Foundations of Data Science: Introduction to unsupervised learning

THE UNIVERSITY of EDINBURGH
informatics

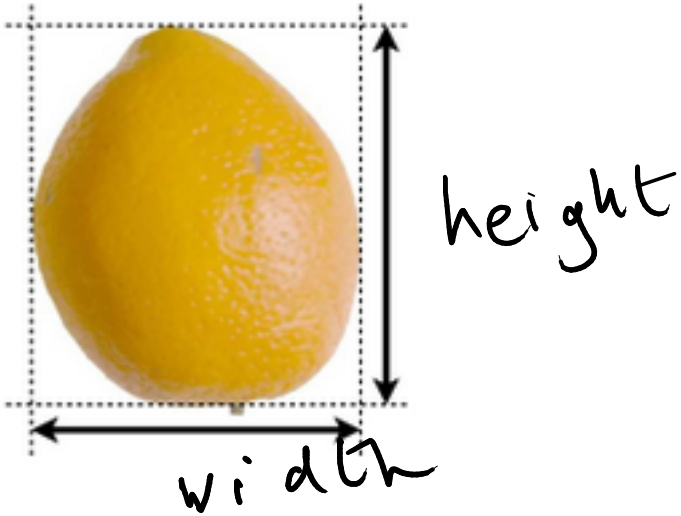
FOUNDATIONS OF DATA SCIENCE

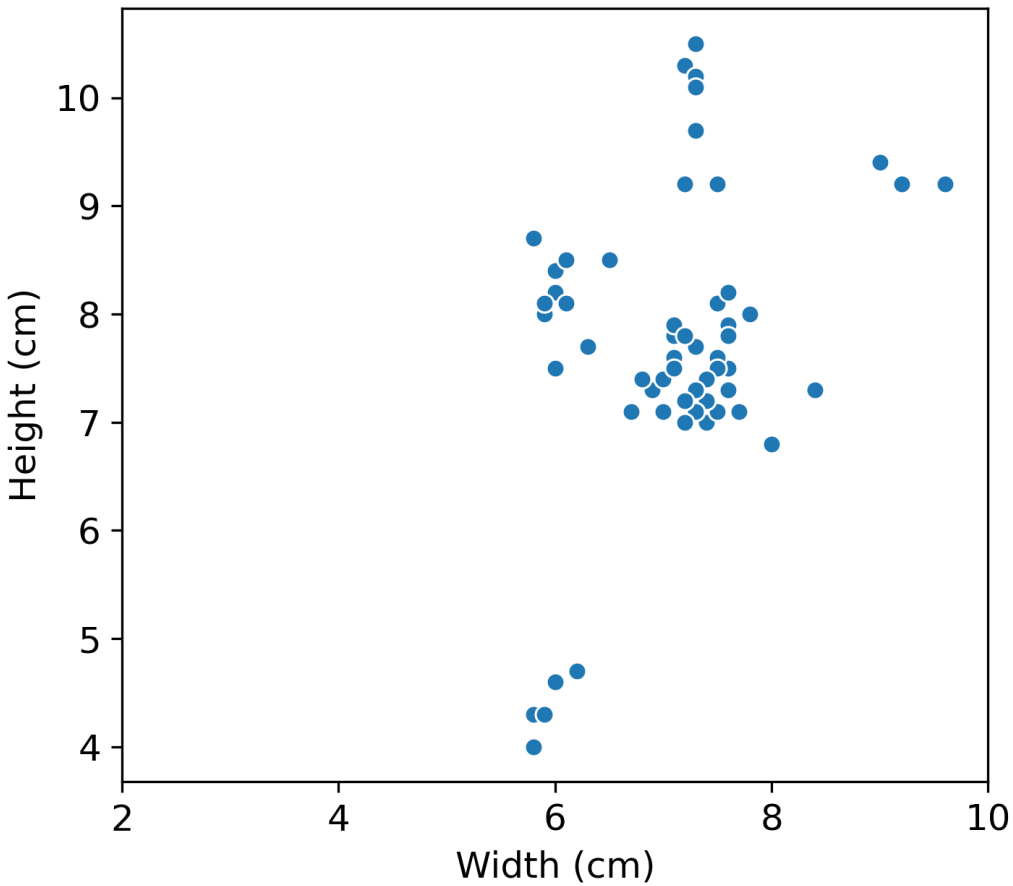
Overview

1. Unsupervised learning, supervised learning, clustering
2. Partitional versus hierarchical clustering
3. K-means
4. Evaluation of K-means

Foundations of Data Science: Introduction to unsupervised learning - Clustering, unsupervised and supervised learning

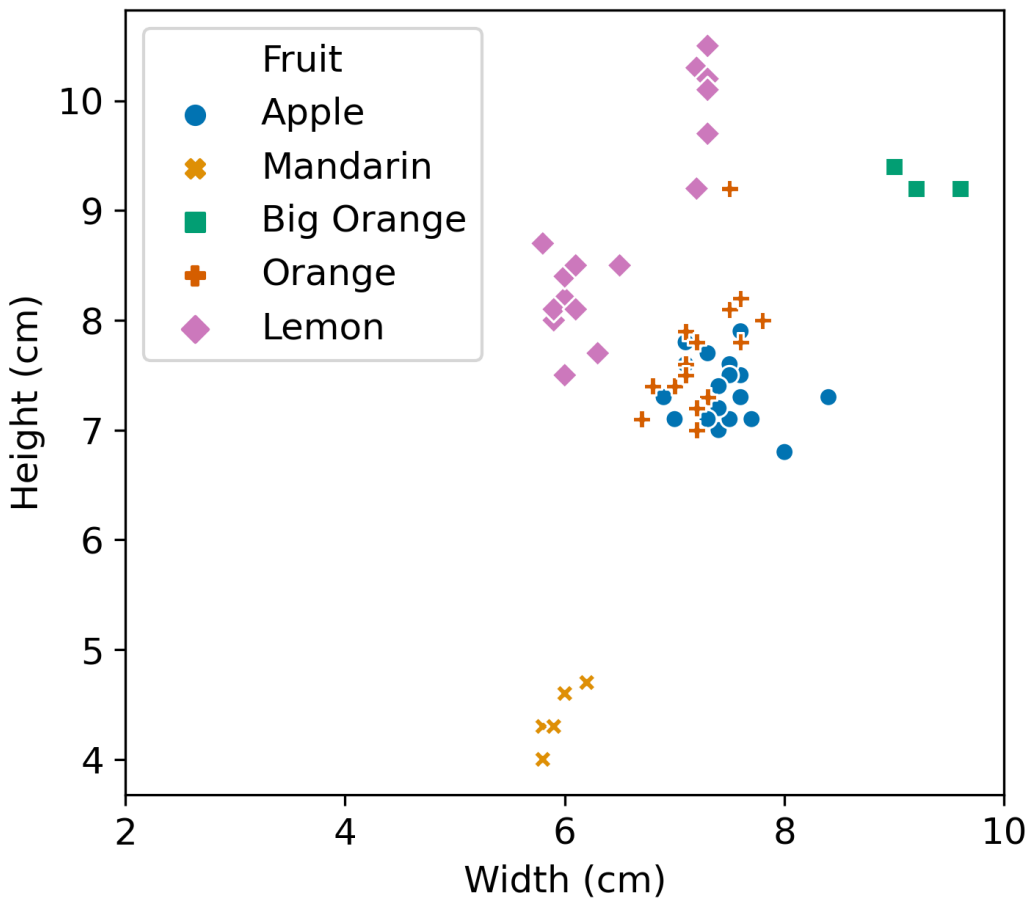
Clustering





Data collected by Iain Murray
https://homepages.inf.ed.ac.uk/imurray2/teaching/oranges_and_lemons

Supervised learning process



For example, classification:

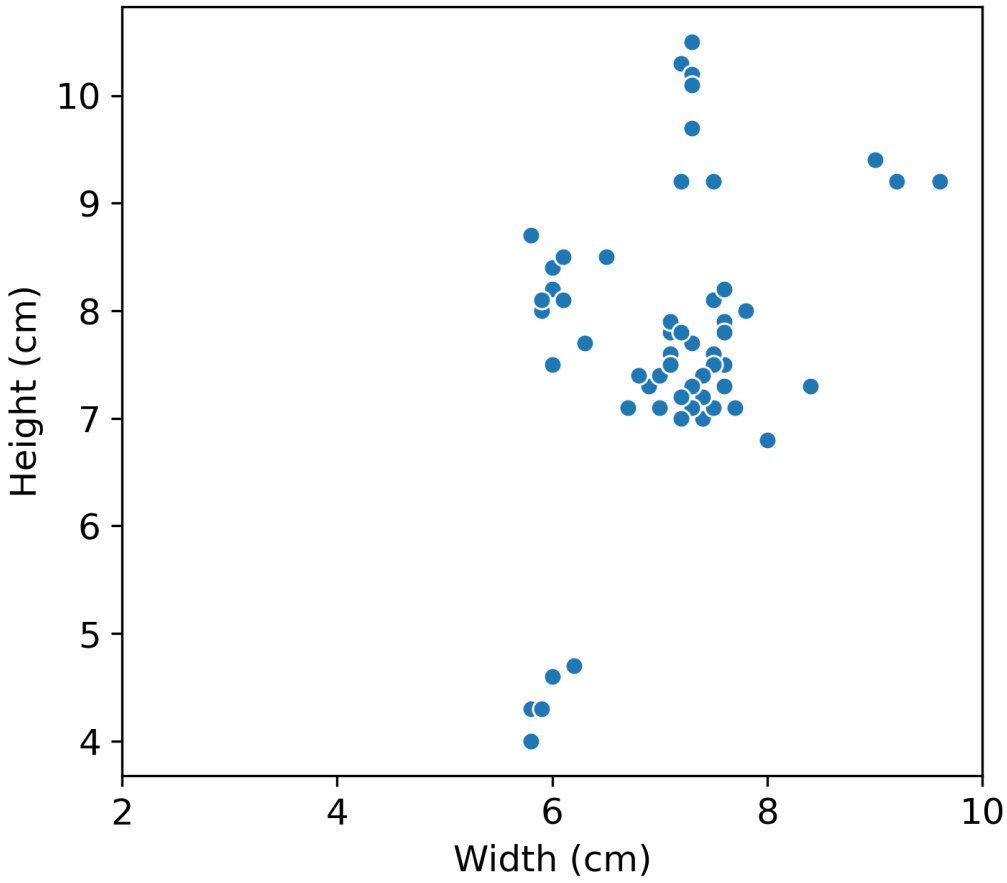
Training set:

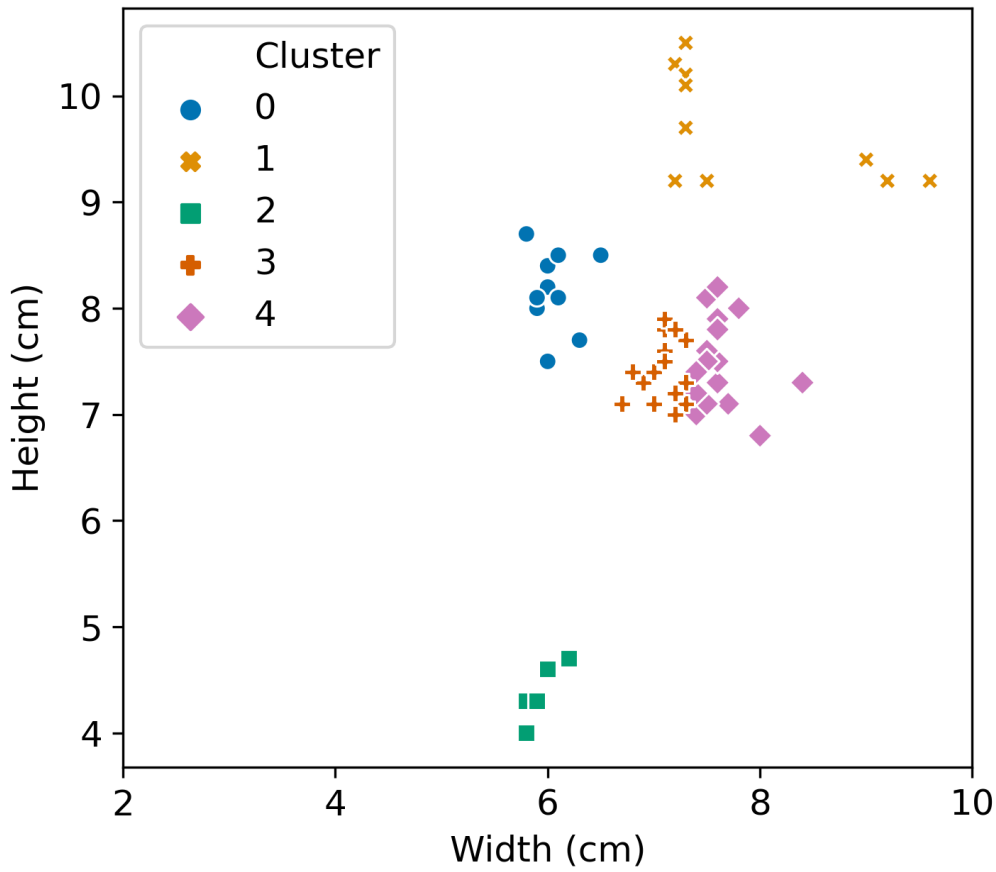
Features: width, height

Label: fruit

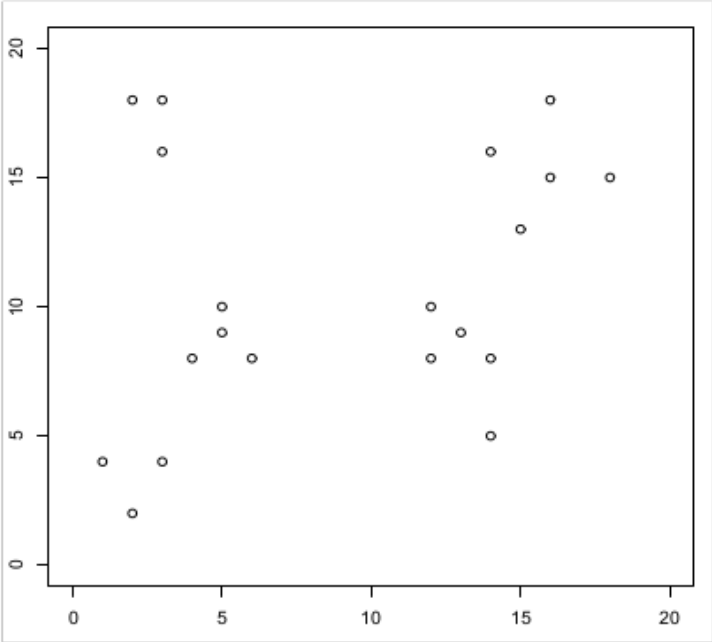
Unsupervised learning process

For example, clustering

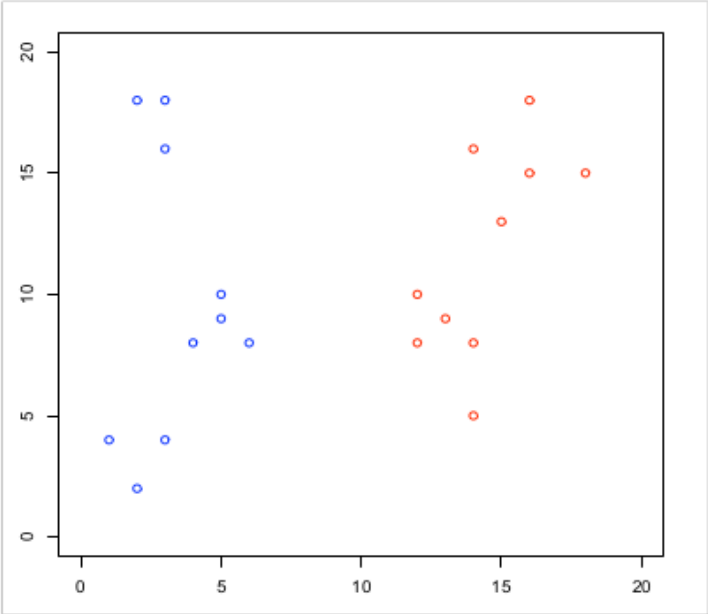




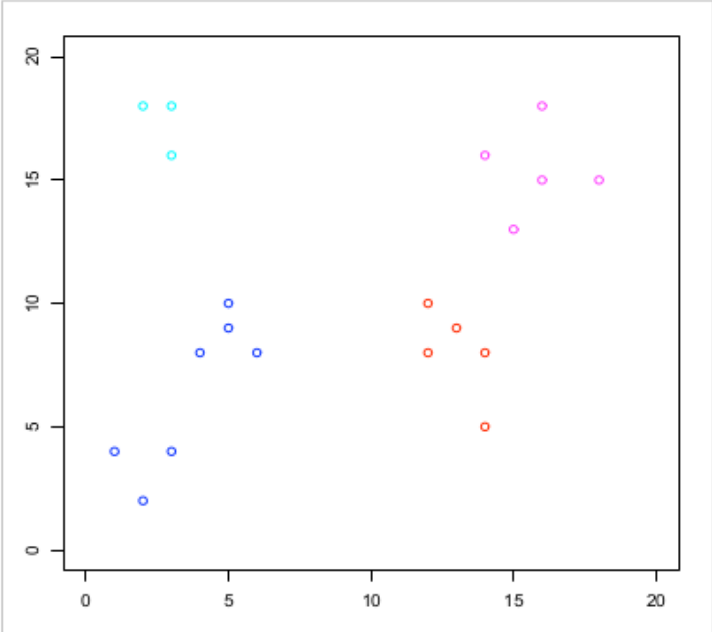
How many clusters are there?



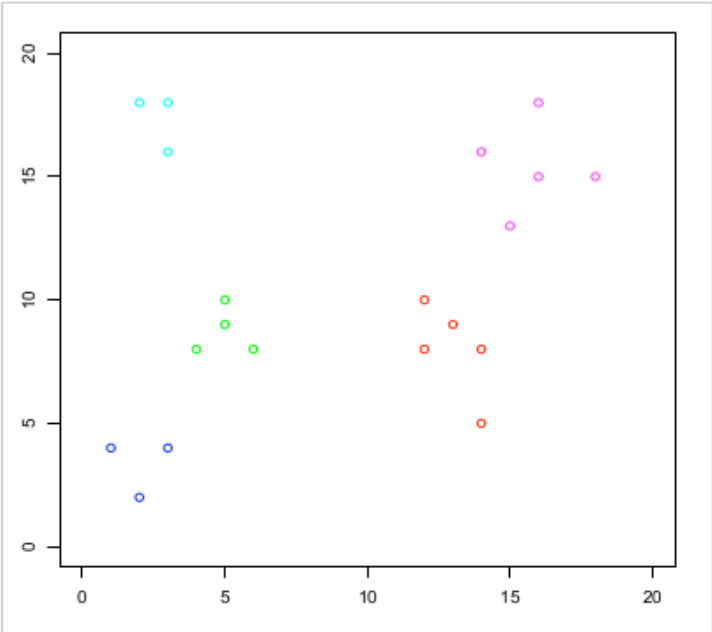
(a) original data



(b) two clusters



(c) four clusters



(d) five clusters

Why cluster?

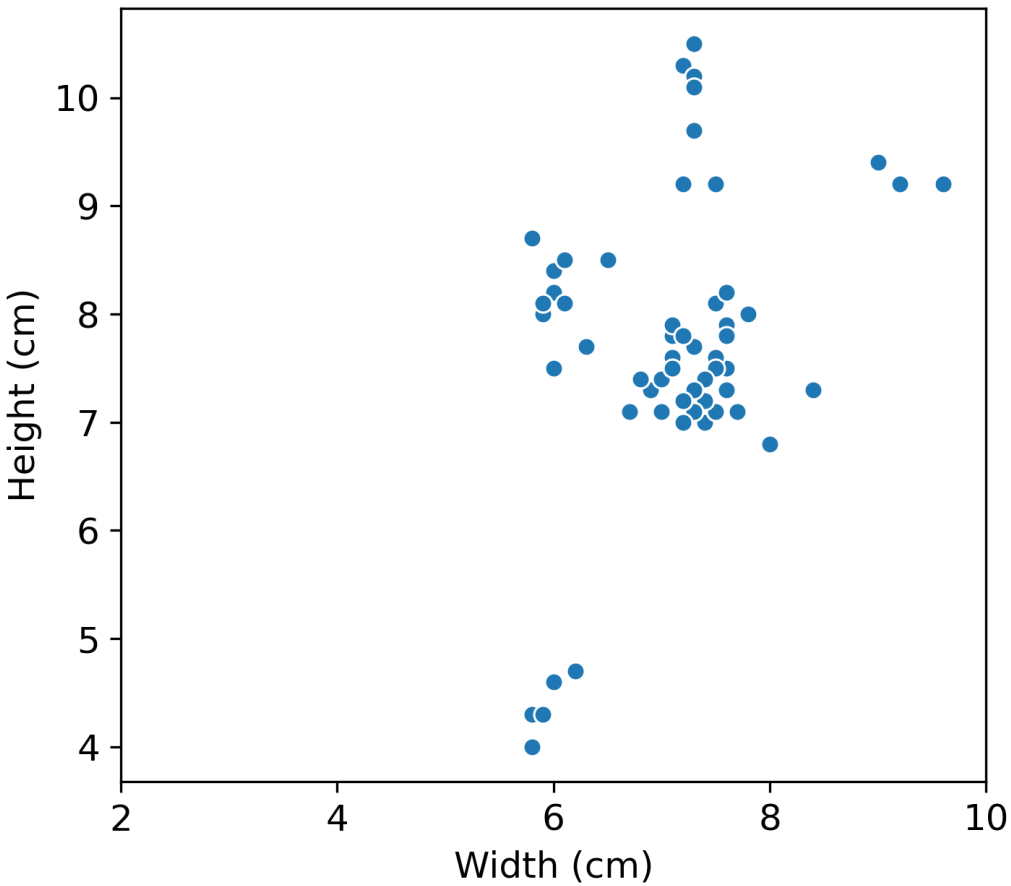
1. Interpretation

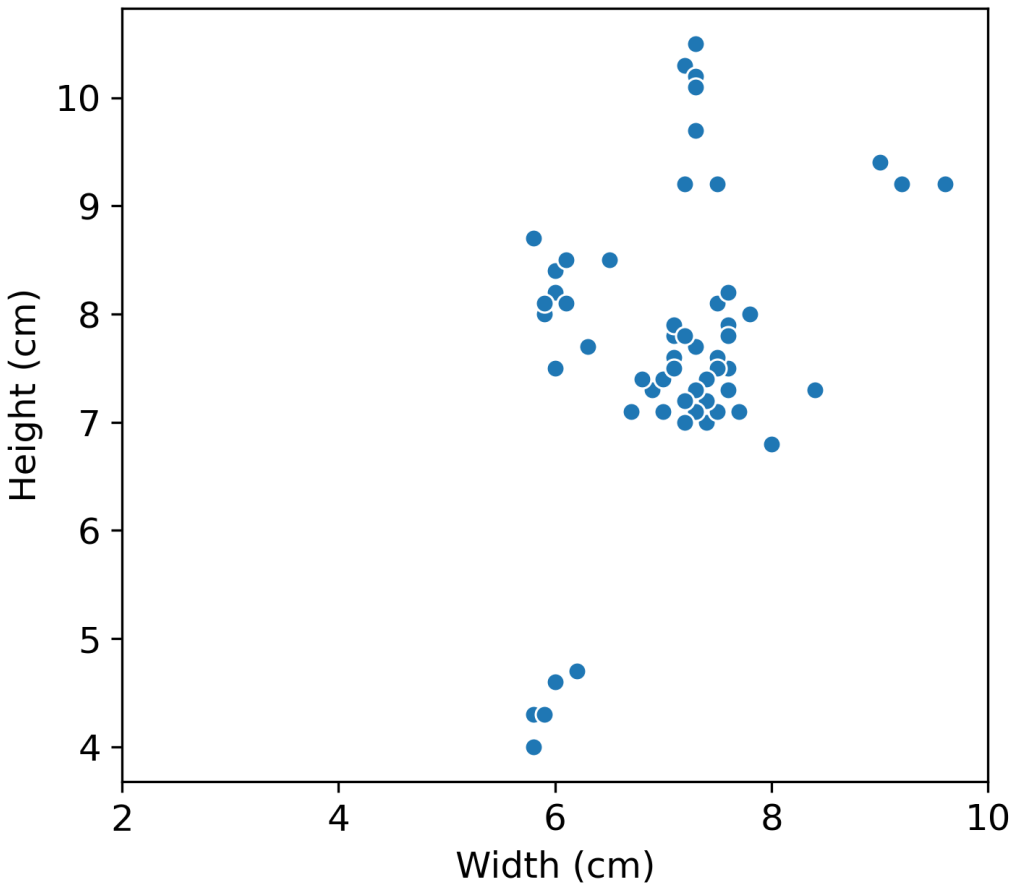
2. Data compression

Foundations of Data Science: Introduction to unsupervised learning - Types of clustering

Types of clustering

Partitional

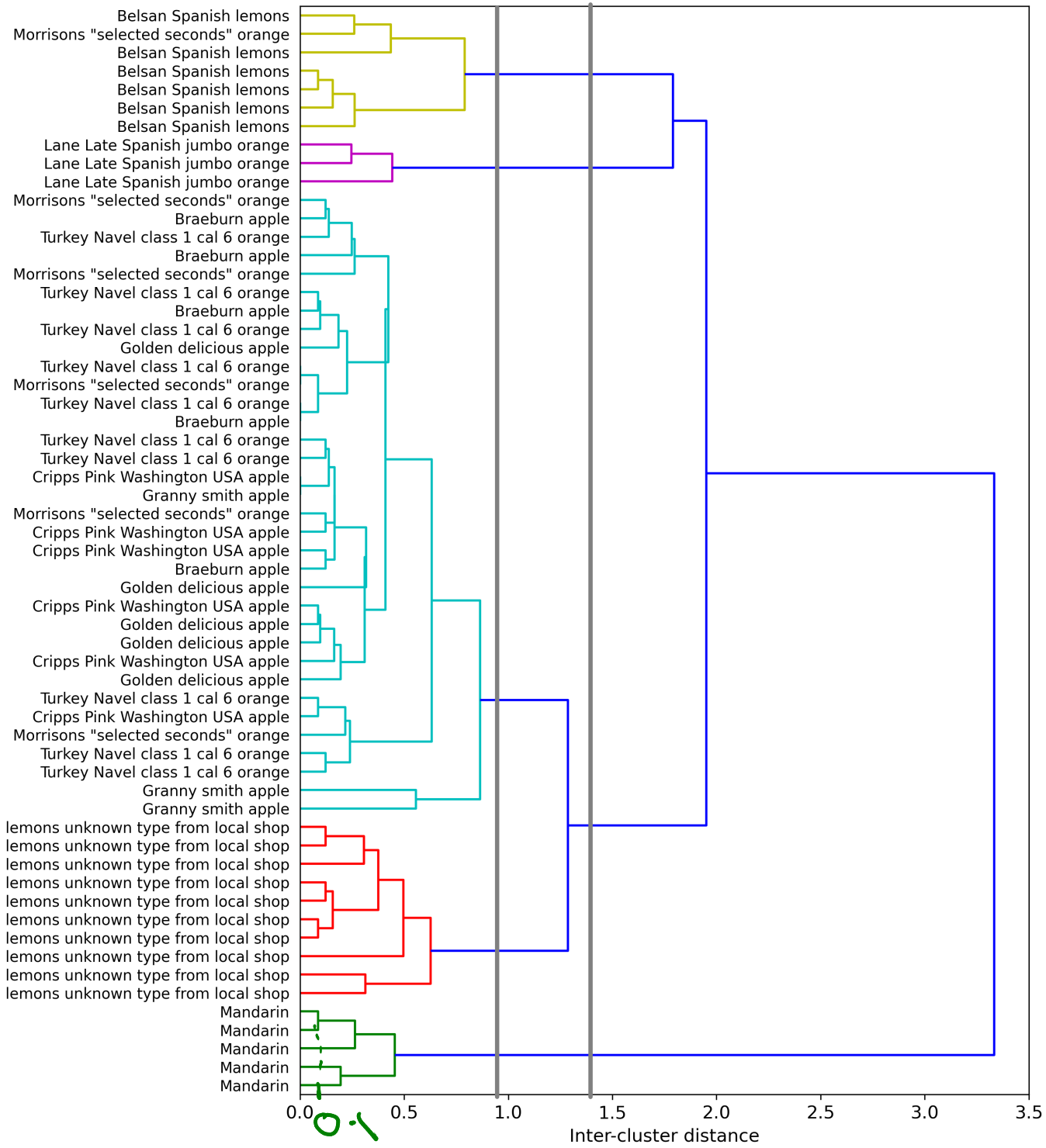




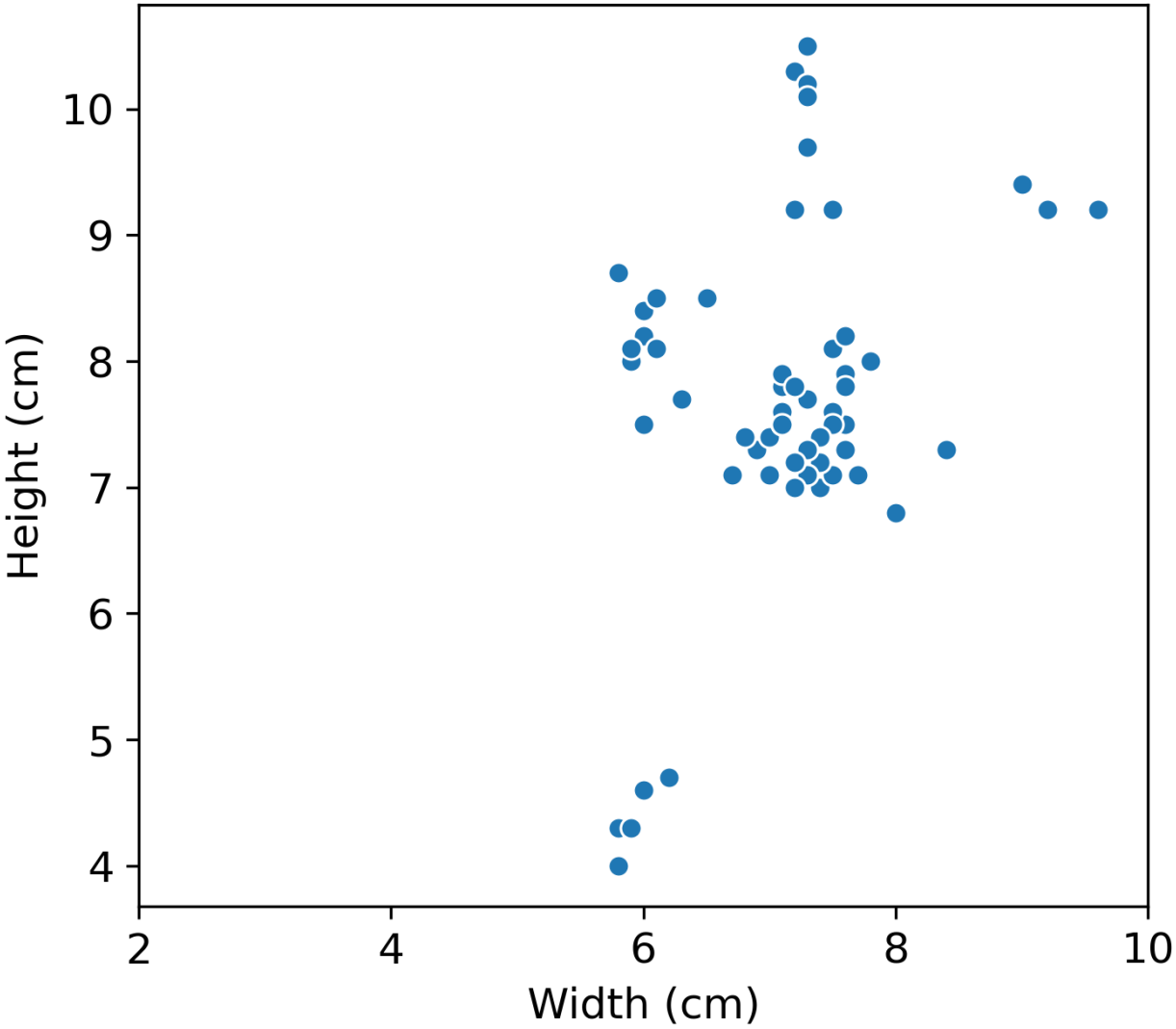
Data collected by Iain Murray
https://homepages.inf.ed.ac.uk/imurray2/teaching/oranges_and_lemons

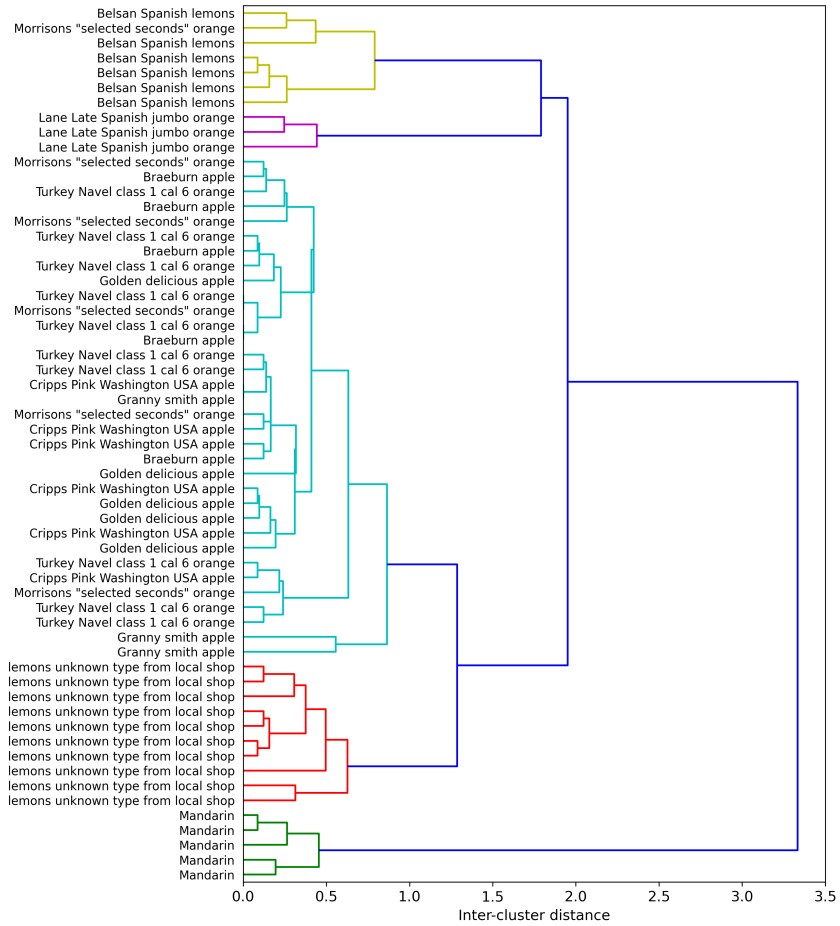
Hierarchical clustering

Dendrogram

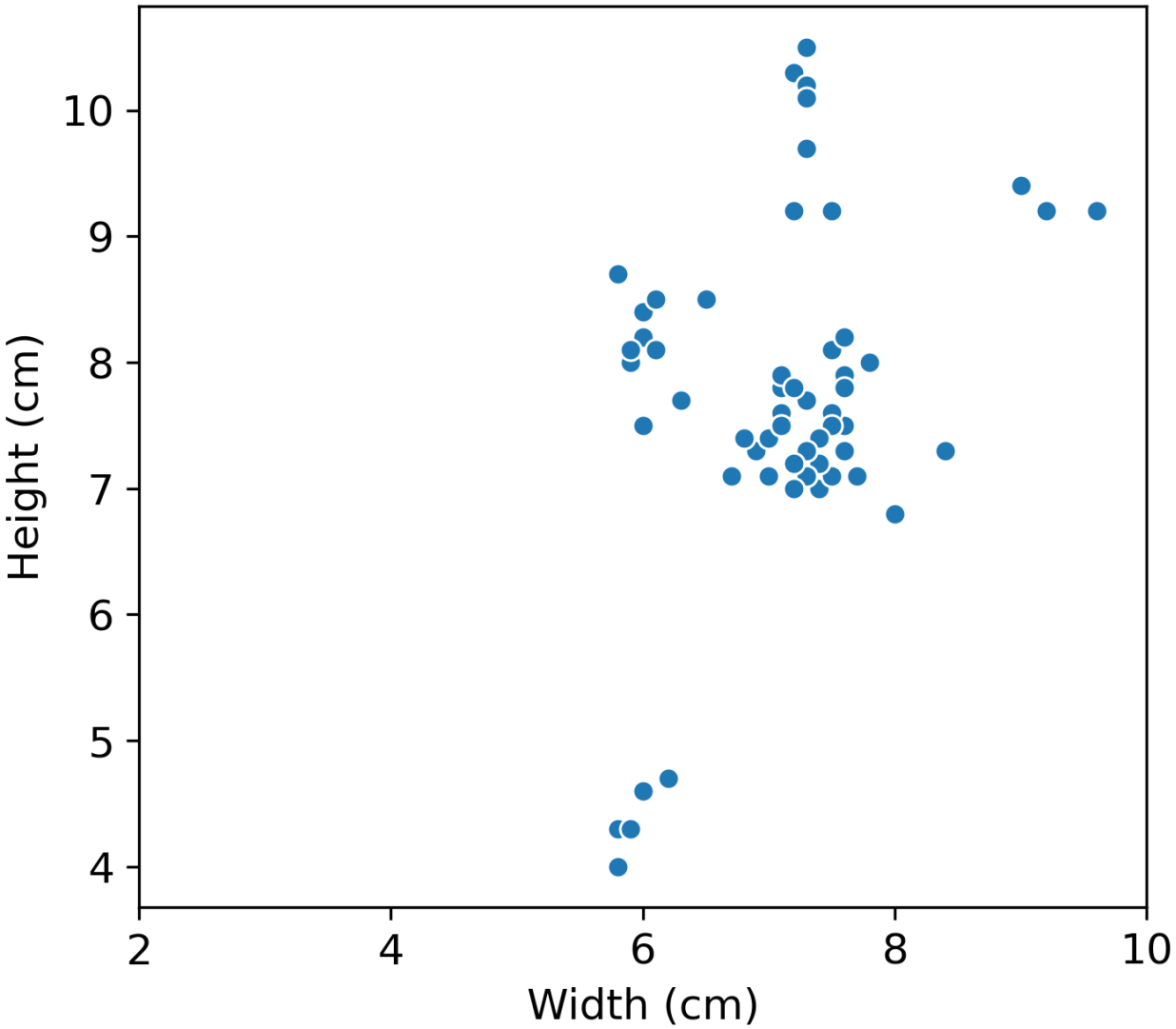


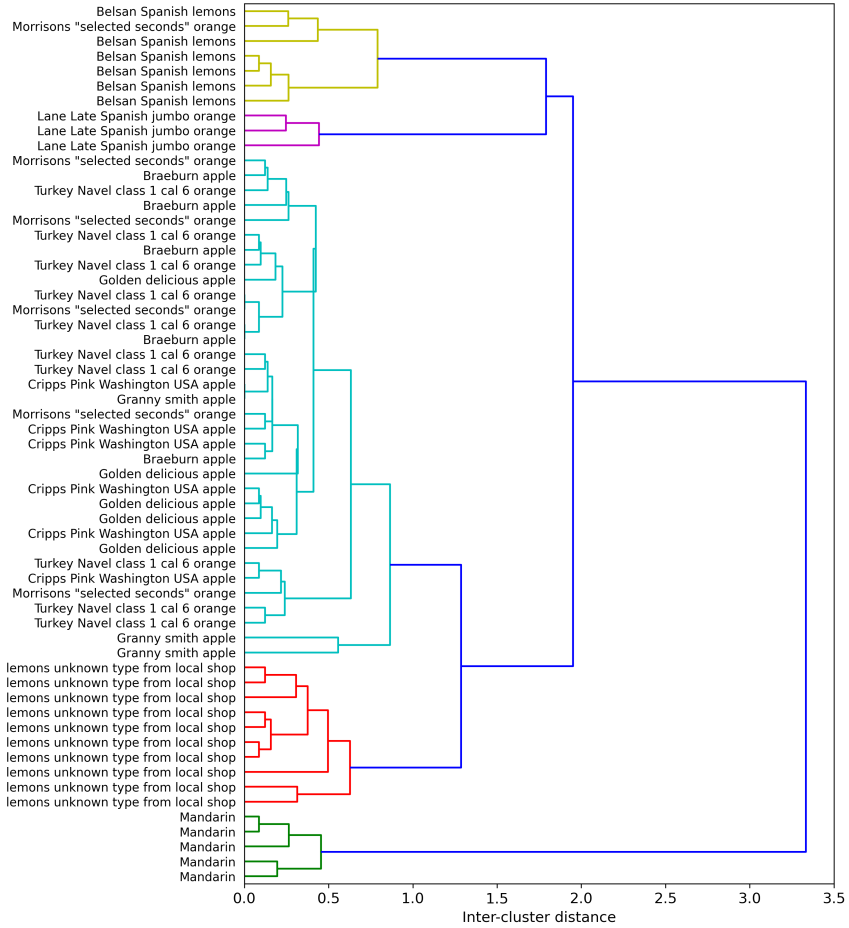
Top-down hierarchical clustering





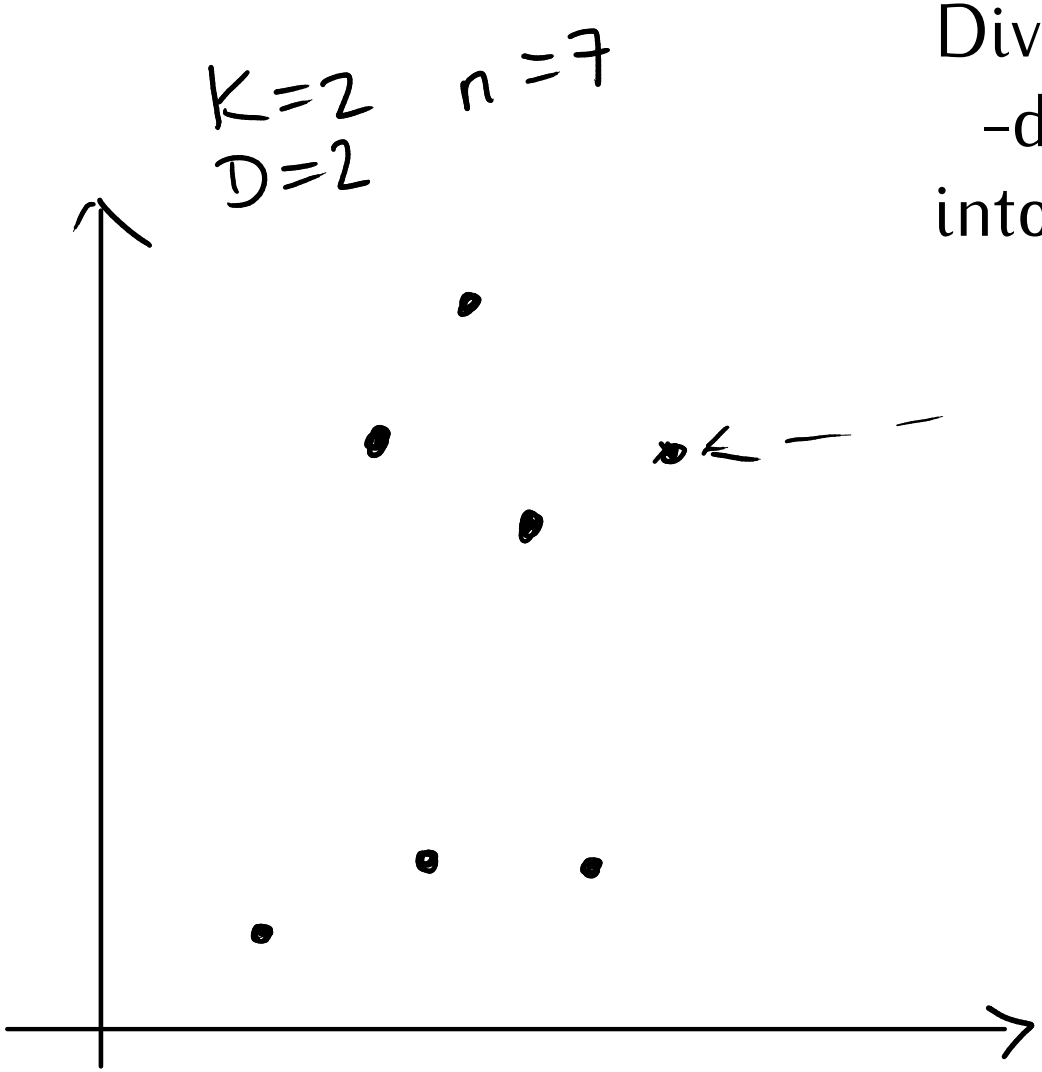
Agglomerative (bottom-up) hierarchical clustering





Foundations of Data Science: Introduction to unsupervised learning - K-means clustering

K-means algorithm



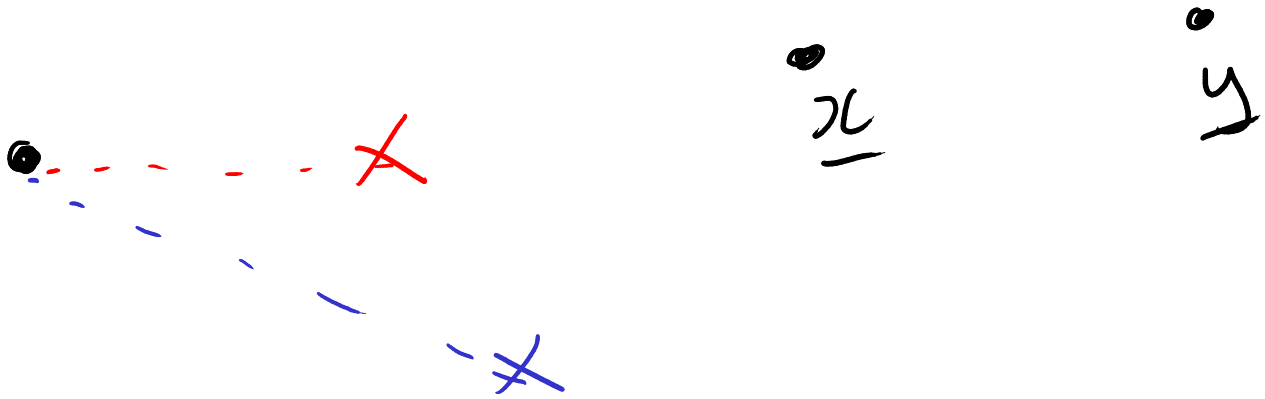
Problem:
Divide a set of
-dimensional data points
into clusters

Algorithm:
1. Initialise cluster centres

2. While not converged
   - Assign

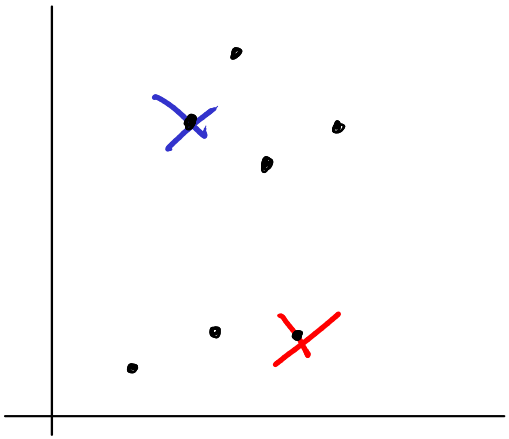
   - Recompute

Distance measure

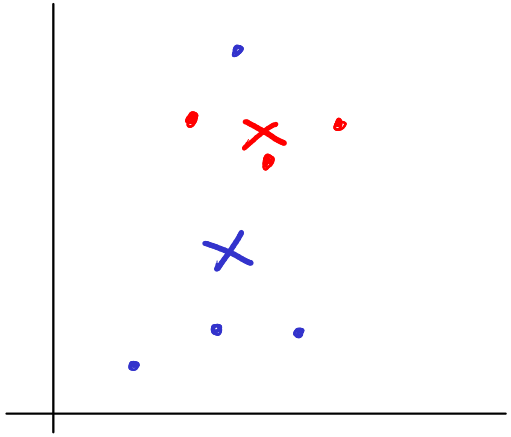


$$d(\underline{x}, \underline{y}) = \| \underline{x} - \underline{y} \| = \sqrt{\sum_{j=1}^{D} (x_j - y_j)^2}$$

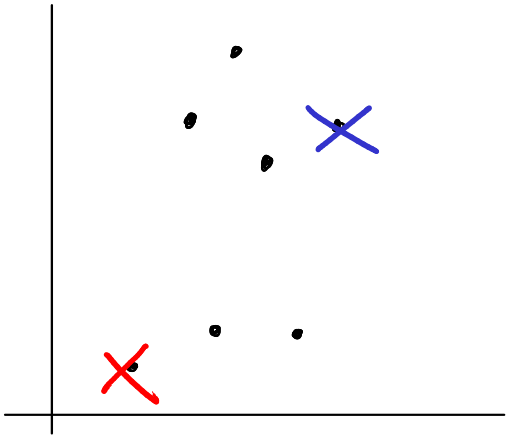
Initialisation methods



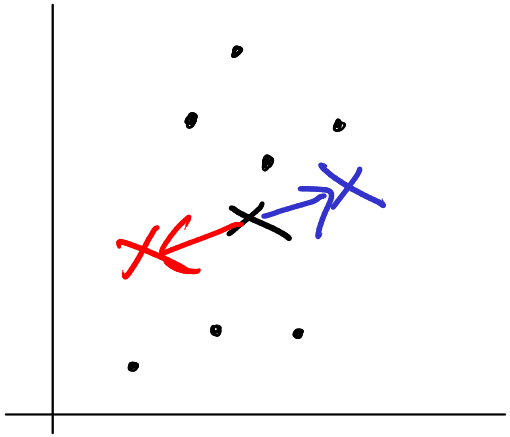
Random data points as cluster centres



Random assigment to clusters



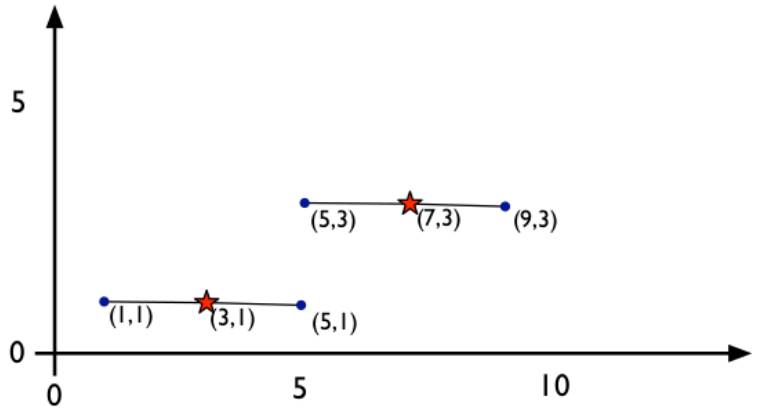
Data points with extreme values

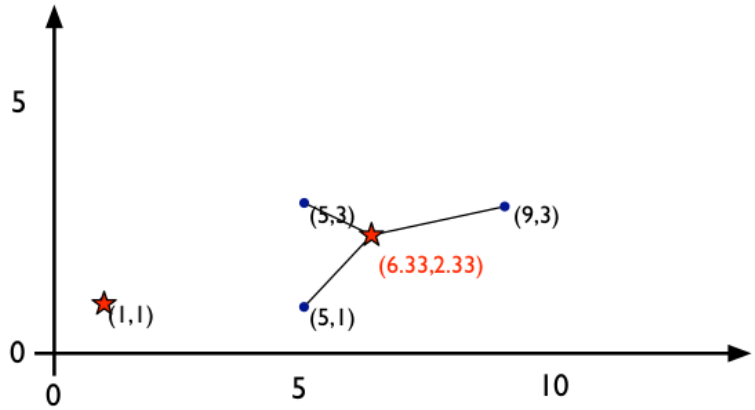


Mean for whole dataset and peturb

Convergence

Convergence is guaranteed...

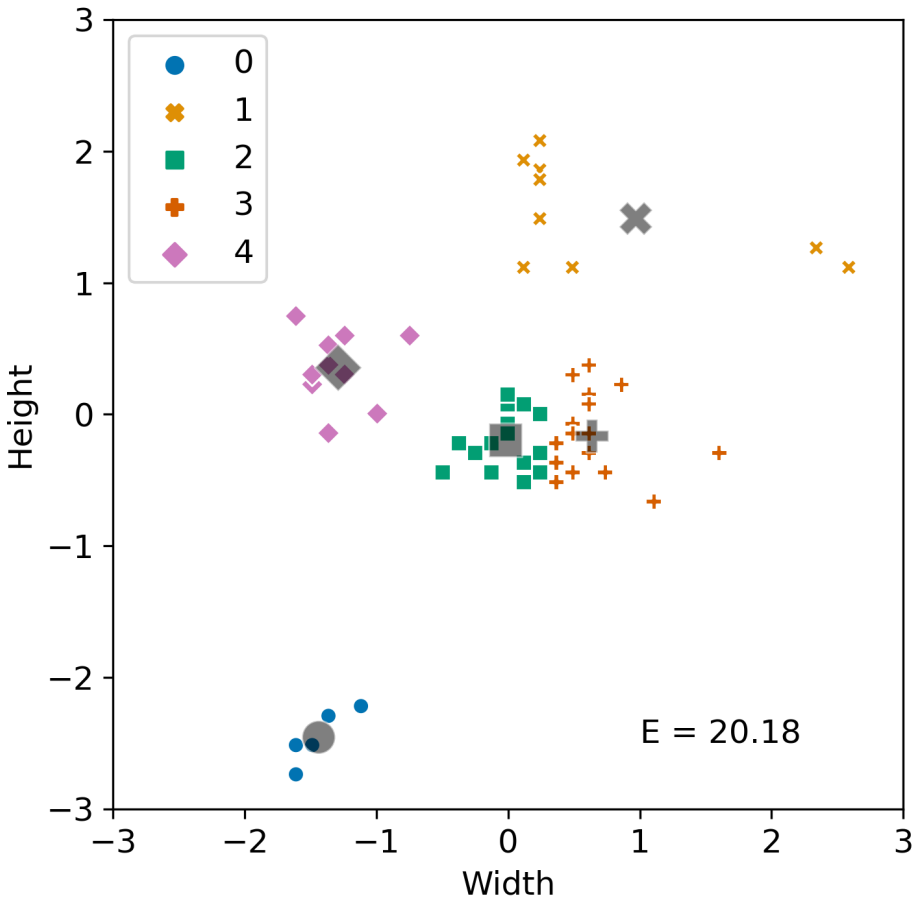


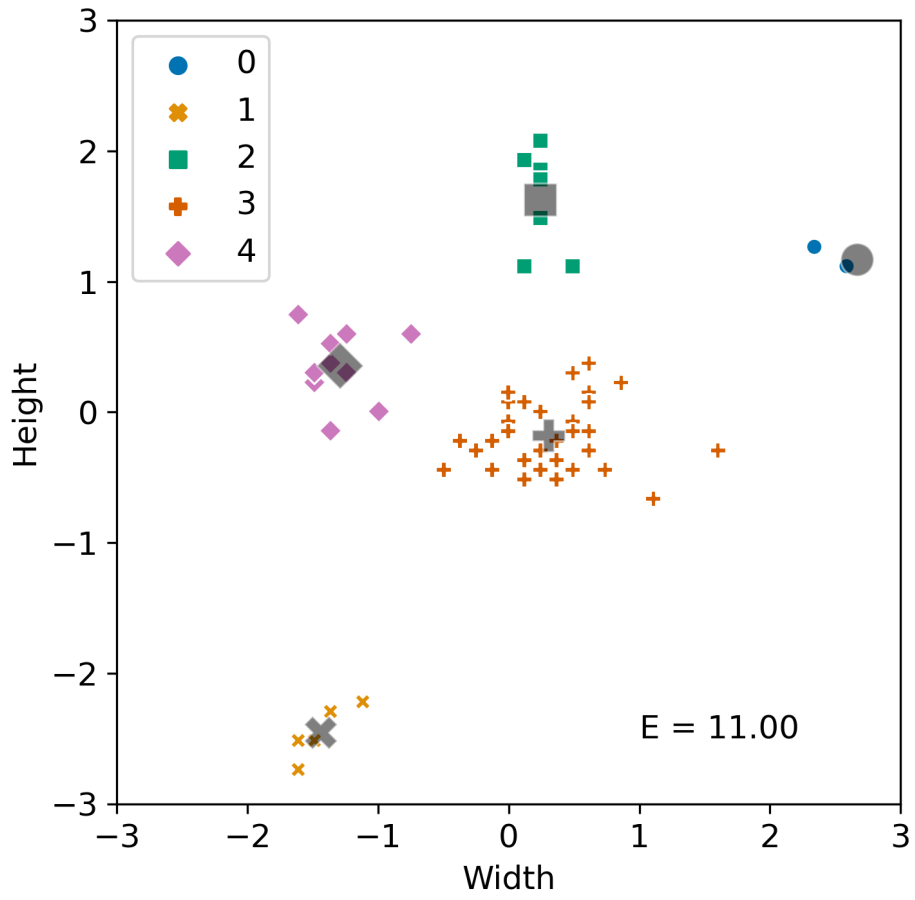


... but unique solutions are not.

Foundations of Data Science: Introduction to unsupervised learning - Evaluation and application of K-means clustering

Multiple solutions





Mean squared error function

AKA inertia

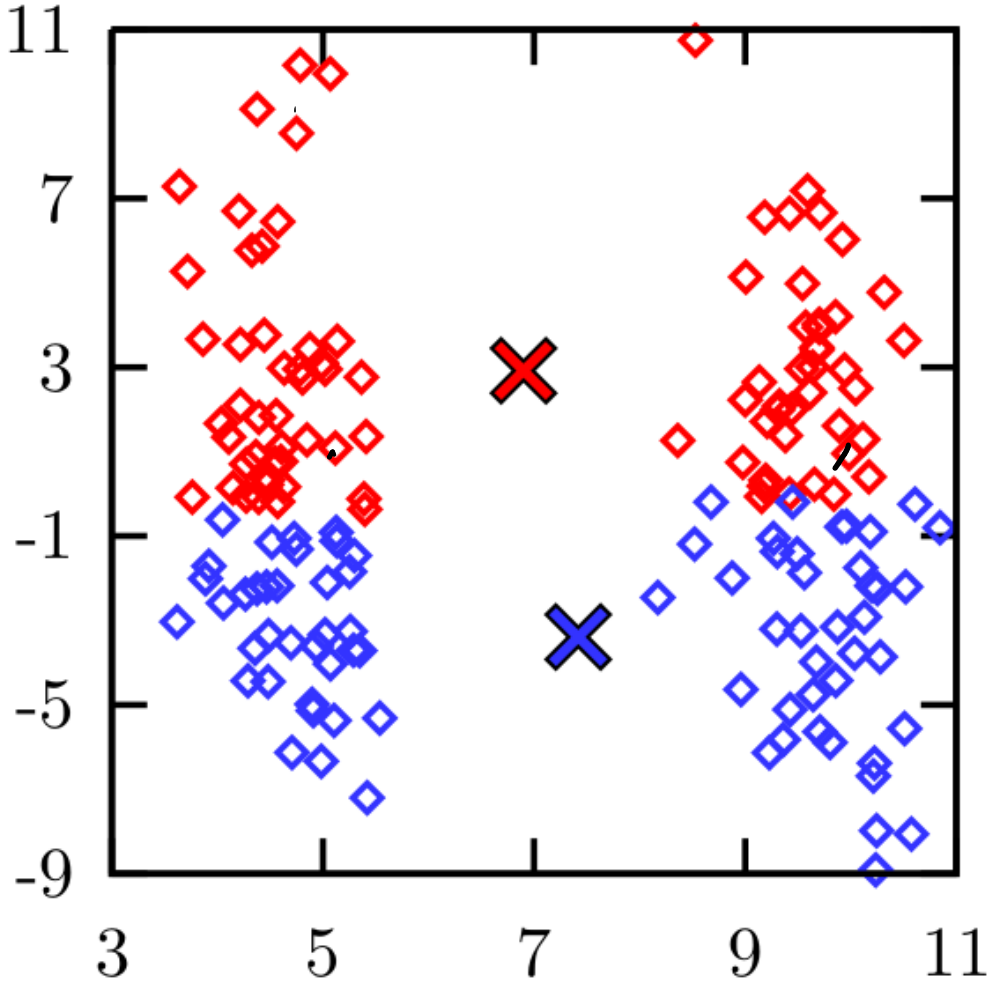
If point i belongs to cluster k

$i \in C_k$

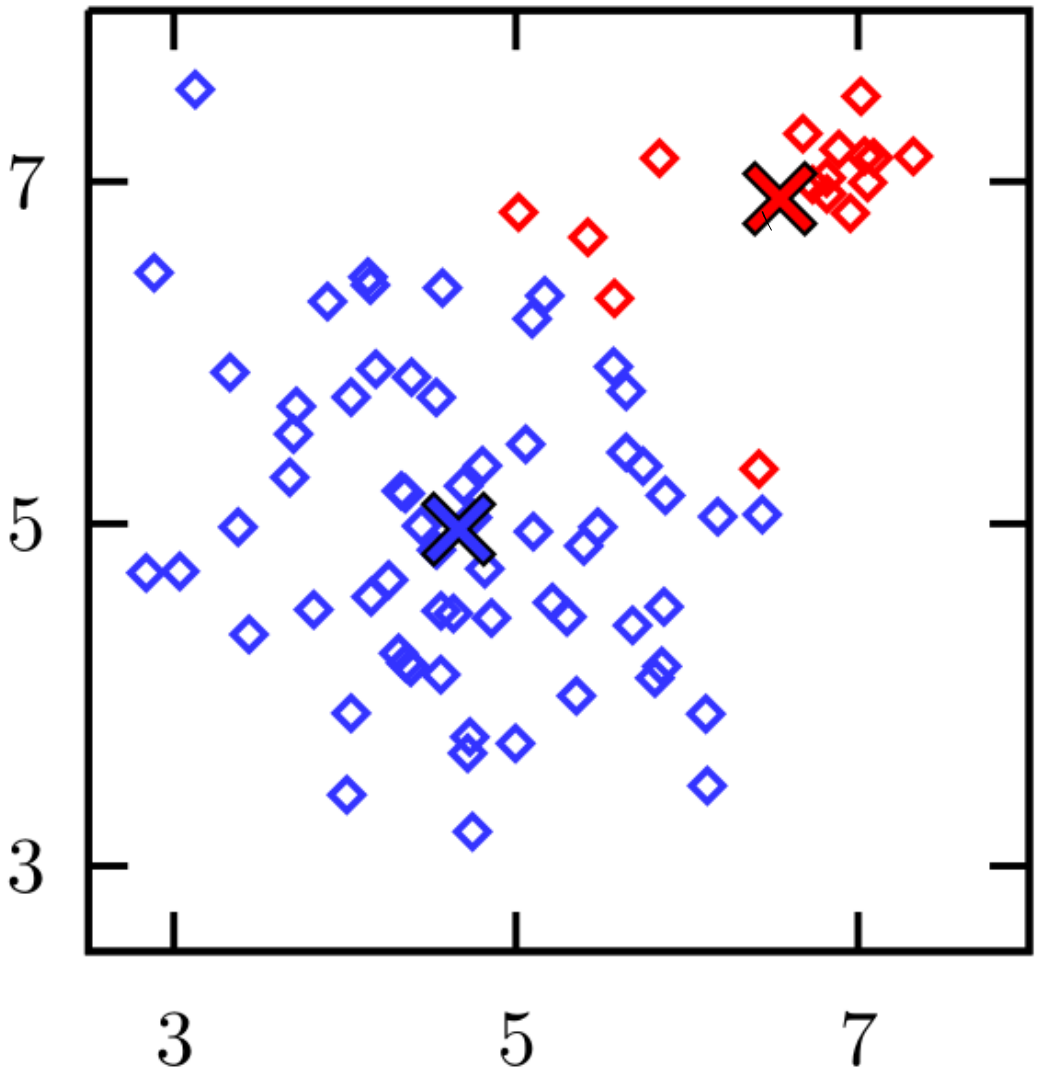
$$E = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in C_k} \| \underline{x}_i - \underline{m}_k \|^2$$

Minimum variance clustering

Failures of K-means (1)

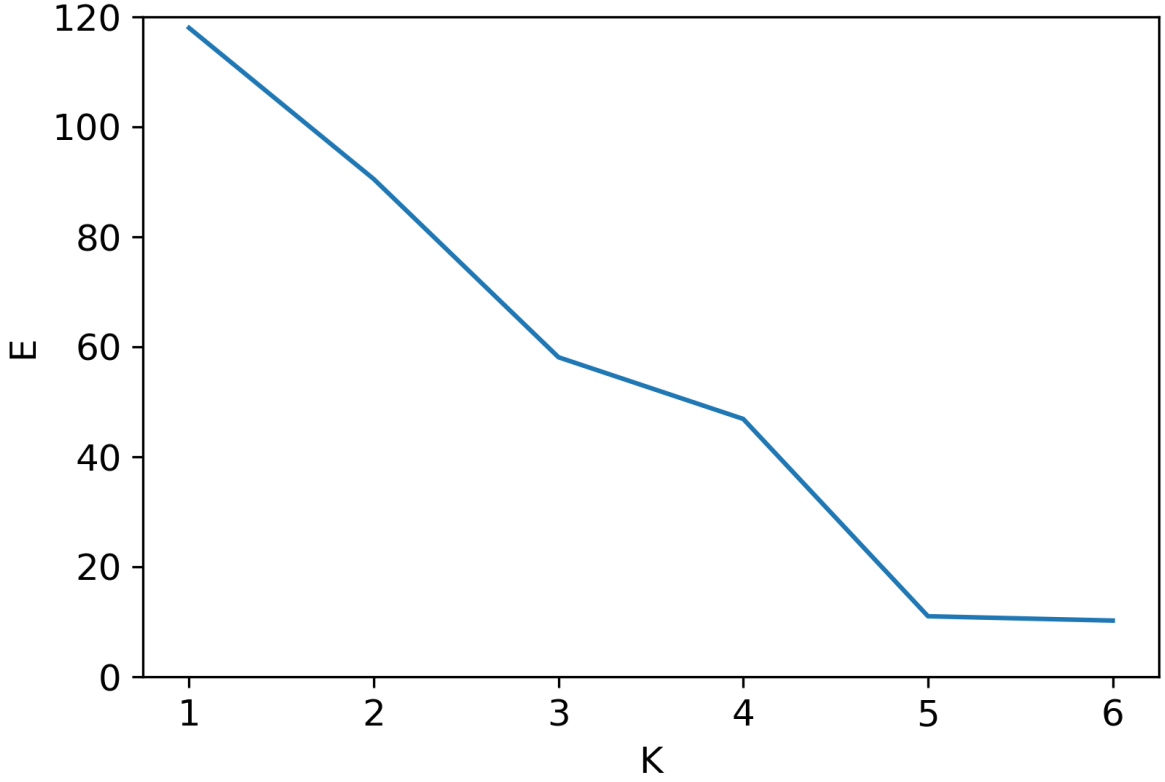


Failure of K-means (2)



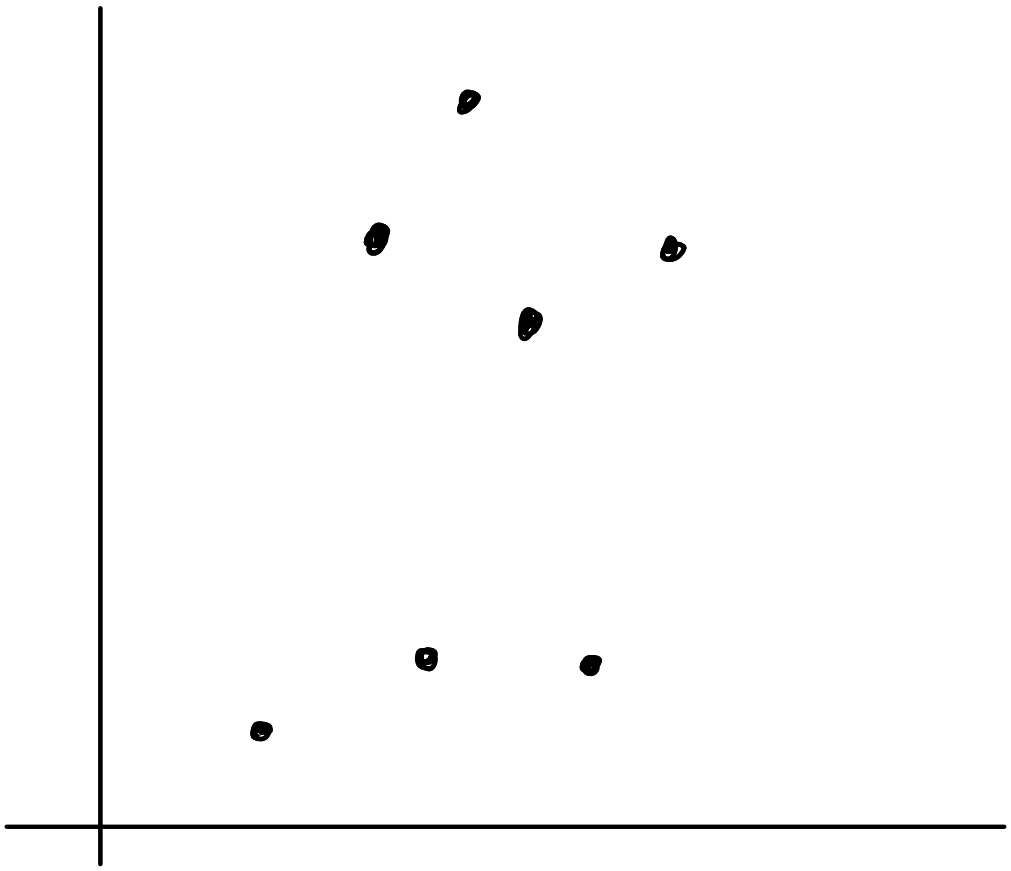
How to choose K?

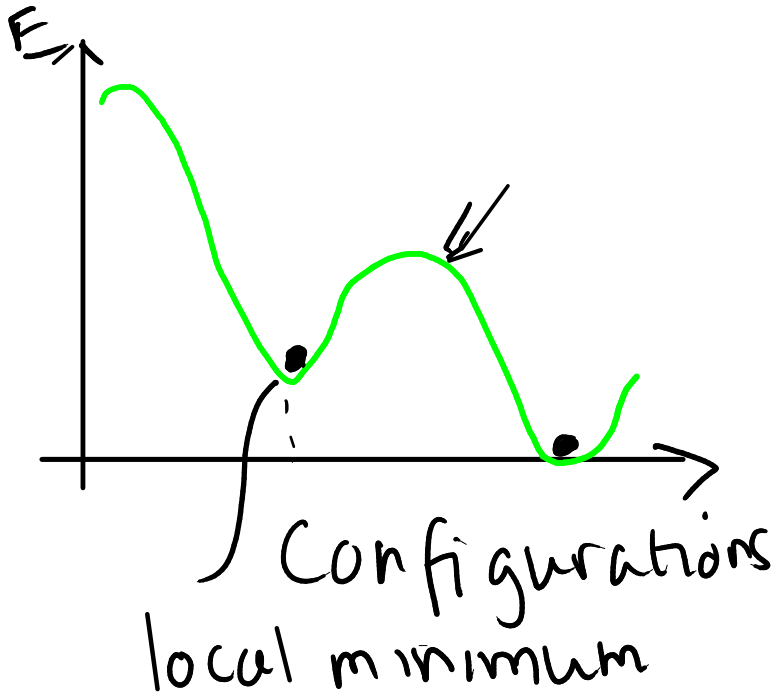
Scree plot



Batch versus online

Online versus batch





The curse of Dimensionality

$\Rightarrow$ Use dimensionality reduction to overcome the curse.