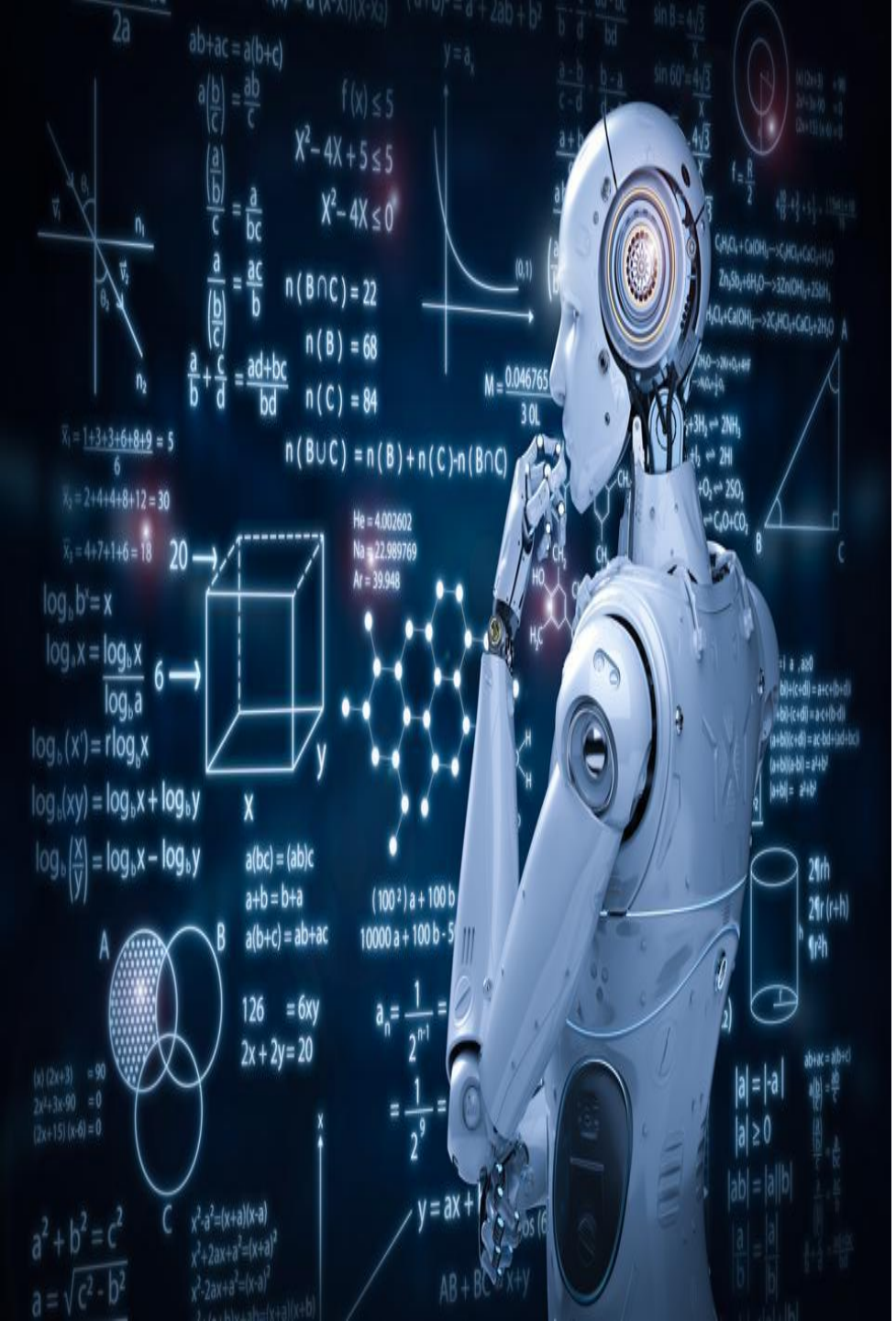


Practical Machine Learning

Day 14: Sep22 DBDA

Kiran Waghmare

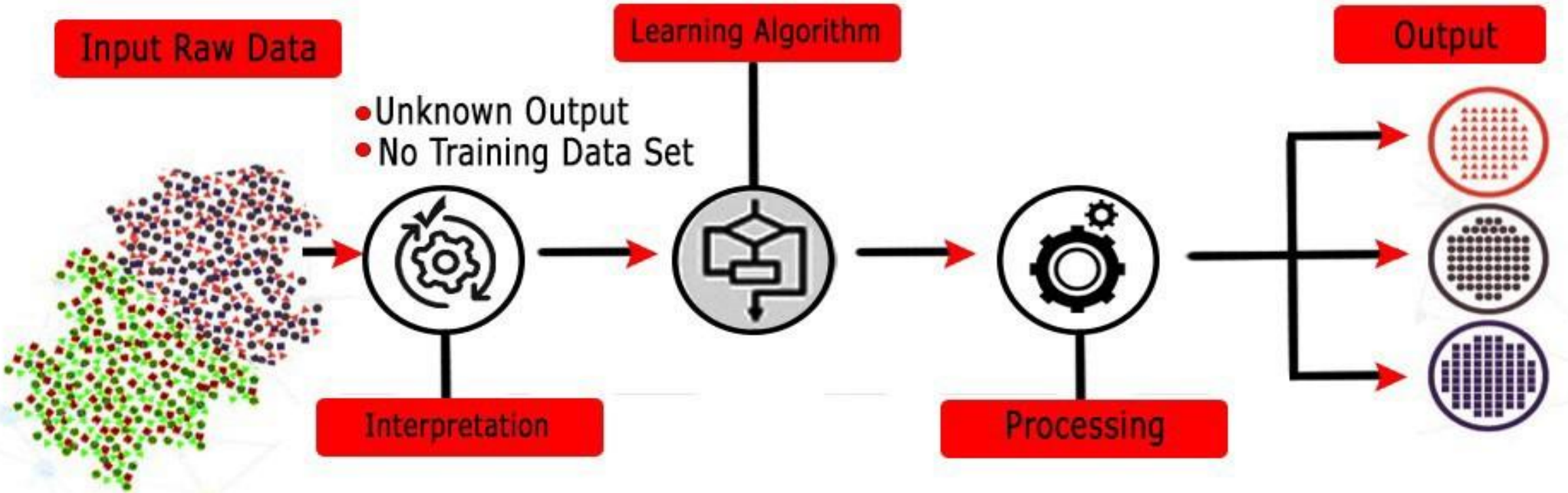
Agenda

- Clustering
- K-Means
- Hierarchical
- DB-SCAN

Machine learning:

- **Supervised vs Unsupervised.**
  - ***Supervised learning*** - the presence of the outcome variable is available to guide the learning process.
    - there **must** be a training data set in which the solution is already known.
  - ***Unsupervised learning*** - the outcomes are unknown.
    - cluster the data to reveal meaningful partitions and hierarchies

Unsupervised Learning

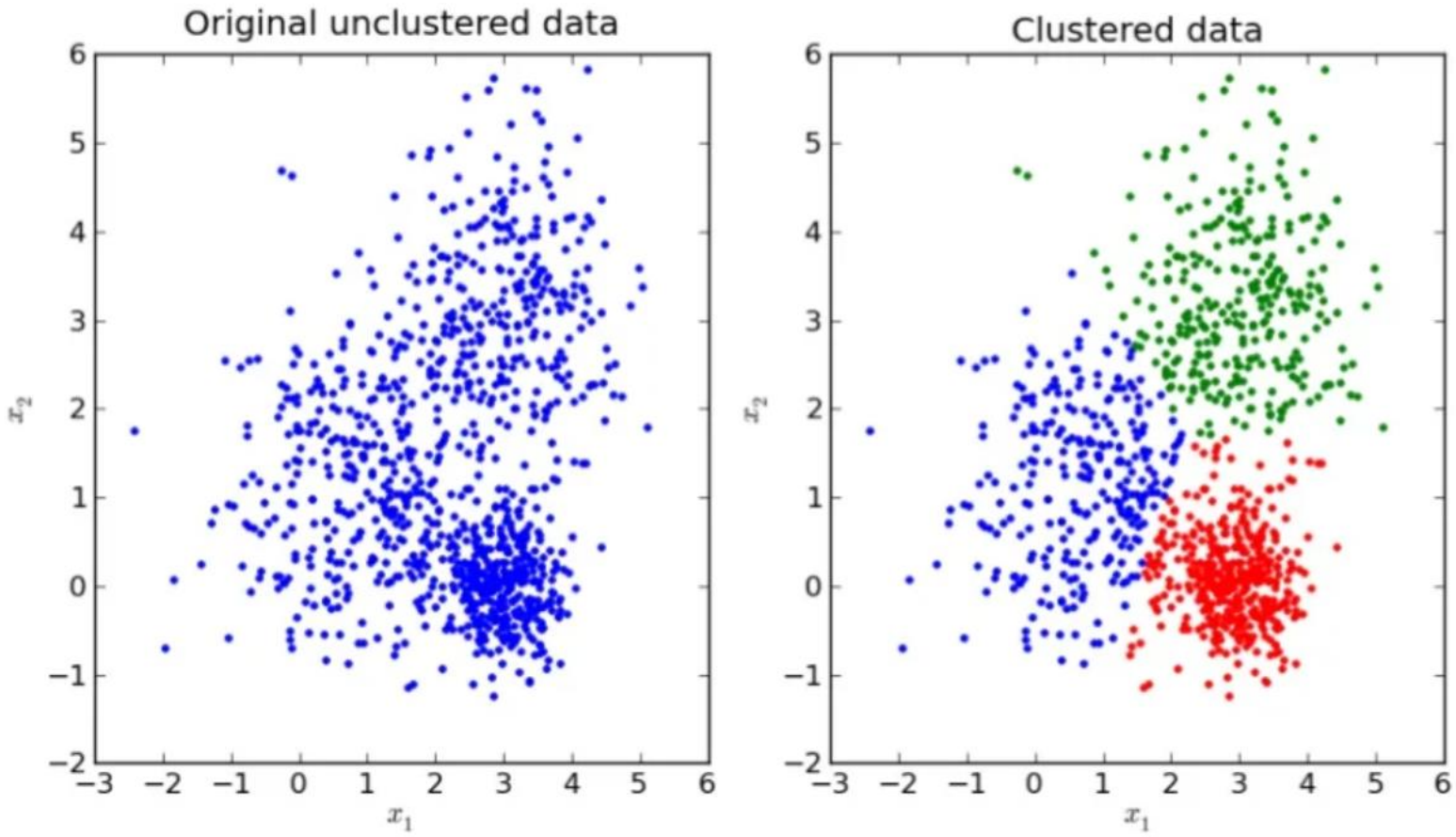


Clustering

Clustering:

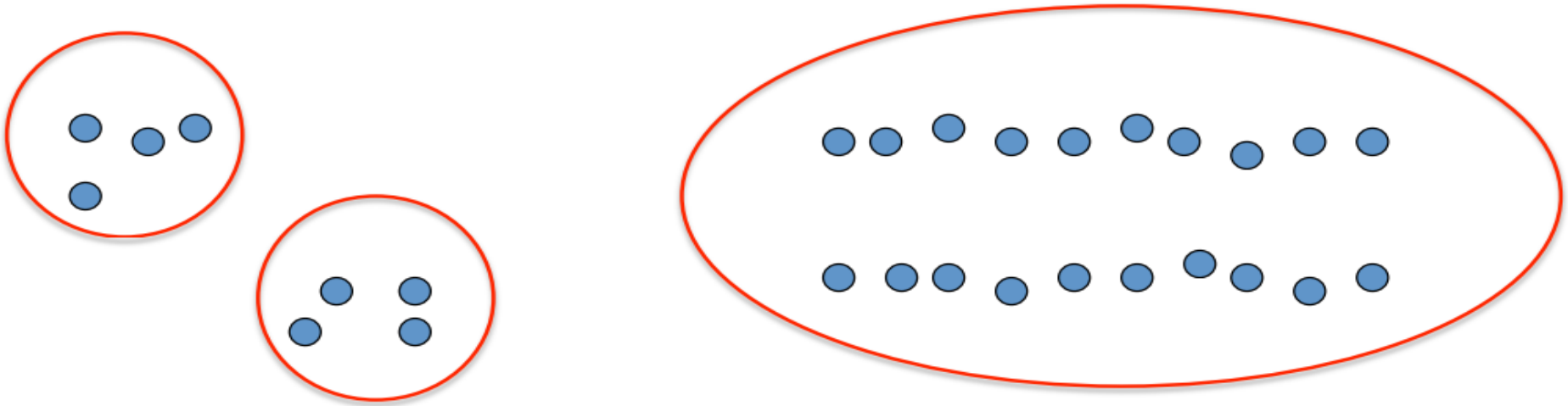
- Unsupervised learning
- Requires data, but no labels
- Detect patterns e.g. in
  - Group emails or search results
  - Customer shopping patterns
  - Regions of images
- Useful when don't know what you're looking for
- But: can get gibberish

- In this case clustering is carried out using the Euclidean distance as a measure.



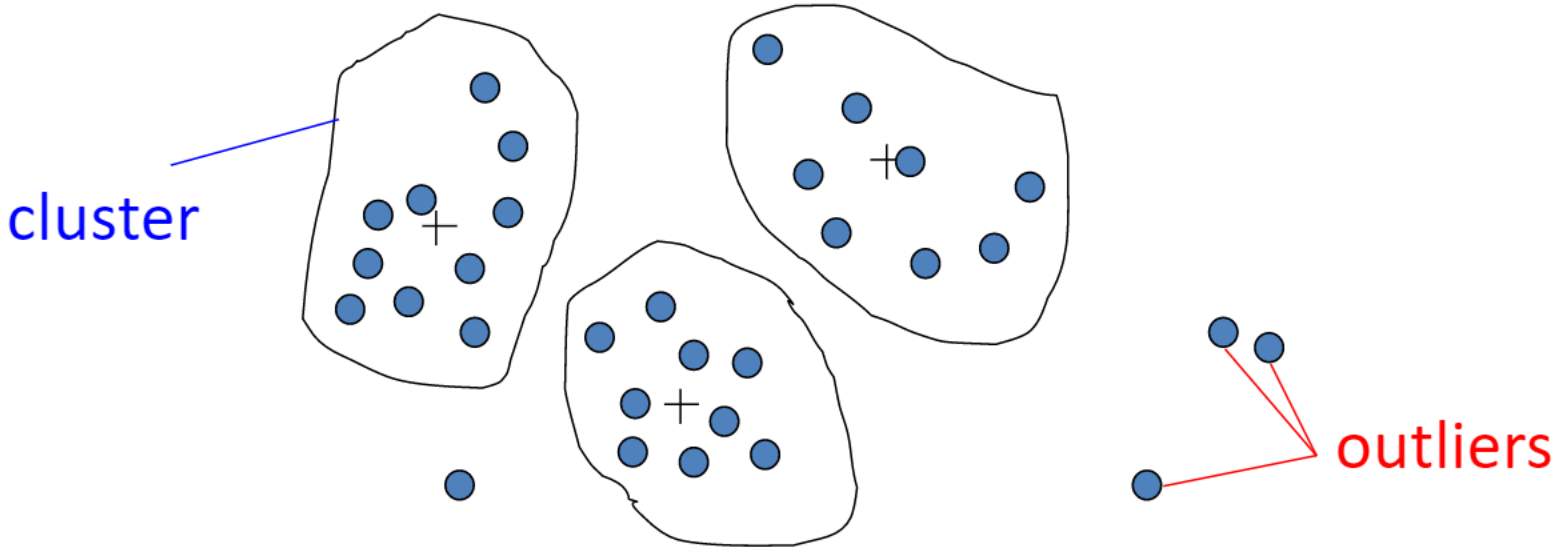
Clustering

- Basic idea: group together similar instances
- Example: 2D point patterns



Outliers

- **Outliers** are **objects that do not belong to any cluster** or form clusters of very small cardinality



- In some applications we are interested in discovering outliers, not clusters (outlier analysis)

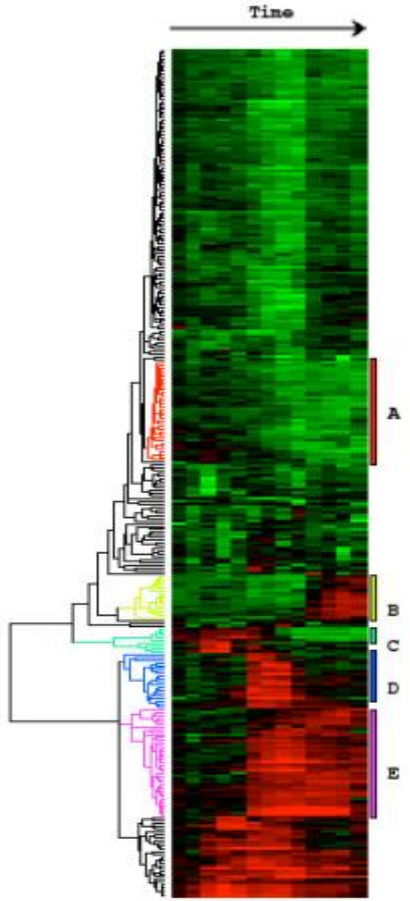
Clustering examples

Image segmentation

Goal: Break up the image into meaningful or perceptually similar regions

Clustering examples

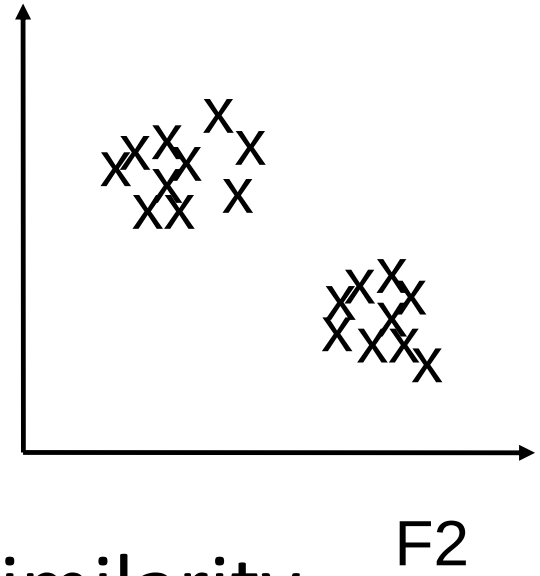
Clustering gene expression data



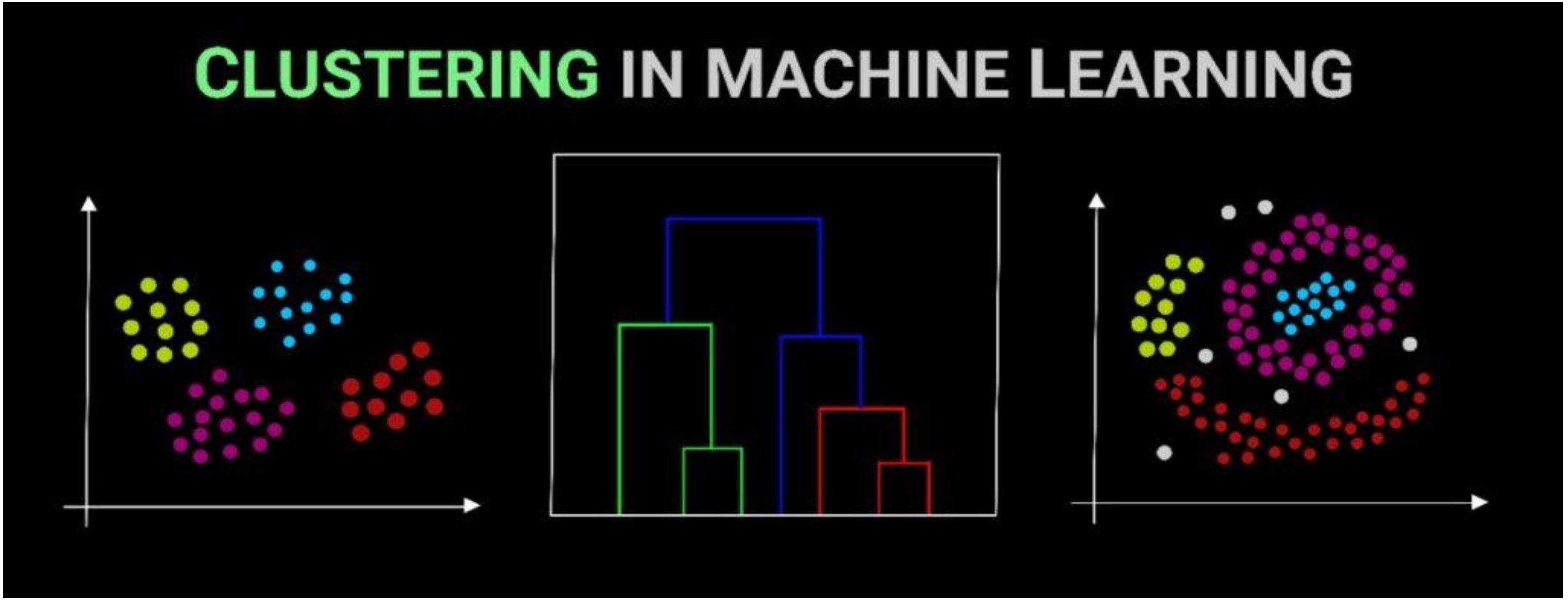
Eisen et al, PNAS 1998

Goal of Clustering

- Given a set of data points, each described by a set of attributes, find clusters such that:
  - Inter-cluster similarity is maximized
  - Intra-cluster similarity is minimized



- Requires the definition of a similarity measure

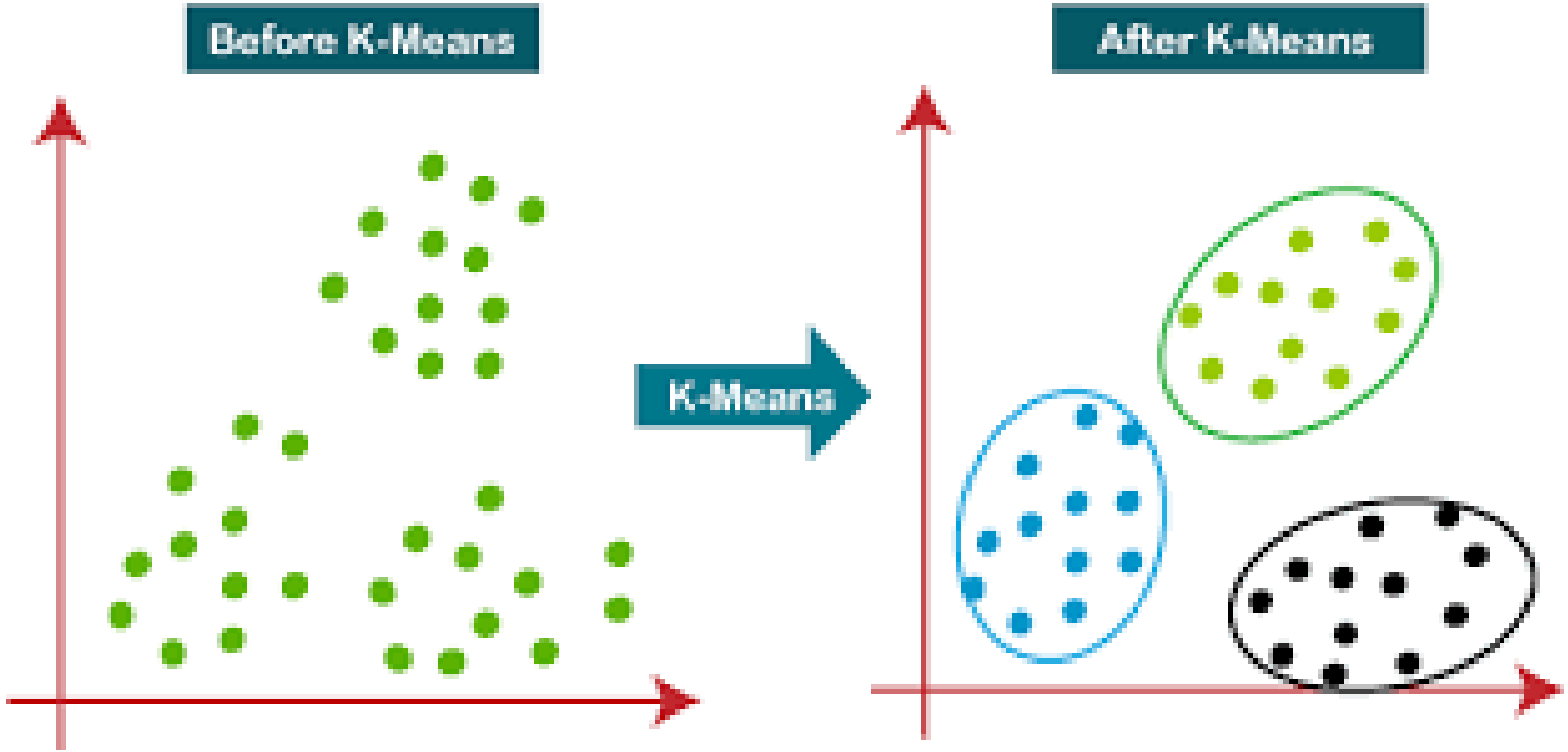

CLUSTERING IN MACHINE LEARNING

K-means algorithm:

1. Given n objects, initialize k cluster centers
2. Assign each object to its closest cluster centre
3. Update the center for each cluster
4. Repeat 2 and 3 until no change in each cluster center

- Experiment: Pack of cards, dominoes

- Apply the K-means algorithm to the Shapley data
  - Change the number of potential cluster and find how the clustering differ


Before K-Means
K-Means
After K-Means