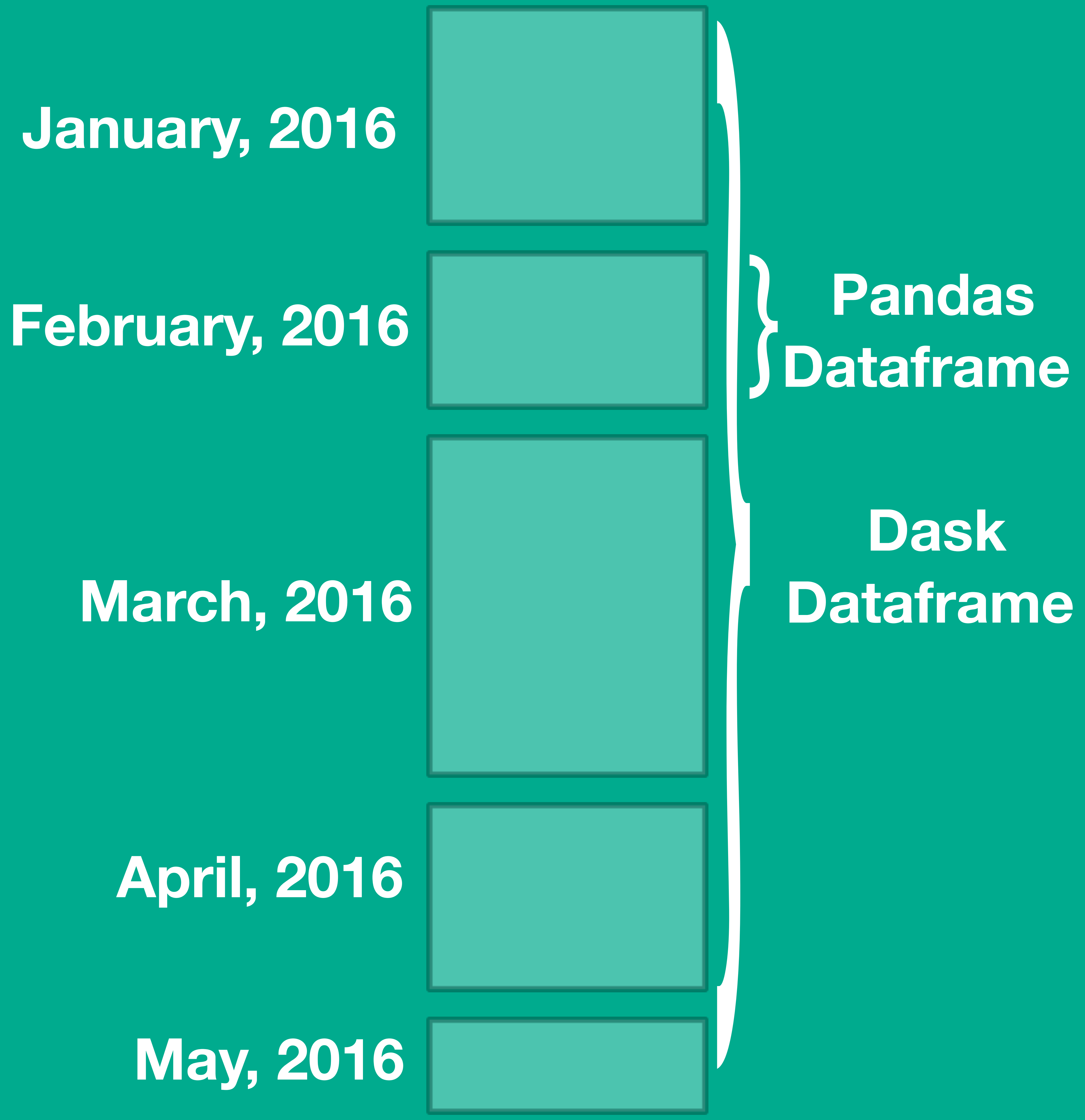
Dask Summit 2021

Scaling geospatial vector data

Partitioning of spatial data

Martin Fleischmann
@martinfleis

How Dask chunks data?



January, 2016

February, 2016

March, 2016

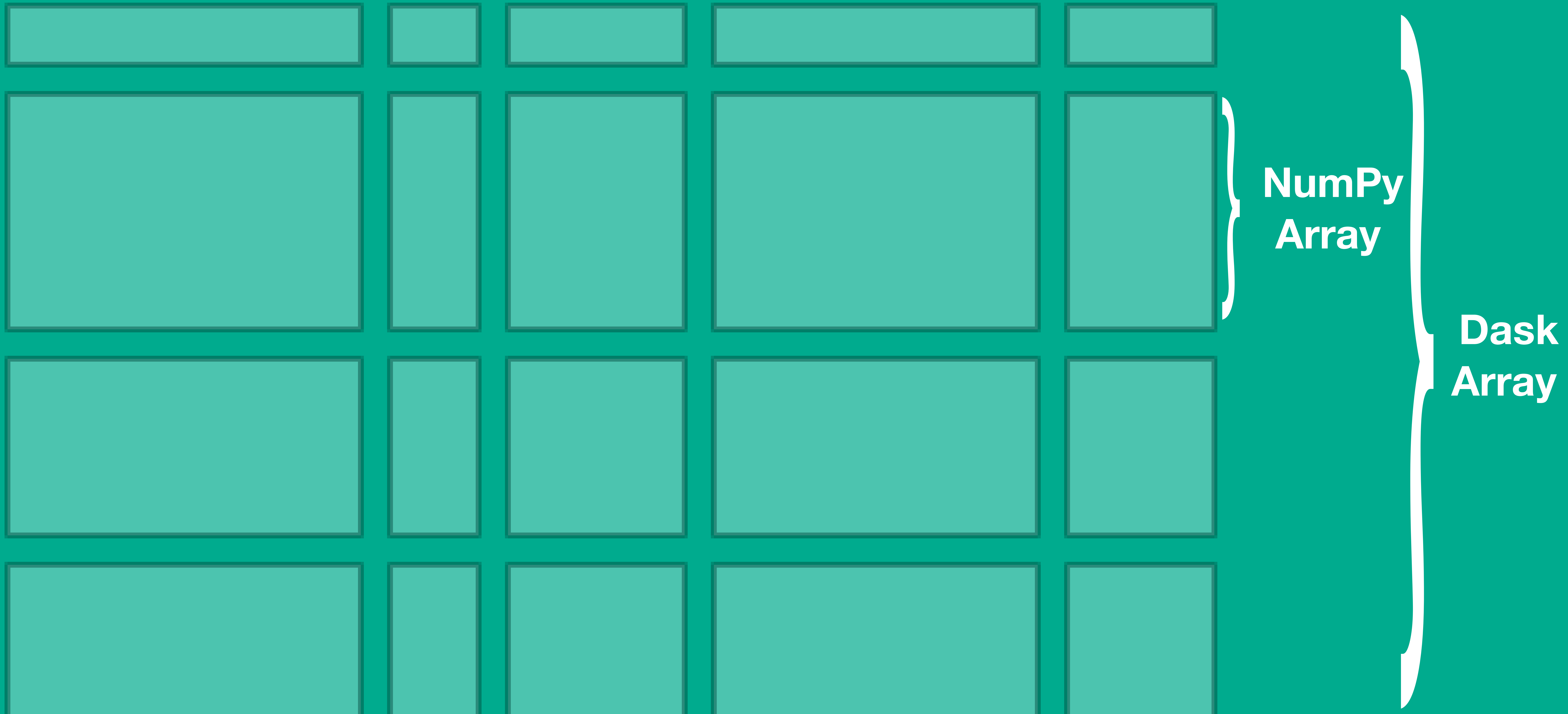
April, 2016

May, 2016

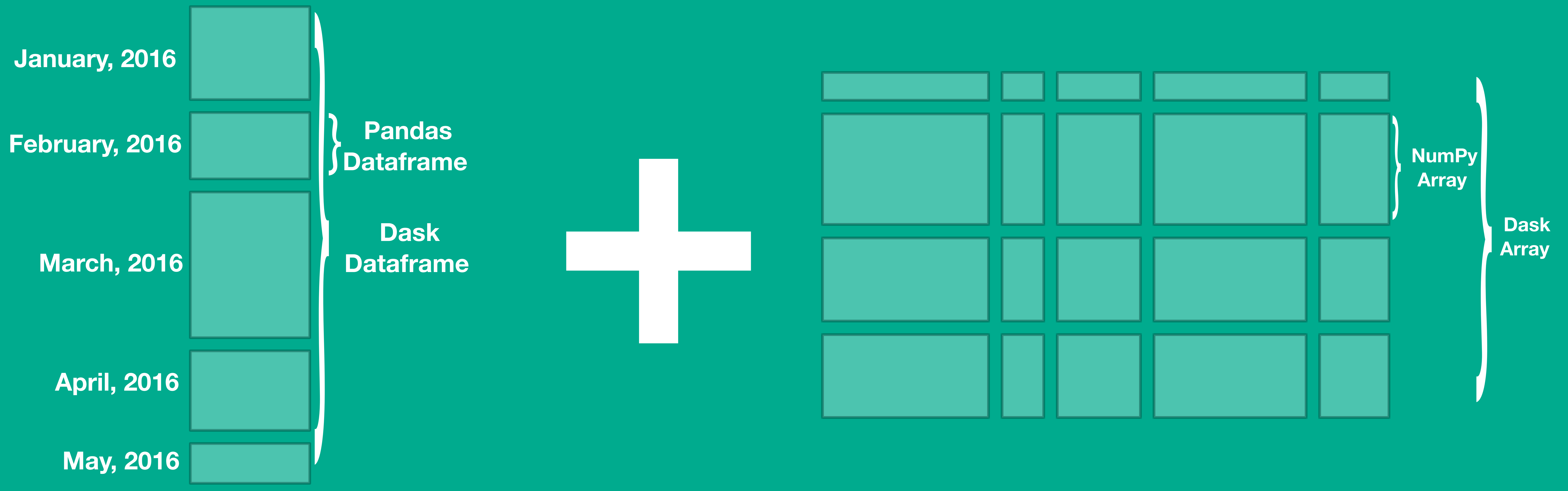
Pandas Dataframe

Dask Dataframe

dask.dataframe



dask.array



January, 2016

February, 2016

March, 2016

April, 2016

May, 2016

Pandas Dataframe

Dask Dataframe

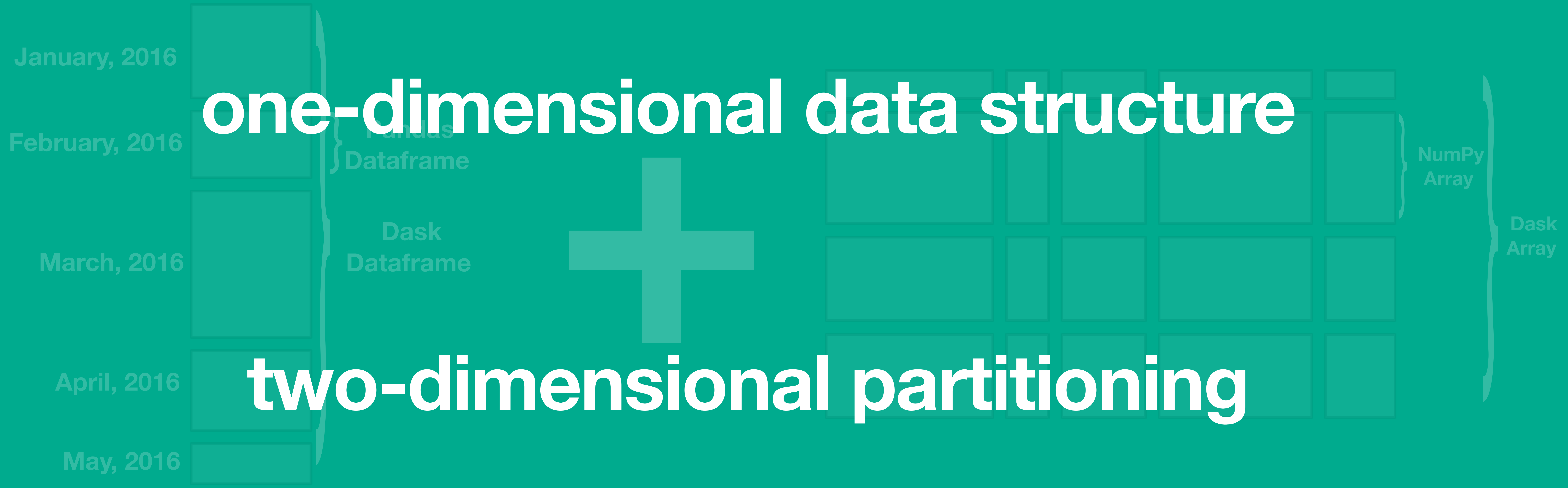
NumPy Array

Dask Array

spatial dataframe

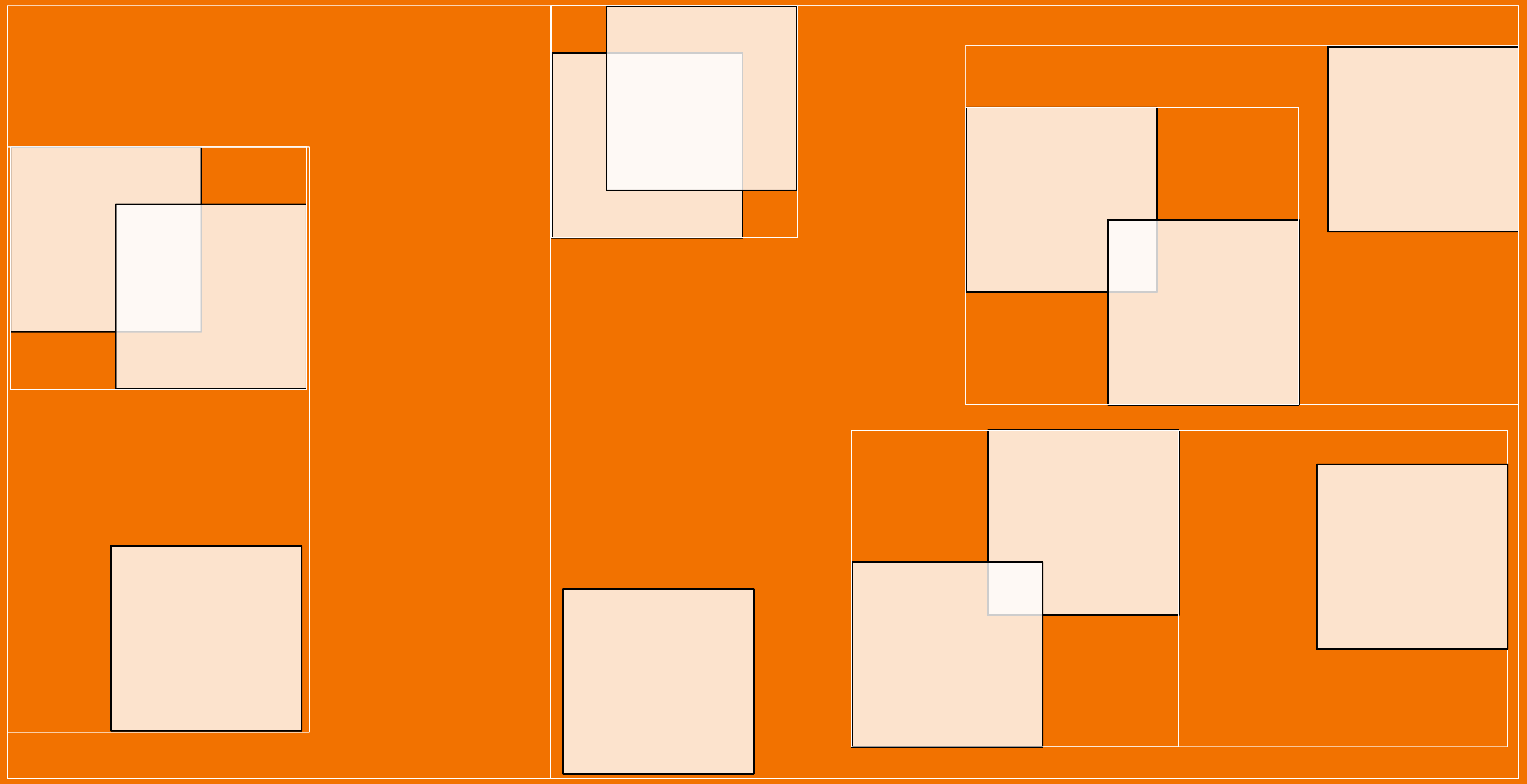
one-dimensional data structure

two-dimensional partitioning

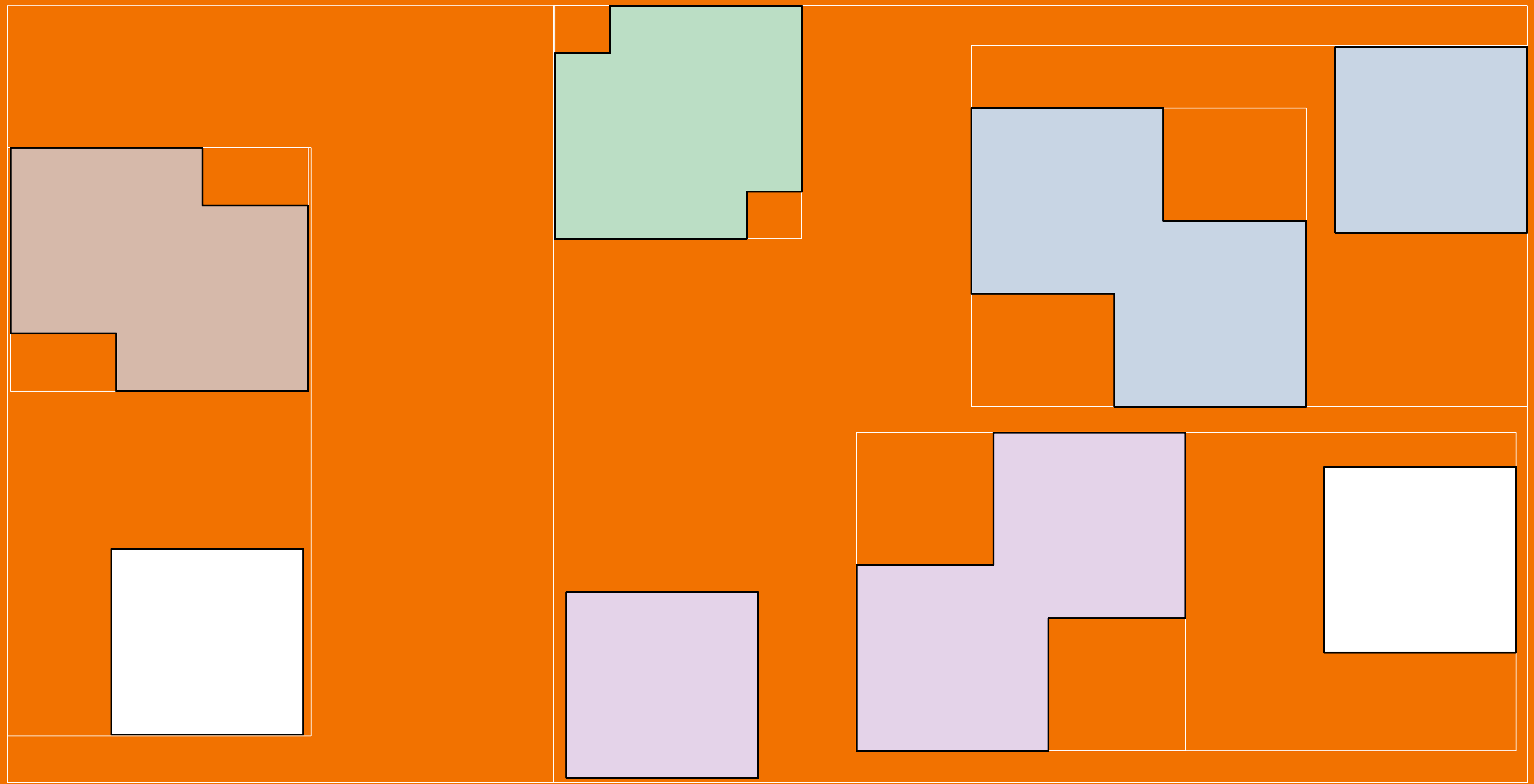


spatial dataframe

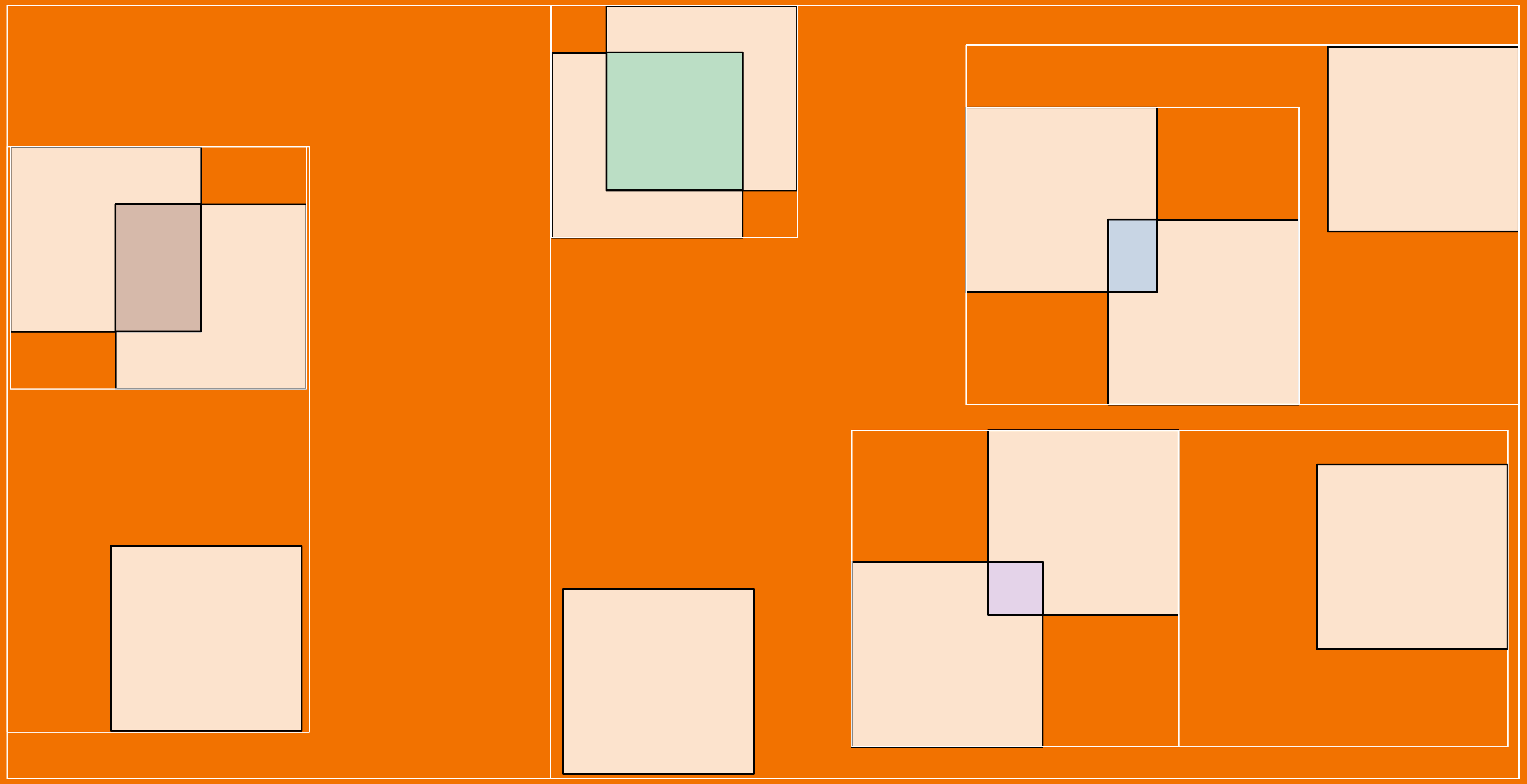
Why we need spatially coherent chunks?



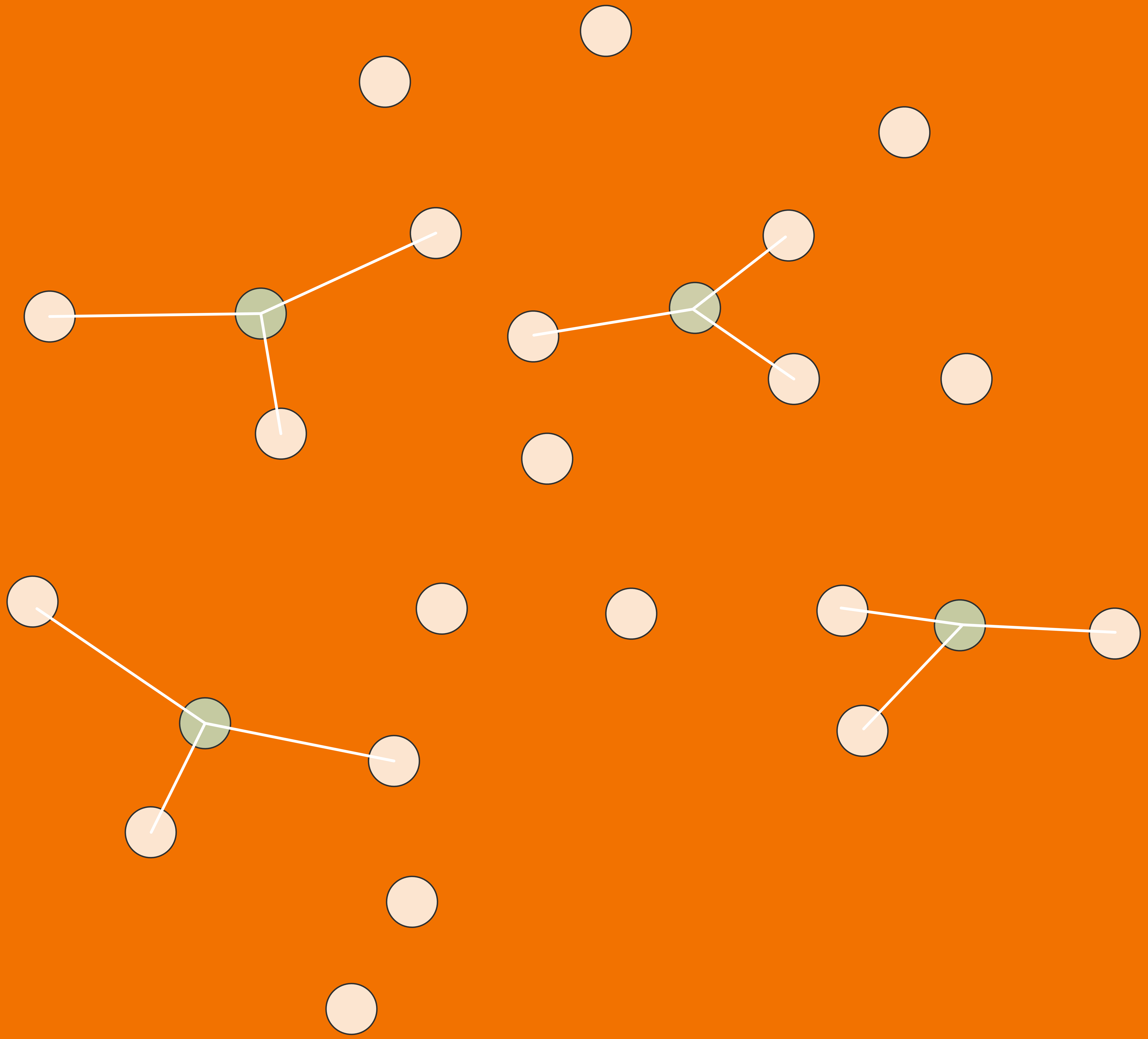
spatial indexing



spatial operations

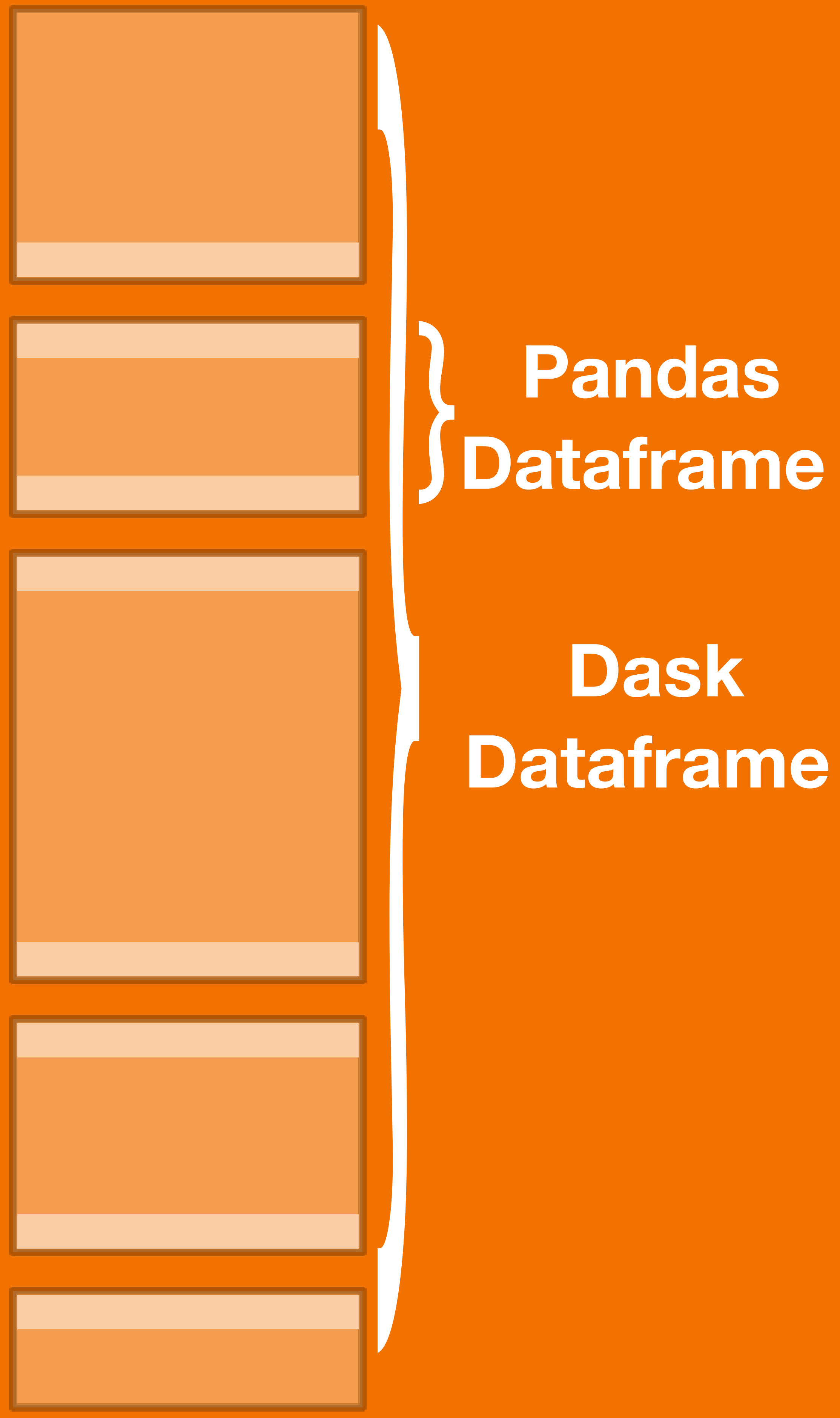


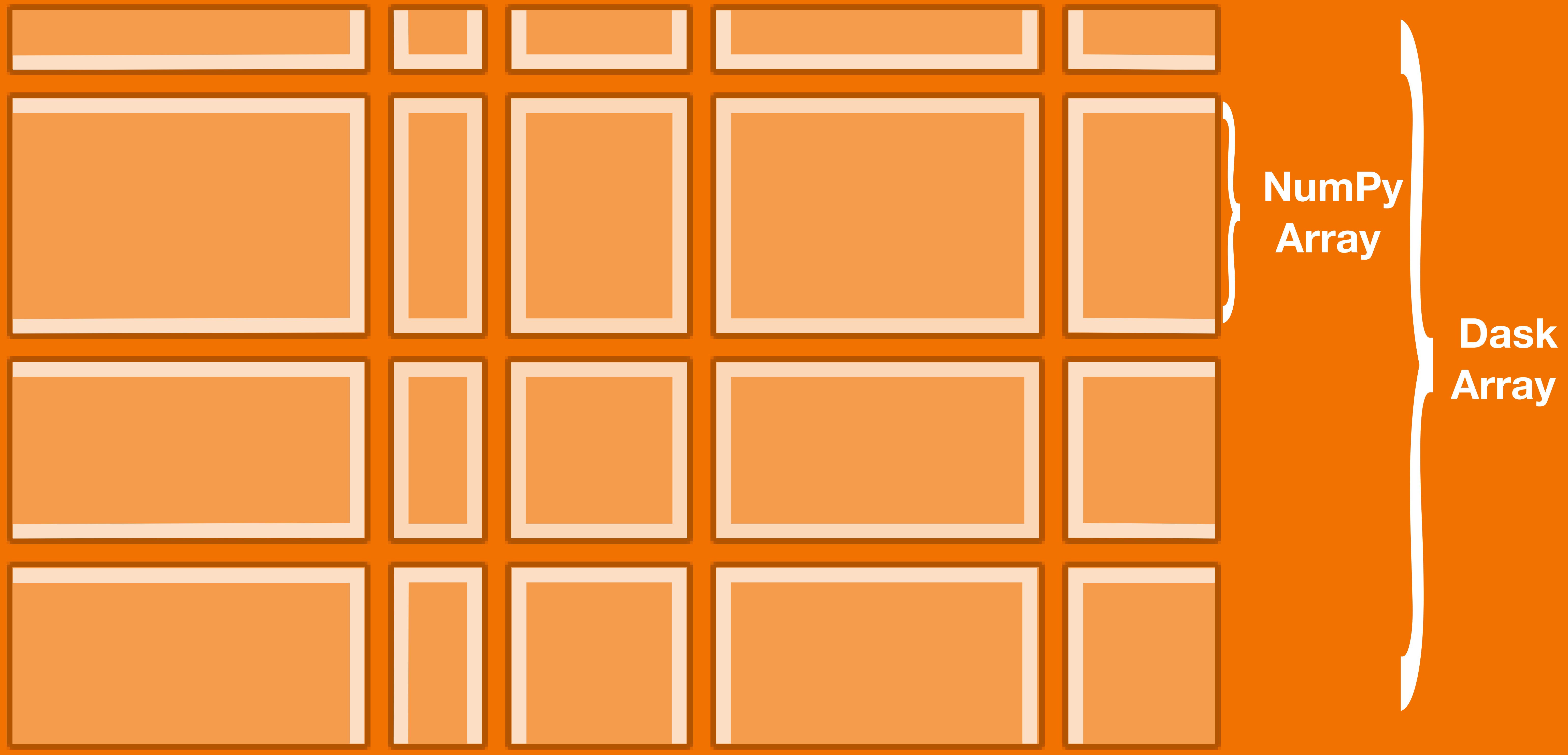
spatial operations



spatial proximity

overlapping computation





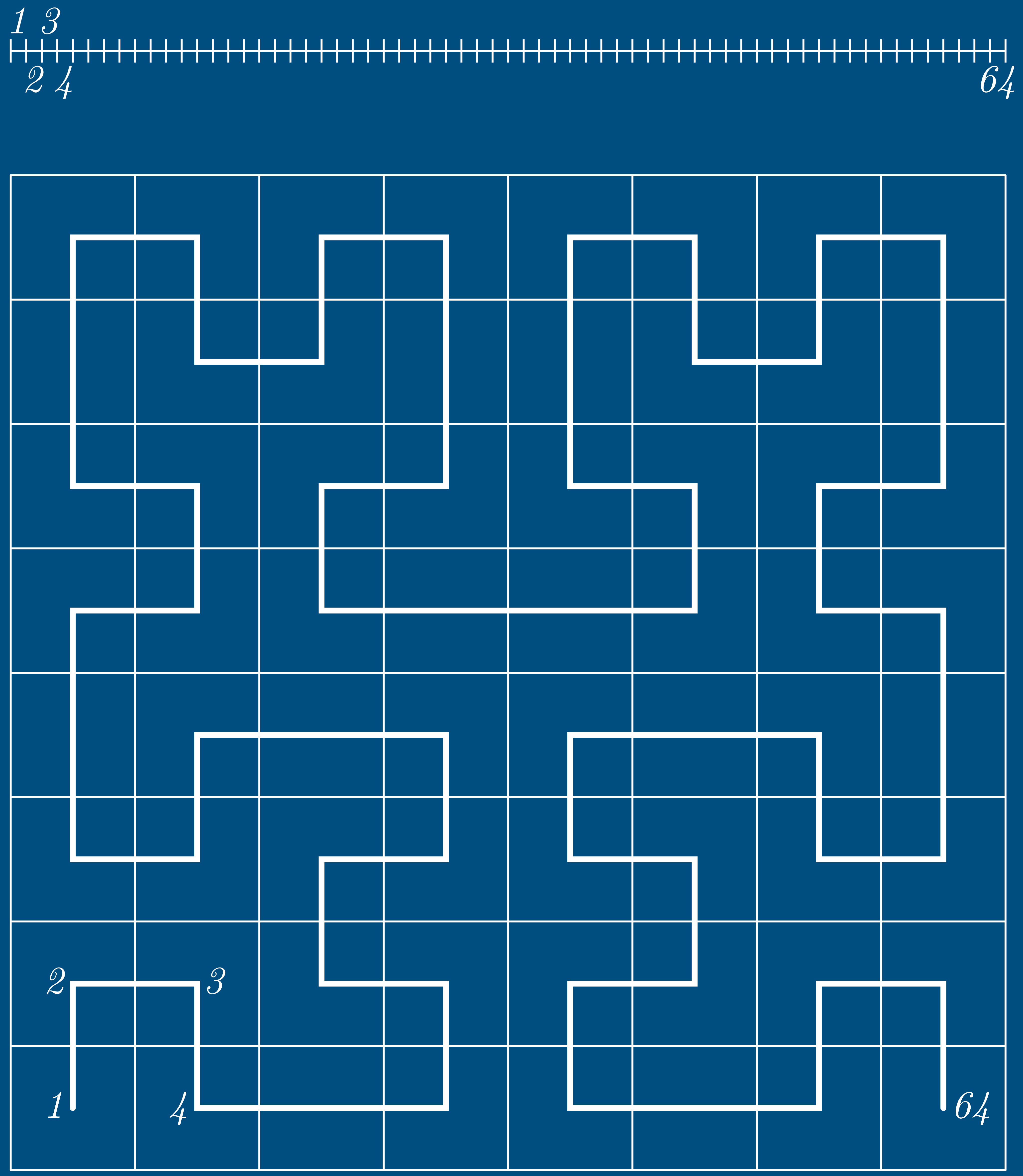
Pandas Dataframe
Dask Dataframe
NumPy Array
Dask Array

distance threshold

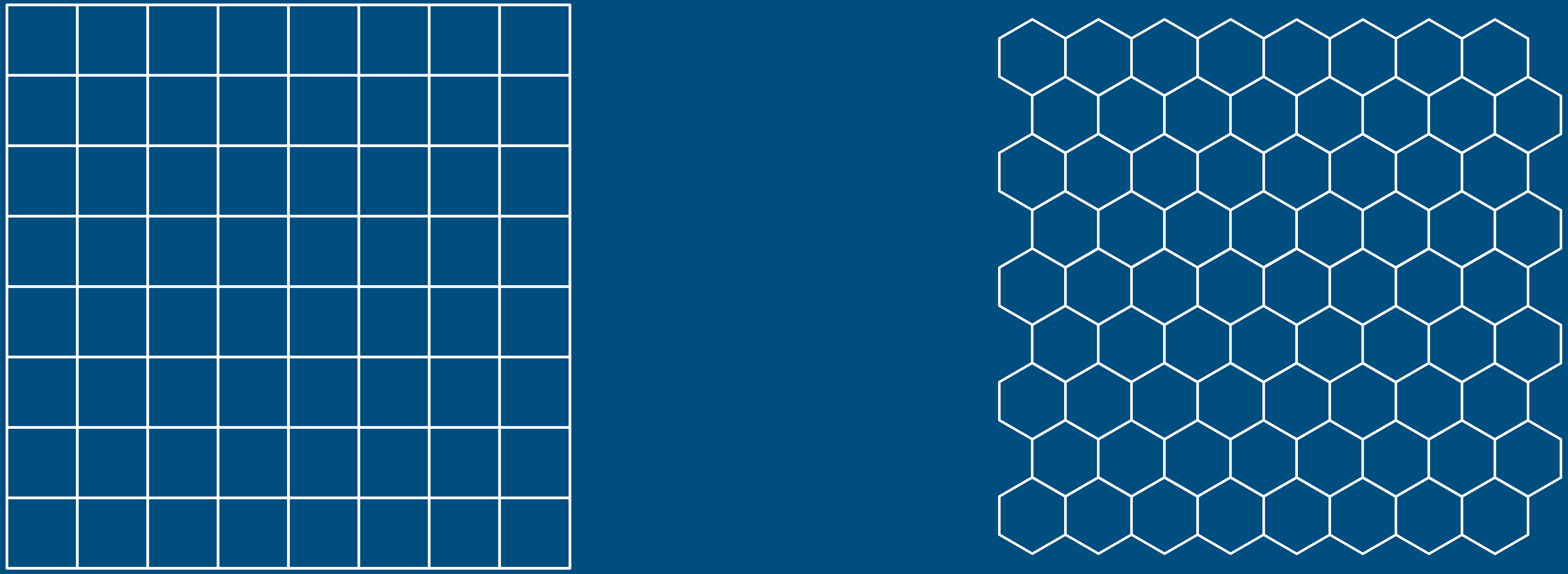
topological threshold

overlapping computation

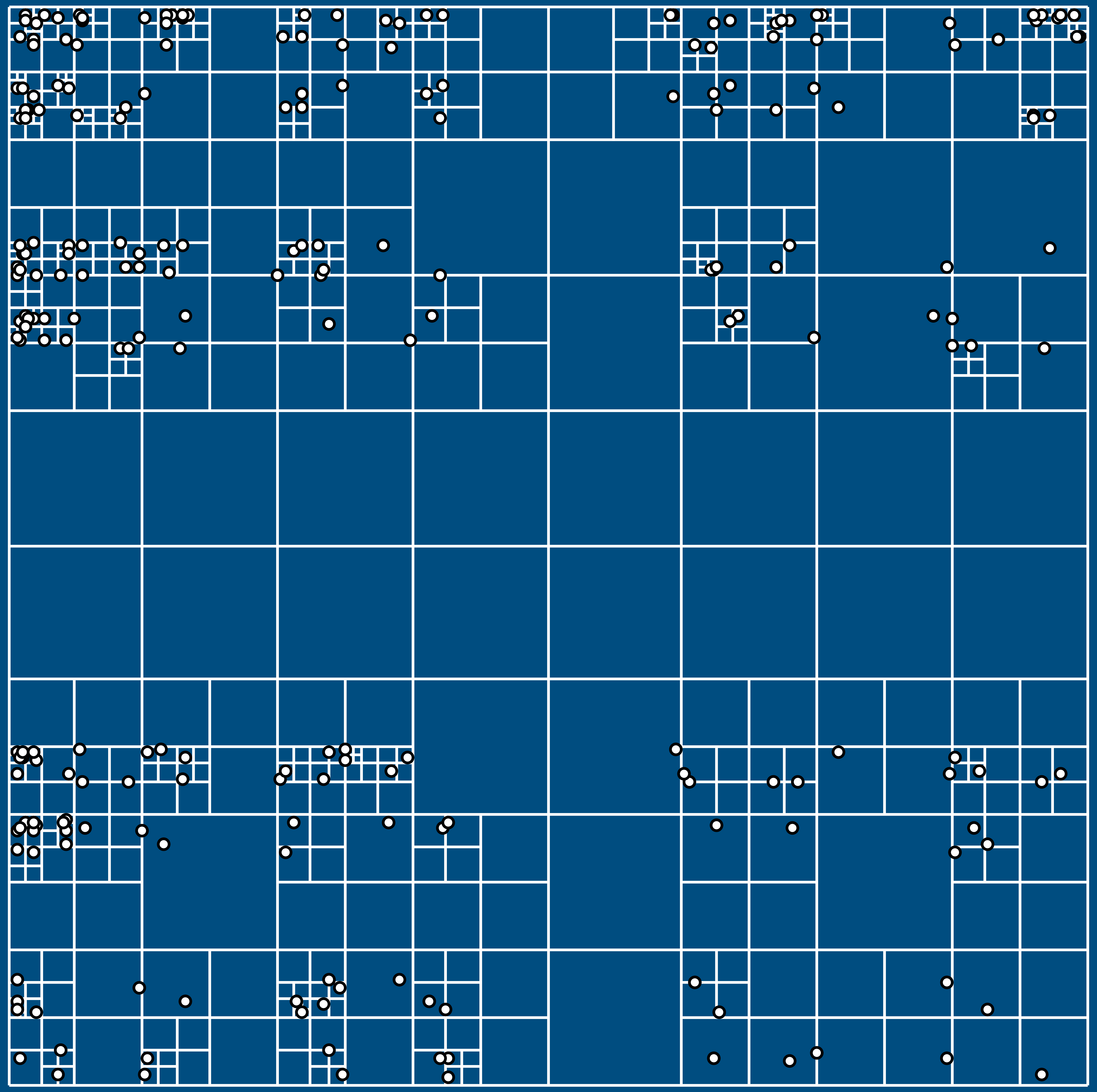
How can we achieve that?



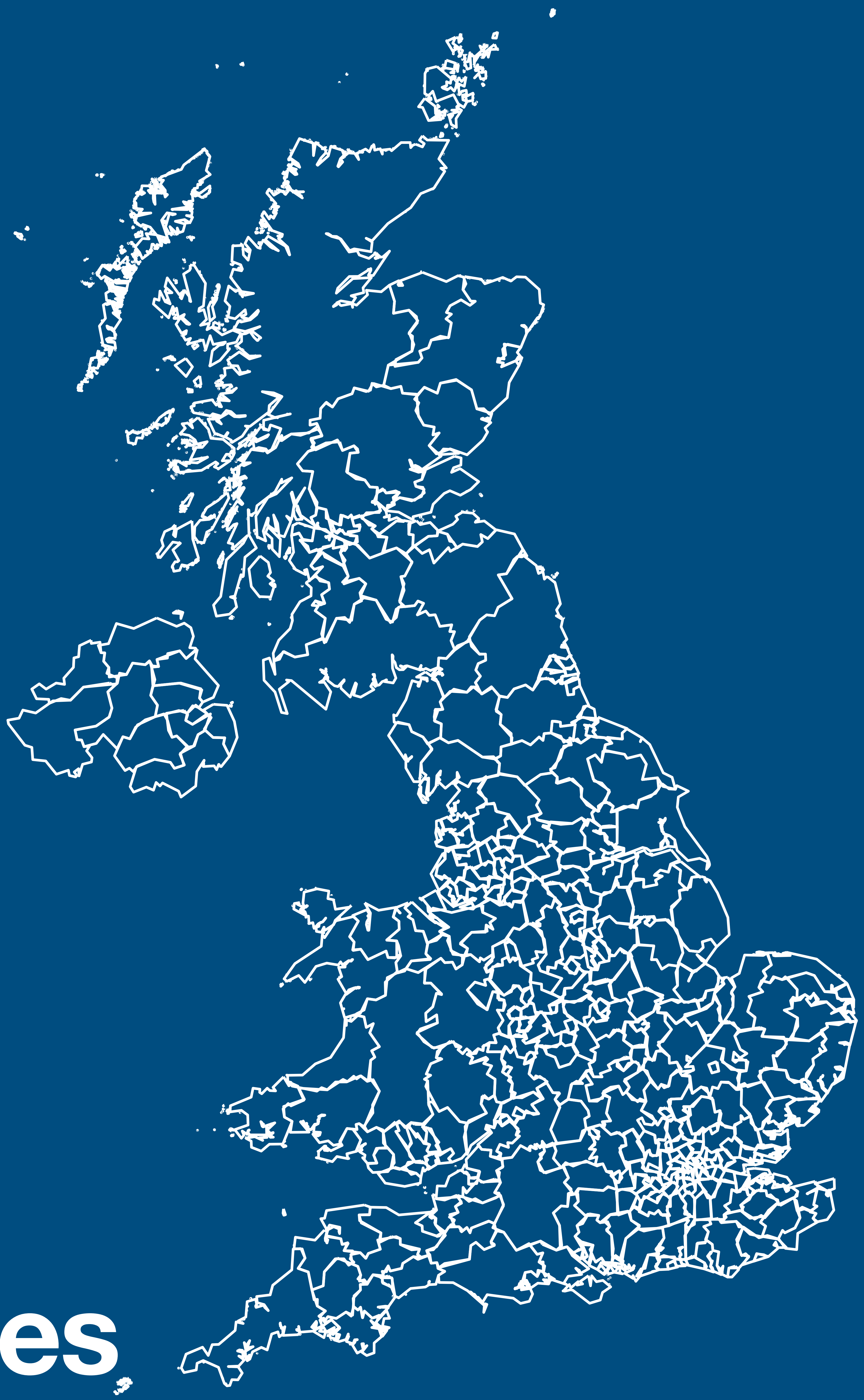
Hilbert curve



(arbitrary) grid



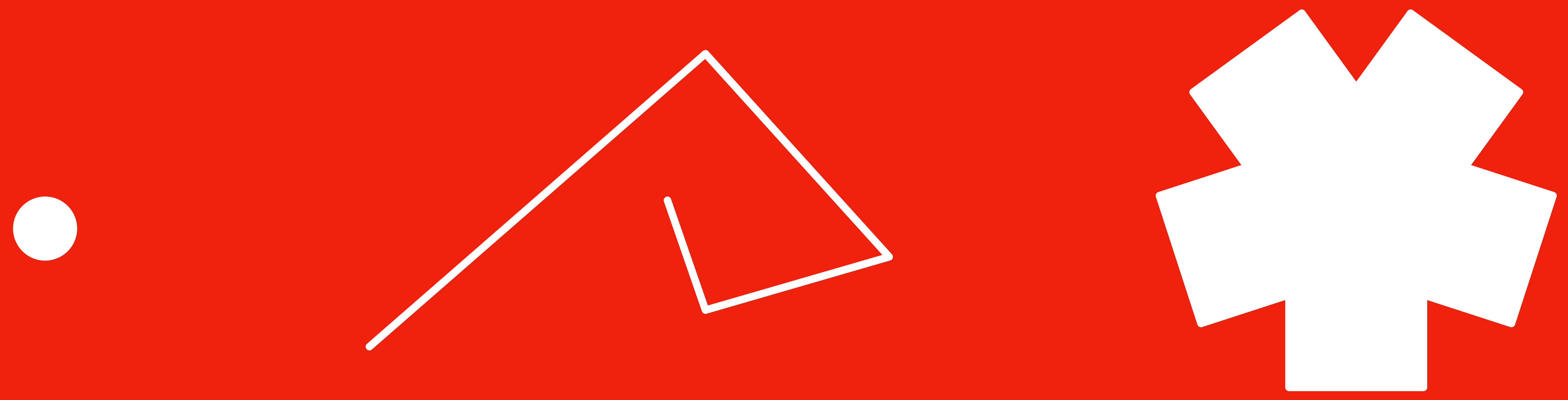
quadtree



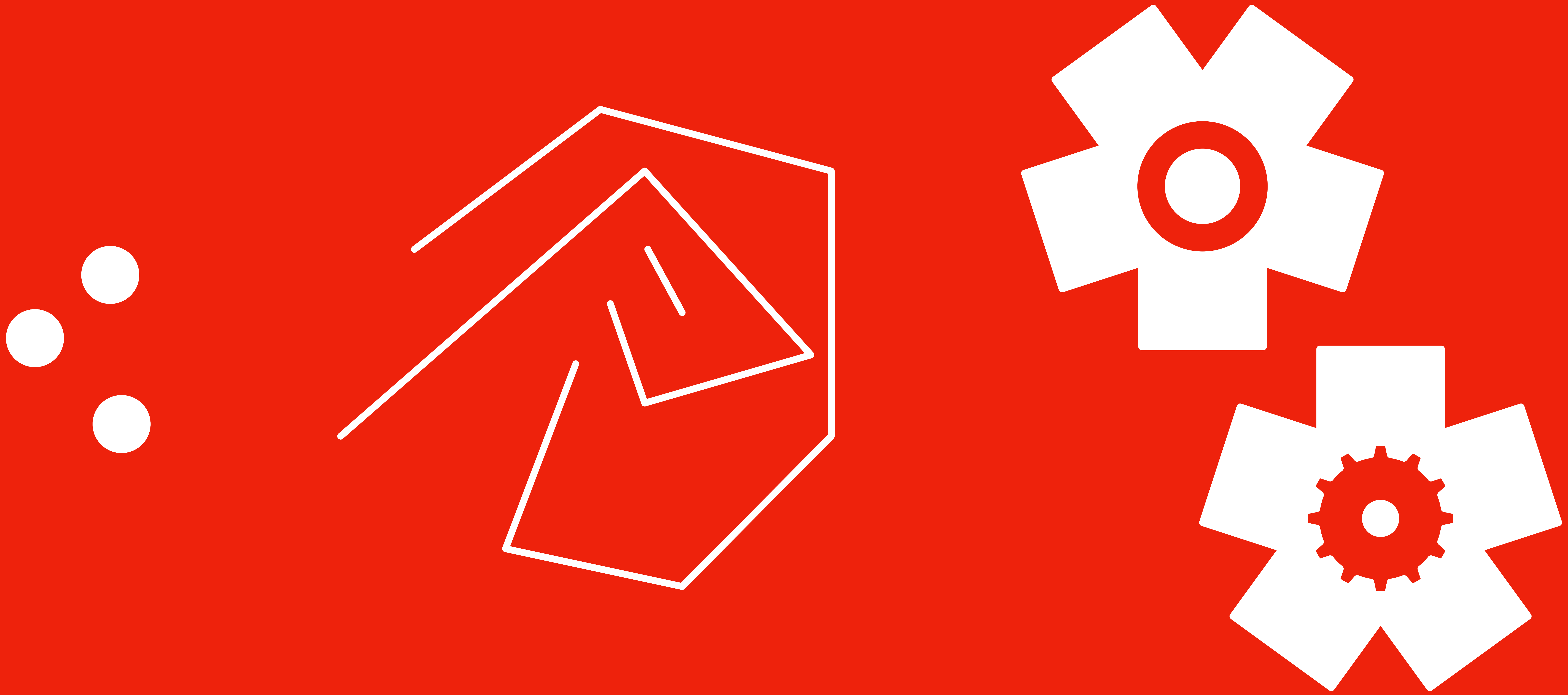
known boundaries

When it gets tricky

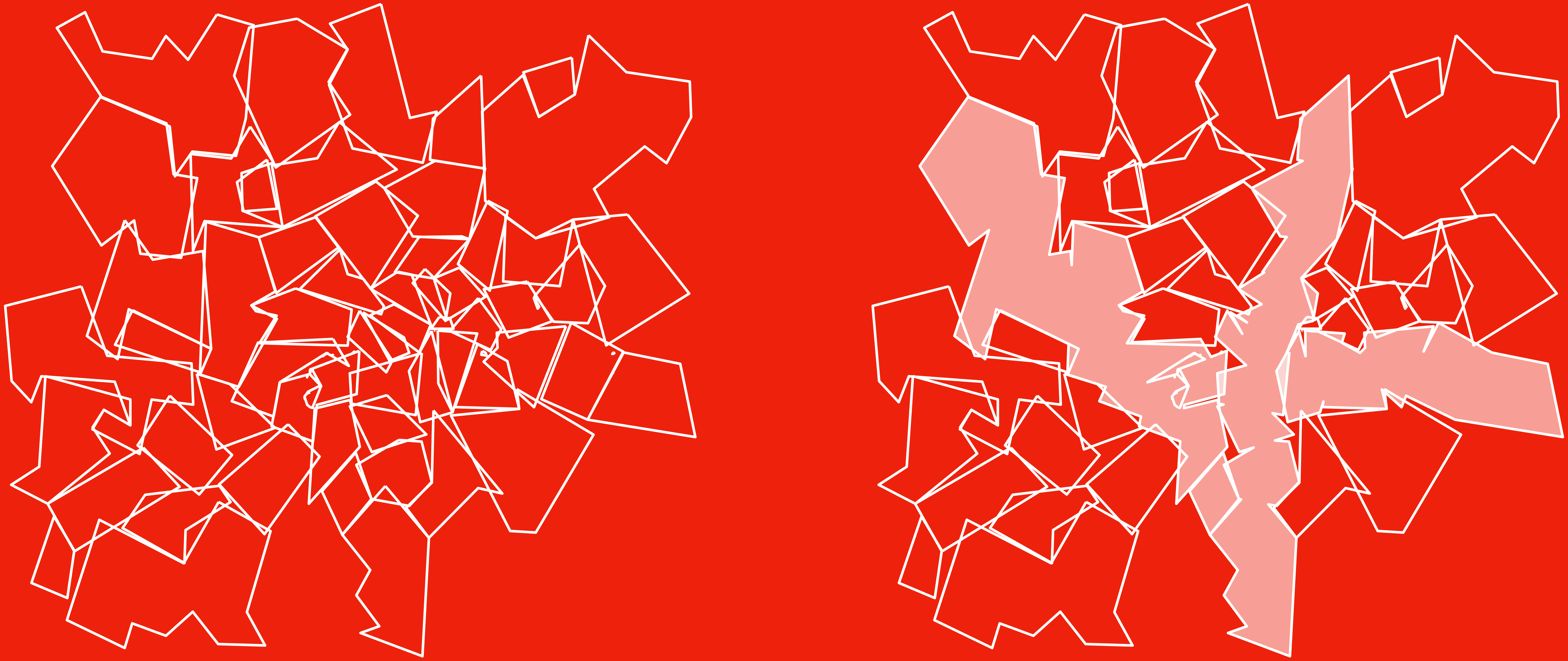
vector data are unpredictable



geometry types

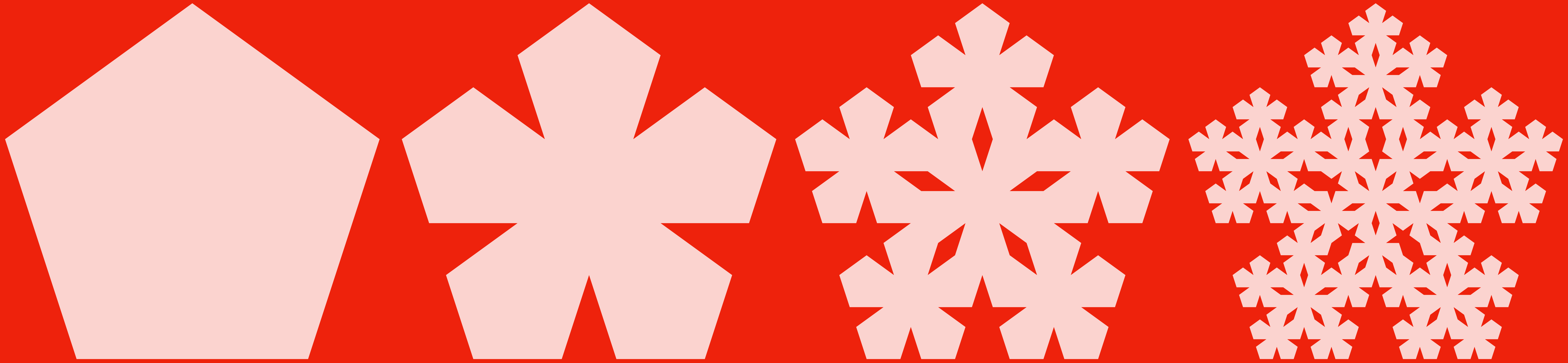


geometry types



boundaries

boundaries



complexity

spatial distribution

What would be nice to have

efficient cross-chunk spatial indexing

support for overlapping computations

Take into account memory demands of different rows?

Questions

How to create *smart* spatial partitions?

How to store spatial partitions?