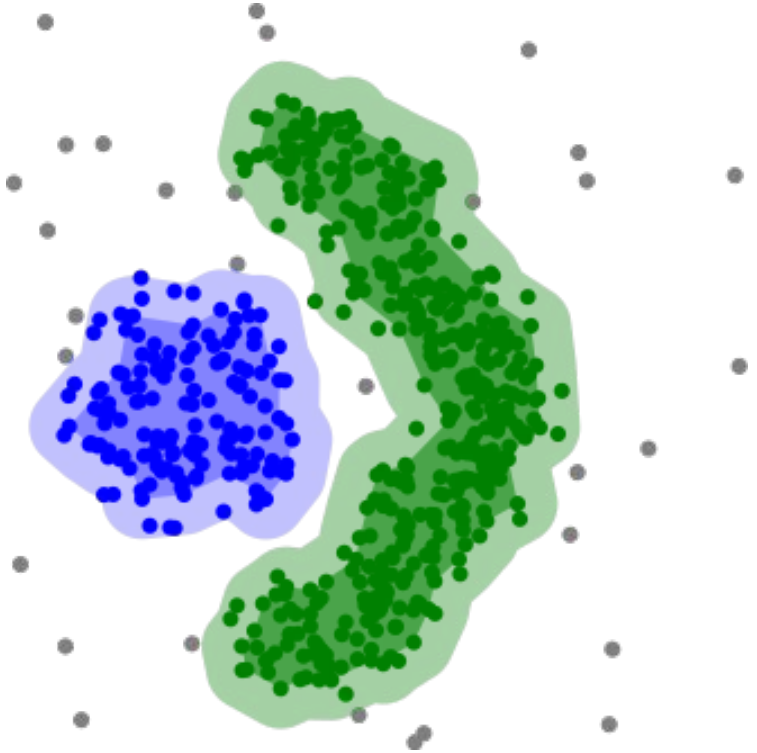
DBSCAN

Density-Based Spatial Clustering of Applications with Noise

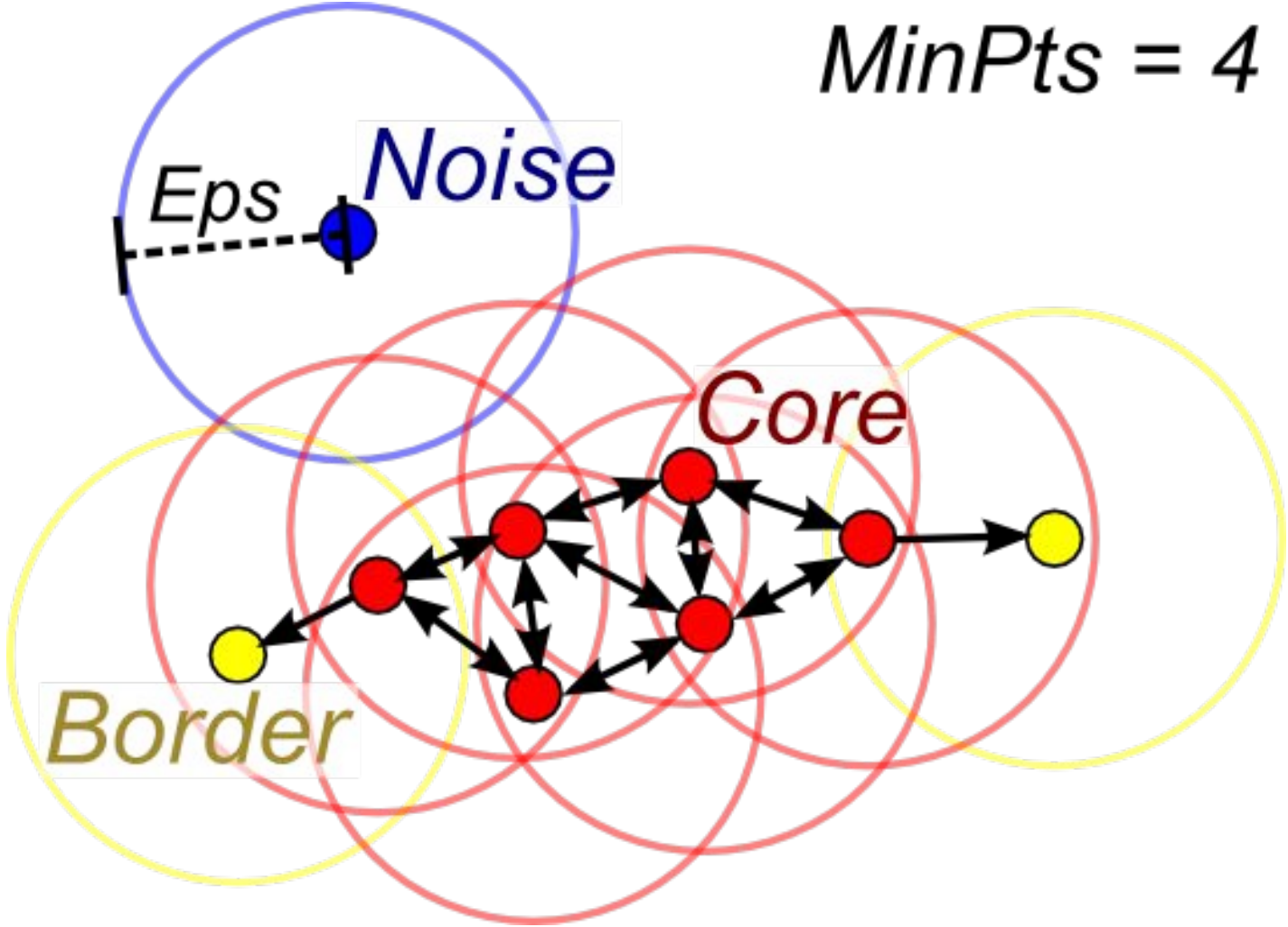
DBSCAN is a data clustering algorithm

- density-based clustering
- non-parametric
- nondeterministic



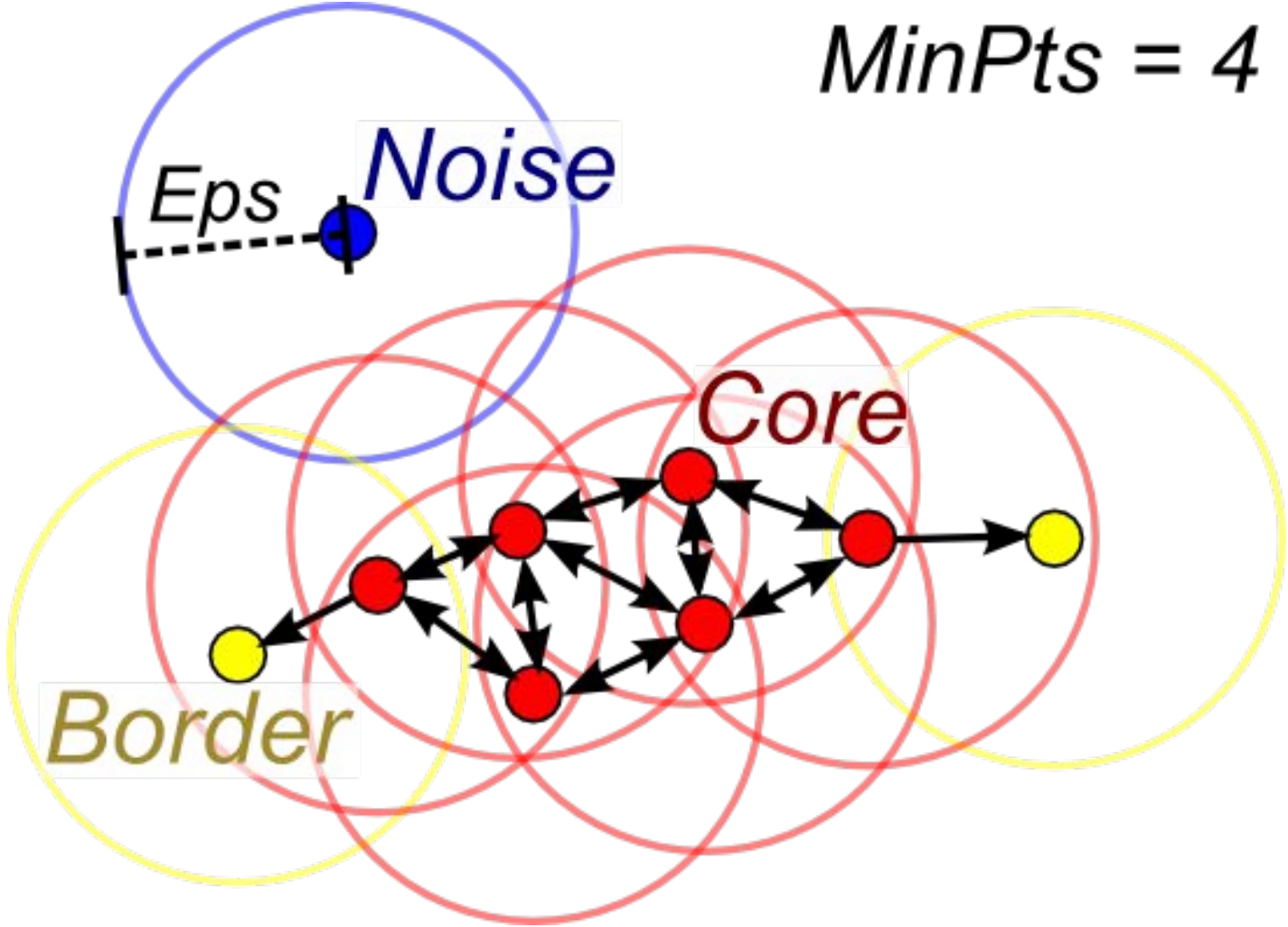
Core points

- A point p is a ***C****ore point* if at least **minPts** points are within distance **ε** of it (including p).


MinPts = 4
Noise
Eps
Core
Border

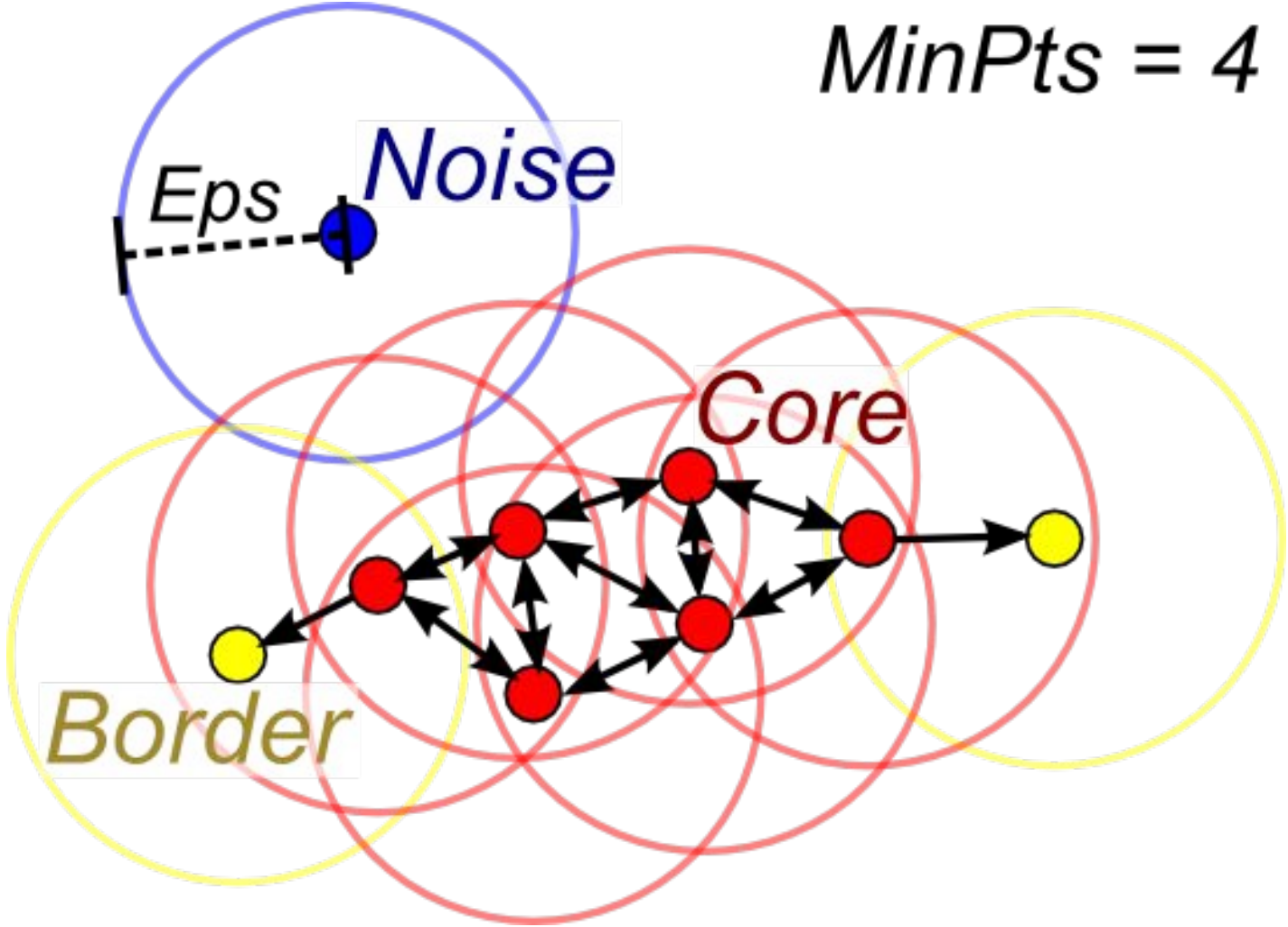
Border points

- A point p is a Border point if :
  - it is in reach of a **C**ore point and
  - less than **minPts** points are within distance **ε** of it (including p).


MinPts = 4
Noise
Eps
Core
Border

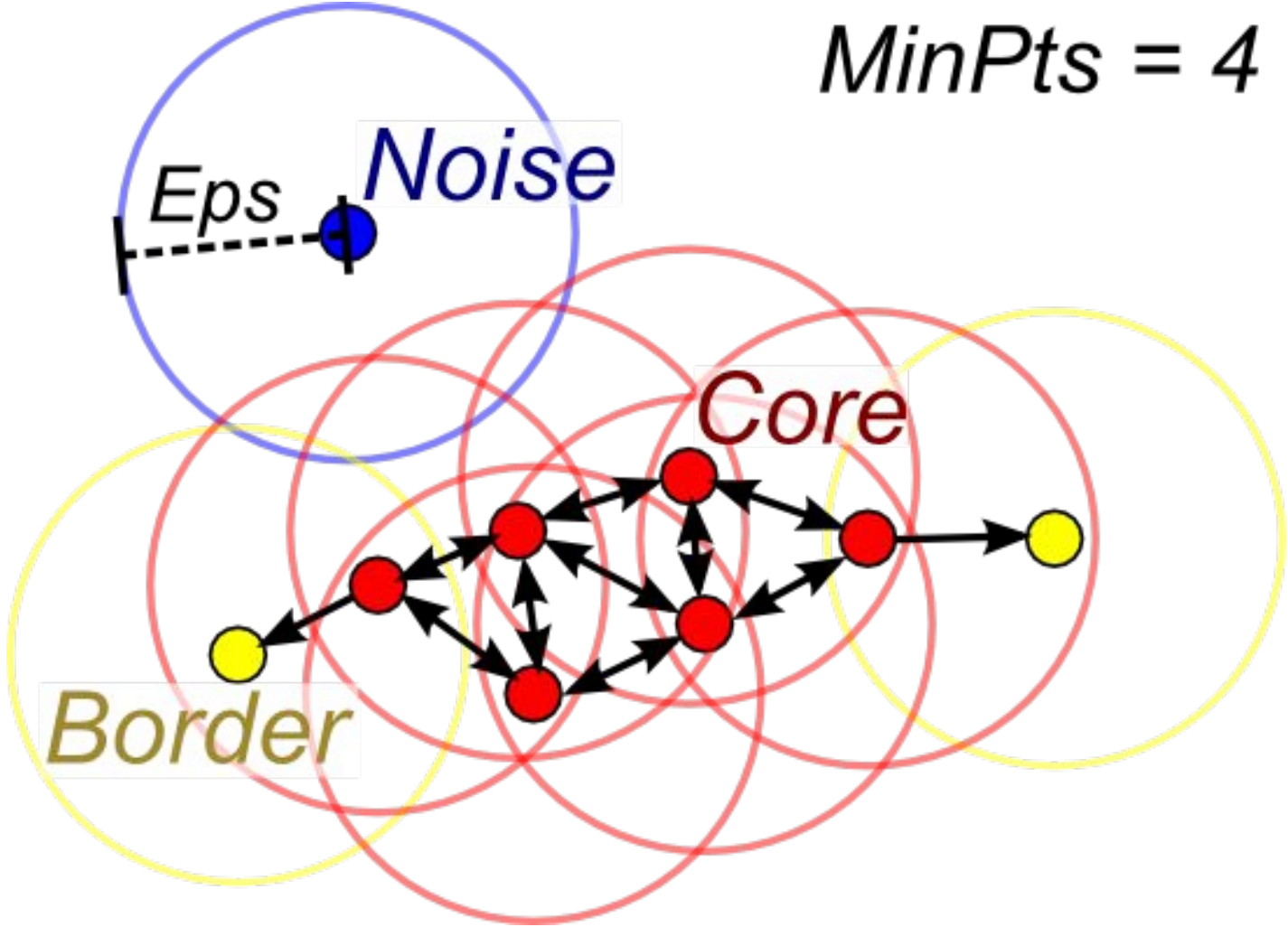
Reach

- A point q is *directly reachable* from p if point q is within distance ε from core point p. Points are only said to be directly reachable from **C**ore points.
- A point q is *reachable* from p if there is a path p1, ..., pn with p1 = p and pn = q, where each pi+1 is *directly reachable* from pi.


MinPts = 4
Noise
Eps
Core
Border

Noise points

- All points not reachable from any other point are outliers or Noise points.


MinPts = 4
Noise
Eps
Core
Border

DBSCAN Algorithm

- Begins by selecting a point randomly
- Checks if the selected point is a **C**ore point
- Then, finds the connected components of all the **C**ore points
- Assign each non-**C**ore point to the nearest cluster if the cluster is its epsilon-neighbor , otherwise assign it to **N**oise.
- Stops when it explores all the points one by one and classifies them as either **C**ore, **B**order or **N**oise point.

Some Density Based Metrics

- A clustering result satisfies homogeneity if all of its clusters contain only data points which are members of a single class.
- A clustering result satisfies completeness if all the data points that are members of a given class are elements of the same cluster.

Complexity

- Average O(n logn)
- Worst case O(n*n)

Avantages

- DBSCAN algorithm is robust to outliers (**N**oise points).
- DBSCAN is great at separating high density clusters from low density clusters.
- Unlike K-means, DBSCAN does not require number of clusters to be specified priorily.
- DBSCAN supports non-globular structures as well.

Disavantages

- DBSCAN does not work well for clusters of varying density.
- DBSCAN algorithm is not deterministic in the sense that it forms different clusters on different trials.
- Sometimes, choosing the value of ‘epsilon’ can be difficult especially when the data is in higher dimensions.

Sources

- https://en.wikipedia.org/wiki/DBSCAN
- https://medium.com/@agarwalvibhor84/lets-cluster-data-points-using-dbscan-278c5459bee5
- https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html
- https://scikit-learn.org/stable/auto_examples/cluster/plot_dbscan.html#sphx-glr-auto-examples-cluster-plot-dbscan-py