




MORE THAN JUST MAPS

Unleashing Geospatial Intelligence 

SQL Saturday San Diego | September 14, 2024

TOPICS

- An introduction to geospatial data
- Implementation in SQL Server, Snowflake, and PostgreSQL
- Loading and processing data
- The part with maps
- A real-world example
- Privacy concerns

HELLO!

Microsoft
CERTIFIED

Solutions Expert

Data Management and
Analytics



PLTW





INTRO TO GEOSPATIAL DATA

WHAT'S SO IMPORTANT?

“It's estimated that over 80% of business data has a location context.”

- Microsoft

Location-enhanced intelligence:



Weather



Community



Economic



Proximity

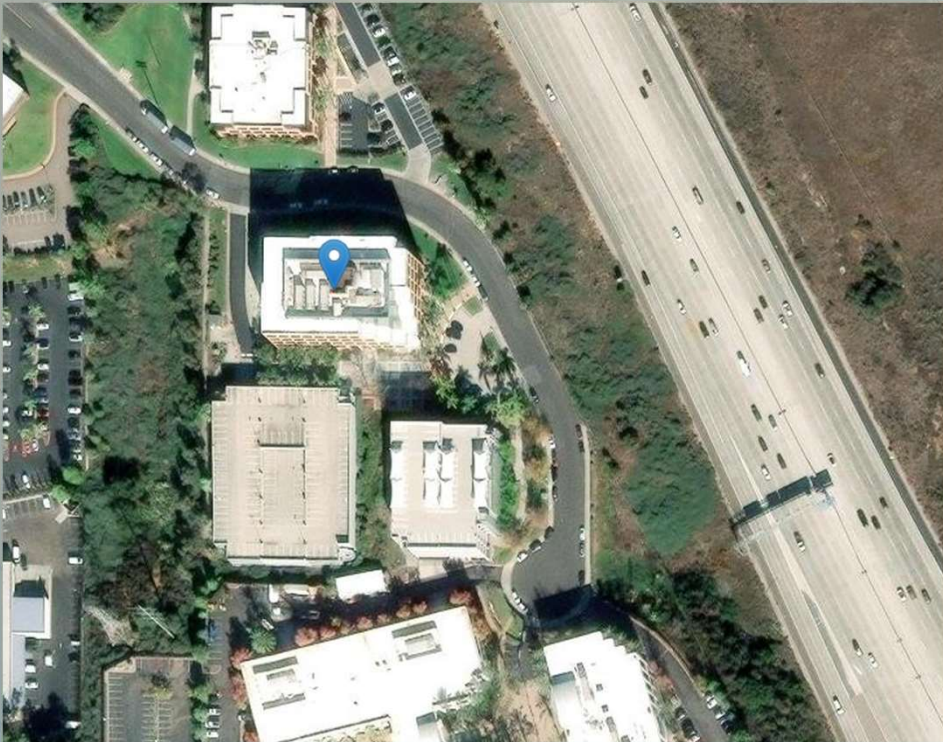


Legal

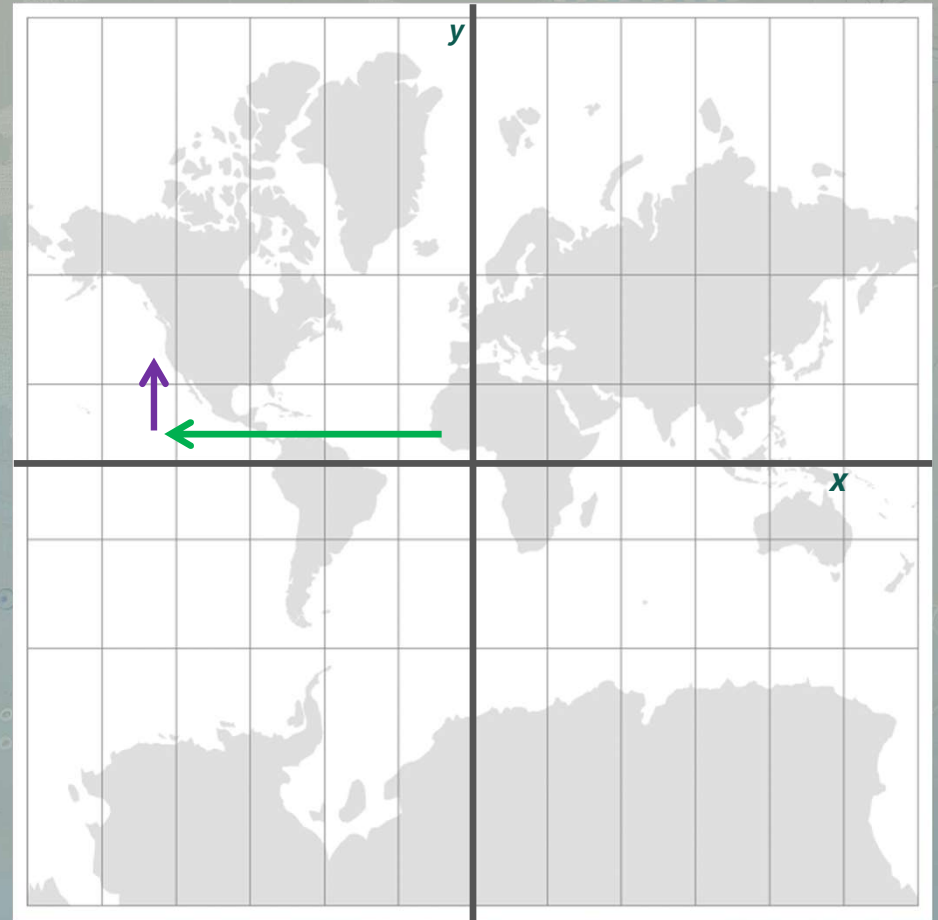
LATITUDE AND LONGITUDE

OR IS THAT LONGITUDE AND LATITUDE?

32.853°N, 117.183°W



(- 117.183, 32.853)



SHAPE DATA

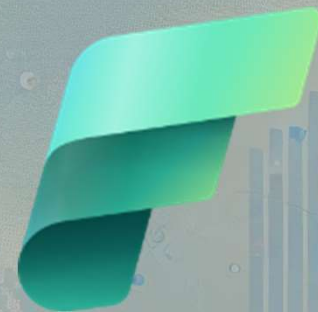
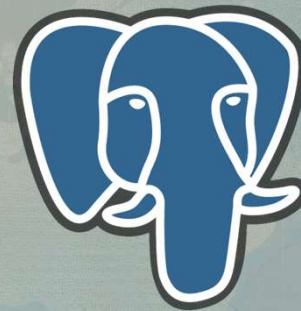
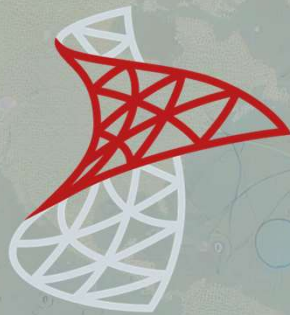
Directly supported in many data platforms

- Well Known Text (WKT) / Well Known Binary (WKB)
 - WKT is human-readable, WKB is more compact
 - Developed by Open Geospatial Consortium
 - Defines shapes **only**
 - Defined record-by-record, supported by pretty much all databases
- GeoJSON
 - One file / one record
 - Defines shape and attributes
 - Use on the rise due to open standard / extensibility
 - Closely-related to **TopoJSON**

SHAPE DATA




NOT directly supported by most data platforms

- ESRI Shapefiles
 - A collection of files in a directory – need .shp, .shx, .dbf
 - The US government publishes geospatial data in this format
- Keyhole Markup Language (KML)
 - Developed for use with Keyhole Viewer → Google Earth
 - XML-based
 - One file



DATA PLATFORM IMPLEMENTATIONS

SQL SERVER / AZURE / FABRIC? /

- GEOGRAPHY type requires Spatial Reference Identifier (SRID)
 - Typically, you'll use **4326** (EPSG 4326)
- INSERT statements use WKT/WKB with the SRID
`geography::STGeomFromText('LINESTRING(-117.1897 32.7336, -86.2944 39.7172)', 4326)`
- Spatial Indexing highly recommended
- Variety of Spatial Temporal (ST) functions
- Azure Database , Synapse SQL 
-  : Use ArcGIS extensions in notebooks or store in TopoJSON

SNOWFLAKE

- GEOGRAPHY type **only supports 4326** / WGS 84
- Supports WKT, WKB and GeoJSON
- Fields subject to the Snowflake 16MB limit
 - We'll discuss some strategies around this later
- Functions are ST_ ← Note the underscore
- Spatial indexing is automatic

POSTGRES SQL

- GEOGRAPHY datatype requires PostGIS module
 - Dates back to 2001
- Includes optional TIGER (Census Bureau) module for geocoding
- Used by Open Street Map
- Like SQL Server, spatial indexing highly recommended
- Same ST_ functions as Snowflake



LOADING GEOSPATIAL DATA

FEATURE TYPES

- **POINT**
 - A single location
- **LINESTRING**
 - Roads, coastlines, creeks
- **POLYGON**
 - Areas like parks, counties, states, Congressional districts, larger rivers, etc.
 - Polygons can have holes
 - Start and end points for a polygon / hole in a polygon must be the same
- **COLLECTION**
 - Can contain multiple objects of different types

LOADING FROM SHAPEFILES

- GeoJSON
 - Snowflake – you're good as long as shapes don't go over 16MB
 - SQL Server – use GeoJSON → WKT converters (online or code libraries)
 - Remember, you'll lose data about your features (e.g., names, codes, etc.)
- ESRI Shapefiles (Gov't sources)
 - SQL Server: shape2sql.exe (sharpgis.net) – not maintained
 - Snowflake: There's a great Medium Article from Snowflake on this; use GeoPandas
 - PostgreSQL: shp2pgsql.
 - It should be relatively easy to convert this output to other DBMS statements – it outputs WKT
- KML
 - Online converters are your best bet

CONSIDERATIONS

- Rounding / Implementation issues
 - In polygons and lines, this can cause a few types of errors:
 - Vertices cross (“edge ... crosses edge ...”)
 - Not an enclosed polygon
 - Ring not **completely** inside outer edge of main polygon
- It’s too much!
 - Snowflake goes beyond its 16MB single field limit

FIXING LOADS

- Some platforms will try to fix your shape for you
- Buffer (grow) a polygon by only a meter or two
- Simplify the shape
 - Minimum accuracy by number of digits:

32	7	1	3	8	5	9°
110 km 70 mi	11 km 7 mi	1.1 km 0.7 mi	111 m 360 ft	11 m 36 ft	1.1 m 43 in	11 cm 4 ¼ in
	Distance to La Jolla Cove	Size of the Gaslamp Quarter	Football field Rugby pitch ☺ Costco building	Average American house	Width of a desk	Height of a cell phone

Avg Phone GPS error
(5 m / 16 ft)



BASIC PROCESSING

CREATE A POINT

Function:

SQL Server	Snowflake / PostgreSQL
STPoint	ST_POINT

- Create a point object from a **longitude** and **latitude**
 - Remember this order!
- If possible, store the point, as it's geospatially indexed
 - Nearest Neighbor Calc Comparison (13K \Leftrightarrow 15K points) in Snowflake:
 - Stored points: 1m, 02s
 - ST_POINT in query: 4m, 07s

Long	Lat
(139.8 , 35.7)	

DISTANCE

Function:

SQL Server	Snowflake / PostgreSQL
STDistance	ST_DISTANCE

- **Minimum** Distance between any two objects
- If the objects overlap, the function returns 0
- **Return Value: in meters**



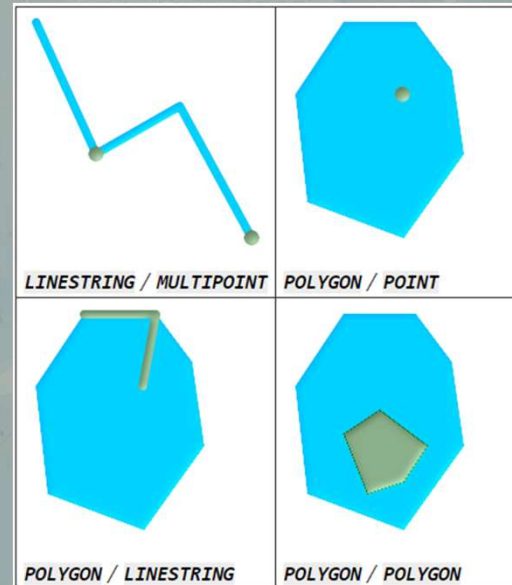
GEOFENCING

Functions:

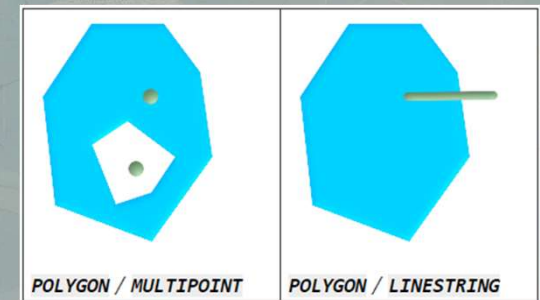
SQL Server	Snowflake / PostgreSQL
STContains STWithin	ST_CONTAINS ST_WITHIN

- Contains: Does object1 COMPLETELY contain object2?
 - “Within” is just the inverse
- **Return Value: true / false**
- **Examples:**
 - What county is this point in?

TRUE



FALSE



USEFUL FUNCTIONS

SQL Server	Snowflake / PostgreSQL	Use
Point	ST_POINT	Create a point (Longitude, Latitude[, SRID in SQL Server])
STxxxFromText	TO_GEOGRAPHY / ST_GeographyFromText	Create a geographic object from WKT <i>SQL Server: xxx = type of geographic object</i>
STArea	ST_AREA	Area of a polygon (sq. meters)
STDistance	ST_DISTANCE	Closest distance between two geographic objects (meters)
STContains	ST_CONTAINS	Does the first object completely contain the second?
STWithin	ST_WITHIN	Inverse of Contains
STUnion	ST_UNION	Combines two objects
STIntersection	ST_INTERSECT	The overlap between two objects
STIntersects	ST_INTERSECTS	Is there any overlap between the two objects?

NEAREST NEIGHBOR

(AKA: What is my closest ...)

Performance Tips

In Snowflake, point construction (ST_POINT) at runtime slows down performance.

- If possible, store the point, as it's automatically geospatially indexed
 - Nearest Neighbor Calc Comparison (13K \Leftrightarrow 15K points) in Snowflake:
 - Stored points: 1m, 02s
 - ST_POINT in query: 4m, 07s

SQL Server: You have to create the spatial index.

Before calculating distance:

If spatial indexed: `WHERE DISTANCE < x`

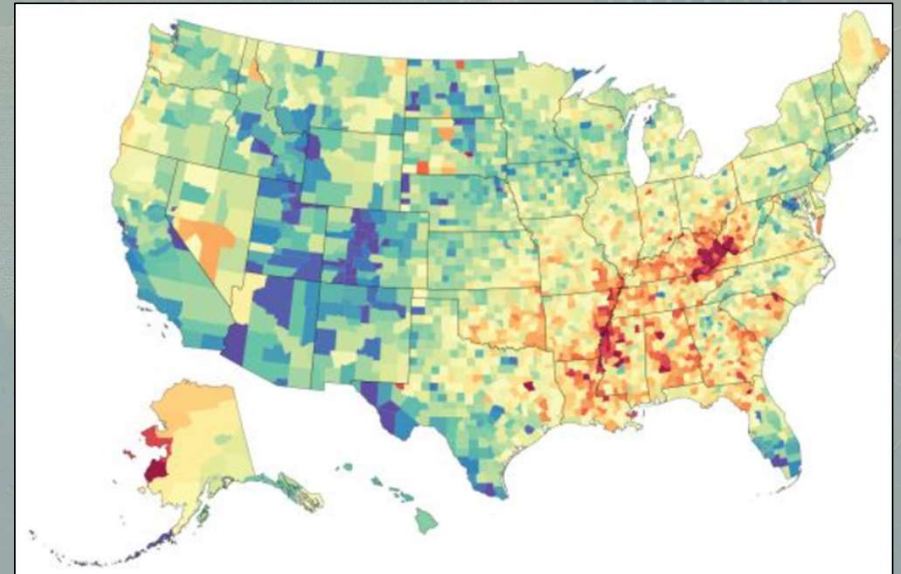
Else: `WHERE ABS(LAT2-LAT1) < x AND ABS(LON2-LON1) < x`

The background of the slide is a world map with a textured, slightly grainy appearance. Overlaid on the map are various data visualization elements. In the top left, there's a circular chart with concentric rings. To its right is a bar chart with several vertical bars of varying heights. In the top right, another bar chart is visible. The bottom left features a network diagram with nodes and connecting lines. The bottom right has a circular chart with segments. The map itself is colored in shades of green and blue, representing land and water. The overall aesthetic is technical and data-driven.

MAXIMIZING MAPS

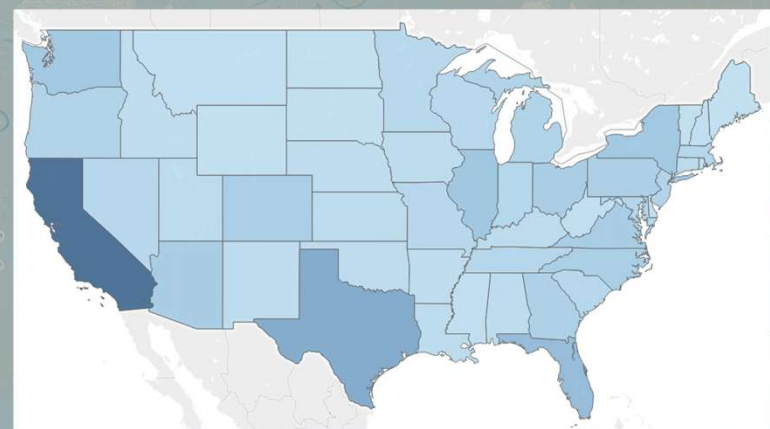
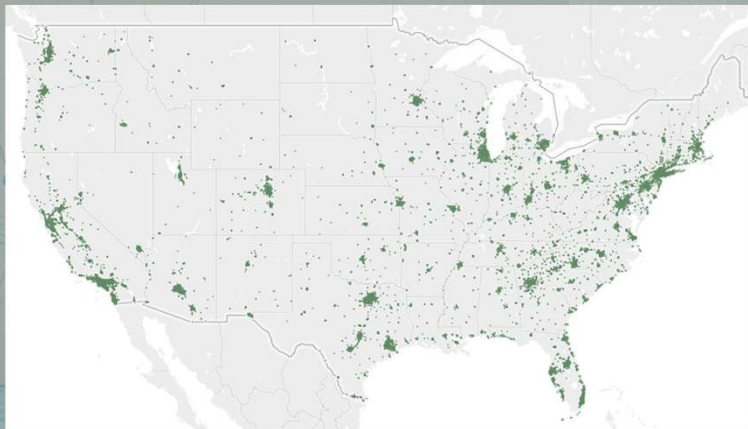
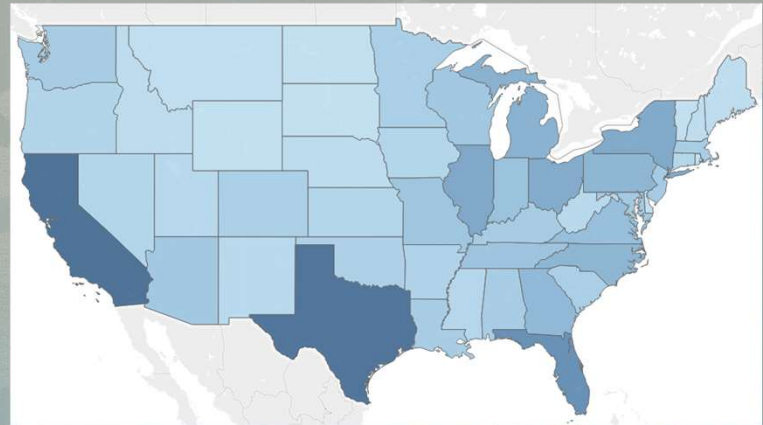
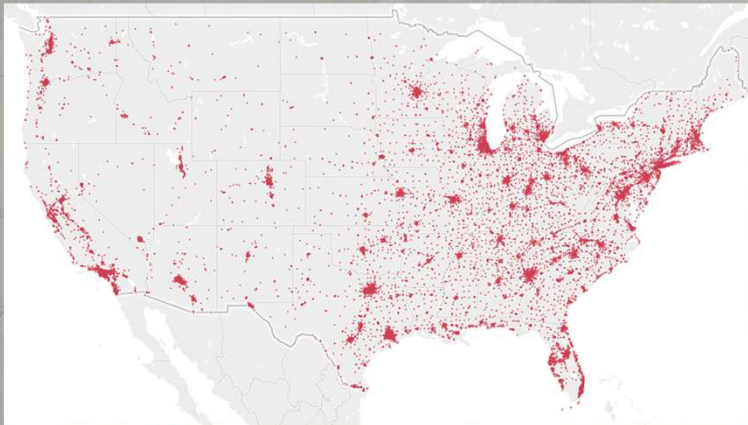
VISUALIZING GEOSPATIAL DATA

- Position is the most effective visual channel (*Dr. Tamara Munzner*)
 - Maps take a lot of space
 - Don't waste or distort it
- Good
 - Highlight spatial differences / correlations
 - Show proximity
 - Visualize the physical (e.g., weather)
- Not as good
 - Top / bottom values
 - Audience not familiar with geography



Age-standardized cancer mortality rates by county
Journal of the American Medical Association

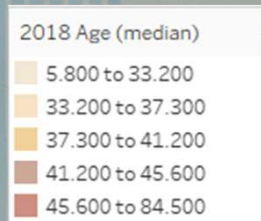
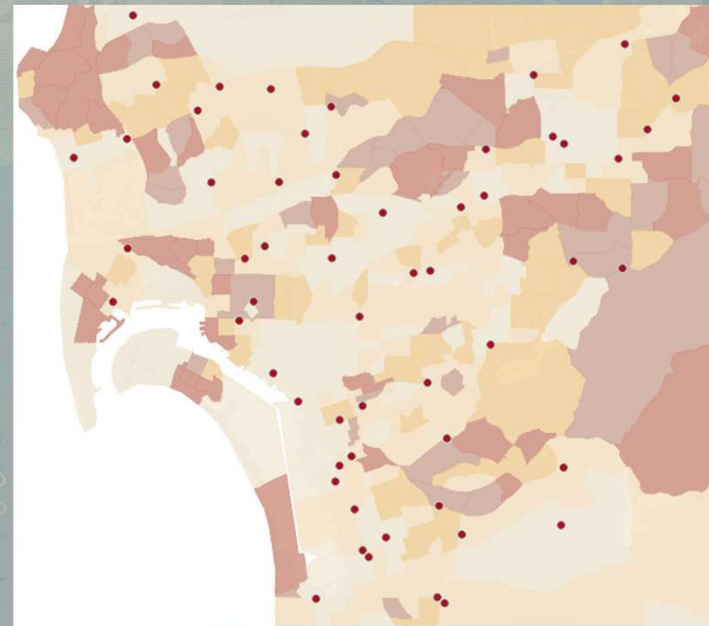
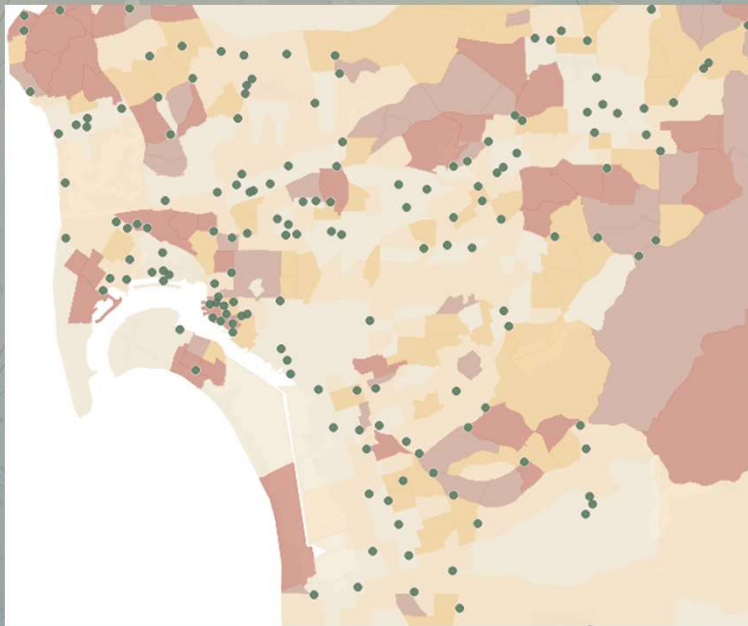
POPULATION DENSITY?



CONSIDERATIONS

Maps look cool.

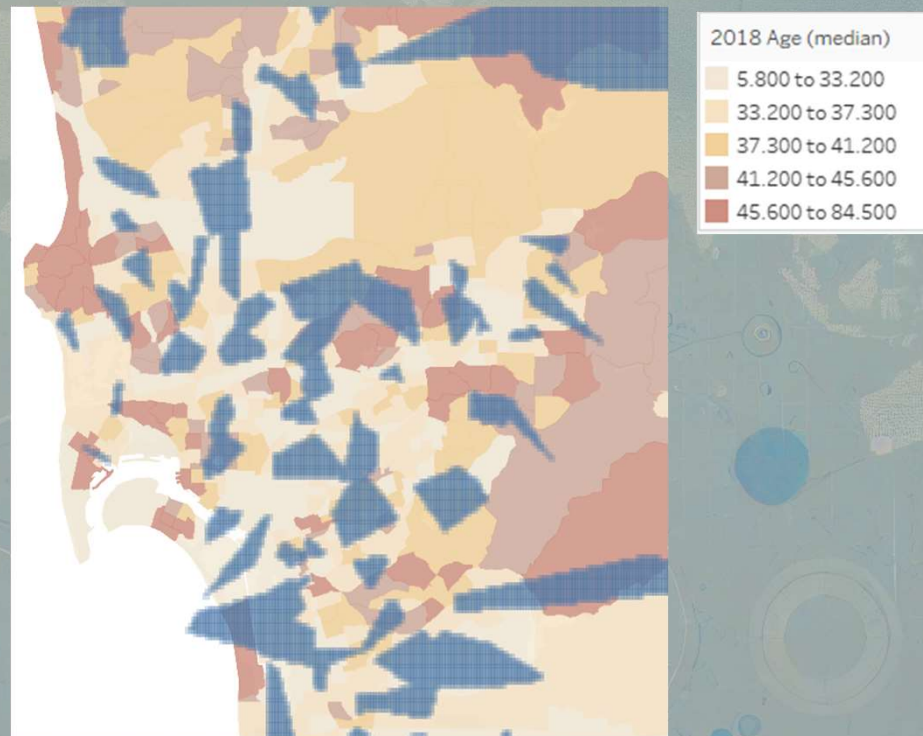
Maps ARE cool **if** there a specific spatial relationship to highlight



Starbucks (left) and McDonald's Locations in the San Diego Area

CONSIDERATIONS

Adding some analysis helps even more



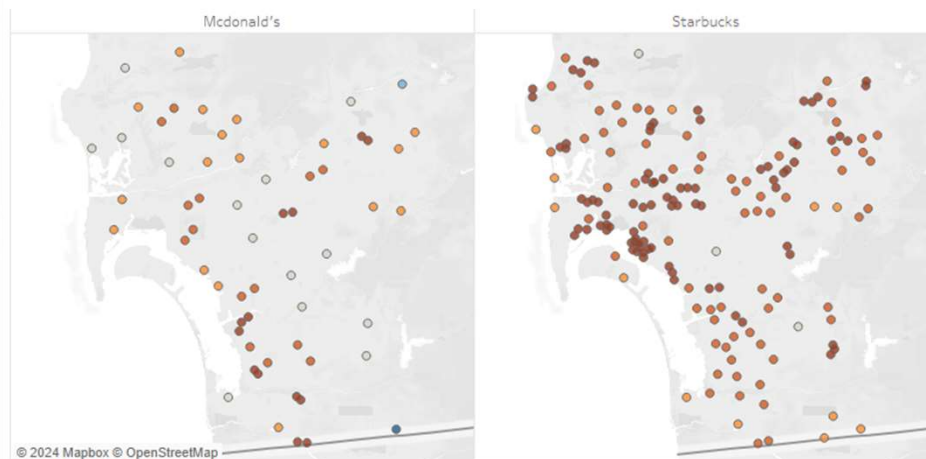
Areas in San Diego (blue) that are closer to a McDonalds than a Starbucks

CONSIDERATIONS

After doing some data set engineering, use traditional charts to complement maps.

Closest Location Brand-to-Brand

Distance to Nearest Same Brand - Map



Average Distance to Nearest...

From Locati...	To Nearest Location
Mcdonald's	Mcdonald's 2.20km
Starbucks	Starbucks 0.44km
Starbucks	Mcdonald's 1.27km
Mcdonald's	Starbucks 1.12km

Distance to Nearest Same Brand - Distribution



CHOROPLETH MAPS

USA

- Putting all states in a single map creates large dead space (> 70%)
- Small states / countries can have big populations, and vice versa



(population)

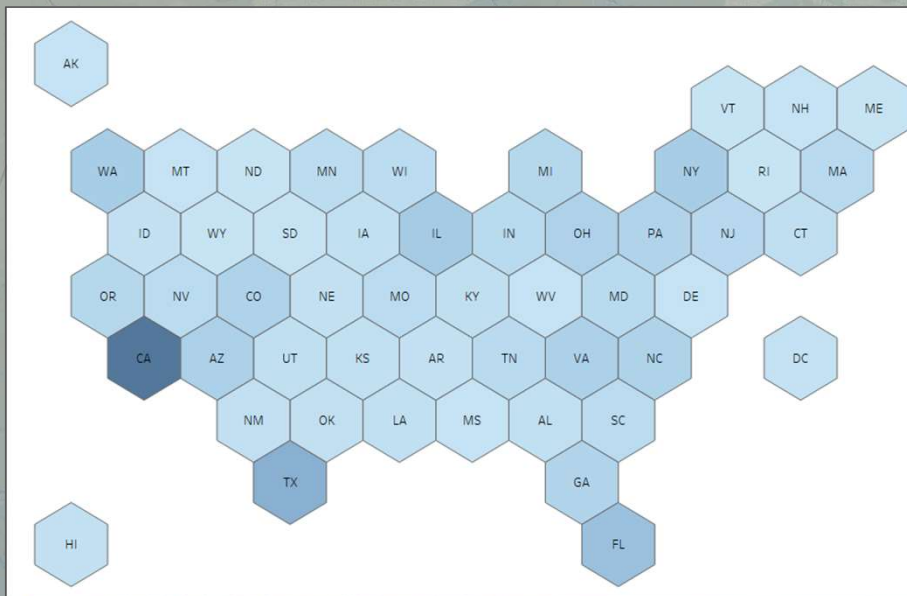


World: “Valeriepieris circle” →



CHOROPLETH CREATIVITY

- USA: Make cutouts of Northeast / AK / HI
- USA: Hex Map (shape file / GeoJSON / WKT on my GitHub)



- World: Cutouts with areas of interest

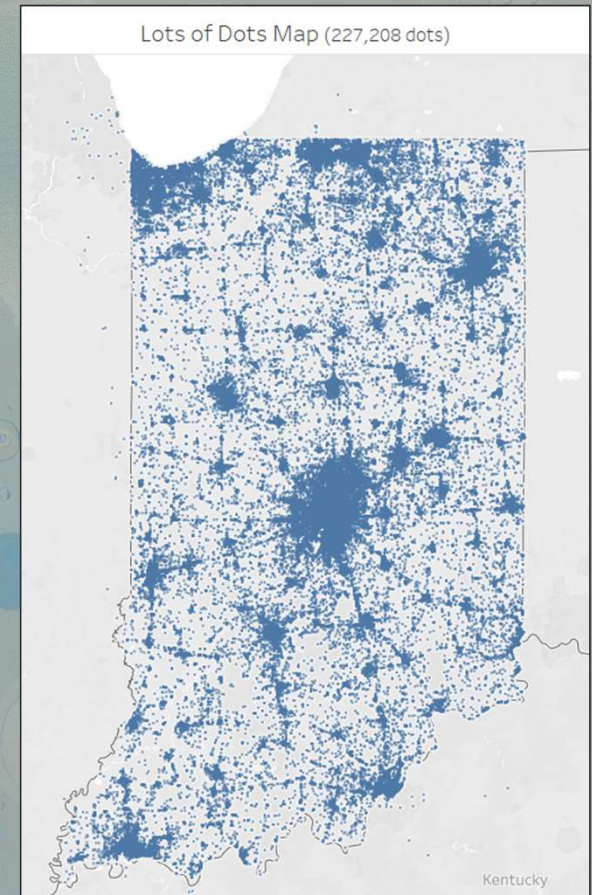
MAPPING IN POWER BI

NATIVE:

- Visualizing maps requires a round-trip to Azure / Azure servers for geocoding
- Can visualize **points**, but not polygons or lines*
 - Maximum: 3,500 (legacy) or 30,000 (Azure maps)
 - Those can take a while to load
- Shapes are limited to geographies like states, counties (TomTom data)
- Custom shapes
 - Shape Maps – conversion to a format called TopoJSON required
 - ArcGIS plugin i\$ available too
- You may be better off visualizing your engineered relationships

MAPPING IN TABLEAU

- Can visualize geographies
 - Understand Postgres, MS SQL Server, Snowflake geographies
 - Also reads shape files and GeoJSON files
- Overlays to Open Street Map
- Supports multiple map layers
 - Built in layers for certain geographical statistics
- Doesn't round-trip to display coordinates, shapes
 - Caches other geographies (states, counties, etc.)





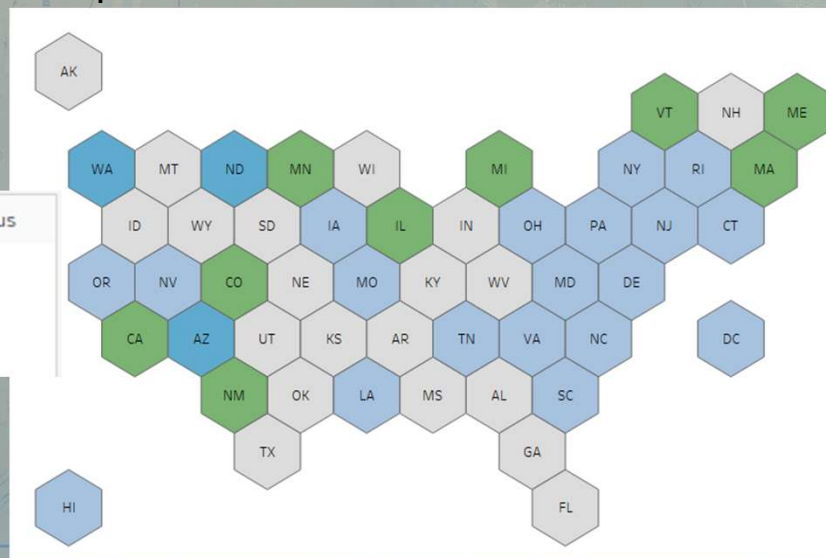
Geospatial + Data Science

REAL-WORLD EXAMPLE (IN PROGRESS)

THE CHALLENGE

A Great thing for kids!

- More states providing free meals to all students (FRPL program)
- Increases food security and academic performance



A challenge to measure

Measuring economic need in a school:

	FRPL %	Title I	SAIPE (Census)
Continuous	✓	✗	✓
School-level	✓	✓	✗
Individual focus	✓	✓	✗
Nationwide	✗	✓	✓

ESTIMATING FRPL ELIGIBILITY

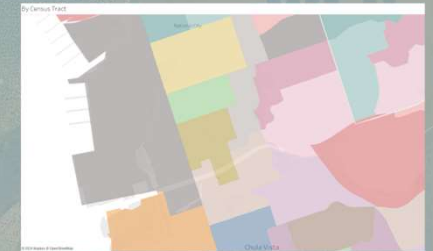
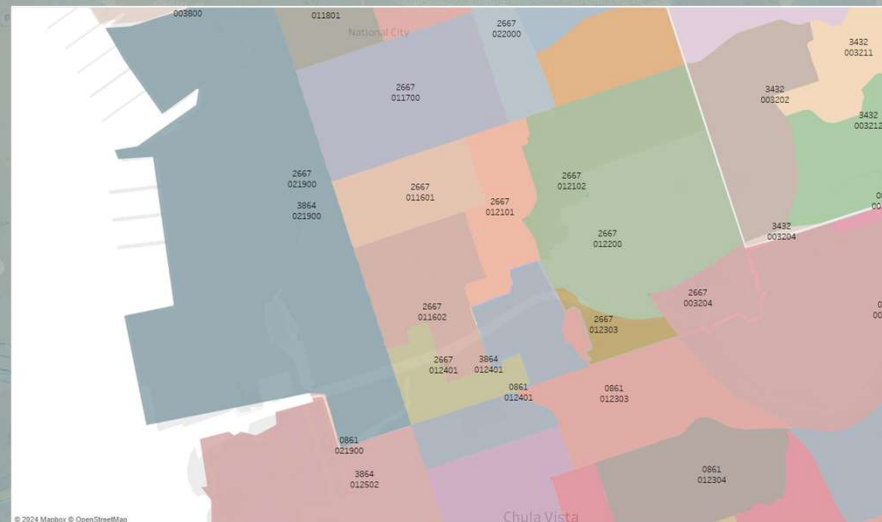
Data Science Project

- Outcome: % of student population < 185% of FPL (FRPL Threshold)
- Input variables:

Variable	Census Tract	School District	School
School location	✓	✓	✓
School / school age population	✓	✓	✓
Children < 185% FPL	✓		
SAIPE (% of students in poverty)		✓	
Income-to-Poverty Ratio (+ S.D.)		✓	

COMBINING DATA

- Need for poverty data more granular than either district or tract
- Solution: overlay them by using ST_INTERSECTION



CALCULATING

- Divvy up population and poverty statistics by land area
 - School age population
 - School age population below Federal Poverty Level (FPL)
 - School age population below 185% FPL
- For each school
 - Take statistics for the smaller area it's in
 - If # children in area < school population: add next closest area
 - Repeat until we've hit the school population
 - Output:
 - School population below FPL
 - School population below 185% FPL

NOW, IT'S JUST A REGRESSION MODEL

Geospatial Poverty Estimate

Raw district poverty number
from Census Bureau

School Title I Status

```
print(LR_statsmodels.summary())
```

OLS Regression Results

```
=====
Dep. Variable:          FRLP Ratio    R-squared:                0.707
Model:                  OLS          Adj. R-squared:            0.707
Method:                 Least Squares   F-statistic:              3098.
Date:                   Mon, 02 Sep 2024   Prob (F-statistic):       0.00
Time:                   18:01:54         Log-Likelihood:           3355.9
No. Observations:       6432           AIC:                     -6700.
Df Residuals:           6426           BIC:                     -6659.
Df Model:               5
Covariance Type:        nonrobust
=====
```

	coef	std err	t	P> t	[0.025	0.975]
const	0.5826	0.009	62.298	0.000	0.564	0.601
Estimated Poverty Ratio	0.1203	0.022	5.356	0.000	0.076	0.164
IPR Estimate	-0.0555	0.001	-41.719	0.000	-0.058	-0.053
IPR Standard Error	-0.0198	0.005	-4.096	0.000	-0.029	-0.010
District Poverty Ratio	0.6030	0.036	16.612	0.000	0.532	0.674
School Wide Title I Indicator	0.1894	0.005	41.402	0.000	0.180	0.198

```
=====
Omnibus:                28.221    Durbin-Watson:           2.028
Prob(Omnibus):           0.000    Jarque-Bera (JB):         34.395
Skew:                   -0.085    Prob(JB):                 3.40e-08
Kurtosis:                3.316    Cond. No.                 84.7
=====
```

These variables explain 70.7%
of variance in FRLP numbers

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

```
targ_pred = LR_statsmodels.predict(feats_test_sm)
targ_pred = pd.DataFrame(targ_pred)
```

We can “expect” to be off in any
single estimate by about 14%

```
print('RMSE:', np.sqrt(metrics.mean_squared_error(targ_test, targ_pred))) # Root Mean Square Error
```

```
RMSE: 0.13984181224181078
```



PRIVACY & ETHICAL CONSIDERATIONS

GEOLOCATION

Your Phone: The greatest tracking device ever

Movement of a “Senior Defense Dept. Official” at the Women’s March (DC / Jan 21,2017)

The New York Times ONE NATION, TRACKED

Opinion | THE PRIVACY PROJECT

**Twelve Million Phones,
One Dataset, Zero Privacy**



CREATING PERSONAL INFORMATION

Personal Information (CCPA):

Information that either:

- Identifies
- relates to
- describes
- is reasonably capable of being associated with
- could reasonably be linked

with a particular person or household

⚠ Many people can be identified by their location at 4:00 AM and 2:00 PM weekdays

What places do I visit?

How much time do I spend there?

What routes do I drive? Or do I take public transportation?

Taxi drivers have been identified from the “anonymous” NYC Taxi Dataset!

REDUCE ORGANIZATIONAL RISK

- Make sure to get consent
 - Exactly what you're collecting
 - The purpose for which you're using the data
 - Whether / how long you intend to store it
 - **If any of the above change, obtain new consent**
- Principle of Data Minimization
 - Don't store raw data any longer than needed
 - Introduce significant rounding / error if you need to keep on to point data
 - Apple offers this in iOS 14+
 - Be careful how you introduce error

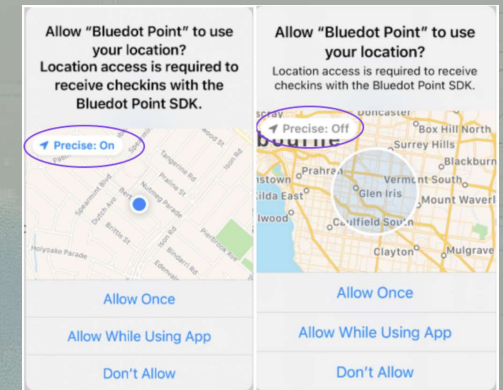
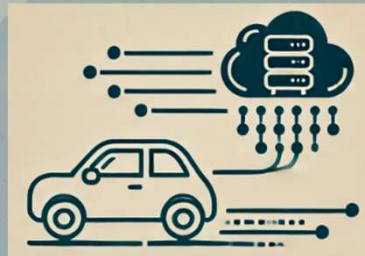
MITIGATING PERSONAL RISK

Does that app really need to know your location?

- Apps from reputable companies
 - These will legally have to tell you what exactly they'll do with your location data
 - Turn “Precise” off if possible
 - iOS: *read the purpose statement for location collection*

Modern connected vehicles

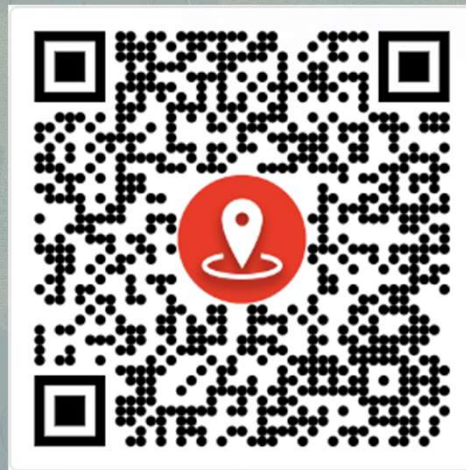
- Read the fine print for what the company can do with your location data
- You have the right to say “no”



THANKS FOR ATTENDING!

Questions?

Resources @ https://github.com/IDreamInSQL/geospatial_resources



 brittongray