Subgroup Discovery Techniques and Applications

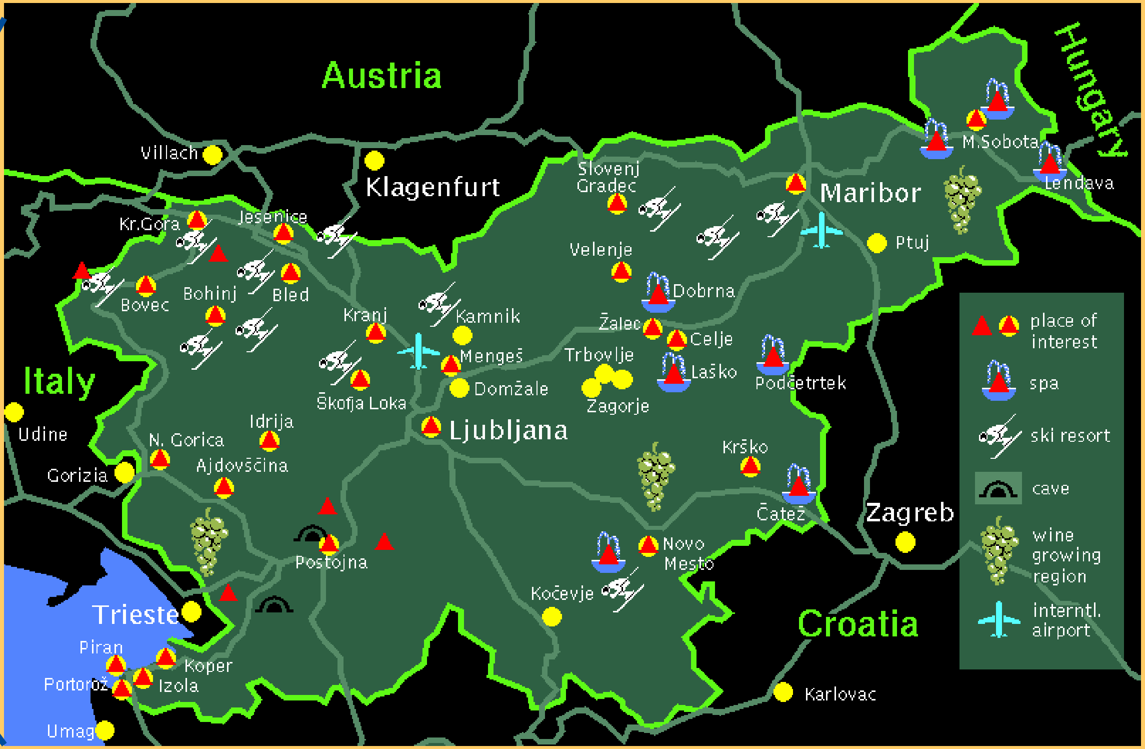
Nada Lavrač

Department of Knowledge Technologies
Jožef Stefan Institute
Ljubljana, Slovenia

PAKDD-2005 paper - joint work with Dragan Gamberger

Slovenia – Ljubljana (capital)

Europe



Ljubljana, Slovenia

Jožef Stefan Institute - Profile

- **Jožef Stefan Institute (founded in 1949) is the leading national research organization in natural sciences and technology**
  - information and communication technologies
  - chemistry, biochemistry & nanotechnology
  - physics, nuclear technology and safety
- **Jožef Stefan International Postgraduate School (founded in 2004) offers MSc and PhD programs**
  - ICT, nanotechnology, ecotechnology
  - research oriented, basic + management courses
  - in English
- **~ 500 researchers and students**

Jožef Stefan - Profile

Jožef Stefan - Profile

$$j = \sigma T^4$$

- Jožef Stefan (1835-1893) was one of the most distinguished physicists of the nineteenth century

- He originated the law of the total radiation from a black body

Department of Knowledge Technologies

- **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 in medicine, ecological modeling, business, virtual enterprises, …**

Department of Knowledge Technologies

- **National funding (50%):**
  - Basic research project on Knowledge Technologies
  - Other national R&D projects, and client applications
- **Current EU funding (50%):**
  - **ECOLEAD** – Collaborative Networked Organizations
  - **SEKT** – Semantically Enabled Knowledge Technologies
  - **IQ** – Constraint-based DM and Inductive databases
  - **ALVIS** – Next Generation Search Engine
  - **SIGMEA** – Data mining in Ecological Modeling
  - **PASCAL** - Network of Excellence in Machine Learning
  - **IST-World** – Supporting EU project analysis
  - **WYS-CEC** – Supporting Women and Youth in Science

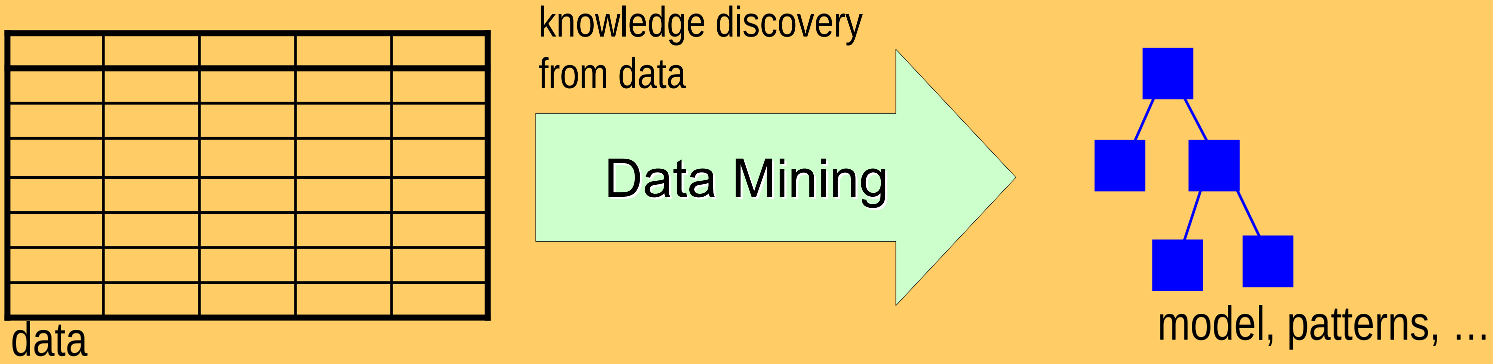
The SolEuNet Project

- Data Mining and Decision Support for Business Competitiveness: A European Virtual Enterprise, EU project, 2000-2003
- 3 Mio EUR, 12 partners (8 academic and 4 business) from 7 countries, coordinated by JSI
- Main SolEuNet objectives: build prototype applications by
  - developing new data mining techniques, by integrating and combining data mining and decision support
  - extending CRISP-DM to collaborative problem solving in virtual organizations

Selected prototype applications

- Mediana – analysis of media research data
- Kline & Kline – improved brand name recognition
- Australian financial house – customer quality evaluation, stock market prediction
- Czech health farm – predict the use of spa resources
- INE Port. statistical buro – Web page access analysis for better Web page organization
- **CHD - Coronary heart disease risk group detection**
- Online Dating – understanding email dating promiscuity
- EC Harris - analysis of building construction projects
- Traffic - analysis of 20 years road traffic accidents in UK
- Project Intelligence - analysis of Web documents describing 5FP IST projects
- …

Basic DM and DS processes



knowledge discovery from data

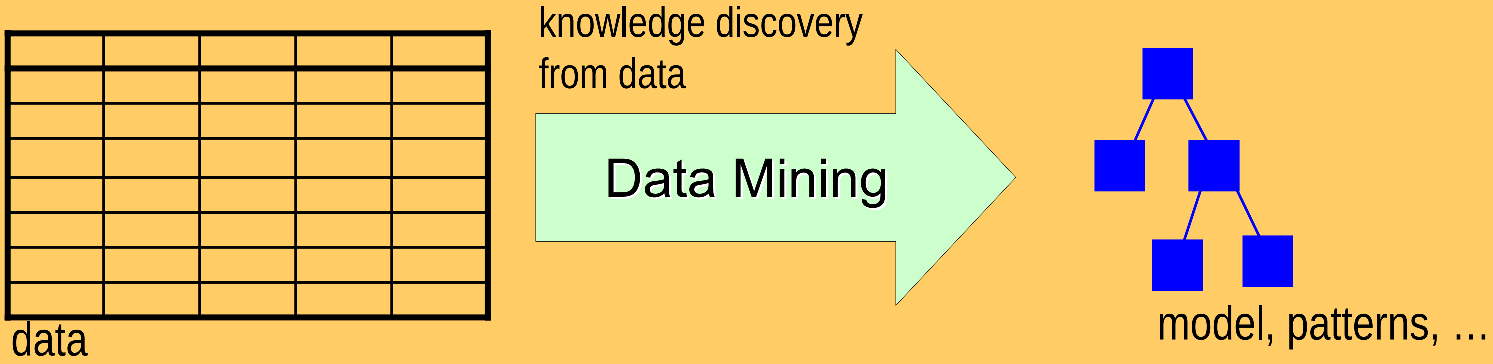
Data Mining

data

model, patterns, …

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

Basic DM and DS processes



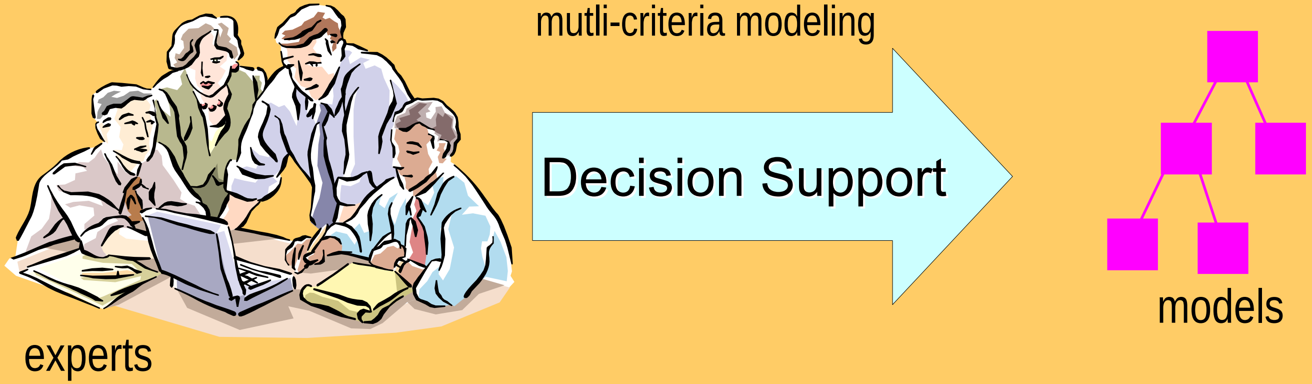
knowledge discovery from data

Data Mining

data

model, patterns, …

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



mutli-criteria modeling

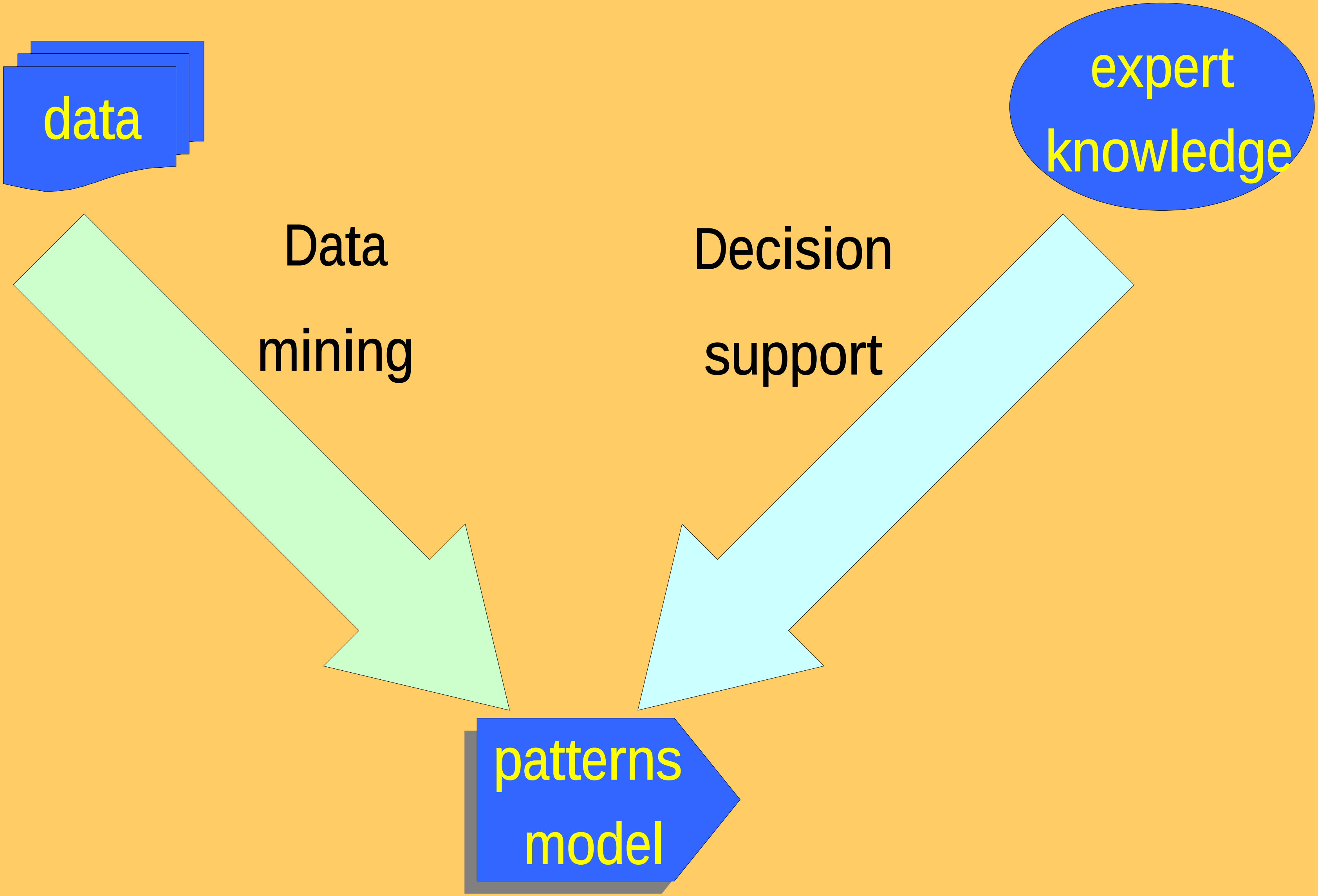
Decision Support

experts

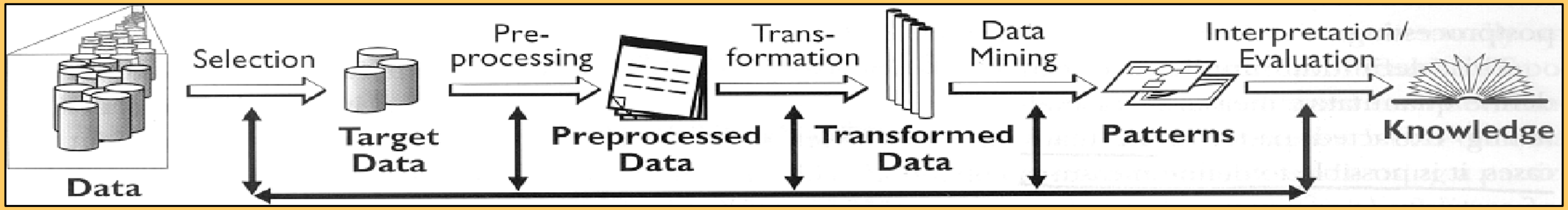
models

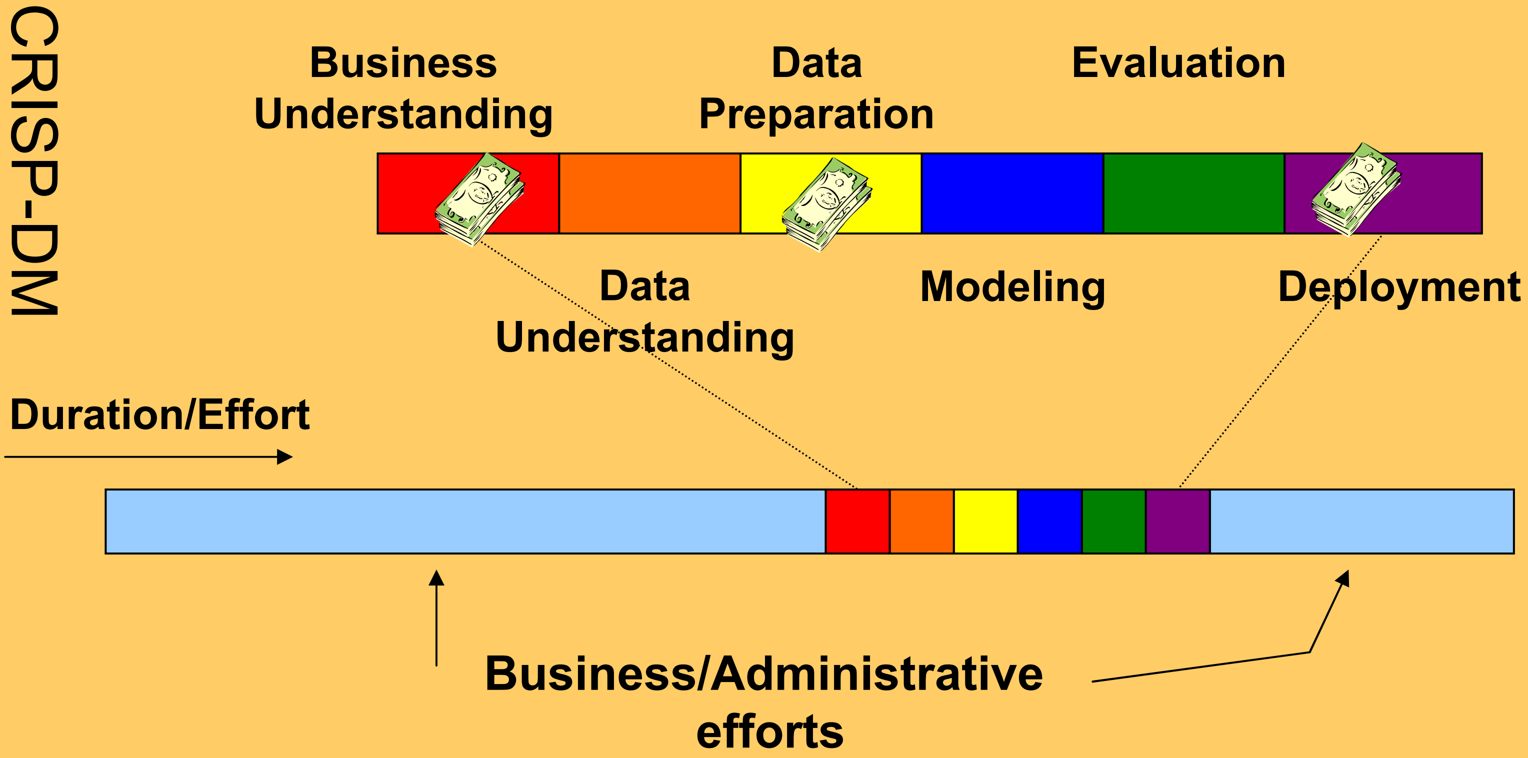
Input: expert knowledge about data and decision alternatives
Goal: construct decision support model – to support the evaluation and choice of best decision alternatives

DM and DS integration



SolEuNet-DM: Extension of the CRISP-DM methodology




CRISP-DM

Business Understanding

Data Preparation

Evaluation

Data Understanding

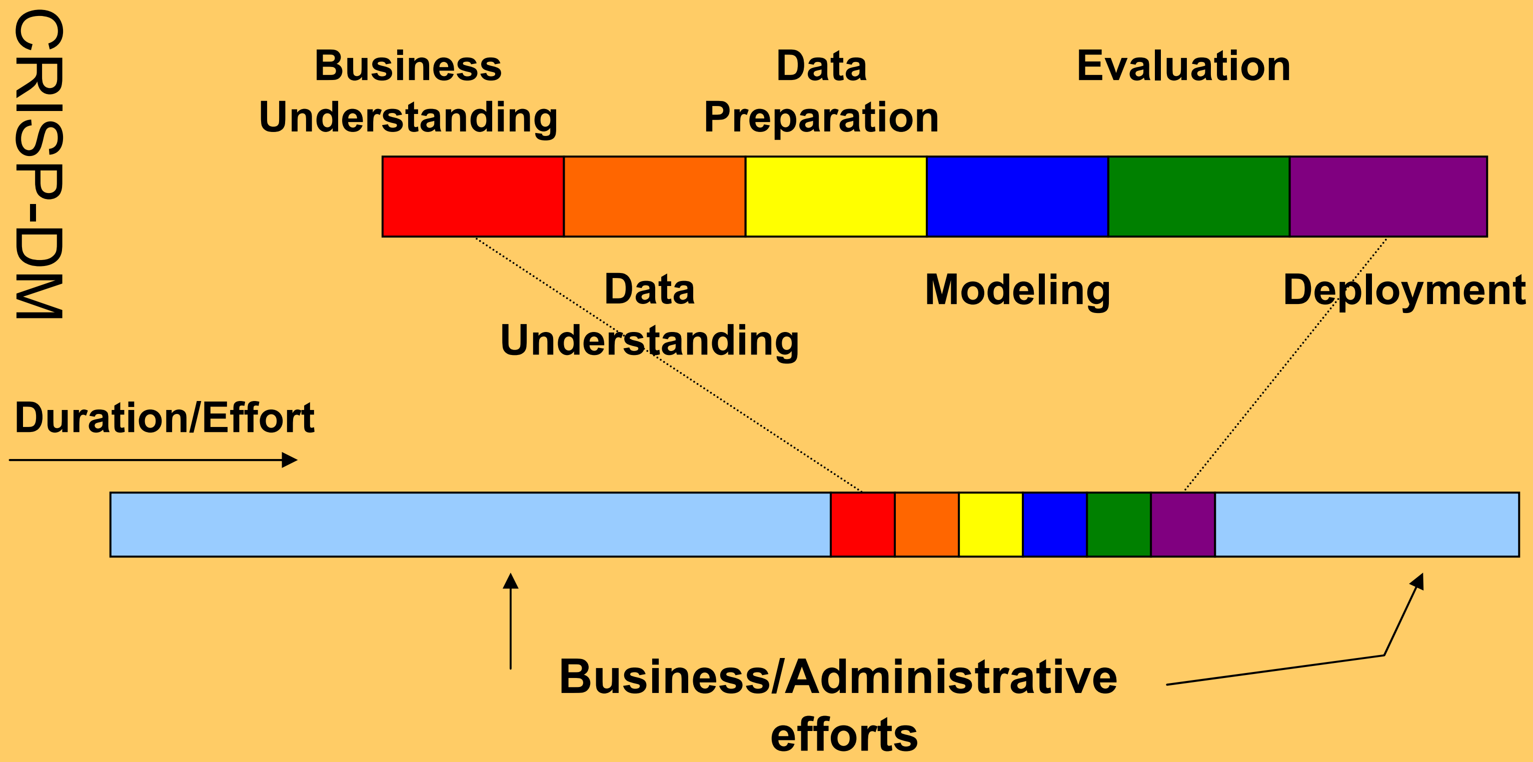
Modeling

Deployment

Duration/Effort

Business/Administrative efforts

SolEuNet-DM in business



- There is MUCH more to a commercial data mining project than simply following CRISP-DM.
- Most of the KDD effort is *NOT* modelling.
- **The VALUE for the client ONLY achieved if results are *DEPLOYED.***

SolEuNet extensions of the CRISP-DM methodology

METHODOLOGY VIEW (CRISP)

- *I. Business Understanding*
- *II. Data Understanding*
- *III. Data Preparation*
- *IV. Modeling*
- *V. Evaluation*
- *VI. Deployment*

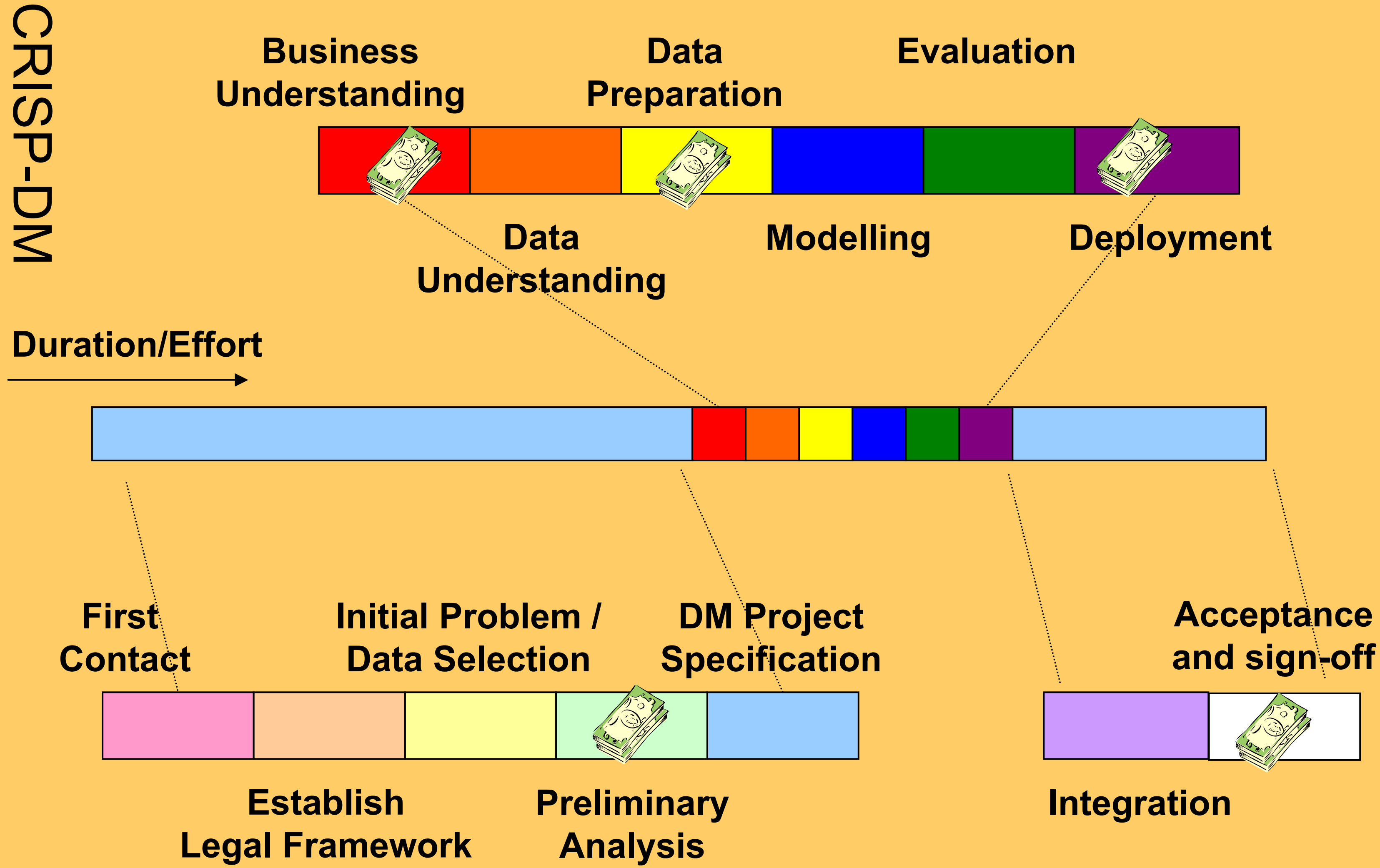
VIRTUAL ORGANISATION VIEW

- *Knowledge Sharing*
- *Collaboration*
- *Trust*
- *Infrastructure*

PROCESS VIEW

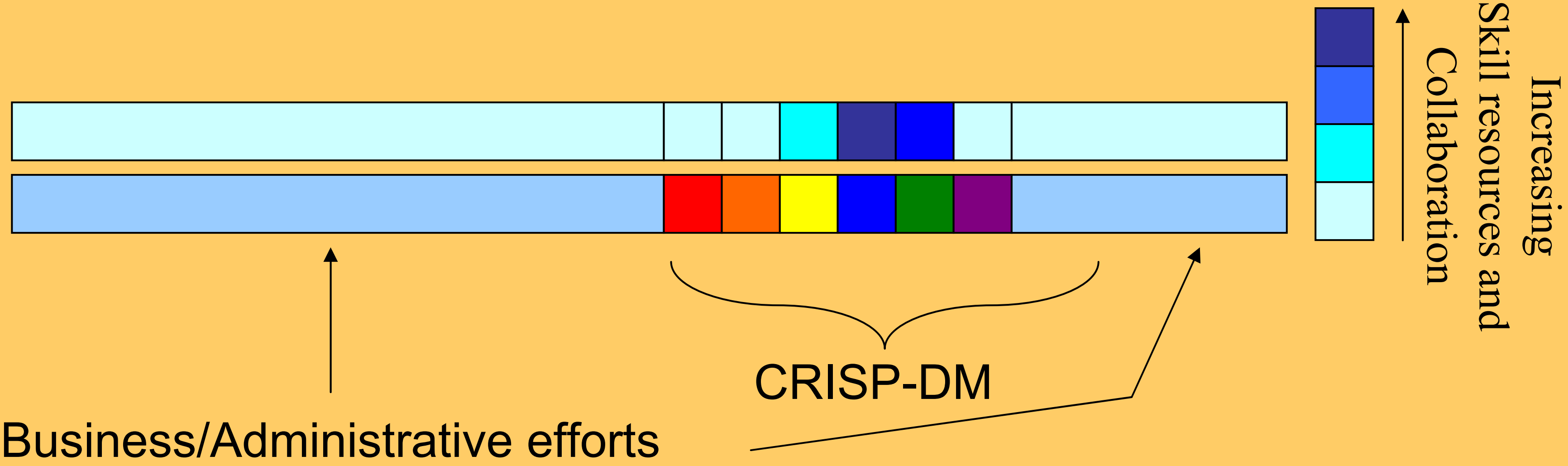
- *A. First Contact*
- *B. Establish the legal framework with the end-user*
- *C. Initial Problem and/or Data Selection*
- *D. Preliminary Analyses*
- *E. Specification of Industrial Pilot Study*
- *F. Negotiate Contractual Terms with the End-user*
- *G. Industrial Project Execution*
- *H. Deployment and further work*

SolEuNet-DM: Extension of the CRISP-DM methodology



Collaboration in SolEuNet-DM

- **What data mining tasks can be executed collaboratively?**
  - The “final stages” of **Data Preparation** assisted by the SumatraTT pre-processing tool
  - **Modeling** (assisted by VizWiz’s model Vizualisation)
  - **Evaluation** (assisted by VizWiz/ROCCON ROC-based model scoring and plotting)


Increasing Skill resources and Collaboration
CRISP-DM
Business/Administrative efforts

Main SolEuNet publications

- **Journal and conference papers**
- **Book: Data Mining and Decision Support: Integration and Collaboration, Kluwer 2003**
  - 4 editors: Mladenić, Lavrač, Bohanec, and Moyle
  - 4 parts, 22 chapters describing main scientific and application results
    - Basic DM and DS techniques
    - DM and DS integration techniques
    - Applications
    - Collaborative DM and lessons learned

Talk outline

- SolEuNet project results
- Subgroup discovery
    - – Motivation: pattern mining applications
        - coronary heart disease patient risk group detection
        - gene expression analysis in functional genomics
        - prediction of mutagenicity of molecules
    - – Approach: subgroup mining techniques
        - SD algorithm (see PAKDD paper)
        - Other algorithms (CN2-SD, APRIORI-SD, RSD)
    - – Focus: the lessons learned

Predictive vs. descriptive induction

- **Predictive induction:** 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 induction:** Discovering interesting regularities in the data, uncovering patterns, ... for solving KDD tasks
  - Symbolic clustering, Association rule learning, Subgroup discovery, ...
  - Exploratory data analysis

Predictive vs. descriptive induction: A rule learning perspective

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

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

Supervised vs. unsupervised learning: A rule learning perspective

- **Supervised learning:** Rules are induced from labeled instances (training examples with class assignment) - usually used in **predictive induction**
- **Unsupervised learning:** Rules are induced from unabeled instances (training examples with no class assignment) - usually used in **descriptive induction**
- **Exception: Subgroup discovery**

  Discovers **individual rules** describing interesting patterns in the data from **labeled** examples

Subgroup discovery

- **Motivation - Medical application**
  - Find and characterize population subgroups with high CHD risk (target class: CHD)

- **Subgroup mining methodology (JAIR 2002)**
  - algorithm SD used for expert-guided subgroup discovery
  - statistical characterization – supporting factors
  - subgroup visualization

Subgroup discovery task

- **Task definition** (Klosgen, 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 task

- **Task definition** (Klosgen, 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)
- **Other (subjective) criteria of interestingness**
  - Surprising to the user, Non-redundant, Simple, Useful - actionable

CHD Risk Group Discovery Task

- **Task:** Find and characterize population subgroups with high CHD risk
- **Input:** Patient records described by **stage A** (anamnestic), stage **B** (an. & lab.), **and stage C** (an., lab. & ECG) attributes
- **Output:** **Best** subgroup descriptions that are **most actionable** for CHD risk screening at primary health-care level

CHD - Anamnestic data (stage A)

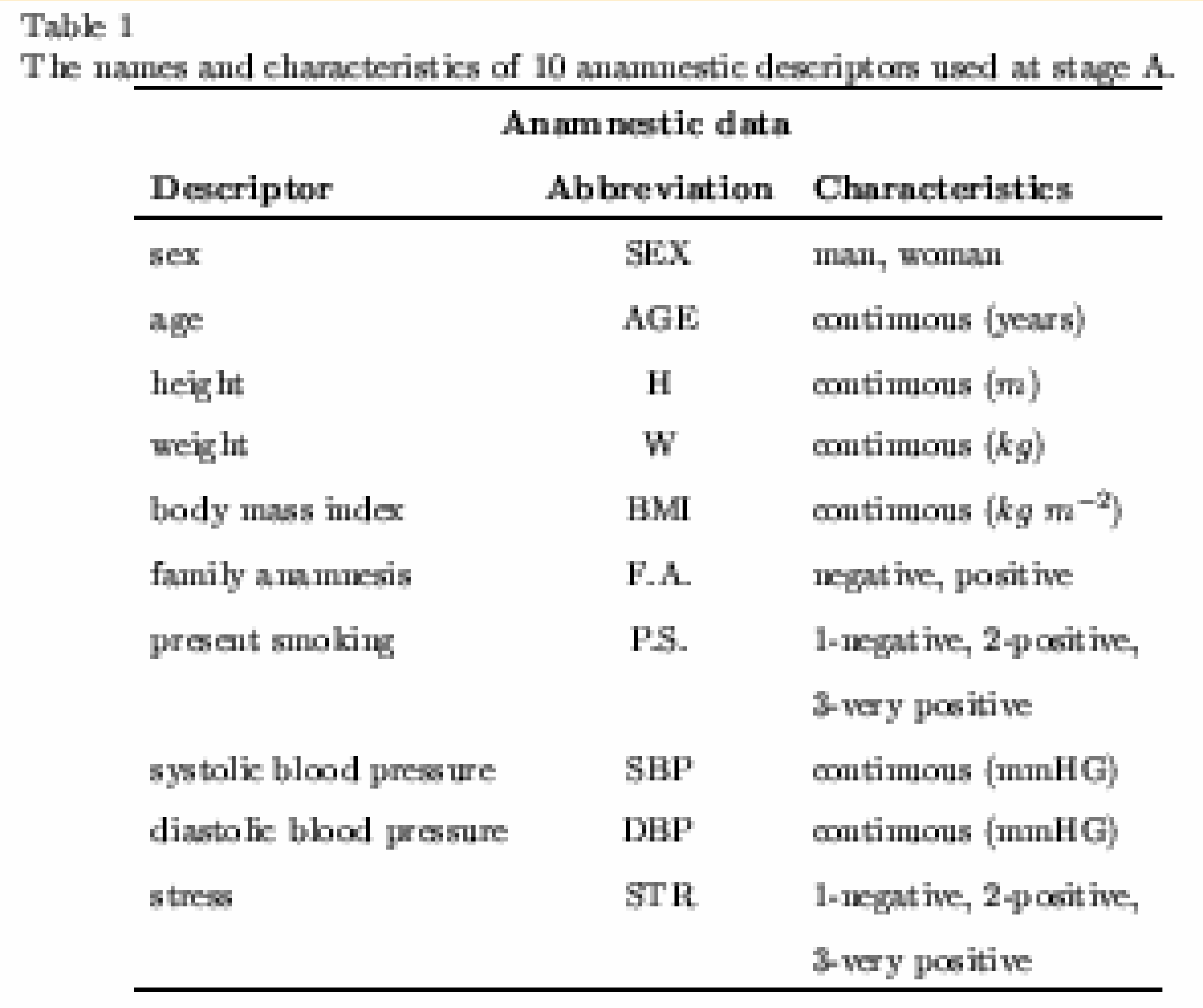


Table 1
The names and characteristics of 10 anamnestic descriptors used at stage A.

Anamnestic data

Descriptor	Abbreviation	Characteristics
sex	SEX	man, woman
age	AGE	continuous (years)
height	H	continuous (m)
weight	W	continuous (kg)
body mass index	BMI	continuous ($kg\ m^{-2}$)
family anamnesis	F.A.	negative, positive
present smoking	P.S.	1-negative, 2-positive, 3-very positive
systolic blood pressure	SBP	continuous (mmHG)
diastolic blood pressure	DBP	continuous (mmHG)
stress	STR	1-negative, 2-positive, 3-very positive

CHD - Laboratory data (stage B)

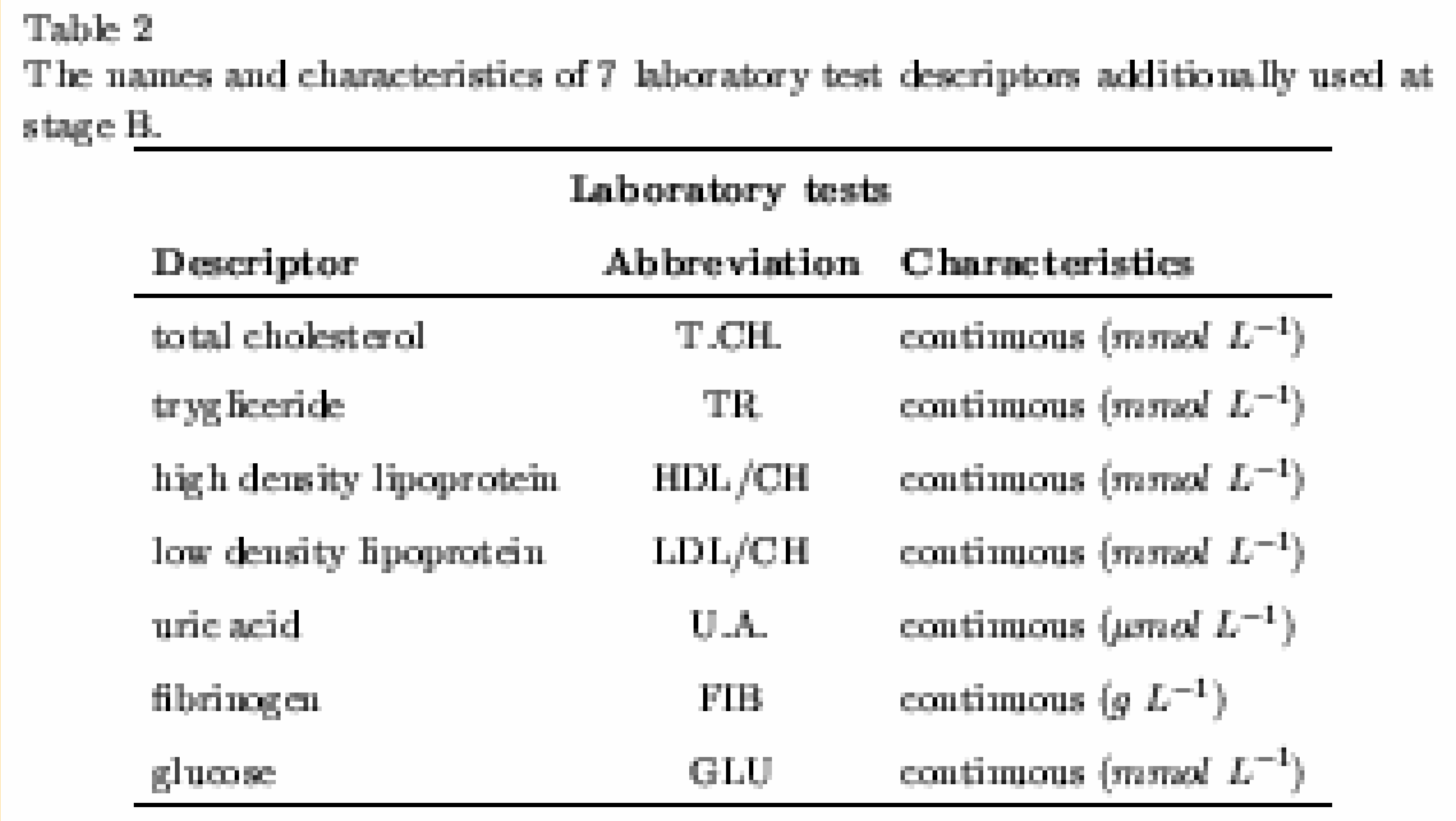


Table 2
The names and characteristics of 7 laboratory test descriptors additionally used at stage B.

Laboratory tests		
Descriptor	Abbreviation	Characteristics
total cholesterol	T.CH.	continuous ($mmol\ L^{-1}$)
trygliceride	TR	continuous ($mmol\ L^{-1}$)
high density lipoprotein	HDL/CH	continuous ($mmol\ L^{-1}$)
low density lipoprotein	LDL/CH	continuous ($mmol\ L^{-1}$)
uric acid	U.A.	continuous ($\mu mol\ L^{-1}$)
fibrinogen	FIB	continuous ($g\ L^{-1}$)
glucose	GLU	continuous ($mmol\ L^{-1}$)

CHD - ECG at rest data (stage C)

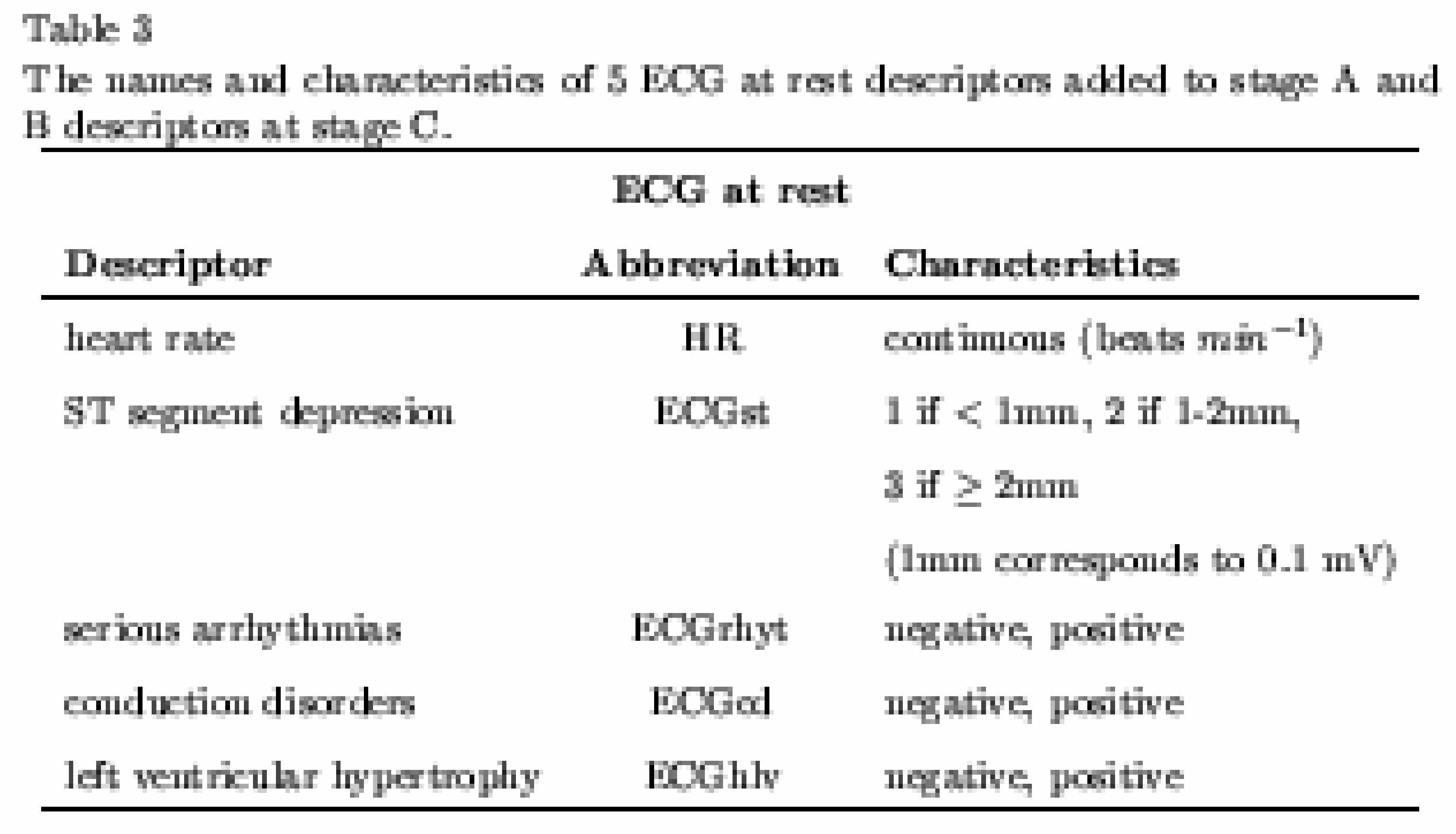


Table 3

The names and characteristics of 5 ECG at rest descriptors added to stage A and B descriptors at stage C.

ECG at rest		
Descriptor	**Abbreviation**	**Characteristics**
heart rate	HR	continuous (beats min^{-1})
ST segment depression	ECGst	1 if < 1mm, 2 if 1-2mm,
		3 if $\geq$ 2mm
		(1mm corresponds to 0.1 mV)
serious arrhythmias	ECGrhyt	negative, positive
conduction disorders	ECGcd	negative, positive
left ventricular hypertrophy	ECGhlv	negative, positive

Results of subgroup discovery in the CHD application

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

A1: CHD ← male & pos. fam. history & age > 46

A2: CHD ← female & bodymassIndex > 25 & age > 63

B1: CHD ← ..., **B2:** CHD ← ..., **C1:** CHD ← ...

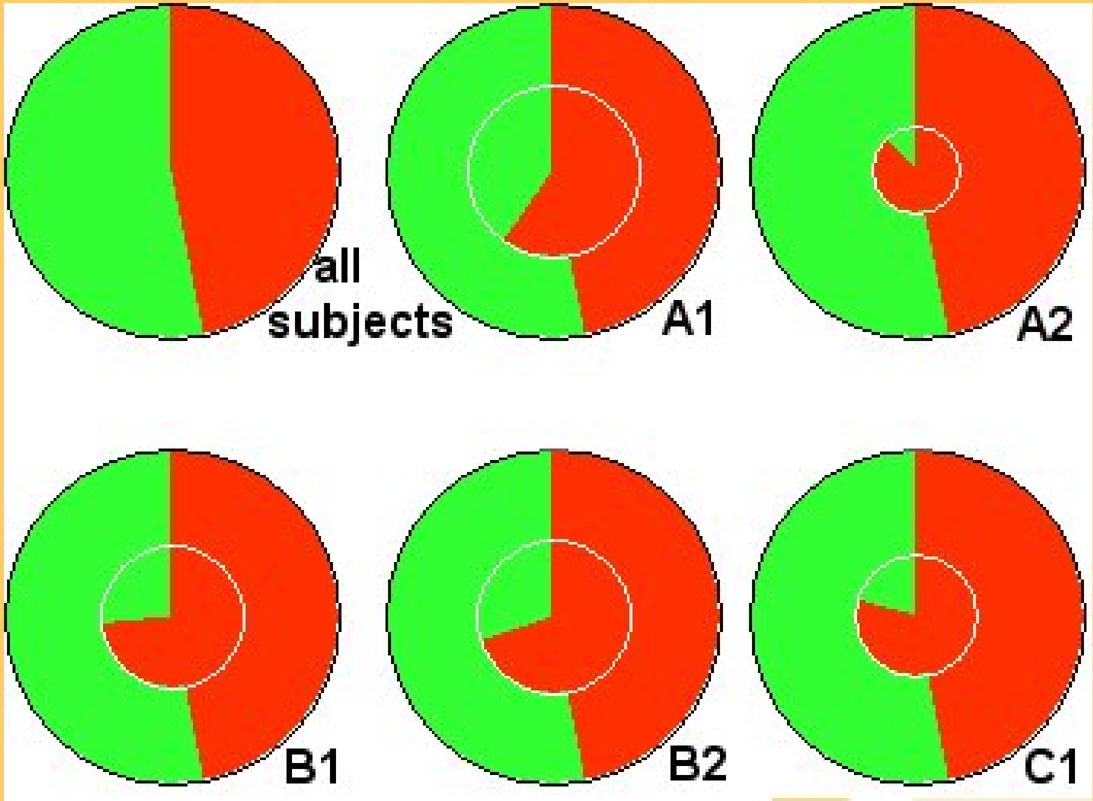
Principal risk factors (found by subgroup mining)

Supporting risk factors (found by statistical analysis):

A1: psychosocial stress, as well as cigarette smoking, hypertension and overweight

A2: …

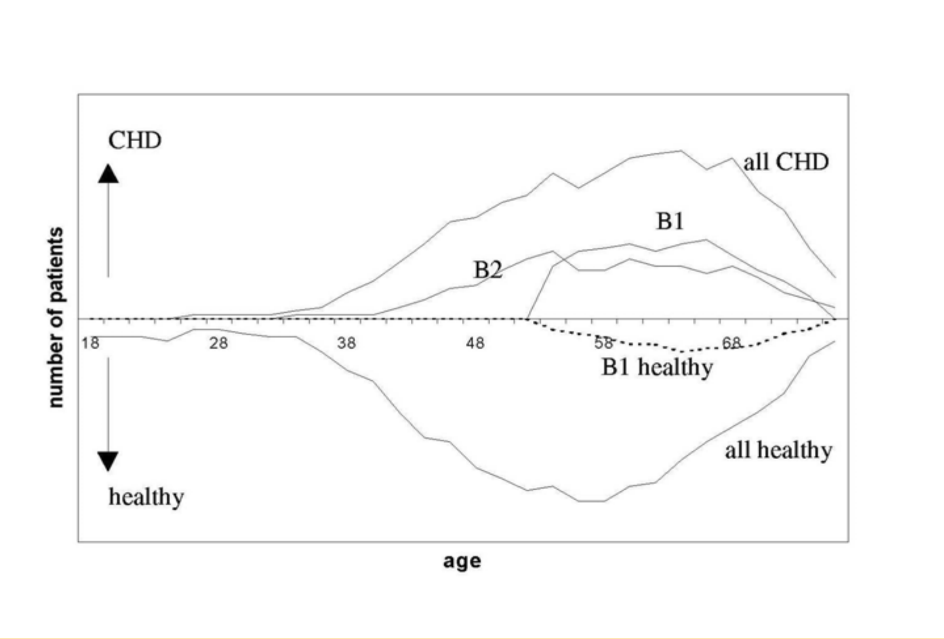
Subgroup visualization



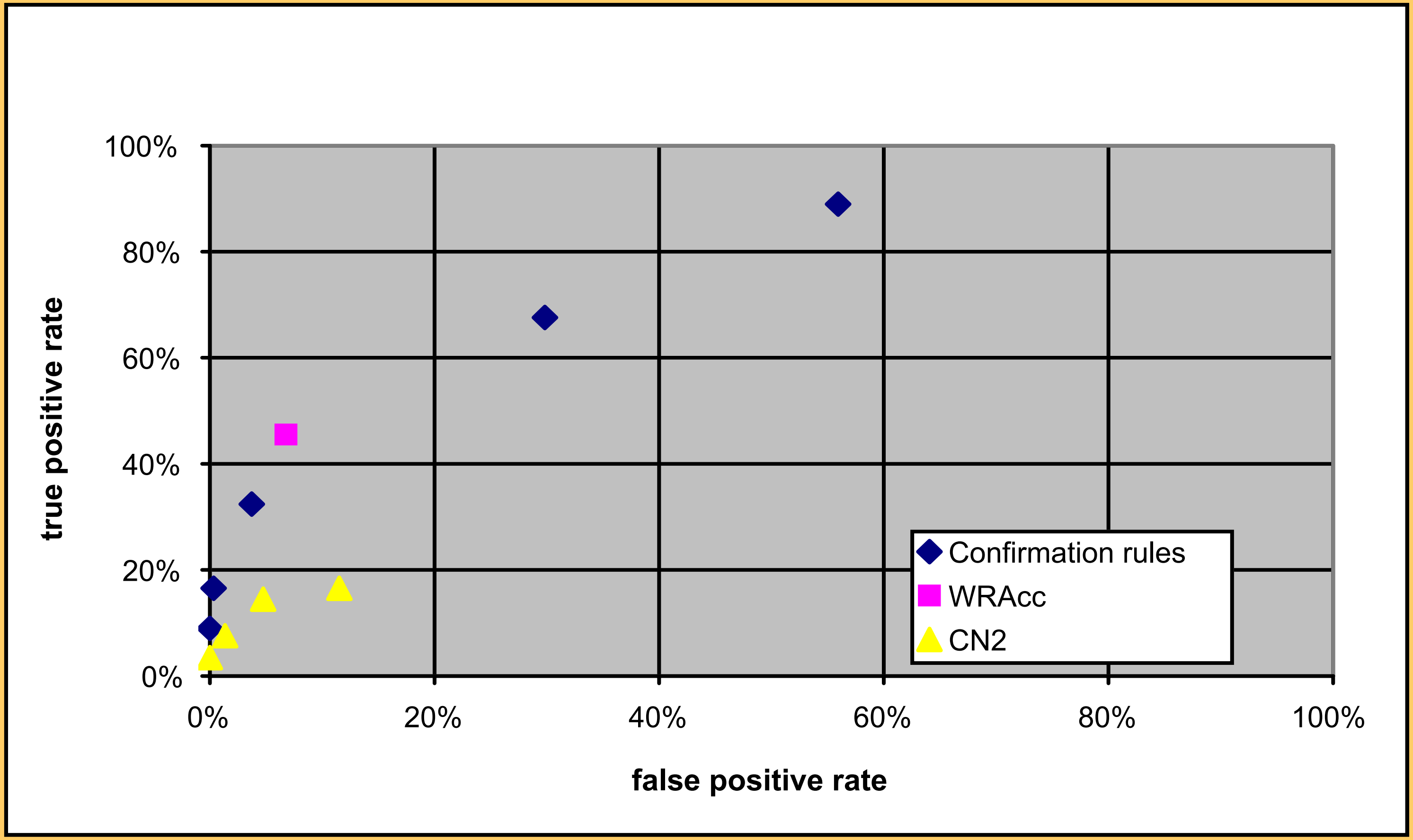
The CHD task: Find and characterize population subgroups with high CHD risk groups (large enough, distributionally unusual, most actionable)

Subgroup visualization

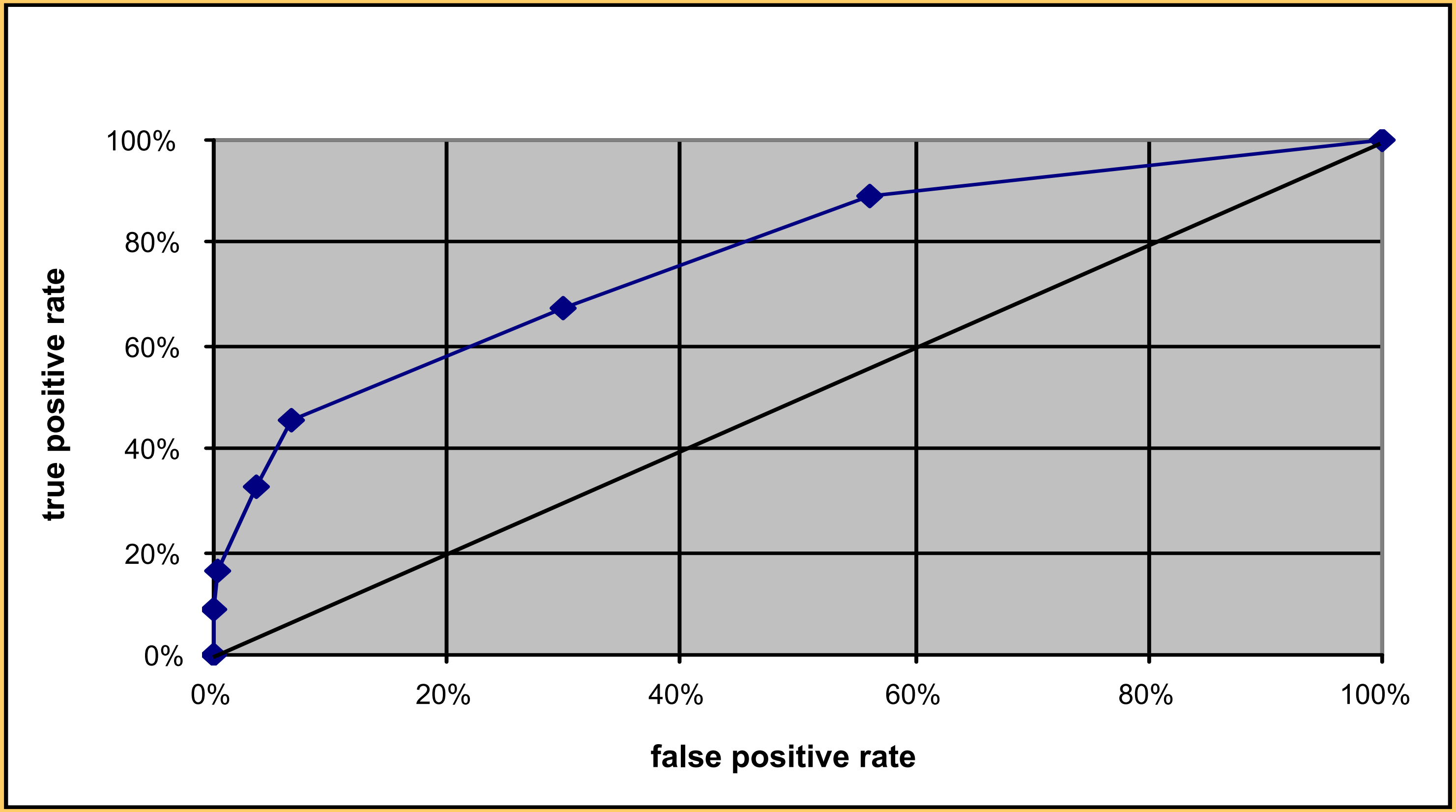
- starts with the subgroup and presents its distribution with respect to a numeric (ordered discrete) attribute of expert's interest
- X axis: selected numeric attribute
- Y+ axis: # of CHD patients
- Y- axis: # of healthy subjects
- Y+ side area under the graph of the subgroup vs. total CHD: sensitivity TPr = TP/Pos = TP/(TP+FN)
- Y- side area under the graph of the subgroup vs. total healthy: false alarm FPr = FP/Neg = FP/(TN+FP)



Rule representation: Points in the ROC space



Best rule subset selection: Points on the ROC convex hull



- Selection/evaluation criterion: optimizing TPr/FPr
- For a single domain: equivalent to optimizing TP/FP

Expert-guided subgroup discovery in the TP/FP space

- The aim of heuristic subgroup discovery is the search for rules with the maximal value of

$$q = TP/(FP+g)$$

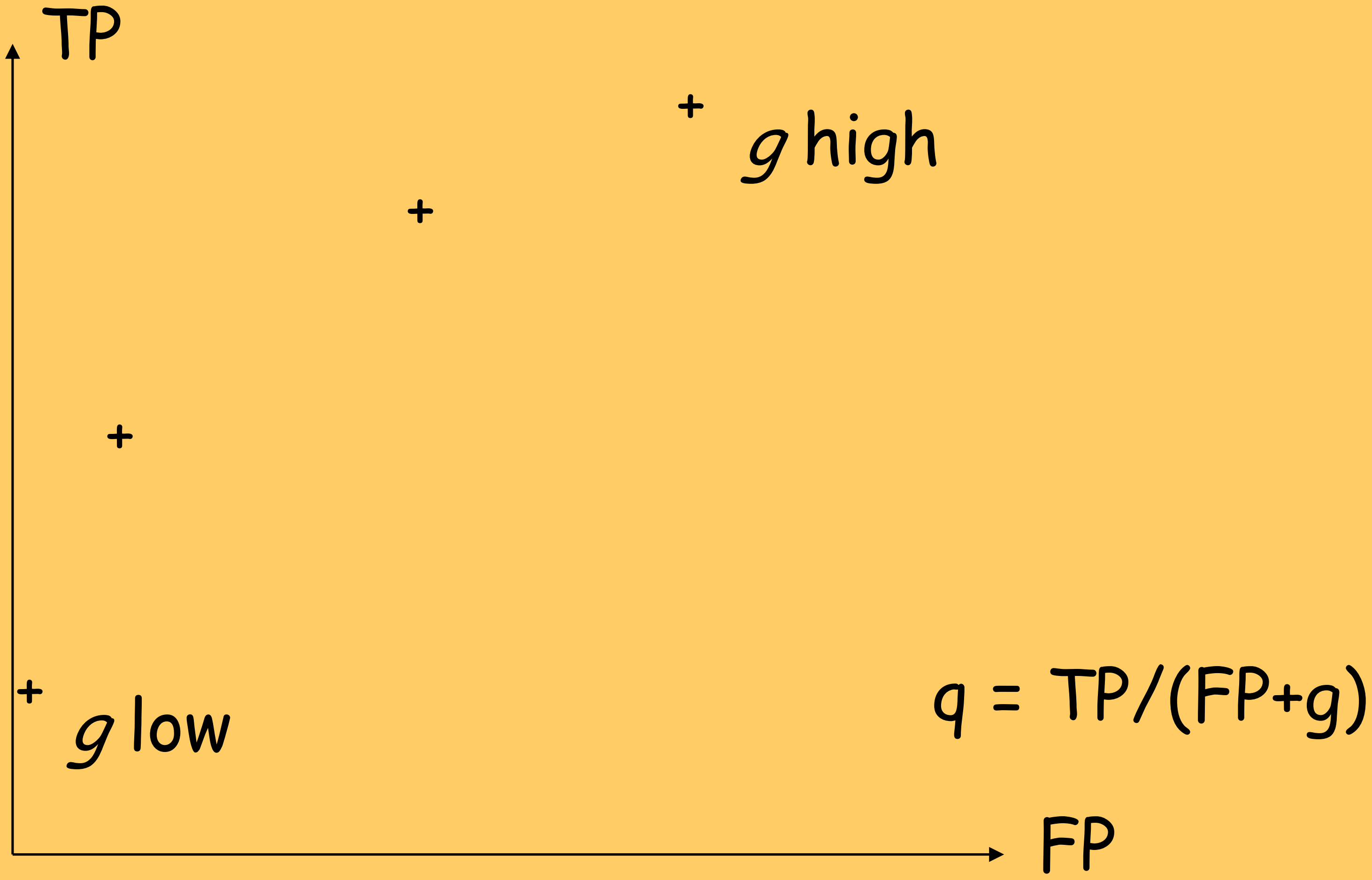
  - TP - true positives
    (CHD patients correctly classified by the rule)
  - FP - false positives
    (healthy subjects incorrectly classified as CHD patients)
  - *g* - generalization parameter

The q measure

$$q = TP/(FP+g)$$

- by searching for rules with high *q* the algorithm tries to find rules covering many target class examples and a low number of non-target class examples
- the tolerated number of covered non-target class examples, relative to the number of covered target class cases, is determined by parameter *g*
  - if *g* value is low (1 or less) then rules cover only few target cases and nearly no non-target class cases -> *high specificity* (low false alarm rate)
  - if *g* value is high (10 or higher) more general rules are generated -> rules with *high sensitivity*

Analysis of the q Measure

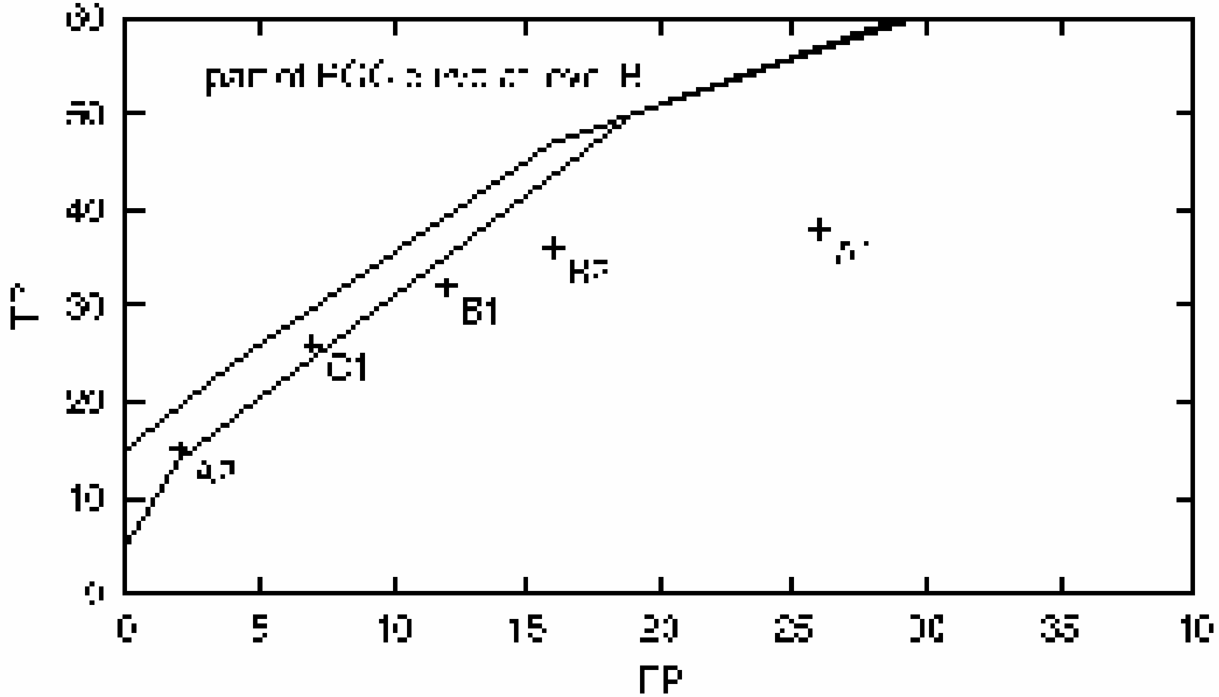


$q = TP/(FP+g)$

Expert-Guided Discovery

- By changing the *g* value (generalization parameter) different models can be induced from the same data set
- Subgroup variation can be achieved also by selecting a subset of features to be used in rule induction
- By combining these techniques many iterations are possible until interesting subgroups have been detected
- NB: the expert may select sub-optimal models !

Expert-Guided Subgroup Discovery in the TP/FP Space



Induced subgroups and their statistical characterization

Subgroup A2 for female patients:

High-CHD-risk **IF**

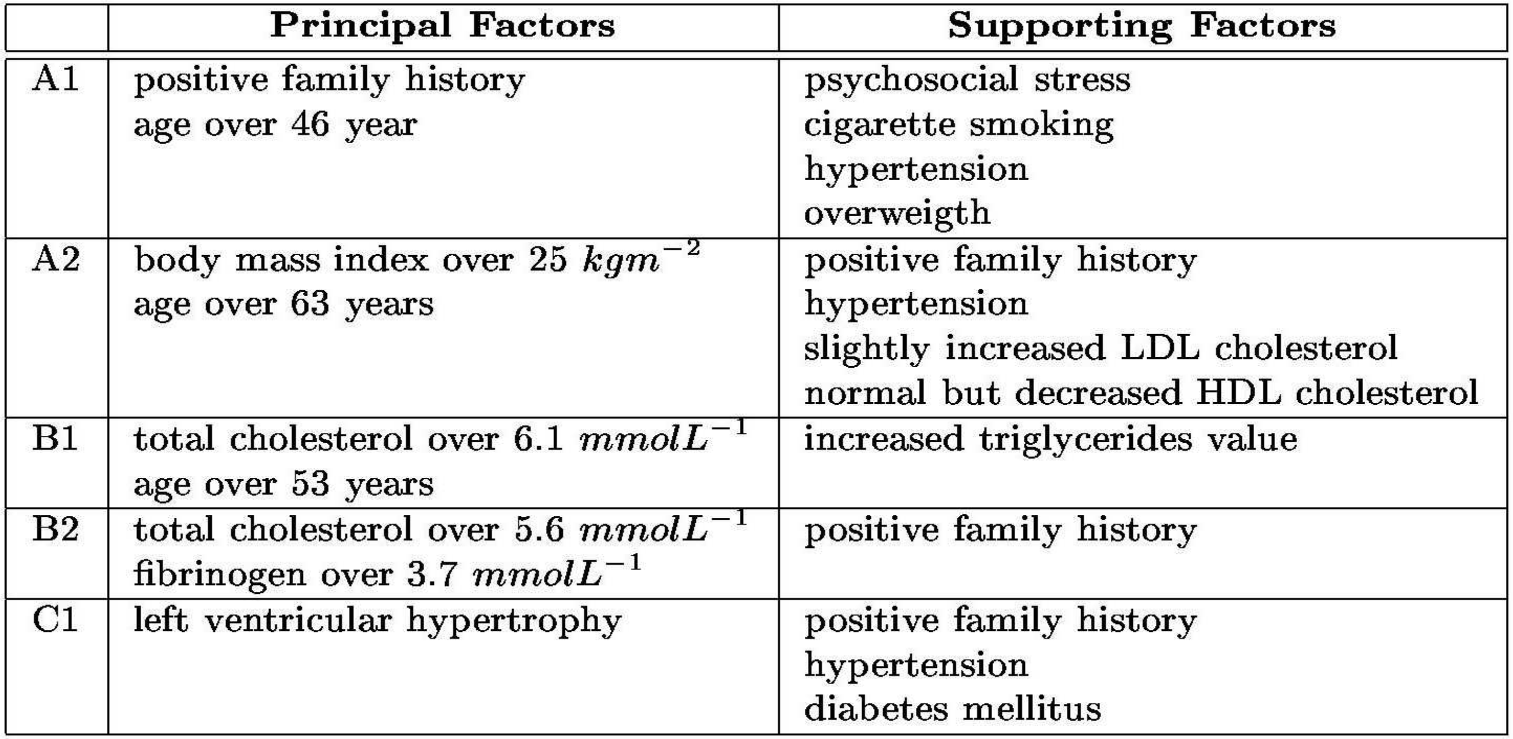
body mass index over 25 kg/m^2 (typically 29)

AND

age over 63 years

Supporting characteristics 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.

Statistical characterization of expert selected subgroups



	Principal Factors	**Supporting Factors**
A1	positive family history age over 46 year	psychosocial stress cigarette smoking hypertension overweigth
A2	body mass index over 25 kgm^{-2} age over 63 years	positive family history hypertension slightly increased LDL cholesterol normal but decreased HDL cholesterol
B1	total cholesterol over 6.1 $mmolL^{-1}$ age over 53 years	increased triglycerides value
B2	total cholesterol over 5.6 $mmolL^{-1}$ fibrinogen over 3.7 $mmolL^{-1}$	positive family history
C1	left ventricular hypertrophy	positive family history hypertension diabetes mellitus

Statistical characterization of subgroups

- starts from induced subgroup descriptions
- statistical significance of all available features (all risk factors) is computed given two populations: true positive cases (CHD patients correctly included into the subgroup) and all negative cases (healthy subjects)
- $\aleph^2$ test with 95% confidence level is used

SD Lessons learned

- In expert-guided subgroup discovery, the expert may decide to choose sub-optimal subgroups, which are the most actionable
- RSS algorithm for rule subset selection, using decreased weights of covered positive examples in rule post-processing, can be used to select a small set of relatively independent patterns
- Additional evidence in the form of supporting factors increases expert’s confidence in rules resulting from automated discovery
- Value-added: Subgroup visualization

Scientific discovery task: DNA microarray data analysis

- Functional genomics: gene expression monitoring by DNA microarrays (“gene chips”) enables:
  - better understanding of many biological processes
  - improved disease diagnosis and prediction in medicine
- Functional genomics is a typical scientific discovery domain characterized by
  - a very large number of attributes (genes) relative to a very small number of examples (observations).
  - typical values: 7000-16000 attributes, 50-150 examples

Machine learning for functional genomics

- Data collected in these applications are not suitable for direct human explanatory analysis
  - as a single DNA micro array experiment results in thousands of measured expression values and
  - due to the lack of existing expert knowledge available for the analysis
- Appropriate for machine learning applications:
  - building predictive models
  - scientific knowledge discovery

Functional genomics domains

- Two-class diagnosis problem of distinguishing between acute lymphoblastic leucemia (ALL, 27 samples) and acute myeloid leukemia (AML, 11 samples), with 34 samples in the test set. Every sample is described with gene expression values for 7129 genes.
- Multi-class cancer diagnosis problem with 14 different cancer types, in total 144 samples in the training set and 54 samples in the test set. Every sample is described with gene expression values for 16063 genes.
- **http://www-genome.wi.mit.edu/cgi-bin/cancer/datasets.cgi.**

Functional genomics domain:
14 class DNA microarray data

	Lung_cancer	Leukemia	Lymphoma	CNS
Profilin mRNA	1012 P	104 P	90 M	25 A
Neurofibromin 2	188 A	39 A	98 M	187 P

signal intensity

continuous values

presence call (signal specificity)

A - absent, P - present, M - marginal

(Affymetrix GENECHIP SW)

SD Lessons learned

- In domains with only 8 training instances overfitting could not be avoided, in domains with 16 and 24 instances good prediction rules were induced.
- Domain specific restrictions in the form of using A,M,P values instead of continuous attributes are useful for avoiding overfitting and because of the ease of their human interpretation.
- Our relevancy filtering approach (using total, absolute and relative relevancy filtering) successfully filtered random features
  - Reducing 16063 features to 4445 relevant features
- Details available (ECAI 2004 and JBI 2004 papers)

Subgroup discovery: our recent related work

- **SD algorithm - JAIR 2002, JBI 2004 (with Gamberger)**
  - Weighted covering algorithm, TP/(FP+g) heuristic
  - Successful (bio)medical applications :CHD and functional genomics, evaluated by experts
- **CN2-SD algorithm - JMLR 2004 (with Kavsek, Flach, Todorovski)**
  - Weighted covering algorithm, WRAcc heuristic, probabilistic classification, ROC evaluation
  - Evaluation on 17 UCI datasets shows improved coverage, significance, AUC, and comparable accuracy w.r.t. CN2
- **APRIORI-SD – IDA 2003, AAI 2005 in press (with Kavsek)**
- **Algorithm RSD – ILP 2002 (with Zelezny)**

CN2-SD: Adapting CN2 to Subgroup Discovery

- Weighted covering algorithm
- Weighted relative accuracy (WRAcc) search heuristics, with added example weights

$$\text{WRAcc}(\text{Cl} \leftarrow \text{Cond}) =$$

$$p(\text{Cond})\ (p(\text{Cl} \mid \text{Cond}) - p(\text{Class}))$$

- Probabilistic classification
- Evaluation with different rule interestingness measures

CN2-SD Lessons learned

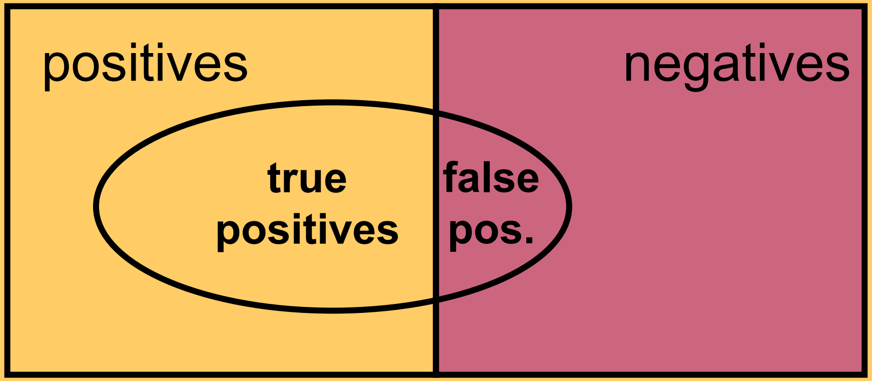
- Can classification rule learning be used for subgroup discovery ?
  - Pro:
    - labeled data available
      (+ patients with CHD, - healthy individuals)
  - Againts:
    - subgroup descriptions should be independent `chunks' of knowledge with high coverage - can not be induced using the covering algorithm used in ruleset induction
    - distribution of + and - can be highly unbalanced

CN2-SD Lessons learned

- Only first few rules induced by the covering algorithm have sufficient support (coverage)
- Subsequent rules are induced from smaller and strongly biased example subsets (pos. examples not covered by previously induced rules), which hinders their ability to detect population subgroups
- ‘Ordered’ rules are induced and interpreted sequentially as a **if-then-else** decision list

CN2-SD Lessons learned

- **Classifiers – models (sets of rules)**
  - Classification rules aim at pure subgroups
  - A set of rules forms a domain model
- **Subgroup descriptions – patterns (individual rules)**
  - Rules describing subgroups aim at significantly higher proportion of positives
  - Each rule (pattern) is an independent chunk of knowledge
- **Link**
  - Subgroup discovery can be viewed as cost-sensitive rule learning, rewarding TP covered, punishing FP covered



CN2-SD Lessons learned: ROC for best rule selection

	Predicted positive	Predicted negative	
Positive examples	**True positives**	**False negatives**	
Negative examples	**False positives**	**True negatives**	

Suppose we have two rules (classifiers) with same classification accuracy (80%).

Classifier 1

	Predicted positive	Predicted negative	
Positive examples	**40**	**10**	50
Negative examples	**10**	**40**	50
	50	50	100

Classifier 2

	Predicted positive	Predicted negative	
Positive examples	**30**	**20**	50
Negative examples	**0**	**50**	50
	30	70	100

Lessons learned:
ROC for best rule selection

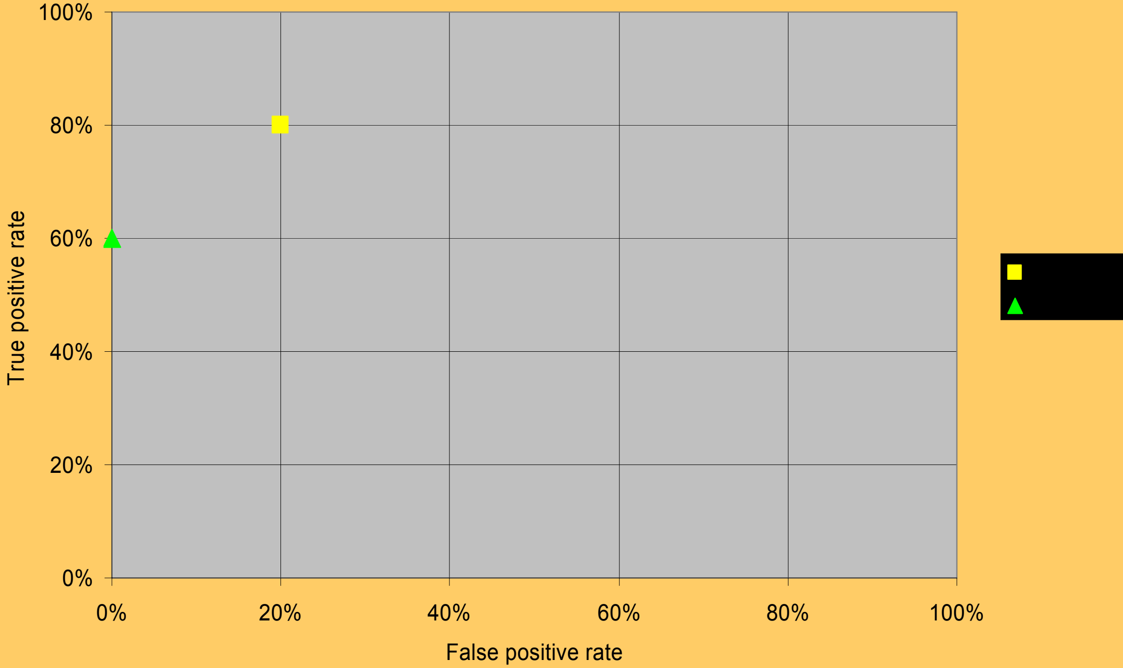
- ***True positive rate*** #true pos. / #pos.
  - $TPr_1 = 40/50 = 80\%$
  - $TPr_2 = 30/50 = 60\%$
- ***False positive rate*** #false pos. / #neg.
  - $FPr_1 = 10/50 = 20\%$
  - $FPr_2 = 0/50 = 0\%$
- ***ROC space*** has
  - FPr on X axis
  - TPr on Y axis

Classifier 1

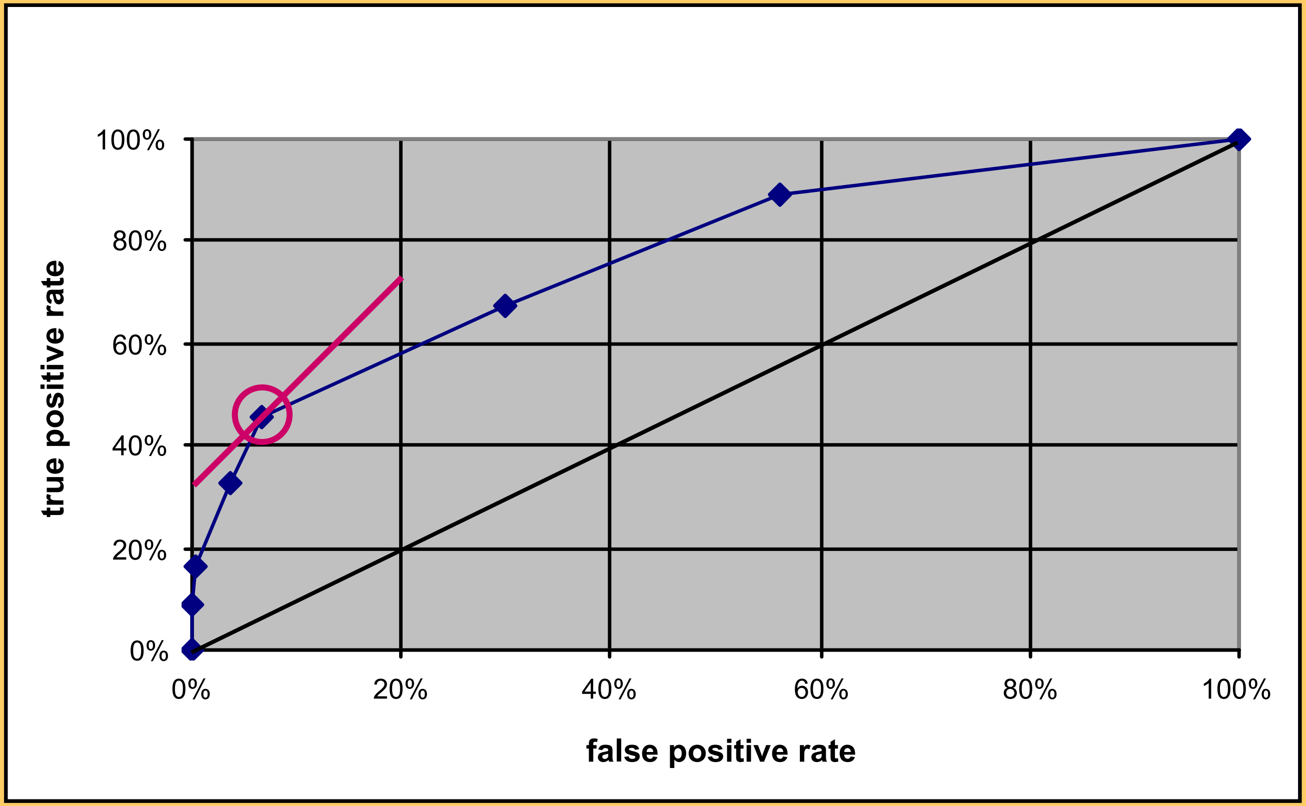
	Predicted positive	Predicted negative	
Positive examples	**40**	**10**	50
Negative examples	**10**	**40**	50
	50	50	100

Classifier 2

	Predicted positive	Predicted negative	
Positive examples	**30**	**20**	50
Negative examples	**0**	**50**	50
	30	70	100



Lessons learned: ROC for best rule selection

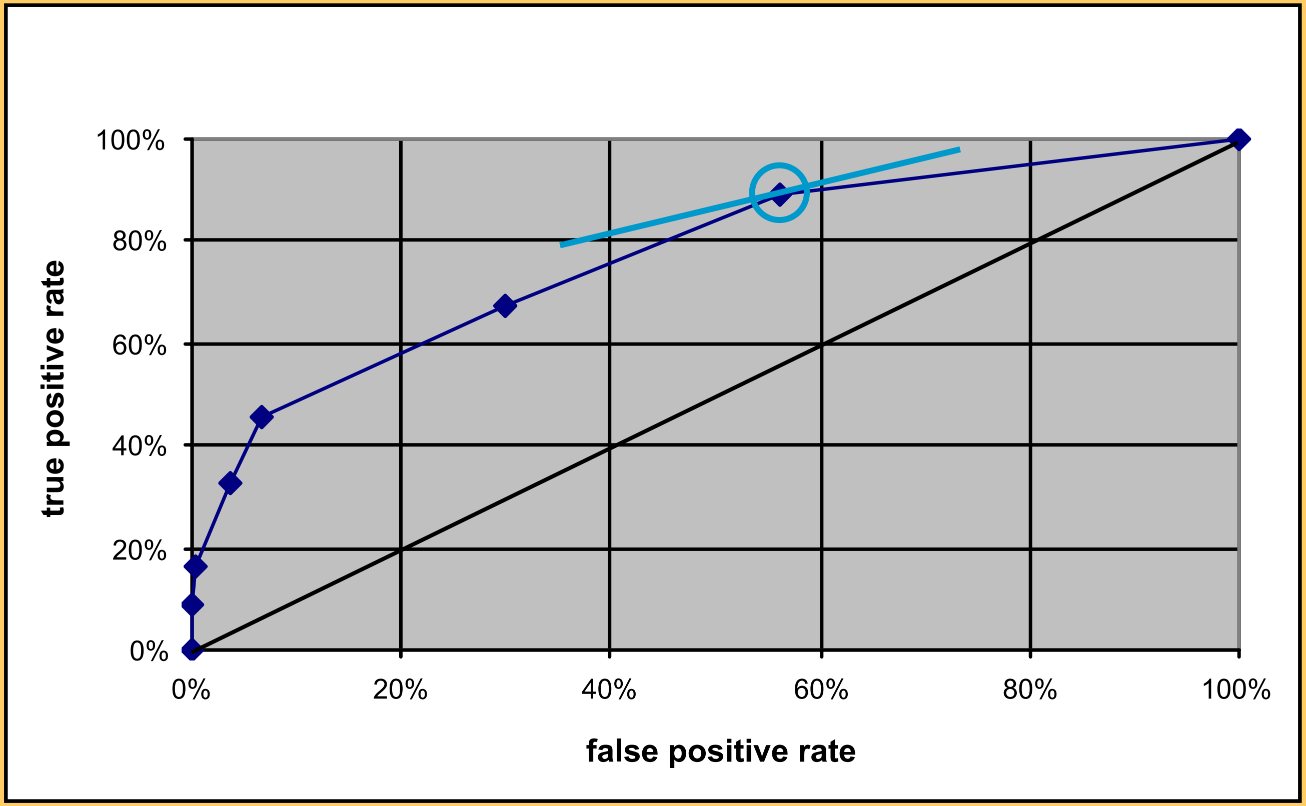


$$\frac{FPcost}{FNcost} = 1/2$$

$$\frac{Neg}{Pos} = 4$$

$$slope = 4/2 = 2$$

Lessons learned: ROC for best rule selection

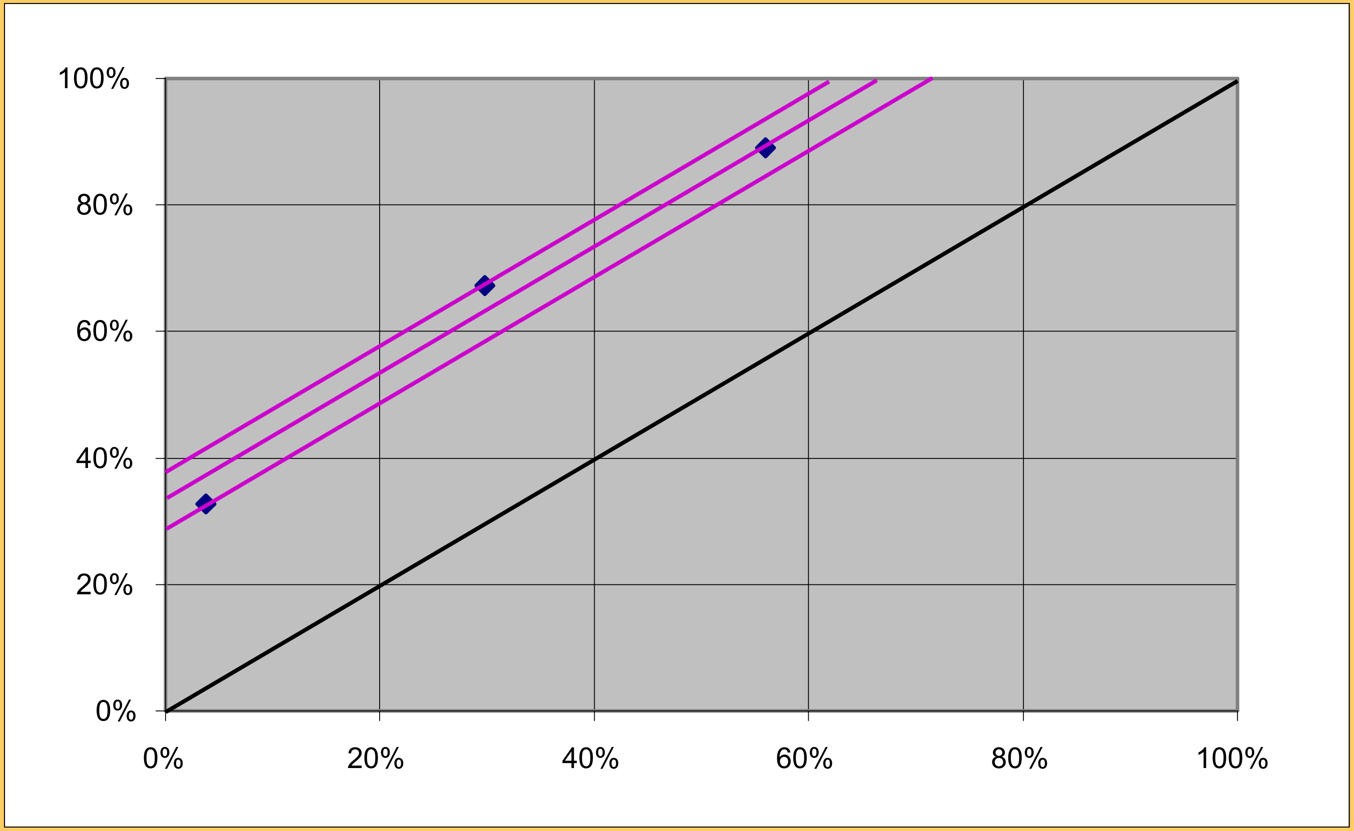


$$\frac{FPcost}{FNcost} = 1/8$$

$$\frac{Neg}{Pos} = 4$$

$$slope = 4/8 = .5$$

Weighted Relative Accuracy



WRAcc = WRAcc(Cl←Cond) = p(Cond) x [p(Cl | Cond) – p(Cl)] =

= coverage x (precision – default precision)

= TP/N – ((TP+FN)/N) x ((TP+FP)/N)

= Pos x Neg x [TPr – FPr]

CN2-SD Lessons learned

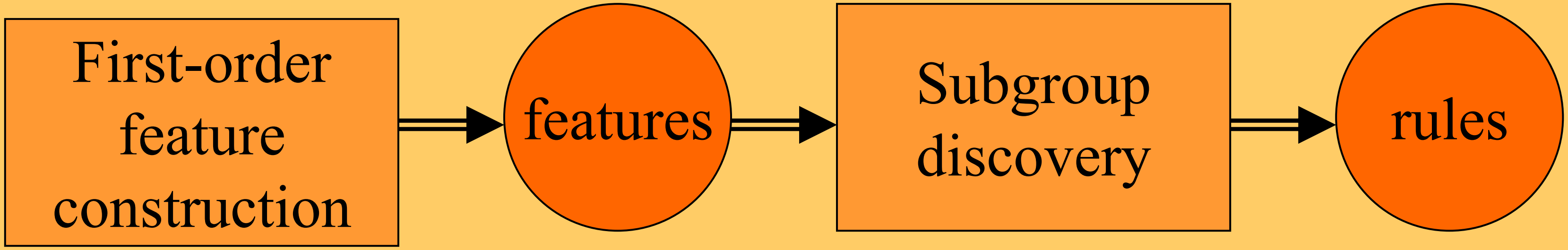
- WRAcc has a clear interpretation in the ROC space
- WRAcc measures rule quality in terms of coverage and accuracy gain
- WRAcc is proportional to significance
- WRAcc is one of the best measures of success in descriptive rule induction, used to measure rule unusualness

APRIORI-SD: Adapting APRIORI to Subgroup Discovery

- Adapting APRIORI to classification rule learning (APRIORI-C)
- Weigthed covering algorithm adapted to rule post-processing
- Best-rule selection using Weighted relative accuracy (WRAcc) heuristics, and based on an optimization approach in terms coverage of the space of training examples
- Analysis of WRAcc in ROC space

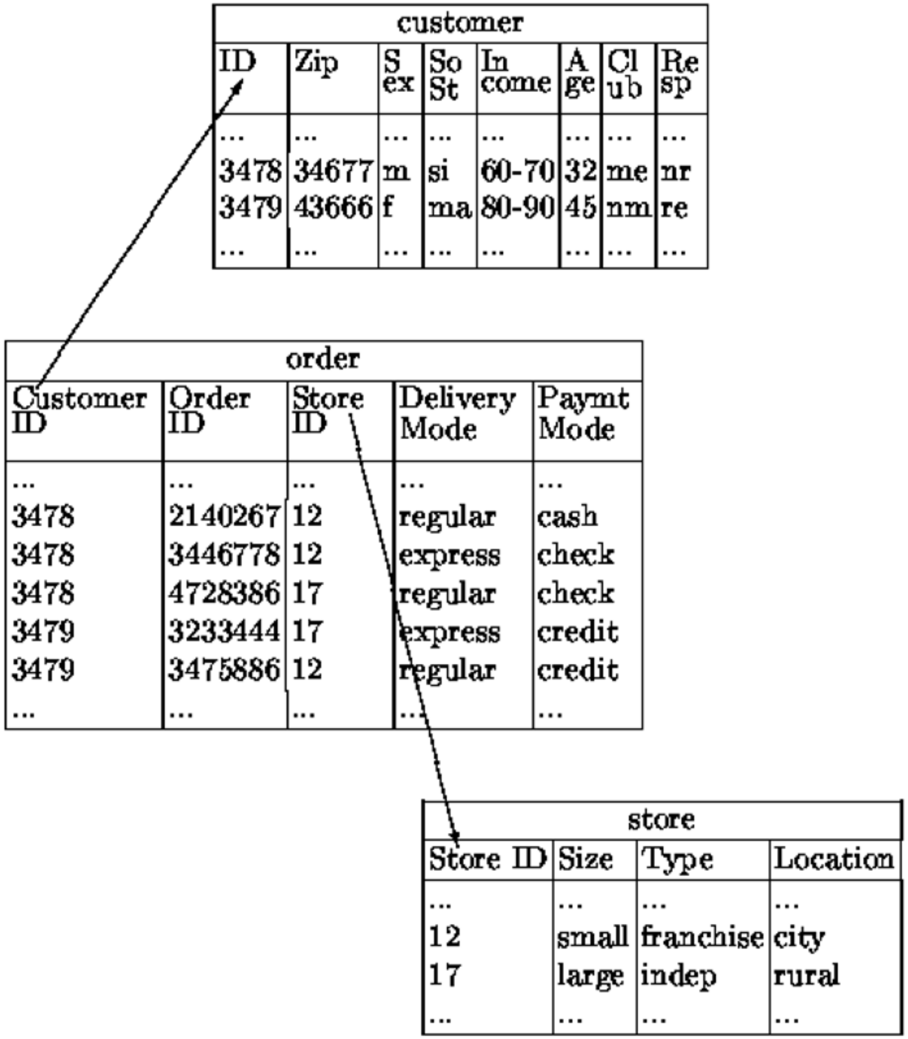
RSD: Upgrading CN2-SD to Relational Subgroup Discovery

- Implementing an propositionalization approach to relational data mining, through efficient first-order feature construction
- Using CN2-SD for propositional subgroup discovery



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, ...



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
...	...	...	...

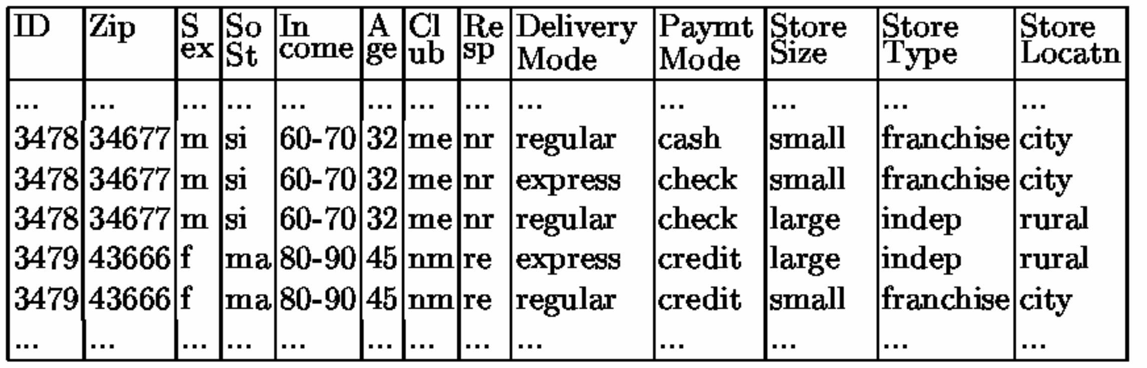
Relational representation of customers, orders and stores.

Complexity of DM problems

- **Propositional** representations: single table with a primary key
  - Data instance is ***fix-sized tuple of constants***
  - Features are those given in the dataset
- **Multiple-instance** representation: Single table without a primary key
  - Data instance corresponds to a `bag' of tuples of constants
- **First-order** representation: multiple tables
  - Data instance is a ***flexible-sized structured object***
    - sequence, set, graph
    - hierarchical: e.g. set of sequences
  - Features need to be **selected** from a potentially infinite set

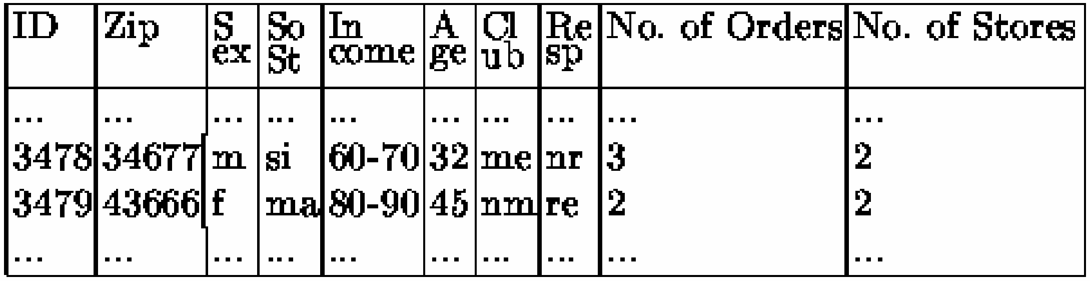
Transforming multi-relational data into a single table

- Creating a single table by joining multiple tables:
  - no primary key
  - multi-instance learning problem
- Making data propositional by aggregates:
  - loss of information



ID	Zip	Sex	SoSt	Income	Age	Club	Resp	Delivery Mode	Paymt Mode	Store Size	Store Type	Store Locatn
...	...	...	...	...	...	...	...	...	...	...	...	...
3478	34677	m	si	60-70	32	me	nr	regular	cash	small	franchise	city
3478	34677	m	si	60-70	32	me	nr	express	check	small	franchise	city
3478	34677	m	si	60-70	32	me	nr	regular	check	large	indep	rural
3479	43666	f	ma	80-90	45	nm	re	express	credit	large	indep	rural
3479	43666	f	ma	80-90	45	nm	re	regular	credit	small	franchise	city
...	...	...	...	...	...	...	...	...	...	...	...	...

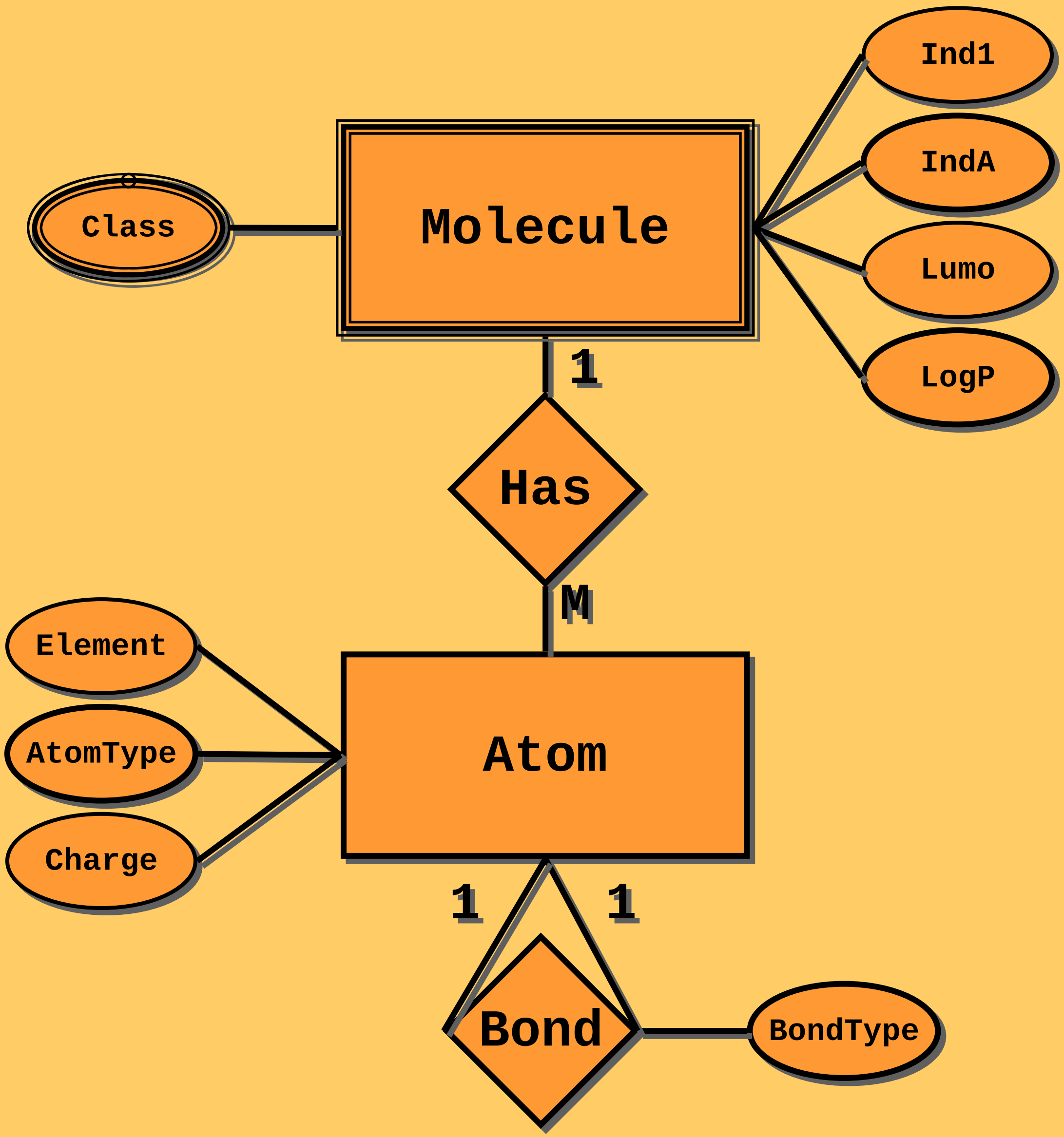
Customer table with multiple orders.



ID	Zip	Sex	SoSt	Income	Age	Club	Resp	No. of Orders	No. of Stores
...	...	...	...	...	...	...	...	...	...
3478	34677	m	si	60-70	32	me	nr	3	2
3479	43666	f	ma	80-90	45	nm	re	2	2
...	...	...	...	...	...	...	...	...	...

Customer table using summary attributes.

Mutagenicity prediction problem



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)
```

Thanks

- To Dragan Gamberger and SolEuNet researchers for joint work in subgroup discovery techniques and applications presented in this talk
- To conference organizers for a high-quality scientific conference and perfect social program
- To Prof. Hiroshi Motoda and Prof. Ho Tu Bao for their invitation to PAKDD-05
- To you for your presence at this last day invited talk