Exercise 4 Solution

1)

	AHC	DBSCAN
Cluster Number	No initial setting of the cluster number required. Cluster hierarchy allows fast exploration of different cluster numbers. The L-Method can be used to find an appropriate cluster number.	Automatic determination of the cluster number
Effort	High run time complexity O (n^2 logn)	O(n^2), with KD-tree O(nlogn)
Differently shaped clusters	Depends on Linkage Method **Single Linkage** is able to handle clusters with arbitrary shapes and tolerates larger differences in cluster sizes. However, this method is sensitive against outliers and strongly affected by the „chain effect". **Complete Linkage** is less sensitive against outliers; however, it tends to split large clusters into smaller clusters. Moreover, it has problems to handle clusters of different shape. Spherical clusters are preferred. **Average Linkage** is an intermediate form of Single and Complete Linkage. However, more compact round clusters are preferred. The **Ward Method** is robust against outliers, but it prefers spherical clusters of similar size.	Problems with clusters of diverging density, can handle clusters with different shapes

2) L-Method, silhouette coefficient
3) **RIS**: Subspace search without clustering „Ranking Interesting Subspaces"(RIS)

- Observations (derived from density-based clustering)
- Interesting subspaces comprise core objects:
  - o is a core object if $|N_{\varepsilon}(o) \geq \text{minPoints}|$
  - Dense regions: set of objects connected to core objects; the more objects dense regions comprise, the more interesting is the subspace.
  - count(S): number of objects in regions around all core objects in S.
  - „Interestingness"= count(S)/Volume (S)

    - Relevance: a subspace is deleted if embedded in another subspace (more dimensions) with higher interestingness.
- How to choose ε and *minPoints*?
    - Suggestion: *minPoints* = *ln (n)* where *n* is the overall number of objects
    - Q: ln(1000)?, ln (10000)?
    - Choice of ε more difficult. Upper bound *lim* for ε determined considering that for a completely uniform distribution not all points are considered core objects.
    - Suggestion: ε = *lim/4*

SURFING (Subspaces Relevant for Clustering)

- Searches for subspaces without clustering
- Analyse the histogram of the *k* nearest neighbour (NN) distances in all subspaces → Subspaces with non-uniform distributions more interesting.
- **Scaleability**: NN computation restricted to e.g. 5% of the elements
- Bottom-up search: An interesting subspace is expanded with further dimensions as long as „interestingness“increases.
- Properties:
    - Does not assume a specific clustering structure
    - Applicable for a wide range of numbers of dimensions
    - Algorithm is stable for *k* between 5-20
    - Resulting high-ranked subspaces tend to be similar/redundant
- „Interestingness “
    - Compute the **diff**erences of the knn-distance to the mean (not squared)
    - Count points with a knn-distance **below** the mean value (points in dense regions)
    - Determine **quality** as normalized differences
- returns relevant subspaces ranked by „interestingness “, e.g., by variance and entropy
- no clustering

CLIQUE (CLustering In QUEst, Agrawal, 1998):

- Pioneering **cell-based method**
- Generates clusters of arbitrary shapes
- Bottom-up method that seeks dense rectangular cells in all subspaces with a high density
    - A grid with constant size overlaid on the data.
    - Results heavily depend on the grid resolution ε and the minimum number of points per cell *minPoints*.
- Clusters represent connected components in a graph where the nodes represent dense units
- Refinements
    - Free positioning of cells (hyper-cubes) with fixed grid size

- Adaptive grid size (depending on the number of dimensions, SCHISM)

SUBCLUE

- density-based method that employs a global density threshold *t*.
- Generates clusters of arbitrary size and shape
- Slightly better quality than the grid-based approach but also dependent on *t*.
- Slower than CLIQUE (performance slows down with > 15 dimensions)
- With a global threshold *t* not all relevant sub clusters are found.
- Bias towards low dimensionality

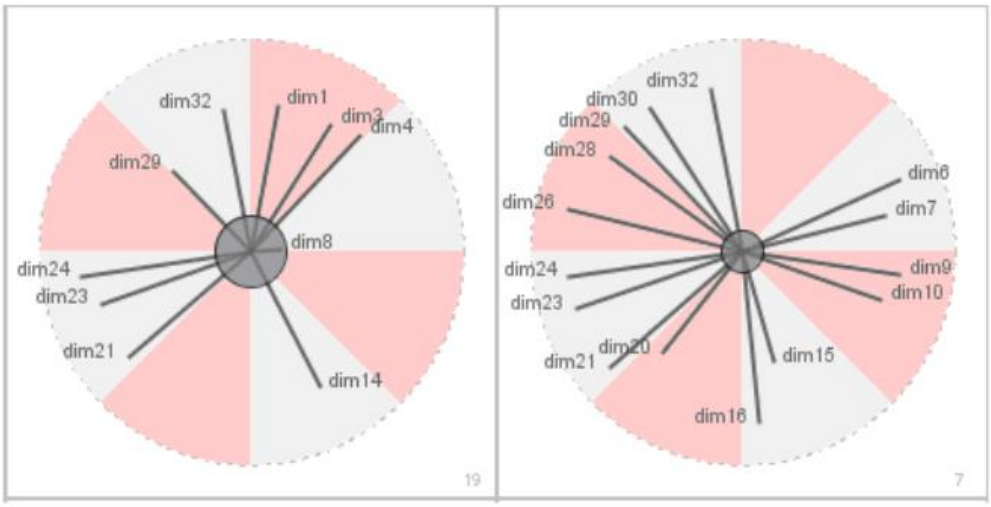
Proclus (PROjected CLUStering)

- **Clustering-based** approach with two parameters: number of clusters *(C)* and average dimensionality *(D)*
- Random initialization of *C* medoids (similar to k-means)
- **Medoids** are representative objects of a data set or a cluster with a data set whose average dissimilarity to all the objects in the cluster is minimal.
- Medoids are similar in concept to means or centroids, but medoids are always restricted to be members of the data set.
- In the refinement stage, for the medoids well-fitting subspaces are searched.
- Medoids are translated until the subspaces do not get better anymore.
- **Properties**: prefers large clusters, efficient, simple, robust against noise

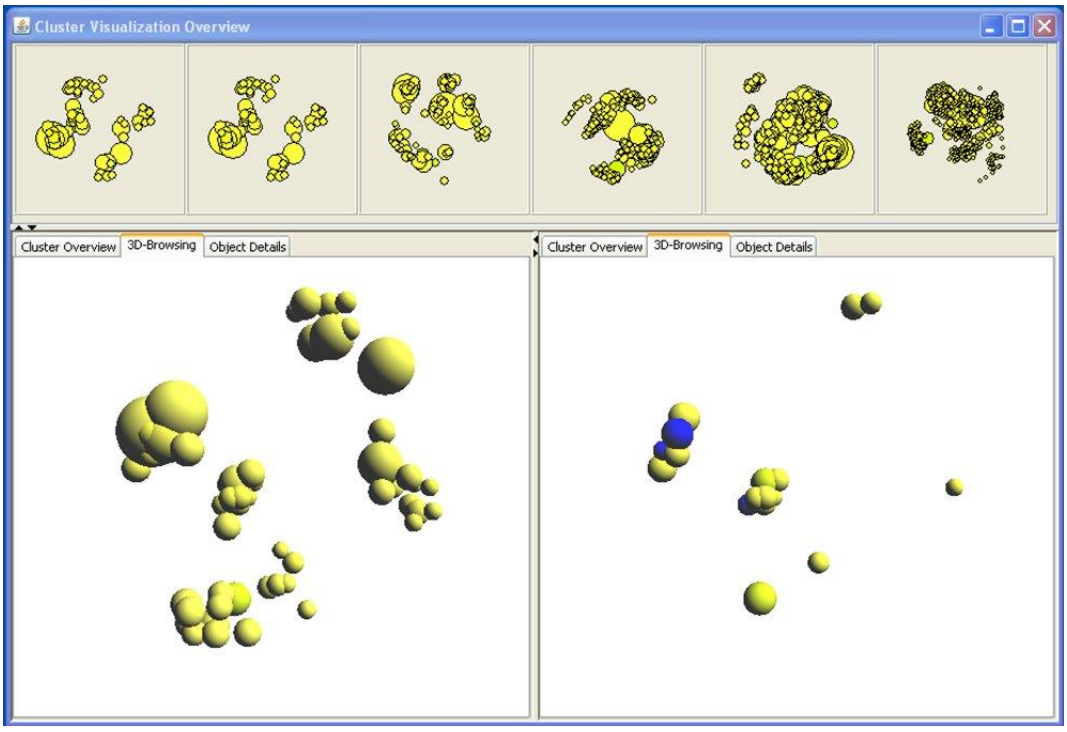
4) SubVis: Comprehensive visualization of clustering results: 2D-Overview of sub-space clusters bottom left (MDS projection). Small multiples represent results with different distance functions. On top: Distribution properties of subspaces. Aggregation tables with aggregated members and table lens for details

ClustNails:

- For a subspace cluster, each dimension that contributes has a weight
- dimensions with high variance have a high weight since they contribute strongly to the separation
- Subspace clusters visualized along with the weights of their dimensions.
- Radial visualization designed such that clusters are comparable w.r.t. involved dimensions
- Length of the spikes („nails“) represents the weights per dimension. Only contributing dimensions labelled. Background colours for comparability



OpenSubspace:



5)

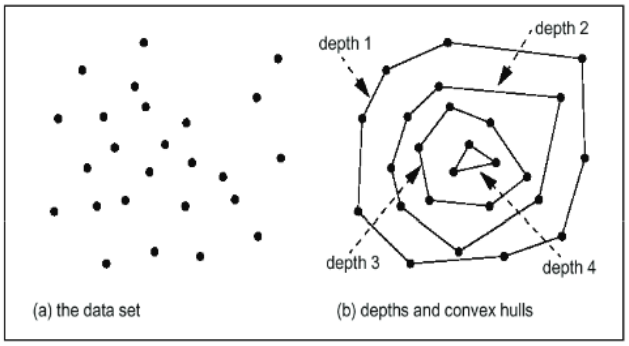
a) **Silhouette Coefficient** (determine distance of object o to all other objects a in Cluster A and to all other objects b in Cluster B and averages the distances → dist(B,o)-dist(A,o)/max(dist(B,o),dist(A,o)))
   i) Runs between -1 und 1
   ii) Coefficient: Sc = 1/nc * Sum(S(o)) → high values are good

b) **Centroid-Based Measure**
   i) Determine centroid for each cluster
   ii) For each point within a cluster: determine distance to centroids
   iii) Distance to the own centroid should be smaller than to all other centroids
   iv) Points that do not fulfil this describe the cluster quality
   v) Overall measure is built from the sum over all clusters → weighted by cluster size → high values preferred
   - low values for narrow and curved clusters; high values for convex compact shapes.
   - low values for split and interwoven (spatially not coherent) clusters.
   - relatively robust against different sizes, densities and numbers of clusters

c) **Grid-Based Measures**
   i) Place a grid over the cluster result
   ii) Count cells with one cluster and multiple clusters
   iii) Good result: little mixed cells
   iv) Cell size: sqrt(n), 3sqrt(n) for 3D
   - Grid-based measure more robust against split clusters and non-convex shapes
   - Influence of the grid resolution: With a too low resolution the results are positive even in case of bad clusters.
   - Grid size must be chosen such that typically several points are in one cell

6) Parallel Coordinates, Scatterplots, Heatmaps, Dendrograms
7) Outlier Detection:

- **Depth-based**
  - Assumption: Outliers at the boundary of a distribution and „normal“ elements in the centre
  - Depth 1: elements are part of the convex hull of the data → likely outliers
  - Depth 2: elements are part of the convex hull after Depth 1 points are removed → less likely outliers

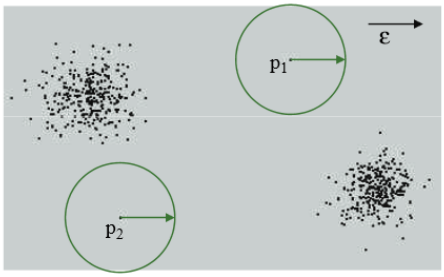


Picture taken from [Preparata and Shamos 1988]

- **Distance-based**
  - Outlier has abnormal distance to its nearest neighbour
  - Not appropriate if the distribution of data varies strongly in local density
  - Given a radius ε and a percentage π
  - A point p is an outlier if at most π percent of all other points have a distance to p less than ε

$$OutlierSet(\varepsilon,\pi) = \{p \mid \frac{Card(\{q \in DB \mid dist(p,q) < \varepsilon\})}{Card(DB)} \leq \pi\}$$

range-query with radius ε



- **Density-based**
  - Outlier has an abnormally low density in the surrounding
  - Detect outliers if they have a locally abnormal distance to the nearest neighbour. In contrast to depth- and distance-based methods, o_2 is an outlier according to the Local Outlier Factor