GUHA Method and the LISp-Miner System

Jan Rauch
Milan Šimůnek

Department of Information and Knowledge Engineering,
University of Economics, Prague

University of Economics, Prague

6 Faculties

- Finance and Accounting
- International Relations
- Business Administration
- Informatics and Statistics
- Economics and Public Administration
- Management

17 000 students

http://www.vse.cz/index-en.php?lang=en

Department of Information and Knowledge Engineering

- Faculty of Informatics and Statistics
- 9 Members
- Bachelor courses of informatics (400 students / year)
- Magister: Information and Knowledge Engineering
- Close cooperation with LISp - Laboratory of Intelligent Systems
- KEG - Knowledge Engineering Group; see http://keg.vse.cz/

http://kizi.vse.cz/KIZI/WCMS_KIZI.nsf/pages/AboutDepartment.html

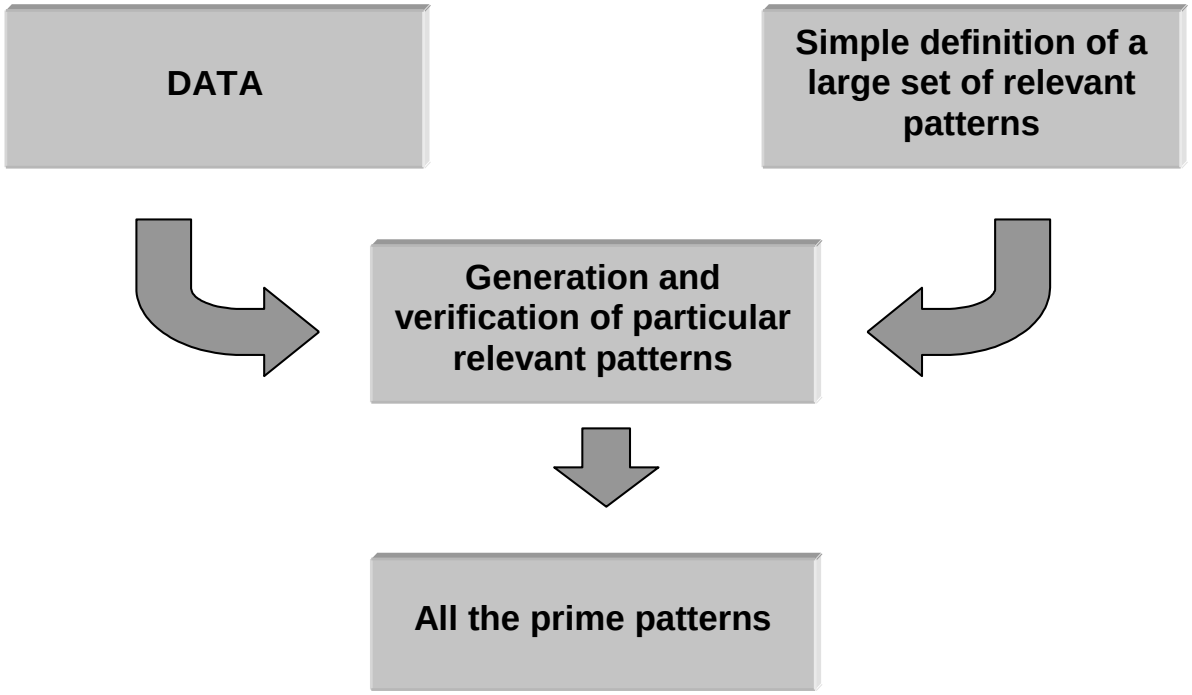
GUHA Method and the LISp-Miner System

- GUHA method and KDD
- LISp-Miner system
- Current research projects
- Observational calculus – logical calculus of KDD patterns

GUHA Method and KDD

- Principle: To offer all interesting patterns true in given data
  - Since 1966, Prague
  - Method of exploratory data analysis
  - P. Hájek, I. Havel, M. Chytil, T. Havránek, …
  - Logical and statistical theory
  - End of 1970s attempt to apply GUHA to data bases
- Today developed as method of KDD
- Implemented by GUHA procedures

GUHA Procedure



- Procedure ASSOC – mines for generalized association rules
- 6 additional procedures in LISp-Miner system,
- additional implementations, …

GUHA – selected implementations (1)

- 1966 - MINSK 22 (I. Havel)
  Boolean data matrix
  simplified version
  association rules
  punch tape

- end of 1960s - IBM 7040 (I. Havel)

- 1976 IBM 370 (I. Havel, J. Rauch)
  Boolean data matrix
  association rules
  statistical quantifiers
  bit strings
  punch cards

GUHA – selected implementations (2)

- Early 1990s – PC-GUHA
  MS DOS
  A. Sochorová, P. Hájek, J. Rauch

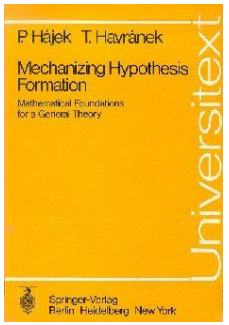
- Since 1995 GUHA+-
  Windows
  D. Coufal + all.

- Since 1996 LISp-Miner
  Windows
  J. Rauch + M. Šimůnek + all.
  6 GUHA procedures
  additional procedures
  … related research

LISp-Miner

- Since 2006 Ferda, M. Ralbovský + all

GUHA – selected publications



- Hájek P., Havel I., Chytil M.: The GUHA method of automatic hypotheses determination, Computing 1 (1966) 293-308.
- Hájek P., Havránek T.: Mechanizing Hypothesis Formation (Mathematical Foundations for a General Theory), Springer-Verlag 1978, 396 pp. see http://www.cs.cas.cz/~hajek/guhabook/

- Hájek P. (guest editor): International Journal of man-Machine Studies, vol. 10, No 1 (special issue on GUHA), 1978.
- Hájek P. (guest editor): International Journal for Man-Machine Studies, vol. 15, No 3 (second special issue on GUHA) 1981.
- Hájek P., Sochorová A., Zvárová J.: GUHA for personal computers, Comp. Stat. and Data Anal. 19 (1995), pp. 149-153.

- Rauch, J. - Šimůnek, M.: An Alternative Approach to Mining Association Rules. In: Lin, T. Y. et all. (ed.): *Foundations of Data Mining and Knowledge Discovery*. Berlin: Springer, 2005, pp. 211 – 232
- Rauch, Logic of Association Rules. Applied Intelligence, 22, 2005
- Rauch, J. - Šimůnek, M.: GUHA Method and Granular Computing. In: Hu, X et al. (Eds.) Proceedings of IEEE conference Granular Computing. Beijing, IEEE Computer Society

GUHA and association rules

http://en.wikipedia.org/wiki/Association_rule_learning#cite_note-

Association rule learning

From Wikipedia, the free encyclopedia

In data mining, **association rule learning** is a popular and well researched method for discovering interesting relations between variables in large databases. Piatetsky-Shapiro [1] describes analyzing and presenting strong rules discovered in databases using different measures of interestingness. Based on the concept of strong rules, Agrawal et al. [2] introduced association rules for discovering regularities between products in large scale transaction data recorded by point-of-sale (POS) systems in supermarkets. For example, the rule $\{onions, potatoes\} \Rightarrow \{beef\}$ found in the sales data of a supermarket would indicate that if a customer buys onions and potatoes together, he or she is likely to also buy beef. Such information can be used as the basis for decisions about marketing activities such as, e.g., promotional pricing or product placements. In addition to the above example from market basket analysis association rules are employed today in many application areas including Web usage mining, intrusion detection and bioinformatics.

Contents [hide]

History [edit]

The concept of association rules was popularised particularly due to the 1993 article of Agrawal [2], which has acquired more than 6000 citations according to Google Scholar, as of March 2008, and is thus one of the most cited papers in the Data Mining field. However, it is possible that what is now called "association rules" is simliar to what appears in the 1966 paper [7] on GUHA, a general data mining method developed by Petr Hájek et al. [8].

History:
The concept of association rules was popularised particularly due to the 1993 article of Agrawal [2], which has acquired more than 6000 citations according to Google Scholar, as of March 2008, and is thus one of the most cited papers in the Data Mining field.

However, it is possible that what is now called "association rules" is simliar to what appears in the 1966 paper [7] on GUHA, a general data mining method developed by Petr Hájek et al. [8].

KDD, GUHA Method and LISp-Miner System

- GUHA method and KDD
- LISp-Miner system
  - Overview
  - 6 GUHA procedures
  - application examples
  - bit string approach
- Current research projects
- Observational calculus – logical calculus of KDD patterns

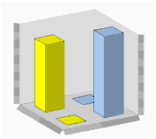
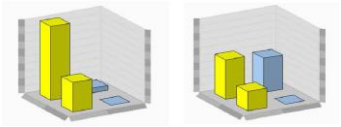
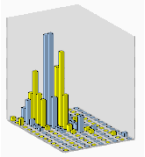
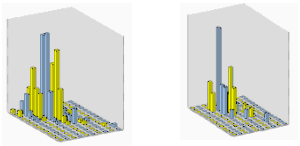
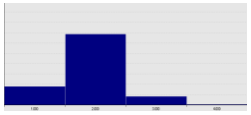
LISp-Miner System

http://lispminer.vse.cz/

- Developed since 1996 at LISp University of Economics, Prague
- Research and teaching of KDD
- Tools
  - Input data transformations - module DataSource
  - 7 GUHA procedures
  - Machine learning procedure KEX
- Research
  - EverMiner
  - SEWEBAR
  - Logical calculi for data mining
  - … , see http://lispminer.vse.cz/research/index.html
- Students work (bachelor, diploma, PhD.)

LISp-Miner, 7 GUHA procedures

http://lispminer.vse.cz

- 4ft-Miner 
- SD4ft-Miner 
- KL-Miner 
- SDKL-Miner 
- CF-Miner 
- SDCF-Miner 
- Ac4ft-Miner - new procedure

LISp-Miner, application examples

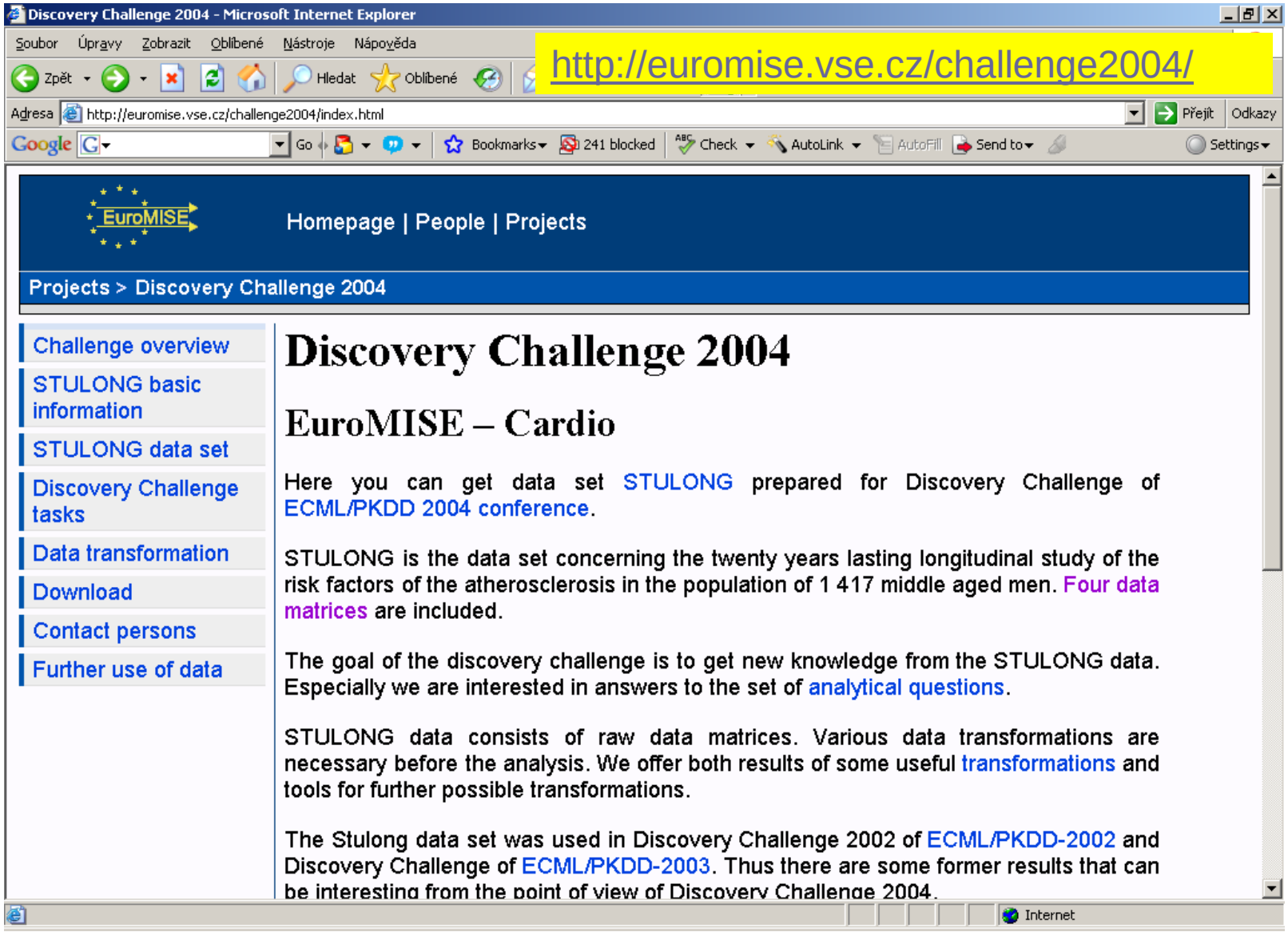
Goals of analysis:

- find new knowledge
- verify if the given knowledge observed in given data

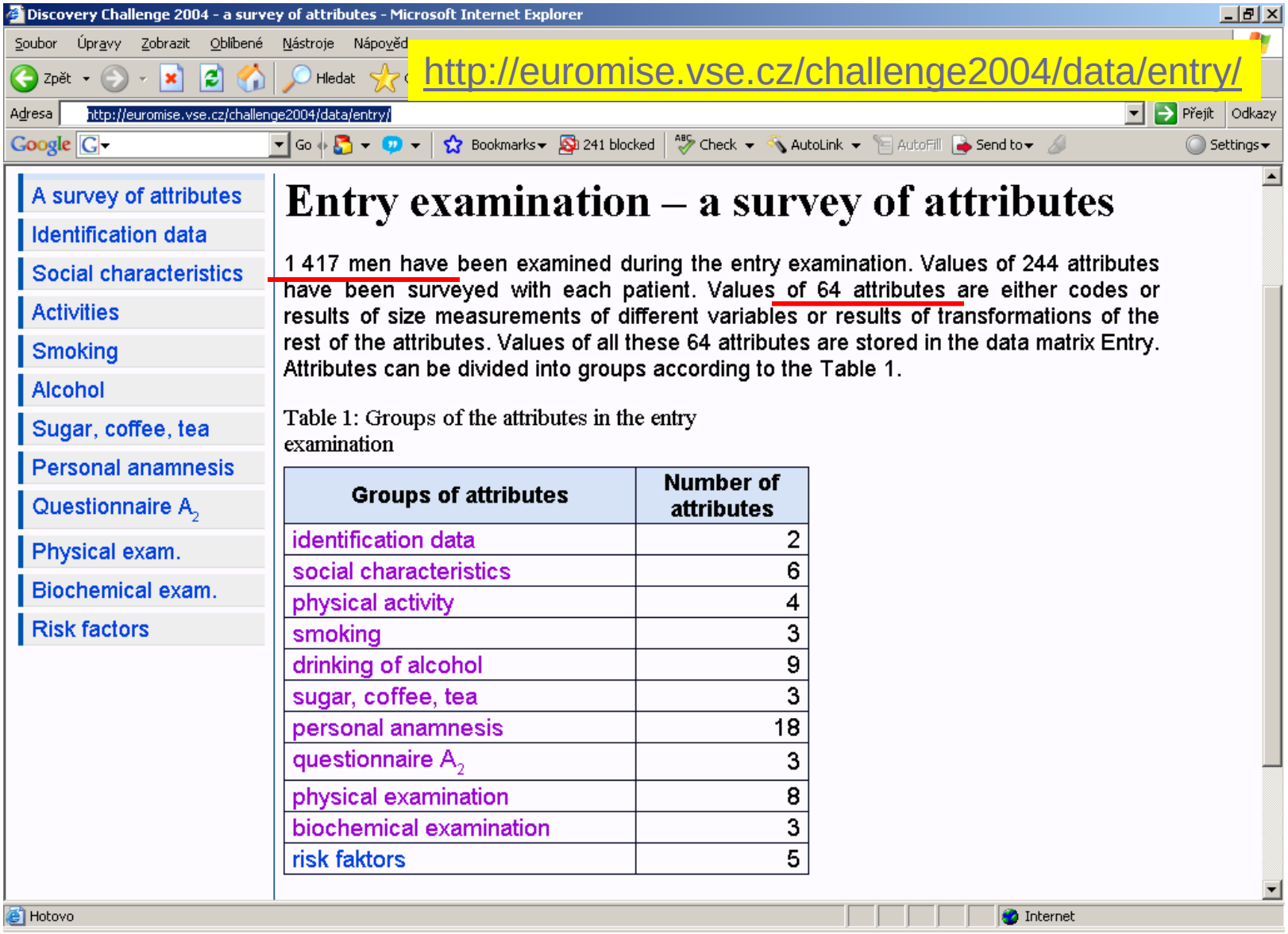
Presented:

- Stulong data set
- Analytical question $\mathcal{B}$ (Physical, Social) $\approx^{?}$ $\mathcal{B}$ (Biochemical)
- Analytical question normal $\otimes$ risk: $\mathcal{B}$ (Physical, Social) $\approx^{?}$ $\mathcal{B}$ (Biochemical)
- Analytical question ? $\uparrow\uparrow$? / $\mathcal{B}$ (*)

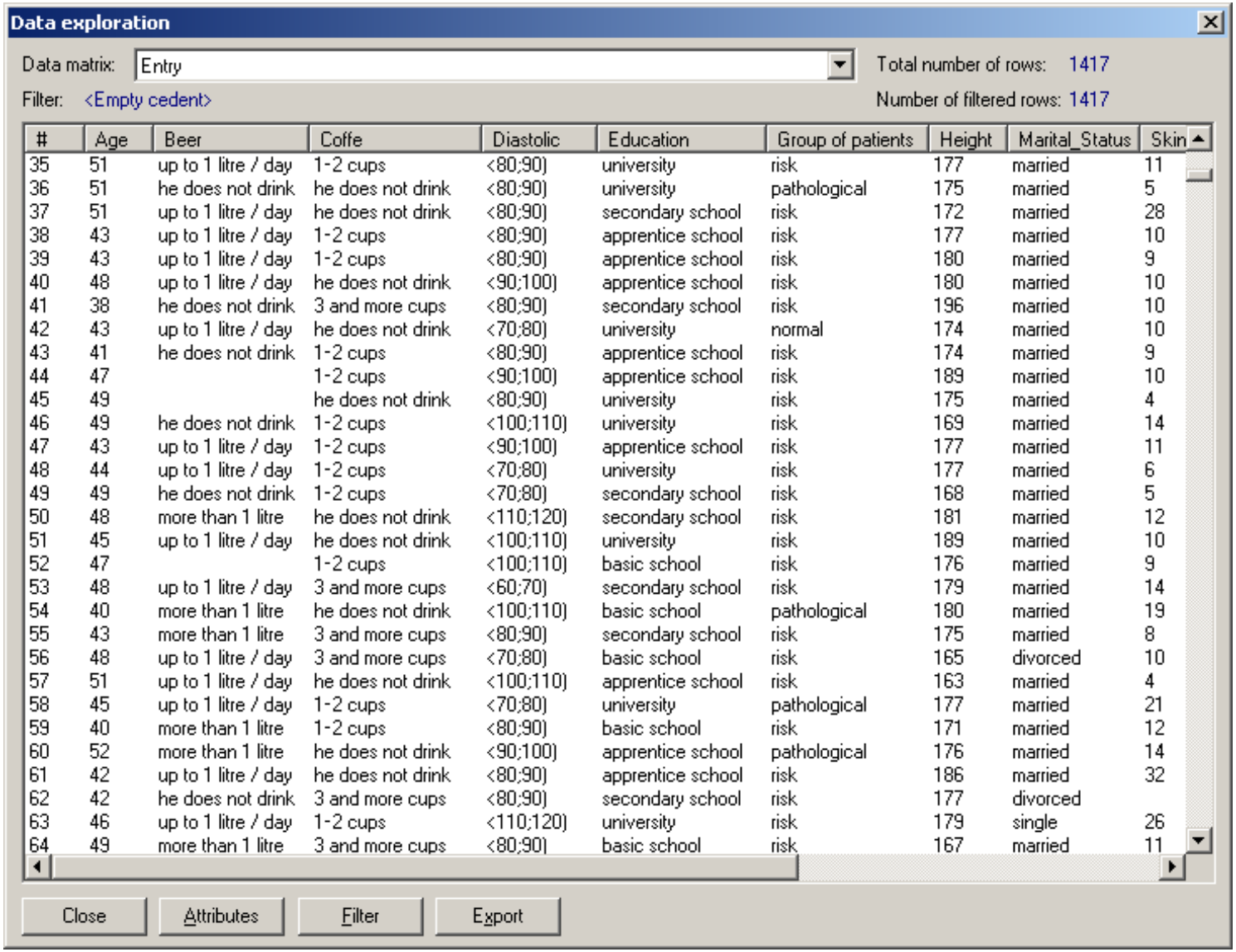
STULONG data set



STULONG data set



Entry data matrix



Data exploration

Data matrix: Entry — Total number of rows: 1417

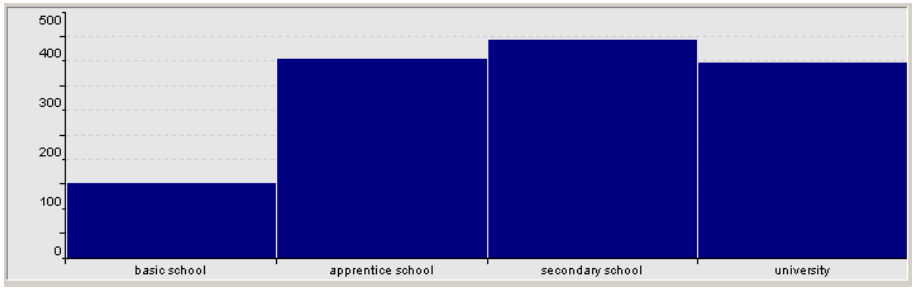
Filter: <Empty cedent> — Number of filtered rows: 1417

#	Age	Beer	Coffe	Diastolic	Education	Group of patients	Height	Marital_Status	Skin
35	51	up to 1 litre / day	1-2 cups	<80;90)	university	risk	177	married	11
36	51	he does not drink	he does not drink	<80;90)	university	pathological	175	married	5
37	51	up to 1 litre / day	he does not drink	<80;90)	secondary school	risk	172	married	28
38	43	up to 1 litre / day	1-2 cups	<80;90)	apprentice school	risk	177	married	10
39	43	up to 1 litre / day	1-2 cups	<80;90)	apprentice school	risk	180	married	9
40	48	up to 1 litre / day	he does not drink	<90;100)	apprentice school	risk	180	married	10
41	38	he does not drink	3 and more cups	<80;90)	secondary school	risk	196	married	10
42	43	up to 1 litre / day	he does not drink	<70;80)	university	normal	174	married	10
43	41	he does not drink	1-2 cups	<80;90)	apprentice school	risk	174	married	9
44	47		1-2 cups	<90;100)	apprentice school	risk	189	married	10
45	49		he does not drink	<80;90)	university	risk	175	married	4
46	49	he does not drink	1-2 cups	<100;110)	university	risk	169	married	14
47	43	up to 1 litre / day	1-2 cups	<90;100)	apprentice school	risk	177	married	11
48	44	up to 1 litre / day	1-2 cups	<70;80)	university	risk	177	married	6
49	49	he does not drink	1-2 cups	<70;80)	secondary school	risk	168	married	5
50	48	more than 1 litre	he does not drink	<110;120)	secondary school	risk	181	married	12
51	45	up to 1 litre / day	he does not drink	<100;110)	university	risk	189	married	10
52	47		1-2 cups	<100;110)	basic school	risk	176	married	9
53	48	up to 1 litre / day	3 and more cups	<60;70)	secondary school	risk	179	married	14
54	40	more than 1 litre	he does not drink	<100;110)	basic school	pathological	180	married	19
55	43	more than 1 litre	3 and more cups	<80;90)	secondary school	risk	175	married	8
56	48	up to 1 litre / day	3 and more cups	<70;80)	basic school	risk	165	divorced	10
57	51	up to 1 litre / day	he does not drink	<100;110)	apprentice school	risk	163	married	4
58	45	up to 1 litre / day	1-2 cups	<70;80)	university	pathological	177	married	21
59	40	more than 1 litre	1-2 cups	<80;90)	basic school	risk	171	married	12
60	52	more than 1 litre	he does not drink	<90;100)	apprentice school	pathological	176	married	14
61	42	up to 1 litre / day	he does not drink	<80;90)	apprentice school	risk	186	married	32
62	42	he does not drink	3 and more cups	<80;90)	secondary school	risk	177	divorced	
63	46	up to 1 litre / day	1-2 cups	<110;120)	university	risk	179	single	26
64	49	more than 1 litre	3 and more cups	<80;90)	basic school	risk	167	married	11

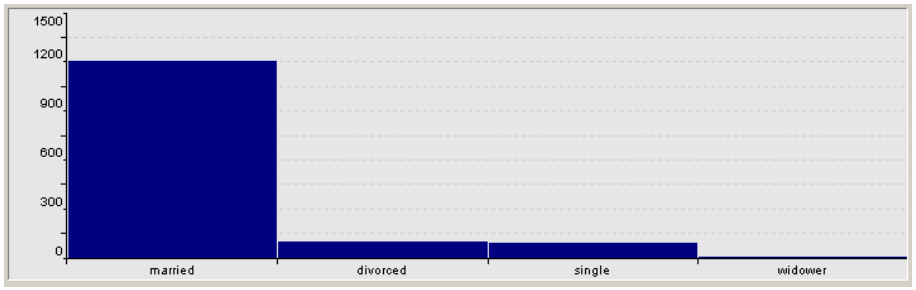
Close | Attributes | Filter | Export

Entry data matrix - Social characteristics

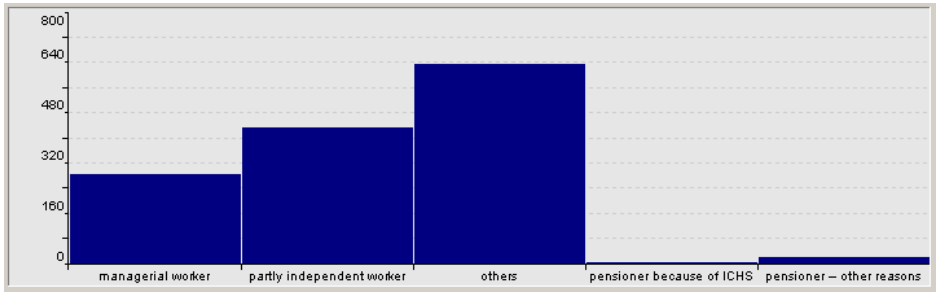
Education



Marital status

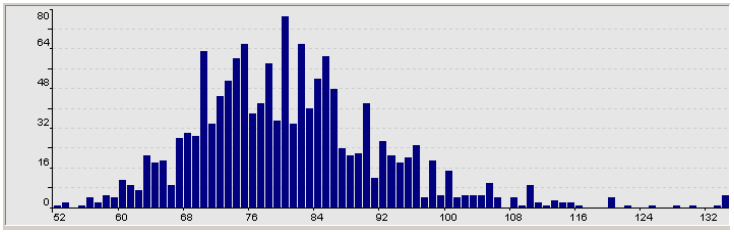


Responsibility in a job

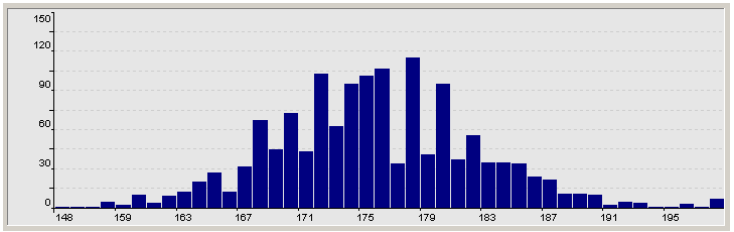


Entry data matrix - Physical examination

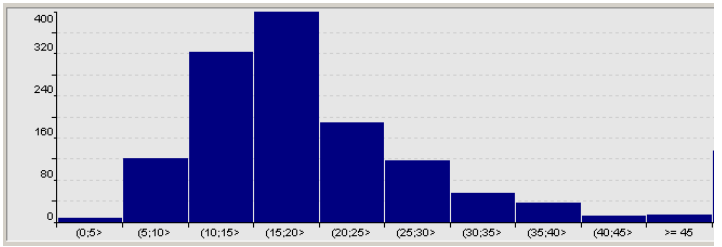
Weight [kg]



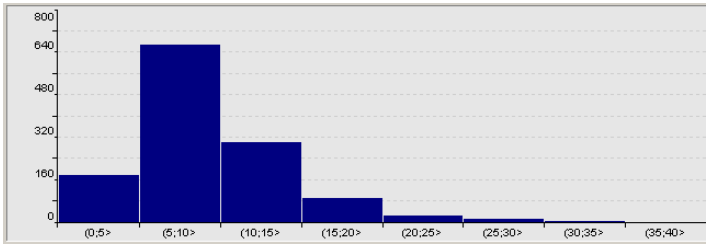
Height [cm]



Skinfold above musculus triceps [mm]



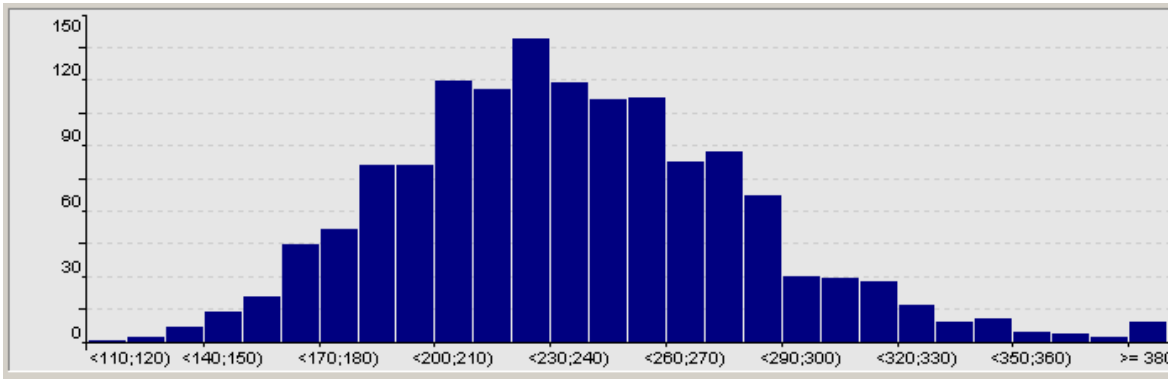
Skinfold above musculus subscapularis [mm]



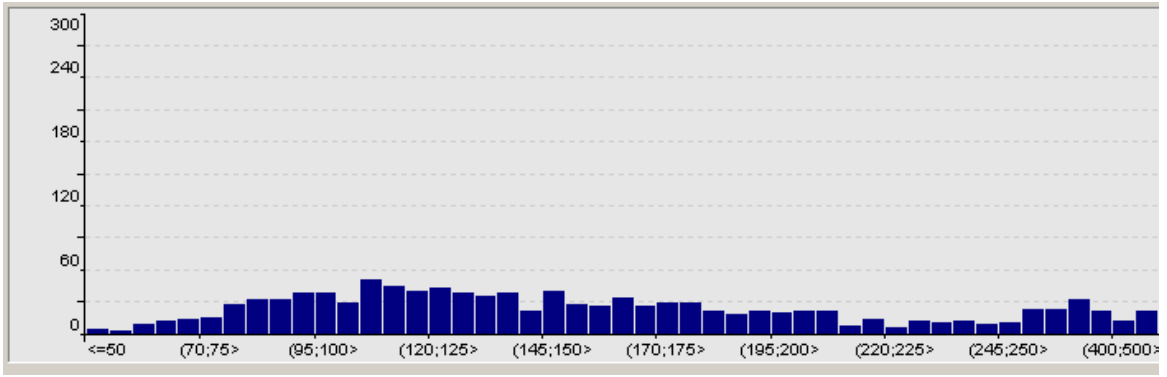
…… additional attributes

Entry data matrix - Biochemical examination

Cholesterol [mg%]



Triglycerides in mg%



LISp-Miner, application examples

Goals of analysis:

- find new knowledge
- verify if the given knowledge observed in given data

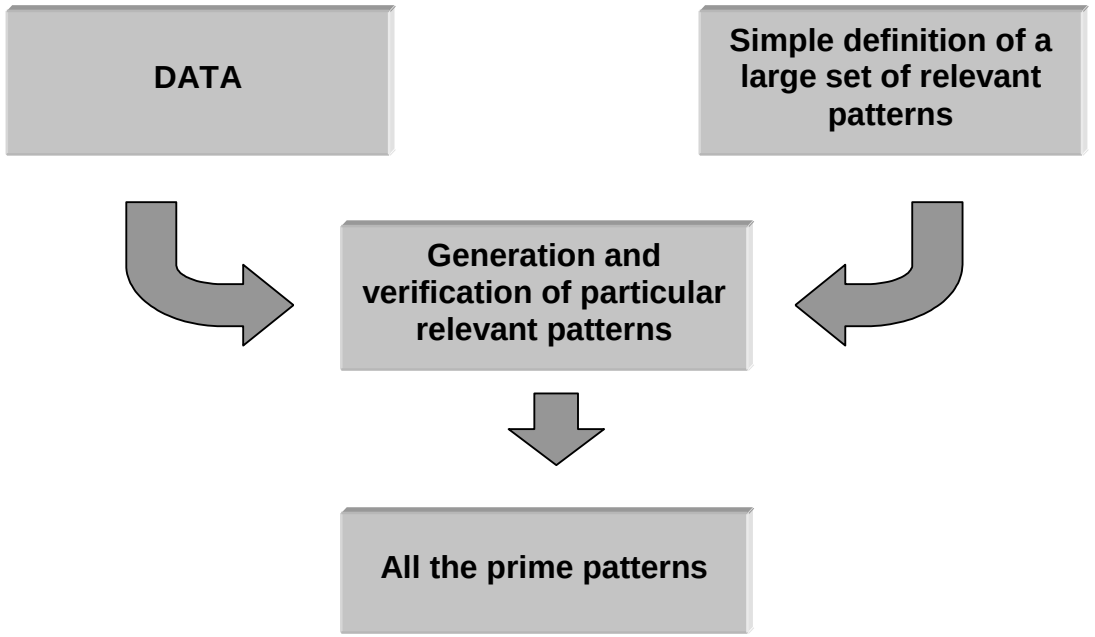
Presented:

- Stulong data set
- Analytical question $\mathcal{B}$ (Physical, Social) $\approx^{?}$ $\mathcal{B}$ (Biochemical)
- Analytical question normal $\otimes$ risk: $\mathcal{B}$ (Physical, Social) $\approx^{?}$ $\mathcal{B}$ (Biochemical)
- Analytical question ? $\uparrow\uparrow$? / $\mathcal{B}$ (*)

$\mathcal{B}$(Social, Physical) $\approx^{?}$ $\mathcal{B}$(Biochemical)

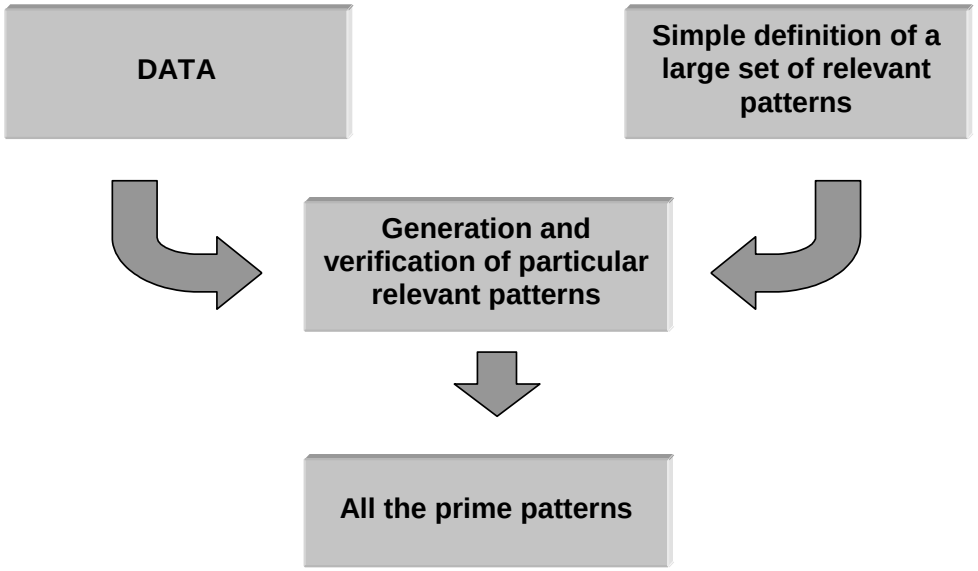
In the ENTRY data matrix, are there some interesting relations between Boolean attributes describing combination of results of Physical examination and Social characteristics and results of Biochemical examination?

GUHA procedure:



$\mathcal{B}$(Social, Physical) $\approx^?$ $\mathcal{B}$(Biochemical)

Applying GUHA procedure 4ft-Miner



Patterns: $\varphi \approx^? \psi$

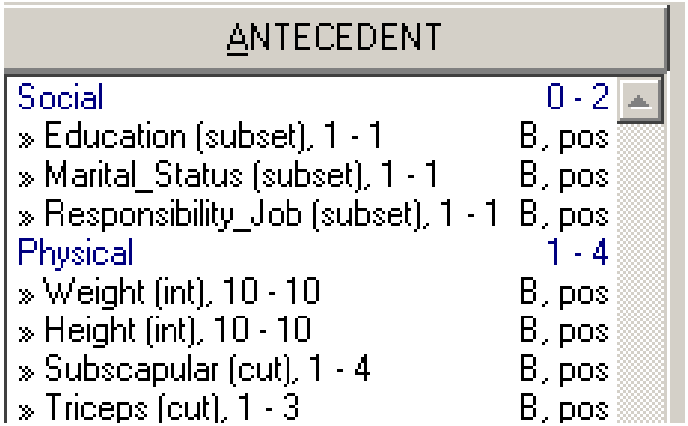
$\varphi \in \mathcal{B}$ (Physical, Social)

$\psi \in \mathcal{B}$ (Biochemical)

$\approx^?$ evaluated using 4-fould table

ENTRY	ψ	$\neg\psi$
φ	a	b
$\neg\varphi$	c	d

Defining $\mathcal{B}$ (Social, Physical) (1)



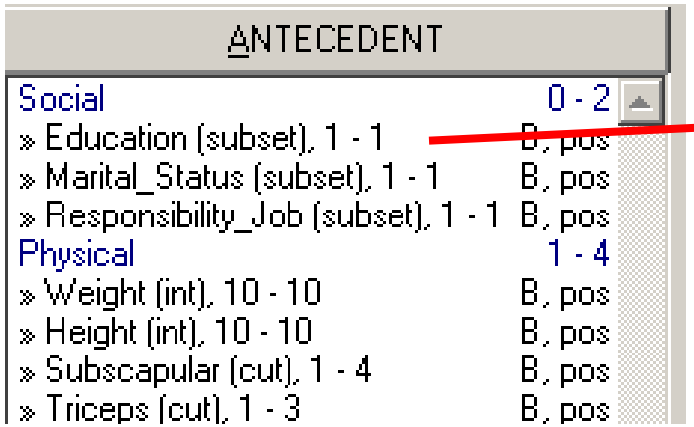
Entry	Succedent	¬ Succedent
Antecedent	a	b
¬ Antecedent	c	d

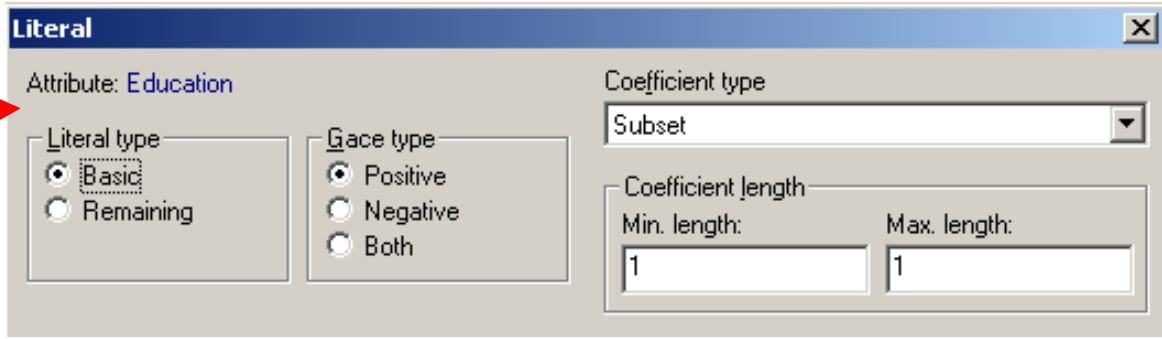
$$\mathcal{B}(\text{Social}, \text{Physical}) = \mathcal{B}(\text{Social}) \wedge \mathcal{B}(\text{Physical})$$

$$\mathcal{B}(\text{Social}) = \bigwedge_{0}^{2} [\mathcal{B}(\text{Education}), \mathcal{B}(\text{Marital Status}), \mathcal{B}(\text{Responsibility\_Job})]$$

$$\mathcal{B}(\text{Physical}) = \bigwedge_{1}^{4} [\mathcal{B}(\text{Weight}), \mathcal{B}(\text{Height}), \mathcal{B}(\text{Subscapular}), \mathcal{B}(\text{Triceps})]$$

Defining $\mathcal{B}$ (Social, Physical) (2)





Set of categories of Education:

basic school, apprentice school, secondary school, university

$\mathcal{B}$ (Education): Subsets of length 1 - 1

Education (basic school), Education (apprentice school)

Education (secondary school), Education (university)

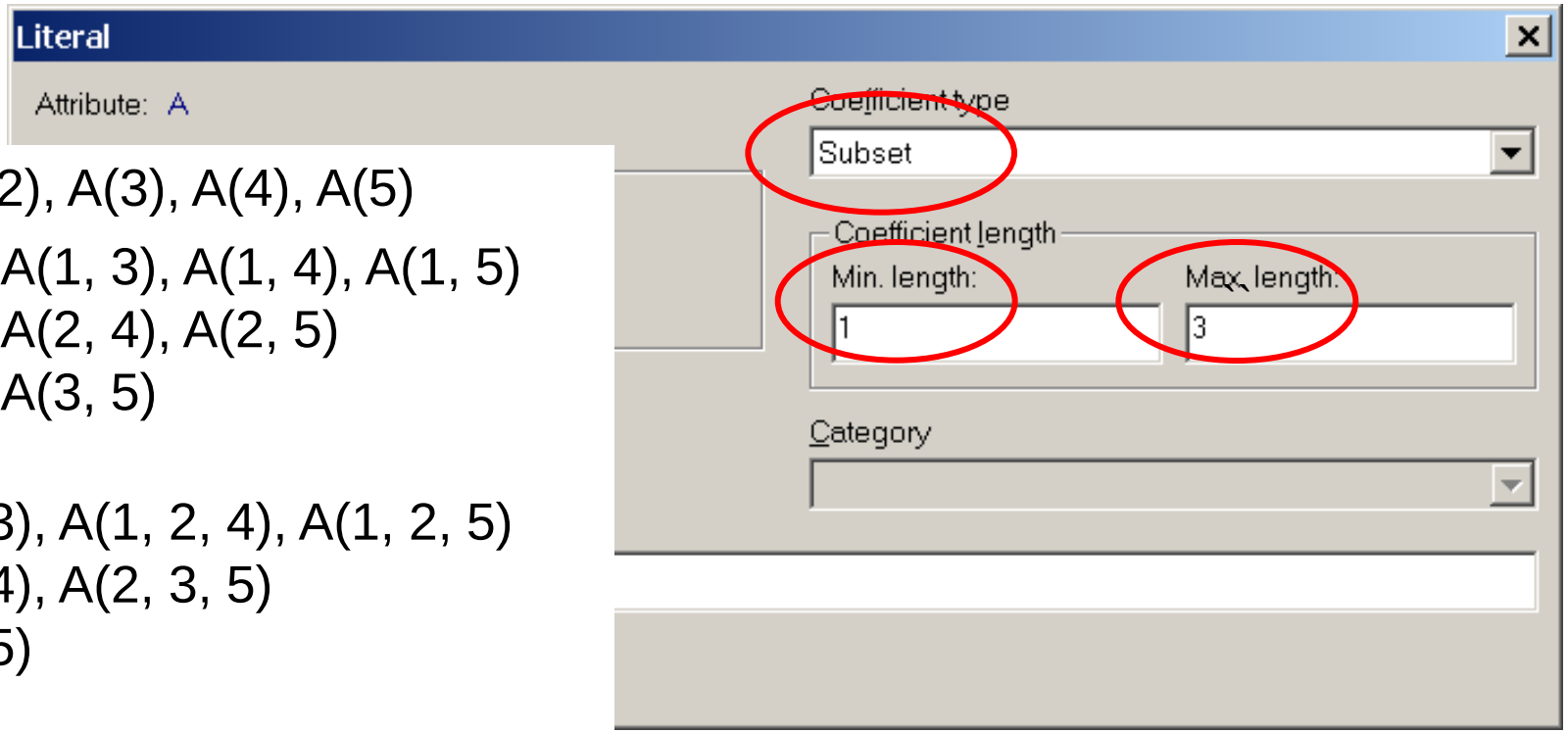
Remark: Attribute A with categories 1, 2, 3, 4, 5

Literals with coefficients Subset (1 – 3):

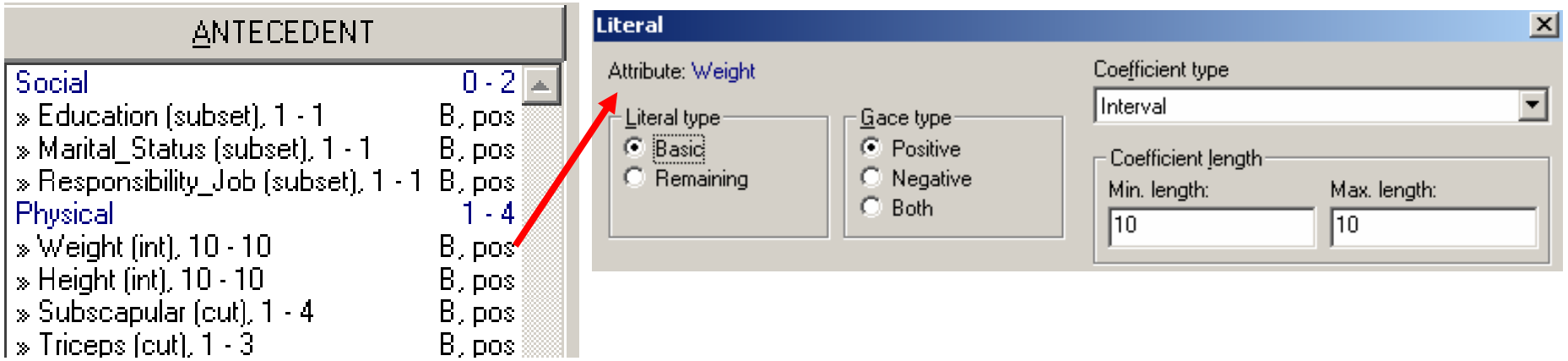
A(1), A(2), A(3), A(4), A(5)

A(1, 2), A(1, 3), A(1, 4), A(1, 5)
A(2, 3), A(2, 4), A(2, 5)
A(3, 4), A(3, 5)
A(4, 5)

A(1, 2, 3), A(1, 2, 4), A(1, 2, 5)
A(2, 3, 4), A(2, 3, 5)
A(3, 4, 5)



Defining $\mathcal{B}$ (Social, Physical) (3)



Set of categories of Weight: 52, 53, 54, 55, ……., 130, 131, 132, 133

$\mathcal{B}$ (Weight): Intervals of length 10 - 10: Weight(52 – 61), Weight(53 – 62), …

52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63,...., 128, 129, 130, 131, 132, 133

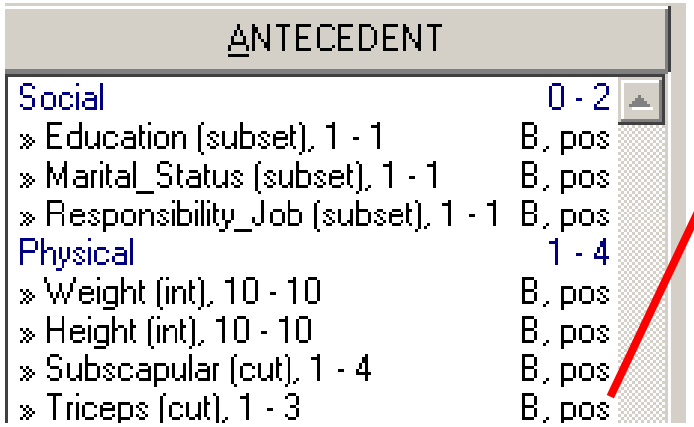
52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63,...., 128, 129, 130, 131, 132, 133

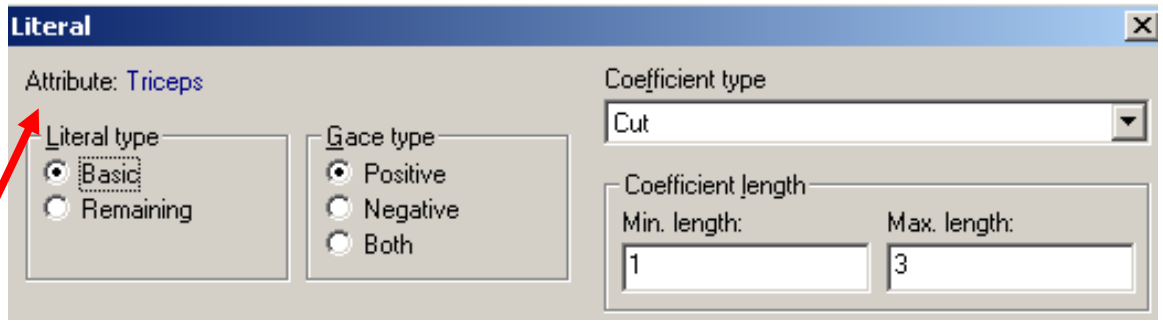
52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63,...., 128, 129, 130, 131, 132, 133

,....,

52, 53, 54, 55, 56, ..., 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133

Defining $\mathcal{B}$ (Social, Physical) (4)





Set of categories of Triceps: (0;5⟩, (5;10⟩, (10;15⟩, …, (25;30⟩ (30;35⟩ (35;40⟩

$\mathcal{B}$ (Triceps): Cuts 1 - 3

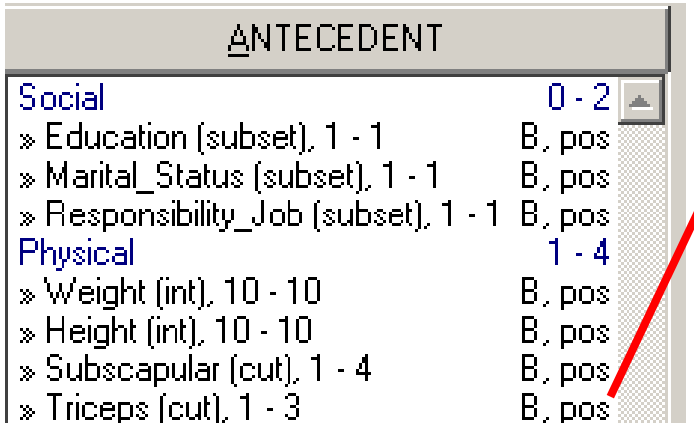
Left cuts 1 – 3

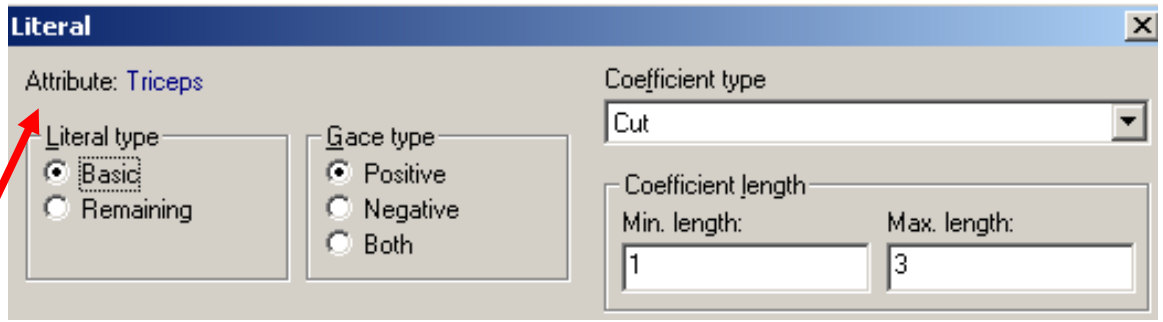
(0;5⟩, (5;10⟩, (10;15⟩, (15;20⟩, (20;25⟩, (25;30⟩, (30;35⟩, (35;40 ⟩ **i.e. Triceps(1 – 5)**

(0;5⟩, (5;10⟩, (10;15⟩, (15;20⟩, (20;25⟩, (25;30⟩, (30;35⟩, (35;40 ⟩ **i.e. Triceps(1 – 10)**

(0;5⟩, (5;10⟩, (10;15⟩, (15;20⟩, (20;25⟩, (25;30⟩, (30;35⟩, (35;40 ⟩ **i.e. Triceps(1 – 15)**

Defining $\mathcal{B}$ (Social, Physical) (5)





Set of categories of Triceps: (0;5⟩, (5;10⟩, (10;15⟩, …, (25;30⟩ (30;35⟩ (35;40⟩

$\mathcal{B}$ (Triceps): Cuts 1 - 3

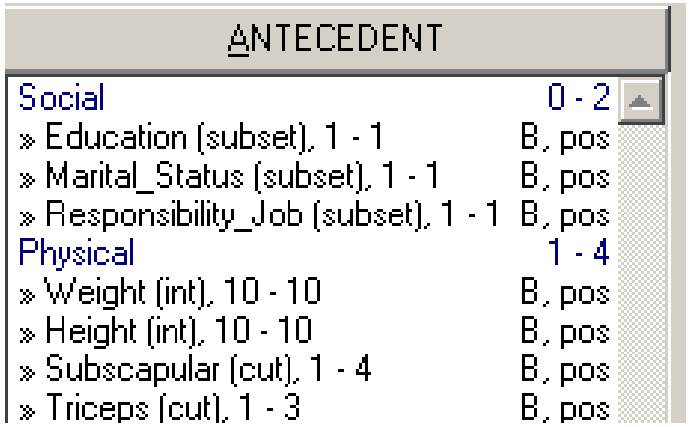
Right cuts 1 – 3

(0;5⟩, (5;10⟩, (10;15⟩, (15;20⟩, (20;25⟩, (25;30⟩, (30;35⟩, (35;40 ⟩ **i.e. Triceps(35 – 40)**

(0;5⟩, (5;10⟩, (10;15⟩, (15;20⟩, (20;25⟩, (25;30⟩, (30;35⟩, (35;40 ⟩ **i.e. Triceps(30 – 40)**

(0;5⟩, (5;10⟩, (10;15⟩, (15;20⟩, (20;25⟩, (25;30⟩, (30;35⟩, (35;40 ⟩ **i.e. Triceps(25 – 45)**

Defining $\mathcal{B}$(Social, Physical) (6)



Examples of $\varphi \in \mathcal{B}$(Social, Physical):

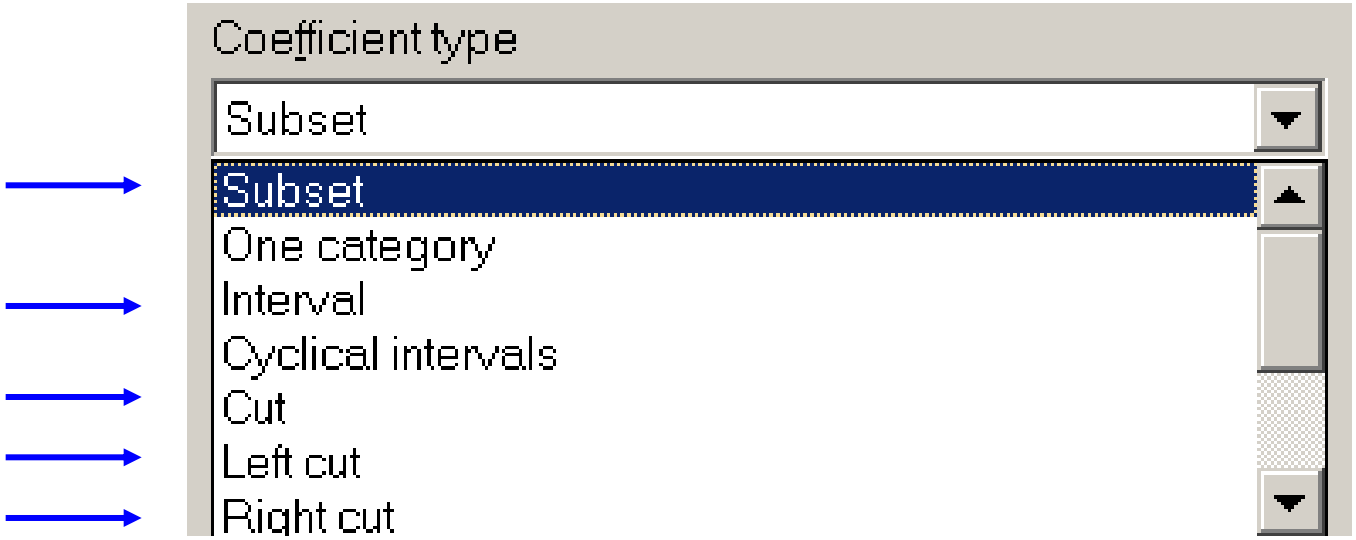
Education (basic school)

Education (university) $\wedge$ Marital_Status(single) $\wedge$ Weight (52 – 61)

Marital_Status(divorced) $\wedge$ Weight (52 – 61) $\wedge$ Triceps (25 – 45)

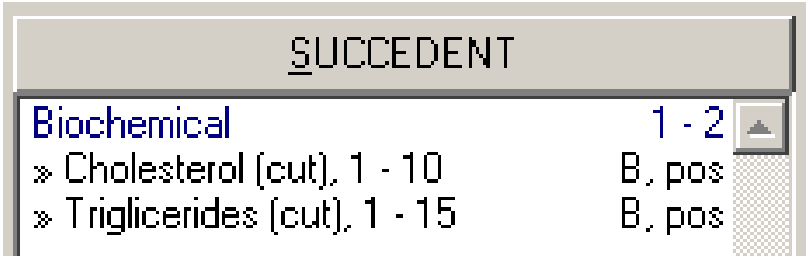
Weight (52 – 61) $\wedge$ Height (52 – 61) $\wedge$ Subscapular(0 – 10) $\wedge$ Triceps (25 – 45)

Remark: Types of coefficients



→ See examples above

Defining $\mathcal{B}$ (Biochemical)



Analogously to $\mathcal{B}$ (Social, Physical)

Examples of $\psi \in \mathcal{B}$ (Biochemical):

Cholesterol (110 – 120), Cholesterol (110 – 130), …, Cholesterol (110 – 210)

Cholesterol ($\geq$ 380), Cholesterol ($\geq$ 370), …, Cholesterol ($\geq$ 290)

Cholesterol ($\geq$ 380) $\wedge$ Triglicerides ($\leq$ 50), …

Cholesterol ($\geq$ 380) $\wedge$ Triglicerides ($\leq$ 300), …

…,

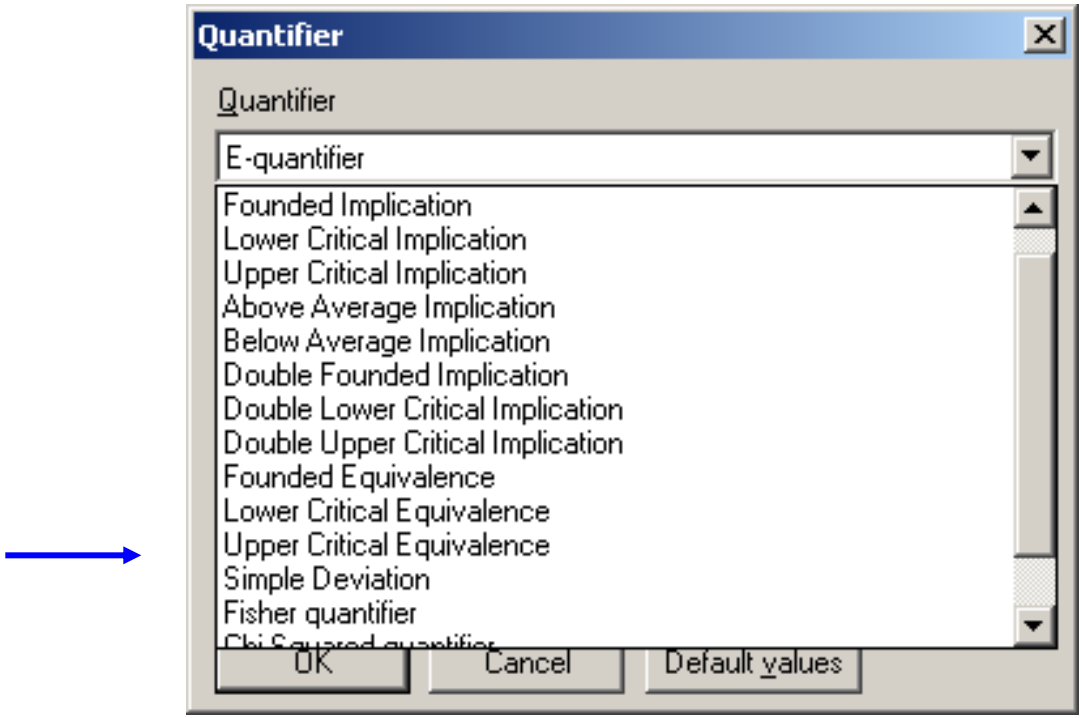
Defining $\approx^?$ in $\varphi \approx^? \psi$

$\approx^?$ is called 4ft – quantifier, it corresponds to a condition concerning

4ft(φ, ψ, $\mathcal{M}$) – four fold table of φ and ψ in $\mathcal{M}$

$\mathcal{M}$	ψ	$\neg\psi$
φ	a	b
$\neg\varphi$	c	d

17 types of 4ft-quantifiers



Two examples of $\approx^?$

$\mathcal{M}$	ψ	$\neg\psi$
φ	a	b
$\neg\varphi$	c	d

Founded implication $\Rightarrow_{p,B}$ $\quad \frac{a}{a+b} \geq p \wedge a \geq B$

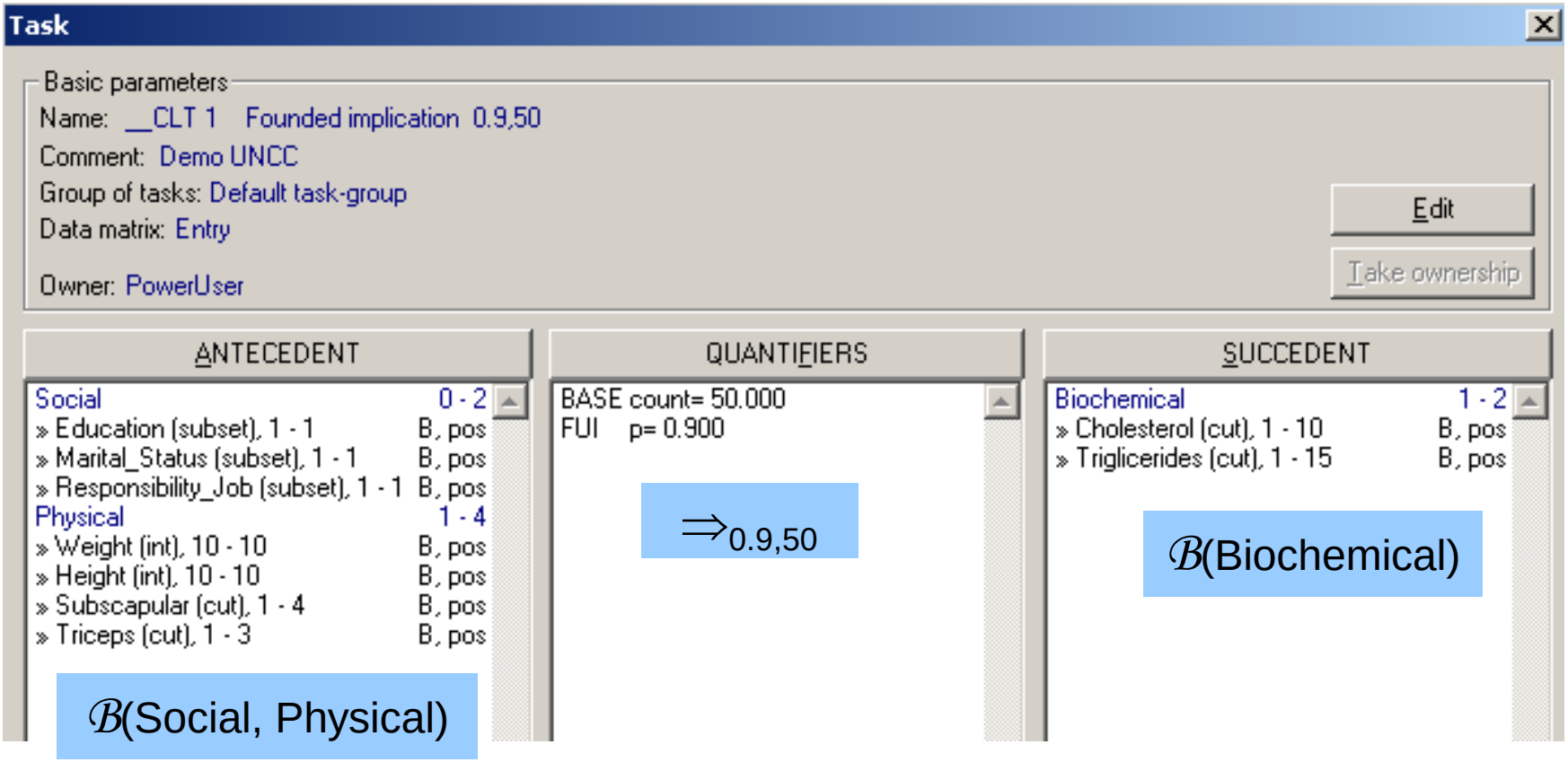
$\varphi \Rightarrow_{p,B} \psi$: at least 100p per cent of objects of $\mathcal{M}$ satisfying φ satisfy also ψ

φ and there are at least Base objects satisfying both φ and ψ

Above average $\Rightarrow^{+}_{p,B}$ $\quad \frac{a}{a+b} \geq (1+p)\frac{a+c}{a+b+c+d} \wedge a \geq B$

$\varphi \Rightarrow^{+}_{p,B} \psi$: the relative frequency of objects of M satisfying ψ among the objects satisfying φ is at least 100p per cent higher than the relative frequency of ψ in the whole data matrix $\mathcal{M}$ and there are at least Base objects satisfying both φ and ψ

Solving $\mathcal{B}$(Social, Physical) $\Rightarrow_{0.9,50}$ $\mathcal{B}$(Biochemical) (1)



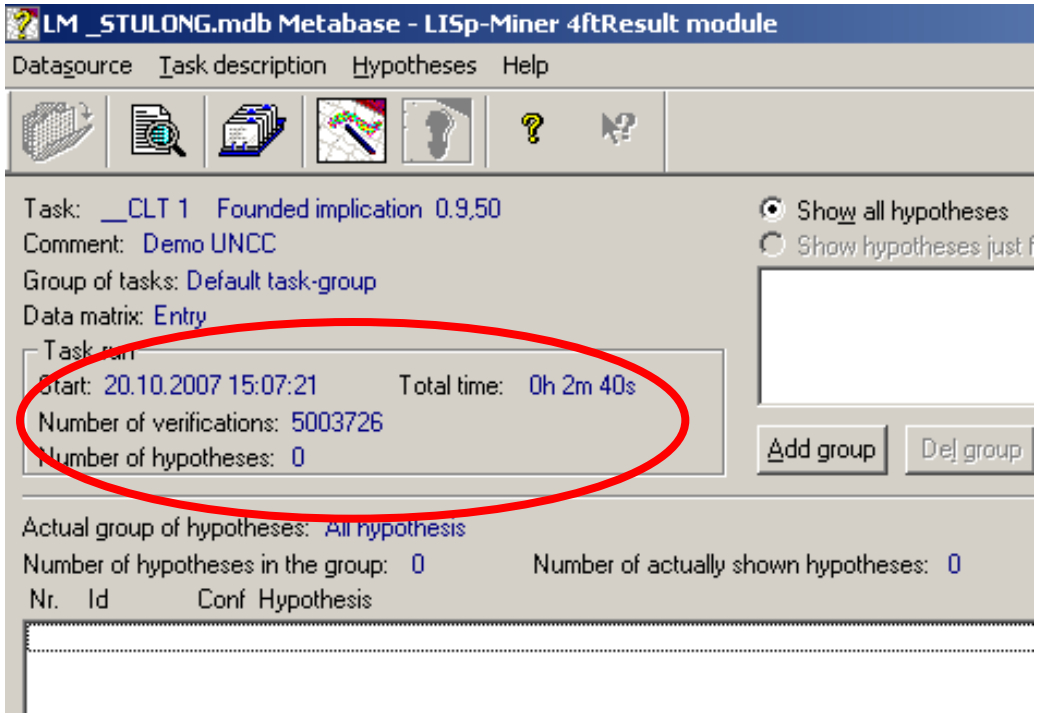
Solving $\mathcal{B}$(Social, Physical) $\Rightarrow_{0.9,50}$ $\mathcal{B}$(Biochemical) (2)

PC with 1.66 GHz, 2 GB RAM

2 min. 40 sec.

5 . 10^6 rules verified

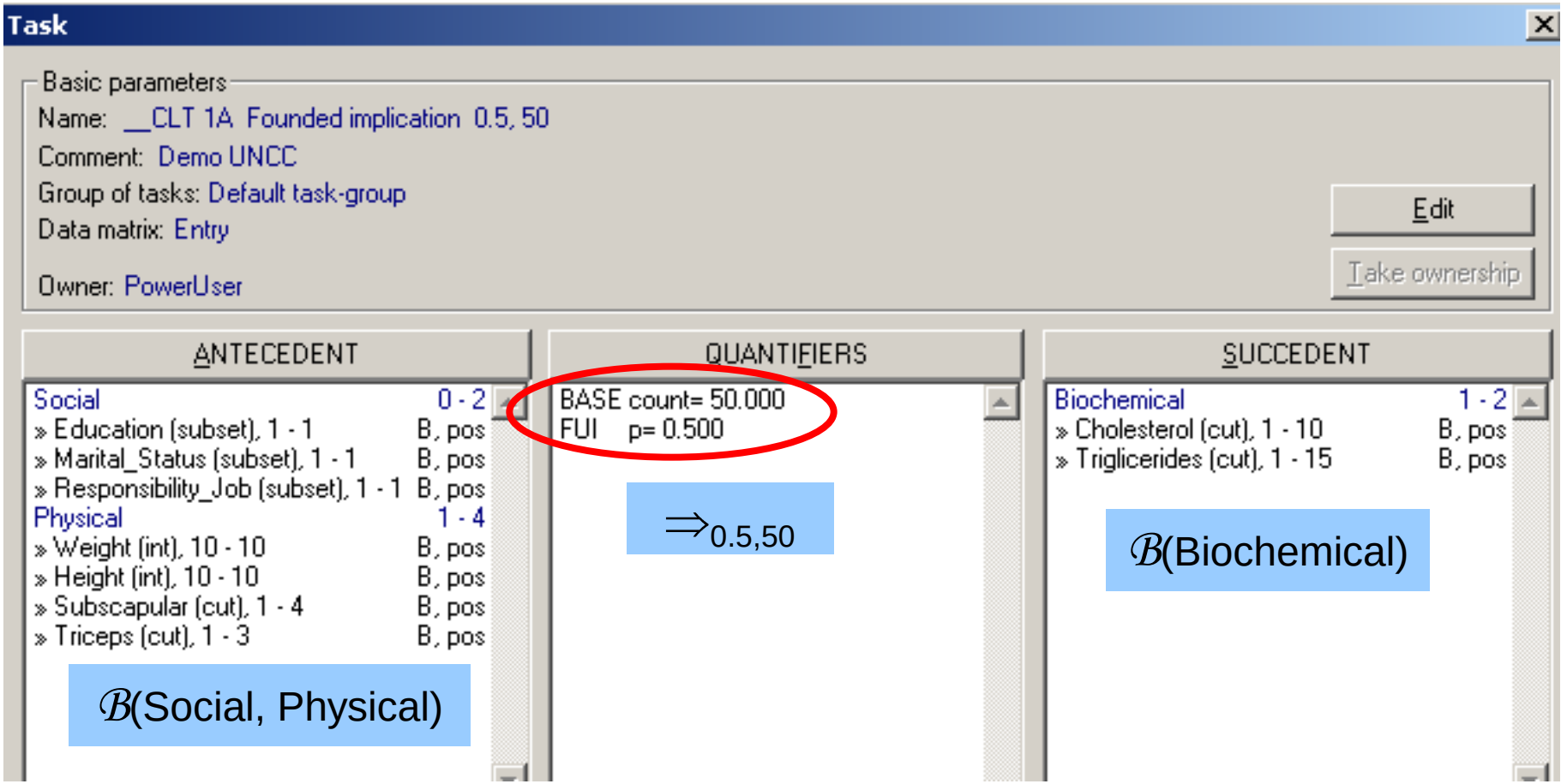
0 true patterns



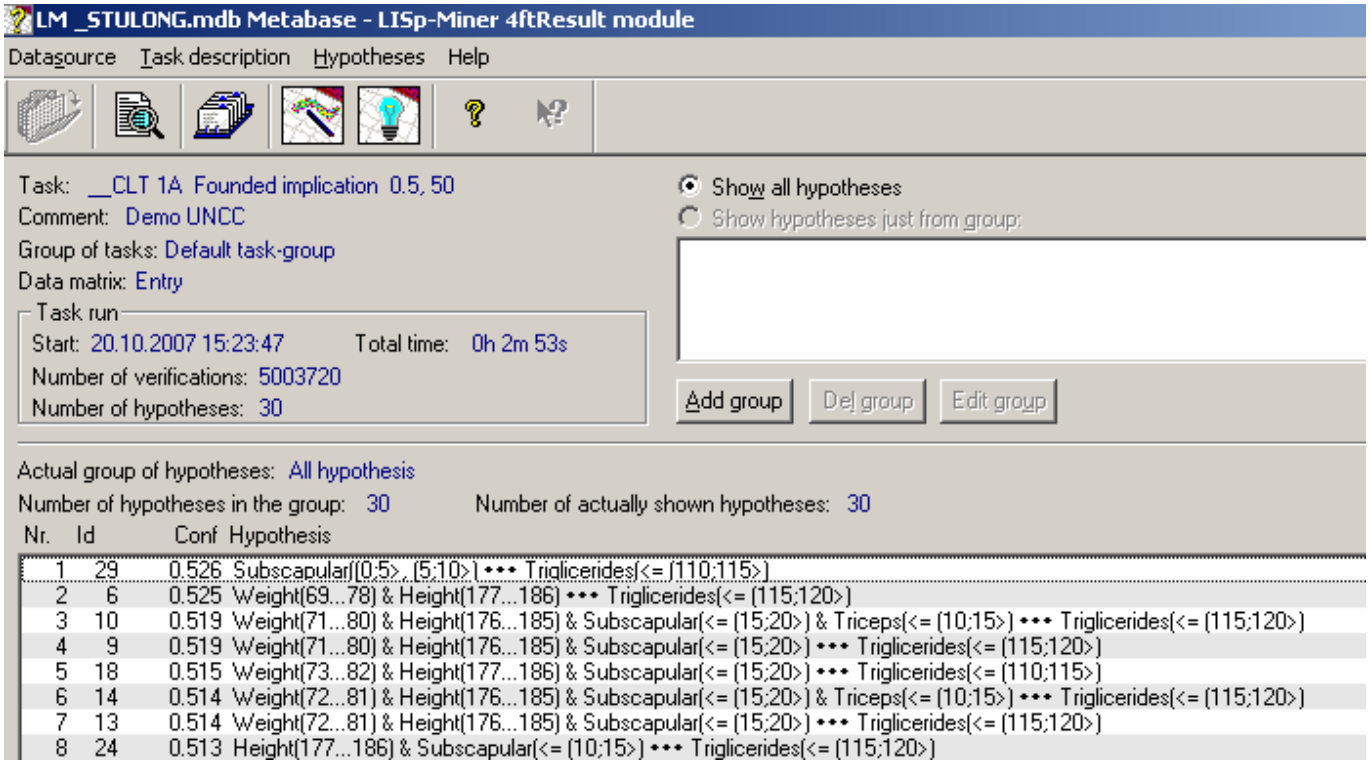
Problem: Confidence 0.9 in $\Rightarrow_{0.9,50}$ too high

Solution: Use confidence 0.5

Solving $\mathcal{B}$(Social, Physical) $\Rightarrow_{0.5,50}$ $\mathcal{B}$(Biochemical) (1)



Solving $\mathcal{B}$(Social, Physical) $\Rightarrow_{0.5,50}$ $\mathcal{B}$(Biochemical) (2)



30 rules with confidence $\geq$ 0.5

Problem: The strongest rule has confidence only 0.526, see detail

Solution: Search for rules expressing 70% higher relative frequency than average

It means to use $\Rightarrow^{+}_{0.7,50}$ instead of $\Rightarrow_{0.5,50}$

Solving $\mathcal{B}$(Social, Physical) $\Rightarrow_{0.5,50}$ $\mathcal{B}$(Biochemical) (3)

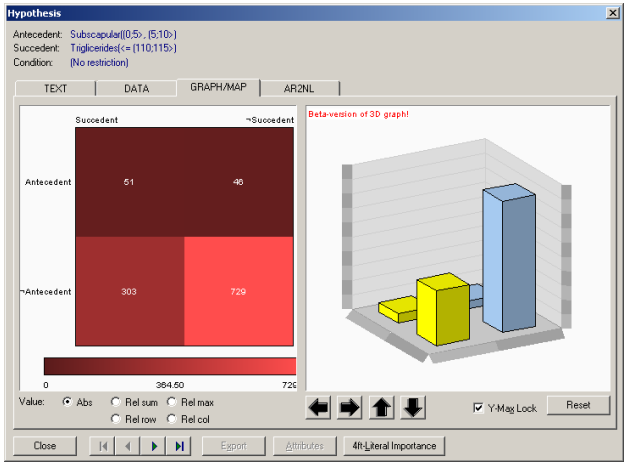
Detail of results - the strongest rule

Antecedent: Subscapular([0;5>, [5;10>)
Succedent: Triglicerides(<= (110;115>)
Condition: (No restriction)

TEXT | DATA | GRAPH/MAP | AR2NL

**** Hypothesis ID: 29

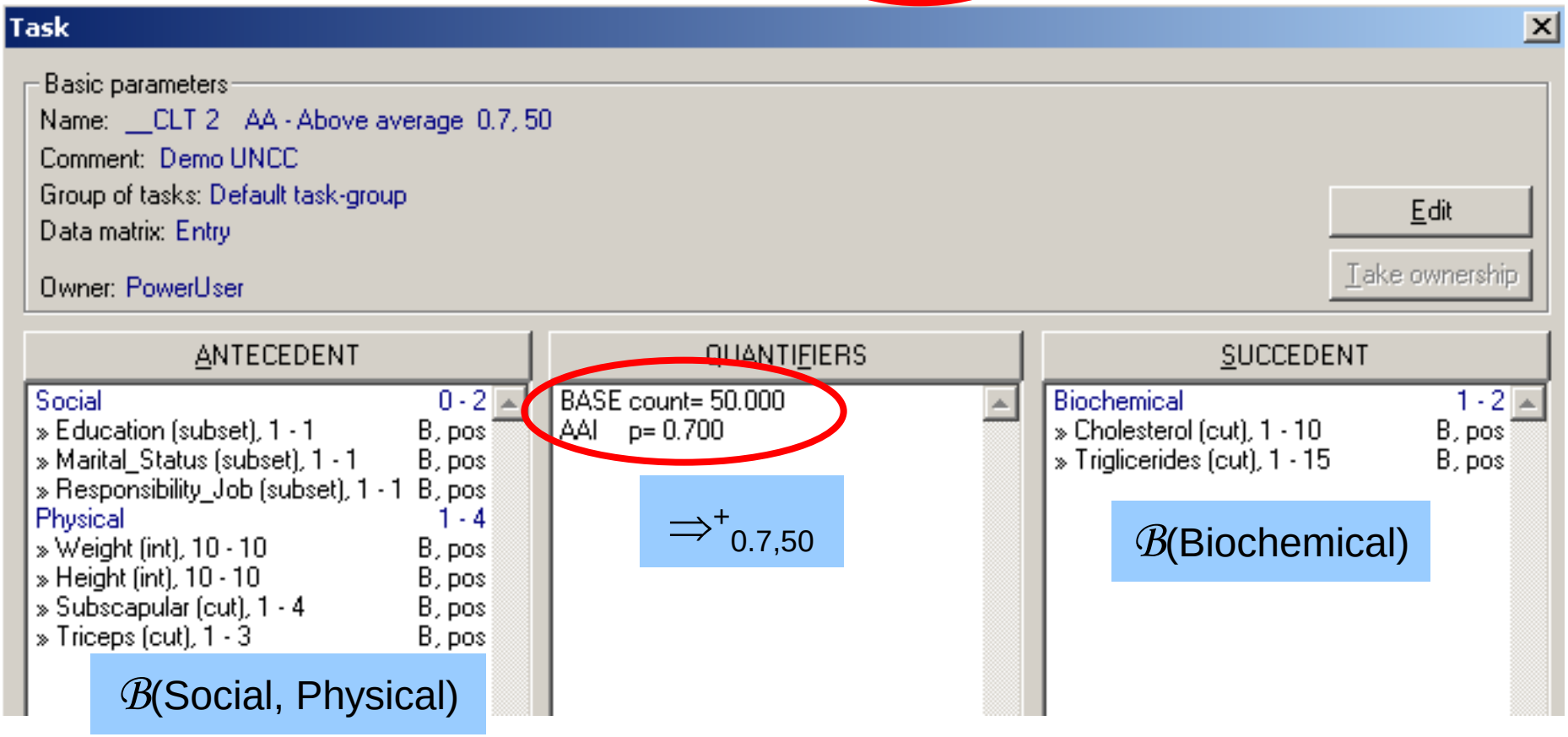
Antecedent	Subscapular	[0;5>, [5;10>	
Succedent	Triglicerides	(<= (110;115>)	
Contingency table			
	Succedent	NOT Succedent	
Antecedent	51	46	97
NOT Antecedent	303	729	1032
	354	775	1129
Values from contingency table:			
a	51	51	a-frequency from the contingency table
b	46	46	b-frequency from the contingency table
c	303	303	c-frequency from the contingency table
d	729	729	d-frequency from the contingency table
r	97	97	r-frequency (a+b) from the contingency table
n	1129	1129	n-frequency (a+b+c+d) from the contingency table
Conf	0.53	0.5257731959	Confidence (validity): a/(a+b)



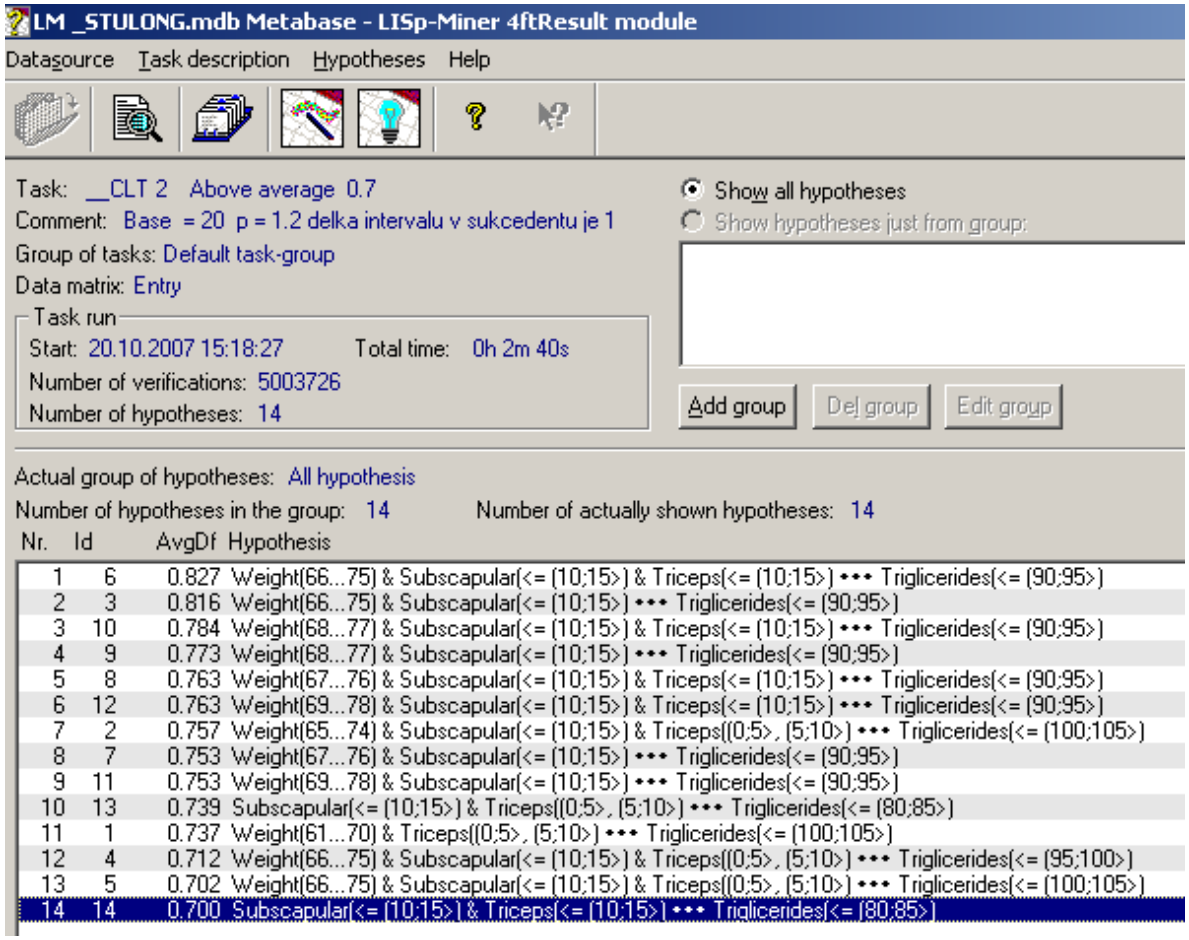
Entry	Triglicerides(≤115)	¬ Triglicerides(≤115)
Subscapular(0;10⟩	51	46
¬ Subscapular(0;10⟩	303	729

Subscapular(0;10⟩ $\Rightarrow_{0.53,\ 51}$ Triglicerides(≤115)

Solving $\mathcal{B}$(Social, Physical) $\Rightarrow^{+}_{0.7,50}$ $\mathcal{B}$(Biochemical) (1)



Solving $\mathcal{B}$(Social, Physical) $\Rightarrow^{+}_{0.7,50}$ $\mathcal{B}$(Biochemical) (2)



14 rules with relative frequency of succedent $\geq$ 0.7 than average

Example – see detail

Solving $\mathcal{B}$(Social, Physical) $\Rightarrow^{+}_{0.7,50}$ $\mathcal{B}$(Biochemical) (3)

Detail of results - the strongest rule (i.e. the greatest AA)

φ: Weight (65;75⟩ $\wedge$ Subscapular($\leq$15) $\wedge$ Triceps($\leq$15)

ψ: Triglicerides ($\leq$ 95)

confidence = 51 / 165 = 0.31 (not interesting!)

Entry	ψ	$\neg\,\psi$	
φ	51	114	165
$\neg\,\varphi$	140	824	964
	191	938	1129

relative frequency of patients satisfying ψ in the whole Entry data matrix: $\frac{51+140}{51+114+140+824} = 0.17$

relative frequency of patients satisfying ψ among the patients satisfying φ: $\frac{51}{51+114} = 0.31$ i.e. 82 % higher

$$\frac{51}{51+114} = (1+0.82)\frac{51+140}{51+114+140+824}$$

thus $\varphi \Rightarrow^{+}_{0.82,51} \psi$

4ft-Miner Summary

- mines rules $\mathcal{B}$(attributes group 1) ≈ $\mathcal{B}$(attributes group 2)
- also onditional rules $\mathcal{B}$(attributes 1) ≈ $\mathcal{B}$(attributes 2) / $\mathcal{B}$(attributes 3)
- very fine tools to define set of Boolean characteristics $\mathcal{B}$ (attributes)
- automatically generates general literals like
  - Subscapular (≤15)
  - Weight (66;75⟩
- various measures of association on $\langle a, b, c, d \rangle$
- works very fast
- output directly filtered by chosen measure
- does not use Apriori, uses bit string approach

LISp-Miner, application examples

Goals of analysis:

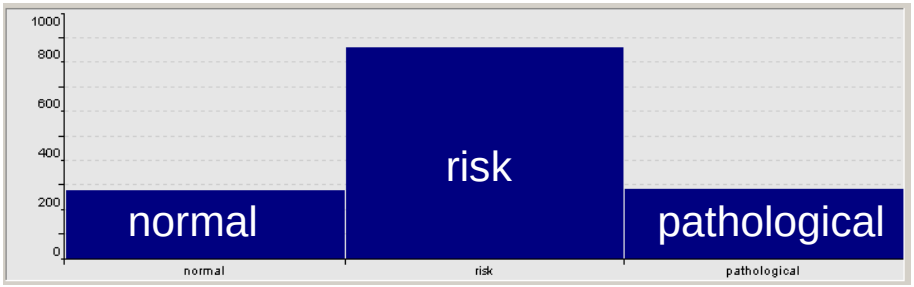
- find new knowledge
- verify if the given knowledge observed in given data

Presented:

- Stulong data set
- Analytical question $\mathcal{B}$ (Physical, Social) $\approx^{?}$ $\mathcal{B}$ (Biochemical)
- Analytical question normal $\otimes$ risk: $\mathcal{B}$ (Physical, Social) $\approx^{?}$ $\mathcal{B}$ (Biochemical)
- Analytical question ? $\uparrow\uparrow$? / $\mathcal{B}$ (*)

normal ⊗ risk: $\mathcal{B}$ (Social, Physical) $\approx^{?}$ $\mathcal{B}$ (Biochemical)

Motivation:



Are there any differences in $\varphi \Rightarrow_{p,\,B} \psi$ between normal and risk ?

Example of difference:

$|\,\text{confidence}_{\text{normal}} - \text{confidence}_{\text{risk}}\,| \geq 0.3$

normal	ψ	$\neg\psi$
φ	a_1	b_1
$\neg\varphi$	c_1	d_1

risk	ψ	$\neg\psi$
φ	a_2	b_2
$\neg\varphi$	c_2	d_2

Condition of interestingness: $\left|\frac{a_1}{a_1+b_1} - \frac{a_2}{a_2+b_2}\right| \geq 0.3 \wedge a_1 \geq 0.3 \wedge a_2 \geq 0.3$

normal ⊗ risk: $\mathcal{B}$ (Social, Physical) $\approx^{?}$ $\mathcal{B}$ (Biochemical)

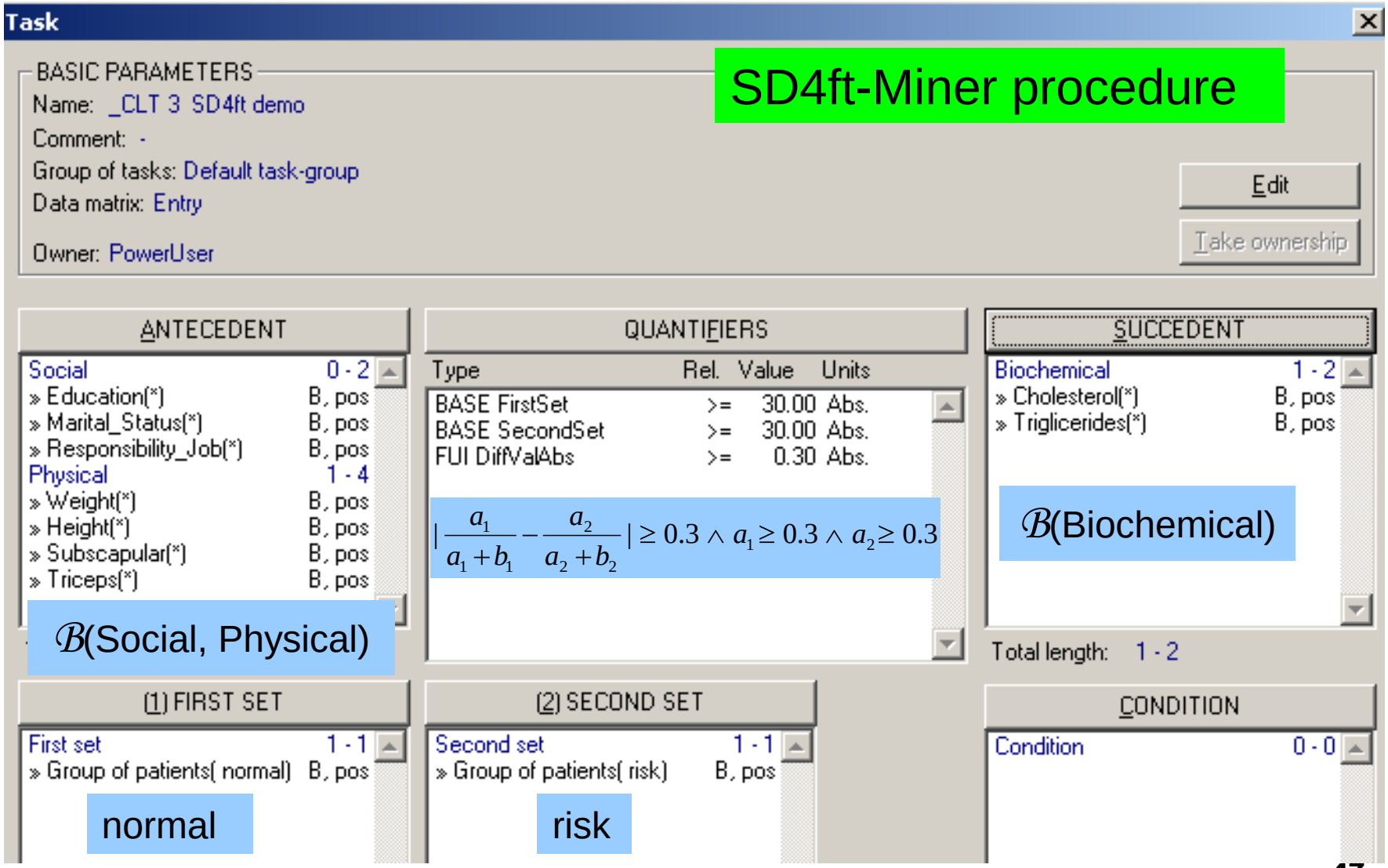
Pattern normal ⊗ risk: $\varphi \Rightarrow^{D}_{0.3,30,30} \psi$ is true in data matrix Entry

if it is: $$\left|\frac{a_1}{a_1+b_1}-\frac{a_2}{a_2+b_2}\right| \geq 0.3 \wedge a_1 \geq 0.3 \wedge a_2 \geq 0.3$$

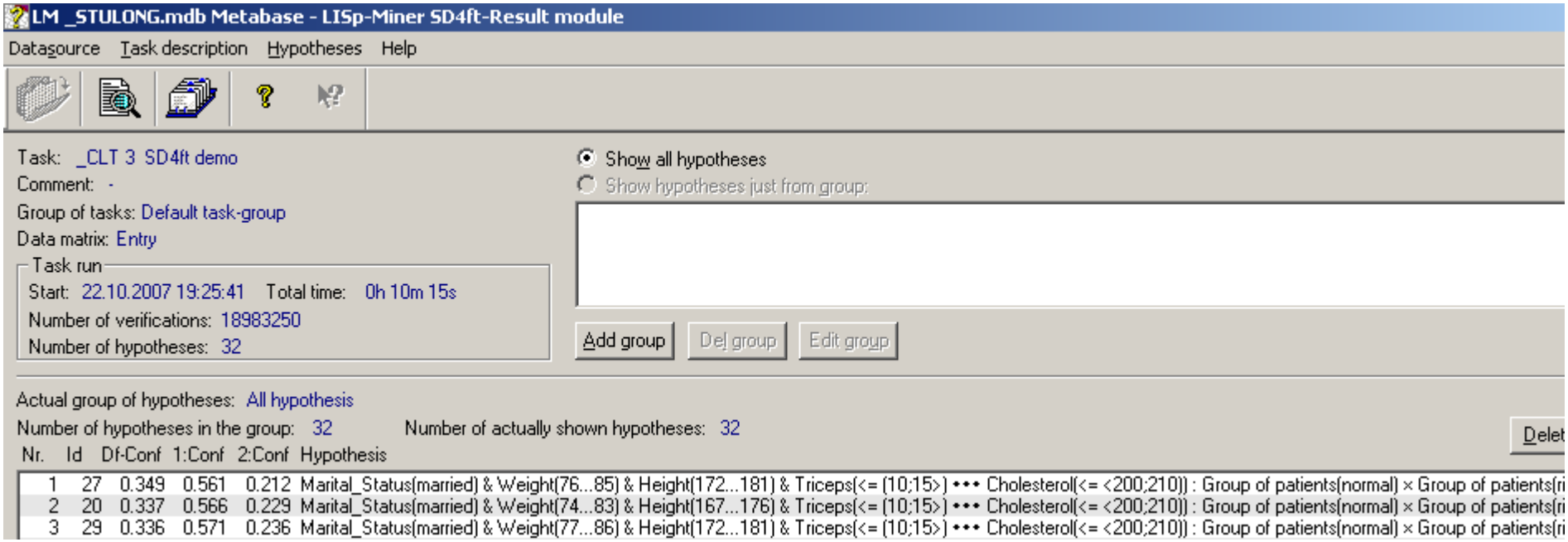
Entry / normal	ψ	$\neg\psi$
φ	a_1	b_1
$\neg\varphi$	c_1	d_1

Entry / risk	ψ	$\neg\psi$
φ	a_2	b_2
$\neg\varphi$	c_2	d_2

Solving normal ⊗ risk: $\mathcal{B}$ (Social, Physical) $\approx^{?}$ $\mathcal{B}$ (Biochemical) (1)



Solving normal $\otimes$ risk: $\mathcal{B}$ (Social, Physical) $\approx^{?}$ $\mathcal{B}$ (Biochemical) (2)



19 000 000 patterns verified in 10 minutes

32 patterns found

The strongest one – see detail

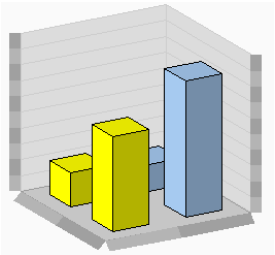
Solving normal ⊗ risk: $\mathcal{B}$ (Social, Physical) $\approx^{?}$ $\mathcal{B}$ (Biochemical) (3)

Detail of results - the strongest pattern

φ: Marital_Status(married) ∧ Weight (75,85⟩ ∧ Height (172,181⟩ ∧ Triceps(≤15)

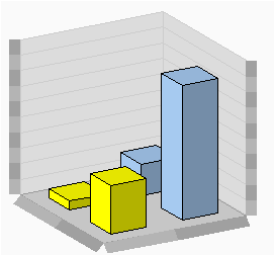
ψ: Cholesterol (≤ 210)

Entry / normal	ψ	¬ψ
φ	*32*	*25*
¬φ	*90*	*129*



confidence$_{normal}$ = 0.56

Entry / risk	ψ	¬ψ
φ	*32*	*119*
¬φ	*188*	*520*



confidence$_{risk}$ = 0.21

confidence$_{normal}$ – confidence$_{risk}$ = 0.35

normal ⊗ risk: $\varphi \Rightarrow^{D}_{0.35,32,32} \psi$

SD4ft-Miner Summary

- mines patterns $\mathcal{B}(\alpha^*) \otimes \mathcal{B}(\beta^*)$: $\mathcal{B}(\varphi^*) \approx \mathcal{B}(\psi^*) / \mathcal{B}(\chi^*)$
- Interpretation: *Are there any differences between some sets α and β what concerns relation of some φ and ψ when some of conditions χ is satisfied?*
- based on the same principles as 4ft-Miner
  - definitions of $\mathcal{B}(*)$
  - measures of association on $\langle a, b, c, d \rangle$
- powerful tool, requires careful applications
- necessity to use background knowledge

LISp-Miner, application examples

Goals of analysis:

- find new knowledge
- verify if the given knowledge observed in given data

Presented:

- Stulong data set
- Analytical question $\mathcal{B}$ (Physical, Social) $\approx^{?}$ $\mathcal{B}$ (Biochemical)
- Analytical question normal $\otimes$ risk: $\mathcal{B}$ (Physical, Social) $\approx^{?}$ $\mathcal{B}$ (Biochemical)
- Analytical question ? $\uparrow\uparrow$? / $\mathcal{B}$ (*)

KL-Miner

GUHA Method and the LISp-Miner System

- GUHA method and KDD
- LISp-Miner system
  - Overview
  - 6 GUHA procedures
  - application examples
  - bit string approach

Bit string

- Current research projects
- Observational calculus – logical calculus of KDD patterns

GUHA Method and the LISp-Miner System

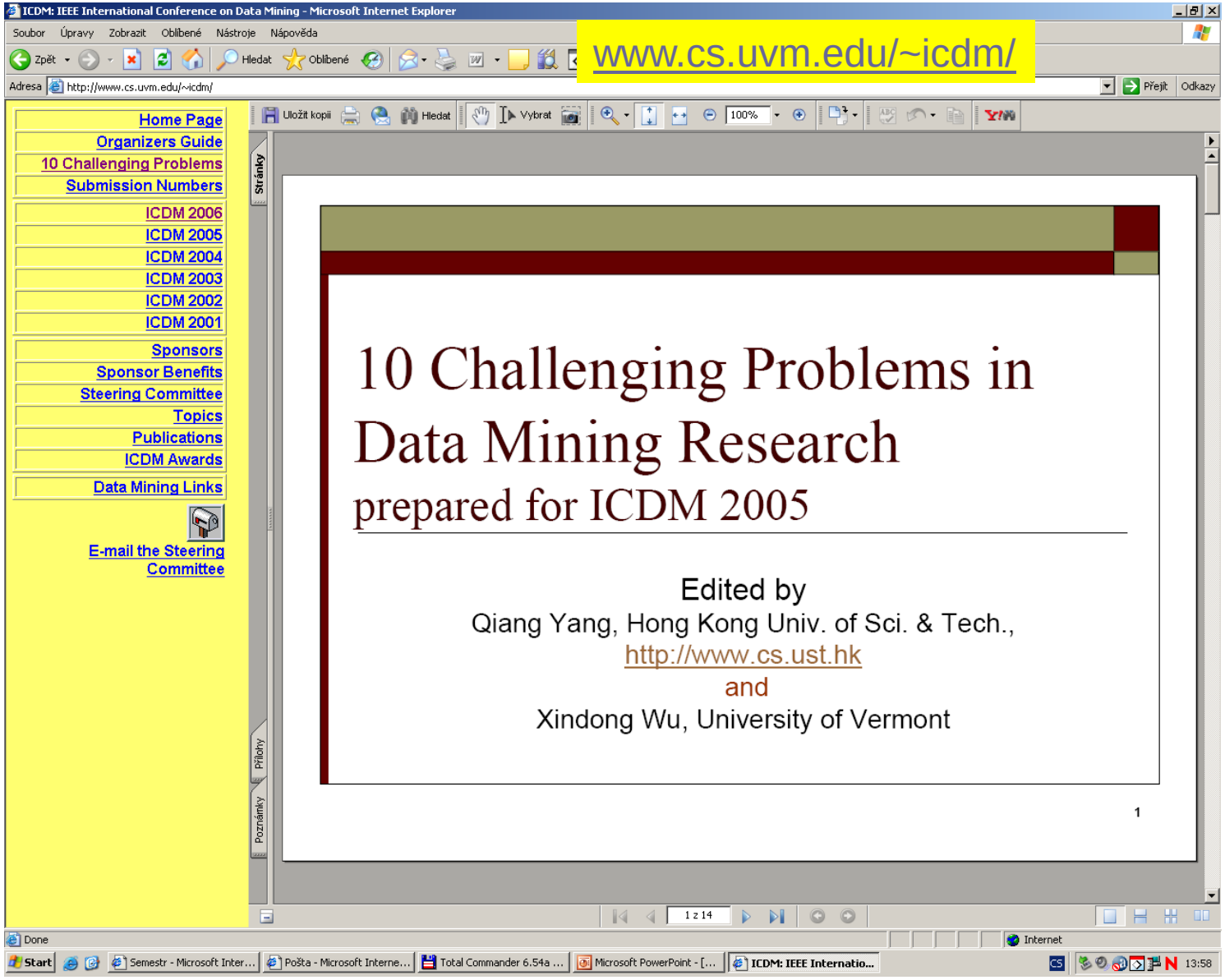
- GUHA method and KDD
- LISp-Miner system
  - Overview
  - 6 GUHA procedures
  - application examples
  - bit string approach
- Current research projects
- Observational calculus – logical calculus of KDD patterns

Current research projects

Starting points:

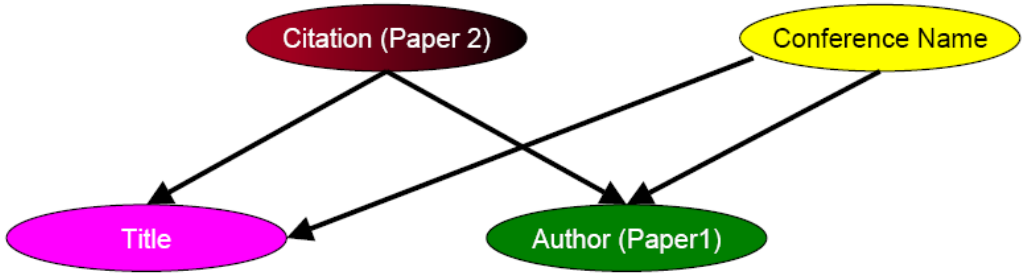
- Hard to use all fine possibilities of particular *-Miners even for specialists
  - ➔ Automated formulation of analytical questions
  - ➔ Project EverMiner
- Particular results of particular *-Miners not too much useful
  - ➔ Automated production of analytical reports
- Necessity to use background knowledge in all phases of data mining
  - ➔ Project LISp-Miner Knowledge Base
- Similar data bases produced and mined in various places
  - ➔ Project SEWEBAR
- See also 10 Challenging Problems in Data Mining Research

10 Challenging Problems in Data Mining Research



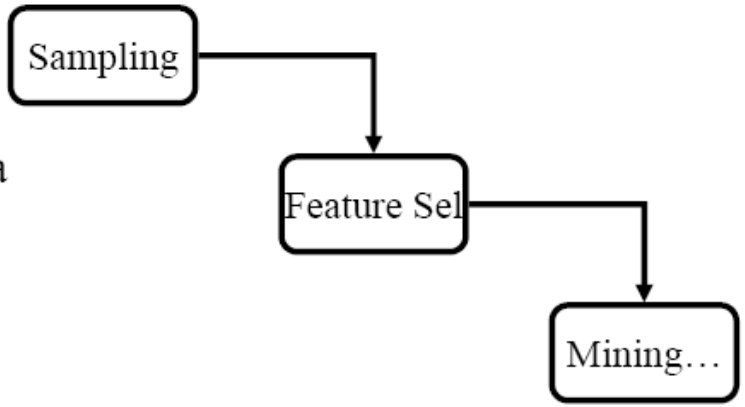
4. Mining Complex Knowledge from Complex Data

- Mining graphs
- Data that are not i.i.d. (independent and identically distributed)
  - many objects are not independent of each other, and are not of a single type.
  - mine the rich structure of relations among objects,
  - E.g.: interlinked Web pages, social networks, metabolic networks in the cell
- Integration of data mining and knowledge inference
  - The biggest gap: unable to relate the results of mining to the real-world decisions they affect - all they can do is hand the results back to the user.
- More research on *interestingness of* knowledge



8. Data-mining-Process Related Problems

- How to automate mining process?
  - the composition of data mining operations
  - Data cleaning, with logging capabilities
  - Visualization and mining automation

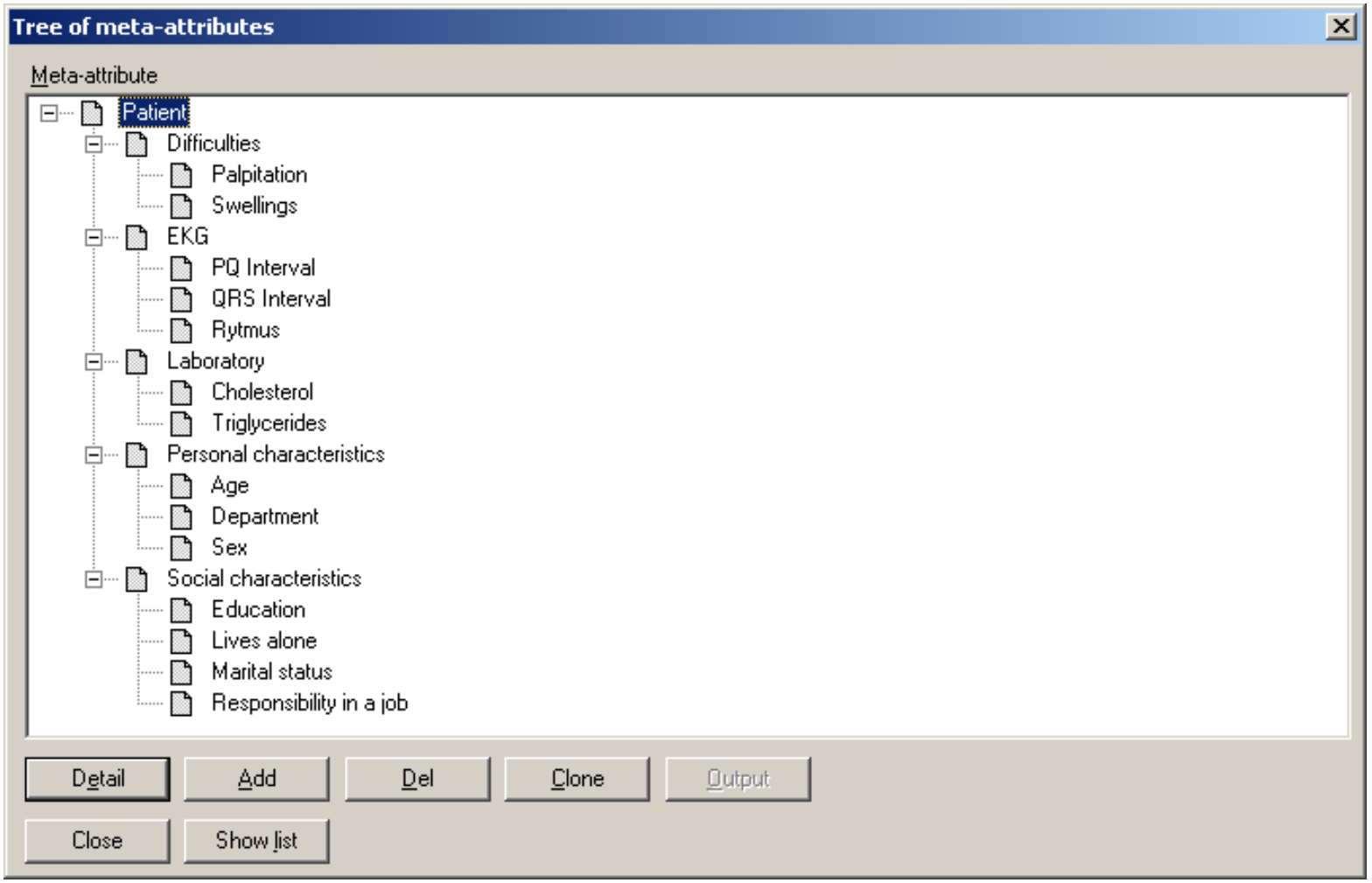


- Need a methodology: help users avoid many data mining mistakes
  - What is a canonical set of data mining operations?

11

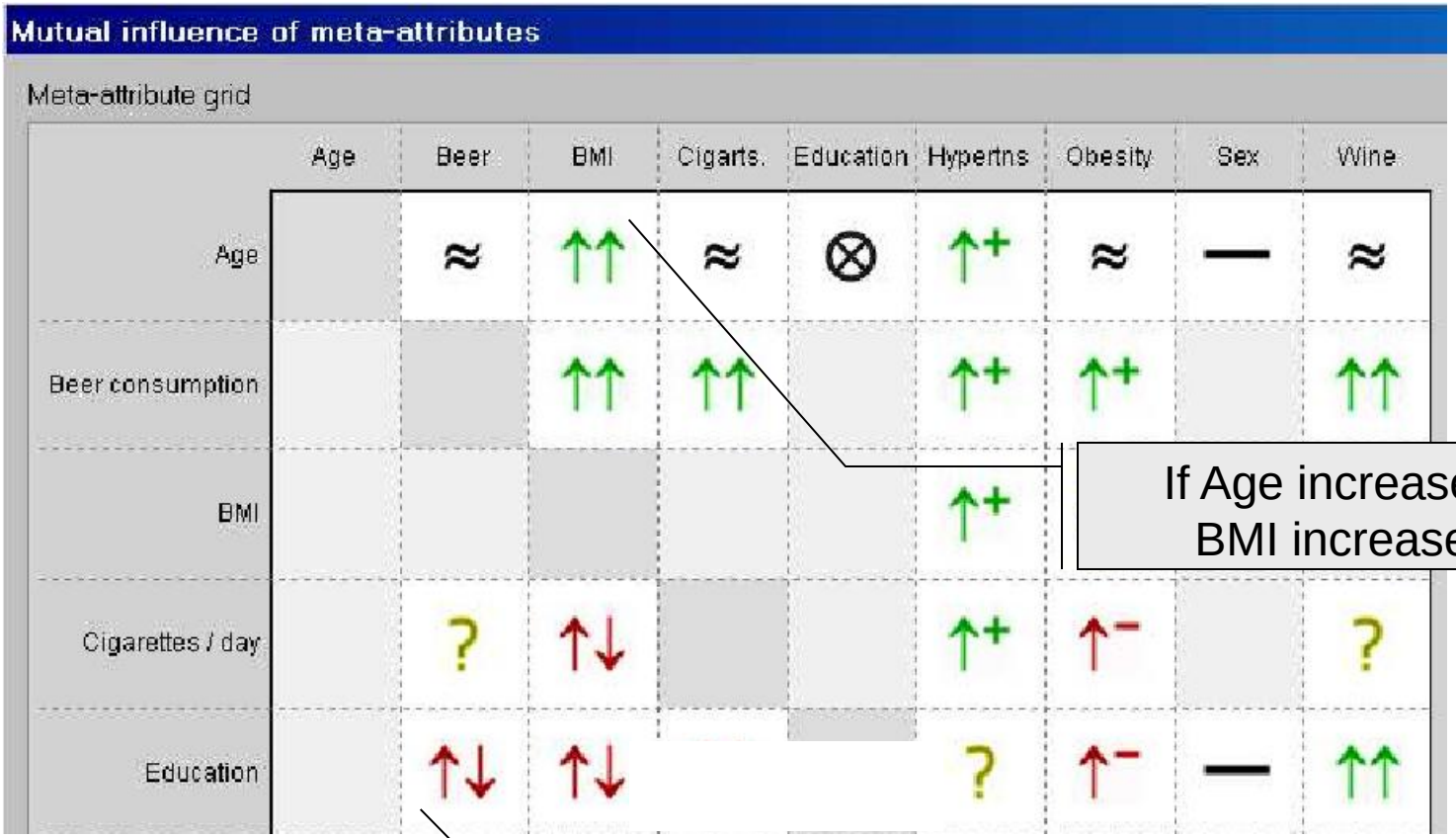
Project LISp-Miner Knowledge Base

Storing and maintaining groups of attributes:



Project LISp-Miner Knowledge Base

Mutual influence of attributes



Analytical report from data mining

- An attempt to bridge the gap between data mining and the real world
- Presents answer to a given analytical question
- Arranged according to
  - user's needs
  - analyzed problem
  - background knowledge
  - analyzed data

Sketch of analytical report for $\mathcal{B}$ [EKG] $\approx^{?}\mathcal{B}$ [Difficulties]

1. Introduction

Informal formulation of analytical question. Description of structure of the report.

2. Analysed data

Overview of basic statistics of used attributes

3. Answering $\mathcal{B}$ [EKG] $\approx^{?}\mathcal{B}$ [Difficulties]

Explanation of $\mathcal{B}$ [EKG], $\mathcal{B}$ [Difficulties], $\Rightarrow_{p, Base}$, $\Rightarrow^{+}_{p, Base}$

4. Results overview

Statistics of all found patterns, suitable statistics on particular attributes occurrences.
Assertions of "second order" like „there is no pattern concerning *Cough*“

5. Particular patterns for $\Rightarrow_{p, Base}$

Structured list of patterns with quantifier $\Rightarrow_{p, Base}$

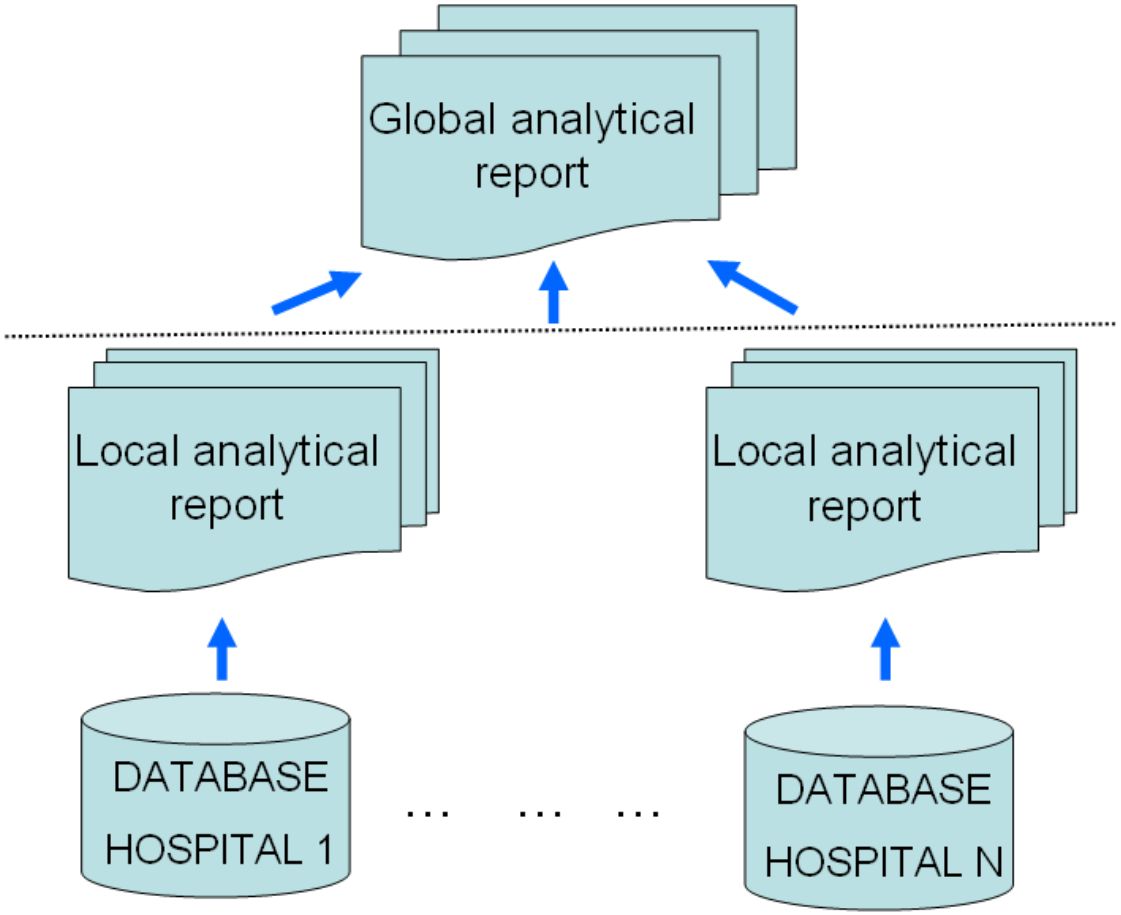
6. Particular patterns for $\Rightarrow^{+}_{p, Base}$

Structured list of patterns with quantifier $\Rightarrow^{+}_{p, Base}$

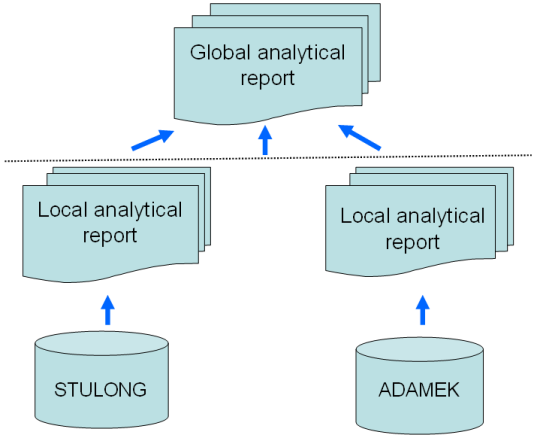
7. Conclusions

Suggestions of additional analysis

Project SEWEBAR – principle (SEmantic WEB and Analytical Reports)



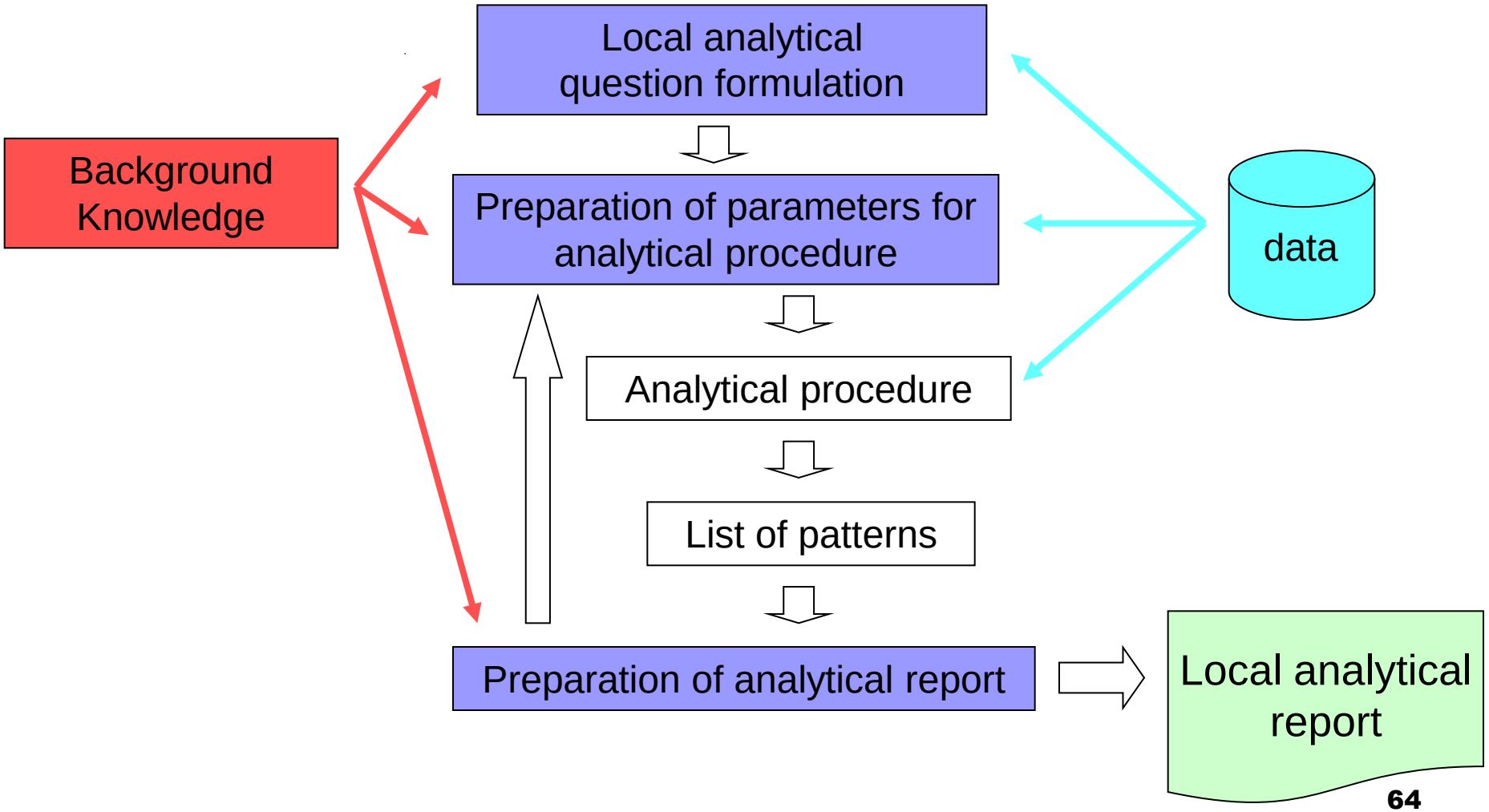
SEWEBAR – pilot project
STULONG and ADAMEK data sets



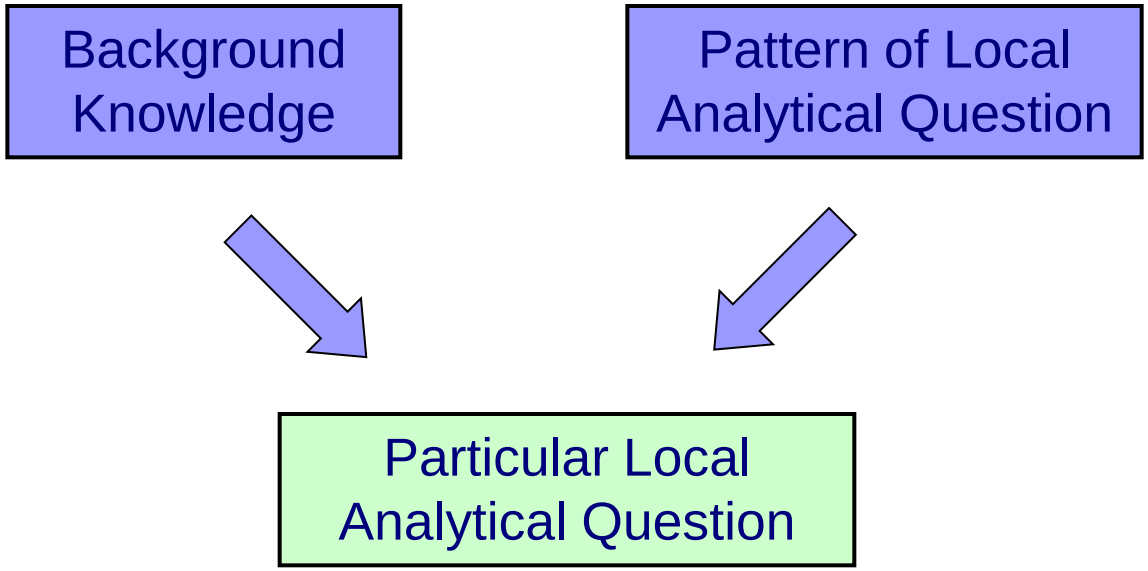
http://sewebar.vse.cz/adamek/index.php/adamek-aslav/72-adamek-aslav-6x4/947-xfarj05-xkoll22-mi-ob-caslav-512

http://sewebar-dev.vse.cz/index.php?option=com_content&view=article&id=46&Itemid=54

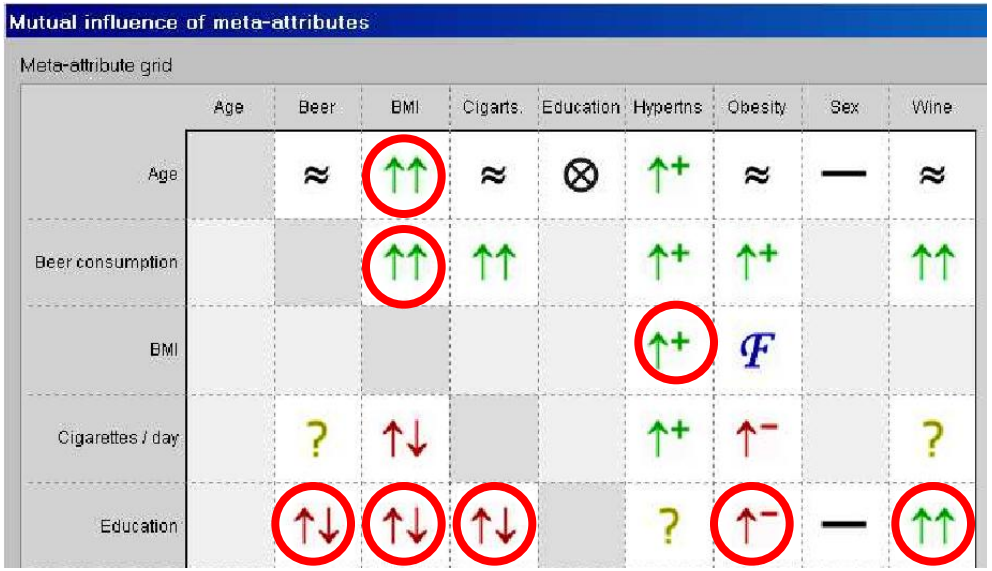
SEWEBAR, 1. step - applying knowledge



Formulating Local Analytical Questions



Formulating Local Analytical Questions – example 1



Is the given item of background knowledge ◯ observed in the ADAMEK data ?

Can be answered using the GUHA procedures implemented in the LISp-Miner system, see e.g.

Rauch J., Šimůnek M: GUHA Method and Granular Computing. In: Hu, X et al (ed.). Proceedings of IEEE conference Granular Computing. 2005, pp. 630-635.

Formulating Local Analytical Questions – example 2

26 Groups of attributes in the ADAMEK data:

Personal information, Family history, Social data, Measures, EKG, Difficulties, ...

What strong relations between Boolean characteristics of two groups X and Y of attributes are observed in the ADAMEK data?

ADAMEK: $\mathcal{B}$ [X] $\approx^{?}$ $\mathcal{B}$ [Y]

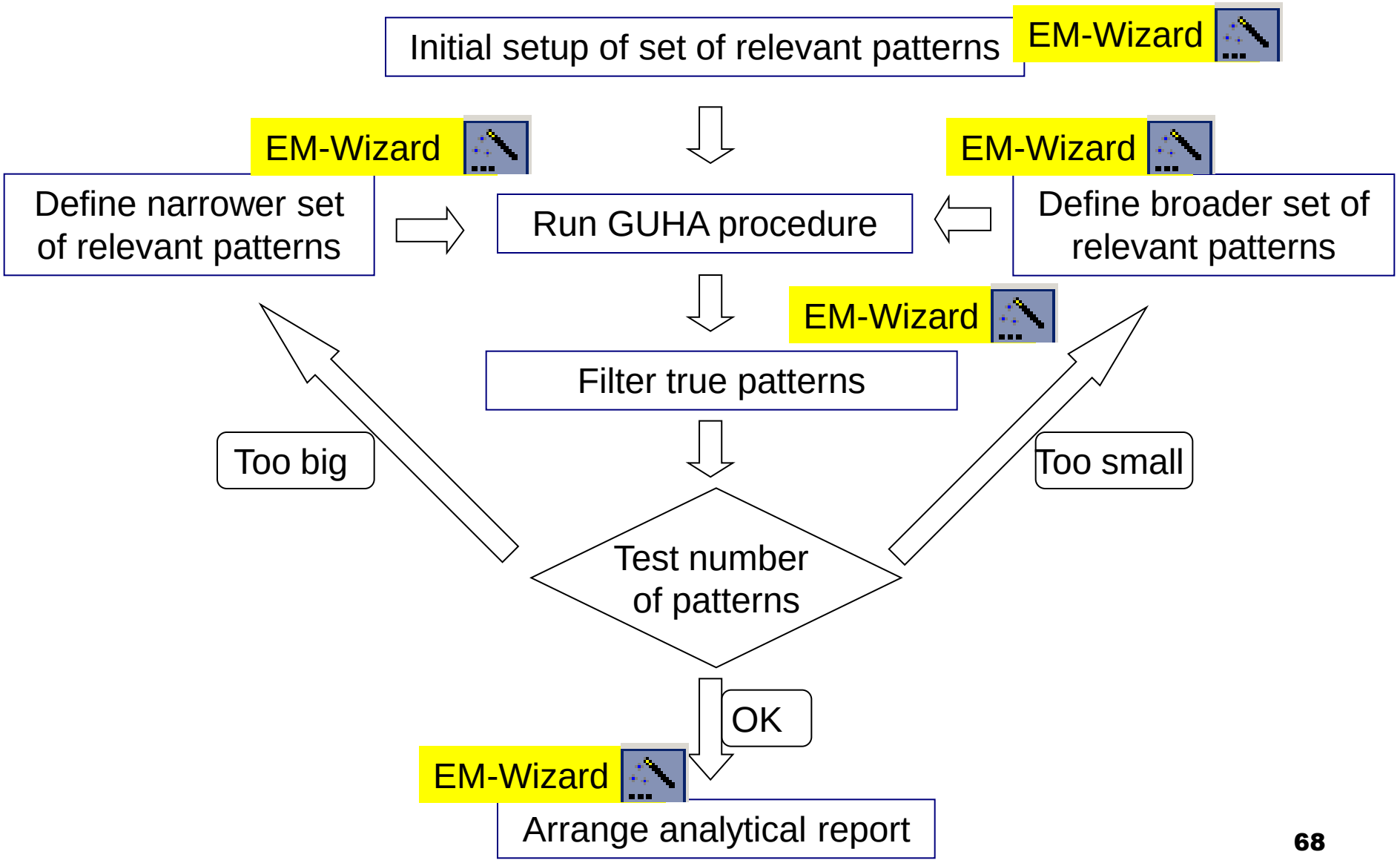
$\mathcal{B}$ [X] – set of all interesting Boolean characteristics of X

$\mathcal{B}$ [Y] – set of all interesting Boolean characteristics of Y

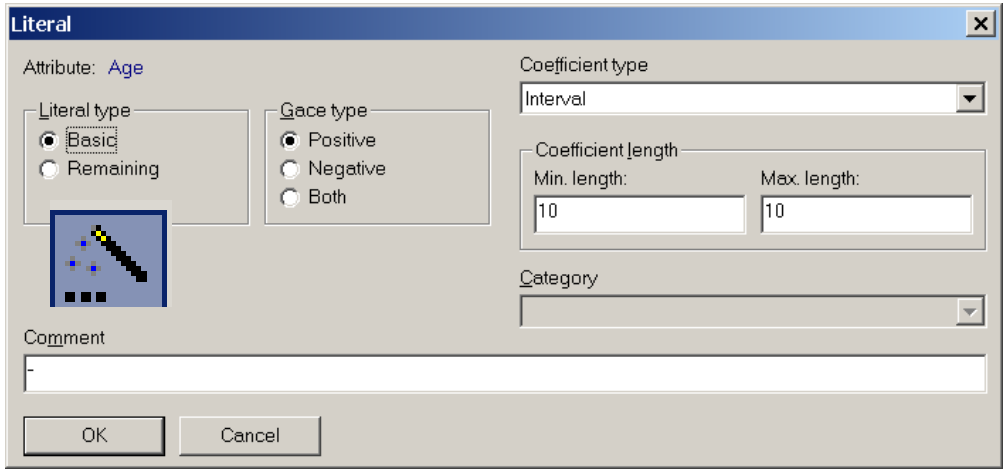
Example:

ADAMEK: $\mathcal{B}$ [EKG] $\approx^{?}$ $\mathcal{B}$ [Difficulties]

Project EverMiner - a vision



Ever Miner - Wizard „Literal“



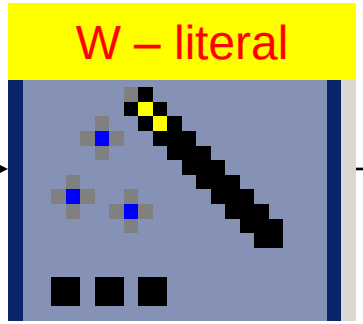
Type of atributes

Number of categories

Frekvencies

Task we solve

….

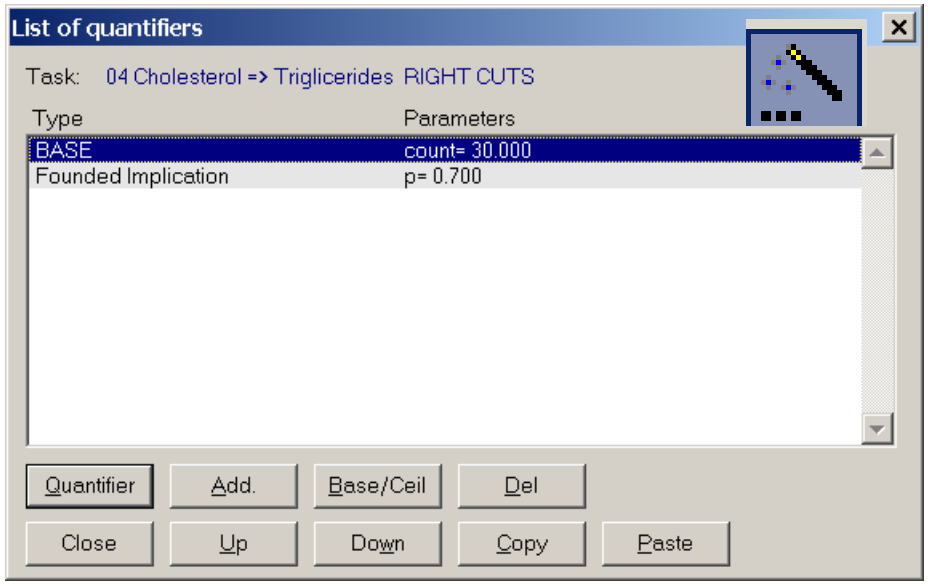


Typ of coefficient

Minimal length

Maximal length

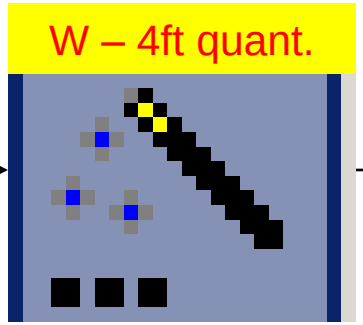
Ever Miner - Wizard „4ft-quantifier “



Solved task

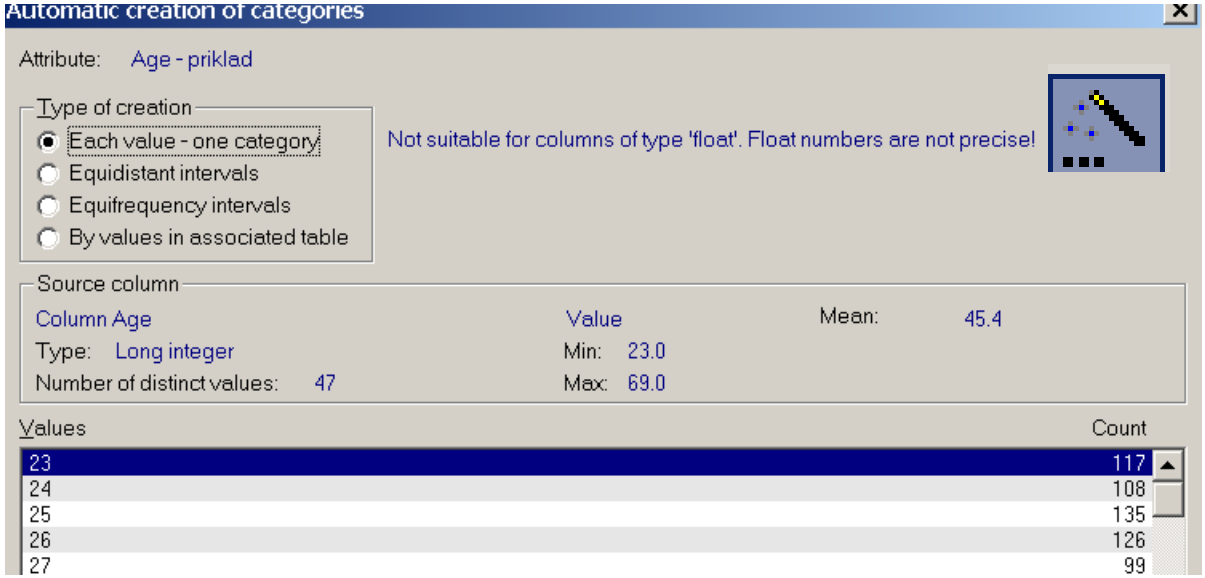
Current results

….



E.g.: $\Rightarrow_{0.95,\ 450}$

Ever Miner - Wizard „categories definition“



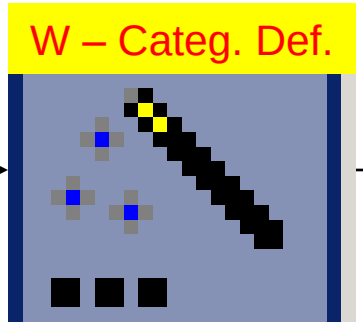
Number of values

Frequencies of values

Task we solve

Current results

….



Typ of categories

Length of intervals

Number of intervals

GUHA Method and the LISp-Miner System

- GUHA method and KDD
- LISp-Miner system
  - Overview
  - 6 GUHA procedures
  - application examples
  - bit string approach
- Current research projects
- Observational calculus – logical calculus of KDD patterns

Observational calculi