Unsupervised Learning

Machine Learning 101

Institute of AI

Why unsupervised learning?

- Supervised learning works great if you have correct labels
- Most of the data in the real world is not labelled
- Labelling using human expertise is expensive

Types of unsupervised learning algorithms

Clustering
Grouping similar instances into clusters

Anomaly detection
Identifying abnormal instances in data

Dimensionality reduction
Generate a representation of data with fewer dimensions

Clustering

Example: **customer segmentation**.
Cluster customers based on their activity on your site so that you can adapt your product and marketing campaigns to each segment.

Example: **semi-supervised learning**.
Only a few samples have labels. Perform clustering and assign the same label to all items in a cluster.

Clustering

Two popular clustering algorithms

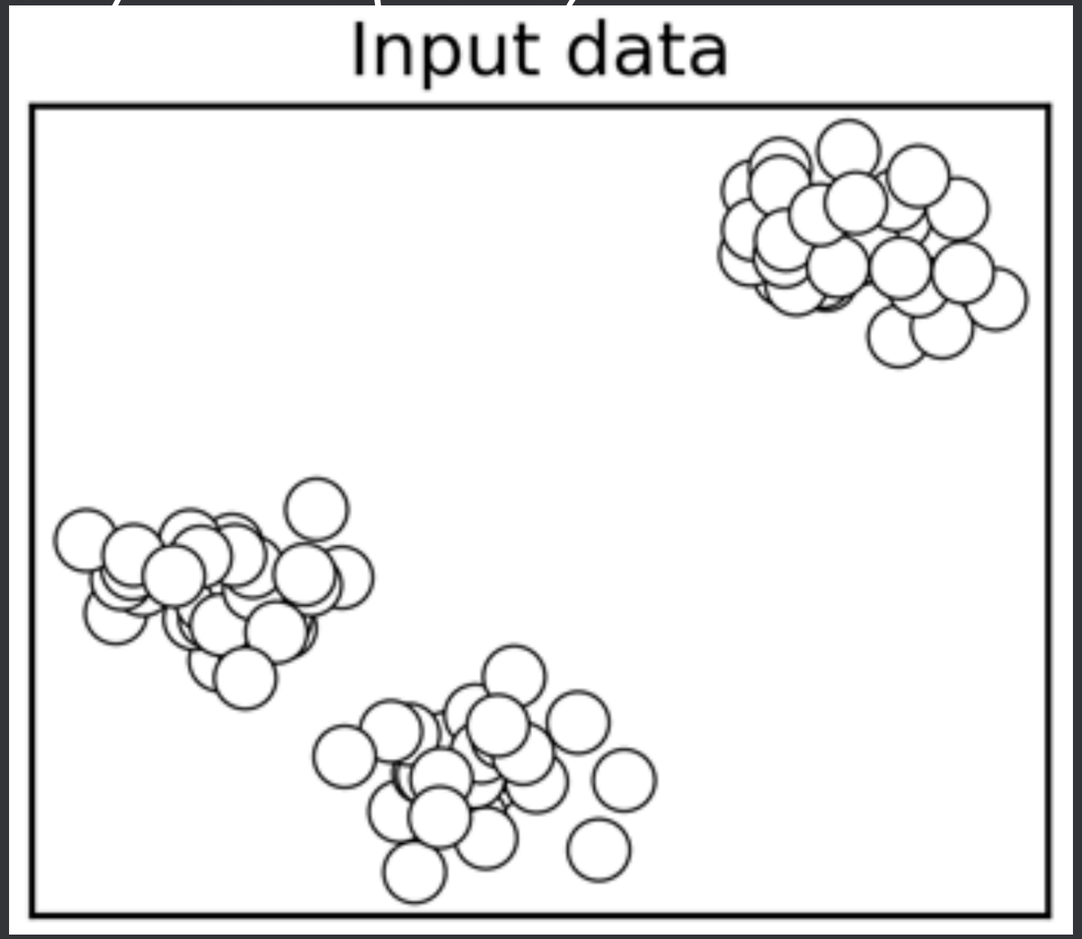
K-Means

DBSCAN

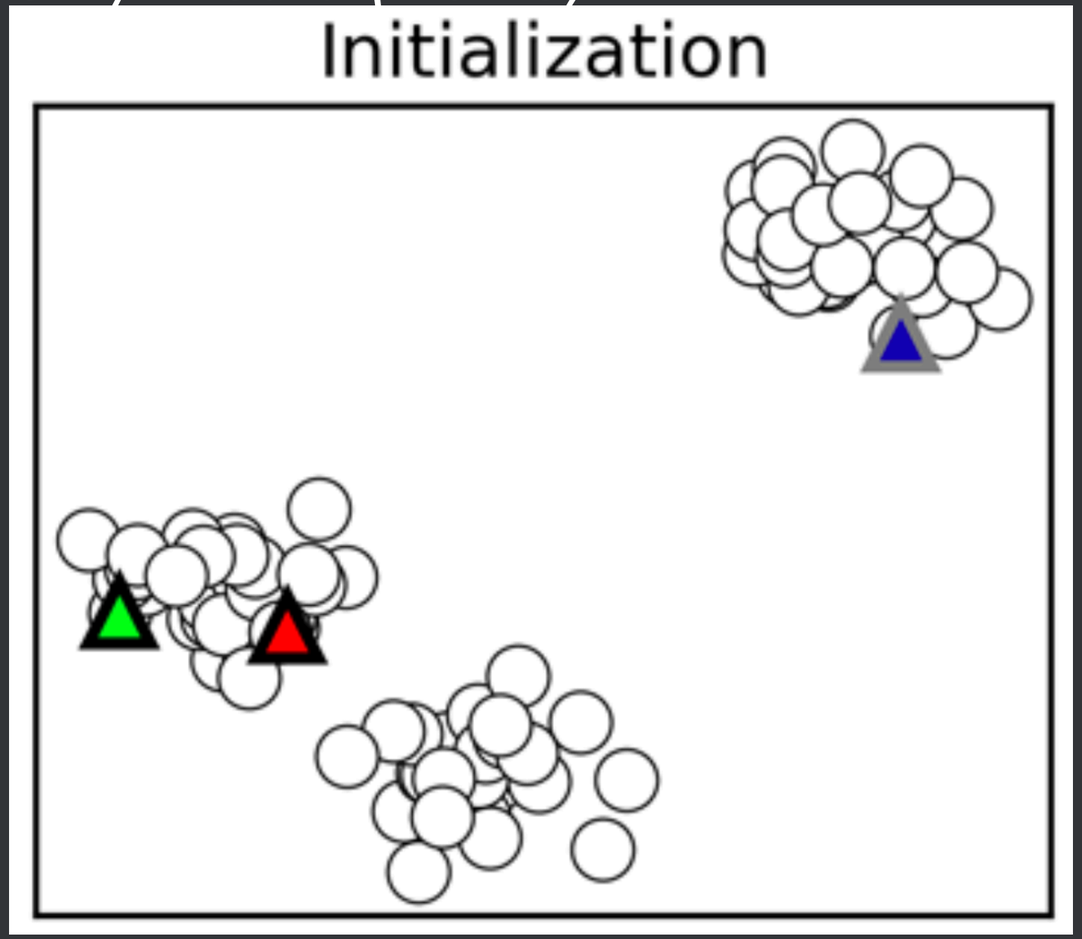
K-Means

Clustering

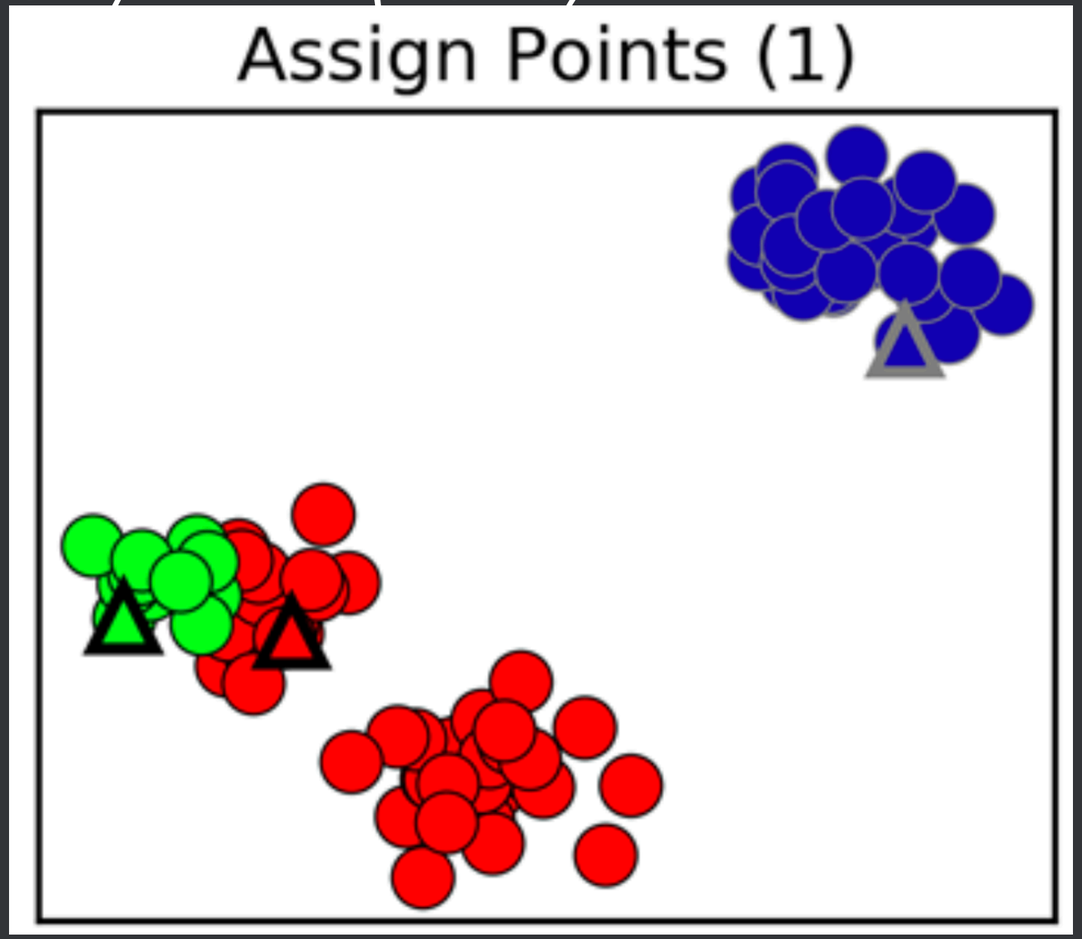
Source: Introduction to Machine Learning with Python, Andreas C. Müller and Sarah Guido, O'Reilly Media (2016)



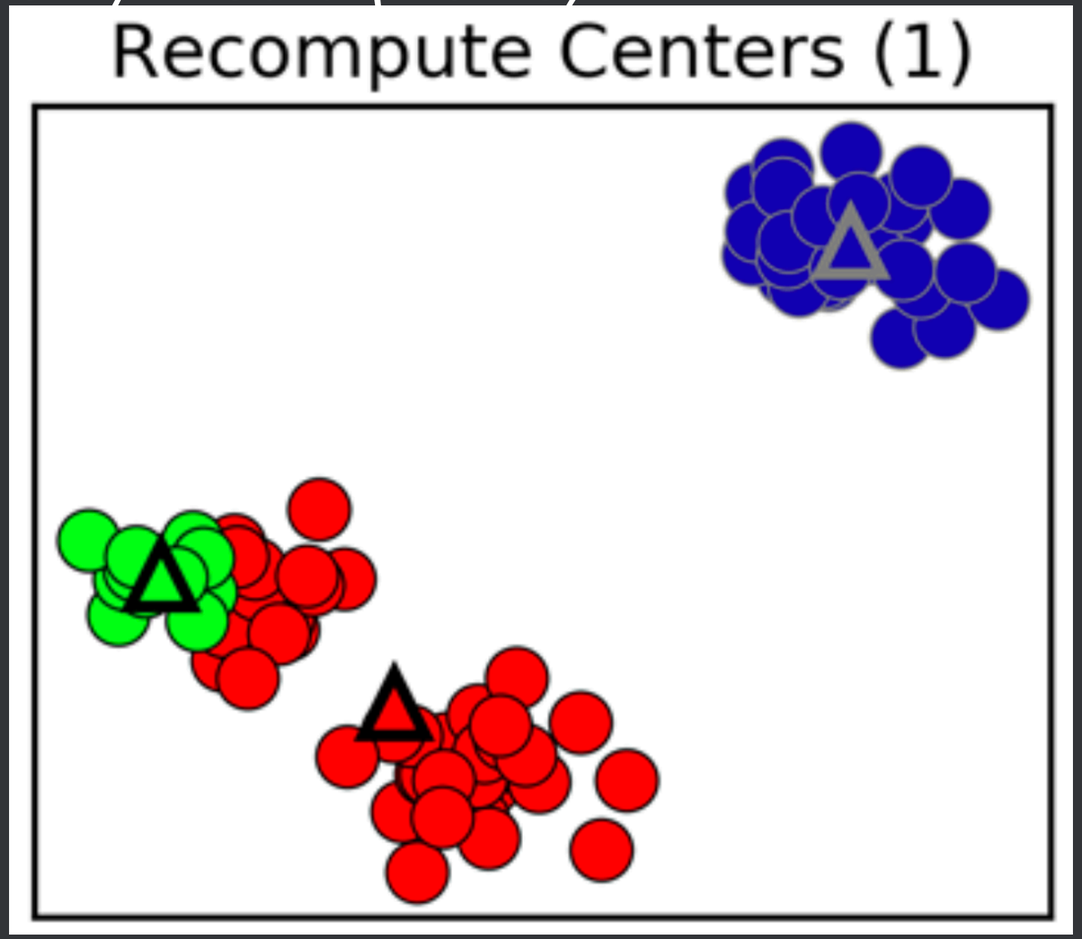
Source: Introduction to Machine Learning with Python, Andreas C. Müller and Sarah Guido, O'Reilly Media (2016)



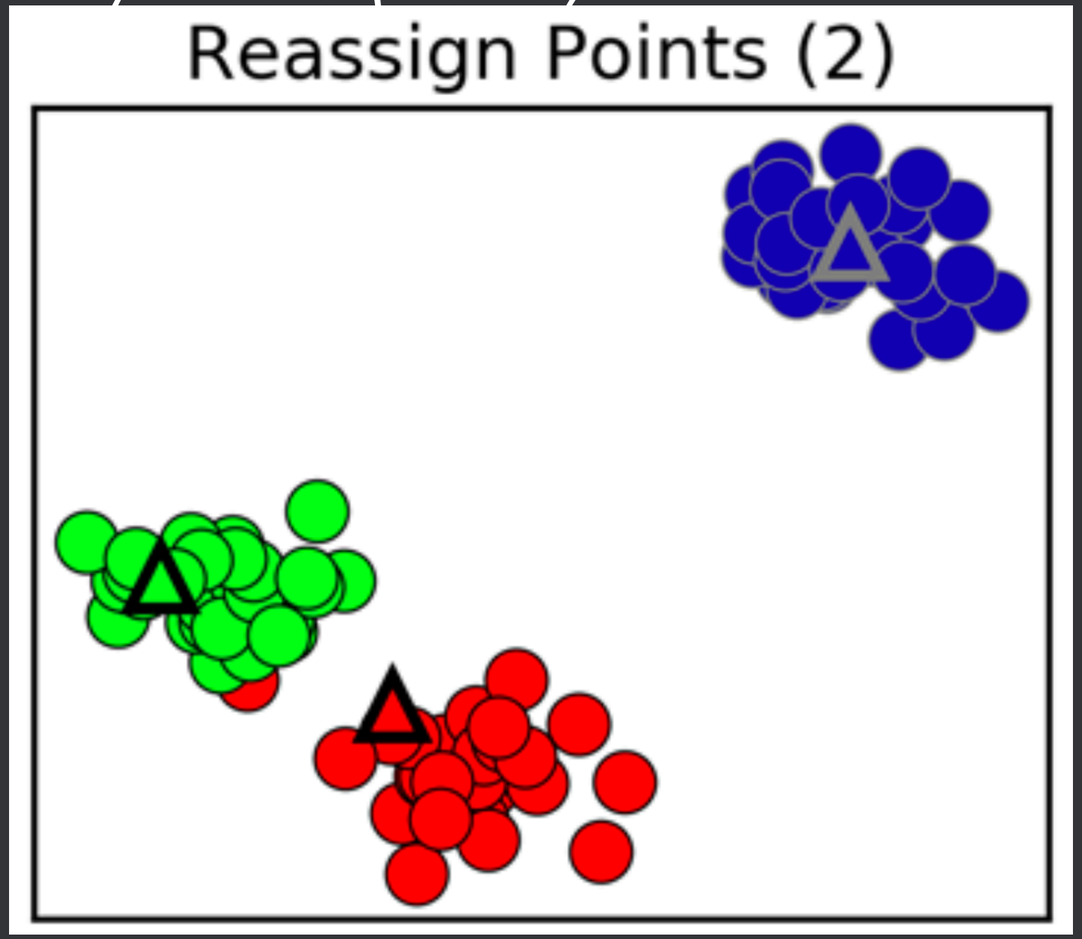
Source: Introduction to Machine Learning with Python, Andreas C. Müller and Sarah Guido, O'Reilly Media (2016)



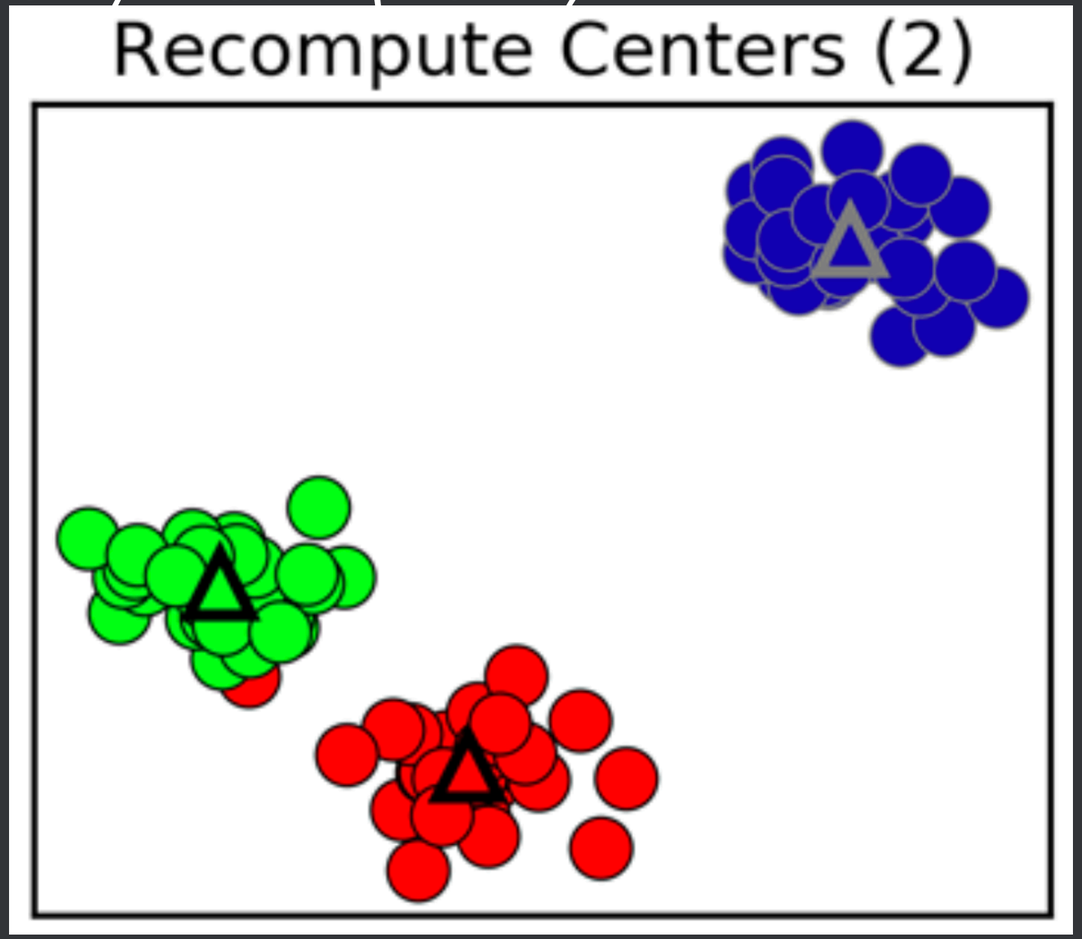
Source: Introduction to Machine Learning with Python, Andreas C. Müller and Sarah Guido, O'Reilly Media (2016)



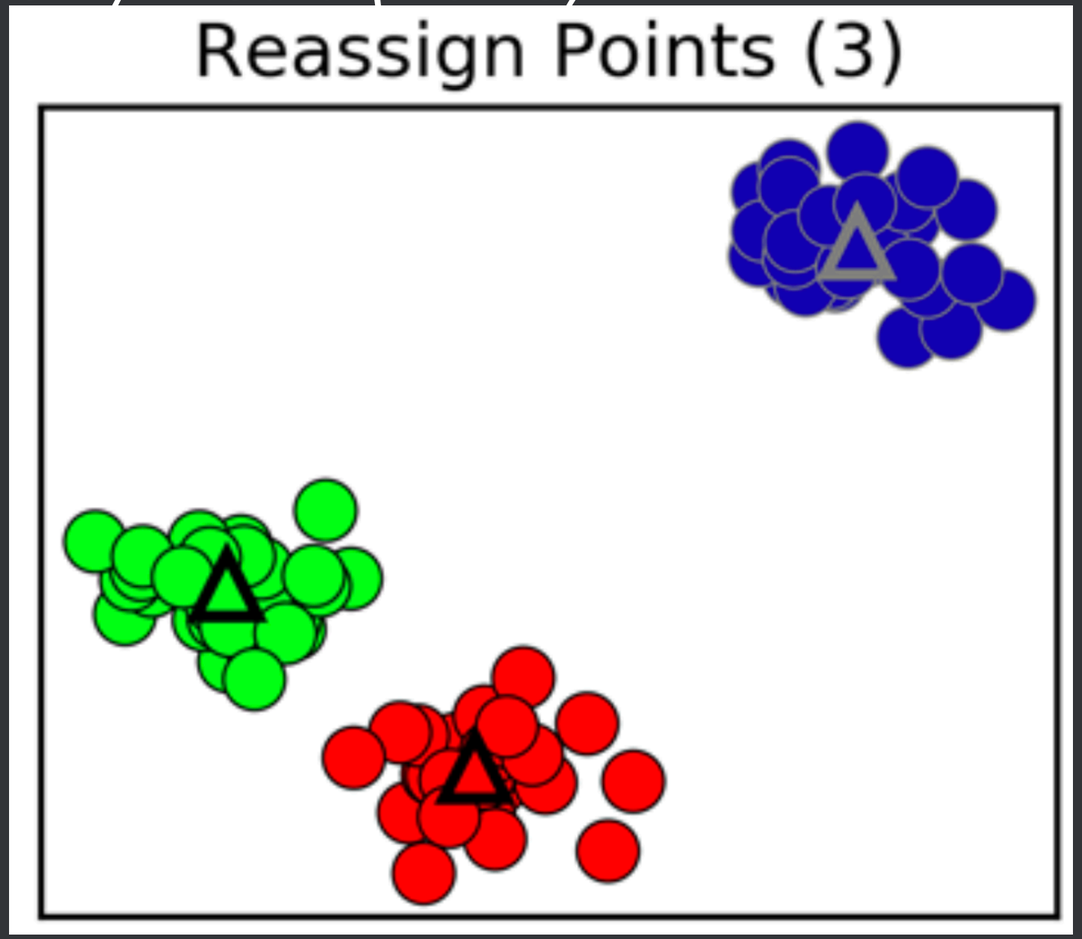
Source: Introduction to Machine Learning with Python, Andreas C. Müller and Sarah Guido, O'Reilly Media (2016)



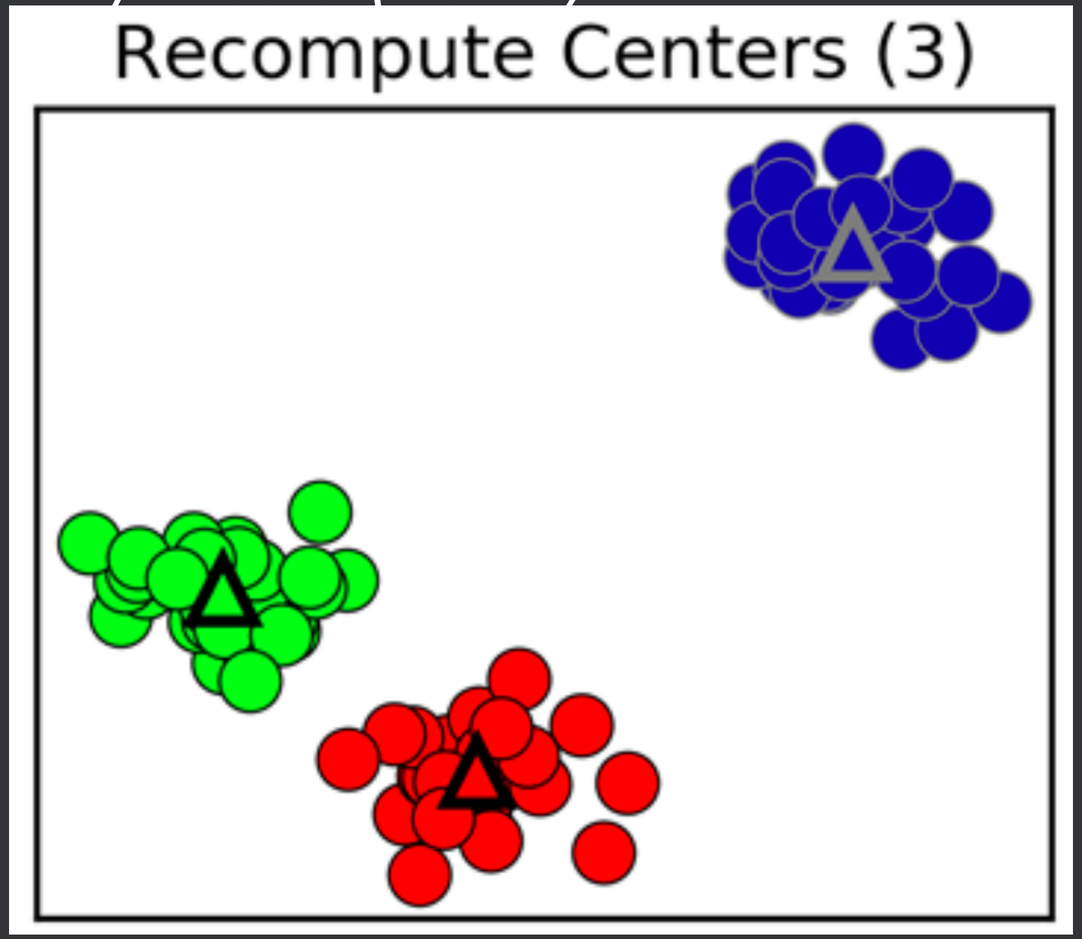
Source: Introduction to Machine Learning with Python, Andreas C. Müller and Sarah Guido, O'Reilly Media (2016)

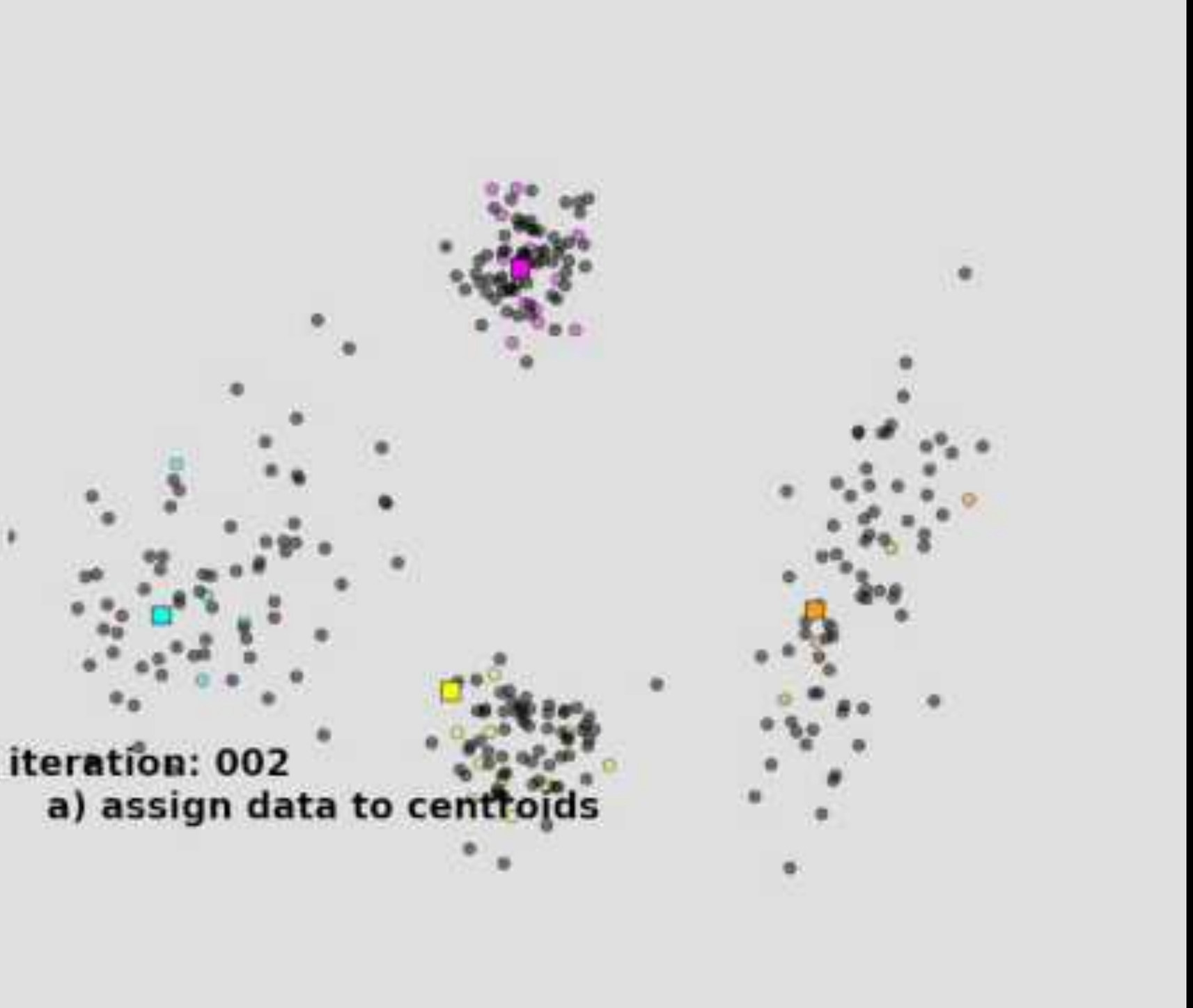


Source: Introduction to Machine Learning with Python, Andreas C. Müller and Sarah Guido, O'Reilly Media (2016)



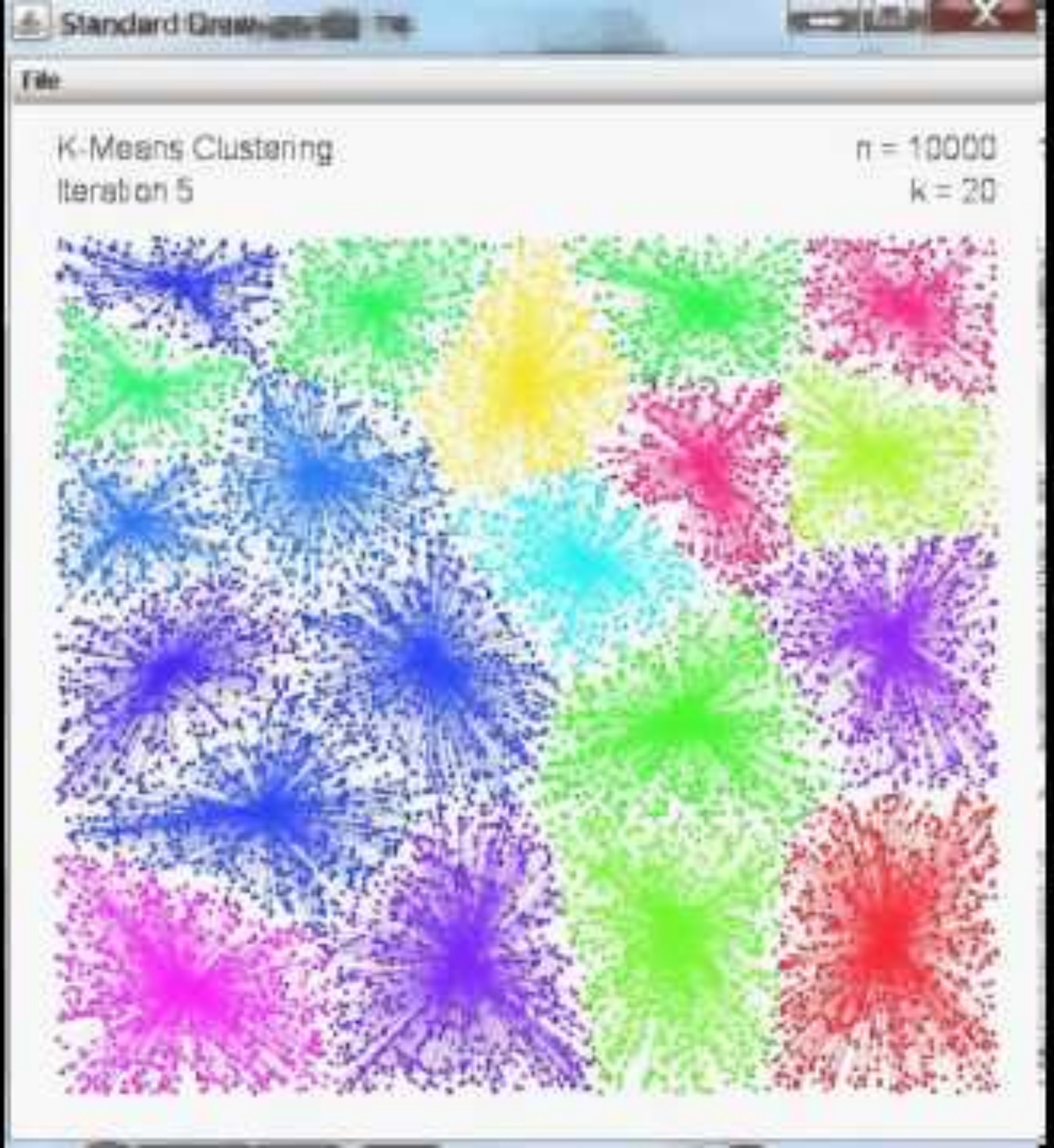
Source: Introduction to Machine Learning with Python, Andreas C. Müller and Sarah Guido, O'Reilly Media (2016)




iteration: 002
a) assign data to centroids

K-Means

Code Walkthrough


Standard Draw
File
K-Means Clustering
Iteration 5
n = 10000
k = 20

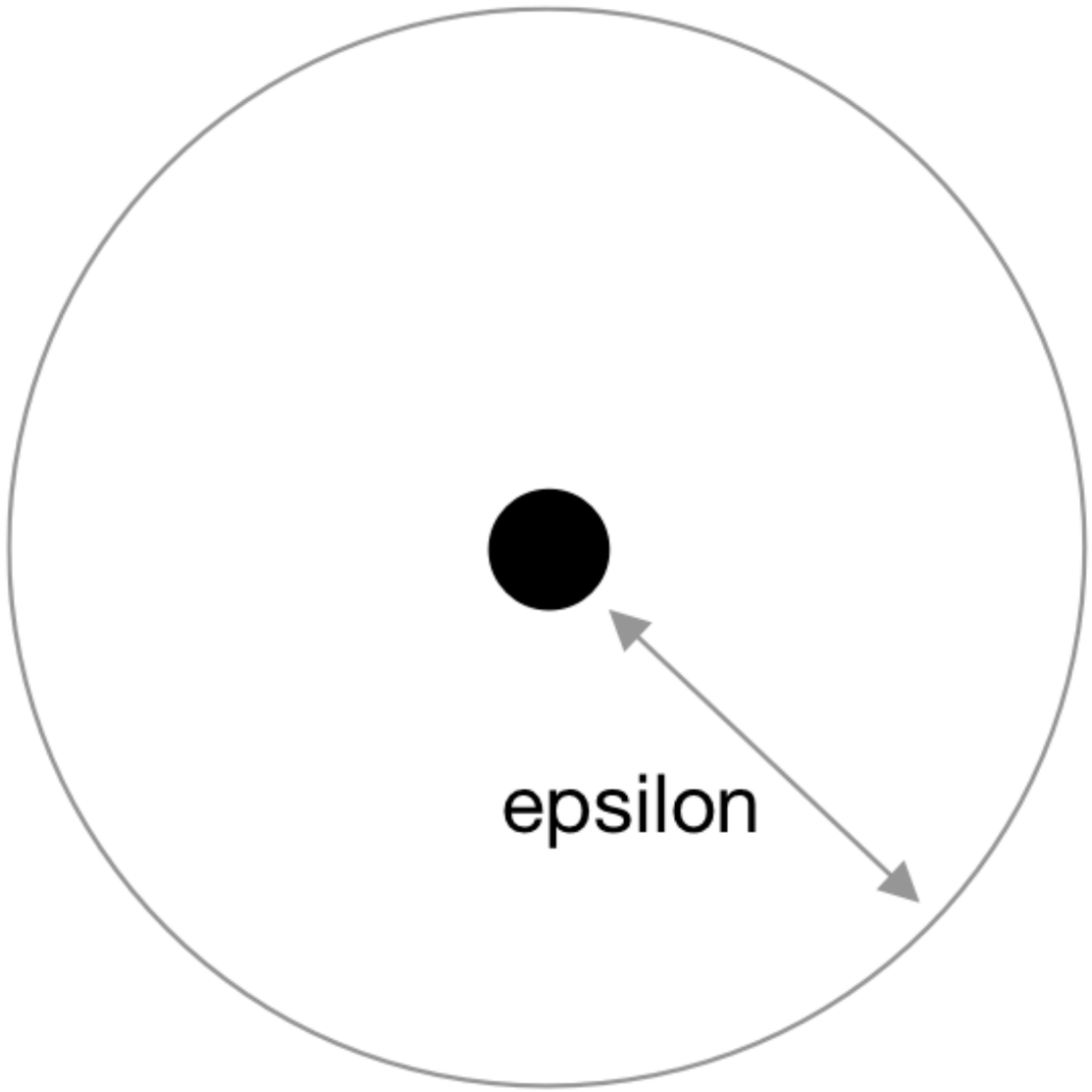
DBSCAN

Density Based Spatial Clustering of Applications with Noise

DBSCAN

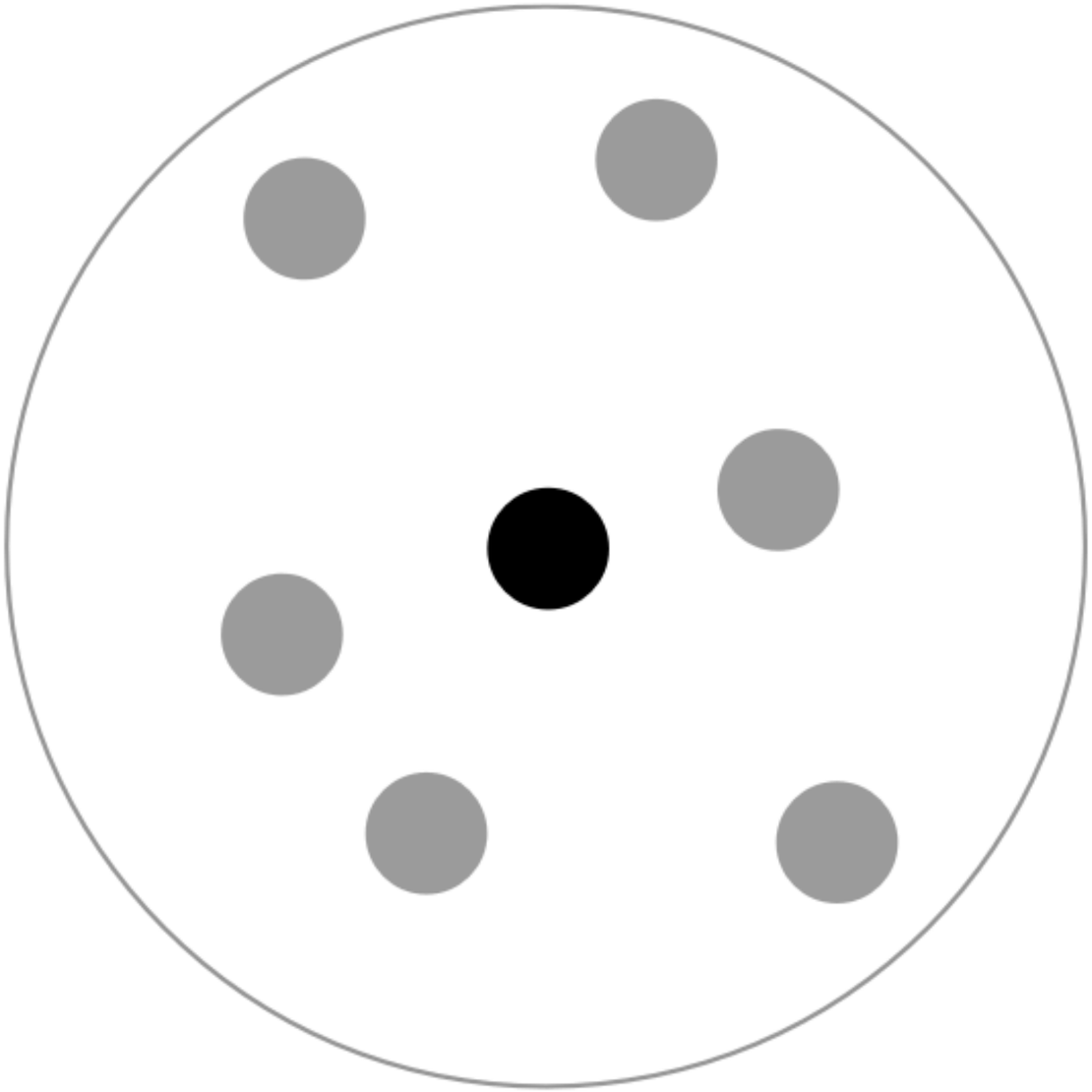
Hyperparameters

1. Epsilon
2. Minimum samples



Epsilon

The distance around a point.



Minimum samples

The number of points that have to be in the ϵ neighborhood to form a cluster.

DBSCAN

The algorithm

1. For each instance, count how many other instances are within epsilon (the ϵ neighborhood).
2. If a neighborhood has more than **"minimum samples"** instances, they form a cluster.
3. The neighborhood of a cluster is a combination of neighborhoods of all points in the cluster.

Visualizing DBSCAN

Link

DBSCAN

Code Walkthrough

Anomaly Detection using DBSCAN

Instances that have a label of `-1` are anomalies