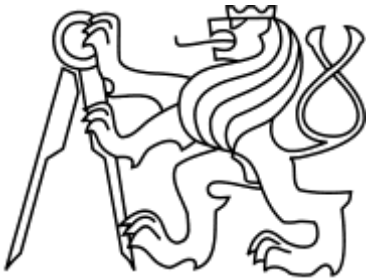
Dobývání a vizualizace znalostí

Olga Štěpánková

Lenka Vysloužilová



Nature Inspired
Technologies Group

Osnova přednášky

- Úvod: data a jejich rostoucí objem
- **Příklady úspěšných aplikací využívání znalostí**
- Typické nástroje pro získávání znalostí
- Vytěžování dat (Data Mining) & dobývání znalostí (Knowledge Discovery)
- Typické postupy DM

Hlavní zdroj inspirace:
www.kdnudgets.com

Dr. Gregory Piatetsky-Shapiro (KDnuggets)
Prof. Gary Parker (Connecticut College)

Kde se bere současná záplava dat?

- Digitální data a archivace.
- Archivace a její meze.

Oblasti:

- Obchodní transakce (obchodní řetězce, banky, pojišťovny ...)
- Telekomunikace, internet a elektronický obchod
- Zdravotnictví
- Věda a výzkum: astronomie, biologie, genomika, …
- Publikace: texty, časopisy a knihy …

Záplava dat?

Prefix	Násobek
mega	10^6
giga	10^9
tera	10^{12}
peta	10^{15}
exa	10^{18}
zetta	10^{21}
yotta	10^{24}

- *Ancestry.com* má asi **600 terabytů** genealogických dat zahrnující *US Census* data z let 1790 až 1930.
- Data předávaná přes Internet: v roce 1993 asi **100 terabytů**. V r. 2008 odhaduje Cisco, Inetrnetová výměna dat činí asi **160 terabytů/s** (tedy asi 5 zettabytů za rok).
- AT&T zpracovává miliardy spojení za den

Země: méně než $3x10^{50}$ atomů.

Vznikající objemy dat nelze skladovat ani prohlížet. → Je nutné z nich vybírat jen to „důležité“! Role znalostí.

Osnova přednášky

- Úvod: data a jejich rostoucí objem
- **Příklady úspěšných aplikací využívání znalostí**
- Typické nástroje pro získávání znalostí
- Vytěžování dat (Data Mining) & dobývání znalostí (Knowledge Discovery)
- Typické postupy DM

Existující teorie

Jaké znalosti potřebujeme při zpracování dat?
Zdroje znalostí?

Dobývání znalostí (DM).

Oblasti DM aplikací v r. 2004

- 13%: Bankovnictví
- 9%: Přímý marketing, odhalování podvodů, analýza vědeckých dat
- 8%: Bioinformatika
- 7%: pojišťovnictví, medicína/farmacie
- 6%: eCommerce/Web, telekomunikace
- 4%: Akciové/investiční trhy, výroba, obchod, bezpečnost
- Méně: doprava, zábava/zprávy, … (celkem asi 10 %)

DM a chování zákazníků

- Základní typy úloh:
  - Předvídání, že zákazník odejde k jinému poskytovateli služby
  - cílený marketing:
    - Získávání nových zákazníků, nabízení dalších produktů (např. Amazon)
  - Odhad rizika u úvěrů
    - Nejdůvěryhodnější zákazníci nepotřebují půjčku – nejzajímavější zákazníci jsou někde „uprostřed" mezi nejlepšími a nejhoršími !
  - Odhalení podvodného chování
- Cílové oblasti
  - bankvonictví, telekomunikace, *maloobchodní prodej*, …
  - Zdravotnictví, …

Customer Attrition: Case Study

Situation:

- Attrition rate at for mobile phone customers is around 25-30% a year!

Task:

- Given customer information for the past N months, predict who is likely to attrite next month.
- Also, estimate customer value and what is the cost-effective offer to be made to this customer.

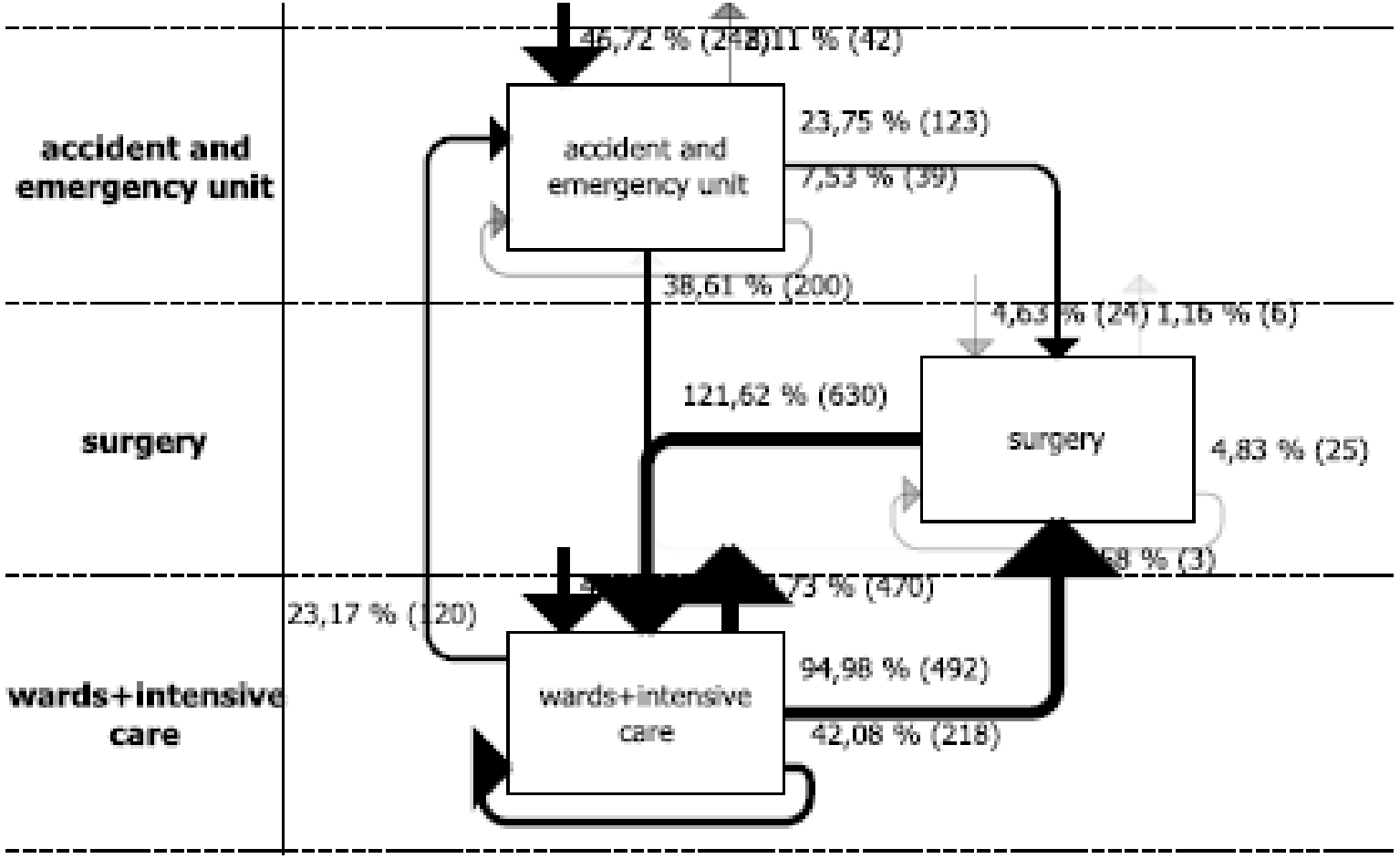
Customer Attrition Results

- Verizon Wireless built a customer data warehouse
- Identified potential attriters
- Developed multiple, regional models
- Targeted customers with high propensity to accept the offer
- Reduced attrition rate from over 2%/month to under 1.5%/month (huge impact, with >30 M subscribers)

(Reported in 2003)

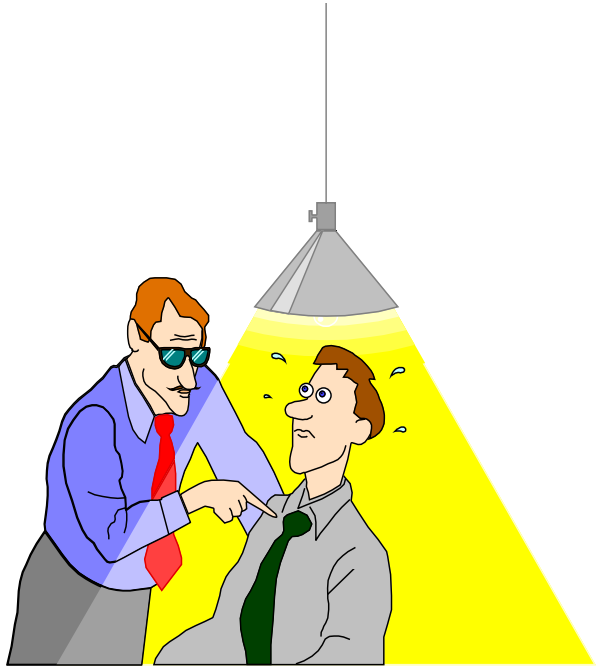
Zdravotnictví

- Na základě kterých fyziologických měření lze provést přesnou diagnózu zvolené choroby?
- Lze nahradit nějaký náročný diagnostický postup souborem jiných měření?
- Volba nejvhodnější medikace či léčebného postupu.
- Včasné upozornění na příznaky, které indikují zvýšené riziko výskytu komplikací někdy v budoucnosti.
- Hledání důvodů (chemické struktury) pro různé chování některých sloučenin (karcinogencita)
- … Analýza probíhajících procesů (ošetření) jako podklad pro zefektivnění péče



Bezpečnost a podezřelé chování

- Krádeže kreditních karet
- Bankovní převody svědčící o praní špinavých peněz
- Analýza struktury telefonních spojení
  - Krádeže telefonů
  - Identifikace teroristických skupin
  - ….
- Otázky etiky a narušování soukromí



Některé příklady z ČVUT

- Krátkodobá predikce požadavků na spotřebu energie
- Včasná prediktivní diagnostika
  - chyb motoru čerpadla
  - zhoršení stavu pacienta (psychické choroby, ..)
  - …
- Podpora diagnostiky na základě přesných měření
- Hledaní charakteristických vzorů chování pro uživatele
- Studie genomických dat
- …

Osnova přednášky

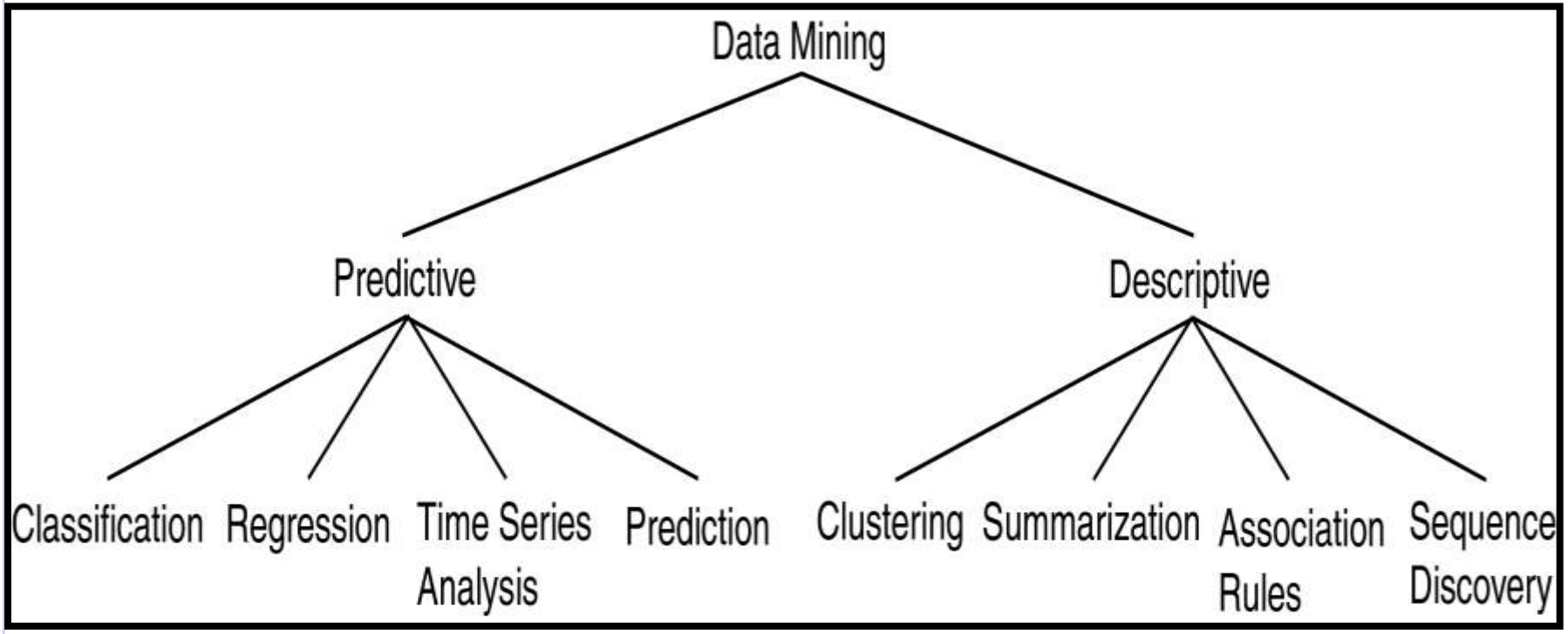
- Úvod: data a jejich rostoucí objem
- Příklady úspěšných aplikací využívání znalostí
- **Typické nástroje pro získávání znalostí**
- Vytěžování dat (Data Mining) & dobývání znalostí (Knowledge Discovery)
- Typické postupy DM

Strojové učení a návrh algoritmů, pomocí nichž lze získat potřebnou znalost

Hlavní typy úloh

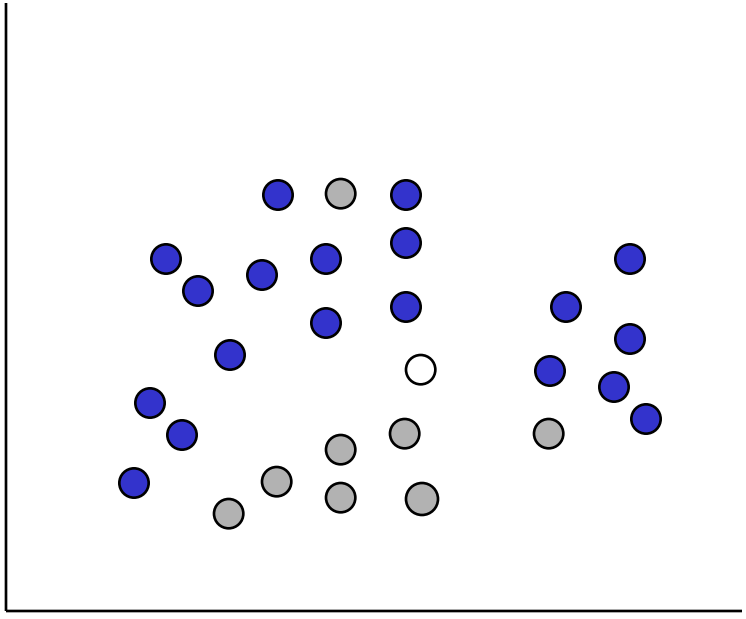
- **Klasifikace:** zařazení případu do jedné z několika tříd
- **Odhad budoucí hodnoty**: pro spojité veličiny
- **Shlukování:** identifikace charakteristických shluků v datech
- **Asociace:** např. analýza nákupního koše (A & B & C se často vyskytují společně)
- **Vizualizace dat:** jako podpora člověka zpracovávajícího data
- Analýza komplexních vztahů a vlivů: např. analýza genetických dat
- …

Často používané modely a úlohy



Klasifikace

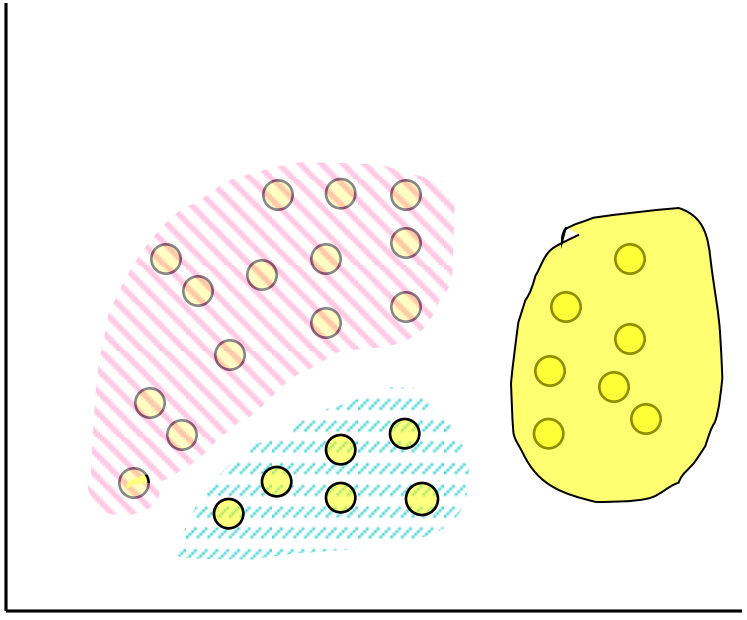
- Úloha: Na základě klasifikovaných trénovacích dat nalezněte „jednoduchou metodu“, jak přiřadit třídu novým případům, pro které známe stejný soubor příznaků



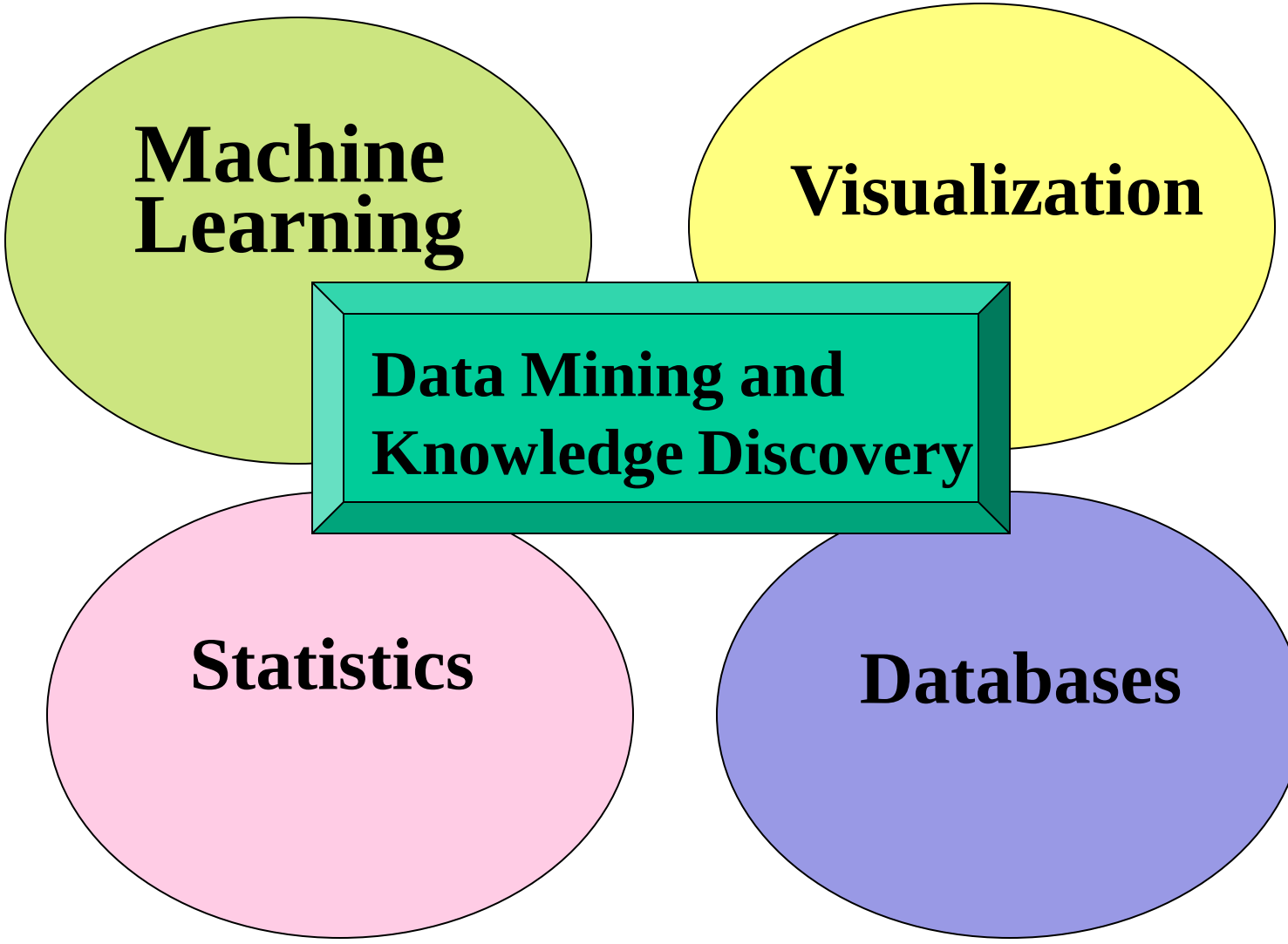
Postupy:
statistika,
Rozhodovací stromy,
Neuronové sítě,
...

Shlukování

- Úloha: Nalezněte „přirozené“ shluky ve zpracovávaných datech, která nemají žádné značky



Related Fields

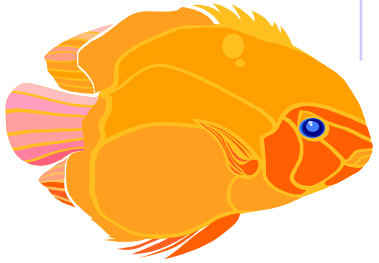


Osnova přednášky

- Úvod: data a jejich rostoucí objem
- Příklady úspěšných aplikací využívání znalostí
- Typické nástroje pro získávání znalostí
- **Vytěžování dat** (Data Mining) **& dobývání znalostí** (Knowledge Discovery)
- Typické postupy DM

Liší se nějak ?

Trocha historie



- Data Fishing, Data Dredging: 1960-
- **Data Mining** :1990 --
  - used DB, business
  -
- Data Archaeology, Information Harvesting, Information Discovery, Knowledge Extraction, ...
- **Knowledge Discovery in Databases** (1989-)
  - komunita zabývající se umělou inteligencí, strojovým učením

Termíny **Vytěžování dat** (Data Mining) a **Dobývání znalostí** (Knowledge Discovery) se chápou jako **synonyma**

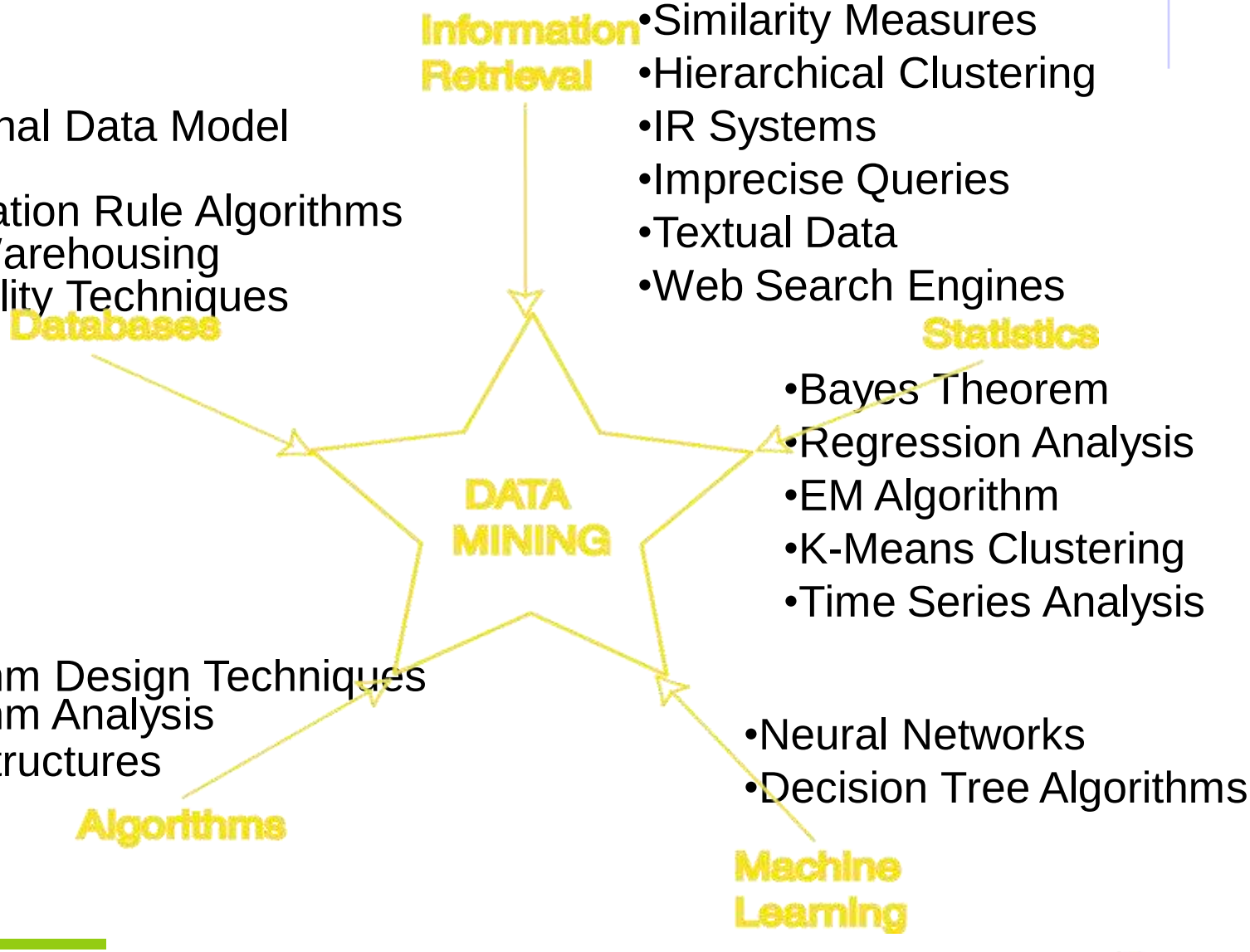
Definice

Dobývání znalostí z dat je *netriviální* process, který vede k identifikaci

- *platných*
- *nových*
- potenciálně *užitečných*
- a *lidskému uživateli srozumitelných vzorů* v datech.

viz *Advances in Knowledge Discovery and Data Mining,* Fayyad, Piatetsky-Shapiro, Smyth, and Uthurusamy, (Chapter 1), AAAI/MIT Press 1996

Podpůrné zdroje

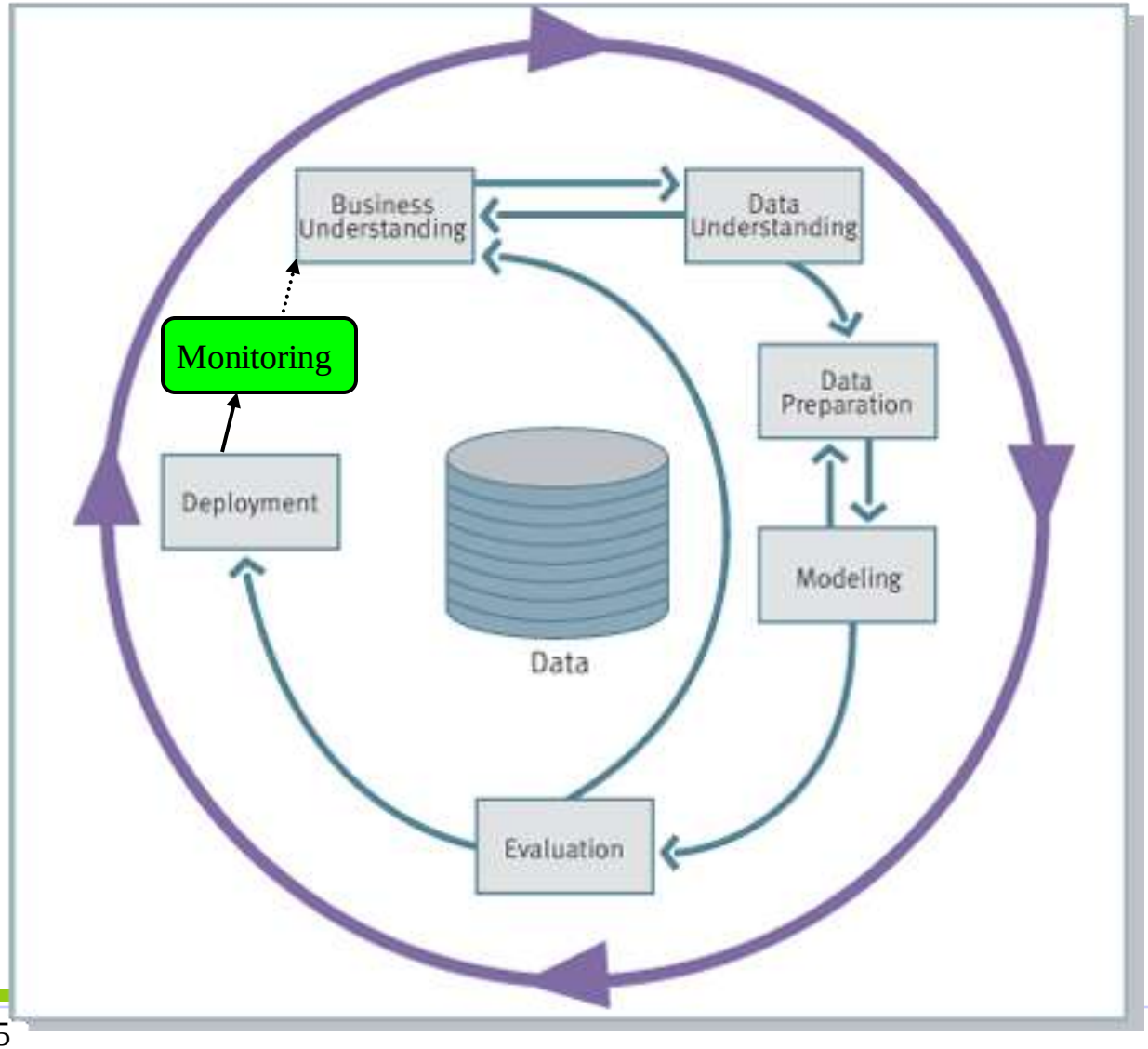


Osnova přednášky

- Úvod: data a jejich rostoucí objem
- Příklady úspěšných aplikací využívání znalostí
- Typické nástroje pro získávání znalostí
- Vytěžování dat (Data Mining) & dobývání znalostí (Knowledge Discovery)
- **Typické postupy DM**

Dobývání znalostí jako proces

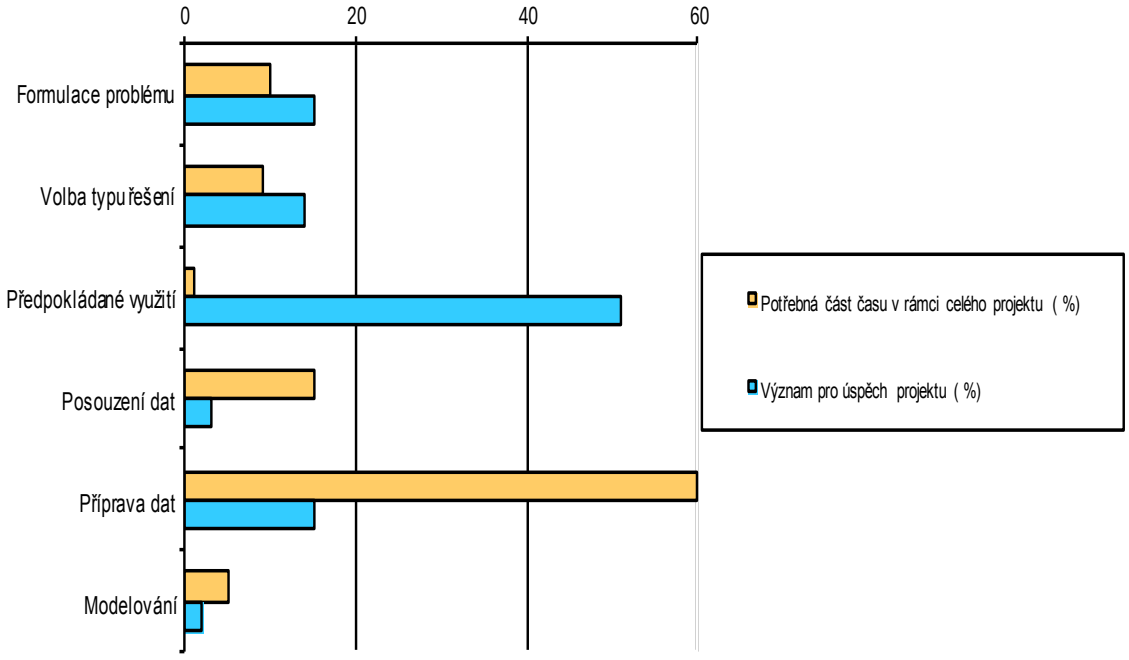
Proces dobývání znalostí CRISP-DM



Metodika **CRISP- DM** CRoss Industry Standard for Process of DM, viz www.crisp-dm.org

Význam kroků CRISP (v %):

celkové časové nároky a úspěch DM řešení



KDD Issues

- Human Interaction
- Overfitting
- Outliers
- Interpretation
- Visualization
- Large Datasets
- High Dimensionality
- Multimedia Data
- Missing Data
- Irrelevant Data
- Noisy Data
- Changing Data
- Integration
- Application

Data Miner Survey 2010:

http://www.rexeranalytics.com/Data-Miner-Survey-Results-2010.html

ALGORITHMS: *Decision trees*, *regression*, and *cluster analysis* remain on the top! *Ensemble Models* are becoming popular - 22% of data miners use them.

MODELS: About one-third of data miners typically build final models with 10 or fewer variables, while about 28% generally construct models with more than 45 variables.

TOOLS: Success of the open source data mining software **R** (used by more data miners – 43% - than any other). **SumatraTT**, **Weka, RapidMiner, LispMiner**

- STATISTICA is selected as the *primary* data mining tool by the most data miners (18%).
- Data miners report using an average of **4.6** software tools overall. STATISTICA, IBM SPSS Modeler, and R received the strongest satisfaction ratings in both 2010 and 2009.

Osnova předmětu

1. Dobývání znalostí – popis a metodika procesu CRISP, motivační příklady
2. Nástroje pro modelování dat a jejich využití I.
3. Porozumění datům a jejich příprava, agregace dat
4. Metody vizualizace dat. Identifikace odlehlých a nesprávných hodnot
5. Práce s časovými řadami
6. Volba relevantních atributů
7. Nástroje pro modelování dat a jejich využití II.
8. Vyhodnocení a využití modelů
9. Metody pro vizualizaci modelů
10. Zpracování přirozeného jazyka jako vstupu
11. *Databáze, anonymizace a ochrana dat*
12. *Slučování dat z heterogenních zdrojů*
13. *Postupy zpracování komplexních dat*

Prerekvizity: Přehled základních pojmů ze statistiky, databáze a datové sklady

Doporučené zdroje

P. Berka: *Dobývání znalostí z databází*, Academia 2003

M. Kubát: Strojové učení v Mařík et al. (eds) *Umělá inteligence* (1), Academia 1993

F.Železný, J.Kléma, O.Štěpánková: Strojové učení v dobývání dat v Mařík et al. (eds) *Umělá inteligence* (4), Academia 2003

S. Few: Simple Visualization Techniques for Quantitative Analysis – Now you see it. Analytics Press 2009

Michael Berthold, David J. Hand: *Intelligent Data Analysis*, Springer 1999, 2003

Daniel T. Larose: *Discovering Knowledge in Data*, Wiley 2005

Daniel T. Larose: *Data Mining: Methods and Models*, Wiley 2006

Oded Maimon, Lior Rokach (eds): *The Data Mining and Knowledge Discovery Handbook*, Springer 2005