Cluster Analysis: Validation, Visualization, Outlier

Outline

- Introduction
- Visualization of Clustering Results
- Validation of Clustering Results

Validation

An analyst may try different clustering techniques and parameterizations. How does she know whether the results are good? What are failures of clustering?

- False negatives. There are essential structures in the data that are not detected.
- False positives. The determined clusters are not meaningful.

One clustering result may contain both false negatives and false positives.

In addition to a global assessment, individual clusters may be analyzed, w.r.t. specific boundaries.

Validation measures consider individual clusters and compose a global score based on (weighted) combinations of individual results.

Validation

If a gold standard is available, e.g., a consensus how experts would manually cluster the data, clustering results are compared to this gold standard.

If no gold standard is available or the gold standard is not convincing, cluster purity measures are used

- Silhoutte coefficient
- Centroid measure
- Grid-based measures

The selection of cluster purity measures is challenging! Depending on the within and between class factors, some measures are not valid.

Validation: Silhouette Coefficient

- For each object o assigned to a cluster A, the distance to all objects a∈A and to all objects b ∈ B (the nearest cluster to o) is computed.
  - dist (o,A), dist (o,B) are the average distances of objects from that cluster to o. This S(o) is computed:

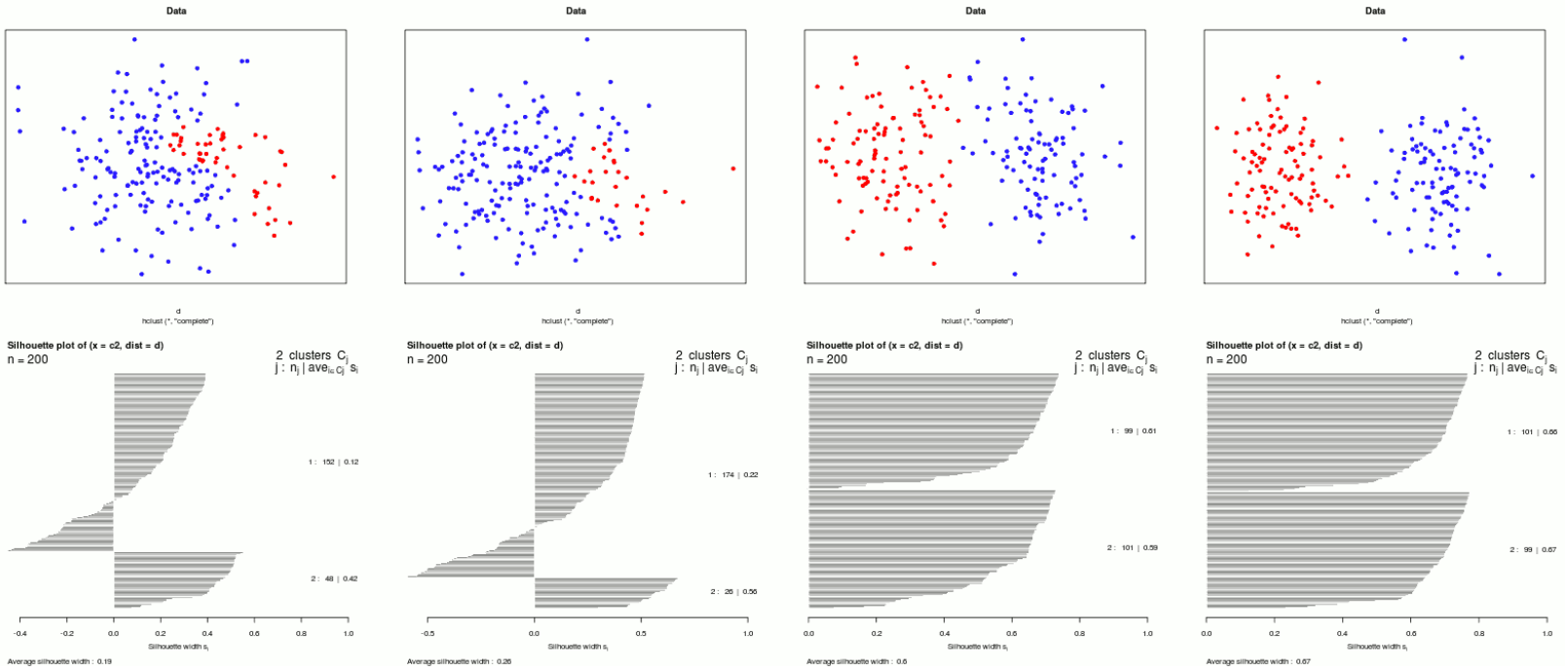
$$S(o) = \begin{cases} 1 \\ \frac{\text{dist}(B,o) - \text{dist}(A,o)}{\max\{\text{dist}(A,o), \text{dist}(B,o)\}} \end{cases}$$

  - S(o) is between -1 and 1. If S(o) ≈ 0; o is between A and B.
  - If S(o) <0, o is closer to B.
  - The silhouette coefficient is defined as follows (with C being the number of clusters):

$$s_C = \frac{1}{n_C} \sum_{o \in C} s(o)$$

Validation: Silhouette Coefficient

- Clustering measures may be used to optimize clustering, e.g. by starting the algorithm again with different parameters or by starting it again with the same parameters, if it is a random algorithm

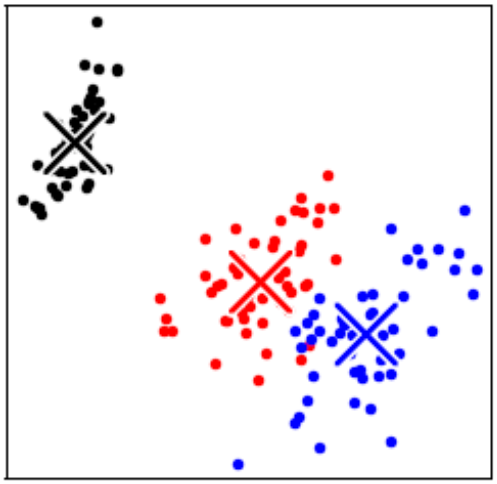


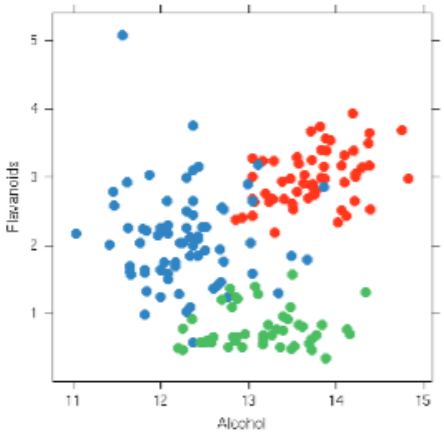
Data, dendrograms and silhouette plots of different data sets. (From Wikimedia, Creative Commons Attribution-Share Alike 3.0 Unported)

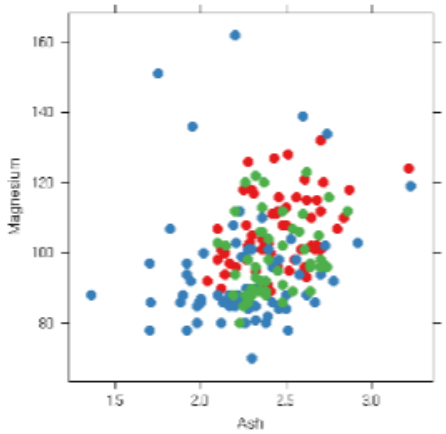
Validation: Centroid-Based Measure

- For all clusters C_i determine a centroid c_i
- For all $points\ p_i \in \{C_i\}$ *determine the distances to all centroids* $c_j\ j \neq i$
- A perfect separation is achieved if the distance from all $p_i \in \{C_i\}$ *to all centroids* $c_j > dist(p_i, c_i)$
- The portion of p_i for which the inequality does not hold indicates the quality of clustering
- The overall centroid measure is determined by summing over all clusters and weighting the result with the size of the cluster $\left| \{C_i\} \right|$ in a scale from 0 to 100

Validation: Centroid-Based Measure







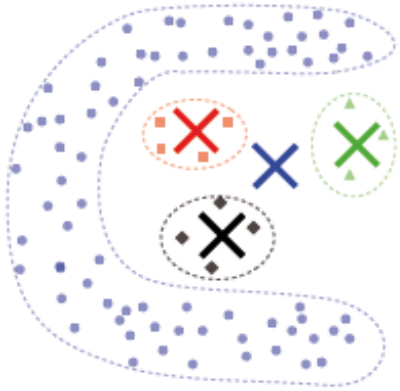
Left: Clustering results in three clusters. Centroids are shown as X (From: Sedlmaier, 2012)

Middle and right: Clustering with good (90) and bad centroid measure (49) (From: Sips, 2009)

Validation: Centroid-Based Measure

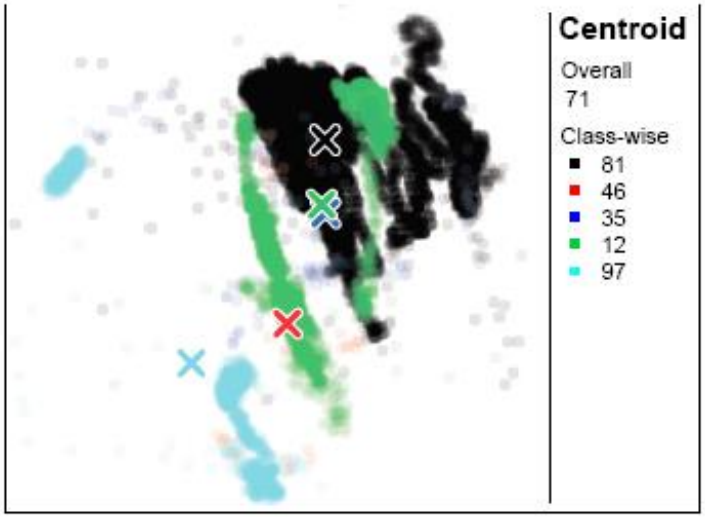
The centroid measure

- produces low values for narrow and curved clusters; high values for convex compact shapes.
- produces low values for split and interwoven (spatially not coherent) clusters.
- is relatively robust against different sizes, densities and numbers of clusters.



The banana-shaped (curvy, narrow) cluster has a centroid outside of the cluster – located between the three other clusters. Although the clustering result is perfect, the centroid measure produces a very low result (From: Sedlmaier, 2012).

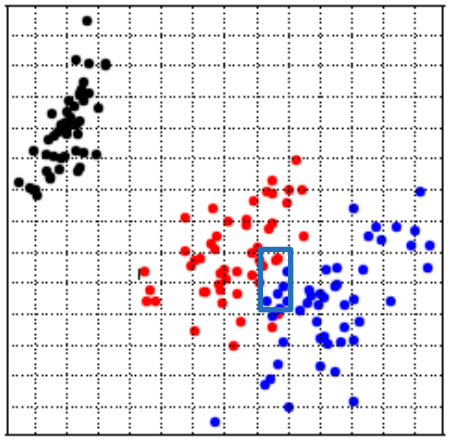
Validation: Centroid-Based Measure



The split green cluster has a very low centroid value altough experts rated it as correct (From: Sedlmaier, 2012)

Validation: Grid-Based Method

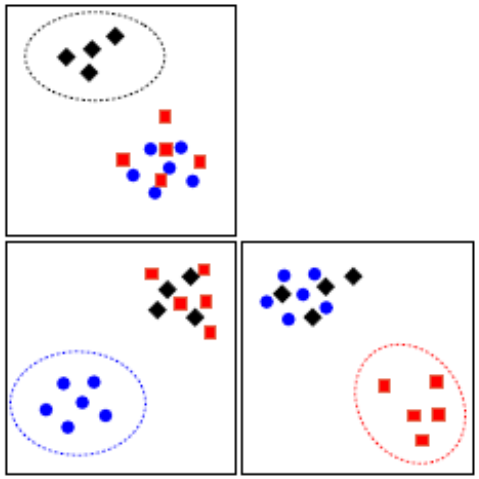
- Grid-based methods overlay a grid on the clustering result and assess whether in the grid cells $cell_{ij}$ are points belonging to one cluster or mixed points of different clusters
- A good clustering has a low portion of cells with mixed points.
- As a suggestion for grid size, for *n* points grid size in 2D should be $\sqrt{n}$ and in 3D $\sqrt[3]{n}$. (Sedlmaier, 2012)
- **Example**: for n=1000, gridsize 31x31 or 10x10x10



Evaluation of clustering with a grid method (From: Sedlmaier, 2012)

Validation

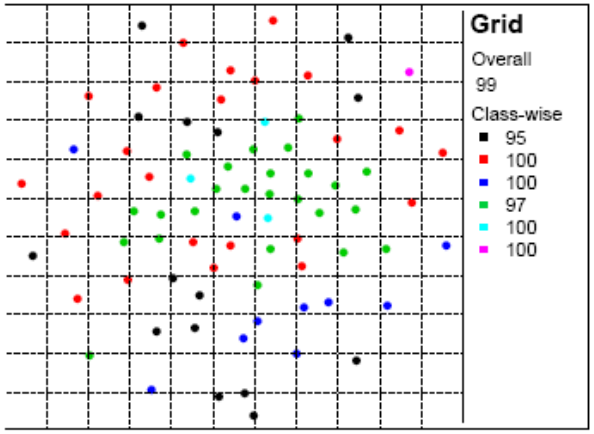
- Centroid- and Grid-based quality can be assessed also for 3D clustering results shown in a 3D scatterplot
- For higher dimensional clusters, scatter plot matrices (SPLOM) are used.
- In SPLOM a good visual separation is achieved if each cluster is at least in one scatterplot clearly separated.



Evaluation of clustering in a SPLOM (From: Sedlmaier, 2012)

Validation

- The grid-based measure is more robust against split clusters and non-convex shapes.
- The grid resolution influences the results. With a too low resolution the results are positive even in case of good clusters.
- Grid size must be chosen such that typically several points are in one cell.



The good grid measure is misleading due to the low grid resolution (From: Sedlmaier, 2012).

Validation

- Validation measures may be used to determine the number of clusters to be detected with k-Means, k-Medoids, fuzzy c-means.
- Thus, for a range of values, e.g. 2-20, validation measures are determined and a parameter *n* is chosen that leads to best validation measures.
- Care is necessary, since validation measures are not independent from the number of clusters; some grow or decrease monotonically with n.
- If for two values n_1 and n_2, the validation measures are very similar, the lower number is preferred.

Visualization of Clustering Results

High-level tasks

- „support an active process of discovery as opposed to passive display“ (Fua, 1999)

Interactive Visualization of clusters serves to (Cao, 2011)

- Interpret clusters
- Evaluate clusters (is the suggested grouping likely to be correct), and to
- Refine clusters (merge or split clusters, add or remove elements, change parameters or even algorithms)

Visualization of Clustering Results

Visualization tasks (in more detail):

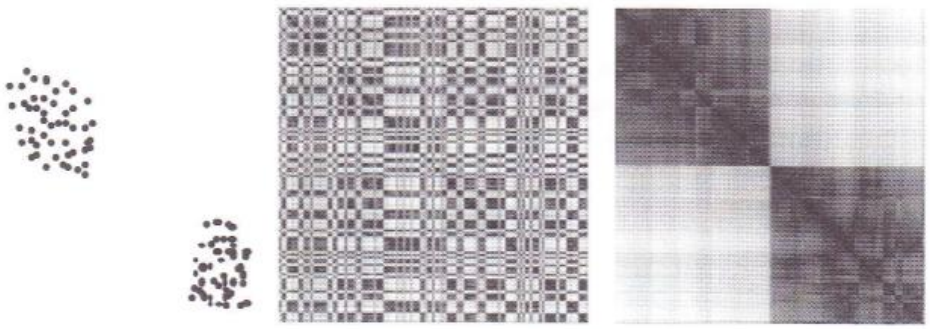
- For 2D data, hard clustering:
  - Show data distribution, provide essential statistical properties of a cluster and indicate per point to which cluster it belongs, show the attributes or a user-selected subset
- For 2D data, fuzzy clustering
  - Show data distribution and indicate per point the probability of belonging to a certain cluster
- Optional additions
  - Display representatives per cluster
  - Local regression lines per cluster
- For HD data, project the points such that distances are roughly preserved after projection in 2D/3D

Visualization of Clustering Results

Visualization Techniques:

- Early Techniques,
- Glyphs,
- Scatterplots (2D, 3D)
- Parallel Coordinates Plot (PC plot)
- Isolines
- Enhanced Dendograms

Early Visualization Techniques

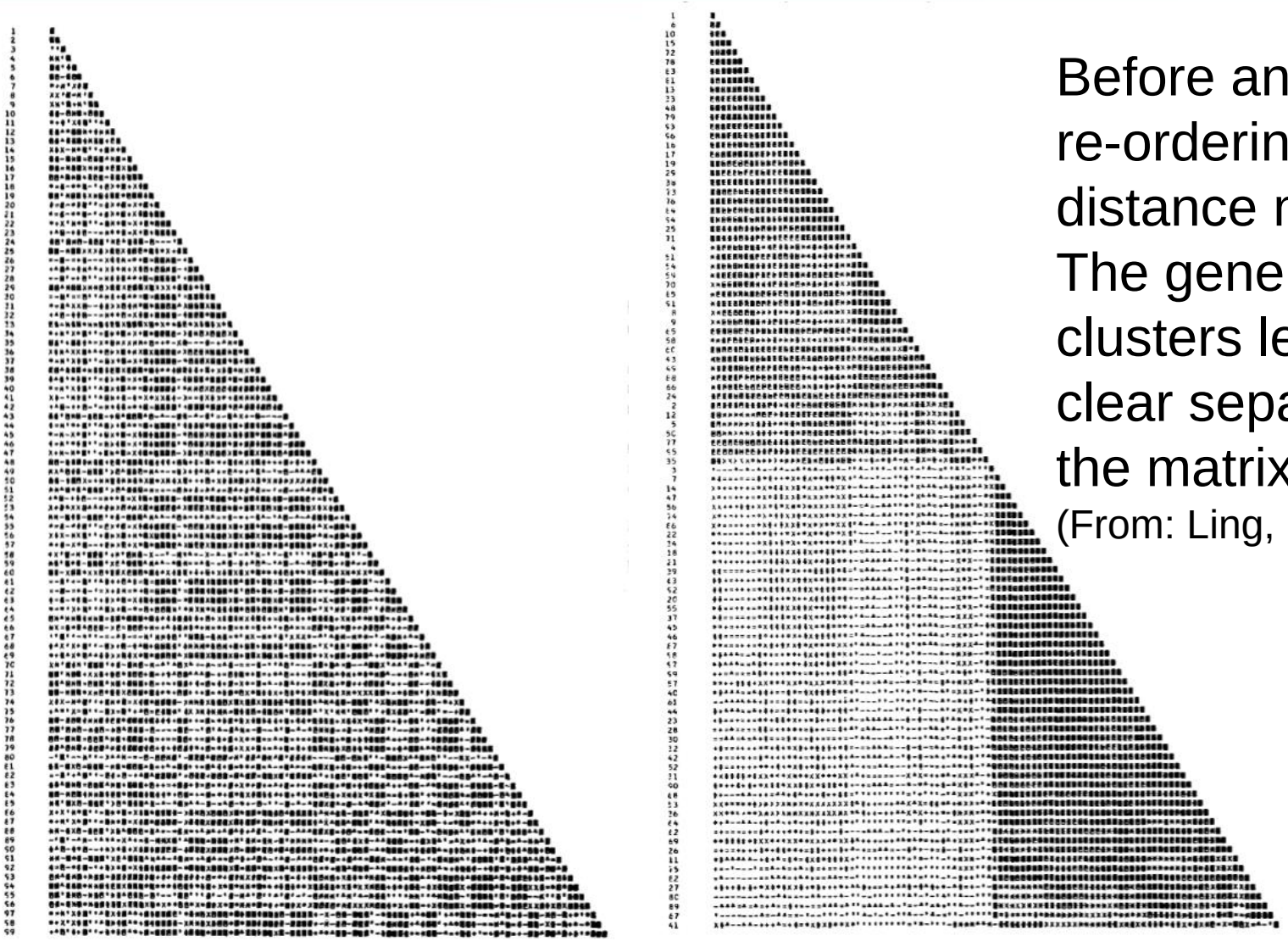


Two clusters were produced. The dis. Matrix is re-ordered accordingly and the clusters are obvious.

(From: Hathaway, 2003)

- The visualization of clustering results for verification of plausibility was treated early ([Rohlfs, 1970], [Ling, 1973]).
- A clustering result is plausible, if the *distance matrix* (based on the chosen distance measure) re-ordered according to cluster membership makes clusters clearly visible (Ling, 1973).
- For visualization, distance was mapped to darkness represented as halftones with alphanumeric signs (a sign becomes dark, if it is overplotted severall times).

Early Visualization Techniques



Before and after re-ordering the distance matrix. The generated two clusters lead to a clear separation in the matrix

(From: Ling, 1973)

The „color scale“ . - : = + ≢ * X X X E E E E E

Early Visualization Techniques

Visualization of Clustering Results: Glyphs

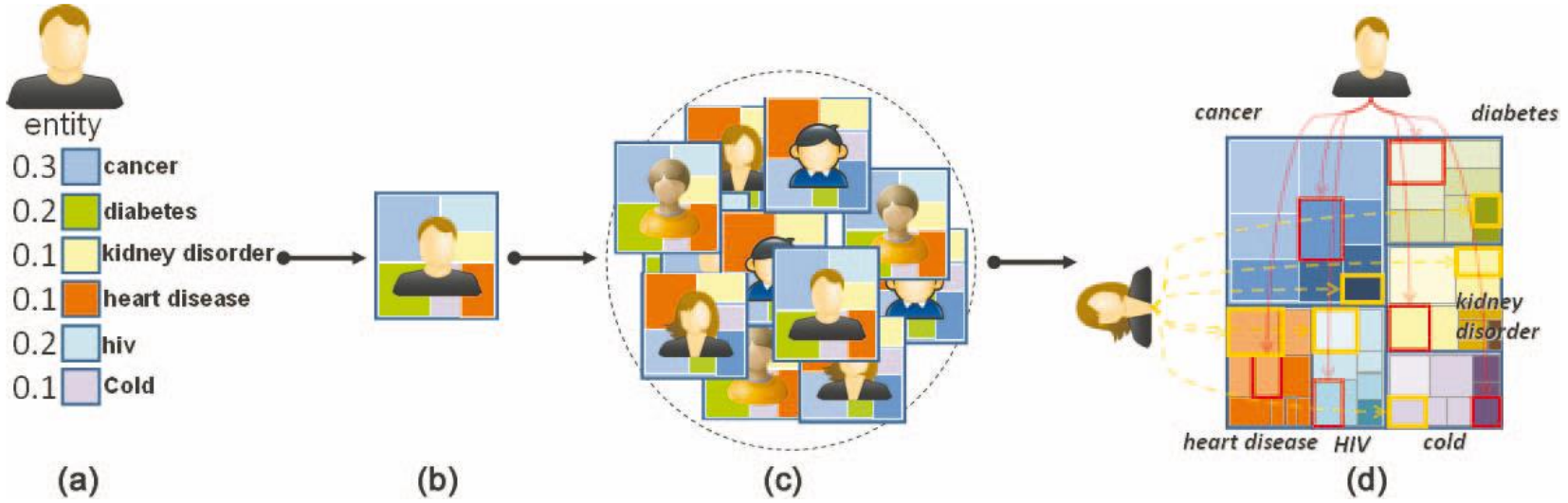
A few design guidelines for glyphs (from Cao,2011):

- Support multiple levels of visual exploration, e.g. features, points (feature vectors), subclusters, clusters
- Provide consistent encodings
  - Normalize data, e.g. to (0,1) ranges
- Design glyphs that are perceived as similar if the underlying data (size of the cluster, …) are similar
- Provide interaction to explore the visualization and refine the clusters

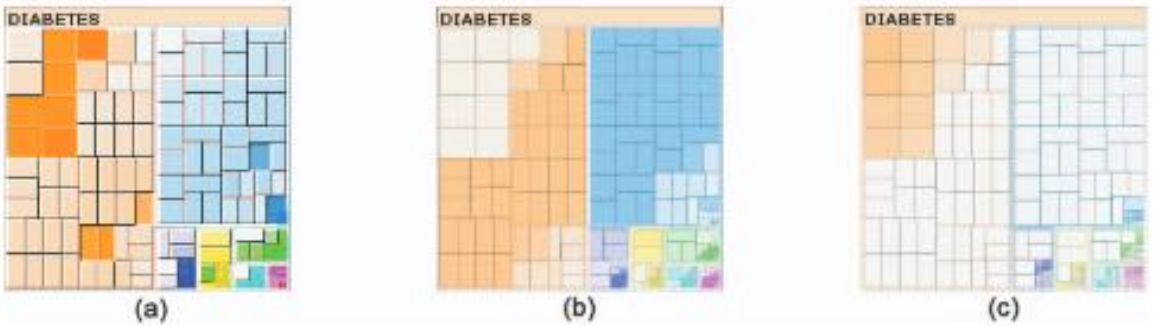
Visualization of Clustering Results: Glyphs

Dynamic icon-based visualization of Multidimensional Clusters (Cao, 2011)

- Glyphs can depict various attributes (size, shape, color) → useful for cluster visualization
- Treemaps depict hierarchies → useful for hierarchical clusters



Visualization of Clustering Results: Glyphs



(From: Cao, 2011)

Three color scales for cluster visualization.

Left: color indicates cluster quality. High saturation for good quality. (Cues?)

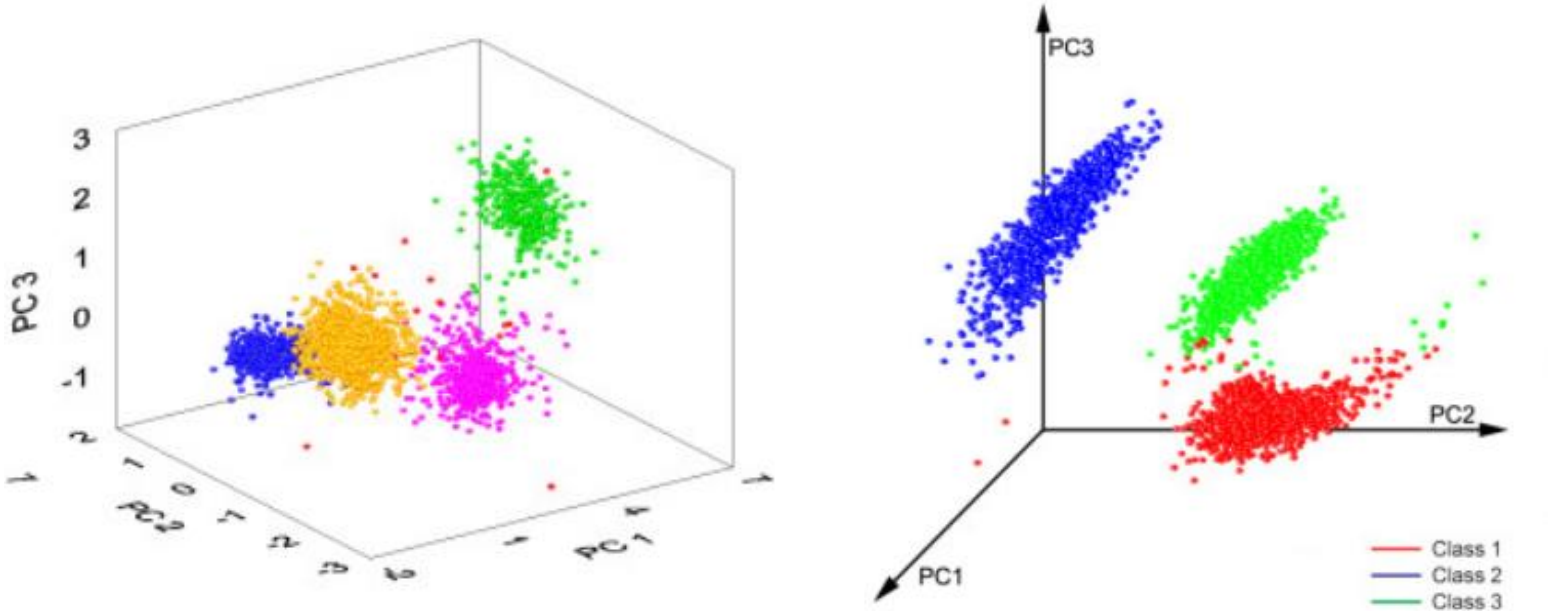
Middle: Color indicates co-occurrence of the current diseases with others (could be relevant for epidemiology data)

Right: Domination cue: Color indicates how strongly the current diseases dominates (multimorbide patients).

More would be possible, e.g. an outlier cue

Visualization: Scatterplots

- Unique colors are typically used to encode memberships to clusters
- Does not scale for many clusters (up to a few dozen)

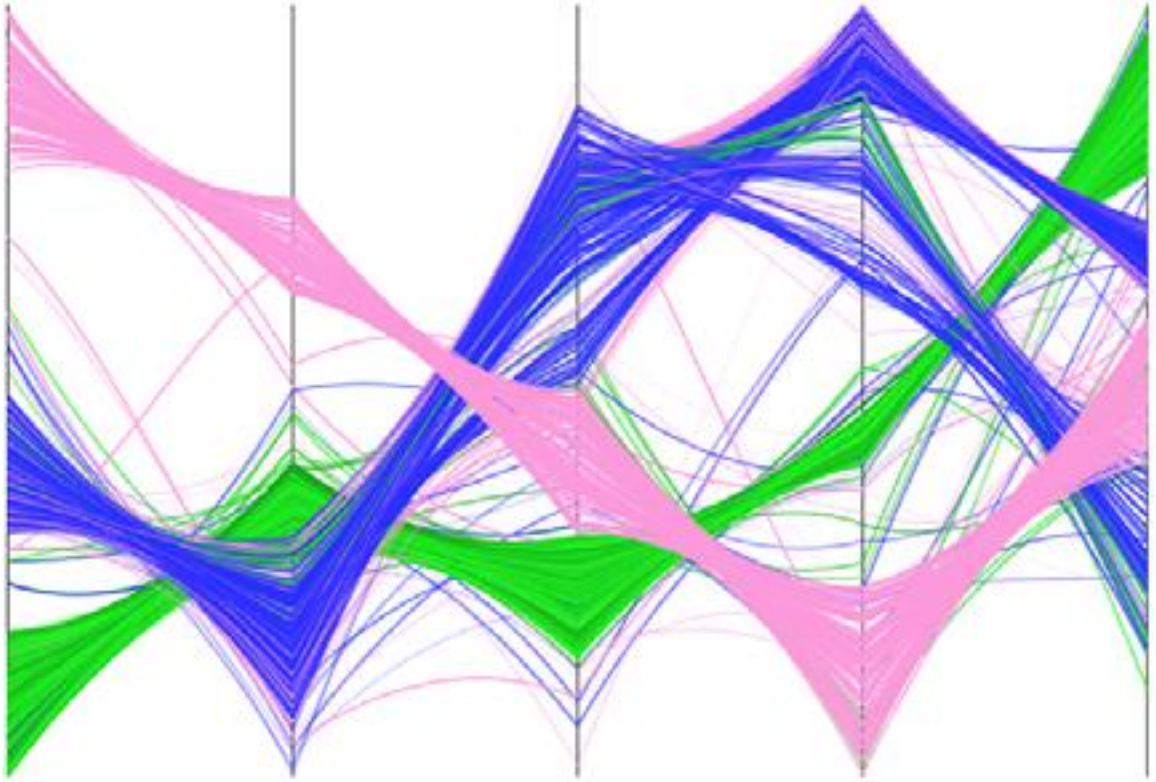


(From: Oliynuk,2012)

Visualization: Scatterplots

- Requires a projection to 2D or 3D space, e.g. with multidimensional scaling or PCA
- The resulting impression depends not only the parameters of the clustering but also on the chosen projection technique.
- For temporal clustering: animated scatterplots may reveal the dynamics of cluster formation, merging and splitting.
  - Careful temporal control is necessary

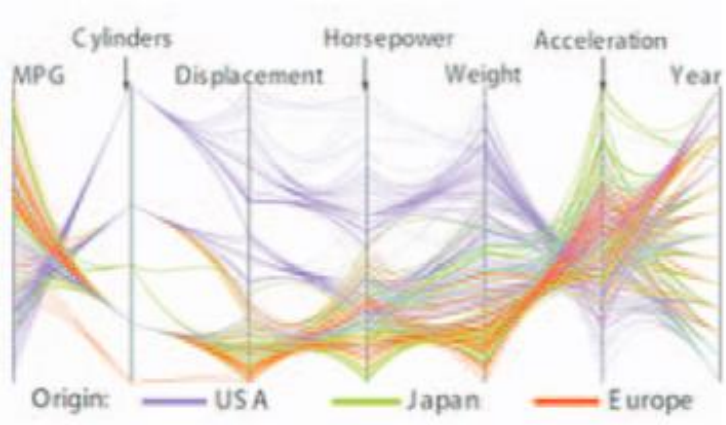
Visualization of Clustering Results: PC plots



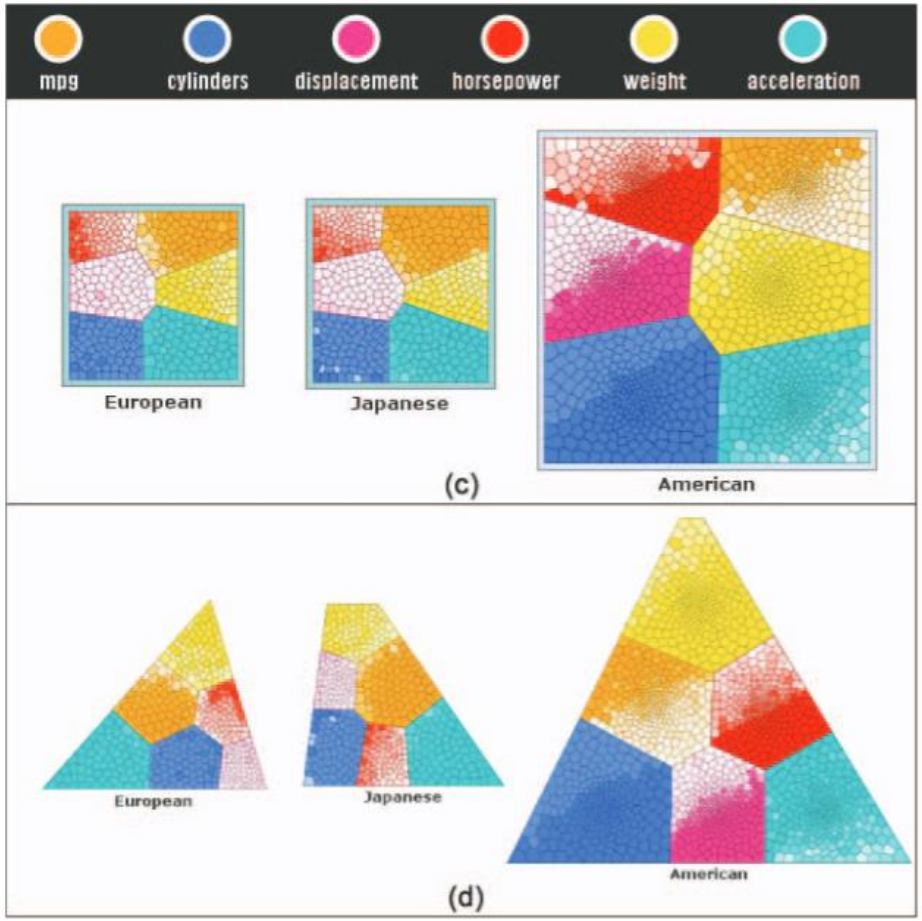
Visualization of clustering results in parallel coordinates with edge clustering (Zhou, 2008).

No projection is involved (full information available); color used again to indicate cluster membership.

Visualization of Clustering Results: PC plots



Visualization of cluster results related to the car dataset (~400 items). Parallel coordinates and icon-based. The bottom row indicates the distribution (skew, kortosis)



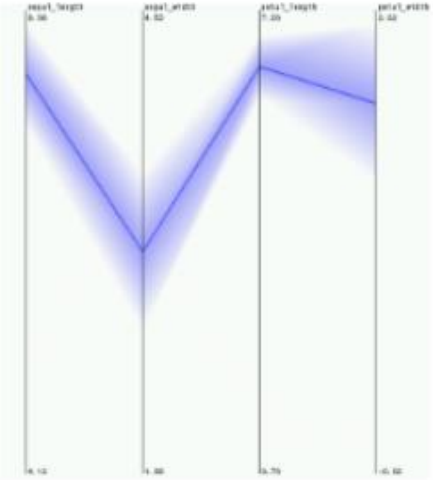
(From: Cao, 2011)

Visualization of Clustering Results: PC plots

Parallel coordinates may be enhanced with

- edge bundling,
- proximity colors,
- opacity adjustment and
- curved lines

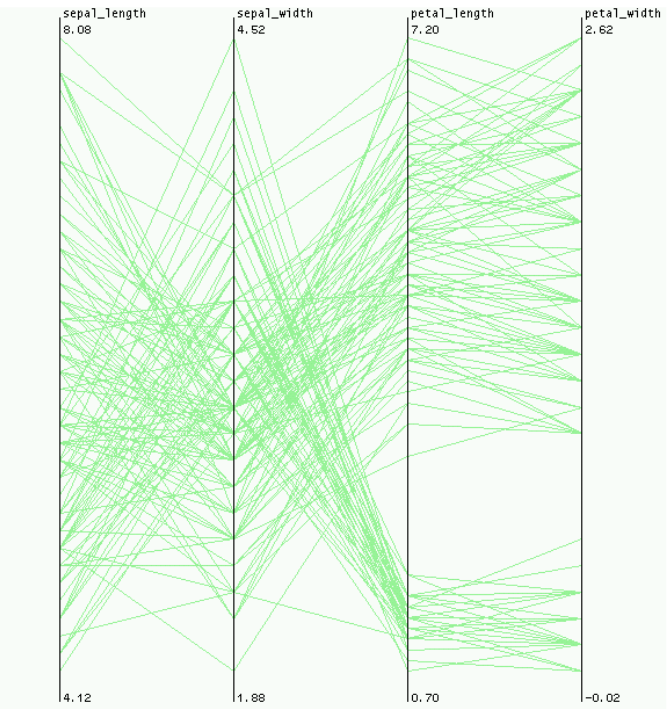
to avoid clutter and to reveal a cluster structure.

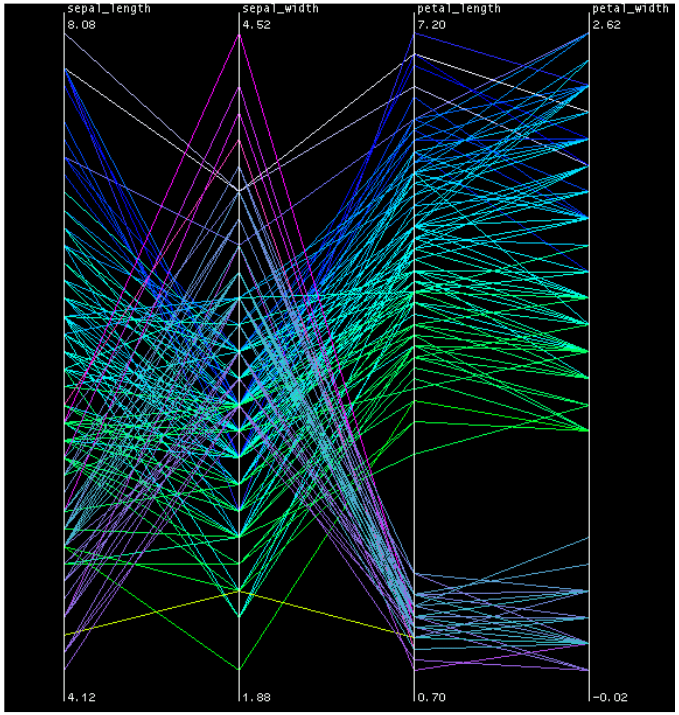


One cluster (4 dimensions) in parallel coordinates with opacity mapped to distance from centroid

(From: Fua, 1999)

Visualization of Clustering Results: PC plots



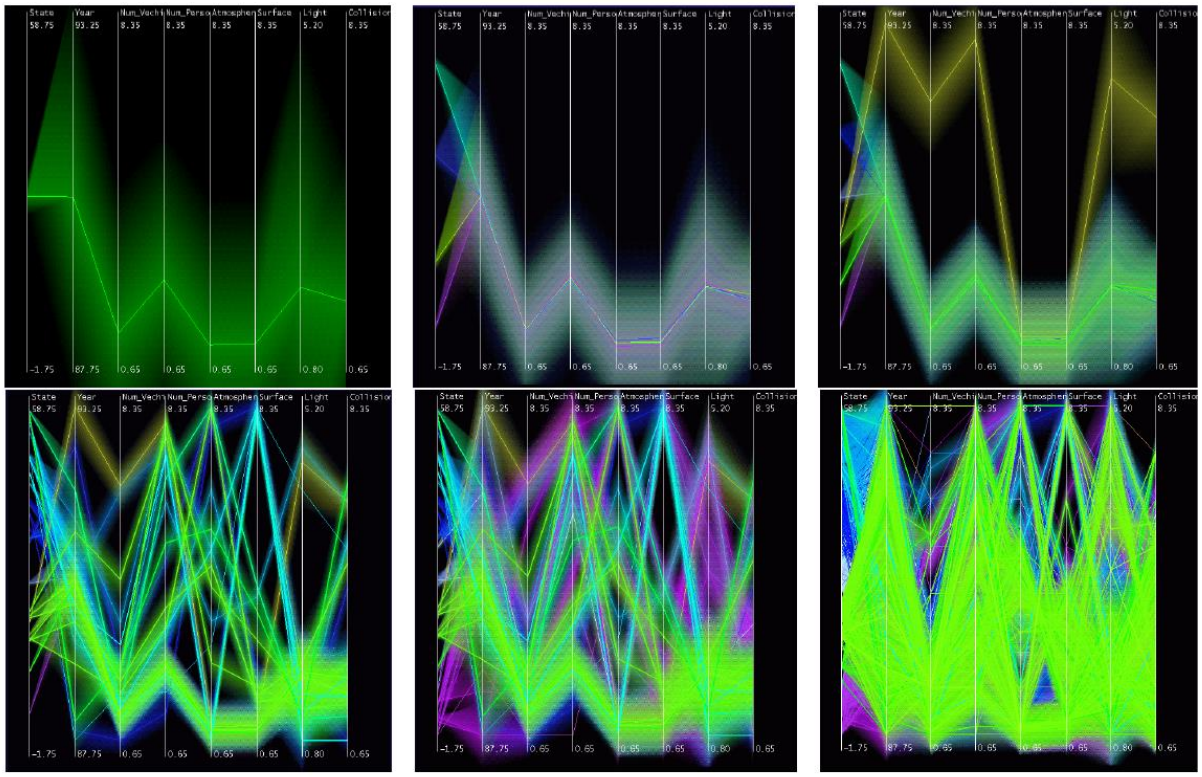


Proximity-based coloring: Hues are chosen to reveal cluster membership. Discrimination of clusters would be hampered in case of many intersecting lines from different clusters (From: Hua, 1999)

Visualization of Clustering Results: PC plots

Parallel coordinates for displaying hierarchical clustering results. Hierachy enables aggregation of items.

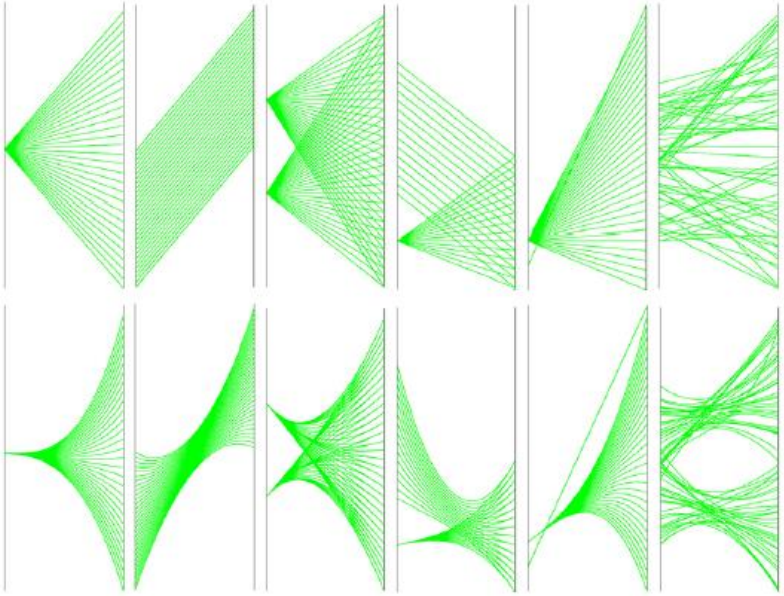
- Clusters at different levels enable multiscale exploration



From top left to bottom right: Root level, intermediate Clustering levels and leaf nodes. Fatal accident dataset with 230,000 items (From: Fua, 1999)

Visualization of Clustering Results: PC plots

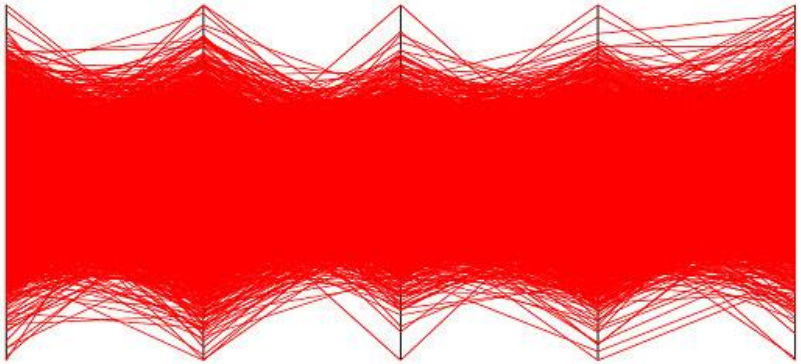
- Visual separation according to cluster results may be supported by bending lines.
- Bending is the result of an energy optimization with gravitation and curvature terms.
- The gravitation term attracts nearby lines, the curvature term is chosen to avoid excessive bending.
- The influence of the curvature is a user-adjustable parameter.

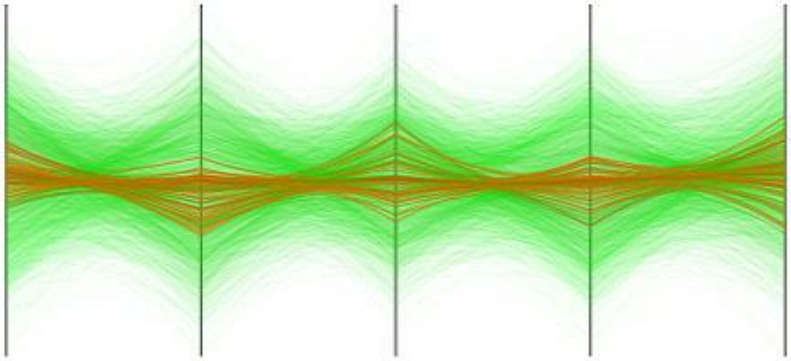


Top row: conventional linear combinations.

Bottom row: after energy optimization clutter is strongly reduced (From: Zhou, 2008)

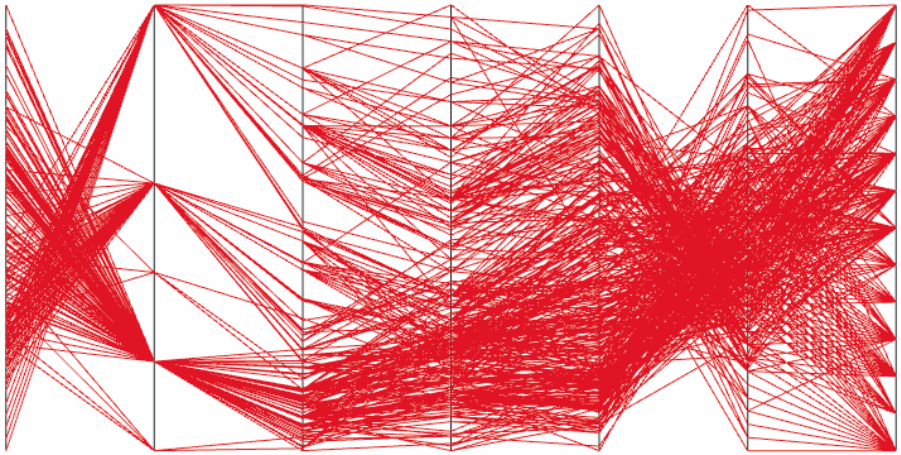
Visualization of Clustering Results: PC plots

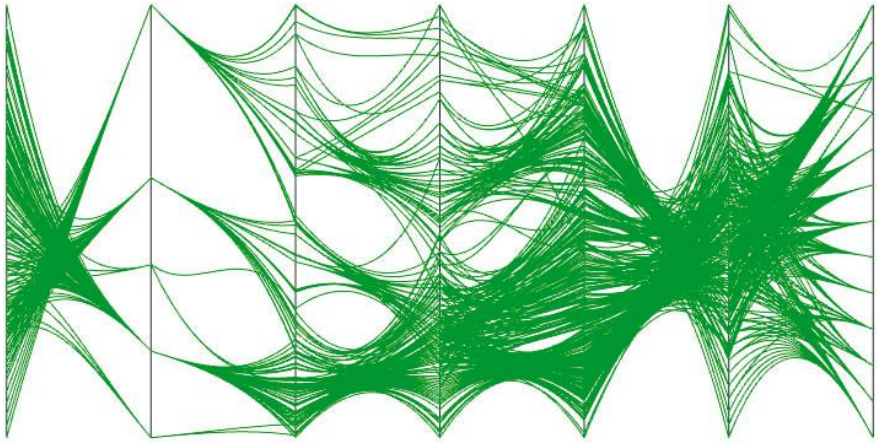


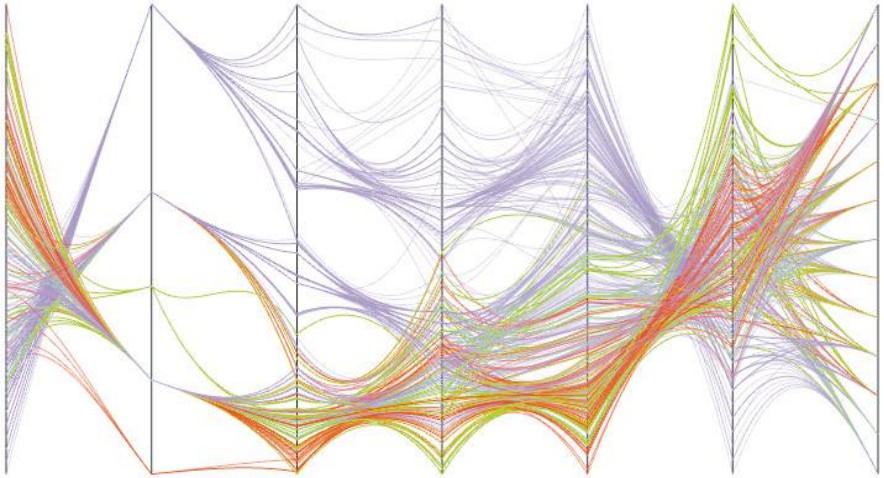


Conventional PC Plot (3848 items) and with color and opacity optimization to reveal clusters (From: Zhou, 2008)

Visualization of Clustering Results: PC plots



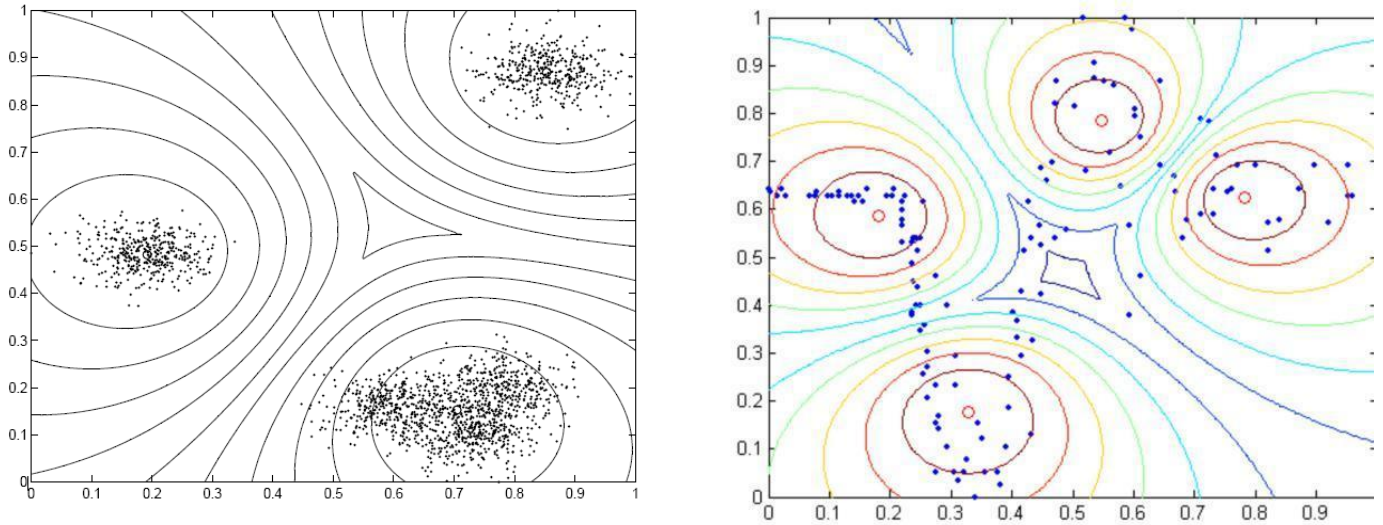




- 392 items with 7 variables (car dataset) Original PC Plot. Energy minimization to support visual clustering.
- Additional use of proximity colors. (From: Zhou, 2008)

Visualization: Isolines

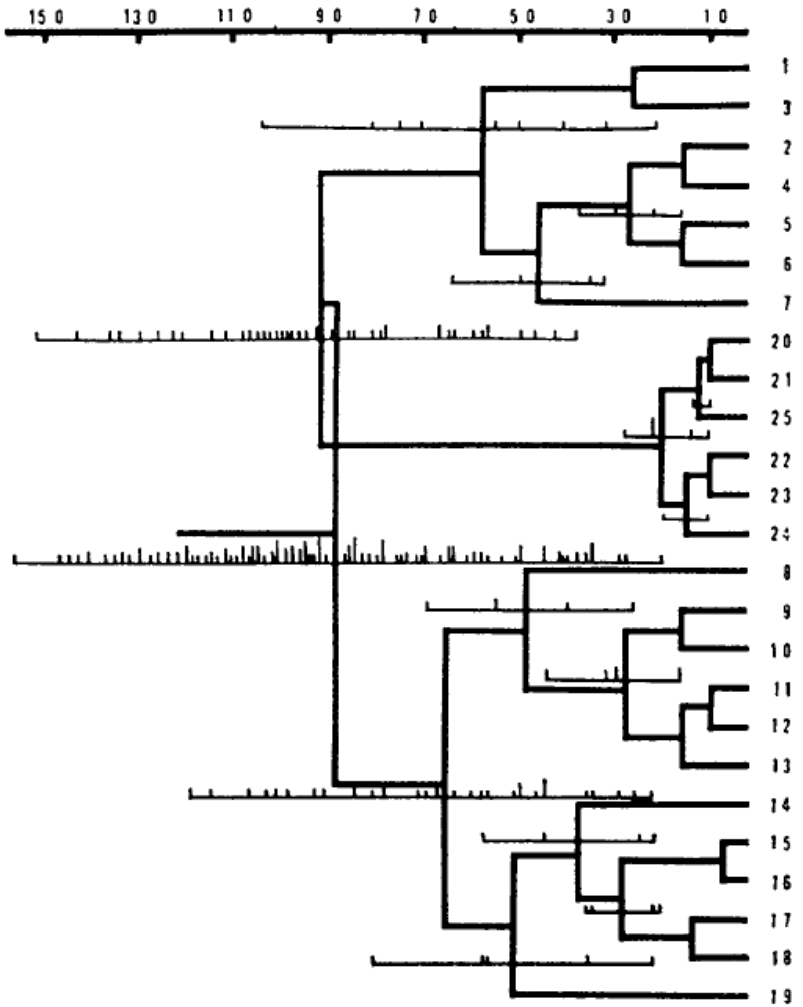
Isolines may represent degree of membership as result of fuzzy clustering.



(From: Fuzzy Clustering and Data Analysis Toolbox, Matlab, Link)

Isolines represent fuzzy cluster membership in a synthetic dataset. Left: after fuzzy c-means clustering, right: after Gustafson-Kessel clustering

Visualization: Enhanced Dendograms



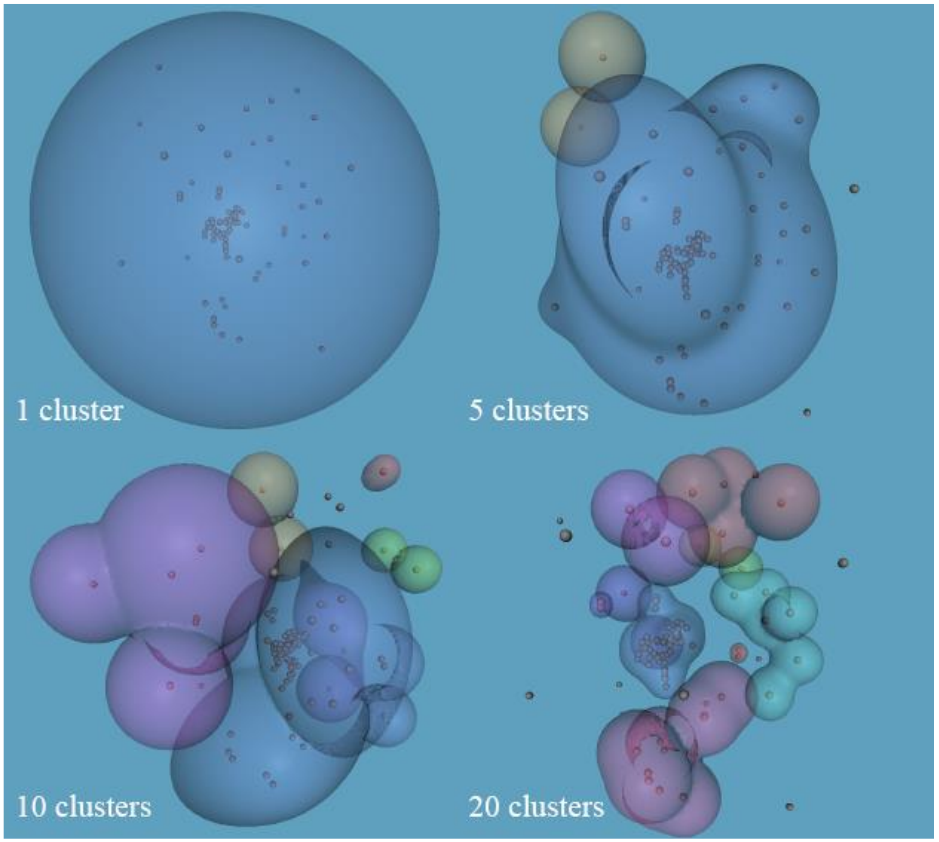
Dendograms

- nicely summarize how the data is grouped.
- Do NOT enable a verification whether this grouping makes sense.

Early concepts were developed to enhance dendograms, e.g. with frequency distributions per cluster (Rohlf, 1970)

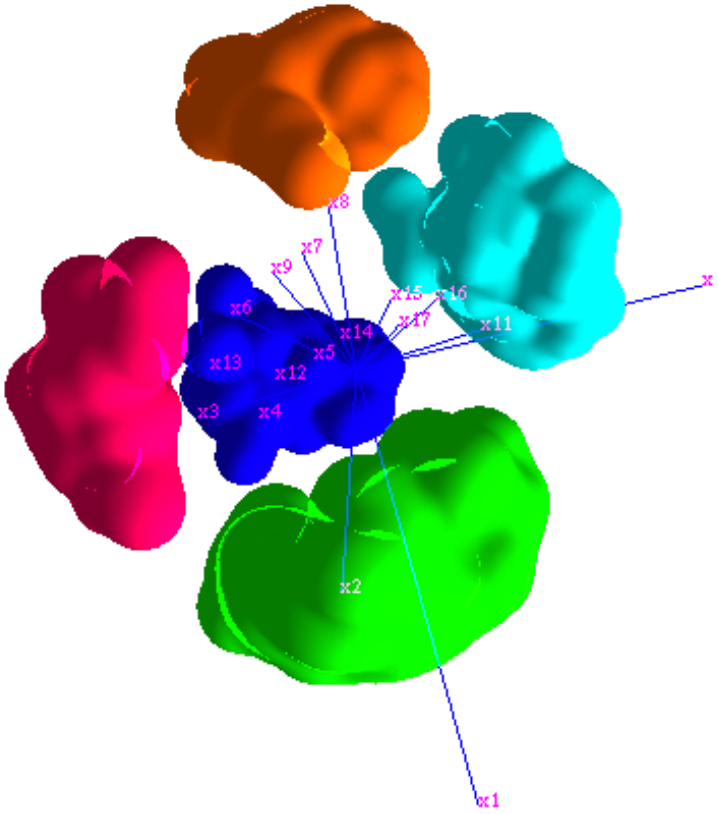
Visualization: 3D Hierarchy Visualizations

Hierarchy may be displayed with semi-transparent objects (nested surfaces). HD data projected to 3D with MDS or a related technique



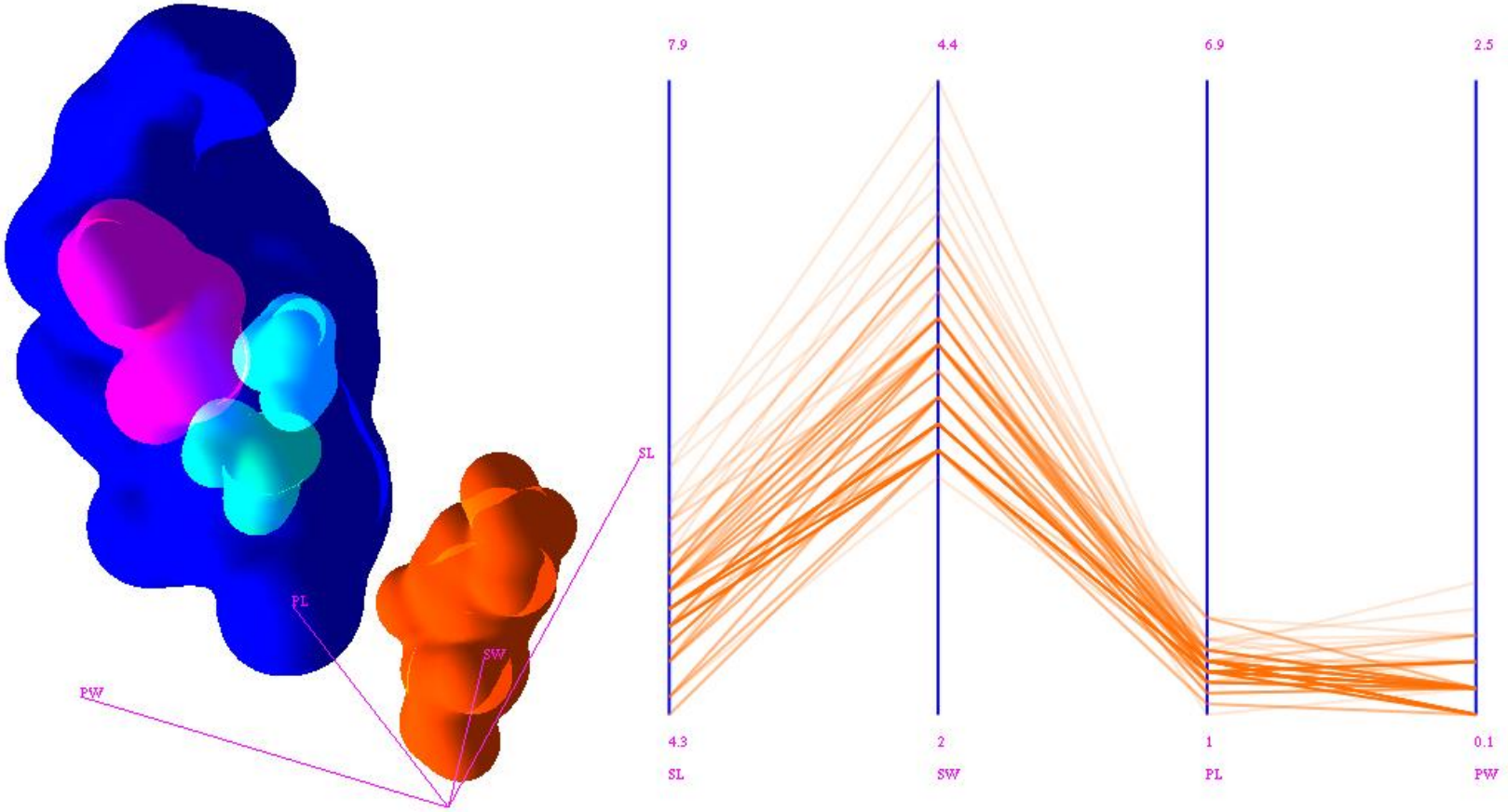
(From: Sprenger, 2000)

Visualization: 3D Hierarchy Visualizations



3D Hierarchy Visualization based on 3D Star coordinates (From: van Long, 2009)

Visualization: 3D Hierarchy Visualizations



3D Hierarchy Visualization based on 3D Star coordinates combined with parallel coordinates. 4D Iris dataset (From: van Long, 2009)

Visualization of Clustering Results

Integrating Cluster Analysis with Feature-Based Analysis:

- 1D Features: skew, curtosity, degree of fit to uniform/Normal/Poisson … distribution, largest gap size, …
- 2D Features: linear, quadratic, logistic … regression, statistical power (F value), number of outliers, …
- Rank by Feature Framework: rank dimensions or pairs of dimensions according to selected 1D or 2D features (Seo, 2004; Seo, 2006)

Visualization of Clustering Results

(From: Seo, 2004)

Outlier Detection

- Subfield of data mining that detects anomalies

  Applications:

  – detection of misuse of credit cards (abnormal transactions), other security measures (network traffic is an abnormality),

  – abnormal values in medicine may indicate pathologies,

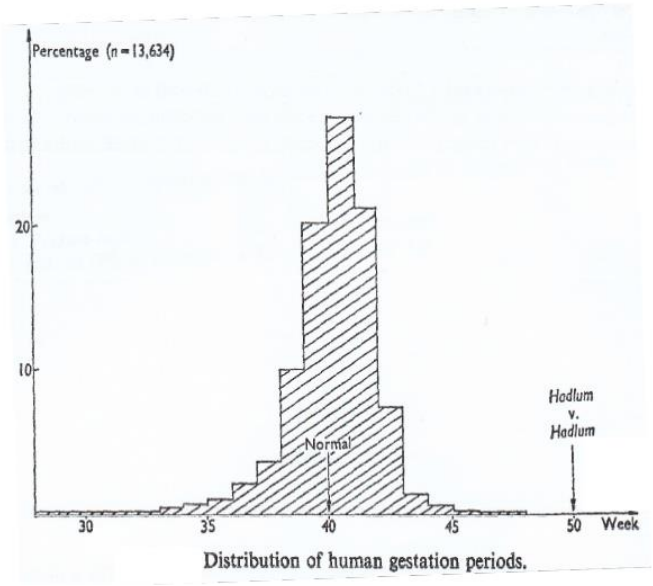
  – Abnormal values of medical parameters in sportsman may identify outstanding sportsman

- **Challenge**: The definition of „abnormality“ is highly context-dependent, e.g. expected values for blood composition depend on age and body mass index.
- Outliers may be particularly interesting or unwanted, e.g. since they influence regression analysis and should be removed.

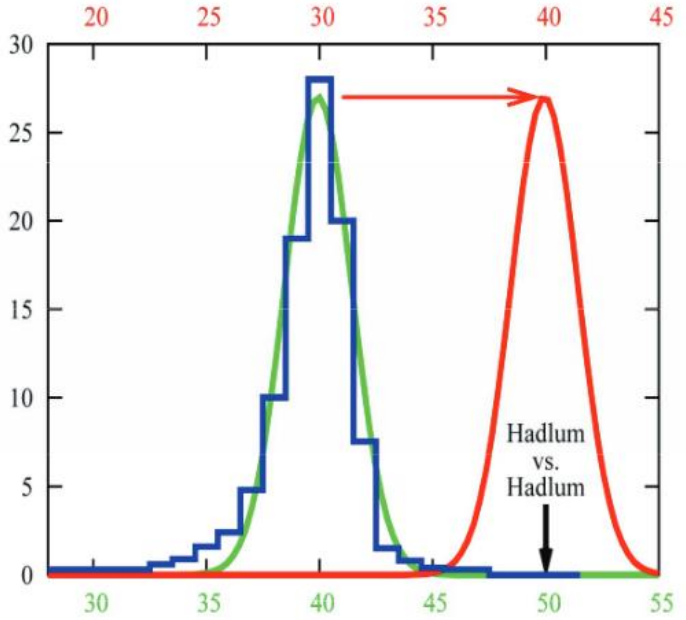
Outlier Detection

- Outliers are often a result of sampling errors. Thus, there would be many datasets with similar values like the outlier but these are not part of the database, e.g. a car database with many heavy cars (a lot of PS, cylinders, high oil consumption) and just one small car.
- The detection of outliers may trigger the collection of more data.

Outlier Detection



Distribution of human gestation periods.



A histogram of gestation periods (Schwangerschaftsdauer) indicates that 349 days (more than 11 months) is an outlier. The birth of a child to Mrs. Hadlum occured 349 days after Mr. Hadlum went to military service.

To assess outliers, a normal distribution is fitted to the data and also a normal distribution centered on the 349 days.

(Source: Kriegel, 2010)

Outlier Detection

- An outlier does not fit to the expectations; it deviates locally strongly from all neighbors
  - One definition is using the interquartile range:
  - *Mild outliers* are outside an interval that is 1.5 x IQ range.
  - *Extreme outliers* are outside an interval that is 3 x IQ range.
- Outlier detection is a typical visual analytics task – it benefits from a combination of automatic analysis and visual exploration/verification
- Box plots explicitly label outliers.
- However, these outliers are just extrema (rare observations) – every distribution of more than several hundred points has them

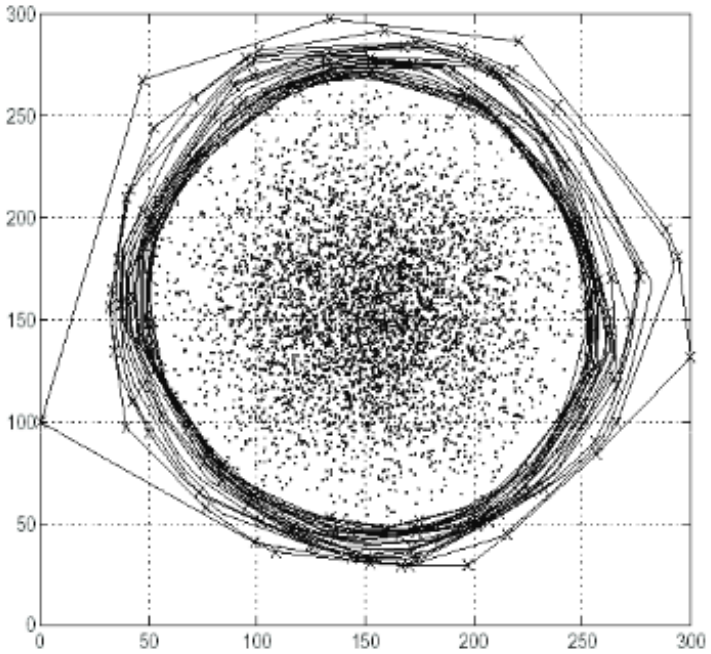
Outlier Detection

- Outliers is multidimensional or high-dimensional space are hard to identify
- In statistics, often certain distributions are assumed (and checked), e.g. normal distribution
  - If the attribute value of an element is extremely unlikely for the given distribution, it is considered an outlier since it is not based on the assumed *generative process*
- Algorithms in data mining differ from implementations of statistical outlier measures for performance reasons
- A variety of clustering algorithms identifies outliers.
- However, „clustering algorithms are optimized to find clusters rather than outliers” (Kriegel, 2009)
- Thus, a set of similar outliers is considered a cluster.

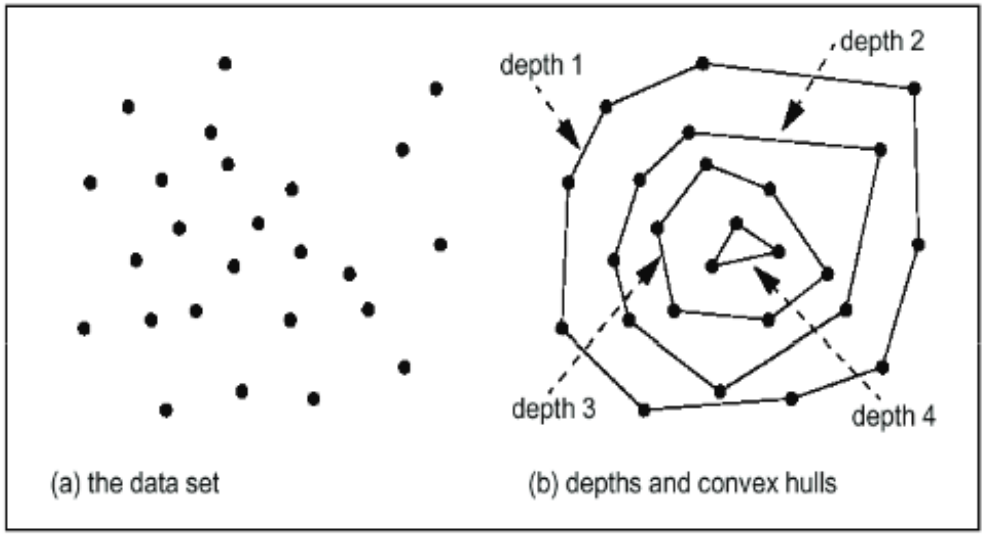
Outlier Detection: Methods

- Depth-based
  - Assumption: Outliers are at the boundary of a distribution and normal elements in the center
  - Depth 1 means that elements are part of the convex hull of the data → likely outliers
  - Depth 2 means that elements are part of the convex hull after Depth 1 points are removed → less likely outliers
- Distance-based
  - An element is considered as an outlier if it has an abnormal distance to its nearest neighbor
  - Not appropriate if the distribution of data varies in local density
- Density-based
  - An element is considered as an outlier if it has an abnormally low density in the surrounding

Outlier Detection: Methods



Picture taken from [Johnson et al. 1998]



Picture taken from [Preparata and Shamos 1988]

Depth-based outlier detection relies on convex hulls. Efficient for 2D and 3D data. Delivers binary output or scoring (depth values as scores)

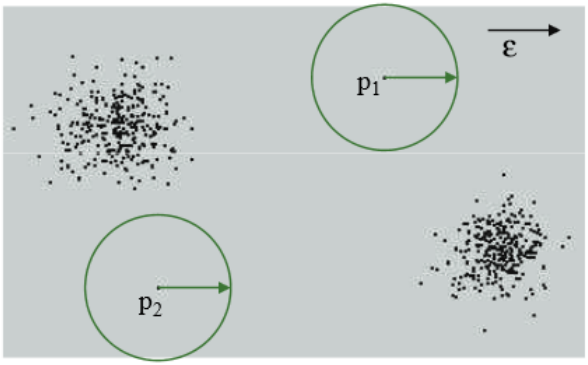
Outlier Detection: Methods

Distance-based outlier

- Given a radius ε and a percentage π
- A point p is considered an outlier if at most π percent of all other points have a distance to p less than ε

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

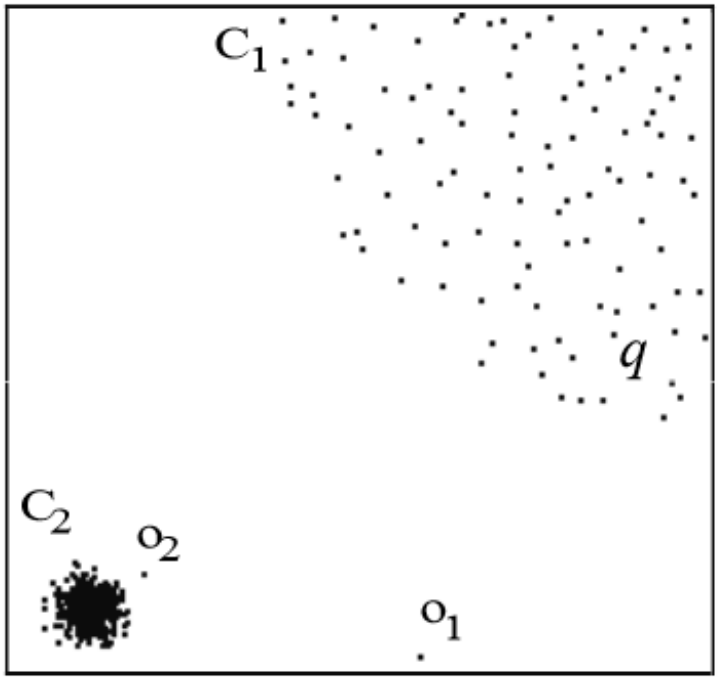
range-query with radius ε



Please note, that p_1 and p_2 may not be considered as outliers with a depth-based method (From: Kriegel, 2009)

Outlier Detection: Methods

Density-based methods detect outliers if they have a locally abnormal distance to the nearest neighbor. In contrast to depth- and distance based methods, o_1 and o_2 are considered outliers according to the Local Outlier Factor (LOF)

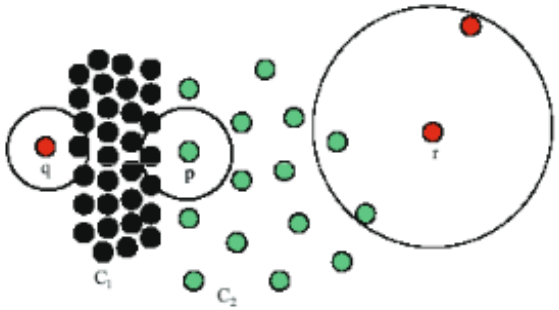


The data has clusters with different density. O_2 has abnormal distances to the C_2 members. (From: Kriegel, 2009)

Outlier Detection: Methods

Discussion: Density-based methods

- Simple techniques do not detect outliers reliable if clusters with different density are not clearly separated.



With simple techniques, p is considered an outlier and q and r are not. (From: Kriegel, 2009)

- With an improved measure, Influenced Outlierness, this can be compensated.
- Advanced density-based methods are computationally very demanding (exponential w.r.t. number of dimensions)

Outlier Detection: Results

Algorithms generate

- **continuous output**, a measure for **outlierness**, or
- **binary output**, some elements are labeled as outliers

If an outlier measure is computed, users are often interested in the top n outliers.

An outlier measure may be used to define a cutoff value where the distance to the next element is large. (The Top 5 outliers are not very reliable if the 4th, 5th, 6th and 7th elements in that sorting have very similar values).

Outlier Detection: Discussion

- Similar to clustering, we have a number of methods available.
- The selection of methods should consider the context, problem and derived model assumptions.
- To provide a small set of methods is better than a single method.
- All methods have parameters, for which good initial suggestions and flexibility to adjust should be provided.

Outlier Detection in HD Data

- In HD data, some dimensions may contain noise, non-discriminative and irrelevant data.
- Outlier detection, in particular with distance-based methods suffers from irrelevant distances
- Often, in a first step (feature selection, dimension reduction) the data is preprocessed to restrict the number of dimensions
- Outlier detection is applied to the corresponding subspaces.

Summary

- Clustering is evaluated by means of quality measures, visualization and expert feedback.
- Cluster visualization relies on projection techniques.
- Overview and detail visualizations provide increasingly more information on cluster members, distributions, …
- Outliers may be deliberately searched for or are a side-effect of some clustering techniques
- Outlier detection is based on an outlier model.

Outlook

Current research focuses on

- temporal clustering (how to show the emergence, development of clusters over time)
- Online clustering (without storing all incoming data, cluster current data)
- Incremental clustering (update clustering results when new data arrive)
- Clustering of complex data, e.g. blood flow, other flow data, fiber tracts

References

Bae, E., & Bailey, J. 2006. COALA: A Novel Approach for the Extraction of an Alternate Clustering of High Quality and High Dissimilarity.In: IEEE International Conference on Data Mining **(**ICDM), pp. 53-62.

James C. Bezdek (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms*.

Nan Cao, David Gotz, Jimeng Sun, Huamin Qu, "DICON: Interactive Visual Analysis of Multidimensional Clusters," *IEEE Trans. Vis. Comput. Graph.* Vol. 17(12):2581–2590, 2011

Matthias Dehmer (2007). *Strukturelle Analyse Web-basierter Dokumente*, DUV Verlag

M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in Proc. of Knowledge Discovery and Data Mining, KDD, pp. 226–231, 1996.

Sylvia Glaßer, Kai Lawonn, Bernhard Preim. Visualization of 3D Cluster Results for Medical Tomographic Image Data. In *Proc. of Conference on Computer Graphics Theory and Applications (VISIGRAPP/GRAPP)*, pp. 169-176, 2014

S. Glaßer, S. Roscher, B. Preim. Adapted Spectral Clustering for Evaluation and Classification of DCE-MRI Breast Tumors, *Bildverarbeitung für die Medizin (BVM),* pp. 198-203, 2014

RJ Hathaway, JC Bezdek. "Visual cluster validity for prototype generator clustering models", Pattern Recognition Letters 24 (9), 1563-1569

Huang Z. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Mining and Knowledge Discovery*. 1998;2(3):283-304

References (II)

Johnson, T., Kwok, I., and Ng, R.T. 1998. Fast computation of 2-dimensional depth contours. In Proc. Int. Conf. on Knowledge Discovery and Data Mining (KDD)

Paul Klemm, Lisa Frauenstein, David Perlich, Katrin Hegenscheid, Henry Völzke, Bernhard Preim. Clustering Socio-demographic and Medical Attribute Data in Cohort Studies, *Bildverarbeitung für die Medizin (BVM),* pp. 180-185, 2014

Kriegel H.-P., Kröger P., Zimek A.: Outlier Detection Techniques. *Tutorial. In: 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining* (PAKDD 2009).

Robert L. Ling. 1973. A computer generated aid for cluster analysis. *Commun. ACM* 16, 6 (1973), 355-361.

S. Salvador and P. Chan, “Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms,” *IEEE Tools with Artificial Intelligence*, pp. 576–584, 2004.

T. C. Sprenger, R. Brunella, and Markus H. Gross. H-blob: a hierarchical visual clustering method using implicit surfaces. *Proc. of IEEE Visualization*, pp. 61–68, 2000

Bart Moberts, Anna Vilanova, Jarke J. van Wijk: Evaluation of Fiber Clustering Methods for Diffusion Tensor Imaging. *IEEE Visualization* 2005: 9-16

J. B. MacQueen: Some Methods for classification and Analysis of Multivariate Observations. In: Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. 1, University of California Press, 1967, S. 281–297

References (III)

E. Müller, S. Günnemann, I. Färber, and T. Seidl. "Discovering Multiple Clustering Solutions: Grouping Objects in Different Views of the Data", Tutorial at the International Conference on Machine Learning, Atlanta, USA (2013)

S. Oeltze, D. J. Lehmann, A. Kuhn, G. Janiga, H. Theisel, B. Preim: Blood Flow Clustering and Applications in Virtual Stenting of Intracranial Aneurysms. *IEEE Trans. Vis. Comput. Graph.* 20(5): 686-701 (2014)

Oliynyk, Andriy; Bonifazzi, Claudio; Montani, Fernando; Fadiga, Luciano, "Automatic online spike sorting with singular value decomposition and fuzzy C-mean clustering", BMC Neuroscience", 2012", Vol. 13(1): 1--19

Preparata, F. and Shamos, M. 1988. Computational Geometry: an Introduction. Springer

F. J. Rohlf. “Adaptive hierarchical clustering schemes”, Systematic Zoology, 19:58-82, 1970

E. M. Reingold and J. S. Tilford. Tidier drawings of trees. *IEEE Transactions on Software Engineering*, 7(2):223–238, 1981

Michael Sedlmair, A. Tatu, Tamara Munzner, Melanie Tory: A Taxonomy of Visual Cluster Separation Factors. Comput. Graph. Forum 31(3): 1335-1344 (2012)

J. Seo, B. Shneiderman: Interactively Exploring Hierarchical Clustering Results. IEEE Computer 35(7): 80-86 (2002)

T Van Long (2009). Visualizing high-density clusters in multidimensional data, PhD thesis, School of Engineering and Science, Jacobs University, Bremen

H. Zhou, X. Yuan, H. Qu, W. Cui, and B. Chen. Visual clustering in parallel coordinates. Computer Graphics Forum, Vol. 27(3):1047–1054, 2008