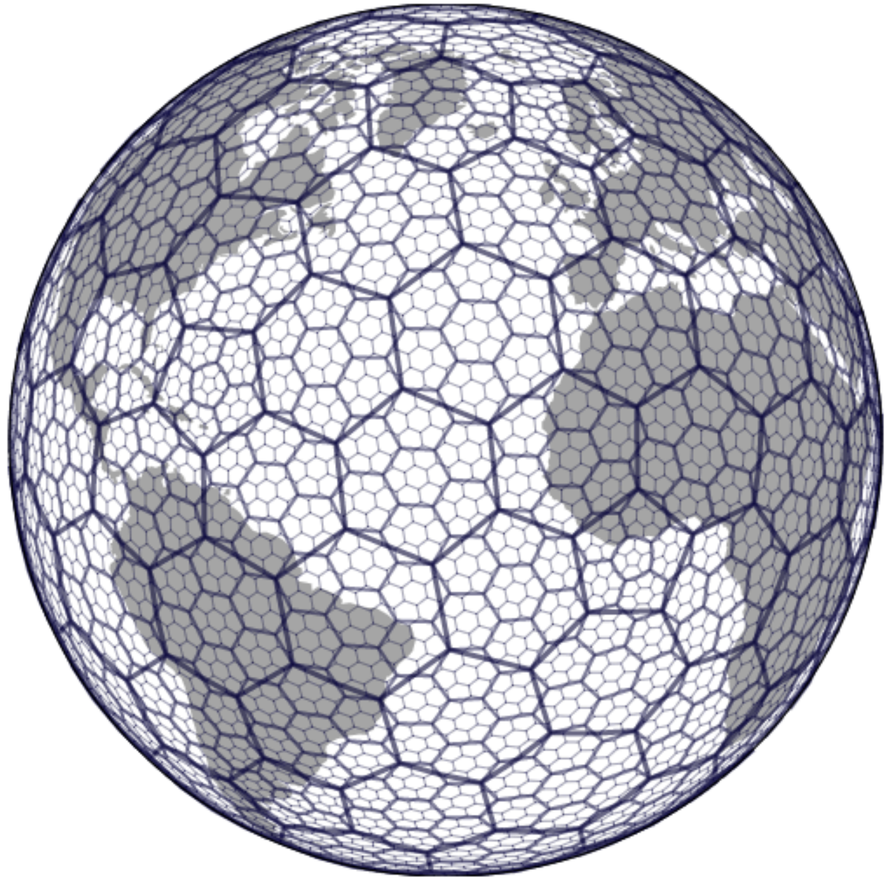
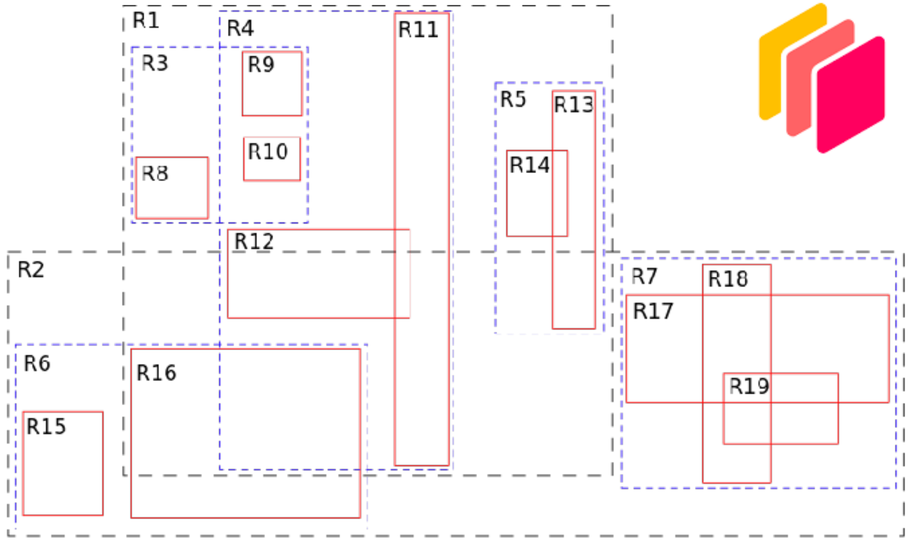
Lecture 13: Big Spatial Data

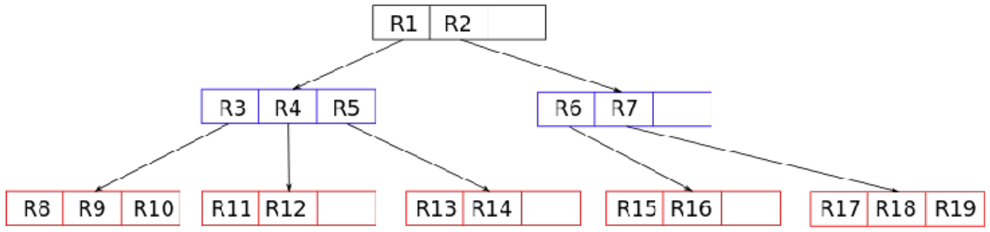
Instructor: Ane Rahbek Vierø
May 1, 2023



Today we will learn about….

Spatial indexes





Tools for large geospatial data

dask +

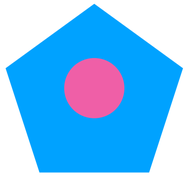
Course Evaluation Survey (Spring 2023)

Spatial is slow

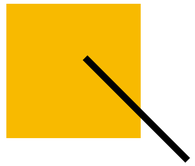
Avoid spatial queries and operations if you can

Spatial queries

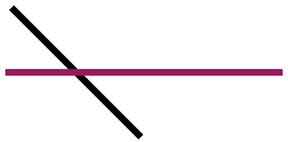
```
point.within(poly)
```



```
line.intersects(poly)
```

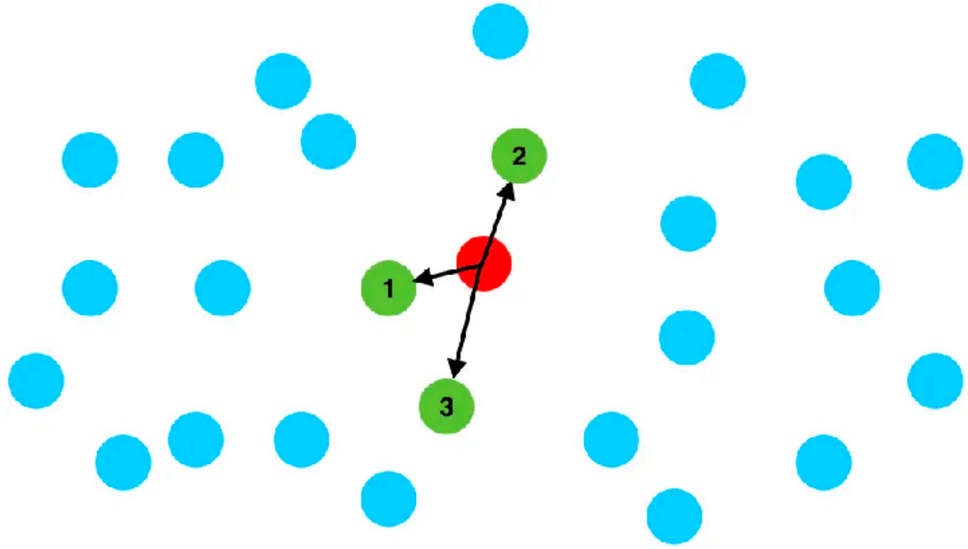


```
line1.crosses(line2)
```



```
poly1.touches(poly2)
```

etc….

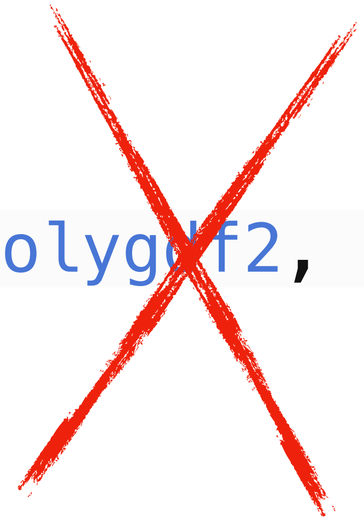


Spatial is slow

```
polygdf.overlay(polygdf2, how="intersection")
```

```
polygdf.loc[poly_gdf["area_name"]=='my_area']
```

Spatial is slow



```
polygdf.overlay(polygdf2, how="intersection")
```

```
polygdf.loc[poly_gdf["area_name"]=='my_area']
```

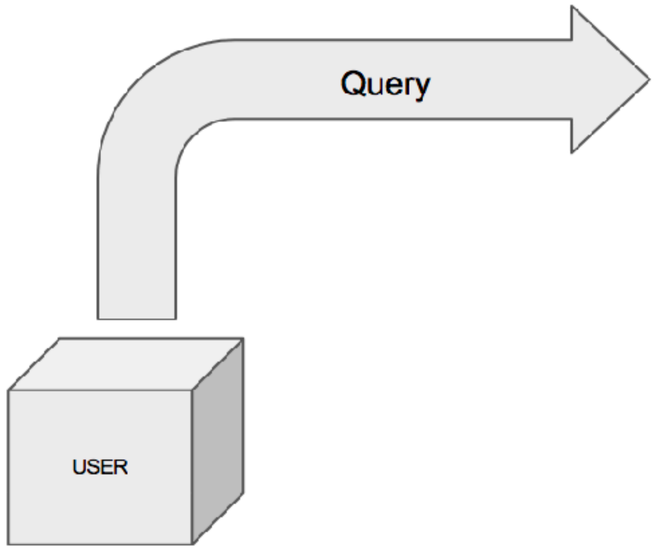
Use non-spatial queries whenever possible

Spatial Indexes

Spatial Indexes

…for efficient spatial access methods

Spatial Indexes



Table

company_id	units	unit_cost
10	12	1.15
12	12	1.05
14	18	1.31
18	18	1.34
11	24	1.15
16	12	1.31
10	12	1.15
12	24	1.3
18	6	1.34
18	12	1.35
14	12	1.95
21	18	1.36
12	12	1.05
20	6	1.31
18	18	1.34
11	24	1.15
14	24	1.05

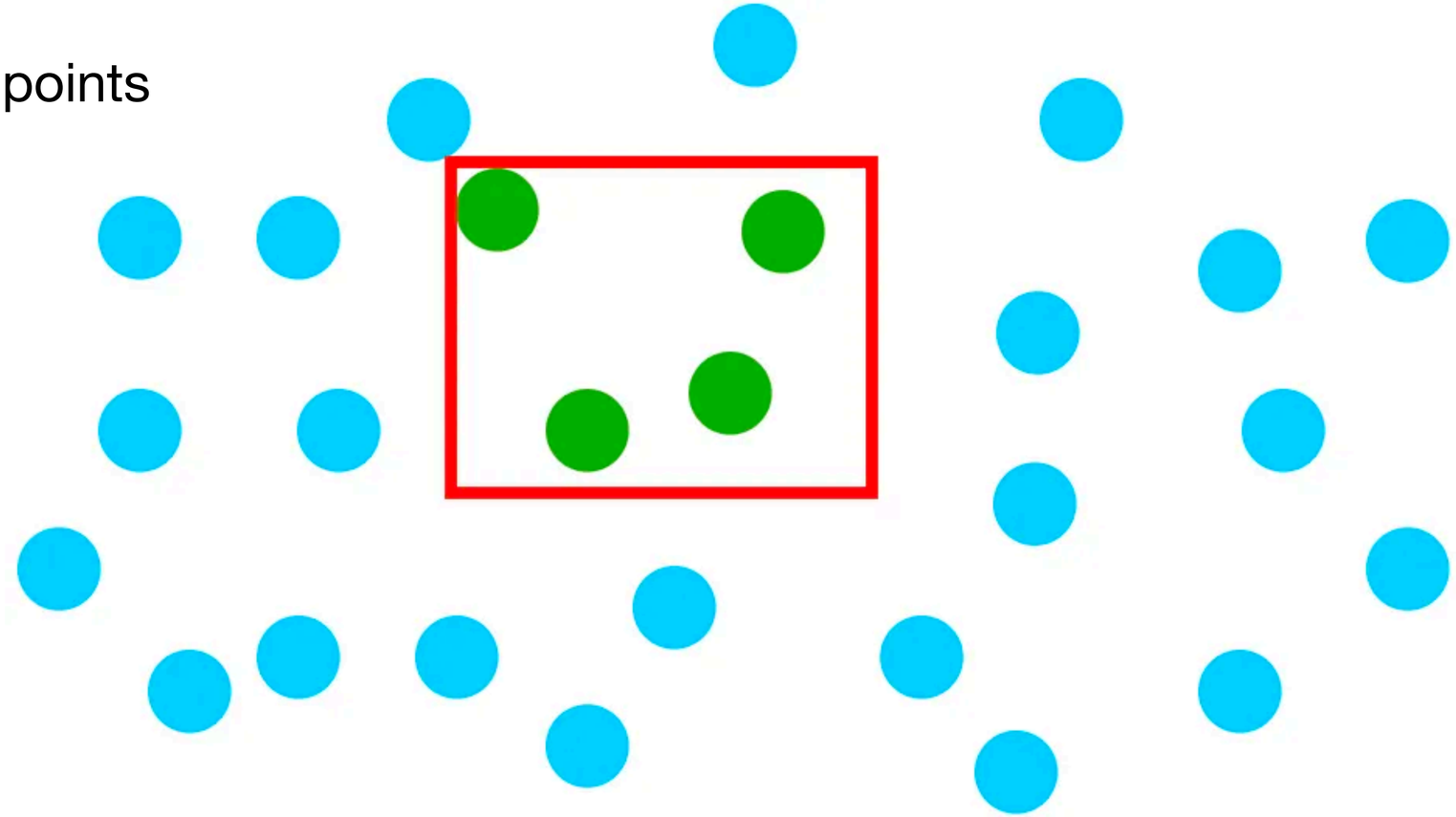
We use indices to speed up data retrieval

Avoid searching every entry during look-up (in worst-case scenario)

Spatial Indexes

Which points are inside this polygon?

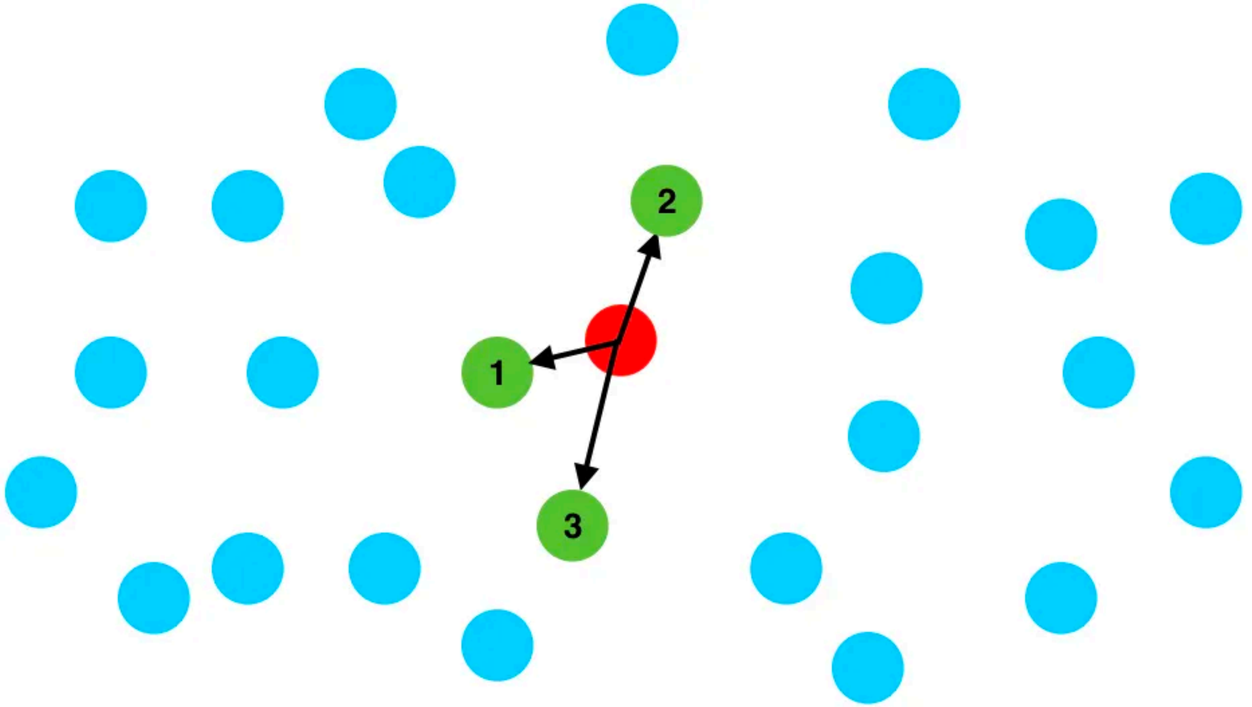
We don’t want to check *all* points



https://blog.mapbox.com/a-dive-into-spatial-search-algorithms-ebd0c5e39d2a

Spatial Indexes

Without a spatial index, K-nearest neighbors requires N distance computations

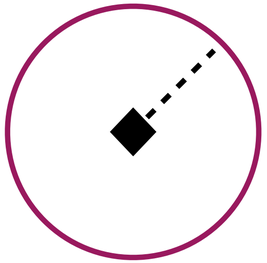


https://blog.mapbox.com/a-dive-into-spatial-search-algorithms-ebd0c5e39d2a

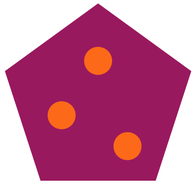
When to use spatial indexes?

If your query involves a:

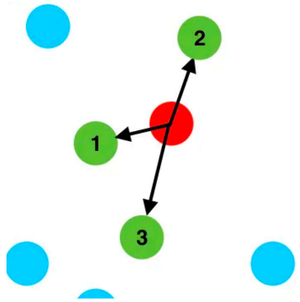
- **Range/radius/window query**



- **Spatial join**



- **K-Nearest Neighbor (KNN)**



…a spatial index is a good idea!

Spatial Indexes

Spatial indexes can be based on either space or data driven structures

Space-driven structures: partitioning the space into grid cells

Data-driven structures: partitioning the collection of spatial objects

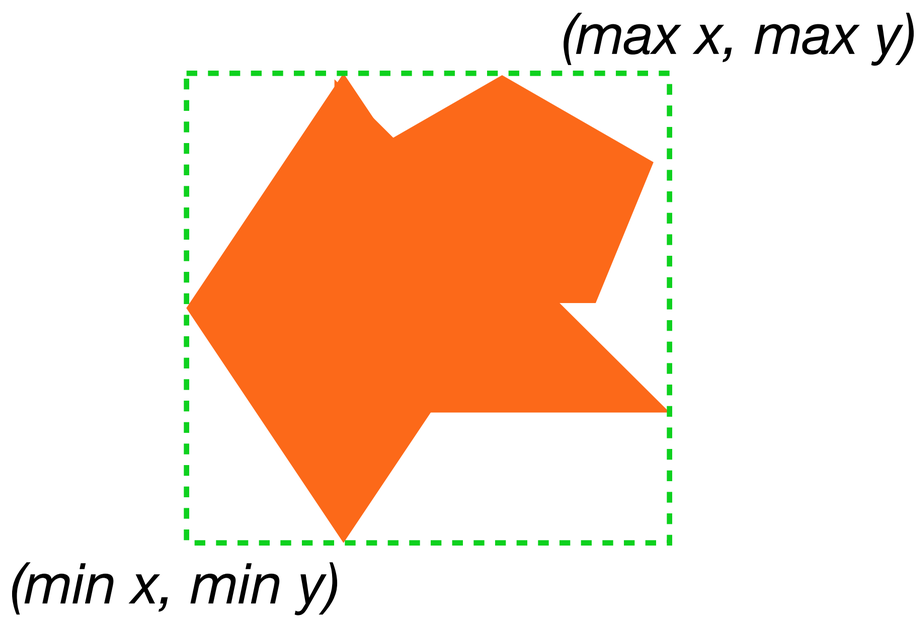
https://gistbok.ucgis.org/bok-topics/spatial-indexing

Minimum Bounding Box

(Or Minimum Bounding Rectangle)

The smallest box that can fit a spatial object

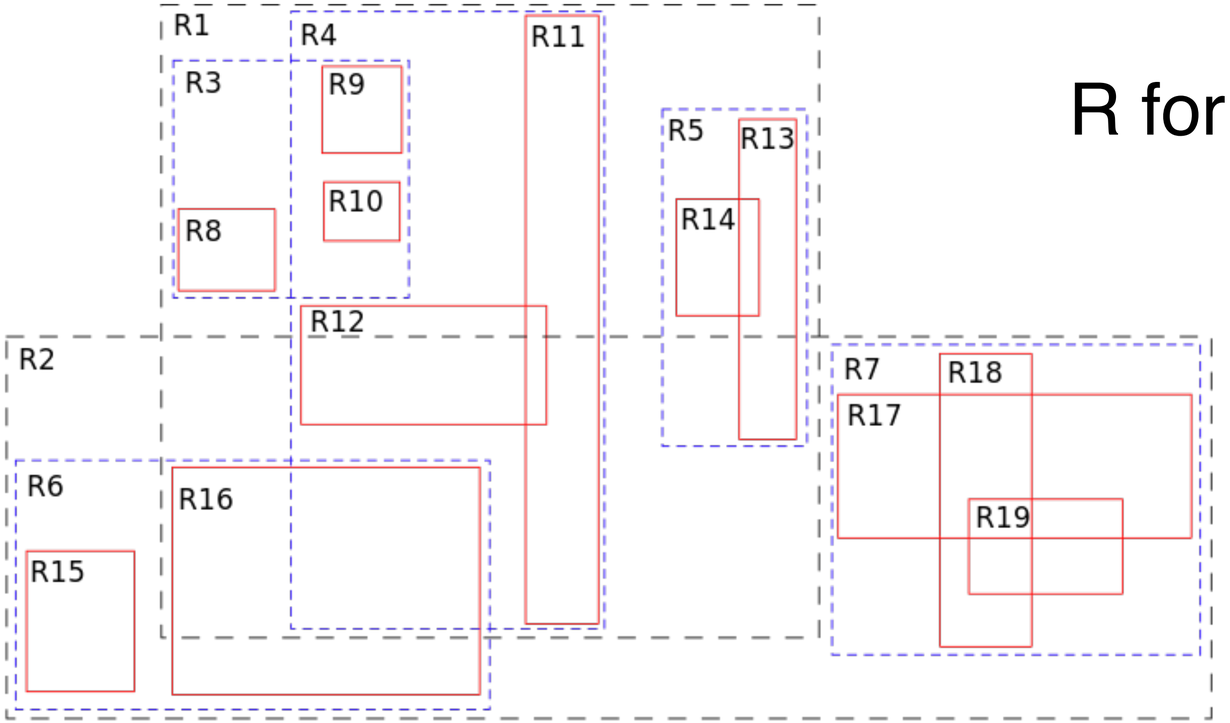
Defined by (min x, min y),(max x, max y)



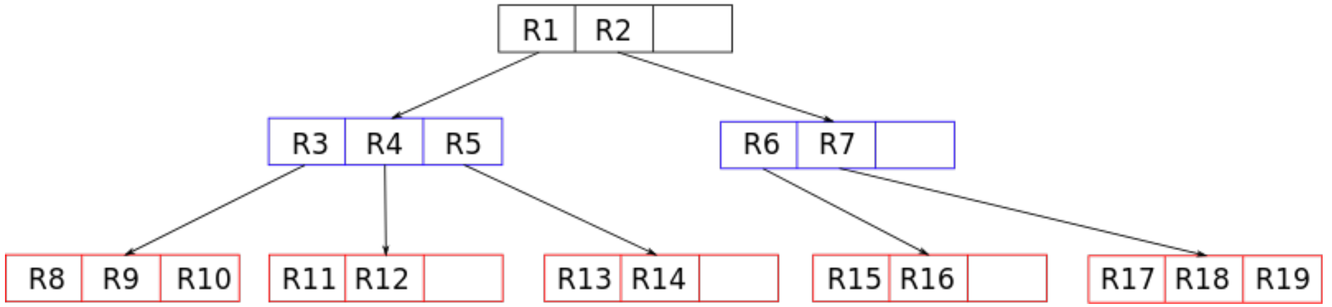
Data-driven Spatial Indexes

…uses spatial containment relationships based on hierarchies of minimum bounding boxes.

R-tree



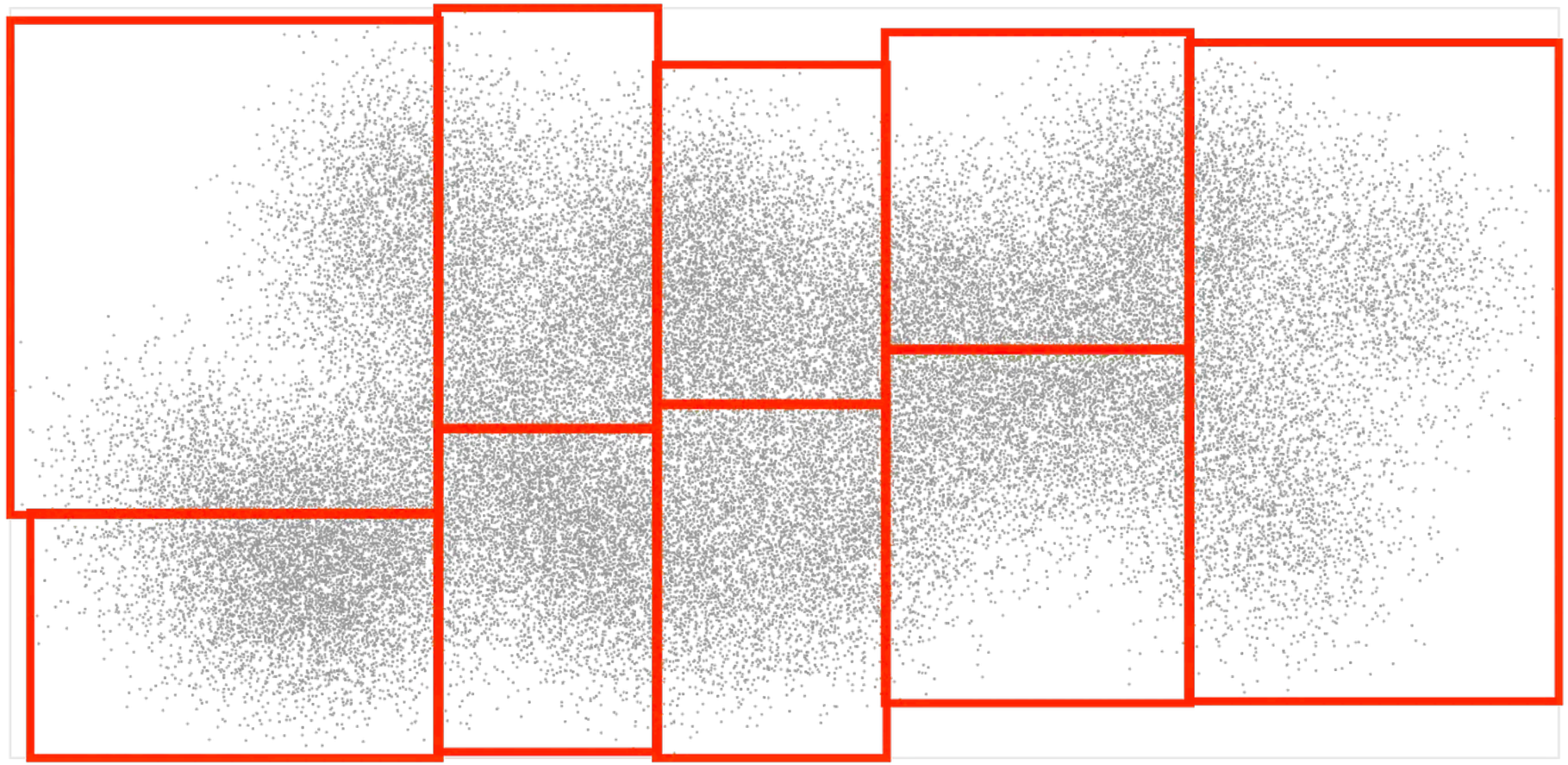
R for ‘Rectangle’



https://en.wikipedia.org/wiki/R-tree

R-tree

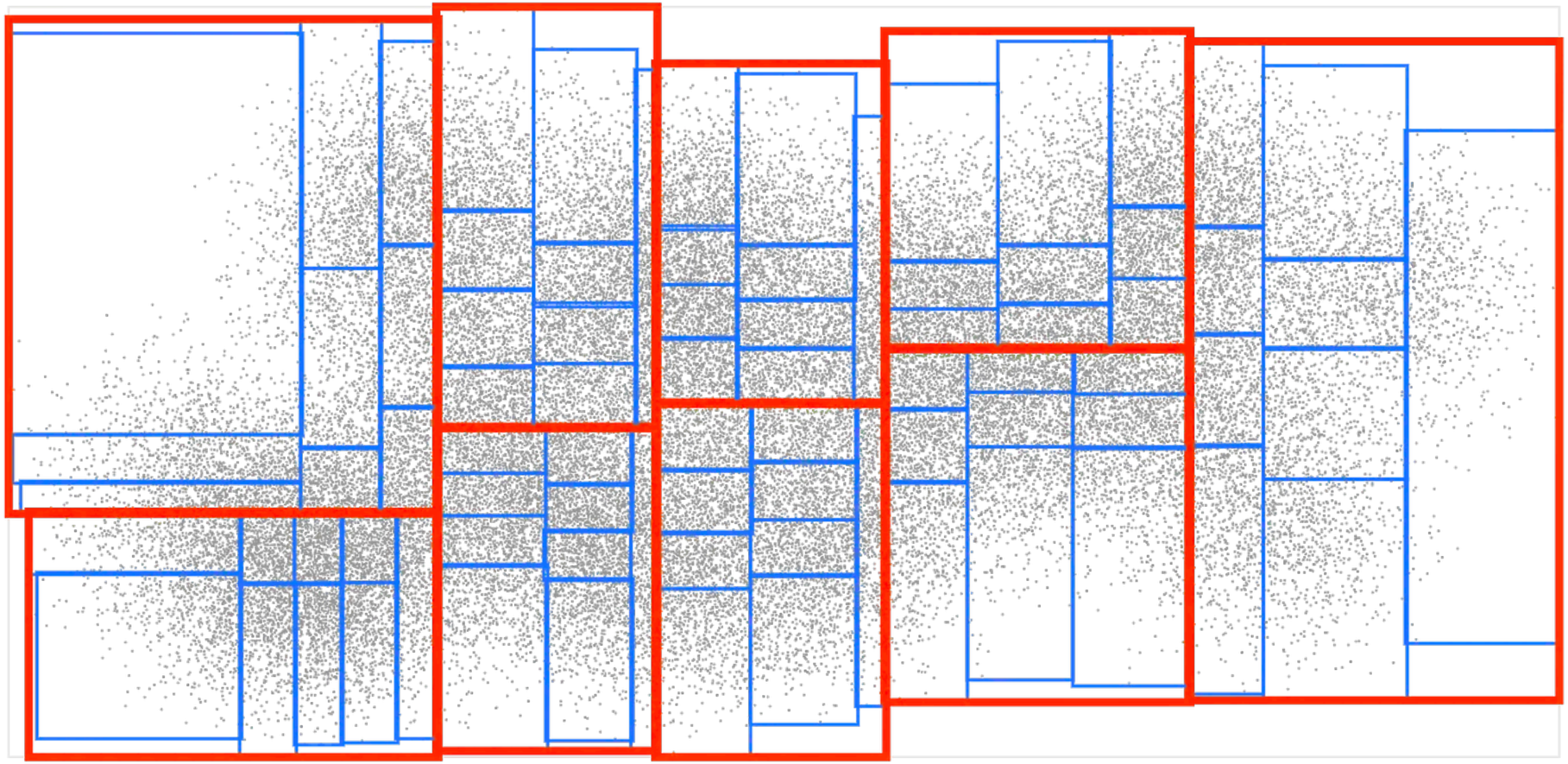
R for ‘Rectangle’



https://blog.mapbox.com/a-dive-into-spatial-search-algorithms-ebd0c5e39d2a

R-tree

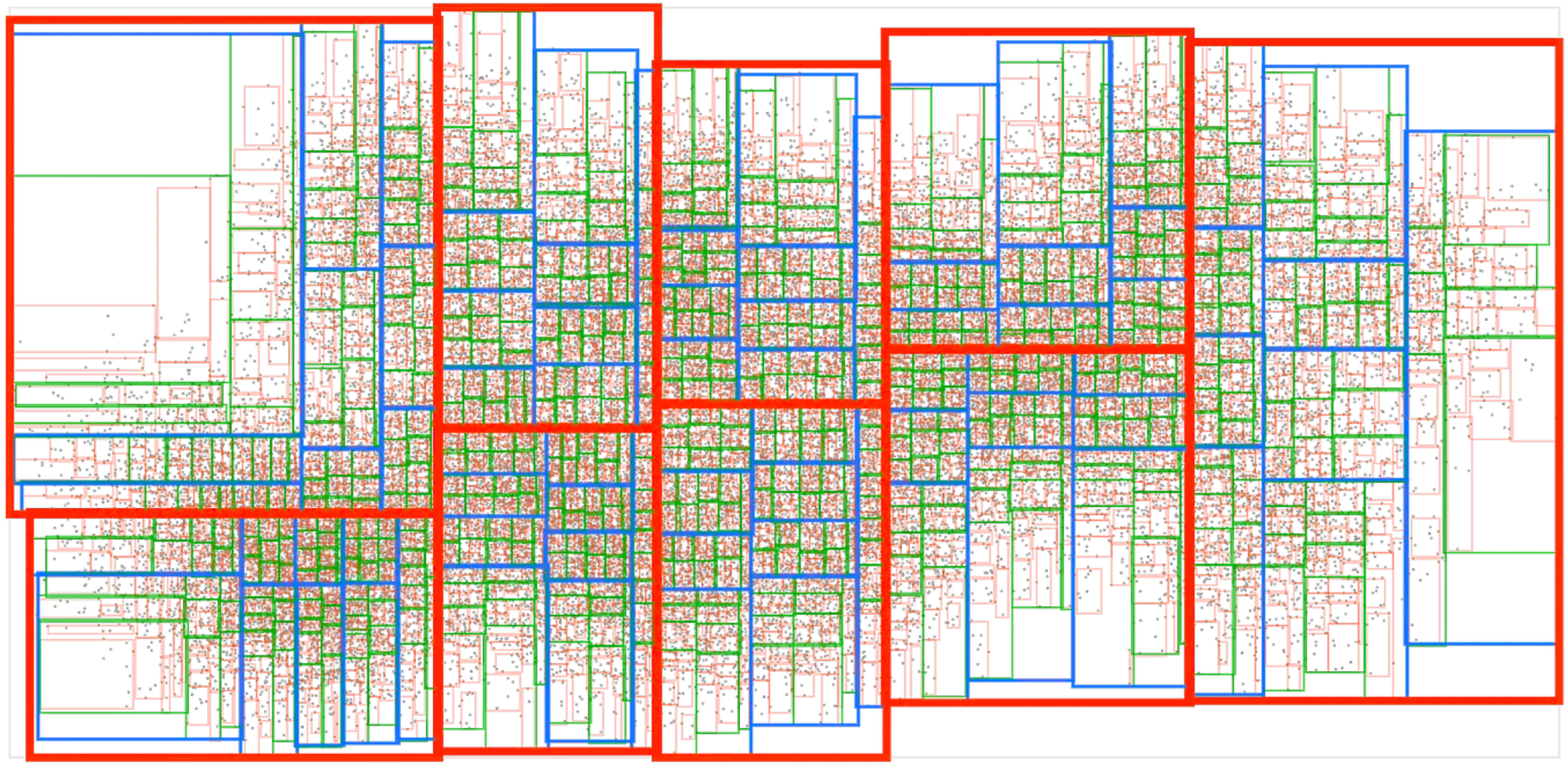
R for ‘Rectangle’



https://blog.mapbox.com/a-dive-into-spatial-search-algorithms-ebd0c5e39d2a

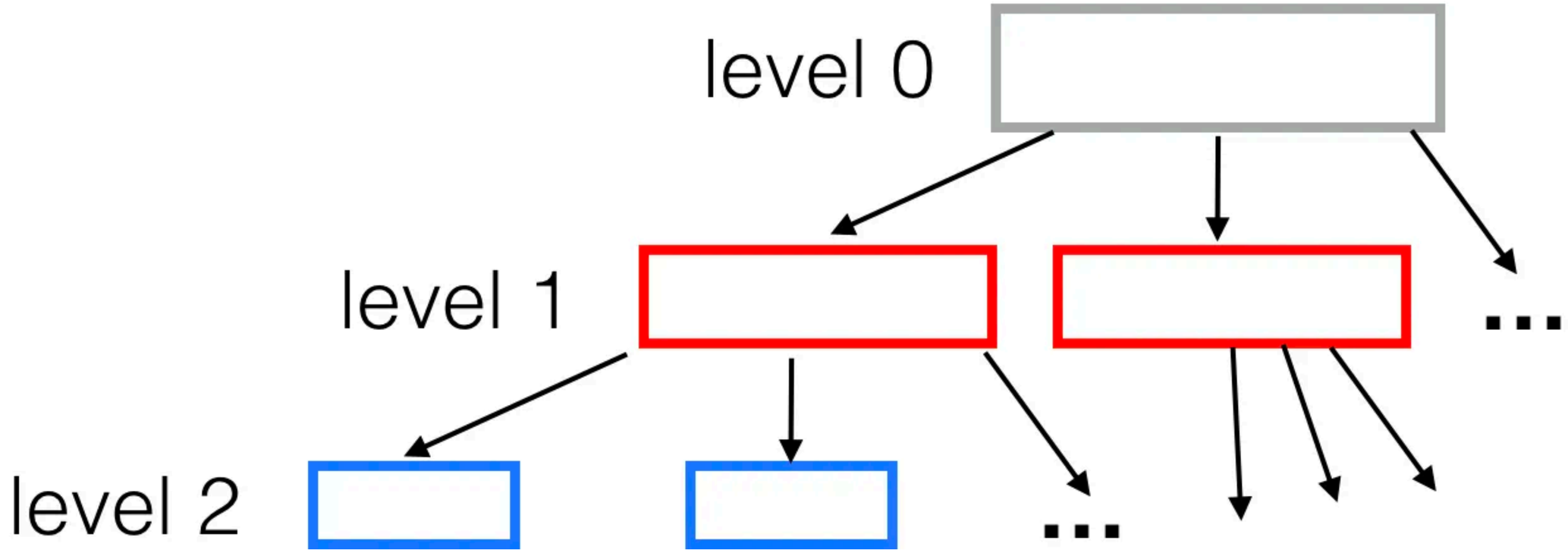
R-tree

R for ‘Rectangle’



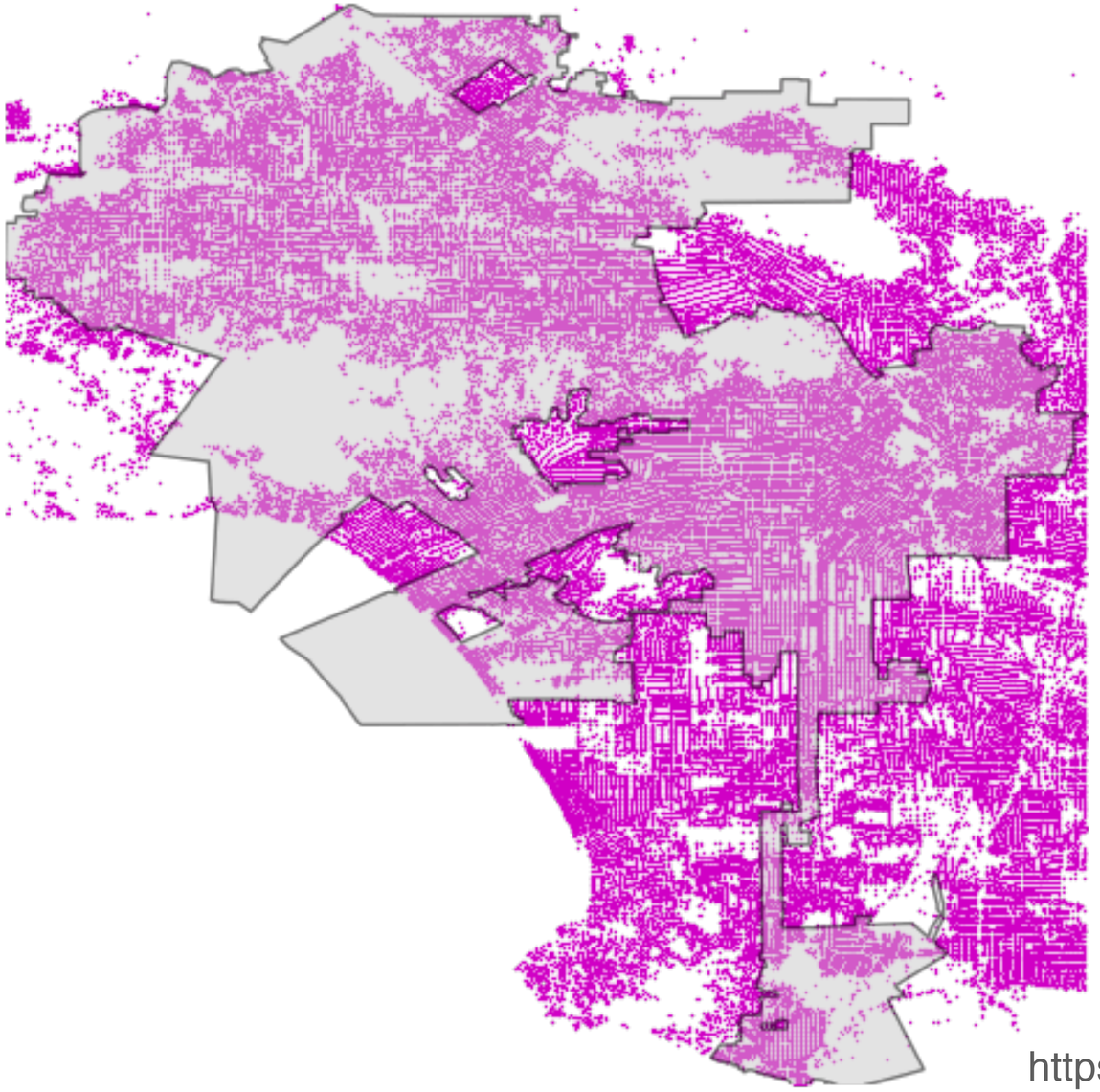
https://blog.mapbox.com/a-dive-into-spatial-search-algorithms-ebd0c5e39d2a

R-tree



R-tree

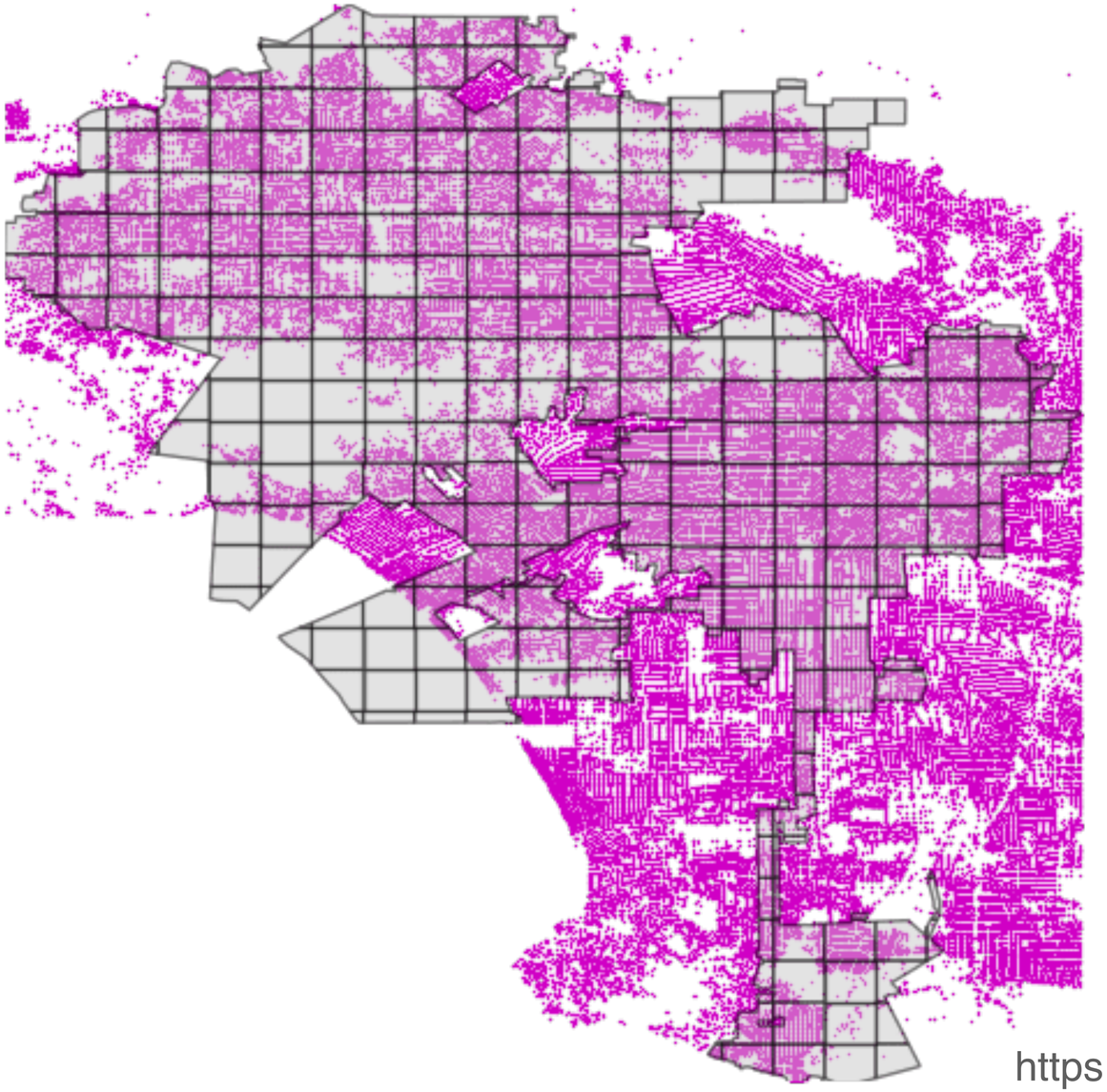
Closely overlapping bounding boxes removes benefit of R-tree



https://geoffboeing.com/2016/10/r-tree-spatial-index-python/

R-tree

Splitting data into subsets can be a solution

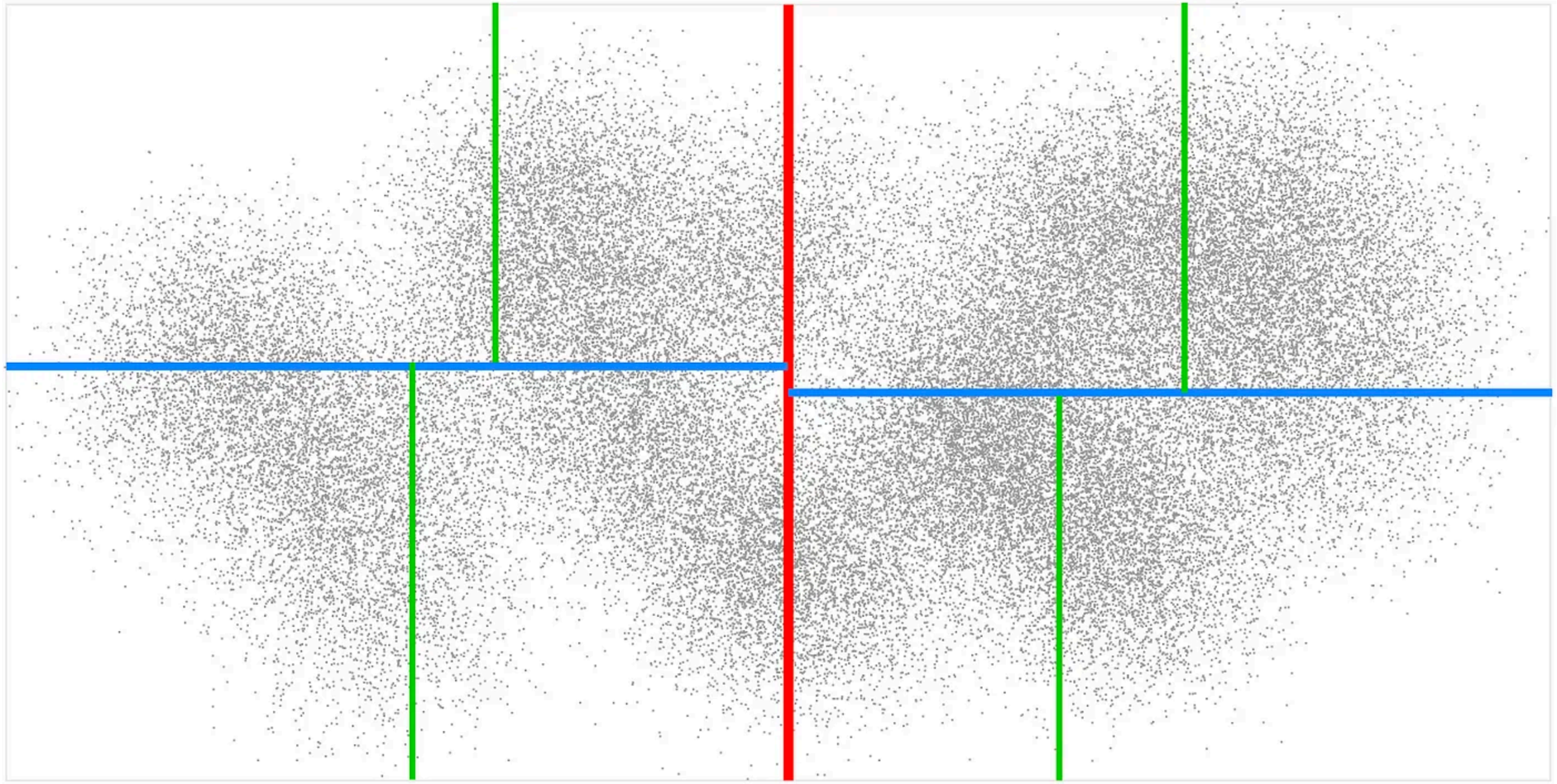


https://geoffboeing.com/2016/10/r-tree-spatial-index-python/

Space-driven Spatial Indexes

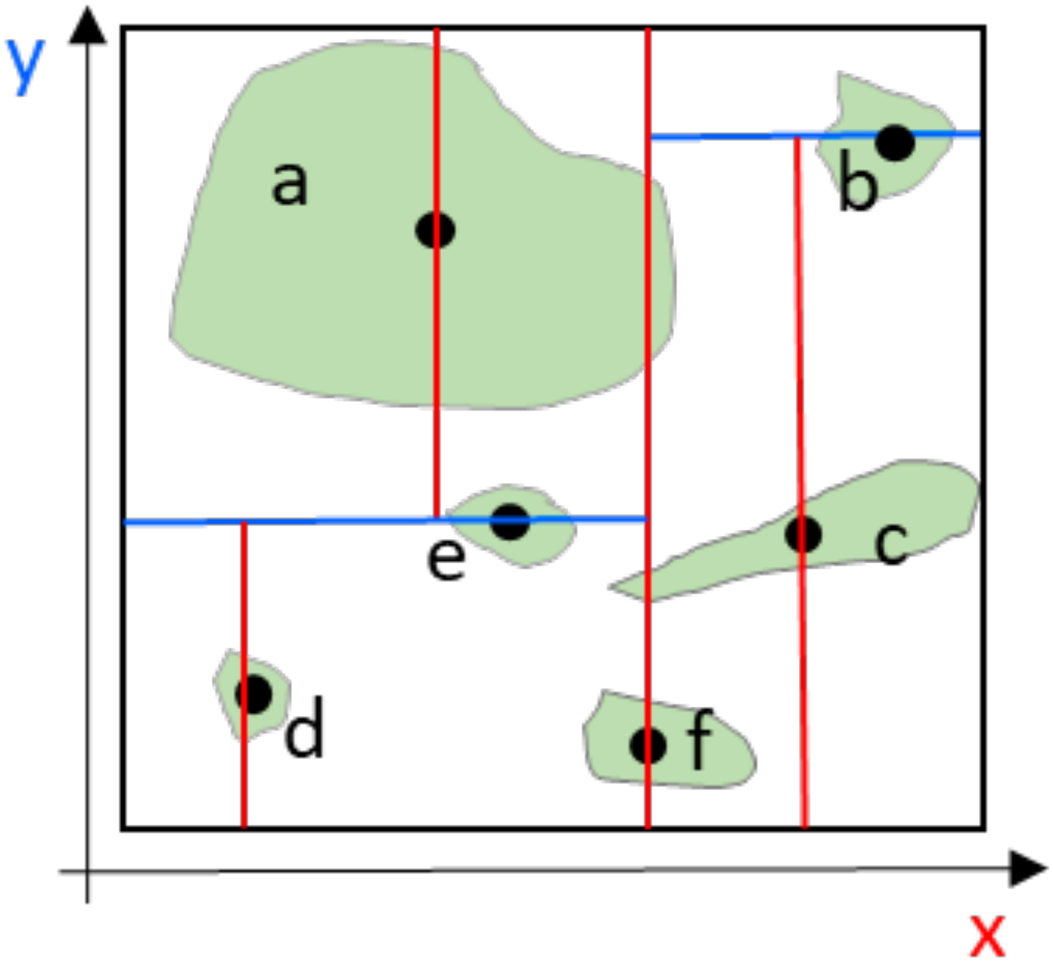
… using partitioning strategies to decompose a 2D plane into cells

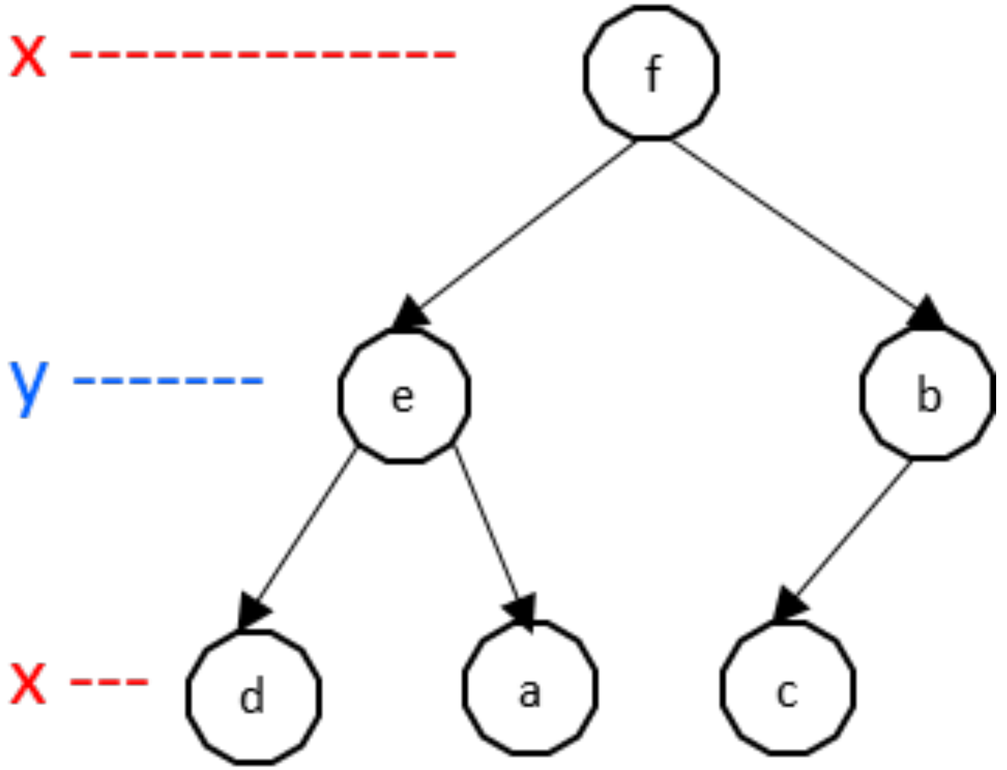
K-d Tree



https://blog.mapbox.com/a-dive-into-spatial-search-algorithms-ebd0c5e39d2a

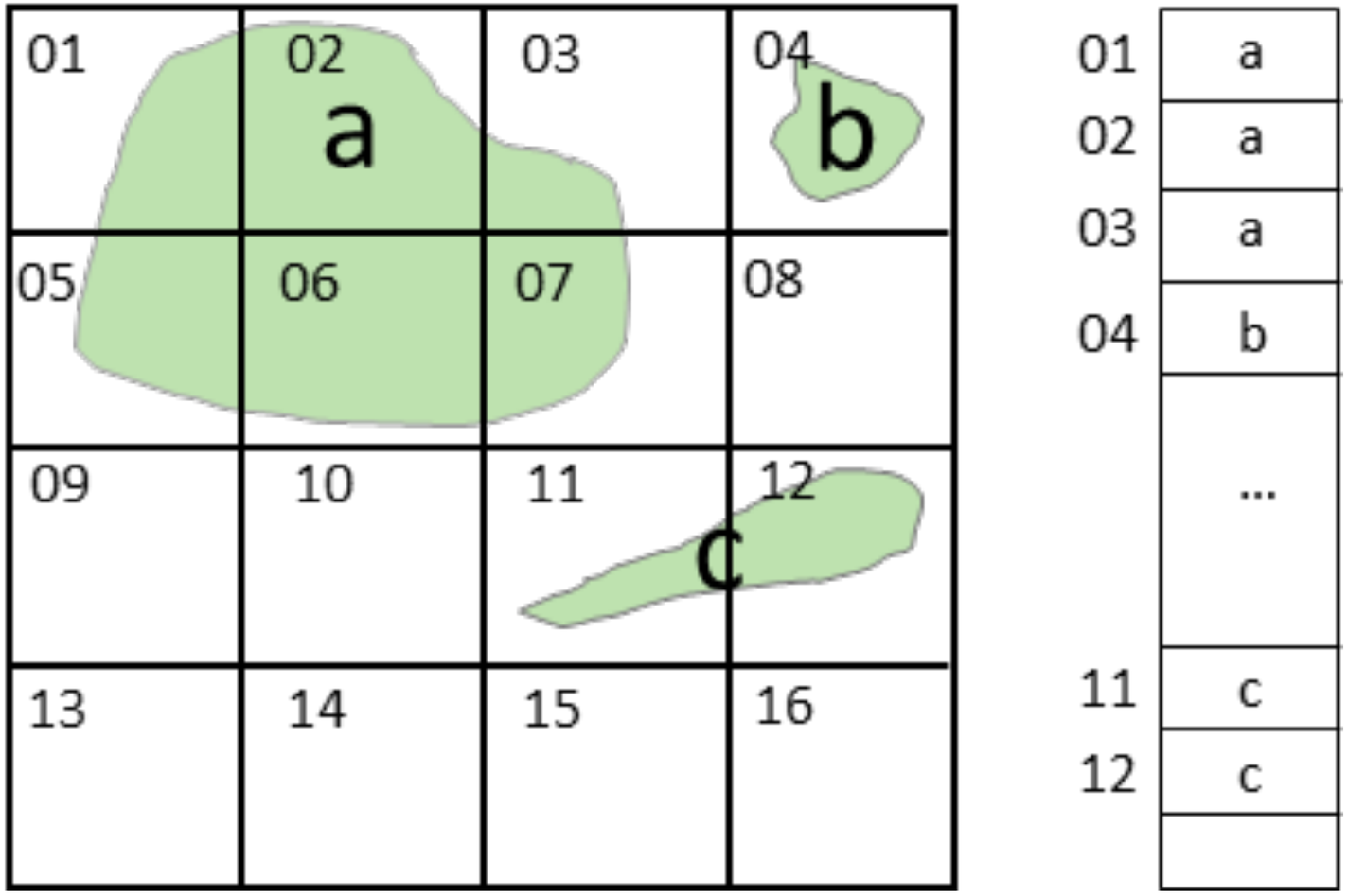
K-d Tree





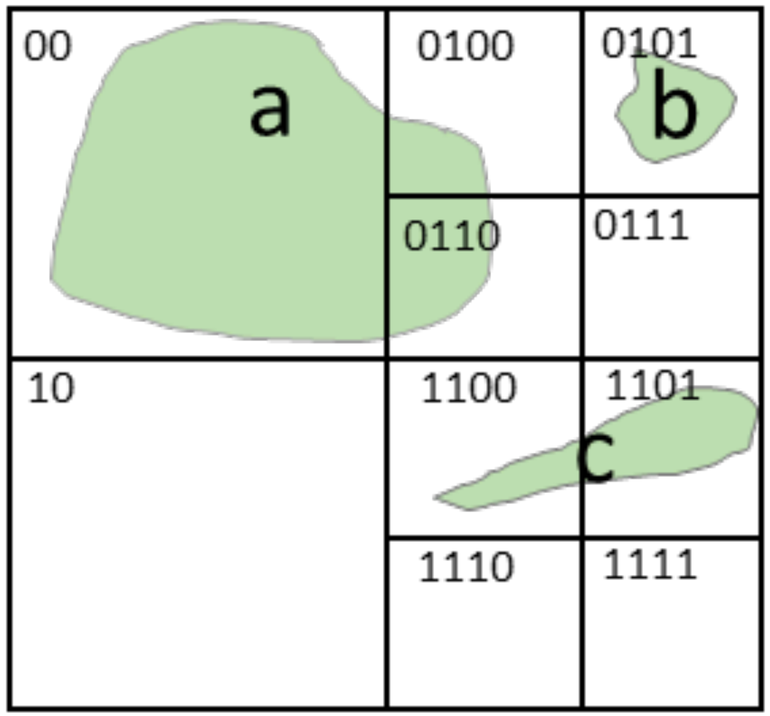
https://gistbok.ucgis.org/bok-topics/2018-quarter-02/spatial-indexing

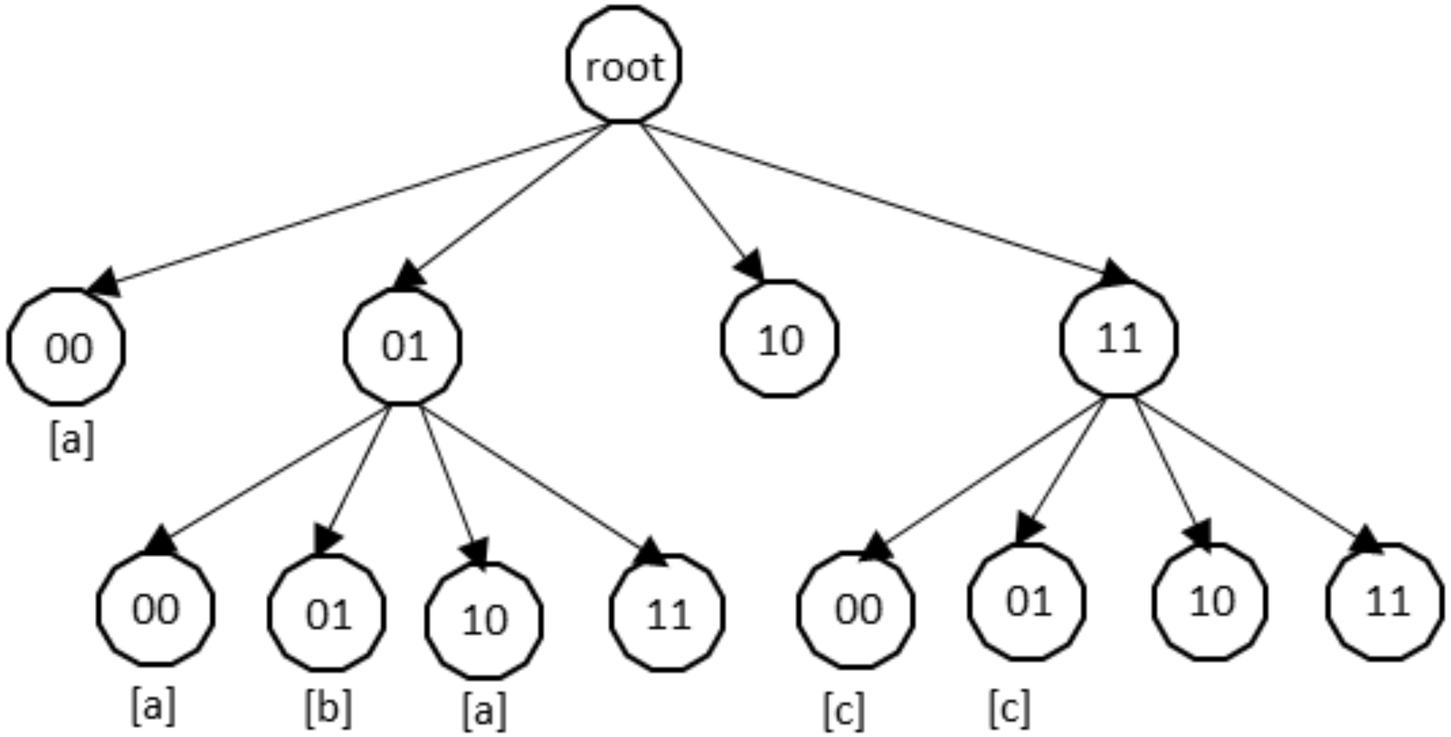
Fixed Grid Index



https://gistbok.ucgis.org/bok-topics/2018-quarter-02/spatial-indexing

QuadTree





https://gistbok.ucgis.org/bok-topics/2018-quarter-02/spatial-indexing

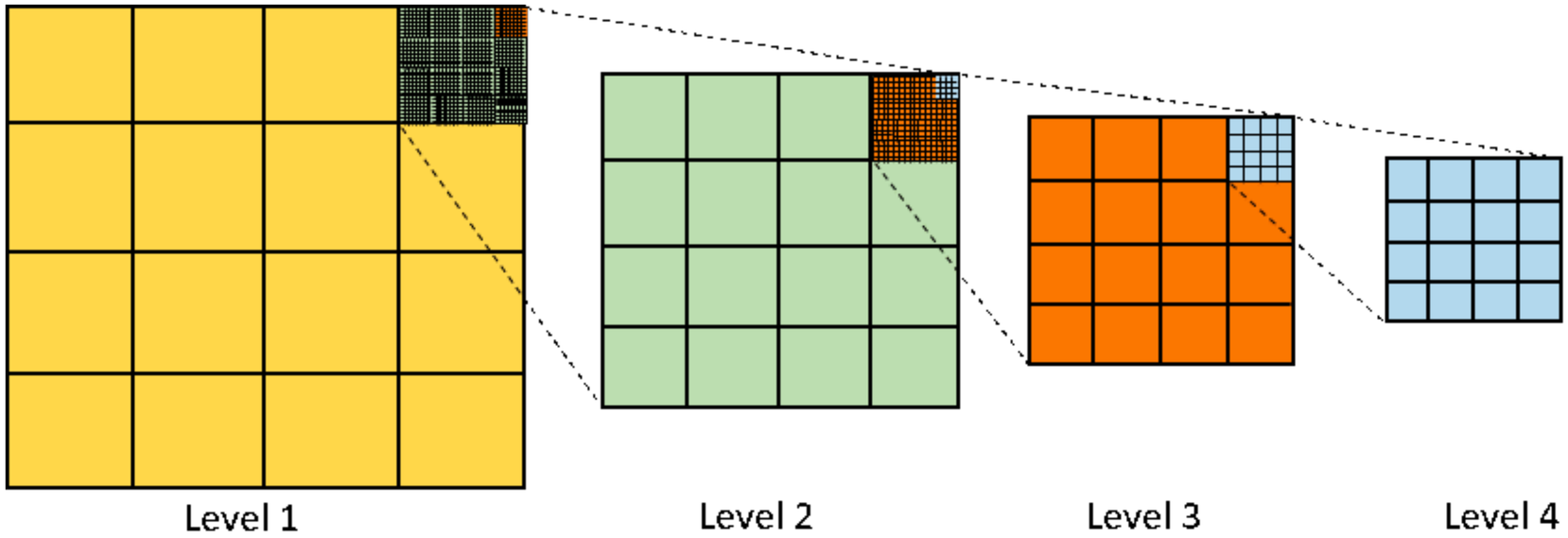
Grid-based Spatial Indexes

- A regular tessellation/mesh
- Divides space into a series of contiguous cells

***

- Often assigned unique identifiers
- For stable spatial indexes
- Also used for data aggregation

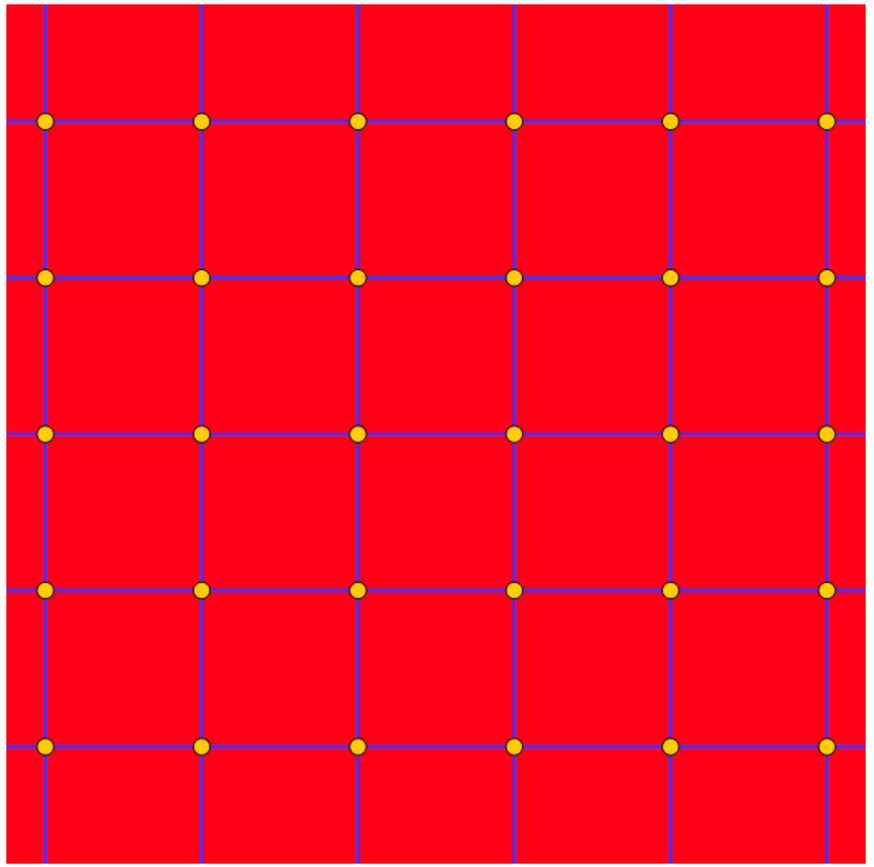
Grid-based Spatial Indexes



https://gistbok.ucgis.org/bok-topics/2018-quarter-02/spatial-indexing

Grid-based Spatial Indexes

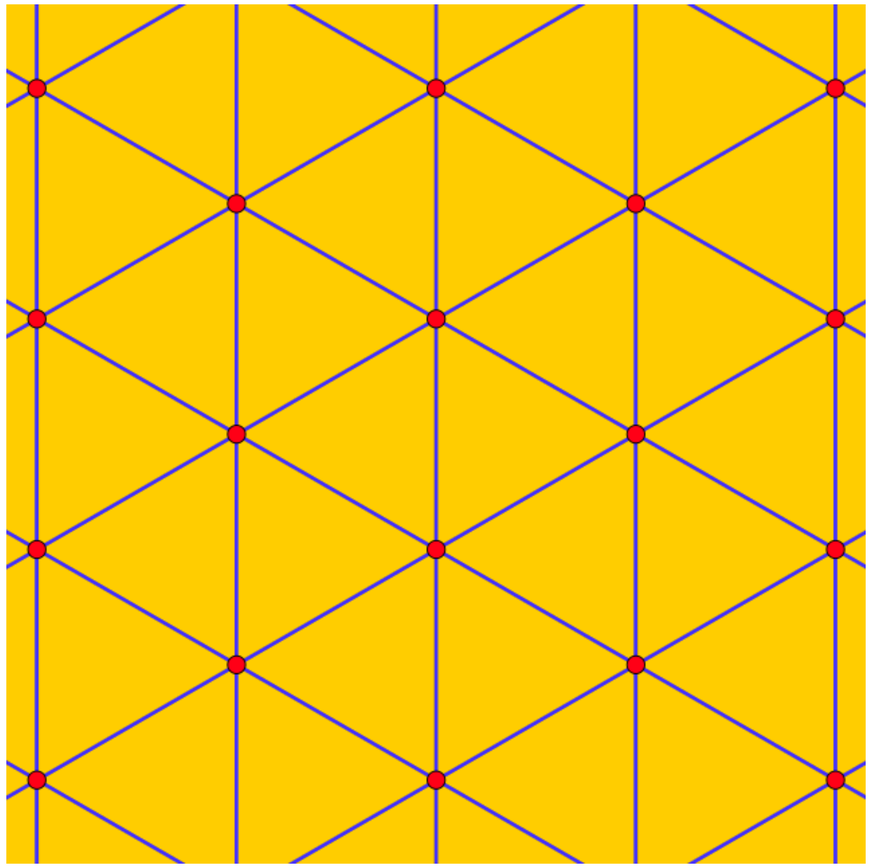
Square grid



https://en.wikipedia.org/wiki/Grid_(spatial_index)

Grid-based Spatial Indexes

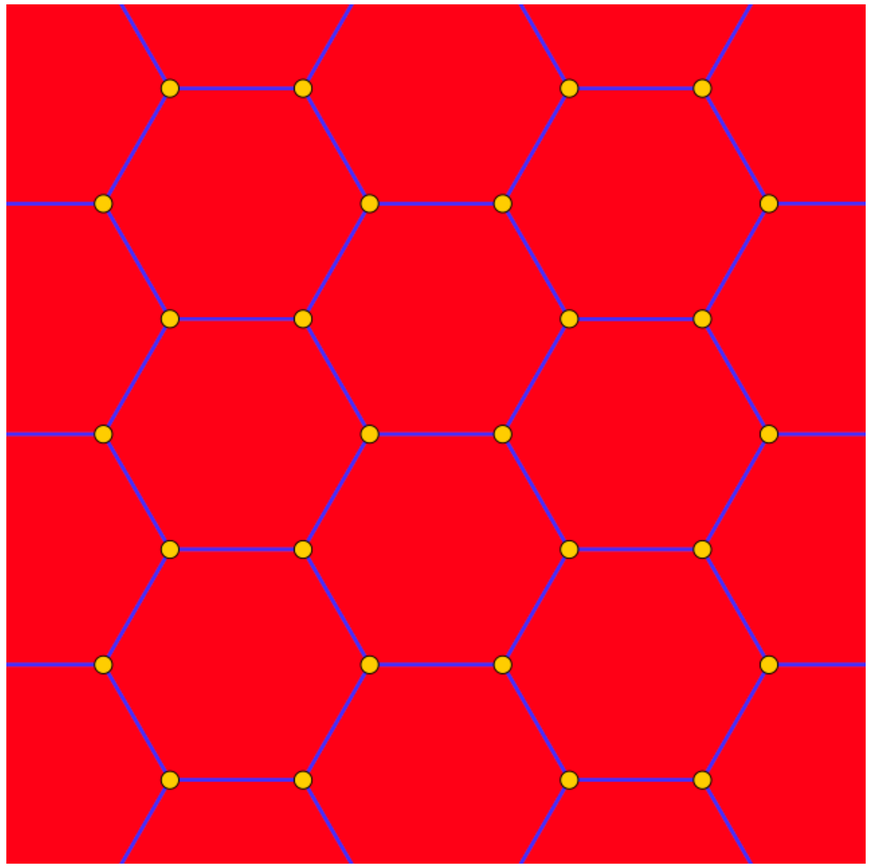
Triangular grid



https://en.wikipedia.org/wiki/Grid_(spatial_index)

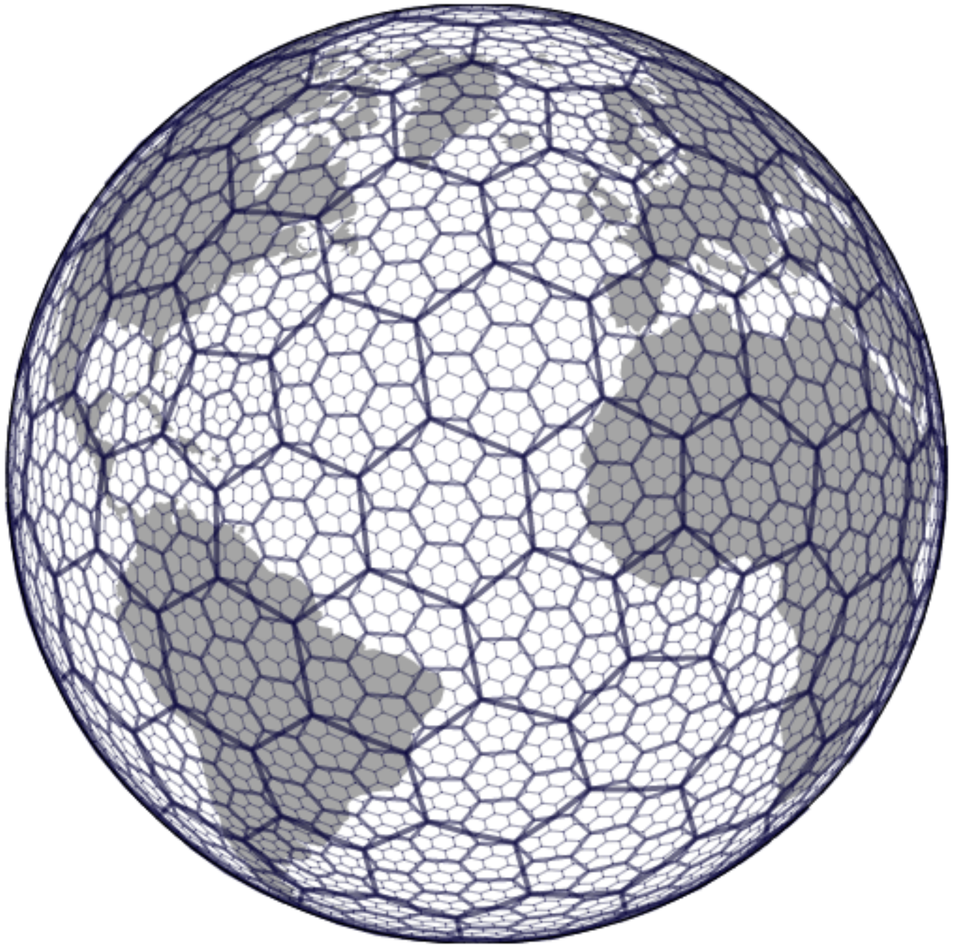
Grid-based Spatial Indexes

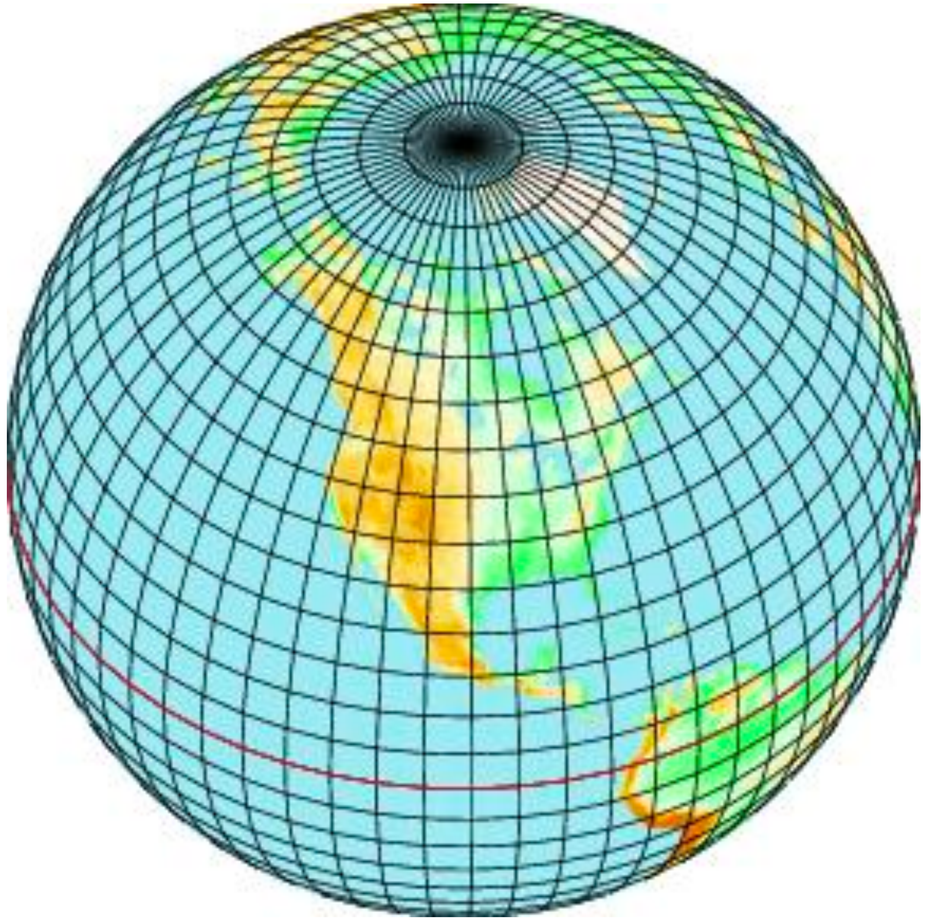
Hex grid



https://en.wikipedia.org/wiki/Grid_(spatial_index)

Grid-based Spatial Indexes





https://h3geo.org/

Choosing your spatial index

Consider:

- Geometry type
- Distribution
- Dimensions
- Query types
- Update frequency
- Stability

Tools for **big** spatial data

Spatial indexes in GeoPandas

`geopandas.sindex.RTreeIndex`

`geopandas.sindex.PyGEOSSTRTreeIndex`

https://geopandas.org/en/stable/docs/reference/sindex.html

H3

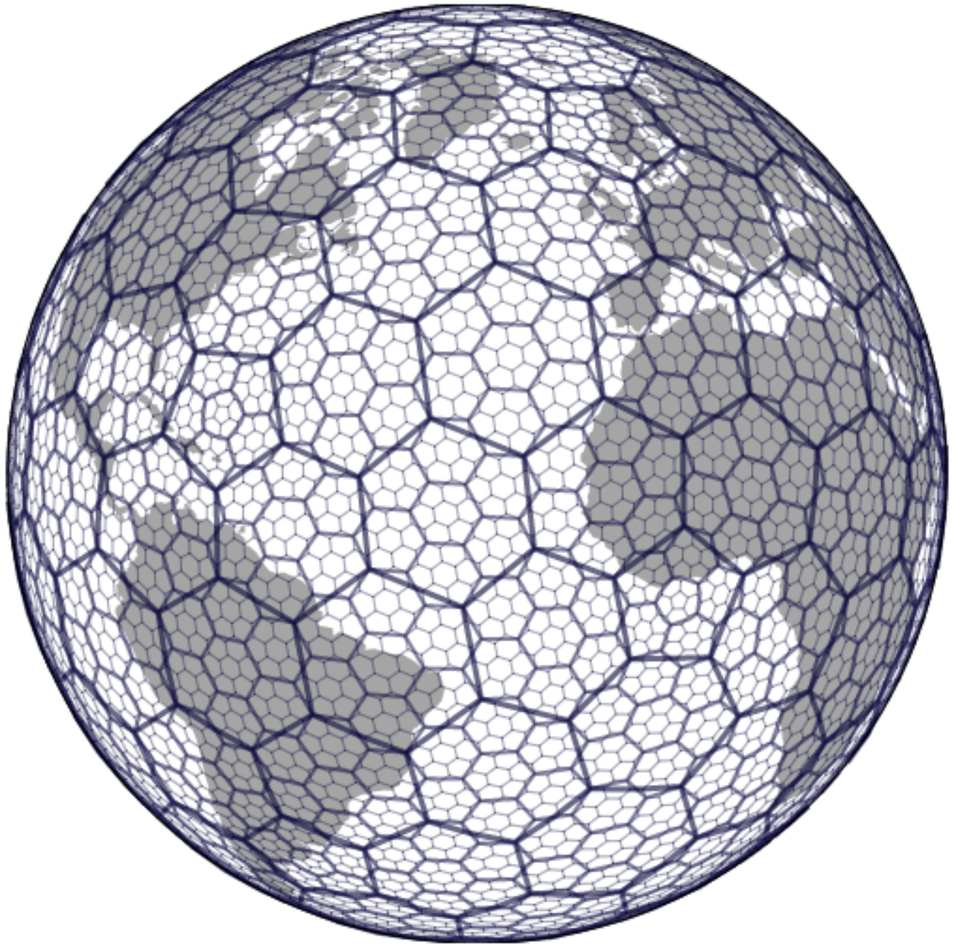
Hexagonal global hierarchical geospatial indexing system

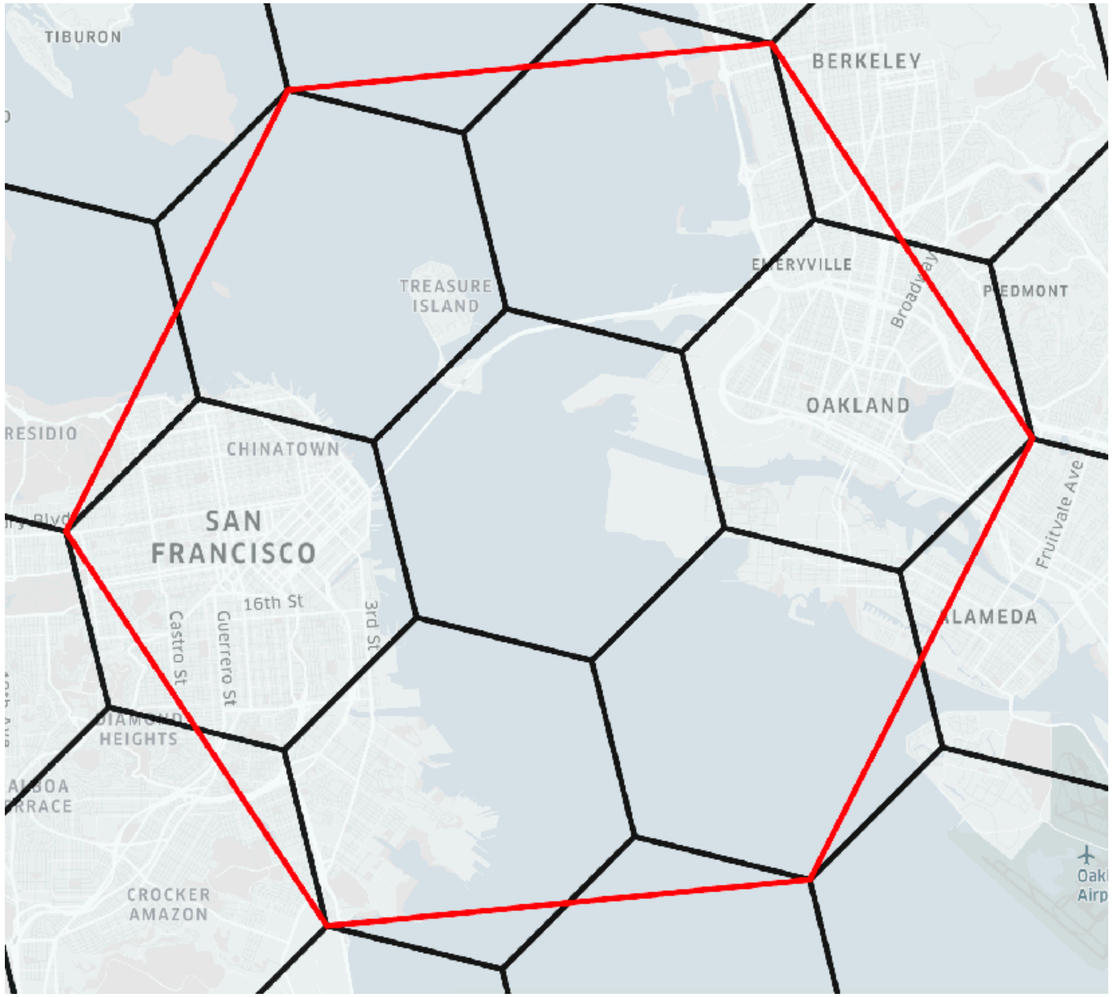
Developed by Uber

For spatial indexing, data aggregation, data joining, flow analysis, etc.

https://h3geo.org/

H3





https://h3geo.org/

H3

Resolution from 0 - 15

Compatible with WGS84

Res	Total number of cells	Number of hexagons	Number of pentagons
0	122	110	12
1	842	830	12
2	5,882	5,870	12
3	41,162	41,150	12
4	288,122	288,110	12
5	2,016,842	2,016,830	12
6	14,117,882	14,117,870	12
7	98,825,162	98,825,150	12
8	691,776,122	691,776,110	12
9	4,842,432,842	4,842,432,830	12
10	33,897,029,882	33,897,029,870	12
11	237,279,209,162	237,279,209,150	12
12	1,660,954,464,122	1,660,954,464,110	12
13	11,626,681,248,842	11,626,681,248,830	12
14	81,386,768,741,882	81,386,768,741,870	12
15	569,707,381,193,162	569,707,381,193,150	12

https://h3geo.org/

H3

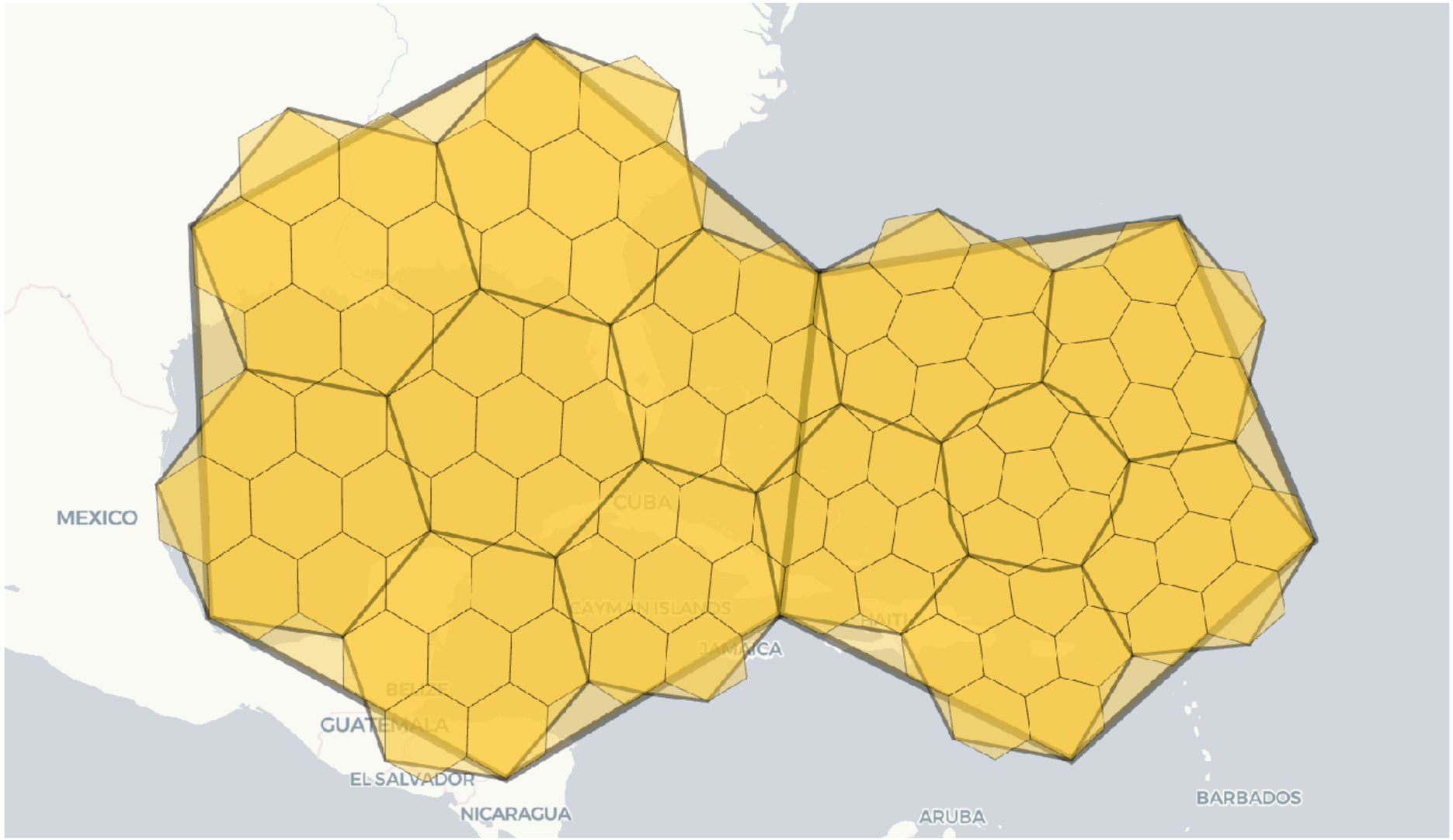
Resolution from 0 - 15

Compatible with WGS84

Res	Average Hexagon Area (km²)	Pentagon Area* (km²)	Ratio (P/H)
0	4,357,449.416078381	2,562,182.162955496	0.5880
1	609,788.441794133	328,434.586246469	0.5386
2	86,801.780398997	44,930.898497879	0.5176
3	12,393.434655088	6,315.472267516	0.5096
4	1,770.347654491	896.582383141	0.5064
5	252.903858182	127.785583023	0.5053
6	36.129062164	18.238749548	0.5048
7	5.161293360	2.604669397	0.5047
8	0.737327598	0.372048038	0.5046
9	0.105332513	0.053147195	0.5046
10	0.015047502	0.007592318	0.5046
11	0.002149643	0.001084609	0.5046
12	0.000307092	0.000154944	0.5046
13	0.000043870	0.000022135	0.5046
14	0.000006267	0.000003162	0.5046
15	0.000000895	0.000000452	0.5046

https://h3geo.org/

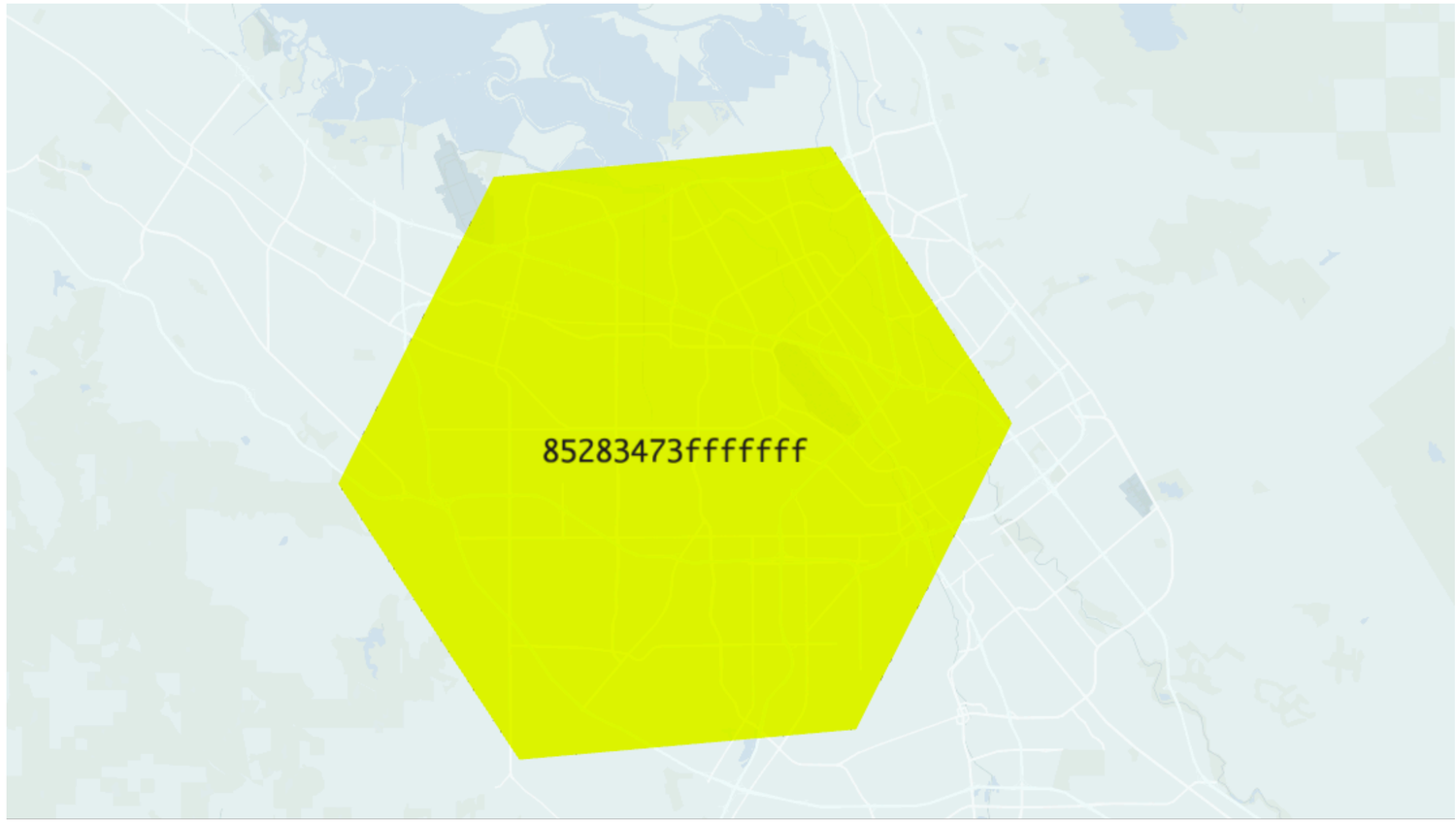
H3



https://h3geo.org/

H3

Each hex is indexed by a unique hex id



https://h3geo.org/

Other examples of global spatial indexes

- S2, an open source, hierarchical, discrete, and global grid system using square cells.
- Geohash, a system for encoding locations using a string of characters, creating a hierarchical, square grid system (a quadtree).
- Placekey, a system for encoding points of interest (POIs) which incorporates H3 in its POI identifier.

https://h3geo.org/docs/comparisons/admin/

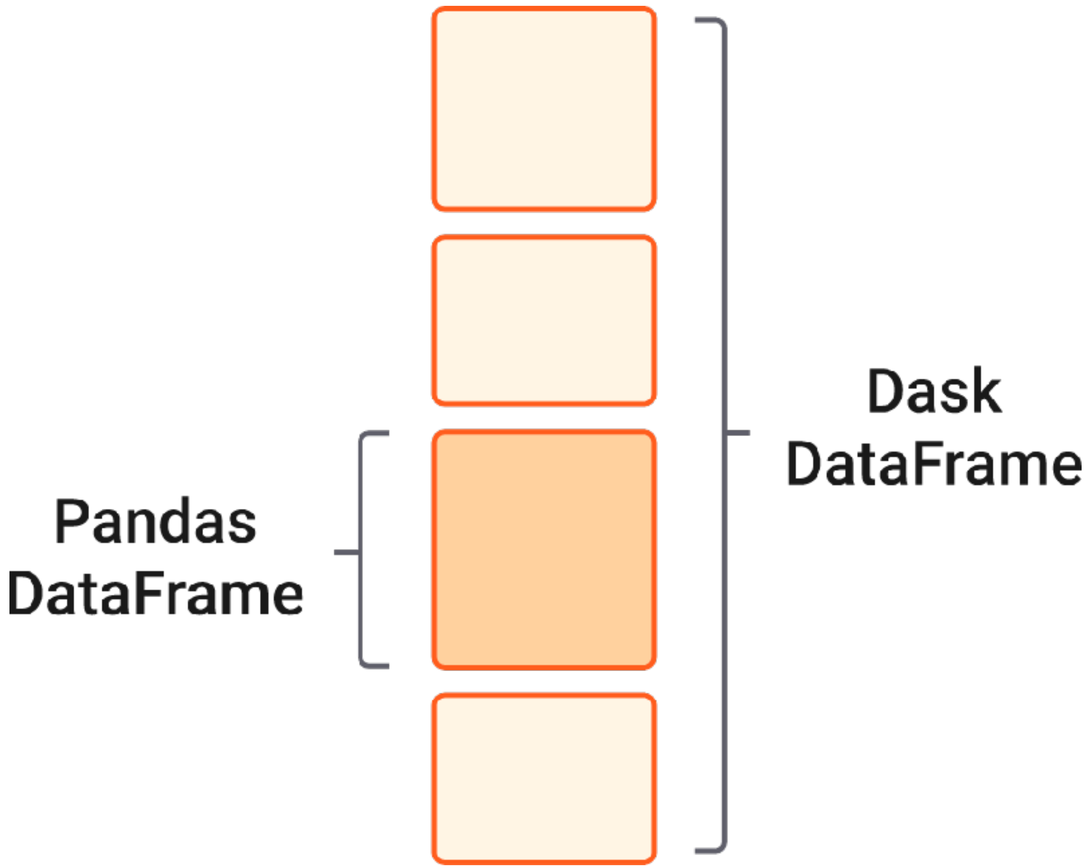
Dask-GeoPandas

“Parallel GeoPandas”

For GeoDataFrames that do not fit in memory or require expensive computation that can easily be parallelized

https://dask-geopandas.readthedocs.io/en/stable/

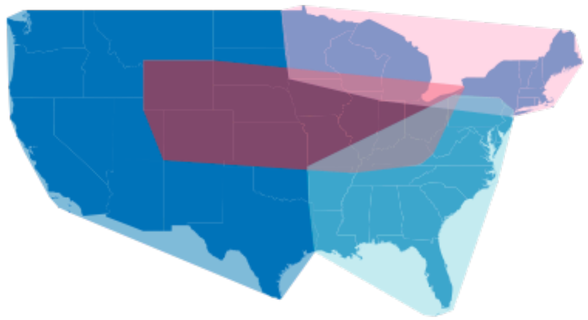
Dask-GeoPandas: Regular Partitioning



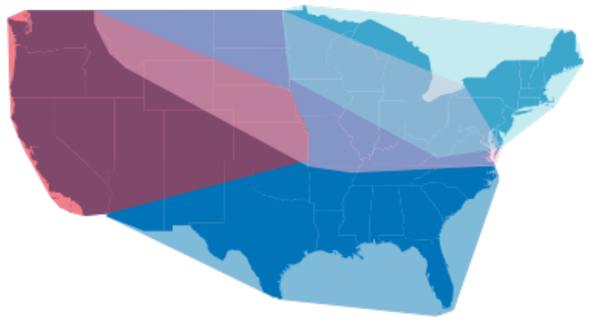
https://dask-geopandas.readthedocs.io/en/stable/guide/spatial-partitioning.html

Dask-GeoPandas: Spatial Partitioning

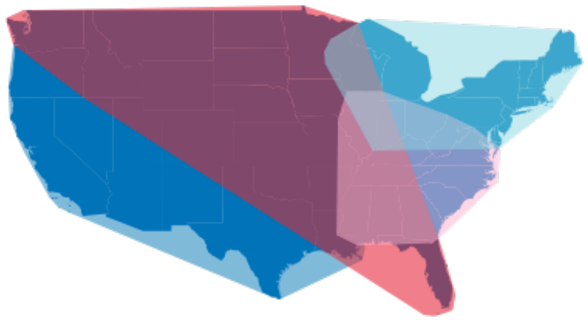
Hilbert



Morton



Geohash



https://dask-geopandas.readthedocs.io/en/stable/guide/spatial-partitioning.html

File formats

GeoParquet

Feather/Apache Arrow

https://geoparquet.org/

https://geopandas.org/en/stable/docs/user_guide/io.html#apache-parquet-and-feather-file-formats

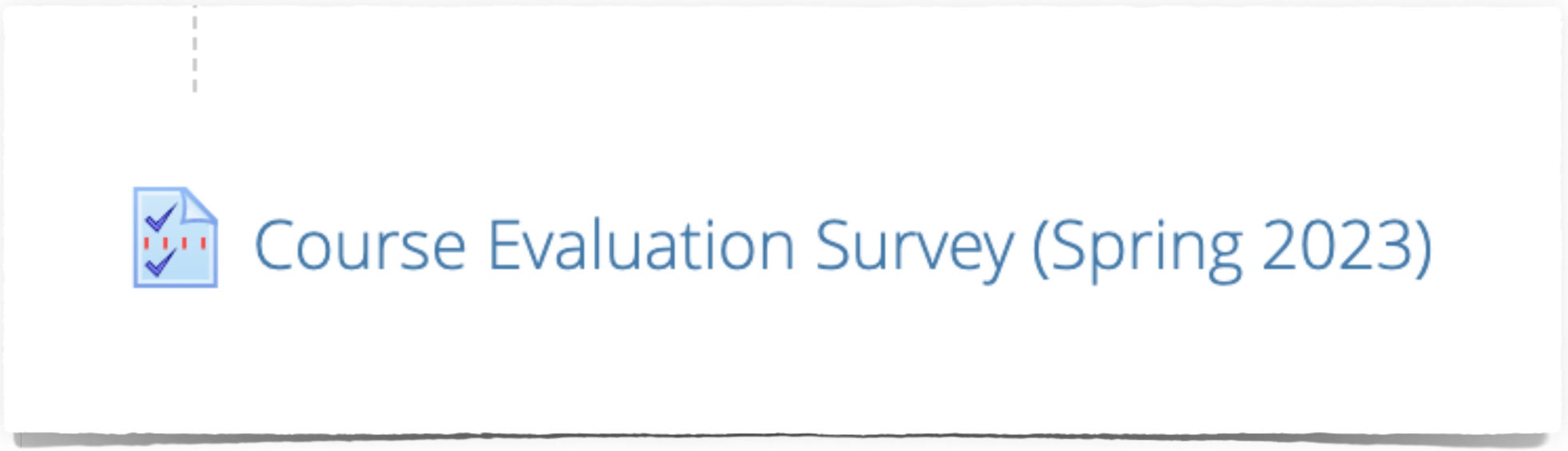
Could you speed up your project?

Discuss in groups:

- Which are the biggest performance barriers in your exam project? What is the problem?
- Where/how could spatial indexes be used to speed up your code?

(5 mins.)

Course evaluation



https://learnit.itu.dk/mod/questionnaire/view.php?id=181151

Jupyter

Using H3 in research projects

Sources and further materials for today's class

https://geoffboeing.com/2016/10/r-tree-spatial-index-python/

https://gistbok.ucgis.org/bok-topics/spatial-indexing

https://blog.mapbox.com/a-dive-into-spatial-search-algorithms-ebd0c5e39d2a

Geodesic Discrete Global Grid Systems

Next week: Spatial Data Visualization