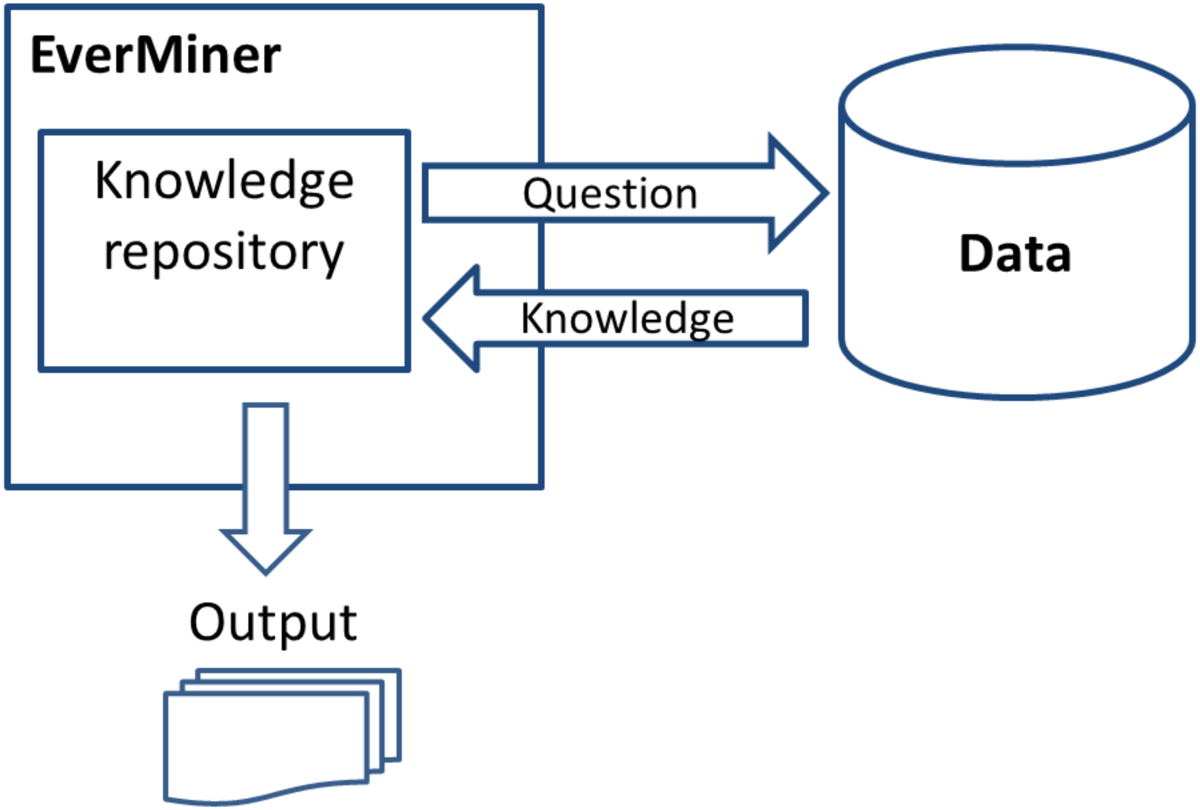
Learning Association Rules from Data through Domain Knowledge and Automation

Jan Rauch, Milan Šimůnek

Faculty of Informatics and Statistics,

University of Economics, Prague

EverMiner – a research idea



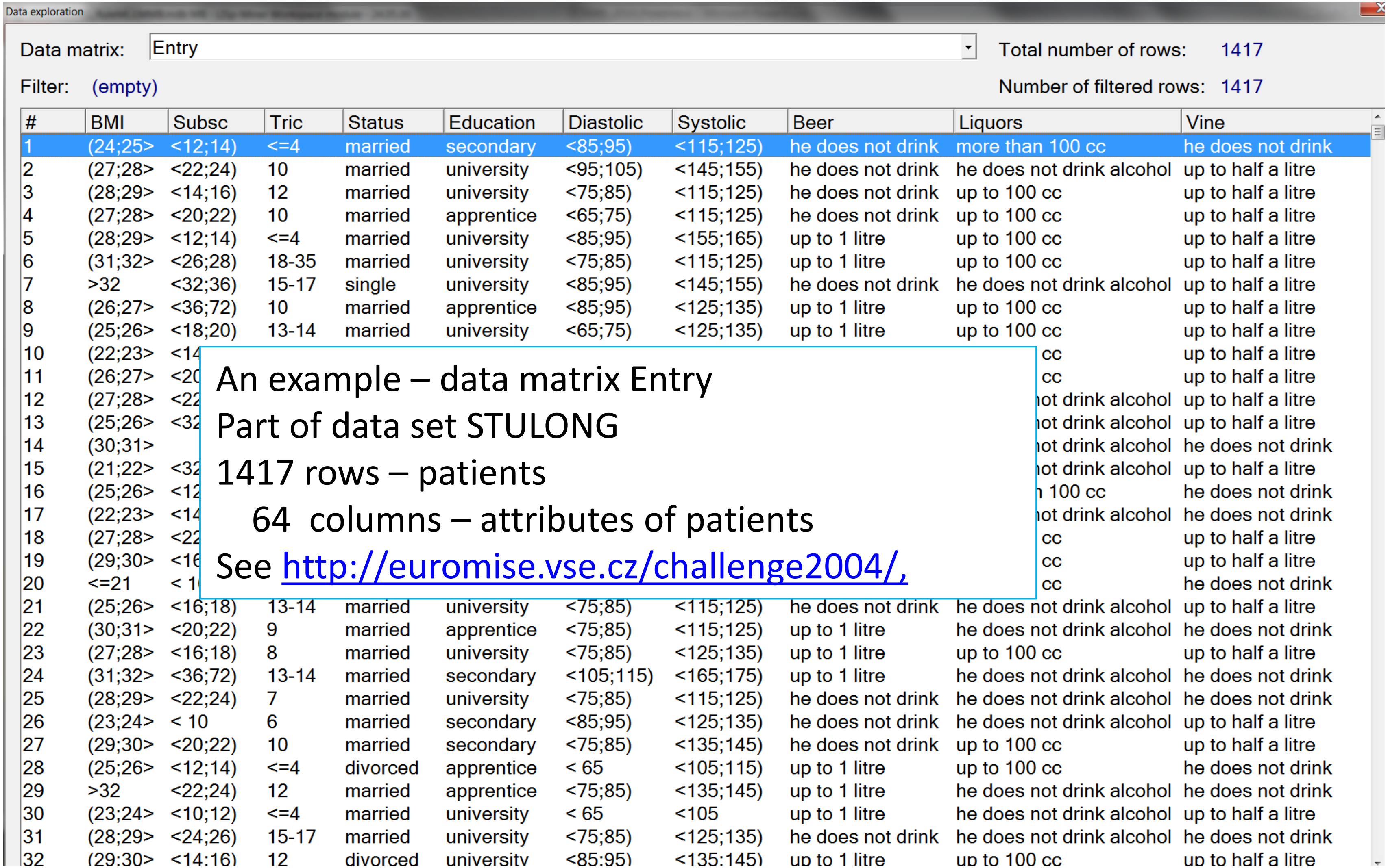
EverMiner – Principles

- Data – data matrix
- Association rules – pairs of Boolean attributes
- General Boolean attributes derived from columns of data matrix
- Domain knowledge not in the form of rules
- Analytical questions formulated using items of domain knowledge
- Set of true rules compared with a set of consequences of items of domain knowledge
- Analytical questions solved by tools of the LISp-Miner system
- The whole process formally described

Laborious process, progress :

- The process described by the LMCL language as an executable program and executed

Analyzed data – Data Matrix



Data matrix: Entry

Total number of rows: 1417

Filter: (empty)

Number of filtered rows: 1417

#	BMI	Subsc	Tric	Status	Education	Diastolic	Systolic	Beer	Liquors	Vine
1	(24;25>	<12;14)	<=4	married	secondary	<85;95)	<115;125)	he does not drink	more than 100 cc	he does not drink
2	(27;28>	<22;24)	10	married	university	<95;105)	<145;155)	he does not drink	he does not drink alcohol	up to half a litre
3	(28;29>	<14;16)	12	married	university	<75;85)	<115;125)	he does not drink	up to 100 cc	up to half a litre
4	(27;28>	<20;22)	10	married	apprentice	<65;75)	<115;125)	he does not drink	up to 100 cc	up to half a litre
5	(28;29>	<12;14)	<=4	married	university	<85;95)	<155;165)	up to 1 litre	up to 100 cc	up to half a litre
6	(31;32>	<26;28)	18-35	married	university	<75;85)	<115;125)	up to 1 litre	up to 100 cc	up to half a litre
7	>32	<32;36)	15-17	single	university	<85;95)	<145;155)	he does not drink	he does not drink alcohol	up to half a litre
8	(26;27>	<36;72)	10	married	apprentice	<85;95)	<125;135)	up to 1 litre	up to 100 cc	up to half a litre
9	(25;26>	<18;20)	13-14	married	university	<65;75)	<125;135)	up to 1 litre	up to 100 cc	up to half a litre
10	(22;23>	<14							cc	up to half a litre
11	(26;27>	<20							cc	up to half a litre
12	(27;28>	<22							ot drink alcohol	up to half a litre
13	(25;26>	<32							ot drink alcohol	up to half a litre
14	(30;31>								ot drink alcohol	he does not drink
15	(21;22>	<32							ot drink alcohol	up to half a litre
16	(25;26>	<12							n 100 cc	he does not drink
17	(22;23>	<14							ot drink alcohol	he does not drink
18	(27;28>	<22							cc	up to half a litre
19	(29;30>	<16							cc	up to half a litre
20	<=21	< 1							cc	he does not drink
21	(25;26>	<16;18)	13-14	married	university	<75;85)	<115;125)	he does not drink	he does not drink alcohol	up to half a litre
22	(30;31>	<20;22)	9	married	apprentice	<75;85)	<115;125)	up to 1 litre	he does not drink alcohol	he does not drink
23	(27;28>	<16;18)	8	married	university	<75;85)	<125;135)	up to 1 litre	up to 100 cc	up to half a litre
24	(31;32>	<36;72)	13-14	married	secondary	<105;115)	<165;175)	up to 1 litre	he does not drink alcohol	he does not drink
25	(28;29>	<22;24)	7	married	university	<75;85)	<115;125)	he does not drink	he does not drink alcohol	he does not drink
26	(23;24>	< 10	6	married	secondary	<85;95)	<125;135)	he does not drink	he does not drink alcohol	up to half a litre
27	(29;30>	<20;22)	10	married	secondary	<75;85)	<135;145)	he does not drink	up to 100 cc	up to half a litre
28	(25;26>	<12;14)	<=4	divorced	apprentice	< 65	<105;115)	up to 1 litre	up to 100 cc	he does not drink
29	>32	<22;24)	12	married	apprentice	<75;85)	<135;145)	up to 1 litre	he does not drink alcohol	he does not drink
30	(23;24>	<10;12)	<=4	married	university	< 65	<105	up to 1 litre	he does not drink alcohol	up to half a litre
31	(28;29>	<24;26)	15-17	married	university	<75;85)	<125;135)	he does not drink	he does not drink alcohol	he does not drink
32	(29;30>	<14;16)	12	divorced	university	<85;95)	<135;145)	up to 1 litre	up to 100 cc	up to half a litre

An example – data matrix Entry

Part of data set STULONG

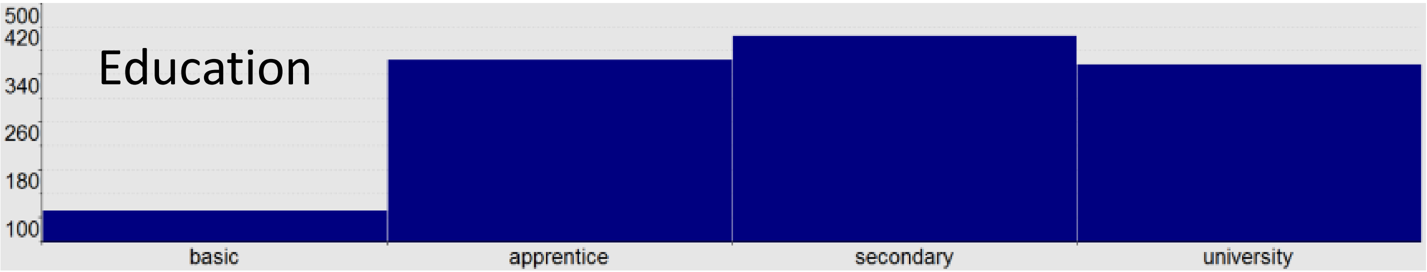
1417 rows – patients

64 columns – attributes of patients

See http://euromise.vse.cz/challenge2004/,

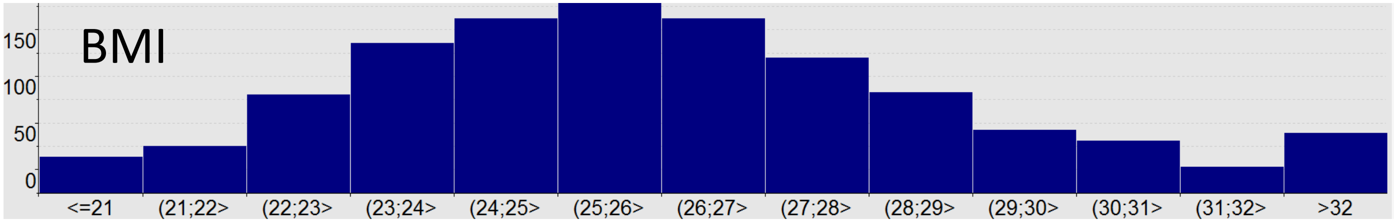
Boolean attributes - examples

Boolean attribute A(α) … α is a subset of values of attribute A



Education(basic)

Education(secondary, university)



BMI(<=21, (21;22〉, (22;23〉) …. BMI(<=23)

BMI((21;22〉, (22;23〉, (23;24〉) …. BMI(21; 24〉

Association rule – pair of Boolean attributes

BMI(21; 24⟩ ∧ Education(secondary, university) $\Rightarrow_{0.9,30}$ Diastolic ⟨65;95)

Entry	Diastolic ⟨65;95)	¬Diastolic ⟨65;95)
BMI(21; 24⟩ ∧ Education(secondary, university)	a	b
¬(BMI(21; 24⟩ ∧ Education(secondary, university))	c	d

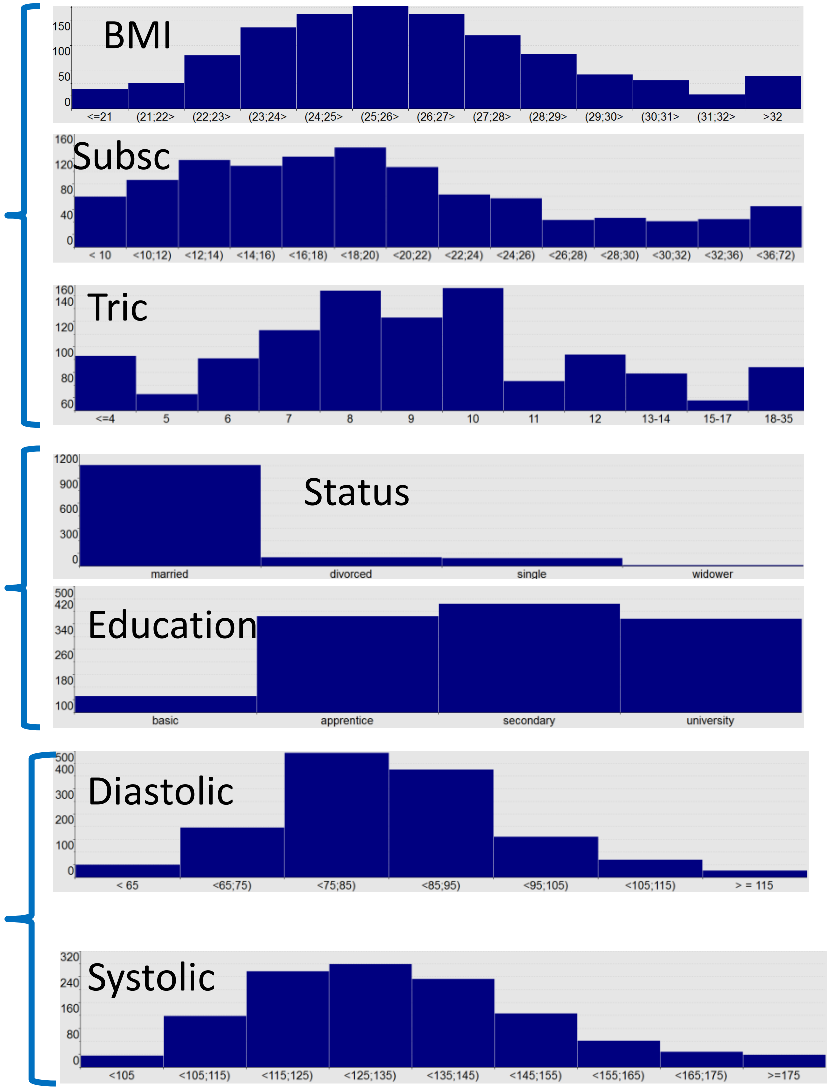
Association rule

BMI(21; 24⟩ ∧ Education(secondary, university) $\Rightarrow_{0.9,30}$ Diastolic ⟨65;95)

is true in data matrix Entry if $\frac{a}{a+b} \geq 0.9 \wedge a \geq 30$

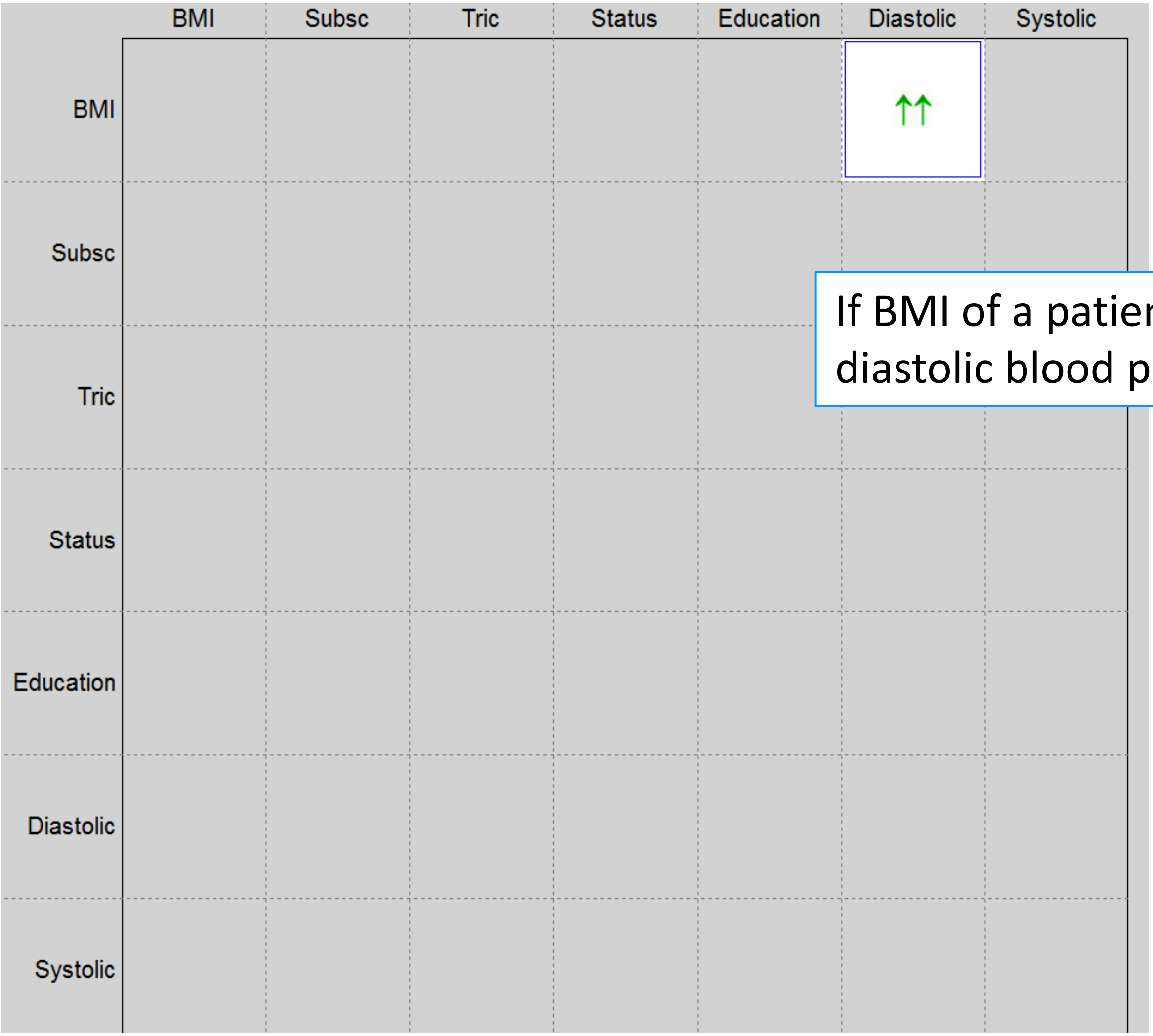
Domain knowledge - Groups of attributes

- Measures
- Personal
- Blood pressure



- Smoking
- Alcohol
- Biochemical
- ....

Domain knowledge - mutual influence of attributes



	BMI	Subsc	Tric	Status	Education	Diastolic	Systolic
BMI						↑↑	
Subsc							
Tric							
Status							
Education							
Diastolic							
Systolic							

BMI ↑↑ Diastolic

If BMI of a patient increases, then his/here diastolic blood pressure increases too.

Analytical questions based on items of domain knowledge

Are there any interesting relations between attributes from group Measures and attributes from group Blood pressure in data matrix Entry? Attributes from group Measures can be combined with attributes from group Personal. Interesting relation is a relation which is strong enough and which is not a consequence of a known dependency BMI ↑↑ Diastolic.

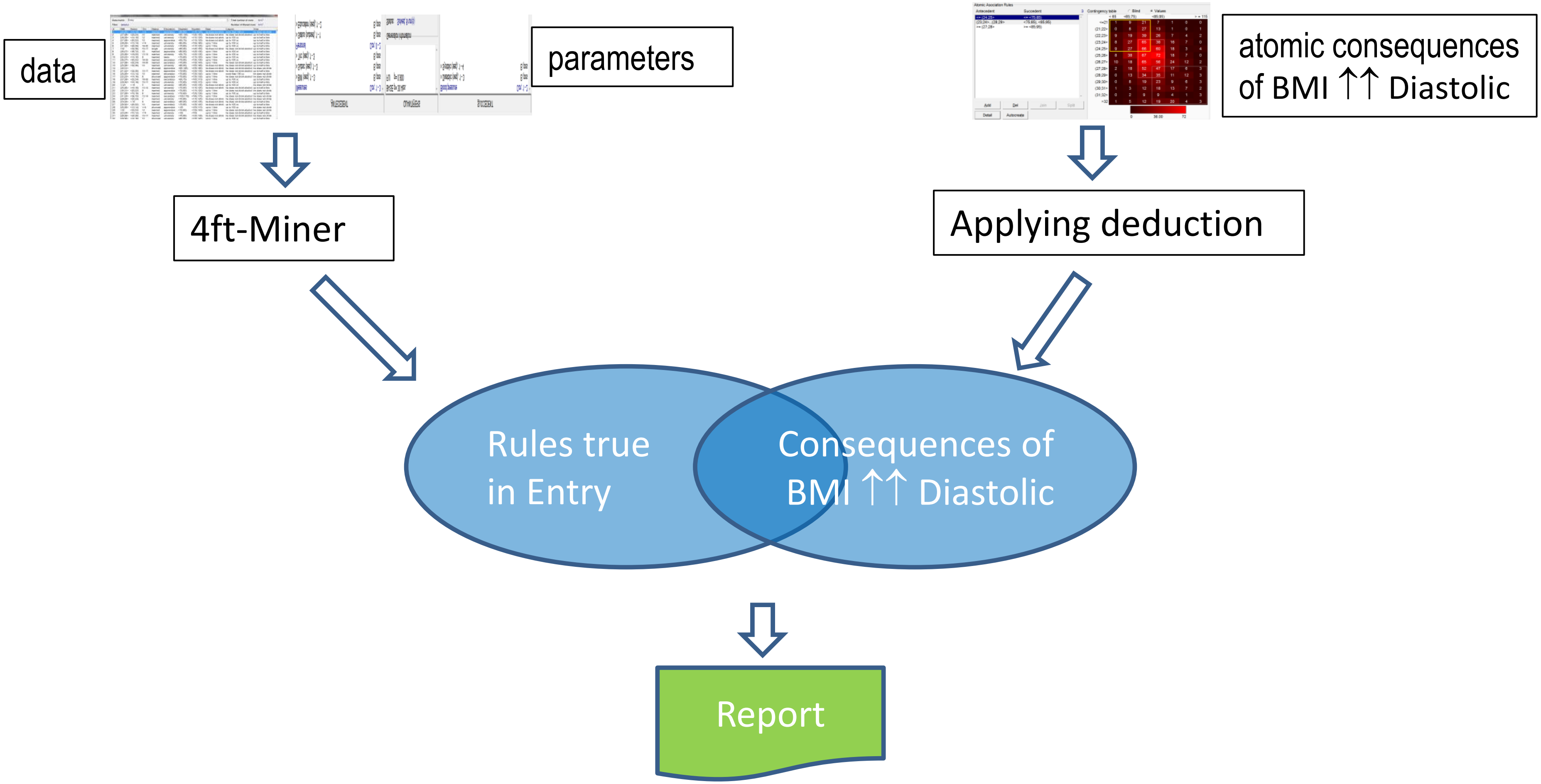
We use asociation rules – pairs of related general Boolean attributes, thus we convert our question to a question concerning association rules:

Entry : (BMI ↑↑ Diastolic) $\not\rightarrow$ $\mathcal{B}$(Measures), $\mathcal{B}$(Personal) ≈ $\mathcal{B}$(Blood pressure)

$\mathcal{B}$(Measures) – a set of Boolean attributes derived from attributes of group Measures
$\mathcal{B}$(Personal) – …
$\mathcal{B}$ (Blood pressure) – …

Set of true rules compared with a set of consequences of items of domain knowledge

Entry : (BMI ↑↑ Diastolic) $\not\rightarrow$ $\mathcal{B}$(Measures), $\mathcal{B}$(Personal) $\Rightarrow_{0.9,30}$ $\mathcal{B}$ (Blood pressure))



data

parameters

4ft-Miner

atomic consequences of BMI ↑↑ Diastolic

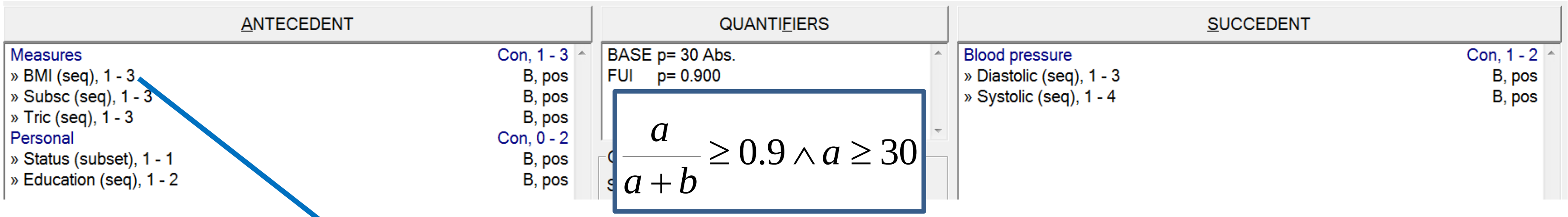
Applying deduction

Rules true in Entry

Consequences of BMI ↑↑ Diastolic

Report

4ft-Miner – input parameters



ANTECEDENT	
Measures	Con, 1 - 3
» BMI (seq), 1 - 3	B, pos
» Subsc (seq), 1 - 3	B, pos
» Tric (seq), 1 - 3	B, pos
Personal	Con, 0 - 2
» Status (subset), 1 - 1	B, pos
» Education (seq), 1 - 2	B, pos

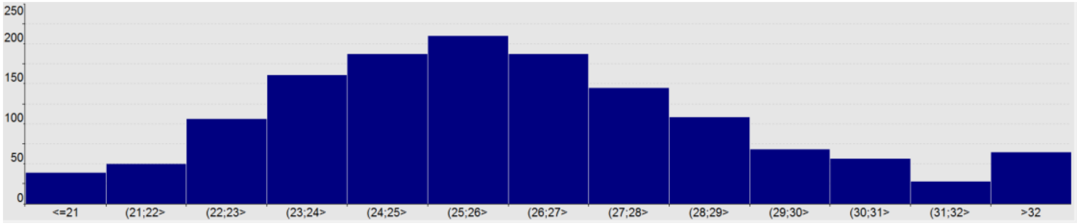
QUANTIFIERS

BASE p= 30 Abs.
FUI p= 0.900

$$\frac{a}{a+b} \geq 0.9 \wedge a \geq 30$$

SUCCEDENT	
Blood pressure	Con, 1 - 2
» Diastolic (seq), 1 - 3	B, pos
» Systolic (seq), 1 - 4	B, pos

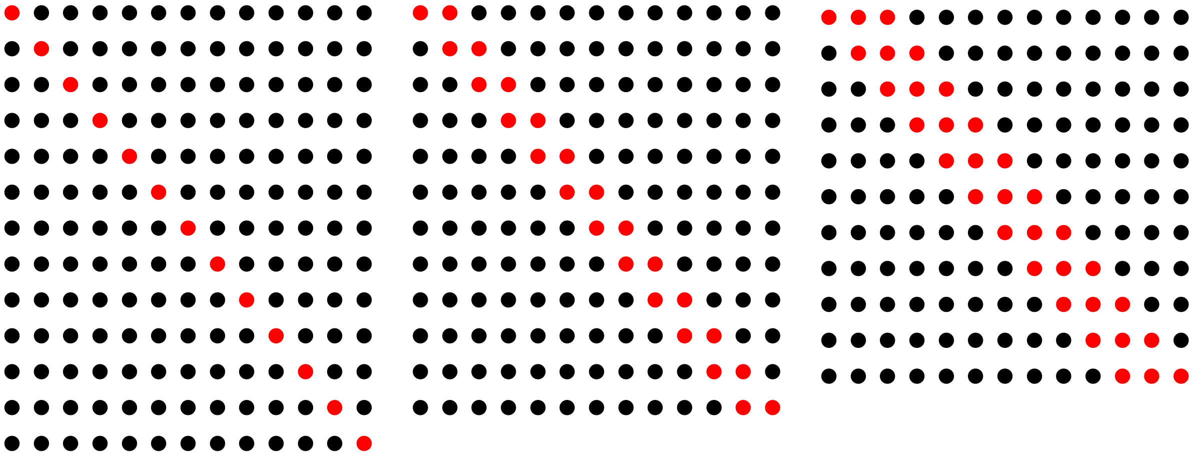
Sequences of categories of BMI, length 1 - 3



Length 1

Length 2

Length 3

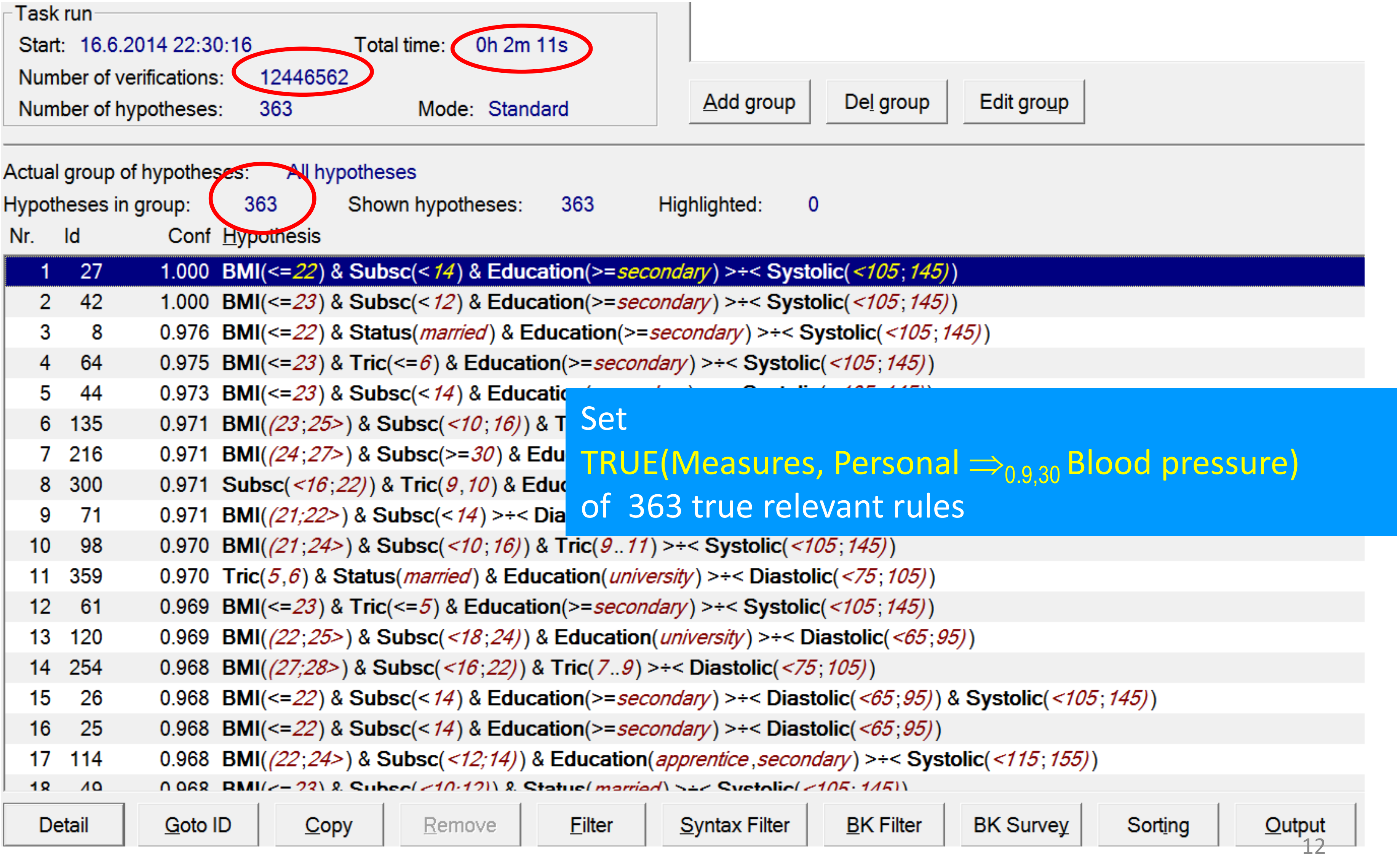


BMI(≤21, (21;22⟩, (22;23⟩) = BMI(≤23)

BMI((21;22⟩, (22;23⟩) = BMI(22;23⟩

13 + 12 + 11 = 36 Boolean attributes

4ft-Miner – output rules



Task run

Start: 16.6.2014 22:30:16 Total time: 0h 2m 11s

Number of verifications: 12446562

Number of hypotheses: 363 Mode: Standard

Actual group of hypotheses: All hypotheses

Hypotheses in group: 363 Shown hypotheses: 363 Highlighted: 0

Nr.	Id	Conf	Hypothesis
1	27	1.000	BMI(<=22) & Subsc(<14) & Education(>=secondary) >÷< Systolic(<105;145))
2	42	1.000	BMI(<=23) & Subsc(<12) & Education(>=secondary) >÷< Systolic(<105;145))
3	8	0.976	BMI(<=22) & Status(married) & Education(>=secondary) >÷< Systolic(<105;145))
4	64	0.975	BMI(<=23) & Tric(<=6) & Education(>=secondary) >÷< Systolic(<105;145))
5	44	0.973	BMI(<=23) & Subsc(<14) & Educatio
6	135	0.971	BMI((23;25>) & Subsc(<10;16)) & T
7	216	0.971	BMI((24;27>) & Subsc(>=30) & Edu
8	300	0.971	Subsc(<16;22)) & Tric(9,10) & Educ
9	71	0.971	BMI((21;22>) & Subsc(<14) >÷< Dia
10	98	0.970	BMI((21;24>) & Subsc(<10;16)) & Tric(9..11) >÷< Systolic(<105;145))
11	359	0.970	Tric(5,6) & Status(married) & Education(university) >÷< Diastolic(<75;105))
12	61	0.969	BMI(<=23) & Tric(<=5) & Education(>=secondary) >÷< Systolic(<105;145))
13	120	0.969	BMI((22;25>) & Subsc(<18;24)) & Education(university) >÷< Diastolic(<65;95))
14	254	0.968	BMI((27;28>) & Subsc(<16;22)) & Tric(7..9) >÷< Diastolic(<75;105))
15	26	0.968	BMI(<=22) & Subsc(<14) & Education(>=secondary) >÷< Diastolic(<65;95)) & Systolic(<105;145))
16	25	0.968	BMI(<=22) & Subsc(<14) & Education(>=secondary) >÷< Diastolic(<65;95))
17	114	0.968	BMI((22;24>) & Subsc(<12;14)) & Education(apprentice,secondary) >÷< Systolic(<115;155))
18	49	0.968	BMI(<=23) & Subsc(<10;12)) & Status(married) >÷< Systolic(<105;145))

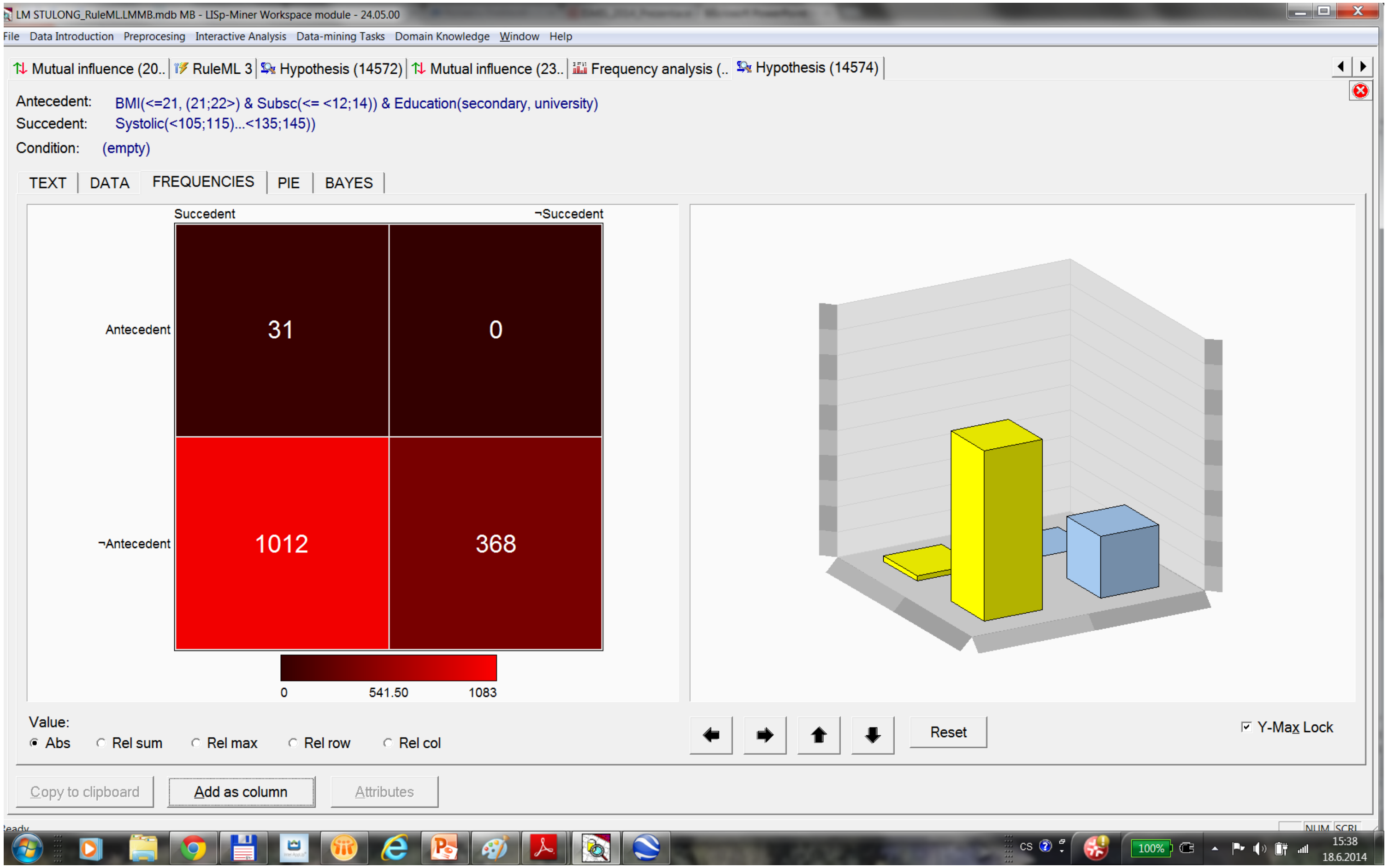
Set

TRUE(Measures, Personal $\Rightarrow_{0.9,30}$ Blood pressure)

of 363 true relevant rules

4ft-Miner – output rule example

$$\mathrm{BMI}(\leq 22) \wedge \mathrm{Subsc}(\leq 14) \wedge \mathrm{Education}(\mathrm{secondary},\ \mathrm{university}) \Rightarrow_{1.0,31} \mathrm{Systolic}\langle 105;145)$$



LM STULONG_RuleML.LMMB.mdb MB - LISp-Miner Workspace module - 24.05.00

Antecedent: BMI(<=21, (21;22>) & Subsc(<= <12;14)) & Education(secondary, university)

Succedent: Systolic(<105;115)...<135;145))

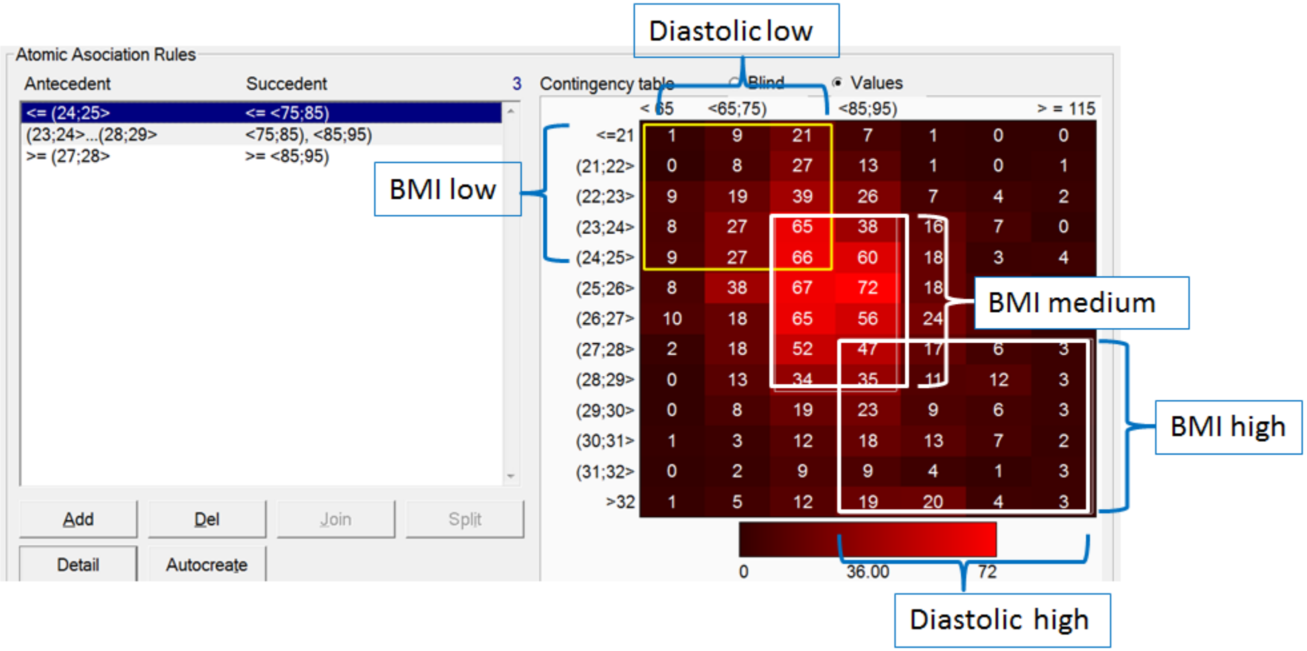
Condition: (empty)

	Succedent	¬Succedent
Antecedent	31	0
¬Antecedent	1012	368

Value:
● Abs ○ Rel sum ○ Rel max ○ Rel row ○ Rel col

☑ Y-Max Lock

Atomic consequences of BMI ↑↑ Diastolic



Atomic Asociation Rules

Antecedent	Succedent
<= (24;25>	<= <75;85)
(23;24>...(28;29>	<75;85), <85;95)
>= (27;28>	>= <85;95)

3 Contingency table ○ Blind ● Values

	< 65	<65;75)		<85;95)			> = 115
<=21	1	9	21	7	1	0	0
(21;22>	0	8	27	13	1	0	1
(22;23>	9	19	39	26	7	4	2
(23;24>	8	27	65	38	16	7	0
(24;25>	9	27	66	60	18	3	4
(25;26>	8	38	67	72	18		
(26;27>	10	18	65	56	24		
(27;28>	2	18	52	47	17	6	3
(28;29>	0	13	34	35	11	12	3
(29;30>	0	8	19	23	9	6	3
(30;31>	1	3	12	18	13	7	2
(31;32>	0	2	9	9	4	1	3
>32	1	5	12	19	20	4	3

low x low : BMI(α) $\Rightarrow_{p,B}$ Diastolic(β) $p \geq 0.9$, $B \geq 30$; $\alpha \in$ BMI low ; $\beta \in$ Diastolic low

medium x medium: …

high x high: …

AC(BMI ↑↑ Diastolic, $\Rightarrow_{0.9,30}$) = low x low $\cup$ medium x medium $\cup$ high x high

Consequences of BMI $\uparrow\uparrow$ Diastolic

Agreed consequences AgC(BMI $\uparrow\uparrow$ Diastolic, $\Rightarrow_{0.9,30}$) : all $\varphi \Rightarrow_{p,B} \psi$:

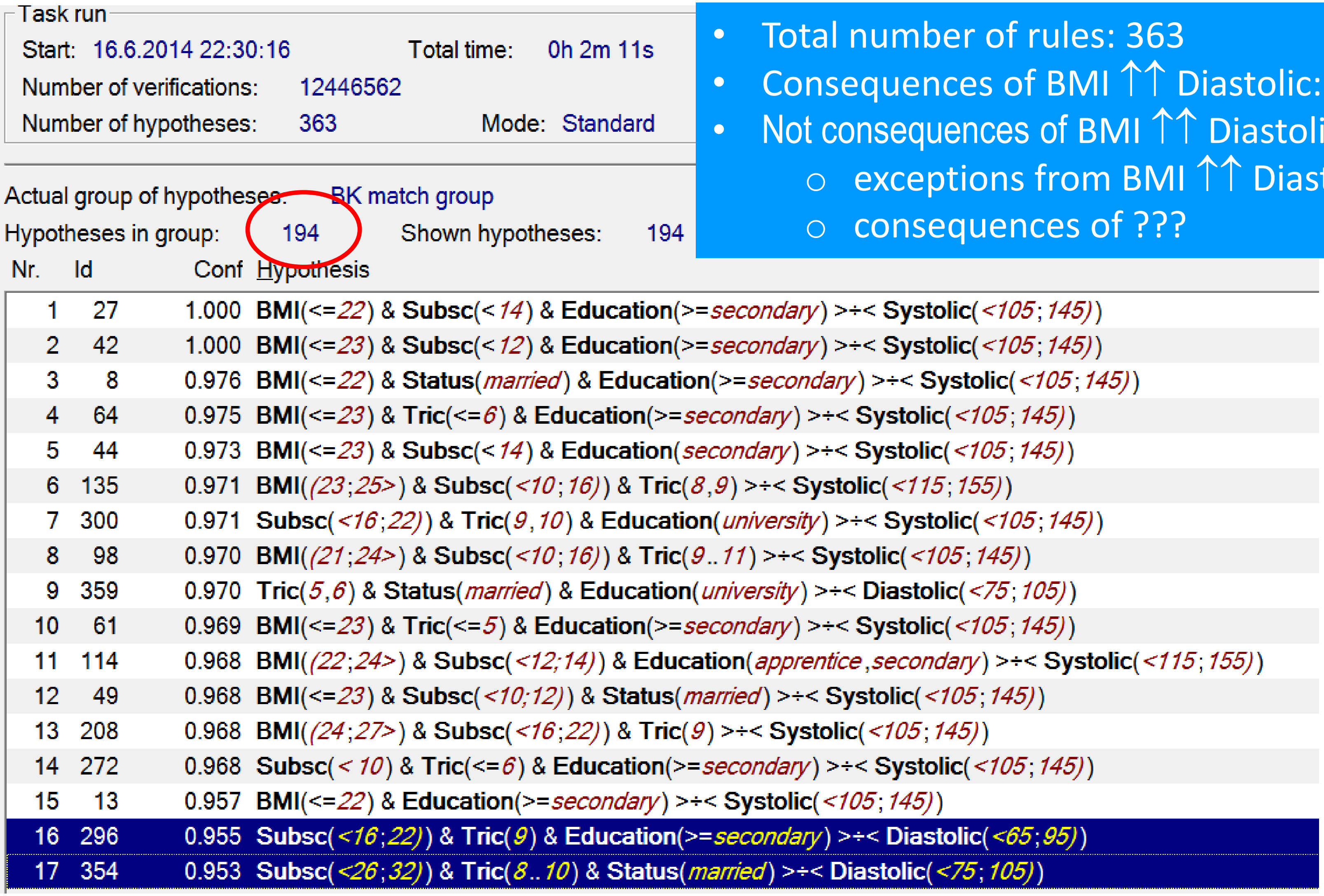
- $\varphi \Rightarrow_{p,B} \psi \notin$ AC(BMI $\uparrow\uparrow$ Diastolic, $\Rightarrow_{0.9,30}$)
- there is $\rho \Rightarrow_{p,B} \sigma \in$ AC(BMI $\uparrow\uparrow$ Diastolic, $\Rightarrow_{0.9,30}$) such that
  - $\varphi \Rightarrow_{p,B} \psi$ does not logically follow from $\rho \Rightarrow_{p,B} \sigma$
  - $\varphi \Rightarrow_{p,B} \psi$ says nothing new new in addition to $\rho \Rightarrow_{p,B} \sigma$
- example: BMI(low) $\wedge$ Education(secondary) $\Rightarrow_{0.9,35}$ Diastolic(low)

Logical consequences LgC(BMI $\uparrow\uparrow$ Diastolic, $\Rightarrow_{0.9,30}$) : all $\varphi \Rightarrow_{p,B} \psi$:

- $\varphi \Rightarrow_{p,B} \psi \notin$ (AC(BMI $\uparrow\uparrow$ Diastolic, $\Rightarrow_{0.9,30}$) $\cup$ AgC(BMI $\uparrow\uparrow$ Diastolic, $\Rightarrow_{0.9,30}$))
- there is $\rho \Rightarrow_{p,B} \sigma \in$ (AC(BMI $\uparrow\uparrow$ Diastolic, $\Rightarrow_{0.9,30}$) $\cup$ AgC(BMI $\uparrow\uparrow$ Diastolic, $\Rightarrow_{0.9,30}$)) such that $\varphi \Rightarrow_{p,B} \psi$ logically follows from $\rho \Rightarrow_{p,B} \sigma$
- example: BMI(low) $\Rightarrow_{1.0,31}$ Diastolic($\langle$75;85),$\langle$85;95))

Cons(BMI $\uparrow\uparrow$ Diastolic, $\Rightarrow_{0.9,30}$) =
AC(BMI $\uparrow\uparrow$ Diastolic, $\Rightarrow_{0.9,30}$) $\cup$ **AgC**(BMI $\uparrow\uparrow$ Diastolic, $\Rightarrow_{0.9,30}$) $\cup$ **LgC**(BMI $\uparrow\uparrow$ Diastolic, $\Rightarrow_{0.9,30}$)

Rules – not consequences of BMI ↑↑ Diastolic



Task run

Start: 16.6.2014 22:30:16 Total time: 0h 2m 11s

Number of verifications: 12446562

Number of hypotheses: 363 Mode: Standard

Actual group of hypotheses: BK match group

Hypotheses in group: 194 Shown hypotheses: 194

Nr.	Id	Conf	Hypothesis
1	27	1.000	BMI(<=22) & Subsc(< 14) & Education(>=secondary) >÷< Systolic(<105;145))
2	42	1.000	BMI(<=23) & Subsc(< 12) & Education(>=secondary) >÷< Systolic(<105;145))
3	8	0.976	BMI(<=22) & Status(married) & Education(>=secondary) >÷< Systolic(<105;145))
4	64	0.975	BMI(<=23) & Tric(<=6) & Education(>=secondary) >÷< Systolic(<105;145))
5	44	0.973	BMI(<=23) & Subsc(< 14) & Education(secondary) >÷< Systolic(<105;145))
6	135	0.971	BMI((23;25>) & Subsc(<10;16)) & Tric(8,9) >÷< Systolic(<115;155))
7	300	0.971	Subsc(<16;22)) & Tric(9,10) & Education(university) >÷< Systolic(<105;145))
8	98	0.970	BMI((21;24>) & Subsc(<10;16)) & Tric(9..11) >÷< Systolic(<105;145))
9	359	0.970	Tric(5,6) & Status(married) & Education(university) >÷< Diastolic(<75;105))
10	61	0.969	BMI(<=23) & Tric(<=5) & Education(>=secondary) >÷< Systolic(<105;145))
11	114	0.968	BMI((22;24>) & Subsc(<12;14)) & Education(apprentice,secondary) >÷< Systolic(<115;155))
12	49	0.968	BMI(<=23) & Subsc(<10;12)) & Status(married) >÷< Systolic(<105;145))
13	208	0.968	BMI((24;27>) & Subsc(<16;22)) & Tric(9) >÷< Systolic(<105;145))
14	272	0.968	Subsc(< 10) & Tric(<=6) & Education(>=secondary) >÷< Systolic(<105;145))
15	13	0.957	BMI(<=22) & Education(>=secondary) >÷< Systolic(<105;145))
16	296	0.955	Subsc(<16;22)) & Tric(9) & Education(>=secondary) >÷< Diastolic(<65;95))
17	354	0.953	Subsc(<26;32)) & Tric(8..10) & Status(married) >÷< Diastolic(<75;105))

- Total number of rules: 363
- Consequences of BMI ↑↑ Diastolic: 169
- Not consequences of BMI ↑↑ Diastolic: 194
    - exceptions from BMI ↑↑ Diastolic?
    - consequences of ???

Additional analytical tasks

Solved task:

TRUE(Measures, Personal $\Rightarrow_{0.9,30}$ Blood pressure) X Cons(BMI ↑↑ Diastolic, $\Rightarrow_{0.9,30}$)

Additional tasks: compare TRUE(Measures, Personal $\Rightarrow_{0.9,30}$ Blood pressure) with

Cons(Subsc ↑↑ Diastolic, $\Rightarrow_{0.9,30}$)

Cons(Tric ↑↑ Diastolic, $\Rightarrow_{0.9,30}$)

Cons(BMI ↑↑ Systolic, $\Rightarrow_{0.9,30}$)

Cons(Subsc ↑↑ Systolic, $\Rightarrow_{0.9,30}$)

Cons(Tric ↑↑ Systolic, $\Rightarrow_{0.9,30}$)

Cons(Tric ↑↓Diastolic, $\Rightarrow_{0.9,30}$)

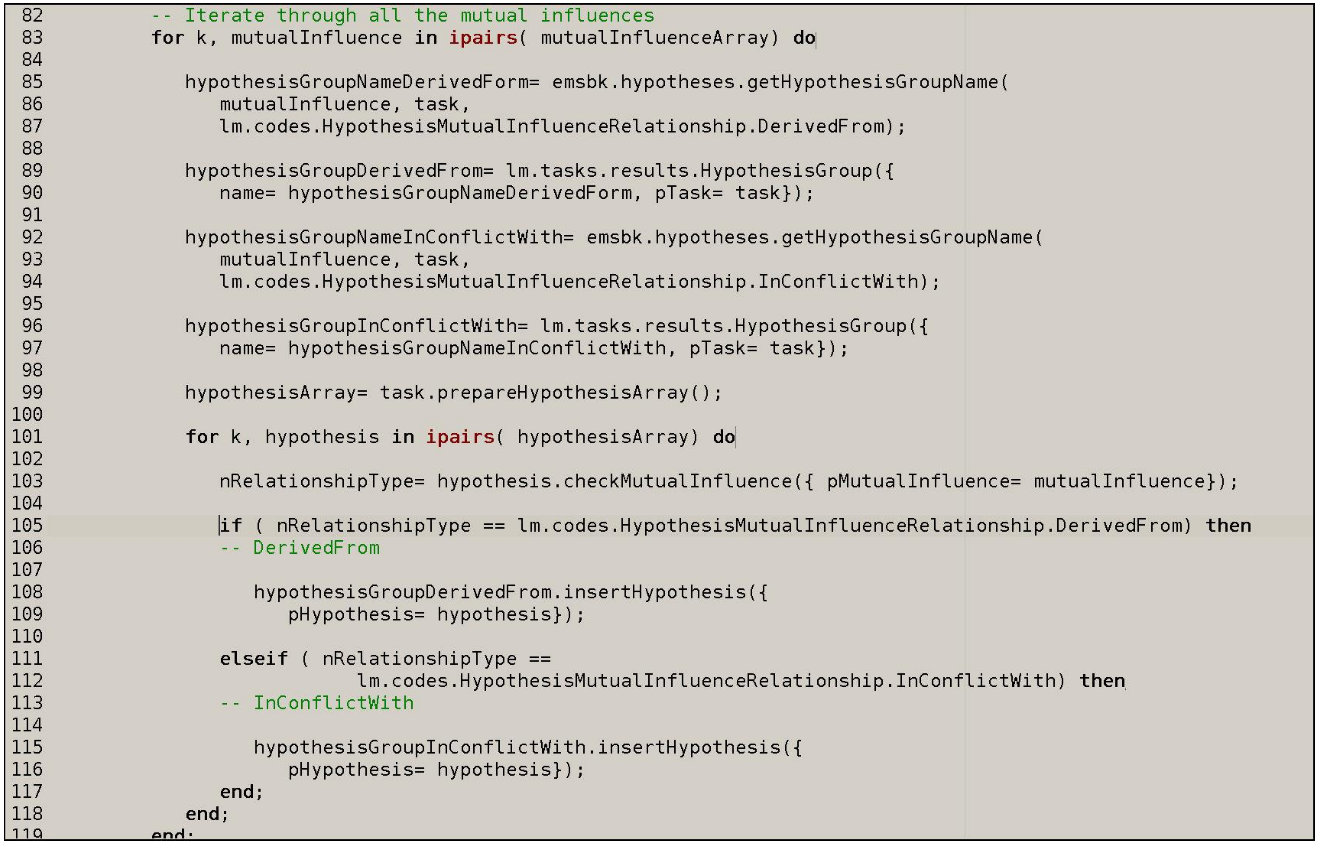
Tric ↑↓ Diastolic … If the skinfold above the musculus triceps increases then diastolic blood pressure decreases.

The same principle, lot of work

Solving additional analytical tasks

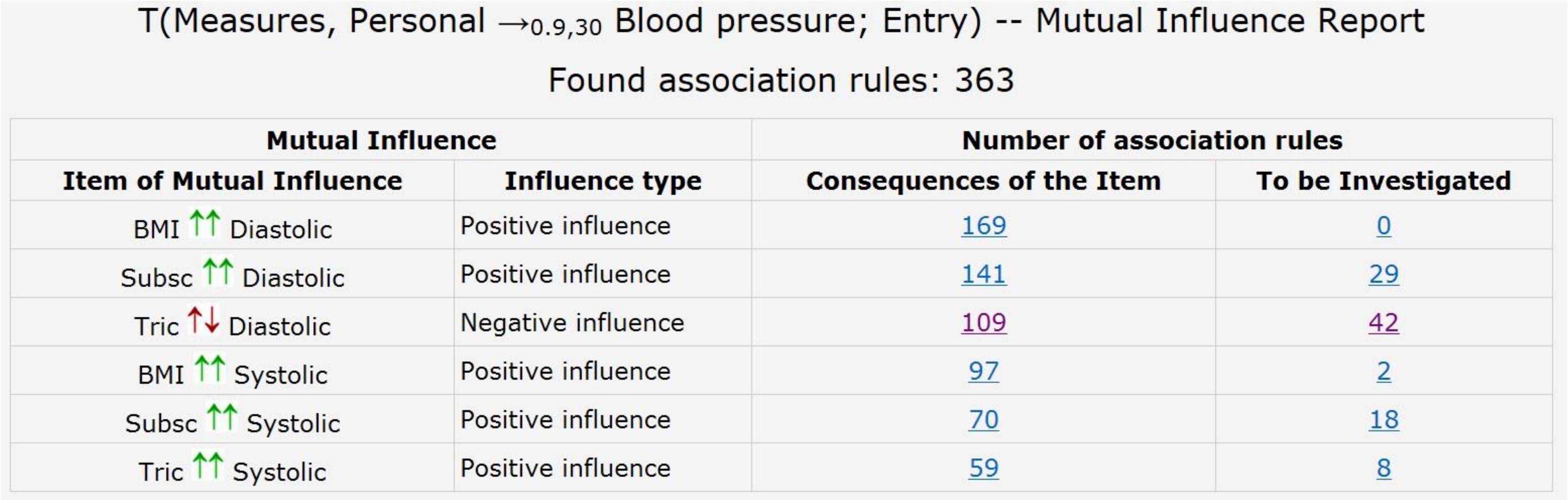
1) Formal description of the solution, formal frame FOFRADAR used
2) Describe the solution in LMCL – LISp-Miner Control Language as an executable program and execute it

LMCL syntax example:



```
-- Iterate through all the mutual influences
for k, mutualInfluence in ipairs( mutualInfluenceArray) do

   hypothesisGroupNameDerivedForm= emsbk.hypotheses.getHypothesisGroupName(
      mutualInfluence, task,
      lm.codes.HypothesisMutualInfluenceRelationship.DerivedFrom);

   hypothesisGroupDerivedFrom= lm.tasks.results.HypothesisGroup({
      name= hypothesisGroupNameDerivedForm, pTask= task});

   hypothesisGroupNameInConflictWith= emsbk.hypotheses.getHypothesisGroupName(
      mutualInfluence, task,
      lm.codes.HypothesisMutualInfluenceRelationship.InConflictWith);

   hypothesisGroupInConflictWith= lm.tasks.results.HypothesisGroup({
      name= hypothesisGroupNameInConflictWith, pTask= task});

   hypothesisArray= task.prepareHypothesisArray();

   for k, hypothesis in ipairs( hypothesisArray) do

      nRelationshipType= hypothesis.checkMutualInfluence({ pMutualInfluence= mutualInfluence});

      if ( nRelationshipType == lm.codes.HypothesisMutualInfluenceRelationship.DerivedFrom) then
      -- DerivedFrom

         hypothesisGroupDerivedFrom.insertHypothesis({
            pHypothesis= hypothesis});

      elseif ( nRelationshipType ==
                  lm.codes.HypothesisMutualInfluenceRelationship.InConflictWith) then
      -- InConflictWith

         hypothesisGroupInConflictWith.insertHypothesis({
            pHypothesis= hypothesis});
      end;
   end;
end;
```

Summary report – one of first experiments



T(Measures, Personal $\rightarrow_{0.9,30}$ Blood pressure; Entry) -- Mutual Influence Report

Found association rules: 363

Mutual Influence		**Number of association rules**	
Item of Mutual Influence	**Influence type**	**Consequences of the Item**	**To be Investigated**
BMI ↑↑ Diastolic	Positive influence	169	0
Subsc ↑↑ Diastolic	Positive influence	141	29
Tric ↑↓ Diastolic	Negative influence	109	42
BMI ↑↑ Systolic	Positive influence	97	2
Subsc ↑↑ Systolic	Positive influence	70	18
Tric ↑↑ Systolic	Positive influence	59	8

A rule is interesting from the point of view of further investigation if it can be considered

- as a conflict (an exception) to the relation of the mutual influence in question.
- as an indication of an additional relation of mutual relation (not used here).

Conclusions and further work

- LISp-Miner Control Language approved as a powerful tool

- **Lot of additional experiments necessary**

- Automatic generation of additional analytical questions
  - all groups of attributes
  - additional types of mutual relations

- Application of additional procedures dealing with additional types of patterns
  - couples of association rules
  - action rules
  - patterns based on contingency tables

Thank you