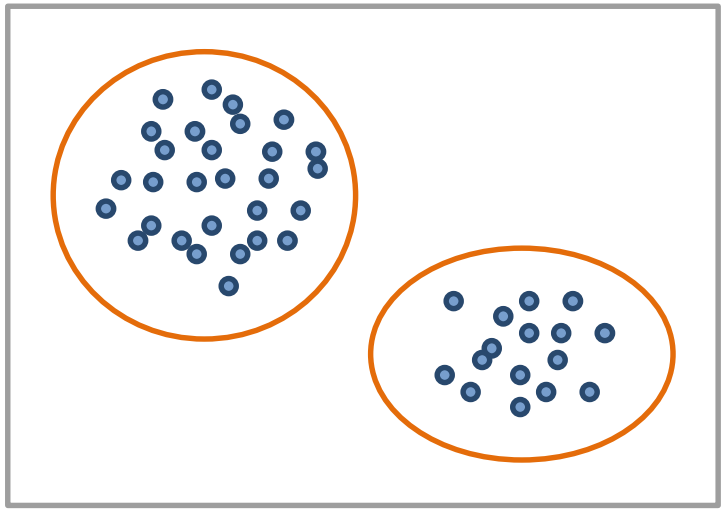
NUMERICKÁ KLASIFIKACE

SHLUKOVÁNÍ

- rozpoznání objektů, které jsou si dostatečně podobné, aby mohly být dány do stejné skupiny

- zjištění odlišností mezi skupinami



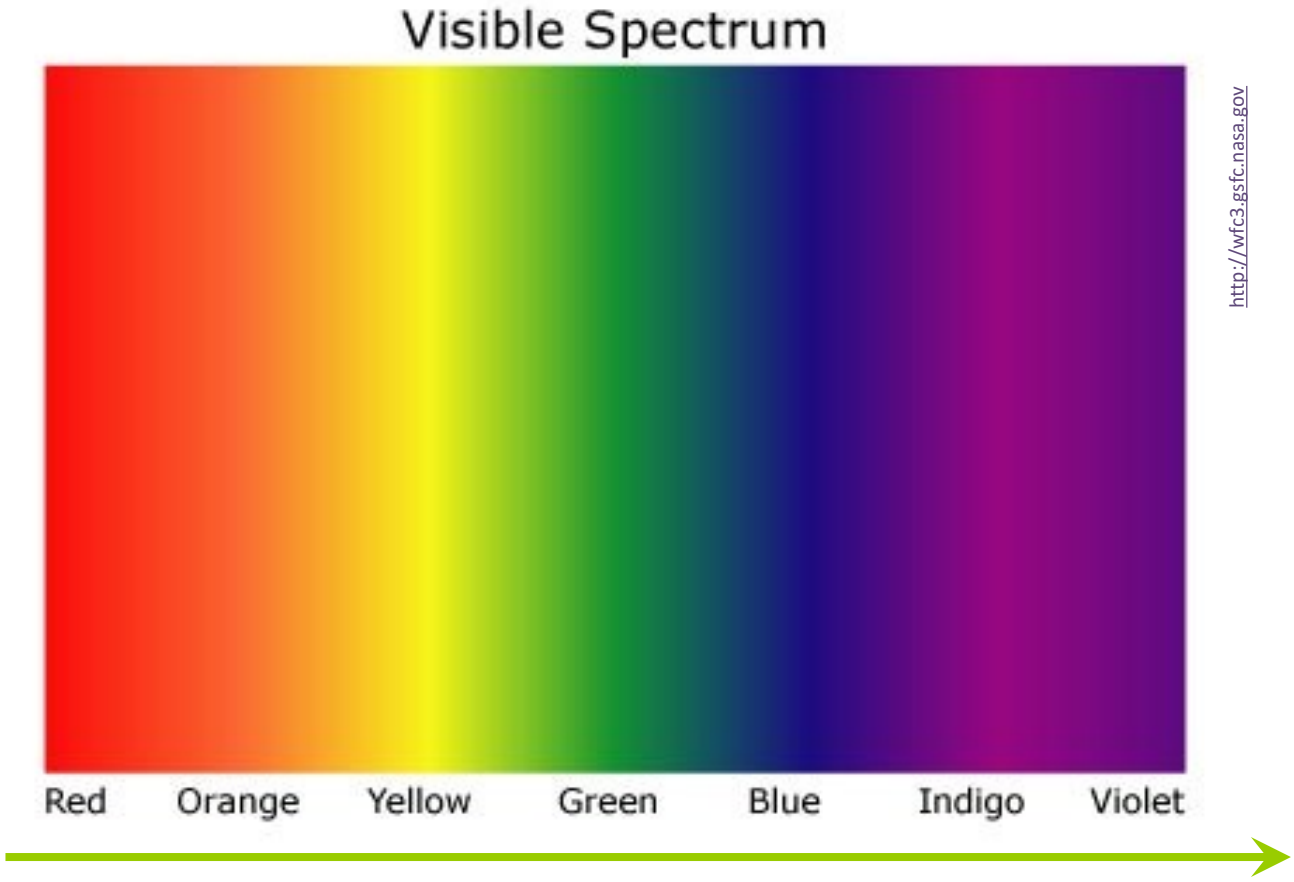
DISKONTINUUM VS. KONTINUUM

- Evoluční teorie predikuje diskontinuum – druhy
  - taxonomové hledají diskontinuity dané odlišnostmi mezi druhy

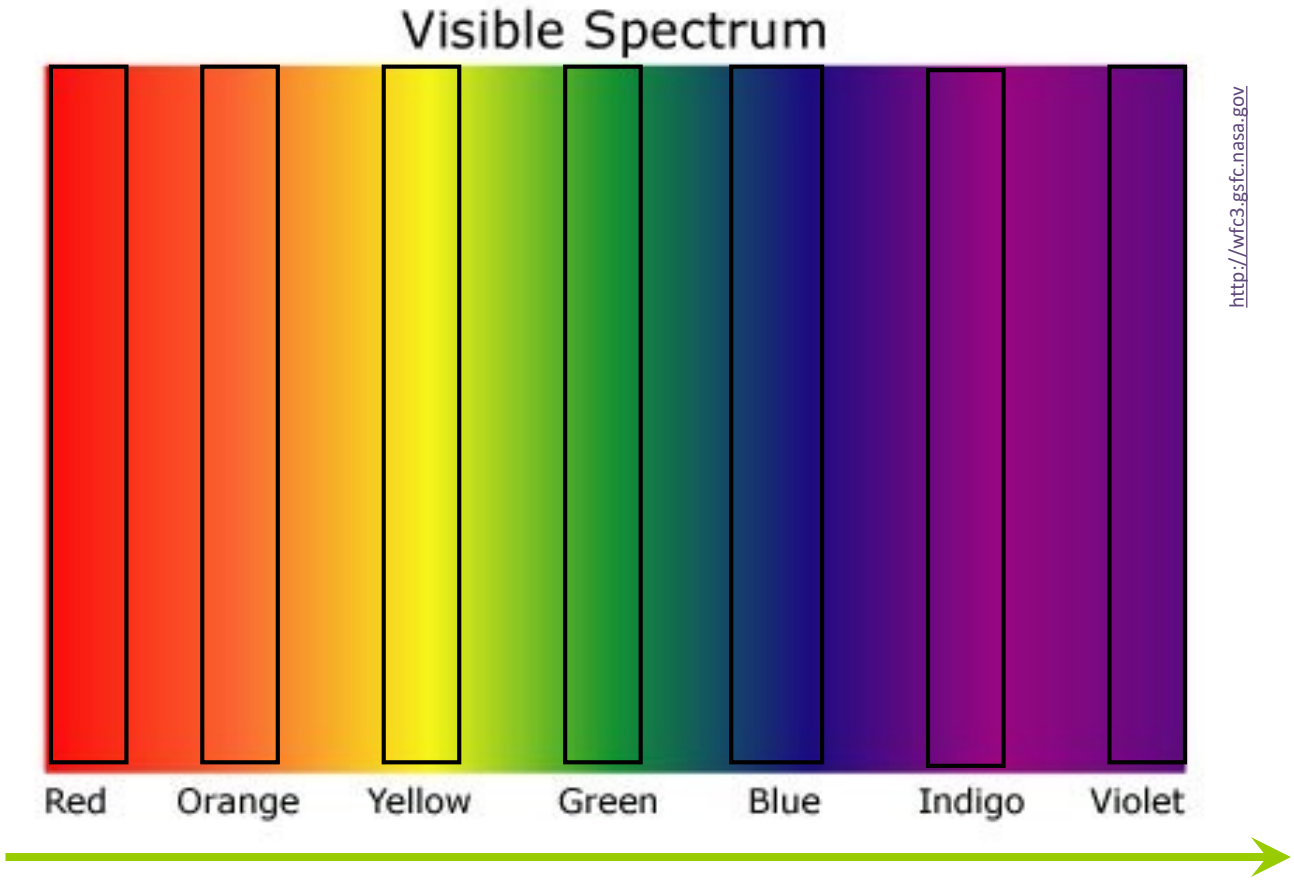
- Svět ekologie nejčastěji kontinuální
  - metody schopné rozpoznat shluky podobných objektů, zatímco ignorují několik hraničních

- Nelze očekávat diskontinuity ve společenstvech, aniž by prostředí bylo diskontinuální (nebo nevzorkujeme opačné konce gradientů) Whittaker 1962

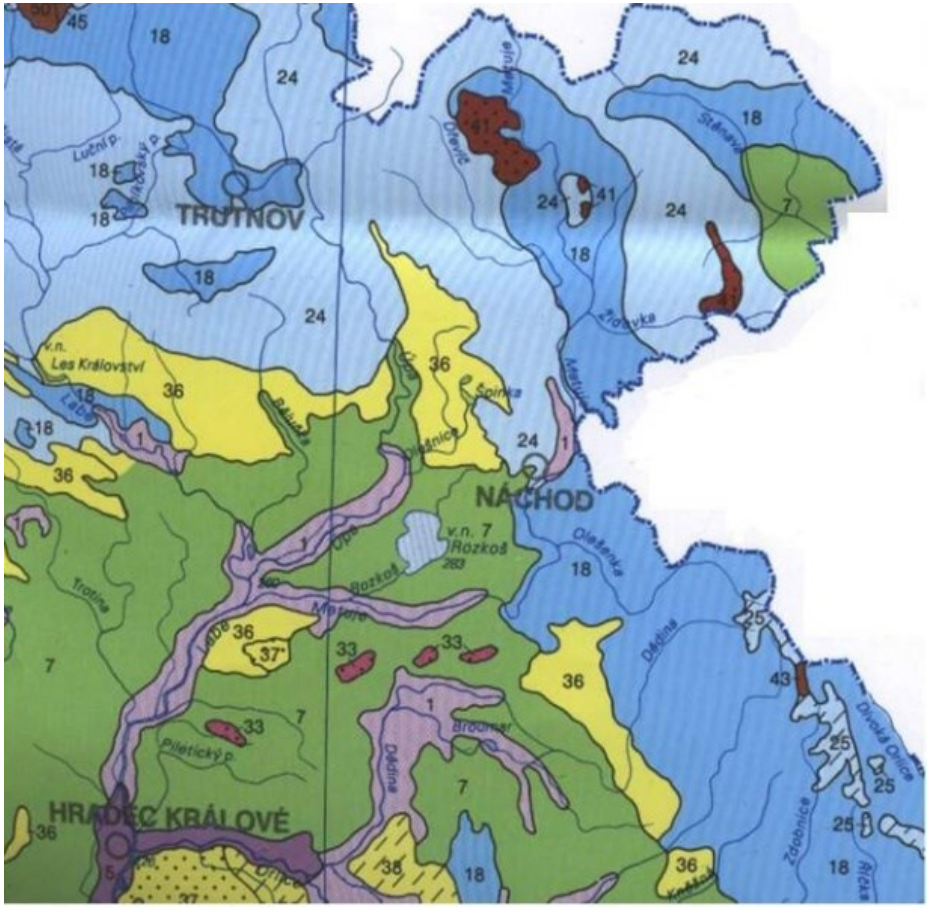
PROČ MÁ SMYSL VĚCI KLASIFIKOVAT?



PROČ MÁ SMYSL VĚCI KLASIFIKOVAT?



PROČ MÁ SMYSL KLASIFIKOVAT?



1-střemchová jasenina /Pruno-Fraxinetum/ místy v komplexu s mokřadními olšinami /Alnion glutinosae/
7-černýšová dubohabřina /Melampyro nemorosi-Carpinetum/
18-bučina s kyčelnicí devítilistou /Dentario enneaphylli-Fagetum/
24-biková bučina /Luzulo-Fagetum/
33-mochnová doubrava /Potentillo albae-Quercetum/
36-biková a/nebo jedlová doubrava /Luzulo albidae-Quercetum petraeae, Abieti-Quercetum/
37-bezkolencová doubrava /Molinio arundinaceae-Quercetum/
41-(sub)montánní a smrčina na balvanitých rozpadech /Betulo carpaticae-Pinetum, Anastrepto piceteeum/

KLASIFIKACE

- smyslem je najít diskontinuity (v jinak často kontinuální realitě), které můžeme pojmenovat – například proto, abychom si usnadnili komunikaci
- cílem je seskupit podobné objekty (vzorky, druhy) do skupin, které jsou vnitřně homogenní, dobře popsatelné a zároveň dobře odlišitelné od ostatních skupin

  - pokud analyzuji vzorky – daná skupina obsahuje vzorky s podobným druhovým složením (např. podobná stanoviště)
  - pokud analyzuji druhy – daná skupina obsahuje druhy s podobným ekologickým chováním

TYPOLOGIE
„SYSTÉM TYPŮ“

- výsledek shlukování objektů na kontinuálním gradientu
- výsledkem typy, pomocí nichž lze popsat kontinuum
- tyto typy samozřejmě nejsou ani „přirozené“, ani jediné „správné“

KLASIFIKACE

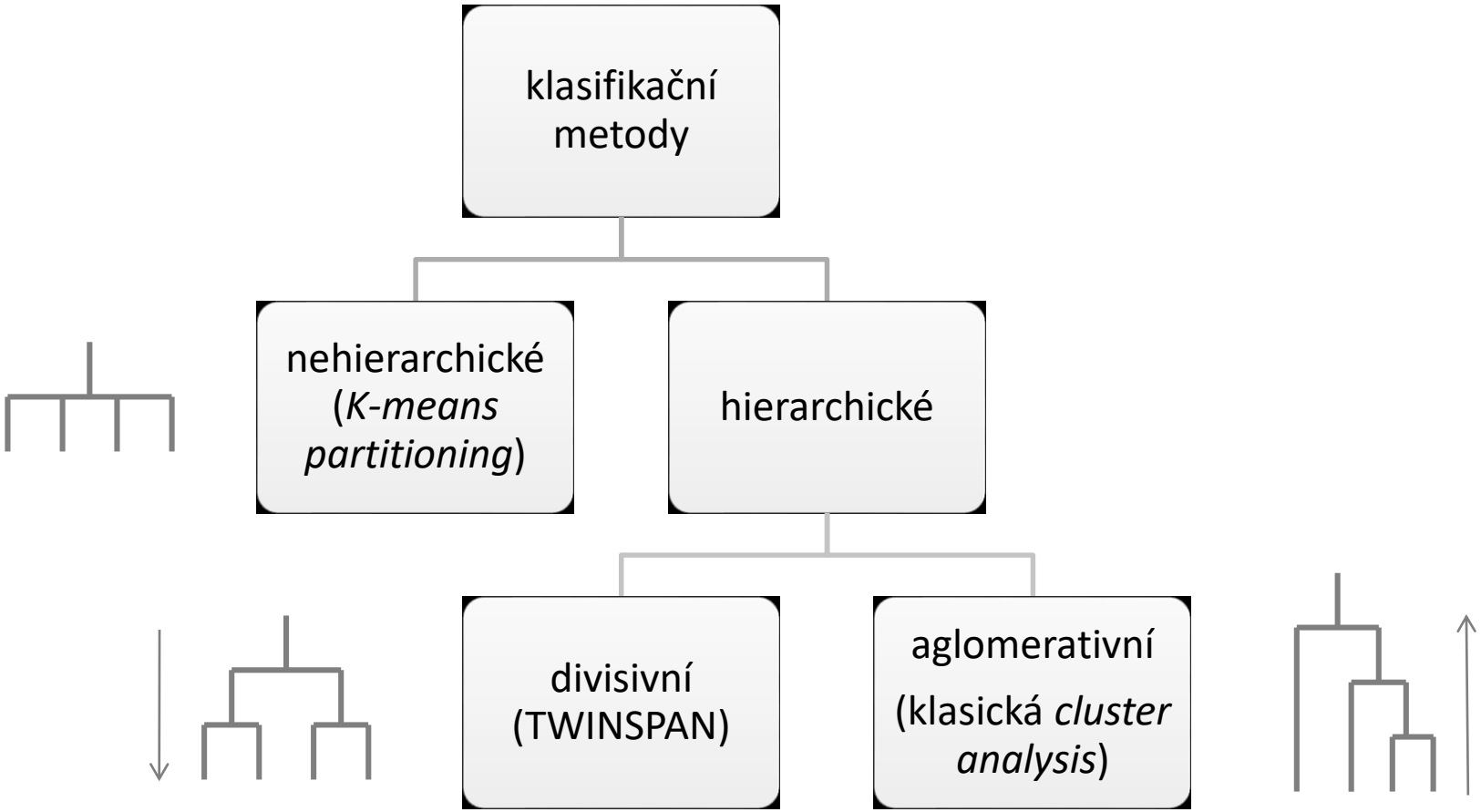
OBECNÉ ROZDĚLENÍ

- subjektivní *vs* ~~objektivní~~
  - v době rozkvětu metod numerické klasifikace se věřilo, že numerické metody přinášejí klasifikaci založenou na objektivních kritériích, tedy tu která „skutečně existuje“ (narozdíl od té subjektivní, která je „výmyslem badatele“)
  - všechny klasifikace jsou ale z principu subjektivní
- neformalizovaná *vs* formalizovaná
  - formalizovaná klasifikace je taková, která je provedena na základě jasných kritérií a díky tomu je možné ji znovu reprodukovat
  - opakem je klasifikace založená na neformálních kritériích (například pocitu), kterou pak není snadné zopakovat

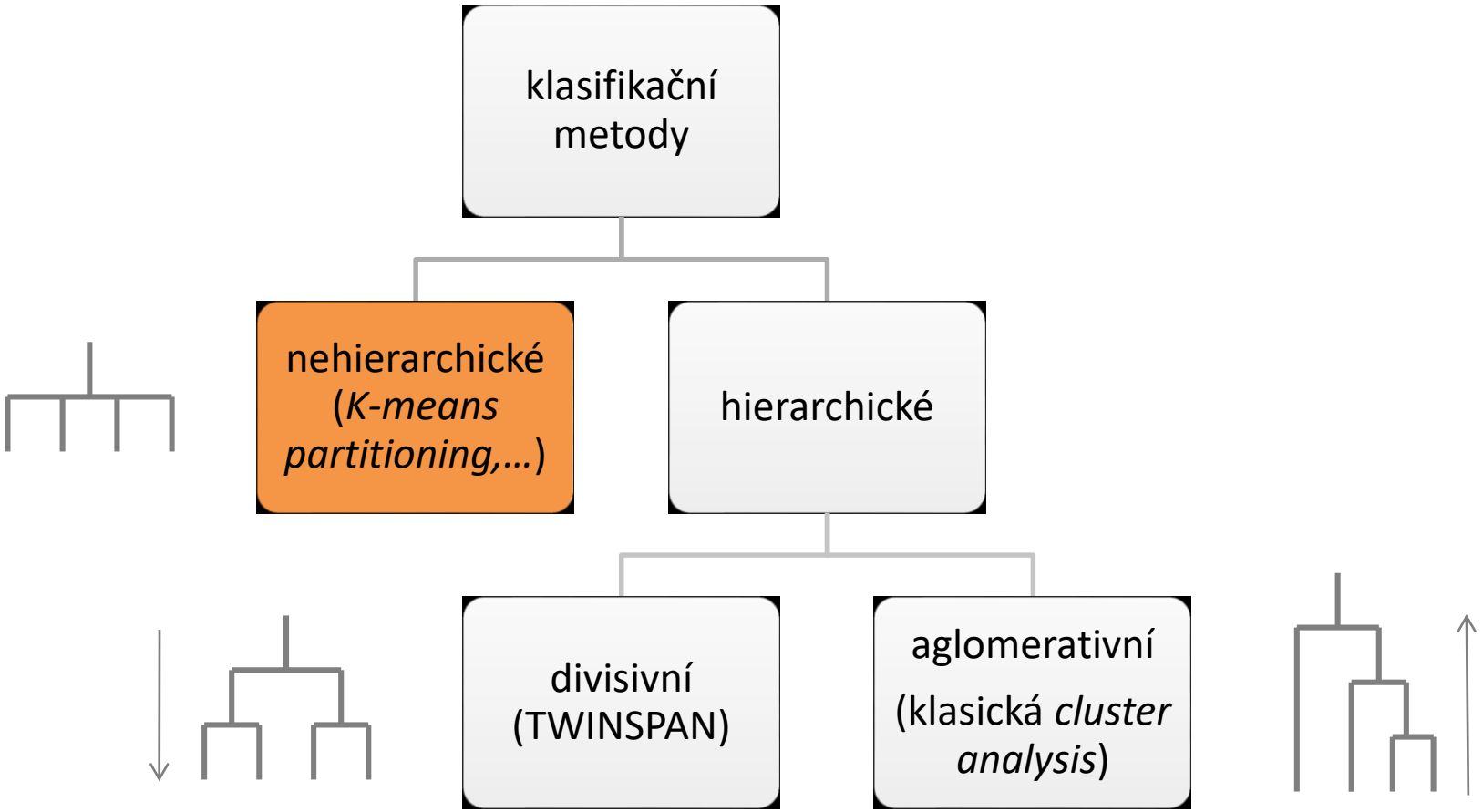
Otázky, které bych si měl položit před tím, než začnu něco klasifikovat

- **Pro jaký účel klasifikaci dělám?**
  - chci klasifikovat můj datový soubor (*srovnat knihy v mojí domácí knihovničce*)
  - chci vytvořit obecný klasifikační systém, který bude použitelný i na další soubory (*vytvořit knihovnický systém kategorizace knih, používaný i v jiných knihovnách*)
- **Podle jakých kritérií budu objekty klasifikovat?**
  - kritérium, podle kterého budu posuzovat, jestli si jsou objekty více či méně podobné (*knihy budu třídit podle obsahové podobnosti nebo např. podle velikosti*)
  - odpovídá výběru indexu podobnosti mezi vzorky
- **Jak stanovím hranice mezi jednotlivými skupinami?**
  - pravidla, podle kterých budu přiřazovat objekty do skupin
  - odpovídá výběru klasifikačního algoritmu

KLASIFIKACE

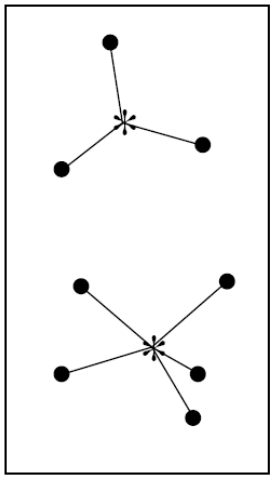


KLASIFIKACE



KLASIFIKACE

NEHIERARCHICKÁ



Legendre & Legendre 1998

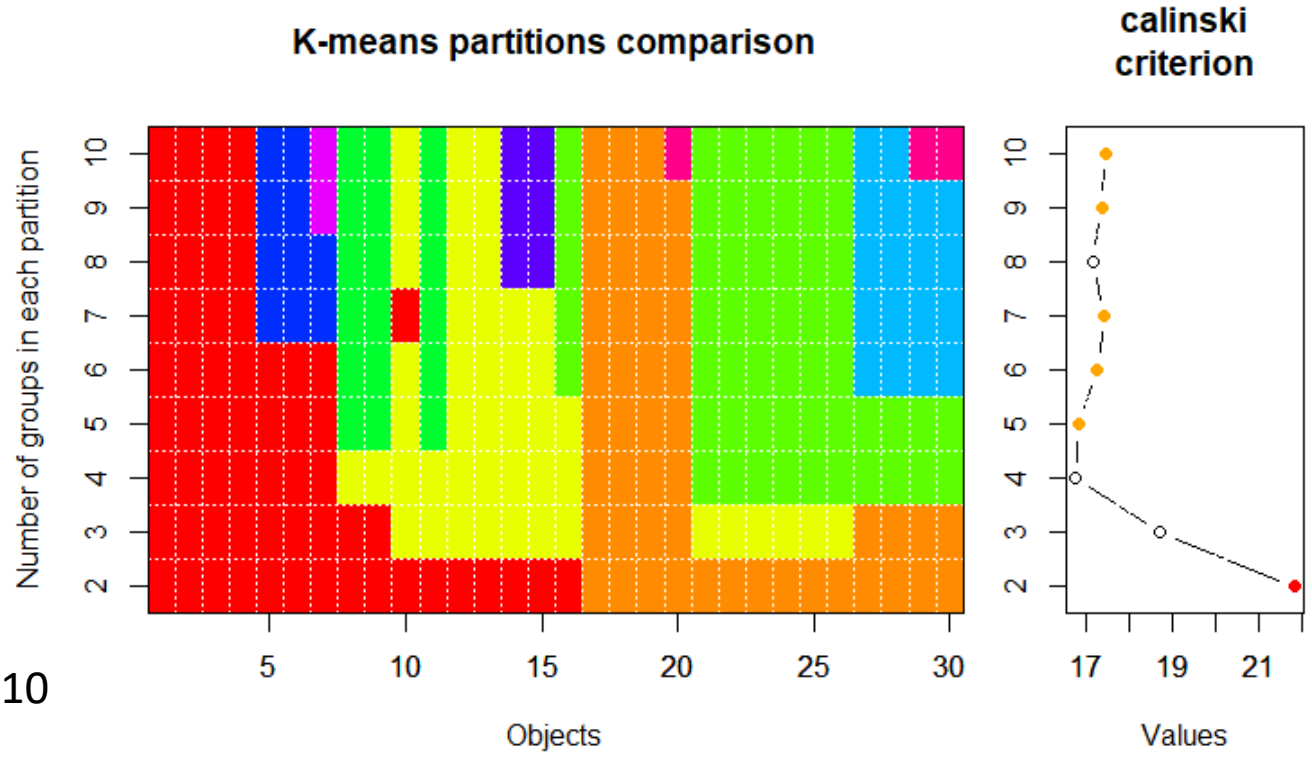
K-means partitioning

(shlukování metodou K-průměrů)

- nehierarchická metoda – všechny shluky jsou si rovny
- minimalizuje sumy čtverců vzdáleností vzorků od centroidů shluku
- na začátku uživatel zvolí počet shluků (*k*)
- iterativní metoda, začne od náhodného přiřazení vzorků do shluků, postupně přehazuje vzorky mezi shluky a hledá optimální řešení
- výsledek do určité míry záleží na počátečním rozmístění shluků do vzorků a je proto dobré proces mnohokrát zopakovat (najít stabilní řešení), protože metoda má tendenci nacházet lokální minima

IDENTIFIKACE "SPRÁVNÉHO K"

- Spuštění kmeans přes cascadeKM
- Calinski – Harabasz criterion (~ F-ratio: $MS_{\text{mezi shluky}}/MS_{\text{uvnitř shluků}}$)
  - Nejvyšší hodnota ~ optimum
  - Hodnota K, při které C-H crit. Vzroste, může být taky zajímavá

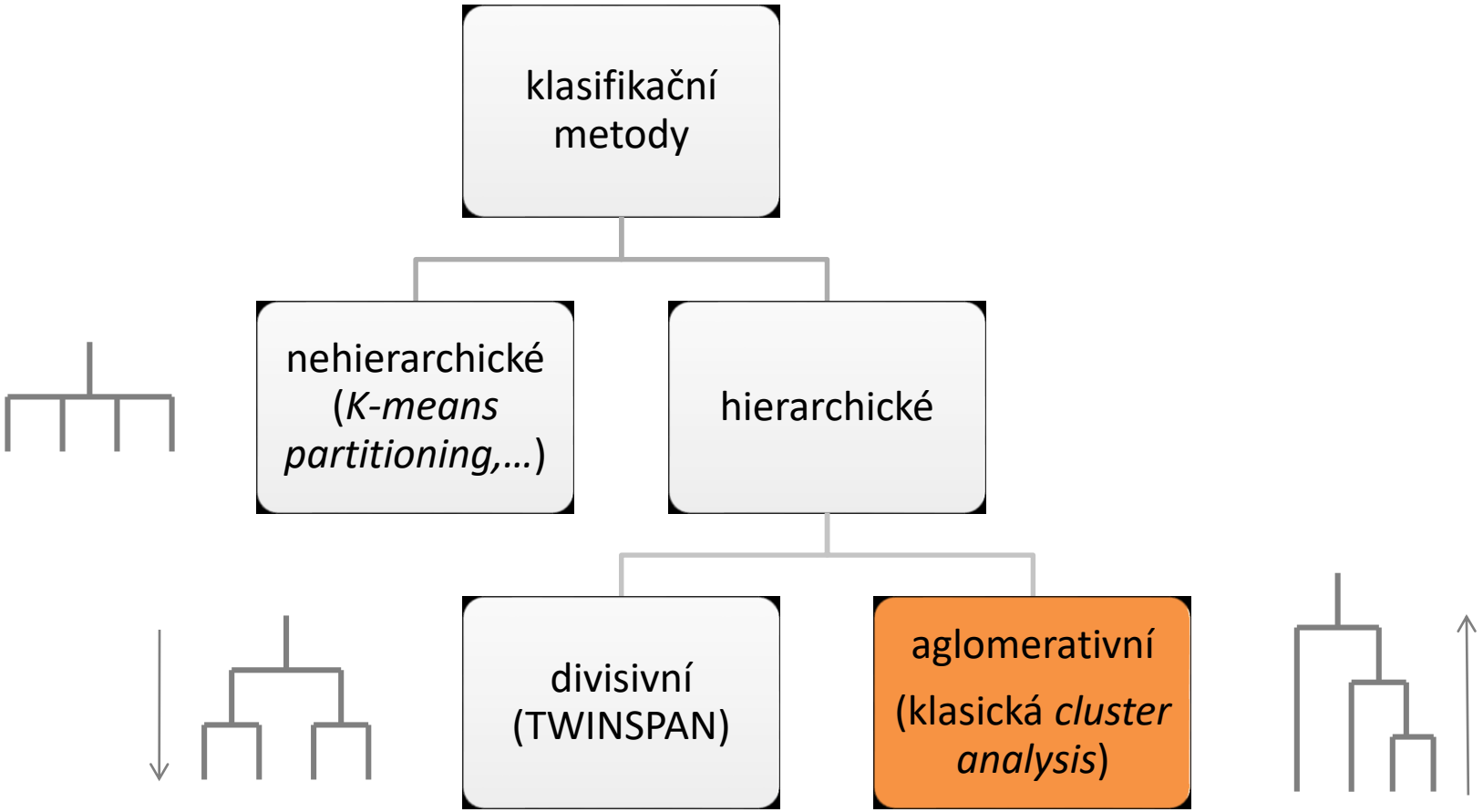


K-means pro k = 2-10

Další metody patritioning

- K-means
  - Kompatibilní pouze s na metrickými distancemi
  - Jinak nutné přepočítat nepodobnosti na vzdálenosti pomocí PCoA
    - Spočte se PCoA, a do K-mean se použije prvních X os, které vyscětlují 90 nebo 95% variability
    - Arbitrární, ale funguje
- Partitioning around the medoids
  - Centrum není průměr, ale nějaký konkrétní bod (medoid)
  - Použitelné na jakoukoliv nepodobnost
  - Má být robustnější než k-means

KLASIFIKACE



KLASIFIKACE

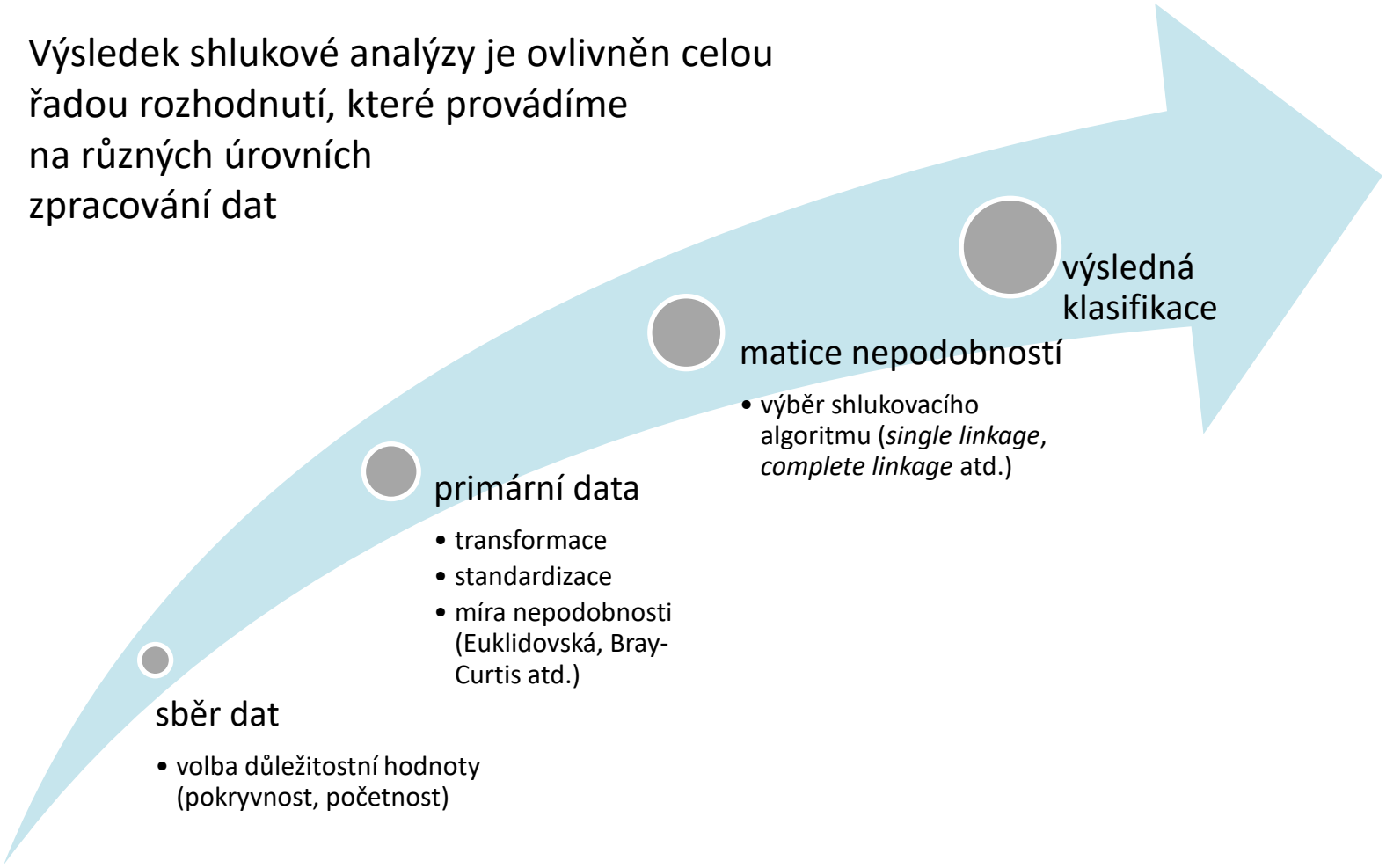
HIERARCHICKÁ A **AGLOMERATIVNÍ**

Shluková analýza (*cluster analysis*)

- hierarchická metoda
  - shluky jsou hierarchicky uspořádány
- aglomerativní metoda
  - shluky jsou tvořeny ‘odspodu’, tzn. postupným shlukováním jednotlivých vzorků do větších skupin
- základní volby:
  - míra nepodobnosti mezi vzorky (*distance measure*)
  - shlukovací (klastrovací) algoritmus (*clustering algorithm*)

SHLUKOVÁ ANALÝZA (*CLUSTER ANALYSIS*)

Výsledek shlukové analýzy je ovlivněn celou řadou rozhodnutí, které provádíme na různých úrovních zpracování dat



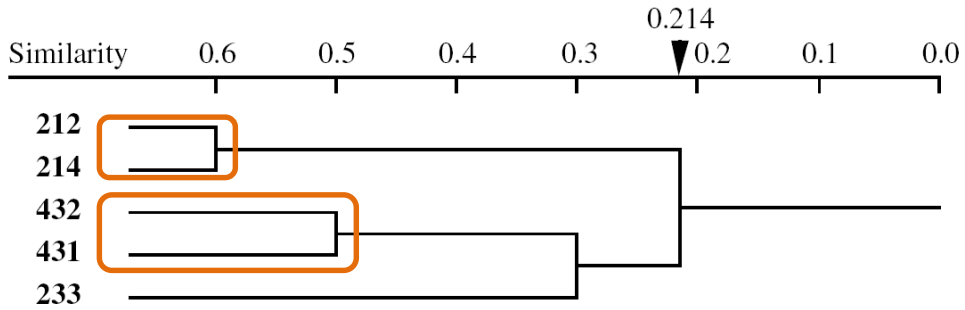
SHLUKOVÁ ANALÝZA (*CLUSTER ANALYSIS*)

SHLUKOVACÍ ALGORITMY

Metoda jednospojná (*single linkage*)

Ponds	Ponds				
	212	214	233	431	432
212	—				
214	0.600	—			
233	0.000	0.071	—		
431	0.000	0.063	0.300	—	
432	0.000	0.214	0.200	0.500	—

matice podobností



výsledný dendrogram

páry vzorků seřazené podle podobností

S_{20}	Pairs formed
0.600	212-214
0.500	431-432
0.300	233-431
0.214	214-432
0.200	233-432
0.071	214-233
0.063	214-431
0.000	212-233
0.000	212-431
0.000	212-432

Legendre & Legendre 1998

SHLUKOVÁ ANALÝZA (*CLUSTER ANALYSIS*)

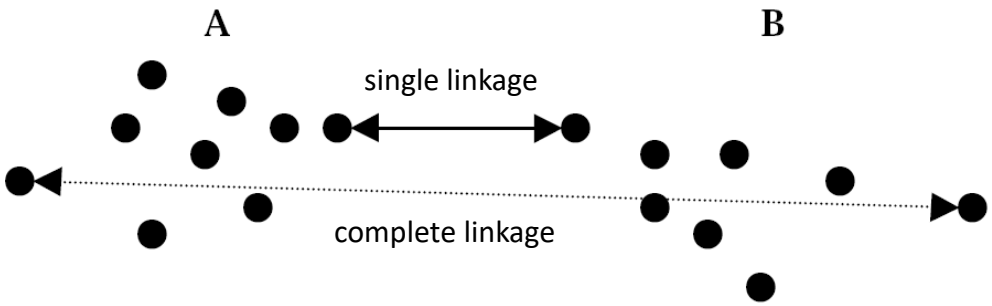
SHLUKOVACÍ ALGORITMY

Metoda jednospojná (*single linkage, nearest neighbour*)

- vzorky se pojí ke shluku, ve kterém je jim nejpodobnější vzorek
- *přidám se ke skupině, ve které je ten, kdo je mí nejvíc sympatický*

Metoda všespojná (*complete linkage, farthest neighbour*)

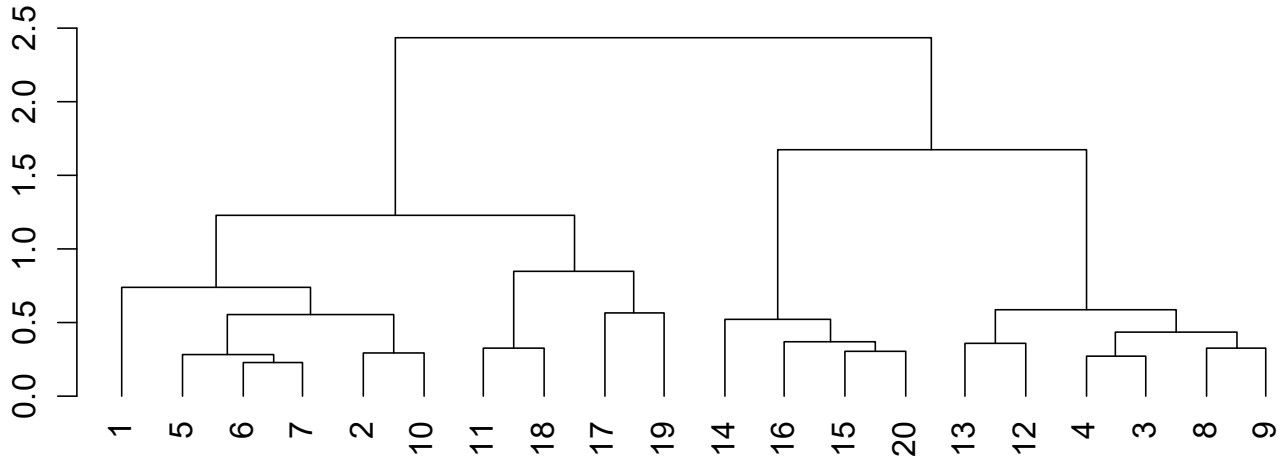
- vzorky se připojí ke shluku až v okamžiku, kdy shluk obsahuje všechny podobné vzorky
- *zjistím nejnesympatičtější jedince ve všech sjkupinách a přidám se ke skupině ve které je ten nejmíň nesympatický*



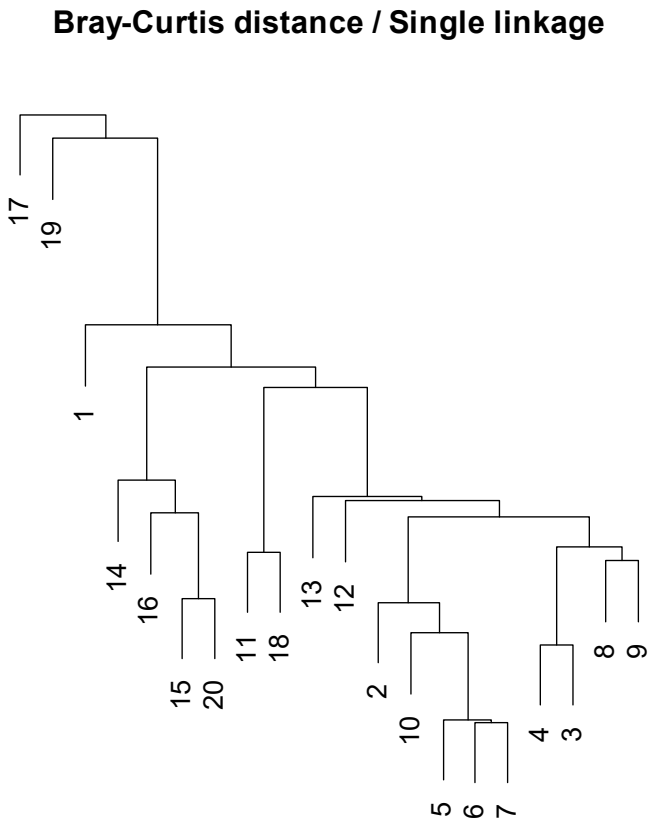
SHLUKOVÁ ANALÝZA (*CLUSTER ANALYSIS*)

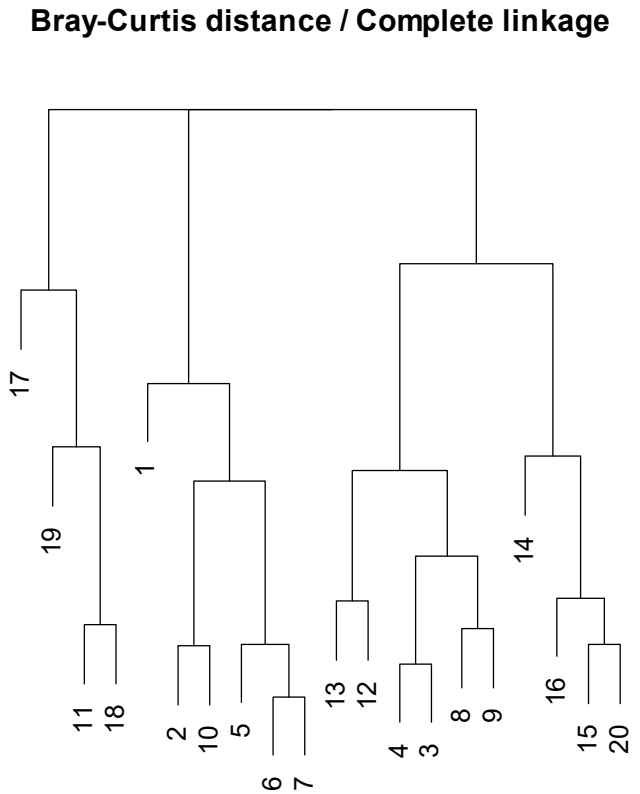
DENDROGRAM

- záleží na tom, které vzorky jsou spojeny na které úrovni
- nezáleží na tom, který vzorek (skupina) je vpravo a který vlevo



METODA JEDNOSPOJNÁ VS VŠESPOJNÁ





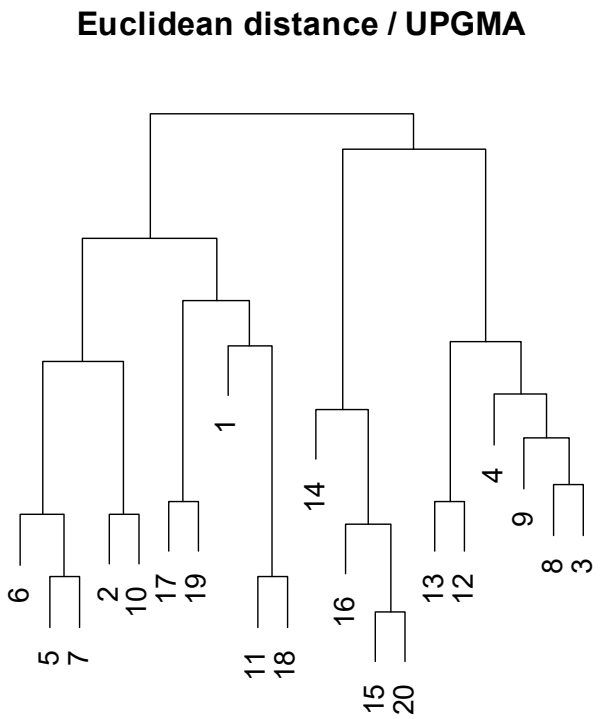
metoda jednospojná se výrazně řetězí

SHLUKOVÁ ANALÝZA (*CLUSTER ANALYSIS*)

SHLUKOVACÍ ALGORITMY

Average linkage (např. UPGMA)

- zahrnuje řadu metod, které stojí mezi *single* a *complete linkage* a v ekologii jsou smysluplnější
- UPGMA *(unweighted pair-group method using arithmetic averages)* – vzorek se připojí ke shluku, ke kterému má největší (neváženou) průměrnou podobnost se všemi jeho vzorky
- *přidám se ke skupině, ve které jsou mi všichni v průměru nejsympatičtější*

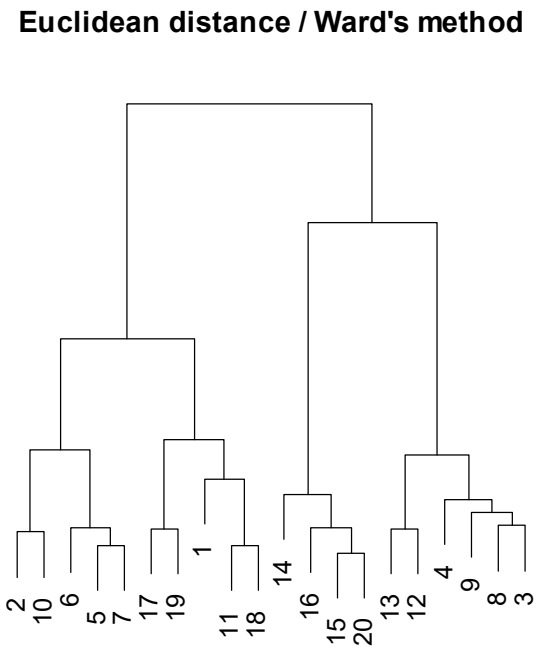


SHLUKOVÁ ANALÝZA (*CLUSTER ANALYSIS*)

SHLUKOVACÍ ALGORITMY

Wardova metoda (*Ward's minimum variance method*)

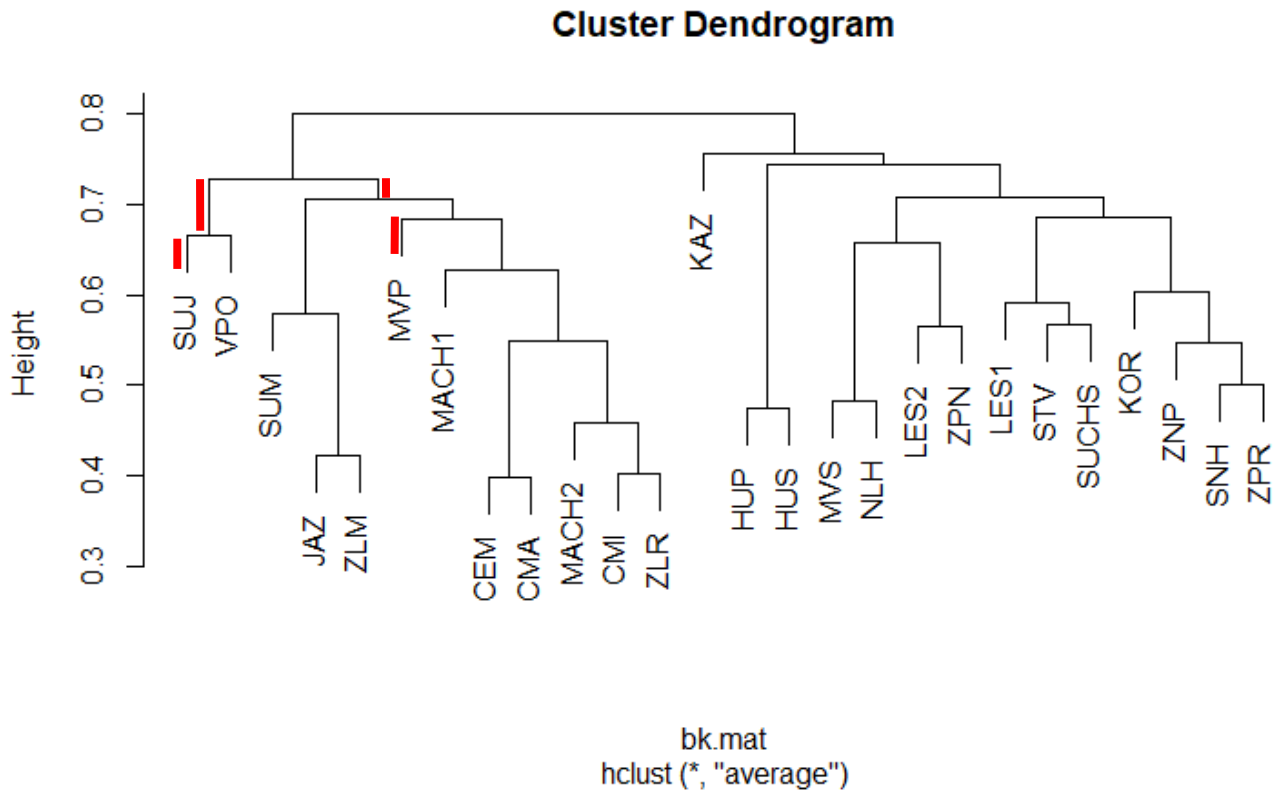
- minimalizuje součet čtverců vzdáleností mezi vzorky a centroidy jejich shluků
- jsou spojovány ty shluky (vzorky) jejichž shluknutí povede k nejmenšímu nárůstu součtu čtverců vnitroshlukových vzdáleností
- výsledné shluky mají tendenci být hypersférické a zhruba stejné velikosti
- neměla by se kombinovat se Sørensenovým (Bray-Curtis) indexem nepodobnosti, možno pouze s metrickými distancemi



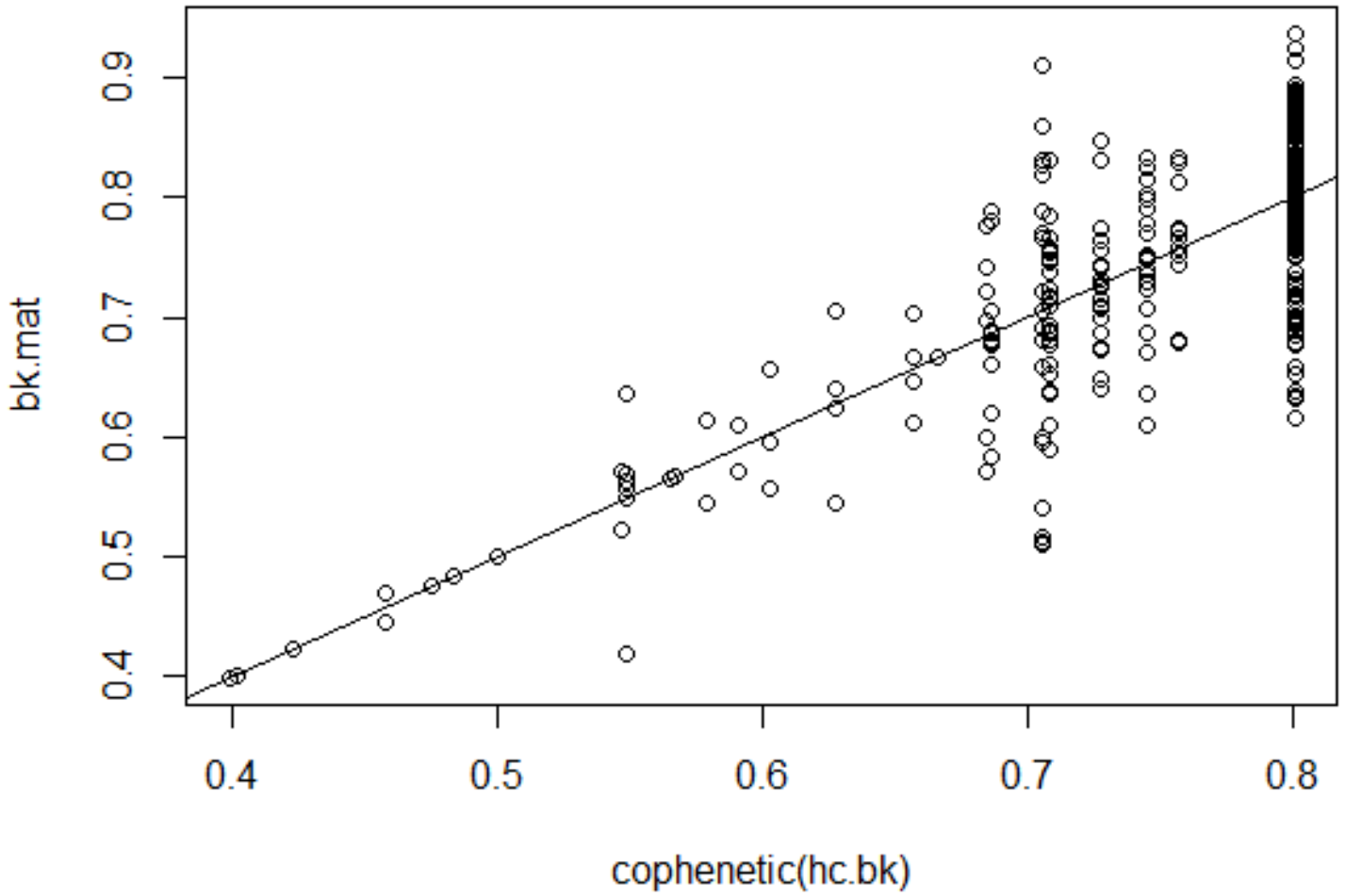
KOFENETICKÁ VZDÁLENOST

- Vzdálenost mezi dvěma vzorky definovaná jako nepodobnost v rámci skupiny v níž jsou dva vzorky spojené do jednoho klastru

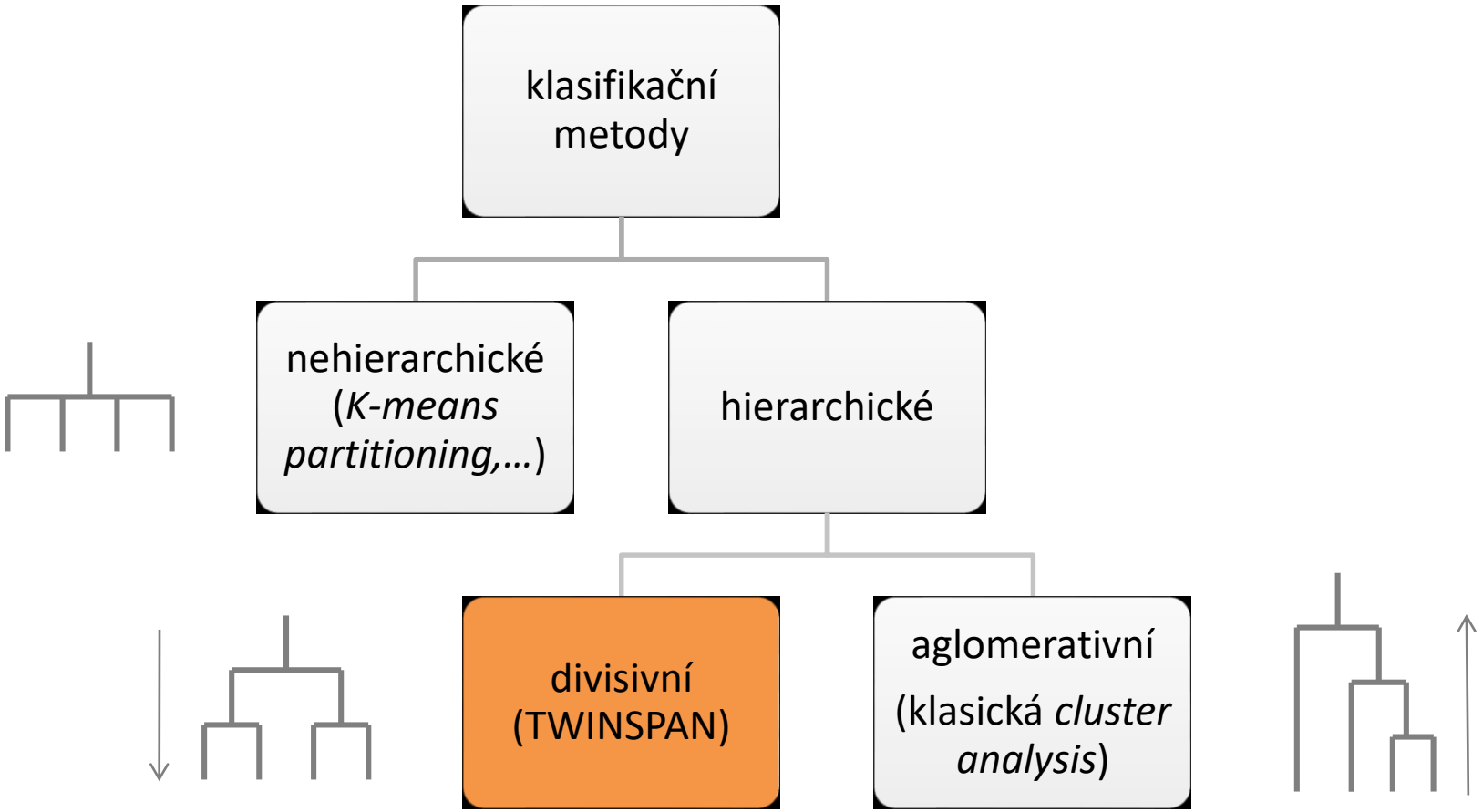
d(SUJ-MVP)



Vztah mezi originální nepodobností a kofenetickou vzdáleností



KLASIFIKACE



… PŘÍŠTĚ