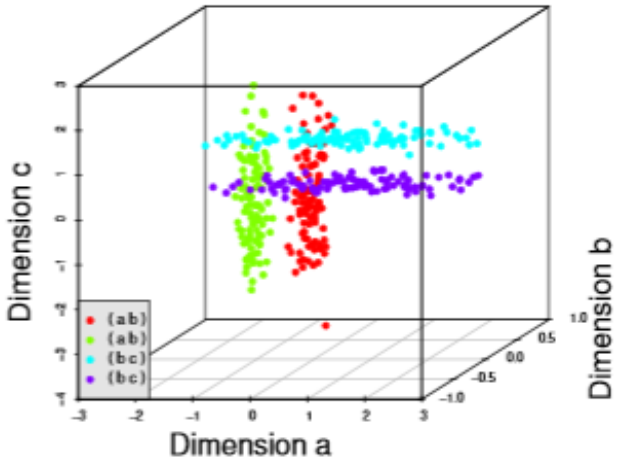
Cluster Analysis and Visualization of Clustering Results: Subspace Clustering



Outline

- Strategies for subspace search and clustering
- Methods for subspace search and clustering
  - Preprocessing (missing data, normalization, transformation)
- Visualization of Clustering Results
  - Scatterplots
  - Parallel coordinates
  - Graphs showing relations between subspaces
  - Multiple linked views
- Applications
- Evaluation

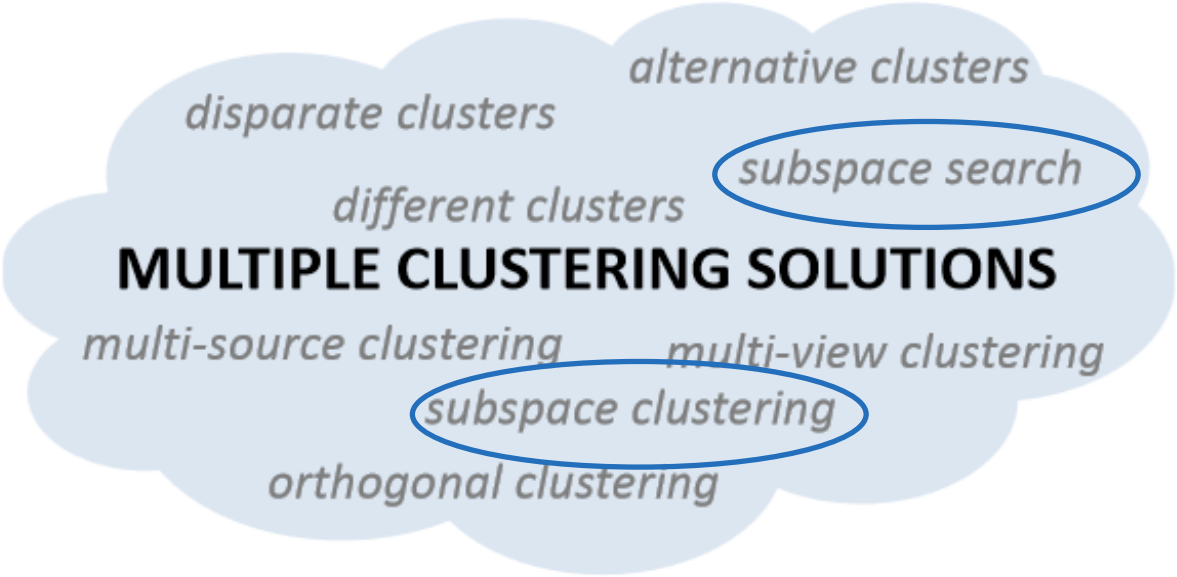
Introduction

Most problems where clustering may detect structures are actually subclustering problems:

- In demographic datasets, many attributes per person are registered.
  - Person A and B may be similar w.r.t. health, but may differ w.r.t. mobility and wealth-related attributes.
- In recommender systems, people are grouped w.r.t. similar interests.
  - Person A and B may have similar interests in science, but A has completely different interests in music.
- In the analysis of gene expression profiles from DNA microarray chips, genes are grouped according to similar expression profiles and related functions (pionieering application, Cheng, 2000).
  - Gene A and B may be similar under certain environmental conditions (in some tissues or before a treatment), but under different conditions Gene B and C are similar (that were not similar originally)

Introduction

Subspace search and subspace clustering are two essential instances of multiple clustering solutions
(From: Müller, 2013)



Subspace Clustering

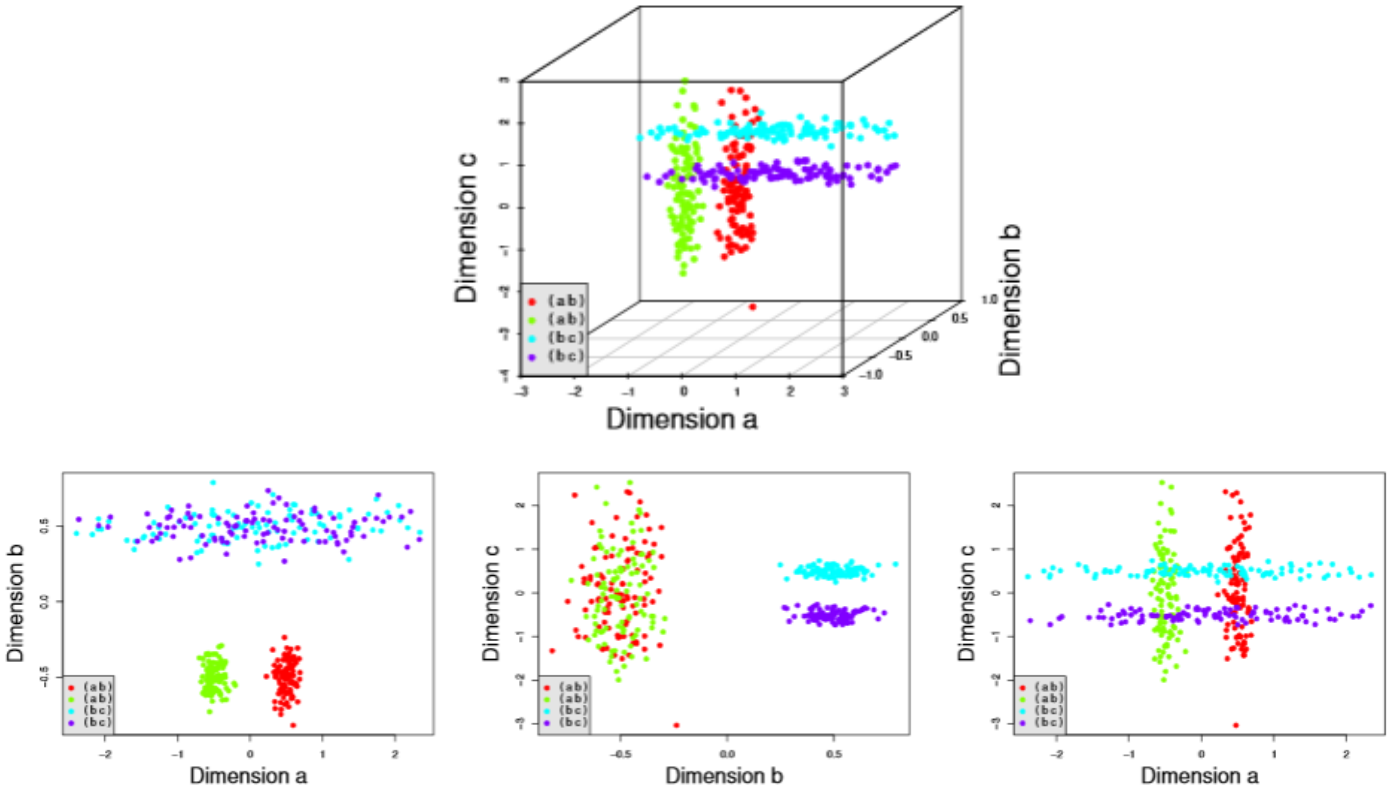
Curse of dimensionality: While data in low-dimensional spaces is rather tight and can be clustered, in high-dimensional (HD, (D > 10 .. 15)) spaces distances are very large (sparse data).

Noise, irrelevant and highly correlated dimensions reduce the quality of global clustering (Hund, 2016)

Subspace search refers to methods aiming at finding low-dimensional representations of a HD dataset useful for grouping (*clusterable* subspaces).

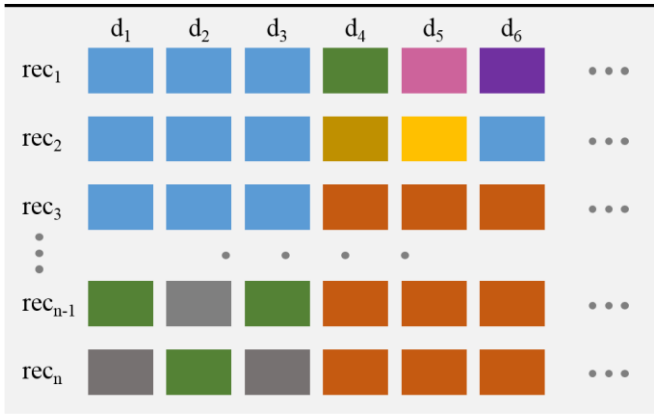
Subspace search requires heuristics to prune the search: For n dimensions: 2^n-1 possible axis-aligned subspaces.

Subspace Clustering



A 3D dataset with no 3D cluster but 4 clusters in different 2D subspaces. In a-c 2D space all clusters are recognize-able but overlap. (From: Parsons, 2004)

Subspace Clustering



With the naked eye, clusters in subspaces are visible, e.g. the first three records are similar in dimensions d_1, d_2, d_3 (From: Hund, 2015). No records are similar in all dimensions.

Introduction

Two terms were employed as synonyms:

- Subspace clustering, introduced by R. Agrawal (1998)
- Projection clustering, introduced by C. Aggarwal (1999)

Meanwhile *subspace clustering* is the dominant term.

Formally:

Let O be a set of objects in a database DB with D dimensions.

A cluster C in a subspace projection S is

C = (O, S) with $O \subseteq \mathrm{DB}, S \subseteq \mathrm{D}$

A clustering result R is defined as:

$R = \{C_1, \ldots, C_k\}, C_i = (O_i, S_i)$

Subspace Clustering

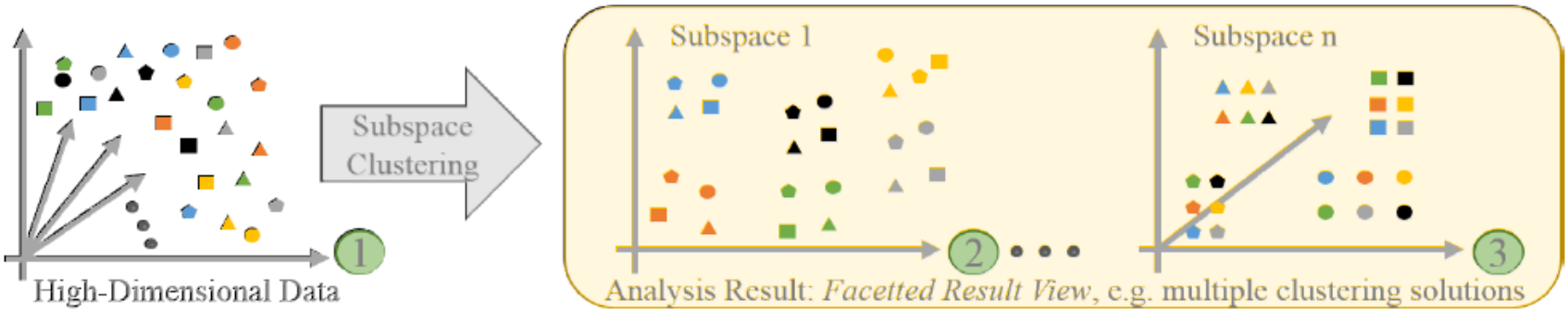
Two tasks to be solved:

- Search relevant subspaces and provide a measure of „interestingness“ for ranking
  - Since this may take long, it is reasonably a non-interactive preprocess
- Cluster data in these subspaces

Related tasks in visual subspace analysis (Tatu, 2009):

- Display the subspaces (involved dimensions)
- Display similarity between subspaces (w.r.t. involved dimensions and topology)
- Display the clustering results in these subspaces
- Show relations between subspace clusters

Subspace Clustering



We are searching for subspaces, represented by certain dimensions and clusters within these subspaces
(From: Hund, 2015).

Subspace Clustering

- Dimension reduction, e.g. feature selection is NOT part of subspace search.
  - Feature selection: based on *entropy* or *information gain;* statistical measures related to the dimension.
  - These GLOBAL techniques are not able to find local clusters (covering only parts of the objects)
  - After PCA, e.g., subspace clusters detected in the original data get no longer detected. Also, the new dimensions are hard to interpret.

- In most cases many subspaces are relevant.
- Subspace search algorithms aim at finding features that correlate locally

Subspace Clustering

Different methods:

- *Subspace search* algorithms only return subspaces (where any global clustering may be applied)
- *Subspace clustering* techniques combine subspace search and clustering

Discussion (see Tatu, 2012):

Decoupling subspace search from the actual clustering

- is more flexible,
- is less biased (clustering strongly relies on assumptions), and
- more effective, since „uninteresting“ subspaces are filtered

Subspace Clustering

Assumptions and Heuristics

- Since the number of possible arbitrarily oriented sub-spaces is infinite, assumptions and heuristics are used.
- Most algorithms search only for axis-aligned subspaces
- Some algorithms (again) expect a certain number of clusters and optimize subspace search accordingly
- Some algorithms are biased towards low dimensional clusters

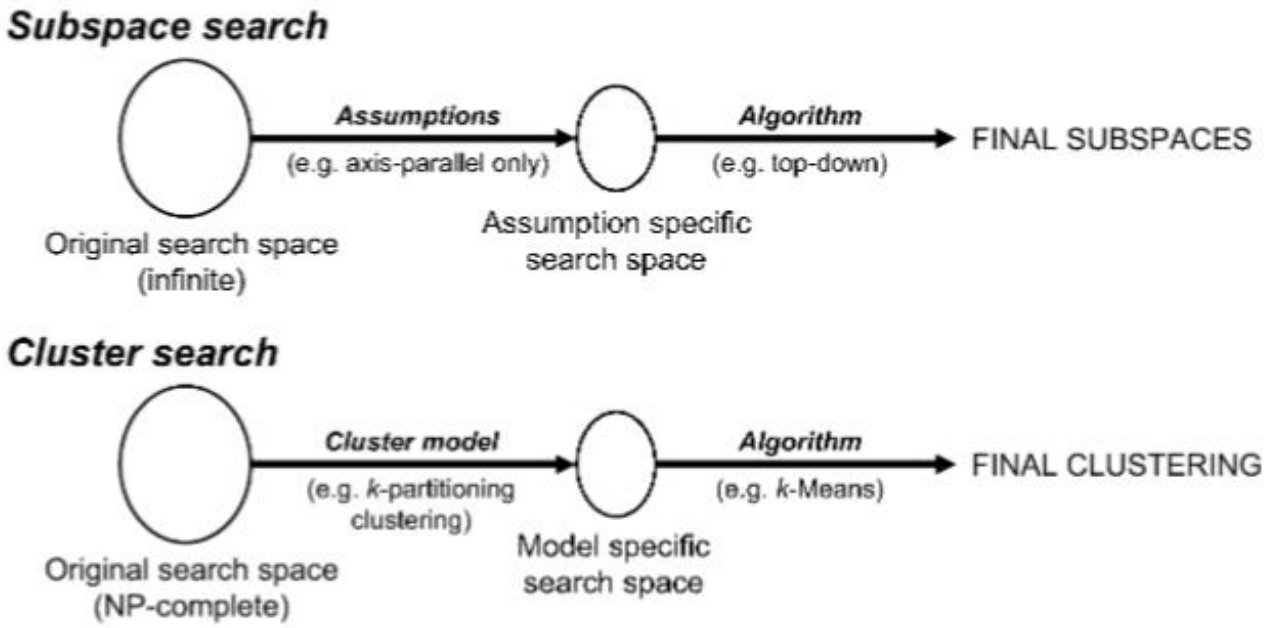
Subspace Clustering

Preprocessing:

- Treatment of incomplete datasets
  - Incomplete datasets may be removed (Hund, 2015)
  - Missing data may be „interpolated“
  - Advanced: consider incomplete data but do not use interpolated values
- If categorical variables are not treated separately, they are transformed to numbers
  - Binary variables are mapped to 0 and 1
  - Variables with e.g. 6 values are mapped to 0, 0.2, …, 1.0
- Normalization
  - For numerical variables in the range [min, max] they are transformed to the range [0,1]

$$normalized\ value_j^{dim_i} = \frac{value_j^{dim_i} - min(dim_i)}{max(dim_i) - min(dim_i)}$$

Subspace Search and Clustering



(From: Kriegel, 2009)

Subspace Search: RIS

Subspace search without clustering

„Ranking Interesting Subspaces“ (RIS) (Kailing, 2003)

Some observations (derived from density-based clustering):

• Interesting subspaces comprise core objects:

o is a core objects if $\left| N_{\varepsilon}(o) \geq \text{minPoints} \right|$

• Dense regions comprise points that are not core objects; the more points dense regions comprise, the more interesting is the subspace.

count(S) is the number of points in regions around all core objects in subspace S.

„Interestingness“ = count(S)/Volume (S)

• For chosing relevant subspaces, a subspace is deleted if it is embedded in another subspace (more dimensions) with higher interestingness.

Subspace Search: RIS

How to choose ε and *minPoints*? (Kailing, 2003)

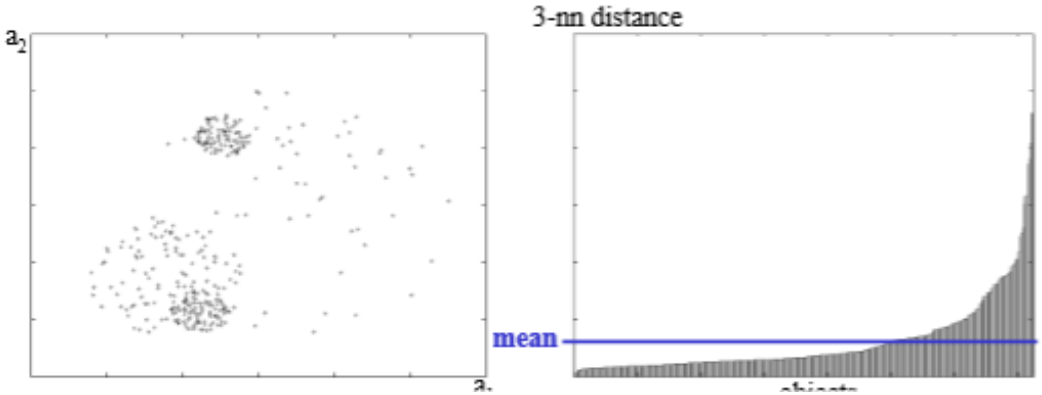
- As suggestion: *minPoints* = *ln (n)* where *n* is the overall number of objects
- The choice of ε is more difficult. Often an upper bound *lim* for ε is determined considering that for a completely uniform distribution not all points are considered core objects.
- As suggestion: ε = *lim/4*

Subspace Search: Surfing

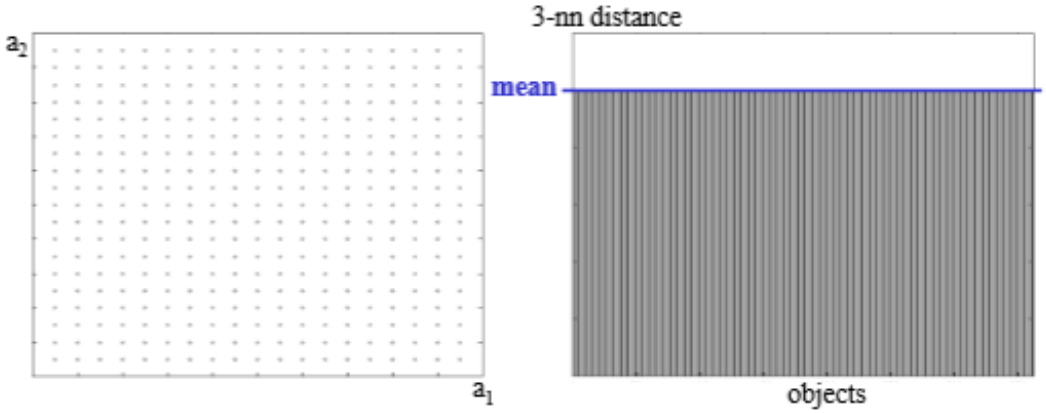
SURFING (Subspaces Relevant For ClusterING, [Baumgartner, 2004])

- Searches for subspaces without clustering
- Analyze the histogram of the *k* nearest neighbor distances in all subspaces. Subspaces with non-uniform distributions are more interesting.
- Scaleability: The nearest neighbor computation may be restricted to e.g. 5% of the elements to save time.
- Bottom-up search: An interesting subspace is increased 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
  - No parameters; algorithm is stable for *k* between 5-20
  - Resulting high-ranked subspaces tend to be similar/redundant

Subspace Search: Surfing



In a clusterable subspace, the histogramm of the 3-nn distances is highly non-uniform



(From: Baumgartner, 2004)

In uniformly distributed data, the distance histogram has equal values. This subspace is not interesting

Subspace Search: Surfing

Specific quality criterion („interestingness“)

- Compute the differences of the k nn-distance to the mean (not squared, like std. dev.)
- Count how many points have a k nn-distance below the mean value
- Determine quality as normalized differences

$$diff_{\mu_S} = \frac{1}{2} \sum_{o \in DB} |\mu_S - nn\text{-}Dist_k^S(o)|.$$

$$Below_S := \{o \in DB \mid nn\text{-}Dist_k^S(o) < \mu_S\}.$$

$$quality(S) = \begin{cases} 0 & if\ Below_S = \emptyset \\ \frac{diff_{\mu_S}}{|Below_S| \cdot \mu_S} & else. \end{cases}$$

Subspace Search: Surfing

- SURFING returns relevant subspaces ranked by „interestingness“ (subspaces are interesting, e.g. because they have a high variance and entropy)
- It does not perform clustering.
- It may be combined with (global) clustering techniques.
- The authors suggest density-based clustering and favor OPTICS (Ankerst, 1999)

Subspace Clustering

Methods for subspace clustering:

- Bottom-up. Start with individual dimensions and merge with others.
  - Monotonicity assumption: If a subspace S contains a cluster C, all subspaces $T \subseteq S$ also contain C. → If T does not contain a cluster, no superspace of T contains a cluster.
  - As a measure for merging clusters, the portion of common points is often used.
- Top down.

Bottom-up approaches are more efficient in HD data (> 20 dimensions) and find subspaces with lower dimensions.

Subspace Clustering

Heuristics to prune the search (Tatu, 2012):

- Prefer subspace with a *high variability* that support partioning (*clusterable spaces*)
  - Subspaces with highest interestingness values are often similar (merely projections of the same pattern in different subspaces)
  - Instead of exploring these similar spaces, they should be grouped and one representative should be explored.
  - In addition, further – not similar – subspaces should be displayed.
- For generating a set of subspaces, avoid high redundancy, e.g. where the involved dimensions overlap strongly

Redundancy w.r.t. involved dimensions, compute the *Tanimoto Similarity* (fraction of dimensions contained in both subspaces)

Subspace Clustering

Different paradigms lead to different algorithms (Müller, 2009):

- **Cell-based** algorithms impose a grid structure over the data and search for grid cells where the number of objects is above a threshold.
- **Density-based** algorithms search for subspaces where the local density is higher than in the surrounding. Resulting subspace clusters may differ in density (depending on the surrounding)
- **Clustering-based** algorithms are steered by global parameters related to the expected result, e.g. the number of clusters to return or the average number of dimensions of these clusters.

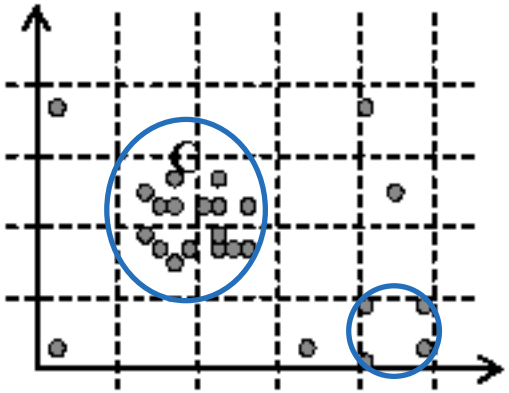
Subspace Clustering: Clique

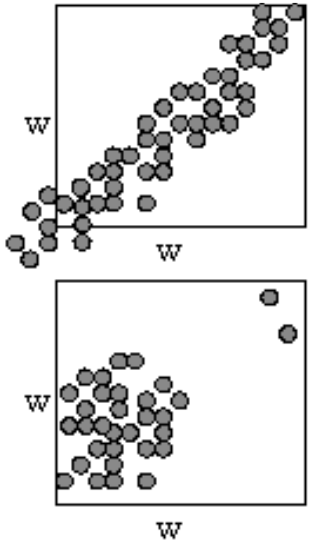
CLIQUE (CLustering In QUEst, Agrawal, 1998):

- Pioneering **cell-based method**
- Generates clusters of arbitrary shapes
- Bottom-up method that seeks dense rectangular cells in all subspaces with a high density of points
  - A grid with constant size is overlaid on the data.
  - The resulting subspaces and clusters 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

Later refinements enabled free positioning of cells (hyper-cubes) with fixed grid size or used an adaptive grid size (depending on the number of dimensions, SCHISM)

Subspace Clustering: Clique





(From: Kailing, 2003)

Left: With minPoints=5 no cluster is returned.

With minPoints=4, two clusters are returned.

The correct result, *one cluster*, cannot be achieved with this gridsize

Right: Other algorithms with hypercubes of fixed sizes have similar problems. Reported clusters do not contain all objects and contain also wrong objects.

Subspace Clustering: SubClue

- SUBCLUE (Kailing, 2004) **density-based method** that employs a global density threshold.
- Generates clusters of arbitrary size and shape
- Slightly better quality than the grid-based approach but also heavily dependent on the threshold.
- Slower than CLIQUE (performance often slows down considerably with > 15 dimensions, [Müller:2009:HSM])
- With a global density threshold not all relevant subclusters are found. There is a bias towards a certain dimensionality

Subspace Clustering: Proclus

Proclus (PROjected CLUStering [Charu, 1999])

- **Clustering-based** approach with two parameters: number of clusters *(C)* and average dimensionality *(D)*
- Random initialization of *C* medoids (similar to k-means)
- 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 agaist noise

Subspace Clustering: Parameters

Algorithms have different parameters:

- Average or maximum number of dimensions,
- Expected number of clusters
- Cluster dominance factor (how much the density in a cluster should exceed the average density?) (Kailing, 2004)
  - Typical values between 1.5 and 2.0
  - Usually applied for grid-based approaches where for each cell the number of elements is compared to the average.

Subspace Clustering: Categorical Data

- The term „subspace clustering“ is unusual for categorical data.
- Instead, mining techniques search for objects that share the same attribute value (over several dimensions) – *frequent item mining* (e.g. to analyze products brought by several customers)
- Few methods deal with hybrid data (continuous numerical and categorical data).
- The HSM method (Müller, 2009) was the first and considers the expected density in subspaces with continuous data and the expected frequency in categorical subspaces to detect patterns.
- Special normalization for categorical, continuous and hybrid data
- Since frequency counting is faster than density estimation, performance increases with more categorical attributes.

Subspace Clustering: Visualization

Challenge:

A multitude of information needs to be conveyed:

- the overlapping dimensions of subspaces,
- the overlapping subspaces,
- the membership of objects to subspaces.

Subspace Clustering: Visualization

Tasks to be supported (Tatu, 2012, TST):

- Reveal properties of individual clusters
  - How many objects? Which dimensions? Which dimensions are removed (and why?) Distributions in each dimension?
- Cluster comparison
  - Difference between clusters w.r.t. involved objects and dimensions
  - Overlap between clusters
- Quality
  - How good is the quality of the clustering?
  - How sensitive is the clustering result w.r.t. parameters?

As a consequence, the navigation through a parameter space needs to be supported.

Subspace Clustering: Visualization

Importance of visualization:

- While for global clustering, established automatic quality assessments are available (e.g. the silhouette coefficient), for partially overlaping subspace clusters these measures are not applicable.

  → Visualization as primary quality assessment
- Even more data and more perspectives needs to be conveyed compared to global clustering.

Subspace Clustering: Visualization

How to assess visualizations of subspace clustering?
(derived from Hund, 2016)

Non-redundancy:

- Are there not too many overlapping dimensions?
- Not too many overlapping instances?

Coverage:

- Are most instances part of any cluster?
- Are most dimensions part of any cluster?

Cluster characteristics:

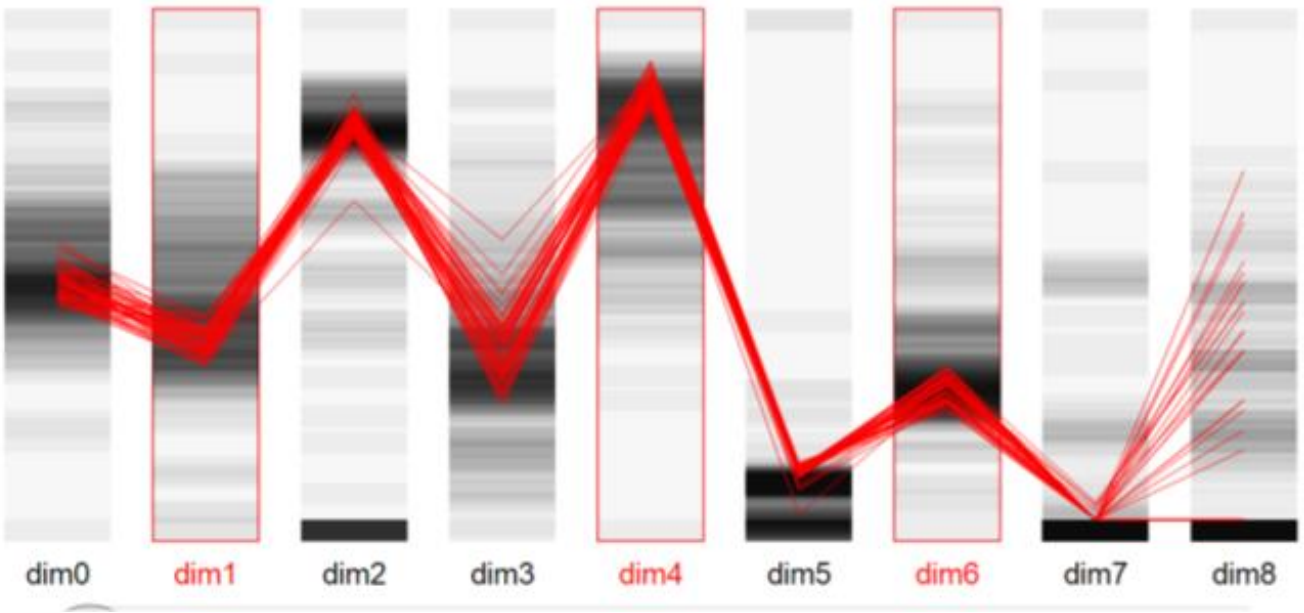
- What number of dimensions the clusters have? (1D or 2D often not interesting)
- Cluster compactness/separability

Subspace Clustering: Visualization

Major techniques:

- Parallel coordinates
- Scatterplot visualizations
- Heatmaps
- Linked views (overview and detailed visualizations)
- Matrix-based visualization (Heidi matrix, [Vadapalli, 2009])
  - Complex visualization that represents all objects and clusters (represented by very few or even single pixels)

Subspace Clustering: Visualization



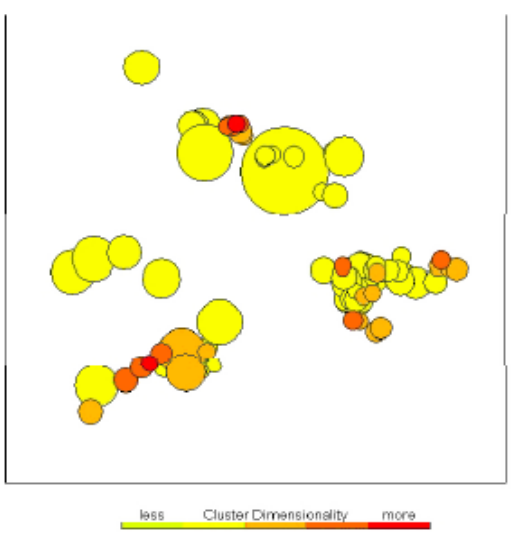
Parallel coordinates reveal the members of a selected subspace cluster. Red dimensions belong to the subspace. As expected, members are more similar in contributing dimensions (From: Hund, 2016).

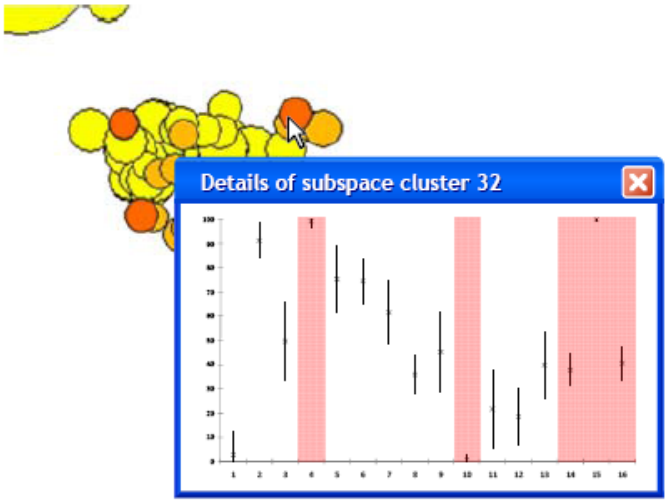
Subspace Clustering: Visualization

Scatterplot-based representations:

- Since most clusters involve >2 dimensions, results need to be projected to 2D
- Projection should preserve proximity, but involves some loss of information
- MDS is often used (Assent, 2007; Tatu, 2012) and performed in experiments better than PCA.
- Comparison with other dimension reduction techniques is essential.

Subspace Clustering: Visualization



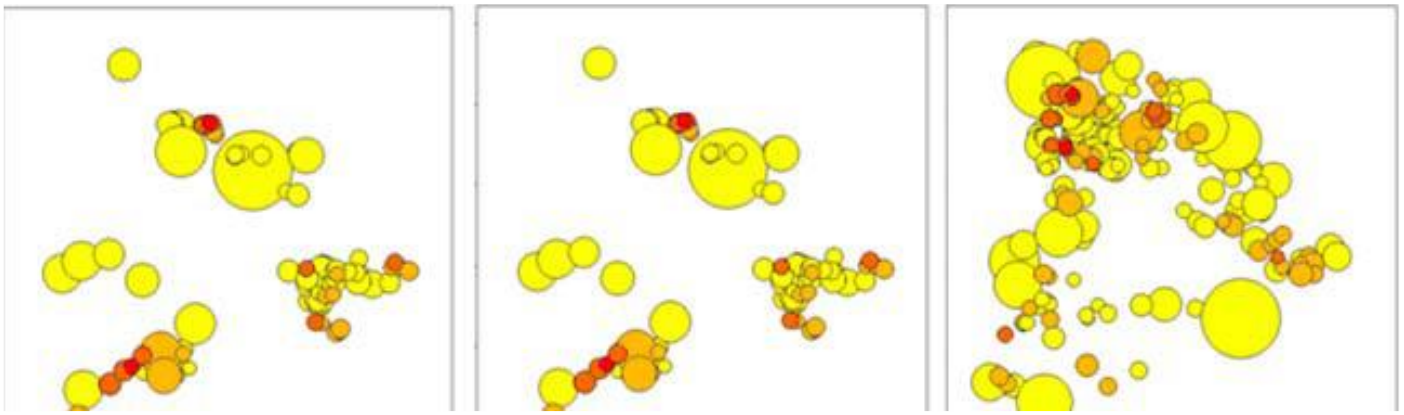


VISA: Visual Subspace Cluster Analysis (Assent, 2007)

Overview: Clusters (determined with a density-based method DUSC) are projected with Multidimensional Scaling and displayed with circles.

- Size represents the number of objects and color represents the dimensionality.
- Spatial proximity reflects similarity between clusters (measured w.r.t. overlapping dimensions and overlapping objects) (From: Assent, 2007)
- On demand, details are displayed for the selected cluster:μ and variance and dimensions involved

Subspace Clustering: Visualization



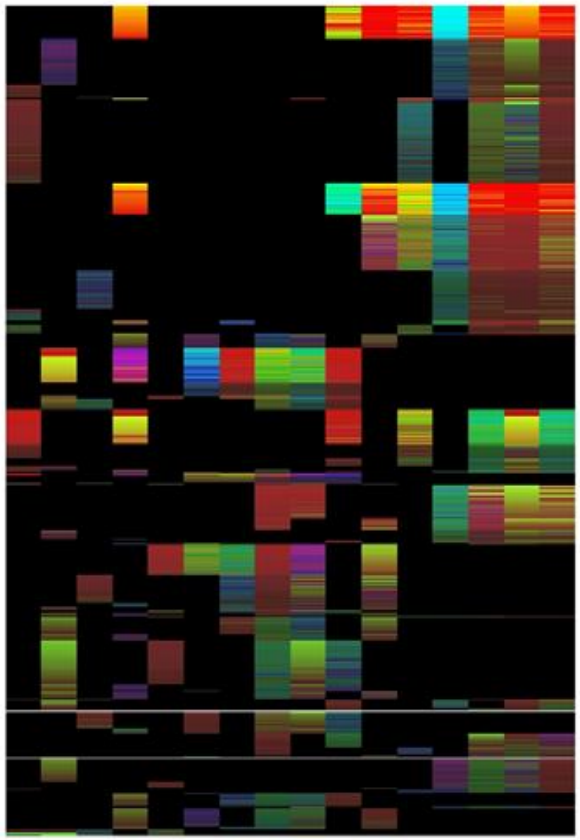
Redundancy between subspaces is computed and used to adjust the redundancy level (From: Assent, 2007).

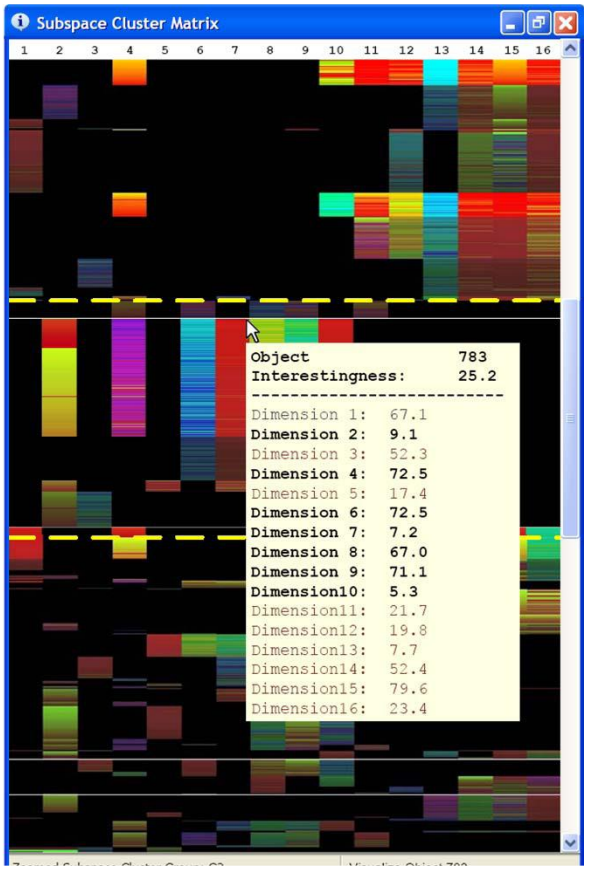
„Bracketing“ is used to support the user in parameter selection – a technique known from photography presenting a number of possible choices in a range

Discussion:

- The distance between clusters can be obscured by a cluster with large size.
- Many clusters lead to a high degree of overlap and clutter.

Subspace Clustering: Visualization

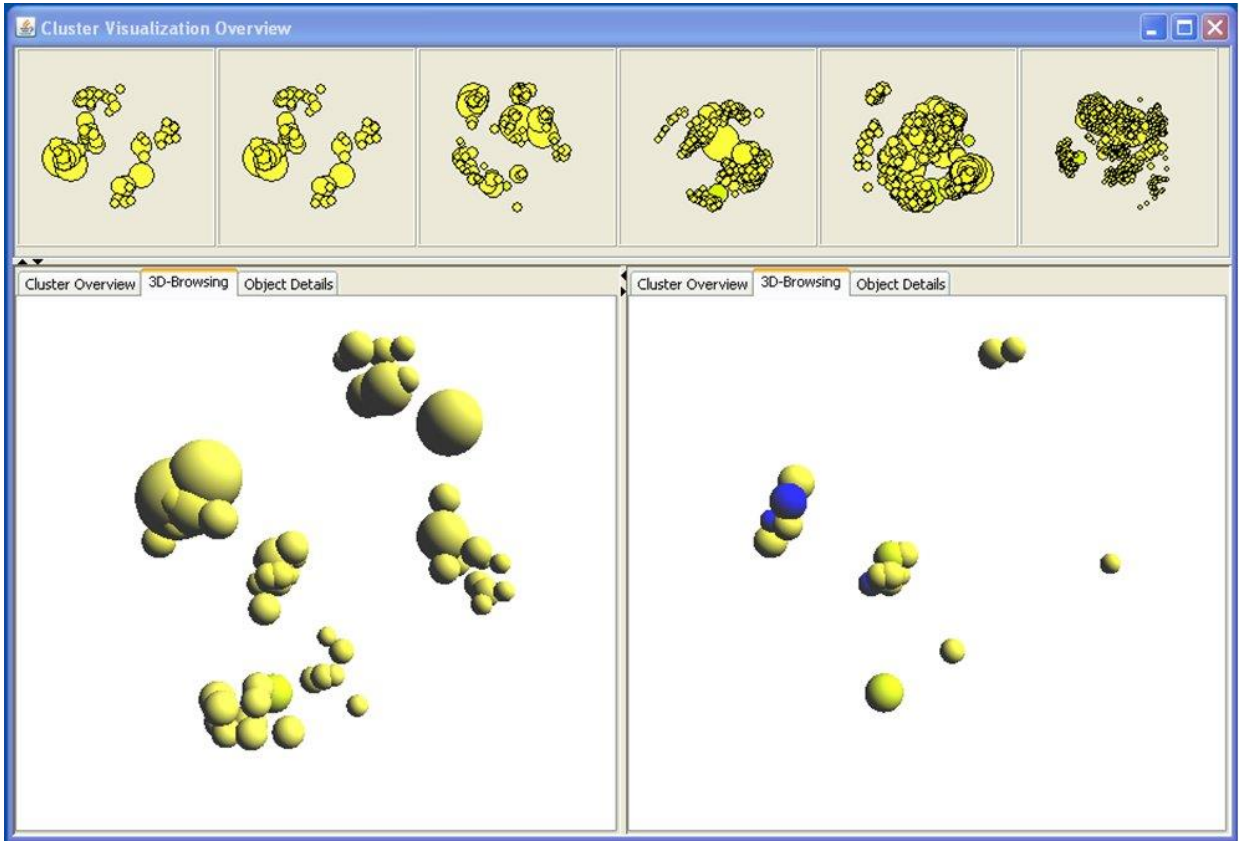




VISA: Visual Subspace Cluster Analysis (Assent, 2007)

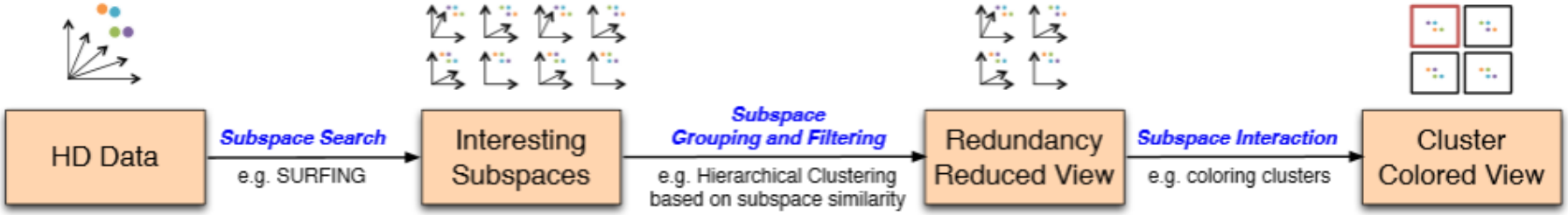
Matrix visualization for detailed analysis: columns represent dimensions and rows represent clusters. On demand detailed information are presented (Assent, 2007). (Are there only 1D clusters shown?)

Subspace Clustering: Visualization



3D overview visualization with rotation/zooming for exploration (From: OpenSubspace).

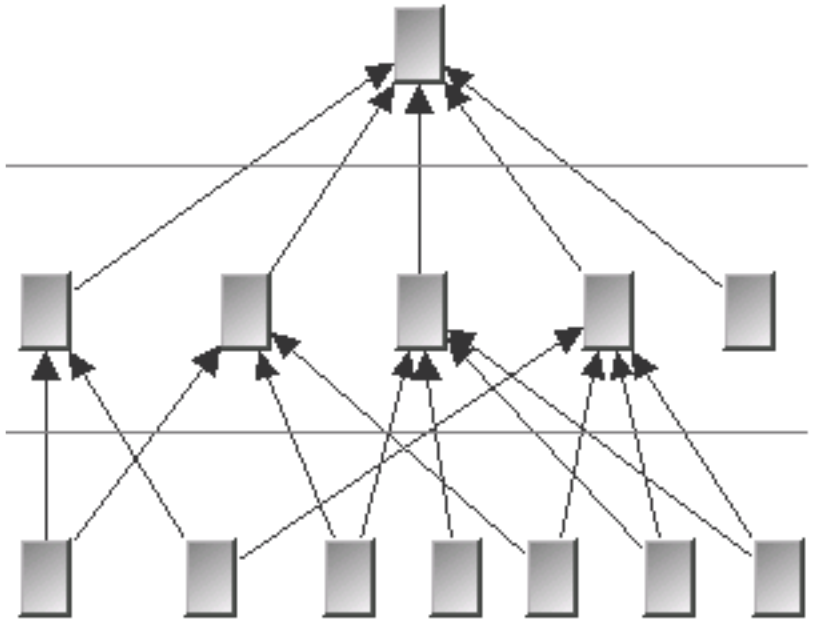
Subspace Clustering



A general workflow for subspace search, clustering and exploration (From: Tatu, 2012).

As preprocessing, often outliers are removed and datasets with missing values. SURFING, e.g., is very sensitive to missing values, but more robust to outliers.

Subspace Clustering: Visualization



Relation between different subspaces displayed as an overview.

Arrows represent subset relations w.r.t. dimensions involved.

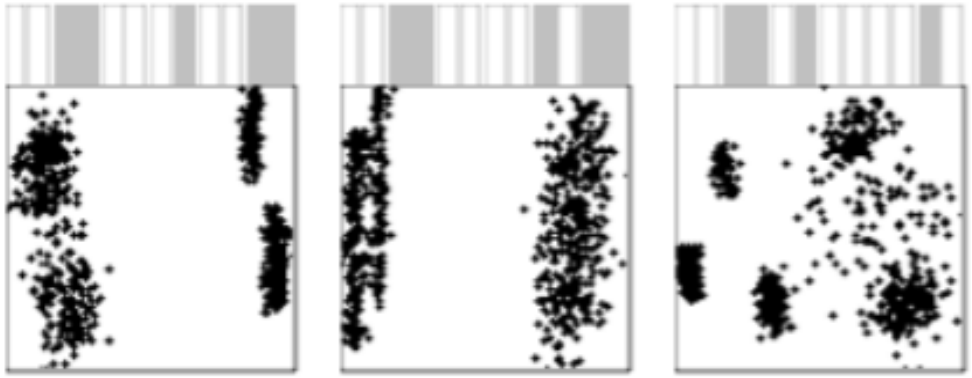
The overview serves to select interesting subspaces.

Note that this is a directed acyclic graph, not a tree (multiple „inheritance“) (From: Achtert, 2007)

Subspace Clustering: Visualization

- Relation between different subspaces by means of graphs and trees does not scale well.
- For larger datasets: „we found that this rarely provides interesting insights and makes the visualization too cluttered“ (Tatu, 2012)

Subspace Clustering: Visualization



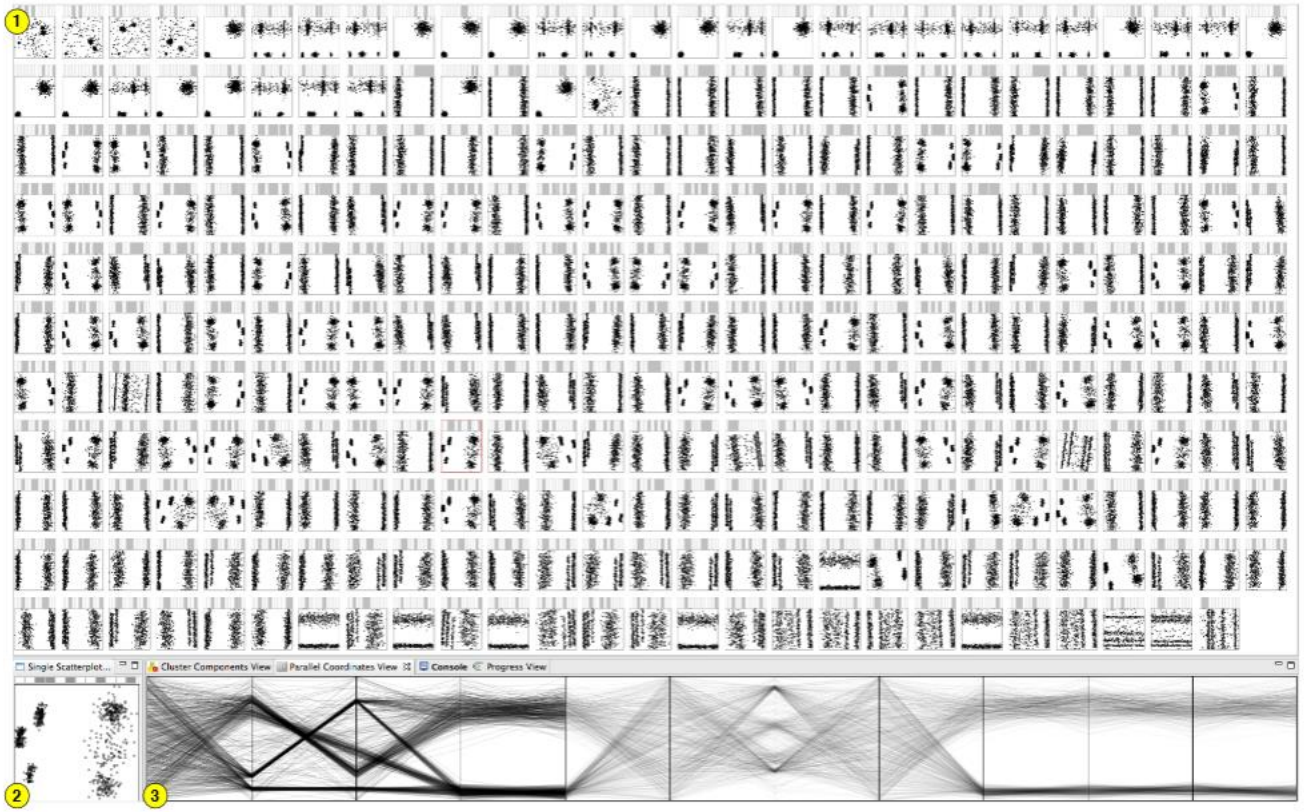
Clusters in three subspaces of a 12D space. The bar on top („*dimension glyph*“) indicates the involved dimensions.

Right view: cluster in 4D: left and middle view 5D clusters.

Data are projected to a 2D scatterplot with multi-dimensional scaling (From: Tatu, 2012)

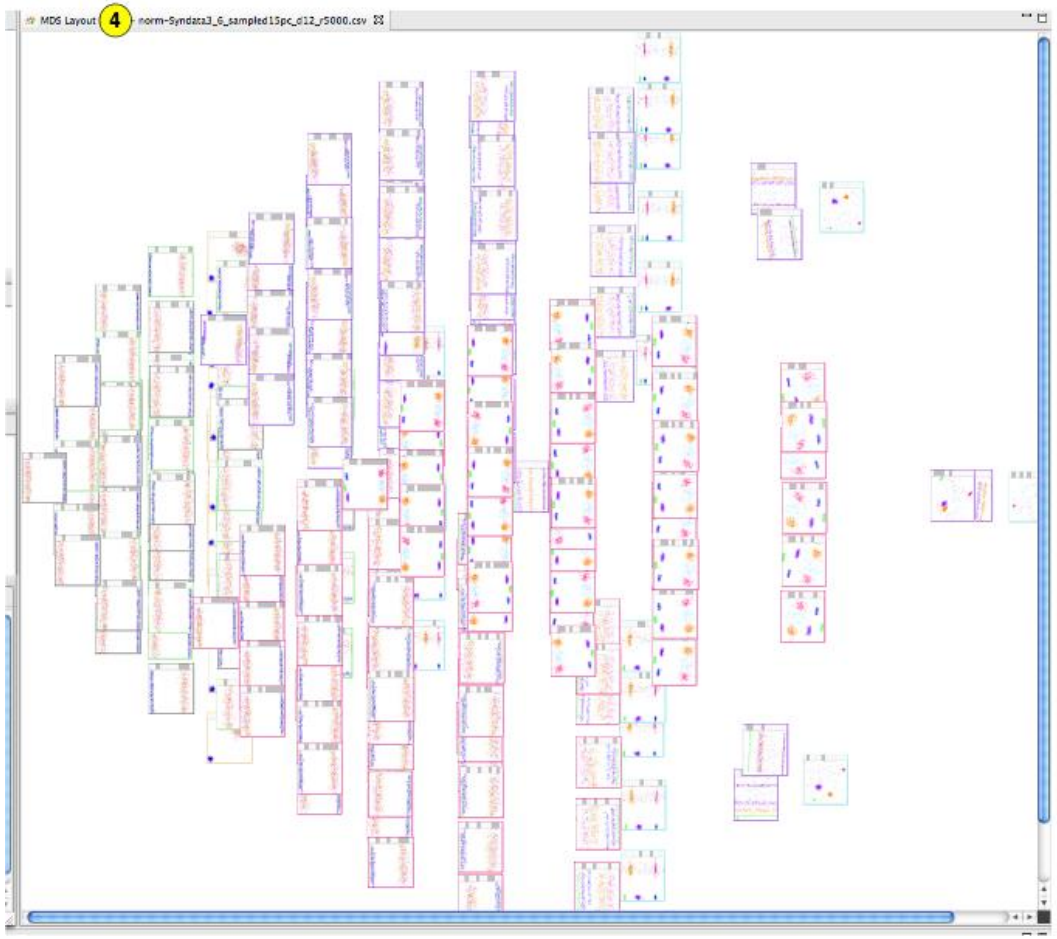
Interaction: Based on his domain knowledge, the user may add/remove dimensions.

Subspace Clustering: Visualization



A 12D space was searched (with SURFING) and 295 subspaces were identified. Subspaces are shown in scatterplots. A single scatter-plot is selected and enlarged (2) and the data of that subspace are also shown in parallel coordinates (3) (with the subspace dimensions and clusters emphasized) (From: Tatu, 2012).

Subspace Clustering: Visualization



Similarity view related to the involved dimensions (spatial proximity indicates high similarity) (From: Tatu, 2012).

For a real application, users should see visited sub-spaces and should be able to record results/findings.

Subspace Clustering: Visualization

Interaction with subspace visualizations

- Lasso interaction to select interesting portions
- Brushing to other (similar) subspaces
- Detail information for the involved dimensions after selection, e.g. histogramm, mean, std. dev.
- Relevant results should be stored (bookmark function)

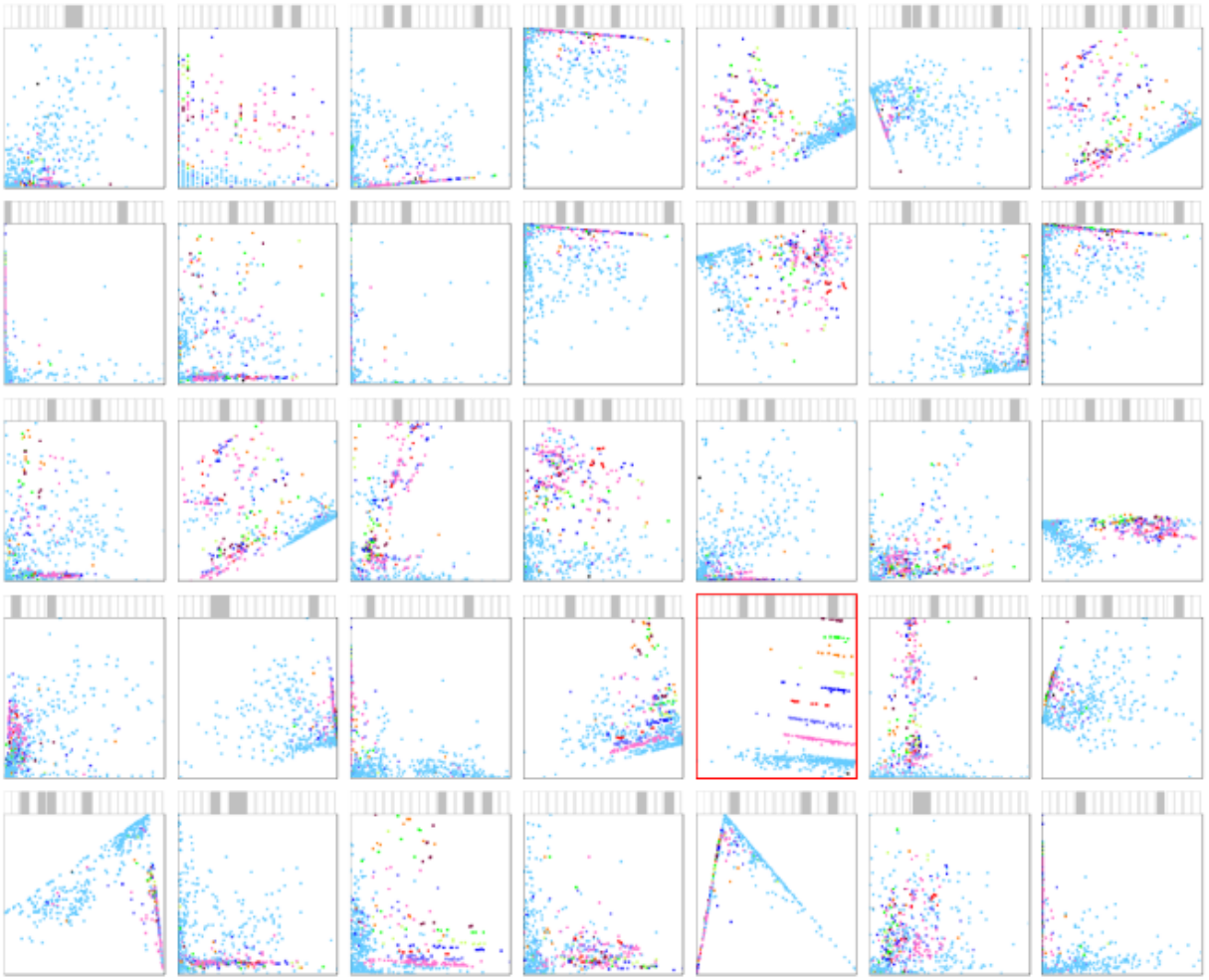
Analysis of clusters:

- The similarity (overlap) of clusters can be analyzed with measures for set comparison, e.g. Dice coefficient and Jaccard Index (measures often used to compare segmentation results).

Application

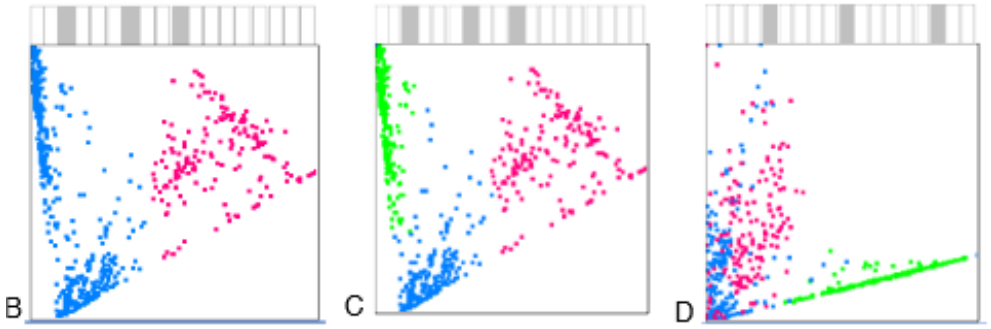
- USDA Food Composition Dataset (Link)
- After preprocessing 722 records (food) with 18 dimensions (indicating which minerals, vitamins, … are part of typical food)
- SURFING produced 216 interesting subspaces (from potentially ~ 262.000)
- Distributions in the high-ranked subspaces often highly skewed
- High redundancy among the subspaces

Application



High-ranked subspaces in a food dataset. The selected subspace has an interesting structure
(From: Tatu, 2012)

Application

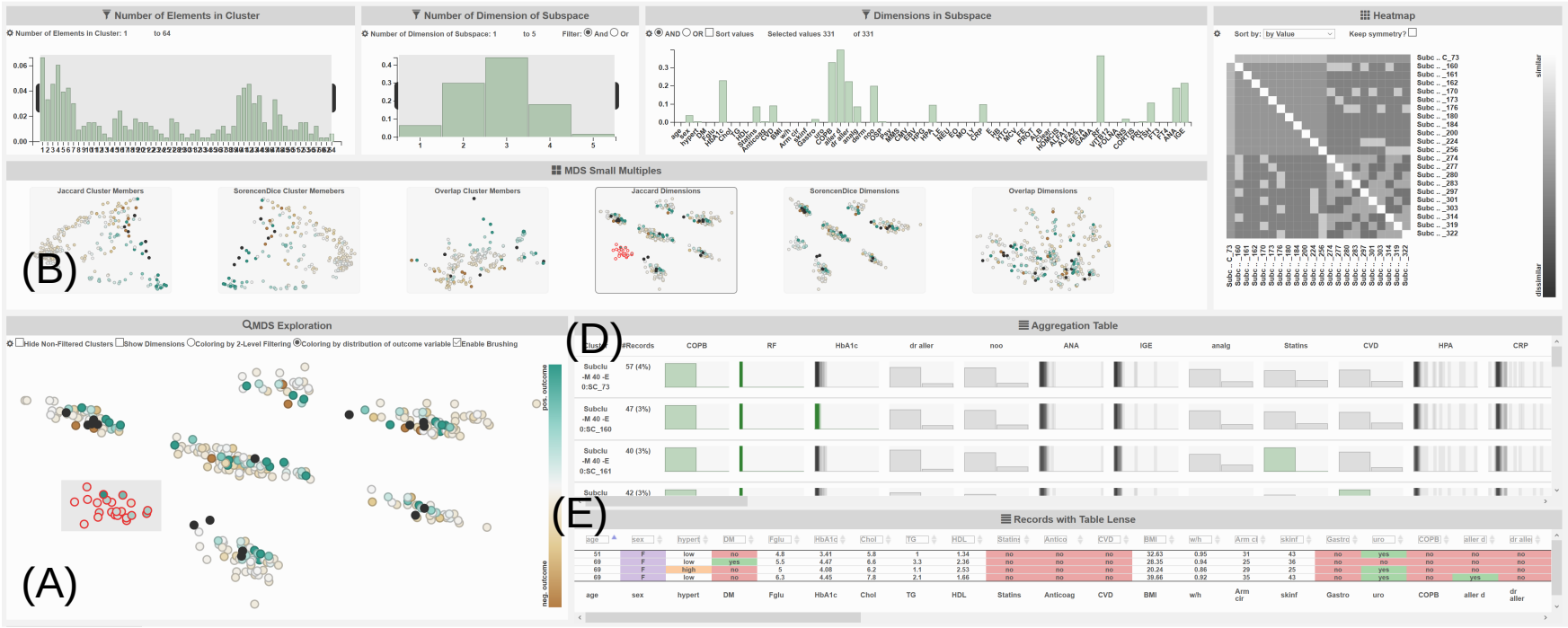


Three subspace clusters from the food dataset:

- Food with carbohydrate, lipid, protein (left)
- An additional cluster marked in the same space (middle)
- Fiber (Ballaststoffe), protein, Vitamin D

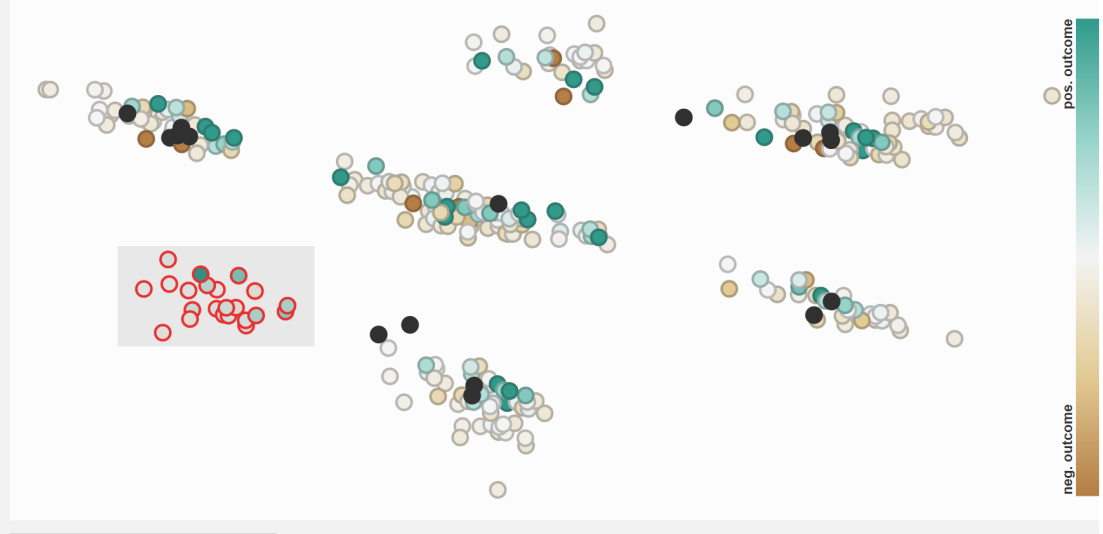
(From: Tatu, 2012)

Subspace Clustering: SubVis



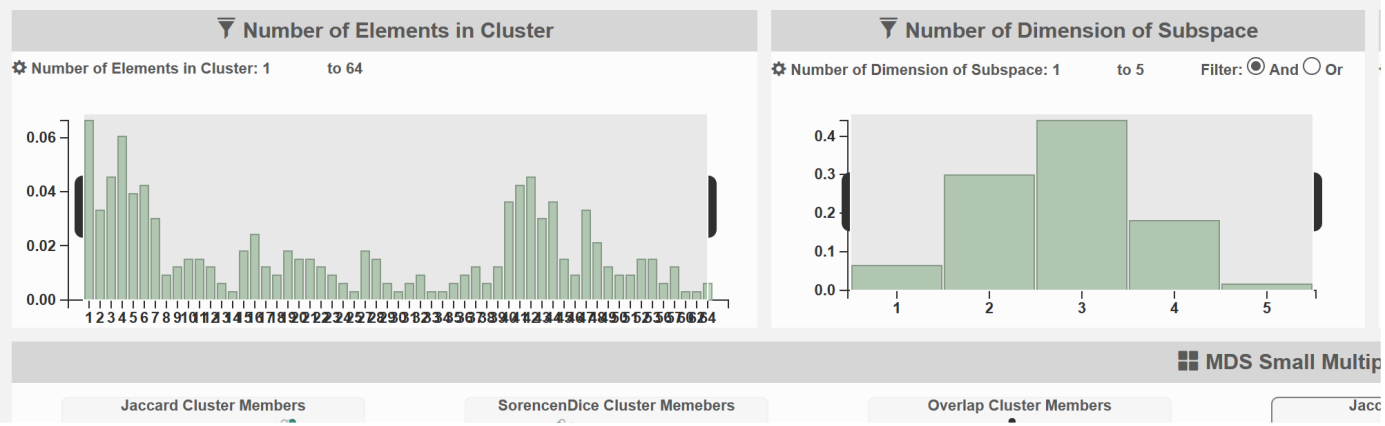
Comprehensive visualization of clustering results: The overview of sub-space clusters bottom left; MDS projection in 2D (A). Small multiples represent results with different distance functions (B). On top: Distribution properties of subspaces (C). Aggregation tables with aggregated members (D) and table lens for details (E). (From: Hund, 2015)

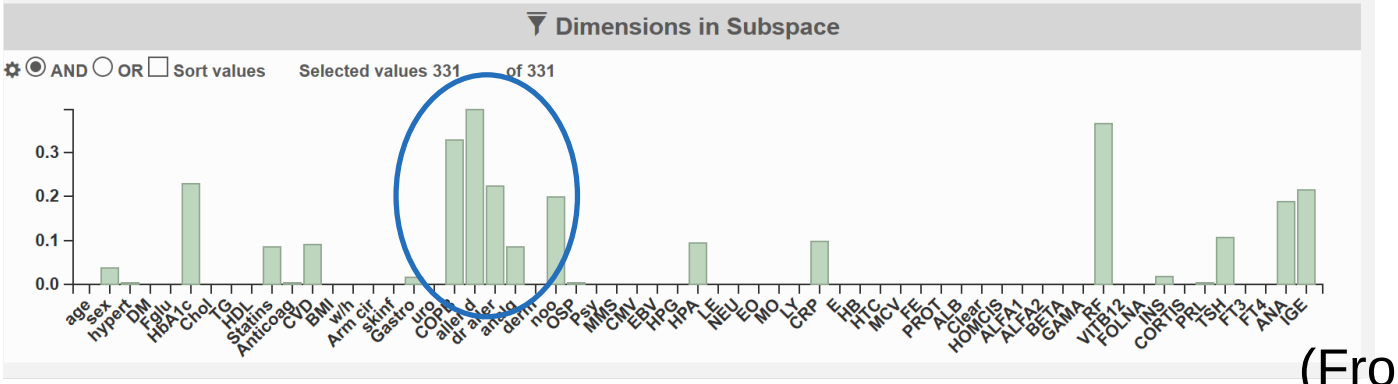
Subspace Clustering: SubVis



MDS overview projection. One cluster is selected. Color-scale very application-specific. Patients with negative or posititve outcome of vaccination for flu (Grippe-Impfung). Coloring may be adjusted to reflect number of cluster members, dimensions (From: Hund, 2015)

Subspace Clustering: SubVis



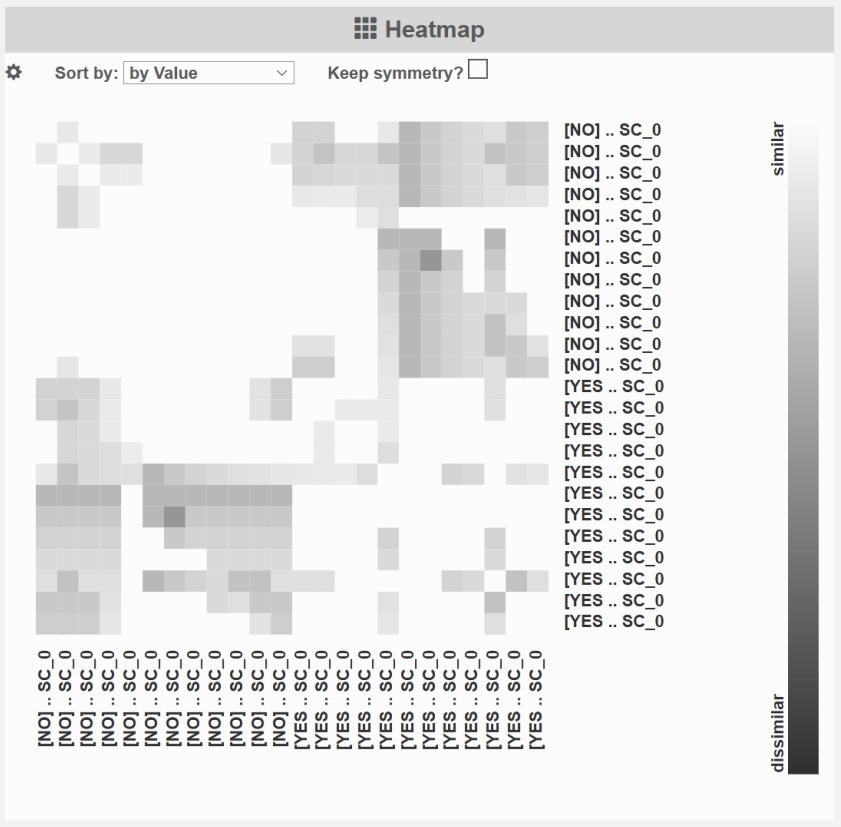


(From: Hund, 2015).

Data for 93 patients were analyzed, only 71 had complete records → the others were removed. Result of Proclus: 64 subspace clusters with 1 to 5 dimensions.

Information related to allergies was involved in most clusters.

Subspace Clustering: SubVis



This heatmap represents how similar the detected subspaces are (w.r.t. overlapping dimensions). Only a few are dissimilar from others (From: Hund, 2015).

Subspace Clustering: ClustNails

Tatu, 2012, Tsinghua Science and Technology Journal

- For a subspace cluster, each dimension that contributes has a weight

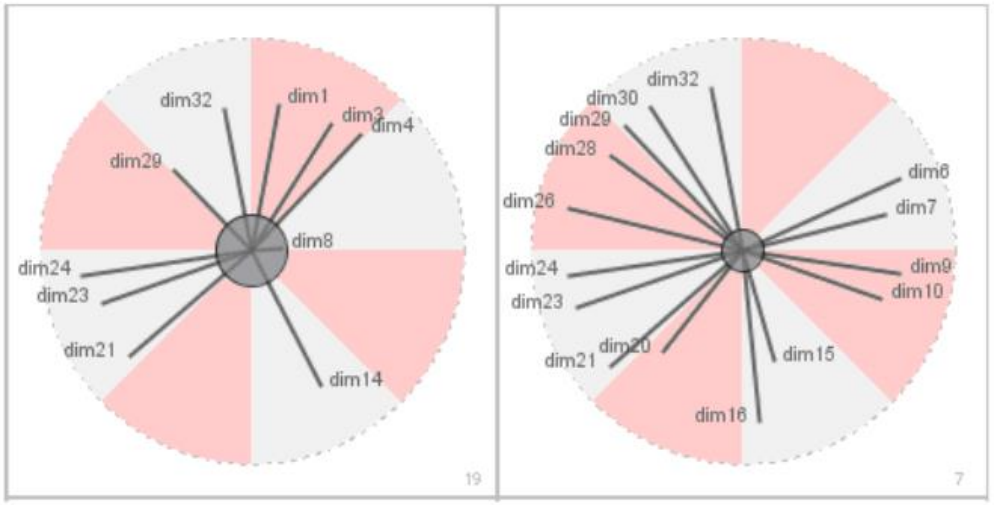
$$w_k^m = \frac{\sum_{x_i^m \in X_k^m} |x_i^m - c_k^m|}{|X_k|}$$

Where $c^m{}_k$ is the center of points in dimension m and $x^m{}_i$ is the value in dimension m and $|X_k|$ is the number of points in this cluster.

- Thus, dimensions with high variance have a high weight since they contribute strongly to the separation.

Subspace Clustering: ClustNails

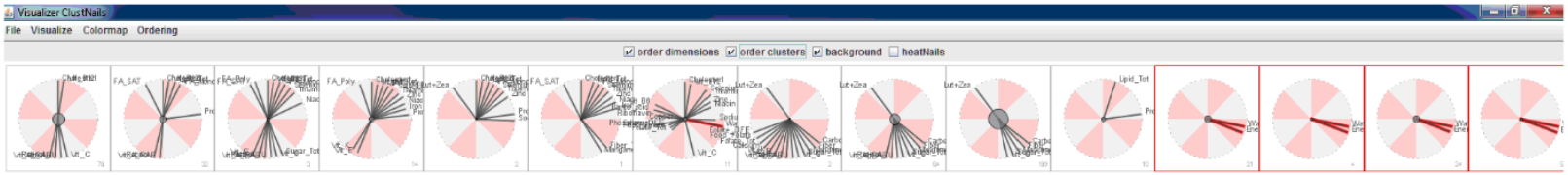
- Subspace clusters are visualized along with the weights of their dimensions.
- The radial visualization is designed such that clusters are comparable w.r.t. involved dimensions.



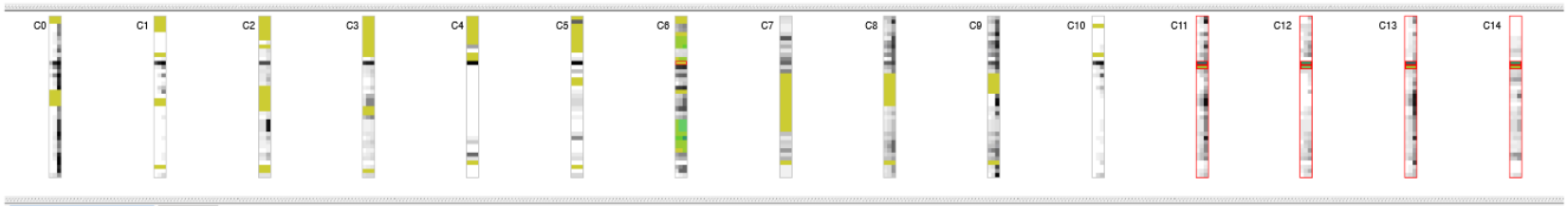
(From: Tatu, 2012)

The length of the spikes („nails“) represents the weights per dimension. Only contributing dimensions are labeled. The background colors serve the comparability.

Subspace Clustering: ClustNails

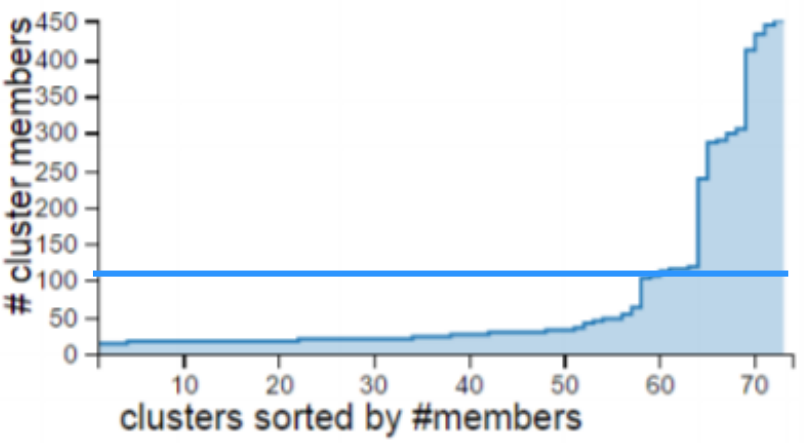


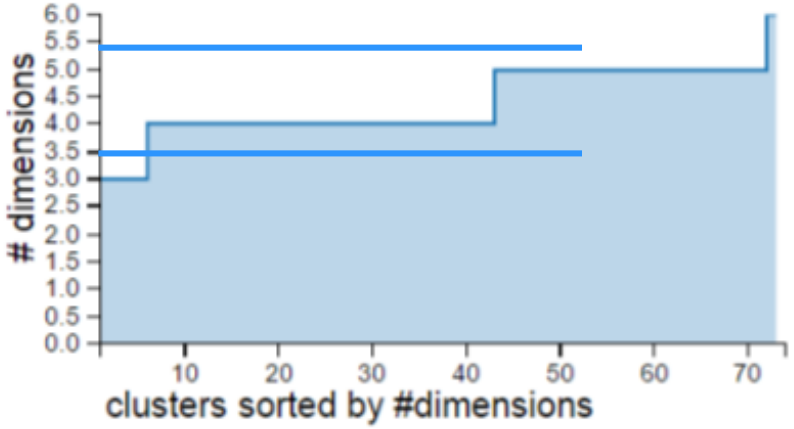
This view (inspired by Star coordinates) serves as an overview of the subspace clusters. It is combined with several detailed views for selected clusters.



At the second level, a histogram view represents the distributions in the relevant dimensions. (From: Tatu, 2012)

Subspace Clustering: Visual Quality



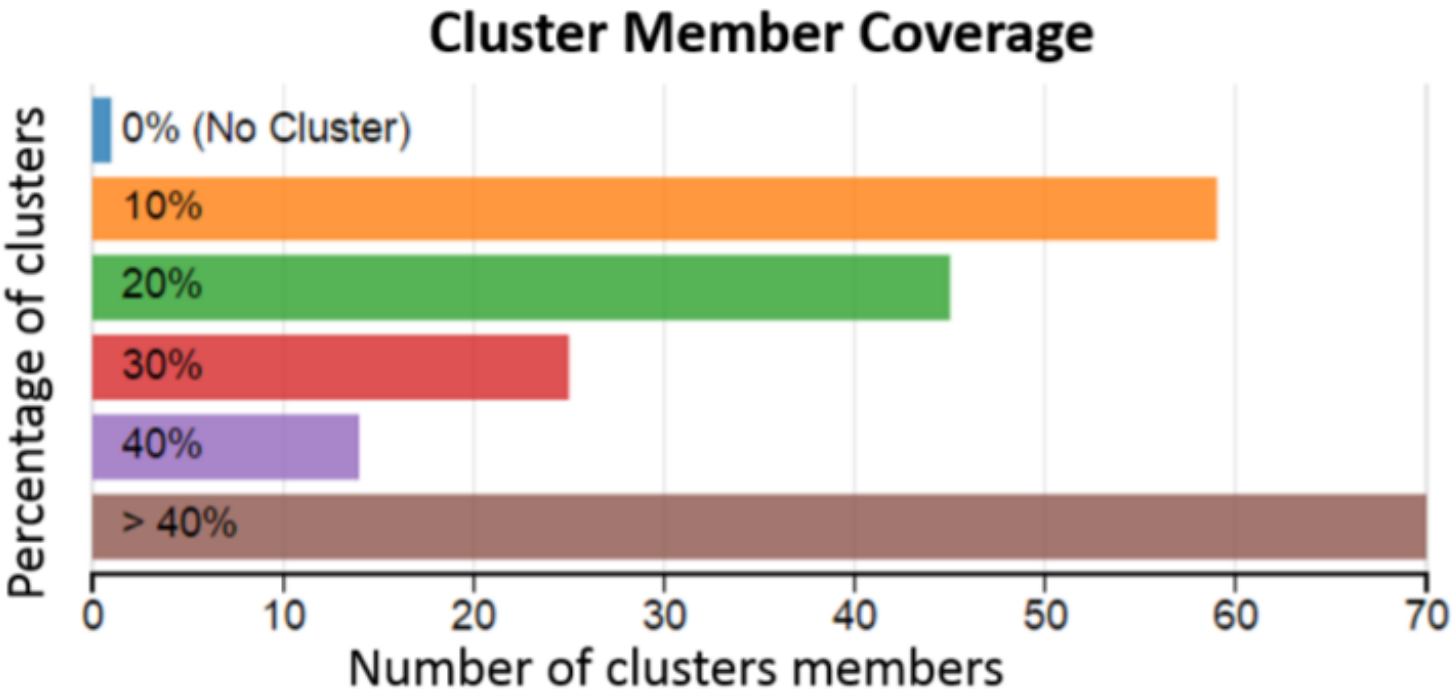


As a first overview, the distribution of cluster members and cluster dimensions per cluster is shown

(From: Hund, 2016).

The histogram views can be combined with interaction: filter for clusters with a minimum number of objects or a certain range of dimensions.

Subspace Clustering: Visual Quality

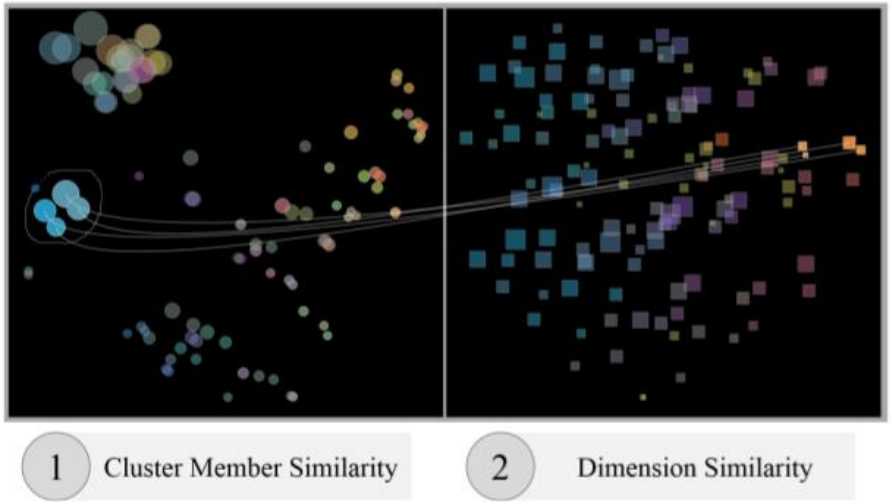


All objects are part of at least one cluster. Most objects are part of <10% of the clusters or >40% of the clusters (From: Hund, 2016).

Subspace Clustering: Visual Quality

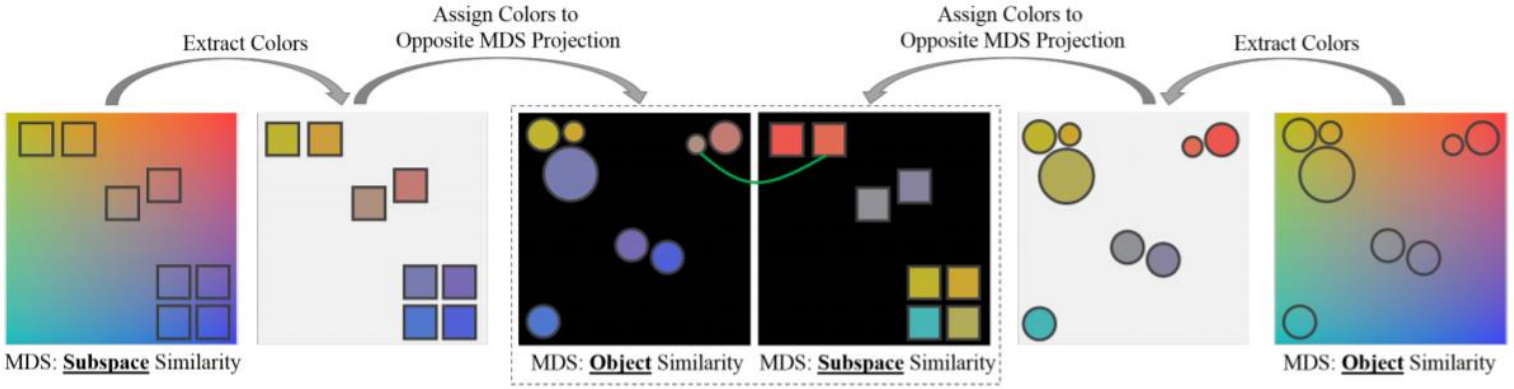
To assess redundancy in subspace clustering, both dimension similarity and cluster member similarity is essential.

Display both features and support links between them (Hund, 2016).



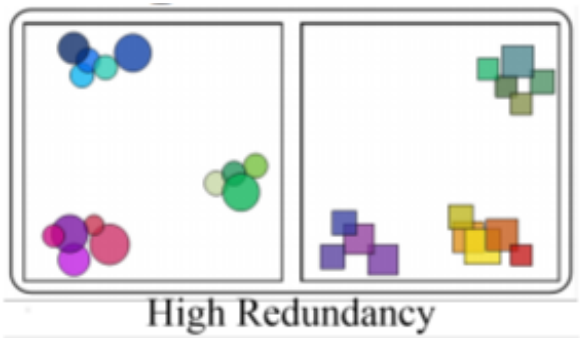
Overview of 132 subspace clusters. Two linked MDS projections. Colors indicate similarity in the opposite view. Thus, for the selected clusters (left) also the dimensions highly overlap – these clusters are truly redundant (From: Hund, 2016).

Subspace Clustering: Visual Quality

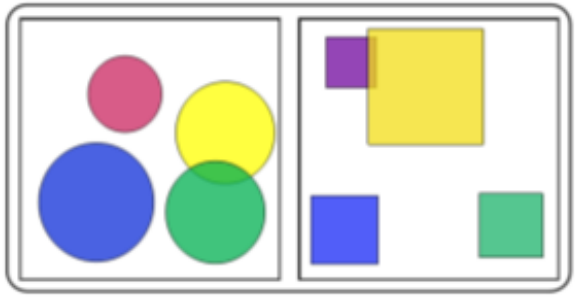


A perceptually motivated 2D color space is employed to encode object and dimension similarity. The primary criterion is proximity and the secondary criterion similarity of colors (From: Hund, 2016).

Subspace Clustering: Visual Quality



Subspace clusters with high and low redundancy (w.r.t. dimensions and instances).

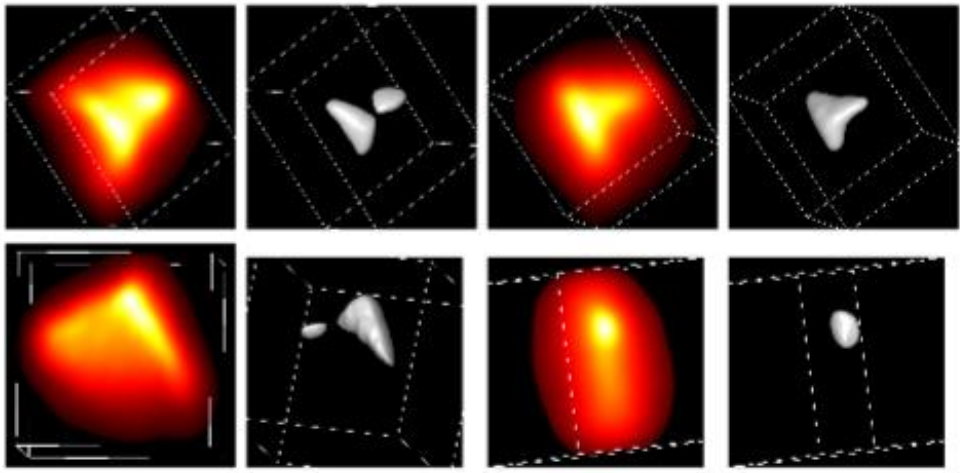


Big (not compact) clusters

(From: Hund, 2016)

Subspace Clustering: 3D Visualization

- Only one publication employs 3D (volume) rendering to reveal results of subspace clustering.
- It is based on astronomy data (galaxy datasets)



Top-ranked subspaces: Top row renderings of 3D subspaces. Bottom row: A 5D subspace is transformed in 3D via PCA and rendered. Attributes relate to color components and log of the mass of galaxies (From: Ferdosi, 2010)

Subspace Clustering: Evaluation

Datasets:

- Artificial data, e.g. generated data with some high-density areas and random noise (Tatu, 2012, used a 12D dataset with two 6D Gaussian clusters).

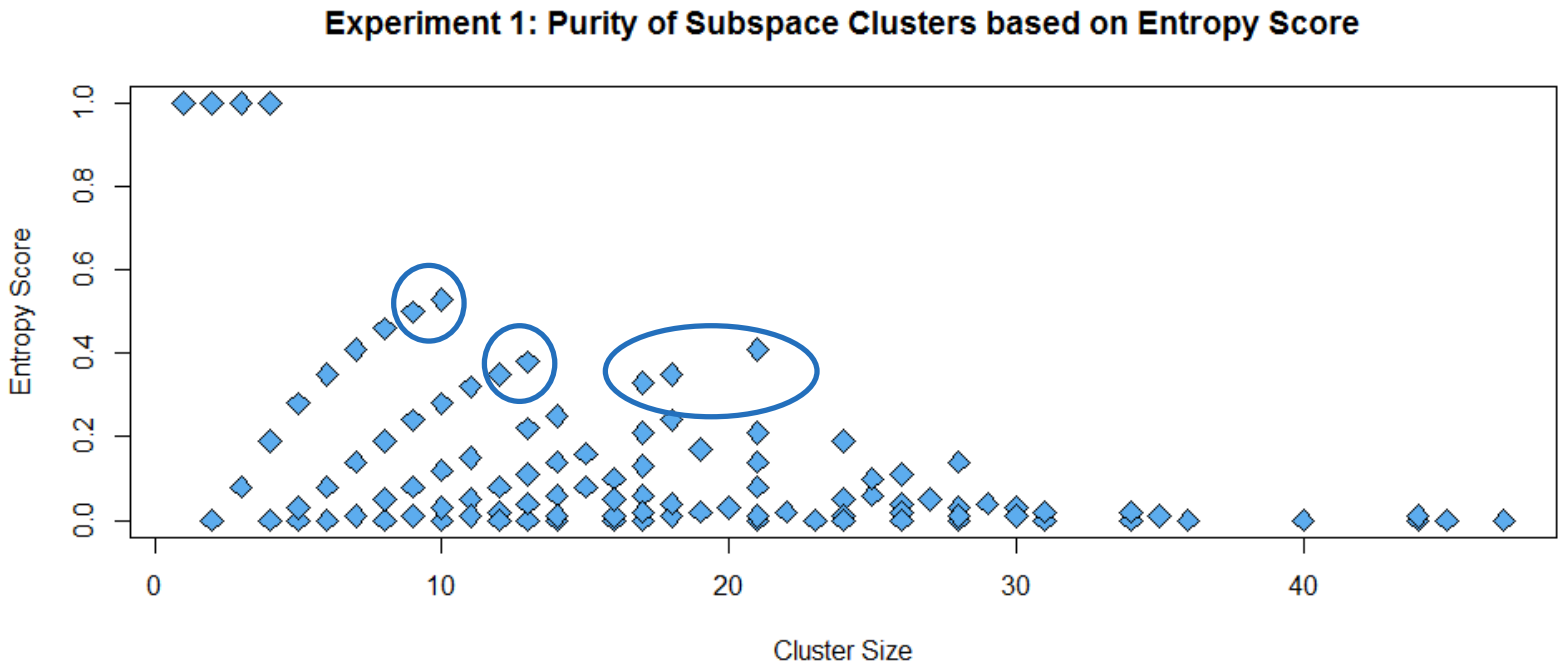
  → Check whether algorithms return known clusters
- Real-World data, e.g. from the machine learning repository, as benchmark data
- Real world data from your customer/research partner to understand the influence on insight generation.

A combination is most interesting and convincing.

Subspace Clustering: Evaluation

- Subspace clustering is based on various parameters. Evaluation often serves to understand whether any parameter combination is better than others or best.
- Clustering evaluation criteria, e.g. Purity, are relevant for subspace clusters as well.
- Subspace clustering may return some excellent, some moderate and some bad clusters.
- Probably not all clusters are very good with one parameter combination.
- Visualization of such measures improves evaluation.

Subspace Clustering: Evaluation



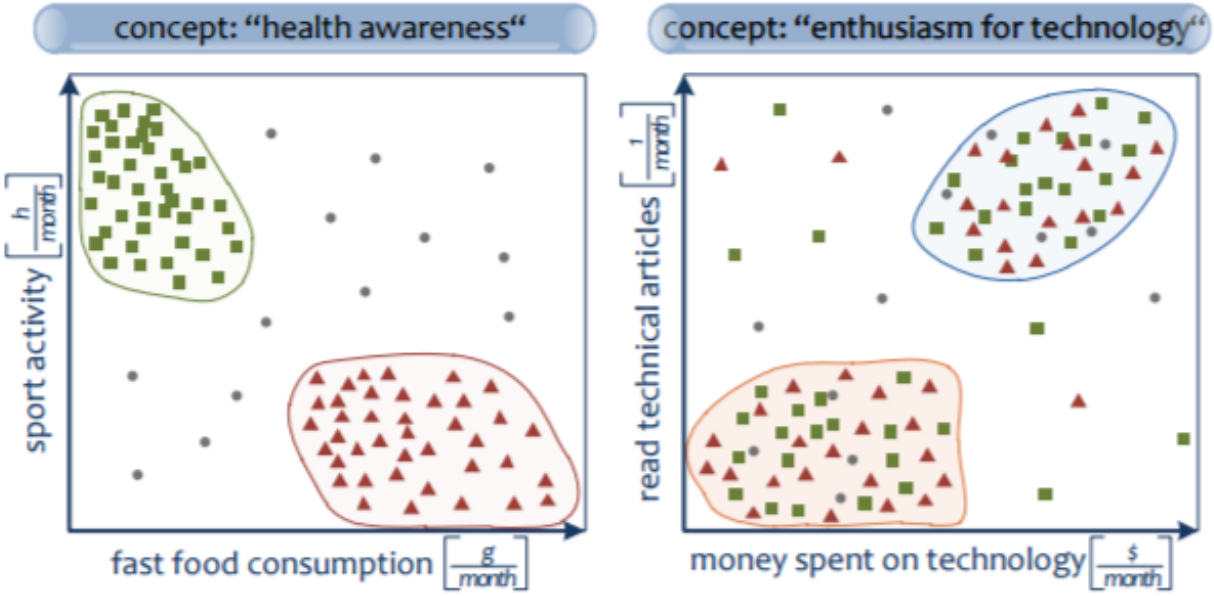
Evaluation of clustering results. Some clusters are not too small and have a good entropy score (From: Hund, 2015).
Application to patient data where vaccination was sometimes successful, sometimes not.

From Subspace Clusters to Concepts

Subspace clusters are determined automatically.

Relevant *concepts* - represented by these clusters - need interpretation from an expert.

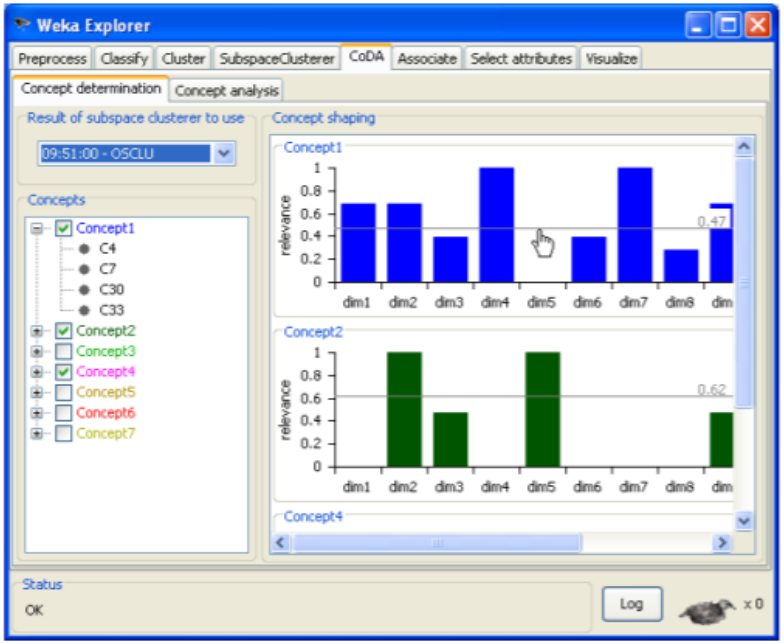
A concept is based on a *subset of the subclusters*, it has *significant dimensions* and a *label* (From: Günnemann, 2010).



From Subspace Clusters to Concepts

Concept discovery and interpretation may be supported by (see Günnemann, 2010).

- Grouping (a concept is represented by clusters with equal or very similar dimensions) → clustering on the clusters
- Ordering dimensions w.r.t. significance



A number of concept candidates was found – each compring a few clusters. The significance of the dimensions is shown supporting the user in setting a threshold (From: Günnemann, 2010).

Subspace Clustering: Evaluation

Datasets.

- There is no reliable ground truth for real world data.
- Often, a combination of synthetic data (hidden clusters and various levels of noise) and real-world data (primarily publicly available data) is used.
- For a specific application, representative datasets of this domain need to be employed, e.g. customer profile data and the clustering results need to be discussed with domain experts w.r.t. cluster quality and **relevance for their decisions**.
- „Representative“ relates to the typical distribution, to the number of objects and dimensions (no „toy“ data).

Summary

- Subspace clustering enables explorative analysis of HD data from multiple perspective
- Most applications are related to rather low HD data (a few dozen dimensions)
- Subspace clustering involves the search for clusterable subspaces, the clustering itself and the presentation of subspace clusters involving the related dimensions.
- The selection of an appropriate algorithm is more complex than in global clustering, since more parameters are involved.

Summary

- According to an *interestingness* measure, the most interesting subspaces are often very similar → clustering on these subspaces to choose a representative.
- Subspace clustering involves complex workflow. → Comprehensive systems with a good trade-off between flexibility and guidance are essential.
- Visualization techniques are solid; not very advanced. How to use transparency, line styles etc. to reveal cluster properties is not clear.
- Clustering identifies one interesting aspect of data: dense areas. Other interesting aspects (correlated dimensions, outliers, other patterns are not found).

Limitations

- Subspace clustering enables the exploration of more dimensions compared to global clustering.
- For truly high-dimensional data, e.g. the > 100 dimensions of patient data, unsupervised subspace clustering produces an enormous amount of information that cannot be feasibly explored.
- Clustering does not employ background knowledge of the user w.r.t. domain and dataset.
- Constrained-based clusterings are more scaleable (see e.g. Hielscher, 2014). Experts specify that
  - some objects should belong to the same cluster (must-link)
  - others must not belong to the same cluster (must-not)

Limitations

- Constraint-based clusters often employ few constraints (e.g. for 1% of the instances) and improve the accuracy of grouping all constraints.
- Must-link constraints are transitive: If d_i must link to d_j and d_j must link to $d_k \rightarrow d_i$ must link to d_k (Wagstaff, 2001)
- In finance data, analysts may label instances of true credit card frauds. In medicine, analysts may label pathologic situations (back pain).
- Constraints may be considered hard – enforcing that all constraints are met or soft, more as hints.

References

Elke Achtert, Christian Böhm, Hans-Peter Kriegel, et al.: „Detection and Visualization of Subspace Cluster Hierarchies“. *Proc. of DASFAA* 2007: 152-163

C. C. Aggarwal, J. L. Wolf, P. S. Yu, C. Procopiuc, J. S. Park. 1999. „Fast algorithms for projected clustering“, *Proc. of ACM SIGMOD Conf. on Management of Data*, pp. 61-72

R. Agrawal, Gehrke J., Gunopulos D., Raghavan P.: „Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications”, *Proc. ACM SIGMOD Conf. on Management of Data,* 1998, pp. 94-10

I. Assent, R. Krieger, E. Müller, T. Seidl: VISA: visual subspace clustering analysis. *SIGKDD Explorations* 9(2): 5-12 (2007)

C. Baumgartner, C. Plant, K. Kailing, H.-P. Kriegel, and P. Kröger. Subspace selection for clustering high-dimensional data. *Proc. of IEEE Conference on Data Mining*, pp. 11–18, 2004

Y. Cheng and G. M. Church. “Biclustering of expression data”. *In International Conference on Intelligent Systems for Molecular Biology*, pp. 93-103, 2000

Ferdosi B, Buddelmeijer H, Trager S, et al. Finding and visualizing relevant subspaces for clustering high-dimensional astronomical data using connected morphological operators. In: *Proc. of IEEE Visual Analytics Science and Technology (VAST*), 2010: 35-42

S. Günnemann, I. Färber, H. Kremer, T. Seidl. „CoDA: Interactive Cluster Based Concept Discovery”, *Proc. of VLDB Endowment (**PVLDB**)* 3(2): 1633-1636 (2010)

T. Hielscher, M. Spiliopoulou, H. Völzke, J.-P. Kühn: Using Participant Similarity for the Classification of Epidemiological Data on Hepatic Steatosis. *Proc. of CBMS* 2014: 1-7

M. Hund, W. Sturm, T. Schreck, T. Ullrich, D. A. Keim, L. Majnaric, A. Holzinger (2015). „Analysis of Patient Groups and Immunization Results Based on Subspace Clustering“. *Brain Informatics*, pp. 358-368

References (II)

M Hund, I Färber, M Behrisch, A Tatu, T Schreck, DA Keim, T Seidl (2016). „Visual Quality Assessment of Subspace Clusterings", *Proc. of KDD Workshop on Interactive Data Exploration and Analytics (IDEA'16)*

Kailing, K. Kriegel, H.P., Kröger, P. "Density-Connected subspace clustering for high-dimensional data", In *Proc. of SIAM International Conference on Data Mining (SDM*), 2004

H.-P. Kriegel, P. Kröger, A. Zimek. „Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering", *ACM Trans. Knowl. Discov. Data* 3, 1, Article 1 (2009), 58 pages

E. Müller, S. Günnemann, I. Assent, and T. Seidl. Evaluating clustering in subspace projections of high dimensional data. *Proc. VLDB Endow.* 2, 1 (2009), 1270-1281

E. Müller, I. Assent, and T. Seidl. HSM: Heterogeneous Subspace Mining in High Dimensional Data. *Proc. of Scientific and Statistical Database Management*, (2009), 497-516

E. Müller, S. Günnemann, I. Färber, T. Seidl. "Discovering Multiple Clustering Solutions: Grouping Ob-jects in Different Views of the Data", *Tutorial at the International Conference on Machine Learning*, 2013

L. Parsons, E. Haque, H. Liu: Subspace clustering for high dimensional data: a review. *SIGKDD Explorations* 6(1): 90-105 (2004)

M. Sips, B. Neubert, J. P. Lewis, P. Hanrahan: Selecting good views of high-dimensional data using class consistency. *Comput. Graph. Forum* 28(3): 831-838 (2009)

References (III)

A. Tatu, F. Maass, I. Färber, E. Bertini, T. Schreck, T. Seidl, D. A. Keim: Subspace search and visualization to make sense of alternative clusterings in high-dimensional data. *IEEE VAST* 2012: 63-72

A. Tatu, G. Albuquerque, M. Eisemann, J. Schneidewind, H. Theisel, M. A. Magnor, D. A. Keim: Combining automated analysis and visualization techniques for effective exploration of high-dimen-sional data. *IEEE VAST* 2009: 59-66

A.Tatu, L. Zhang, E. Bertini, T. Schreck, D. Keim, S. Bremm, T. von Landesberger, "ClustNails: Visual analysis of subspace clusters," in *Tsinghua Science and Technology*, vol. 17(4): 419-428, 2012.

S. Vadapalli, K. Karlapalem. 2009. Heidi matrix: nearest neighbor driven high dimensional data visualization. In *Proc. of ACM SIGKDD Workshop on Visual Analytics and Knowledge Discovery* (VAKD '09), 83-92.

K. Wagstaff, C. Cardie, S. Rogers and S. Schroedl (2001). “Constrained K-means Clustering with Background Knowledge”, *Proc. of Machine Learning*, 2001, p. 577–584