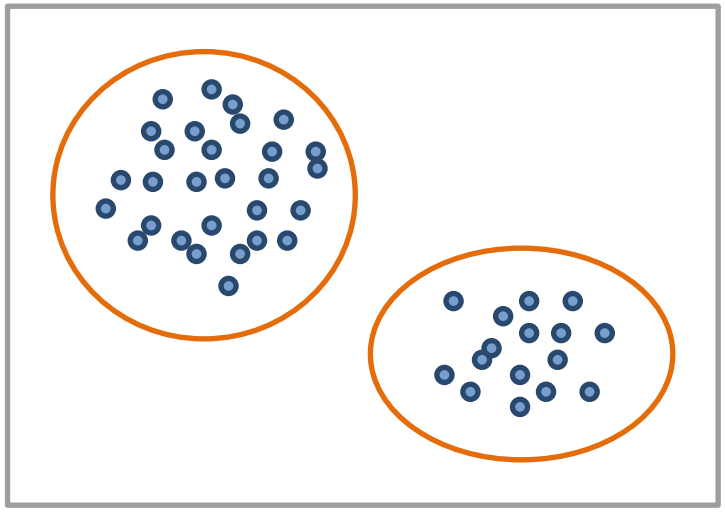
NUMERICKÁ KLASIFIKACE

SHLUKOVÁNÍ

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

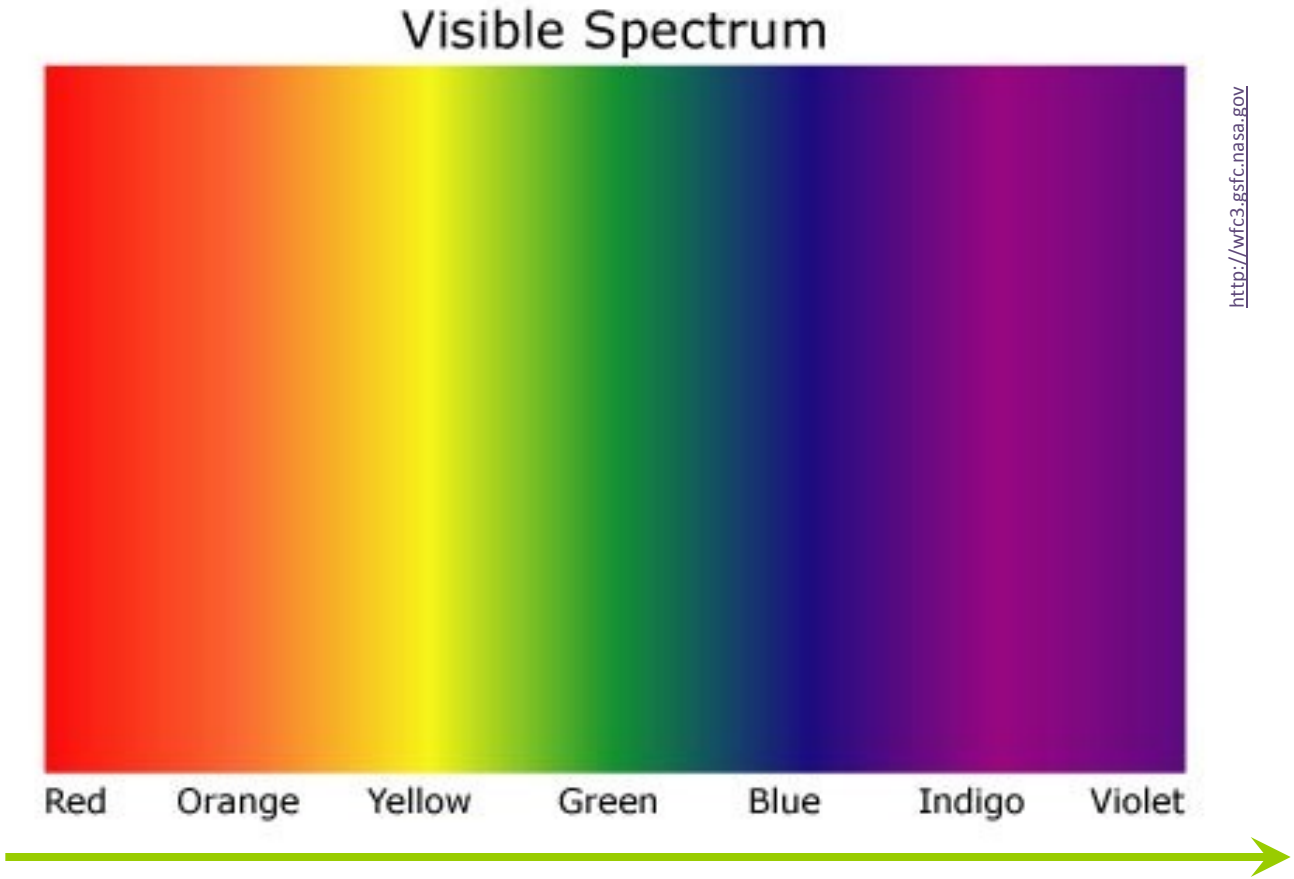
- zjištění odlišností mezi skupinami



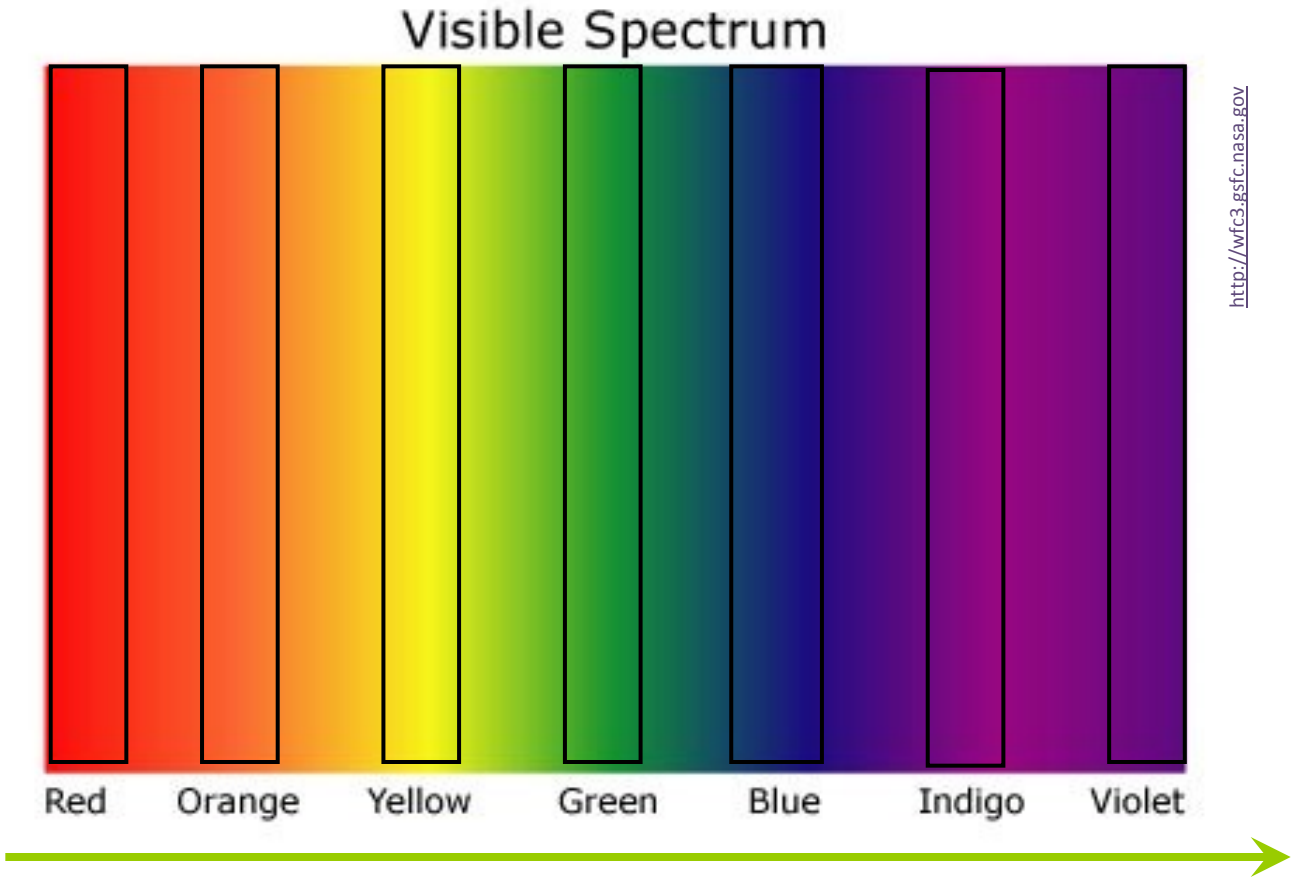
Diskontinuum vs. kontinuum

- 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?

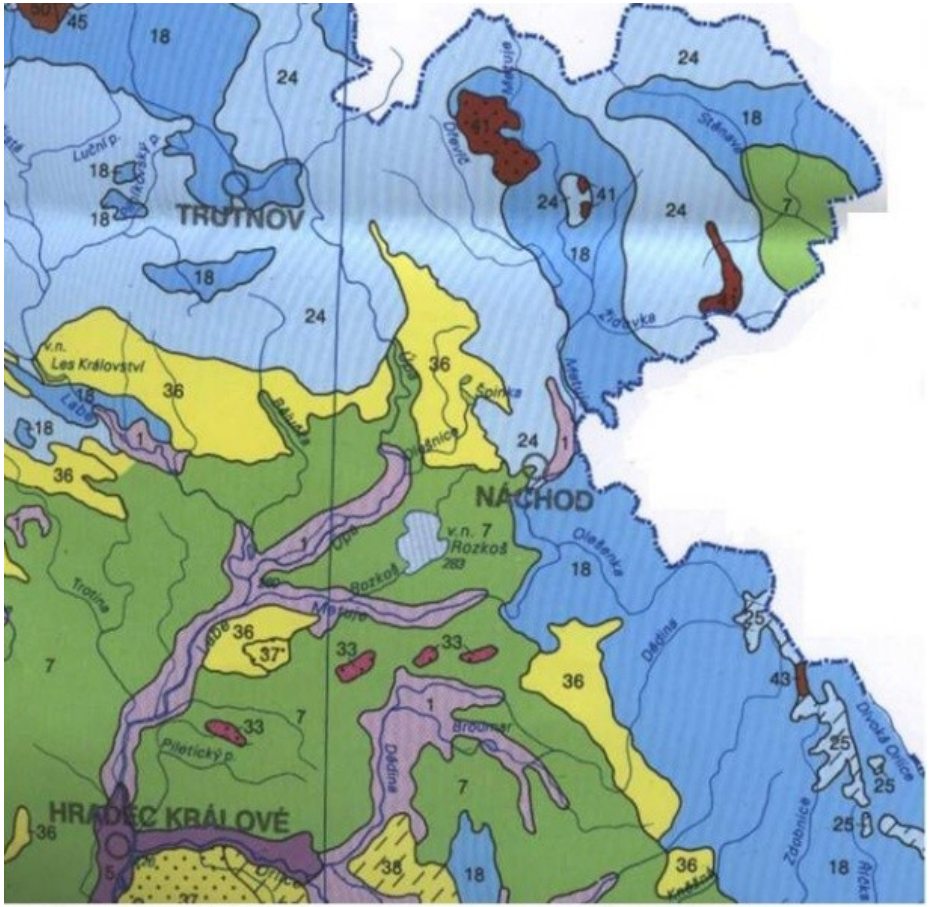


Proč má smysl věci klasifikovat?



vlnová délka (~ ekologický gradient)

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
- Výsledné shluky lze považovat za „typy“
  - Umožňují popsat kontinuum
  - Vzhledem k subjektivitě klasifikací nemají tyto typy nárok na označení ani „přirozené“, ani „jediné správné“

KLASIFIKACE

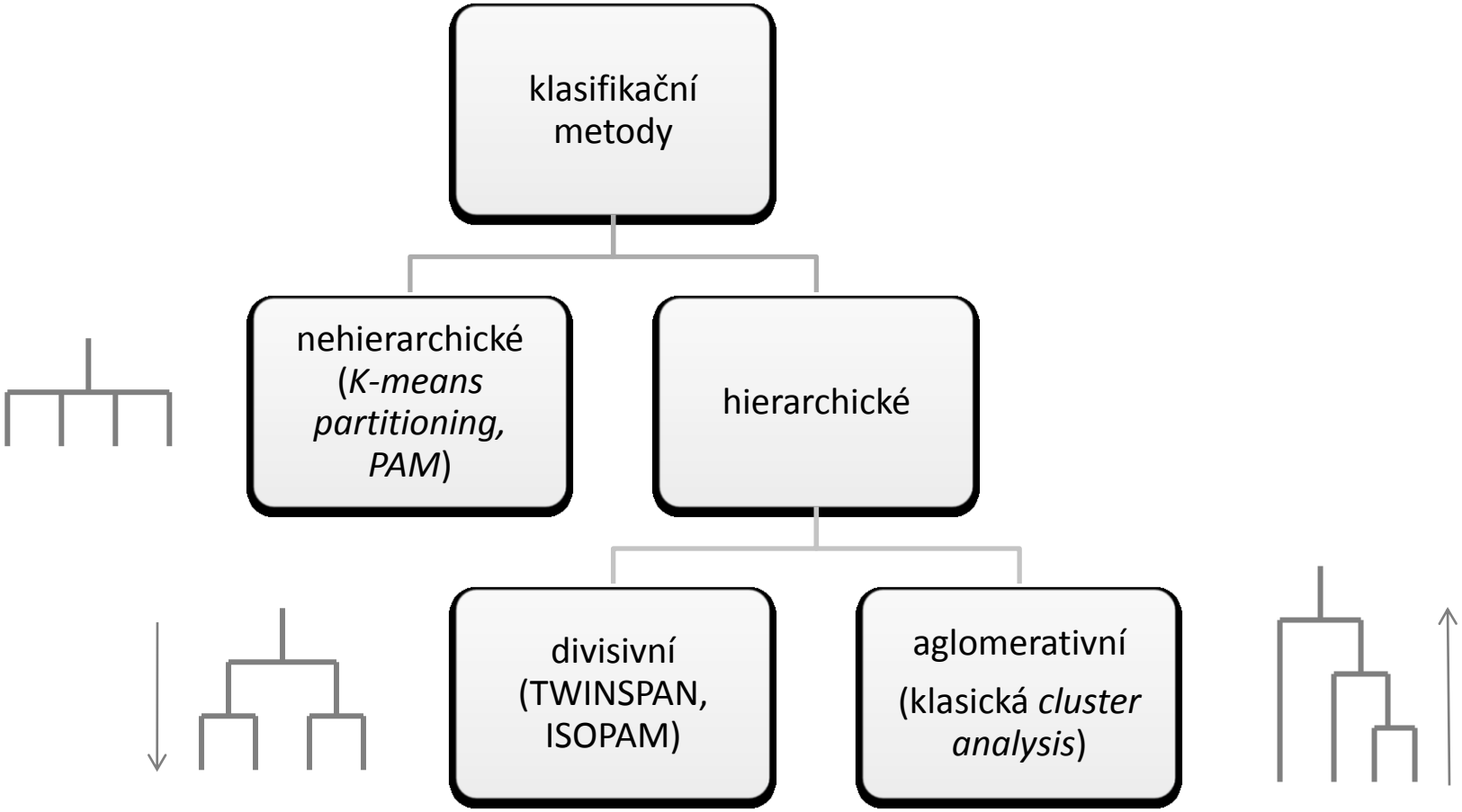
OBECNÉ ROZDĚLENÍ

- 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“ (na rozdíl 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

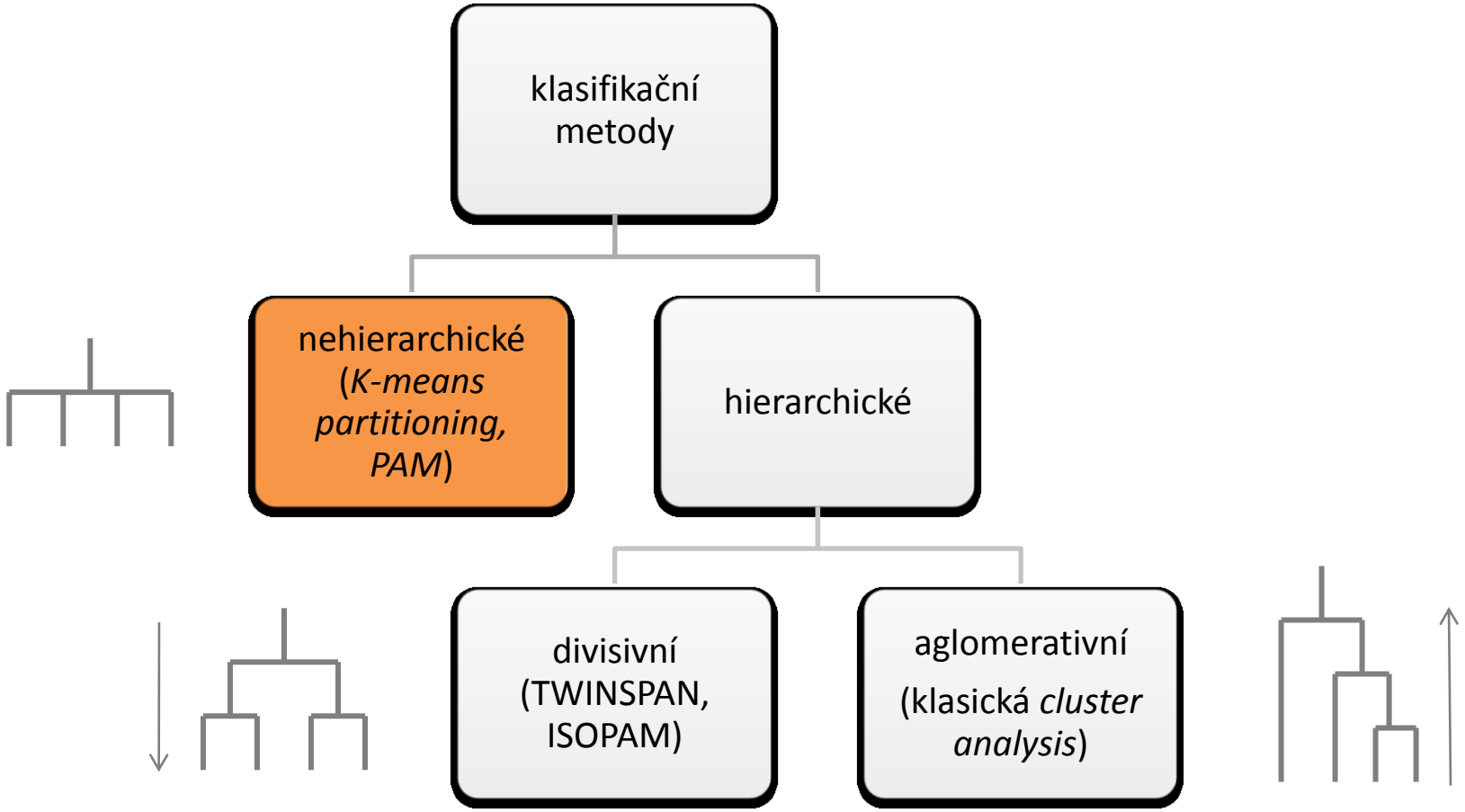
Otázky, které bych si měl položit před tím, než začnu něco klasifikovat

- **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

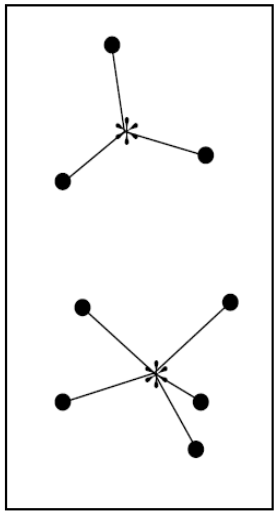
SYSTÉM KLASIFIKAČNÍCH METOD



KLASIFIKACE



K-MEANS PARTITIONING



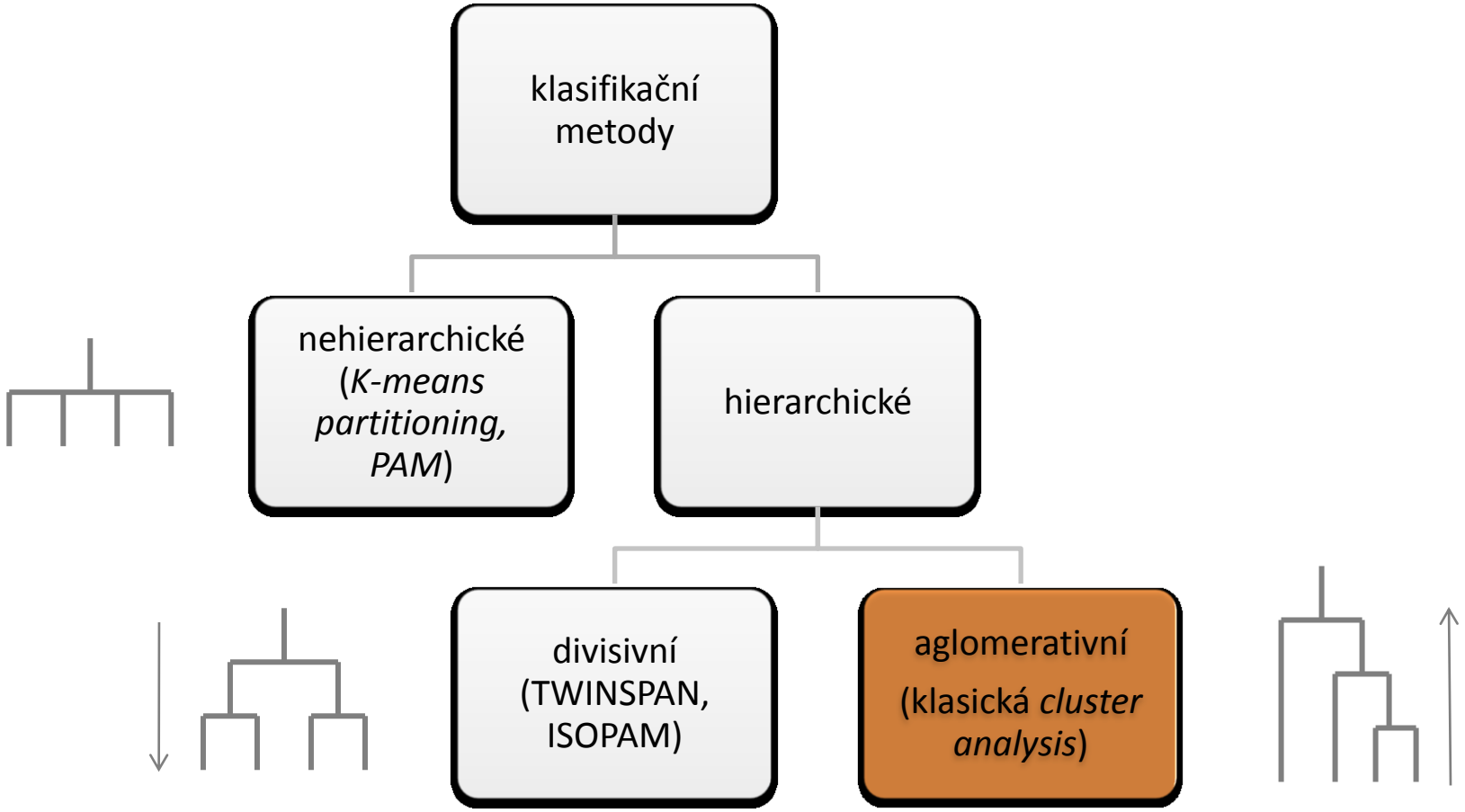
Legendre & Legendre 1998

- minimalizuje sumy čtverců vzdáleností vzorků od centr shluku
  - Vyžaduje metrické nepodobnosti
  - Sørensenovým (Bray-Curtis) indexem nepodobnosti třeba odmocnit
  - Nelze-li použít metrické d., lze použít PCoA osy
- na začátku uživatel zvolí počet shluků (*k*)
  - Analýza se obvykle zkouší pro nějaký rozsah k (k = 2 – xx)
  - Na základě této zkoušky se vybere vhodné k – subjektivně nebo na dákladě diagnostiky
- 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

PARTITIONING AROUND MEDOIDS - PAM

- Obdoba *k*-means
- Místo centroidů se shluky staví okolo konkrétních bodů (= reprezentativních pozorování, medoidů) v datasetu
- Cílem nalézt rozdělení do skupin, které minimalizuje sumu vzdáleností mezi medoidy a jednotlivými pozorováními
- Řešení obvykle stabilnější než k-means (to nutně neznamená lepší)
- Umožňuje pracovat s libovolnými nepodobnostmi

KLASIFIKACE



Proč hierarchie?

- Nehierarchické klasifikační metody dovedou dobře popsat shluky podél jednoho nebo dvou gradientů
  - Klasifikace po celou dobu uvažuje vztahy se všemi vzorky v datasetu
- Variabilita ve složení společenstev je často složitější
  - Např. na první úrovni les-bezlesí
  - Dále gradienty v lese a bezlesí, které mohou fungovat jinak
- Hierarchie umožňuje zacílení na menší podsoubor v rámci datasetu, přičemž ostatní vzorky jsou ignorovány
- Ordinační osy jsou taky hierarchické

KLASIFIKACE

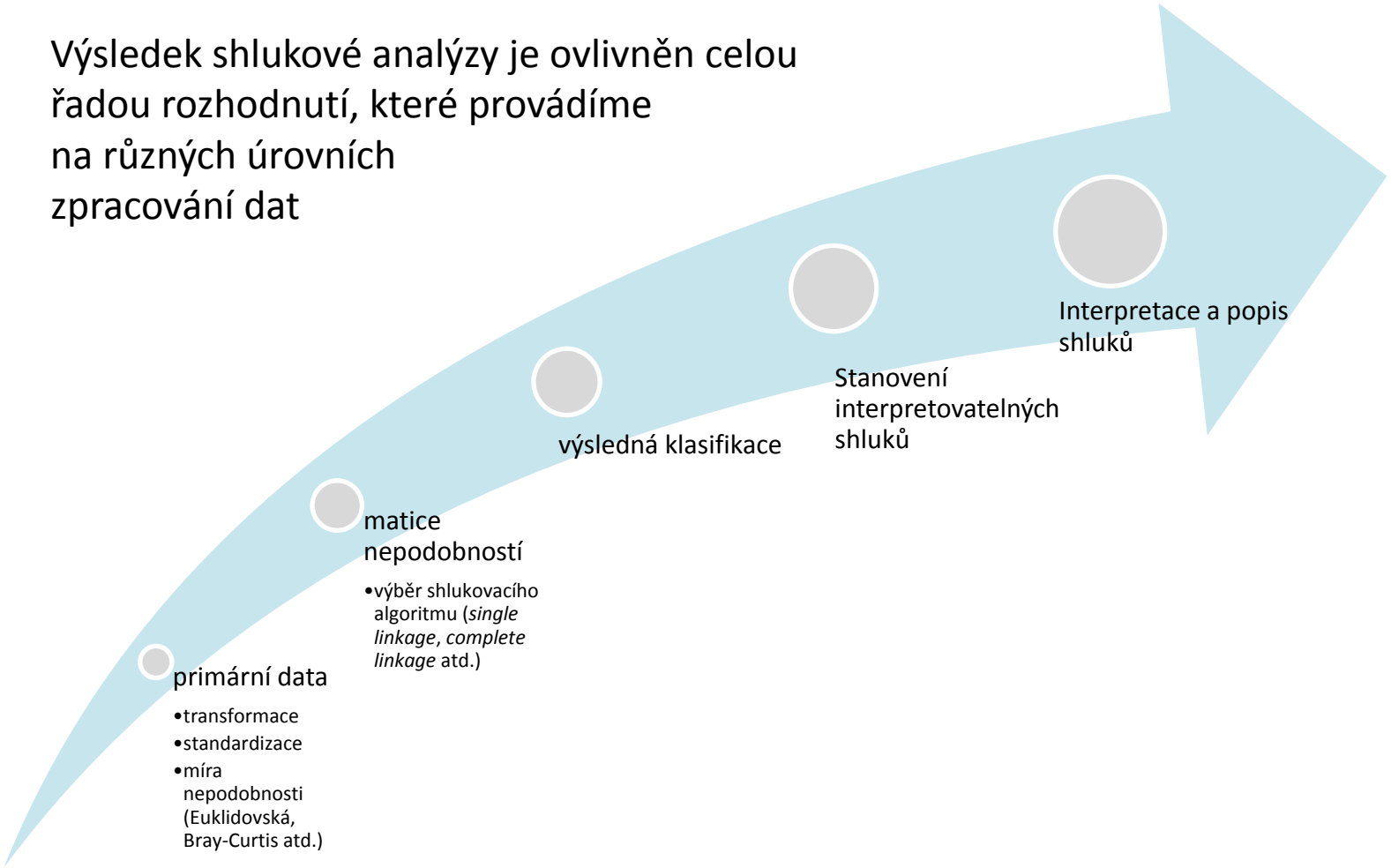
HIERARCHICKÁ A **AGLOMERATIVNÍ**

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*)
  - Definice interpretovatelných shluků na dendrogramu

SHLUKOVÁ ANALÝZA (*CLUSTER ANALYSIS*)

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



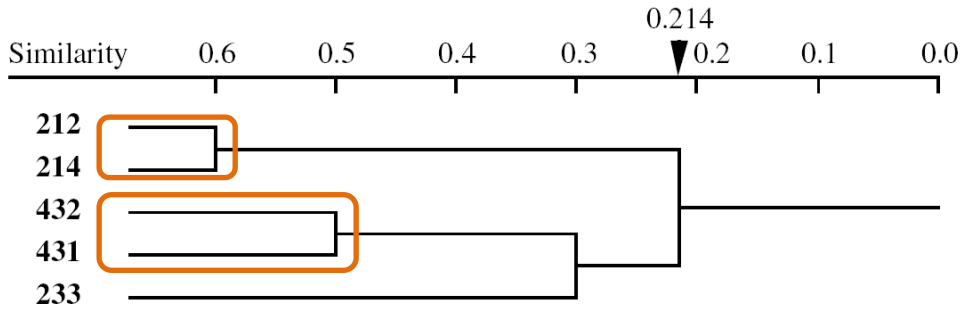
SHLUKOVÁ ANALÝZA (*CLUSTER ANALYSIS*)

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*)

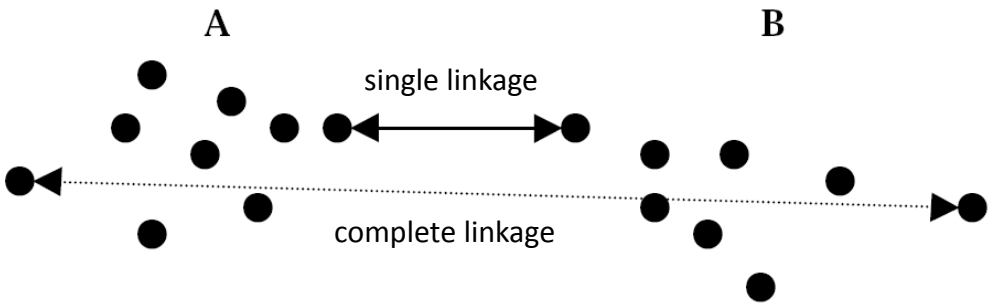
SHLUKOVACÍ ALGORITMY

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*)

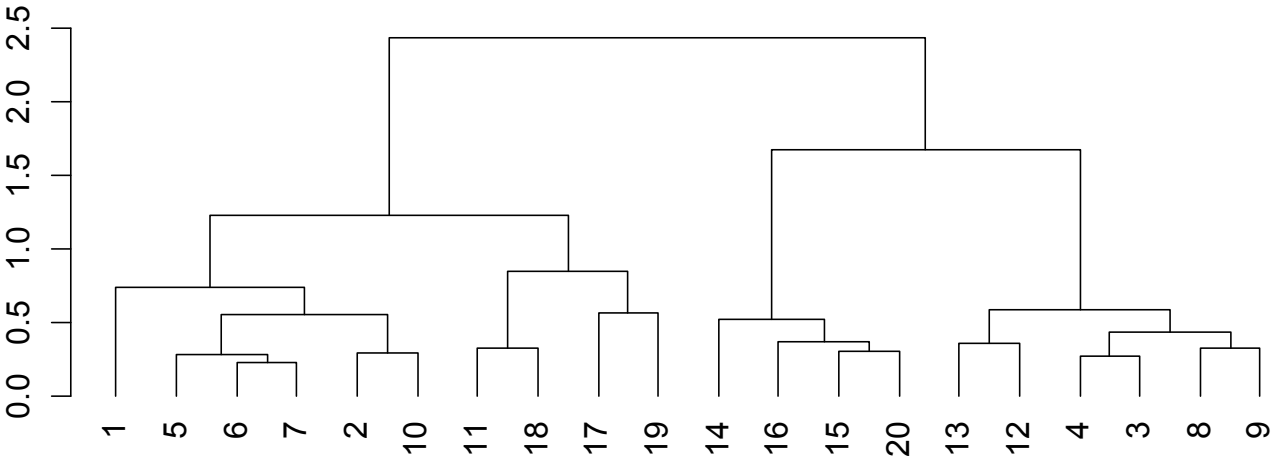
- vzorky se připojí ke shluku až v okamžiku, kdy shluk obsahuje všechny podobné vzorky
- *zjistím nejnesympatičtější jedince ve všech sjkupinách a přidám se ke skupině ve které je ten nejmíň nesympatický*



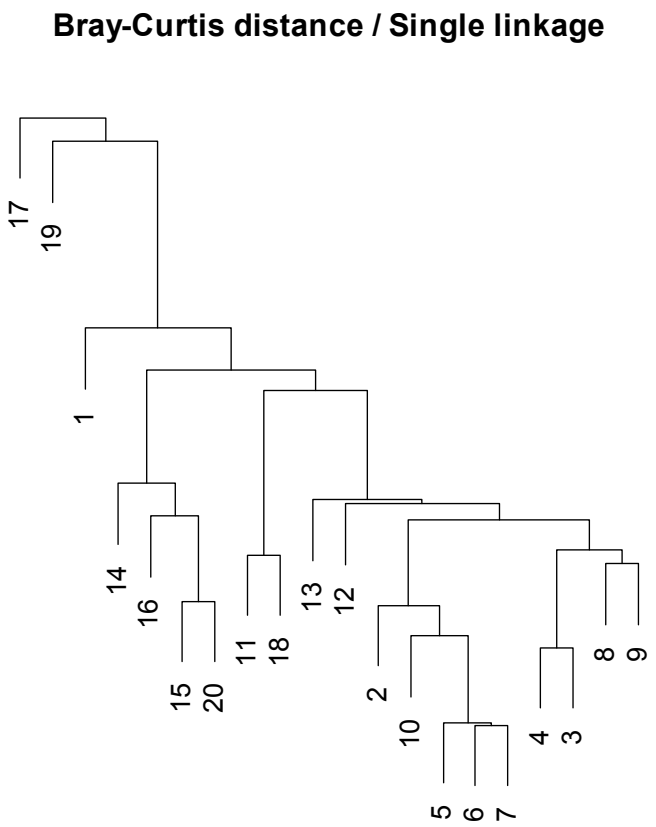
SHLUKOVÁ ANALÝZA (*CLUSTER ANALYSIS*)

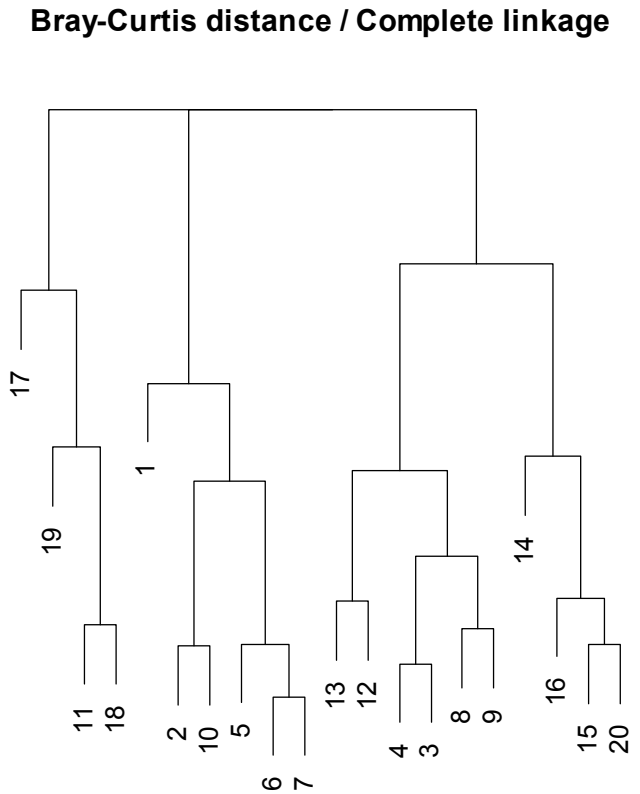
DENDROGRAM

- 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





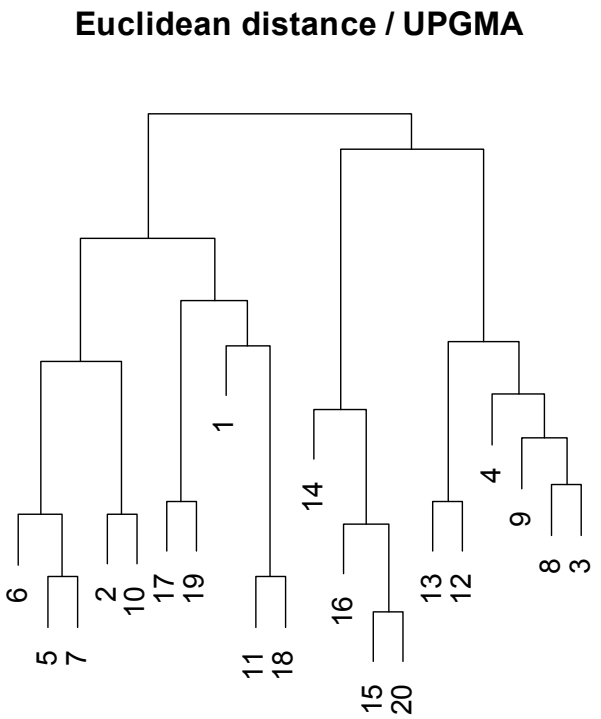
metoda jednospojná se výrazně řetězí

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*

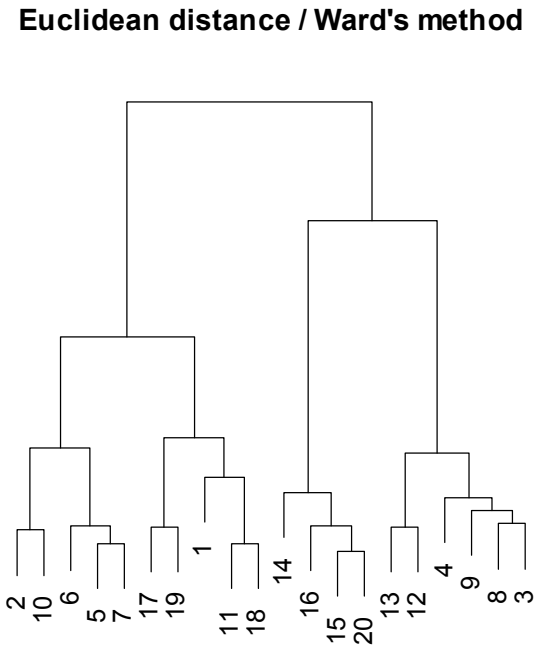


SHLUKOVÁ ANALÝZA (*CLUSTER ANALYSIS*)

SHLUKOVACÍ ALGORITMY

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

- minimalizuje součet čtverců vzdáleností mezi vzorky a centroidy jejich shluků
  - Vyžaduje metrické distance (spočte se i s jinými, ale výsledek je diskutabilní)
  - Se Sørensenovým (Bray-Curtis) indexem pouze po odmocnění nepodobností
- 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
  - To obvykle chceme



Stanovení interpretovatelných shluků

- Důležitá je hrubá topologie dendrogramu, ne detaily na koncích větví
- Shluky stanovíme „seříznutím konců větví“
  - Buď definujeme k (počet shluků)
  - Nebo výšku dendrogramu, kde se provede řez a podle toho se definují shluky

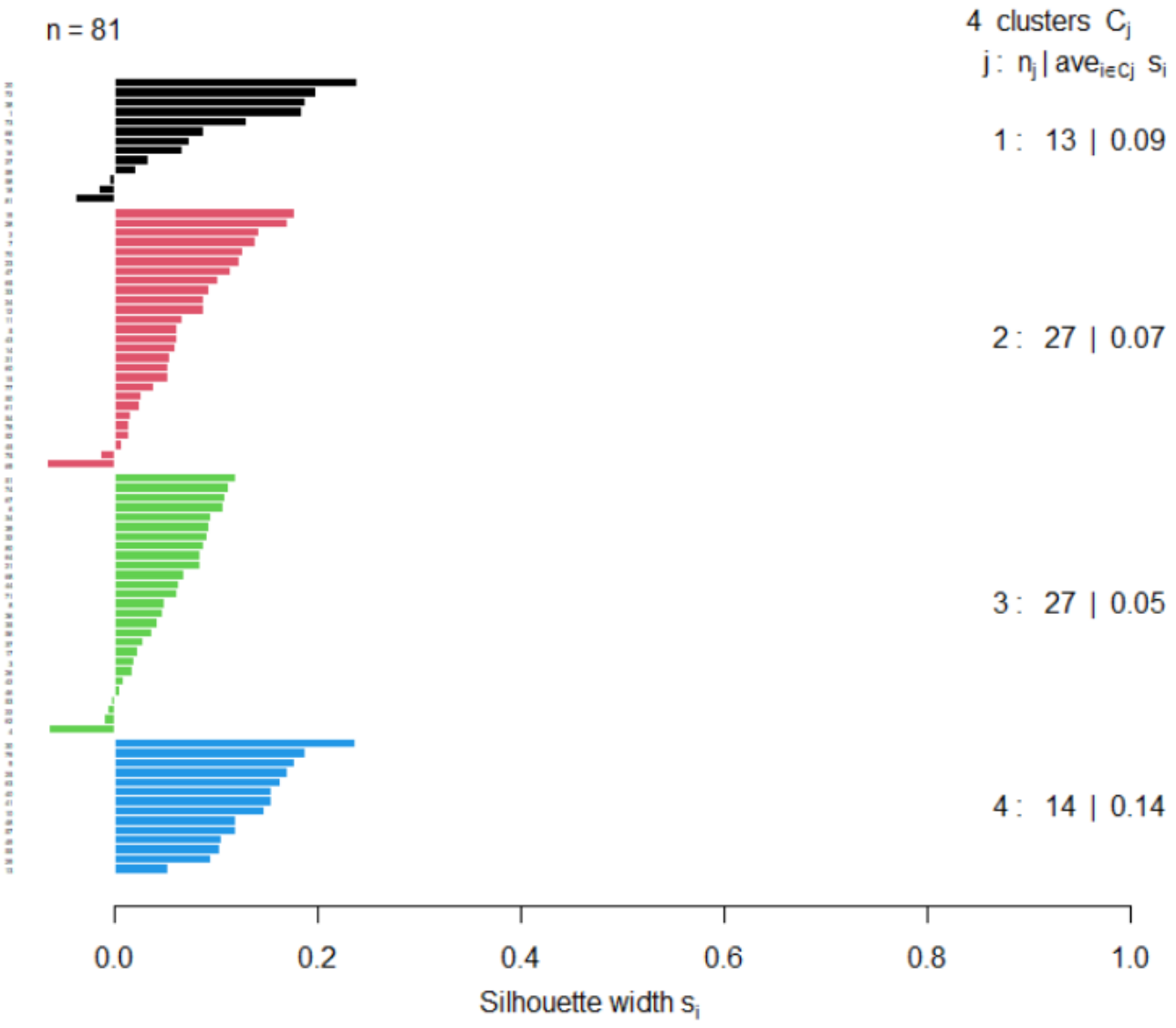
Diagnostika klasifikací

- Informuje o kvalitě klasifikace
- Umožňuje stanovit k v k-means a PAM
- Umožňuje stanovit interpretovatelné shluky v hierarchických metodách
- Přístupů je řada
  - Shoda mezi příslušností do shluků a nepodobností v původní matici
  - Analýza indikačních druhů

Diagnostika pomocí šířky siluety (silhouette width)

- Klíčový parametr šířka siluety (silhouette width)
  - Definovaný pro jednotlivé body
  - s = (b – a) / max (a, b)
    - a – průměrná nepodobnost mezi daným bodem a dalšími body ve **shluku kam patří**
    - b – průměrná nepodobnost mezi daným bodem a **sousedním** shlukem (kam daný bod nepatří)
  - S = 1: ideální klasifikace (bod leží ve středu svého shluku)
  - S = +- 0: hraniční body; S = 0, je-li bod ve shluku sám
  - S < 0: nesprávně klasifikované body (mají blíž k jinému shluku než ke svému)
- Průměrná SW charakterizuje celkovou kvalitu celé klasifikace
  - Lze porovnávat různá k nebo počty shluků v hclust
  - Lze porovnávat různé metody (např. PAM vs. hclust), i třeba různé indexy nepodobnosti.
- Velmi univerzální metoda

SILHOUETTE PLOT



Average silhouette width : 0.08

POPIS VLASTNOSTÍ SHLUKŮ

- Boxploty
- Jednocestná ANOVA

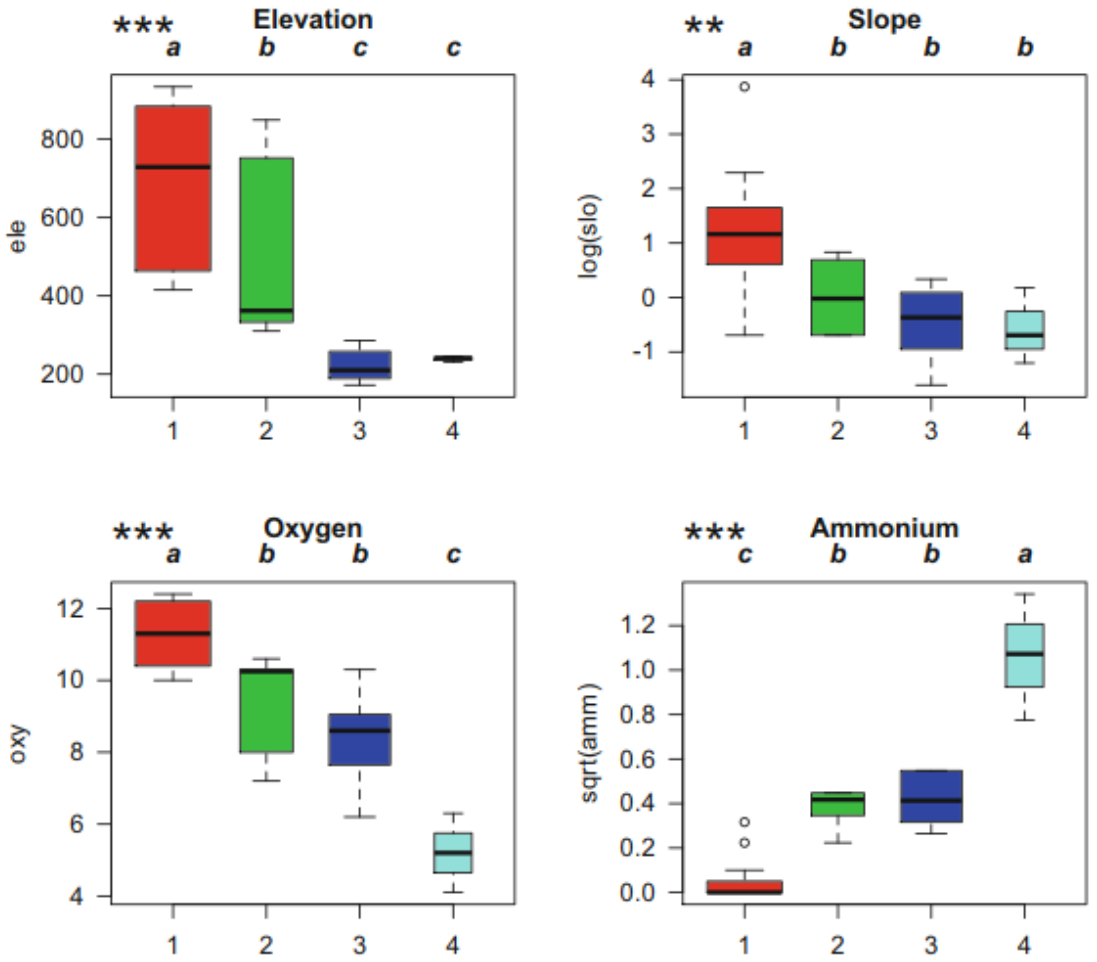


Fig. 4.24 Boxplots of four environmental variables grouped according to the four species-based groups from the optimized Ward clustering. Stars indicate the significance of the differences among groups for each environmental variable. The letters indicate which group means are significantly different

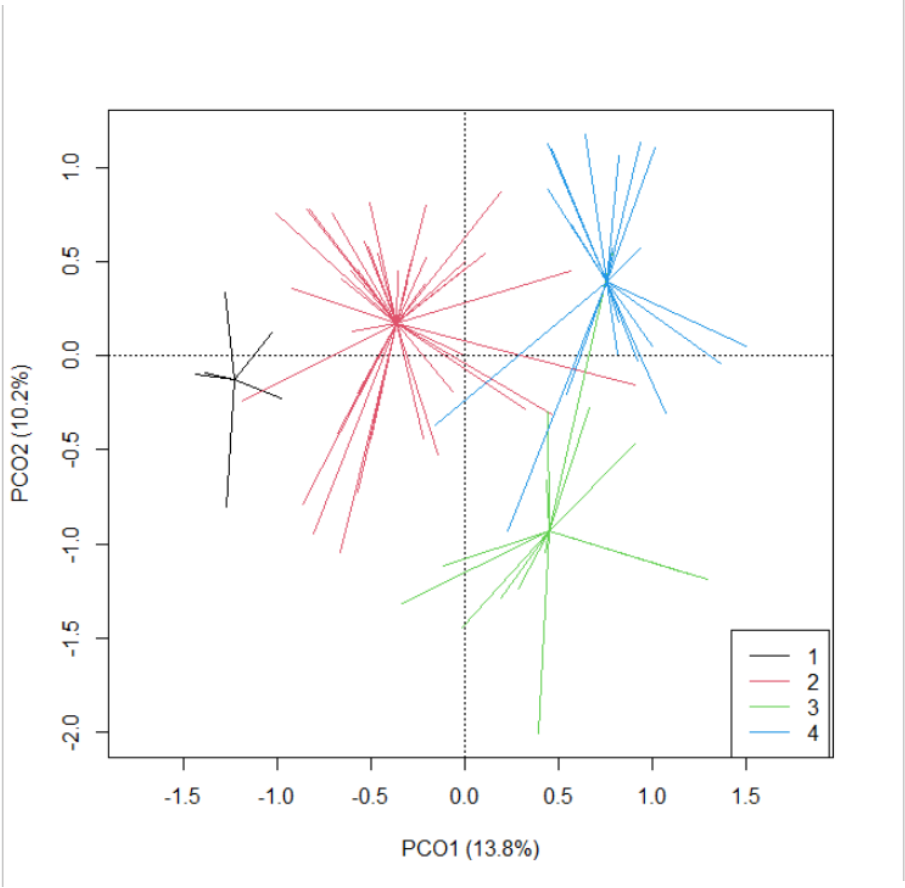
Analýza diagnostických druhů

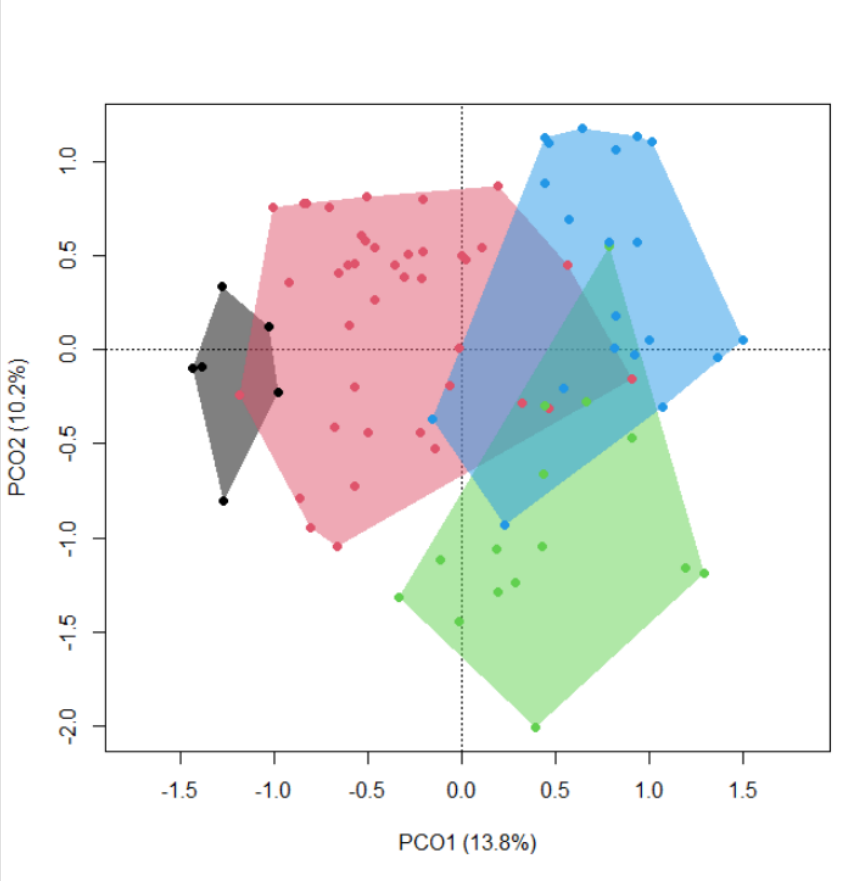
- Korelace druhů se shluky
  - Např. phi-coeficient (= Pearson r pro 0/1 data)
- Test signifikance
  - Fisher exact
  - Permutační
  - P-hodnoty by se měly upravit kvůli mmnohonásobnému prorovnání

```
Association function: r.g
Significance level (alpha): 0.05

Total number of species: 117
Selected number of species: 42
Number of species associated to 1 group: 42
Number of species associated to 2 groups: 0
Number of species associated to 3 groups: 0

List of species associated to each combination:

 Group 1  #sps.  11
                        stat p.value
Sinapis alba           0.609   0.005 **
Fallopia convolvulus   0.588   0.005 **
Phacelia tanacetifolia 0.553   0.005 **
Chenopodium album      0.479   0.005 **
Camelina sativa        0.478   0.005 **
Linaria vulgaris       0.458   0.005 **
Festuca sp.            0.429   0.010 **
Atriplex patula        0.346   0.045 *
Digitaria ischaemum    0.346   0.030 *
Polygonum aviculare    0.307   0.005 **
Pisum sativum          0.296   0.045 *

 Group 2  #sps.  10
                        stat p.value
Festuca rubra          0.501   0.005 **
```

PROMÍTNUTÍ VÝSLEDKŮ NUMERICKÉ KLASIFIKACE DO ORDINAČNÍHO DIAGRAMU

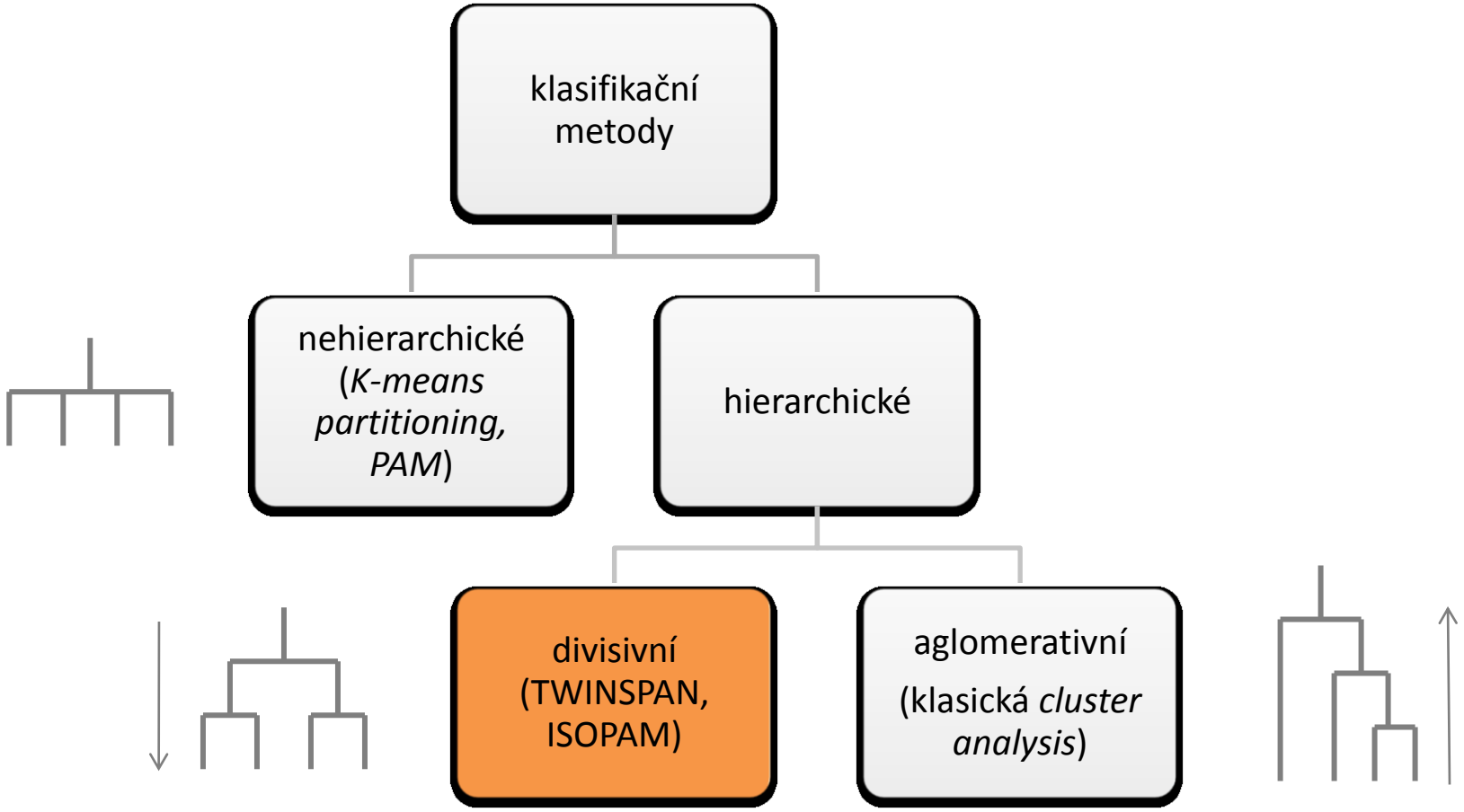
PCoA (Bray-Curtis) + Hclust (Ward-sqrt(Bray-Curtis))





Je vhodné, aby míra nepodobnosti mezi vzorky byla v obou metodách (numerické klasifikaci i ordinační analýze) stejná.

KLASIFIKACE



... PŘÍŠTĚ