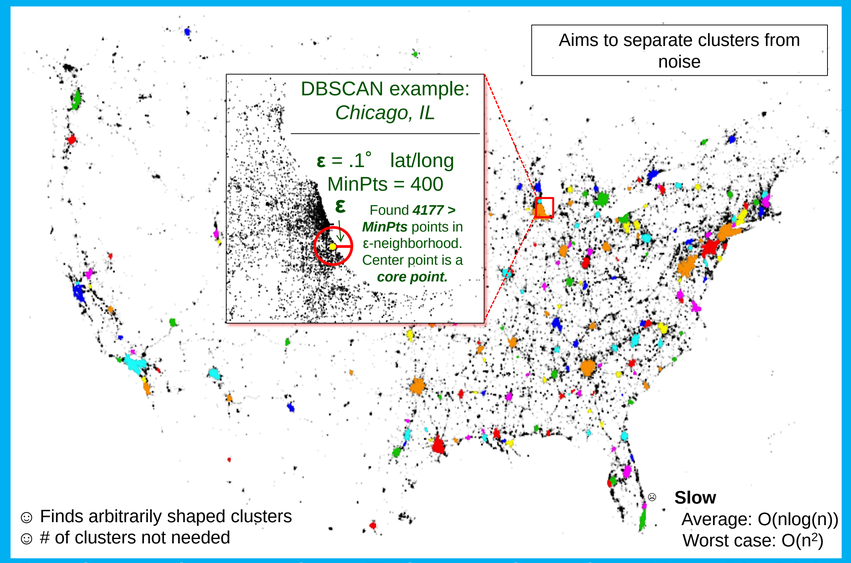
Mr. Scan

Density-Based Clustering for Extreme Scale Systems

ParaDyn™

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

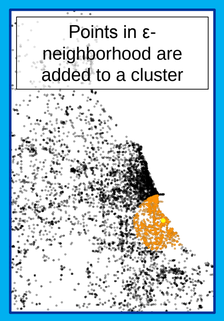
DBSCAN

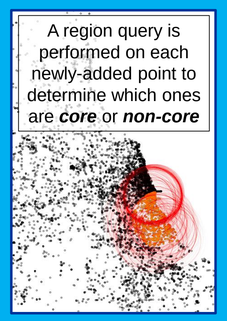


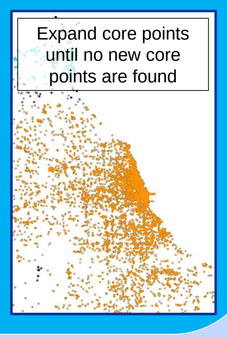
Geo-located tweets from Twitter gathered on Nov. 10, 2011

DBSCAN Scaling

- DBSCAN's serial performance limits feasible calculations to tens of millions of points
- Need parallelism to cluster datasets with billions of points



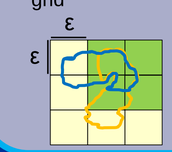


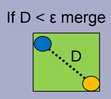


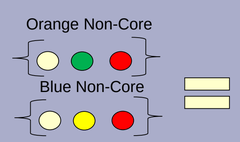
Mr. Scan Execution

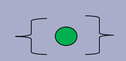
3. Merge Overlapping Clusters (CP)

1. Detect possible cluster overlap with grid
2. Check for representative point overlap
3. Calculate the diff of the non-core points
4. Calc distance from representative points
   Merge if at least ONE's distance from ● < ε



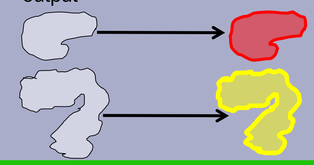






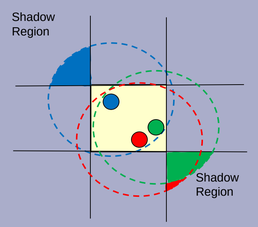
4. Color the Clusters (FE)

- Color each cluster
- Send coloring to BE for cluster output



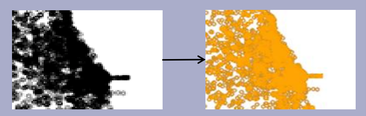
2. Pick Representative Points (BE)

- Representative points are a small finite set of points that represent all core points in a ε x ε box
- Used to detect cluster overlap
- A core point that would cause a merge has to fall within ε of a representative point



Representative points are the set of points which represent the furthest extension into each shadow region (Max of 2 per shadow region)

1. Run DBSCAN (BE)



- Run DBSCAN to classify points into clusters on local partition

0. Prep Input Data

- Divide data into ε x ε boxes
- Form roughly equal partitions from boxes
- Outline each partition with "shadow area"
- Target # points per BE:
  800K – GPU DBSCAN
  50K – CPU DBSCAN