

Clustering

South-African Council for Automation and Control

Exploratory Data Analysis workshop

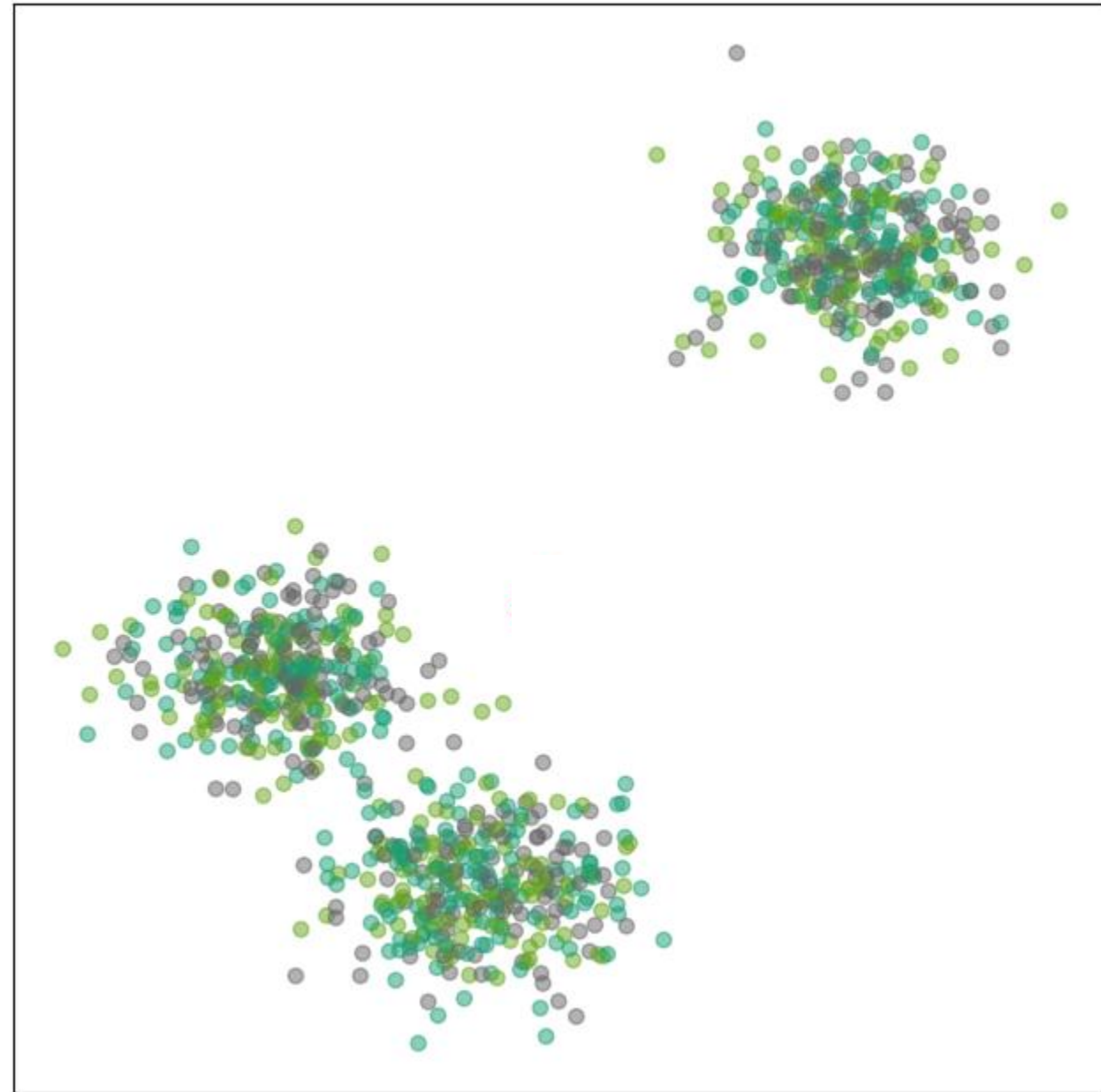
May 2025

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

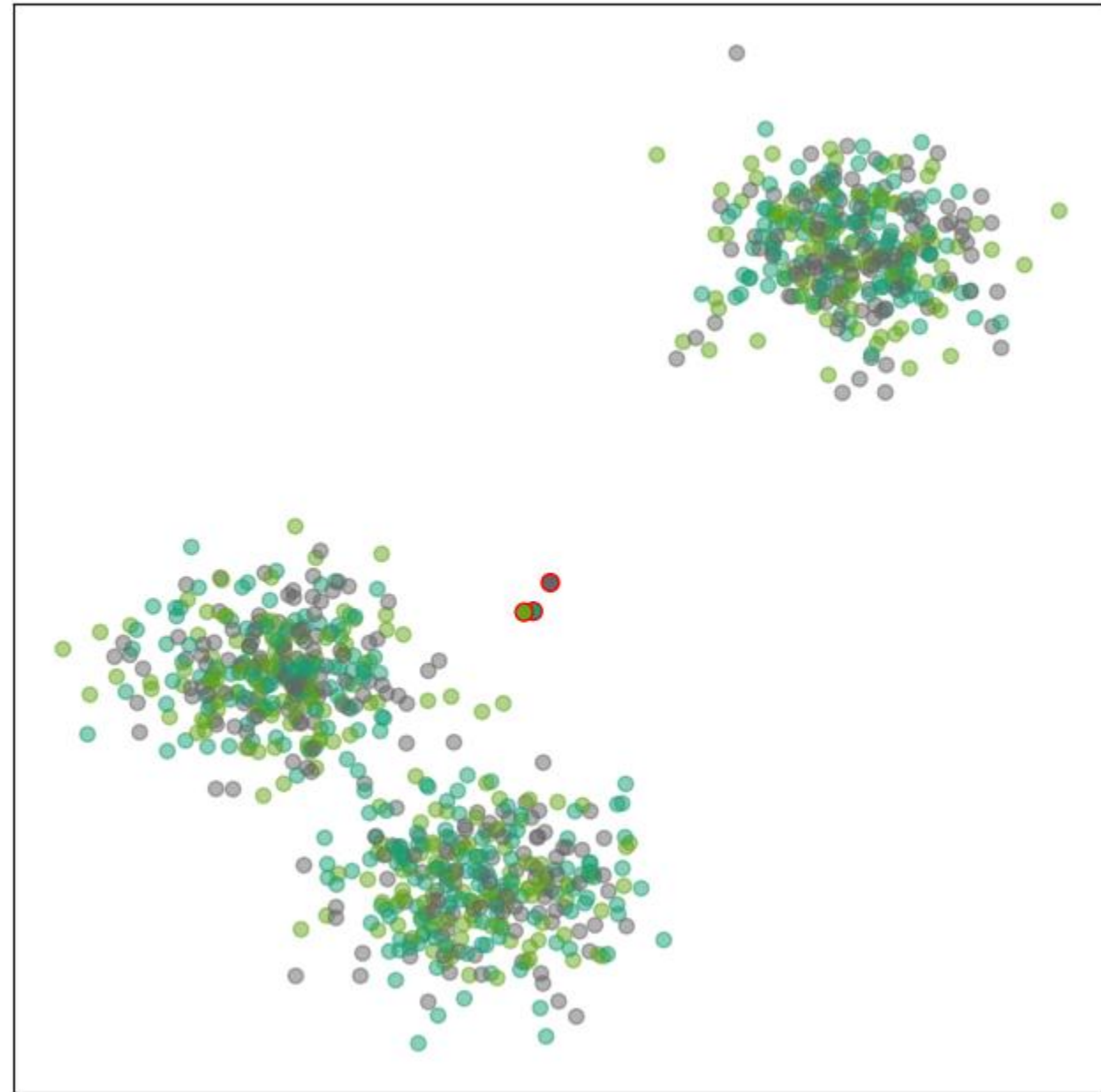
K-means clustering

- 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



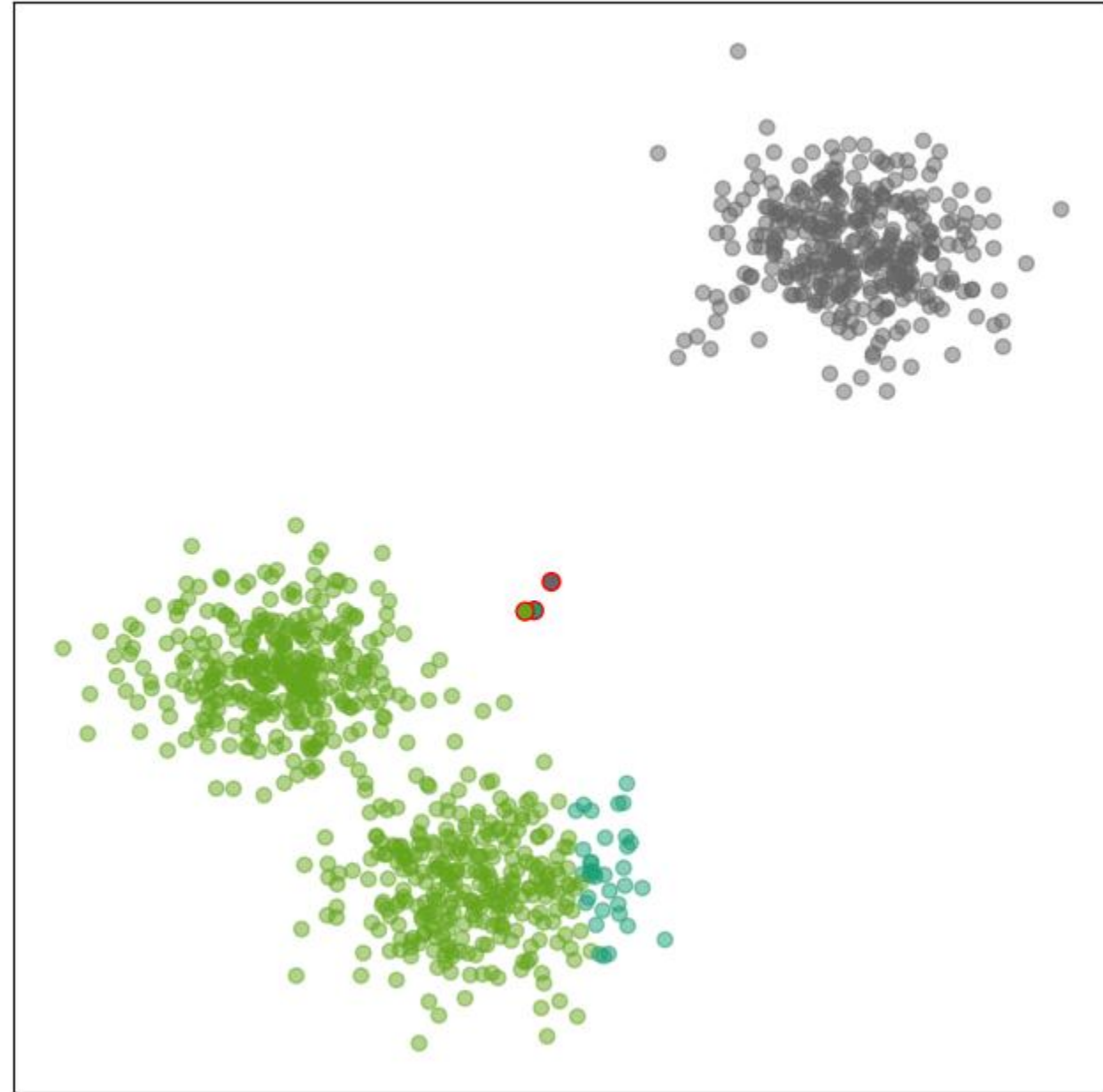
K-means clustering

- 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



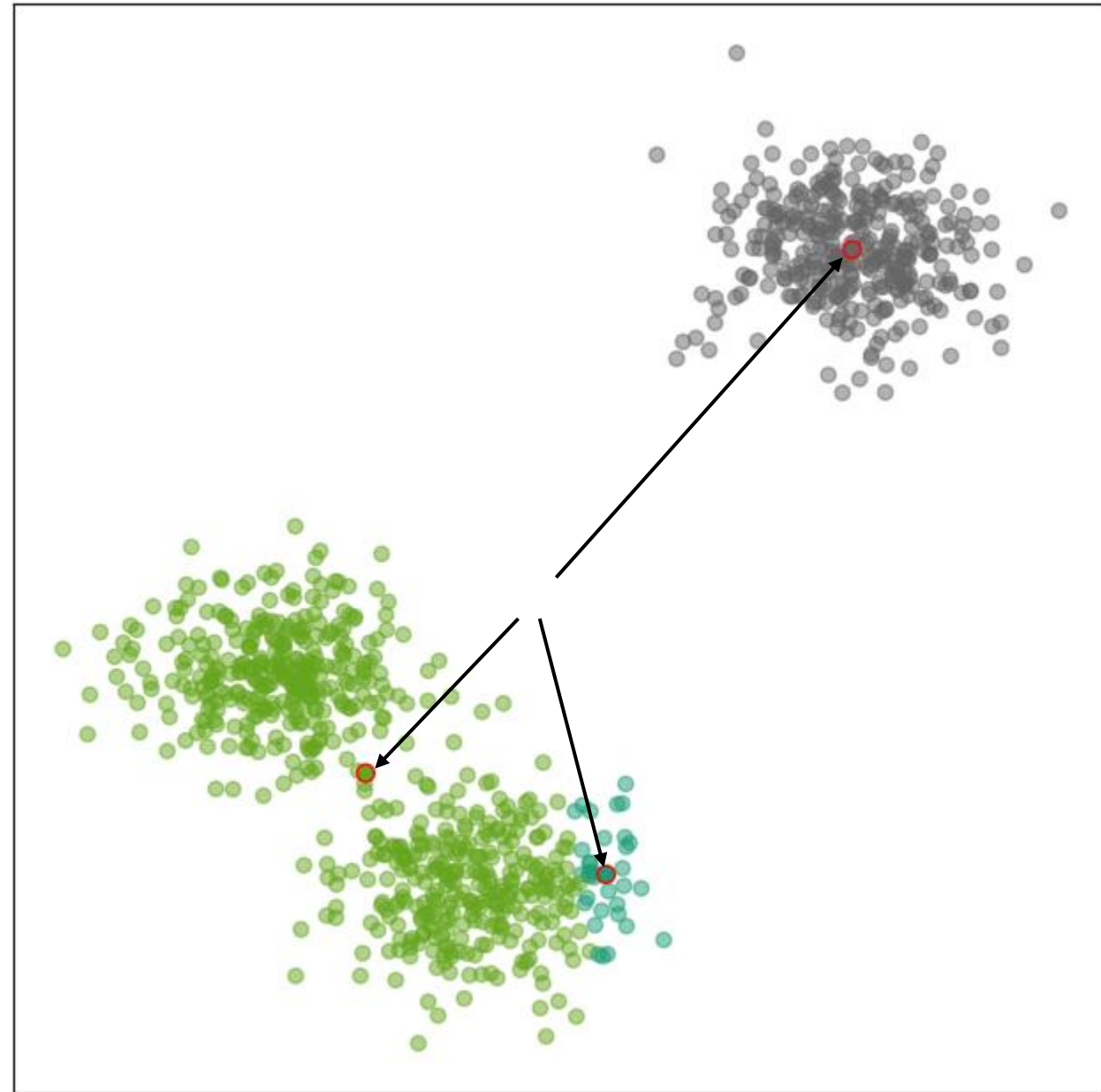
K-means clustering

- 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



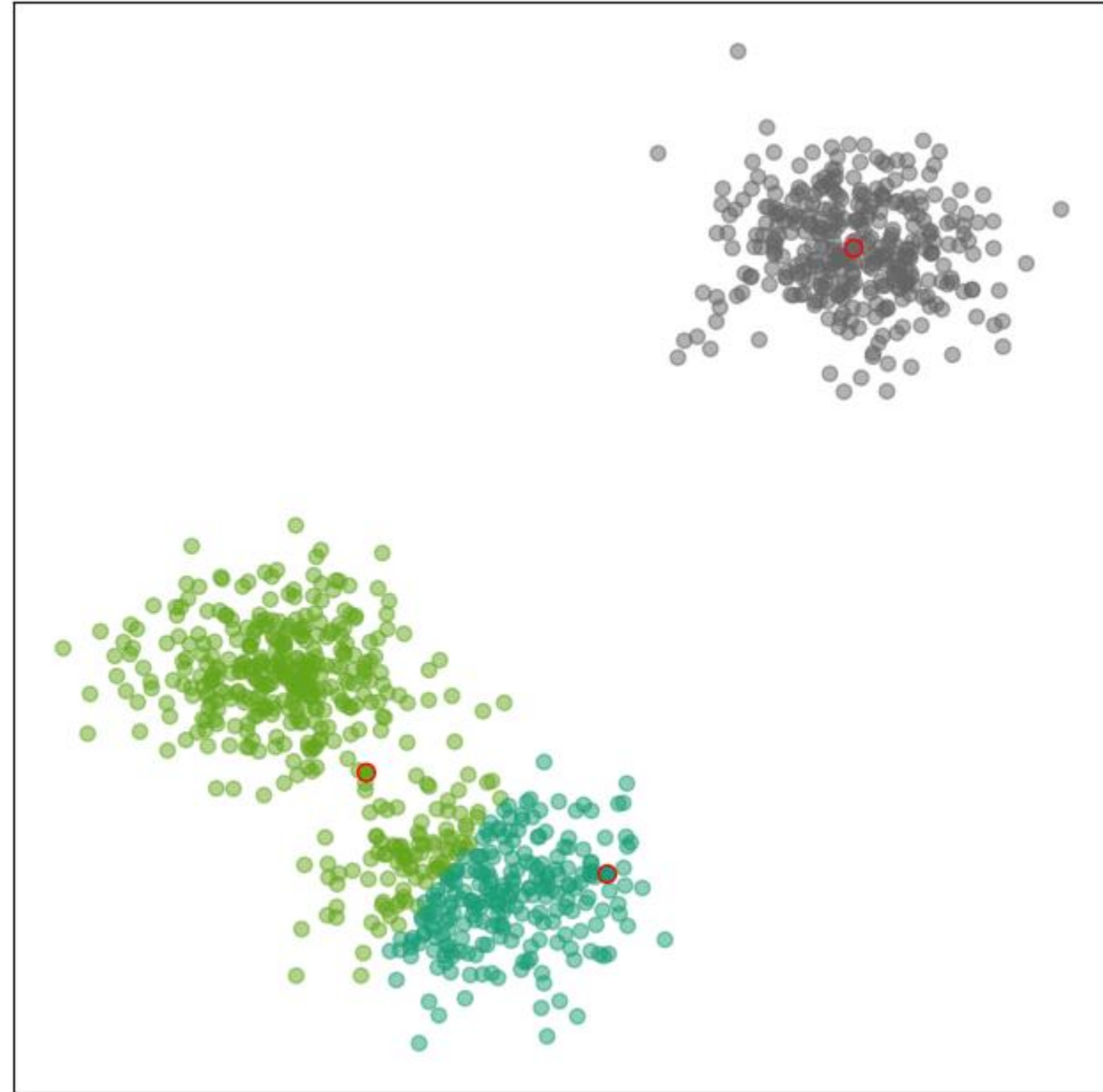
K-means clustering

- 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



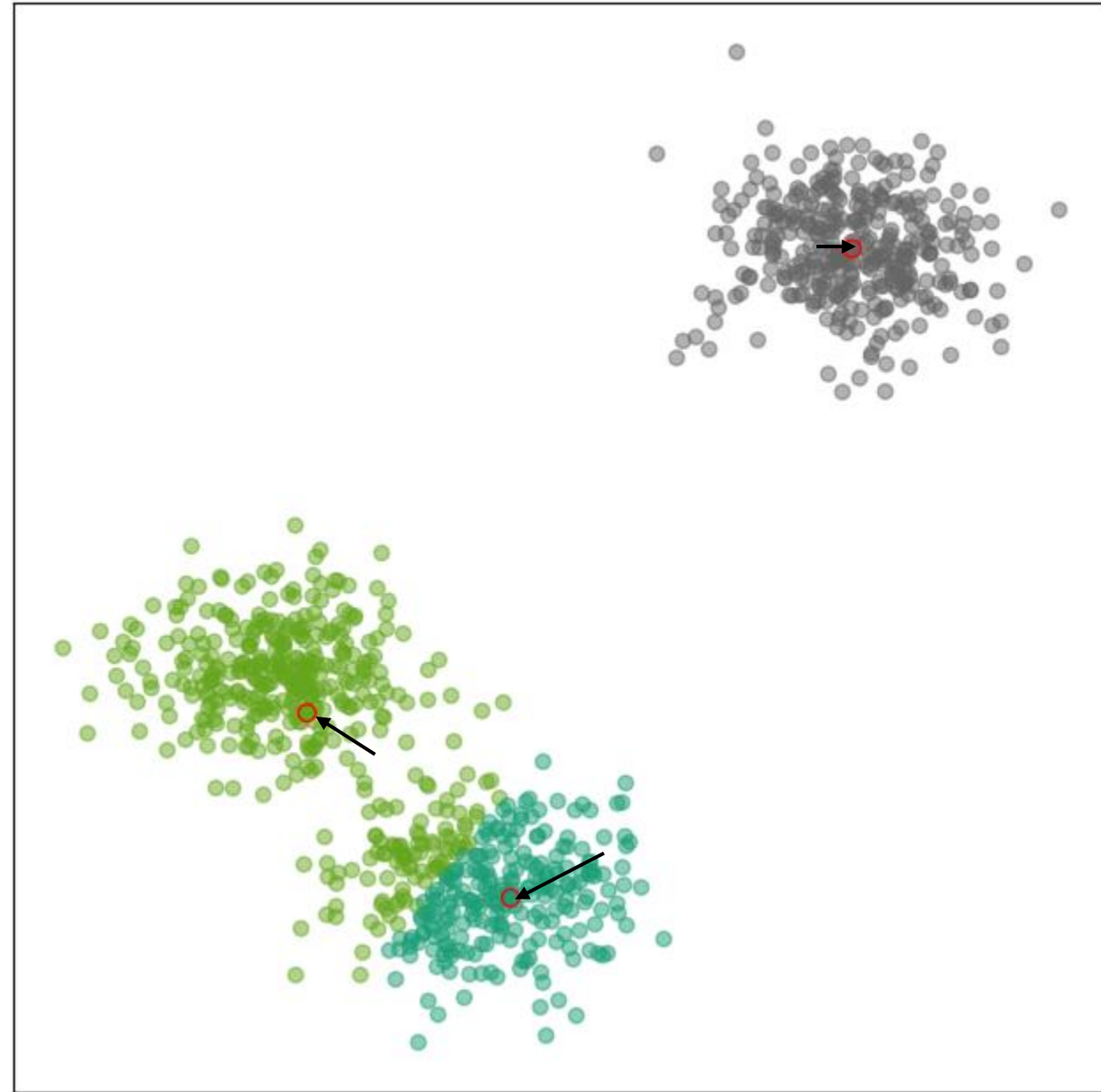
K-means clustering

- 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



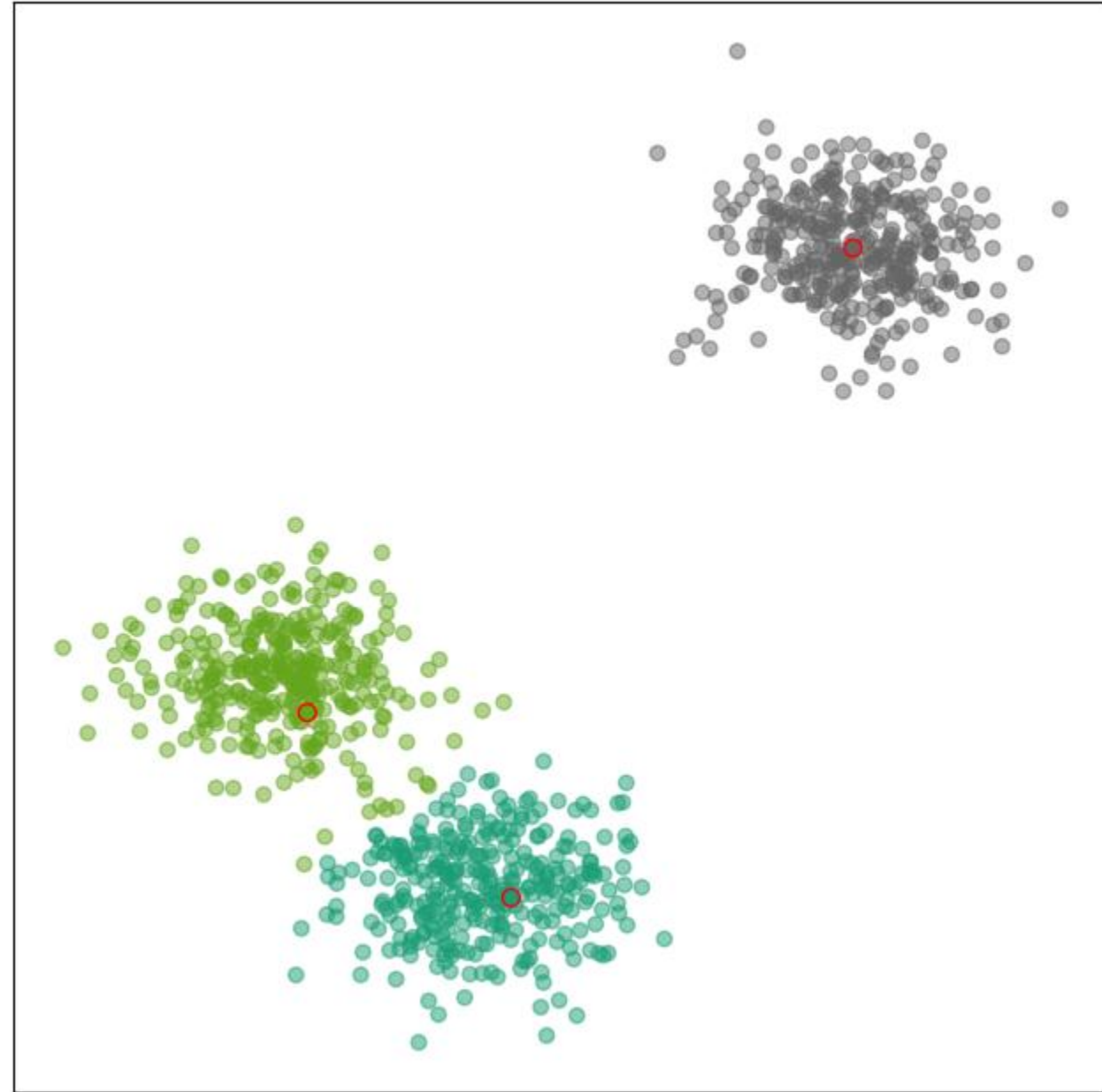
K-means clustering

- 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



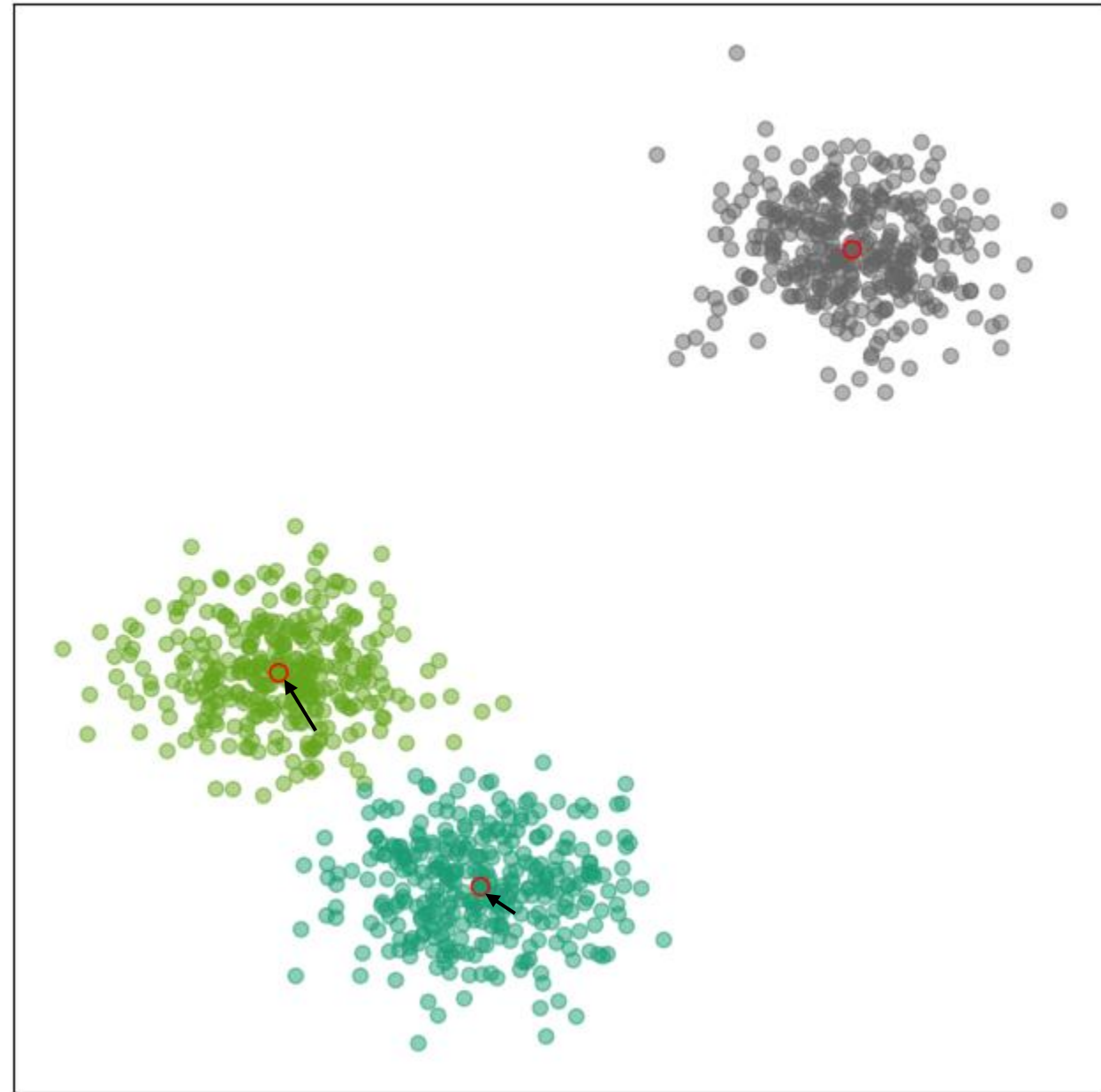
K-means clustering

- 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



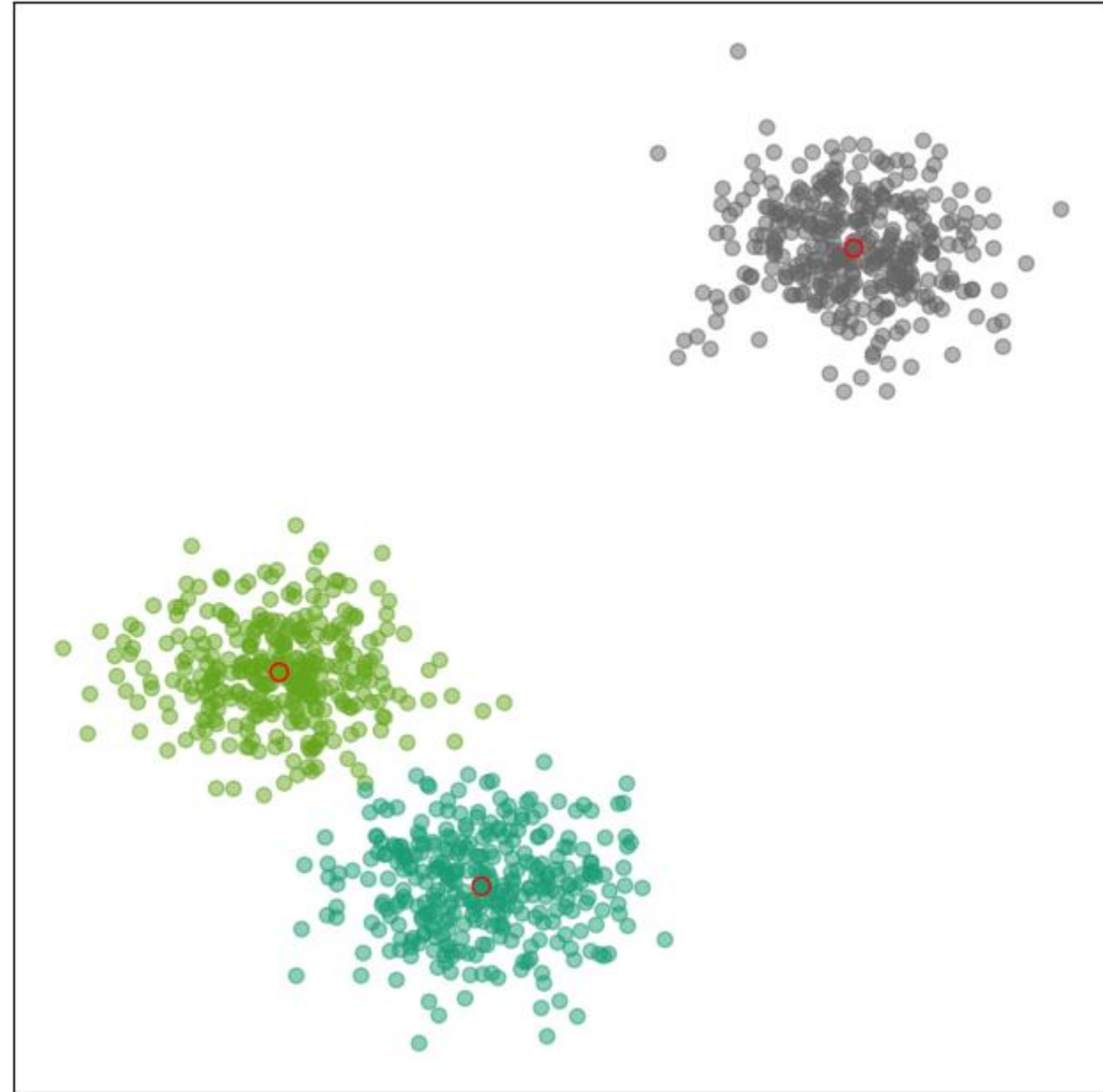
K-means clustering

- 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



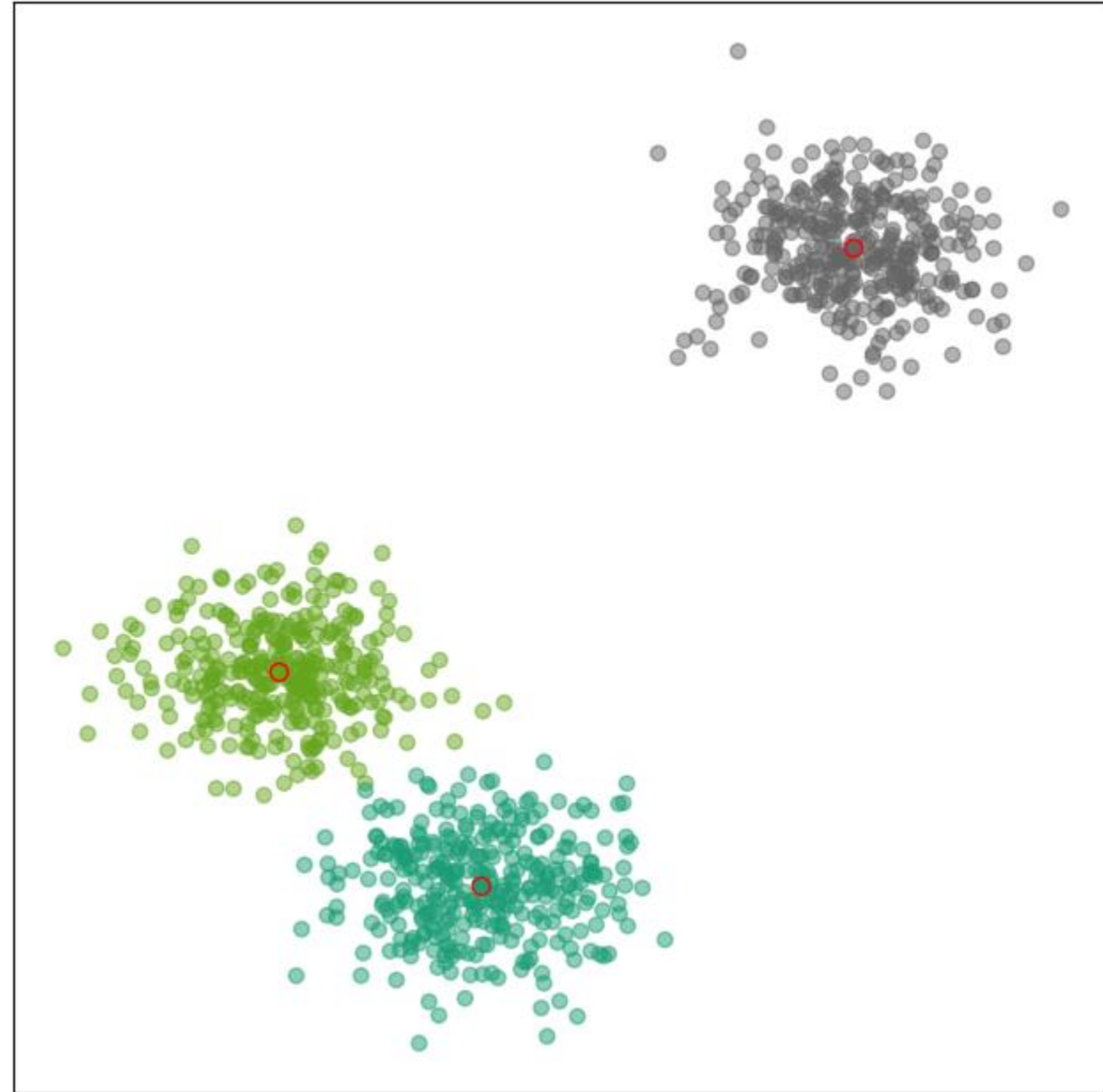
K-means clustering

- 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**



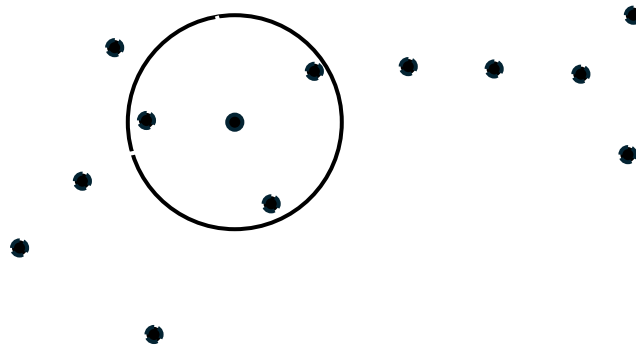
DBSCAN

- 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



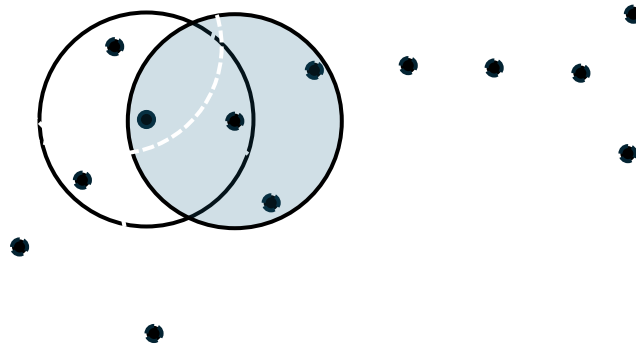
DBSCAN

- 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



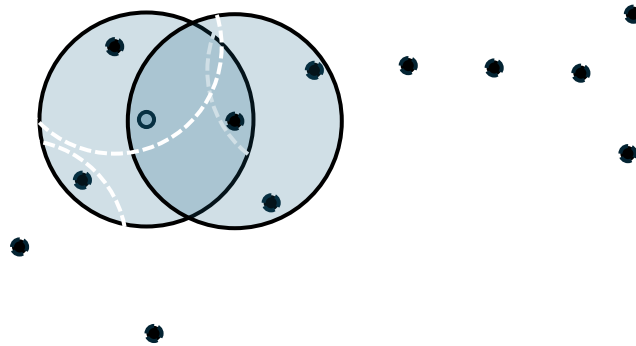
DBSCAN

- 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



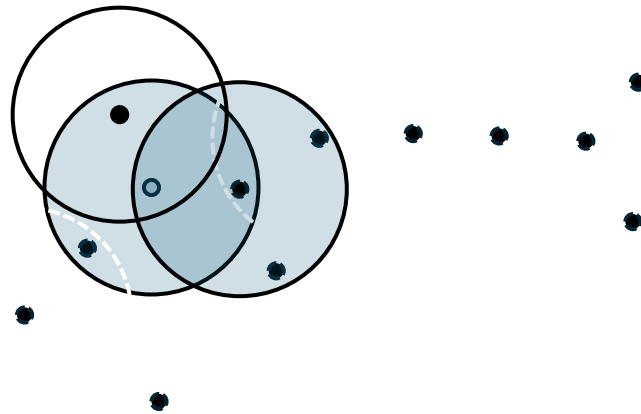
DBSCAN

- 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



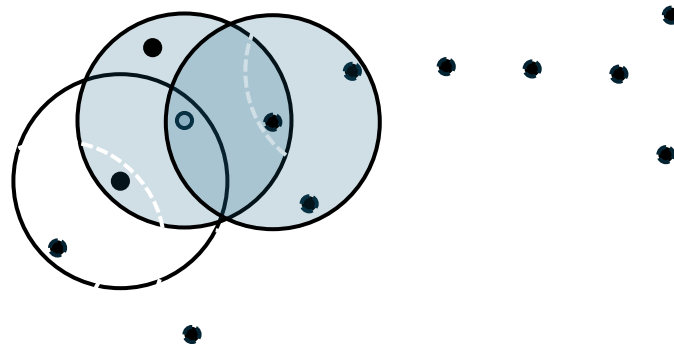
DBSCAN

- 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



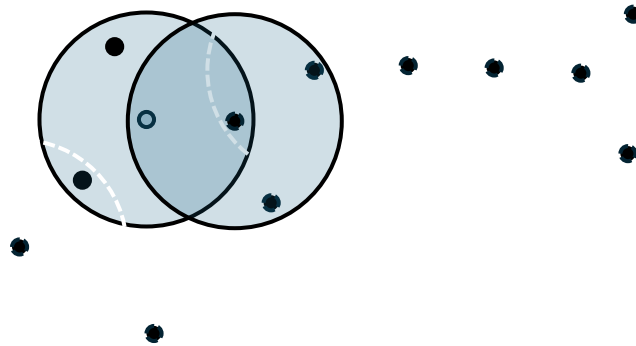
DBSCAN

- 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



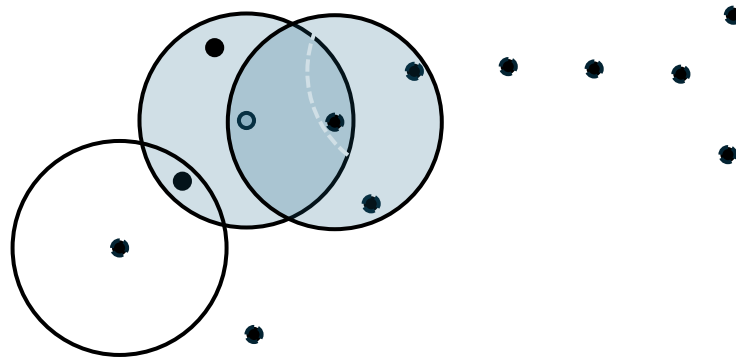
DBSCAN

- 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



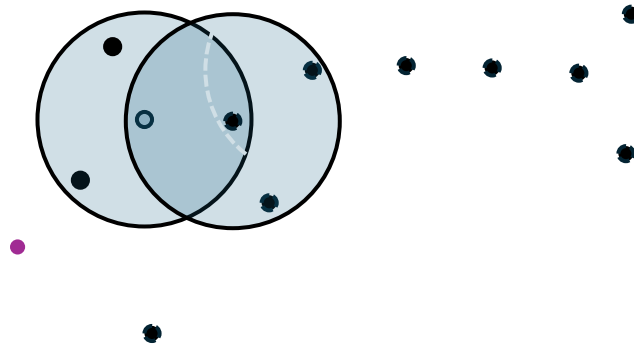
DBSCAN

- 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



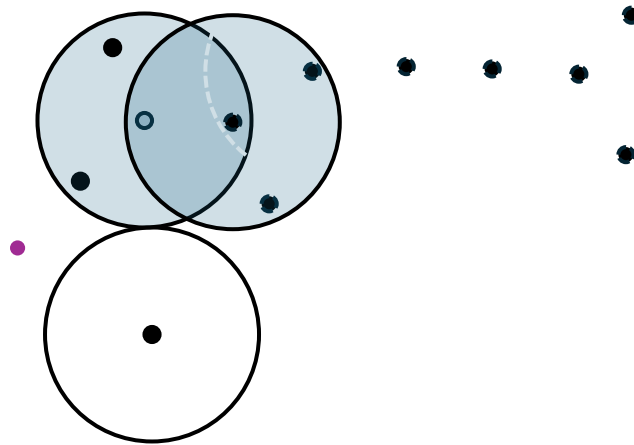
DBSCAN

- 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



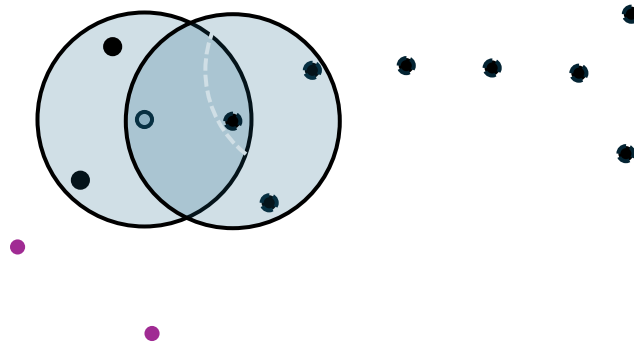
DBSCAN

- 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



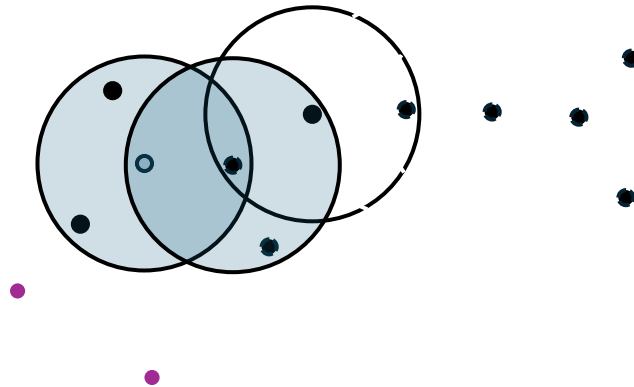
DBSCAN

- 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



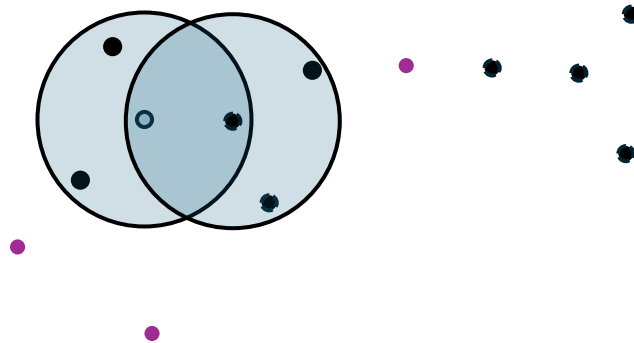
DBSCAN

- 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



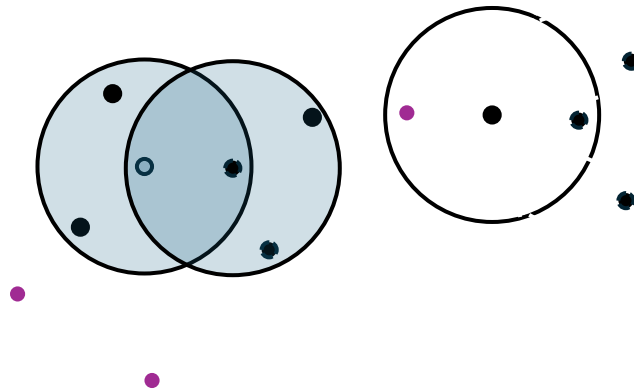
DBSCAN

- 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



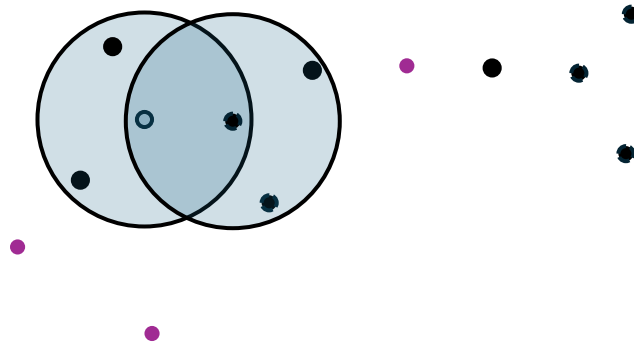
DBSCAN

- 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



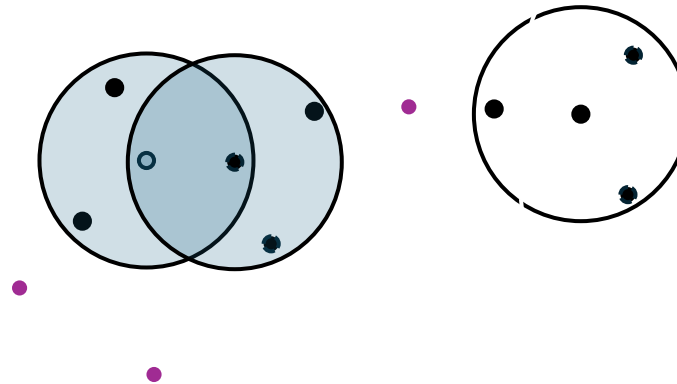
DBSCAN

- 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



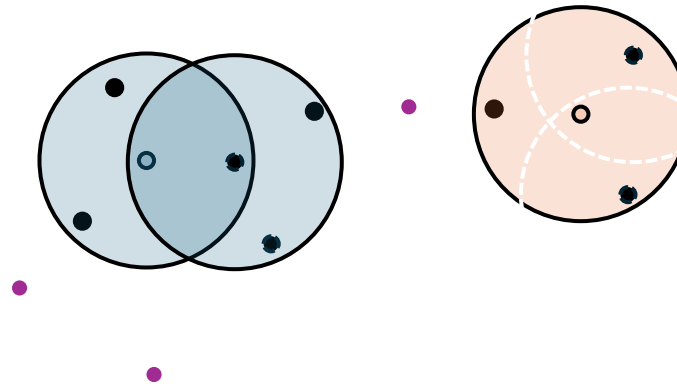
DBSCAN

- 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



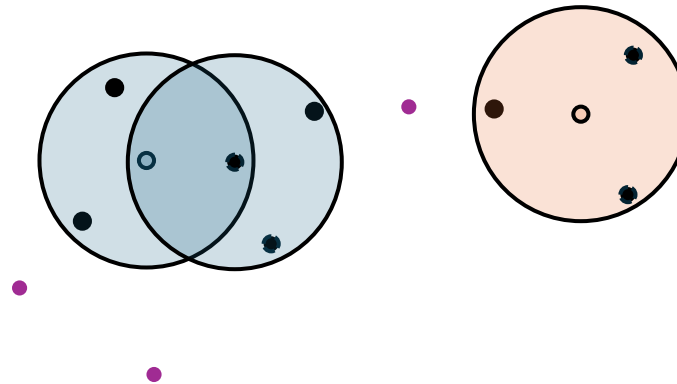
DBSCAN

- 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



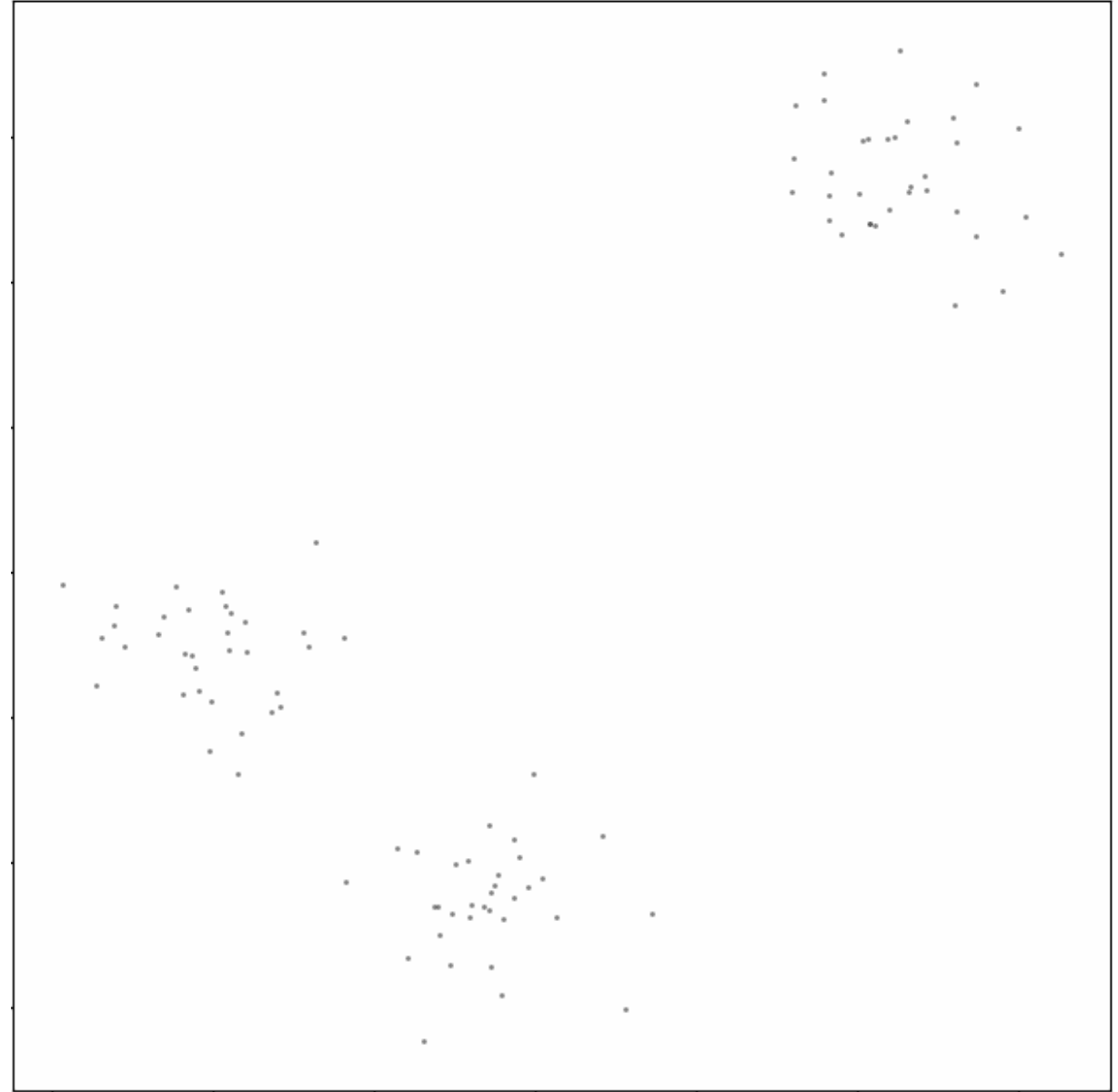
DBSCAN

- 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



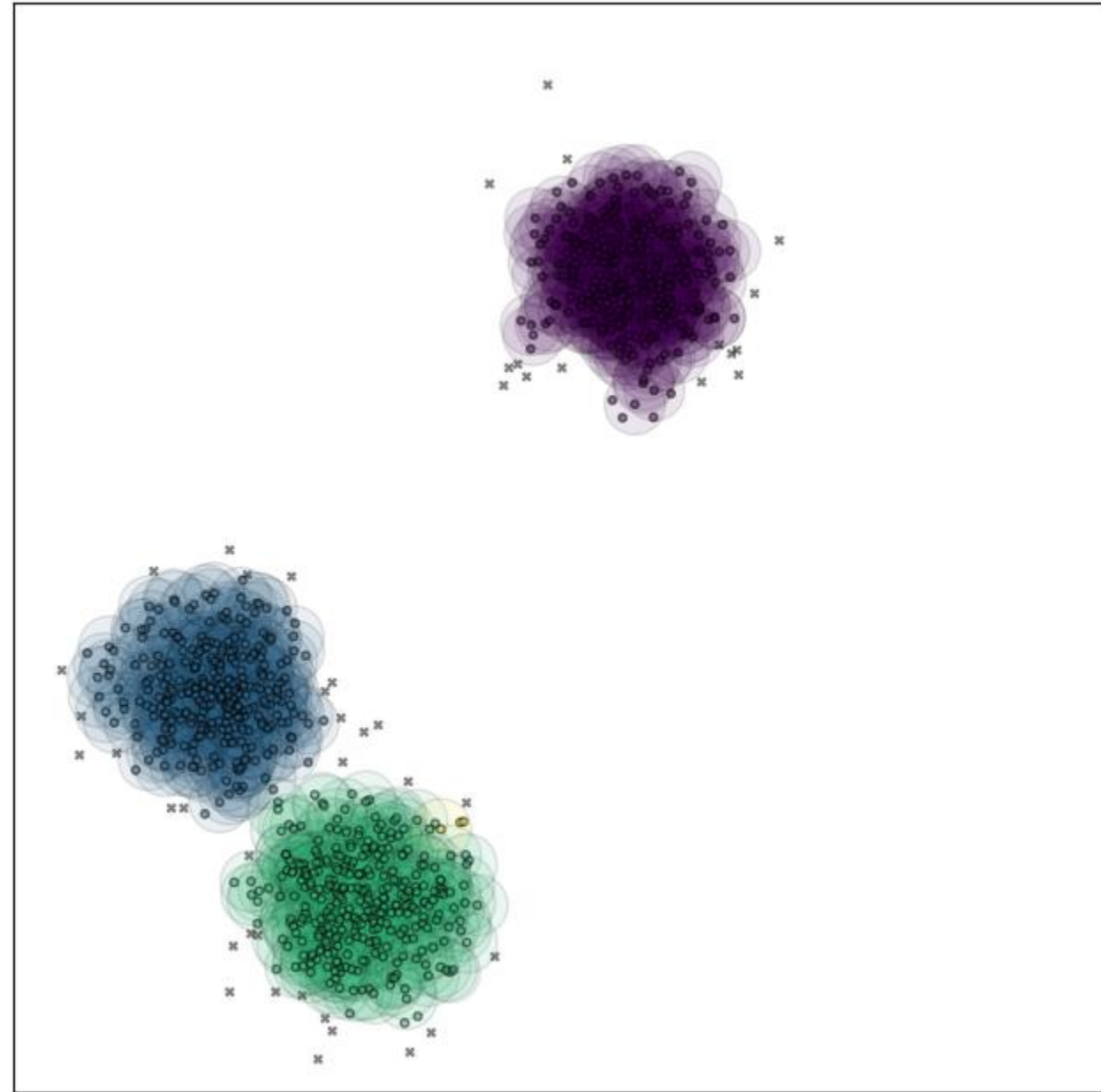
DBSCAN

- 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



DBSCAN

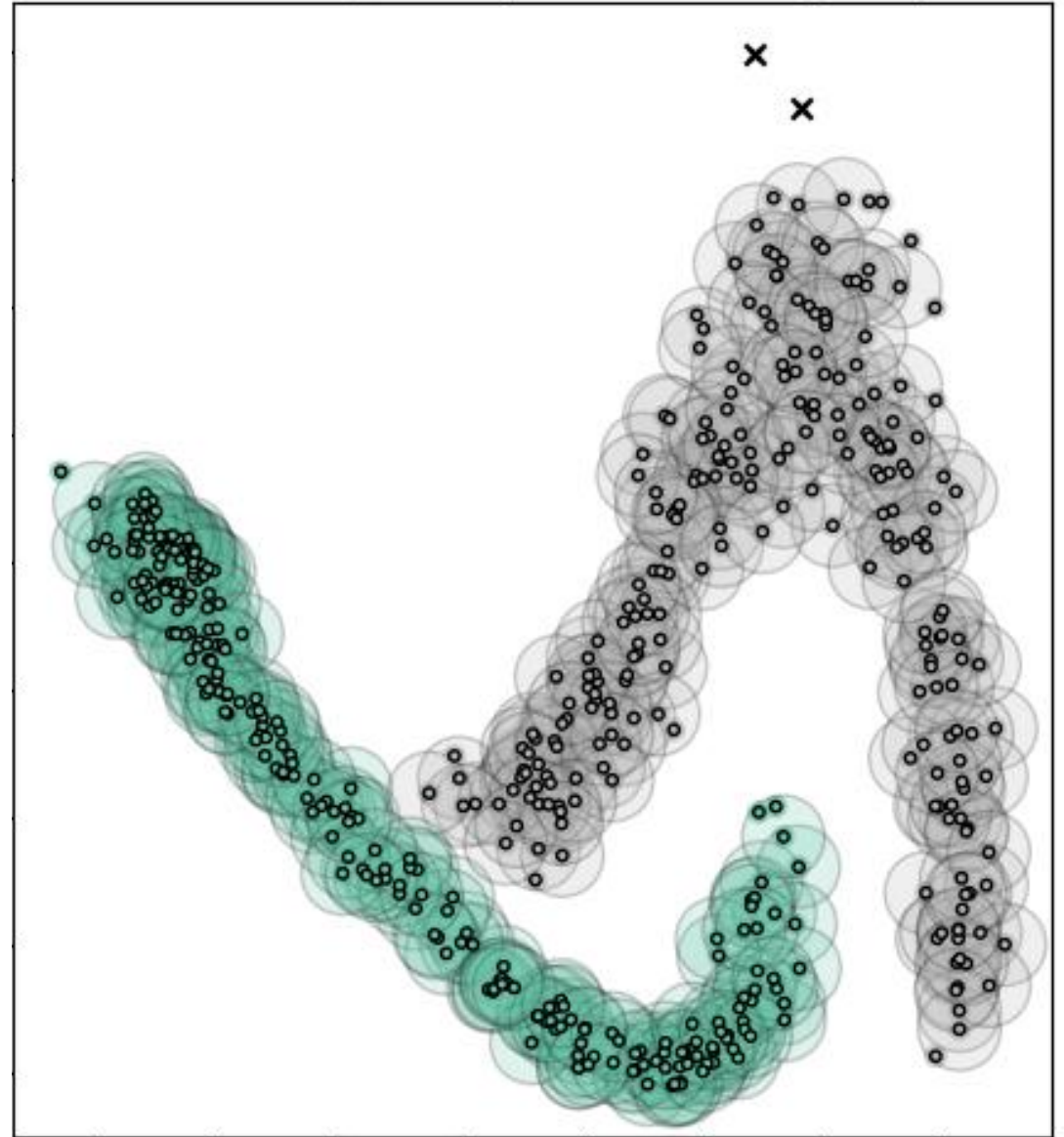
- Density-Based Spatial Clustering of Applications with Noise
- Identifies cluster points based on neighbourhood radius ϵ and a minimum number of samples in the neighbourhood



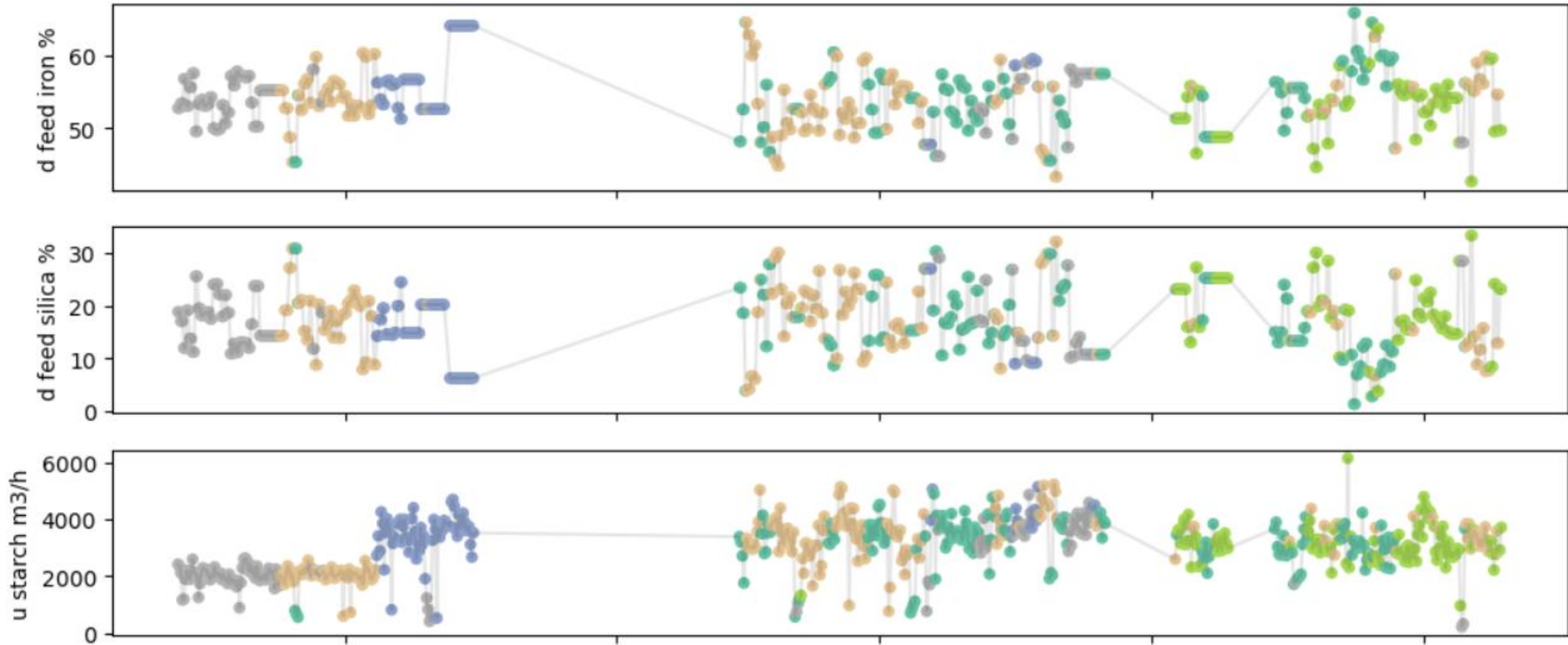
DBSCAN

- 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

DBSCAN clustering with $\epsilon = 0.01$, $\text{min_samples} = 5$

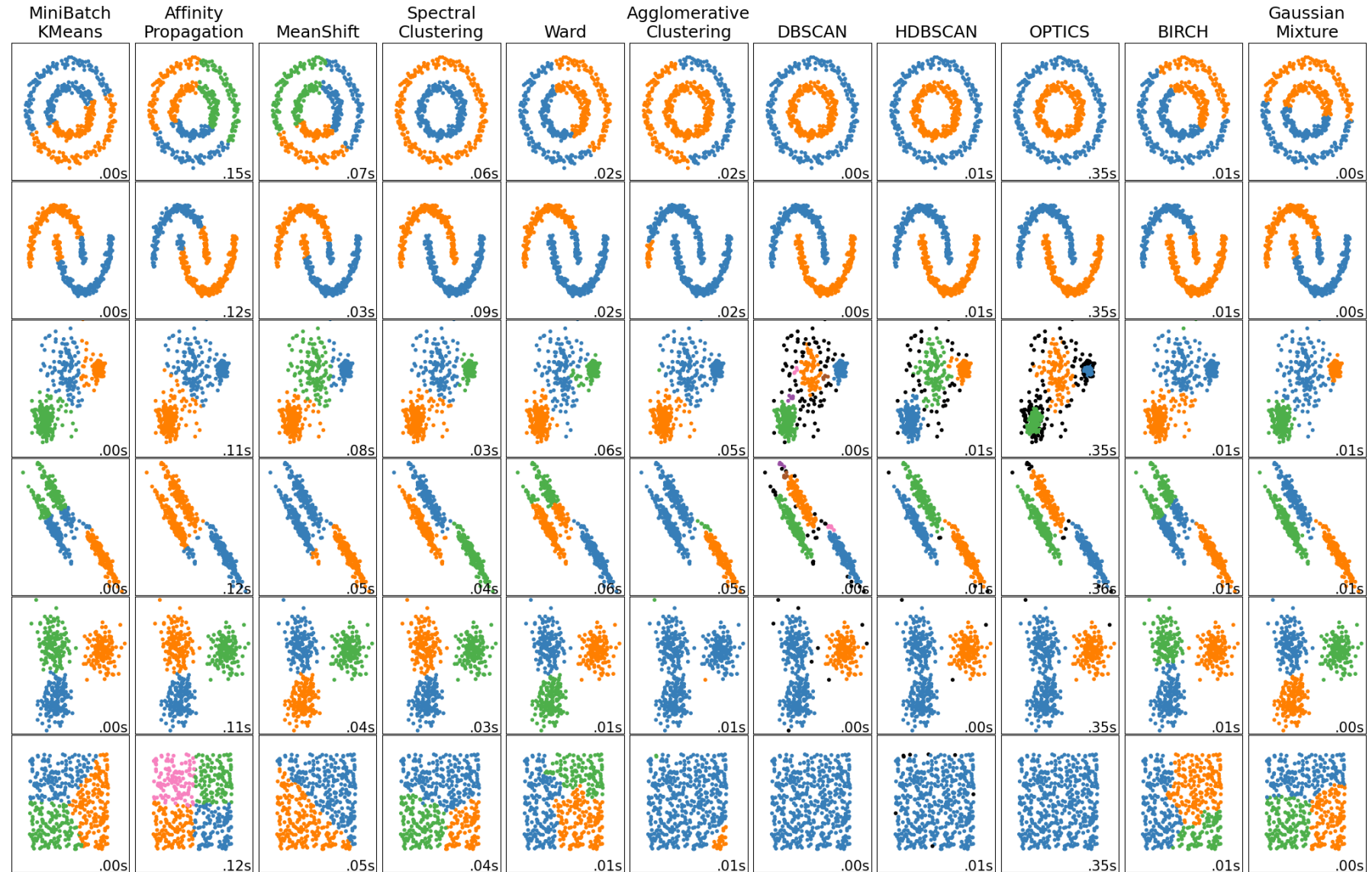


More informative time series plots



Clustering

- Many other methods available on scikit-learn

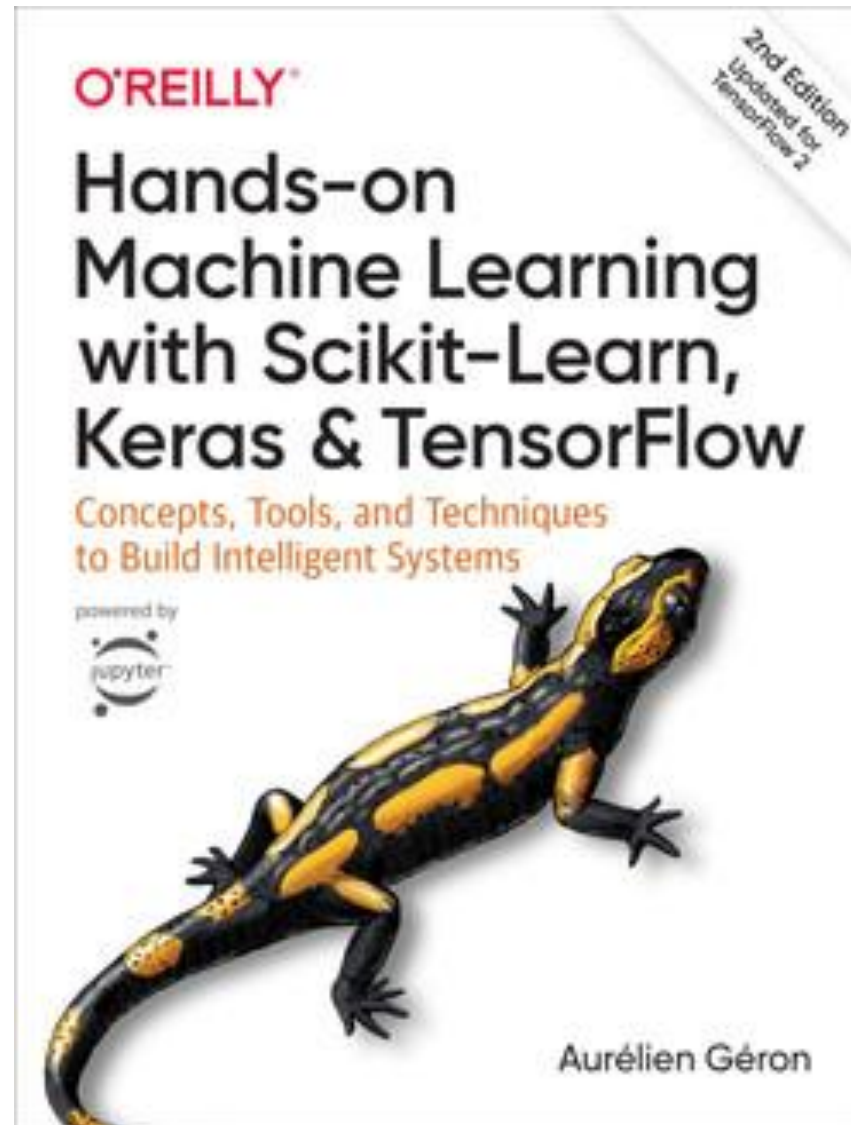


<https://scikit-learn.org/stable/modules/clustering.html>

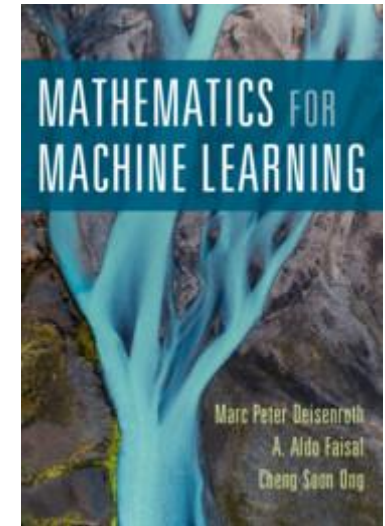
Resources



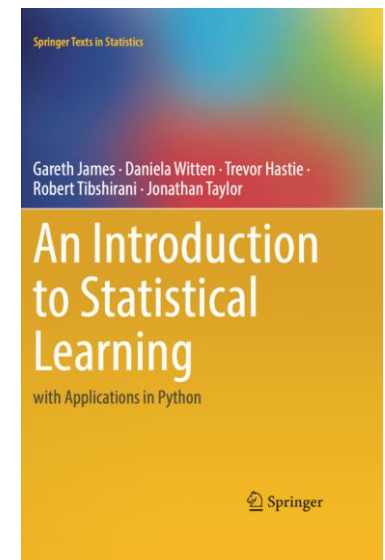
https://scikit-learn.org/stable/unsupervised_learning.html



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



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



<https://www.statlearning.com/>