EVML3

UNSUPERVISED ML

JEROEN VEEN

HAN_UNIVERSITY
OF APPLIED SCIENCES

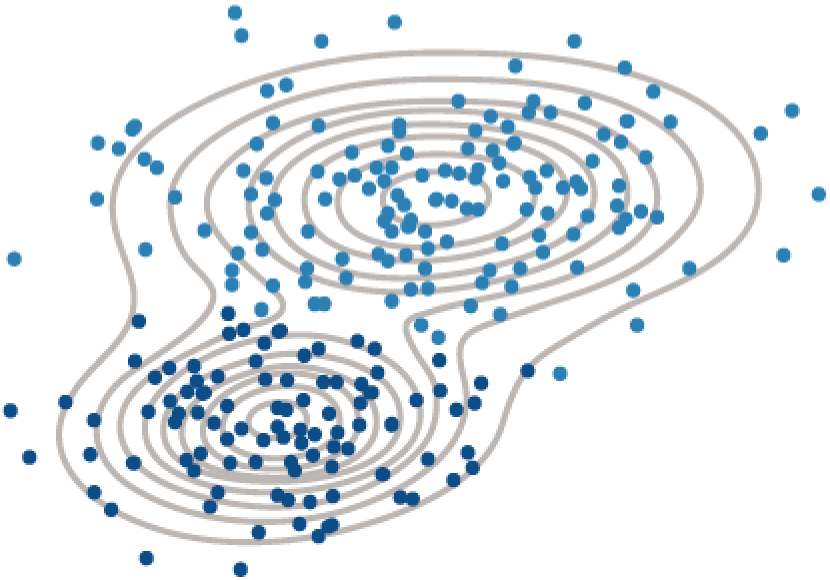
CONTENTS

- Clustering
  - K-means
  - Expectation maximization

- Dimensionality reduction
  - Principal component analysis

CLUSTER ANALYSIS

- Data is partitioned into groups based on some measure of similarity
- The notion of a "cluster" cannot be precisely defined, which is one of the reasons why there are so many clustering algorithms.

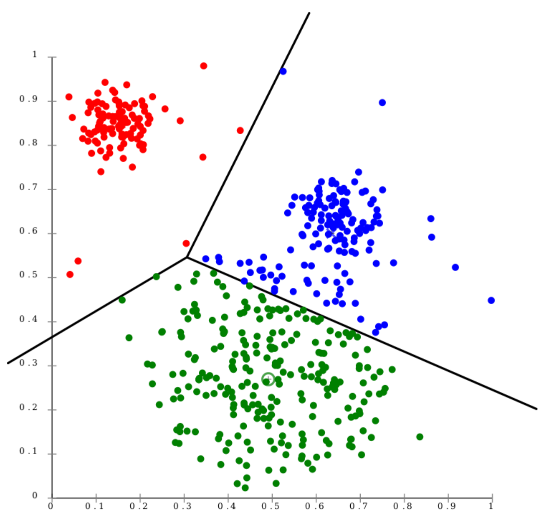


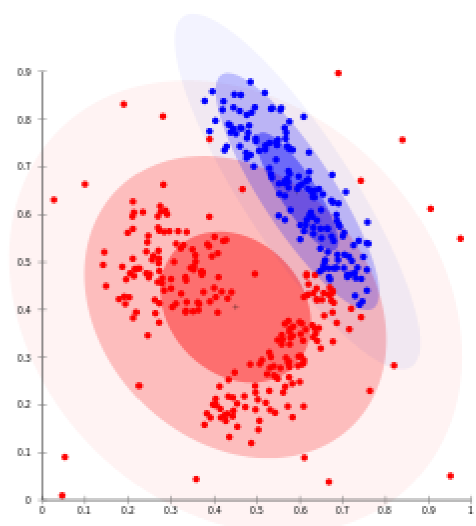
Gaussian mixture model used to separate data into two clusters.

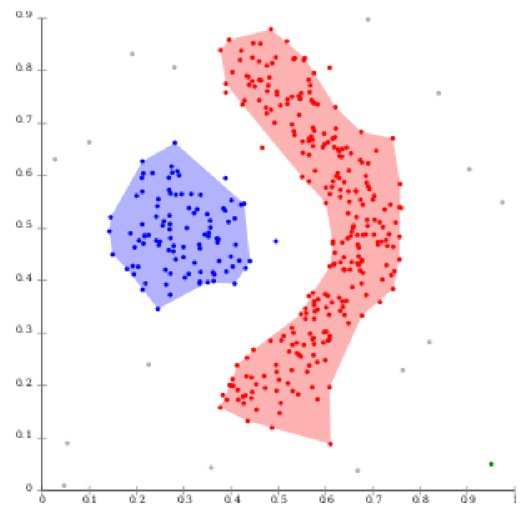
Source: Mathworks, Applying Unsupervised Learning

CLUSTERING ALGORITHMS

- Connectivity
- Centroids
- Distribution
- Density
- …







HARD VS SOFT CLUSTERING

- Hard clustering, where each data point belongs to only one cluster
- Soft clustering, where each data point can belong to more than one cluster

If you don't yet know how the data might be grouped:

- Use self-organizing feature maps or hierarchical clustering to look for possible structures in the data.
- Use cluster evaluation to look for the "best" number of groups for a given clustering algorithm.

K-MEANS

- Partitions data into k number of mutually exclusive clusters.
- How well a point fits into a cluster is determined by the distance from that point to the cluster's center (*inertia)*

Source: Mathworks, Applying Unsupervised Learning

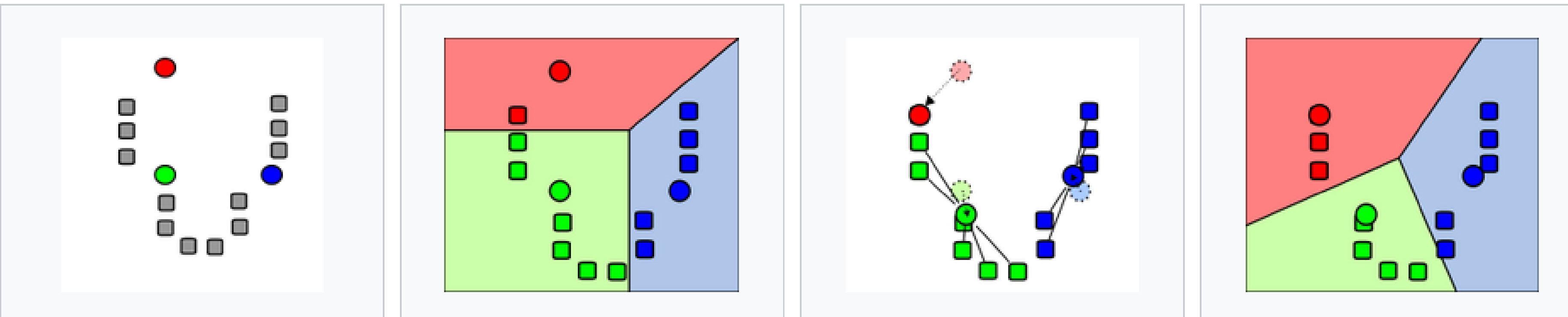
- Best used when the number of clusters is known
- For fast clustering of large data sets

K-MEANS

- https://www.youtube.com/watch?v=_aWzGGNrcic

- Do not mistake K-means, which is an unsupervised machine learning, with K-NN which is supervised machine learning.

K-MEANS ALGORITHM



1. *k* initial "means" (in this case *k*=3) are randomly generated within the data domain (shown in color).

2. *k* clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.

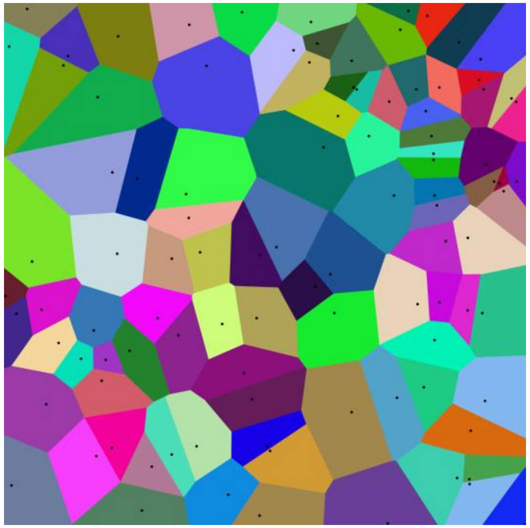
3. The centroid of each of the *k* clusters becomes the new mean.

4. Steps 2 and 3 are repeated until convergence has been reached.

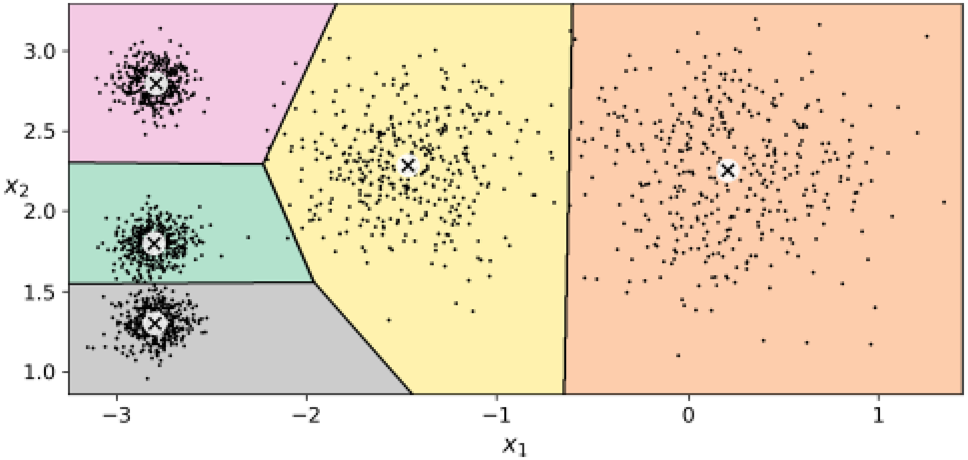
Source: https://en.wikipedia.org/wiki/K-means_clustering

VORONOI TESSELLATION

- Decision boundaries for hard clustering



Source: https://nl.wikipedia.org/wiki/Voronoi-diagram



Source: Géron, ISBN: 9781492032632

IMPROVING THE CENTROID INITIALIZATION

- Problem: Convergence to a local optimum (suboptimal solution)

- Possible solutions:
  1. Supply approximate centroids (initial guesses)

  2. Run the algorithm multiple times with different random initialization and keep the 'best' solution

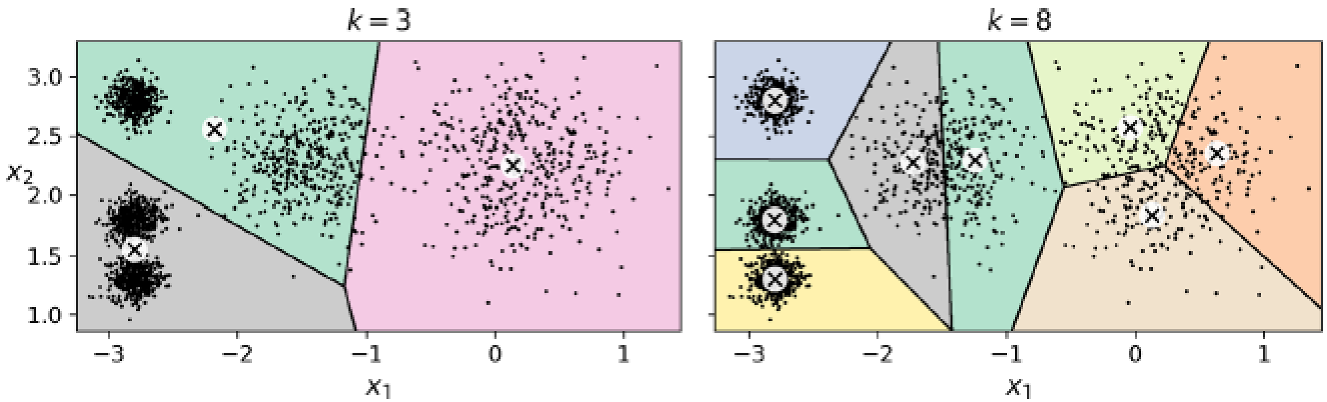
  3. Select centroids within the dataset that are distant from one another *(K-Means++)*

FURTHER IMPROVEMENTS

- Accelerated K-Means
  - avoiding many unnecessary distance calculations by exploiting the triangle inequality

- Mini-batch K-Means
  - Keeping smaller data sets in memory

OPTIMAL NUMBER OF CLUSTERS



Source: Géron, ISBN: 9781492032632

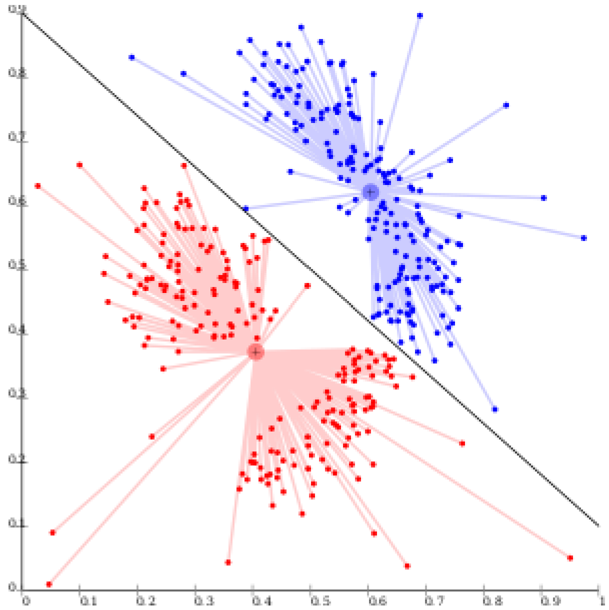
- Inertia is not a proper metric

- Silhouette: measure how similar an object is to its own cluster (cohesion) compared to other clusters (separation)

HAN_UNIVERSITY OF APPLIED SCIENCES

LIMITATIONS OF K-MEANS

- does not behave very well when the clusters have varying sizes.
- cannot represent density-based clusters



Source: ahttps://en.wikipedia.org/wiki/Cluster_analysis#/media/File:KMeans-density-data.svg

HAN_UNIVERSITY OF APPLIED SCIENCES

CLUSTERING APPLICATIONS

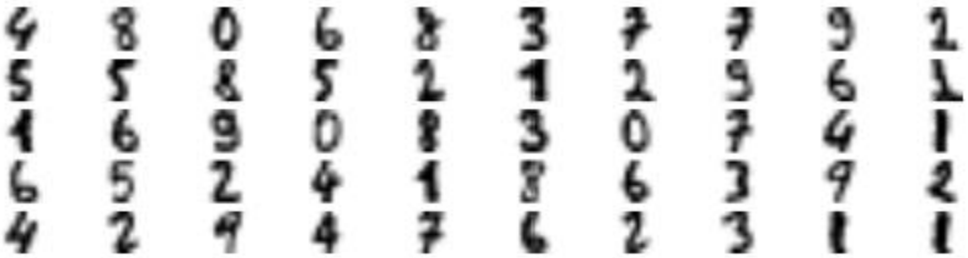
- Image segmentation
- Preprocessing
  - Dimensionality reduction
  - Semi-supervised learning

APPLICATION: COLOR SEGMENTATION

Source: Géron, ISBN: 9781492032632

APPLICATION: SEMI-SUPERVISED LEARNING

- Dimensionality reduction
- Label propagation
- Active learning, e.g. uncertainty sampling



Clustered MNIST representatives

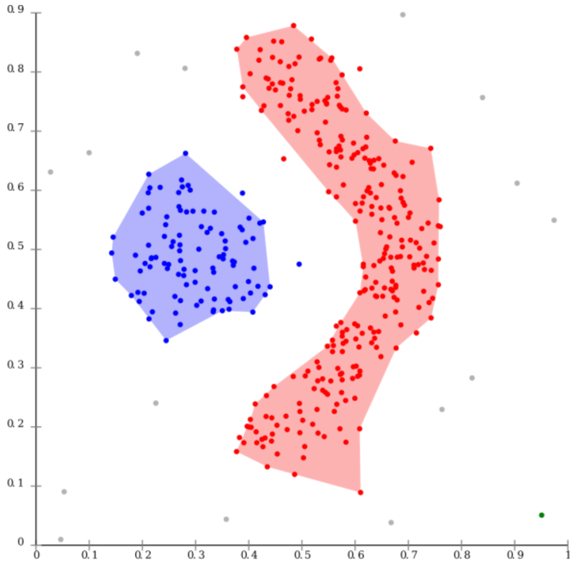
DENSITY-BASED SPATIAL CLUSTERING OF APPLICATIONS WITH NOISE

- Aka DBSCAN
- ε-neighborhood
- core instance

- if all the clusters are dense enough and if well separated

- Finds non-linearly separable clusters on which k-means or Gaussian Mixture EM clustering fails

Special (efficient) variant of spectral clustering: Connected components



Source: https://en.wikipedia.org/wiki/DBSCAN#/media/File:DBSCAN-density-data.svg

HAN_UNIVERSITY OF APPLIED SCIENCES

POPULAR CLUSTERING ALGORITHMS

- Agglomerative clustering
- Mean-Shift
- Affinity propagation
- OPTICS
- Etc.

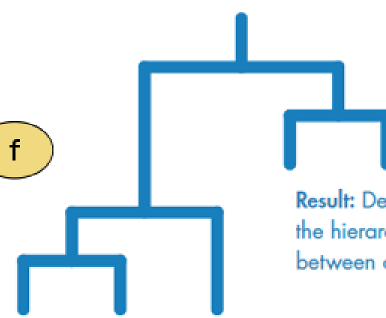
Hierarchical Clustering

How it Works

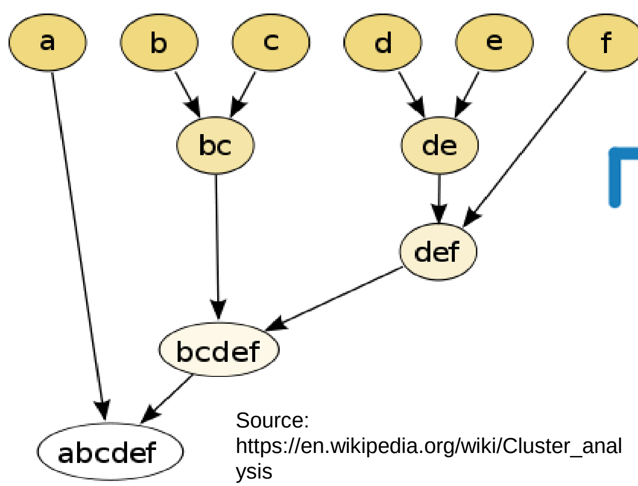
Produces nested sets of clusters by analyzing similarities between pairs of points and grouping objects into a binary, hierarchical tree.

Best Used...

- When you don't know in advance how many clusters are in your data
- You want visualization to guide your selection



Result: Dendrogram showing the hierarchical relationship between clusters



Source: https://en.wikipedia.org/wiki/Cluster_analysis

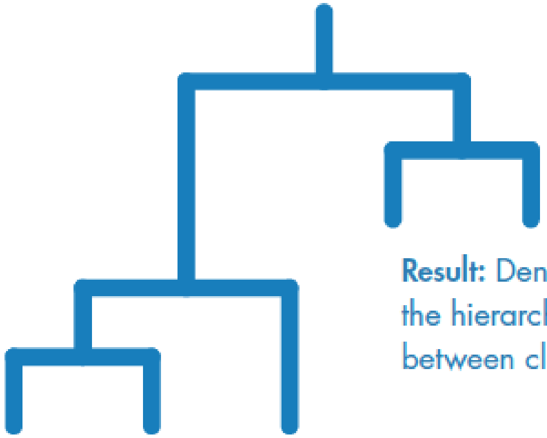
Hierarchical Clustering

How it Works

Produces nested sets of clusters by analyzing similarities between pairs of points and grouping objects into a binary, hierarchical tree.

Best Used...

- When you don't know in advance how many clusters are in your data
- You want visualization to guide your selection



Result: Dendrogram showing the hierarchical relationship between clusters

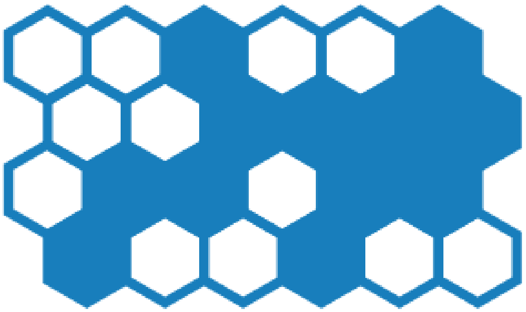
Self-Organizing Map

How It Works

Neural-network based clustering that transforms a dataset into a topology-preserving 2D map.

Best Used...

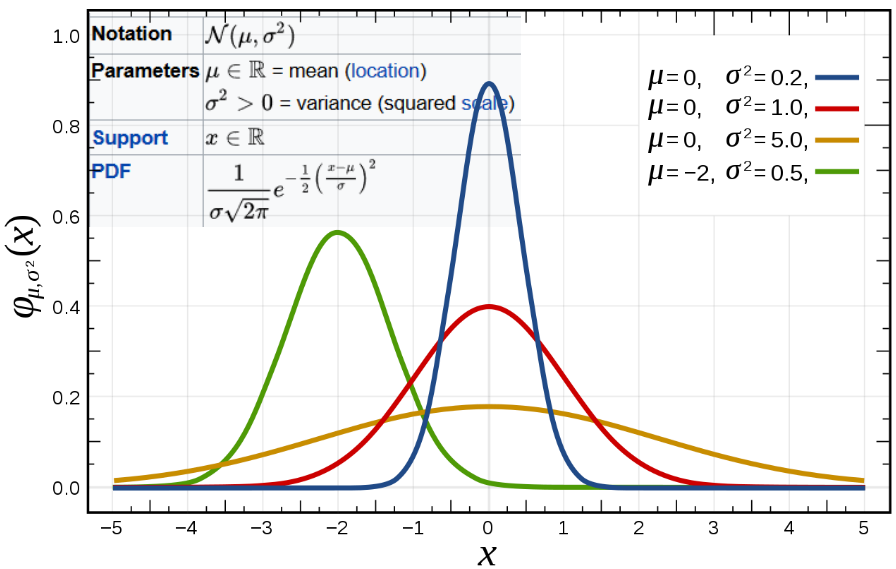
- To visualize high-dimensional data in 2D or 3D
- To deduce the dimensionality of data by preserving topology (shape)



Result: Lower-dimensio (typically 2D) representation

GAUSSIAN MIXTURE MODELS (GMM)

- Probabilistic model, assuming normally distributed subpopulations



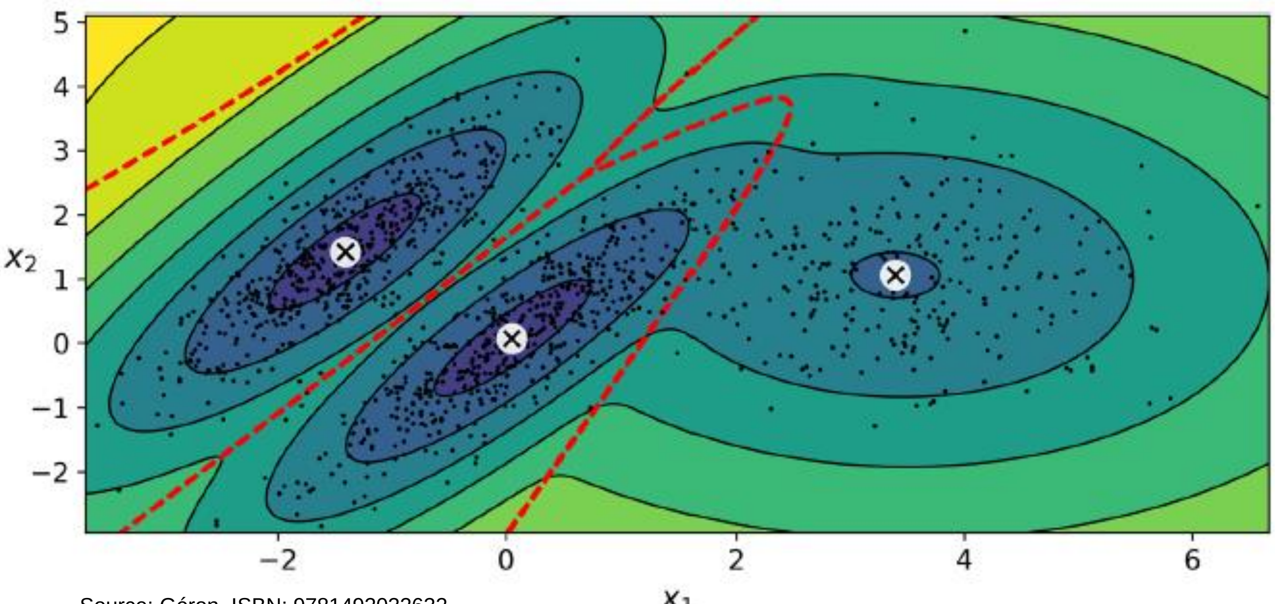
Source: https://en.wikipedia.org/wiki/Normal_distribution

Source: Mathworks, Applying Unsupervised Learning

HAN_UNIVERSITY OF APPLIED SCIENCES

EXPECTATION MAXIMIZATION (EM)

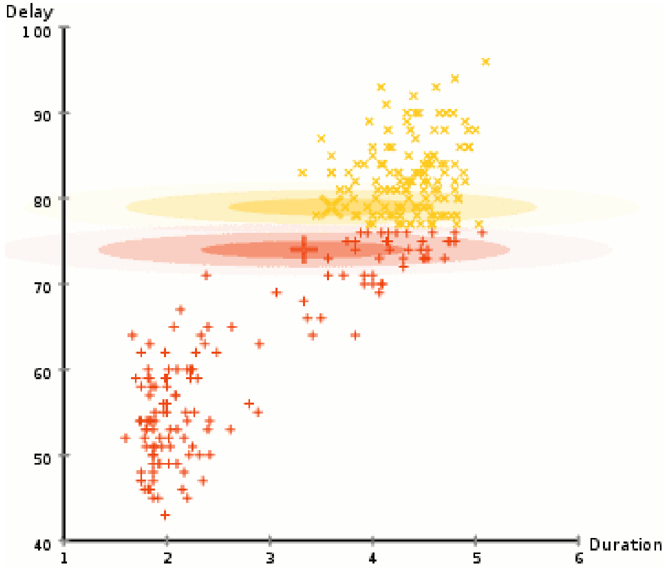
- *a priori* given number of components
- Generalization of K-Means
- Soft cluster assignments



Source: Géron, ISBN: 9781492032632

EXAMPLE: EM CLUSTERING OF GEYSER ERUPTION DATA

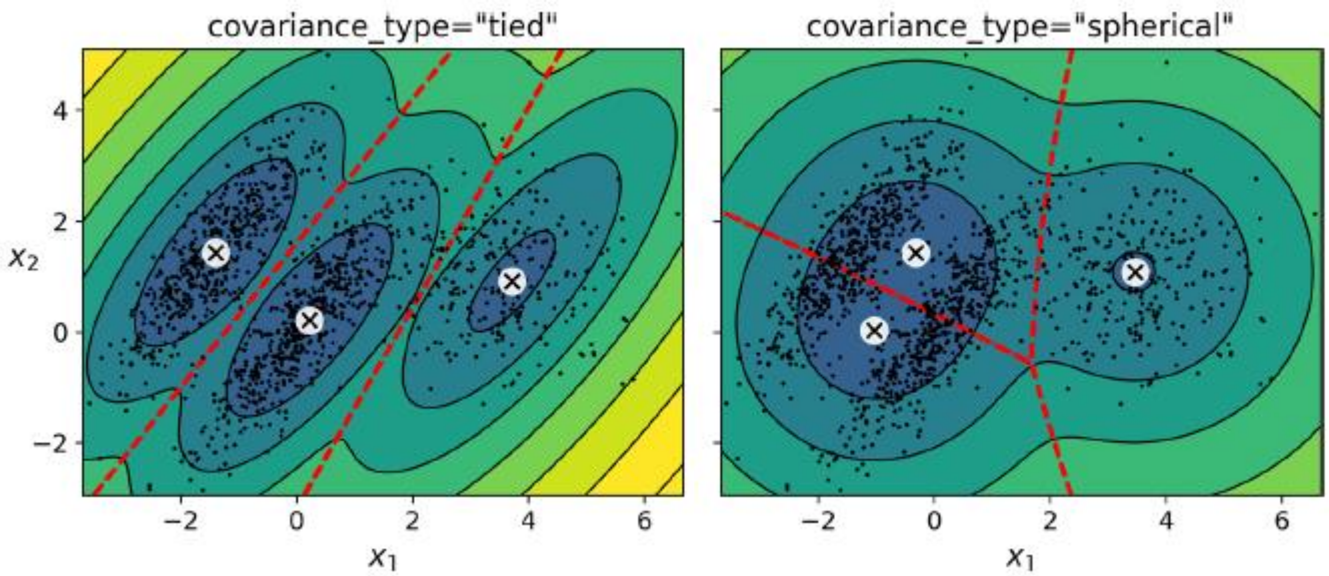
- Highly predictable geothermal feature
- Erupted every 44 minutes to two hours since 2000



Source: https://en.wikipedia.org/wiki/Expectation%E2%80%93maximization_algorithm

APPLYING CONSTRAINTS

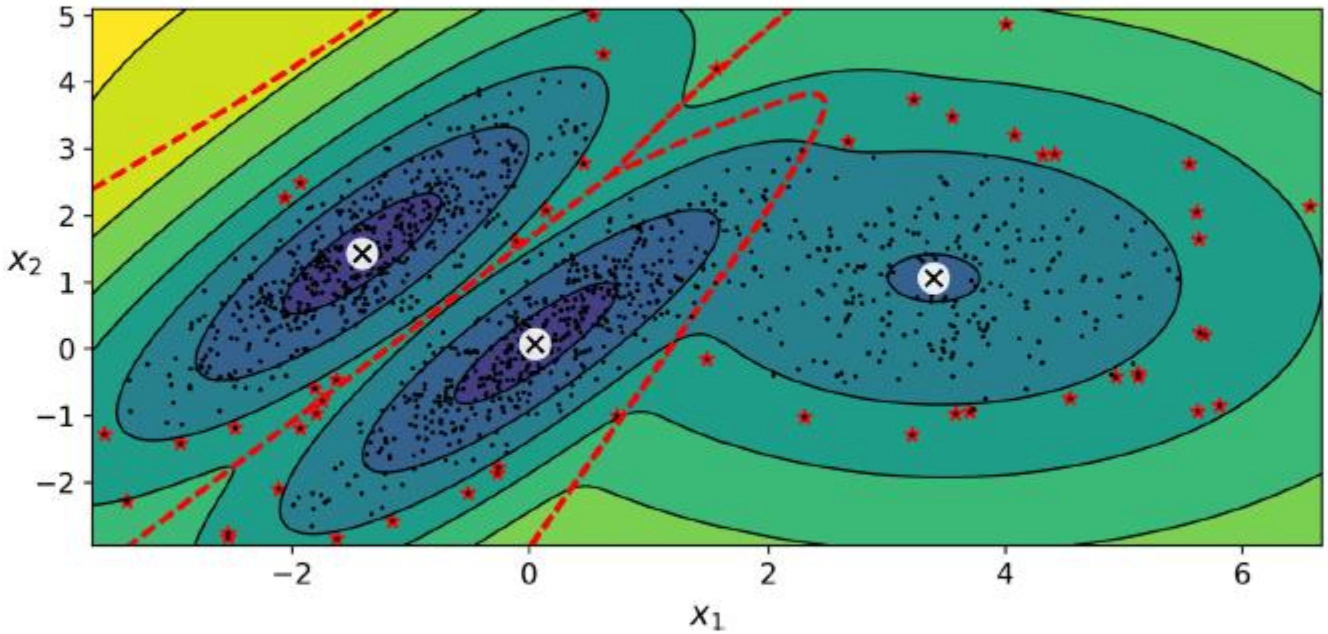
- When there are many dimensions, or many clusters, or few instances, EM can struggle to converge to the optimal solution.
- You might need to reduce the difficulty of the task by limiting the number of parameters that the algorithm has to learn.



Source: Géron, ISBN: 9781492032632

APPLICATION: ANOMALY DETECTION

- Aka outlier detection



- -> Dbscan (Density Based Spatial Clustering of Applications with Noise)

SELECTING THE NUMBER OF CLUSTERS

- Inertia and silhouette are not proper metrics

- Model selection criteria from statistics, e.g.
  - Bayesian information criterion (BIC)
  - Akaike information criterion (AIC)

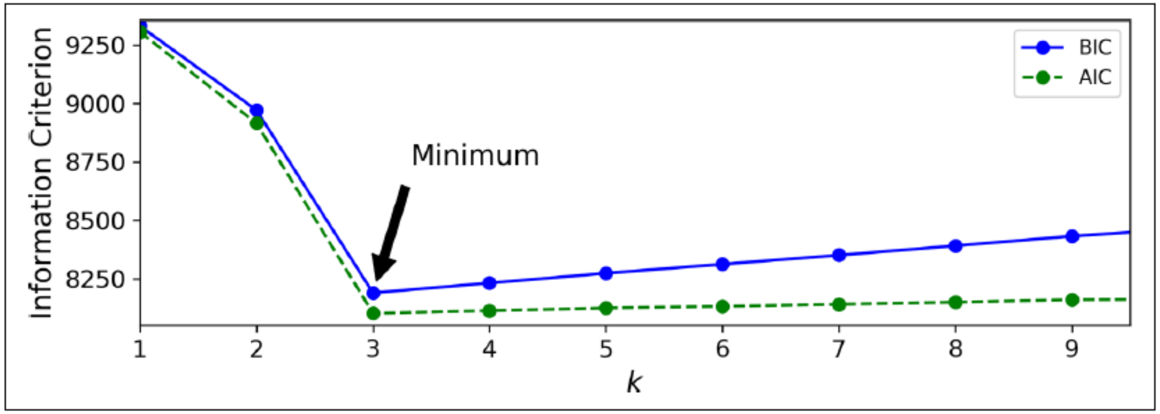
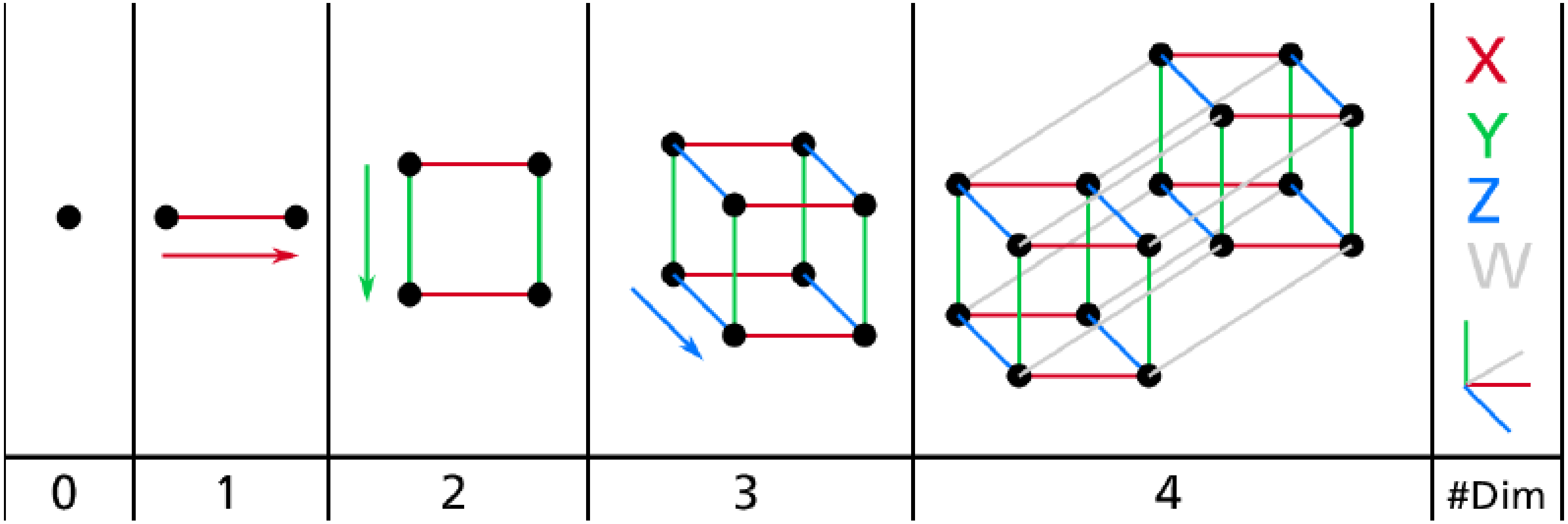


Figure 9-21. AIC and BIC for different numbers of clusters k

DIMENSIONALITY REDUCTION

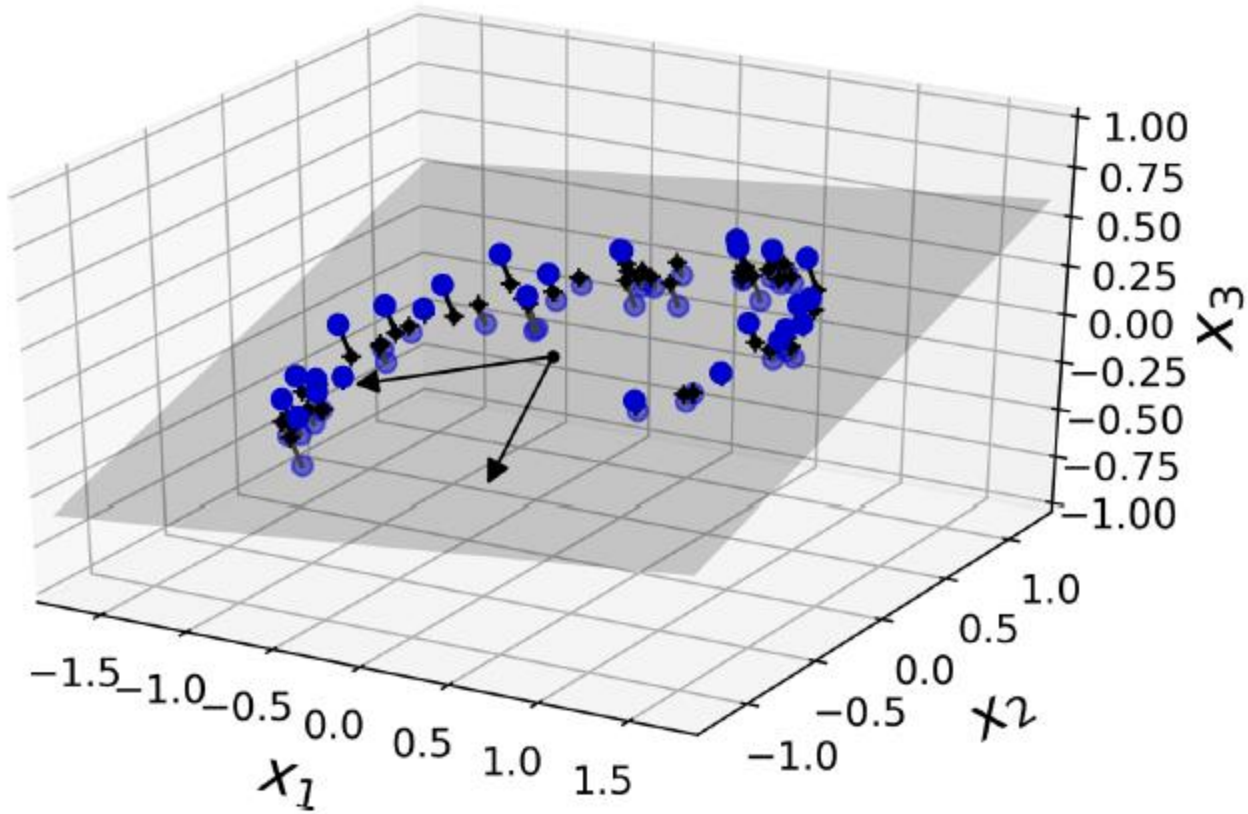
- curse of dimensionality



Source: Géron, ISBN: 9781492032632

- high-dimensional datasets are at risk of being very sparse

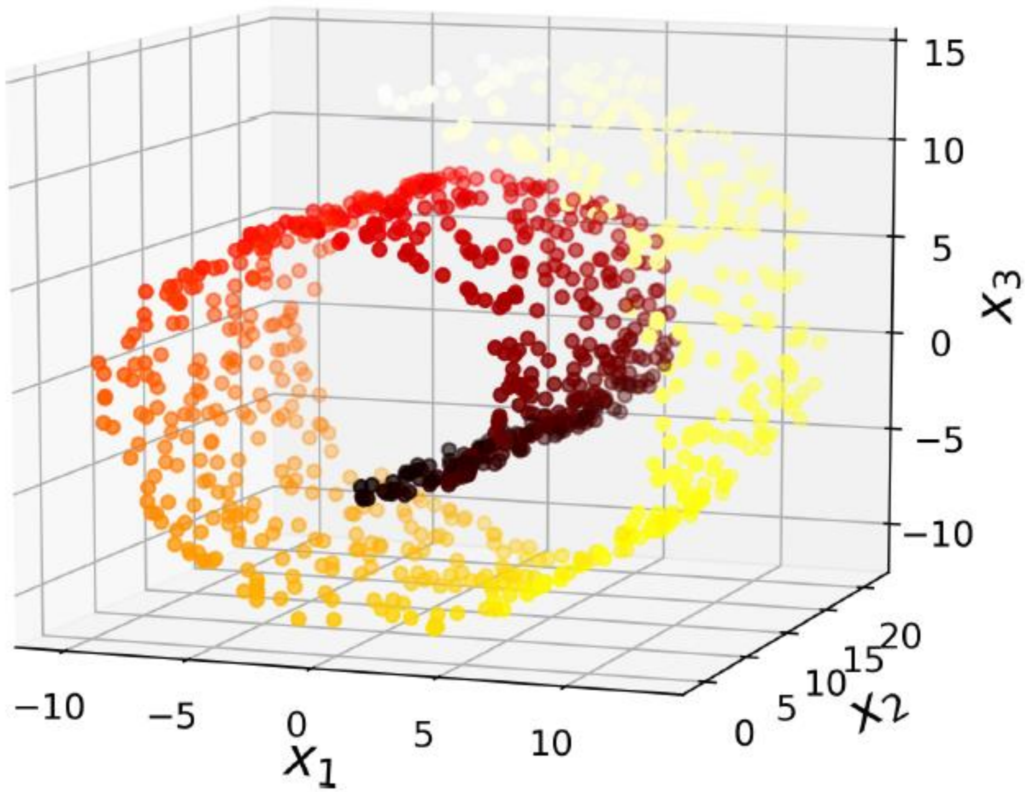
PROJECTION



1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
x_3
1.0
0.5
0.0
−0.5
−1.0
x_2
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
x_1

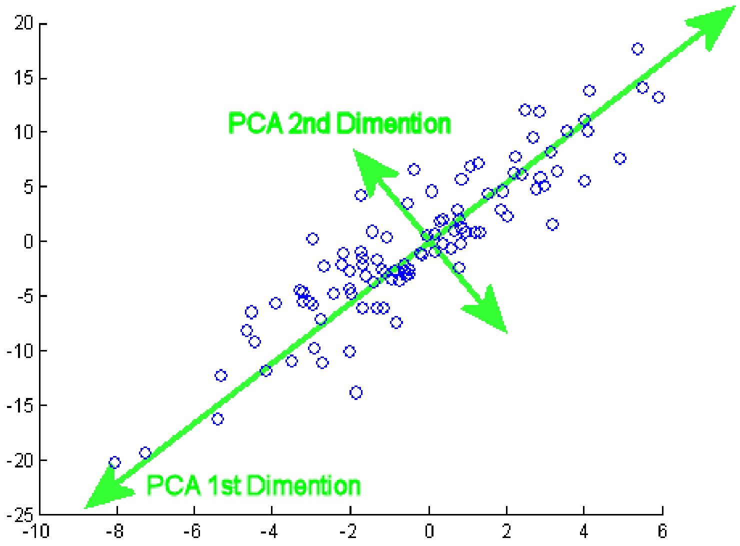
MANIFOLD LEARNING

- Assumption: most real-world high-dimensional datasets lie close to a much lower-dimensional manifold.



PRINCIPAL COMPONENT ANALYSIS (PCA)

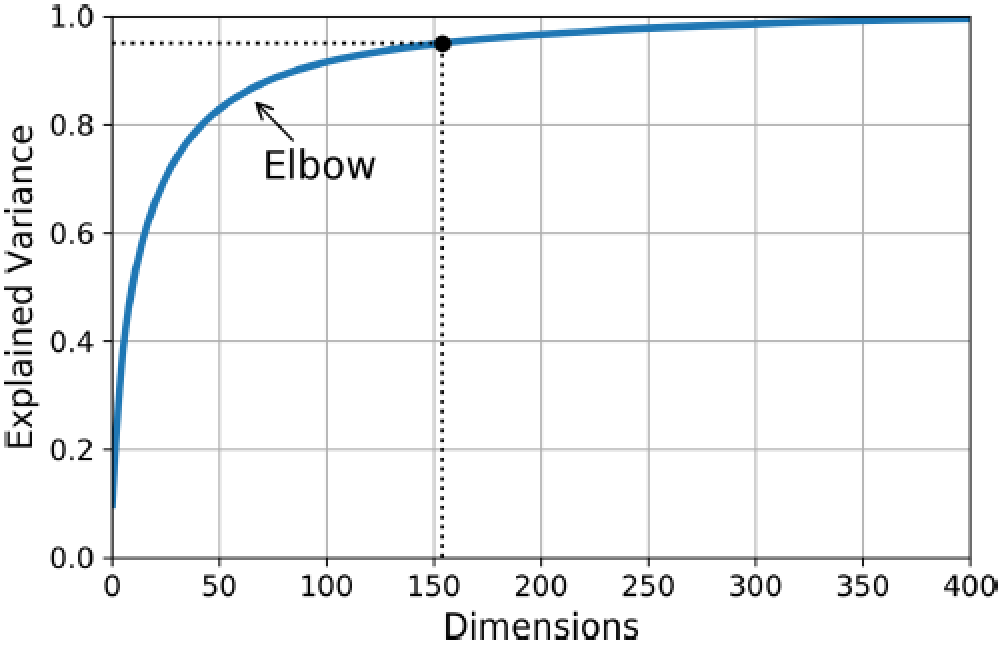
- Choose the right hyperplane
- Most of the information in high-dimensional dataset is captured by the first few principal components (PCs)
- Maximum variance
- PCA finds a zero-centered unit vector pointing in the direction of the PC



HAN_UNIVERSITY OF APPLIED SCIENCES

CHOOSING THE NUMBER OF DIMENSIONS

- Explain sufficiently large portion of the variance (e.g., 95%).
- Or for visualization, reduce to 2 or 3
- Elbow point



HAN_UNIVERSITY OF APPLIED SCIENCES

EXAMPLE: MNIST DATASET COMPRESSION

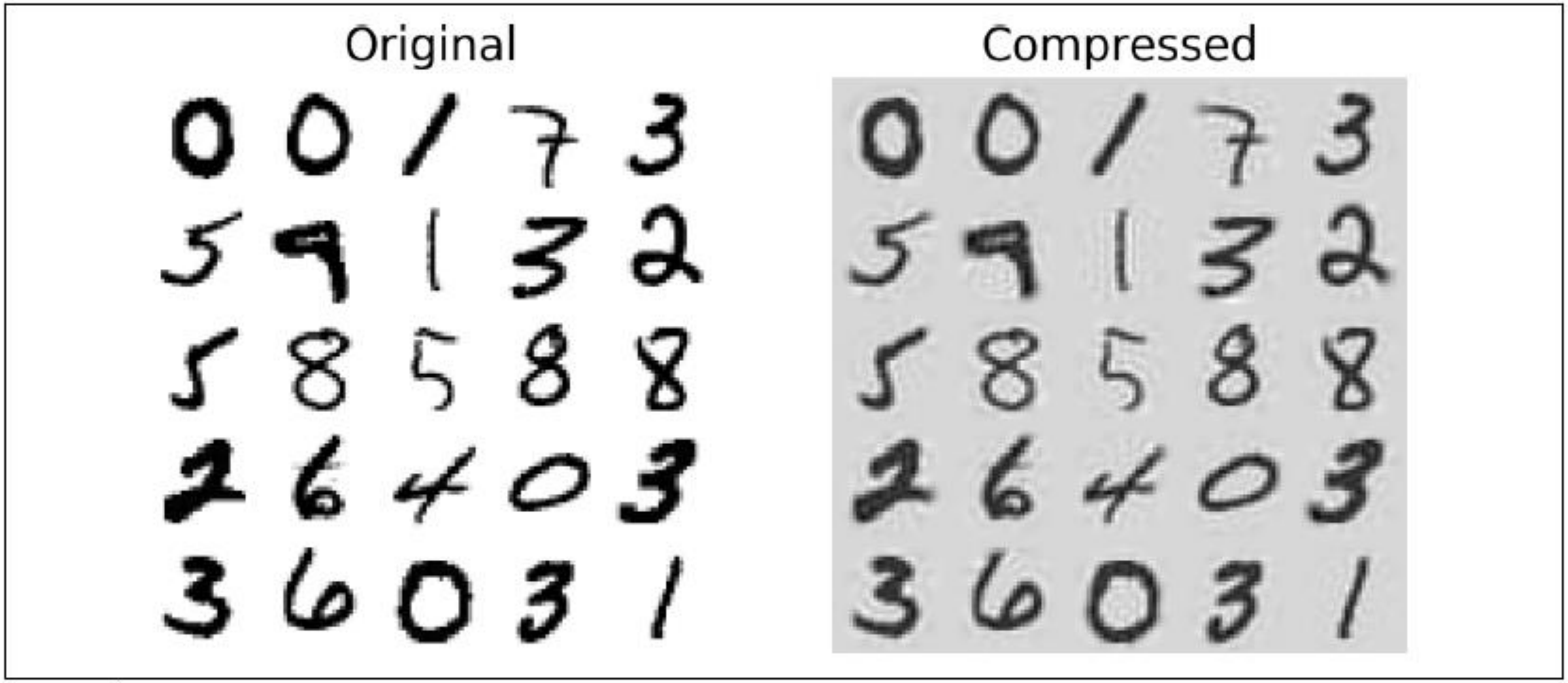


Figure 8-9. MNIST compression that preserves 95% of the variance