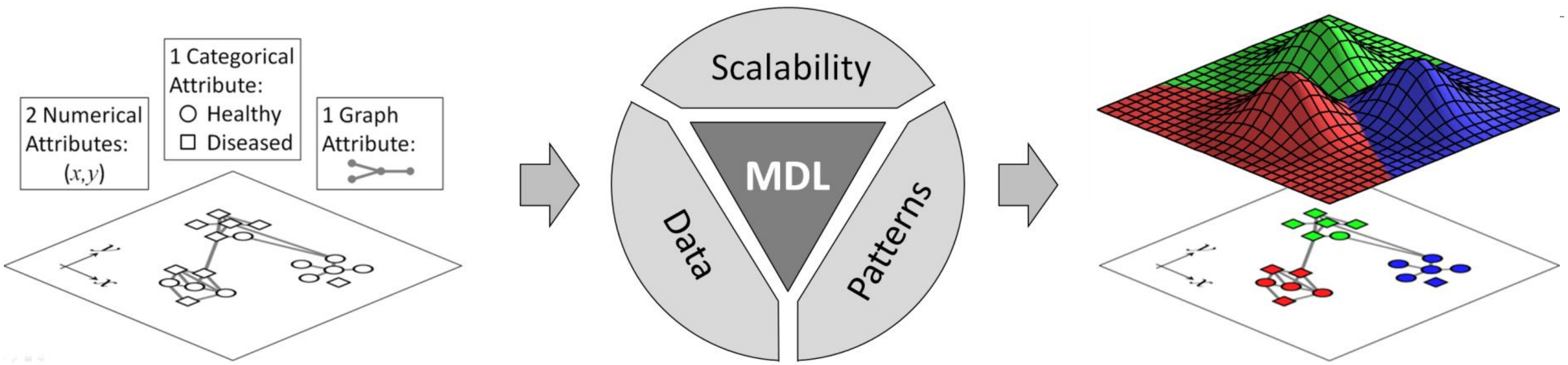
HelmholtzZentrum münchen
Deutsches Forschungszentrum für Gesundheit und Umwelt

Current Topics in Information-theoretic Data Mining

NINA HUBIG, ANNIKA TONCH

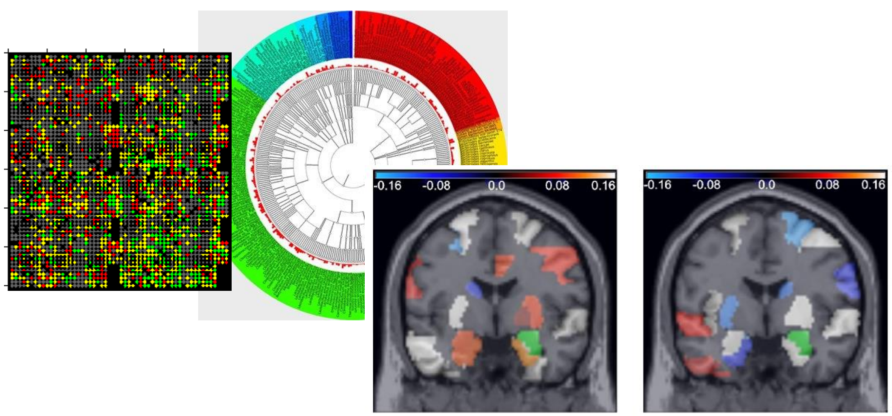
Helmholtz-Hochschul research group iKDD



Heterogeneous Input Data

Information Factory

Valid Output Knowledge



Applications:
Neuroscience,
Diabetes research.

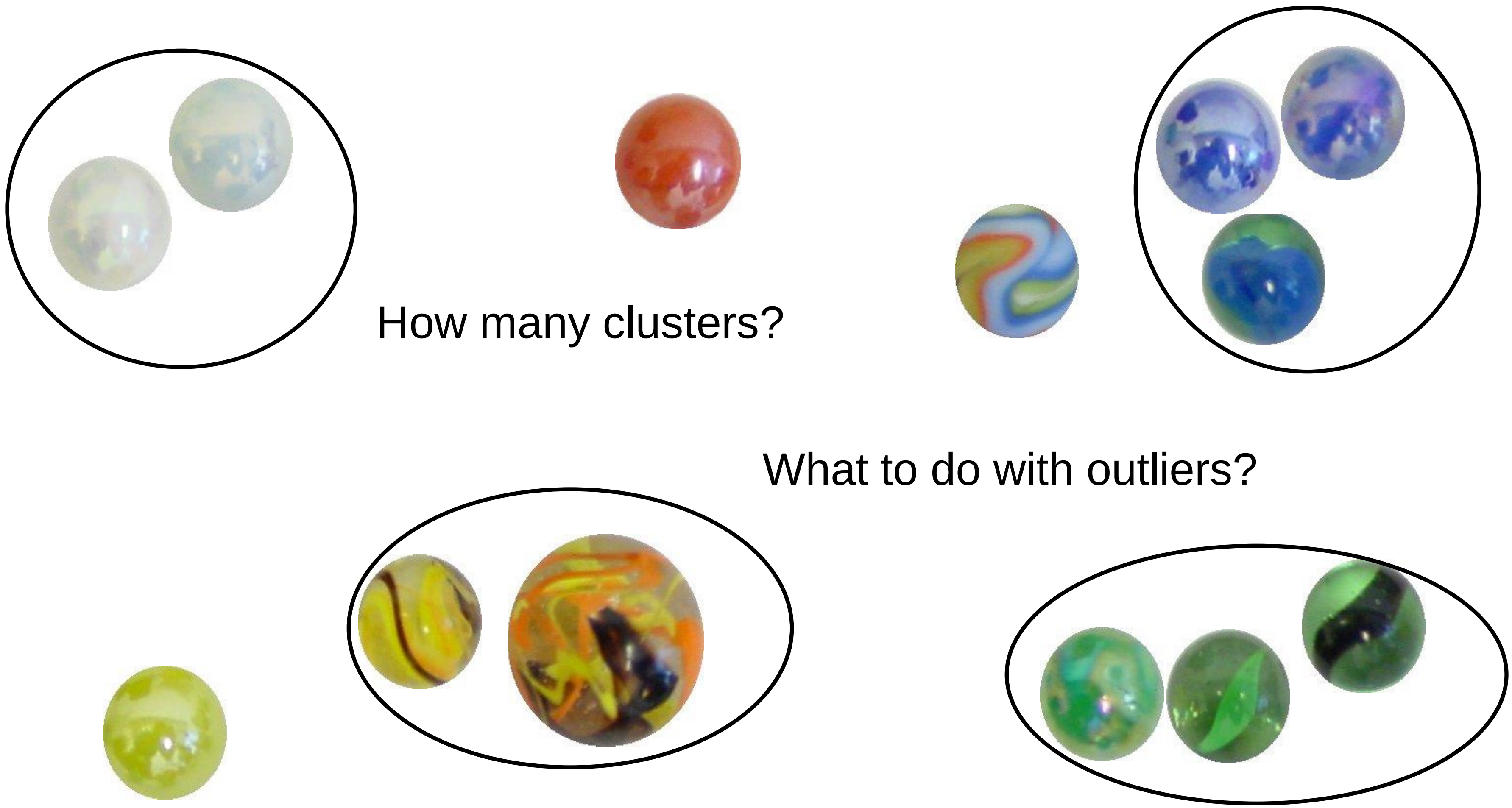
Outline

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

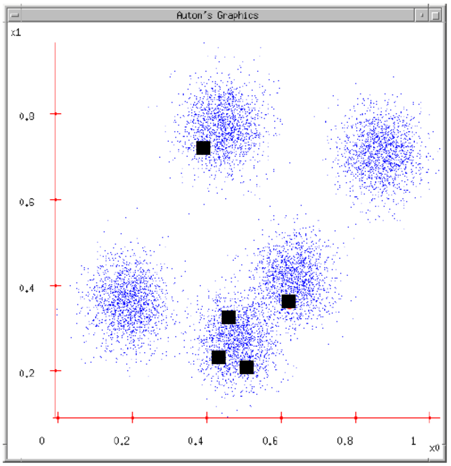
Information-theoretic Data Mining

INTRODUCTION

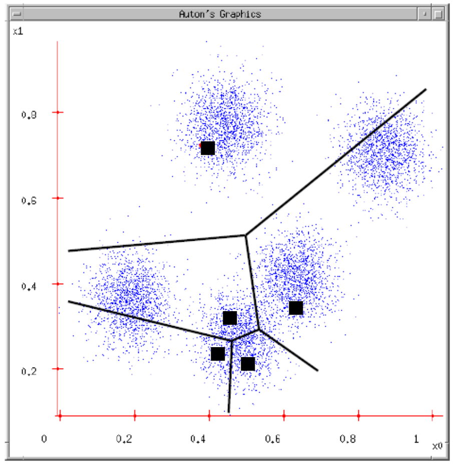
Example Clustering:
Find a natural grouping of the data objects.



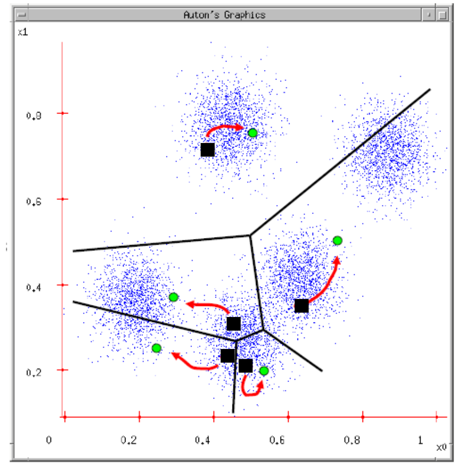
The Algorithm K-Means



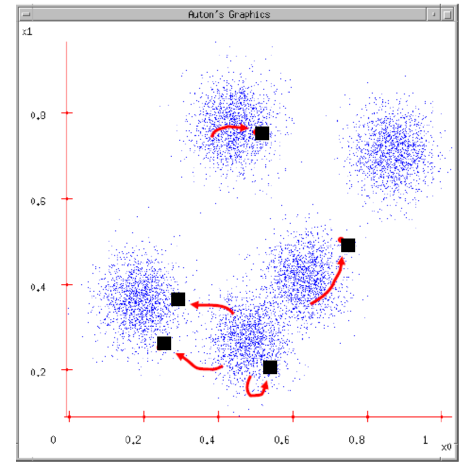
1) **Initialize**
K cluster centers randomly.



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



3) **Update** centers.

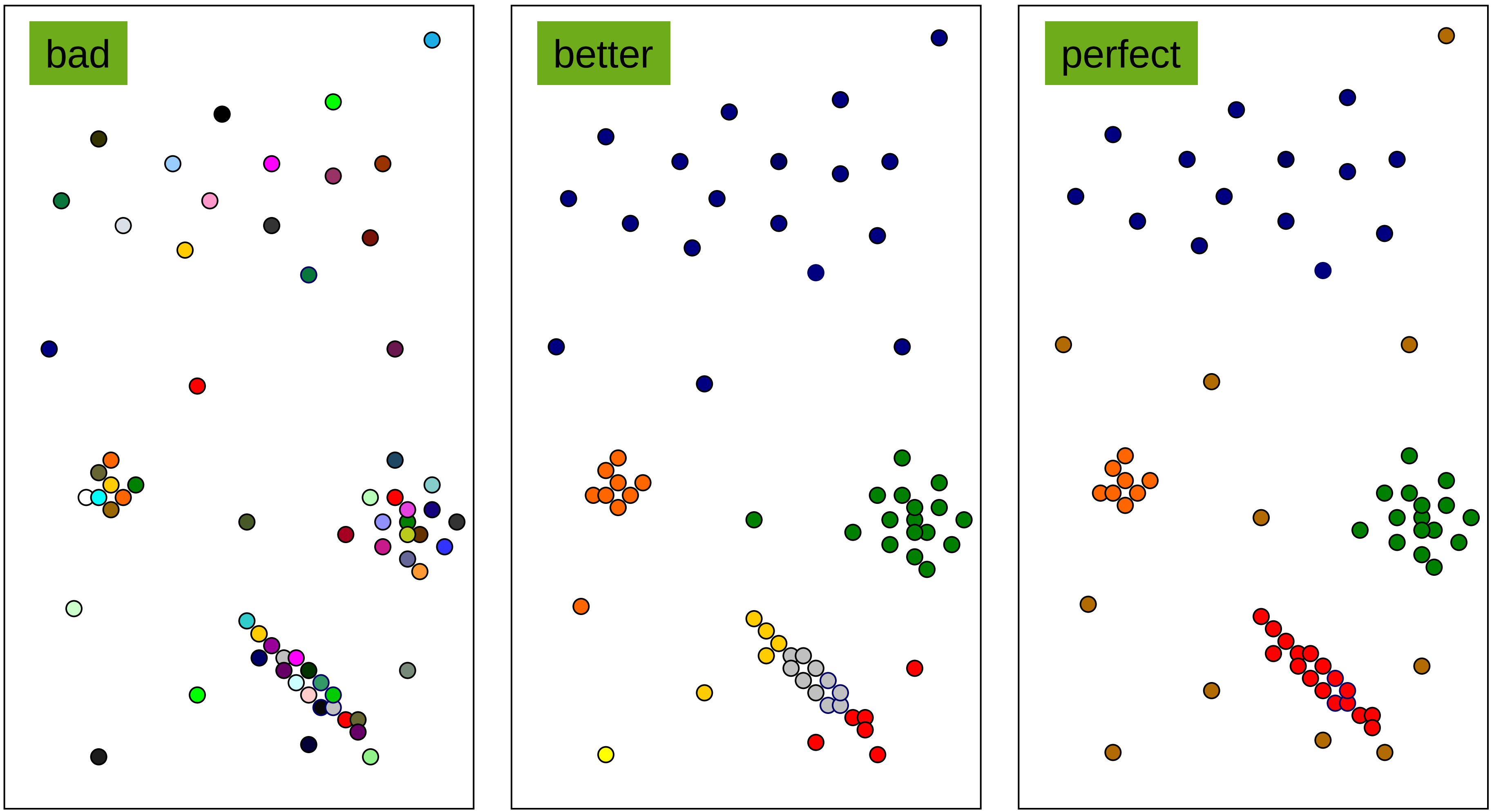


4) **Iterate**
2) und 3) until convergence.

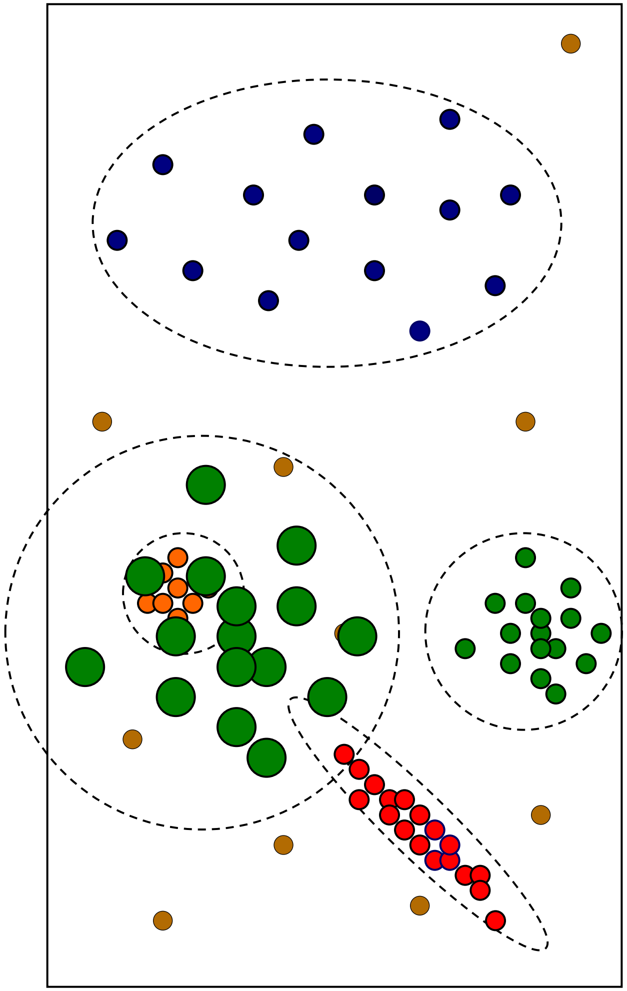
+ fast convergence,
+ well-defined objective function,
+ gives a model describing the result.

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

We need a quality criterion for clustering



Measuring Clustering Quality by Data Compression

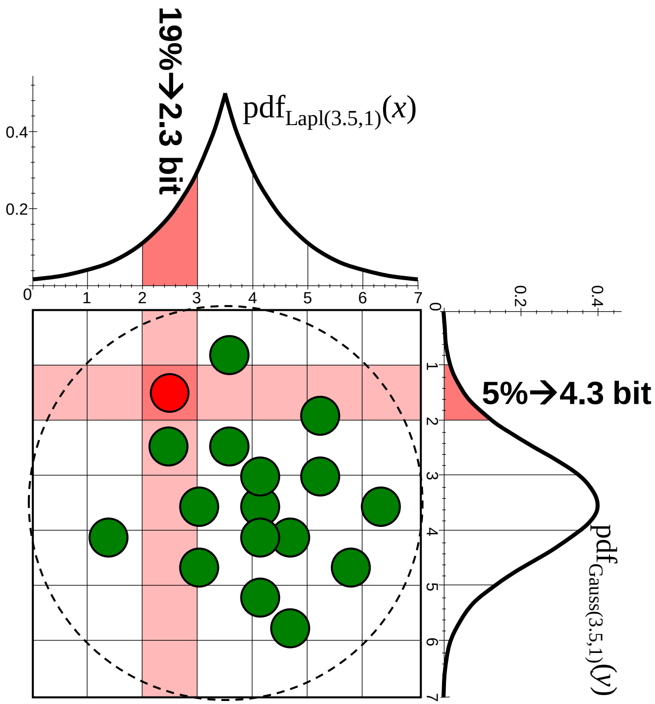


Data compression is a good criterion for…

- 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



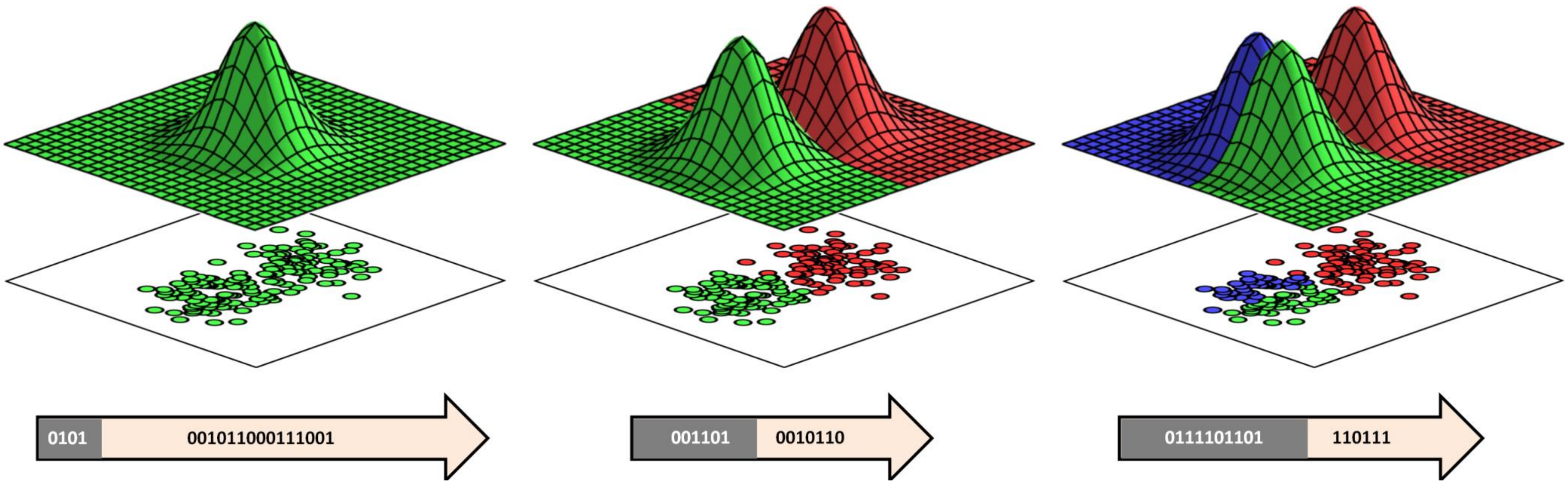
Data compression is a good criterion for…

- 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?

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

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.

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

- **ECTS**: 3 Credits (Bachelor), 6 Credits (Master)
- Master students get the harder papers ;)
- **Presentation**: 20 min presentation/10 min questions. Download the template from the seminar web page
- Write a **report** (max 8 pages).
  - 3-4 pages Bachelor students
  - 5-6 pages Master students
- **Attendance** and **participation** of the seminar meetings
  - ASK the lecturers ;)

- **Seminar days: February 19 -20, time to be announced at the website.**

Contents of the Report

Follow the structure of a scientific publication.

- **Abstract and Introduction**
  - General motivation.
- **State of the Art and Contributions**
  - How is this paper different from (SoA)? e.g What is new? What is better? What is faster?
- **Problem statement**
  - Mathematical formulation
- **Method**
  - Overview: input, output.
  - Method/Algorithm.
- **Results**
  - 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**
  - **YOUR PERSONAL CONCLUSION & IDEAS**
- **References**

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!!)

Topic Selection

FIND YOUR OWN PAPER

HelmholtzZentrum münchen
Deutsches Forschungszentrum für Gesundheit und Umwelt

Mining Numerical and Mixed Data

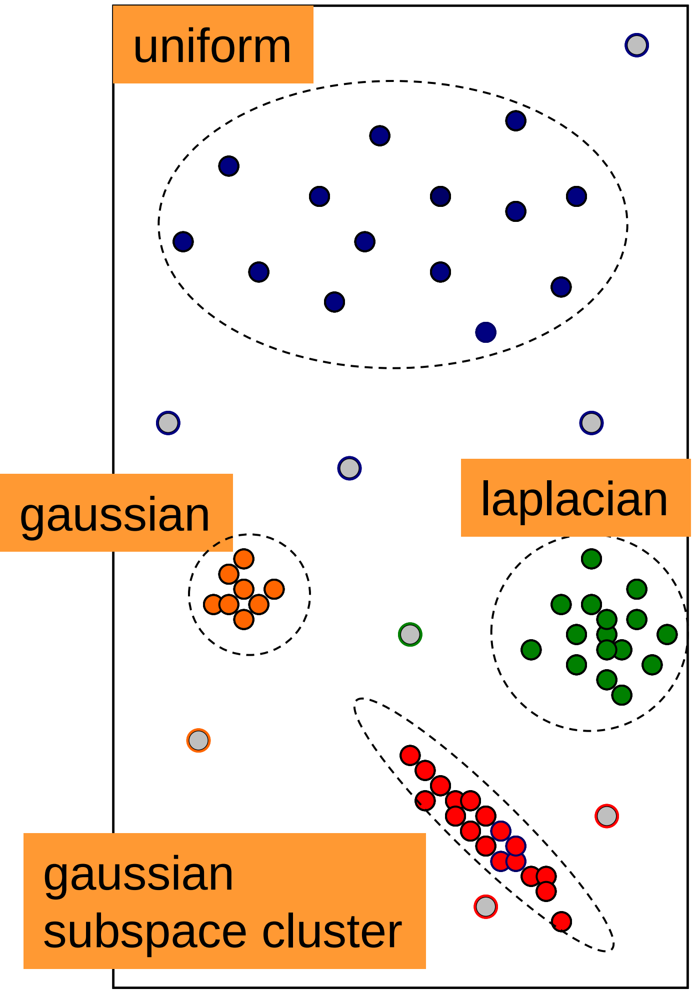
BASIC CLUSTERING

FINDING ALTERNATIVE CLUSTERINGS

MIXED (NUMERICAL, CATEGORICAL DATA)

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

Easy



Start with an arbitrary partitioning

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

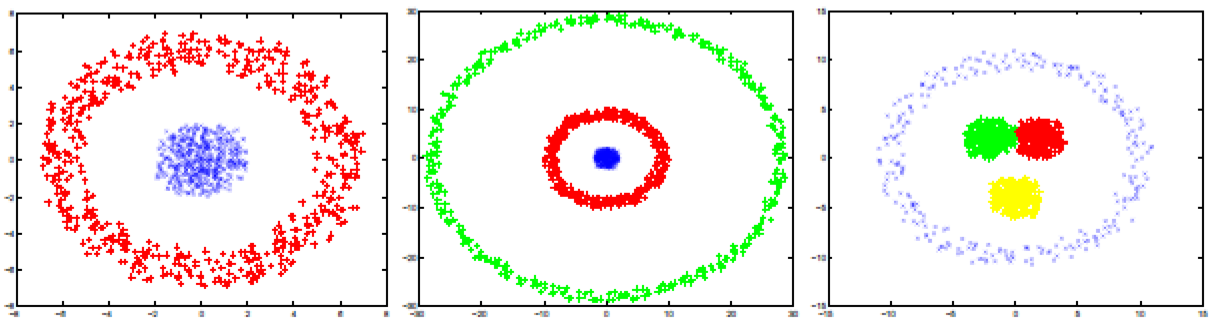
2. Cluster Merging (CM):
Stiches clusters which match well together.

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

Free from sensitve parameter settings !

Hard

A Nonparametric Information- Theoretic Clustering Algorithm

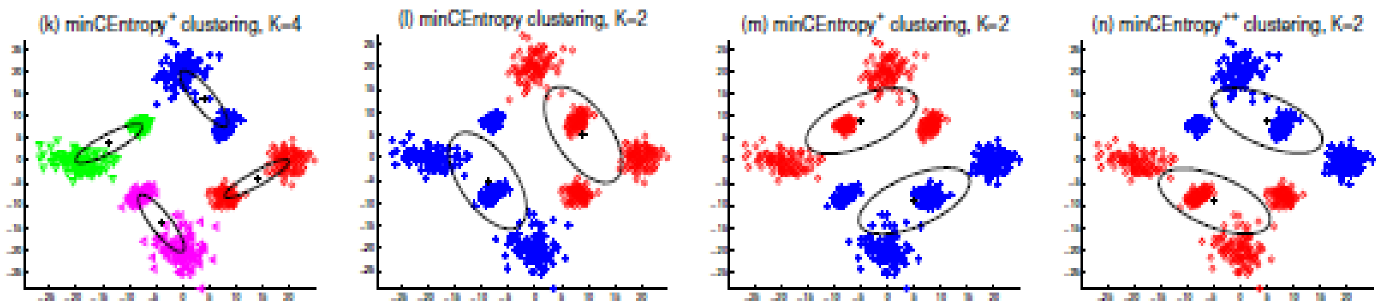


- first google pick for information theoretical clustering ;)
- close to **machine learning**
- uses entropy and **mutual information** as quality function

  ➔ a bit different than our MDL-based approaches!

Easy

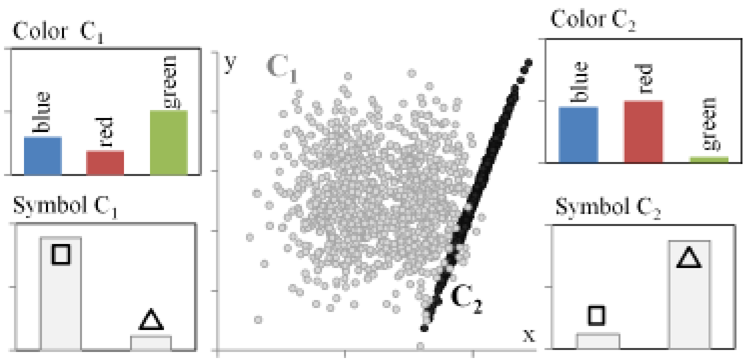
minCEntropy: a Novel Information Theoretic Approach for the Generation of Alternative Clusterings



- 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)

Easy

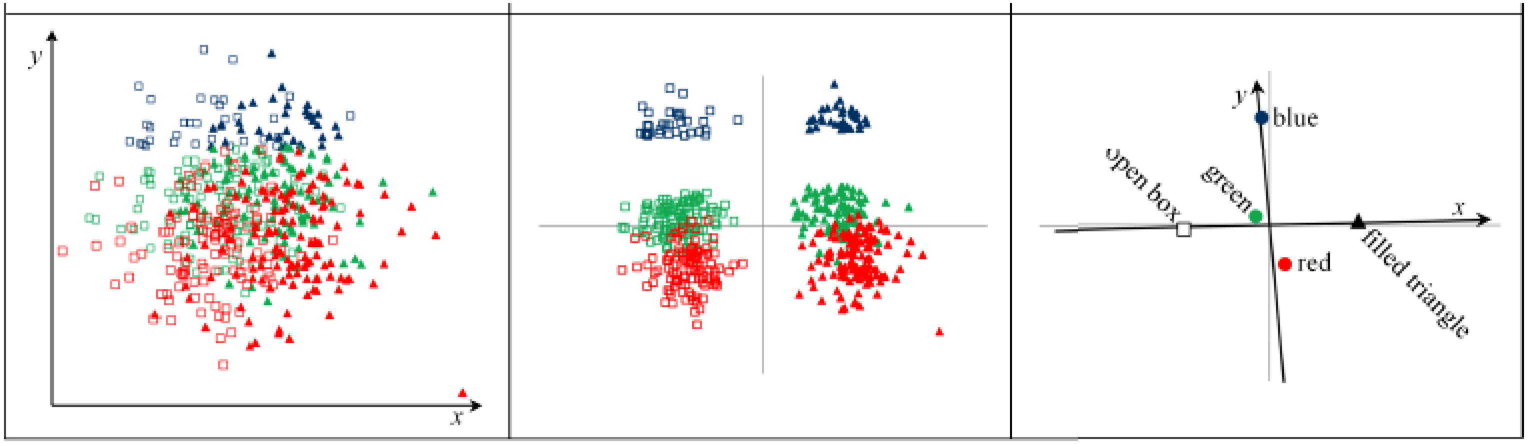
INCONCO: Interpretable Clustering of Numerical and Categorical Objects



- 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

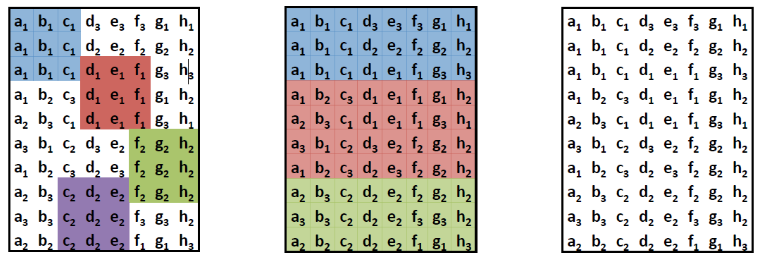
Hard

Dependency Clustering across measurement scales



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

Relevant overlapping subspace clusters on categorical data



Subspace clustering
6076 bits

Full-D clustering
6147 bits

No clustering
6671 bits

- Focus on subspace clustering on **categorical** data.
- Non redundant approach
- Parameter free /automized

HelmholtzZentrum münchen
Deutsches Forschungszentrum für Gesundheit und Umwelt

Graph Mining

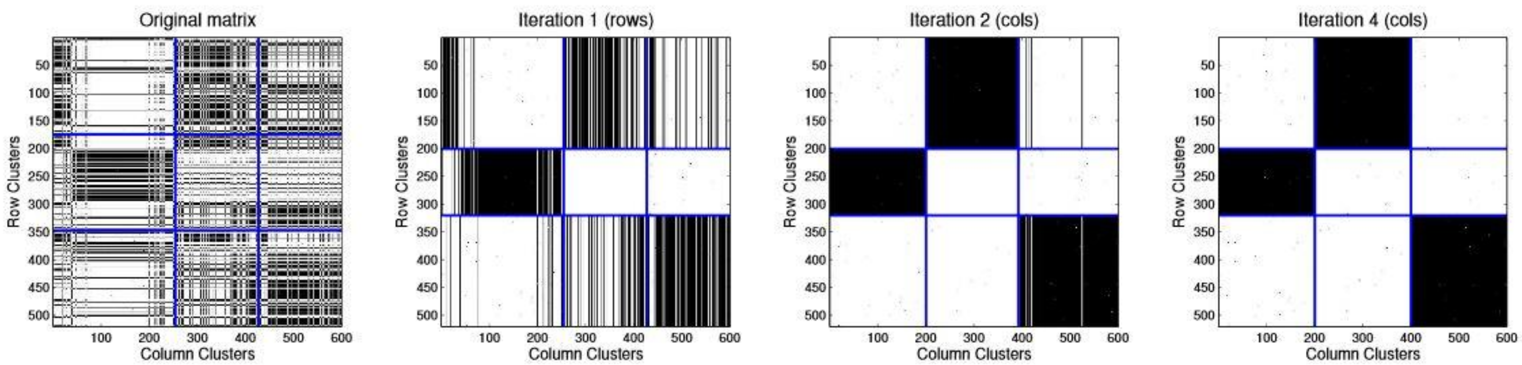
CLUSTERING

WEIGHTED GRAPHS

SUMMARIZATION, STRUCTURE MINING

Hard

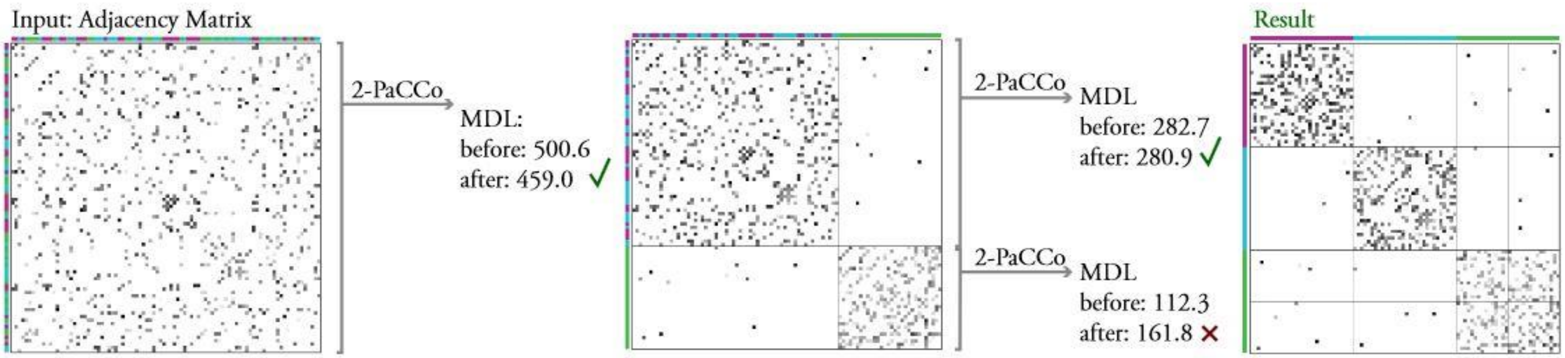
Fully Automatic Cross-Associations



- 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

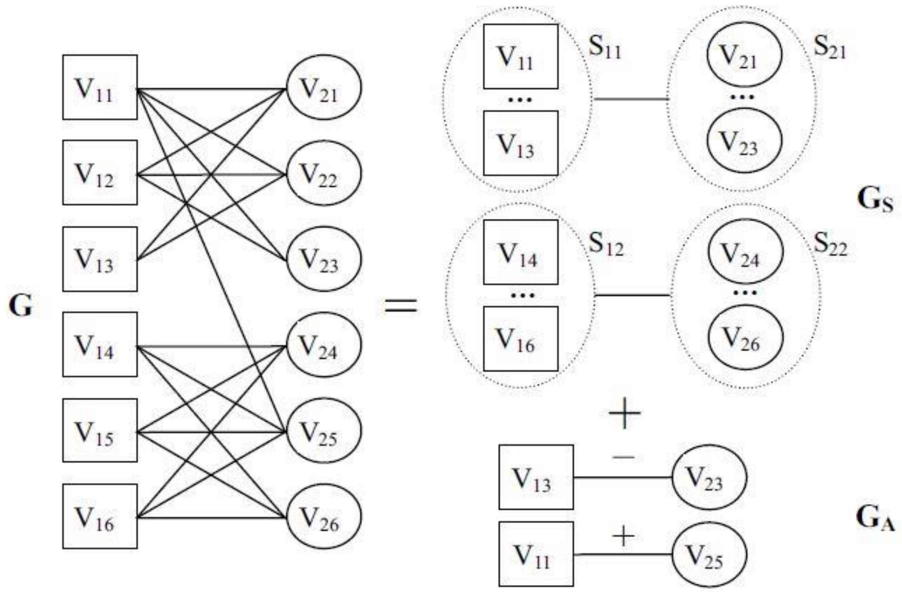
Weighted Graph Compression for Parameter-free Clustering With PaCCo

Easy



- 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

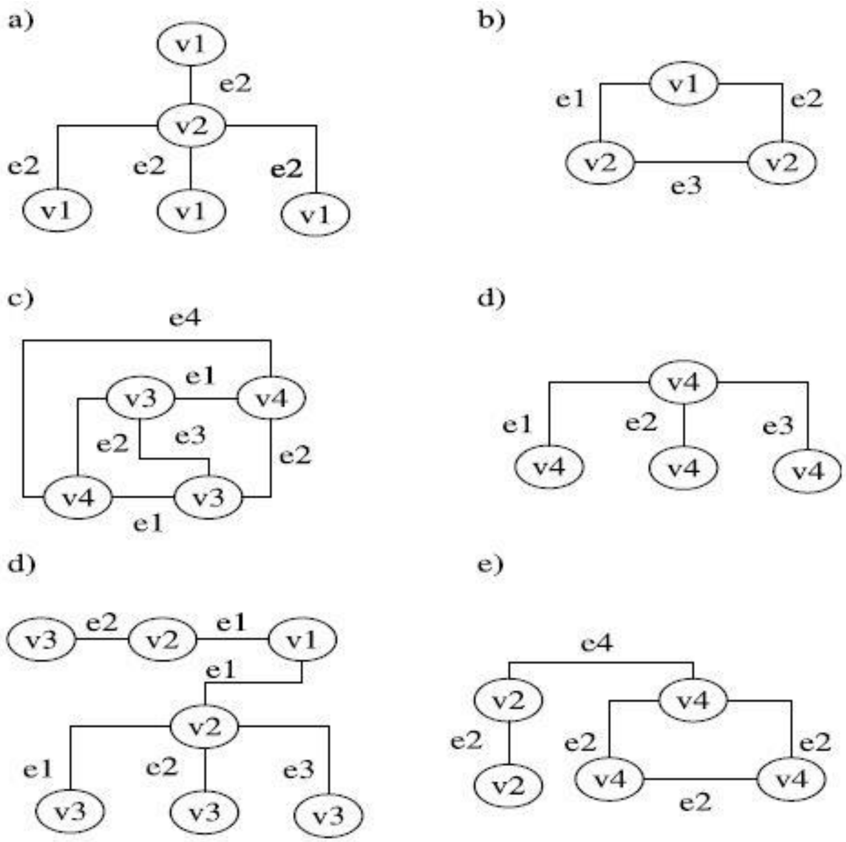
Summarization-based Mining Bipartite Graphs



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

Hard

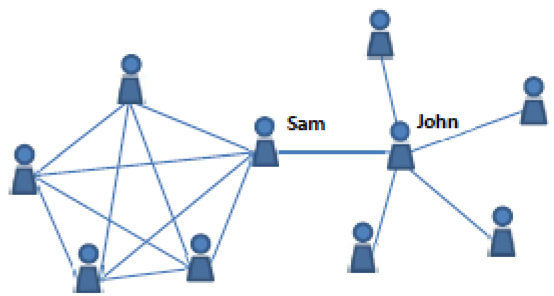
Subdue: Compression-Based Frequent Pattern Discovery in Graph Data



- 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

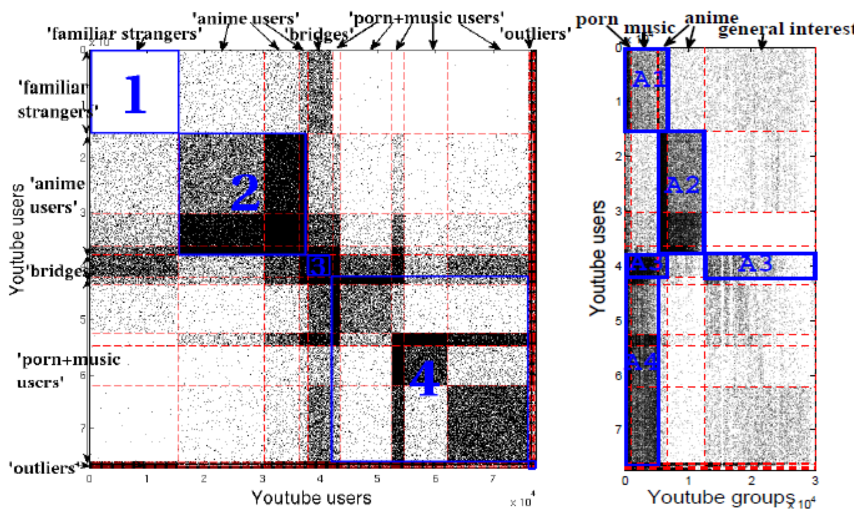
Easy

Compression-based Graph Mining Exploiting Structure Primitives



- Graph clusterer that distinguishes different pattern in graphs
- Suitable for sparse graphs
- Minimum Description Length compression leads to favorizing „stars“ or „cliques“

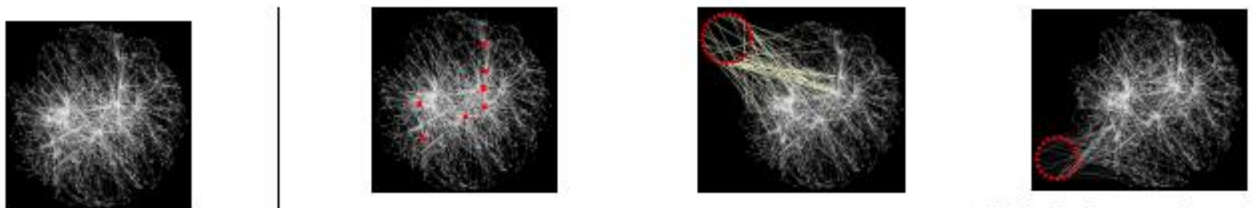
PICS: Parameter-free Identification of Cohesive Subgroups in Large Attributed Graphs



- Summarizes Graphs with node Attributes
- Fully Automatic
- Linear runtime

Hard

VOG: Summarizing and Understanding Large Graphs



(a) Original Wikipedia Controversy graph (with 'spring embedded' layout [15]). No structure stands out.

(b) VoG: 8 out of the 10 most informative structures are stars (their centers in red - Wikipedia editors, heavy contributors etc.).

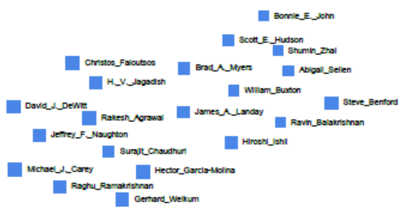
(c) VoG: The most informative bipartite graph - 'edit war' - warring factions (one of them, in the top-left red circle), changing each-other's edits.

(d) VoG: the second most informative bipartite graph - another 'edit war', between vandals (bottom left circle of red points) vs responsible editors (in white).

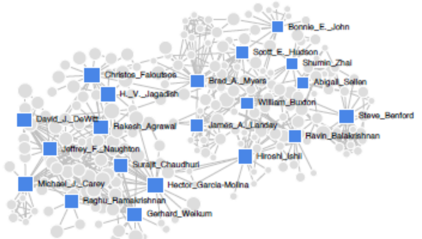
- Compressing a graph with structure patterns: cliques, hubs, chains
- near linear runtime
- Newest paper on the line ;)

Hard

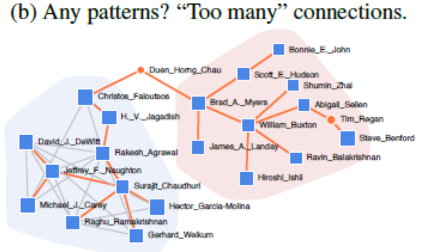
Mining Connection Pathways for Marked Nodes in Large Graphs



(a) What to say about this "list" of authors?



(b) Any patterns? "Too many" connections.



(c) The "right" connections → Better sensemaking

- determining connection pathways ➔ different ways of link analysis
- NP hard problem (travelling salesman)
- Uses minimum description length

Vielen Dank für die Aufmerksamkeit