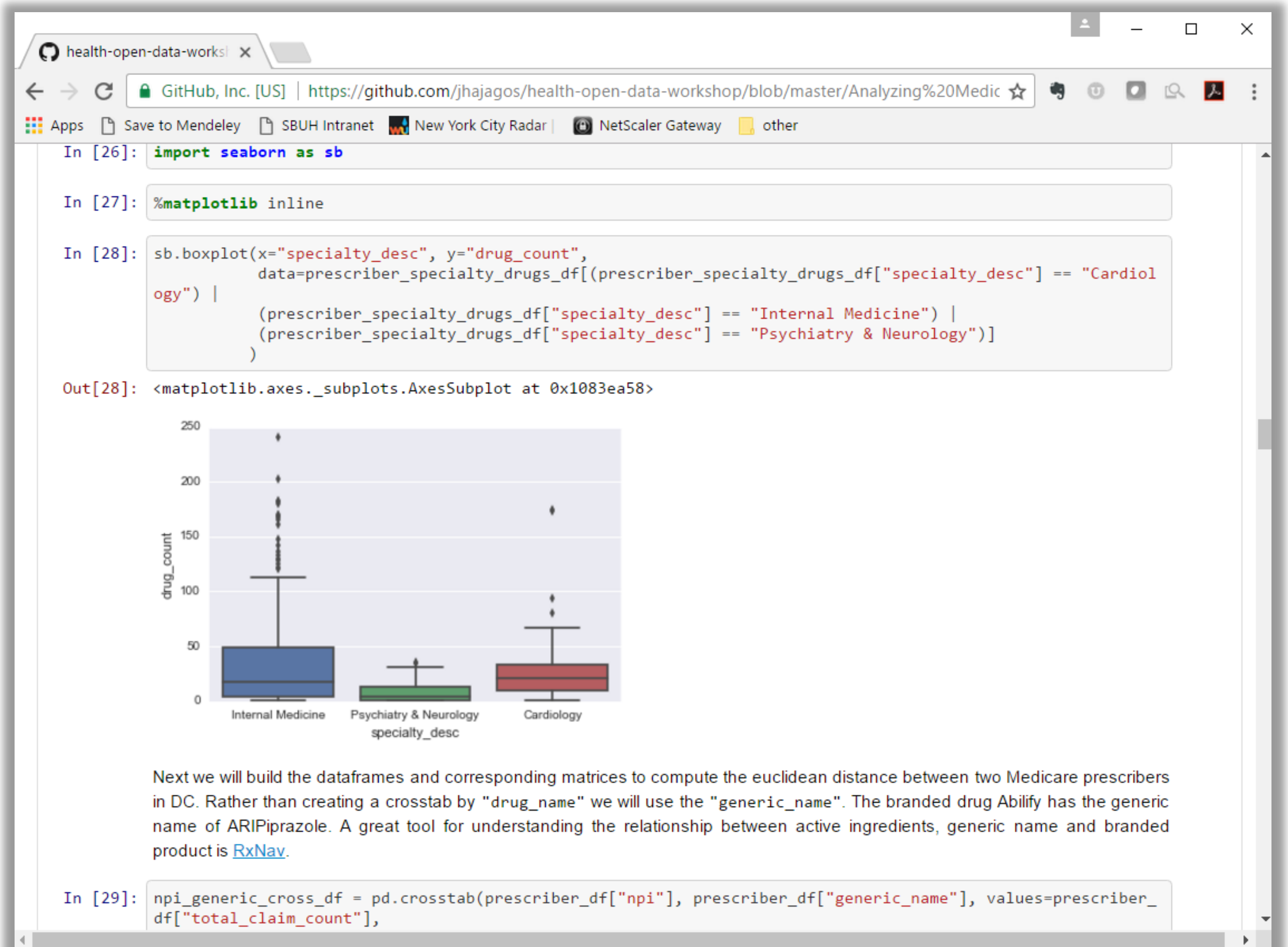
Gephi 0.9.1 - Project 1
File Workspace Tools Window Help
Overview
Data Laboratory
Preview
Workspace 1
Workspace 2
Data Table
Nodes Edges Configuration Add node Add edge Search/Replace Import Spreadsheet Export table More actions
Filter: is_family_medicine
Id | Label | credential | taxonomy_code | taxonomy_name | 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 | 207RI0011X | 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 - Cardi… | leaf | 1 | 0 | 1 | BAKERSFIELD | Cardiovascular Disease | 0.563766
1649382920 | BAER, HAROLD MD FA… | MD FACP | 207RN0300X | Allopathic & Osteopathic Physicians - Internal Medicine - Nephr… | 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
Cardiovascular Disease 0.511712
0.49374
Nephrology 0.481389
0.480749
Infectious Disease 0.480135
Interventional Cardiol… 0.475924
0.471769
Nephrology 0.470468
0.470364
Nuclear Radiology 0.465012
Critical Care Medicine 0.451344
1356304539 | NAIR, SHYAM MD | MD | 174400000X | Other Service Providers - Specialist | leaf | 1 | 0 | 1 | BAKERSFIELD | | 0.4312
1629077755 | VARANASI, UMAMAHE… | MD | 207RN0300X | Allopathic & Osteopathic Physicians - Internal Medicine - Nephr… | leaf | 1 | 0 | 1 | BAKERSFIELD | Nephrology | 0.428627
1174592547 | KHAN, NASSER MD | MD | 207RC0000X | Allopathic & Osteopathic Physicians - Internal Medicine - Cardi… | leaf | 1 | 0 | 1 | BAKERSFIELD | Cardiovascular Disease | 0.428519
1295919371 | SINGH, SARABJEET MD | MD | 207RI0011X | 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 - Nephr… | 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, FAKHRUDDIN … | 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 - Cardi… | leaf | 1 | 0 | 1 | BAKERSFIELD | Cardiovascular Disease | 0.381854
1487690830 | NARANG, YADVINDER… | MD FACC | 207RC0000X | Allopathic & Osteopathic Physicians - Internal Medicine - Cardi… | leaf | 1 | 0 | 1 | BAKERSFIELD | Cardiovascular Disease | 0.379136
1659340339 | AZIZ, HANY MD | MD | 207RC0200X | Allopathic & Osteopathic Physicians - Internal Medicine - Critica… | 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 | BHAMBI, BRIJESH MD | MD | 207RC0000X | Allopathic & Osteopathic Physicians - Internal Medicine - Cardi… | leaf | 1 | 0 | 1 | BAKERSFIELD | Cardiovascular Disease | 0.3588
1841448859 | SANDHU, SUKHDEEP 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
Id | Label | credential | taxonomy_code
1558372896 | NICHOLS, EDWARD MD | MD | 207R00000X
1043290778 | AGGARWAL, ATUL MD | MD | 207RI0011X
1588619464 | ALIM, MUHAMMAD MD | MD | 207R00000X
1134195647 | SINGH, SARABJIT MD | MD | 207RC0000X
1649382920 | BAER, HAROLD MD FA… | MD FACP | 207RN0300X
specialization | Eigenvector Centrality
| 0.662349
Interventional Cardiol… | 0.568149
| 0.566924
Cardiovascular Disease | 0.563766
Nephrology | 0.53562
Add column
Merge columns
Delete column
Clear column
Copy data to other column
Fill column with a value
Duplicate column
Create a boolean column from regex match
Create column with list of regex matching groups
Negate boolean values
Convert 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

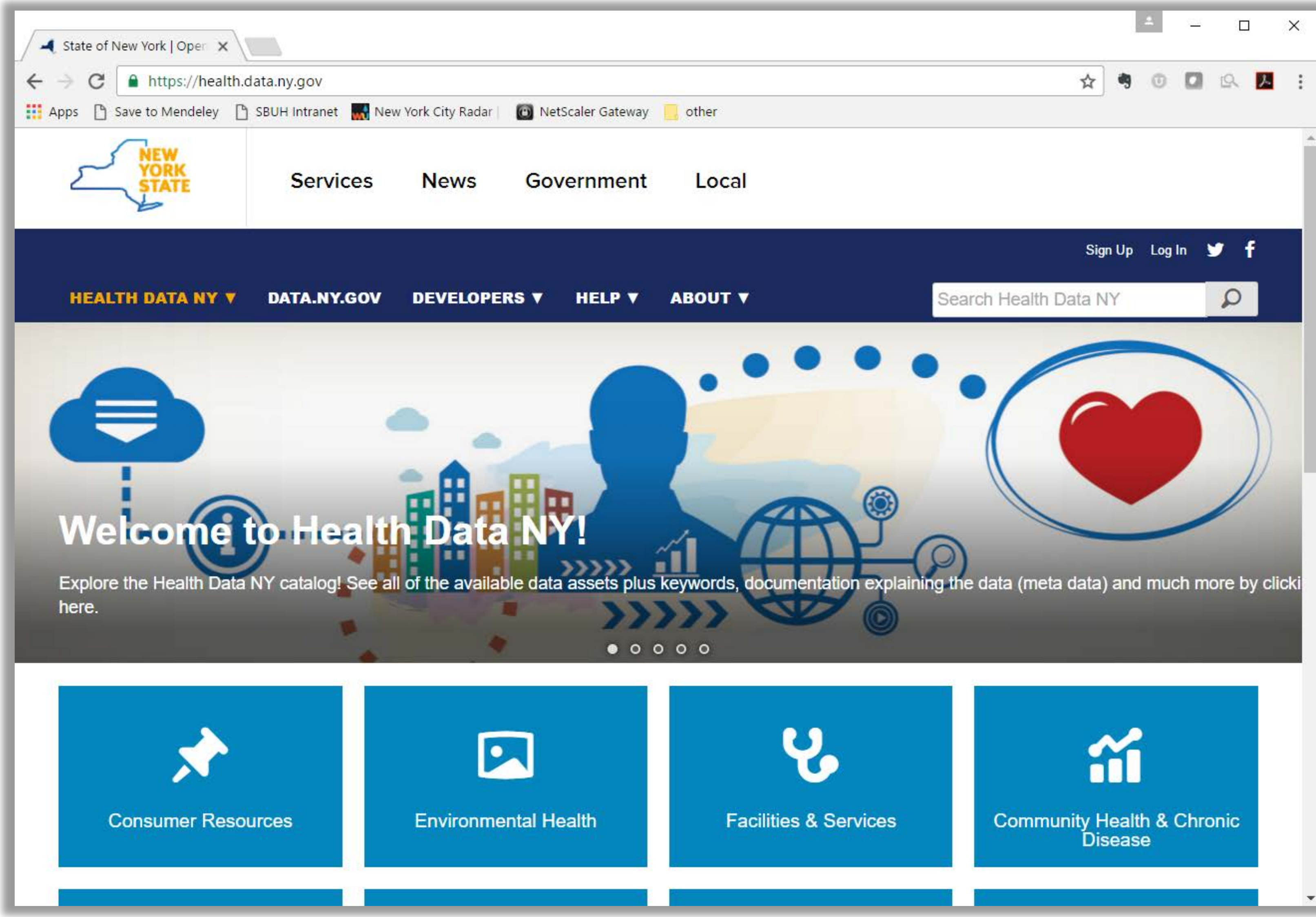
Educating Data Scientists to work with healthcare data

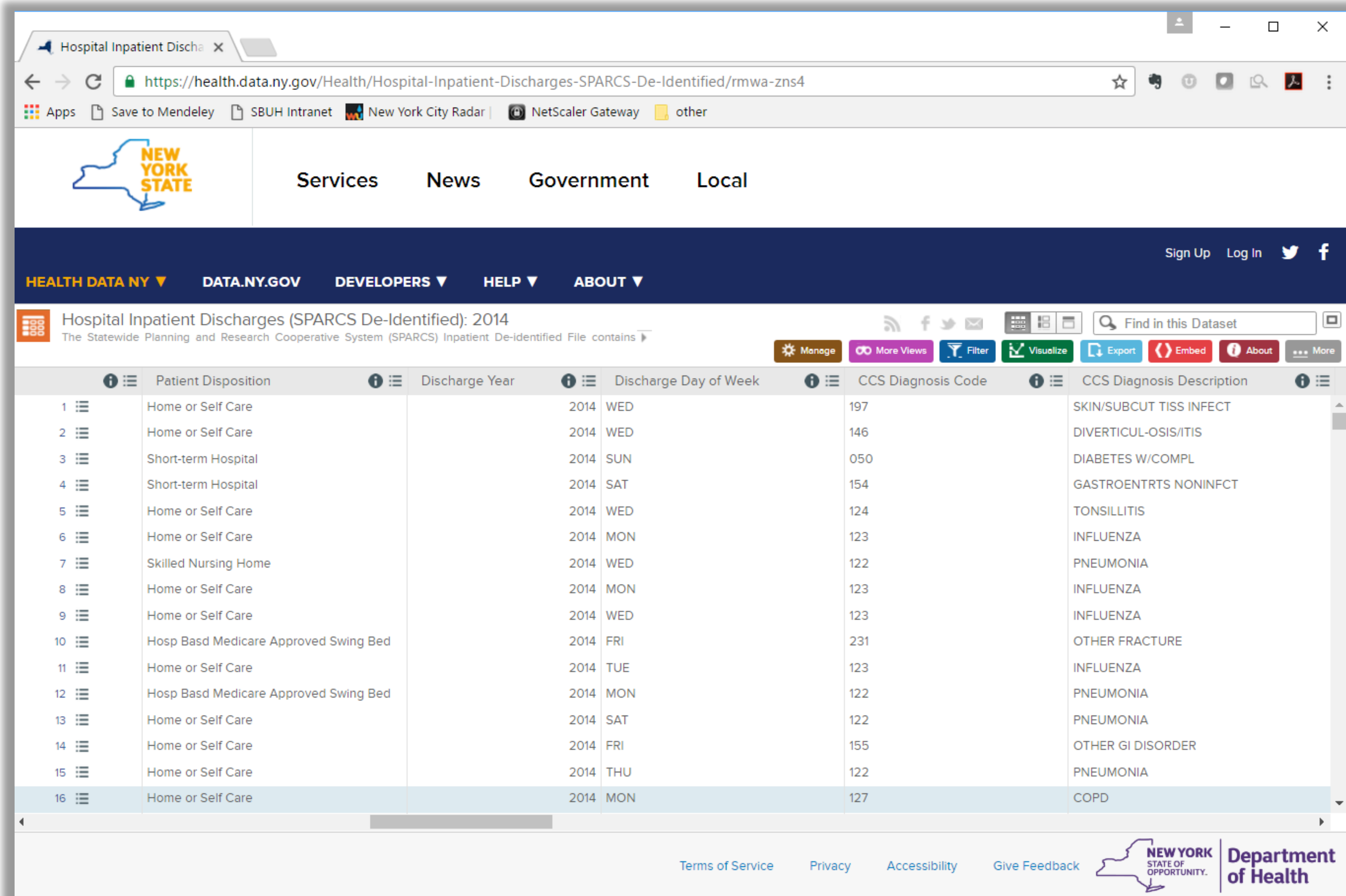
Jupyter Notebooks for Analytic Reproducible Analysis



Example 3: Kidney transplants in NY State

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


State of New York | Open
https://health.data.ny.gov
Apps
Save to Mendeley
SBUH Intranet
New York City Radar
NetScaler Gateway
other
NEW YORK STATE
Services
News
Government
Local
Sign Up
Log In
HEALTH DATA NY
DATA.NY.GOV
DEVELOPERS
HELP
ABOUT
Search Health Data NY
Welcome to Health Data NY!
Explore the Health Data NY catalog! See all of the available data assets plus keywords, documentation explaining the data (meta data) and much more by clicki
here.
Consumer Resources
Environmental Health
Facilities & Services
Community Health & Chronic Disease



Hospital Inpatient Discha

https://health.data.ny.gov/Health/Hospital-Inpatient-Discharges-SPARCS-De-Identified/rmwa-zns4

Apps | Save to Mendeley | SBUH Intranet | New York City Radar | NetScaler Gateway | other

NEW YORK STATE

Services | News | Government | Local

Sign Up | Log In

HEALTH DATA NY ▼ | DATA.NY.GOV | DEVELOPERS ▼ | HELP ▼ | ABOUT ▼

Hospital Inpatient Discharges (SPARCS De-Identified): 2014

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

Find in this Dataset

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
```

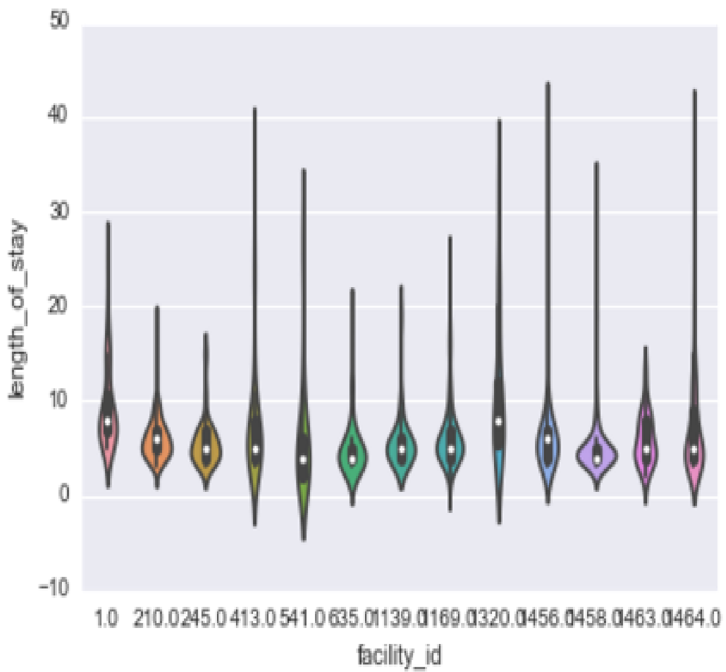
In [30]:
```
kidney_transplants_df.groupby(["facility_name_with_id"])["length_of_stay"].mean()
```

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_sta
y"] <= 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>
```



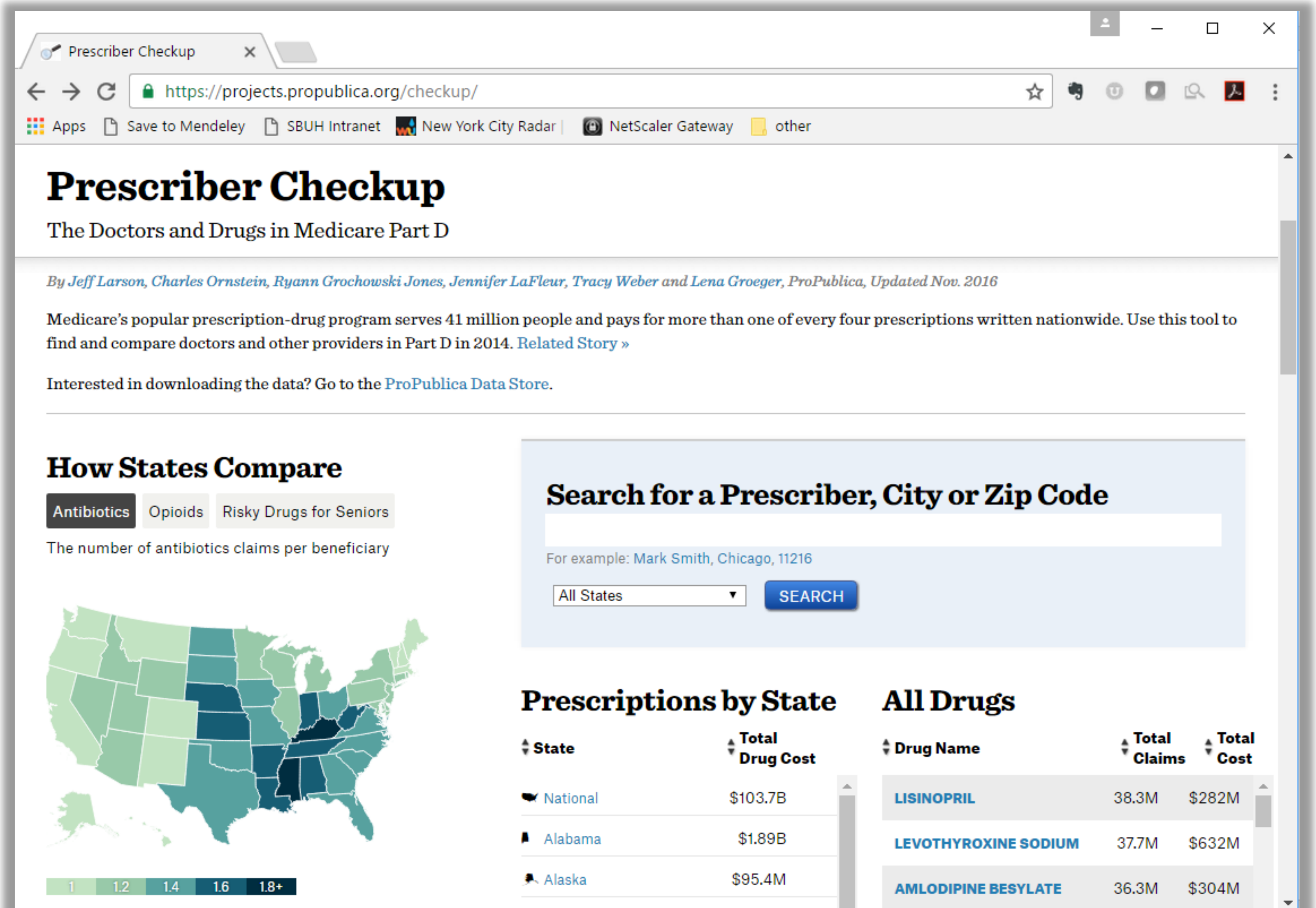
Example 3: Data sources and tools

- 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*

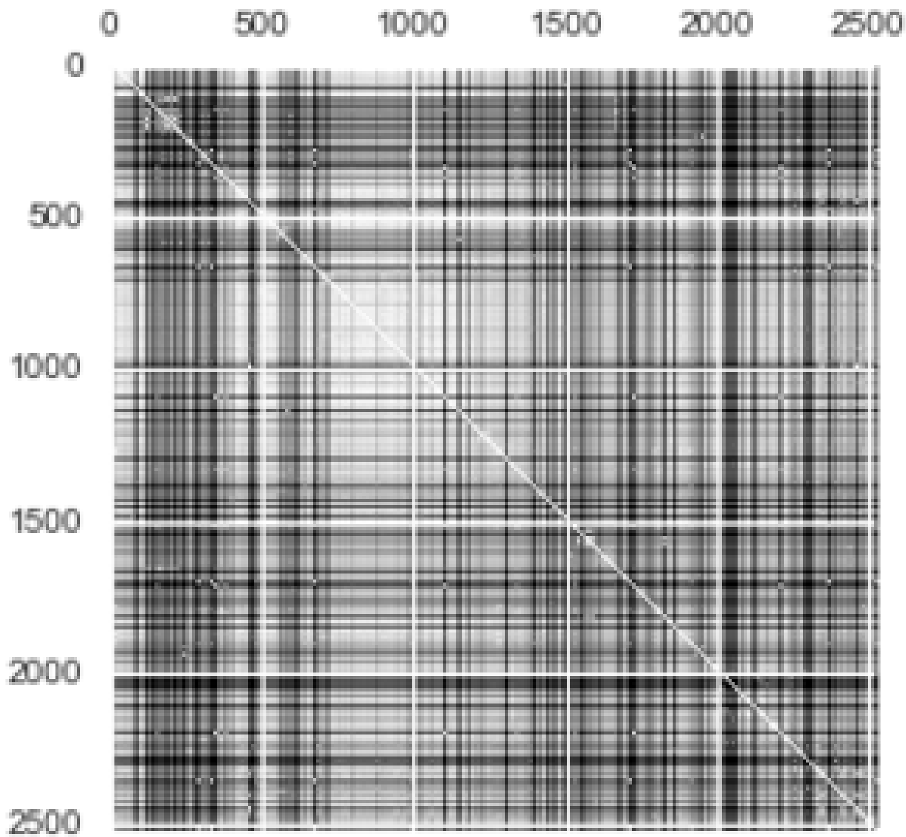
Euclidean distance between providers:

Prescriber 1 and 2 is Sqrt(1) = 1

Prescriber 1 and 3 is Sqrt(6) = 2.44

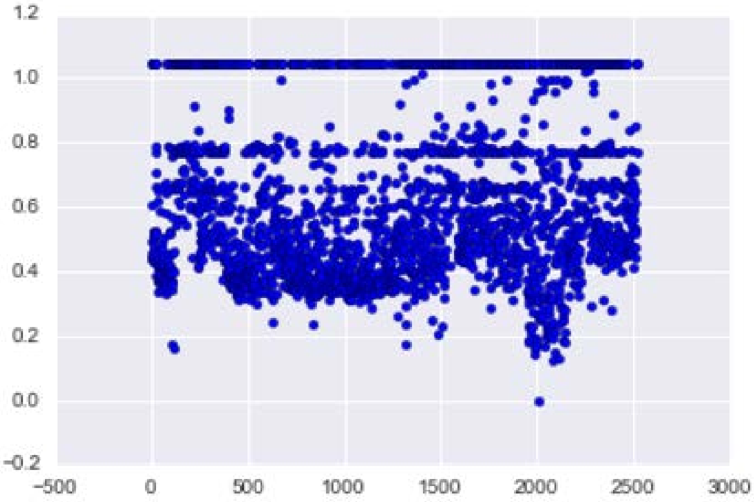
Prescriber 2 and 3 is Sqrt(7) = 2.65

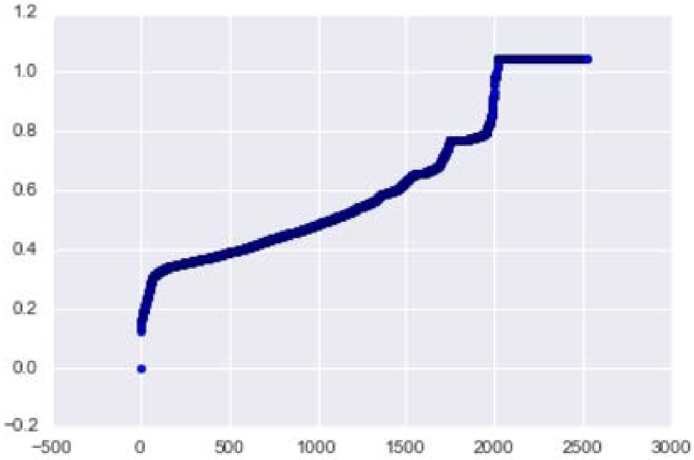
Prescriber Distance Matrix



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
Enumeration date:
2008-06-16

Last update date:
2013-06-21
Current as of:

Is sole proprietor?:
No
Is organization subpart?:
No

In PECOS?:
Yes
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