

---

# Moderní systémy pro získávání znalostí z informací a dat

---

Jan Žižka

IBA – Institut biostatistiky a analýz  
PřF & LF, Masarykova universita  
Kamenice 126/3, 625 00 Brno

Email: [zizka@iba.muni.cz](mailto:zizka@iba.muni.cz)

# MATEMATICKÁ BIOLOGIE & ICT

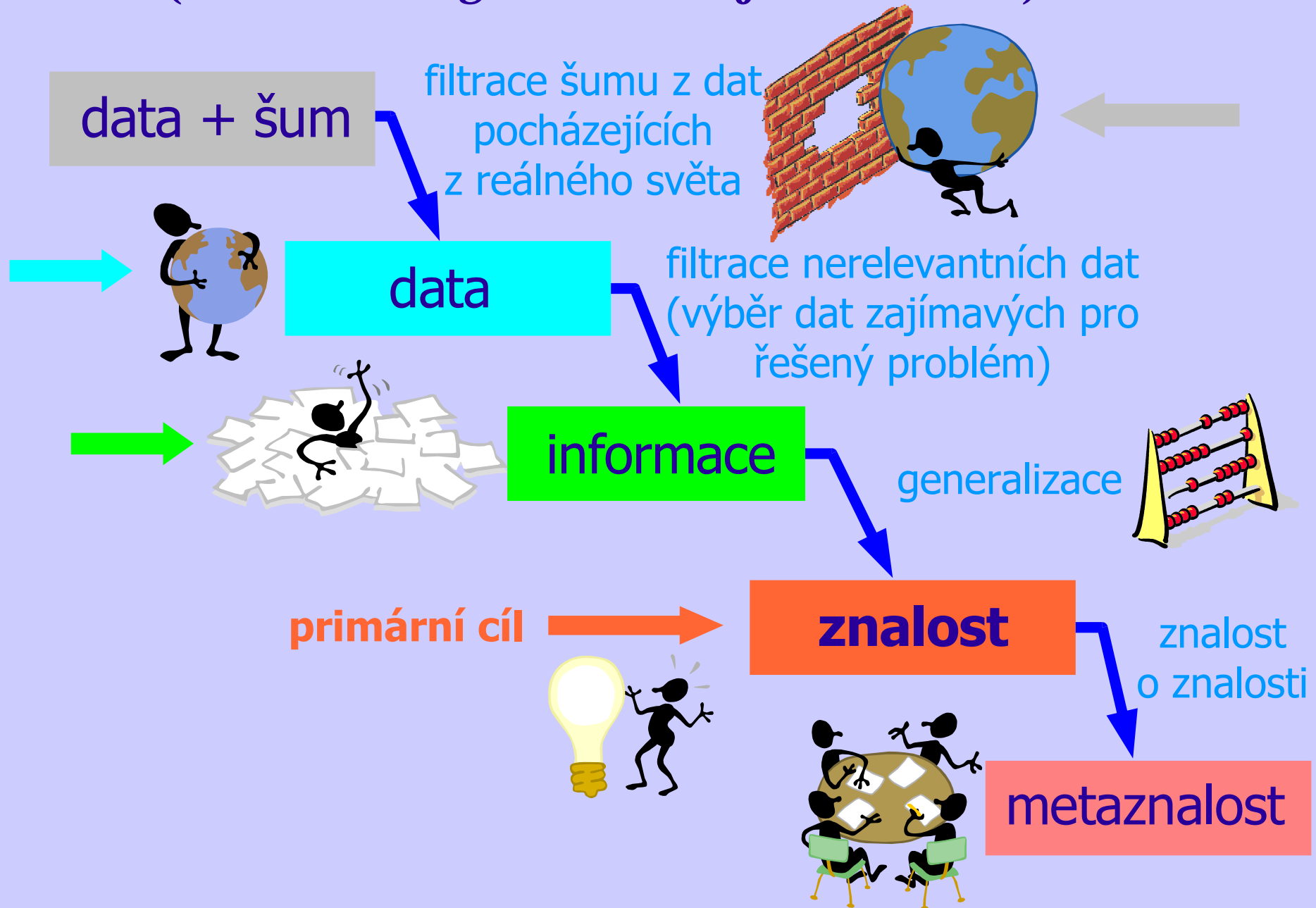
**Bioinformatika:** Aplikace výpočetních a statistických technik na zpracování a analýzu biologických dat.

**Strojové učení** (*machine learning, ML*), **umělá inteligence** (*artificial intelligence, AI*), **dolování z dat** (*data mining*):

Moderní systémy pro zpracování informace a získávání **znalostí** z dat. Rozšiřují a doplňují tradiční aplikace matematických a informatických metod také na biomedicínská data.

V komplikovaných případech, typických pro realitu, slouží jako alternativní metody, inspirované zpracováním informace inteligentními biologickými systémy.

# Hierarchický vztah *data* → *informace* → *znalost* (z hlediska algoritmů strojového učení)

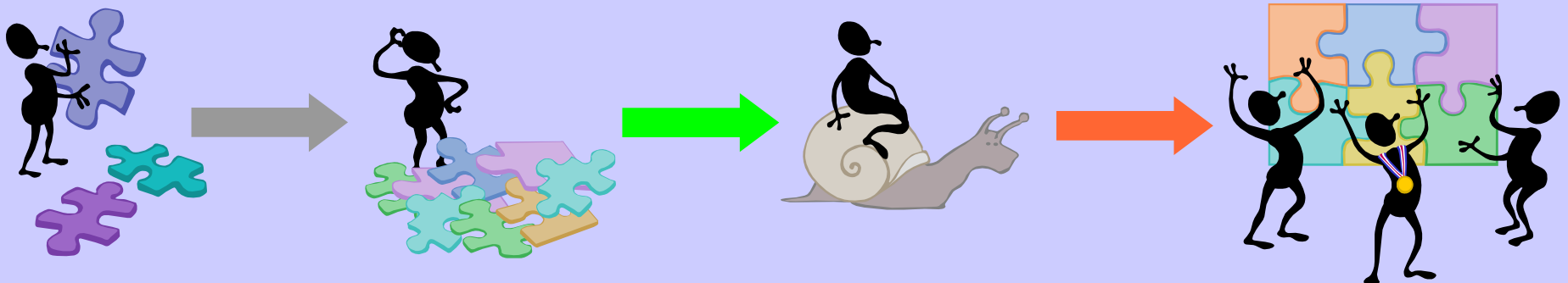


# MATEMATICKÁ BIOLOGIE & ICT

**Moderní přístupy umělé inteligence** se zaměřují na vyhledávání stanoveného cíle ve vysoce složitých prostorech obsahujících takové množství stavů, že z praktického hlediska nelze použít systematické prohledávání.

**Induktivní strojové učení** využívá možnost objevovat znalost na základě zobecnění omezeného množství vzorů.

**Dolování znalostí z dat** zahrnuje přípravu dat, hledání účinného algoritmu pro zobecnění, a nakonec interpretaci.



# MATEMATICKÁ BIOLOGIE & ICT

Vzdoruje-li reálný problém tradičním analytickým metodám, matematickému modelování, apod., pak lze k řešení použít *simulaci* přístupu *inteligentních biologických systémů schopných se učit a zobecňovat*.



Hledání skutečné znalosti v datech se často podobá hledání nejvyššího vrcholku kopce ve velmi zvlněné zamížené krajině (lokální extrémy, globální extrém, nelinearita, nespojitě funkce, apod.).



# MATEMATICKÁ BIOLOGIE & ICT

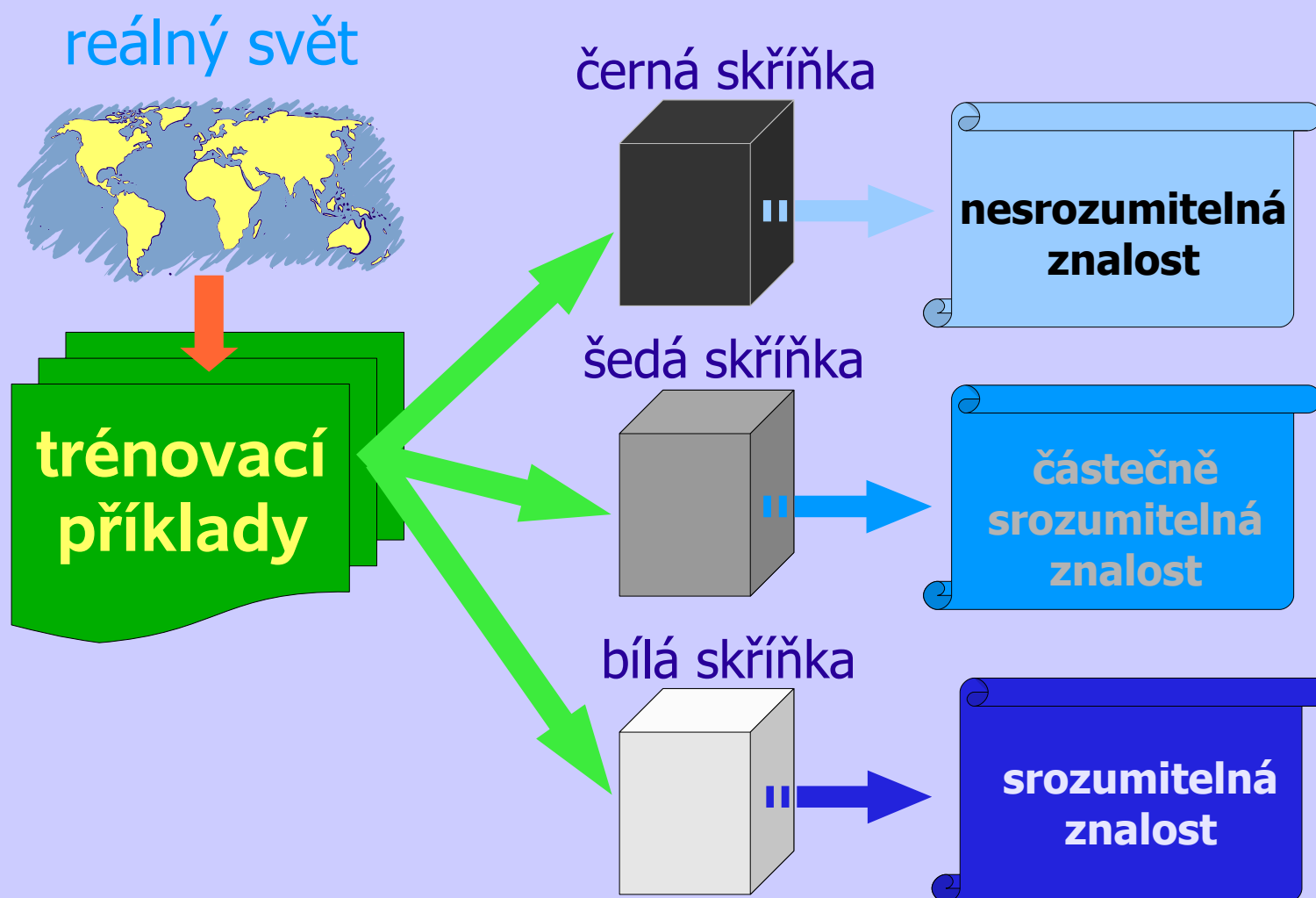
„Vytěžit“ použitelnou znalost ze „surových“ dat vyžaduje pochopit vlastnosti disponibilních metod, navrhnout a provést řadu časově náročných experimentů (výpočetní složitost – čas a paměť) a správně interpretovat získané znalosti pro jejich použití.

Induktivní učení z příkladů poskytne trénovaným algoritmům potřebné parametry. Natrénované algoritmy pak lze použít pro náročné regresní a klasifikační problémy.



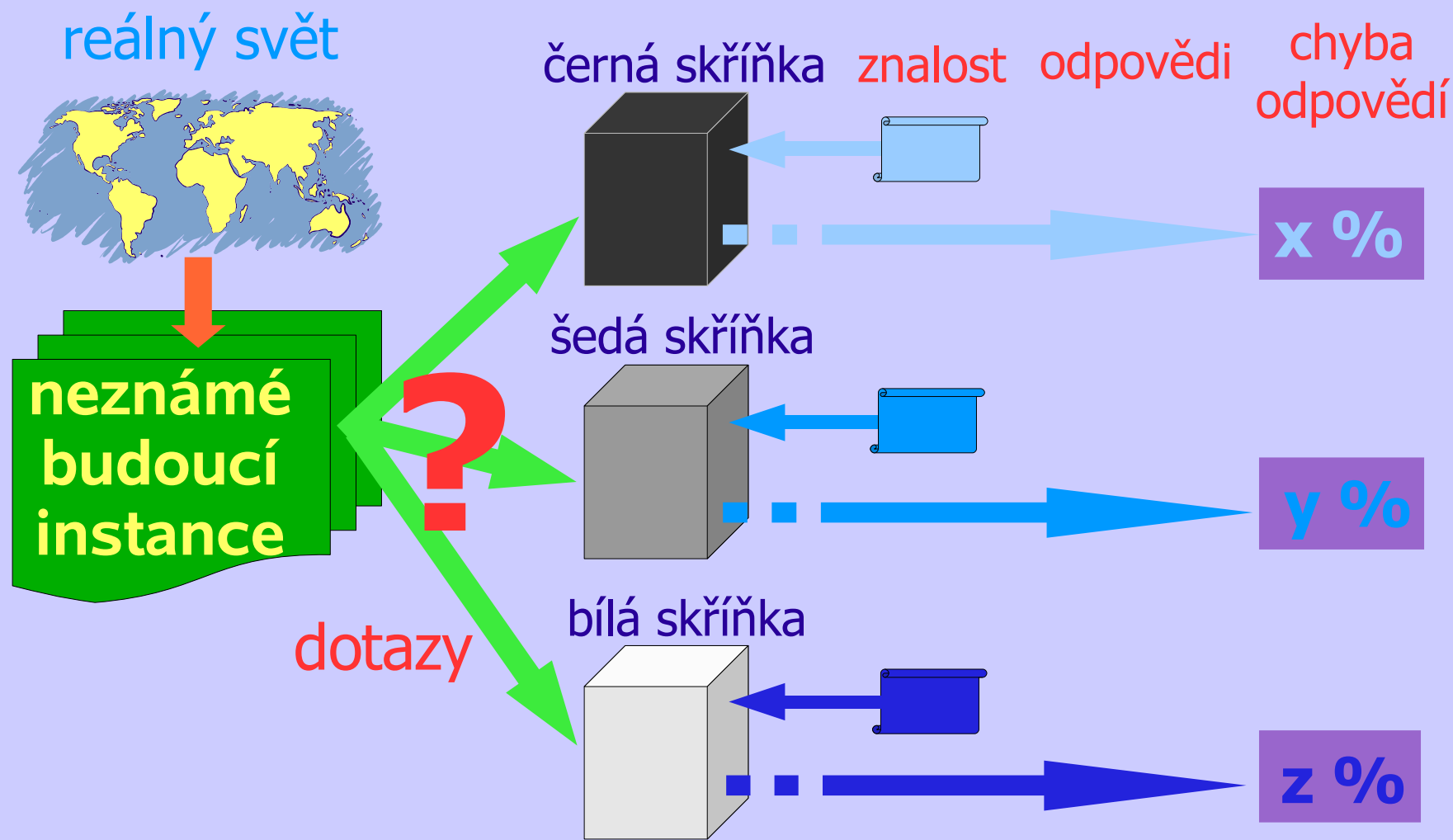
# MATEMATICKÁ BIOLOGIE & ICT

Natrénované algoritmy lze rozdělit podle typu poskytované znalosti, která se aplikuje na případy v budoucnosti:



# MATEMATICKÁ BIOLOGIE & ICT

Funkčnost algoritmů ovšem nemusí (ale i může) odpovídat srozumitelnosti znalosti získané trénováním:





# MATEMATICKÁ BIOLOGIE & ICT

Algoritmy lze také rozdělit podle typu učení:



# MATEMATICKÁ BIOLOGIE & ICT

Data jsou nejčastěji uspořádána formou tabulky, kde *řádky* představují *instance* (příklady, vzorky, ...) a *sloupce atributy* (dimenze, parametry, proměnné, vlastnosti, ...):

← jeden z atributů

Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
5	1	1	1	2	1	3	1	1	2
5	4	4	5	7	10	3	2	1	2
3	1	1	1	2	2	3	1	1	2
6	8	8	1	3	4	3	7	1	2
4	1	1	3	2	1	3	1	1	2
8	10	10	8	7	10	9	7	1	4
1	1	1	1	2	10	3	1	1	2
2	1	2	1	2	1	3	1	1	2
2	1	1	1	2	1	1	1	5	2
4	2	1	1	2	1	2	1	1	2
1	1	1	1	1	1	3	1	1	2
2	1	1	1	2	1	2	1	1	2
5	3	3	3	2	3	4	4	1	4
1	1	1	1	2	3	3	1	1	2
8	7	5	10	7	9	5	5	4	4
7	4	6	4	6	1	4	3	1	4
4	1	1	1	2	1	2	1	1	2
4	1	1	1	2	1	3	1	1	2
10	7	7	6	4	10	4	1	2	4
6	1	1	1	2	1	3	1	1	2
7	2	2	10	5	10	5	4	4	4

← názvy atributů

← klasifikační třída

← jeden z příkladů

(Wisconsin breast-cancer data)

# MATEMATICKÁ BIOLOGIE & ICT

V současnosti existuje již řada uživatelsky pohodlných nástrojů pro dolování znalostí strojovým učením, např. WEKA:

The screenshot displays the Weka GUI Explorer interface. On the left, the 'Weka GUI Chooser' window shows the 'Waikato Environment for Knowledge Analysis' version 3.5.3, with a 'GUI' section containing buttons for 'Simple CLI', 'Explorer', 'Experimenter', 'KnowledgeFlow', 'ArffViewer', and 'Log'. The main 'Weka Explorer' window has a menu bar with 'Preprocess', 'Classify', 'Cluster', 'Associate', 'Select attributes', and 'Visualize'. Below the menu are buttons for 'Open file...', 'Open URL...', 'Open DB...', 'Generate...', 'Undo', 'Edit...', and 'Save...'. A 'Filter' section contains a 'Choose' button and a text field set to 'None', with an 'Apply' button. The 'Current relation' section shows 'Relation: Breast-Cancer-weka.filters.unsupervised.attribute.Remove-R1' with 699 instances and 10 attributes. The 'Attributes' section has buttons for 'All', 'None', 'Invert', and 'Pattern', followed by a table of attributes with checkboxes. The 'Selected attribute' section shows 'Name: Clump Thickness', 'Type: Numeric', 'Missing: 0 (0%)', 'Distinct: 10', and 'Unique: 0 (0%)'. Below this is a table of statistics for 'Clump Thickness':

Statistic	Value
Minimum	1
Maximum	10
Mean	4.418
StdDev	2.816

The 'Class' section shows 'Class: Class (Nom)' and a 'Visualize All' button. Below this is a histogram for 'Clump Thickness' with a blue area under the curve and red bars on top. The x-axis ranges from 1 to 10, and the y-axis shows counts for each bin: 195, 108, 80, 130, 34, 23, 46, 83.

The 'Status' bar at the bottom shows 'OK' and a 'Log' button with a small bird icon and 'x 0'.

# MATEMATICKÁ BIOLOGIE & ICT

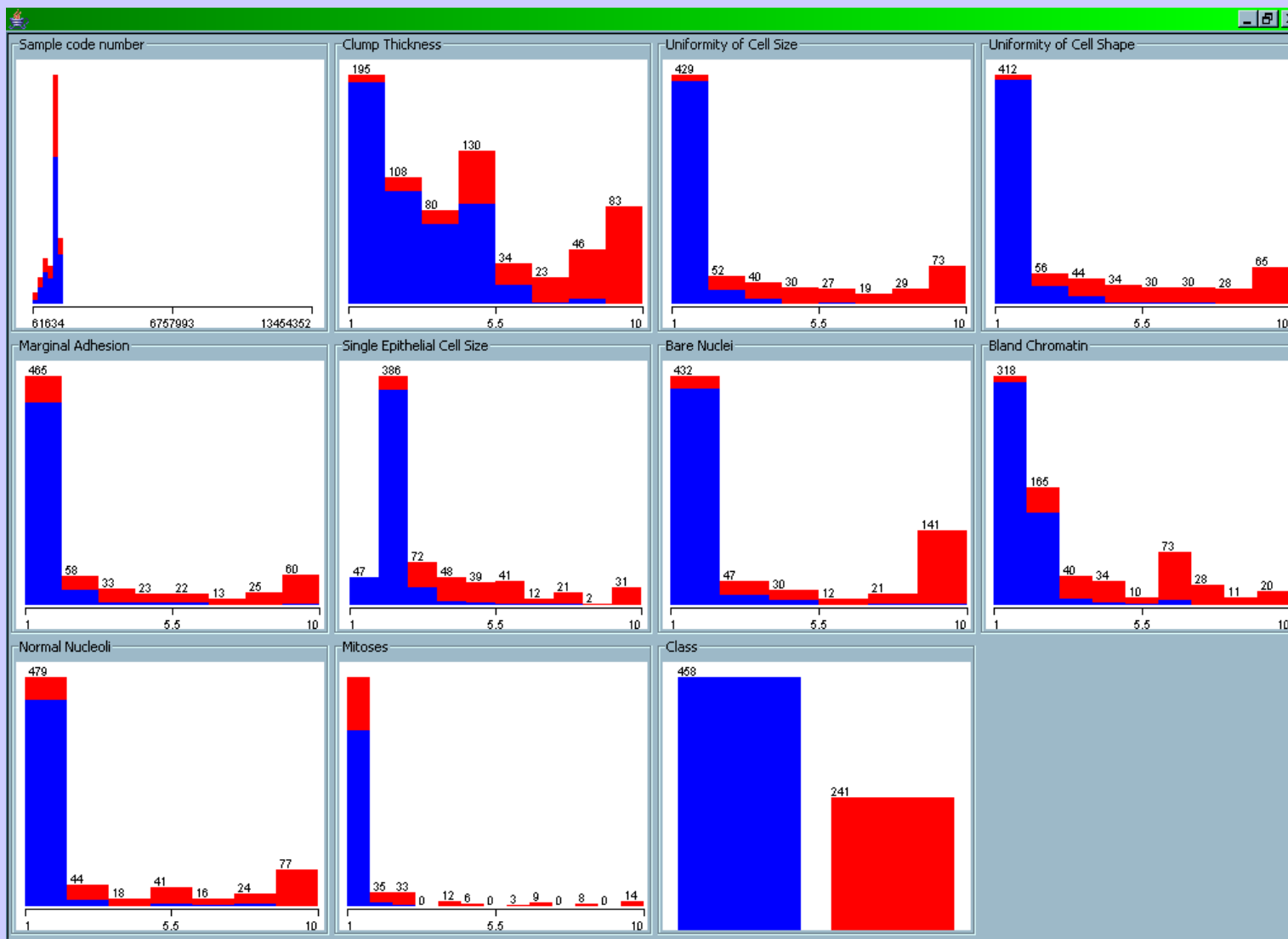
WEKA obsahuje i editor dat typu *spreadsheet*, který **nemá** typická omezení (např. pouze 256 sloupců a 65 536 řádků):

Relation: Breast-Cancer

No.	Sample code number Numeric	Clump Thickness Numeric	Uniformity of Cell Size Num	Uniformity of Cell Shape	Marginal Adhesion Numeric	Single Epithelial Cell Size Numeric	Bare Nuclei Numeric	Bland Chromatin Numeric	Normal Nucleoli Numeric	Mitoses Numeric	Class Nom
1	1000025.0	5.0			1.0	2.0	1.0	3.0	1.0	1.0	2
2	1002945.0	5.0			5.0	7.0	10.0	3.0	2.0	1.0	2
3	1015425.0	3.0			1.0	2.0	2.0	3.0	1.0	1.0	2
4	1016277.0	6.0			1.0	3.0	4.0	3.0	7.0	1.0	2
5	1017023.0	4.0			3.0	2.0	1.0	3.0	1.0	1.0	2
6	1017122.0	8.0			8.0	7.0	10.0	9.0	7.0	1.0	4
7	1018099.0	1.0			1.0	2.0	10.0	3.0	1.0	1.0	2
8	1018561.0	2.0			1.0	2.0	1.0	3.0	1.0	1.0	2
9	1033078.0	2.0			1.0	2.0	1.0	1.0	1.0	5.0	2
10	1033078.0	4.0			1.0	2.0	1.0	2.0	1.0	1.0	2
11	1035283.0	1.0			1.0	1.0	1.0	3.0	1.0	1.0	2
12	1036172.0	2.0			1.0	2.0	1.0	2.0	1.0	1.0	2
13	1041801.0	5.0			3.0	2.0	3.0	4.0	4.0	1.0	4
14	1043999.0	1.0	1.0	1.0	1.0	2.0	3.0	3.0	1.0	1.0	2
15	1044572.0	8.0	7.0	5.0	10.0	7.0	9.0	5.0	5.0	4.0	4
16	1047630.0	7.0	4.0	6.0	4.0	6.0	1.0	4.0	3.0	1.0	4
17	1048672.0	4.0	1.0	1.0	1.0	2.0	1.0	2.0	1.0	1.0	2
18	1049815.0	4.0	1.0	1.0	1.0	2.0	1.0	3.0	1.0	1.0	2
19	1050670.0	10.0	7.0	7.0	6.0	4.0	10.0	4.0	1.0	2.0	4
20	1050718.0	6.0	1.0	1.0	1.0	2.0	1.0	3.0	1.0	1.0	2
21	1054590.0	7.0	3.0	2.0	10.0	5.0	10.0	5.0	4.0	4.0	4
22	1054593.0	10.0	5.0	5.0	3.0	6.0	7.0	7.0	10.0	1.0	4
23	1056784.0	3.0	1.0	1.0	1.0	2.0	1.0	2.0	1.0	1.0	2
24	1057013.0	8.0	4.0	5.0	1.0	2.0		7.0	3.0	1.0	4
25	1059552.0	1.0	1.0	1.0	1.0	2.0	1.0	3.0	1.0	1.0	2
26	1065726.0	5.0	2.0	3.0	4.0	2.0	7.0	3.0	6.0	1.0	4
27	1066373.0	3.0	2.0	1.0	1.0	1.0	1.0	2.0	1.0	1.0	2
28	1066979.0	5.0	1.0	1.0	1.0	2.0	1.0	2.0	1.0	1.0	2
29	1067444.0	2.0	1.0	1.0	1.0	2.0	1.0	2.0	1.0	1.0	2
30	1070935.0	1.0	1.0	3.0	1.0	2.0	1.0	1.0	1.0	1.0	2
31	1070935.0	3.0	1.0	1.0	1.0	1.0	1.0	2.0	1.0	1.0	2
32	1071760.0	2.0	1.0	1.0	1.0	2.0	1.0	3.0	1.0	1.0	2
33	1072179.0	10.0	7.0	7.0	3.0	8.0	5.0	7.0	4.0	3.0	4
34	1074610.0	2.0	1.0	1.0	2.0	2.0	1.0	3.0	1.0	1.0	2
35	1075123.0	3.0	1.0	2.0	1.0	2.0	1.0	2.0	1.0	1.0	2
36	1079304.0	2.0	1.0	1.0	1.0	2.0	1.0	2.0	1.0	1.0	2
37	1080185.0	10.0	10.0	10.0	8.0	6.0	1.0	8.0	9.0	1.0	4
38	1081791.0	6.0	2.0	1.0	1.0	1.0	1.0	7.0	1.0	1.0	2
39	1084584.0	5.0	4.0	4.0	9.0	2.0	10.0	5.0	6.0	1.0	4

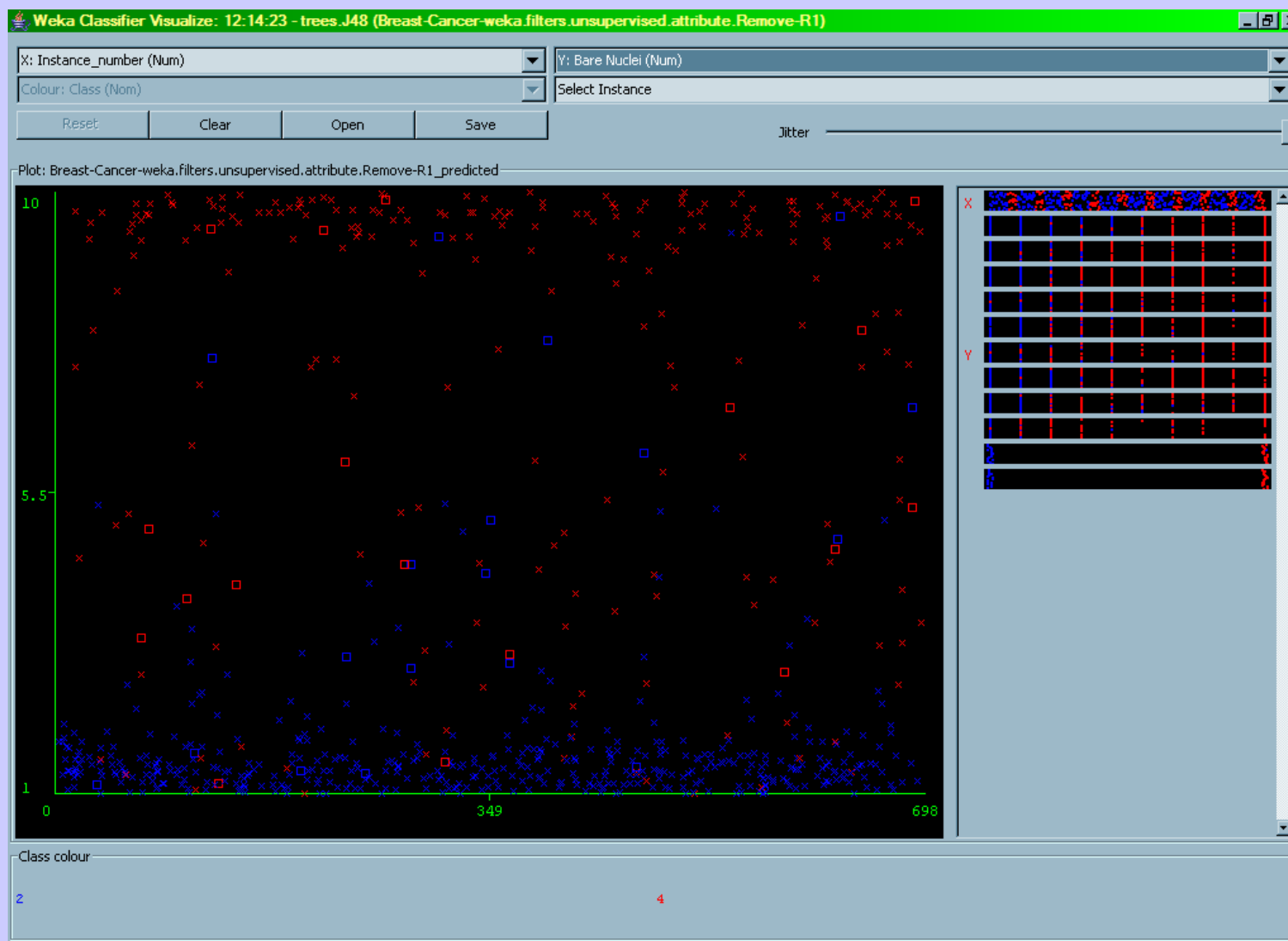
# MATEMATICKÁ BIOLOGIE & ICT

WEKA podporuje také zobrazování, např. rozložení hodnot všech atributů včetně klasifikační třídy:



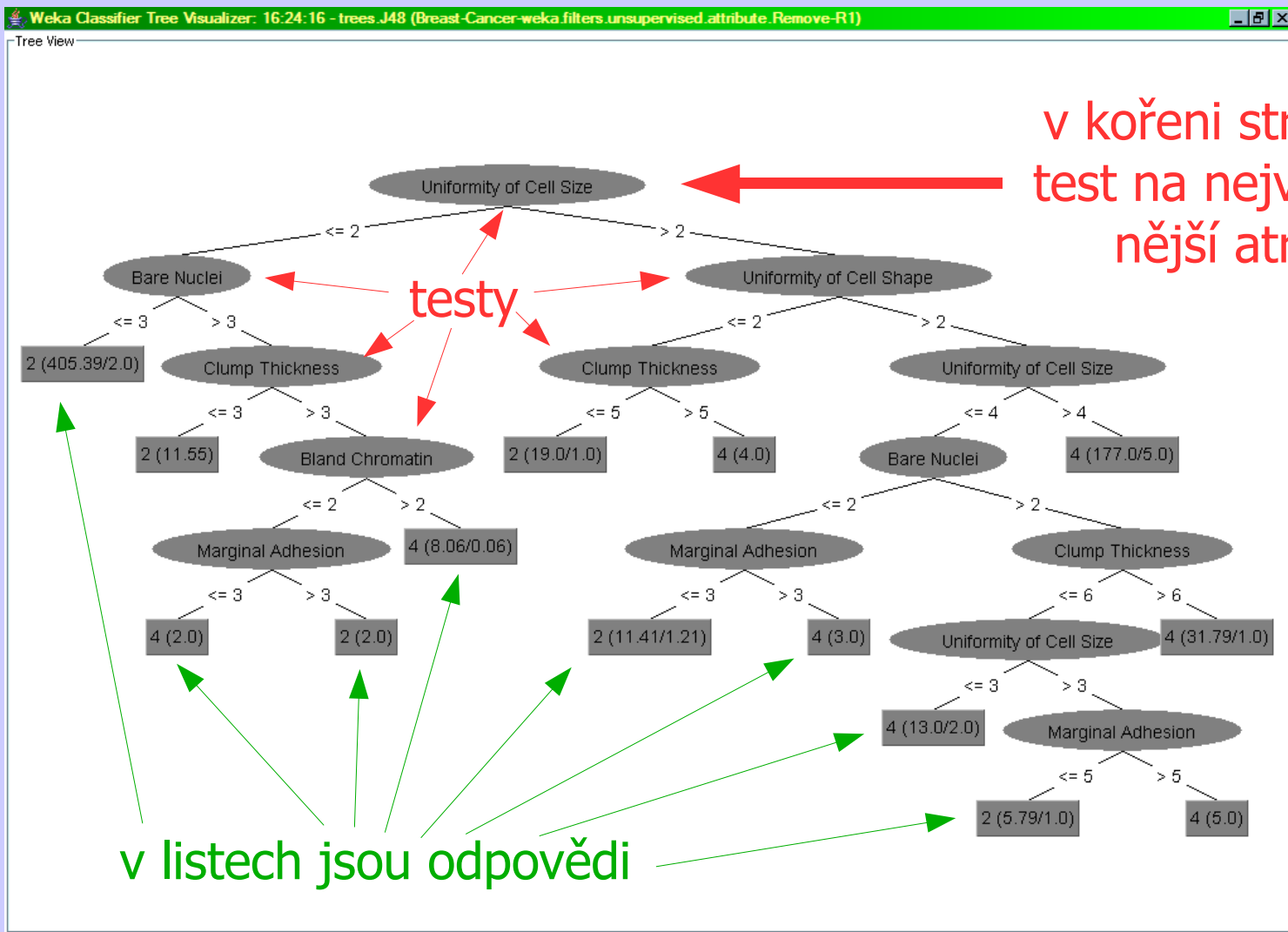
# MATEMATICKÁ BIOLOGIE & ICT

Lze zobrazit třeba i klasifikační chyby jednotlivých příkladů pro zvolené atributy (□ je chybně, x je správně):



# MATEMATICKÁ BIOLOGIE & ICT

Příklad automaticky generovaného rozhodovacího stromu pro reálná data *Wisconsin breast-cancer* (klasifikace dle vlastností odebraného vzorku buněk) algoritmem J48 systému WEKA:



# MATEMATICKÁ BIOLOGIE & ICT

Obdobný systém YALE (Yet Another Learning Environment) také umožňuje vytvořit složitý proces dolování z dat:

The screenshot displays the YALE software interface. On the left, a tree view shows the project structure: Root, Experiment, Input (ExampleSource), and MultilayerPerceptron. The main window is divided into a 'Key' table and a 'Value' table. The 'Key' table lists various configuration parameters, and the 'Value' table shows their corresponding values. A 'Neural Network' window is open, showing a diagram of a neural network with 6 input nodes (att1 to att6), 3 hidden nodes, and 2 output nodes labeled 'negative' and 'positive'. Below the diagram, a 'Controls' panel includes buttons for 'Start' and 'Accept', and displays training parameters: Epoch 10000, Num Of Epochs 10000, Error per Epoch = 0.0175566, Learning Rate = 0.3, and Momentum = 0.1. A status bar at the bottom shows the current task as '[1] MultilayerPerceptron 125 s' and the time as '23:08:25'.

Key	Value
configure_operator	Start configuration wizard...
attributes	C:\Program Files\YALE\yale-3.4\sample\data\weighting.xml Edit ...
sample_ratio	1.0
sample_size	
datamanagement	
column_separators	
comment_chars	
decimal_point_character	
use_quotes	
permutate	
local_random_seed	



# MATEMATICKÁ BIOLOGIE & ICT

Optimalizace genetickými algoritmy umožňuje mj. řešit úlohy, které lze převést na *problém obchodního cestujícího*, např. hledat nejúčinnější a neekonomičtější stanovení druhů a pořadí testů vyšetření:

**GA-TSP v0.1 (Genetic algorithm for Travelling Salesman Problem) - Copyright 2004 Tomáš Černý / mazy**

**Zobrazení**

Tlustě  
 Graf

Generací: 1  
Zobrazuj po: 1

**Kontrola simulace**

Další generace Start  
Ukonči Začni znovu

Minimální doba [ms]: 100

**Nastavení algoritmu**

Velikost populace (nemění se během simulace): 300

4 x 2 x 1 x 1/2 x x 2 Elita [%]: 5

**Křížení**

Šance [%]: 60

Selekce rodičů: Ruleta podle pořadí

Partially Matched Cross. (PMX)  
 Greedy Subtour Cross. (GSX)  
 Greedy Crossover (GXO)

**Mutace / Evoluční alg. / Heuris.**

Šance [%]: 5

Prohození dvou  
 Inverze podsekvence  
 Částečná 2-Opt 10  
 2-Opt heuristika  
 2-Opt prohození  
 Hladové prohození po cestě  
 Max. jedna mutace na dítě

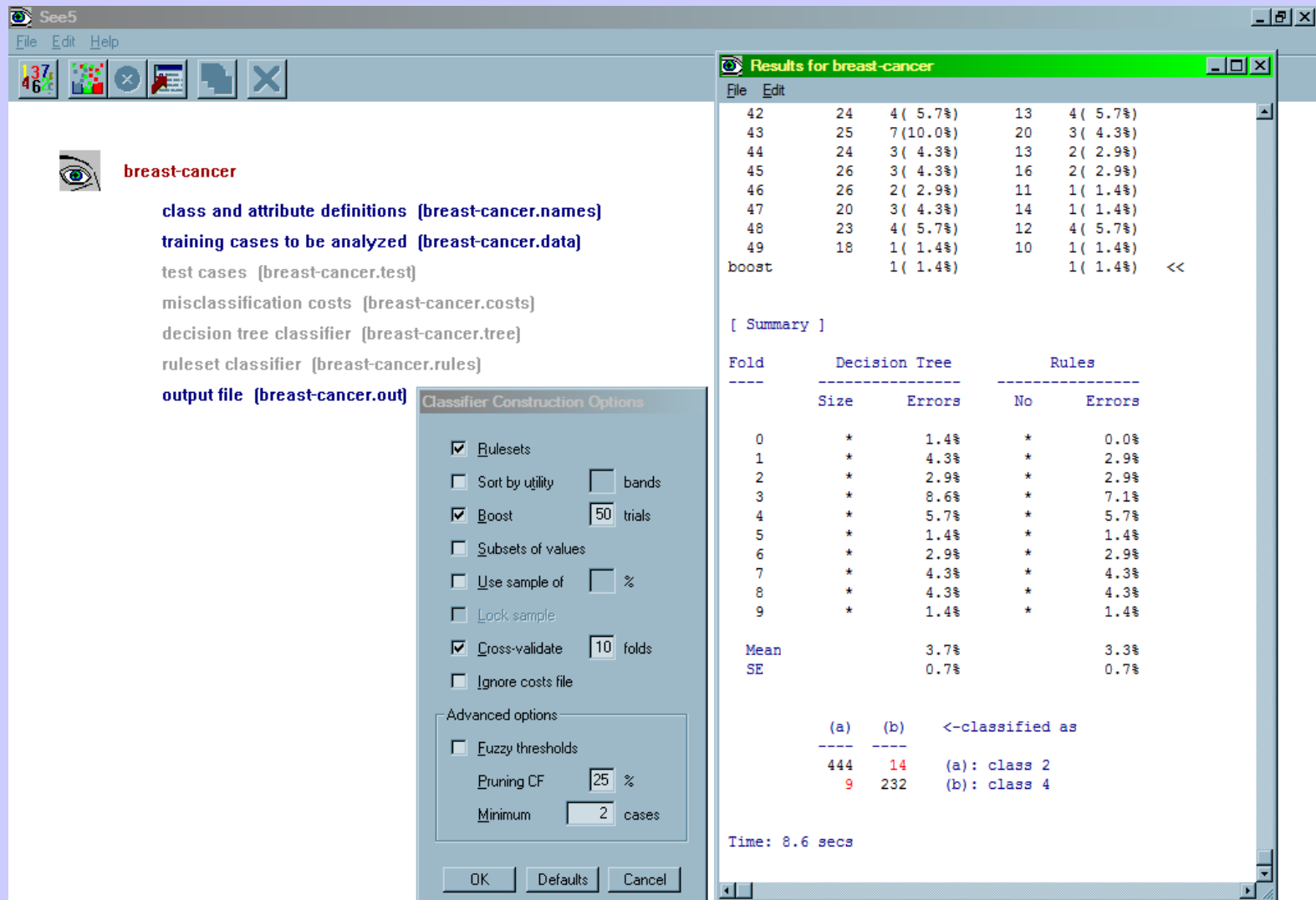
**Informace**

Měst :	100	Křížení :	51766
Generace :	305	Mutací :	8591
Změna v :	204	Nejlepší :	8034
		Počáteční :	45982
		Průměrná :	8124
		Nej. [%] :	17%
		Nej. [% z řešení] :	0%

Info :

# MATEMATICKÁ BIOLOGIE & ICT

Vysoce efektivní profesionální generátor rozhodovacích stromů a pravidel je systém C5/See5, používaný pro různé aplikace:



**breast-cancer**

- class and attribute definitions (breast-cancer.names)
- training cases to be analyzed (breast-cancer.data)
- test cases (breast-cancer.test)
- misclassification costs (breast-cancer.costs)
- decision tree classifier (breast-cancer.tree)
- ruleset classifier (breast-cancer.rules)
- output file (breast-cancer.out)

**Classifier Construction Options**

- Rulesets
- Sort by utility  bands
- Boost  trials
- Subsets of values
- Use sample of  %
- Lock sample
- Cross-validate  folds
- Ignore costs file

**Advanced options**

- Fuzzy thresholds
- Pruning CF  %
- Minimum  cases

**Results for breast-cancer**

File	Edit				
42	24	4 ( 5.7%)	13	4 ( 5.7%)	
43	25	7 (10.0%)	20	3 ( 4.3%)	
44	24	3 ( 4.3%)	13	2 ( 2.9%)	
45	26	3 ( 4.3%)	16	2 ( 2.9%)	
46	26	2 ( 2.9%)	11	1 ( 1.4%)	
47	20	3 ( 4.3%)	14	1 ( 1.4%)	
48	23	4 ( 5.7%)	12	4 ( 5.7%)	
49	18	1 ( 1.4%)	10	1 ( 1.4%)	
boost		1 ( 1.4%)		1 ( 1.4%)	<<

[ Summary ]

Fold	Decision Tree		Rules	
	Size	Errors	No	Errors
0	*	1.4%	*	0.0%
1	*	4.3%	*	2.9%
2	*	2.9%	*	2.9%
3	*	8.6%	*	7.1%
4	*	5.7%	*	5.7%
5	*	1.4%	*	1.4%
6	*	2.9%	*	2.9%
7	*	4.3%	*	4.3%
8	*	4.3%	*	4.3%
9	*	1.4%	*	1.4%
Mean		3.7%		3.3%
SE		0.7%		0.7%

Time: 8.6 secs

(a) (b) <-classified as

444	14	(a): class 2
9	232	(b): class 4