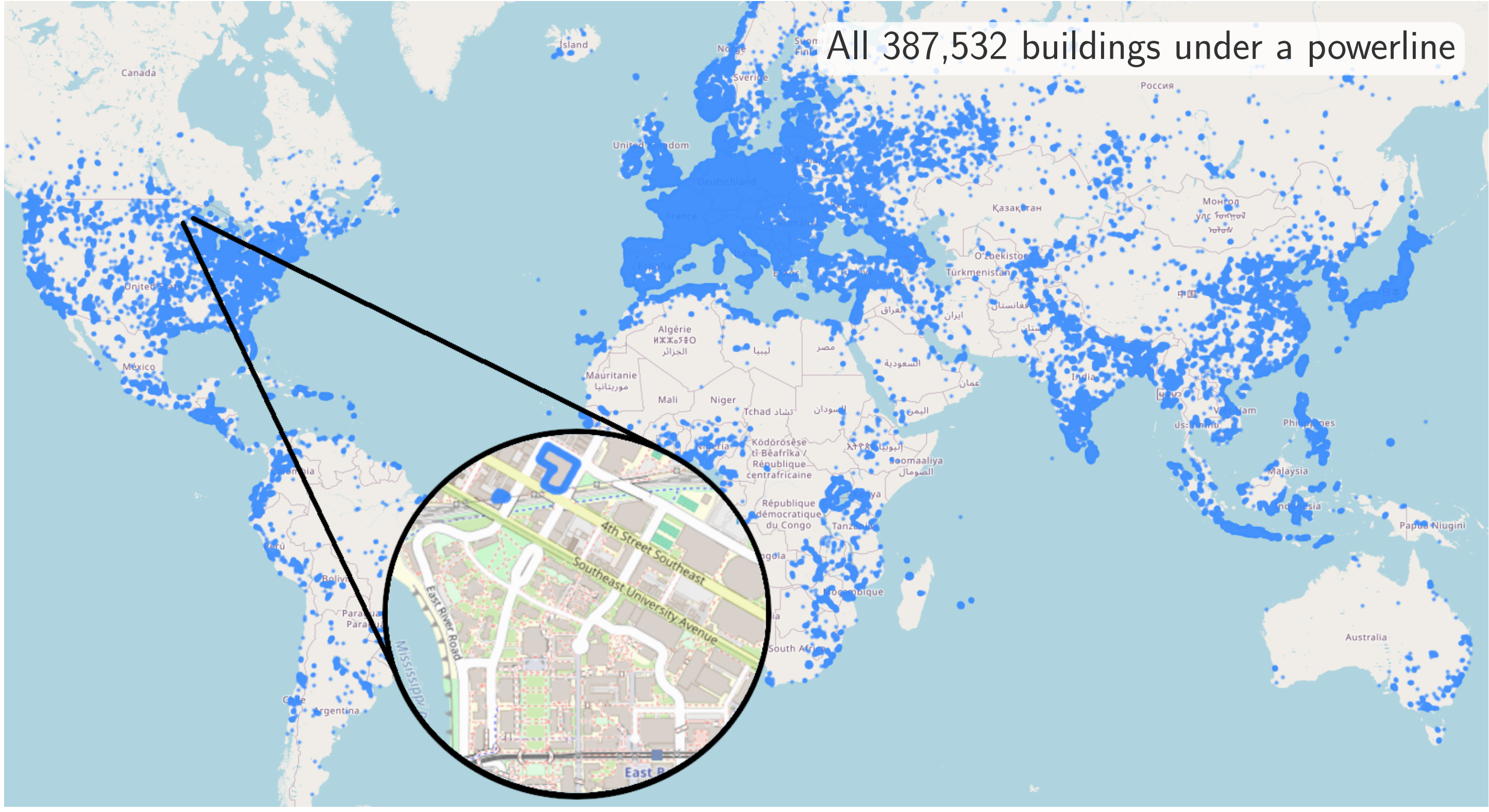


Efficient Spatial Joins on Large Geometry Sets

Hannah Bast, Patrick Brosi, Johannes Kalmbach
University of Freiburg



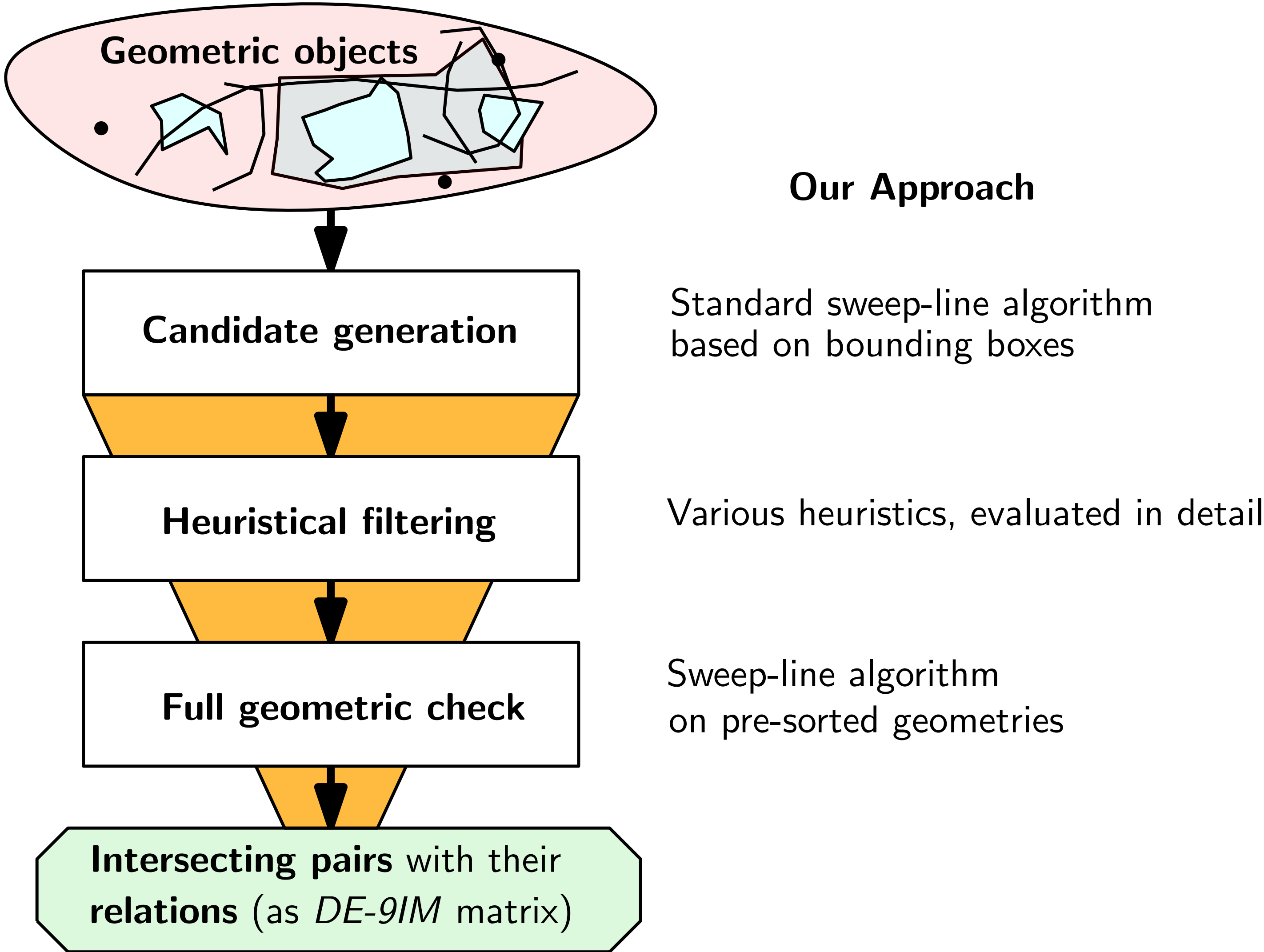
Spatial Joins



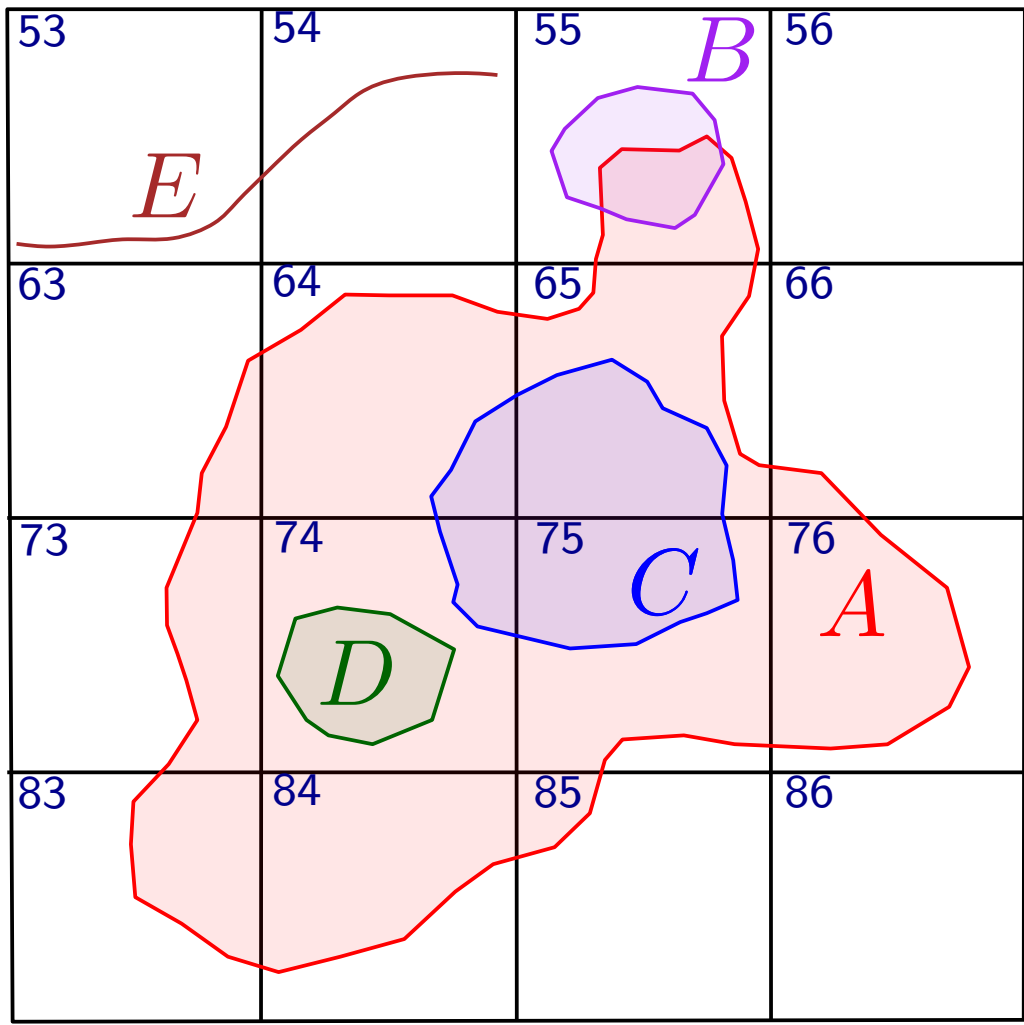
- Q1 All streets in Germany
- Q2 All buildings under a powerline
- Q3 All residential street crossings
- Q4 All residential streets per region

Core GIS
functionality

Typical Approach



Heuristic: Cell IDs **C**



- A [-55, -63, -64, -65, -66, -73, +74, -75, -76, -83, -84, -85]
- B [-55]
- C [-64, -65, -74, -75]
- D [-74]
- E [-53, -54]

- Plane-covering grid
- Cells identified by an ID
- Precompute sorted list of cell IDs for each geometry
- Positive ID: cell is contained, Negative ID: only intersects cell
- Base (some) geometric relations on result of cell ID list intersection

Examples

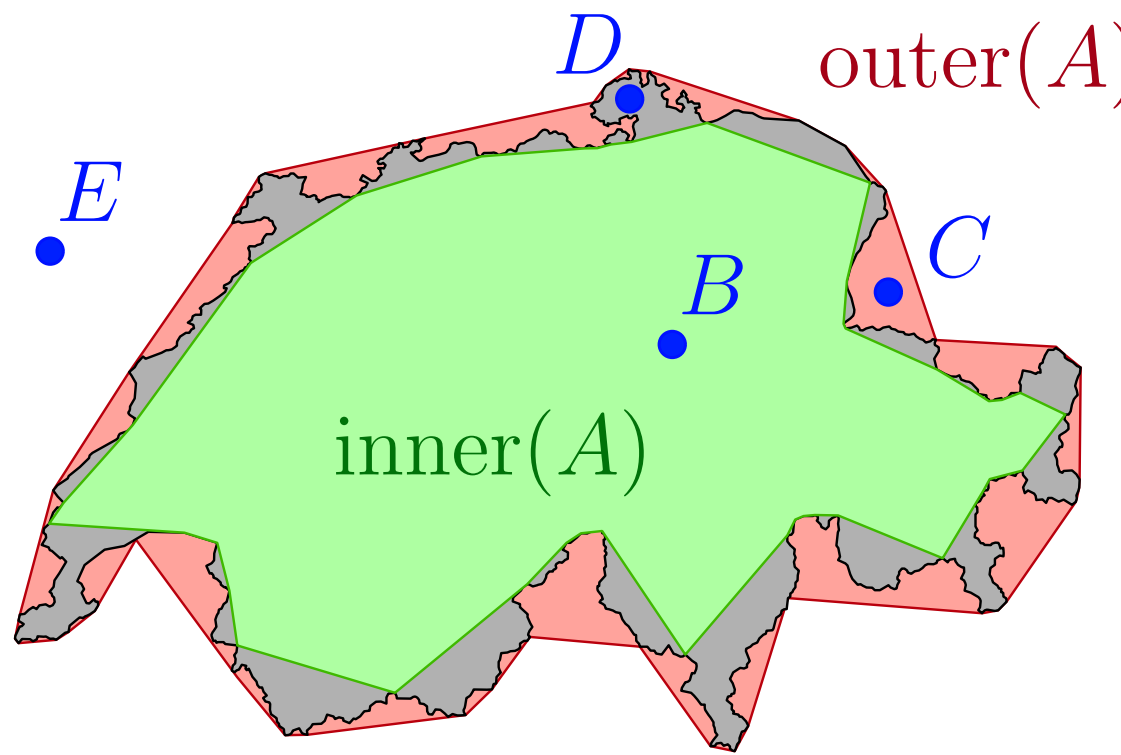
E and A are disjoint (no cell ID shared)
A contains D, because shared cell +74 (the only cell of D) is contained in A.

Heuristics: Approximate Geometries **I D O**

Goal

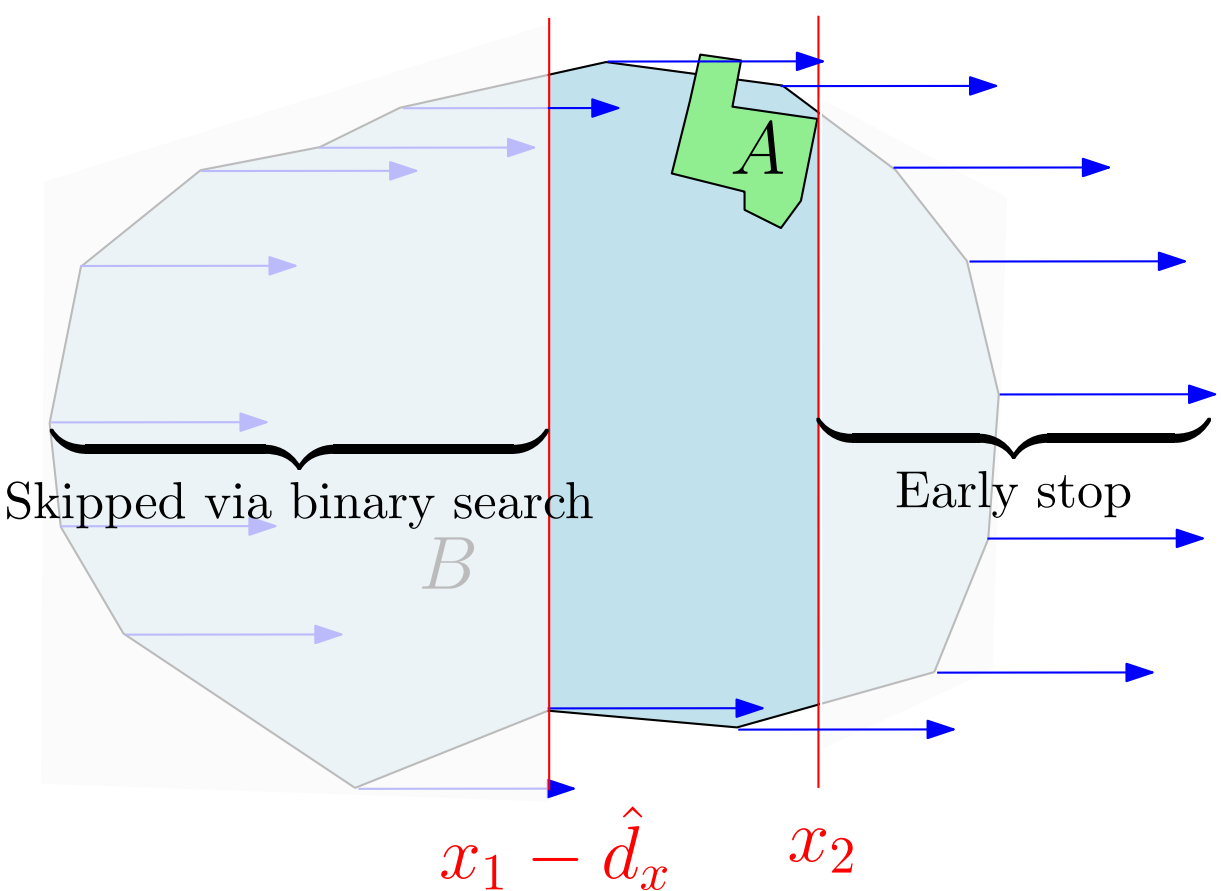
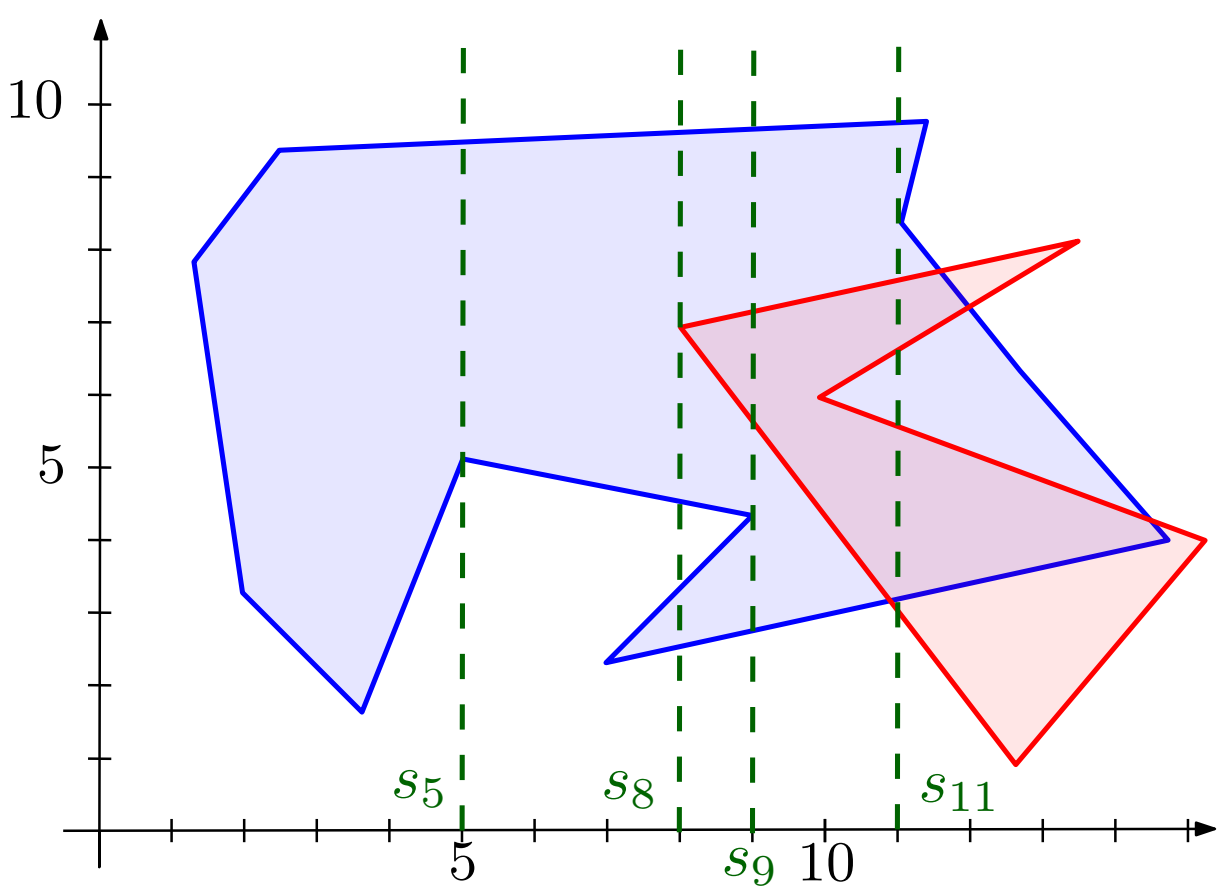
Based on simplified geometries, quickly decide **contained** or **disjoint**.

- Simplified **inner** and **outer** geometries **I**
- Diagonal bounding box **D**
- Oriented bounding box **O**



Additional heuristic **S**: precompute polygon surface areas to quickly decide that a polygon cannot be contained in another

Sweep-Line Checks on Pre-Sorted Geometries



- Full checks between (open) **polygons** based on sweep-line approach over segments, sorted by left x-coordinate
- Geometries are stored **pre-sorted**: sweep over two geometries requires no sorting or copying
- Allows for **fast skip** of irrelevant segments on the left of smaller geometry **via binary search**, and early abort on right
- Fast skip can also be used for **point/polygon tests**

Pre-sorting is crucial when heuristics are ineffective

Results (commodity PC, 16 cores)

Full self-joins

heuristics	FIN		OSM	
	presort	GEOS	presort	GEOS
	0.8 min	465.9 min	1055 min	>30 h
C	0.7 min	97.4 min	181 min	>30 h
C S	0.7 min	97.8 min	188 min	>30 h
C S D	0.6 min	98.3 min	165 min	>30 h
C S O	0.7 min	97.5 min	186 min	>30 h
C S I	0.7 min	97.6 min	176 min	>30 h

Non-self joins Q1-Q4, with best performing heuristics

	#results	#cands	#exact	presort	GEOS
Q1	2.0 M	2.5 M	10 k	1.1 min	1.2 min
Q2	1.4 M	73 M	12 M	1.2 min	1.2 min
Q3	84 M	100 M	100 M	1.2 min	1.4 min
Q4	298 M	417 M	34 M	2.7 min	6.4 min