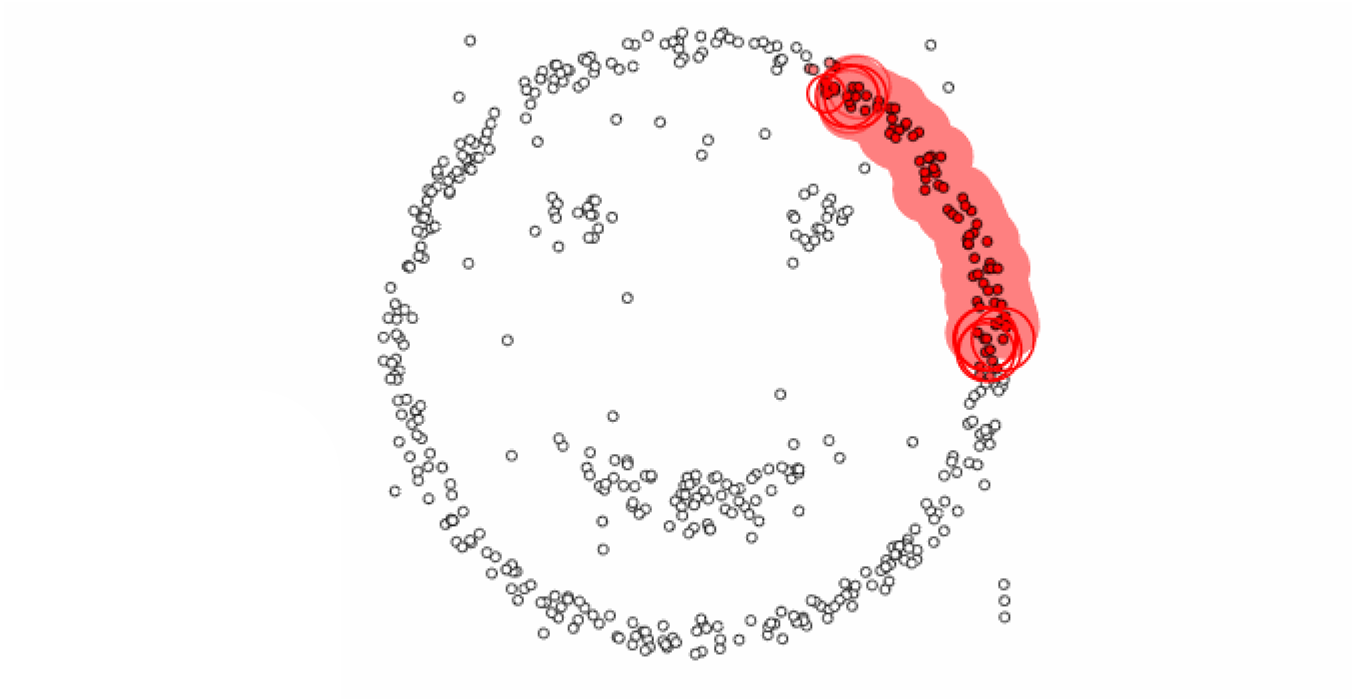
DBSCAN



A Density Based Clustering Method

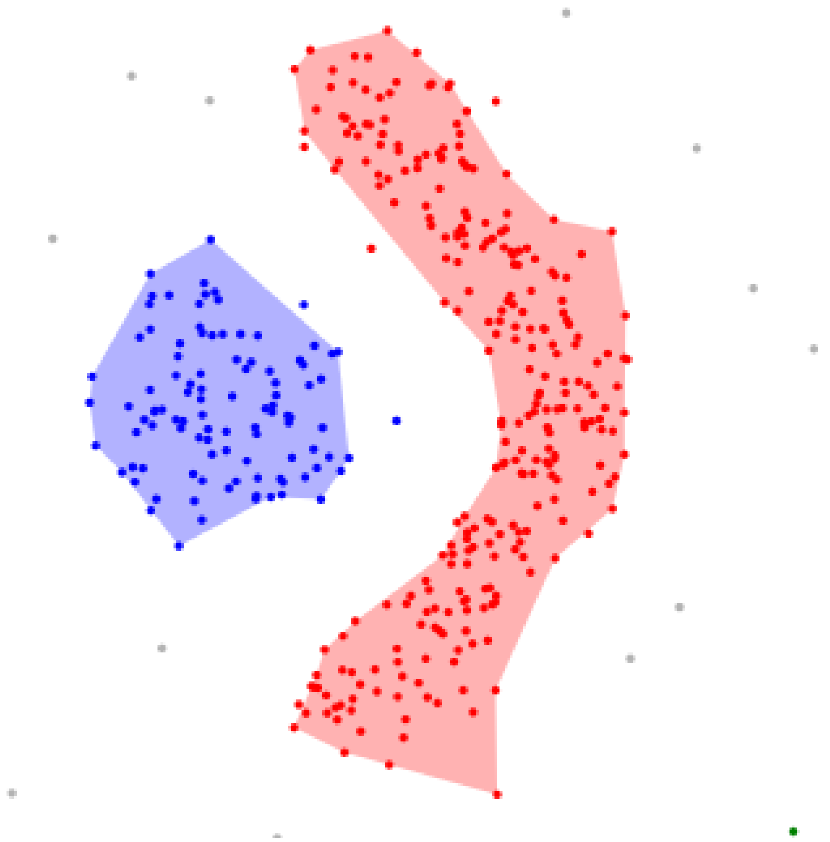
Liam Haas-Neill; November 30, 2018

https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68

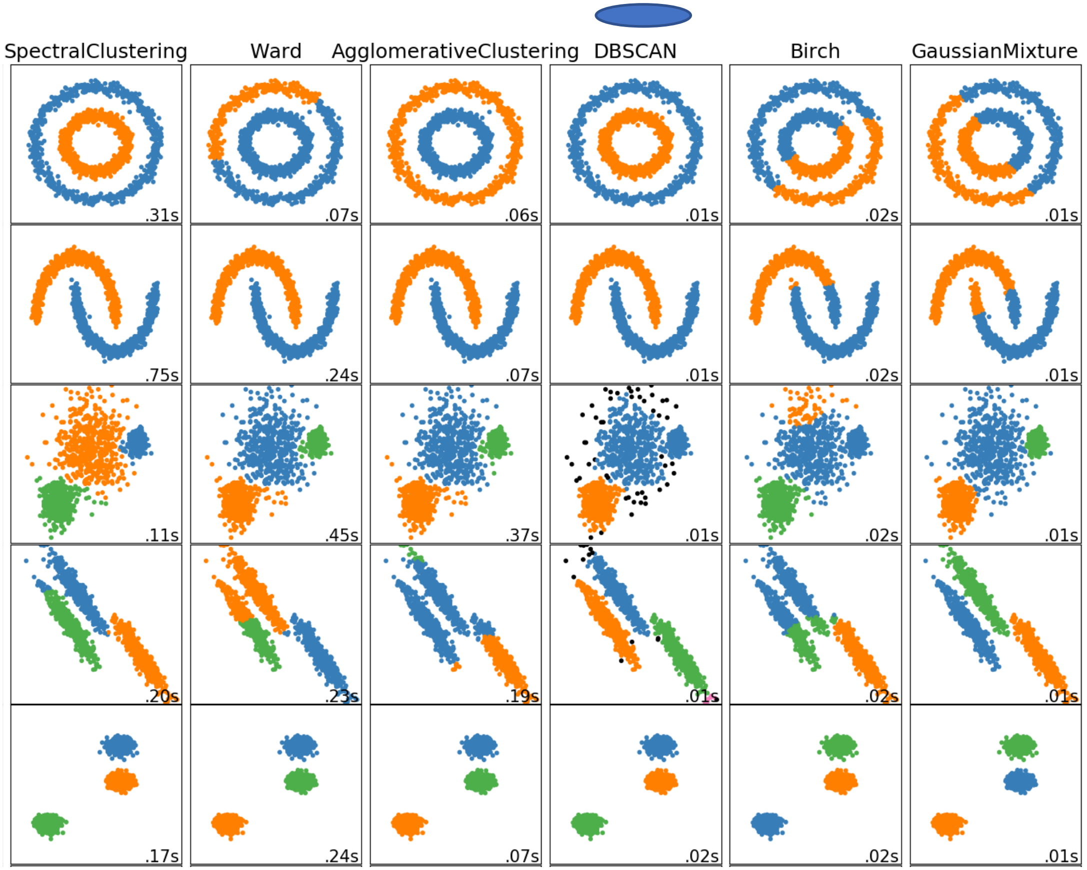
DBSCAN

- Density
- Based
- ~~SCAN~~
- Spatial
- Clustering of
- Applications with
- Noise

DBSCAN clusters points that are connected to each other by regions of high density data

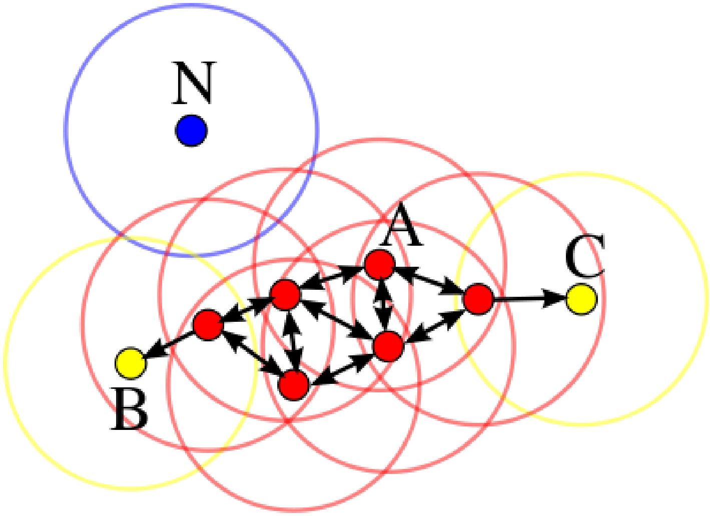


https://en.wikipedia.org/wiki/DBSCAN#/media/File:DBSCAN-density-data.svg



https://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_comparison.html

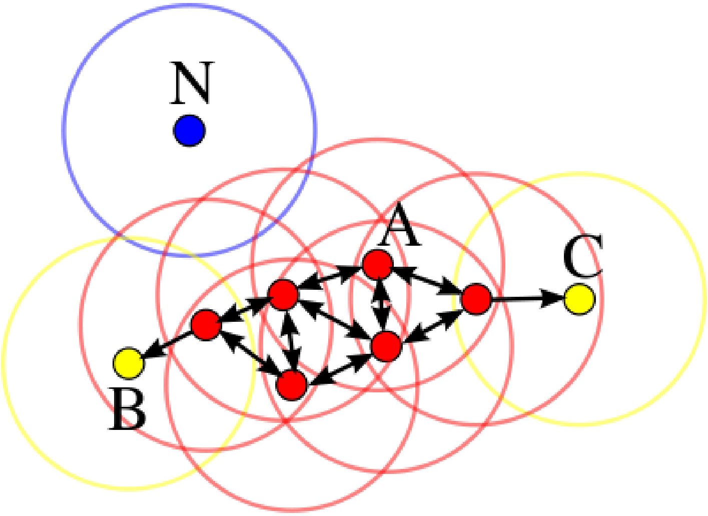
Some Terms



https://en.wikipedia.org/wiki/DBSCAN

- *Core point*: a point which is close to a lot of other points
- *Directly Reachable Point*: A point that is close to the point you are looking at
- *Reachable Point*: A point that is connected to the point you are looking at by a path of directly reachable points, and is not a core point
- *Outlier*: A point that is not reachable from a core point

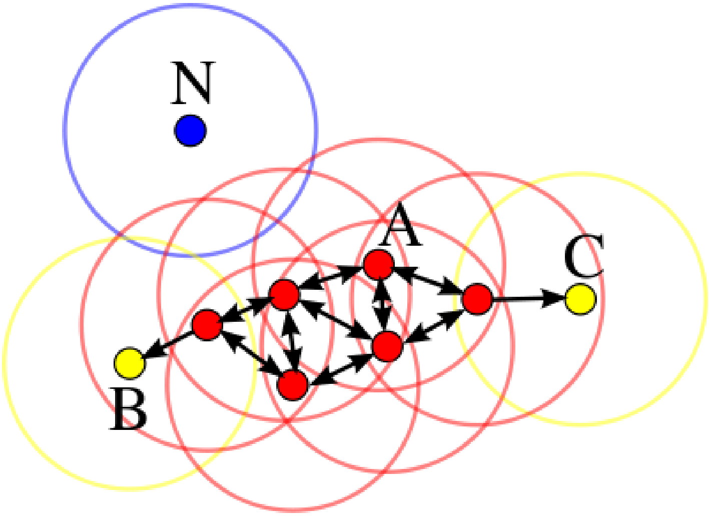
Hyperparameters



https://en.wikipedia.org/wiki/DBSCAN

- **d** – distance which defines what it means for a point to be "close"
  - The circle at point P with radius d is called the *neighbourhood* of the point
- **Nmin** – number defining the minimum number of points in the neighbourhood of a point for it to be considered a core point

Better Terms



https://en.wikipedia.org/wiki/DBSCAN

- *Core point*: a point which has at least **Nmin** neighbours a distance of **d** or less away

- *Directly Reachable Point*: Point Q is directly reachable from point P if |Q-P|<**d**

- *Reachable Point*: a point Q is reachable from core point P if there exists a path of n points p(i), such that p(i+1) is directly reachable from p(i), and P=p(1) and Q=p(n)

- *Outlier*: A point that is not directly reachable from a core point

The Algorithm

- Pick **d** and **Nmin**
- Visit an unvisited point
- Determine if it is a core point
- If it is not: label as noise, move to new point
- If it is: add it and its nearest neighbours to a cluster
- Select a new point and repeat

Runtime: O(n^2) ; Clever stuff: O(nlogn)

Advantages

- Don’t need to specify N-Clusters
- Finds arbitrary shaped clusters
- Robust to noise, finds outliers
- Only 2 hyperparameters

Disadvantages

- Border points don’t have a determined cluster when they are reached by 2 core points of different clusters
- Dependent on distance measure
- Cannot clusterize with data containing “clusters” of widely varied density
- Can be difficult to choose **Nmin** & **d**