Clustering

South-African Council for Automation and Control

Exploratory Data Analysis workshop

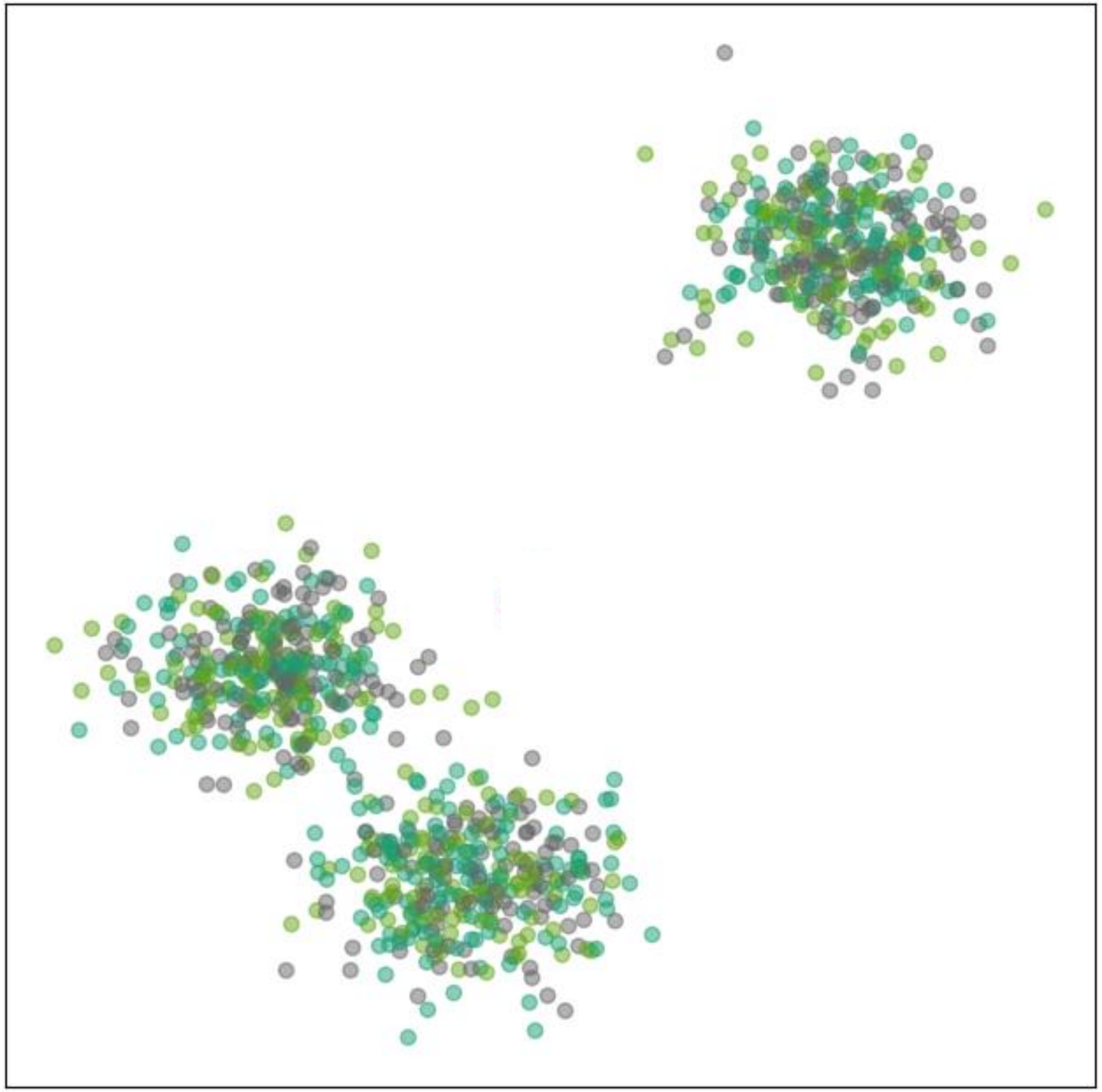
March 2024

Why clustering?

- Clustering is an unsupervised learning technique that groups "similar" data points together
- Similarity is often distance based in feature space (*consider scaling, curse of dimensionality*)
- Clustering can help identify structure in data, and assist in translating identified structures from dimensionality reduction back to time series plots

K-means clustering

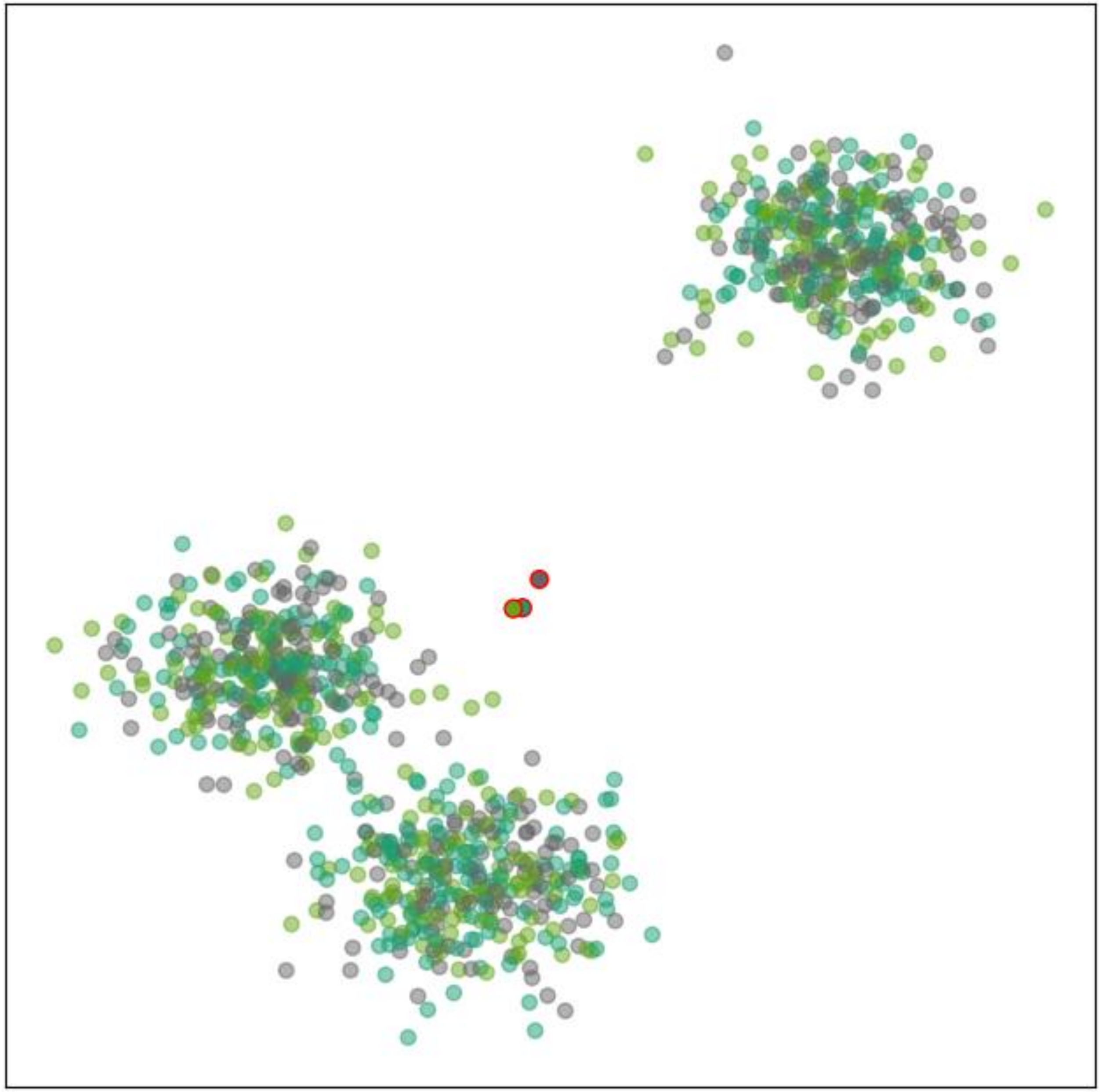
- Specify k number of clusters *a priori*
- **Randomly split data points amongst clusters**
  1. Calculate mean value of each cluster: *cluster centre*
  2. Reassign data points to closest cluster centre
  3. Repeat to convergence



https://scikit-learn.org/stable/modules/clustering.html

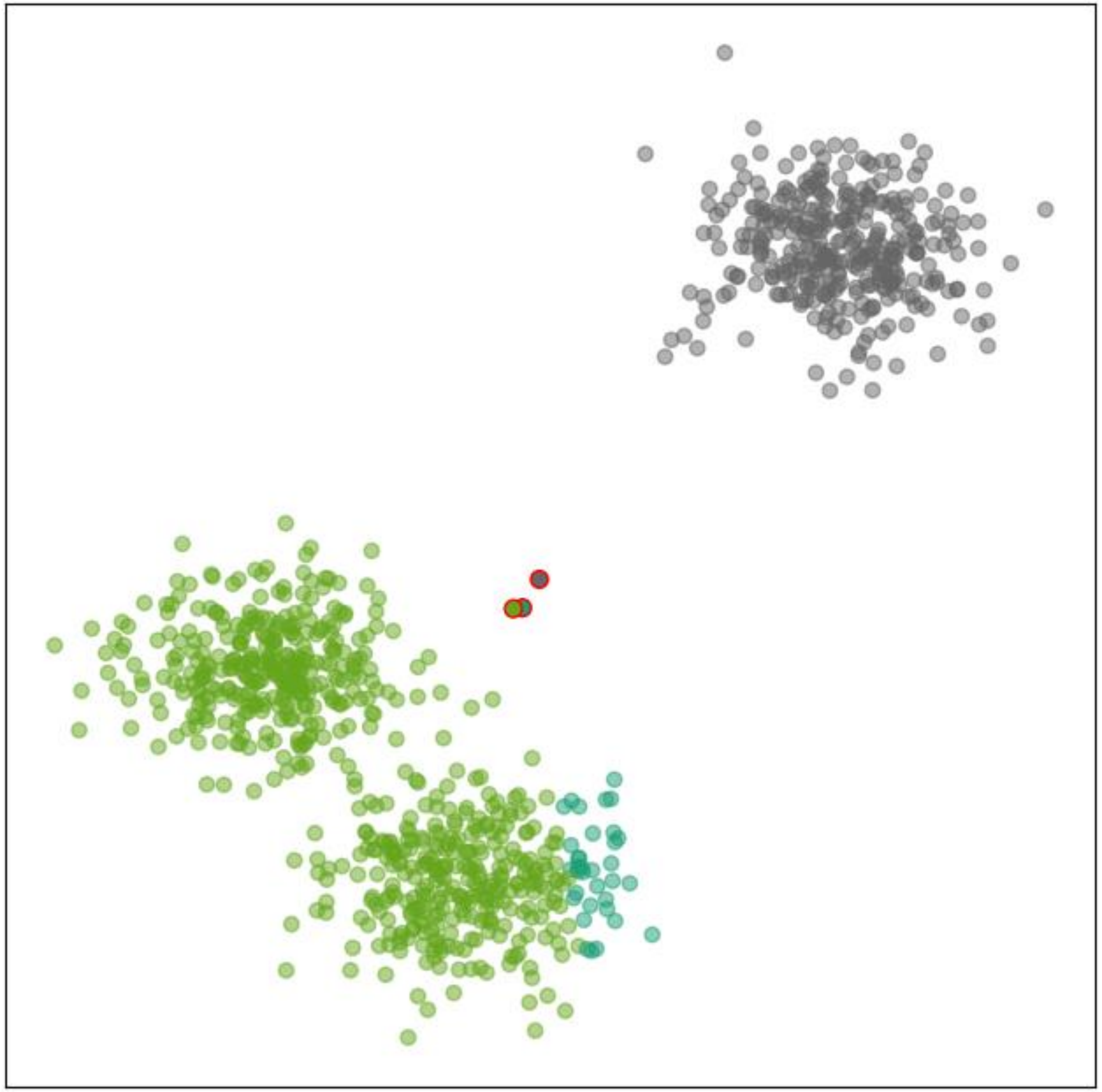
K-means clustering

- Specify *k* number of clusters *a priori*
- Randomly split data points amongst clusters
  1. **Calculate mean value of each cluster: *cluster centre***
  2. Reassign data points to closest cluster centre
  3. Repeat to convergence



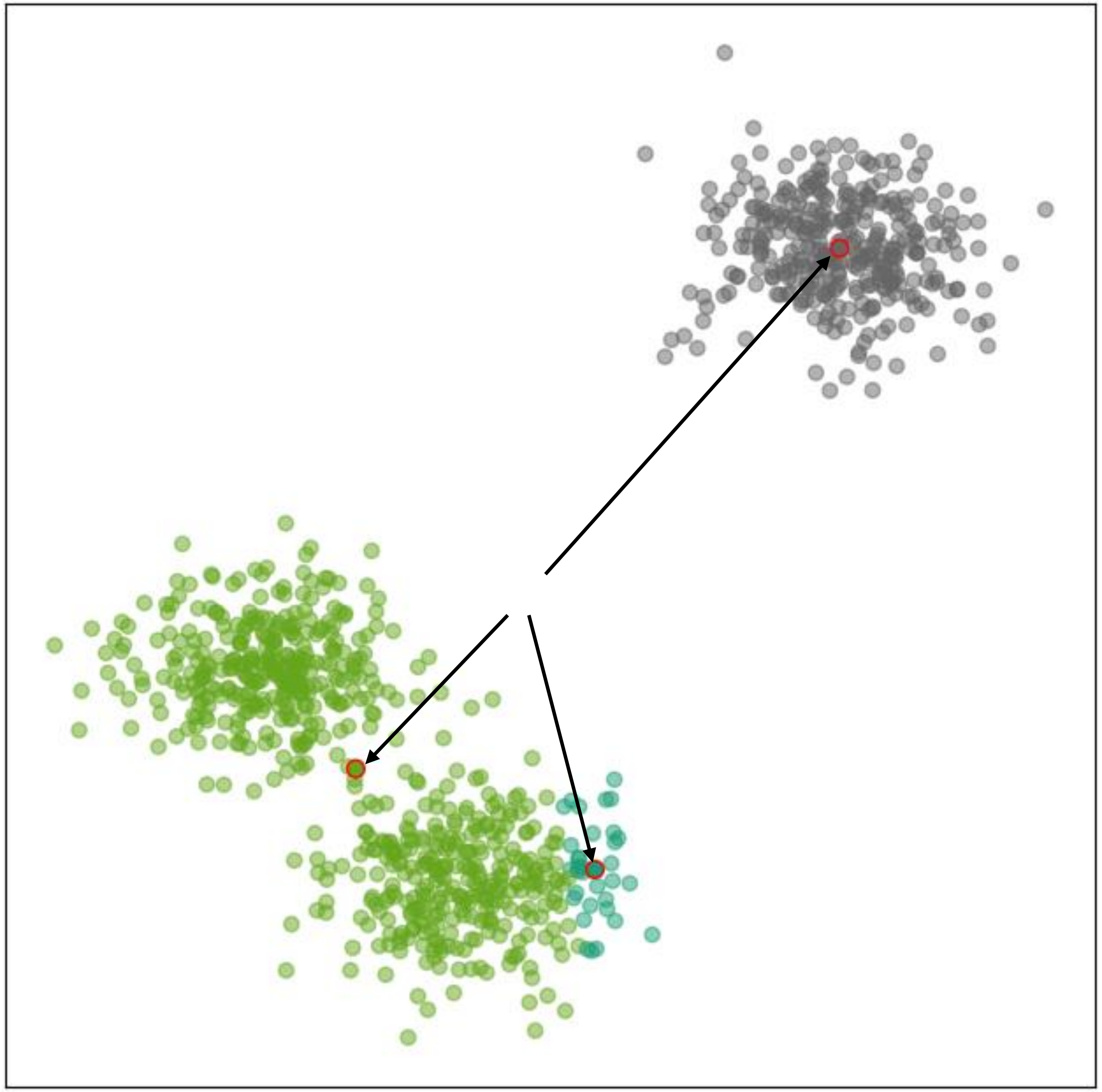
K-means clustering

- Specify k number of clusters *a priori*
- Randomly split data points amongst clusters
  1. Calculate mean value of each cluster: *cluster centre*
  2. **Reassign data points to closest cluster centre**
  3. Repeat to convergence



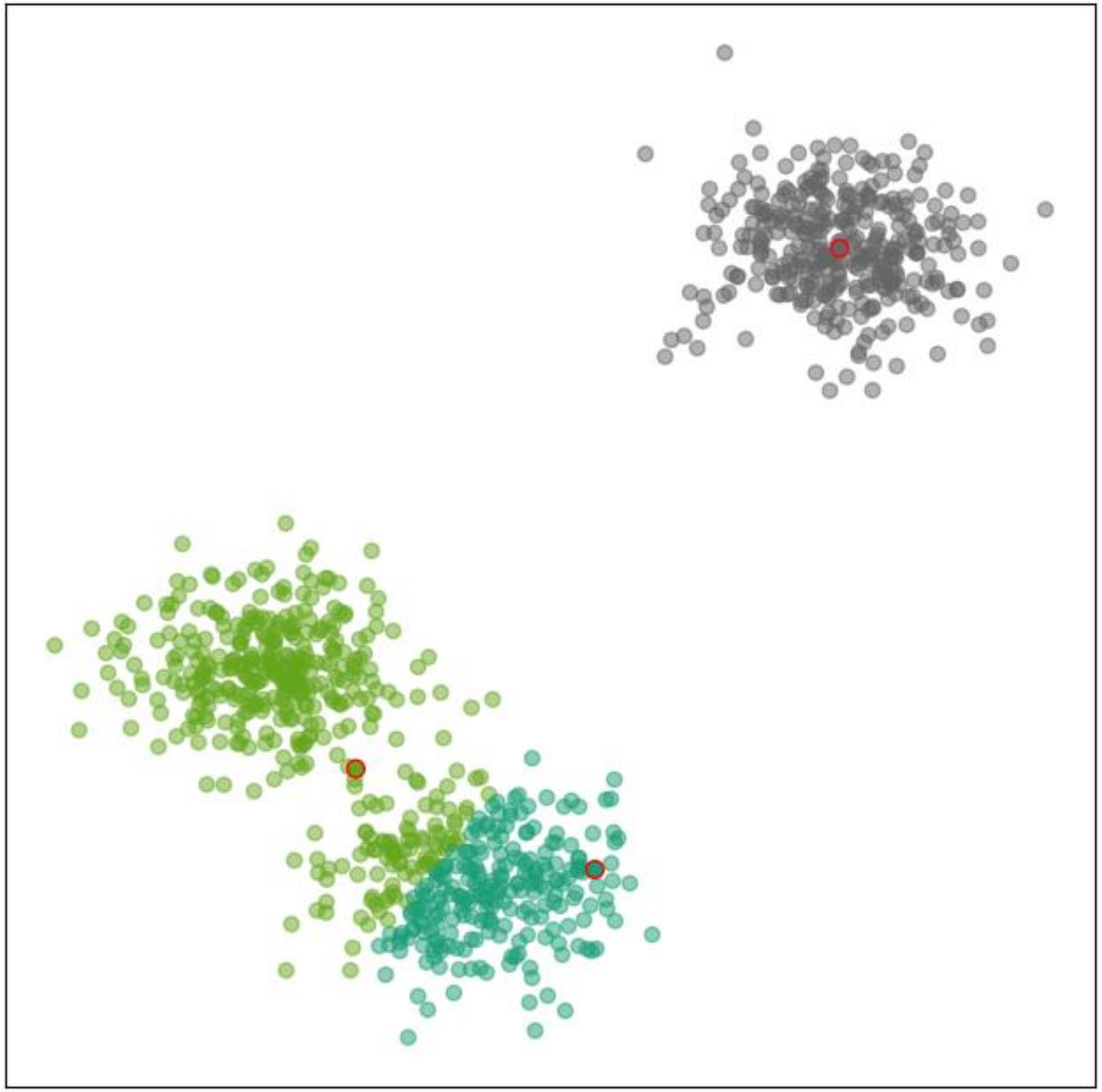
K-means clustering

- Specify *k* number of clusters *a priori*
- Randomly split data points amongst clusters
  1. **Calculate mean value of each cluster: *cluster centre***
  2. Reassign data points to closest cluster centre
  3. Repeat to convergence



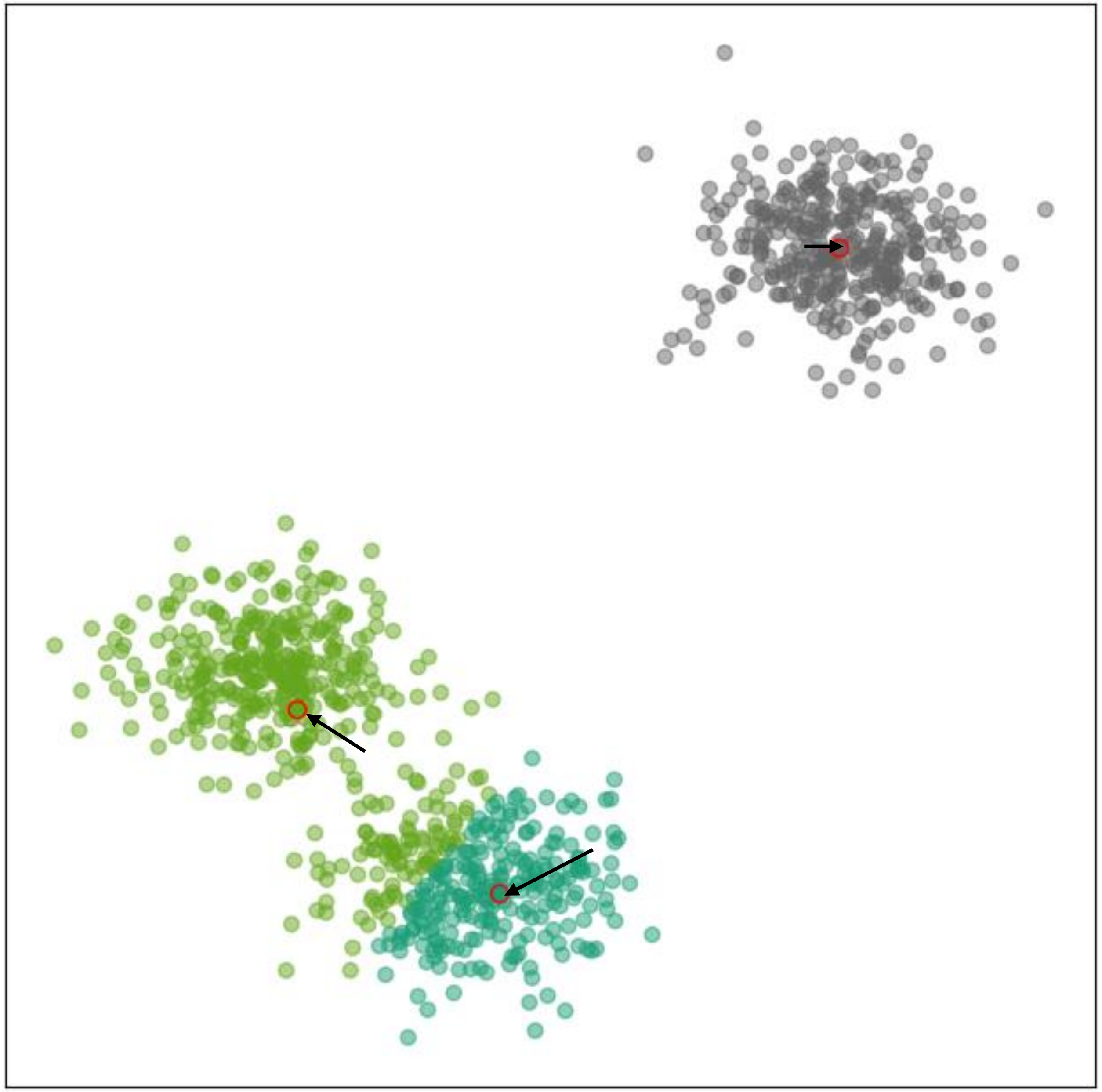
K-means clustering

- Specify k number of clusters *a priori*
- Randomly split data points amongst clusters
  1. Calculate mean value of each cluster: *cluster centre*
  2. **Reassign data points to closest cluster centre**
  3. Repeat to convergence



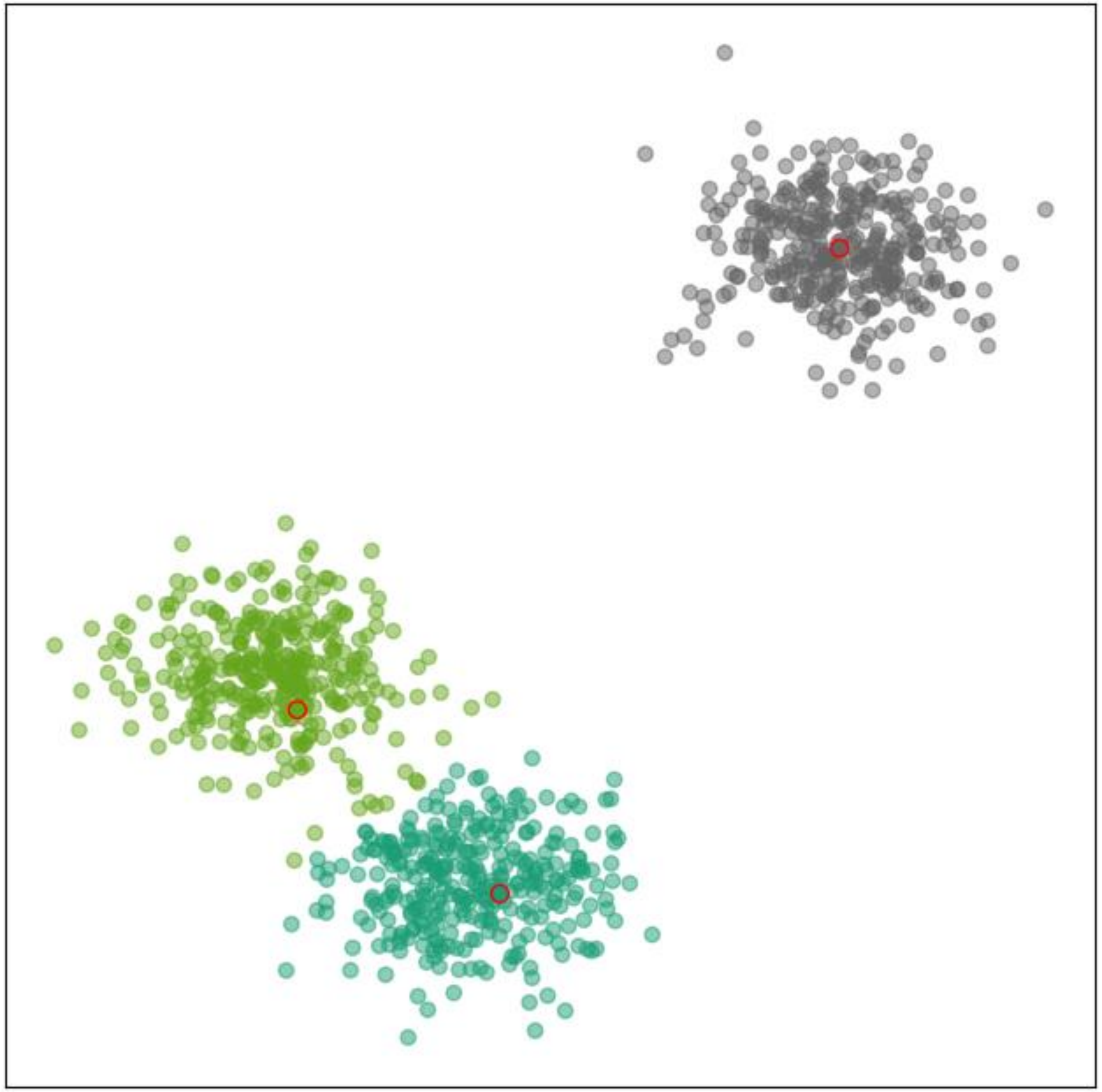
K-means clustering

- Specify k number of clusters *a priori*
- Randomly split data points amongst clusters
  1. **Calculate mean value of each cluster: *cluster centre***
  2. Reassign data points to closest cluster centre
  3. Repeat to convergence



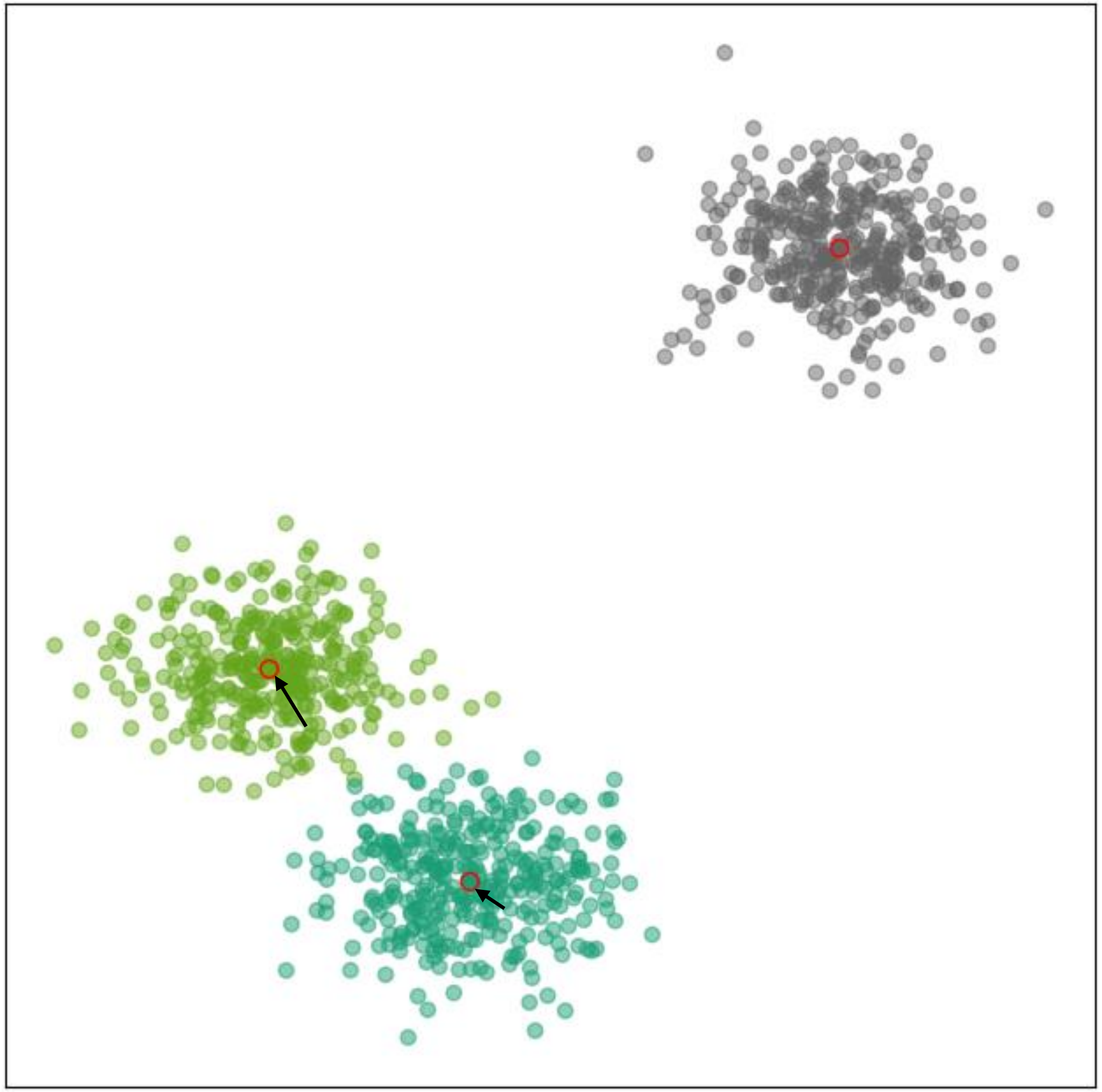
K-means clustering

- Specify k number of clusters *a priori*
- Randomly split data points amongst clusters
  1. Calculate mean value of each cluster: *cluster centre*
  2. **Reassign data points to closest cluster centre**
  3. Repeat to convergence



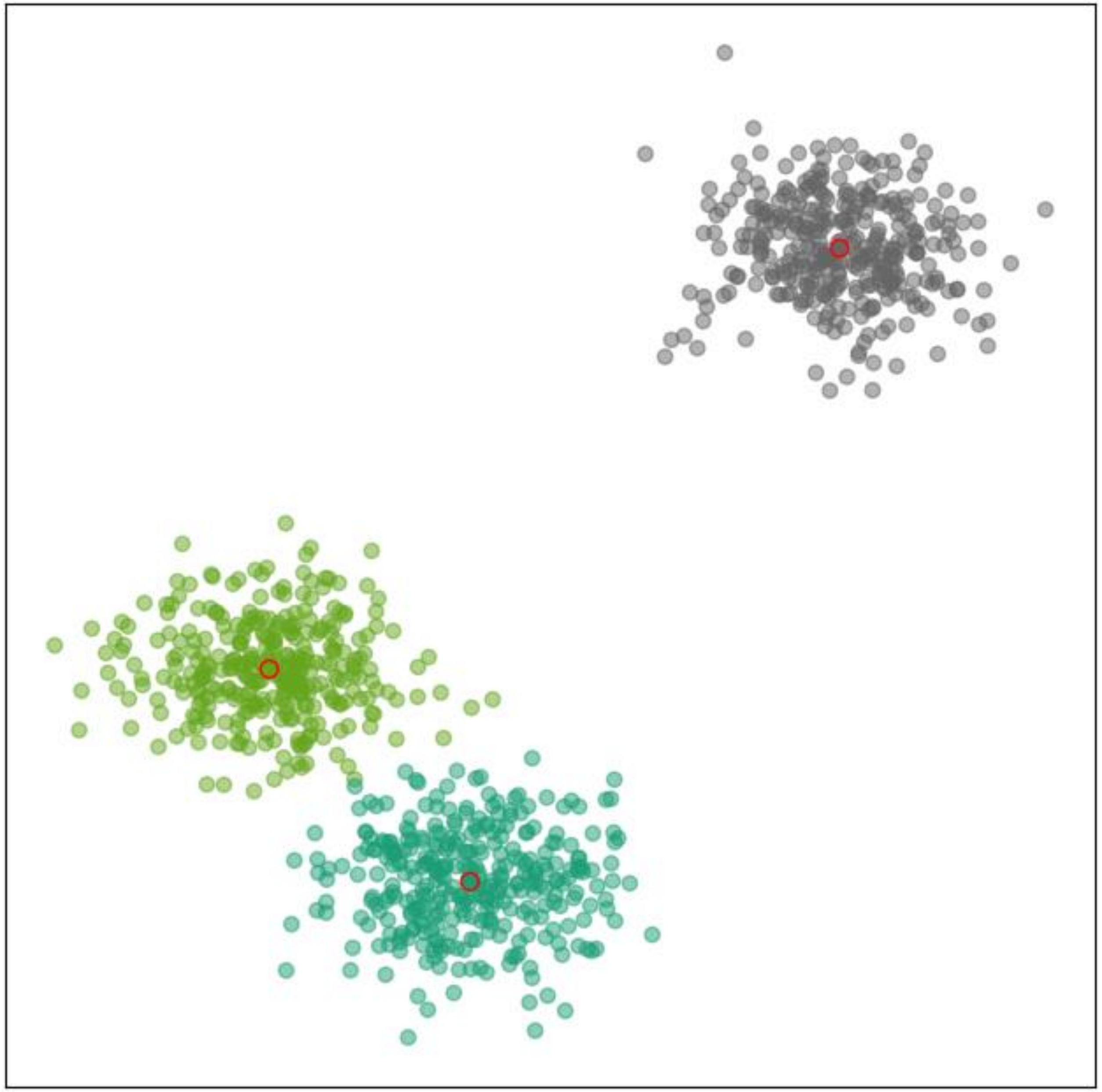
K-means clustering

- Specify k number of clusters *a priori*
- Randomly split data points amongst clusters
  1. **Calculate mean value of each cluster: *cluster centre***
  2. Reassign data points to closest cluster centre
  3. Repeat to convergence



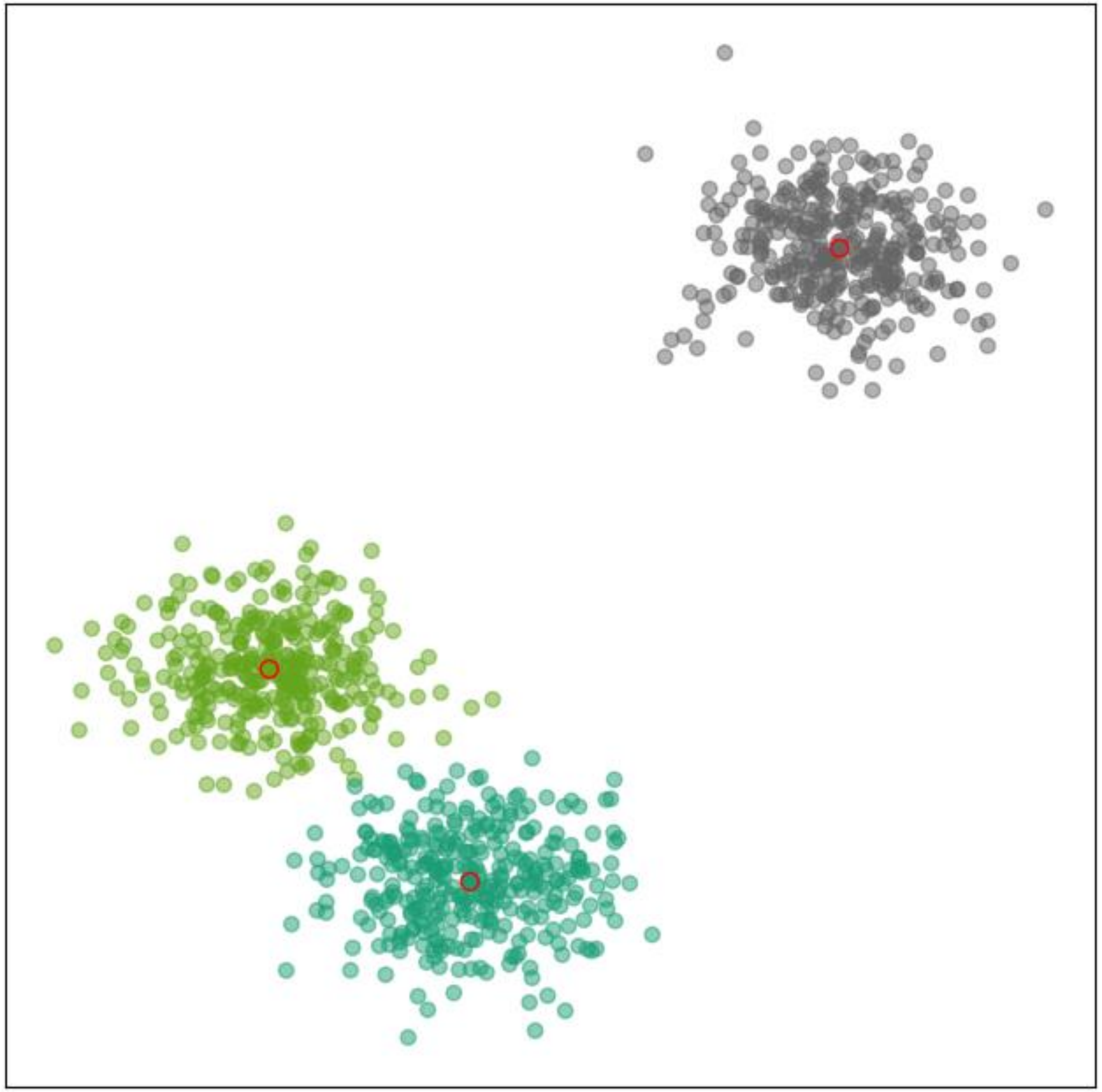
K-means clustering

- Specify *k* number of clusters *a priori*
- Randomly split data points amongst clusters
  1. Calculate mean value of each cluster: *cluster centre*
  2. **Reassign data points to closest cluster centre**
  3. Repeat to convergence



K-means clustering

- Specify k number of clusters *a priori*
- Randomly split data points amongst clusters
  1. Calculate mean value of each cluster: *cluster centre*
  2. Reassign data points to closest cluster centre
  3. Repeat to **convergence**



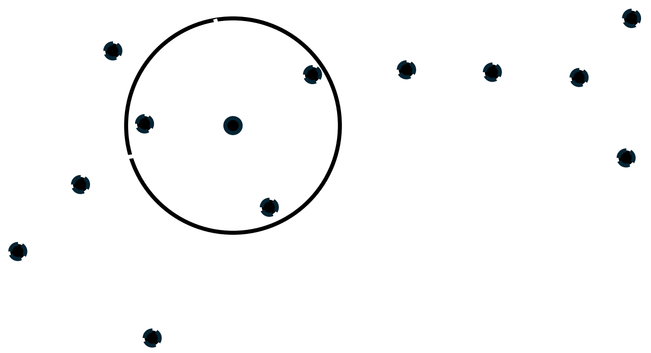
DBSCAN

- Density-Based Spatial Clustering of Applications with Noise
- Identifies cluster points based on neighbourhood radius ε and a minimum number of samples in the neighbourhood

https://scikit-learn.org/stable/modules/clustering.html

DBSCAN

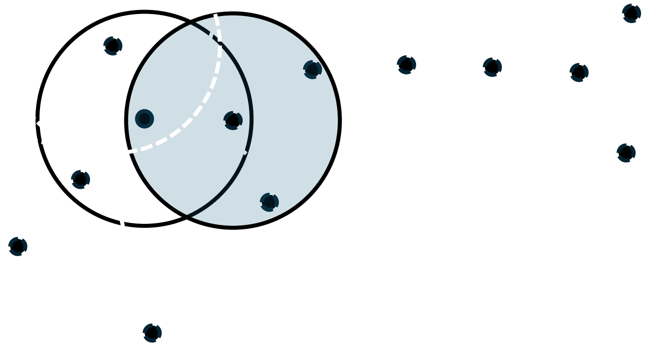
- Density-Based Spatial Clustering of Applications with Noise
- Identifies cluster points based on neighbourhood radius ε and a minimum number of samples in the neighbourhood



https://scikit-learn.org/stable/modules/clustering.html

DBSCAN

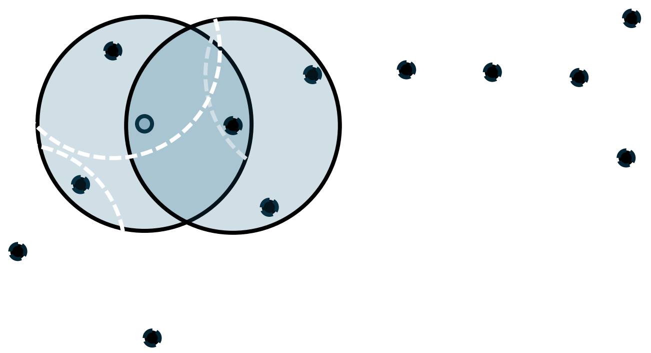
- Density-Based Spatial Clustering of Applications with Noise
- Identifies cluster points based on neighbourhood radius ε and a minimum number of samples in the neighbourhood



https://scikit-learn.org/stable/modules/clustering.html

DBSCAN

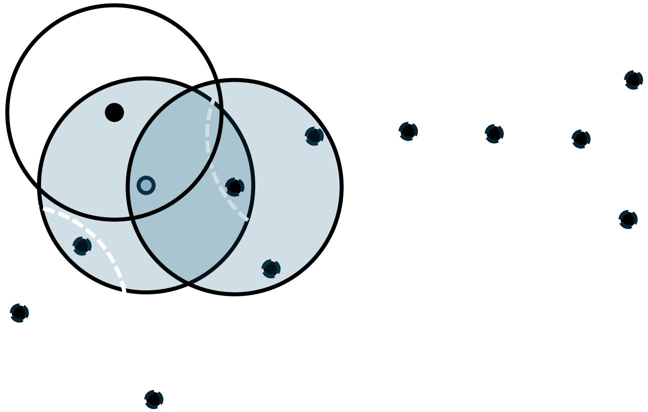
- Density-Based Spatial Clustering of Applications with Noise
- Identifies cluster points based on neighbourhood radius ε and a minimum number of samples in the neighbourhood



https://scikit-learn.org/stable/modules/clustering.html

DBSCAN

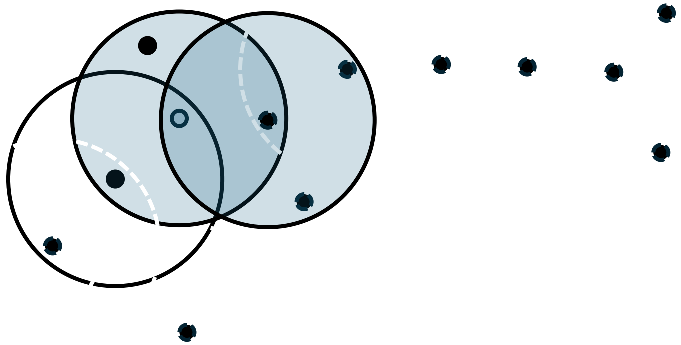
- Density-Based Spatial Clustering of Applications with Noise
- Identifies cluster points based on neighbourhood radius ε and a minimum number of samples in the neighbourhood



https://scikit-learn.org/stable/modules/clustering.html

DBSCAN

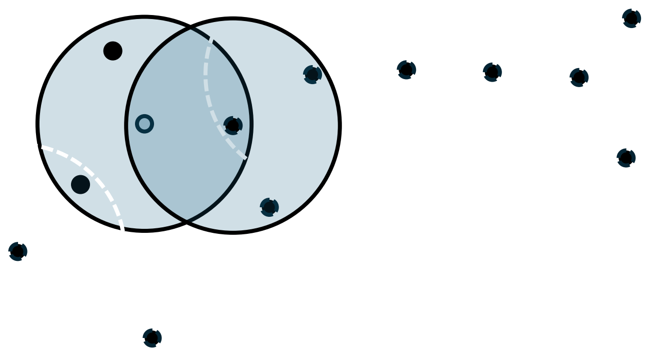
- Density-Based Spatial Clustering of Applications with Noise
- Identifies cluster points based on neighbourhood radius ε and a minimum number of samples in the neighbourhood



https://scikit-learn.org/stable/modules/clustering.html

DBSCAN

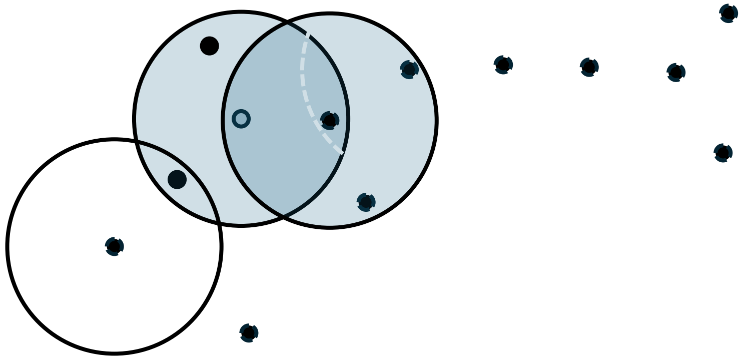
- Density-Based Spatial Clustering of Applications with Noise
- Identifies cluster points based on neighbourhood radius ε and a minimum number of samples in the neighbourhood



https://scikit-learn.org/stable/modules/clustering.html

DBSCAN

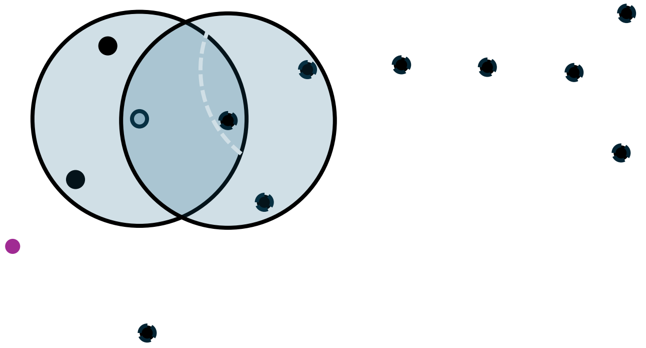
- Density-Based Spatial Clustering of Applications with Noise
- Identifies cluster points based on neighbourhood radius ε and a minimum number of samples in the neighbourhood



https://scikit-learn.org/stable/modules/clustering.html

DBSCAN

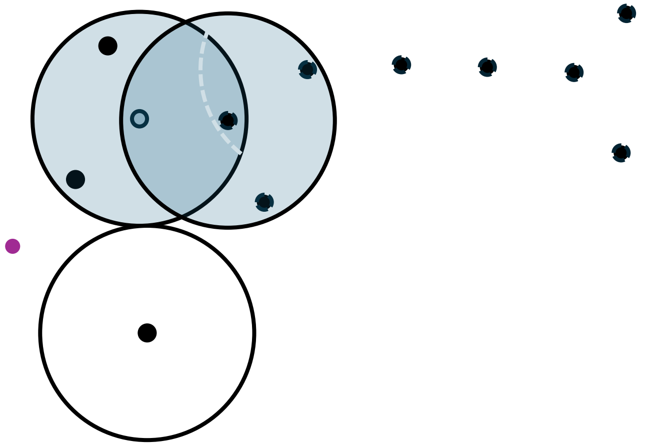
- Density-Based Spatial Clustering of Applications with Noise
- Identifies cluster points based on neighbourhood radius ε and a minimum number of samples in the neighbourhood



https://scikit-learn.org/stable/modules/clustering.html

DBSCAN

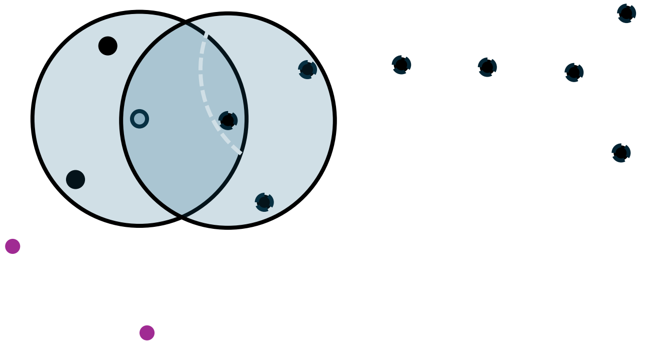
- Density-Based Spatial Clustering of Applications with Noise
- Identifies cluster points based on neighbourhood radius ε and a minimum number of samples in the neighbourhood



https://scikit-learn.org/stable/modules/clustering.html

DBSCAN

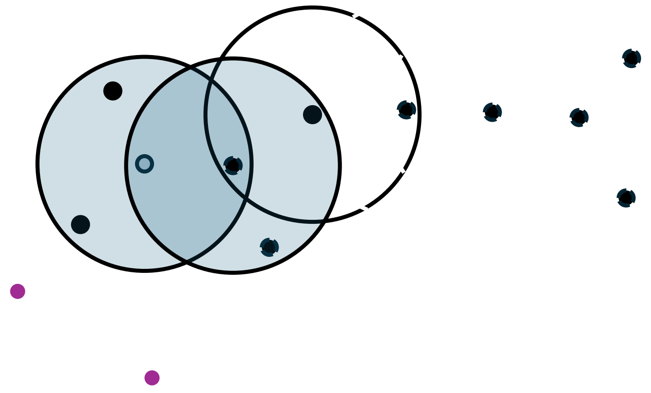
- Density-Based Spatial Clustering of Applications with Noise
- Identifies cluster points based on neighbourhood radius ε and a minimum number of samples in the neighbourhood



https://scikit-learn.org/stable/modules/clustering.html

DBSCAN

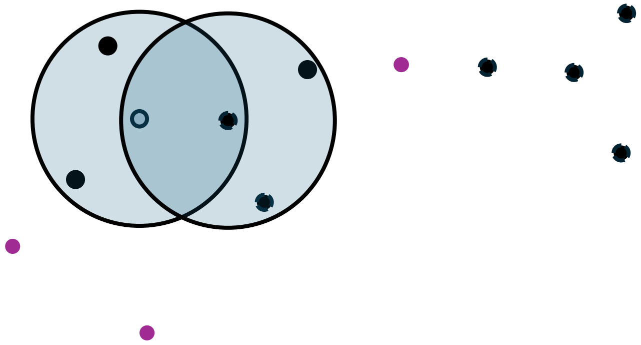
- Density-Based Spatial Clustering of Applications with Noise
- Identifies cluster points based on neighbourhood radius ε and a minimum number of samples in the neighbourhood



https://scikit-learn.org/stable/modules/clustering.html

DBSCAN

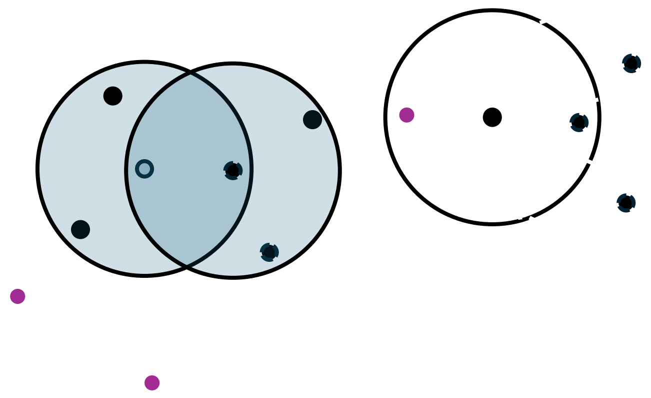
- Density-Based Spatial Clustering of Applications with Noise
- Identifies cluster points based on neighbourhood radius ε and a minimum number of samples in the neighbourhood



https://scikit-learn.org/stable/modules/clustering.html

DBSCAN

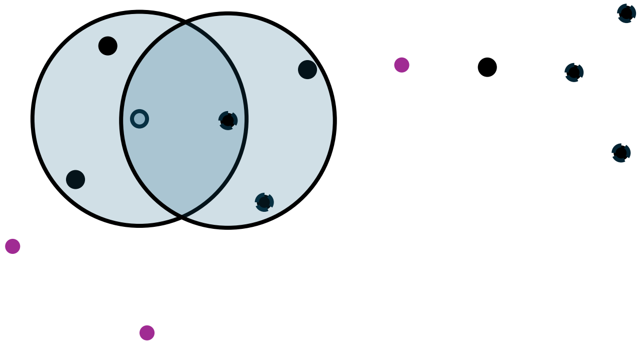
- Density-Based Spatial Clustering of Applications with Noise
- Identifies cluster points based on neighbourhood radius ε and a minimum number of samples in the neighbourhood



https://scikit-learn.org/stable/modules/clustering.html

DBSCAN

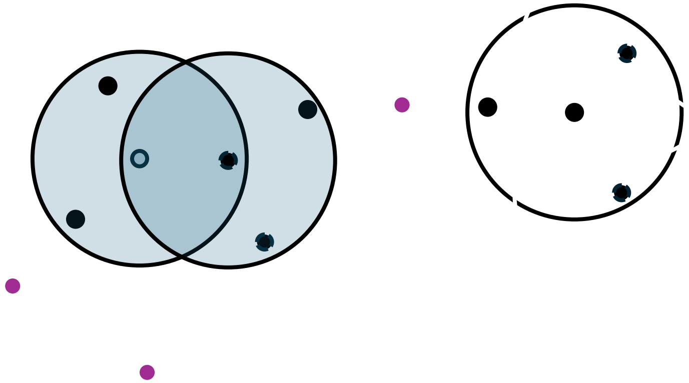
- Density-Based Spatial Clustering of Applications with Noise
- Identifies cluster points based on neighbourhood radius ε and a minimum number of samples in the neighbourhood



https://scikit-learn.org/stable/modules/clustering.html

DBSCAN

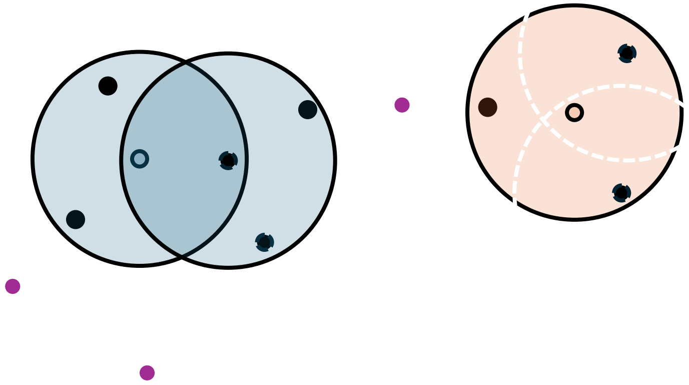
- Density-Based Spatial Clustering of Applications with Noise
- Identifies cluster points based on neighbourhood radius ε and a minimum number of samples in the neighbourhood



https://scikit-learn.org/stable/modules/clustering.html

DBSCAN

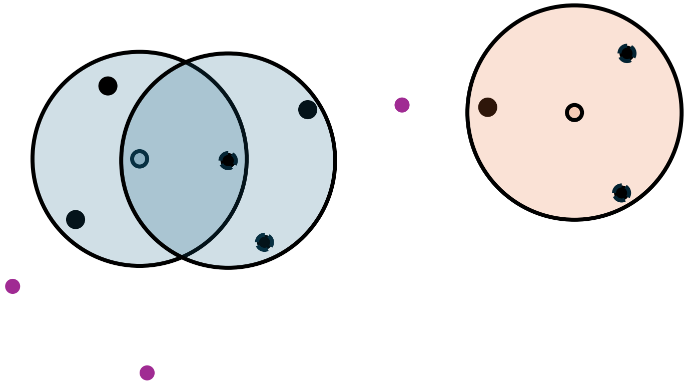
- Density-Based Spatial Clustering of Applications with Noise
- Identifies cluster points based on neighbourhood radius ε and a minimum number of samples in the neighbourhood



https://scikit-learn.org/stable/modules/clustering.html

DBSCAN

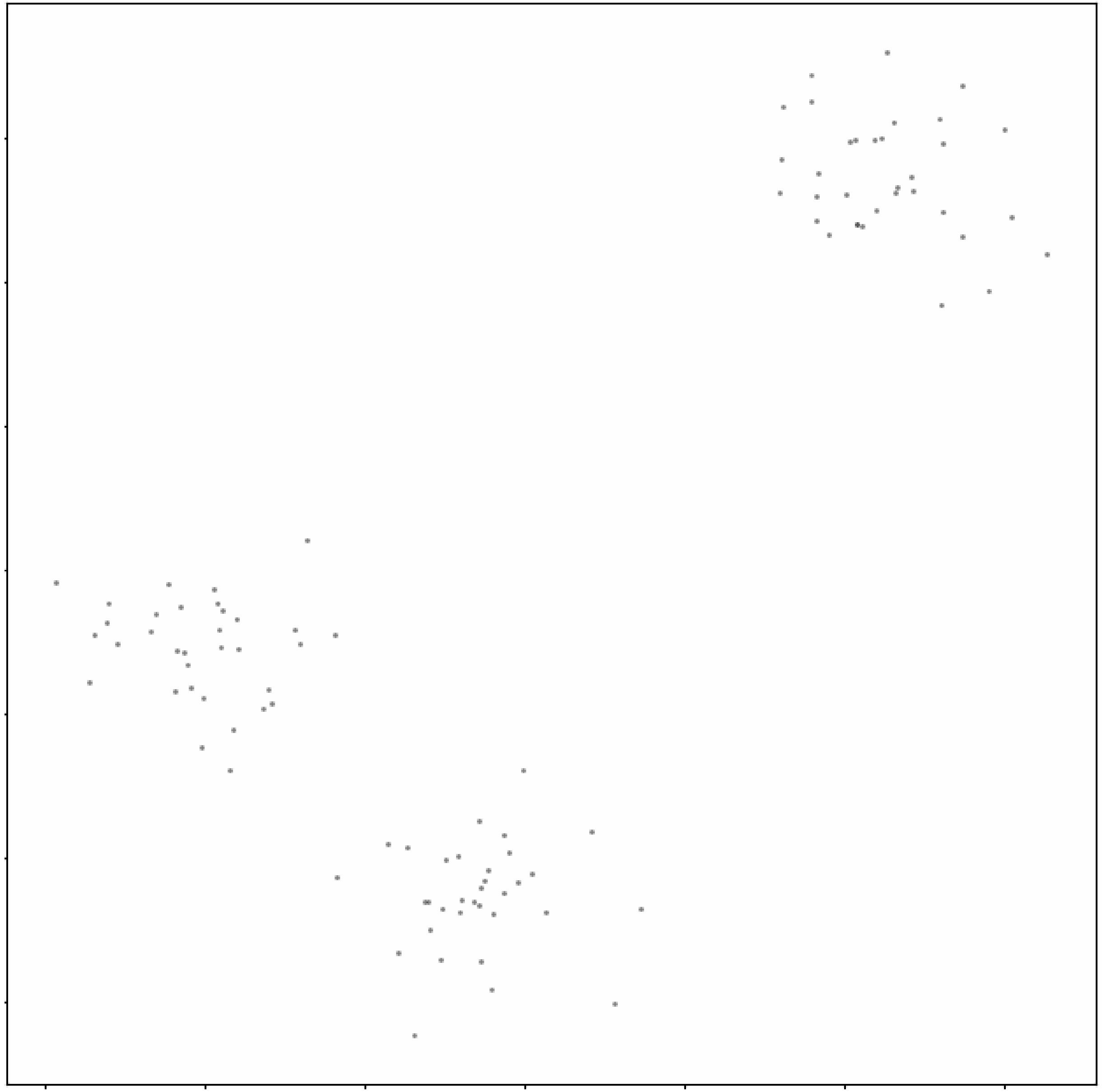
- Density-Based Spatial Clustering of Applications with Noise
- Identifies cluster points based on neighbourhood radius ε and a minimum number of samples in the neighbourhood



https://scikit-learn.org/stable/modules/clustering.html

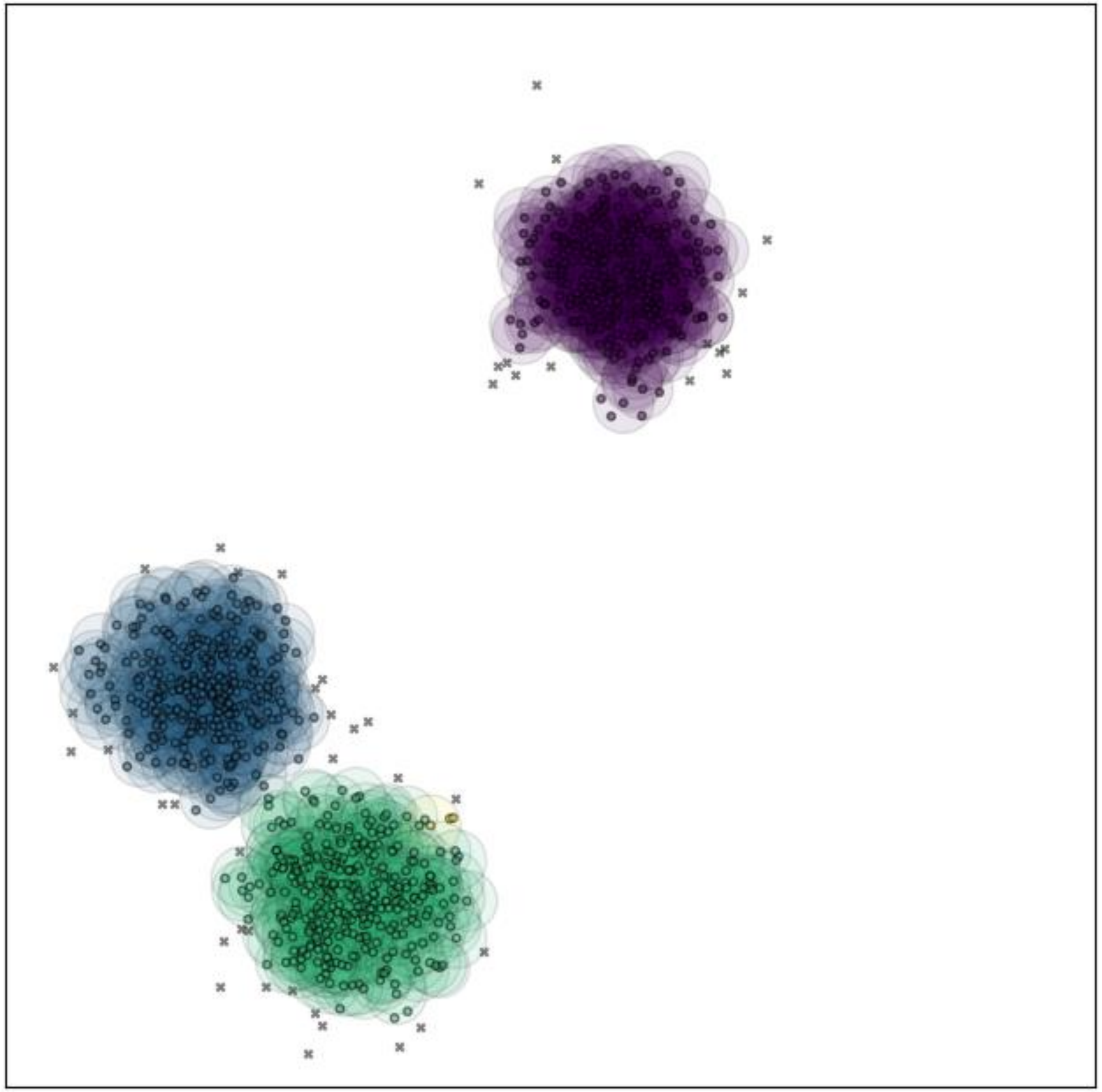
DBSCAN

- Density-Based Spatial Clustering of Applications with Noise
- Identifies cluster points based on neighbourhood radius ε and a minimum number of samples in the neighbourhood



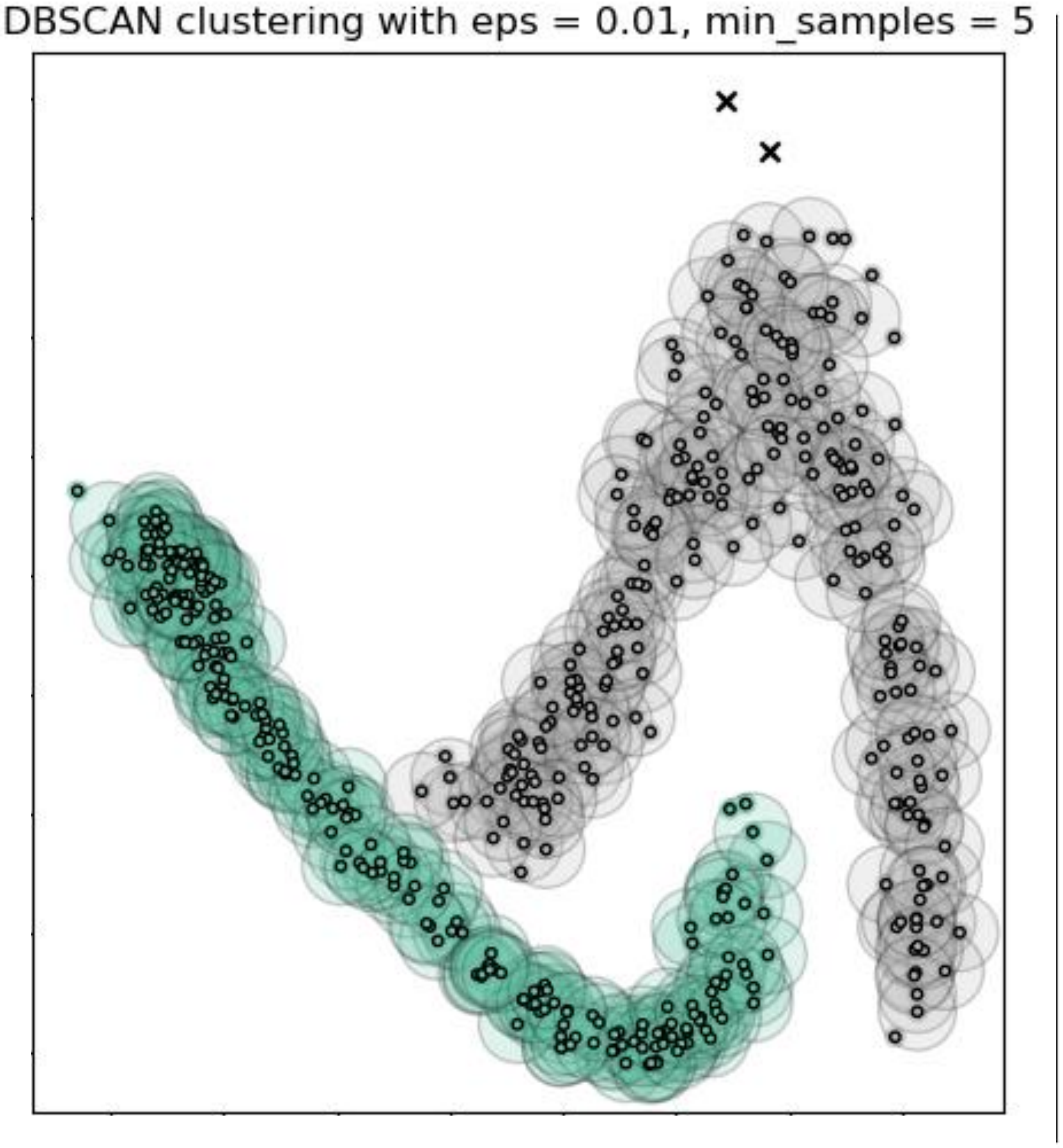
DBSCAN

- Density-Based Spatial Clustering of Applications with Noise
- Identifies cluster points based on neighbourhood radius ε and a minimum number of samples in the neighbourhood

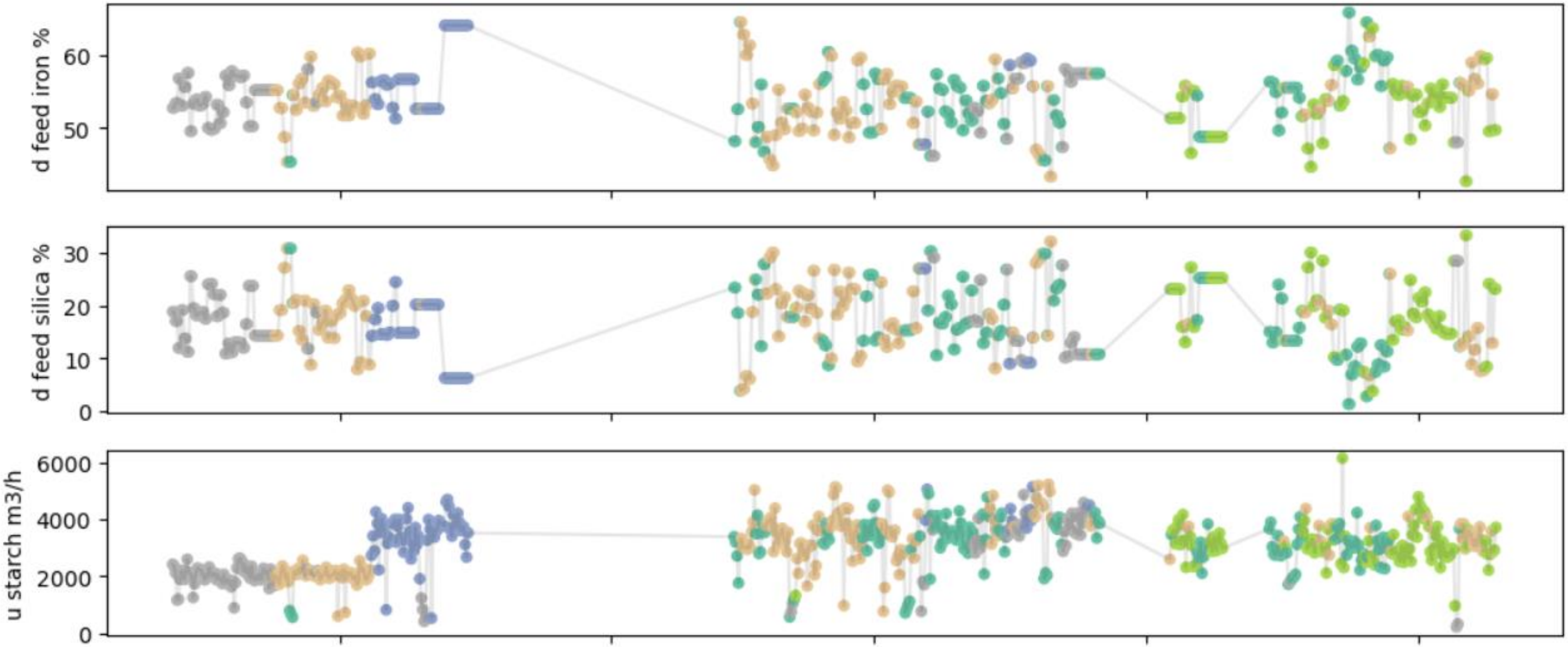


DBSCAN

- Density-Based Spatial Clustering of Applications with Noise
- Identifies cluster points based on neighbourhood radius ε and a minimum number of samples in the neighbourhood
- Capable of identifying non-convex clusters

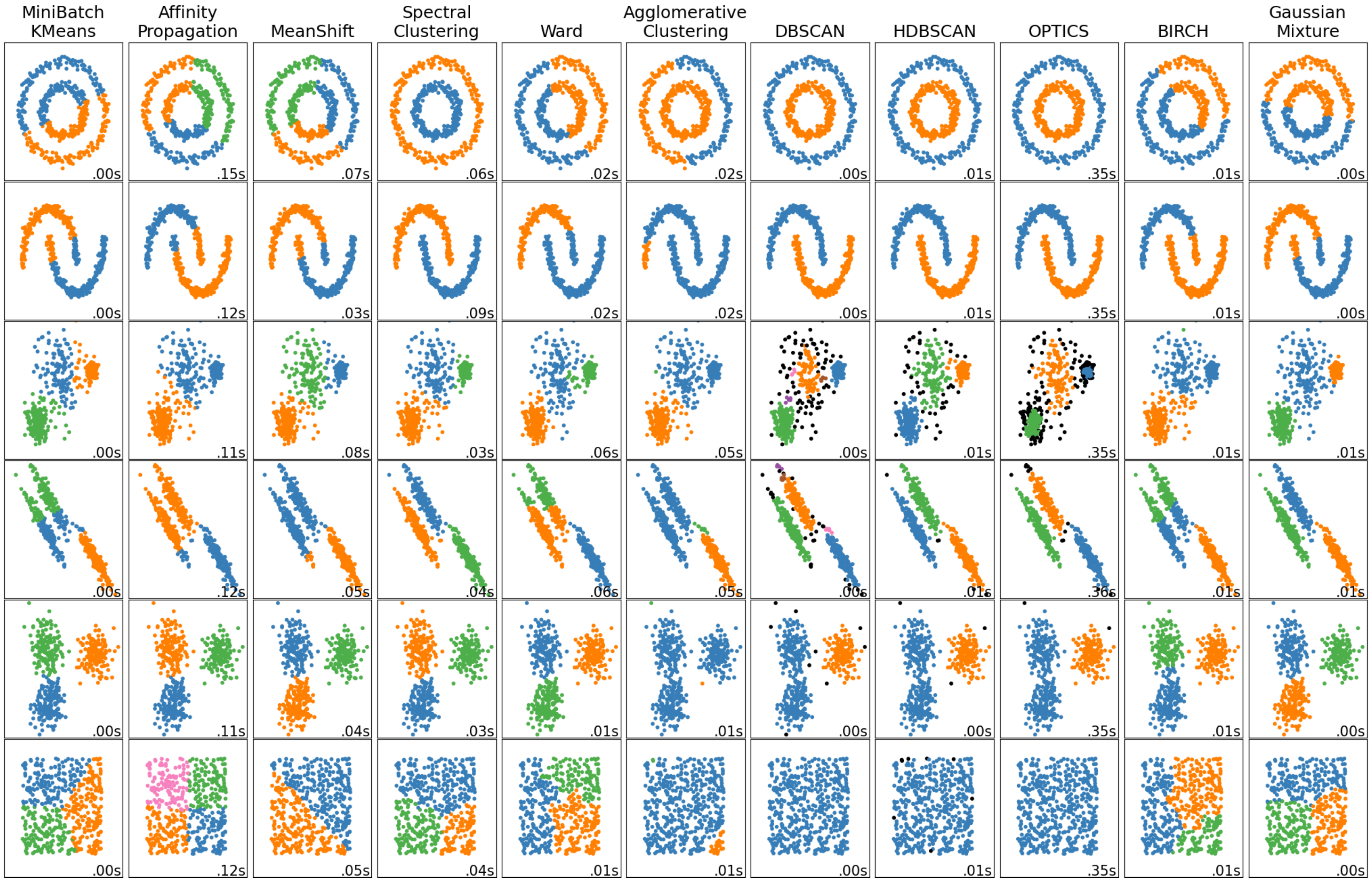


More informative time series plots



Clustering

- Many other methods available on scikit-learn

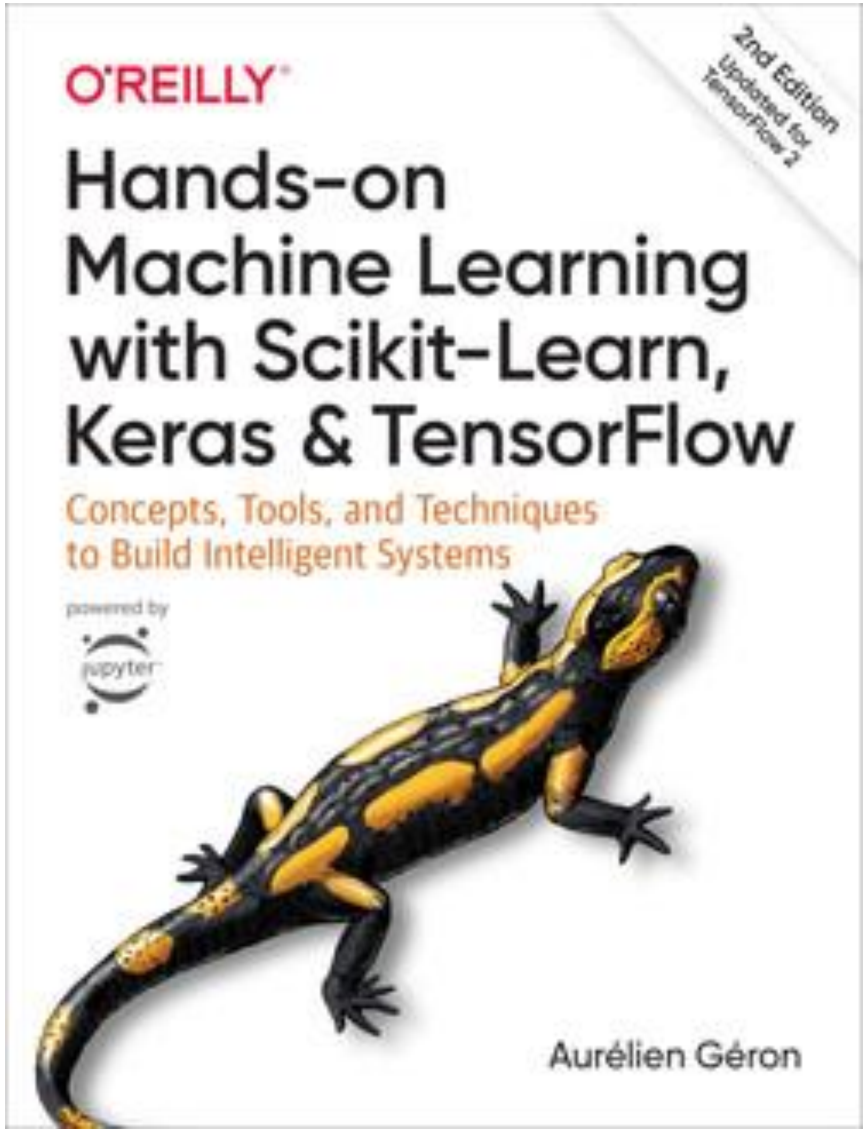


https://scikit-learn.org/stable/modules/clustering.html

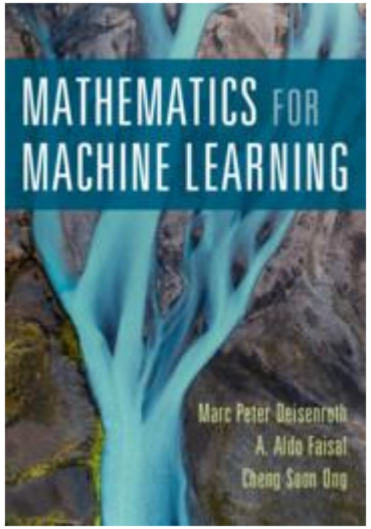
Resources

scikit learn

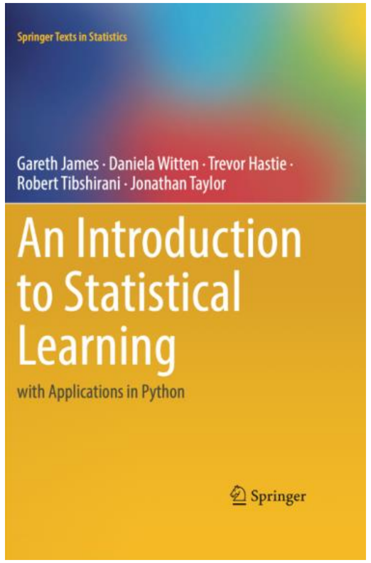
https://scikit-learn.org/stable/unsupervised_learning.html


O'REILLY
2nd Edition
Updated for TensorFlow 2
Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow
Concepts, Tools, and Techniques to Build Intelligent Systems
powered by
jupyter
Aurélien Géron

https://www.oreilly.com/library/view/hands-on-machine-learning/9781492032632/



https://mml-book.github.io/


Springer Texts in Statistics
Gareth James · Daniela Witten · Trevor Hastie · Robert Tibshirani · Jonathan Taylor
An Introduction to Statistical Learning
with Applications in Python
Springer

https://www.statlearning.com/