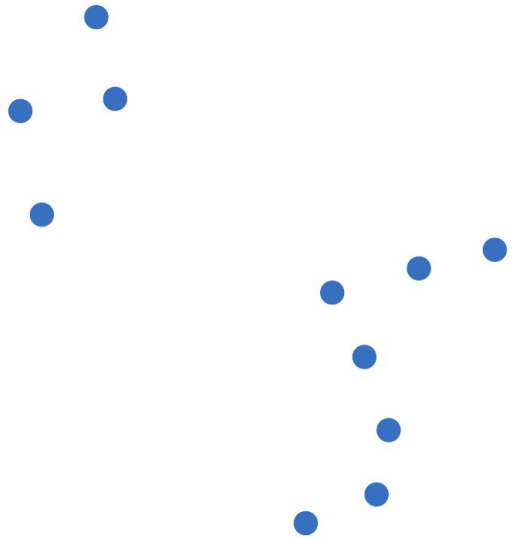
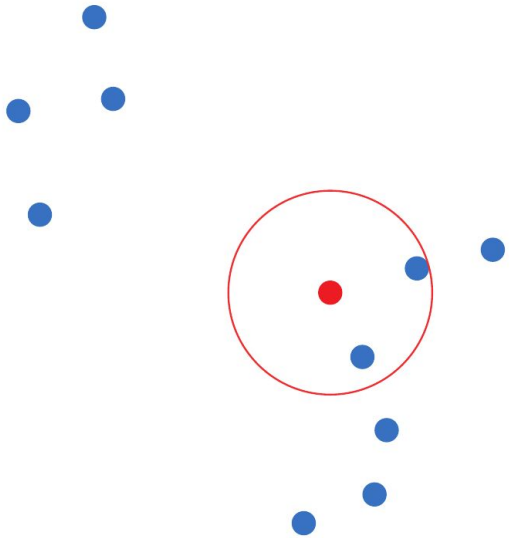
DBSCAN

DBSCAN



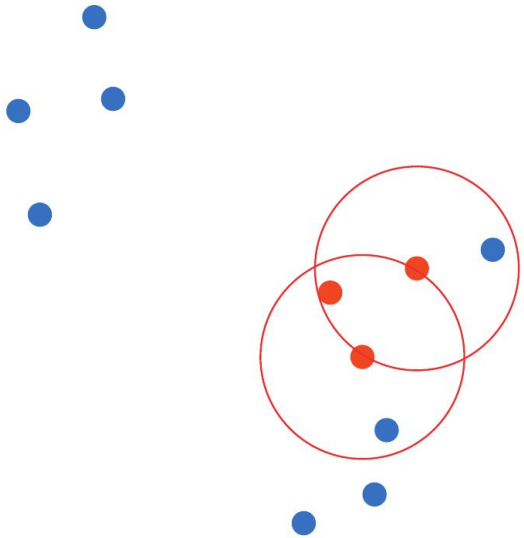
CAMBRIDGE SPARK

DBSCAN



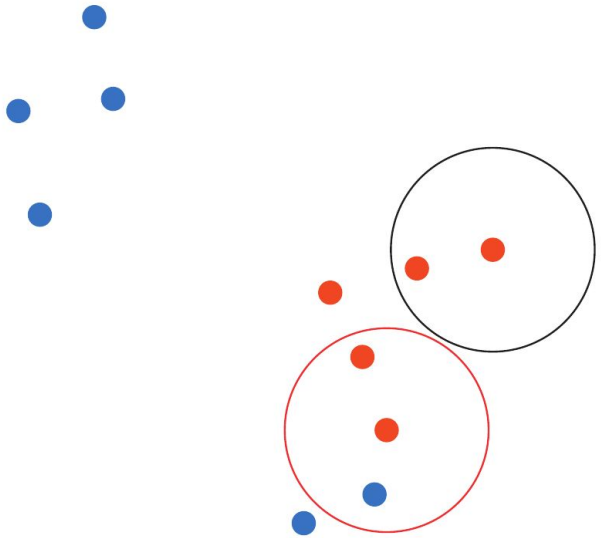
CAMBRIDGE SPARK

DBSCAN



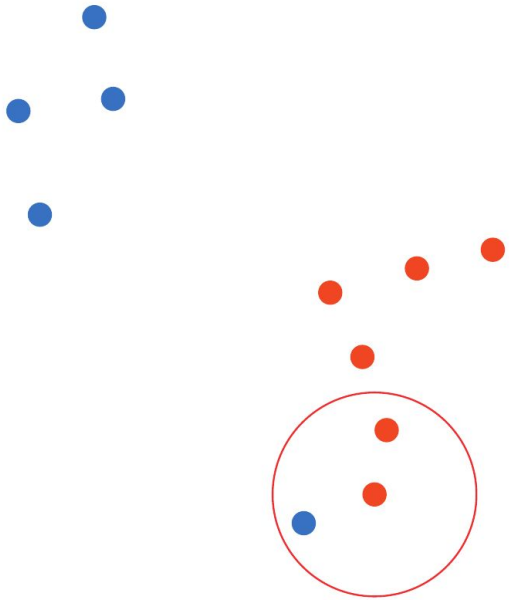
CAMBRIDGE SPARK

DBSCAN



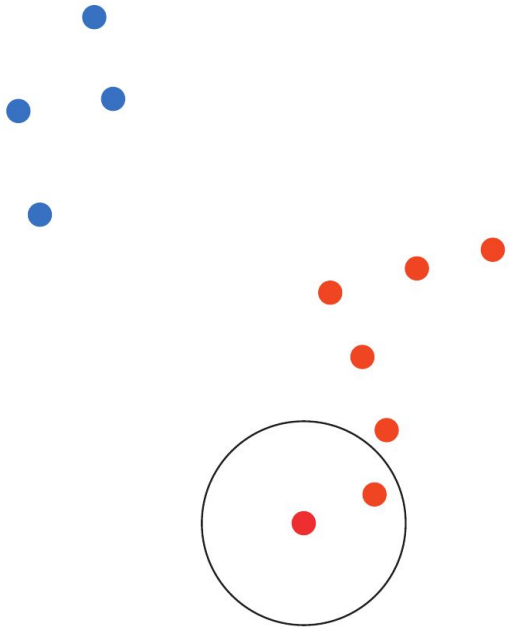
CAMBRIDGE SPARK

DBSCAN



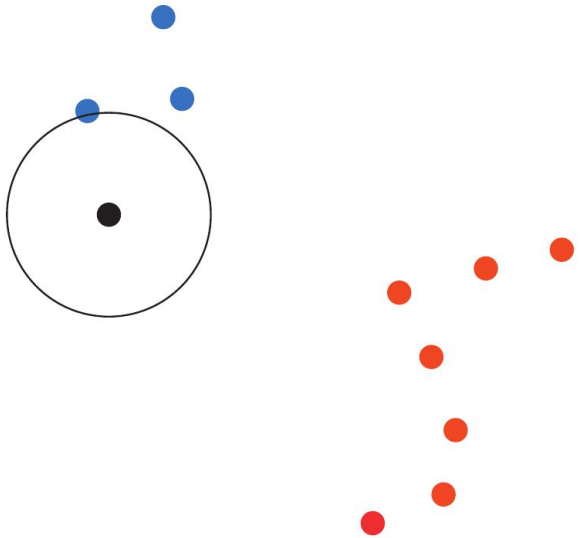
CAMBRIDGE SPARK

DBSCAN



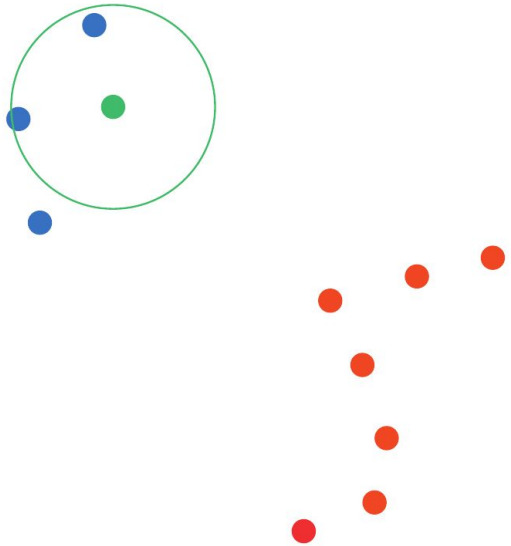
CAMBRIDGE SPARK

DBSCAN



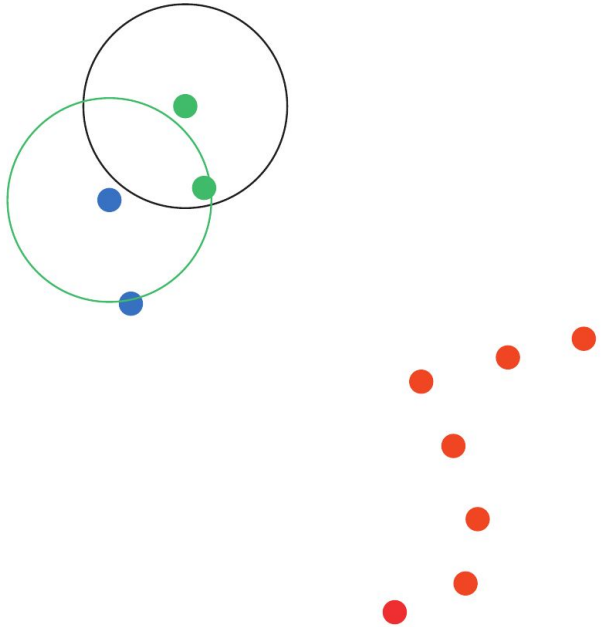
CAMBRIDGE SPARK

DBSCAN



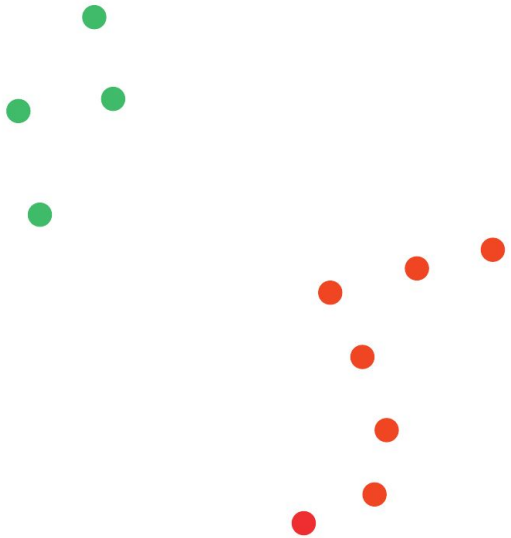
CAMBRIDGE SPARK

DBSCAN



CAMBRIDGE SPARK

DBSCAN



CAMBRIDGE SPARK

DBSCAN

- Clusters = Zones of high-density
- Two parameters: min_samples and eps

Algorithm:

- Start at a random point, consider all points within radius eps
- If that covers min_samples, keep that ball
  - Expend by considering esp-balls around every point of the current ball and iterate
- Otherwise mark the point as noise

… let’s see this in action

DBSCAN: pros and cons

Pros

- Clusters are not necessarily globular
- No choice of number of clusters
- Very efficient implementations exist
- Robust to noise

Cons

- The eps and min_samples can be hard to tune
- If clusters have significantly different densities it is hard to find a meaningful eps, min_samples

Hands-on session

dbscan.ipynb

Comparisons of clustering algorithms