Data Mining
and Knowledge Discovery

ICT3 Programme

2017 / 2018

Nada Lavrač

Jožef Stefan Institute

Ljubljana, Slovenia

Outline

- **JSI & Knowledge Technologies**
- **Introduction to Data Mining and KDD**
  - Data Mining and KDD process
  - DM standards, tools and visualization
  - Classification of Data Mining techniques: Predictive and descriptive DM
- **Selected Data Mining techniques: Advanced subgroup discovery techniques and applications**
- **Relation between data mining and text mining**

Jožef Stefan Institute

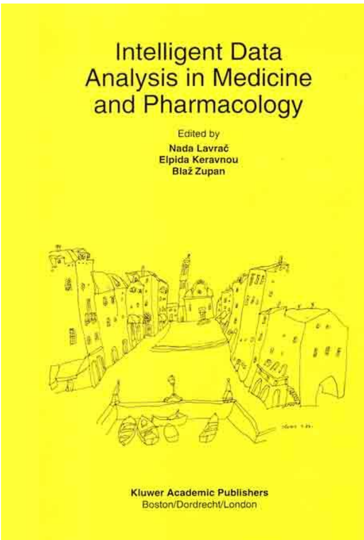
- **Jožef Stefan Institute (JSI, founded in 1949)**
  - named after a distinguished physicist Jožef Stefan (1835-1893)

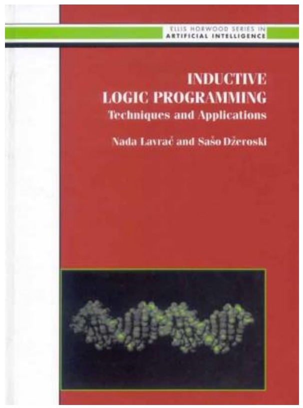
    $j = \sigma T^4$
  - leading national research organization in natural sciences and technology (~700 researchers and students)
- **JSI research areas**
  - information and communication technologies
  - chemistry, biochemistry & nanotechnology
  - physics, nuclear technology and safety
- **Jožef Stefan International Postgraduate School (IPS, founded in 2004)**
  - offers MSc and PhD programs (ICT, nanotechnology, ecotechnology)
  - research oriented, basic + management courses
  - in English

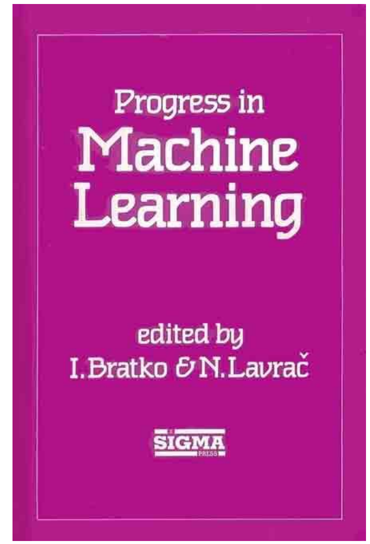
Department of Knowledge Technologies

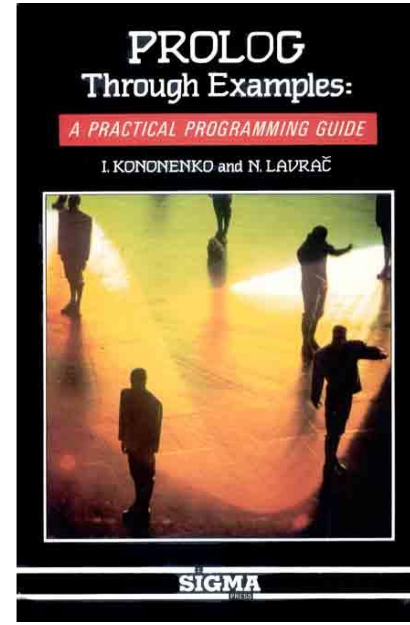
- **Head:** Nada Lavrač,
- **Staff:** 45 (30 researchers, 10 students, 5 tech/admin)
- **Machine learning & Data mining**
  - ML (decision tree and rule learning, subgroup discovery, …)
  - Text and Web mining
  - Relational data mining - inductive logic programming
  - Equation discovery
- **Other research areas:**
  - Semantic Web and Ontologies
  - Knowledge management
  - Decision support
  - Human language technologies
- **Applications:**
  - Medicine, Bioinformatics, Public Health
  - Ecology, Finance, …

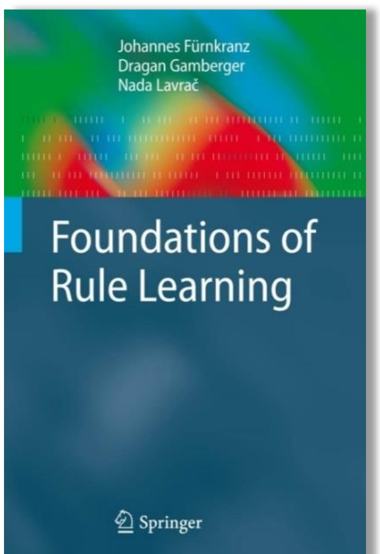
Selected Publications

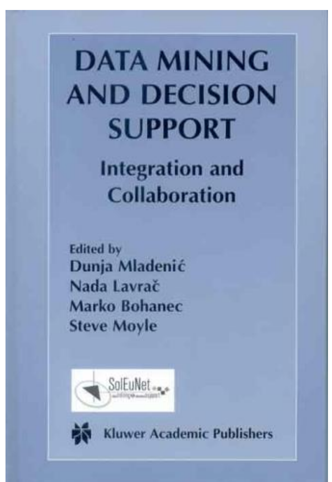


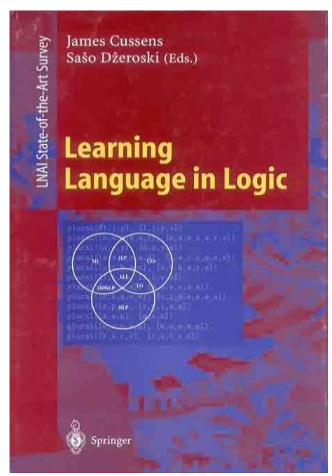


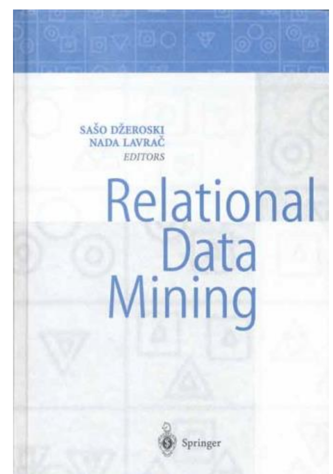


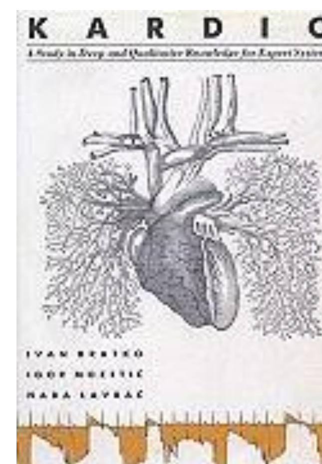












Data Mining 2017/2018 Logistics: Course participants

Contacts: http://kt.ijs.si/petra_kralj/dmkd.html

- Nada Lavrač, Bojan Cestnik, Petra Kralj Novak, Martin Žnidaršič
- Petra Kralj Novak: petra.kralj.novak@ijs.si

IPS ICT3 students **10 ECTS** Data mining and knowledge discovery Knowledge Technologies Module	Mišel Cevzar Darko Dujić David Gojo Aljoša Vodopija

Course Schedule – 2017/18

Wednesday	18.10.2017	17-19h	**prof. Nada Lavrač**	IPS Lecture hall
Wednesday	25.10.2016	15-17h	**prof. Bojan Cestnik**	IPS Lecture hall
		17h-19h	**dr. Petra Kralj Novak**	
			dr. Petra Kralj Novak: Reading club **Dr. Martin Žnidaršič: Reading club**	IPS Lecture hall
		15-16h	**written exam (2 ECTS) - dr. Petra Kralj Novak**	IPS Lecture hall
		16-18h	**seminar proposals topic discussion - dr. Petra Kralj Novak**	
		15-19h	**Seminars**	IPS Lecture hall
		15-19h	**spare term for seminars**	IPS Lecture hall

Data Mining: PhD Credits and Coursework

- Attending lectures
- Attending reading club
- Optional: Attending ICT2 theory exercises and hands-on (intro to WEKA by dr. Petra Kralj Novak)
- **Written exam (40%)**
- **Seminar (60%):**
  - Data analysis of your own data (e.g., using WEKA for questionnaire data analysis)
  - Implementing a selected data mining workflow in the ClowdFlows data mining platform
  - …. own initiative is welcome …

Data Mining: PhD Credits and coursework

Exam: Written exam (60 minutes) - Theory

Seminar: topic selection + results presentation

- One hour available for seminar topic discussion – one page written proposal defining the task and the selected dataset
- Deliver written report + electronic copy (4 pages in Information Society paper format, instructions on the web)
  - Report on data analysis of own data needs to follow the CRISP-DM methodology
  - Report on DM SW development needs to include SW compatible with the ClowdFlows I/O requirements
  - Presentation of your seminar results (15 minutes each: 10 minutes presentation + 5 minutes discussion)

Outline

- **JSI & Knowledge Technologies**
- **Introduction to Data Mining and KDD**
  – Data Mining and KDD process
  – DM standards, tools and visualization
  – Classification of Data Mining techniques: Predictive and descriptive DM
- **Selected Data Mining techniques: Advanced subgroup discovery techniques and applications**
- **Relation between data mining and text mining**

Part I. Introduction

- Data Mining in a Nutshell
- Data Mining and the KDD process
- DM standards and tools

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

Machine Learning and Data Mining

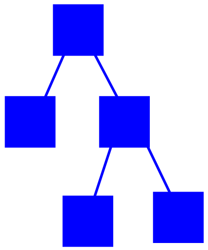
data

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

data

knowledge discovery
from data

Machine Learning
Data Mining



model, patterns, …

Given: class labeled data

Find: a classification model, a set of interesting patterns

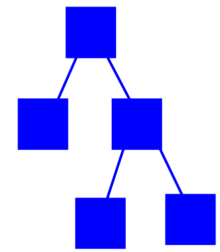
Machine Learning and Data Mining

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

data

knowledge discovery
from data

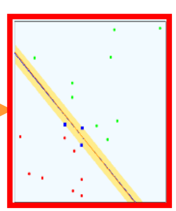


model, patterns, …

Given: class labeled data

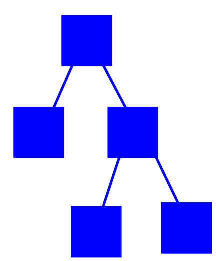
Find: a classification model, a set of interesting patterns

new unclassified instance



classified instance

black box classifier
no explanation



symbolic model
symbolic patterns

explanation

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

Pattern discovery in 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

PATTERN

Rule:

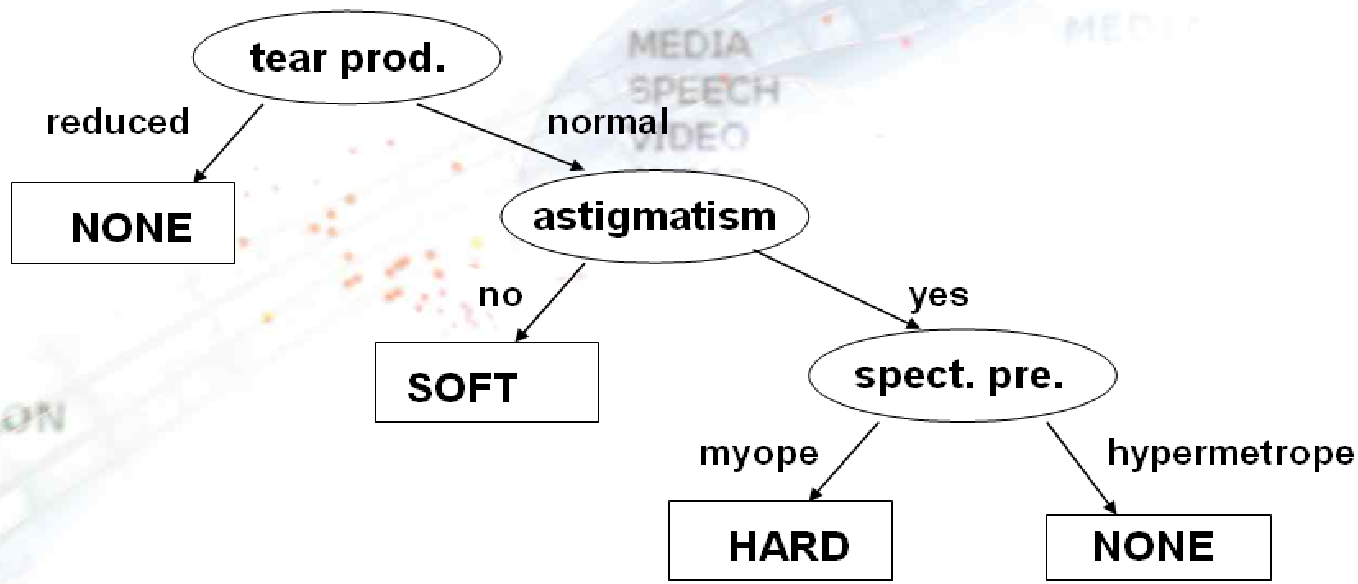
IF
Tear prod. =
reduced

THEN
Lenses =
NONE

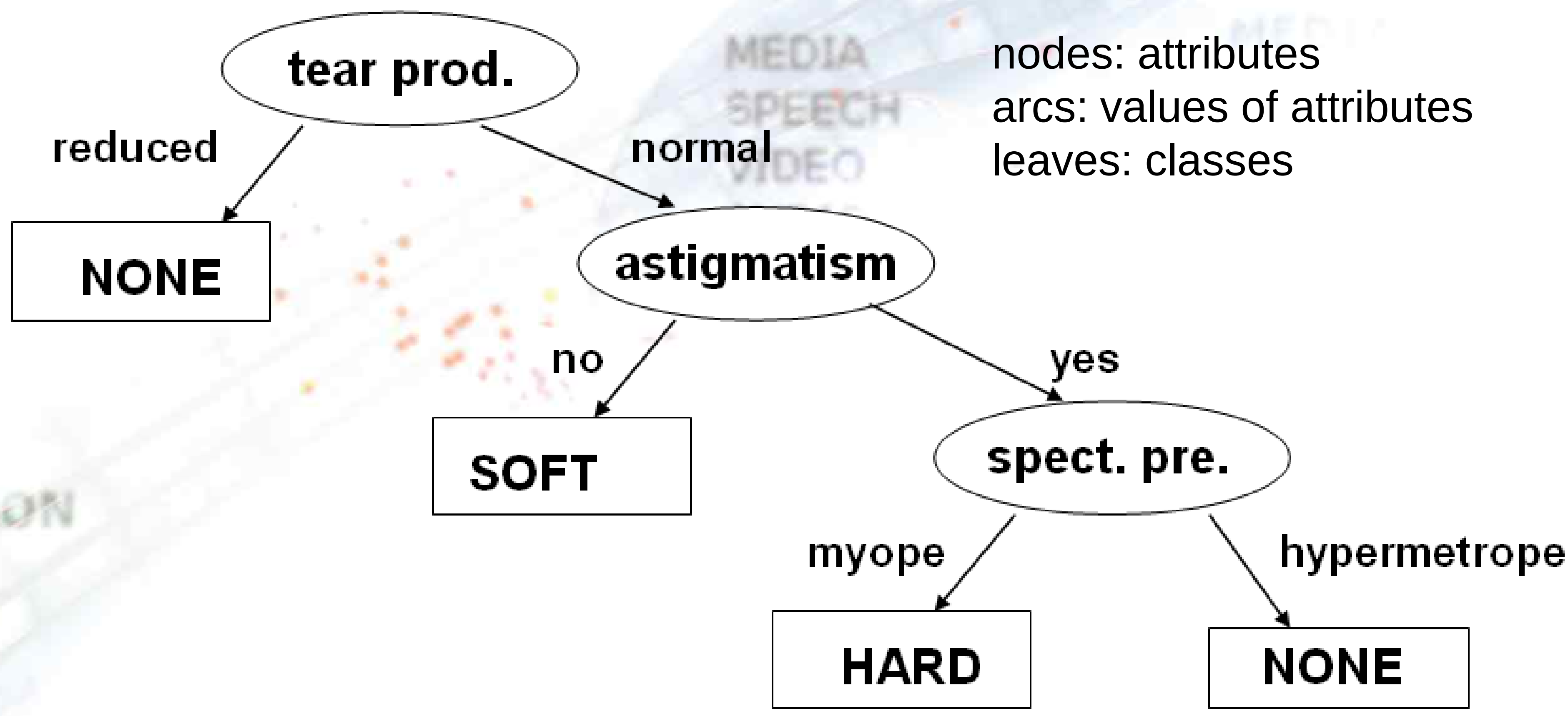
Learning a classification model from contact lens data

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

Data Mining

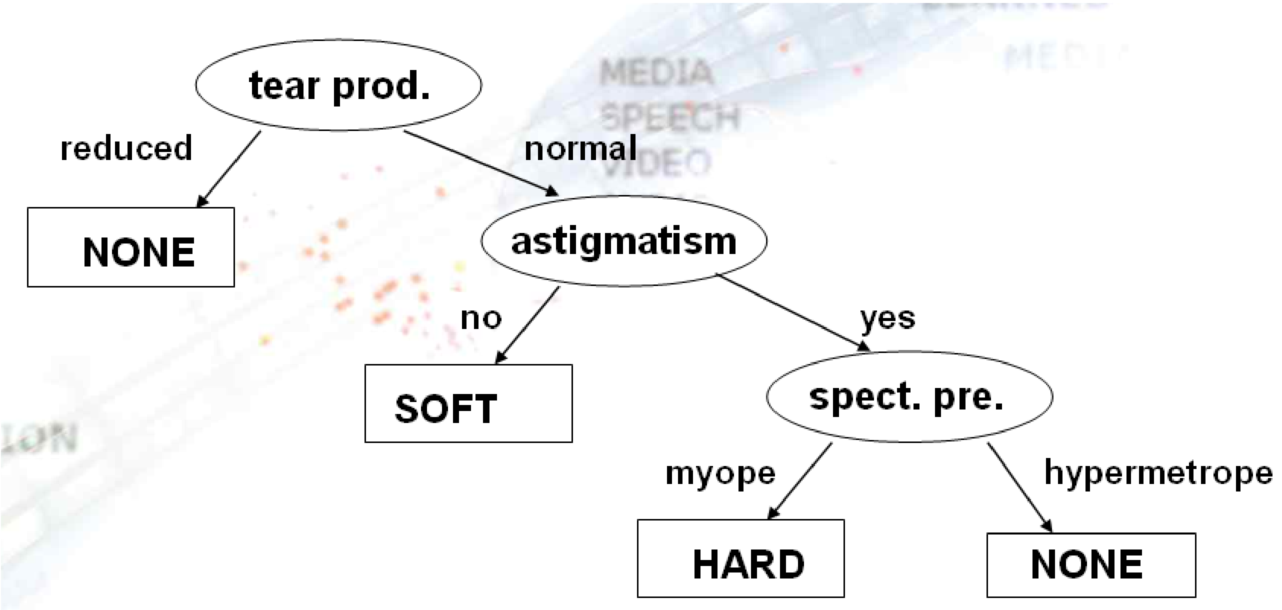


Decision tree classification model learned from contact lens data



nodes: attributes
arcs: values of attributes
leaves: classes

Learning a decision tree classification model



Using Gain(S,A) heuristic for determining the most informative attribute

$$Gain(S,A) = E(S) - \sum_{v \in Values(A)} p_v \cdot E(S_v)$$

Gain(S,A) estimates the reduction of entropy of set S after splitting into subsets based on values of attribute A

Entropy

- **S** - training set, $\mathbf{C_1,...,C_N}$ - classes
- **Entropy E(S)** – measure of the impurity of training set S

$$E(S) = -\sum_{c=1}^{N} p_c . \log_2 p_c$$

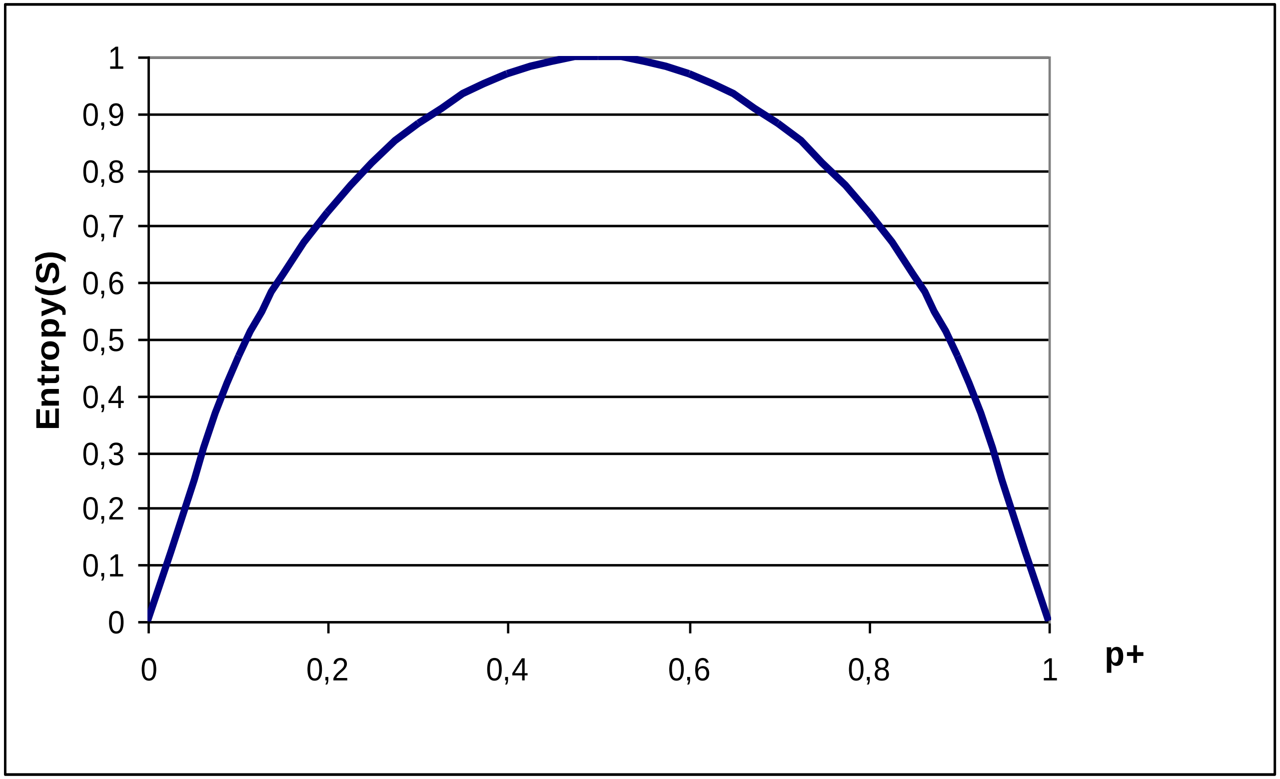
$\mathbf{p_c}$ - prior probability of class $\mathbf{C_c}$
(relative frequency of $\mathbf{C_c}$ in **S**)

- Entropy in binary classification problems

$$\mathbf{E(S) = - p_+ \log_2 p_+ - p_- \log_2 p_-}$$

Entropy

- $E(S) = - p_+ \log_2 p_+ - p_- \log_2 p_-$
- The entropy function relative to a Boolean classification, as the proportion $\mathbf{p_+}$ of positive examples varies between 0 and 1



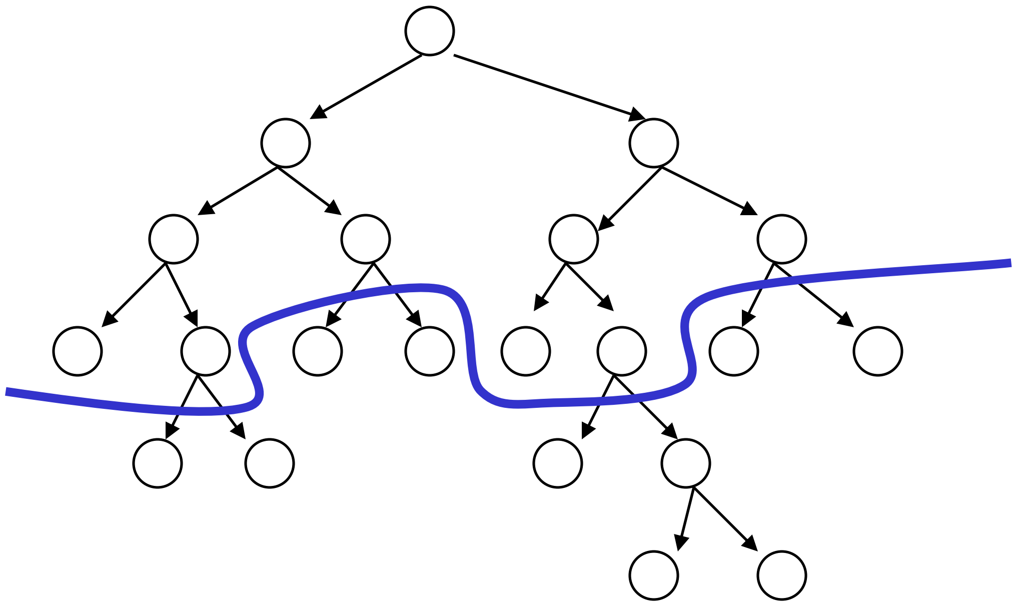
Entropy – why ?

- **Entropy E(S) =** expected amount of information (in bits) needed to assign a class to a randomly drawn object in S (under the optimal, shortest-length code)
- Why ?
- Information theory: optimal length code assigns $-\log_2 p$ bits to a message having probability p
- So, in binary classification problems, the expected number of bits to encode + or – of a random member of S is:

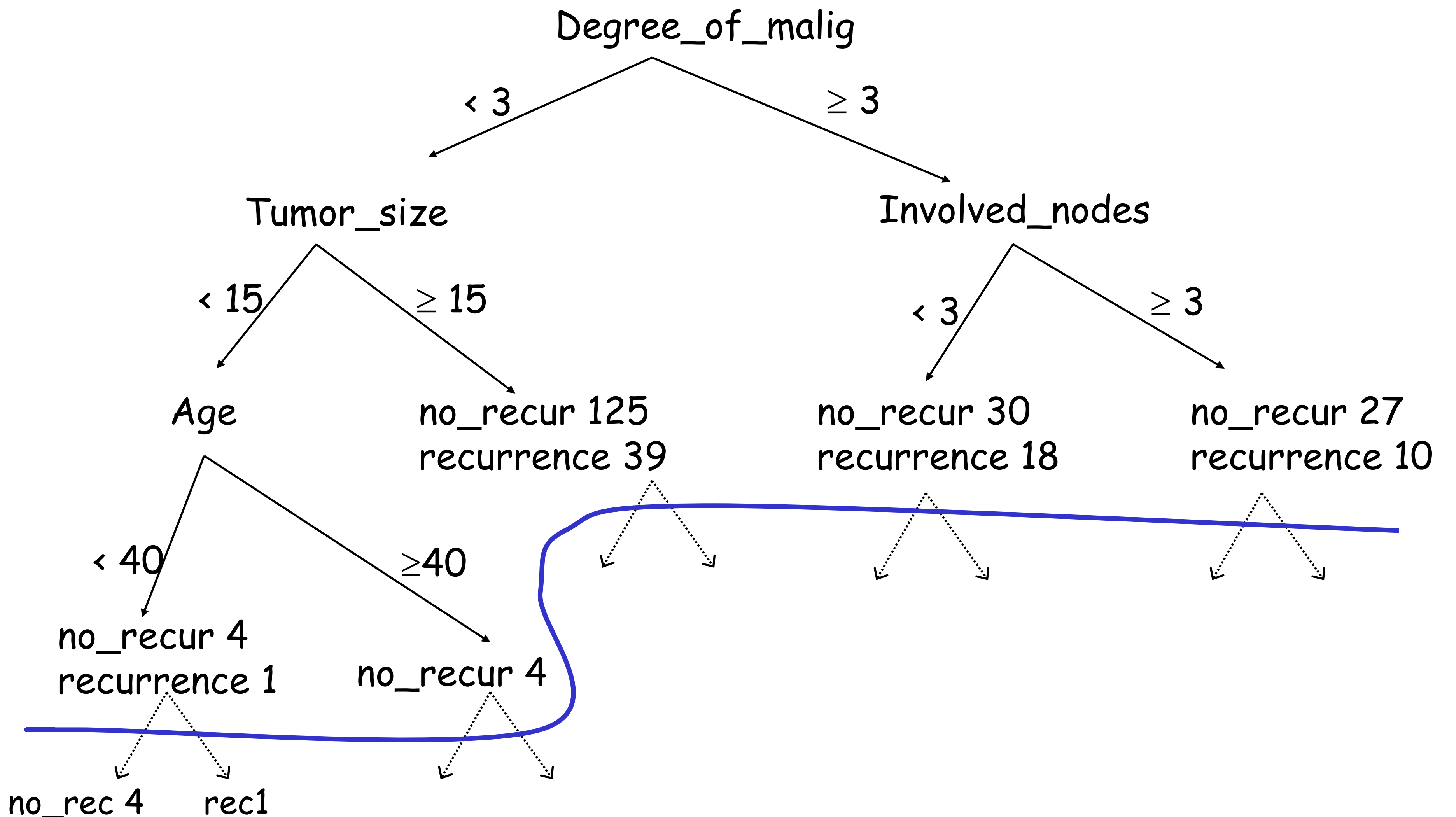
$$p_+ (- \log_2 p_+) + p_- (- \log_2 p_-) = - p_+ \log_2 p_+ - p_- \log_2 p_-$$

Pruning of decision trees

- Avoid overfitting the data by tree pruning
- Pruned trees are
  - less accurate on training data
  - more accurate when classifying unseen data



Prediction of breast cancer recurrence: Tree pruning

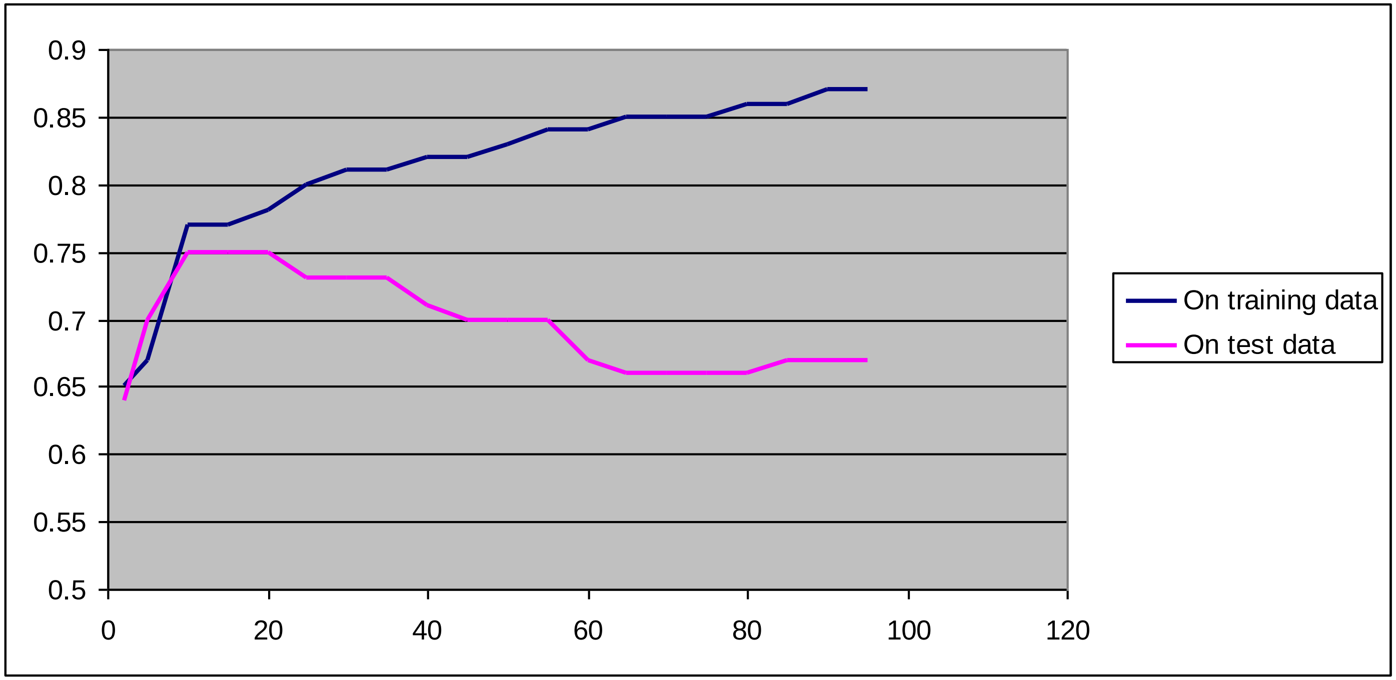


Accuracy and error

- Accuracy: percentage of correct classifications
  - on the training set
  - on unseen instances
- How accurate is a decision tree when classifying unseen instances
  - An estimate of accuracy on unseen instances can be computed, e.g., by averaging over 4 runs:
    - split the example set into training set (e.g. 70%) and test set (e.g. 30%)
    - induce a decision tree from training set, compute its accuracy on test set
- Error = 1 - Accuracy
- High error may indicate data overfitting

Overfitting and accuracy

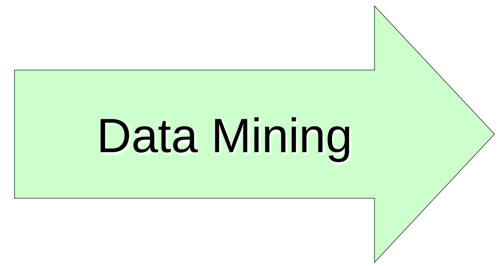
- Typical relation between tree size and accuracy

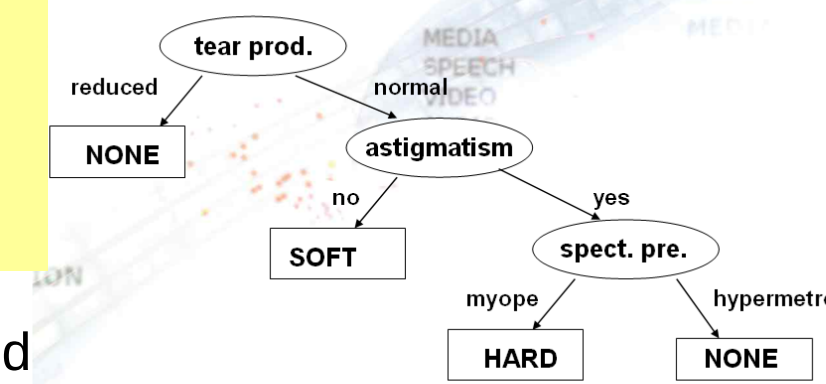


- Question: how to prune optimally?

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





lenses=NONE ← tear production=red

lenses=NONE ← tear production=normal AND astigmatism=yes AND spect. pre.=hypermetrope

lenses=SOFT ← tear production=normal AND astigmatism=no

lenses=HARD ← tear production=normal AND astigmatism=yes AND spect. pre.=myope

lenses=NONE ←

Classification rules model learned from contact lens data

lenses=NONE ← tear production=reduced

lenses=NONE ← tear production=normal AND
astigmatism=yes AND
spect. pre.=hypermetrope

lenses=SOFT ← tear production=normal AND
astigmatism=no

lenses=HARD ← tear production=normal AND
astigmatism=yes AND
spect. pre.=myope

lenses=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 tasks
    - classification and class description
    - "one vs. all" multi-class learning

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

Why learn and use symbolic models

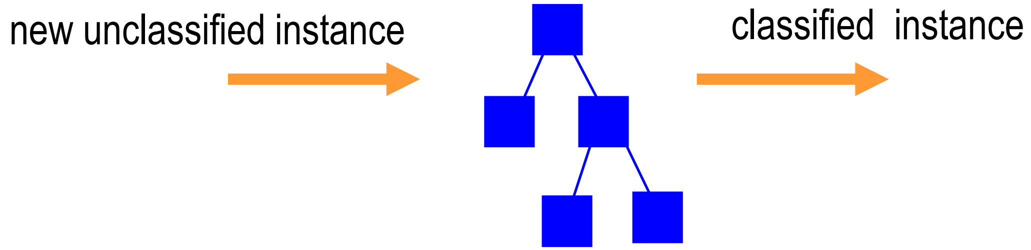
Given: the learned classification model
(a decision tree or a set of rules)

Find: the class label for a new unlabeled instance

Why learn and use symbolic models

Given: the learned classification model
(a decision tree or a set of rules)

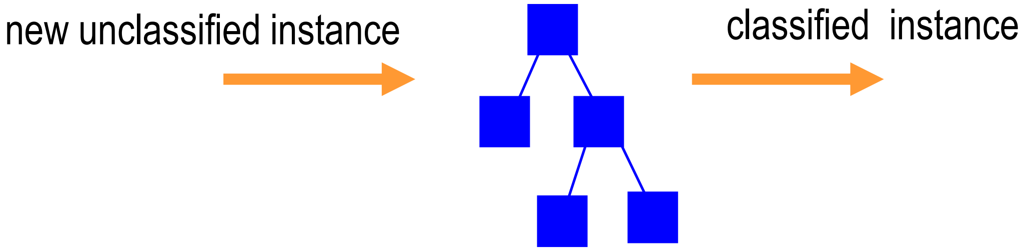
Find: the class label for a new unlabeled instance



Why learn and use symbolic models

Given: the learned classification model
(a decision tree or a set of rules)

Find: - the class label for a new unlabeled instance



- use the model for the explanation of classifications of new data instances
- use the discovered patterns for data exploration

First Generation Data Mining

- **First machine learning algorithms for**
  - Decision tree and classification rule learning in 1970s and early 1980s, by Quinlan, Michalski et al., Breiman et al., …
- **Characterized by**
  - Learning from simple tabular data
  - Relatively small set of instances and attributes
- **Lots of ML research followed in 1980s**
  - Numerous conferences ICML, ECML, … and ML sessions at AI conferences IJCAI, ECAI, AAAI, …
  - Extended set of learning tasks and algorithms addressed

Part I. Introduction

- Data Mining in a Nutshell
- Data Mining and the KDD process
- DM standards and tools

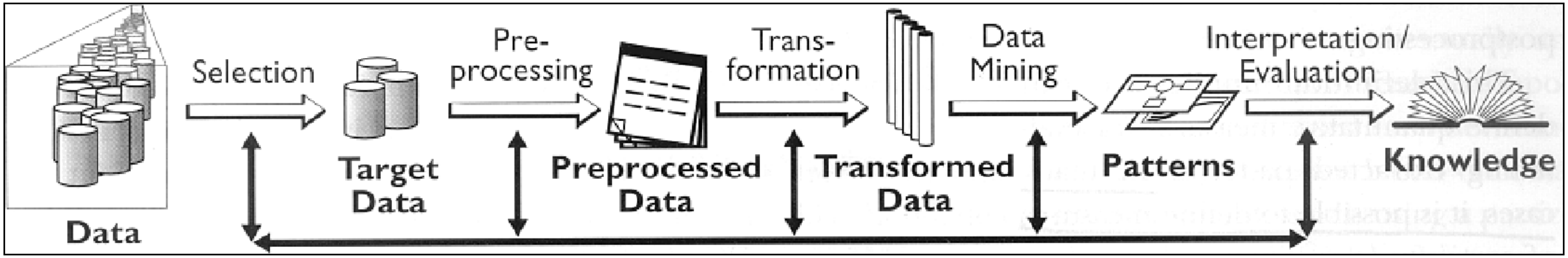
Data Mining and KDD

- 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

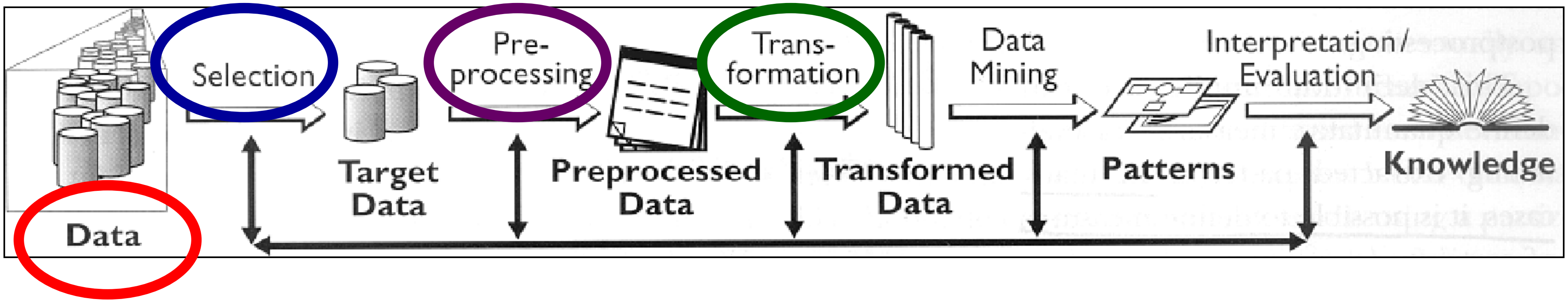
KDD Process

KDD process of discovering useful knowledge from data



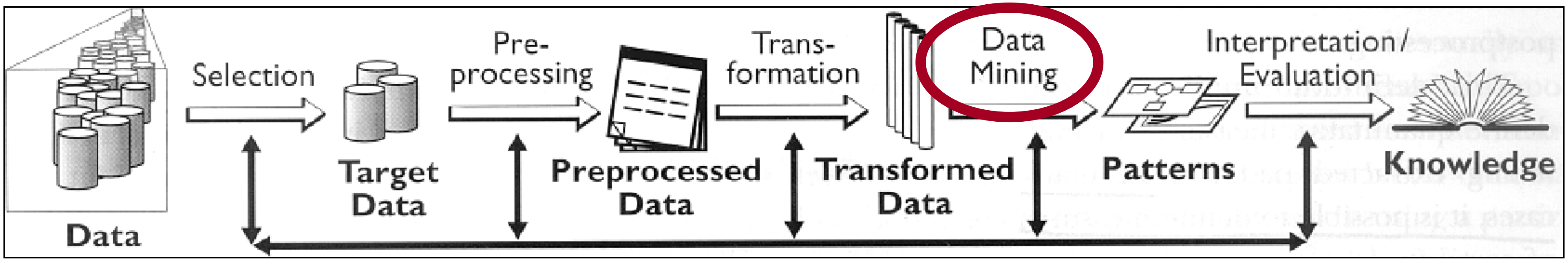
- KDD process involves several phases:
  - data preparation
  - data mining (machine learning, statistics)
  - evaluation and use of discovered patterns
- Data mining is the key step, but represents only 15%-25% of the entire KDD process

MEDIANA – analysis of media research data



- Questionnaires about journal/magazine reading, watching of TV programs and listening of radio programs, about 1200 questions. Yearly publication: frequency of reading/listening/watching, distribution w.r.t. Sex, Age, Education, Buying power,..
- Data for one year, about 8,000 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.

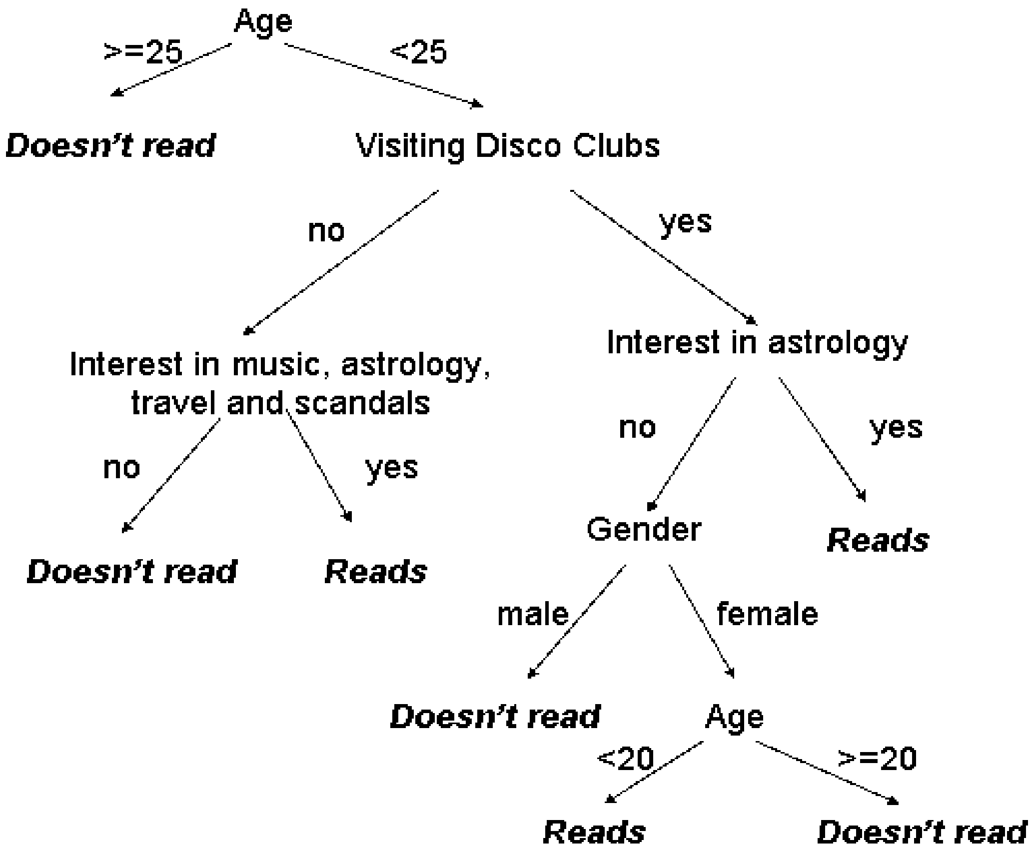
Simplified association rules

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.

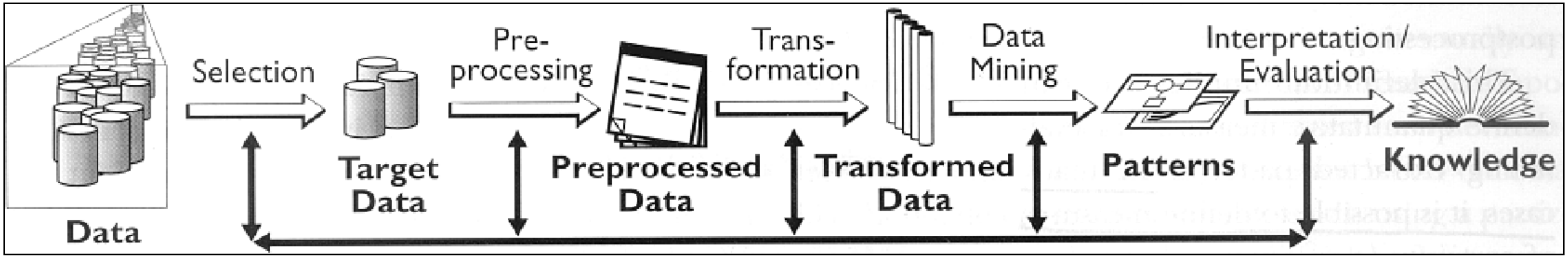


Part I. Introduction

- Data Mining in a Nutshell
- Data Mining and the KDD process
- DM standards and tools

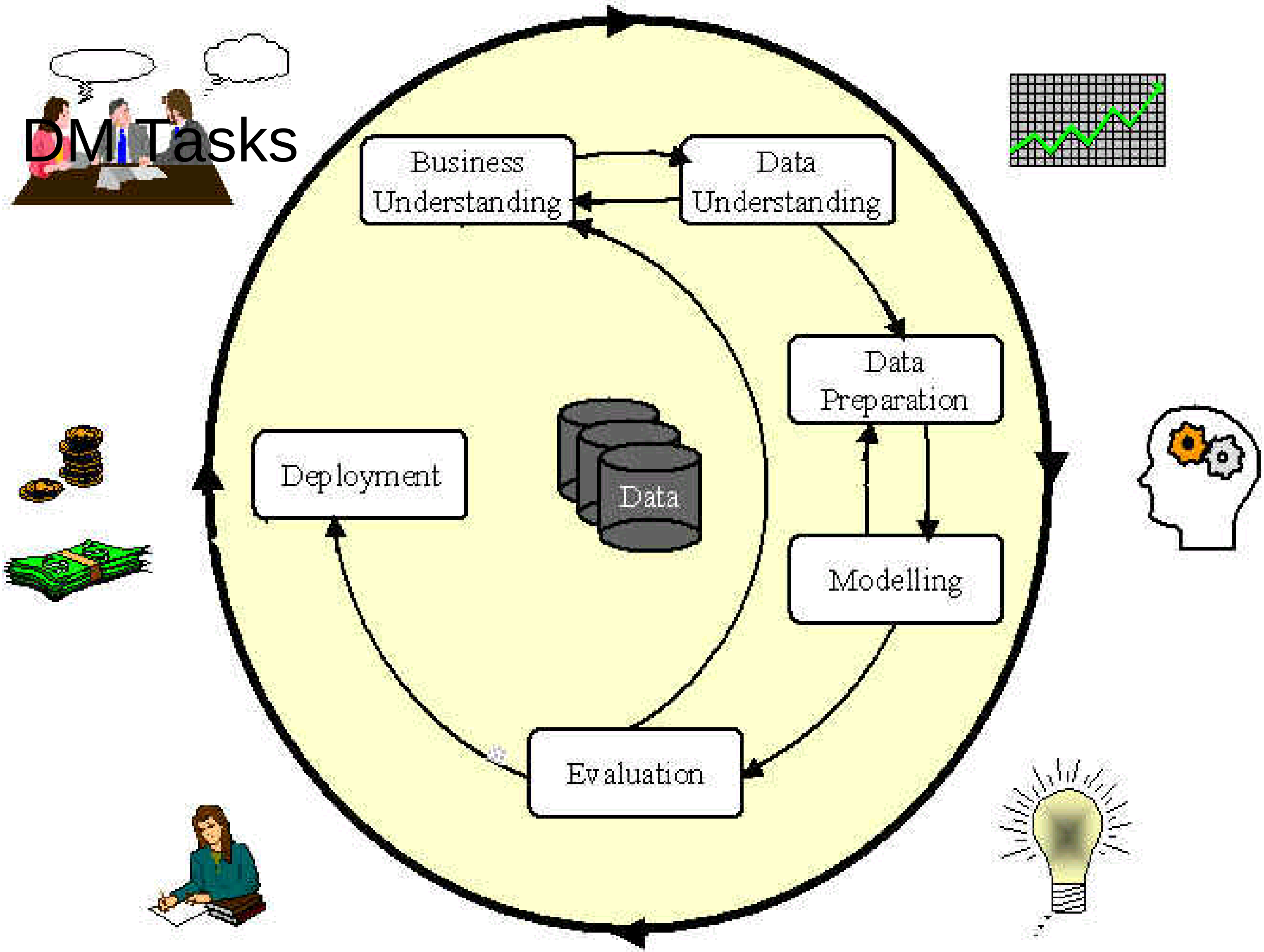
CRISP-DM

- 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):

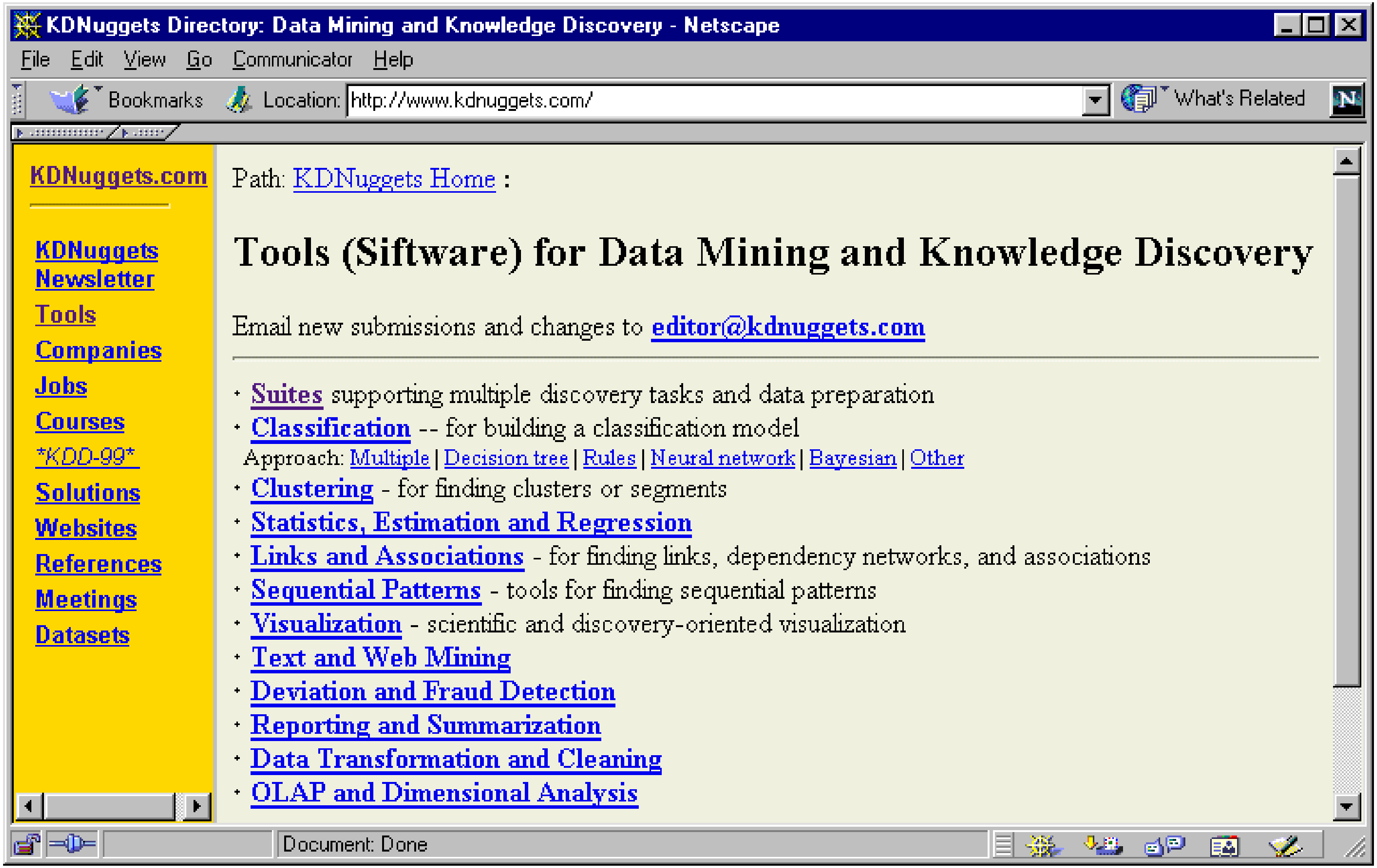


CRISP Data Mining Process

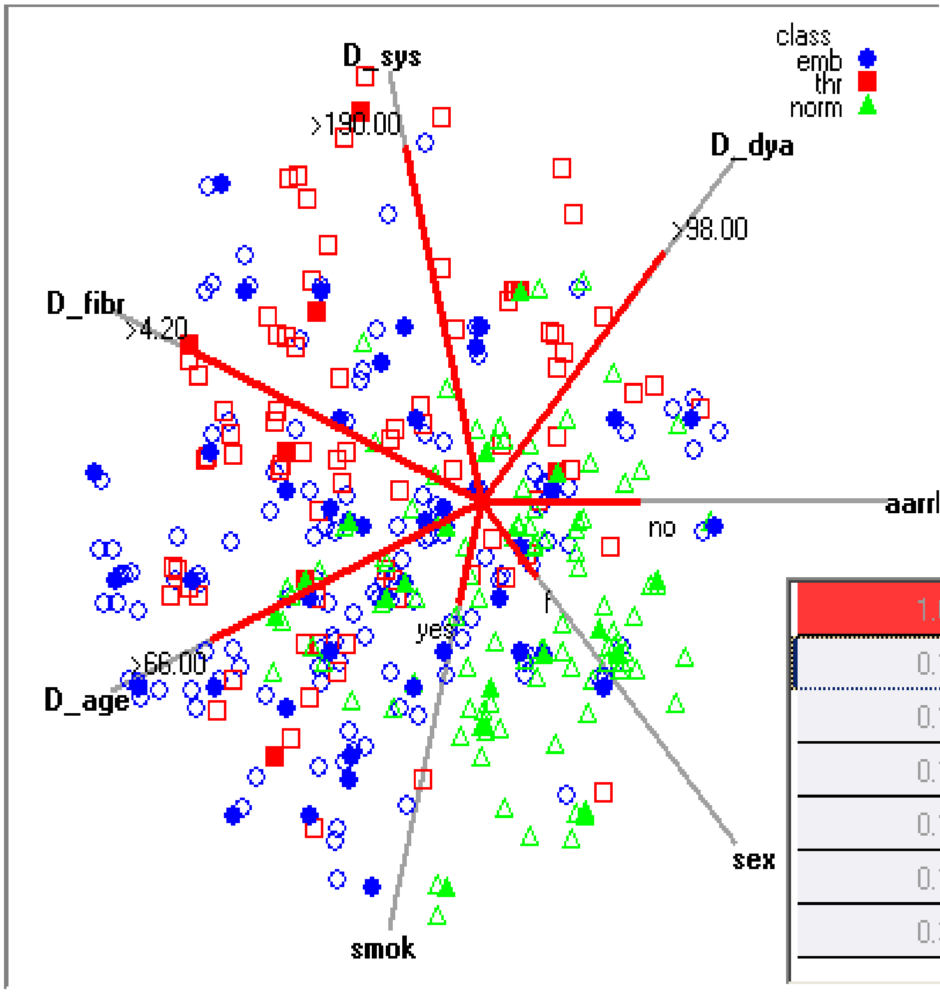
- DM Tasks

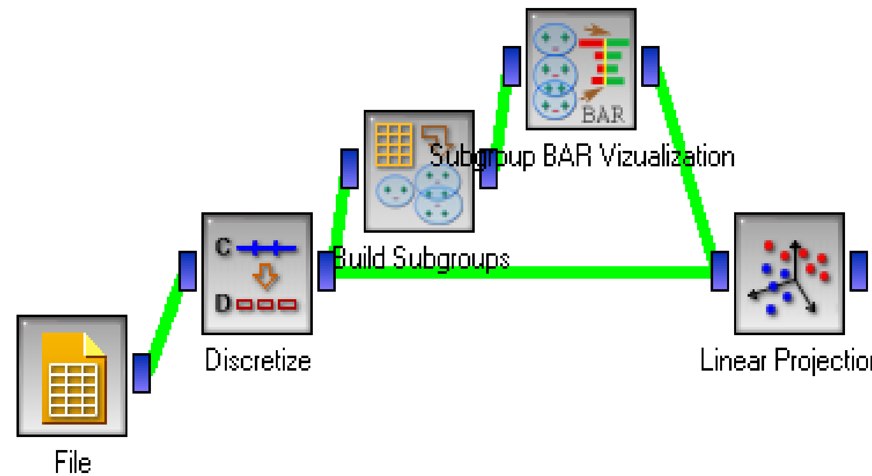


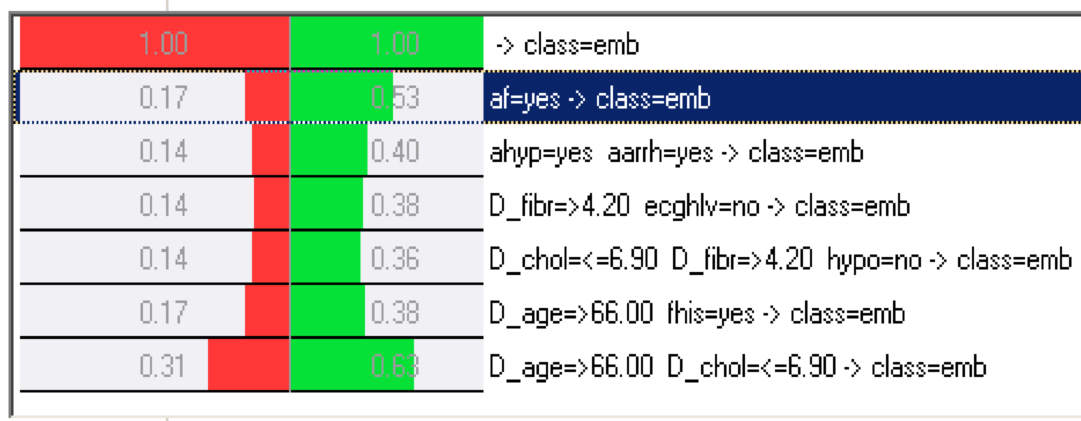
DM tools


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

Orange: Visual programming and visualization

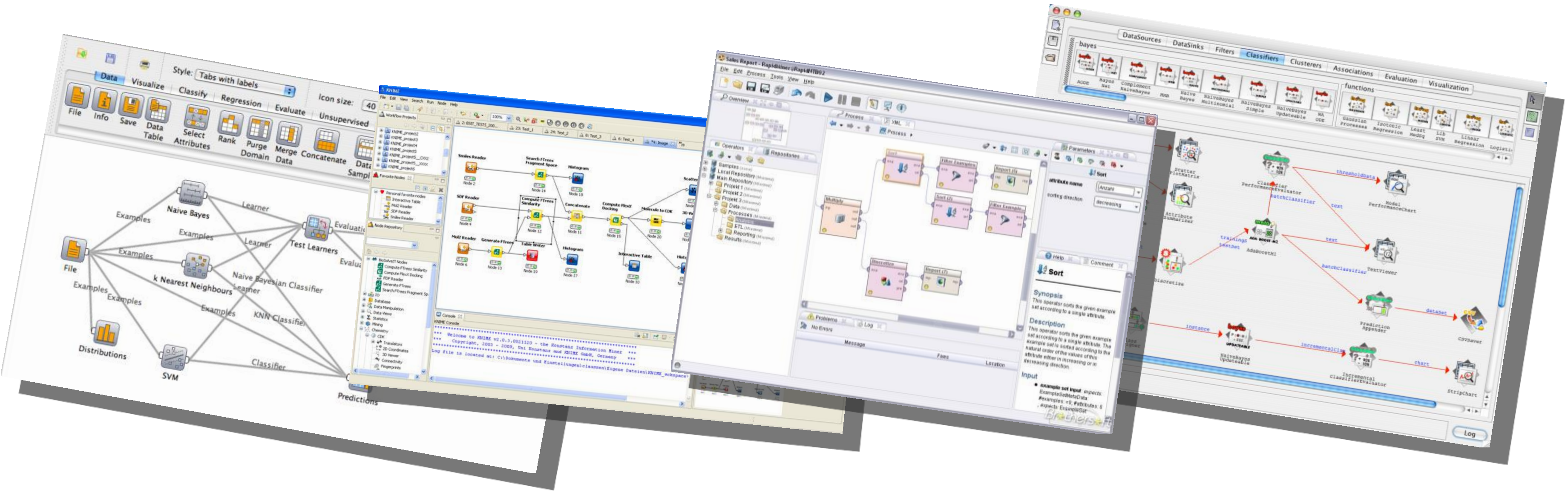






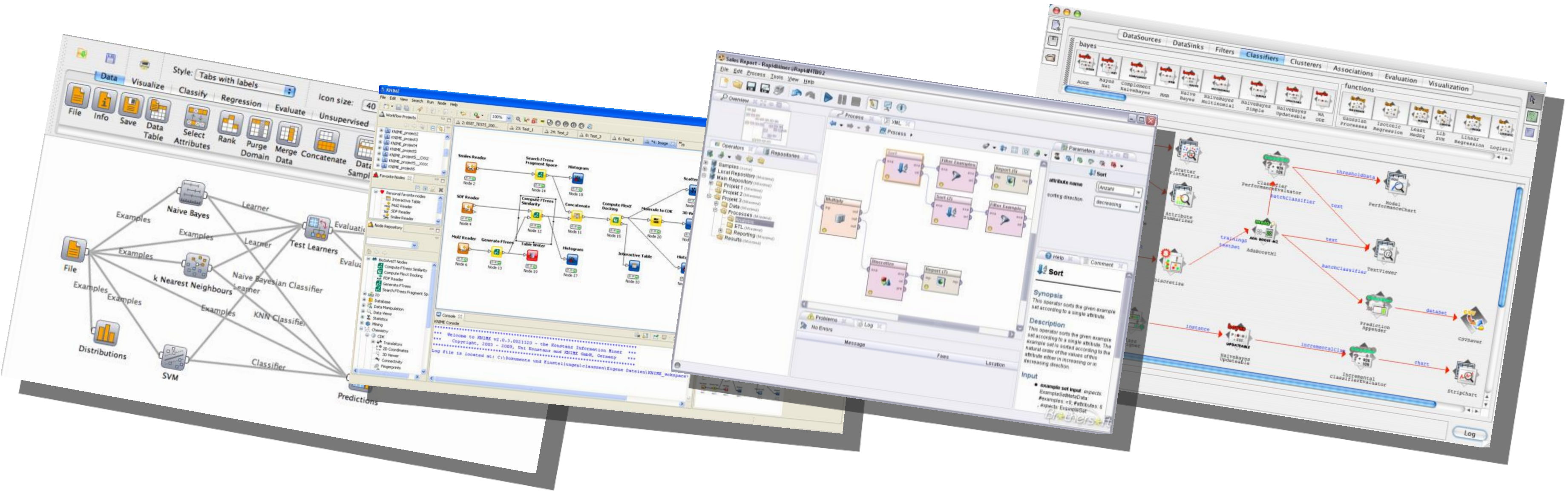
Other Second Generation Data Mining Platforms

WEKA, KNIME, RapidMiner, Orange, …



Other Second Generation Data Mining Platforms

WEKA, KNIME, RapidMiner, Orange, …



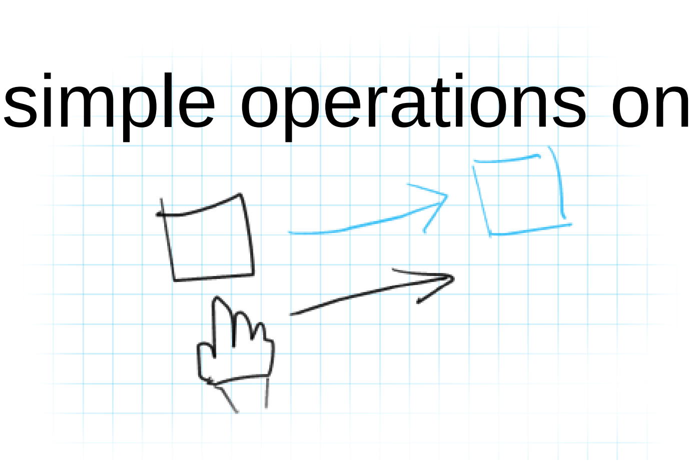
– include numerous data mining algorithms

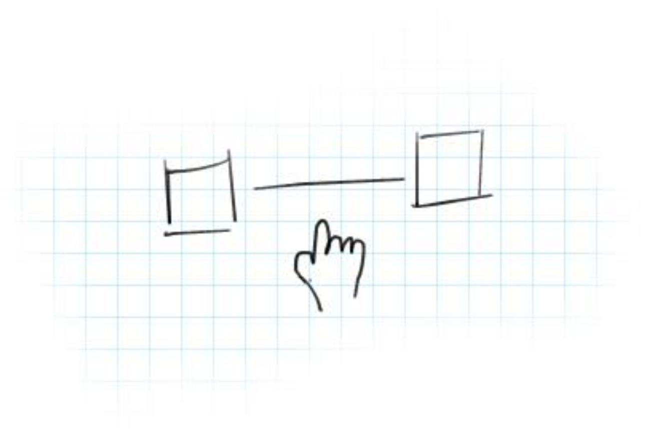
– enable data and model visualization

– like Taverna, WEKA, KNIME, RapidMiner, Orange also enable complex **workflow** construction

Building scientific workflows

- consists of simple operations on workflow elements
  - drag
  - drop
  - connect

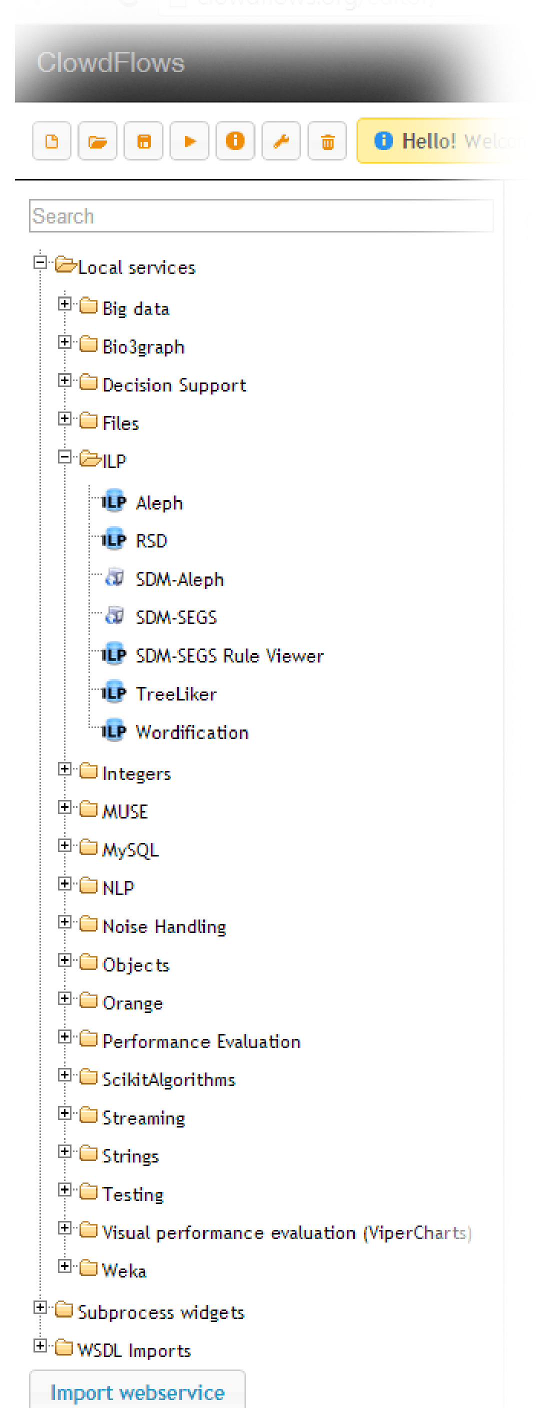




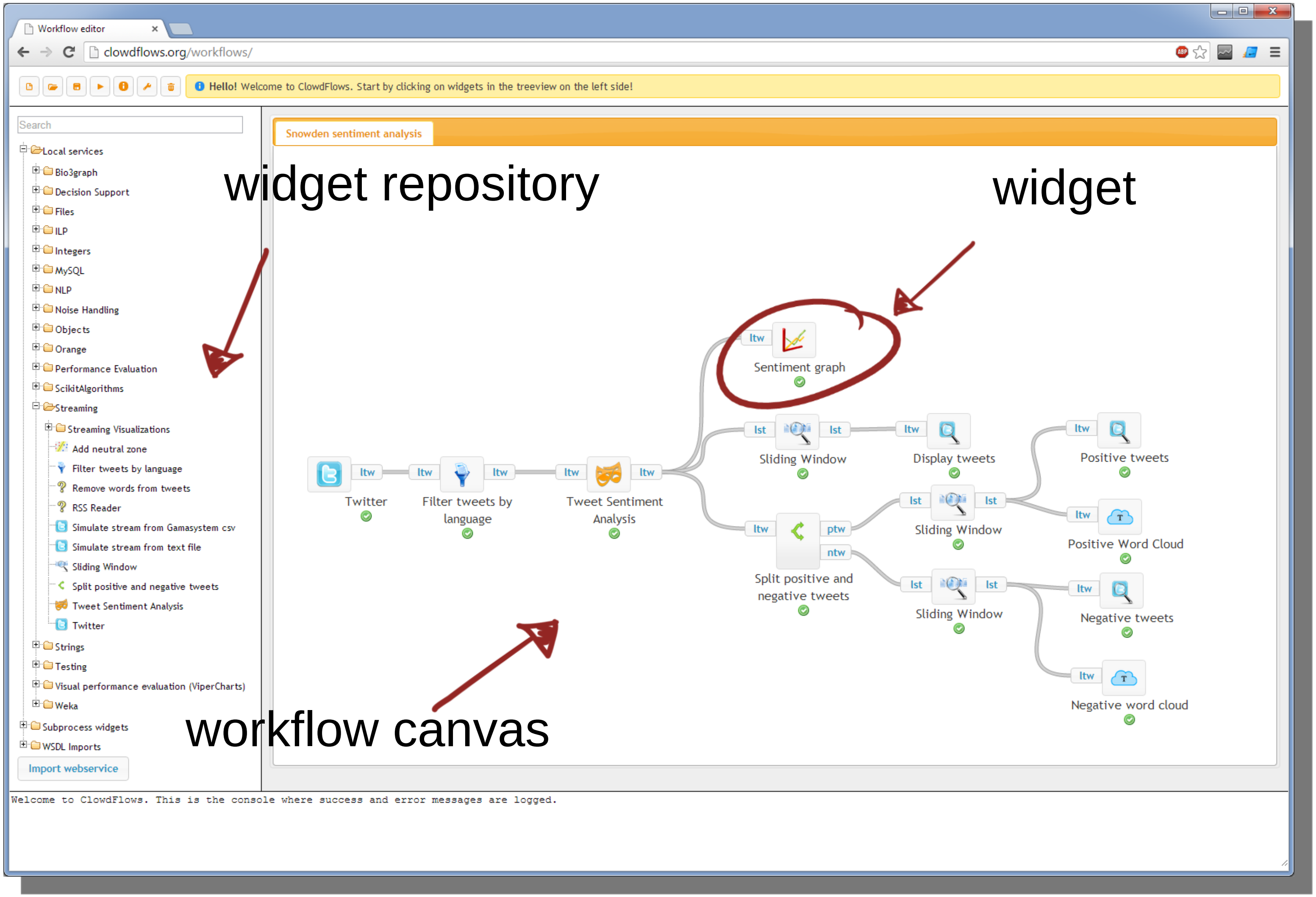
- suitable for non-experts
- good for representing complex procedures
- allow users to publicly upload their workflows so that they are available to a wider audience, perfect for experiment replication

ClowdFlows platform

- **Large algorithm repository**
  - Relational data mining
  - All Orange algorithms
  - WEKA algorithms as web services
  - Data and results visualization
  - Text analysis
  - Social network analysis
  - Analysis of big data streams
- **Large workflow repository**
  - Enables access to our technology heritage

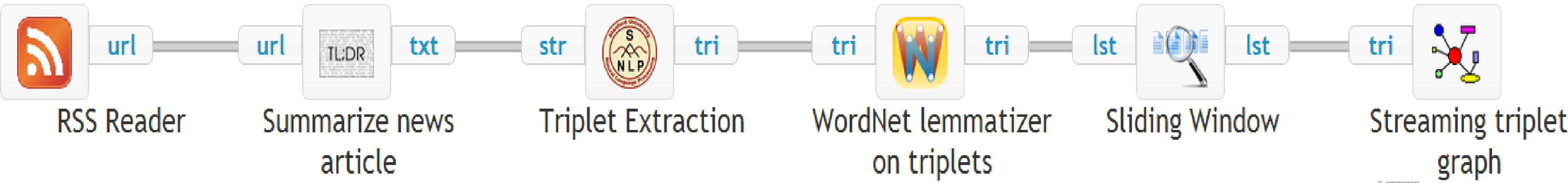


ClowdFlows user interface



“Big Data” Use Case

- Real-time analysis of big data streams
- Example: semantic graph construction from news streams. http://clowdflows.org/workflow/1729/.

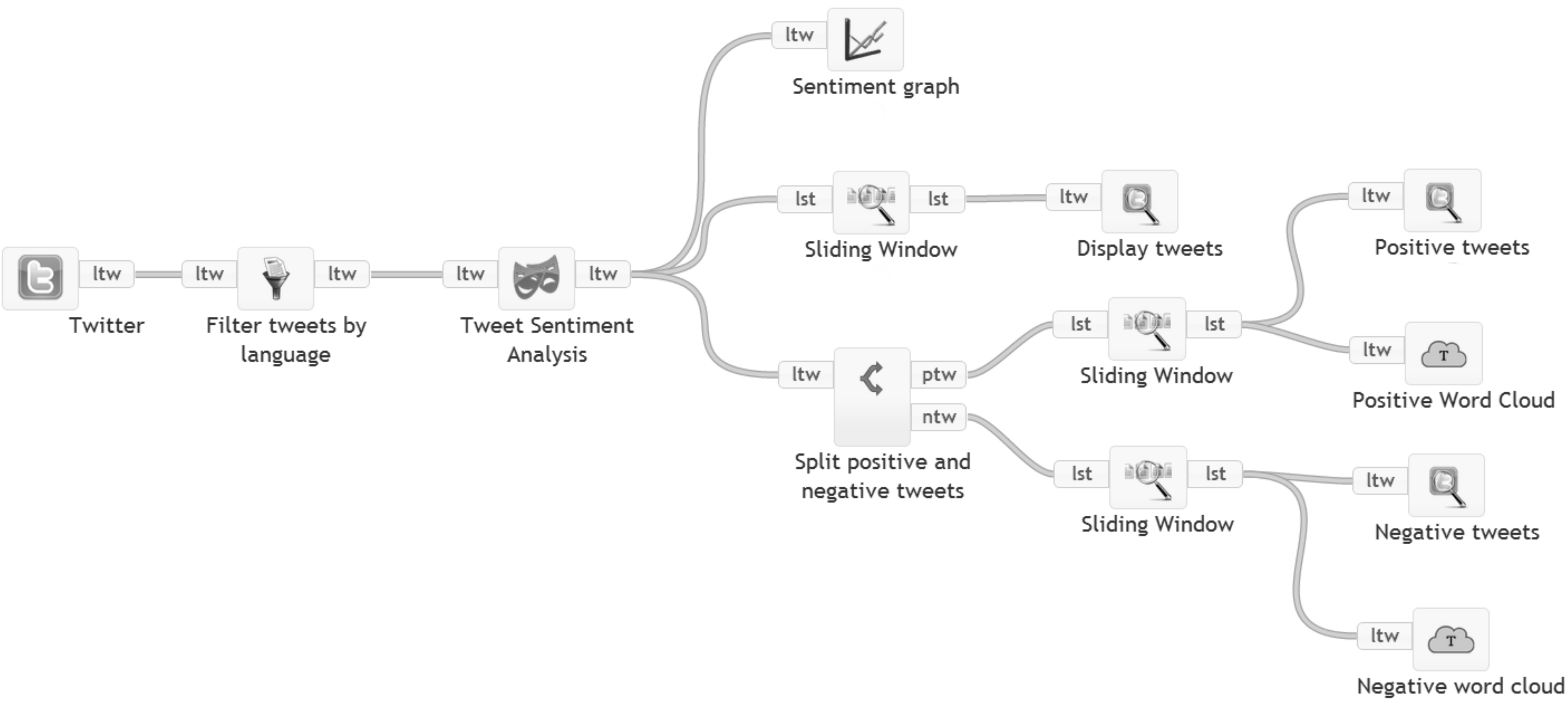


- Example: news monitoring by graph visualization (graph of CNN RSS feeds)

http://clowdflows.org/streams/data/31/1

“Big Data” Use Case

- Analysis of positive/negative sentiment of tweets in real time: http://clowdflows.org/workflow/1041/.



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

Outline

- **JSI & Knowledge Technologies**
- **Introduction to Data mining and KDD**
  - Data Mining and KDD process
  - DM standards, tools and visualization
  - Classification of Data Mining techniques: Predictive and descriptive DM
- **Selected data mining techniques: Advanced subgroup discovery techniques and applications**
- **Relation between data mining and text mining**

Selected Data Mining Techniques Outline

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

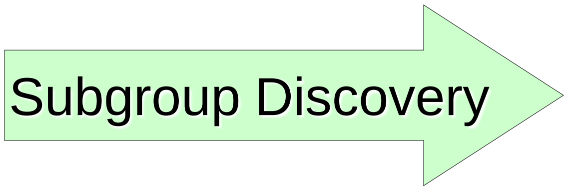
Task reformulation: Binary Class Values

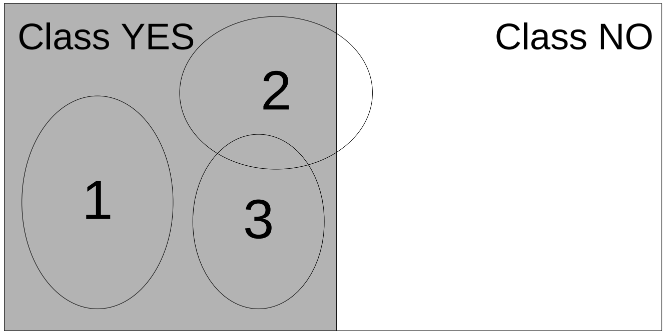
Person	Age	Spect. presc.	Astigm.	Tear prod.	Lenses
O1	17	myope	no	reduced	NO
O2	23	myope	no	normal	YES
O3	22	myope	yes	reduced	NO
O4	27	myope	yes	normal	YES
O5	19	hypermetrope	no	reduced	NO
O6-O13	...	...	...	...	...
O14	35	hypermetrope	no	normal	YES
O15	43	hypermetrope	yes	reduced	NO
O16	39	hypermetrope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



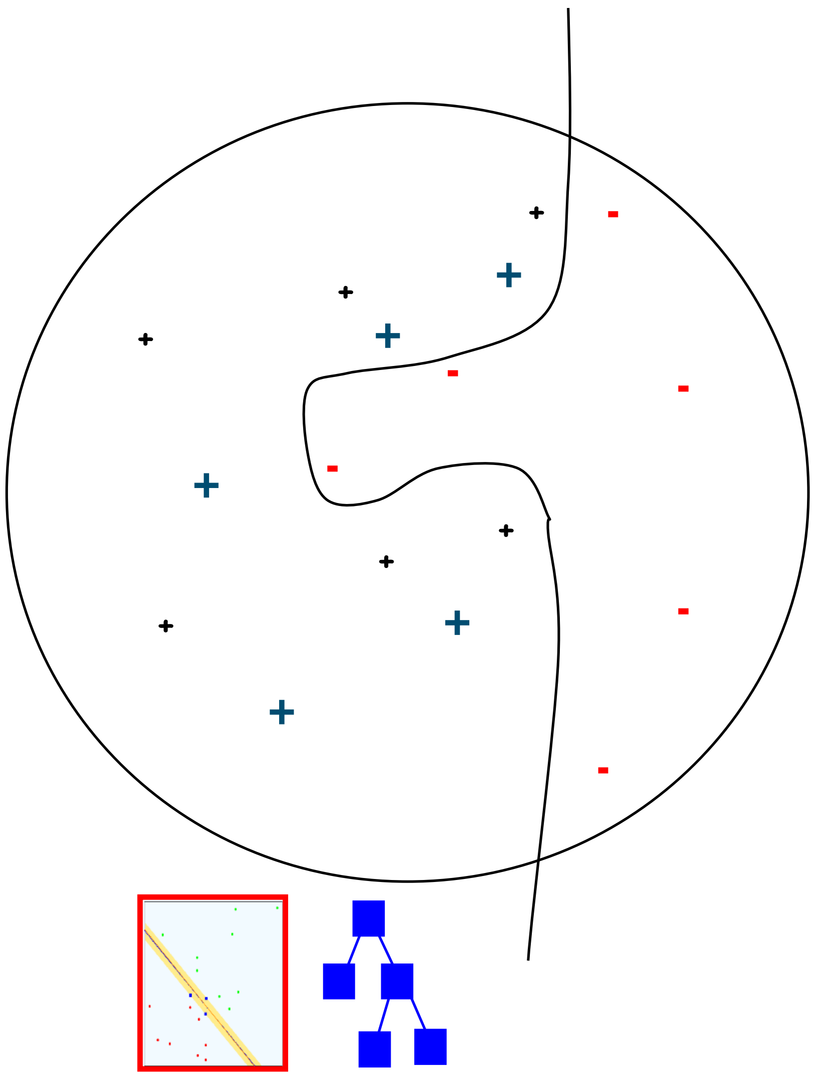


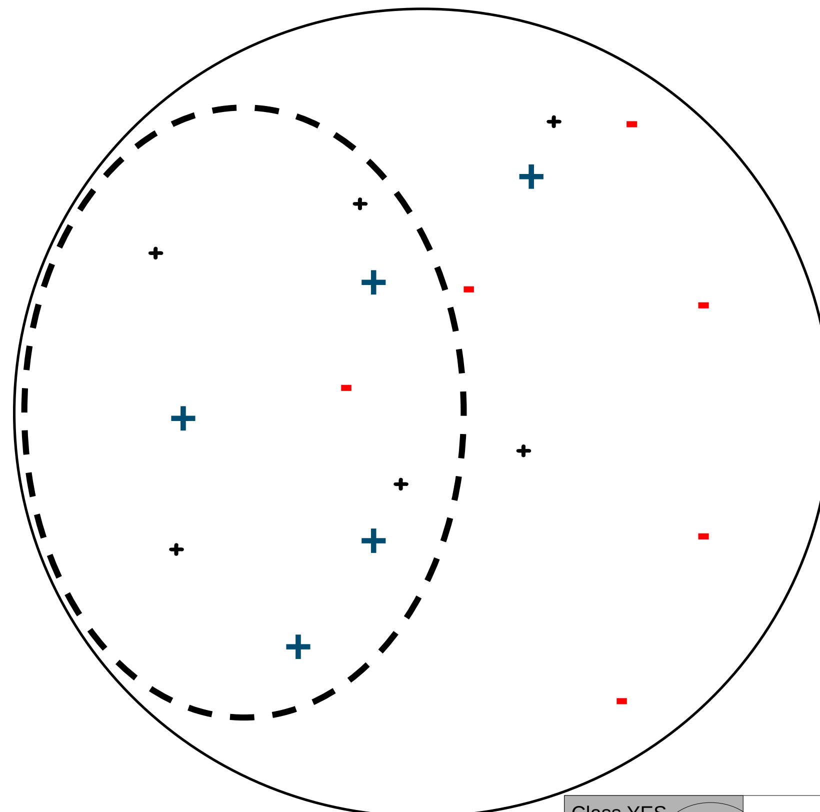
- A task in which individual interpretable patterns in the form of rules are induced from data, labeled by a predefined property of interest.
- SD algorithms learn several independent rules that describe groups of target class examples
  - subgroups must be large and statistically significant

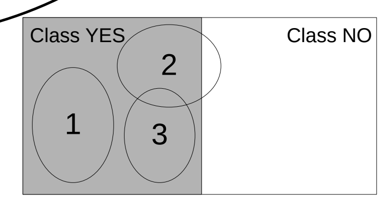
Classification versus Subgroup Discovery

- **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

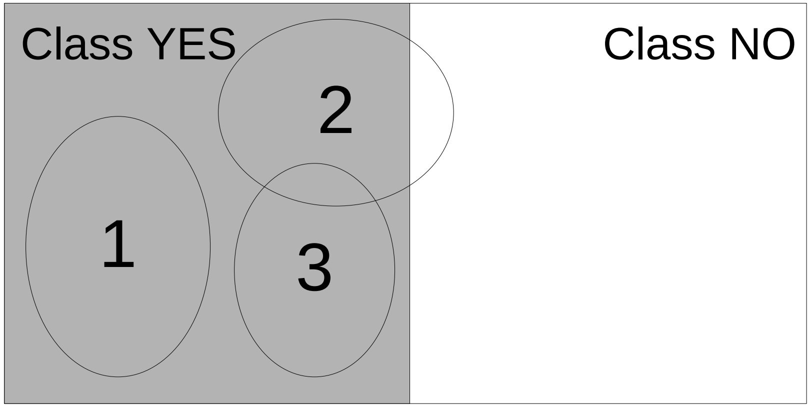
CHD-risk ← ...

CHD-risk ← ...

CHD-risk ← ...

Characteristics of SD Algorithms

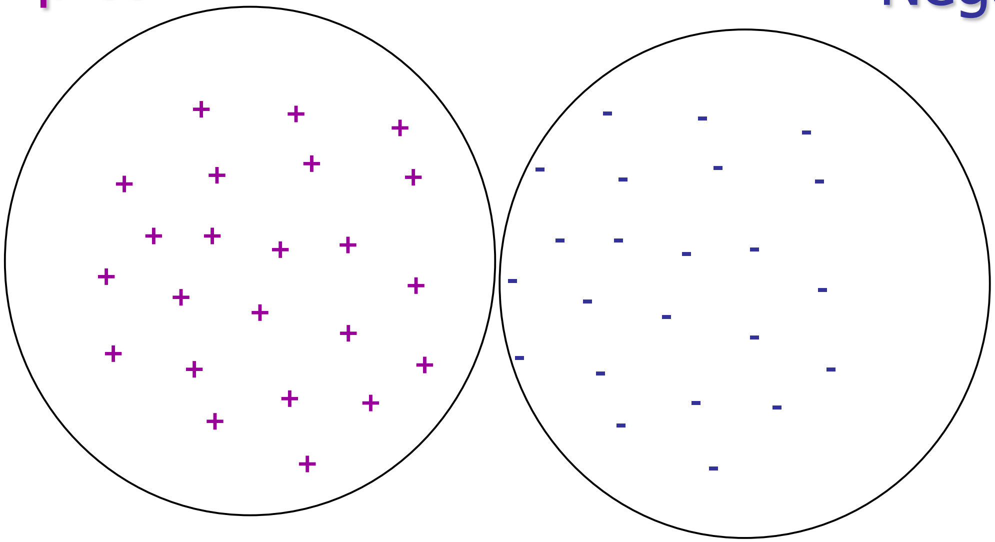
- SD algorithms do not look for a single complex rule to describe all examples of target class YES (all CHD-risk patients), but several rules that describe parts (subgroups) of YES.
- Standard rule learning approach: Using the covering algorithm for rule set construction



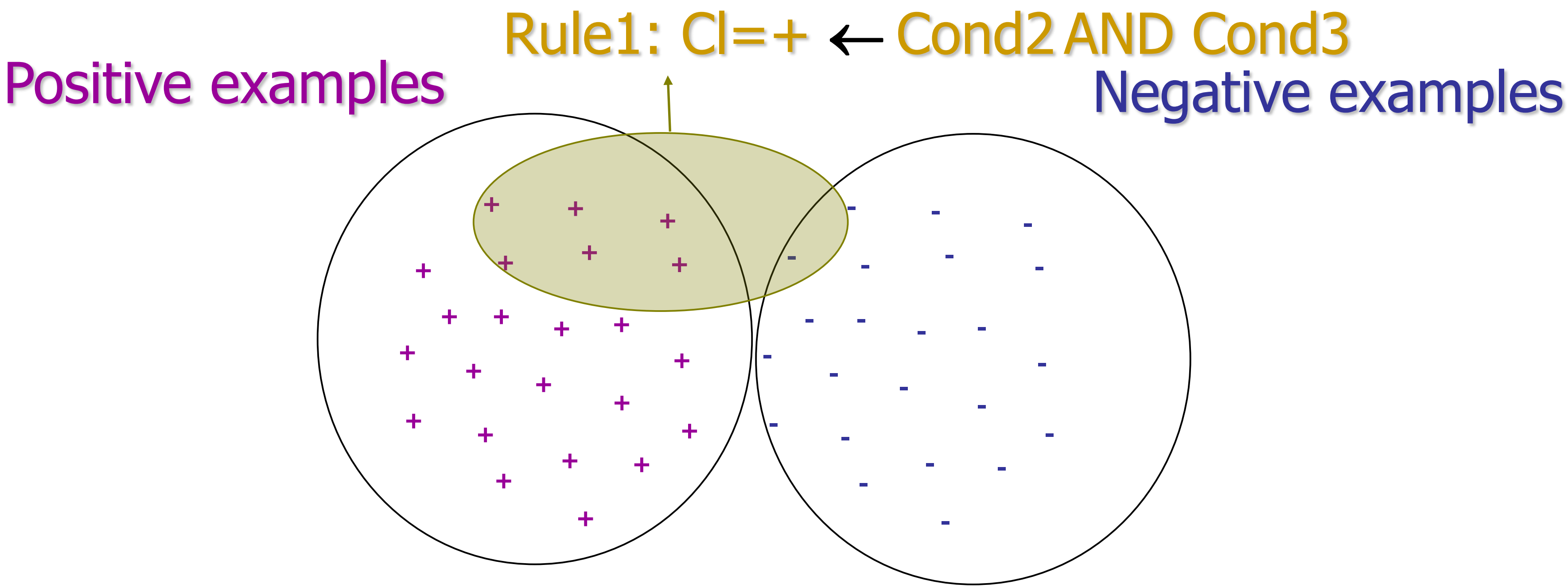
Covering algorithm

Positive examples

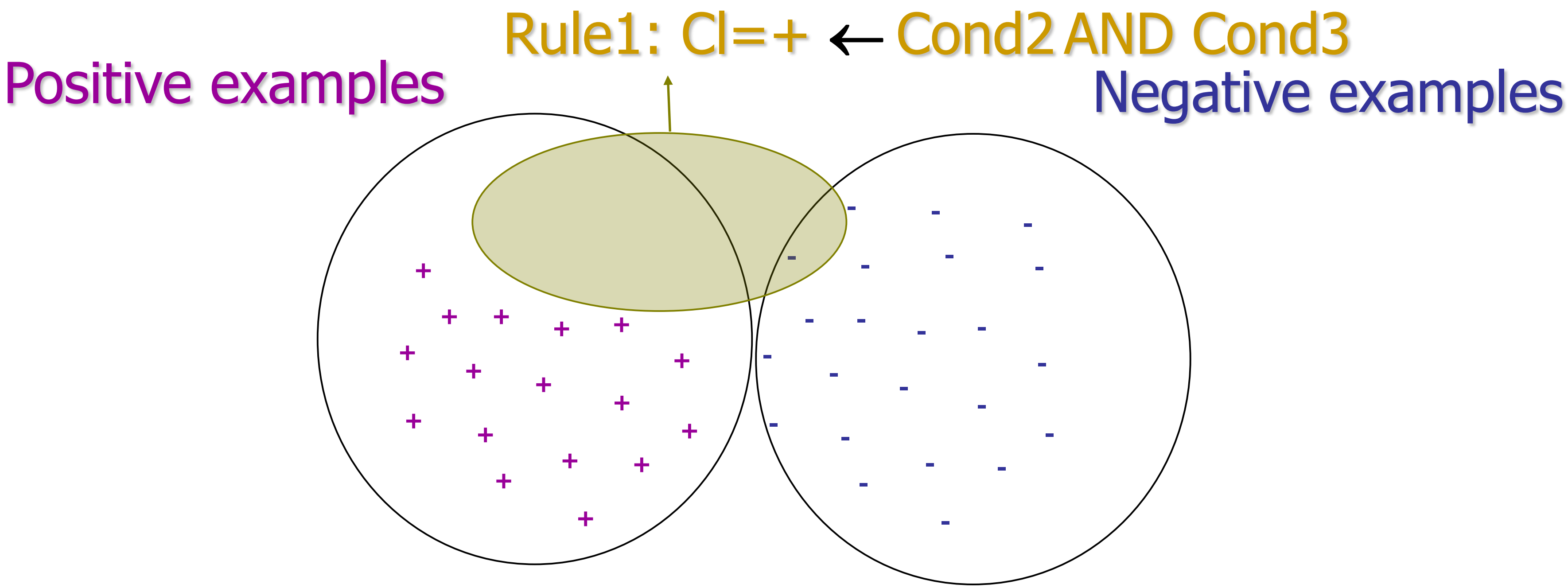
Negative examples



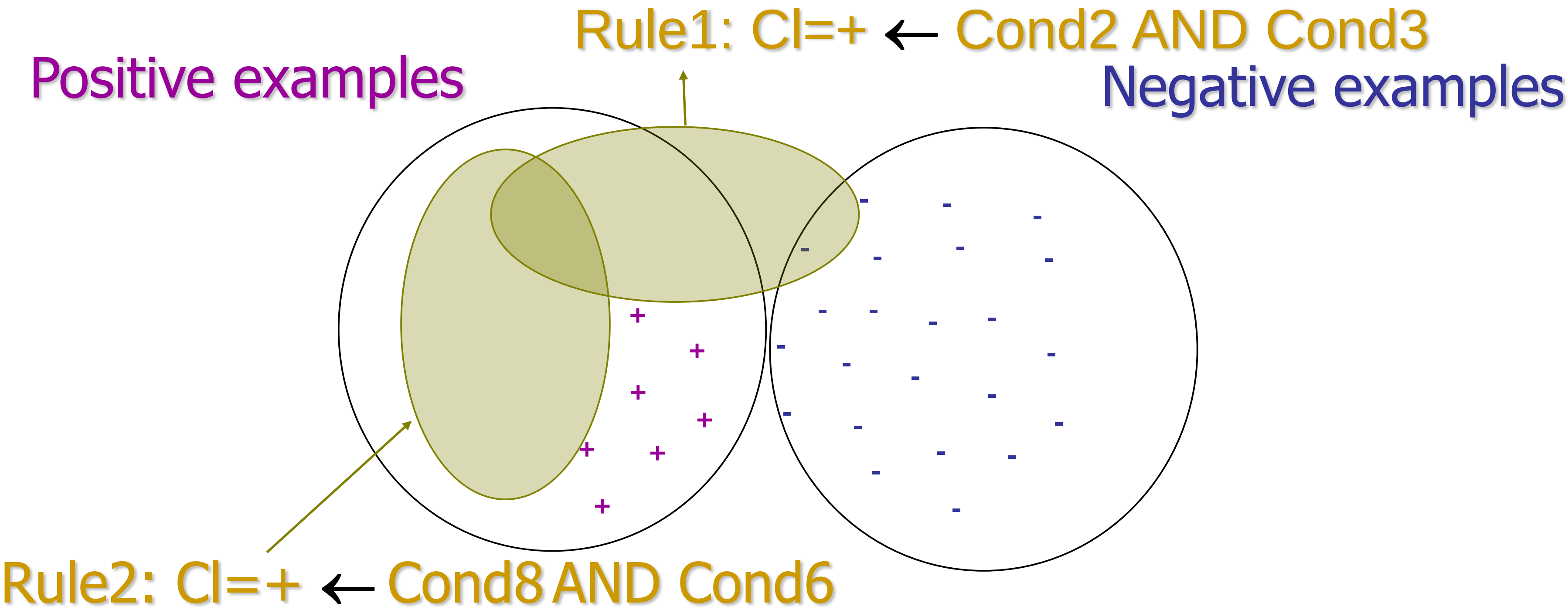
Covering algorithm



Covering algorithm

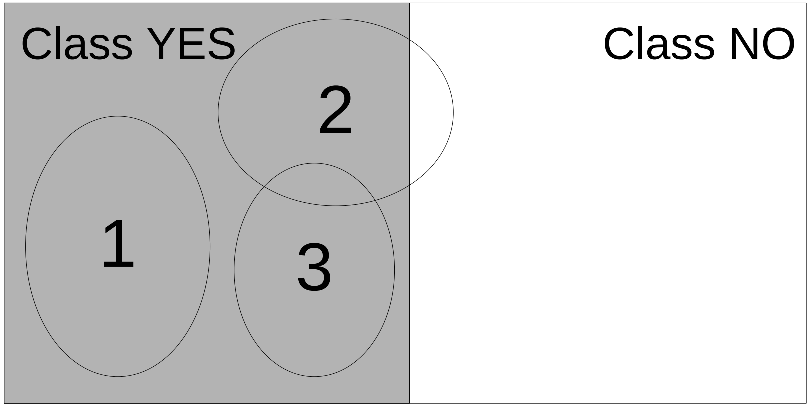


Covering algorithm

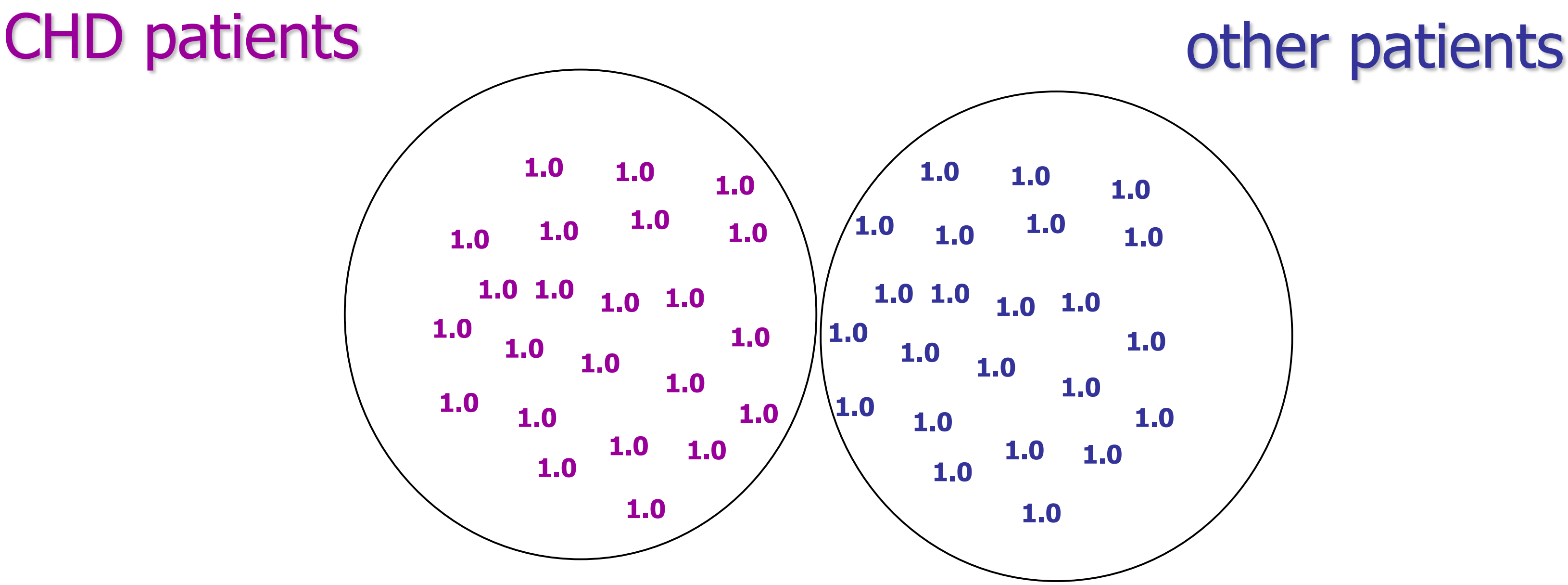


Characteristics of SD Algorithms

- 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.



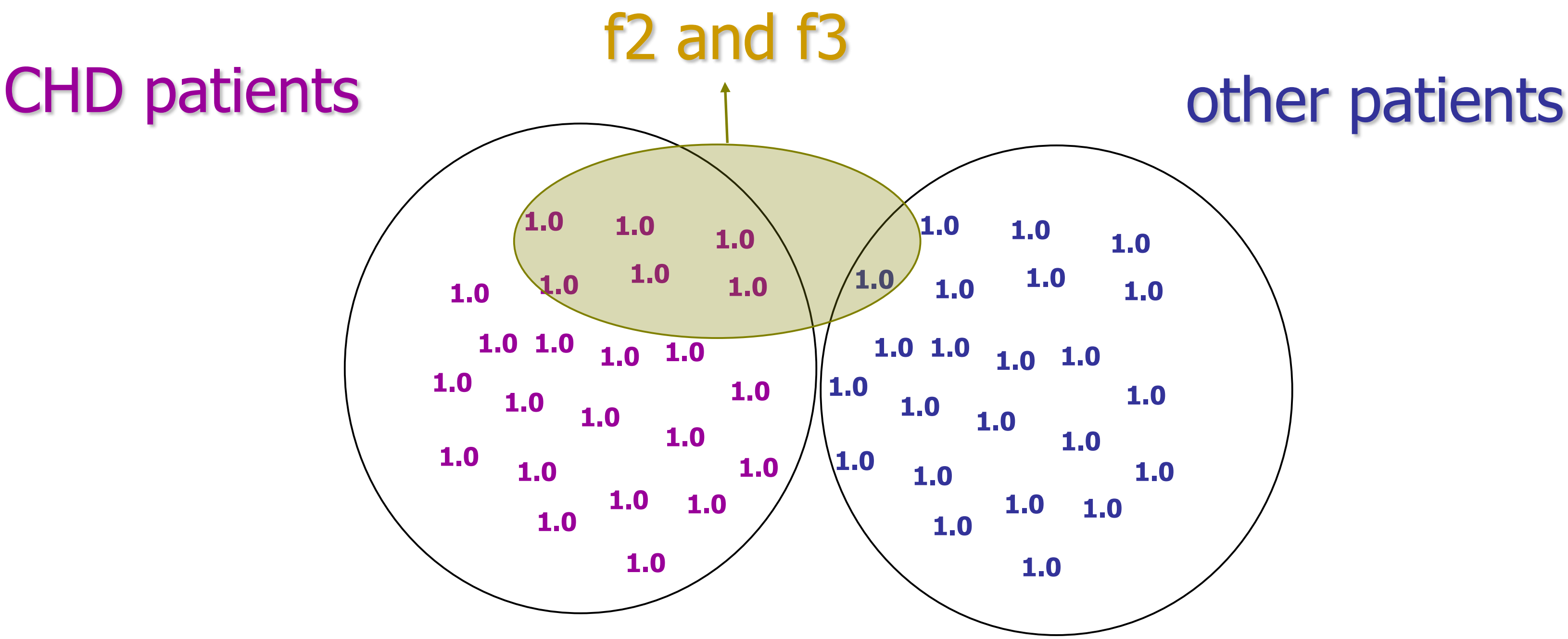
Weighted covering algorithm for rule set construction



- 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

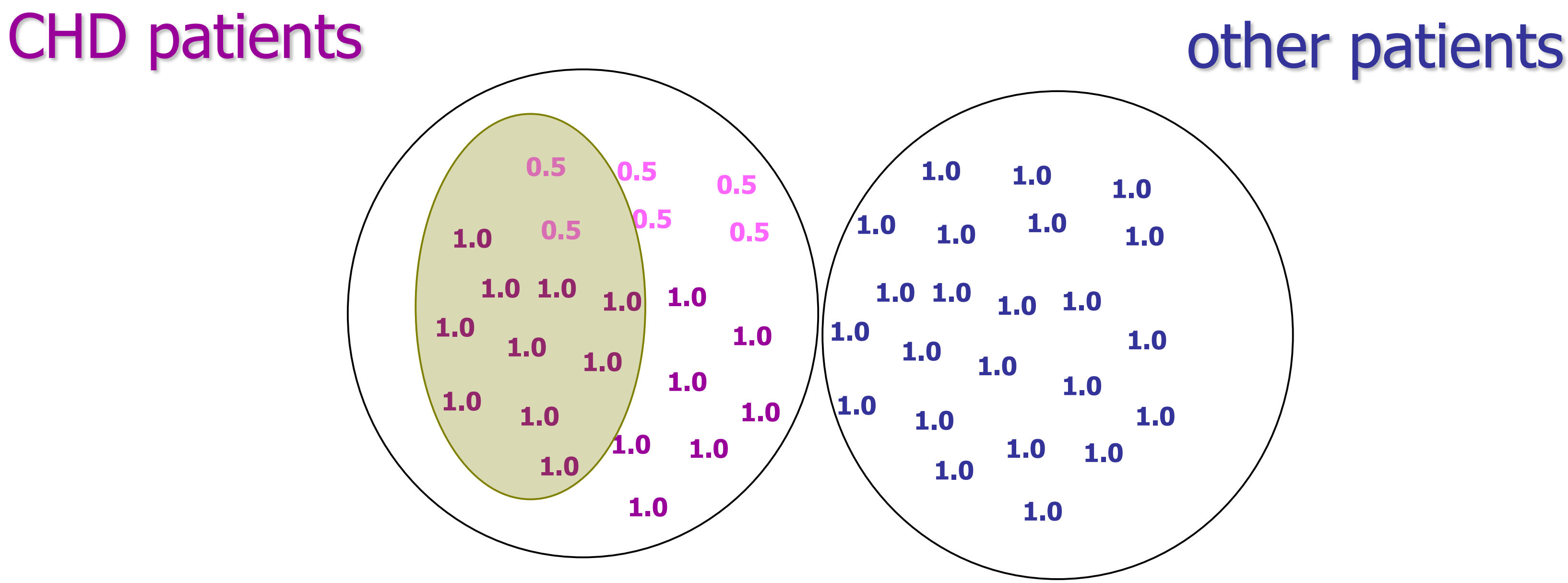


Rule quality measure in SD: q(Cl ← Cond) = TP/(FP+g)

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

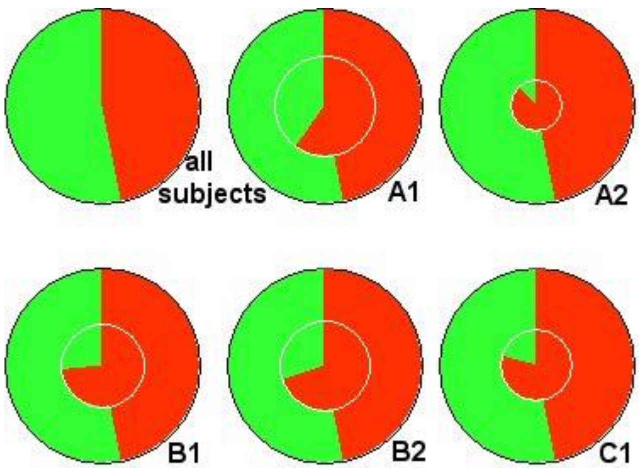
*__Coverage__ = sum of the covered weights, *__Precision__ = purity of the covered examples

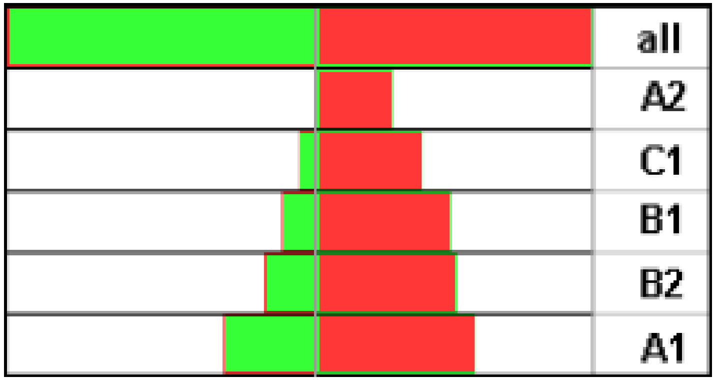
Weighted covering algorithm for rule set construction

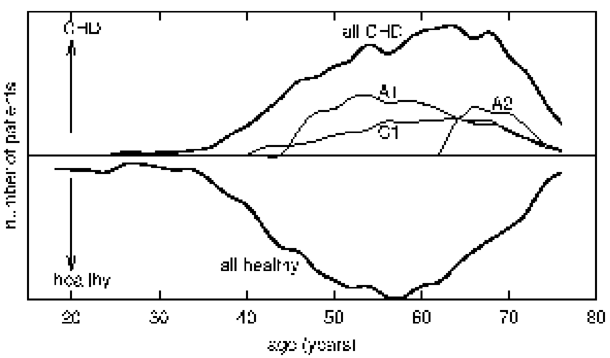


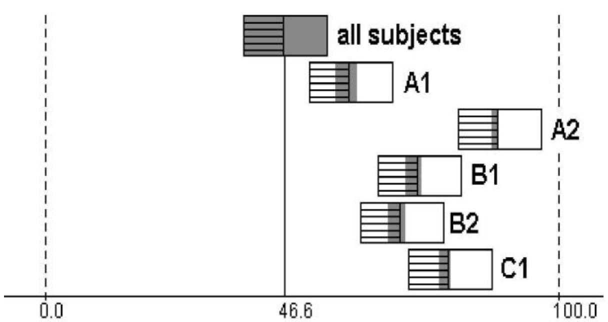
In contrast with classification rule learning algorithms (e.g. CN2), the covered positive examples are not deleted from the training set in the next rule learning iteration; they are re-weighted, and a next ‘best’ rule is learned.

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)

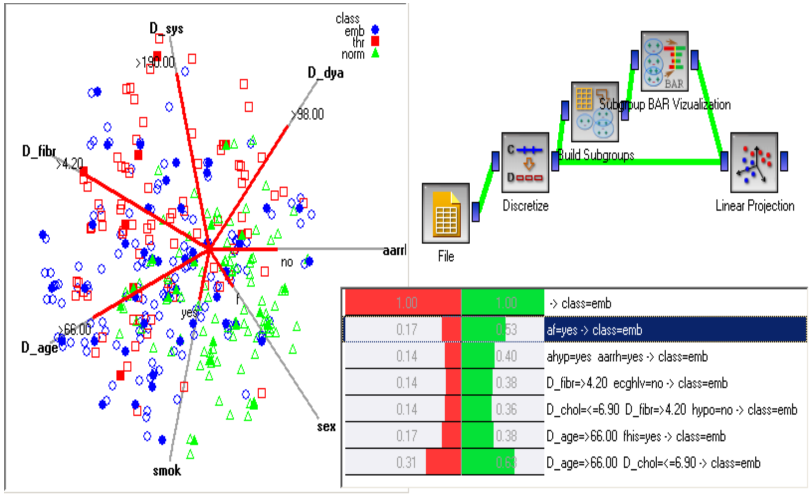
AND

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.

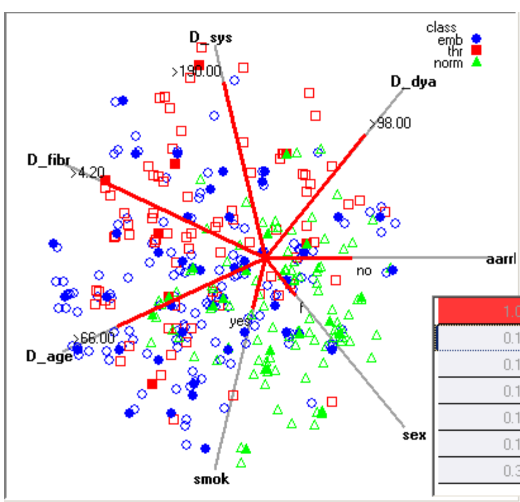
SD algorithms in the Orange DM Platform

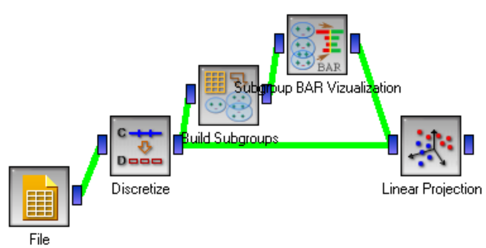
- **SD Algorithms in Orange**
  - SD (Gamberger & Lavrač, JAIR 2002
  - APRIORI-SD (Kavšek & Lavrač, AAI 2006
  - CN2-SD (Lavrač et al., JMLR 2004): Adapting CN2 classification rule learner to Subgroup Discovery
    - Weighted covering algorithm
    - Weighted relative accuracy (WRAcc) search heuristics,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

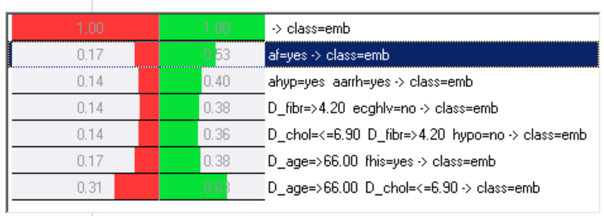


SD algorithms in Orange and Orange4WS

- **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







Selected Data Mining Techniques Outline

- Subgroup discovery
- 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	So St	In come	Age	Club	Resp
...	...	...	...	...	...	...	...
3478	34677	m	si	60-70	32	me	nr
3479	43666	f	ma	80-90	45	nm	re
...	...	...	...	...	...	...	...

order

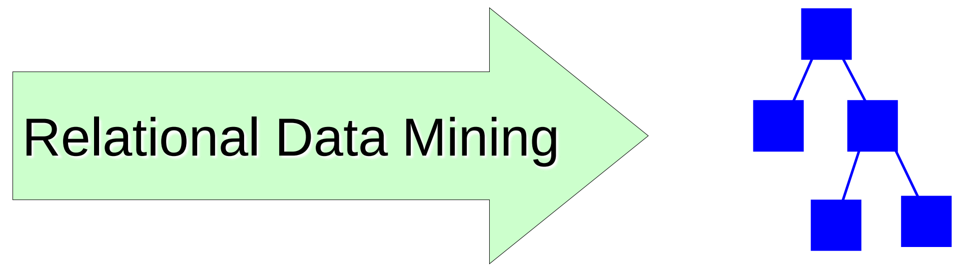
Customer ID	Order ID	Store ID	Delivery Mode	Paymt Mode
...	...	...	...	...
3478	2140267	12	regular	cash
3478	3446778	12	express	check
3478	4728386	17	regular	check
3479	3233444	17	express	credit
3479	3475886	12	regular	credit
...	...	...	...	...

store

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

Relational representation of customers, orders and stores.

knowledge discovery
from data



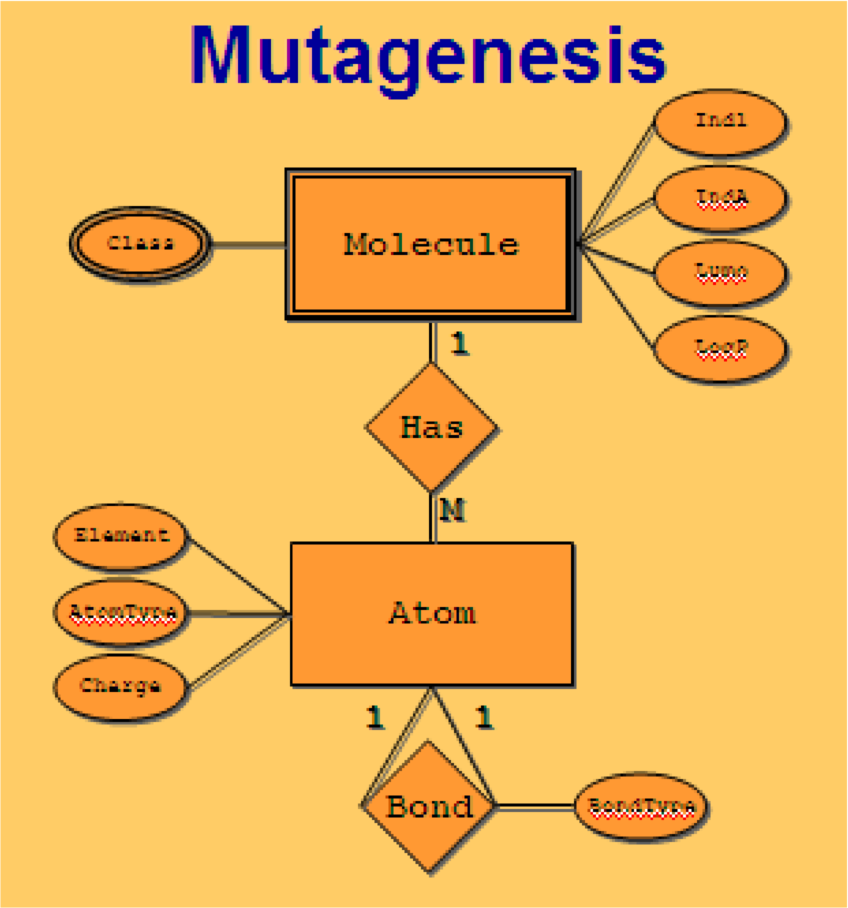
model, patterns, …

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

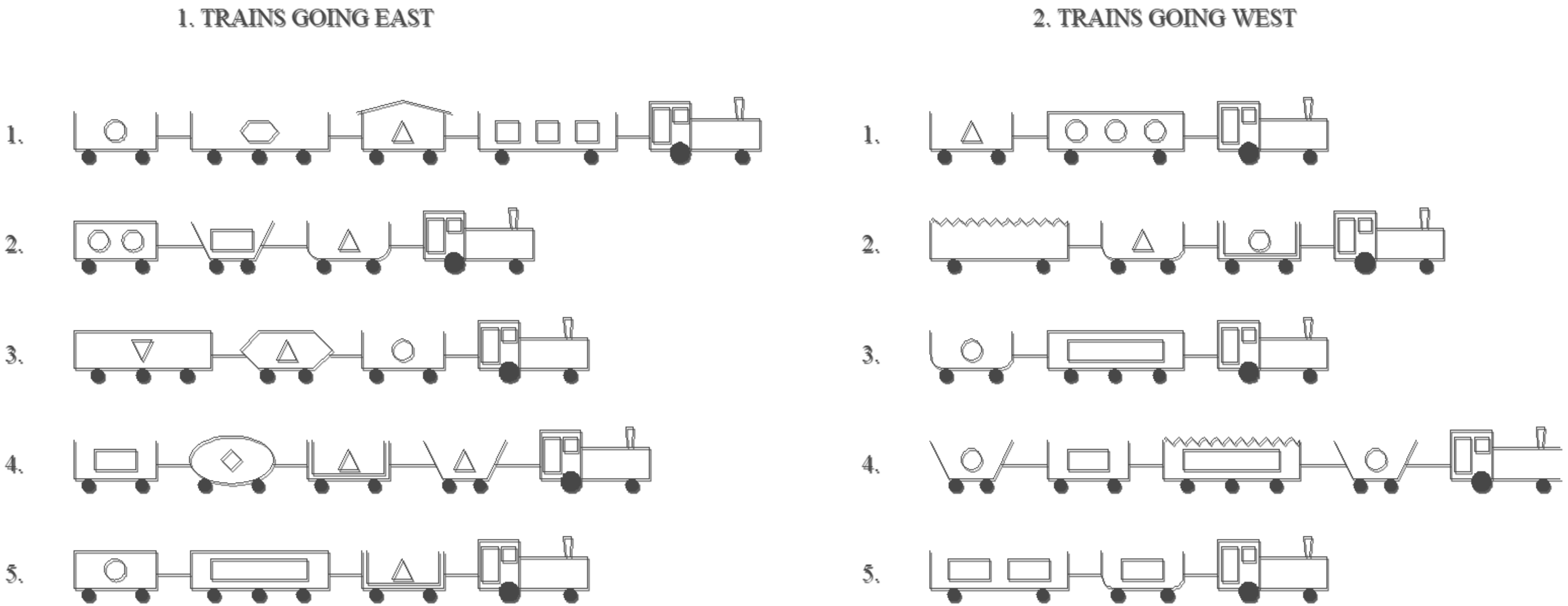
Find: a classification model, a set of interesting patterns

Relational Data Mining (ILP)

- Learning from multiple tables
  - 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

TRAIN_TABLE

TRAIN	EASTBOUND
t1	TRUE
t2	TRUE
…	…
t6	FALSE
…	…

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

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
…	…	…			…

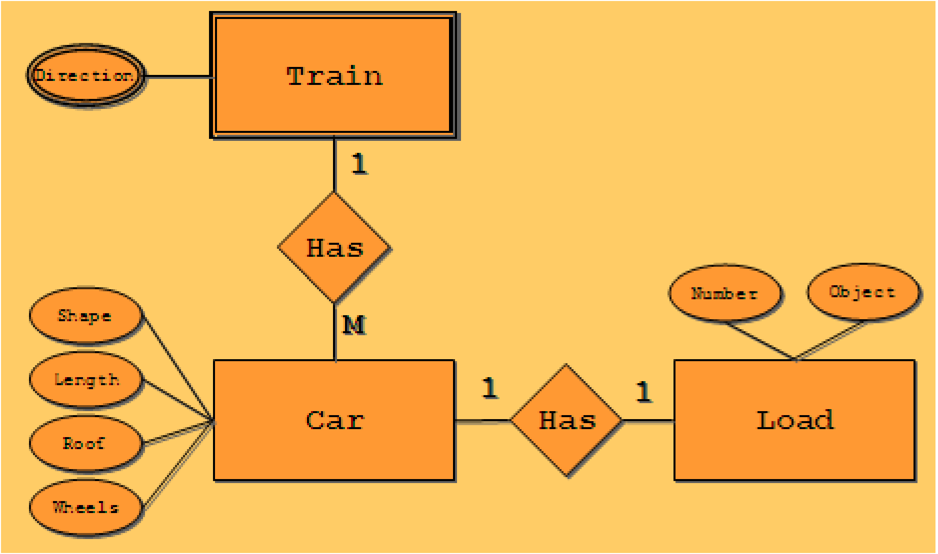
Relational data representation

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

TRAIN_TABLE

TRAIN	EASTBOUND
t1	TRUE
t2	TRUE
…	…
t6	FALSE
…	…

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

Propositionalization task

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), …

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

TRAIN_TABLE

TRAIN	EASTBOUND
t1	TRUE
t2	TRUE
…	…
t6	FALSE
…	…

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

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) $\leftarrow$ f1(T), f4(T)

Relational interpretation:

eastbound(T) $\leftarrow$
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

TRAIN	EASTBOUND
t1	TRUE
t2	TRUE
...	...
t6	FALSE
...	...

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	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

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

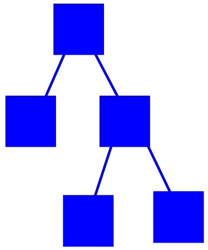
Propositionalization

	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

	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



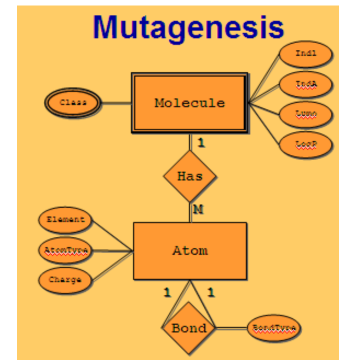
model, patterns, …

Relational Data Mining in Orange4WS

- Propositionalization workflow in Orange4WS
- RSD as a service for propositionalization through efficient first-order feature construction

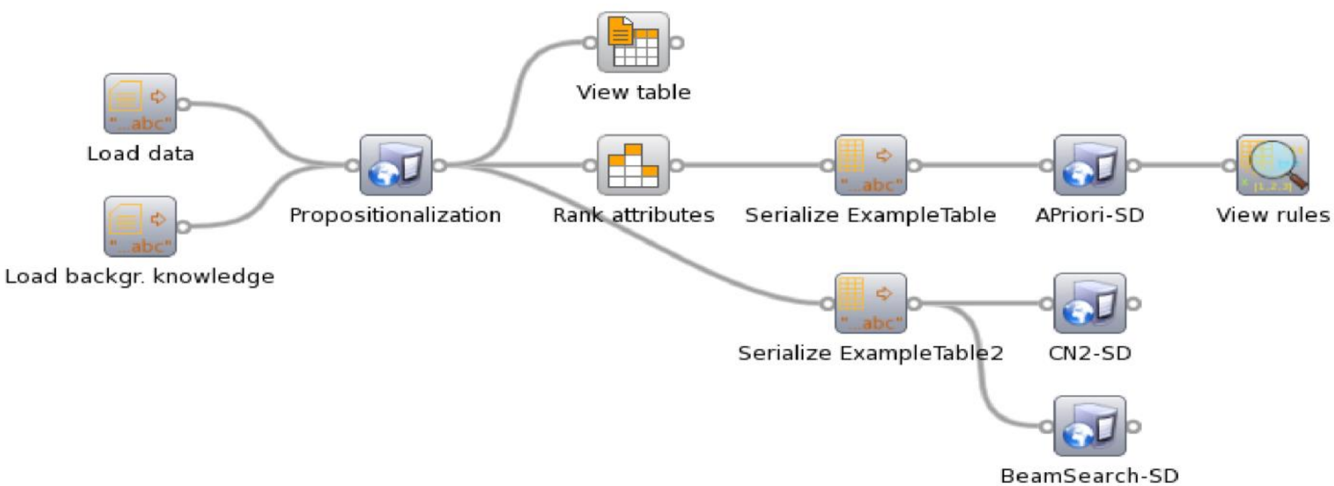
  f121(M):- hasAtom(M,A), atomType(A,21)

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



- subgroup discovery using CN2-SD

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



Selected Data Mining Techniques
Outline

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

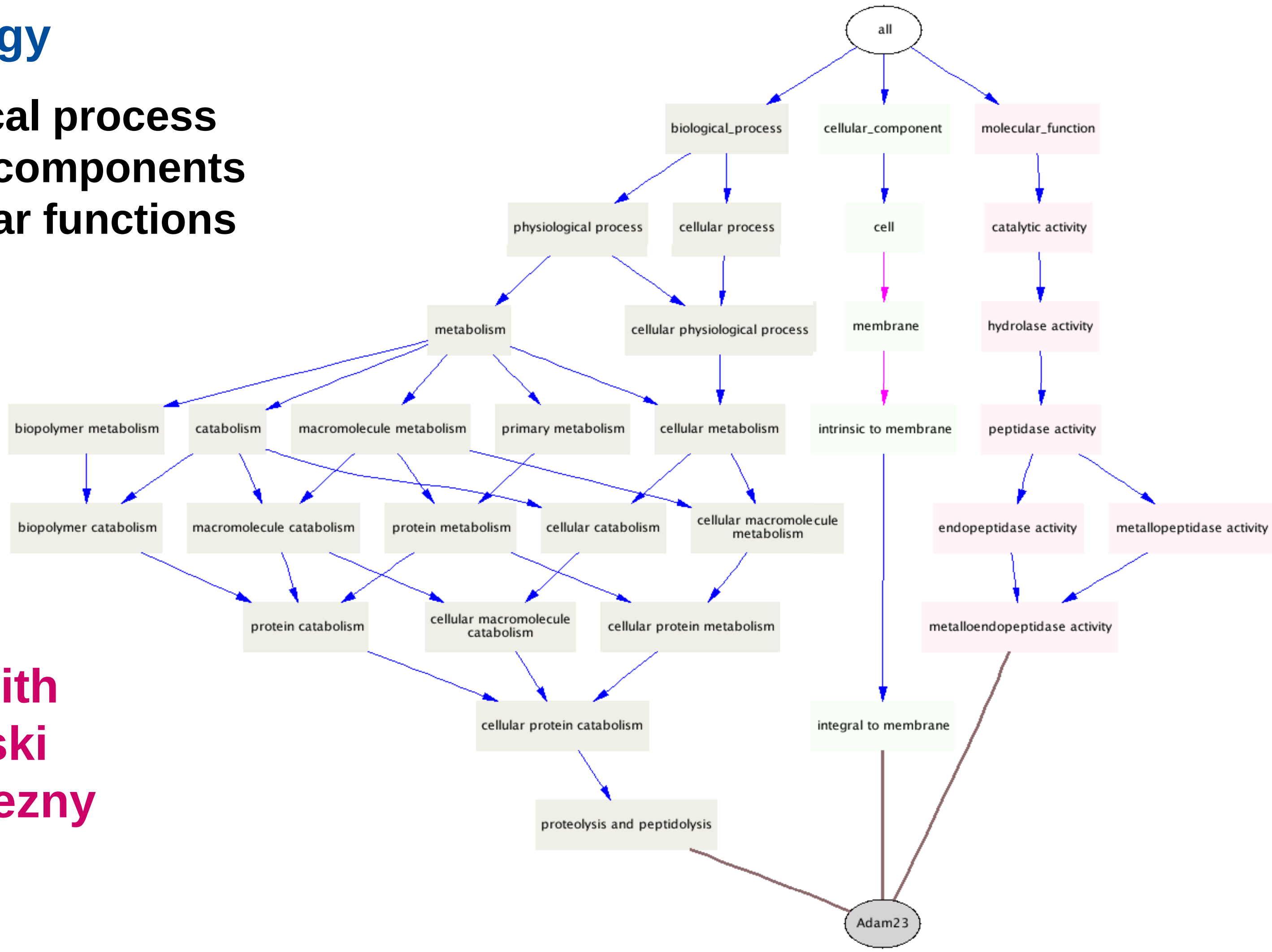
Semantic Data Mining in Orange4WS

- 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

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

Gene Ontology

12093 biological process
1812 cellular components
7459 molecular functions

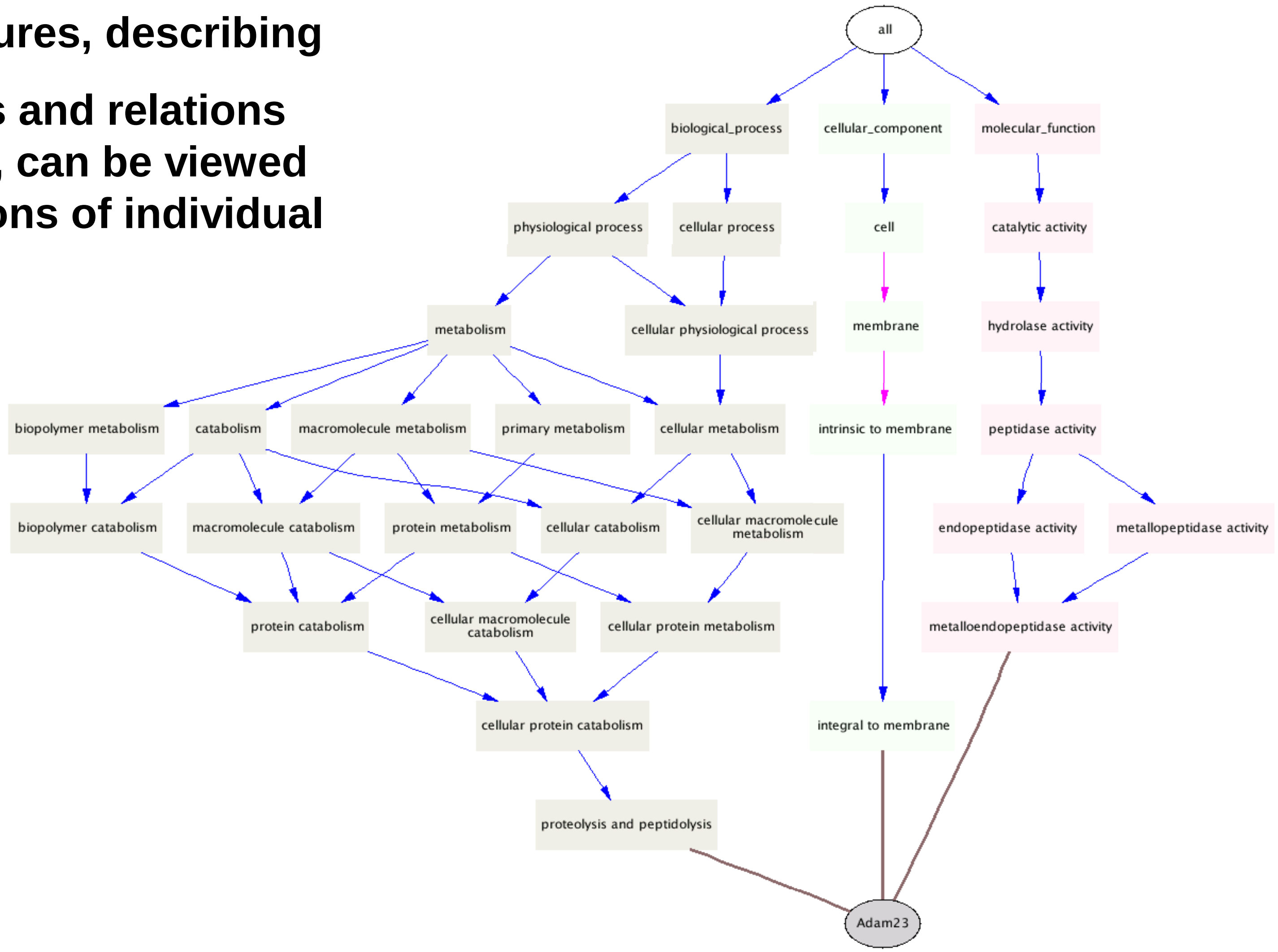


Joint work with Igor Trajkovski and Filip Zelezny

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').
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').

existential

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

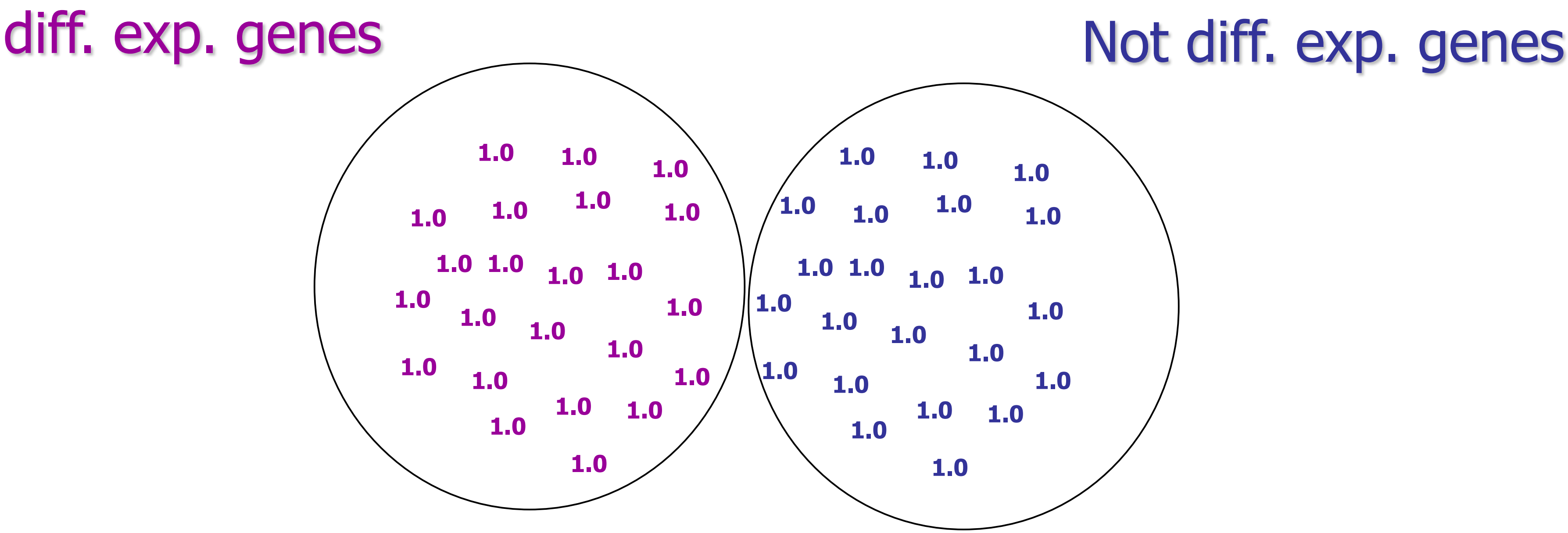
Propositional learning: subgroup discovery

	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

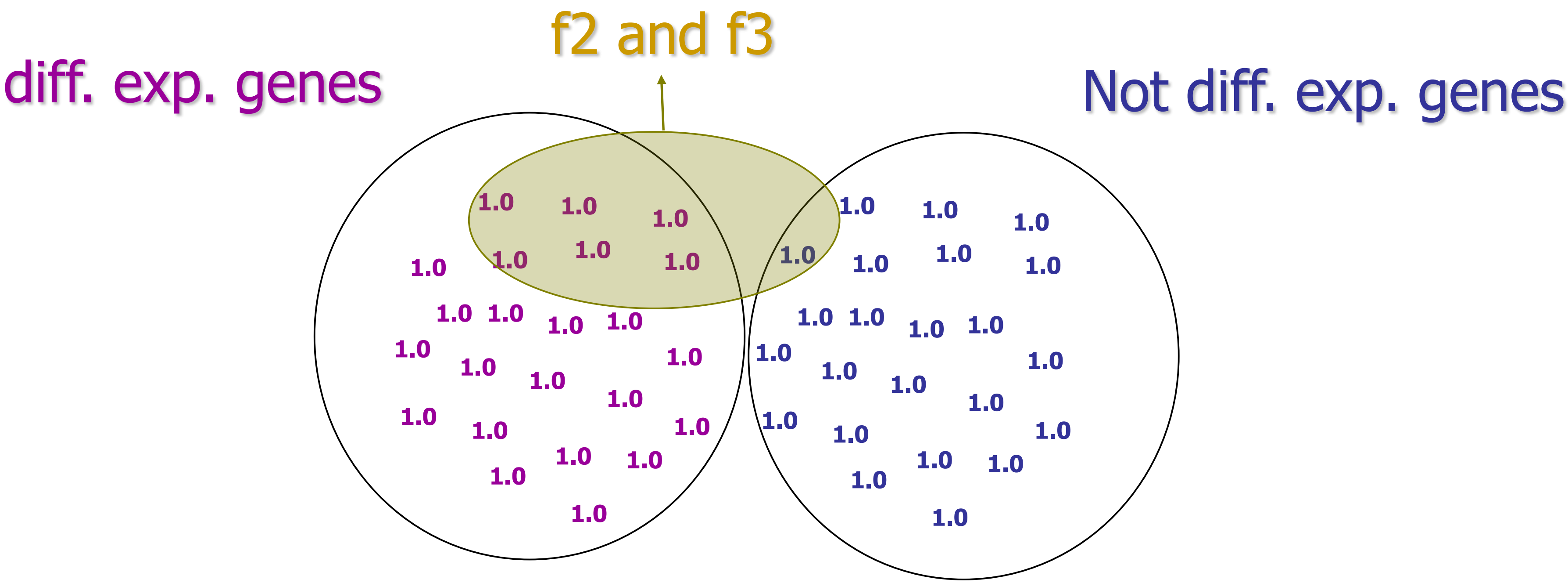
f2 and f3

[4,0]

Subgroup Discovery



Subgroup Discovery



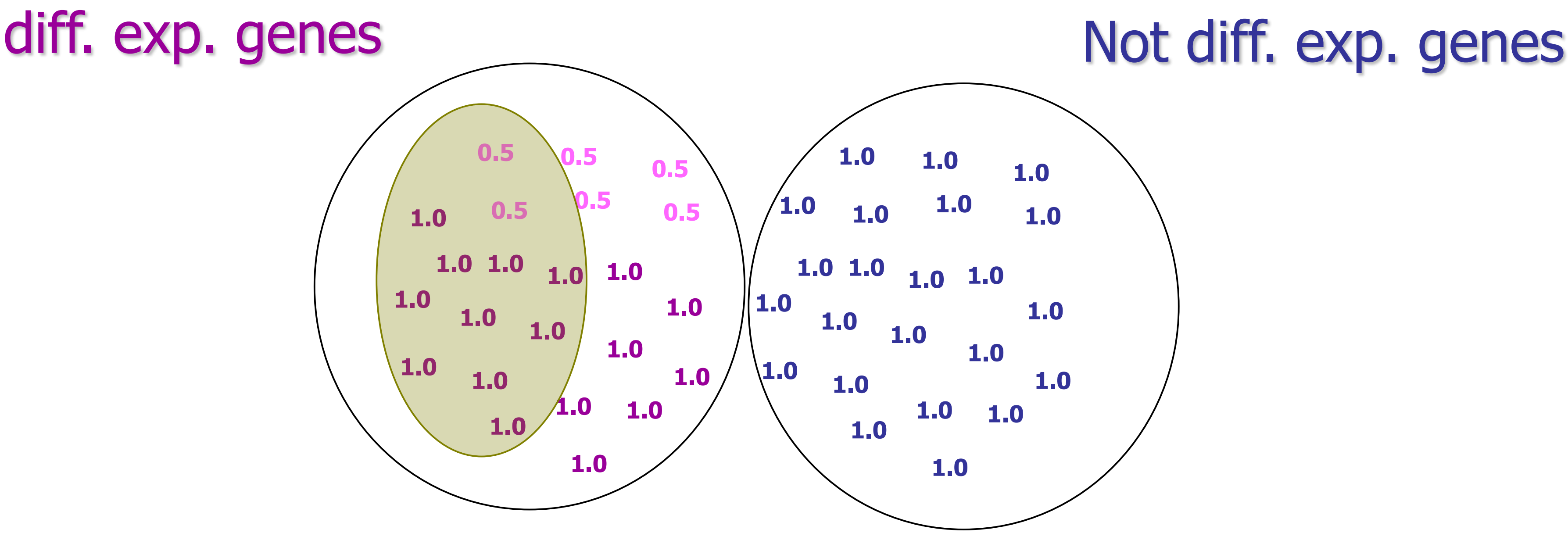
In RSD (using propositional learner CN2-SD):

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 on semantic data mining in Orange4WS, (generalizing SEGS to g-SEGS, SDM-SEGS, and SDM-Aleph) done in collaboration with A. Vavpetič

Outline

- **JSI & Knowledge Technologies**
- **Introduction to Data mining and KDD**
  - Data Mining and KDD process
  - DM standards, tools and visualization
  - Classification of Data Mining techniques: Predictive and descriptive DM
- **Selected data mining techniques: Advanced subgroup discovery techniques and applications**
- **Relation between data mining and text mining**

Data mining vs. text mining

Data mining:

- instances are objects, belonging to different classes
- instances are feature vectors, described by attribute values
- classification model is learned using data mining algorithms

Task reformulation: Binarization

Person	Young	Myope	Astigm.	Reuced tea	Lenses
O1	1	1	0	1	NO
O2	1	1	0	0	YES
O3	1	1	1	1	NO
O4	1	1	1	0	YES
O5	1	0	0	1	NO
O6-O13	...	...	...	...	...
O14	0	0	0	0	YES
O15	0	0	1	1	NO
O16	0	0	1	0	NO
O17	0	1	0	1	NO
O18	0	1	0	0	NO
O19-O23	...	...	...	...	...
O24	0	0	1	0	NO

Binary features and class values

Data mining vs. text mining

Data mining:

- instances are objects, belonging to different classes
- instances are feature vectors, described by attribute values
- classification model is learned using data mining algorithms

Text mining:

- instances are text documents
- text documents need to be transformed into feature vector representation in data preprocessing
- data mining algorithms can then be used for learning the model

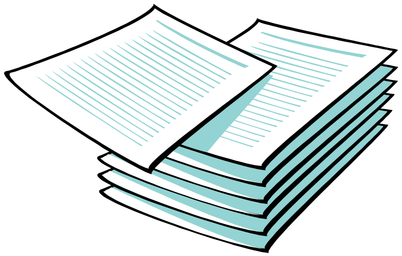
Text mining: Words/terms as binary features

Document	Word1	Word2	…	WordN	Class
d1	1	1	0	1	NO
d2	1	1	0	0	YES
d3	1	1	1	1	NO
d4	1	1	1	0	YES
d5	1	0	0	1	NO
d6-d13	...	...	...	...	...
d14	0	0	0	0	YES
d15	0	0	1	1	NO
d16	0	0	1	0	NO
d17	0	1	0	1	NO
d18	0	1	0	0	NO
d19-d23	...	...	...	...	...
d24	0	0	1	0	NO

Instances = documents

Words and terms = Binary features

Text mining



Step 1

BoW vector construction

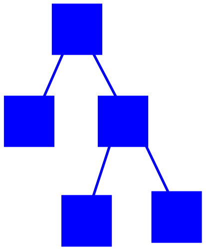
Document	Word1	Word2	…	WordN	Class
d1	1	1	0	1	NO
d2	1	1	0	0	YES
d3	1	1	1	1	NO
d4	1	1	1	0	YES
d5	1	0	0	1	NO
d6-d13	…	…	…	…	…
d14	0	0	0	0	YES
d15	0	0	1	1	NO
d16	0	0	1	0	NO
d17	0	1	0	1	NO
d18	0	1	0	0	NO
d19-d23	…	…	…	…	…
d24	0	0	1	0	NO

1. BoW features construction
2. Table of BoW vectors construction

Document	Word1	Word2	…	WordN	Class
d1	1	1	0	1	NO
d2	1	1	0	0	YES
d3	1	1	1	1	NO
d4	1	1	1	0	YES
d5	1	0	0	1	NO
d6-d13	…	…	…	…	…
d14	0	0	0	0	YES
d15	0	0	1	1	NO
d16	0	0	1	0	NO
d17	0	1	0	1	NO
d18	0	1	0	0	NO
d19-d23	…	…	…	…	…
d24	0	0	1	0	NO

Step 2

Data Mining



model, patterns, clusters, …

Text Mining process

- Text preprocessing for feature construction
  - StopWords elimination
  - Word stemming or lemmatization
  - Term construction by frequent N-Grams construction
  - Terms obtained from thesaurus (e.g., WordNet)

- BoW vector construction

- Data Mining of BoW vector table
  - Text Categorization, Clustering, Summarization, …

Text Mining from unlabeled data

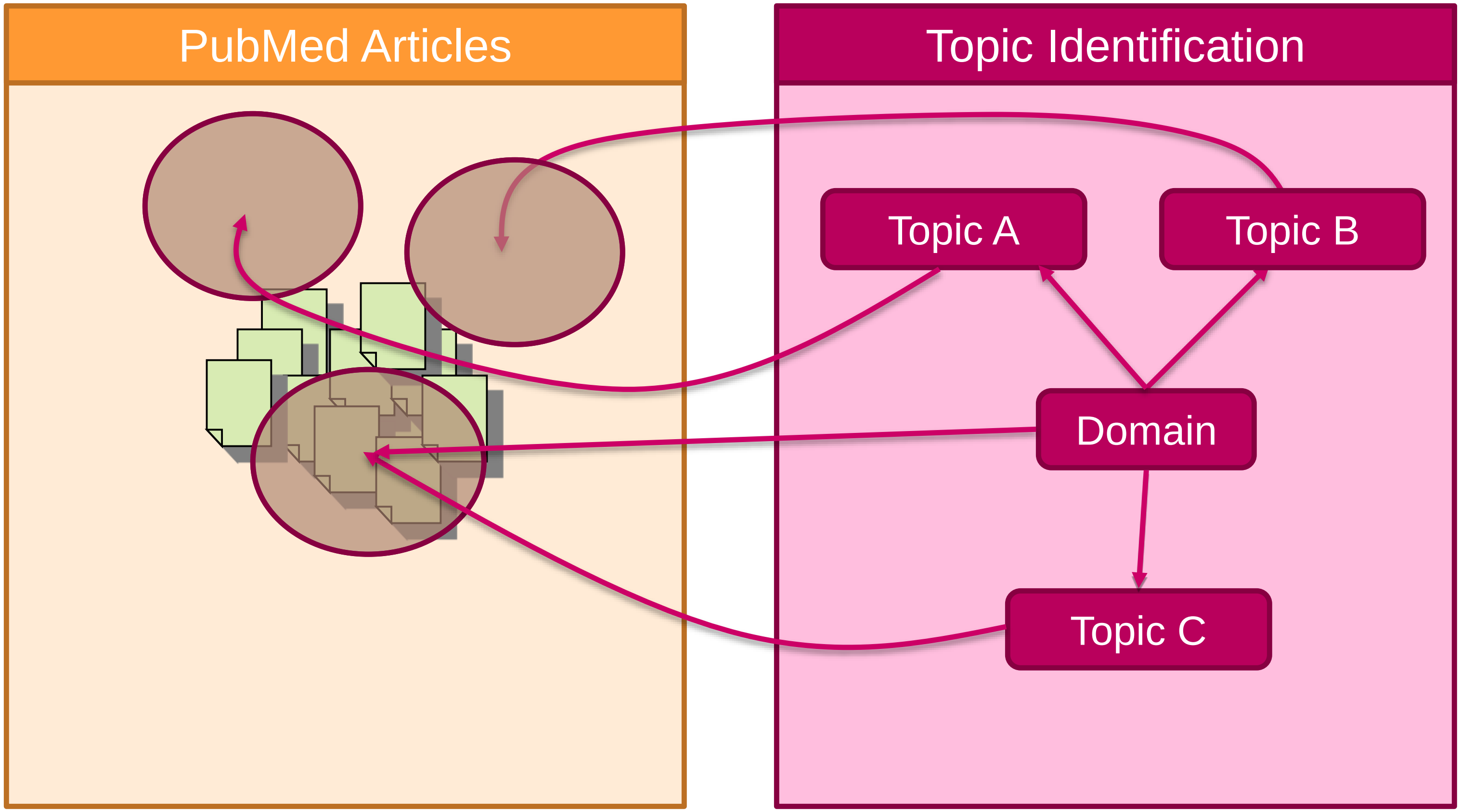
Document	Word1	Word2	…	WordN	Class
d1	1	1	0	1	NO
d2	1	1	0	0	YES
d3	1	1	1	1	NO
d4	1	1	1	0	YES
d5	1	0	0	1	NO
d6-d13	...	...	...	...	...
d14	0	0	0	0	YES
d15	0	0	1	1	NO
d16	0	0	1	0	NO
d17	0	1	0	1	NO
d18	0	1	0	0	NO
d19-d23	...	...	...	...	...
d24	0	0	1	0	NO

Unlabeled data - clustering: grouping of similar instances

Scientific literature in PubMed: source of knowledge for Text Mining

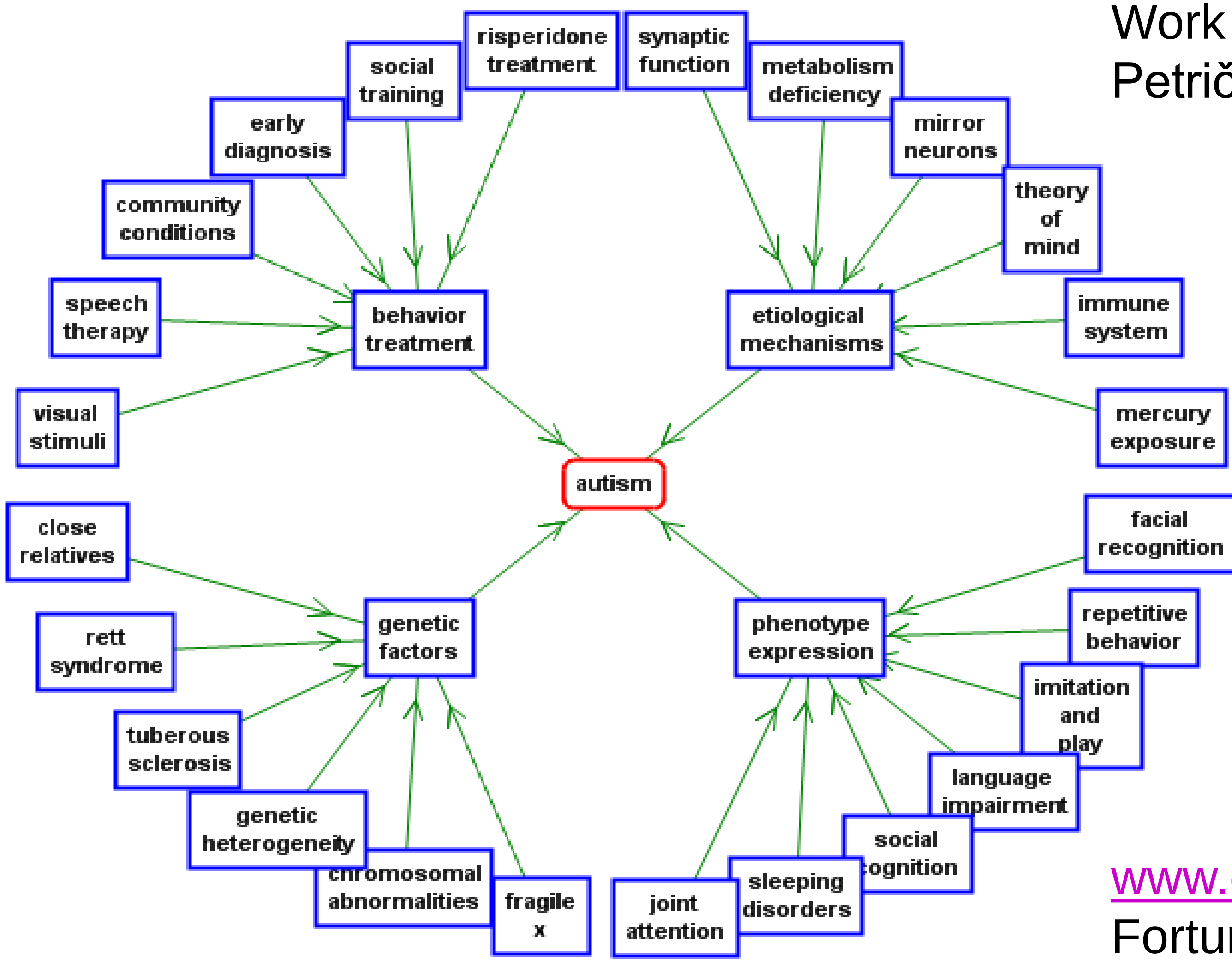
- Biomedical bibliographical database PubMed
- US National Library of Medicine
- More than 21M citations
- More than 5,600 journals
- 2,000 – 4,000 references added each working day!

Text Mining Example: Clustering of PubMed Articles



Slide adapted from D. Mladenić, JSI

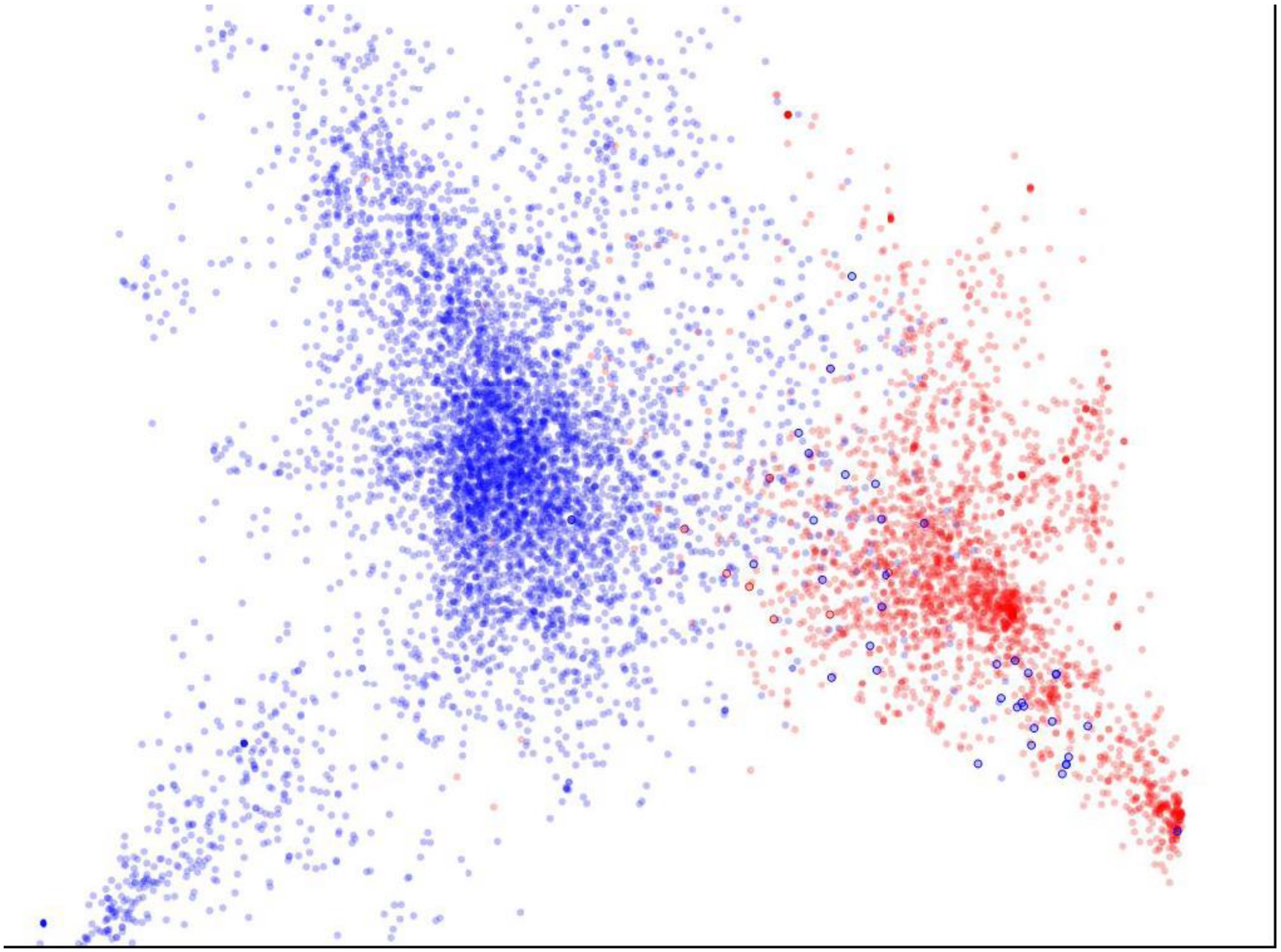
OntoGen Applied to Clustering of PubMed Articles on Autistic Spectrum Disorders



Work by
Petrič et al. 2009

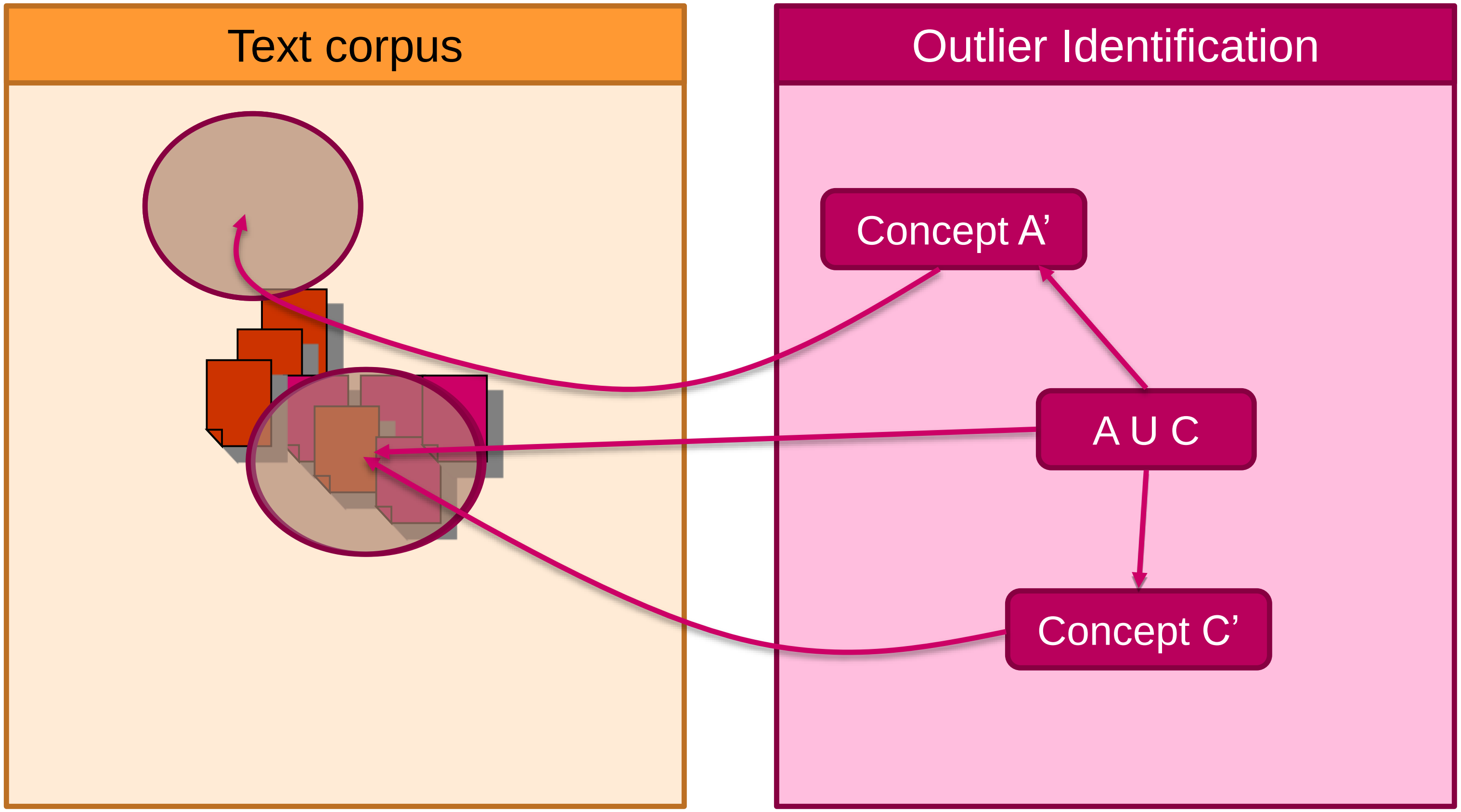
www.ontogen.si
Fortuna, Mladenić,
Grobelnik 2006

Outlier analysis from two document sets



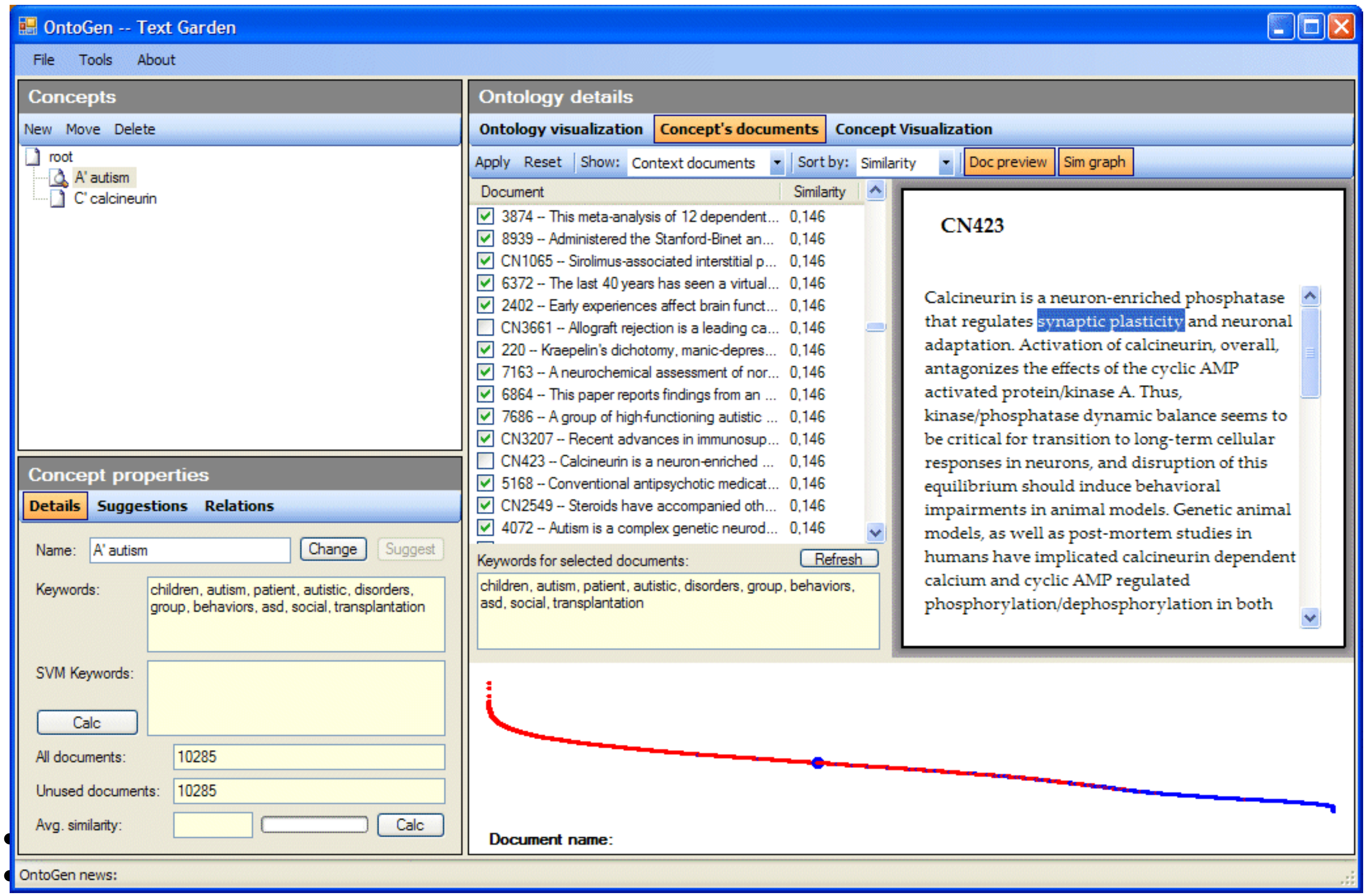
2-dimensional projection of documents (about autism (red) and calcineurin (blue). Outlier documents are bolded for the user to easily spot them.

Using OntoGen for Outlier Document Identification



Slide adapted from D. Mladenić, JSI

Using OntoGen on autism-calcineurin data: Outlier calcineurin document CN423



Work by
Petrič et al. 2010

Summary

- **JSI & Knowledge Technologies**
- **Introduction to Data Mining and KDD**
  - Data Mining and KDD process
  - DM standards, tools and visualization
  - Classification of Data Mining techniques: Predictive and descriptive DM
- **Selected Data Mining techniques: Advanced subgroup discovery techniques and applications**
- **Relation between data mining and text mining**