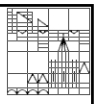
Nycomed Chair for
Bioinformatics & Information Mining

Universität Konstanz



Learning in Parallel Universes

Bernd Wiswedel

15 September, 2008

Overview

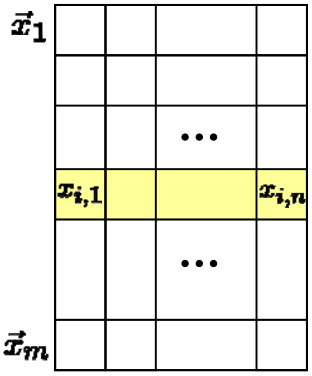
- What are Parallel Universes?
- Application Scenarios
- One sample approach: Neighborgrams
- Connection to LeGo

Motivation

- Data Mining as application to analyse huge amounts of data
- One focus of Data Mining: Find interesting patterns in a data set, e.g. cluster
- Often data very complex, sometimes multiple representations of data available → Parallel Universes

What are Parallel Universes?

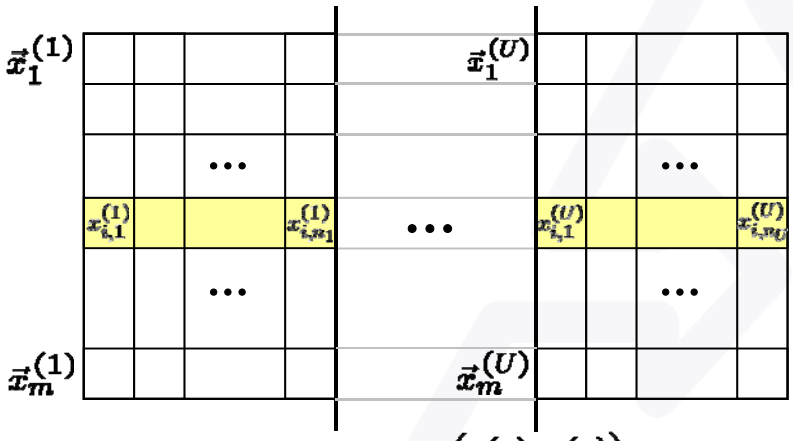
- Usually: Data given in a single feature space
  - Mostly high-dimensional and numeric representation



  - Definition of one, global distance measure $d\left(\vec{x}_i, \vec{x}_j\right)$

What are Parallel Universes?

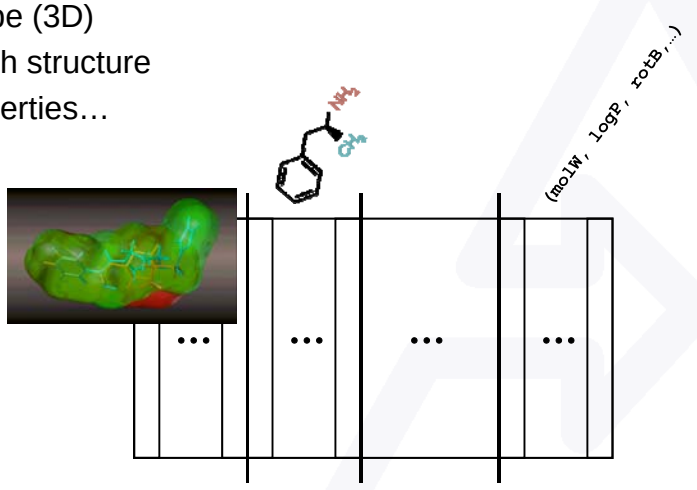
- Parallel Universes
  - Different object representations



  - Different distance measures $d^{(u)}\left(\vec{x}_i^{(u)}, \vec{x}_j^{(u)}\right)$

Why Parallel Universes?

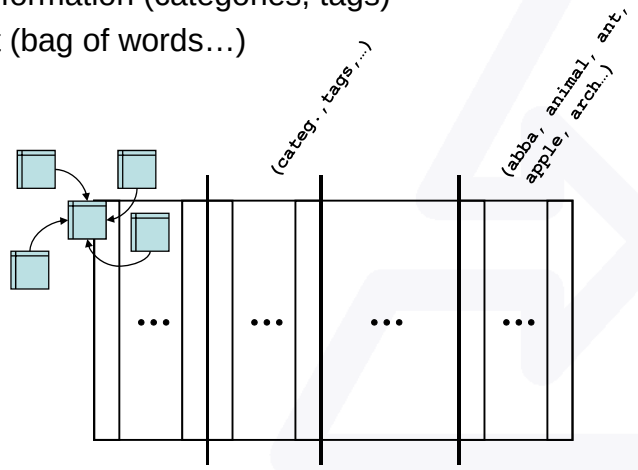
- Example 1: Chemistry - universes encode, e.g.
  - shape (3D)
  - graph structure
  - properties…



see also: A. Bender, R. Glen: ***Molecular similarity: a key technique in molecular informatics***, Org. Biomol. Chem., 2:3204-3218, 2004

Why Parallel Universes?

- Example 2: Web - universes encode, e.g.
  - link structure
  - meta information (categories, tags)
  - content (bag of words…)

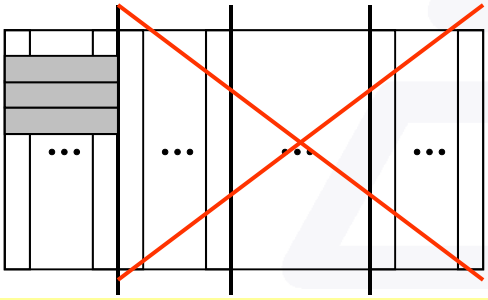


Why Parallel Universes?

- More examples:
  - Music - universes encode
    - semantic meta information (composer, artist, genre,…)
    - groupings (style, category,…)
    - other properties (tempo, beat, key, …)
  - Image or 3D object recognition – universes encode
    - properties (has door, has wheels…)
    - texture information
    - histogram or intensity/color distributions

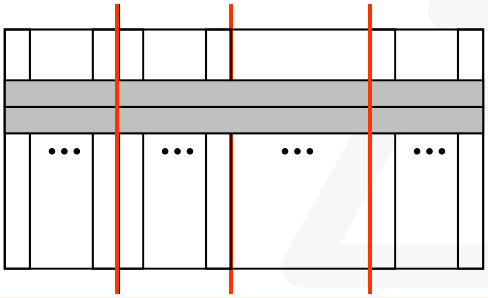
Learning in Parallel Universes

- Naive Approach:
  - Consider only one universe at a time:
    Ignores information in other universes
  - Construct joint feature space:
    often impossible, introduces artifacts.
- Better:
  - Consider all universes at once
  - Allow to identify (local) models that occur only in few (one) universes



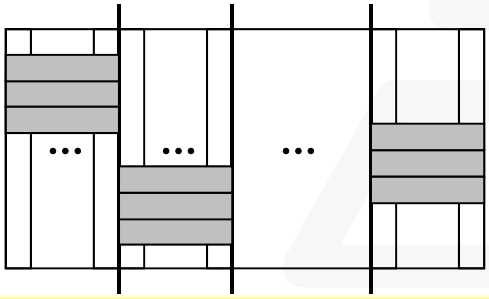
Learning in Parallel Universes

- Naive Approach:
  - Consider only one universe at a time:
    Ignores information in other universes
  - Construct joint feature space:
    often impossible, introduces artifacts.
- Better:
  - Consider all universes at once
  - Allow to identify (local) models that occur only in few (one) universes



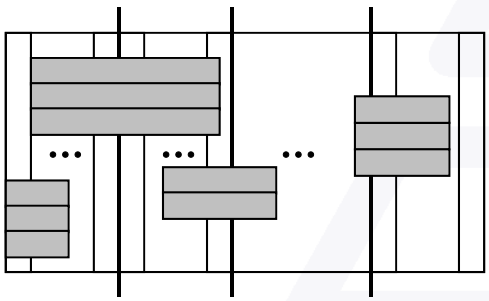
Learning in Parallel Universes

- Naive Approach:
  - Consider only one universe at a time:
    Ignores information in other universes
  - Construct joint feature space:
    often impossible, introduces artifacts.
- Better:
  - Consider all universes at once
  - Allow to identify (local) models that occur only in few (one) universes



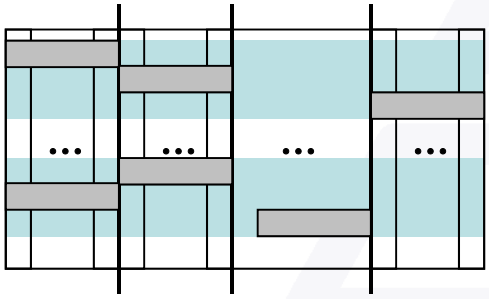
Related Approaches: Subspace Clustering

- choose subset of data and attributes for each cluster
  - usually no interpretation of subspaces possible
  - selects from one, large universe
  - first finds also overlapping clusters
  - most prominent approaches: CLIQUE, COSA



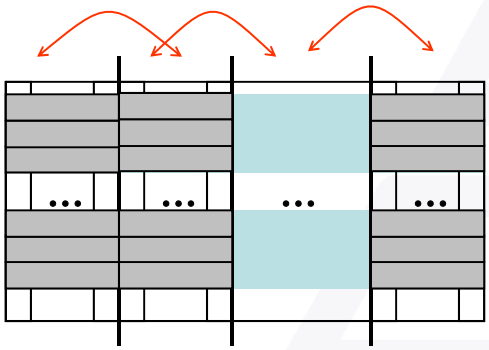
Related Approaches: Multi-Instance Learning

- each object has several possible representations in same space (e.g. molecular confirmations in 3D)
  - universes all possess the same semantics
  - two extremes: similar in all universes, similar in at least one universe.
  - number of universes per object can vary.



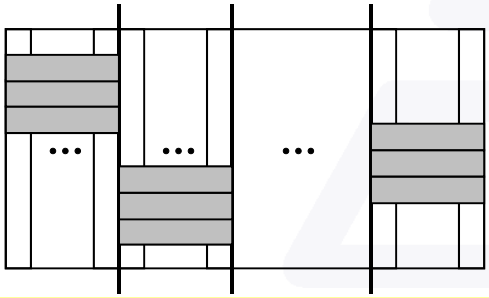
Related Approaches: Multi-View Learning

- each object has several possible representations in different spaces
  - universes with different semantics
  - independent and complete models in each universe (learning algorithms may assist each other)



Learning in Parallel Universes

- Clear separation of Universes (a-priori given)
- Each individual universe does not suffice for learning
- Allow to identify (local) models that occur only in few (one) universes
- Identify overlaps



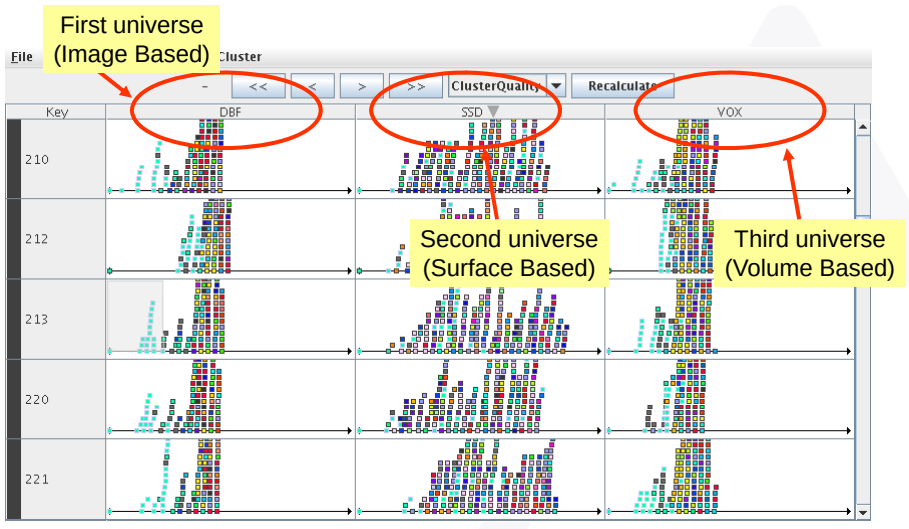
One sample approach: Neighborgrams

- Supervised approach
- Construct local neighorhood histogram („Neighborgrams“) for objects of interest in all universes
- Derive quality values for individual neighborgrams
- Covering-like approach to construct classification model

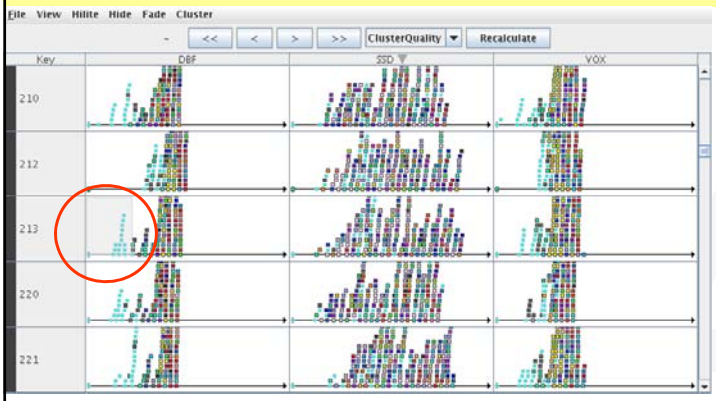
- Intuitive visualization allows for interactive exploration and user-controlled model construction

Neighborgrams on KN-DB

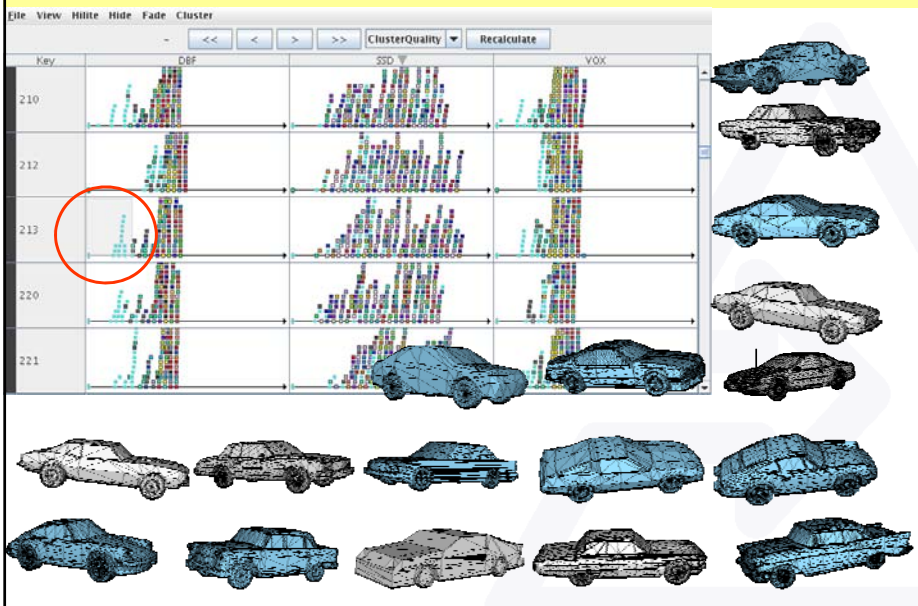
Neighborgrams on KN-DB



Neighborgrams on KN-DB

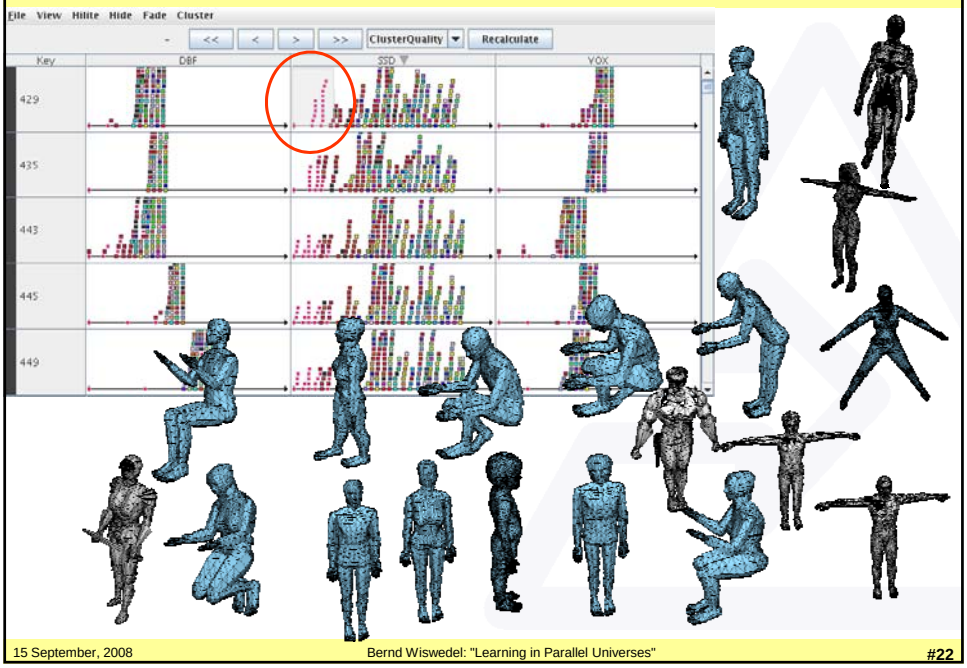


Neighborgrams on KN-DB



Neighborgrams on KN-DB

Neighborgrams on KN-DB



Summary Neighborgrams

- Visualization tool for interactive exploration of clusters
- Works well for small size data sets or to model minority class
- Manual clustering
- Semi-Automatic clustering
  - Inspect proposed cluster
  - Discard, accept or fine-tune cluster
- Fully automatic clustering
  - Sequential covering approach
  - Identify greedily the next best cluster, remove covered objects, restart

Connection to LeGo

- Output is selected set of Neighborgram Clusters, spread over different universes
- Such clusters can be considered as local patterns
- Open problem: Construction of a global model as opposed to a simply aggregation of clusters
- Special focus on identifying overlaps among universes (often of special interest)

Summary

- Learning in Parallel Universes as simultaneous analysis of multiple descriptor spaces
- Encompasses identification of patterns that:
  - are specific to individual universes and
  - span multiple universes (not necessarily all)
- Final model construction comprises all previously identified patterns

Thanks!