Získavanie znalostí z databáz ak sú znalosťami fuzzy pravidlá

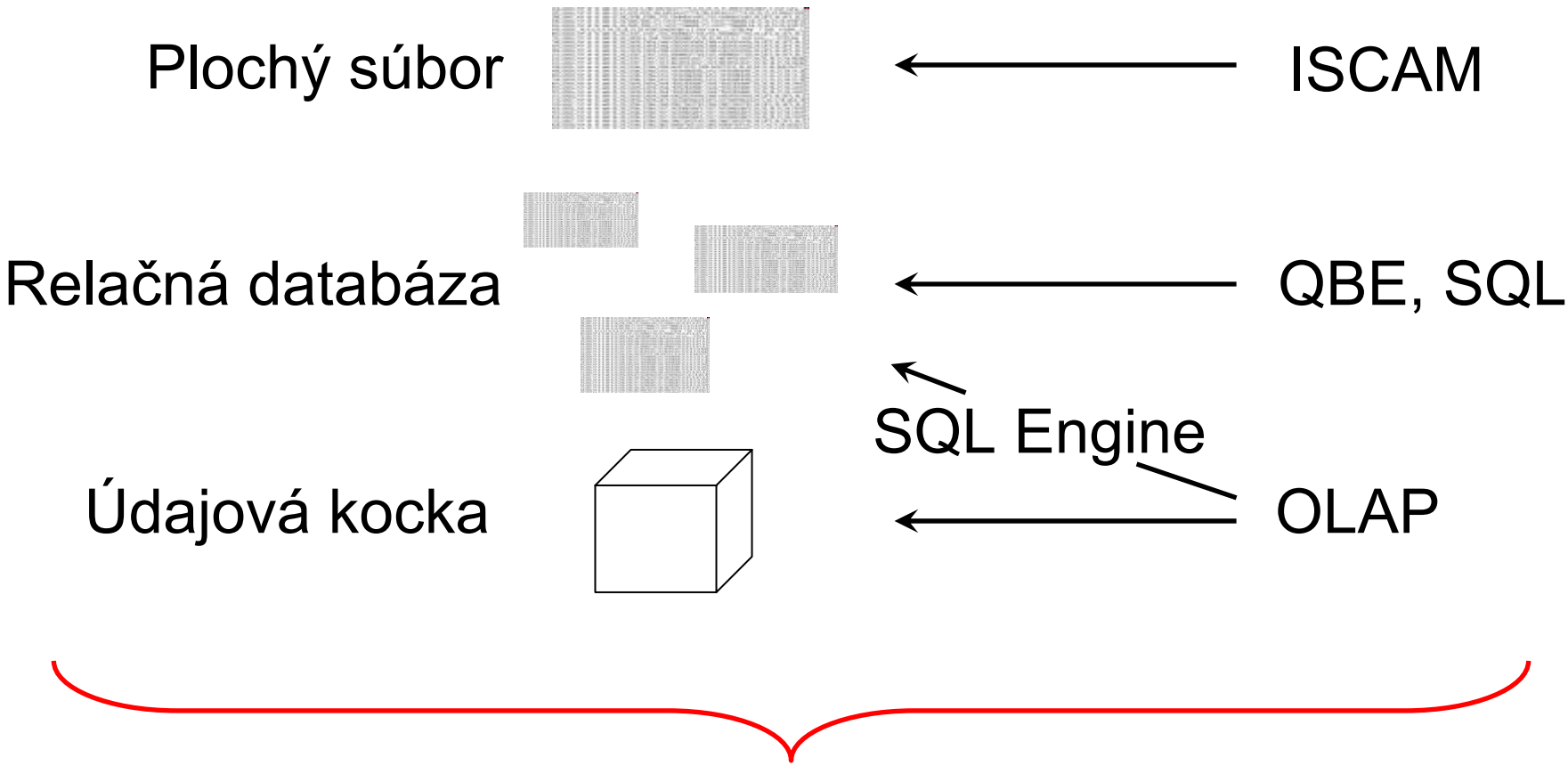
Systémová integrácia 2008
13. a 14. november

© Ján Boháčik
Jan.Bohacik@gmail.com

O čom budeme hovoriť?

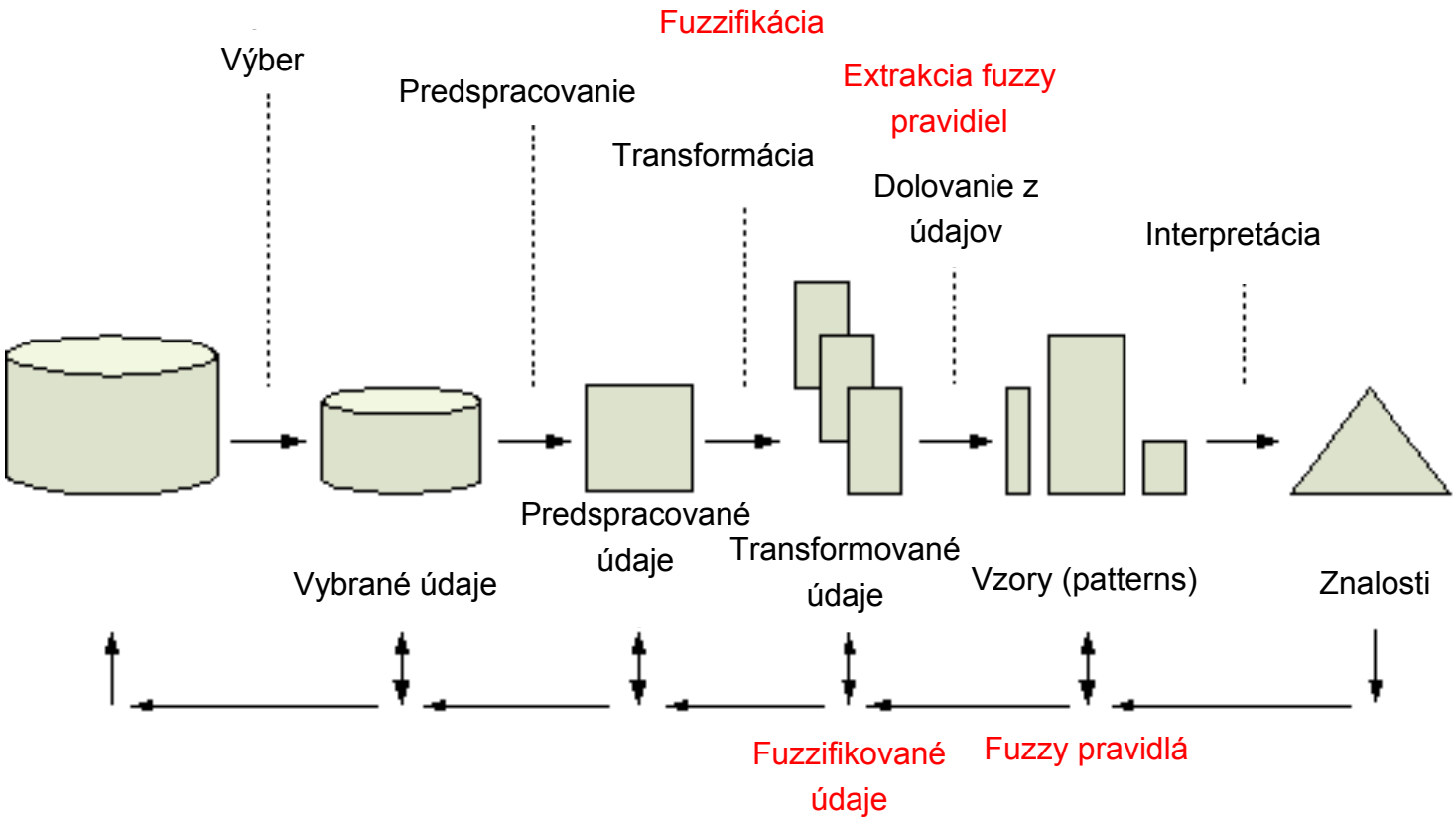
- Vymedzenie a definícia problému.
  - ✓ Kladenie dotazov. Získavanie znalostí z databáz.
  - ✓ Prečo používať fuzzy logiku?
  - ✓ Čo sú znalosti? Klasifikácia pravidiel.
- Získavanie fuzzy pravidiel z databáz.
  - ✓ Výber a predspracovanie údajov.
  - ✓ Fuzzifikácia údajov (transformácia údajov).
  - ✓ Interpretácia a vyhodnotenie fuzzy pravidiel.
- Zhrnutie a ďalšie plány.

Databázy a kladenie dotazov



Použiteľné iba ak sa vopred vie, čo sa hľadá!

Získavanie znalostí z databáz [Fayyad et al., 1996]



[Fayyad et al., 1996] Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R.: Advances in Knowledge Discovery and Data Mining, AAAI Press/MIT Press, Menlo Park, 1996, pp. 37-54.

Spôsoby reprezentácie znalostí

DB

atď.

Štatistické metódy

Rozhodovacie tabuľky

Pravidlá IF-THEN

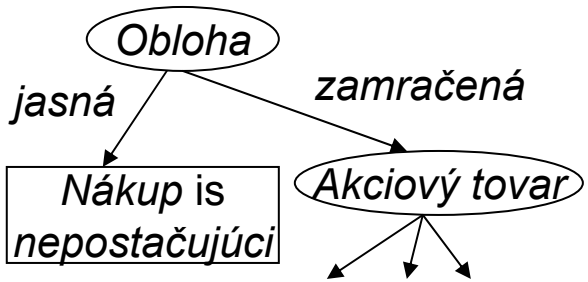
IF *Obloha* is *jasná* **THEN** *Nákup* is *postačujúci*

IF *Obloha* is *zamračená* **AND** *Akciový tovar* is *žiadny* **THEN** *Nákup* is *postačujúci*

Neurónová sieť

- Klasifikácia
- Nezrozumiteľné znalosti
- Preučenie
- Numerické dáta

Rozhodovací strom



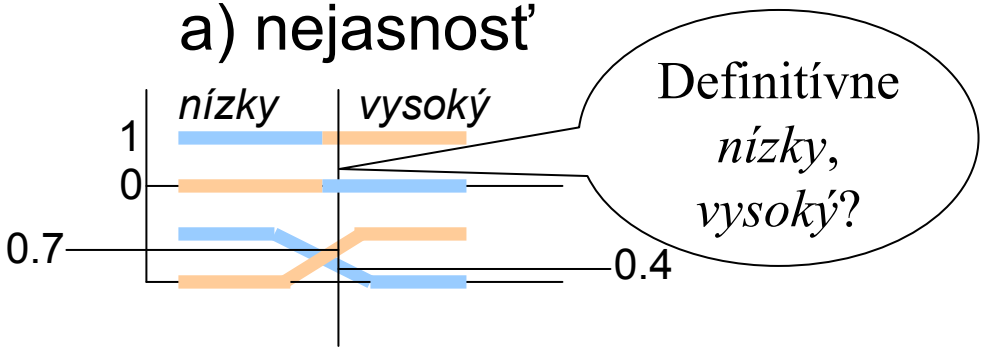
- Klasifikácia
- Dôležité atribúty
- Na pravidlá
- Kategoriálne triedy

- Klasifikácia, závislosti
- Blízke ľudskému uvažovaniu
- Vytvoriteľné a použiteľné tak ľuďmi ako aj počítačmi
- Mnoho použiteľných variant
- V zásade kategoriálne dáta

Prečo používať fuzzy logiku?

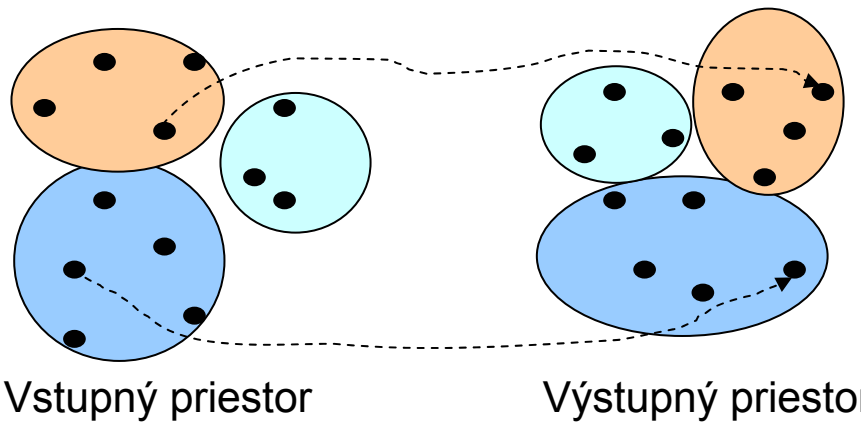
- Neurčitosť poznávacieho procesu

a) nejasnosť



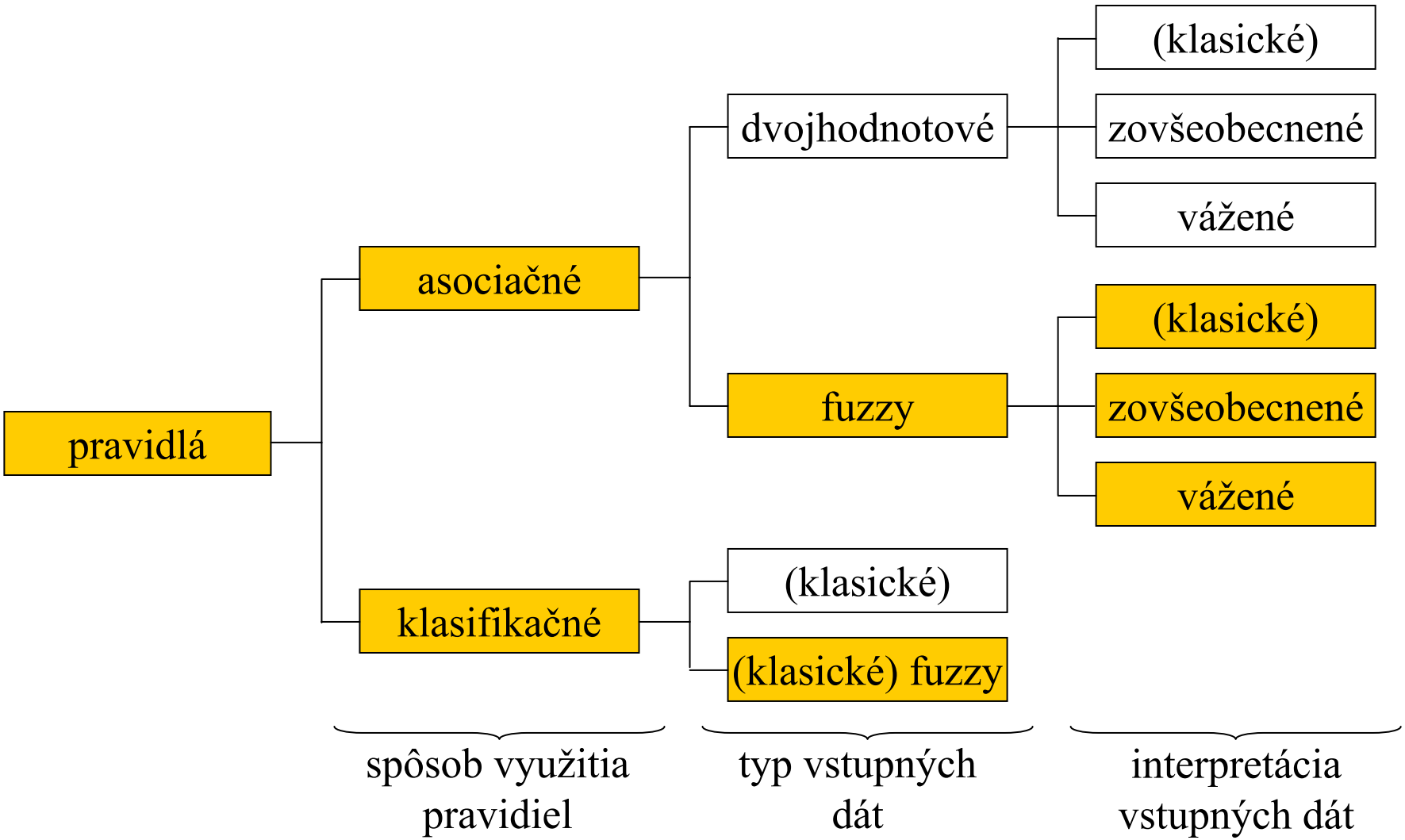
b) nejednoznačnosť

- Redukcia zložitosti



- Ľudia často vôbec nepoužívajú presné údaje

Klasifikácia pravidiel



Výber a predspracovanie údajov (1)

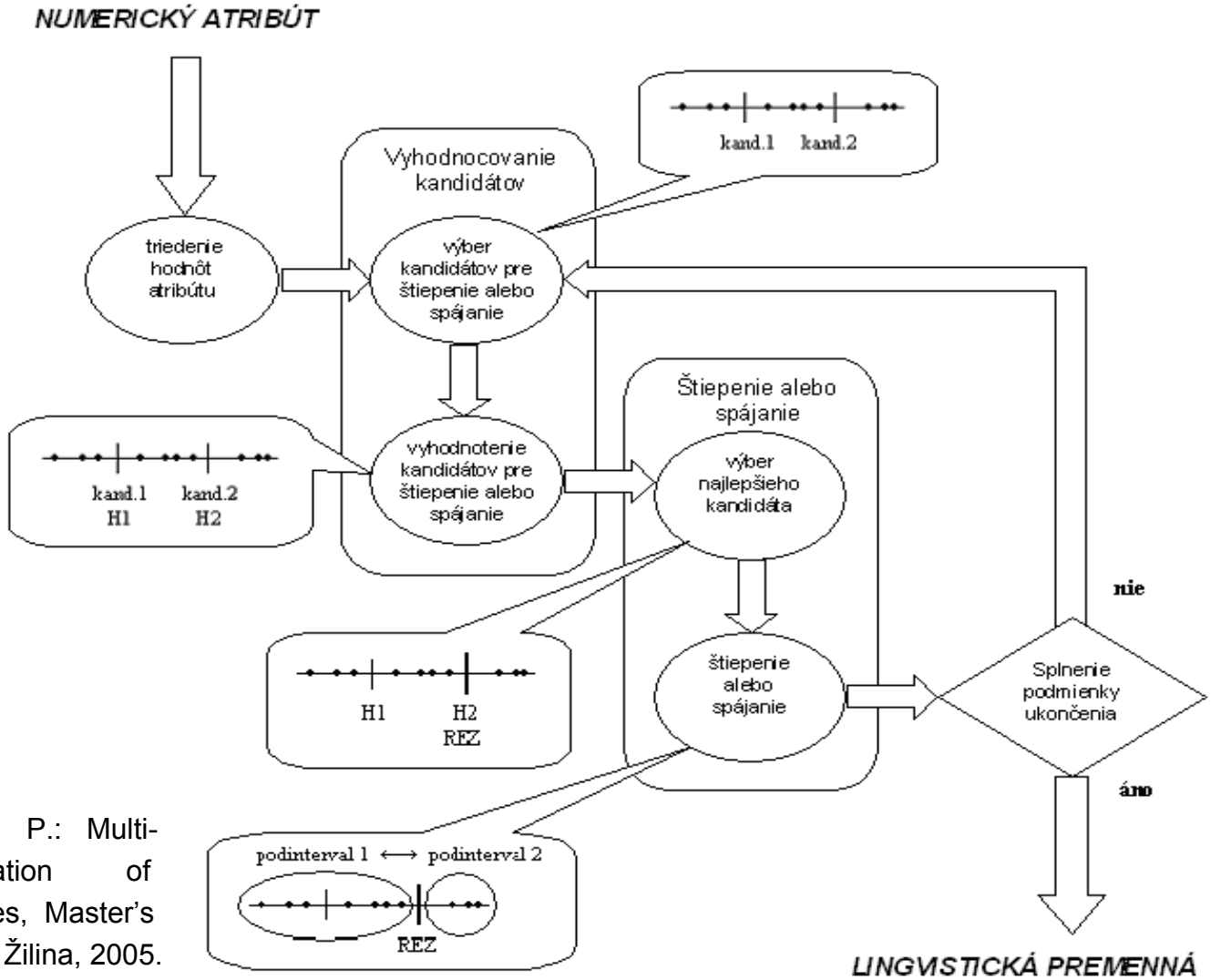
1. Vytvorenie jednej tabuľky,
2. Odstránenie nadbytočných inštancií a atribútov,
3. Kontrola formátu údajov,
4. Identifikácia odchýlok,
5. Odstránenie chýbajúcich hodnôt,
6. Identifikácia chybných kategoriálnych hodnôt,
7. Tvorba odvodených atribútov.

Výber a predspracovanie údajov (2)

Inštancia	*Teplota*	*Obloha*			*Novinka*	*Nákup*
		daždivá	*zamračená*	*jasná*		
(e_1)	15,4543	0,1	0,8	0,1	*áno*	20 000
(e_2)	16,0138	0,7	0,3	0,0	*nie*	15 000
(e_3)	11,9943	0,9	0,1	0,0	*áno*	14 000
(e_4)	14,2764	0,2	0,2	0,6	*nie*	14 800

Fragment údajov po výbere a predspracovaní

Fuzzifikácia údajov (1)



[Pazúrik., 2005] Pazúrik, P.: Multi-Interval Discretization of Continuous Attributes, Master's Thesis, University of Žilina, 2005.

Fuzzifikácia údajov (2)

Inštancia	*Teplota*			*Obloha*			*Novinka*		*Nákup*	
	nízka	*stredná*	*vysoká*	*daždivá*	*zamračená*	*jasná*	*nie*	*áno*	*nepostačujúci*	*postačujúci*
(e_1)	0,0	0,9	0,1	0,1	0,8	0,1	0,0	1,0	0,1	0,9
(e_2)	0,2	0,7	0,1	0,7	0,3	0,0	1,0	0,0	0,2	0,8
(e_3)	0,5	0,5	0,0	0,9	0,1	0,0	1,0	0,0	0,9	0,1
(e_4)	0,1	0,6	0,3	0,2	0,2	0,6	1,0	0,0	0,9	0,1

Fragment údajov po fuzzifikácií

Fragment databázy web obchodu

V	A_1			A_2			A_3			A_4		C	
	$a_{1,1}$	$a_{1,2}$	$a_{1,3}$	$a_{2,1}$	$a_{2,2}$	$a_{2,3}$	$a_{3,1}$	$a_{3,2}$	$a_{3,3}$	$a_{4,1}$	$a_{4,2}$	c_1	c_2
e_1	0.2	0.7	0.1	0.3	0.7	0.0	0.2	0.8	0.0	0.0	1.0	0.4	0.6
e_2	0.9	0.1	0.0	1.0	0.0	0.0	0.8	0.1	0.1	0.6	0.5	0.2	0.8
e_3	0.8	0.2	0.0	0.6	0.4	0.0	0.0	1.0	0.0	1.0	1.0	0.0	0.7
e_4	0.0	0.7	0.3	0.8	0.2	0.0	0.1	0.9	0.0	0.8	0.2	0.1	0.9
e_5	0.0	0.1	0.9	0.7	0.3	0.0	0.3	0.4	0.3	0.5	0.5	1.0	0.0
e_6	0.0	0.7	0.3	0.0	0.3	0.7	0.7	0.3	0.0	0.8	0.2	0.8	0.2
e_7	0.9	0.1	0.0	0.2	0.8	0.0	0.1	0.9	0.0	0.0	1.0	1.0	0.0
e_8	0.0	0.9	0.1	0.0	0.9	0.1	0.1	0.9	0.0	0.7	0.0	1.0	0.0
e_9	0.0	0.0	1.0	0.0	0.0	1.0	0.6	0.0	0.4	0.8	0.2	1.0	0.0
e_{10}	1.0	0.0	0.0	0.5	0.5	0.0	0.0	1.0	0.0	1.0	0.0	0.0	1.0
e_{11}	0.0	0.3	0.7	0.0	0.0	1.0	0.0	1.0	0.1	0.9	0.1	1.0	0.0
e_{12}	0.0	1.0	0.0	0.0	0.2	0.8	0.2	0.8	0.0	1.0	0.0	0.7	0.3
e_{13}	1.0	0.0	0.0	1.0	0.0	0.0	0.3	0.0	0.7	0.6	0.4	0.2	0.8
e_{14}	0.9	0.1	0.0	0.0	0.3	0.7	0.0	1.0	0.9	0.1	0.9	0.7	0.3
e_{15}	0.7	0.3	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.8	0.2	0.3	0.7
e_{16}	0.2	0.6	0.2	0.0	1.0	0.0	0.0	0.7	0.3	0.7	0.3	0.4	0.6

$e_i \in V \subseteq U$

V – známe inštancie

U – úplná množina

A = {A_1; A_2; A_3; A_4} = {*Teplota; Obloha*; *Akciový tovar*; *Návštevnosť*}

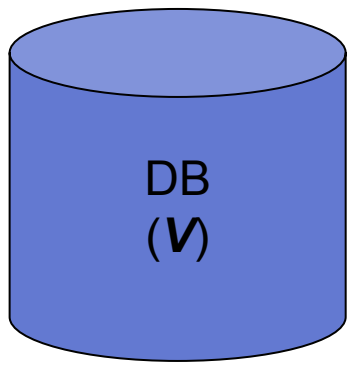
A_1 = *Teplota* = {$a_{1,1}$; $a_{1,2}$; $a_{1,3}$} = {*vysoká*; *stredná*; *nízka*}

A_2 = *Obloha* = {$a_{2,1}$; $a_{2,2}$; $a_{2,3}$} = {*jasná*; *zamračená*; *daždivá*}

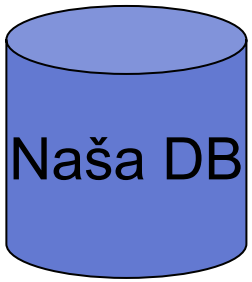
A_3 = *Akciový tovar* = {$a_{3,1}$; $a_{3,2}$; $a_{3,3}$} = {*žiaden*; *slabý*; bohatý}

A_4 = *Návštevnosť* = {$a_{4,1}$; $a_{4,2}$} ={*malá*; *vysoká*}

C = *Nákup* = {c_1; c_2} = {*postačujúci*; *nepostačujúci*}



Interpretácia a vyhodnotenie (1)

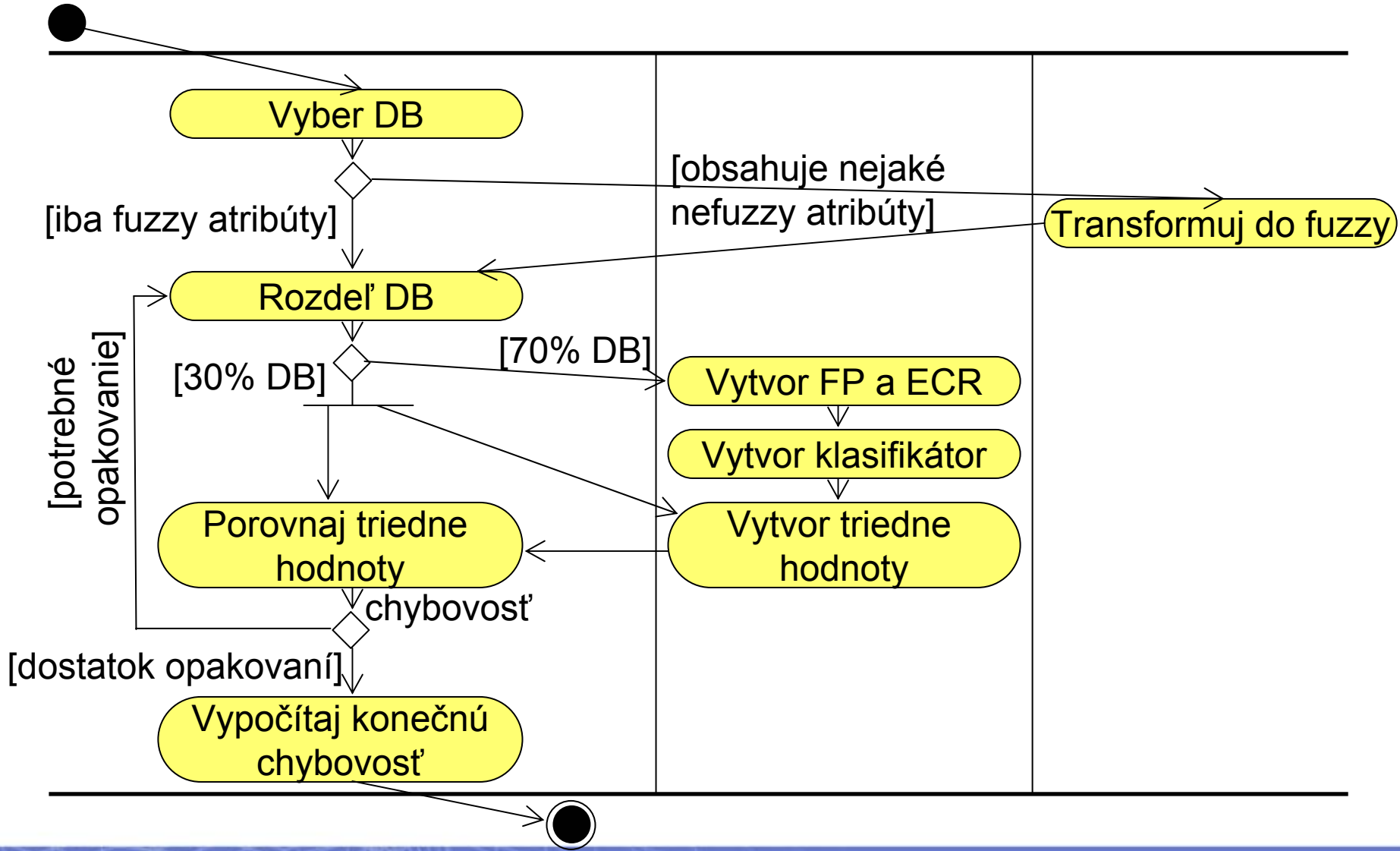


Určujú sa triedne hodnoty novej inštancie $e_{nová}$:

U	A_1			A_2			A_3			A_4		**Nákup** (C)	
	$a_{1,1}$	$a_{1,2}$	$a_{1,3}$	$a_{2,1}$	$a_{2,2}$	$a_{2,3}$	$a_{3,1}$	$a_{3,2}$	$a_{3,3}$	$a_{4,1}$	$a_{4,2}$	postačujúci (c_1)	nepostačujúci (c_2)
$e_{nová}$	0.2	0.7	0.1	0.3	0.7	0.0	0.2	0.8	0.0	0.0	1.0	?	?

IF *Obloha* is *jasná* **THEN** *Nákup* is *nepostačujúci*
IF *Obloha* is *zamračená* **AND** *Akciový tovar* is *žiaden* **THEN** *Nákup* is *postačujúci*
IF *Obloha* is *zamračená* **AND** *Teplota* is *vysoká* **AND** *Návštevnosť* is *malá*
THEN *Nákup* is *nepostačujúci*
IF *Obloha* is *zamračená* **AND** *Teplota* is *vysoká* **AND** *Návštevnosť* is *vysoká*
THEN *Nákup* is *postačujúci*
IF *Obloha* is *zamračená* **AND** *Teplota* is *stredná* **THEN** *Nákup* is *postačujúci*
IF *Teplota* is *nízka* **THEN** *Nákup* is *postačujúci*
IF *Obloha* is *zamračená* **AND** *Akciový tovar* is *slabý* **THEN** *Nákup* is *postačujúci*
IF *Obloha* is *daždivá* **THEN** *Nákup* is *postačujúci*

Interpretácia a vyhodnotenie (2)



Interpretácia a vyhodnotenie (3)

IF *Návštevnosť* is *vysoká* **THEN** *Nákup* is *postačujúci* **AND** *Teplota* is *nízka*

IF *Obloha* is *daždivá* **THEN** *Nákup* is *postačujúci*

.

.

.

IF *Nákup* is *postačujúci* **AND**

Akciový tovar is *žiaden* **THEN** *Obloha* is *zamračená*

Interpretácia a vyhodnotenie (4)

- Jednoduchosť
- Užitočnosť

$$\text{Užitočnosť(R)} = \frac{1}{\mathbf{M}(V)} \sum_{e \in V} \mathbf{T}(\mu_{Predpoklad}(e); \mu_{Záver}(e)) \in \langle 0; 1 \rangle$$

- Určitosť

$$\text{Určitosť(R)} = \frac{\sum_{e \in V} \mathbf{T}(\mu_{Predpoklad}(e); \mu_{Záver}(e))}{\sum_{e \in V} \mu_{Predpoklad}(e)} \in \langle 0; 1 \rangle$$

- Originalita
- Zisk

$$\text{Zisk(R)} = \text{Určitosť(R)} - \frac{1}{\mathbf{M}(V)} \sum_{e \in V} \mu_{Záver}(e)$$

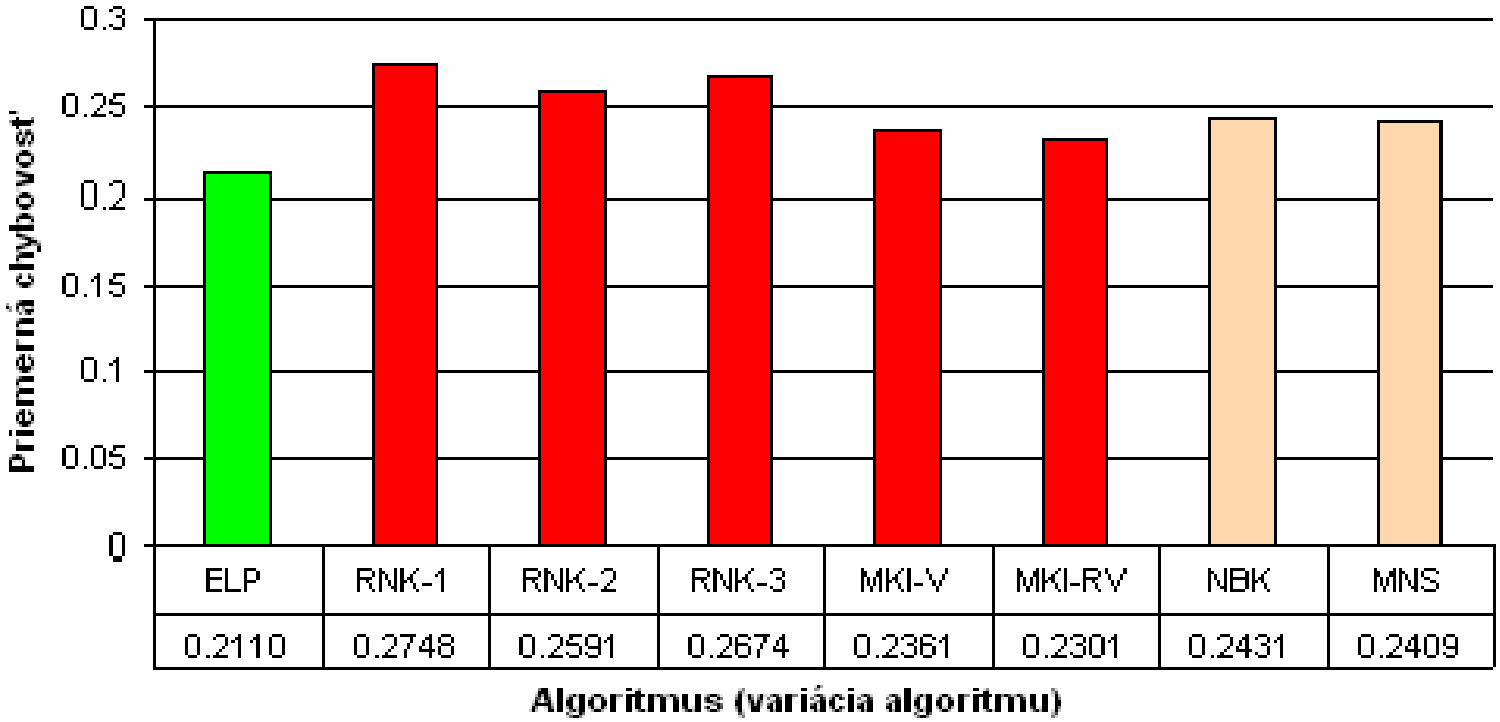
Eliminácia fuzzy atribútov [Boháčik, 2007]

{fuzzy pravidlo} = vytvorPravidlá(α; β; **A**; C; V)	
Krok 1	Inicializuj dočasne vytvorené pravidlá $\boldsymbol{R} := \theta$, dočasné trénovacie inštancie $\boldsymbol{I} := V$, momentálne uvažované atribúty $\boldsymbol{P} := \mathbf{A}$, maximálny počet povolených atribútov v predpoklade pravidla *maxDĺžka* := **M**(**A**), *dĺžka* := 1 a *ponechanýAtribút* := *false*.
Krok 2	Nastav eliminovaný atribút $A_{eliminovaný} := \text{argmax}\{\mathbf{NEJED}_\alpha(A_k; C; \boldsymbol{I}) \mid A_k \in \boldsymbol{P}\}$, $\boldsymbol{I}_1 := \theta$, $\boldsymbol{P}_1 := \boldsymbol{P} - A_{eliminovaný}$, $\boldsymbol{I}_2 := \theta$, $\boldsymbol{P}_2 := \boldsymbol{P}$. Ak *ponechanýAtribút* = *false*, potom nastav $\boldsymbol{Z} := V$. Inak polož $\boldsymbol{Z} := \boldsymbol{I}$.
Krok 3	Pre každé $\boldsymbol{e} \in \boldsymbol{I}$ urob: ak neexistuje žiadne $\boldsymbol{f} \in \boldsymbol{Z}$, $\boldsymbol{f} \neq \boldsymbol{e}$, také, že pre všetky $A_k \in \boldsymbol{P}_1$ sa $\text{argmax}\{\mu_{a_{k,l}}(\boldsymbol{f}) \mid a_{k,l} \in A_k\} = \text{argmax}\{\mu_{a_{k,l}}(\boldsymbol{e}) \mid a_{k,l} \in A_k\}$ a $\text{argmax}\{\mu_{c_j}(\boldsymbol{e}) \mid c_j \in C\}$, potom $\boldsymbol{I}_1 := \boldsymbol{I}_1 \cup \{\boldsymbol{e}\}$. Inak, $\boldsymbol{I}_2 := \boldsymbol{I}_2 \cup \{\boldsymbol{e}\}$.
Krok 4	Ak *dĺžka* + 1 < *maxDĺžka*, choď na Krok 5. Ak nie je, vytvor jedno pravidlo pre každé $\boldsymbol{e} \in \boldsymbol{I}$ a vlož ich bezduplicitne do $\boldsymbol{R}$. Forma pravidiel je: IF A_{k_1} is $\text{argmax}\{\mu_{a_{k_1,l}}(\boldsymbol{e}) \mid a_{k_1,l} \in A_{k_1}\}$ AND A_{k_2} is $\text{argmax}\{\mu_{a_{k_2,l}}(\boldsymbol{e}) \mid a_{k_2,l} \in A_{k_2}\}$ AND … AND A_{k_q} is $\text{argmax}\{\mu_{a_{k_q,l}}(\boldsymbol{e}) \mid a_{k_q,l} \in A_{k_q}\}$ THEN C is $\text{argmax}\{\mu_{c_j}(\boldsymbol{e}) \mid c_j \in C\}$, kde $\{A_{k_1}, A_{k_2}, \ldots, A_{k_q}\} = \boldsymbol{P}_1$ ak $\boldsymbol{e} \in \boldsymbol{I}_1$ a $\{A_{k_1}, A_{k_2}, \ldots, A_{k_q}\} = \boldsymbol{P}_2$ ak $\boldsymbol{e} \in \boldsymbol{I}_2$.
Krok 5	Vykonaj Kroky 2-5 jeden krát s $\boldsymbol{I} := \boldsymbol{I}_1$, $\boldsymbol{P} := \boldsymbol{P}_1$, *dĺžka* := *dĺžka* +1, *ponechanýAtribút* := *false* a jeden krát s $\boldsymbol{I} := \boldsymbol{I}_2$, $\boldsymbol{P} := \boldsymbol{P}_2$, *dĺžka* := *dĺžka* + 1, *ponechanýAtribút* := *true*.
Krok 6	Vypočítaj $\mathbf{UP}_0(\boldsymbol{E}_i; c; V)$ pre každé pravidlo IF $\boldsymbol{E}_i$ THEN C is c_j v $\boldsymbol{R}$. Ako výsledok algoritmu vráť pravidlá, pre ktoré $\mathbf{UP}_0(\boldsymbol{E}_i; c; V) \geq \beta$.

[Boháčik., 2007] Boháčik, J.: Induction by fuzzy attribute elimination, Journal of Information, Control and Management Systems, Vol. 5, No. 2, 2007, pp. 291-301.

Porovnanie algoritmov

- UCI Repository of ML Databases [Asuncion and Newman, 2007],
- Fuzzifikácia algoritmom [Lee et al., 2001].



[Asuncion and Newman, 2007] UCI Machine Learning Repository [http://www.ics.uci.edu/~mlearn/MLRepository.html], Irvine, CA: University of California, School of Information and Computer Science, 2007.

[Lee et al., 2001] Lee, H. M., Chen, C. M., Chen, J. M., JOU, Y. L.: An Efficient Fuzzy Classifier with Feature Selection Based on Fuzzy Entropy. In: IEEE Transactions on Systems, Man, and Cybernetics – Cybernetics 3, 2001, pp.426-432.

Zhrnutie a záver

- Proces Získavania znalostí z databáz:
  - Definovaný v 1996.
  - Odvtedy vzniklo viacero systémov na jeho riešenie, napr. Clementine, DB2 Intelligent Miner, alebo Enterprise Miner.
- Momentálne v tejto oblasti:
  - Vedci sa sústreďujú na metódy pre výber a predspracovanie údajov, interpretáciu či vyhodnotenie získaných znalostí, a metódy pre zdokonalenie existujúcich algoritmov.
  - Ukazuje sa, že využitie fuzzy množín a ďalších princípov fuzzy logiky pomáha odstrániť ich nedostatky.
- Hlavné prínosy:
  - Súhrnné rozpracovanie procesu Získavania znalostí z databáz s ohľadom na vytváranie fuzzy IF-THEN pravidiel vo fáze dolovania.
  - Poukázanie na jednotlivé čiastkové výskumné úlohy, ktoré treba v jednotlivých fázach tohto procesu riešiť.