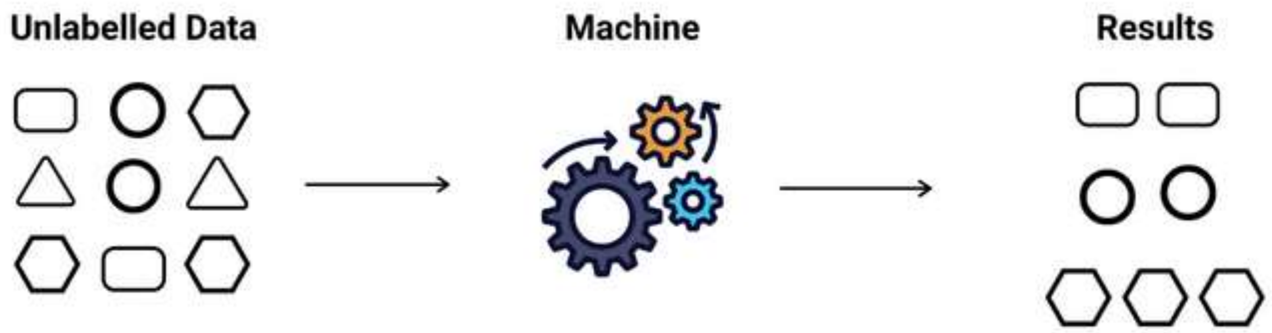
ML Fundamentals: Session 3

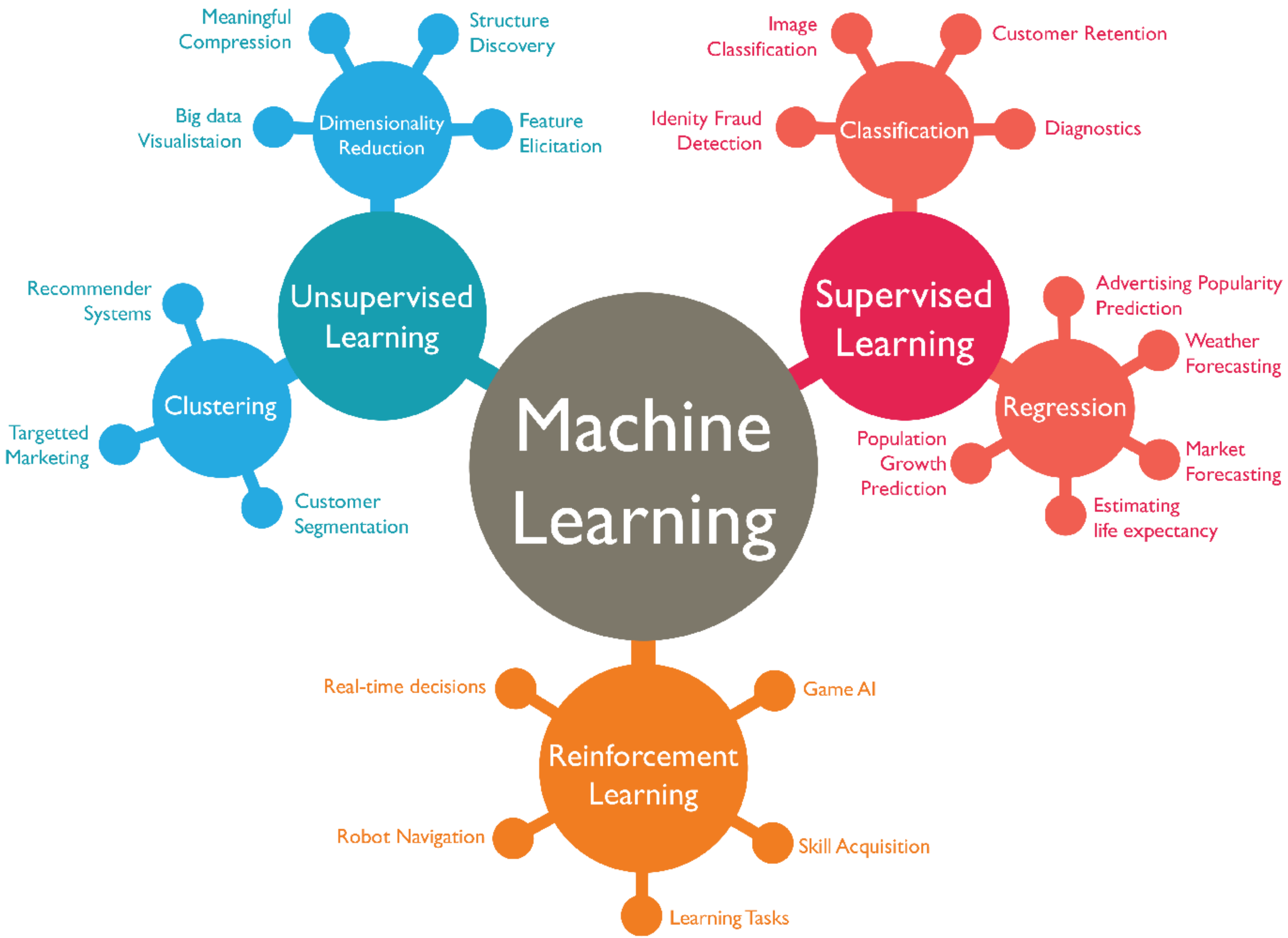
Unsupervised Learning with scikit-learn

Alia Hamwi

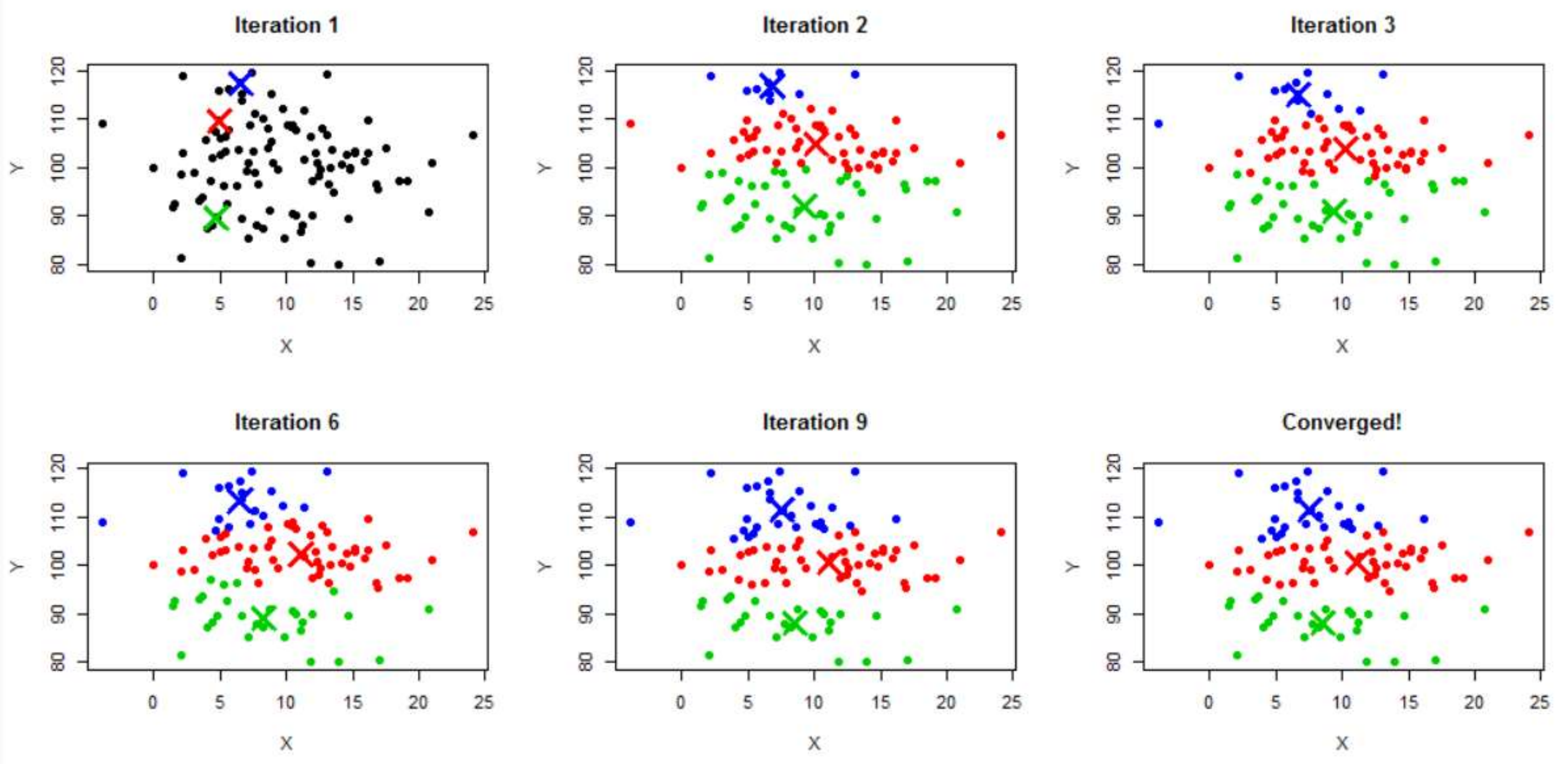
Types of Learning

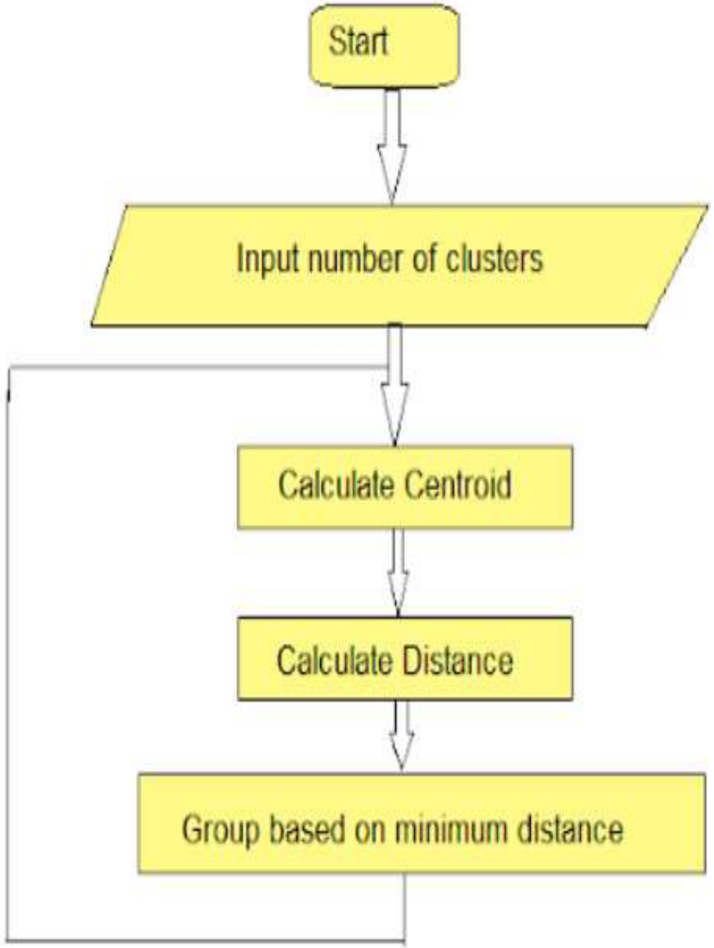
- Unsupervised learning
  -Given: training data (without desired outputs)



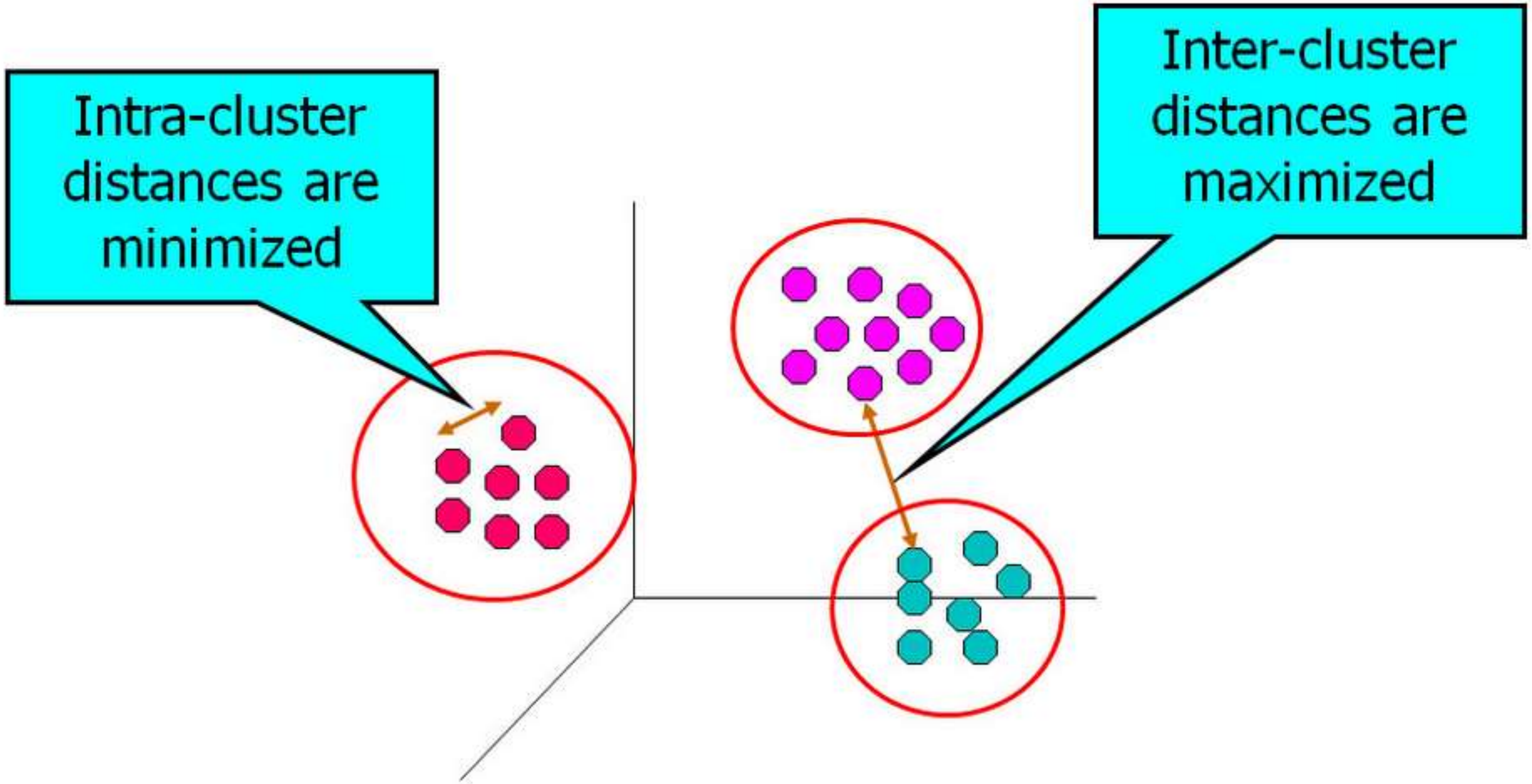

Machine Learning
Unsupervised Learning
Dimensionality Reduction
Meaningful Compression
Structure Discovery
Big data Visualistaion
Feature Elicitation
Clustering
Recommender Systems
Targetted Marketing
Customer Segmentation
Supervised Learning
Classification
Image Classification
Customer Retention
Idenity Fraud Detection
Diagnostics
Regression
Advertising Popularity Prediction
Weather Forecasting
Population Growth Prediction
Market Forecasting
Estimating life expectancy
Reinforcement Learning
Real-time decisions
Game AI
Robot Navigation
Skill Acquisition
Learning Tasks

Clustering: K-means



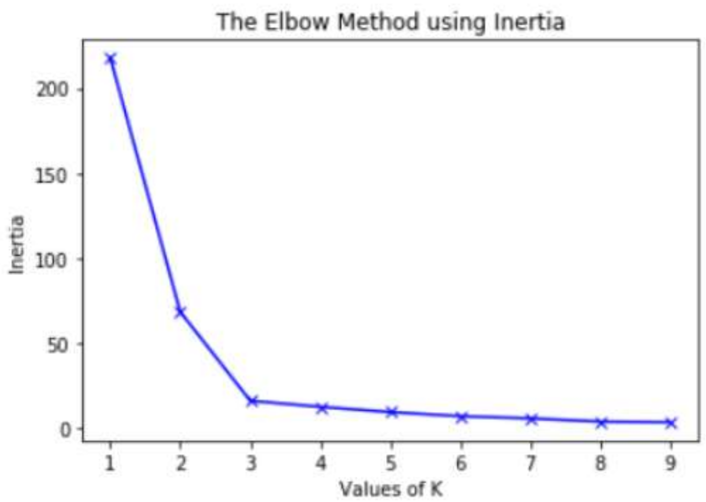


Measuring Performance of K-means



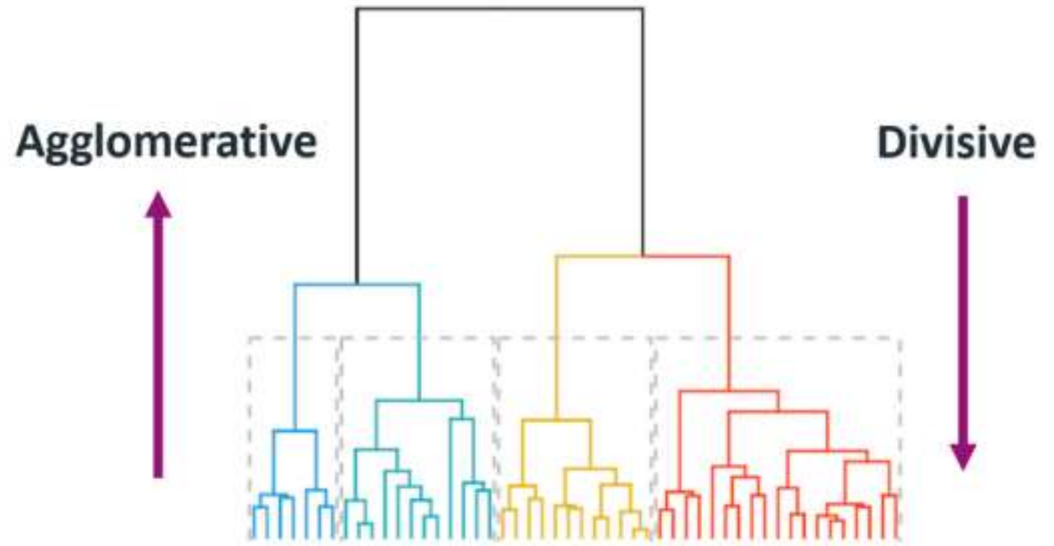
How to Choose K (# of Clusters)

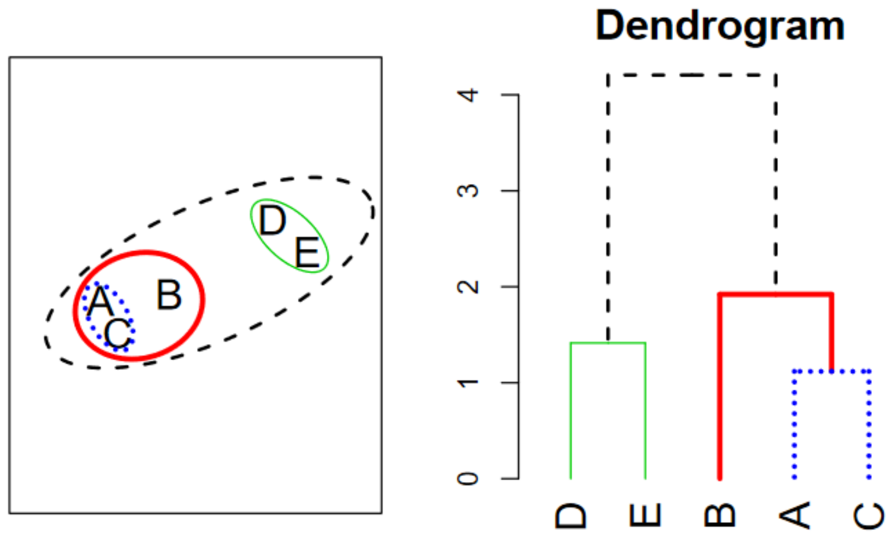
- To determine the optimal number of clusters, we have to select the value of k at the “elbow” ie the point after which the distortion/inertia start decreasing in a linear fashion. Thus for the given data, we conclude that the optimal number of clusters for the data is **3**.
- inertia tells how far away the points within a cluster are. Therefore, a small of inertia is aimed for.



Hierarchical Clustering

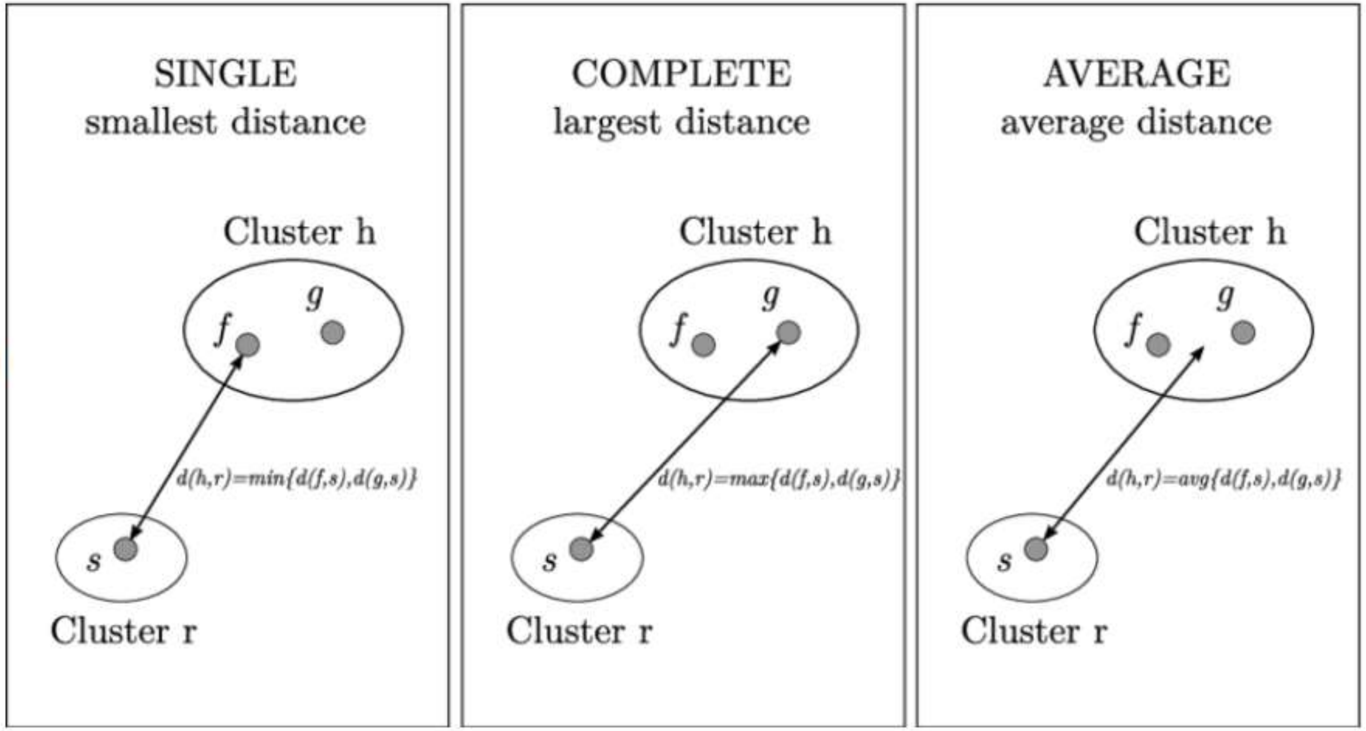
- The approach in words:
  - Start with each point in its own cluster.
  - Identify the closest two clusters and merge them.
  - Repeat.
  - Ends when all points are in a single cluster.





Hierarchical Clustering

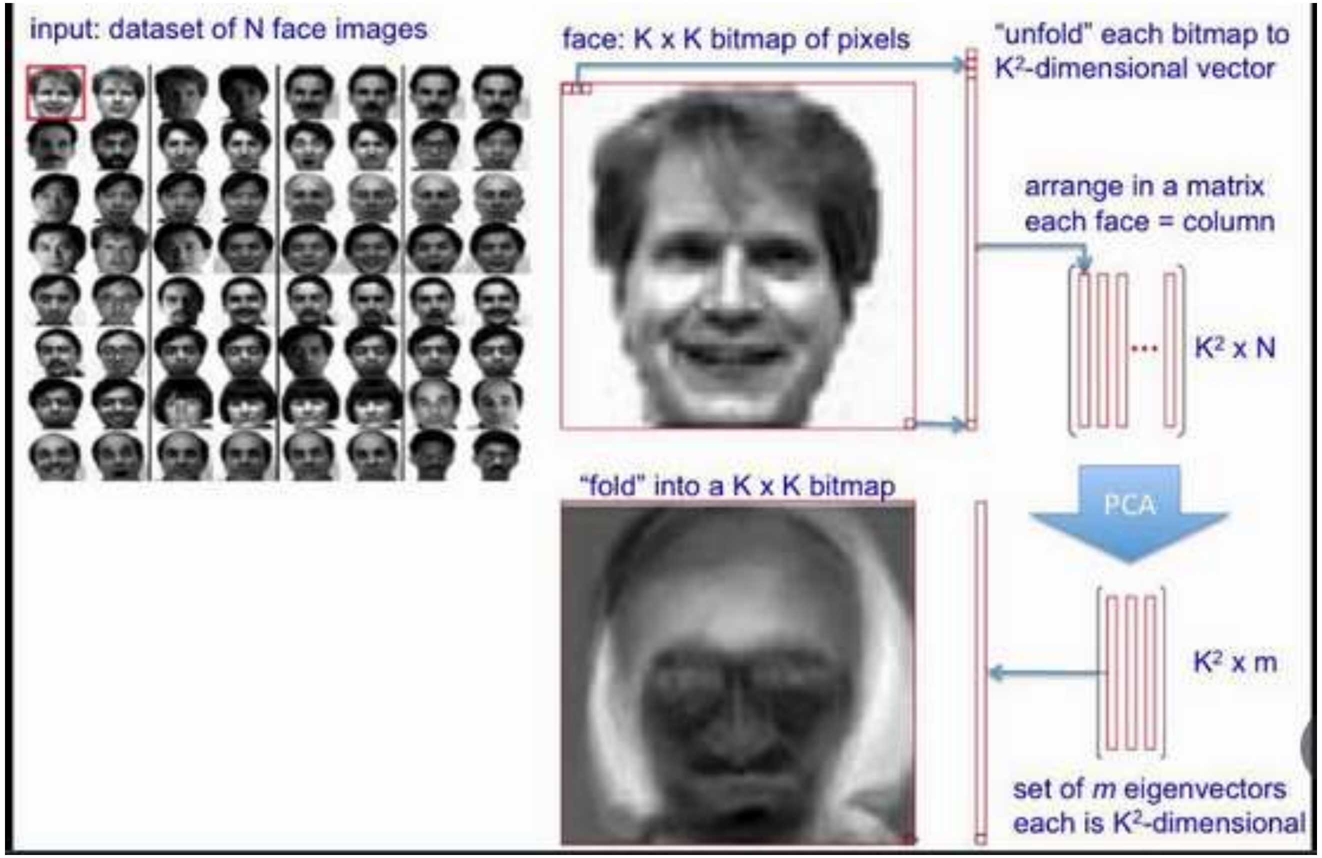
- Different linkage methods for hierarchical clustering



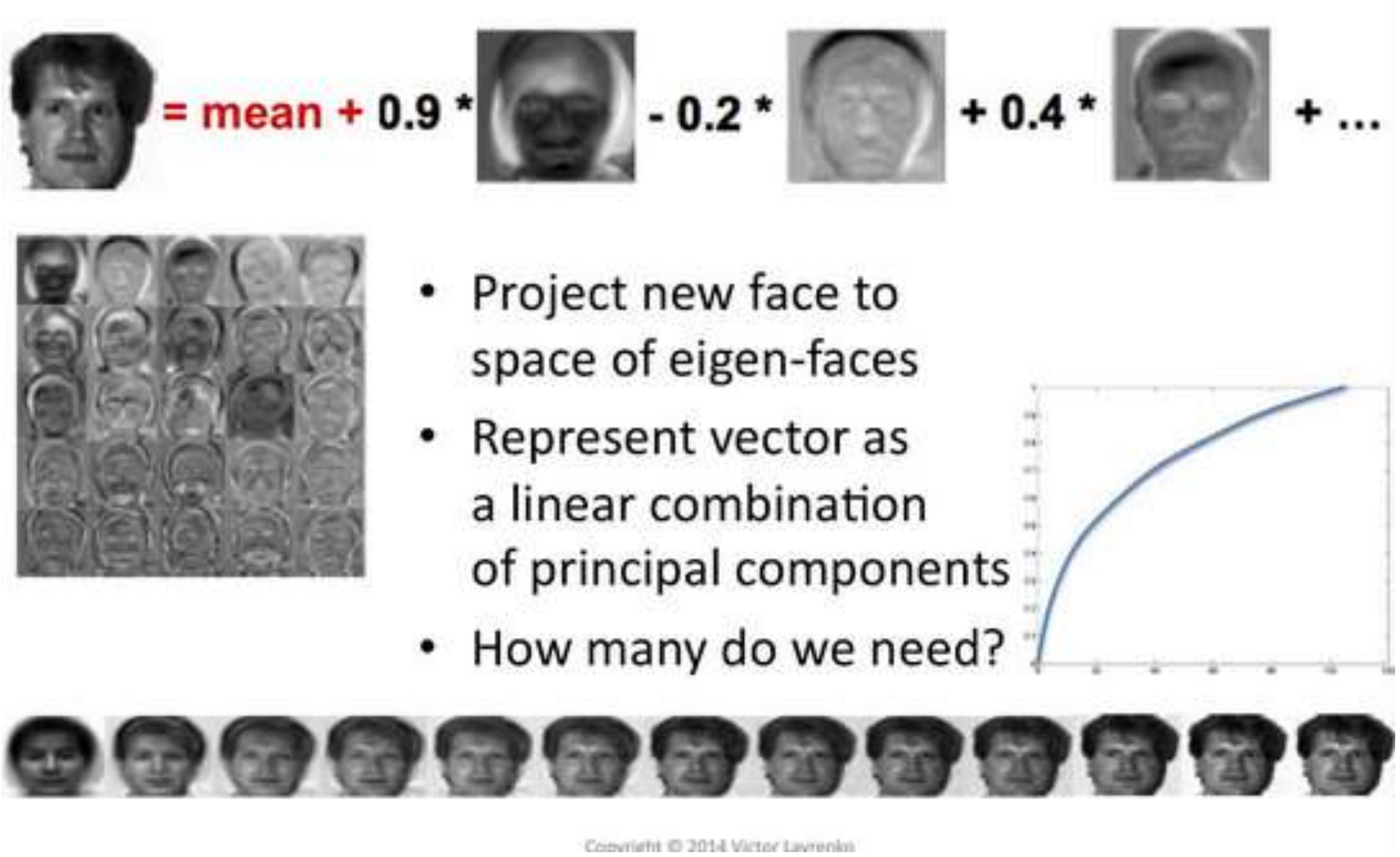
Principal component analysis (PCA)

- PCA produces a low-dimensional representation of a dataset. It finds a sequence of linear combinations of the variables that have maximal variance, and are <u>mutually uncorrelated</u>
- Apart from producing derived variables for use in supervised learning problems, PCA also serves as a tool for data visualization.

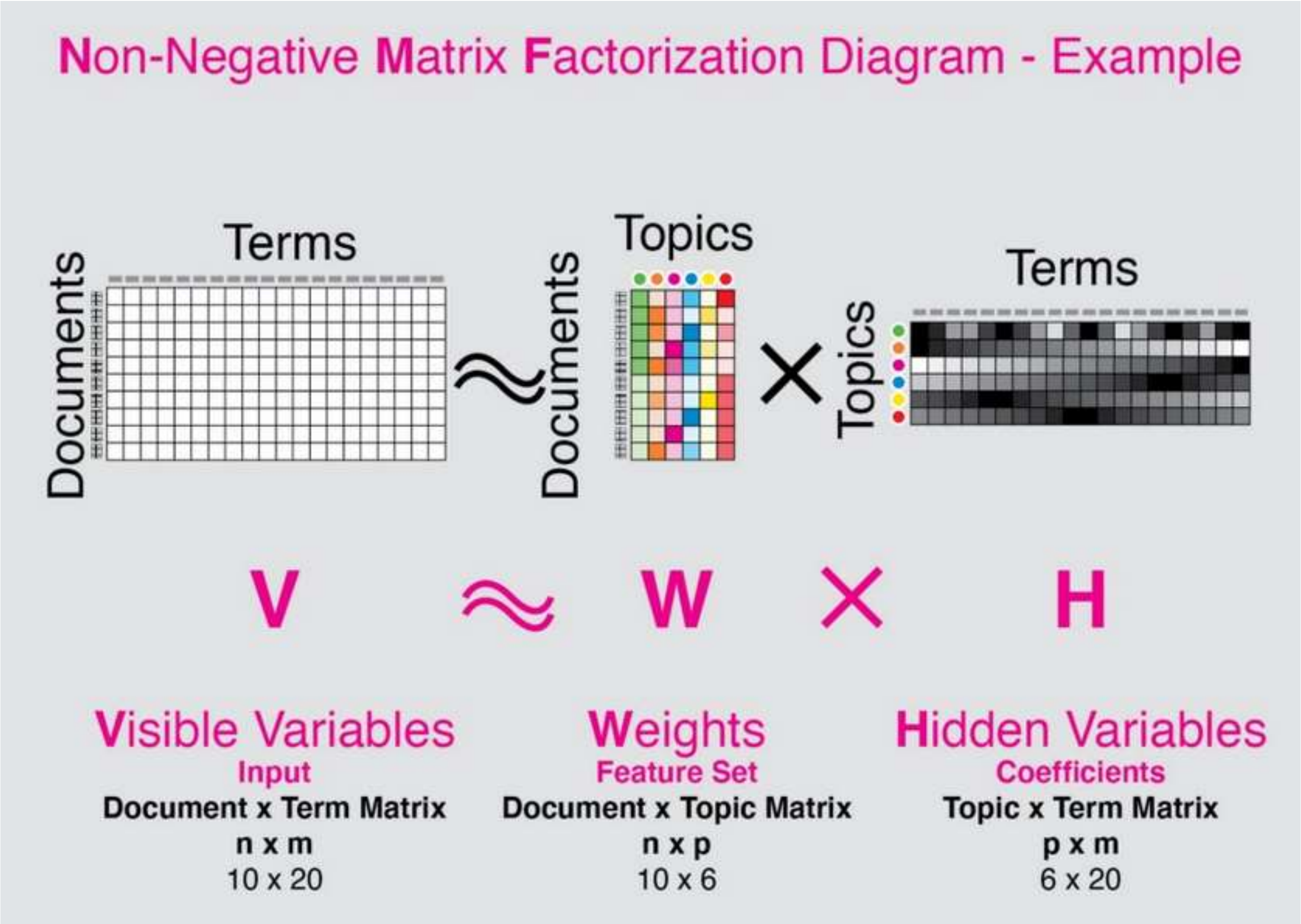
Face recognition: Eigenfaces



Face recognition: Eigenfaces



Non-negative Matrix-Factorization



Non-negative Matrix-Factorization

- In the example above, the Topics (p) are set to 6. Each column of the W matrix represents a probability that the topic is in the document. Each row of the W matrix represents a distribution of topic frequencies in each Document. Each row of the H matrix represent the distribution of term frequencies in each topic, and can be seen as the degree to which each term is activated in each topic.

References

- https://towardsdatascience.com/nmf-a-visual-explainer-and-python-implementation-7ecdd73491f8
- https://www.pyimagesearch.com/2021/05/10/opencv-eigenfaces-for-face-recognition/

Thank You