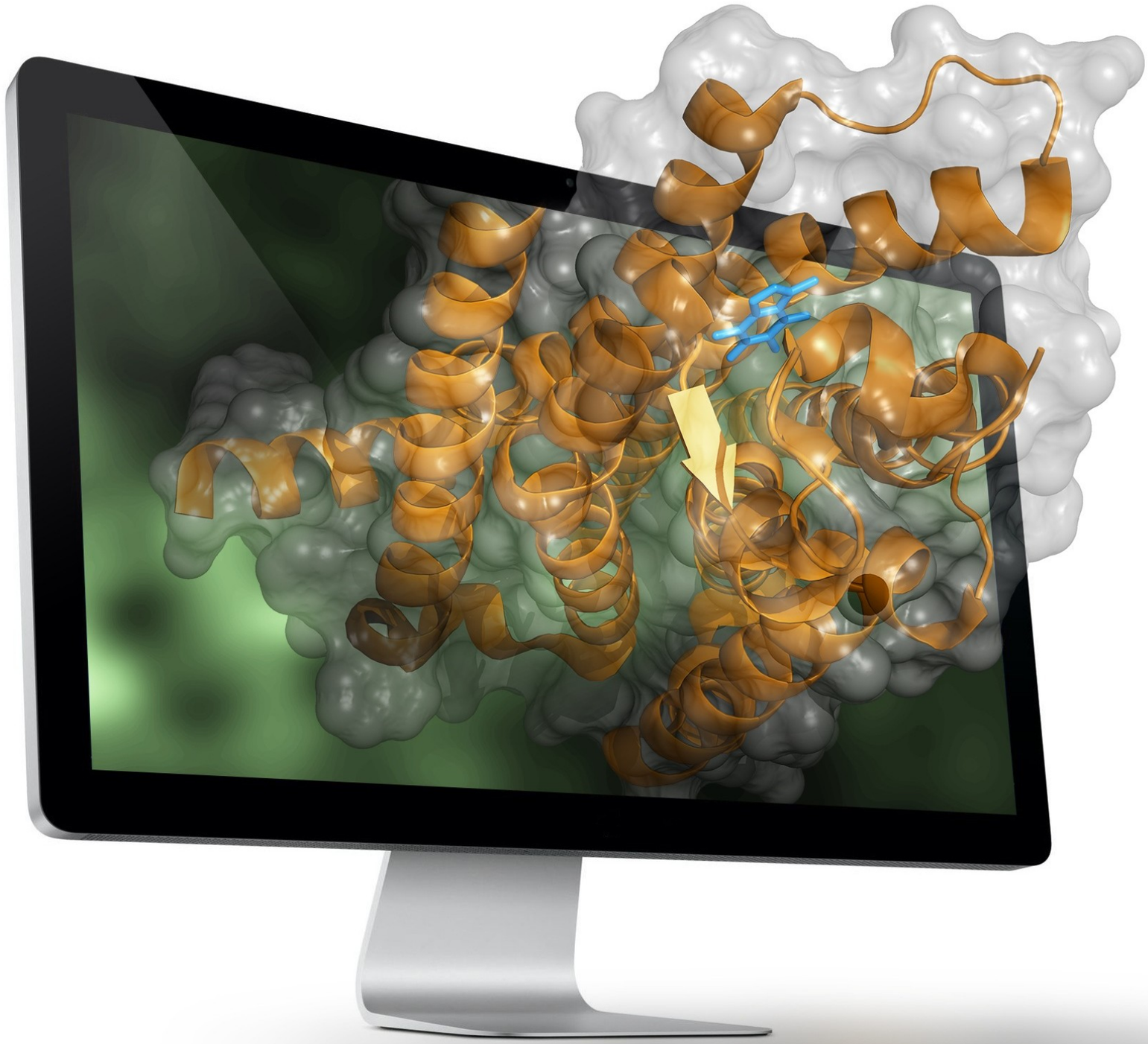


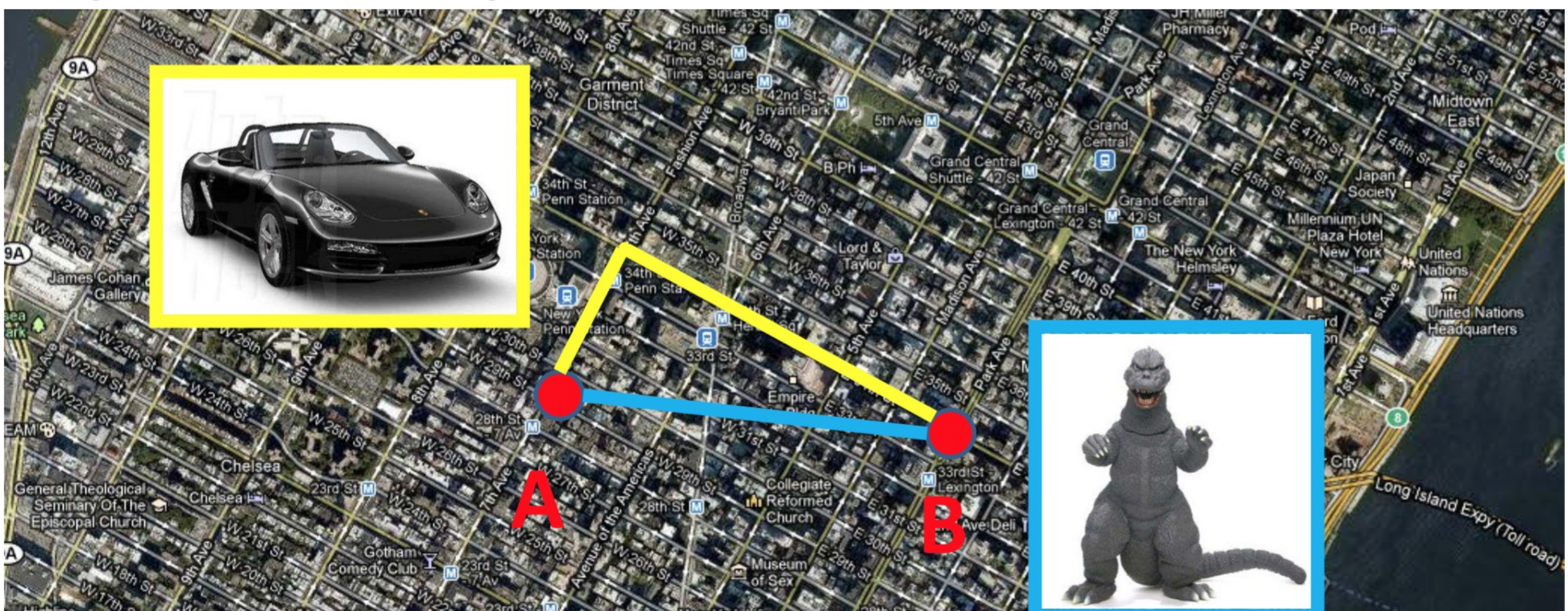
Pokročilá chemoinformatika

Neparametrické modely
únor 2017

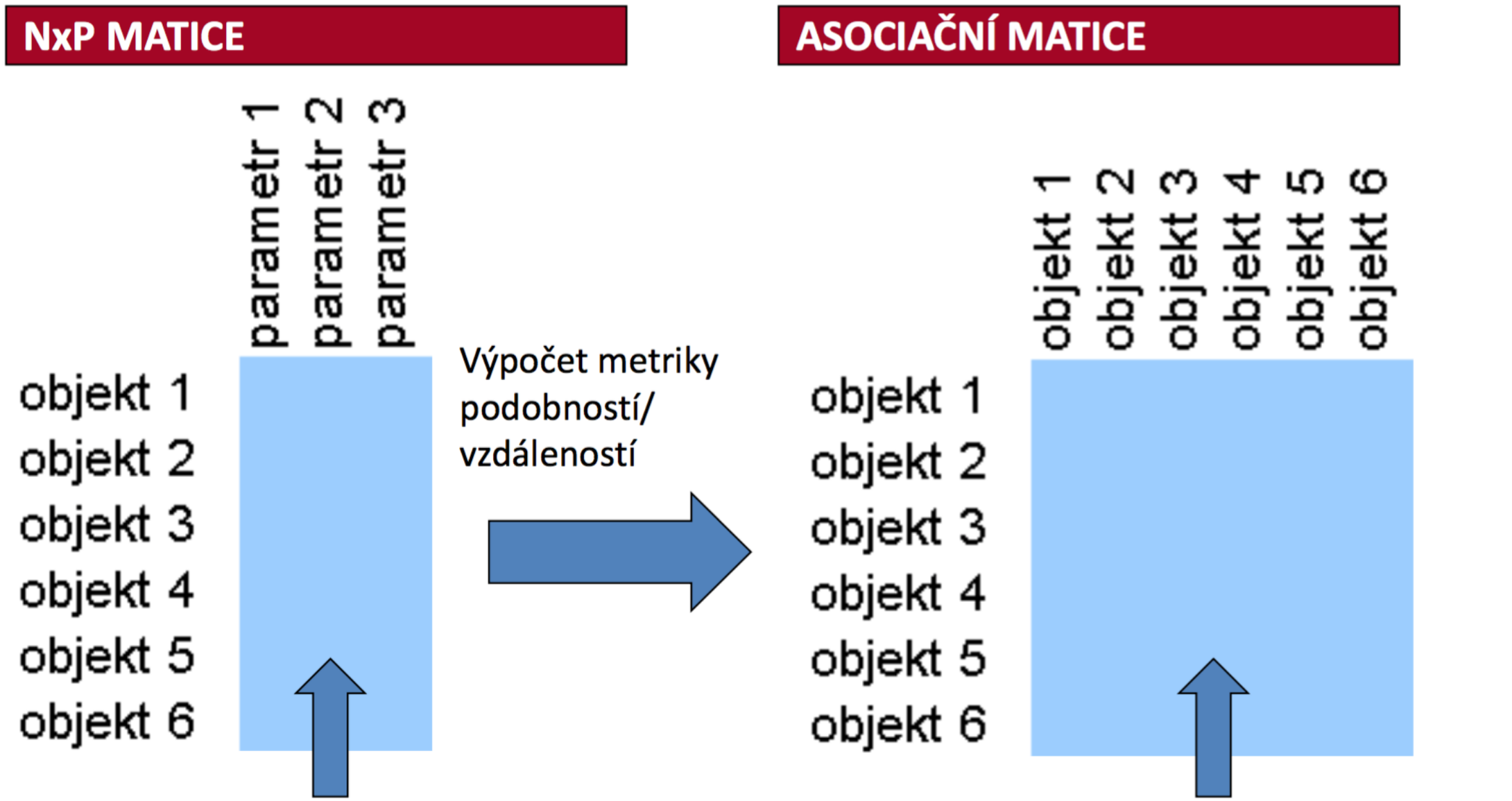
Vzdálenost

- Každá molekula je definována v N-dimenzionálním prostoru (N je počet deskriptorů).

- Mezi každou molekulou můžeme spočítat vzdálenost a pro celou sadu pak asociační matici.



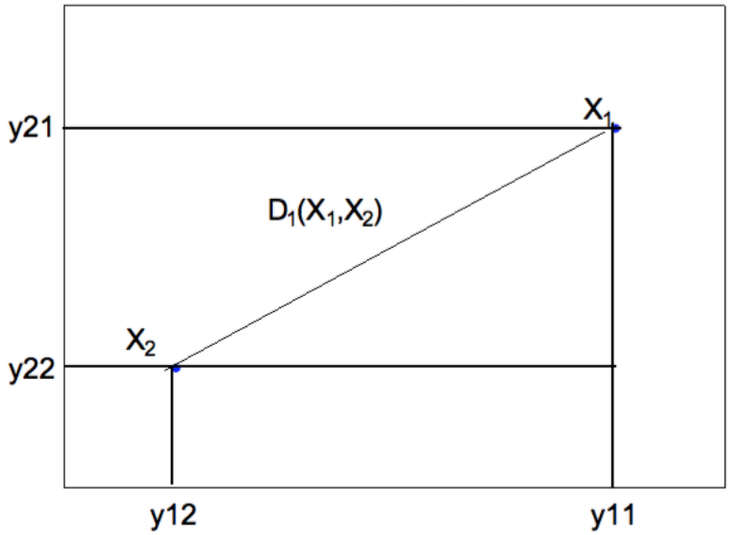
Asociační matice



Základní vzdálenosti

- Euklidovská vzdálenost

$$D_1^2(x_1, x_2) = \sum_{j=1}^{p} (y_{1j} - y_{2j})^2$$



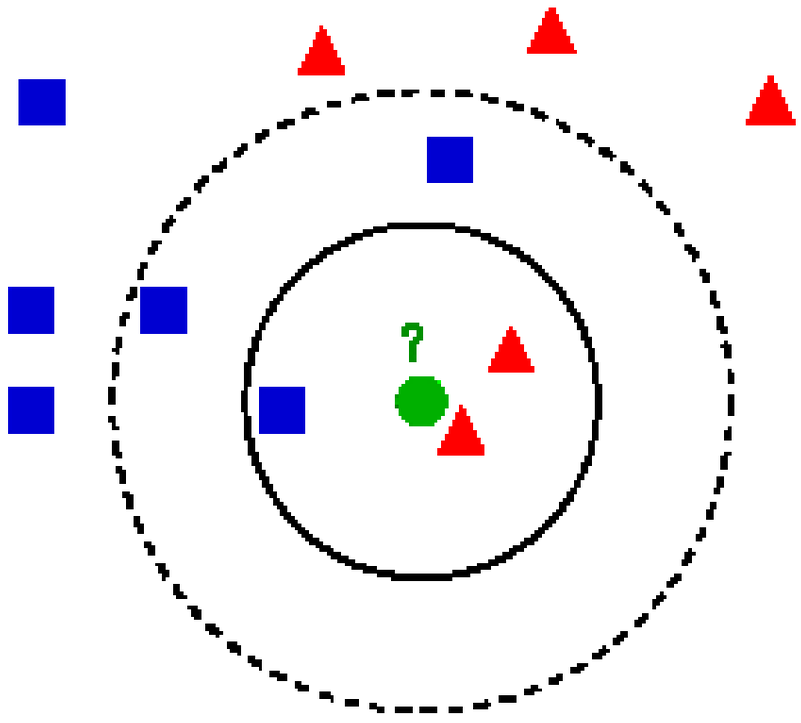
- Průměrná vzdálenost

$$D_2^2(x_1, x_2) = \frac{1}{p} \sum_{j=1}^{p} (y_{1j} - y_{2j})^2$$

- Chord distance
- Manhattanska lenost

KNN (K-Nearest Neighbors)
Algoritmus k-nejbližších sousedů

- patří mezi nejjednodušší “machine lerning” algoritmy
- Je to jedna z mnoha hierachickým metod shlukování
- může sloužit ke klasifikaci nebo predikci
- pokud budeme klasifikovat zelené kolečko podle 3-KK (bereme v úvahu 3 sousedy) bude patřit do trojúhelníků, v 5-KK do čtverců
- v případě regrese se hodnota bude počítat na základě nejbližších sousedů a vzdálenosti

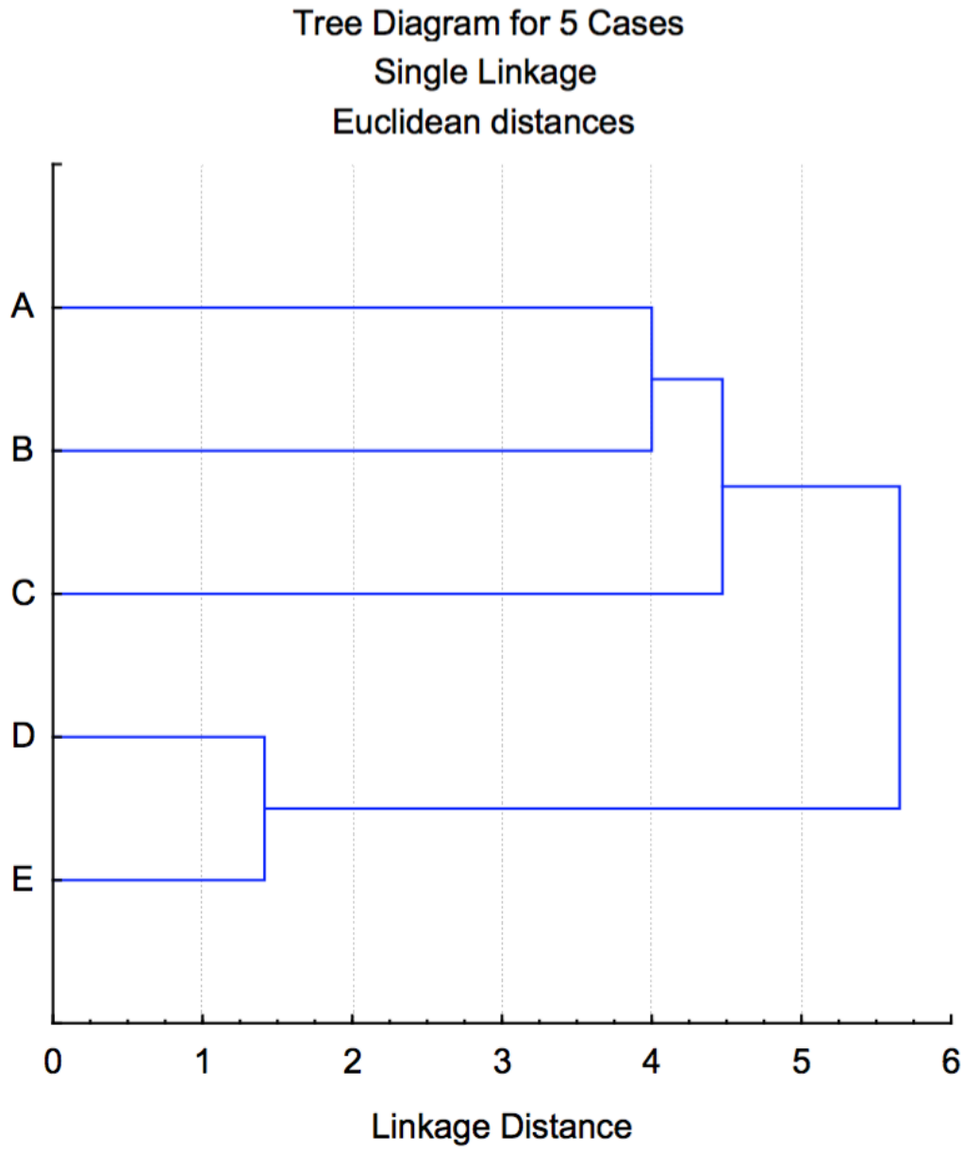


Tvorba KNN

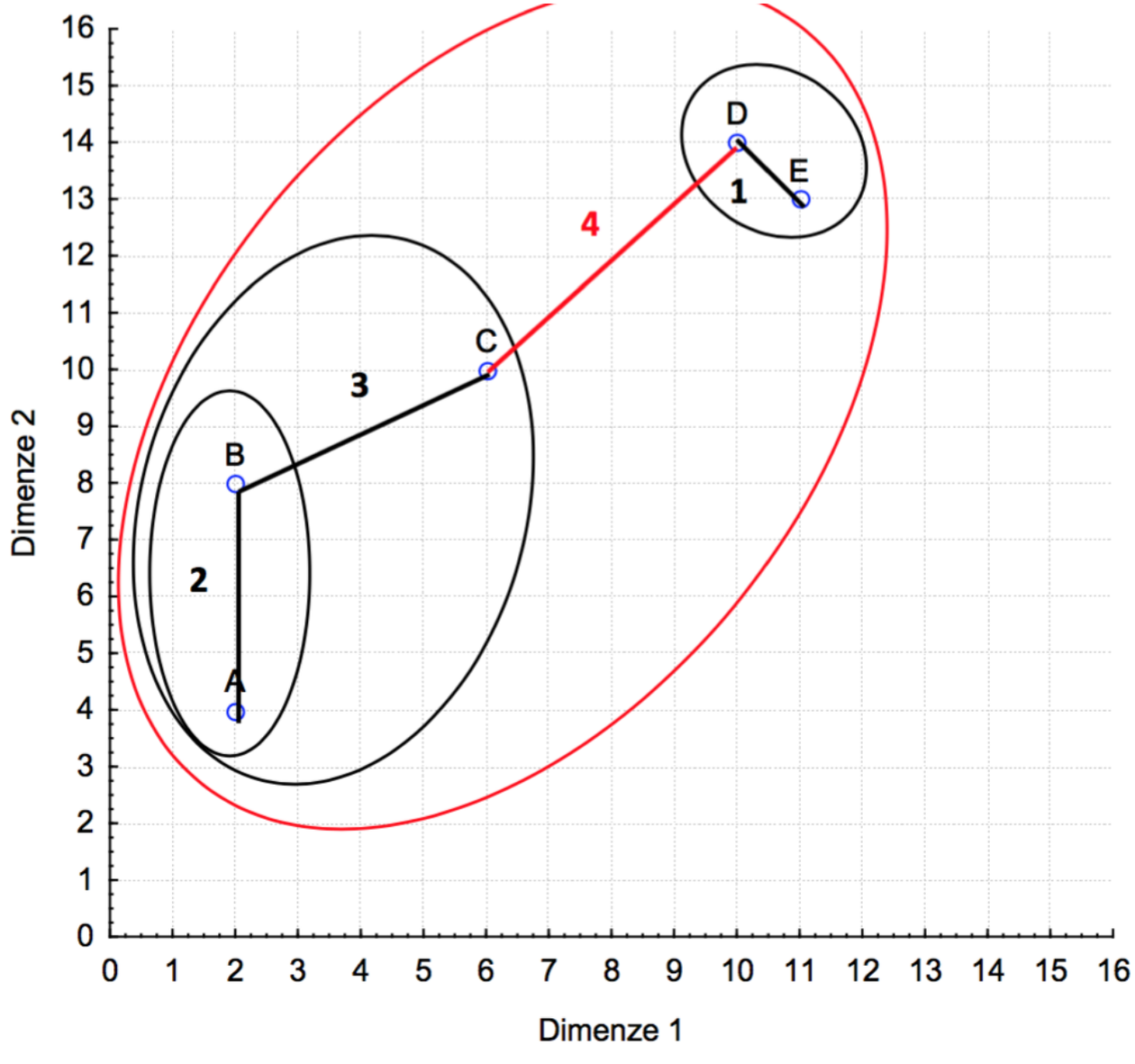
1. Výpočet asociační matice
2. Definice prvního shluku (nebližší prvky v matici)
3. Přepočet matice, nový shluk je brán jako nový prvek
4. Opakujeme kroky 2 - 3, dokud není pouze jeden “prvek”

Vizualizace shlukové analýzy

dendogram



klastry



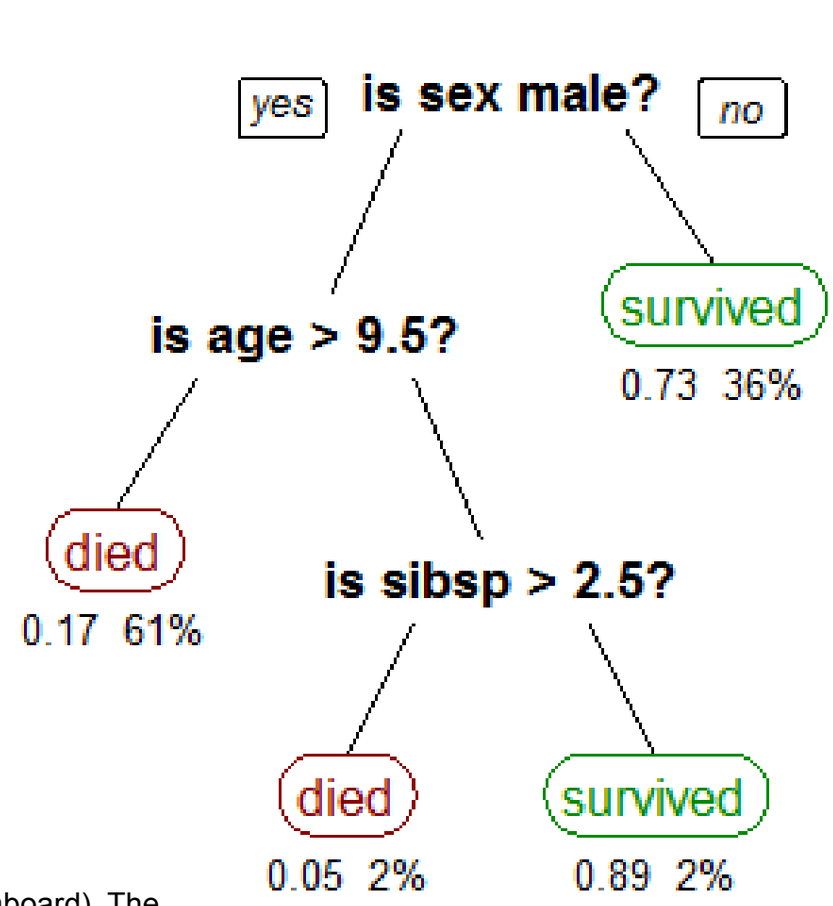
Výhody a nevýhody KNN

- Výhody:
  - Robustní na nesourodé data
  - Efektivní při velkém množství dat
- Nevýhody:
  - Potřeba definovat parametr K
  - Potřeba výběru metody výpočtu vzdálenosti
  - Výpočetně náročné – nutnost výpočet všech vzdáleností

Decision Tree
Rozhodovací stromy

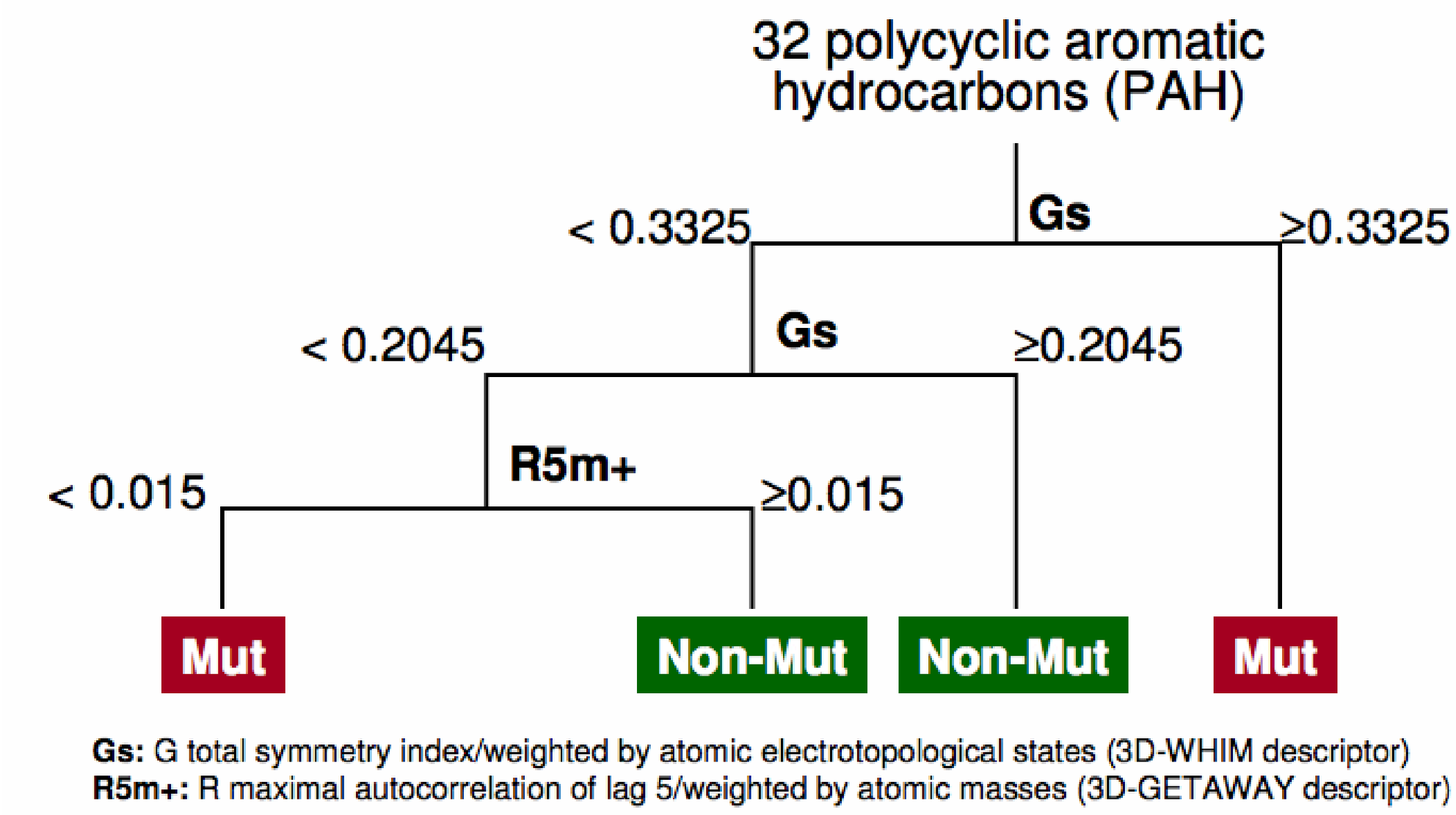
- může sloužit ke klasifikaci (classification trees) nebo predikci (regression trees)

- V těchto stromových strukturách představují listy (leafs) třídy a větve představují spojky mezi třídami.



Zdroj: wiki | A tree showing survival of passengers on the Titanic ("sibsp" is the number of spouses or siblings aboard). The figures under the leaves show the probability of survival and the percentage of observations in the leaf.

Decission tree – příklad predikce mutagenity



Zdroj: P. Gramatica, E. Papa, A. Marrocchi, L. Minuti, A. Taticchi, Ecotoxicology and Environmental Safety 2007, 66(3), 353-361.

Tvorba rozhodovacího stromu CART

1. Regresní stromy
   Kriteriální statistika: Least square devition
   $Q = 1/n \sum (y_i - \bar{y})^2$

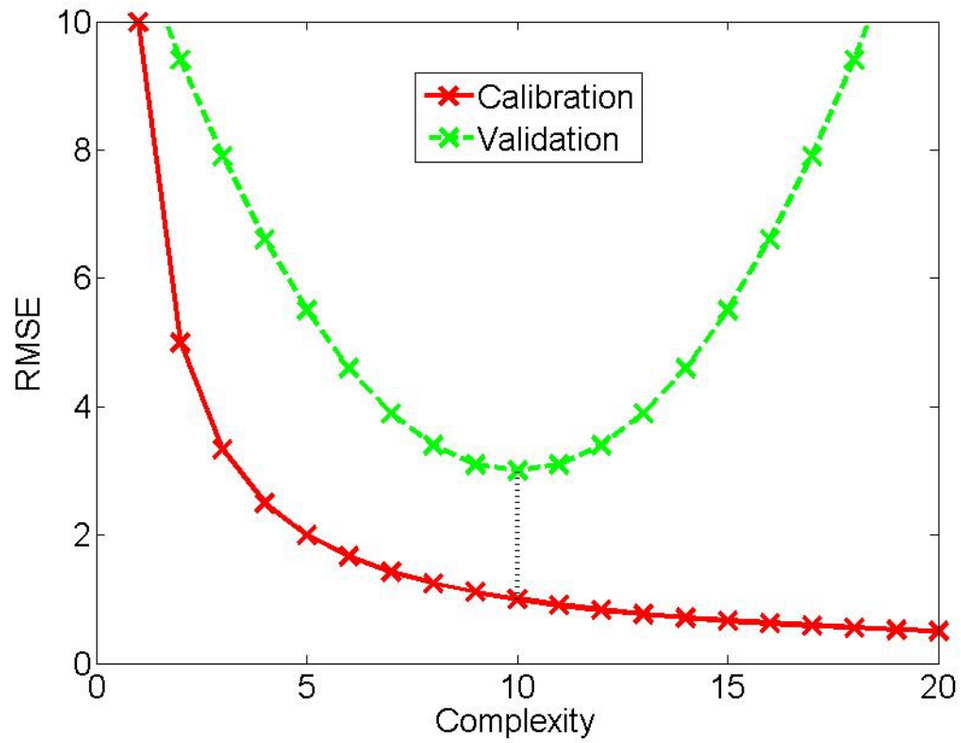
2. Klasifikační stromy
   Kriteriální statistika: Gini index

- Podle kriteriální statistiky je provedene nejvhodnější rozdělení na dceřinné uzly

- Růst je zastaven podle zvolených kriterií (nebo vyčerpání možností)

Testování výsledného stromu a jeho prořezání

- Každý strom je nutné testovat pomocí externí validace nebo křížové validace!



- Výsledná komplexita stromu je redukována dle výsledku testování.

Výhody a nevýhody rozhodovacích stromů

- Výhody:
  - Snadné grafické znazornění
  - Odolný vůči odlehlým hodnotám
  - Vhodné použití pro regresi i klasifikaci
- Nevýhody:
  - Nestabilní
  - Stromy jsou nevhodné pro malý počt vzorků
  - Vyžaduje znalosti s nastavením stromů

Náhodné lesy

- Tato metoda používá pro klasifikaci nebo regresi velké množství stromů
- Vhodné metoda pro použití v případě velkého množství deskriptorů a dobře funguje i na malých datových sadách
- Možnosti dalších analýz:
  - Měření významnosti proměnných
  - Efekt proměnných na predikci
  - Prototyp kategorií a možnost jejich překryvu
  - Detekce odlehlých hodnot
  - Doplenění chybějících hodnot

Genetický algoritmus

- TODO