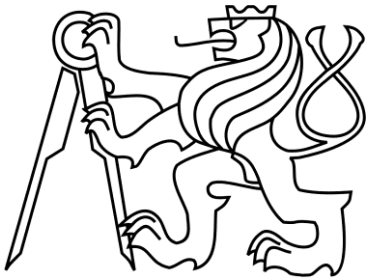
Dobývání a vizualizace znalostí

Olga Štěpánková

et al.



Nature Inspired Technologies Group

Osnova předmětu

- Dobývání znalostí - popis a metodika procesu CRISP a objasnění základních pojmů
- Nástroje pro modelování klasifikovaných dat a jejich využití I (učení s učitelem): algoritmy pro tvorbu rozhodovacího stromu
- Nástroje pro modelování neklasifikovaných dat a jejich využití II (učení bez učitele): algoritmy pro shlukování
- Vyhodnocení modelů: křížová validace, bootstrapping, ROC křivka, křivka učení.
- Metody vizualizace dat (box graf, histogram, scatter plot matrix, paralelní souřadnice, RadViz) a jejich použití.
- Porozumnění datům a jejich příprava
- Další nástroje pro modelování III: konstrukce asociačních pravidel
- Další nástroje pro modelování dat IV: neuronové sítě a jejich aplikace
- Tvorba modelu kombinací více základních modelů - bagging, boosting, AdaBoost.
- Práce s časovými řadami.
- Zpracování přirozeného jazyka jako vstupu: text mining
- Využití doménové znalosti a relační strojové učení

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

Osnova přednášky

- Úvod: data a jejich rostoucí objem
- Vytěžování dat (Data Mining) & dobývání znalostí (Knowledge Discovery) a související pojmy
- Typické postupy DM – metodika CRISP-DM a její fáze

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, Internetová výměna dat činí asi **160 terabytů/s** (tedy asi 5 zettabytů za rok).
- AT&T zpracovává miliardy spojení za den

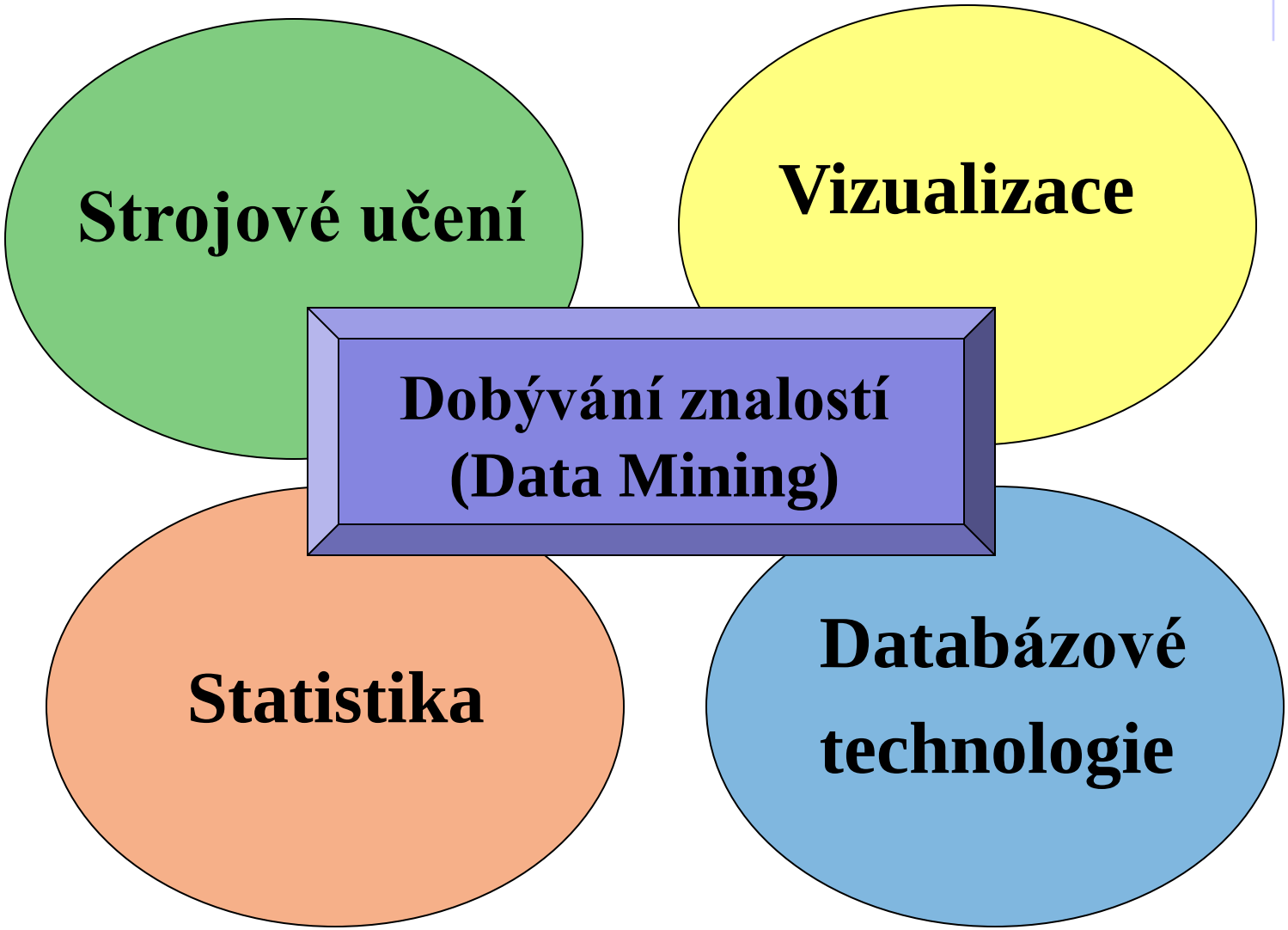
Země: méně než $3x10^{50}$ atomů.

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

Dobývání znalostí z dat

- **Cíl:** částečná automatizace procesu získání **zajímavých vzorů chování z reálných dat:** tvorba jejich modelů - pomocí nástrojů strojového učení, statistiky, databázových technologií,...
- Nové slibné odvětví SW průmyslu, jehož cílem je využít existující data pro **zlepšení rozhodovacích procesů a získání nových znalostí**

Souvislosti



Dobývání znalostí z dat

Příklady aplikací:

- průmysl (diagnostika poruch, predikce spotřeby …)
- obchod (marketing, bankovnictví)
- věda (charakterizace karcinogenních látek)
- medicína (mapování lidského genomu, personalizace medicíny)

Definice dobývání znalostí

Data Mining is the

non-trivial process of identifying

- *valid*
- *novel*
- potentially *useful*
- and ultimately *understandable patterns* in data.

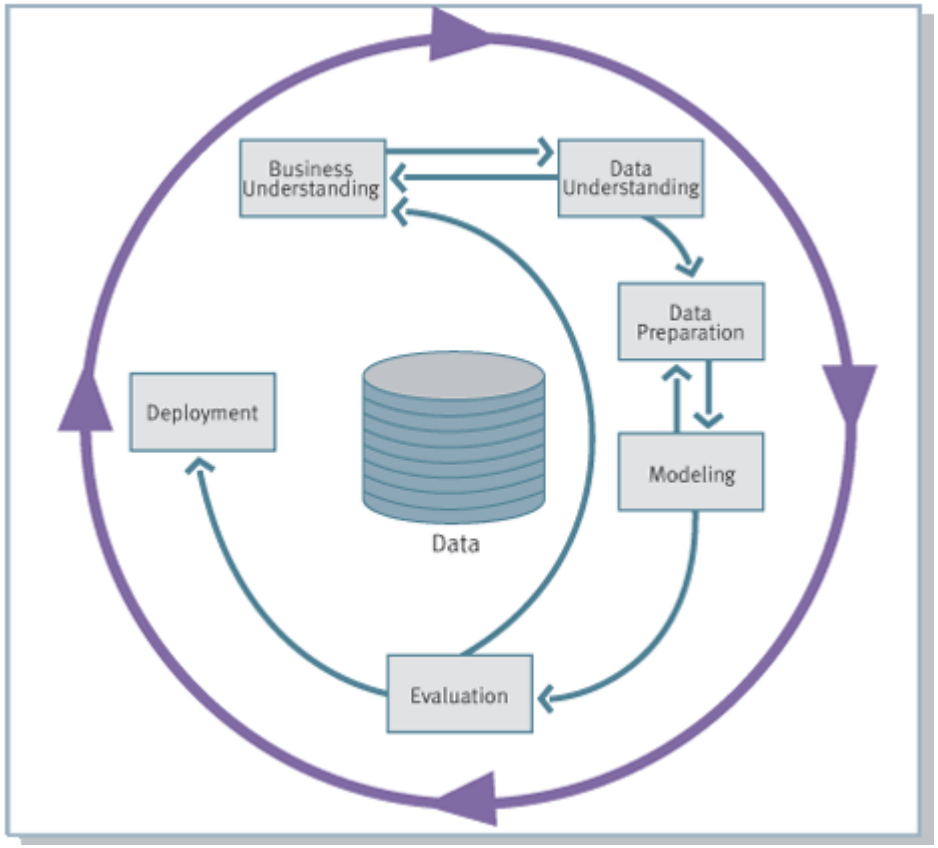
from *Advances in Knowledge Discovery and Data Mining,* Fayyad, Piatetsky-Shapiro, Smyth, and Uthurusamy, (Chapter 1), AAAI/MIT Press 1996

Terminologie

- **Koncept:** oblast zájmu – co chceme
  - předpověď počasí
- **Instance(pozorování):** nezávislé pozorované jednotky
  - data o počasí jednoho konkrétního dne
- **Atributy(příznaky):** jednotlivé vlastnosti instance
  - teplota, tlak, množství srážek
- **Příznakový prostor:** prostor, jehož dimenze jsou definovány jednotlivými příznaky
  - pozorování jsou body v příznakovém prostoru
- **Matice pozorování:** řádky jsou instance a sloupce příznaky

Metodika CRISP-DM

(www.crisp-dm.org)



Zadaní – Business Understanding

- pochopení cílů úlohy
- náklady
- hodnotí se přínos
- stanovení předběžného plánu
- forma předání dat
  - anonymizace dat
  - formát dat

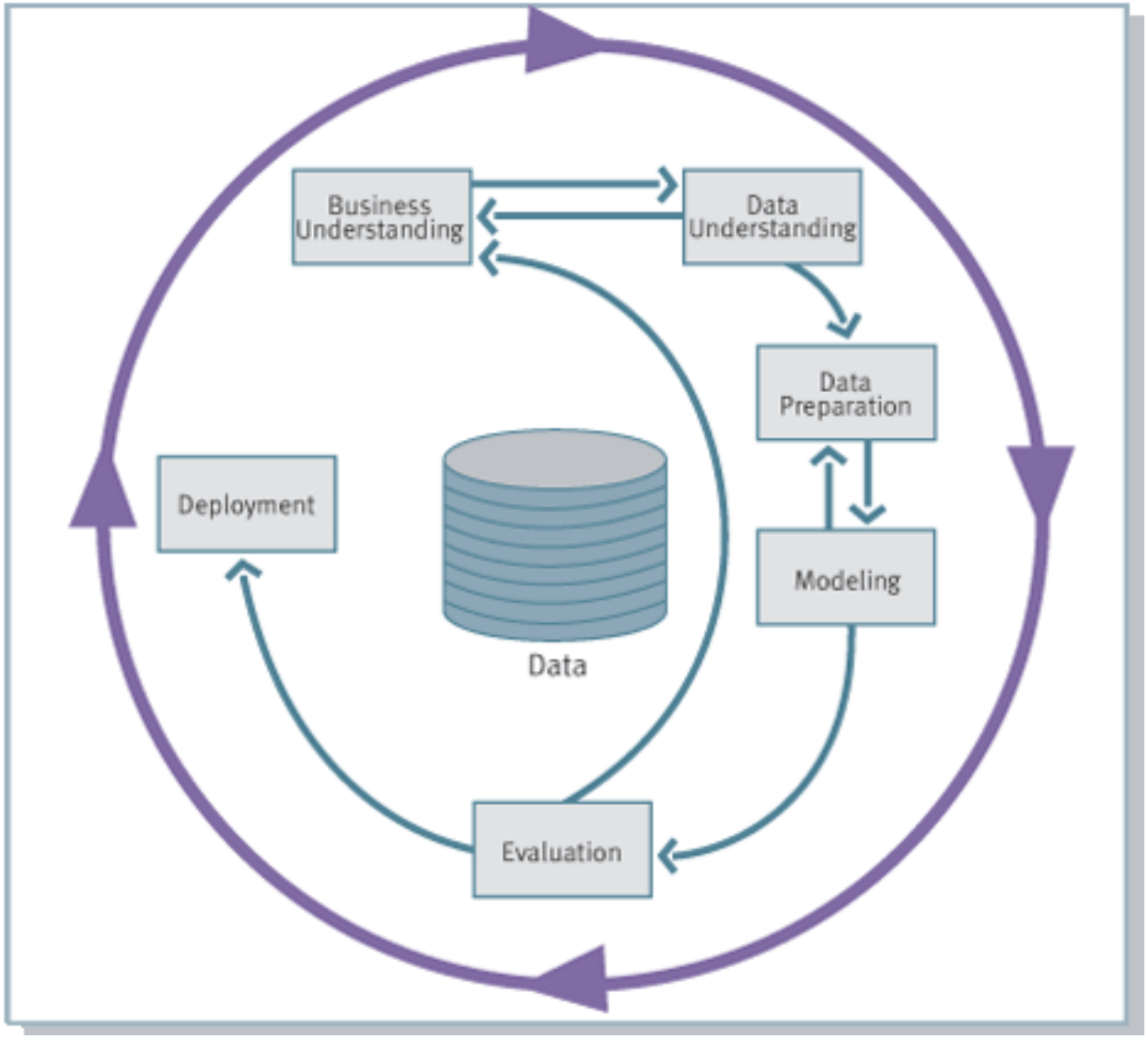
Problémy reálných dat?

- Data obsahují **špatné údaje** způsobené chybami měřicích přístrojů i lidské obsluhy
- **Nevyplněné údaje**
- Data jsou popsána pomocí **příliš mnoha atributů** - není zřejmé, které z nich jsou pro řešení zvolené úlohy relevantní. Úspěch modelování závisí na volbě vhodné množiny atributů (PAC učení)
- Data mají formu **složitého relačního schématu**, nikoliv jediné tabulky předpokládané atributovými metodami strojového učení

Analýza dat – Data Understanding

- získání základní představy o datech
- kvalita dat (chybějící údaje)
- descriptivní charakteristiky dat
  - četnosti hodnot (histogramy)
  - minima, maxima, průměry
- použití vizualizačních technik

Metodika CRISP-DM



Příprava dat – Data Preparation

- příprava dat pro modelování
  - selekce atributů – výběr relevantních atributů
  - čištění dat
  - získávání odvozených atributů
  - převod typů dat
  - transformace dat do jedné velké tabulky
  - formátování pro jednotlivé modelovací techniky
- nejpracnější část celého procesu
- často se provádí opakovaně

Modelování - Modeling

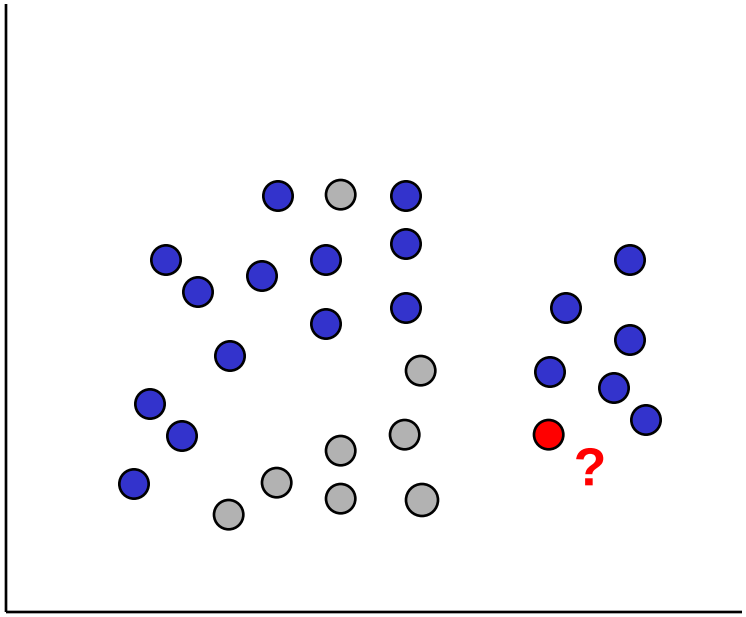
- použití analytických metod (metody strojového učení)
- zkouší se více metod
- příklady metod
  - rozhodovací stromy
  - asociační pravidla
  - shluková analýza
  - statistické metody
- často návrat zpět k přípravě dat

Typy úloh

- **Klasifikace:** přiřazení třídy objektu
- **Predikce:** předpověď chování objektu v čase
- **Asociace:** hledání vazeb mezi objekty
- **Shlukování:** seskupování podobných objektů

Učení s učitelem

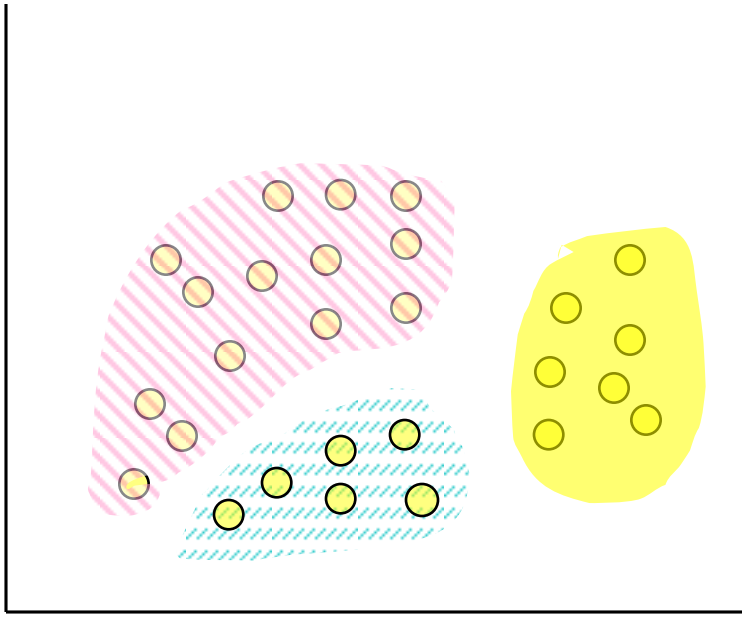
- Úloha: Na základě učitelem klasifikovaných trénovacích dat nalezněte „jednoduchou metodu“, jak přiřadit třídu novým případům, pro které známe stejný soubor příznaků



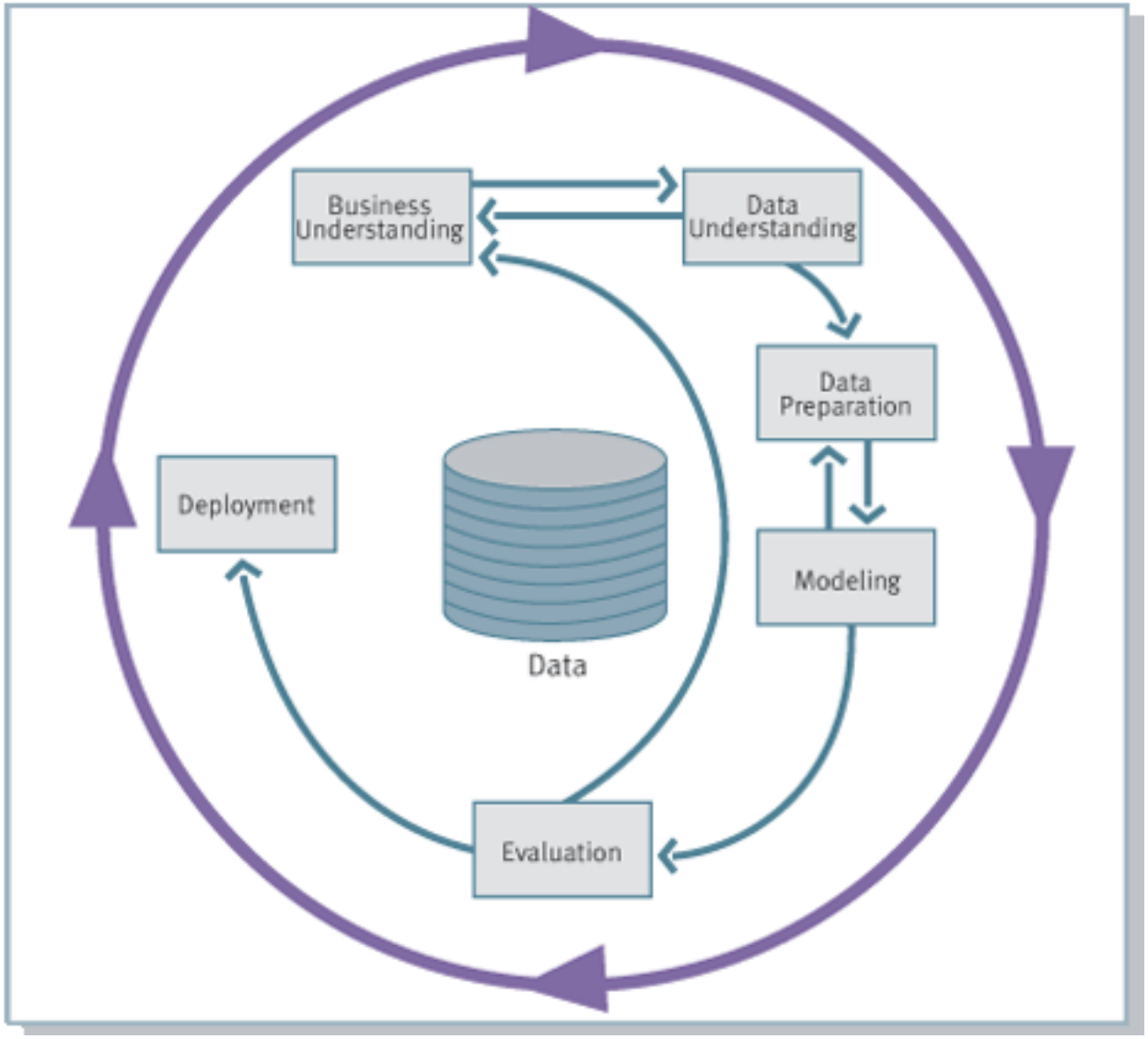
Postupy:
statistika,
Rozhodovací stromy,
Neuronové sítě,
...

Učení bez učitele

- Úloha: Nalezněte podobné příklady (instance) ve zpracovávaných datech, která nemají žádné značky.



Metodika CRISP-DM



Vyhodnocení - Evaluation

- zhodnocení dosažených výsledků modelování
- zhodnocení výsledků z pohledu zadání
- použití vizualizačních technik

- často návrat zpět na začátek celého procesu a stanovení nových cílů (úprava zadaní)

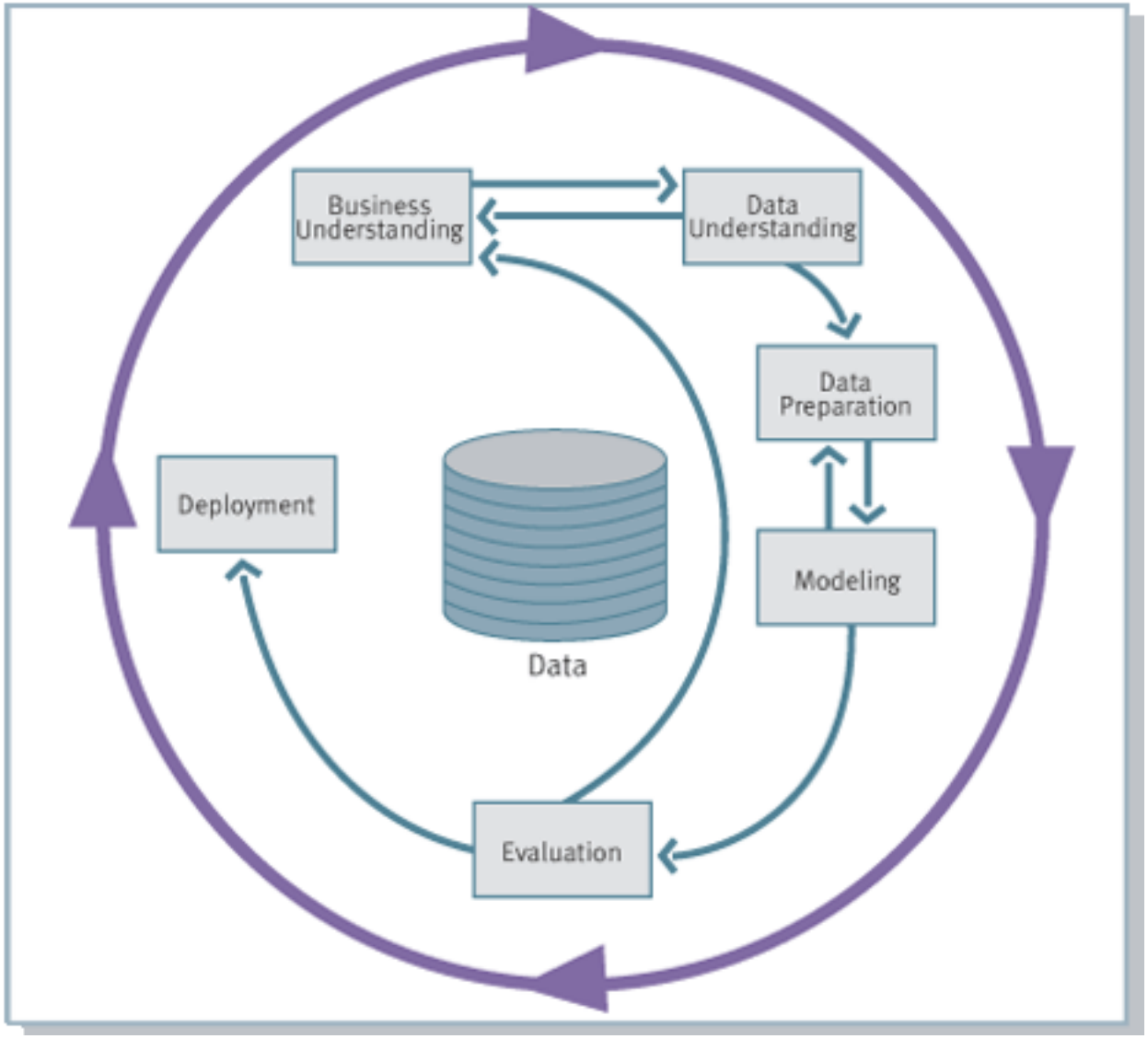
Testování modelů

- Q: Jak dobře funguje (klasifikuje) model, který jsme vytvořili?
- Chyba, s jakou model klasifikuje na trénovacích datech **není** dobrým odhadem pro chování modelu na dosud neznámých datech.
- Q: Proč?
- Nová data nebudou přesně stejná jako ta použitá pro učení! A navíc i náhodně vygenerovaný konečný soubor dat lze popsat nějakým modelem (třeba samotnou výchozí tabulkou).

Testování pro "ROZSÁHLÁ" data

- Máme-li hodně dat (tisíce instancí), které obsahují pro každou třídu dostatek vzorků (stovky instancí), pak stačí provést jednoduché testování:
  - Rozděl výchozí data náhodně do 2 množin: **trénovací** (asi 2/3 dat) a **testovací** (zbytek, tedy asi 1/3 dat)
  - Vytvoř klasifikační model nad *trénovací množinou* a proveď hodnocení (např.pomocí relativní chyby) na *testovací množině*
  - **Relativní chyba**: procentuální podíl chybných instancí vůči mohutnosti celé uvažované množiny instancí

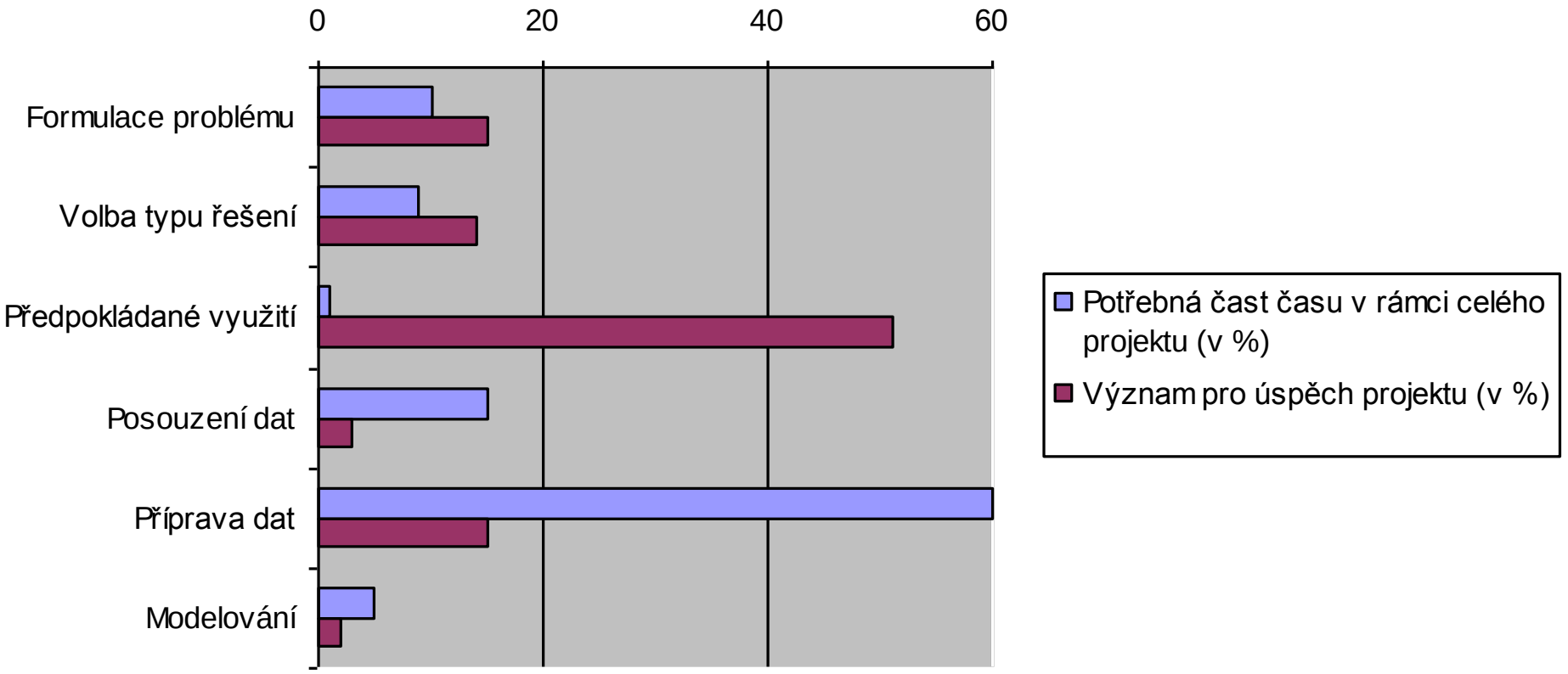
Metodika CRISP-DM



Použití - Deployment

- úprava získaných znalostí do srozumitelné formy pro zadavatele
- případně pomoc s implementací výsledků do praxe

Časové nároky procesu?



Shrnutí

- Co je dobývání znalostí ?
- Co je koncept, instance, pozorování, atribut, příznak?
- Co je klasifikace, shlukování?
- Co je metodika CRISP-DM a jaké jsou její jednotlivé fáze?
- Jaký je rozdíl mezi učením s učitelem a učením bez učitele?
- Proč při testování rozdělujeme data na trénovací a testovací množinu?