HelmholtzZentrum münchen
Deutsches Forschungszentrum für Gesundheit und Umwelt

Technische Universität München

Current Topics in Information-theoretic Data Mining

CLAUDIA PLANT,

NINA HUBIG, SAM MAURUS, ANNIKA TONCH

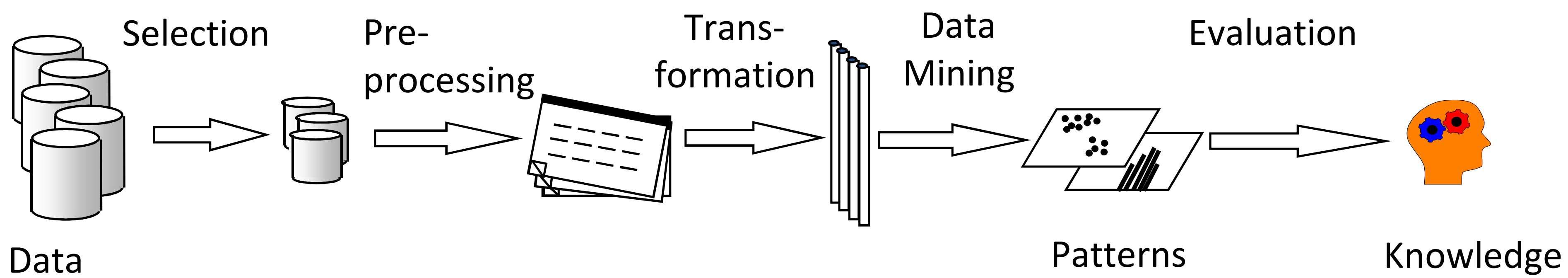
Outline

1. Introduction
2. General Information
3. Short Presentation of Topics
4. Selection of Topics

Information-theoretic Data Mining

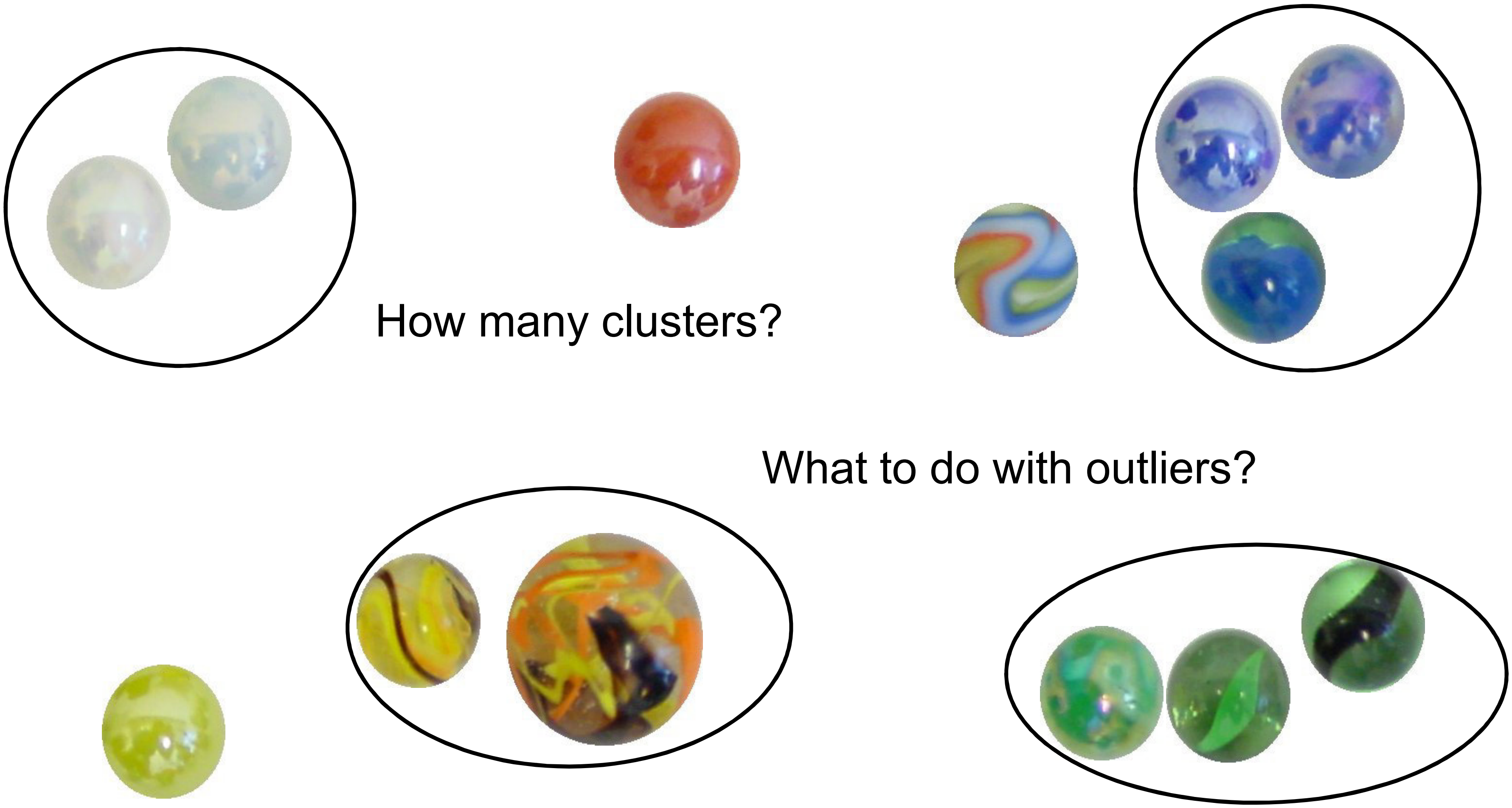
INTRODUCTION

What is Data Mining?

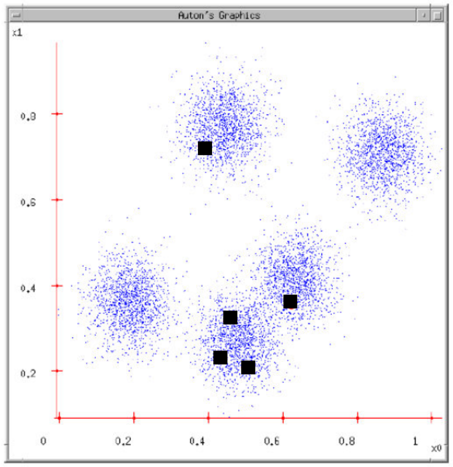


Knowledge Discovery in Databases *is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in the data.* [Fayyad et al. 2nd KDD Conference 1996]

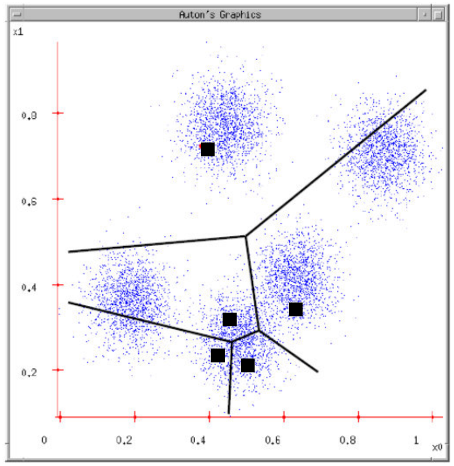
Example Clustering:
Find a natural grouping of the data objects.



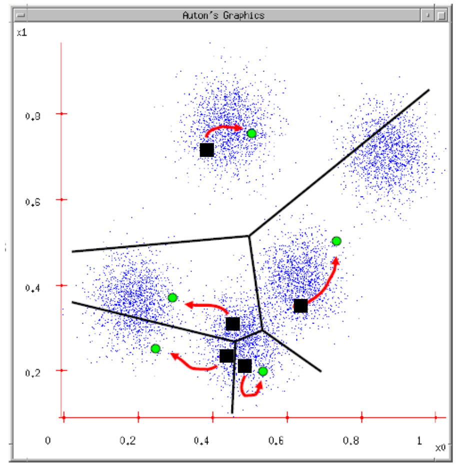
The Algorithm K-Means



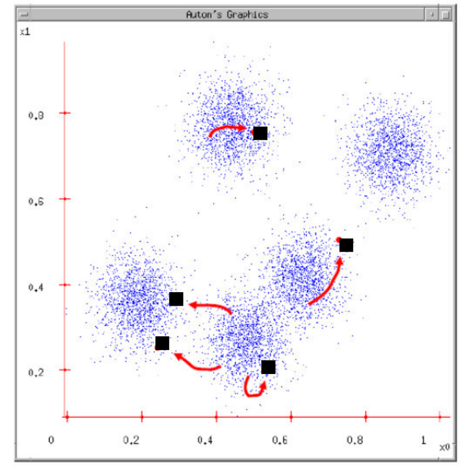
1) **Initialize**
K cluster centers randomly.



2) **Assign** points to the closest center.



3) **Update** centers.

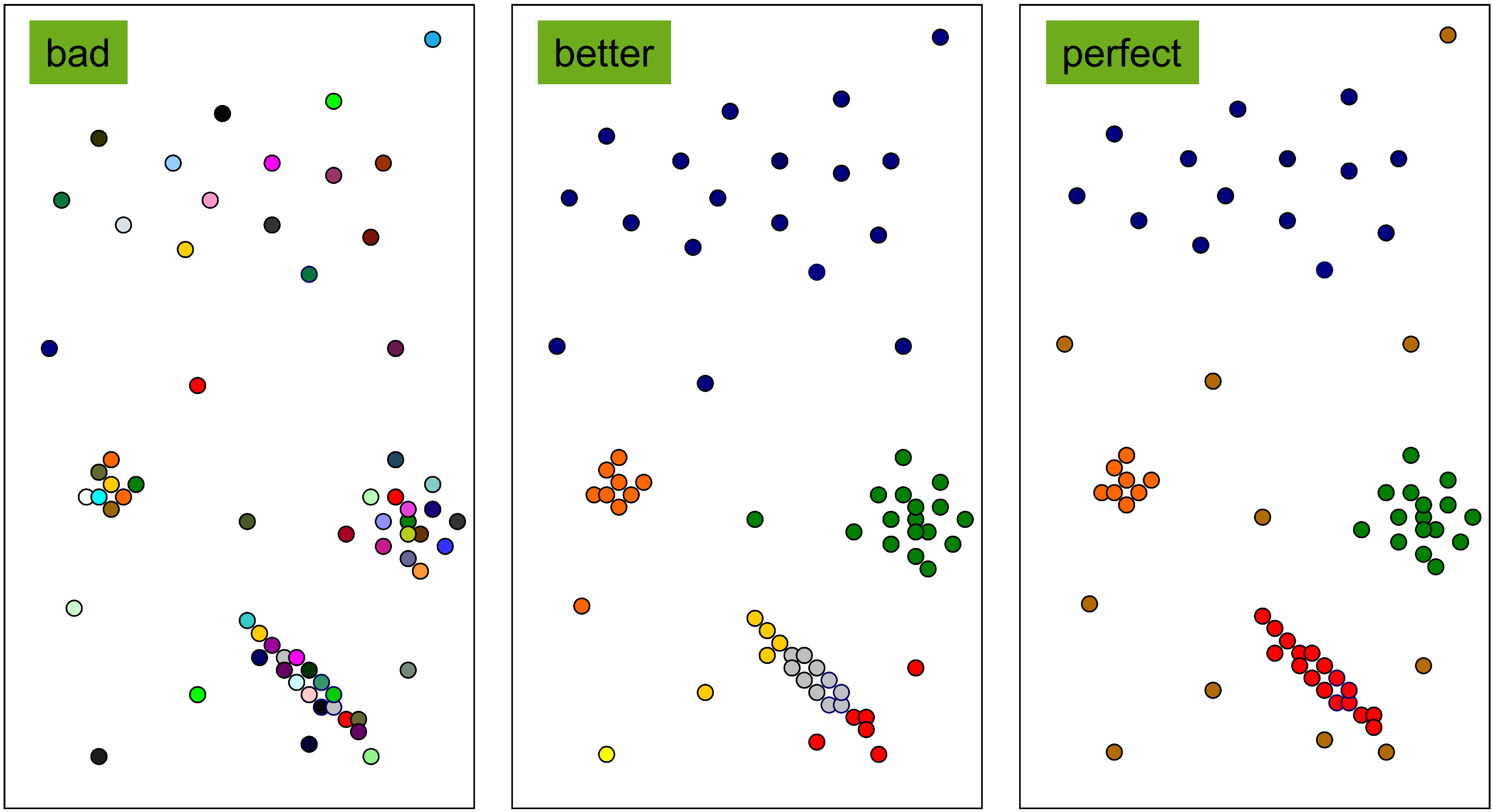


4) **Iterate**
2) und 3) until convergence.

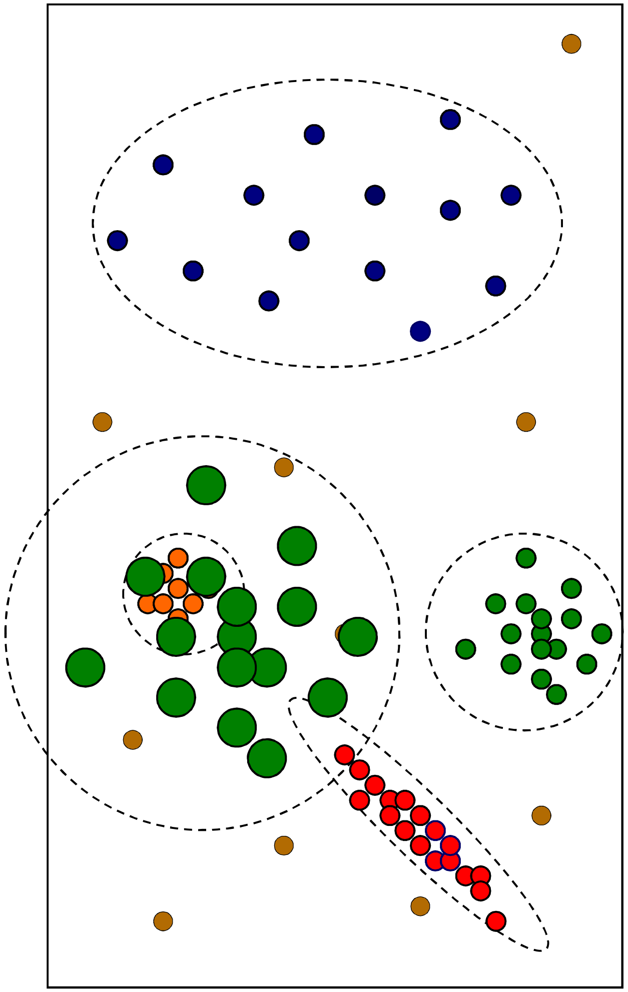
+ fast convergence,
+ well-defined objective function,
+ gives a model describing the result.

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

We need a quality criterion for clustering



Measuring Clustering Quality by Data Compression

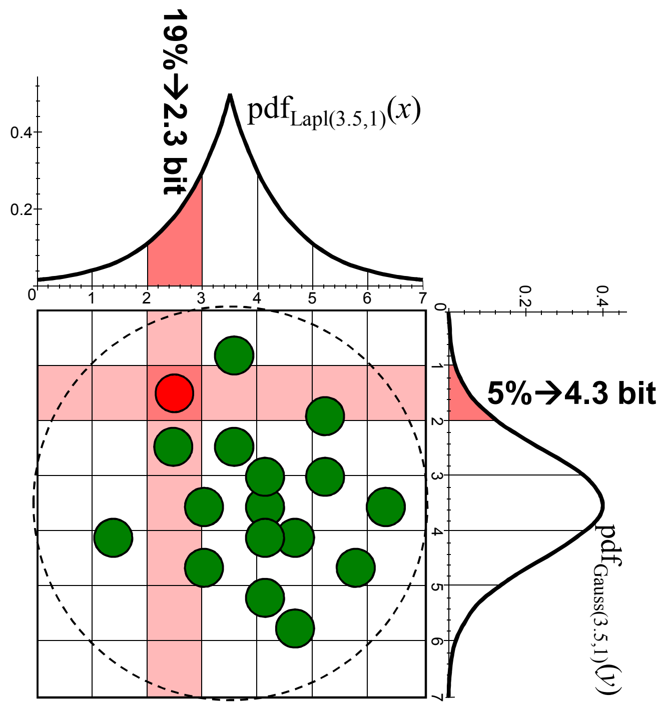


Data compression is a good criterion for…

- the required number of clusters
- the goodness of a cluster structure
- the quality of a cluster description

How can a cluster be compressed?

Measuring Clustering Quality by Data Compression



Data compression is a good criterion for…

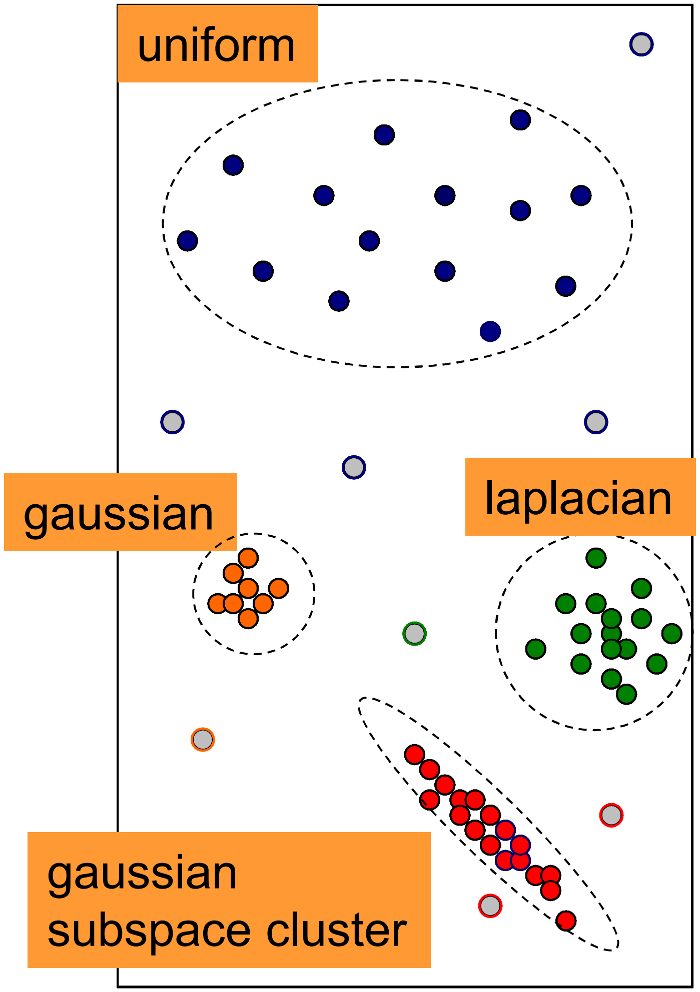
- the required number of clusters
- the goodness of a cluster structure
- the quality of a cluster description by a pdf

How can a cluster be compressed?

- Huffman-coded coordinates of points
- Huffman-coded cluster-id for each point
- if necessary: decorrelation matrix
- type and parameters of the pdf
  (e.g. Gaussian, $\mu=3.5$, $\sigma=1.0$)

Minimum Description Length (MDL) Principle:
Automatic balance of
Goodness-of-fit and model complexity

Algorithm RIC:
Robust Information-theoretic Clustering (KDD 2006)



Start with an arbitrary partitioning

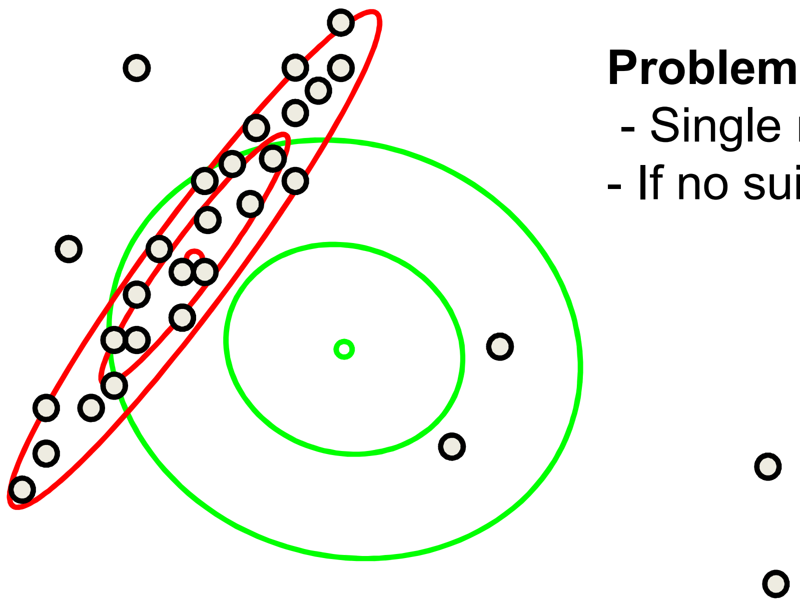
1. Robust Fitting (RF):
Purifies individual clusters from noise, determines a stable model.

2. Cluster Merging (CM):
Stiches clusters which match well together.

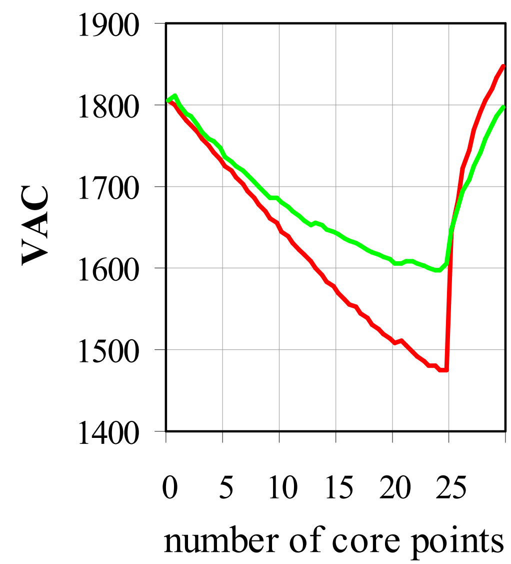
Additional value-add:
Description of the cluster content by assigning model distribution functions to the individual coordinates.

Free from sensitve parameter settings !

Robust Fitting

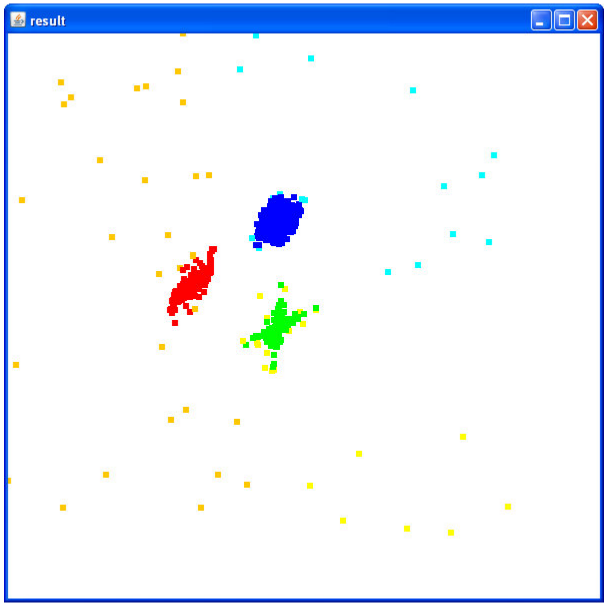


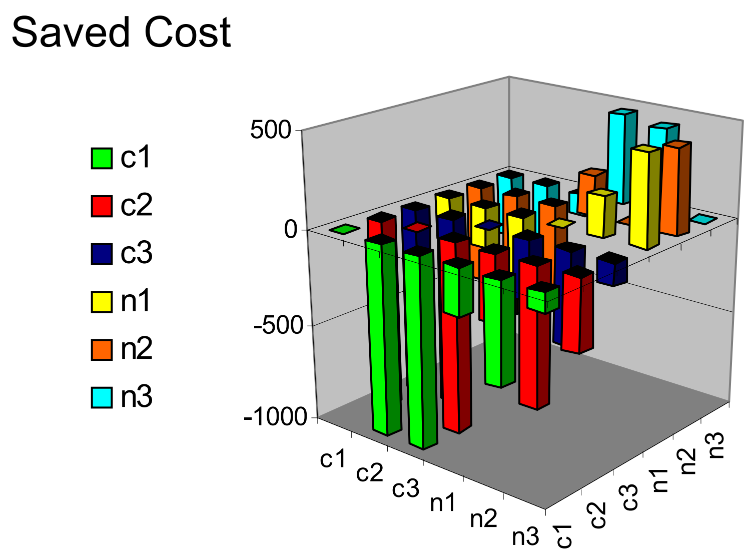
Problem:
- Single noise points may spoil model functions
- If no suitable model is availlable noise cannot be filtered



Solution:
- Outlier-robust PCA based on the median
- Select the best partitioning into cluster and Noise points with VAC.

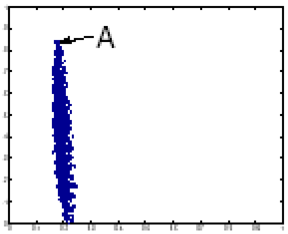
Cluster Merging



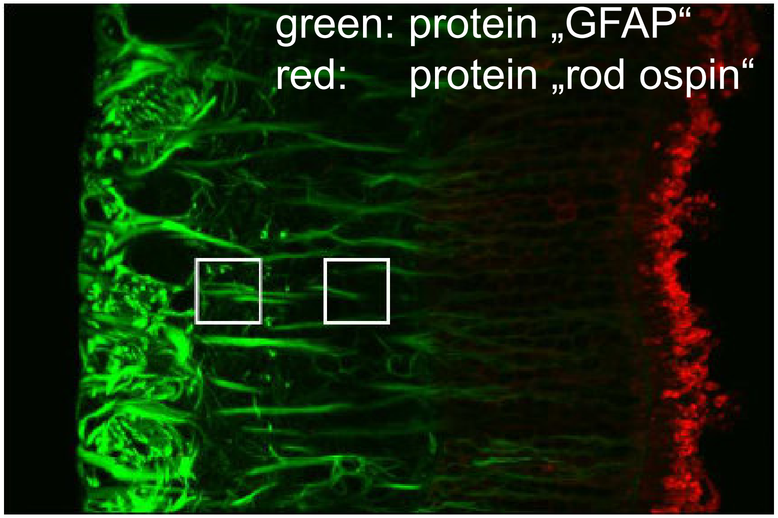


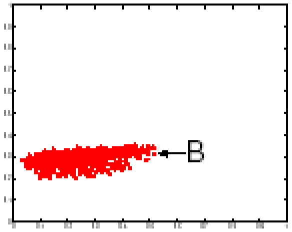
Cluster Merging:
Cost table with |C|/2 Einträge.
Greedy algorithm: Merge in each step pair of clusters with maximum saved costs.
Example: Merge first n_2 with n_3 and then *with* n_1.

Results on Cat Retina Images: Biological Interpretation of Selected Clusters

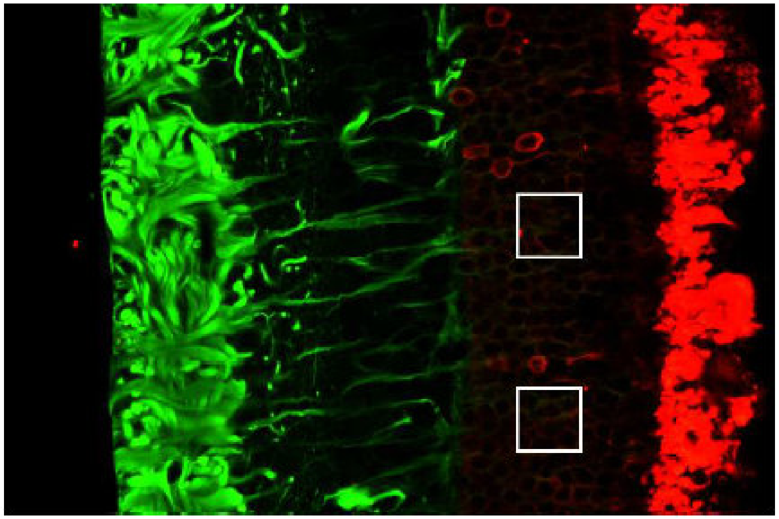


- Layer-detached retina treated with oxygen exposition.
- Tiles represent „Müller Cells“ with protein GFAP propagated to the inner layer of the retina.

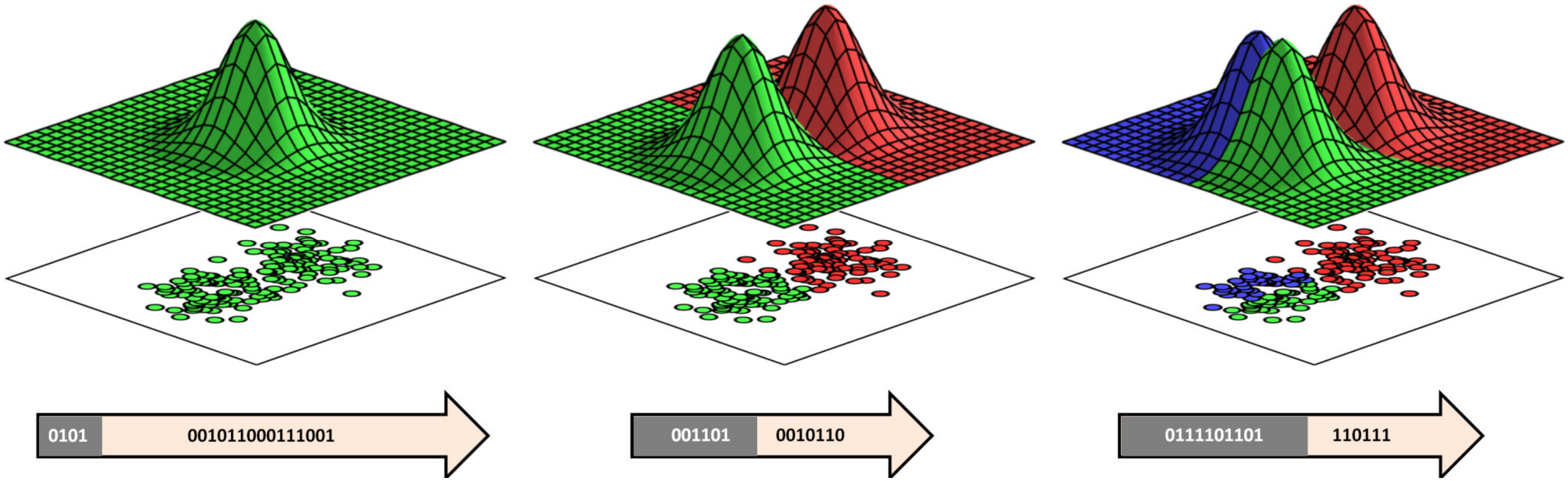




•3 month of layer detachment.
•Tiles are „rod photoreceptors“ with the protein rod ospin redistributed into the cell bodies, characteristic for detached retinas.



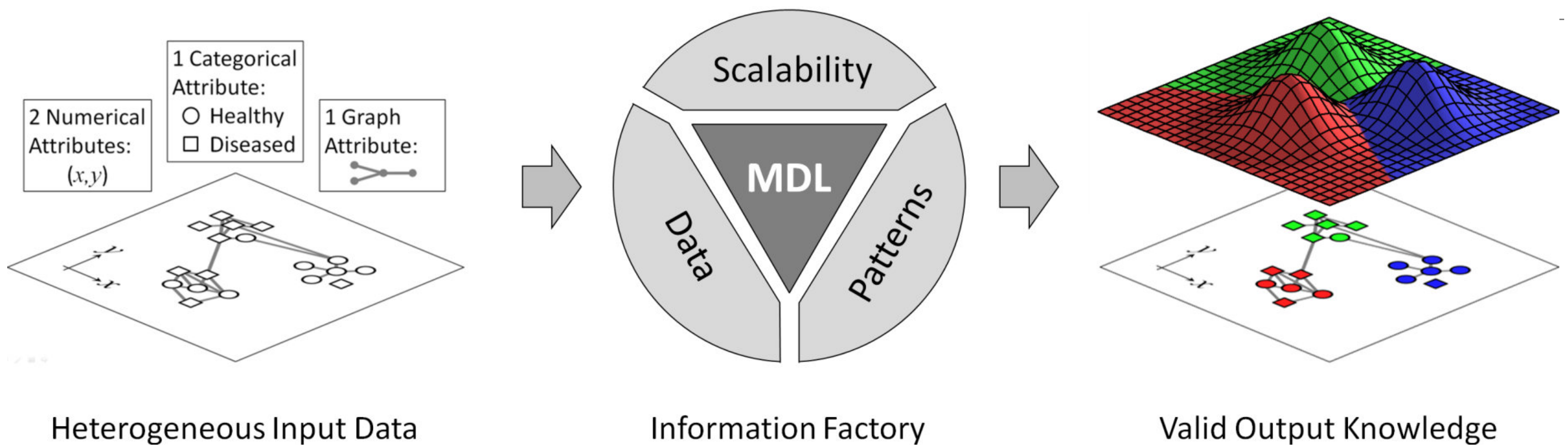
Key Idea

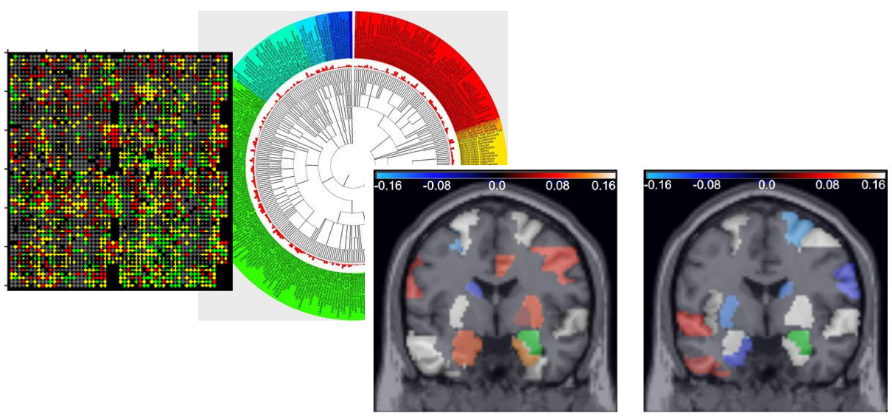


Data compression is a very general measure for:

- The amount of any kind of non-random information in any kind of data,
- The success of any kind of data mining technique.

Helmholtz-Hochschul research group iKDD





Applications:
Neuroscience,
Diabetes research.

General Information

ABOUT THE SEMINAR

Goals of the Seminar

Learn how to:

- Read scientific papers
- Discover the state-of-the-art on a specific topic
- Write a scientific report
- Do a scientific presentation

The Seminar in Practice

- **SWS**: 2+0, **ECTS**: 4 Credits
- **Presentation**: 20 min presentation/10 min questions. Download the template from the seminar web page
  - Slides must be in **english**, presentation can be hold in german
- Write a **report** (max 8 pages). Size can vary between bachelor and master students.
  - Report can either be written in german or english
- **Attendance** and **participation** of the seminar meetings
  - Participation: read the abstract, see figures, read introduction and conclusion
  - Prepare questions
- **Grade: 60% presentation, 40% report.**
- **Seminar days: May 15 -16, time to be announced at the website.**

Contents of the Report

Follow the structure of a scientific publication.

- **Abstract and Introduction (~1 page)**
  - General motivation.
- **State of the Art and Contributions (~2 pages)**
  - How is this paper different from (SoA)? e.g What is new? What is better? What is faster?
- **Problem statement (~1 page)**
  - Mathematical formulation
- **Method (~2 pages)**
  - Overview: input, output.
  - Method/Algorithm.
- **Results (~1 page)**
  - Summary of experiments and results (what type of data and validation).
  - YOUR CRITIQUE of the methodology, set-up and validation (what else could have been done?, is it enough to demonstrate the contribution?, is the data biased?, are there non mentioned assumptions?, can it be easily reproduced?)
- **Conclusion (~1 page)**
  - **YOUR PERSONAL CONCLUSION & IDEAS**
- **References (~1 page)**

Contents of the Presentation

As a rule of thumb: max 1 slide per minute (max 20 slides for 20 mins)

- **Present the paper**
  - Type and year of publication: journal, conference, workshop, etc.
  - Authors/Institution
- **Motivation and Goal**
  - What is the problem that the authors try to solve?
  - Name potential applications: what for?
  - General motivation: why is it interesting?
- **Related Work (state of the art)**
  - Mention most similar approaches and explain how your paper is different from them?
  - Citing/Referencing other people's work [Lastname-Conference-Year].
- **Method**
  - Overview (1 or 2 slides): input, output, contribution (the proposed new elements).
  - Method/Algorithm (Only key ideas).
- **Results (short version)**
  - Explain the type of **data** used.
  - Validation: what is being validated and how.
- **Conclusion** (include your own conclusions!!)

Selection of Topics

NINA, SAM, ANNIKA

HelmholtzZentrum münchen
Deutsches Forschungszentrum für Gesundheit und Umwelt

Technische Universität München

Mining Numerical and Mixed Data

BASIC CLUSTERING

FINDING ALTERNATIVE CLUSTERINGS

MIXED (NUMERICAL, CATEGORICAL DATA)

nina.hubig@helmholtz-muenchen.de

Vector: Basic Clustering

Finding alternative clusterings

Mixed (numerical, categorical data)

Binary: Dimensionality reduction

Matrix factorization

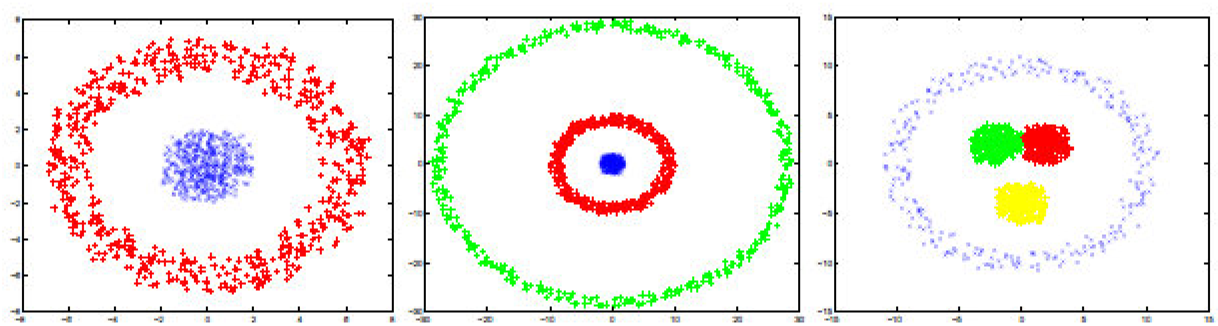
pattern extraction

Graphs: Clustering

Weighted graphs

Summarization, Structure mining

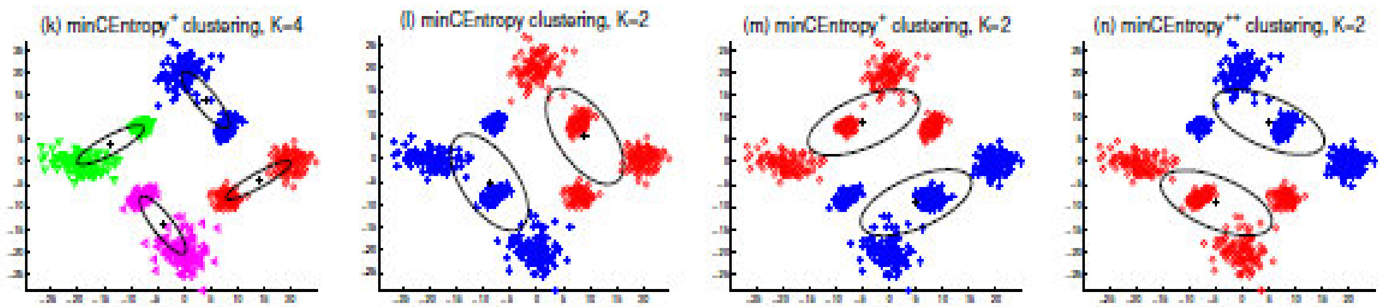
A Nonparametric Information- Theoretic Clustering Algorithm



- first google pick for information theoretical clustering ;)
- close to **machine learning**
- uses entropy and **mutual information** as quality function

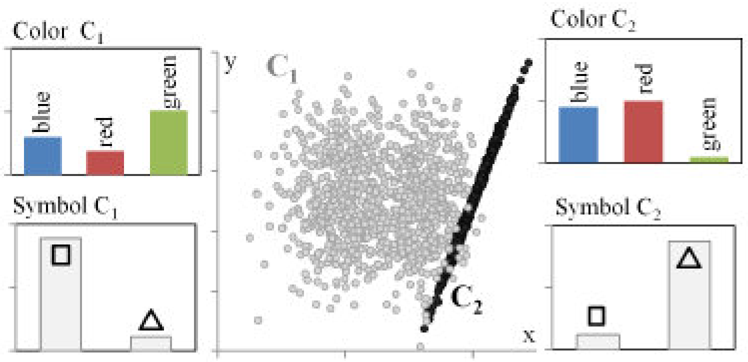
  ➔ a bit different than our MDL-based approaches!

minCEntropy: a Novel Information Theoretic Approach for the Generation of Alternative Clusterings



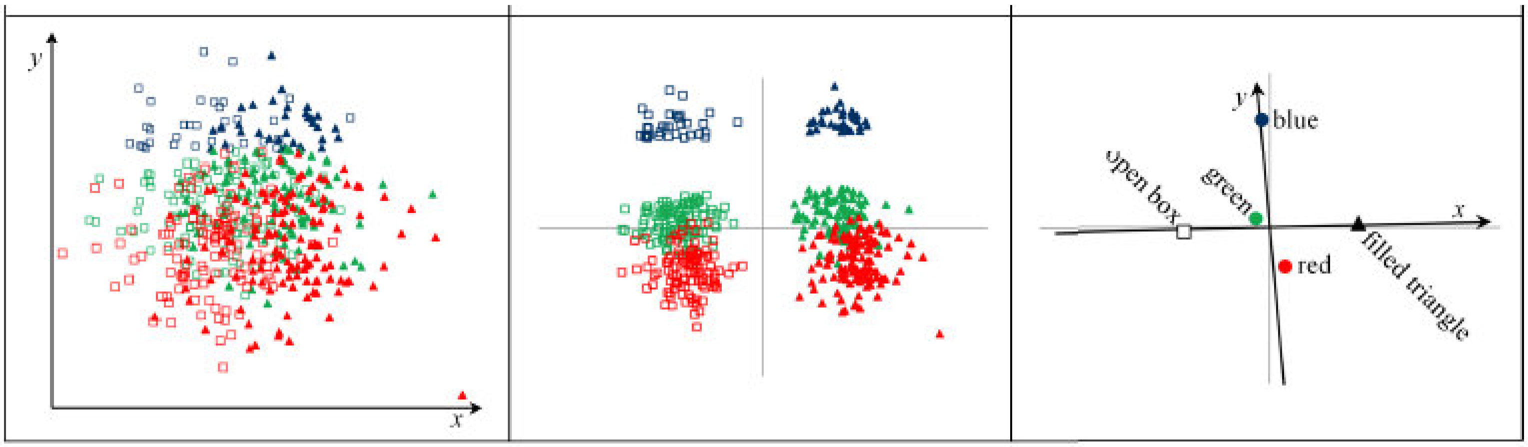
- Aims at finding different **alternative clusterings** for the same data set
- Uses a **general entropy** as objective function (not Shannon)
- can also be used semi-supervised (close to machine learning topics)

INCONCO: Interpretable Clustering of Numerical and Categorical Objects



- Uses Minimum Description Length (MDL) ;)
- Tackles mixed-type attributes: numerical, categorical data
- Clusters by revealing „dependency patterns“ among attributes by using and extended Cholesky decomposition

Dependency Clustering across measurement scales



- Uses MDL ;)
- supports mixed-type attributes
- finds **attribute dependencies regardless the measurement scale**

HelmholtzZentrum münchen
Deutsches Forschungszentrum für Gesundheit und Umwelt

Technische Universität München

Mining Binary Data

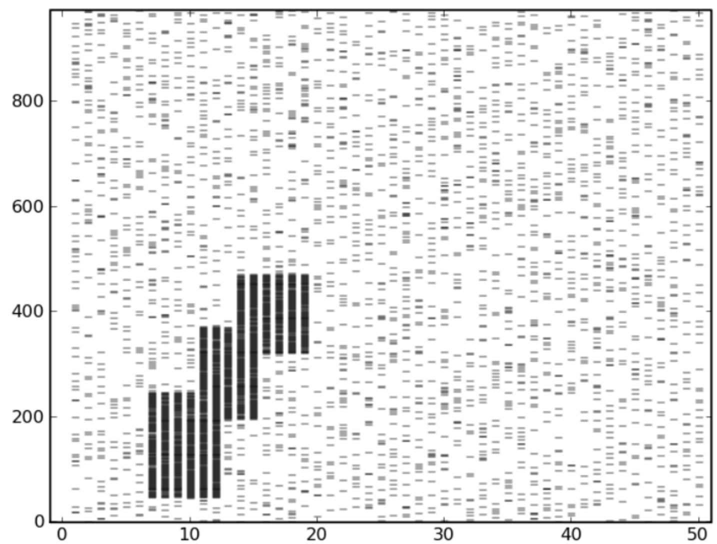
DIMENSIONALITY REDUCTION

MATRIX FACTORIZATION

PATTERN EXTRACTION

samuel.maurus@helmholtz-muenchen.de

Mining Top-K Patterns from Binary Datasets in presence of Noise



- Rows can be thought of as transactions. Items (columns) are either present (1) or absent (0) in a transaction.
- Can we approximately summarize the data set using just K *base* transactions?
- What if the patterns are overlapping? How can we be robust against noise yet not fall victim to overfitting?

Lucchese et al.

Model-order Selection for Boolean Matrix Factorization

Binary basis vectors ("questionnaire themes")

$$\begin{array}{c} \begin{array}{ccc} \text{Q1} & \text{Q2} & \text{Q3} \end{array} \\ \begin{bmatrix} Y & Y & N \\ Y & Y & Y \\ N & Y & Y \end{bmatrix} \end{array} = \begin{bmatrix} Y & N \\ Y & Y \\ N & Y \end{bmatrix} \circ \begin{bmatrix} Y & Y & N \\ N & Y & Y \end{bmatrix}$$

Original data matrix

Binary usage matrix

Aim: Intuitive dimensionality reduction

Here K=2, but why did we choose that? Can we somehow automate the selection of K?

Miettinen et al.

Directly Mining Descriptive Patterns

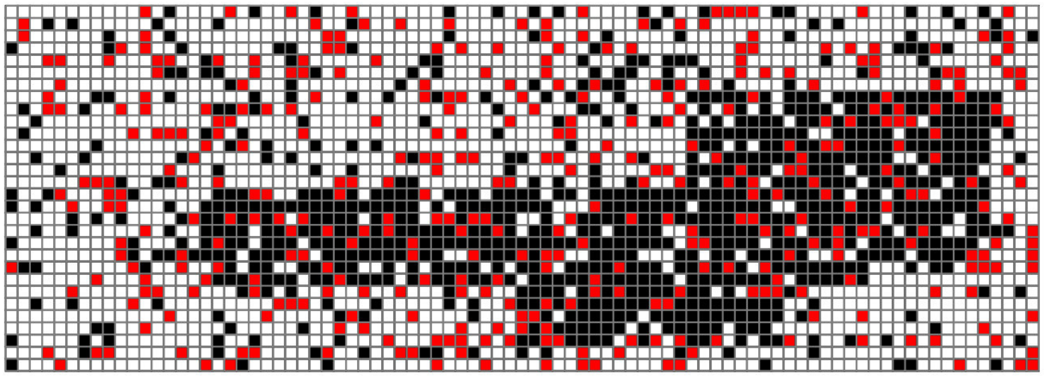
Transactions of items at the checkout

Millions of transactions in a central database

Can we generate a small set of quality patterns (groups of items or "frequent itemsets") that together describe the database? What do we mean by "quality"?

Smets et al.

Filling in the Blanks – Missing Values in Binary Data Sets



Binary data set (□,■) with missing values(■).

How can we intelligently predict the values that are missing?

Vreeken et al.

HelmholtzZentrum münchen
Deutsches Forschungszentrum für Gesundheit und Umwelt

Technische Universität München

Graph Mining

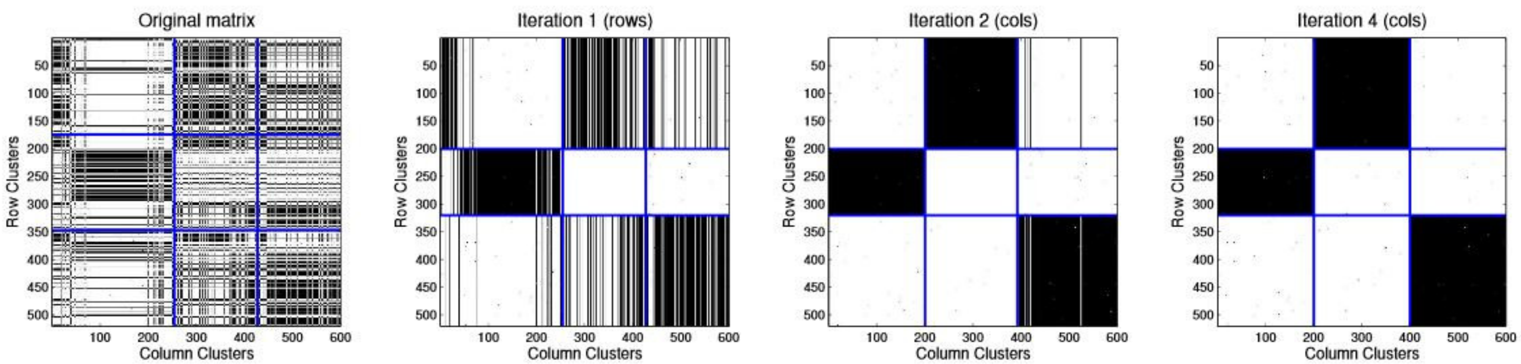
CLUSTERING

WEIGHTED GRAPHS

SUMMARIZATION, STRUCTURE MINING

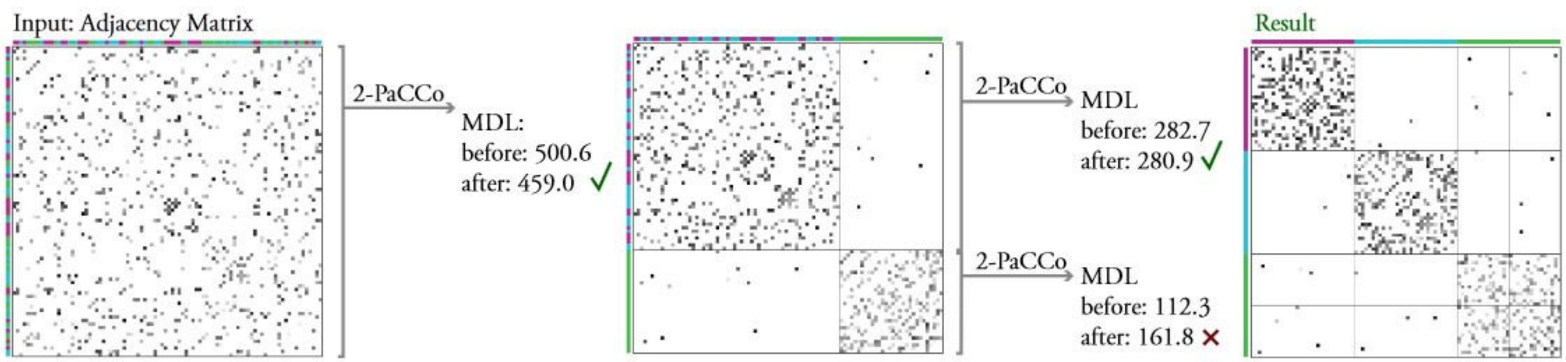
annika.tonch@helmholtz-muenchen.de

Fully Automatic Cross-Associations



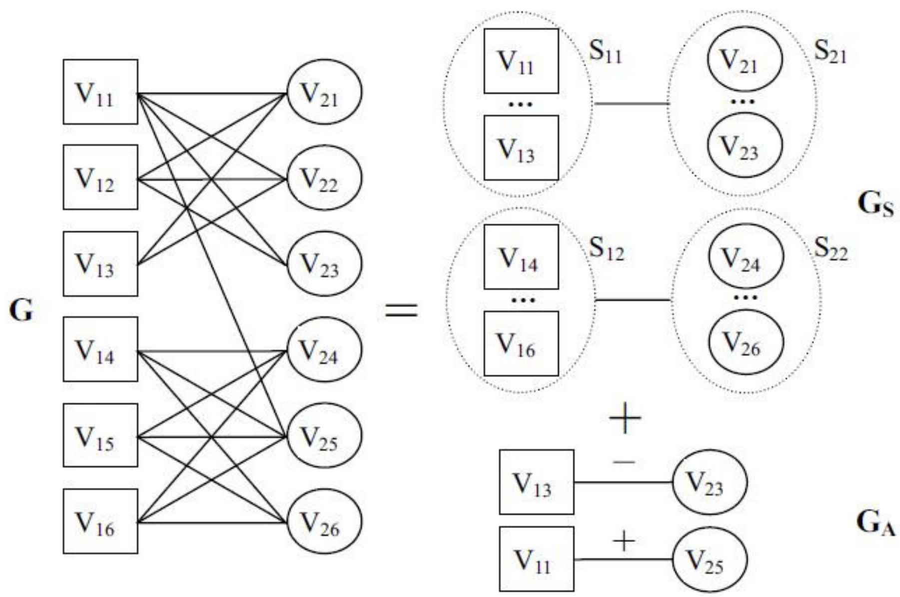
- Finding structures in datasets (parameter-free, fully automatic, scalable to very large matrices)
- Input data: binary matrix (for example gained by graph data)
- Rearrangement of rows and columns according to the smallest coding costs suggested by MDL

Weighted Graph Compression for Parameter-free Clustering With PaCCo



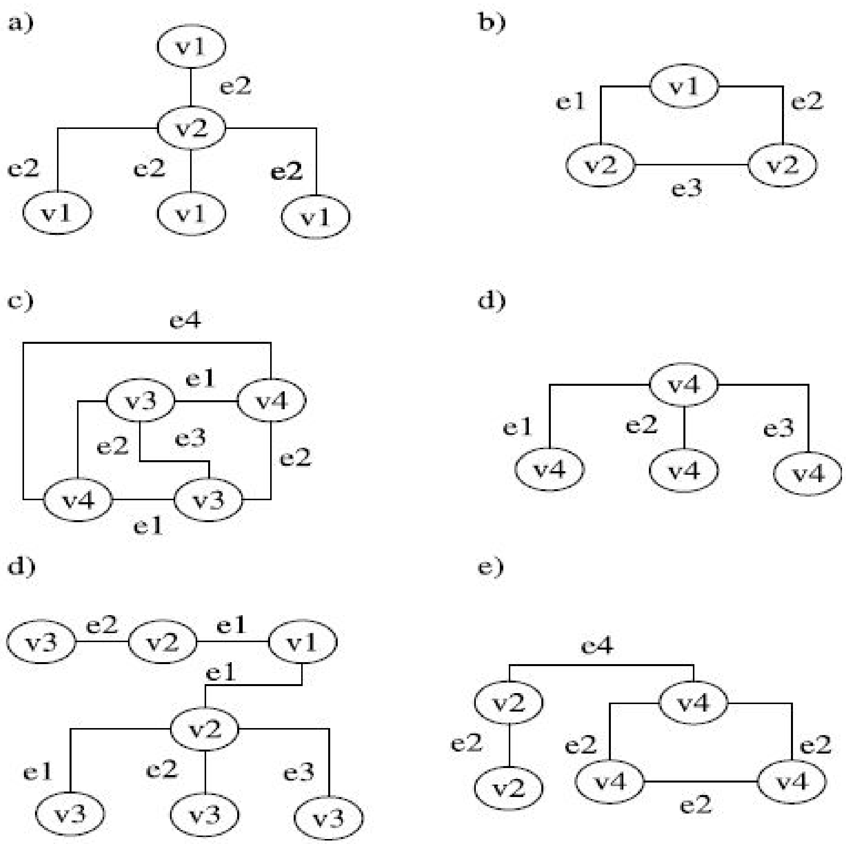
- Clustering weighted graphs (parameter-free, fully automatic, reduced runtime)
- Input data: adjacency matrix (containing weight information)
- Downsplitting of the clusters according to the smallest coding costs suggested by MDL

Summarization-based Mining Bipartite Graphs



- Mining bipartite graphs
- Transforming the original graph into a compact summary graph controlled by MDL
- Contributions: Clustering, hidden structure Mining, link prediction

Subdue: Compression-Based Frequent Pattern Discovery in Graph Data



- Discovering interesting patterns
- Input data: single graph or set of graphs (labeled or unlabeled)
- Outputting substructures that best compress the input data set according to MDL