Ludwig Maximilians Universität München
Institut für Informatik
Lehr- und Forschungseinheit für Datenbanksysteme

# Clustering in Subspaces of High-Dimensional Data

*Talk at RWTH Aachen, 2.3.2010*

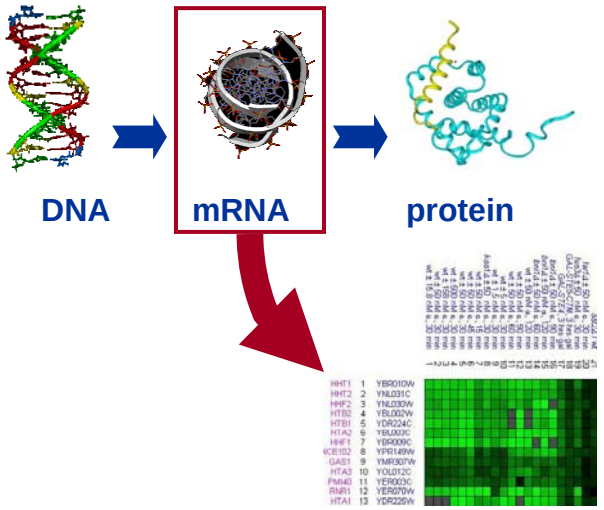
Arthur Zimek
Ludwig-Maximilians-Universität München
Munich, Germany
http://www.dbs.ifi.lmu.de/~zimek
zimek@dbs.ifi.lmu.de

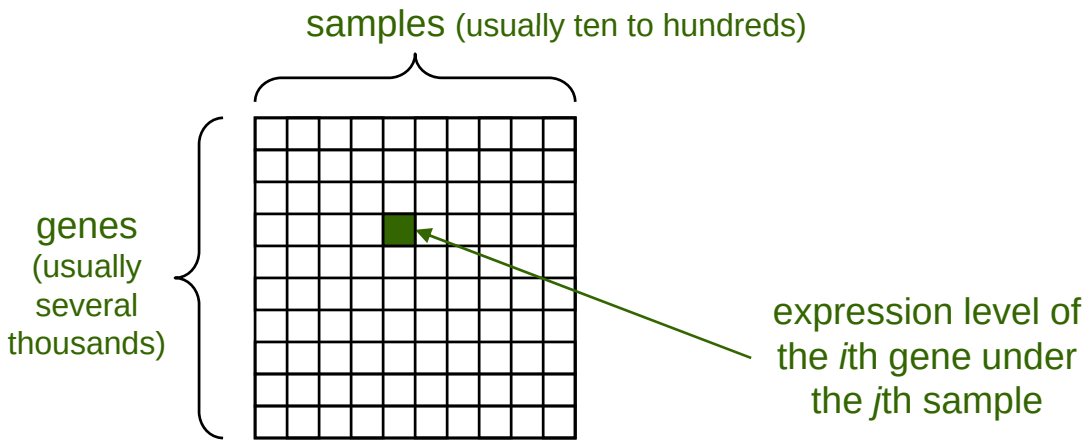
# Outline

1. Sample Applications
2. General Problems and Challenges: the Curse of Dimensionality
3. A First Taxonomy of Approaches
4. Arbitrarily-oriented Subspace Clustering
    1. PCA-Based Approaches
    2. Correlation Clustering Based on the Hough-Transform

# Sample Applications

- Gene Expression Analysis
  - Data:
    - Expression level of genes under different samples such as
      - different individuals (patients)
      - different time slots after treatment
      - different tissues
      - different experimental environments
    - Data matrix:



# Sample Applications

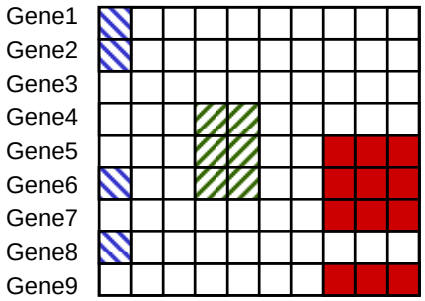
- Task 1: Cluster the rows (i.e. genes) to find groups of genes with similar expression profiles indicating homogeneous functions
  - *Challenge*:

    genes usually have

    different functions

    under varying

    (combinations of) conditions




Cluster 1: {G1, G2, G6, G8}

Cluster 2: {G4, G5, G6}

Cluster 3: {G5, G6, G7, G9}

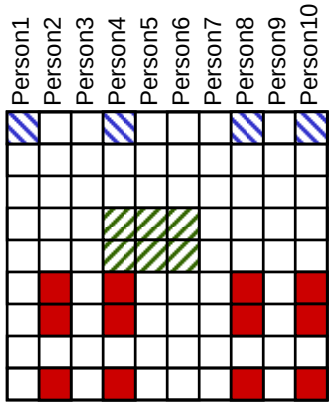
- Task 2: Cluster the columns (e.g. patients) to find groups with similar expression profiles indicating homogeneous phenotypes
  - *Challenge*:

    different phenotypes

    depend on different

    (combinations of)

    subsets of genes



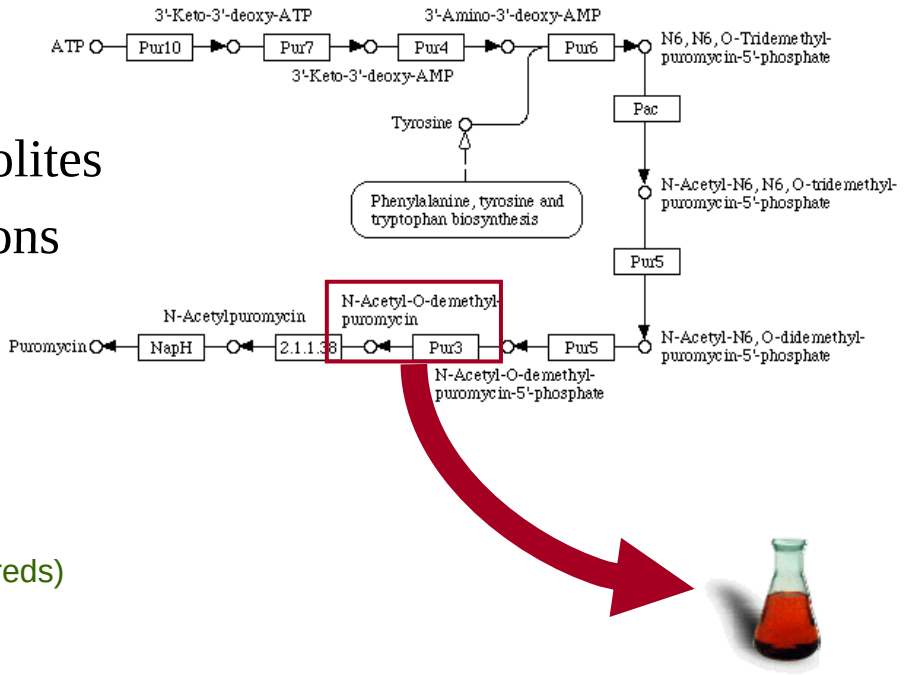

Cluster 1: {P1, P4, P8, P10}

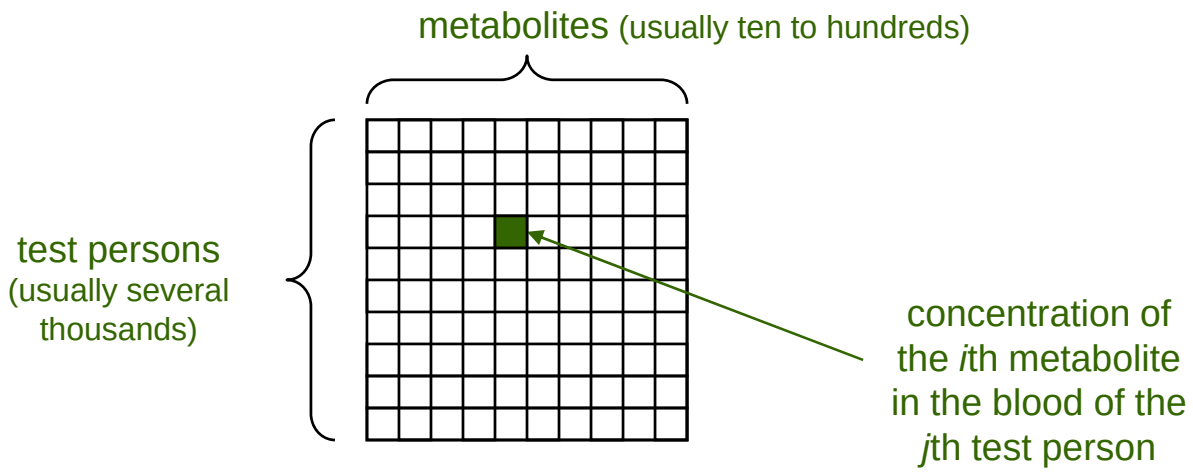
Cluster 2: {P4, P5, P6}

Cluster 3: {P2, P4, P8, P10}

# Sample Applications

- Metabolic Screening
  - Data
    - Concentration of different metabolites in the blood of different test persons
    - Example: *Bavarian Newborn Screening*
    - Data matrix:

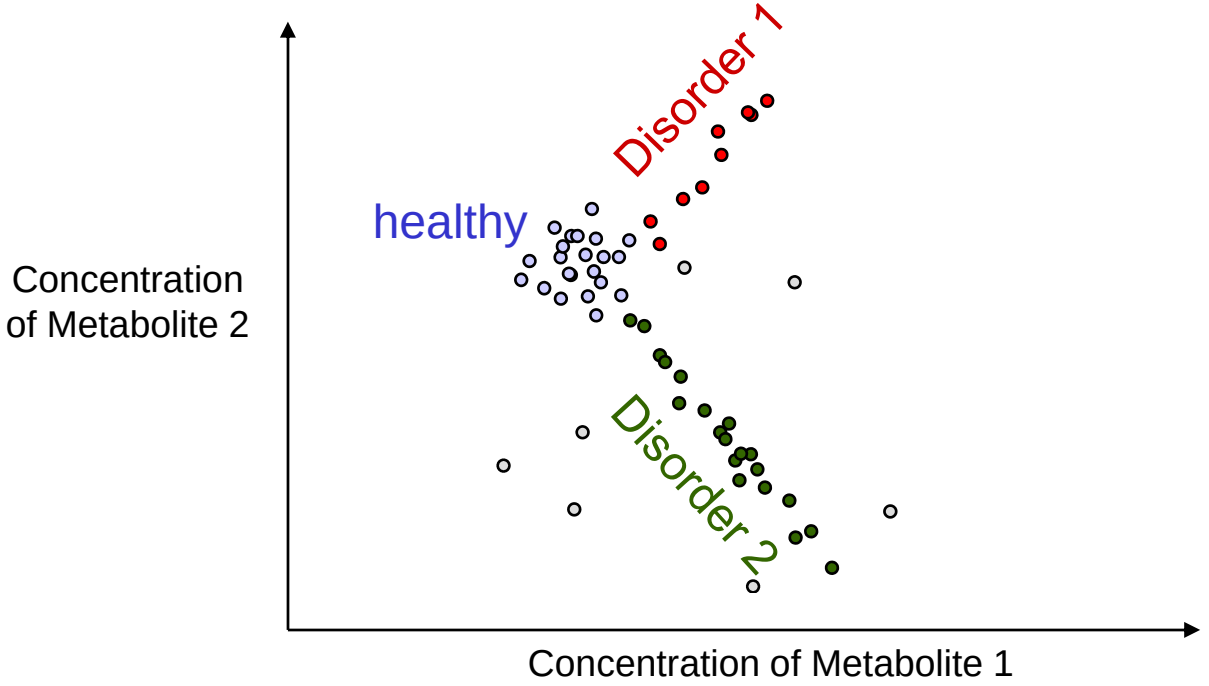


# Sample Applications

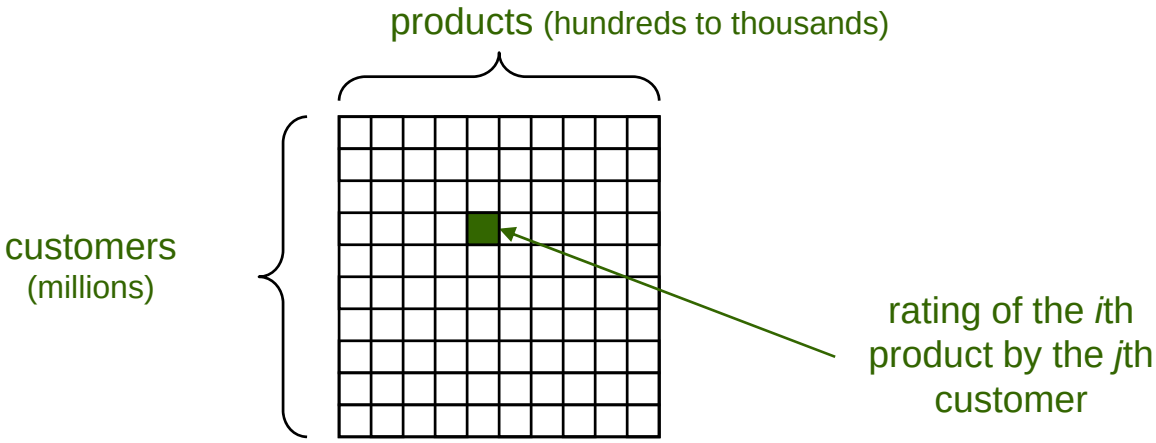
- Task: Cluster test persons to find groups of individuals with similar correlation among the concentrations of metabolites indicating homogeneous metabolic behavior (e.g. disorder)
  - *Challenge*:

    different metabolic disorders appear through different correlations of (subsets of) metabolites



# Sample Applications

- Customer Recommendation / Target Marketing
  - Data
    - Customer ratings for given products
    - Data matrix:




  - Task: Cluster customers to find groups of persons that share similar preferences or disfavor (e.g. to do personalized target marketing)
    - *Challenge*:

      customers may be grouped differently according to different preferences/disfavors, i.e. different subsets of products

# Outline

1. Sample Applications

2. General Problems and Challenges: the Curse of Dimensionality

3. A First Taxonomy of Approaches

4. Arbitrarily-oriented Subspace Clustering
    1. PCA-Based Approaches
    2. Correlation Clustering Based on the Hough-Transform

# General Problems & Challenges

The "*curse of dimensionality*": one buzzword for many problems

- First aspect: *Optimization Problem* (Bellman).

  *"[The] curse of dimensionality [… is] a malediction that has plagued the scientists from earliest days."* [Bel61]

  - The difficulty of any global optimization approach increases exponentially with an increasing number of variables (dimensions).
  - General relation to clustering: fitting of functions (each function explaining one cluster) becomes more difficult with more degrees of freedom.
  - Direct relation to subspace clustering: number of possible subspaces increases dramatically with increasing number of dimensions.

# General Problems & Challenges

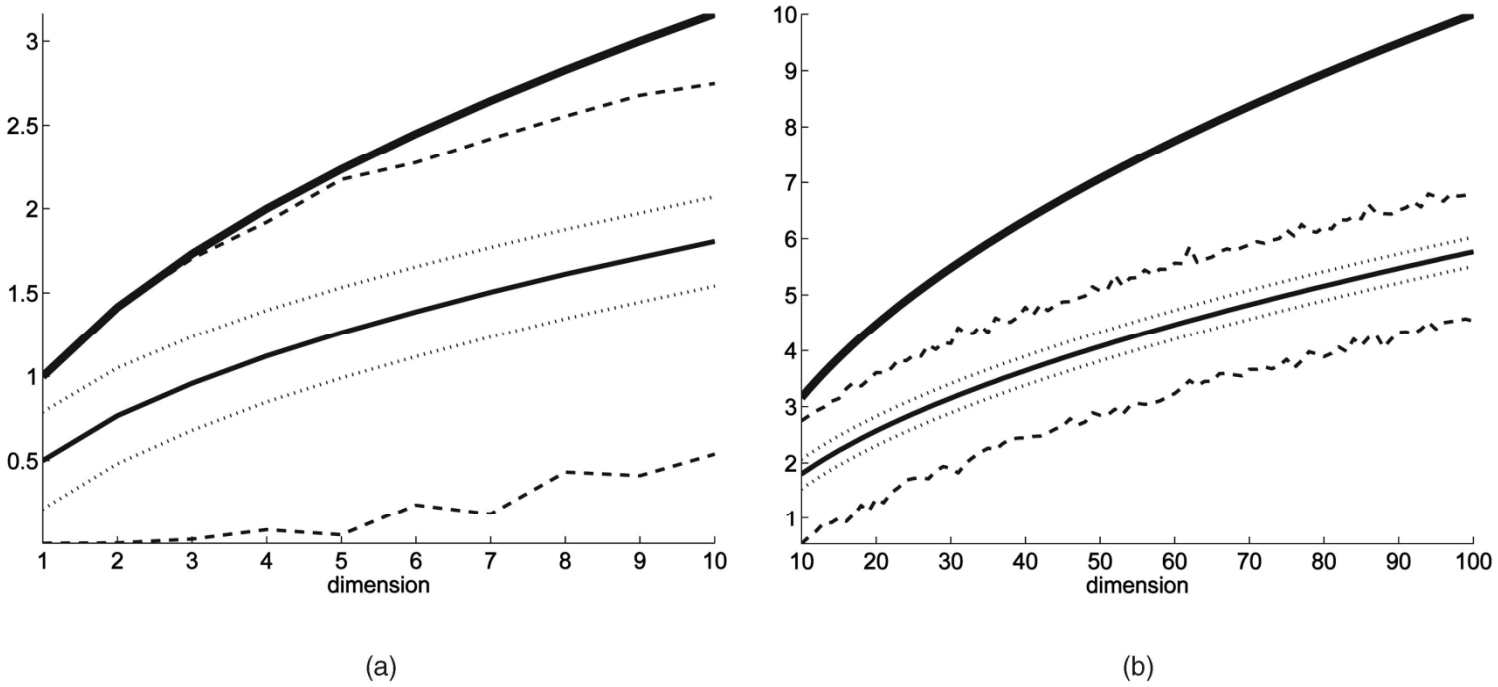
- Second aspect: *Concentration effect of $L_p$-norms*
  - In [BGRS99,HAK00] it is reported that the ratio of ($\text{Dmax}_d$ – $\text{Dmin}_d$) to $\text{Dmin}_d$ converges to zero with increasing dimensionality *d*
    - $\text{Dmin}_d$ = distance to the nearest neighbor in *d* dimensions
    - $\text{Dmax}_d$ = distance to the farthest neighbor in *d* dimensions

    Formally:

$$\forall \varepsilon > 0 : \lim_{d \to \infty} \mathrm{P}\left[ dist_d\left( \frac{\text{Dmax}_d - \text{Dmin}_d}{\text{Dmin}_d}, 0 \right) \leq \varepsilon \right] = 1$$

    - This holds true for a wide range of data distributions and distance functions

# General Problems & Challenges

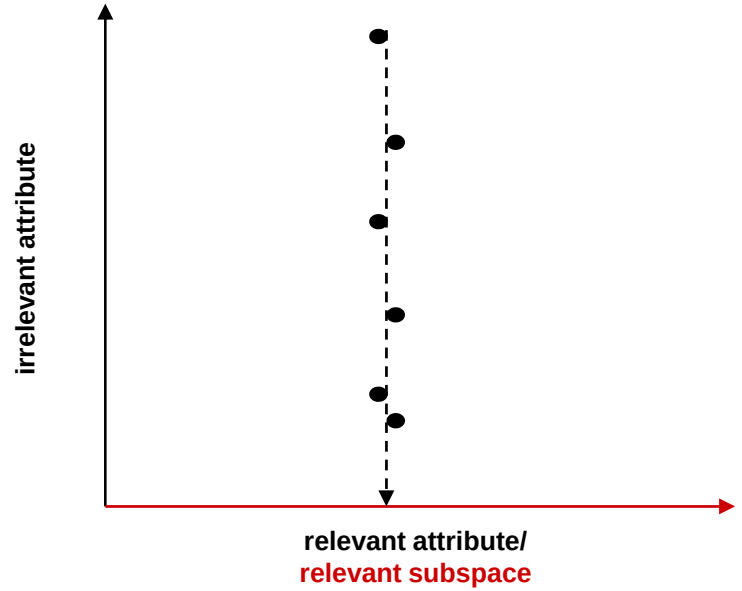



From bottom to top: minimum observed value, average minus standard deviation, average value, average plus standard deviation, maximum observed value, and maximum possible value of the Euclidean norm of a random vector. The expectation grows, but the variance remains constant. A small subinterval of the domain of the norm is reached in practice. (Figure and caption: [FWV07])

- The observations stated in [BGRS99,HAK00] are valid *within* clusters but *not between different* clusters as long as the clusters are well separated [BFG99,FWV07,HKK+10].
- This is *not* the main problem for subspace clustering, although it should be kept in mind for range queries.

# General Problems & Challenges

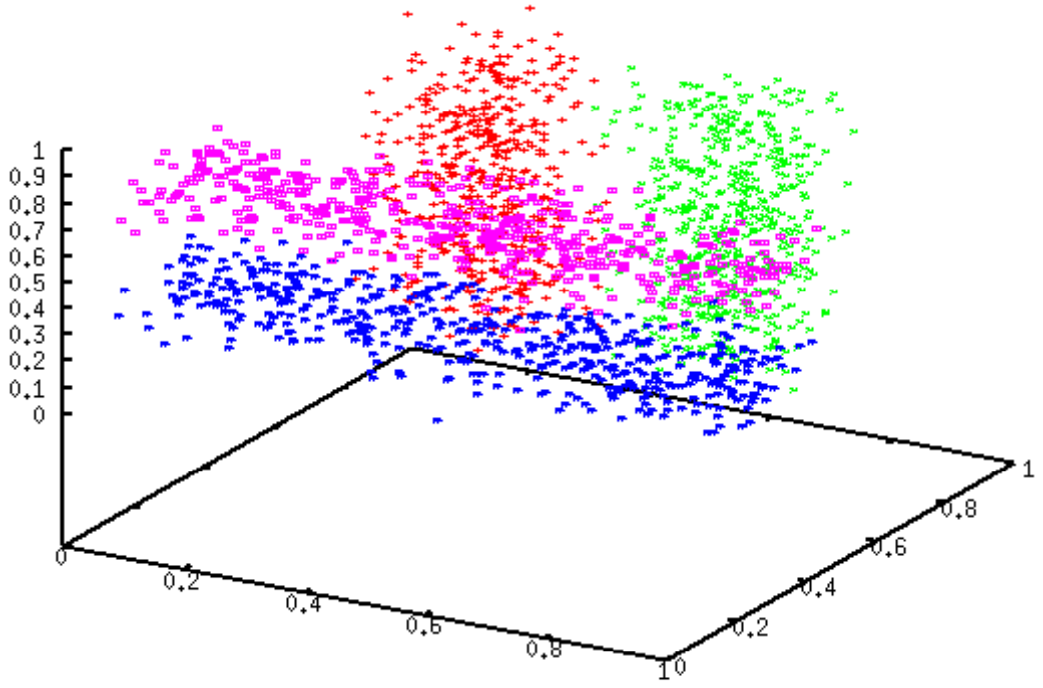
- Third aspect: *Relevant and Irrelevant attributes*
  - A subset of the features may be relevant for clustering
  - Groups of similar (“dense”) points may be identified when considering these features only




  - Different subsets of attributes may be relevant for different clusters

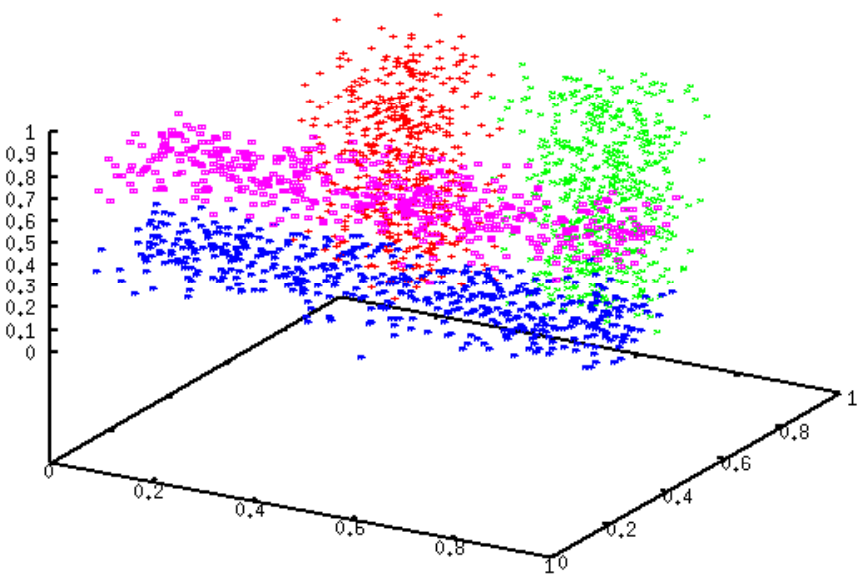
# General Problems & Challenges

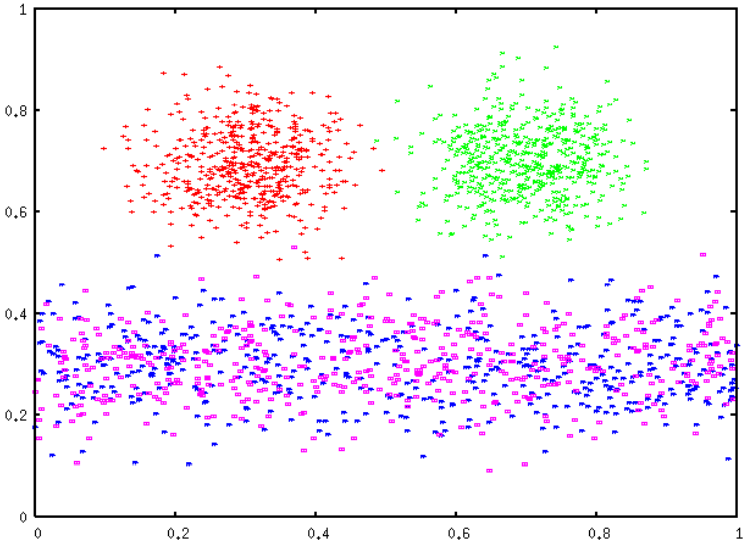
- Effect on clustering:
  - Usually the distance functions used give equal weight to all dimensions
  - However, not all dimensions are of equal importance
  - Adding irrelevant dimensions ruins any clustering based on a distance function that equally weights all dimensions

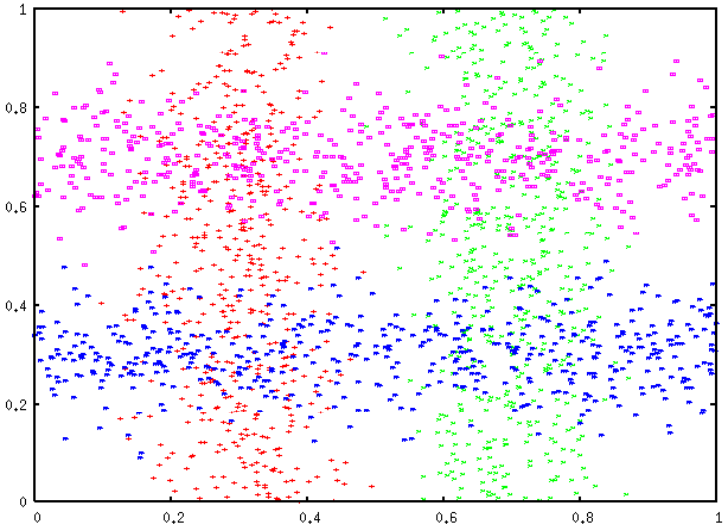


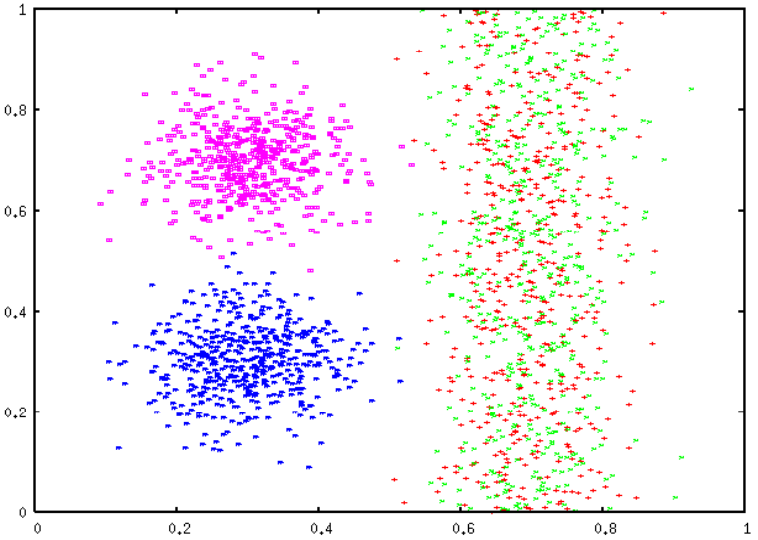
# General Problems & Challenges

- again: different attributes are relevant for different clusters



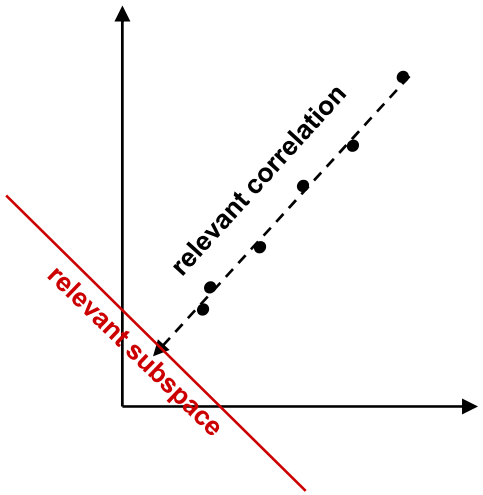



# General Problems & Challenges

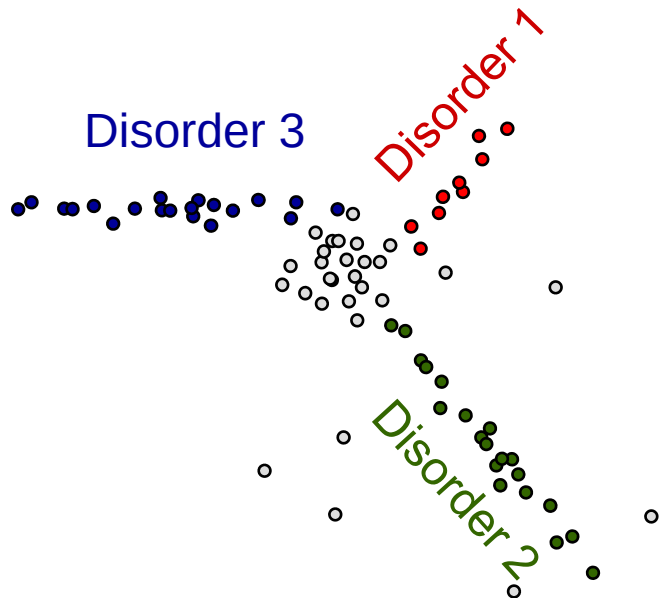
- Fourth aspect: *Correlation among attributes*
  - A subset of features may be correlated
  - Groups of similar (“dense”) points may be identified when considering this correlation of features only




  - Different correlations of attributes may be relevant for different clusters

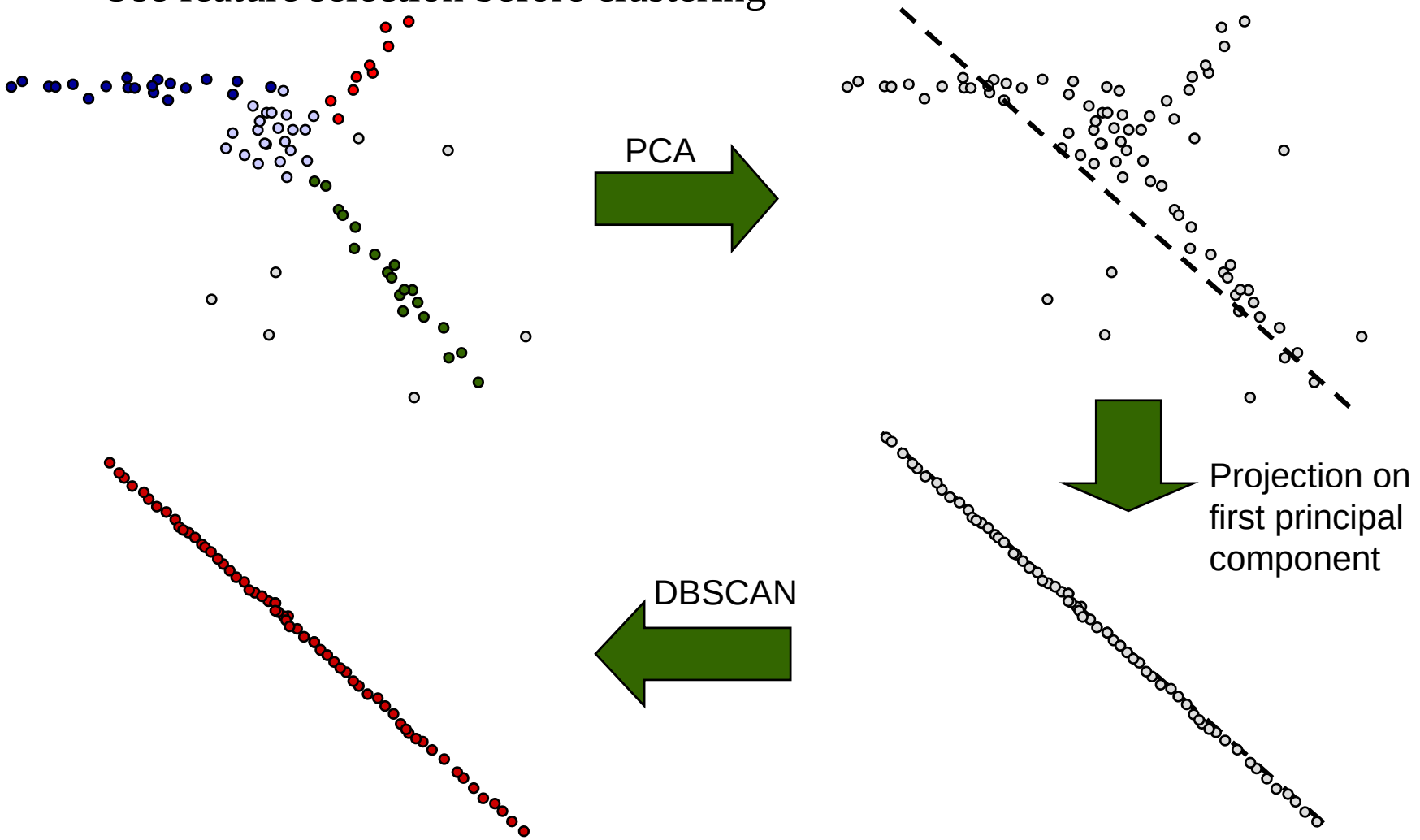
# General Problems & Challenges

- Why not feature selection?
  - (Unsupervised) feature selection is global (e.g. PCA)
  - We face a local feature relevance/correlation: some features (or combinations of them) may be relevant for one cluster, but may be irrelevant for a second one



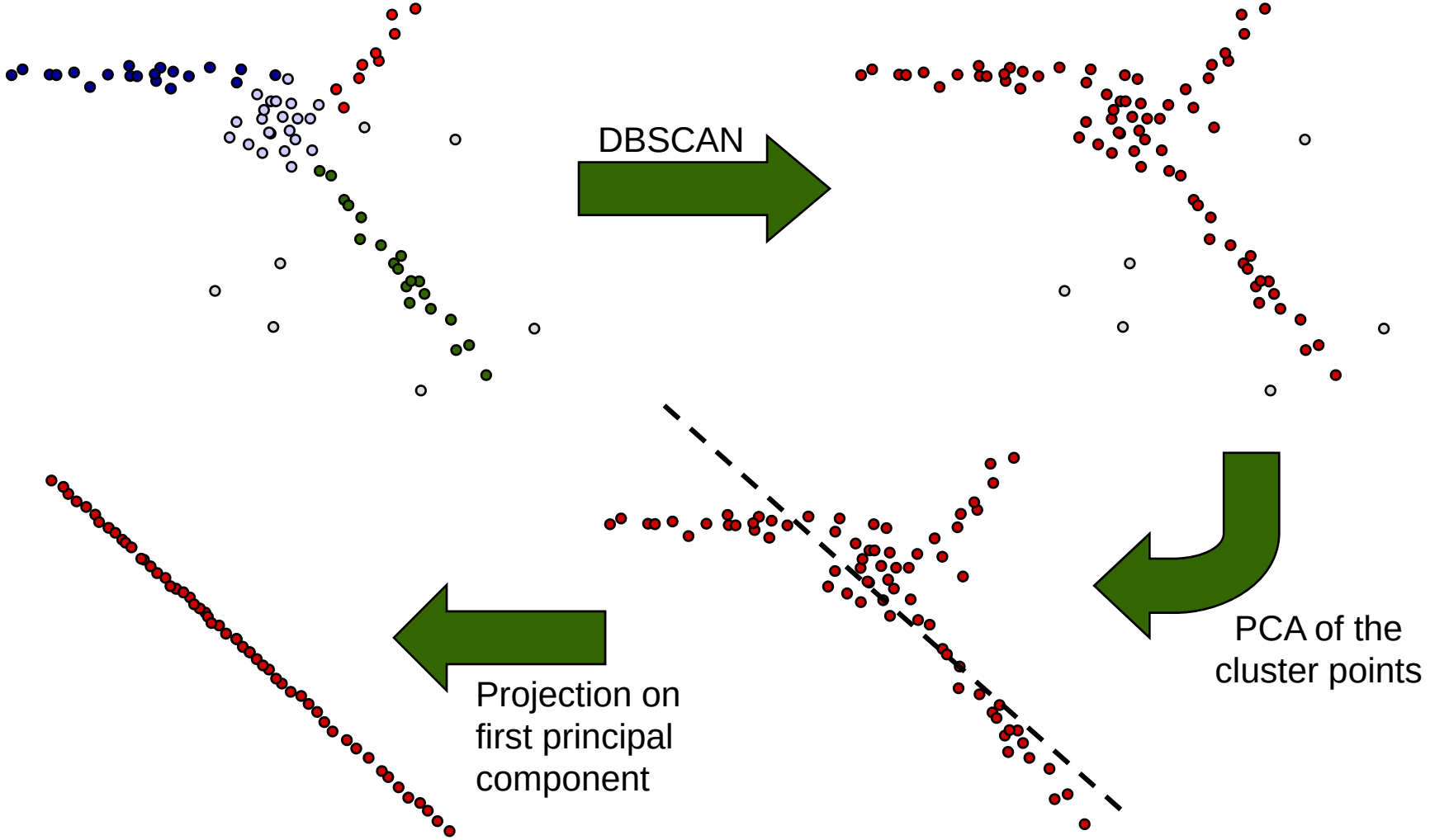
# General Problems & Challenges

- Use feature selection before clustering



# General Problems & Challenges

- Cluster first and then apply PCA



# General Problems & Challenges

- Problem Summary
  - Curse of dimensionality/Feature relevance and correlation
    - Usually, no clusters in the full dimensional space
    - Often, clusters are hidden in subspaces of the data, i.e. only a subset of features is relevant for the clustering
    - E.g. a gene plays a certain role in a subset of experimental conditions
  - Local feature relevance/correlation
    - For each cluster, a different subset of features or a different correlation of features may be relevant
    - E.g. different genes are responsible for different phenotypes
  - Overlapping clusters
    - Clusters may overlap, i.e. an object may be clustered differently in varying subspaces
    - E.g. a gene plays different functional roles depending on the environment

# General Problems & Challenges

- General problem setting of clustering high dimensional data

***Search for clusters in***
***(in general arbitrarily oriented) subspaces***
***of the original feature space***

- Challenges:
  - Find the correct subspace of each cluster
    - Search space:
      - all possible arbitrarily oriented subspaces of a feature space
      - infinite
  - Find the correct cluster in each relevant subspace
    - Search space:
      - “Best” partitioning of points (see: minimal cut of the similarity graph)
      - NP-complete [SCH75]

# General Problems & Challenges

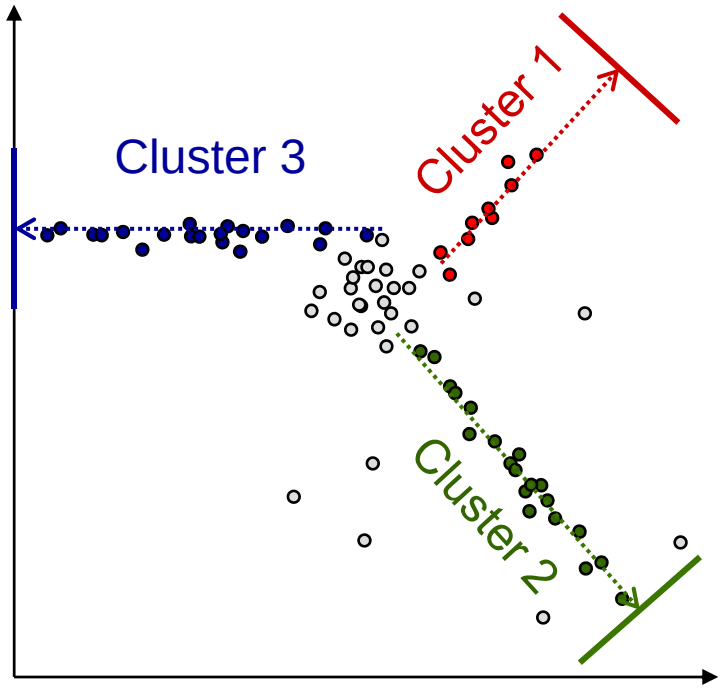
- Even worse: ***Circular Dependency***
  - Both challenges depend on each other
  - In order to determine the correct subspace of a cluster, we need to know (at least some) cluster members
  - In order to determine the correct cluster memberships, we need to know the subspaces of all clusters

- How to solve the circular dependency problem?
  - Integrate subspace search into the clustering process
  - Thus, we need heuristics to solve
    - the clustering problem
    - the subspace search problem

    ***simultaneously***

# General Problems & Challenges

- Solution: integrate variance / covariance analysis into the clustering process
  - Variance analysis:
    - Find clusters in axis-parallel subspaces
    - Cluster members exhibit low variance along the relevant dimensions
  - Covariance/correlation analysis:
    - Find clusters in arbitrarily oriented subspaces
    - Cluster members exhibit a low covariance w.r.t. a given combination of the relevant dimensions (i.e. a low variance along the dimensions of the arbitrarily oriented subspace corresponding to the given combination of relevant attributes)



# Outline

1. Sample Applications

2. General Problems and Challenges: the Curse of Dimensionality

3. A First Taxonomy of Approaches

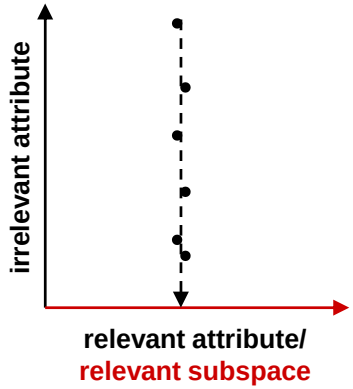
4. Arbitrarily-oriented Subspace Clustering
   1. PCA-Based Approaches
   2. Correlation Clustering Based on the Hough-Transform

# A First Taxonomy of Approaches

- So far, we can distinguish between
  - Clusters in axis-parallel subspaces

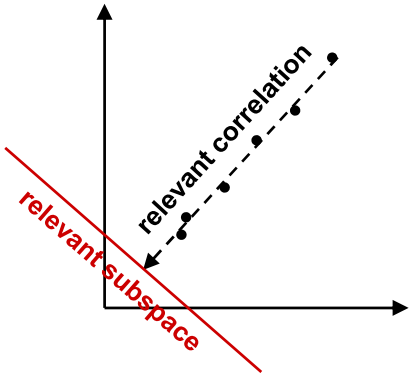
    (common assumption to restrict the search space)

    Approaches are usually called
    - "subspace clustering algorithms"
    - "projected clustering algorithms"
    - "bi-clustering or co-clustering algorithms"



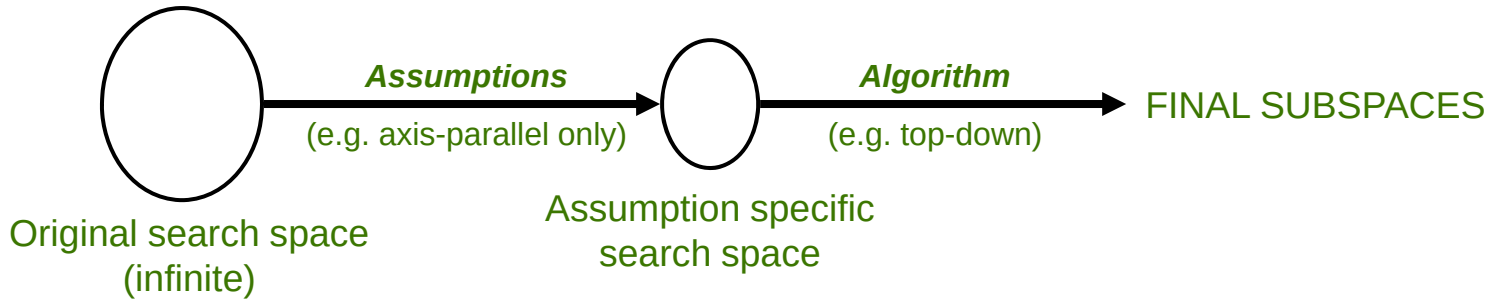

  - Clusters in arbitrarily oriented subspaces

    Approaches are usually called
    - "bi-clustering or co-clustering algorithms"
    - "pattern-based clustering algorithms"
    - "correlation clustering algorithms"

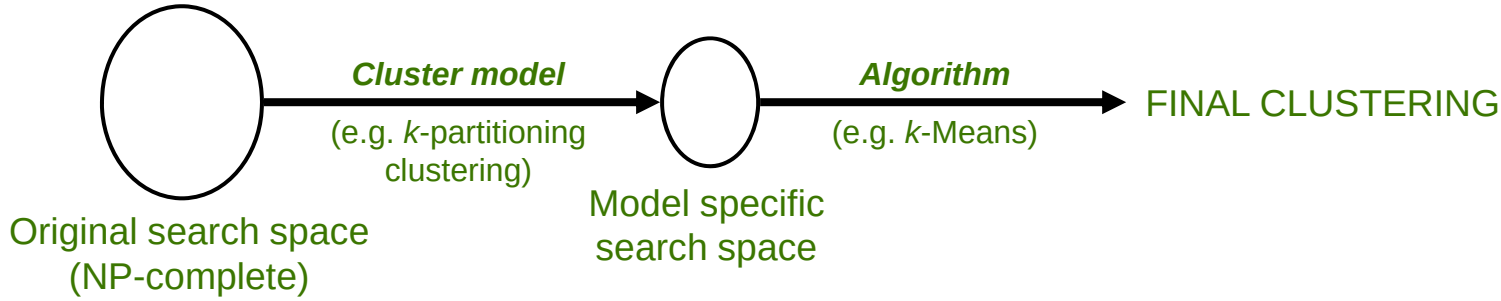


# A First Taxonomy of Approaches

- A first big picture
  - We have two problems to solve
  - For both problems we need heuristics that have huge influence on the properties of the algorithms
    - ***Subspace search***




    - ***Cluster search***



# A First Taxonomy of Approaches

- Restricted on *axis-parallel subspaces* – what are we searching for?
  - Overlapping clusters: points may be grouped differently in different subspaces

    => "***subspace clustering***"
  - Disjoint partitioning: assign points uniquely to clusters (or noise)

    => "***projected clustering***"

  *Notes:*

  - The terms **subspace** clustering and **projected** clustering are not used in a unified or consistent way in the literature
  - These two problem definitions are products of the presented algorithms:
    - The first "projected clustering algorithm" integrates a distance function accounting for clusters in subspaces into a "flat" clustering algorithm (k-medoid) => DISJOINT PARTITION
    - The first "subspace clustering algorithm" is an application of the APRIORI algorithm => ALL CLUSTERS IN ALL SUBSPACES

# A First Taxonomy of Approaches

- Restricted on *axis-parallel subspaces* – how are we searching?
- Basically, there are two different ways to efficiently navigate through the search space of possible subspaces

Bottom-up:

If the cluster criterion implements the downward closure, one can use any bottom-up frequent itemset mining algorithm (e.g. APRIORI [AS94])
*Key*: downward-closure property OR merging-procedure
Example approaches:
[AGGR98, CFZ99, NGC01, KKK04, KKRW05, MSE06, ABK+07a]

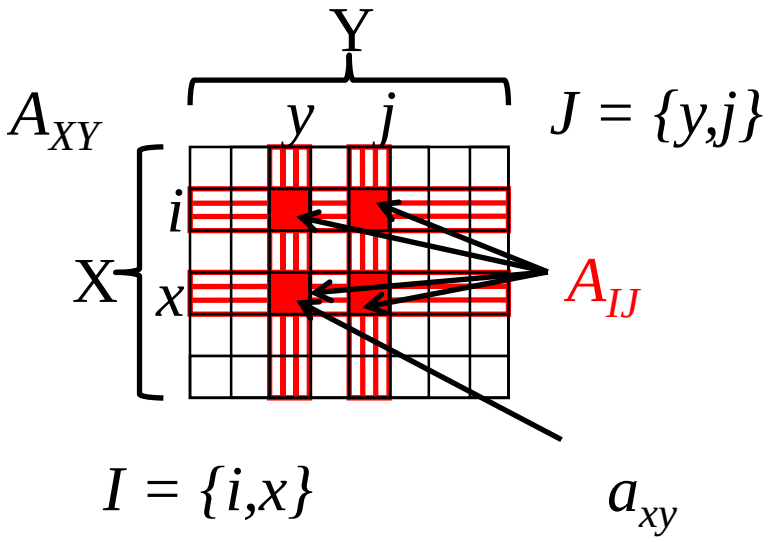
Top-down:

The search starts in the full *d*-dimensional space and iteratively learns for each point or each cluster the correct subspace
*Key*: procedure to learn the correct subspace
Example approaches: [APW+99, BKKK04]

# A First Taxonomy of Approaches

*Pattern-based clustering* relies on patterns in the data matrix.

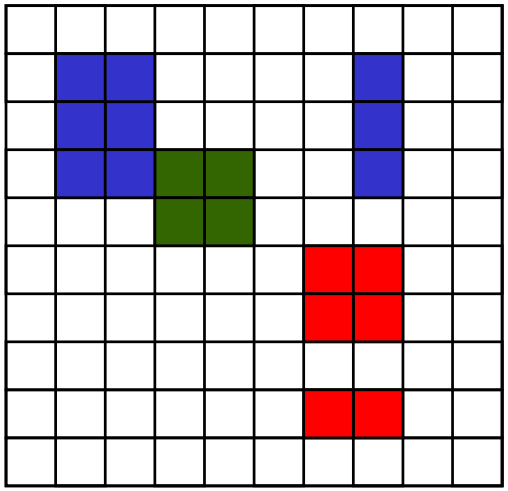
- Simultaneous clustering of rows and columns of the data matrix (hence *bi*clustering).
  - Data matrix $\boldsymbol{A}$ = (X,Y) with set of rows X and set of columns Y
  - $a_{xy}$ is the element in row $x$ and column $y$.
  - submatrix $A_{IJ}$ = (I,J) with subset of rows I ⊆ X and subset of columns J ⊆ Y contains those elements $a_{ij}$ with $i \in$ I und $j \in$ J



# A First Taxonomy of Approaches

General aim of biclustering approaches:

Find a set of submatrices $\{(I_1,J_1),(I_2,J_2),\ldots,(I_k,J_k)\}$ of the matrix $\boldsymbol{A}=(X,Y)$ (with $I_i \subseteq X$ and $J_i \subseteq Y$ for $i = 1,\ldots,k$) where each submatrix (= bicluster) meets a given homogeneity criterion.



Sounds similar to subspace clustering but:
the *homogeneity criterion* is completely *different*!

# A First Taxonomy of Approaches

- Some values often used by bicluster models:
  - mean of row *i*:

$$a_{iJ} = \frac{1}{|J|} \sum_{j \in J} a_{ij}$$

  - mean of column *j*:

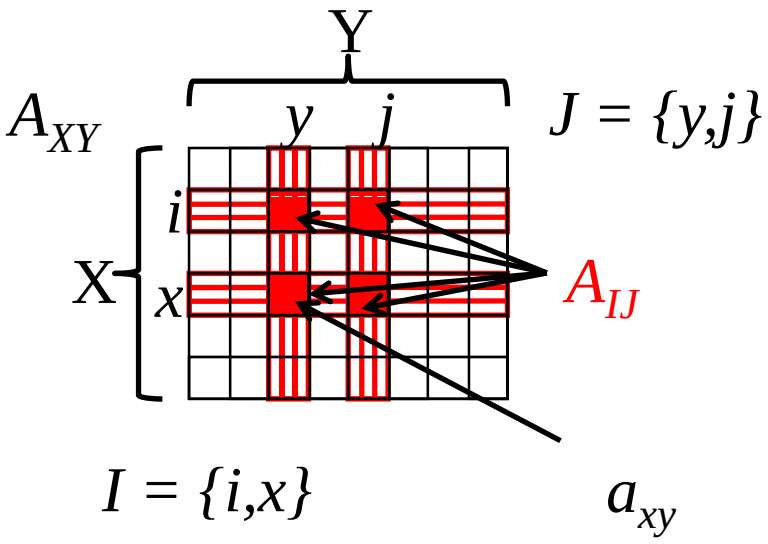
$$a_{Ij} = \frac{1}{|I|} \sum_{i \in I} a_{ij}$$

  - mean of all elements:

$$a_{IJ} = \frac{1}{|I||J|} \sum_{i \in I, j \in J} a_{ij}$$

$$= \frac{1}{|J|} \sum_{j \in J} a_{Ij}$$

$$= \frac{1}{|I|} \sum_{i \in I} a_{iJ}$$



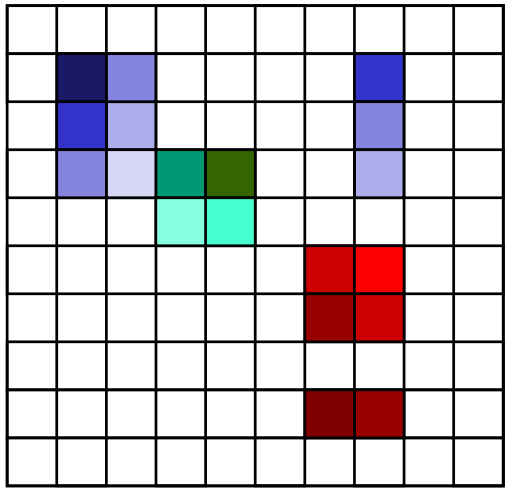
# A First Taxonomy of Approaches

most common model (following Cheng & Church [CC00]):
*biclusters* with *coherent values*

- based on a particular form of covariance between rows and columns

$$a_{ij} = \mu + r_i + c_j$$

$$\forall i \in I, j \in J$$

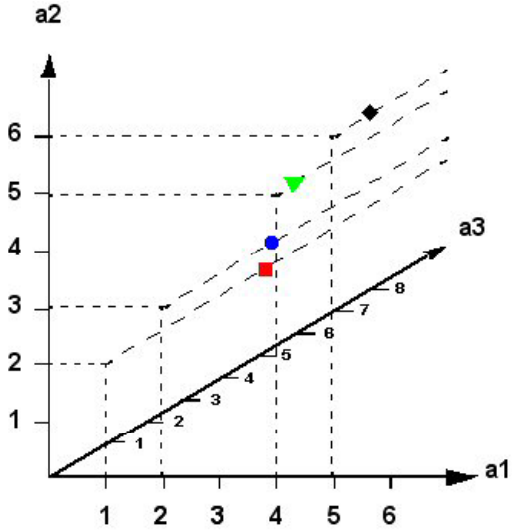


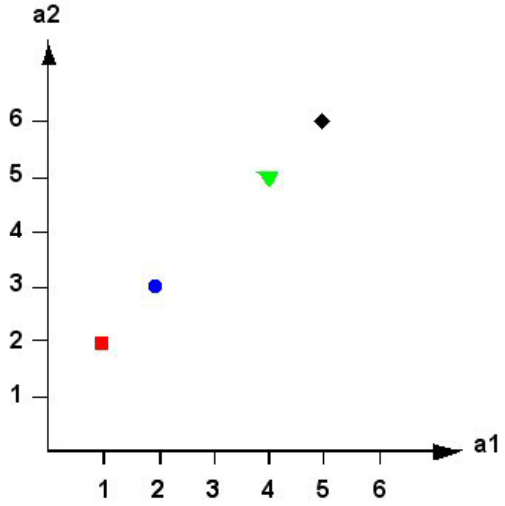
- special cases:
  - $c_j = 0$ for all $j$ → constant values on rows
  - $r_i = 0$ for all $i$ → constant values on columns

# A First Taxonomy of Approaches

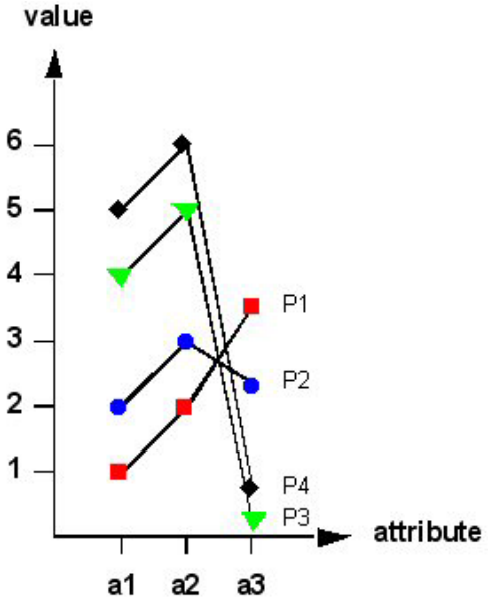
- embedding space: sparse hyperplane parallel to axes of irrelevant attributes

| | a1 | a2 | a3 |
|---|---|---|---|
| P1 | 1 | 2 | 3.5 |
| P2 | 2 | 3 | 2.3 |
| P3 | 4 | 5 | 0.2 |
| P4 | 5 | 6 | 0.7 |



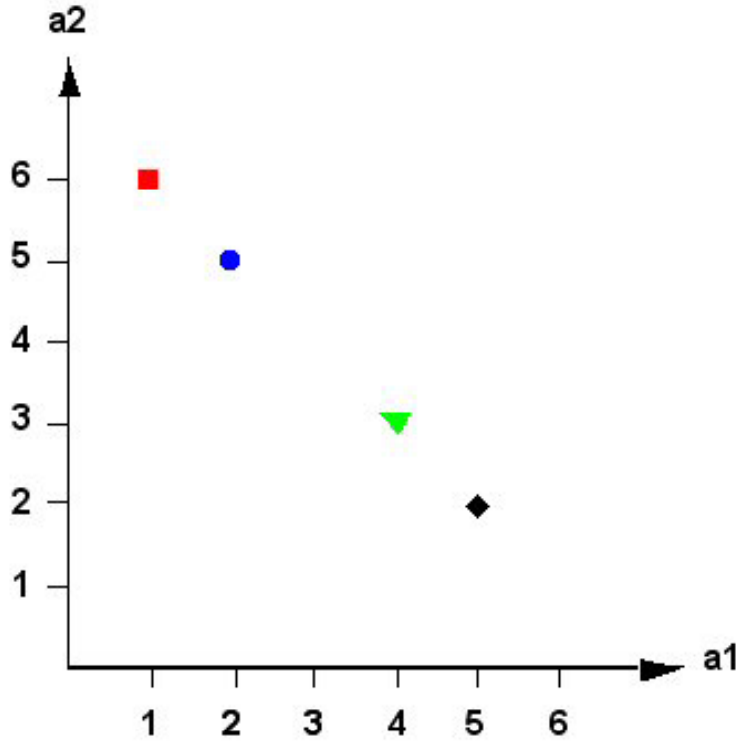



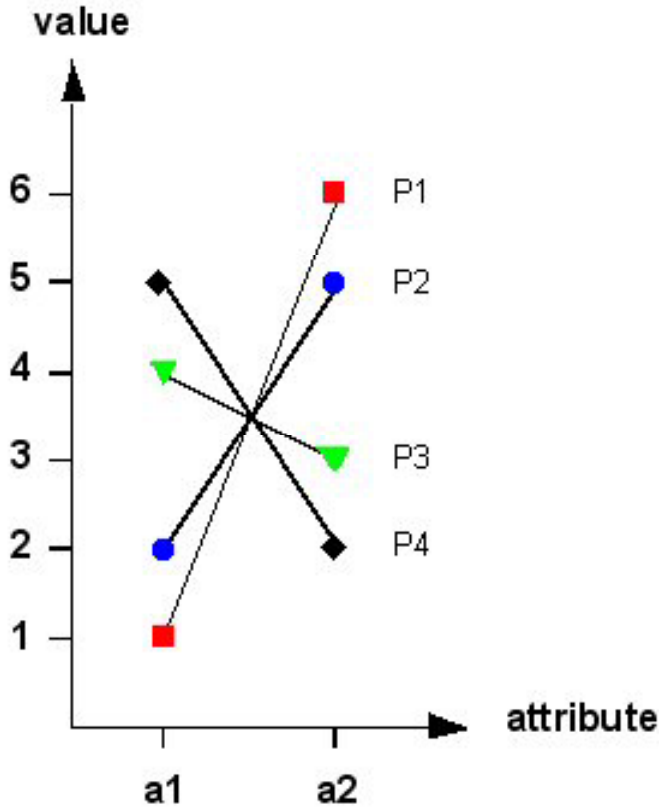

- subspace: increasing one-dimensional line
- pattern (parallel coordinates-plot): parallel lines



# A First Taxonomy of Approaches

- Pattern-based approaches find simple positive correlations
- negative correlations: no additive pattern



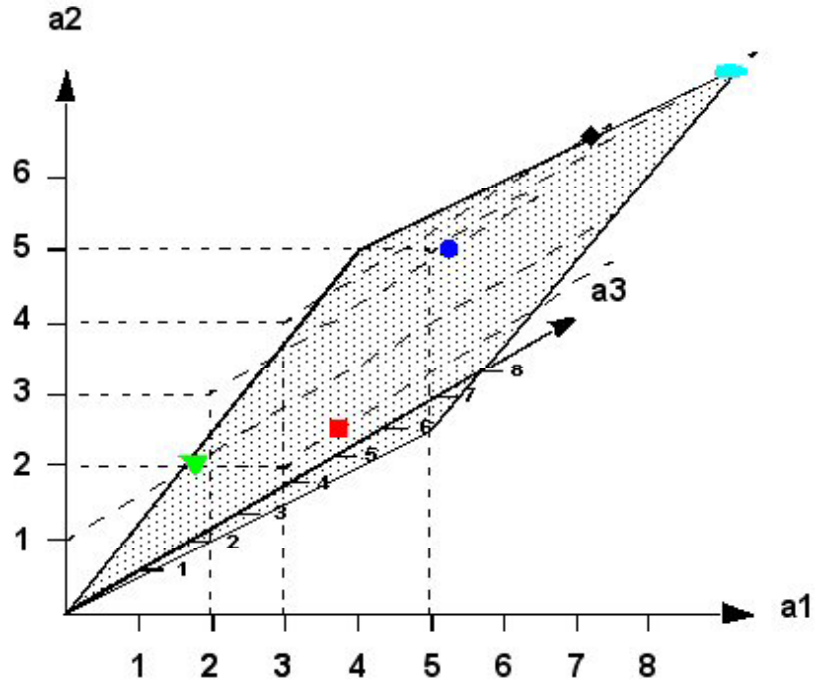



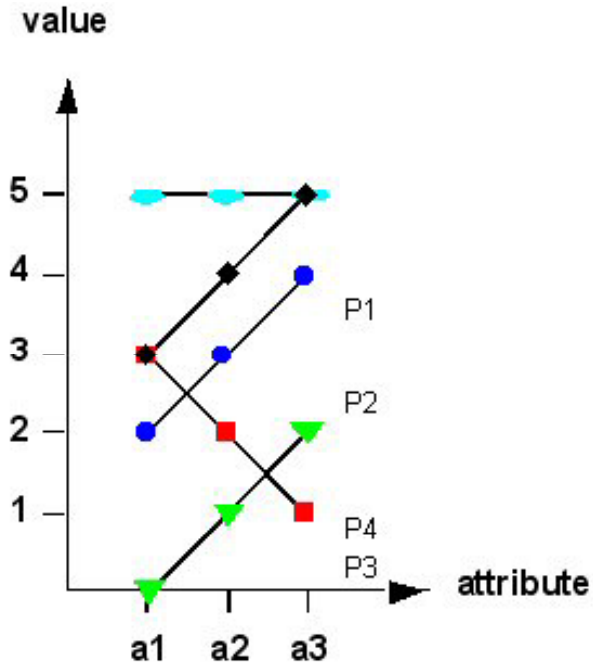
# A First Taxonomy of Approaches

- more complex correlations: out of scope of pattern-based approaches

$a1 - 2 \cdot a2 + a3 = 0$

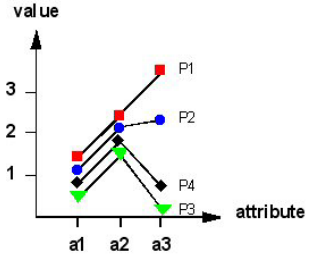
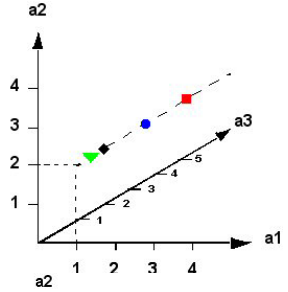
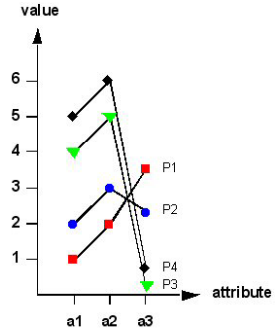
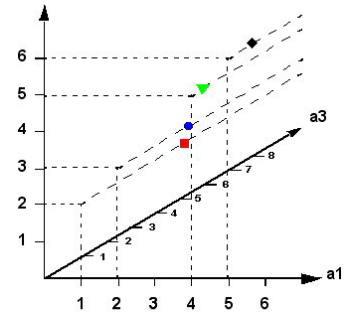
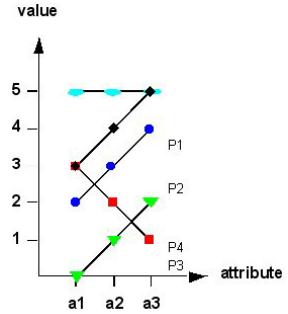
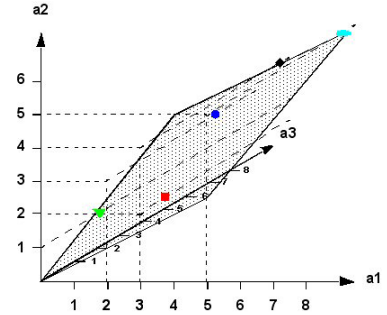
| | a1 | a2 | a3 |
|---|---|---|---|
| P1 | 3 | 2 | 1 |
| P2 | 2 | 3 | 4 |
| P3 | 0 | 1 | 2 |
| P4 | 3 | 4 | 5 |
| P5 | 5 | 5 | 6 |




- interesting subspace is arbitrarily oriented, related to complex correlations among attributes → *Correlation Clustering*

# A First Taxonomy of Approaches

| Matrix-Pattern | | Problem | Spatial Pattern | |
|---|---|---|---|---|
|  | Constant values in columns, change of values only on rows | Subspace / Projected Clustering | Axis-parallel hyperplanes |  |
|  | From constant values in rows and columns (no change of values) to arbitrary change of values in common direction | Pattern-based / Bi-Clustering | Special cases of axis-parallel to special cases of arbitrarily oriented hyperplanes |  |
|  | No particular pattern | Correlation Clustering | Arbitrarily oriented hyperplanes |  |

# A First Taxonomy of Approaches

- Note: this taxonomy considers only the subspace search space
- the clustering search space is equally important
- other important aspects for classifying existing approaches are e.g.
  - The underlying cluster model that usually involves
    - Input parameters
    - Assumptions on number, size, and shape of clusters
    - Noise (outlier) robustness
  - Determinism
  - Independence w.r.t. the order of objects/attributes
  - Assumptions on overlap/non-overlap of clusters/subspaces
  - Efficiency

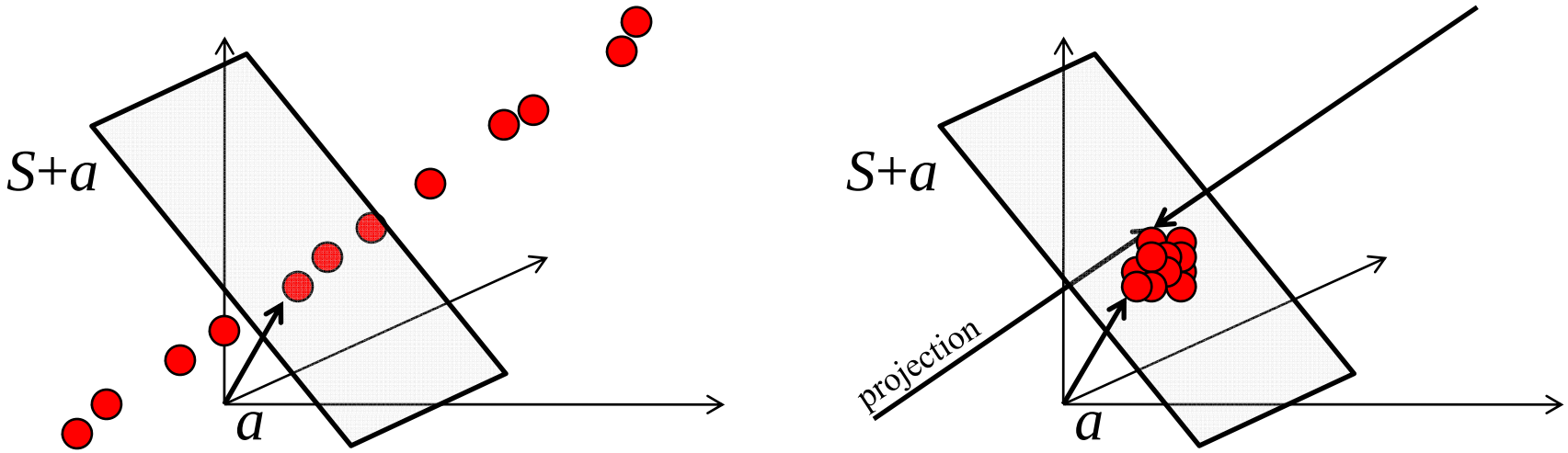
Extensive survey: [KKZ09]
http://doi.acm.org/10.1145/1497577.1497578

# Outline

1. Sample Applications
2. General Problems and Challenges: the Curse of Dimensionality
3. A First Taxonomy of Approaches
4. Arbitrarily-oriented Subspace Clustering
    1. PCA-Based Approaches
    2. Correlation Clustering Based on the Hough-Transform

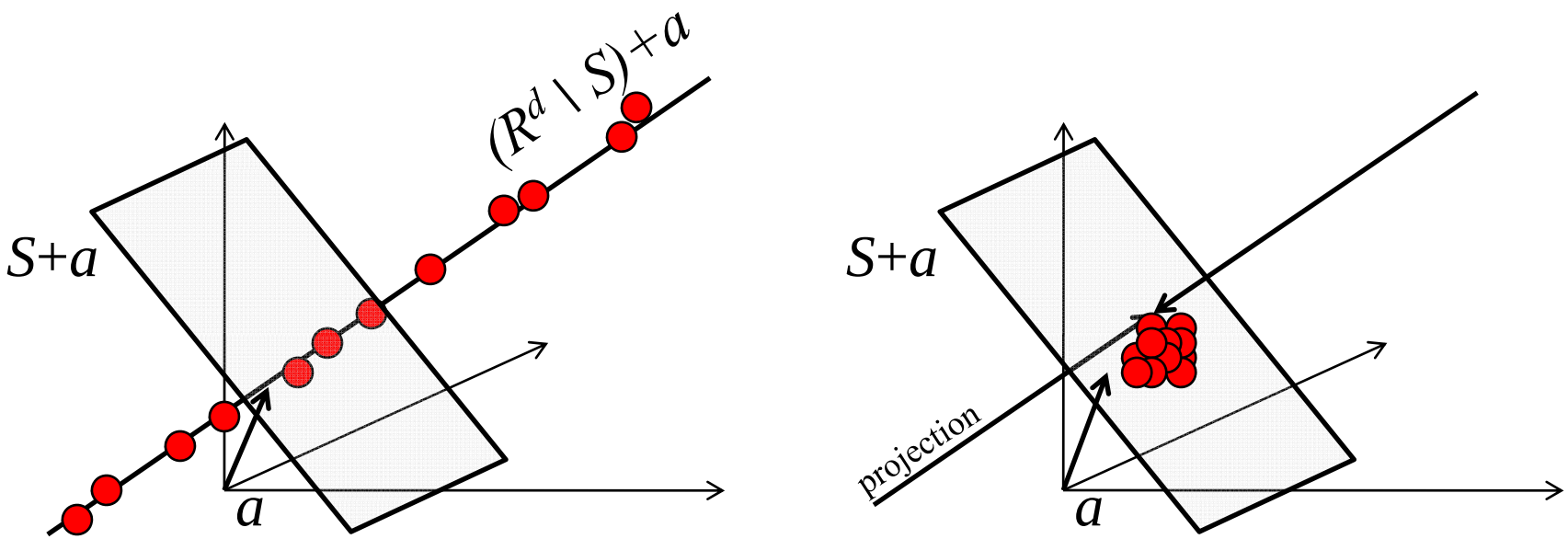
# PCA-based Approaches

- Pattern-based approaches find pairwise positive correlations
- More general approach: oriented clustering aka. generalized subspace/projected clustering aka. correlation clustering
- Assumption: any cluster is located in an arbitrarily oriented affine subspace $S+a$ of $R^d$



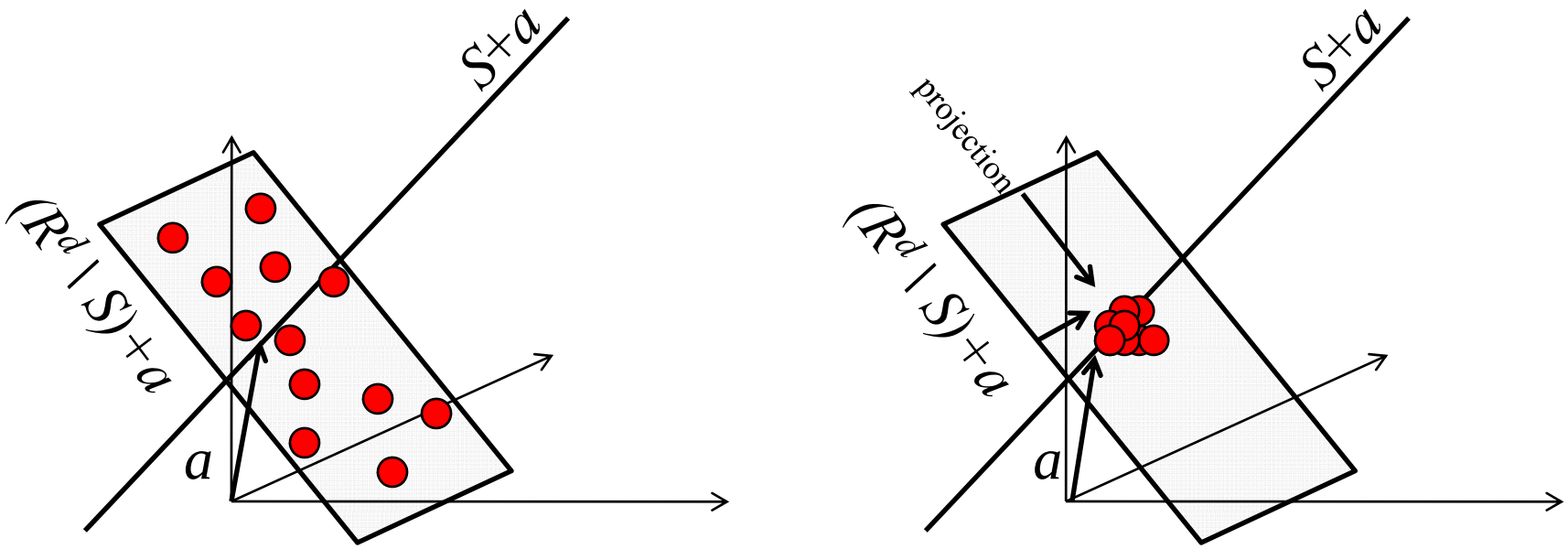
# PCA-based Approaches

- Affine subspace $S+a$, $S \subset R^d$, *affinity* $a \in R^d$ is interesting if a set of points clusters within this subspace
- Points may exhibit high variance in perpendicular subspace $(R^d \setminus S)+a$



# PCA-based Approaches

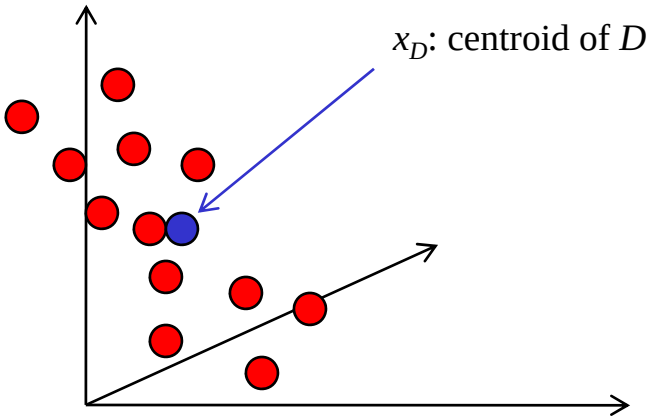
- high variance in perpendicular subspace $(R^d \setminus S)+a$ → points form a hyperplane within $R^d$ located in this subspace $(R^d \setminus S)+a$
- Points on a hyperplane appear to follow linear dependencies among the attributes participating in the description of the hyperplane



# PCA-based Approaches

- Directions of high/low variance: PCA (local application)
- locality assumption: local selection of points sufficiently reflects the hyperplane accommodating the points
- general approach: build covariance matrix $\Sigma_D$ for a selection $D$ of points (e.g. $k$ nearest neighbors of a point)

$$\Sigma_D = \frac{1}{|D|}\sum_{x \in D}(x - x_D)(x - x_D)^T$$




properties of $\Sigma_D$:

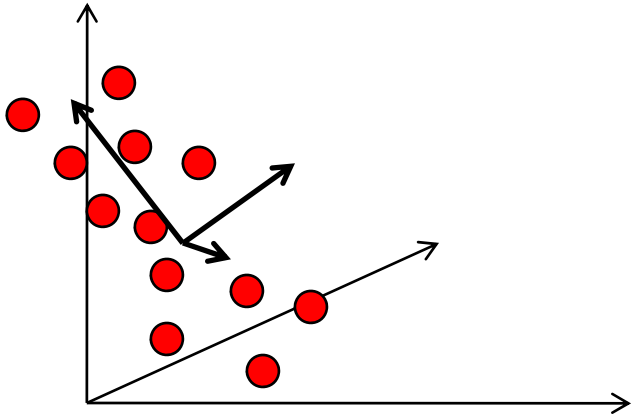
- $d$ x $d$
- symmetric
- positive semidefinite
- $\sigma_{D_{ij}}$ (value at row $i$, column $j$) = covariance between dimensions $i$ and $j$
- $\sigma_{D_{ii}}$ = variance in $i$th dimension

# PCA-based Approaches

- decomposition of $\Sigma_D$ to eigenvalue matrix $E_D$ and eigenvector matrix $V_D$:

$$\Sigma_D = V_D E_D V_D^{\mathrm{T}}$$

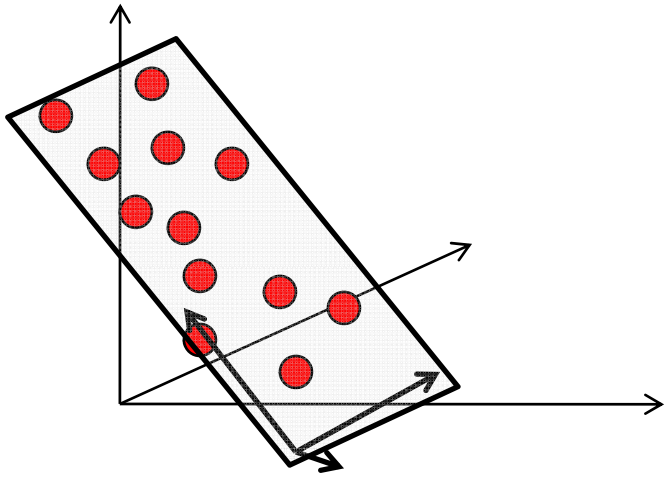
- $E_D$ : diagonal matrix, holding eigenvalues of $\Sigma_D$ in decreasing order in its diagonal elements
- $V_D$ : orthonormal matrix with eigenvectors of $\Sigma_D$ ordered correspondingly to the eigenvalues in $E_D$



- $V_D$ : new basis, first eigenvector = direction of highest variance
- $E_D$ : covariance matrix of D when represented in new axis system $V_D$

# PCA-based Approaches

- points forming λ-dimensional hyperplane → hyperplane is spanned by the first λ eigenvectors (called "strong" eigenvectors – notation: $\breve{V}_D$ )
- subspace where the points cluster densely is spanned by the remaining *d*-λ eigenvectors (called "weak" eigenvectors – notation: $\hat{V}_D$ )



for the eigensystem, the sum of the smallest *d*-λ eigenvalues $\sum_{i=\lambda+1}^{d} e_{D_{ii}}$ is minimal under all possible transformations → points cluster optimally dense in this subspace

# PCA-based Approaches

model for correlation clusters [ABK+06]:

- $\lambda$-dimensional hyperplane accommodating the points of a correlation cluster $C \subset R^d$ is defined by an equation system of $d$-$\lambda$ equations for $d$ variables and the affinity (e.g. the mean point $x_C$ of all cluster members):

$$\hat{V}_C^{\mathrm{T}} x = \hat{V}_C^{\mathrm{T}} x_C$$

- equation system approximately fulfilled for all points $x \in C$
- quantitative model for the cluster allowing for probabilistic prediction (classification)
- Note: correlations are observable, linear dependencies are merely an assumption to explain the observations – predictive model allows for evaluation of assumptions and experimental refinements

Examples of PCA based correlation clustering:

[AY00, BKKZ04, ABK+07c,ABK+07b]

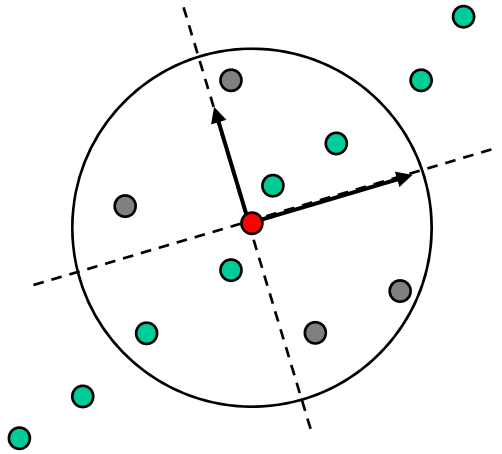
# Outline

1. Sample Applications
2. General Problems and Challenges: the Curse of Dimensionality
3. A First Taxonomy of Approaches
4. Arbitrarily-oriented Subspace Clustering
    1. PCA-Based Approaches
    2. Correlation Clustering Based on the Hough-Transform

# Correlation Clustering Based on the Hough-Transform

different correlation primitive: Hough-transform

- problems of PCA based approaches: locality assumption
  - characteristic neighborhood?
  - PCA sensitive for outliers in local neighborhoods
  - choice of $\lambda$?
  - “locality assumption” questionable in view of the “curse of dimensionality”



# Correlation Clustering Based on the Hough-Transform

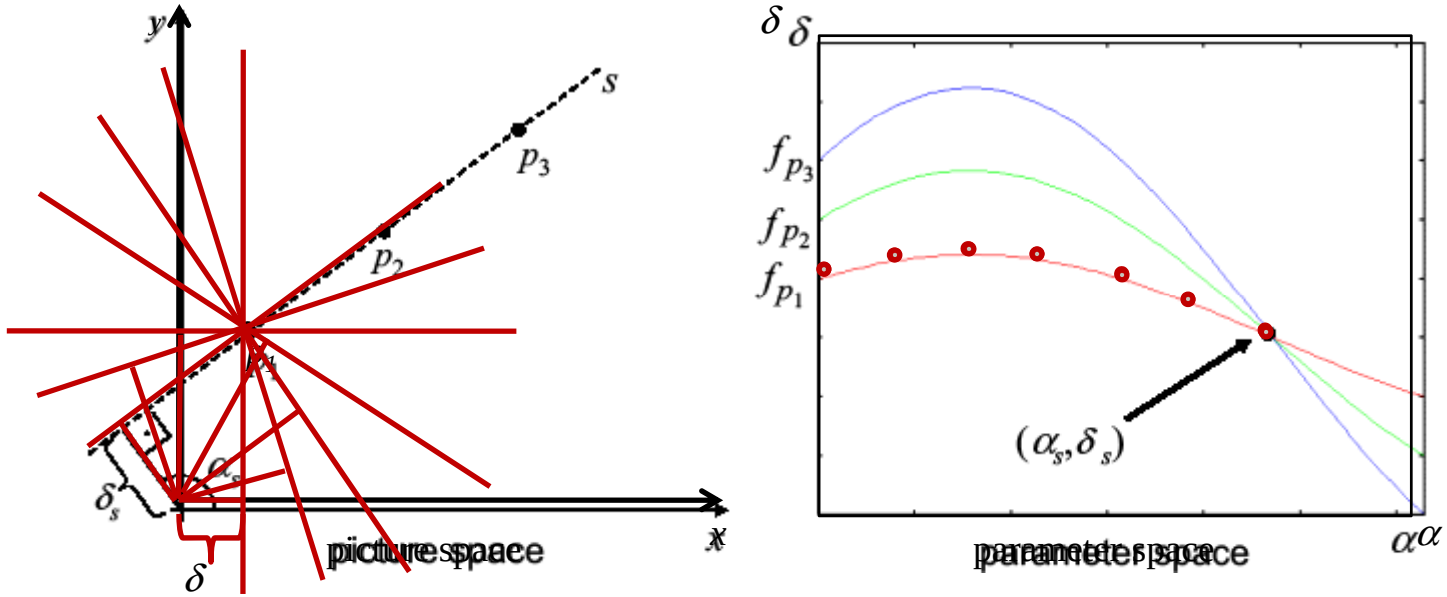
- Hough-transform:
  - developed in computer-graphics
  - 2-dimensional (image procesing)
- CASH: Clustering in Arbitrary Subspaces based on the Hough-Transform [ABD+08]
  - generalization to *d*-dimensional spaces
  - transfer of the clustering to a new space ("Parameter-space" of the Hough-transform)
  - restriction of the search space

    (from innumerable infinite to $O(n!)$)
  - common search heuristic for Hough-transform: $O(2^d)$

    $\rightarrow$ efficient search heuristic

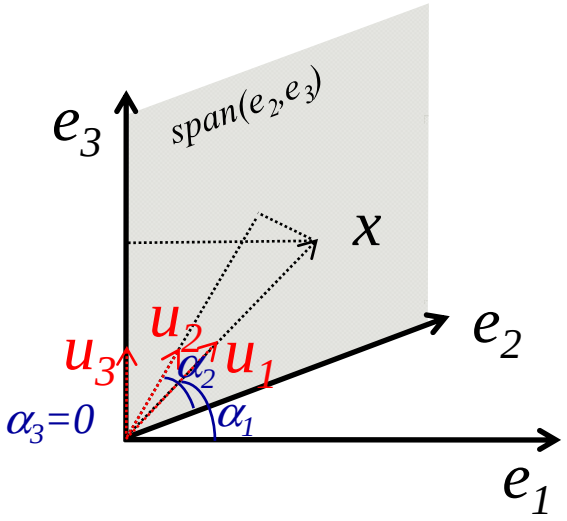
# Correlation Clustering Based on the Hough-Transform

- given: $D \subseteq \Re^d$
- find linear subspaces accommodating many points
- Idea: map points from data space (picture space) onto functions in parameter space



# Correlation Clustering Based on the Hough-Transform

- $e_i$, $1 \leq i \leq d$: orthonormal-basis
- $x = (x_1,\ldots,x_d)^T$: $d$-dimensional vector onto hypersphere around the origin with radius $r$
- $u_i$: unit-vector in direction of projection of $x$ onto subspace $span(e_i,\ldots,e_d)$
- $\alpha_1,\ldots,\alpha_{d-1}$: $\alpha_i$ angle between $u_i$ and $e_i$

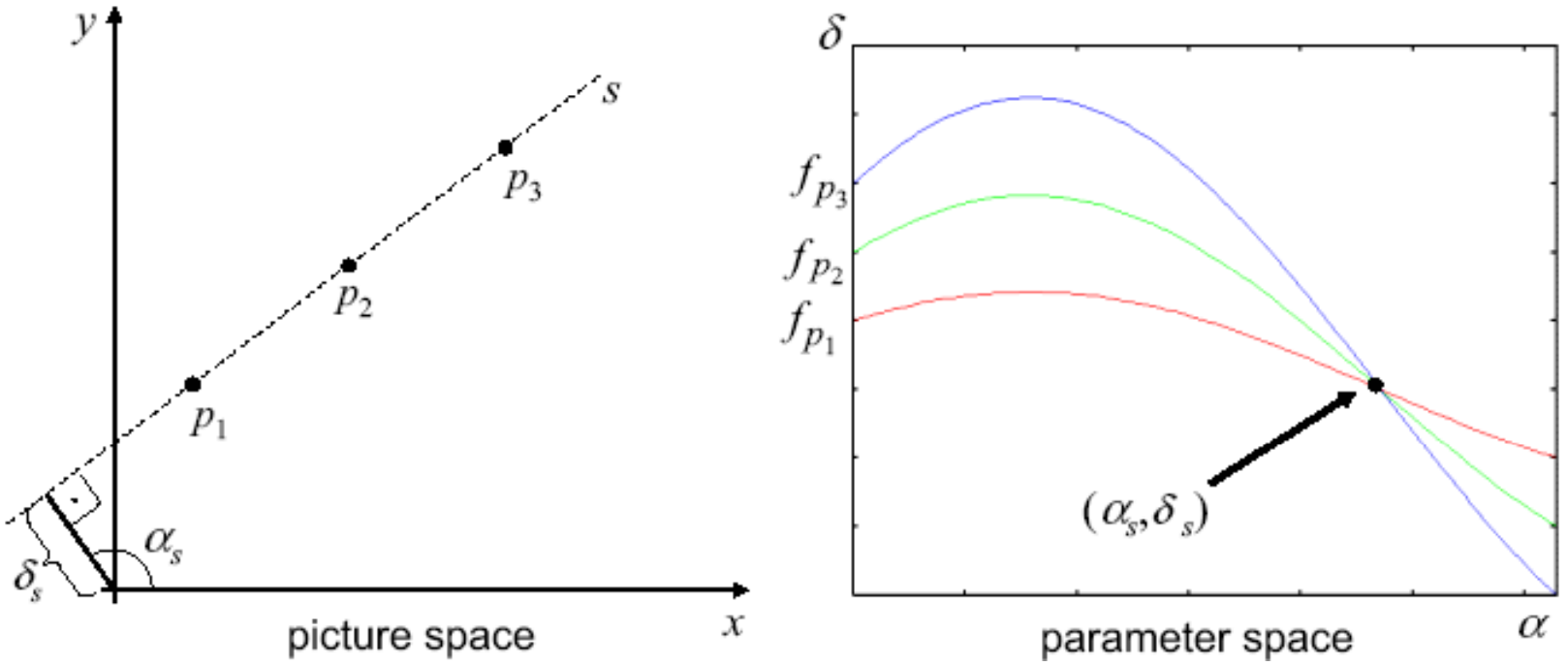



$$x_i = r \cdot \left( \prod_{j=1}^{i-1} \sin(\alpha_j) \right) \cdot \cos(\alpha_i)$$

# Correlation Clustering Based on the Hough-Transform

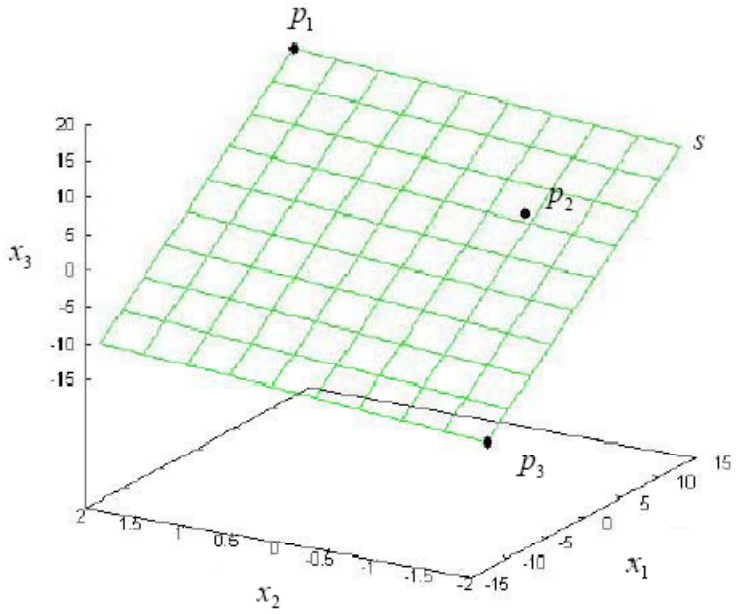
Length $\delta$ of the normal vector $\delta \cdot \vec{n}$ with $\|\vec{n}\| = 1$ and angles $\alpha_1,\ldots,\alpha_{d-1}$ for the line through point $p$:

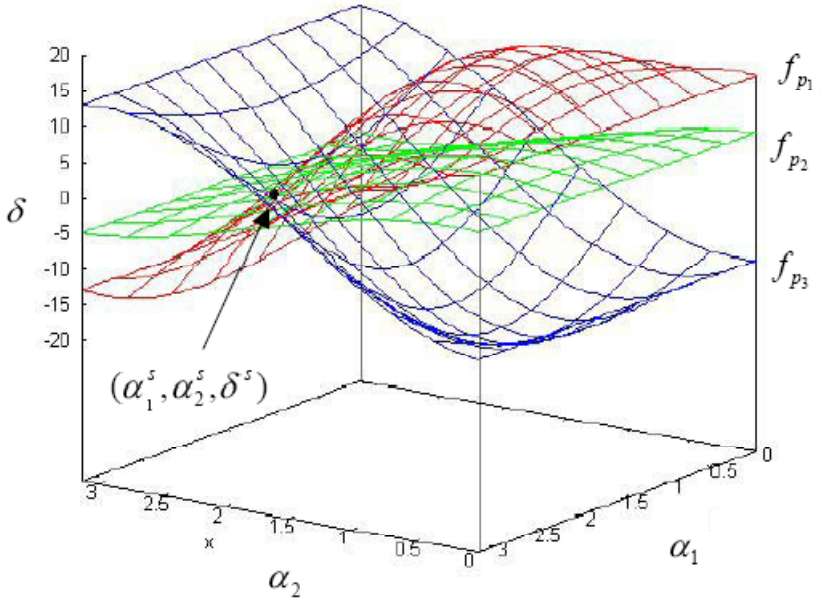
$$f_p(\alpha_1,\ldots,\alpha_{d-1}) = \langle p, n \rangle = \sum_{i=1}^{d} p_i \cdot \left( \prod_{j=1}^{i-1} \sin(\alpha_j) \right) \cdot \cos(\alpha_i)$$



# Correlation Clustering Based on the Hough-Transform

- Properties of the transformation
  - Point in the data space = sinusoidal curve in parameter space
  - Point in parameter space = hyper-plane in data space
  - Points on a common hyper-plane in data space = sinusoidal curves through a common point in parameter space
  - Intersections of sinusoidal curves in parameter space = hyper-plane through the corresponding points in data space

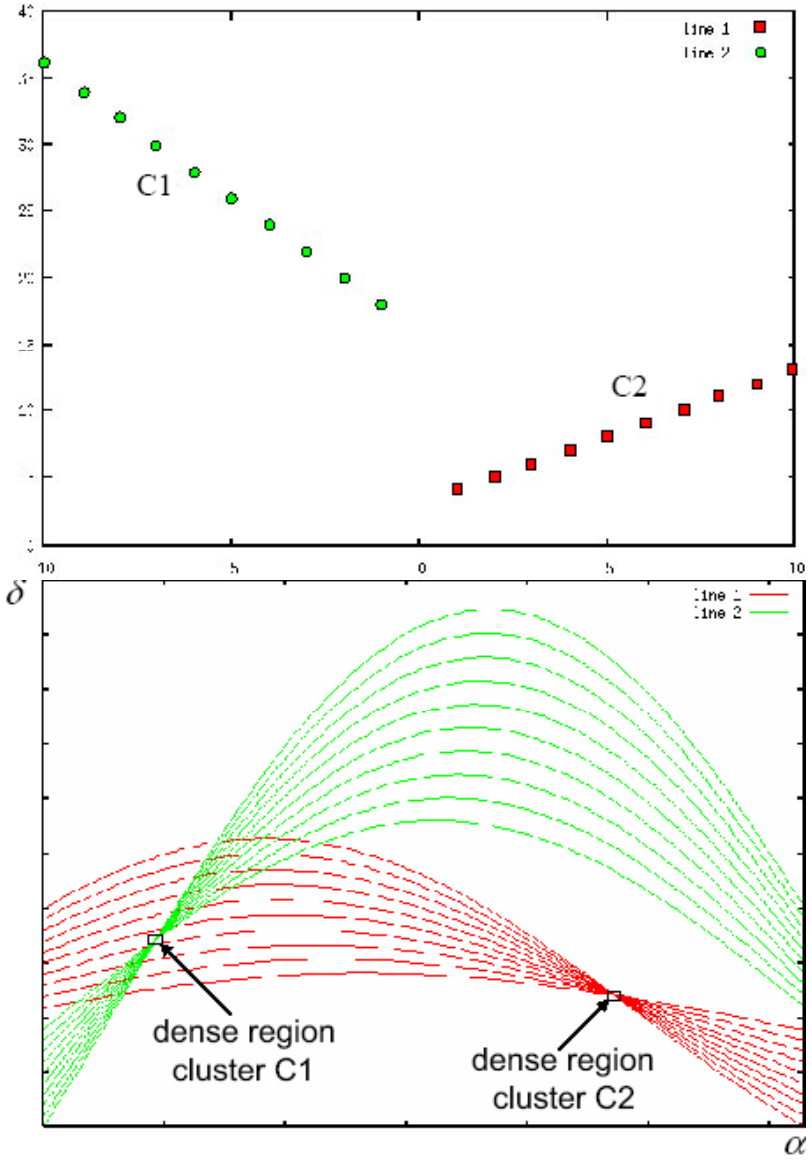


# Correlation Clustering Based on the Hough-Transform

- dense regions in parameter space $\Leftrightarrow$ linear structures in data space (hyperplanes with $\lambda \leq d\text{-}1$)
- exact solution: find all intersection points
  - infeasible
  - to exact
- approximative solution: grid-based clustering in parameter space

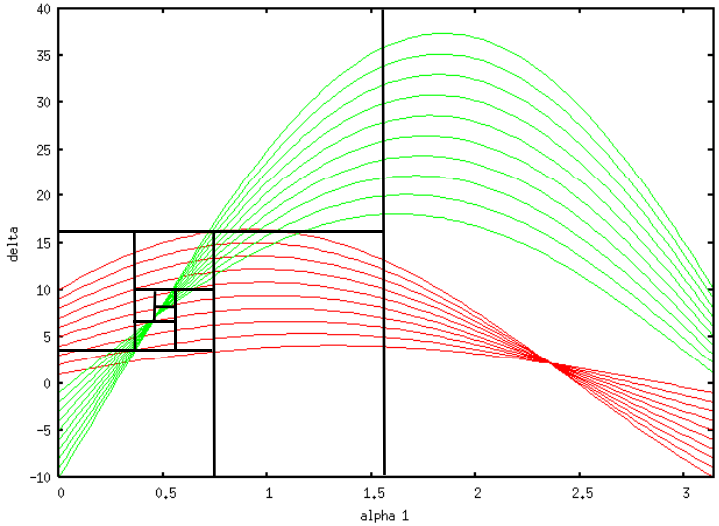
  $\rightarrow$ find grid cells intersected by at least $m$ sinusoids
  - search space bounded but in $O(r^d)$
  - pure clusters require large value for $r$ (grid solution)



# Correlation Clustering Based on the Hough-Transform

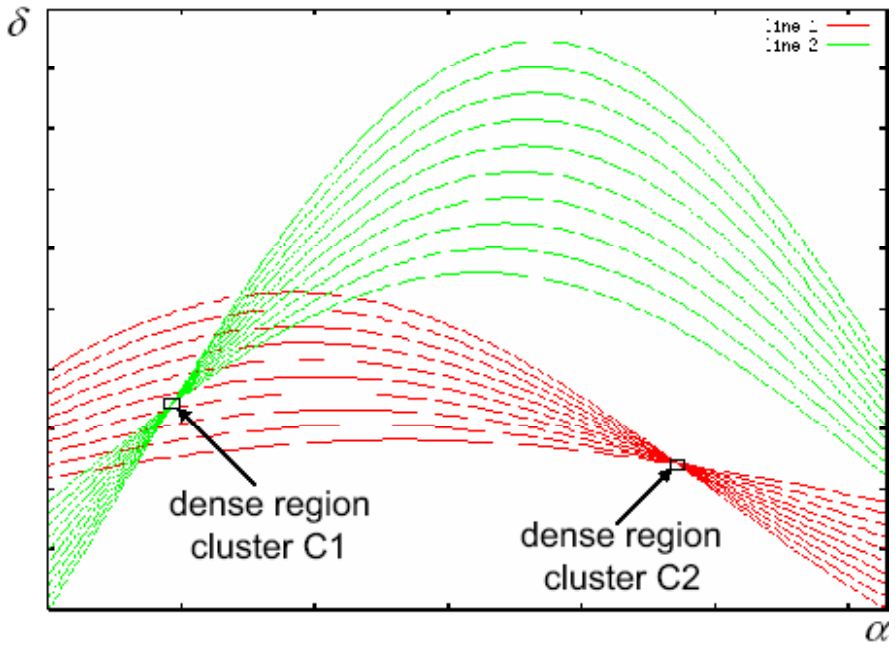
efficient search heuristic for dense regions in parameter space

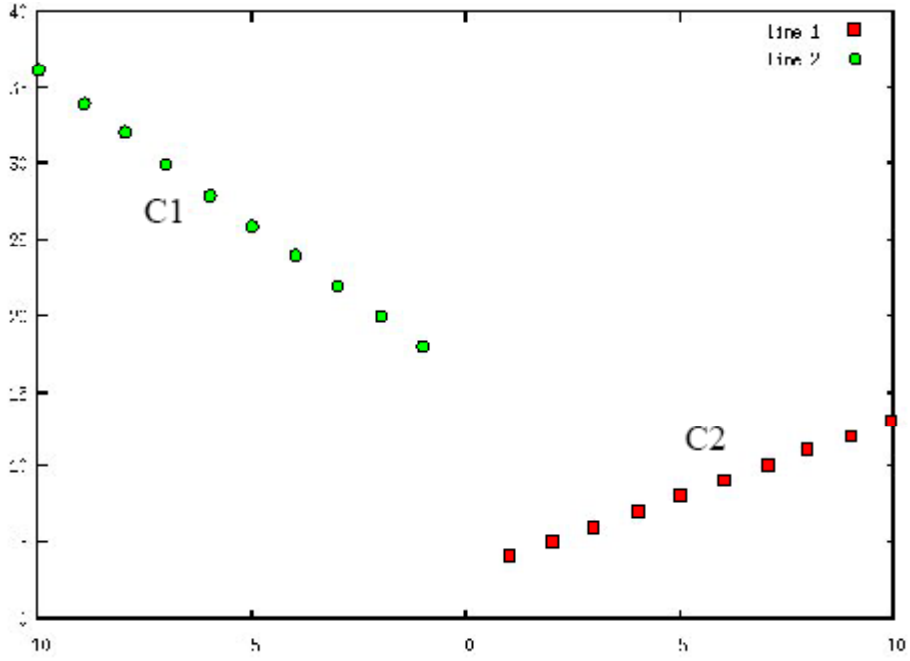
- construct a grid by recursively splitting the parameter space (best-first-search)
- identify dense grid cells as intersected by many parametrization functions
- dense grid cell represents (*d-1*)-dimensional linear structure
- transform corresponding data objects in corresponding (*d-1*)-dimensional space and repeat the search recursively



# Correlation Clustering Based on the Hough-Transform

- grid cell representing less than *m* points can be excluded
  $\rightarrow$ early pruning of a search path
- grid cell intersected by at least *m* sinusoids after *s* recursive splits represents a correlation cluster (with $\lambda \leq d\text{-}1$)
  - remove points of the cluster (and corr. sinusoids) from remaining cells



# Correlation Clustering Based on the Hough-Transform

properties:

- finds arbitrary number of clusters
- requires specification of depth of search (number of splits per axis)
- requires minimum density threshold for a grid cell
- Note: this minimum density does not relate to the locality assumption: CASH is a global approach to correlation clustering

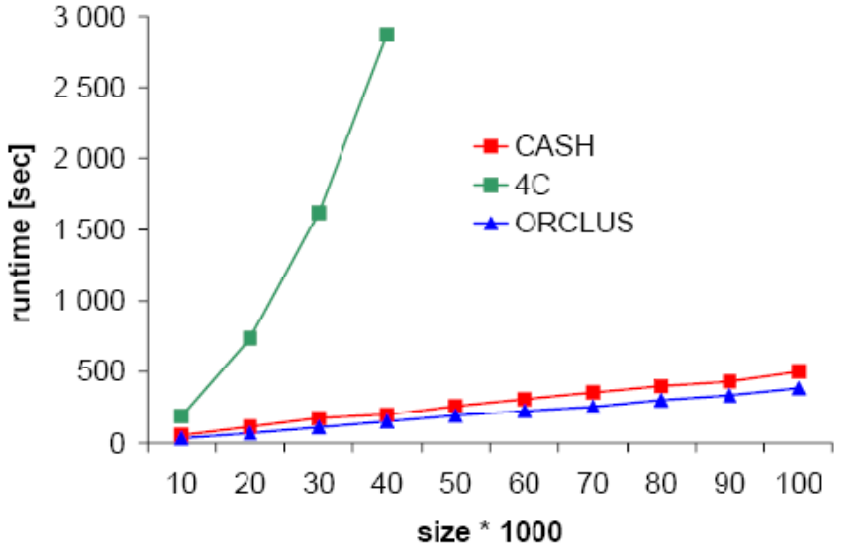
# Correlation Clustering Based on the Hough-Transform

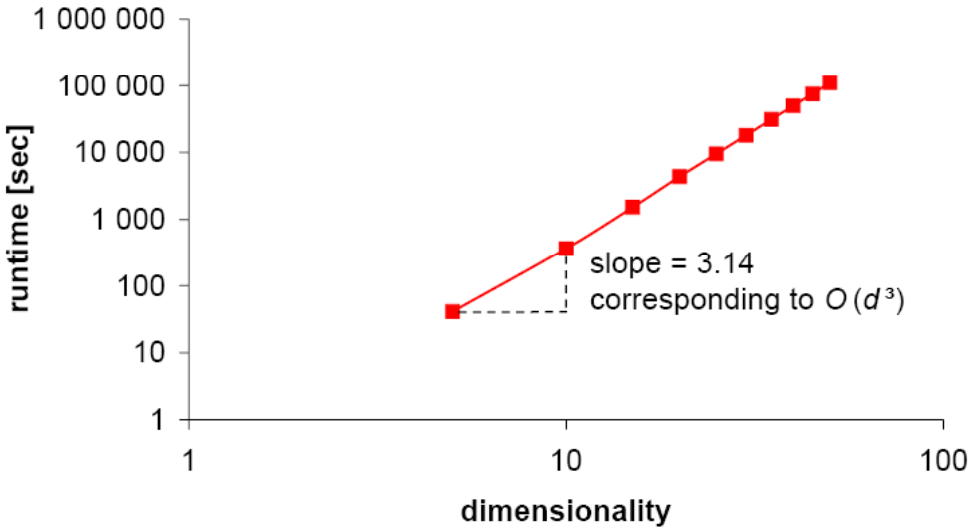
- search heuristic: linear in number of points, but ~ $O(d^3)$

  depth of search $s$, number $c$ of pursued paths (ideally: $c$ cluster):

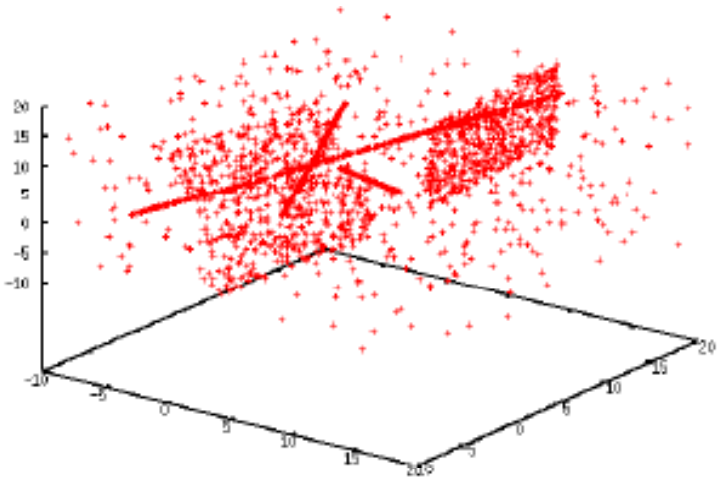
  - priority search: $O(s \cdot c)$
  - determination of curves intersecting a cell: $O(n \cdot d^3)$
  - overall: $O(s \cdot c \cdot n \cdot d^3)$

    (note: PCA generally in $O(d^3)$)

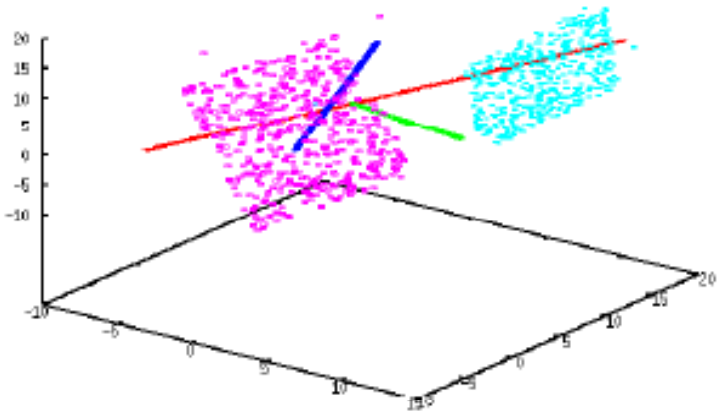


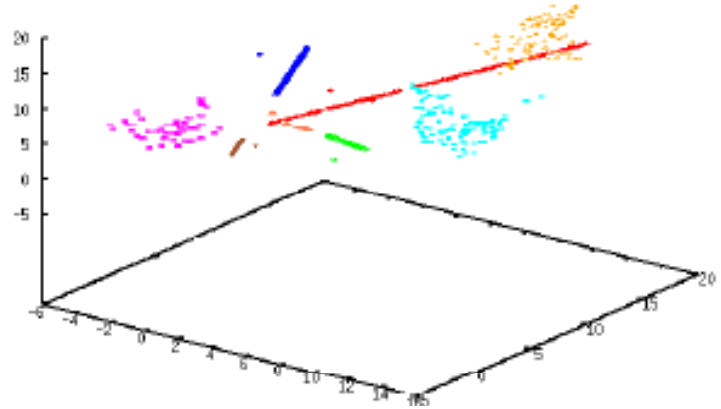
# Correlation Clustering Based on the Hough-Transform



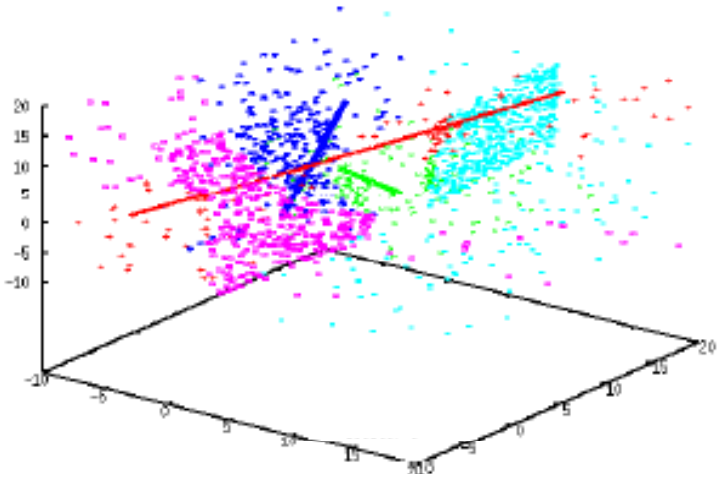
(a) Data set



(b) CASH: Cluster 1-5



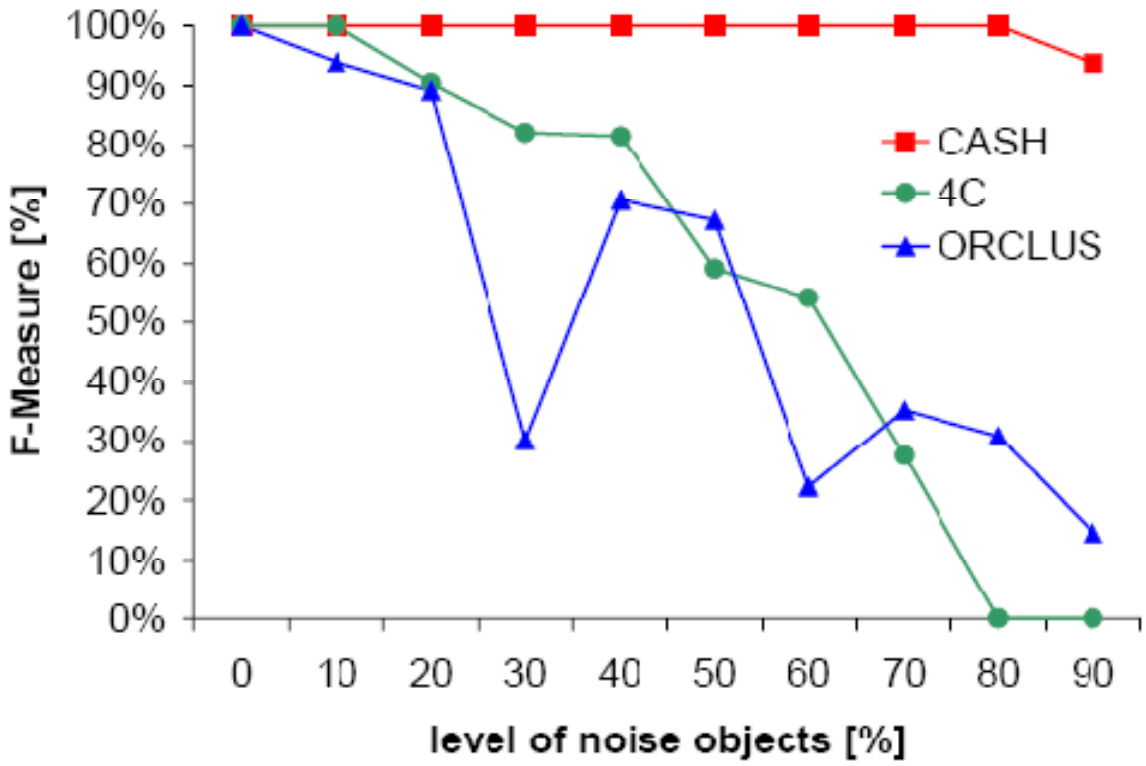
(c) 4C: Cluster 1-8



(d) ORCLUS: Cluster 1-5

# Correlation Clustering Based on the Hough-Transform

- stability with increasing number of noise objects



# Summary and Perspectives

- PCA: mature technique, allows construction of a broad range of similarity measures for local correlation of attributes
- drawback: all approaches suffer from locality assumption
- successfully employing PCA in correlation clustering in “really” high-dimensional data requires more effort henceforth
- new approach based on Hough-transform:
  - does not rely on locality assumption
  - but worst case again complete enumeration

# Summary and Perspectives

- some preliminary approaches base on concept of self-similarity (intrinsic dimensionality, fractal dimension): [BC00,PTTF02,GHPT05]
- interesting idea, provides quite a different basis to grasp correlations in addition to PCA
- drawback: self-similarity assumes locality of patterns even by definition

# Literature


[ABD+08] E. Achtert, C. Böhm, J. David, P. Kröger, and A. Zimek.
**Robust clustering in arbitrarily oriented subspaces.**
In Proceedings of the 8th SIAM International Conference on Data Mining (SDM), Atlanta, GA, 2008

[ABK+06] E. Achtert, C. Böhm, H.-P. Kriegel, P. Kröger, and A. Zimek.
**Deriving quantitative models for correlation clusters**.
In Proceedings of the 12th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Philadelphia, PA, 2006.

[ABK+07a] E. Achtert, C. Böhm, H.-P. Kriegel, P. Kröger, I. Müller-Gorman, and A. Zimek.
**Detection and visualization of subspace cluster hierarchies**.
In Proceedings of the 12th International Conference on Database Systems for Advanced Applications (DASFAA), Bangkok, Thailand, 2007.

[ABK+07b] E. Achtert, C. Böhm, H.-P. Kriegel, P. Kröger, and A. Zimek.
**On exploring complex relationships of correlation clusters**.
In Proceedings of the 19th International Conference on Scientific and Statistical Database Management (SSDBM), Banff, Canada, 2007.

[ABK+07c] E. Achtert, C. Böhm, H.-P. Kriegel, P. Kröger, and A. Zimek.
**Robust, complete, and efficient correlation clustering**.
In Proceedings of the 7th SIAM International Conference on Data Mining (SDM), Minneapolis, MN, 2007.

# Literature


[AGGR98] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan.
**Automatic subspace clustering of high dimensional data for data mining applications**.
In Proceedings of the ACM International Conference on Management of Data (SIGMOD), Seattle, WA, 1998.

[AHK01] C. C. Aggarwal, A. Hinneburg, and D. Keim.
**On the surprising behavior of distance metrics in high dimensional space.**
In Proceedings of the 8th International Conference on Database Theory (ICDT), London, U.K., 2001.

[APW+99] C. C. Aggarwal, C. M. Procopiuc, J. L. Wolf, P. S. Yu, and J. S. Park.
**Fast algorithms for projected clustering**.
In Proceedings of the ACM International Conference on Management of Data (SIGMOD), Philadelphia, PA, 1999.

[AS94] R. Agrawal and R. Srikant. **Fast algorithms for mining association rules**.
In Proceedings of the ACM International Conference on Management of Data (SIGMOD), Minneapolis, MN, 1994.

[AY00] C. C. Aggarwal and P. S. Yu.
**Finding generalized projected clusters in high dimensional space**.
In Proceedings of the ACM International Conference on Management of Data (SIGMOD), Dallas, TX, 2000.

# Literature


[BC00] D. Barbara and P. Chen.
**Using the fractal dimension to cluster datasets**.
In Proceedings of the 6th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Boston, MA, 2000.

[BDCKY02] A. Ben-Dor, B. Chor, R. Karp, and Z. Yakhini.
**Discovering local structure in gene expression data: The order-preserving submatrix problem**.
In Proceedings of the 6th Annual International Conference on Computational Molecular Biology (RECOMB), Washington, D.C., 2002.

[Bel61] R. Bellman.
**Adaptive Control Processes. A Guided Tour.**
Princeton University Press, 1961.

[BFG99] K. P. Bennett, U. Fayyad, and D. Geiger.
**Density-based indexing for approximate nearest-neighbor queries.**
In Proceedings of the 5th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), San Diego, CA, 1999.

[BGRS99] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft.
**When is “nearest neighbor” meaningful?**
In Proceedings of the 7th International Conference on Database Theory (ICDT), Jerusalem, Israel, 1999.

# Literature


[BKKK04] C. Böhm, K. Kailing, H.-P. Kriegel, and P. Kröger.
**Density connected clustering with local subspace preferences**.
In Proceedings of the 4th International Conference on Data Mining (ICDM), Brighton, U.K., 2004.

[BKKZ04] C. Böhm, K. Kailing, P. Kröger, and A. Zimek.
**Computing clusters of correlation connected objects**.
In Proceedings of the ACM International Conference on Management of Data (SIGMOD), Paris, France, 2004.

[CC00] Y. Cheng and G. M. Church.
**Biclustering of expression data**.
In Proceedings of the 8th International Conference Intelligent Systems for Molecular Biology (ISMB), San Diego, CA, 2000.

[CDGS04] H. Cho, I. S. Dhillon, Y. Guan, and S. Sra.
**Minimum sum-squared residue co-clustering of gene expression data**.
In Proceedings of the 4th SIAM International Conference on Data Mining (SDM), Orlando, FL, 2004.

[CFZ99] C. H. Cheng, A. W.-C. Fu, and Y. Zhang.
**Entropy-based subspace clustering for mining numerical data**.
In Proceedings of the 5th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), San Diego, CA, pages 84–93, 1999.

# Literature


[CST00] A. Califano, G. Stolovitzky, and Y. Tu.
**Analysis of gene expression microarrays for phenotype classification**.
In Proceedings of the 8th International Conference Intelligent Systems for Molecular Biology (ISMB), San Diego, CA, 2000.

[EKSX96] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu.
**A density-based algorithm for discovering clusters in large spatial databases with noise**.
In Proceedings of the 2nd ACM International Conference on Knowledge Discovery and Data Mining (KDD), Portland, OR, 1996.

[FM04] J. H. Friedman and J. J. Meulman.
**Clustering objects on subsets of attributes**.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 66(4):825–849, 2004.

[FWV07] D. Francois, V. Wertz, and M. Verleysen.
**The concentration of fractional distances.**
IEEE Transactions on Knowledge and Data Engineering, 19(7): 873-886, 2007.

[GHPT05] A. Gionis, A. Hinneburg, S. Papadimitriou, and P. Tsaparas.
**Dimension induced clustering**.
In Proceedings of the 11th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Chicago, IL, 2005.

# Literature


[GLD00] G. Getz, E. Levine, and E. Domany.
**Coupled two-way clustering analysis of gene microarray data**.
Proceedings of the National Academy of Sciences of the United States of America, 97(22):12079–12084, 2000.

[GRRK05] E. Georgii, L. Richter, U. Rückert, and S. Kramer.
**Analyzing microarray data using quantitative association rules**.
Bioinformatics, 21(Suppl. 2):ii1–ii8, 2005.

[GW99] B. Ganter and R. Wille.
**Formal Concept Analysis**.
Mathematical Foundations. Springer, 1999.

[HAK00] A. Hinneburg, C. C. Aggarwal, and D. A. Keim.
**What is the nearest neighbor in high dimensional spaces?**
In Proceedings of the 26th International Conference on Very Large Data Bases (VLDB), Cairo, Egypt, 2000.

[Har72] J. A. Hartigan.
**Direct clustering of a data matrix**.
Journal of the American Statistical Association, 67(337):123–129, 1972.

[HKK+10] M. Houle, H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek.
**Can Shared-Neighbor Distances Defeat the Curse of Dimensionality?**
In Proceedings of the 22nd International Conference on Scientific and Statistical Data Management (SSDBM), Heidelberg, Germany, 2010.

# Literature


[IBB04] J. Ihmels, S. Bergmann, and N. Barkai.
**Defining transcription modules using large-scale gene expression data**.
Bioinformatics, 20(13):1993–2003, 2004.

[Jol02] I. T. Jolliffe.
**Principal Component Analysis**.
Springer, 2nd edition, 2002.

[KKK04] K. Kailing, H.-P. Kriegel, and P. Kröger.
**Density-connected subspace clustering for highdimensional data**.
In Proceedings of the 4th SIAM International Conference on Data Mining (SDM), Orlando, FL, 2004.

[KKRW05] H.-P. Kriegel, P. Kröger, M. Renz, and S. Wurst.
**A generic framework for efficient subspace clustering of high-dimensional data**.
In Proceedings of the 5th International Conference on Data Mining (ICDM), Houston, TX, 2005.

[KKZ09] H.-P. Kriegel, P. Kröger, and A. Zimek.
**Clustering High Dimensional Data: A Survey on Subspace Clustering, Pattern-based Clustering, and Correlation Clustering.**
ACM Transactions on Knowledge Discovery from Data (TKDD), Volume 3, Issue 1 (March 2009), Article No. 1, pp. 1-58, 2009.

# Literature


[LW03] J. Liu and W. Wang.
**OP-Cluster: Clustering by tendency in high dimensional spaces**.
In Proceedings of the 3th International Conference on Data Mining (ICDM), Melbourne, FL, 2003.

[MK03] T. M. Murali and S. Kasif.
**Extracting conserved gene expression motifs from gene expression data**.
In Proceedings of the 8th Pacific Symposium on Biocomputing (PSB), Maui, HI, 2003.

[MO04] S. C. Madeira and A. L. Oliveira.
**Biclustering algorithms for biological data analysis: A survey**.
IEEE Transactions on Computational Biology and Bioinformatics, 1(1):24–45, 2004.

[MSE06] G. Moise, J. Sander, and M. Ester.
**P3C: A robust projected clustering algorithm**.
In Proceedings of the 6th International Conference on Data Mining (ICDM), Hong Kong, China, 2006.

[NGC01] H.S. Nagesh, S. Goil, and A. Choudhary.
**Adaptive grids for clustering massive data sets**.
In Proceedings of the 1st SIAM International Conference on Data Mining (SDM), Chicago, IL, 2001.

# Literature


[PBZ+06] A. Prelic, S. Bleuler, P. Zimmermann, A. Wille, P. Bühlmann, W. Guissem, L. Hennig, L. Thiele, and E. Zitzler.
**A systematic comparison and evaluation of biclustering methods for gene expression data**.
Bioinformatics, 22(9):1122–1129, 2006.

[Pfa07] J. Pfaltz.
**What constitutes a scientific database?**
In Proceedings of the 19th International Conference on Scientific and Statistical Database Management (SSDBM), Banff, Canada, 2007.

[PHL04] L. Parsons, E. Haque, and H. Liu.
**Subspace clustering for high dimensional data: A review**.
SIGKDD Explorations, 6(1):90–105, 2004.

[PJAM02] C. M. Procopiuc, M. Jones, P. K. Agarwal, and T. M. Murali.
**A Monte Carlo algorithm for fast projective clustering**.
In Proceedings of the ACM International Conference on Management of Data (SIGMOD), Madison, WI, 2002.

[PTTF02] E. Parros Machado de Sousa, C. Traina, A. Traina, and C. Faloutsos.
**How to use fractal dimension to find correlations between attributes**.
In Proc. KDD-Workshop on Fractals and Self-similarity in Data Mining: Issues and Approaches, 2002.

# Literature


[PZC+03] J. Pei, X. Zhang, M. Cho, H. Wang, and P. S. Yu.
**MaPle: A fast algorithm for maximal pattern-based clustering**.
In Proceedings of the 3th International Conference on Data Mining (ICDM), Melbourne, FL, 2003.

[RRK04] U. Rückert, L. Richter, and S. Kramer.
**Quantitative association rules based on half-spaces: an optimization approach**.
In Proceedings of the 4th International Conference on Data Mining (ICDM), Brighton, U.K., 2004.

[SCH75] J.L. Slagle, C.L. Chang, S.L. Heller.
**A Clustering and Data-Reorganization Algorithm**.
IEEE Transactions on Systems, Man and Cybernetics, 5: 121-128, 1975

[SLGL06] K. Sim, J. Li, V. Gopalkrishnan, and G. Liu.
**Mining maximal quasi-bicliques to co-cluster stocks and financial ratios for value investment.**
In Proceedings of the 6th International Conference on Data Mining (ICDM), Hong Kong, China, 2006.

[SMD03] Q. Sheng, Y. Moreau, and B. De Moor.
**Biclustering microarray data by Gibbs sampling**.
Bioinformatics, 19(Suppl. 2):ii196–ii205, 2003.

# Literature


[STG+01] E. Segal, B. Taskar, A. Gasch, N. Friedman, and D. Koller.
**Rich probabilistic models for gene expression**.
Bioinformatics, 17(Suppl. 1):S243–S252, 2001.

[SZ05] K. Sequeira and M. J. Zaki.
**SCHISM: a new approach to interesting subspace mining**.
International Journal of Business Intelligence and Data Mining, 1(2):137–160, 2005.

[TSS02] A. Tanay, R. Sharan, and R. Shamir.
**Discovering statistically significant biclusters in gene expression data**.
Bioinformatics, 18 (Suppl. 1):S136–S144, 2002.

[TXO05] A. K. H. Tung, X. Xu, and C. B. Ooi.
**CURLER: Finding and visualizing nonlinear correlated clusters**.
In Proceedings of the ACM International Conference on Management of Data (SIGMOD), Baltimore, ML, 2005.

[Web01] G. I. Webb.
**Discovering associations with numeric variables**.
In Proceedings of the 7th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), San Francisco, CA, pages 383–388, 2001.

[WLKL04] K.-G. Woo, J.-H. Lee, M.-H. Kim, and Y.-J. Lee.
**FINDIT: a fast and intelligent subspace clustering algorithm using dimension voting**.
Information and Software Technology, 46(4):255–271, 2004.

# Literature

[WWYY02] H. Wang, W. Wang, J. Yang, and P. S. Yu.
**Clustering by pattern similarity in large data sets**.
In Proceedings of the ACM International Conference on Management of Data (SIGMOD), Madison, WI, 2002.

[YWWY02] J. Yang, W. Wang, H. Wang, and P. S. Yu.
**δ-clusters: Capturing subspace correlation in a large data set**.
In Proceedings of the 18th International Conference on Data Engineering (ICDE), San Jose, CA, 2002.