



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

**MS  
MT**  
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 KARVINĚ

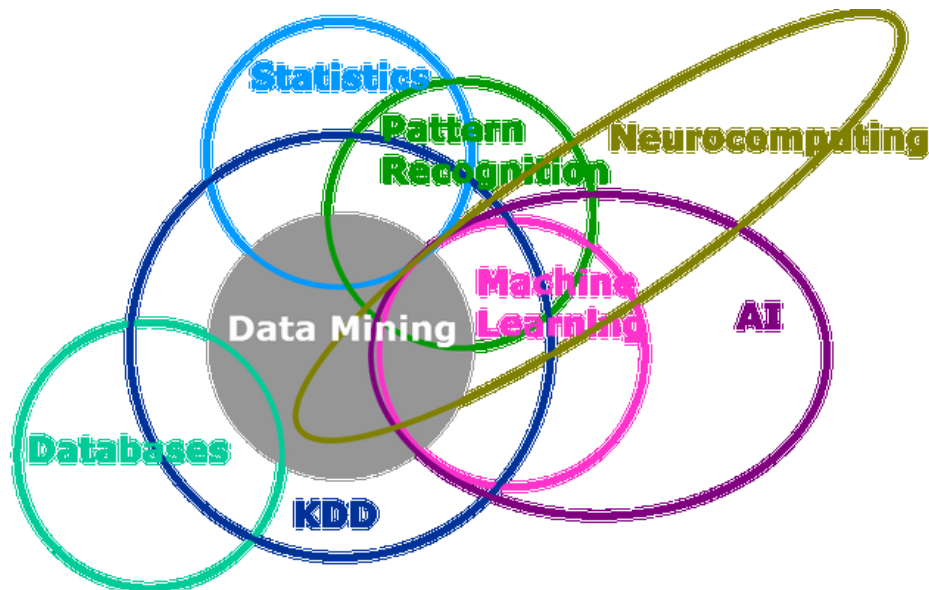
# Dolování dat (Data mining)

---



SLEZSKÁ  
UNIVERZITA  
OBCHODNĚ PODNIKATELSKÁ  
FAKULTA V KARVÍNĚ

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



# The Rise of Deep Learning

## 'Deep Voice' Software Can Clone Anyone's Voice With Just 3.7 Seconds of Audio

Using snippets of voices, Baidu's 'Deep Voice' can generate new speech, accents, and tones.



with DEEPMIND STARCRIFT TRIUMPH FOR

Let There Be Sight: How Deep Learning Is Helping the Blind 'See'



## Technology outpacing security measures

Facial Recognition | Features and Interviews



## AI beats docs in cancer spotting

A new study provides a fresh example of machine learning as an important diagnostic tool. Paul Binger reports.



## AI Can Help In Predicting Cryptocurrency Value



## 'Creative' AlphaZero leads way for chess computers and, maybe, science

Former chess world champion Garry Kasparov likes what he has to say about a computer that could be used to find cures for diseases



## How an A.I. 'Cat-and-Mouse Game' Generates Believable Fake Photos

By CADE METZ and KEITH HOLENBERG - JAN. 2, 2018



AI: Faked Data



## Human faces show how far AI image generation has come in just four years

People on the right aren't real: they're the product of machine learning



## Stock Predictions Based On AI: Is the Market Truly Predictable?



## Neural networks everywhere

New chip reduces neural networks' power consumption by up to 95 percent, making them practical for battery-powered devices.

DeepMind, Google - Boston - Comment by Kerry Walker - Digital Reporter - @RandPhugase

## After Millions of Trials, These Simulated Humans Learned to Do Perfect Backflips and Cartwheels

George Drachler - MIT News Office - FEBRUARY 2, 2018



## Researchers introduce a deep learning method that converts mono audio recordings into 3D sounds using video scenes



## Automation And Algorithms: De-Risking Manufacturing With Artificial Intelligence

Sarah Goehke - Contributor - Manufacturing - 1 item in the Industrialization of additive manufacturing

TWEET THIS The two key applications of AI in manufacturing are pricing and manufacturability feedback

Complex of bacteria-infecting viral proteins modeled in CASP-13. The complex consists of 10 proteins that were modeled individually. PROTEIN DATA BANK

## Google's DeepMind acs protein folding

By Robert F. Service | Dec. 6, 2018, 12:05 PM



## **Dolování dat:**

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

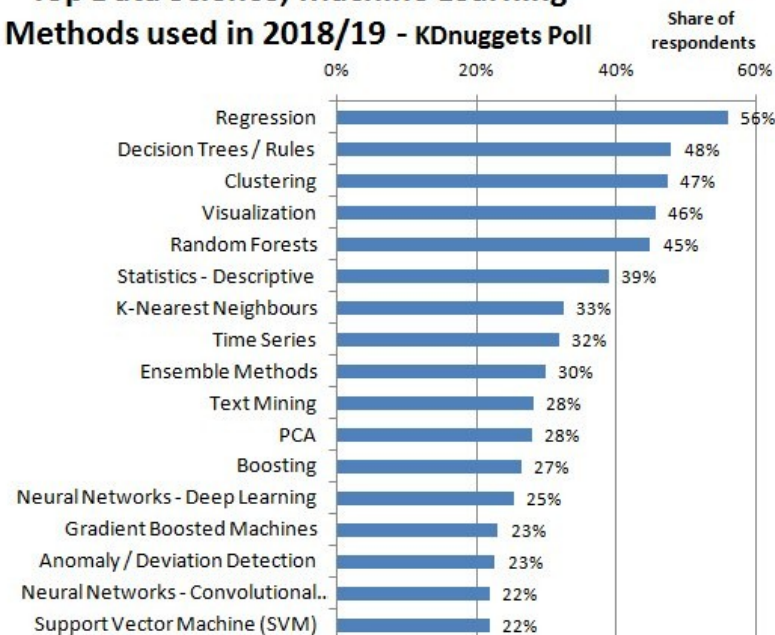
# Stručná anotace předmětu



**SLEZSKÁ  
UNIVERZITA**  
OBCHODNĚ PODNIKATELSKÁ  
FAKULTA V KARVINĚ

- **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

**Top Data Science, Machine Learning  
Methods used in 2018/19 - KDnuggets Poll**



# Požadavky na absolvování předmětu

---



**SLEZSKÁ  
UNIVERZITA**  
OBCHODNĚ PODNIKATELSKÁ  
FAKULTA V KARVINĚ

- **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**

---



- **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](https://files.opf.slu.cz/))**
-

## 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

---



- **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**





- **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 KARVÍNĚ

**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 KARVINĚ

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



... Dolování dat

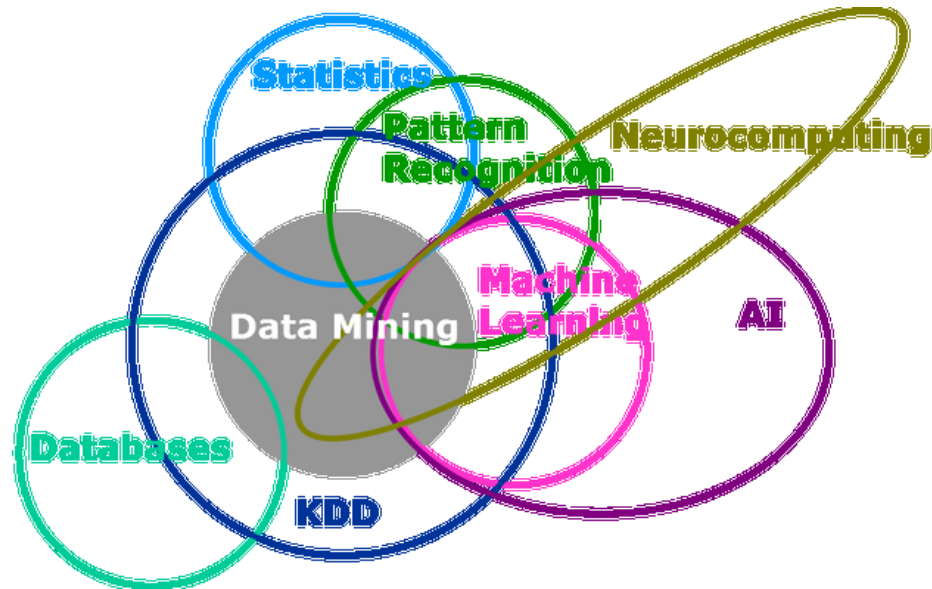
# Dolování dat (Data mining)

---



SLEZSKÁ  
UNIVERZITA  
OBCHODNĚ PODNIKATELSKÁ  
FAKULTA V KARVÍNĚ

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 KARVINĚ

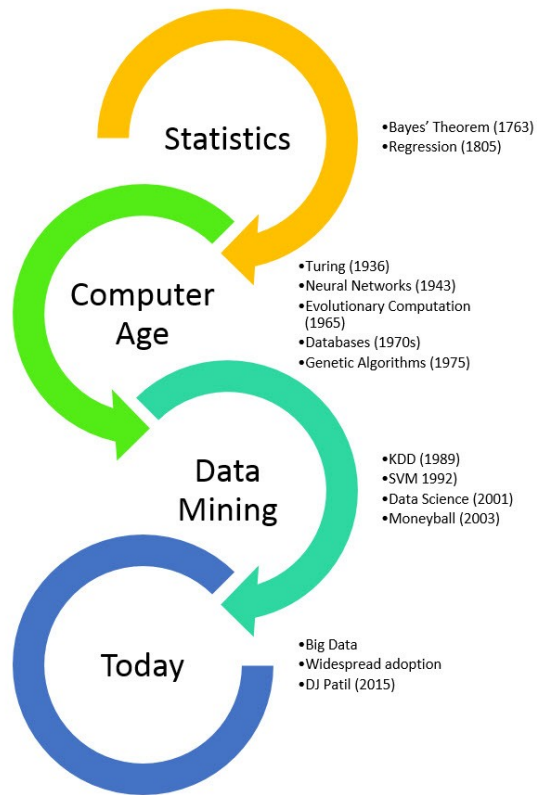
- Analysis of observational data sets to find unsuspected relationships and summarize data in novel ways that are both understandable and useful to the data owner (Hand, Manilla, Smyth, 2001)

# Trocha historie



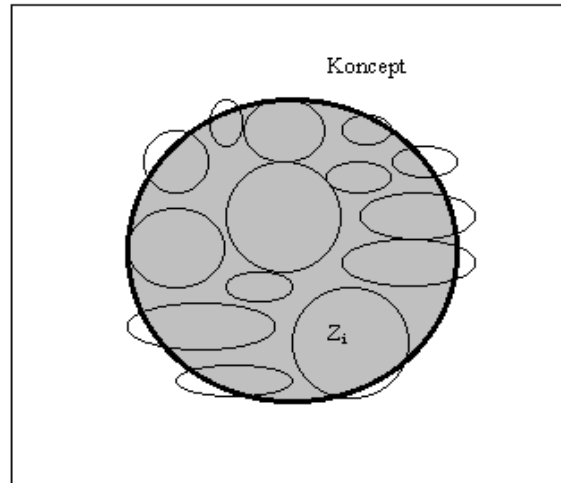
**SLEZSKÁ  
UNIVERZITA**  
OBCHODNĚ PODNIKATELSKÁ  
FAKULTA V KARVINĚ

## Data Mining

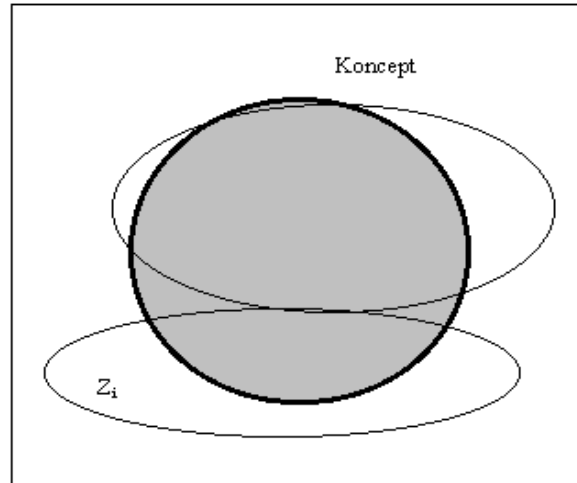




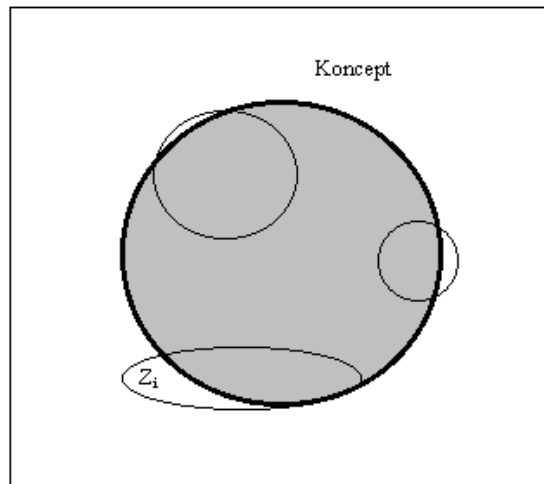
**klasifikace/predikce:** cílem je nalézt znalosti  
použitelné pro klasifikaci nových případů



**deskripce:** cílem je nalézt dominantní strukturu  
nebo vazby



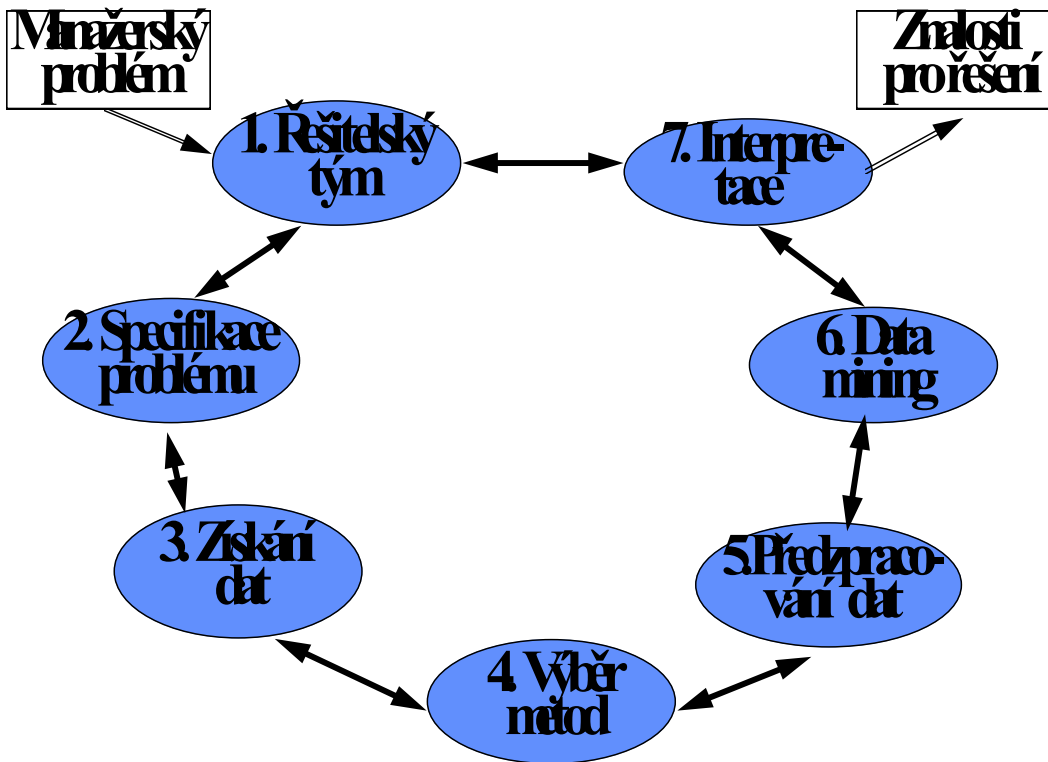
**hledání „nugetů“: cílem je nalézt dílčí  
překvapivé znalosti**



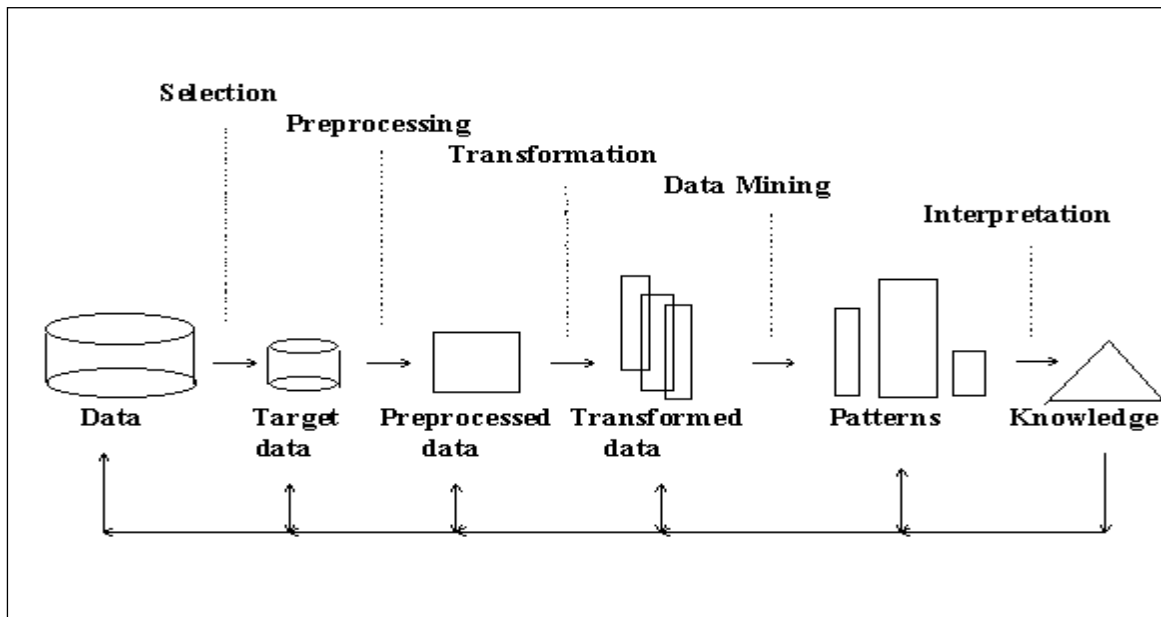
# Manažerský pohled



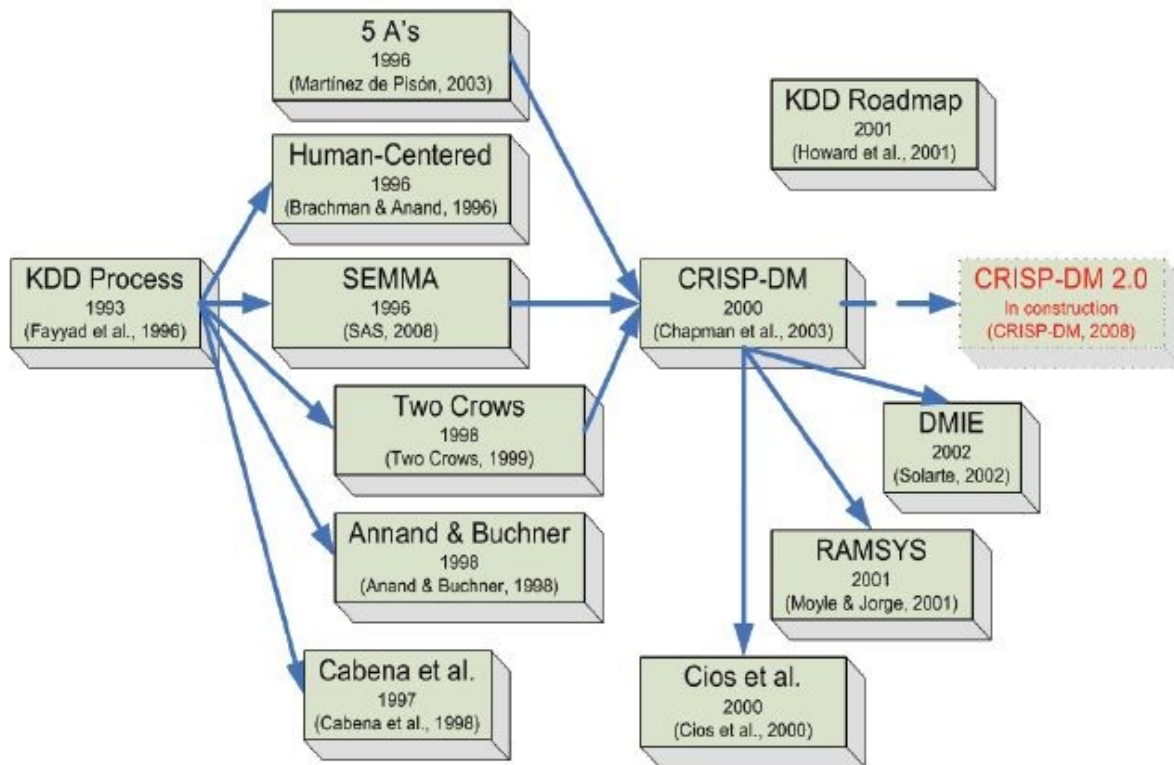
SLEZSKÁ  
UNIVERZITA  
OBCHODNĚ PODNIKATELSKÁ  
FAKULTA V KARVÍNĚ



# Pohled zpracování dat



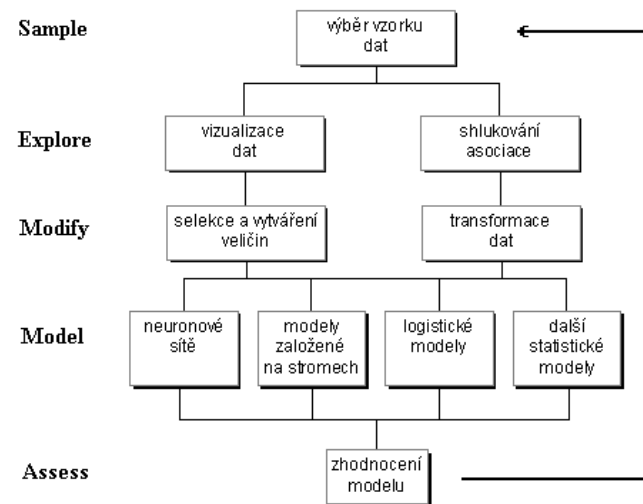
# Standardy pro dobývání znalostí - Metodiky



(Marban a kol, 2009)

Navržená pro Enterprise Miner firmy SAS:

- **Sample** (vybrání vhodných objektů),
- **Explore** (vizuální explorace a redukce dat),
- **Modify** (seskupování objektů a hodnot atributů, datové transformace),
- **Model** (analýza dat: neuronové sítě, rozhodovací stromy, statistické techniky, asociace a shlukování),
- **Assess** (porovnání modelů a interpretace).

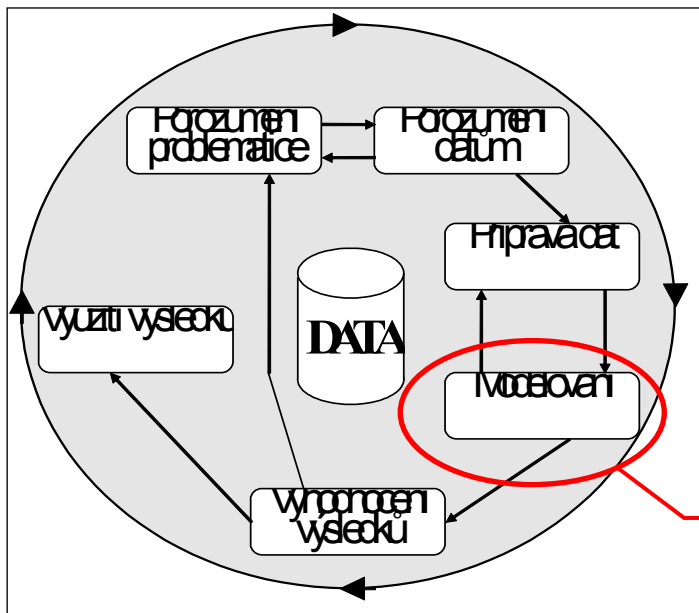


# Metodika CRISP-DM



**SLEZSKÁ  
UNIVERZITA**  
OBCHODNĚ PODNIKATELSKÁ  
FAKULTA V KARVINĚ

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



Data Mining

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



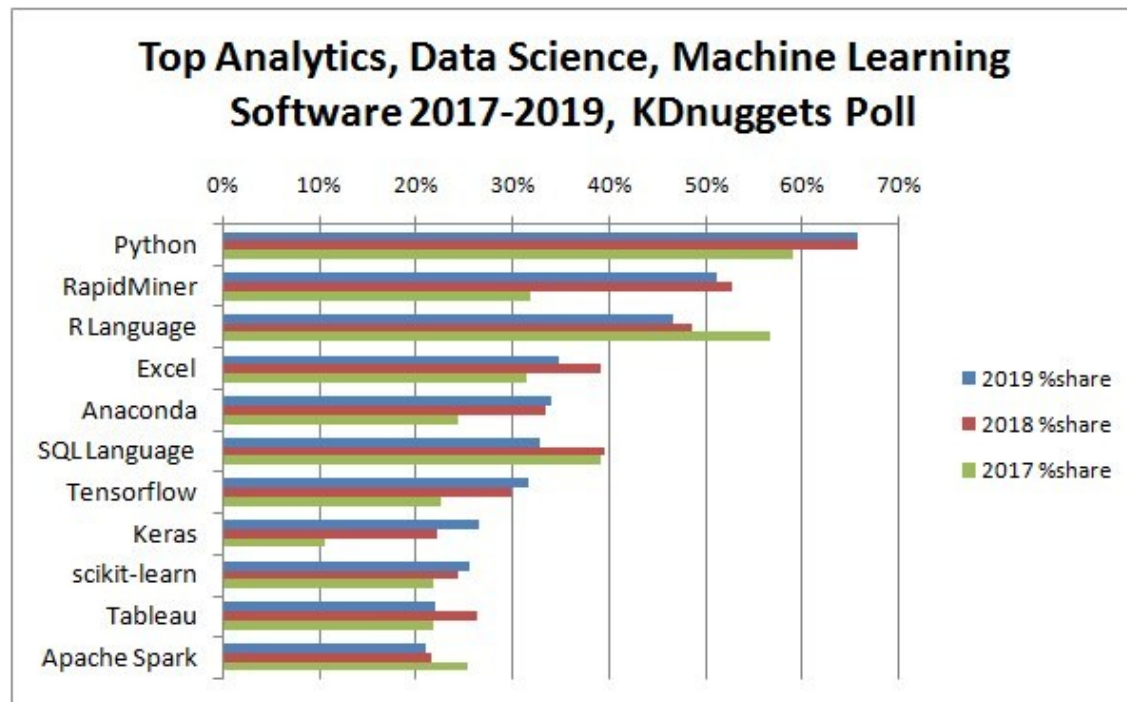


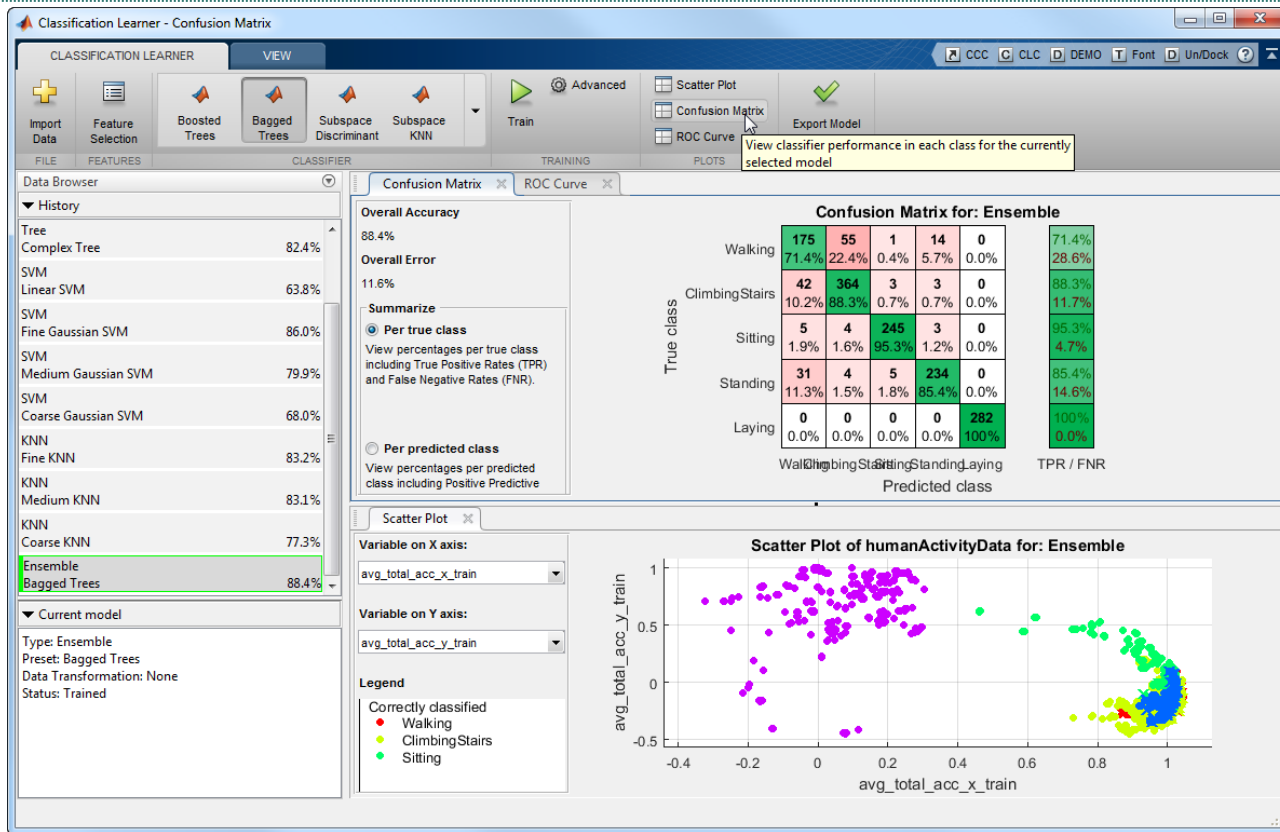
- pokrývají celý proces dobývání znalostí (od předzpracování po interpretaci),
  - nabízejí více algoritmů pro analýzu (než „jednoúčelové“ systémy strojového učení),
  - kladou důraz na vizualizaci (ve způsobu práce se systémem i při interpretaci výsledků).
-

# Systemy pro dobývání znalostí z databází



SLEZSKÁ  
UNIVERZITA  
OBCHODNĚ PODNIKATELSKÁ  
FAKULTA V KARVINĚ





# Rapid Miner



**SLEZSKÁ  
UNIVERZITA**  
OBCHODNĚ PODNIKATELSKÁ  
FAKULTA V KARVINĚ

The screenshot displays the Rapid Miner software interface. The main window shows a workflow diagram with the following operators: Retrieve, Normalization, MissingValue..., LibSVM, Nominal2Bino..., and Nominal2Hum... The workflow starts with an input 'inp' leading to 'Retrieve', which outputs 'exa'. This connects to 'Normalization', which outputs 'exa' and 'tra'. 'tra' goes to 'LibSVM', which outputs 'mod' and 'exa'. 'exa' from 'Normalization' and 'exa' from 'LibSVM' go to 'MissingValue...', which outputs 'exa' and 'pre'. 'exa' from 'MissingValue...' goes to 'Nominal2Bino...', which outputs 'exa' and 'pre'. 'exa' from 'Nominal2Bino...' goes to 'Nominal2Hum...', which outputs 'exa' and 'pre'. The final output is 'res'.

On the left, the 'Samples' tree shows a list of datasets: Golf, Golf-Testset, Iris, Labor-Negotiations, Market-Data, Polynomial, Ripley-Set, Sonar, Transactions, and Weighting. The 'DB' section shows 'LaptopSMCOnderzoek (bokhove)'.

On the right, the 'Parameters' panel for 'LibSVM (Support Vector Machine (LibSVM))' shows the following settings:

- svm type: C-SVC
- kernel type: rbf
- gamma: 22644346174132
- C: 85795083818439
- epsilon: 0.0010
- calculate confidences

A warning message indicates '4 hidden expert parameters'.

At the bottom, the 'Problems' panel shows '2 potential problems':

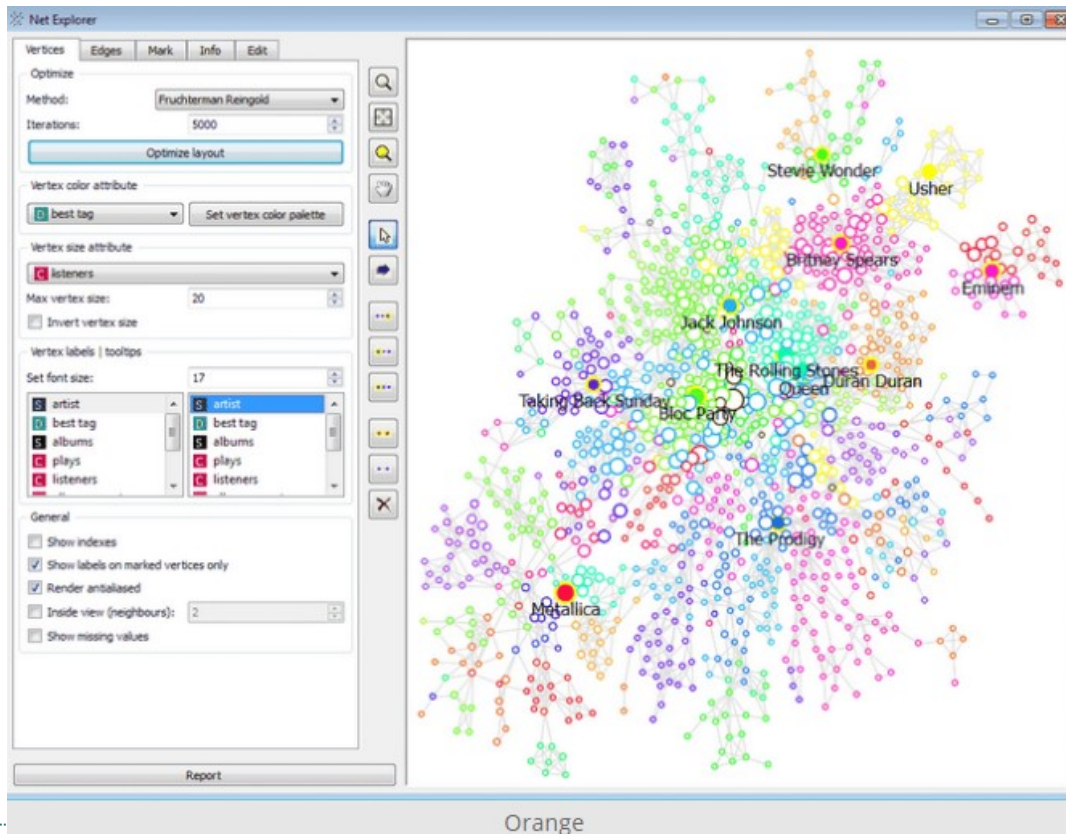
Message	Fixes	Location
Attribute filter does not match any attributes.	Select all attributes.	Nominal2Binomin...
Attribute filter does not match any attributes.	Select all attributes.	Nominal2Numeric...

The 'Support Vector Machine (LibSVM)' synopsis is also visible at the bottom right.

# Python (Orange)



**SLEZSKÁ  
UNIVERZITA**  
OBCHODNĚ PODNIKATELSKÁ  
FAKULTA V KARVINĚ





# Aplikační oblasti pro dobývání znalostí

---



SLEZSKÁ  
UNIVERZITA  
OBCHODNĚ PODNIKATELSKÁ  
FAKULTA V KARVINĚ

- 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

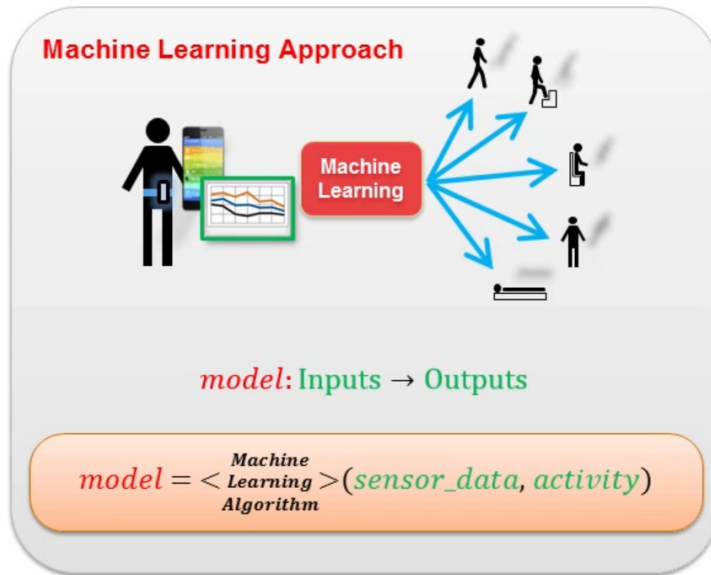
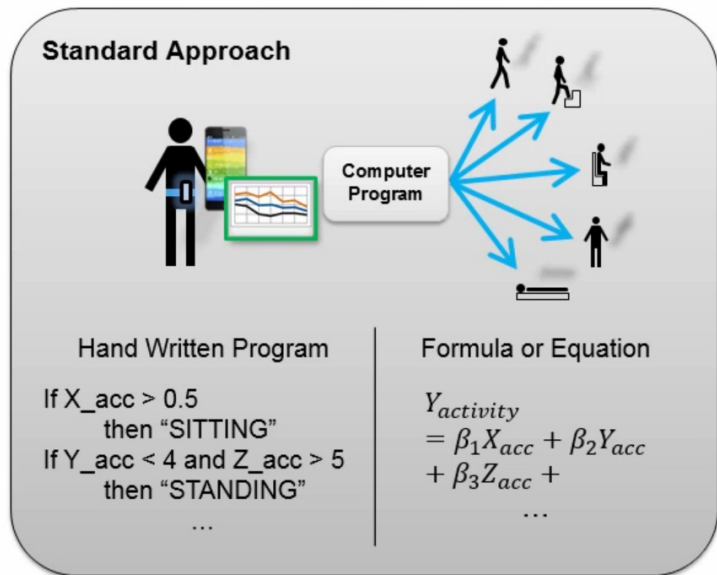


SLEZSKÁ  
UNIVERZITA  
OBCHODNĚ PODNIKATELSKÁ  
FAKULTA V KARVÍNĚ

## Machine Learning

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

**Task:** Human Activity Detection



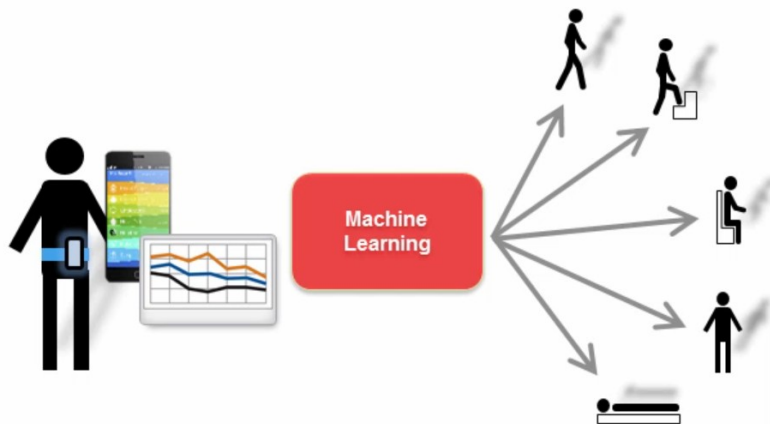


# Rozpoznání činnosti uživatele



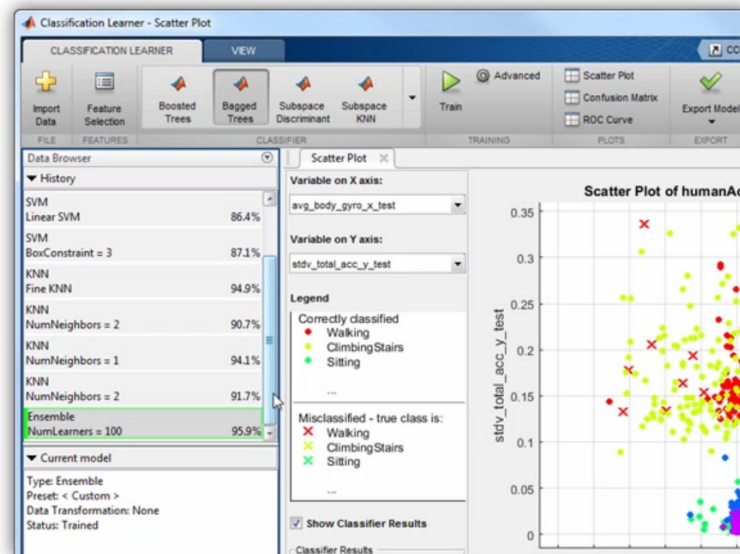
**SLEZSKÁ  
UNIVERZITA**  
OBCHODNĚ PODNIKATELSKÁ  
FAKULTA V KARVÍNĚ

## Example: Human Activity Learning Using Mobile Phone Data



Data:

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



# Děkuji za pozornost

Některé snímky převzaty od:  
prof. Ing. Petr Berka, CSc. [berka@vse.cz](mailto:berka@vse.cz)