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

# Příprava dat


Jan Górecki


SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

# Obsah přednášky

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

- Co to je příprava dat
- Příklady strukturovaných dat
- Odvozené atributy
- Příliš mnoho objektů
- Příliš mnoho atributů
- Numerické atributy
- Kategoriální atributy
- Chybějící hodnoty

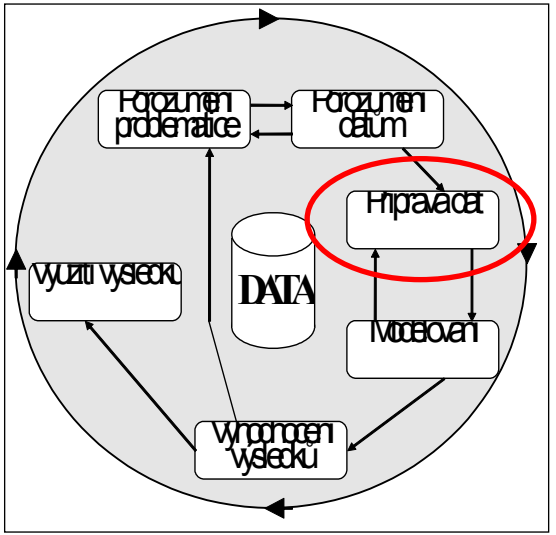
# Příprava dat

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

Cíl:

1. vybrat (nebo vytvořit) z dostupných dat ty údaje, které jsou relevantní pro zvolenou úlohu dobývání znalostí,
2. reprezentovat tyto údaje v podobě, která je vhodná pro zpracování zvoleným algoritmem.

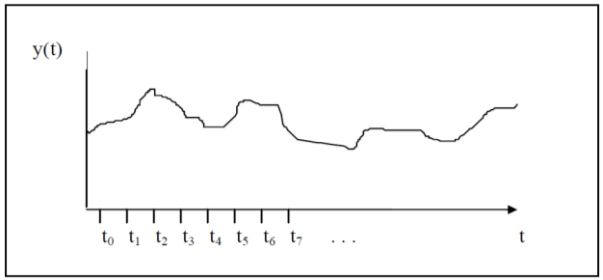
CRISP-DM



# Strukturovaná data

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

- časová data (např. časové řady kurzů akcií)




Původní časová řada

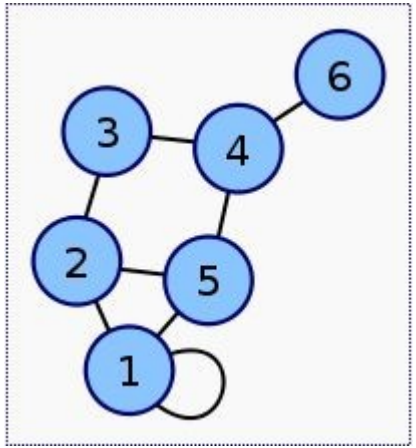
| vstupy | | | | výstup |
|---|---|---|---|---|
| $y(t_0)$ | $y(t_1)$ | $y(t_2)$ | $y(t_3)$ | $y(t_4)$ |
| $y(t_1)$ | $y(t_2)$ | $y(t_3)$ | $y(t_4)$ | $y(t_5)$ |
| $y(t_2)$ | $y(t_3)$ | $y(t_4)$ | $y(t_5)$ | $y(t_6)$ |
| . . . | | | | |

Časová řada po transformaci

# Strukturovaná data

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

- grafy
  - graf reprezentován seznamem hran spojujících dva uzly (adjacency matrix - matice sousednosti)




$$\begin{pmatrix} 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

# Strukturovaná data

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

- texty
  - text reprezentován seznamem slov v dokumentu (bag-of-words, vektorový model)

| | I | love | dogs | hate | and | knitting | is | my | hobby | passion |
|---|---|---|---|---|---|---|---|---|---|---|
| Doc 1 | 1 | 1 | 1 | | | | | | | |
| Doc 2 | 1 | | 1 | 1 | 1 | 1 | | | | |
| Doc 3 | | | | | 1 | 1 | 1 | 2 | 1 | 1 |

| | are | call | from | hello | home | how | me | money | now | tomorrow | win | you |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| **1** | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 2 | 0 |
| **2** | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| **3** | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |

# Odvozené atributy

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

- Z atributů přítomných v původních datech lze vytvářet odvozené atributy
- Plyne z doménových znalostí
  Např.:
  1) převod rodného čísla klienta na věk a pohlaví

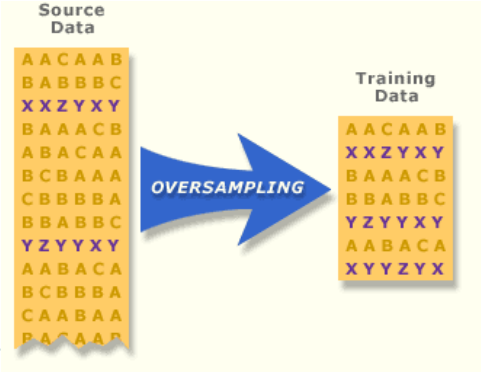
| Rodné číslo | Rok narození | Měsíc narození | Den narození | Pohlaví |
|---|---|---|---|---|
| 870412/xxxx | 1987 | 4 | 12 | muž |
| 035708/xxxx | 2003 | 7 | 8 | žena |
| 096224/xxxx | 2009 | 12 | 24 | žena |

  2) převod data na den v týdnu – *14.11.2019 -> čtvrtek*
- Opět se tedy jedná o operaci, kdy je třeba úzce spolupracovat s expertem.

# Příliš mnoho objektů

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

1. použít jen určitý vzorek (sample) vybraný z celých dat,
   - náhodný výběr
   - stratifikovaný výběr – ve vzorku je stejné rozdělení příkladů do tříd jako v celých datech
   - „oversampling“ – ve vzorku se preferují příklady minoritní třídy




SAS EM demo

# Příliš mnoho objektů



2. vytvořit více modelů na základě podmnožin objektů a modely poté zkombinovat
3. rozdělit data do podmnožin a z každé podmnožiny vybrat pár ji reprezentujících znaků (průměr, směrodatná odchylka, atd.

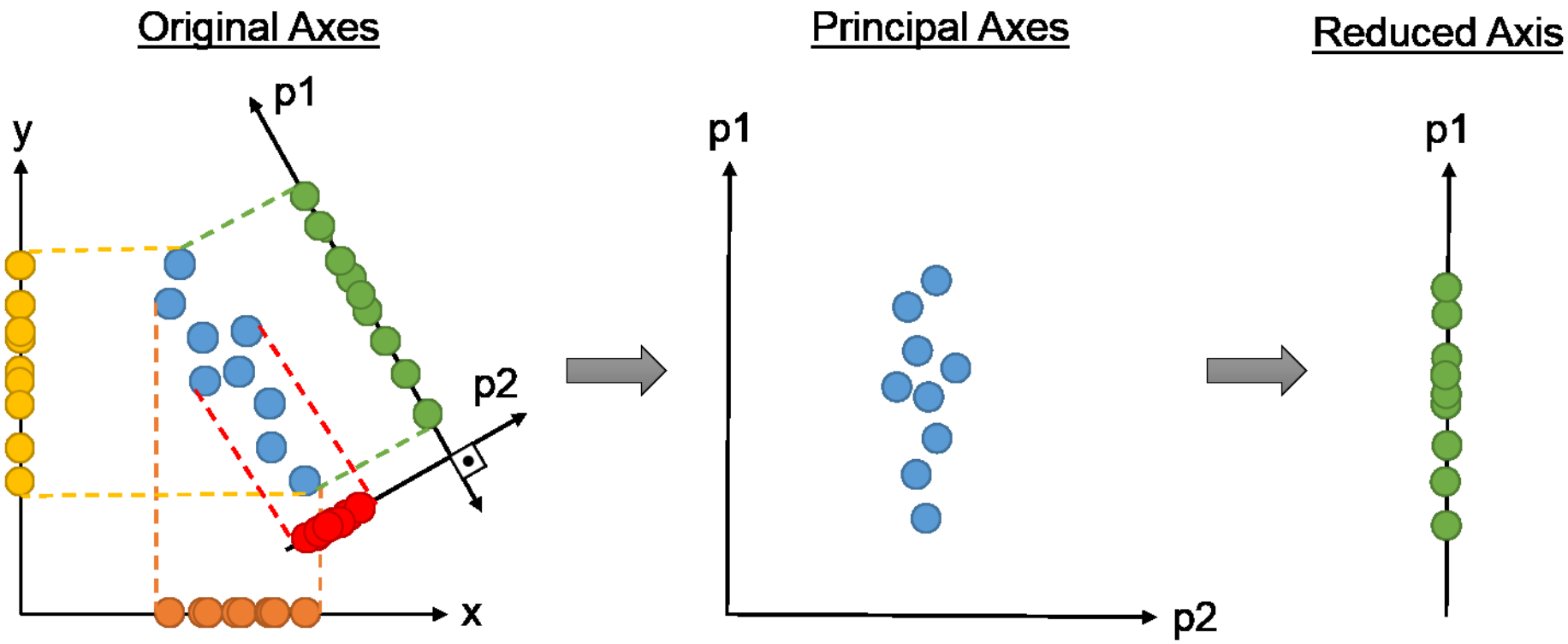
# Příliš mnoho atributů

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

- ***transformace*** - z existujících atributů vytvoříme menší počet atributů nových (analýza hlavních komponent),
- ***selekce*** - z existujících atributů vybereme jen ty nejdůležitější

# Analýza hlavních komponent (*Principal Component Analysis*, PCA)

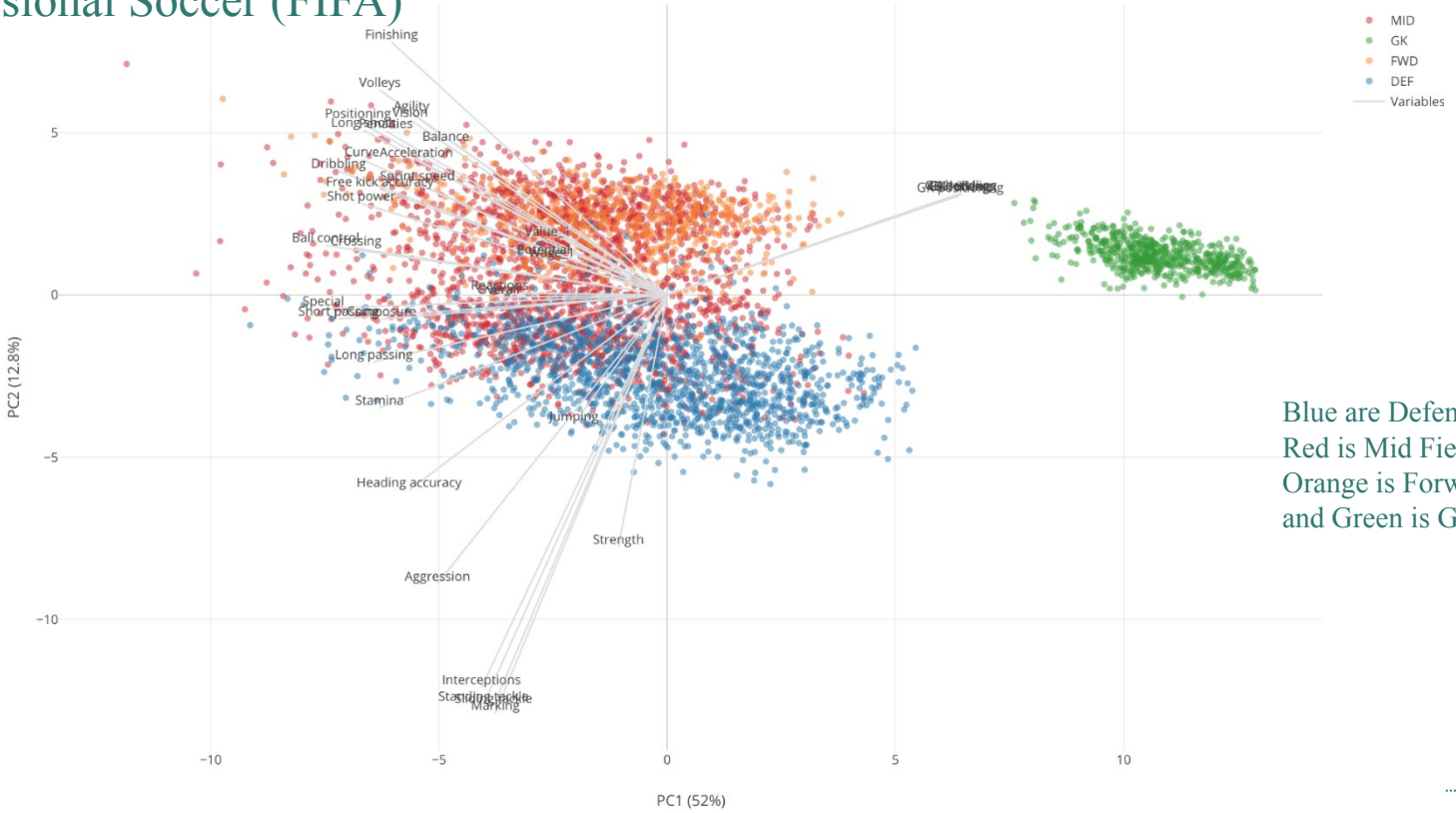
SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ



# Analýza hlavních komponent (*Principal Component Analysis*, PCA)

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

## Professional Soccer (FIFA)




Blue are Defense players (DEF),
Red is Mid Field players (MID),
Orange is Forward players (FWD),
and Green is Goal Keepers (GK)

# Automatická selekce

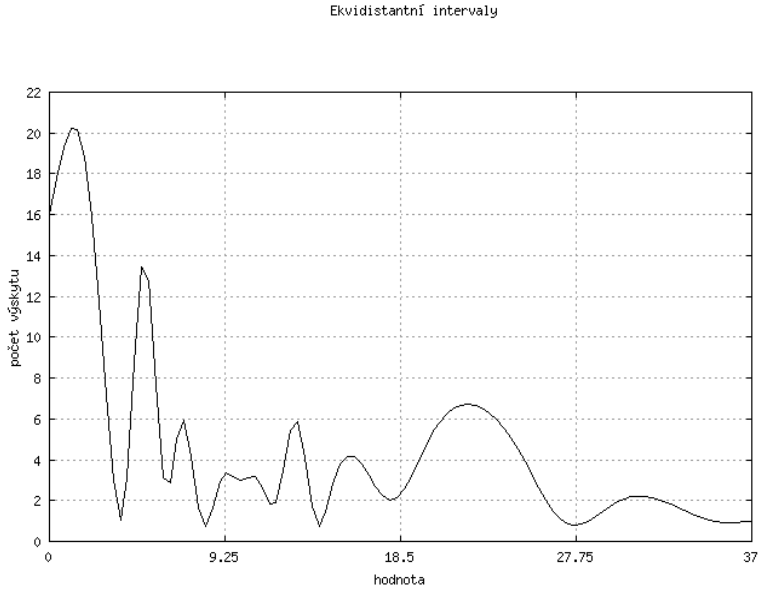
SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

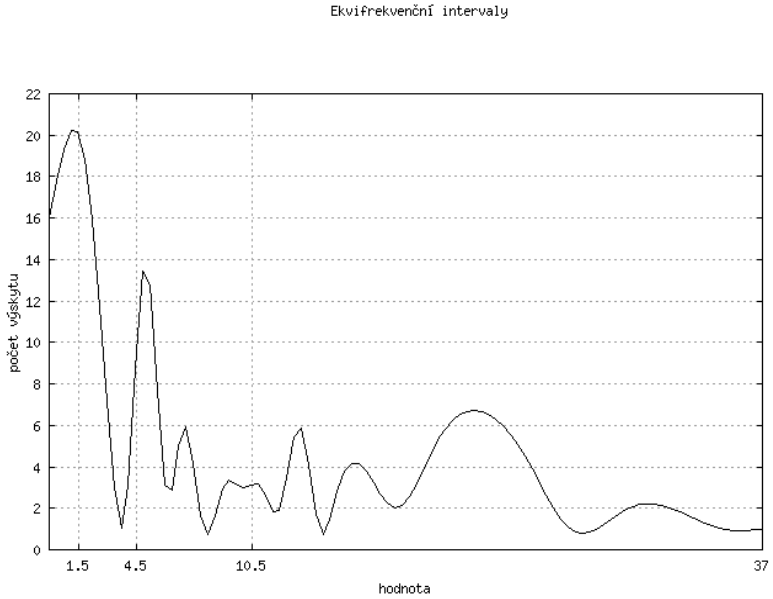
- nalezení atributů, které nejlépe přispějí ke klasifikaci objektů do tříd
  - ***metoda filtru*** – filter approach
  - ***metoda obálky*** – wrapper approach

# Numerické atributy

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

Cena akcie: z reálného čísla do množiny kategorií (velmi nízká, nízká, vysoká, velmi vysoká)



# Kategoriální atributy

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

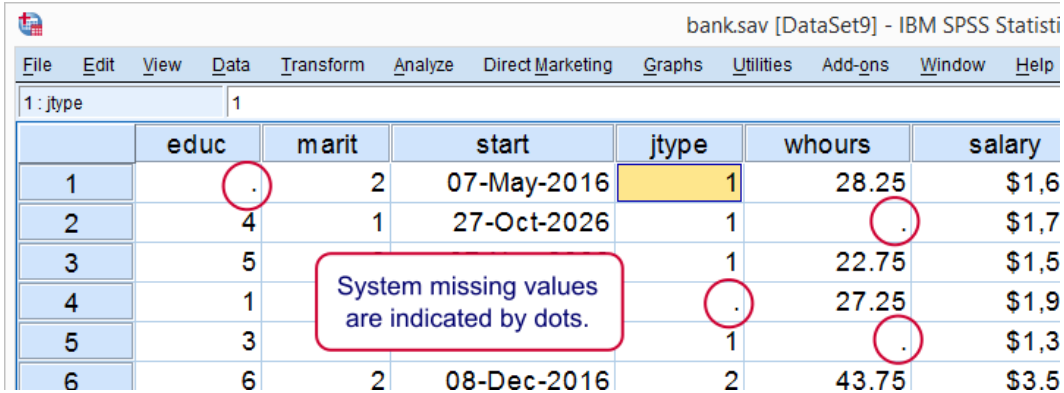
- kódování čísly (+ binarizace)

- Chceme-li použít algoritmy, které používají numerické vstupy (např. neuronové sítě), musíme hodnoty kategoriálních atributů zakódovat pomocí čísel.
  - V případě binárních atributů (např. ano, ne) to mohou být např. hodnoty 0, 1 nebo –1, 1.
  - V případě ordinálních atributů (např. malý, střední, velký) to může být pořadové číslo hodnoty.
  - V případě nominálních atributů (Čech, Slovák, Němec) je potřeba vytvořit tolik aktributů, kolik je kategorií, a pak binarizovat, např.
    národnost = Slovák ->
    národnost_Čech = 0, národnost_Slovák = 1, národnost_Němec = 0,

# Ošetření chybějících hodnot

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

1) ignorovat objekt s nějakou chybějící hodnotou,
2) nahradit chybějící hodnotu novou hodnotou „nevím“,
3) nahradit chybějící hodnotu některou z existujících hodnot atributu a sice:
   a) nejčetnější hodnotou,
   b) proporcionálním podílem všech hodnot,
   c) libovolnou hodnotou,
   d) „predikovanou“ hodnotou




| | educ | marit | start | jtype | whours | salary |
|---|---|---|---|---|---|---|
| 1 | . | 2 | 07-May-2016 | 1 | 28.25 | $1,6 |
| 2 | 4 | 1 | 27-Oct-2026 | 1 | . | $1,7 |
| 3 | 5 | | | 1 | 22.75 | $1,5 |
| 4 | 1 | | | . | 27.25 | $1,9 |
| 5 | 3 | | | 1 | . | $1,3 |
| 6 | 6 | 2 | 08-Dec-2016 | 2 | 43.75 | $3.5 |

# Děkuji za pozornost

Některé snímky převzaty od:

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