AIcademy

Unsupervised Machine Learning / Clustering

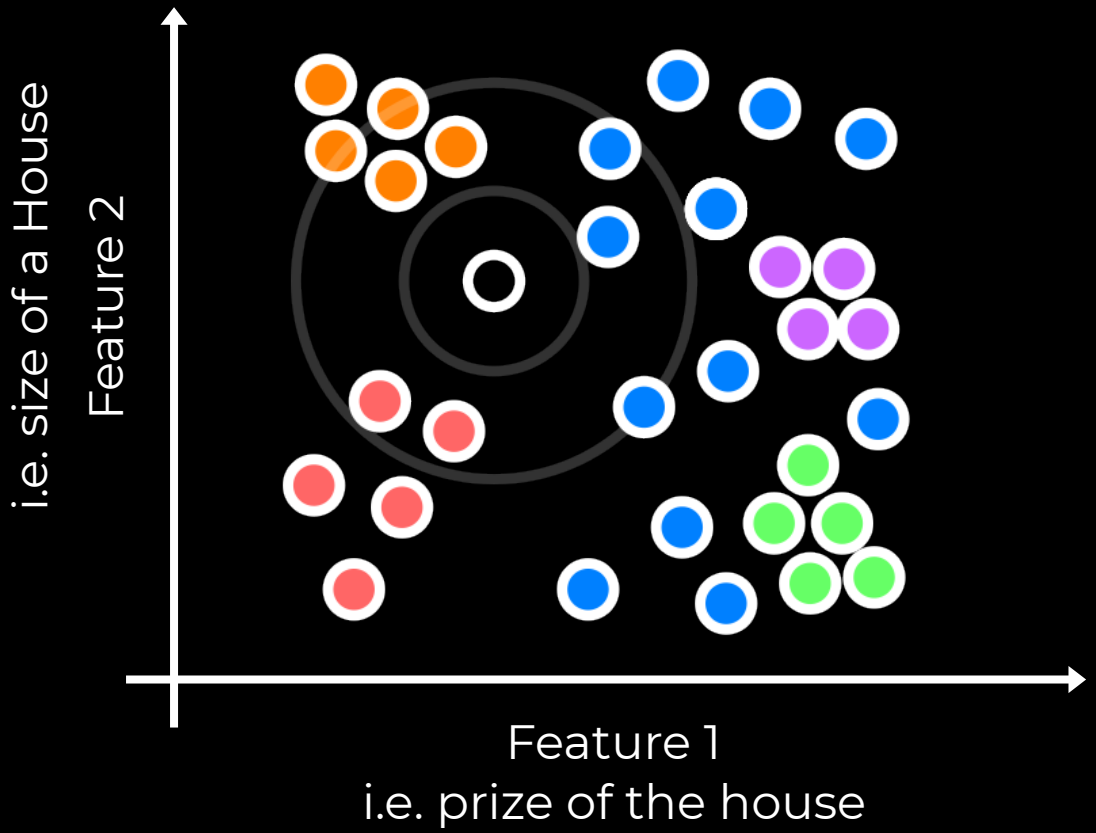
Unsupervised Machine Learning

Allows us to discover new patterns in datasets.

In most cases unsupervised machine learning is equivalent to clustering.

Clustering groups “samples” of a “similar” kind together.

Sample Space

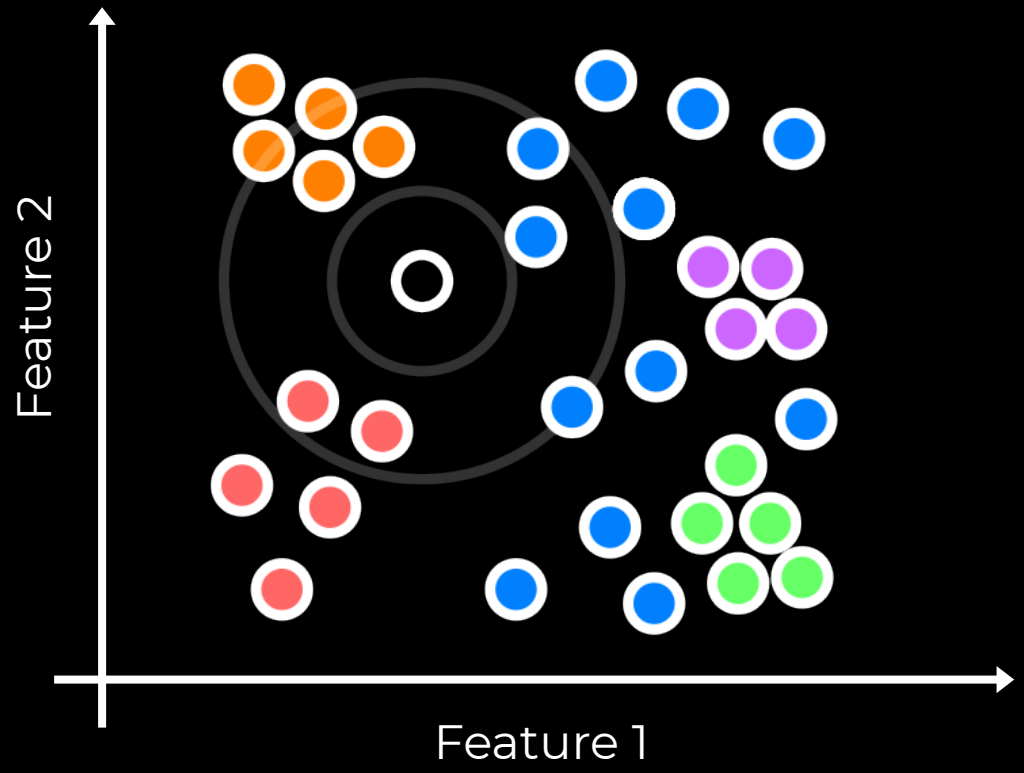


Allows to find patterns that one did not know about.

In most cases unsupervised machine learning is equivalent to clustering.

Clustering groups “samples” of a “similar” kind together.

Sample Space



We see already in the diagram that here several “samples” from our dataset “cluster” together.

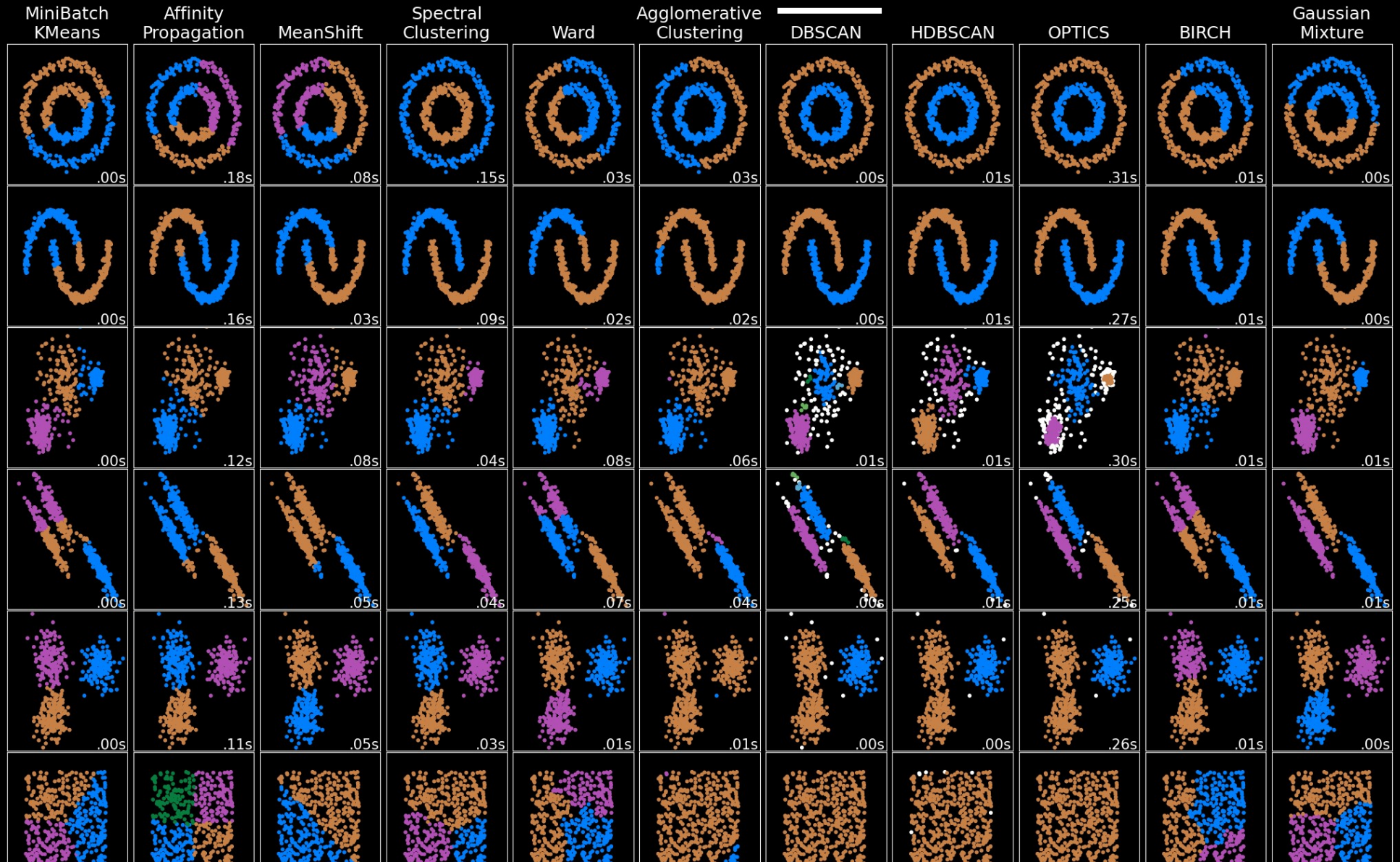
With multiple features finding such an result is not as easy

Clustering Algorithms

Many different clustering algorithms do exist.

Here we will discuss one of the most useful and natural, density based clustering method.

DBSCAN:
Density Based Spatial Clustering with Applications of Noise


MiniBatch KMeans
Affinity Propagation
MeanShift
Spectral Clustering
Ward
Agglomerative Clustering
DBSCAN
HDBSCAN
OPTICS
BIRCH
Gaussian Mixture
.00s .18s .08s .15s .03s .03s .00s .01s .31s .01s .00s
.00s .16s .03s .09s .02s .02s .00s .01s .27s .01s .00s
.00s .12s .08s .04s .08s .06s .01s .01s .30s .01s .01s
.00s .13s .05s .04s .07s .04s .00s .01s .25s .01s .01s
.00s .11s .05s .03s .01s .01s .00s .00s .26s .01s .00s

DBSCAN Algorithm

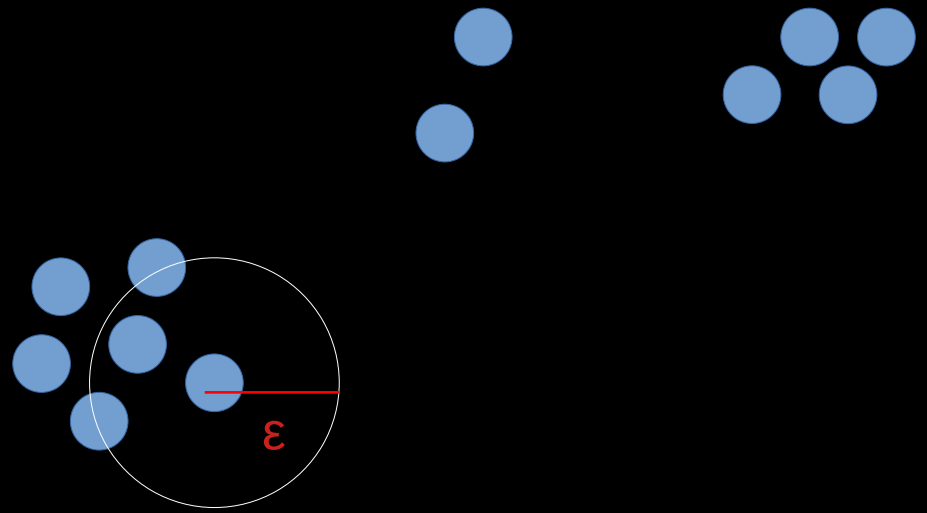
- Published in 1996
- Density Based Algorithm for Discovering Clusters in Large Spatial Datasets with noise.
- Finds members of a density connected region

$$\rho > \rho(\varepsilon, minpts) = minpts / V(\varepsilon)$$

Number of a sequences within radius ε

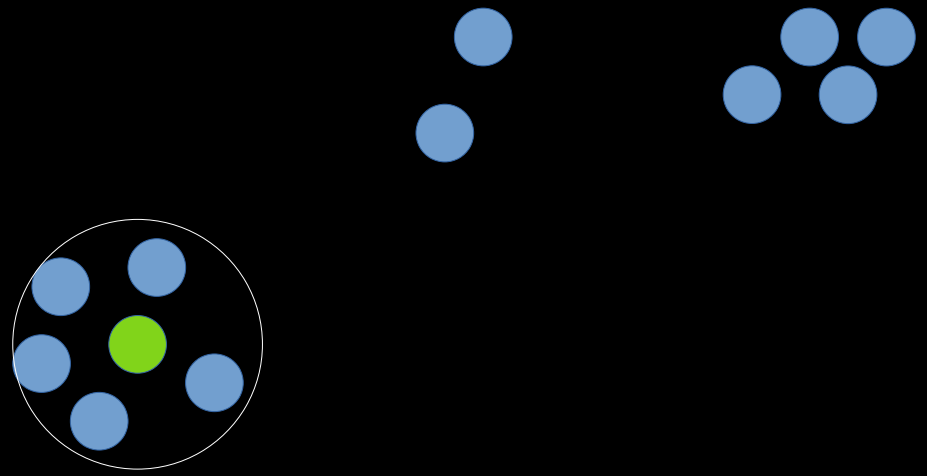
Volume of a ball with radius ε

How it works



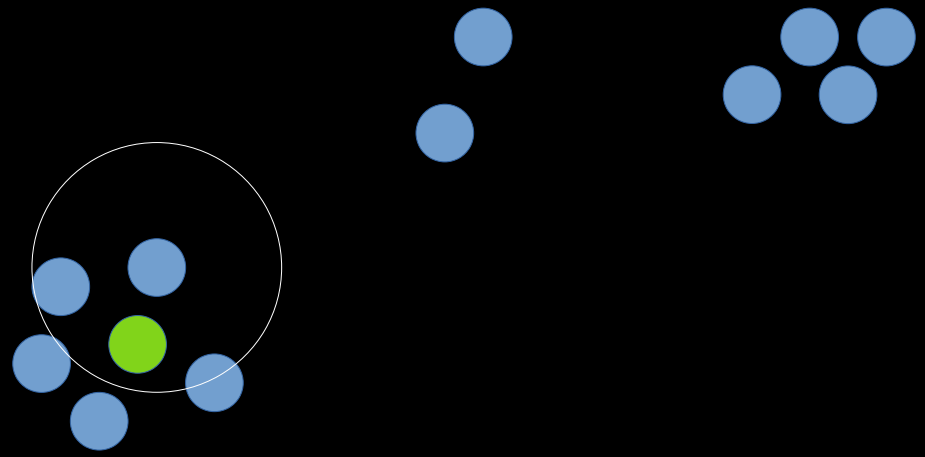
Run with minpts 3

How it works



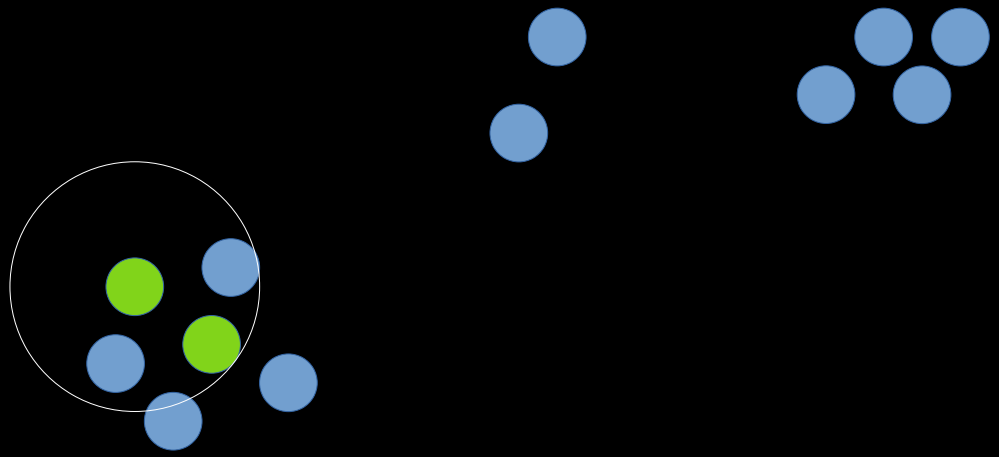
Run with minpts 3

How it works



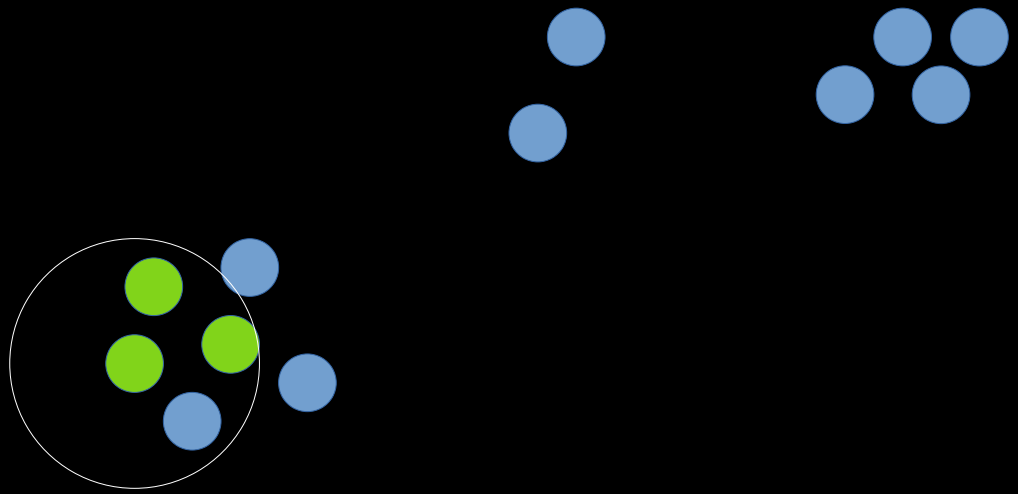
Run with minpts 3

How it works



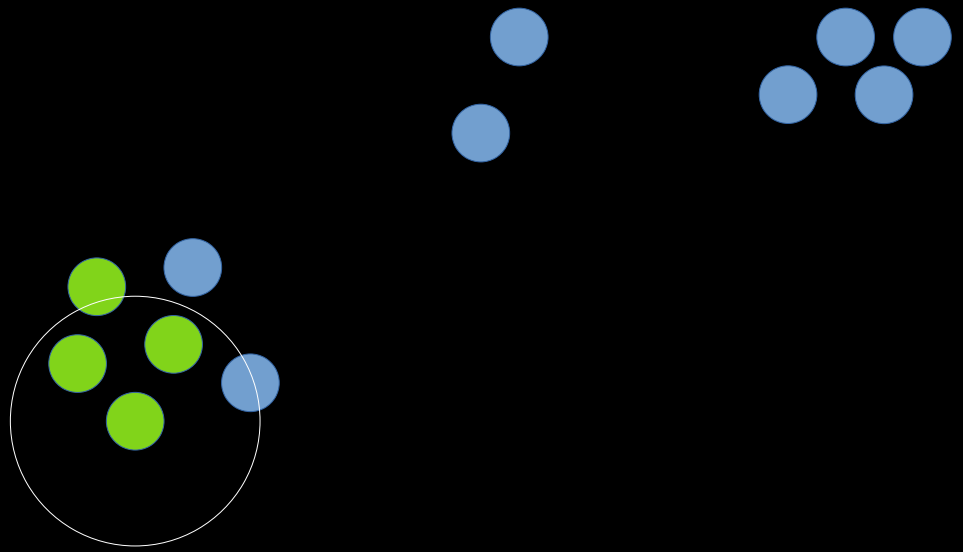
Run with minpts 3

How it works



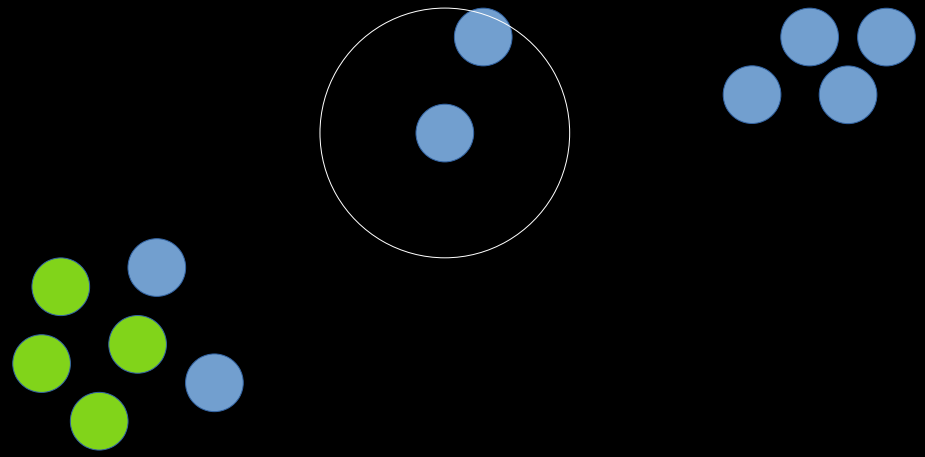
Run with minpts 3

How it works



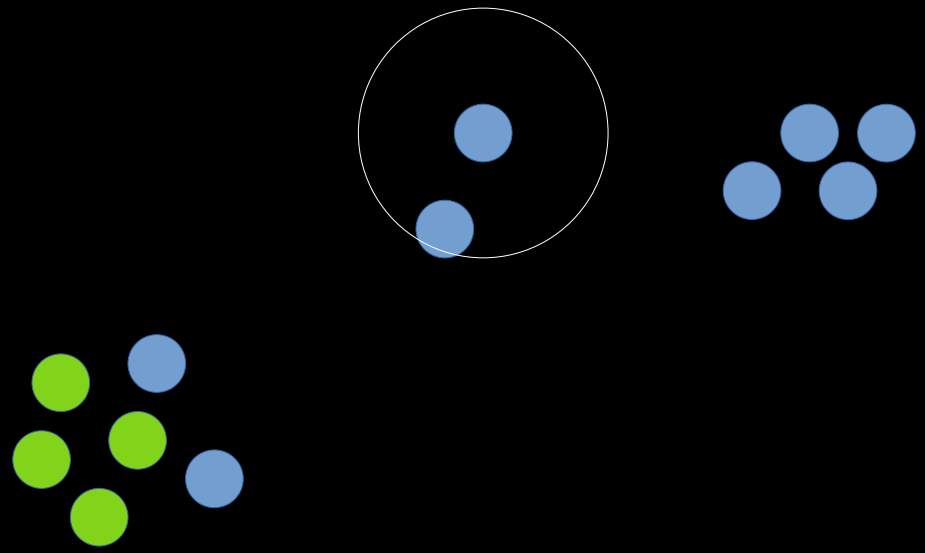
Run with minpts 3

How it works



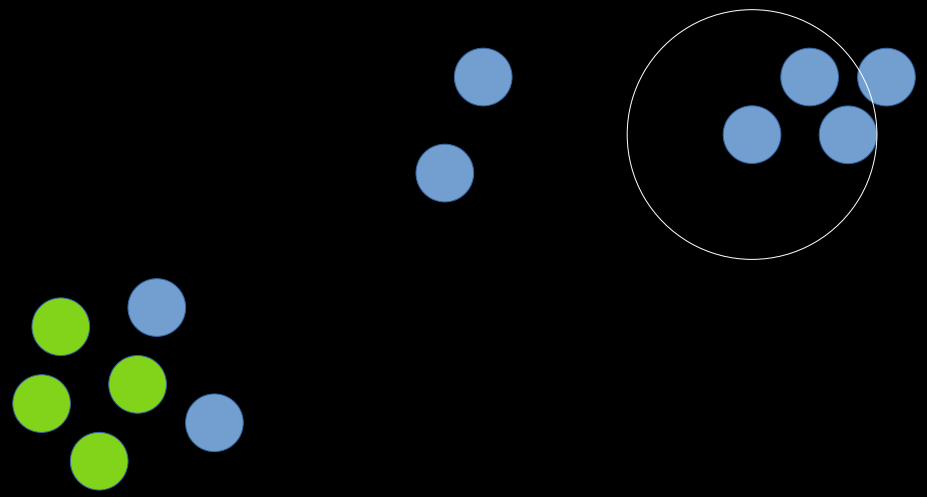
Run with minpts 3

How it works



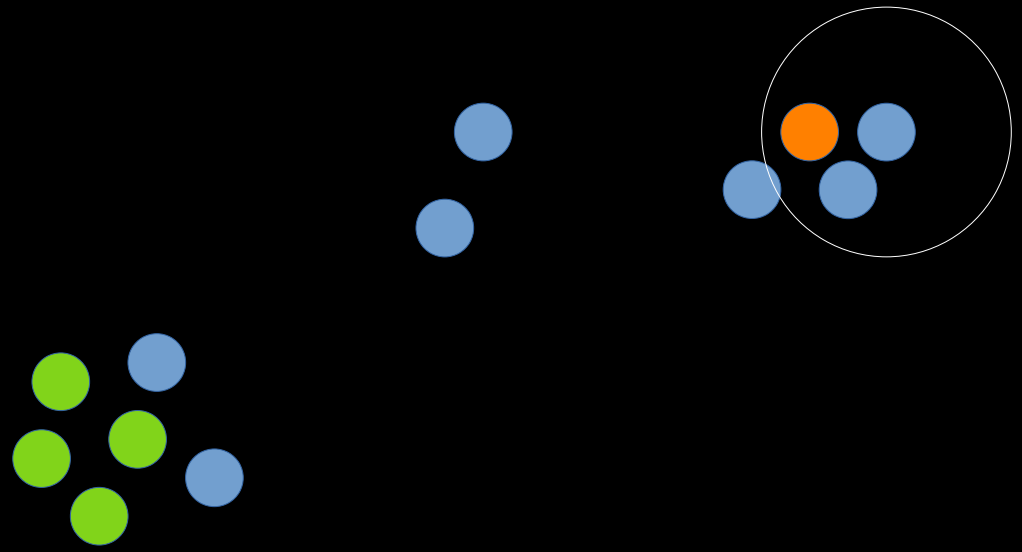
Run with minpts 3

How it works



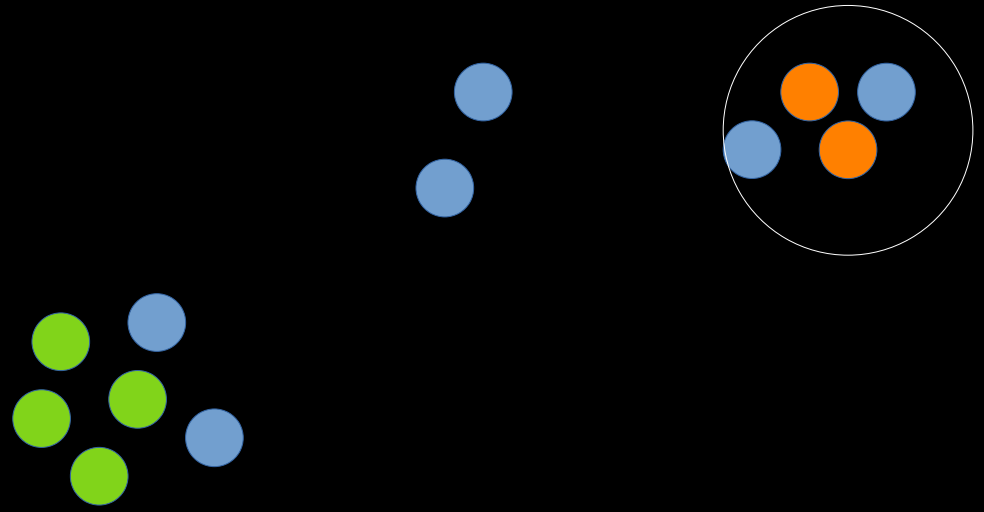
Run with minpts 3

How it works



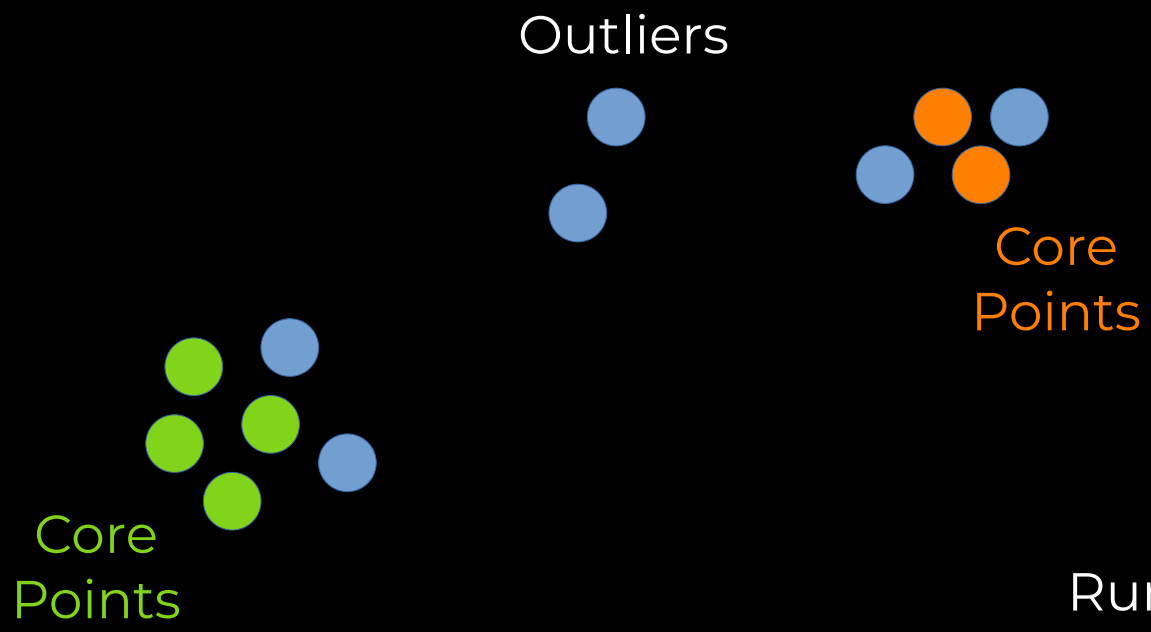
Run with minpts 3

How it works



Run with minpts 3

How it works



Run with minpts 3

DBSCAN Results

Using DBSCAN one distinguishes:
Core Points, Neighbors and Outliers.

The result are density connected regions.

DBSCAN

Guessing parameters ε and minpts can be tricky.

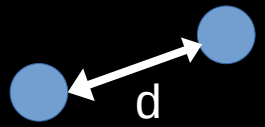
Guides:

- Plotting and guessing by scale
- Taking prior knowledge into account
- In the worst case, run DBSCAN multiple times with different ε and minpts parameters

DBSCAN and DISTANCE

Depending on the dataset you may want to adapt the distance measure between the samples.

The distance function:
d(sample1, sample2)
is strongly influencing the results of DBSCAN and has to be carefully designed / chosen.



Practical

We will use DBSCAN to investigate climate zones.

For this we have a dataset of weather conditions in different French cities.

You will learn how to draw data on a map and how clustering algorithms may allow you to find the unexpected.