



MACHINE LEARNING, ARTIFICIAL INTELLIGENCE, DATA MINING

MATEMATICKÁ BIOLOGIE & ICT

Moderní systémy pro získávání znaností z informací a dat

Jan Žižka

MATEMATICKÁ BIOLOGIE & ICT

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

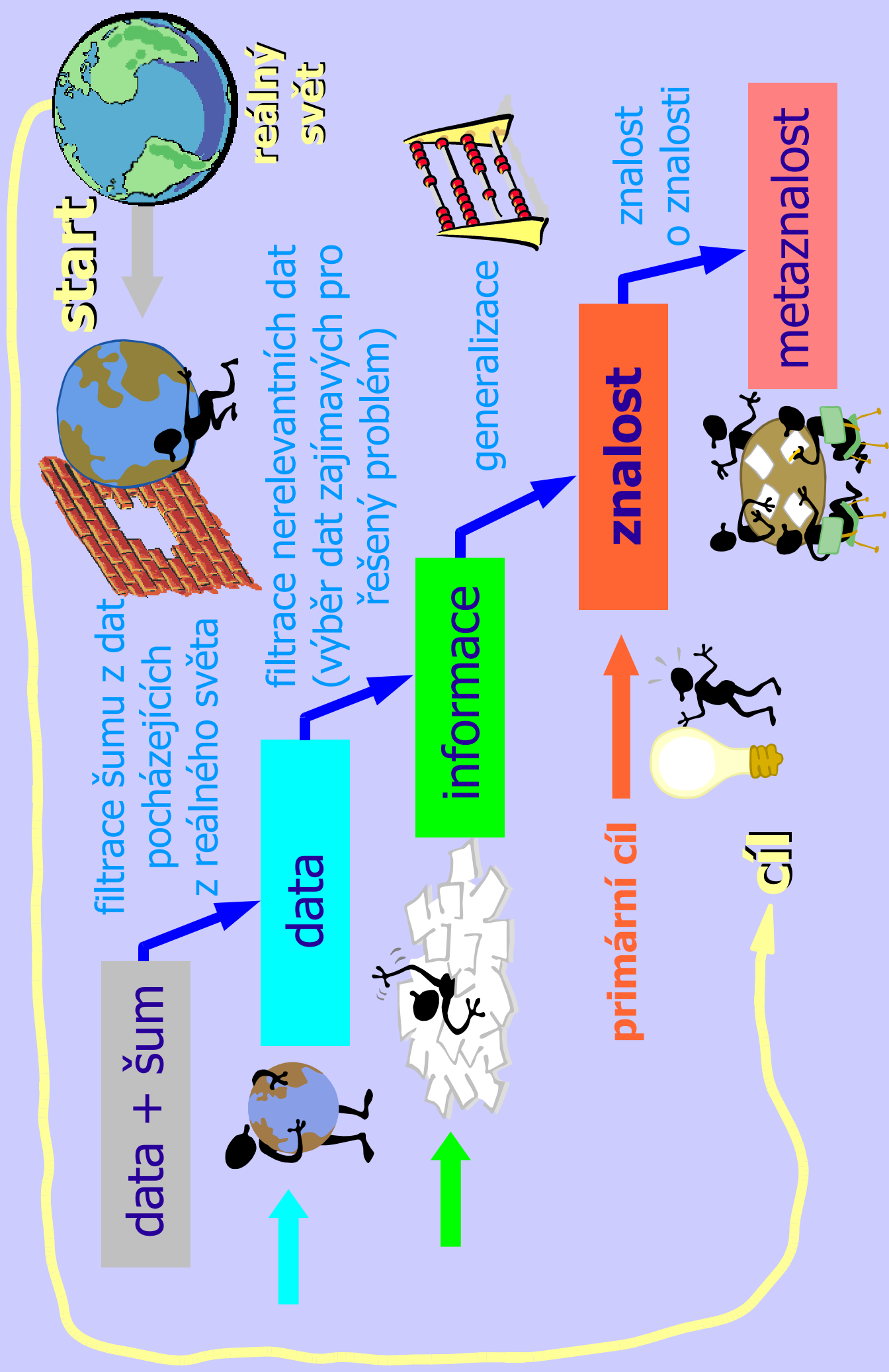
Strojové učení (*machine learning, ML*), **umělá inteligence** (*artificial intelligence, AI*), **dolování z dat** (*data mining*):

Moderní systémy pro zpracování informace a získávání **znalostí** z dat. Rozšiřují a doplňují tradiční aplikace matematických a inforatických 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)

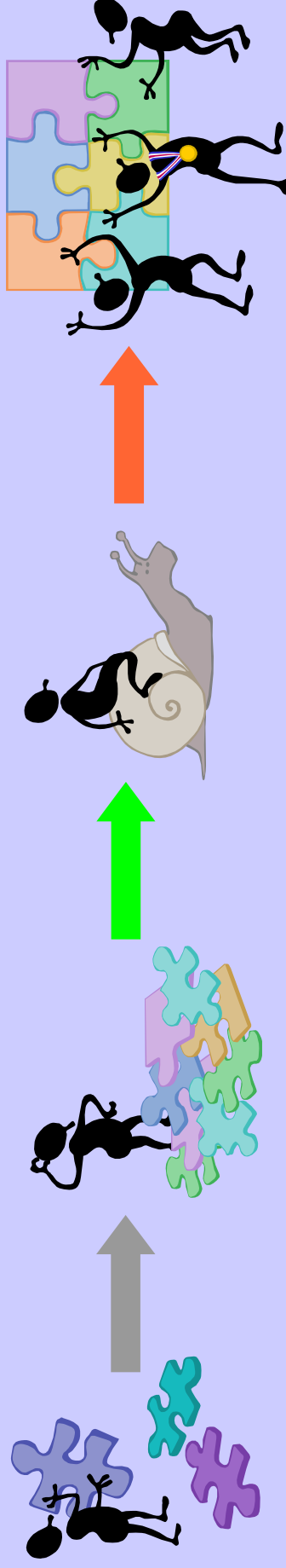


MATEMATICKÁ BIOLOGIE & ICT

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.



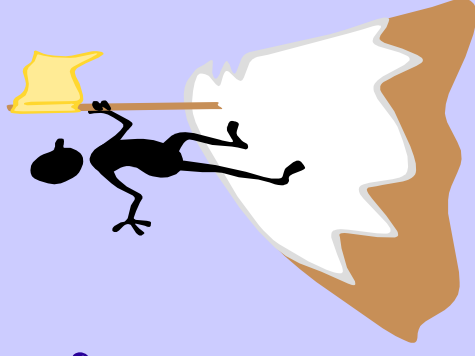
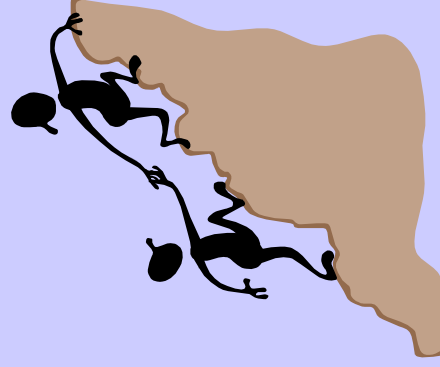
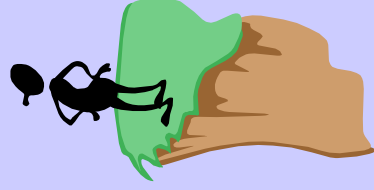
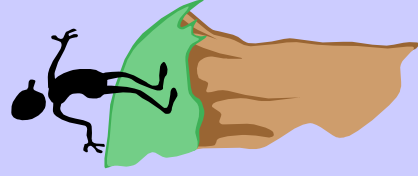
MATEMATICKÁ BIOLOGIE & ICT

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*.

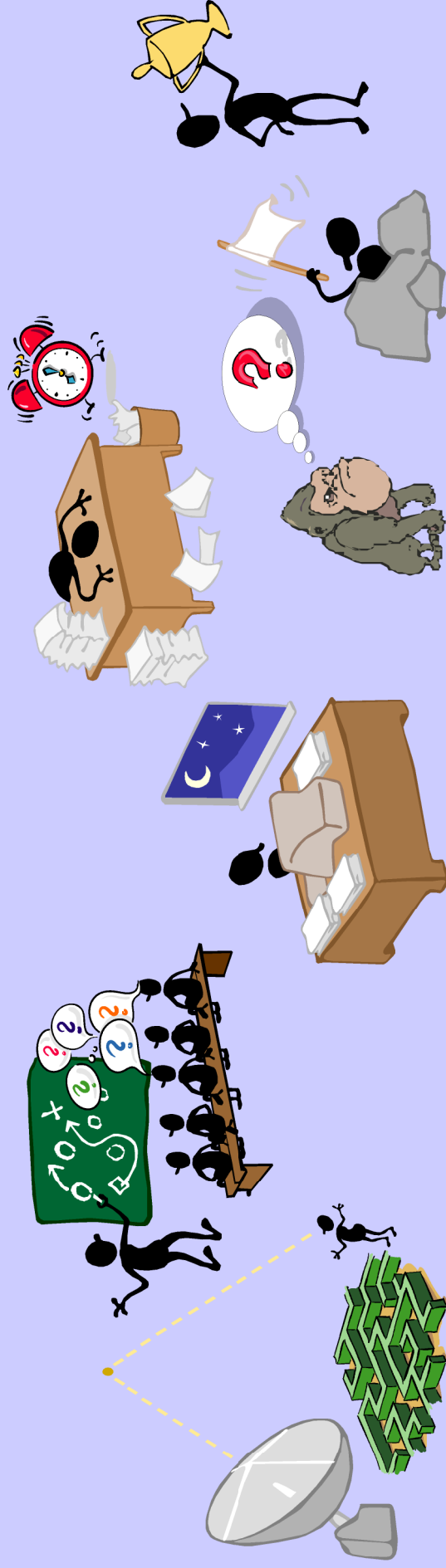
Hledání skutečné znalosti v datech se často podobá hledání nejvyššího vrcholku kopce ve velmi zvlněné zmlžené krajině (lokální extrém, globální extrém, nelinearita, nespojitě funkce, apod.).



MATEMATICKÁ BIOLOGIE & ICT

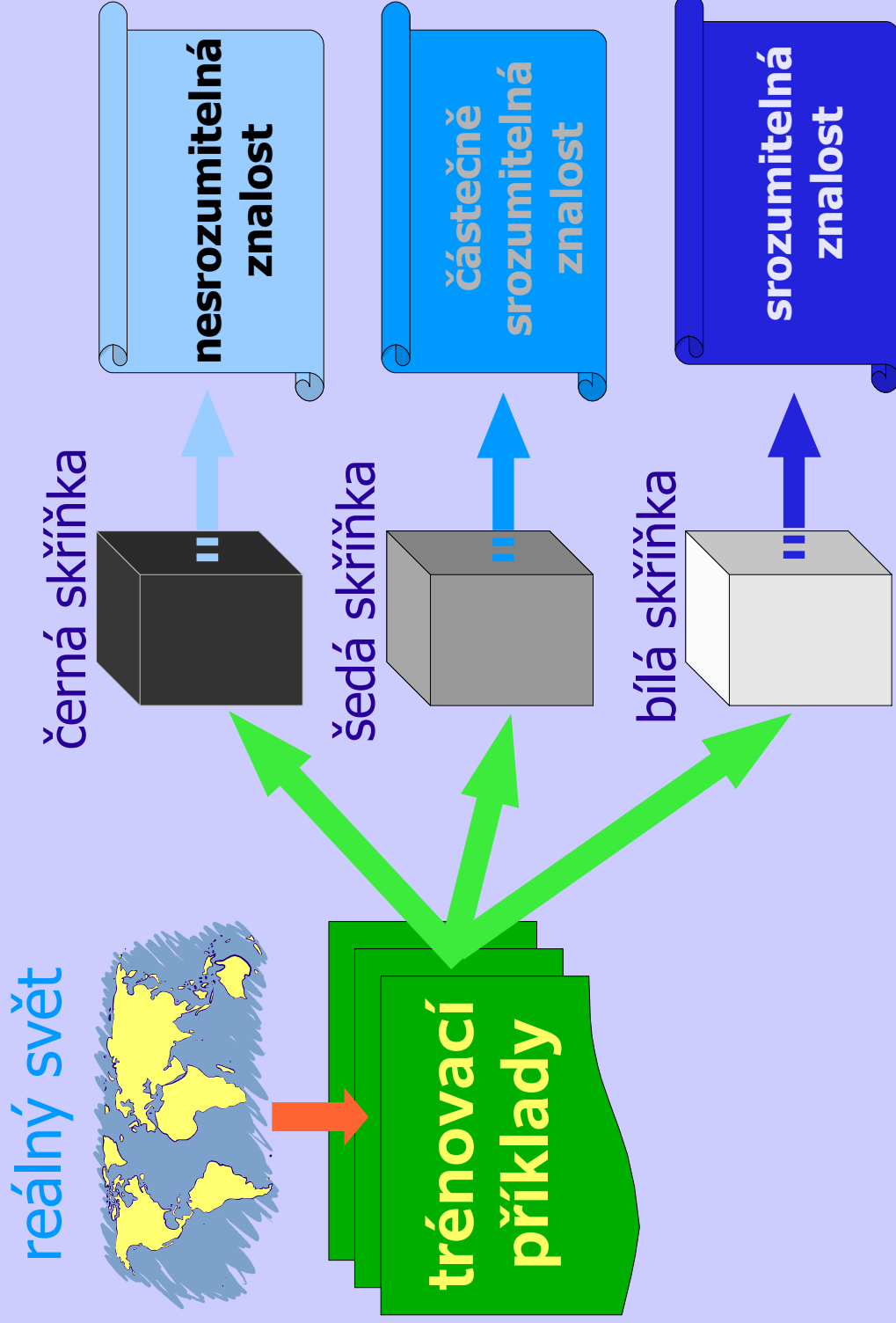
„Vytěžit“ použitelnou znalost ze „surových“ dat vyžaduje pochopit vlastnosti disponibilních metod, navrhnout a provést řadu časově náročných experimentů (výpočetní složitost – čas a paměť) a správně interpretovat získané znalosti pro jejich použití.

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



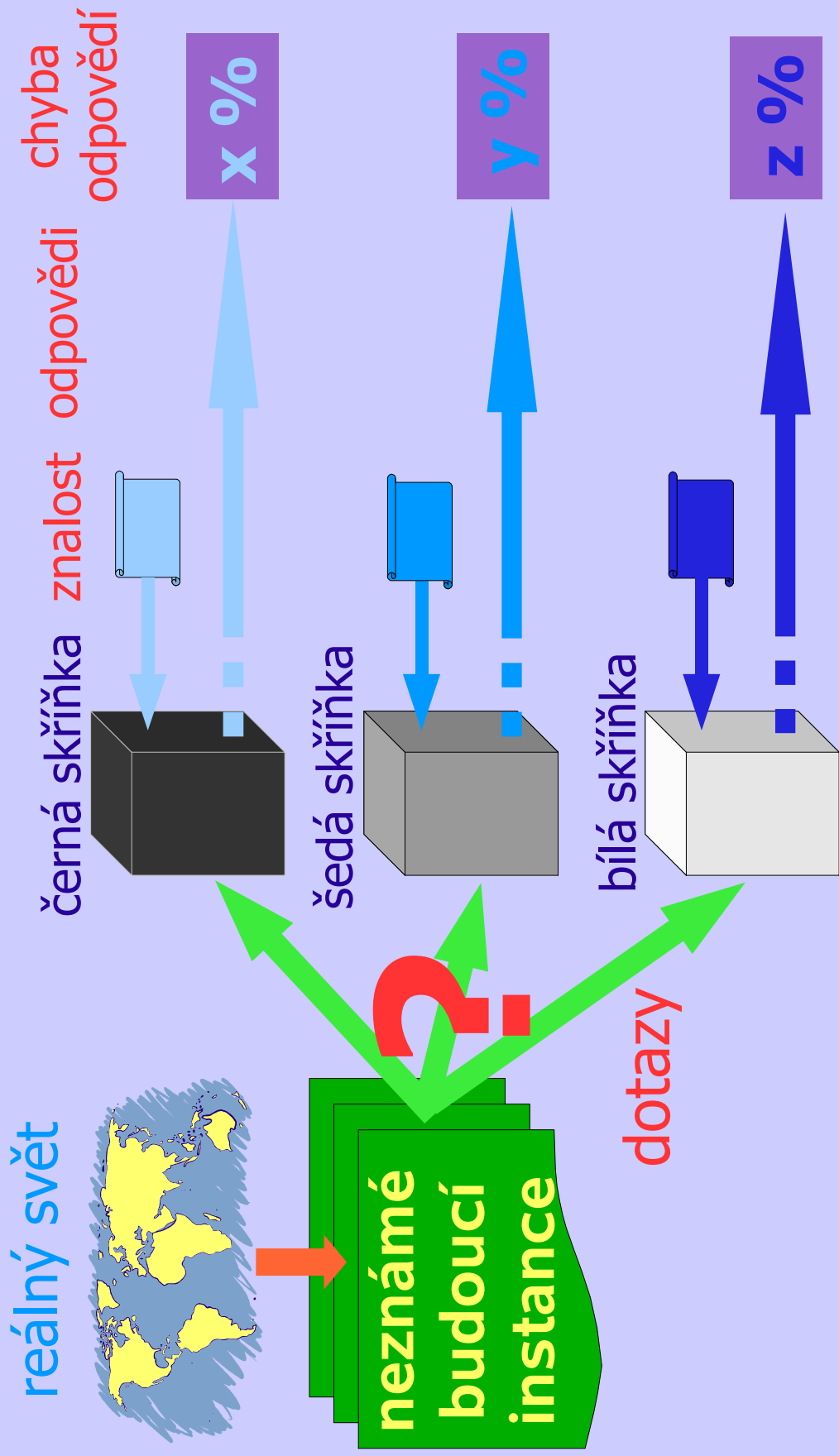
MATEMATICKÁ BIOLOGIE & ICT

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



MATEMATICKÁ BIOLOGIE & ICT

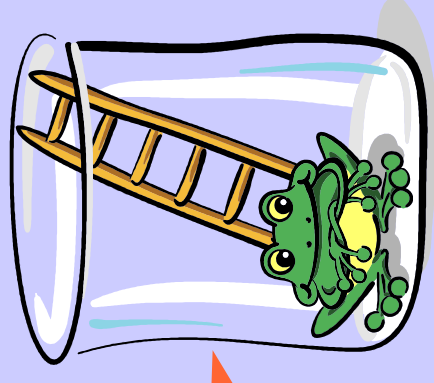
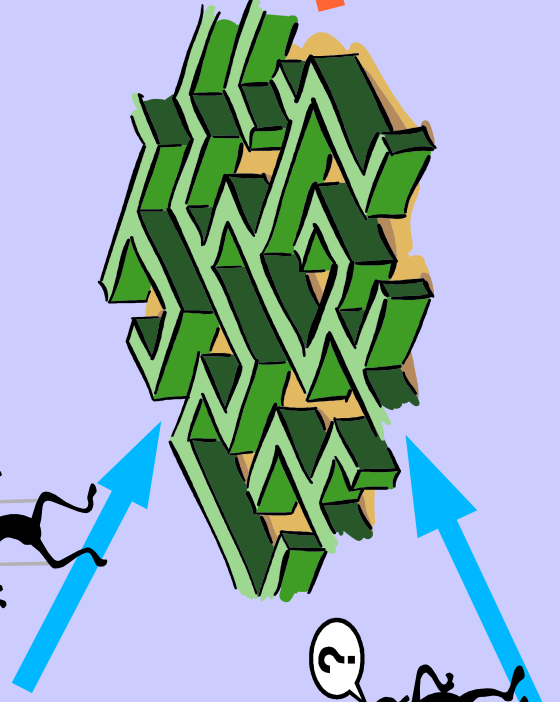
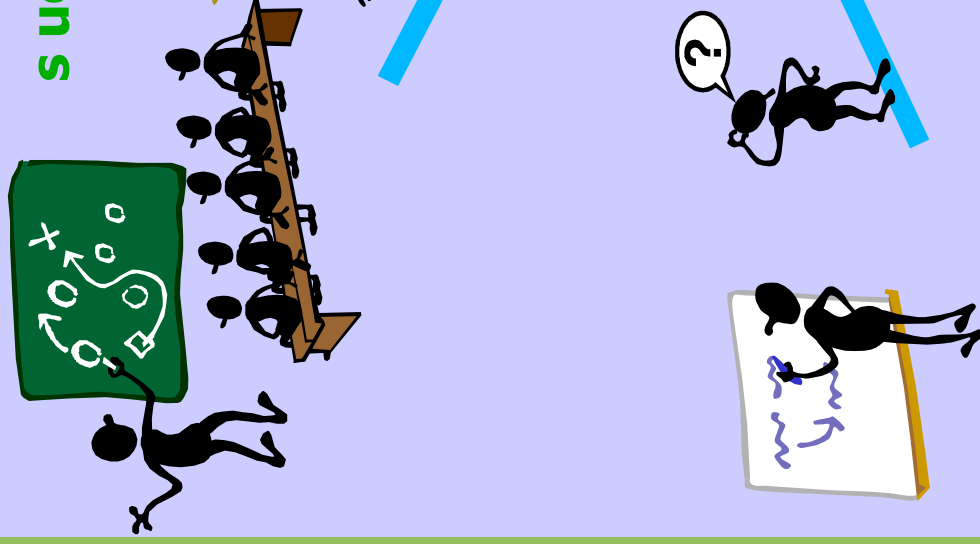
Funkčnost algoritmů ovšem nemusí (ale i může) odpovídat srozumitelnosti znalosti získané trénováním:



MATEMATICKÁ BIOLOGIE & ICT

Algoritmy lze také rozdělit podle typu učení:

s učitelem (zpětná vazba, oprava chyb vně algoritmu: např. umělé neuronové sítě trénované zpětným šířením chyb)



Predikce pro případy neznámé při trénování

bez učitele (oprava chyb uvnitř algoritmu: např. shlukování, Kohonenovy mapy, adaptivní rezonanční teorie)

MATEMATICKÁ BIOLOGIE & ICT

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

Diagram illustrating the structure of a data table with annotations:

- Annotations:
 - ← jeden z atributů (points to the 'Bland Chromatin' column)
 - ← názvy atributů (points to the column headers)
 - ← klasifikační třída (points to the 'Class' column)
 - ← jeden z příkladů (points to a row in the table)

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	3	3	10	5	10	5	4	4	4

(Wisconsin breast-cancer data)

MATEMATICKÁ BIOLOGIE & ICT

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

The screenshot displays the Weka Explorer interface. The top-left pane shows the 'Weka GUI Chooser' with options: Simple CLI, Experiment, ArrffViewer, Explorer, KnowledgeFlow, and Log. The main window is titled 'Weka Explorer' and contains several panels:

- Preprocess**: Buttons for Classify, Cluster, Associate, Select attributes, and Visualize.
- Open file...**, **Open URL...**, **Open DB...**, **Generate...**, **Undo**, **Edit...**, **Save...**
- Filter**: Set to 'None'.
- Current relation**: 'Breast-Cancer-weka.filters.unsupervised.attribute.Remove-R.1', Instances: 699, Attributes: 10.
- Attributes**: A list of 10 attributes with checkboxes. Attribute 1, 'Clump Thickness', is selected.
- Selected attribute**: 'Clump Thickness', Type: Numeric, Unique: 0 (0%), Distinct: 10. A table shows statistics: Minimum (1), Maximum (10), Mean (4.418), StdDev (2.816).
- Class: Class (Nom)**: A bar chart showing the distribution of the 'Class' attribute. The x-axis ranges from 1 to 10. The y-axis shows counts: 195 (Class 1), 108 (Class 2), 80 (Class 3), 130 (Class 4), 34 (Class 5), 23 (Class 6), 46 (Class 7), 83 (Class 8).
- Status**: OK.

MATEMATICKÁ BIOLOGIE & ICT

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

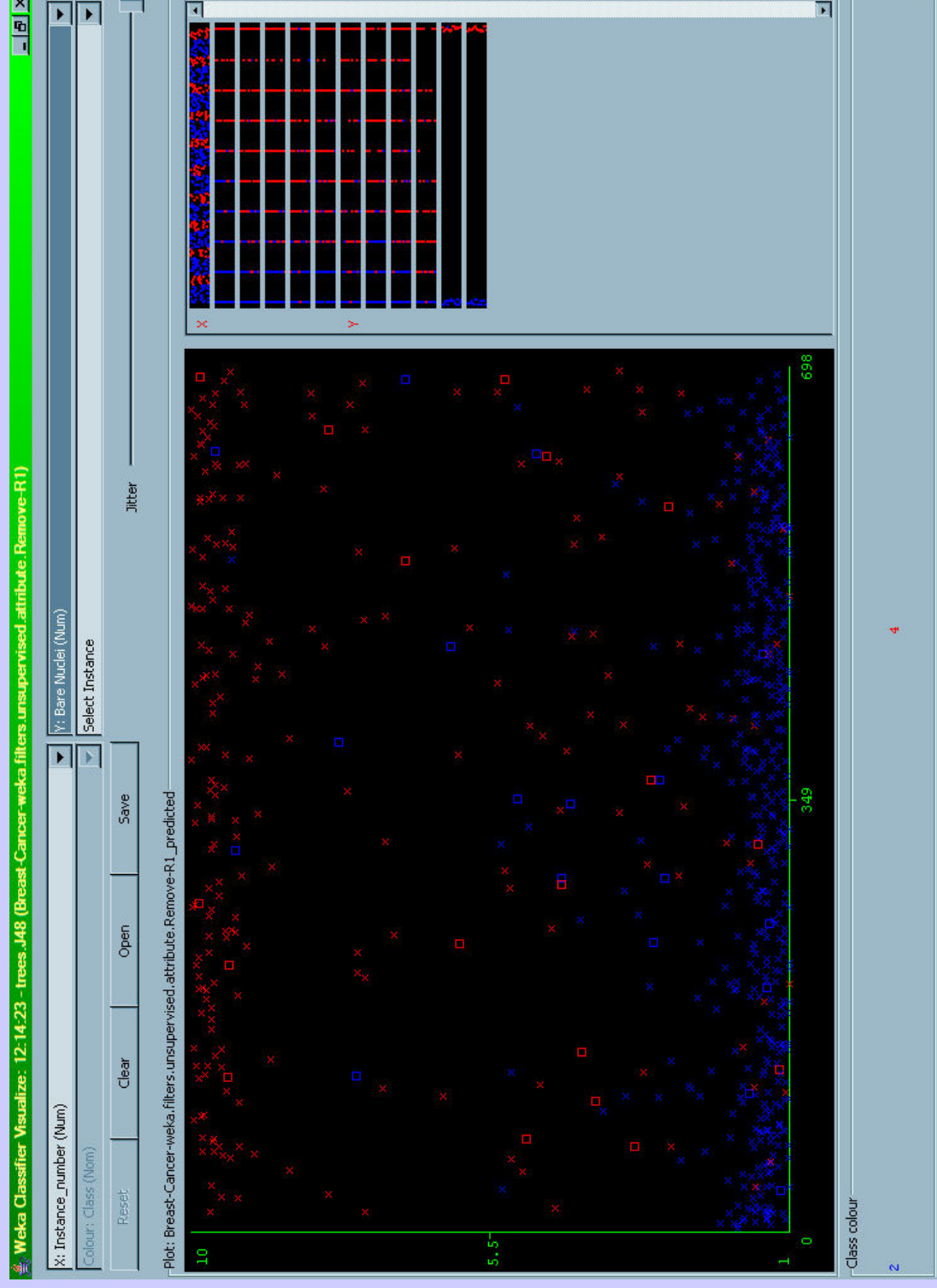
Viewer

Relation: Breast-Cancer

No.	Sample code number Numeric	Clump Thickness Numeric	Uniformity of Cell Size Num	Uniformity of Cell Shape Numeric	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		Get mean...	1.0	2.0	1.0	3.0	1.0	1.02	
2	1002945.0	5.0		Set all values to...	5.0	7.0	10.0	3.0	3.0	2.0	1.02
3	1015425.0	3.0		Set missing values to...	1.0	2.0	2.0	3.0	3.0	1.0	1.02
4	1016277.0	6.0		Replace values with...	1.0	3.0	4.0	3.0	3.0	7.0	1.02
5	1017023.0	4.0		Rename attribute...	3.0	2.0	1.0	3.0	3.0	1.0	1.02
6	1017122.0	8.0		Attribute as Class	8.0	7.0	10.0	9.0	7.0	7.0	1.04
7	1018099.0	1.0		Delete attribute	1.0	2.0	10.0	3.0	3.0	1.0	1.02
8	1018561.0	2.0		Delete attributes...	1.0	2.0	1.0	1.0	1.0	1.0	1.02
9	1033078.0	2.0		Sort data (ascending)	1.0	2.0	1.0	2.0	1.0	1.0	5.02
10	1033078.0	4.0		Optimal column width (current)	1.0	1.0	1.0	3.0	3.0	1.0	1.02
11	1035283.0	1.0		Optimal column width (all)	1.0	2.0	1.0	2.0	2.0	1.0	1.02
12	1036172.0	2.0			1.0	2.0	1.0	2.0	2.0	1.0	1.02
13	1041801.0	5.0			3.0	2.0	3.0	4.0	4.0	4.0	1.04
14	1043999.0	1.0	1.0		1.0	2.0	3.0	3.0	3.0	1.0	1.02
15	1044572.0	8.0	5.0		10.0	7.0	9.0	5.0	7.0	5.0	4.04
16	1047630.0	7.0	4.0		4.0	6.0	1.0	4.0	4.0	3.0	1.04
17	1048672.0	4.0	1.0		1.0	2.0	1.0	2.0	1.0	1.0	1.02
18	1049815.0	4.0	1.0		1.0	2.0	1.0	3.0	3.0	1.0	1.02
19	1050670.0	10.0	7.0		6.0	4.0	10.0	4.0	4.0	1.0	2.04
20	1050718.0	6.0	1.0		1.0	2.0	1.0	3.0	3.0	1.0	1.02
21	1054590.0	7.0	3.0		10.0	5.0	10.0	5.0	5.0	4.0	4.04
22	1054593.0	10.0	5.0		3.0	6.0	7.0	7.0	7.0	10.0	1.04
23	1055784.0	3.0	1.0		1.0	2.0	1.0	2.0	1.0	1.02	
24	1057013.0	8.0	4.0		1.0	2.0	2.0	7.0	7.0	3.0	1.04
25	1059552.0	1.0	1.0		1.0	2.0	1.0	1.0	3.0	1.0	1.02
26	1065726.0	5.0	2.0		4.0	2.0	7.0	3.0	3.0	6.0	1.04
27	1066373.0	3.0	2.0		1.0	1.0	1.0	1.0	2.0	1.0	1.02
28	1066979.0	5.0	1.0		1.0	2.0	1.0	2.0	2.0	1.0	1.02
29	1067444.0	2.0	1.0		1.0	2.0	1.0	1.0	2.0	1.0	1.02
30	1070935.0	1.0	1.0		1.0	2.0	1.0	1.0	1.0	1.0	1.02
31	1070935.0	3.0	1.0		1.0	1.0	1.0	1.0	2.0	1.0	1.02
32	1071760.0	2.0	1.0		1.0	2.0	1.0	1.0	3.0	1.0	1.02
33	1072179.0	10.0	7.0		3.0	8.0	5.0	7.0	4.0	4.0	3.04
34	1074610.0	2.0	1.0		2.0	2.0	1.0	3.0	3.0	1.0	1.02
35	1075123.0	3.0	1.0		1.0	2.0	1.0	1.0	2.0	1.0	1.02
36	1079304.0	2.0	1.0		1.0	2.0	1.0	2.0	2.0	1.0	1.02
37	1080185.0	10.0	10.0		8.0	6.0	1.0	8.0	9.0	9.0	1.04
38	1081791.0	6.0	2.0		1.0	2.0	1.0	1.0	7.0	1.0	1.02
39	1084584.0	5.0	4.0		9.0	2.0	10.0	5.0	6.0	6.0	1.04

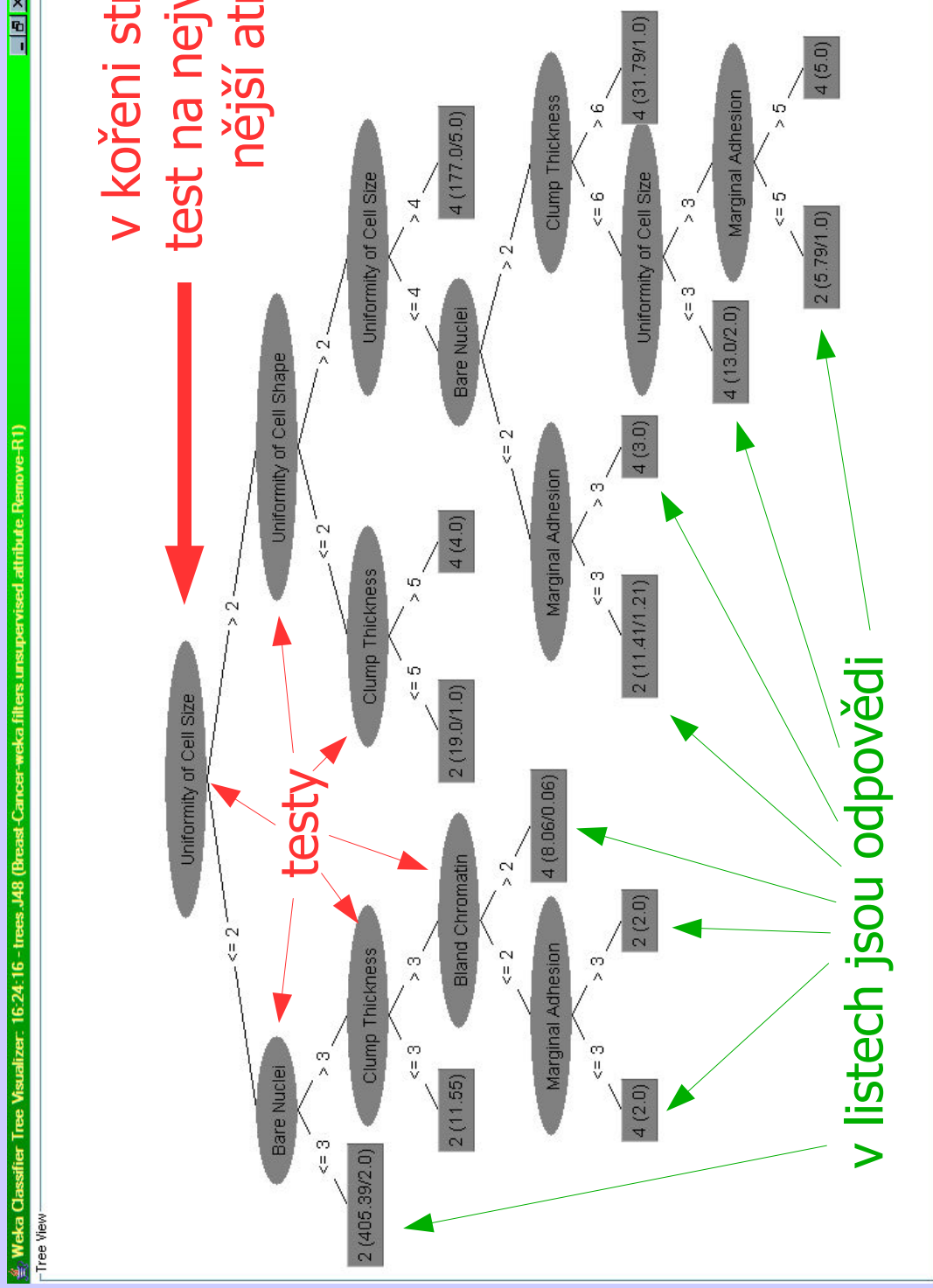
Undo OK Cancel

Lze zobrazit třeba i klasifikační chyby jednotlivých příkladů pro zvolené atributy (\square je chybně, \times 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 for configuring a neural network. The main window is titled "Start configuration wizard..." and shows a table with the following data:

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

Below the table, a "Neural Network" diagram is shown, featuring six input nodes (att1 to att6) and two output nodes (negative and positive). The diagram is a fully connected feedforward network. To the right of the diagram is a "Controls" panel with the following settings:

- Epoch: 10000
- Num Of Epochs: 10000
- Error per Epoch: 0.0175566
- Learning Rate: 0.3
- Momentum: 0.1

At the bottom of the interface, a log window shows the following output:

```

12.11.2006 23:06:20: Experiment initialised
12.11.2006 23:06:20: Experiment starts
12.11.2006 23:06:20: Experiment:
Root[0] (Experiment)
+- Input[0] (ExampleSource)
+- MultiLayerPerceptron[0] (MultiLayerPerceptron)
  
```

The status bar at the bottom right indicates "[1] MultiLayerPerceptron 125 s" and the time "23:08:25".

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

Ilustrace Start

Graf Ukončí Začni znovu

Generací: 1 Minimální doba [ms]: 100

Zobrazuji po: 1

Nastavení algoritmu

Velikost populace (nemění se během simulace): 300

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

Mutace / Evoluční alg. / Heuris: Šance [%]: 5

Selekce rodičů: Ruleta podle pořadí

Prohození dvou: Inverze podsekvence: Částečná 2-Opt: 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%

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)

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

(a) (b) <-classified as
 444 14 (a): class 2
 9 232 (b): class 4

Time: 8.6 secs

Classifier Construction Options

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

Advanced options
 Fuzzy thresholds
 Pruning CF 25 %
 Minimum 2 cases

Umělé neuronové sítě mohou v řadě případů nalézt ve složitých mnohorozměrných prostorech oddělovací hranice mezi určitými skupinami datových instancí. Hranice může být tvořena velmi komplikovanou nelineární funkcí, která nemusí být hladká, spojitá, apod., a kterou nelze analytickými metodami odhadnout ani přibližně.

Podobně jako u dalších algoritmů strojového učení, návrh efektivní umělé neuronové sítě není snadný a hledání správných parametrů, včetně architektury sítě, bývá časově náročné.

Podarí-li se však najít přijatelné řešení, natrénovaný algoritmus poskytuje kvalitní a rychlou podporu při zkoumání budoucích, v době tréninku neznámých datových instancí – to platí obecně i pro ostatní algoritmy, i když často v různé míře.

MATEMATICKÁ BIOLOGIE & ICT

The screenshot displays the NetVisualiser software interface. At the top, the title bar reads "Untitled - NetVisualiser". The menu bar includes "Zobrazit", "Data", "Síť", "Učení", "Nápověda", and "Konec". The toolbar contains various icons for file operations, network editing, and training. The main workspace is divided into three panels:

- Struktura neuronové sítě:** A diagram of a neural network with 1 input node, 2 hidden nodes, and 2 output nodes. Connections are shown in red and green.
- Parametry sítě:** A control panel with the following settings:
 - Učící konstanta: 0.6
 - Moment: 0.05
 - Nepoužito: 0
 - Nepoužito: 0Buttons for "Použít" are present at the bottom of the panel.
- Chybová funkce:** A graph showing the error function. The y-axis is labeled "Chyba" and ranges from -1.000 to 1.000. The x-axis represents iterations. The graph shows two data series: "Trenovací data" (blue squares) and "Testovací data" (red squares). The error values are currently 0.0 for both. Buttons for "Použít", "Smazat", and "Překreslit" are located at the bottom of the graph.

At the bottom right of the interface, the text "Iterace: 0, Příchod daty: 0" is displayed.

MATEMATICKÁ BIOLOGIE & ICT

Intuitivně očekávaná hranice

Struktura neuronové sítě

Parametry sítě

- Učící konstanta: 0.6
- Moment: 0.05
- Nepoužito: 0
- Nepoužito: 0

Chybová funkce

Chyba: -1.000

100 Použít
Smazat
Překreslit

Trenovací data
Testovací data

Delta učení

Iterace: 0, Příchod daty: 0

MATEMATICKÁ BIOLOGIE & ICT

**náhodný odhad
sítě na počátku**

Struktura neuronové sítě

Parametry sítě

Učící konstanta	0.6
Moment	0.05
Nepoužito	0
Nepoužito	0

Chybová funkce

Chyba	0.040
100	100

Trénovací data 0.03368
Testovací data 0.00000

Délka učení

MATEMATICKÁ BIOLOGIE & ICT

sítí nalezená hranice na konci výpočtu

Struktura neuronové sítě

Parametry sítě

Učební konstanta	0.1	<input type="checkbox"/>
Moment	0.01	<input type="checkbox"/>
Nepronážitelná	0	<input checked="" type="checkbox"/>
Nepronážitelná	0	<input checked="" type="checkbox"/>

Chybová funkce

Chyba	0.040		
100	2100	4100	6100
Trenovací data	0.00045		
Testovací data	0.00000		

iterace: 887262, Průchod daty: 7783

MATEMATICKÁ BIOLOGIE & ICT

Metod a algoritmů pro vyhledávání znalosti z dat a z informace existují minimálně desítky, s modifikacemi stovky a více.

V současné době i v dohledné budoucnosti jsou a budou tyto nástroje intenzivně rozvíjeny a aplikovány. Důvodem je extrémně silný nárůst množství dat v nejrůznějších oborech a zároveň potřeba tato data nejen ukládat, ale i netriviálním způsobem zpracovávat pomocí strojů – lidé je zpracovávat nemohou kvůli obrovskému rozsahu a složitosti.

Na lidech je ovšem nalézt metody zpracování a vyhodnotit výsledky včetně rozhodnutí, co, jak, kdy a kde použít.

Dolování znalosti z dat je složitý a časově náročný proces, kde neuvážená, povrchní aplikace algoritmů bez jejich pochopení může vést ke špatným výsledkům v realitě.