Clustering

Aalto-yliopisto
Aalto-universitetet
Aalto University

Martin Trapp

26.06.2023

Outline

What is clustering?

Example algorithm (K-Means)

Challenges and possible remedies

Summary

Learning Goals

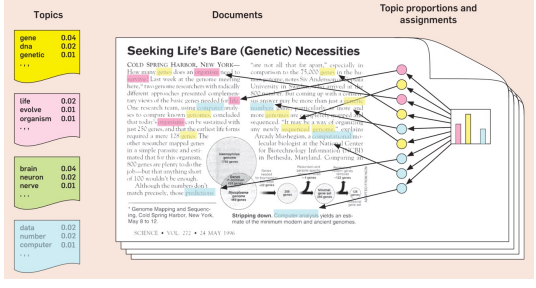
Challenges associated with clustering

Know about the K-Means algorithm

Know about the challenges of K-Means and possible solutions

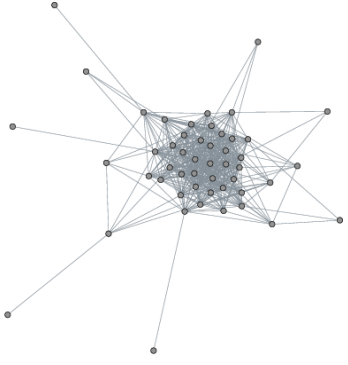
Applications are Everywhere

Clustering Documents (Topic Modelling)



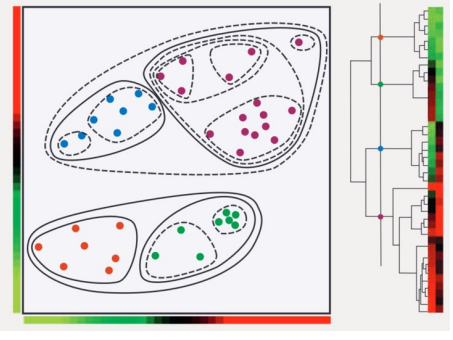
Blei 2012

Clustering Social Networks



Orbanz 2013

Gene Expression Clustering



D'haeseleer 2005

What is clustering?

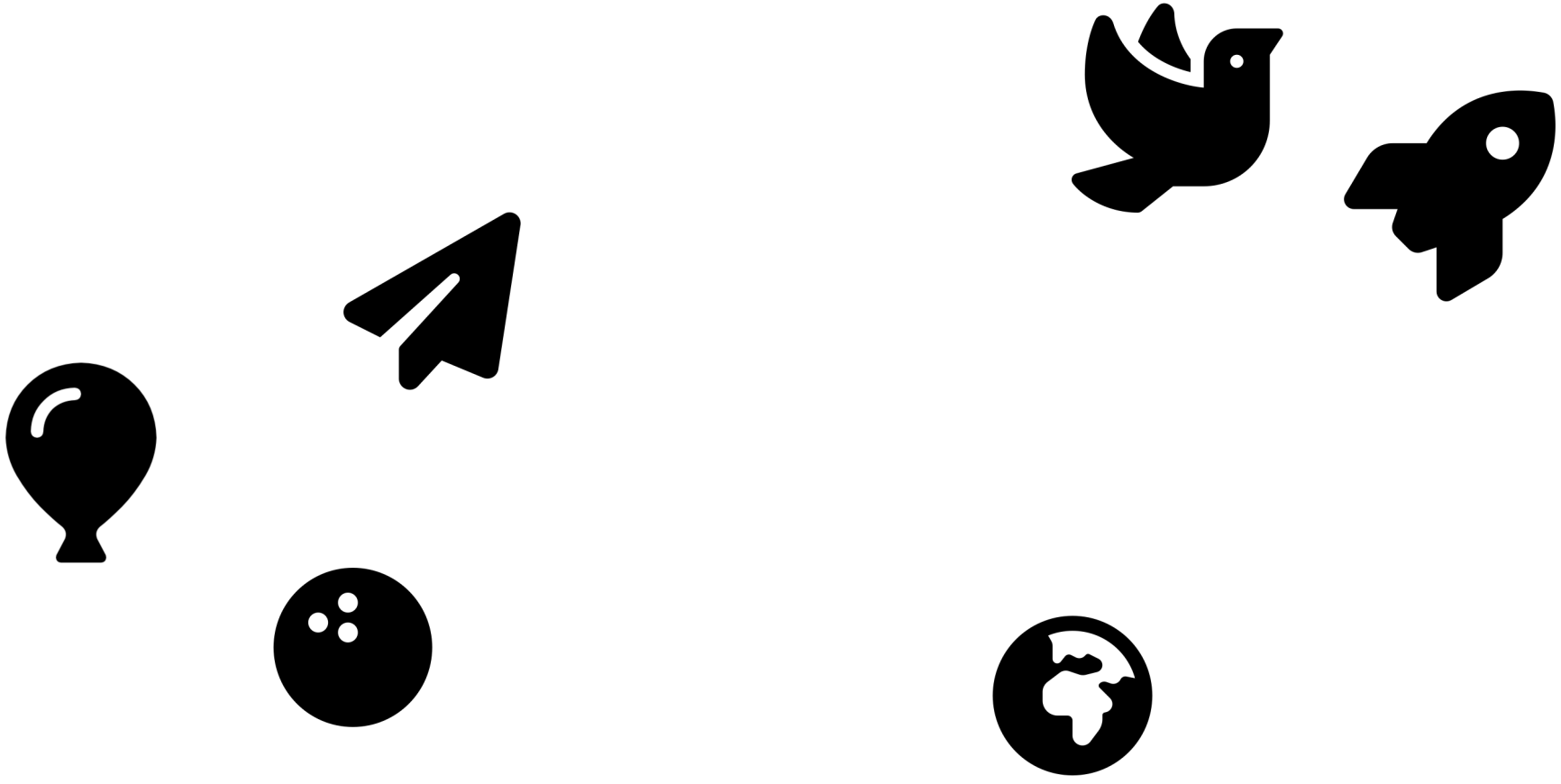
Clustering by Definition

Definition: Cluster Analysis

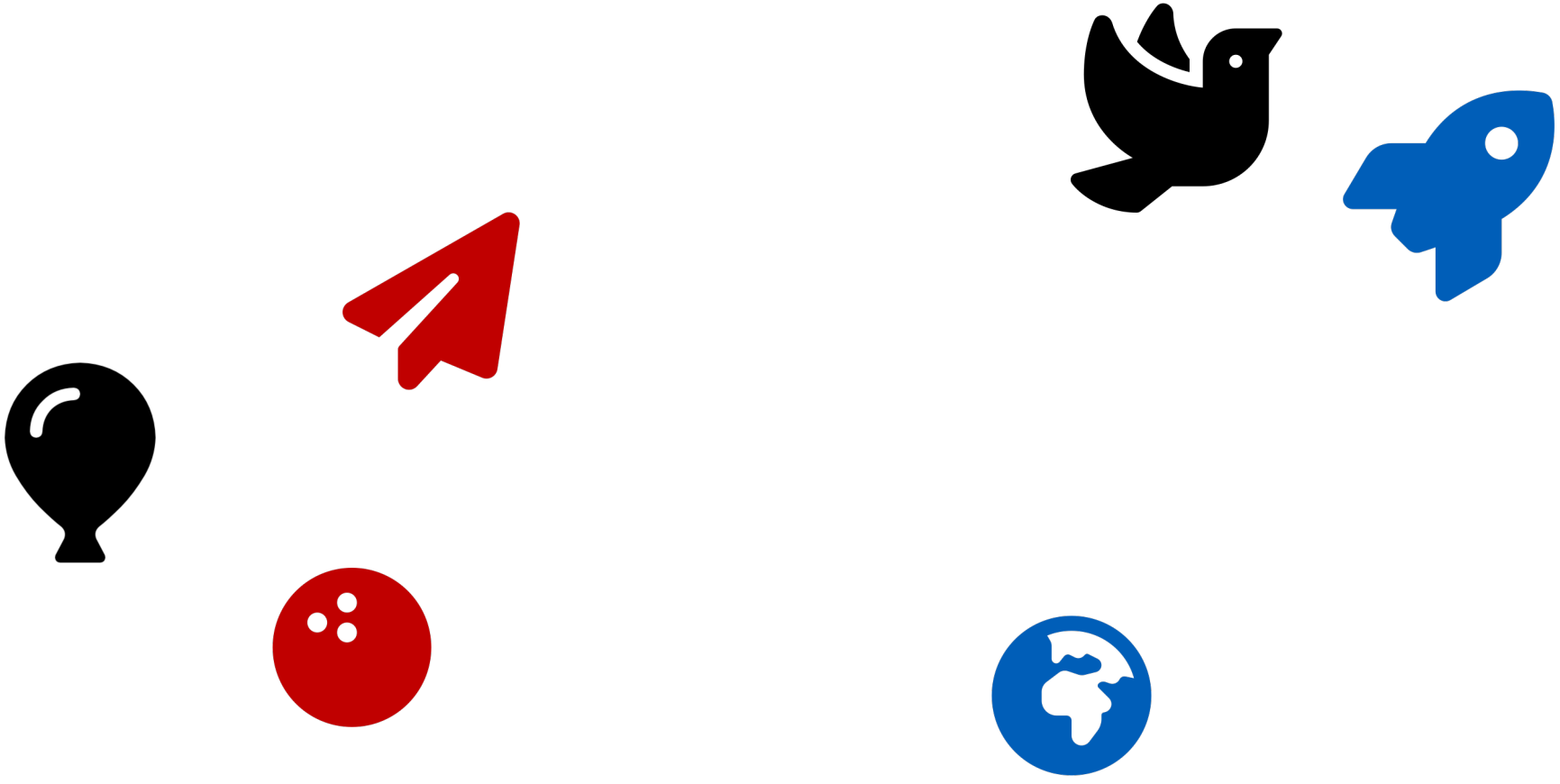
A set of methods for constructing a (hopefully) sensible and informative classification of an initially unclassified set of data, using the variable values observed on each individual.

B. S. Everitt and A. Skrondal (2010). The Cambridge Dictionary of Statistics (4th ed.) Cambridge University Press.

Clustering by Example



Clustering by Example



Clustering is not easy

- We can have many sensible clusterings/groupings
- Grouping depends on the features/characteristics we consider
- We can have a hierarchy of clusters
- It can be hard to measure the quality of the clustering

Clustering Methods

Rough Overview

Density-based methods

- **Assumption**: clusters located in high-density regions
- **Examples**: DBSCAN, DENCLUE, …

Partitional methods

- **Assumption**: data can be partitioned into disjoint groups
- **Examples**: K-Means, K-medoids, ...

Hierarchical methods

- **Assumptions**: hierarchically merge pairs of similar data points
- **Examples**: Agglomerative methods, divisive methods, …

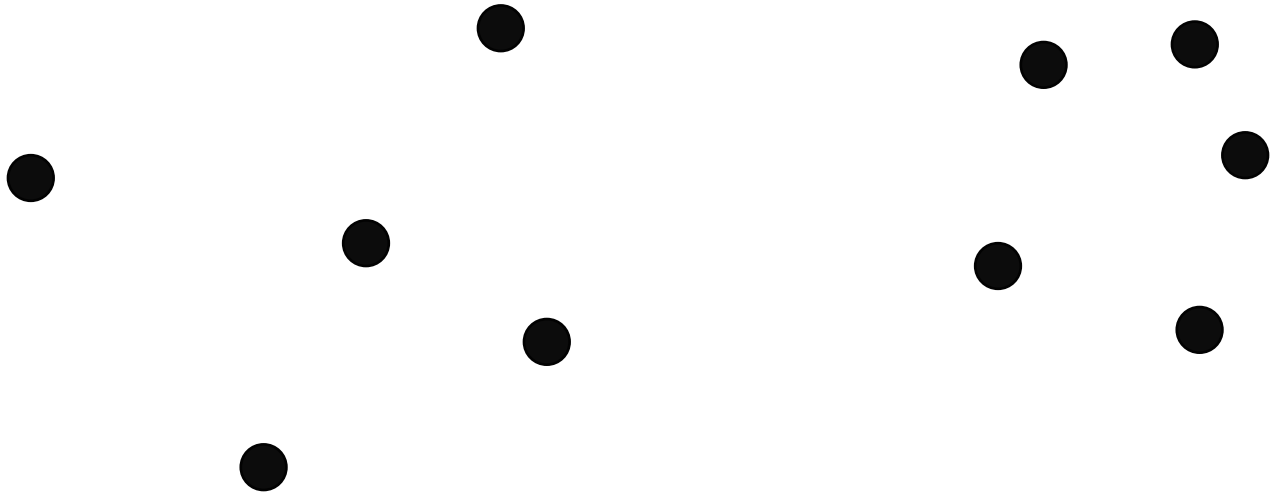
K-Means in a Nutshell

- **Represent clusters by their centroid/mean**
- **Find initial centroids (e.g., random, farthest traversal)**
- **Iteratively adjust centroids**

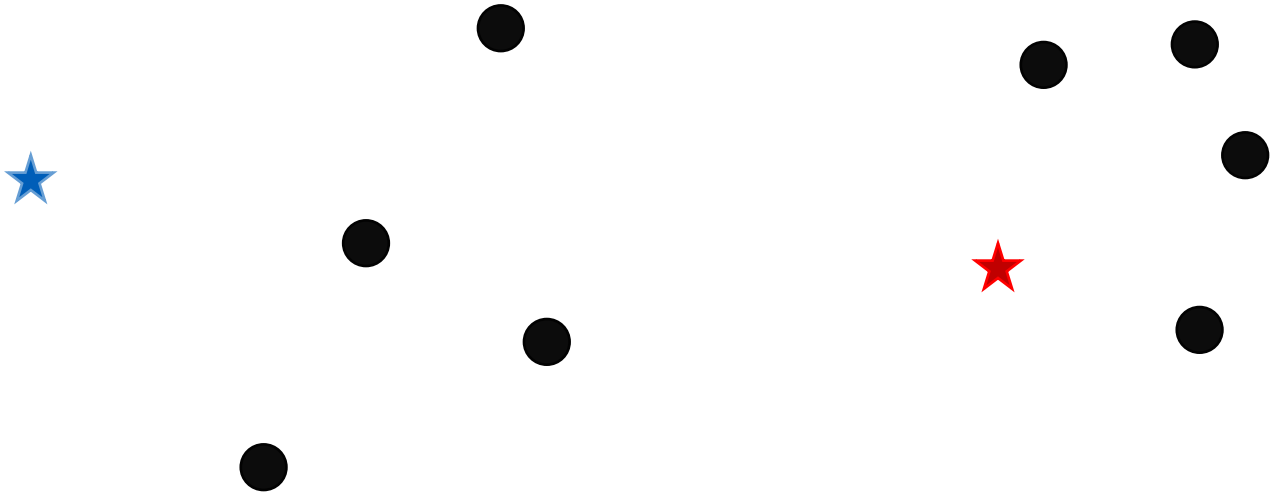
Procedure:

1. Assign each datum to its closest centroid
2. Recompute centroids based on assignments
3. Go back to (1.)

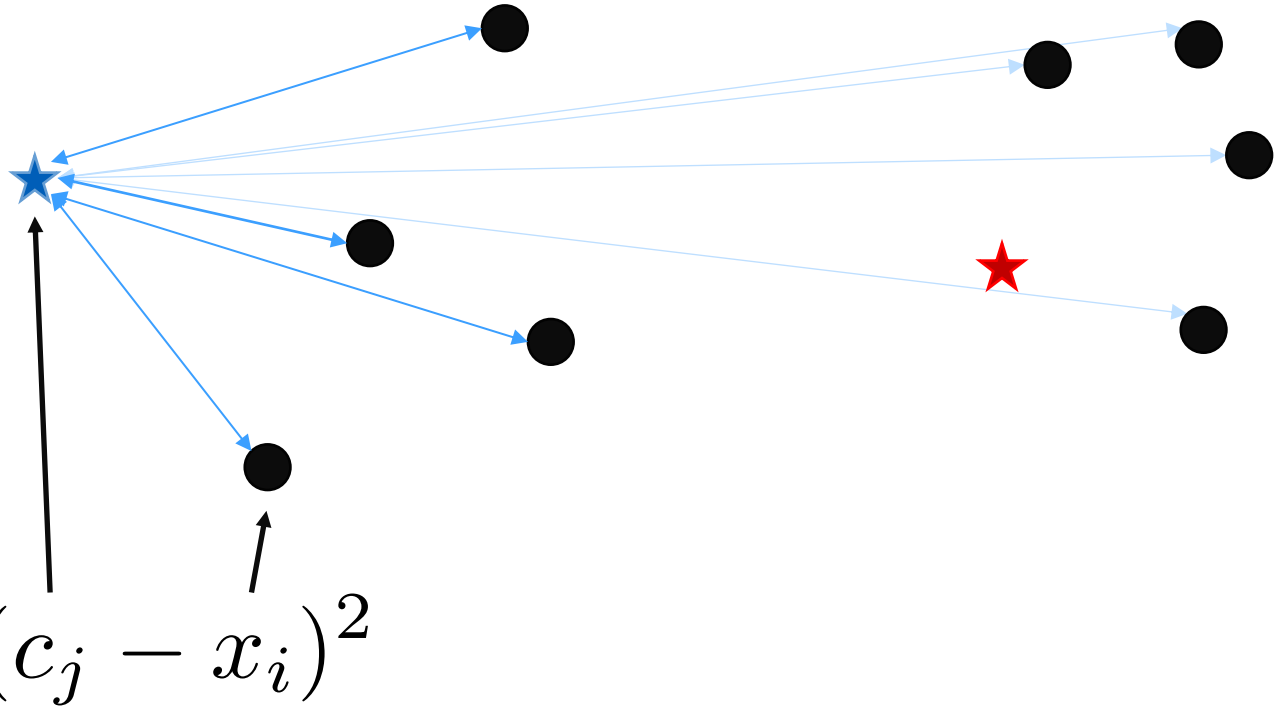
K-Means in a Nutshell



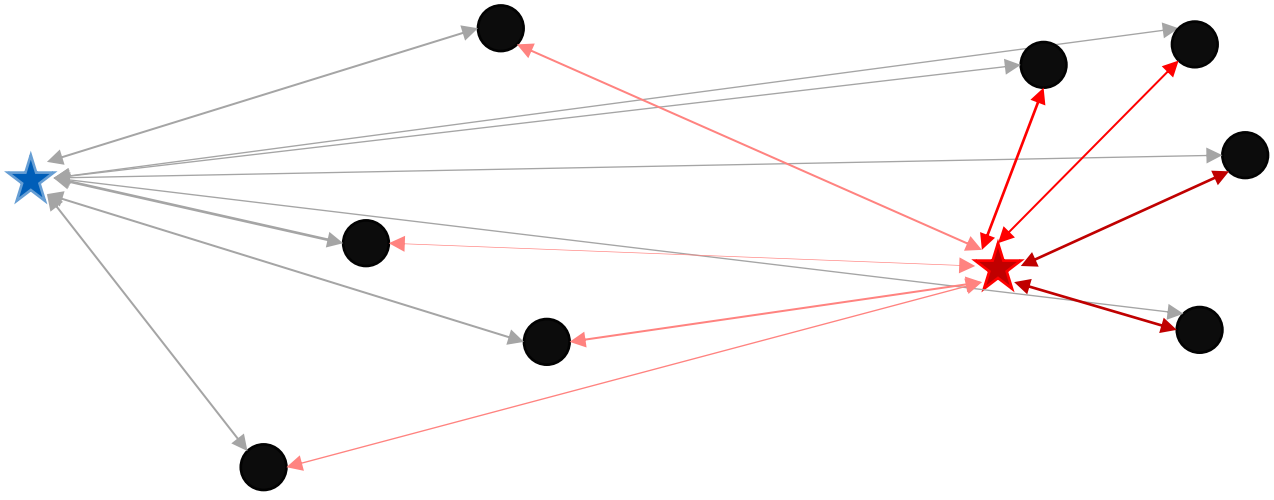
K-Means in a Nutshell



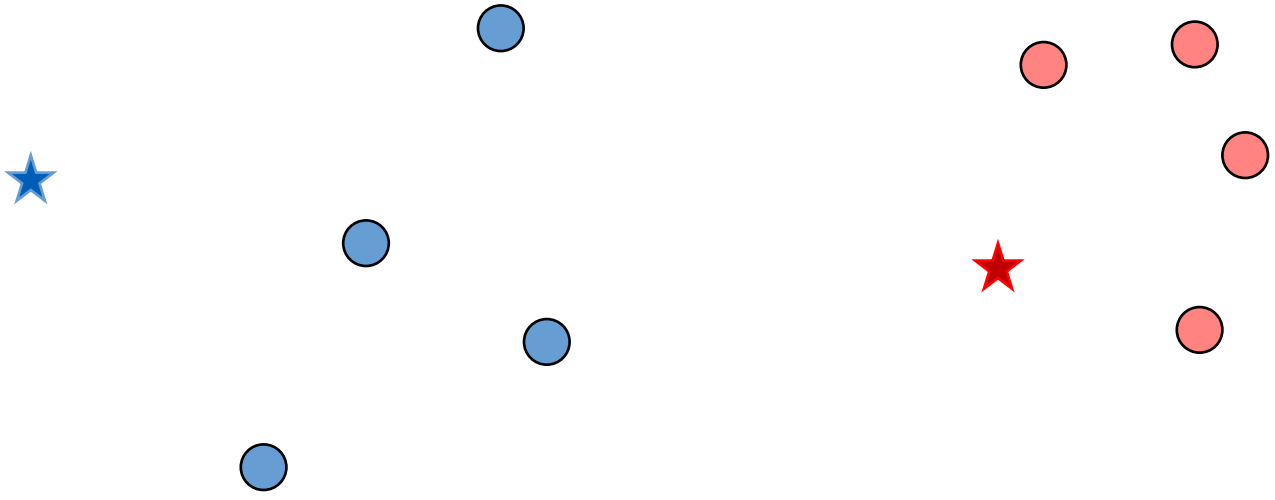
K-Means in a Nutshell



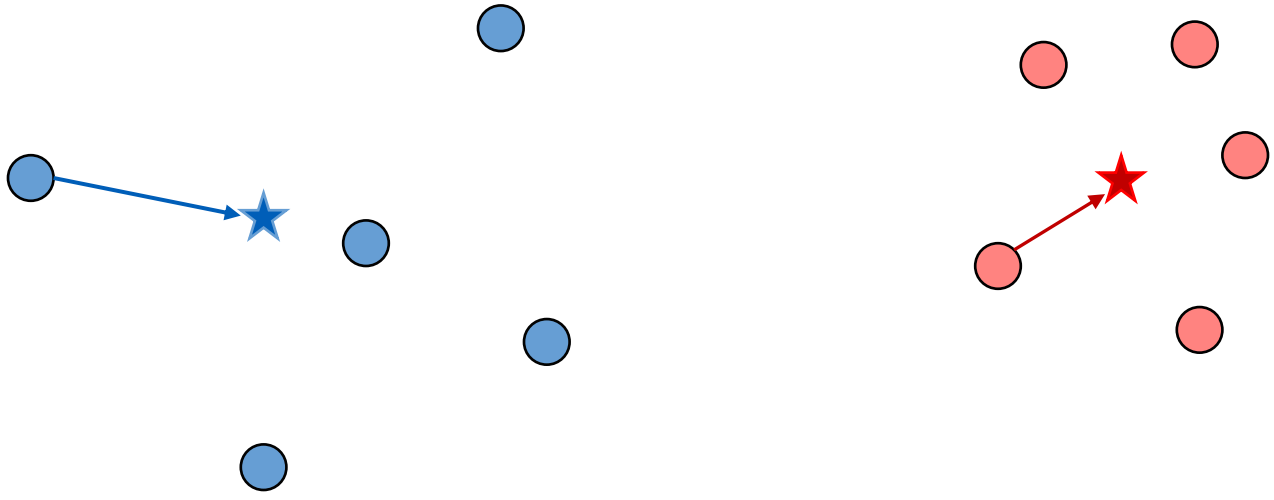
K-Means in a Nutshell



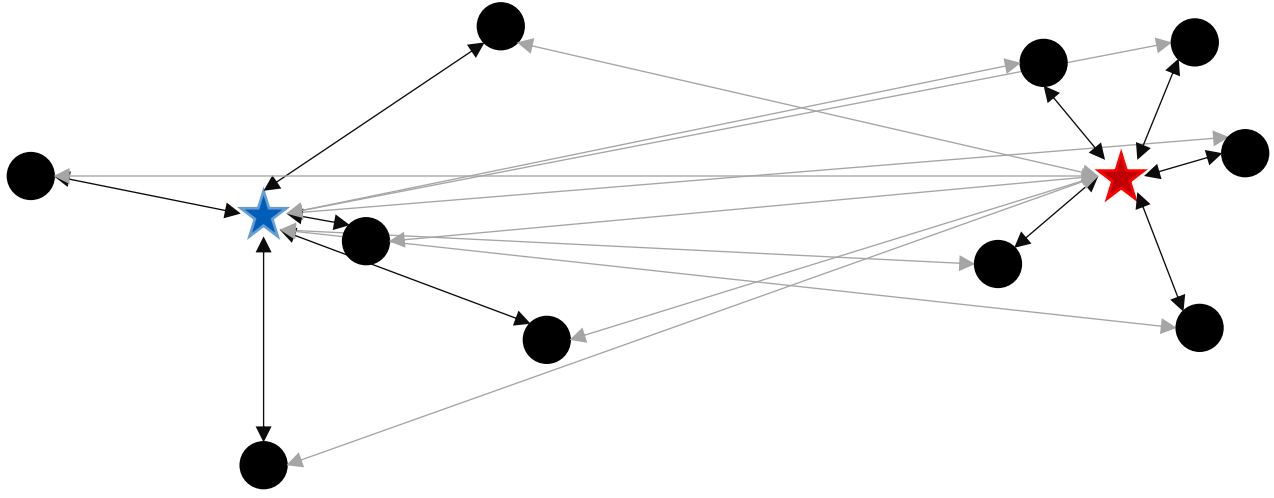
K-Means in a Nutshell



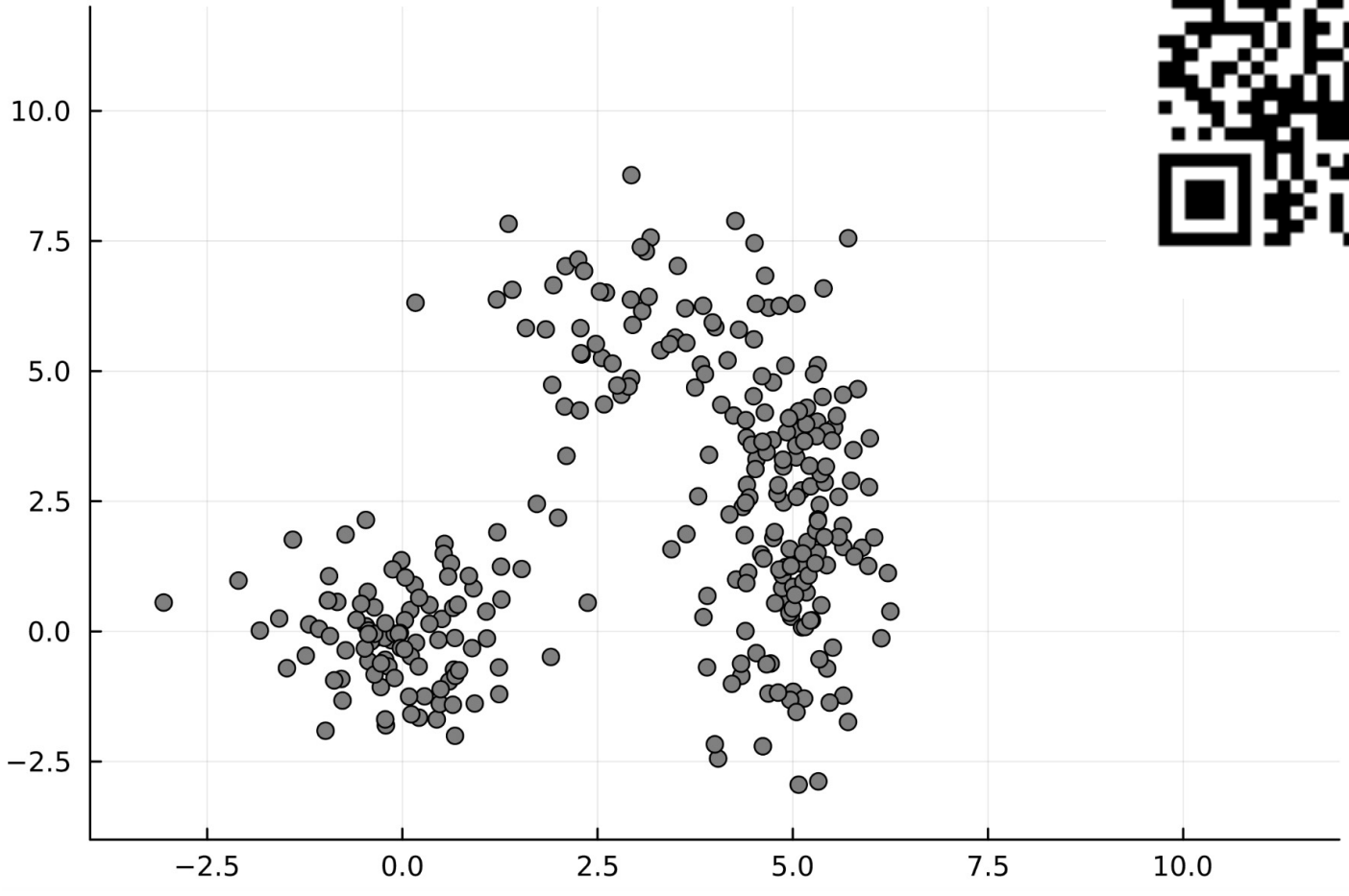
K-Means in a Nutshell



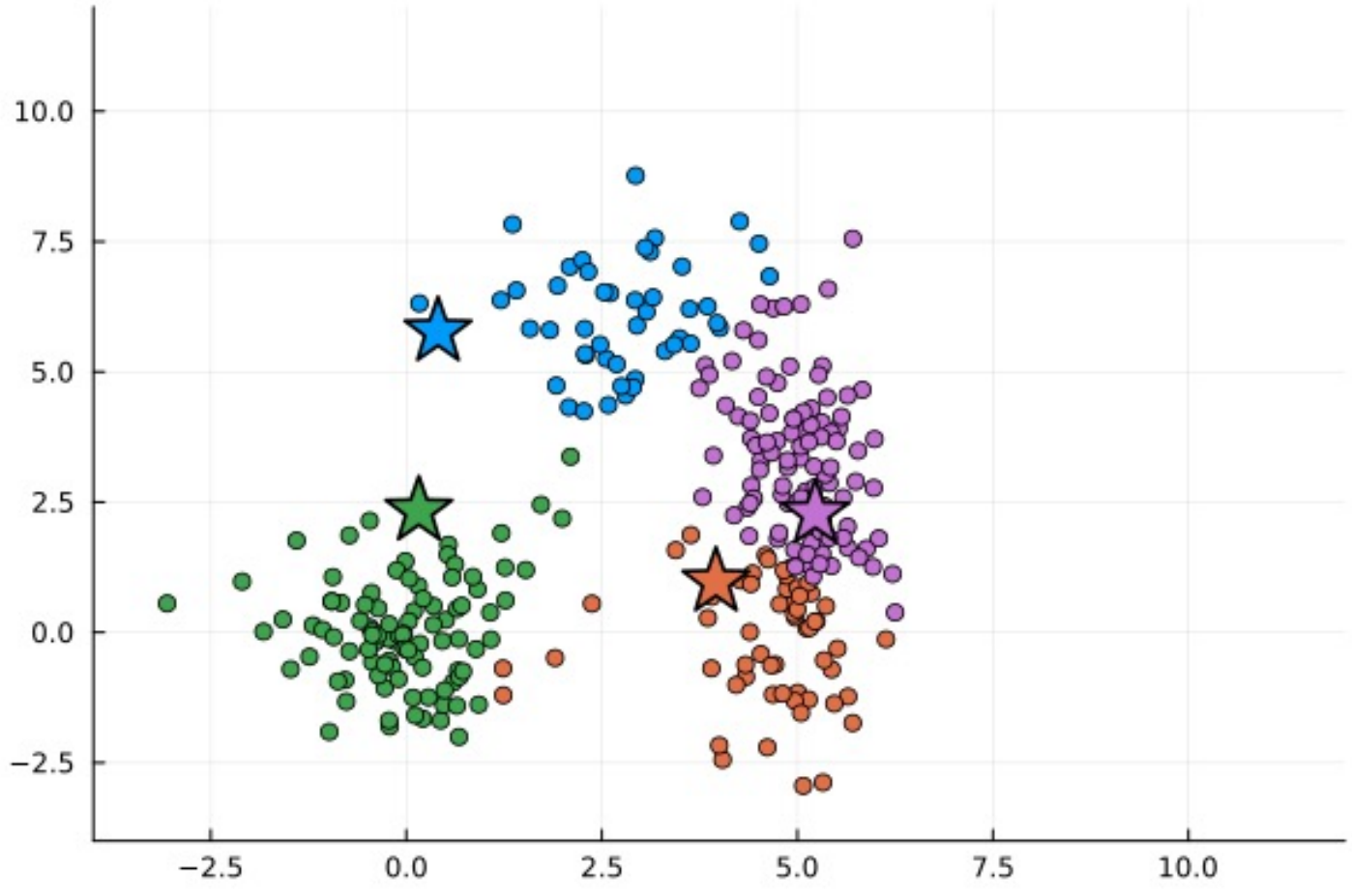
K-Means in a Nutshell



Let’s cluster!



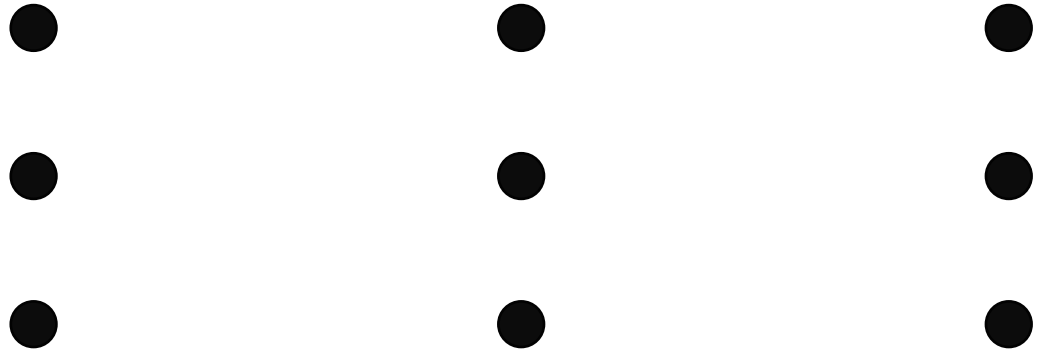
Let’s cluster!



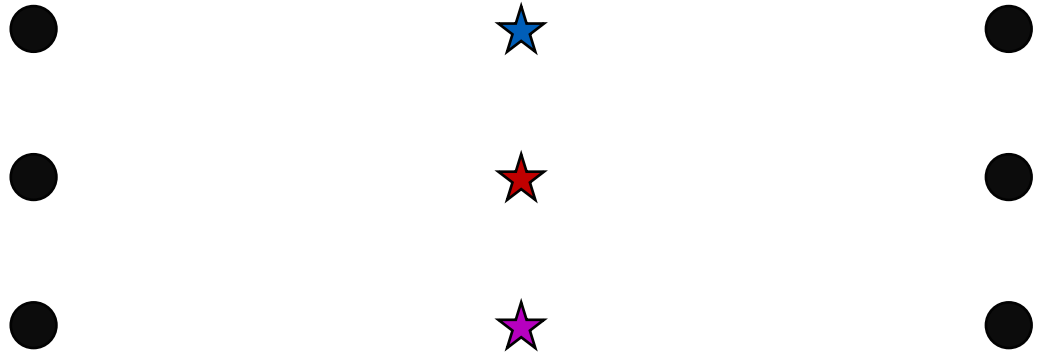
Challenges with K-Means

- Sensitivity to the initialisation
- We need to know how many clusters we expect
- Sensitive to outliers because of the squared Euclidean distance
- We assume that we can perform “hard” clustering

Sensitivity to the initialisation

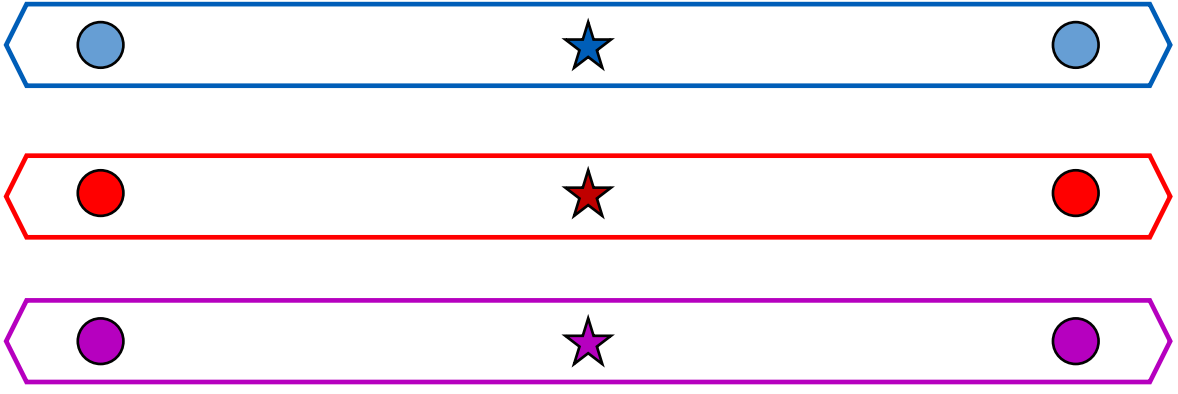


Sensitivity to the initialisation



Sensitivity to the initialisation

random initialization can be arbitrarily bad



remedy through more sophisticated heuristics

Challenges & Remedies

- **Sensitivity to the initialisation**
  - More sophisticated initialisations (kmeans++) can help
- **We need to know how many clusters we expect**
  - Nonparametric methods can help (Chinese restaurant process)
- **Sensitive to outliers because of the squared Euclidean distance**
  - Truncation of the data set or change of distance can help

Summary

- **Clustering aims to find a classification of unlabelled data**
- **Clustering can be ambiguous and depend on various factors**

- **K-Means:**
  - Use centroids to represent clusters
  - Classify data based on the “closest” centroid
  - Adjust centroid locations iteratively
- **K-Means comes with various challenges such as:**
  - Initialization of the centroids, similarity measure, …
- **Remedies to the challenges exist**

Thank you for listening!