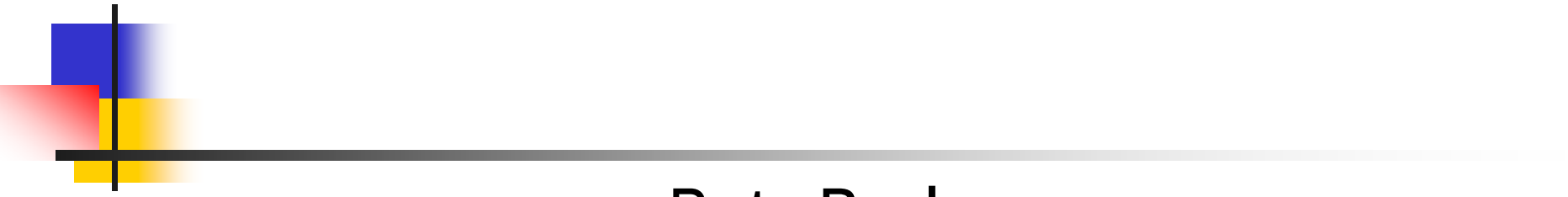
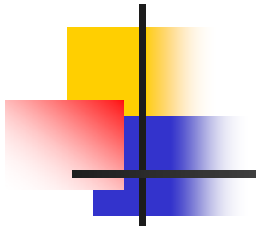
Mining for Association Meta-Rules



Petr Berka
University of Economics
Prague
berka@vse.cz



Outline

- Association rules
- Association meta-rules
  - Typology
  - Running example
  - Further experiments

Association rules (1/3)

- Market basket analysis

Expressions:

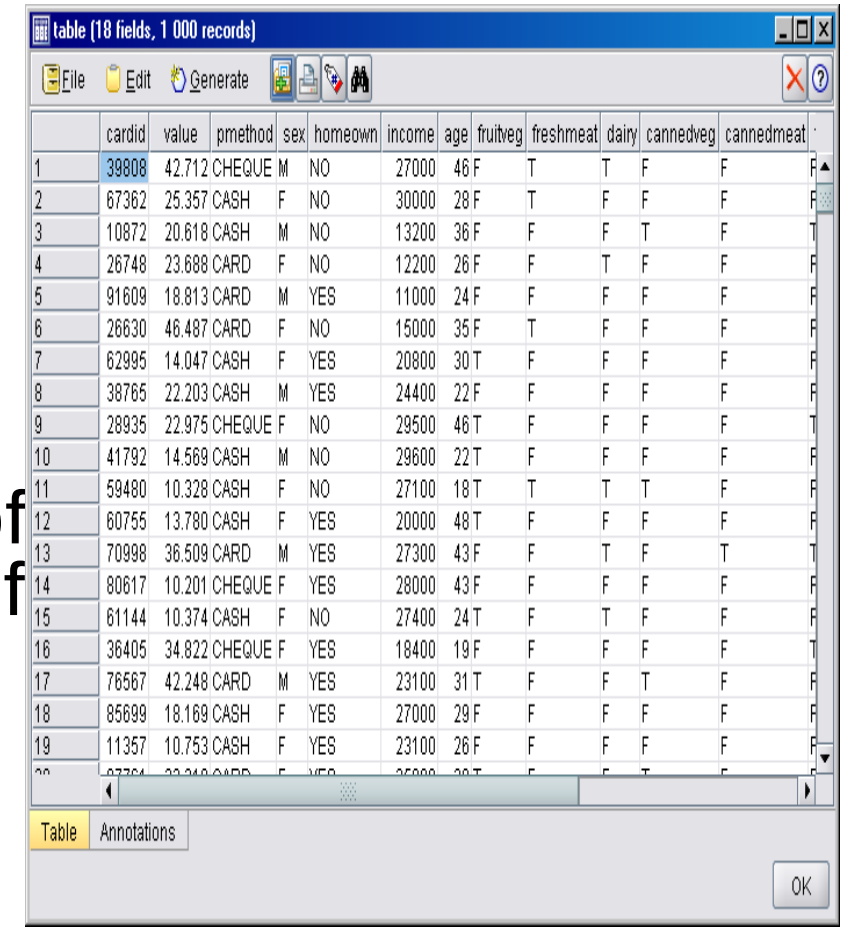
$$X \rightarrow Y$$

Meaning:

transactions containing items of set X tend to contain items of set Y

Example:

{eggs, bacon} → {cheese}



	cardid	value	pmethod	sex	homeown	income	age	fruitveg	freshmeat	dairy	cannedveg	cannedmeat
1	39808	42.712	CHEQUE	M	NO	27000	46	F	T	T	F	F
2	67362	25.357	CASH	F	NO	30000	28	F	T	F	F	F
3	10872	20.618	CASH	M	NO	13200	36	F	F	F	T	F
4	26748	23.688	CARD	F	NO	12200	26	F	F	T	F	F
5	91609	18.813	CARD	M	YES	11000	24	F	F	F	F	F
6	26630	46.487	CARD	F	NO	15000	35	F	T	F	F	F
7	62995	14.047	CASH	F	YES	20800	30	T	F	F	F	F
8	38765	22.203	CASH	M	YES	24400	22	F	F	F	F	F
9	28935	22.975	CHEQUE	F	NO	29500	46	T	F	F	F	F
10	41792	14.569	CASH	M	NO	29600	22	T	F	F	F	F
11	59480	10.328	CASH	F	NO	27100	18	T	T	T	T	F
12	60755	13.780	CASH	F	YES	20000	48	T	F	F	F	F
13	70998	36.509	CARD	M	YES	27300	43	F	F	T	F	T
14	80617	10.201	CHEQUE	F	YES	28000	43	F	F	F	F	F
15	61144	10.374	CASH	F	NO	27400	24	T	F	T	F	F
16	36405	34.822	CHEQUE	F	YES	18400	19	F	F	F	F	F
17	76567	42.248	CARD	M	YES	23100	31	T	F	F	T	F
18	85699	18.169	CASH	F	YES	27000	29	F	F	F	F	F
19	11357	10.753	CASH	F	YES	23100	26	F	F	F	F	F

Association rules (2/3)

Apriori-like: IF balance=high THEN loan=yes

client	income	balance	sex	unemployed	loan
k1	high	high	female	no	yes
k2	high	high	male	no	yes
k3	low	low	male	no	no
k4	low	high	female	yes	yes
k5	low	high	male	yes	yes
k6	low	low	female	yes	no
k7	high	low	male	no	yes
k8	high	low	female	yes	yes
k9	low	medium	male	yes	no
k10	high	medium	female	no	yes
k11	low	medium	female	yes	no
k12	low	medium	male	no	yes

	SUC	¬SUC	Σ
ANT	**a(4)**	b(0)	**4**
¬ANT	c(4)	d(4)	8
Σ	8	4	**12**

- Support $a/(a+b+c+d) = 4/12$
- Confidence $a/(a+b) = 4/4$

Association rules (3/3)

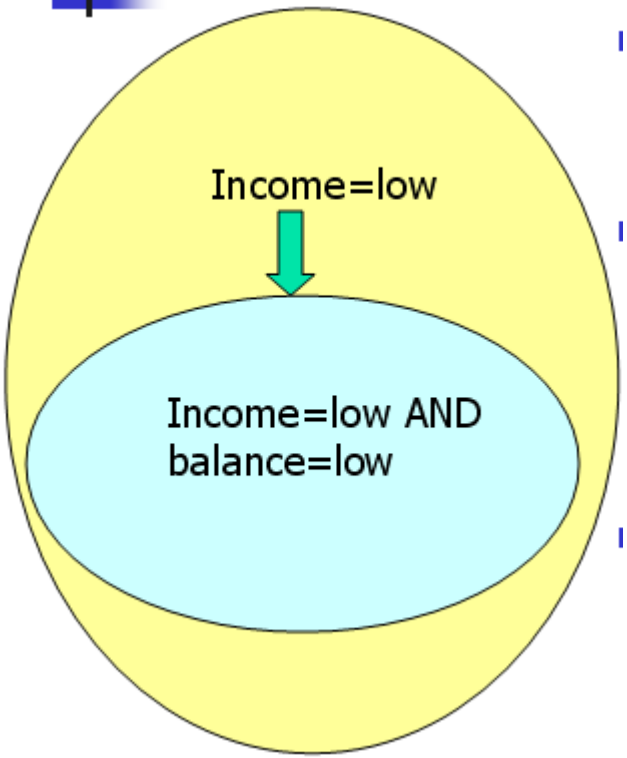
GUHA-like: balance(high OR medium) AND NOT(unemployed(yes)) $\Rightarrow_{0.9}$ loan(yes) / sex(male)

client	income	balance	sex	unemployed	loan
k1	high	high	female	no	yes
k2	high	high	male	no	yes
k3	low	low	male	no	no
k4	low	high	female	yes	yes
k5	low	high	male	yes	yes
k6	low	low	female	yes	no
k7	high	low	male	no	yes
k8	high	low	female	yes	yes
k9	low	medium	male	yes	no
k10	high	medium	female	no	yes
k11	low	medium	female	yes	no
k12	low	medium	male	no	yes

	SUC	¬SUC	Σ
ANT	**a(2)**	b(0)	**2**
¬ANT	c(2)	d(2)	4
Σ	4	2	**6**

- Support $a/(a+b+c+d) = 2/6$
- Confidence $a/(a+b) = 2/2$

Association rules mining algorithm (1/2)



- Generating syntactically correct combinations (323 for our toy example)

combination
. . .
4a
4n
5a
5n
1n 2n
1n 2s
1n 2v
1n 3m
1n 3z
. . .

combination
1n
1n 2n
1n 2n 3m
1n 2n 3m 4a
1n 2n 3m 4a 5a
1n 2n 3m 4a 5n
1n 2n 3m 4n
1n 2n 3m 4n 5a
1n 2n 3m 4n 5n
1n 2n 3m 5a
. . .

Association rules mining algorithm (2/2)

- Creating rules from combination
- Testing if rules fulfill quantitative characteristics (length, support, confidence)

- Algorithms (and implementations)
  - apriori (weka, Clementine, Enterprise Miner, ...)
  - GUHA method (LISp-Miner)

Association rules mining results

- Large list of association rules that should be evaluated by domain experts

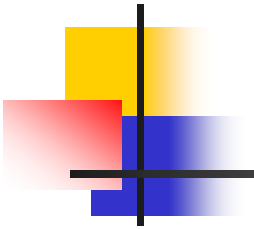
1. income=high 5 ==> loan=yes 5 conf:(1)
2. loan=no 4 ==> income=low 4 conf:(1)
3. **balance=high 4 ==> loan=yes 4 conf:(1)**
4. income=high unemployed=no 4 ==> loan=yes 4 conf:(1)
5. income=high sex=female 3 ==> loan=yes 3 conf:(1)
6. income=low sex=female 3 ==> unemployed=yes 3 conf:(1)
7. unemployed=yes loan=no 3 ==> income=low 3 conf:(1)
8. balance=high unemployed=no 2 ==> income=high 2 conf:(1)
9. income=high balance=high 2 ==> unemployed=no 2 conf:(1)
10. income=high balance=high 2 ==> loan=yes 2 conf:(1)

...

Association rules postprocessing

- sorting, selecting, searching the output
- vizualization

- eliminating redundant or irrelevant rules
- clustering
- . . .
- **association rule mining**



Association meta-rules

- Postprocessed (standard) association rules using association rule mining
- Inspired by meta-learning where results of individual classifiers are used as input for subsequent learning step

Typology of association meta-rules

- Qualitative

$$Ant \Rightarrow Suc$$

Where Ant and Suc are cedents (conjunctions of attribute-value pairs in the simplest case)

- Quantitative

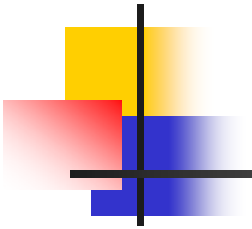
$$Ant \Rightarrow Q$$

$$Q1 \Rightarrow Q2$$

Where Ant is cedent and Qi are quantitative characteristics

- Frequent cedents

Conj



Association meta-rules mining

- Data
  - Encoded standard association rules
- Algorithm
  - (some) algorithm for association rule mining

    (e.g. apriori implemented in weka)
- Results
  - List of association meta-rules

Encoding association rules (1/2)

IF balance=high THEN loan=yes, supp(4), conf(1)

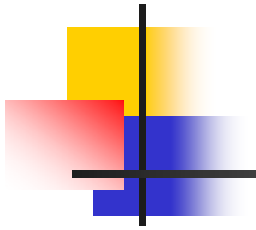
- Encoding cedents (composed of attribute value pairs):
  - Binary attributes (balance_ANY, loan_ANY)
  - Binary attributes (balance_high, loan_yes) ... k binary attributes for k values of original attribute
  - **Nominal attributes (balance, loan)**
  - Binary attributes w.r.t cedents
    - (An_balance_ANY, Suc_loan_ANY)
    - (Ant_balance_high, Suc_loan_yes)
  - Nominal attributes w.r.t cedents (Ant_balance, Suc_loan)
- Encoding quantitative characteristics:
  - Discretized numeric attributes

Encoding association rules (2/2)

1. income=high 5 ==> loan=yes 5 conf:(1)
2. loan=no 4 ==> income=low 4 conf:(1)
3. **balance=high 4 ==> loan=yes 4 conf:(1)**

id	true	income	balance	sex	unemployed	loan	support	confidence
1	t	high	?	?	?	yes	'(2.5-inf)'	'(0.915-inf)'
2	t	low	?	?	?	no	'(2.5-inf)'	'(0.915-inf)'
3	t	?	high	?	?	yes	'(2.5-inf)'	'(0.915-inf)'
4	t	high	?	?	no	yes	'(2.5-inf)'	'(0.915-inf)'
5	t	high	?	female	?	yes	'(2.5-inf)'	'(0.915-inf)'
6	t	low	?	female	yes	?	'(2.5-inf)'	'(0.915-inf)'
7	t	low	?	?	yes	no	'(2.5-inf)'	'(0.915-inf)'
8	t	high	high	?	no	?	'(-inf-2.5]'	'(0.915-inf)'
9	t	high	high	?	no	?	'(-inf-2.5]'	'(0.915-inf)'
10	t	high	high	?	?	yes	'(-inf-2.5]'	'(0.915-inf)'

Interpreting meta-rules (do they make sense?)



- Give meta-rules better insight into the list of standard (ordinary) association rules?

- Is the list of meta-rules easier to evaluate?

Running example: Standard rules

1. income=high 5 ==> loan=yes 5 conf:(1)
2. loan=no 4 ==> income=low 4 conf:(1)
3. **balance=high 4 ==> loan=yes 4 conf:(1)**
4. income=high unemployed=no 4 ==> loan=yes 4 conf:(1)
5. income=high sex=female 3 ==> loan=yes 3 conf:(1)
6. income=low sex=female 3 ==> unemployed=yes 3 conf:(1)
7. unemployed=yes loan=no 3 ==> income=low 3 conf:(1)
8. balance=high unemployed=no 2 ==> income=high 2 conf:(1)
9. income=high balance=high 2 ==> unemployed=no 2 conf:(1)
10. income=high balance=high 2 ==> loan=yes 2 conf:(1)

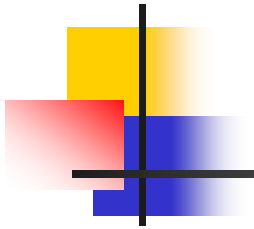
. . .

72. income=high 5 ==> unemployed=no loan=yes 4 conf:(0.8)

Running example: Qualitative rules

1. income=low loan=yes 7 ==> balance=high 7 conf:(1)
2. unemployed=yes loan=yes 7 ==> balance=high 7 conf:(1)
3. balance=high unemployed=yes 9 ==> income=low 8 conf:(0.89)
4. income=low balance=high 9 ==> unemployed=yes 8 conf:(0.89)
5. balance=high unemployed=no 8 ==> income=high 7 conf:(0.88)
6. income=high balance=high 8 ==> unemployed=no 7 conf:(0.88)
7. balance=medium loan=no 8 ==> unemployed=yes 7 conf:(0.88)
8. balance=medium unemployed=yes 8 ==> loan=no 7 conf:(0.88)
9. loan=no 18 ==> income=low 15 conf:(0.83)
10. **balance=high 22 ==> loan=yes 18 conf:(0.82)**
11. income=high 26 ==> loan=yes 21 conf:(0.81)

Running example: Quantitative rules



1. support=(-inf-2.5] 59 ==> confidence=(0.915-inf) 59 conf:(1)
2. balance=high 22 ==> confidence='(0.915-inf)' 22 conf:(1)
3. loan=no 18 ==> confidence='(0.915-inf)' 18 conf:(1)
4. **balance=high loan=yes 18 ==> confidence='(0.915-inf)' 18 conf:(1)**
5. sex=female 15 ==> confidence='(0.915-inf)' 15 conf:(1)
6. income=low loan=no 15 ==> confidence='(0.915-inf)' 15 conf:(1)
7. balance=medium 12 ==> support='(-inf-2.5]' 12 conf:(1)
8. balance=medium 12 ==> confidence='(0.915-inf)' 12 conf:(1)
9. sex=male 12 ==> support='(-inf-2.5]' 12 conf:(1)
10. sex=male 12 ==> confidence='(0.915-inf)' 12 conf:(1)
11. balance=medium 12 ==> support='(-inf-2.5]' confidence='(0.915-inf)' 12 conf:(1)

confidence=(0.915-inf) 66 ==> support=(-inf-2.5] 59 conf:(0.89)

Running example: Frequent cedents

1. true=t 72 ==> loan=yes 37 conf:(0.51)
2. true=t 72 ==> income=low 30 conf:(0.42)
3. true=t 72 ==> unemployed=no 28 conf:(0.39)
4. true=t 72 ==> unemployed=yes 27 conf:(0.38)
5. true=t 72 ==> income=high 26 conf:(0.36)
6. true=t 72 ==> balance=high 22 conf:(0.31)
7. true=t 72 ==> income=high loan=yes 21 conf:(0.29)
8. true=t 72 ==> income=low unemployed=yes 21 conf:(0.29)
9. true=t 72 ==> income=high unemployed=no 20 conf:(0.28)
10. true=t 72 ==> unemployed=no loan=yes 20 conf:(0.28)
11. true=t 72 ==> loan=no 18 conf:(0.25)
12. true=t 72 ==> balance=high loan=yes 18 conf:(0.25)

. . .

Initial experiments to evaluate the output size

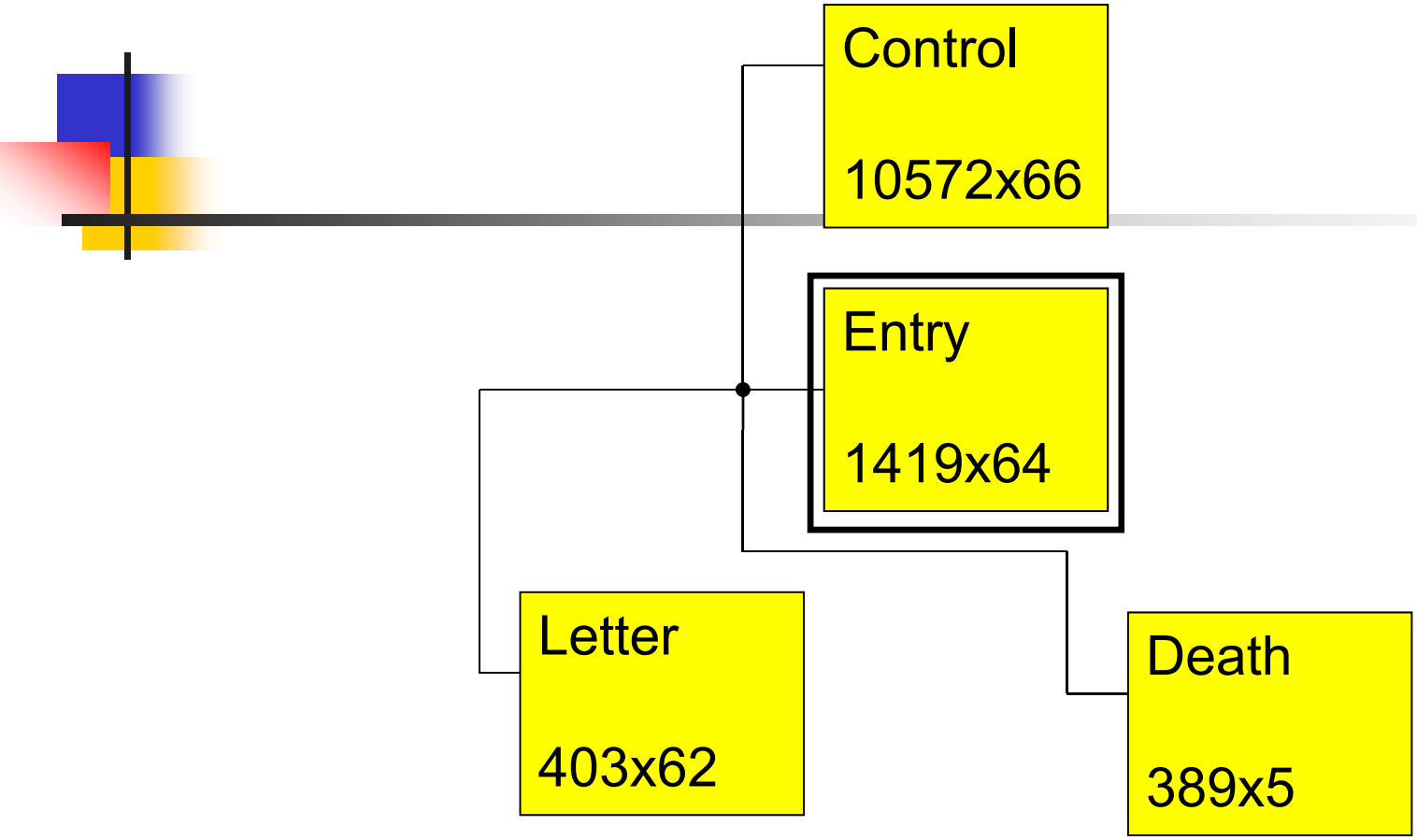
Data set	examples	attributes	ordinary rules	qualitative rules	quantitative rules	frequent cedents
Breast cancer	286	10	18742	167	341	80
Lenses	24	5	89	13	47	34
Mushroom	8124	23	100000	135	109	550
Vote	435	17	100000	6007	12	150
Monk	123	7	124	29	30	33
Tic-tac-toe	958	10	506	69	30	24
Tumor	339	18	100000	234	633	66

Atherosclerosis risk factors study

Longitudinal (1975-2000) study of atherosclerosis risk factors in the population of middle-aged men divided into three groups (normal, risk, pathological).

- to identify atherosclerosis risk factors prevalence in a population of middle-aged men,
- to follow the development of these risk factors and their impact on the examined men health, especially with respect to atherosclerotic CVD,
- to study the impact of complex risk factors intervention on development of risk factors and CVD mortality,
- to compare (after 10-12 years) risk factors profile and health of the selected men in different groups.

Data STULONG

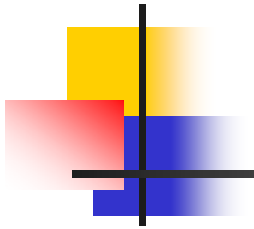


STULONG Tasks

- T1 relations between social characteristics, smoking, drinking of alcohol, coffee or tea
- T2 relations between smoking, drinking of alcohol, coffee or tea and risk factors
- T3 relations between smoking, drinking of alcohol, coffee or tea and physical examinations (blood pressure, BMI, skinfold)
- T4 relation between smoking, drinking of alcohol, coffee or tea and biochemical examinations (cholesterol, triglycerides, urine)

STULONG experiments

task	attributes	ordinary rules	qualitative rules	quantitative rules	frequent cedents
T1	18	12389	36	28	142
T2	14	5186	23	35	162
T3	17	2369	16	24	59
T4	13	3391	6	6	110

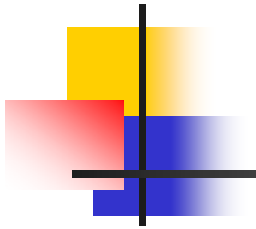


Further work (1/2)

- Association meta-rules in the GUHA (LISp-Miner) framework:
  - More complicated association rules
  - Cedents composed of literals

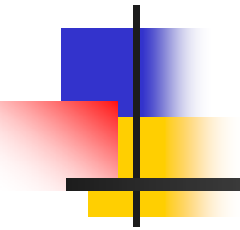
(balance(high OR medium) AND NOT(unemployed(yes)) $\Rightarrow_{0.9}$ loan(yes) / sex(male))

- Binary attributes (balance_ANY, loan_ANY)
- Binary attributes (balance(coef), $\neg$balance(coef))... 2 x 2^k binary attributes for k values of original attribute
- Nominal attributes (balance, loan) with original values + transforming of association rules
- ???



Further work (2/2)

- More complicated meta-rules
- Postprocessing of meta-rules (meta-meta learning?)



Thank you