Data Commons

datacommons.org

Context

There is a lot of data …

From census (US, India, UK, …), RBI, USGS, BLS, BEA, UDISE, World Bank, …

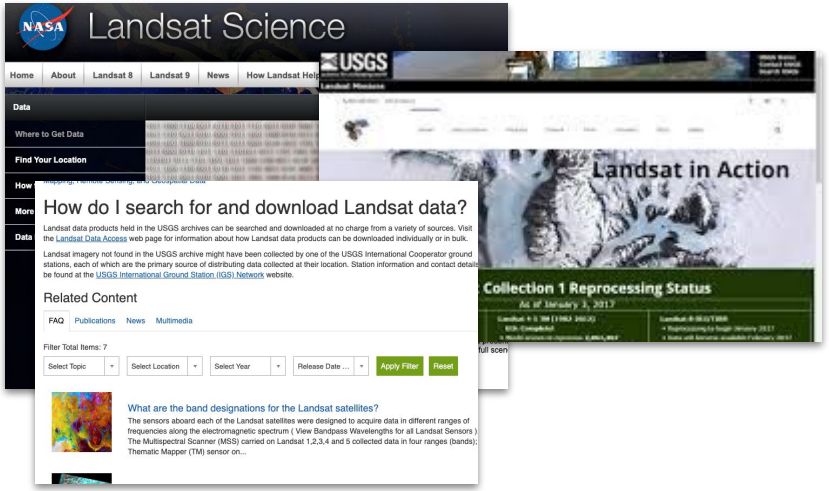
This data is essential

For science, journalism, policy and our users

Using this data is extremely painful

Repeated expensive data wrangling

Analogy with Satellite Imagery and Google Earth/Maps



Our goal for data

CA counties most at food risk from climate change?

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

to

Just ask in Natural Language

Journalists

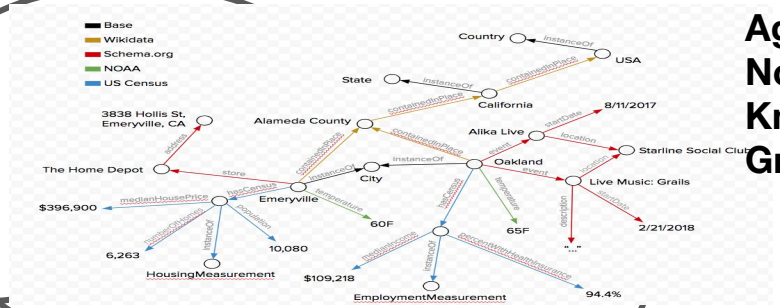
Students

Search, News

Researchers

Cloud

APIS



Aggregated Normalized Knowledge Graph

CDC

NOAA

HURDAT

BLS

Census

IPCC

EPA

OECD

World Bank

WHO

EIA

What we have built: Code

Open source:

- GCP infrastructure for creating, storing, serving, KG
- Visualization tools

In Google's Search Stack: Integration with Google search

What we have built: Data

People & places, …

- **Demographics**: Census (US, India, …) , Eurostat, …
- **Economics**: BLS, BEA, WorldBank, …
- **Health**: CDC, DEA, WHO, ICD, …
- **Food, Crime**, **Education, Elections**, **Trade**, **…**

Climate Change & Sustainability:

- **Climate:** IPCC, EPA, HURDAT, NOAA
- **Energy:** EIA, NREL, …
- **Water, Agriculture:** USGS, WRIS (India), USDA, FAO
- **Emissions**: EPA, EIA, ..

3.5+ Billion time series
1.7 trillion triples
2.4 M places
100+ k variables

3x search KG
5x FRED

What we have built: Data

Biomed

- **Biochemistry:** PharmGKG, ChEMBL, Drugs@FDA …
- **Protein:** UniProt, MINT, HUPO, Tissue Atlas
- **Sequence:** clinVar, ENCODE, dbSNP, GTEx, NCBI Gene…
- **Species:** Uniprot controlled vocab of species
- **Omics:** HMDB, Human1D, Virtual Metabolic Human
- **Misc:** SIDER, CDC Wonder, ICD-10, Disease Ontology…

44M Gene variants
3.4M Bio specimens
1.95M compounds

Principal Investigator is Samantha Piekos

Work started at Stanford, now at The Institute for Systems Biology

Application: In Google Search

NL interface in Google search

Energy use per capita in India

CO2 emissions in Sweden

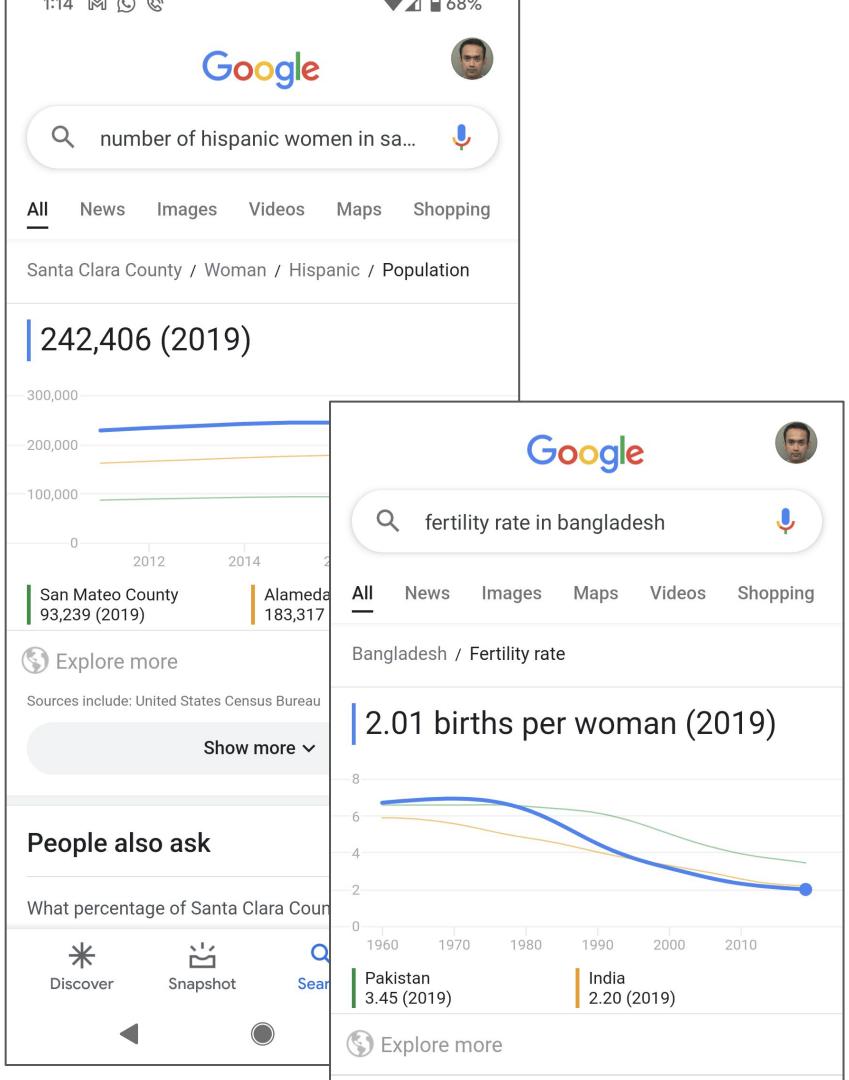
Number of unemployed in California

Population growth rate in Germany

Fertility rate in bangladesh

Number of women phds in Mountain View

Number of poor hispanic women in Santa Clara



Sustainability Data Commons

Climate Change: two facets

The Physical World: materials science, chemistry, grid, ...

The Information World

Understanding the impact of the changing climate to prepare

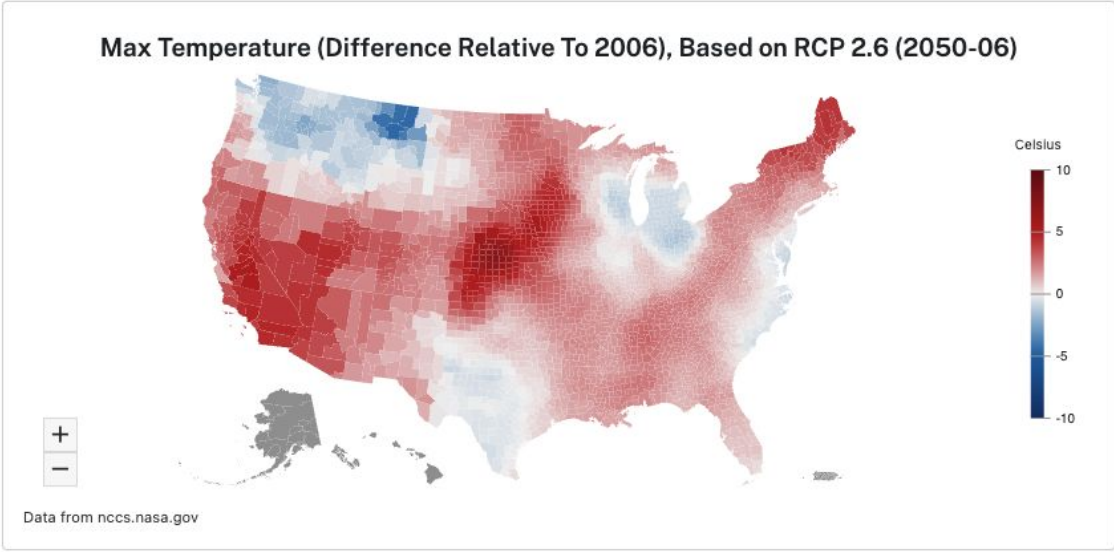
Reducing waste

Carbon accounting

Climate Change: Temperature

Climate change is not as simple as 1.5 vs 2 degree.

Climate deltas vary widely across places



Top Places

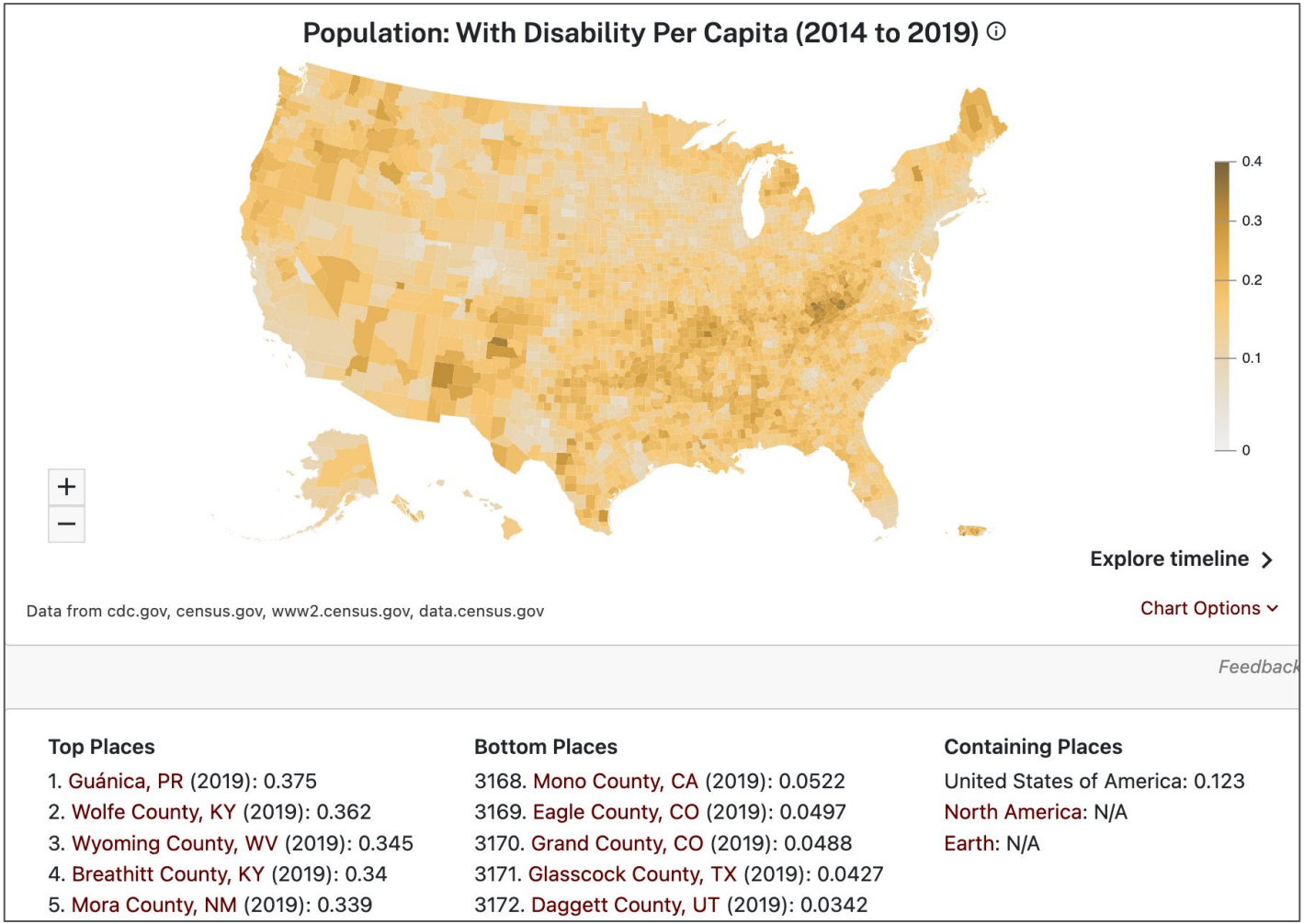
1. Phillips County, KS (2050-06): 7.22 Celsius
2. Norton County, KS (2050-06): 7.17 Celsius
3. Harlan County, NE (2050-06): 7.12 Celsius
4. Phelps County, NE (2050-06): 7.08 Celsius
5. Furnas County, NE (2050-06): 6.99 Celsius

Bottom Places

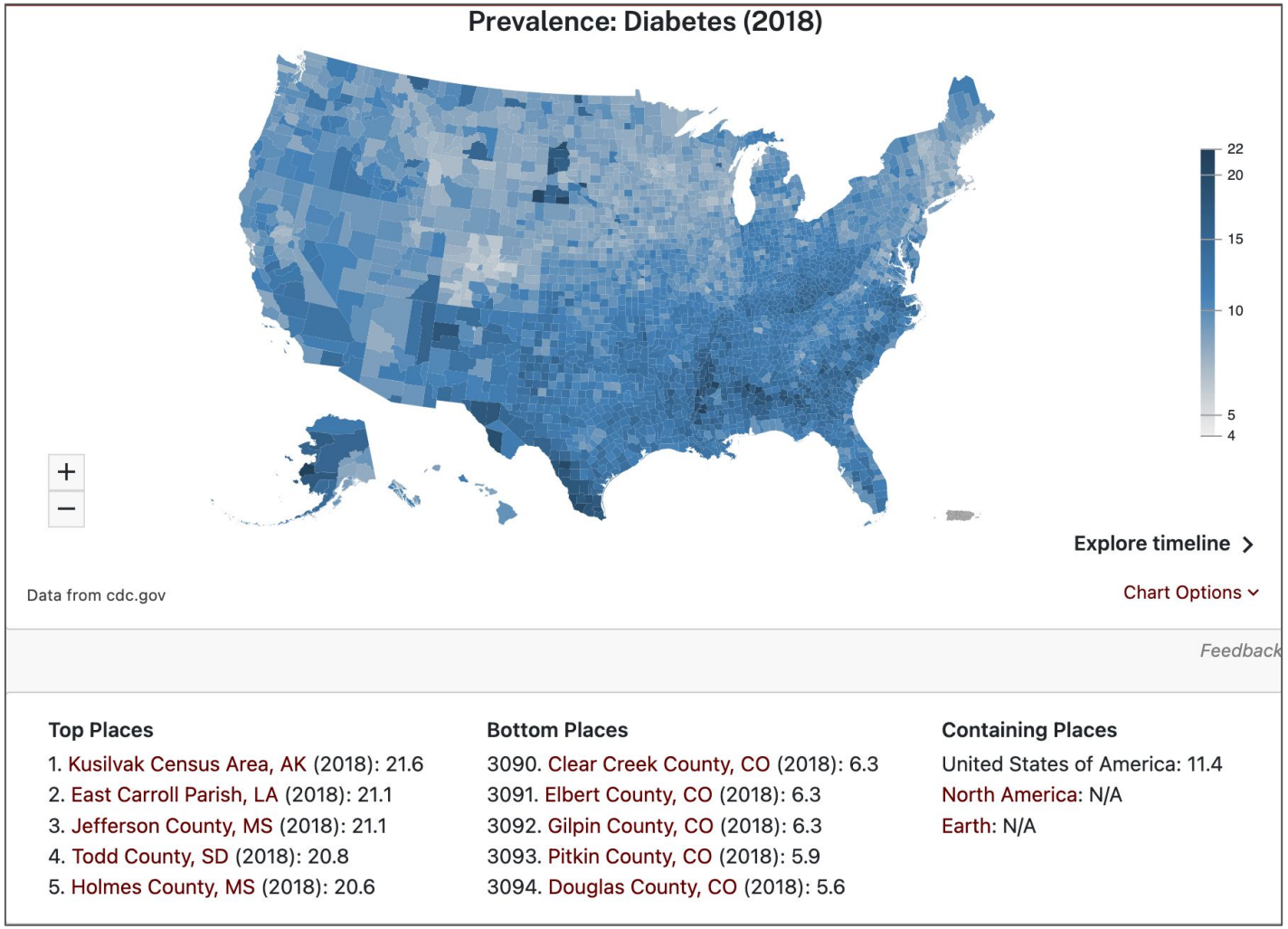
3059. Petroleum County, MT (2050-06): -3.8 Celsius
3060. Chouteau County, MT (2050-06): -4.13 Celsius
3061. Blaine County, MT (2050-06): -4.27 Celsius
3062. Phillips County, MT (2050-06): -4.43 Celsius
3063. Fergus County, MT (2050-06): -4.48 Celsius

There are many inequities

Fraction of population with Disabilities



Percentage of population with Diabetes



Top Places

1. Kusilvak Census Area, AK (2018): 21.6
2. East Carroll Parish, LA (2018): 21.1
3. Jefferson County, MS (2018): 21.1
4. Todd County, SD (2018): 20.8
5. Holmes County, MS (2018): 20.6

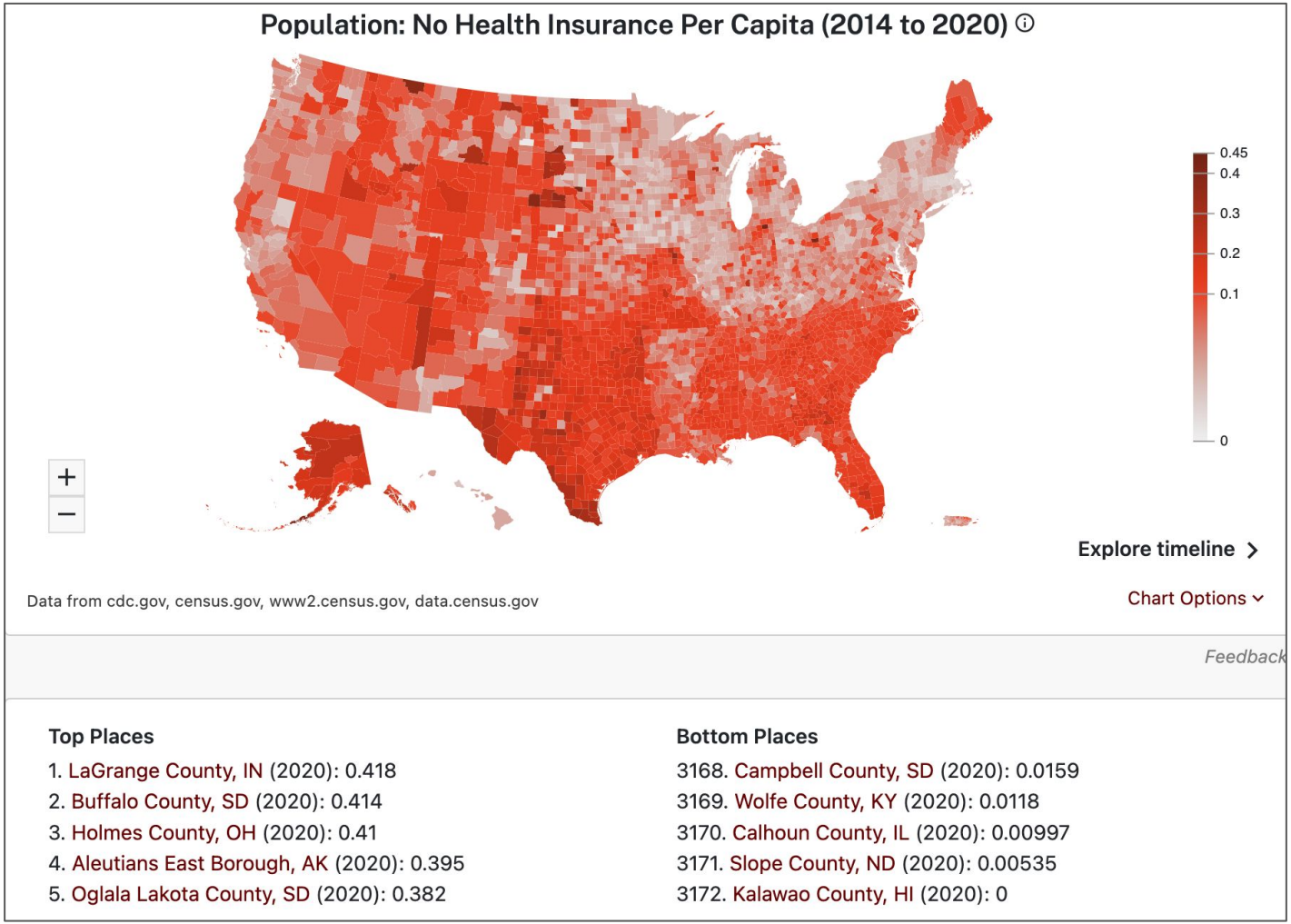
Bottom Places

3090. Clear Creek County, CO (2018): 6.3
3091. Elbert County, CO (2018): 6.3
3092. Gilpin County, CO (2018): 6.3
3093. Pitkin County, CO (2018): 5.9
3094. Douglas County, CO (2018): 5.6

Containing Places

United States of America: 11.4
North America: N/A
Earth: N/A

Without Health Insurance



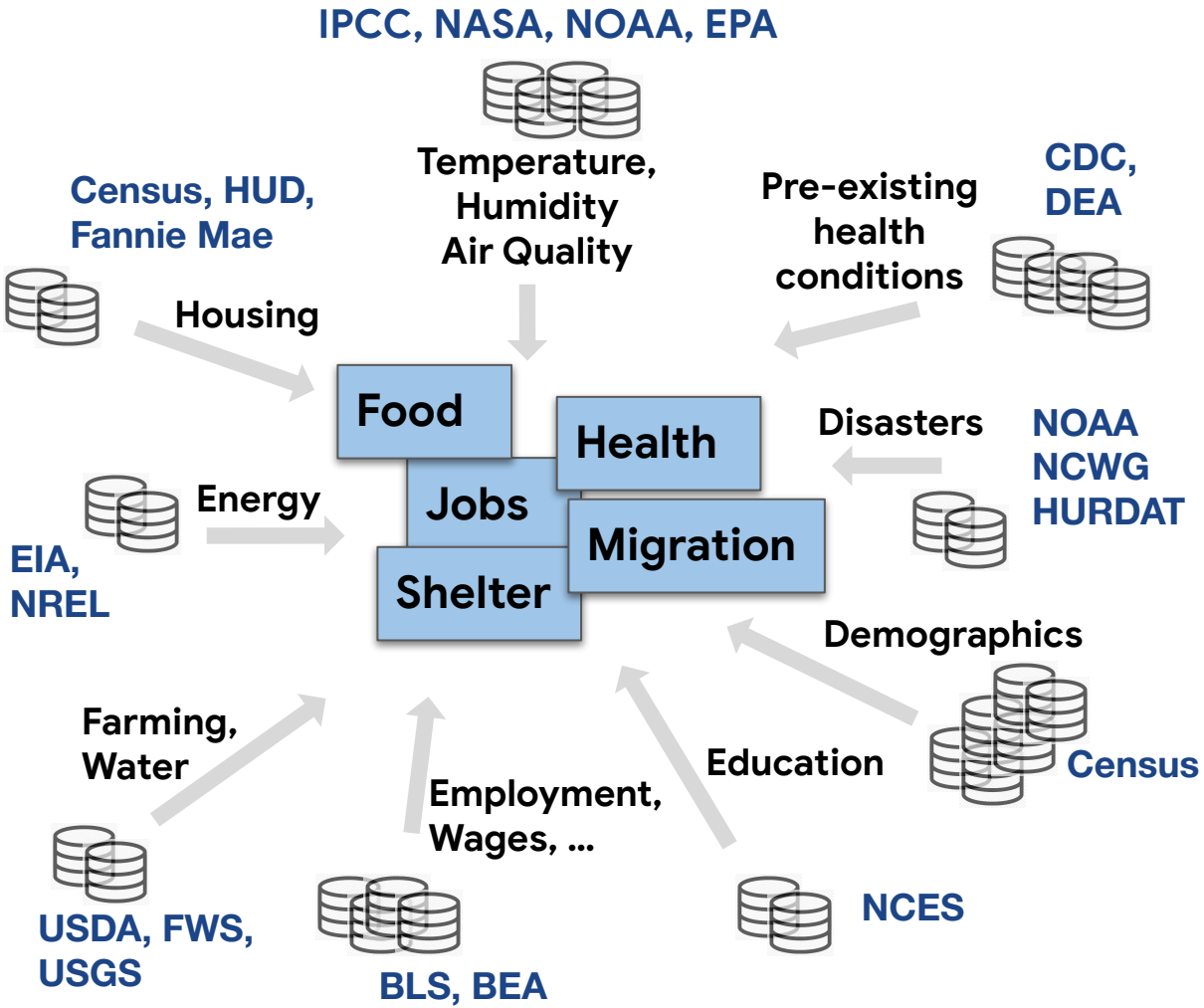
Climate change will worsen these …

We have to prepare

To prepare, we have to know who is going to be most affected: 10, 20, 30 years from now

Messy Data Ecosystems

We need data not just about climate, but about data about health, food, farming, water, employment, …



The Sustainability Data Commons Mission

- `Organize and make easily accessible' all the sustainability data for everyone
  - researchers, policy folks, administrators, students … all of our users!
- Sustainability Data Commons is the base data layer for many applications
  - Different interfaces, data augmentations, ML tools and more
  - From many different companies, universities, …
- This 'base data layer' is open and built together with other players from industry, academia and government (like html, schema.org, etc.)

Many topics already in and many more coming

Climate, Water, Agriculture, Air quality

Fuels, Energy Production

Natural Disasters (Storms, fires, floods, ...)

Mortality causes, Poverty, Social Benefits

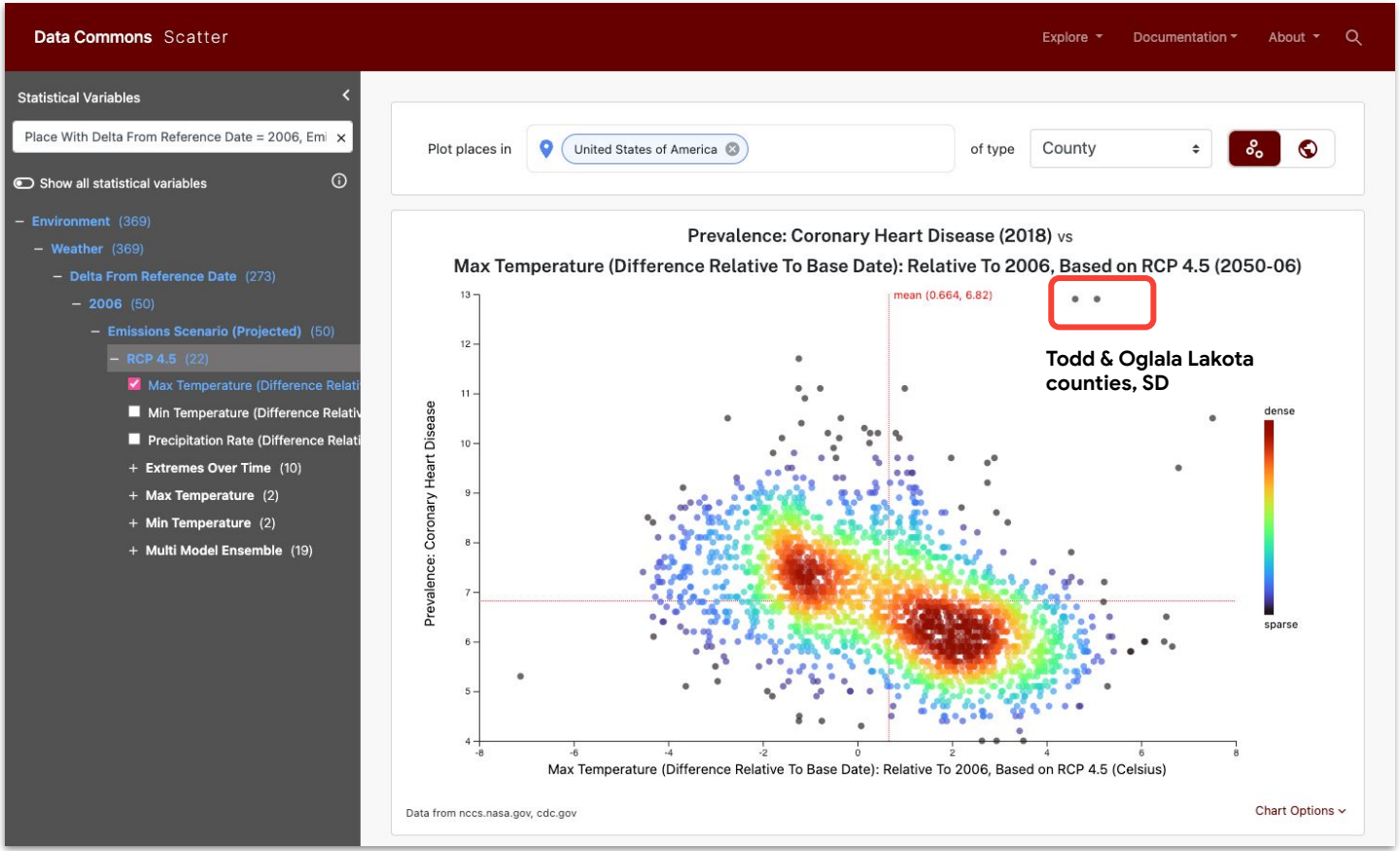
Education (K-12 and college)

Employment & labor markets, Businesses

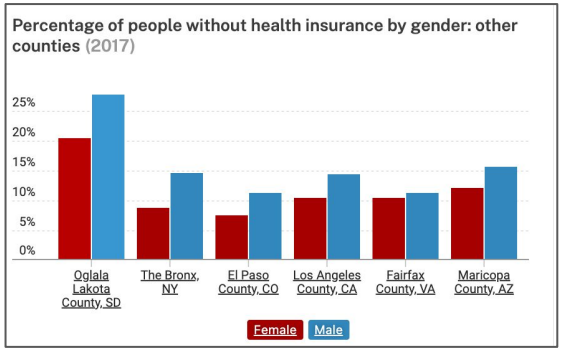
Housing, Commute, Insurance, Disabilities, Immigration, Crime

Application: The Resiliency Dashboard

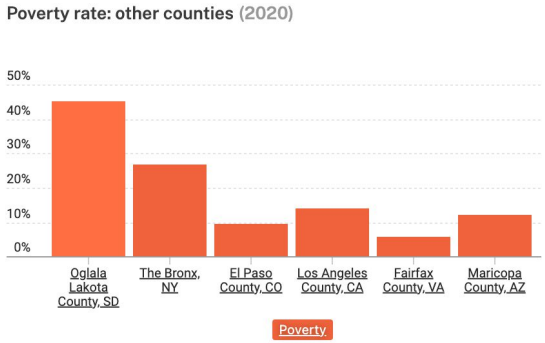
Health: Cardiac conditions --- who is in most danger from temperature rise?



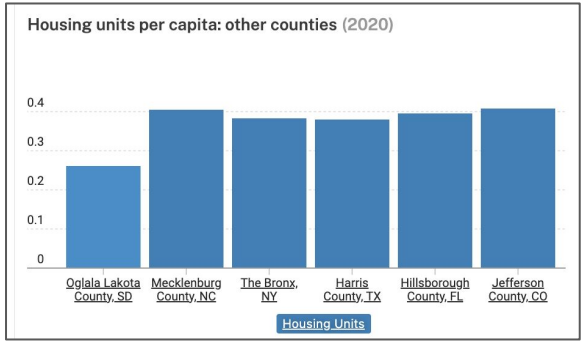
Oglala Lakota County --- other vulnerability factors



High rate of uninsured



High Level of Poverty

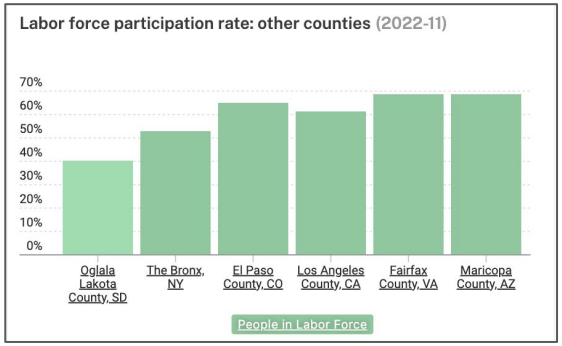


Fewer homes per capita

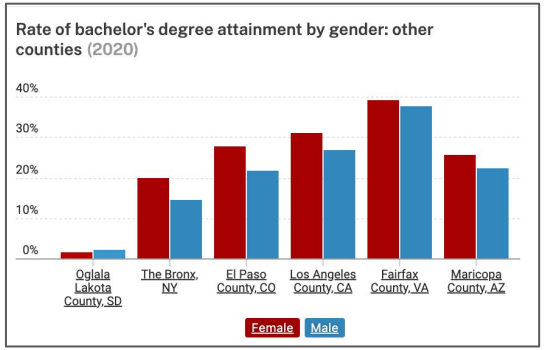
Oglala Lakota County --- the bigger picture

Rankings (in)	South Dakota	United States
Largest Population	15 of 67	2186 of 3108
Highest Median Income	64 of 67	3110 of 3187
Highest Median Age	62 of 67	3168 of 3187
Highest Unemployment Rate	2 of 66	193 of 3185

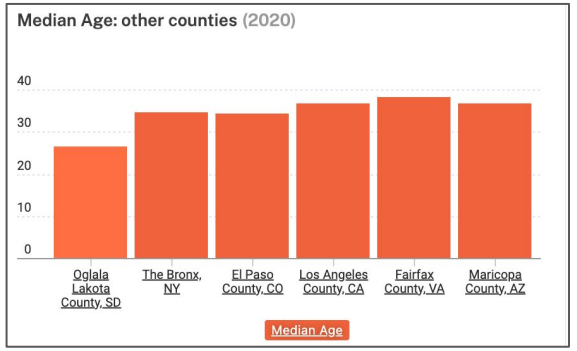
Data from census.gov, fbi.gov, bls.gov



Low labor force participation



Low levels of college edu.



Very young!

Oglala Lakota County

County in South Dakota, United States of America, North America

Population: 13,586 (2021) *www2.census.gov*

Enter a country, state, county or city

Graph Browser

OVERVIEW

Rankings (in)	South Dakota	United States of America
Largest Population	15 of 67	2186 of 3108
Highest Median Income	64 of 67	3110 of 3187
Highest Median Age	62 of 67	3168 of 3187
Highest Unemployment Rate	3 of 66	301 of 3185

Data from census.gov, fbi.gov, bls.gov

Places in Oglala Lakota County

Towns

Batesland

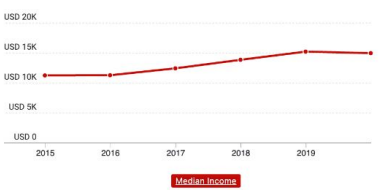
Cities

Kyle, Manderson-White Horse Creek, Oglala, Pine Ridge, Porcupine, Wounded Knee

ECONOMICS

MORE CHARTS ›

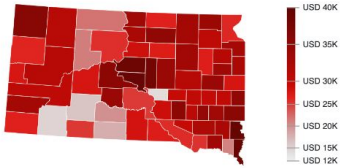
Median individual income in Oglala Lakota County



Data from census.gov Export Explore More ›

Feedback

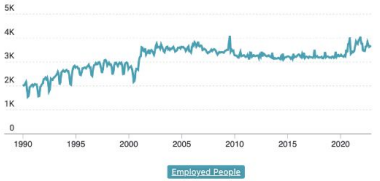
Median individual income: counties near Oglala Lakota County (2014 – 2020)



Data from census.gov Export Explore More ›

Feedback

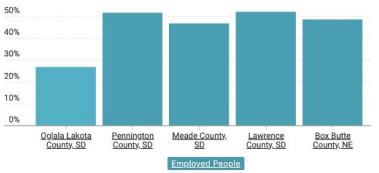
Number of people employed in Oglala Lakota County



Data from bls.gov Export Explore More ›

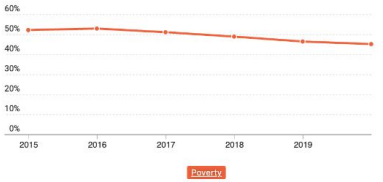
Feedback

Percentage of people employed: counties near Oglala Lakota County (2022-12)



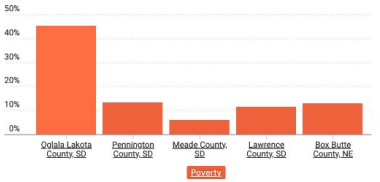
Data from www2.census.gov, bls.gov Export Explore More ›

Poverty rate in Oglala Lakota County



Data from www2.census.gov, census.gov Export Explore More ›

Poverty rate: counties near Oglala Lakota County (2020)



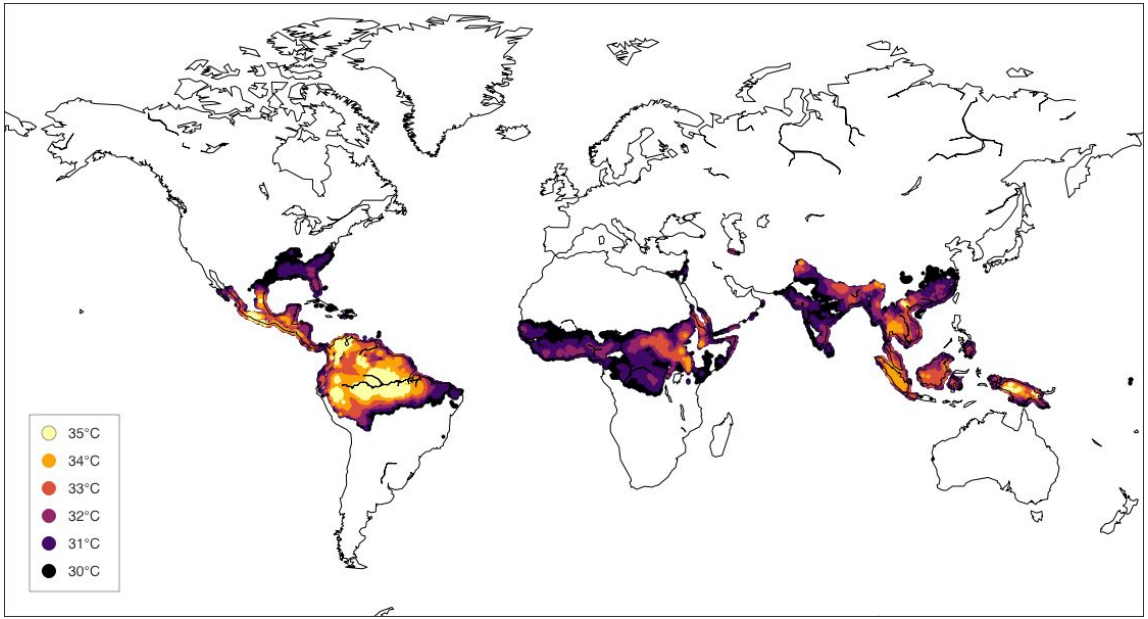
Data from www2.census.gov, census.gov Export Explore More ›

Not just the US …

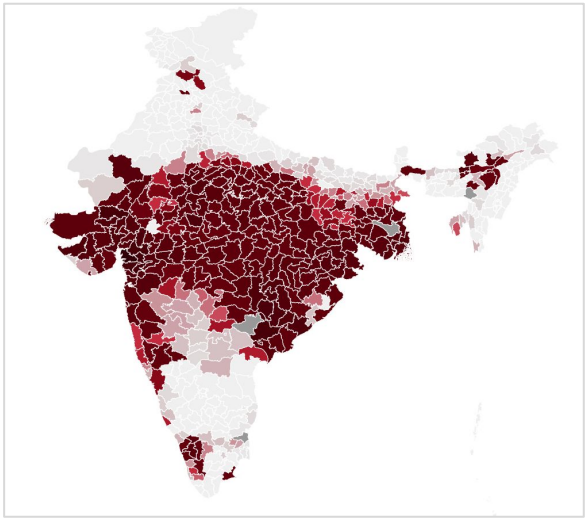
Wet Bulb Temperature

Unlike raw temperature, no IPCC data

Need models of humidity, intra-day variation in humidity, etc. to project max wet bulb temperature.



Focussing on India ...

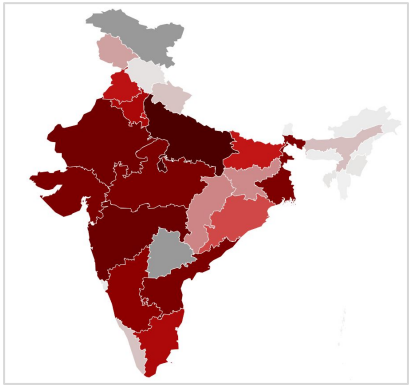


Regions of India that will experience days with wet bulb temp > 34 deg

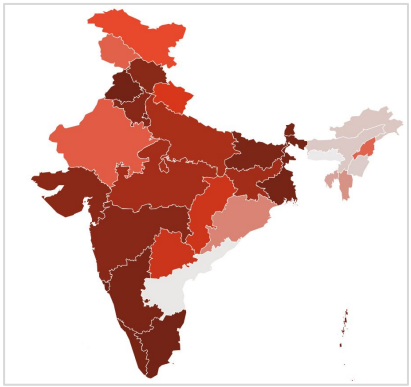
(color indicates number of days)
(RCP 4.5, NASA_Mean_ACCESS1-0)

Who will be affected most?

Homeless Population



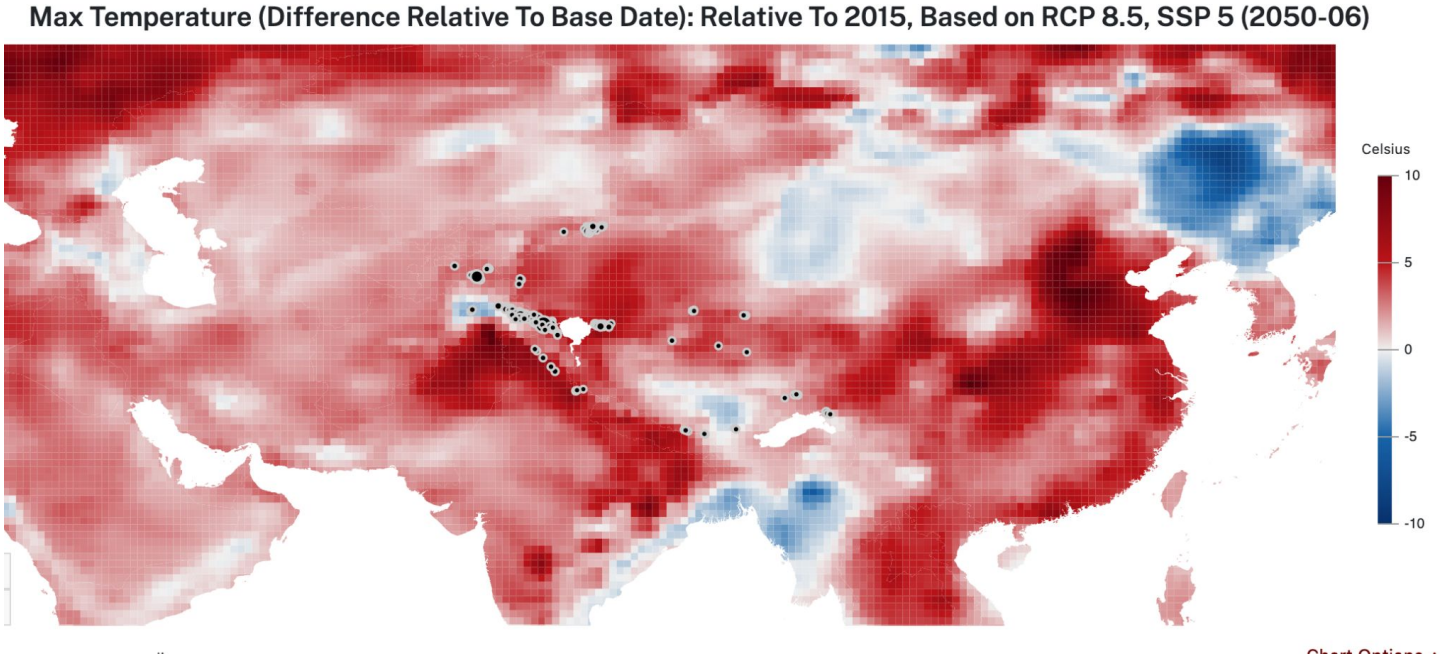
Census IN (2011)



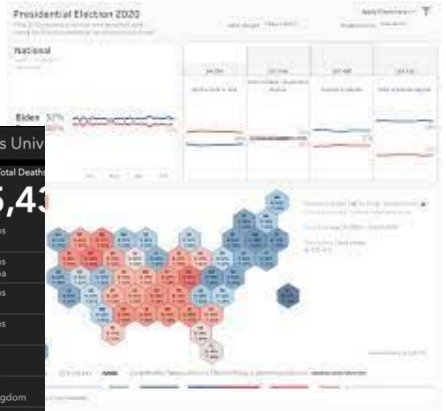
UDISE IN (2020)

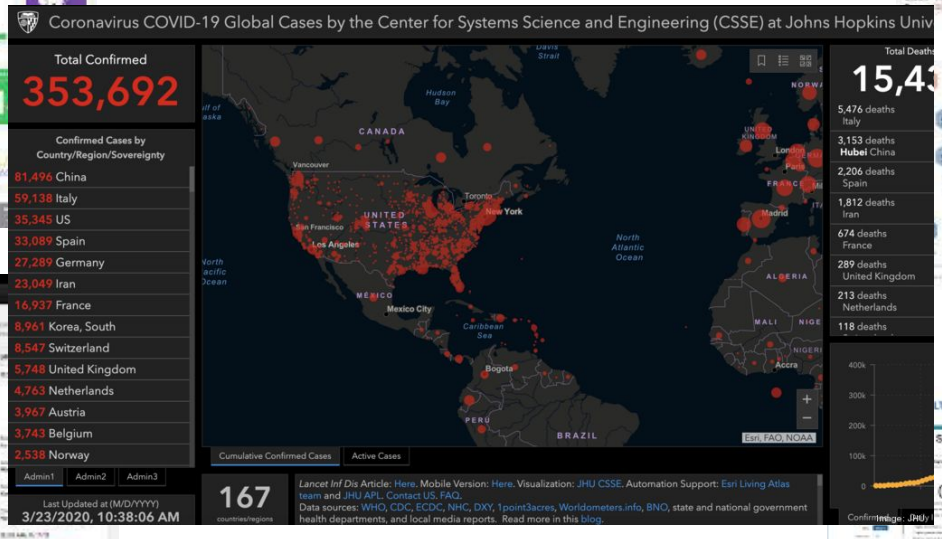
Percentage of schools with functioning toilets

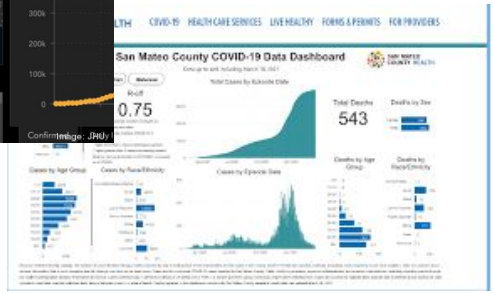
Glaciers feeding the Indus, Ganga & Brahmaputra



Dashboards focus attention







JHU Dashboards

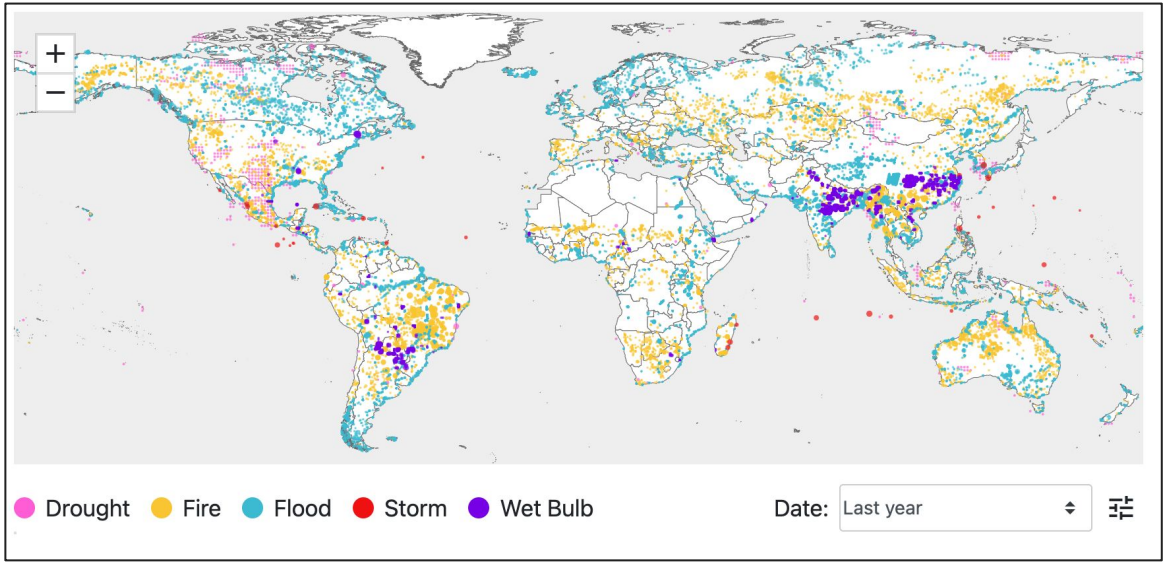
Coming Soon: Google-Stanford Resilience Platform

The platform that will inform about who is most at what kind of risk

A set of dashboards --- But more than that --- a platform on top of which many applications can be built

Part of the Google-Stanford

Data Commons partnership



Education

Data is the platform

- Berkeley DS100
- Being used data science/ML courses (in CS & Econ) at Berkeley, Harvard & MIT

Case Study: Prevalence of Obesity in 500 US Cities

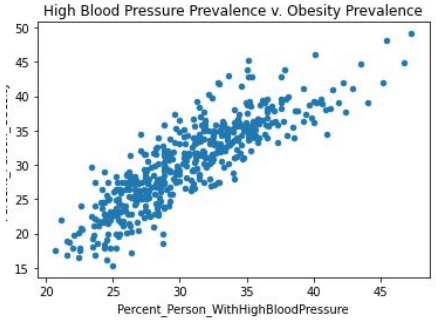
Obesity is well known to correlate with health factors such as high blood pressure, but is also known to correlate with economic factors s low-income, unemployment, etc [1][2]. The Center for Disease Control (CDC) provides prevalence percentages on health conditions such a obesity, high blood pressure, 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 ac the United States.

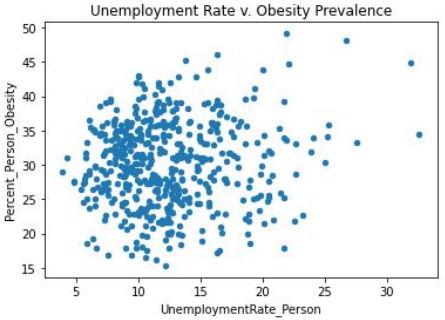
Even though these statistics come from different datasets across different government agencies with different storage formats, Data Commons surfaces each of these in a single, uniform knowledge graph. In fact, you can see this in the browser by looking at the *provenar* column. Let's use the data in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

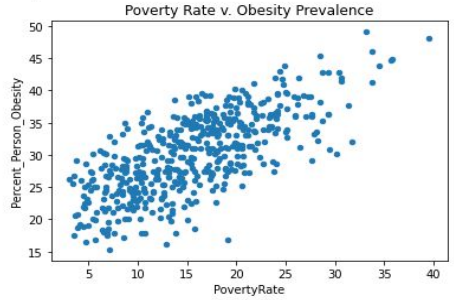
- Prevalence of high blood pressure
- Unemployment rate
- Percent of population living with income below the poverty line

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 pr employment statistics (you can see this by navigating to the "employment" and "employmentStatus" sections for San Francisco and obse the different provenances). Our choice of using statistics from the Bureau of Labor Statistics is purely demonstrative, but it would be inte to see if similar results can be reproduced using US Census employment statistics.

```
[ ]  !pip install datacommons_pandas --upgrade --quiet
```







Data Commons in Core ML Curriculum (MIT)

- MIT's large undergraduate course: Intro to Machine Learning (6.036)
- Course offered every semester
- > 500 students (~400 Undergraduates; 100+ Masters students)
- Core, Required Course for ML/AI tracks
- All integration work was done by course TA's with minimal help from the core DC team

Data Literacy

Data Literacy with Data Commons

- Introductory, few prerequisites
- Pedagogical Goals:
  - **Use *Real* Data**
  - Be Interactive
  - Easy to Adapt
  - **Focus on storytelling**
  - **Assume little programming and statistics background**
- Target Audience:
  - **Intro courses** in data analysis/literacy/science
  - Specialized skills development, e.g. **decision making** for public policy, healthcare etc.
  - PreReqs: **no** programming, **descriptive** statistics

Data Science with *Real* Data

- Intermediate / Advanced
- Pedagogical Goals:
  - **Use *Real* Data**
  - Be Interactive
  - Easy to Adapt
- Target Audience:
  - **Foundational courses** in data analysis/literacy/science
  - Specialized skills development, e.g. **analytical tools** for public policy, healthcare etc.
  - PreReqs: programming, statistics

Biomedical Data Commons

Biomed DC: the data

Biochemistry Data

- antibodyABCD, ChEMBL, Drugs@FDA, FDA - Pharmacologic Class, PharmGKB

Protein Data

- The Human Protein Tissue Atlas, HUPO - Proteomics Standard Initiative, Molecular INteraction Database (MINT), UniProt

Sequencing Data

- dbSNP, ENCODE, GTEx, NCBI - Assembly, NCBI - Gene, clinVar

Disease / Healthcare Data

- CDC 500 Cities / CDC Places, CDC Diabetes Surveillance System, CDC Wonder Mortality, ICD-10 Codes, Medical Subject Headings (MeSH), Side Effect Resource (SIDER)

Omics Data (coming soon)

- The Human Metabolome Database (HMDB), Human1D, Virtual Metabolic Human

Sample Application: Personalized medicine

Purpose: Clinician support for interpreting patient genetic information by putting into the context of disease and treatment.

English: For a genetic variant, identify the clinical significance and the associated disease. Then identify the disease description, symptoms, drug treatments, and drug dosages.

SPARQL:

```
GENETIC_VARIANT = 'rs33930977' # declare genetic variant

query = '''
SELECT ?gv_dcid ?disease_name
WHERE {{
?gv typeOf GeneticVariant .
?gv name "{0}" .
?gv dcid ?gv_dcid .
?gv diseaseName ?disease_name
}}
'''.format(GENETIC_VARIANT)
result = dc.query(query)[0] # get disease name
gv_dcid, disease_name = result['?gv_dcid'], result['?disease_name'].lower()
# get variant clinical significance
sigLink = list(dc.get_property_values([gv_dcid],\
                                      'clinicalSignificance').values())[0]
sig = list(dc.get_property_values(sigLink, 'name').values())[0][0]
query = '''
SELECT ?disease_dcid ?description
WHERE {{
?disease typeOf Disease .
?disease name "{0}" .
?disease dcid ?disease_dcid .
?disease description ?description
}}
'''.format(disease_name)
result = dc.query(query)[0]  # get disease description
disease_dcid, des = result['?disease_dcid'], result['?description']
```

```
# identify symptoms associated with the disease
dis_symp_dcids = list(dc.get_property_values([disease_dcid], \
                      'diseaseOntologyID',  out=False, limit=1000).values())[0]
dis_symp_dcids = [dcid for dcid in dis_symp_dcids if "cooccurrence" in dcid]
symptoms_dcids = list(dc.get_property_values(dis_symp_dcids,
                                             'medicalSubjectHeadingID').values())
symptoms_dcids = list(itertools.chain.from_iterable(symptoms_dcids))
symptoms_names = list(dc.get_property_values(symptoms_dcids, 'name').values())
symptoms_names = list(itertools.chain.from_iterable(symptoms_names))
# identify drug treatments
disease_treat_dcids = list(dc.get_property_values([disease_dcid],\
                           'diseaseID',  out=False, limit=1000).values())[0]
disease_treat_dcids = [dcid for dcid in disease_treat_dcids if "CTD_" in dcid]
treatment_dcids = list(dc.get_property_values(disease_treat_dcids, \
                                              'compoundID').values())
treatment_dcids = list(itertools.chain.from_iterable(treatment_dcids))
treatment_names = list(dc.get_property_values(treatment_dcids,\
                                              'commonName').values())
# get drug dosage forms
dosage_form = list(dc.get_property_values(treatment_dcids,\
                                          'dosageForm').values())

info = [sig, disease_name, des, symptoms_names, treatment_names, dosage_form]
for item in info:  # print desired info
  print(item)
```

Uses data from 6 sources: ClinVar, DiseaseOntology, MedicalSubjectHeadings, UCSF SPOKE, ChEMBL, Drugs@FDA

Beyond A Single DataCommons.org

Limits of Google's datacommons.org

There is a lot of data that cannot come into Google's datacommons.org

Open data with non-commercial use licenses

Data govts not willing to share with Google

Open data behind walled gardens: D&B, Bloomberg, ...

Private: Patient records, sales, inventory

Bringing Data Commons benefits to this data

Many Data Commons, from different organizations

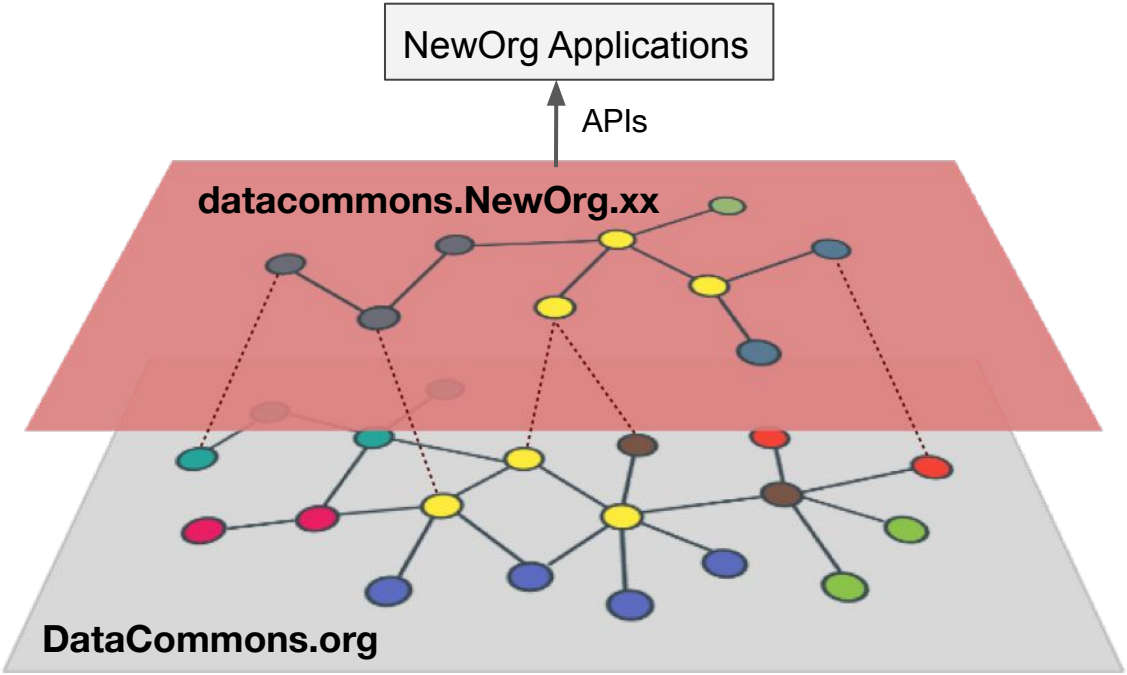
Different topics, some behind firewalls, some behind paywalls, ...

Like the Web

- All have the same schema and same API
- Same tools work on all the Data Commons

Why would someone put up a data commons?

- More value can be derived from their data



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 ...

Feeding America and the Meal Gap Index

Case Study: Feeding America Data Commons

FEEDING AMERICA Data Commons

TIMELINE EXPLORER | SCATTER EXPLORER | MAP EXPLORER | ABOUT | FAQ

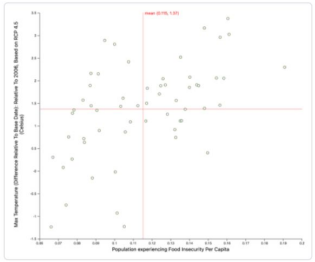
MAP THE MEAL GAP

Feeding America Data Commons provides access to data from Map the Meal Gap, overlayed with data from a wide range of additional sources into a single portal under a common scheme. Learn more ›

Explore the data

CLIMATE CHANGE AND FOOD INSECURITY IN CALIFORNIA

While there are significant urban populations in California with food insecurity challenges today e.g. Los Angeles County, when food insecurity numbers are mapped with anticipated temperature rises, the people of Kern County, the heart of California's agricultural ecosystem that serves the world, will likely experience the greatest strife in the years to come.

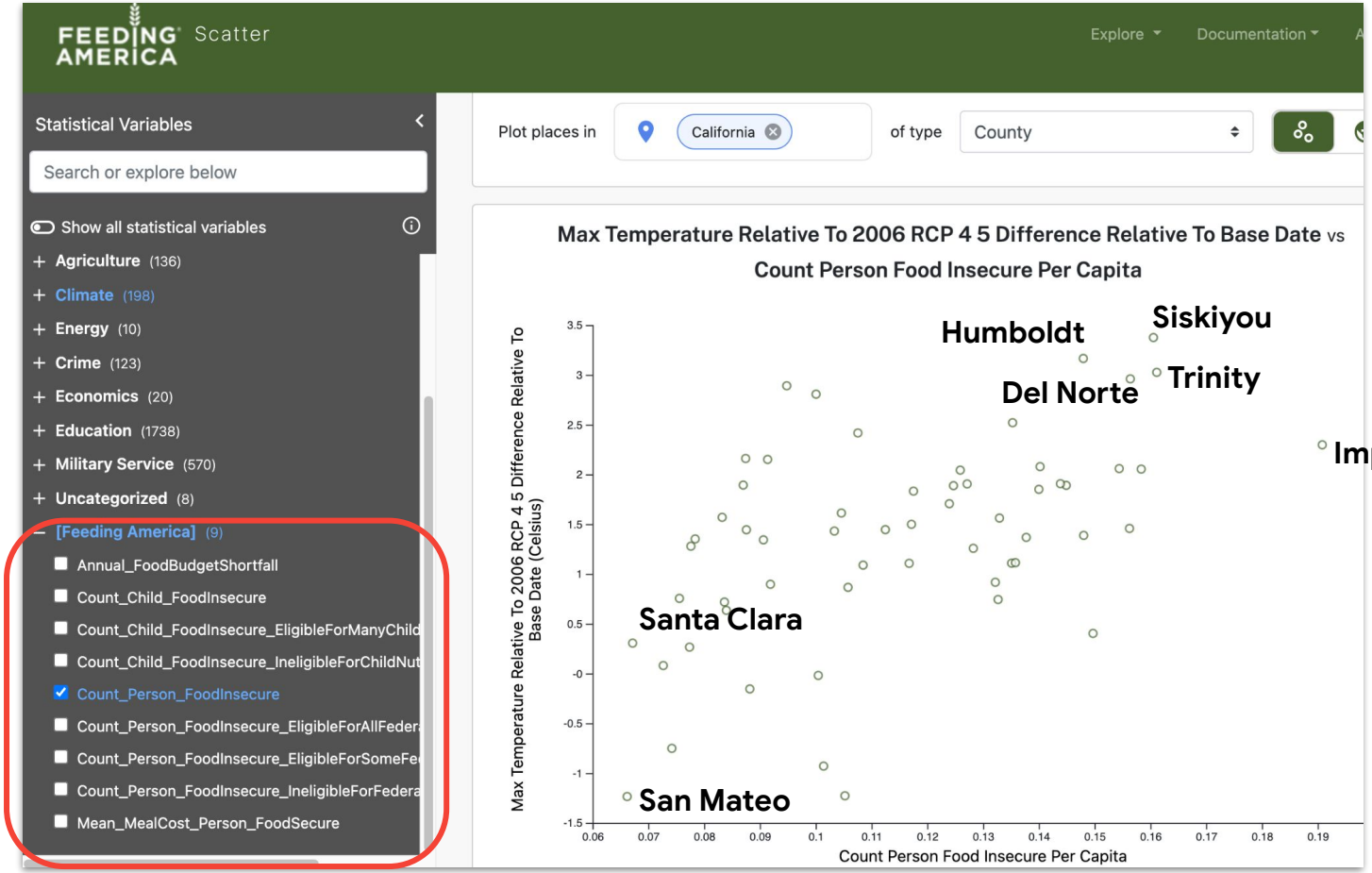


Max projected summer temperatures (RCP 4.5)

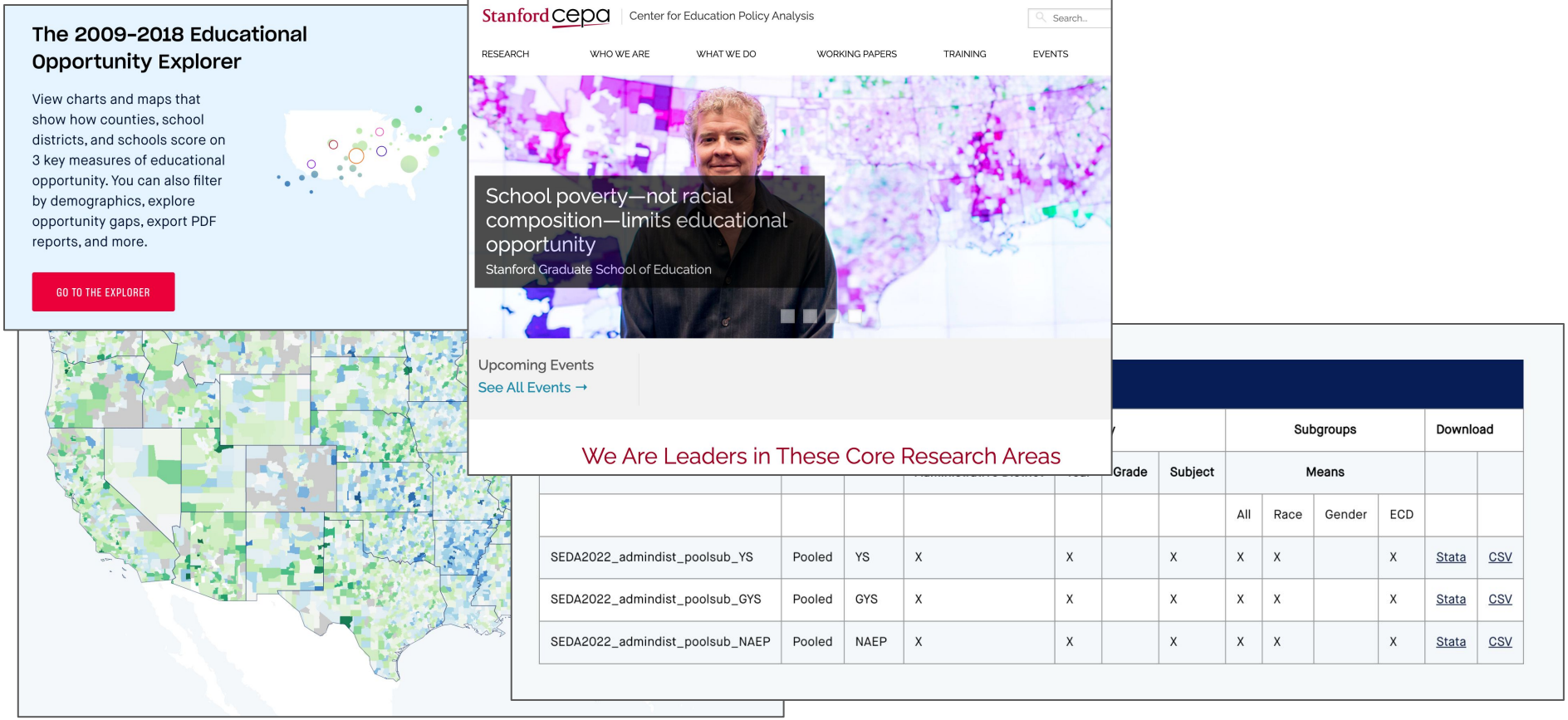
CA counties most at food risk from rising temperatures

From DataCommons.org

From FeedingAmerica

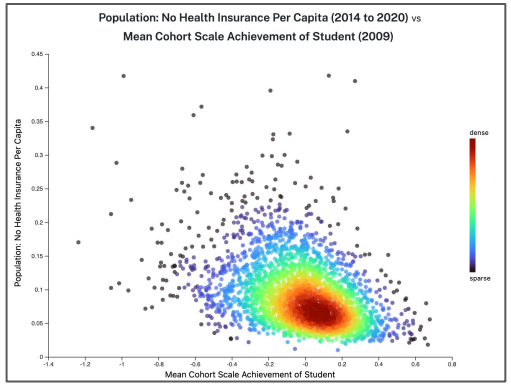


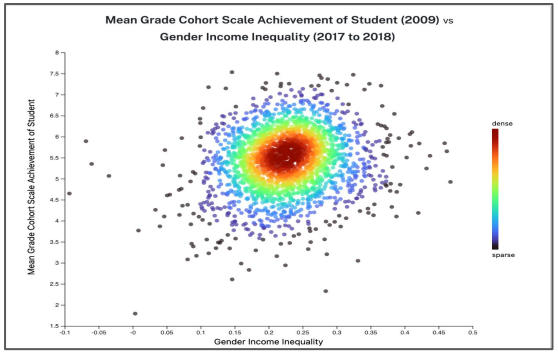
Case Study: Stanford CEPA



How does this correlate with …

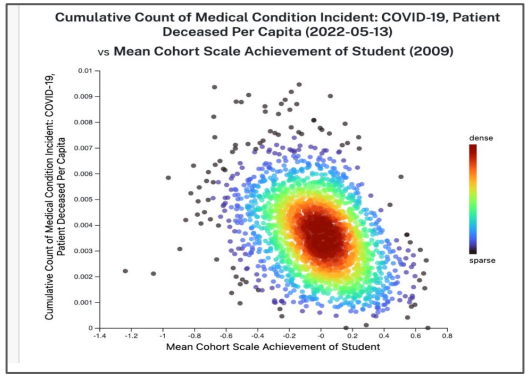
Covid deaths

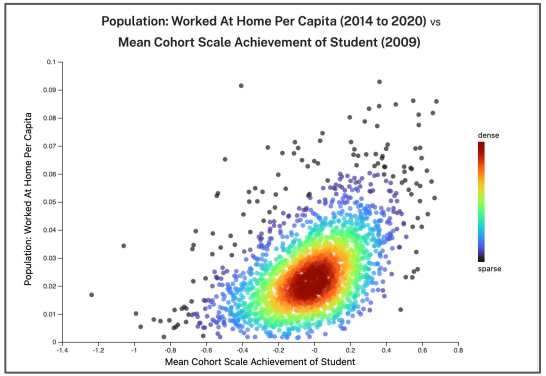




GINI Coefficient

Uninsured





Working from home

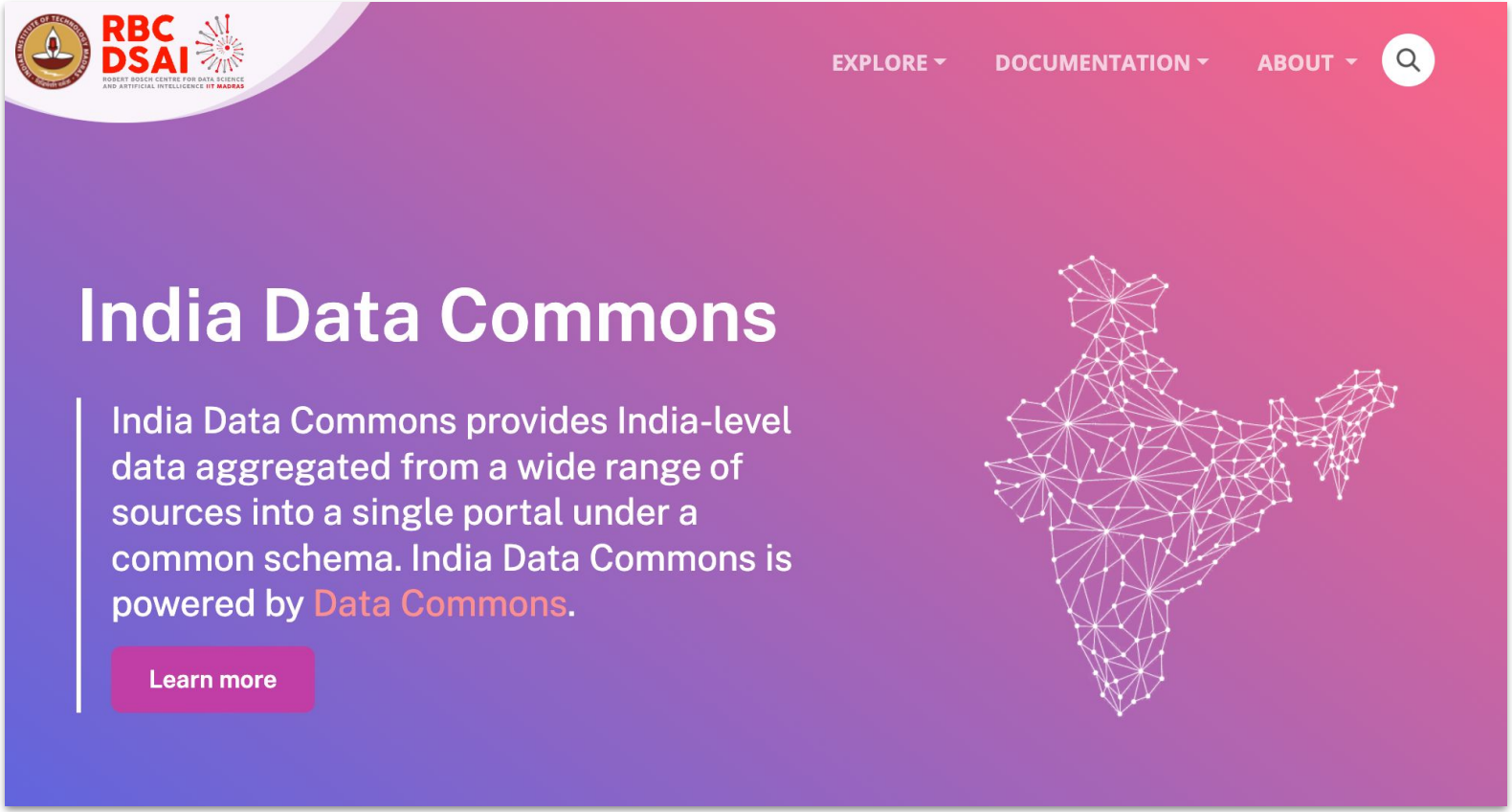
What if anyone could discover this in 10 min? Makes the CEPA data so much more useful!

Covid

Rate:
- Race
- Household size

Morbidity:
- Per Capita prescriptions

India Data Commons

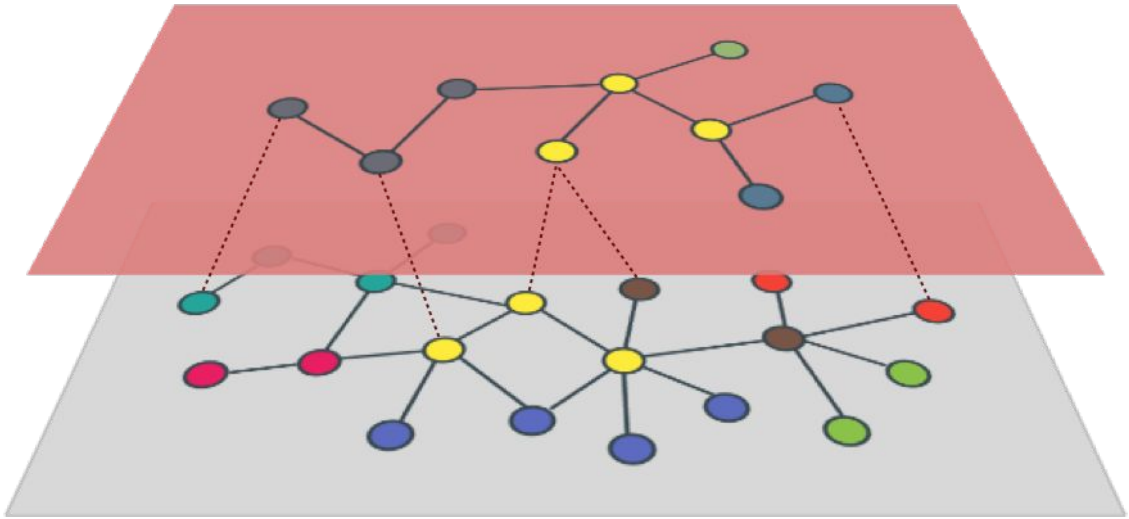


Thanks

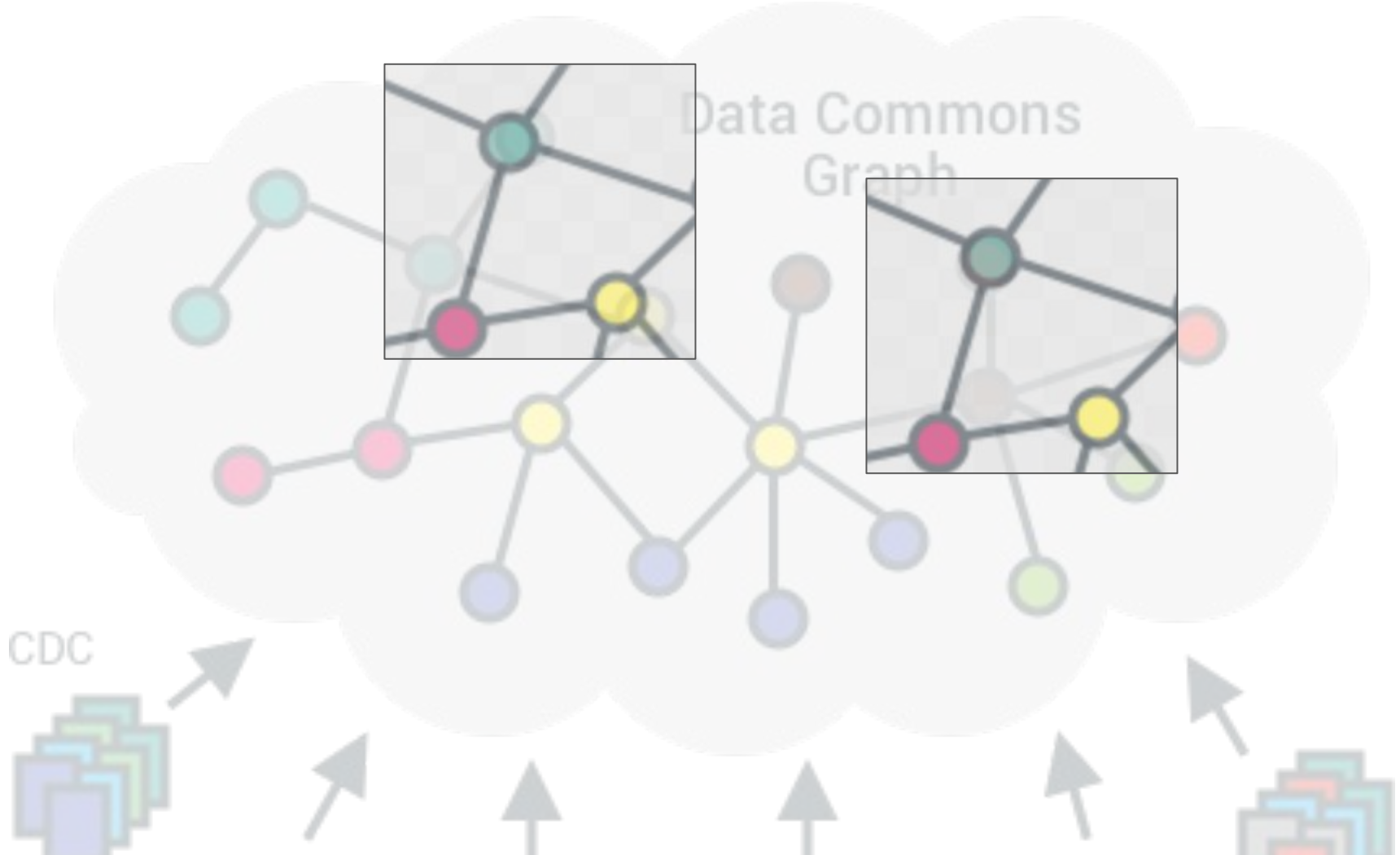
Thanks!

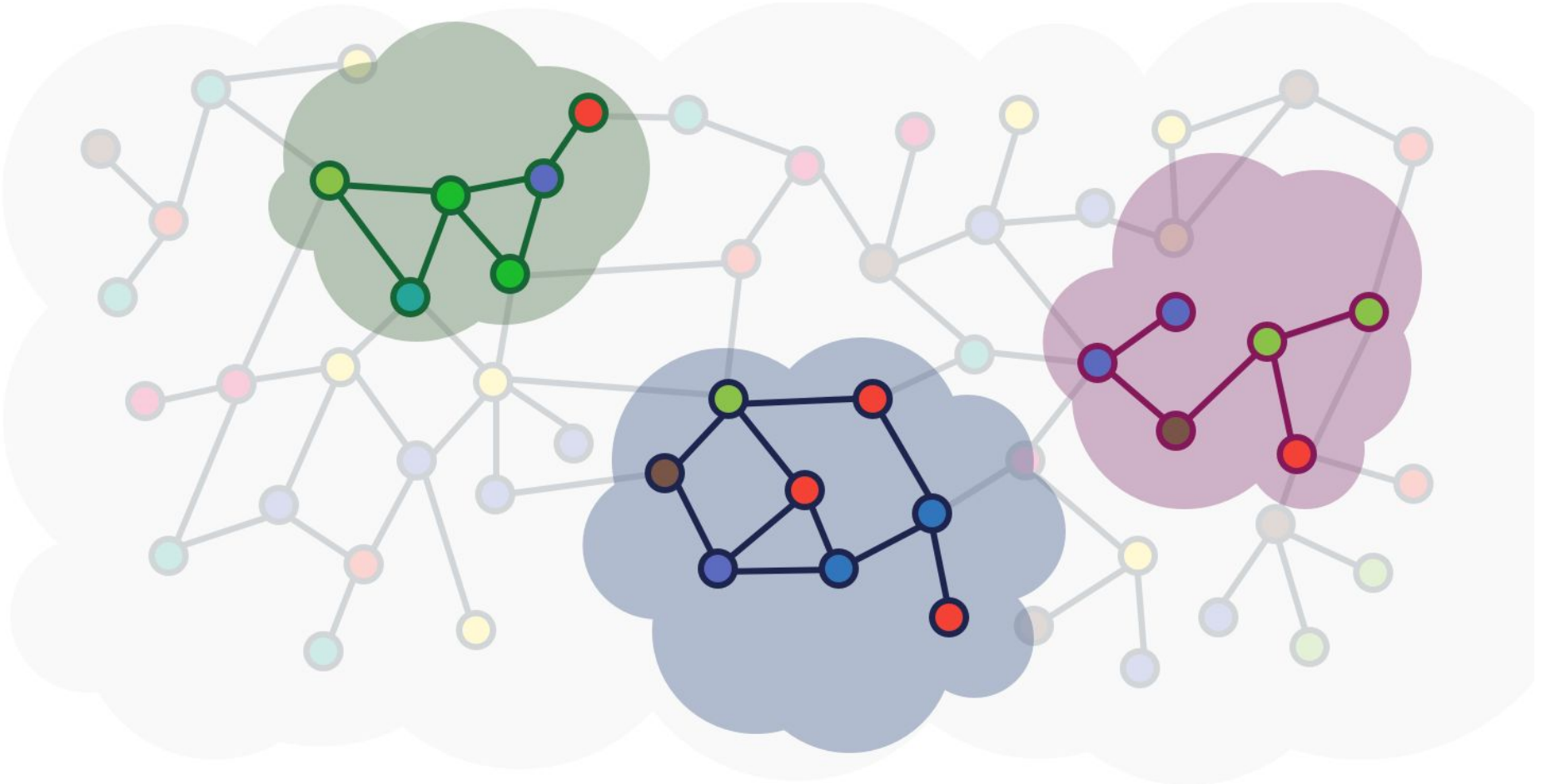
Many Data Commons, one schema, one API

- 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 ...


Data Commons
Graph
CDC



Stanford University Natural Disasters and Severe Weather Dashboard

Disasters Documentation About

Filter events Show

Drought

Severity

Earthquakes

Magnitude 5+

Floods

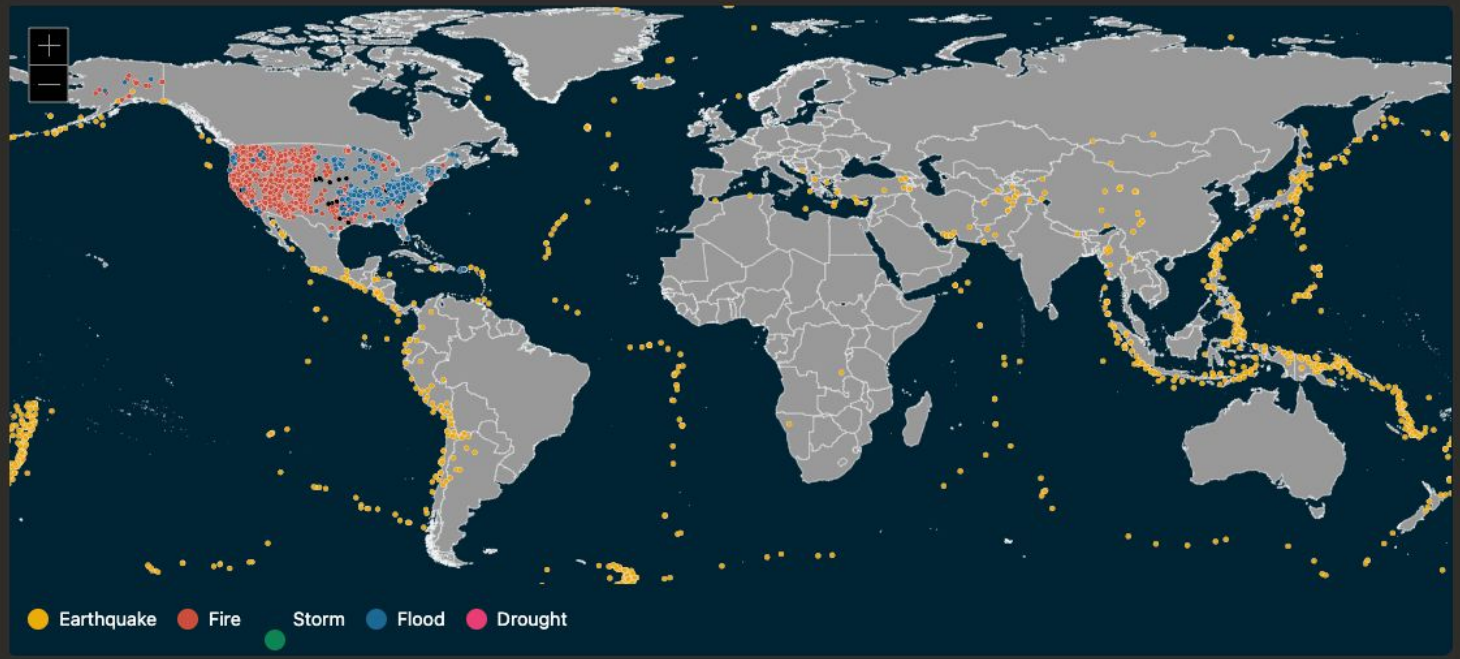
Severity

Fires

Severity

Event dates

2022

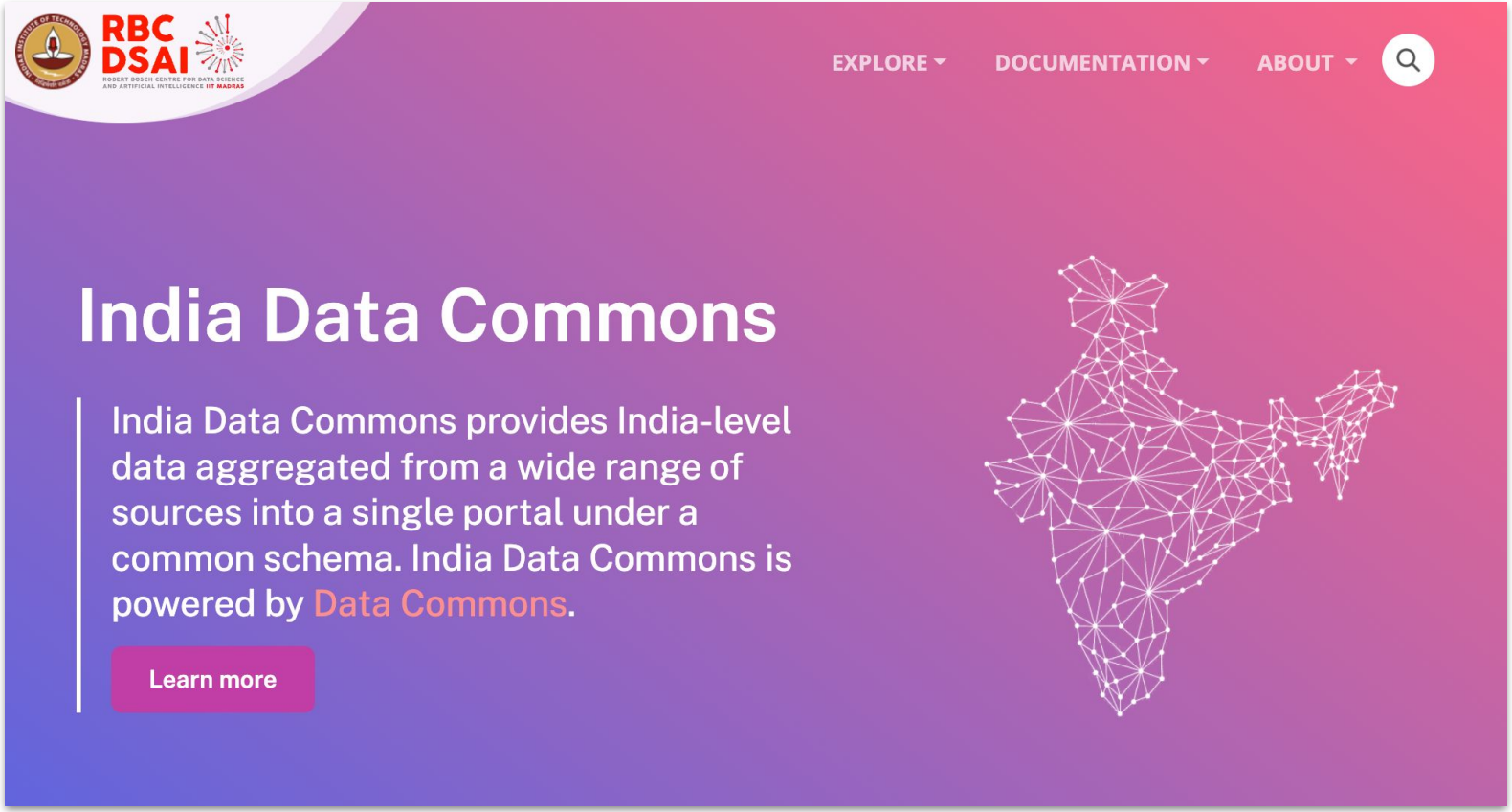


Earthquakes in 2022

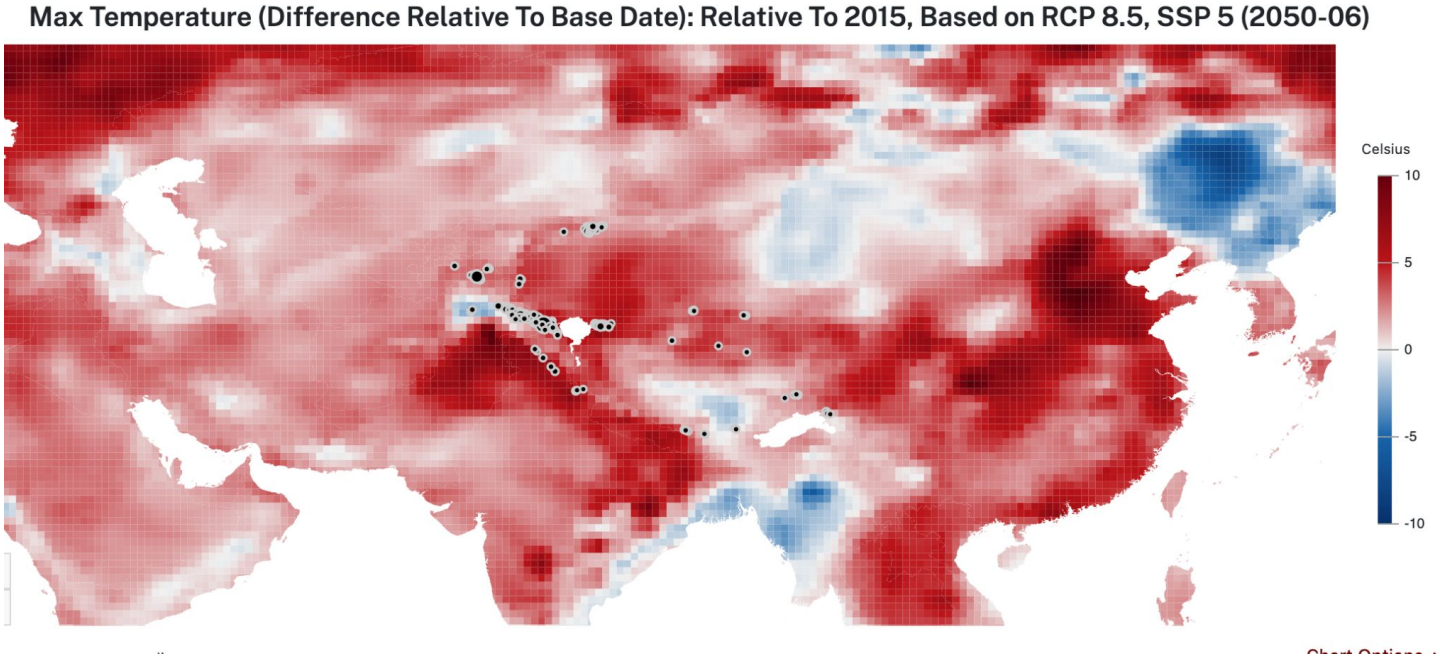
Fires in 2022

Storms in 2022

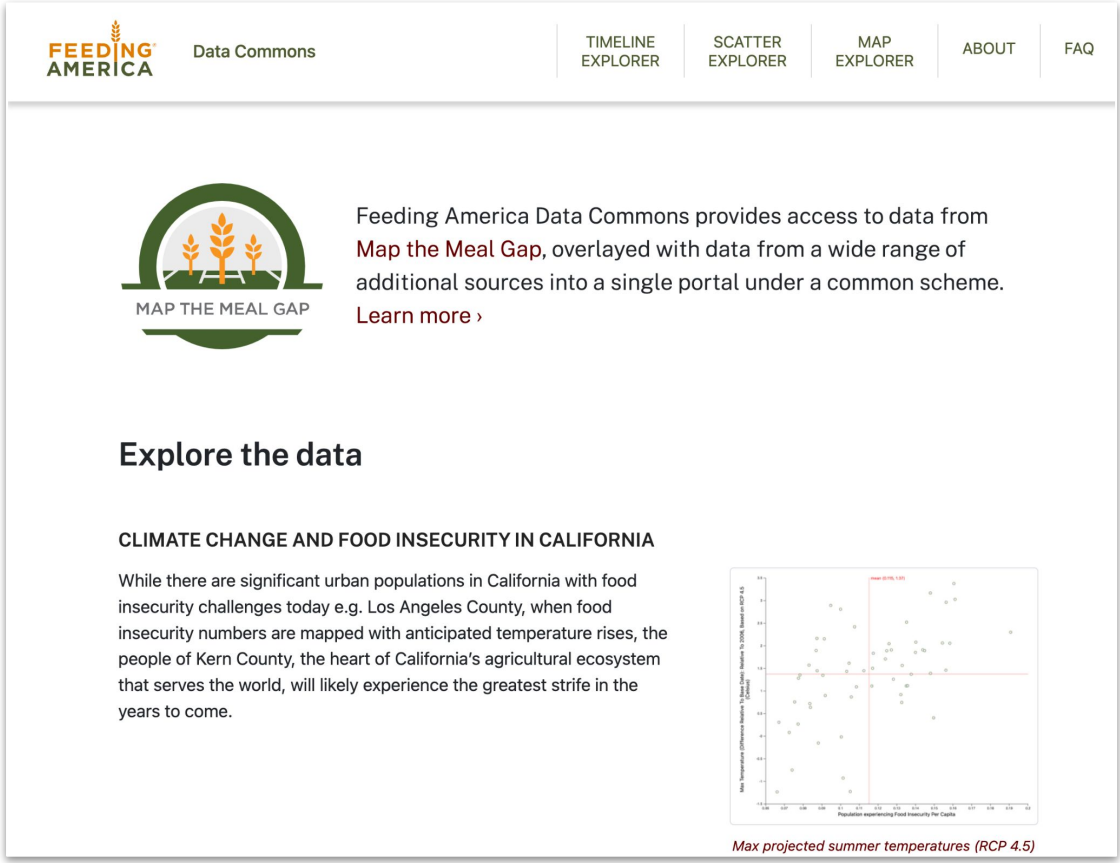
India Data Commons



Glaciers feeding the Indus, Ganga & Brahmaputra

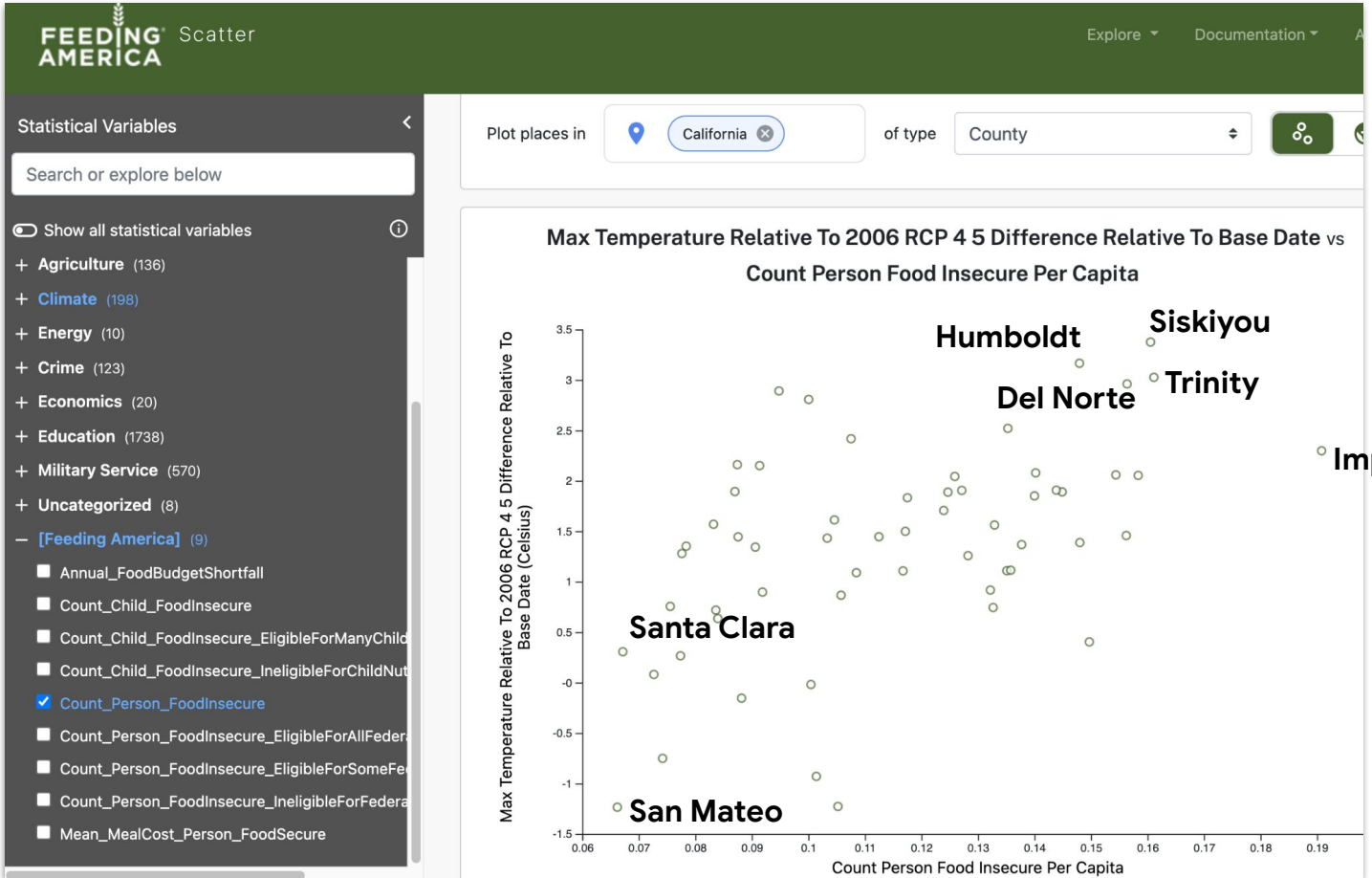


Feeding America Data Commons

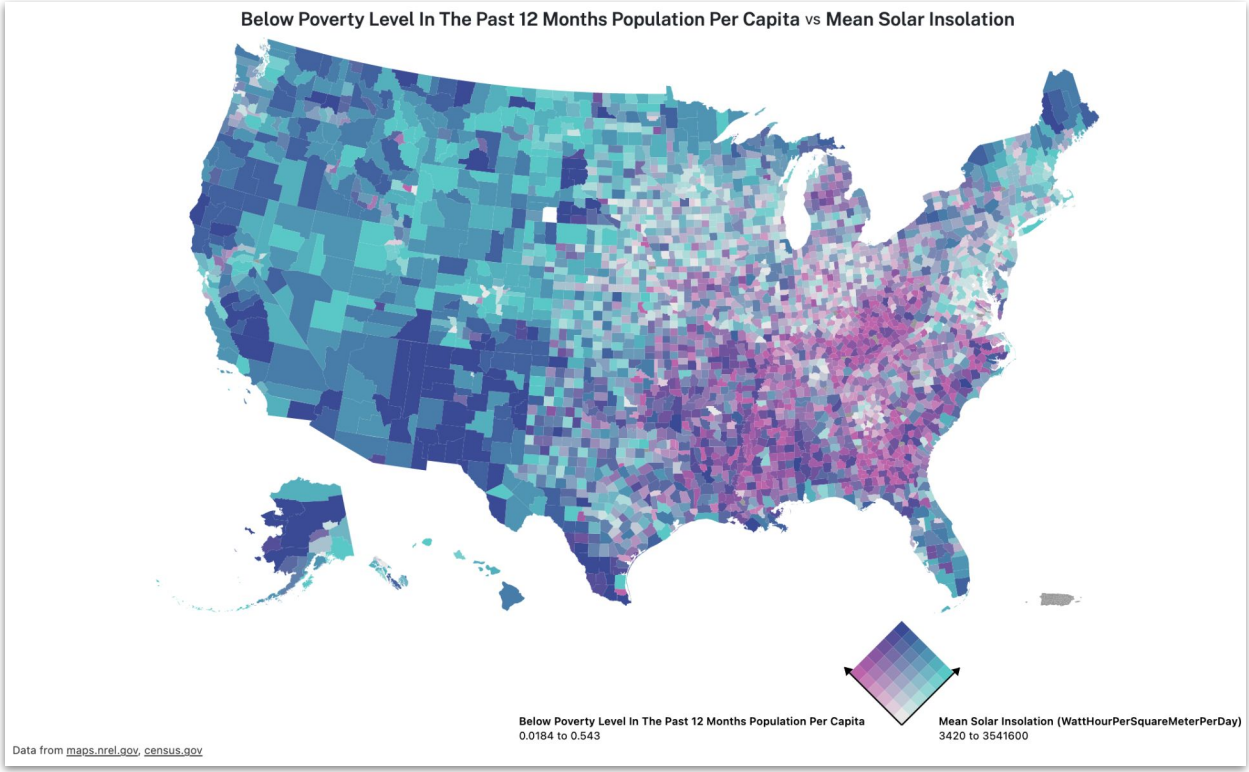


Max projected summer temperatures (RCP 4.5)

CA counties most at food risk from rising temperatures



On a positive note: Where can renewable investment have maximal impact --- Solar Energy Potential vs. Poverty

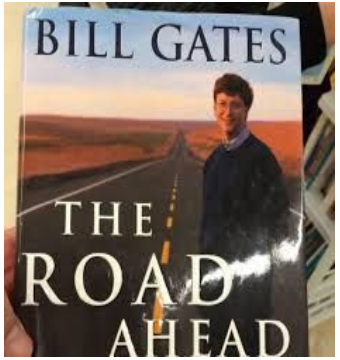


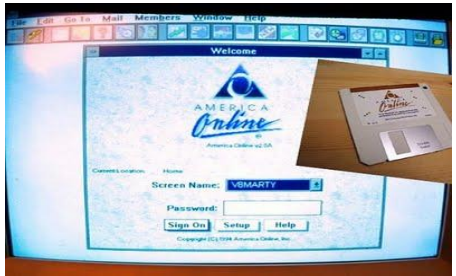
Thank you!

A lesson from the past …

In 1993, a number of players wanted to ‘own’ the net

- Microsoft was planning to spend $2B on MSN network
- AOL, the leader, was giving away 50m accounts
- IBM invested heavily in Prodigy
- Impressive tools & content partnerships for these networks
- Myhrvold’s 10 reasons for why ‘this web thing won’t work
  - identity, security, commerce, …





The Web

- Very amateurish
  - Mostly students in universities, researchers, enthusiasts
  - No authoring tool, no commerce, no security, …
- But, by end of ‘95, game over
  - May 26, ‘95, ‘Internet Tidal Wave’ memo
  - December ‘95, AOL, MSFT kill proprietary services

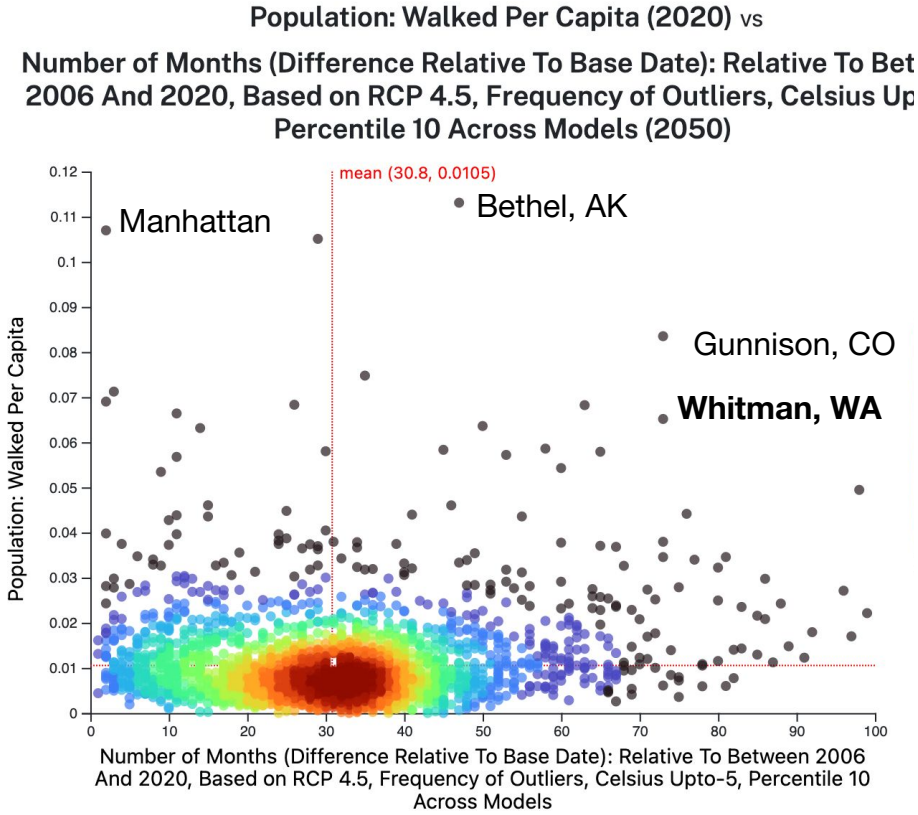
How did *that* happen?

The Web actually had far more investment!

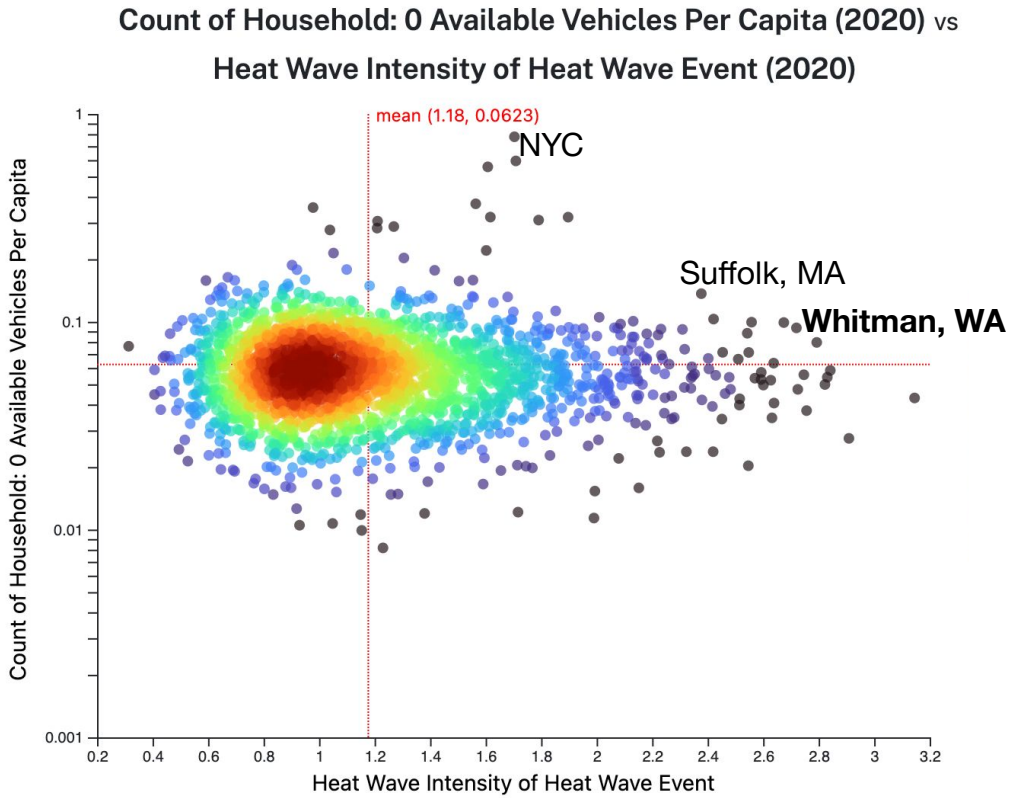
- Web recruited millions of people contributing content
  - Even the billions spent by AOL/MSFT could not compete with this
- Everybody was putting up `interesting' pages and convincing 10 others to get on the web to see their page
- Almost no barrier to entry
  - Permissionless innovation --- no biz dev deals required
  - Gradual investment/learning curves

Back in the US, so many ‘vulnerability vectors’

What about people who walk to work?



What about people without cars?

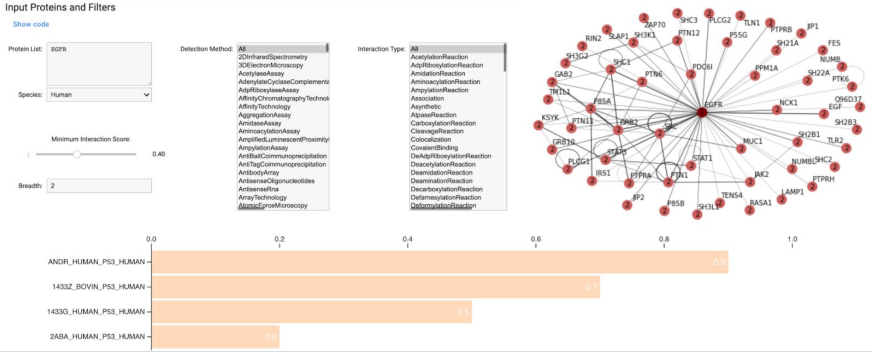


Biomedical Data Commons site (coming Q1 2023)

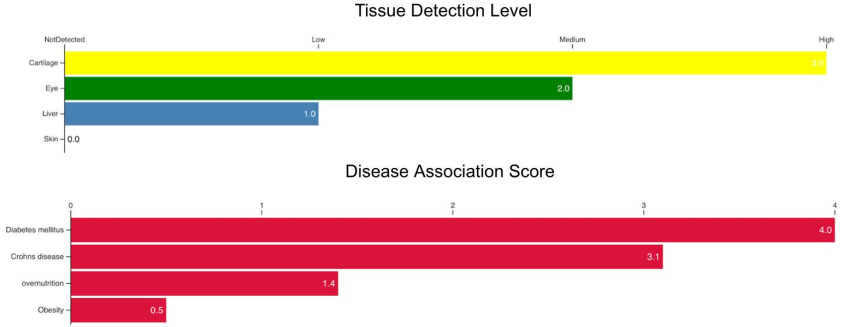
Public Health - Austin, TX

Pages for each disease, protein, place, …

Protein Browser
Protein-Protein Interactions: EGFR and P53



Protein Browser
Tissue Detection and Associated Diseases



Covid

Rate:
- Race
- Household size

Morbidity:
- Per Capita prescriptions

Other Ongoing Sustainability Applications

Bridging information gaps can reduce food waste

Schema.org collaboration for open standards for reporting Scope 1/2/3 emissions