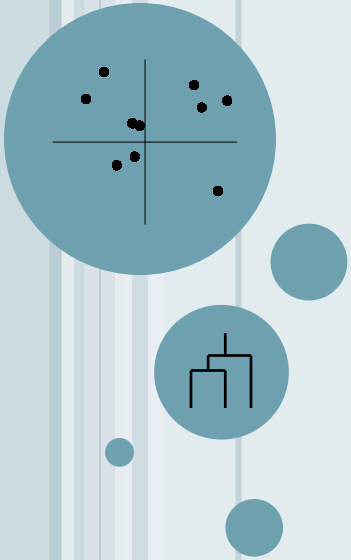
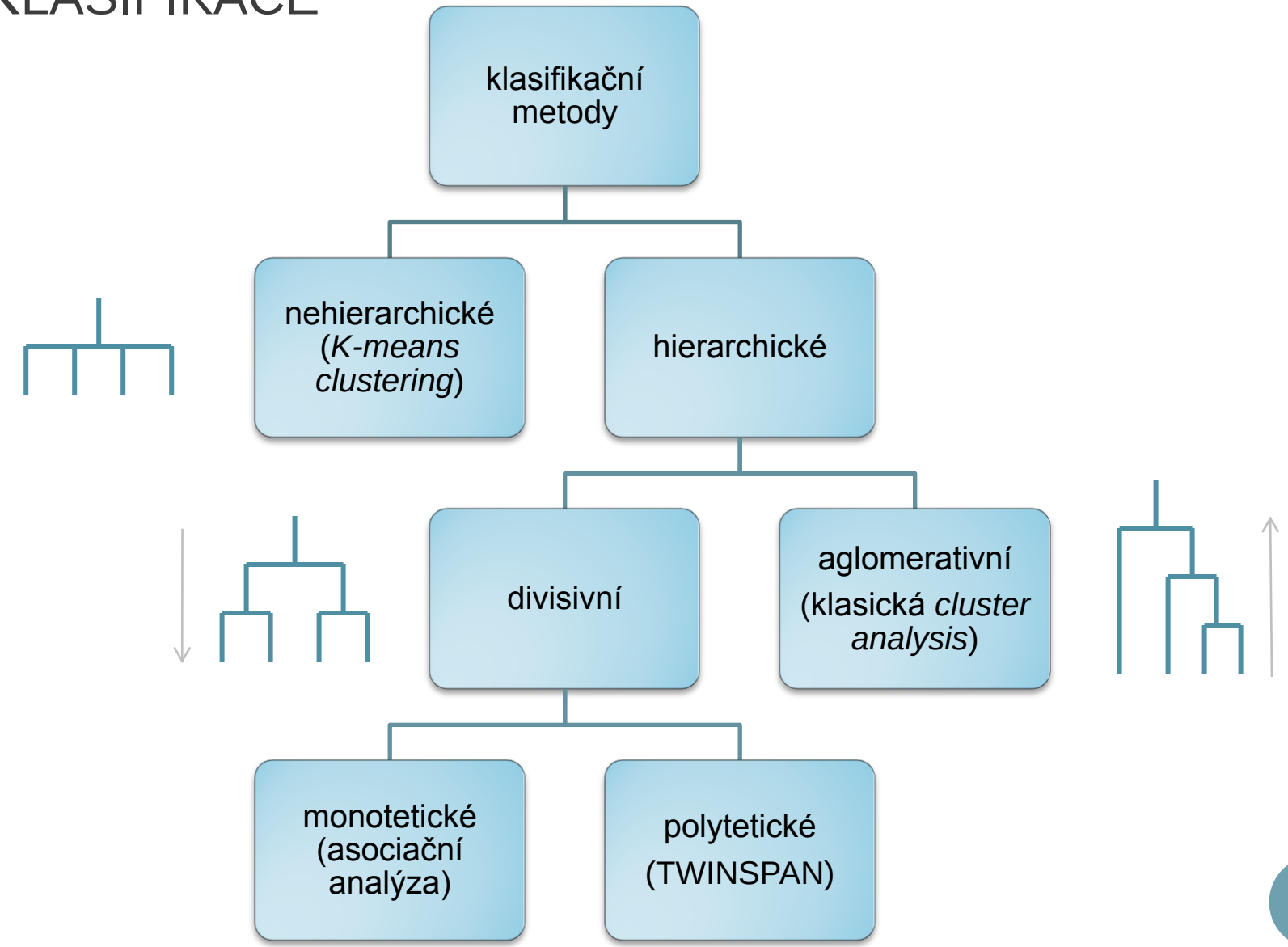


KLASIFIKACE

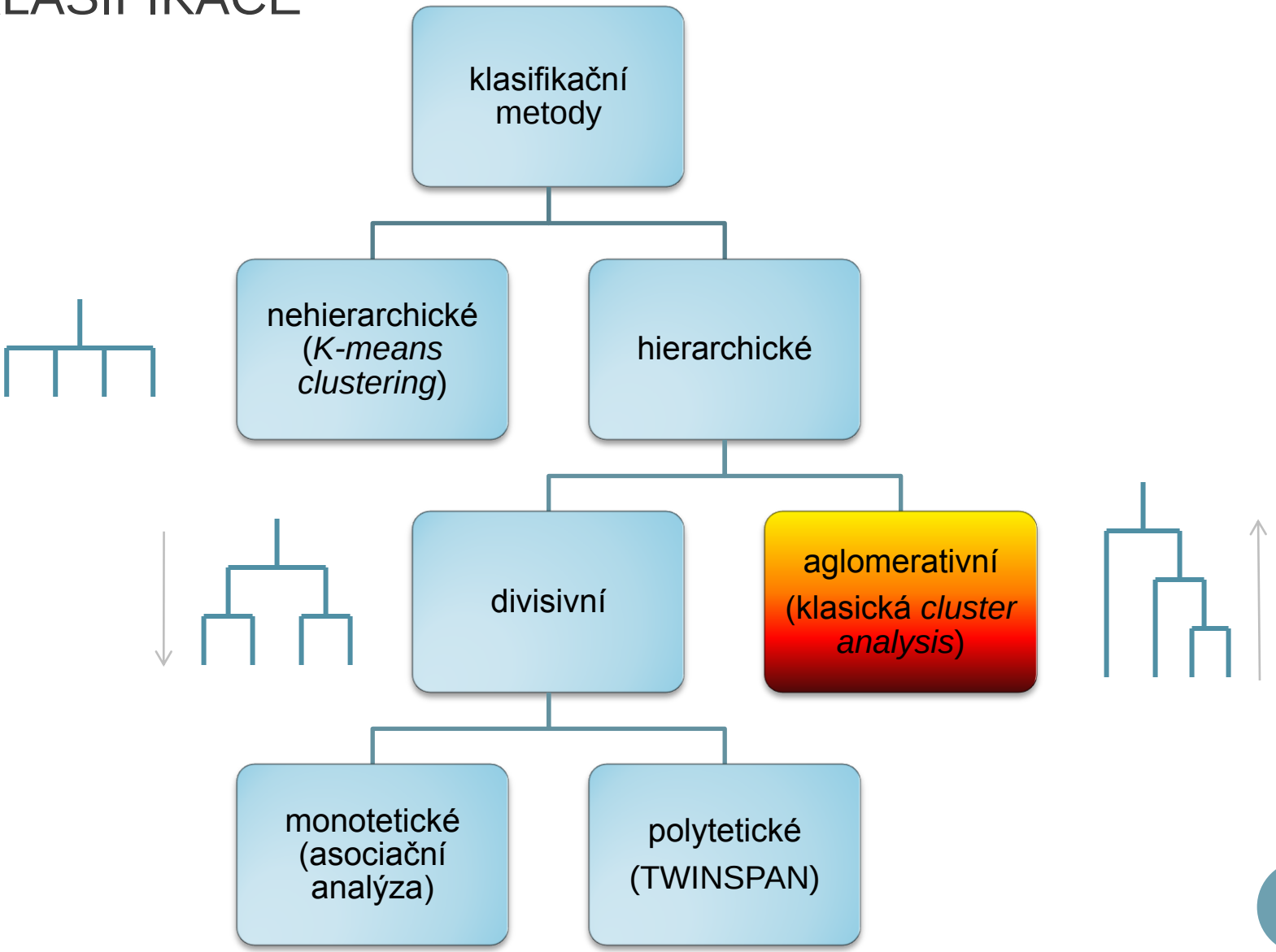
KLASIFIKACE

- cílem je seskupit podobné vzorky (případně druhy) do skupin, které jsou vnitřně homogenní, dobře popsatelné a odlišitelné od jiných 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

KLASIFIKACE



KLASIFIKACE



KLASIFIKACE

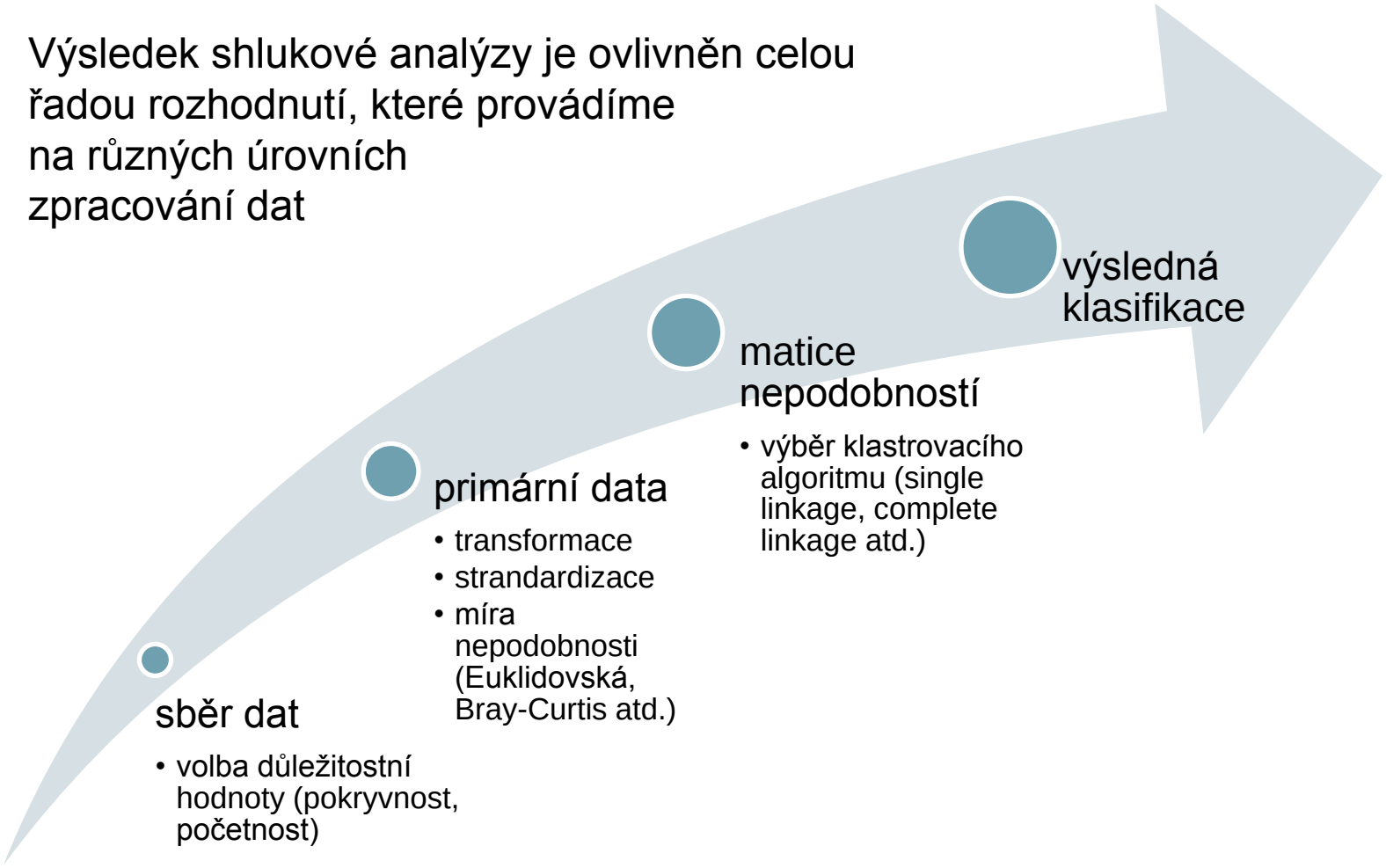
HIERARCHICKÁ A **AGLOMERATIVNÍ**

***Cluster analysis* – shluková analýza**

- 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*)
- pozor – nejde o OBJEKTIVNÍ metodu klasifikace (ta neexistuje), protože výsledná podoba klasifikace je ovlivněna řadou našich SUBJEKTIVNÍCH rozhodnutí

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

Single linkage, nearest neighbour

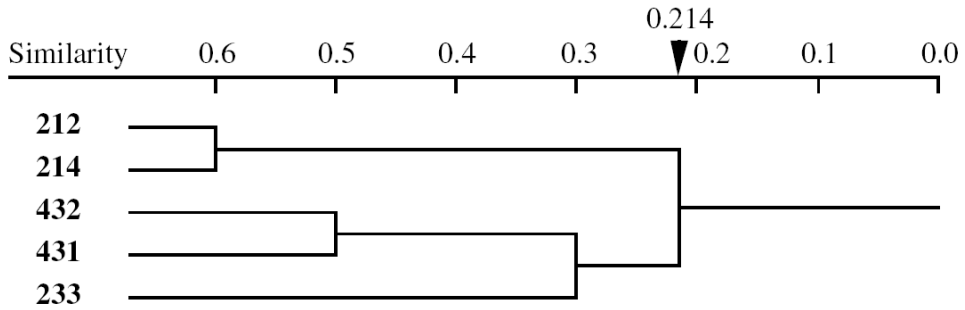
Ponds	Ponds				
	212	214	233	431	432
212	—				
214	0.600	—			
233	0.000	0.071	—		
431	0.000	0.063	0.300	—	
432	0.000	0.214	0.200	0.500	—

matice podobností

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



výsledný dendrogram

SHLUKOVÁ ANALÝZA (*CLUSTER ANALYSIS*)

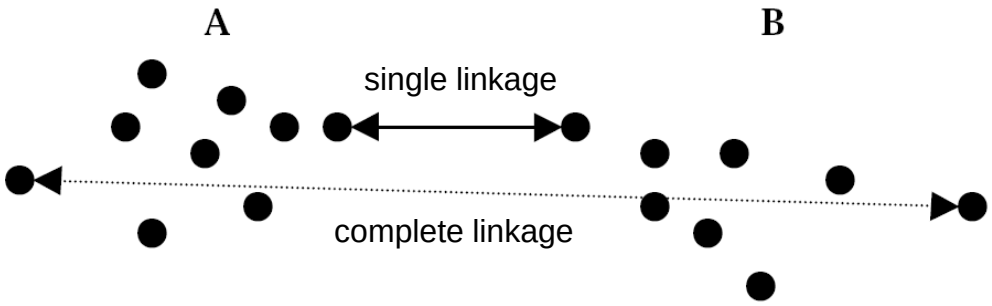
SHLUKOVACÍ ALGORITMY

Single linkage (nearest neighbour*, metoda jednospojná*)

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

***Complete linkage* (*farthest neighbour*, metoda všespojná)**

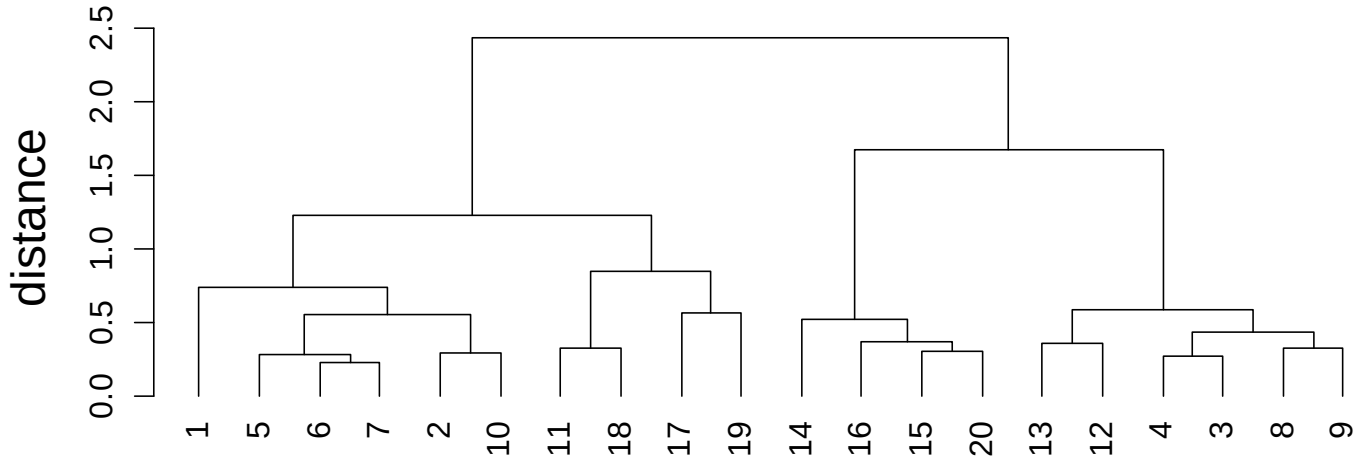
- vzorky se připojí ke shluku až v okamžiku, kdy shluk obsahuje všechny podobné vzorky
- *přidám se ke skupině ve které je ten, kdo je mi nejmíň nesympatický*



SHLUKOVÁ ANALÝZA (*CLUSTER ANALYSIS*)

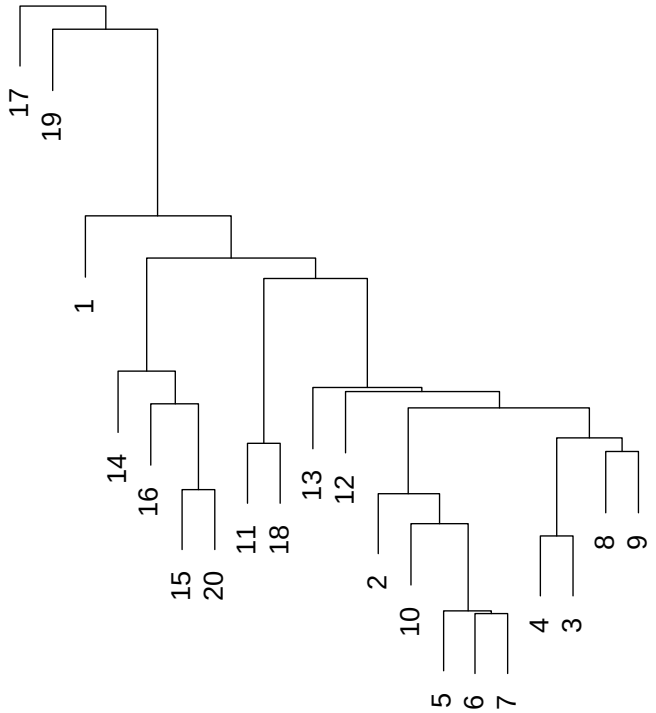
DENDROGRAM

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

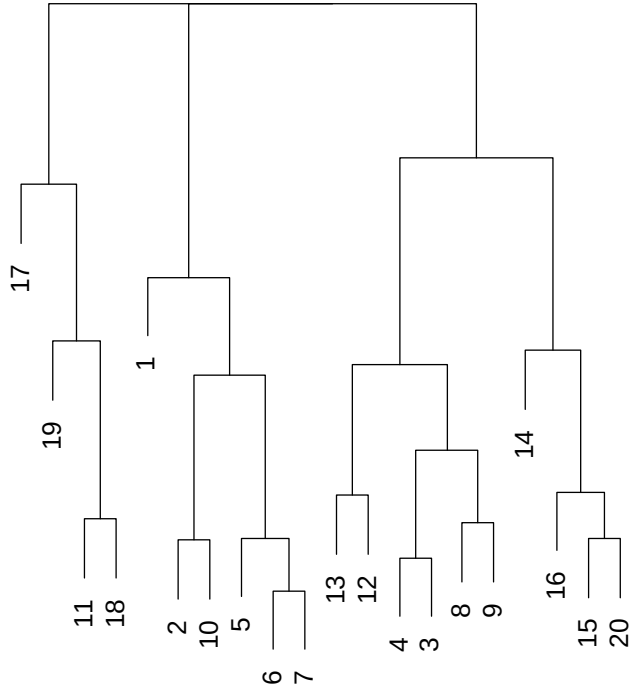


Single linkage vs *Complete linkage*

Bray-Curtis distance / Single linkage



Bray-Curtis distance / Complete linkage

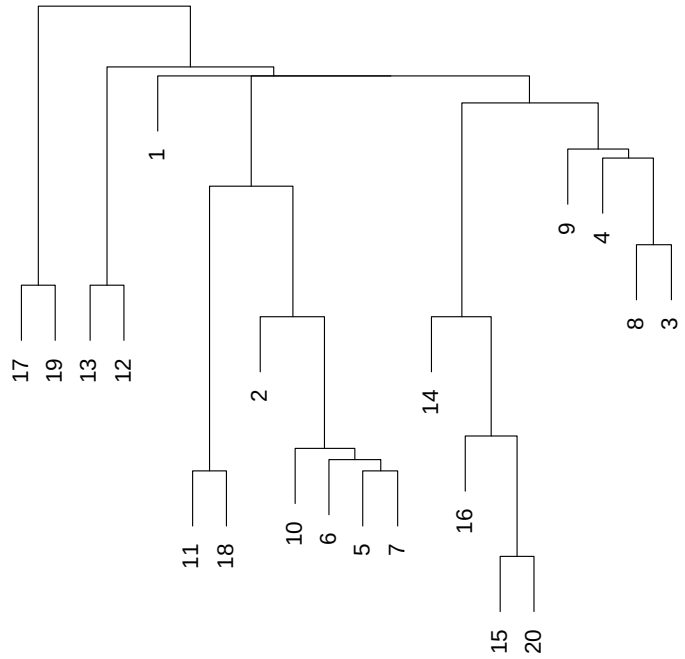


single linkage se výrazně řetězí

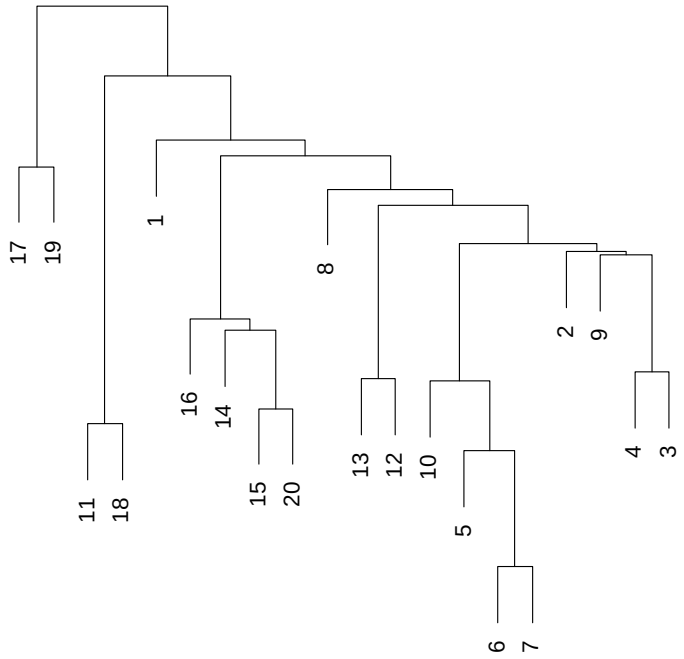
SINGLE LINKAGE

VLIV TRANSFORMACE DRUHOVÝCH DAT

Single linkage / Euclidean distance / no transformation



Single linkage / Euclidean distance / LOG transformation



transformace dat (např. logaritmická) může výrazně ovlivnit výsledný dendrogram – v případě euklidovských vzdáleností a *single linkage* metody obzvlášť

SHLUKOVÁ ANALÝZA (*CLUSTER ANALYSIS*)

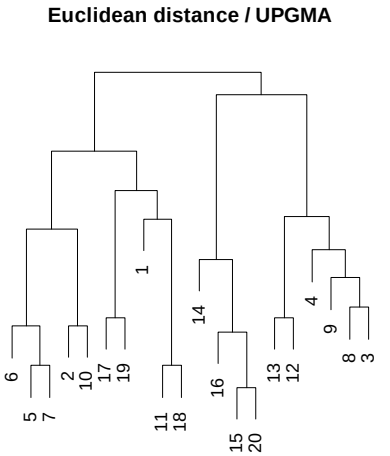
SHLUKOVACÍ ALGORITMY

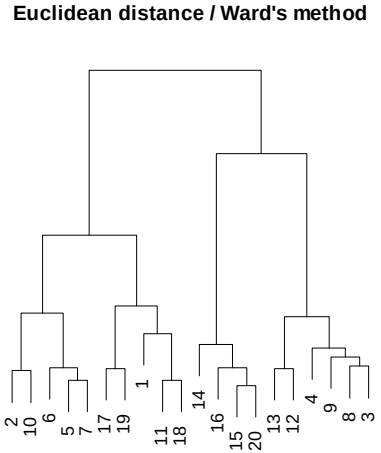
Average linkage (např. UPGMA)

- 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 všemy jeho vzorky
- *přidám se ke skupině, ve které jsou mi všichni v průměru nejvíc sympatičtí*

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

- ke shluku se připojí vzorek, jehož vzdálenost od centroidu shluku je nejmenší (počítáno přes čtverce vzdáleností mezi vzorky a centroidy shluků)
- nesmí se kombinovat se Sørensenovým (Bray-Curtis) indexem podobnosti



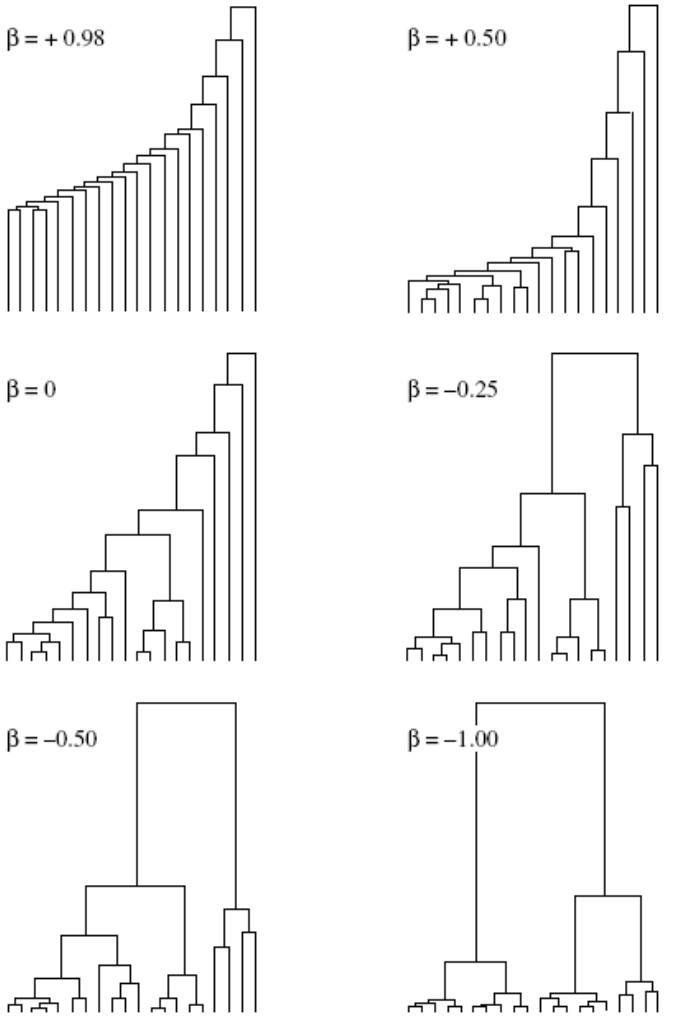


SHLUKOVÁ ANALÝZA (*CLUSTER ANALYSIS*)

SHLUKOVACÍ ALGORITMY

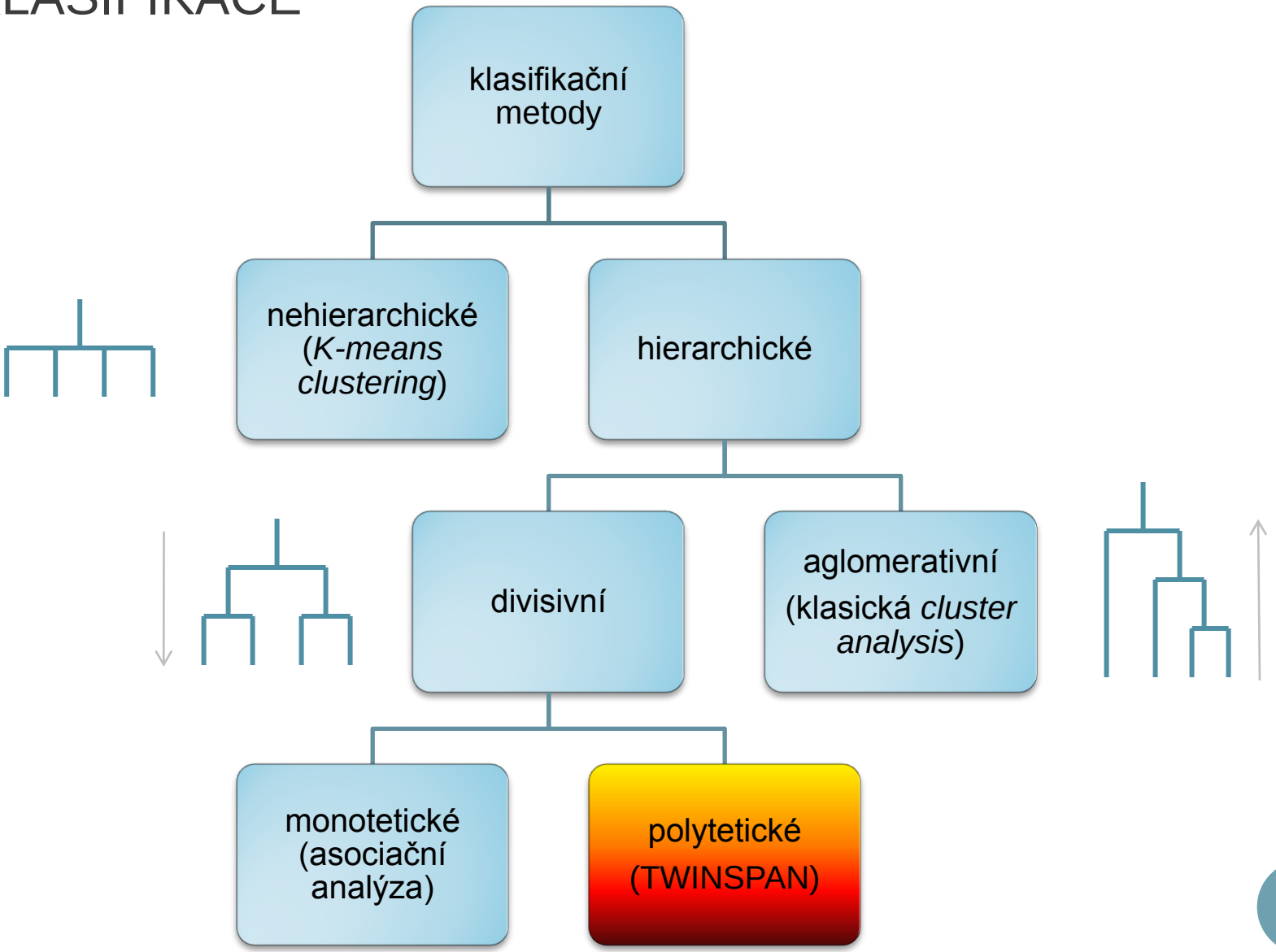
Flexible clustering (beta flexible)

- nastavení parametru β ovlivňuje řetězení dendrogramu
- nejvíc se řetězí pro β ~ 1, nejméně pro β = -1
- optimální reprezentace vzdáleností mezi vzorky je při β = -0,25



Legendre & Legendre 1998

KLASIFIKACE



KLASIFIKACE

HIERARCHICKÁ A **DIVISIVNÍ**

TWINSPAN (*Two Way INdicator Species ANalysis*)

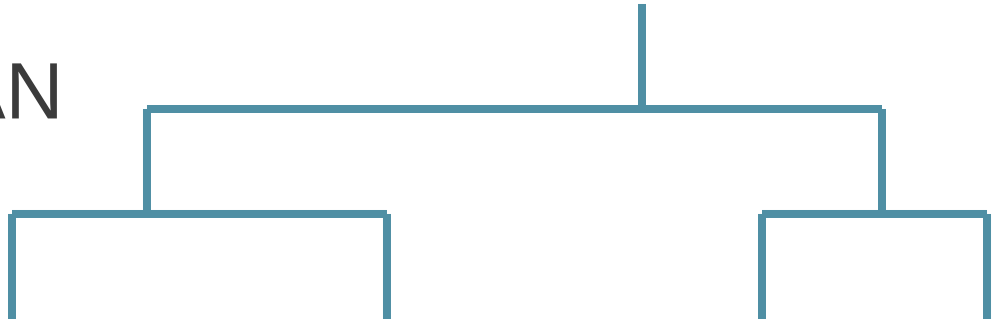
- divisivní metoda
  - začíná dělením celého souboru vzorků a postupuje směrem dolů
- polytetická metoda
  - každé dělení závisí na **několika (indikačních) druzích** (x monotetická metoda – dělení ovlivňuje jediný druh)
- metoda velmi oblíbená mezi vegetačními ekology
  - ale – algoritmus je poměrně složitý, s řadou arbitrárních kroků, a proto má také řadu zarytých odpůrců
- 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
- stejný postup je následně aplikován i na jednotlivé shluky

KLASIFIKACE
HIERARCHICKÁ A **DIVISIVNÍ**

TWINSPAN (*Two Way INdicator Species ANalysis*)

- *pseudospecies*
  - metoda primárně funguje pro kvalitativní data
  - kvantitativní informace se dodává rozdělením druhů na *pseudospecies* podle abundance (*cut levels*)
- 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
- PC-ORD, JUICE

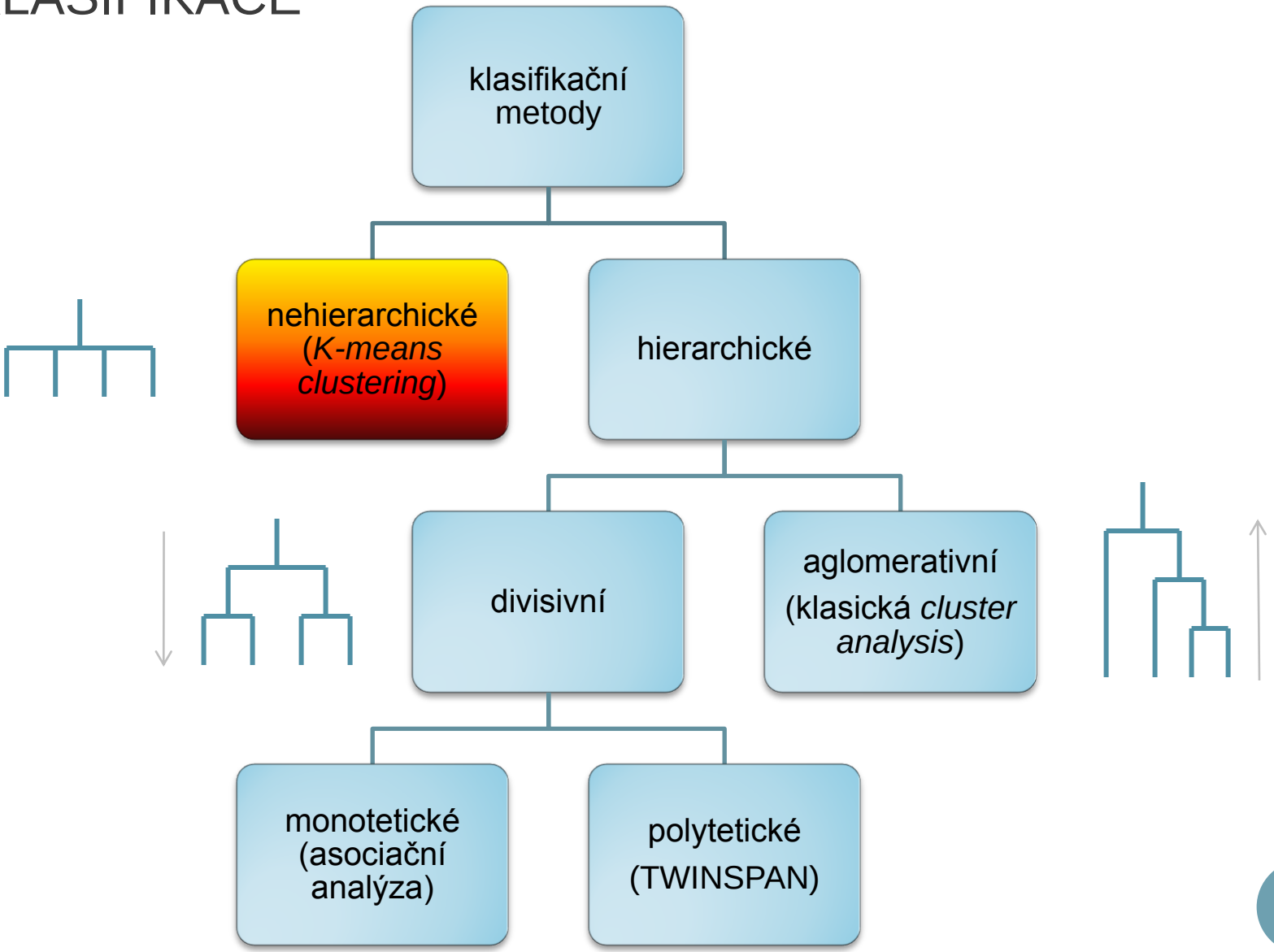
TWINSPAN



TWO-WAY ORDERED TABLE

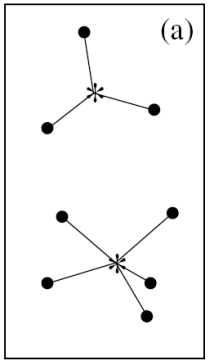
```
                  1123333445677889991 1122222233555555555566666666677777788888999111122233444447788991      113344448
                 5615201674954282604776703678945012346789012356789013459013452350289145890126678916123489342335787

  44  Galespe6  -22-22--112--2----1--2-22-222--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-------  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---2225242--22-2-2----25-----------------------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---1-  00
 124  Dryofil6  --2--2-222--22222222242-22222222222422222222222-22222242325433232--2--1-----2--2--22----4--21-2-22-  00
 149  Dryodil6  --21--------12222---2--22222-------2-22-22222-22--22---2-22221----211---------1--1----------1----  00
 215  Oxalace6  ----22--2--22222212----2------2-2222222222222222-222-22-2212-----2-----2-------2---------------  00
 227  Abiealb1  -----------2---22---------22----222442552222555222---5533455-3---------------2-2---------------  00
   8  Coryave4  -225544-55--445455--22-5-22222-2222222--2552--52224-24343353422222---2-2222--22-2-2252422-22---42  01
  16  Galebif6  2----2--1------22---22-1-1221---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--22222-22-2----1---2222--1---42---2--------2-----22--2----1-22--22  01
  56  Impapar6  -2222222--222222-3222222224222222--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-------21-2122222----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-----224222222244222222  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------------2222222-25------22----222----------22--2------2222---22-2-2---25222222-  10
  63  Polyvul6  -------------22-2-22---22--------1--1-----221--1--122------24222--2-2-----1-222222---1-----111222  10
 126  Sorbauc7  --------2-----2---2-1-------------11-2-1222-222--1---2-21222--2-21--2---122---2222----------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---224-1222255----------2--12  10
  21  Luzuluz6  --2-----2----------222-2---2---2222--------1--2-222-222-2-2-23-222-222222-2--222-2-2-22----22223  11
  24  Pinusyl1  ---------2-------2--21-----2-----2-------------------------2---22--122224224332-2-222222-----2--2  11
  41  Avenfle6  --2-1---21-1--212---2-22--------222--------1-1--22-2--2-2-3--2222222422222323322-2--22-----22-3  11
  50  Querpet1  ---22-2-45-2-2------22---2-2422-2---------------4554----------2445-54545552------52-222522525--  11
 162  Convmaj6  --------21-------2--221-2-------2--------------22222--22-----2--52-2212-22----22------2-2-22--2  11

                  0000000000000000000000000000000000000000000000000000000000000000000011111111111111111111111111111111
                  0000000000000000000111111111111111111111111111111111111111111111111100000000000000000001111111111111
                   ********** TWINSPAN completed **********
```

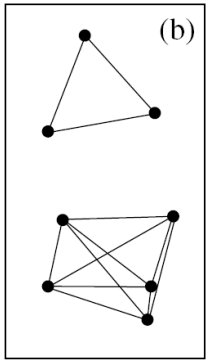
KLASIFIKACE



KLASIFIKACE

NEHIERARCHICKÁ





Legendre & Legendre 1998

K-means clustering

(shlukování metodou K-průměrů)

- nehierarchická metoda – všechny shluky jsou si rovny
- minimalizuje sumy čtverců vzdáleností mezi vzorky uvnitř shluků
- na začátku uživatel zvolí počet shluků
- 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)
- STATISTICA, SYN-TAX 2000

KLASIFIKACE

OBECNÉ ROZDĚLENÍ

Subjektivní

- založená na subjektivních kritériích badatele, špatně (pokud vůbec) reprodukovatelná

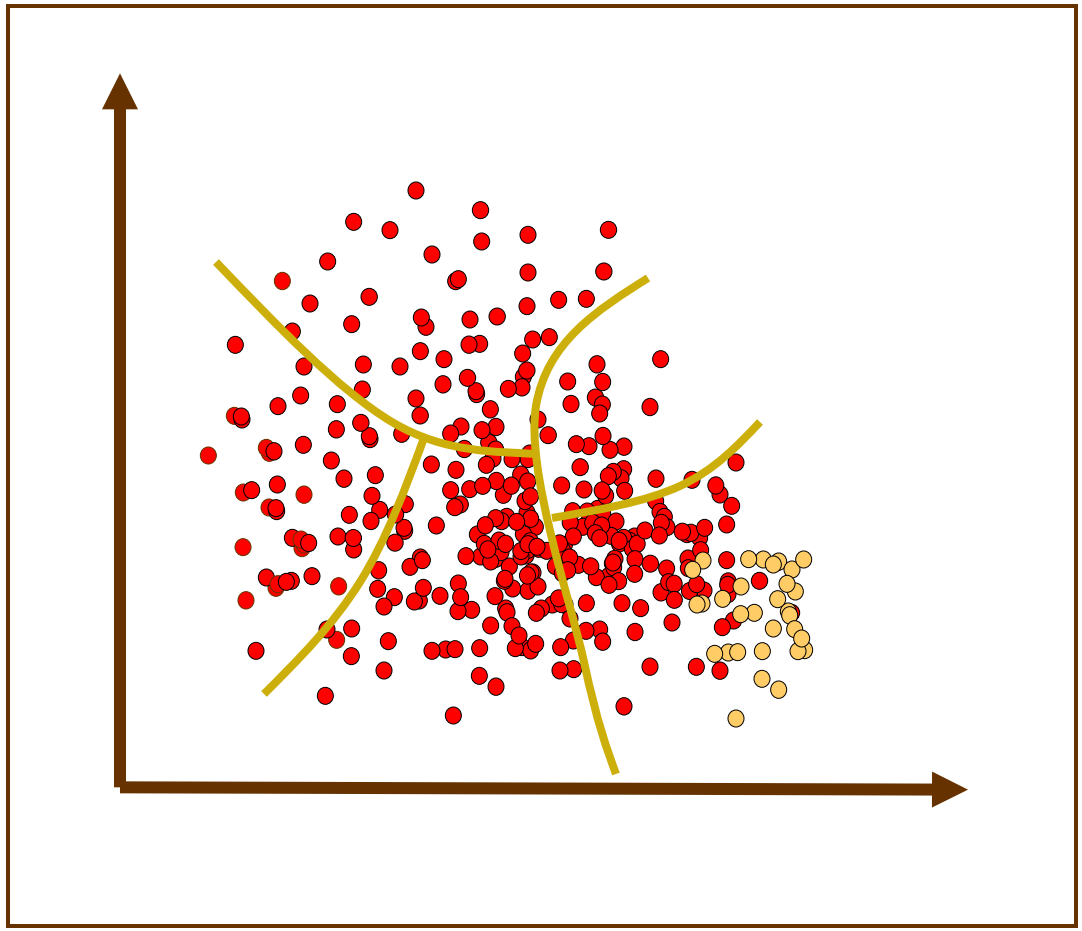
Formalizovaná (ne objektivní!)

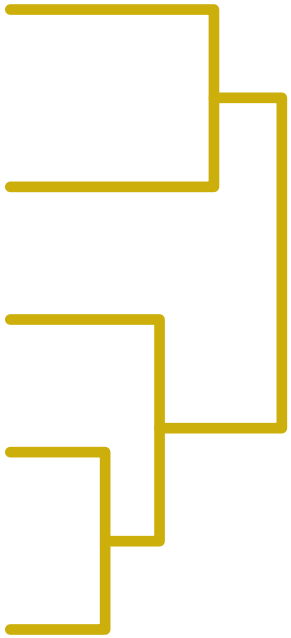
- omezený výběr jasně specifikovaných (formálních) kritérií, dobře reprodukovatelná

- neřízená (*unsupervised*)
  - numerické metody klasifikace (*cluster analysis*, TWINSPAN)
  - výslednou klasifikaci můžeme ovlivnit pouze výběrem metody (kombinace klasifikačního algoritmu a míry podobnosti), případně požadovaného počtu shluků
- řízená (*supervised*)
  - ANN – *artificial neural networks*, klasifikační stromy, náhodné lesy (*random forests*), COCKTAIL
  - klasifikační systém musíme nejdříve naučit, jak má vypadat výsledná klasifikace (*training*), a systém ji pak reprodukuje na dalších vzorcích

Nevýhody neřízené klasifikace

(přiřazení vzorků do skupin se mění podle kontextu)

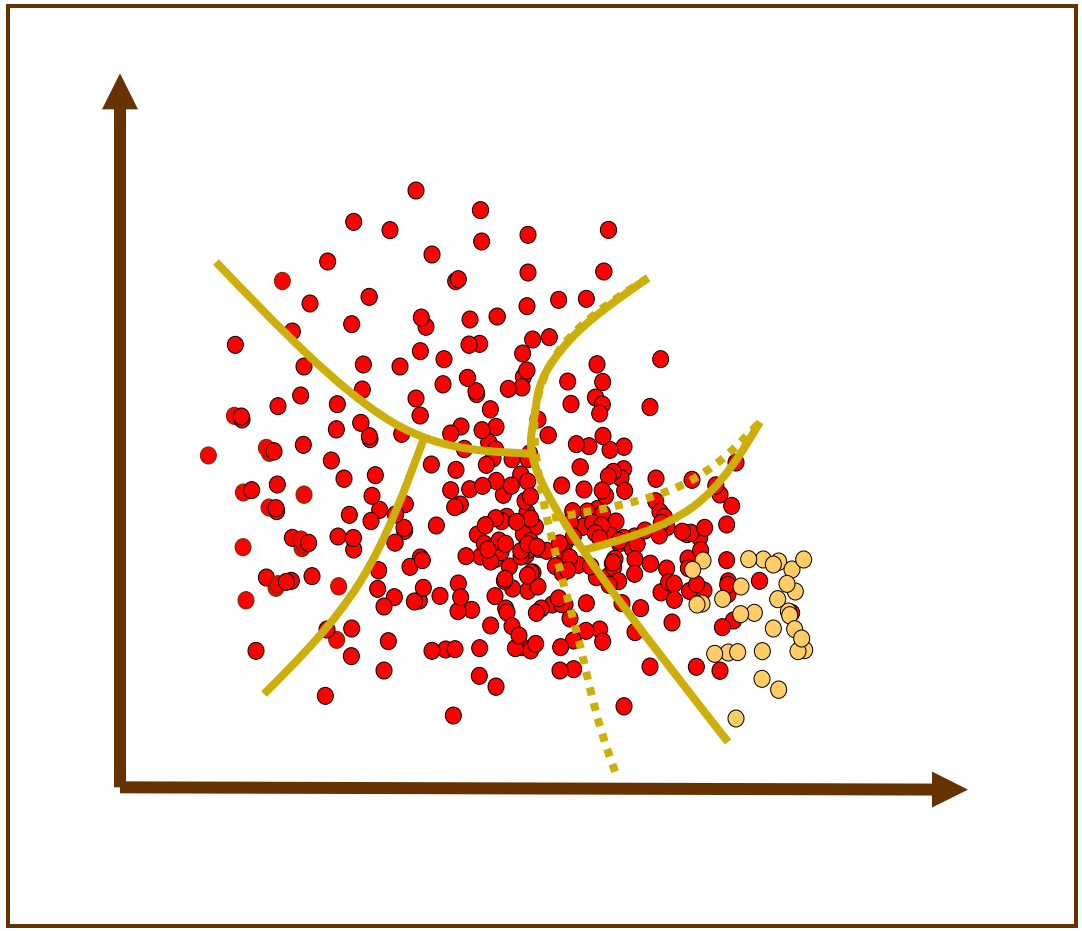




M. Chytrý - Vácrátót

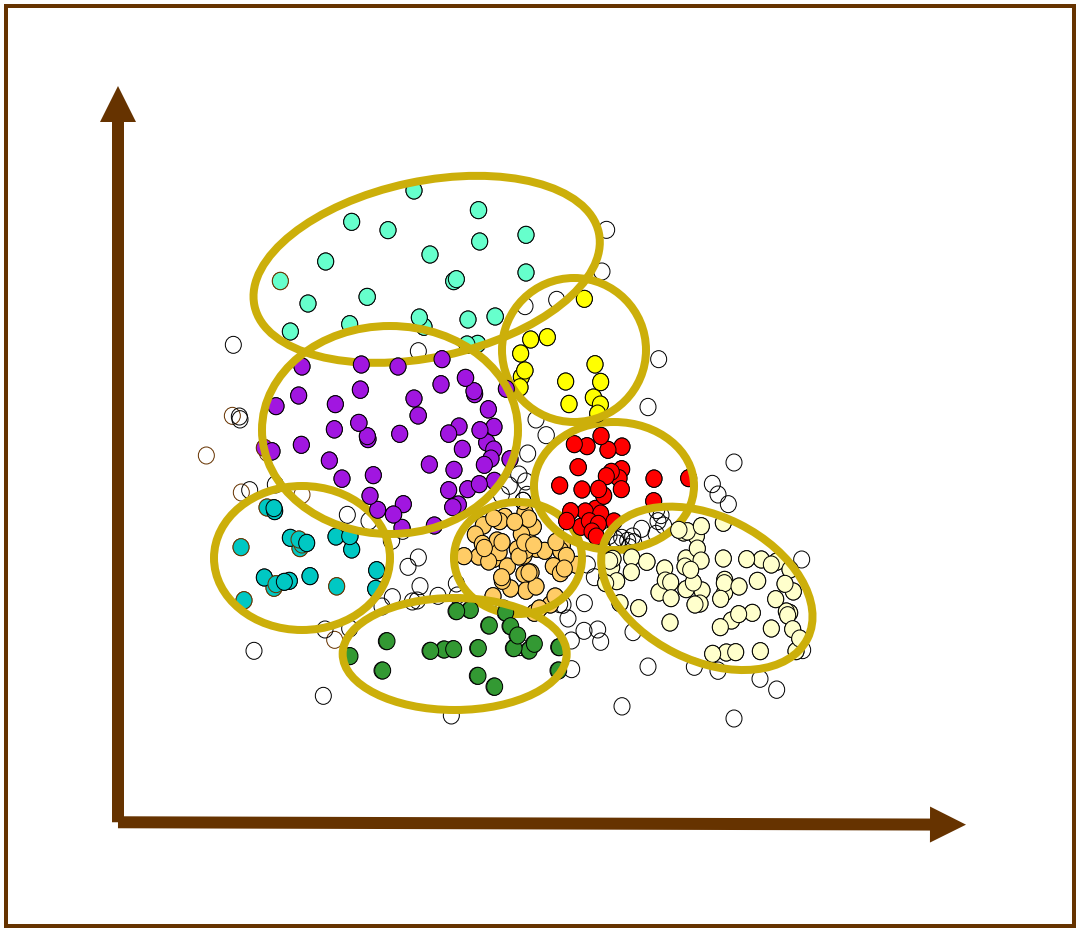
NEVÝHODY NEŘÍZENÉ KLASIFIKACE

(PŘIŘAZENÍ VZORKŮ DO SKUPIN SE MĚNÍ PODLE KONTEXTU)



M. Chytrý - Vácrátót

Řízená klasifikace



M. Chytrý - Vácrátót

Řízená klasifikace