🇨🇱 Data Council: Austin

Scaling Spatial Analytics with Spatial Indexes and the Cloud

Matt (Matthew) Forrest
CARTO

CARTO

Our agenda

CARTO

Why we are here today
GIS and spatial thinking
Welcome spatial indexes
Workshop

- Data exploration
- Performance face-off
- Visualizing indexed data
- Spatial analysis
- Spatial Autocorrelation

Outcomes and wrap-up

CARTO

Why we are here today

CARTO

Geospatial data is being produced at a massive scale

1. Size
2. Velocity
3. Complexity

What is the largest geospatial dataset?

Here are some examples from BigQuery Public Data (from July 18, 2022)

🛰️ NOAA Global Forecast System: 9B rows, 193TB

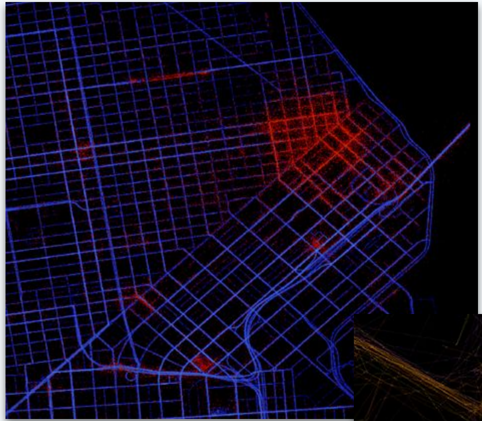
🗺️ OpenStreetMap: 993M rows, 327GB

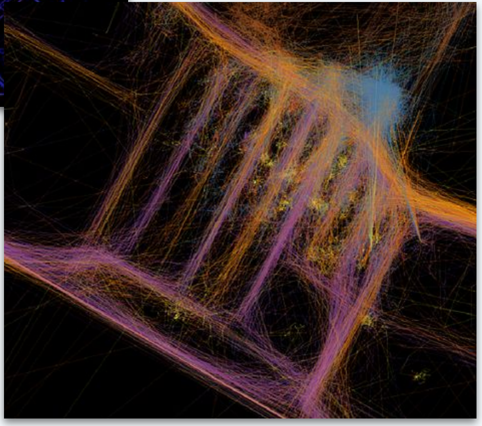
🚕 NYC Taxis Dataset: ~115M rows, ~15 to 18GB (per year, per taxi type - i.e. Yellow, Green, Uber/Lyft)

👫 WorldPop 1km Population Grids: 4.6B rows, 858GB

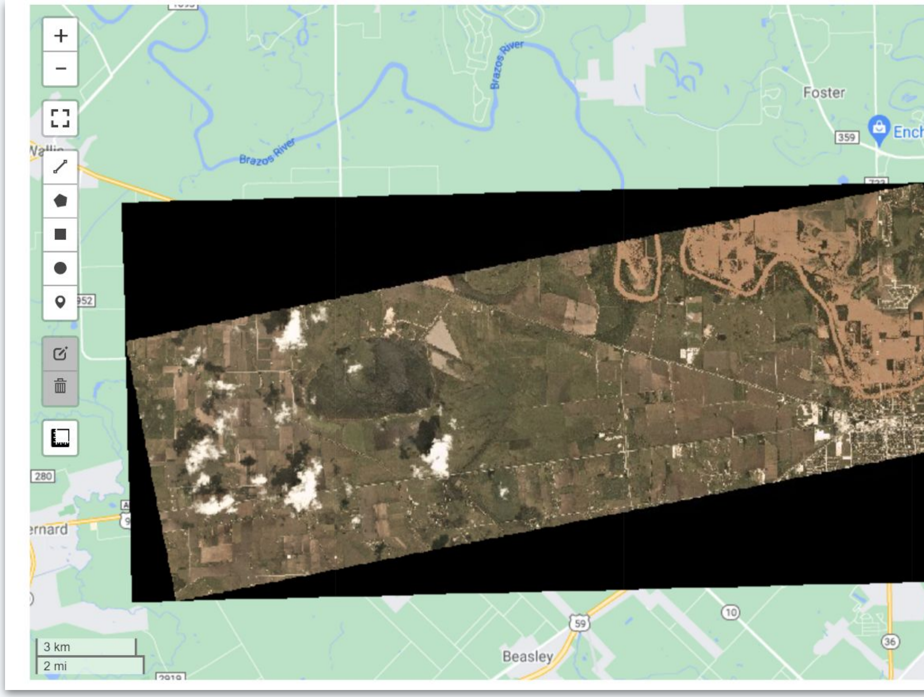
forrest.nyc *mbforr*

Larger volumes of vector (or vector ready) data





Greater reliance on raster data



CARTO

🇨🇱 Data Council: Austin

2000s

GDAL

2010s

Apache Spark

Fiona

DECK.GL

PyPROJ

Shapely

Leaflet

GeoPandas

kepler.gl

Rasterio

2020s

Apache Sedona

Leafmap

CARTO

🇨🇱 Data Council: Austin

Databases

Apache Sedona

Apache Spark

Desktop

Web

kepler.gl

Python

GeoPandas

Leafmap

Developer

DECK.GL

Leaflet

CARTO

Tools

GDAL

PYPROJ

Shapely

Fiona

Rasterio

Data +

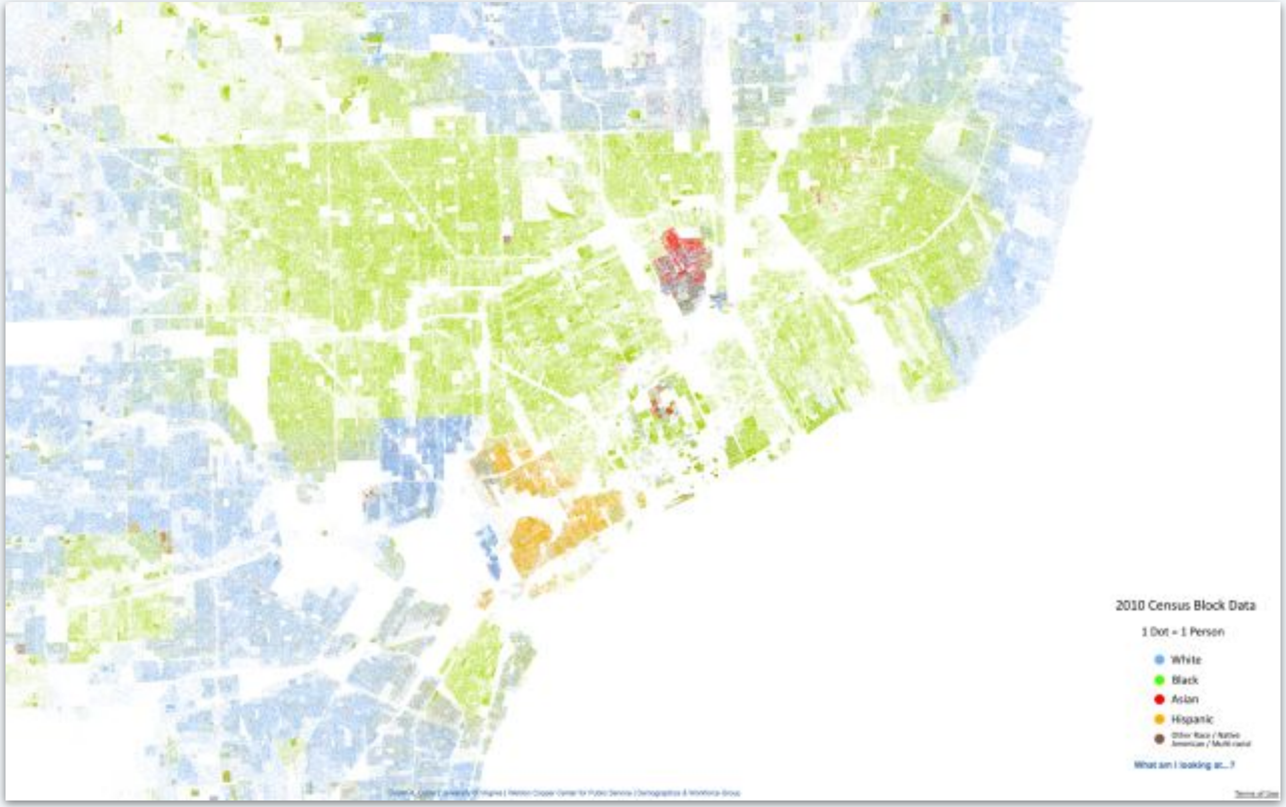
GIS and spatial thinking

CARTO

"Everything is related to everything else, but near things are more related than distant things."

“The phenomenon external to a geographic area of interest affects what goes on inside.”

🇨🇱 Data Council: Austin



CARTO

NATIONAL GEOGRAPHIC | Education

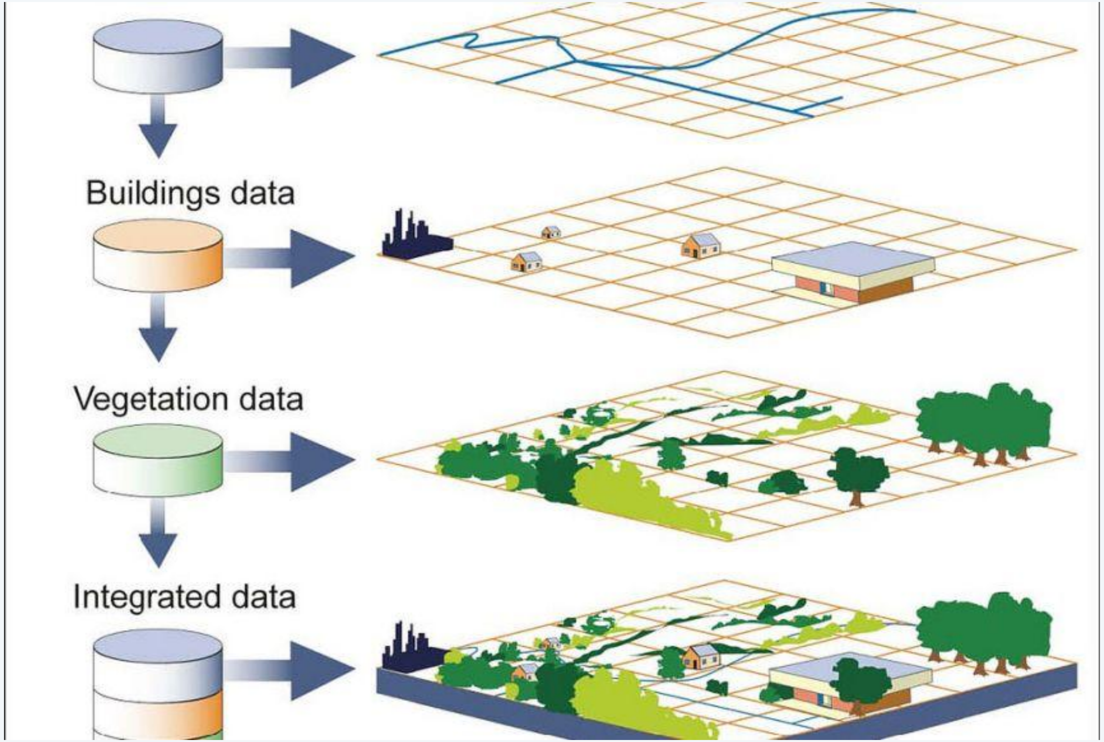
RESOURCE | ARTICLE

GIS (Geographic Information System)

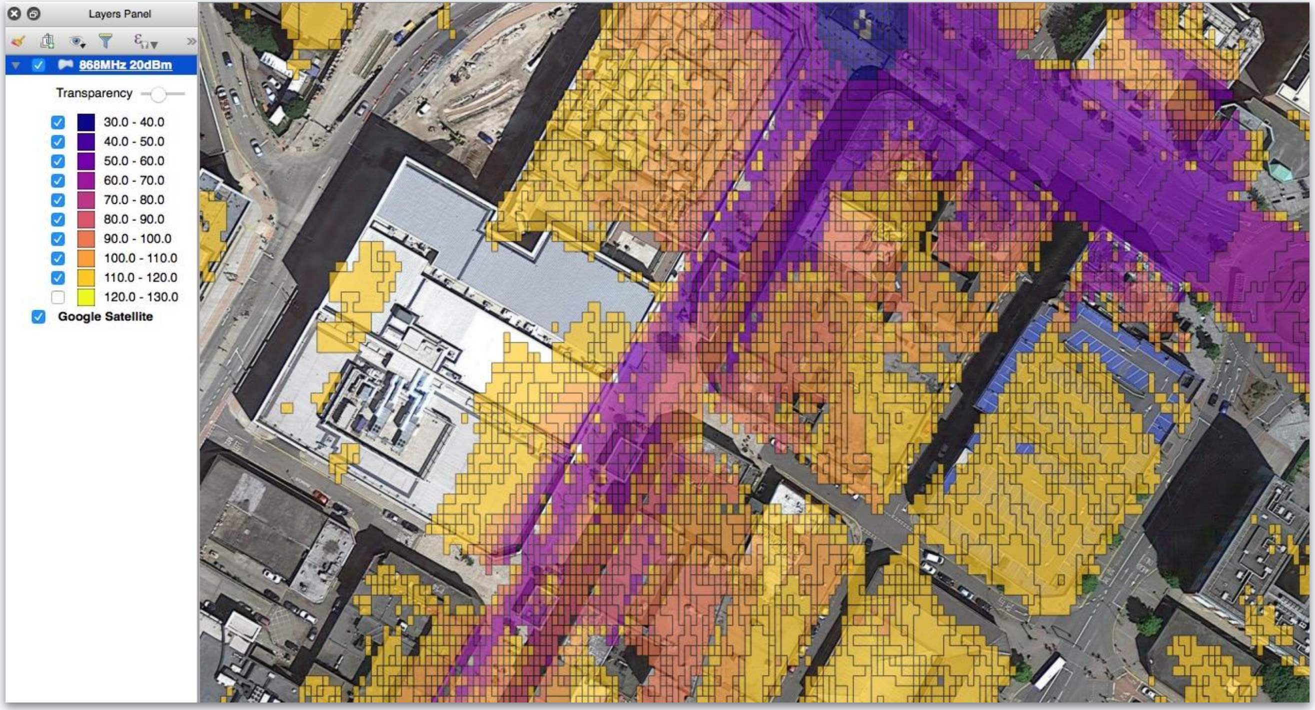
A geographic information system (GIS) is a computer system for capturing, storing, checking, and displaying data related to positions on Earth's surface.

GRADES
4 - 12+

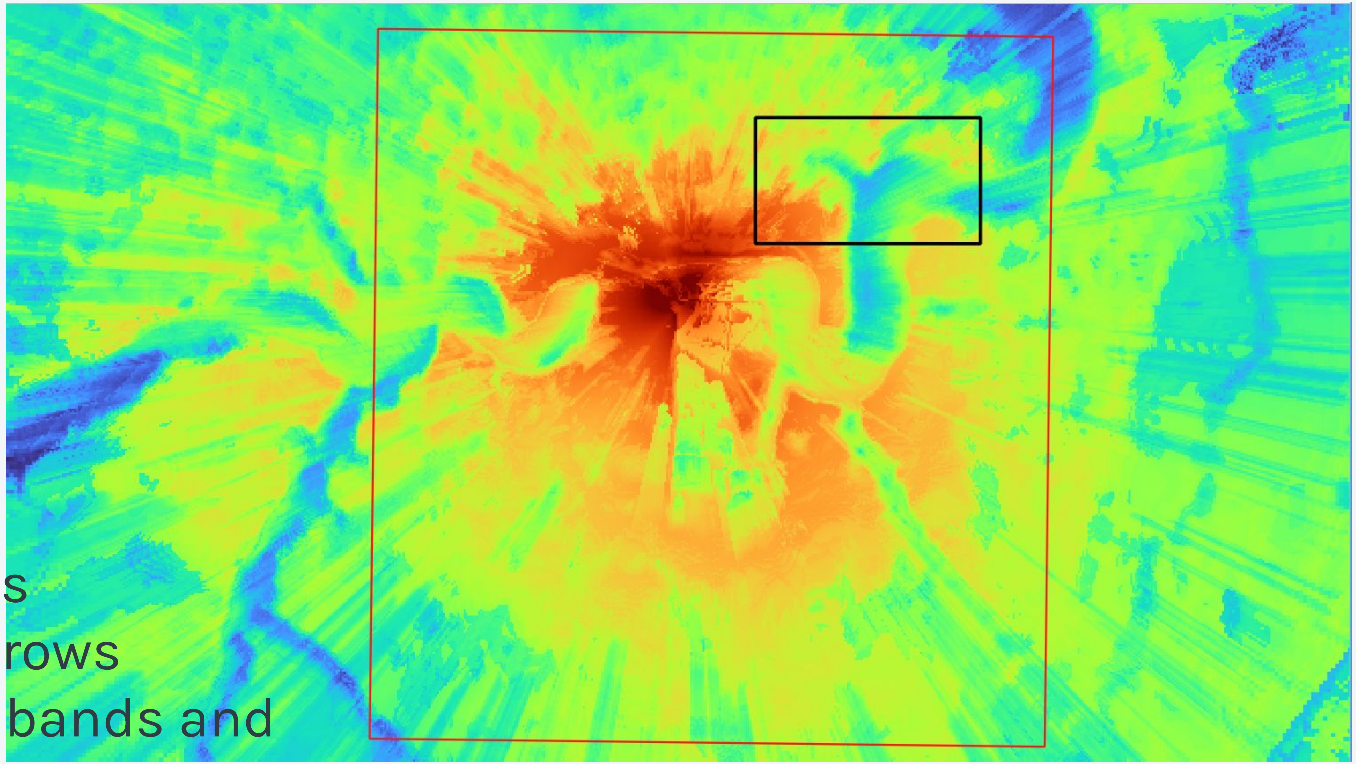
SUBJECTS
Geography, Geographic Information Systems (GIS), Physical Geography



CARTO


Layers Panel
868MHz 20dBm
Transparency
30.0 - 40.0
40.0 - 50.0
50.0 - 60.0
60.0 - 70.0
70.0 - 80.0
80.0 - 90.0
90.0 - 100.0
100.0 - 110.0
110.0 - 120.0
120.0 - 130.0
Google Satellite

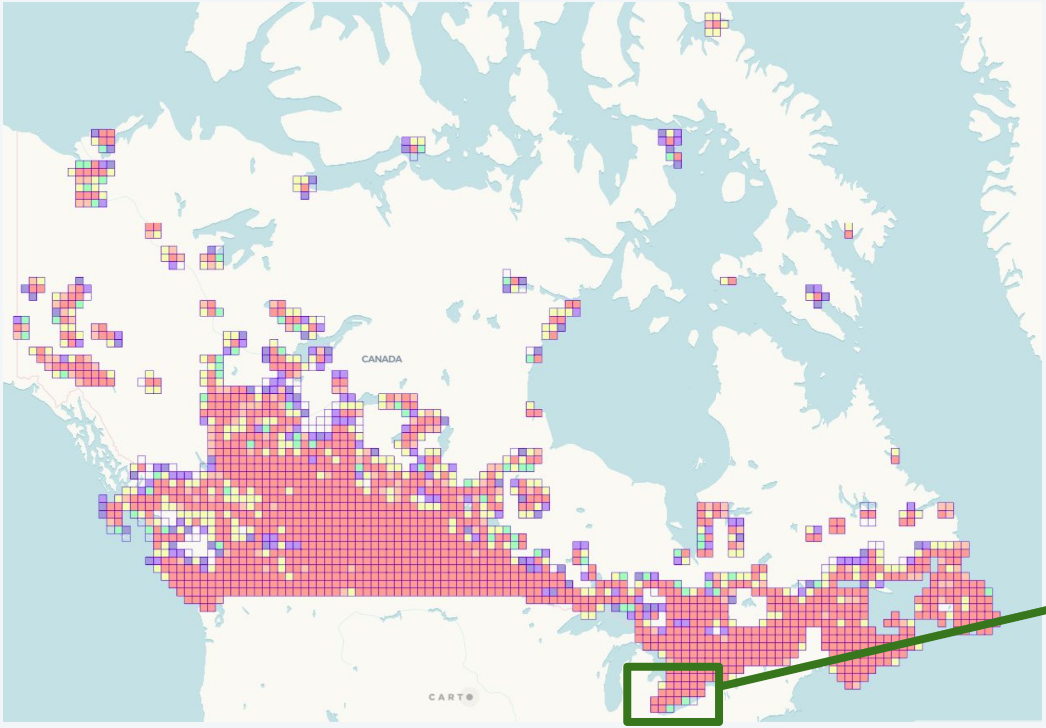
🇨🇱 Data Council: Austin

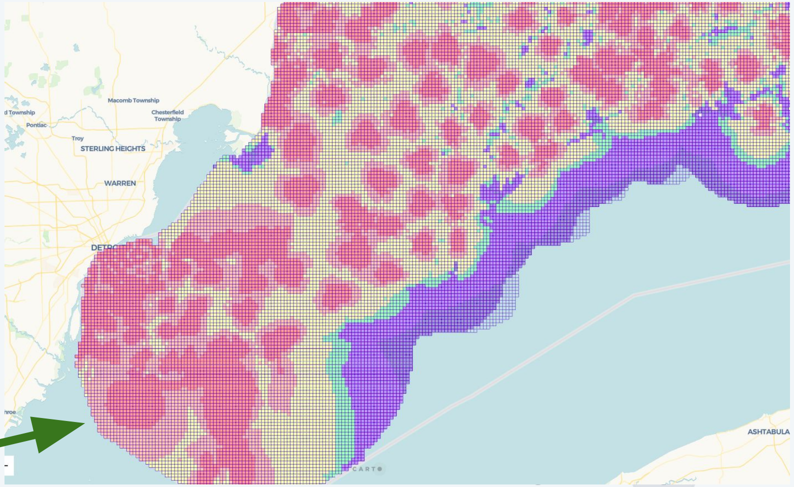


30m cells
9 billion rows
Multiple bands and signal strengths

CARTO

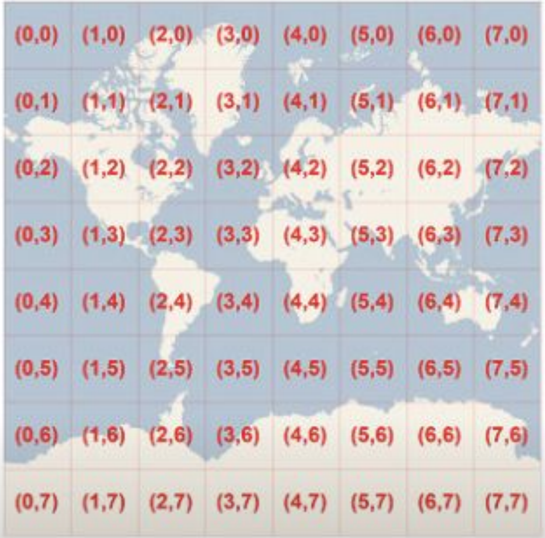
🇨🇱 Data Council: Austin

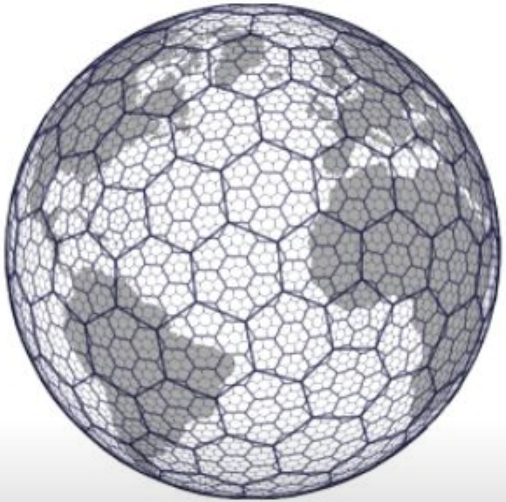


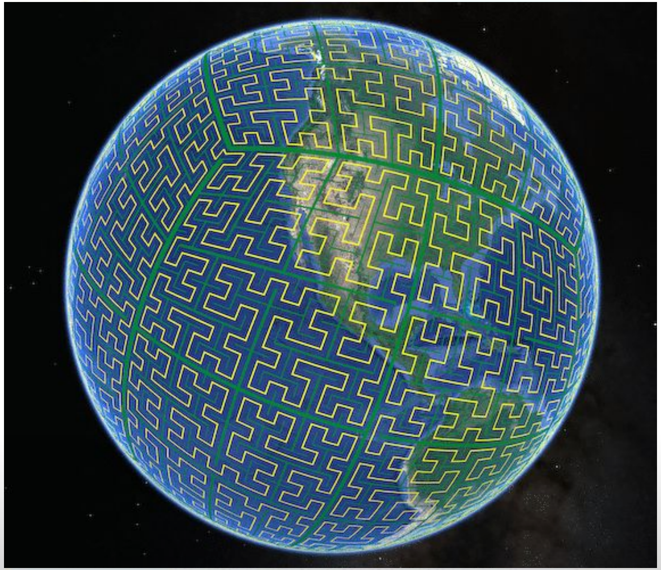


255GB to 28 GB
26:05 to 0:03 spatial join

Welcome spatial indexes







Quadkey (source)

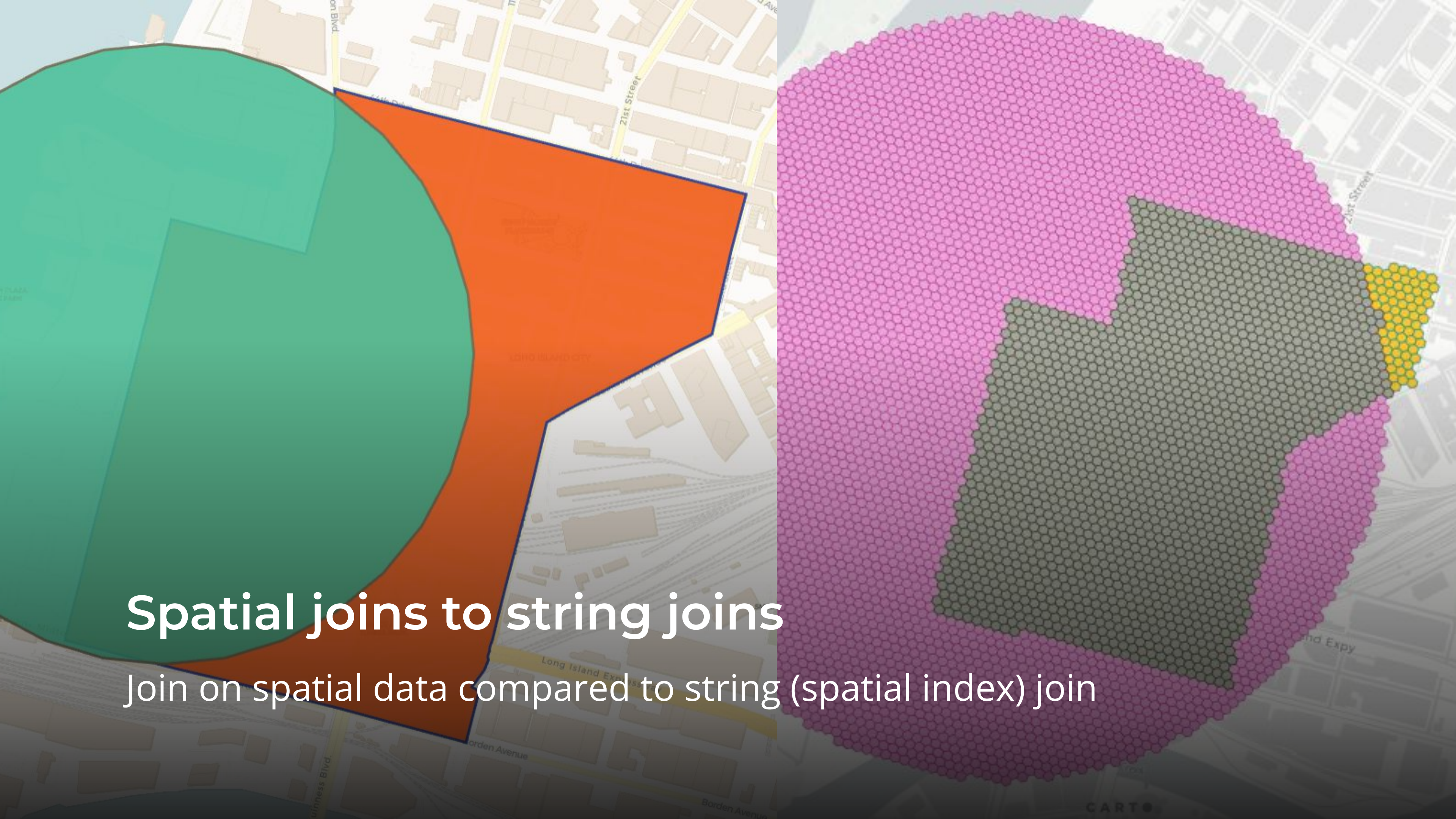
Uber's H3 (source)

S2 (source)

Geospatial Hierarchical Indexes

Different strategies to partition the space intro discrete grids

1. Query performance
2. Storage
3. Visualization performance
4. Intuitive visualization
5. Neighbors and children



Spatial joins to string joins

Join on spatial data compared to string (spatial index) join

Geometries vs Spatial Indexes:

What do they look like?

POLYGON((-96.196141 41.125515, -96.195606 41.125514, -96.181864 41.125507, -96.177078 41.125474, -96.167733 41.125456, -96.160565 41.125456, -96.154682 41.125429, -96.151094 41.125414, -96.138848 41.125395, -96.138454 41.125394, -96.138381 41.125394, -96.137158 41.125391, -96.130043 41.125377, -96.1301...))*

*This represents about a **10th** of the geometric description of a census tract, which we have truncated for readability

8a2aa84ec307fff

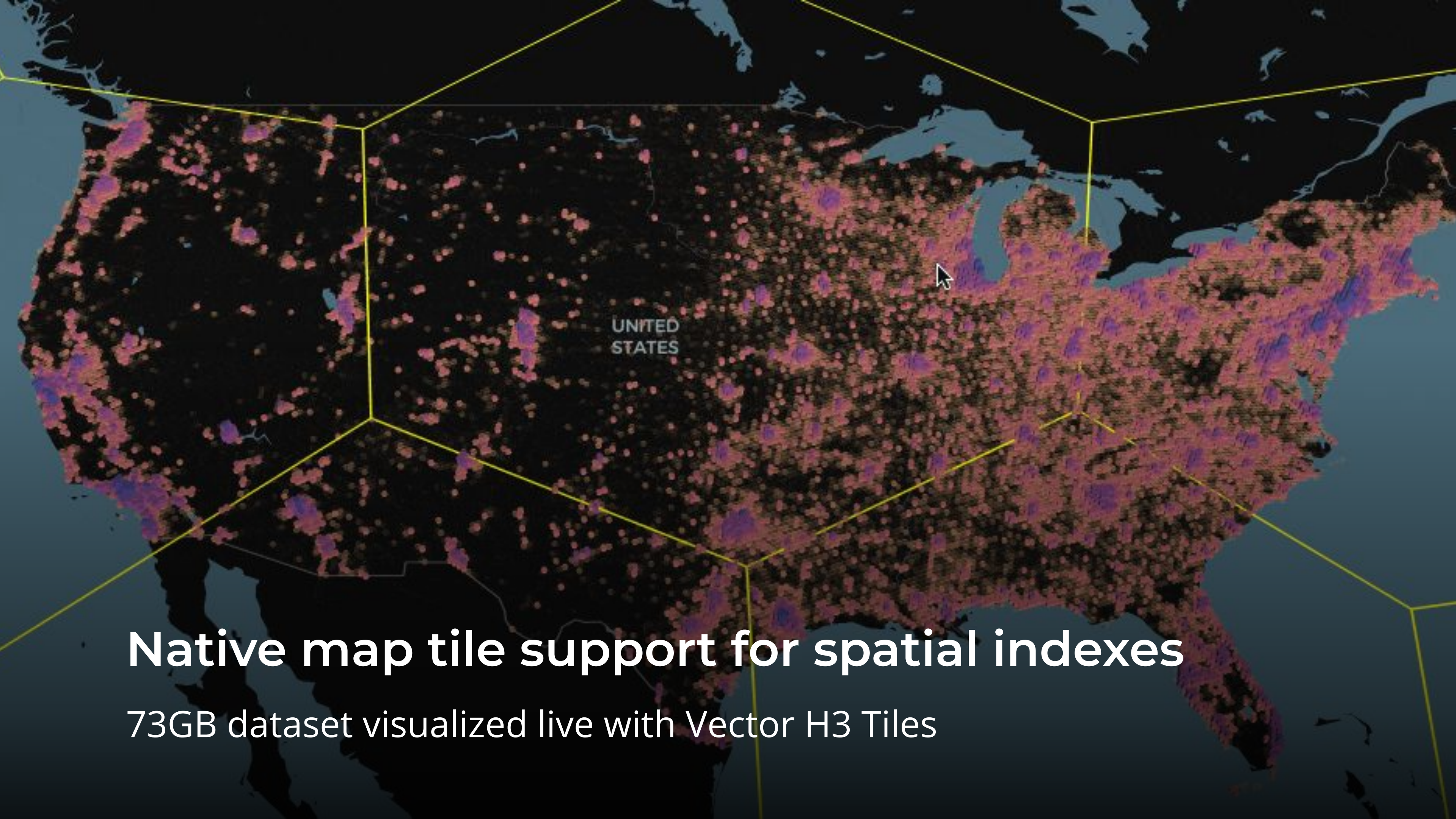
🇨🇱 Data Council: Austin

Performance comparisons

Example ETL use case	Geometries	Spatial Indexes	Gain of spatial index over geometries
Processing time	12 days	7 hours	**98% time saved**
Data transferred into the Database	4 TB	1.5 TB	**62% less data transferred**
RAM to process the largest file	256 GB	28 GB	**89% reduction in RAM**
Time to process a spatial join with population	26 minutes	3 seconds	**99% less time**
Time to generate a tileset	23 minutes	1.5 minutes	**94% time saved**
Population coverage	15.48%	15.48%	**0% coverage lost**

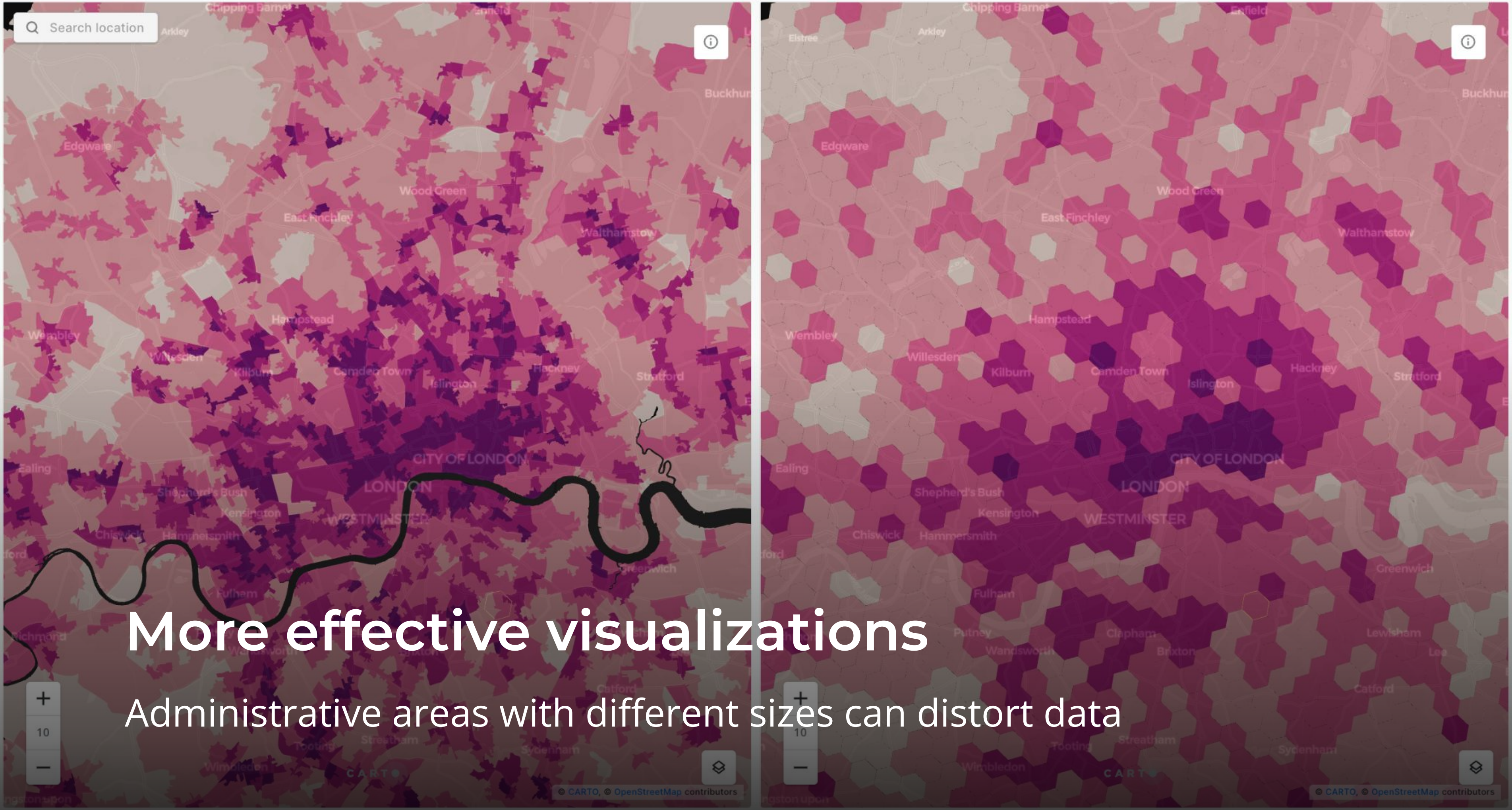
Estimated reduction in cloud data warehouse bill by 85%

CARTO



Native map tile support for spatial indexes

73GB dataset visualized live with Vector H3 Tiles



More effective visualizations

Administrative areas with different sizes can distort data

Analyze parents, children, and neighbors

Easy to scale up and down and analyze relationships

1. Loss of raw data
2. Precise spatial coverage
3. Original data quality and precision
4. Boundary effects

🇨🇱 Data Council: Austin

CARTO

Workshop

1. Data exploration
2. Performance face-off
3. Visualizing indexed data
4. Enriching grids
5. Spatial analysis
6. Spatial Autocorrelation

Workshop docs here

Q&A

Thank you!

Matt Forrest
matt@carto.com
in mbforr