

# Data Mining

//



## M9DM2 Data mining II

- S novou akreditací změna rozsahu na 2hodinový seminář a ukončení kolokviem.
- Pokračování předmětu Data mining 1 s důrazem na praktické použití metod.
- Prohloubení znalostí z kurzu Data mining I.
- Vybrané semináře vedené odborníky z praxe.

# M9DM2 Data mining II – plán seminářů

- 17.9. a 24.9. **Diskriminační analýza** (Navrátil)
- 1.10. a 8.10. **Organizace dat. tabulek, SQL** (Pokora)
- 15.10., 22.10. a 29.10. **Text mining** (Buček)
- 5.11. a 12.11. **Business intelligence** (Nagy)
- 19.11. a 26.11. **Čištění a příprava dat** (Zlámal)
- 3.12. a 10.12. (Selingerová)
- 17.12. **Prezentace projektů**

# M9DM2 Data mining II – kolokvium

- Aktivní účast na seminářích (maximálně 2 neomluvené absence).
- Vypracování domácích úkolů během semestru.
- Prezentace studentského projektu na závěrečném semináři.

# M9DM2 Data mining II – projekt

- Utvořte maximálně tříčlenný tým, vyberte si vhodný datový soubor a položte otázky, na které se budete snažit odpovědět.
- O této skutečnosti nás informujte e-mailem - uveďte, prosím, složení týmu, název projektu a jednu až dvě věty, co budete dělat.
- Proveďte vlastní analýzu (v libovolném softwaru).
- Připravte krátkou prezentaci (cca 10 min.).
- V prezentaci publikum seznámte s vaším problémem, jak jste jej řešili a na co jste přišli.
- Rozhodně není nutné popisovat použité metody a jiné technické záležitosti, zaměřte se hlavně na výsledky a jejich interpretaci.
- Pro zpříjemnění předvánoční atmosféry můžete přinést cukroví či jiné pohoštění.

# Data Mining



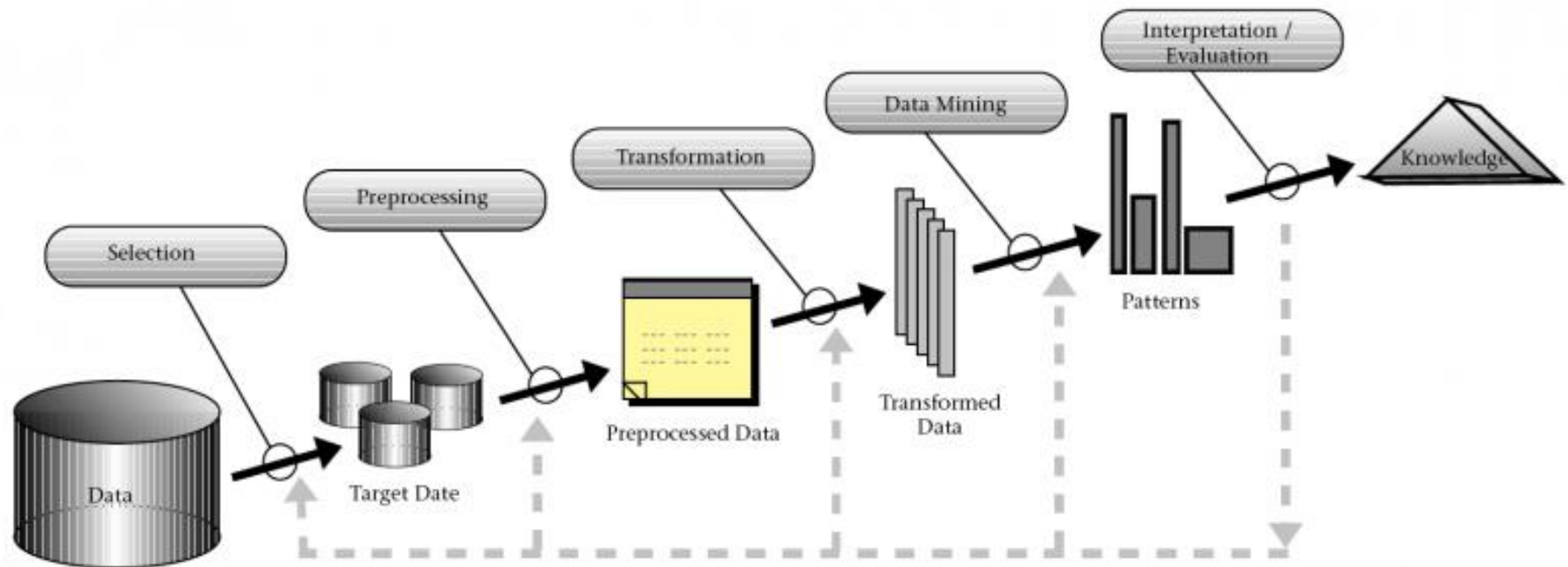
Kde jsme skončili?

# Co je to data mining?

**Data mining** je analytická metodologie získávání netriviálních skrytých a **potenciálně užitečných informací** z dat.

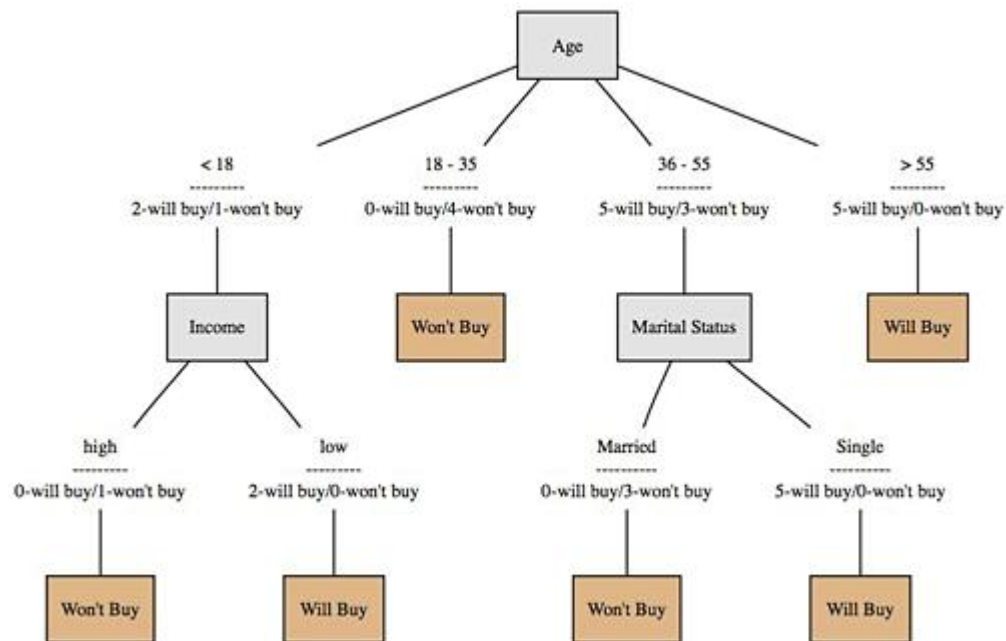
*[wikipedia]*

# Data mining



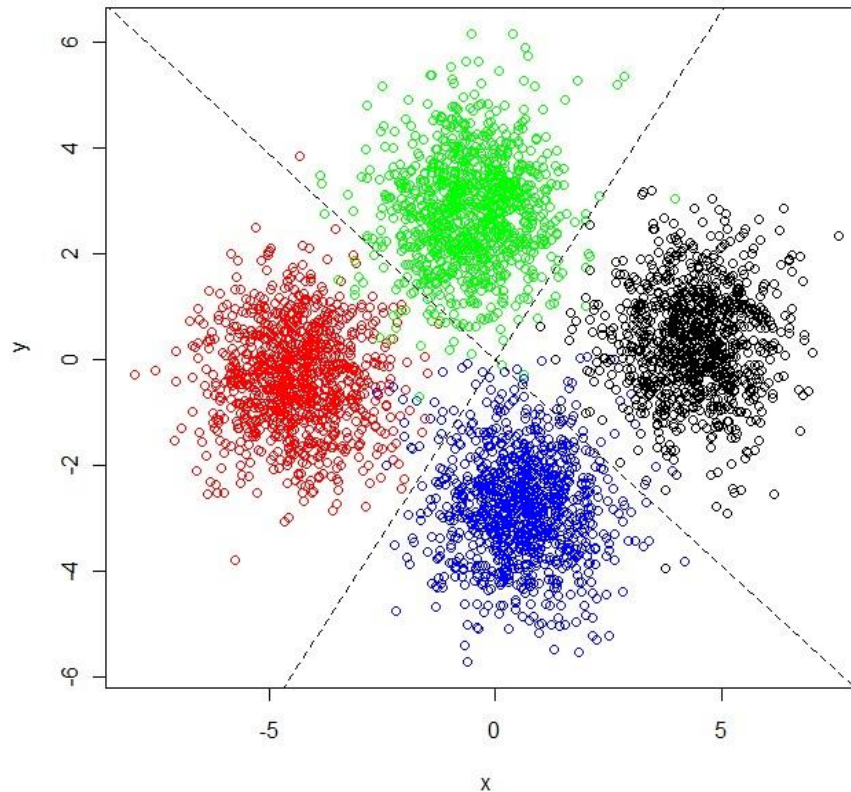


# Rozhodovací stromy



# Diskriminační analýza

Linear Discriminant Analysis (LDA)



# Data a cíl

$(x_{i1}, \dots, x_{ip}, Y_i), i=1, \dots, n$

$x_i$  je vektor **spojitých** prediktorů

$Y_i$  udává příslušnost pozorování k dané skupině (kategoriální proměnná) – skupiny označme  $1, 2, \dots, J$ .

Úkol: Na základě dat zkonstruovat rozhodovací pravidlo, které bude co nejlépe klasifikovat nová pozorování  $(x_{i1}^*, \dots, x_{ip}^*)$  do příslušné skupiny.

# Úvod a historie

- Úloha supervised learning (učení s učitelem).
- Zakladatel: R. A. Fisher (1936) – klasifikace kosatců (iris).
- Způsoby odvození klasifikačního pravidla:
  - Kanonická diskriminační analýza
  - Pravděpodobnostní modely:
    - Parametrické metody (LDA, QDA)
    - Neparametrické metody (k-nearest neighbors, metody založené na jádrových odhadech hustoty, na hloubce, apod.)

# Kritéria pro hodnocení kvality modelu

- Matice záměn.
- Správnost klasifikace, chyba klasifikace.
- Správnost klasifikace do dané třídy, chyba klasifikace do dané třídy.
- Specificita a senzitivita.
- ROC – křivky.