

SDMX Meets Data Commons:

Integrating SDMX Semantics into an Open and Large Data Commons Ecosystem

Speakers:

Luis G. González Morales (UNSD)

Jehangir Amjad (Google)

Agenda

- Data Commons
- Global SDG Database
- SDMX \longleftrightarrow Data Commons (Schema.org) Interop

Every Year, Billions Spent Collecting and Sharing Data

A Myriad of Data Sources



eurostat



UDISE



Every Year, Billions Spent Collecting and Sharing Data

A Myriad of
Data Sources



This Data Is
Essential for:





-  Science
-  Journalism
-  Policy
-  Law
-  Academia
-  Users

Every Year, Billions Spent Collecting and Sharing Data

A Myriad of Data Sources



This Data Is Essential for:

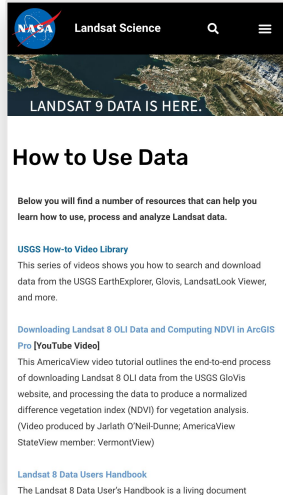
-  Science
-  Journalism
-  Policy
-  Law
-  Academia
-  Users

Using this Data Is Extremely Painful

- Forage for data
- Track down assumptions
- Clean
- Normalize
- Join
- Costly data wrangling

Analogy: Landsat Imagery vs. Google Maps

NASA Landsat / Satellite Imagery Complex / Inaccessible



NASA Landsat Science

LANDSAT 9 DATA IS HERE.

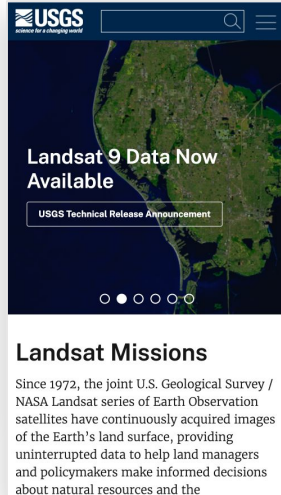
How to Use Data

Below you will find a number of resources that can help you learn how to use, process and analyze Landsat data.

USGS How-to Video Library
This series of videos shows you how to search and download data from the USGS EarthExplorer, Glovis, LandsatLook Viewer, and more.

Downloading Landsat 8 OLI Data and Computing NDVI in ArcGIS Pro [YouTube Video]
This AmericaView video tutorial outlines the end-to-end process of downloading Landsat 8 OLI data from the USGS GloVis website, and processing the data to produce a normalized difference vegetation index (NDVI) for vegetation analysis. (Video produced by Jarlath O'Neil-Dunne, AmericaView StateView member, VermontView)

Landsat 8 Data Users Handbook
The Landsat 8 Data User's Handbook is a living document



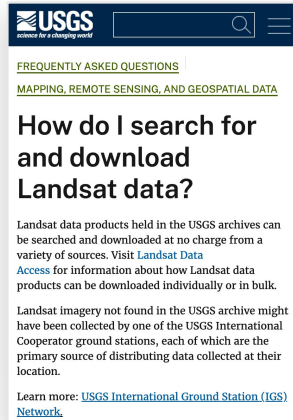
USGS

Landsat 9 Data Now Available

USGS Technical Release Announcement

Landsat Missions

Since 1972, the joint U.S. Geological Survey / NASA Landsat series of Earth Observation satellites have continuously acquired images of the Earth's land surface, providing uninterrupted data to help land managers and policymakers make informed decisions about natural resources and the



USGS

FREQUENTLY ASKED QUESTIONS
MAPPING, REMOTE SENSING, AND GEOSPATIAL DATA

How do I search for and download Landsat data?

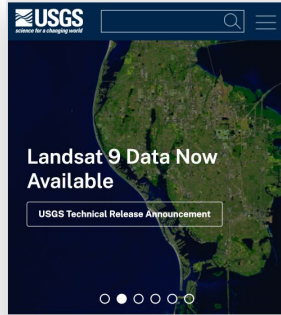
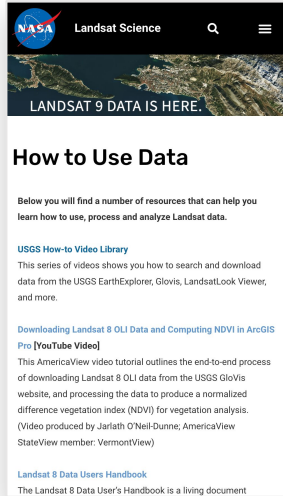
Landsat data products held in the USGS archives can be searched and downloaded at no charge from a variety of sources. Visit [Landsat Data Access](#) for information about how Landsat data products can be downloaded individually or in bulk.

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.

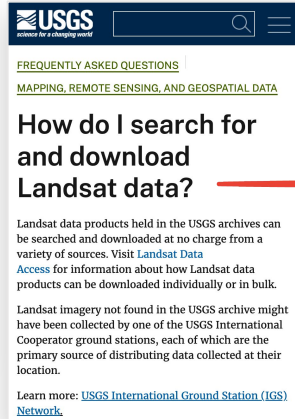
Learn more: [USGS International Ground Station \(IGS\) Network](#).

Analogy: Landsat Imagery vs. Google Maps

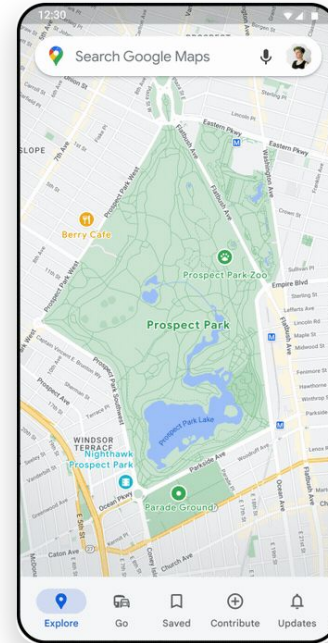
NASA Landsat / Satellite Imagery
Complex / Inaccessible



Landsat Missions
Since 1972, the joint U.S. Geological Survey / NASA Landsat series of Earth Observation satellites have continuously acquired images of the Earth's land surface, providing uninterrupted data to help land managers and policymakers make informed decisions about natural resources and the



Google Maps
Accessible / New Paradigm



Our Goal for Data

From Search for Datasets,
Download, Clean, Normalize, Join...

IES/NCES National Center for Education Statistics

- 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 (0.04 MB)
- ECPP SPSS file (4.27 MB)
- ECPP State file (4.12 MB)
- ECPP R file (4.28 MB)
- ECPP ASCII file (4.53 MB)

WHO UNEP

IPCC Plenary
IPCC Bureau
Executive Committee

Working Group II
Impacts, Adaptation, and Vulnerability
TSU

Working Group III
Mitigation of Climate Change
TSU

Authors, Contributors, Reviewers

Bureau of Economic Analysis

What is the Interactive Data Application?

Public-Use Data Files and Documentation

National Health and Medical Research Council

Public-Use Data Files and Documentation

United States Census Bureau

American Community Survey

Data Tables & Tools

Data Profiles Selector

Narrative Profiles

Subject Tables

Our Goal for Data

From Search for Datasets,
Download, Clean, Normalize, Join...

IES/NCES National Center for Education Statistics

- 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

- ECPP SAS file (0.04 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)

IPCC Plenary Executive Committee

Working Group I: Impacts, Adaptation, and Vulnerability

Working Group II: Mitigation of Climate Change

Working Group III: Systemic Risks

Task Force on High Level Aspects of Systemic Risks

Authors, Contributors, Reviewers

BEA Interactive Data Application

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/or manipulate data by providing common tools and features across national, international, regional and industry releases. The application makes the data more accessible, easy to use and export. The underlying features are robust and visually appealing. The application also allows for data sharing with others via a number of social tools.

BEA updates its data in near real time. During BEA news releases there might be a slight delay in accessing the most recent data but access to supplemental data files is always available.

The interactive data application program is available on both PC and Mac. The application uses a "table" browser experience common to many e-commerce and other standard Web sites. Navigating between datasets and accessing and changing query parameters is easy as they are similar across all datasets.

Video: Navigating BEA Interactive Data

Bureau of Economic Analysis

Public-Use Data Files and Documentation

The National Center for Health Statistics (NCHS) presents other downloadable public-use data files through the Center for Disease Control and Prevention's (CDC) FTP file server. Users of this service have access to data on: demographics and characteristics from birth surveys and data collection systems. Downloading instructions are available in "Tutorial" 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 manner appropriate for their analysis. NCHS makes every effort to release data collected through surveys and data systems in a timely manner.

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

Related Sites

National Health and

United States Census Bureau

American Community Survey

Data Tables & Tools

Data Profiles Selector

Data Profiles consist of four tables (Social, Economic, Housing, and Demographic) for a particular geography. We provide an easy selector for the most popular geographies: state, county and data.census.gov, provides additional geographies for this tab.

Narrative Profiles

The only place to find the current Narrative Profiles is right here on this website. Narrative Profiles contain much of the same information as the Subject Tables, but with a text-based report with plenty of colorful graphs and charts. The dropdown boxes to generate a Narrative Profile for your geography.

Subject Tables

More interested in a particular topic than a particular geographic area? Subject Tables have both numbers and percent choice for data seekers. You can search/filter the listing, and you can change geographies and go back in time! Choose your geography and year.

Supplemental Tables

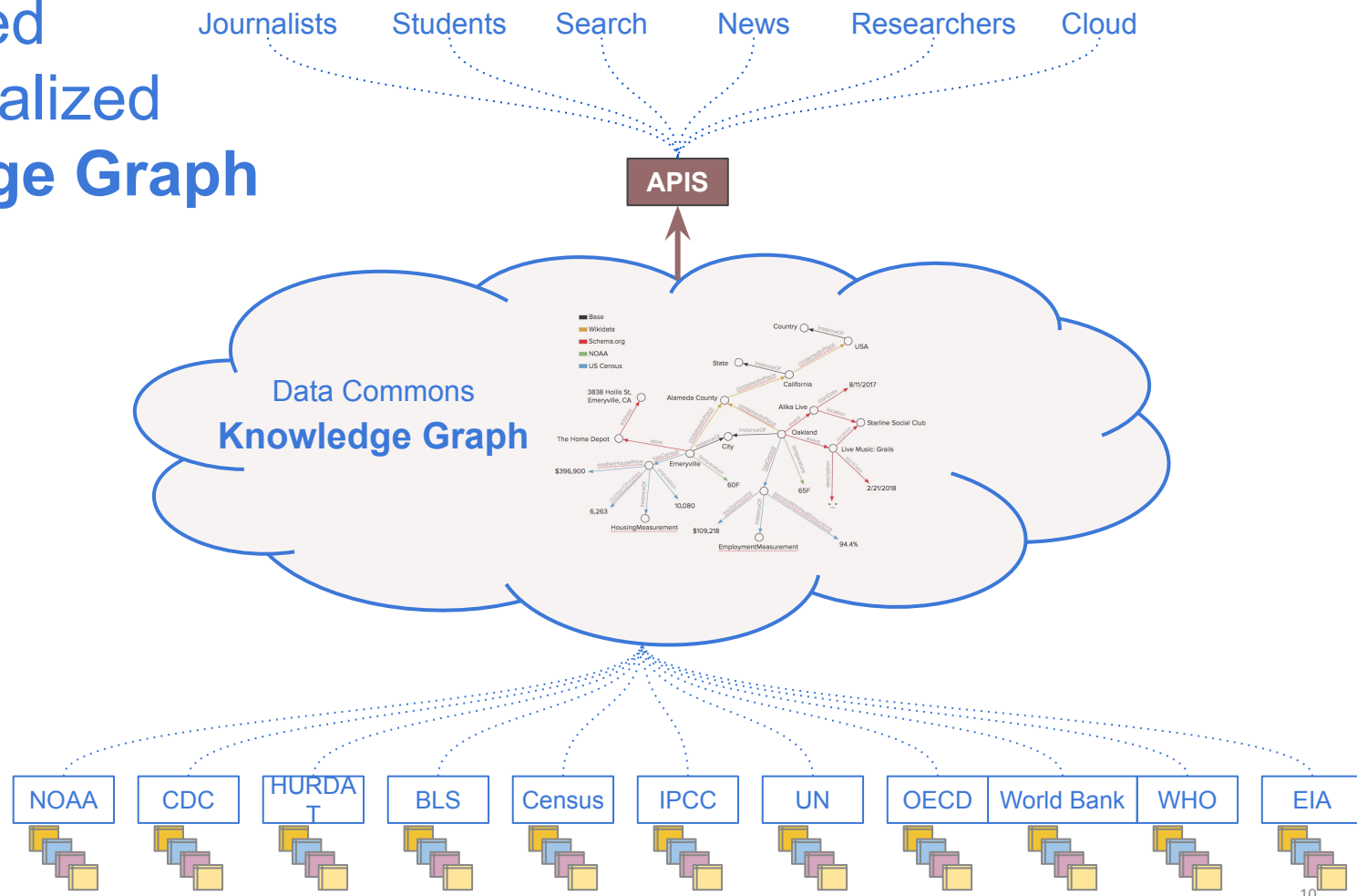
Looking for statistics about people and households located in a particular geographic area? Supplemental Tables! These recent ACS statistics at a lower population threshold than the fact files, they are the only source of 1-year data for geographies with a population of 64,999. Use the geography selector to get links to the tables and data.census.gov.

Data Commons Goal
Just Ask in Natural Language

What California counties are most at food risk from climate change?



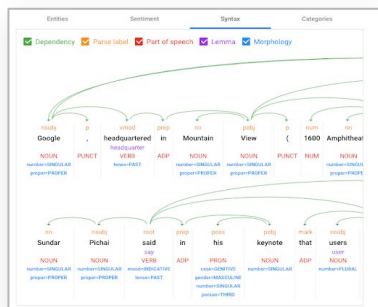
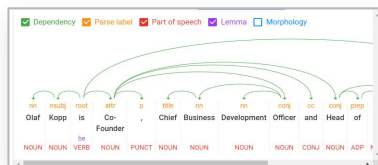
Aggregated and Normalized Knowledge Graph



Target Audiences

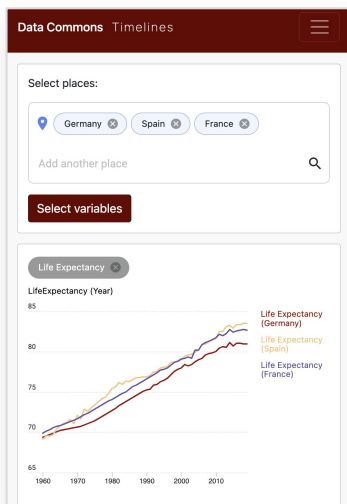
Every Day Consumers

Natural language interface in search, in KPs



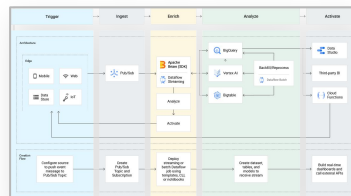
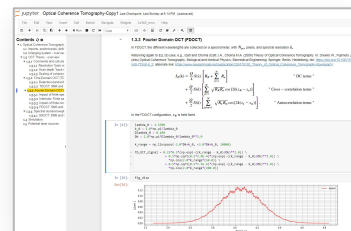
Policy and Journalists

Dashboards/visualizations on datacommons.org



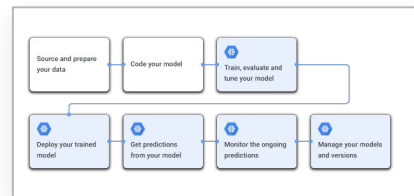
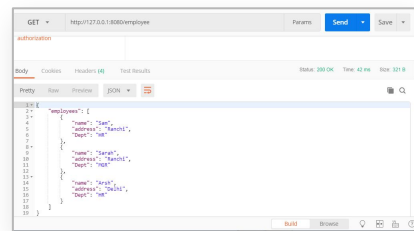
Application Builders

REST APIs, Python, Python Notebooks, BigQuery



AI / ML Modelers

Algorithms + Compute + **DATA**



What We Built

Data Commons Code

Open Source
GCP infrastructure for
creating, storing,
serving, KG.

Visualization tools

NL interface to data

Ability to Extend to
Custom Data Commons
Knowledge Graphs

Open Data

Demographics
Census (US, India, ...),
Eurostat...

Economics
BLS, BEA, WorldBank...

Health
CDC, DEA, WHO, ICD...

Climate
IPCC, EPA, HURDAT,
NOAA

Energy
EIA, NREL...

Food, Crime, Education,
Elections, Trade...

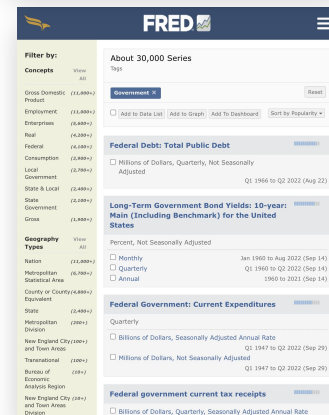
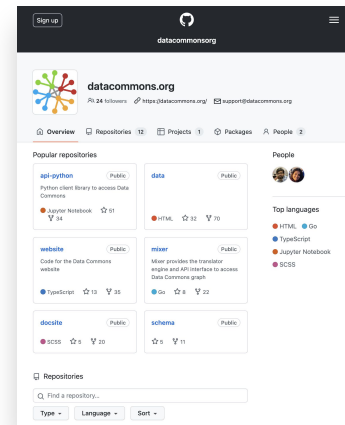
Data Commons Snapshot

3.5B+ time series

3.3M+ places

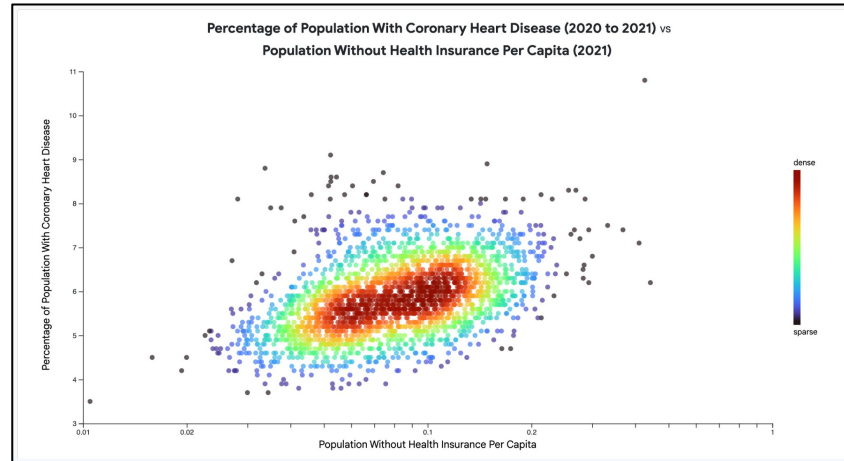
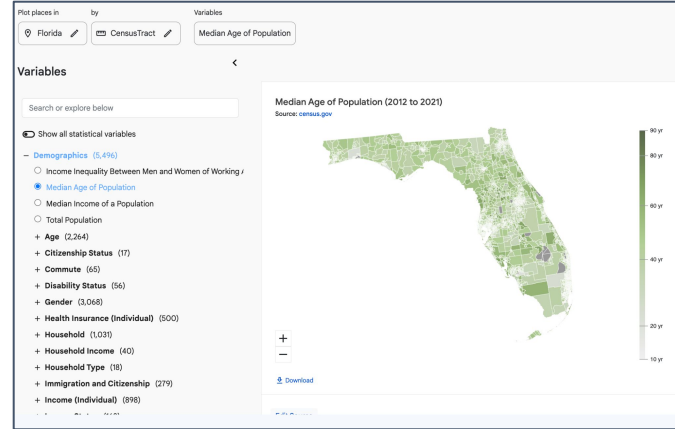
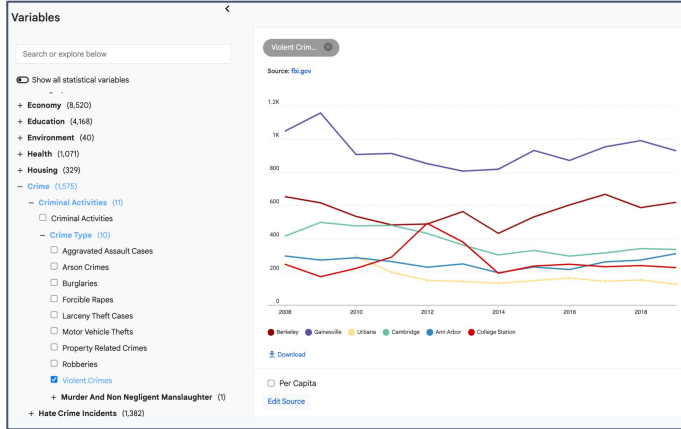
130K+ variables

4x the size of FRED



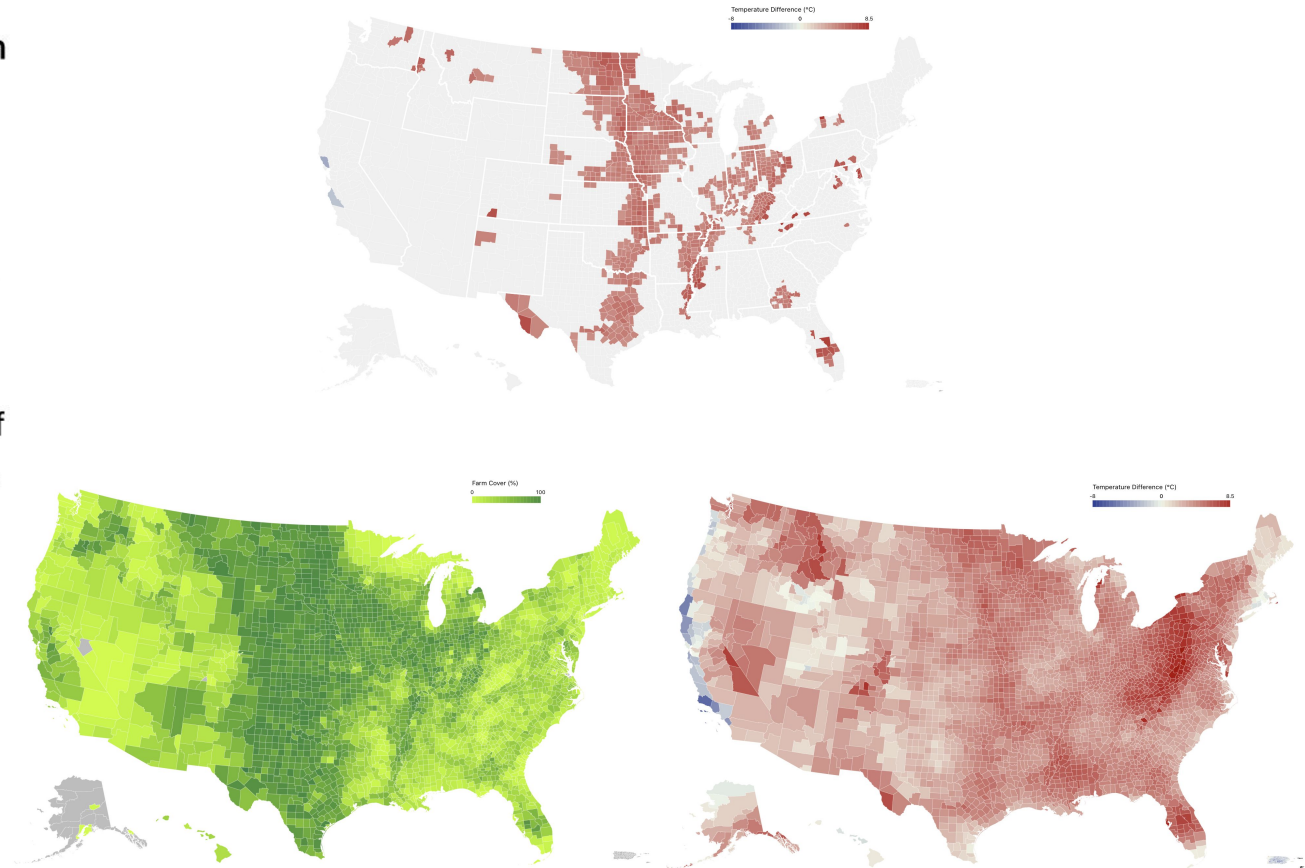
Examples

Visualization Tools



Climate Change x Farm Cover

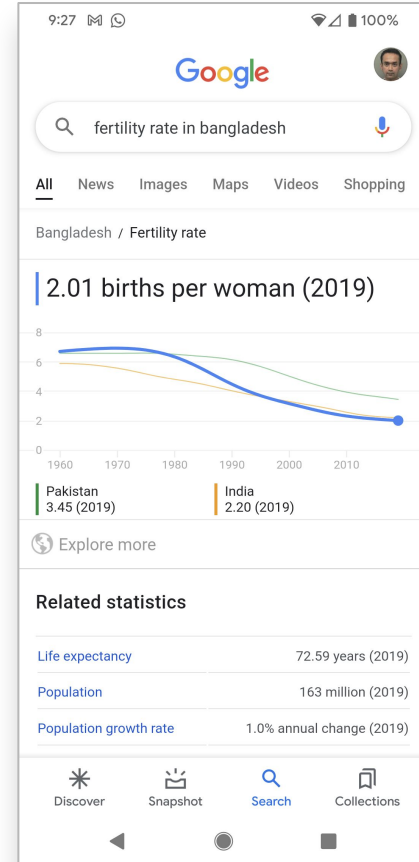
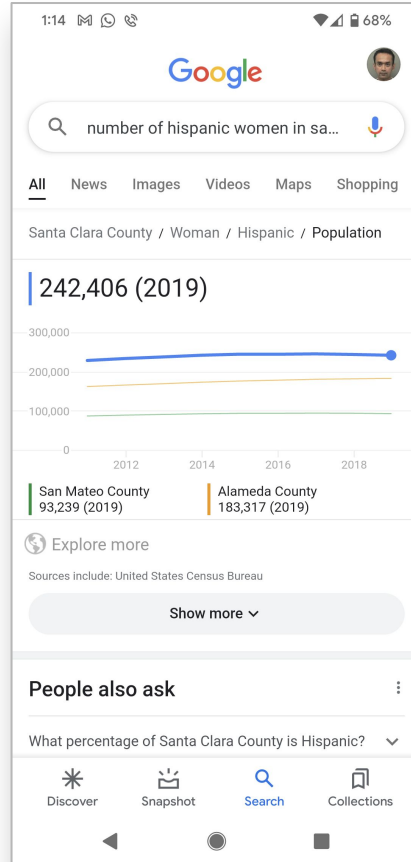
To illustrate farming regions in the US which may be most impacted by climate change, the map now highlights counties with more than 50% farm cover that have extreme projected temperature differences (increase of more than 4 degrees or decrease of more than 2 degrees Celsius).



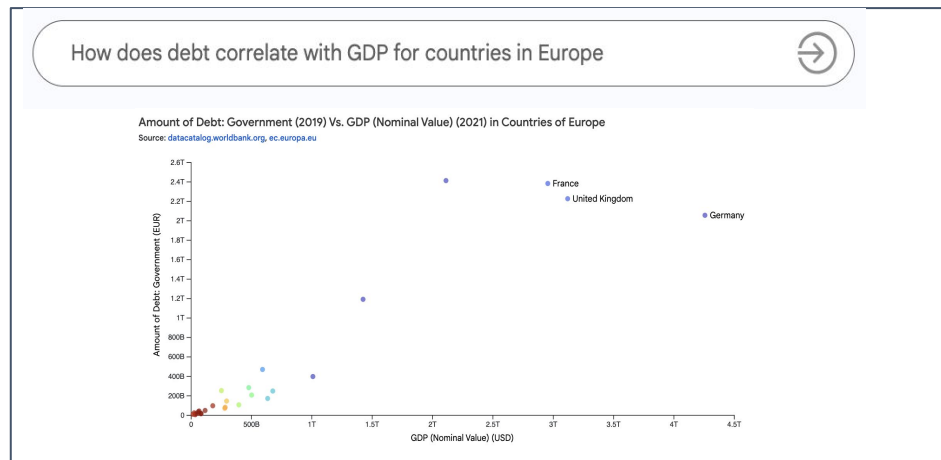
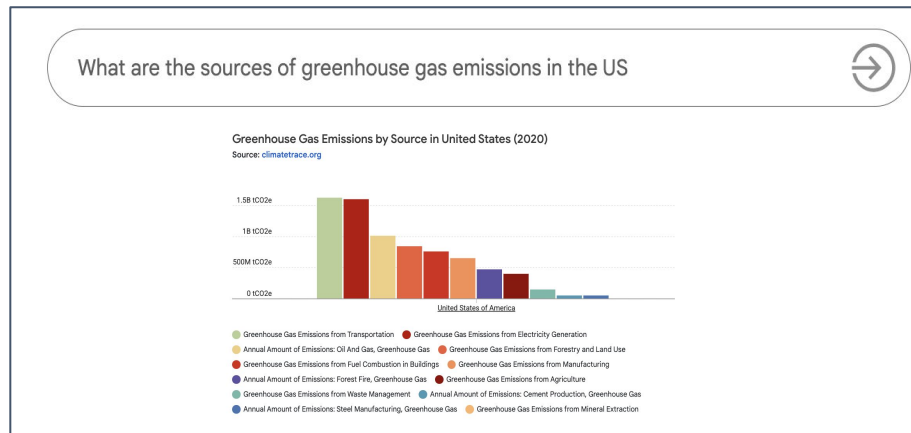
Natural Language (NL) Interface in Google Search

A Few Sample Search Queries

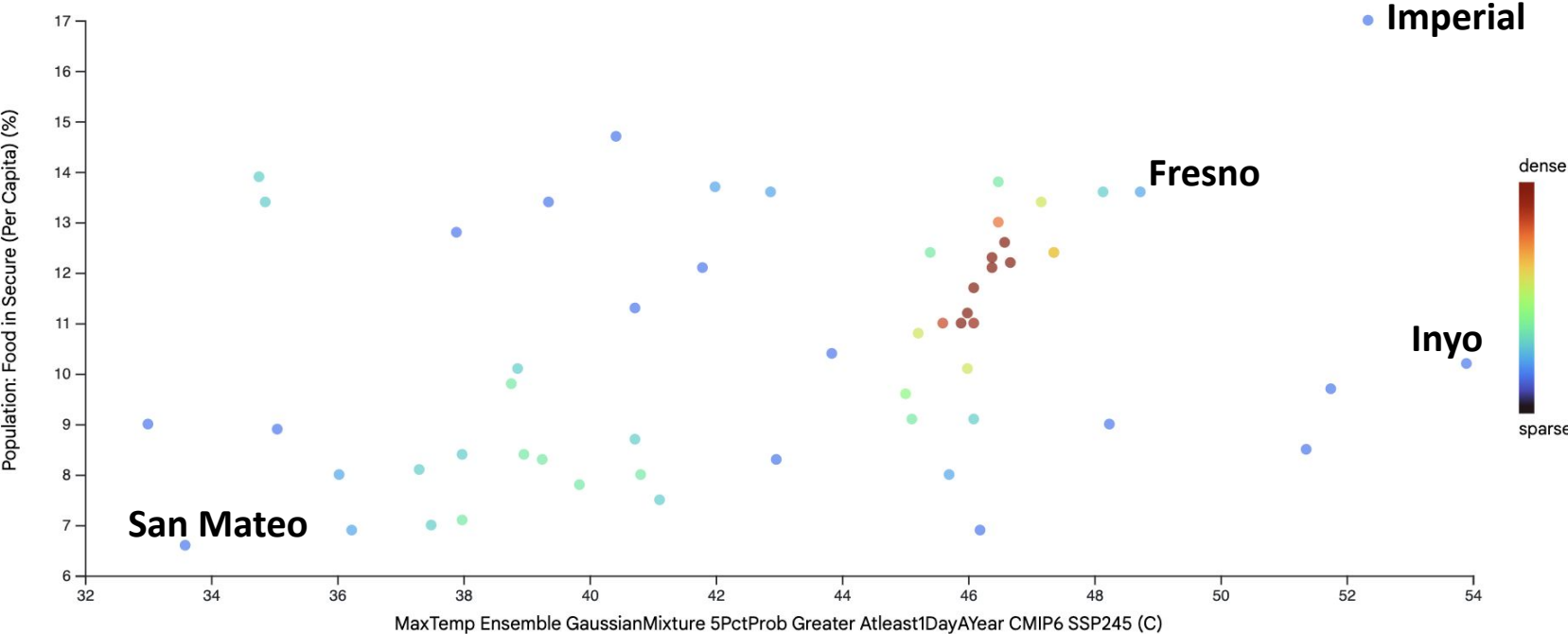
- *“Energy use per capita in India”*
- *“CO2 emissions in Sweden”*
- *“Number of unemployed in California”*
- *“Population growth rate in Germany”*
- *“Fertility rate in Bangladesh”*
- *“Number of poor Hispanic women in Santa Clara”*



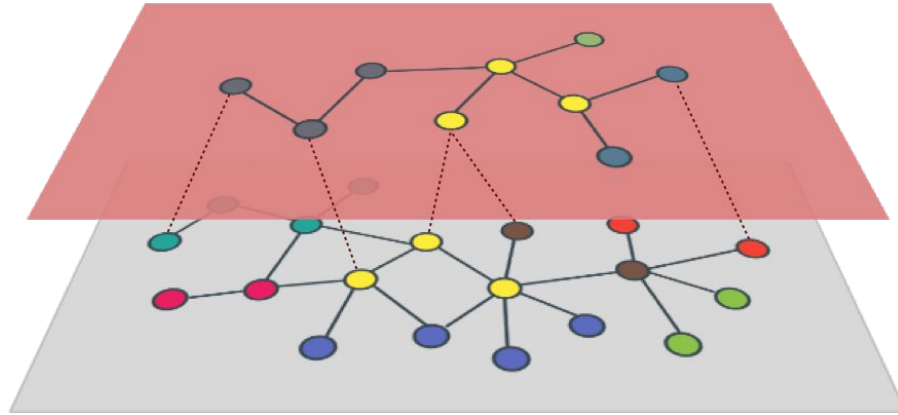
Just Ask (NL)



California Counties Most At Risk From Rising Temperatures



Many Data Commons, One Schema, One API



- Different topics, different ACLs, free vs paid, ...
- An overlay, on top of which both the overlay and base Data Commons can be accessed with the same, single API — overlaid data could be private or semi-public or ...



Home

Countries / Areas

Goals

Search



UN Data Commons

for the SDGs

Introducing the new UN Data Commons for the SDGs – a platform integrating authoritative SDG data and information resources from across the UN System into a public repository with advanced search functionality and a modern, user-friendly interface.

Explore SDG Data by Countries or Areas

Select a country or area

Learn about country and SDG region progress on the UN SDGs through the UN Data Commons.

UN Data Commons for the SDGs

- Developed in partnership between UNSD and Google Data Commons
- Establish explicit and implicit links to external resources, making data more easily findable, searchable, and usable.
- Build applications that efficiently access related data across multiple domains using linked open data techniques
- Generate insights by reasoning over complex relationships
- Incrementally add new data and evolve the data schema to accommodate new data types and new use cases.

The Global SDG Database

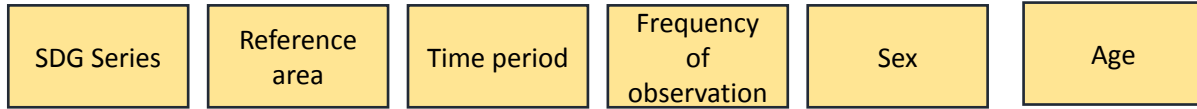
- Cross-domain repository of SDG-related data from many different sources.
- Maintained by UN Statistics Division in collaboration with 40+ custodian agencies across the UN System
- As of October 2023, data for more than 200 unique indicators, with more than 2 million data records
- Covers both country-level data and regional aggregates
- Key role in facilitating SDG data sharing, accessibility, and transparency.

SDMX Model for the SDGs

- Presents all relevant data in a simple, self-contained tabular view
- Each data point is characterized by
 - **Measures**: Observed values on one or more variables interest
 - **Dimensions**: Set of uniquely identifying characteristics
 - **Attributes**: Set of additional characteristics that further describe it

SDMX for the SDGs

Dimensions:



Proportion of population below international poverty line (SI_POV_DAY1)

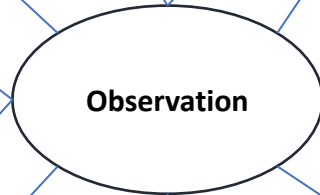
World (1)

2019

Annual (A)

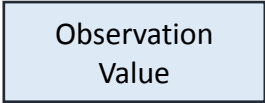
Total or no breakdown (T)

Total or no breakdown (T)



8.5

Measure:



(A)
Normal value

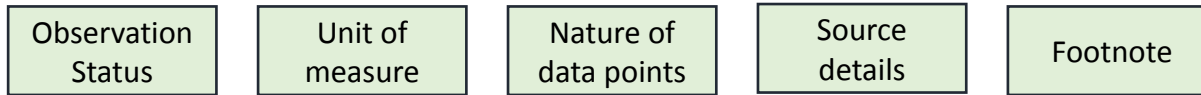
(PERCENT)
Percentage

(G)
Global monitoring data

“Poverty and Inequality Portal, World Bank”

“Accessed March 21, 2023”

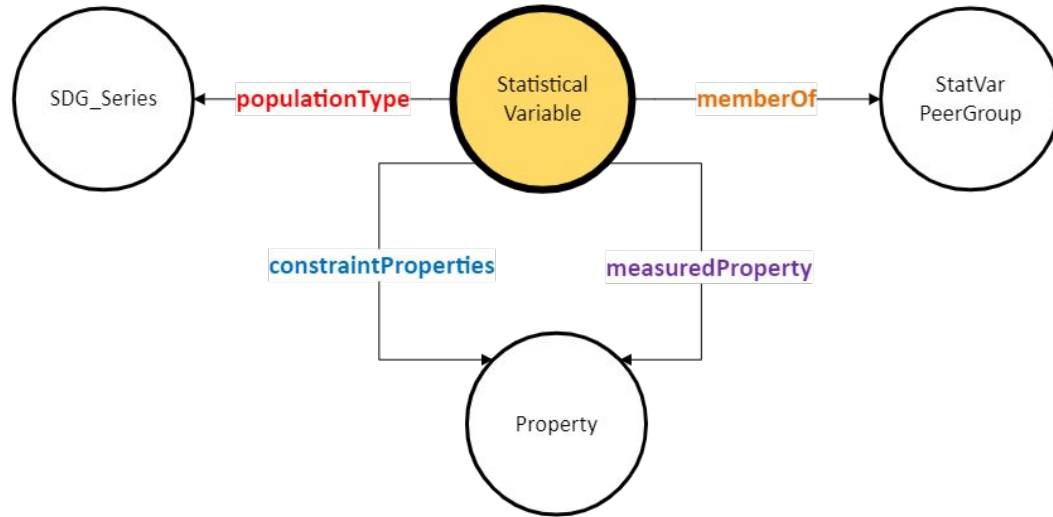
Attributes:



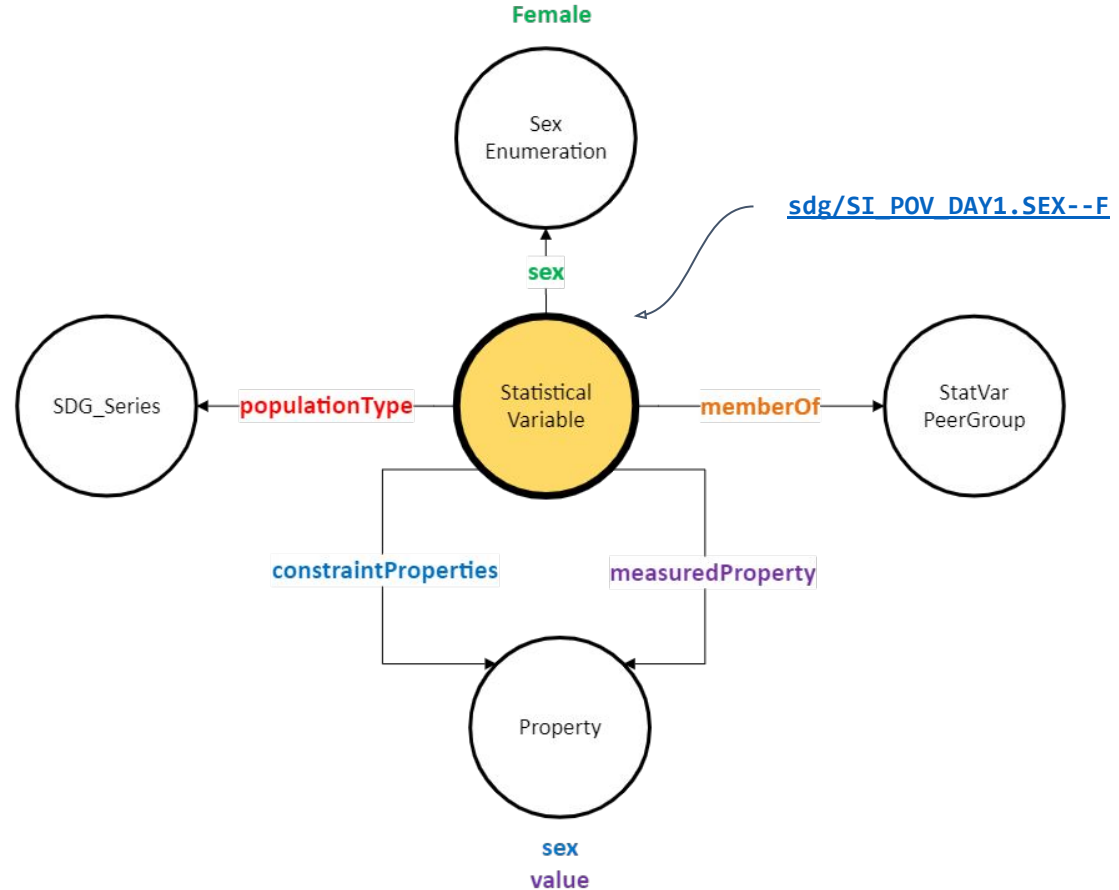
Data Commons Schema

- The model captures the following elements of a data point:
 - **Entity**: The object or thing being measured
 - **Variable**: The specific measurement being taken
 - **Observation**: The value of the variable for a given entity
 - **Provenance**: The source of the data

DC Model for Statistical Variables



DC Model for Statistical Variables



From SDG “Series” to “Statistical Variable”

An indicator “Proportion of population below international poverty line” (SI_POV_DAY1) that is disaggregated only by sex, can be split into 3 slices or “Statistical Variables”, namely:

SDMX Series	SDMX slice definition	population type	statistical variable	constraint property	sex	member of
SI_POV_DAY1	SEX=“M”	SDG_SI_POV_DAY1	sdg/SI_POV_DAY1.SEX--M	Sex	Male	dc/g/SDGSIPOVDAY1_sdgsex
SI_POV_DAY1	SEX=“F”	SDG_SI_POV_DAY1	sdg/SI_POV_DAY1.SEX--F	Sex	Female	dc/g/SDGSIPOVDAY1_sdgsex
SI_POV_DAY1	SEX=“_T”	SDG_SI_POV_DAY1	sdg/SI_POV_DAY1			dc/g/SDGSIPOVDAY1
Type:		SDG_Series	StatisticalVariable	Property		StatVarGroup

Importing SDG Datasets into Data Commons

- Mapping SDG entities to Data Commons entities
- Map CSV content to graph model via Template MCF (TMCF) file

Example:

Mapping Between SDMX Code Lists and DC Enumerations

subject_id	subject_label	predicate_id	object_id	object_label
sdg-geography:4	Afghanistan	skos:exactMatch	dc:country/AFG	Afghanistan
sdg-geography:248	Åland Islands	skos:exactMatch	dc:country/AFG	Åland Islands
sdg-geography:8	Albania	skos:exactMatch	dc:country/AFG	Albania
sdg-geography:12	Algeria	skos:exactMatch	dc:country/AFG	Algeria
sdg-geography:16	American Samoa	skos:exactMatch	dc:country/AFG	American Samoa
sdg-geography:20	Andorra	skos:exactMatch	dc:country/AFG	Andorra
sdg-geography:24	Angola	skos:exactMatch	dc:country/AFG	Angola
sdg-geography:660	Anguilla	skos:exactMatch	dc:country/AFG	Anguilla
sdg-geography:28	Antigua and Barbuda	skos:exactMatch	dc:country/AFG	Antigua and Barbuda
sdg-geography:32	Argentina	skos:exactMatch	dc:country/AFG	Argentina
sdg-geography:51	Armenia	skos:exactMatch	dc:country/AFG	Armenia
sdg-geography:533	Aruba	skos:exactMatch	dc:country/AFG	Aruba
sdg-geography:36	Australia	skos:exactMatch	dc:country/AFG	Australia
sdg-geography:40	Austria	skos:exactMatch	dc:country/AFG	Austria
sdg-geography:31	Azerbaijan	skos:exactMatch	dc:country/AFG	Azerbaijan
sdg-geography:44	Bahamas	skos:exactMatch	dc:country/AFG	Bahamas
sdg-geography:48	Bahrain	skos:exactMatch	dc:country/AFG	Bahrain
sdg-geography:50	Bangladesh	skos:exactMatch	dc:country/AFG	Bangladesh
sdg-geography:52	Barbados	skos:exactMatch	dc:country/AFG	Barbados

Thank you!
Questions

datacommons.org

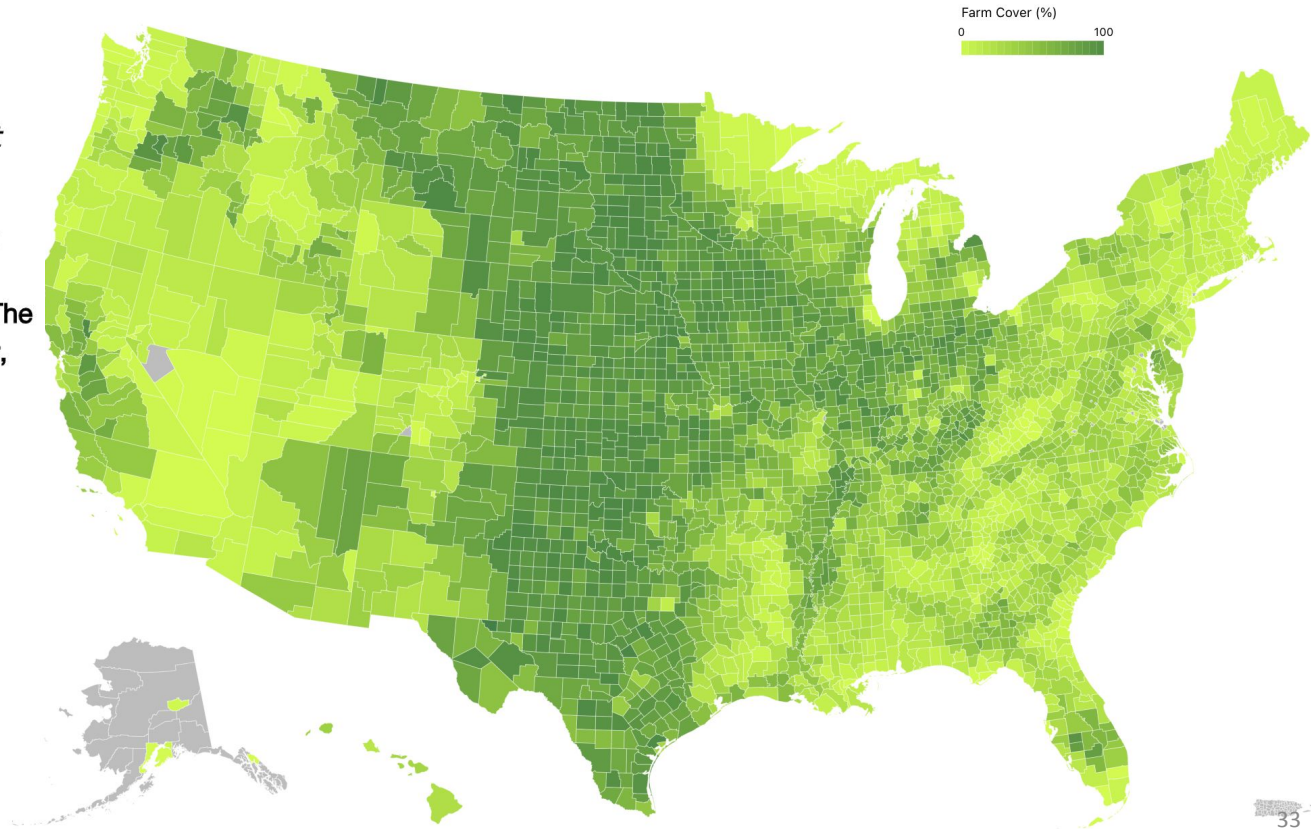
unstats.un.org/UNSDWebsite/undatacommons/sdgs

Appendix

Farm Cover

The United States is one of the largest food producing countries

The US is one of the largest food producing countries in the world, with many people whose livelihood is dependent on farming. **The central states have the most farmland cover, as illustrated in dark green.**

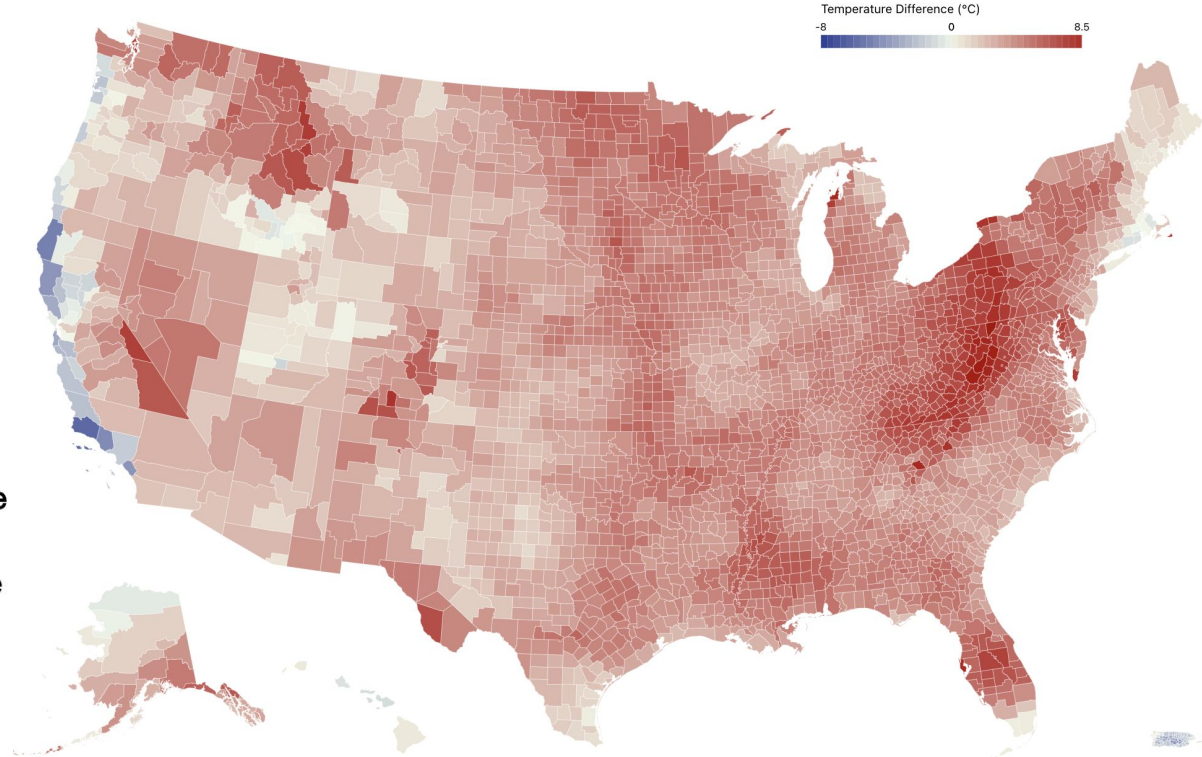


Climate Change: Temperature

As temperatures increase, our food production environments will potentially face newer challenges.

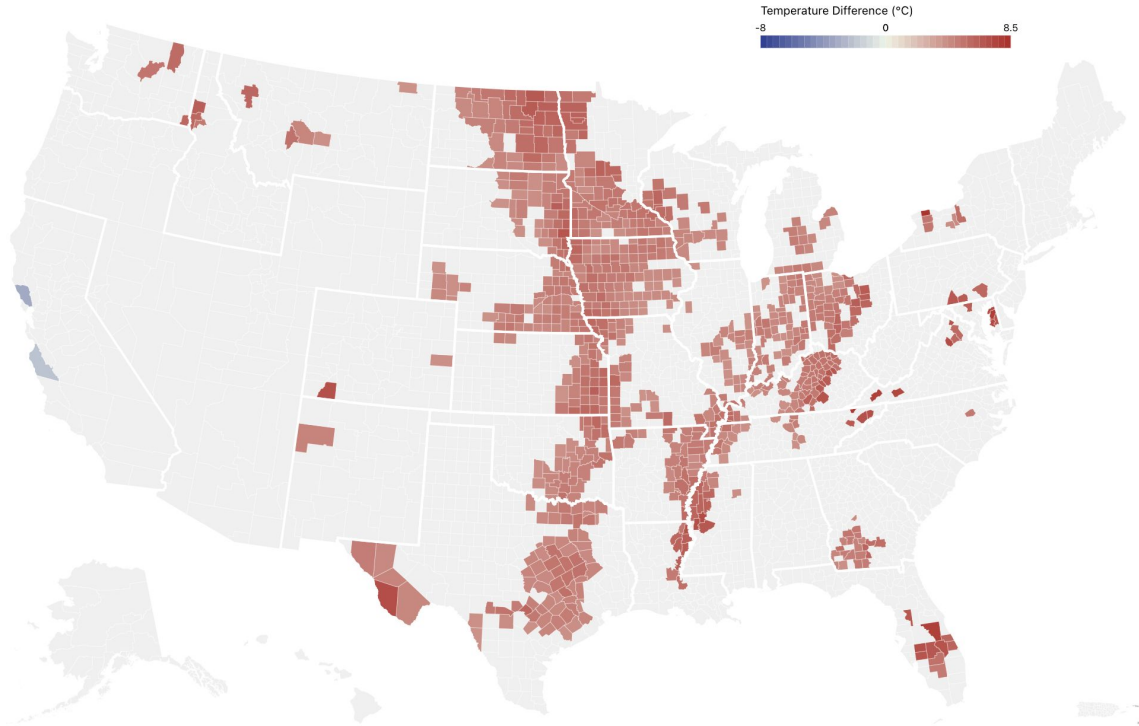
This map is based on the estimated temperature results from the CMIP6 modeling ensemble, under the SSP245 scenario. **This scenario assumes some climate protection measures will be taken and is considered a medium estimate.**

The map shows the predicted difference in the average yearly maximum temperatures between 1980-2010 and 2040-2050. For more info on climate change models see our [about page](#).



Climate Change x Farm Cover

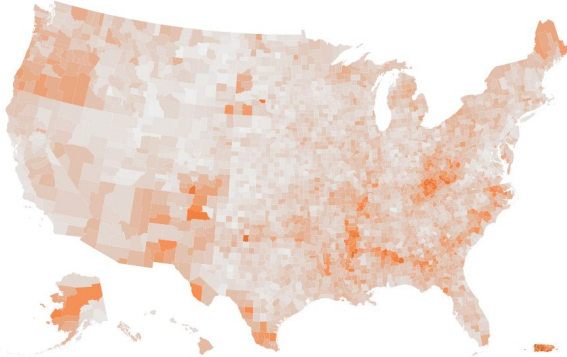
To illustrate **farming regions in the US which may be most impacted by climate change**, the map now highlights counties with more than 50% farm cover that have extreme projected temperature differences (increase of more than 4 degrees or decrease of more than 2 degrees Celsius).



Many Other Existing Inequities

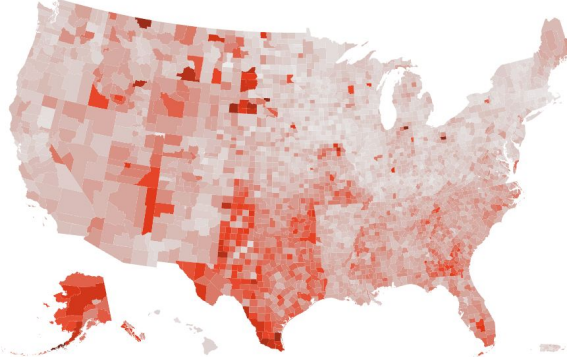
Hunger

Households on Food Stamps Per Capita



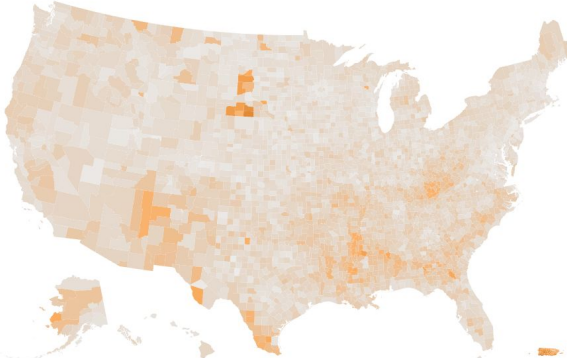
Health Insurance

Population without Insurance, Per Capita



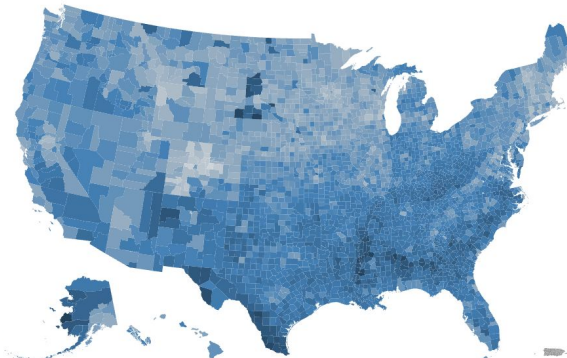
Poverty

Population Living in Poverty, Per Capita



Diabetes

Prevalence of Diabetes



SDMX ↔ Data Commons Semantics

- Natural Synergies with between SDMX Semantics and Data Commons Semantics
- Both organized around Time Series of (statistical) Observations about places/entities
- Concepts (dimensions, attributes) can be mapped to **properties** in the DC Knowledge Graph.
- Special/Custom relationships can easily be represented/encoded

Existing SDMX Data in Data Commons

1. UN Stats: UN SDG data

- UN Stats API
- UN Stats git submodule which does much of the processing
- Geographical Entities and their relationship to each other (e.g. country contained in a region)
- Scripts:
<https://github.com/datacommonsorg/data/tree/master/scripts/un/sdg>

2. OECD

- Bulk import of a few hundred datasets
- OECD API
- Imported in a "schemaless" manner (minimal schema mappings)
- Scripts:
<https://github.com/datacommonsorg/data/tree/master/scripts/oecd/sdmx>

UN Stats Data: Schema

Schema

- The series `SDG_<series code>` is used as the `StatisticalVariable` `populationType`
- The `TIME_PERIOD` dimension is used for `observationDate`
- The `NATURE`, `OBS_STATUS`, and `REPORTING_TYPE` dimensions are used for `measurementMethod`
- The `UNIT_MEASURE` and `BASE PERIOD` dimensions are used for `unit`
- The `UNIT_MULT` dimension is used for `scalingFactor`
- All other dimensions are used for the `StatisticalVariable` definition: for each dimension we define a new `sdg_<dimension>` **property** and corresponding enumeration
- The `StatisticalVariable` `dcid` is formatted like `sdg/<series code>.<dimension 1>--<value 1>__<dimension 2>--<value 2>`

Sample Template Mappings

- Preprocess data to produce a flat CSV
- Map CSV headers to Data Commons properties

```
Node: E:SDG->E0
typeOf: dcs:StatVarObservation
variableMeasured: C:SDG->VARIABLE_CODE
observationAbout: C:SDG->GEOGRAPHY_CODE
observationDate: C:SDG->TIME_PERIOD
value: C:SDG->OBS_VALUE
unit: C:SDG->UNIT_MEASURE
scalingFactor: C:SDG->UNIT_MULT
measurementMethod: C:SDG->MEASUREMENT_METHOD
```

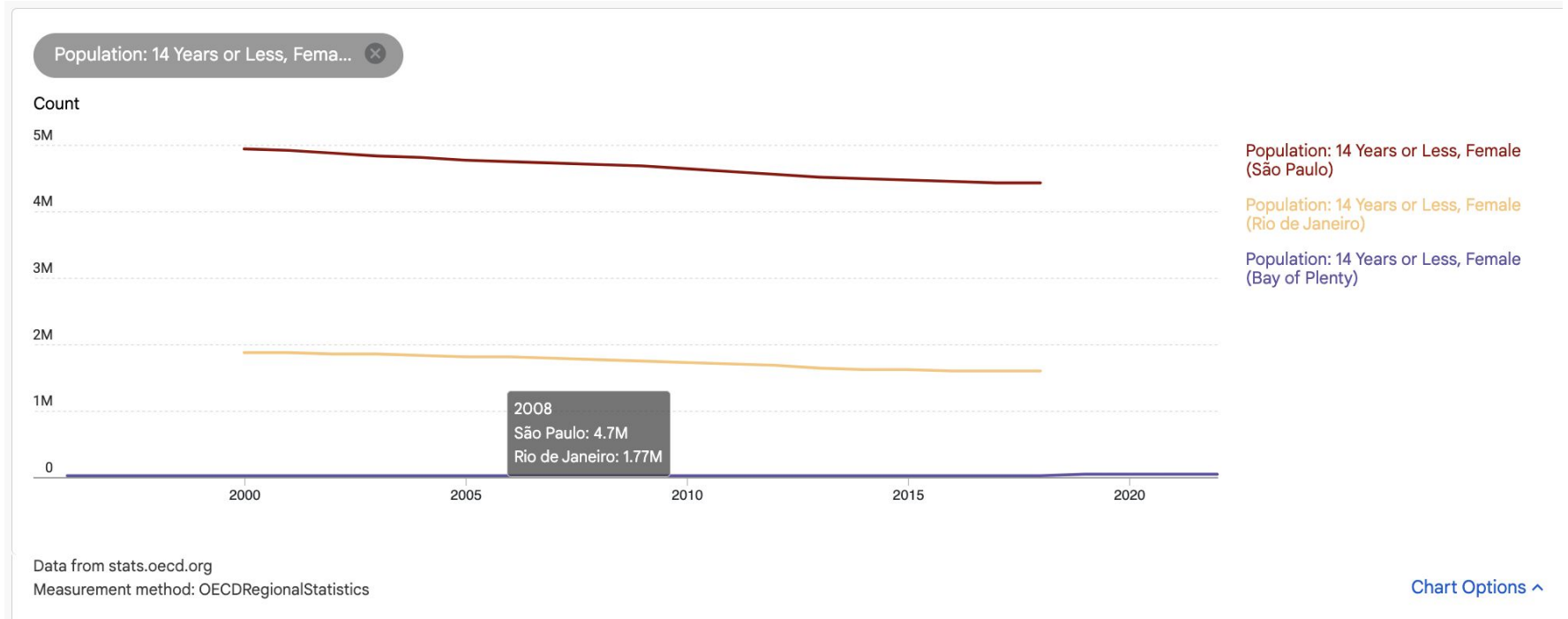

Mapping Places

unDataCode,unDataLabel,dcid,dc_name,containedInPlace,typeOf

undata-geo:G00002290,Pakistan,country/PAK,Pakistan,"[{ 'dcid': 'Earth', 'name': 'World'}, { 'dcid': 'SouthernAsia', 'name': 'Southern Asia'}, { 'dcid': 'asia', 'name': 'Asia'}]","[{ 'dcid': 'Country', 'name': 'Country'}]"

undata-geo:G00129000,Eastern Europe,EasternEurope,Eastern Europe (including Northern Asia),"[{ 'dcid': 'europe', 'name': 'Europe'}]","[{ 'dcid': 'UNGeoRegion', 'name': 'UNGeoRegion'}]"

Data Example (OECD)



Interoperability and Data Modelling

- Interoperability is highly dependent on data and metadata modelling decisions and practices
- Same content can be represented in a variety of ways
- No single “right” way of representing information
- Some data structures are better suited for data exchange processes
- Others are better suited for analyzing and communicating data to users

What is Data Modeling?

A process focused on:

1. Clearly and unambiguously **identifying things** that a dataset aims to capture
2. Selecting the key properties that should be captured to **describe those things** in a meaningful way
3. Deciding **how things relate** to each other
4. Deciding how this information should be **formally codified**