Dobývání a vizualizace znalostí

Olga Štěpánková, Tomáš Sieger, Jiří Anýž

https://cw.fel.cvut.cz/wiki/courses/a6m33dvz/start

Nature Inspired
Technologies Group

Osnova úvodní přednášky

- Úvod: data, objem, reprezentace a základní terminologie
- Vytěžování dat (Data Mining) & dobývání znalostí (Knowledge Discovery) a používané techniky (předpokládané dovednosti)
- Přehled základních přístupů k modelování dat
- Typické postupy DM – metodika CRISP-DM
- Průzkumová analýza dat a nejjednodušší vizualizační techniky
- Plán semestru a další zdroje informací

Kde se bere současná záplava dat?

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

Největší zdroje a oblasti:

- Obchodní transakce (obchodní řetězce, banky, pojišťovny ...)
- Telekomunikace, internet a elektronický obchod, sociální sítě
- Zdravotnictví a používané senzory
- 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

Planeta Země: méně než $3x10^{50}$ atomů.

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

Terminologie

- **Instance(pozorování):** - nezávislé pozorované jednotky
  - údaje o počasí jednoho konkrétního dne
- **Atributy(příznaky):** - jednotlivé údaje, které se pro instanci zaznamenávají (*teplota, tlak, množství srážek, ..*)
- **Reprezentace dat**
  - **Matice pozorování:** řádky jsou instance a sloupce příznaky
  - **Relační databáze**, grafy, ... (sociální sítě, ...)
- **Příznakový prostor:** prostor, jehož dimenze jsou definovány jednotlivými příznaky
  - pozorování jsou body v příznakovém prostoru
- **Koncept:** oblast zájmu – podmožina příznakového prostoru (která má nějaký výnam), např. „léto"
- **Model:** popis konceptu nebo alg.odpovídání na dotazy, …

Definice dobývání znalostí

Dobývání znalostí (Data Mining) je *netriviální* proces zpracování dat, který vede k identifikaci či vyhledání takových **srozumitelných vzorů** v příslušných datech, které jsou

- validní,
- nové,
- a potenciálně užitečné (použitelné).

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

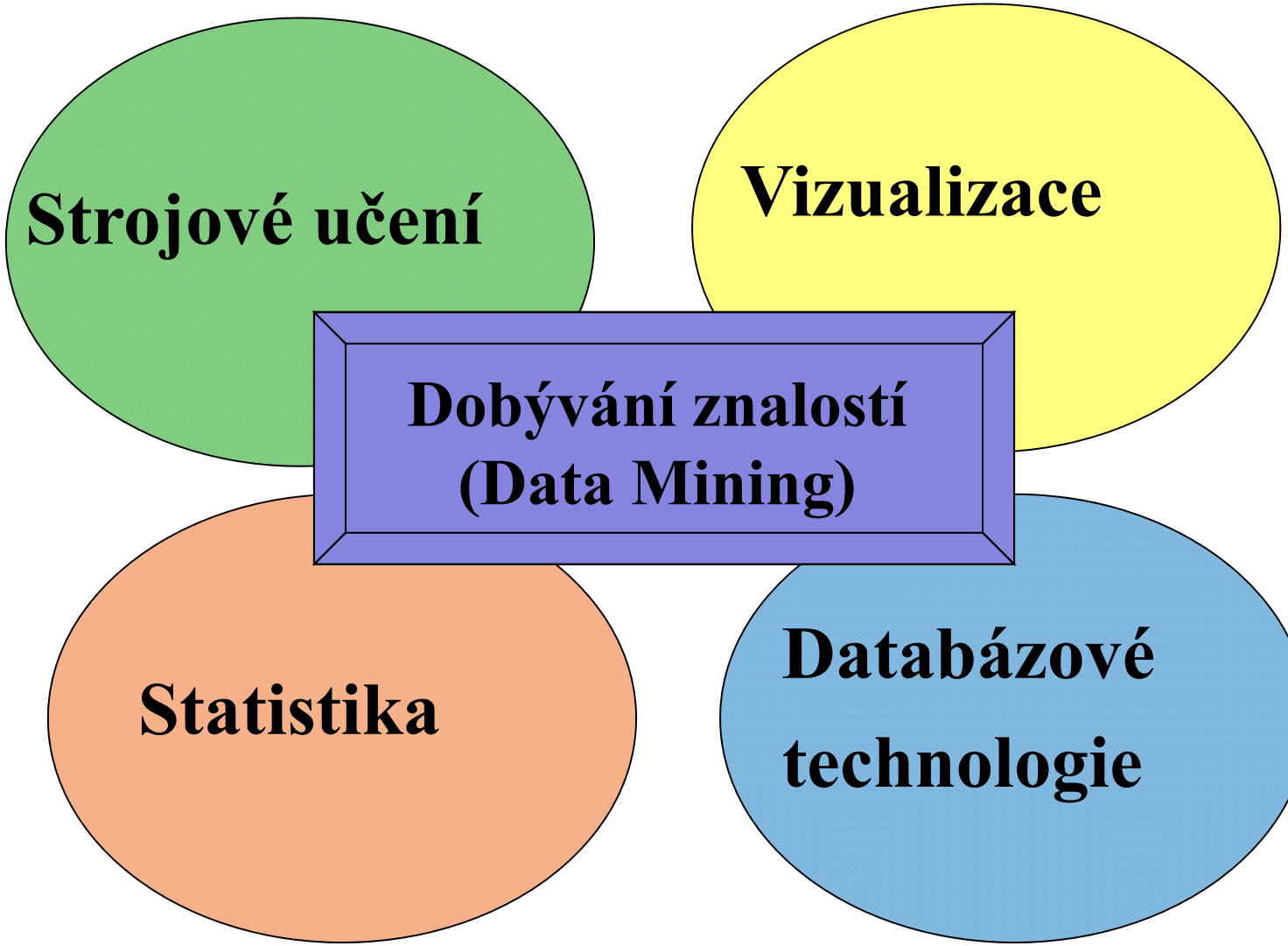
Dobývání znalostí z dat

- **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**

Jinak: ***Aplikace strojového učení v praxi***

Souvislosti



Dobývání znalostí z dat

Příklady aplikací:

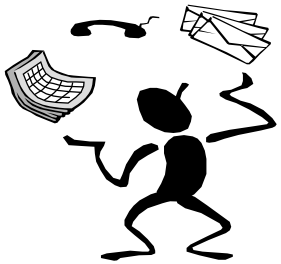
- průmysl (diagnostika poruch, predikce spotřeby …)
- obchod (marketing, bankovnictví)
- věda (charakterizace karcinogenních látek)
- medicína (mapování lidského genomu)
- analýza sociálních sítí (LinkedIn, ...)

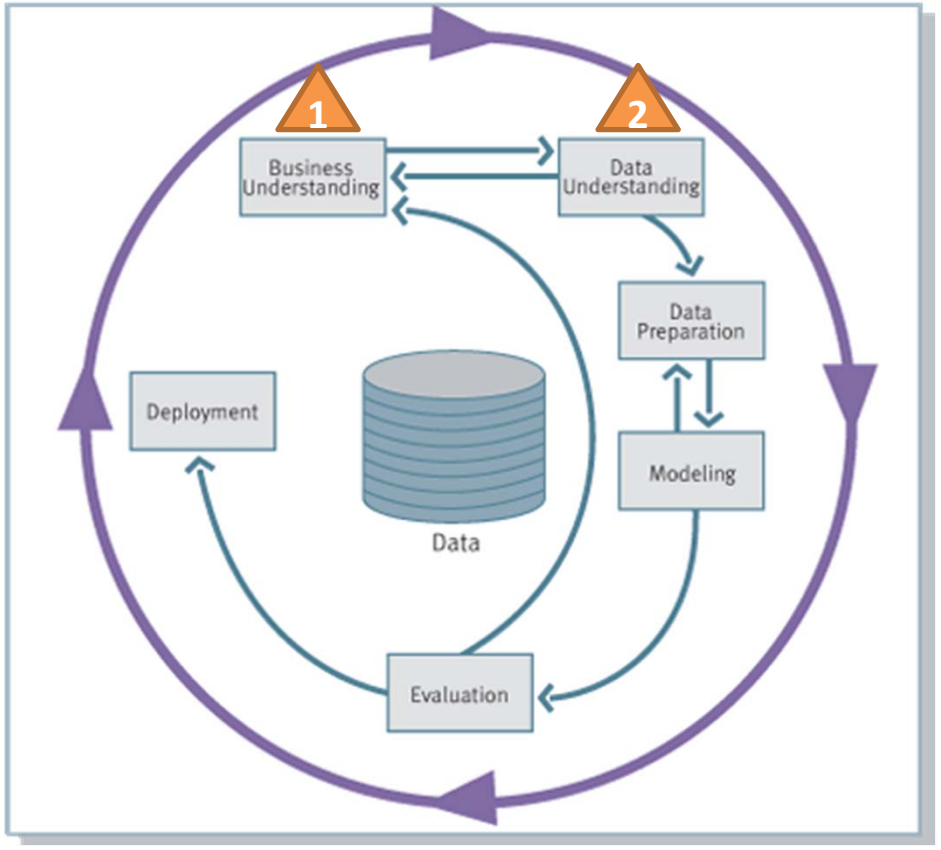
Typické úlohy a způsoby jejich řešení (modely):

- Hledání instancí, které jsou si navzájem „podobné"- **shlukování**
- Hledání typických vzorů chování – **asociační pravidla**
- Předpovídání hodnoty některého příznaku pro novou (dosud nezaznamenanou) instanci – **regrese**, **rozhodovací stromy**, .., **NN**

Metodika CRISP-DM

(www.crisp-dm.org)





Nature Inspired Technologies Group

Zadaní – 1. Business Understanding

- Pochopení cílů úlohy a problémů, které potřebuje majitel dat/zákazník (**zadavatel**) řešit
  - náklady
  - hodnotí se potenciální přínos vzniklého řešení
  - stanovení předběžného plánu
- Výchozí data a forma jejich předání
  - anonymizace dat
  - formát dat
- Způsob komunikace mezi zadavatelem a řešitelem (forma, frekvence, ..)

Osobní údaje

- Identifikační údaje
  - Jméno a přijmení
  - Adresa
  - Datum narození, rodné číslo
  - Identifikační číslo např. v nemocničním informačním systému
- Citlivé osobní údaje
  - Národnostní, rasový nebo etnický původ
  - Politické postoje
  - Náboženství
  - Zdravotní stav
  - Biometrické údaje

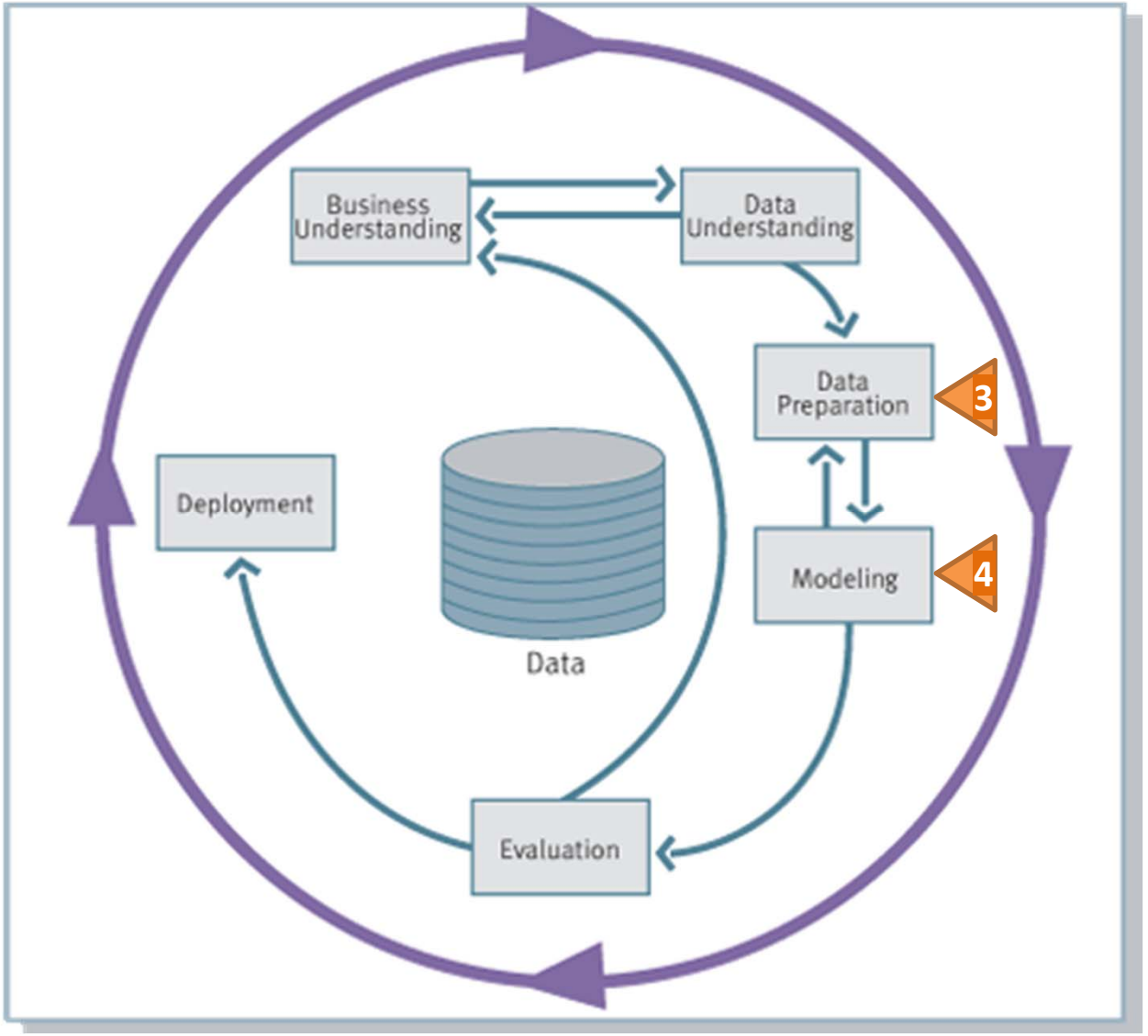
Problémy reálných dat?

- 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 (probably approximately correct learning = PAC learning)
- Data mají formu **složitého relačního schématu**, nikoliv jediné tabulky předpokládané atributovými metodami strojového učení
- POZOR na přísně důvěrná data (nejen osobní údaje)!!!

Analýza dat – 2. Data Understanding

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

Metodika CRISP-DM



Nature Inspired Technologies Group

Příprava dat – 3. Data Preparation

- příprava dat pro modelování
  - selekce atributů – výběr relevantních atributů
  - čištění dat
  - doplnění dat (např. z veřejně přístupných zdrojů)
  - 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 – 4. Modeling

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

- UCI Machine Learning Repository = data pro testování modelů
  - http://archive.ics.uci.edu/ml/

Typy úloh

Informované metody lze použít, pokud

- je jasné, jaký koncept nás zajímá
- a výchozí data obsahují odpovídající atribut.

Pak hovoříme o **učení s učitelem**. Metody jsou např.

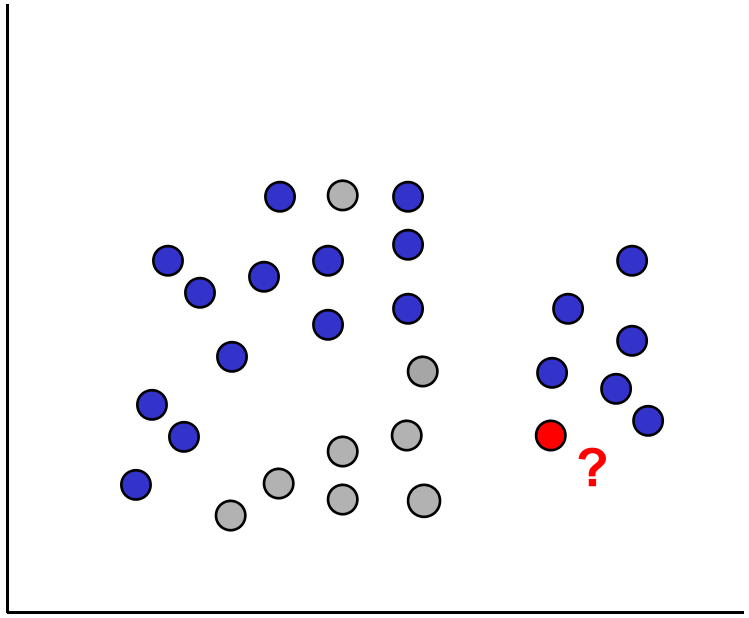
- **Klasifikace**: přiřazení třídy novému objektu (instanci)
- **Predikce**: předpověď chování objektu v čase

Neinformované metody – učení bez učitele

- **Asociace**: hledání vazeb mezi objekty
- **Shlukování**: seskupování podobných objektů
- Hledání **nejpodobnějšího objektu**

Učení s učitelem

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

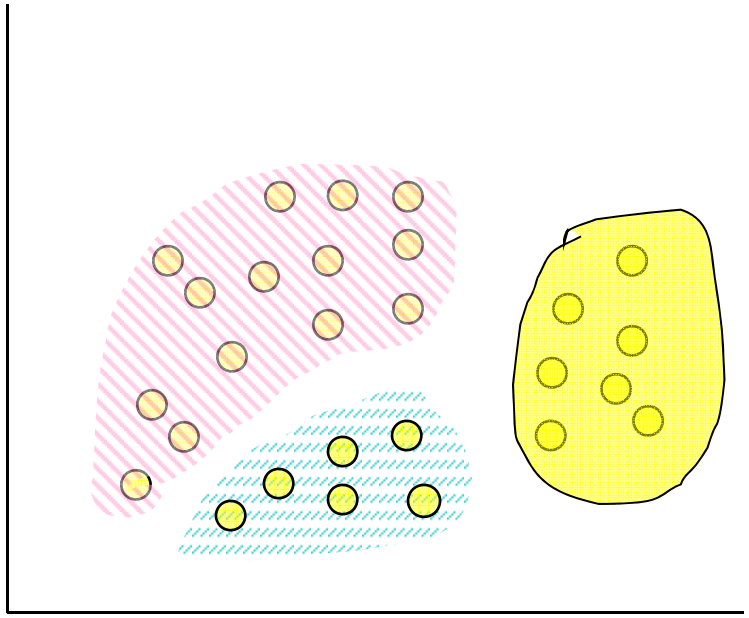


Třída 1
Třída 2

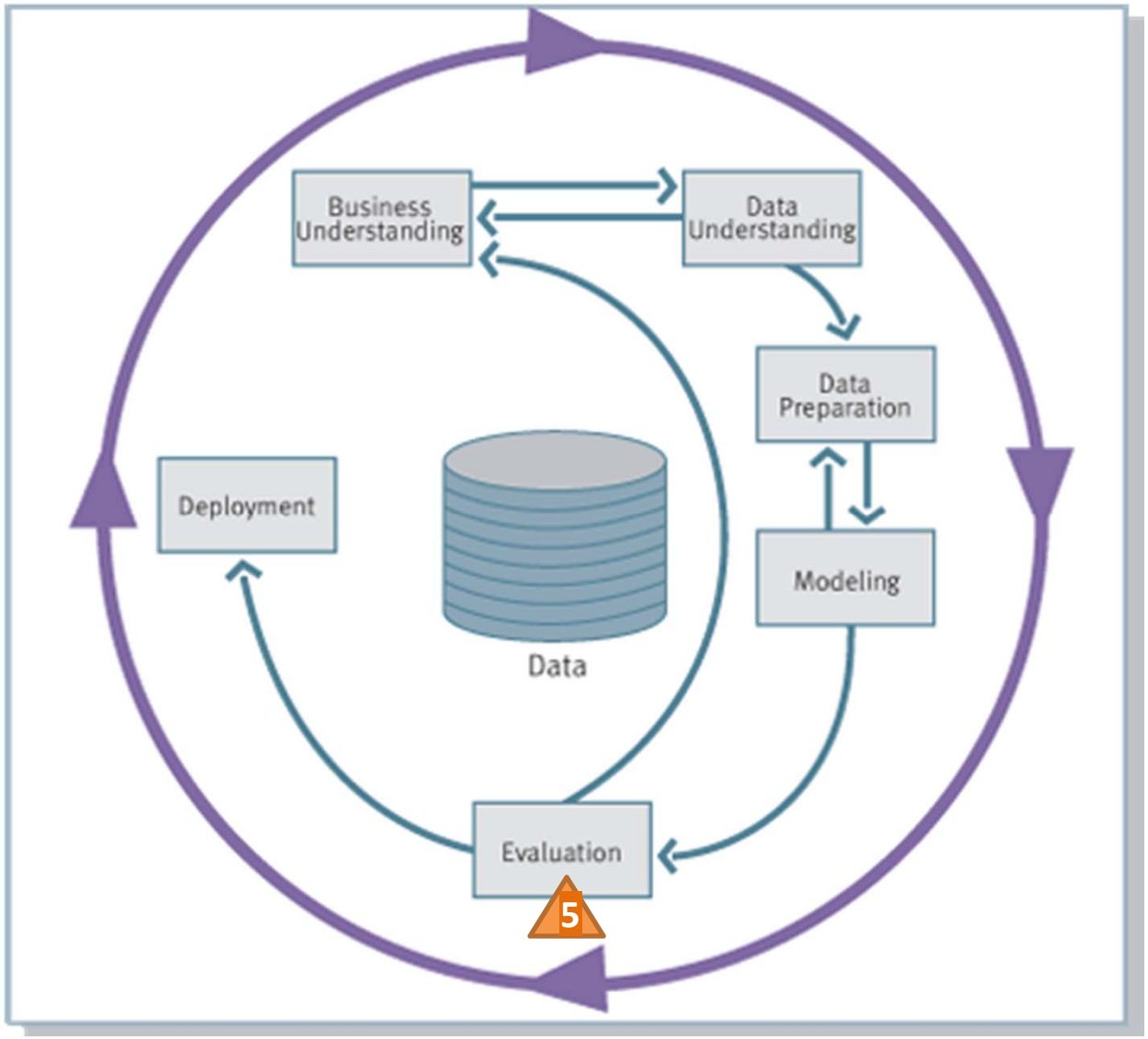
Postupy:
Nejbližší soused
Statistika,
Rozhodovací stromy,
Neuronové sítě,
...

Učení bez učitele

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



Metodika CRISP-DM



Nature Inspired Technologies Group

Vyhodnocení – 5. Evaluation

- Zhodnocení dosažených výsledků modelování z hlediska statistických ukazatelů pomocí testování
- Zhodnocení výsledků z pohledu zadání
- Pro komunikace se zadavatelem se často používají vizualizační techniky

První výsledky nebývají zcela uspokojivé – spíš naznačují směr dalšího hledání a upozorňují na data, která by se hodilo doplnit !

- Často je nutné se vrátit zpět na začátek celého procesu a stanovit nové cíle (upřesnění nebo úprava zadaní)

Testování modelů

- Q: Jak dobře funguje (klasifikuje) model, který jsme vytvořili?

- POZOR! Chyba, s jakou model klasifikuje na trénovacích datech **není** dobrým odhadem pro chování modelu na dosud neznámých datech

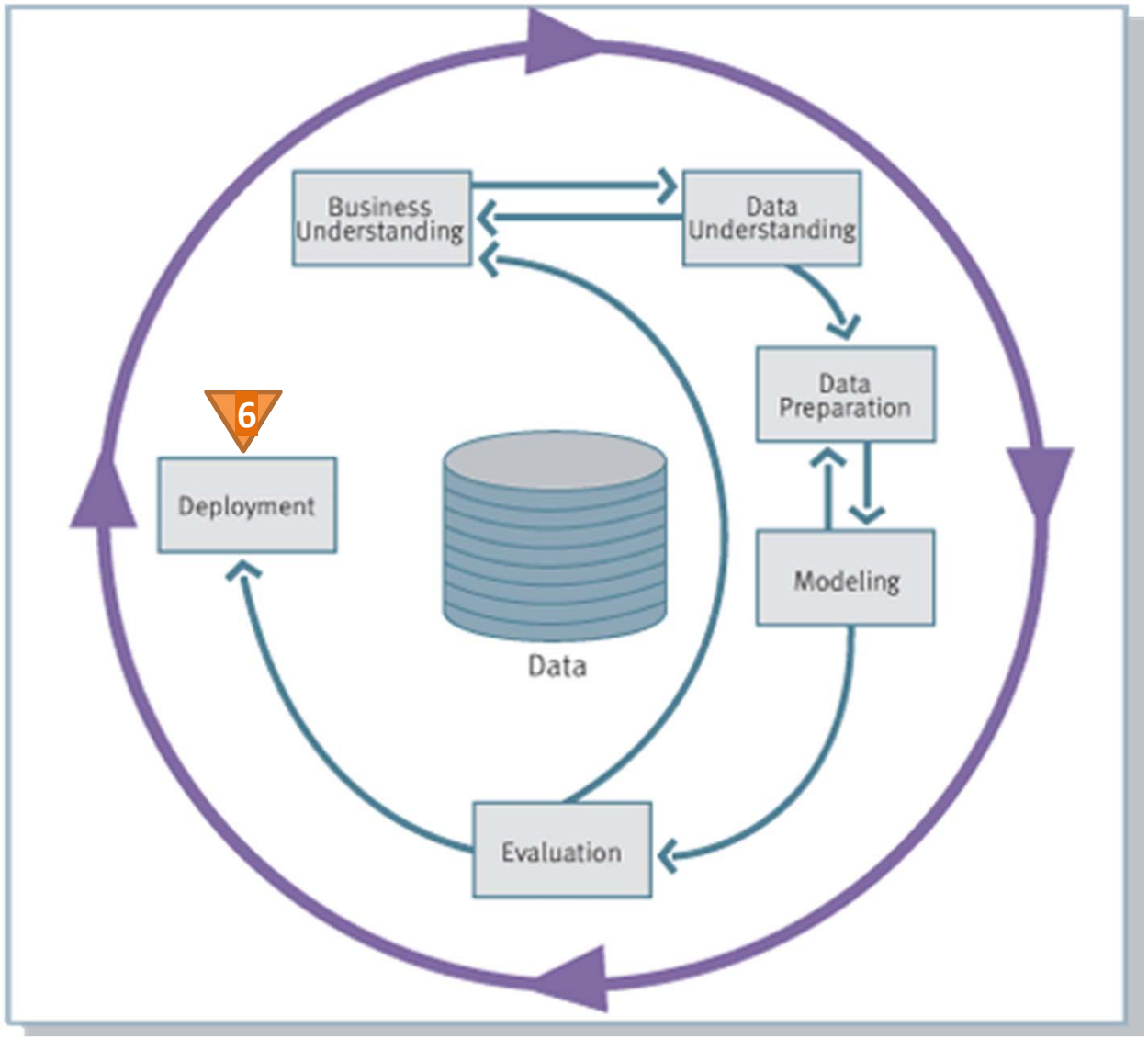
- Q: Proč?

- 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

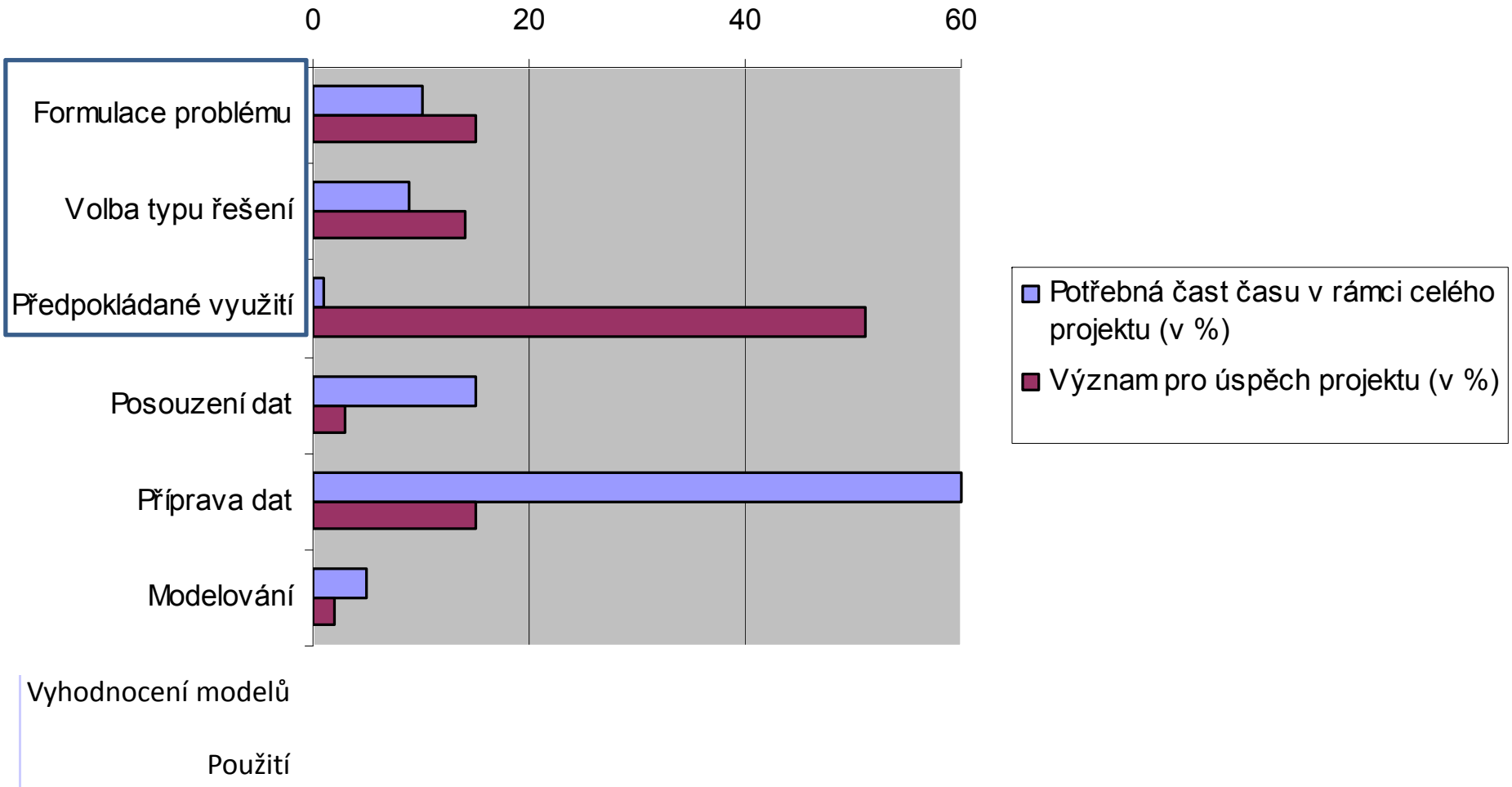


Nature Inspired Technologies Group

Použití – 6. Deployment

- Úprava získaných znalostí do srozumitelné formy, kterou zadavatel může prakticky využít
- Mnohdy následuje i pomoc s implementací výsledků do praxe

Časové nároky procesu DM?



Nature Inspired Technologies Group

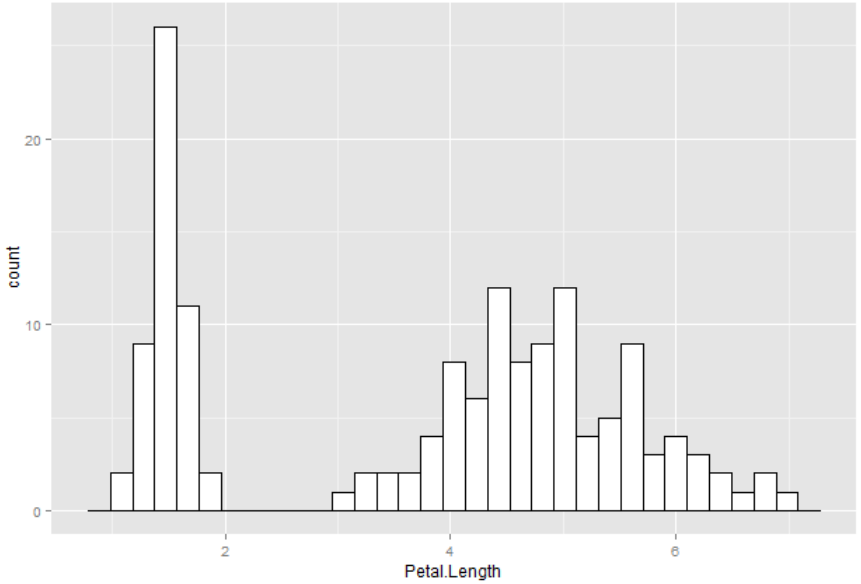
Shrnutí

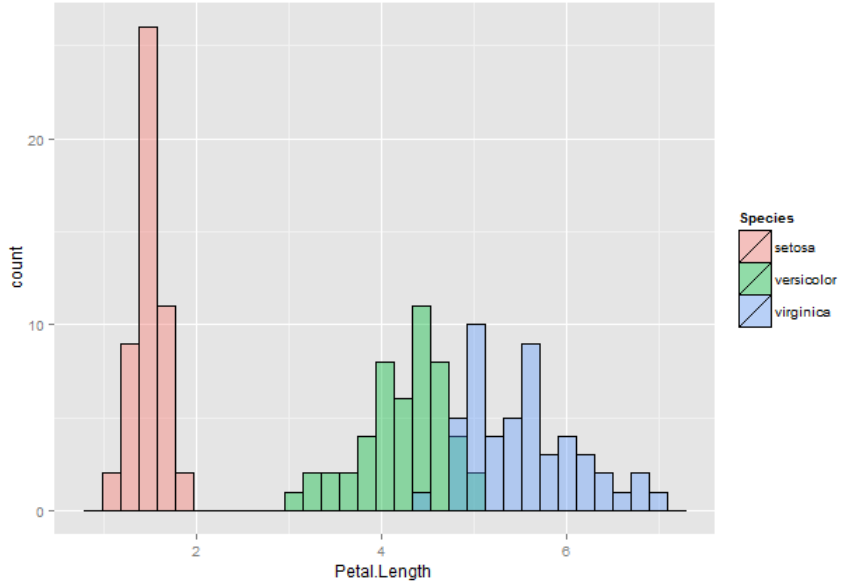
- Co je dobývání znalostí ?
- Co je koncept, pozorování(instance), příznak(atribut), příznakový prostor, matice pozorování?
- Co je metodika CRISP-DM a jaké jsou její jednotlivé fáze?
- Jaký je rozdíl mezi učením s učitelem a učením bez učitele?
- Proč při testování rozdělujeme data na trénovací a testovací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

Průzkumná analýza dat – Data Understanding

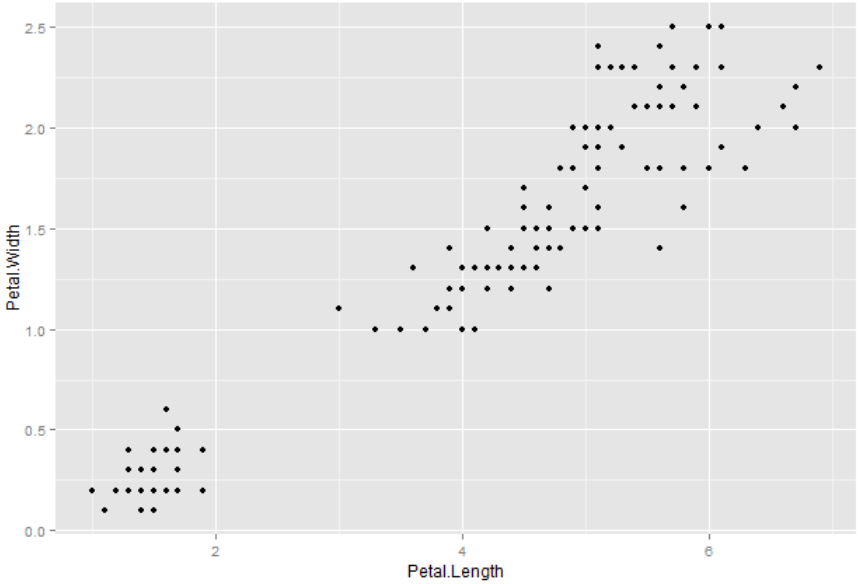
- získání základní představy o datech
  - Reprezentace dat (datová matice nebo relační DB?)
  - Formát dat: počty instancí, atributů
- chybějící hodnoty
- deskriptivní charakteristiky dat podle typu dat
  - četnosti hodnot (histogramy)
  - minima, maxima, průměry
  - odlehlé hodnoty
- Vizualizace dat

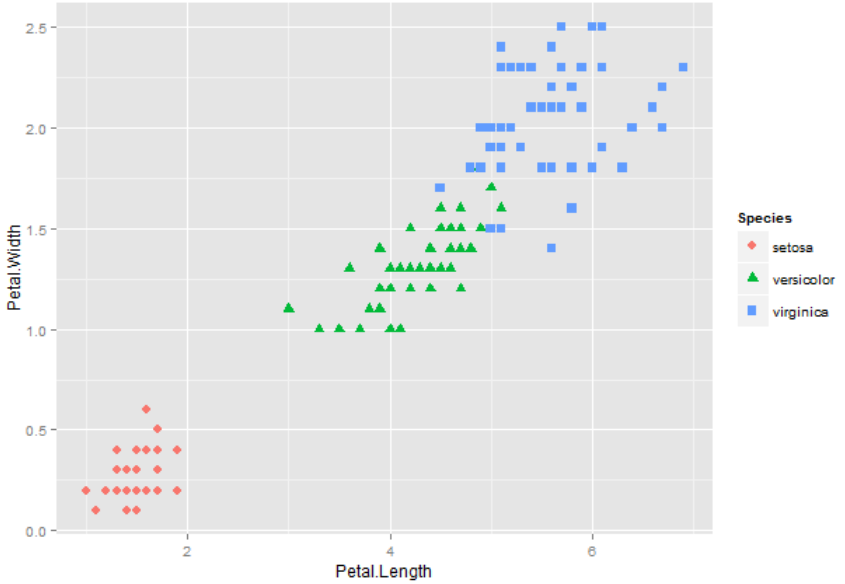
Histogram


count
20
10
0
2
4
6
Petal.Length


count
20
10
0
2
4
6
Petal.Length
Species
setosa
versicolor
virginica

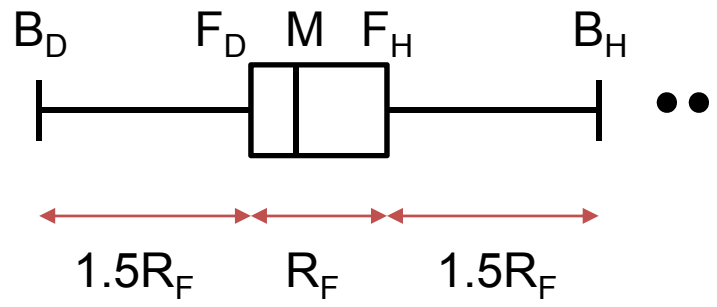
Scater plot = XY graf





Nature Inspired Technologies Group

Box graf



M **medián**

F_D dolní kvartil

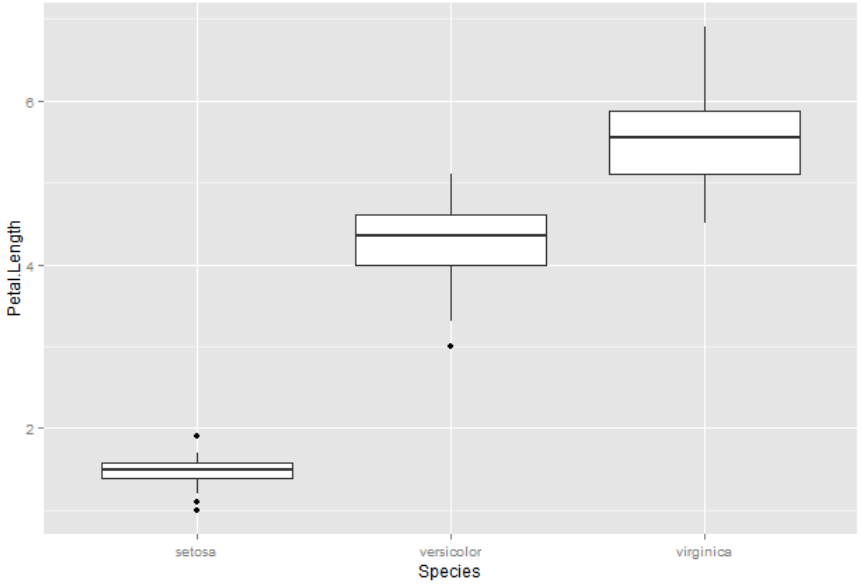
F_H horní kvartil

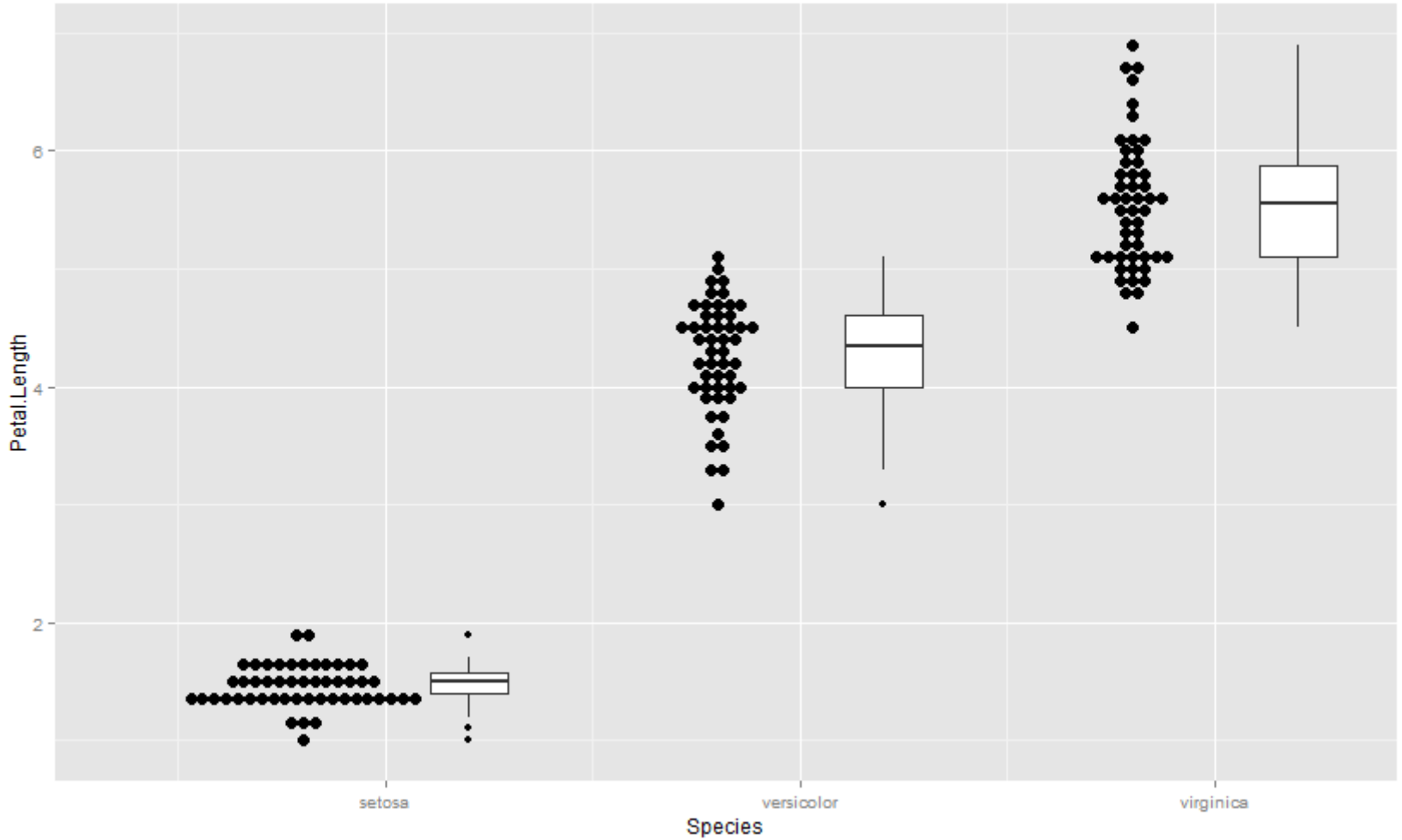
$R_F = F_H - F_D$

mezikvartilové rozpětí

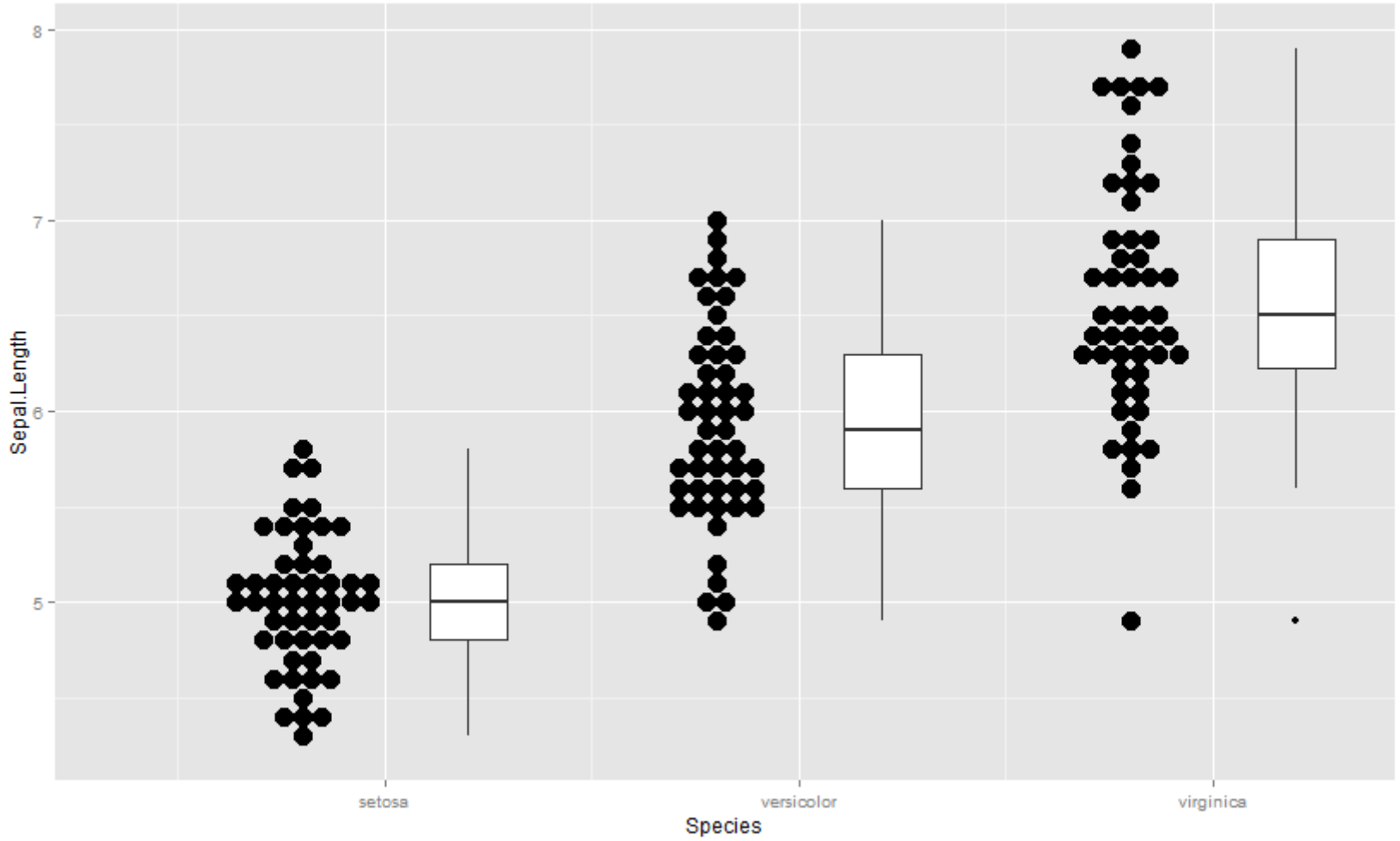
$B_D = F_D - 1{,}5 * R_F$ („vousy“)

F_D (resp. ***F_H***) bod, ve kterém distribuční funkce náhodné proměnné prochází hodnotou **0,25** (resp. **0,75**)





Nature Inspired
Technologies Group


8
7
6
5
Sepal.Length
setosa
versicolor
virginica
Species

Shrnutí

- Co je to kvartil?
- Jaký je rozdíl mezi výběrovým průměrem a mediánem?
- Co je to histogram?
- Jak se zobrazuje odlehlá hodnota v box grafu?

Osnova předmětu

1. Dobývání znalostí - popis a metodika procesu a objasnění základních pojmů
2. Nástroje pro modelování neklasifikovaných dat a jejich využití
3. Nástroje pro modelování klasifikovaných dat a jejich využití
4. Vyhodnocení a využití modelů
5. Porozumnění datům a jejich příprava, vizualizace dat
6. Selekce a extrakce příznaků
7. Konstrukce asociačních pravidel (s využitím Apriori algoritmu).
8. Tvorba modelu kombinací více základních modelů
9. Neuronové sítě, volba parametrů a jejich aplikace.
10. Práce s časovými řadami.
11. Zpracování přirozeného jazyka jako vstupu
12. “Text mining” a podpora kreativity
13. Prezentace semestrálních prací
14. Zajímavé aplikace

Prerekvizity: Přehled základních pojmů ze statistiky

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

Charu C. Aggarwal: Data Mining. The Textbook. Springer 2015, http://link.springer.com/book/10.1007/978-3-319-14142-8

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