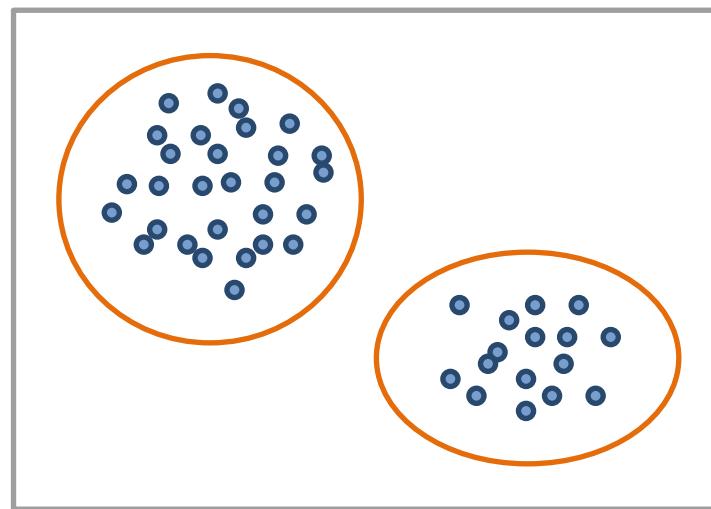


NUMERICKÁ KLASIFIKACE

SHLUKOVÁNÍ

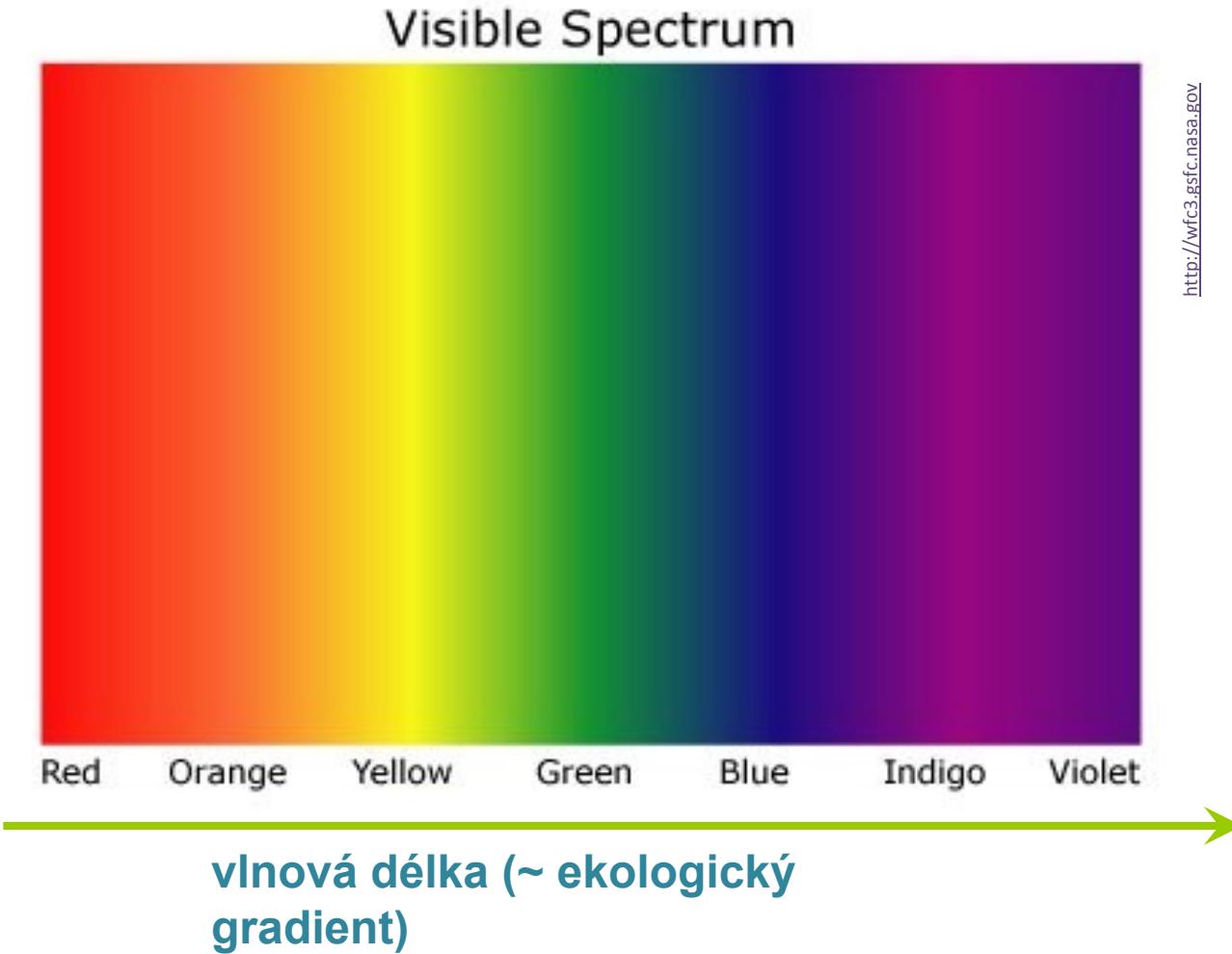
- rozpoznání objektů, které jsou si dostatečně podobné, aby mohly být dány do stejné skupiny
- zjištění odlišností mezi skupinami



DISKONTINUUM VS. KONTINUUM

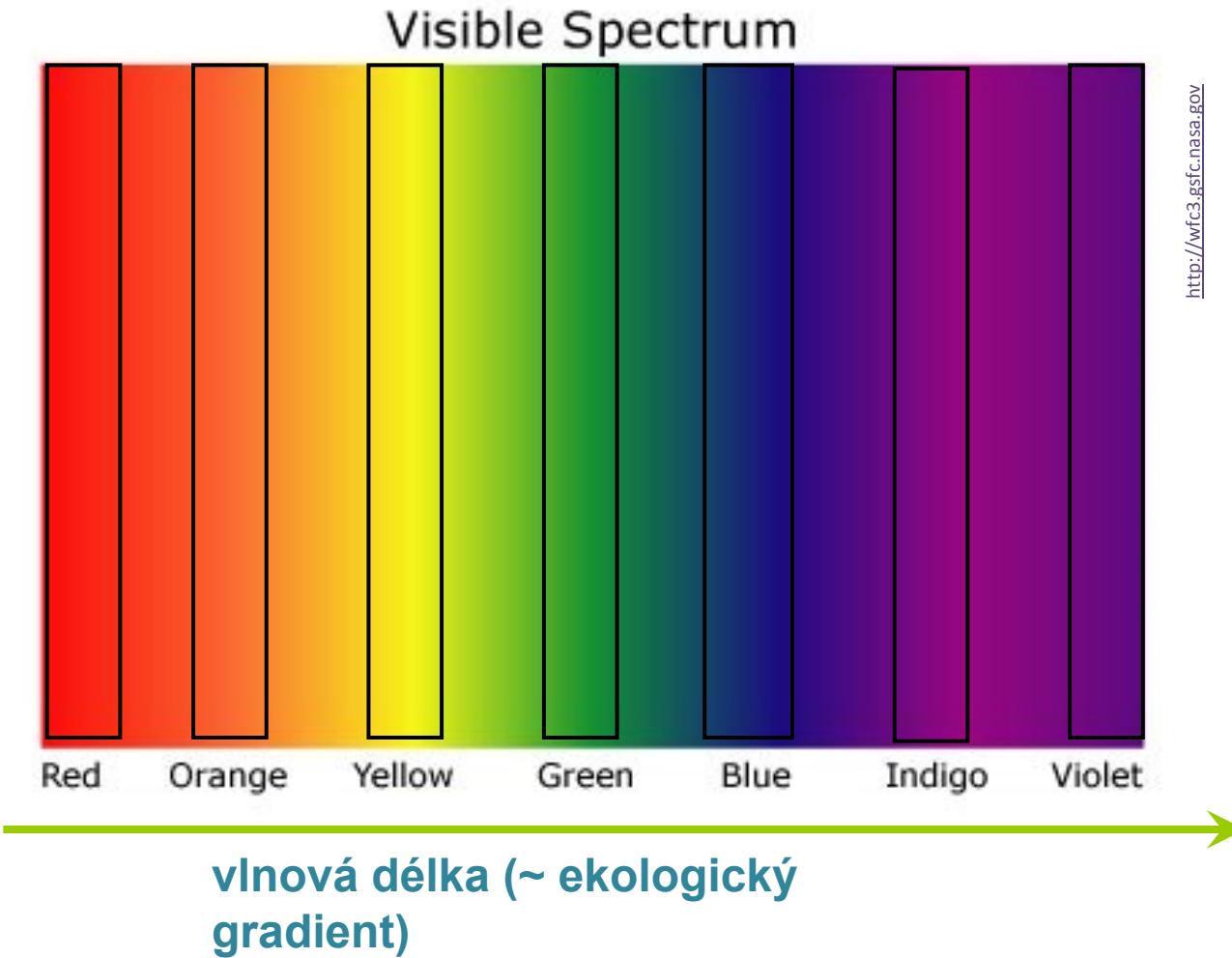
- Evoluční teorie predikuje diskontinuum – druhy
 - taxonomové hledají diskontinuity dané odlišnostmi mezi druhy
- Svět ekologie nejčastěji kontinuální
 - metody schopné rozpoznat shluky podobných objektů, zatímco ignorují několik hraničních
- Nelze očekávat diskontinuity ve společenstvech, aniž by prostředí bylo diskontinuální (nebo nevzorkujeme opačné konce gradientů) Whittaker 1962

PROČ MÁ SMYSL VĚCI KLASIFIKOVAT?

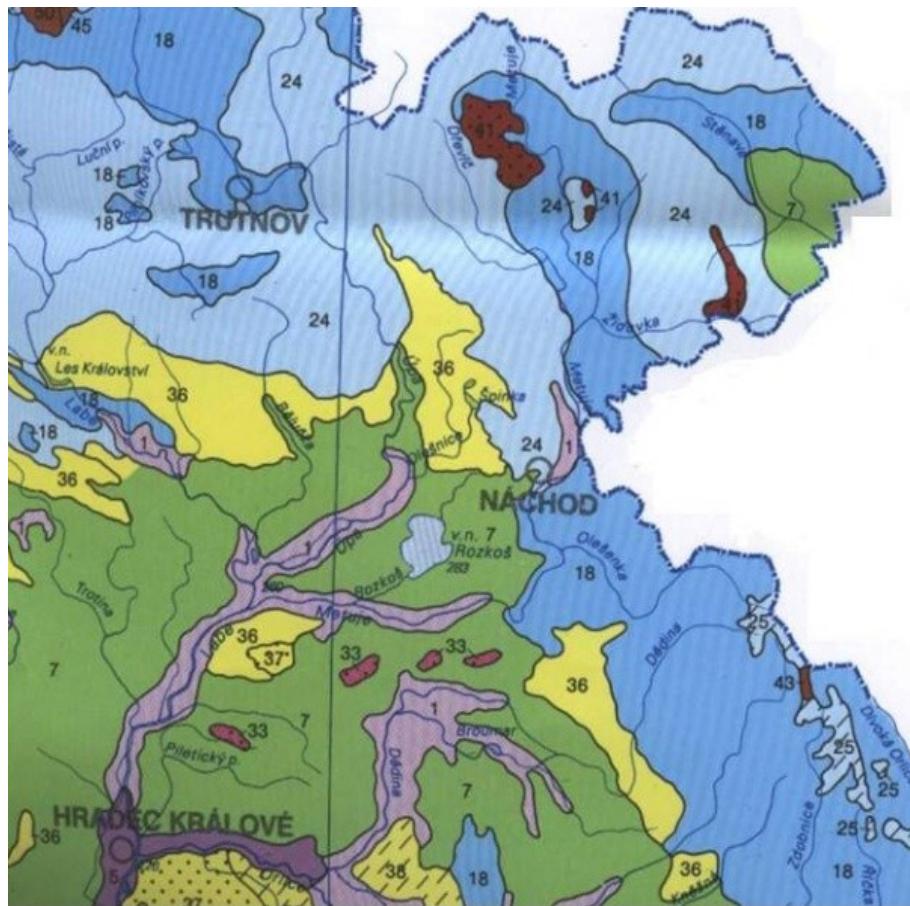


<http://wfc3.gsfc.nasa.gov>

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 kyčelníci 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 carpaticaе-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*)

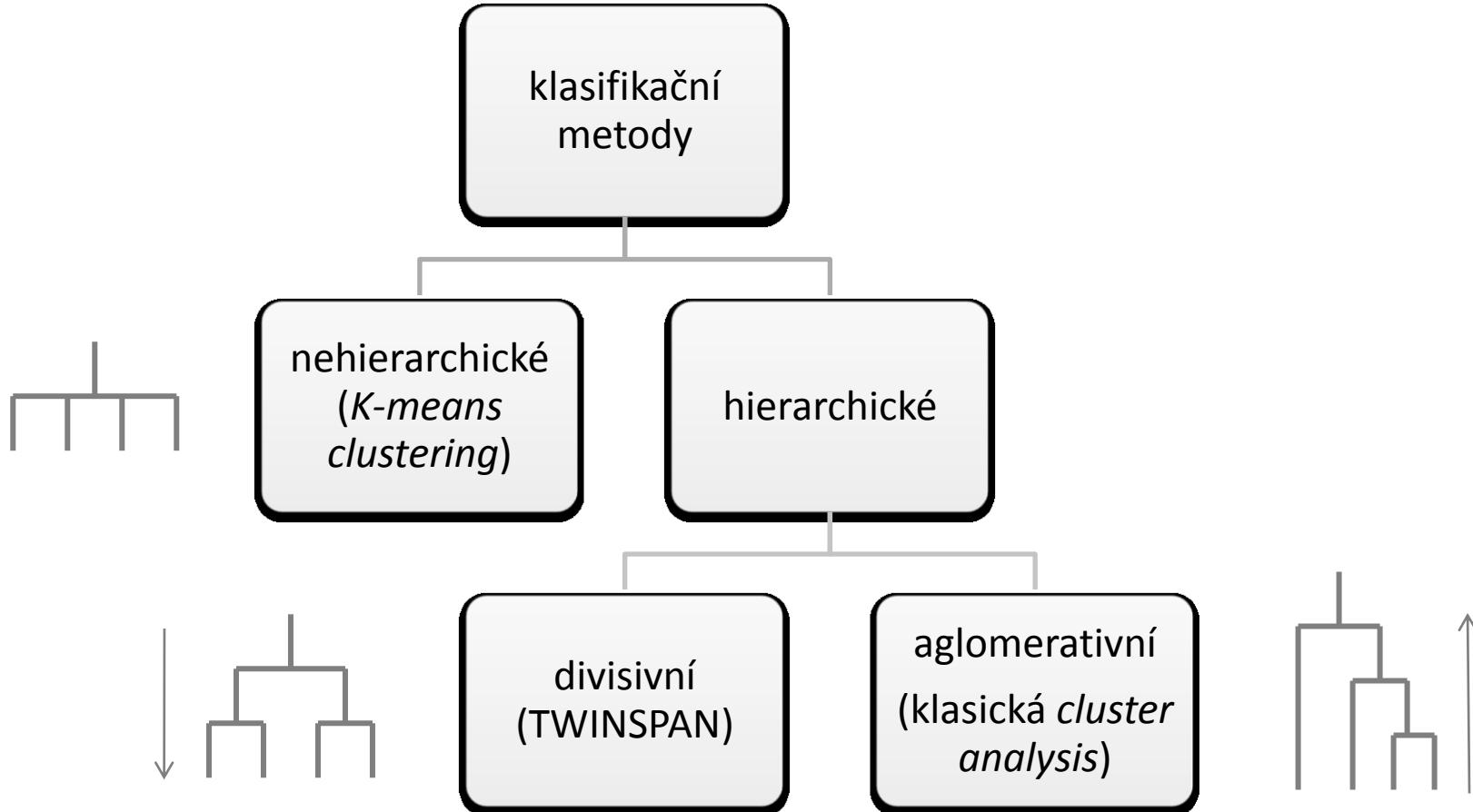
○ **Podle jakých kritérií budu objekty klasifikovat?**

- 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

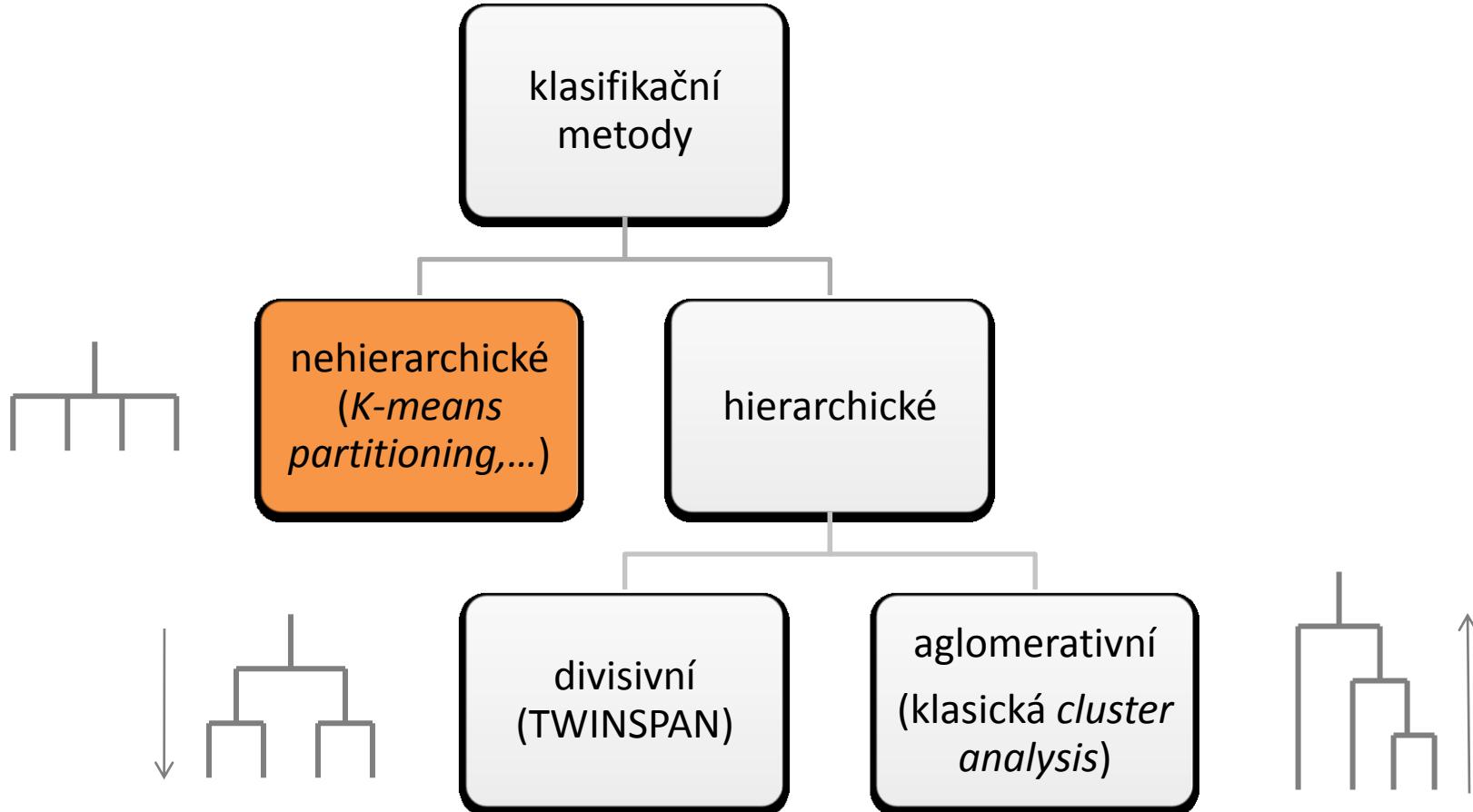
○ **Jak stanovím hranice mezi jednotlivými skupinami?**

- pravidla, podle kterých budu přiřazovat objekty do skupin
- odpovídá výběru klasifikačního algoritmu

KLASIFIKAČE



KLASIFIKAČE



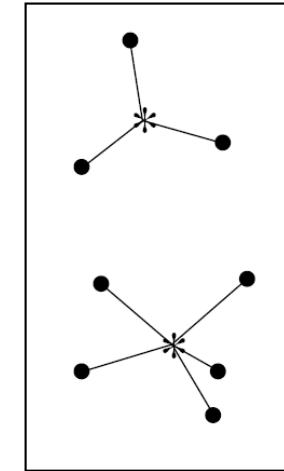
KLASIFIKACE

NEHIERARCHICKÁ

K-means partitioning

(shlukování metodou K-průměrů)

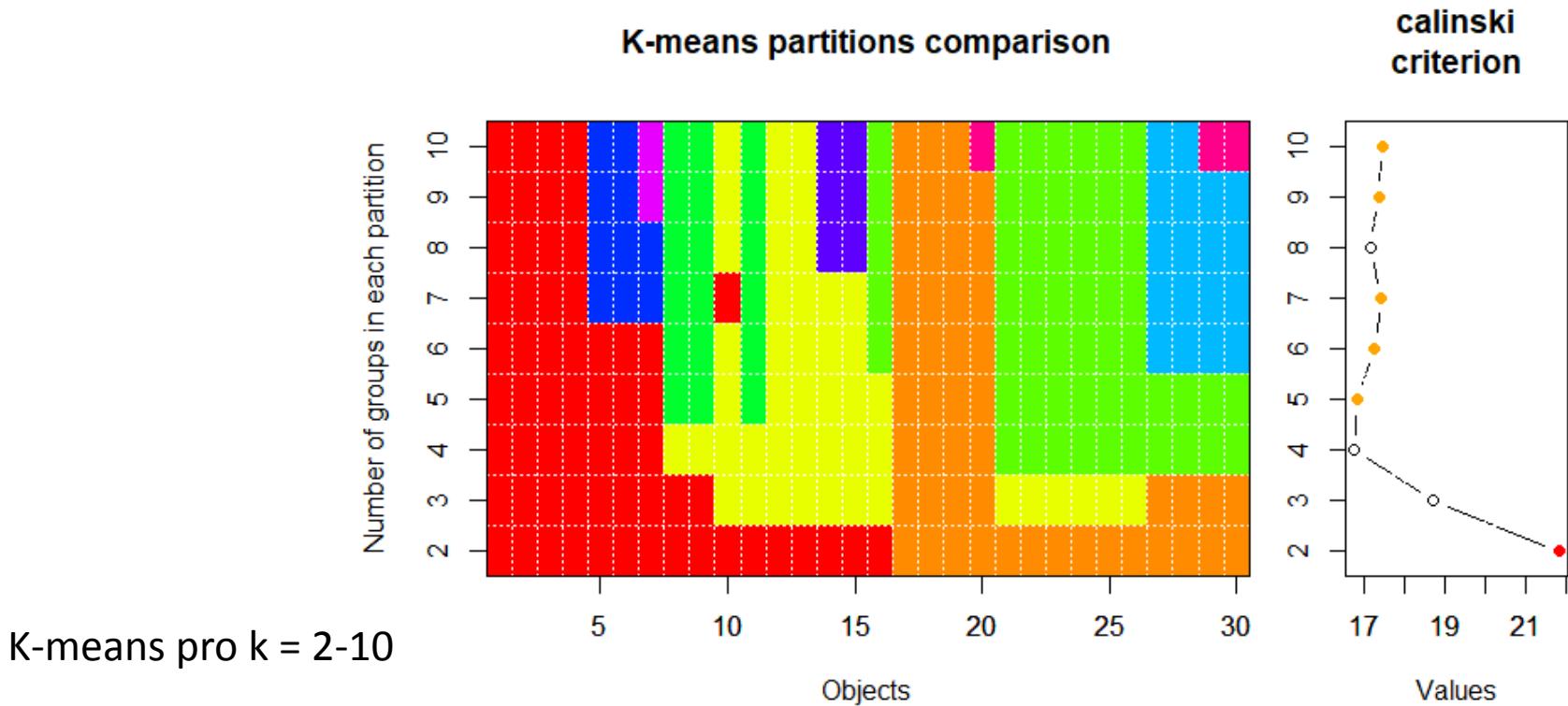
- nehierarchická metoda – všechny shluky jsou si rovny
- minimalizuje sumy čtverců vzdáleností vzorků od centroidů shluku
- na začátku uživatel zvolí počet shluků (k)
- iterativní metoda, začne od náhodného přiřazení vzorků do shluků, postupně přehazuje vzorky mezi shluky a hledá optimální řešení
- výsledek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



Legendre & Legendre 1998

IDENTIFIKACE "SPRÁVNÉHO K"

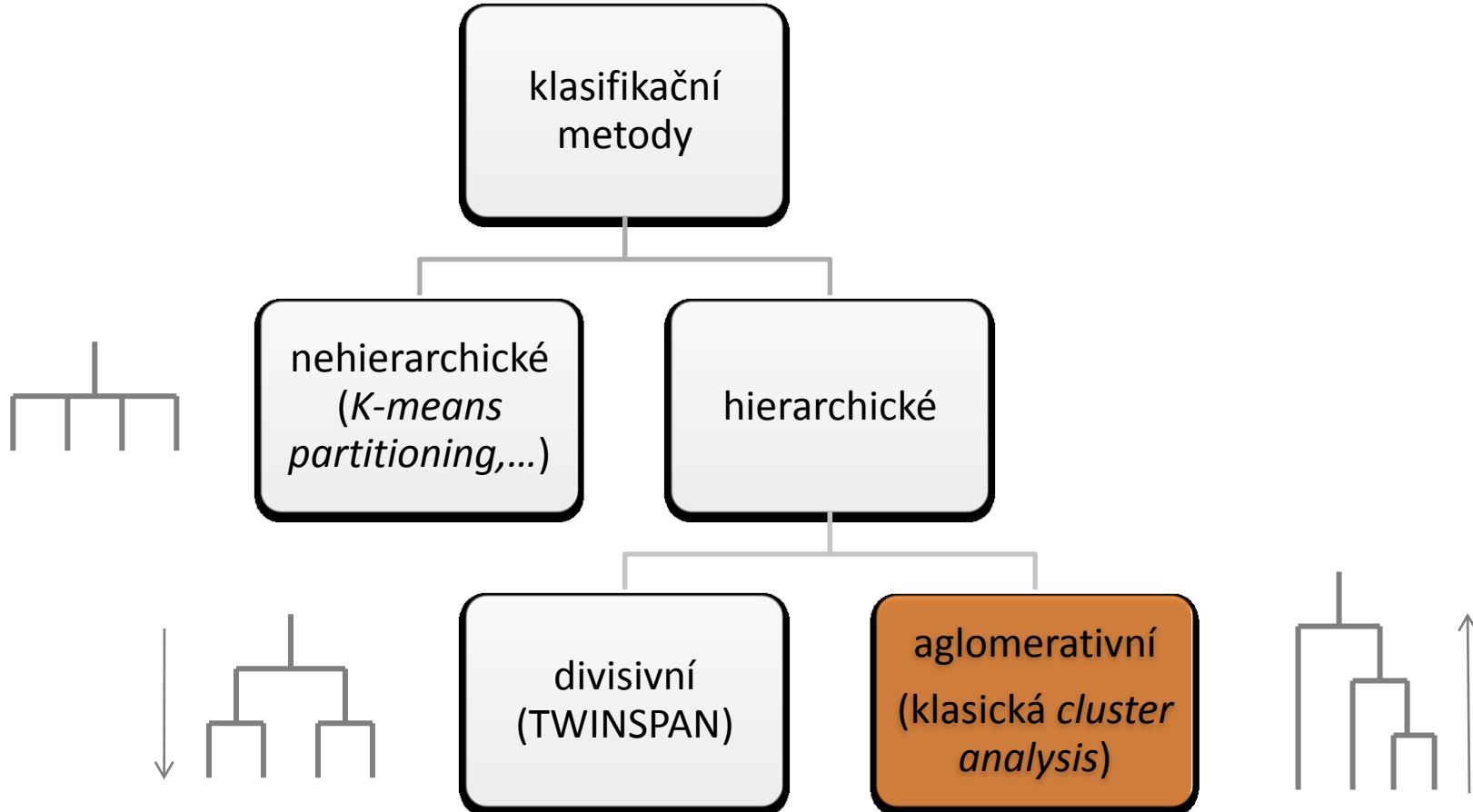
- Spuštění kmeans přes
- Calinski – Harabasz criterion ($\sim F\text{-ratio: } \text{MS}_{\text{mezi shluky}}/\text{MS}_{\text{uvnitř shluků}}$)
 - Nejvyšší hodnota \sim optimum
 - Hodnota K, při které C-H crit. Vzroste, může být taky zajímavá



DALŠÍ METODY PARTITIONING

- K-means
 - Limitované použití na metrické distance
 - Jinak nutné přepočítat nepodobnosti na vzdálenosti pomocí PCoA
- Partitioning around the medoids
 - Centrum není průměr, ale nějaký konkrétní bod (medoid)
 - Použitelné na jakoukoliv nepodobnost
 - Má být robustnější než k-means

KLASIFIKAČE



KLASIFIKACE

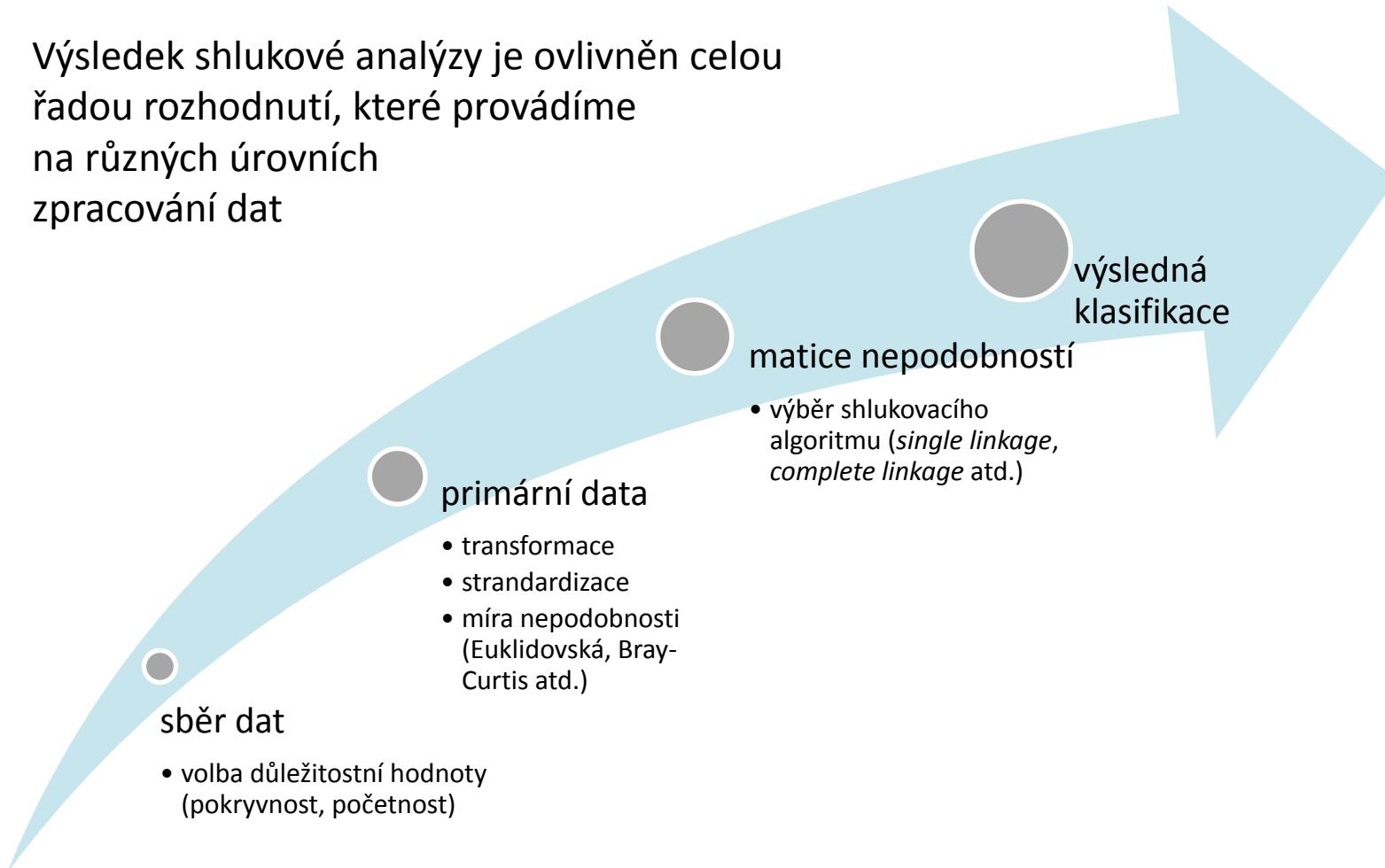
HIERARCHICKÁ A AGLOMERATIVNÍ

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



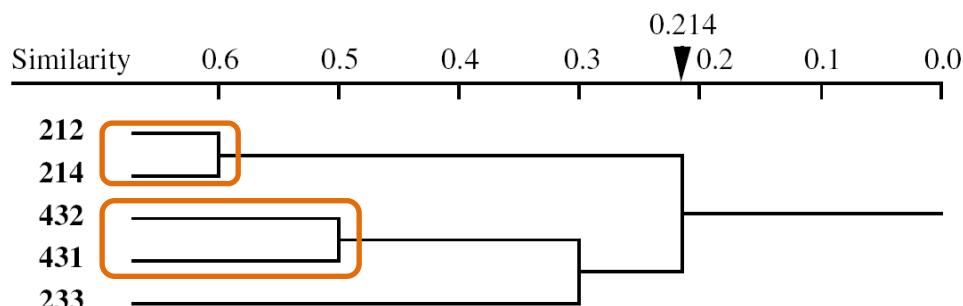
SHLUKOVÁ ANALÝZA (CLUSTER ANALYSIS)

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 podobnosti



výsledný dendrogram

páry vzorků seřazené podle podobnosti

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

SHLUKOVÁ ANALÝZA (CLUSTER ANALYSIS)

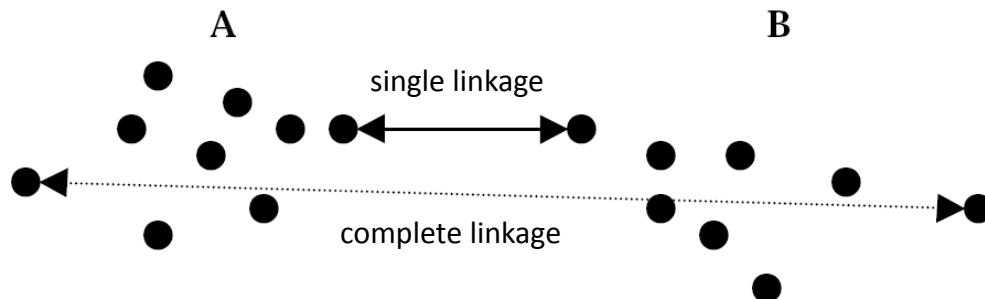
SHLUKOVACÍ ALGORITMY

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 nejvíce sympatický

Metoda všespojná (*complete linkage, farthest neighbour*)

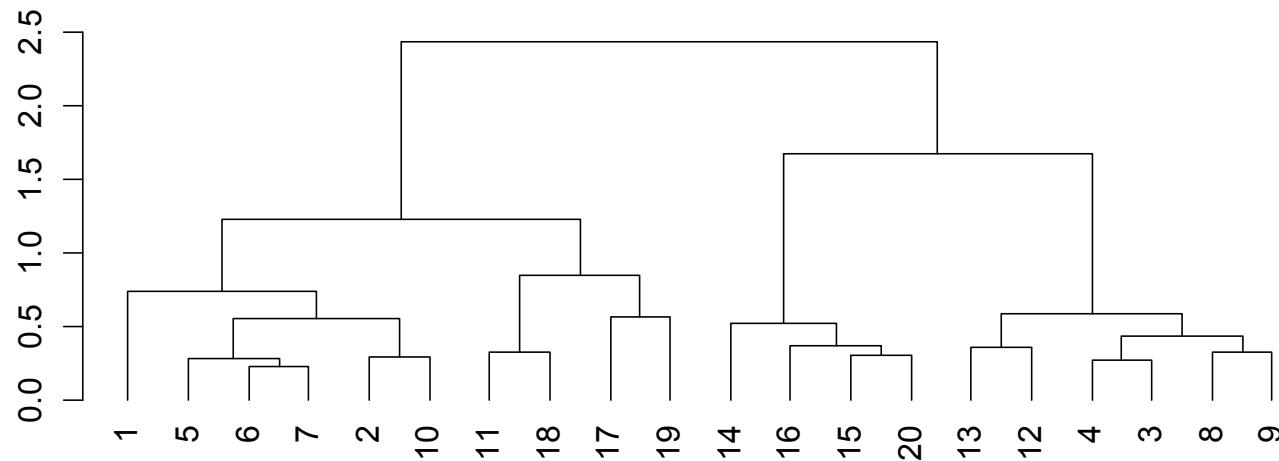
- vzorky se připojí ke shluku až v okamžiku, kdy shluk obsahuje všechny podobné vzorky
- zjistím nejnesympatičtější jedince ve všech skupinách a přidám se ke skupině ve které je ten nejmíň nesympatický



SHLUKOVÁ ANALÝZA (*CLUSTER ANALYSIS*)

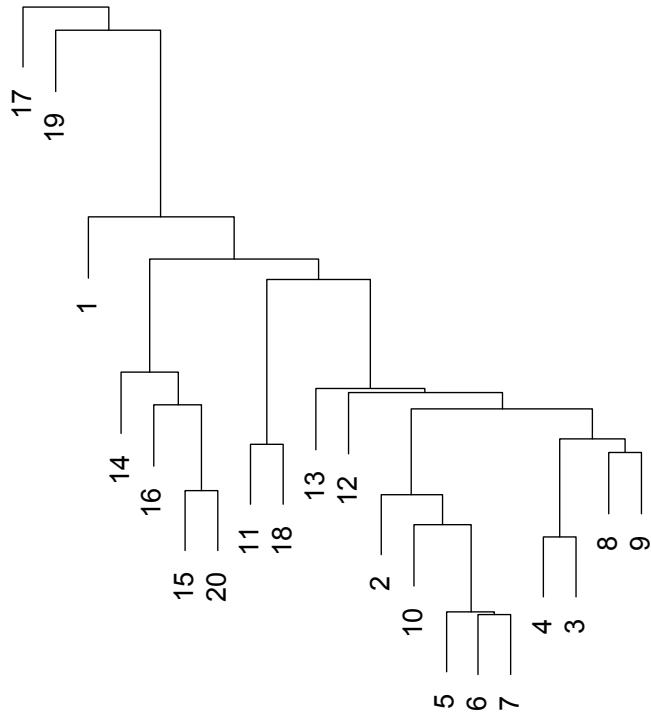
DENDROGRAM

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

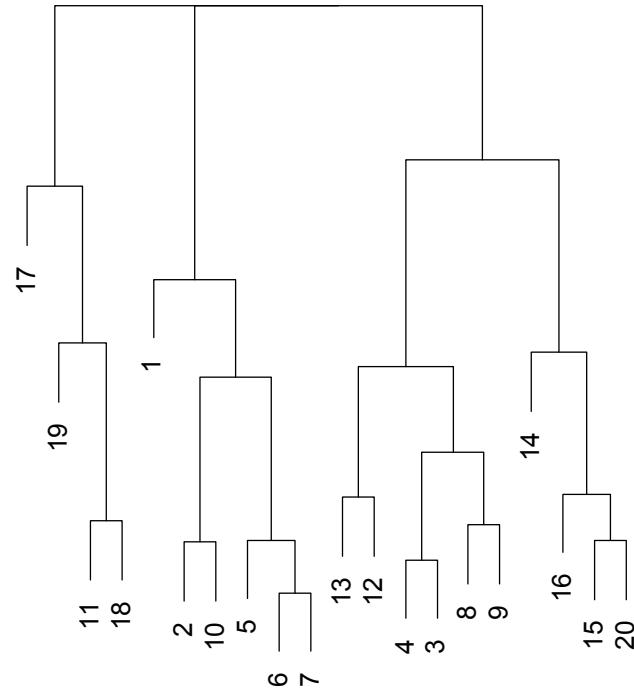


METODA JEDNOSPOJNÁ VS VŠESPOJNÁ

Bray-Curtis distance / Single linkage



Bray-Curtis distance / Complete linkage



metoda jednospojná se výrazně řetězí

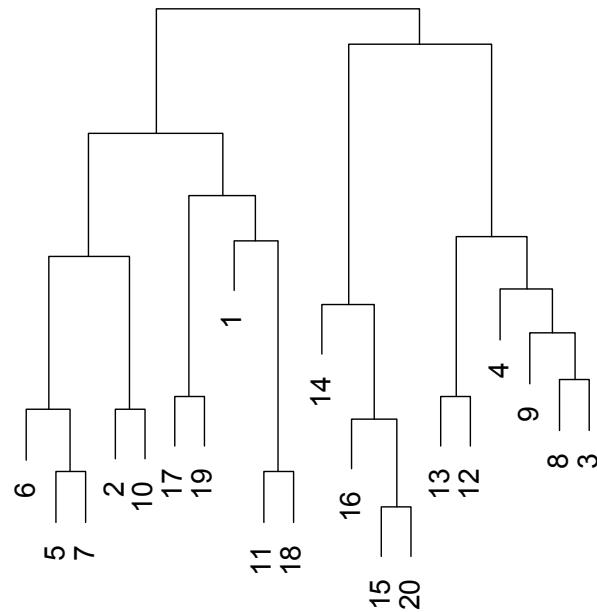
SHLUKOVÁ ANALÝZA (CLUSTER ANALYSIS)

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*

Euclidean distance / UPGMA



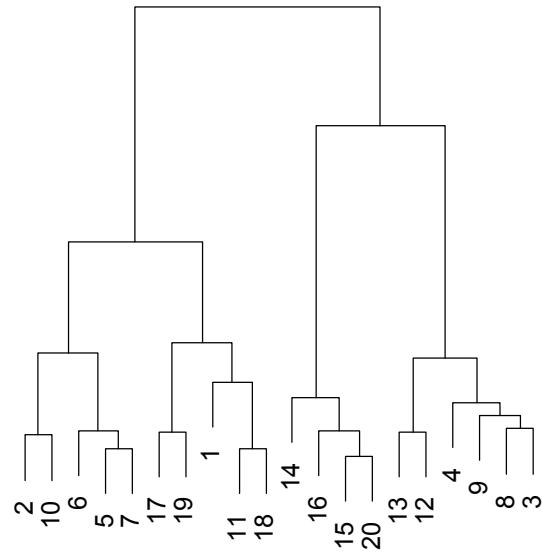
SHLUKOVÁ ANALÝZA (CLUSTER ANALYSIS)

SHLUKOVACÍ ALGORITMY

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

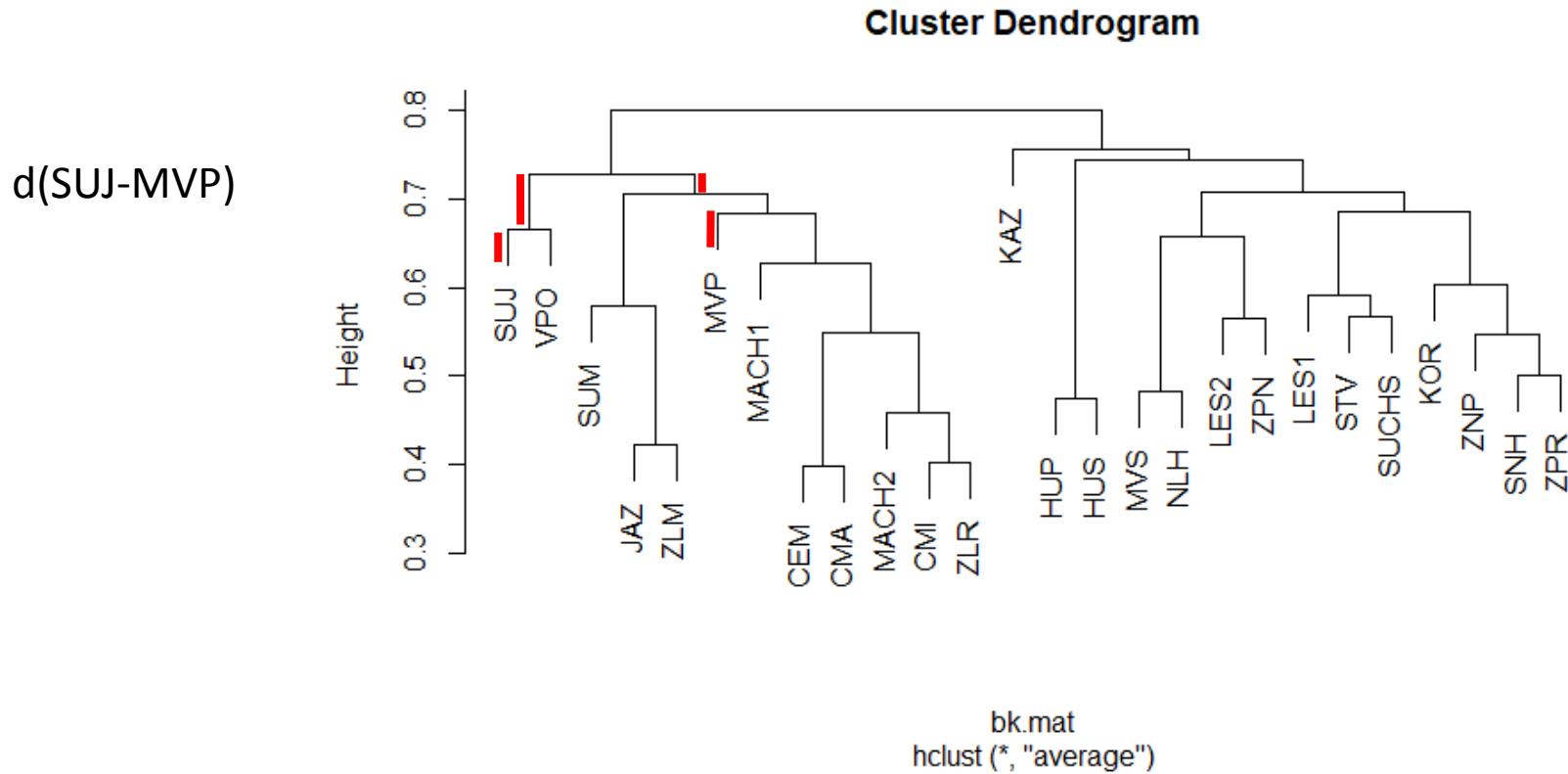
- minimalizuje součet čterců vzdáleností mezi vzorky a centroidy jejich shluků
- jsou spojovány ty shluky (vzorky) jejichž shluknutí povede k nejmenšímu nárůstu součtu čtverců vnitroshlukových vzdáleností
- výsledné shluky mají tendenci být hypersférické a zhruba stejné velikosti
- neměla by se kombinovat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

Euclidean distance / Ward's method

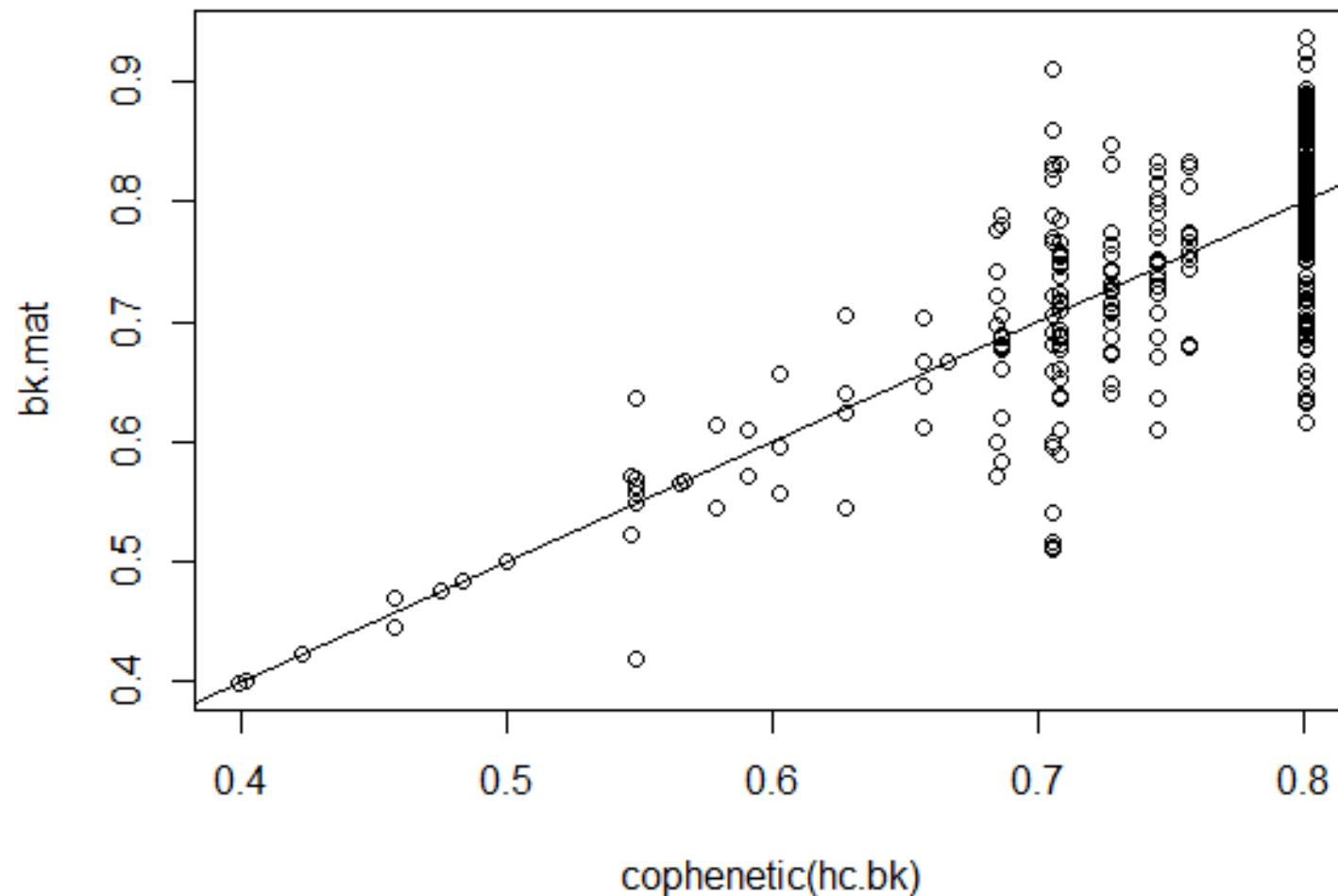


KOFENETICKÁ VZDÁLENOST

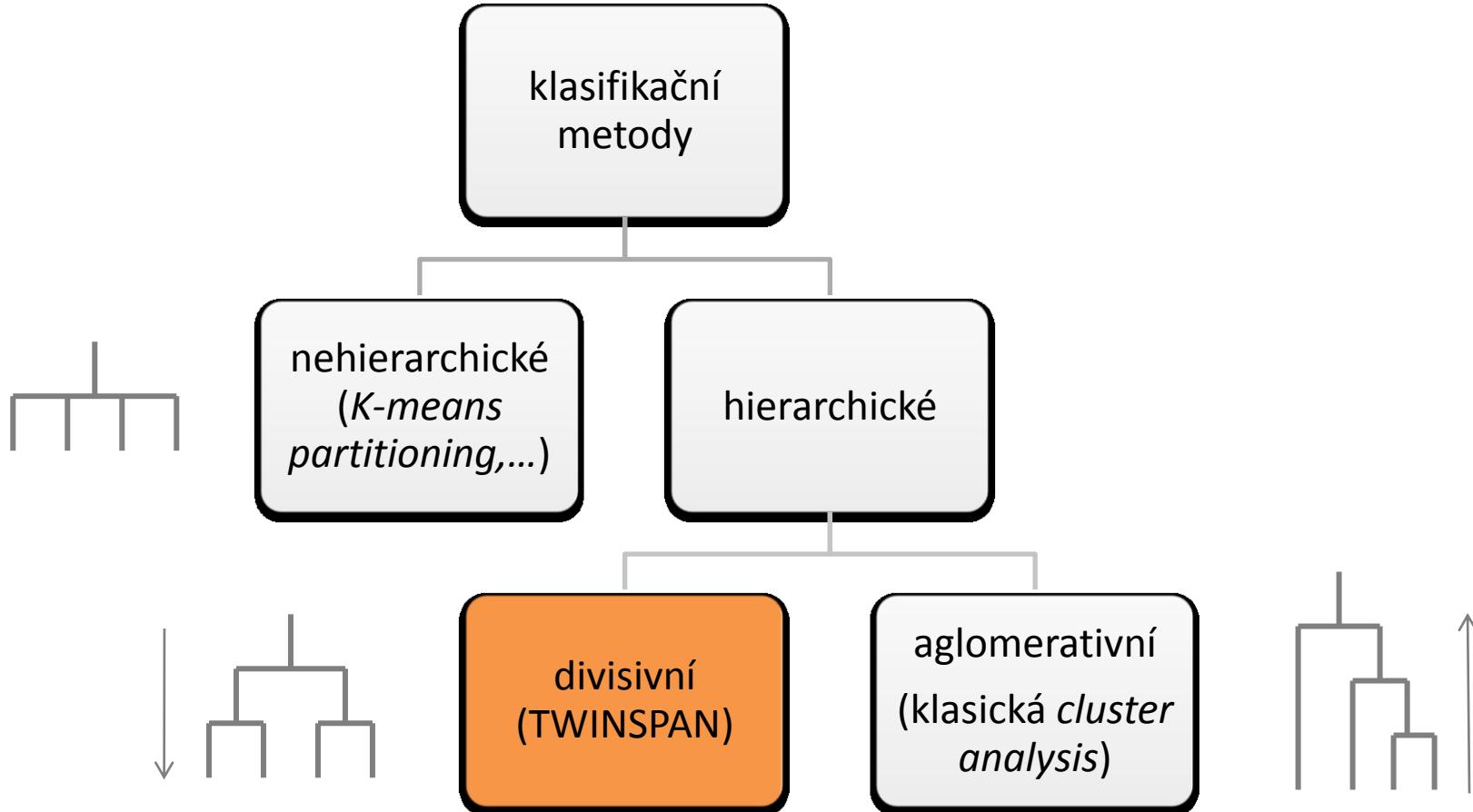
- Vzdálenost mezi dvěma vzorky definovaná jako nepodobnost v rámci skupiny v níž jsou dva vzorky spojené do jednoho klastru



VZTAH MEZI ORIGINÁLNÍ NEPODOBOSTÍ A KOFENETICKOU VZDÁLENOSTÍ



KLASIFIKAČE



TWINSPAN

Two Way INDICATOR SPECIES ANALYSIS

- divisivní metoda
 - začíná dělením celého souboru vzorků a postupuje směrem dolů
 - skupina se dále nedělí, pokud je příliš malá, nebo bylo dosaženo dělení do požadované úrovně
- polytetická metoda
 - každé dělení závisí na **několika (indikačních) druzích** (x monotetická metoda – dělení ovlivňuje jediný druh)
- metoda velmi oblíbená mezi vegetačními ekology
 - ale – algoritmus je poměrně složitý, né zcela popsáný a s řadou arbitrárních kroků. Proto má také řadu zarytých odpůrců: "TWINSPAN too unstable and tricky: Better avoided." (Jari Oksanen)
- vzorky jsou uspořádány podle první osy korespondenční analýzy (CA, DCA) a podle ní jsou rozděleny do dvou shluků (vzorky s pozitivním skóre a negativním skóre)
- metoda ošetří vzorky, které leží blízko středu osy, a které tak mají velkou pravděpodobnost, že budou špatně klasifikovány

TWINSPAN

Two Way INDICATOR SPECIES ANALYSIS

- *pseudospecies*

- metoda primárně funguje pro **kvalitativní** data
- kvantitativní informace se dodává rozdelením druhů na *pseudospecies* podle relativní abundance (*cut levels*), např. 1, 5, 10, 20 %.

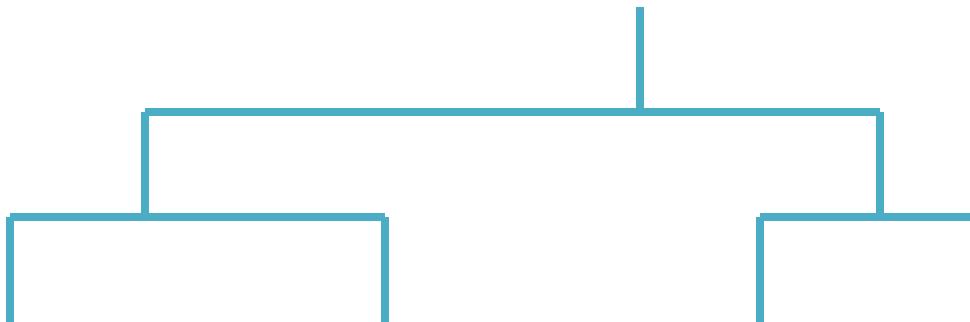
	Species	Sample 1	Sample 2
Original table	<i>Cirsium oleraceum</i>	0	1
	<i>Glechoma hederacea</i>	6	0
	<i>Juncus tenuis</i>	15	25
Table with pseudo-species used in TWINSPAN	Cirsoler1	0	1
	Glechede1	1	0
	Glechede2	1	0
	Junctenu1	1	1
	Junctenu2	1	1
	Junctenu3	1	1
	Junctenu4	0	1

TWINSPAN

Two Way INDICATOR SPECIES ANALYSIS

- *pseudospecies*
 - metoda primárně funguje pro **kvalitativní** data
 - kvantitativní informace se dodává rozdelením druhů na *pseudospecies* podle relativní abundance (*cut levels*), např. 1, 5, 10, 20 %.
- výsledkem je (mimo jiné) tabulka podobná fytocenologické
 - snímky z určitých klastrů a druhy s vysokou fidelitou k dané skupině jsou seskupeny dohromady
- metoda vhodná v případě, že jsou data strukturovaná podle jednoho výrazného gradientu
- vhodné na hledání (několika málo) ekologicky interpretovatelných skupin v datech

TWINSPAN



TWO-WAY ORDERED TABLE

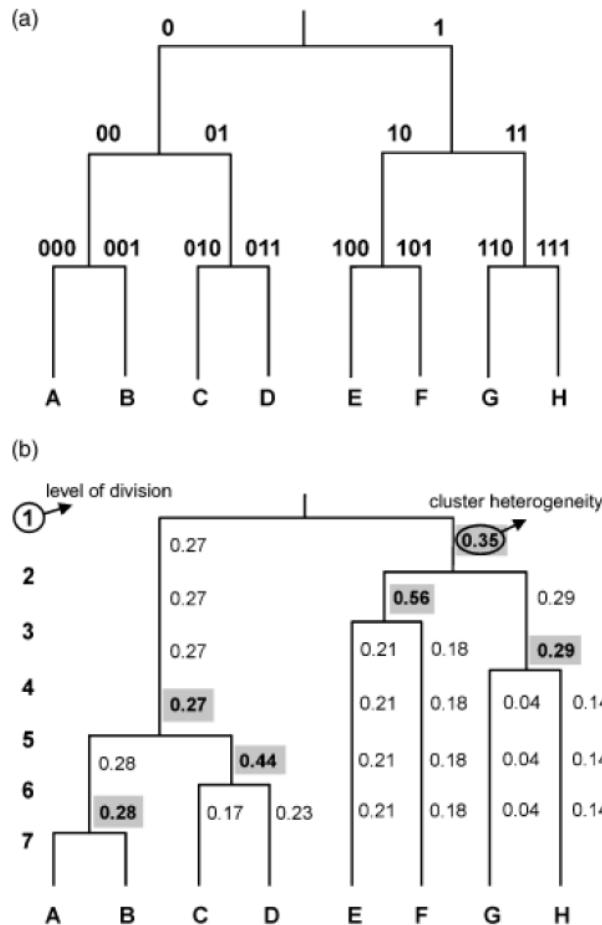
112333344567788999 1122222335555555666666677777888889991112233444778899 1133444488
5615201674954282604776703678945012346789012356789013459013452350289145890126678916123489342335787

44	Galespe6	-22-22--112--2---1-2-22-2222-2222--2-1---221-2-1-2-2-2-1---22-----12-----2-1-00
48	Impagla6	5222242--22522222221-1-12-2-1-----1---2-----1-1-----1-1-----1-1-----00
61	Seneova6	-1--2---21-22-2-1-----222-22222222221-1---222222-2-----1-----2-----00
86	Urtidio6	2222222222522-2-2221-222---22-222-1222---2-2-22-2-2-212-22-----2-----2-----00
94	Galemon6	-22-422-2-2-2---22---52-22-52---225242--22-2-2---25-----2-----2-----00
113	Sambnig4	-22-2222-2-22-----2-2-2522-2-2-2---255---25554-22222---2-22-22---2-1---2-2---2-2-----00
121	Asareur6	---2222-2-22-2222-22-2-----2222222222-22---2-2-2---2-----2-----2-1-----00
124	Dryofil6	--2-2-222-222222242-22222222224222222222-2222242325433232-2-1---2-2-22---4-21-2-22-00
149	Dryodil6	--21-----12222---2-22222---2-22-2222-22-22---2-2221---211-----1-----1-----00
215	Oxalace6	---22-2-22222212---2-----2-2222222222222222-222-22-2212---2-----2-----00
227	Abiealb1	-----2-22-----22---22244255222255522---5533455-3-----2-2-----00
8	Coryave4	-225544-55-445455-22-5-22222-222222-2552-52224-2434335342222---2-2222-22-2-2252422-22-42 01
16	Galebif6	2---2-1---22---2-1-1221---122---12-1-1-2221---221221---22222---22-----2-2-2-12-1-----01
55	Gerarob6	2-2-222-2-22-22122-22222-2-2-2222-22-2-1---2222-1-1---42---2-----2-----2-2-2---1-22-22 01
56	Impapar6	-222222-222222-32222222422222-1-22-2222-2222-222-22-2-222-122222---2-----22-2-2-22222---01
65	Tilicor1	-545---4-22-45-54---255-4-522-525522---222---24242---5-4-453-5-----2-----3---2-2-25445252 01
89	Acerpse7	-2-222-----21-212222---22-2-2-2-----22-2-1---222-22-22---1-----122---22---2-1-----01
151	Stelhol6	--222222-2-222-22-----22222-----2-2-----2-----2-----2222-221 01
26	Poa nem6	2-22-222---22-2-2-22---2-22-2222-----22222-2222-22-2-22---22-----224222222442222 10
33	Solivir6	1-1-----2-11222---2221-1---112---12---12-2---21-1---2-211-----222-21-1-----1-221 10
34	Tilicor4	-----222-2-2-----222222-25-----22-----2222-----22-2-----2222-22-2-2-25222222-10
63	Polyvul6	-----22-2-222-22-----1-1---221-1-122-----24222-2-2-----1-222222-1-----11222 10
126	Sorbauc7	-----2---2-2-1-----11-2-1222-222-1-2-21222-2-21-2---122-----22222-----21-2 10
128	Fagusyl1	-----2-----5-22---25---5-----52-----55-5-----5-----422---5453---2-----2 10
195	Calaaru6	-----2---2-22-1-22---222-1-2-----2222-2222-233-25-22-2-2---224-122225-----2-12 10
21	Luzuluz6	-2-----2-----222-2---22222-----1-2-222-222-2-2-23-222-222222-2-222-2-2-22-----22223 11
24	Pinusyl1	-----2-----2-21-----2-----2-----2-----2-----222-1222424332-2-222222-2-2-2 11
41	Avenfle6	-2-1---21-1-212---2-22-----222-----1-1-22-2-2-2-3-----222222224222223322-2-22-----2-23 11
50	Querpel1	--22-2-45-2-2-----22---2-2422-2-----4554-----2445-54545552-----52-222522525---11
162	Convmaj6	-----21-----2-221-2-----2-----2-----22222-22-----2-52-2212-22-----22-----2-2-22-2 11

***** TWINGAN *****

***** TWINSPAN completed *****

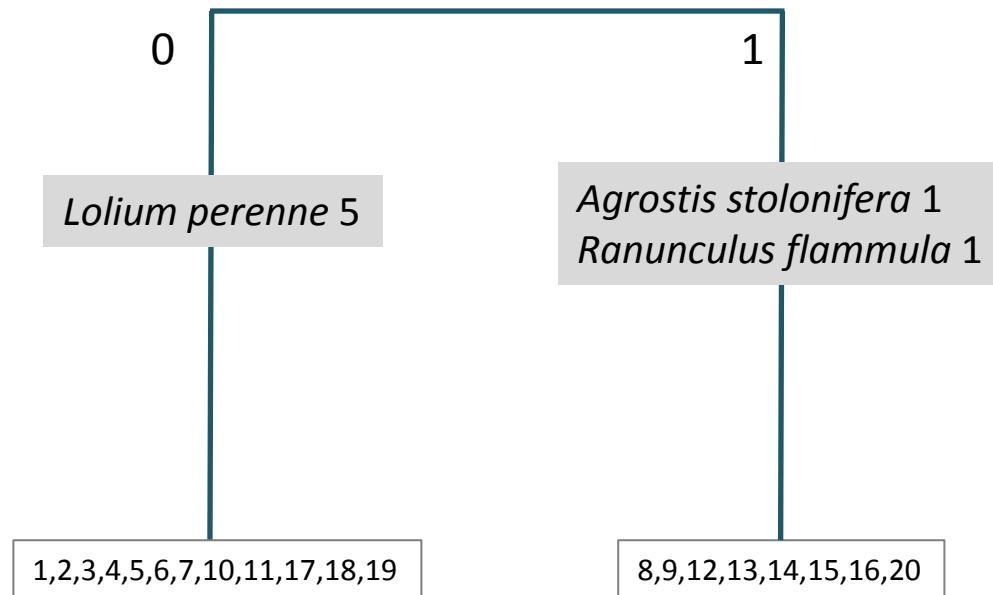
MODIFIKOVANÝ TWINSPLAN (ROLEČEK ET AL. 2009)



- na rozdíl od původního algoritmu (a) umožňuje modifikovaný TWINSPAN (b) dopředu stanovit cílový počet skupin
 - algoritmus se po každém dělení na dvě skupiny rozhoduje, kterou ze skupin bude dále dělit – vybere tu, která je více „heterogenní“ na základě její betadiverzity
 - míru betadiverzity je nutné zvolit (např. Jaccardův index podobnosti)
 - JUICE, R

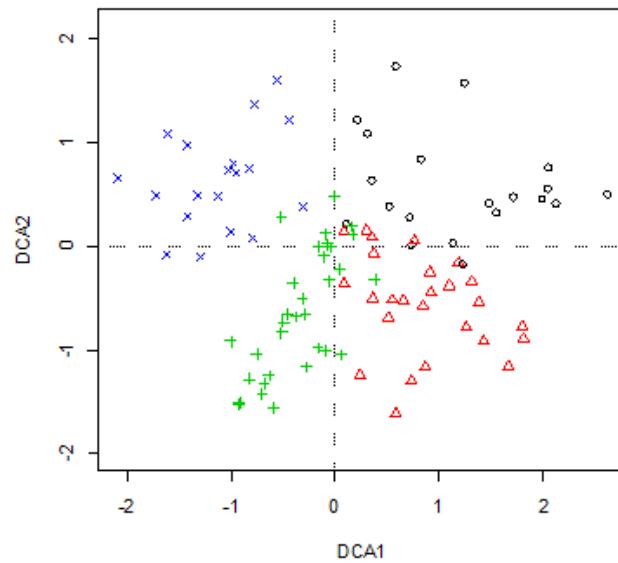
TWINSPAN

- kde je to možné, dělení je doplněno indikátorovými druhy

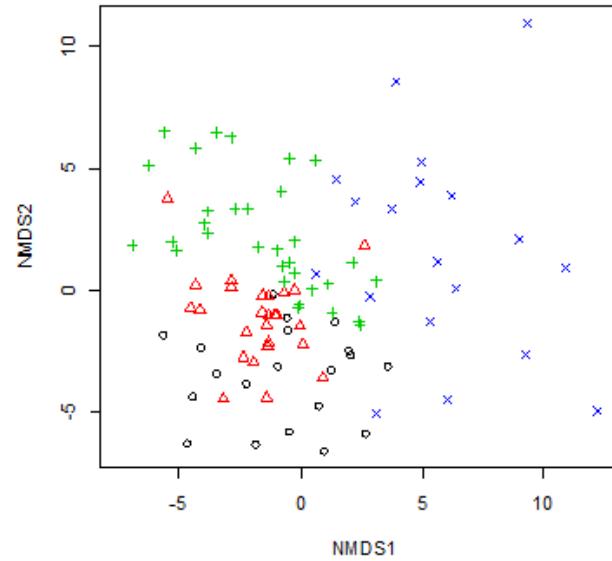


PROMÍTNUTÍ VÝSLEDKŮ NUMERICKÉ KLASIFIKACE DO ORDINAČNÍHO DIAGRAMU

DCA + TWINSPAN



NMDS (Bray-Curtis) + TWINSPAN



Je vhodné, aby míra nepodobnosti mezi vzorky byla v obou metodách (numerické klasifikaci i ordinační analýze) stejná (ze zvolených příkladů ten vlevo je vhodné řešení, vpravo nevhodné)

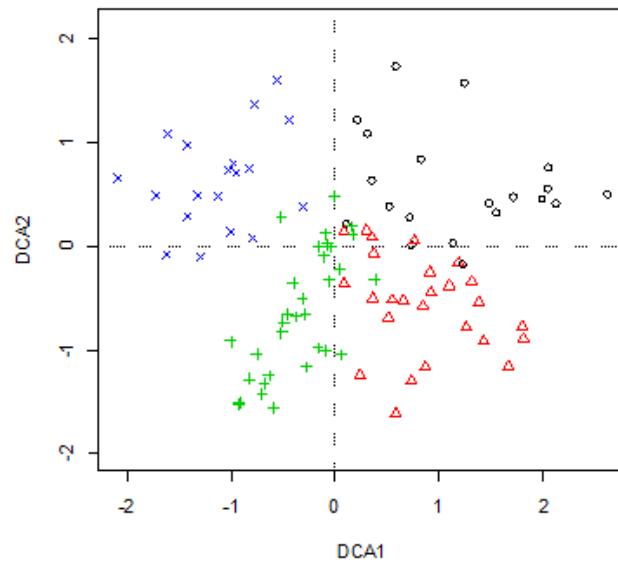
data z údolí Vltavy, klasifikace metodou **TWINSPAN** (Zelený & Chytrý 2007)

INTERPRETACE VÝSLEDKŮ NUMERICKÉ KLASIFIKACE

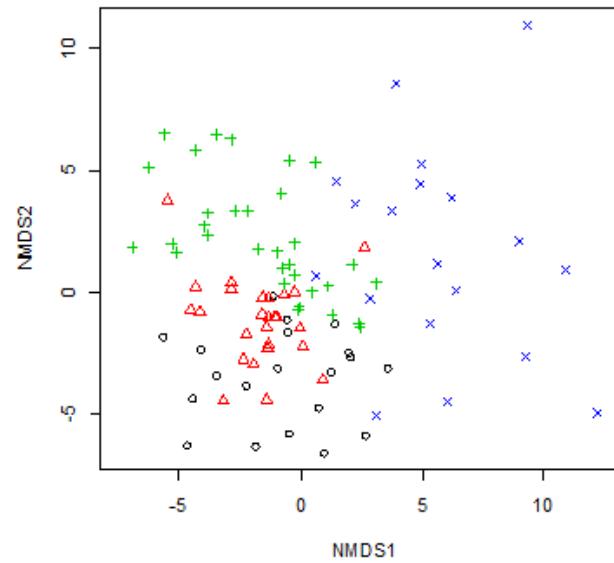
- promítnutí výsledků do ordinačního diagramu
- porovnání skupin na základě externích kritérií (např. měřených proměnných prostředí)
- porovnání skupin na základě druhového složení – stanovení charakteristických druhů

PROMÍTNUTÍ VÝSLEDKŮ NUMERICKÉ KLASIFIKACE DO ORDINAČNÍHO DIAGRAMU

DCA + TWINSPAN



NMDS (Bray-Curtis) + TWINSPAN



Je vhodné, aby míra nepodobnosti mezi vzorky byla v obou metodách (numerické klasifikaci i ordinační analýze) stejná (ze zvolených příkladů ten vlevo je vhodné řešení, vpravo nevhodné)

data z údolí Vltavy, klasifikace metodou **TWINSPAN** (Zelený & Chytrý 2007)