Data Mining and Knowledge Discovery

Part of Jožef Stefan IPS Programme - ICT3 and UL Programme - Statistics

2012 / 2013

Nada Lavrač

Jožef Stefan Institute

Ljubljana, Slovenia

Course Outline

I. Introduction

- Data Mining in a Nutshell
- Predictive and descriptive DM techniques
- Data Mining and KDD process
- DM standards, tools and visualization

  (Mladenić et al. Ch. 1 and 11)

II. Predictive DM Techniques

- Bayesian classifier (Kononenko Ch. 9.6)
- Decision Tree learning

  (Mitchell Ch. 3, Kononenko Ch. 9.1)
- Classification rule learning (Kononenko Ch. 9.2)
- Classifier Evaluation

  (Bramer Ch. 6)

III. Regression

(Kononenko Ch. 9.4)

IV. Descriptive DM

- Predictive vs. descriptive induction
- Subgroup discovery
- Association rule learning (Kononenko Ch. 9.3)
- Hierarchical clustering (Kononenko Ch. 12.3)

– V. Relational Data Mining

- RDM and Inductive Logic Programming (Dzeroski & Lavrac Ch. 3, Ch. 4)
- Propositionalization approaches
- Relational subgroup discovery

Introductory seminar lecture

X. JSI & Department of Knowledge Technologies

I. Introduction: First generation data mining

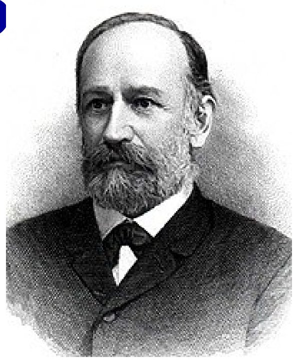
- Data Mining in a nutshell
- Predictive and descriptive DM techniques
- Data Mining and KDD process
- DM standards, tools and visualization

  (Mladenić et al. Ch. 1 and 11)

XX. Selected data mining techniques: Advanced subgroup discovery techniques and applications

XXX. Recent advances: Cross-context link discovery

Jožef Stefan Institute and IPS



- **Jožef Stefan Institute (JSI, founded in 1949)**
  - named after a distinguished physicist Jožef Stefan (1835-1893)

    $$j = \sigma T^4$$

  - leading national research organization in natural sciences and technology (~700 researchers and students)
- **JSI research areas**
  - information and communication technologies
  - chemistry, biochemistry & nanotechnology
  - physics, nuclear technology and safety
- **Jožef Stefan International Postgraduate School (IPS, founded in 2004)**
  - offers MSc and PhD programs (ICT, nanotechnology, ecotechnology)
  - research oriented, basic + management courses
  - in English

Jožef Stefan Institute
Department of Knowledge Technologies

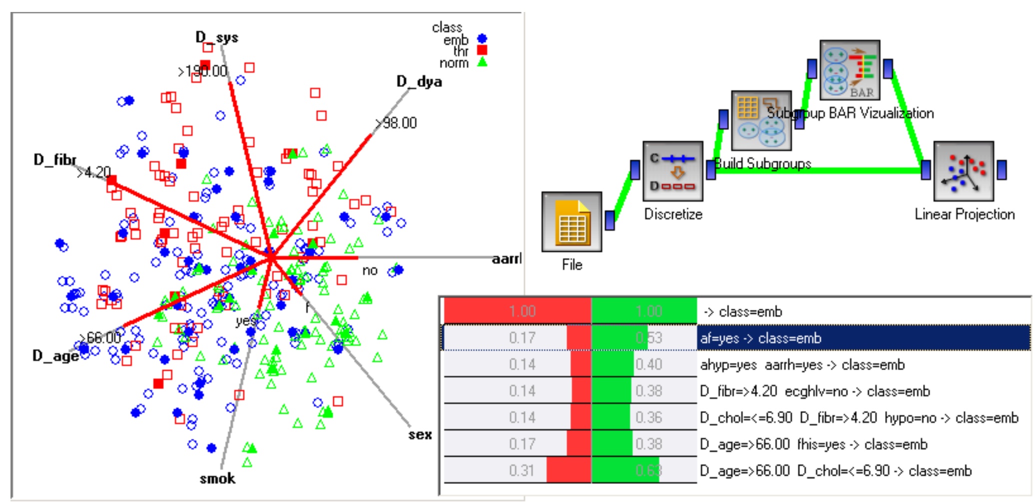
- **Head:** Nada Lavrač, **Staff:** 30 researchers, 10 students
- **Machine learning & Data mining**
  - ML (decision tree and rule learning, subgroup discovery, …)
  - Text and Web mining
  - Relational data mining - inductive logic programming
  - Equation discovery
- **Other research areas:**
  - Knowledge management
  - Decision support
  - Human language technologies
- **Applications:**
  - Medicine, Bioinformatics, Public Health
  - Ecology, Finance, …

Basic Data Mining Task

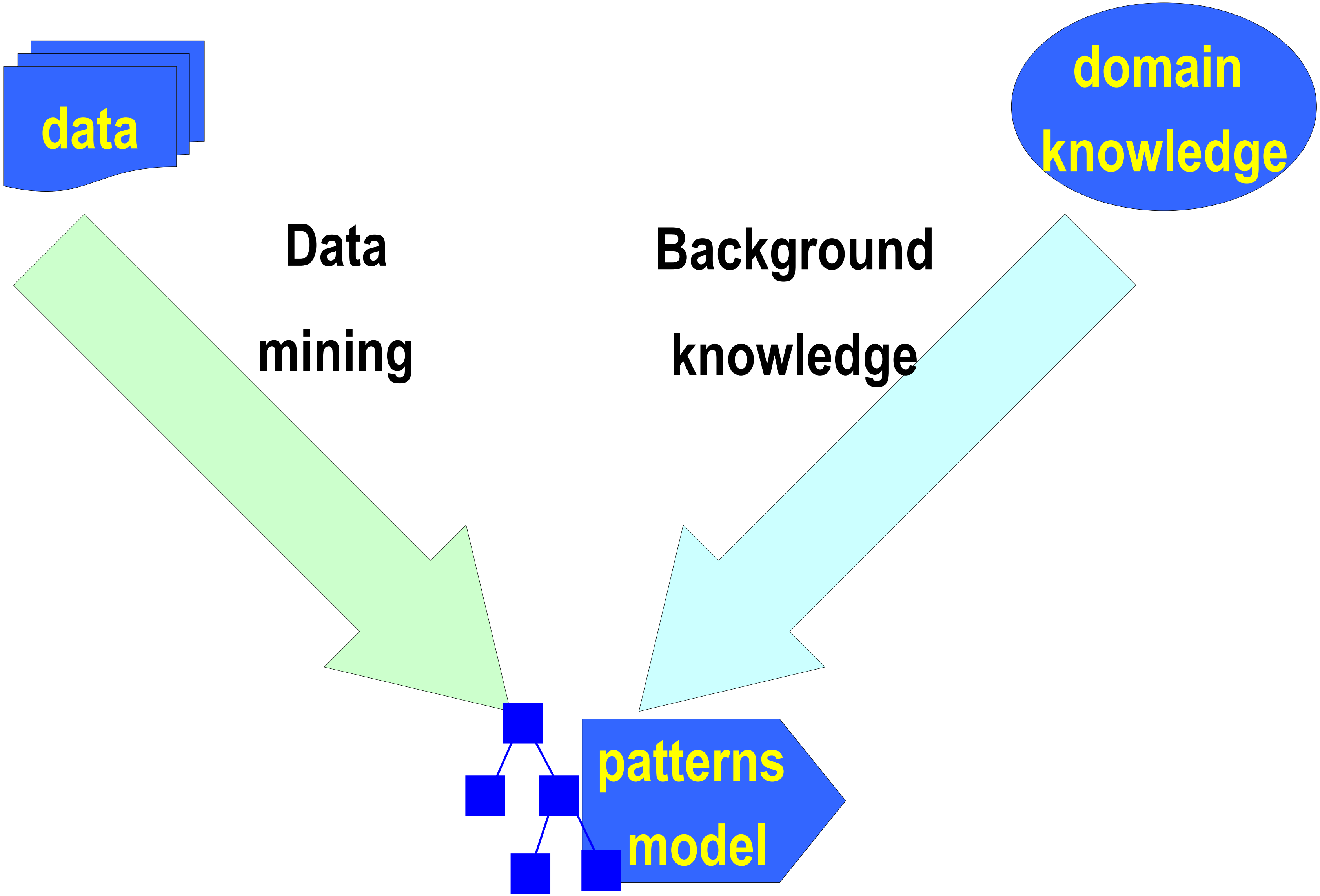
Input: transaction data table, relational database, text documents, Web pages
Goal: build a classification model, find interesting patterns in data, ...

Data Mining and Machine Learning

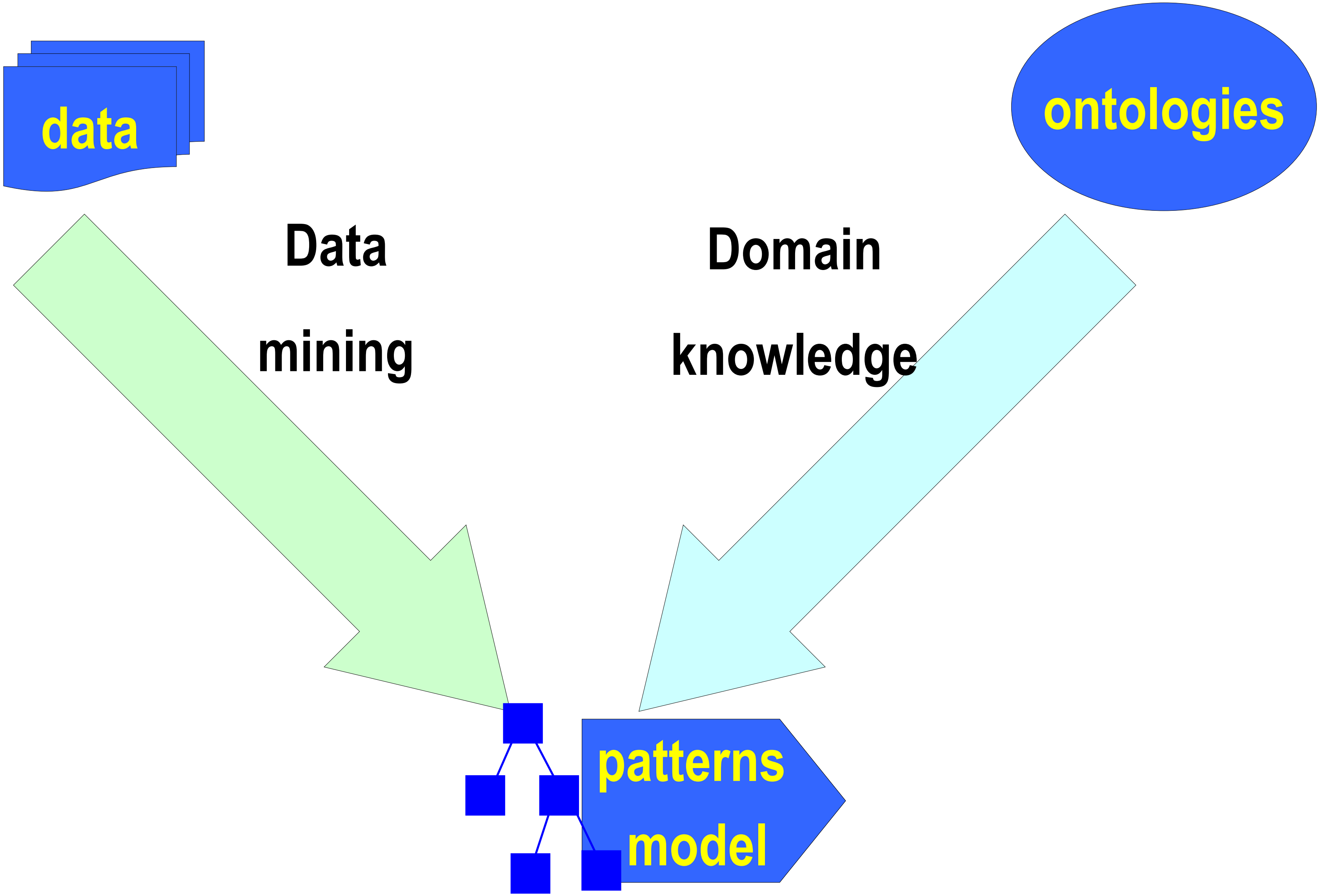
- Machine learning techniques
  - classification rule learning
  - subgroup discovery
  - relational data mining and ILP
  - equation discovery
  - inductive databases
- Data mining and decision support integration
- Data mining applications
  - medicine, health care
  - ecology, agriculture
  - knowledge management, virtual organizations



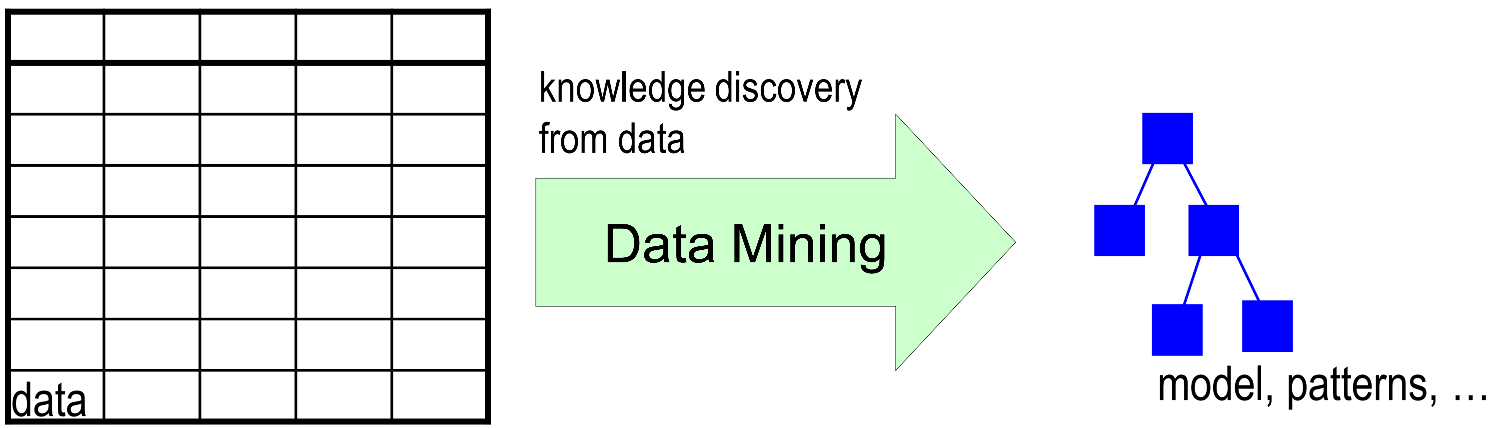
Relational data mining: domain knowledge = relational database



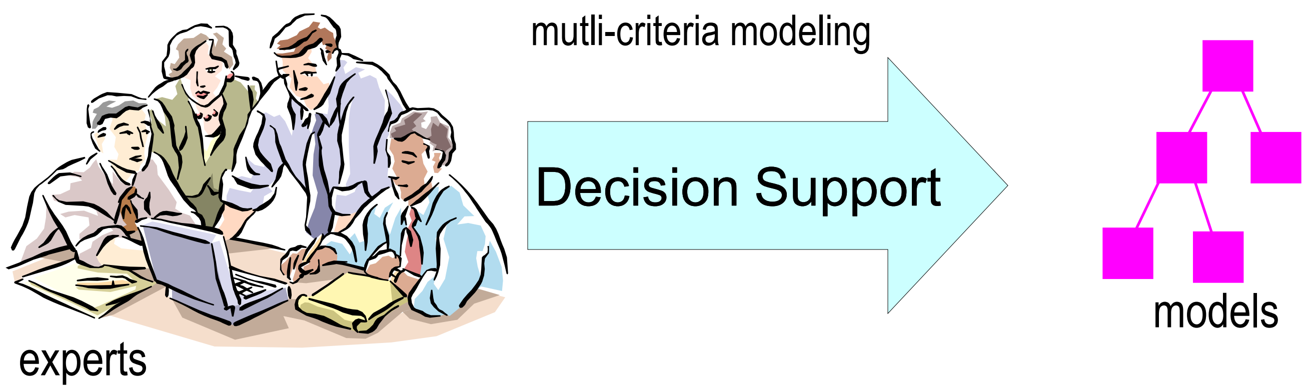
Semantic data mining: domain knowledge = ontologies



Basic DM and DS Tasks

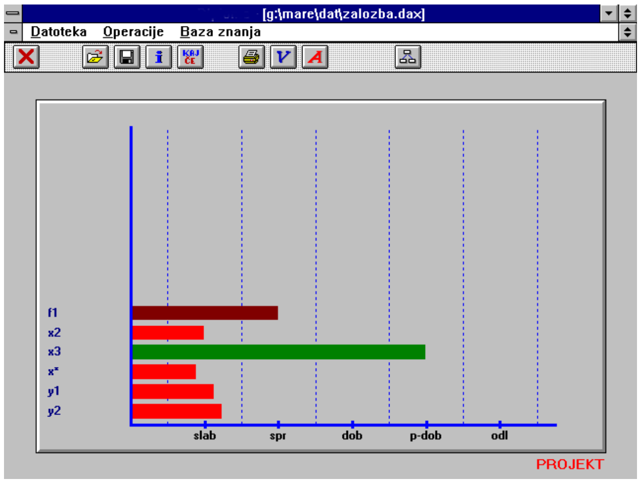


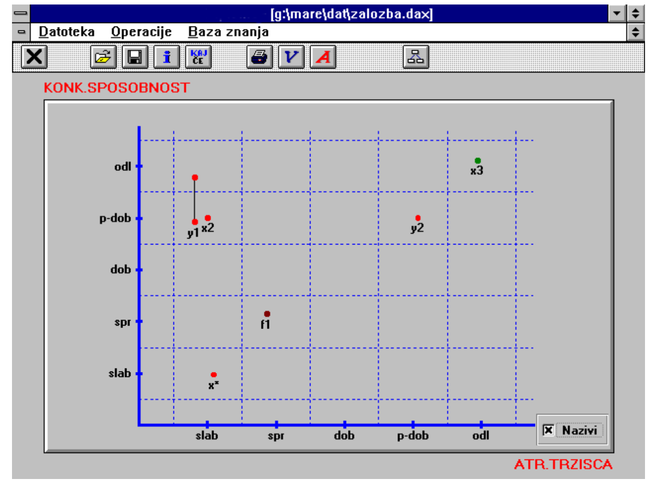
Input: transaction data table, relational database, text documents, Web pages
Goal: build a classification model, find interesting patterns in data, ...

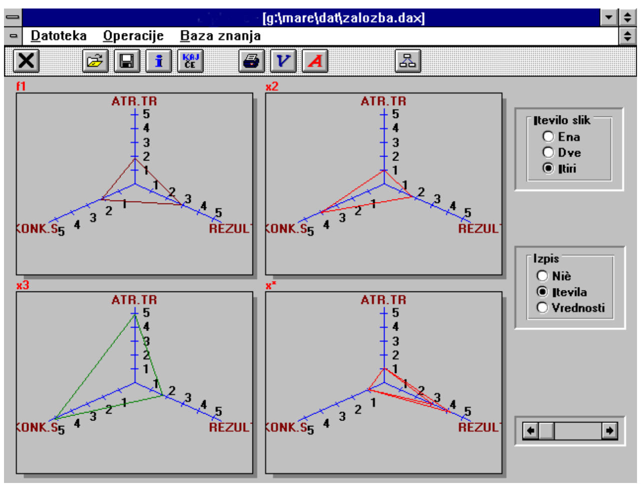


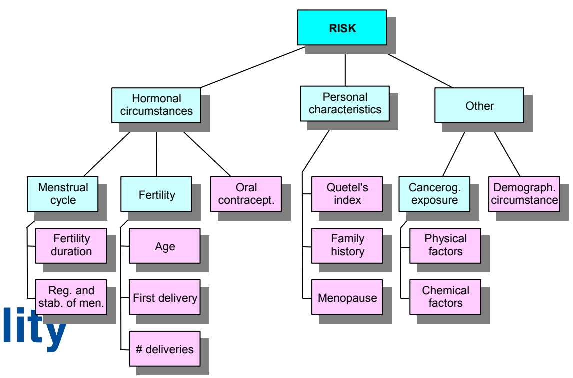
Input: expert knowledge about data and decision alternatives
Goal: construct decision support model – to support the evaluation and choice of best decision alternatives

Decision support tools: DEXi





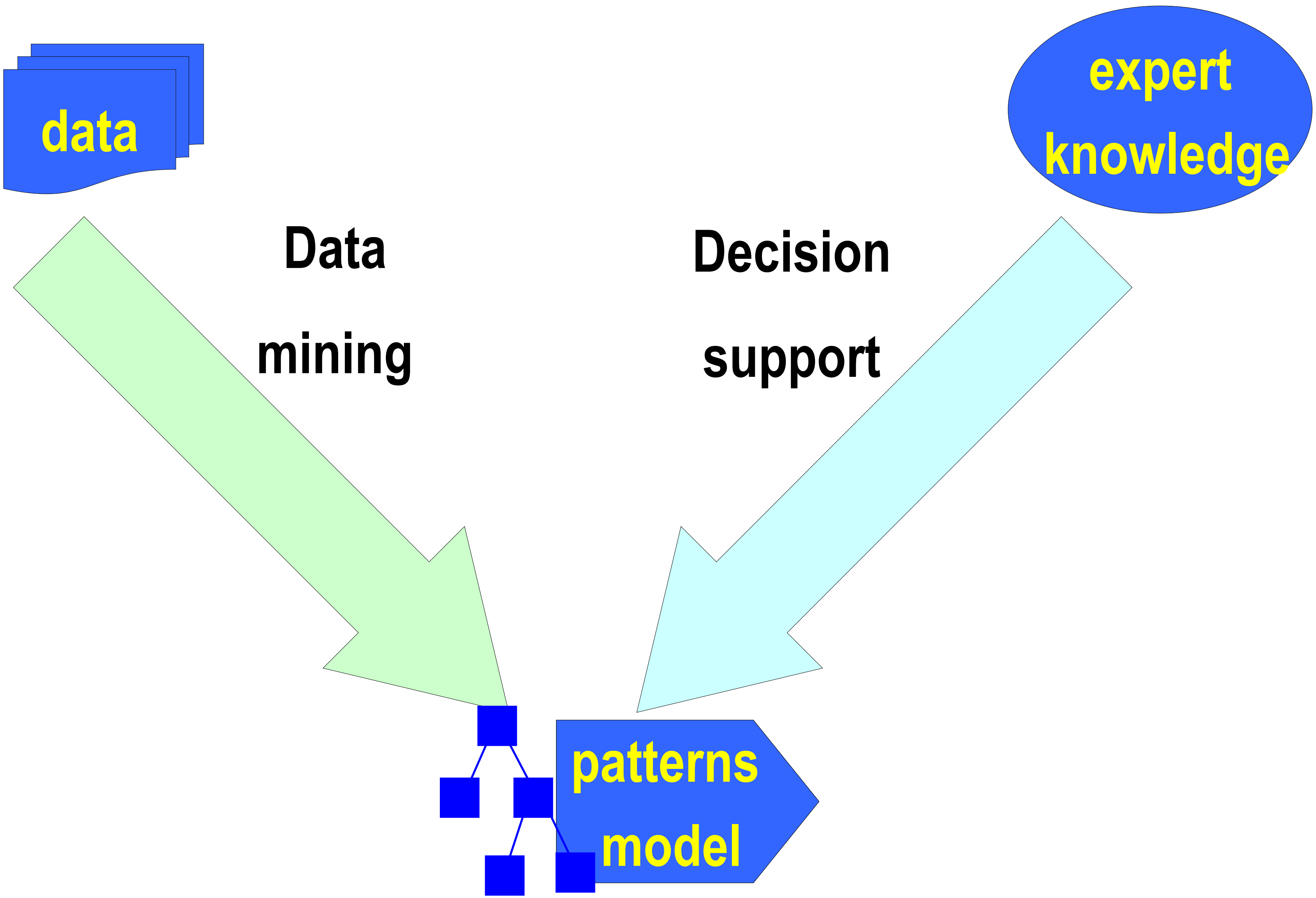




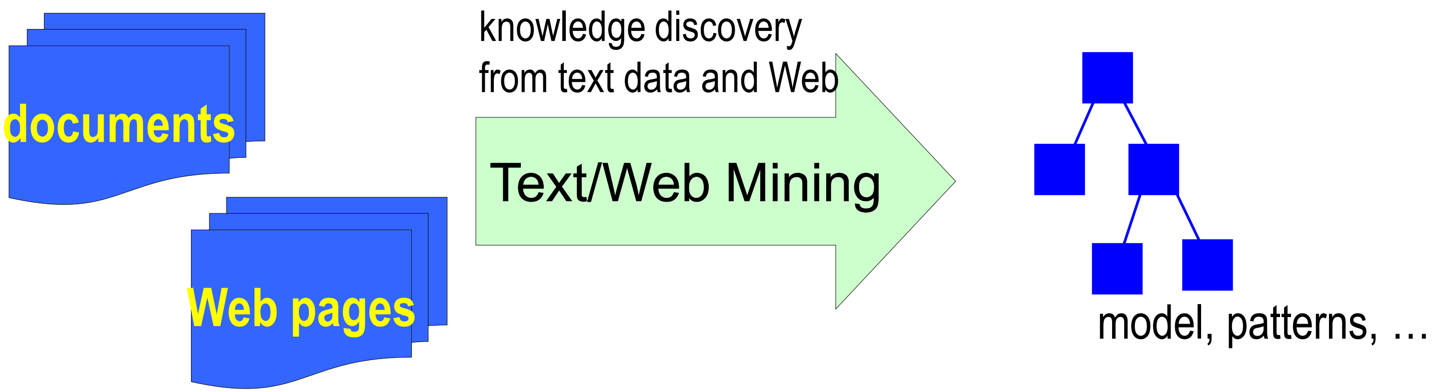
DEXi supports :

- *if-then* analysis
- analysis of stability
- Time analysis
- *how* explanation
- *why* explanation

DM and DS integration



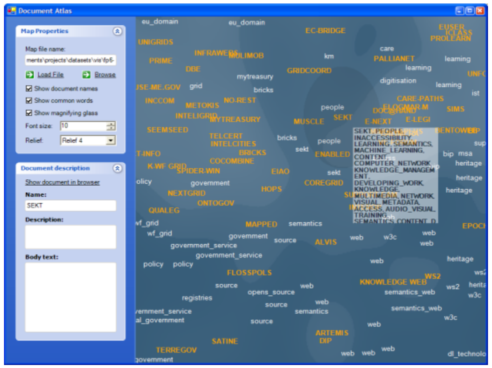
Basic Text and Web Mining Task



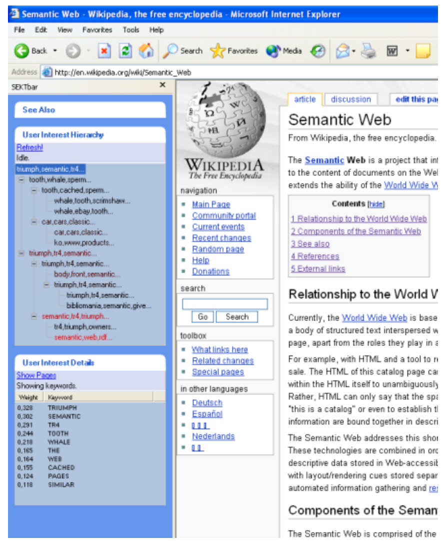
Input: text documents, Web pages
Goal: text categorization, user modeling, data visualization...

Text Mining (lectures by D. Mladenić)

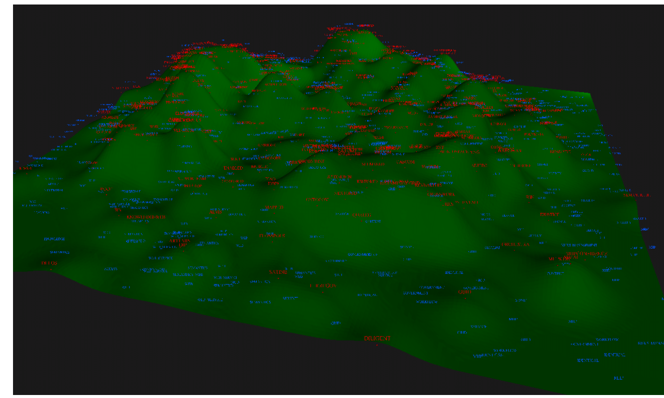
Document-Atlas

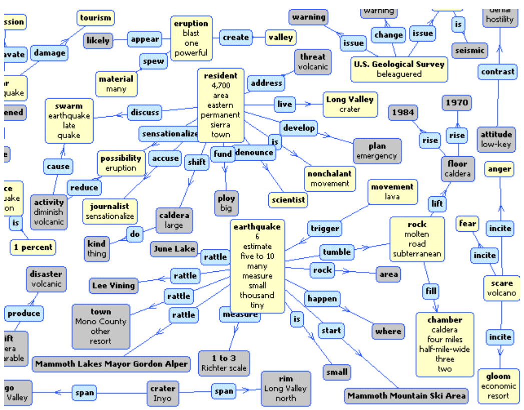


SEKTbar

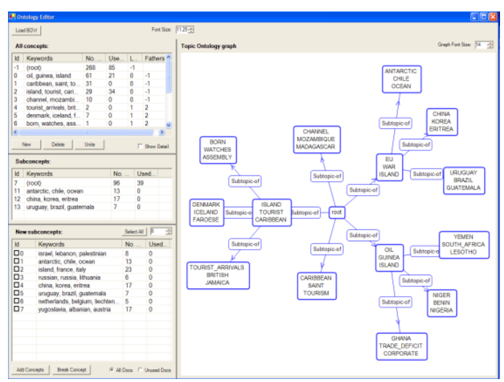


Content-Land

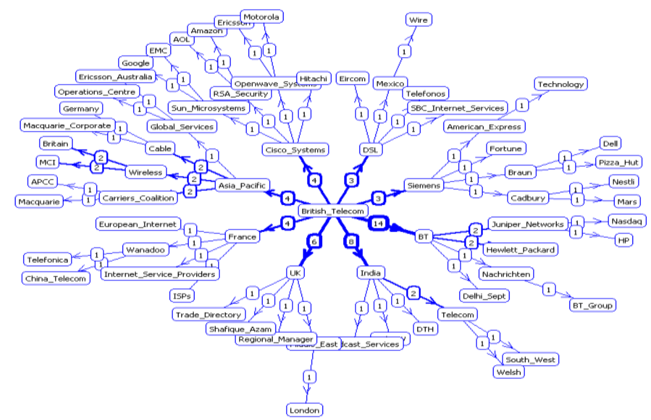




Semantic-Graphs

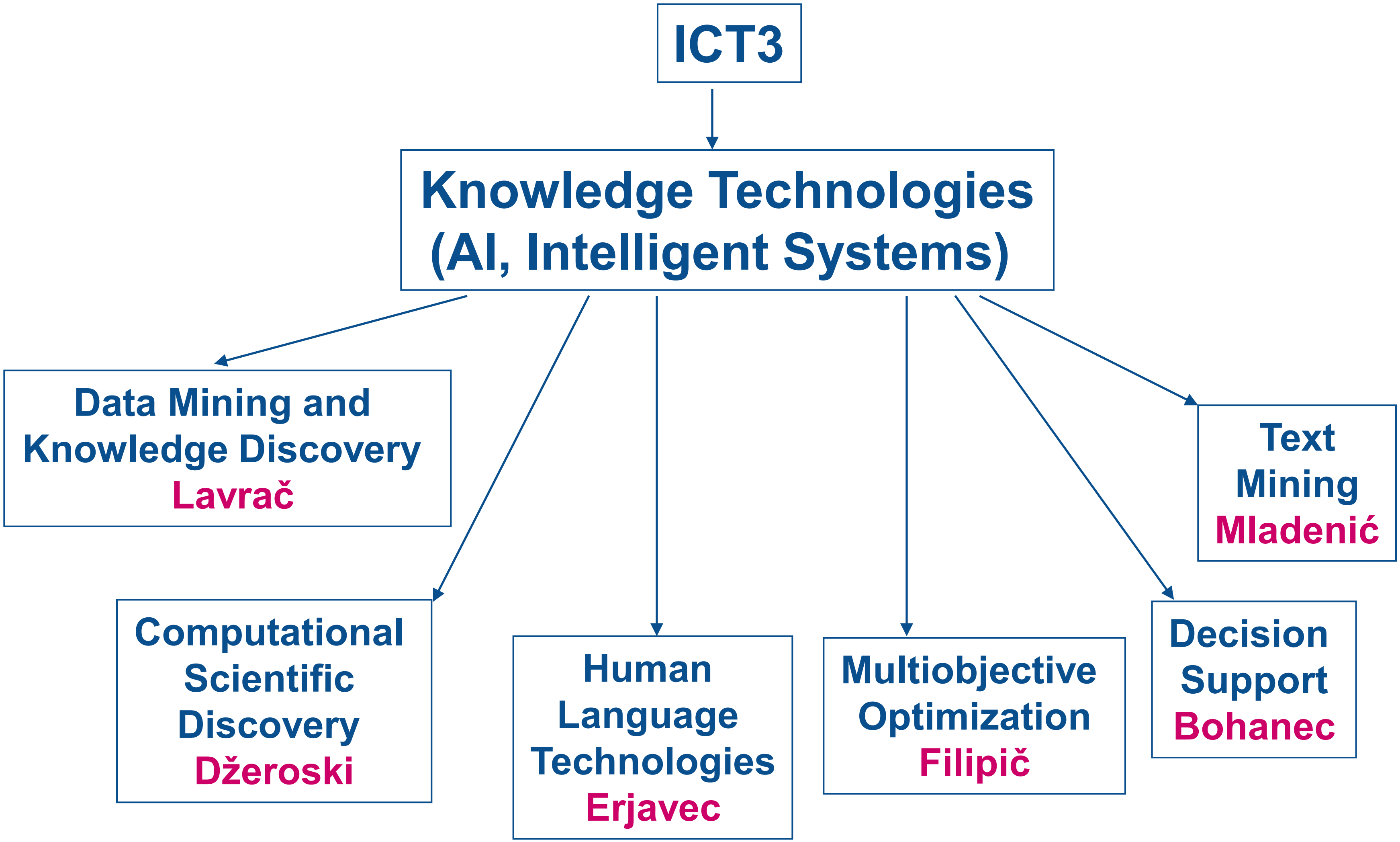


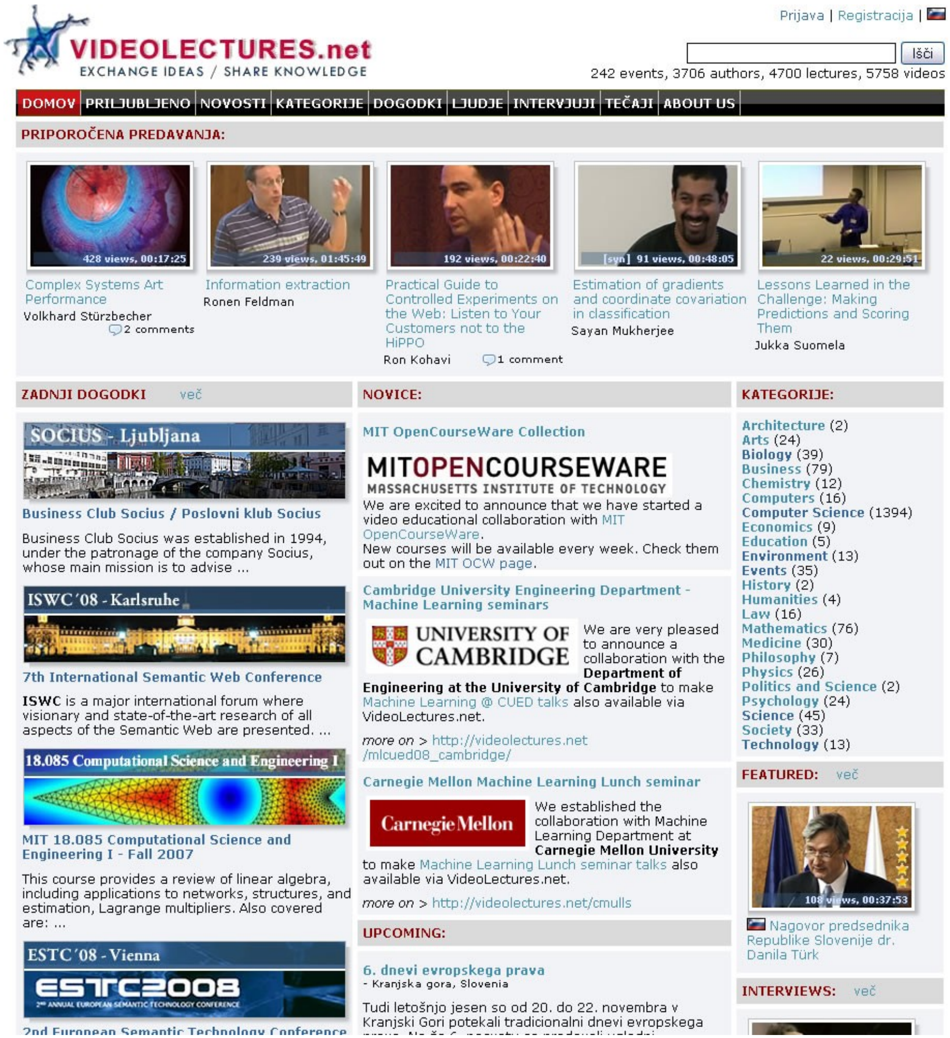
OntoGen



Contexter

Knowledge Technologies: Main research areas & IPS lectures



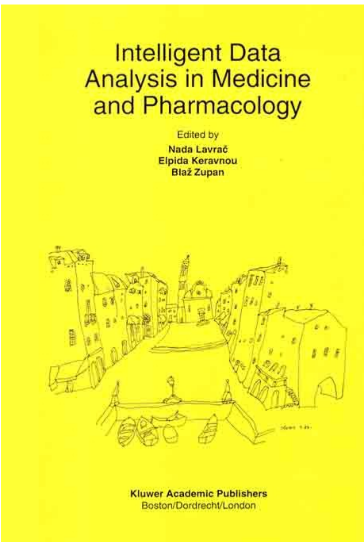


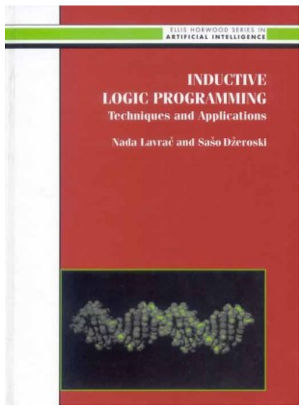
videolectures.net portal

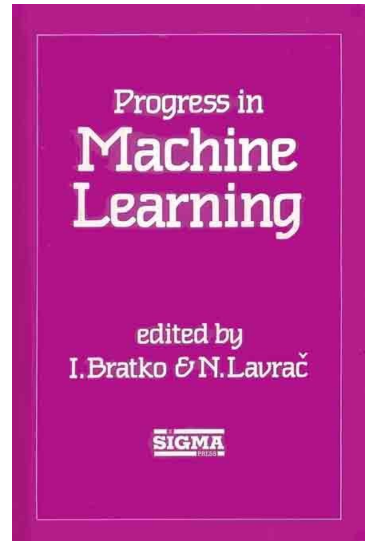
~ 10,000 lectures

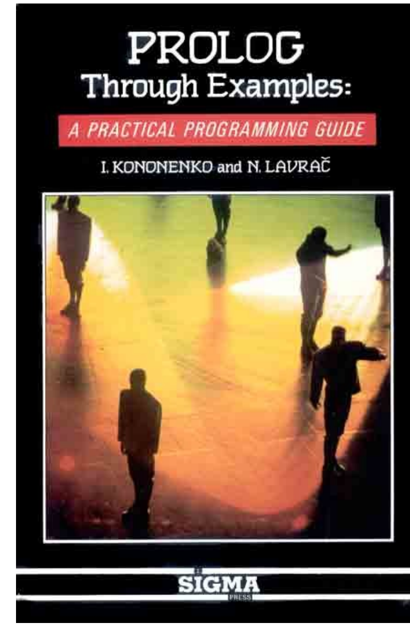
http//:videolectures.net

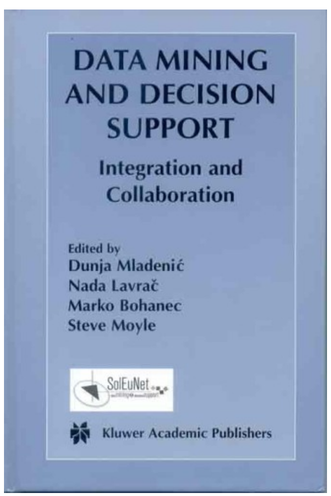
Selected Publications

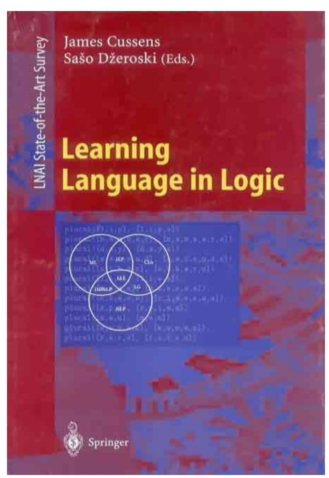


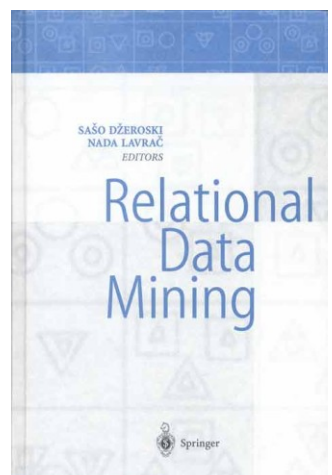


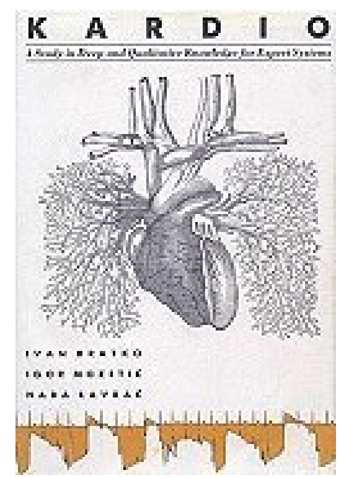












Introductory seminar lecture

X. JSI & Knowledge Technologies

I. Introduction

- Data Mining in a Nutshell
- Predictive and descriptive DM techniques
- Data Mining and the KDD process
- DM standards, tools and visualization

  (Mladenić et al. Ch. 1 and 11, Kononenko & Kukar Ch. 1)

XX. Selected data mining techniques: Advanced subgroup discovery techniques and applications

XXX. Recent advances: Cross-context link discovery

Part I. Introduction

- Data Mining in a Nutshell
- Predictive and descriptive DM techniques
- Data Mining and the KDD process
- DM standards, tools and visualization

What is DM

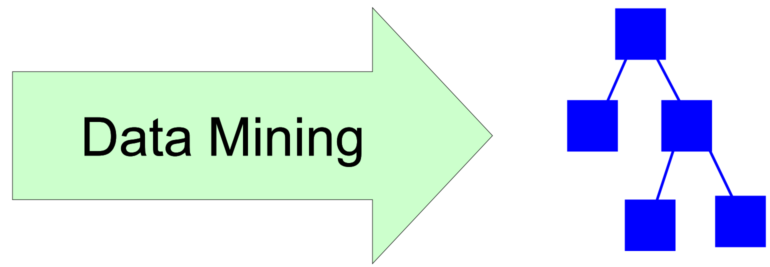
- Extraction of useful information from data: discovering relationships that have not previously been known
- The viewpoint in this course: Data Mining is the application of Machine Learning techniques to solve real-life data analysis problems

Data Mining in a Nutshell

Person	Age	Spect. presc.	Astigm.	Tear prod.	Lenses
O1	17	myope	no	reduced	NONE
O2	23	myope	no	normal	SOFT
O3	22	myope	yes	reduced	NONE
O4	27	myope	yes	normal	HARD
O5	19	hypermetrope	no	reduced	NONE
O6-O13	...	...	...	...	...
O14	35	hypermetrope	no	normal	SOFT
O15	43	hypermetrope	yes	reduced	NONE
O16	39	hypermetrope	yes	normal	NONE
O17	54	myope	no	reduced	NONE
O18	62	myope	no	normal	NONE
O19-O23	...	...	...	...	...
O24	56	hypermetrope	yes	normal	NONE

data

knowledge discovery
from data



model, patterns, …

Given: transaction data table, relational database, text documents, Web pages

Find: a classification model, a set of interesting patterns

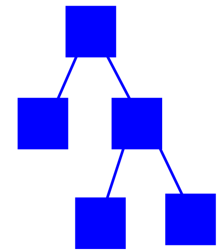
Data Mining in a Nutshell

Person	Age	Spect. presc.	Astigm.	Tear prod.	Lenses
O1	17	myope	no	reduced	NONE
O2	23	myope	no	normal	SOFT
O3	22	myope	yes	reduced	NONE
O4	27	myope	yes	normal	HARD
O5	19	hypermetrope	no	reduced	NONE
O6-O13	...	...	...	...	...
O14	35	hypermetrope	no	normal	SOFT
O15	43	hypermetrope	yes	reduced	NONE
O16	39	hypermetrope	yes	normal	NONE
O17	54	myope	no	reduced	NONE
O18	62	myope	no	normal	NONE
O19-O23	...	...	...	...	...
O24	56	hypermetrope	yes	normal	NONE

data

knowledge discovery
from data

Data Mining

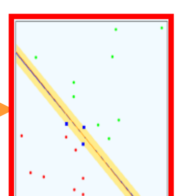


model, patterns, …

Given: transaction data table, relational database, text documents, Web pages

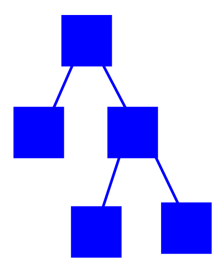
Find: a classification model, a set of interesting patterns

new unclassified instance



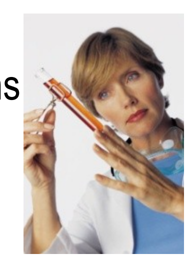
classified instance

black box classifier
no explanation



symbolic model
symbolic patterns

explanation

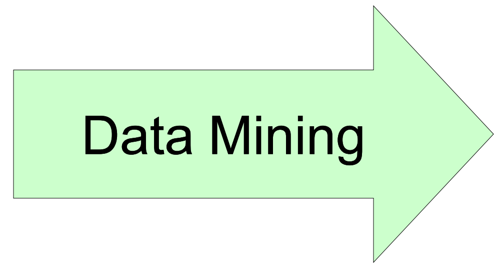


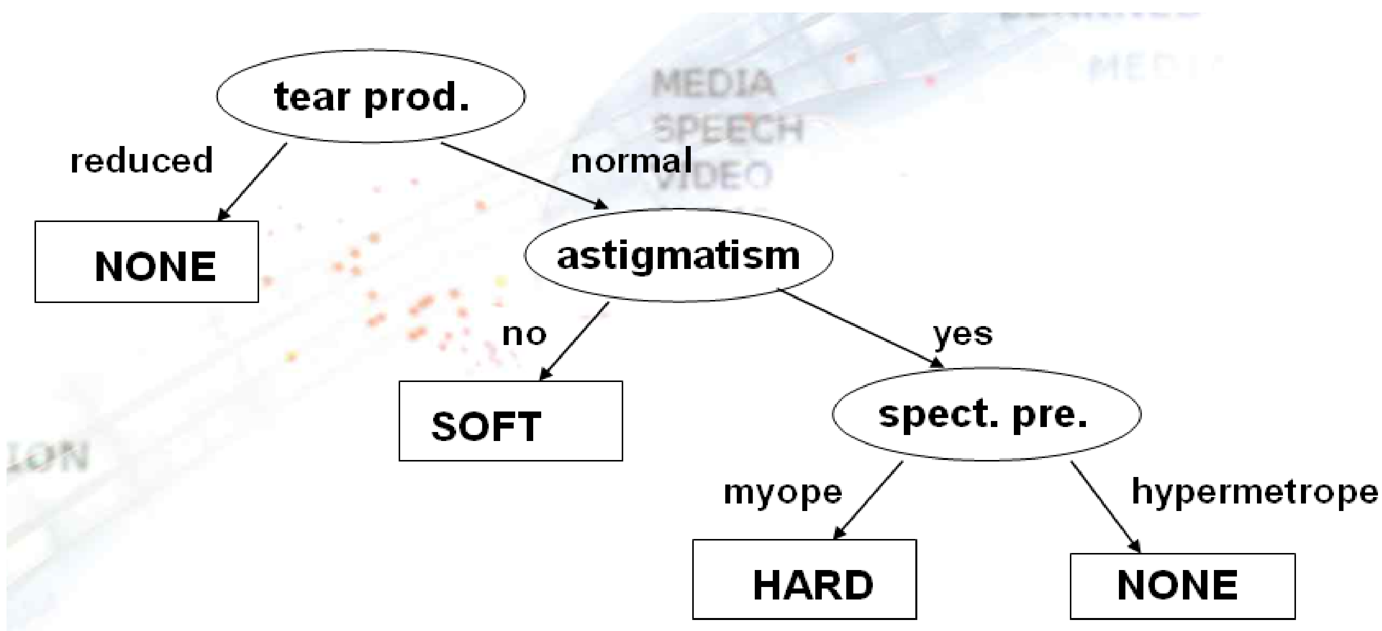
Simplified example: Learning a classification model from contact lens data

Person	Age	Spect. presc.	Astigm.	Tear prod.	Lenses
O1	17	myope	no	reduced	NONE
O2	23	myope	no	normal	SOFT
O3	22	myope	yes	reduced	NONE
O4	27	myope	yes	normal	HARD
O5	19	hypermetrope	no	reduced	NONE
O6-O13	...	...	...	...	...
O14	35	hypermetrope	no	normal	SOFT
O15	43	hypermetrope	yes	reduced	NONE
O16	39	hypermetrope	yes	normal	NONE
O17	54	myope	no	reduced	NONE
O18	62	myope	no	normal	NONE
O19-O23	...	...	...	...	...
O24	56	hypermetrope	yes	normal	NONE

Simplified example: Learning a classification model from contact lens data

Person	Age	Spect. presc.	Astigm.	Tear prod.	Lenses
O1	young	myope	no	reduced	NONE
O2	young	myope	no	normal	SOFT
O3	young	myope	yes	reduced	NONE
O4	young	myope	yes	normal	HARD
O5	young	hypermetrope	no	reduced	NONE
O6-O13	...	...	...	...	...
O14	pre-presbyo	hypermetrope	no	normal	SOFT
O15	pre-presbyo	hypermetrope	yes	reduced	NONE
O16	pre-presbyo	hypermetrope	yes	normal	NONE
O17	presbyopic	myope	no	reduced	NONE
O18	presbyopic	myope	no	normal	NONE
O19-O23	...	...	...	...	...
O24	presbyopic	hypermetrope	yes	normal	NONE





Task reformulation: Binary Class Values

Person	Age	Spect. presc.	Astigm.	Tear prod.	Lenses
O1	17	myope	no	reduced	NO
O2	23	myope	no	normal	YES
O3	22	myope	yes	reduced	NO
O4	27	myope	yes	normal	YES
O5	19	hypermetrope	no	reduced	NO
O6-O13	...	...	...	...	...
O14	35	hypermetrope	no	normal	YES
O15	43	hypermetrope	yes	reduced	NO
O16	39	hypermetrope	yes	normal	NO
O17	54	myope	no	reduced	NO
O18	62	myope	no	normal	NO
O19-O23	...	...	...	...	...
O24	56	hypermetrope	yes	normal	NO

Binary classes (positive vs. negative examples of Target class)
- for Concept learning – classification and class description
  - for Subgroup discovery – exploring patterns characterizing groups of instances of target class

Learning from Numeric Class Data

Person	Age	Spect. presc.	Astigm.	Tear prod.	LensPrice
O1	17	myope	no	reduced	0
O2	23	myope	no	normal	8
O3	22	myope	yes	reduced	0
O4	27	myope	yes	normal	5
O5	19	hypermetrope	no	reduced	0
O6-O13	...	...	...	...	...
O14	35	hypermetrope	no	normal	5
O15	43	hypermetrope	yes	reduced	0
O16	39	hypermetrope	yes	normal	0
O17	54	myope	no	reduced	0
O18	62	myope	no	normal	0
O19-O23	...	...	...	...	...
O24	56	hypermetrope	yes	normal	0

Numeric class values – regression analysis

Learning from Unlabeled Data

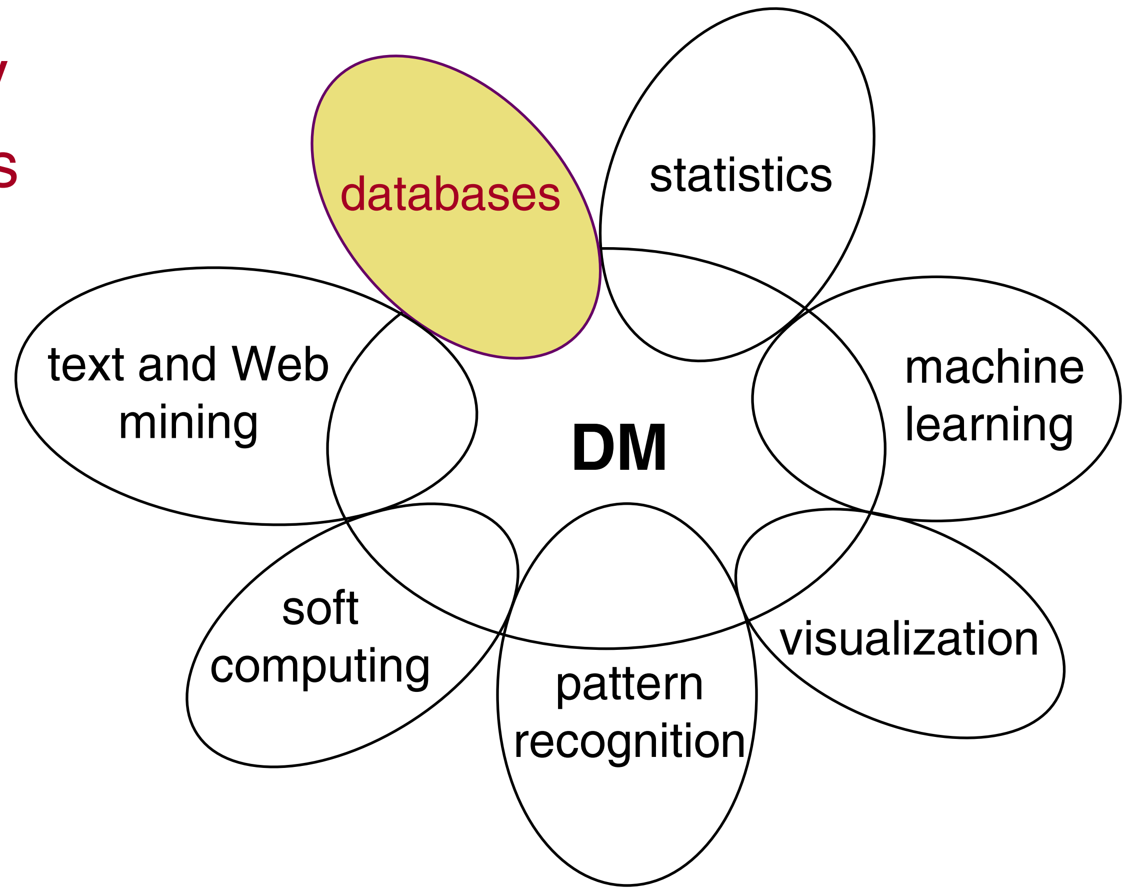
Person	Age	Spect. presc.	Astigm.	Tear prod.	~~Lenses~~
O1	17	myope	no	reduced	~~NONE~~
O2	23	myope	no	normal	~~SOFT~~
O3	22	myope	yes	reduced	~~NONE~~
O4	27	myope	yes	normal	~~HARD~~
O5	19	hypermetrope	no	reduced	~~NONE~~
O6-O13	...	...	...	...	~~...~~
O14	35	hypermetrope	no	normal	~~SOFT~~
O15	43	hypermetrope	yes	reduced	~~NONE~~
O16	39	hypermetrope	yes	normal	~~NONE~~
O17	54	myope	no	reduced	~~NONE~~
O18	62	myope	no	normal	~~NONE~~
O19-O23	...	...	...	...	~~...~~
O24	56	hypermetrope	yes	normal	~~NONE~~

Unlabeled data - clustering: grouping of similar instances
- association rule learning

Data Mining: Related areas

Database technology and data warehouses

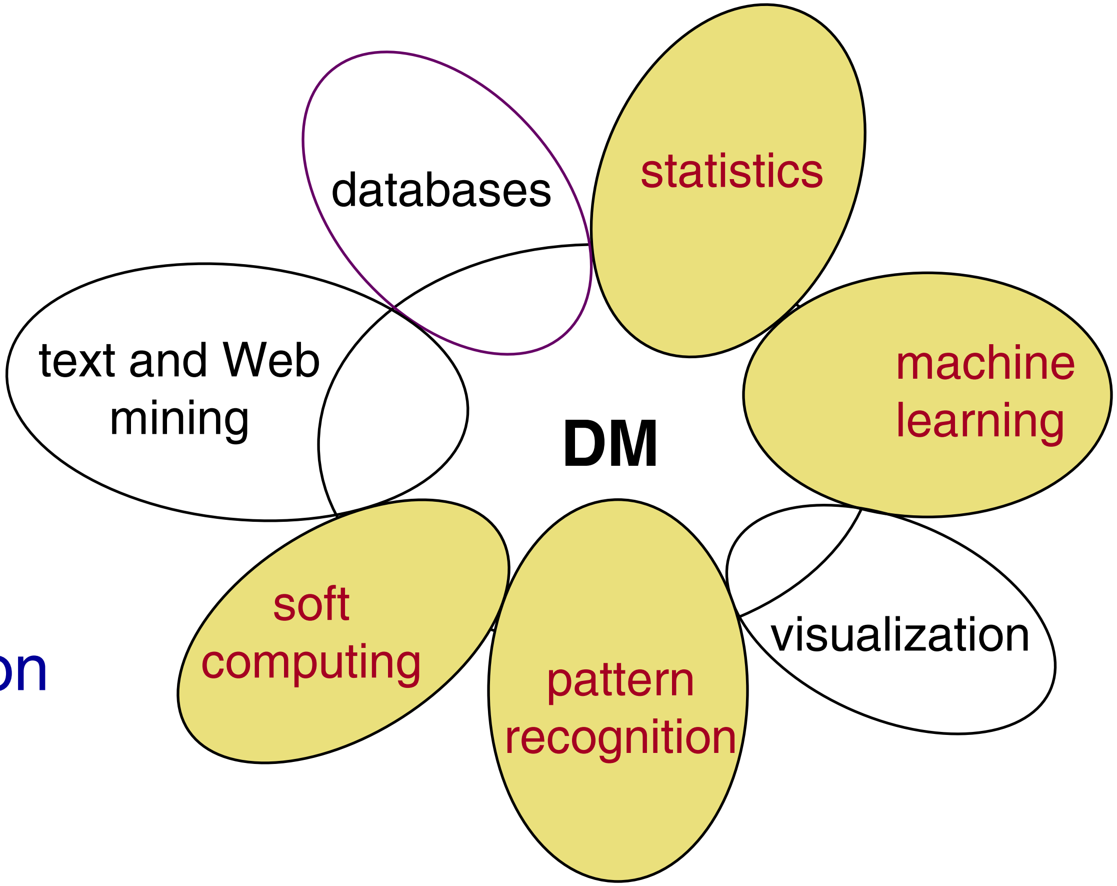
- efficient storage, access and manipulation of data



Related areas

Statistics,
machine learning,
pattern recognition
and soft computing*

- classification techniques and techniques for knowledge extraction from data

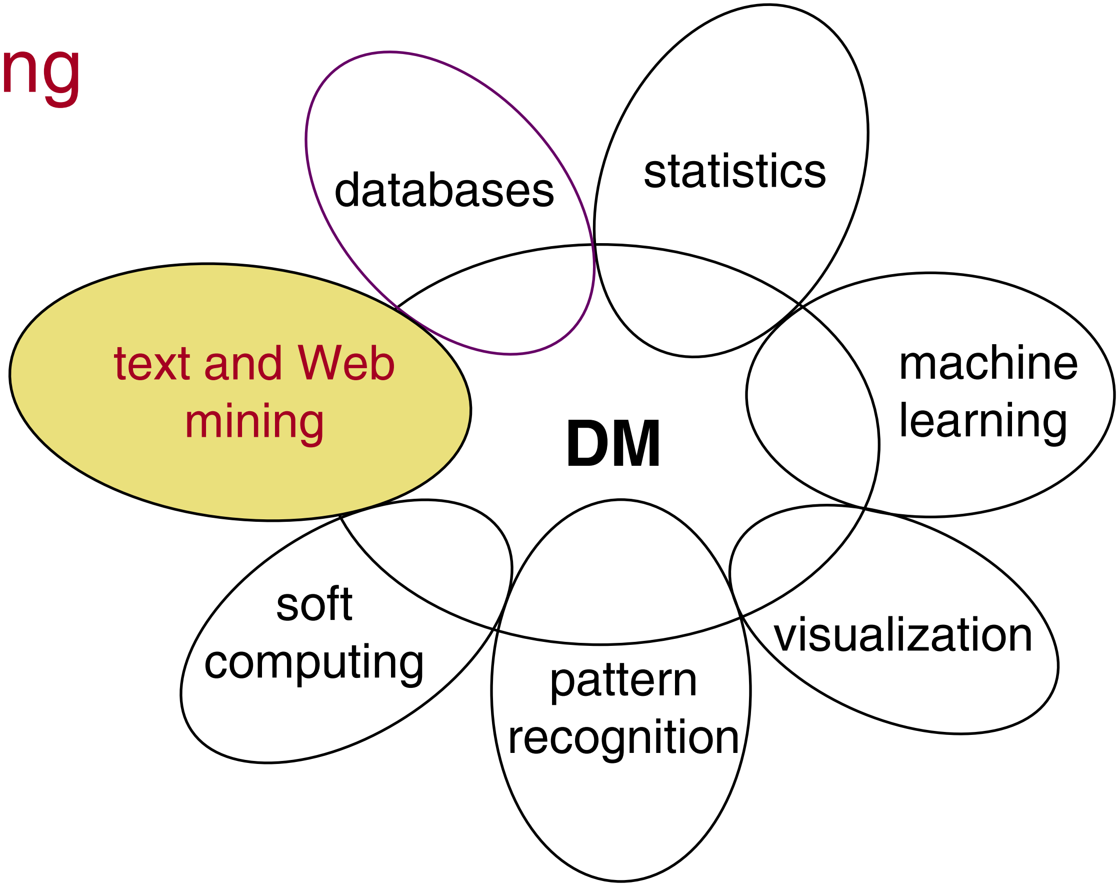


*neural networks, fuzzy logic, genetic algorithms, probabilistic reasoning

Related areas

Text and Web mining

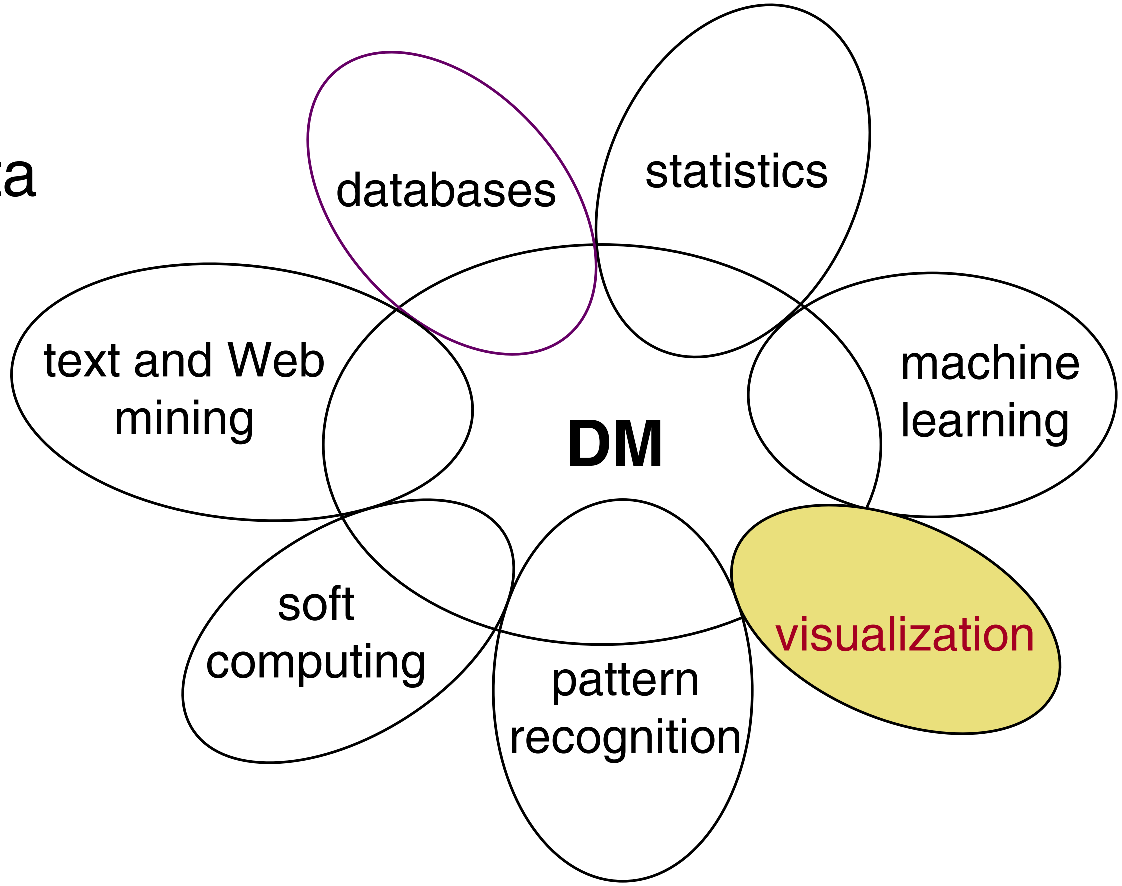
- Web page analysis
- text categorization
- acquisition, filtering and structuring of textual information
- natural language processing



Related areas

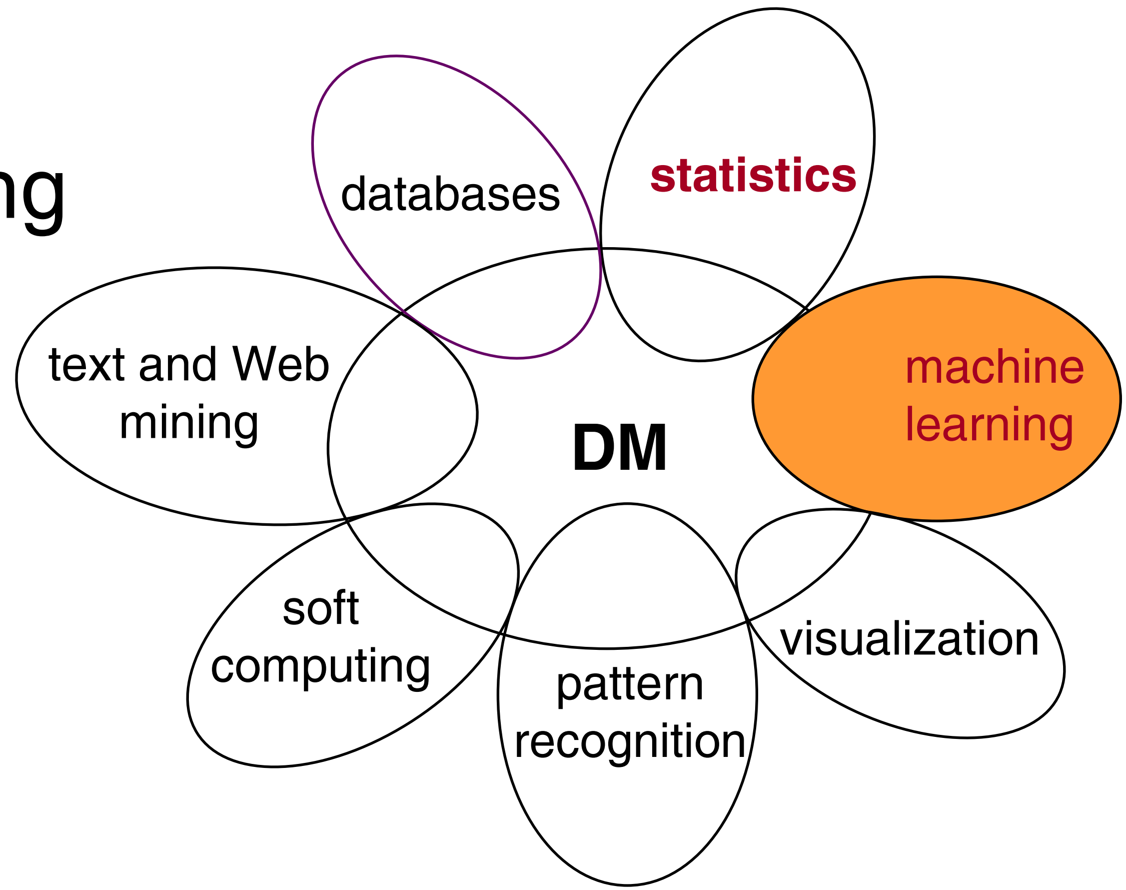
Visualization

- visualization of data and discovered knowledge



Point of view in this course

Knowledge discovery using machine learning methods



Data Mining, ML and Statistics

- All three areas have a long tradition of developing inductive techniques for data analysis.
  - reasoning from properties of a data sample to properties of a population
- **DM vs. ML - Viewpoint in this course**:
  - Data Mining is the application of Machine Learning techniques to hard real-life data analysis problems

Data Mining, ML and Statistics

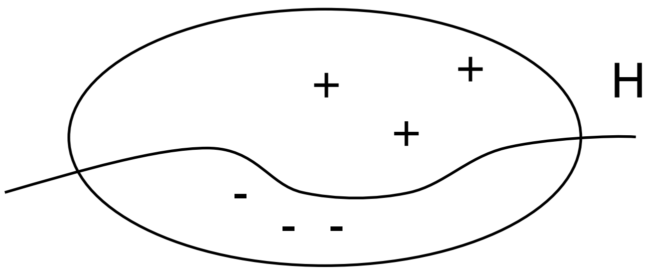
- All three areas have a long tradition of developing inductive techniques for data analysis.
  - reasoning from properties of a data sample to properties of a population
- **DM vs. Statistics:**
  - **Statistics**
    - Hypothesis testing when certain theoretical expectations about the data distribution, independence, random sampling, sample size, etc. are satisfied
    - Main approach: best fitting all the available data
  - **Data mining**
    - Automated construction of understandable patterns, and structured models
    - Main approach: structuring the data space, heuristic search for decision trees, rules, … covering (parts of) the data space

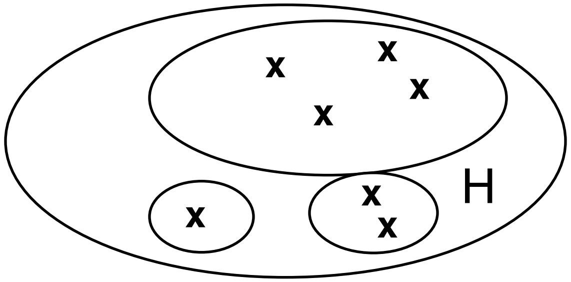
Part I. Introduction

- Data Mining in a Nutshell
- Predictive and descriptive DM techniques
- Data Mining and the KDD process
- DM standards, tools and visualization

Types of DM tasks

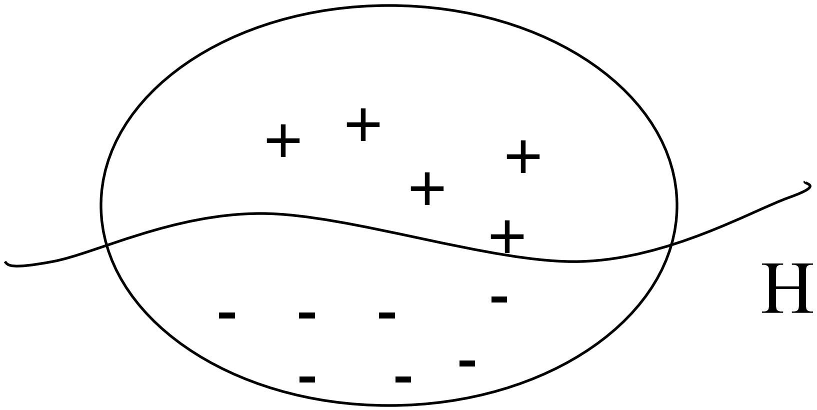
- **Predictive DM:**
  - Classification (learning of rules, decision trees, ...)
  - Prediction and estimation (regression)
  - Predictive relational DM (ILP)
- **Descriptive DM:**
  - description and summarization
  - dependency analysis (association rule learning)
  - discovery of properties and constraints
  - segmentation (clustering)
  - subgroup discovery



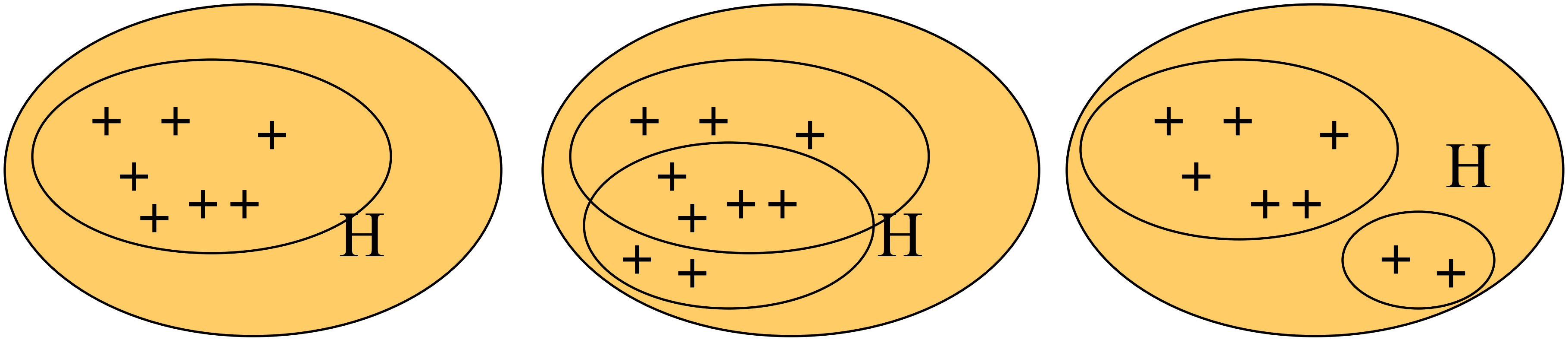


Predictive vs. descriptive DM

Predictive DM



Descriptive DM



Predictive vs. descriptive DM

- **Predictive DM:** Inducing classifiers for solving classification and prediction tasks,
  - Classification rule learning, Decision tree learning, ...
  - Bayesian classifier, ANN, SVM, ...
  - Data analysis through hypothesis generation and testing
- **Descriptive DM:** Discovering interesting regularities in the data, uncovering patterns, ... for solving KDD tasks
  - Symbolic clustering, Association rule learning, Subgroup discovery, ...
  - Exploratory data analysis

Predictive DM formulated as a machine learning task:

- Given a set of labeled **training examples** (n-tuples of attribute values, labeled by class name)

	A1	A2	A3	Class
example1	$v_{1,1}$	$v_{1,2}$	$v_{1,3}$	C_1
example2	$v_{2,1}$	$v_{2,2}$	$v_{2,3}$	C_2

. .

- By performing generalization from examples (induction) find a **hypothesis** (classification rules, decision tree, …) which explains the training examples, e.g. rules of the form:

$$(A_i = v_{i,k}) \ \& \ (A_j = v_{j,l}) \ \& \ \ldots \Rightarrow \text{Class} = C_n$$

Predictive DM - Classification

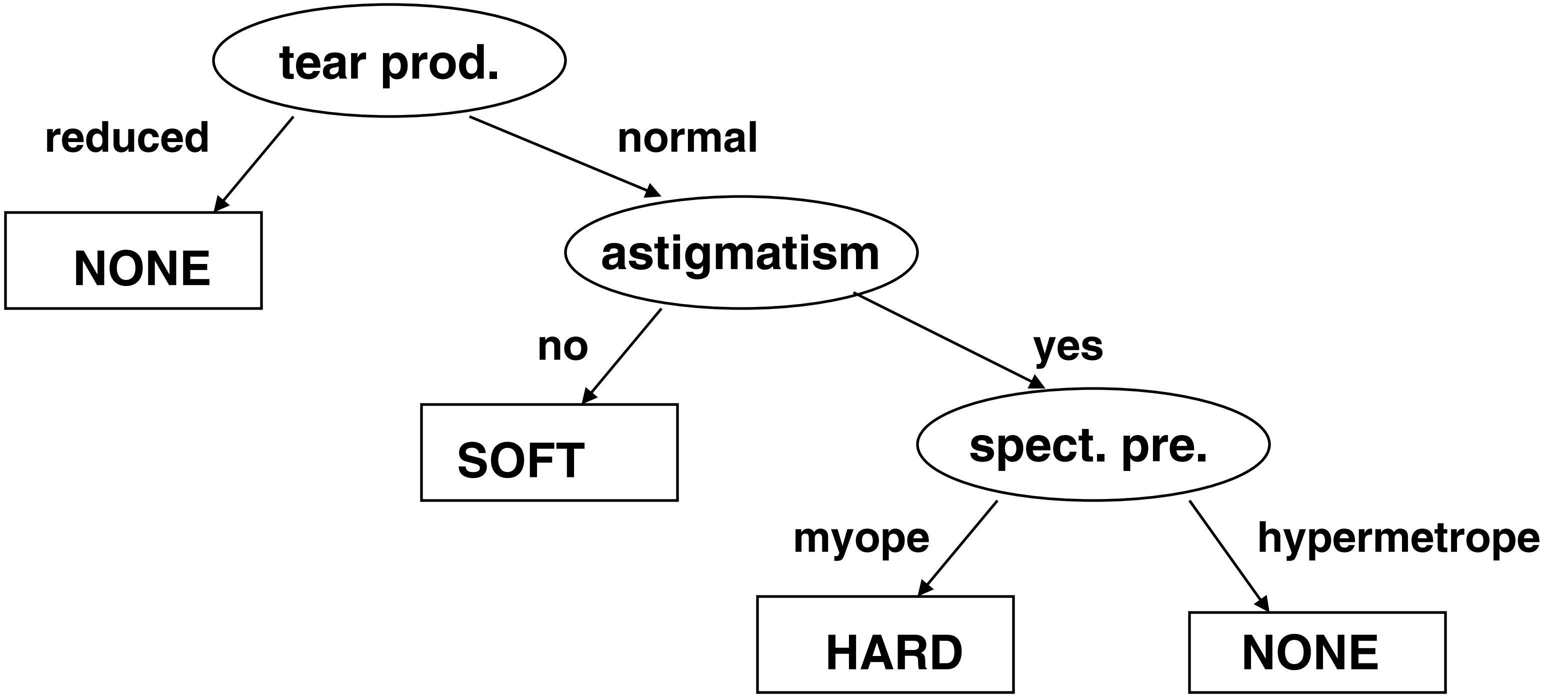
- data are objects, characterized with attributes - they belong to different classes (discrete labels)
- given objects described with attribute values, induce a model to predict different classes
- decision trees, if-then rules, discriminant analysis, ...

Data mining example
Input: Contact lens data

Person	Age	Spect. presc.	Astigm.	Tear prod.	Lenses
O1	young	myope	no	reduced	NONE
O2	young	myope	no	normal	SOFT
O3	young	myope	yes	reduced	NONE
O4	young	myope	yes	normal	HARD
O5	young	hypermetrope	no	reduced	NONE
O6-O13	...	...	...	...	...
O14	pre-presbyo	hypermetrope	no	normal	SOFT
O15	pre-presbyo	hypermetrope	yes	reduced	NONE
O16	pre-presbyo	hypermetrope	yes	normal	NONE
O17	presbyopic	myope	no	reduced	NONE
O18	presbyopic	myope	no	normal	NONE
O19-O23	...	...	...	...	...
O24	presbyopic	hypermetrope	yes	normal	NONE

Contact lens data: Decision tree

Type of task: prediction and classification
Hypothesis language: decision trees
(nodes: attributes, arcs: values of attributes, leaves: classes)



Contact lens data: Classification rules

Type of task: prediction and classification
Hypothesis language: rules X ➔ C, if X then C
X conjunction of attribute values, C class

tear production=reduced → lenses=NONE
tear production=normal & astigmatism=yes & spect. pre.=hypermetrope → lenses=NONE
tear production=normal & astigmatism=no → lenses=SOFT
tear production=normal & astigmatism=yes & spect. pre.=myope → lenses=HARD
DEFAULT lenses=NONE

Task reformulation: Concept learning problem (positive vs. negative examples of Target class)

Person	Age	Spect. presc.	Astigm.	Tear prod.	Lenses
O1	young	myope	no	reduced	NO
O2	young	myope	no	normal	YES
O3	young	myope	yes	reduced	NO
O4	young	myope	yes	normal	YES
O5	young	hypermetrope	no	reduced	NO
O6-O13	...	...	...	...	...
O14	pre-presbyo	hypermetrope	no	normal	YES
O15	pre-presbyo	hypermetrope	yes	reduced	NO
O16	pre-presbyo	hypermetrope	yes	normal	NO
O17	presbyopic	myope	no	reduced	NO
O18	presbyopic	myope	no	normal	NO
O19-O23	...	...	...	...	...
O24	presbyopic	hypermetrope	yes	normal	NO

Contact lens data: Classification rules in concept learning

Type of task: prediction and classification
Hypothesis language: rules X ➔ C, if X then C
X conjunction of attribute values, C target class

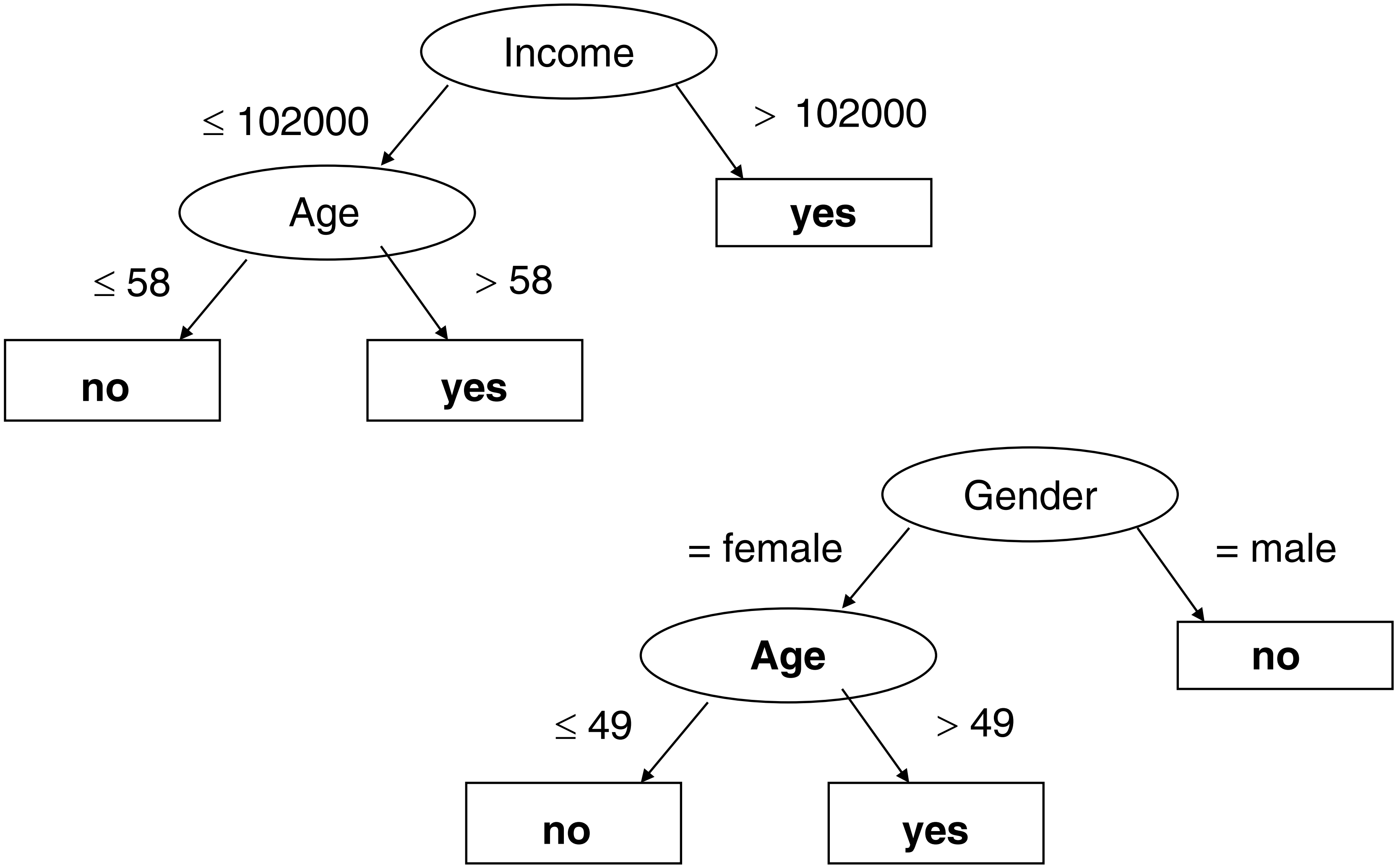
Target class: yes

tear production=normal & astigmatism=no → lenses=YES
tear production=normal & astigmatism=yes & spect. pre.=myope → lenses=YES
else NO

Illustrative example: Customer data

Customer	Gender	Age	Income	Spent	BigSpender
c1	male	30	214000	18800	yes
c2	female	19	139000	15100	yes
c3	male	55	50000	12400	no
c4	female	48	26000	8600	no
c5	male	63	191000	28100	yes
O6-O13	...	...	...	...	...
c14	female	61	95000	18100	yes
c15	male	56	44000	12000	no
c16	male	36	102000	13800	no
c17	female	57	215000	29300	yes
c18	male	33	67000	9700	no
c19	female	26	95000	11000	no
c20	female	55	214000	28800	yes

Customer data: Decision trees



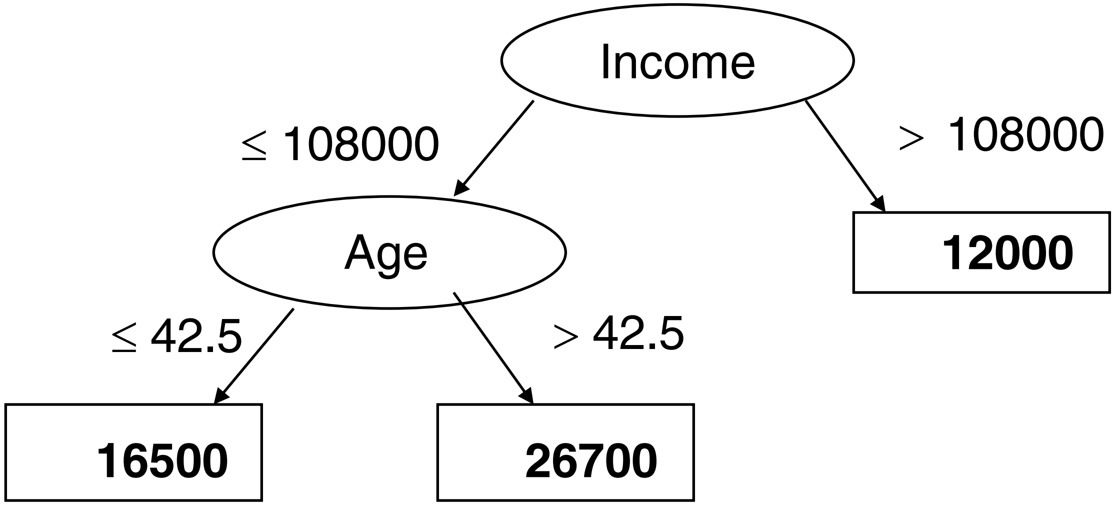
Predictive DM - Estimation

- often referred to as regression
- data are objects, characterized with attributes (discrete or continuous), classes of objects are continuous (numeric)
- given objects described with attribute values, induce a model to predict the numeric class value
- regression trees, linear and logistic regression, ANN, kNN, ...

Estimation/regression example: Customer data

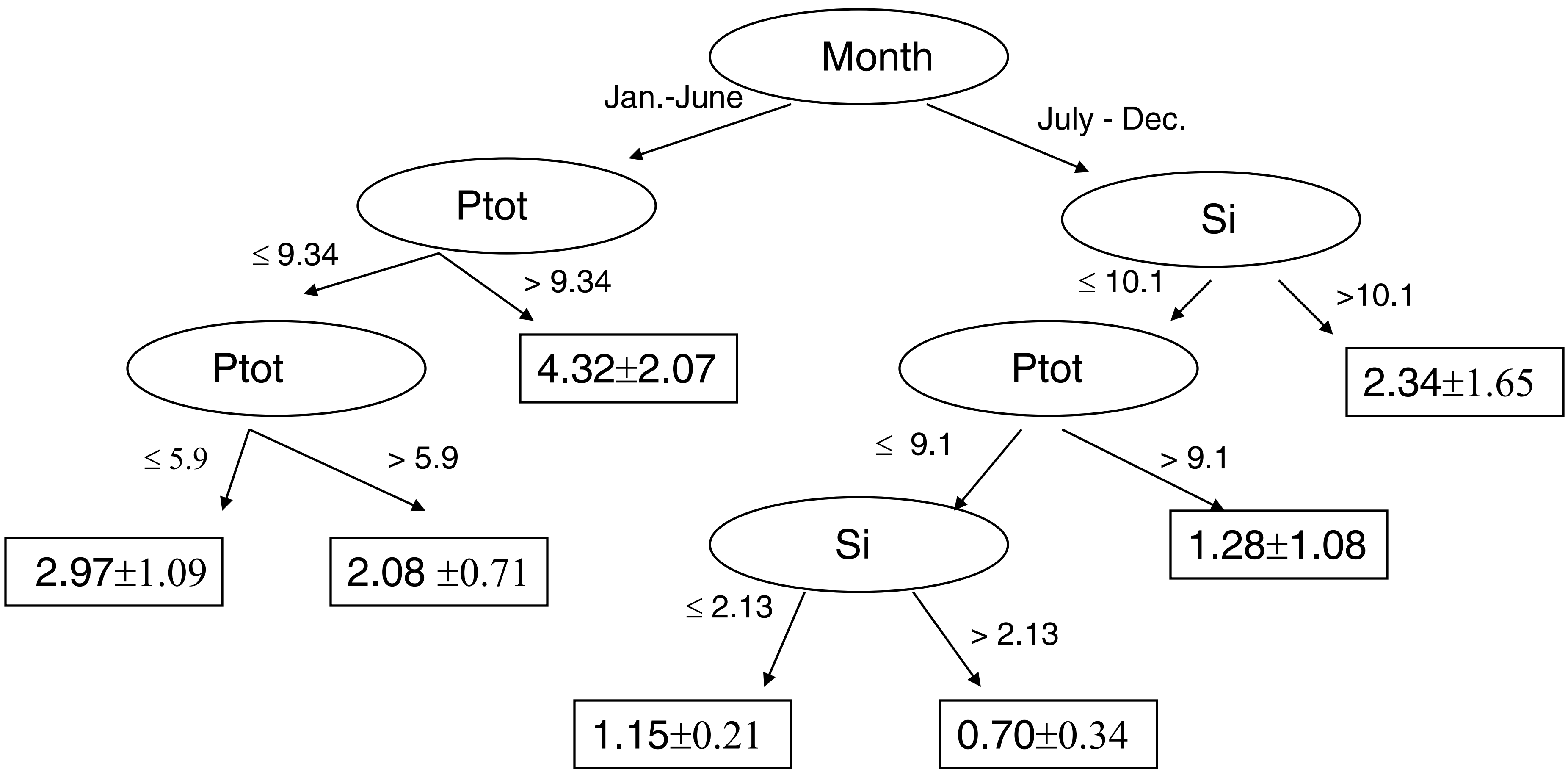
Customer	Gender	Age	Income	Spent	
c1	male	30	214000	18800	
c2	female	19	139000	15100	
c3	male	55	50000	12400	
c4	female	48	26000	8600	
c5	male	63	191000	28100	
O6-O13	...	...	...	...	
c14	female	61	95000	18100	
c15	male	56	44000	12000	
c16	male	36	102000	13800	
c17	female	57	215000	29300	
c18	male	33	67000	9700	
c19	female	26	95000	11000	
c20	female	55	214000	28800	

Customer data: regression tree



In the nodes one usually has Predicted value +- st. deviation

Predicting algal biomass: regression tree



Descriptive DM: Subgroup discovery example - Customer data

Customer	Gender	Age	Income	Spent	BigSpender
c1	male	30	214000	18800	yes
c2	female	19	139000	15100	yes
c3	male	55	50000	12400	no
c4	female	48	26000	8600	no
c5	male	63	191000	28100	yes
O6-O13	...	...	...	...	...
c14	female	61	95000	18100	yes
c15	male	56	44000	12000	no
c16	male	36	102000	13800	no
c17	female	57	215000	29300	yes
c18	male	33	67000	9700	no
c19	female	26	95000	11000	no
c20	female	55	214000	28800	yes

Customer data: Subgroup discovery

Type of task: description (pattern discovery)

Hypothesis language: rules **X ➔ Y,** if X then Y

X is conjunctions of items, Y is target class

Age > 52 & Sex = male ➔ BigSpender = no

Age > 52 & Sex = male & Income ≤ 73250
➔ BigSpender = no

Descriptive DM: Clustering and association rule learning example - Customer data

Customer	Gender	Age	Income	Spent	~~BigSpender~~
c1	male	30	214000	18800	~~yes~~
c2	female	19	139000	15100	~~yes~~
c3	male	55	50000	12400	~~no~~
c4	female	48	26000	8600	~~no~~
c5	male	63	191000	28100	~~yes~~
O6-O13	...	...	...	...	~~...~~
c14	female	61	95000	18100	~~yes~~
c15	male	56	44000	12000	~~no~~
c16	male	36	102000	13800	~~no~~
c17	female	57	215000	29300	~~yes~~
c18	male	33	67000	9700	~~no~~
c19	female	26	95000	11000	~~no~~
c20	female	55	214000	28800	~~yes~~

Descriptive DM: Association rule learning example - Customer data

Customer	Gender	Age	Income	Spent	BigSpender
c1	male	30	214000	18800	yes
c2	female	19	139000	15100	yes
c3	male	55	50000	12400	no
c4	female	48	26000	8600	no
c5	male	63	191000	28100	yes
O6-O13	...	...	...	...	...
c14	female	61	95000	18100	yes
c15	male	56	44000	12000	no
c16	male	36	102000	13800	no
c17	female	57	215000	29300	yes
c18	male	33	67000	9700	no
c19	female	26	95000	11000	no
c20	female	55	214000	28800	yes

Customer data: Association rules

Type of task: description (pattern discovery)

Hypothesis language: rules **X → Y,** if X then Y

X, Y conjunctions of items

1. Age > 52 & BigSpender = no → Sex = male
2. Age > 52 & BigSpender = no → Sex = male & Income ≤ 73250
3. Sex = male & Age > 52 & Income ≤ 73250 → BigSpender = no

Predictive vs. descriptive DM: Summary from a rule learning perspective

- **Predictive DM:** Induces **rulesets** acting as classifiers for solving classification and prediction tasks
- **Descriptive DM:** Discovers **individual rules** describing interesting regularities in the data

- **Therefore:** Different goals, different heuristics, different evaluation criteria

Relational Data Mining (Inductive Logic Programming) in a Nutshell

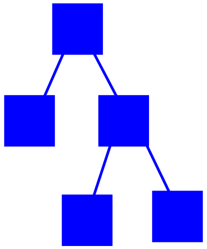
customer							
ID	Zip	Sex	So St	In come	A ge	Cl ub	Re sp
...	...	...	...	...	...	...	...
3478	34677	m	si	60-70	32	me	nr
3479	43666	f	ma	80-90	45	nm	re
...	...	...	...	...	...	...	...

order				
Customer ID	Order ID	Store ID	Delivery Mode	Paymt Mode
...	...	...	...	...
3478	2140267	12	regular	cash
3478	3446778	12	express	check
3478	4728386	17	regular	check
3479	3233444	17	express	credit
3479	3475886	12	regular	credit
...	...	...	...	...

store			
Store ID	Size	Type	Location
...	...	...	...
12	small	franchise	city
17	large	indep	rural
...	...	...	...

Relational representation of customers, orders and stores.

knowledge discovery
from data

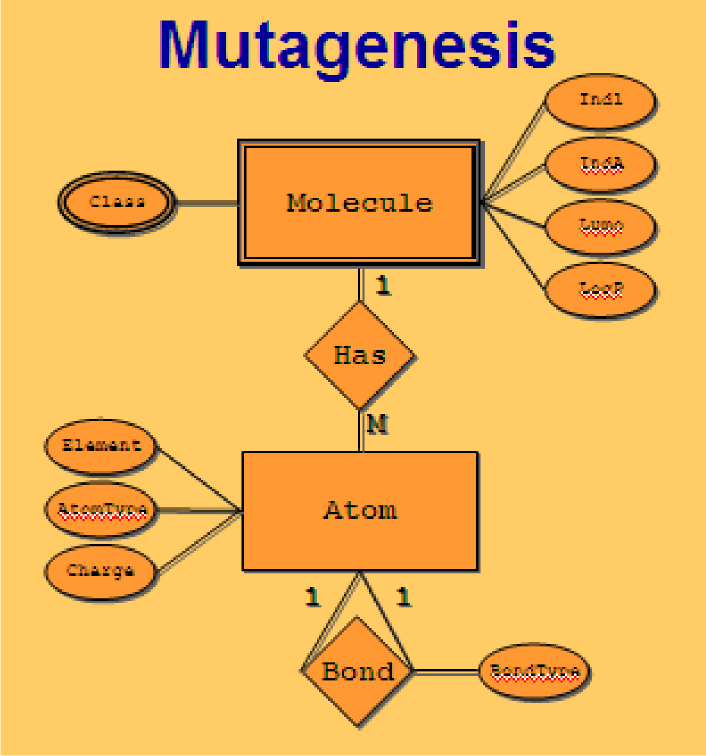


model, patterns, …

Given: a relational database, a set of tables. sets of logical facts, a graph, …
Find: a classification model, a set of interesting patterns

Relational Data Mining (ILP)

- Learning from multiple tables
- Complex relational problems:
  - temporal data: time series in medicine, trafic control, ...
  - structured data: representation of molecules and their properties in protein engineering, biochemistry, ...



Relational Data Mining (ILP)

customer							
ID	Zip	Sex	SoSt	Income	Age	Club	Resp
...	...	...	...	...	...	...	...
3478	34677	m	si	60-70	32	me	nr
3479	43666	f	ma	80-90	45	nm	re
...	...	...	...	...	...	...	...

order				
Customer ID	Order ID	Store ID	Delivery Mode	Paymt Mode
...	...	...	...	...
3478	2140267	12	regular	cash
3478	3446778	12	express	check
3478	4728386	17	regular	check
3479	3233444	17	express	credit
3479	3475886	12	regular	credit
...	...	...	...	...

store			
Store ID	Size	Type	Location
...	...	...	...
12	small	franchise	city
17	large	indep	rural
...	...	...	...

Relational representation of customers, orders and stores.

customer							
ID	Zip	Sex	So St	Income	Age	Club	Resp
...	...	...	...	...	...	...	...
3478	34677	m	si	60-70	32	me	nr
3479	43666	f	ma	80-90	45	nm	re
...	...	...	...	...	...	...	...

order				
Customer ID	Order ID	Store ID	Delivery Mode	Paymt Mode
...	...	...	...	...
3478	2140267	12	regular	cash
3478	3446778	12	express	check
3478	4728386	17	regular	check
3479	3233444	17	express	credit
3479	3475886	12	regular	credit
...	...	...	...	...

store			
Store ID	Size	Type	Location
...	...	...	...
12	small	franchise	city
17	large	indep	rural
...	...	...	...

Relational representation of customers, orders and stores.

ID	Zip	Sex	Soc St	Income	Age	Club	Resp
...	...	...	...	...	...	...	...
3478	34667	m	si	60-70	32	me	nr
3479	43666	f	ma	80-90	45	nm	re
...	...	...	...	...	...	...	...

Basic table for analysis

ID	Zip	Sex	Soc St	Income	Age	Club	Resp
...	...	...	...	...	...	...	...
3478	34667	m	si	60-70	32	me	nr
3479	43666	f	ma	80-90	45	nm	re
...	...	...	...	...	...	...	...

Data table presented as logical facts (Prolog format)

customer(Id,Zip,Sex,SoSt,In,Age,Club,Re)

Prolog facts describing data in Table 2:

customer(3478,34667,m,si,60-70,32,me,nr).

customer(3479,43666,f,ma,80-90,45,nm,re).

Expressing a property of a relation:

customer(_,_,f,_,_,_,_,_).

Relational Data Mining (ILP)

Data bases:

- Name of relation p
- Attribute of p
- n-tuple $< v_1, ..., v_n >$ = row in a relational table
- relation p = set of n-tuples = relational table

customer							
ID	Zip	Sex	So St	Income	Age	Club	Resp
...	...	...	...	...	...	...	...
3478	34677	m	si	60-70	32	me	nr
3479	43666	f	ma	80-90	45	nm	re
...	...	...	...	...	...	...	...

order				
Customer ID	Order ID	Store ID	Delivery Mode	Paymt Mode
...	...	...	...	...
3478	2140267	12	regular	cash
3478	3446778	12	express	check
3478	4728386	17	regular	check
3479	3233444	17	express	credit
3479	3475886	12	regular	credit
...	...	...	...	...

store			
Store ID	Size	Type	Location
...	...	...	...
12	small	franchise	city
17	large	indep	rural
...	...	...	...

Relational representation of customers, orders and stores.

Logic programming:

- Predicate symbol p
- Argument of predicate p
- Ground fact $p(v_1, ..., v_n)$
- Definition of predicate p
  - Set of ground facts
  - Prolog clause or a set of Prolog clauses

Example predicate definition:

```
good_customer(C) :-
customer(C,_,female,_,_,_,_,_),
   order(C,_,_,_,creditcard).
```

Part I. Introduction

- Data Mining in a Nutshell
- Predictive and descriptive DM techniques
- Data Mining and the KDD process
- DM standards, tools and visualization

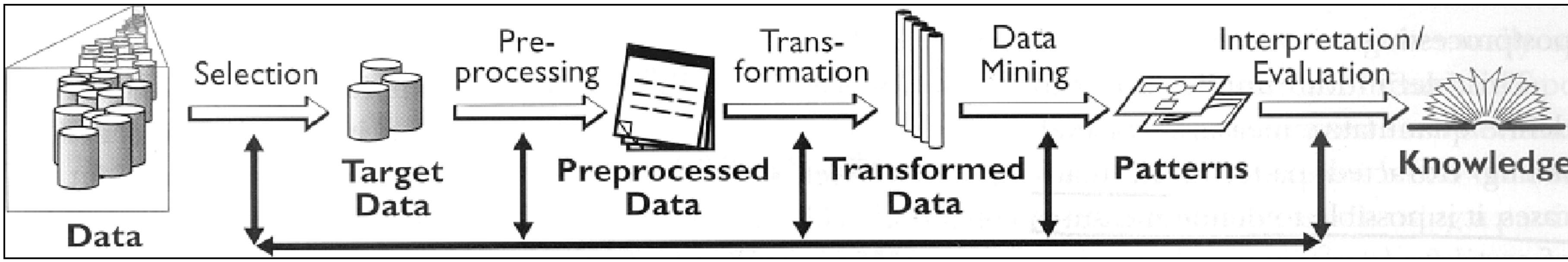
Data Mining and KDD

- KDD is defined as "the process of identifying valid, novel, potentially useful and ultimately understandable models/patterns in data." *
- Data Mining (DM) is the key step in the KDD process, performed by using data mining techniques for extracting models or interesting patterns from the data.

Usama M. Fayyad, Gregory Piatesky-Shapiro, Pedhraic Smyth: The KDD Process for Extracting Useful Knowledge form Volumes of Data. Comm ACM, Nov 96/Vol 39 No 11

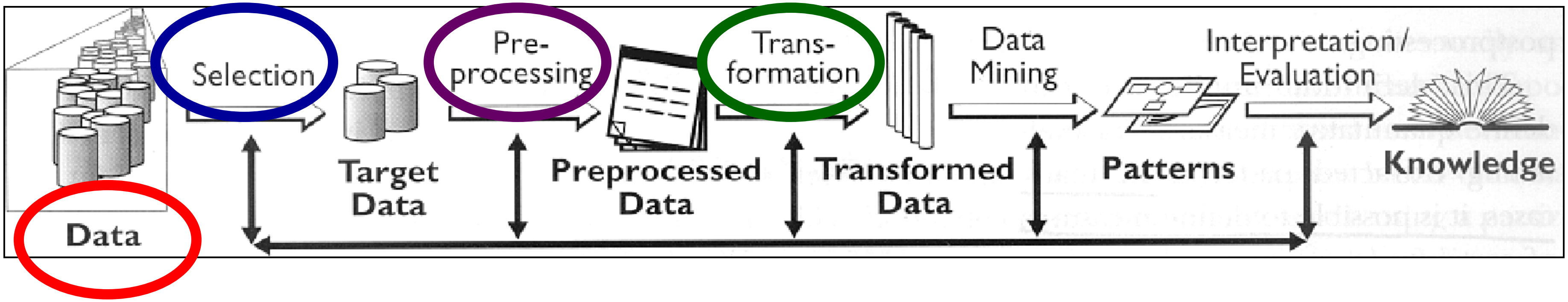
KDD Process

KDD process of discovering useful knowledge from data



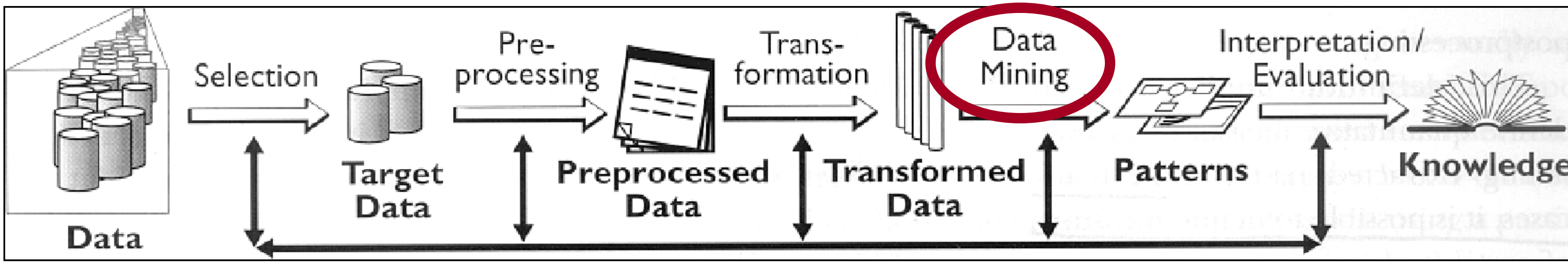
- KDD process involves several phases:
  - data preparation
  - data mining (machine learning, statistics)
  - evaluation and use of discovered patterns
- Data mining is the key step, but represents only 15%-25% of the entire KDD process

MEDIANA – analysis of media research data



- Questionnaires about journal/magazine reading, watching of TV programs and listening of radio programs, since 1992, about 1200 questions. Yearly publication: frequency of reading/listening/watching, distribution w.r.t. Sex, Age, Education, Buying power,..
- Data for 1998, about 8000 questionnaires, covering lifestyle, spare time activities, personal viewpoints, reading/listening/watching of media (yes/no/how much), interest for specific topics in media, social status
- good quality, "clean" data
- table of n-tuples (rows: individuals, columns: attributes, in classification tasks selected class)

MEDIANA – media research pilot study



- Patterns uncovering regularities concerning:
  - Which other journals/magazines are read by readers of a particular journal/magazine ?
  - What are the properties of individuals that are consumers of a particular media offer ?
  - Which properties are distinctive for readers of different journals ?
- Induced models: description (association rules, clusters) and classification (decision trees, classification rules)

Simplified association rules

Finding profiles of readers of the Delo daily newspaper

1. reads_Marketing_magazine 116 ➔ reads_Delo 95 (0.82)
2. reads_Financial_News (Finance) 223 ➔ reads_Delo 180 (0.81)
3. reads_Views (Razgledi) 201 ➔ reads_Delo 157 (0.78)
4. reads_Money (Denar) 197 ➔ reads_Delo 150 (0.76)
5. reads_Vip 181 ➔ reads_Delo 134 (0.74)

Interpretation: Most readers of Marketing magazine, Financial News, Views, Money and Vip read also Delo.

Simplified association rules

1. reads_Sara 332 ➔ reads_Slovenske novice 211 (0.64)
2. reads_Ljubezenske zgodbe 283 ➔
   reads_Slovenske novice 174 (0.61)
3. reads_Dolenjski list 520 ➔
   reads_Slovenske novice 310 (0.6)
4. reads_Omama 154 ➔ reads_Slovenske novice 90 (0.58)
5. reads_Delavska enotnost 177 ➔
   reads_Slovenske novice 102 (0.58)

Most of the readers of Sara, Love stories, Dolenjska new, Omama in Workers new read also Slovenian news.

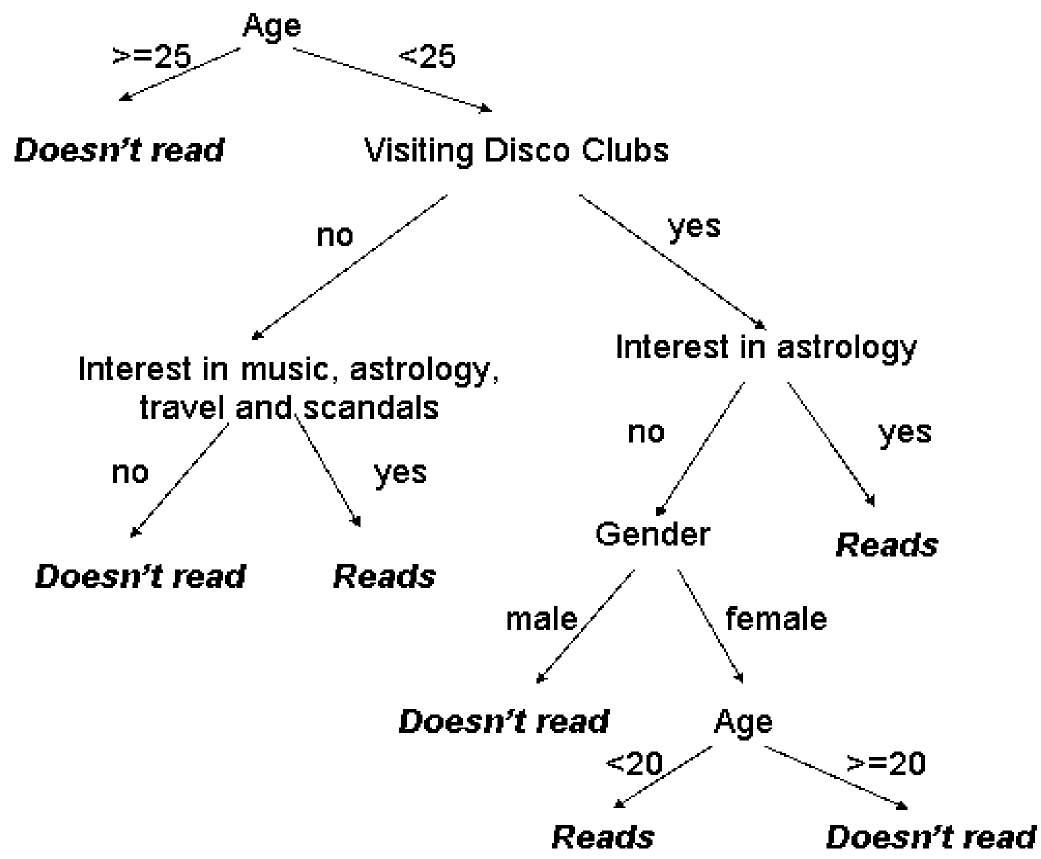
Simplified association rules

1. reads_Sportske novosti 303 ➔
   reads_Slovenski delnicar 164 (0.54)
2. reads_Sportske novosti 303 ➔
   reads_Salomonov oglasnik 155 (0.51)
3. reads_Sportske novosti 303 ➔
   reads_Lady 152 (0.5)

More than half of readers of Sports news reads also Slovenian shareholders magazine, Solomon advertisements and Lady.

Decision tree

Finding reader profiles: decision tree for classifying people into readers and non-readers of a teenage magazine Antena.

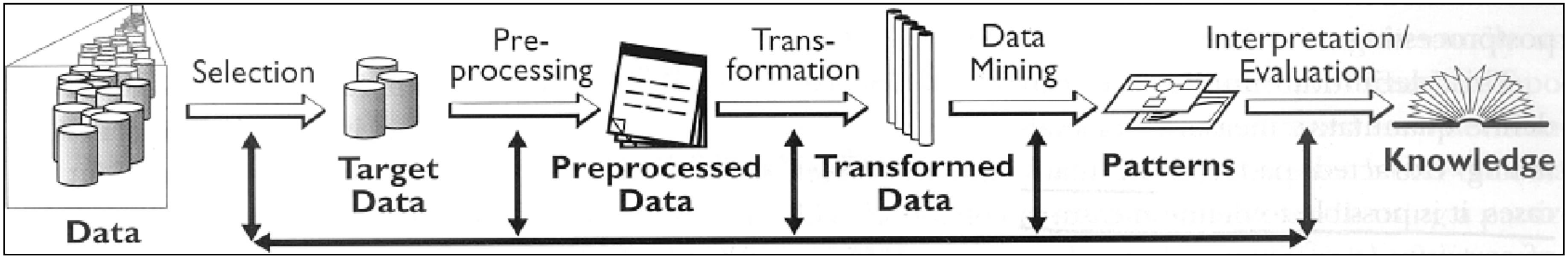


Part I. Introduction

- Data Mining in a Nutshell
- Predictive and descriptive DM techniques
- Data Mining and the KDD process
- DM standards, tools and visualization

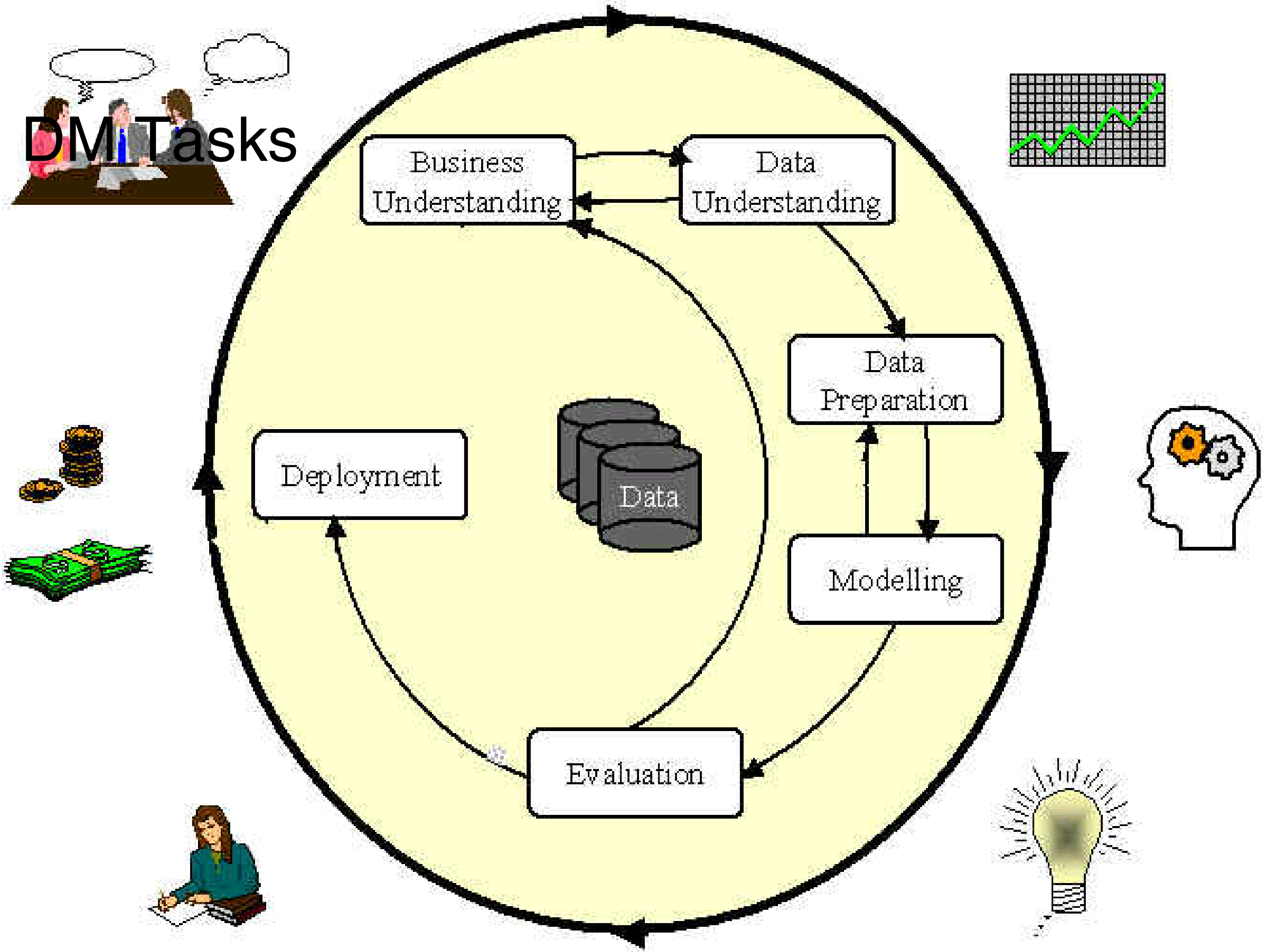
CRISP-DM

- Cross-Industry Standard Process for DM
- A collaborative, 18-months partially EC founded project started in July 1997
- NCR, ISL (Clementine), Daimler-Benz, OHRA (Dutch health insurance companies), and SIG with more than 80 members
- DM from art to engineering
- Views DM more broadly than Fayyad et al. (actually DM is treated as KDD process):

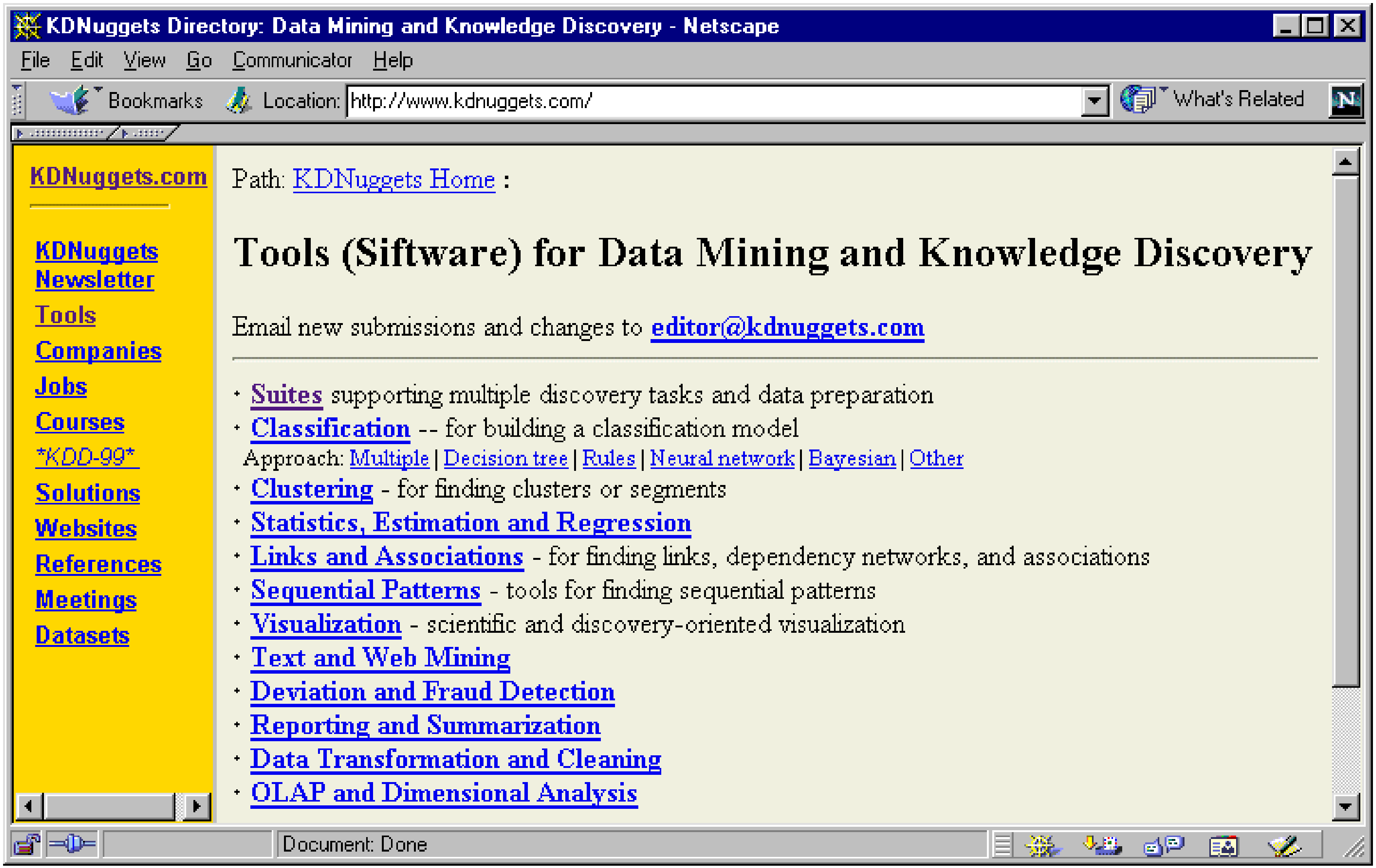


CRISP Data Mining Process

- DM Tasks

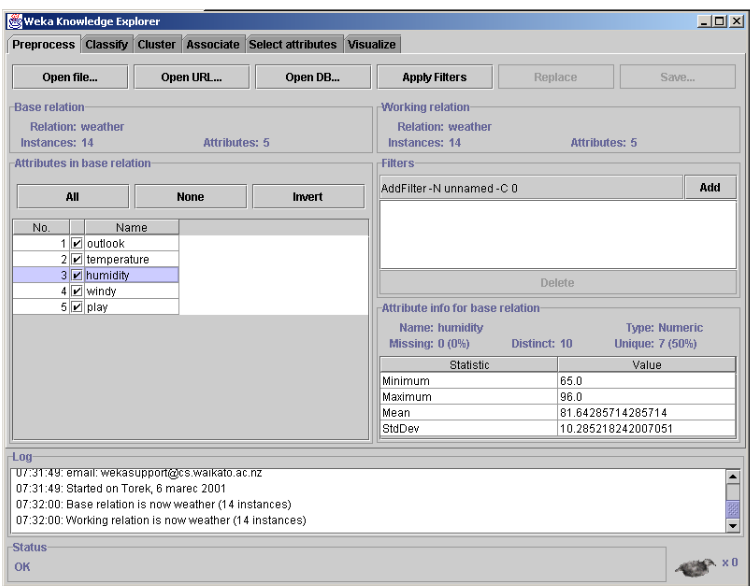


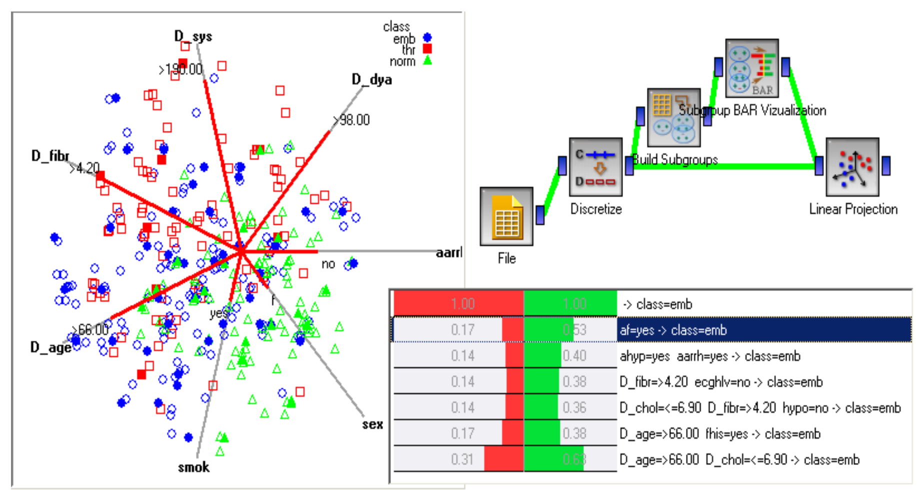
DM tools


KDNuggets Directory: Data Mining and Knowledge Discovery - Netscape
File Edit View Go Communicator Help
Bookmarks Location: http://www.kdnuggets.com/ What's Related
KDNuggets.com
KDNuggets Newsletter
Tools
Companies
Jobs
Courses
KDD-99
Solutions
Websites
References
Meetings
Datasets
Path: KDNuggets Home :
Tools (Siftware) for Data Mining and Knowledge Discovery
Email new submissions and changes to editor@kdnuggets.com
Suites supporting multiple discovery tasks and data preparation
Classification -- for building a classification model
Approach: Multiple | Decision tree | Rules | Neural network | Bayesian | Other
Clustering - for finding clusters or segments
Statistics, Estimation and Regression
Links and Associations - for finding links, dependency networks, and associations
Sequential Patterns - tools for finding sequential patterns
Visualization - scientific and discovery-oriented visualization
Text and Web Mining
Deviation and Fraud Detection
Reporting and Summarization
Data Transformation and Cleaning
OLAP and Dimensional Analysis
Document: Done

Public DM tools

- WEKA - Waikato Environment for Knowledge Analysis
- KNIME - Konstanz Information Miner
- R – Bioconductor, …
- Orange, Orange4WS, ClowdFlows

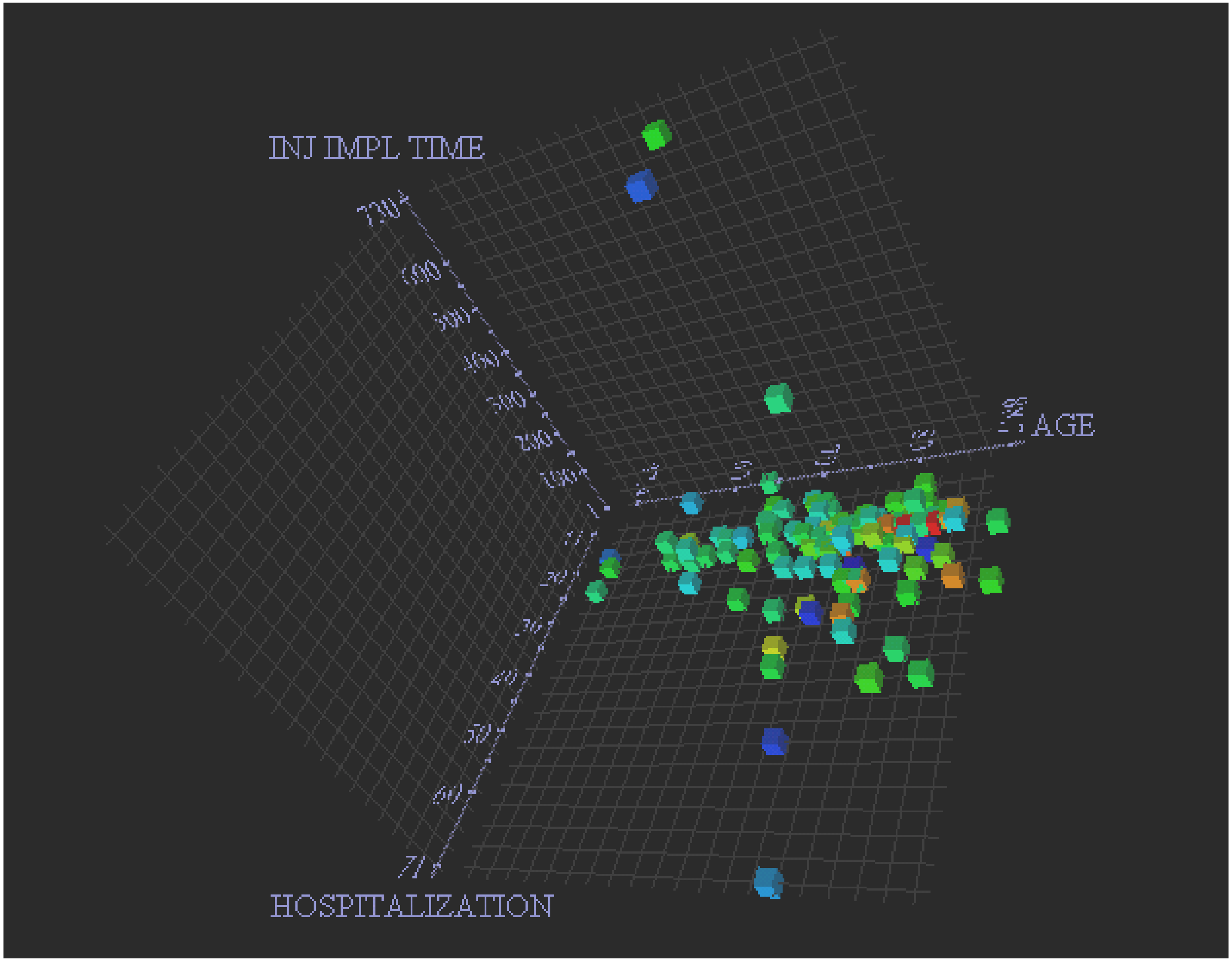




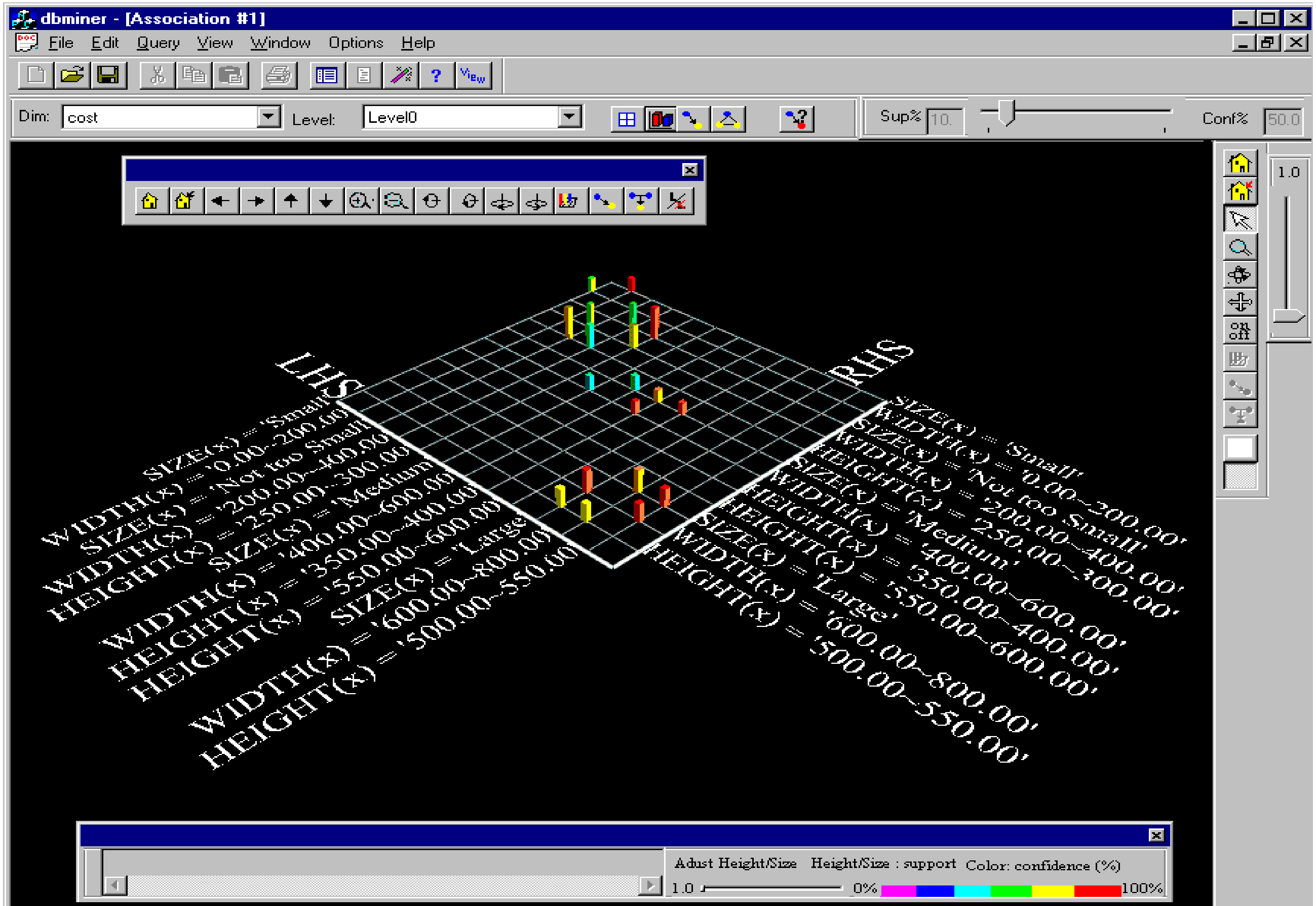
Visualization

- can be used on its own (usually for description and summarization tasks)
- can be used in combination with other DM techniques, for example
  - visualization of decision trees
  - cluster visualization
  - visualization of association rules
  - subgroup visualization

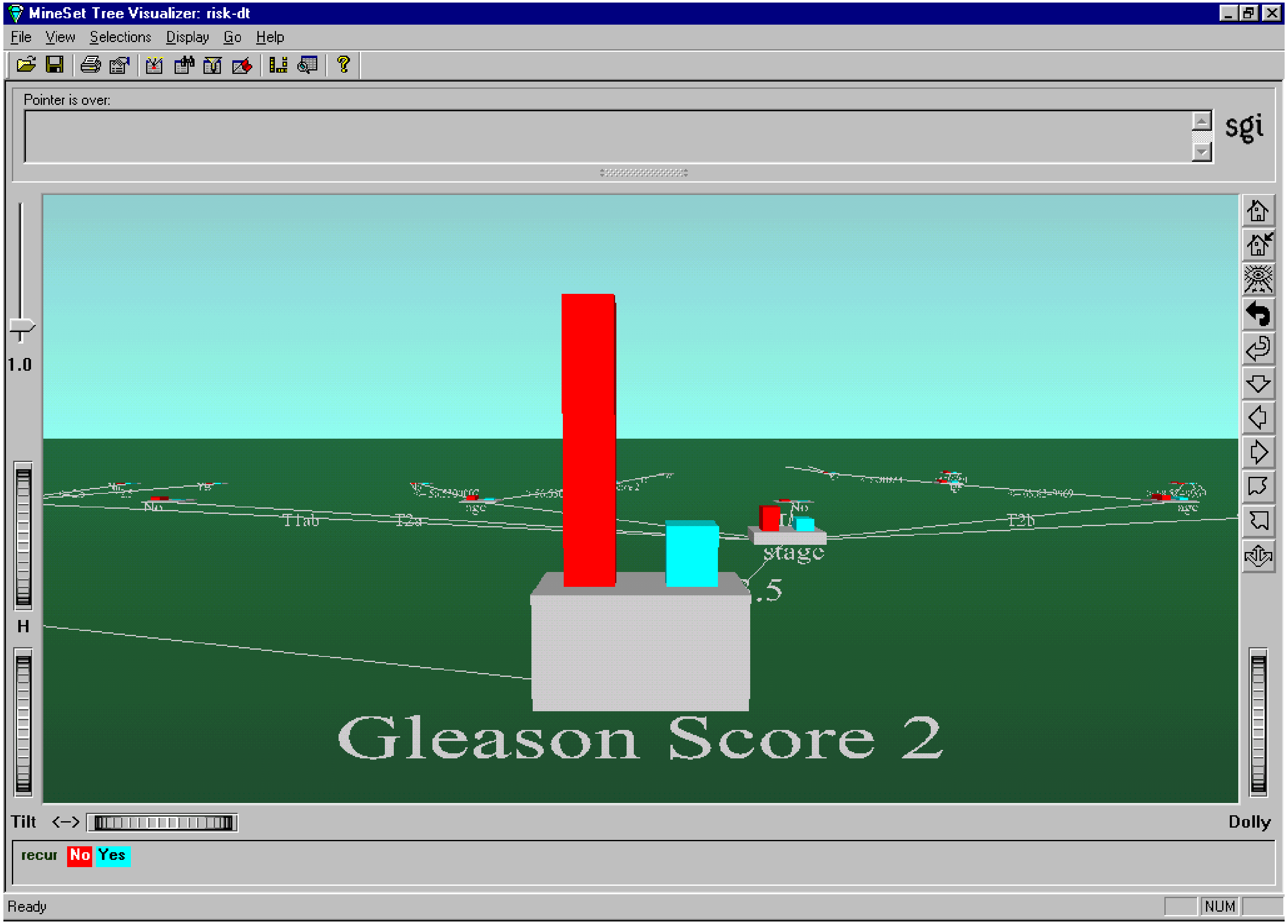
Data visualization: Scatter plot



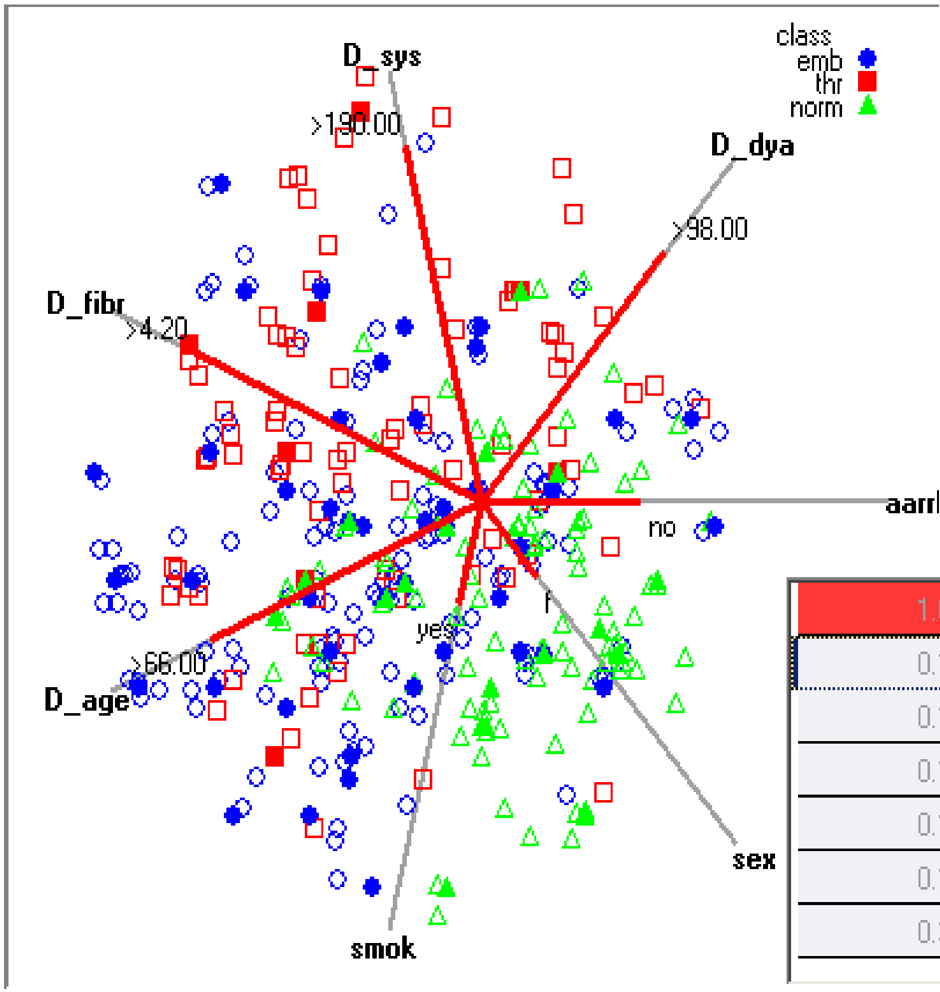
DB Miner: Association rule visualization

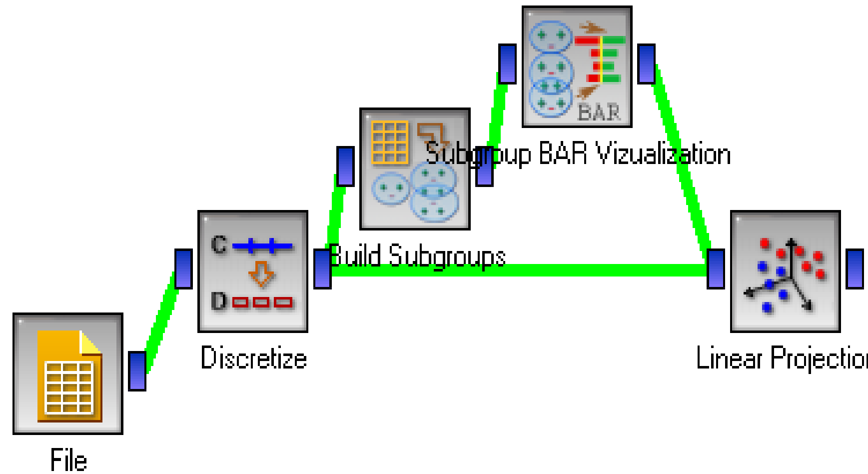


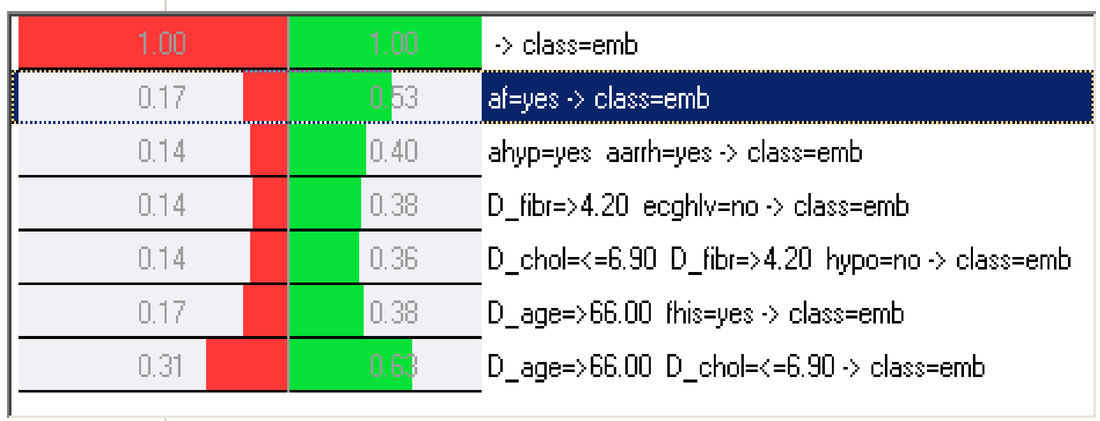
MineSet: Decision tree visualization



Orange: Visual programming and subgroup discovery visualization







Part I: Summary

- KDD is the overall process of discovering useful knowledge in data
  - many steps including data preparation, cleaning, transformation, pre-processing
- Data Mining is the data analysis phase in KDD
  - DM takes only 15%-25% of the effort of the overall KDD process
  - employing techniques from machine learning and statistics
- Predictive and descriptive induction have different goals: classifier vs. pattern discovery
- Many application areas, many powerful tools available

Introductory seminar lecture

X. JSI & Knowledge Technologies

I. Introduction: First generation data mining

- Data Mining in a nutshell
- Data Mining and KDD process
- DM standards, tools and visualization
- Classification of Data Mining techniques: Predictive and descriptive DM

  (Mladenić et al. Ch. 1 and 11, Kononenko & Kukar Ch. 1)

XX. Selected data mining techniques: Advanced subgroup discovery techniques and applications

XXX. Recent advances: Cross-context link discovery

XX. Talk outline

- Subgroup discovery in a nutshell
- Relational data mining and propositionalization in a nutshell
- Semantic data mining: Using ontologies in SD

Task reformulation: Binary Class Values

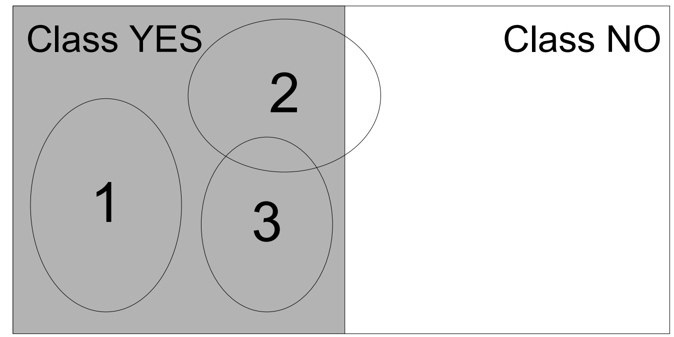
Person	Age	Spect. presc.	Astigm.	Tear prod.	Lenses
O1	17	myope	no	reduced	NO
O2	23	myope	no	normal	YES
O3	22	myope	yes	reduced	NO
O4	27	myope	yes	normal	YES
O5	19	hypermetrope	no	reduced	NO
O6-O13	...	...	...	...	...
O14	35	hypermetrope	no	normal	YES
O15	43	hypermetrope	yes	reduced	NO
O16	39	hypermetrope	yes	normal	NO
O17	54	myope	no	reduced	NO
O18	62	myope	no	normal	NO
O19-O23	...	...	...	...	...
O24	56	hypermetrope	yes	normal	NO

Binary classes (positive vs. negative examples of Target class)

- for Concept learning – classification and class description
  - for Subgroup discovery – exploring patterns characterizing groups of instances of target class

Subgroup Discovery

Person	Age	Spect. presc.	Astigm.	Tear prod.	Lenses
O1	17	myope	no	reduced	NO
O2	23	myope	no	normal	YES
O3	22	myope	yes	reduced	NO
O4	27	myope	yes	normal	YES
O5	19	hypermetrope	no	reduced	NO
O6-O13	...	...	...	...	...
O14	35	hypermetrope	no	normal	YES
O15	43	hypermetrope	yes	reduced	NO
O16	39	hypermetrope	yes	normal	NO
O17	54	myope	no	reduced	NO
O18	62	myope	no	normal	NO
O19-O23	...	...	...	...	...
O24	56	hypermetrope	yes	normal	NO

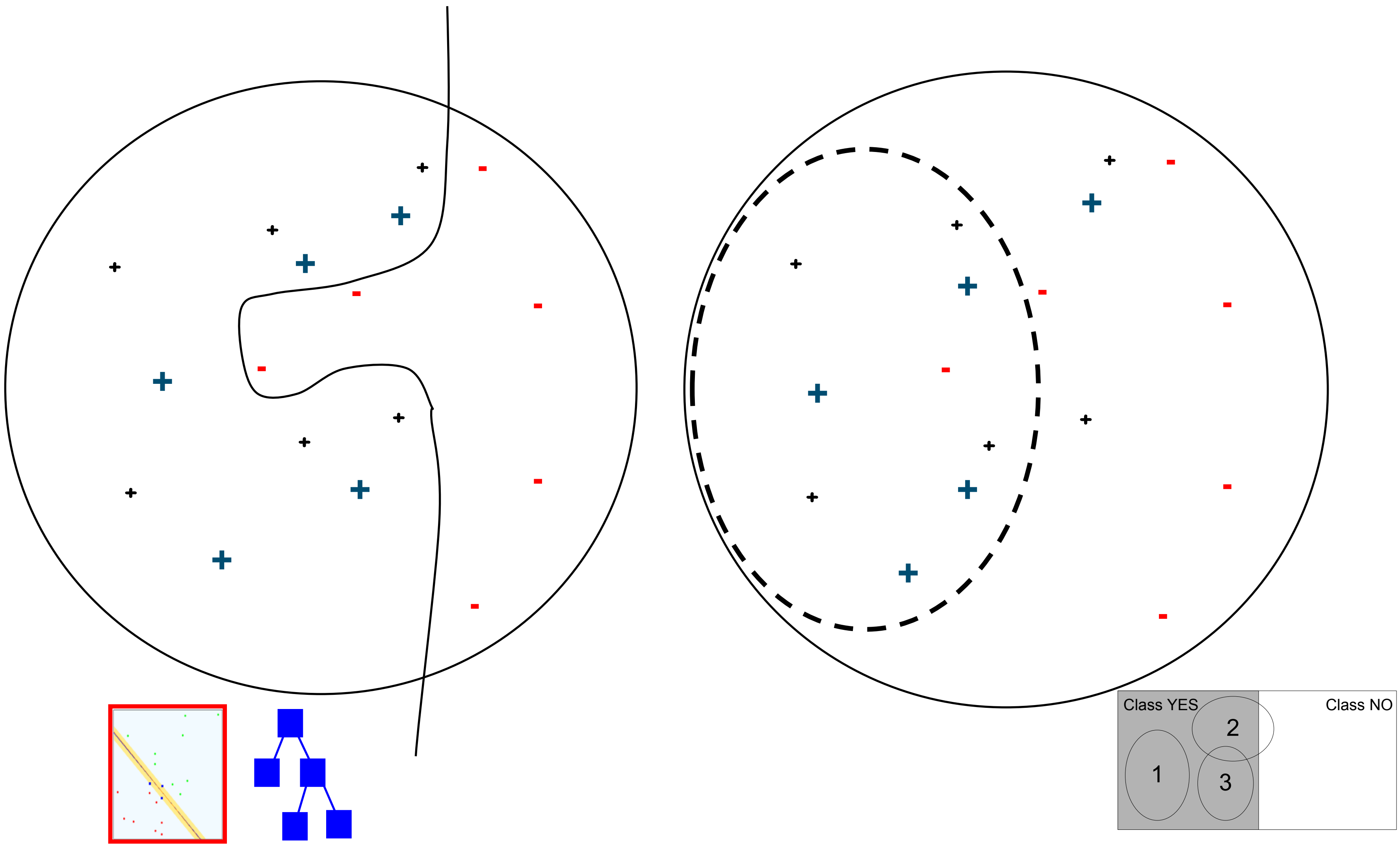


- A task in which individual interpretable patterns in the form of rules are induced from data, labeled by a predefined property of interest.
- SD algorithms learn several independent rules that describe groups of target class examples
  - subgroups must be large and significant

Classification versus Subgroup Discovery

- **Classification (predictive induction) - constructing sets of classification rules**
  - aimed at learning a model for classification or prediction
  - rules are dependent
- **Subgroup discovery (descriptive induction) – constructing individual subgroup describing rules**
  - aimed at finding interesting patterns in target class examples
    - large subgroups (high target class coverage)
    - with significantly different distribution of target class examples (high TP/FP ratio, high significance, high WRAcc
  - each rule (pattern) is an independent chunk of knowledge

Classification versus Subgroup discovery



Subgroup discovery task

Task definition (Kloesgen, Wrobel 1997)

- **Given:** a population of individuals and a property of interest (target class, e.g. CHD)
- **Find:** `most interesting' descriptions of population subgroups
  - are as large as possible

    (high target class coverage)
  - have most unusual distribution of the target property

    (high TP/FP ratio, high significance)

Subgroup discovery example: CHD Risk Group Detection

Input: Patient records described by **stage A** (anamnestic), stage **B** (an. & lab.), **and stage C** (an., lab. & ECG) attributes

Task: Find and characterize population subgroups with high CHD risk (large enough, distributionally unusual)

From **best induced descriptions**, five were selected by the expert as **most actionable** for CHD risk screening (by GPs):

CHD-risk ← male & pos. fam. history & age > 46

CHD-risk ← female & bodymassIndex > 25 & age > 63

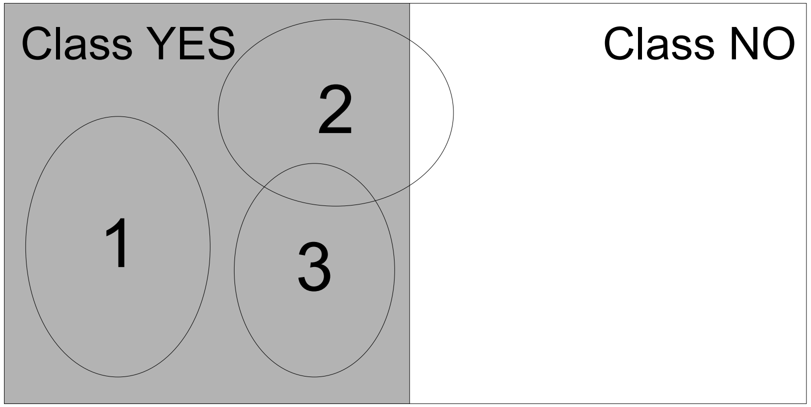
CHD-risk ← ...

CHD-risk ← ...

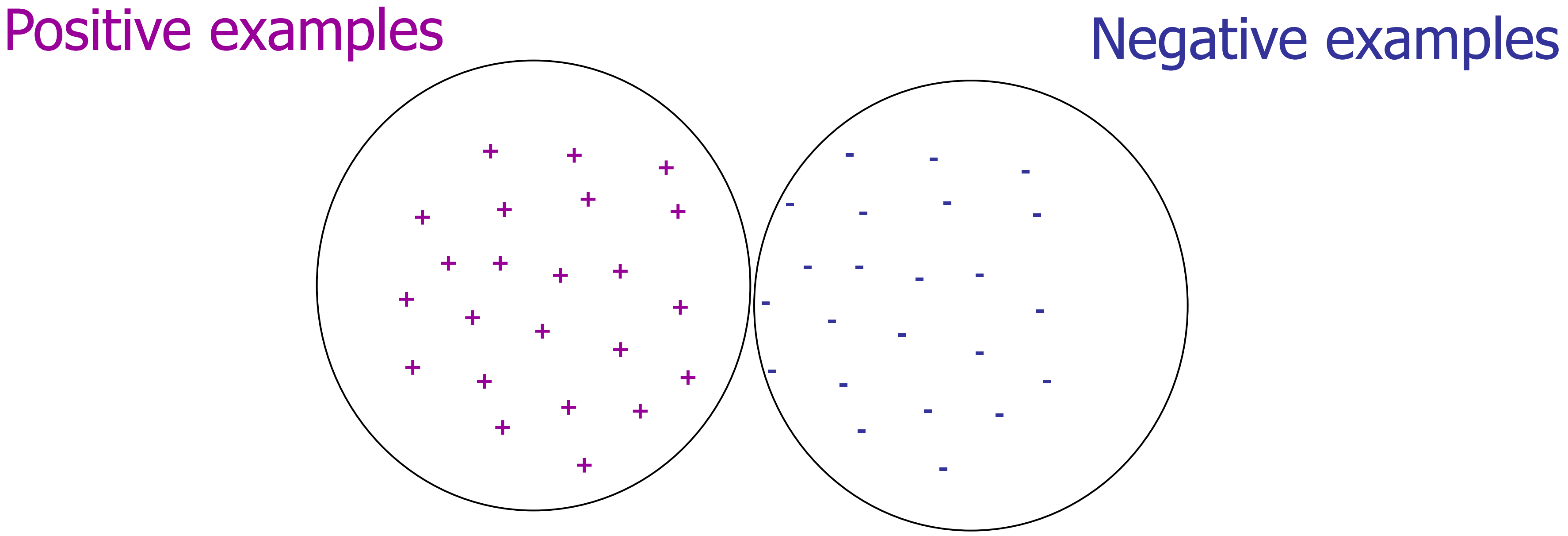
CHD-risk ← ...

Characteristics of SD Algorithms

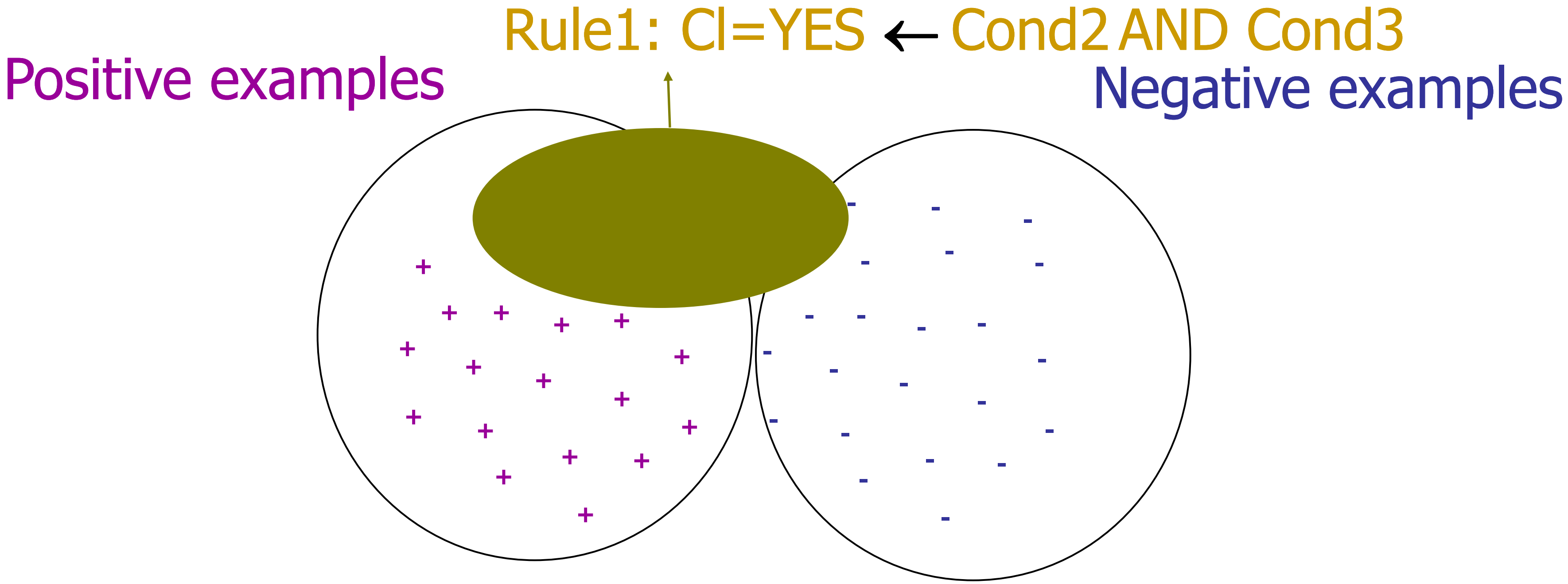
- SD algorithms do not look for a single complex rule to describe all examples of target class YES (all CHD-risk patients), but several rules that describe parts (subgroups) of YES.
- Standard rule learning approach: Using the covering algorithm for rule set construction



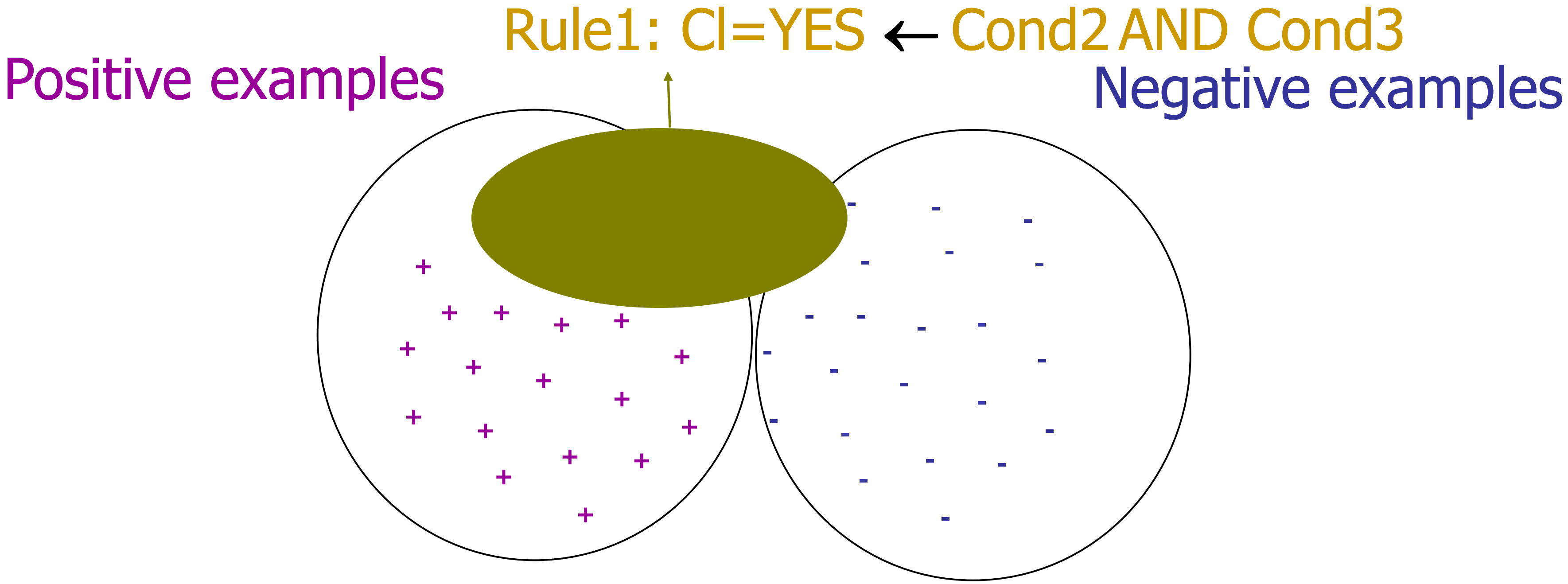
Covering algorithm



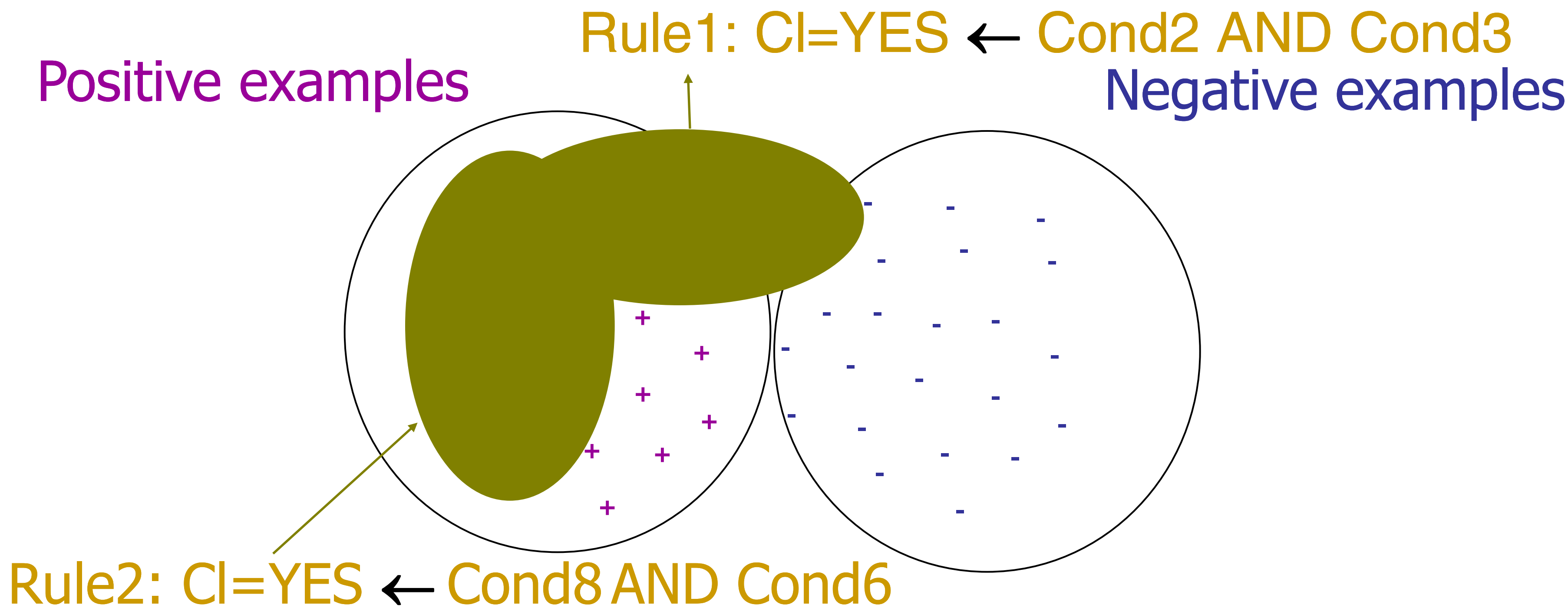
Covering algorithm



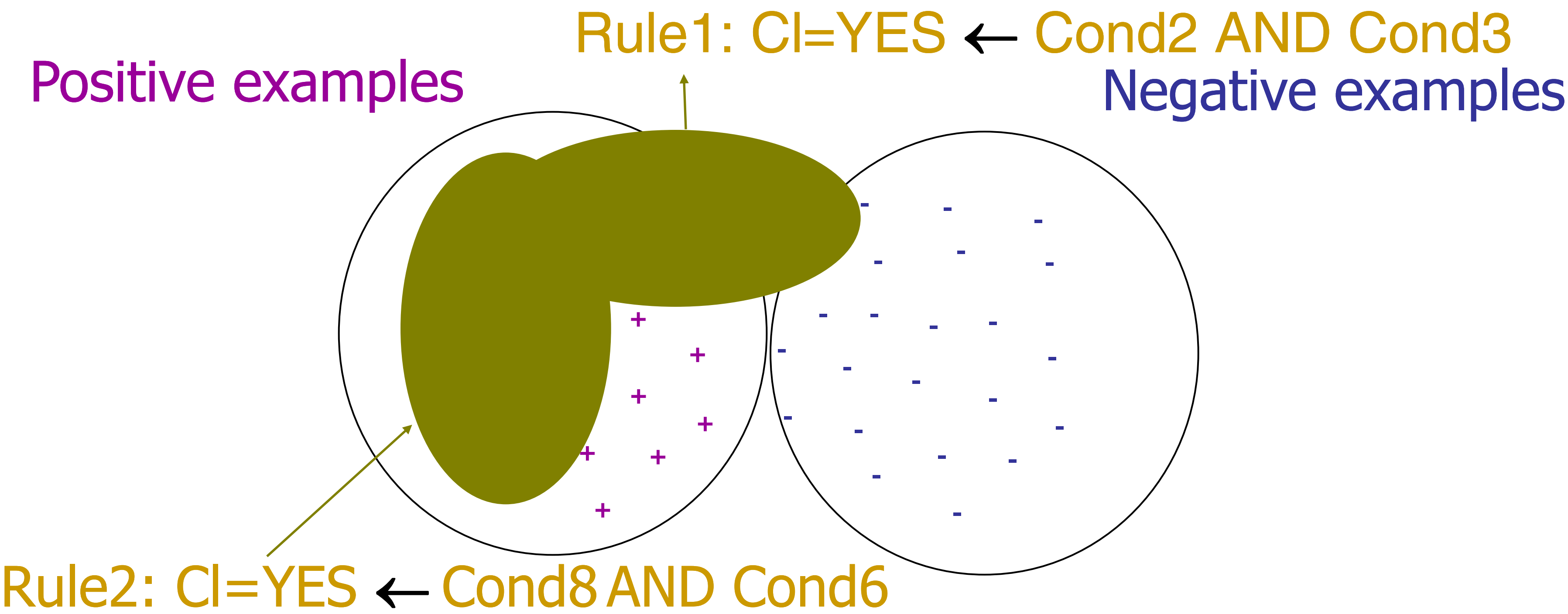
Covering algorithm



Covering algorithm

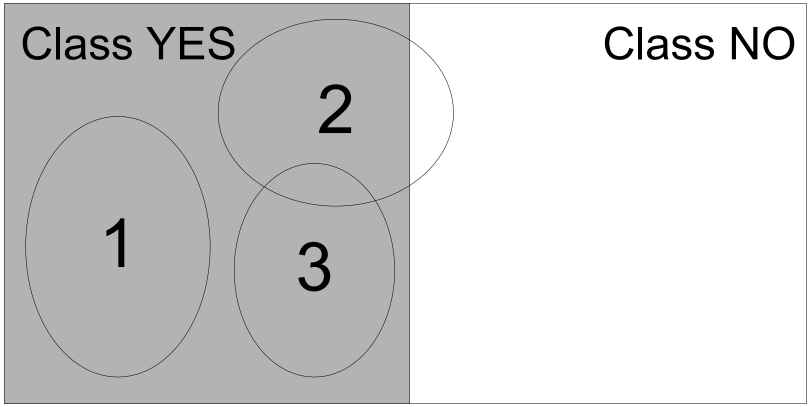


Covering algorithm



Characteristics of SD Algorithms

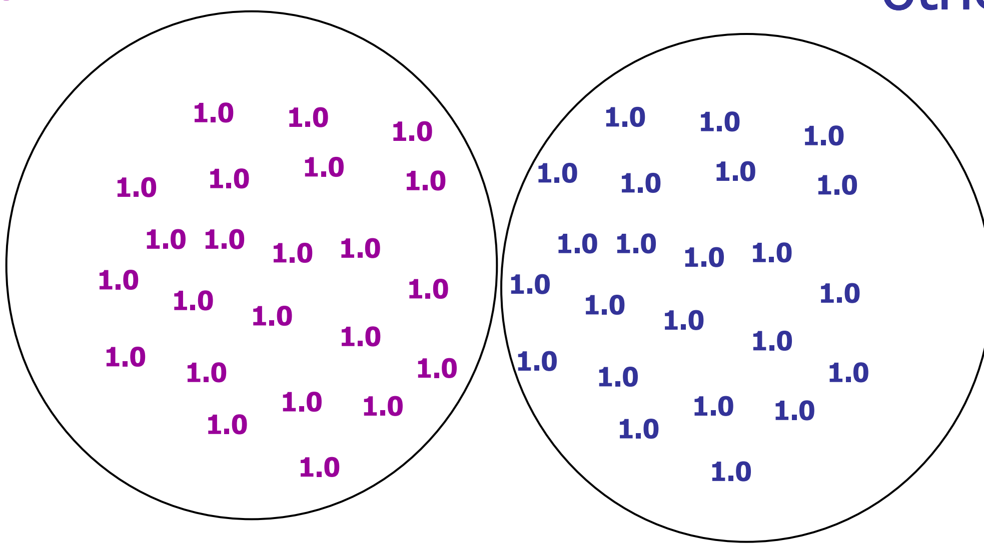
- SD algorithms do not look for a single complex rule to describe all examples of target class YES (all CHD-risk patients), but several rules that describe parts (subgroups) of YES.
- Advanced rule learning approach: using example weights in the weighted covering algorithm for repetitive subgroup construction and in the rule quality evaluation heuristics.



Weighted covering algorithm for rule set construction

CHD patients

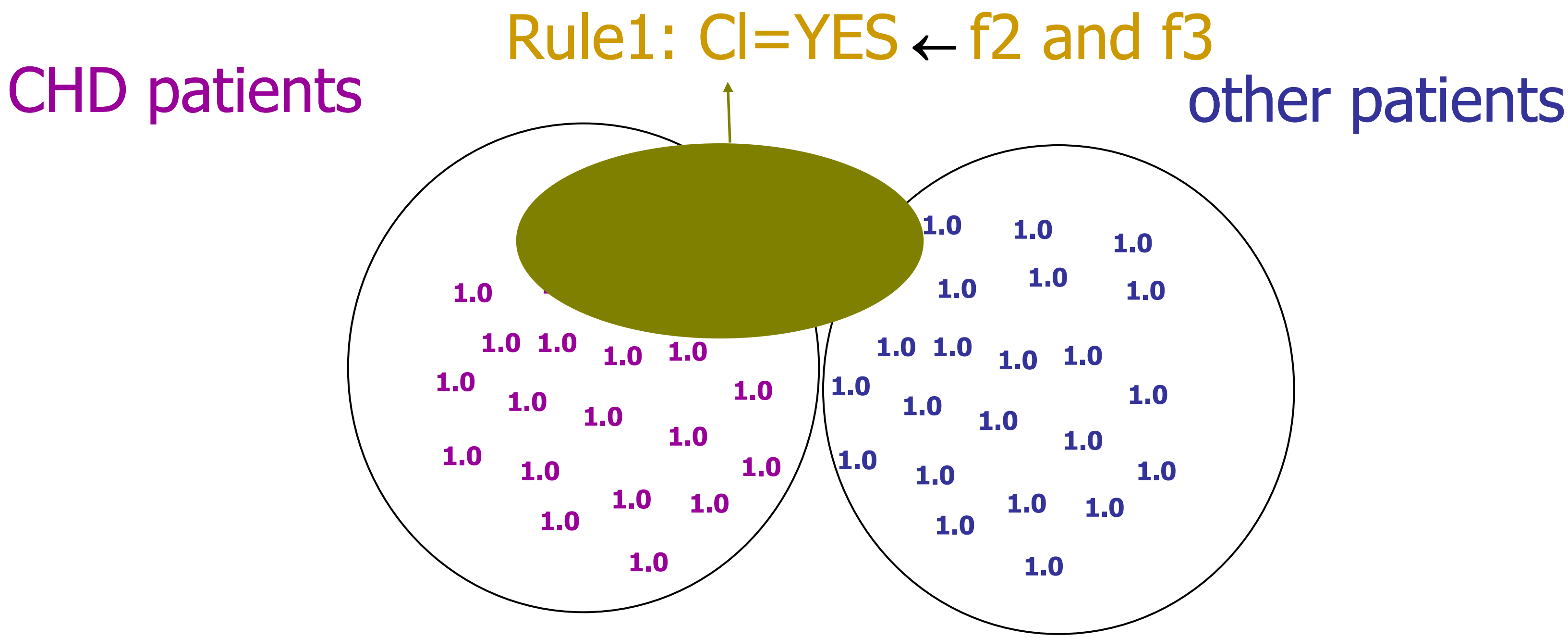
other patients



- For learning a set of subgroup describing rules, SD implements an iterative weigthed covering algorithm.

- Quality of a rule is measured by trading off coverage and precision.

Weighted covering algorithm for rule set construction

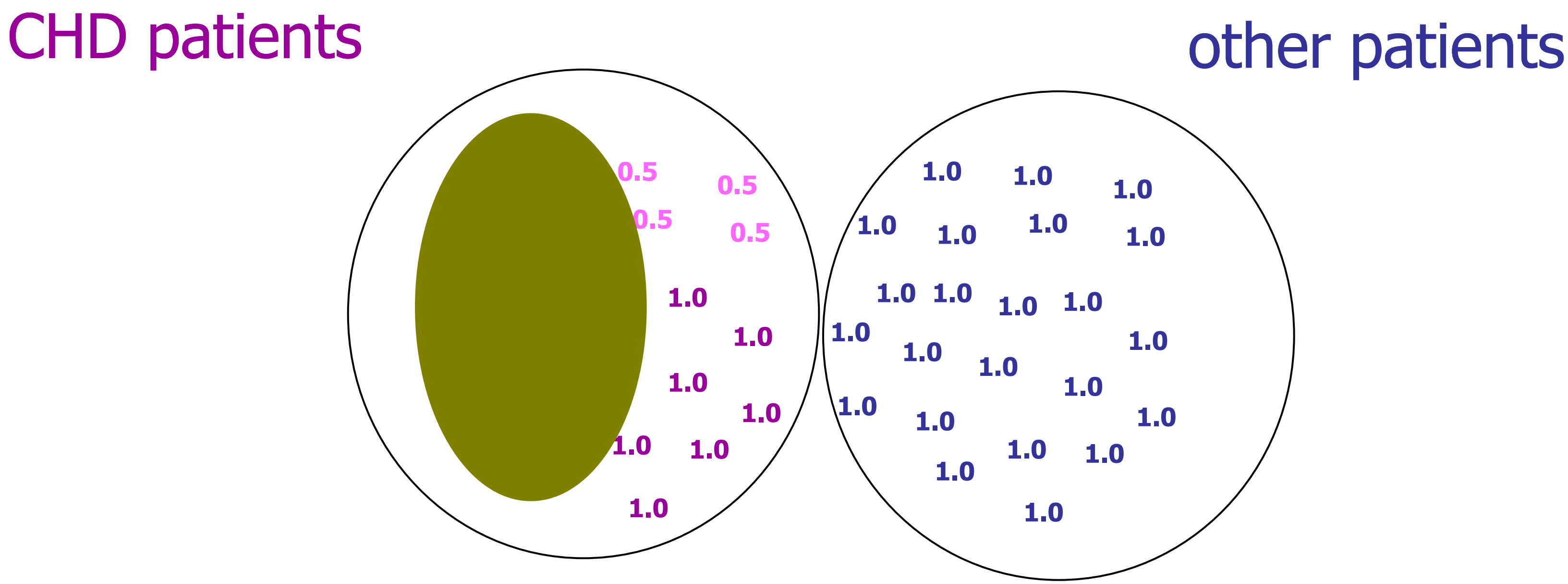


Rule quality measure in SD: q(Cl ← Cond) = TP/(FP+g)

Rule quality measure in CN2-SD: WRAcc(Cl ←Cond) = p(Cond) x [p(Cl | Cond) – p(Cl)] = coverage x (precision – default precision)

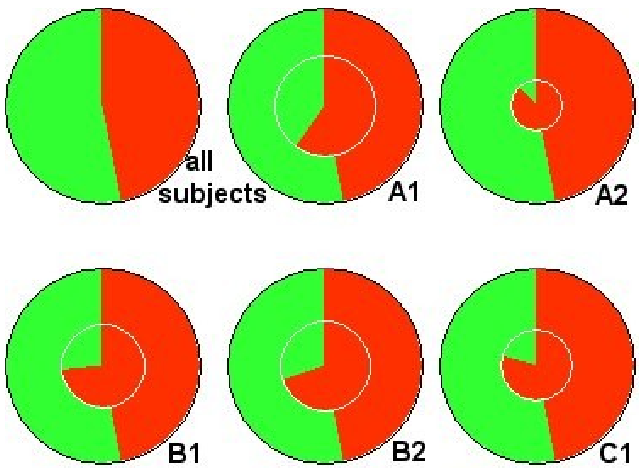
***Coverage** = sum of the covered weights, ***Precision** = purity of the covered examples

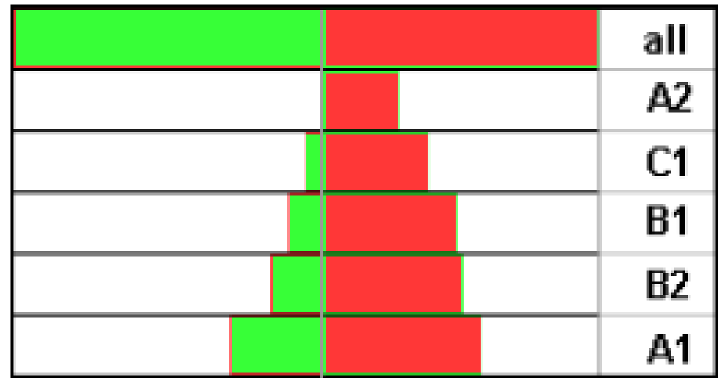
Weighted covering algorithm for rule set construction

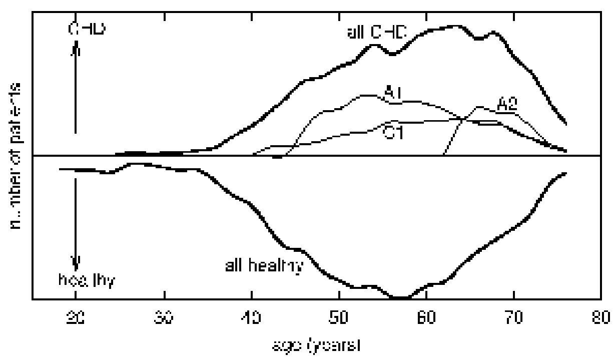


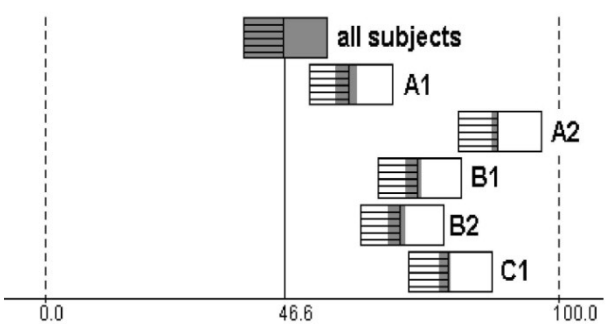
In contrast with classification rule learning algorithms (e.g. CN2), the covered positive examples are not deleted from the training set in the next rule learning iteration; they are re-weighted, and a next ‘best’ rule is learned.

Subgroup visualization









The CHD task: Find, characterize and visualize population subgroups with high CHD risk (large enough, distributionally unusual, most actionable)

Induced subgroups and their statistical characterization

Subgroup A2 for femle patients:

High-CHD-risk **IF**

body mass index over 25 kg/m^2 (typically 29)

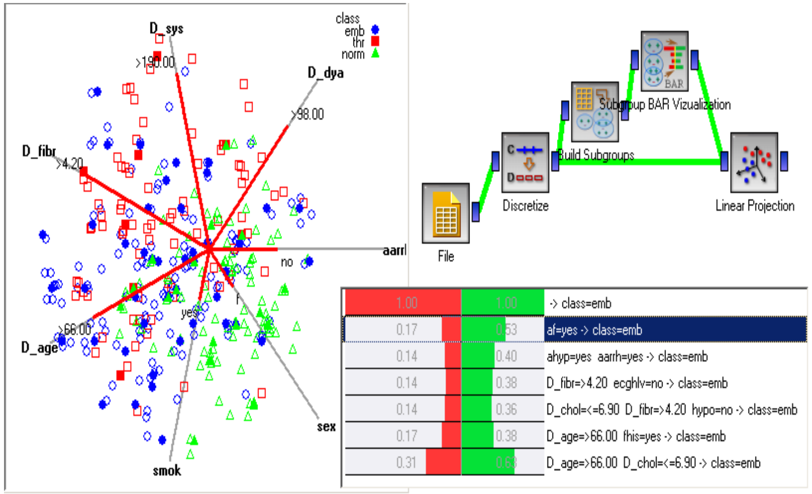
AND

age over 63 years

Supporting characteristics (computed using ℵ2 statistical significance test) are: positive family history and hypertension. Women in this risk group typically have slightly increased LDL cholesterol values and normal but decreased HDL cholesterol values.

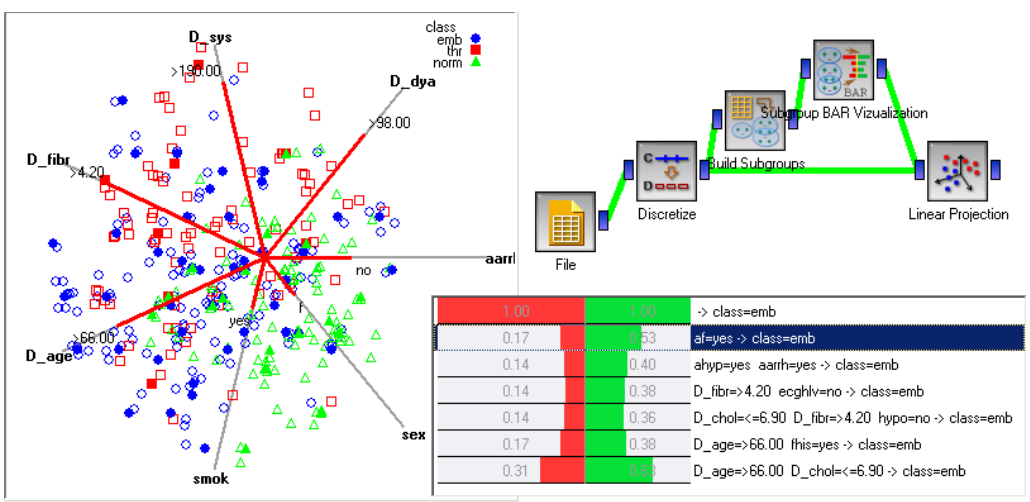
SD algorithms in the Orange DM Platform

- **SD Algorithms in Orange**
  - SD (Gamberger & Lavrač, JAIR 2002
  - APRIORI-SD (Kavšek & Lavrač, AAI 2006
  - CN2-SD (Lavrač et al., JMLR 2004): Adapting CN2 classification rule learner to Subgroup Discovery
    - Weighted covering algorithm
    - Weighted relative accuracy (WRAcc) search heuristics, with added example weights



SD algorithms in Orange and Orange4WS

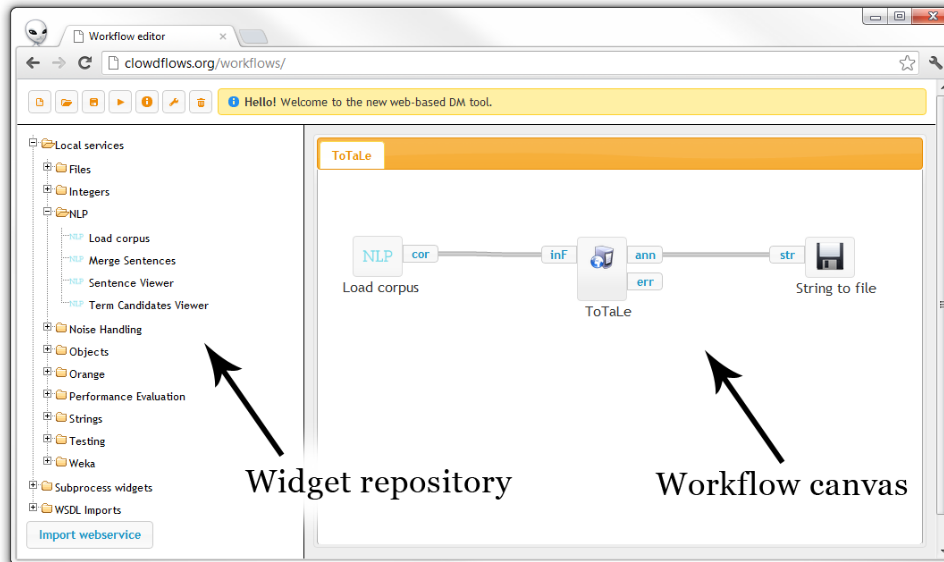
- **Orange**
  - classification and subgroup discovery algorithms
  - data mining workflows
  - visualization
  - developed at FRI, Ljubljana

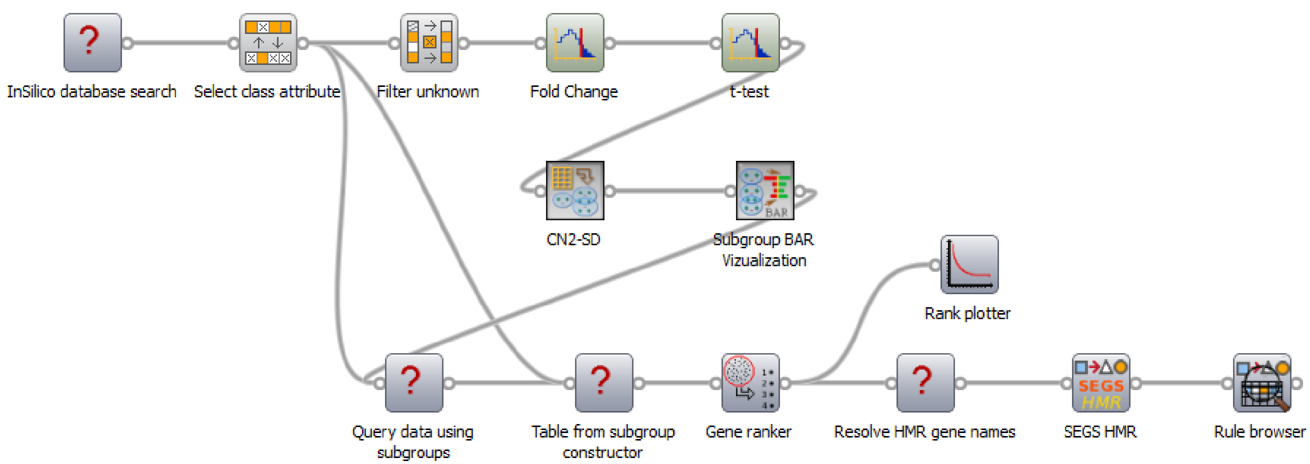


- **Orange4WS** (Podpečan 2010)
  - Web service oriented
  - supports workflows and other Orange functionality
  - includes also
    - WEKA algorithms
    - relational data mining
    - semantic data mining with ontologies
  - Web-based platform is under construction

Current platform and workflow developments

- CrowdFlows browser-based DM platform (Kranjc et al. 2012)
- Semantic Subgroup Discovery workflows (Vavpetič et al., 2012)





XX. Talk outline

- Subgroup discovery in a nutshell
- Relational data mining and propositionalization in a nutshell
- Semantic data mining: Using ontologies in SD

Relational Data Mining (Inductive Logic Programming) in a nutshell

customer

ID	Zip	Sex	SoSt	Income	Age	Club	Resp
...	...	...	...	...	...	...	...
3478	34677	m	si	60-70	32	me	nr
3479	43666	f	ma	80-90	45	nm	re
...	...	...	...	...	...	...	...

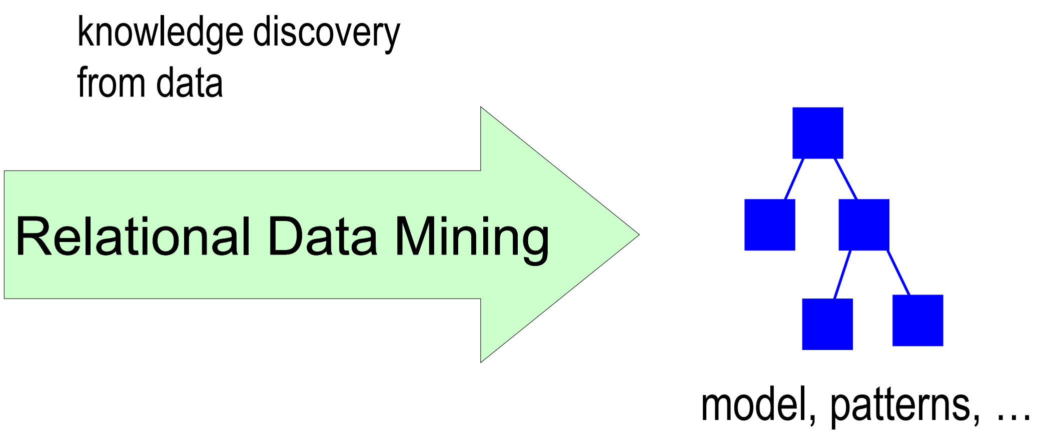
order

Customer ID	Order ID	Store ID	Delivery Mode	Paymt Mode
...	...	...	...	...
3478	2140267	12	regular	cash
3478	3446778	12	express	check
3478	4728386	17	regular	check
3479	3233444	17	express	credit
3479	3475886	12	regular	credit
...	...	...	...	...

store

Store ID	Size	Type	Location
...	...	...	...
12	small	franchise	city
17	large	indep	rural
...	...	...	...

Relational representation of customers, orders and stores.

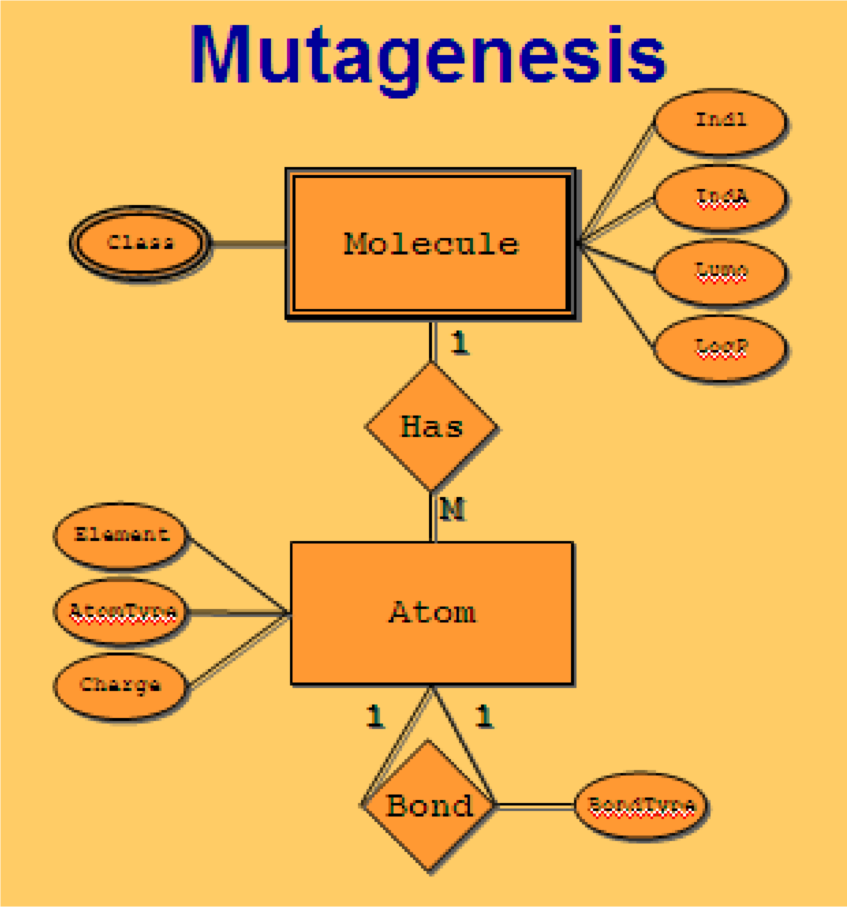


Given: a relational database, a set of tables. sets of logical facts, a graph, …

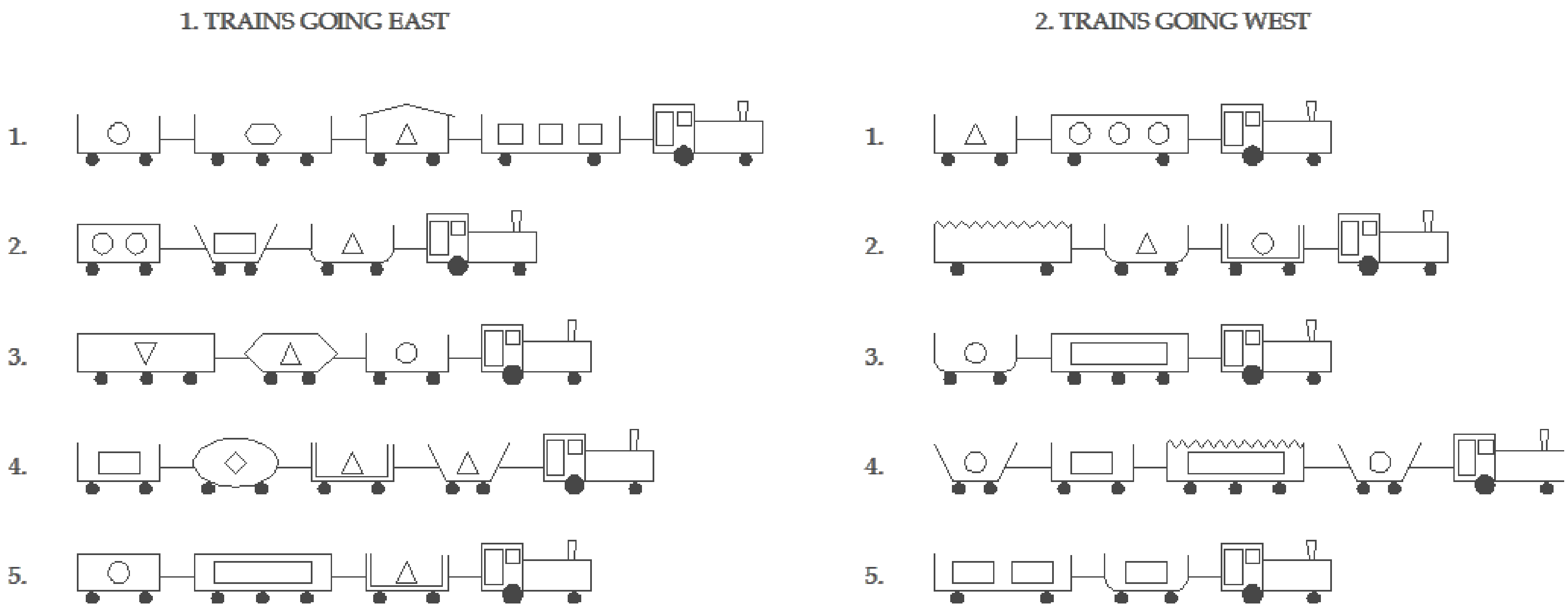
Find: a classification model, a set of interesting patterns

Relational Data Mining (ILP)

- Learning from multiple tables
  - patient records connected with other patient and demographic information
- Complex relational problems:
  - temporal data: time series in medicine, ...
  - structured data: representation of molecules and their properties in protein engineering, biochemistry, ...



Sample ILP problem: East-West trains



Relational data representation

CAR	OBJECT	NUMBER
c1	circle	1
c2	hexagon	1
c3	triangle	1
c4	rectangle	3
...	...	

TRAIN_TABLE

TRAIN	EASTBOUND
	UE
	UE
	SE

CAR	TRAIN	SHAPE	LENGTH	ROOF	WHEELS
c1	t1	rectangle	short	none	2
c2	t1	rectangle	long	none	3
c3	t1	rectangle	short	peaked	2
c4	t1	rectangle	long	none	2
...	...	...			...

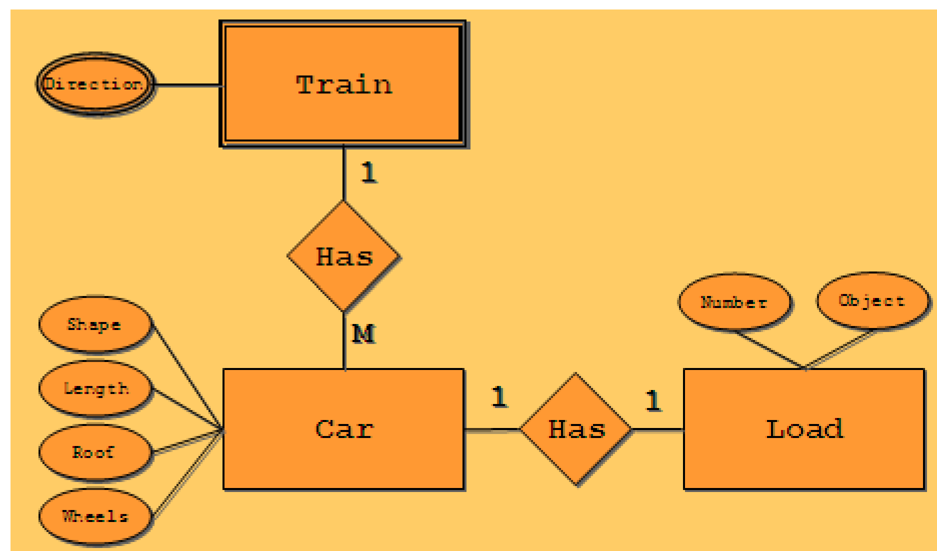
Relational data representation

AR	OBJECT	NUMBER
c1	circle	1
c2	hexagon	1
c3	triangle	1
c4	rectangle	3
...	...	

TRAIN_TABLE

TRAIN	EASTBOUND
	UE
	UE
	SE

CAR	TRAIN	SHAPE	LENGTH	ROOF	WHEELS
c1	t1	rectangle	short	none	2
c2	t1	rectangle	long	none	3
c3	t1	rectangle	short	peaked	2
c4	t1	rectangle	long	none	2
...	...	...			...



Propositionalization in a nutshell

CAR	OBJECT	NUMBER
c1	circle	1
c2	hexagon	1
c3	triangle	1
c4	rectangle	3
...	...	

TRAIN_TABLE

TRAIN	EASTBOUND
	...UE
	...UE
	...SE

CAR	TRAIN	SHAPE	LENGTH	ROOF	WHEELS
c1	t1	rectangle	short	none	2
c2	t1	rectangle	long	none	3
c3	t1	rectangle	short	peaked	2
c4	t1	rectangle	long	none	2
...	...	...			...

Transform a multi-relational (**multiple-table**) representation to a propositional representation (**single table**)

Proposed in ILP systems LINUS (Lavrac et al. 1991, 1994), 1BC (Flach and Lachiche 1999), …

Propositionalization in a nutshell

Main propositionalization step:
first-order feature construction

f1(T):-hasCar(T,C),clength(C,short).

f2(T):-hasCar(T,C), hasLoad(C,L),
loadShape(L,circle)

f3(T) :- ....

Propositional learning:

t(T) ← f1(T), f4(T)

Relational interpretation:

eastbound(T) ←
hasShortCar(T),hasClosedCar(T).

LOAD	CAR	OBJECT	NUMBER
l1	c1	circle	1
l2	c2	hexagon	1
l3	c3	triangle	1
l4	c4	rectangle	3
...	...	...	

TRAIN_TABLE

	UND
	JE
	JE
	SE

CAR	TRAIN	SHAPE	LENGTH	ROOF	WHEELS
c1	t1	rectangle	short	none	2
c2	t1	rectangle	long	none	3
c3	t1	rectangle	short	peaked	2
c4	t1	rectangle	long	none	2
...	...	...			...

PROPOSITIONAL TRAIN_TABLE

train(T)	f1(T)	f2(T)	f3(T)	f4(T)	f5(T)
t1	t	t	f	t	t
t2	t	t	t	t	t
t3	f	f	t	f	f
t4	t	f	t	f	f
...	...	...			...

Relational Data Mining through Propositionalization

customer

ID	Zip	Sex	So St	Income	Age	Club	Resp
...	...	...	...	...	...	...	...
3478	34677	m	si	60-70	32	me	nr
3479	43666	f	ma	80-90	45	nm	re
...	...	...	...	...	...	...	...

order

Customer ID	Order ID	Store ID	Delivery Mode	Paymt Mode
...	...	...	...	...
3478	2140267	12	regular	cash
3478	3446778	12	express	check
3478	4728386	17	regular	check
3479	3233444	17	express	credit
3479	3475886	12	regular	credit
...	...	...	...	...

store

Store ID	Size	Type	Location
...	...	...	...
12	small	franchise	city
17	large	indep	rural
...	...	...	...

Relational representation of customers, orders and stores.

Step 1

Propositionalization

	f1	f2	f3	f4	f5	f6	...				...	fn
g1	1	0	0	1	1	1	0	0	1	0	1	1
g2	0	1	1	0	1	1	0	0	0	1	1	0
g3	0	1	1	1	0	0	1	1	0	0	0	1
g4	1	1	1	0	1	1	0	0	1	1	1	0
g5	1	1	1	0	0	1	0	1	1	0	1	0
g1	0	0	1	1	0	0	0	1	0	0	0	1
g2	1	1	0	0	1	1	0	1	0	1	1	1
g3	0	0	0	0	1	0	0	1	1	1	0	0
g4	1	0	1	1	1	0	1	0	0	1	0	1

Relational Data Mining through Propositionalization

customer							
ID	Zip	Sex	SoSt	Income	Age	Club	Resp
...	...	...	...	...	...	...	...
3478	34677	m	si	60-70	32	me	nr
3479	43666	f	ma	80-90	45	nm	re
...	...	...	...	...	...	...	...

order				
Customer ID	Order ID	Store ID	Delivery Mode	Paymt Mode
...	...	...	...	...
3478	2140267	12	regular	cash
3478	3446778	12	express	check
3478	4728386	17	regular	check
3479	3233444	17	express	credit
3479	3475886	12	regular	credit
...	...	...	...	...

store			
Store ID	Size	Type	Location
...	...	...	...
12	small	franchise	city
17	large	indep	rural
...	...	...	...

Relational representation of customers, orders and stores.

Step 1

Propositionalization

1. constructing relational features
2. constructing a propositional table

	f1	f2	f3	f4	f5	f6	...				...	fn
g1	1	0	0	1	1	1	0	0	1	0	1	1
g2	0	1	1	0	1	1	0	0	0	1	1	0
g3	0	1	1	1	0	0	1	1	0	0	0	1
g4	1	1	1	0	1	1	0	0	1	1	1	0
g5	1	1	1	0	0	1	0	1	1	0	1	0
g1	0	0	1	1	0	0	0	1	0	0	0	1
g2	1	1	0	0	1	1	0	1	0	1	1	1
g3	0	0	0	0	1	0	0	1	1	1	0	0
g4	1	0	1	1	1	0	1	0	0	1	0	1

Relational Data Mining through Propositionalization

customer							
ID	Zip	S ex	So St	In come	A ge	Cl ub	Re sp
...	...	...	...	...	...	...	...
3478	34677	m	si	60-70	32	me	nr
3479	43666	f	ma	80-90	45	nm	re
...	...	...	...	...	...	...	...

order				
Customer ID	Order ID	Store ID	Delivery Mode	Paymt Mode
...	...	...	...	...
3478	2140267	12	regular	cash
3478	3446778	12	express	check
3478	4728386	17	regular	check
3479	3233444	17	express	credit
3479	3475886	12	regular	credit
...	...	...	...	...

store			
Store ID	Size	Type	Location
...	...	...	...
12	small	franchise	city
17	large	indep	rural
...	...	...	...

Relational representation of customers, orders and stores.

Step 1

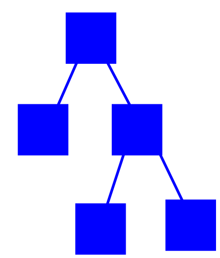
Propositionalization

	f1	f2	f3	f4	f5	f6	...				...	fn
g1	1	0	0	1	1	1	0	0	1	0	1	1
g2	0	1	1	0	1	1	0	0	0	1	1	0
g3	0	1	1	1	0	0	1	1	0	0	0	1
g4	1	1	1	0	1	1	0	0	1	1	1	0
g5	1	1	1	0	0	1	0	1	1	0	1	0
g1	0	0	1	1	0	0	0	1	0	0	0	1
g2	1	1	0	0	1	1	0	1	0	1	1	1
g3	0	0	0	0	1	0	0	1	1	1	0	0
g4	1	0	1	1	1	0	1	0	0	1	0	1

	f1	f2	f3	f4	f5	f6	...				...	fn
g1	1	0	0	1	1	1	0	0	1	0	1	1
g2	0	1	1	0	1	1	0	0	0	1	1	0
g3	0	1	1	1	0	0	1	1	0	0	0	1
g4	1	1	1	0	1	1	0	0	1	1	1	0
g5	1	1	1	0	0	1	0	1	1	0	1	0
g1	0	0	1	1	0	0	0	1	0	0	0	1
g2	1	1	0	0	1	1	0	1	0	1	1	1
g3	0	0	0	0	1	0	0	1	1	1	0	0
g4	1	0	1	1	1	0	1	0	0	1	0	1

Step 2

Data Mining



model, patterns, …

Relational Data Mining through Propositionalization

customer							
ID	Zip	Sex	SoSt	Income	Age	Club	Resp
...	...	...	...	...	...	...	...
3478	34677	m	si	60-70	32	me	nr
3479	43666	f	ma	80-90	45	nm	re
...	...	...	...	...	...	...	...

order				
Customer ID	Order ID	Store ID	Delivery Mode	Paymt Mode
...	...	...	...	...
3478	2140267	12	regular	cash
3478	3446778	12	express	check
3478	4728386	17	regular	check
3479	3233444	17	express	credit
3479	3475886	12	regular	credit
...	...	...	...	...

store			
Store ID	Size	Type	Location
...	...	...	...
12	small	franchise	city
17	large	indep	rural
...	...	...	...

Relational representation of customers, orders and stores.

Step 1

Propositionalization

	f1	f2	f3	f4	f5	f6	...				...	fn
g1	1	0	0	1	1	1	0	0	1	0	1	1
g2	0	1	1	0	1	1	0	0	0	1	1	0
g3	0	1	1	1	0	0	1	1	0	0	0	1
g4	1	1	1	0	1	1	0	0	1	1	1	0
g5	1	1	1	0	0	1	0	1	1	0	1	0
g1	0	0	1	1	0	0	0	1	0	0	0	1
g2	1	1	0	0	1	1	0	1	0	1	1	1
g3	0	0	0	0	1	0	0	1	1	1	0	0
g4	1	0	1	1	1	0	1	0	0	1	0	1

	f1	f2	f3	f4	f5	f6	...				...	fn
g1	1	0	0	1	1	1	0	0	1	0	1	1
g2	0	1	1	0	1	1	0	0	0	1	1	0
g3	0	1	1	1	0	0	1	1	0	0	0	1
g4	1	1	1	0	1	1	0	0	1	1	1	0
g5	1	1	1	0	0	1	0	1	1	0	1	0
g1	0	0	1	1	0	0	0	1	0	0	0	1
g2	1	1	0	0	1	1	0	1	0	1	1	1
g3	0	0	0	0	1	0	0	1	1	1	0	0
g4	1	0	1	1	1	0	1	0	0	1	0	1

Step 2

Data Mining

```
target(A) :-
    'Doctor'(A), 'Italy'(A).

target(A) :-
    'Public'(A), 'Gold'(A).

target(A) :-
    'Poland'(A), 'Deposit'(A), 'Gold'(A).

target(A) :-
    'Germany'(A), 'Insurance'(A).

target(A) :-
    'Service'(A), 'Germany'(A).
```

patterns (set of rules)

RSD Lessons learned

Efficient propositionalization can be applied to individual-centered, multi-instance learning problems:

- one free global variable (denoting an individual, e.g. molecule M)
- one or more structural predicates: (e.g. has_atom(M,A)), each introducing a new existential local variable (e.g. atom A), using either the global variable (M) or a local variable introduced by other structural predicates (A)
- one or more utility predicates defining properties of individuals or their parts, assigning values to variables

feature121(M):- hasAtom(M,A), atomType(A,21)

feature235(M):- lumo(M,Lu), lessThr(Lu,-1.21)

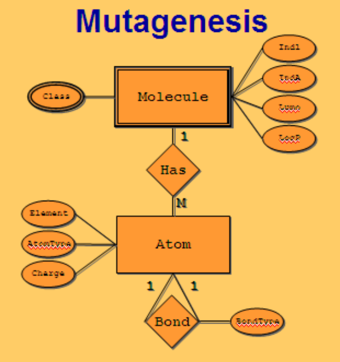
mutagenic(M):- feature121(M), feature235(M)

Relational Data Mining in Orange4WS

- service for propositionalization through efficient first-order feature construction (Železny and Lavrač, MLJ 2006)

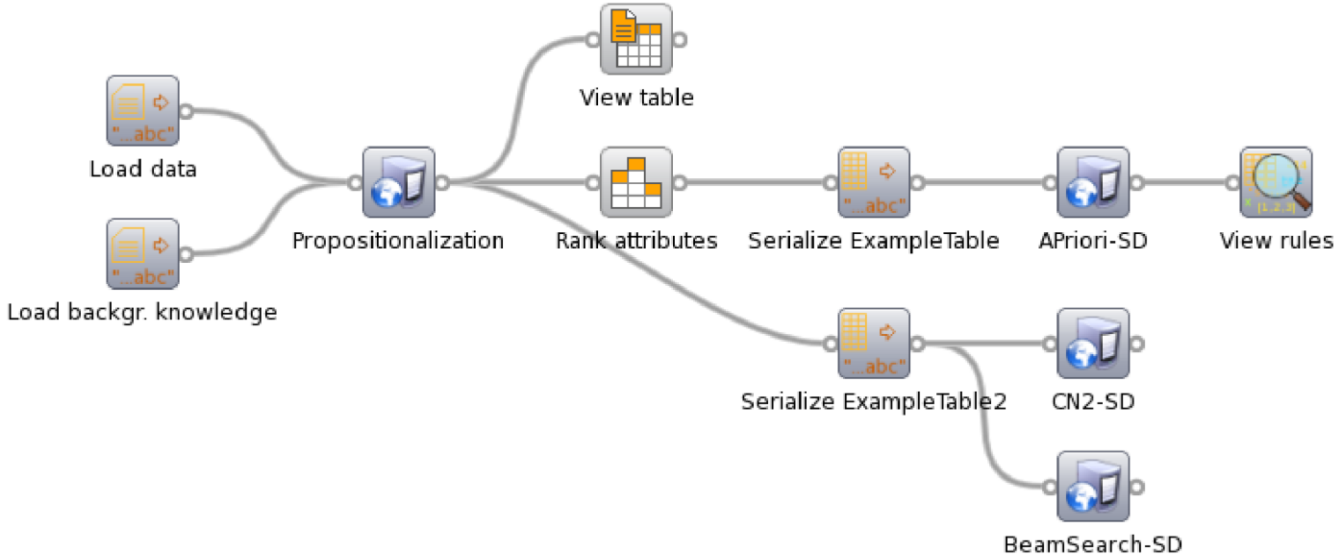
  f121(M):- hasAtom(M,A), atomType(A,21)

  f235(M):- lumo(M,Lu), lessThr(Lu,1.21)



- subgroup discovery using CN2-SD

  mutagenic(M) ← feature121(M), feature235(M)



Talk outline

- Subgroup discovery in a nutshell
- Relational data mining and propositionalization in a nutshell
- Semantic data mining: Using ontologies in SD

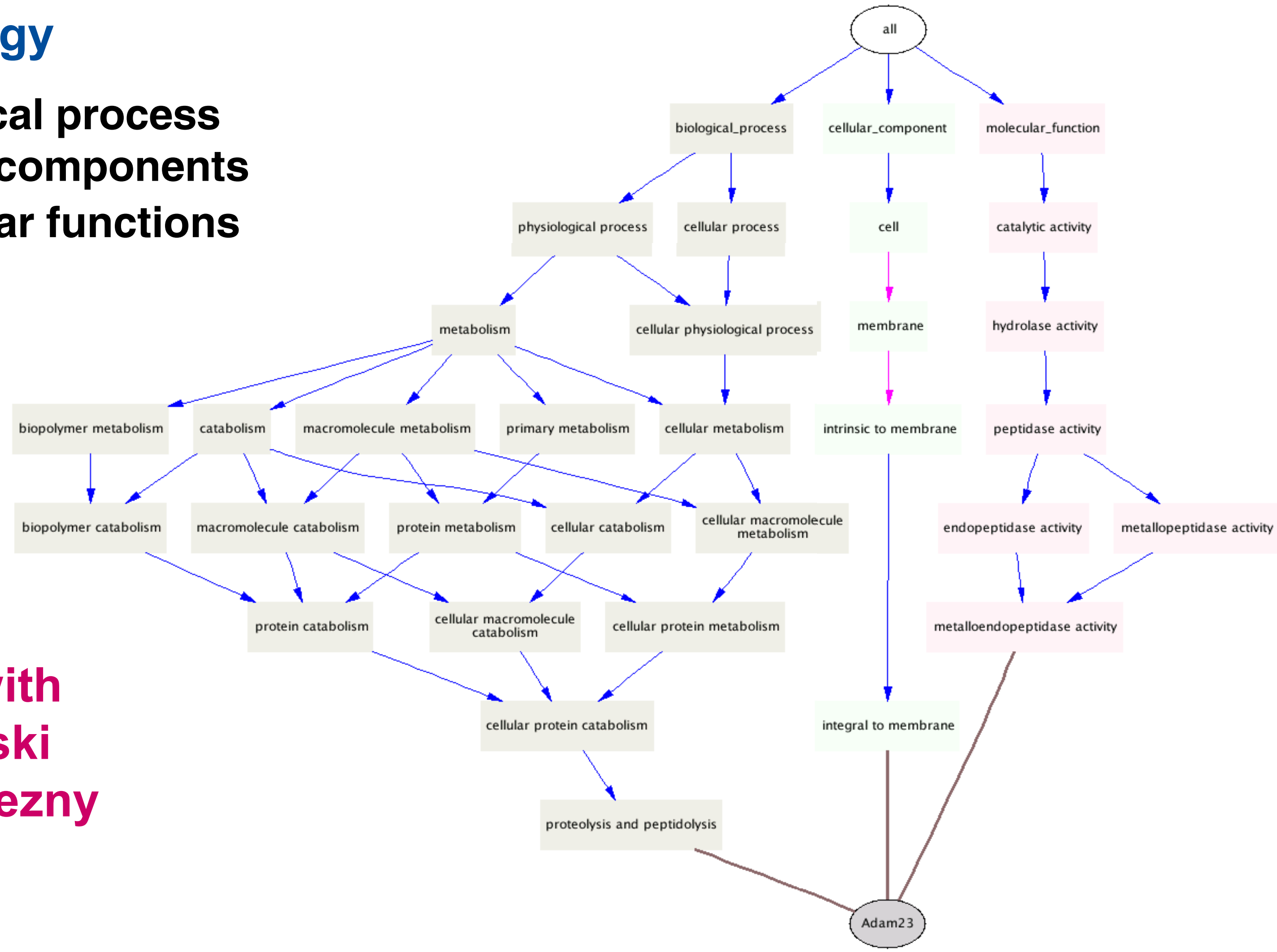
Semantic Data Mining in Orange4WS

- Exploiting semantics in data mining
  - Using **domain ontologies** as background knowledge for data mining
- Semantic data mining technology: a two-step approach
  - Using propositionalization through first-order feature construction
  - Using subgroup discovery for rule learning
- Implemented in the SEGS algorithm

Using domain ontologies (e.g. Gene Ontology) as background knowledge for Data Mining

Gene Ontology

12093 biological process
1812 cellular components
7459 molecular functions

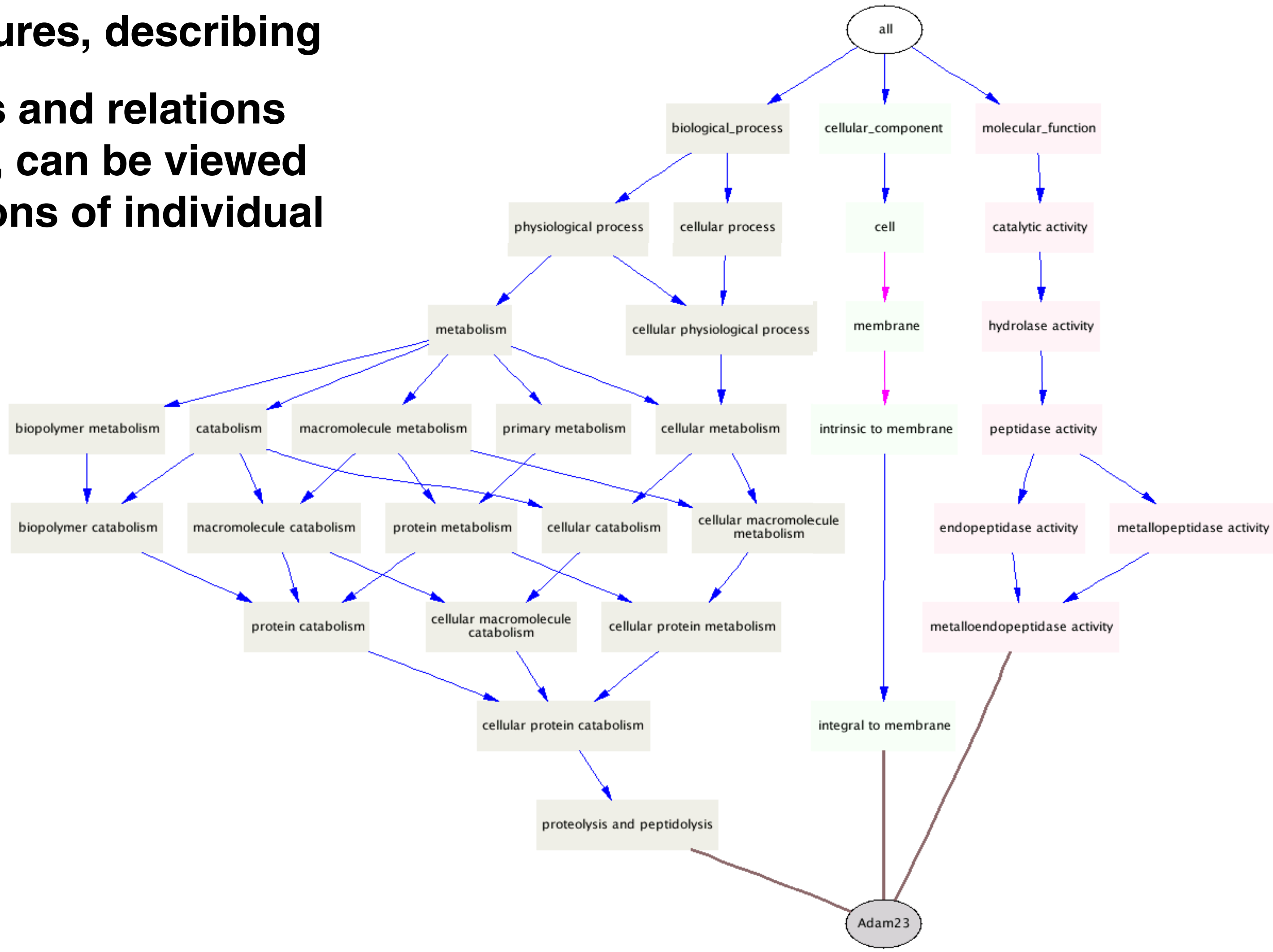


Joint work with Igor Trajkovski and Filip Zelezny

Using domain ontologies (e.g. Gene Ontology) as background knowledge for Data Mining

First-order features, describing

gene properties and relations between genes, can be viewed as generalisations of individual genes



First order feature construction

First order features with support > *min_support*

f(7,A):-function(A,'GO:0046872').
f(8,A):-function(A,'GO:0004871').
f(11,A):-process(A,'GO:0007165').
f(14,A):-process(A,'GO:0044267').
f(15,A):-process(A,'GO:0050874').
f(20,A):-function(A,'GO:0004871'), process(A,'GO:0050874').
f(26,A):-component(A,'GO:0016021').
f(29,A):- function(A,'GO:0046872'), component(A,'GO:0016020').
f(122,A):-interaction(A,B),function(B,'GO:0004872').

existential

f(223,A):-interaction(A,B),function(B,'GO:0004871'),
process(B,'GO:0009613').
f(224,A):-interaction(A,B),function(B,'GO:0016787'),
component(B,'GO:0043231').

Propositionalization

diffexp g1 (gene64499)
diffexp g2 (gene2534)
diffexp g3 (gene5199)
diffexp g4 (gene1052)
diffexp g5 (gene6036)
....

random g1 (gene7443)
random g2 (gene9221)
random g3 (gene2339)
random g4 (gene9657)
random g5 (gene19679)
....

	f1	f2	f3	f4	f5	f6	…				…	fn
g1	1	0	0	1	1	1	0	0	1	0	1	1
g2	0	1	1	0	1	1	0	0	0	1	1	0
g3	0	1	1	1	0	0	1	1	0	0	0	1
g4	1	1	1	0	1	1	0	0	1	1	1	0
g5	1	1	1	0	0	1	0	1	1	0	1	0
g1	0	0	1	1	0	0	0	1	0	0	0	1
g2	1	1	0	0	1	1	0	1	0	1	1	1
g3	0	0	0	0	1	0	0	1	1	1	0	0
g4	1	0	1	1	1	0	1	0	0	1	0	1

Propositional learning: subgroup discovery

	f1	f2	f3	f4	f5	f6	…				…	fn
g1	1	0	0	1	1	1	0	0	1	0	1	1
g2	0	1	1	0	1	1	0	0	0	1	1	0
g3	0	1	1	1	0	0	1	1	0	0	0	1
g4	1	1	1	0	1	1	0	0	1	1	1	0
g5	1	1	1	0	0	1	0	1	1	0	1	0
g1	0	0	1	1	0	0	0	1	0	0	0	1
g2	1	1	0	0	1	1	0	1	0	1	1	1
g3	0	0	0	0	1	0	0	1	1	1	0	0
g4	1	0	1	1	1	0	1	0	0	1	0	1

Over-expressed

IF

f2 and f3

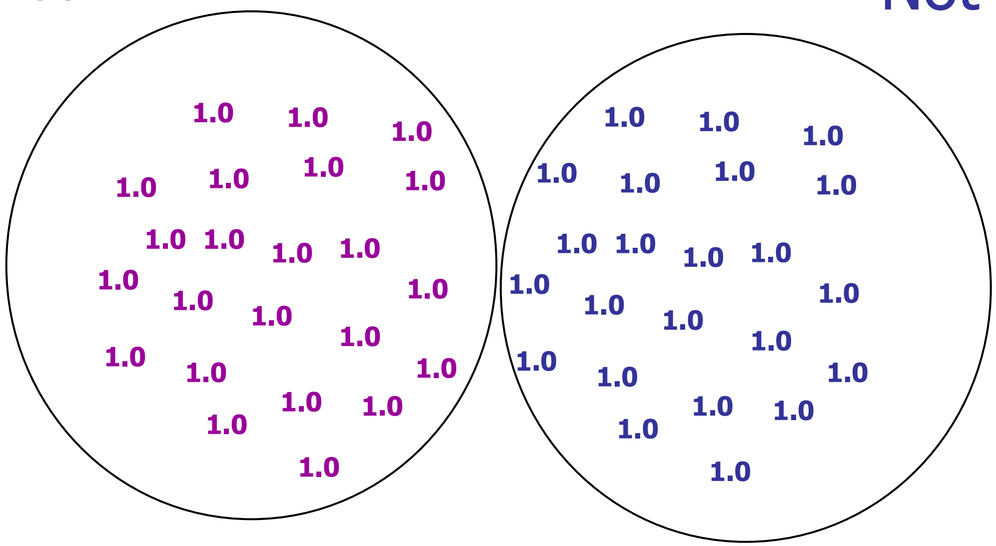
[4,0]

diffexp(A) :- interaction(A,B) & function(B,'GO:0004871')

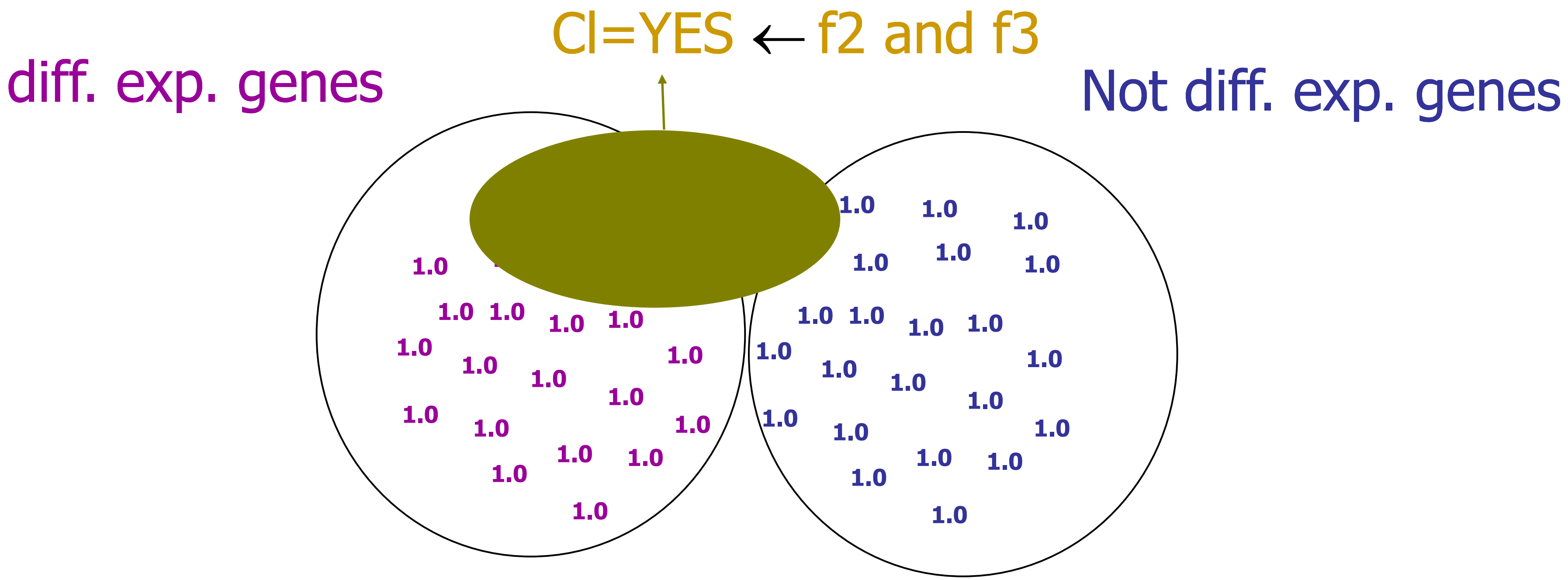
Subgroup Discovery

diff. exp. genes

Not diff. exp. genes



Subgroup Discovery



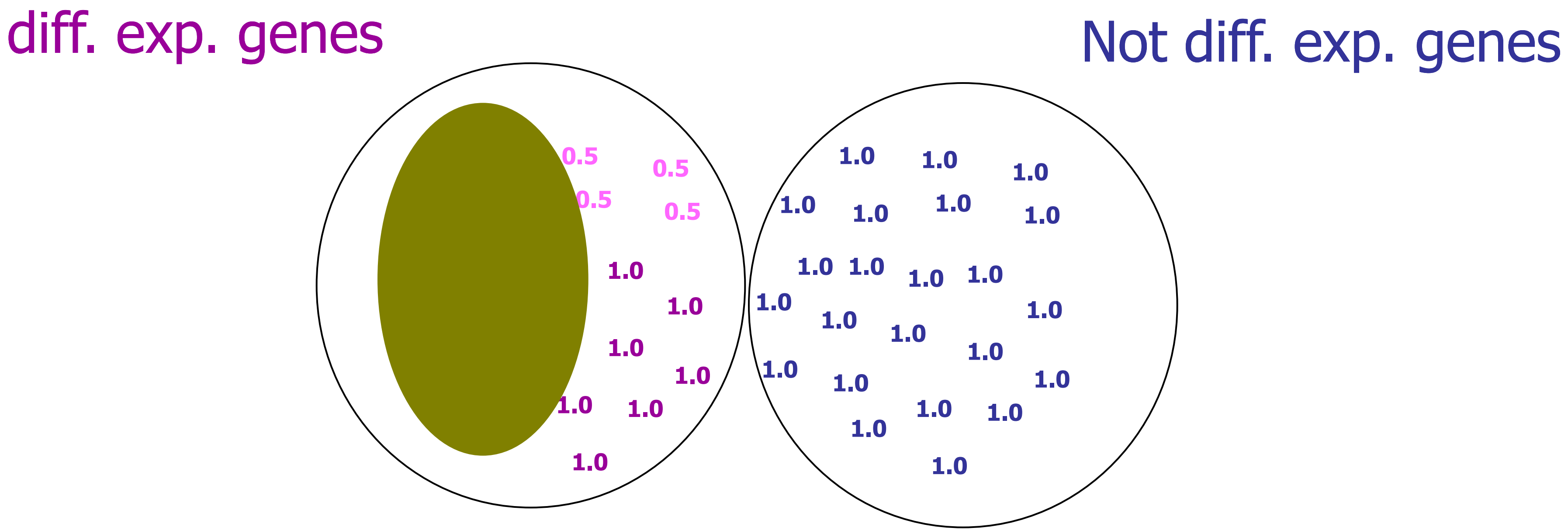
In RSD (using propositional learner CN2-SD):

Quality of the rules = Coverage x Precision

*Coverage = sum of the covered weights

*Precision = purity of the covered genes

Subgroup Discovery



RSD naturally uses gene weights in its procedure for repetitive subgroup generation, via its heuristic rule evaluation: weighted relative accuracy

Semantic Data Mining in two steps

- **Step 1: Construct relational logic features** of genes such as

  interaction(g, G) & function(G, protein_binding)

  (*g interacts with another gene whose functions include protein binding*)

  and **propositional table construction** with features as attributes
- **Step 2:** Using these features to **discover and describe subgroups of genes** that are differentially expressed (e.g., belong to class DIFF.EXP. of top 300 most differentially expressed genes) in contrast with RANDOM genes (randomly selected genes with low differential expression).
- Sample subgroup description:

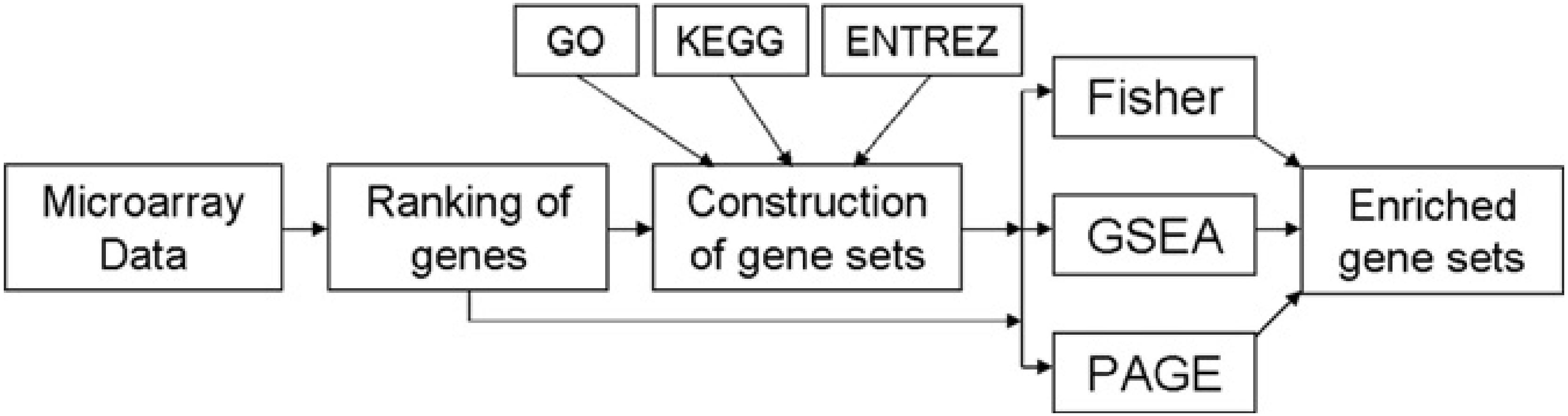
```
diffexp(A) :- interaction(A,B) AND
              function(B,'GO:0004871') AND
              process(B,'GO:0009613')
```

Summary: SEGS, using the RSD approach

- The SEGS approach enables to discover new medical knowledge from the combination of gene expression data with public gene annotation databases
- The SEGS approach proved effective in several biomedical applications (JBI 2008, …)
  - The work on semantic data mining - using ontologies as background knowledge for subgroup discovery with SEGS - was done in collaboration with I.Trajkovski, F. Železny and J. Tolar
- Recent work: Semantic subgroup discovery implemented in Orange4WS

Semantic subgroup discovery with SEGS

- SEGS workflow is implemented in the Orange4WS data mining environment

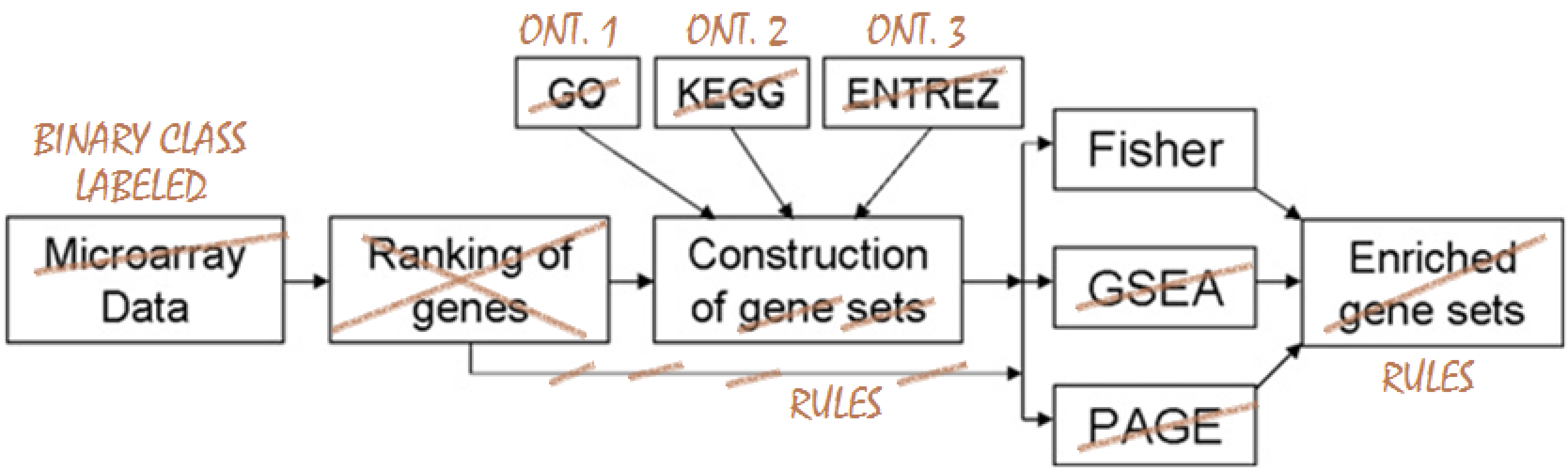


- SEGS is also implemented also as a Web applications

(Trajkovski et al., IEEE TSMC 2008, Trajkovski et al., JBI 2008)

From SEGS to SDM-SEGS: Generalizing SEGS

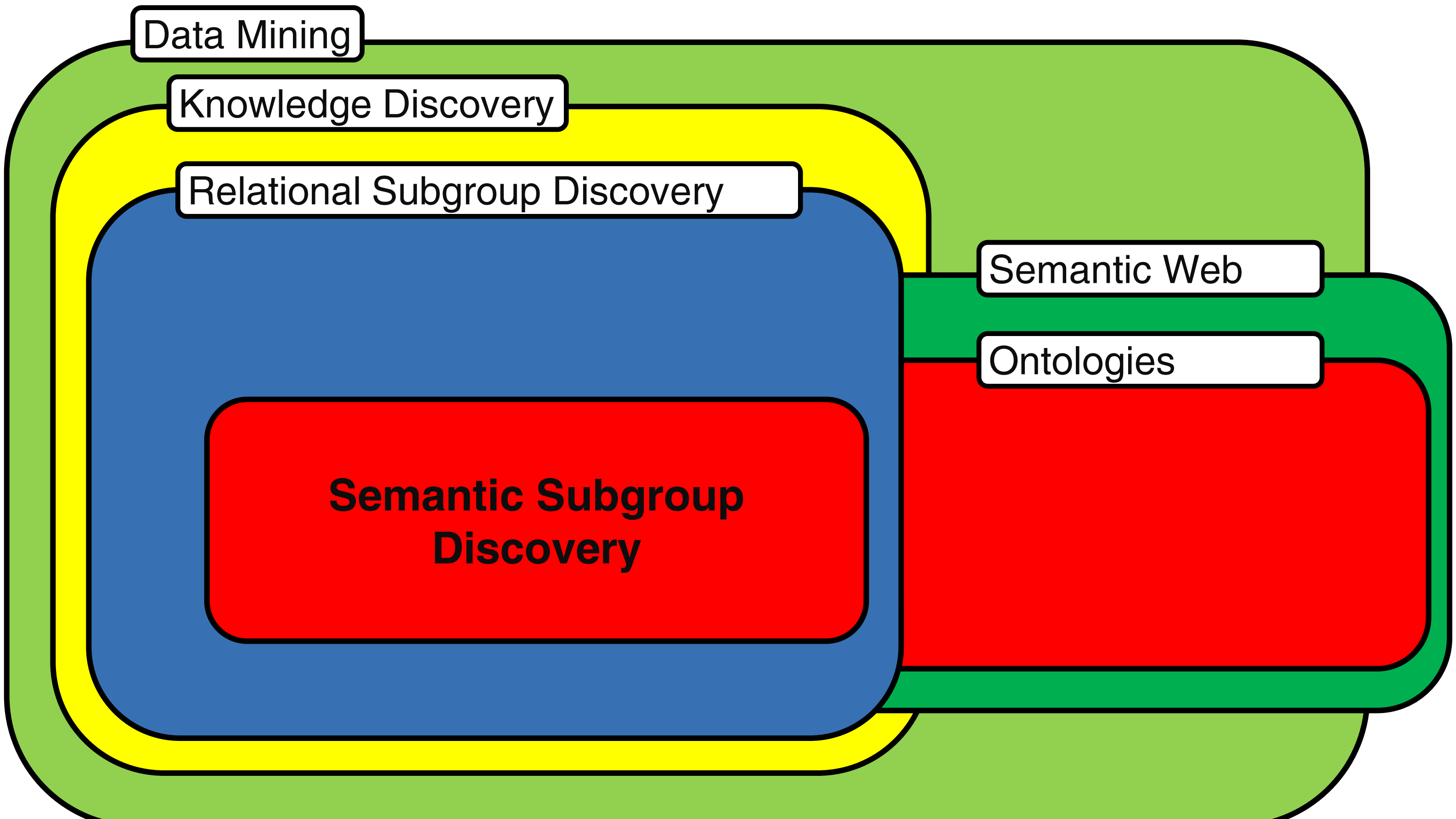
- SDM-SEGS: a general semantic data mining



- Discovers subgroups both for ranked and labeled data
- Exploits input ontologies in OWL format
- Is also implemented in Orange4WS

Semantic Data Mining

- Semantic subgroup discovery (Vavpetič et al., 2012)

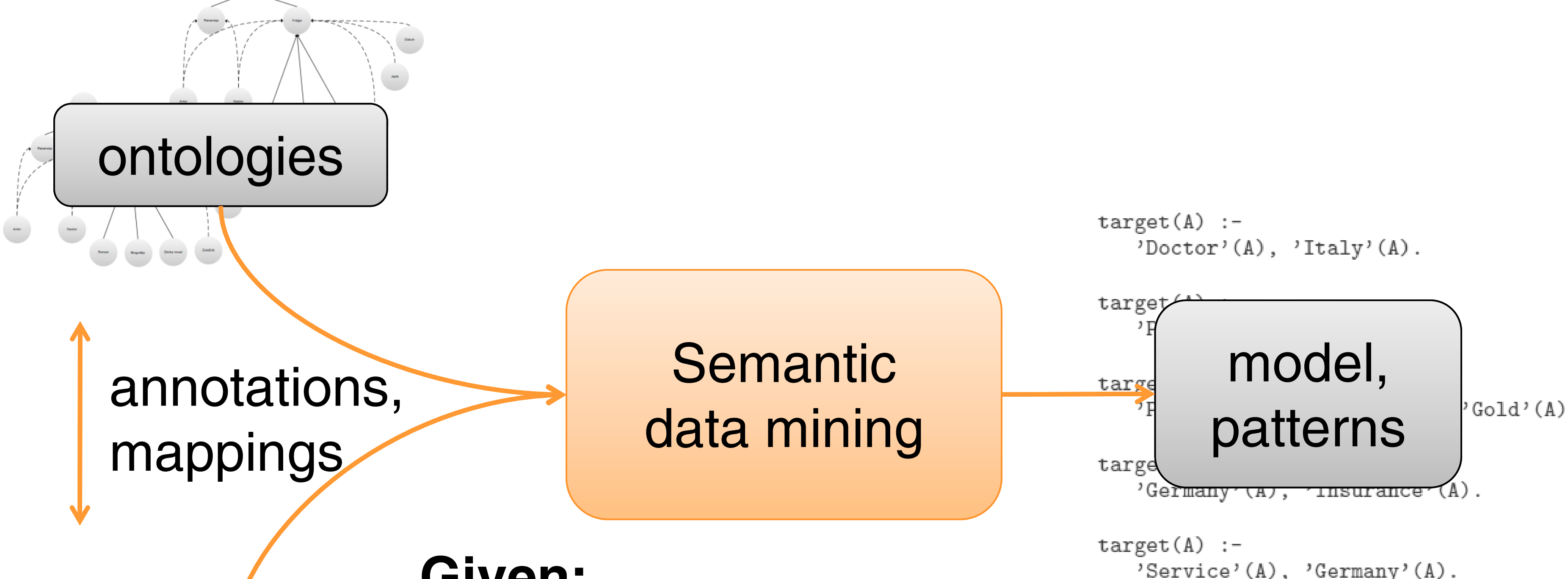


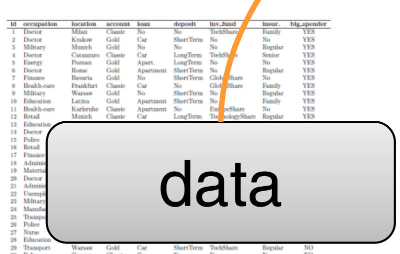
What is Semantic Data Mining

- Ontology-driven (semantic) data mining is an emerging research topic – the topic of this tutorial
- Semantic Data Mining (SDM) - a new term denoting:
  – the new challenge of mining semantically annotated resources, with ontologies used as background knowledge to data mining
  – approaches with which semantic data are mined

What is Semantic Data Mining

SDM task definition





Given:

- transaction data table, relational database, text documents, Web pages, …
- one or more domain ontologies

Find: a classification model, a set of patterns

Introductory seminar lecture

X. JSI & Knowledge Technologies

I. Introduction

- Data Mining and KDD process
- DM standards, tools and visualization
- Classification of Data Mining techniques: Predictive and descriptive DM

  (Mladenić et al. Ch. 1 and 11, Kononenko & Kukar Ch. 1)

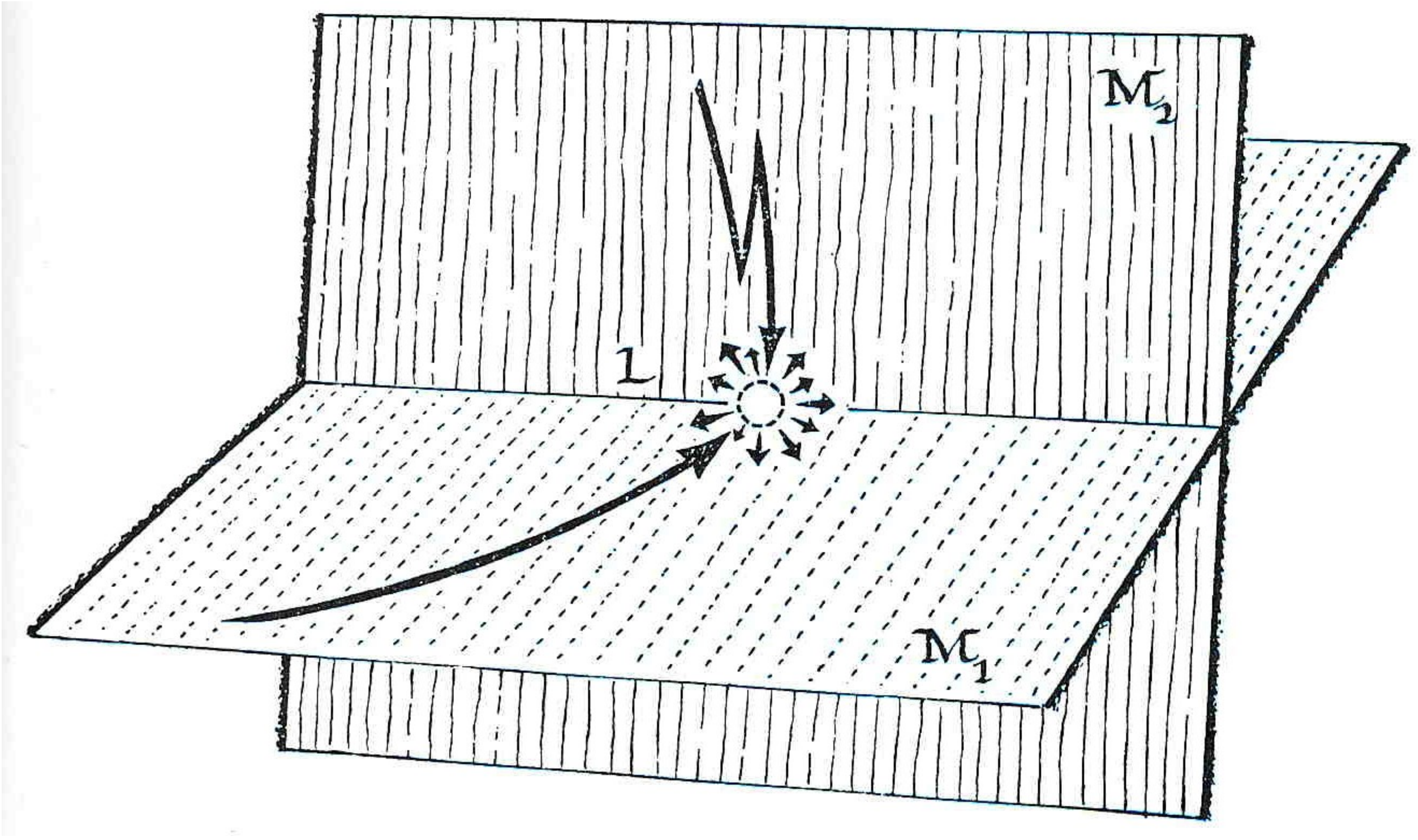
XX. Selected data mining techniques: Advanced subgroup discovery techniques and applications

XXX. Recent advances: Cross-context link discovery

The BISON project

- EU project: Bisociation networks for creative information discovery (www.bisonet.eu), 2008-2010
- Exploring the idea of bisociation (Arthur Koestler, The act of creation, 1964):
  - The mixture - in one human mind – of **two different contexts** or **different categories of objects,** that are normally considered **separate categories** by the processes of the mind.
  - The **thinking process** that is the functional basis of **analogical or metaphoric thinking** as compared to logical or associative thinking.
- Main challenge: Support humans to find **new interesting associations accross domains**

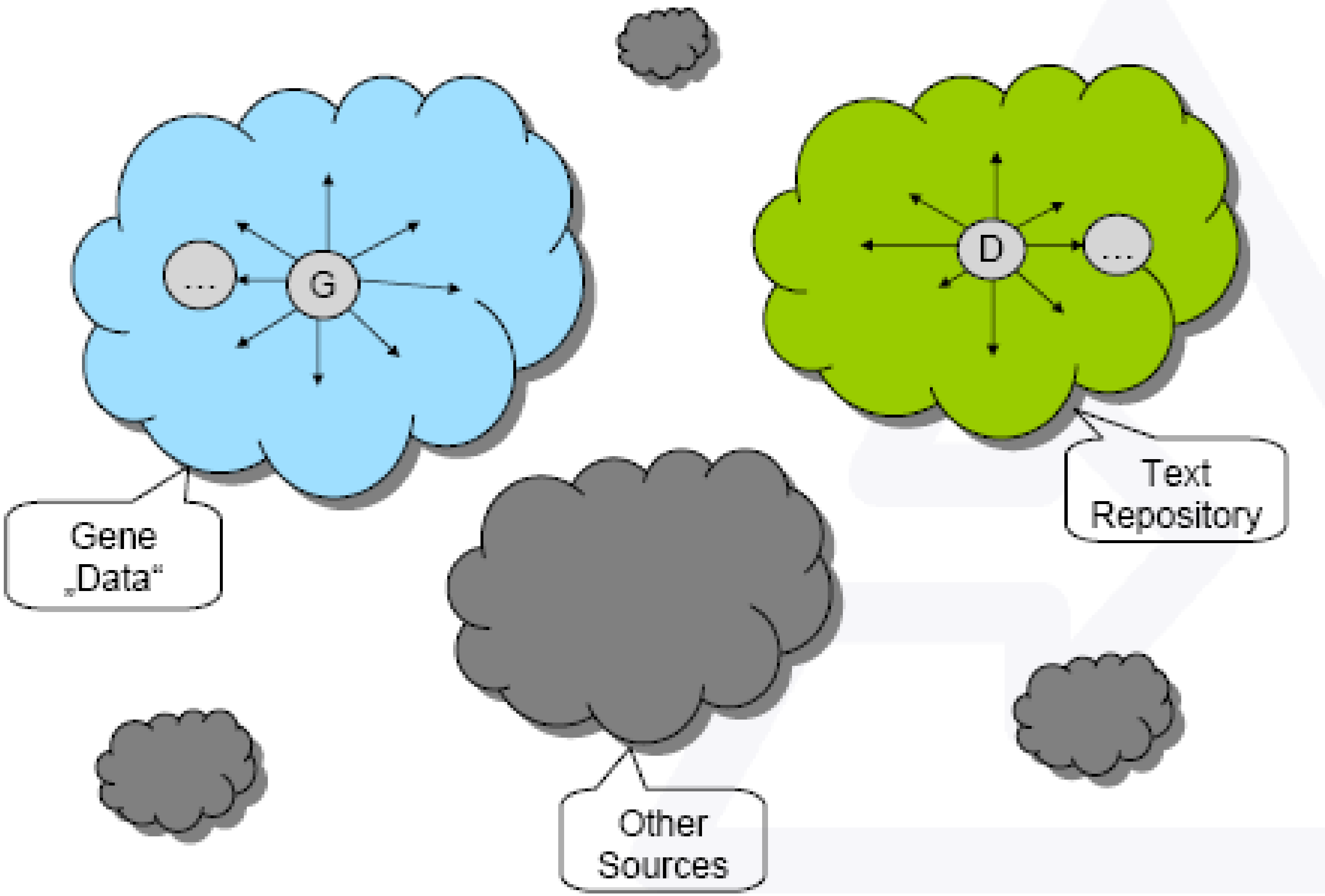
Bisociation (A. Koestler 1964)



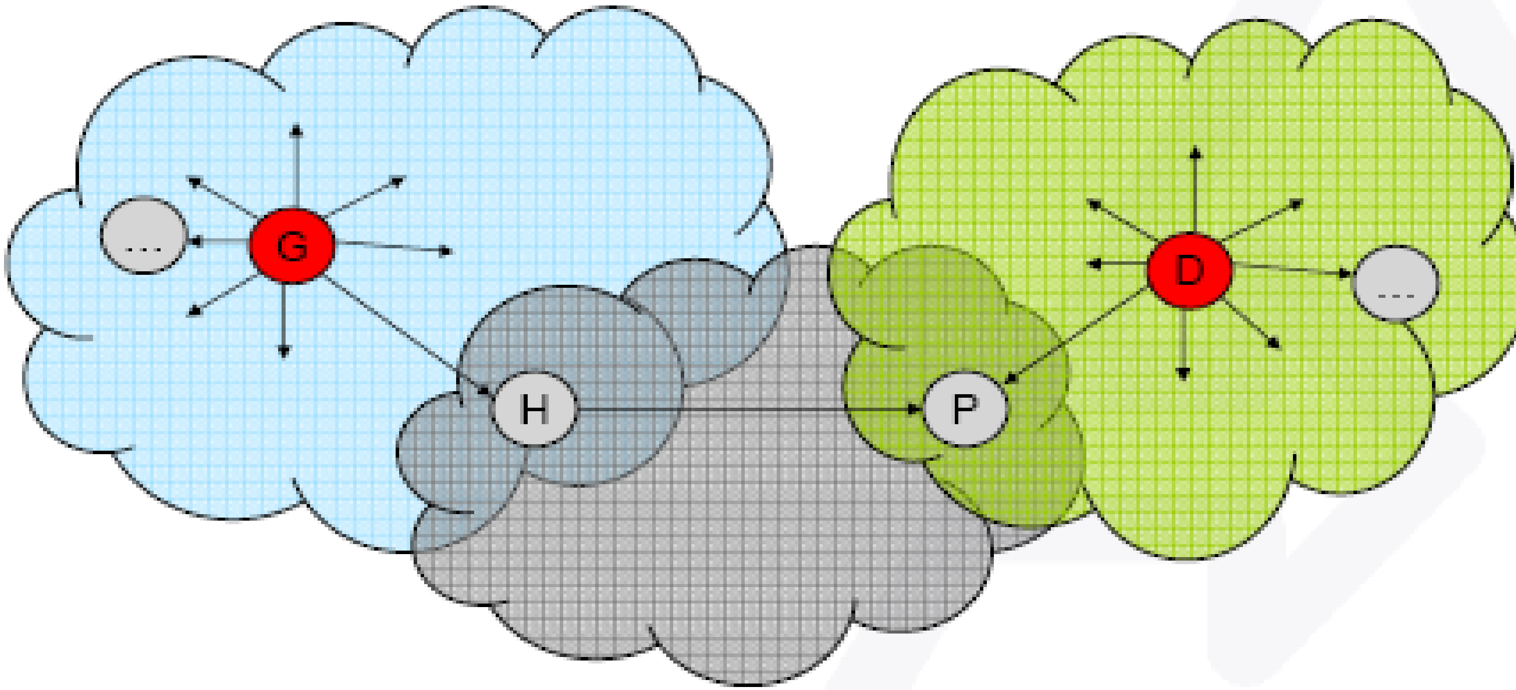
The BISON project

- BISON challenge: Support humans to find **new, interesting links** **accross domains,** named **bisociations**
  - across different contexts
  - across different types of data and knowledge sources
- Open problems:
  - Fusion of heterogeneous data/knowledge sources into a joint representation format - a large information network named BisoNet (consisting of nodes and relatioships between nodes)
  - Finding unexpected, previously unknown links between BisoNet nodes belonging to different contexts

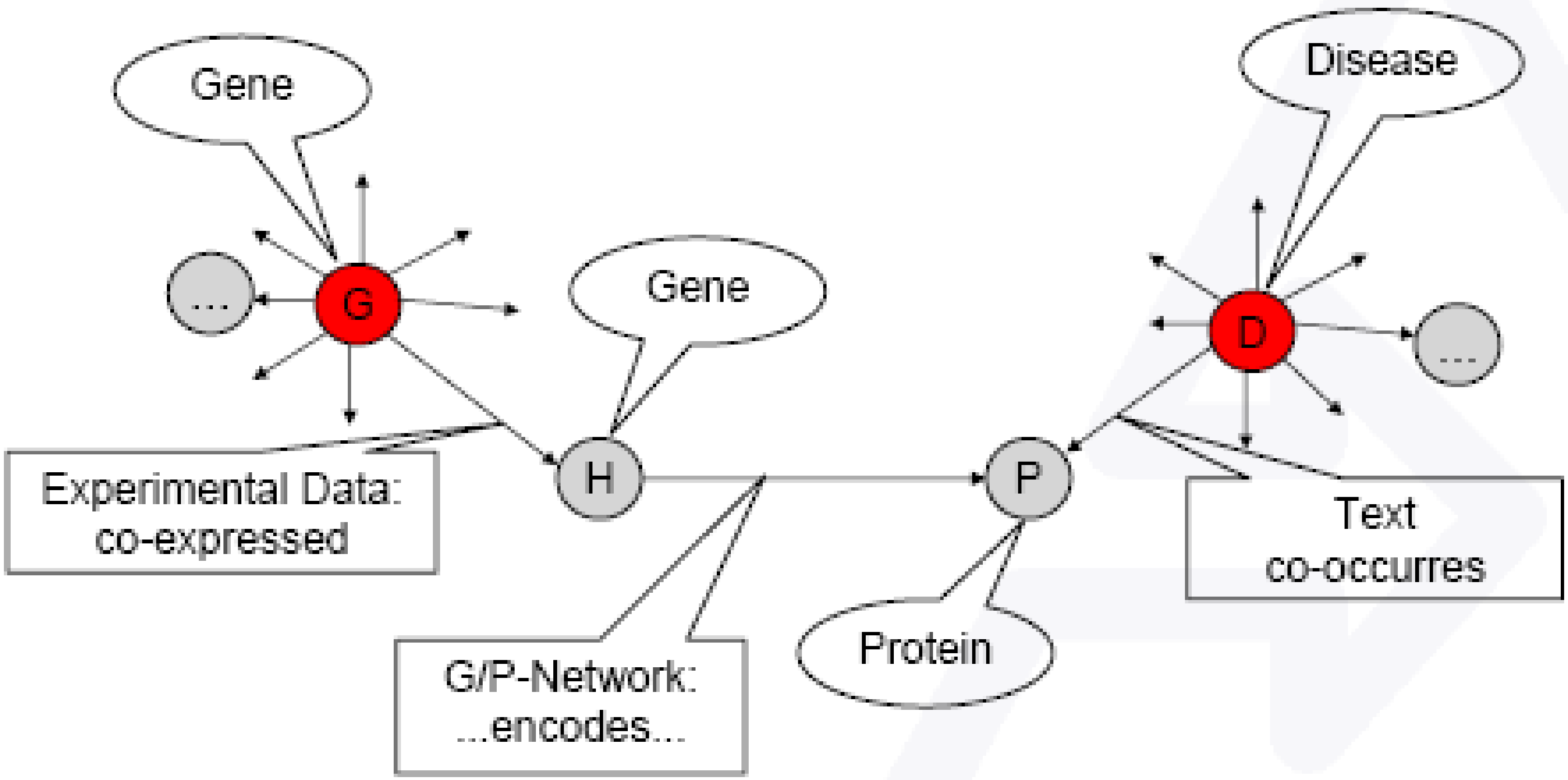
Heterogeneous data sources (BISON, M. Berthold, 2008)



Bridging concepts (BISON, M. Berthold, 2008)



Chains of associations across domains (BISON, M. Berthold, 2008)



Semantic Data Mining for DNA Microarray Data Analysis

- Semantic data mining integrates public gene annotation data through relational features
- It is implemented in the SEGS algorithm (Trajkovski, Železny, Lavrač and Tolar, JBI 2008), available in Orange4WS
- It can be combined with additional biomedical resources (BioMine), providing additional means for creative knowledge discovery from publicly available data sources

Biomine graph exploration (Toivonnen et al., Uni. Helsinki)

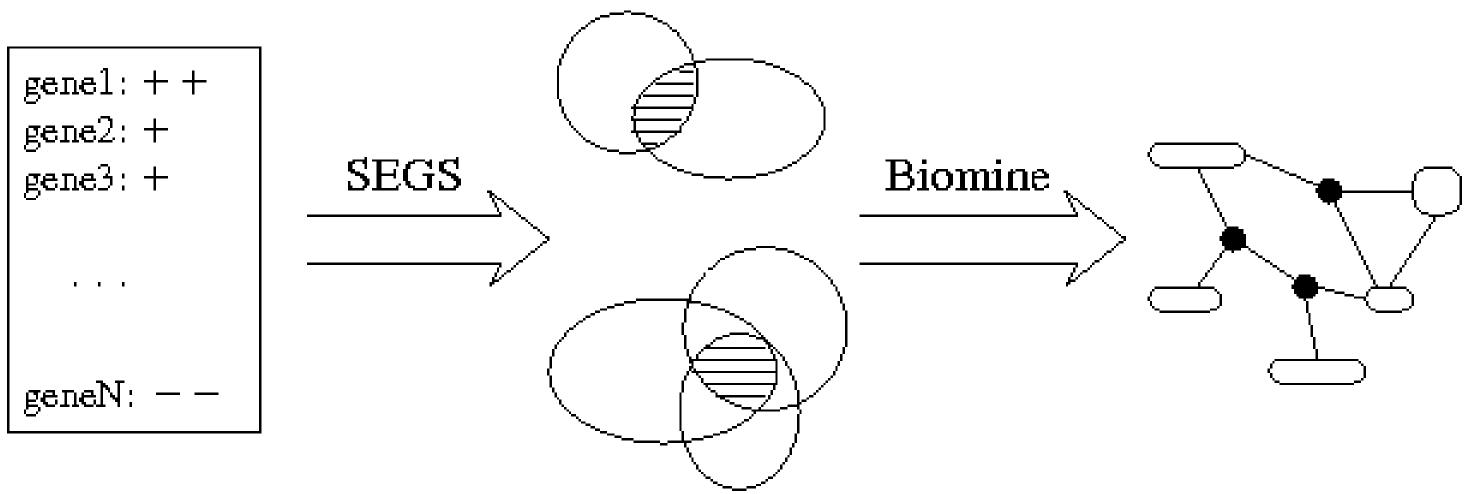
- **BioMine graph** contains information from public databases, including annotated sequences, proteins, orthology groups, genes and gene expressions, gene and protein interactions, PubMed articles, and different ontologies.
  - **nodes (~1 mio)** correspond to different concepts (such as gene, protein, domain, phenotype, biological process, tissue)
  - **semantically labeled edges (~7 mio)** connect related concepts
- **BioMine query engine** answers queries to potentially discover new links between entities by sophisticated graph exploration algorithms

The SEGS + BioMine Methodology

Microarray

Gene sets

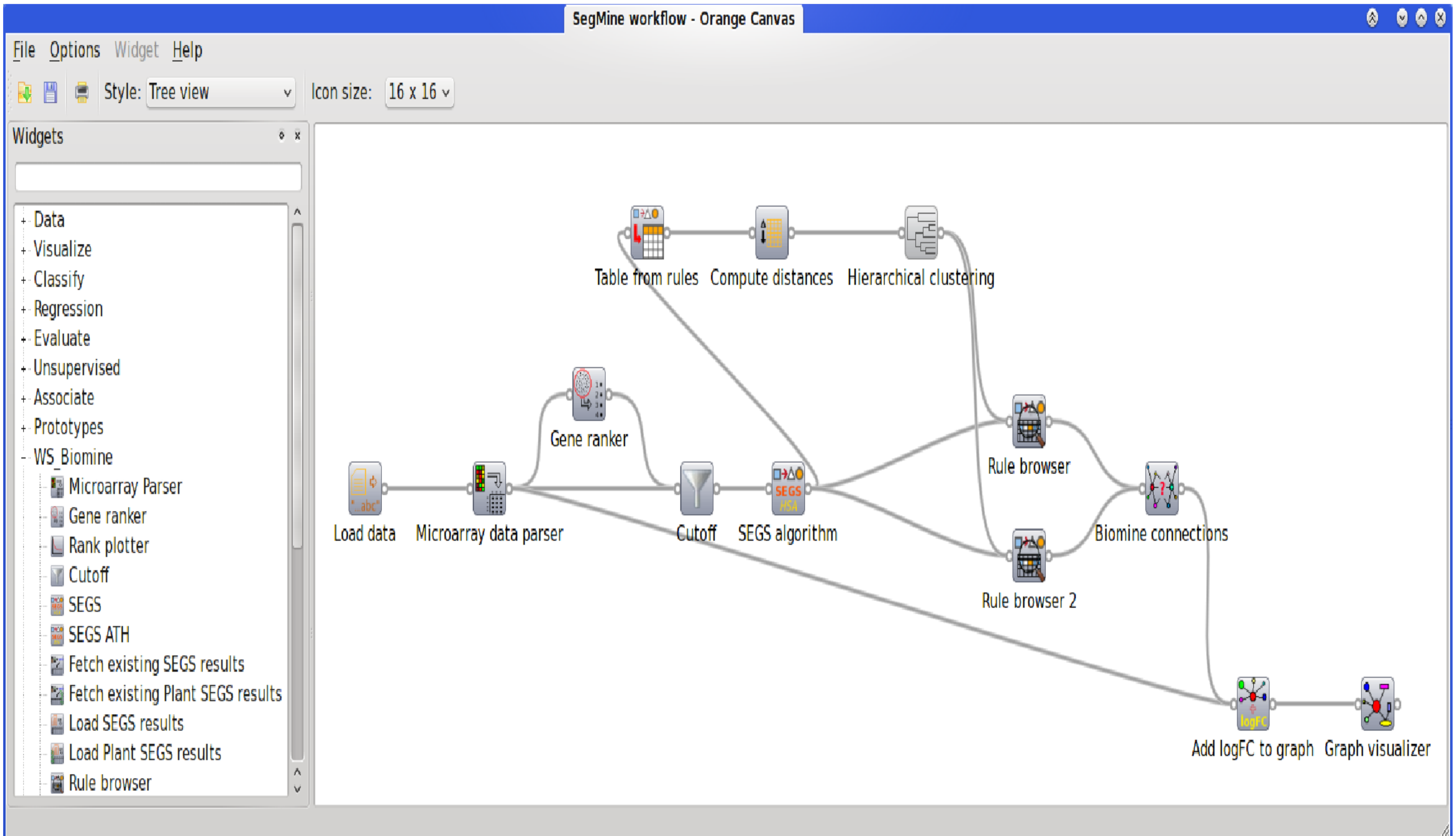
Exploratory
link discovery



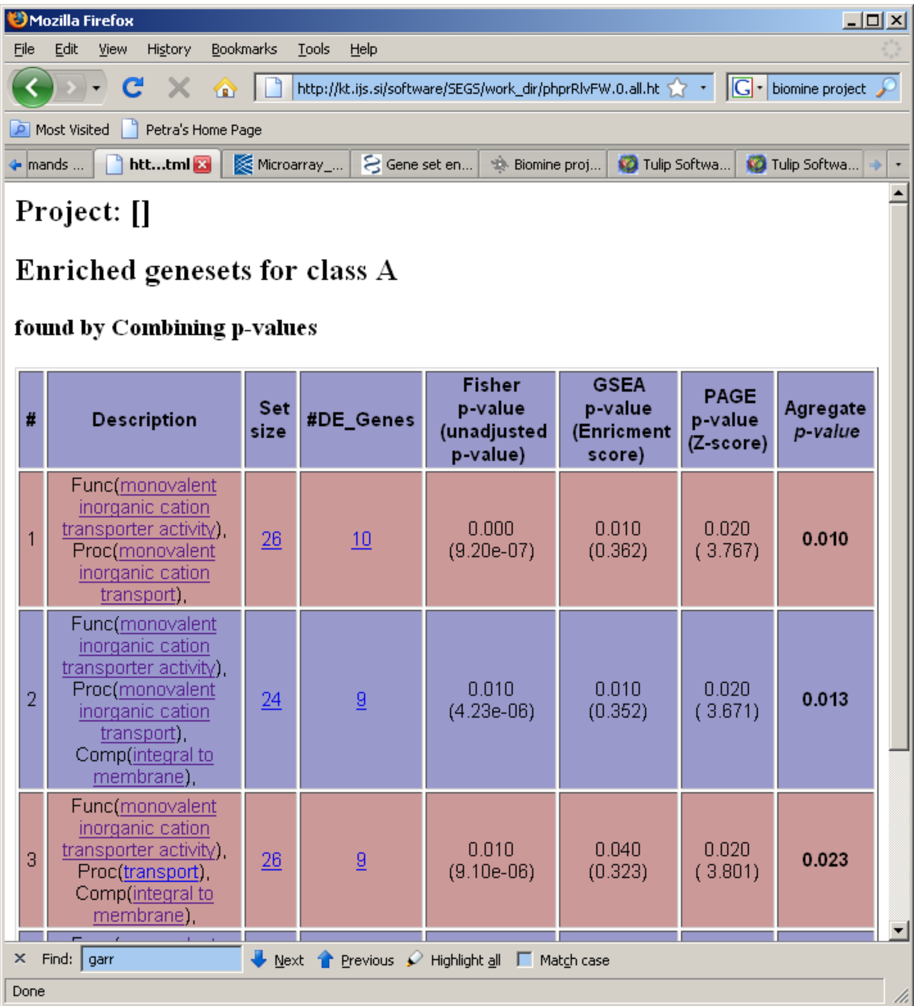
e.g. slow-vs-fast
cell growth

Work by
Lavrač et al. 2009, 2010
Podpečan et al. 2010

Semantic Data Mining in Orange4WS: SEGS + BioMine workflow implementation



SEGS output:



Mozilla Firefox

http://kt.ijs.si/software/SEGS/work_dir/phprRlvFW.0.all.ht

Project: []

Enriched genesets for class A

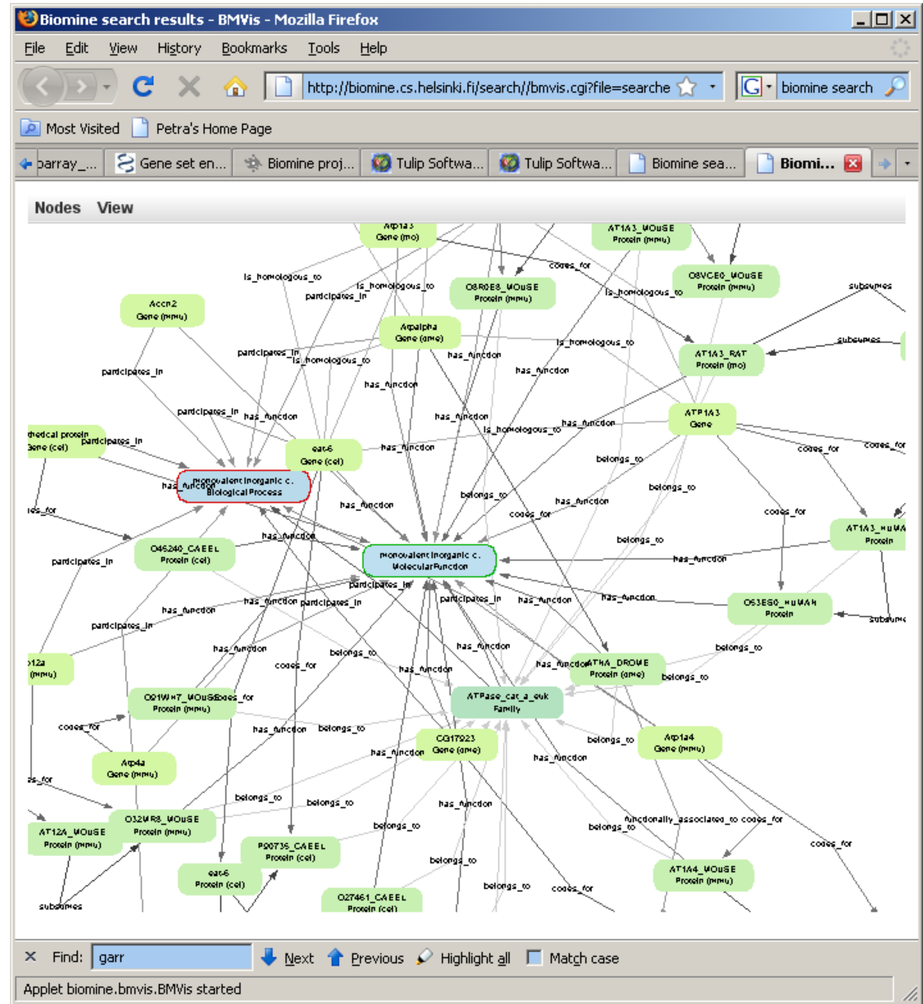
found by Combining p-values

#	Description	Set size	#DE_Genes	Fisher p-value (unadjusted p-value)	GSEA p-value (Enricment score)	PAGE p-value (Z-score)	Agregate *p-value*
1	Func(monovalent inorganic cation transporter activity), Proc(monovalent inorganic cation transport),	26	10	0.000 (9.20e-07)	0.010 (0.362)	0.020 (3.767)	**0.010**
2	Func(monovalent inorganic cation transporter activity), Proc(monovalent inorganic cation transport), Comp(integral to membrane),	24	9	0.010 (4.23e-06)	0.010 (0.352)	0.020 (3.671)	**0.013**
3	Func(monovalent inorganic cation transporter activity), Proc(transport), Comp(integral to membrane),	26	9	0.010 (9.10e-06)	0.040 (0.323)	0.020 (3.801)	**0.023**

Find: garr

Done

BioMine output:



Summary of SEGS + BioMine

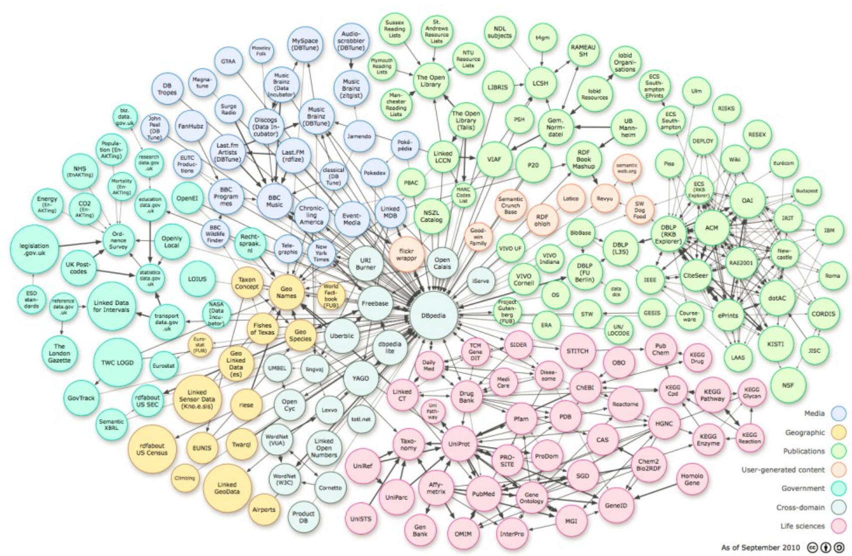
- Semantic Data Mining algorithm SEGS discovers interesting gene group descriptions as conjunctions of concepts from three ontologies: GO, KEGG and Entrez
- Biomine finds cross-context links (paths) between concepts discovered by SEGS, using other ontologies, PubMed and other biomedical resources
- Initial results in stem cell microarray data analysis (EMBC 2009) indicate that the SEGS+Biomine methodology may lead to new insights – in vitro experiments are in progress at NIB to verify and validate the preliminary insights
- A general purpose Semantic Data Mining algorithm g-SEGS is also available in Orange4WS
- New developments concern SDM implementation in ClowdFlows

Future work

- Current Semantic data mining scenario: Mining empirical data with ontologies as background knowledge
    - abundant empirical data, but
    - scarce background knowledge
- Future Semantic data mining scenario:
    - envisioning a growing amount of semantic data
        - abundance of ontologies and semantically anotated data collections
        - e.g. Linked Data
            - over 6 billion RDF triples
            - over 148 million links

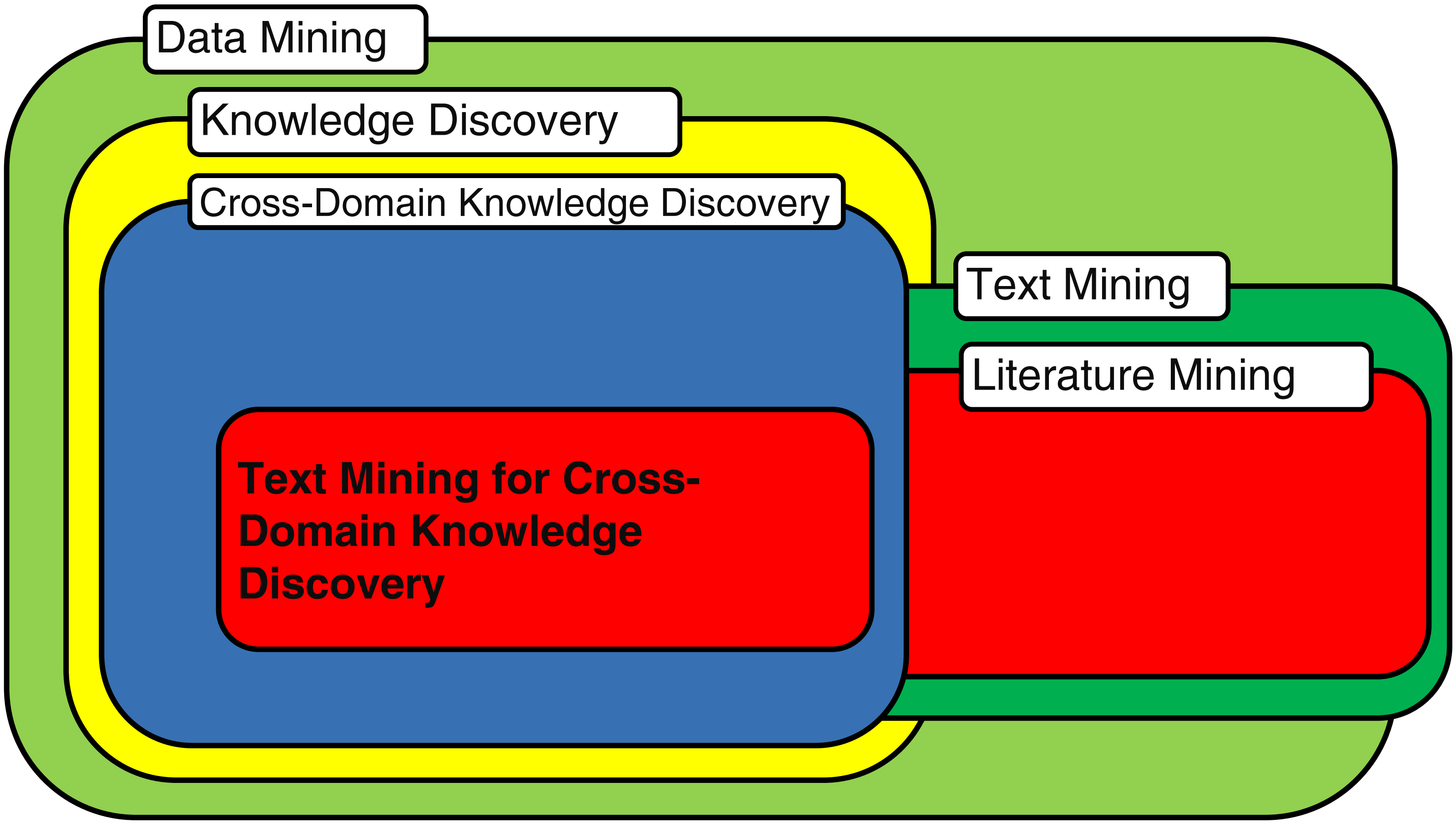
Future work

- We may envision a paradigm shift from data mining to knowledge mining
- The envisioned future Semantic data mining scenario in mining the Semantic Web:
  - mining knowledge encoded in domain ontologies,
  - constrained by annotated (empirical) data collections.



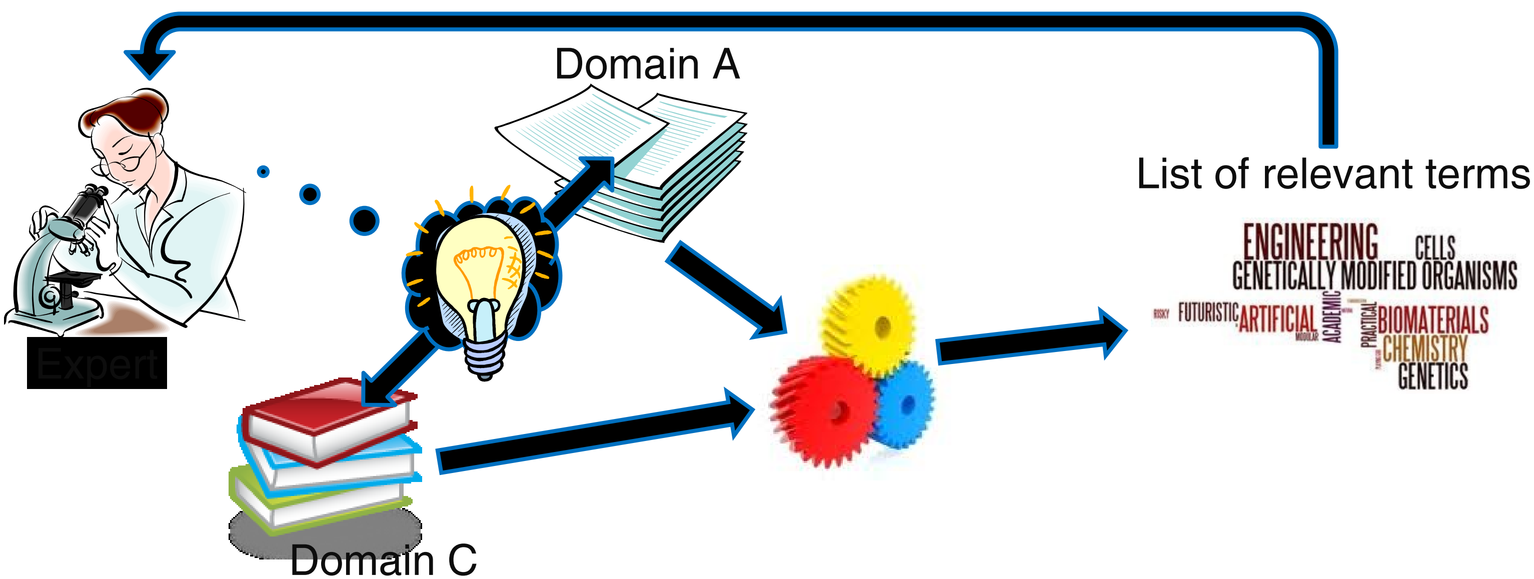
Current and future work

- Cross-domain literature mining: Fidning bridging concepts with CrossBee (Juršič et al., 2012)



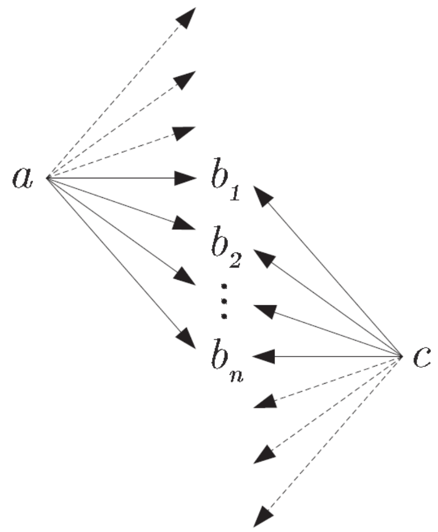
Current and future work

- Cross-domain literature mining: Fidning bridging concepts with CrossBee (Juršič et al., 2012)
  - Help experts in cross-domain bisociative discovery for unknown facts



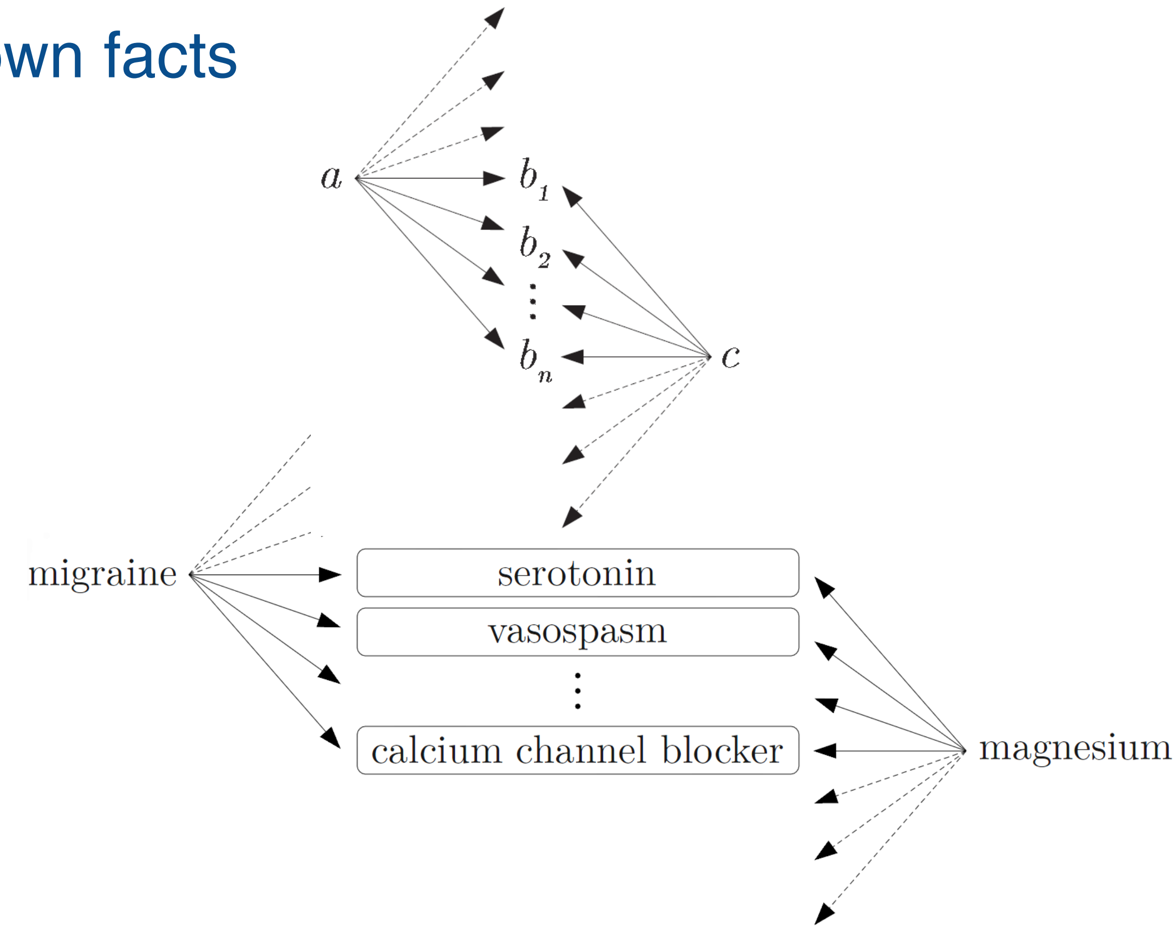
Current and future work

- Cross-domain literature mining: Fidning bridging concepts with CrossBee (Juršič et al., 2012)
  - Help experts in cross-domain bisociative discovery for unknown facts

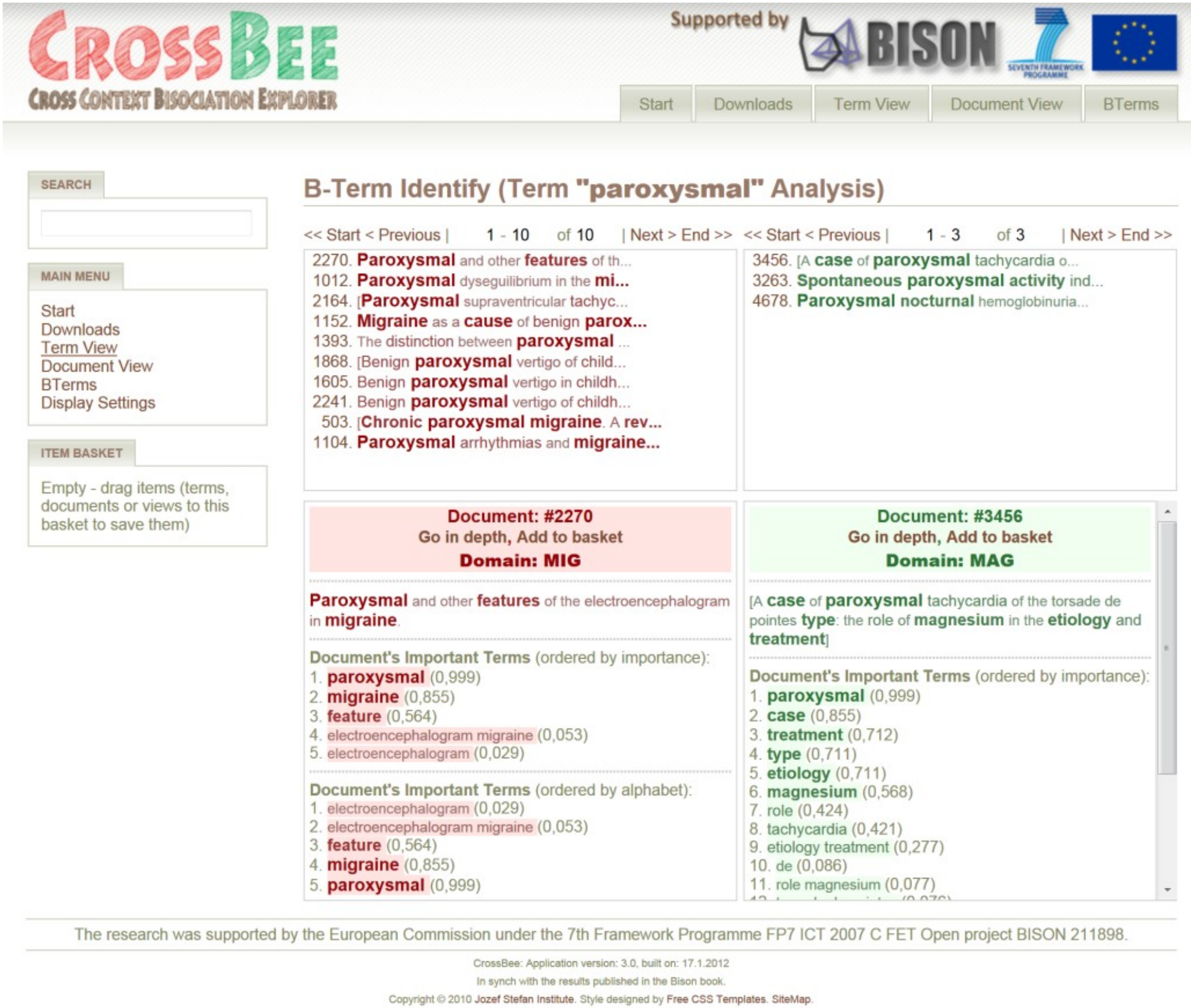


Current and future work

- Cross-domain literature mining: Fidning bridging concepts with CrossBee (Juršič et al., 2012)
  - Help experts in cross-domain bisociative discovery for unknown facts



CrossBee system



Introductory seminar lecture: Summary

- **JSI & Knowledge Technologies**
- **Introduction to Data mining and KDD**
  - Data Mining in a Nutshell
  - Predictive and descriptive DM techniques
  - Data Mining and KDD process
  - DM standards, tools and visualization
- **Selected data mining techniques: Advanced subgroup discovery techniques and applications**
- **Recent advances: Cross-context link discovery**