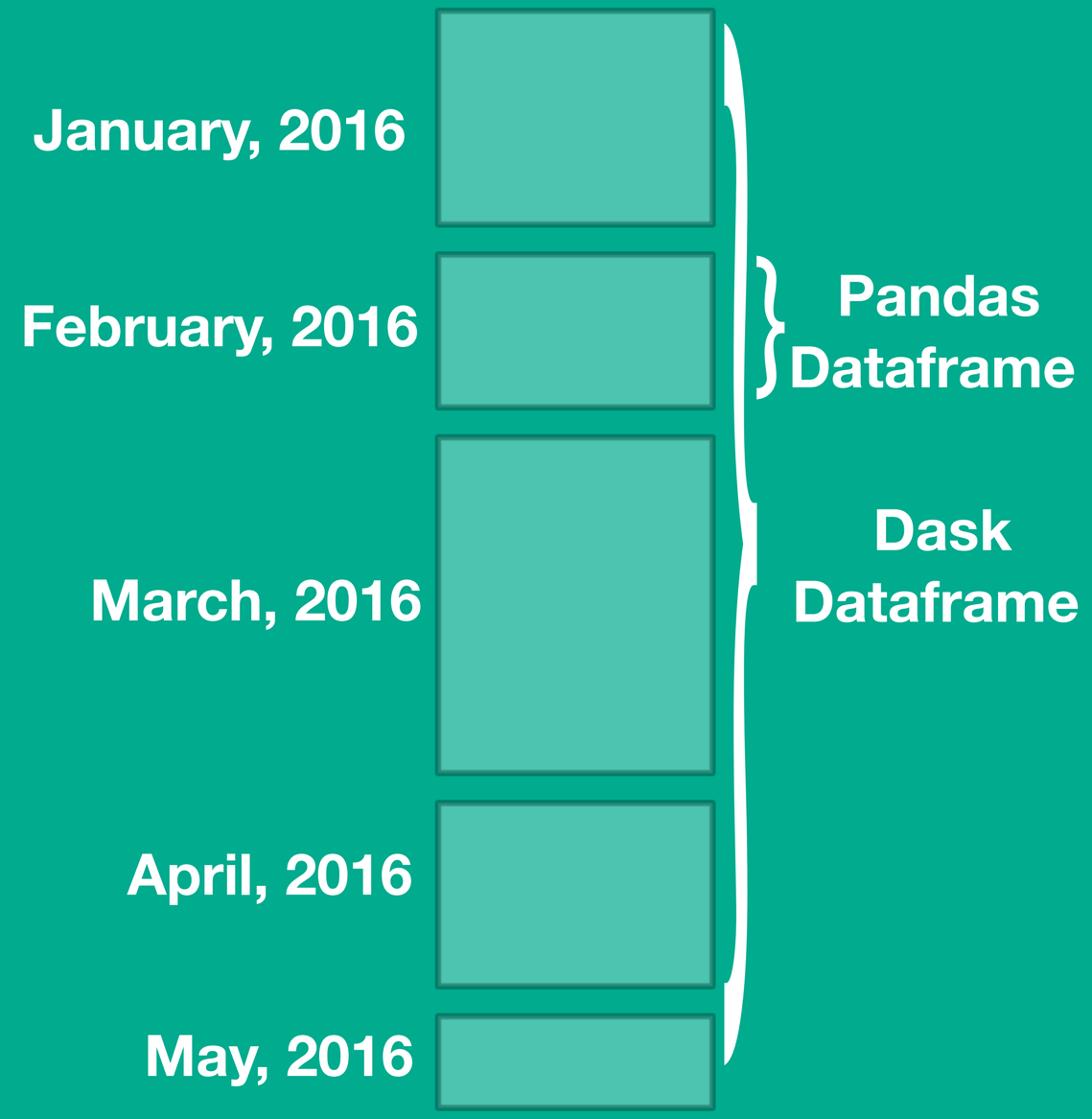
Dask Summit 2021

Scaling geospatial vector data

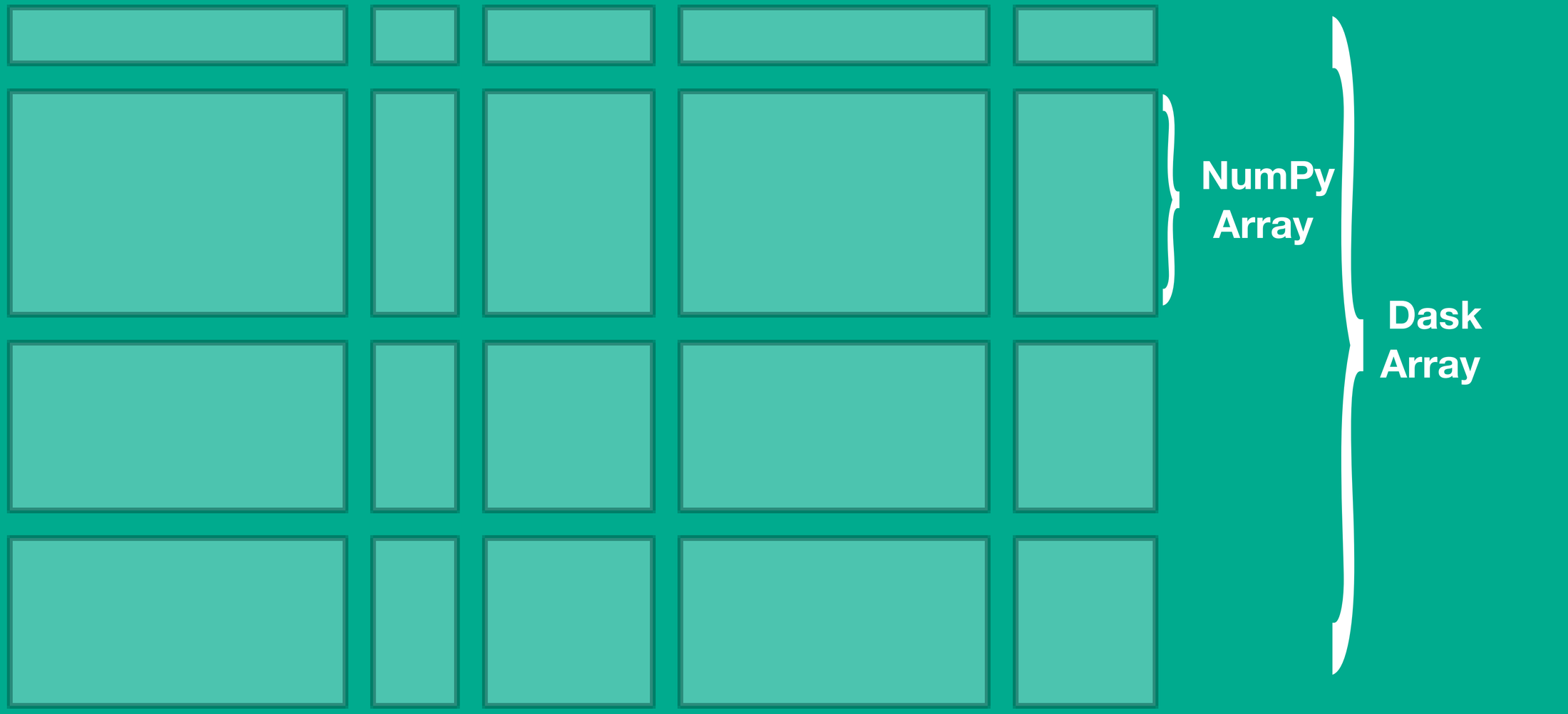
Partitioning of spatial data

Martin Fleischmann
@martinfleis

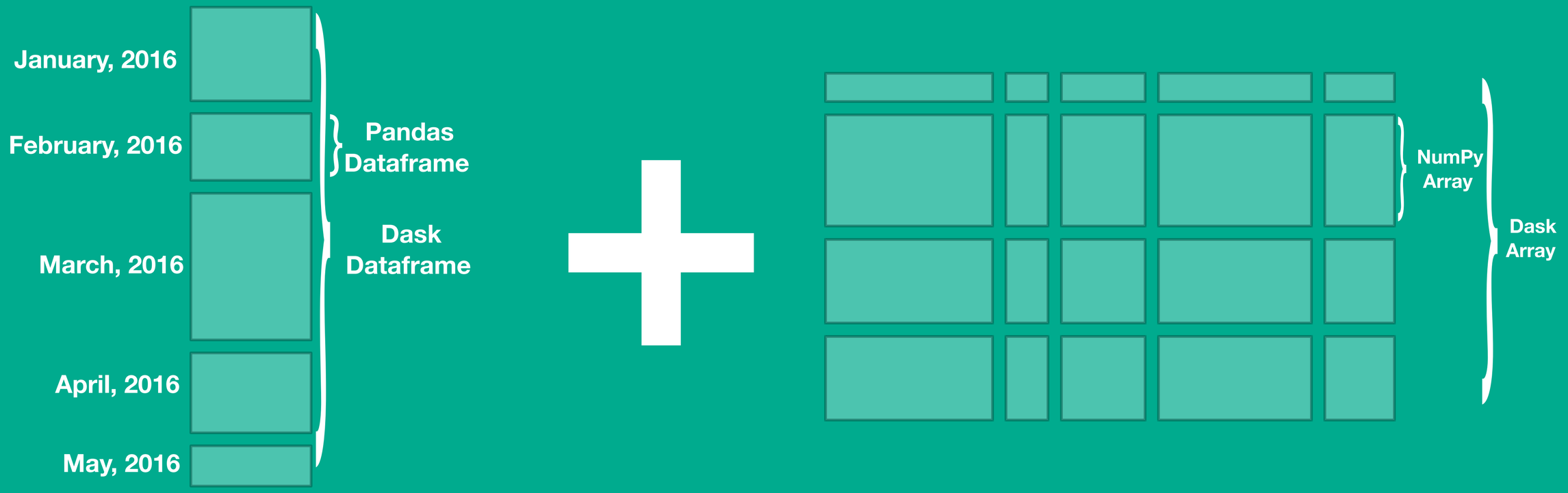
How Dask chunks data?



dask.dataframe



dask.array



spatial dataframe

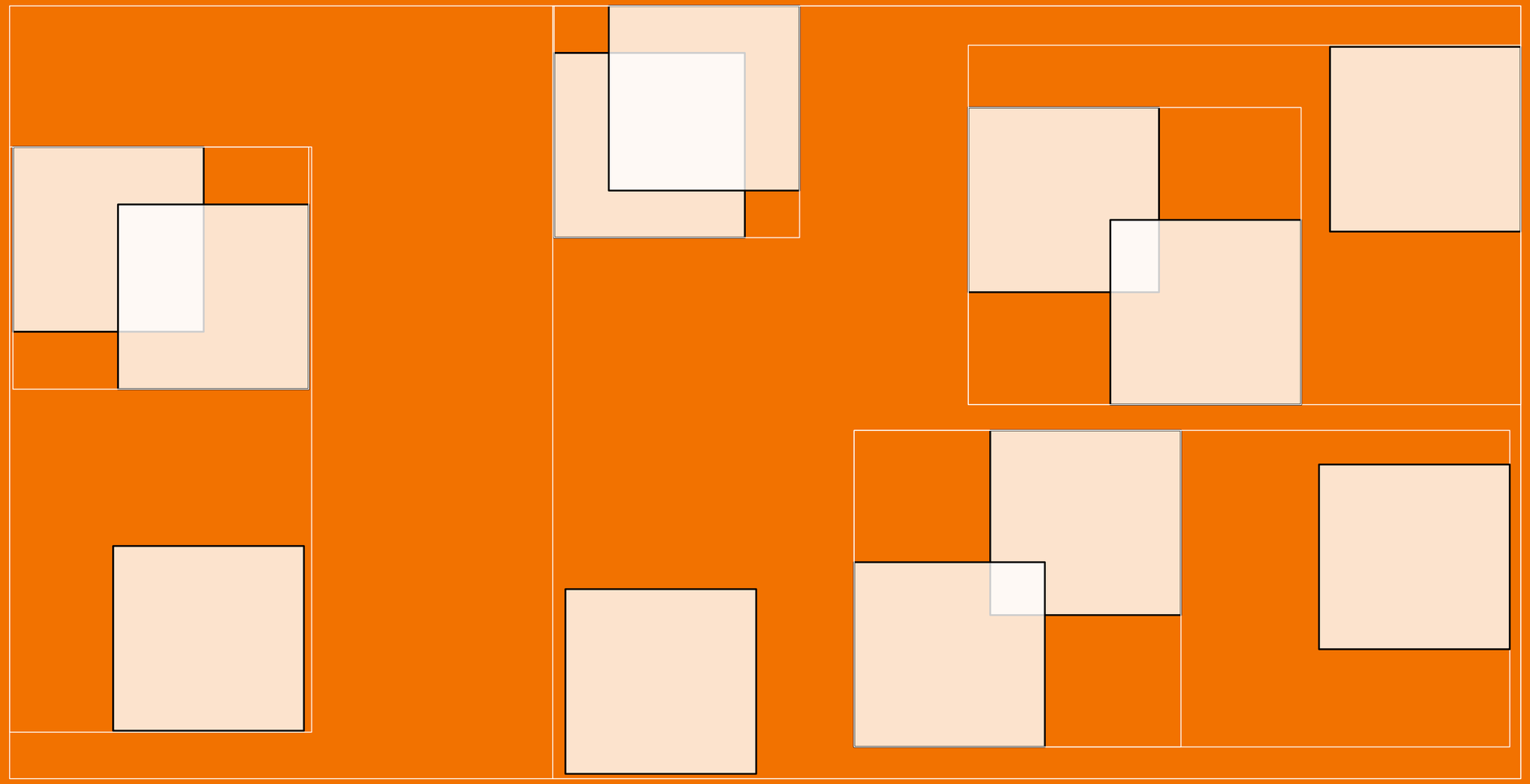
one-dimensional data structure

+

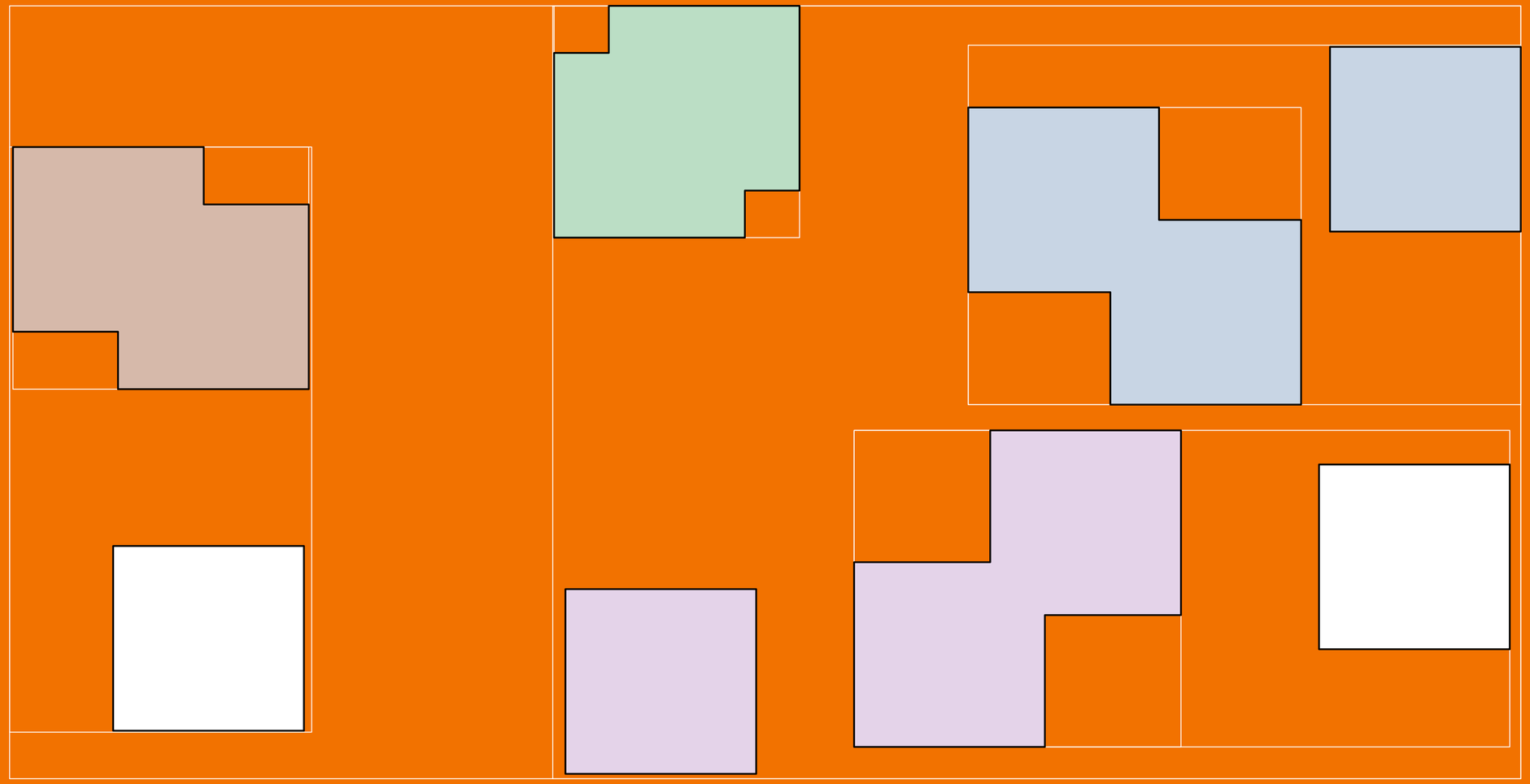
two-dimensional partitioning

spatial dataframe

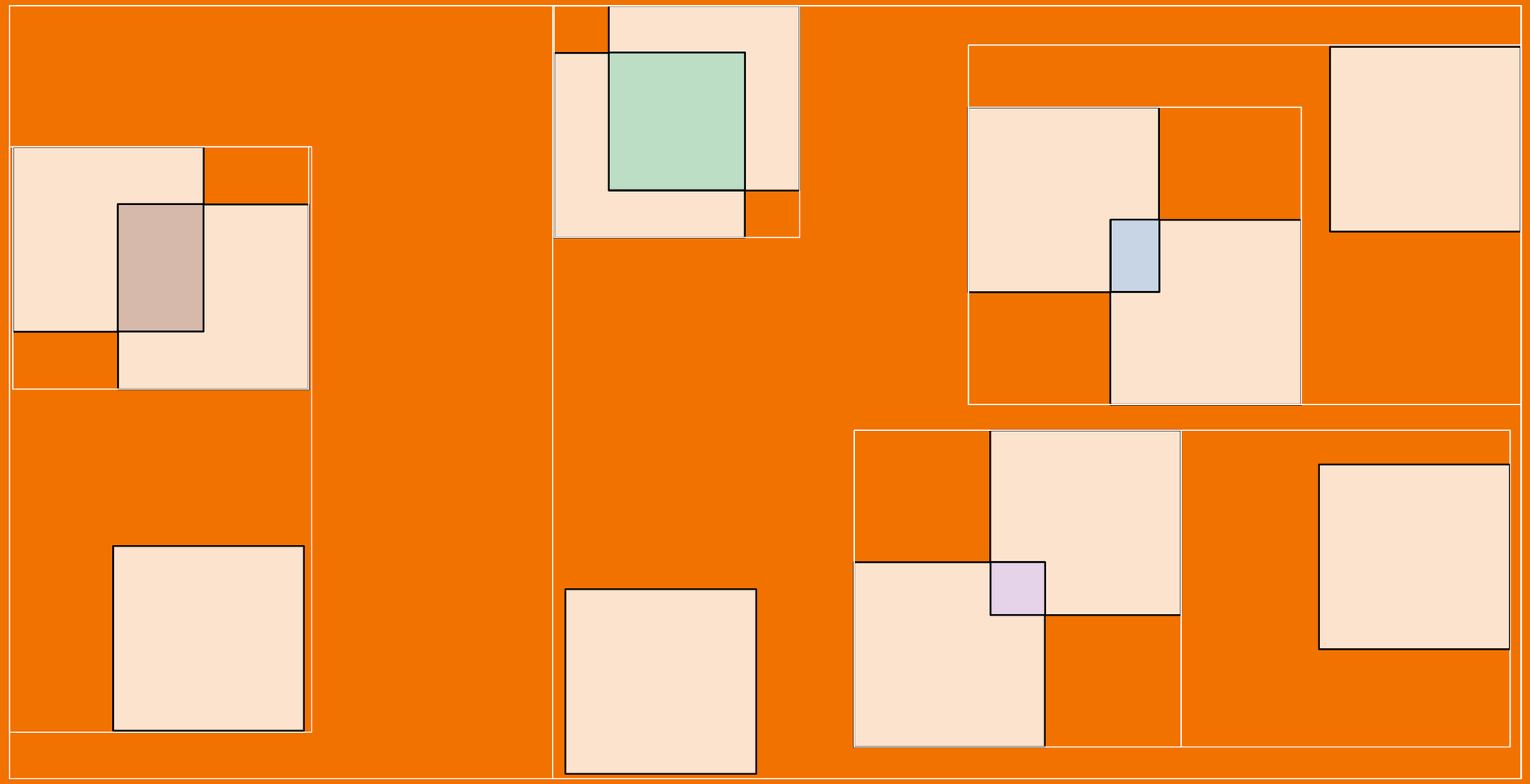
Why we need spatially coherent chunks?



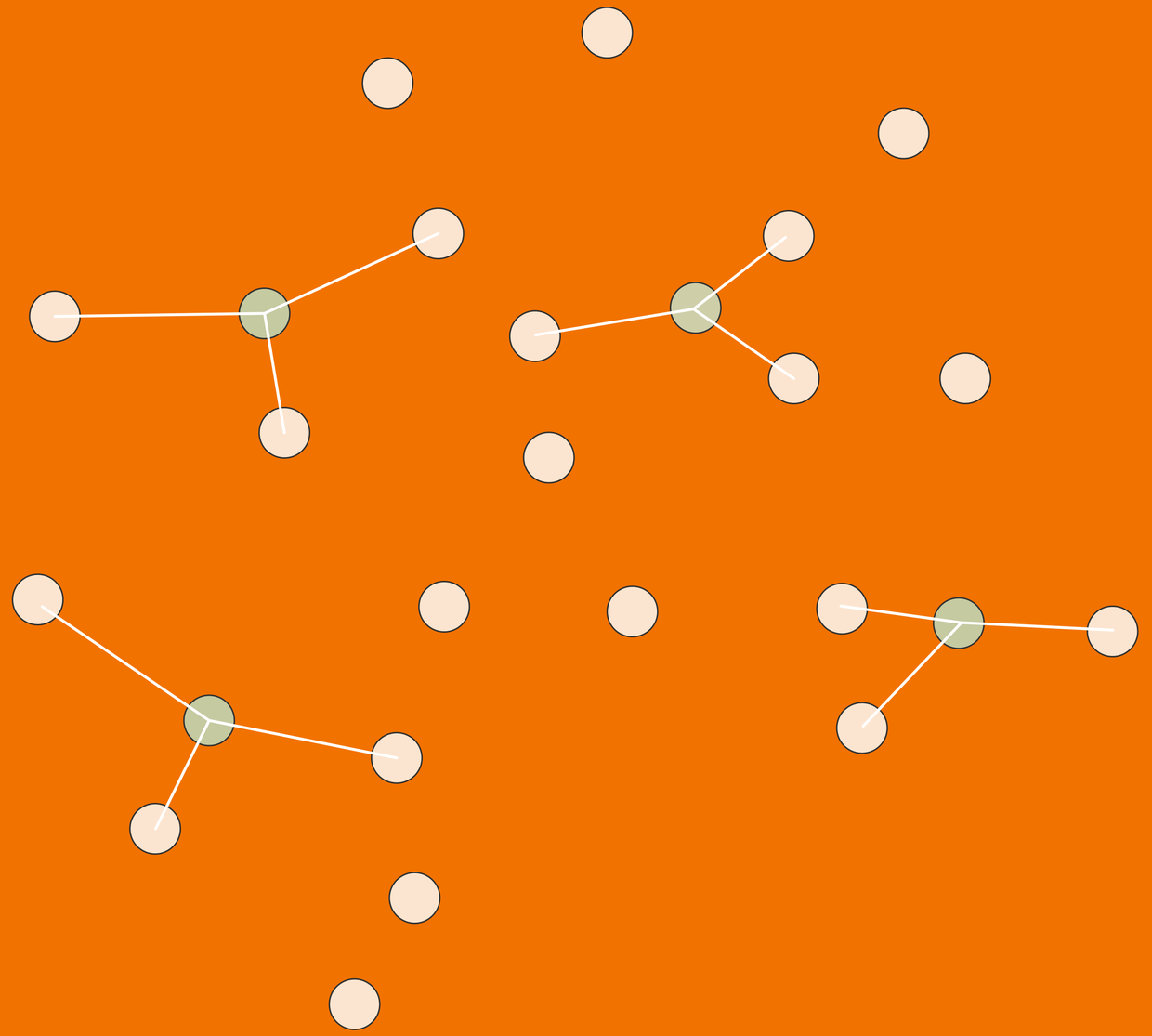
spatial indexing



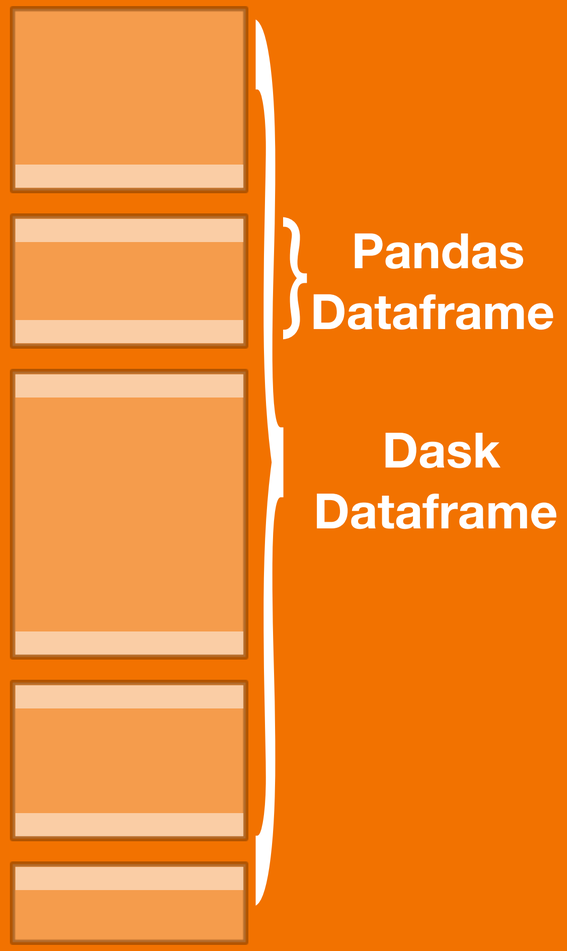
spatial operations

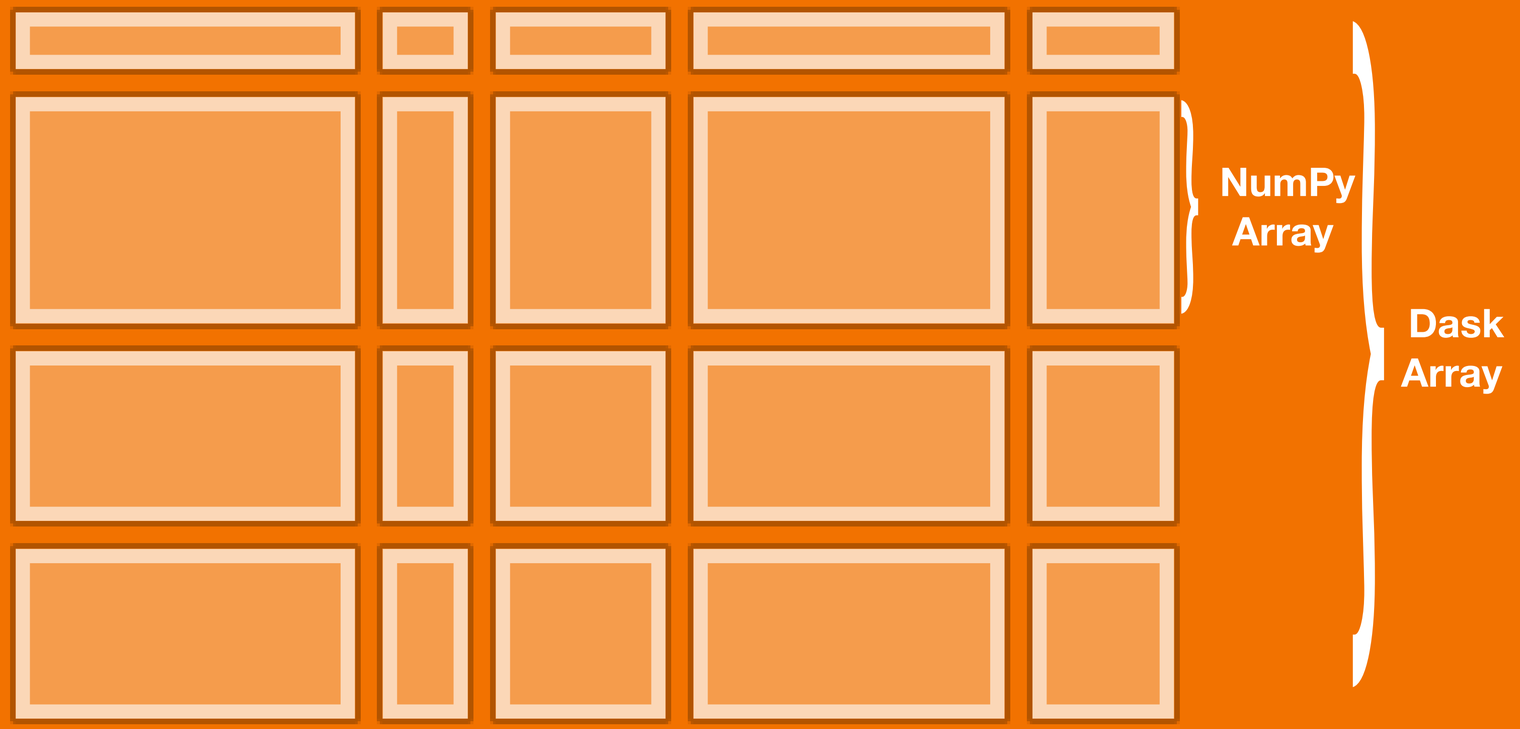


spatial operations



spatial proximity





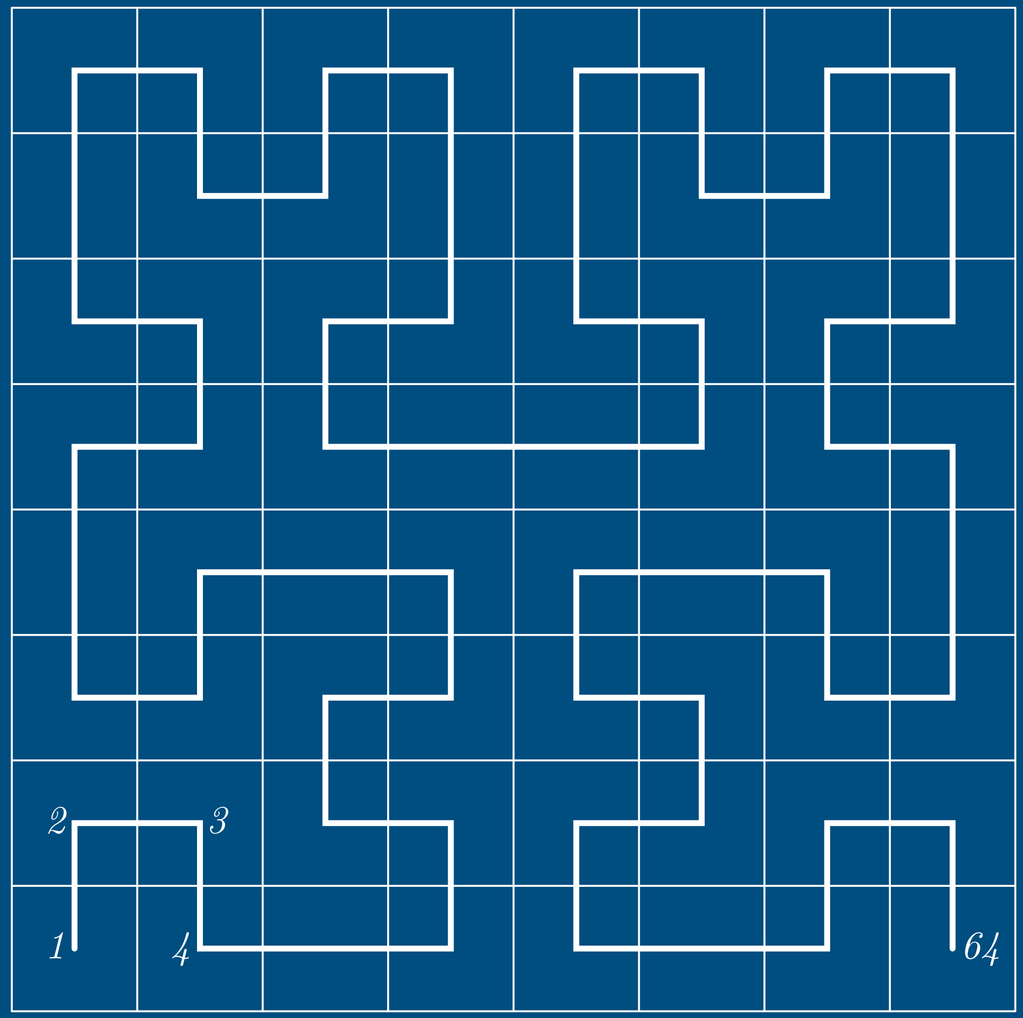
overlapping computation

distance threshold

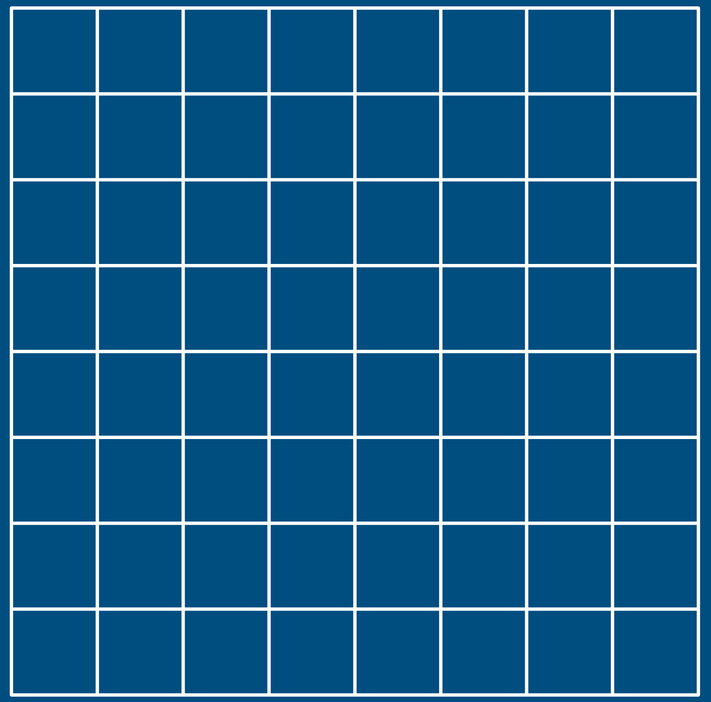
topological threshold

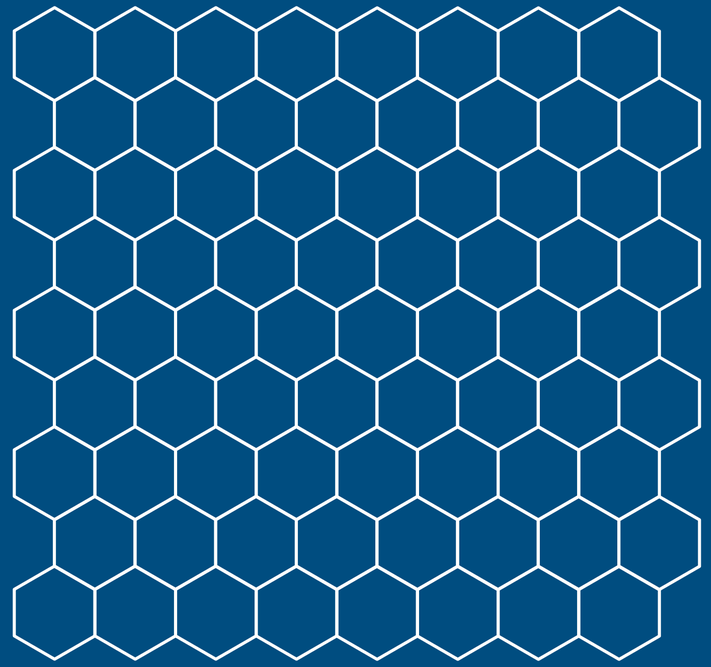
overlapping computation

How can we achieve that?

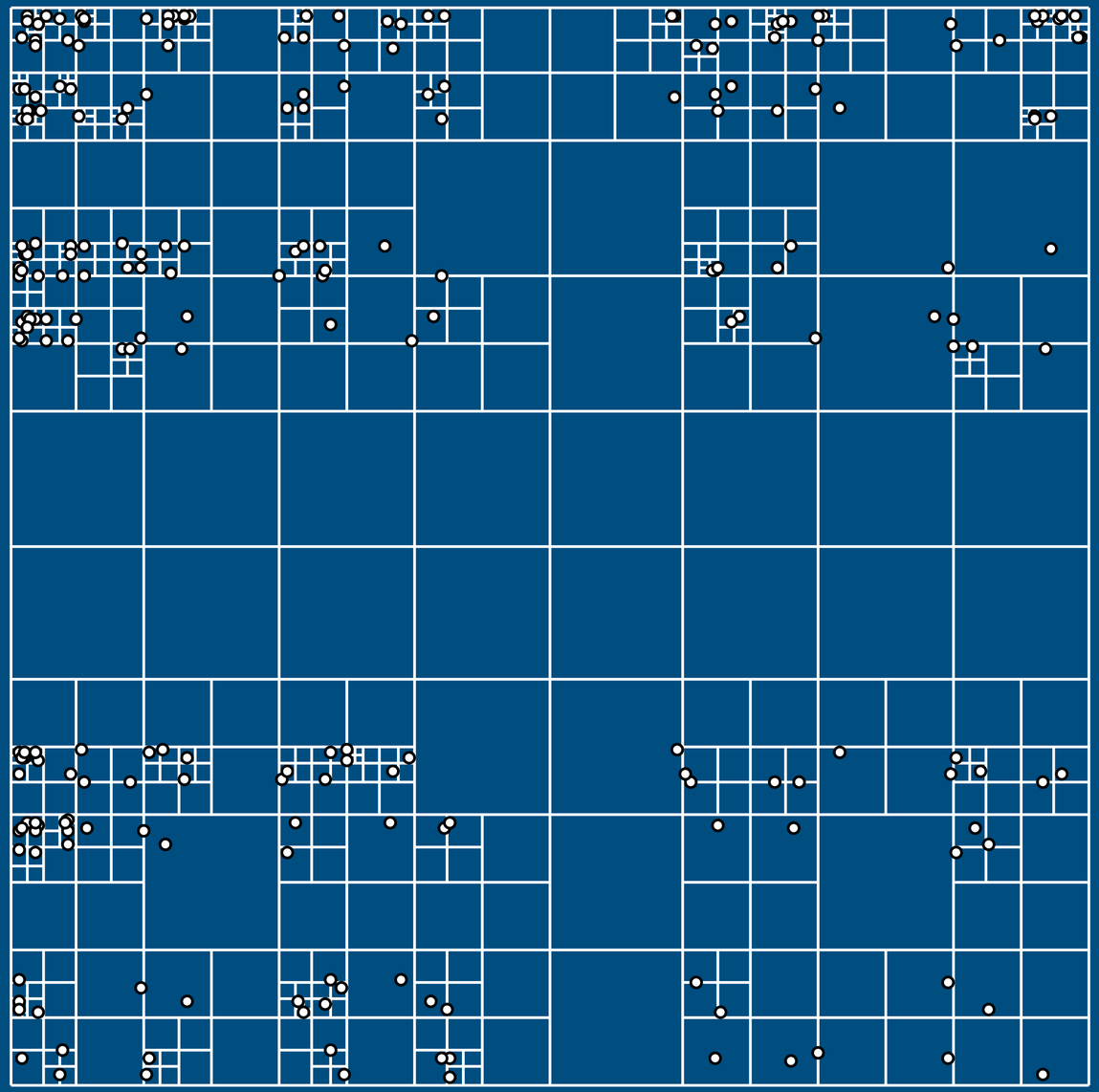


Hilbert curve

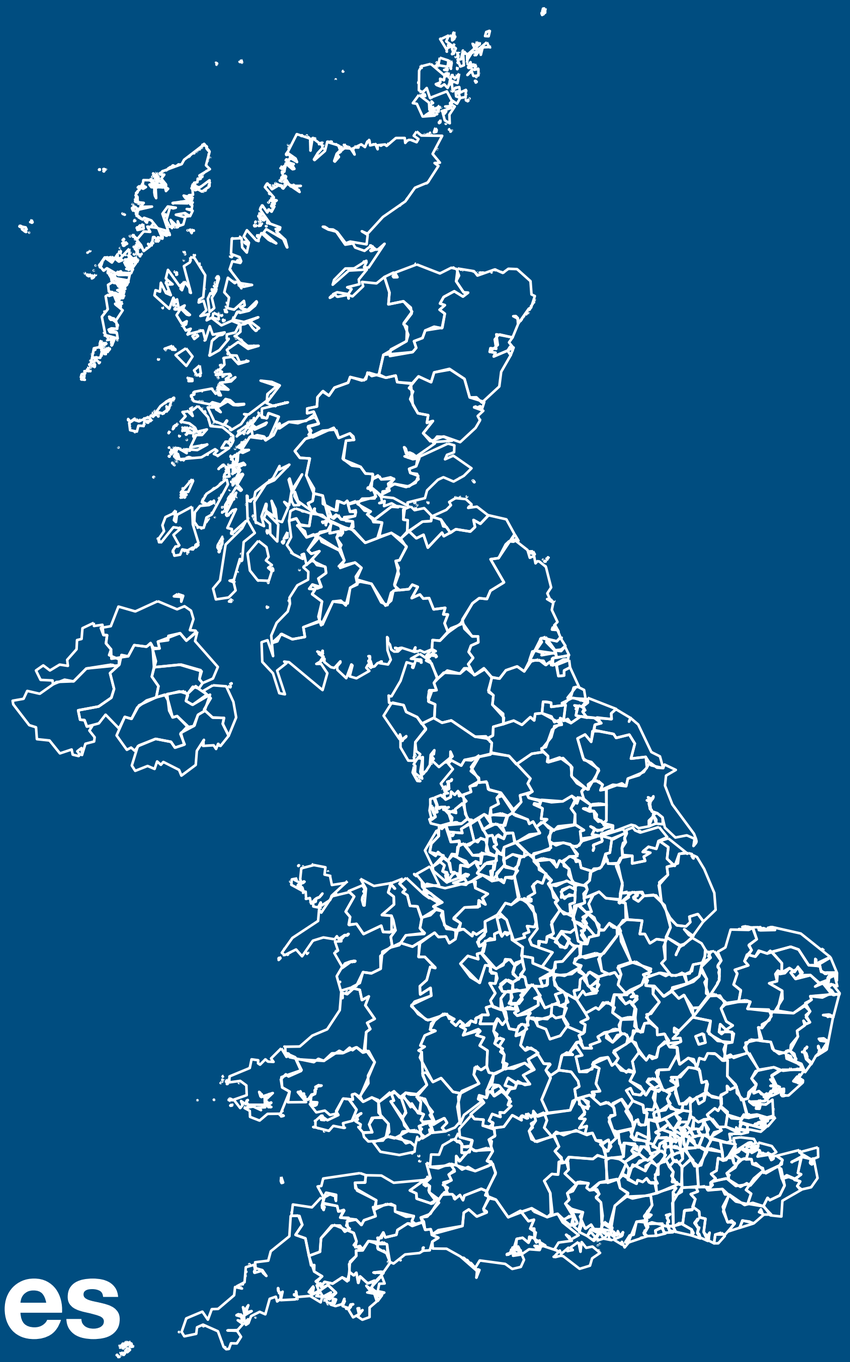




(arbitrary) grid



quadtree



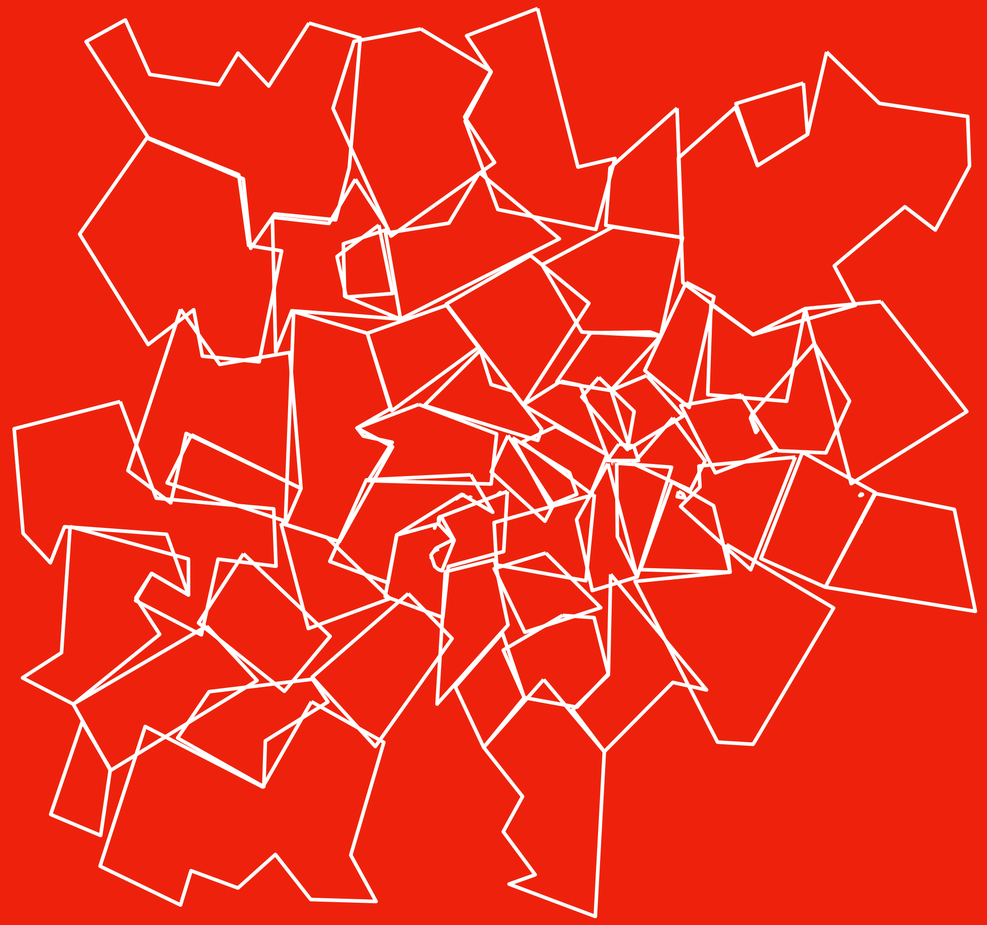
known boundaries

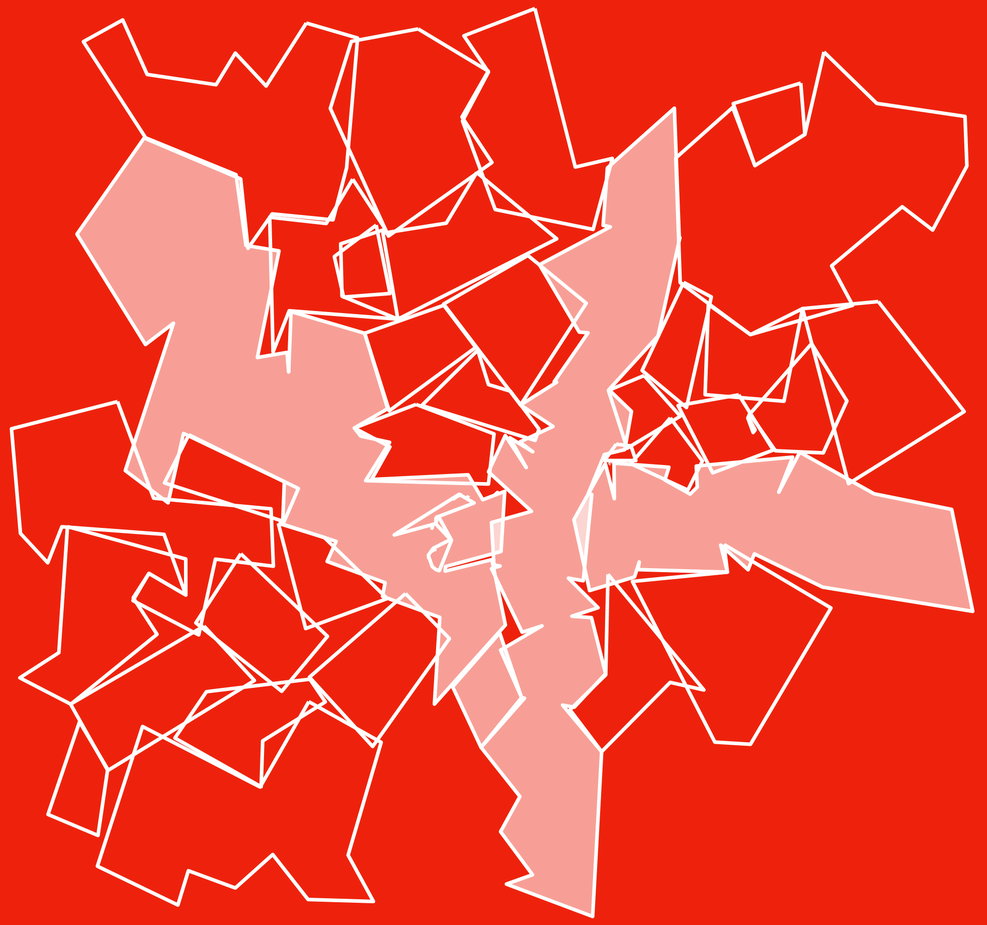
When it gets tricky

vector data are unpredictable

geometry types

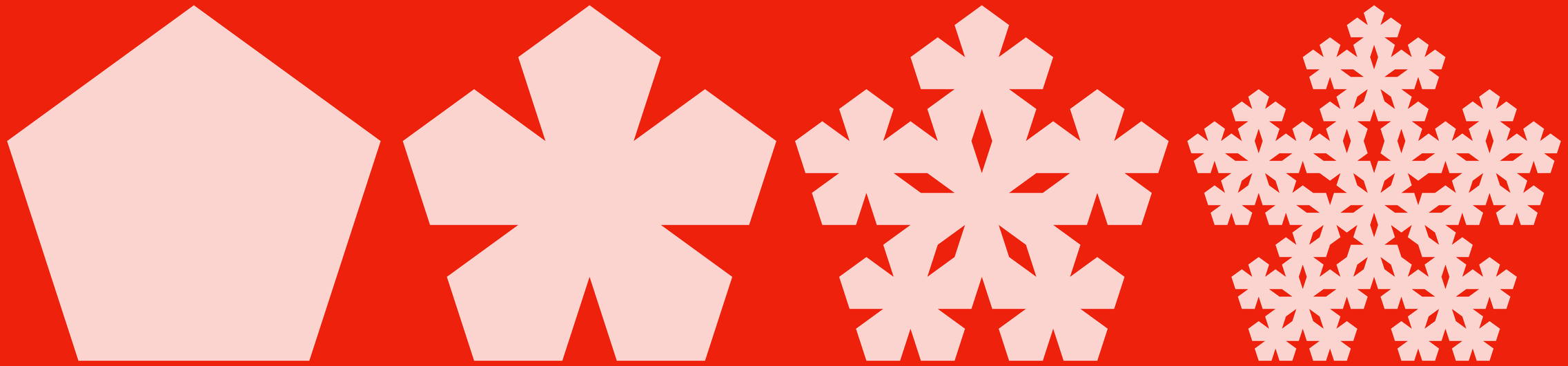
geometry types





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 spatial partitions?

How to store spatial partitions?