



Dobývání a vizualizace znalostí

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

Olga Štěpánková, Tomáš Sieger,
Lenka Vysloužilová, Milan Němý



Osнова úvodní přednášky



- ❖ Úvod: data, objem, reprezentace a základní terminologie
- ❖ Vytěžování dat (Data Mining) & dobývání znalostí (Knowledge Discovery), 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 ...

Prague Health Heekaton, 26.-28.10.2018, <https://www.ceehacks.com/>, Challenges



- 1. Hack the Veteran.** Improve the life of a real ex-soldier with service-related injuries.
- 2. Big Data & AI.** Making patients healthier through data analysis and Artificial Intelligence.
- 3. Smart Hospital.** Novel solutions for hospitals & practices in the domains of information systems, e.g. orientation of patient in a hospital, every day medical professional's work.
- 4. Diseases of Affluence including Social Media Addiction.** Cure Diabetes, Hypertension, Cancer & Addictions through data, software and hardware.
- 5. Abnormal Involuntary Movement Smartphone App = AIMSA.** Design an app for detecting early signs of anomalous involuntary rhythmic facial and tongue movements ...
- 6. Adherence and Compliance.** How to Make Patients Follow the Prescription & Therapy Better.
- 7. Smart Rehabilitation.** Digitize the stretching and flexing of muscles and make rehab fun for kids.
- 8. Predictive Health Diagnostics.** Design a neural network able to understand patient query and match it with the right answer using Q&A database from online medical consultation platform.
- 9. Communicating with the Deaf.** How could the deaf communicate in life emergency through sign language using digital communication channels combined with the power of AI?
- 10. Understanding brain functions and dysfunctions.** Predictive neuroscience, functional brain mapping, brain image analysis, classifying and simulating diseases of the brain, motor control and movement disorders.

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 3×10^{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**
 - ❖ **Maticе 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 – podmnožina příznakového prostoru (která má nějaký význam), 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

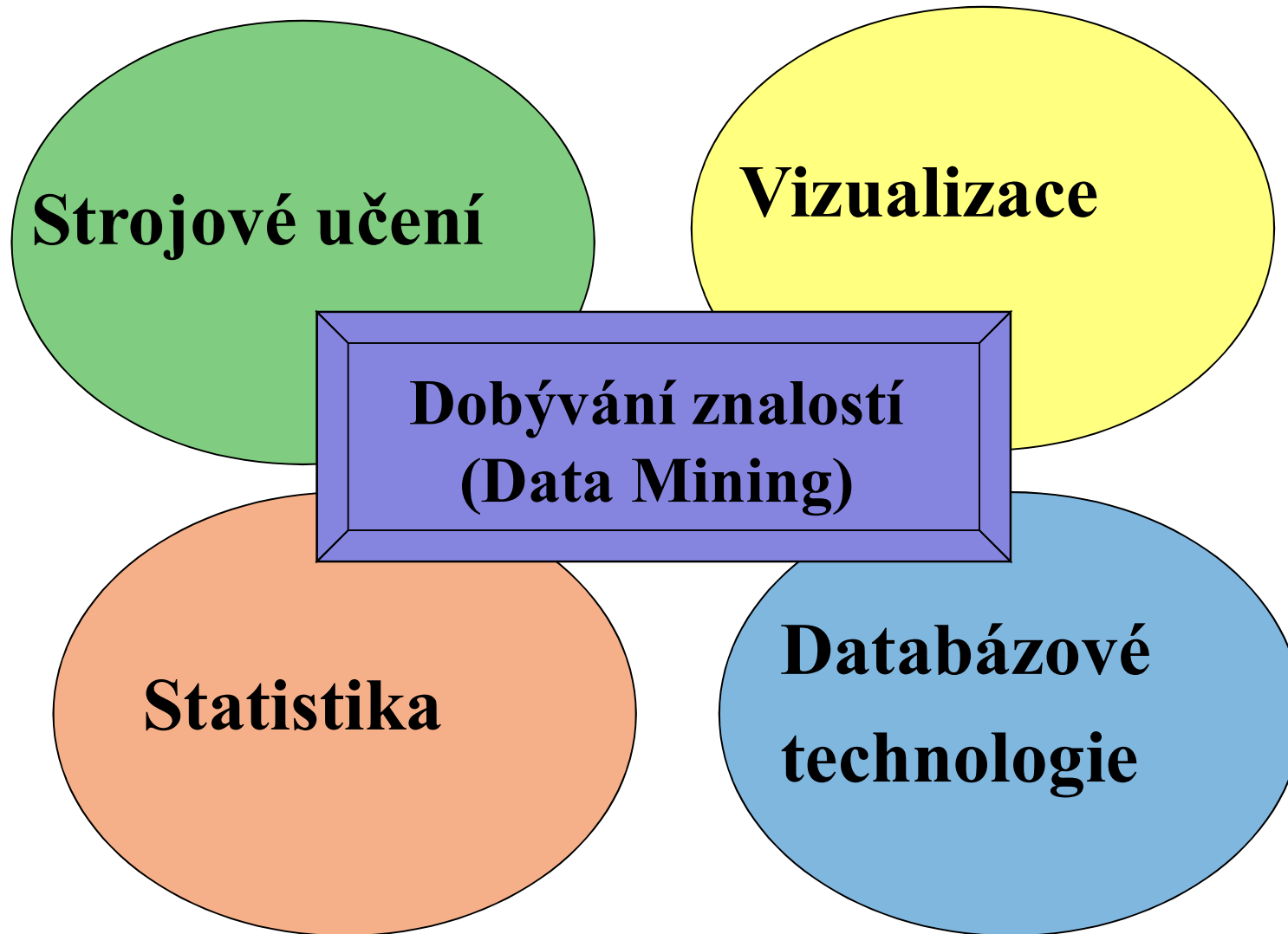


❖ **Cíl:** částečná automatizace procesu získání **zajímavých vzorů chování z reálných dat:** tvorba jejich modelů - pomocí nástrojů strojového učení, statistiky, databázových technologií,...

❖ 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, bankovníctví)
- ◆ věda (charakterizace karcinogenních látek)
- ◆ medicína (mapování lidského genomu)
- ◆ analýza sociálních sítí (LinkedIn, ...)

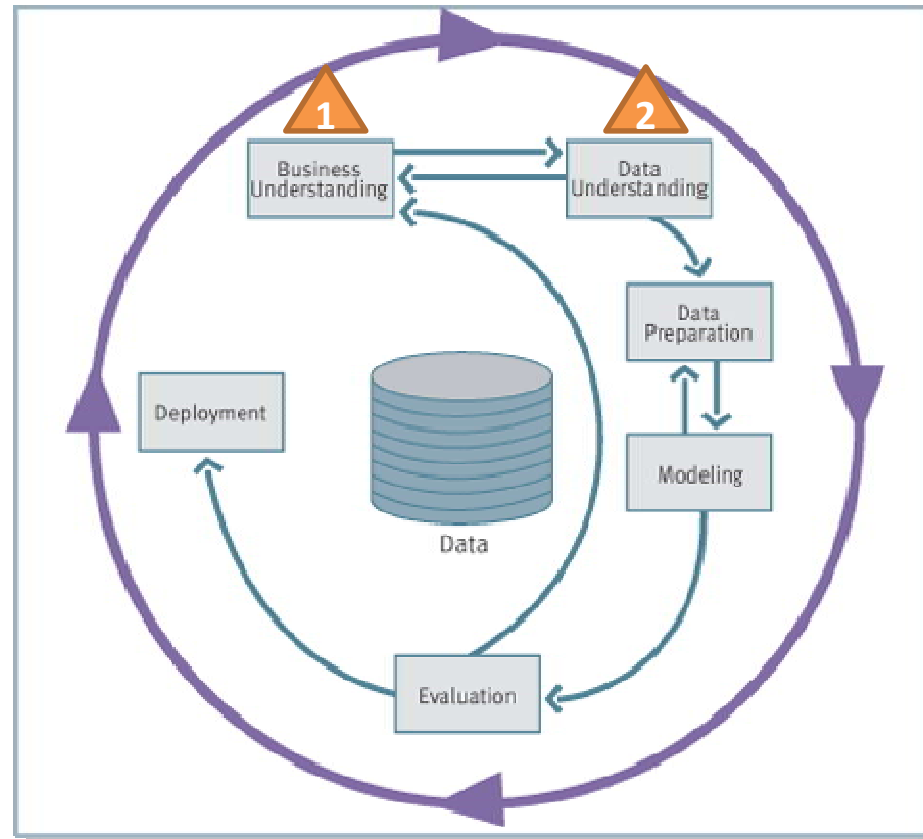
❖ Typické úlohy a způsoby jejich řešení (modely):

- ◆ 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)



Zadání – 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 příjmení
- ❖ 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

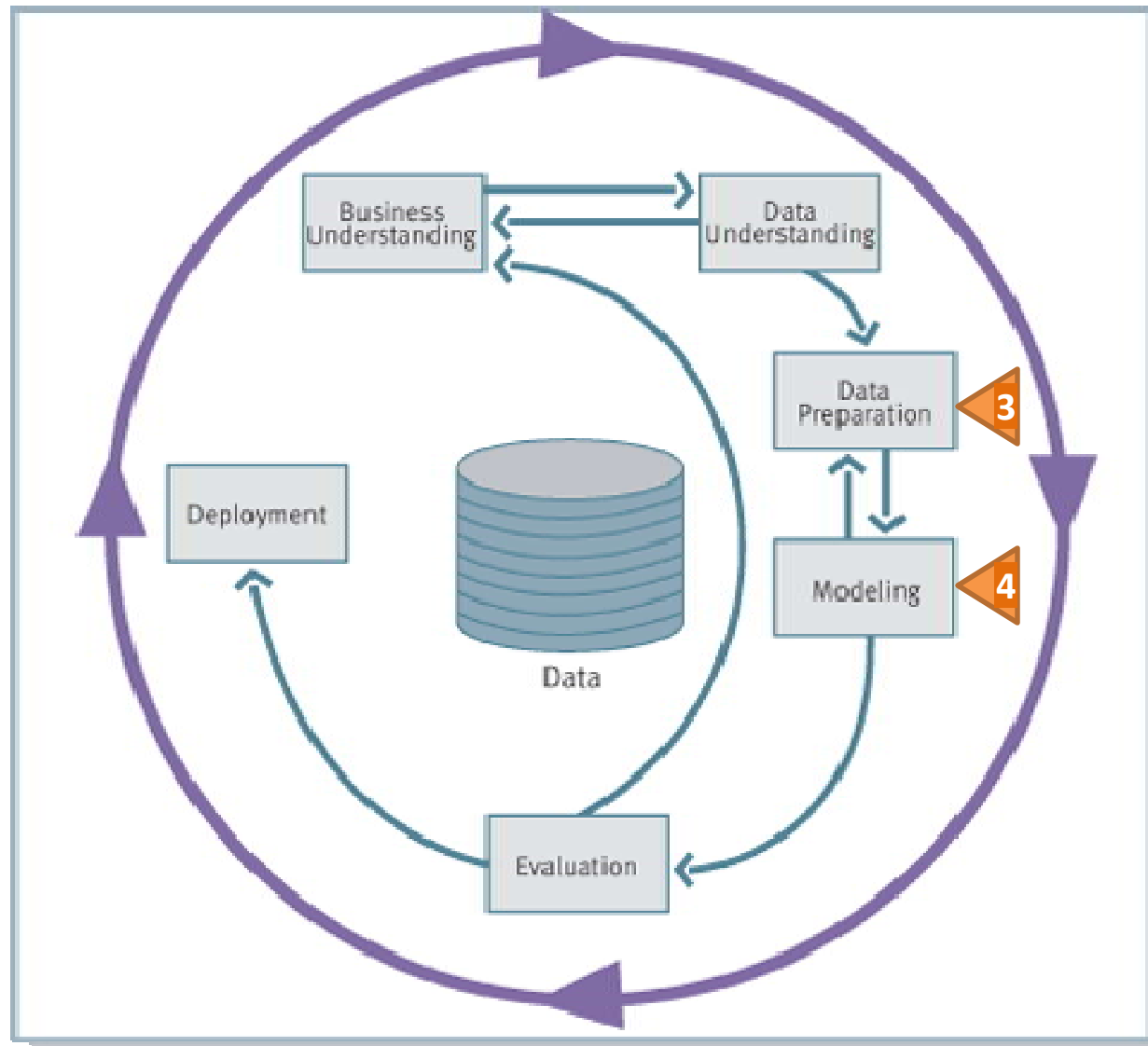
❖ **POZOR** na přísně důvěrná data (nejen osobní údaje)!!!

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



— 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

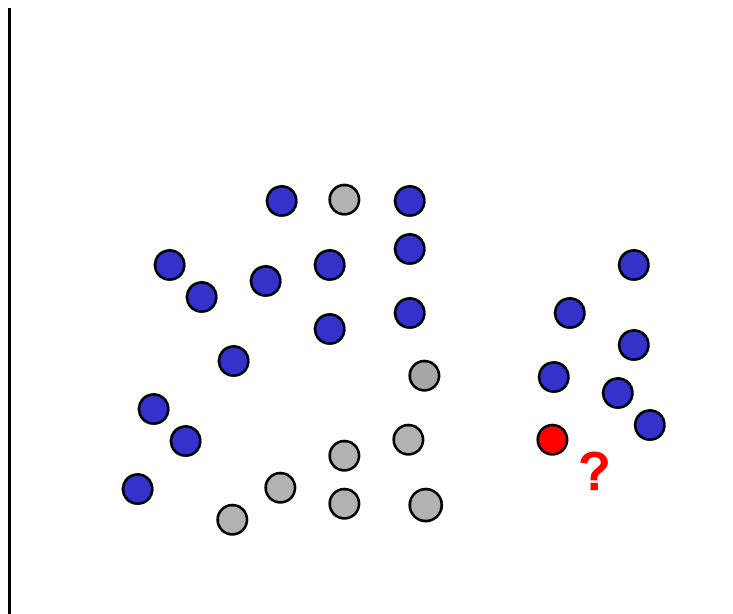
Neinformované metody – učení bez učitele

- ❖ **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



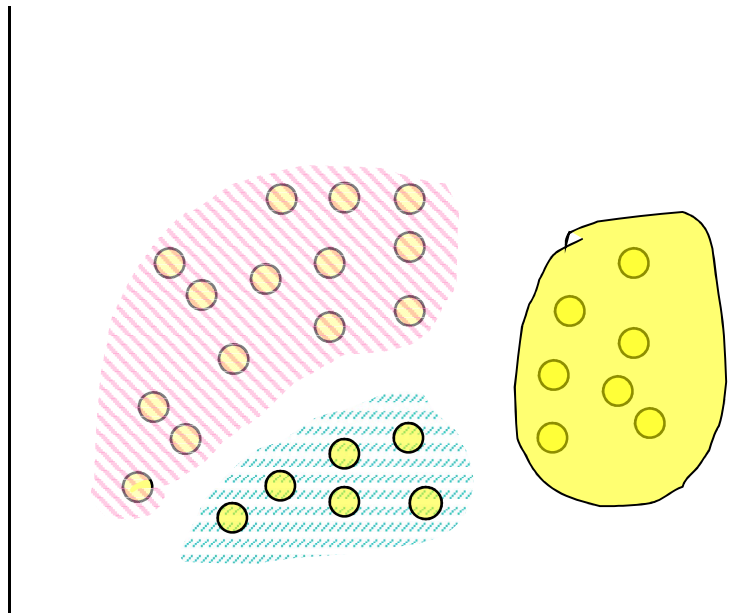
Třída 1
Třída 2

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

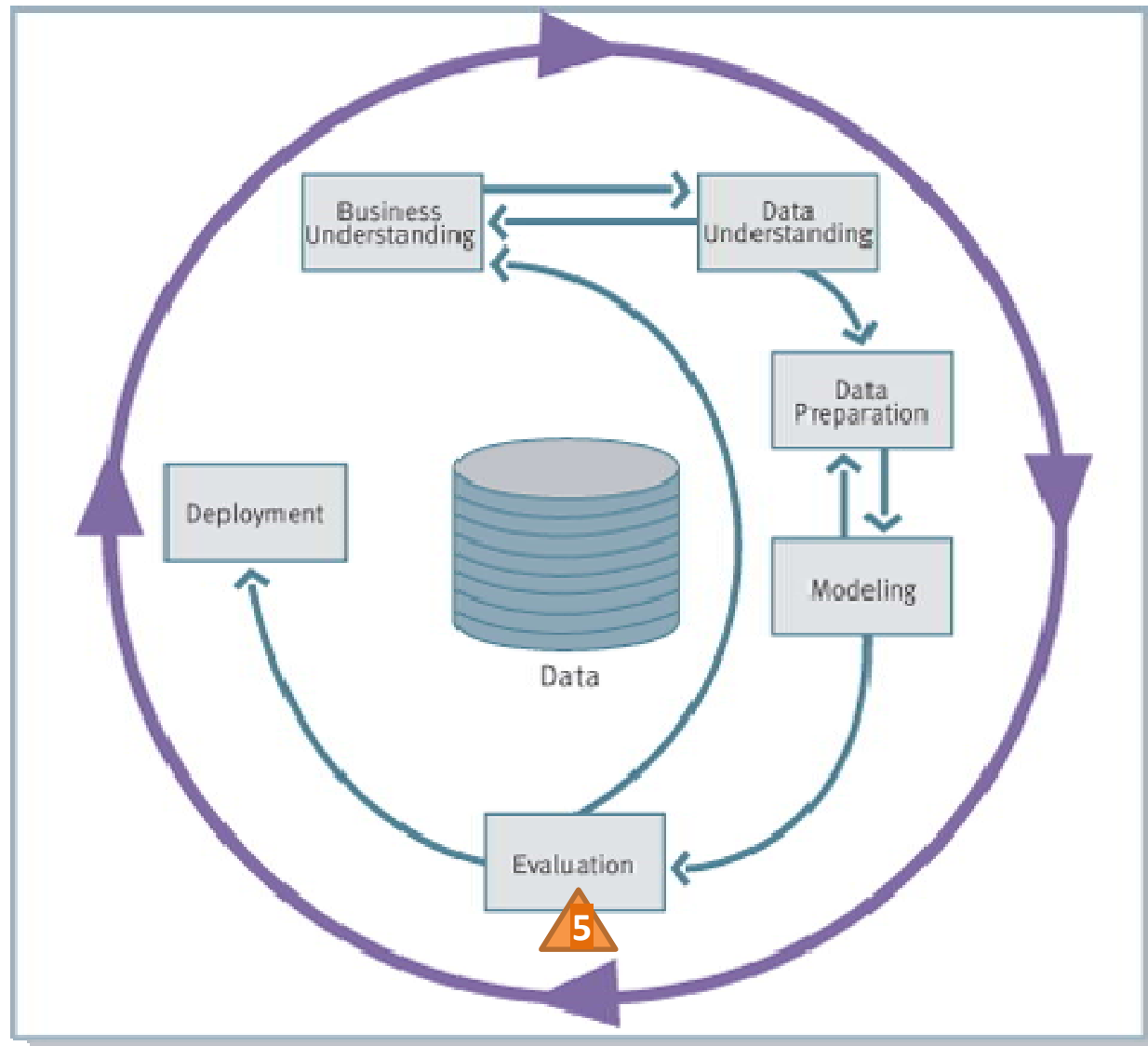
Učení bez učitele



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



Metodika CRISP-DM



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 zadání)

Testování modelů



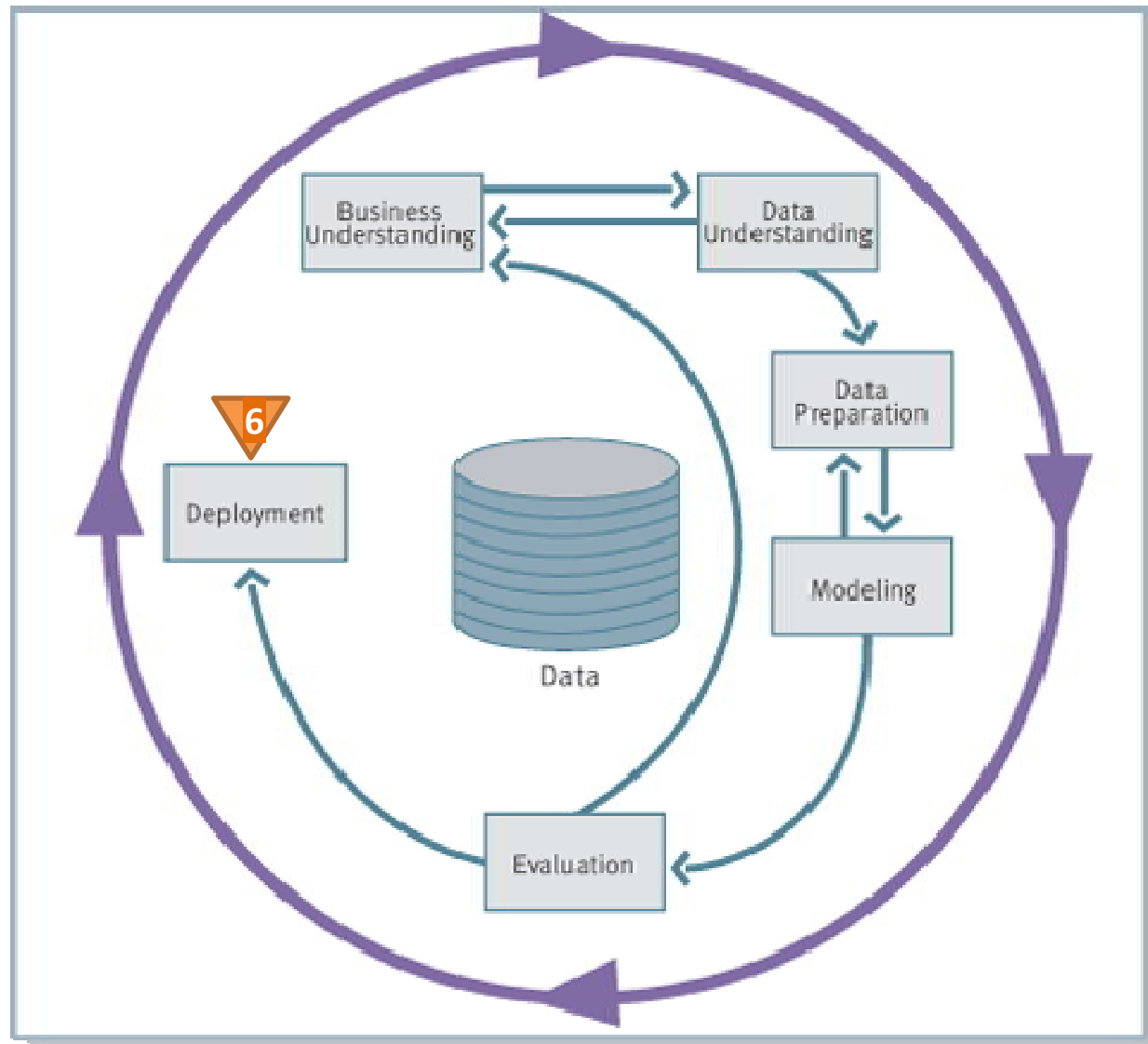
- ❖ Q: Jak dobře funguje (klasifikuje) model, který jsme vytvořili?
- ❖ POZOR! Chyba, s jakou model klasifikuje na trénovacích datech **není** dobrým odhadem pro chování modelu na dosud neznámých datech
- ❖ Q: Proč?
- ❖ Nová data nebudou přesně stejná jako ta použitá pro učení! A navíc i náhodně vygenerovaný 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:
 - ❖ Rozděli 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 proved' hodnocení (např. pomocí relativní chyby) na *testovací množině*
 - ❖ **Relativní chyba**: procentuální podíl chybných instancí vůči mohutnosti celé uvažované množiny instancí

Metodika CRISP-DM

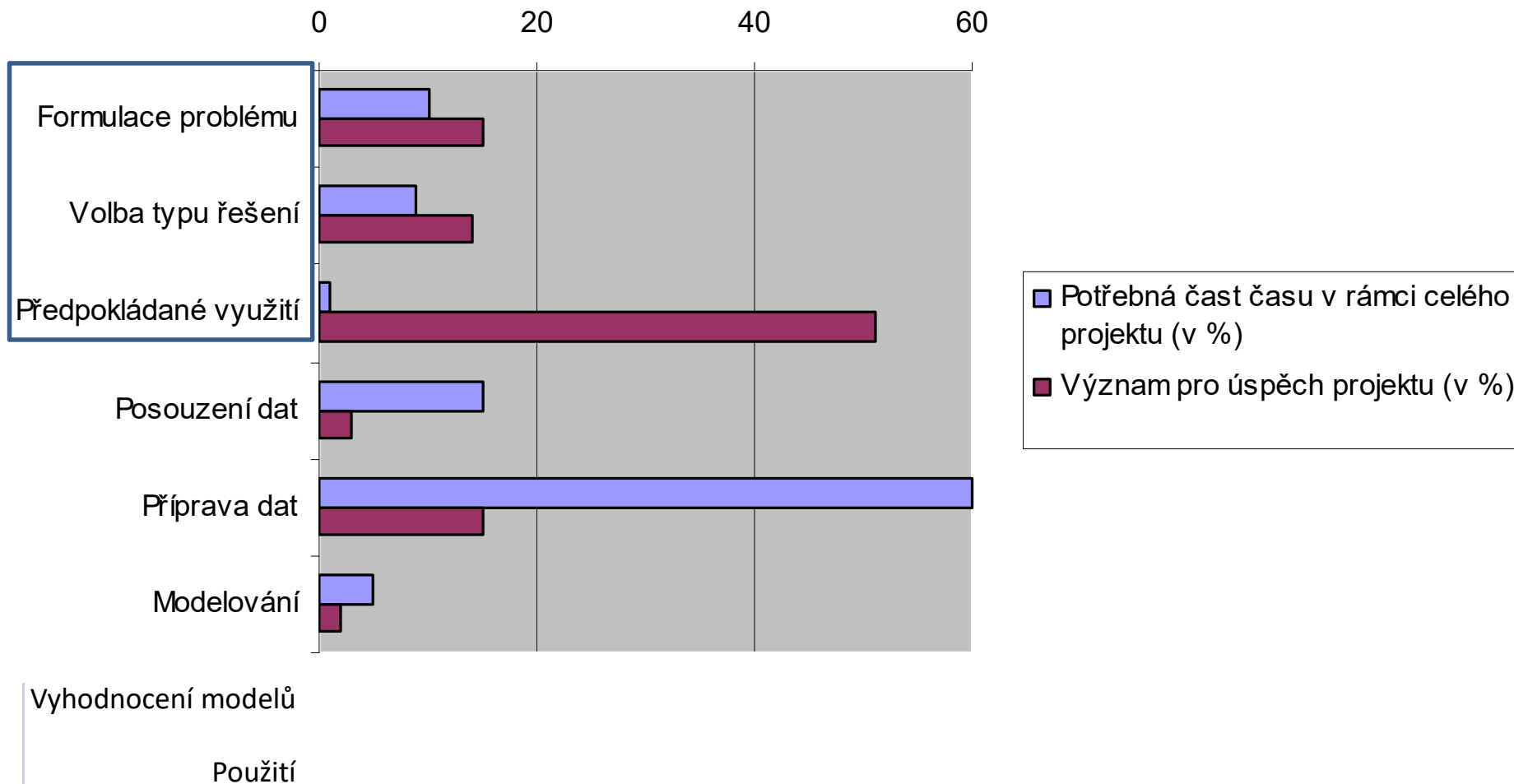


— Použití – 6. Deployment



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

Časové nároky procesu DM?



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?

Osнова předmětu



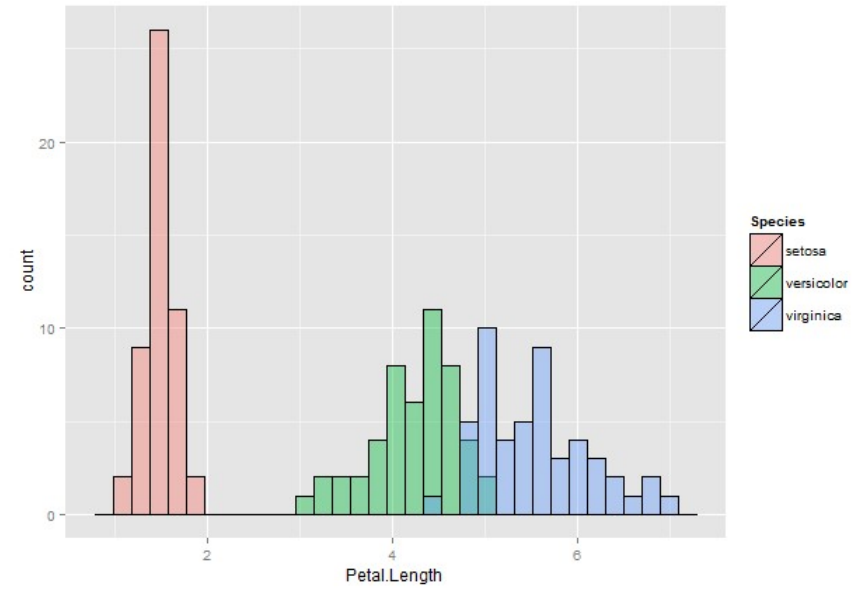
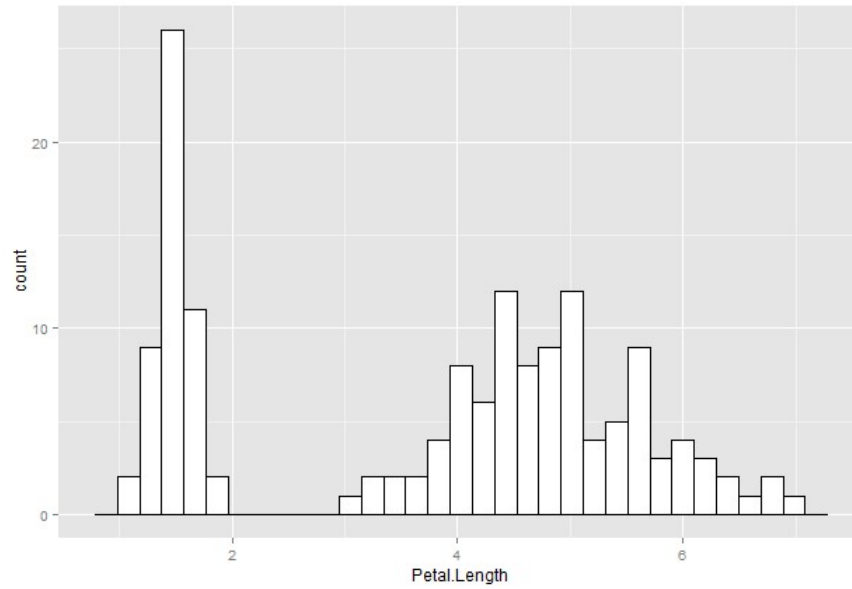
1. Dobývání znalostí - popis a metodika procesu a objasnění základních pojmů
2. Nástroje pro modelování neklasifikovaných dat a jejich využití
3. Nástroje pro modelování klasifikovaných dat a jejich využití
4. Vyhodnocení a využití modelů
5. Porozumnění datům a jejich příprava, vizualizace dat
6. Selektce 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

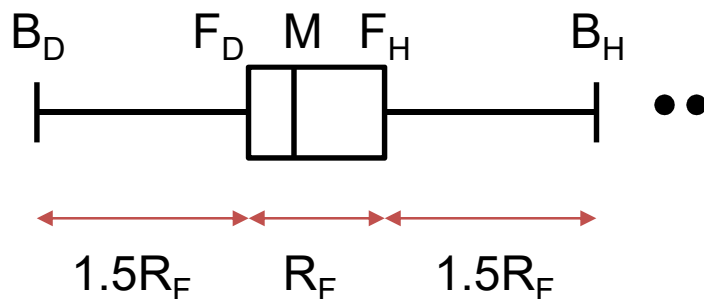


- ❖ získání základní představy o datech
 - ❖ Reprezentace dat (datová matice nebo relační DB?)
 - ❖ Formát dat: počty instancí, atributů
- ❖ chybějící hodnoty
- ❖ deskriptivní charakteristiky dat podle typu dat
 - ❖ četnosti hodnot (histogramy)
 - ❖ minima, maxima, průměry
 - ❖ odlehlé hodnoty
- ❖ Jednoduchá vizualizace dat - informace o
 - ❖ hodnotách 1 příznaku: **histogram, box-plot**
 - ❖ vztahu dvou nebo tří příznaků: **rovinný** nebo **prostorový graf**

Histogram



Box graf



M medián

F_D dolní kvartil

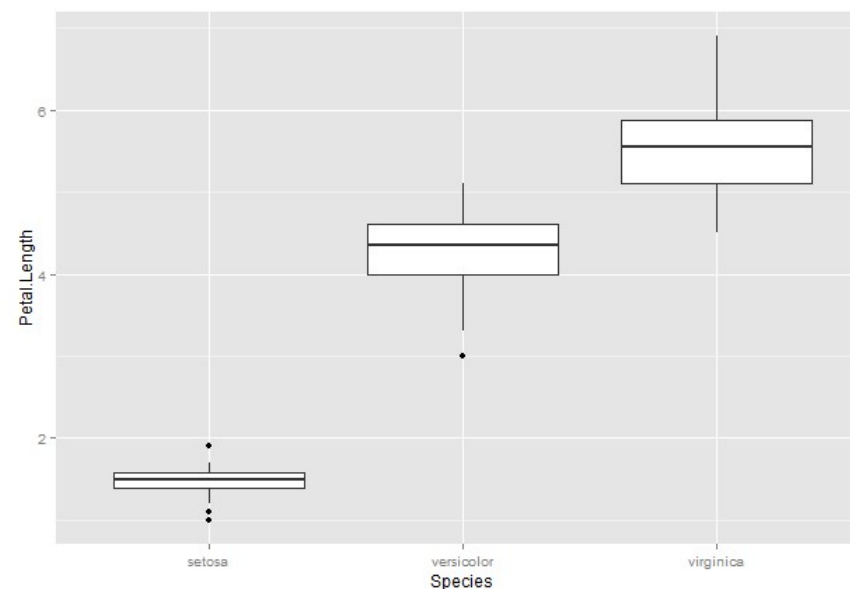
F_H horní kvartil

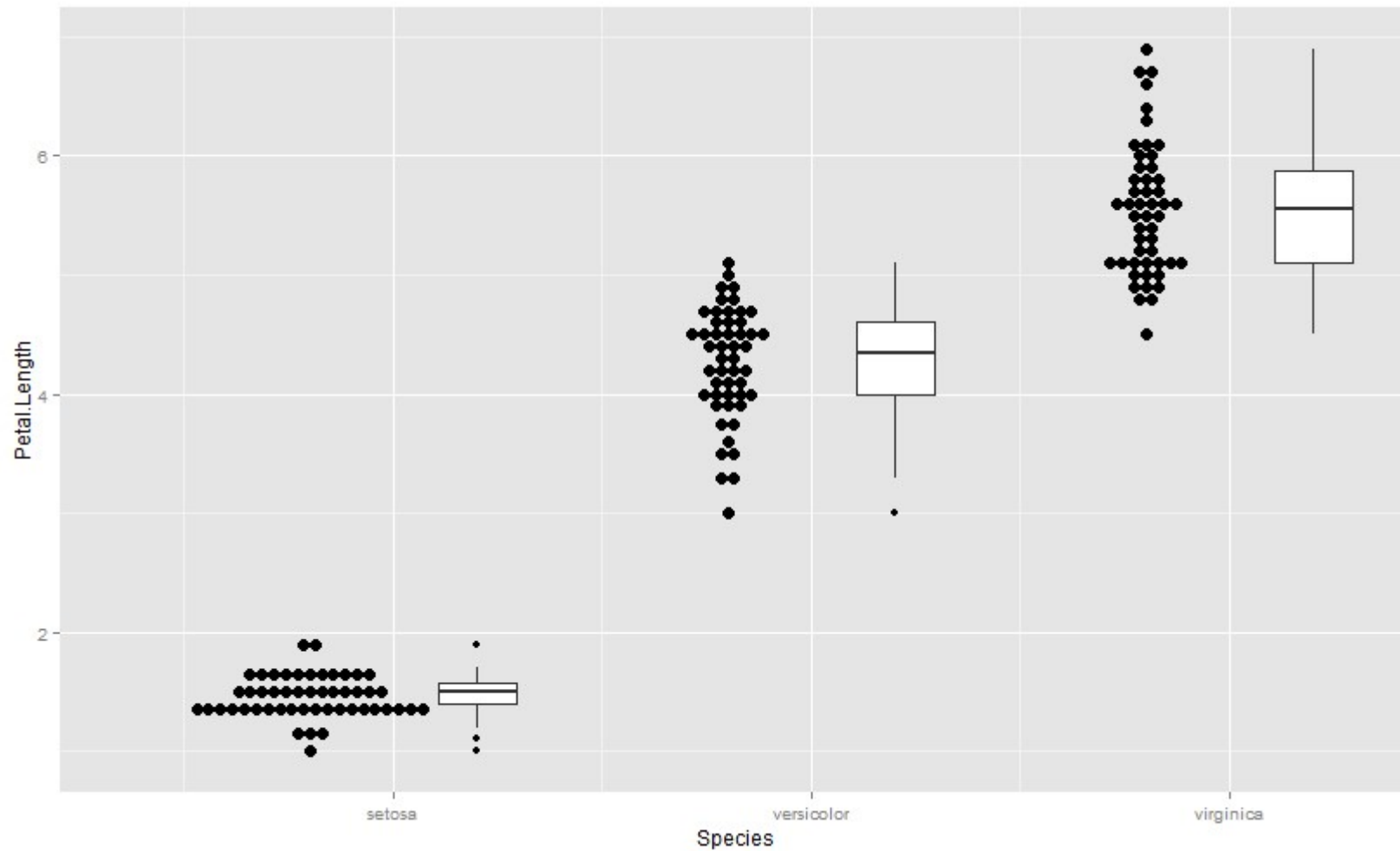
$$R_F = F_H - F_D$$

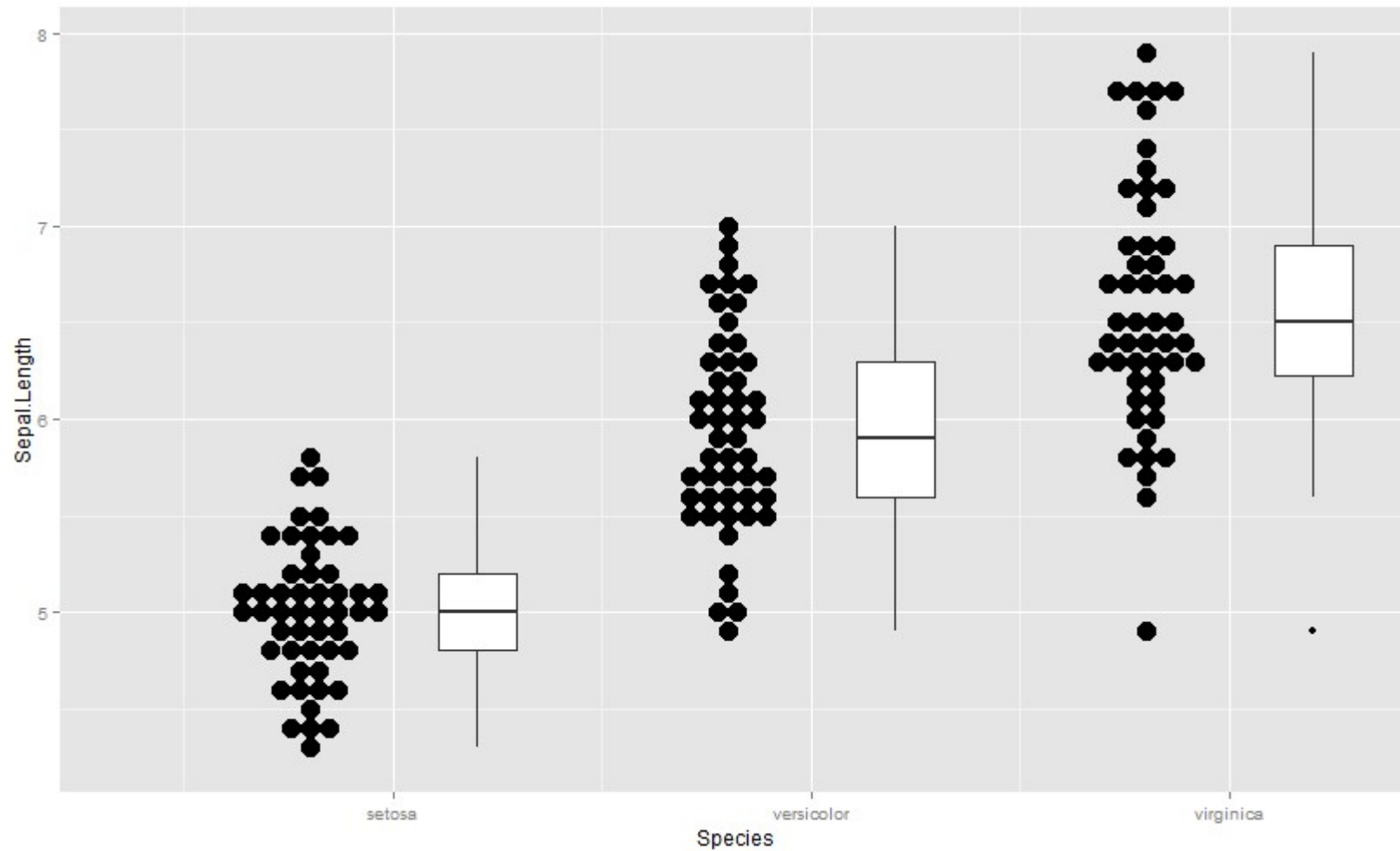
mezikvartilové rozpětí

$$B_D = F_D - 1,5 * R_F \text{ („vousy“)}$$

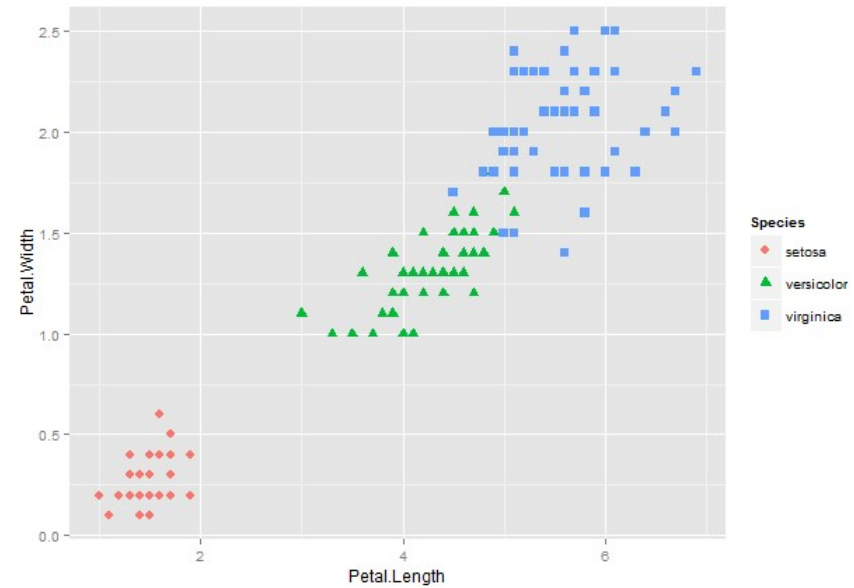
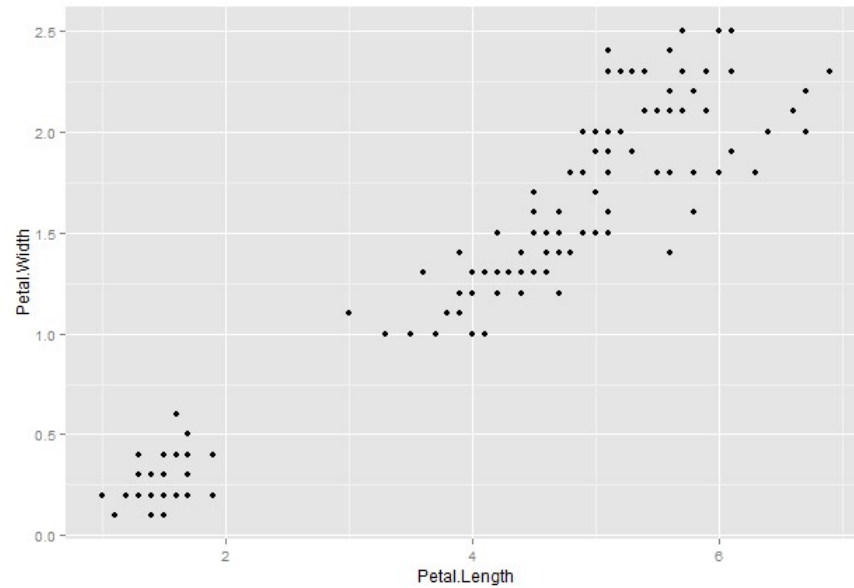
F_D (resp. F_H) bod, ve kterém distribuční funkce náhodné proměnné prochází hodnotou 0,25 (resp. 0,75)







Scater plot = (rovinný) XY graf



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?

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