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

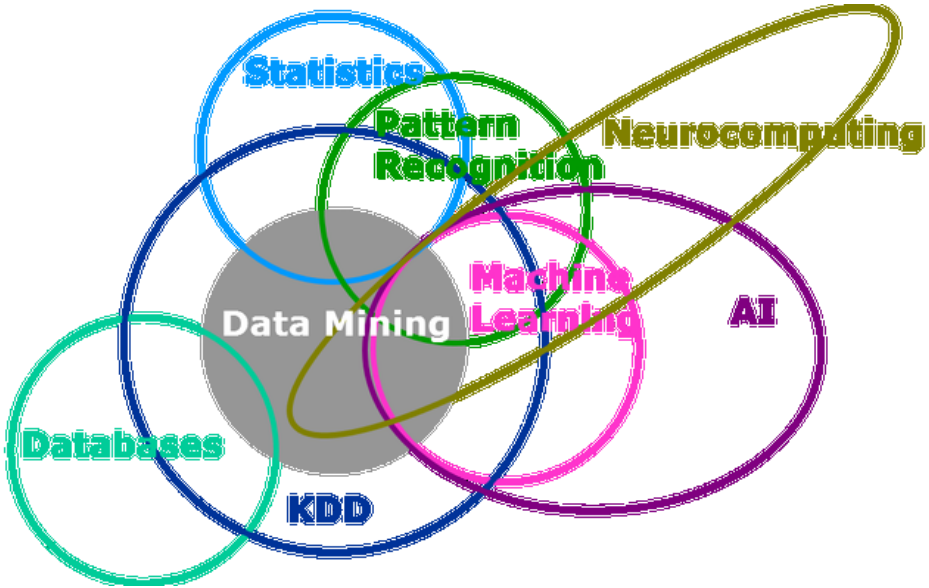
Jan Górecki

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:

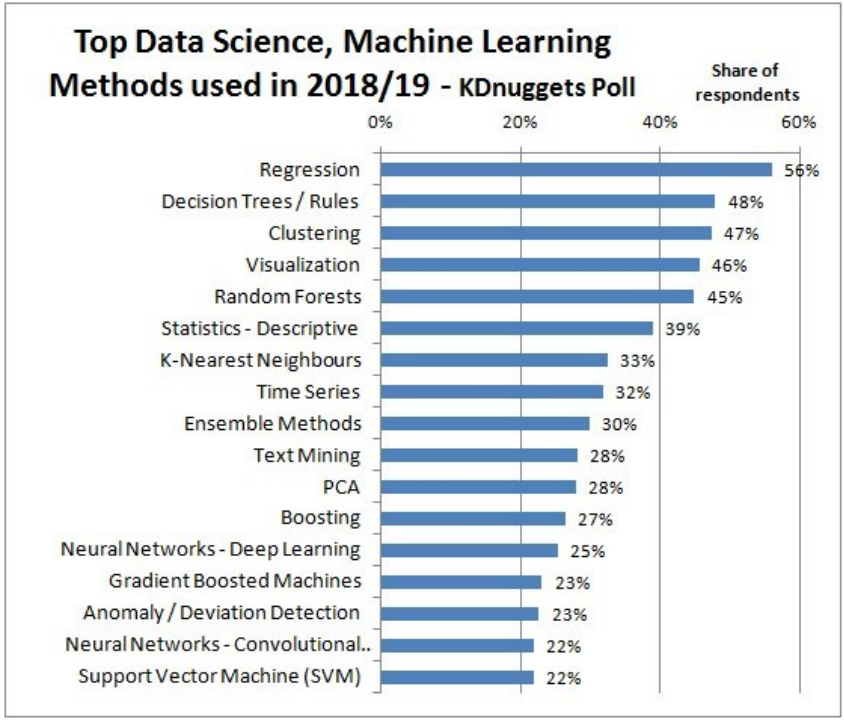
- **Prezenční forma: 13 přednášek a 12 seminářů,**
- **Kombinovaná forma: 3 přednášky**

- **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 – prezenční forma

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) – odevzdání přes odevzdávárnu v IS SU do 20.12.2023 23:55**
- **zkouška (60% hodnocení)**

Celkem maximum: 100

Požadované minimum: 60

Požadavky na absolvování předmětu – kombinovaná forma

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

- **docházka se nevyžaduje (ale je hodnocena až 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) – odevzdání přes odevzdávárnu v IS SU do 20.12.2023 23:55**
- **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
- **Kaggle: Your Home for Data Science**
  https://www.kaggle.com
- **KEEL - dataset repository**
  http://www.keel.es a tam KEEL-dataset

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

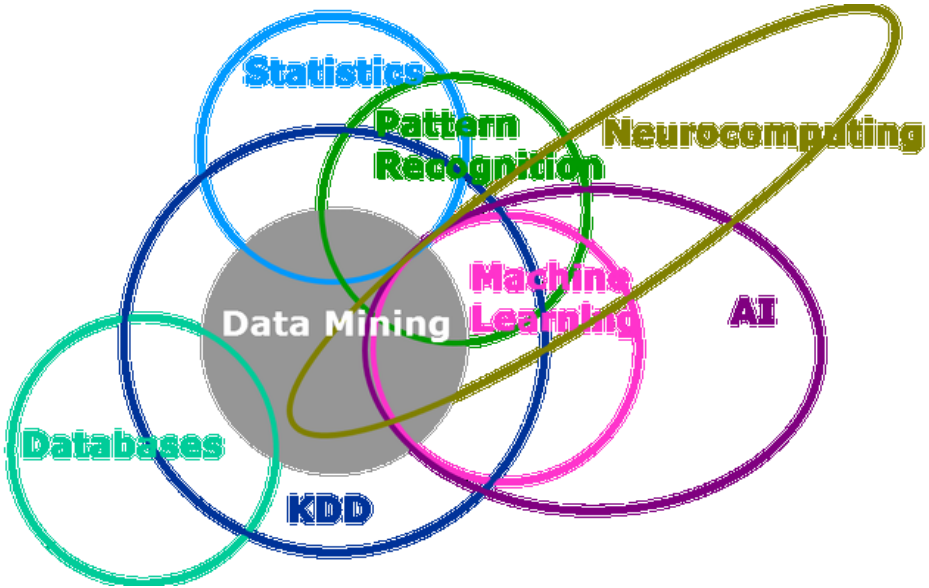
- Definice …
- Historie …
- Úlohy …
- Příklad …
- Postupy (metodiky) …
- Software pro …

} … Dolování dat

Dolování dat (Data mining)

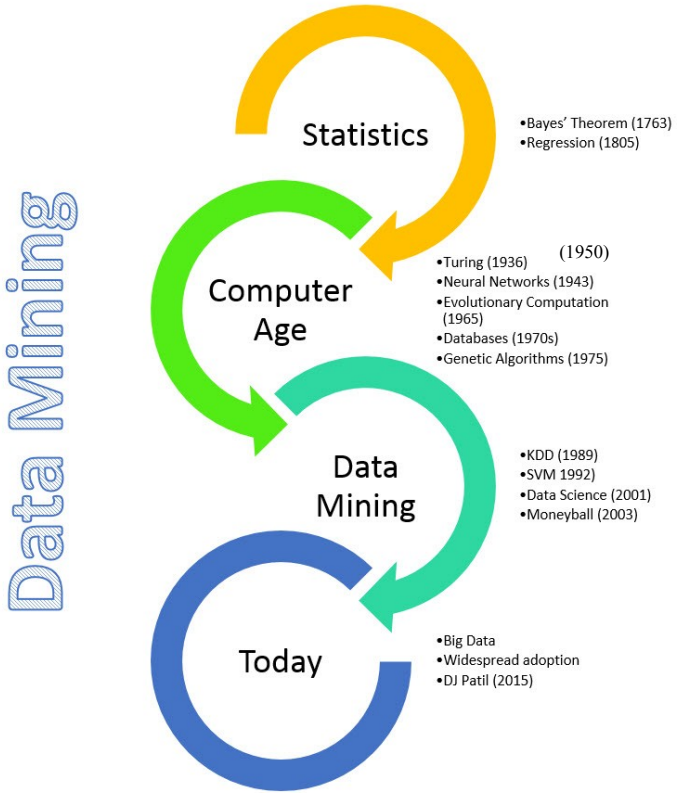
SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

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



Trocha historie

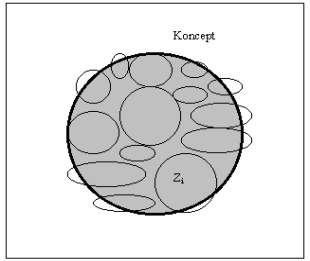
SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

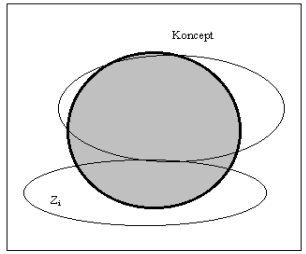


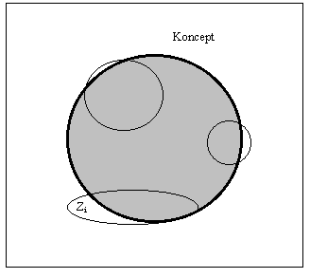
Úlohy dobývání znalostí

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

- **Klasifikace/Predikce**
  - **Cíl:** Nalézt znalosti použitelné pro klasifikaci nových případů.
  - **Příklad:** Predikce akciových cen, klasifikace e-mailů jako spam nebo ne-spam.
- **Deskripce**
  - **Cíl:** Nalézt dominantní strukturu nebo vazby skryté v datech. Dáváme přednost menšímu počtu méně přesných avšak srozumitelných pravidel.
  - **Příklad:** Analýza sociálních médií odhalila, že pozitivní recenze na film často korelují s vyššími prodeji vstupenek v prvním týdnu promítání.
- **Hledání „Nugetů“**
  - **Cíl:** Nalézt dílčí překvapivé (vzácné, cenné) znalosti.
  - **Příklad:** Odhalení neobvyklých nákupních vzorců v datech o prodeji.







Detekce Spamu - Shromažďování a Předzpracování Dat

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

1. Shromažďování Trénovacích Dat:

- **Příklad:** Sbíráme 1000 e-mailů, z nichž 500 je označeno jako spam a 500 jako ne-spam. Data jsou anonymizována, aby byly odstraněny všechny osobní informace.

2. Předzpracování Dat:

- **Příklad:** E-maily jsou pročištěny od nepotřebných dat, jako jsou hlavičky e-mailů, a text je převeden na malá písmena. "KUP TEĎ!!!" se stane "kup teď".

Detekce Spamu - Analýza Dat (metoda Naivní Bayes)

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

3. Analýza Dat s Naivním Bayesem:

- **Základní Idea:** Naivní Bayes je statistický klasifikační model založený na Bayesově teorému. Je "naivní", protože předpokládá nezávislost mezi jednotlivými slovy (nebo rysy) ve zprávě, což ve skutečnosti nemusí být vždy pravda.

Příklad Použití Naivního Bayese:

- **Trénovací Data:** Máme trénovací data, kde jsou e-maily již označeny jako spam nebo ne-spam. Model se učí z těchto dat, jak rozpoznat charakteristiky spamových a ne-spamových e-mailů.
- **Výpočet Pravděpodobnosti:** Model vypočítá pravděpodobnost, že daný e-mail je spam nebo ne, na základě frekvence slov v e-mailu. Například, e-mail obsahující slova jako "sleva", "klikněte" a "zdarma" může mít vyšší pravděpodobnost být označen jako spam.
- **Klasifikace:** E-mail je klasifikován jako spam nebo ne-spam na základě vypočítané pravděpodobnosti.

Detekce Spamu - Aplikace Modelu a Výsledek

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

4. Výsledek Analýzy:

- **Příklad:** Model je nyní schopen s přesností 95% identifikovat, zda je nový e-mail spam nebo ne, na základě jeho obsahu a charakteristik.

5. Aplikace Modelu:

- **Příklad:** Model je integrován do e-mailového systému. Když přijde nový e-mail obsahující "kup teď", je automaticky přesunut do složky se spamem.

6. Aktualizace Modelu:

- **Příklad:** Model se pravidelně aktualizuje s novými daty, aby se zlepšila jeho přesnost. Pokud některé spamové e-maily projdou, uživatelé je mohou manuálně označit jako spam, což pomáhá modelu se učit a zlepšovat.

7. Výsledek:

- **Příklad:** Uživatelé vidí výrazné snížení spamu v jejich doručené poště, což jim umožňuje se soustředit na důležité e-maily a zvyšuje jejich produktivitu.

Aplikační oblasti pro dobývání znalostí

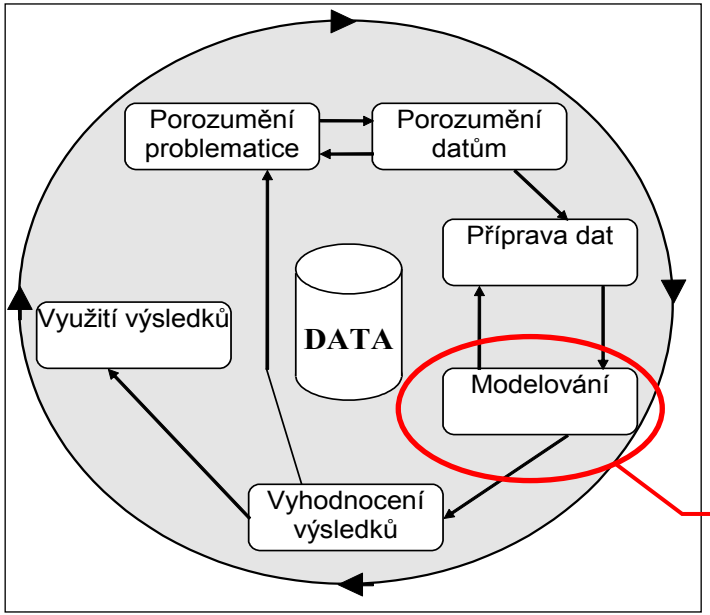
SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

- 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,

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

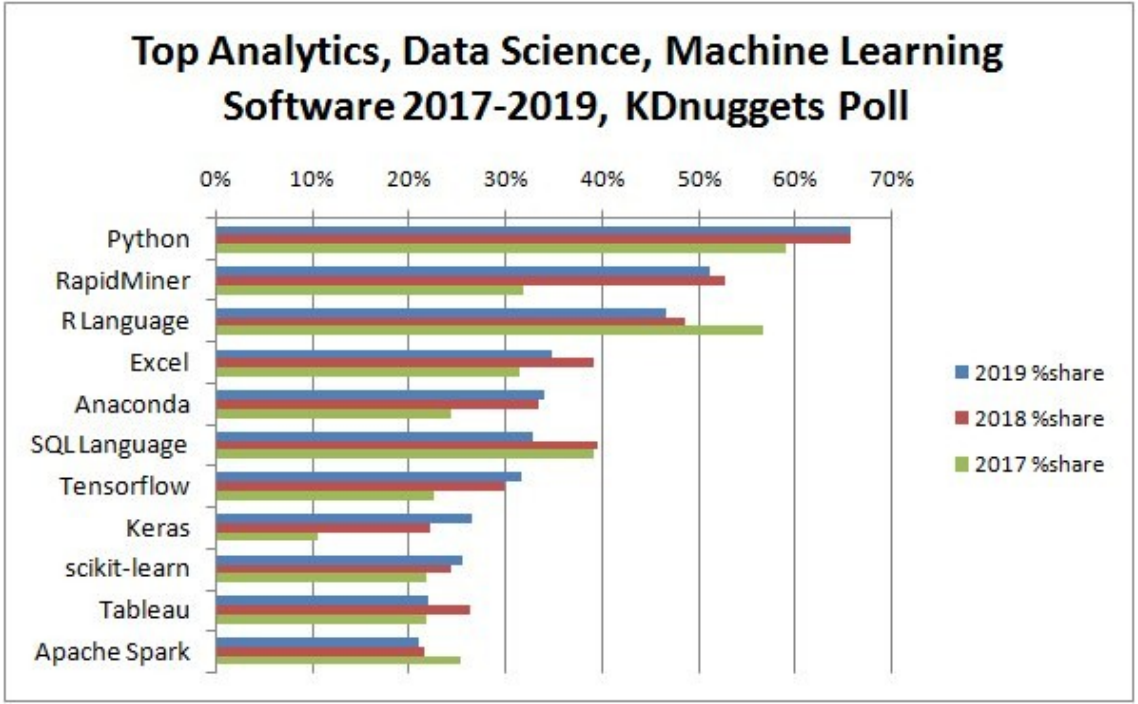
Gartner peerinsights customers' choice 2019

MathWorks

ANACONDA.

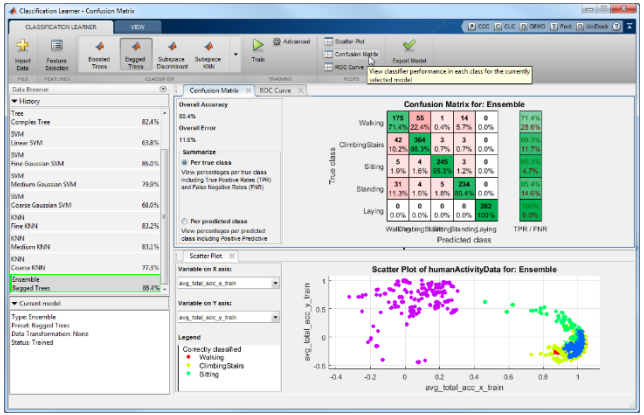
rapidminer

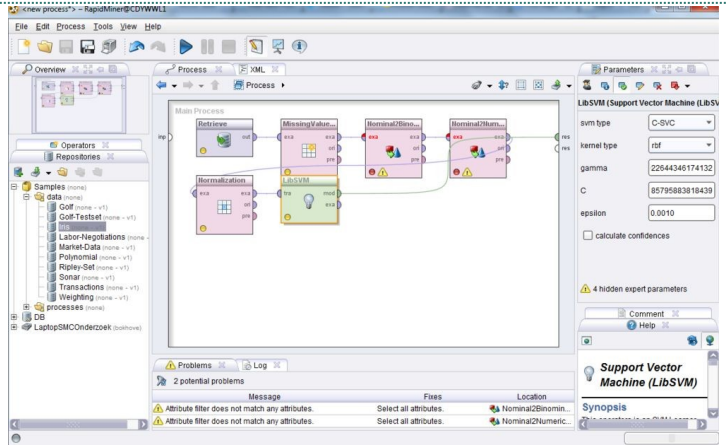
alteryx The Thrill of Solving

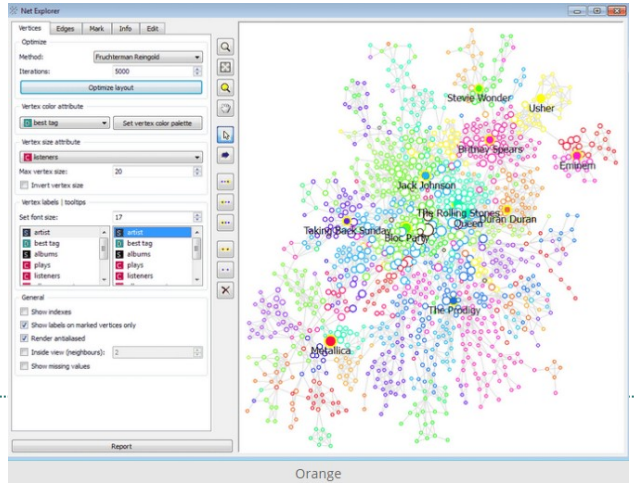


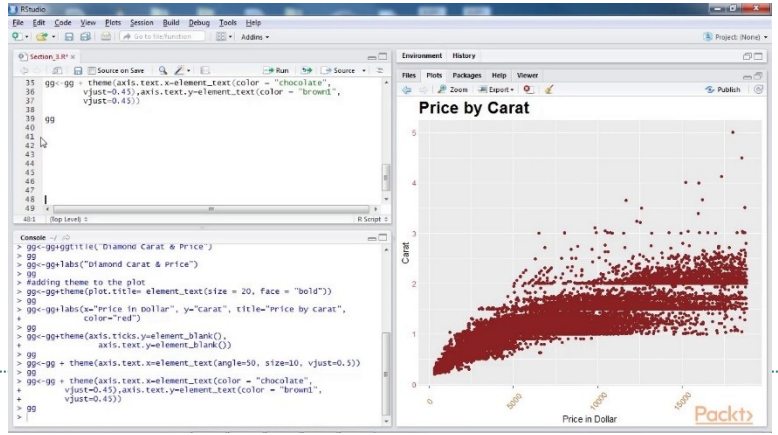
MATLAB, Rapid Miner, Python, R

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ









Děkuji za pozornost

Některé snímky převzaty od:

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

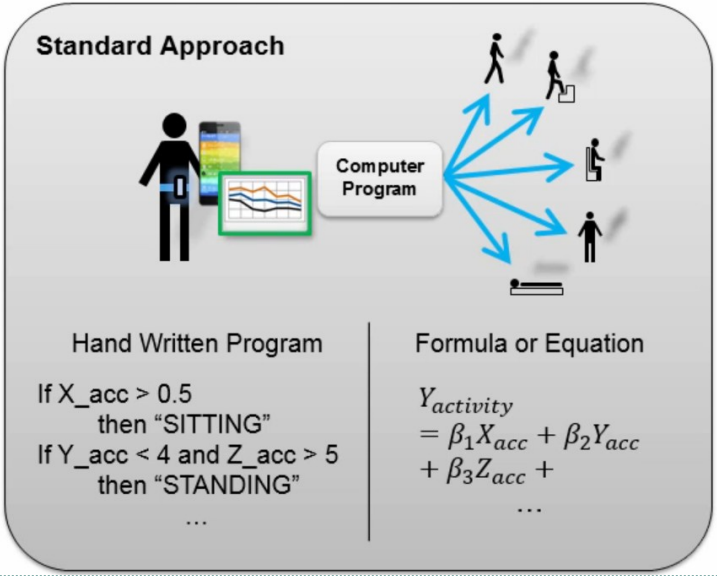
Rozpoznání činnosti uživatele

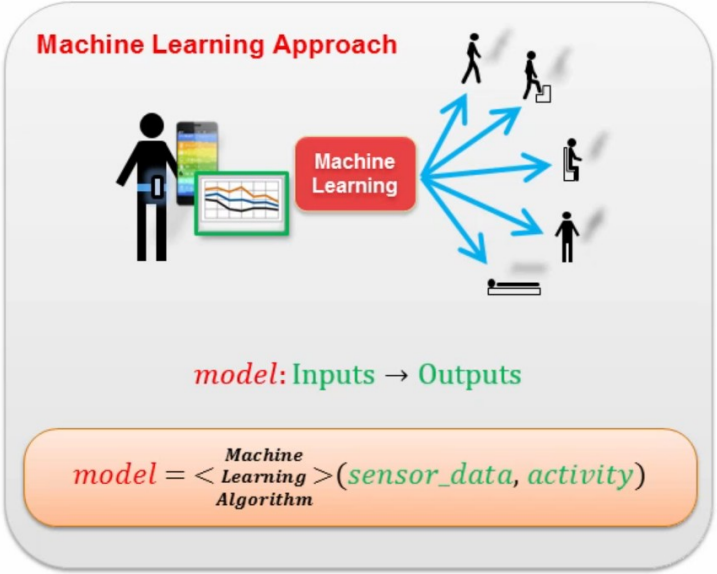
SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

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