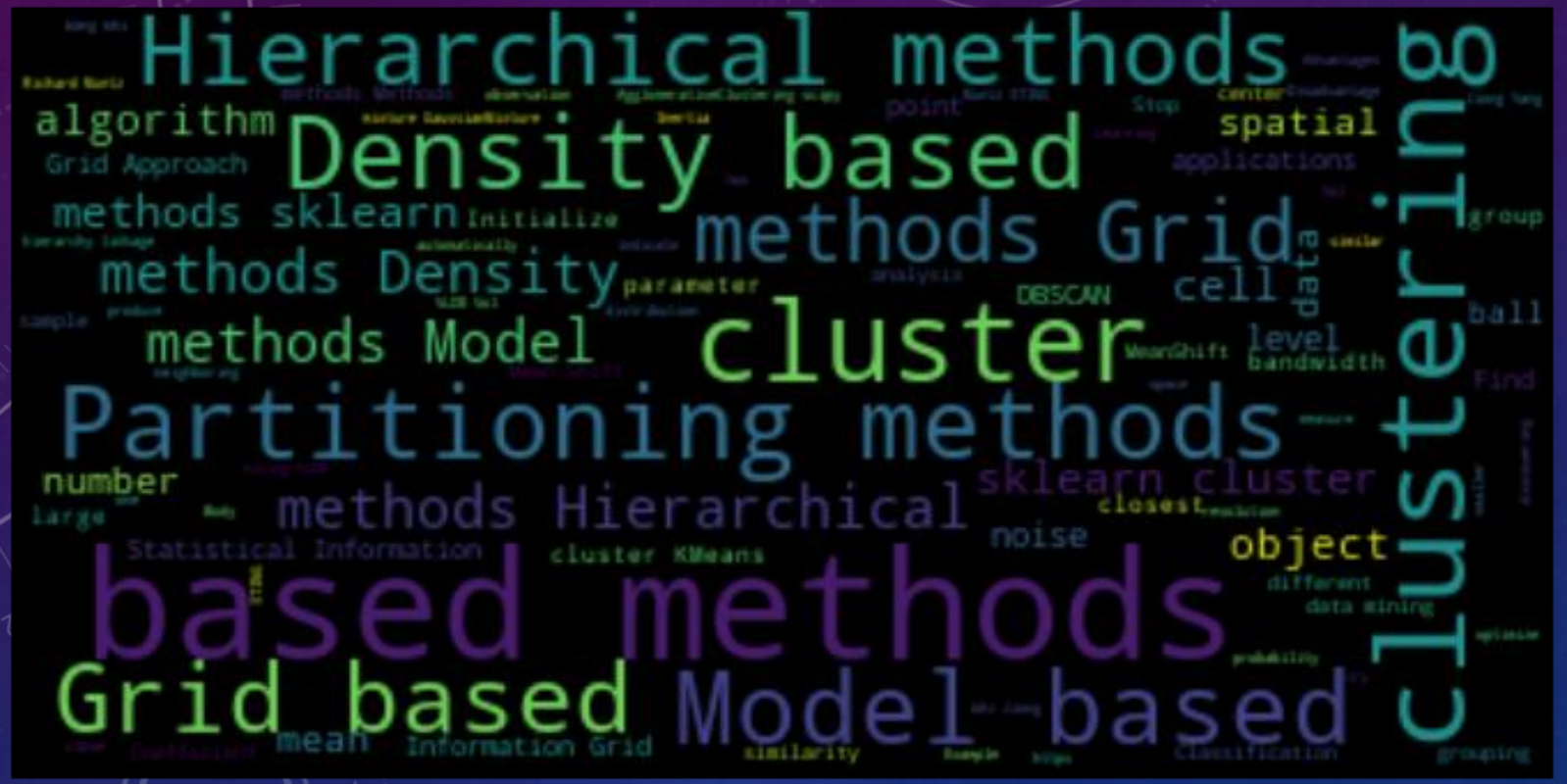
DATA SCIENCE



Prof. Lev Muchnik

2025

Jerusalem School of Business Administration

AGENDA

- **Unsupervised Learning**
- **Clustering Models**
  - Partitioning methods
  - Hierarchical methods
  - Density-based methods
  - Grid-based methods
  - Model-based methods
- The curse of dimensionality
- Dimensionality Reduction for visualization:
  - Principal Components
  - TSNE (t-distributed Stochastic Neighbor Embedding)
  - U-map
  - pacmap

Clustering_and_dimensionality_reduction_Part_A.ipynb
Clustering_and_dimensionality_reduction_Part_B.ipynb
Online Sales.ipynb
Chicago Crime Clusters.ipynb
PedestrianSafetyNYC.ipynb
Clustering_and_dimensionality_reduction_Part_C.ipynb

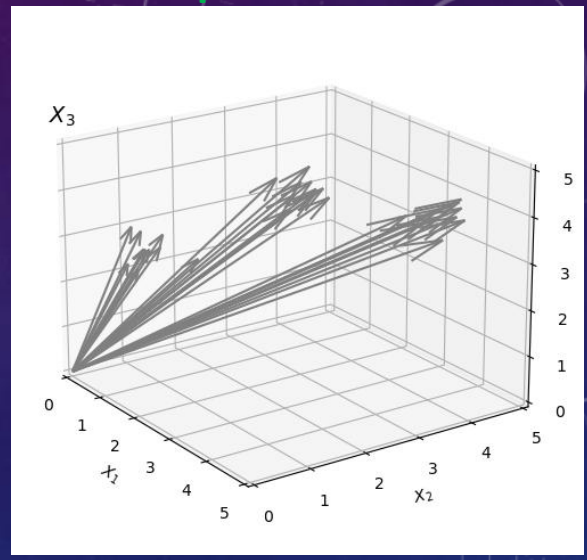
Home Assignment

Feature Vector / Feature Space representation

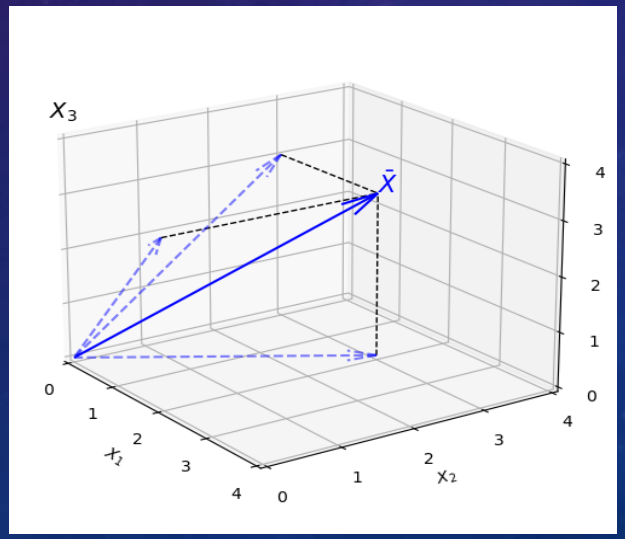
- Each observation is typically represented by a **feature vector**

$$\bar{X}_i = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_d \end{bmatrix}$$

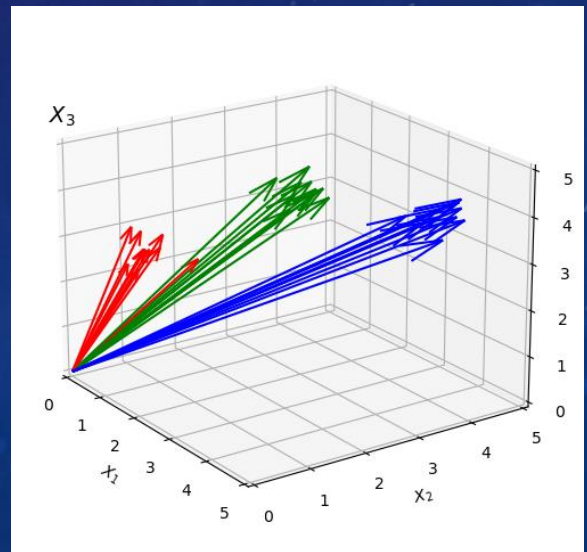
- Feature Vectors form **Feature Space**



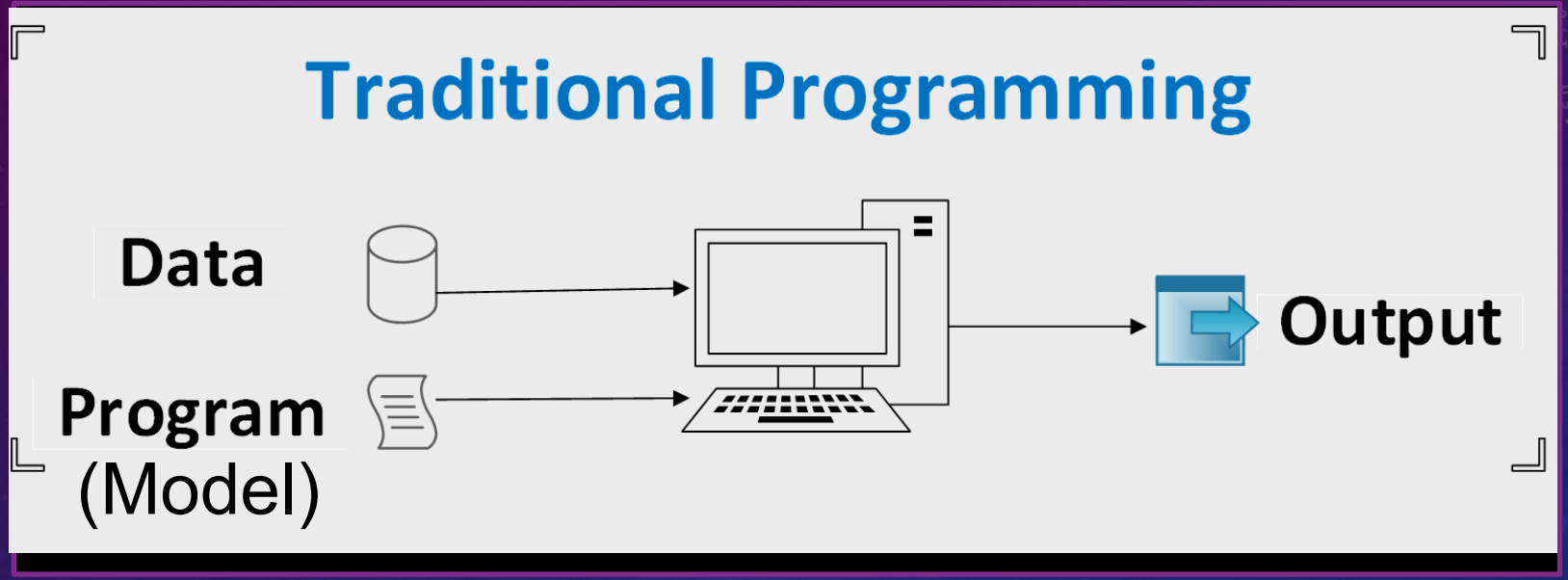
- Each **feature vector** $\bar{X}_i$ can be placed (embedded) in a multi-dimensional space



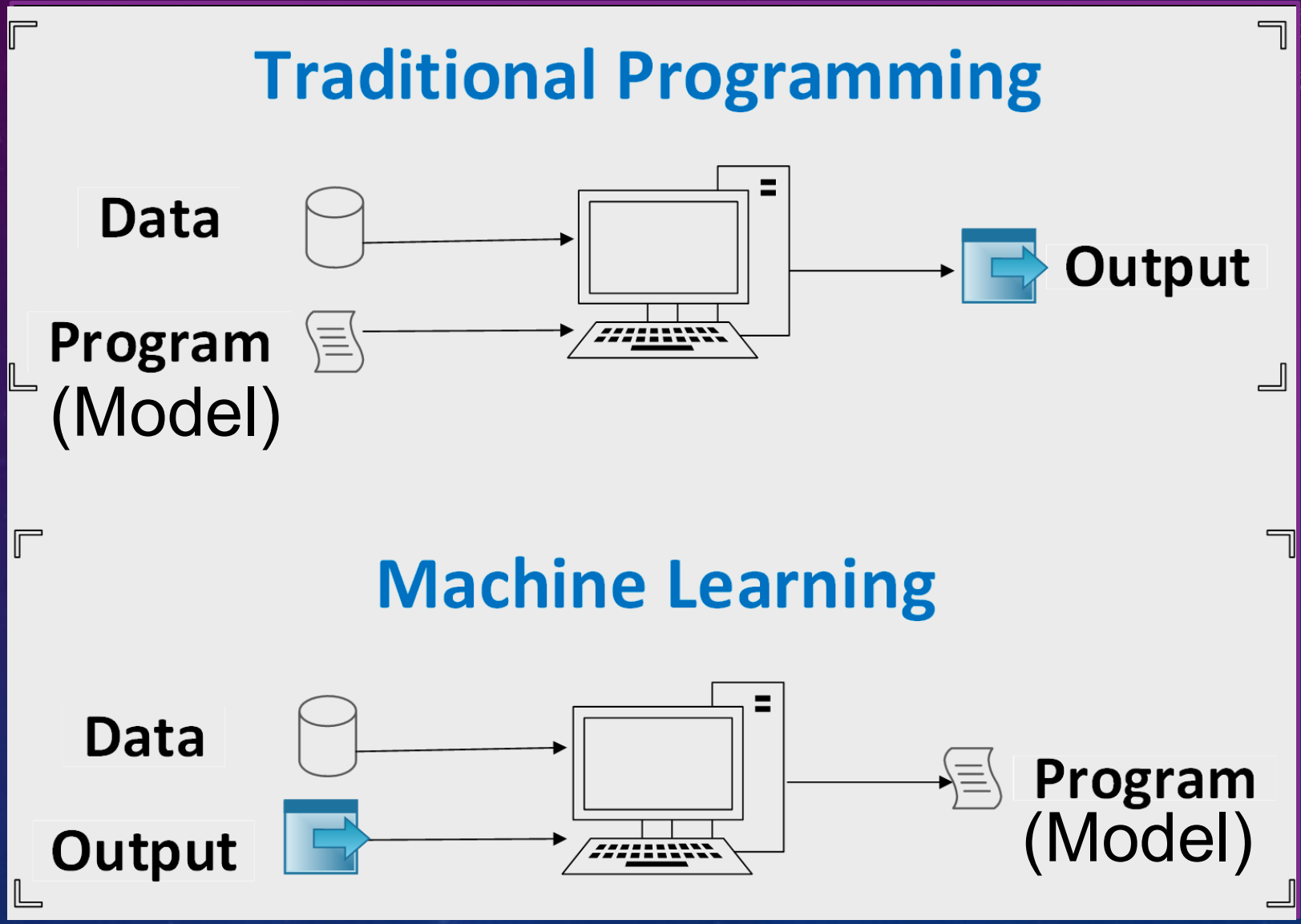
- Machine Learning Models are trained to characterize feature vectors



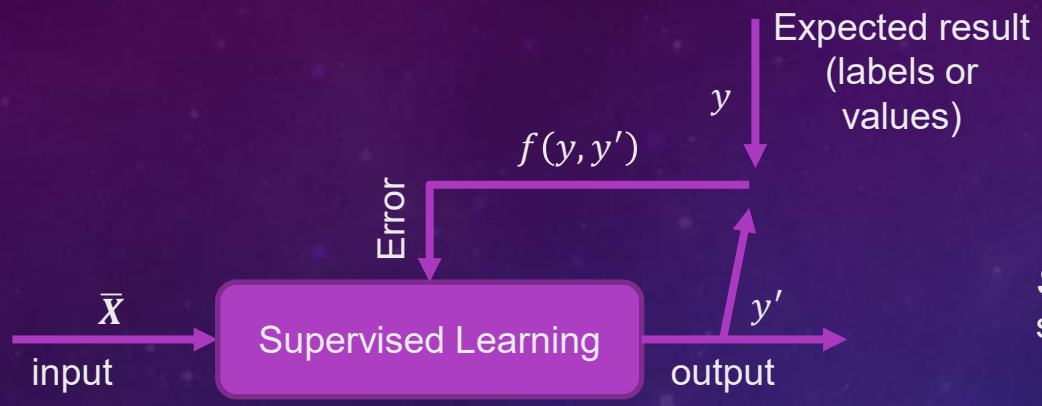
ML: Conceptual Shift In Problem solving



ML: Conceptual Shift In Problem solving



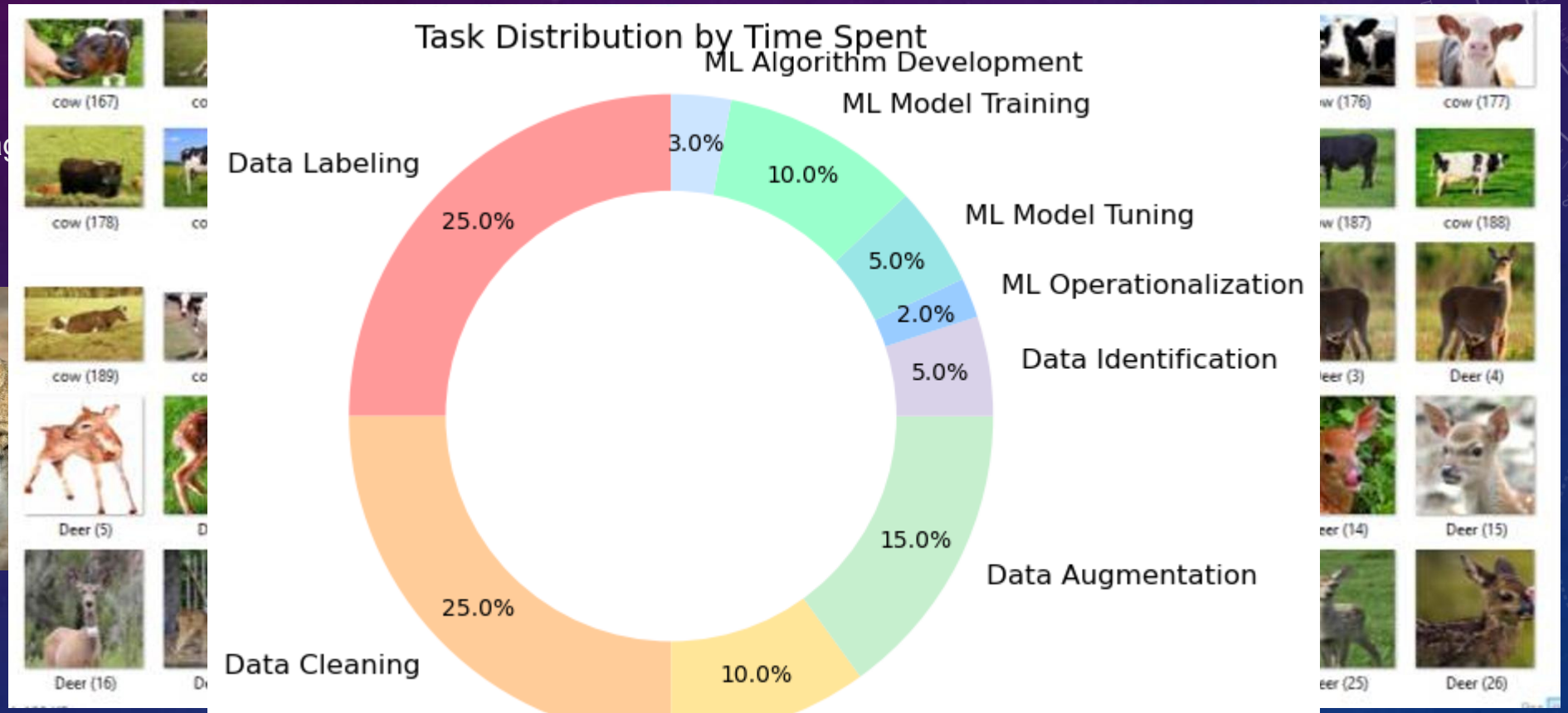
Methods for Training Models:



Supervised learning – the system is given an example to follow at each step

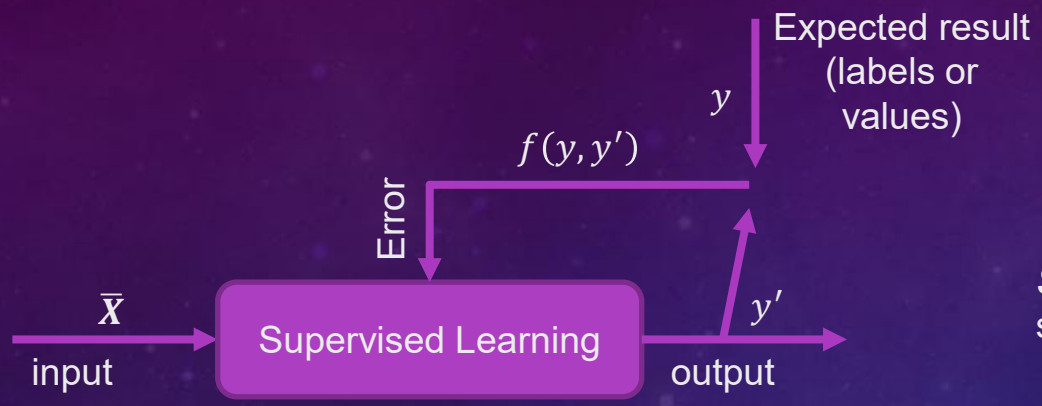
Methods for Training Models: Supervised

Imag

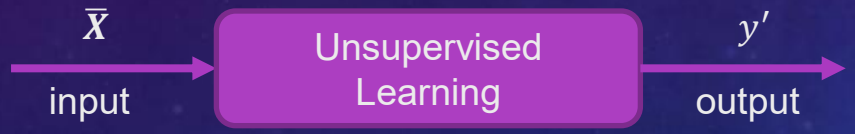


Source: Cognylitica; Factordaily, 2024

Methods for Training Models:



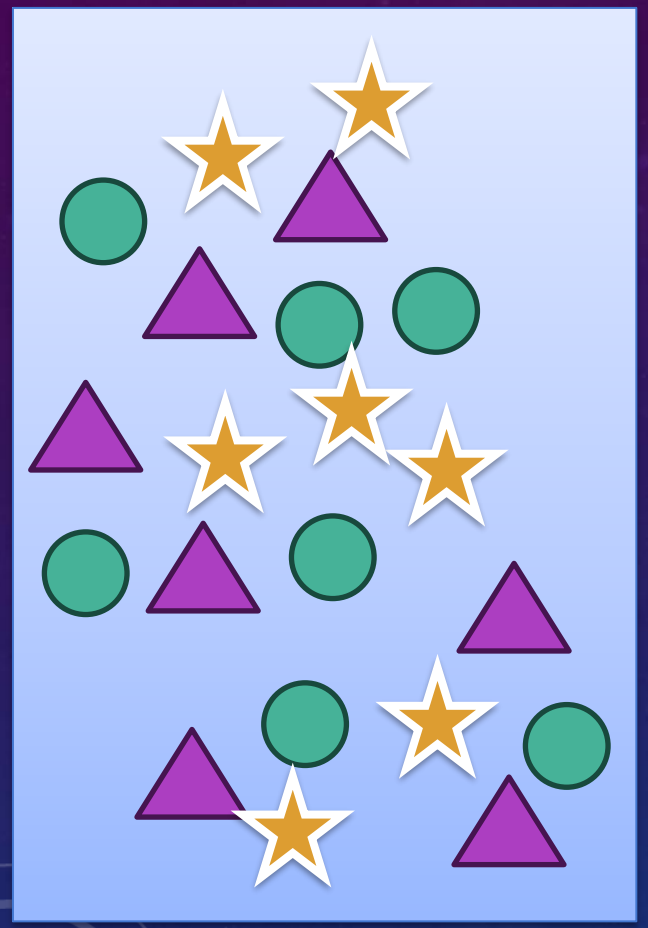
Supervised learning – the system is given an example to follow at each step



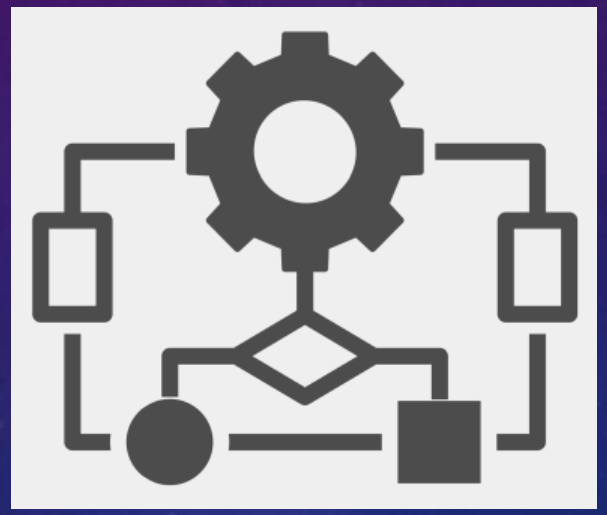
Unsupervised learning – the system infers knowledge from patterns in the data

Unsupervised learning

Unlabeled dataset



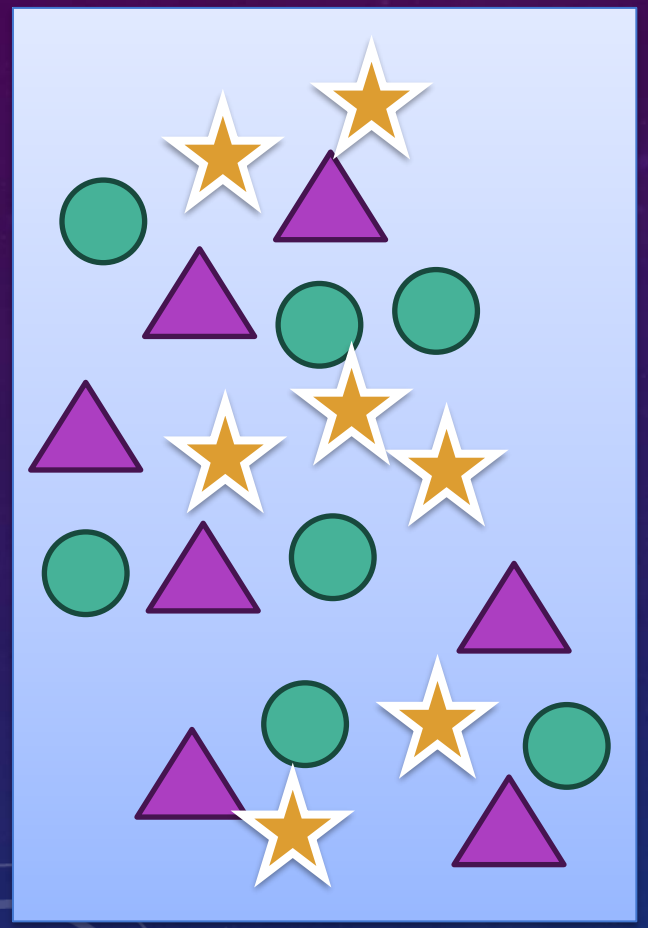
Clustering Algorithm



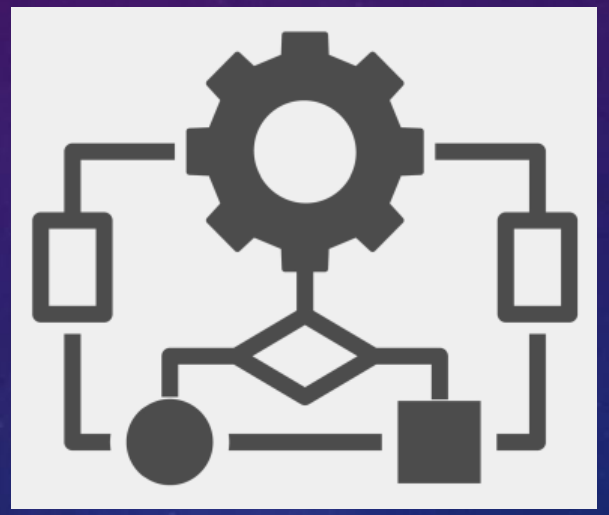
Algorithm Output

Unsupervised learning

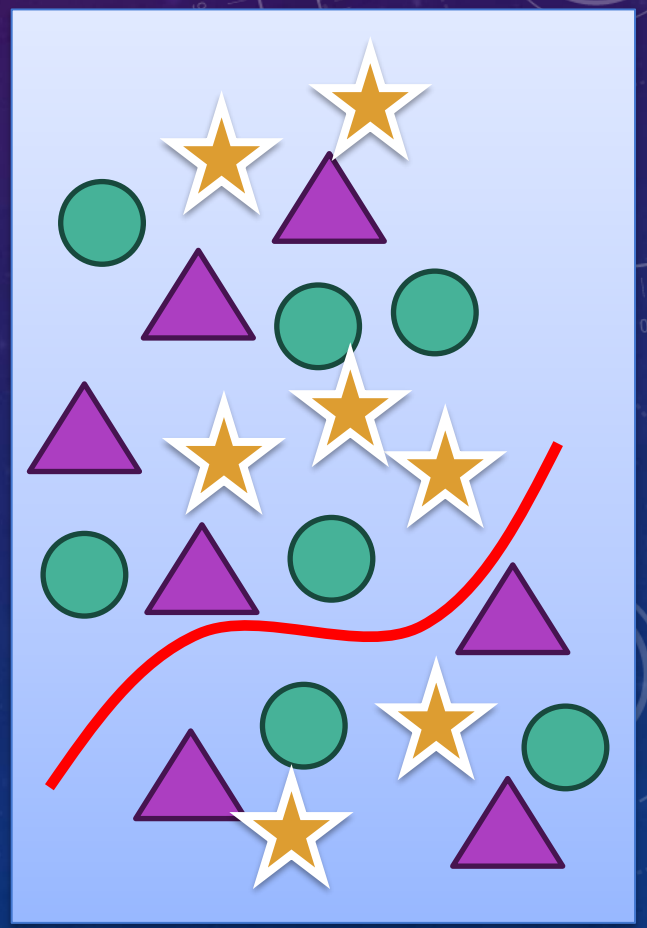
Unlabeled dataset



Clustering Algorithm

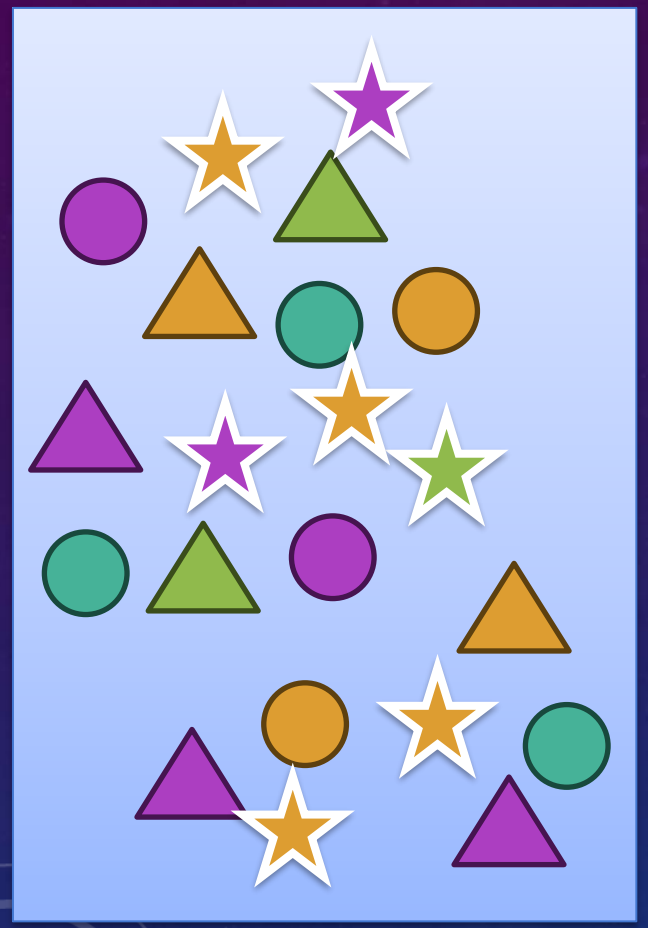


Algorithm Output

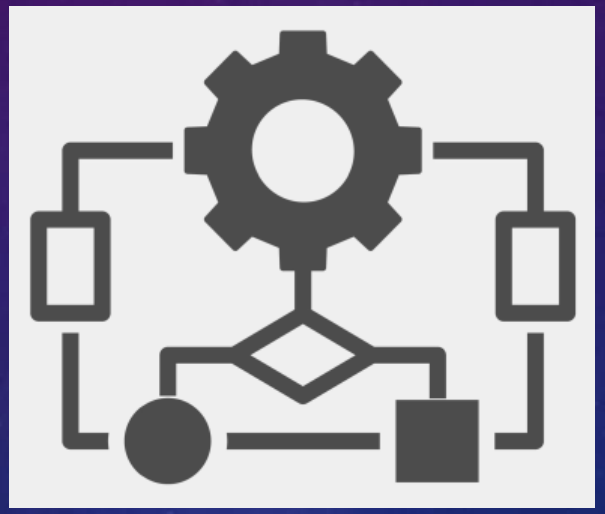


Unsupervised learning

Unlabeled dataset

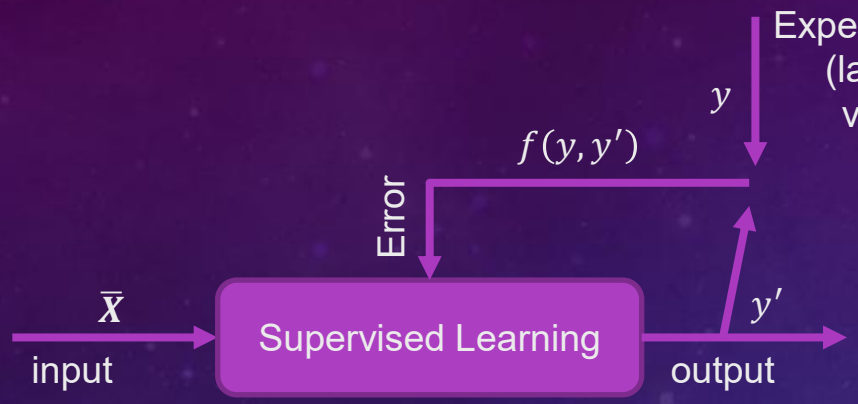


Clustering Algorithm

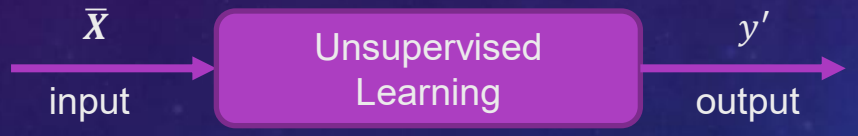


Algorithm Output

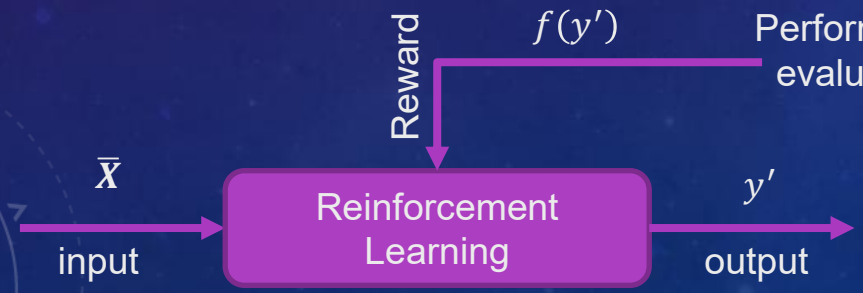
METHODS FOR TRAINING MODELS:



Supervised learning – the system is given an example to follow at each step

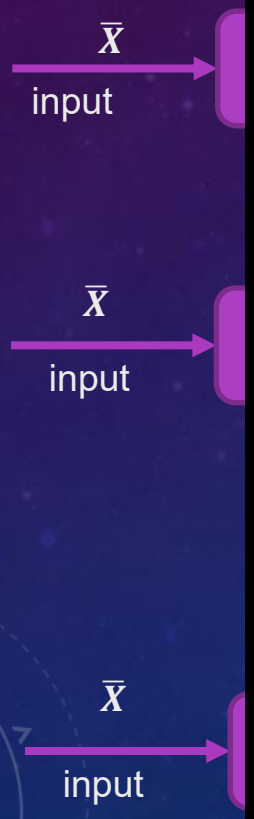


Unsupervised learning – the system infers knowledge from patterns in the data



Reinforcement learning – the algorithm is rewarded for successful steps and penalized for unsuccessful ones.

Methods for Training Models:



example to follow at each

owledge from patterns in

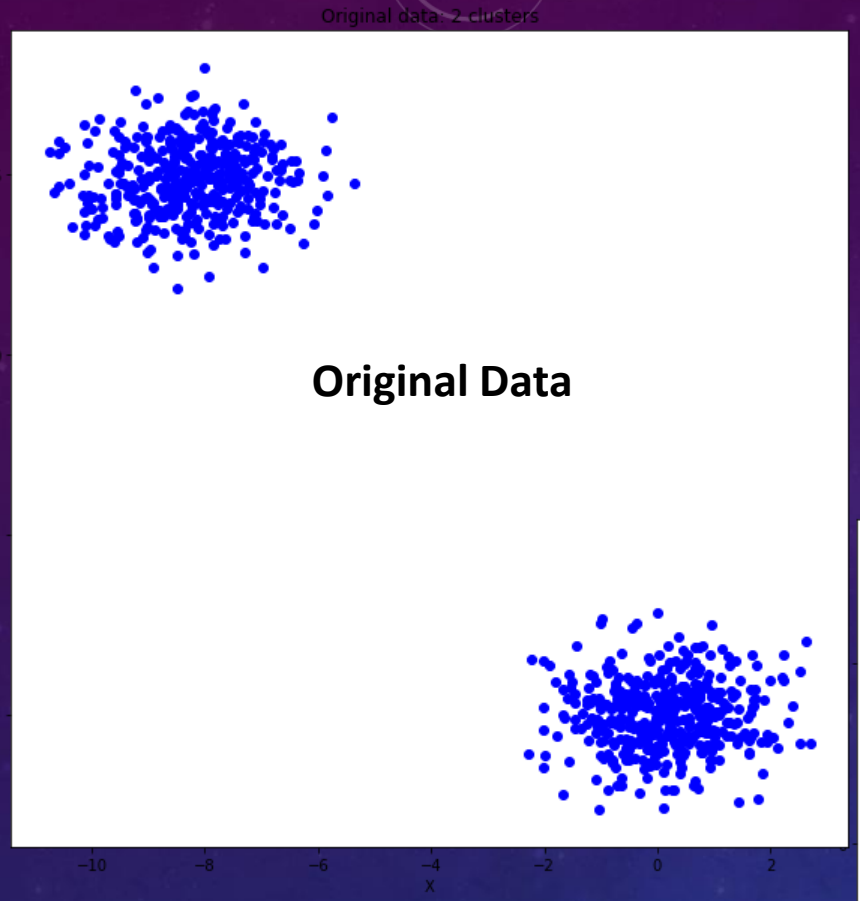
warded for successful steps

MACHINE LEARNING ALGORITHMS

Clustering – grouping of data based on some measure of similarity

- Clustering is related to **classification**, which is an analogous supervised task.
- The objective of classification is to determine to which class/category each observation belongs.
- Classification is performed by algorithm called classifier (e.g. Logit is a classifier)

Self-supervised – hybrid between supervised and unsupervised
Semi-supervised - small amount of labeled data with a lot of unlabeled.
Active learning - select the most informative data points for labeling
Online vs. Offline learning – whether the model can learn incrementally



Group points by similarity

CLUSTERING

$$\bar{X}_i = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_d \end{bmatrix}$$

- What?

  Clustering (or cluster analysis) is a task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups.

- Why?

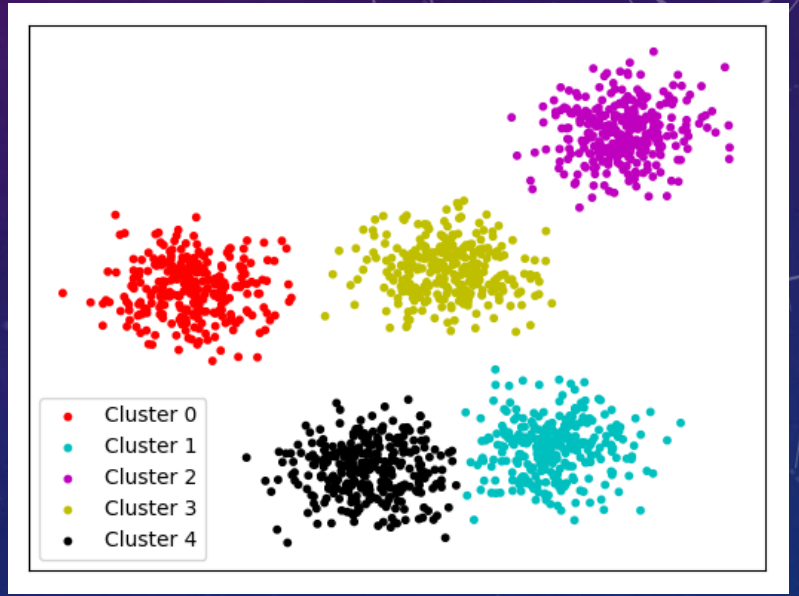
  Used in exploratory data mining to:

  produce customer segmentation

  group products (e.g. movies) for recommendation

  detect outliers (e.g. fraud detection)

  group e-mails or search results



- How?

  Two types of criteria:

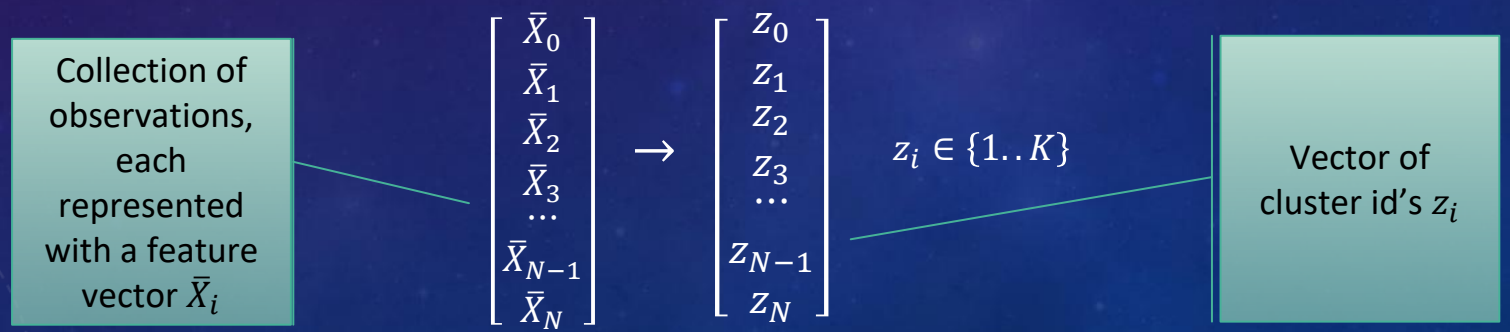
  - Intra-class similarity: objects in the same cluster are similar to objects in the same cluster
  - Inter-class similarity: objects in the same cluster are dissimilar to objects in other clusters

  - *Critical to understand feature engineering and distance metric!*

When? (for example)
- When you don't know what you are looking for
- When it's too expensive to label data
- When you need to systematically explore the data
- Before manually labeling data
- Risk: can produce meaningless results

CLUSTERING

- What do we mean by clustering in practice?
- Observe characteristics of some objects $\bar{X}_i = \{x_1, x_2, \ldots x_N\}$
- Clustering labels objects with K distinct labels $(1..K)$: $\{z_1, z_2, \ldots z_N\}$
  - z_k is a label from 1 to K, suggesting the cluster.
  - If i and j are in the same cluster, then $z_i = z_j$



METHODS OF CLUSTERING

- Partitioning methods
- Hierarchical methods
- Density-based methods
- Grid-based methods
- Model-based methods

METHODS OF CLUSTERING: PARTITIONING METHODS (ALSO CALLED: CENTROID-BASED)

- **Partitioning methods**
- Hierarchical methods
- Density-based methods
- Grid-based methods
- Model-based methods

K-means clustering

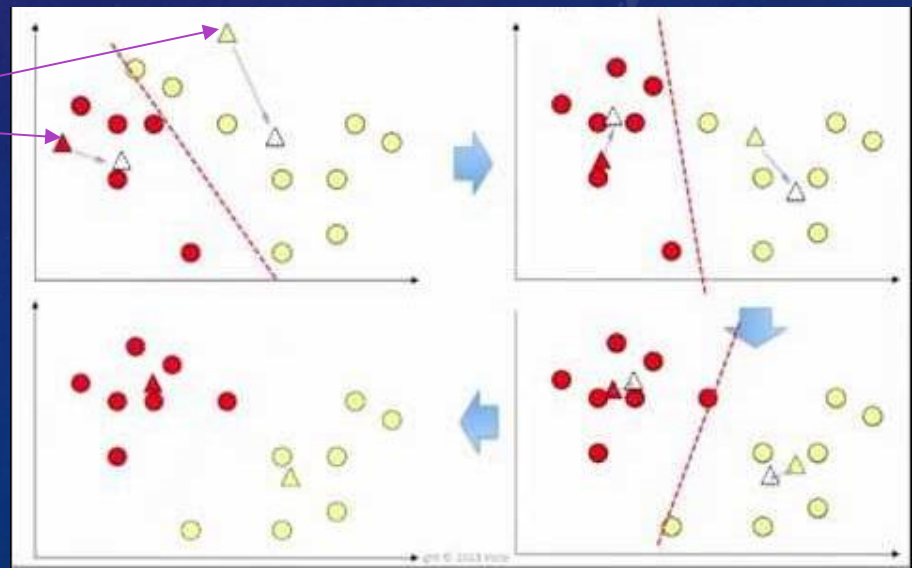
Initialize: Pick K points as cluster centers

Body:

1. Assign each data point to the closest cluster center
2. Recompute cluster center of each cluster

Stop: continue until point assignment no longer change

initialize



MacQueen, James. "Some methods for classification and analysis of multivariate observations." *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. Vol. 1. No. 14. 1967.

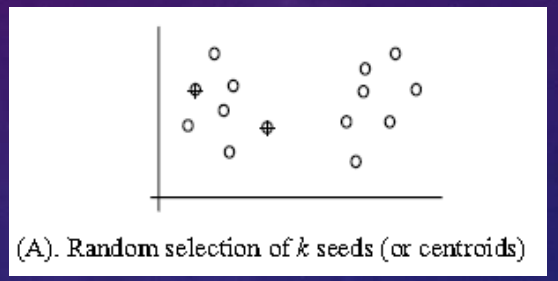
sklearn.cluster.KMeans

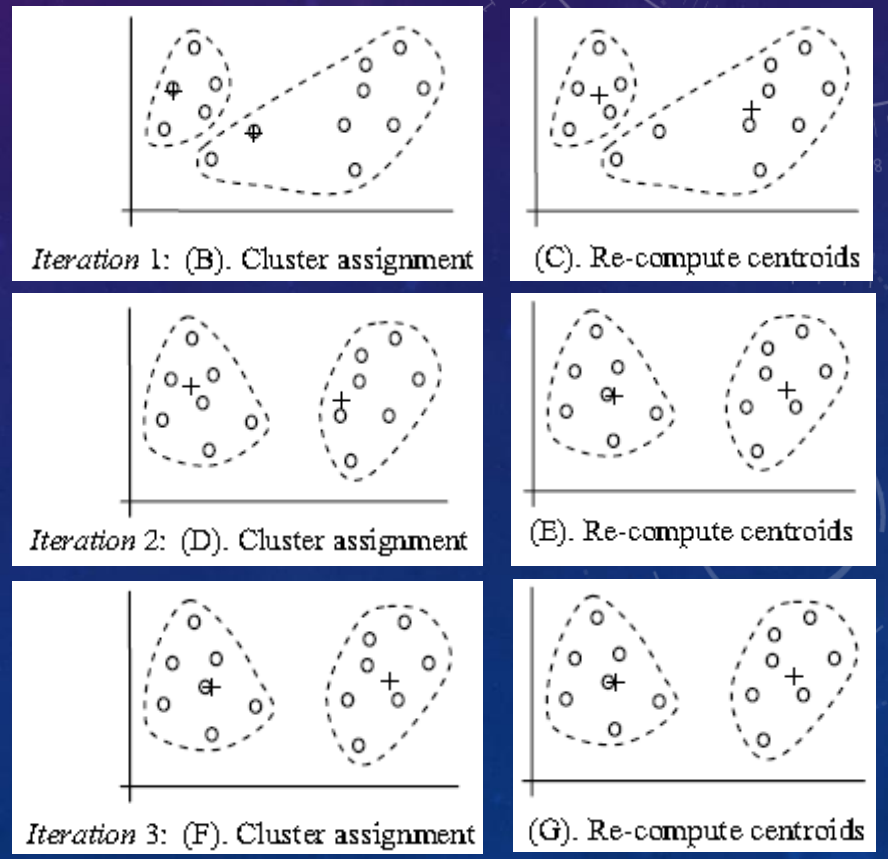
METHODS OF CLUSTERING: PARTITIONING METHODS

- **Partitioning methods**
- Hierarchical methods
- Density-based methods
- Grid-based methods
- Model-based methods

sklearn.cluster.KMeans

K-means clustering simulation





Source: Liu (2007)

METHODS OF CLUSTERING: PARTITIONING METHODS

- **Partitioning methods**
- Hierarchical methods
- Density-based methods
- Grid-based methods
- Model-based methods

K-means clustering demo

sklearn.cluster.KMeans

- https://www.youtube.com/watch?v=5FmnJVv73fU
- https://www.youtube.com/watch?v=BVFG7fd1H30

METHODS OF CLUSTERING: PARTITIONING METHODS

- **Partitioning methods**
- Hierarchical methods
- Density-based methods
- Grid-based methods
- Model-based methods

K-means clustering

- Iterative (as opposed to single path)
- Advantages:
  - Commonly used, frequently, the first method to try
  - Predefined number of clusters
  - Scalable to very large number of objects

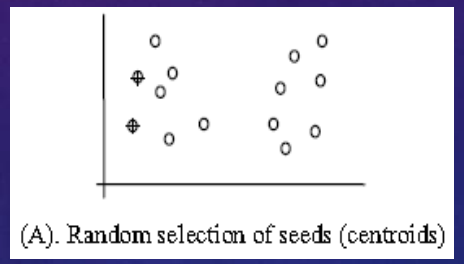
sklearn.cluster.KMeans

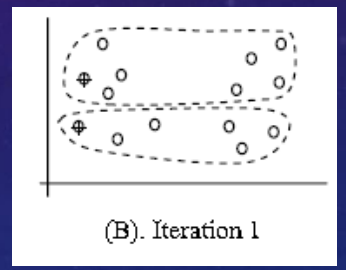
METHODS OF CLUSTERING: PARTITIONING METHODS

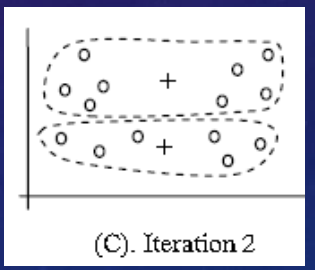
- **Partitioning methods**
- Hierarchical methods
- Density-based methods
- Grid-based methods
- Model-based methods

K-means clustering Weaknesses

- We need to determine our own k
- The algorithm is sensitive to initial seeds.


(A). Random selection of seeds (centroids)

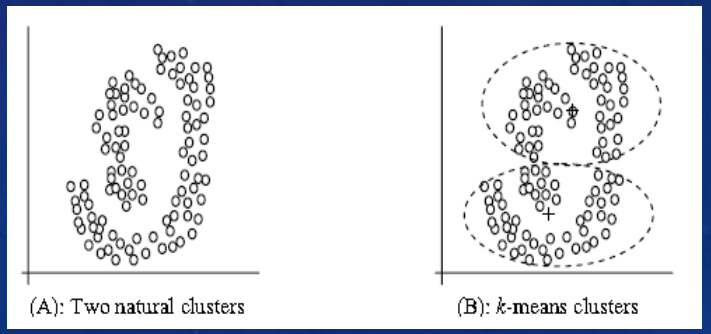

(B). Iteration 1


(C). Iteration 2

- The k-means algorithm is not suitable for discovering clusters that are not hyper-ellipsoids (or hyper-spheres).

Clustering_and_dimensionality_reduction_Part_A.ipynb

sklearn.cluster.KMeans


(A): Two natural clusters

(B): *k*-means clusters

Source: Liu (2007)

DETERMINING K WITH INERTIA

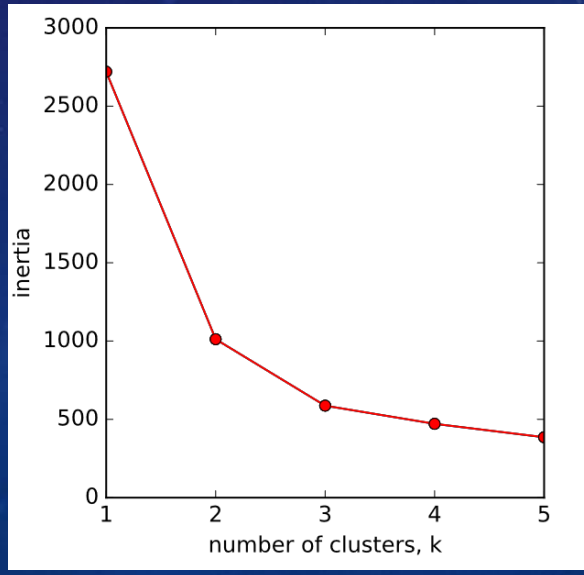
- Inertia is the sum of square distances of samples to their closest cluster center.
  - A measure of variation, or "spread".

$$\sum_{j=1}^{K} \sum_{i \in cluster_j} \left\| x_i - \bar{x}_j \right\|^2$$

When plotting the inertia vs. the number of clusters, we usually see an "elbow".

This elbow indicates when it is no longer necessary to add more clusters.

Clustering_and_dimensionality_reduction_Part_A.ipynb

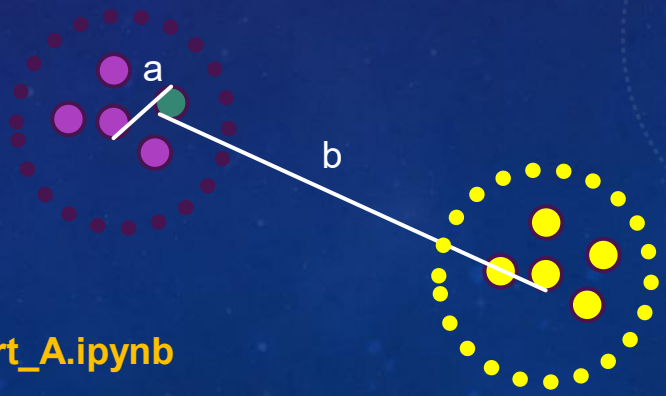


DETERMINING K WITH SILHOUETTE COEFFICIENTS

- Silhouette measure of how close each point in one cluster is to points in the neighboring clusters.
- Can be computed for individual point of average for the entire clustering
- Range of [-1, 1]
  - Coefficients near +1 indicate that the sample is far away from the neighboring clusters.
  - Coefficient of 0 indicates that the sample is on or very close to the decision boundary between two neighboring clusters.
  - Negative coefficients imply that the samples have been assigned to the wrong cluster.

$$\frac{b - a}{\max(a, b)}$$

a – mean distance to observations in own cluster
b – mean distance to observations in nearest cluster



sklearn.metrics.silhouette_score
sklearn.metrics.silhouette_samples

Clustering_and_dimensionality_reduction_Part_A.ipynb

K-MEANS CLUSTERING – (ANOTHER) SIMPLE EXAMPLE OF CUSTOMER SEGMENTATION

Online Sales.ipynb

- Given: record of transaction: customer_id, Quantity, Date, Unit Price.
- Data Source: https://archive.ics.uci.edu/dataset/352/online+retail
- Approach: Segment customers using **RFM (Recency, Frequency, Monetary)** Model
- Outline:
  - Load and clean the dataset
  - Build RFM Features for each customer:
    - Recency (days since last purchase), Frequency (number of purchases), Monetary (total amount spent)
  - Normalize the data (e.g. *sklearn.preprocessing.StandardScaler*)
  - Apply K-means clustering
  - Visualize the results (*sklearn.cluster.KMeans*)

```
scaler = StandardScaler()
rfm_scaled = scaler.fit_transform(rfm)
```

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
375599	569469	22621	TRADITIONAL KNITTING NANCY	6	2011-10-04 12:22:00	1.65	16360	United Kingdom
427	536406	82494L	WOODEN FRAME ANTIQUE WHITE	6	2010-12-01 11:33:00	2.55	17850	United Kingdom
374619	569384	22210	WOOD STAMP SET BEST WISHES	12	2011-10-03 16:47:00	0.83	15159	United Kingdom
11688	537262	22114	HOT WATER BOTTLE TEA AND SYMPATHY	1	2010-12-06 11:26:00	3.95	15039	United Kingdom
464246	576081	21506	FANCY FONT BIRTHDAY CARD,	12	2011-11-14 08:16:00	0.42	15203	United Kingdom

METHODS OF CLUSTERING: HIERARCHICAL METHODS

- Partitioning methods
- **Hierarchical methods**
- Density-based methods
- Grid-based methods
- Model-based methods

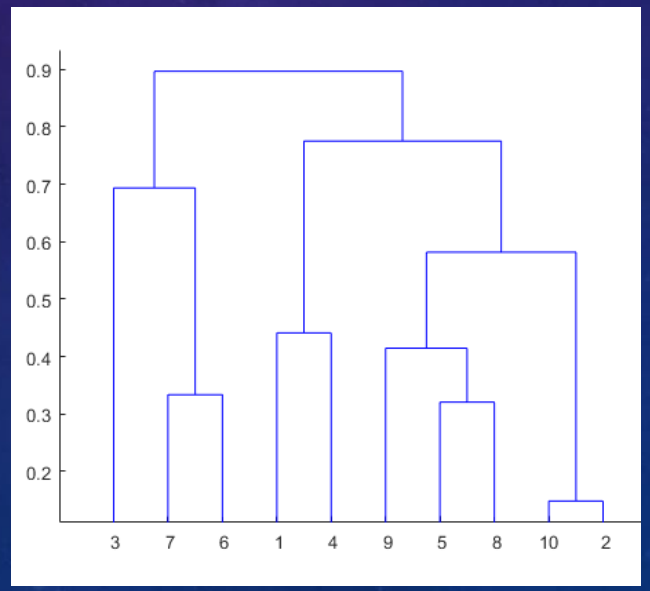
Aggregative clustering algorithm:

Initialize: Declare every object to be a cluster of its own

Body:

1. Find a pair of closest clusters and merge them into a new (super)cluster

Stop: Continue until more that 1 cluster remains



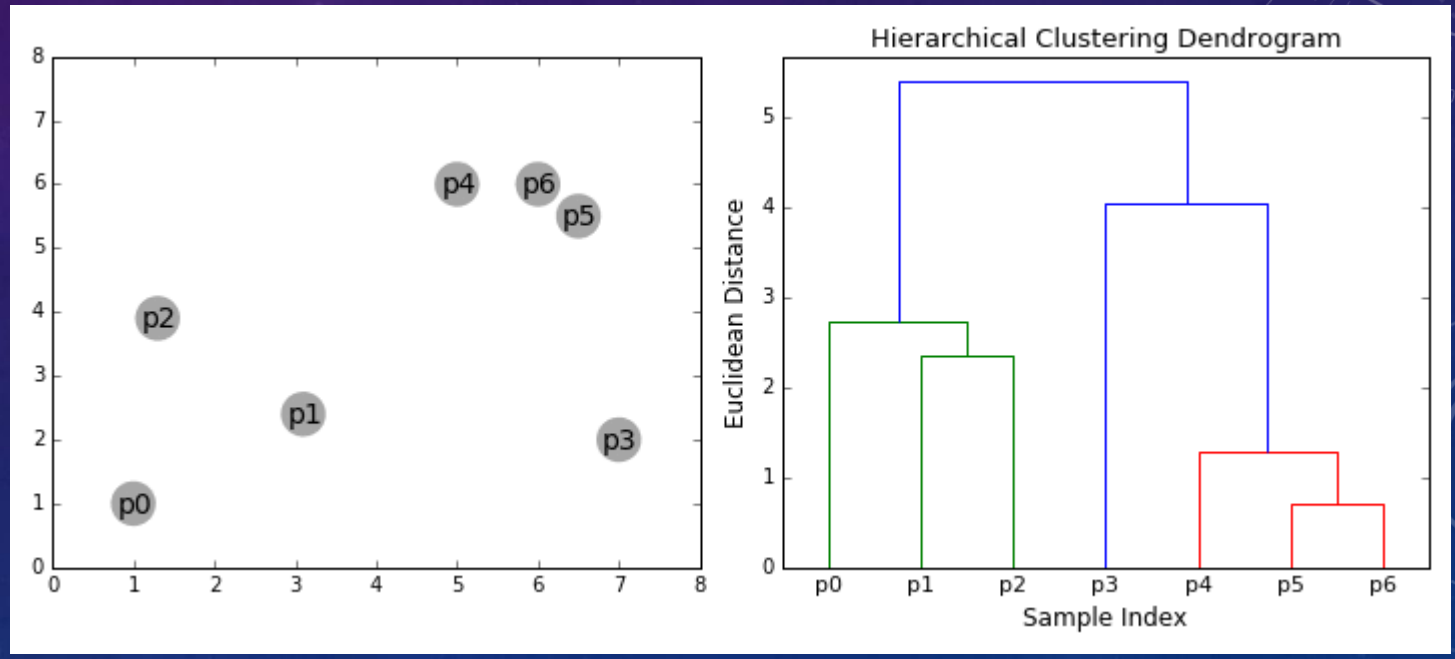
Ward Jr, Joe H. "Hierarchical grouping to optimize an objective function." Journal of the American statistical association58.301 (1963): 236-244.

sklearn.cluster.AgglomerativeClustering

scipy.cluster.hierarchy.linkage

METHODS OF CLUSTERING: HIERARCHICAL METHODS

- Partitioning methods
- **Hierarchical methods**
- Density-based methods
- Grid-based methods
- Model-based methods



sklearn.cluster.AgglomerativeClustering
scipy.cluster.hierarchy.linkage

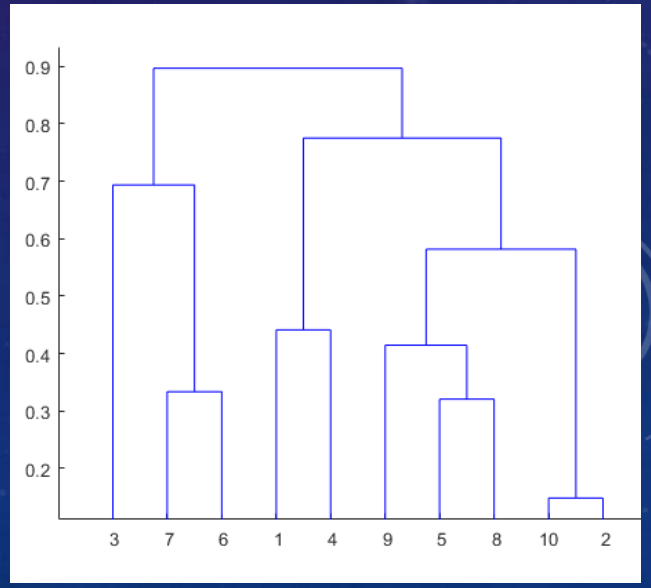
METHODS OF CLUSTERING: HIERARCHICAL METHODS

- Partitioning methods
- **Hierarchical methods**
- Density-based methods
- Grid-based methods
- Model-based methods

- Produces hierarchy of clusters (unlike partitioning)
- Each object belongs to a sequence of nested clusters
- There is a divisive method that splits large clusters into smaller ones

Hierarchical methods differ by:

- Linkage: How they identify the closest clusters to merge (or split)
- How they maintain record of clusters



sklearn.cluster.AgglomerativeClustering
scipy.cluster.hierarchy.linkage

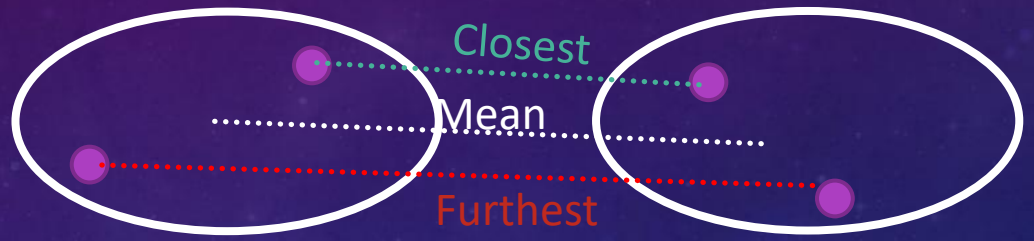
METHODS OF CLUSTERING: HIERARCHICAL METHODS

- Partitioning methods
- **Hierarchical methods**
- Density-based methods
- Grid-based methods
- Model-based methods

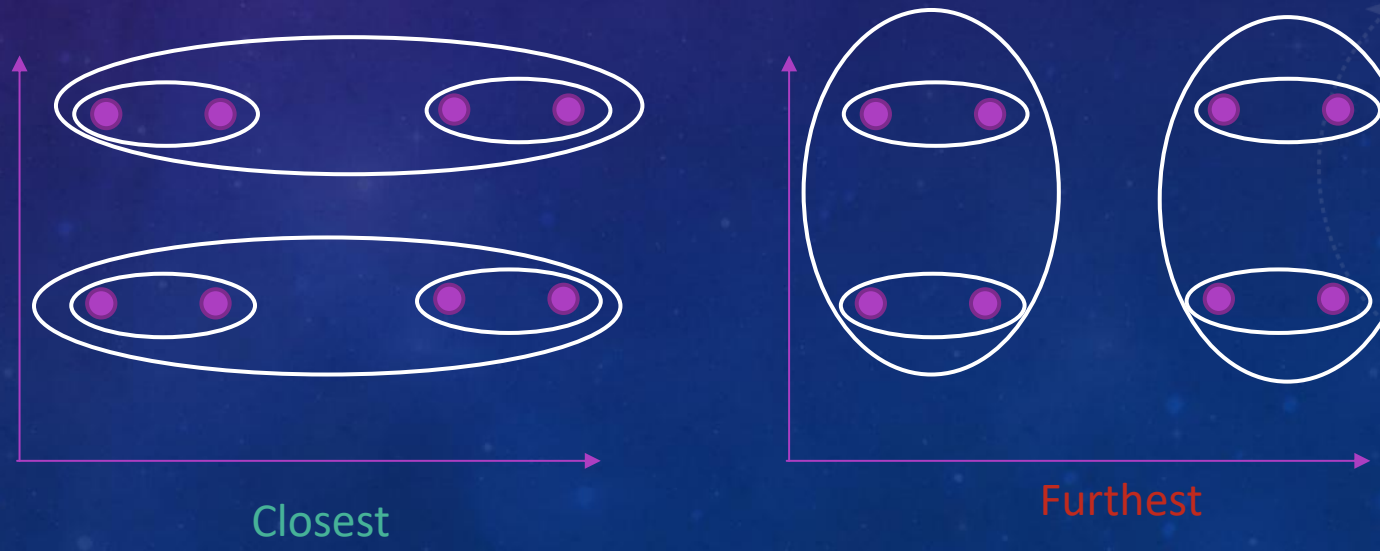
SKLearn default: **ward** – minimizes variance of the clusters being merged

sklearn.cluster.AgglomerativeClustering

The definition of “Closest” may affect clustering outcome



Which one to optimize?

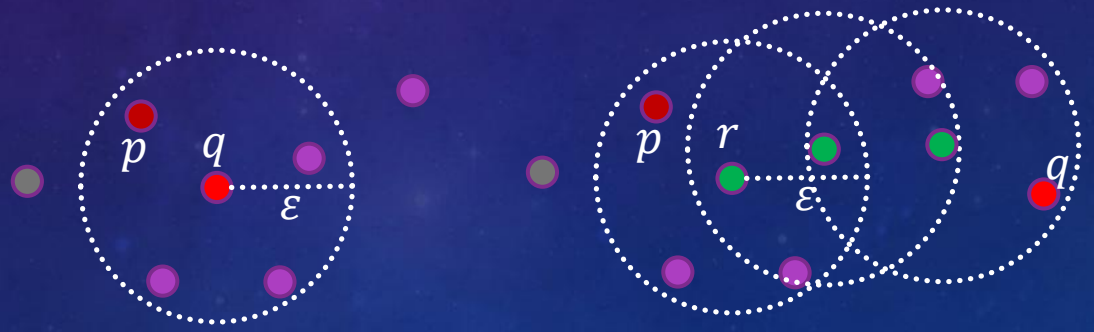


METHODS OF CLUSTERING: DENSITY-BASED METHODS

- Partitioning methods
- Hierarchical methods
- **Density-based methods**
- Grid-based methods
- Model-based methods

- Clusters are defined as areas of *higher density* than the remainder of the data set.
- Objects in the sparse area that separates clusters are considered noise/outliers or border objects
- Density Reachability: point p is density-reachable from q if
  - p is within ε from q and
  - q has sufficient number of points in its neighborhood.

q is not density-reachable by p if q is lonely (even if they are close)



- Density Connected: (chaining) point p is density-connected to q if there exists point r which:
  - Has sufficient number of points in its neighborhoods
  - Both p and q are within ϵ from its neighborhoods

Density-Connectivity: allows connecting overlapping dense regions and to spread beyond ε

METHODS OF CLUSTERING: DENSITY-BASED METHODS

- Partitioning methods
- Hierarchical methods
- **Density-based methods**
- Grid-based methods
- Model-based methods

Ester, Martin, et al. "A density-based algorithm for discovering clusters in large spatial databases with noise." *Kdd*. Vol. 96. No. 34. 1996.

sklearn.cluster.DBSCAN

DBSCAN: Density-Based Spatial Clustering of Applications with Noise

- Requires two parameters: ε –distance
  minPts – minimal number of points

Initialize: Start with arbitrary point p

Body:

1. Extract neighborhood N_p of the chosen point (within distance ε of p)
2. If size of the neighborhood, $|N_p| > minPts$:

   Start clustering points

   Else:

   p is labeled as noise (can later still become part of a cluster)

   Mark p as visited
3. If point p becomes part of a cluster, then its neighborhood N_p is also part of the cluster

Stop: Continue until unvisited points p are still available

METHODS OF CLUSTERING: DENSITY-BASED METHODS

- Partitioning methods
- Hierarchical methods
- **Density-based methods**
- Grid-based methods
- Model-based methods

DBSCAN: Density-Based Spatial Clustering of Applications with Noise

Clustering points (if $|N_p| > minPts$):

1. If **none** of the points in N_p belongs to a cluster (new cluster)
   - assign all points in N_p to a new cluster
2. If **one** of the points in N_p belongs to some cluster C (attach to cluster)
   - Assign all points in N_p to that cluster
3. If points in N_p belong to **more than one** cluster (merge clusters)
   - Pick one of the clusters (say, with minimal index C)
   - Assign all points in N_p to C
   - Re-assign points in all other clusters observed in N_p to C

sklearn.cluster.DBSCAN

METHODS OF CLUSTERING: DENSITY-BASED METHODS

- Partitioning methods
- Hierarchical methods
- **Density-based methods**
- Grid-based methods
- Model-based methods

DBSCAN: Density-Based Spatial Clustering of Applications with Noise

Advantages

1. Doesn't require a-priori specification of the number of clusters
2. Insensitive to noise. Identifies noise points
3. Finds clusters of arbitrary (and uneven!) size and shape

1. Disadvantages
    1. Fails if density of clusters varies. Alternatives:
        1. HDBSCAN (Hierachical DBSCAN)
        2. OPTICS (Ordering Points To Identify the Clustering Structure)
    2. Doesn't work well with high-dimensional data

sklearn.cluster.DBSCAN

Ester, Martin, et al. "A density-based algorithm for discovering clusters in large spatial databases with noise." *Kdd*. Vol. 96. No. 34. 1996.

METHODS OF CLUSTERING: DENSITY-BASED METHODS

- Partitioning methods
- Hierarchical methods
- **Density-based methods**
- Grid-based methods
- Model-based methods

Mean Shift algorithm

Input: *bandwidth* (spatial size of the cluster)

python can try and automatically) estimate this parameter by running another clustering algorithm and measuring typical cluster bandwidth.

Basic step: locate the density function maxima

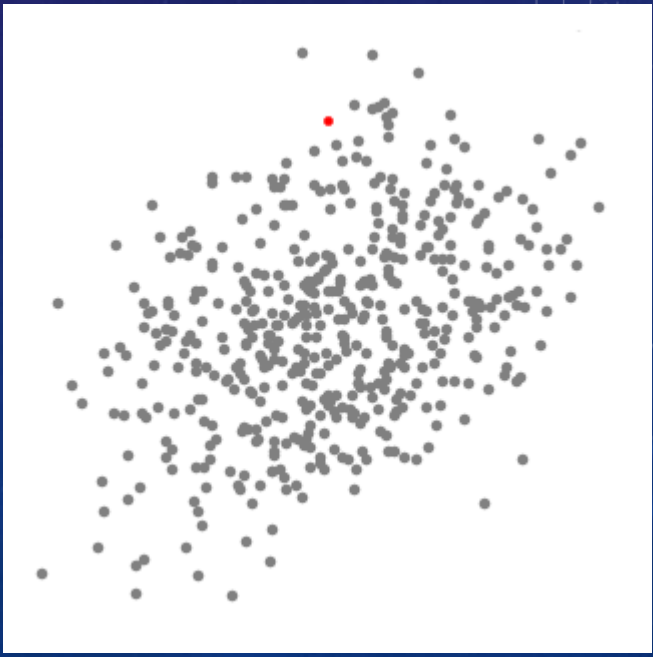
Initialize: Place ball of given bandwidths (r) at a random point

Body:

Shift the ball center to mean of all points

Stop: Continue until the ball stops moving.

The ball will be moving towards highest density region



sklearn.cluster.MeanShift

Fukunaga, Keinosuke, and Larry Hostetler. "The estimation of the gradient of a density function, with applications in pattern recognition." IEEE Transactions on information theory 21.1 (1975): 32-40.

METHODS OF CLUSTERING: DENSITY-BASED METHODS

- Partitioning methods
- Hierarchical methods
- **Density-based methods**
- Grid-based methods
- Model-based methods

Mean Shift algorithm

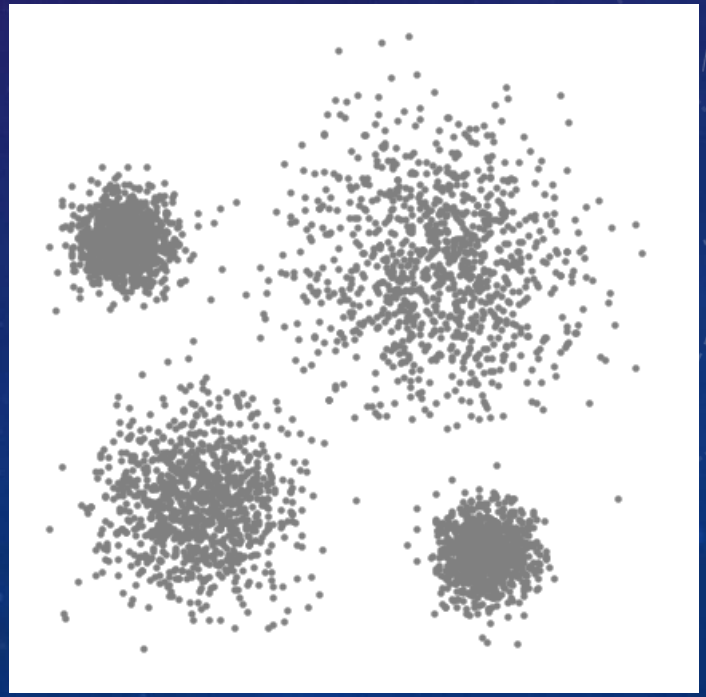
Initialize: Drop many balls of bandwidth r in random locations

Body:

1. Let each ball shift its center to its own mean until it (nearly) stops
2. If balls overlap, the ball containing most points is preserved

Stop: Continue until balls stop moving

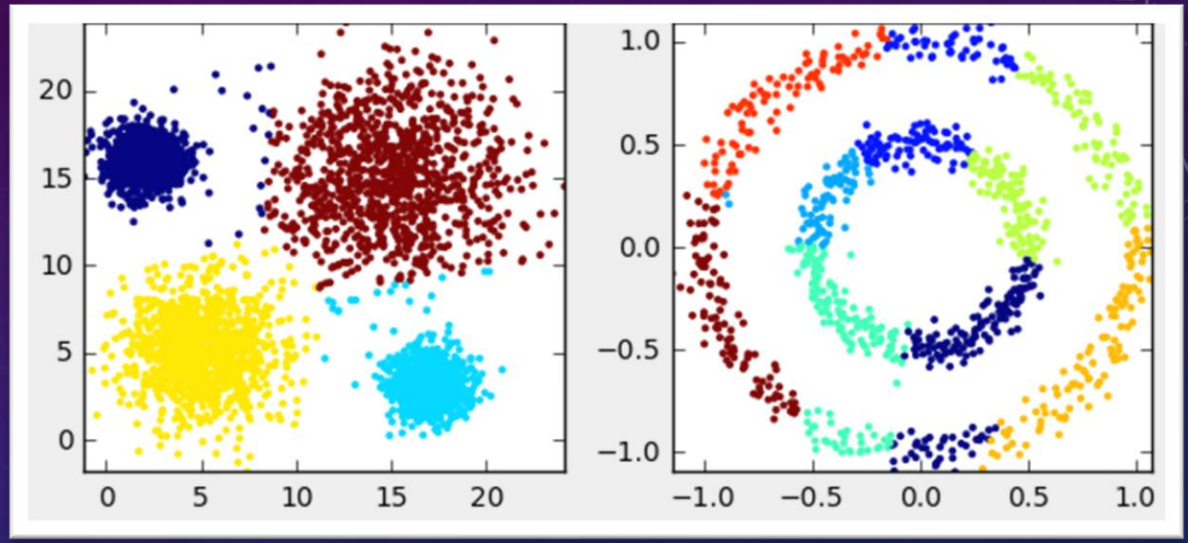
Observations are clustered according to their ball



sklearn.cluster.MeanShift

METHODS OF CLUSTERING: DENSITY-BASED METHODS

- Partitioning methods
- Hierarchical methods
- **Density-based methods**
- Grid-based methods
- Model-based methods



Advantages:

1. Simple
2. Determines the number of clusters automatically

Disadvantages:

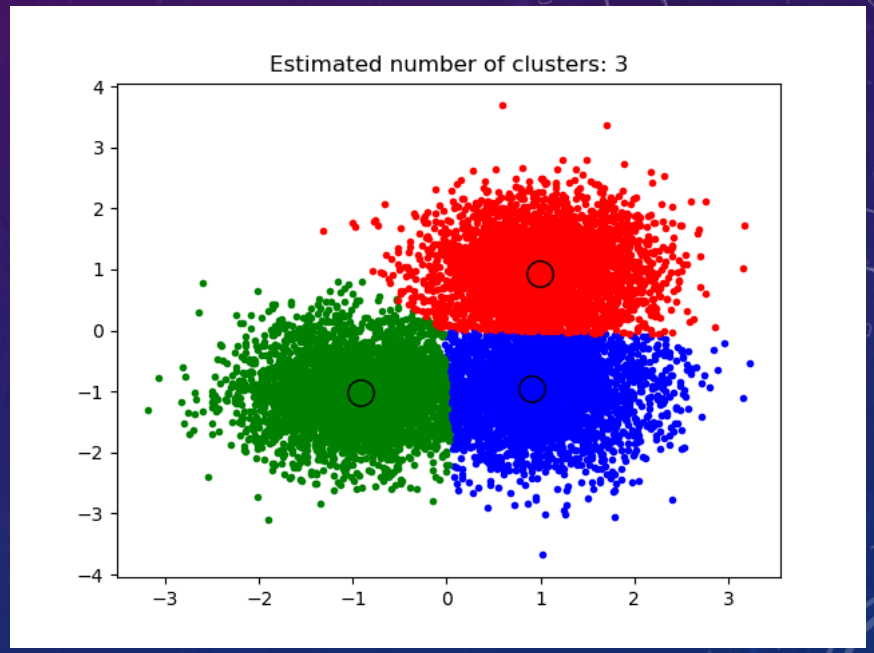
1. Computationally demanding: $O(T * n^2)$
   T – number of iterations, n – number of points
1. Can't detect cluster of any shape
2. Bandwidth is difficult to guess
3. Same bandwidth may not fit all clusters

sklearn.cluster.MeanShift

METHODS OF CLUSTERING: DENSITY-BASED METHODS

- Partitioning methods
- Hierarchical methods
- **Density-based methods**
- Grid-based methods
- Model-based methods

- Mean shift clustering algorithm can find centers of mass
- Detects the number of clusters automatically



python can try and automatically) estimate this parameter by running another clustering algorithm and measuring typical cluster bandwidth.

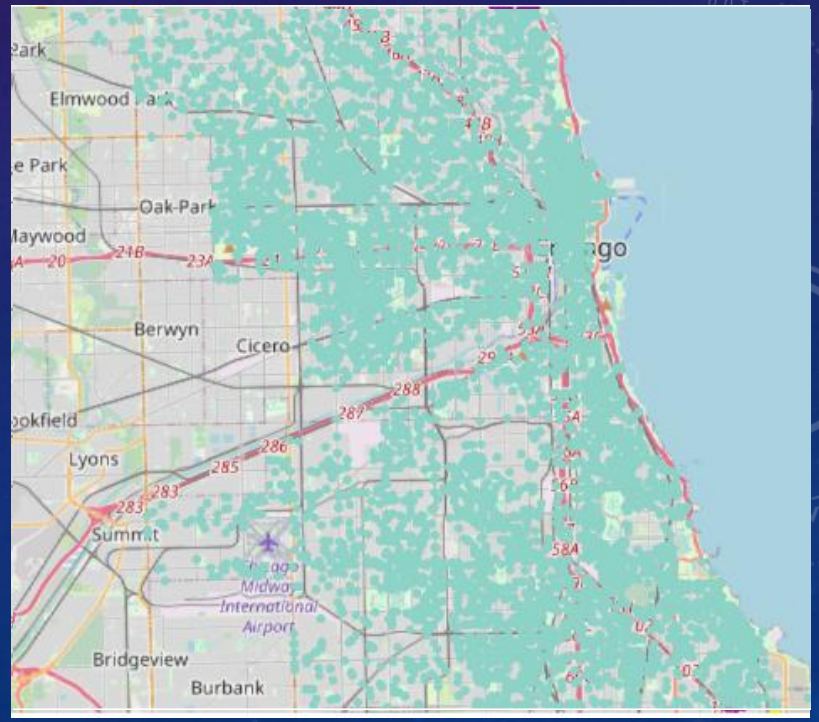
from sklearn.cluster import estimate_bandwidth
bandwidth = estimate_bandwidth(X, quantile=0.2, n_samples=500)

sklearn.cluster.MeanShift

DBSCAN EXAMPLE: CHICAGO CRIME

Chicago Crime Clusters.ipynb

- Density-based methods are particularly useful for clustering geo-data
- Data: City of Chicago Crime Data

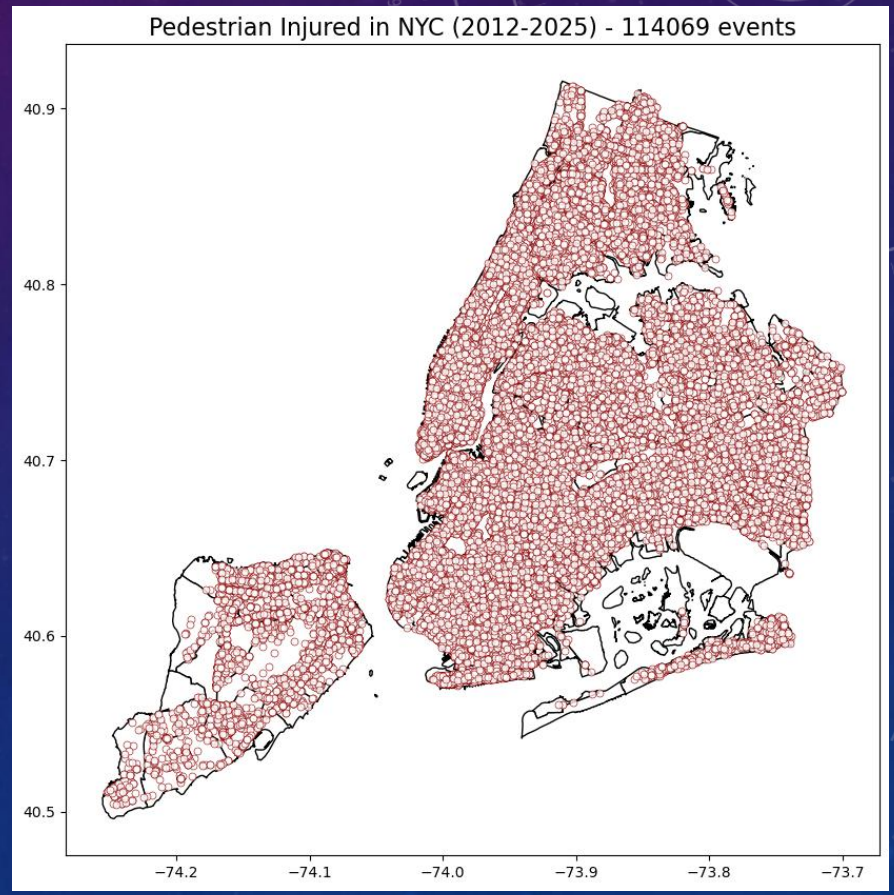


NYC PEDESTRIAN CASUALTIES

- Objective: Find least safe areas for NYC pedestrians

- NYC Vehicle Collision Data:
  - https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95/about_data
  - 1.9M car crash events with coordinates and damage
- Method:
  - Use DBSCAN to cluster accidents with injuries or fatalities involving pedestrians

PedestrianSafetyNYC.ipynb



METHODS OF CLUSTERING: GRID-BASED METHODS

- Partitioning methods
- Hierarchical methods
- Density-based methods
- **Grid-based methods**
- Model-based methods

Grid-based clustering methods:

- Partition the data space into cells to form a grid structure
- Find densely populated cells (clusters)
- Refine the grid to find finer clusters

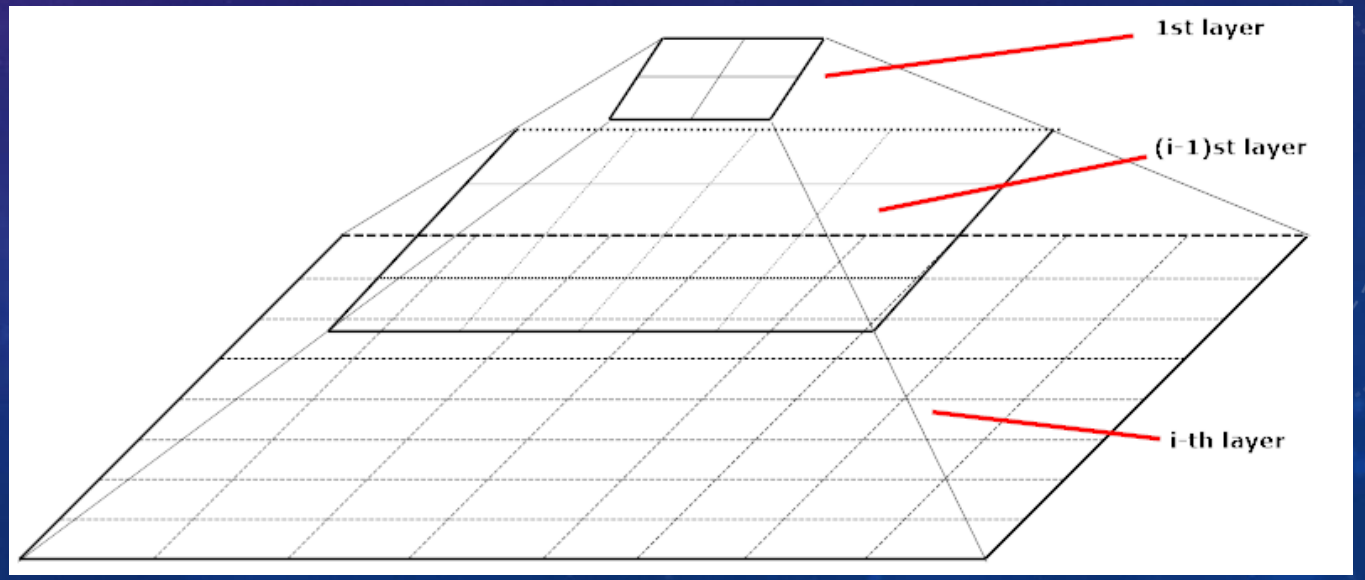
METHODS OF CLUSTERING: GRID-BASED METHODS

- Partitioning methods
- Hierarchical methods
- Density-based methods
- **Grid-based methods**
- Model-based methods

Statistical Information Grid Approach (STING)

Initialize: The space is divided into cells of different levels of resolution.

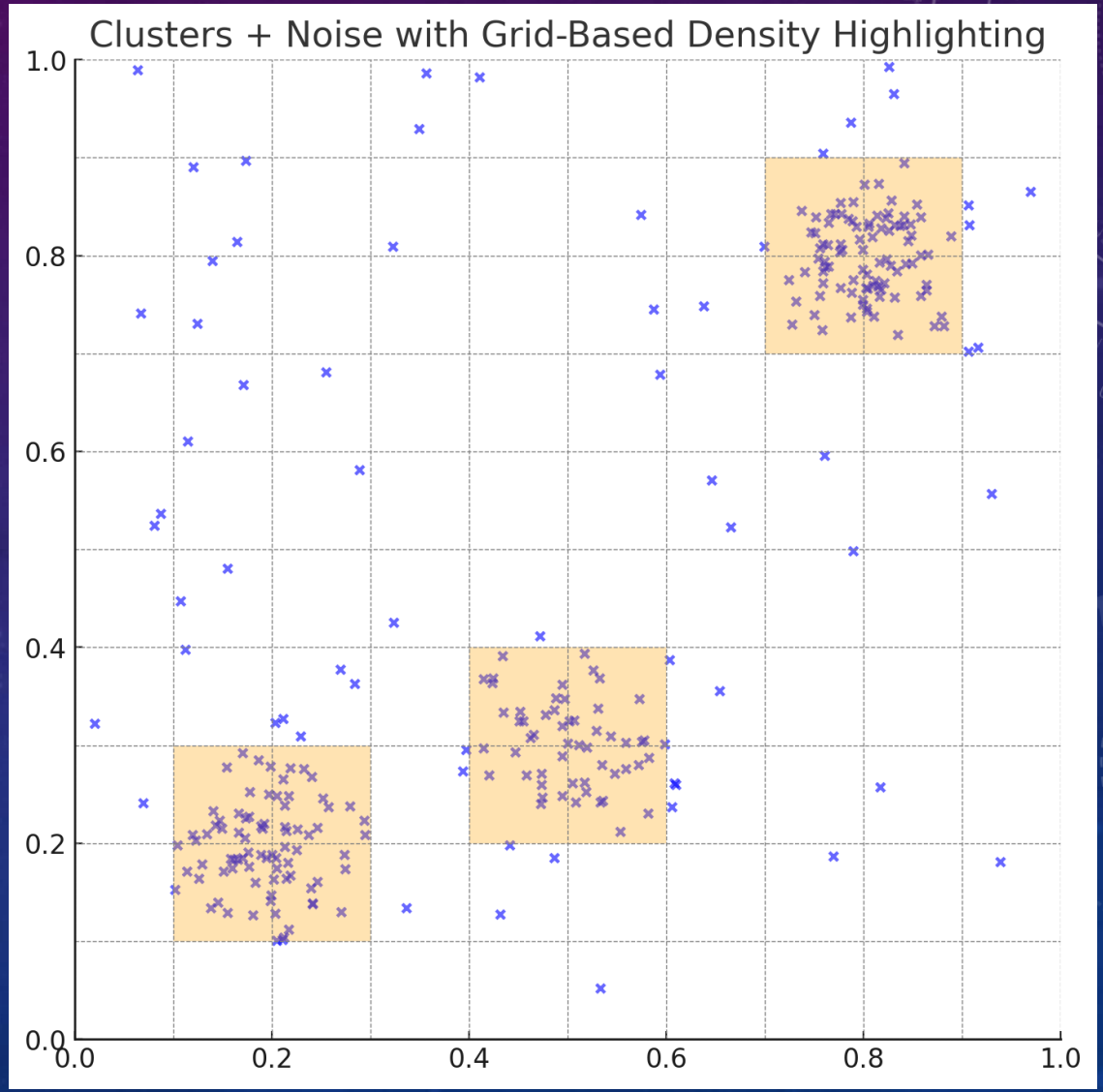
- Different levels of cells correspond to different resolutions
- These cells form a **tree structure**: Higher-level cell (parent) is partitioned into lower-levels cells (children)



Wang, Wei, Jiong Yang, and Richard Muntz. "STING: A statistical information grid approach to spatial data mining." *VLDB*. Vol. 97. 1997.

METHODS OF CLUSTERING: GRID-BASED METHODS

- Partitioning methods
- Hierarchical methods
- Density-based methods
- **Grid-based methods**
- Model-based methods



Gaussian Clusters

Wang, Wei, Jiong Yang, and Richard Muntz. "STING: A statistical information grid approach to spatial data mining." *VLDB*. Vol. 97. 1997.

METHODS OF CLUSTERING: GRID-BASED METHODS

- Partitioning methods
- Hierarchical methods
- Density-based methods
- **Grid-based methods**
- Model-based methods

Statistical Information Grid Approach (STING)

Initialize:

1. The space is divided into cells of different levels of resolution.
2. Statistical info for each cell is computed and stored ‘within’ the cell.
   - Count, min, max, ….
   - Type of distribution (uniform, normal, ..)
3. Parameters of the parent cell is calculated from parameters of the child cells

Body: Use top-down approach to answer spatial queries

1. Start from a relatively high level
2. For each cell, compute the probability that it contains a cluster.
   1. Remove irrelevant cells from consideration
   2. Move to the next level, considering only children of the relevant cells

Wang, Wei, Jiong Yang, and Richard Muntz. "STING: A statistical information grid approach to spatial data mining." *VLDB*. Vol. 97. 1997.

METHODS OF CLUSTERING: GRID-BASED METHODS

- Partitioning methods
- Hierarchical methods
- Density-based methods
- **Grid-based methods**
- Model-based methods

Clustering In QUEst (CLIQUE)

Subspace Clustering:

Finds clusters in subsets of dimensions — important when clusters exist only in certain combinations of features.

Grid Partitioning:

Each dimension is divided into equal-width intervals - forms multi-dimensional grid cells (units).

Density Threshold:

A unit is **dense** if it contains more than a minimum number of points.

Wang, Wei, Jiong Yang, and Richard Muntz. "STING: A statistical information grid approach to spatial data mining." *VLDB*. Vol. 97. 1997.

METHODS OF CLUSTERING: GRID-BASED METHODS

- Partitioning methods
- Hierarchical methods
- Density-based methods
- **Grid-based methods**
- Model-based methods

```
pip install pyclustering

from pyclustering.cluster.clique import clique
```

Clustering In QUEst (CLIQUE)

Partition Feature Space:

Each dimension is divided into ξ intervals (e.g., 10 bins).The full space becomes a grid of hyper-rectangular units.

Identify Dense Units in 1D:

For each dimension, find bins with enough points. Mark those as 1D dense units.

Bottom-Up Subspace Search:

Combine dense units from lower-dimensional subspaces (Apriori-like).

Search for dense units in 2D, 3D, ..., kD.

Only explore combinations of previously found dense units.

Cluster Formation:

Merge connected dense units into clusters. Points inside these connected regions are assigned to the same cluster.

Wang, Wei, Jiong Yang, and Richard Muntz. "STING: A statistical information grid approach to spatial data mining." *VLDB*. Vol. 97. 1997.

METHODS OF CLUSTERING: GRID-BASED METHODS

- Partitioning methods
- Hierarchical methods
- Density-based methods
- **Grid-based methods**
- Model-based methods

Statistical Information Grid Approach (STING)

- Advantages:
  1. Easy to parallelize
  2. Very fast – O(K), where K is the number of cells at the bottom level

     (speed is defined by the number of cells, which is smaller than the number of observations)
- Disadvantage:
  1. All cluster boundaries are either horizontal or vertical. Diagonal boundaries are not detected.
- Note: this algorithm can be considered grid-based or model-based clustering

Wang, Wei, Jiong Yang, and Richard Muntz. "STING: A statistical information grid approach to spatial data mining." *VLDB*. Vol. 97. 1997.

METHODS OF CLUSTERING: MODEL-BASED METHODS

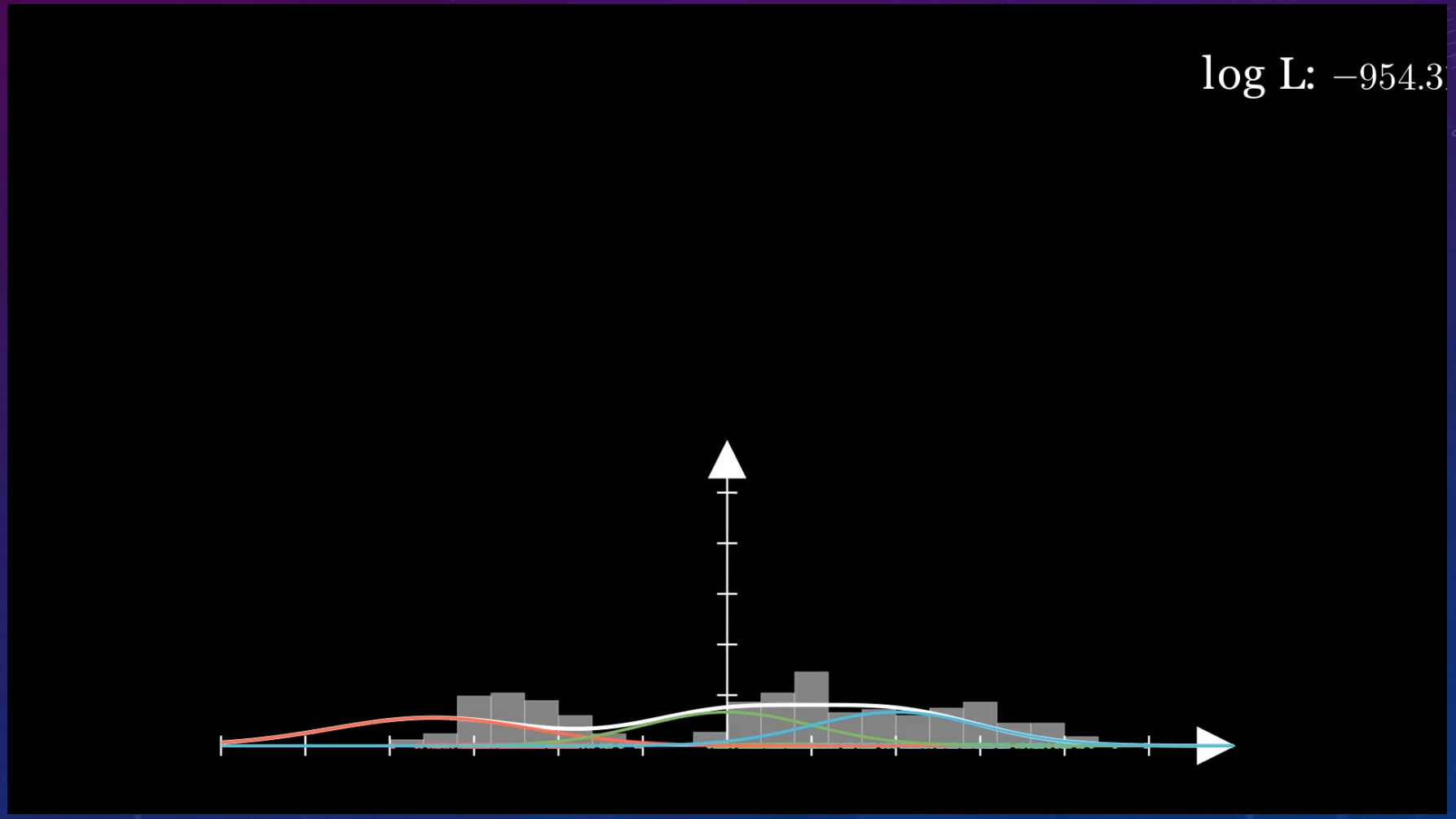
- Partitioning methods
- Hierarchical methods
- Density-based methods
- Grid-based methods
- **Model-based methods**

- All previous clustering algorithms rely on some kind of heuristic:
  - find closest cluster centroid (k-means)
  - move towards the center of mass (mean-shift)
- This is not necessary a disadvantage (clustering is largely exploratory)
- Model-based clustering methods:
  - Attempt to optimize the fit between the data and some mathematical model
  - Based on the assumption that the data are generated by the underlying probability distribution

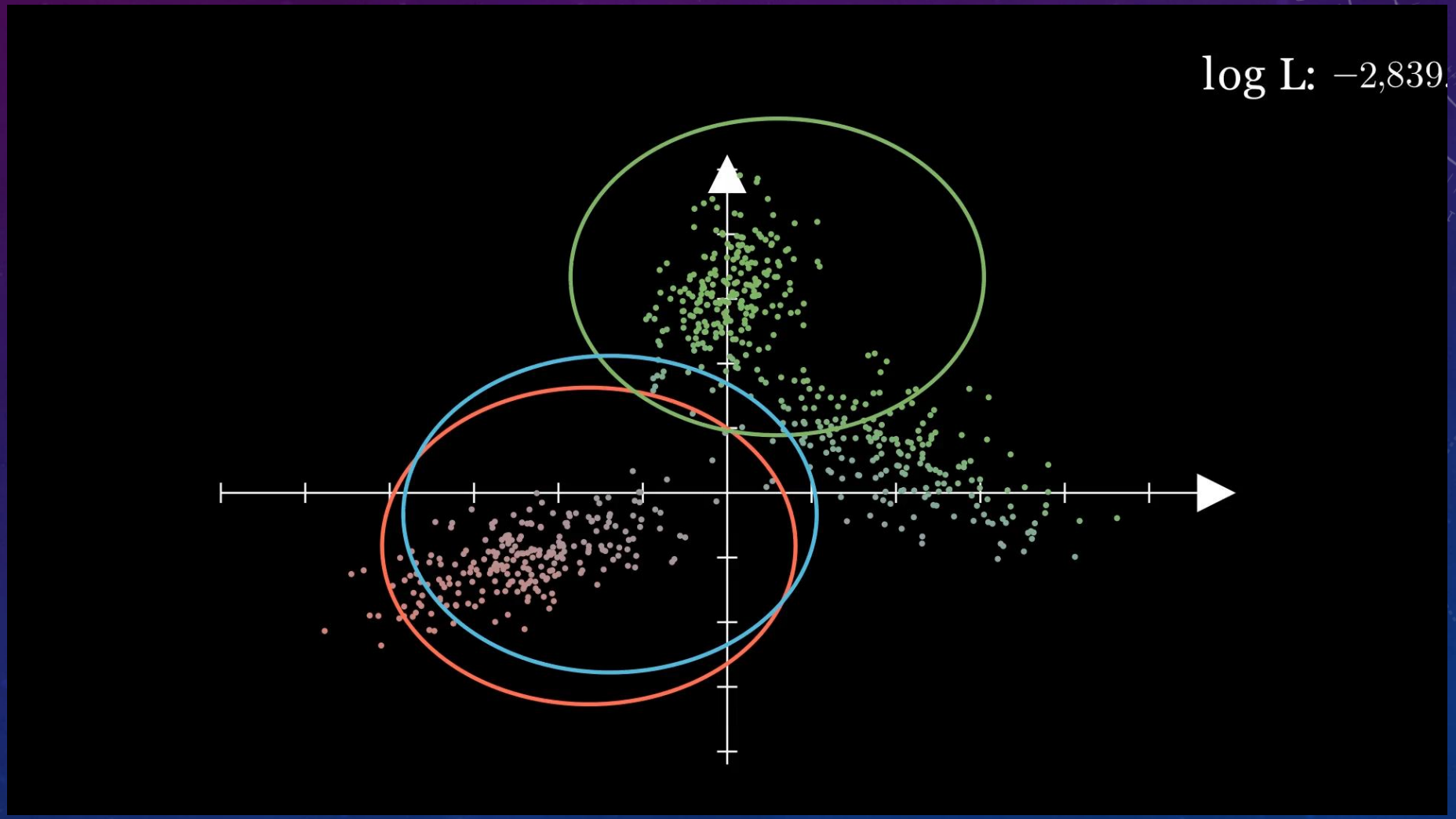
1. Banfield, Jeffrey D., and Adrian E. Raftery. "Model-based Gaussian and non-Gaussian clustering." *Biometrics* (1993): 803-821.
2. Melnykov, Volodymyr, and Ranjan Maitra. "Finite mixture models and model-based clustering." *Statistics Surveys* 4 (2010): 80-116.

sklearn.mixture.GaussianMixture

METHODS OF CLUSTERING: MODEL-BASED METHODS



METHODS OF CLUSTERING: MODEL-BASED METHODS



METHODS OF CLUSTERING: MODEL-BASED METHODS

- Partitioning methods
- Hierarchical methods
- Density-based methods
- Grid-based methods
- **Model-based methods**

Gaussian Expectation Maximization

- Iterative refinement (similar to k-means)

Initialize: Randomly assign **K** cluster centers

Body:

1. **Expectation step**:

   assign each data point x_i to cluster C_k with probability $P(C_k|x_i) = \frac{P(C_k)P(x_i|C_k)}{P(x_i)}$

2. Maximization Step:

   (re)Estimate model parameters: $m_k = \frac{1}{N}\sum_{i=1}^{N}\frac{X_i P(C_k|x_i)}{\sum_j P(C_j|x_i)}$

Stop: continue until the distribution of clusters stops changing significantly.

sklearn.mixture.GaussianMixture

METHODS OF CLUSTERING: MODEL-BASED METHODS

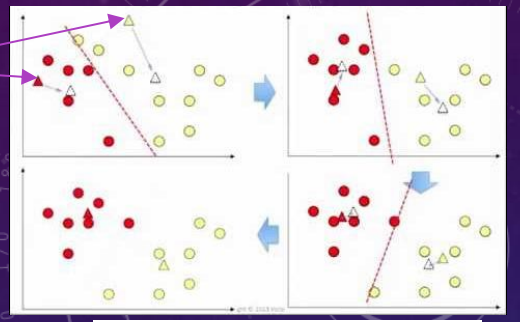
- Partitioning methods
- Hierarchical methods
- Density-based methods
- Grid-based methods
- **Model-based methods**

Gaussian Expectation Maximization

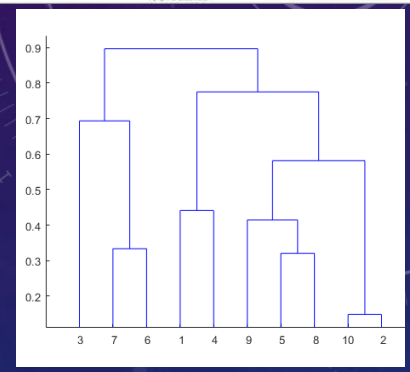
- Features:
  - User can define probability distributions (Gaussian or any other?)
  - "Soft" clustering: each point x_i is assigned with every cluster C_k with probability with probability $P(C_k|x_i)$
- Advantages:
  - Converges fast
- Disadvantages:
  - May converge to local optima

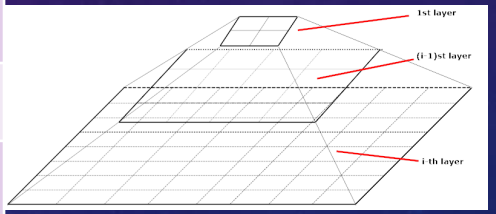
sklearn.mixture.GaussianMixture

METHODS OF CLUSTERING: SUMMARY



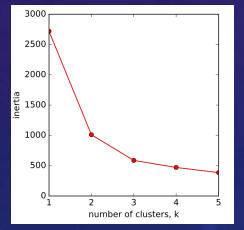
Family	Methods
Partitioning methods	sklearn.cluster.KMeans
Hierarchical methods	sklearn.cluster.AgglomerativeClustering
Density-based methods	sklearn.cluster.DBSCAN sklearn.cluster.MeanShift
Grid-based methods	
Model-based methods	sklearn.mixture.GaussianMixture, $P(C_k\|x_i) = \frac{P(C_k)P(x_i\|C_k)}{P(x_i)}$



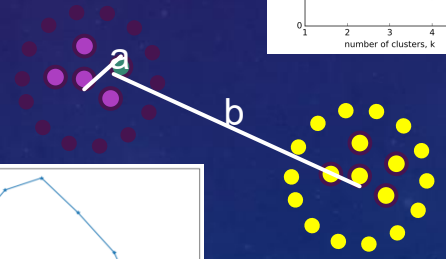


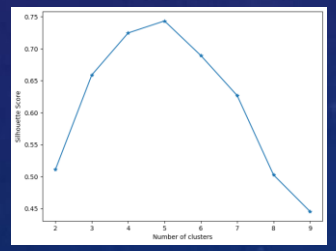
Clustering Evaluation:

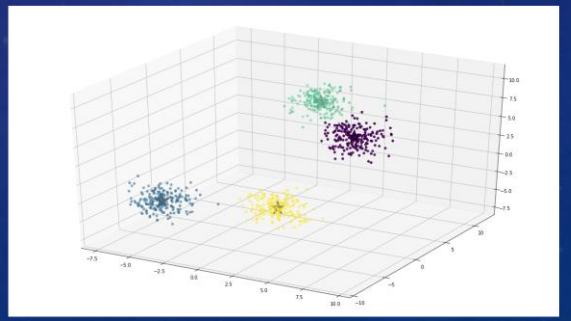
- Inertia: $\sum_{j=1}^{K} \sum_{i \in cluster_j} \|x_i - \bar{x}_j\|^2$



- Silhouette: $\frac{b-a}{\max(a,b)}$



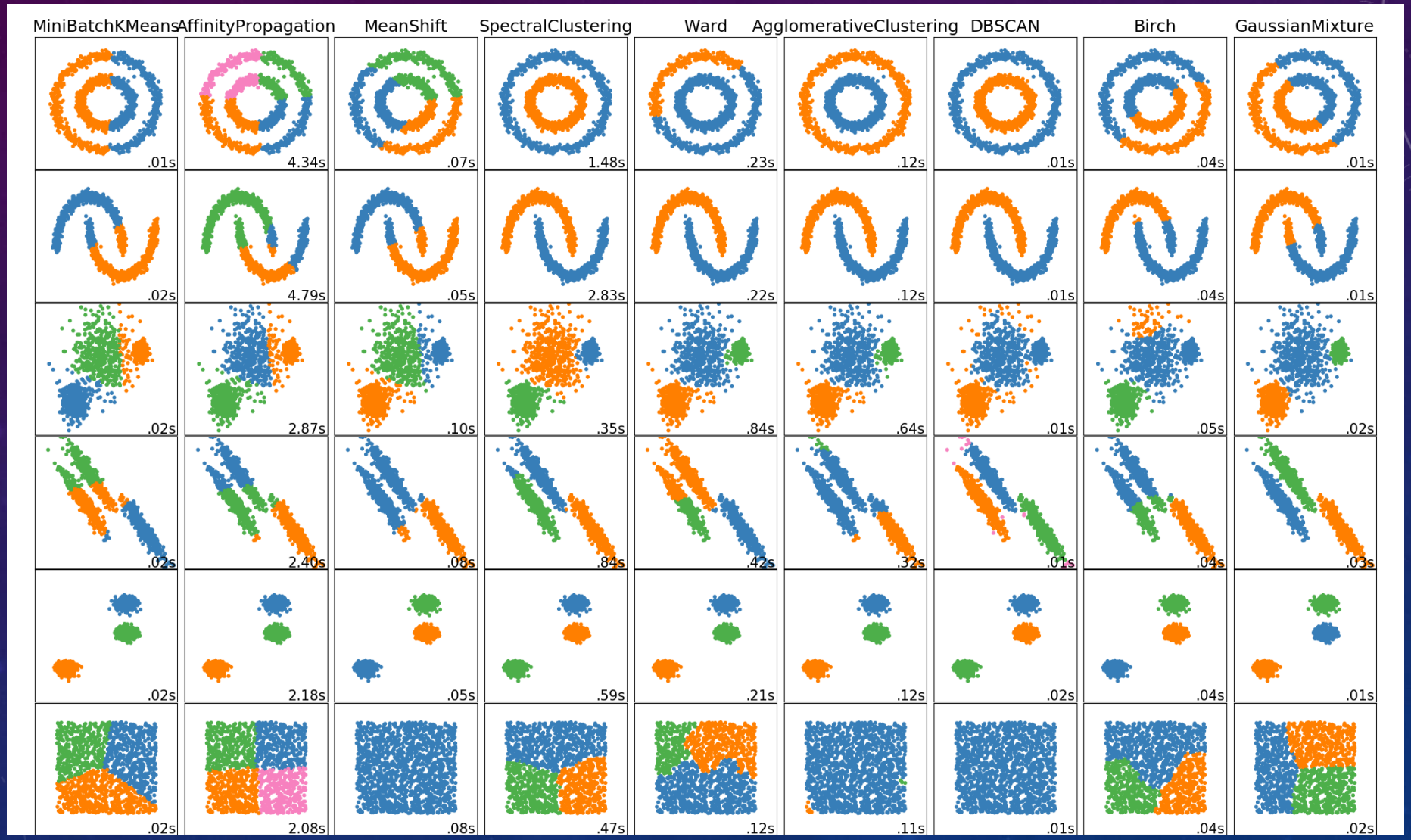




Clustering_and_dimensionality_reduction_Part_A.ipynb

Demonstrate clustering with synthetic example

SKLEARN CLUSTERING METHODS



https://scikit-learn.org/stable/modules/clustering.html

DIMENSIONALITY REDUCTION - AGENDA

- The curse of dimensionality
- Principal component analysis (PCA)

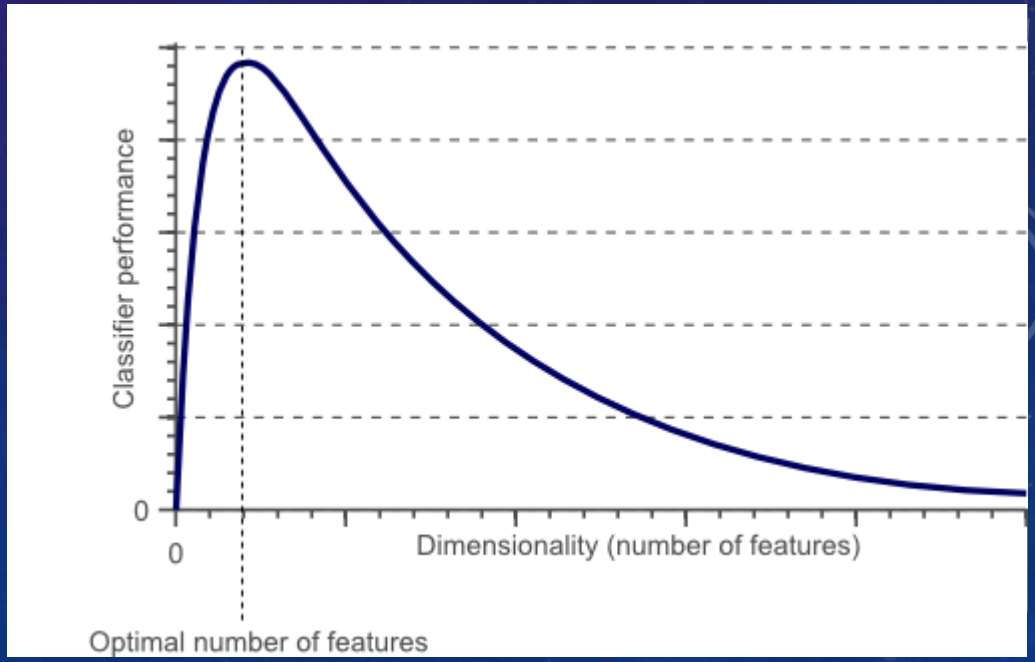
PCA - Synthetic example.ipynb
Clustering_and_dimensionality_reduction_Part_B.ipynb
PCA\US_Elections_Dimensionality_Reduction.ipynb
Dimentionality_Reduction_Knesset_Votings.ipynb
Clustering_and_dimensionality_reduction_Part_C.ipynb
sports_tsne.ipynb
Enron_network_analysis.ipynb

THE CURSE OF DIMENSIONALITY

- In favor of additional features:
  - Richer, detailed information
- Against additional features:
  - Less comprehendible
  - Correlative to already present features (redundant)
  - Could add noise
  - Difficult / Expensive to collect
  - Require additional RAM & add computational complexity
  - Data-hungry (exponentially growing requirements for data):

$$number\ of\ samples \sim sample\ density^{dimensions}$$

Danger of overfitting!



DIMENSIONALITY REDUCTION

- The objective is to summarize data with many variables (p) by a smaller set of k, synthetic variables.
- The algorithm will remove correlated piece and find the transformation that preserves most of the information
- The data will be easier to present and classify

DIMENSIONALITY REDUCTION

- "**Residual**" – variation in information in A that is not retained in X
- Dimensionality reduction must strive to *fewer dimensions*
  - Lead to clarity of representation and ease of understanding
- But avoid oversimplification due to loss of significant information

Dimensionality reductions is always a
Tradeoff between the number of reduced dimensions (in X) and the loss of data

DIMENSIONALITY REDUCTION: PCA

- PCA – Principal Component Analysis
- Closely related to Principal Factor Analysis
- Invented in 1901 (Pearson)

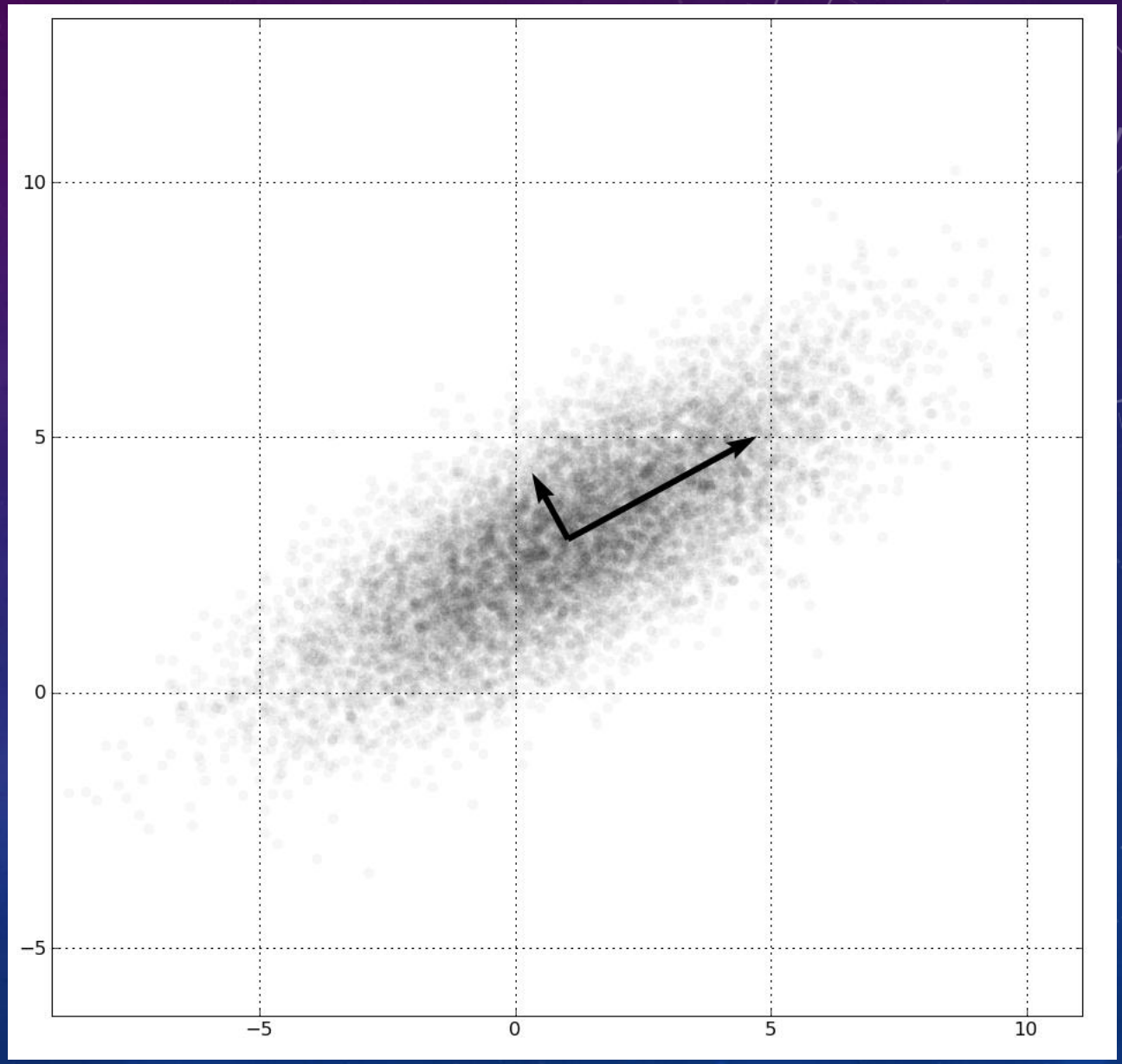
- PCA transforms $n \times p$ matrix A in which different dimensions p can be **correlated** into another matrix A' with **uncorrelated dimensions** (principal components).
- Each of the new axis is a linear combination of the original p variables.
- Top k components retain maximal possible variation of the original data (i.e. top k dimensions retain maximal information about A)
- PCA can be used to transform A into X which is $n \times k \quad (k \leq p)$ in which k dimensions are uncorrelated.

WHAT PCA DOES?

PCA transforms the variables of the dataset to find the most ‘informative’/‘relevant’ (principal) projections

Some dimensions are more important than others, hence PCA can be used to reduce data dimensionality.

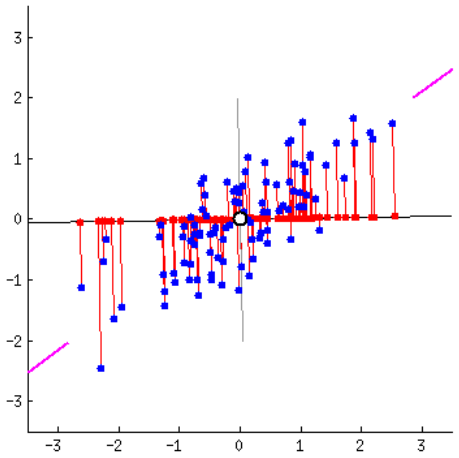
In other words: PCA projects data onto a lower dimensional space



WHAT PCA DOES?

PCA - Synthetic example.ipynb

- Principal components represent the directions of the data that explain a maximal amount of variance.
  - The lines that capture most information of the data
- Think of principal components as new axes that provide the best angle to see and evaluate the data, so that the differences between the observations are better visible.

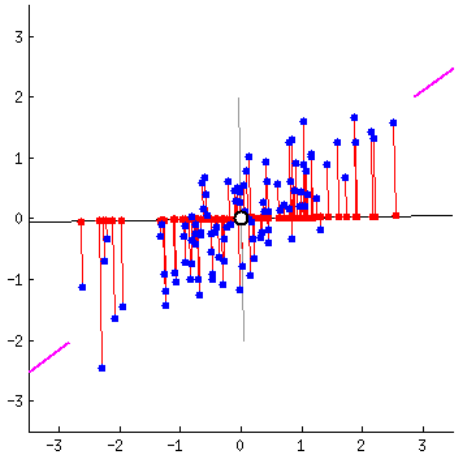


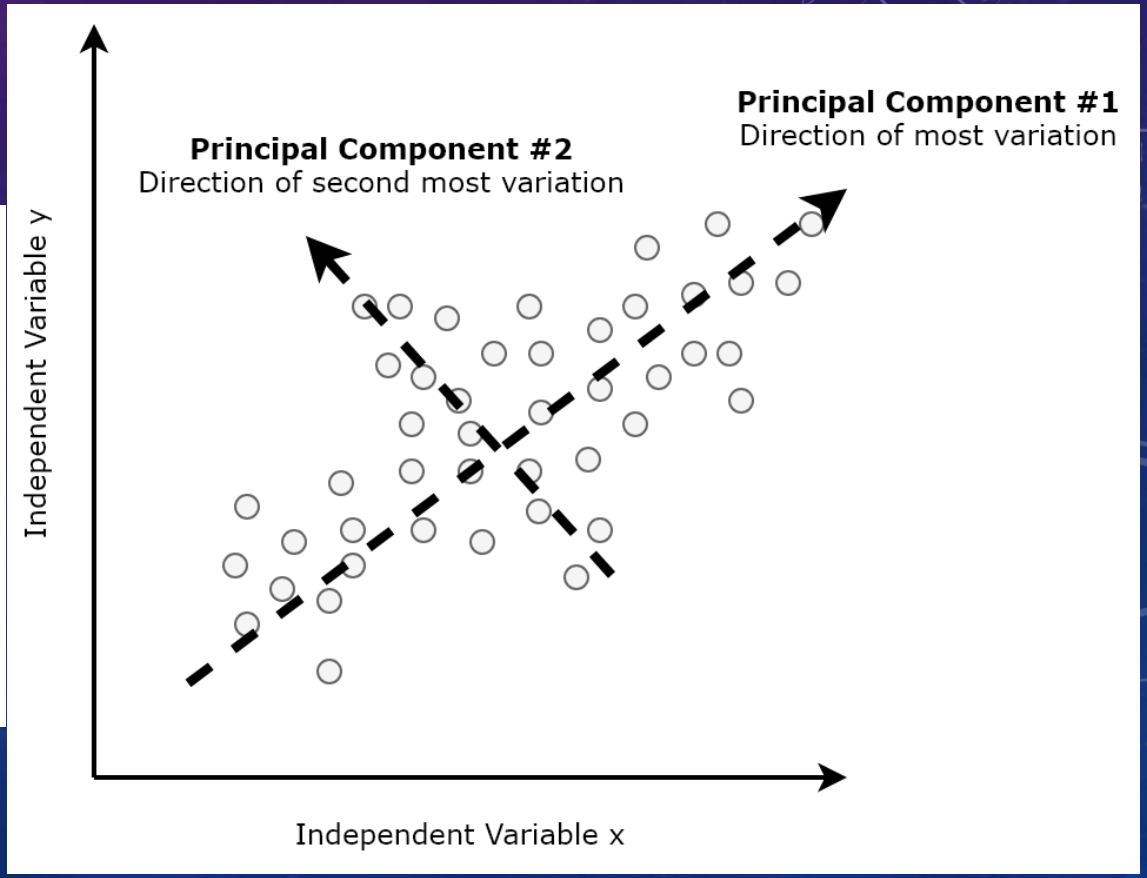
- The 1st component is approximately the line that matches the purple marks because it goes through the origin and it's the line in which the projection of the points (red dots) is the most spread out.
  - The line that maximizes the variance.
- The 2nd principal component is calculated in the same way, with the condition that it is uncorrelated with the first principal component and that it accounts for the next highest variance.

https://builtin.com/data-science/step-step-explanation-principal-component-analysis

SUMMARY OF UNSUPERVISED LEARNING: DIMENSIONALITY REDUCTION WITH PCA

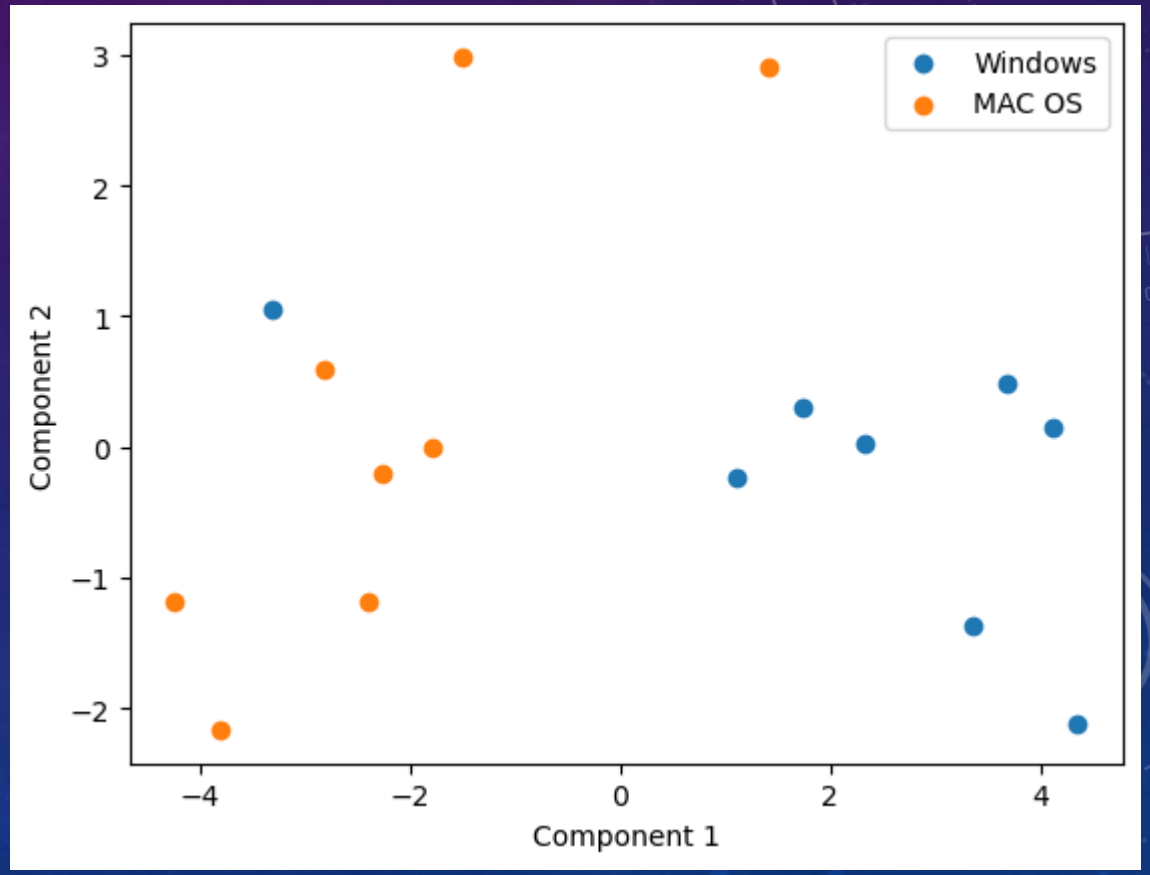
- The curse of dimensionality





SUMMARY OF UNSUPERVISED LEARNING: DIMENSIONALITY REDUCTION WITH PCA

Participant	Price	Software	Aesthetics	Brand	OS
P1	6	5	3	4	0
P2	7	3	2	2	0
P3	6	4	4	5	0
P4	5	7	1	3	0
P5	7	7	5	5	1
P6	6	4	2	3	0
P7	5	7	2	1	0
P8	6	5	4	4	0
P9	3	5	6	7	1
P10	1	3	7	5	1
P11	2	6	6	7	0
P12	5	7	7	6	1
P13	2	4	5	6	1
P14	3	5	6	5	1
P15	1	6	5	5	1
P16	2	3	7	7	1



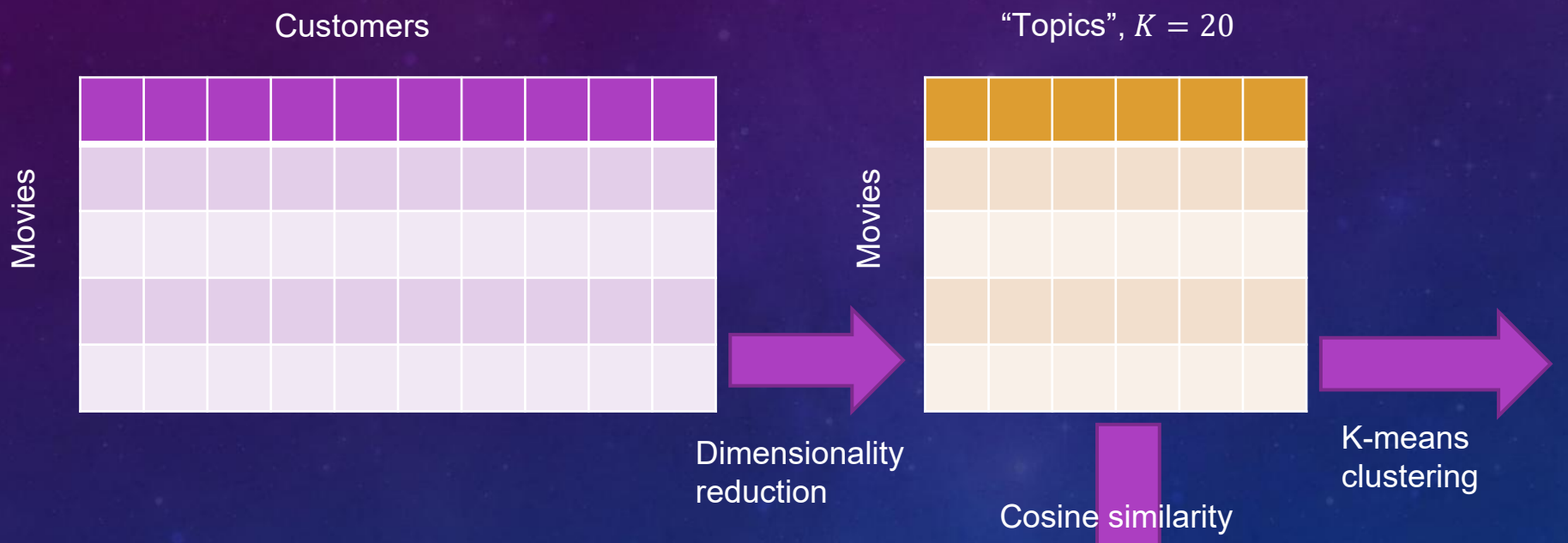
PCA - Synthetic example.ipynb

CUSTOMER-BASED MOVIE CLUSTERING

1. Customers rank movies

2. Dimensionality Reduction

3. Clustering



Recommender system:
Movies you'll like if you like that movie

Clustering_Movies.ipynb

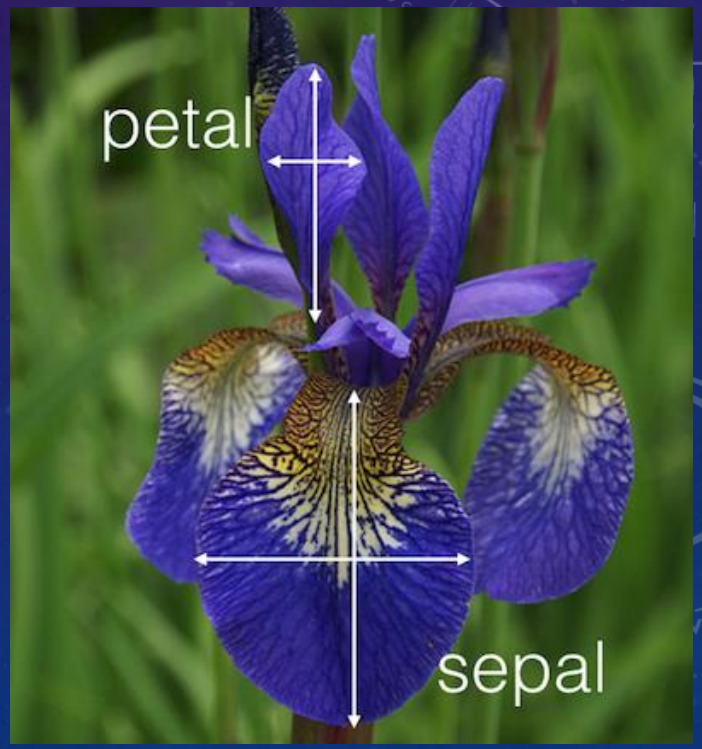
IRIS DATASET

Clustering_and_dimensionality_reduction_Part_B.ipynb

Iris - Dimentionality Reduction

- The data set contains **3 classes** of **50 instances each**, where each class refers to a type of iris plant. One class is linearly separable from the other 2; the latter are NOT linearly separable from each other.
- Predicted attribute: class of iris plant.

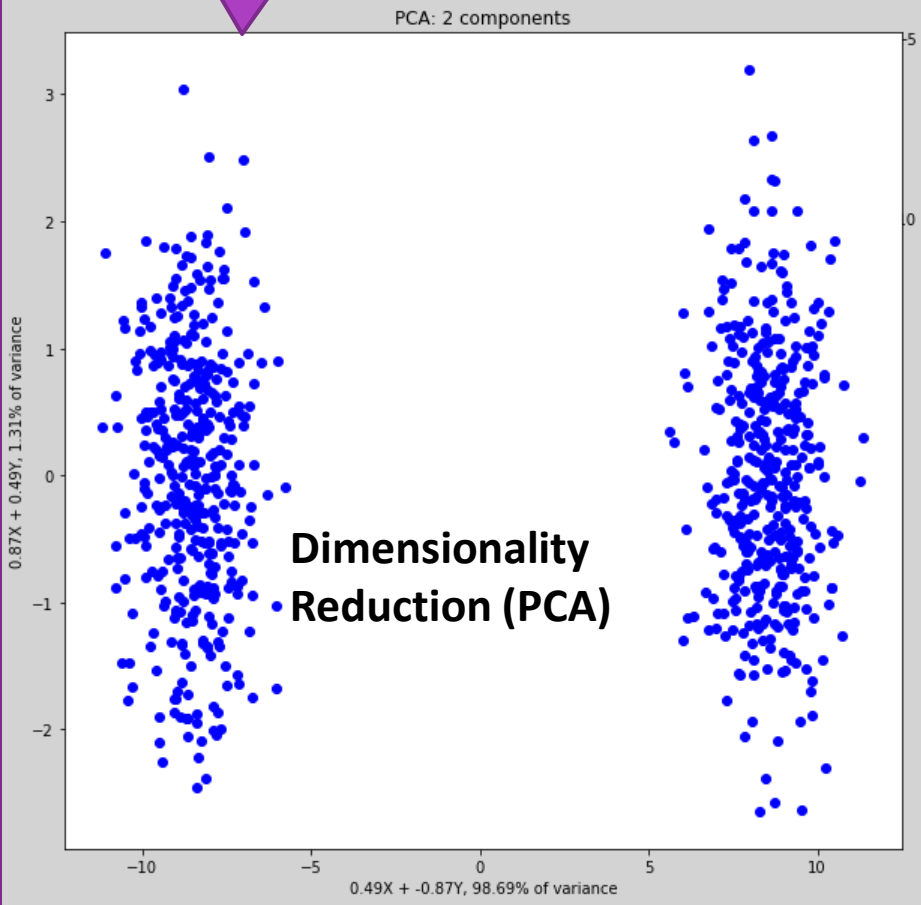
Dataset attributes
1. sepal length in cm
2. sepal width in cm
3. petal length in cm
4. petal width in cm
5. class:
-- Iris Setosa
-- Iris Versicolour
-- Iris Virginica

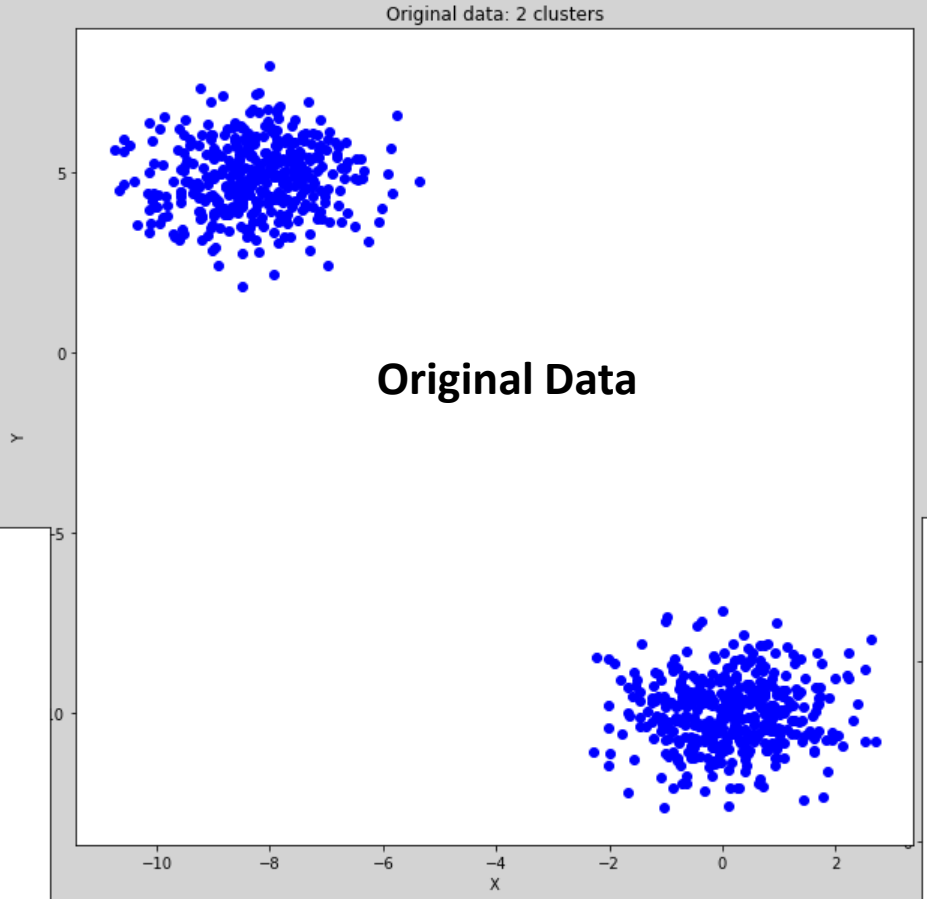


Data Set Characteristics:	Multivariate	**Number of Instances:**	150
Attribute Characteristics:	Real	**Number of Attributes:**	4
Associated Tasks:	Classification	**Missing Values?**	No

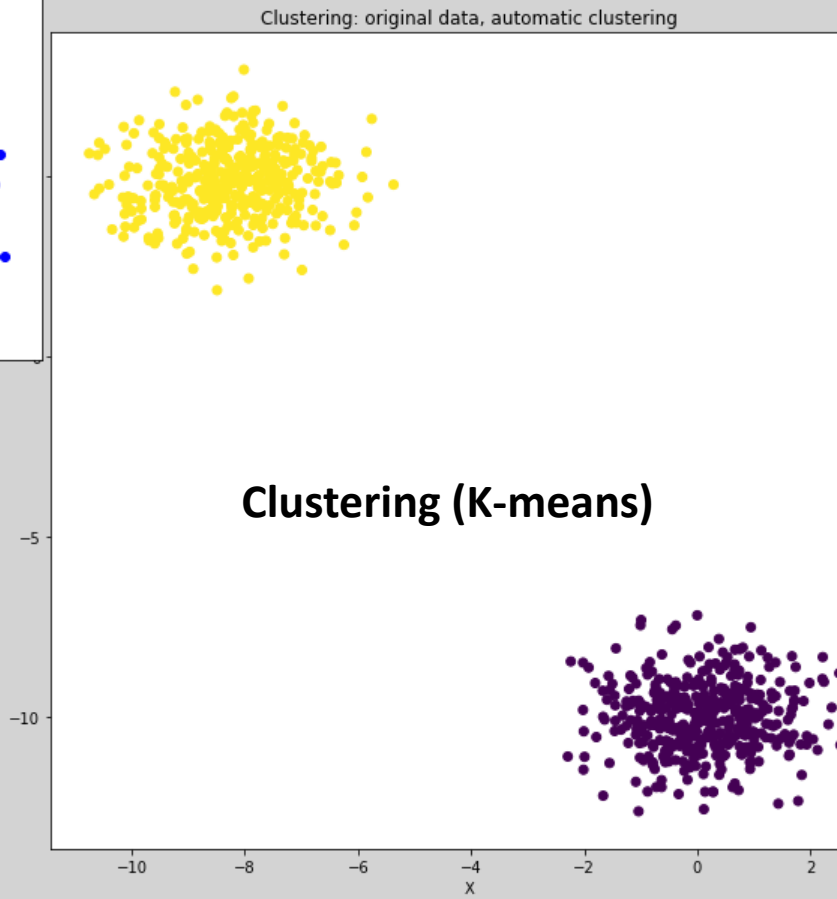
Ronald Fisher (1936), *Use of multiple measurements in taxonomic problems*

Find “appropriate” coordinates



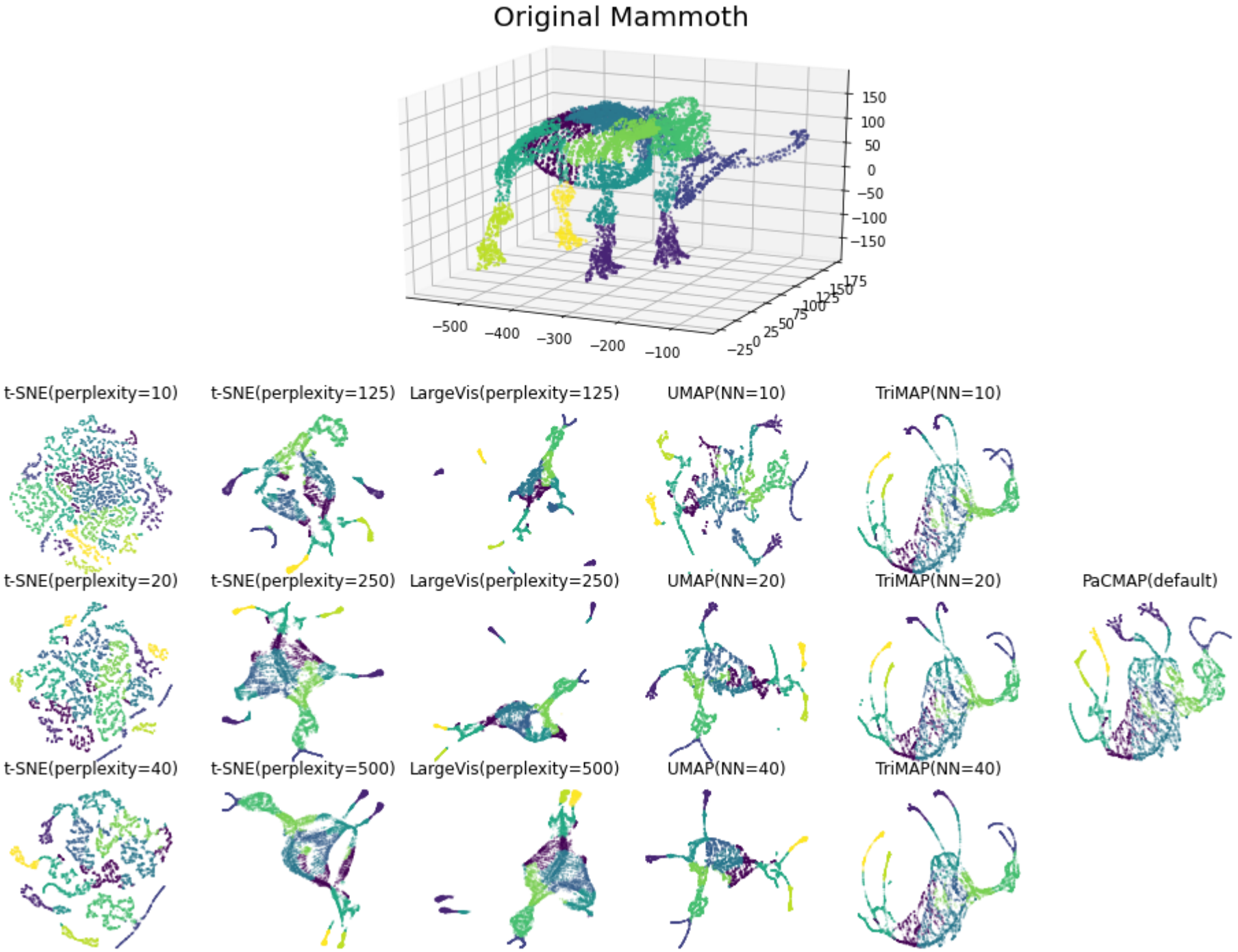


Group points by similarity



DIMENSIONALITY REDUCTION FOR VISUALIZATION

Method Name	Features	Implementation
Principal Component Analysis	very fast doesn't produce insightful visualizations	pca = sklearn.decomposition.PCA(*n_components=2)* *X2d = Pca.fit_transform(X)*
T-distributed Stochastic Neighbor Embedding (t-SNE)	Non-linear, graph-based. Key parameters: perplexity & number of neighbors These parameters balance global and local structure Very popular and tends to work very well Sklearn implementation is slow, even the new multicore version. Other implementations are faster	sklearn.manifold.TSNE openTSNE - https://opentsne.readthedocs.io/en/latest/ MulticoreTSNE - https://github.com/DmitryUlyanov/Multicore-TSNE
Uniform Manifold Approximation and Projection (umap)	A good alternative to t-SNE Tends to be faster than t-SNE	https://umap-learn.readthedocs.io/en/latest/ pip install umap-learn
Pairwise Controlled Manifold Approximation (PacMAP)	A new algorithm. First captures global structure, and then fine-tines local structure.	https://github.com/YingfanWang/PaCMAP pip install pacmap



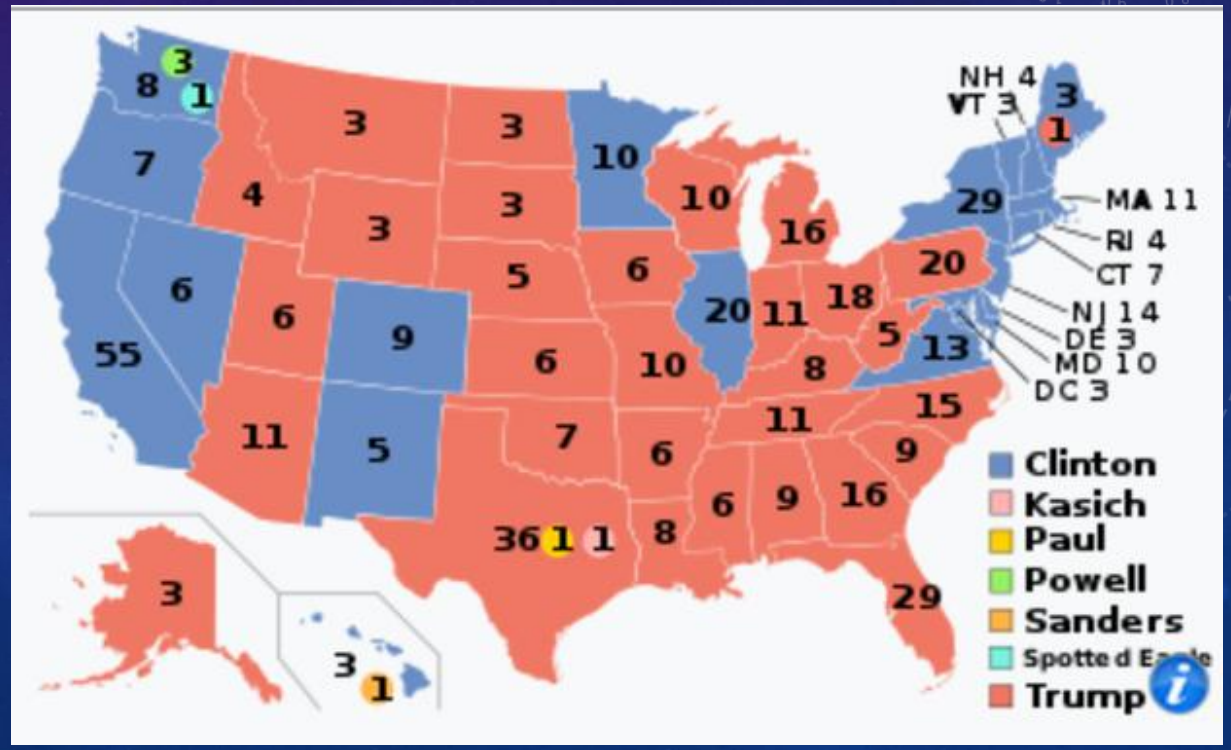
Wang, Yingfan, Haiyang Huang, Cynthia Rudin, and Yaron Shaposhnik. "Understanding How Dimension Reduction Tools Work: An Empirical Approach to Deciphering t-SNE, UMAP, TriMap, and PaCMAP for Data Visualization." *J. Mach. Learn. Res.* 22, no. 201 (2021): 1-73.

US PRESIDENTIAL ELECTIONS 2016: POSTDICTION

Use state-level public data to explain vote patterns

PCA\US_Elections_Dimensionality_Reduction.ipynb

Variable	Source
Income	List of US States by Income
Education	List of US States by Educational Attainment
Religiosity	List of U.S. states by religiosity
Life Expectancy	List of US States by Life Expectancy
Urbanization	Urbanization in the United States



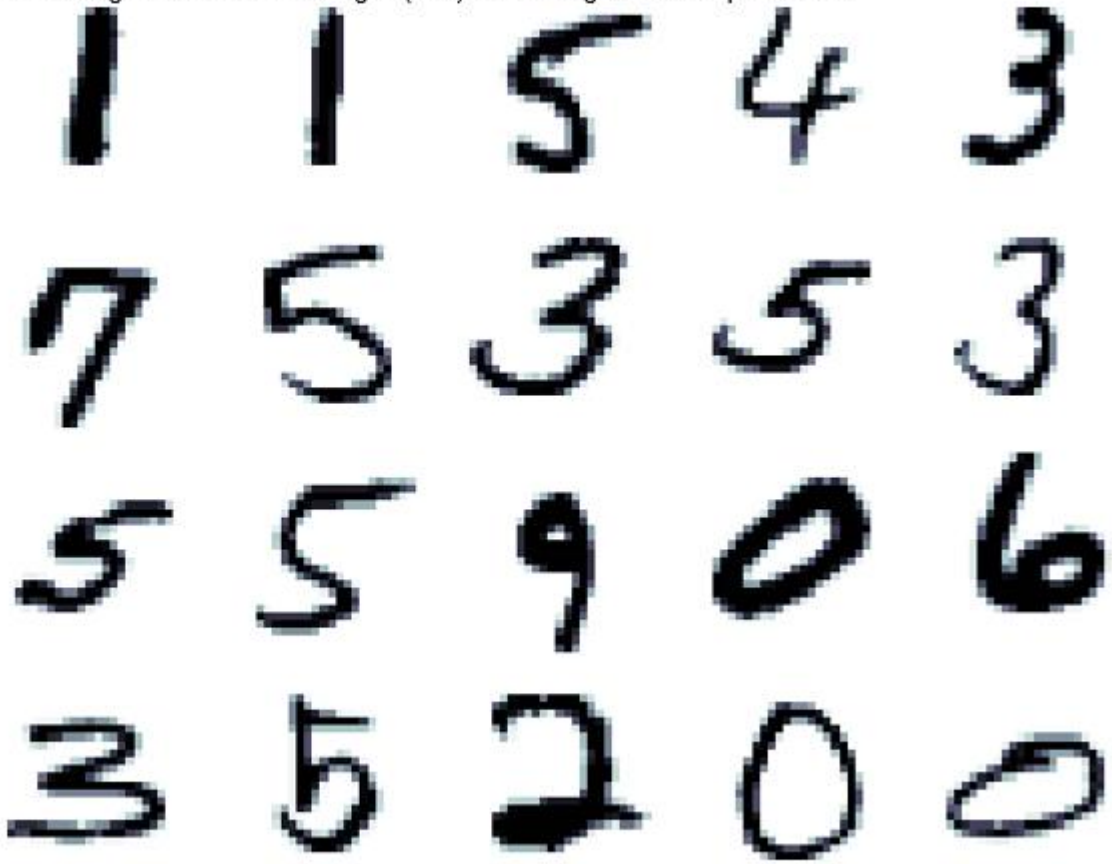
Outcome: Electoral Vote

KNESSET MEMBERS VOTING PATTERNS

Dimentionality_Reduction_Knesset_Votings.ipynb

Clustering Handwritten Digits

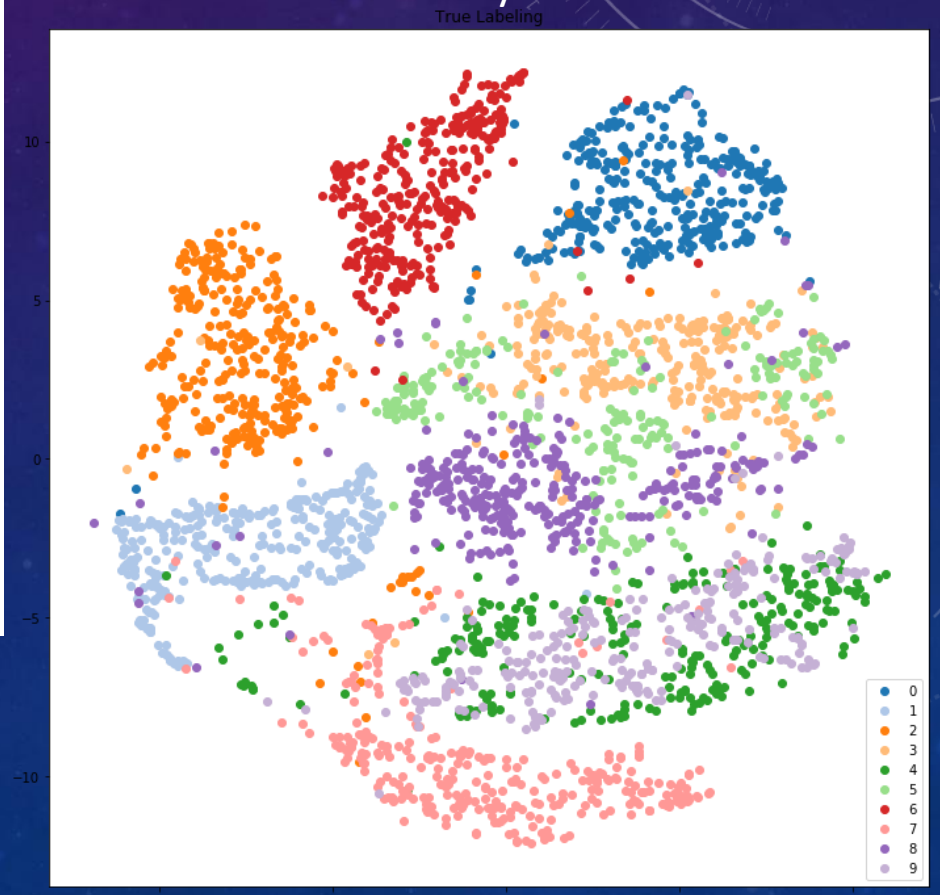
MNIST is a dataset of 70,000 labeled images of handwritten digits (0..9) Each image is 28x28 pixels size



Can we segment digits into clusters?

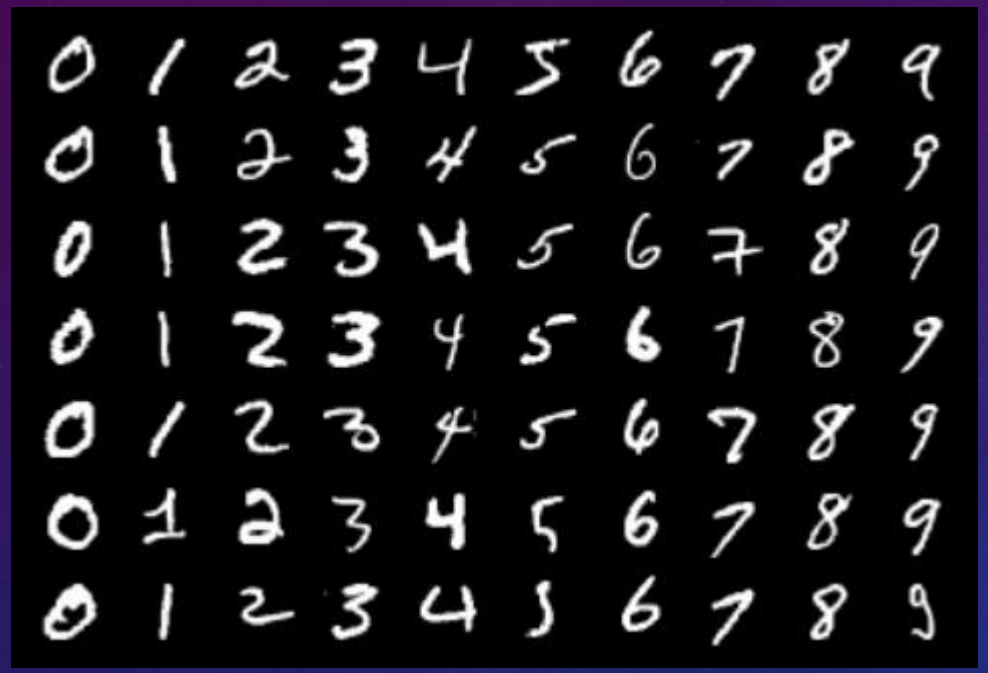
Clustering_and_dimensionality_reduction_Part_C.ipynb

Dimensionality reduction

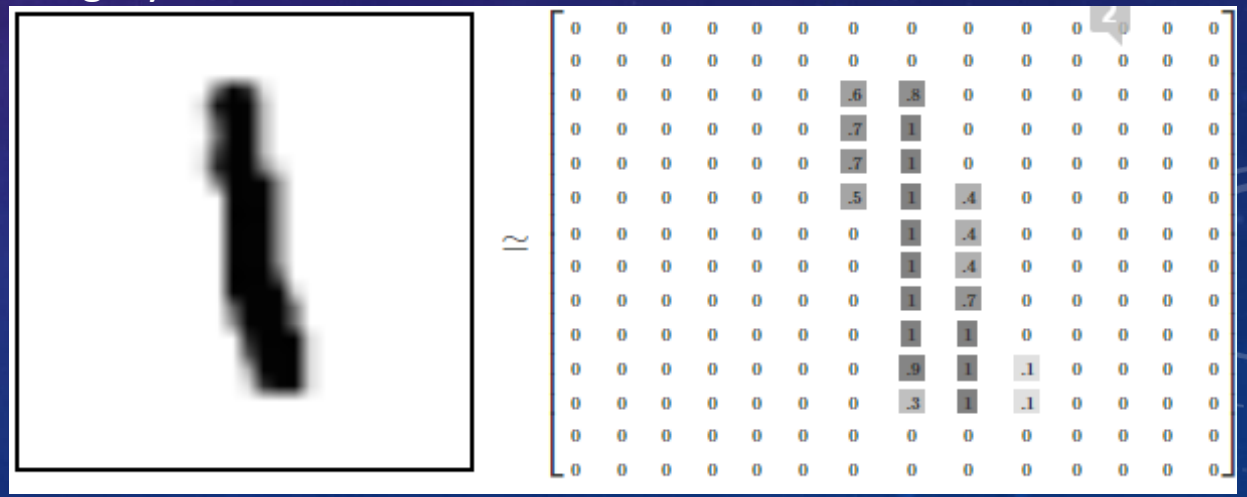


DIMENSIONALITY REDUCTION: HANDWRITTEN DIGITS

Clustering_and_dimensionality_reduction_Part_C.ipynb



- 70,000 images of handwritten digits
- by 500 writers
- Labeled
- 28x28 pixels (784 pixels)
- 256 grey levels



Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. "Gradient-based learning applied to document recognition." *Proceedings of the IEEE*, 86(11):2278-2324, November 1998

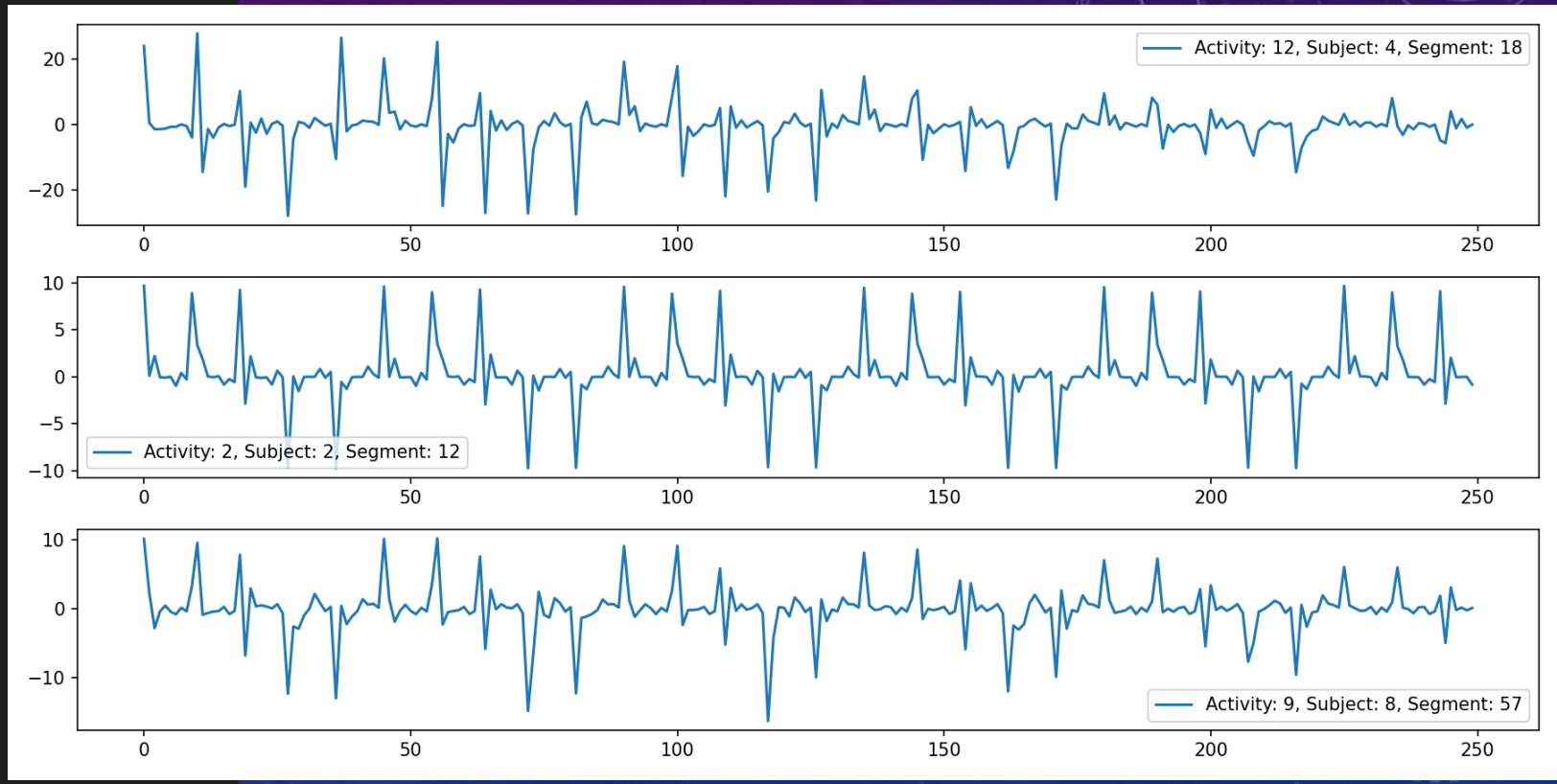
Data

- Daily and Sports Activities Dataset link
- **Subjects:** 8 (4 male, 4 female)
- **Duration:** 5625 samples (5 min) of activity records per subject
- **Data size** 9120 time series : 19 activities × 8 subjects × 60 segments per subject

Activity ID	Activity Description
A1	Sitting
A2	Standing
A3	Lying on back
A4	Lying on right side
A5	Ascending stairs
A6	Descending stairs
A7	Standing in an elevator still
A8	Moving around in an elevator
A9	Walking in a parking lot
A10	Walking on treadmill (4 km/h, flat)
A11	Walking on treadmill (4 km/h, 15° incline)
A12	Running on treadmill (8 km/h)
A13	Exercising on a stepper
A14	Exercising on a cross trainer
A15	Cycling on exercise bike (horizontal)
A16	Cycling on exercise bike (vertical)
A17	Rowing
A18	Jumping
A19	Playing basketball

DIMENSIONALITY REDUCTION OF SPORT ACTIVITY

sports_tsne.ipynb

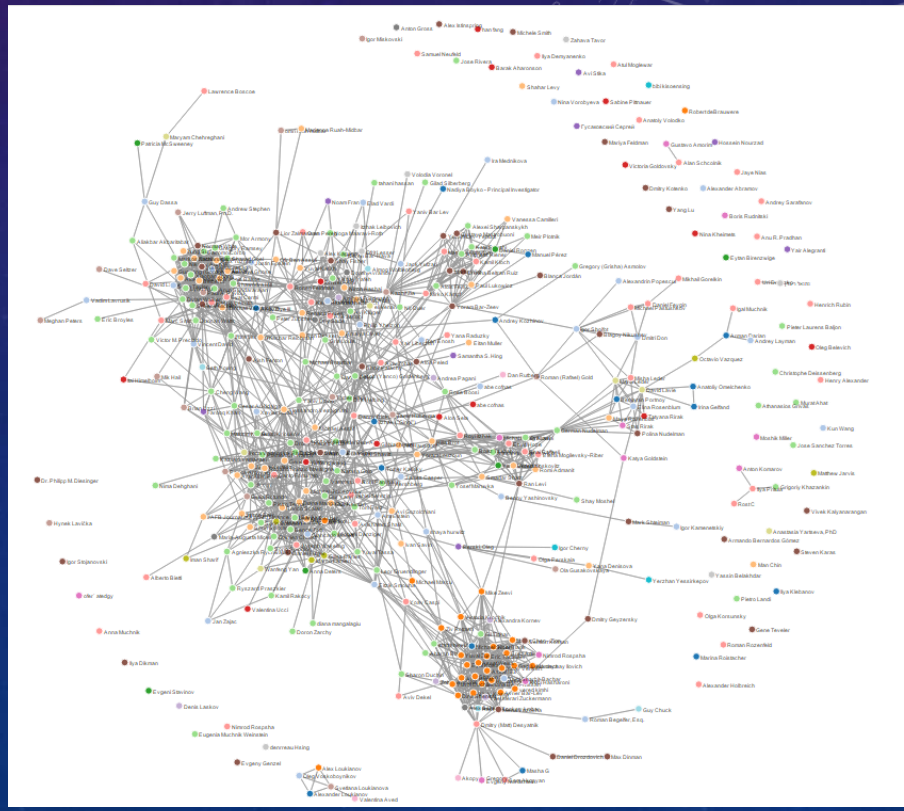


https://archive.ics.uci.edu/dataset/seitivitca+strops+dna+yliad/256

METHODS OF CLUSTERING: NETWORKS AND COMMUNITY DETECTION

- Graph Partitioning or Community detection is a task aimed to break a network down into a ***set of tightly-knit regions***, with ***sparser interconnections*** between the regions.

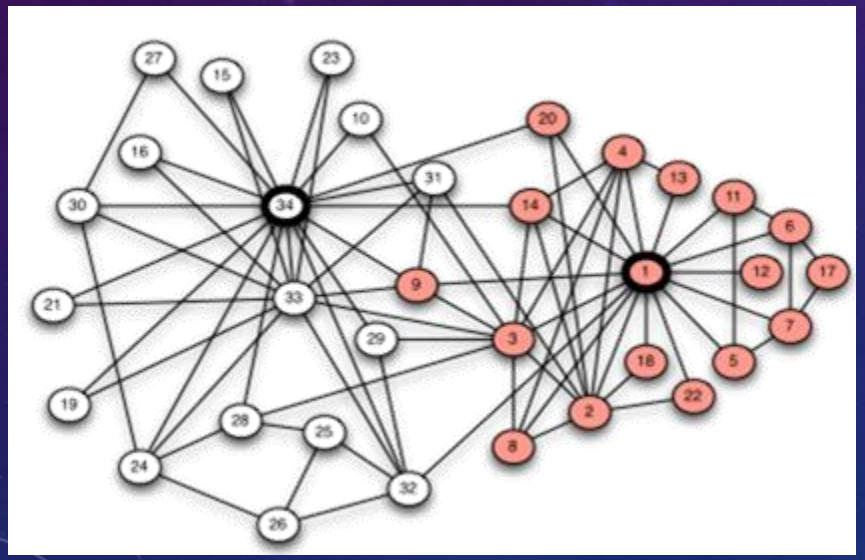
- One could represent many datasets as a (weighted) network
- Network or graph is a set of interconnected nodes
- Mathematically: Graph is an **ordered pair of nodes and edges**



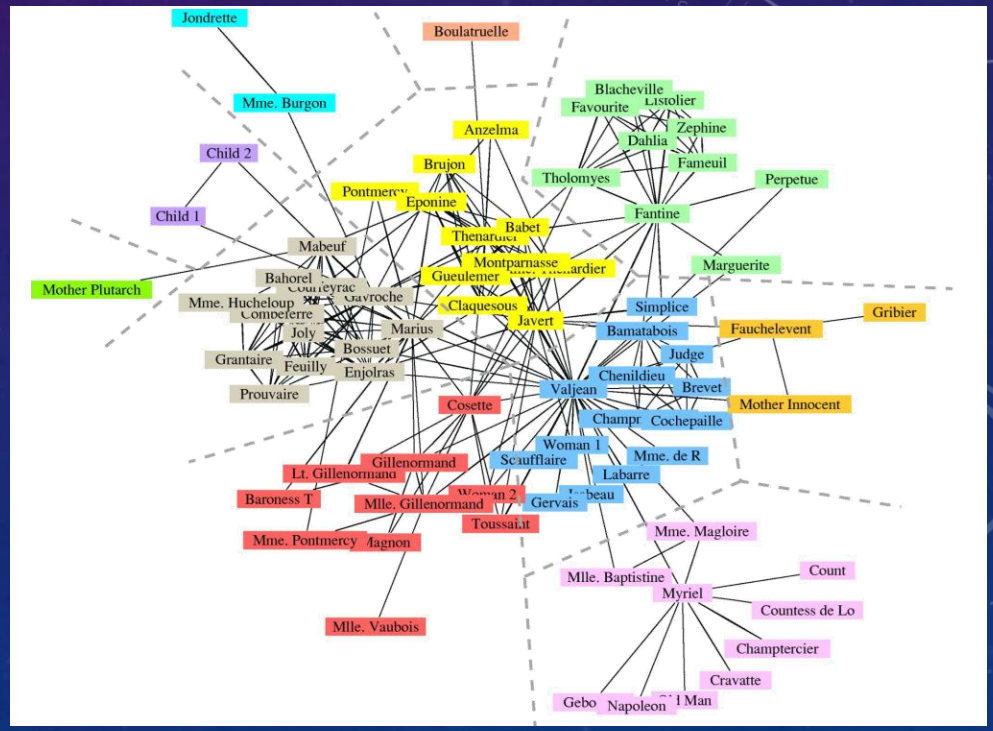
Graph $G = (V, E)$
V – set of nodes/vertices, $i = 1, \ldots n$
E – set of links/edges, $(i, j), m$

METHODS OF CLUSTERING: NETWORKS AND COMMUNITY DETECTION

Graph Partitioning or Community detection is a task aimed to break a network down into a ***set of tightly-knit regions***, with ***sparser interconnections*** between the regions.



Zachary Karate club



The network of interactions between major characters in the novel *Les Mise´rables* by Victor Hugo

OBTAINING NETWORK FROM A LIST OF OBSERVATIONS

1. Using a distance function.
    - Nodes: each observation (e.g. a person, a customer, a product) is represented by a node
    - Edges between pairs of nodes have weights (Inverse) proportional to similarity (distance) between the corresponding nodes
    - Complication: this process yields fully connected network (every node is connected to all other nodes). This challenges community detection algorithms.
      Use threshold to drop weaker links and emphasize clusters.
2. Correlation matrix
    - In the cases when a node is represented by time series or relationship to many other items (e.g. cluster titles, base overlap of content with many documents), one can generate a correlation matrix
    - Convert correlation matrix to graph.

GRAPH PARTITIONING: MODULARITY

Modularity:

$$Q = (number\ of\ edges\ within\ groups)\ - (expected\ number\ within\ groups)$$

All edges in one group: $Q = 0$ (trivial division)

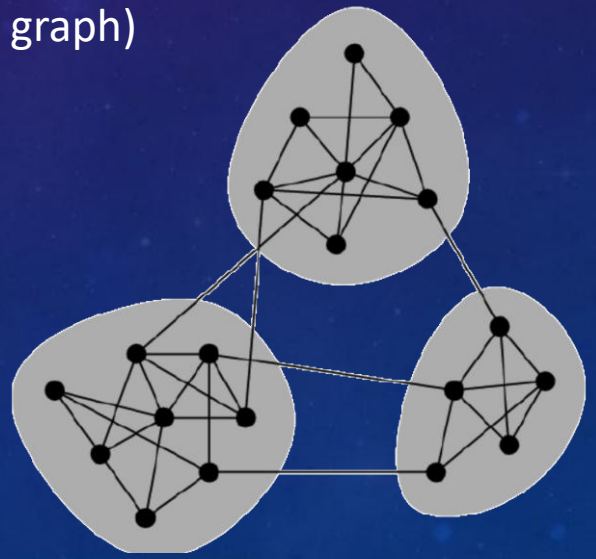
k_i- degree of the node i

$$A_{i,j} = \begin{cases} 1, there\ is\ an\ edge\ between\ i\ \&\ j \\ 0, there\ is\ \boldsymbol{no}\ edge\ between\ i\ \&\ j \end{cases}$$

$P_{i,j}$ - Probability of an edge between randomly chosen i & j (in randomized graph)

$$P_{i,j} \approx \frac{k_i * k_j}{2 * m}$$

$$\boldsymbol{Q = \sum_{groups} (A_{ij} - P_{i,j} |\ \ i,j\ in\ the\ same\ groups)}$$



COMMUNITY "NULL MODEL": THE CONFIGURATION MODEL

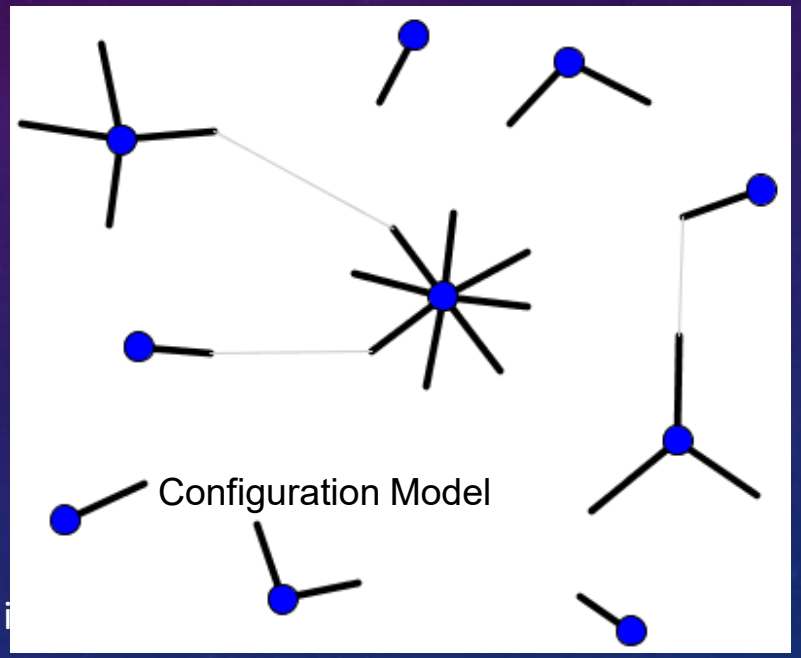
Probability of an edge between randomly chosen i & j:

$$P_{i,j} \approx \frac{k_i * k_j}{2 * m}$$

where $\mathrm{m} = |\{E\}|$, is the number of edges in the graph

Why?

- Cut each edge into two halves (stubs)
  - There would be $\sum k_i = 2m$ stubs
- Reconnect stubs randomly
- The resulting network preserves the degree of each node but "kills" the communi structure.
- $P_{i,j} = \frac{Is\ the\ edge\ between\ i\ \&\ j\ present?}{total\ number\ of\ possible\ rewirings} \approx \frac{k_i * k_j}{2 * m_i}$



$$\boldsymbol{Q = \sum_{groups} \left(A_{ij} - \frac{k_i * k_j}{2*\mathrm{m}} \middle| \quad i,j\ in\ the\ same\ groups \right)}$$

GRAPH PARTITIONING: MODULARITY

Modularity: $Q = (\#\ of\ edges\ within\ groups)\ -(expected\ \#of\ edges\ within\ groups)$

$$\boldsymbol{Q} = \sum_{\boldsymbol{groups}} \left(\boldsymbol{A_{ij}} - \frac{\boldsymbol{k_i} * \boldsymbol{k_j}}{2*\mathrm{m}} \mid \ \boldsymbol{i,j\ in\ the\ same\ groups} \right)$$

Modularity : $-1 \leq Q \leq 1$

Larger values of Q indicate stronger community structure

Some community detection algorithms maximize Q directly.

LABEL PROPAGATION ALGORITHM

- Initialize

  Enumerate nodes with running labels (e.g. $i's$ node would be assigned a label i).

- Run
  - Iterate over the nodes in random order
    - Update label of the current node with the label of the majority of its neighbors
  - Stop when no node changed its label during the iteration

LABEL PROPAGATION ALGORITHM

LABEL PROPAGATION ALGORITHM

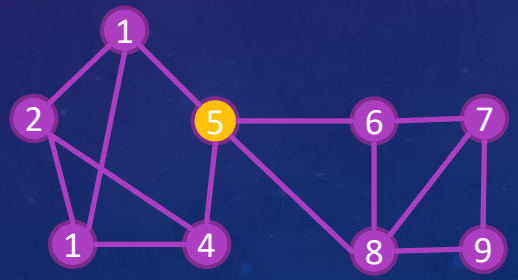
0. Initialize nodes with labels

1. Select a random node (without replacement)

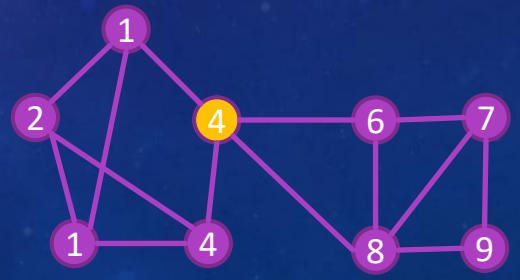
2. Update the node's label with the label of the majority of the node's neighbors

#1

3. Move the next (random) node

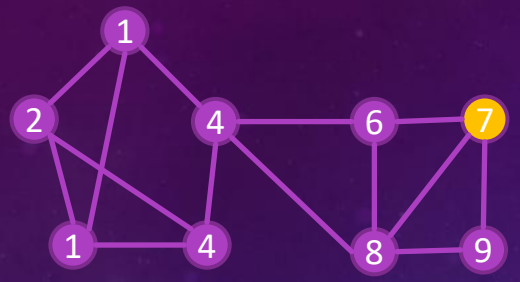


4. Update node label

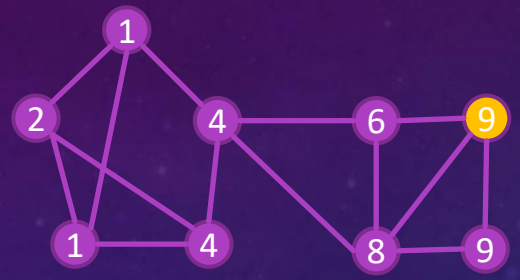


#2

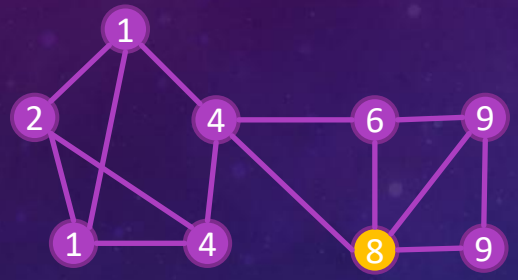
LABEL PROPAGATION ALGORITHM



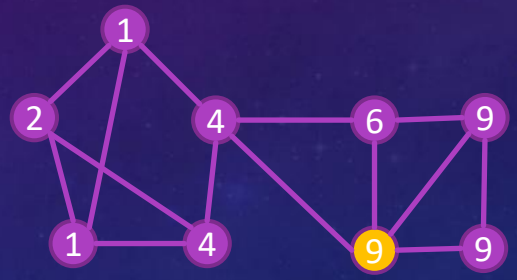
#3



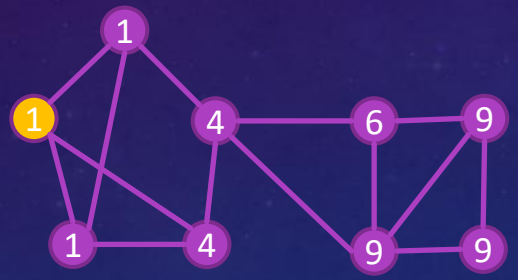
#4



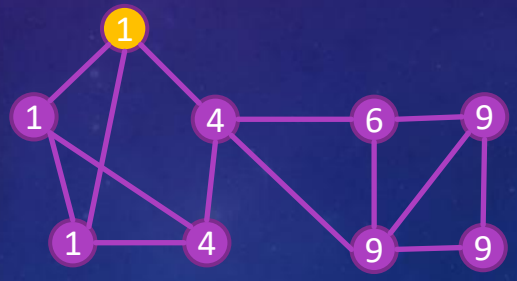
#5



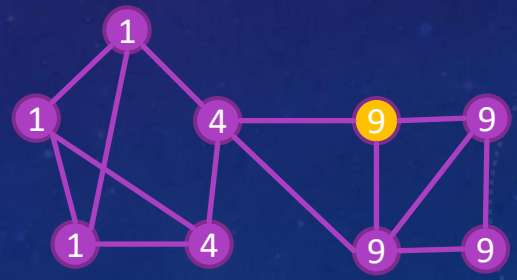
#6



#7

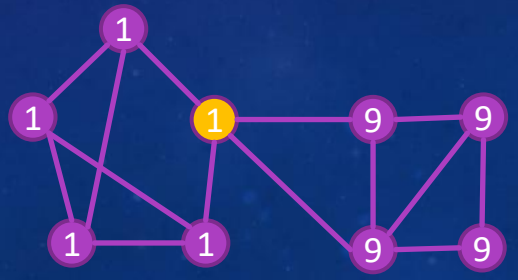


#8 – 9



#10

#11 – 12



LABEL PROPAGATION ALGORITHM AND PROBABILISTIC COMMUNITY STRUCTURE

- Label Propagation Algorithm is stochastic. Every run may yield different results.
- Nodes residing in the same community are very likely to be assigned to the same community
- Assignment of the nodes not embedding into communities may vary.
- Label Propagation Algorithm can be executed multiple times to measure the probability that any two nodes are in the same community.
  - This will generate a probabilistic model of community structure
  - Detect boundary nodes

SEMI-SUPERVISED VERSION OF THE LABEL PROPAGATION ALGORITHM

- Assume that few of the node labels are known
- Initiate Label Propagation Algorithm with:
  - Known labels for existing nodes
  - Sequential (arbitrary) labels for the remaining nodes
- Run Label Propagation Algorithm without updating labels of the labeled nodes.

- Besides inferring labels for unlabeled nodes, the algorithm can:
  - Detect communities for which no label was collected (i.e. can guide data collection)
  - Determine nodes with mixed / unclear labels.

Enron_network_analysis.ipynb

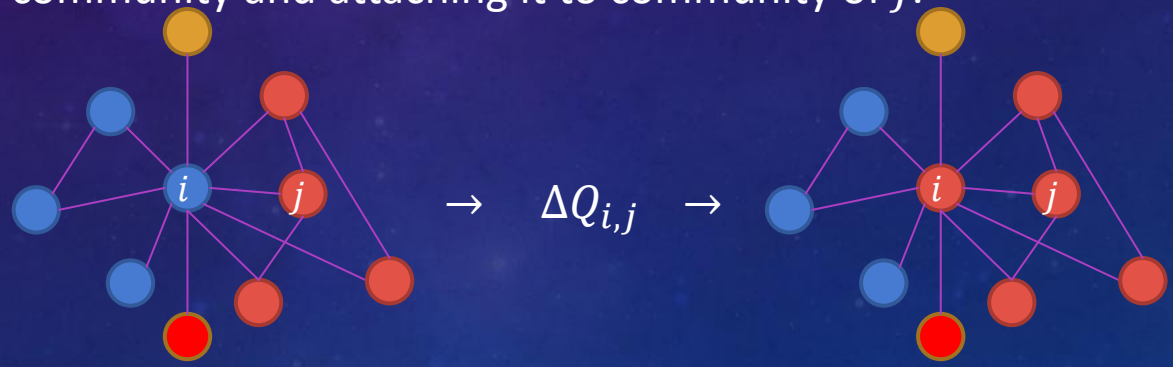
LOUVAIN COMMUNITY DETECTION

- The method directly maximizes network modularity Q
- The algorithm is greedy and is very fast ($O\left(n \cdot log(n)\right)$).
- Extremely efficient – can operate on very large graphs.

Blondel, Vincent D., et al. "Fast unfolding of communities in large networks." *Journal of statistical mechanics: theory and experiment* 2008.10 (2008): P10008.

LOUVAIN COMMUNITY DETECTION

- Find small communities by optimizing modularity locally on all nodes:
  - Assign each node to an individual community.
  - For each node i and its neighbor j, compute the change in modularity $\Delta Q_{i,j}$ due to removal of i from its community and attaching it to community of j.



- Place i in the community of the node j for which $\Delta Q_{i,j}$ is maximal
- Group each community into one node. Preserve links between communities.
- Iterate until only one community left.

Community Detection.ipynb