Data Commons

Context

Data powers everything

- policy,
- journalism,
- health,
- science

How do we make it easier?

Problem is not a shortage of data

Demographics: ACS, Housing Survey, Community Survey,

Economics: BLS, BEA, World Bank, OECD, …

Health: CDC Wonder, CDC Diabetes, County Health, …

Climate: NOAA, Hurricanes, Flooding, …

Genomics: NCBI, ENCODE, ...

Too many formats, schemas, ...

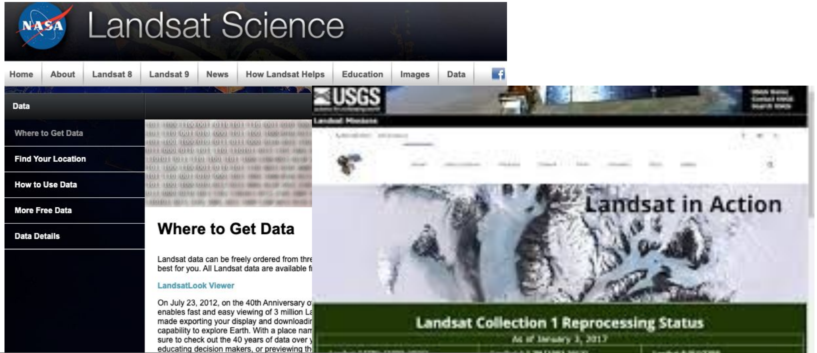
Using Data: current model

Forage for data sources, track down assumptions/provenance

Clean it up, compile data sources, figure out storage, …

High upfront costs, sparse ecosystem, few tools, …

Google maps made satellite imagery part of everyone's life



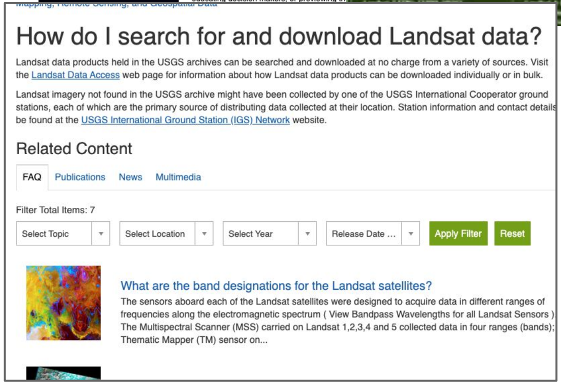
Landsat Science

Where to Get Data

Landsat data can be freely ordered from thre
best for you. All Landsat data are available f

LandsatLook Viewer

On July 23, 2012, on the 40th Anniversary o
enables fast and easy viewing of 3 million La
made exporting your display and downloadin
capability to explore Earth. With a place nam
sure to check out the 40 years of data over y
educating decision makers, or previewing th



How do I search for and download Landsat data?

Landsat data products held in the USGS archives can be searched and downloaded at no charge from a variety of sources. Visit the Landsat Data Access web page for information about how Landsat data products can be downloaded individually or in bulk.

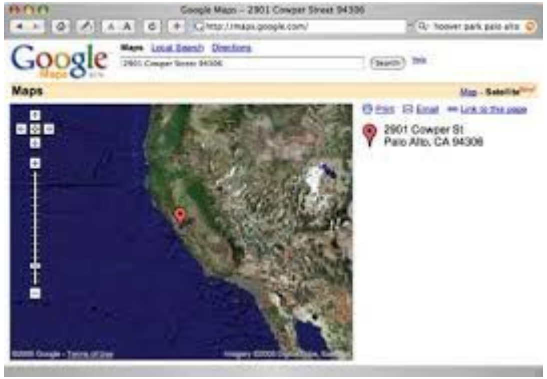
Landsat imagery not found in the USGS archive might have been collected by one of the USGS International Cooperator ground stations, each of which are the primary source of distributing data collected at their location. Station information and contact details be found at the USGS International Ground Station (IGS) Network website.

Related Content

Filter Total Items: 7

What are the band designations for the Landsat satellites?

The sensors aboard each of the Landsat satellites were designed to acquire data in different ranges of frequencies along the electromagnetic spectrum (View Bandpass Wavelengths for all Landsat Sensors) The Multispectral Scanner (MSS) carried on Landsat 1,2,3,4 and 5 collected data in four ranges (bands); Thematic Mapper (TM) sensor on...

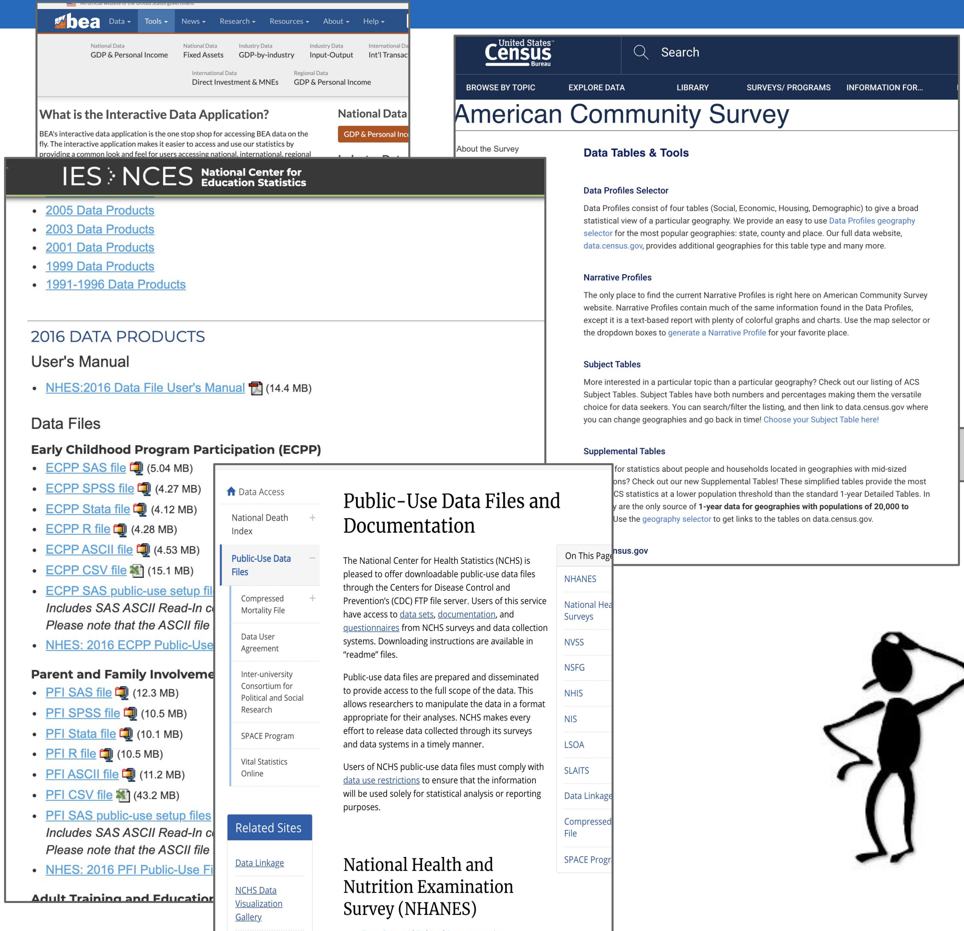


Google Maps – 2901 Cowper Street 94306

2901 Cowper St
Palo Alto, CA 94306

©2005 Google - Terms of Use
Imagery ©2005 DigitalGlobe, EarthSat

We want to do that for data



National Data GDP & Personal Income | National Data Fixed Assets | Industry Data GDP-by-industry | Industry Data Input-Output | International Da Int'l Transac

International Data Direct Investment & MNEs | Regional Data GDP & Personal Income

What is the Interactive Data Application?

BEA's interactive data application is the one stop shop for accessing BEA data on the fly. The interactive application makes it easier to access and use our statistics by providing a common look and feel for users accessing national, international, regional

National Data

GDP & Personal Inc

American Community Survey

About the Survey

Data Tables & Tools

Data Profiles Selector

Data Profiles consist of four tables (Social, Economic, Housing, Demographic) to give a broad statistical view of a particular geography. We provide an easy to use Data Profiles geography selector for the most popular geographies: state, county and place. Our full data website, data.census.gov, provides additional geographies for this table type and many more.

Narrative Profiles

The only place to find the current Narrative Profiles is right here on American Community Survey website. Narrative Profiles contain much of the same information found in the Data Profiles, except it is a text-based report with plenty of colorful graphs and charts. Use the map selector or the dropdown boxes to generate a Narrative Profile for your favorite place.

Subject Tables

More interested in a particular topic than a particular geography? Check out our listing of ACS Subject Tables. Subject Tables have both numbers and percentages making them the versatile choice for data seekers. You can search/filter the listing, and then link to data.census.gov where you can change geographies and go back in time! Choose your Subject Table here!

Supplemental Tables

for statistics about people and households located in geographies with mid-sized ons? Check out our new Supplemental Tables! These simplified tables provide the most CS statistics at a lower population threshold than the standard 1-year Detailed Tables. In y are the only source of **1-year data for geographies with populations of 20,000 to** Use the geography selector to get links to the tables on data.census.gov.

nsus.gov

- 2005 Data Products
- 2003 Data Products
- 2001 Data Products
- 1999 Data Products
- 1991-1996 Data Products

2016 DATA PRODUCTS

User's Manual

- NHES:2016 Data File User's Manual (14.4 MB)

Data Files

Early Childhood Program Participation (ECPP)

- ECPP SAS file (5.04 MB)
- ECPP SPSS file (4.27 MB)
- ECPP Stata file (4.12 MB)
- ECPP R file (4.28 MB)
- ECPP ASCII file (4.53 MB)
- ECPP CSV file (15.1 MB)
- ECPP SAS public-use setup fil *Includes SAS ASCII Read-In c* *Please note that the ASCII file*
- NHES: 2016 ECPP Public-Use

Parent and Family Involveme

- PFI SAS file (12.3 MB)
- PFI SPSS file (10.5 MB)
- PFI Stata file (10.1 MB)
- PFI R file (10.5 MB)
- PFI ASCII file (11.2 MB)
- PFI CSV file (43.2 MB)
- PFI SAS public-use setup files *Includes SAS ASCII Read-In c* *Please note that the ASCII file*
- NHES: 2016 PFI Public-Use Fi

Adult Training and Educatior

Data Access

- National Death Index
- Public-Use Data Files
  - Compressed Mortality File
  - Data User Agreement
  - Inter-university Consortium for Political and Social Research
  - SPACE Program
  - Vital Statistics Online

Related Sites

- Data Linkage
- NCHS Data Visualization Gallery

Public-Use Data Files and Documentation

The National Center for Health Statistics (NCHS) is pleased to offer downloadable public-use data files through the Centers for Disease Control and Prevention's (CDC) FTP file server. Users of this service have access to data sets, documentation, and questionnaires from NCHS surveys and data collection systems. Downloading instructions are available in "readme" files.

Public-use data files are prepared and disseminated to provide access to the full scope of the data. This allows researchers to manipulate the data in a format appropriate for their analyses. NCHS makes every effort to release data collected through its surveys and data systems in a timely manner.

Users of NCHS public-use data files must comply with data use restrictions to ensure that the information will be used solely for statistical analysis or reporting purposes.

National Health and Nutrition Examination Survey (NHANES)

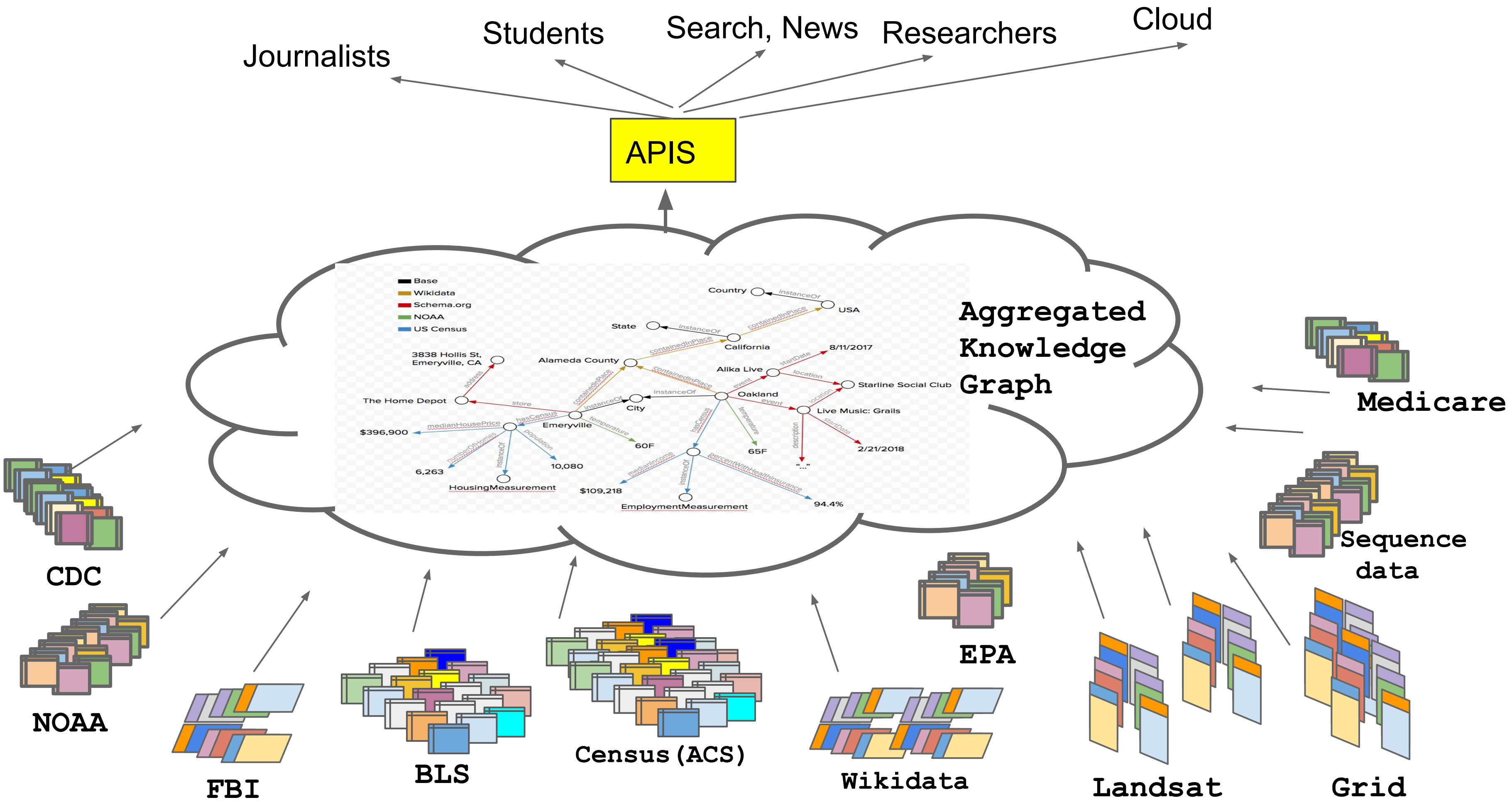
On This Page

- NHANES
- National Hea Surveys
- NVSS
- NSFG
- NHIS
- NIS
- LSOA
- SLAITS
- Data Linkage
- Compressed File
- SPACE Progr

From search for dataset, download, clean, normalize, join …

to

Just ask Google



Journalists

Students

Search, News

Researchers

Cloud

APIS

Aggregated Knowledge Graph

CDC

NOAA

FBI

BLS

Census (ACS)

Wikidata

EPA

Landsat

Grid

Medicare

Sequence data

Data Commons vs collection of datasets

Collections of datasets (ala NIH Data Commons). Solves the problem of finding the dataset.

But the remaining problems --- cleaning, joining … remain

Datacommons --- a single KG built by cleaning, normalizing, joining all these datasets

Data Commons v 0.9

- People, places, …: Integrated view of Census (ACS), CDC, BLS, BEA, FEC, NOAA, DEA, DOL,… on average 5k variables for every state, county, city, zip, school district, congressional district, …
- Education: College Scoreboard, NCES, ...
- Disasters: earthquakes, hurricanes, floods, fires

- Scientific Collections: Bronx Botanical Gardens, ENCODE, parts of NCBI, GTEx

Data Commons v 0.9

Applications for 4 categories

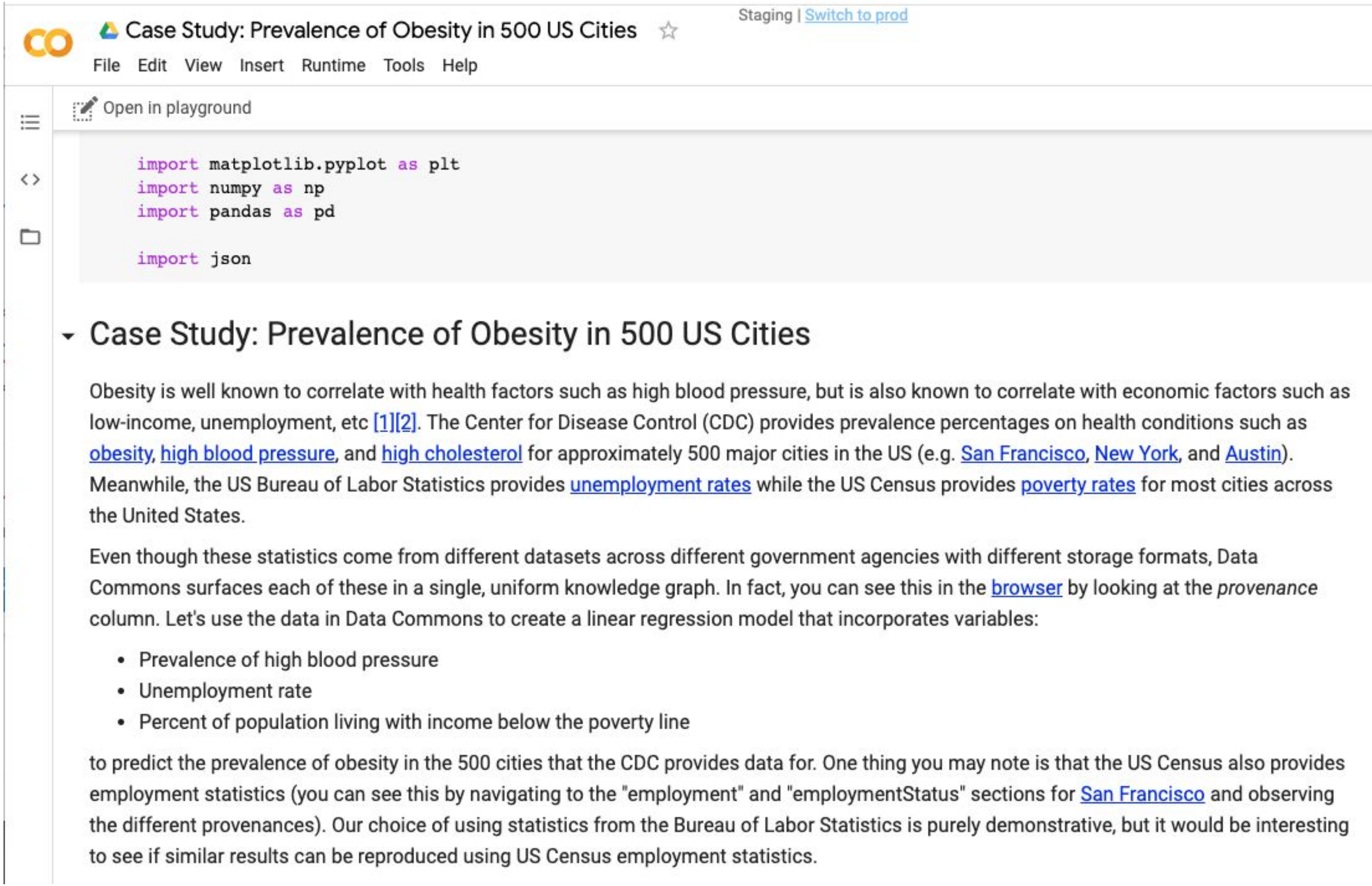
- Researchers
- Students
- Journalists
- Google Users

APIs

APIs in

- REST
- Python / Python Notebooks
- SPARQL
- SQL against Big Query
- Google Sheets

Python Notebooks for students & researchers


CO
Case Study: Prevalence of Obesity in 500 US Cities
Staging | Switch to prod
File Edit View Insert Runtime Tools Help
Open in playground

```
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd

import json
```

Case Study: Prevalence of Obesity in 500 US Cities

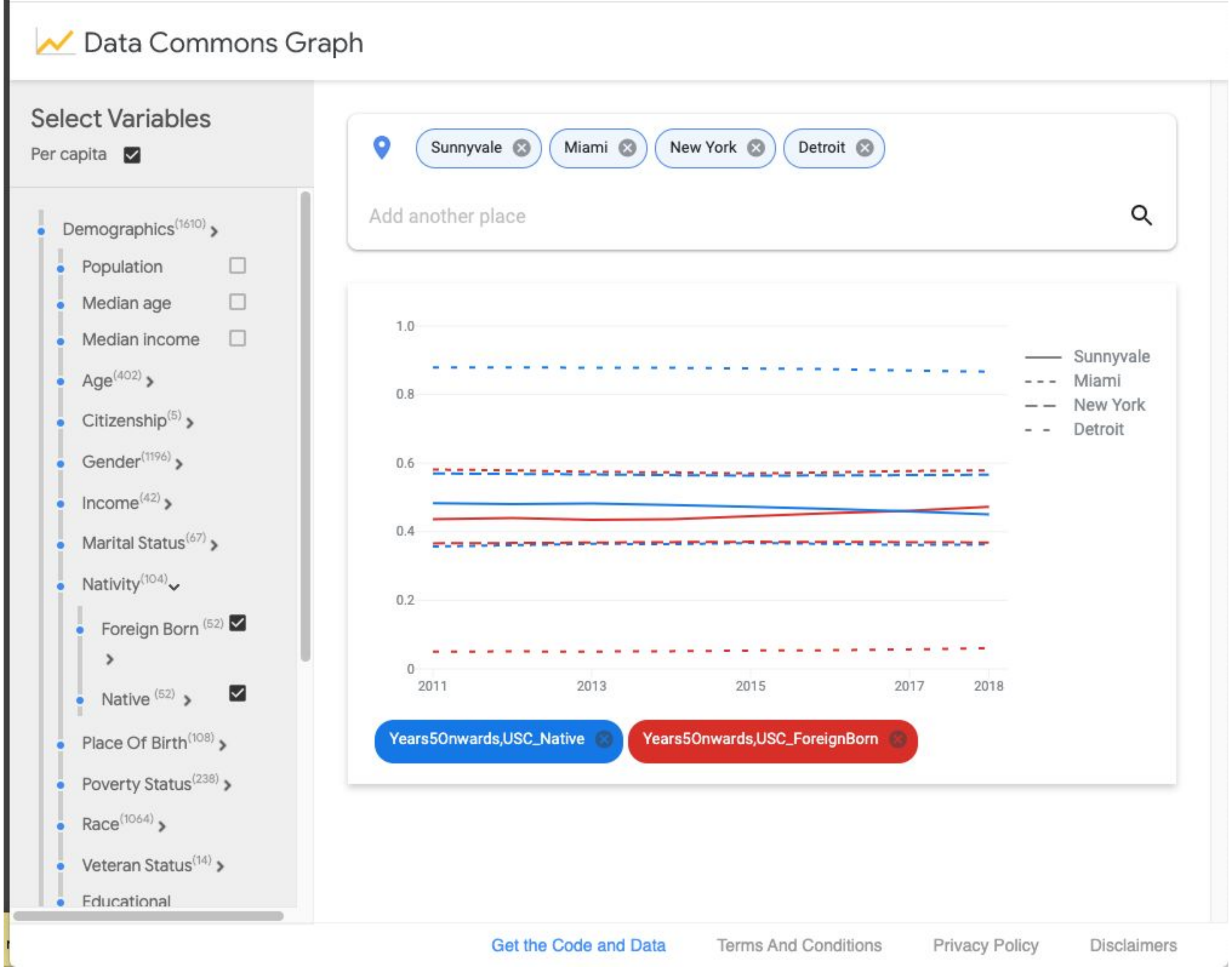
Obesity is well known to correlate with health factors such as high blood pressure, but is also known to correlate with economic factors such as low-income, unemployment, etc [1][2]. The Center for Disease Control (CDC) provides prevalence percentages on health conditions such as obesity, high blood pressure, and high cholesterol for approximately 500 major cities in the US (e.g. San Francisco, New York, and Austin). Meanwhile, the US Bureau of Labor Statistics provides unemployment rates while the US Census provides poverty rates for most cities across the United States.

Even though these statistics come from different datasets across different government agencies with different storage formats, Data Commons surfaces each of these in a single, uniform knowledge graph. In fact, you can see this in the browser by looking at the *provenance* column. Let's use the data in Data Commons to create a linear regression model that incorporates variables:

- Prevalence of high blood pressure
- Unemployment rate
- Percent of population living with income below the poverty line

to predict the prevalence of obesity in the 500 cities that the CDC provides data for. One thing you may note is that the US Census also provides employment statistics (you can see this by navigating to the "employment" and "employmentStatus" sections for San Francisco and observing the different provenances). Our choice of using statistics from the Bureau of Labor Statistics is purely demonstrative, but it would be interesting to see if similar results can be reproduced using US Census employment statistics.

Simple Charting Tool



Data Commons Graph

Select Variables

Per capita ☑

- Demographics(1610)
  - Population □
  - Median age □
  - Median income □
  - Age(402)
  - Citizenship(5)
  - Gender(1196)
  - Income(42)
  - Marital Status(67)
  - Nativity(104)
    - Foreign Born (52) ☑
    - Native (52) ☑
  - Place Of Birth(108)
  - Poverty Status(238)
  - Race(1064)
  - Veteran Status(14)
  - Educational

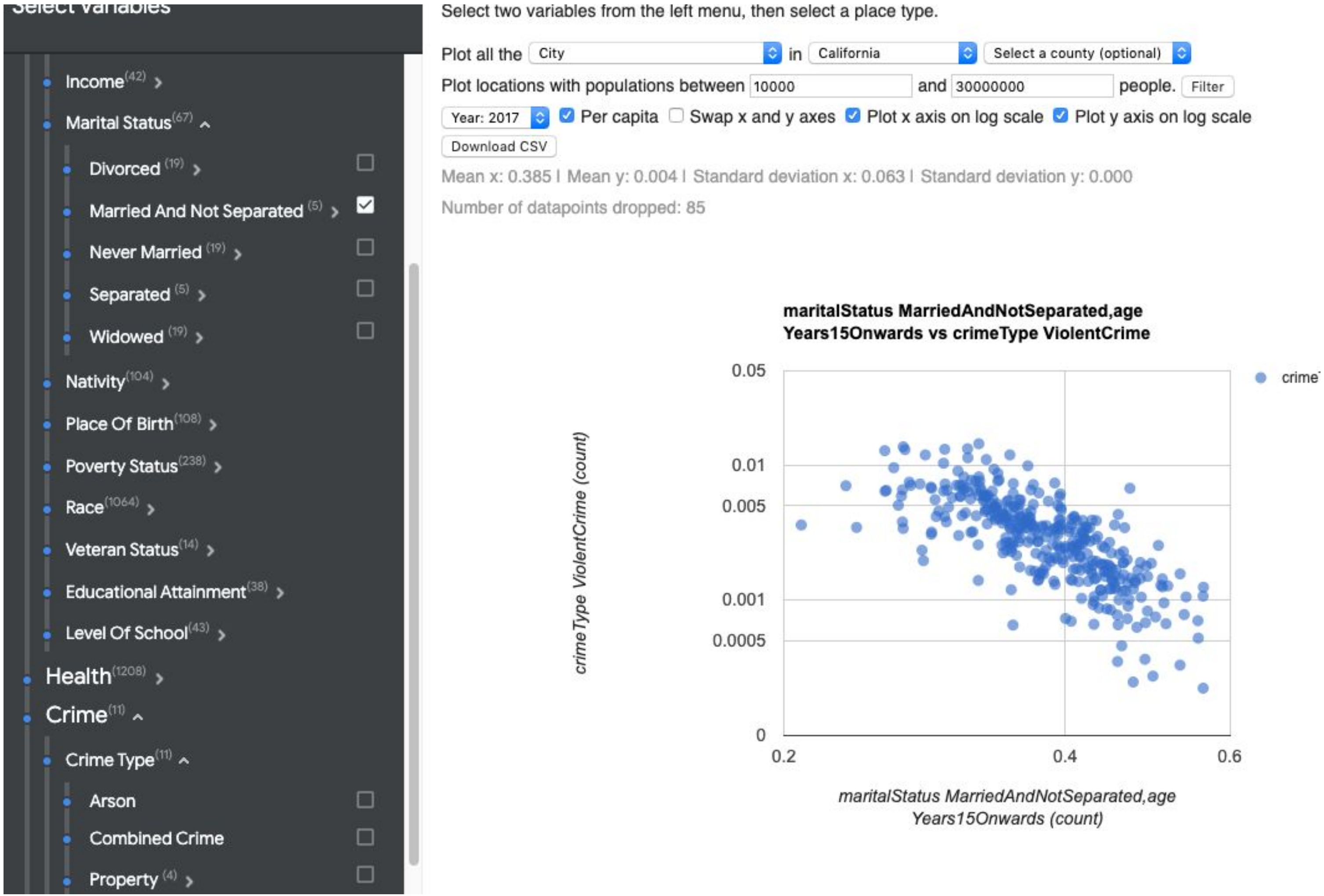
Sunnyvale ⊗ Miami ⊗ New York ⊗ Detroit ⊗

Add another place

Years5Onwards,USC_Native ⊗

Years5Onwards,USC_ForeignBorn ⊗

Simple Correlation Tool



Select two variables from the left menu, then select a place type.

Plot all the City in California Select a county (optional)

Plot locations with populations between 10000 and 30000000 people. Filter

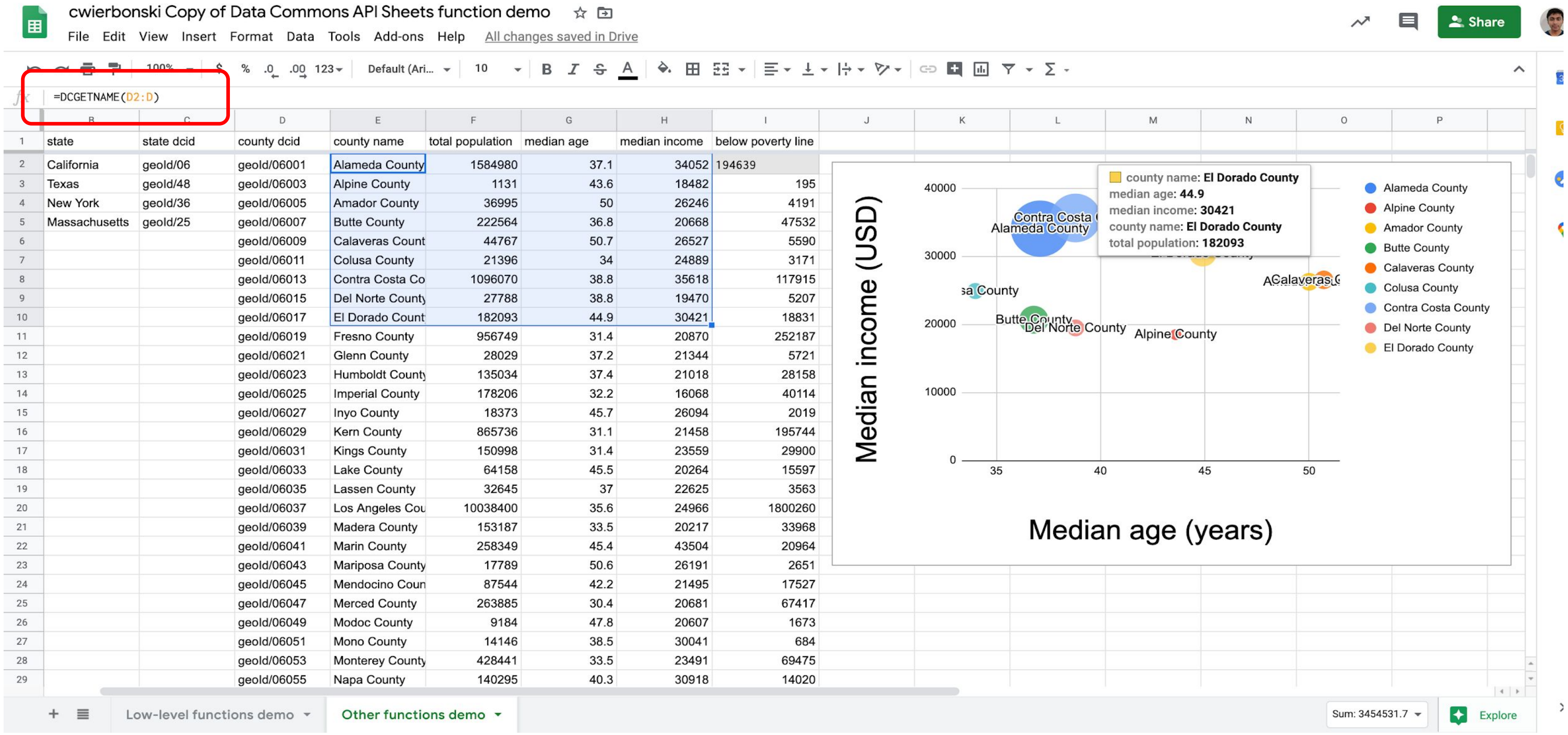
Year: 2017 ☑ Per capita ☐ Swap x and y axes ☑ Plot x axis on log scale ☑ Plot y axis on log scale

Download CSV

Mean x: 0.385 | Mean y: 0.004 | Standard deviation x: 0.063 | Standard deviation y: 0.000

Number of datapoints dropped: 85

Google Sheets API



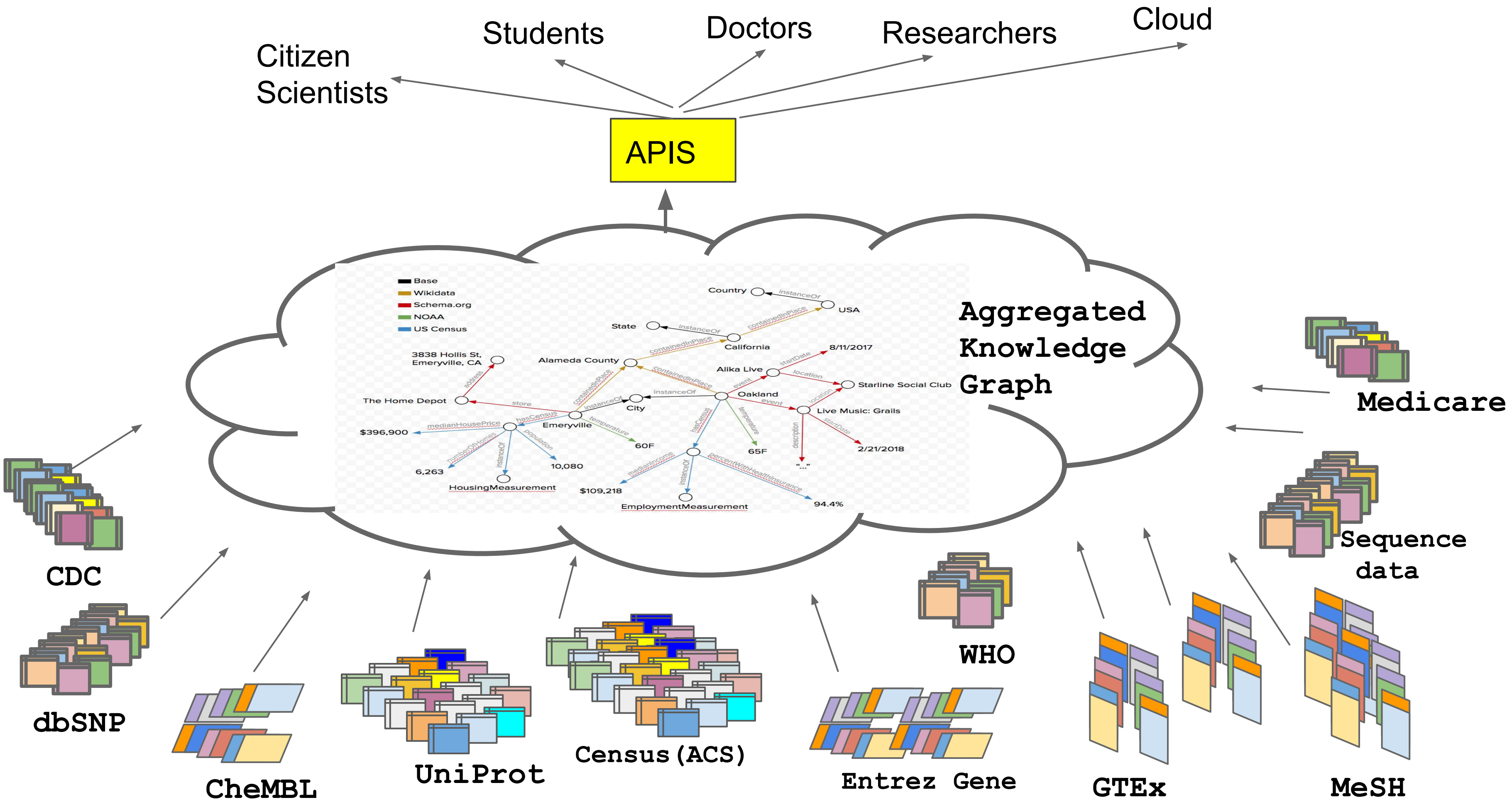
cwierbonski Copy of Data Commons API Sheets function demo

```
=DCGETNAME(D2:D)
```

	B	C	D	E	F	G	H	I
1	state	state dcid	county dcid	county name	total population	median age	median income	below poverty line
2	California	geoId/06	geoId/06001	Alameda County	1584980	37.1	34052	194639
3	Texas	geoId/48	geoId/06003	Alpine County	1131	43.6	18482	195
4	New York	geoId/36	geoId/06005	Amador County	36995	50	26246	4191
5	Massachusetts	geoId/25	geoId/06007	Butte County	222564	36.8	20668	47532
6			geoId/06009	Calaveras Count	44767	50.7	26527	5590
7			geoId/06011	Colusa County	21396	34	24889	3171
8			geoId/06013	Contra Costa Co	1096070	38.8	35618	117915
9			geoId/06015	Del Norte County	27788	38.8	19470	5207
10			geoId/06017	El Dorado Count	182093	44.9	30421	18831
11			geoId/06019	Fresno County	956749	31.4	20870	252187
12			geoId/06021	Glenn County	28029	37.2	21344	5721
13			geoId/06023	Humboldt County	135034	37.4	21018	28158
14			geoId/06025	Imperial County	178206	32.2	16068	40114
15			geoId/06027	Inyo County	18373	45.7	26094	2019
16			geoId/06029	Kern County	865736	31.1	21458	195744
17			geoId/06031	Kings County	150998	31.4	23559	29900
18			geoId/06033	Lake County	64158	45.5	20264	15597
19			geoId/06035	Lassen County	32645	37	22625	3563
20			geoId/06037	Los Angeles Cou	10038400	35.6	24966	1800260
21			geoId/06039	Madera County	153187	33.5	20217	33968
22			geoId/06041	Marin County	258349	45.4	43504	20964
23			geoId/06043	Mariposa County	17789	50.6	26191	2651
24			geoId/06045	Mendocino Coun	87544	42.2	21495	17527
25			geoId/06047	Merced County	263885	30.4	20681	67417
26			geoId/06049	Modoc County	9184	47.8	20607	1673
27			geoId/06051	Mono County	14146	38.5	30041	684
28			geoId/06053	Monterey County	428441	33.5	23491	69475
29			geoId/06055	Napa County	140295	40.3	30918	14020

🟨 county name: **El Dorado County**
median age: **44.9**
median income: **30421**
county name: **El Dorado County**
total population: **182093**

Biomedical Data Commons



Citizen Scientists

Students

Doctors

Researchers

Cloud

APIS

Aggregated Knowledge Graph

Medicare

Sequence data

CDC

dbSNP

CheMBL

UniProt

Census(ACS)

Entrez Gene

WHO

GTEx

MeSH

Biomedical Data Commons Datasets

- CDC - 500 Cities
- CDC - Diabetes Atlas
- CDC - Wonder
- ChEMBL*
- ClinVar
- dbSNP
- Dartmouth Medicare Atlas

*Source: UCSF SPOKE

- Disease Ontology*
- FDA - Pharmacologic Class*
- GTEx
- ENCODE
- Entrez Gene
- MeSH*
- NY Times - COVID-19 Cases + Deaths

- SIDER*
- SPOKE*
- UCSC Genome Browser
- US Census - SAHIE
- UniProt*
- WHO - COVID-19 Cases + Deaths
- WHO - ICD-10 Codes

Eg: Extract 3 data points on given variants

Given a list of genetic variants

Find:

1. Clinical Significance
2. Functional Category
3. Significant Gene Associations in Whole Blood

Current Methods to find information

Option 1:
Site Search

Use needed database browser and search all variants individually

Option 2:
Download and Analyze Data

Download data, read it into memory, and parse it for the needed info programmatically

Option 3:
Data Commons

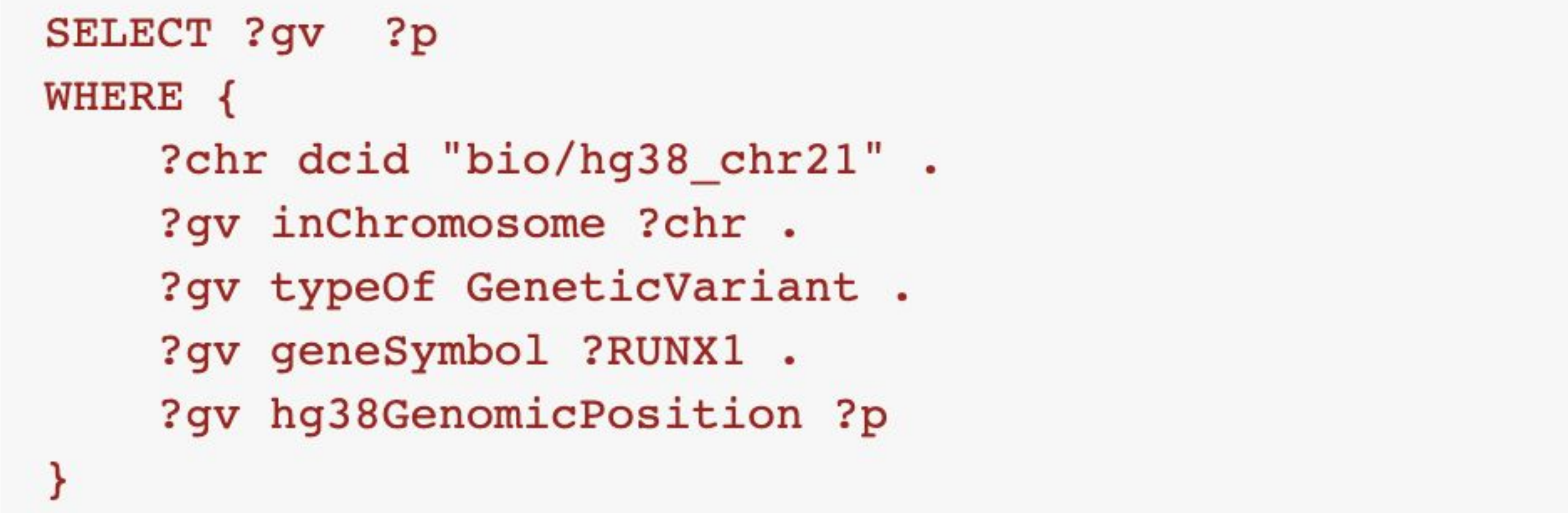
Run 1 query on Data Commons

More on Option 2

- I want to find out more about a list of genetic variants
- Download ClinVar and analyze clinical significance of the variants - **a few hours**
- Download dbSNP and identify the type of genomic region these SNPs are located - **a couple of days**
- Download GTEx and identify which genes a genetic variant is significantly associated in a given tissue - **a week**

Example Data Commons Query

Issue Query to Data Commons



```
SELECT ?gv  ?p
WHERE {
    ?chr dcid "bio/hg38_chr21" .
    ?gv inChromosome ?chr .
    ?gv typeOf GeneticVariant .
    ?gv geneSymbol ?RUNX1 .
    ?gv hg38GenomicPosition ?p
}
```

Time to run query: 21 seconds

Many Other Commons ...

- Economics
- Covid
- Energy

....

Technical Challenges

- Representation
- Infrastructure
- Inference

- General thoughts on KGs ...

Representational Issues

Representing time, geo, measures, provenance, ...

Stat. aggregates exacerbate the problem of long predicates: maleLatinoPopulationUnder...

Triples are easy to understand, relatively widely understood

But real systems need more expressiveness

Infrastructure

Data Commons has about 50B triples

Bio part is another 100B triples

Wide range of queries and latency requirements

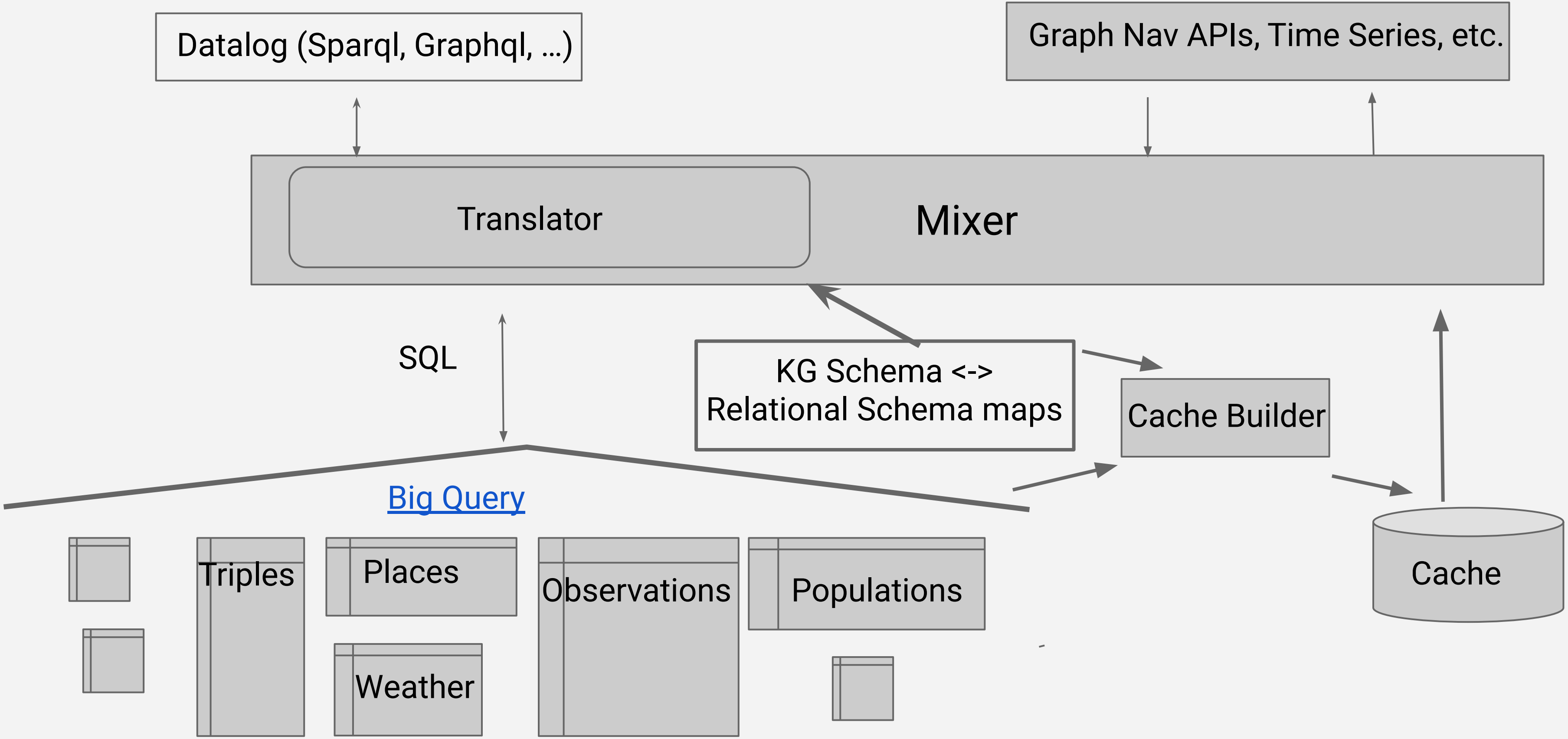
- simple queries that require millisecond responses
- complex sparql queries
- everything in between

Data Commons approach

Multiple ‘backing’ stores, that support different classes of queries with different latencies

- Big Query … in memory hash tables
- KG is a view on top of more native encodings
- DLG + function terms + contexts + simple kinds of inference is the real ‘language’ --- serves as the ‘epistemological/knowledge level’

Serving System



Datalog (Sparql, Graphql, …)
Graph Nav APIs, Time Series, etc.
Translator
Mixer
SQL
KG Schema <-> Relational Schema maps
Cache Builder
Big Query
Triples
Places
Observations
Populations
Weather
Cache

Ending thoughts

- Vocabulary creep: Google KG, Cyc, Wikidata all have many tens of thousands of ‘schema’ terms. Human language manages with few thousand terms. How do we bring the compositionality of NL to KR?

- The problems of old AI haven’t been solved. They just can’t be expressed clearly in today’s ML formalisms. e.g., World’s tallest mountain yesterday?