Data Mining

Co je to Data Mining?

⌘ Data mining (DM), nebo také dolování z dat či vytěžování dat, je analytická metodologie získávání netriviálních skrytých a **potenciálně užitečných informací**.

Aplikace

- Bankovnictví: schvalování úvěrů/kreditních karet
  - Predikce dobrých zákazníků.
- CRM:
  - Identifikace zákazníků, kteří mají v úmyslu přejít ke konkurenci.
  - Cross-selling.
  - Up-selling.
- Cílený marketing:
  - Identifikace pravděpodobných respondentů na nabídku.
- Detekce fraudu: telecomunikace, finanční transakce
  - Online identifikace podvodného chování.

Aplikace

- Medicína: efektivita léčebné péče
  - Analýza pacientovy historie (předchozí nemoci a jejich průběh): nalezení vztahu mezi nemocemi.
- Farmacie: identifikace nových léků
- Vědecká analýza dat:
  - Identifikace nových galaxií.
- Design webových stránek:
  - Nalezení vztahu návštěvníka stránek a příslušná změna podoby stránek.

Aplikace

- Rozpoznávání psaného textu, řeči, obrázků.
- Supermarkety
  - Identifikace současně nakupovaného zboží
- Průmysl:
  - automatické přenastavení ovládacích prvků při změně parametrů procesu.
- Sport:
  - NBA-optimalizace herní strategie
- další...

Aplikace - Rozmístění zboží v supermarketech

⌘ Cíl: identifikovat zboží, které je nakupováno souběžně dostatečným množstvím zákazníků.

⌘ Výsledek: Jestliže zákazník nakupuje dětské pleny a mléko, pak si velmi pravděpodobně koupí i pivo.

Aplikace - Rozmístění zboží v supermarketech

Data mining a princip indukce

Indukce vs. Dedukce

- Dedukce zachovává platné vztahy:
  1. Koně jsou savci.
  2. Všichni savci mají plíce.
  3. Proto platí, že všichni koně mají plíce.

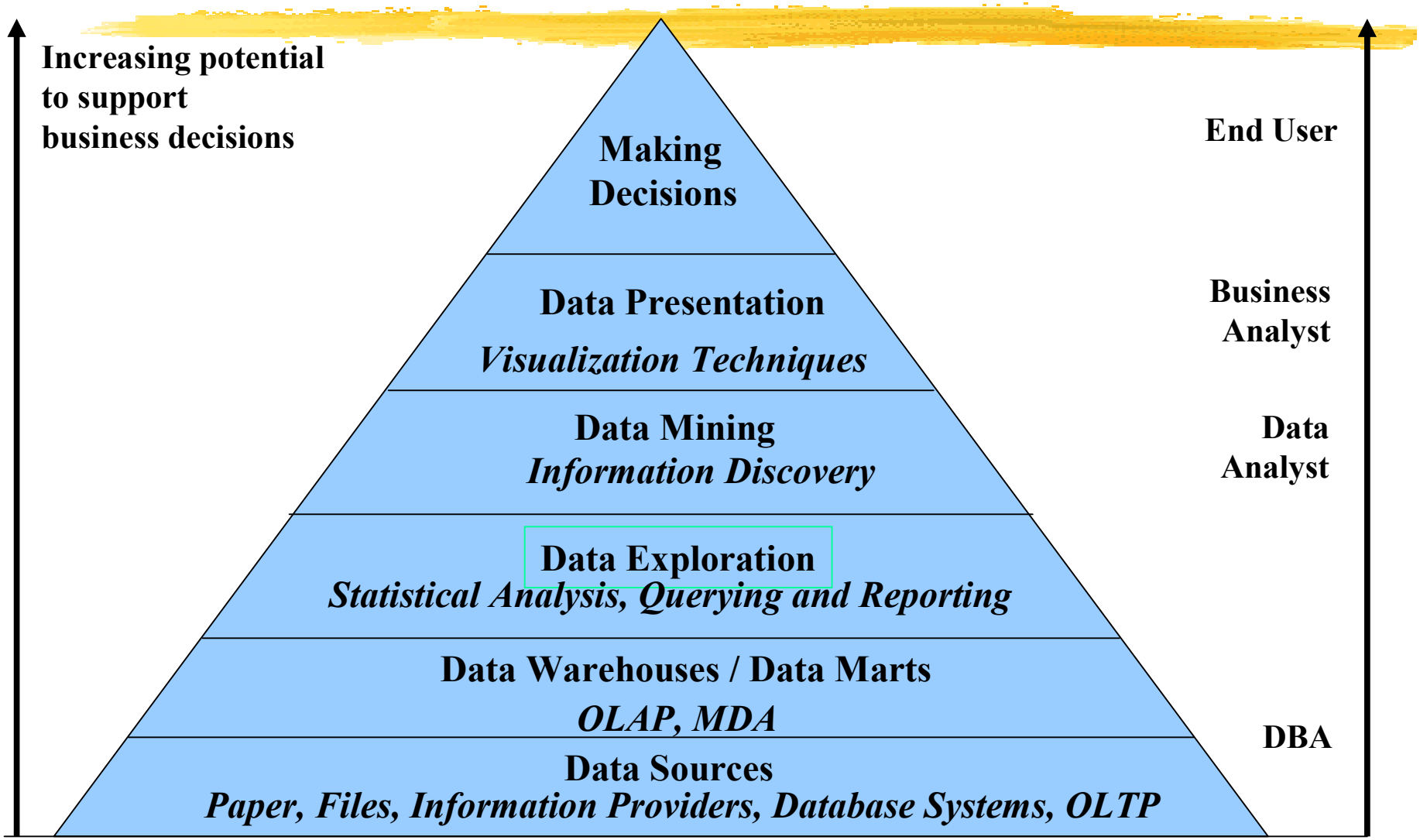
- Indukce přidává informace:
  1. Všichni doposud pozorovaní koně mají plíce.
  2. Proto platí, že všichni koně mají plíce.

Problém s indukcí

- Z platných faktů můžeme vyvodit nepravdivé tvrzení (model).

- Příklad:
  - Evropské labutě jsou bílé
  - Indukce: „Labutě jsou bílé" jakožto obecné pravidlo.
  - Objevením Austrálie se objevili i černé labutě…
  - Problém: množina pozorování nebyla náhodná a tudíž reprezentativní.

Data mining –podpora business rozhodnutí



Historie názvu

1960 Data Fishing, Data Dredging (bagrování):

- užíváno statistiky

1989 Knowledge Discovery (KD, KDD):

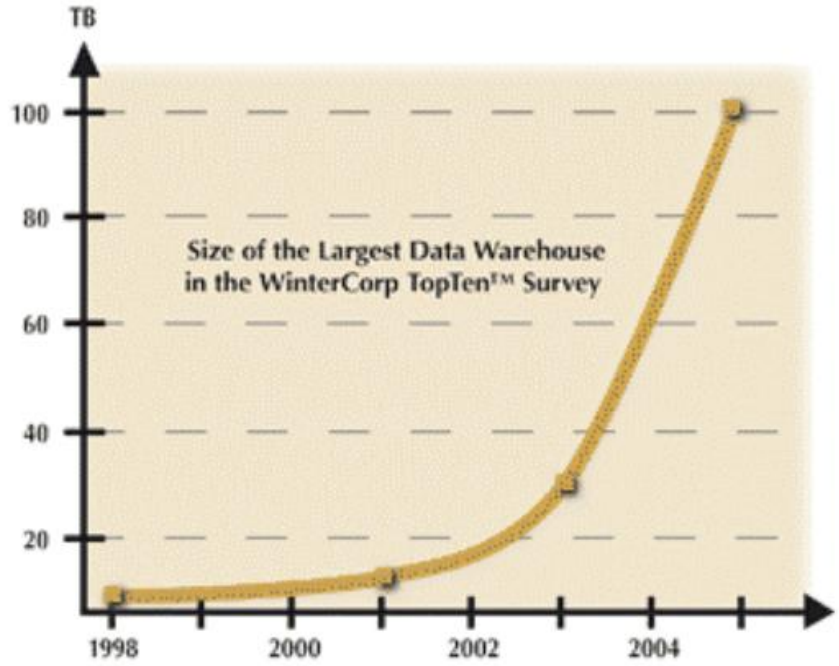
- užíváno komunitou zabývající se umělou inteligencí a strojovým učením

1990 Data Mining (DM):

- užíváno v komerční sféře a databázové komunitě

Další názvy: Data Archaeology, Information Harvesting, Information Discovery, Knowledge Extraction, ...

Data mining – nutnost?



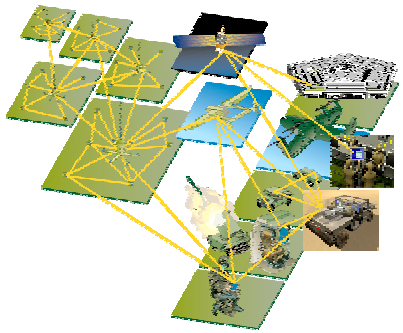
Největší světové databáze v r. 2005:

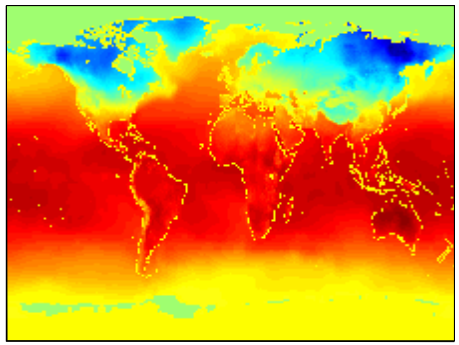
- Max Planck Inst. for Meteorology ~ 222 TB
- Yahoo ~ 100 TB
- AT&T ~ 94 TB

Data mining – nutnost?

- Terabytes -- 10^12 bytes: data obchodních řetězců, bank,…
- Petabytes -- 10^15 bytes: geografická data
- Exabytes -- 10^18 bytes: národní databáze zdravotních záznamů
- Zettabytes -- 10^21 bytes: databáze meteo-snímků
- Zottabytes -- 10^24 bytes: video-databáze

Data mining – nutnost?





Proč data mining? Proč dnes?

- Data jsou produkována
- Data jsou skladována
- Výpočetní síla je dostupná
- Výpočetní síla je cenově dostupná
- Konkurenční tlak je velice silný
- Komerční produkty jsou k dispozici

Data mining vs. Statistická analýza

Data Mining

- Původně vyvinuto pro expertní systémy automaticky řešící zadané problémy.
- Neklade takový důraz na přesné porozumění použité metody.
- Pokud něco dává smysl, pak to použijme!
- Žádné předpoklady o datech.
- Funguje i pro velmi rozsáhlá data.
- Vyžaduje porozumění problému z datovému a business pohledu.

Statistická analýza

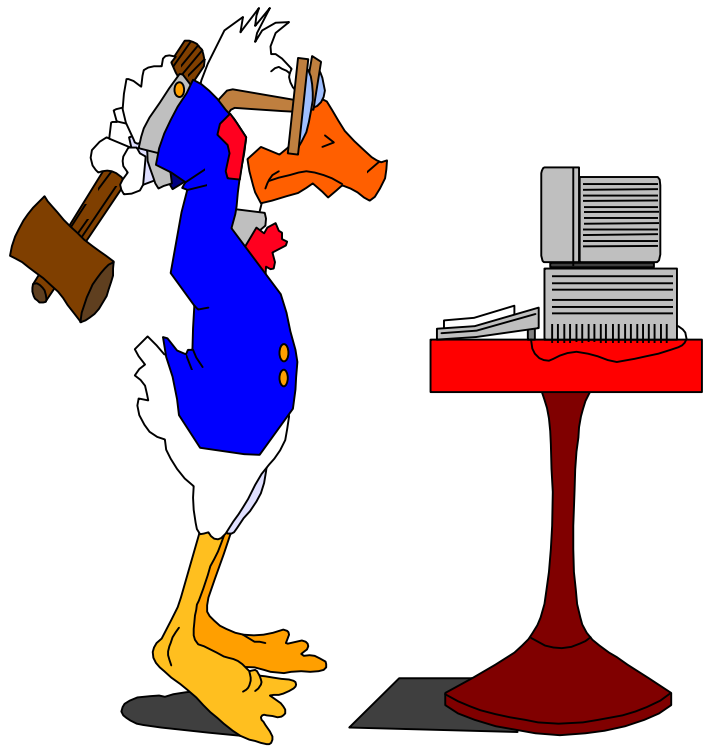
- Testuje se statistická korektnost modelu.
  - Jsou statistické předpoklady modelu modelu splněny?
- Testování Hypotéz.
- Intervalové odhady.
- Pracuje se s výběrem hodnot.
- Standardní metody nejsou optimalizovány pro rozsáhlá data.
- Vyžaduje pokročilé statistické znalosti.

Data mining

- Proces (polo-) automatické analýzy (rozsáhlých) databází k identifikaci vztahů, které jsou:
  - validní: platí na nových datech s určitou jistotou obecné platnosti
  - nové: doposud neznámé
  - užitečné: dají se v praxi nějak použít
  - srozumitelné: (vždy) se nalezený vztah dá nějak vysvětlit

Data mining není:

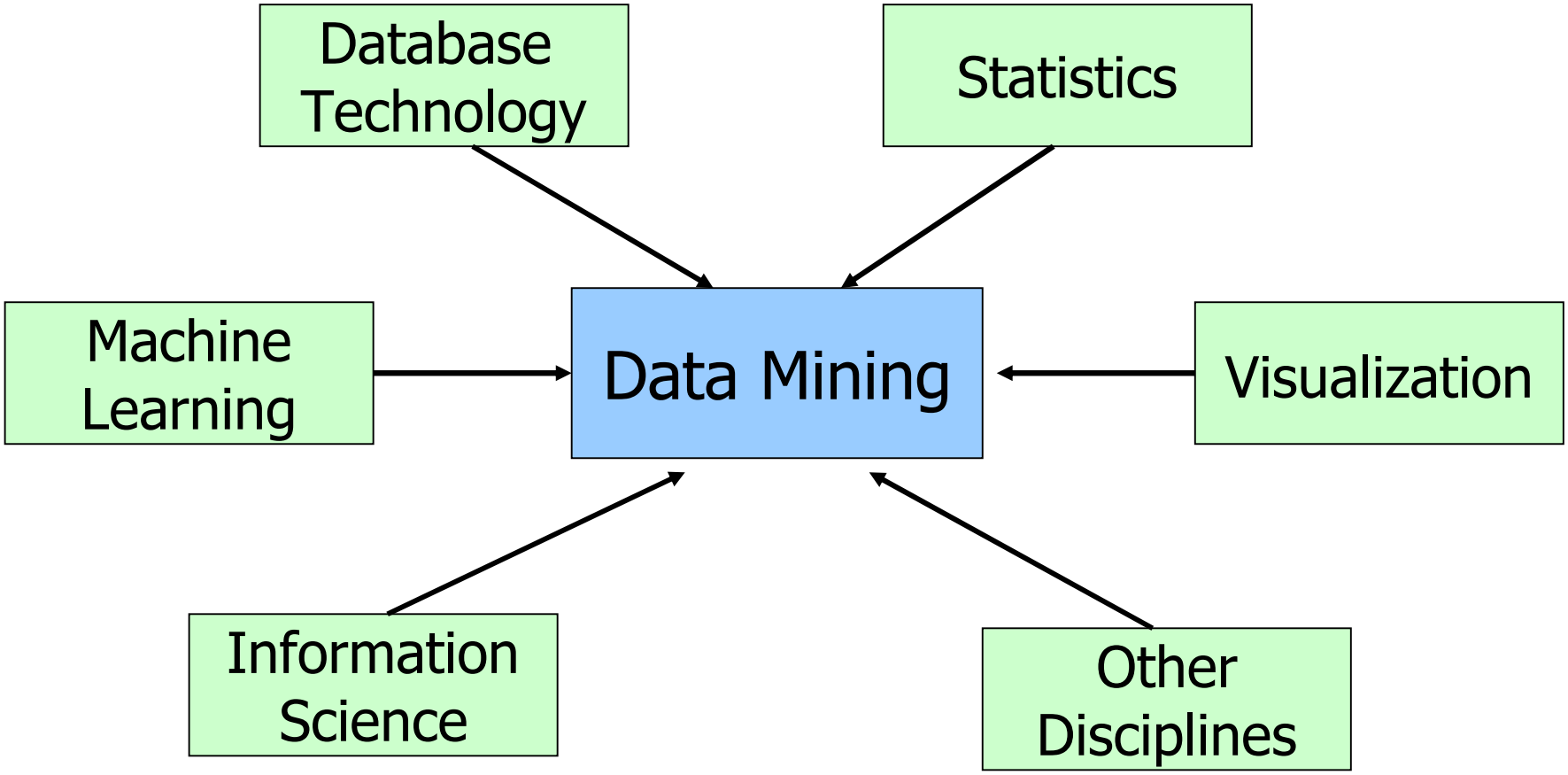
- Brutální hromadné zpracování dat.
- Slepé použití algoritmů.
- Hledání vztahů tam, kde žádné neexistují.



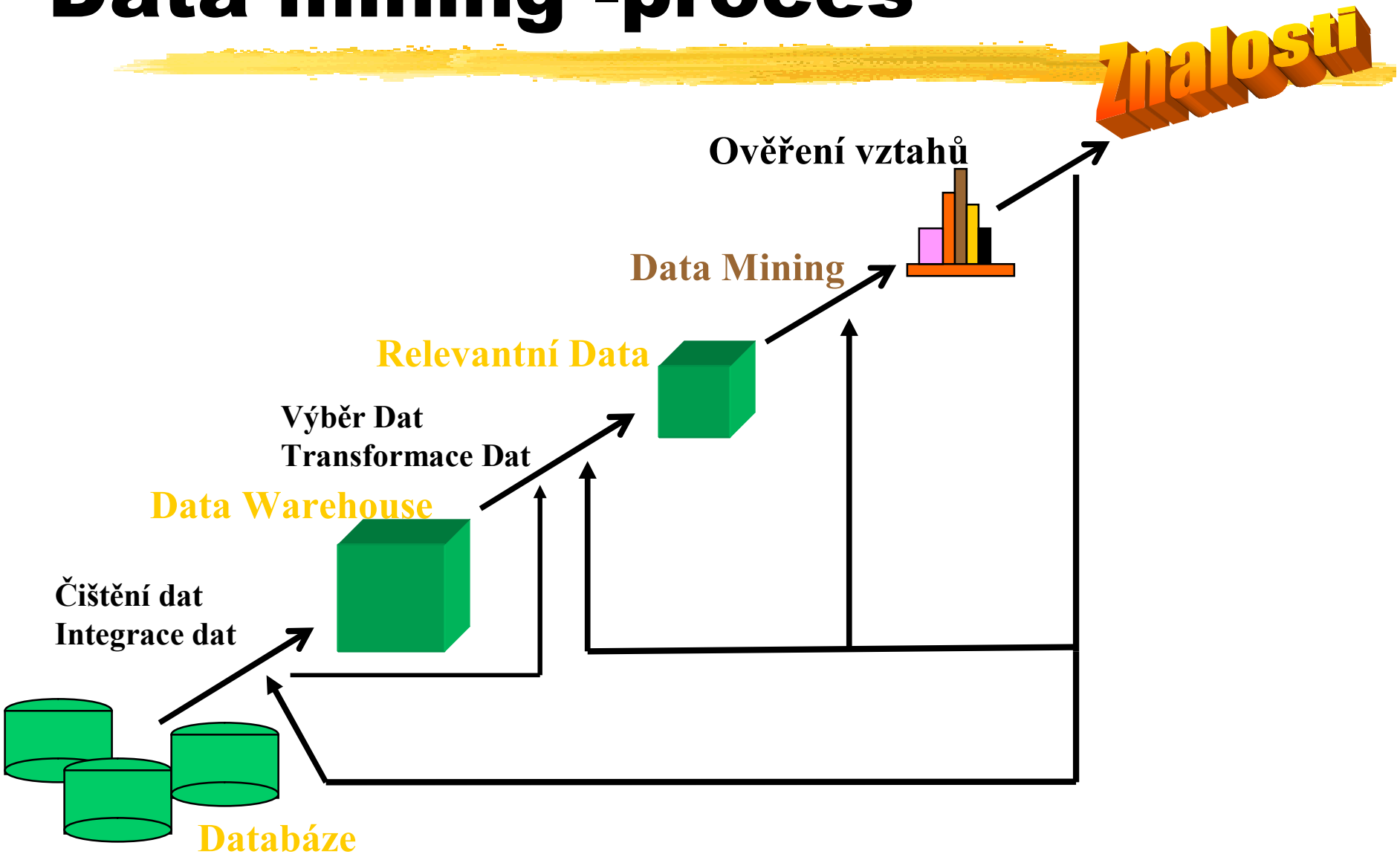
Známé ≠ Zajímavé

- Zajímavé jsou ty vztahy, které se liší od obecných očekávání
- Data mining se vyplácí právě díky objevování dosud neznámých a překvapivých vztahů

Vztah s ostatními disciplínami

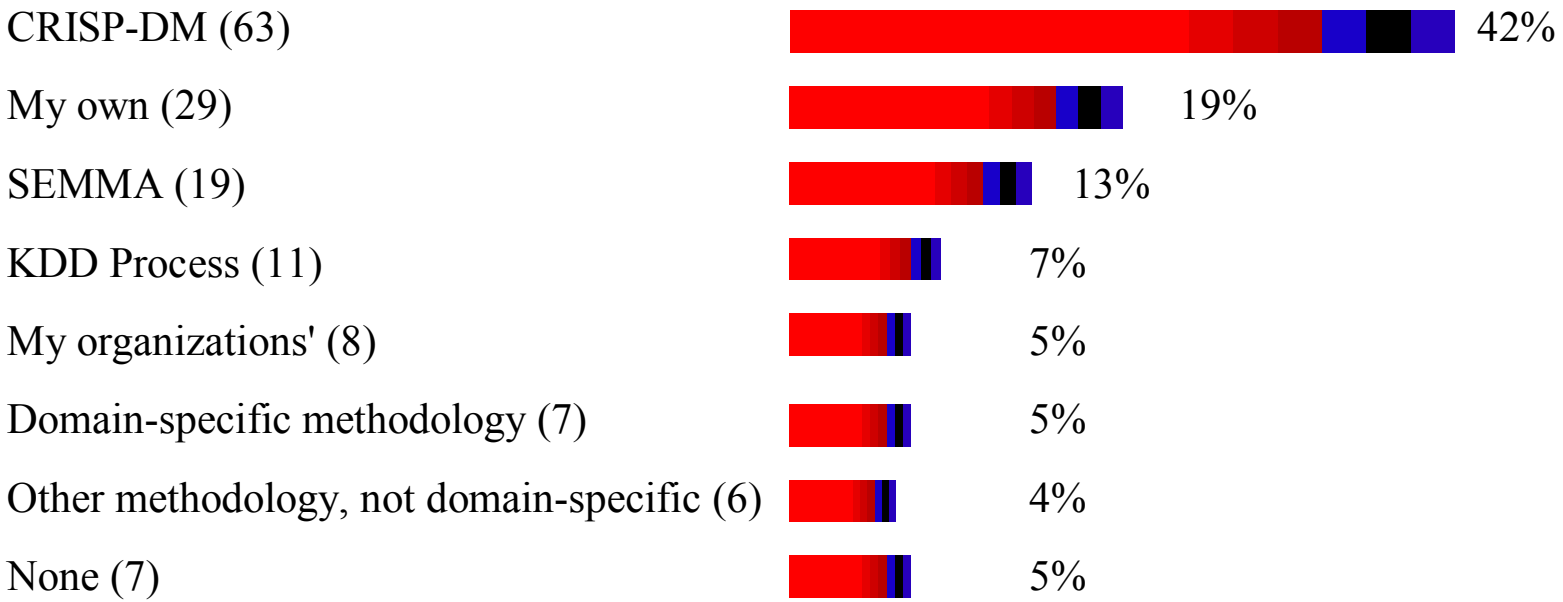


Data mining -proces



Data Mining Methodology (Aug 2007)

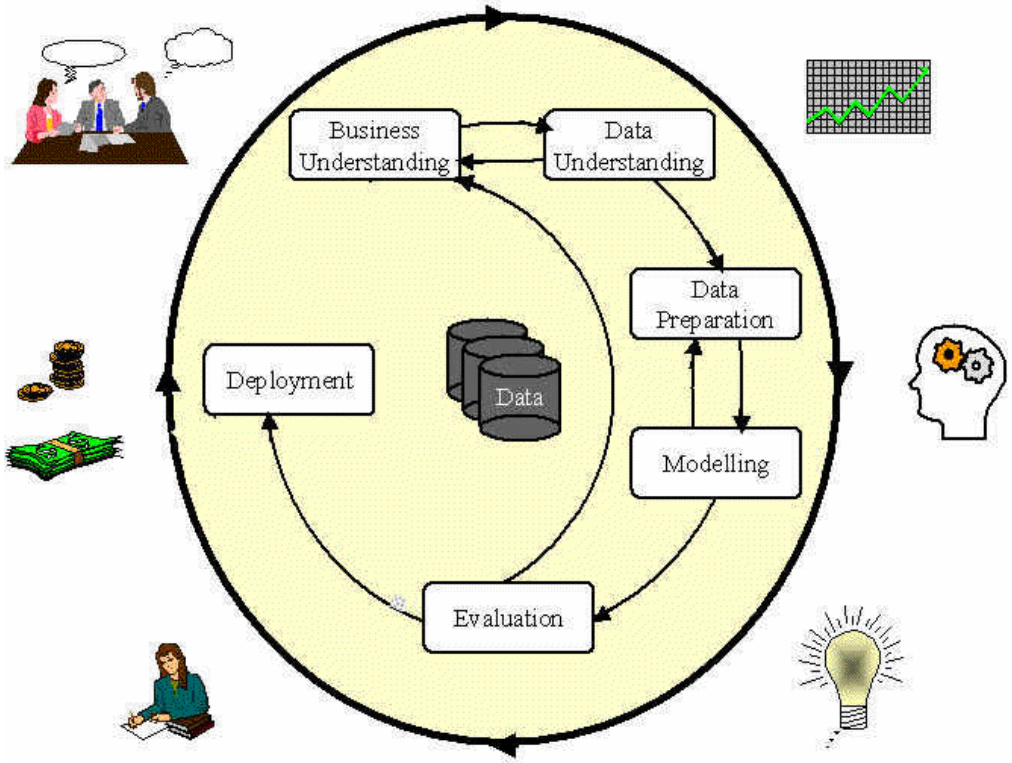
What main methodology are you using for data mining?



CRISP-DM

(*CRoss Industry Standard Process for Data Mining)*

1. pochopení obchodních souvislostí
2. pochopení dat
3. příprava dat
4. modelování
5. vyhodnocení modelu
6. nasazení modelu do obchodního procesu



SEMMA

(Sample, Explore, Modify, Model, Assess)

• **Sample** (optional) your data by extracting a portion of a large data set big enough to contain the significant information, yet small enough to manipulate quickly. For optimal cost and performance, SAS Institute advocates a sampling strategy, which applies a reliable, statistically representative sample of large full detail data sources. Mining a representative sample instead of the whole volume reduces the processing time required to get crucial business information. If general patterns appear in the data as a whole, these will be traceable in a representative sample. If a niche is so tiny that it's not represented in a sample and yet so important that it influences the big picture, it can be discovered using summary methods. We also advocate creating partitioned data sets with the Data Partition node:

Training -- used for model fitting.
Validation -- used for assessment and to prevent over fitting.
Test -- used to obtain an honest assessment of how well a model generalizes.

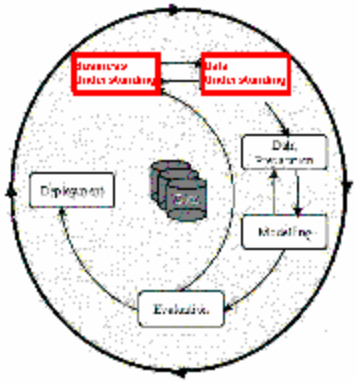
• **Explore** your data by searching for unanticipated trends and anomalies in order to gain understanding and ideas. Exploration helps refine the discovery process. If visual exploration doesn't reveal clear trends, you can explore the data through statistical techniques including factor analysis, correspondence analysis, and clustering. For example, in data mining for a direct mail campaign, clustering might reveal groups of customers with distinct ordering patterns. Knowing these patterns creates opportunities for personalized mailings or promotions.

• **Modify** your data by creating, selecting, and transforming the variables to focus the model selection process. Based on your discoveries in the exploration phase, you may need to manipulate your data to include information such as the grouping of customers and significant subgroups, or to introduce new variables. You may also need to look for outliers and reduce the number of variables, to narrow them down to the most significant ones. You may also need to modify data when the "mined" data change. Because data mining is a dynamic, iterative process, you can update data mining methods or models when new information is available.

• **Model** your data by allowing the software to search automatically for a combination of data that reliably predicts a desired outcome. Modeling techniques in data mining include neural networks, tree-based models, logistic models, and other statistical models -- such as time series analysis, memory-based reasoning, and principal components. Each type of model has particular strengths, and is appropriate within specific data mining situations depending on the data. For example, neural networks are very good at fitting highly complex nonlinear relationships.

• **Assess** your data by evaluating the usefulness and reliability of the findings from the data mining process and estimate how well it performs. A common means of assessing a model is to apply it to a portion of data set aside during the sampling stage. If the model is valid, it should work for this reserved sample as well as for the sample used to construct the model. Similarly, you can test the model against known data. For example, if you know which customers in a file had high retention rates and your model predicts retention, you can check to see whether the model selects these customers accurately. In addition, practical applications of the model, such as partial mailings in a direct mail campaign, help prove its validity.

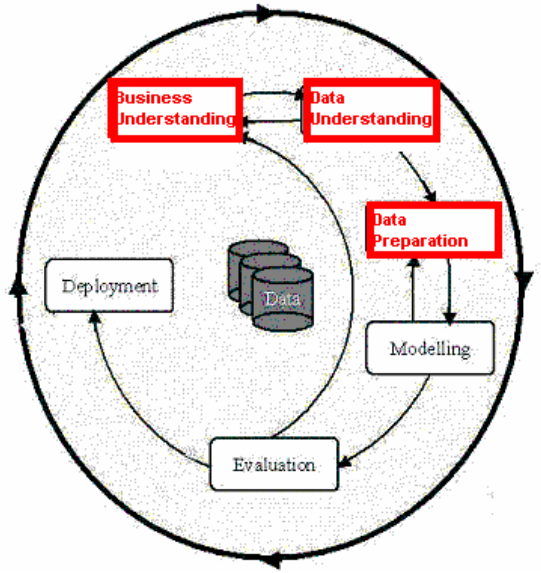
Phases in the DM Process (1 & 2)



- Business Understanding:
  - Statement of Business Objective
  - Statement of Data Mining objective
  - Statement of Success Criteria
- Data Understanding
  - Explore the data and verify the quality
  - Find outliers

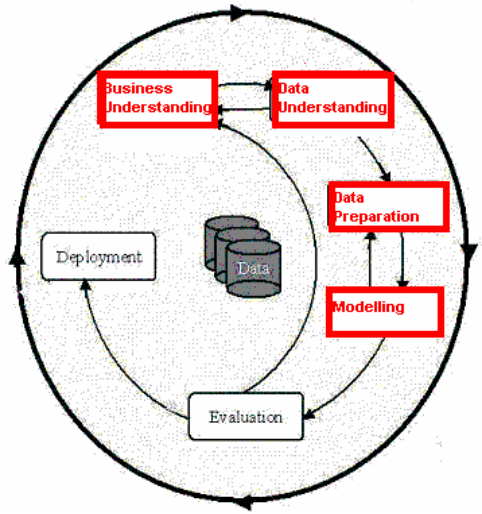
Fáze DM procesu (3)

- Příprava dat:
  - Obvykle zabírá přes 90% celkové času
    - Sběr dat
    - Konsolidace a čištění
      - Vazební tabulky, agregace, chybějící hodnoty,…
    - Selekce
      - Ignorování neužitečných dat?
      - Odlehlá pozorování?
      - Výběr dat?
      - Vizualizační nástroje.
    - Transformace – vytváření nových odvozených proměnných



Phases in the DM Process (4)

- Model building
  - Selection of the modeling techniques is based upon the data mining objective
  - Modeling is an iterative process - different for *supervised* and *unsupervised learning*

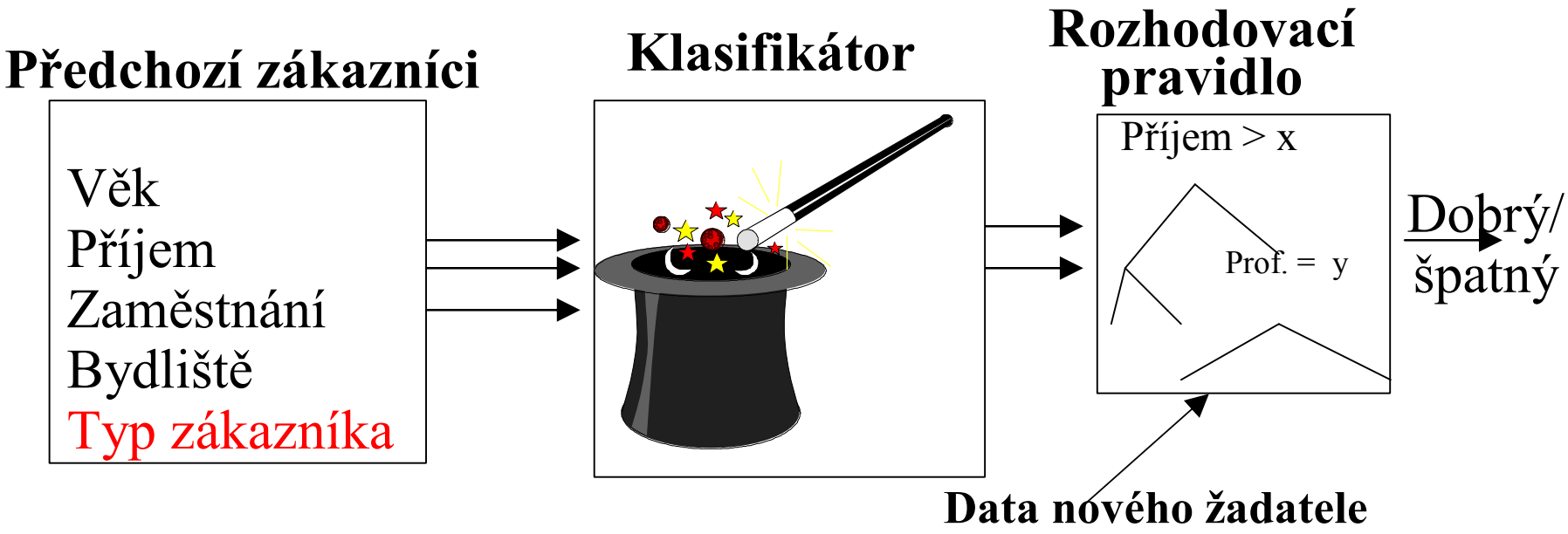


Základní přístupy k modelování

- Prediktivní:
  - Regrese/ Klasifikace
  - Analýza časových řad
- Deskriptivní:
  - Klastrová analýza
  - Asociační pravidla
  - Detekce deviací/zlomů

Klasifikace

- Na základě známých údajů o „starých“ zákaznících a jejich platební morálce máme predikovat platební způsobilost nového žadatele o úvěr.

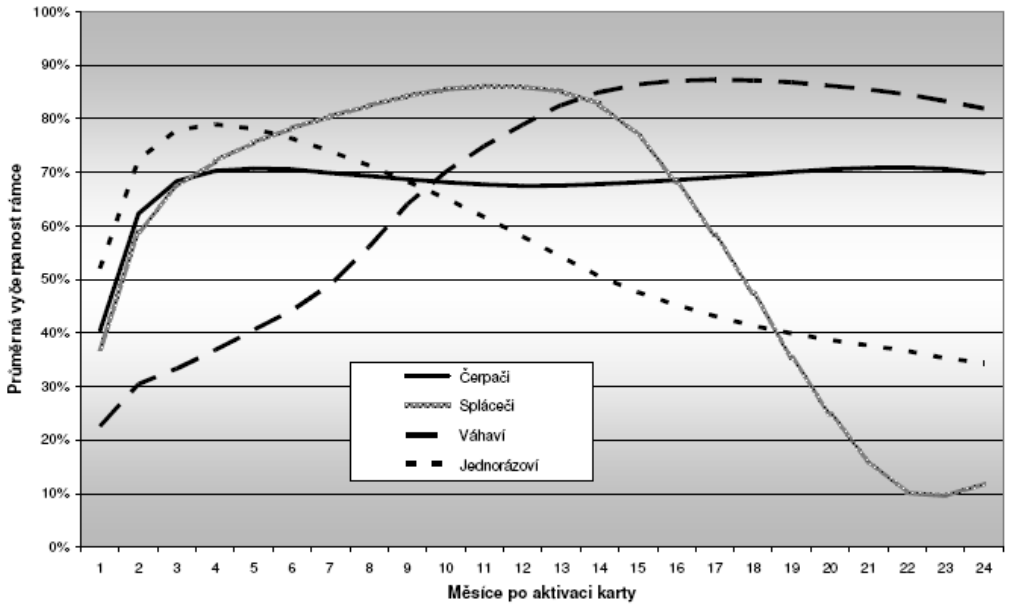


Klasifikační metody

- Cíl: Predikovat třídu Ci = f(x1, x2, .. Xn)
- Regrese: (lineární nebo polynomiální)
  - a*x1 + b*x2 + c = Ci.
- Metody nejbližšího souseda.
- Rozhodovací stromy
- Pravděpodobnostní modely (GLM)
- Neuronové sítě

Klastrová analýza

- Máme nalézt skupiny/ klastry stávajících zákazníků na základě platební historie tak, aby podobní klienti byli ve stejné skupině/ klastru.
- Základní požadavek: Kvalitní míra podobnosti (http://cs.wikipedia.org/wiki/Shluková_analýza).

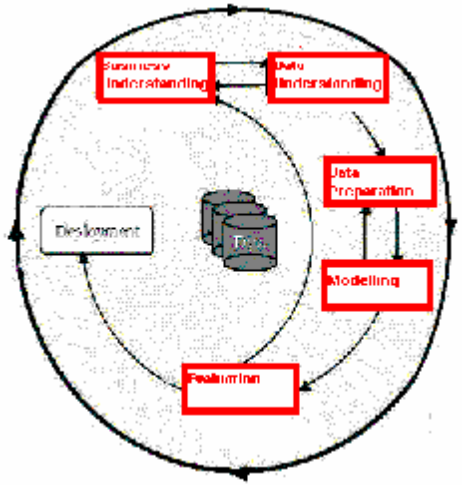


Klastrovací metody

- Hierarchická klastrová analýza
  - agglomerativní / divizivní
  - Jednospojová (single link) /všespojová (complete link)
- K-means
- …

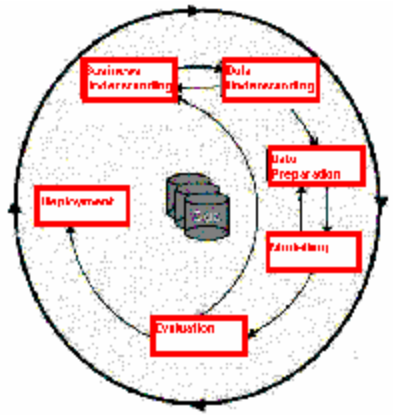
Phases in the DM Process (5)

- Model Evaluation
  - Evaluation of model: how well it performed on test data
  - Methods and criteria depend on model type:
    - e.g., coincidence matrix with classification models, mean error rate with regression models
  - Interpretation of model: important or not, easy or hard depends on algorithm



Phases in the DM Process (6)

- Deployment
  - Determine how the results need to be utilized
  - Who needs to use them?
  - How often do they need to be used
- Deploy Data Mining results by:
  - Scoring a database
  - Utilizing results as business rules
  - interactive scoring on-line



Miningový software

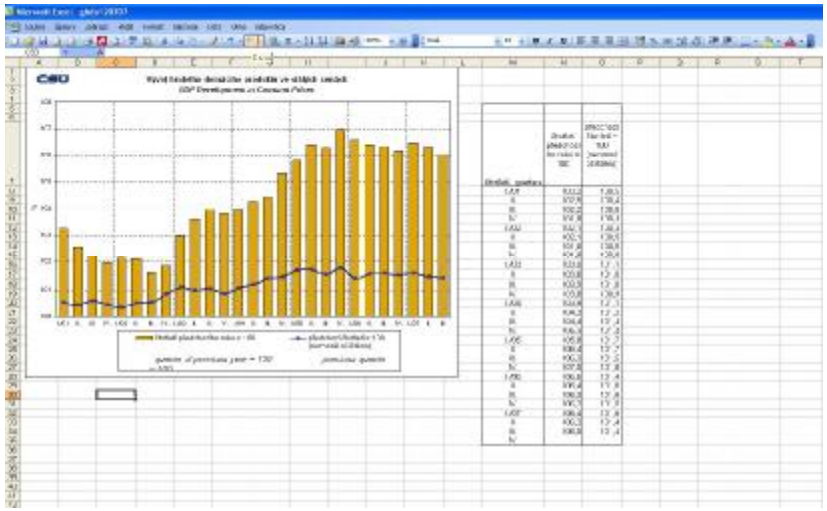
- Cca 20 až 30 dodavatelů
- Hlavní hráči na trhu:
  - Clementine,
  - IBM's Intelligent Miner,
  - SGI's MineSet,
  - SAS's Enterprise Miner.
- Řada vestavěných produktů:
  - fraud detection:
  - electronic commerce applications,
  - health care,
  - customer relationship management

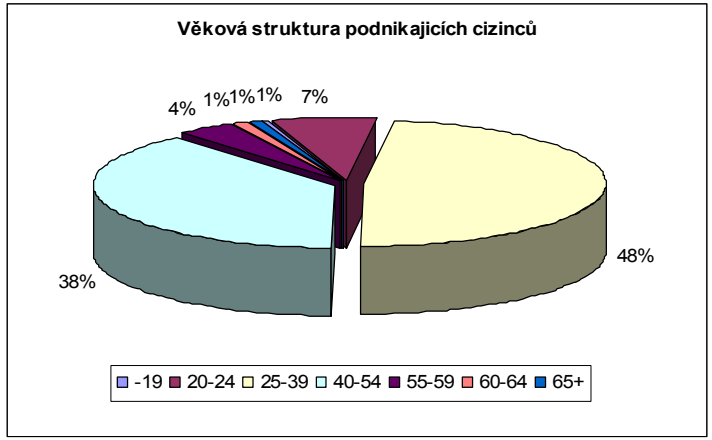
Software

AcaStat	GAUSS	MRDCL	RATS	StatsDirect
ADaMSoft	GAUSS	NCSS	RKWard[4]	Statistix
Analyse-it	GenStat	OpenEpi	SalStat	SYSTAT
ASReml	Golden Helix	Origin	SAS	The Unscrambler
Auguri	gretl	Ox programming language	SOCR	UNISTAT
BioStat	JMP	OxMetrics	Stata	VisualStat
BrightStat	MacAnova	Origin	Statgraphics	Winpepi
Dataplot	Mathematica	Partek	STATISTICA	WinSPC
EasyReg	Matlab	Primer	StatIt	XLStat
Epi Info	MedCalc	PSPP	StatPlus	XploRe
EViews	modelQED	R	SPlus	
Excel	Minitab	R Commander[4]	SPSS	

Software

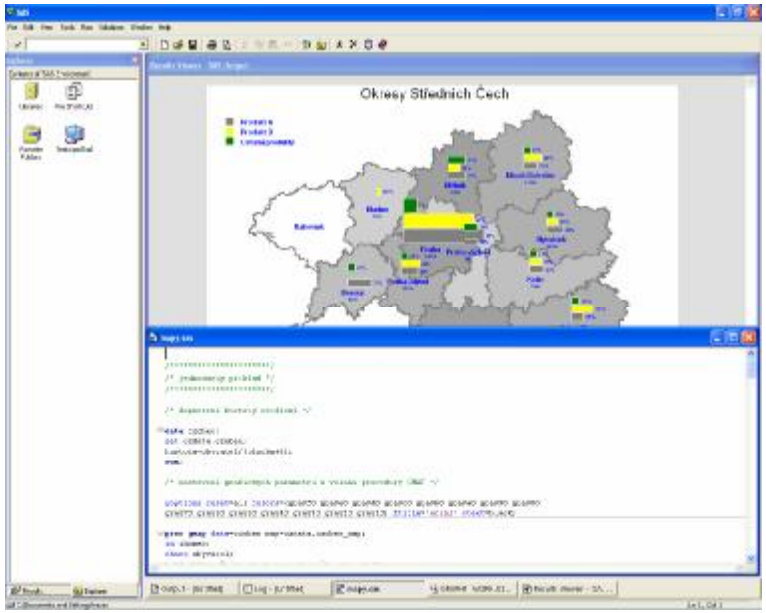
- MS Excel: office.microsoft.com/en-us/excel

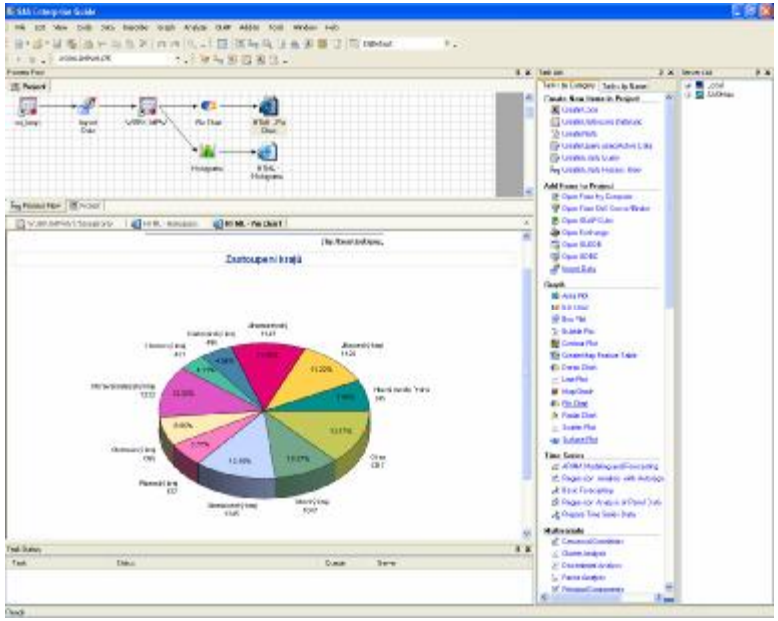




Software

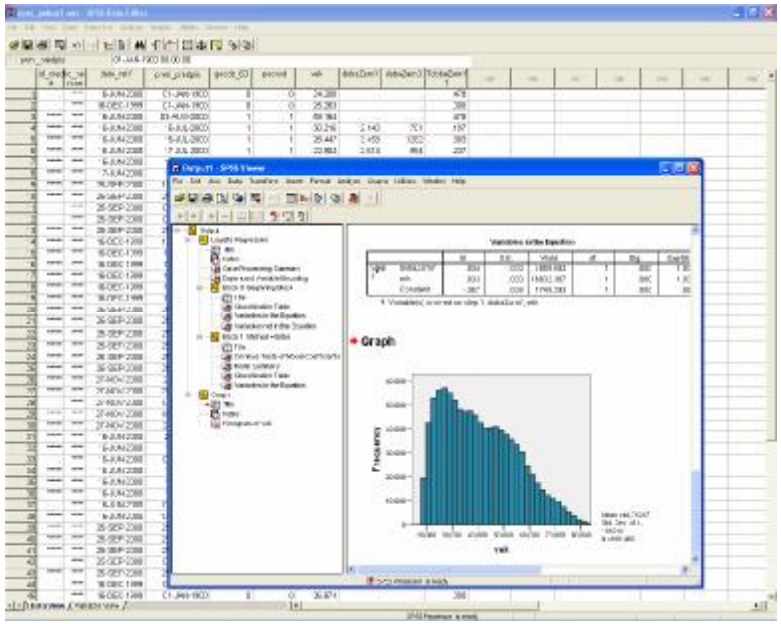
⌘ SAS : www.sas.com

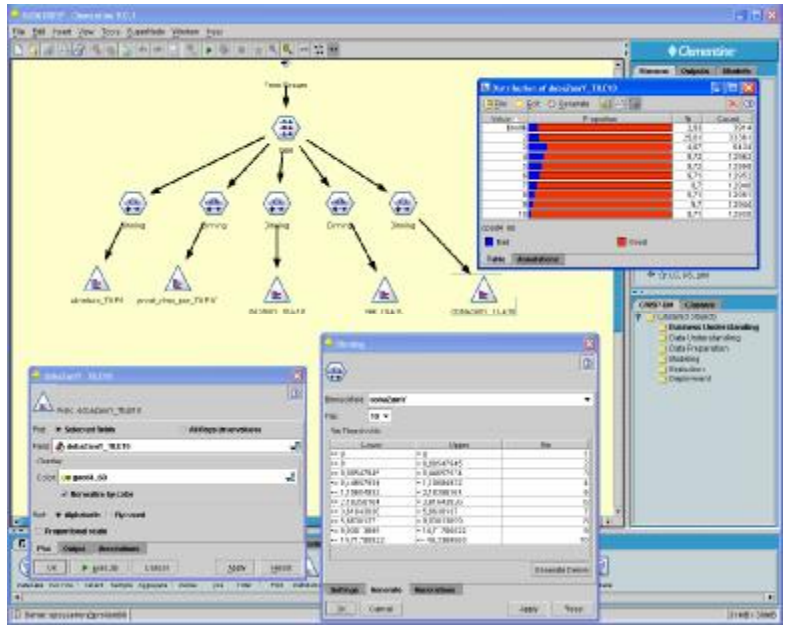




Software

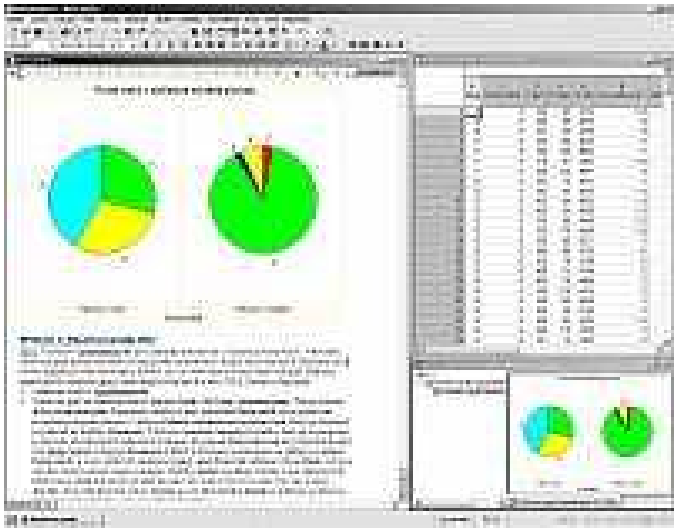
⌘ SPSS : www.spss.cz

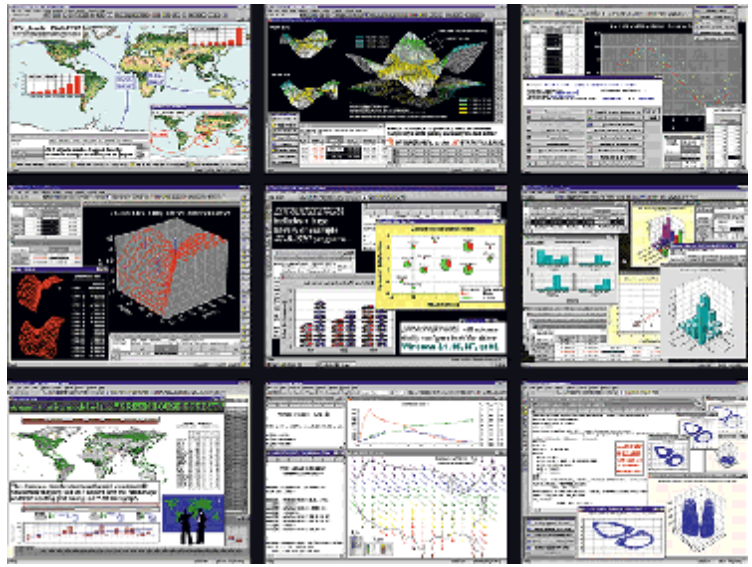




Software

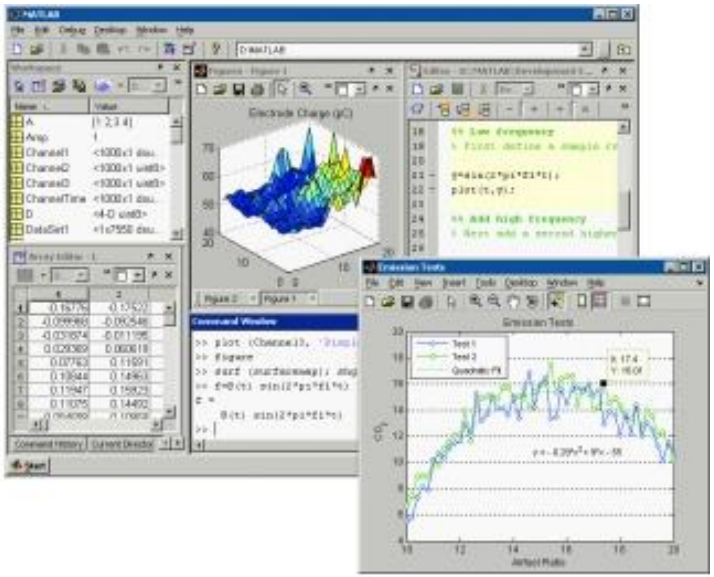
- Statistica: www.statistica.cz

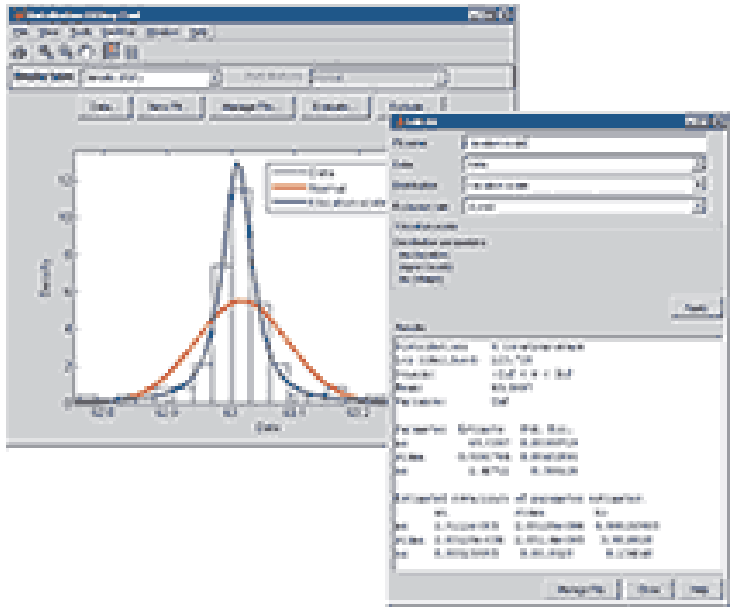




Software

- Matlab: www.mathworks.com, www.humusoft.cz





Software

- Eviews: ww.eviews.com