EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

Název projektu	Rozvoj vzdělávání na Slezské univerzitě v Opavě
Registrační číslo projektu	CZ.02.2.69/0.0./0.0/16_015/0002400

Dolování dat

Úvodní informace a požadavky na absolvování

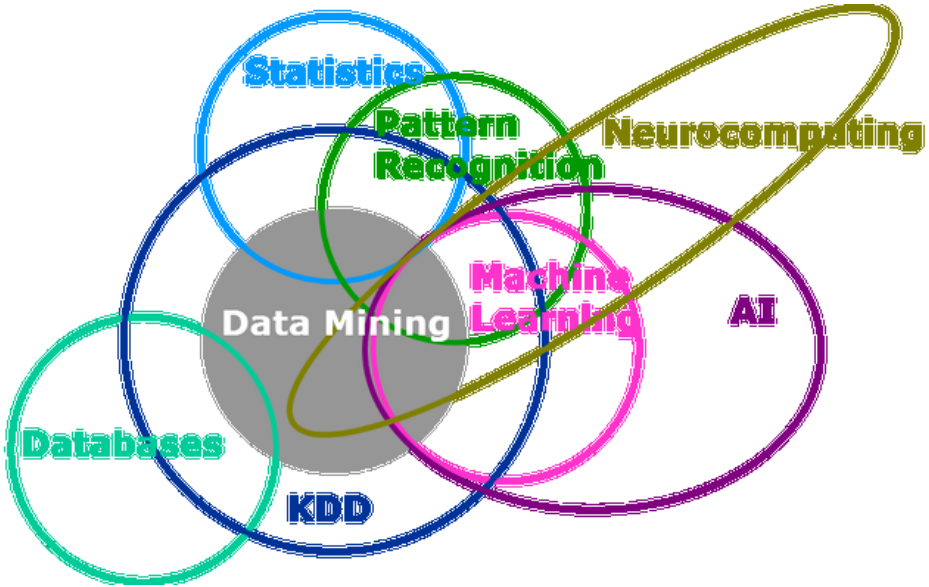
Jan Górecki

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

Dolování dat (Data mining)

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

Non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns from data (Fayyad a kol., 1996)



© MIT 6.S191: Introduction to Deep Learning

The Rise of Deep Learning

Obecné informace o předmětu

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

Dolování dat:

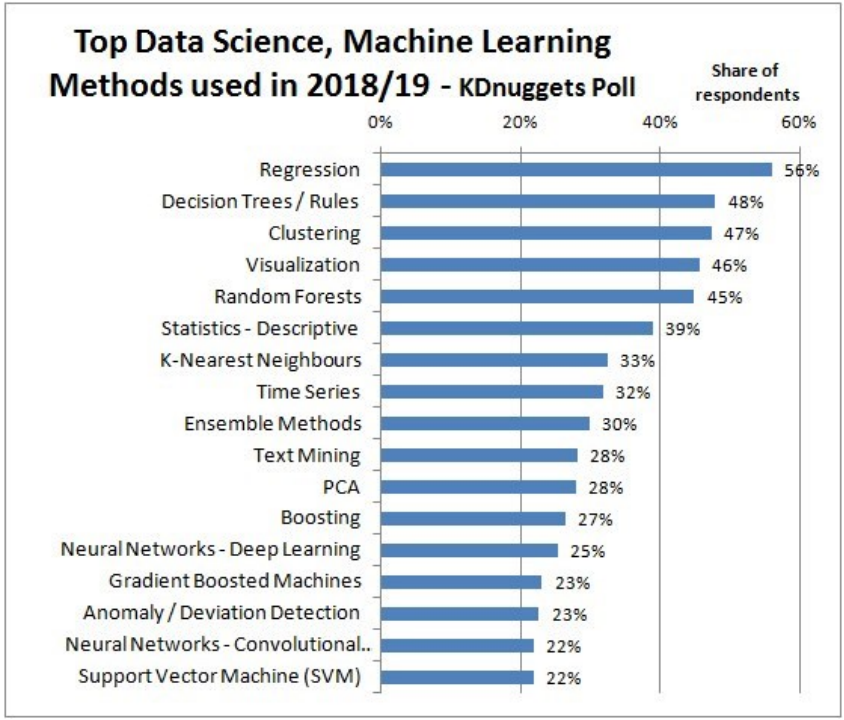
- **Prezenční forma: 13 přednášek a 12 seminářů,**
- **Kombinovaná forma: 3 přednášky**

- **zakončena zkouškou**

Stručná anotace předmětu

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

- **Proces dolování dat**
  Dolování dat, úlohy dolování dat, metodiky pro dolování dat
- **Statistika v kontextu dolování dat**
  Kontingenční tabulky, regresní analýza, diskriminační analýza, shluková analýza
- **Strojové učení**
  Základní pojmy, principy strojového učení, typy strojového učení, formy strojového učení, trénovací data, atributy, chybová funkce
- **Metody dolování dat**
  Rozhodovací stromy, Rozhodovací pravidla, Neuronové sítě, Genetické algoritmy, bayesovské metody, metody založené na analogii
- **Evaluace modelů**
  kritéria, deskriptivní úlohy, klasifikační úlohy, vizualizace modelů, vizualizace klasifikací, porovnávání modelů, volba nejvhodnějšího algoritmu, kombinování modelů
- **Předzpracování dat**
  Příprava dat, strukturovaná data, více vzájemně propojených tabulek, odvozené atributy, příliš mnoho objektů, příliš mnoho atributů, numerické atributy, kategoriální atributy, chybějící hodnoty



Požadavky na absolvování předmětu

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

- **docházka na semináře min. 60% (10 % hodnocení),**
- **zpracování seminární práce (30% hodnocení),**
  - **Analýza vybraných dat dle metodiky CRISP-DM pomocí metod dolování dat (alespoň 5 metod celkově, z nichž alespoň 2 statistické a alespoň 3 ze strojového učení)**
- **zkouška (60% hodnocení)**

Celkem maximum: 100

Požadované minimum: 60

Výukové materiály

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

- **Veškeré elektronické materiály je možné nalézt na školní síti: L:\gorecki\public\NPDOD-NKDOD \ (přes https://raimundo.opf.slu.cz/NetStorage/ popř. files.opf.slu.cz)**

Literatura

Povinná:

- BERKA, P. a GÓRECKI, J., 2017. *Dolování dat*. Skripta SU OPF, Karviná.
- BERKA, P., 2003. *Dobývání znalostí z databází*. Praha: Academia. ISBN 80-200-1062-9.

Doporučená:

- CLARK, B., E. FOKOUE a H. H. ZHANG, 2009. *Principles and theory for data mining and machine learning*. New York: Springer. ISBN 978-0-387-98134-5.
- MURPHY, K. P., 2012. *Machine learning: A probabilistic perspective.* London, England: The MIT Press. ISBN 978-0-262-01802-9.

Software

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

- **MATLAB**
  - **Statistics and Machine Learning Toolbox**
  - **https://www.mathworks.com/solutions/data-science.html**
  - **trial verze z mathworks.com**
  - **Octave – free verze MATLABu**
- **Python**
- **R**
- **RapidMiner**

MathWorks®

ANACONDA.

rapidminer

alteryx | The Thrill of Solving

Data

SLEZSKÁ
UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ
FAKULTA V KARVINÉ

- **Nejlépe vlastní**
- **UC Irvine Machine Learning Repository**
  https://archive.ics.uci.edu/ml/index.php
- **Kaggle: Your Home for Data Science**
  https://www.kaggle.com/
- **KEEL - dataset repository**
  http://www.keel.es/datasets.php

Kontakty

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

Ing. Jan Górecki, Ph.D.
gorecki@opf.slu.cz
A407

- **konzultace po domluvě emailem**

Sekretariát Katedry informatiky a matematiky
A402

EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

Název projektu	Rozvoj vzdělávání na Slezské univerzitě v Opavě
Registrační číslo projektu	CZ.02.2.69/0.0./0.0/16_015/0002400

Dolování dat

Dolování dat

Jan Górecki

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

Obsah přednášky

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

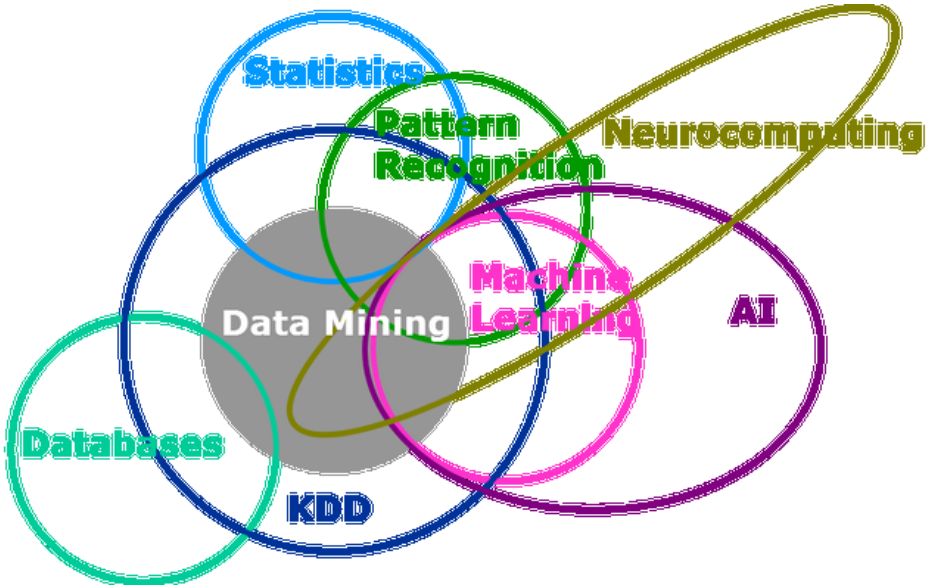
- Definice …
- Historie …
- Úlohy …
- Pohledy na …
- Postupy (metodiky) …
- Software pro …
- Příklad …

… Dolování dat

Dolování dat (Data mining)

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

Non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns from data (Fayyad a kol., 1996)



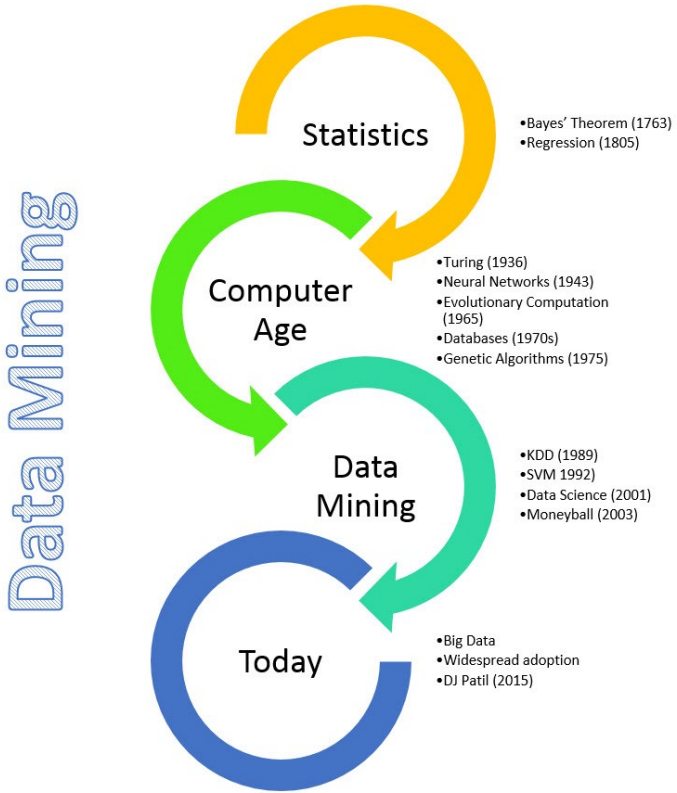
Dolování dat (Knowledge Discovery in Databases, Data Mining, ..., Knowledge Destilery, ...)

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

- Analysis of observational data sets to find unsuspected relationships and summarize data in novel ways that are both understandable and useful to the data owner (Hand, Manilla, Smyth, 2001)

Trocha historie

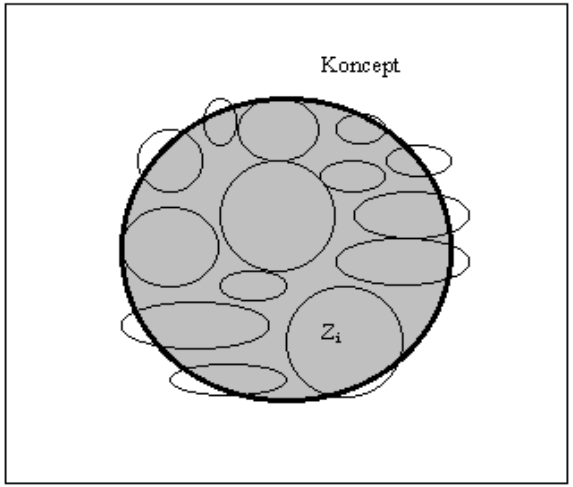
SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ



Úlohy dobývání znalostí

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

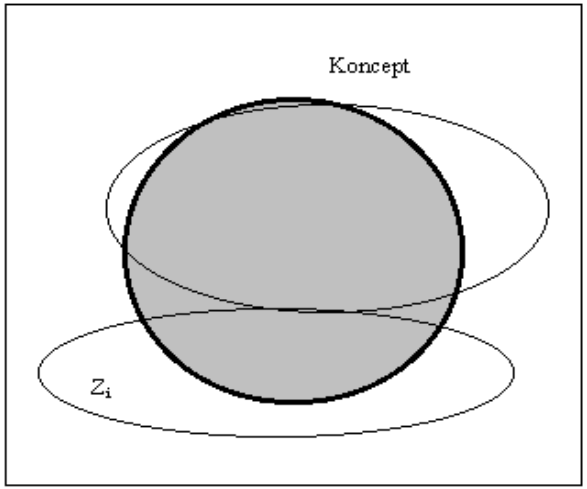
klasifikace/predikce: cílem je nalézt znalosti použitelné pro klasifikaci nových případů



Úlohy dobývání znalostí

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

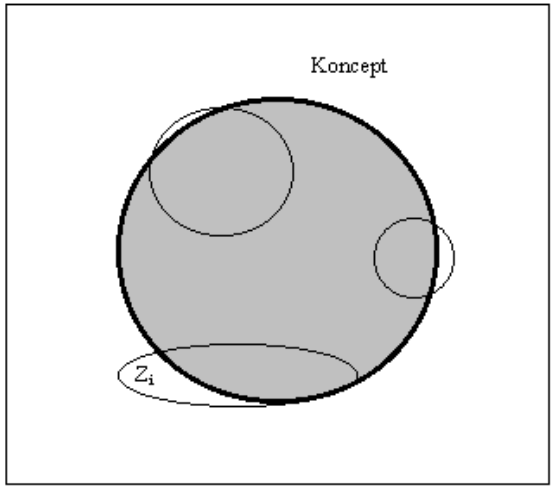
deskripce: cílem je nalézt dominantní strukturu nebo vazby



Úlohy dobývání znalostí

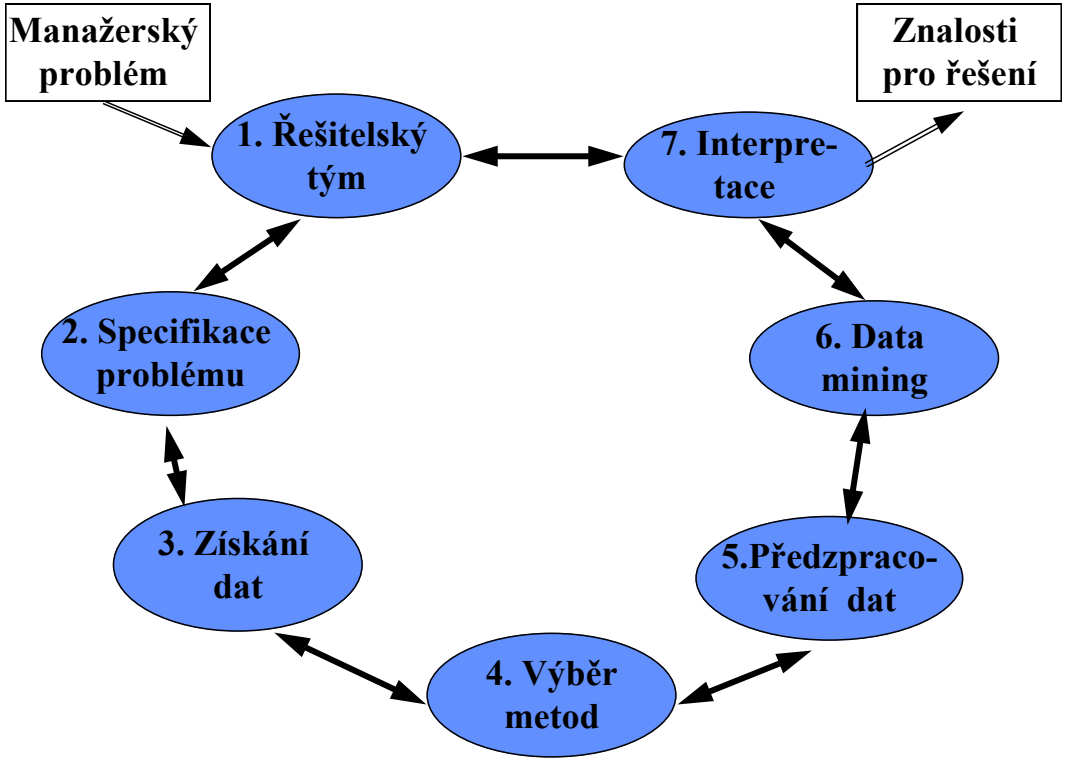
SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

hledání „nugetů“: cílem je nalézt dílčí překvapivé znalosti



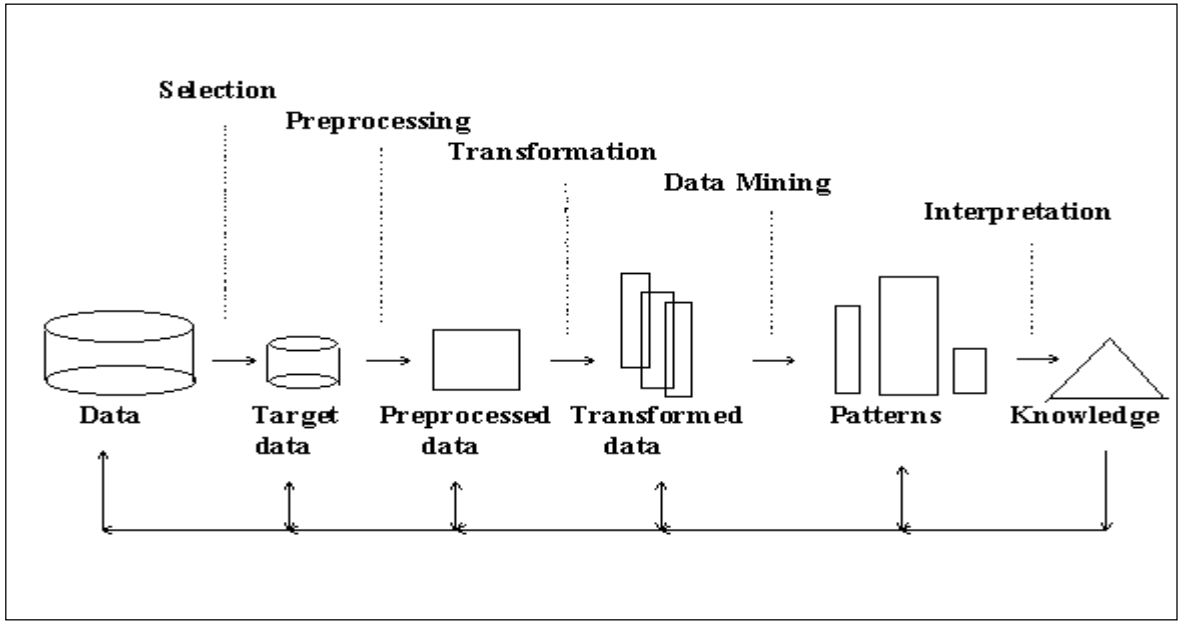
Manažerský pohled

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ



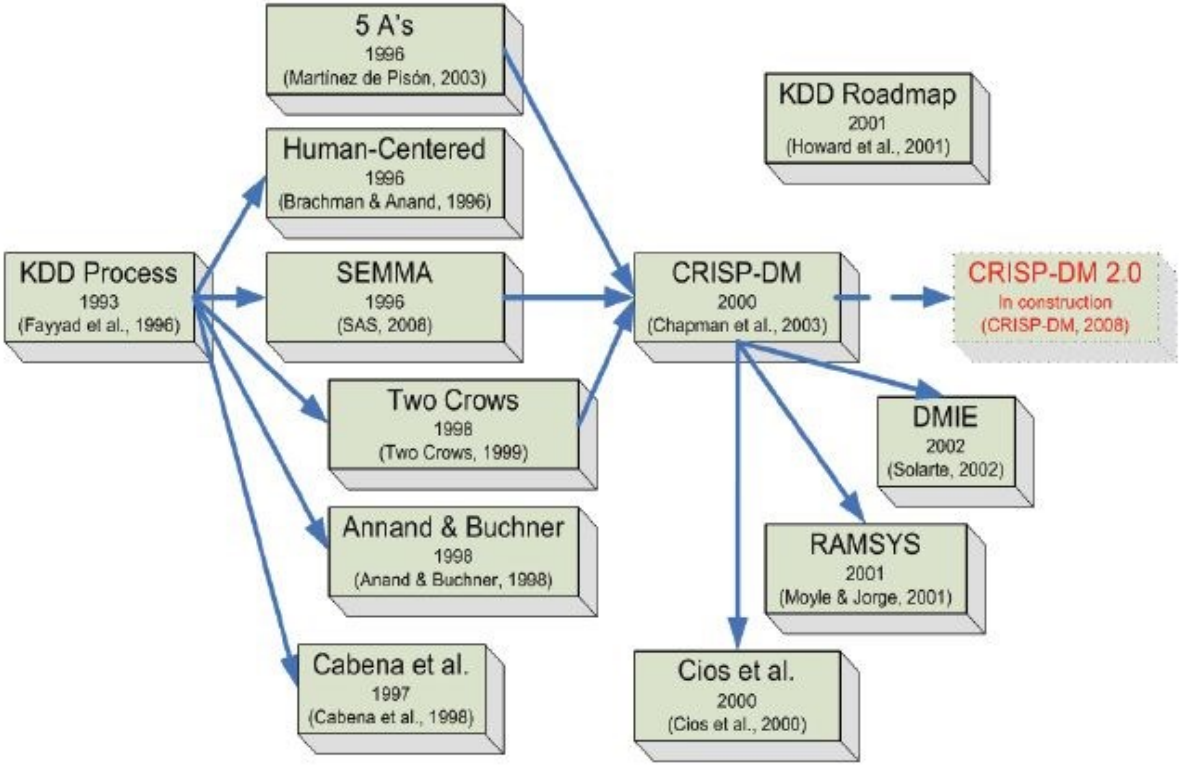
Pohled zpracování dat

SLEZSKÁ
UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ
FAKULTA V KARVINÉ



Standardy pro dobývání znalostí - Metodiky

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ



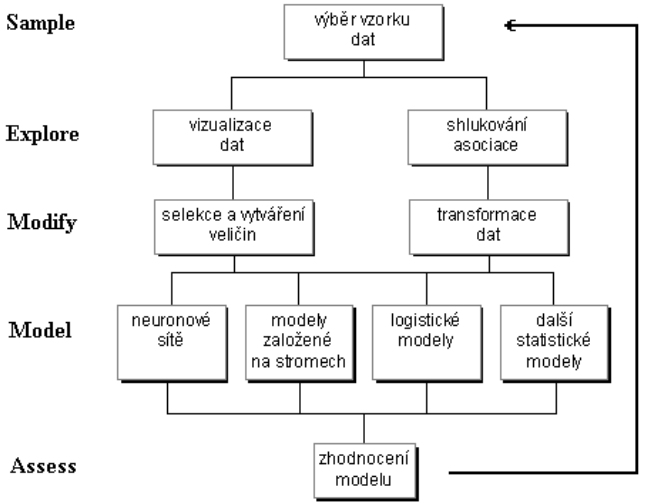
(Marban a kol, 2009)

Metodika SEMMA

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

Navržená pro Enterprise Miner firmy SAS:

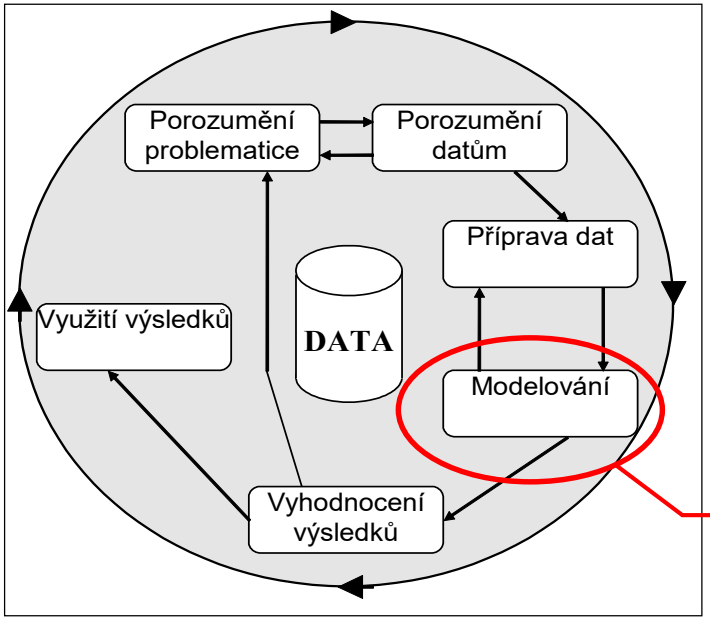
- **S**ample (vybrání vhodných objektů),
- **E**xplore (vizuální explorace a redukce dat),
- **M**odify (seskupování objektů a hodnot atributů, datové transformace),
- **M**odel (analýza dat: neuronové sítě, rozhodovací stromy, statistické techniky, asociace a shlukování),
- **A**ssess (porovnání modelů a interpretace).



Metodika CRISP-DM

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

V současnosti de-facto standard podporovaný většinou systémů pro dobývání znalostí



Data Mining

Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine Business Objectives *Background* *Business Objectives* *Business Success Criteria* **Assess Situation** *Inventory of Resources* *Requirements, Assumptions, and Constraints* *Risks and Contingencies* *Terminology* *Costs and Benefits* **Determine Data Mining Goals** *Data Mining Goals* *Data Mining Success Criteria* **Produce Project Plan** *Project Plan* *Initial Assessment of Tools and Techniques*	**Collect Initial Data** *Initial Data Collection Report* **Describe Data** *Data Description Report* **Explore Data** *Data Exploration Report* **Verify Data Quality** *Data Quality Report*	*Data Set* *Data Set Description* **Select Data** *Rationale for Inclusion / Exclusion* **Clean Data** *Data Cleaning Report* **Construct Data** *Derived Attributes* *Generated Records* **Integrate Data** *Merged Data* **Format Data** *Reformatted Data*	**Select Modeling Technique** *Modeling Technique* *Modeling Assumptions* **Generate Test Design** *Test Design* **Build Model** *Parameter Settings* *Models* *Model Description* **Assess Model** *Model Assessment* *Revised Parameter Settings*	**Evaluate Results** *Assessment of Data Mining Results w.r.t. Business Success Criteria* *Approved Models* **Review Process** *Review of Process* **Determine Next Steps** *List of Possible Actions* *Decision*	**Plan Deployment** *Deployment Plan* **Plan Monitoring and Maintenance** *Monitoring and Maintenance Plan* **Produce Final Report** *Final Report* *Final Presentation* **Review Project** *Experience Documentation*

Systémy pro dobývání znalostí z databází

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

- pokrývají celý proces dobývání znalostí (od předzpracování po interpretaci),
- nabízejí více algoritmů pro analýzu (než „jednoúčelové" systémy strojového učení),
- kladou důraz na vizualizaci (ve způsobu práce se systémem i při interpretaci výsledků).

Systémy pro dobývání znalostí z databází

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

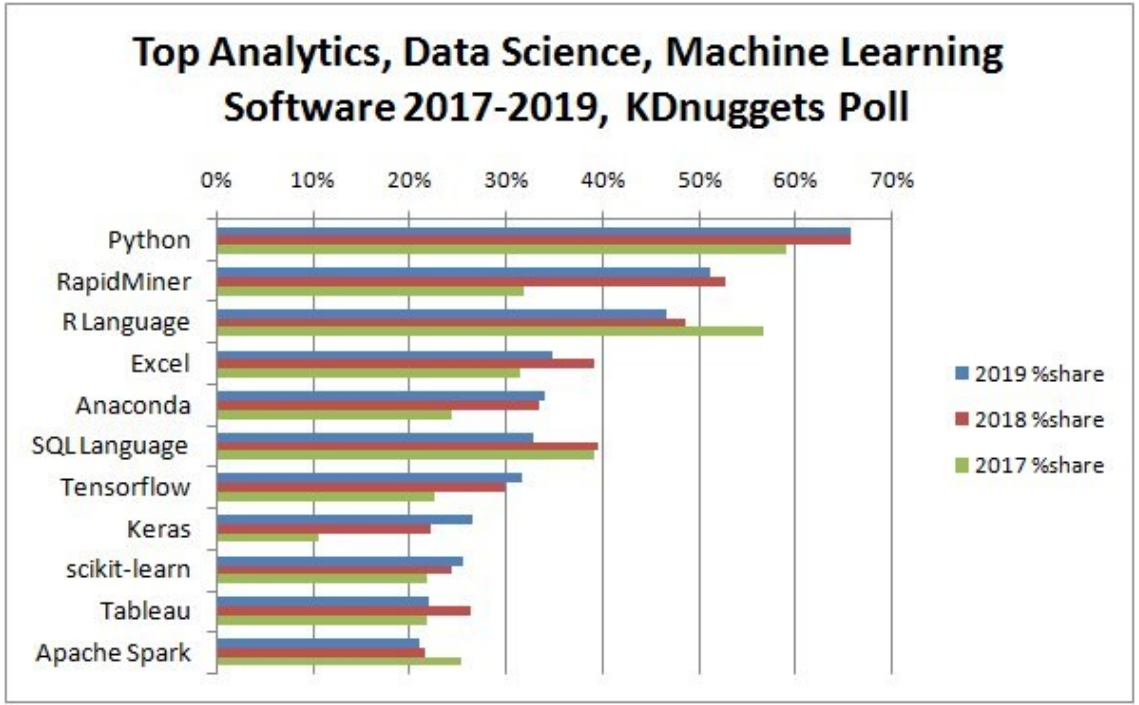
Gartner peerinsights customers' choice 2019

MathWorks

ANACONDA.

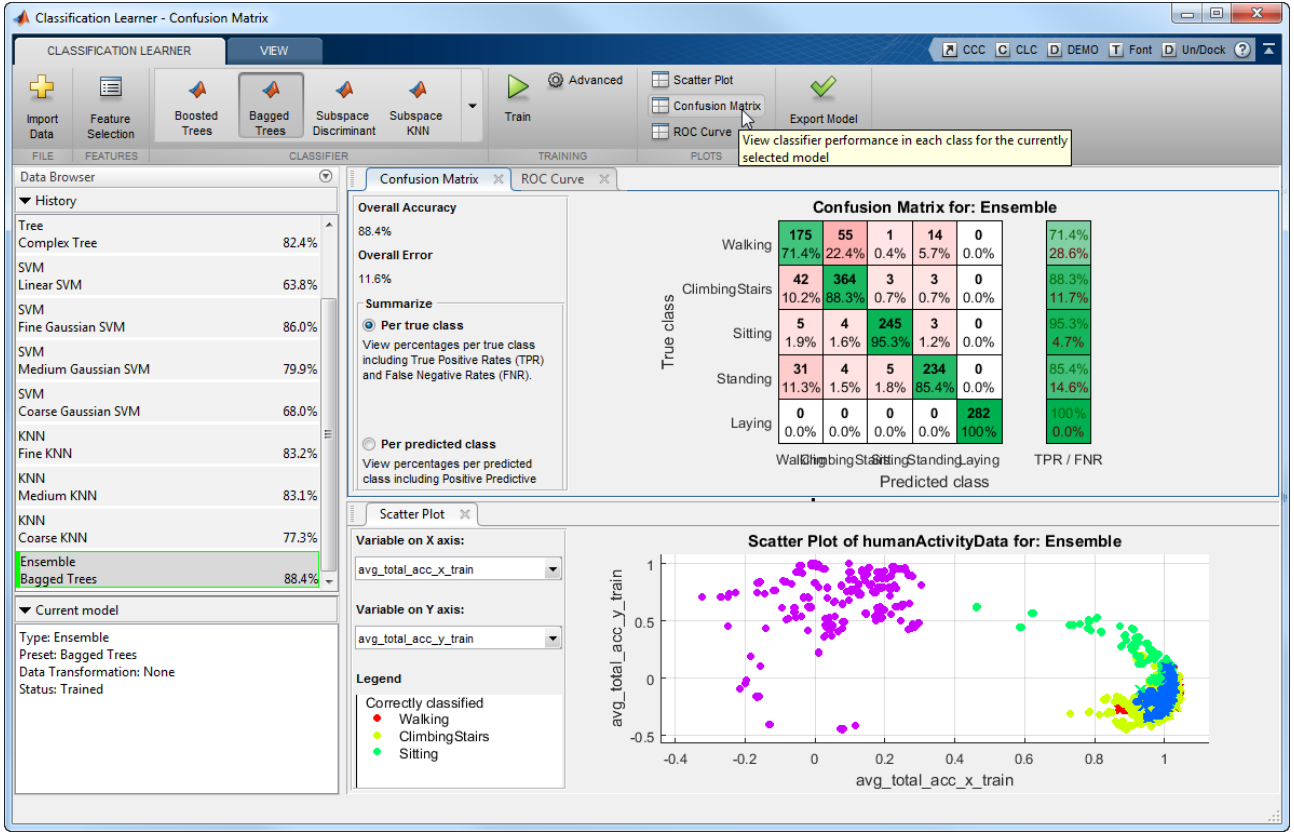
rapidminer

alteryx The Thrill of Solving



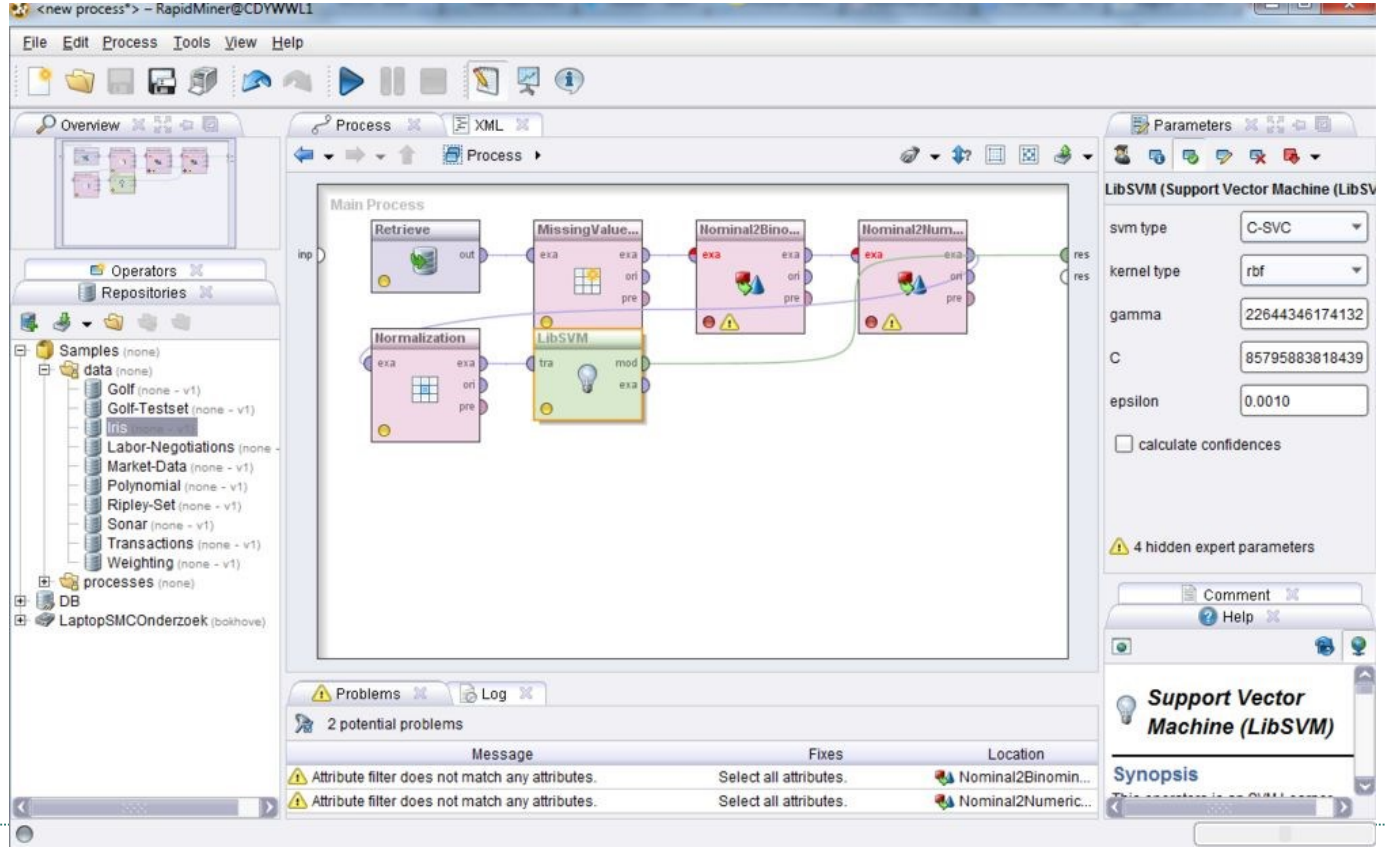
MATLAB

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ



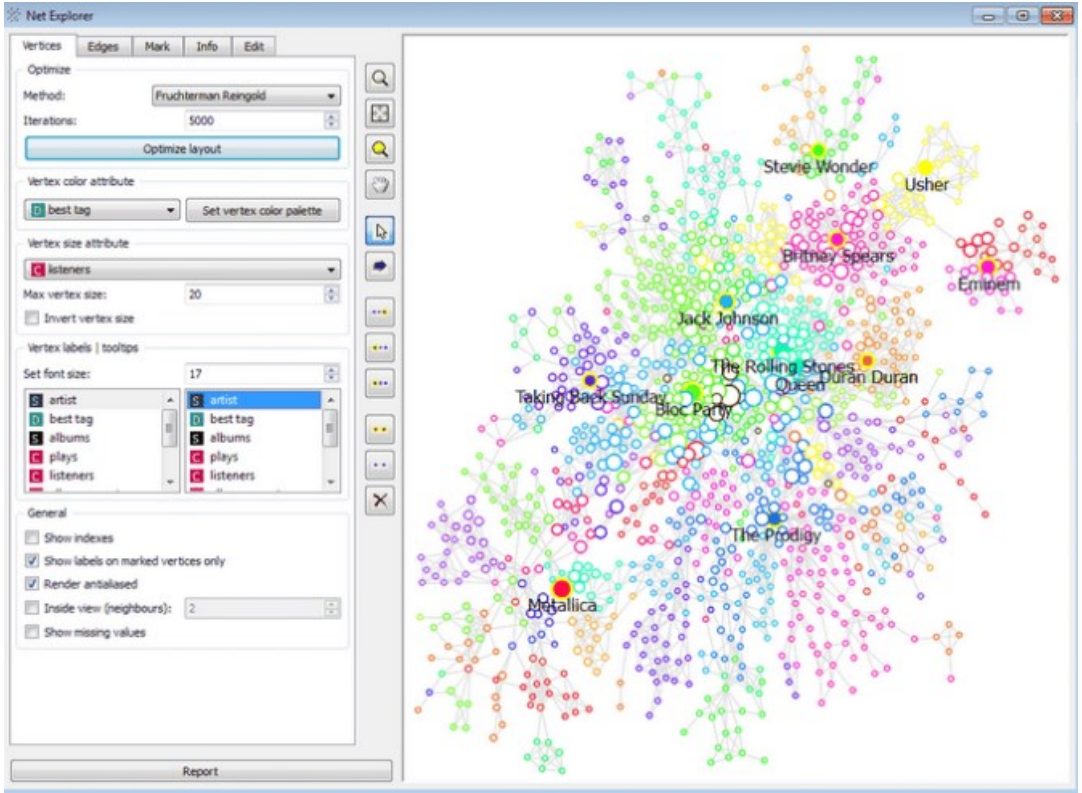
Rapid Miner

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ



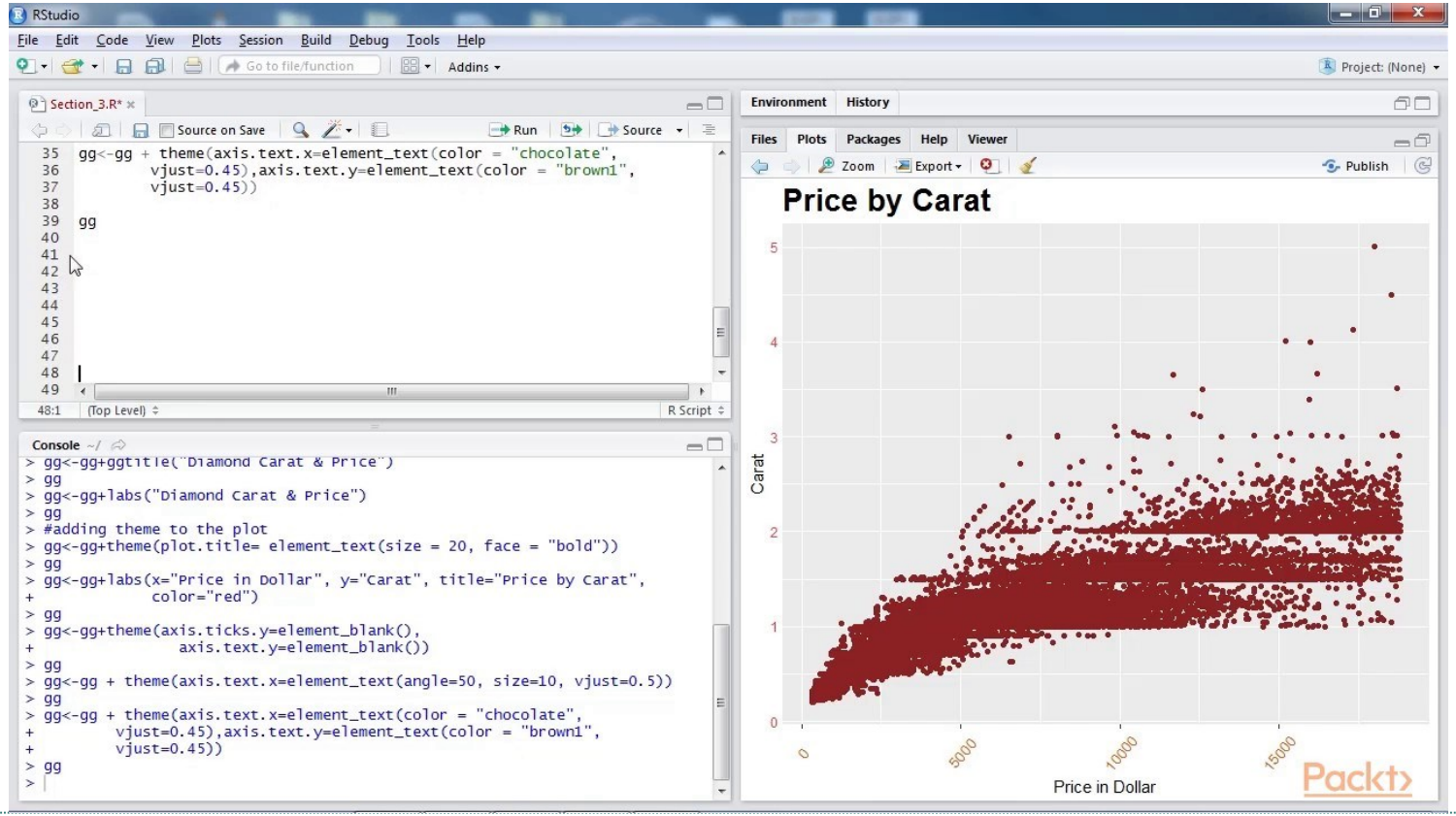
Python (Orange)

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ



Orange

R


RStudio
File Edit Code View Plots Session Build Debug Tools Help
Addins
Project: (None)
Section_3.R*
Source on Save
Run
Source
35 gg<-gg + theme(axis.text.x=element_text(color = "chocolate",
36 vjust=0.45),axis.text.y=element_text(color = "brown1",
37 vjust=0.45))
38
39 gg
40
41
42
43
44
45
46
47
48
49
48:1 (Top Level)
R Script
Console ~/
> gg<-gg+ggtitle("Diamond Carat & Price")
> gg
> gg<-gg+labs("Diamond Carat & Price")
> gg
> #adding theme to the plot
> gg<-gg+theme(plot.title= element_text(size = 20, face = "bold"))
> gg
> gg<-gg+labs(x="Price in Dollar", y="Carat", title="Price by Carat",
+ color="red")
> gg
> gg<-gg+theme(axis.ticks.y=element_blank(),
+ axis.text.y=element_blank())
> gg
> gg<-gg + theme(axis.text.x=element_text(angle=50, size=10, vjust=0.5))
> gg
> gg<-gg + theme(axis.text.x=element_text(color = "chocolate",
+ vjust=0.45),axis.text.y=element_text(color = "brown1",
+ vjust=0.45))
> gg
>
Environment History
Files Plots Packages Help Viewer
Zoom Export Publish
Price by Carat
Carat
0 1 2 3 4 5
0 5000 10000 15000
Price in Dollar
Packt>

Aplikační oblasti pro dobývání znalostí

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

- Segmentace a klasifikace klientů banky (např. rozpoznání problémových nebo naopak vysoce bonitních klientů),
- Predikce vývoje kursů akcií,
- Predikce spotřeby elektrické energie,
- Analýza příčin poruch v telekomunikačních sítích,
- Analýza důvodů změny poskytovatele nějakých služeb (internet, mobilní telefony),
- Segmentace a klasifikace klientů pojišťovny,
- Určení příčin poruch automobilů,
- Rozbor databáze pacientů v nemocnici,
- **Rozpoznání činnosti uživatele pomocí senzorů z mobilního telefonu.**

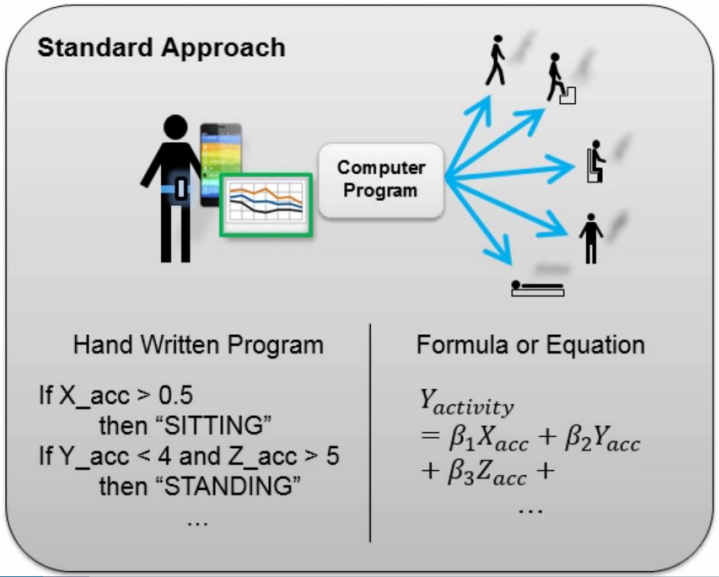
Rozpoznání činnosti uživatele

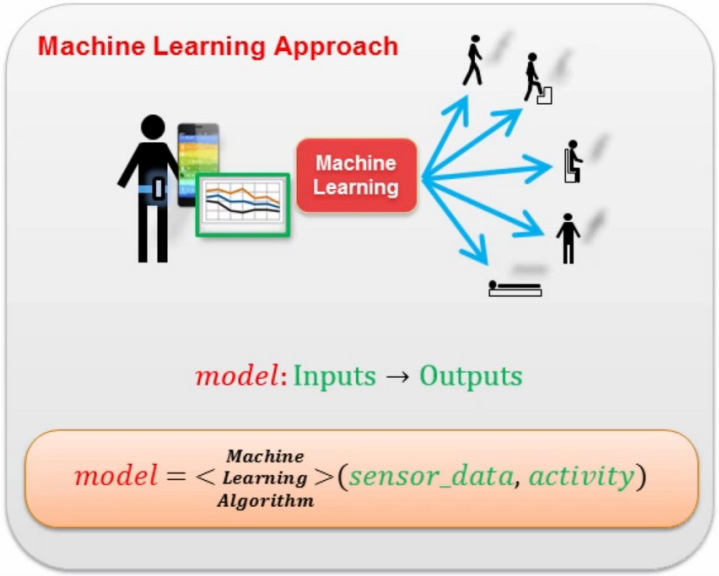
SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

Machine Learning

Machine learning uses **data** and produces a **program** to perform a **task**

Task: Human Activity Detection



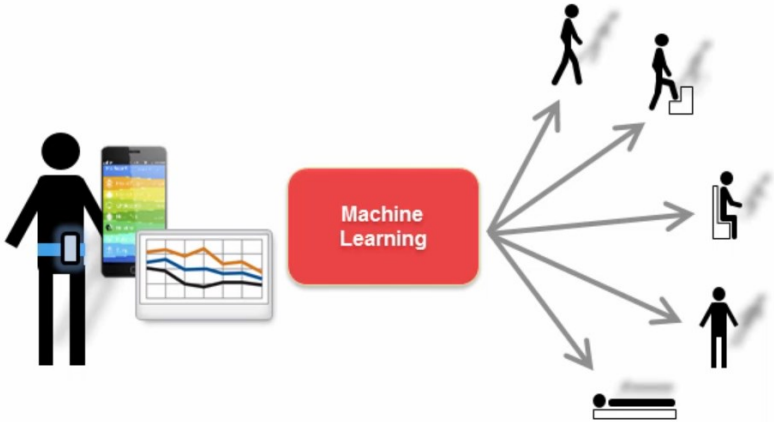


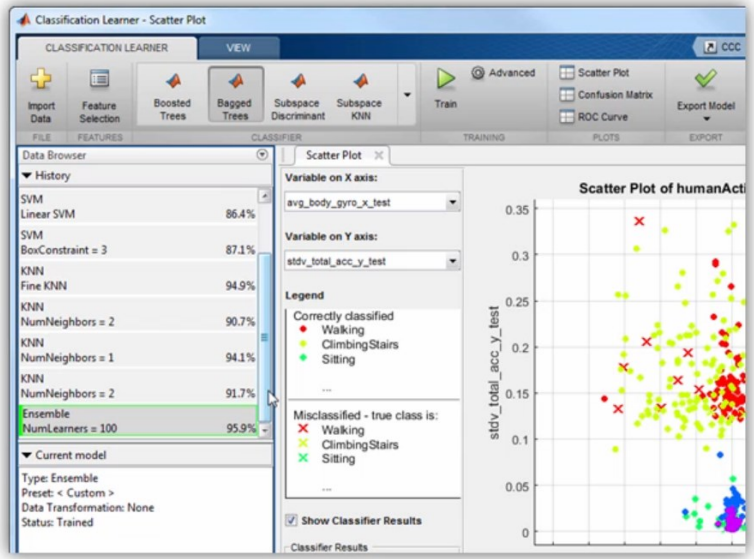
https://www.mathworks.com/videos/machine-learning-with-matlab-100694.html?elqsid=1569409667237&potential_use=Education

Rozpoznání činnosti uživatele

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

Example: Human Activity Learning Using Mobile Phone Data





Data:

- 3-axial Accelerometer data
- 3-axial Gyroscope data

https://www.mathworks.com/videos/machine-learning-with-matlab-100694.html?elqsid=1569409667237&potential_use=Education

Děkuji za pozornost

Některé snímky převzaty od:

prof. Ing. Petr Berka, CSc. berka@vse.cz