Moderní systémy pro získávání znalostí z informací a dat

Jan Žižka

Matematická biologie & ICT

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

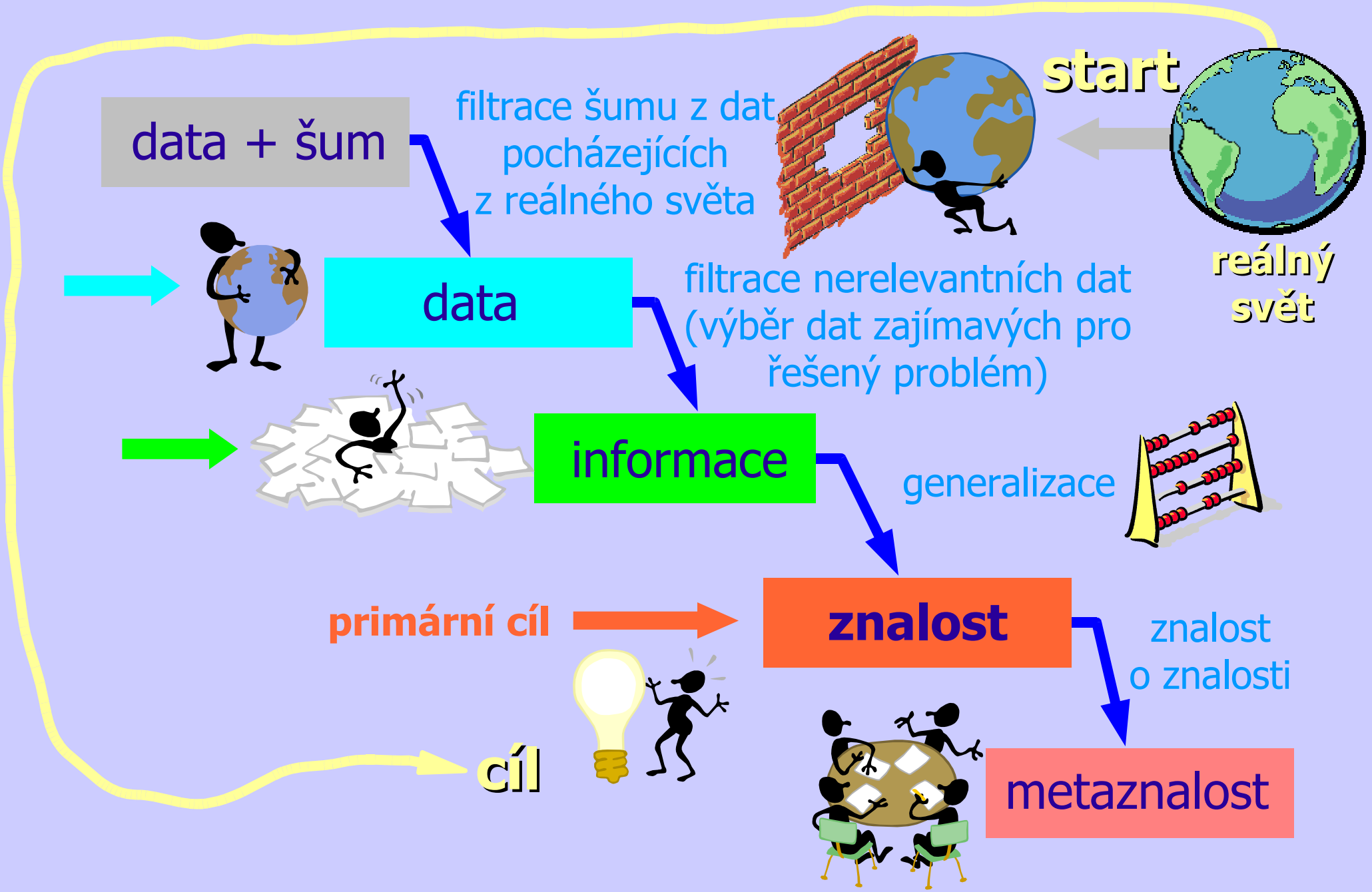
Strojové učení *(machine learning, ML),* ***umělá inteligence*** *(artificial intelligence, AI),* ***dolování z dat*** *(data mining):*

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

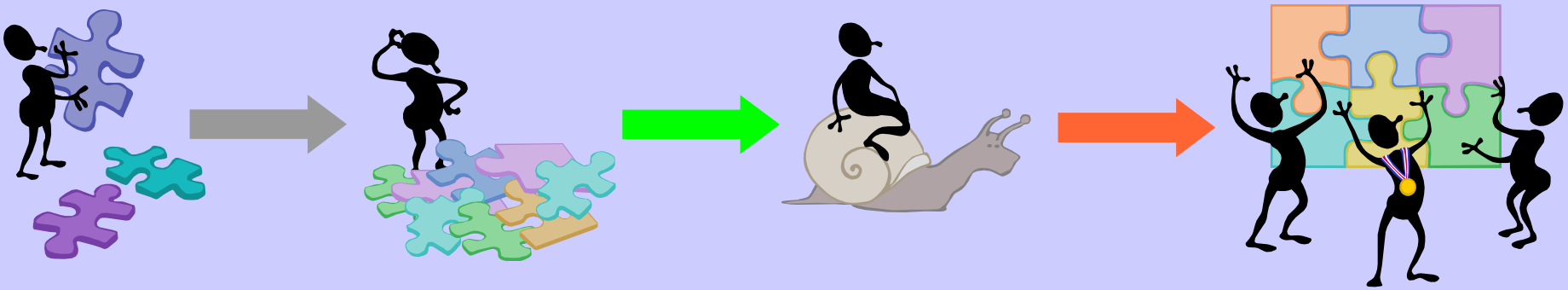


Matematická biologie & ICT

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.

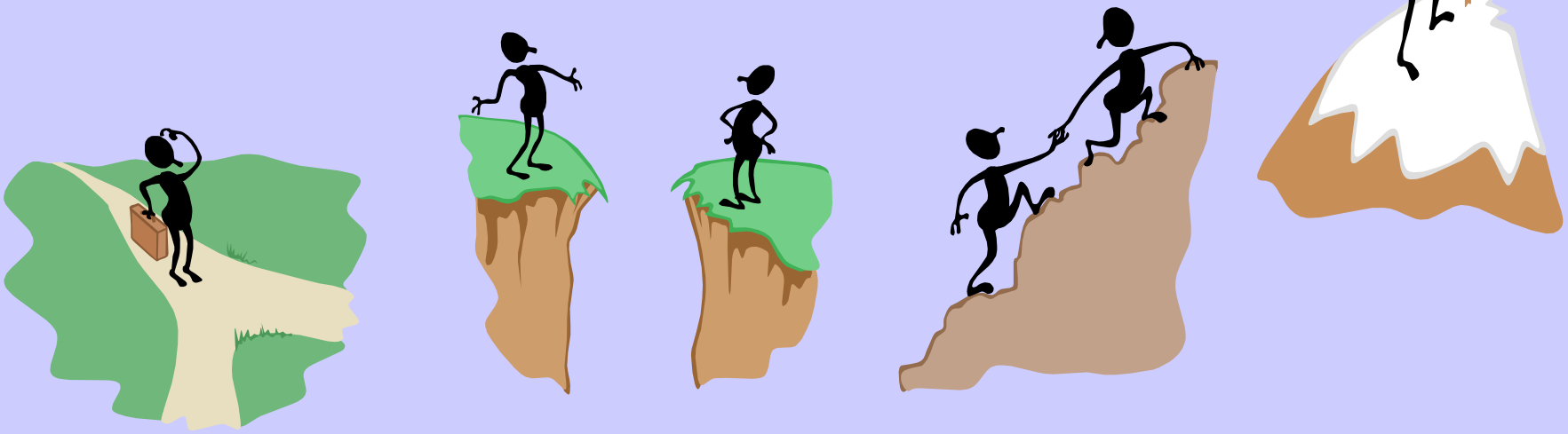
Dolování znalostí z dat zahrnuje přípravu dat, hledání účinného algoritmu pro zobecnění, a nakonec interpretaci.



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 zamlžené krajině (lokální extrémy, globální extrém, nelinearita, nespojité funkce, apod.).



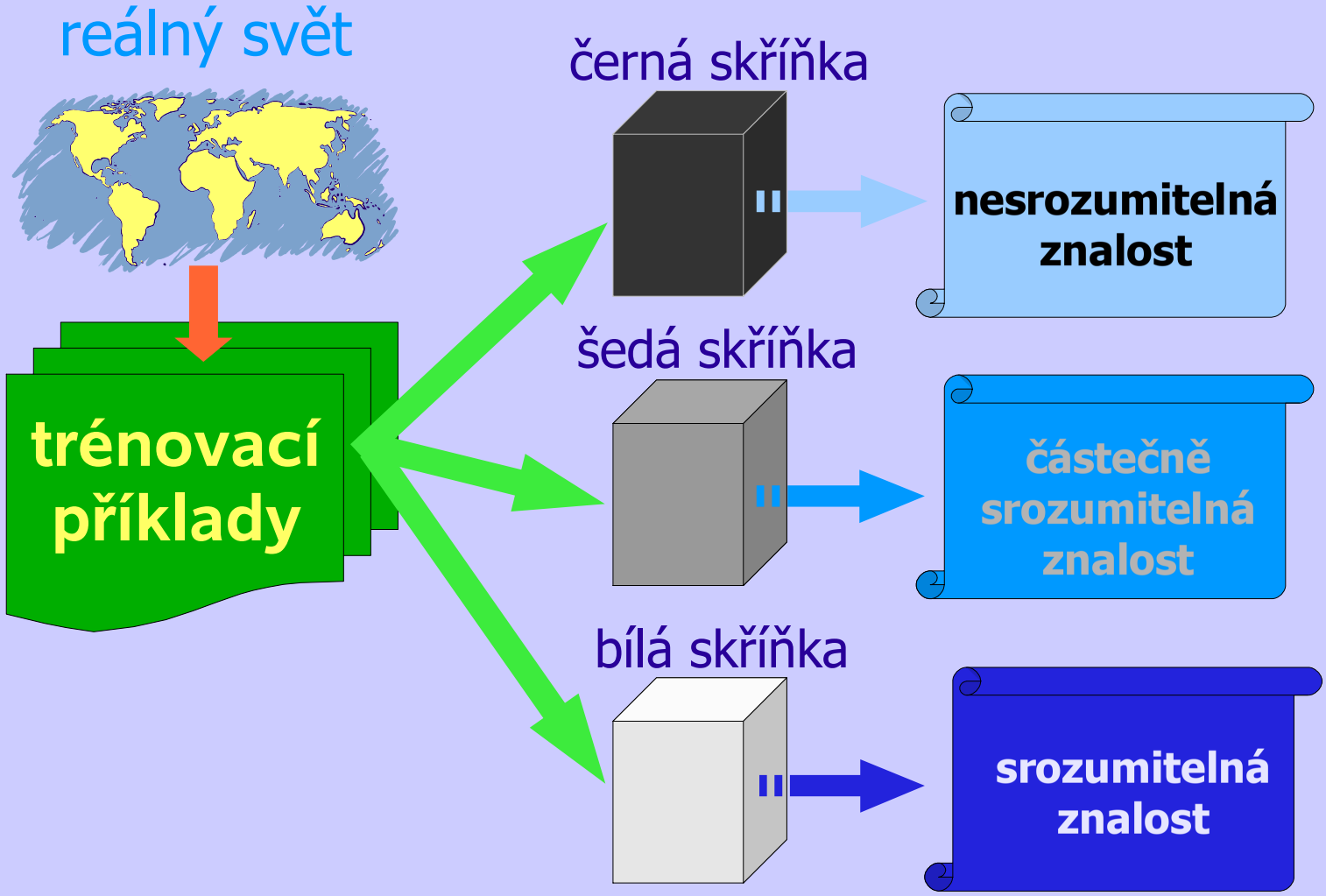
Matematická biologie & ICT

„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.

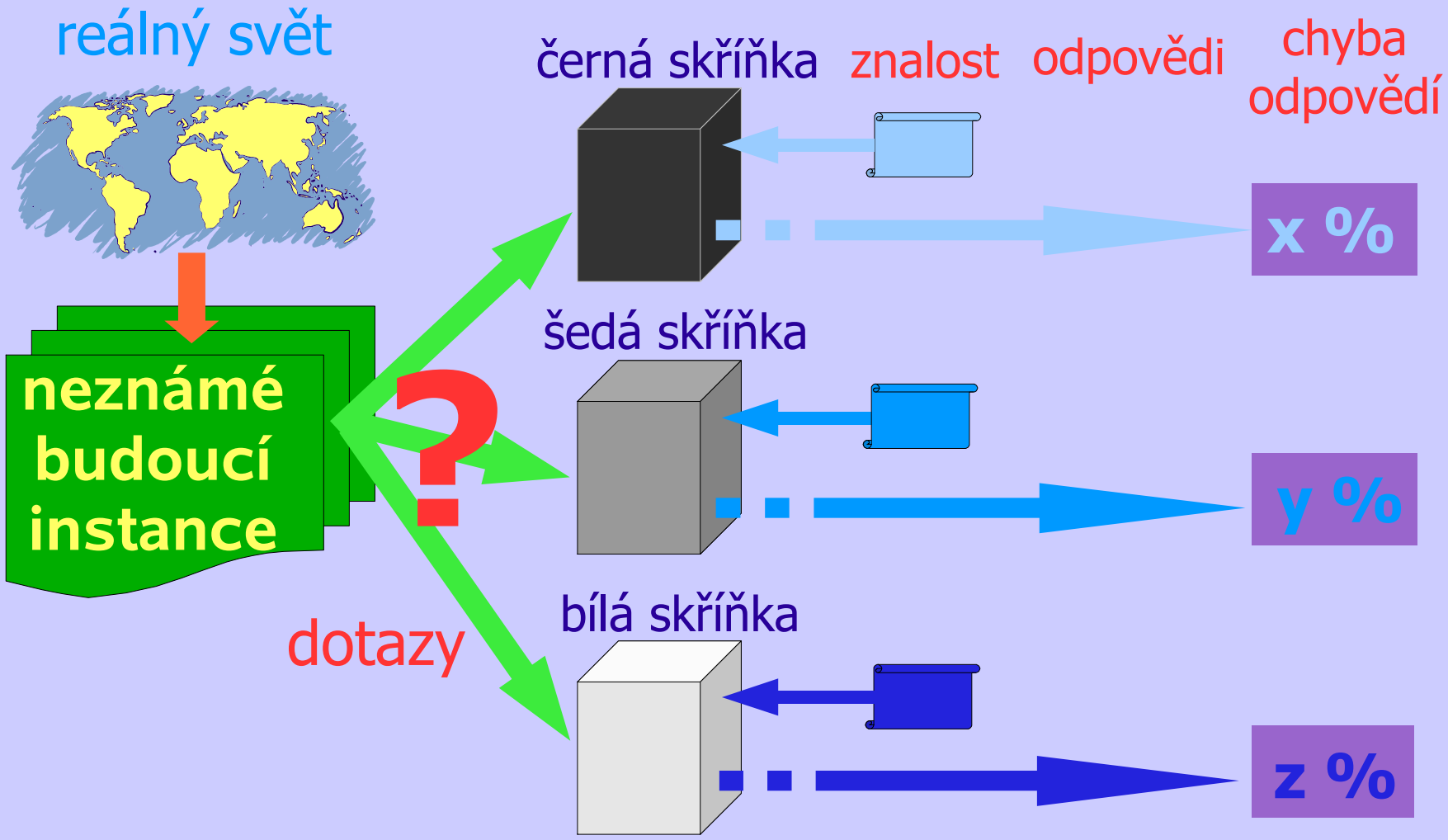
Matematická biologie & ICT

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



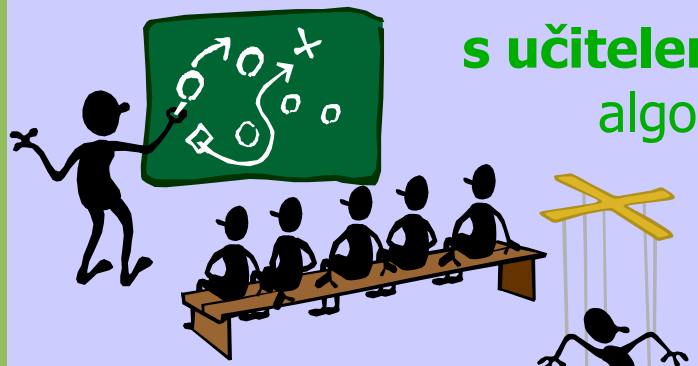
Matematická biologie & ICT

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

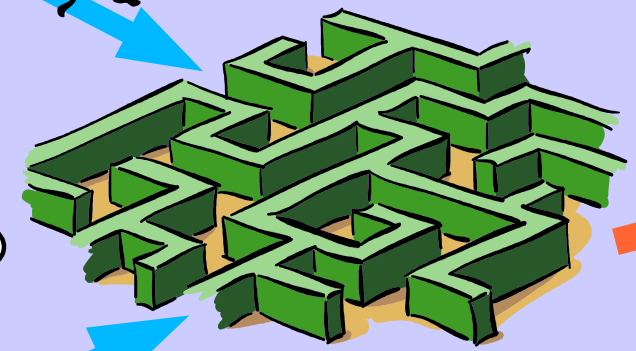


Matematická biologie & ICT

Algoritmy lze také rozdělit podle typu učení:



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



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

Predikce pro případy neznámé při trénování

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 Thick-ness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhe-sion	Single Epithelial Cell Size	Bare Nuclei	Bland Chro-matin	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

názvy atributů

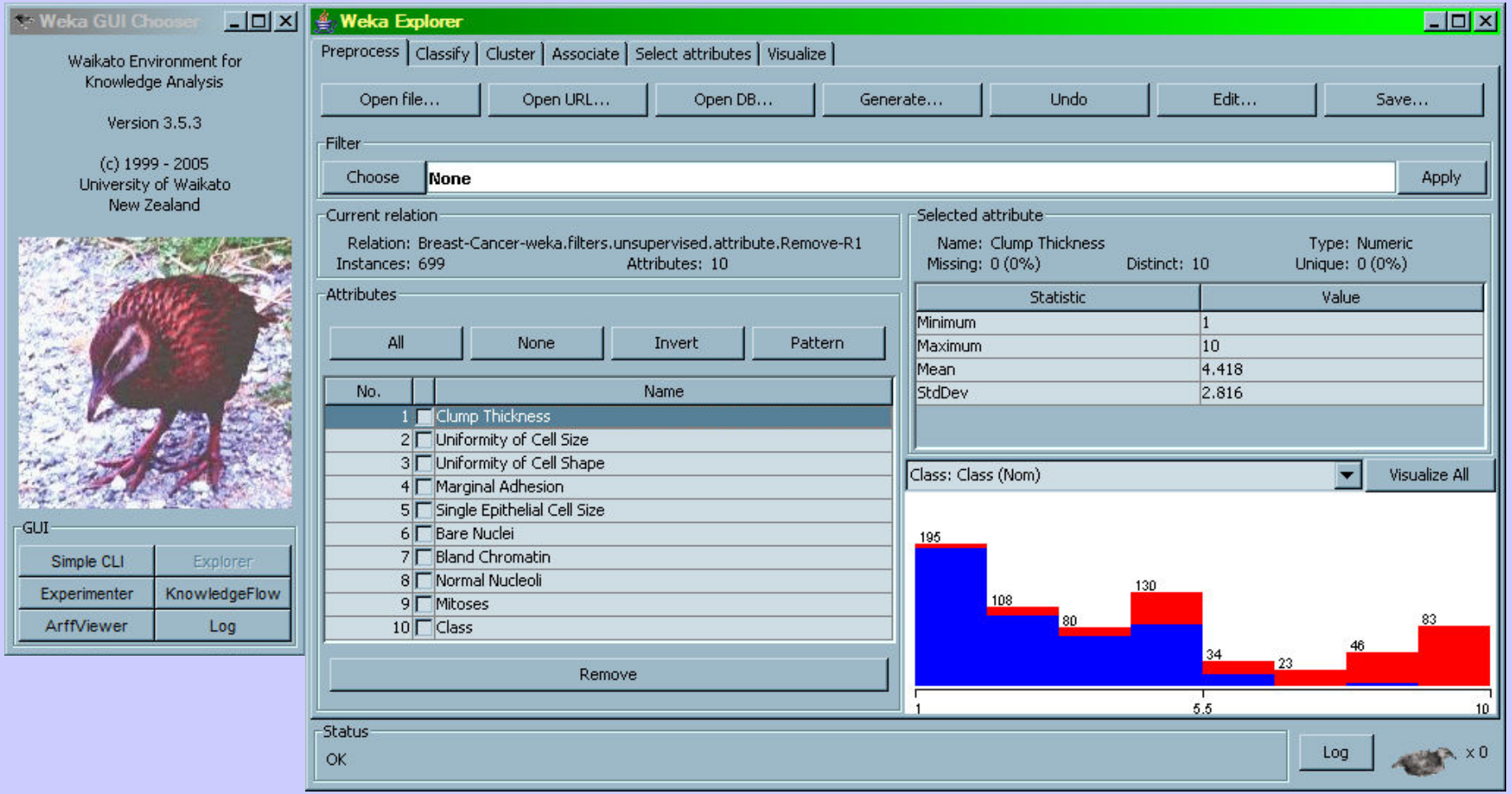
klasifikační třída

jeden z příkladů

(Wisconsin breast-cancer data)

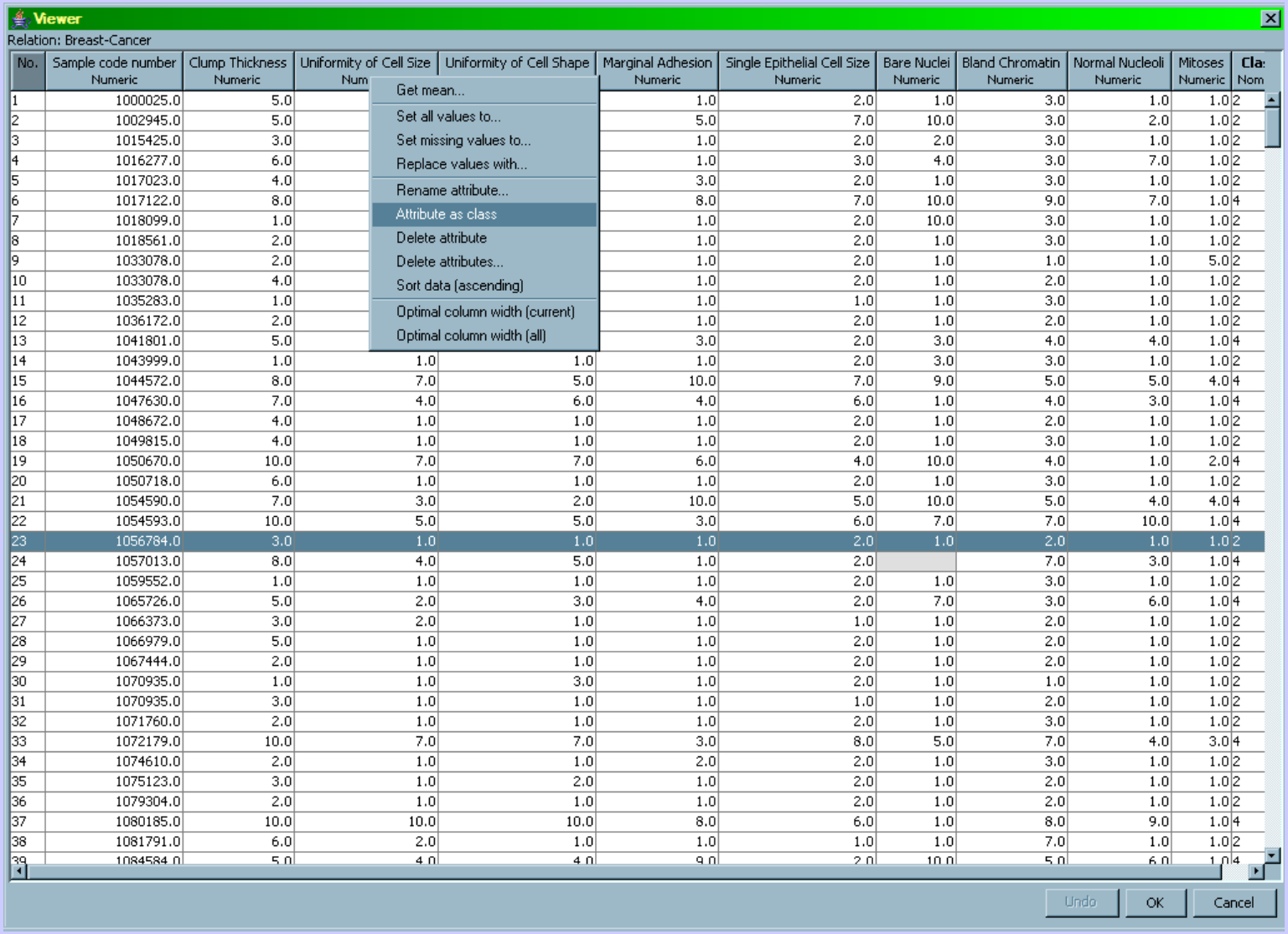
Matematická biologie & ICT

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



Matematická biologie & ICT

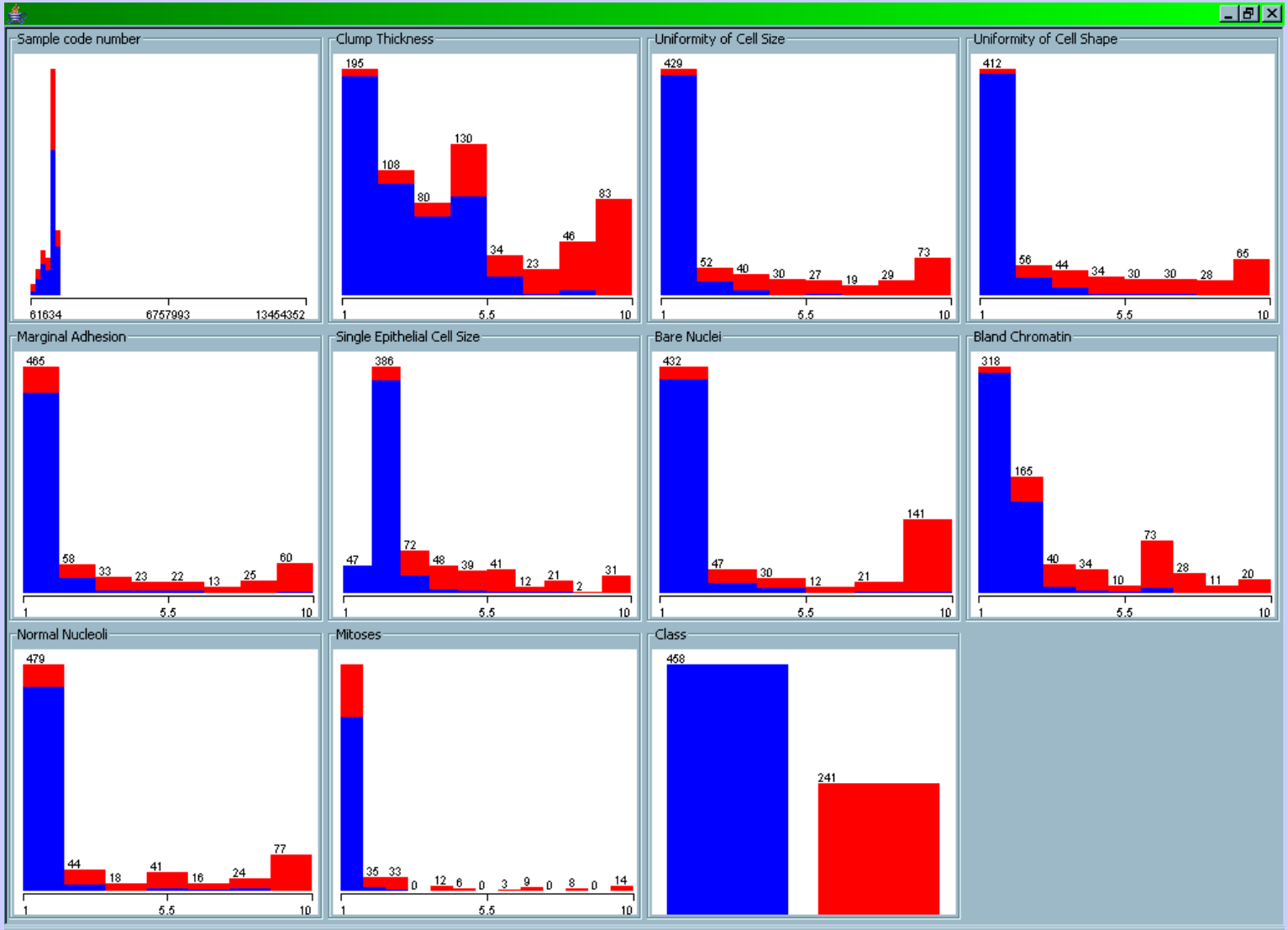
WEKA obsahuje i editor dat typu *spreadsheet*, který ***nemá*** typická omezení (např. pouze 256 sloupců a 65 536 řádků):



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	Cla: 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

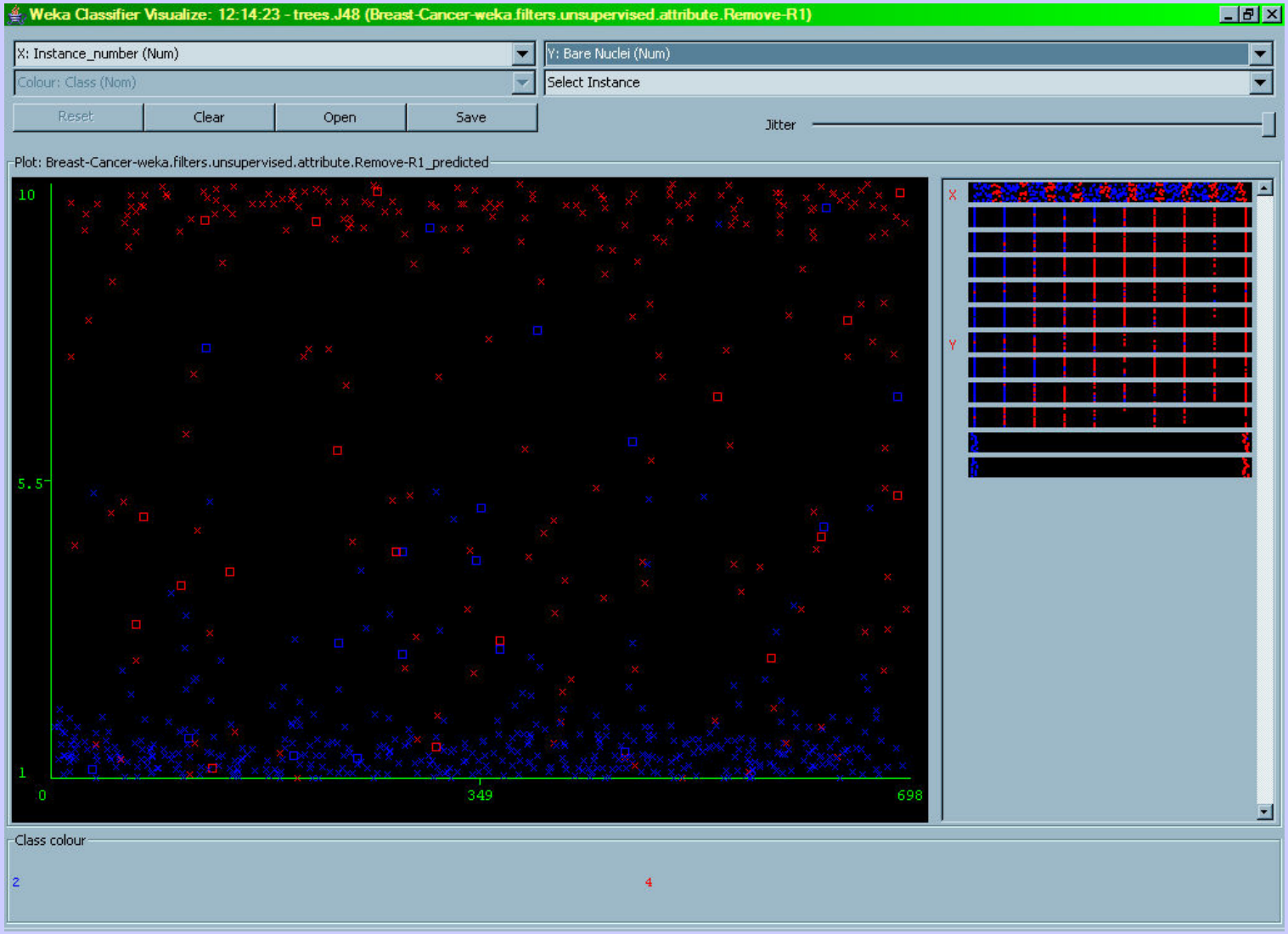
Matematická biologie & ICT

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



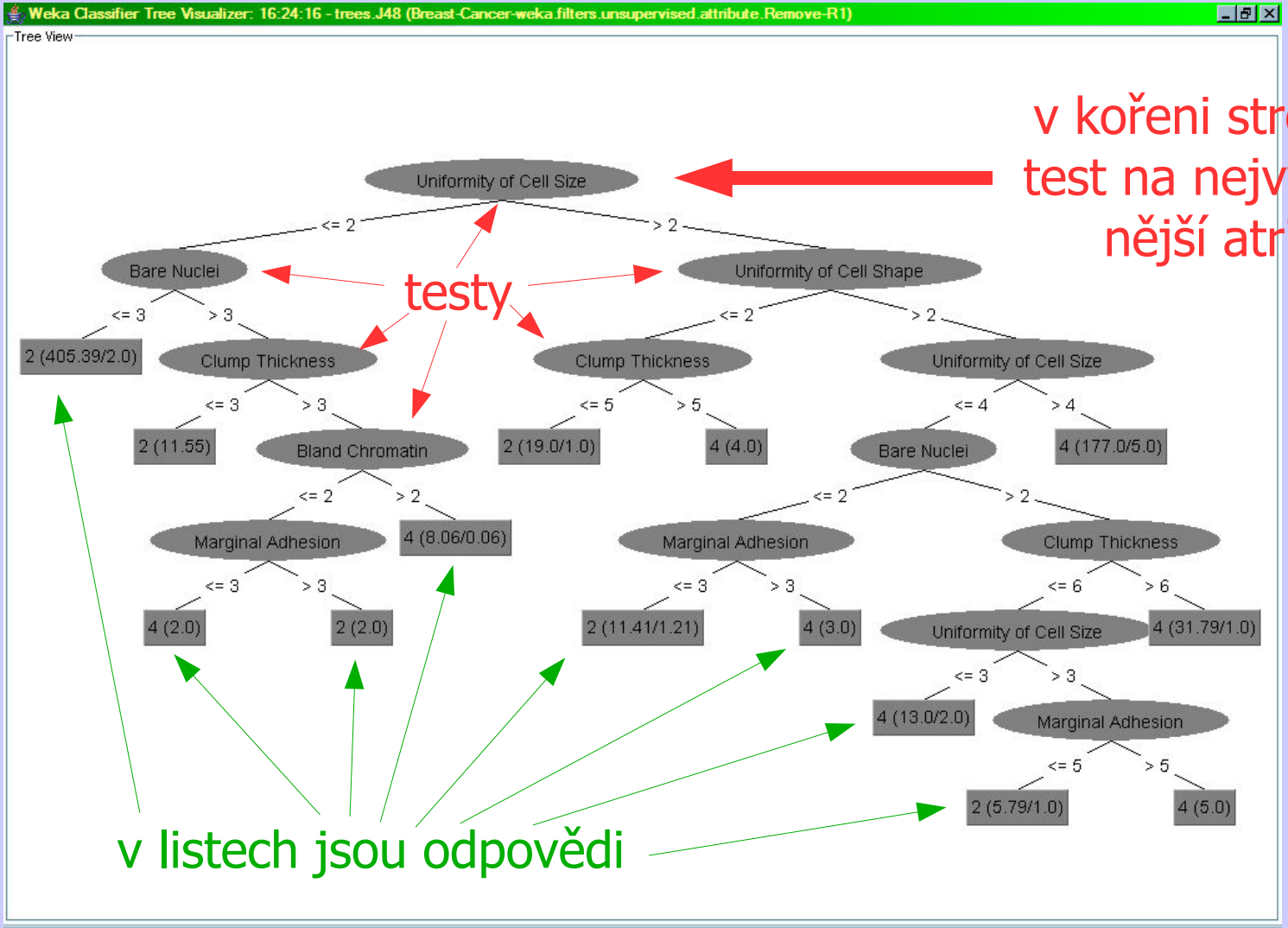
Matematická biologie & ICT

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



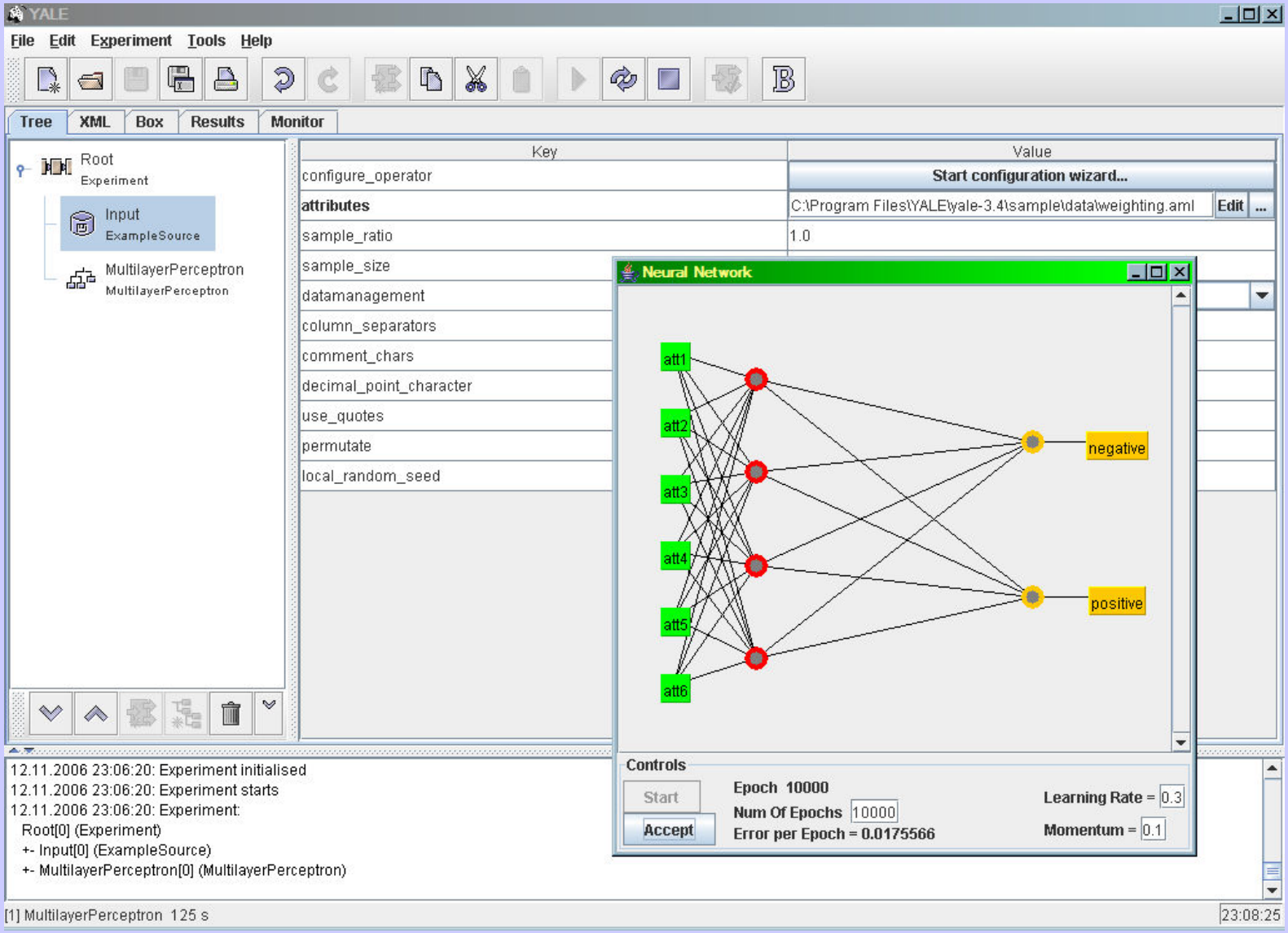
Matematická biologie & ICT

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



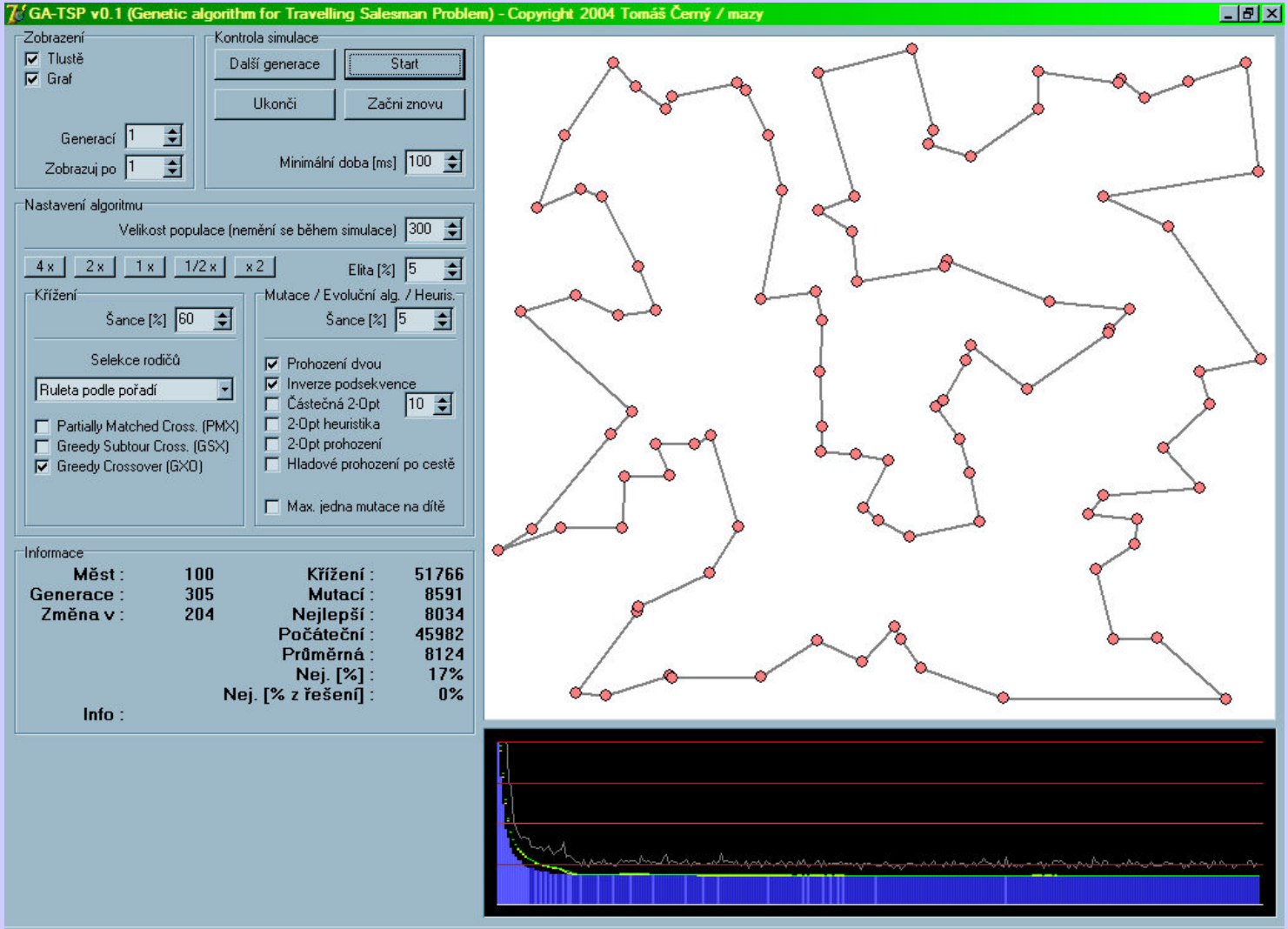
Matematická biologie & ICT

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



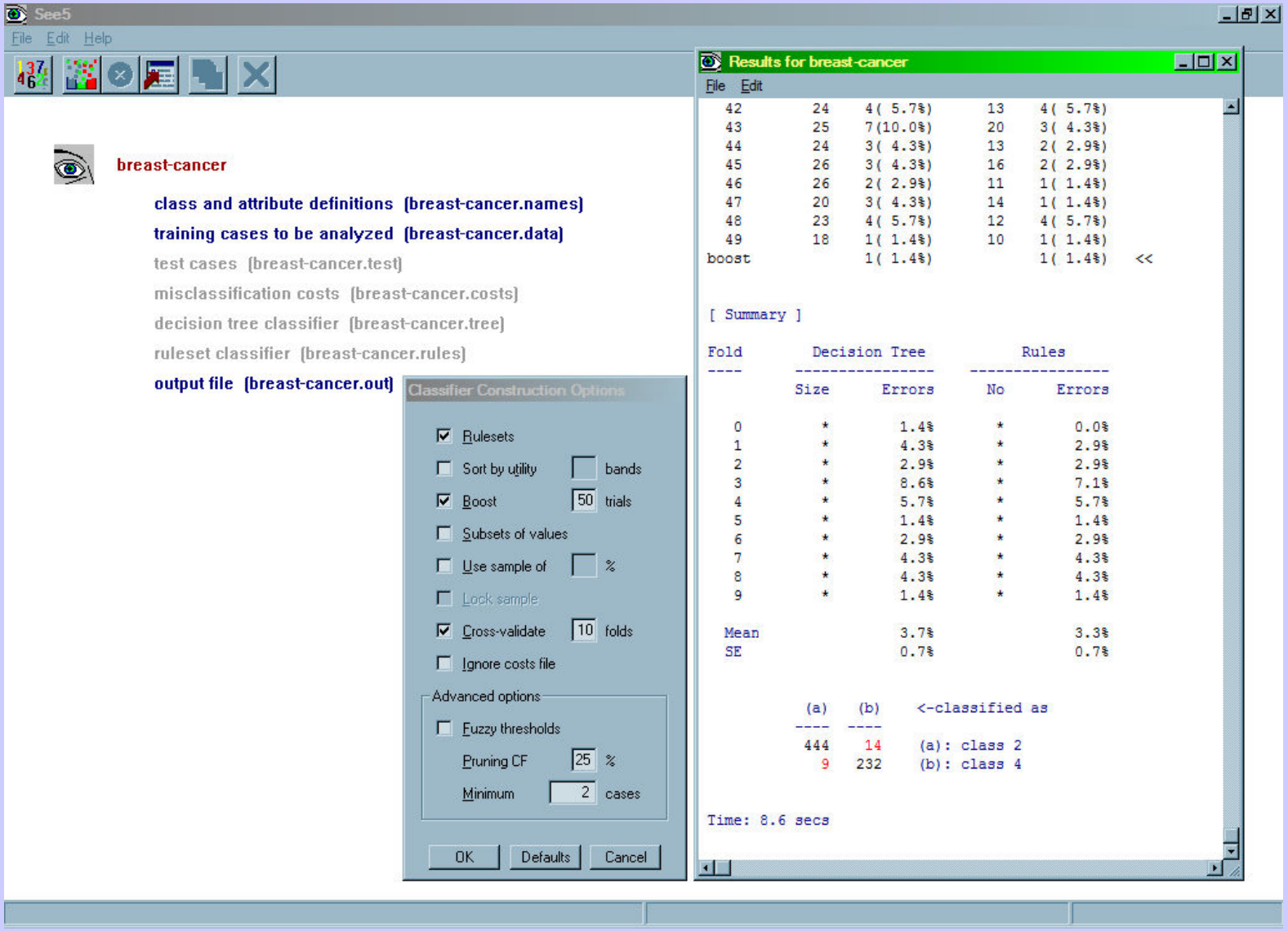
Matematická biologie & ICT

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



MATEMATICKÁ BIOLOGIE & ICT

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



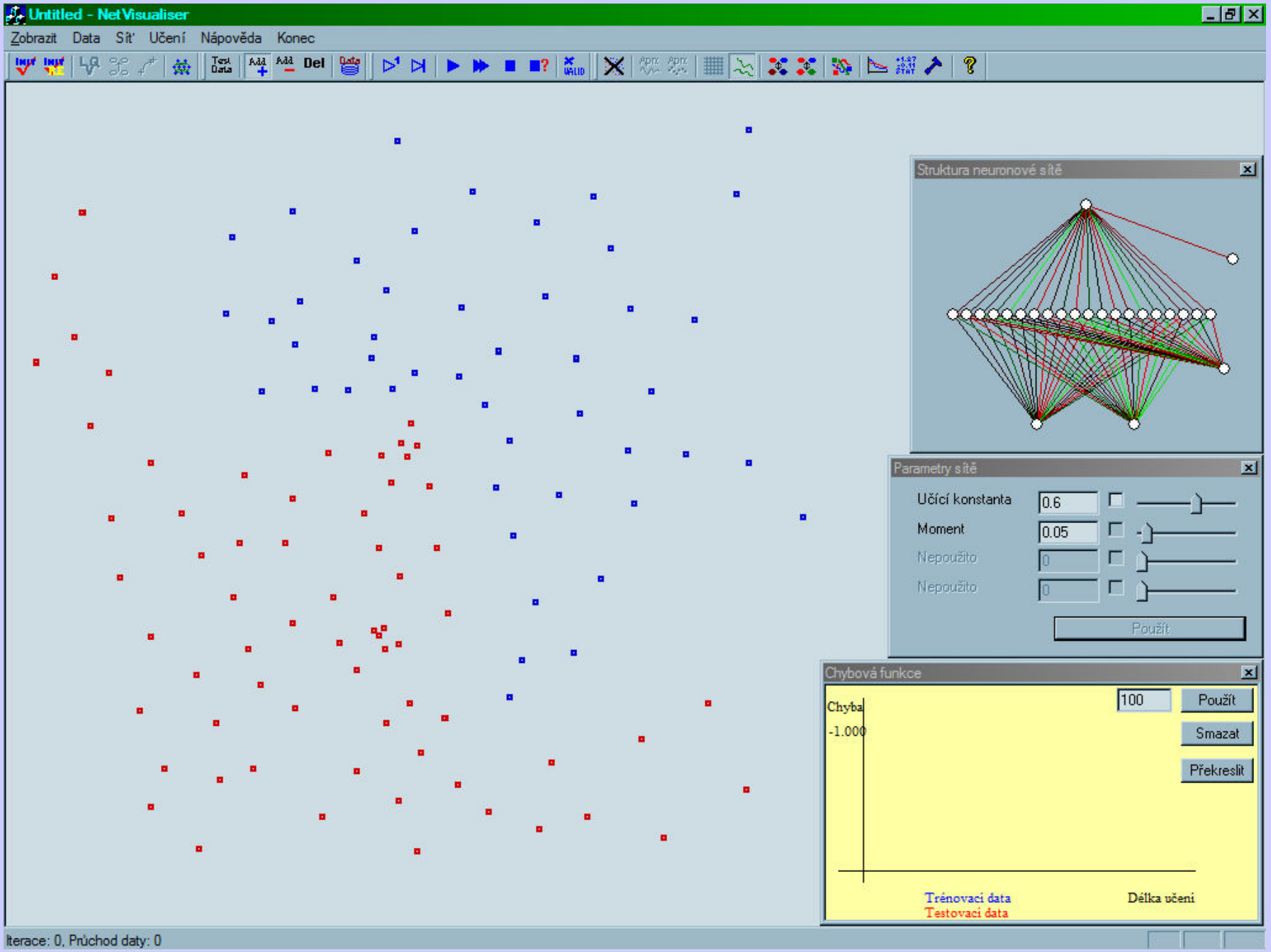
Matematická biologie & ICT

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.

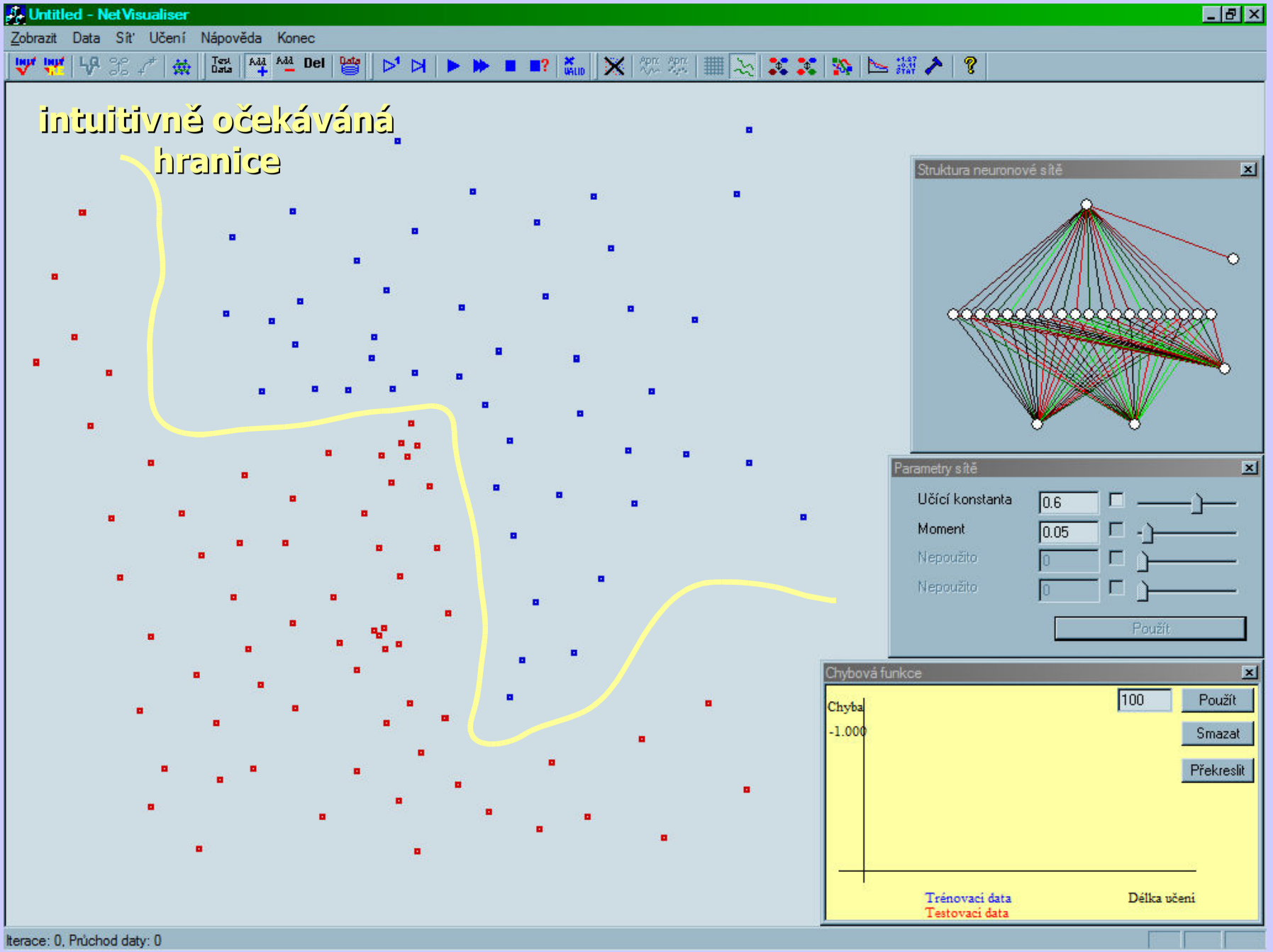
Podobně jako u dalších algoritmů strojového učení, návrh efektivní umělé neuronové sítě není snadný a hledání správných parametrů, včetně architektury sítě, bývá časově náročné.

Podaří-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.

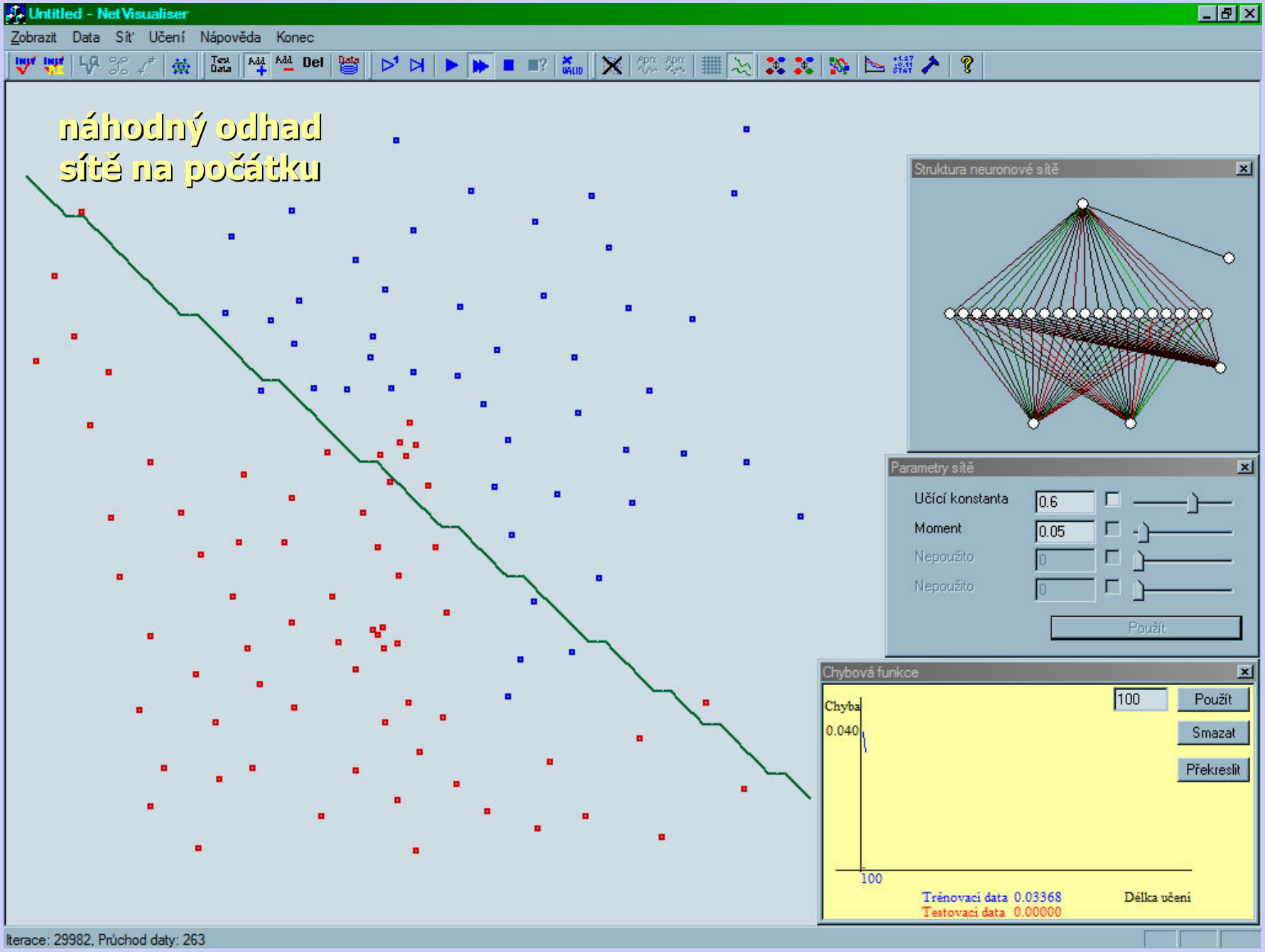
Matematická biologie & ICT



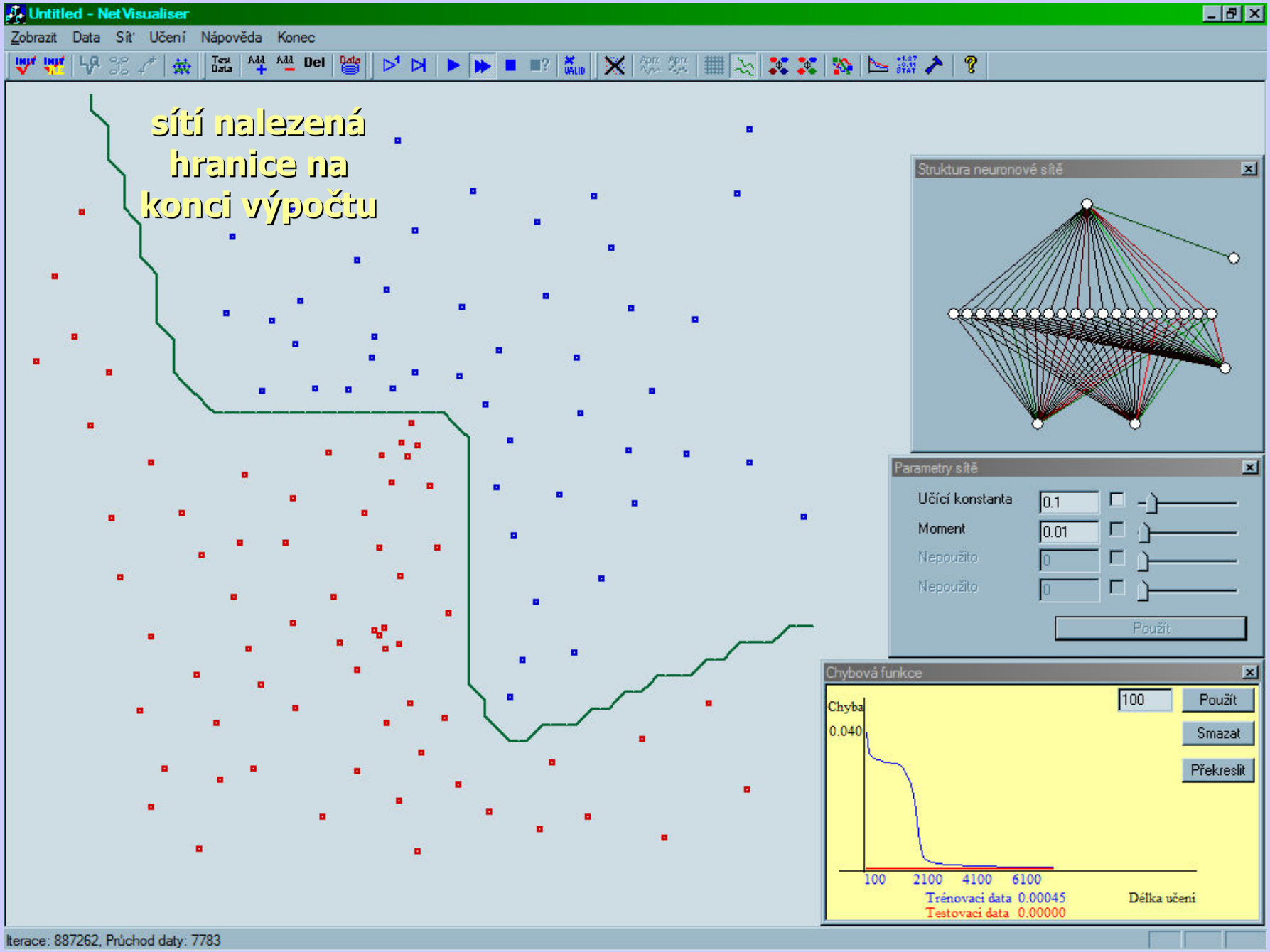
Matematická biologie & ICT



Matematická biologie & ICT



Matematická biologie & ICT



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.