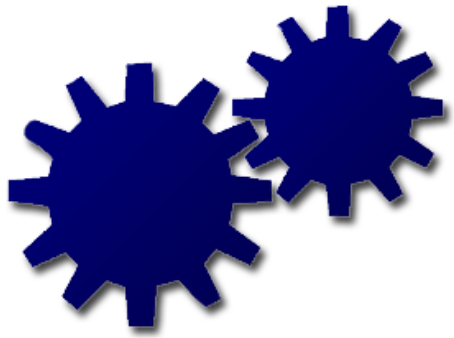


DOBÝVÁNÍ ZNALOSTÍ Z DATABÁZÍ

Úvod a oblasti aplikací

Martin Plchút

plchut@e-globals.net

DEFINICE A POJMY

Netriviální extrakce implicitních, dříve neznámých a potenciálně užitečných informací z dat.

[Fayyad a kol, 1996]

… knowledge discovery, data mining, data warehouse, information harvesting, data archeology, data destilery, business intelligence …

dobývání znalostí, dolování z dat (data mining)

Knowledge Discovery in Databases (KDD)
Data mining (DM)

MULTIOBOROVÁ DISCIPLÍNA

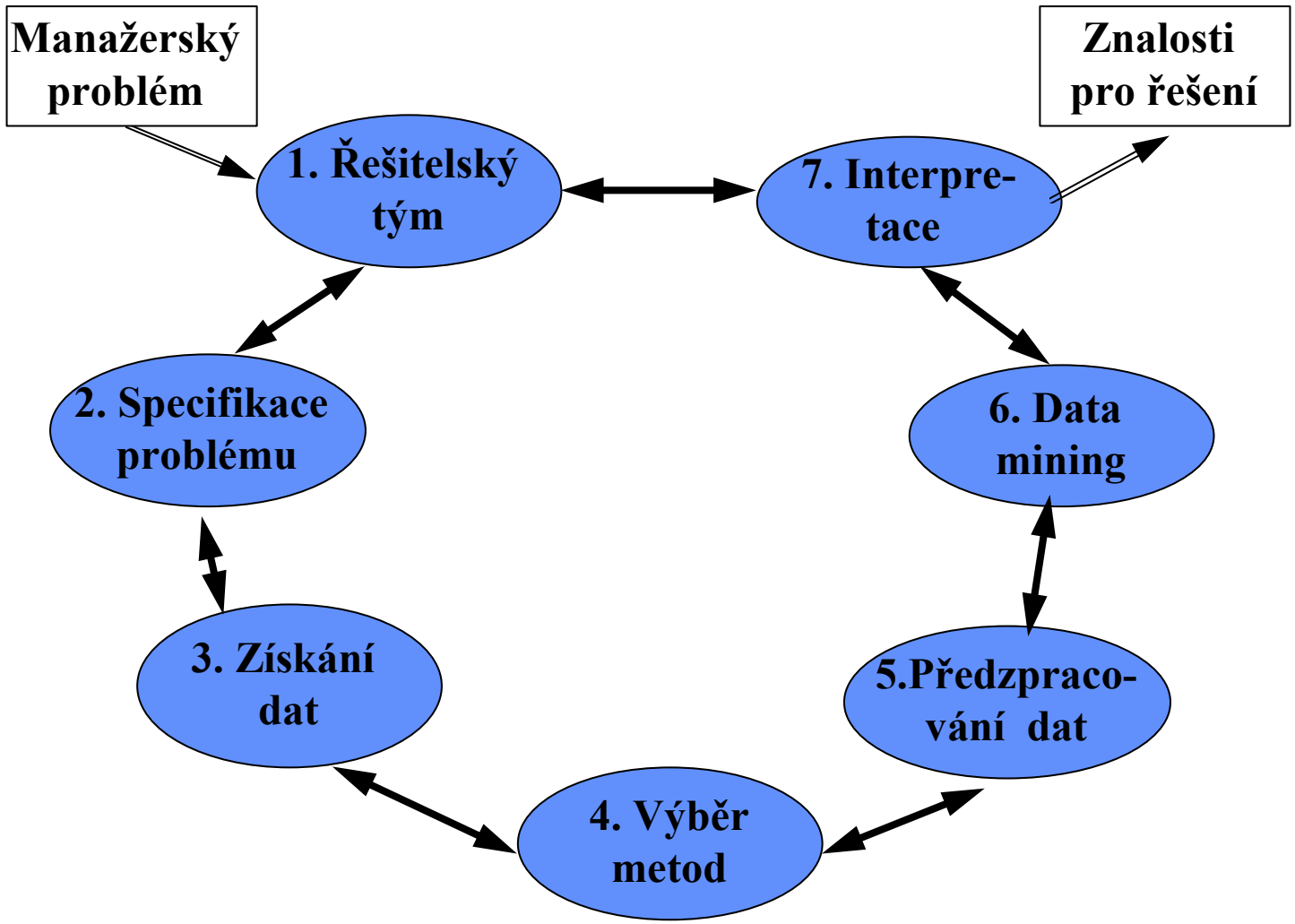
Tato nová oblast informatiky má své zdroje v těchto disciplínách:

- ❑ databáze
- ❑ statistika
- ❑ umělá inteligence (strojové učení)

Vývoj těchto disciplín probíhal nezávisle do chvíle kdy:

- rozsah automaticky sbíraných dat začínal uživatelům přerůstat přes hlavu
- vznikla potřeba používat tato data pro podporu (strategického) rozhodování ve firmách
- metody strojového učení se začaly využívat v oblasti ekonomie

PROCES KDD – manažerský pohled



PROCES KDD – technologický pohled

Znalosti

Vyhodnocení, interpretace

Vzorky

"Dolování"

Transformovaná data

Selekce a transformace

Data warehouse

Cílová data

Filtrace, integrace

Databáze

ALGORITMY PRO KDD

- **Statistika**

  - diskriminační analýza
  - regresní analýza
  - shluková analýza

- **Symbolické metody umělé inteligence**

  - rozhodovací stromy
  - rozhodovací pravidla
  - asociační pravidla
  - případové usuzování

- **Subsymbolické metody umělé inteligence**

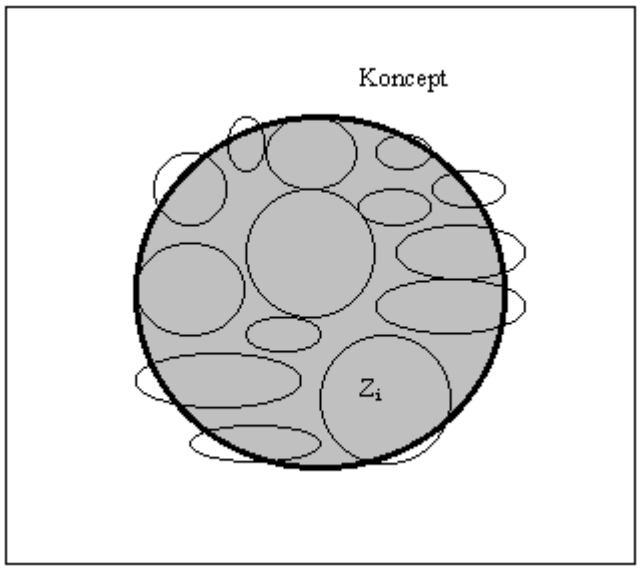
  - neuronové sítě
  - genetické algoritmy
  - bayesovské sítě

ÚLOHY KDD

V případě dobývání znalostí z databází můžeme mluvit o různých typech úloh. Jsou to především:

- klasifikace / predikce
- deskripce
- hledání nugetů

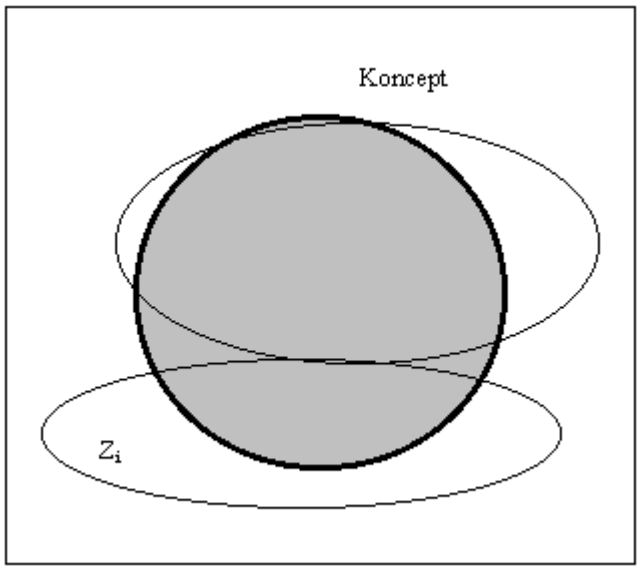
ÚLOHY KDD – klasifikace / predikce



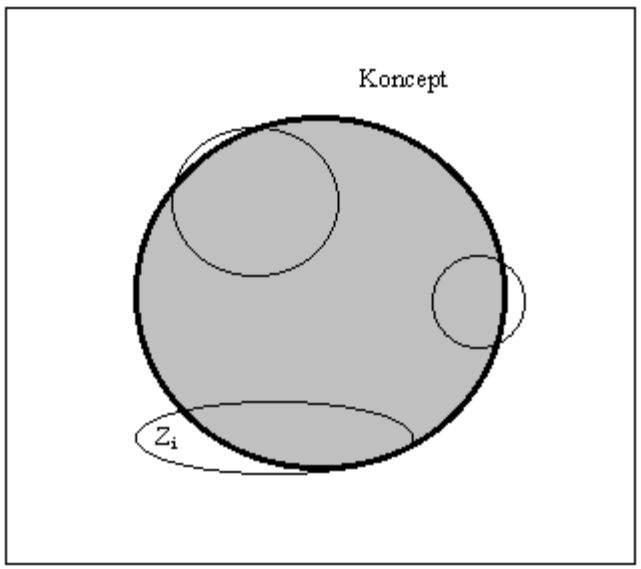
Při **klasifikaci/predikci** je cílem nalézt znalosti použitelné pro klasifikaci nových případů - zde požadujeme, aby získané znalosti co nejlépe odpovídaly danému konceptu; dáváme přednost přesnosti pokrytí na úkor jednoduchosti (připouštíme větší množství méně srozumitelných dílčích znalostí.

Rozdíl mezi klasifikací a predikcí spočívá v tom, že u predikce hraje důležitou roli čas; ze starších hodnot nějaké veličiny se pokoušíme odhadnout její vývoj v budoucností (např. předpověď počasí nebo pohybu cen akcií).

ÚLOHY KDD – deskripce a hledání nugetů



Při **deskripci (popisu)** je cílem nalézt dominantní strukturu nebo vazby, které jsou skryté v daných datech. Požadujeme srozumitelné znalosti pokrývající daný koncept. Dáváme tedy přednost menšímu množství méně přesných znalostí.



Hledáme-li nugety, požadujeme zajímavé (nové, překvapivé) znalosti, které nemusí plně pokrývat daný koncept.

APLIKAČNÍ OBLASTI

- Segmentace a klasifikace klientů banky (např. rozpoznání problémových nebo naopak vysoce bonitních klientů),
- Predikce vývoje kursů akcií,
- Analýza důvodů změny poskytovatele nějakých služeb (internet, mobilní telefony),
- Segmentace a klasifikace klientů pojišťovny,
- Analýza nákupního košíku (Market Basket Analysis).

- Predikce spotřeby elektrické energie,
- Analýza příčin poruch v telekomunikačních sítích,
- Rozbor databáze pacientů v nemocnici,

- Veřejné mínění a sčítání lidu.

METODOLOGIE – Metoda „5A“

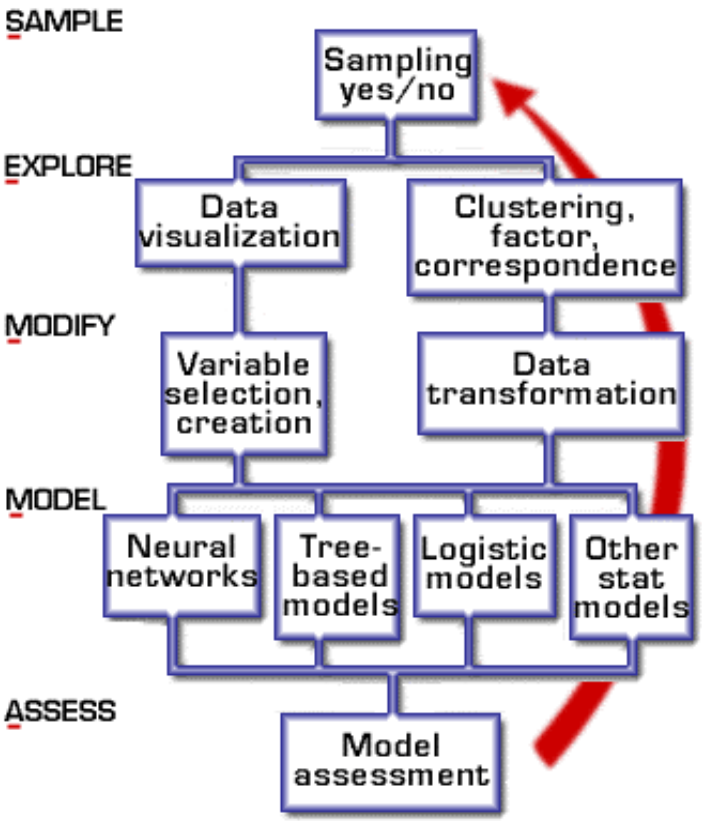
S postupem doby začaly vznikat **metodologie**, které si kladou za cíl poskytnout uživatelům jednotný rámec pro řešení různých úloh z oblasti dobývání znalostí. Tyto metodologie umožňují sdílet a přenášet zkušenosti z úspěšných projektů.

Metodologii „5A“ nabízí firma SPSS jako svůj pohled na proces dobývání znalostí. Název metodologie je akronymem pro jednotlivé prováděné kroky:

- Assess – posouzení potřeb projektu,
- Access – shromáždění potřebných dat,
- Analyze – provedení analýz,
- Akt – přeměna znalostí na akční znalosti,
- Automate – převedení výsledků analýzy do praxe.

METODOLOGIE - SEMMA

Enterprise Miner, softwarový produkt firmy SAS, vychází z vlastní metodologie pro dobývání znalostí z databází. Název metodologie opět charakterizuje jednotlivé prováděné kroky:



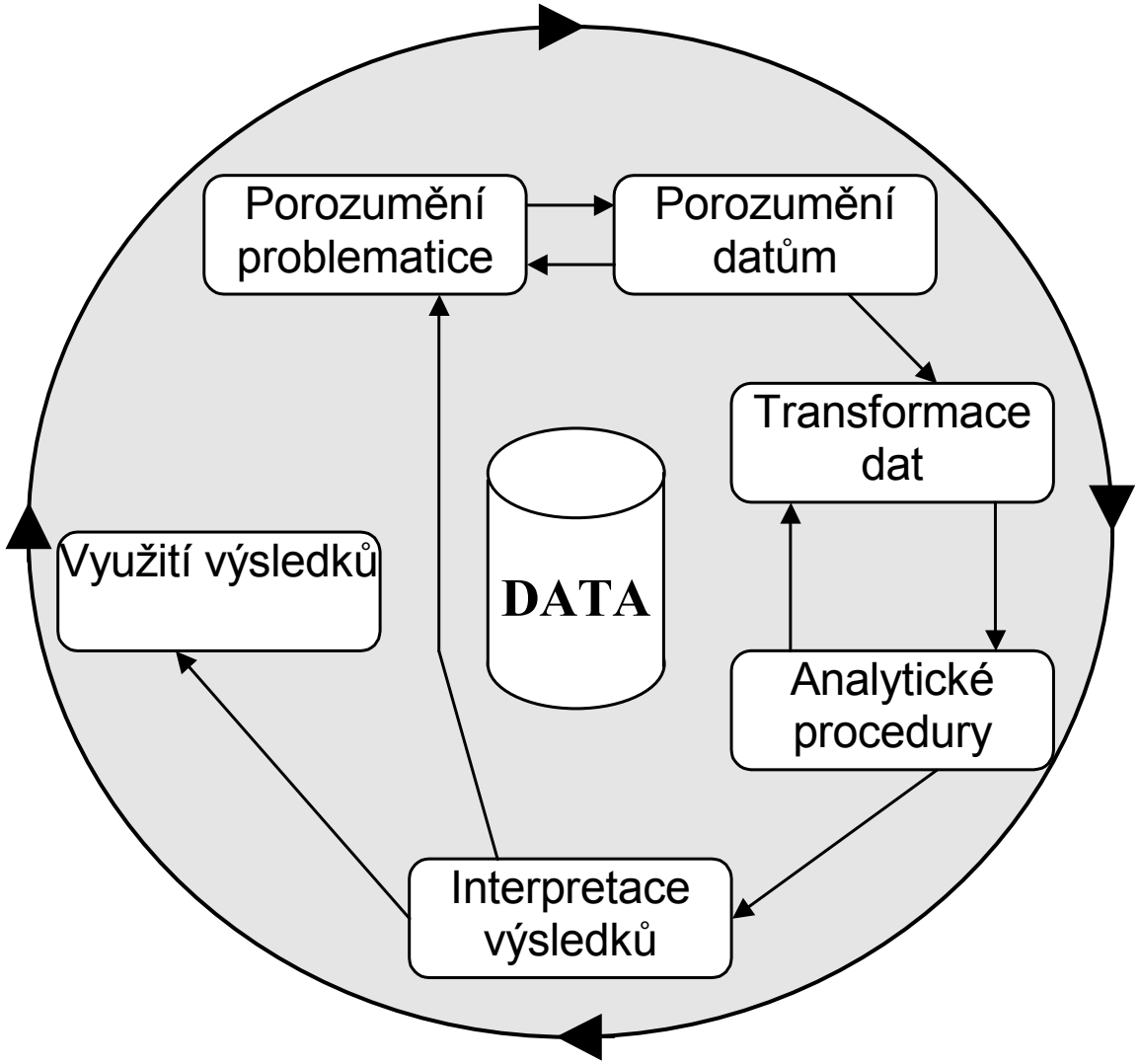
- Sample (vybrání vhodných objektů),
- Explore (vizuální explorace a redukce dat),
- Manipulate (seskupování objektů a hodnot atributů, datové transformace),
- Model (analýza dat: neuronové sítě, rozhodovací stromy, statistické techniky, asociace a shlukování),
- Assess (porovnání modelů a interpretace).

METODOLOGIE – CRISP DM

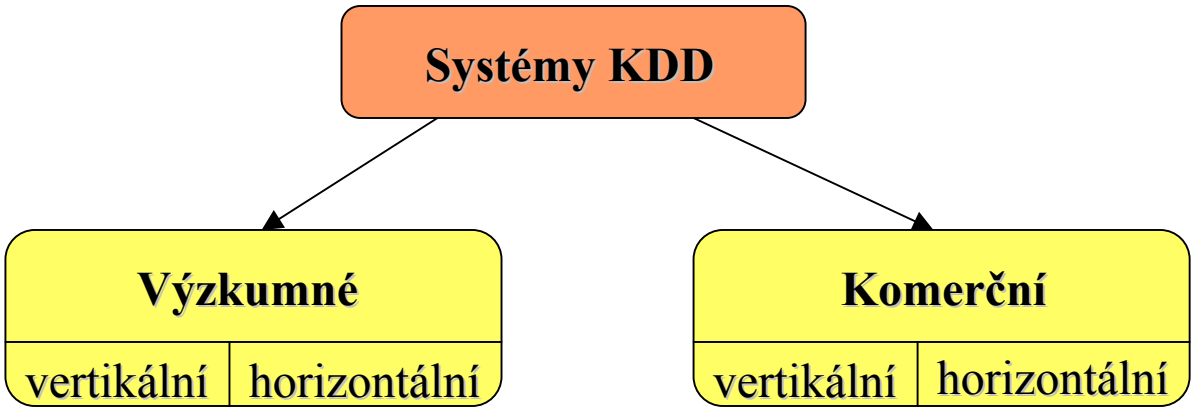
Metodologie **CRISP-DM (CRoss-Industry Standard Process for Data Mining)** vznikla v rámci výzkumného projektu Evropské komise. Cílem projektu je navrhnout univerzální postup (tzv. standardní model procesu dobývání znalostí z databází), který bude použitelný v nejrůznějších komerčních aplikacích. Vytvoření takovéto metodologie umožní řešit rozsáhlé úlohy dobývání znalostí rychleji, efektivněji, spolehlivěji a s nižšími náklady. Kromě návrhu standardního postupu má CRISP-DM nabízet „průvodce“ potenciálními problémy a řešeními, které se mohou vyskytnout v reálných aplikacích.

Na projektu spolupracují firmy NCR (přední dodavatel datových skladů), DaimlerChrysler, ISL (tvůrce systému Clementine) a OHRA (velká holandská pojišťovna). Všechny tyto firmy mají bohaté zkušenosti s reálnými úlohami dobývání znalostí z databází.

METODOLOGIE – CRISP DM



SYSTÉMY DOBÝVÁNÍ ZNALOSTÍ



Systémy pro dobývání znalostí z databází tedy:

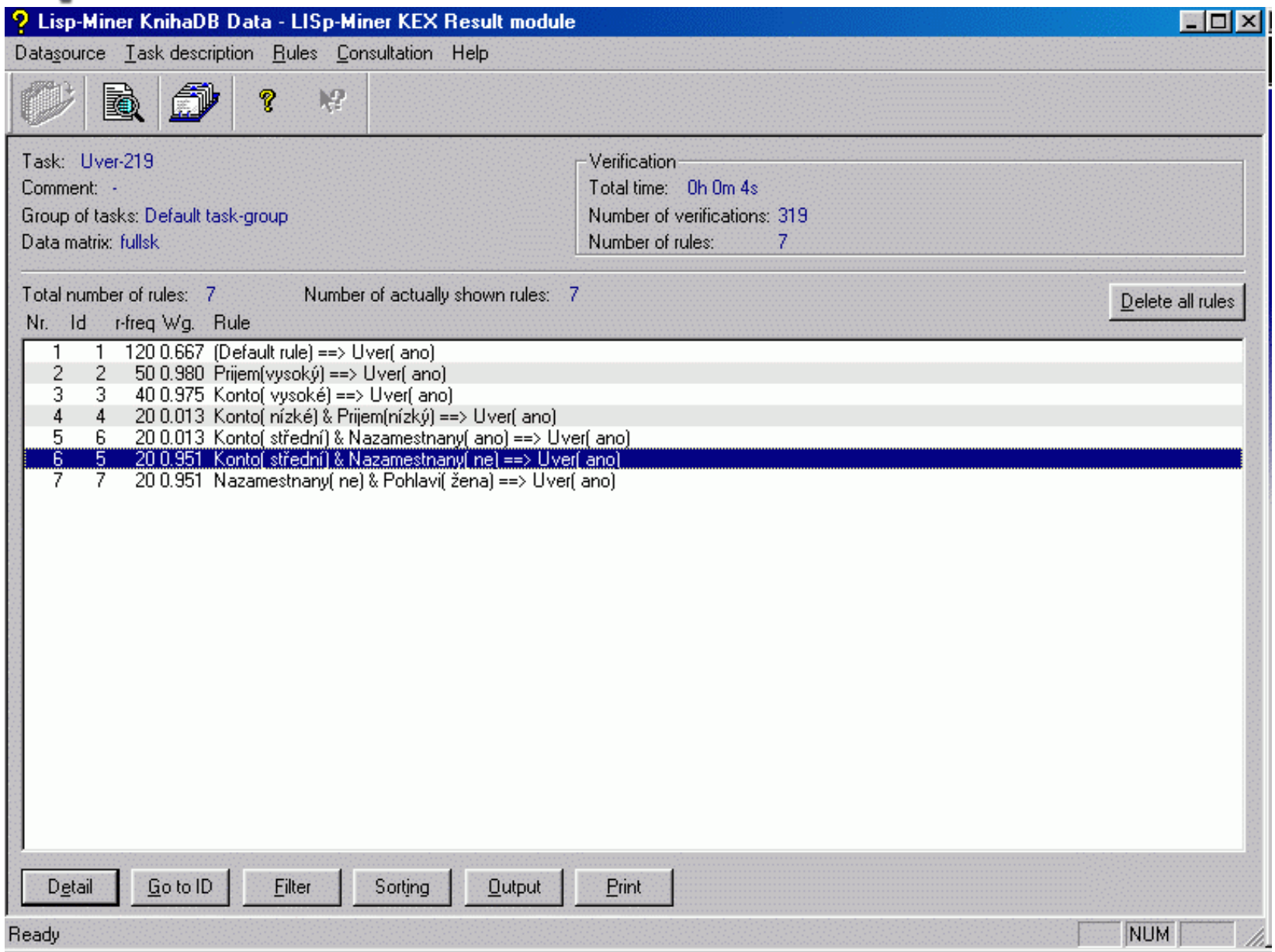
- pokrývají celý proces dobývání znalostí (od přezpracování po interpreraci)
- nabízejí více algoritmů pro analýzu (než „jednoúčelové“ systémy strojového učení)
- kladou důraz na vizualizaci (ve způsobu práce se systémem i při interpretaci výsledků).

SYSTÉMY DOBÝVÁNÍ ZNALOSTÍ

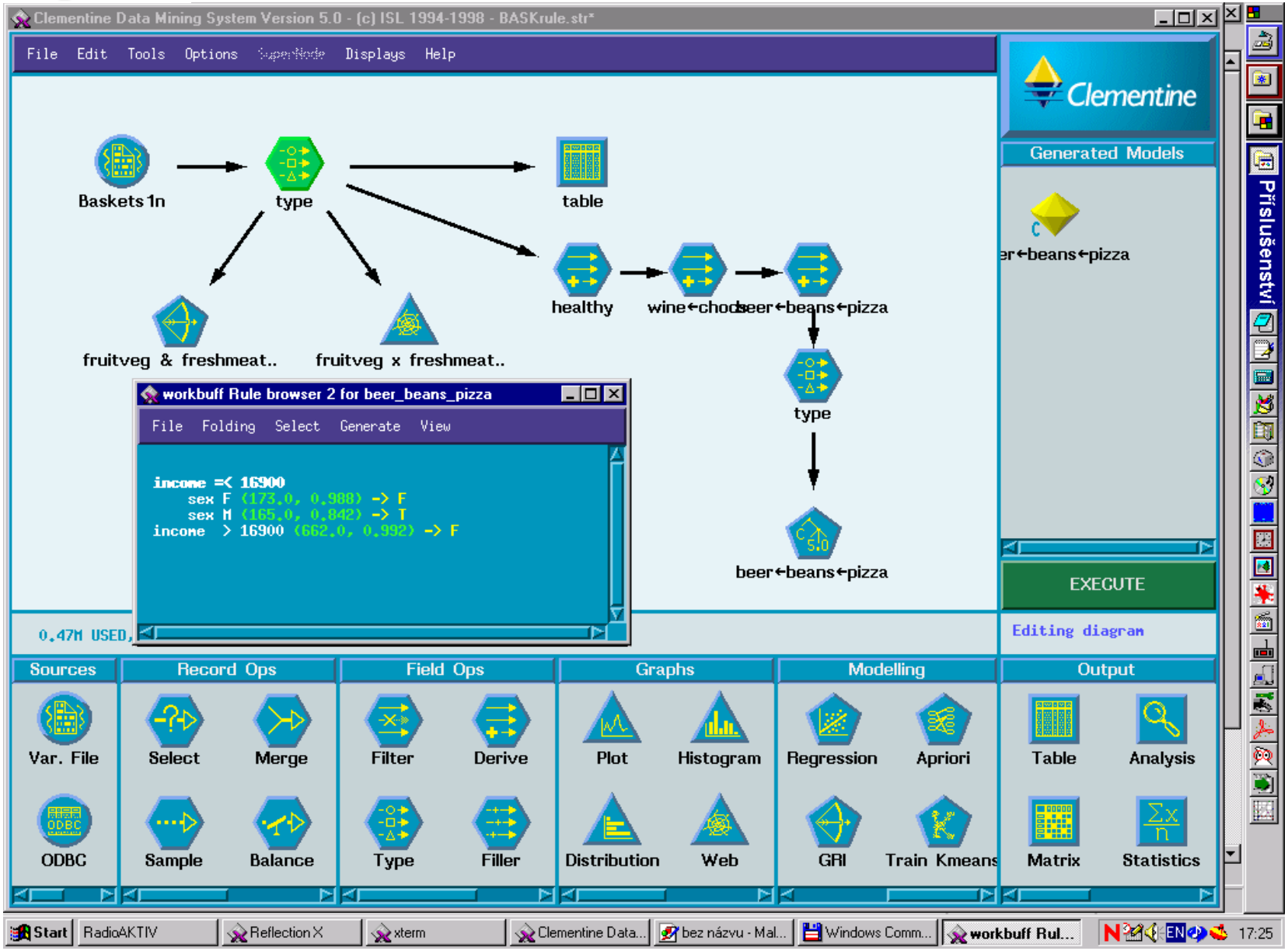
Systém	*Výrobce*	*URL*
CART	Salford Systems	http://www.salford-systems.com
Clementine	Integral Solutions (SPSS)	http://www.isl.co.uk/clem.html
Enterprise Miner	SAS Institute	http://www.sas.com/software/components/miner.html
Intelligent Miner	IBM	http://www-4.ibm.com/software/data/iminer
Kepler	Dialogis	http://www.dialogis.de
KnowledgeStudio	Angoss	http://www.angoss.com
LISp Miner	VŠE	http://lispminer.vse.cz
MineSet	Silicon Graphics	http://www-europe.sgi.com/software/mineset
See5	RuleQuest Research	http://www.rulequest.com/see5-info.html
Weka	University of Waikato	http://www.cs.waikato.ac.nz/~ml/weka
WizWhy	WizSoft	http://www.wizsoft.com/why.html

Systém	*Rozhodovací stromy*	*Rozhodovací pravidla*	*Asociační pravidla*	*Neuronové sítě*	*Lineární statistické metody*	*Nejbližší soused*
CART	✓	×	×	×	×	×
Clementine	✓	✓	✓	✓	✓	✓
Enterprise Miner	✓	×	✓	✓	✓	✓
Intelligent Miner	✓	×	✓	✓	✓	✓
Kepler	✓	✓	✓	×	×	✓
KnowledgeStudio	✓	✓	×	×	×	×
LISp Miner	×	✓	✓	×	×	×
MineSet	✓	×	✓	×	×	×
See5	✓	✓	×	×	×	×
Weka	✓	✓	✓	✓	✓	✓
WizWhy	×	✓	×	×	×	×

SYSTÉMY DOBÝVÁNÍ ZNALOSTÍ



SYSTÉMY DOBÝVÁNÍ ZNALOSTÍ



APLIKAČNÍ OBLASTI

Projekt ***STULONG*** sleduje dva hlavní cíle:

- ❑ podrobné studium rizikových faktorů aterosklerózy u mužů středního věku,
- ❑ demonstrace možností aplikace metod dobývání znalostí z databází v medicínské oblasti.

Projekt Stulong započal v první polovině sedmdesátých let jako rozsáhlá epidemiologické studie primární prevence aterosklerózy. Sběr dat probíhal na několika pracovištích. Prvotní data byla později upravena do elektronické podoby a zpracována základními statistickými postupy. Poté byla analyzována v rámci výzkumu aplikací metod dobývání znalostí v medicíně.

Rizikové faktory aterosklerózy byly sledovány **u 1 419 mužů v letech 1976 - 1999** v souladu s celkovou metodikou projektu. Při vstupním vyšetření byly zjišťovány hodnoty **244** atributů. Hodnoty **219** z nich byly číselné kódy nebo výsledky měření. Dále bylo provedeno **10 610** kontrolních vyšetření, při kterých byly zjišťovány hodnoty **66** atributů.