Advanced Topics in Pattern Mining
- Introduction -

PhD Course – Szeged, 2013

Tamás Horváth

University of Bonn &
Fraunhofer IAIS, Sankt Augustin, Germany
tamas.horvath@iais.fraunhofer.de

Slides 2-10 are taken from Stefan Wrobel

universitätbonn
RHEINISCHE FRIEDRICH-WILHELMS-UNIVERSITÄT

Fraunhofer
IAIS

Fraunhofer IAIS: Intelligent Analysis and Information Systems

''From sensor data to business intelligence, from media analysis to visual information systems:
Our research allows companies to do more with data''

- 240 people: scientists, project engineers, technical and administrative staff
- Located on Fraunhofer Campus Schloss Birlinghoven/Bonn
- Joint research groups and cooperation with

universitätbonn
RHEINISCHE FRIEDRICH-WILHELMS-UNIVERSITÄT

b-it

Fachhochschule Bonn-Rhein-Sieg

Director: Prof. Dr. Stefan Wrobel

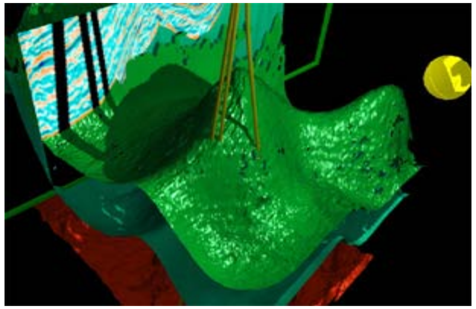
universitätbonn
RHEINISCHE FRIEDRICH-WILHELMS-UNIVERSITÄT

Fraunhofer IAIS

Fraunhofer IAIS: Intelligent Analysis and Information Systems

Core research areas:

- machine learning and adaptive systems
- data mining and business intelligence
- automated media analysis
- interactive access and exploration
- autonomous systems



Why is Knowledge Discovery becoming more and more important? -- Four Current Trends

Convergence

Ubiquitous intelligent system`s

Users as producers

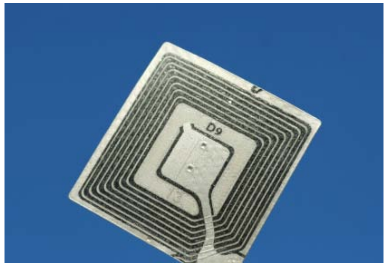
Networked autonomy

Convergence

- Universal digital representation of any media content
  - Web, MP3, digital cameras, Video
- Internet formats replace traditional delivery channels
  - Online Magazines, Blogs, Podcasts, Webradio, IPTV, Video on Demand
- Explosive growth of accessible media assets
  - digitalisation, crosslinking, swapping

- Automated search, structuring, classification and selection are of central relevance

Ubiquitous Intelligent Systems



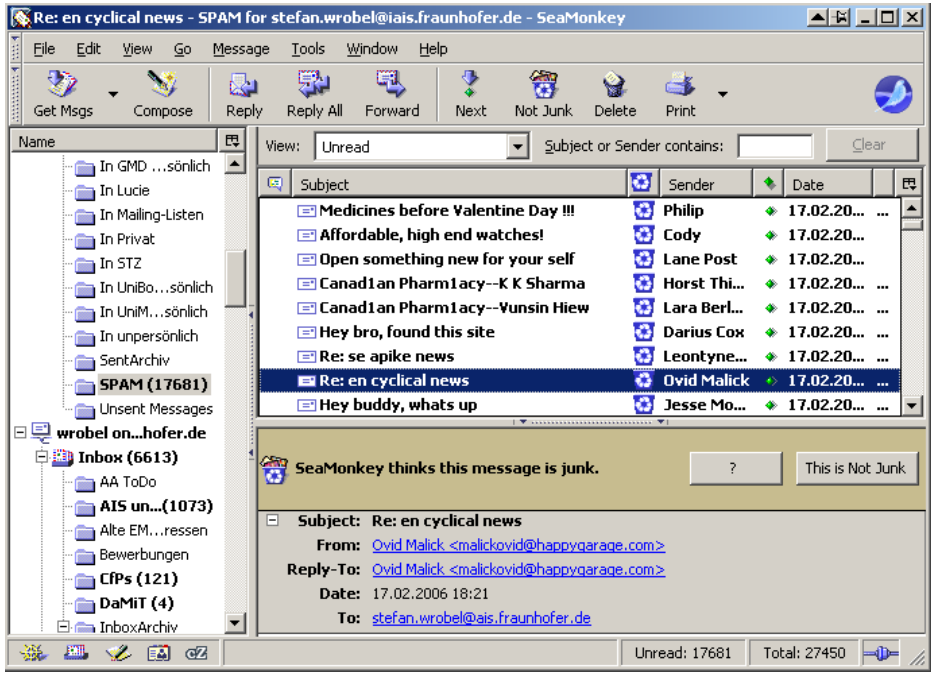
- Personal devices, integrated processors (Factor 20 – 30 above PCs)
- Interactivity, Sensors, Actuators
- Enormous production of data
- Physical and virtual worlds merge

Users as Producers

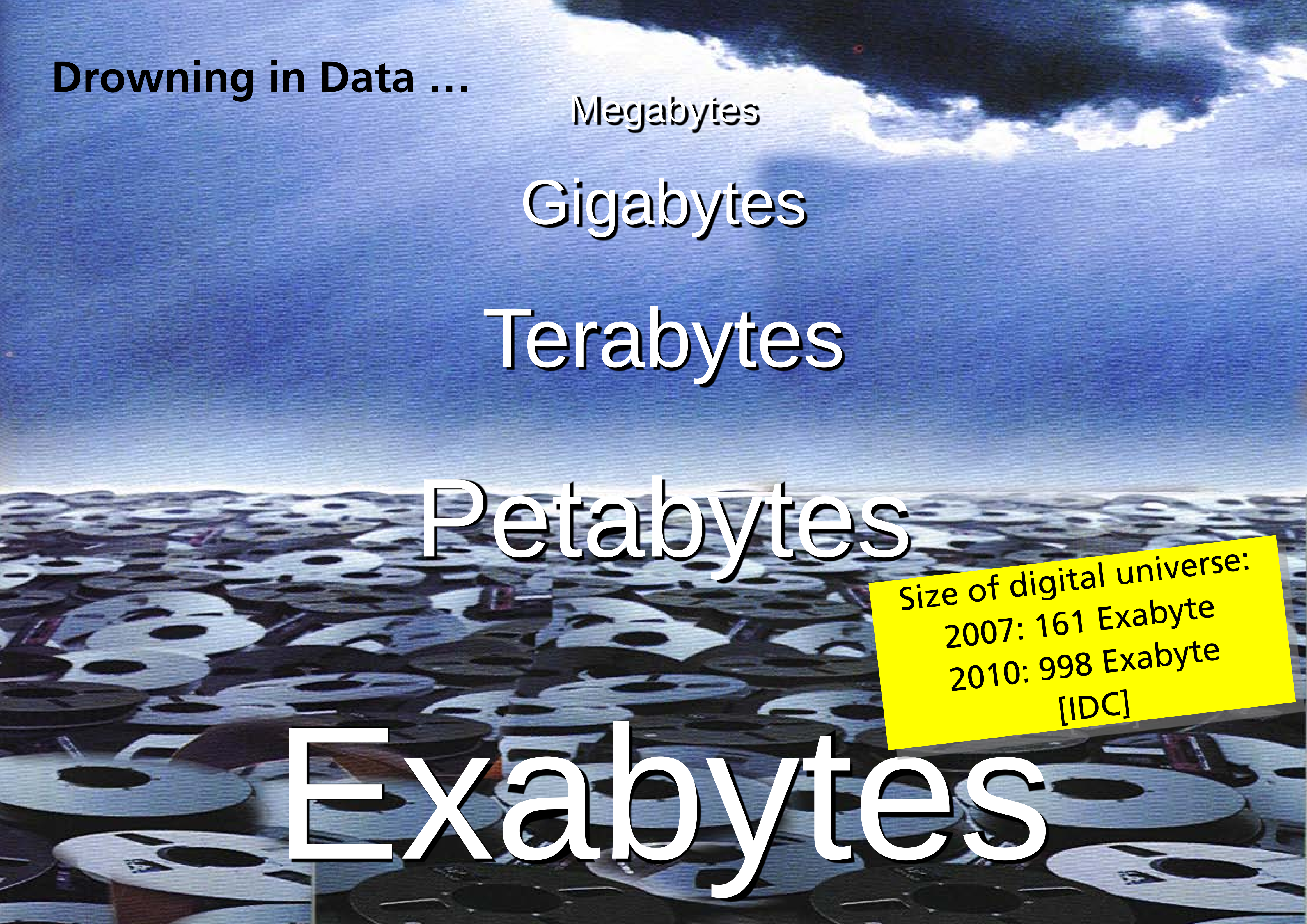
- Web 2.0, Social Web, Crowdsourcing
- Exploding growth of content
- Media providers transform from content to confidence providers, competing with social communities
- Users expect full interactivity and control
- Quality control, confidence, choice and searching are becoming central

Networked Autonomy

- Growing readiness to use loosely controlled systems (autonomous agents)
- Loosely coupled company structures
- Service orientation (SOA) in IT systems
- First mobile autonomous systems
- Flexibility and capability for autonomous decisions on the basis of observations and goals is becoming central



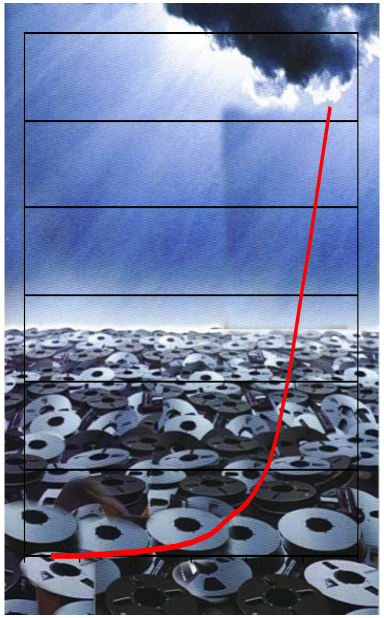
Drowning in Data ...



Challenges and Research Opportunities

- Amount and variety of available data is growing with enormous dynamics
- Systems, people and organizations cannot handle them but must use the knowledge hidden in those data is crucial for making the right decisions!
- Autonomous agents and systems must process sensor data and make intelligent decisions

→ We need **machine learning** and **data mining**! More than ever.



Machine Learning and Data Mining

Machine Learning

- "Machine learning refers to a system capable of the autonomous acquisition and integration of knowledge. This capacity to learn from experience, analytical observation, and other means, results in a system that can continuously self-improve and thereby offer increased efficiency and effectiveness." [AAAI Webpage]

Knowledge Discovery/Data Mining

- "Knowledge Discovery in Databases is the nontrivial process of identifying valid, novel, potentially useful and ultimately understandable **patterns** in large databases." [Fayyad et.al., 1996]

Global vs. Local Models

- **machine learning:** usually searches for **global** models
  - **global patterns** (e.g., decision trees, separating hyperplanes, etc.)
    - given any possible object, a global pattern (e.g., a decision tree) can be used to make a prediction
- **(descriptive) data mining:** usually searches for **local** models
  - **local patterns** (e.g., association rules, subgroups etc.)
    - for many objects, the model simply "does not apply" (contains no information)
    - for those where it does apply, it reports a *descriptive* characteristic which is not necessarily sufficient to make a prediction

Two Problem Examples

1. **machine learning:**
   - on-line learning of conjunctive concepts from examples in the mistake bound model
     - global predictive pattern

2. **data mining:**
   - association rule mining
     - local descriptive pattern

Formal Models of Learning

a formal model of learning can be defined by specifying the following components:

1. **Learner:** Who is doing the learning?
2. **Domain:** What is being learned?
3. **Information Source:** From what is the learner learning?
4. **Prior knowledge:** What does the learner know about the domain initially?
5. **Performance Criteria:** How do we know whether, or how well, the learner has learned? What is the learner‘s output?

Components of Formal Models of Learning

1. **Learner:** Who is doing the learning?

   typically a computer program that may be restricted, e.g.

   - it must work in *polynomial time*
   - it must use only *finite memory*
   - …

Components of Formal Models of Learning

2. Domain: What is being learned? E.g.,

- an unknown *concept*
  (rule separating positive examples from negative examples)
- an unknown *function*
- an unknown *language*
- …

universitätbonn
RHEINISCHE FRIEDRICH-WILHELMS-UNIVERSITÄT
Fraunhofer IAIS

Components of Formal Models of Learning

3. Information Source: From what is the learner learning? E.g.,

a) The learner is given +/- *labeled examples*
(can be chosen at random, arbitrarily, maliciously by some adversary, by a helpful teacher, etc.)

b) The learner may ask *questions,* e.g.,

- **membership** queries (e.g., $w \in L$? Answer YES/NO)
- **equivalence** queries (e.g. $L' = L$? Answer YES/counterexample)

Is the information corrupted by **noise**?

c) …

Components of Formal Models of Learning

4. **Prior knowledge:**

   What does the learner know about the domain initially?

   (e.g., the unknown concept is representable in a certain way)

Components of Formal Models of Learning

5. Performance Criteria:

How do we know whether, or how well, the learner has learned? What is the learner‘s output?

- off-line vs. on-line measures
- descriptive output vs. predictive output
- accuracy (error rate, number of mistakes during learning)
- …

Example 1: On-line learning of conjunctive concepts with mistake-bound measure

The Model:

1. **Learner:** *computer program* (no time/space restriction)
2. **Domain:** $X = \{0,1\}^n$; *unknown concept* $c : X \to \{0,1\}$
3. **Source:** $\langle \vec{x}_1, c(\vec{x}_1)\rangle, \langle \vec{x}_2, c(\vec{x}_2)\rangle, \ldots$; arbitrary, noise-free
4. **Prior knowledge:** c is a conjunction (e.g., $v_3\overline{v}_5v_8$)
5. **Performance:** on-line prediction (#prediction mistakes)

On-line learning of conjunctive concepts

Algorithm 1:

1. **let** $h = v_1\overline{v}_1v_2\overline{v}_2\ldots v_n\overline{v}_n$
2. **for all** $\vec{x}_i$ with $c(\vec{x}_i) \neq h(\vec{x}_i)$
3. remove from h all literals that are false in $\vec{x}_i$

example: $n = 3$, $c = v_1\overline{v}_3$ (unknown!), $\langle 110, 1\rangle, \langle 111, 0\rangle, \langle 100, 1\rangle, \ldots$

- $h = v_1\overline{v}_1v_2\overline{v}_2v_3\overline{v}_3$

  $\langle 110, 1\rangle$**:** $c(110) \neq h(110) \Rightarrow h = v_1v_2\overline{v}_3$

  $\langle 111, 0\rangle$**:** $c(111) = h(111) \Rightarrow$ **no change**

  $\langle 100, 1\rangle$**:** $c(100) \neq h(100) \Rightarrow h = v_1\overline{v}_3$

On-line learning of conjunctive concepts

Theorem 1: On-line learning of conjunctive concepts can be done with at most $n+1$ prediction mistakes

Proof Sketch: The proof follows from Lemmas 1-3, noting that worst-case occurs when the target concept *c to be learned* is **true.**

Lemma 1:(*correctness*)**:** No literal in *c* is ever removed from *h*.

Lemma 2: Each mistake causes at least one literal to be removed from *h***.** (Note that *mistakes are only made on positive examples*!)

Lemma 3: The first mistake causes *n* literals to be removed from *h*.

Example II (Data Mining): Mining Association Rules

Example: Market-basket transactions

Analysis of purchase "basket" data (items purchased together) in a department store

TID	*Items*
1	**Bread, Milk**
2	**Bread, Diaper, Beer, Eggs**
3	**Milk, Diaper, Beer, Coke**
4	**Bread, Milk, Diaper, Beer**
5	**Bread, Milk, Diaper, Coke**

Example of Association Rules:

{Diaper} → {Beer}
{Milk, Bread} → {Eggs,Coke}
{Beer, Bread} → {Milk}

Implication means co-occurrence, not causality!

Association Rules: Notions and Notations

- I: set of items
- itemset: collection of one or more items
- transaction: itemset
- transaction database D: multiset of transactions

TID	*Items*
1	**Bread, Milk**
2	**Bread, Diaper, Beer, Eggs**
3	**Milk, Diaper, Beer, Coke**
4	**Bread, Milk, Diaper, Beer**
5	**Bread, Milk, Diaper, Beer**

Association Rules: Notions and Notations

support set $D[X]$ of an itemset X:

$$D[X] = \{T : T \in D \text{ and } X \subseteq T\}$$

– multiset of sets

support (s): fraction of transactions that contain an itemset, i.e., for $X \subseteq I$

$$s(X) = \frac{|D[X]|}{|D|}$$

frequent itemset: itemset with support greater than or equal to a threshold $minsup$

TID	*Items*
1	**Bread, Milk**
2	**Bread, Diaper, Beer, Eggs**
3	**Milk, Diaper, Beer, Coke**
4	**Bread, Milk, Diaper, Beer**
5	**Bread, Milk, Diaper, Coke**

example:

- $s(\{Milk, Diaper, Beer\}) = 2/5$

Association Rules

- **association rule**
  - implication expression of the form $X \to Y$, where X and Y are itemsets
    - **example:** {Milk, Diaper} → {Beer}

- **rule evaluation metrics**
  - support (s): fraction of transactions that contain both X and Y
  - confidence (c): fraction of transactions that contain both X and Y relative to the transactions that contain X

TID	*Items*
1	**Bread, Milk**
2	**Bread, Diaper, Beer, Eggs**
3	**Milk, Diaper, Beer, Coke**
4	**Bread, Milk, Diaper, Beer**
5	**Bread, Milk, Diaper, Coke**

example: $R = \{Milk, Diaper\} \to \{Beer\}$

$$s(R) = \frac{|D[\{Milk, Diaper, Beer\}]|}{|D|} = \frac{2}{5}$$

$$c(R) = \frac{|D[\{Milk, Diaper, Beer\}]|}{|D[\{Milk, Diaper\}]|} = \frac{2}{3}$$

universitätbonn

RHEINISCHE FRIEDRICH-WILHELMS-UNIVERSITÄT

Fraunhofer IAIS

Applications of Association Rules

- cross-marketing
- attached mailing
- catalog design
- loss-leader analysis
- store layout
- customer segmentation based on buying patterns

Mining Association Rules

Given

- a *transaction database* D over a set I of items and
- *thresholds* $minsup$ and $minconf$

find all association rules $X \to Y$ satisfying

$$s(X \to Y) \geq minsup \text{ and } c(X \to Y) \geq minconf$$

Brute-Force Approach

1. list all possible association rules
2. compute the support and confidence for each rule
3. prune rules that fail the *minsup* and *minconf* thresholds

computationally prohibitive

- total number of *possible* association rules is exponential in the cardinality of the set of all items

⇨ **exponential delay** in worst case

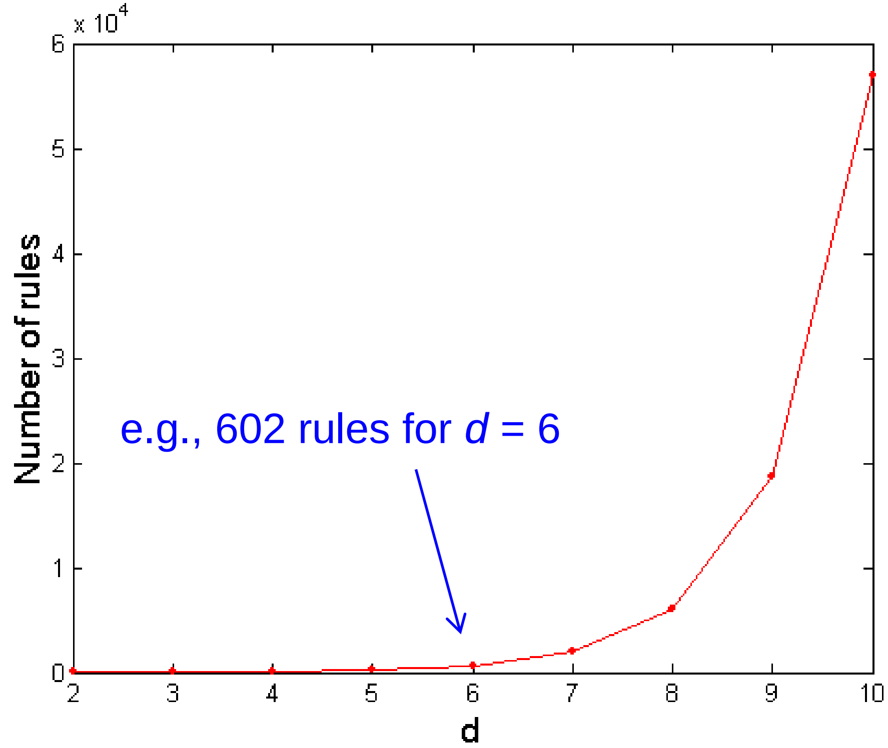
Upper Bound on the Number of Association Rules

let $d = |I|$

$\Rightarrow$ total number of itemsets is 2^d

$\Rightarrow$ total number of possible association rules is $3^d - 2^{d+1} + 1$

Proof: *exercise*



Mining Association Rules

two-step approach:

1. **frequent itemset** generation
   - generate all itemsets whose support ≥ *minsup*
   - use e.g. the Apriori or the FP-Growth Algorithm

2. **rule** generation
   - generate association rules of confidence ≥ *minconf* from each frequent itemset *X* by binary partitioning of *X*

Input to a Typical Machine Learning/Data Mining Problem

single relation

- can be represented by a **single table of fixed length**
  - **rows:** objects/instances
  - **columns:** attributes

previous two examples:

- *learning conjunctions:* each training example is a binary vector of length $n+1$
  - +1 column: *target value (i.e., c*)
- *association rule mining:* each transaction is a binary vector of length n
  - n: number of items

Problem

classical machine learning/data mining methods

- developed for **single** relational problem settings

many applications

- deal with graphs and/or
- require multiple relations

remark:

- graphs can be considered as (special) **relational** structures!

problem:

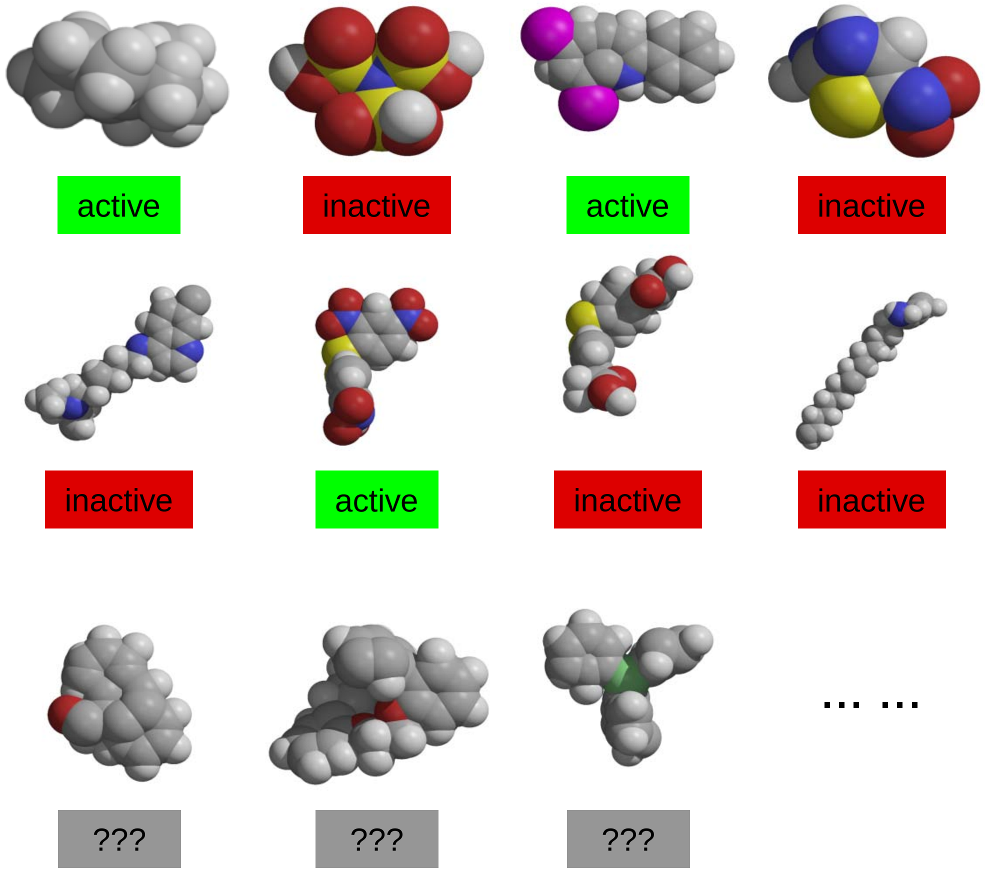
- **no** (natural) representation of graphs and (multi-)relational structures by a single table of **fixed width**

An Application Example: Virtual Screening in Drug Discovery

- select a limited number of candidate compounds from **millions** of database molecules that are most likely to possess a desired biological activity

training molecules

test molecules

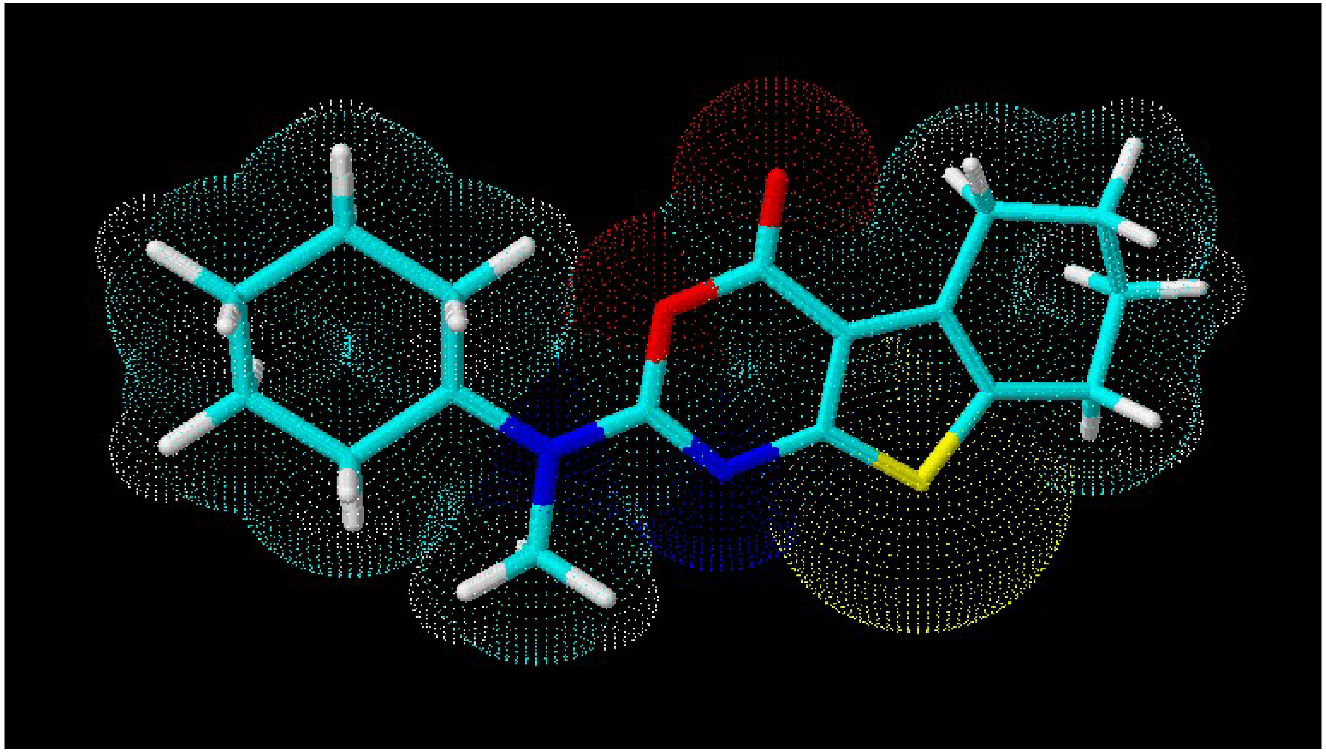


universitätbonn
RHEINISCHE FRIEDRICH-WILHELMS-UNIVERSITÄT

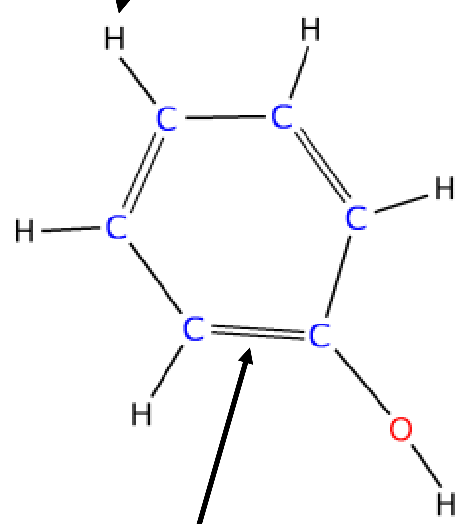
Fraunhofer IAIS

An Application Example: Virtual Screening in Drug Discovery

Molecules and their Molecular Graphs



vertex label



edge label
"double"

molecules give rise to **labeled undirected graphs**

This PhD Course

algorithmic aspects of local pattern mining in

- **single** table representations,
- **graphs** and **relational structures**

topics:

- the theory extraction problem
- itemset/association rule mining
- graph mining
- local/global pattern mining in relational structures