CLUSTERING ALGORITHMS

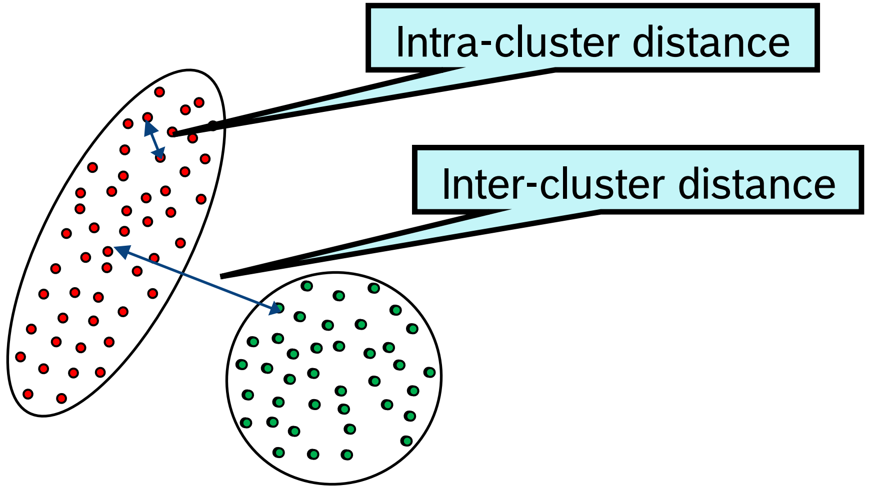
SONAL KUMARI,
CR/PJ-AI

BOSCH

Clustering Techniques
Outline

Clustering Techniques
What is Clustering?

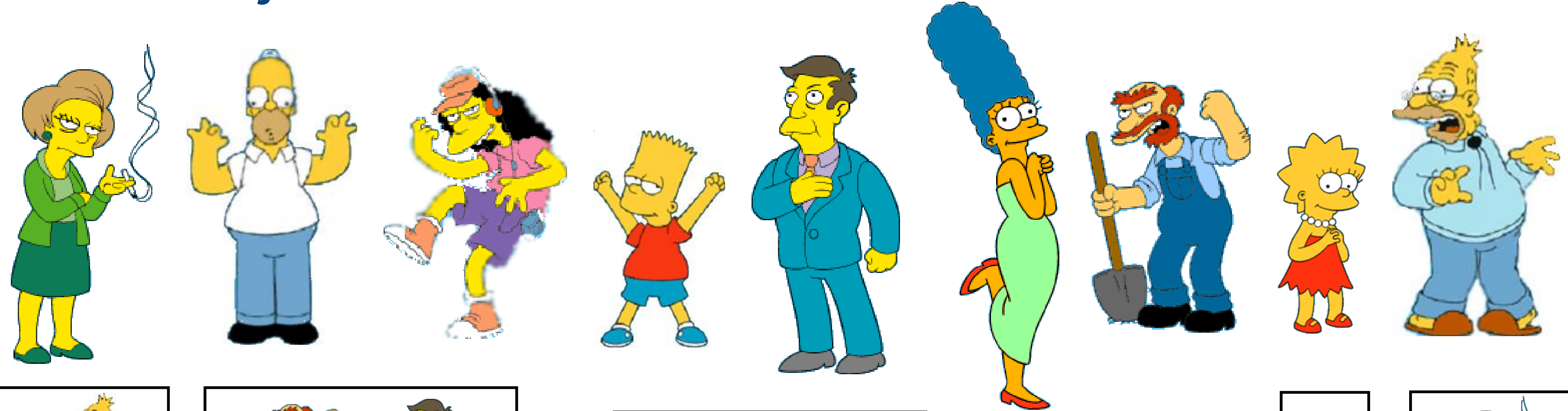
- Unsupervised learning
  - No a priori knowledge about data (class-label is unknown)
  - Finding pattern or structure in the given data (data exploration)
  - Find class-label and number of classes from data
- Grouping similar objects together
  - High intra-class similarity (within a cluster)
  - Low inter-class similarity (between different clusters)
- Clustering results depends on similarity
- How to define similarity?
  - Expressed through a distance metric
- Distance metric:
  - Symmetry: $d(x,y)=d(y,x)$
  - Positivity: $d(x,y)\geq 0$
  - Triangle inequality: $d(x,y)\leq d(x,z)+d(z,y)$

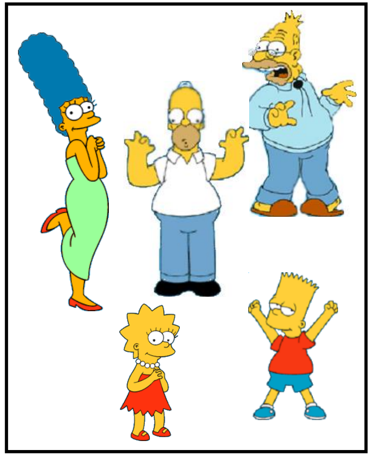


Cluster-1

Cluster-2

Clustering Techniques
How these objects should be clustered?

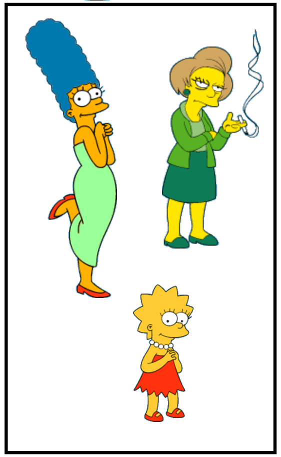




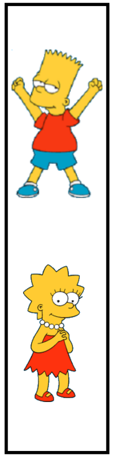
Simpson's Family

School Employees

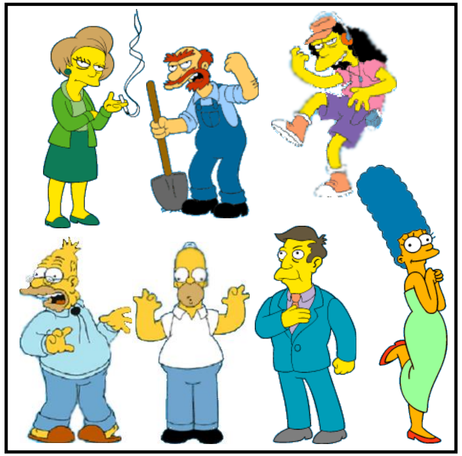
Males



Females



Kids



Adults

Clustering Techniques
Critical Steps in Clustering

1. Which feature should be selected?
   - Depends on the use-case
2. Pre-processing
   - Data cleaning, Binning, Data reduction, Normalization (z-transformation, mean-adjustment, etc.)
   - Variable weight adjustment: depends on selected features [optional]
3. How to select ***distance metric*** for similarity/dissimilarity?
   - Depends on variable type, use-case, choice of clustering algorithm, etc.
4. Choice of clustering algorithm?
   - Depends on variable type (binary, continuous, categorical, mixed, etc.)
   - Presence/absence of Noise or outlier, dimensionality of data
   - Overlapping (fuzzy/soft clustering, probabilistic clustering) or disjoint/exclusive groups
   - Also depends on use-case

Feature Selection

↓

Pre-processing

↓

Similarity Metric

↓

Clustering Algorithm

↓

Clustering Techniques
Clustering Applications

- ❑ Market analysis: e.g. customer segmentation based on their behaviors
- ❑ Pattern recognition: grouping of houses based on geographical location, etc.
- ❑ Image processing: object detection in an image
- ❑ Text mining: document clustering to improve search recall for search engine
- ❑ Medical field: e.g. identification of gene which is responsible for disease
- ❑ Data reduction: summarization & compression
- ❑ etc.

Clustering Techniques
Distance Metrics (1/2)

- Euclidean: $d_E(x,y) = \sqrt{\sum_i^n (x_i - y_i)^2}$
  - Scale variant, Sensitive to data dimensionality: Normalization (scaling) can solve this issue
- Squared Euclidean: $d_E{}^2(x,y) = \sum_i^n (x_i - y_i)^2$
  - Tends to give more weight to outliers in comparison to Euclidean
- Standard Euclidean: $d_{ES}(x,y) = \sqrt{\sum_i^n \frac{1}{S_i^2}(x_i - y_i)^2}$ where S_i^2 is i-dimensional variance
- Manhattan (City-block): $d_{CB}(x,y) = \sum_i^n |x_i - y_i|$
  - Sensitive to outliers but comparatively less in comparison to Euclidean
- Minkowski (generalization of Euclidean and Manhattan): $dis = \sqrt[m]{\sum_i^n (x_i - y_i)^m}$
- Chebyshev: $d_C(x,y) = \max_i |x_i - y_i|$, very sensitive to outliers & noise
- Jaccard (used for binary data): $dis_J = 1 - \frac{\sum_i \min(x_i, y_i)}{\sum_i \max(x_i, y_i)}$
- Hamming (used for binary data): $dis_H = \sum_i^n |x_i - y_i|$, *x* and *y* are two strings

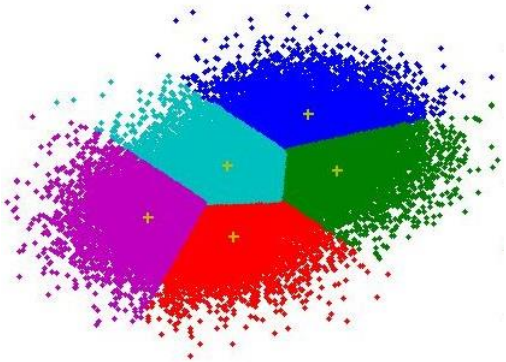
Clustering Techniques
Distance Metrics (2/2)

- Mahalanobis: $d_M(x,y) = \sqrt{(x_i - y_i)^T C^{-1}(x_i - y_i)}$ where C is covariance matrix
  - Address the issues of Euclidean distance metrics, takes care of correlated (redundant) feature
- Cosine: $dis_{\cos} = \frac{\sum_i^n x_i y_i}{\sqrt{\sum_i^n x_i^2}}$
  - Only consider angle, not magnitude (rotation invariant) & used for **text** high dimensional data
- Pearson Correlation: $dis_{PC} = 1 - \frac{\sum_i^n (x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_i^n (x_i - \mu_x)^2 \sum_i^n (y_i - \mu_y)^2}}$
  - Scale & shift invariant (mean subtraction), used to find trends or overall shape rather than magnitude,
  - Used for high dimensional data, but not suitable for low dimensional data.
- Chi-square (histogram comparison): $dis_{cs} = \sum_i^n \frac{(x_i - y_i)^2}{(x_i + y_i)}$
- Hellinger distance: to differentiate between two probability distributions, used for skewed data

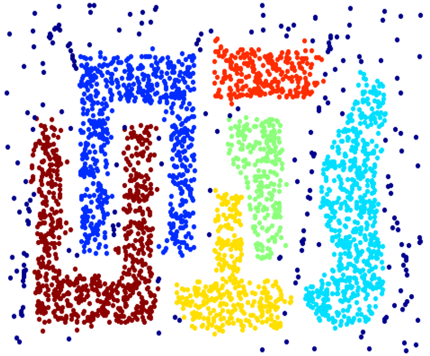
Clustering Techniques
Which Distance Metric is the best?

- Distance metric influence the clustering results
- Euclidean is most widely used for low dimensional continuous data
- Similarly, Pearson is used for high dimensional continuous data
- For categorical variable, hamming distance (similar to Manhattan distance) is used

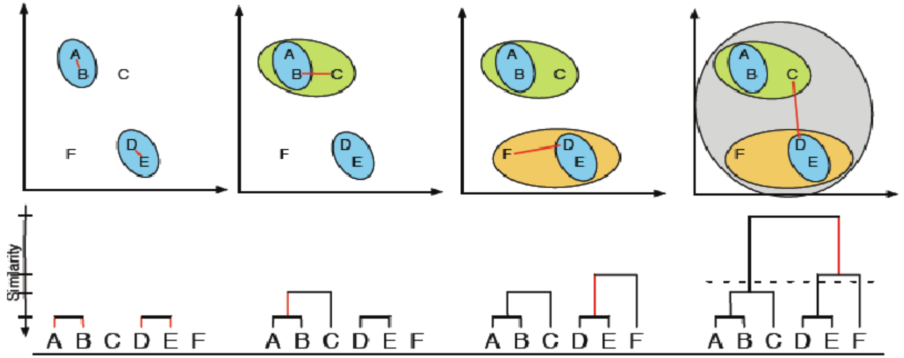
Clustering Techniques
Major type of Clustering Algorithms



Partition Based Clustering

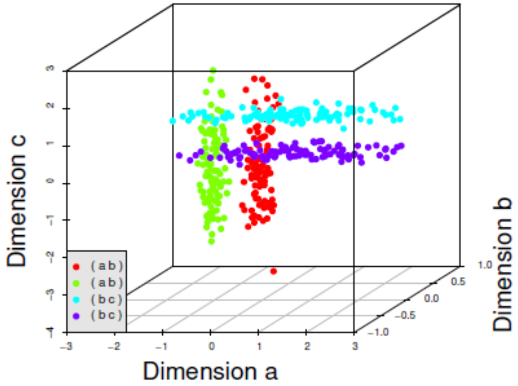


Density Based Clustering

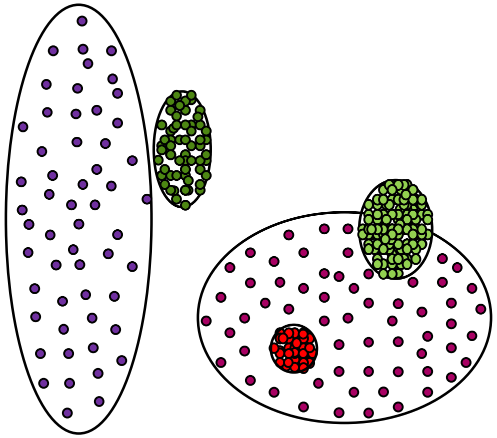


Hierarchical Clustering

Clustering Techniques
Hybrid Clustering Techniques

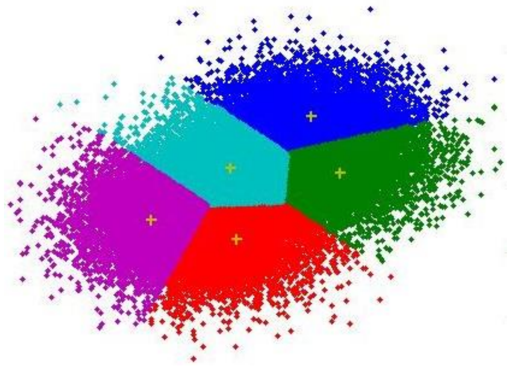


Subspace Based Clustering



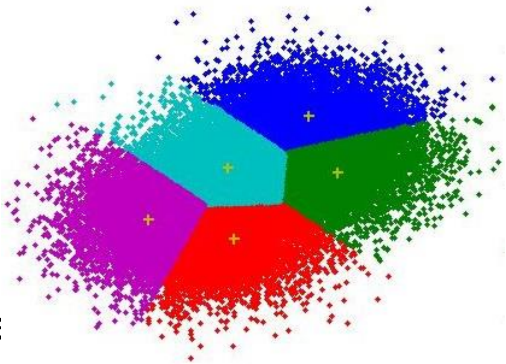
Shared Nearest Neighbor Based Clustering

Clustering Techniques
Partition-based Clustering Algorithms (1/3)



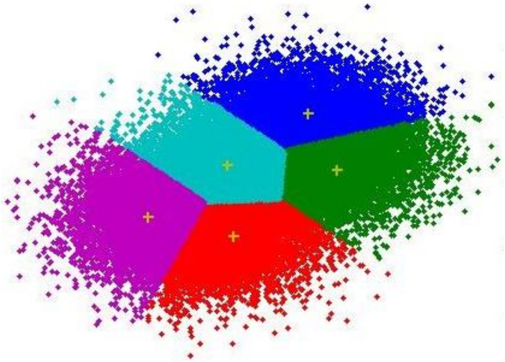
- *k*-means
  - Minimize sum of squared error
  - Time and memory efficient
  - Optimal *k*: knee or elbow-method or, Average Silhouette method (maximize),
  - **Cons:** Converges to local minima, mean is not defined for categorical data , cannot handle noise/outliers, assume features are not correlated (PCA), unable to find non-convex shaped clusters, clustering results depends on initial seed selection
- *k*-medoid or PAM (Partitioning Around Medoids)
  - Similar to *k*-means, but uses medoid as cluster representatives & minimizes sum of dissimilarities
  - Handle noise/outlier better than *k*-means but does not scale for large data

Clustering Techniques
Partition-based Clustering Algorithms (2/3)



- CLARA (Clustering LARge Applications)
  - Select multiple samples, apply PAM on each sample, and give best clustering
  - **Cons:** Biased towards selected sample, because sample may not represent the whole
- CLARANS (Clustering Large Applications based on RANdom Search)
  - Dynamically search in neighbors
- *k*-Modes
  - Uses dissimilarity instead of distance and mode instead of mean
  - Handle categorical data very well
- *k*-prototype (hybrid of *k*-means and *k*-modes)
  - Handle mixed (categorical and numerical) data well

Clustering Techniques
Partition-based Clustering Algorithms (3/3)

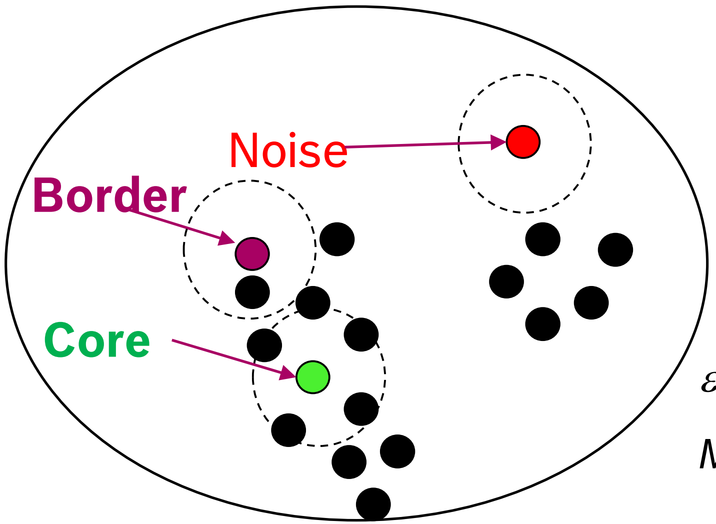


- Nearest Neighbor Clustering
  - Incremental approach and suitable for streaming
  - Uses a threshold to decide if new object is going to merge with existing cluster or
  - **Cons:** Highly order dependent, difficult to decide threshold in advance
- Birch
  - Uses in-memory ***R*-tree** to store points that are being clustered
  - **Increment** approach: insert a point to the existing cluster of *R*-tree if within *threshold* else create new cluster
  - If R-tree size does not fit in the memory, then merge some nearest clusters
  - At the end, keep on merging nearest clusters iteratively until desired number of clusters are found

Clustering Techniques
Density-based Clustering (1/2)

DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

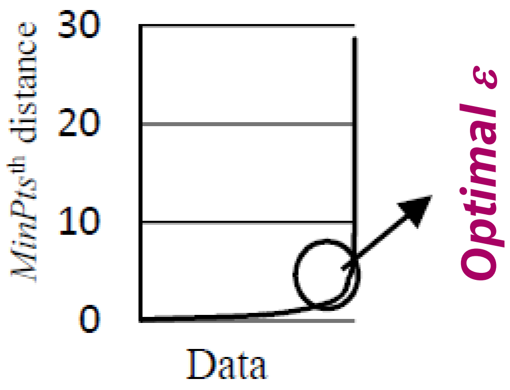
- Arbitrary shaped clusters
- Handles Noise/Outliers
- Optimal ε : sharp change in distances



ε = 1cm

MinPts = 5

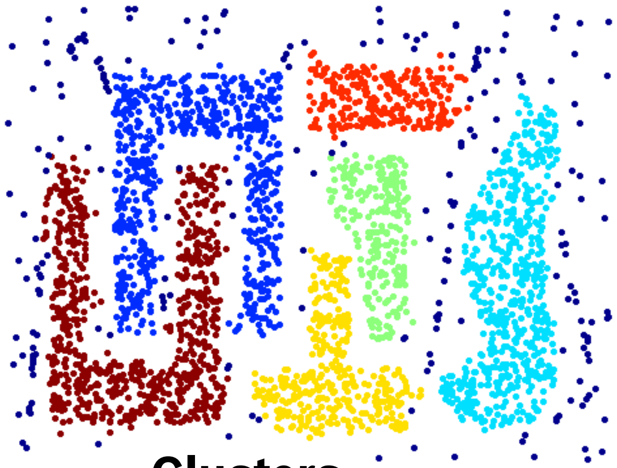
Clusters



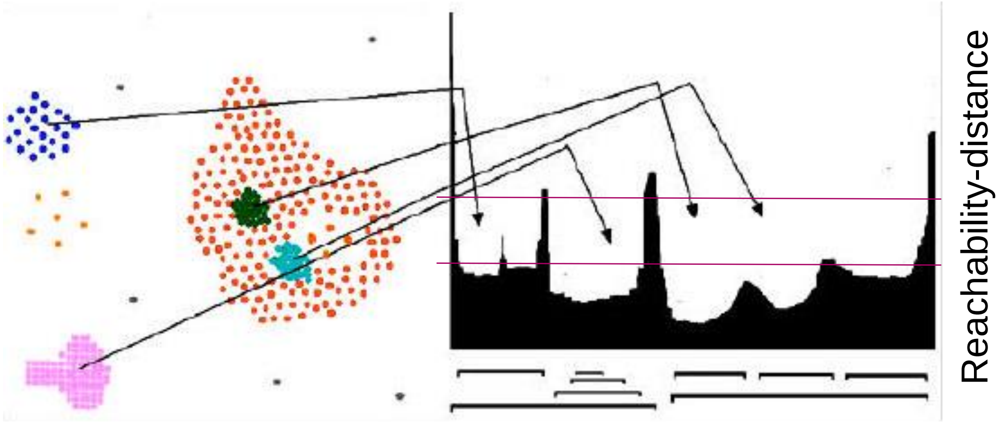
Clustering Techniques
Density-based Clustering (2/2)

OPTICS (Ordering points to identify the clustering structure)

- $\varepsilon \leq \varepsilon'$
- Ordering (PQ)
- Reachability
- Core-distance



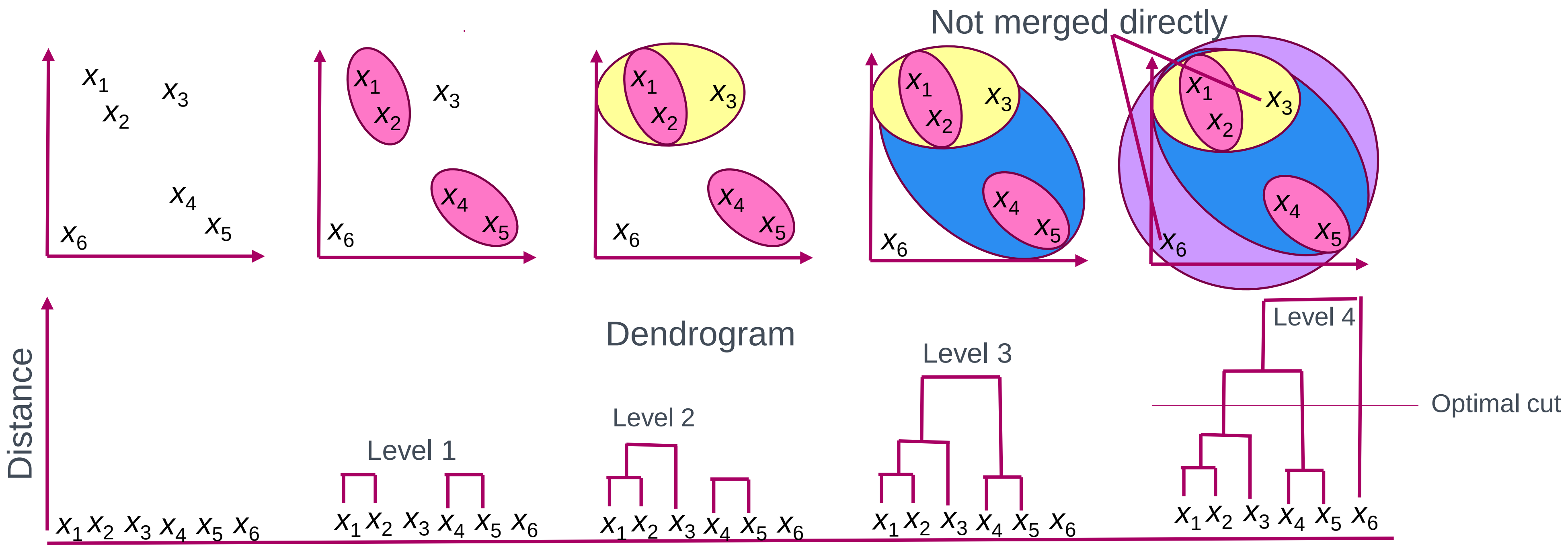
Clusters



Source: http://scialert.net/fulltext/?doi=itj.2009.476.485

Clustering Techniques
Hierarchical Clustering (1/2)

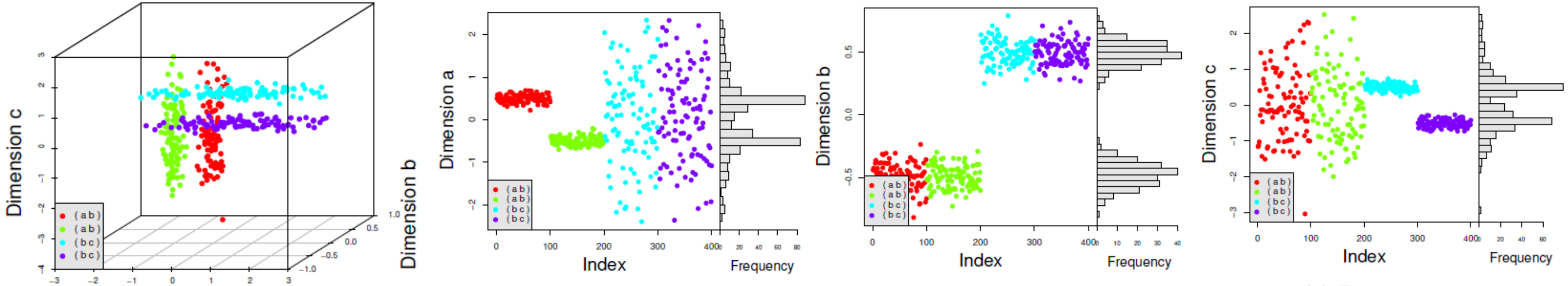
Agglomerative (bottom-up): SLINK, CLINK, Average Link, Ward's, etc.
Divisive (top-down): Bisecting *k*-means
How to find optimal cut in the dendrogram?



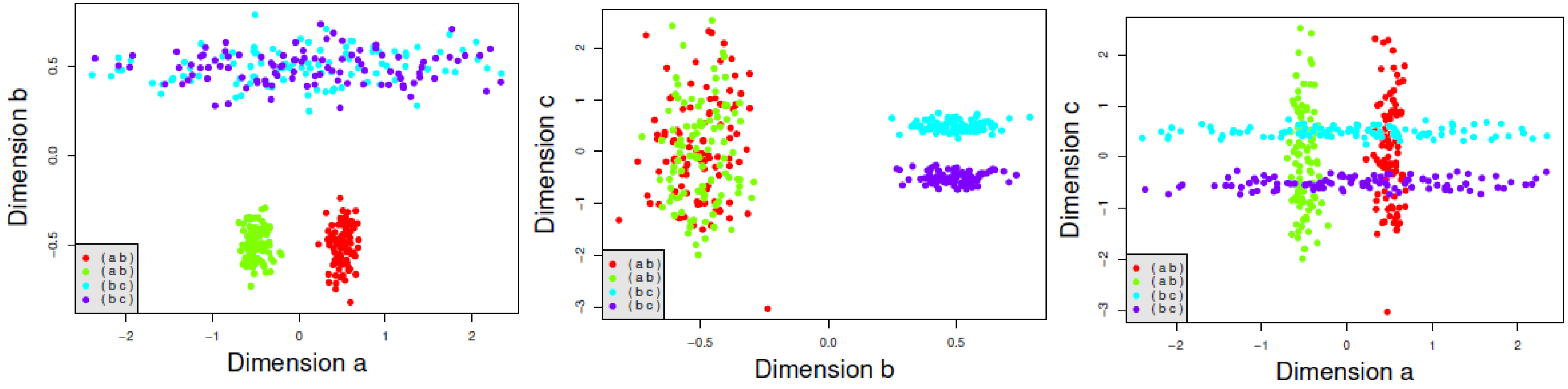
Clustering Techniques
Hierarchical Clustering Algorithms (2/2)

- SLINK (Single Linkage)
  - Distance between two clusters is determined by the distance of the two closest objects (nearest neighbors) in those two clusters
  - Produces long loose clusters which sometimes results into chaining effect (data dependent)
- CLINK (Complete Linkage)
  - Distance between two clusters is determined by the greatest distance between any two objects (farthest neighbors) in two different clusters
  - Produce tight clusters, but sensitive to outliers
- Group-average Linkage
  - Distance between two clusters is determined by taking average distance between all pairs of the objects in two different clusters
- Centroid based: Minimize the variance of the merged clusters
- Wards Linkage: Minimize the variance of the merged clusters

Clustering Techniques
Subspace Clustering (1/2)



(a) Dimension a (b) Dimension b (c) Dimension c



(a) Dims a & b (b) Dims b & c (c) Dims a & c

Source: L. Parsons, L. Parsons, E. Haque, E. Haque, H. Liu, and H. Liu, "Subspace clustering for high dimensional data: A review," ACM SIGKDD Explor. Newsl., vol. 6, no. 1, pp. 90–105, 2004.

Clustering Techniques
Subspace Clustering (2/2)

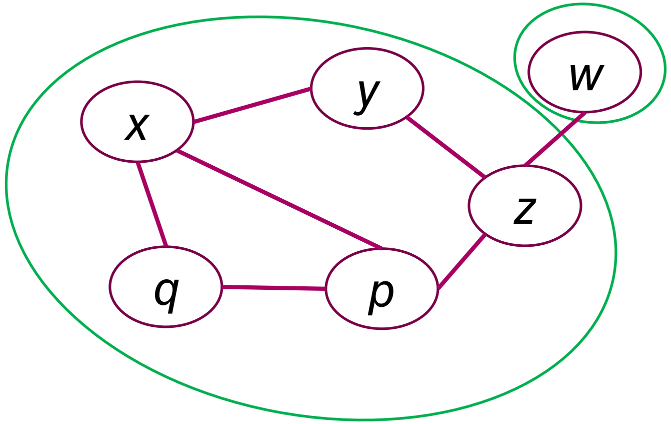
Two types:

- Bottom-up
  - Starts finding clusters in 1-dimensional space and keep on increasing dimensional space
  - Exhaustive approach
  - Time and memory intensive
  - Density-based: DUSC, SUBCLUE, FIRES, INSCY, etc.
  - Grid-based: CLIQUE, ENCLUS, MAFIA, etc.

- Top-down
  - Find important subspaces and then find clusters
  - Time and memory efficient
  - PROCLUS, ORCLUS, FINDIT, etc.

Clustering Techniques
Graph Clustering

- Partition the graph so that edges within a group have large weights and edges across groups have small weights



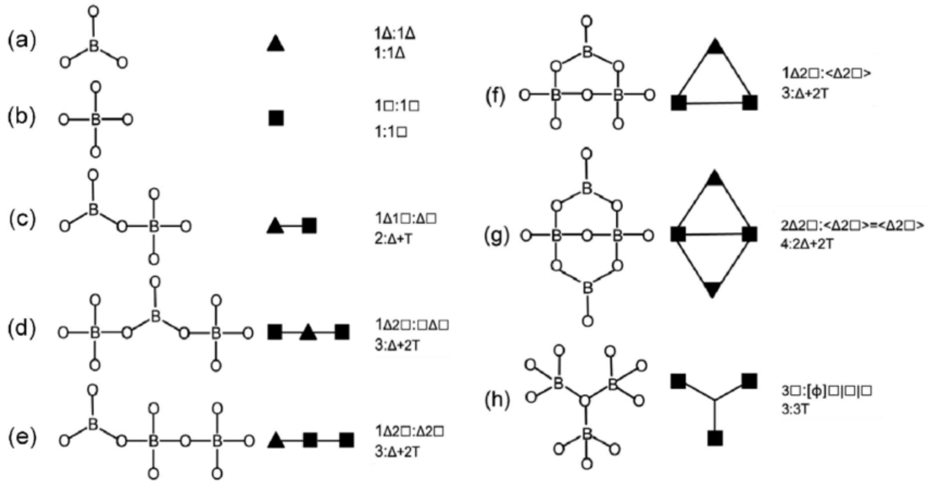
- **Pros:** Fast for sparse data & good clustering results
- **Cons:** Sensitive to the choice of parameter & computationally expensive for large data

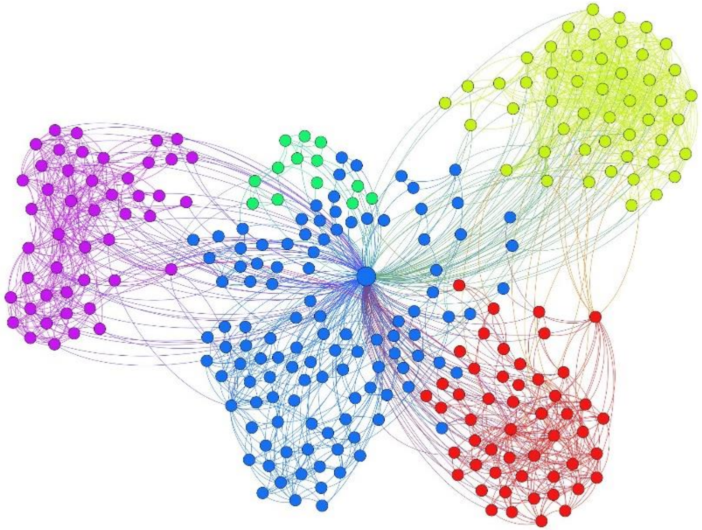
- Graph construction techniques:
  1. Fully connected graph
  2. ε-neighborhood graph
  3. *k*-nearest neighbor graph

NO$_x$ diagnosis with real world driving
Graph Clustering Types

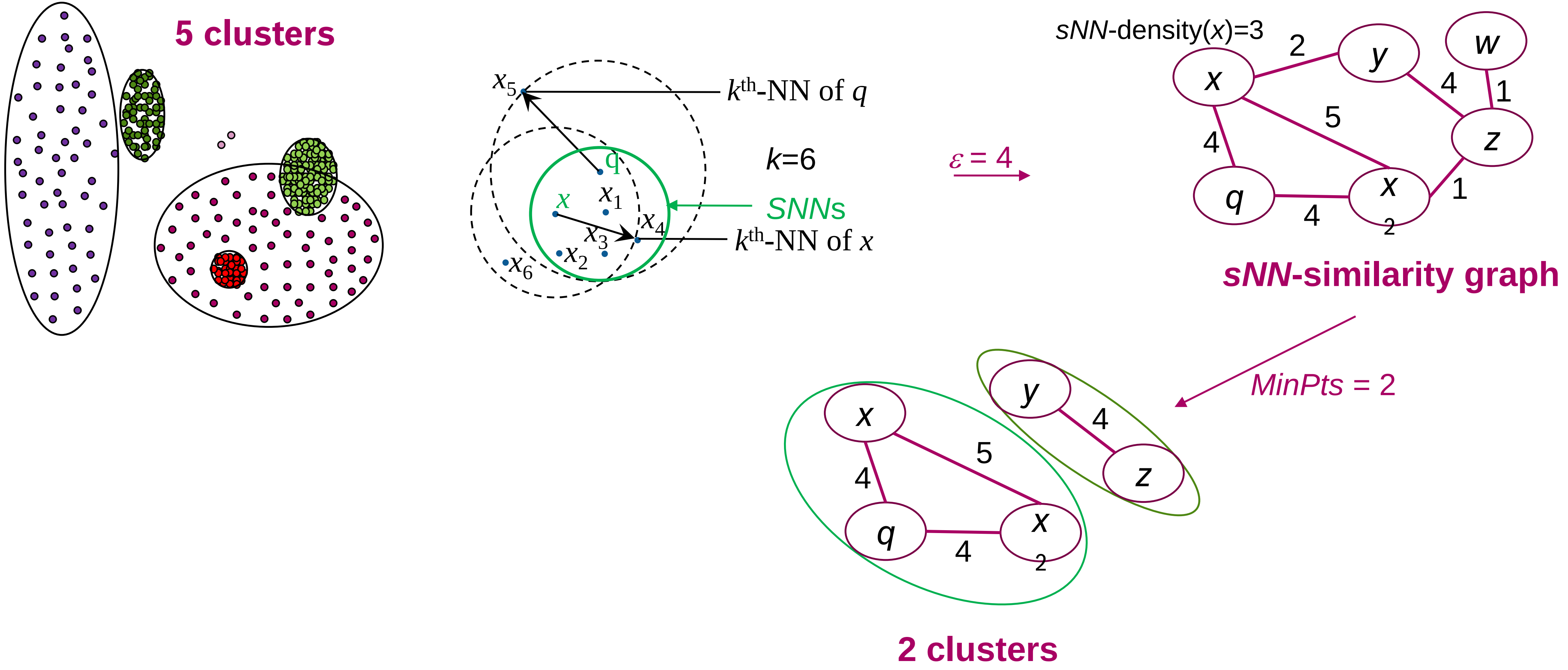
Two major types:

- Between graph
  - Divides set of graphs into different groups
  - Chemicals can be grouped based on their structure similarity
- Within graph
  - Divides the nodes of a graph into clusters
  - In social networking, people with similar behavior can be grouped together
  - Many links within a cluster & fewer links between clusters
  - Hierarchical, Clique, SNN, Spectral, etc.





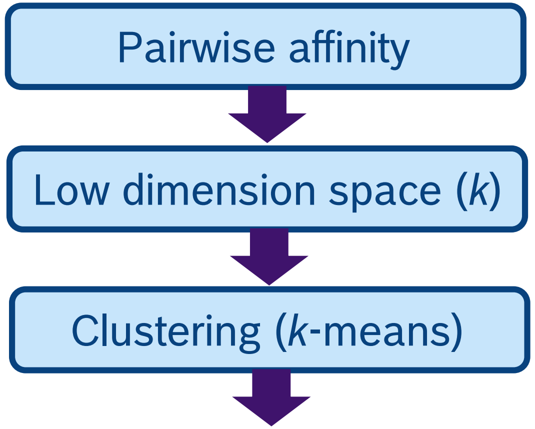
Clustering Techniques
SNN Clustering

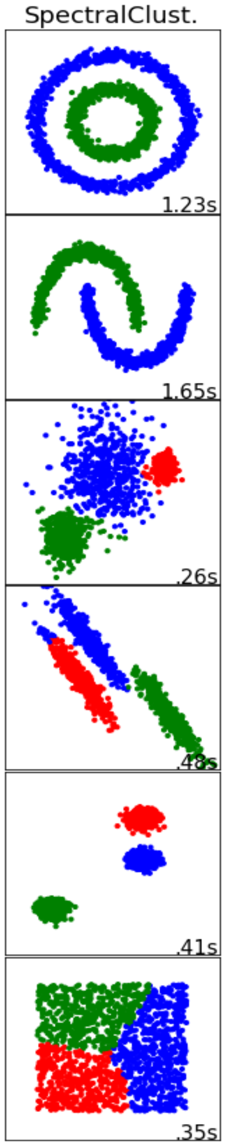


BOSCH

Clustering Techniques
Spectral Clustering

- Also fall in the category of subspace clustering
- Capable to identify arbitrary shaped clusters efficiently (based on connectivity)
- Applications: image/document data, audio data, etc.
- Affinity is inversely proportional to distance
- Algorithm:
  - Construct **pairwise** affinity matrix: $A_{i,j} = exp\left(-\frac{\|x_i - x_j\|}{2\sigma^2}\right)$
  - Construct degree **matrix** $\boldsymbol{D}$=diag($\boldsymbol{d_1},\ldots,\boldsymbol{d}_{\mathrm{n}}$)
  - Compute Laplacian $\boldsymbol{L}$=$\boldsymbol{D}-\boldsymbol{A}$ (unormalized)
  - Compute the first k eigen-vectors $u_1,\ldots,u_k$ of $\boldsymbol{L}$
  - Let $\boldsymbol{U} \in \mathbb{R}^{N\times k}$ contain the vectors $u_1,\ldots,u_k$ as columns
  - Let $\boldsymbol{y_i} \in \mathbb{R}^{k}$ be the vector corresponding to the $\boldsymbol{i}$-th row of $\boldsymbol{U}$
  - Cluster the points ($\boldsymbol{y_i}$) into $\boldsymbol{k}$ clusters with k-means





Clustering Techniques
Challenges in Spectral Clustering

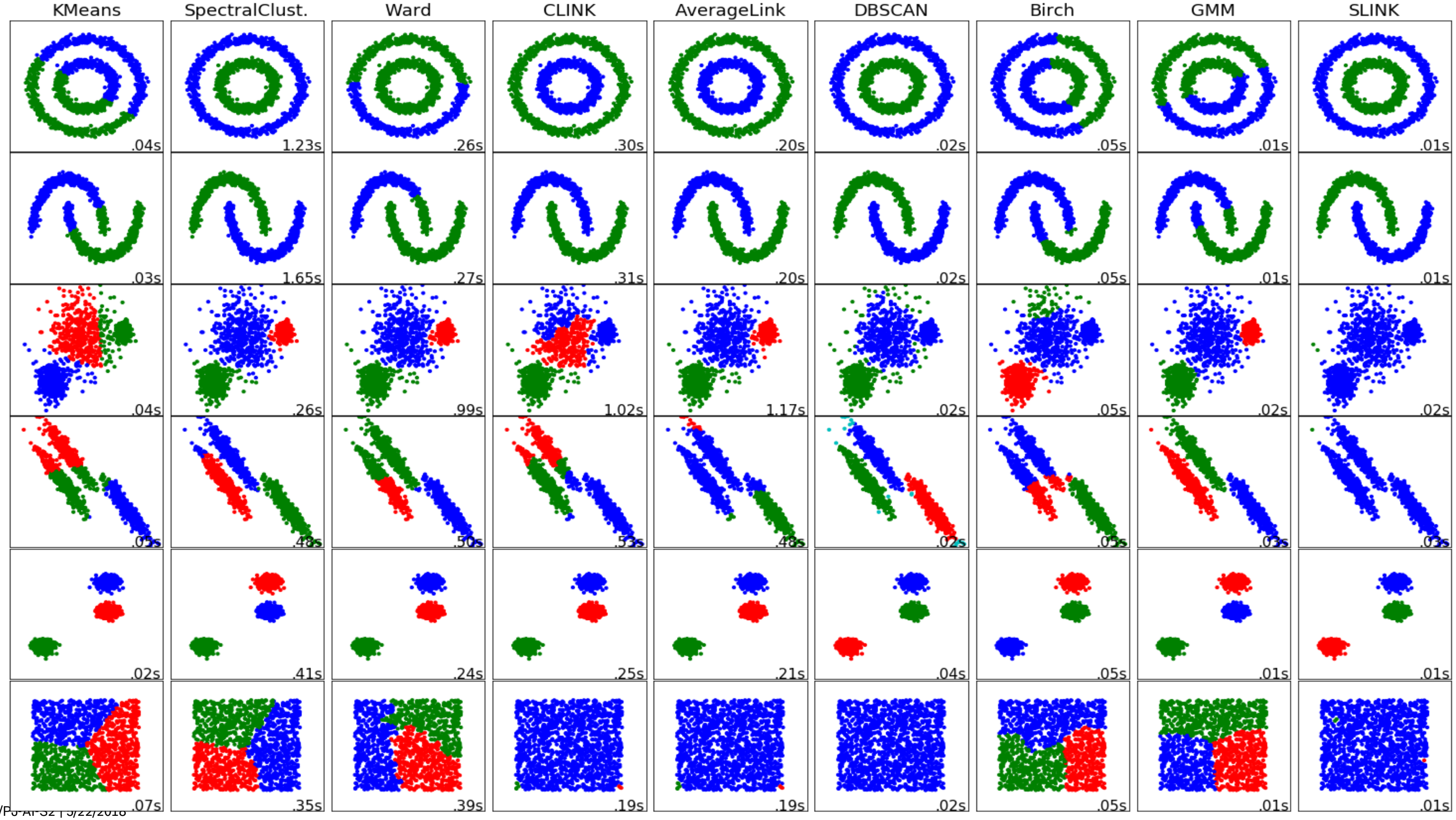
- Number of cluster:
  - The number of eigenvalues of magnitude 0 is equal to the number of clusters (*k*), but this works for well separated clusters
  - Incrementally select a single eigen-vector

- Limitations:
  - Time and memory intensive

Clustering Techniques
Model based Clustering

- Expectation maximization clustering:
  - Similar to k-means
  - At each iteration, assign each object to a cluster with a probability
  - Re-estimate model parameter

Clustering Techniques
Clustering Algorithms Comparison



Clustering Techniques
Cluster Validation

- Overall similarity score (intra-cluster similarity): should be high

- F-measure (high is better) & entropy (low is better):
  - Benchmarked data is required

- Rand-index & Omega-index
  - Benchmarked data is required

Clustering Techniques
Conclusions

- Critical steps involved in clustering
- Various distance metrics
- Different type of Clustering approaches
- Clustering validation approaches

- Following clustering techniques have not been covered:
  - Semi-supervised clustering
  - fuzzy/soft clustering: Each object belong to every cluster with some weight varying from 0-1

Thank You.
-Any Questions?

BOSCH