

E8601: Pokročilé statistické metody

4. cvičení



Asociační matice – opakování
Shluková analýza

Asociační matice – Q mode analýza



NxP datová tabulka

	p_1	p_2	p_3
n_1			
n_2			
n_3			
n_4			
n_5			

Hodnota subjektu n_5 v parametru p_1

Výpočet metriky vzdálenosti



NxN asociační matice

	n_1	n_2	n_3	n_4	n_5
n_1	0				
n_2		0			
n_3			0		
n_4				0	
n_5					0

Vzdálenost subjektu n_5 od subjektu n_1 .

	p_1	p_2	p_3
n_1			
n_2			
n_3			
n_4			
n_5			

Hodnota subjektu n_5 v parametru p_1

Výpočet metriky podobnosti



	n_1	n_2	n_3	n_4	n_5
n_1	1				
n_2		1			
n_3			1		
n_4				1	
n_5					1

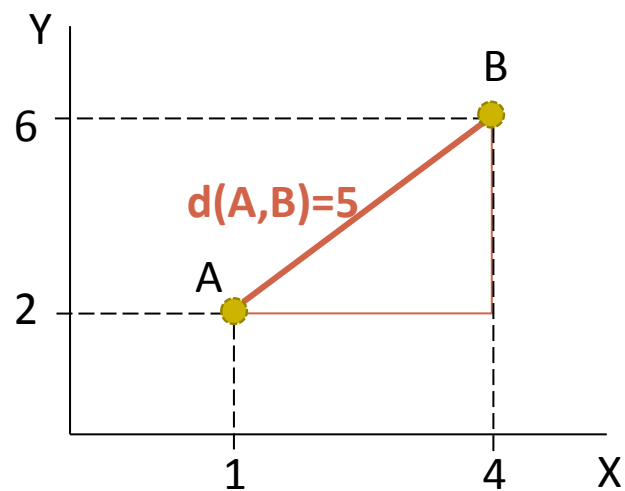
Podobnost subjektu n_5 se subjektem n_1 .

Euklidova vzdálenost I



- Euklidova vzdálenost vychází z Pythagorovy věty:

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{k=1}^m (x_{ik} - x_{jk})^2}$$



Euklidova vzdálenost II



	plat	počet cigaret/den
n_1	15 000	10
n_2	25 000	15
n_3	20 000	20
n_4	13 000	25
n_5	18 000	10

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{k=1}^m (x_{ik} - x_{jk})^2}$$

POZOR!

- Proměnné s číselně většími hodnotami budou mít větší váhu při shlukování!!!
- Např. pokud budeme hodnotit výšku (150–200 cm) a cholesterol (do 5 mmol/l), výška bude mít větší váhu při shlukování – objekty budou rozděleny do shluků podle jejich výšky.
- Data s nesrovnatelnými hodnotami proměnných je potřeba před analýzou **standardizovat**. Jak?
 - Např. standardizace na z-skóre.
 - Jak byste popsali rozložení z-skóre?

$$z = \frac{x - \mu}{\sigma}$$

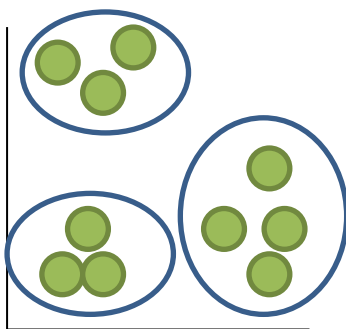
Shluková analýza



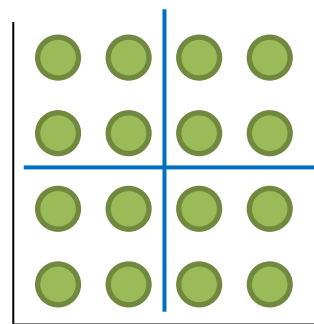
Shluková analýza – jaký je cíl?



- Seskupení objektů do shluků podle toho, jak si jsou podobné – chceme co nejpodobnější objekty v rámci shluků a co nejodlišnější mezi shluky.
- Shluková analýza vychází z asociační matice vzdáleností objektů (Q mode) *nebo* závislosti parametrů (R mode).
- Můžeme provést dvě hlavní chyby: špatný výběr metriky a špatný výběr algoritmu shlukování.
- Smysluplnost výsledků shlukování závisí jednak na objektivní existenci shluků v datech, jednak na arbitrárně nastavených kritériích definice shluků.

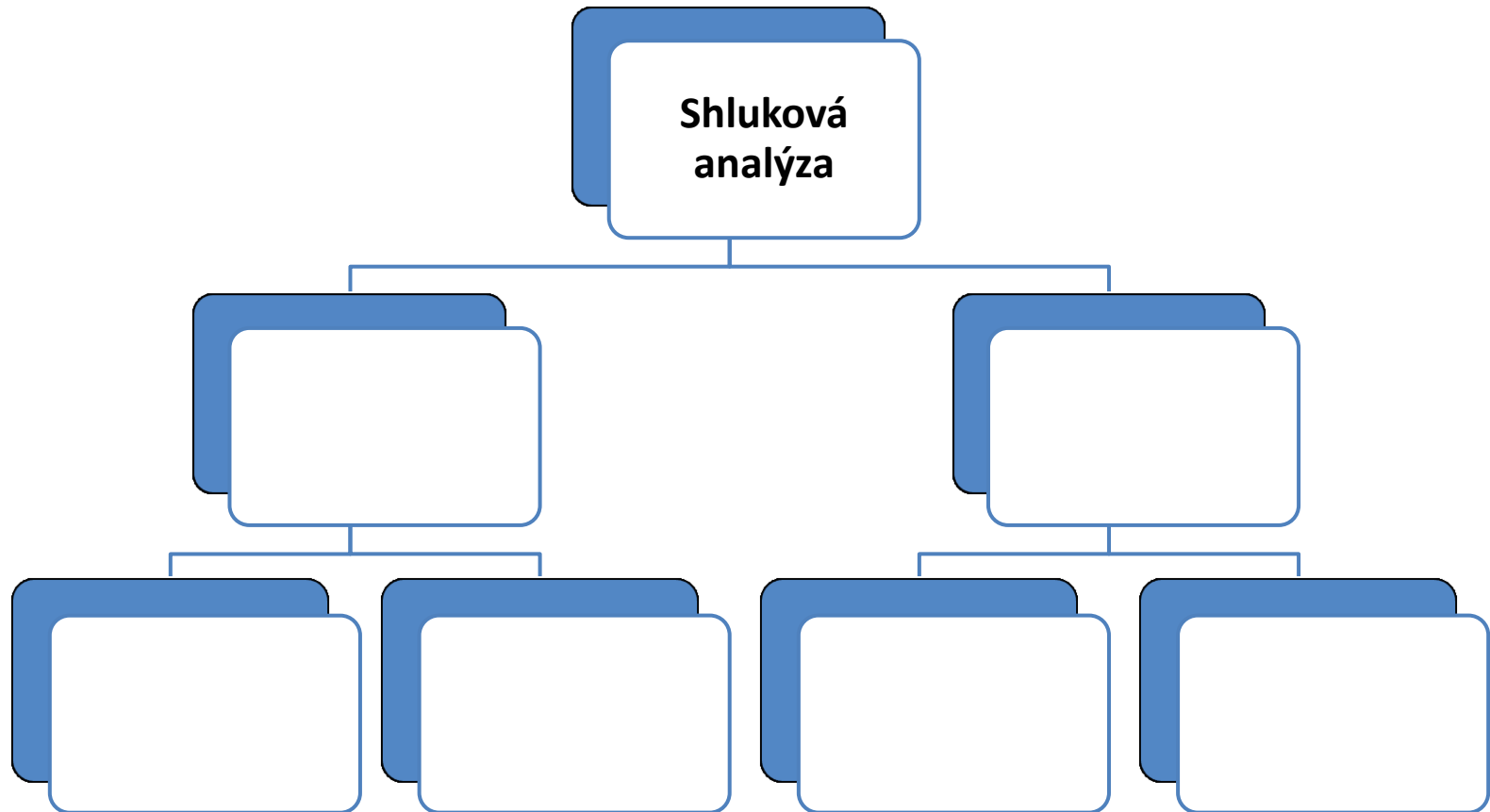


Jednoznačné odlišení
existujících shluků v datech

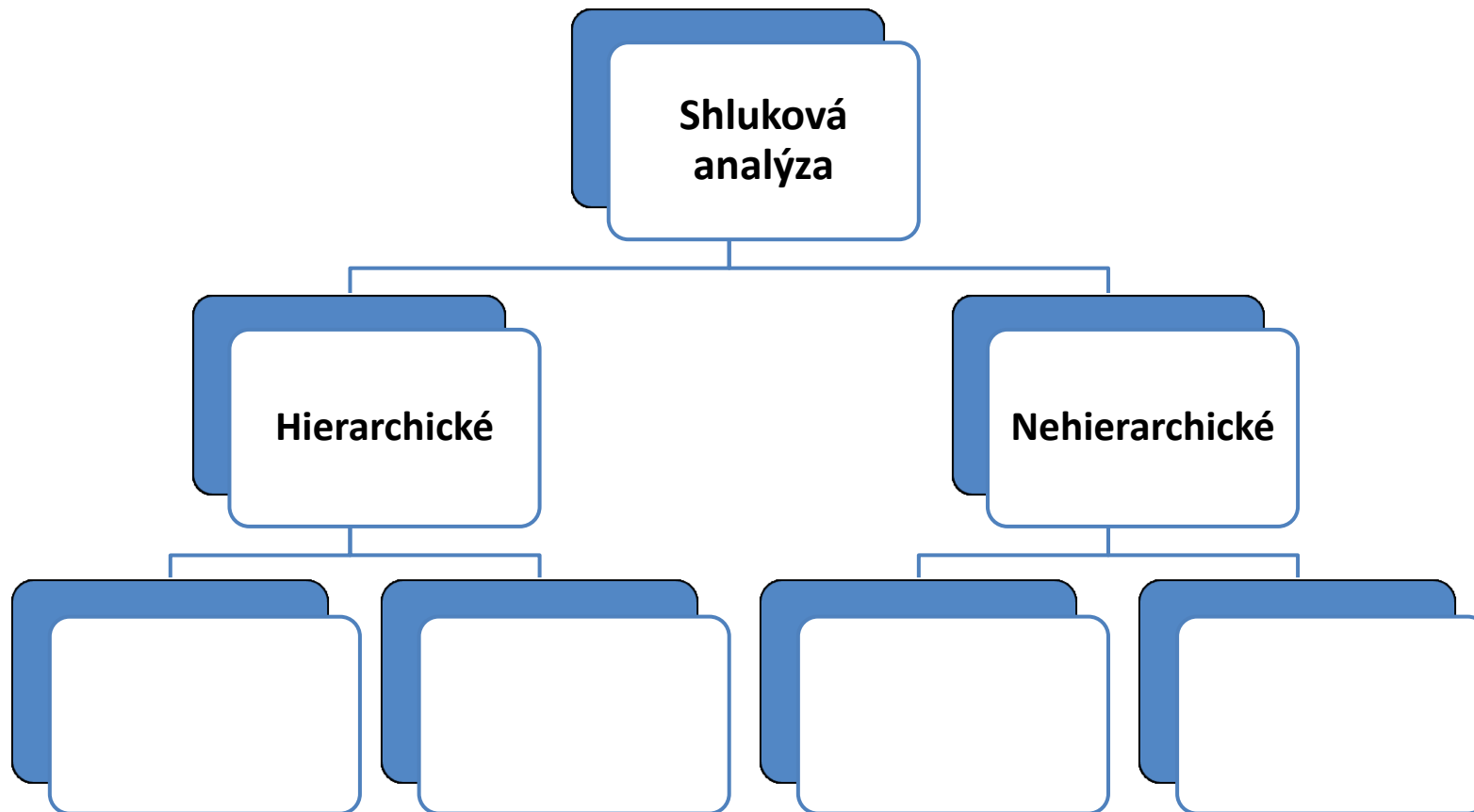


Shlukovou analýzu lze provést i na
datech bez objektivní existence shluků

Shluková analýza: typy metod



Shluková analýza: typy metod



Shluková analýza: typy metod



Hierarchické

shluky jsou definovány postupným skládáním objektů

Divizivní

Objekty jsou nejprve rozděleny do dvou shluků, tyto shluky jsou dále rozděleny atd.

Aglomerativní

Po spojení první dvojice objektů dochází k postupnému napojování dalších objektů.

Nehierarchické

Shluky jsou definovány v jednom kroku

Divizivní

Objekty jsou rozděleny do předem nastaveného počtu shluků.

Aglomerativní
sít spojených bodů

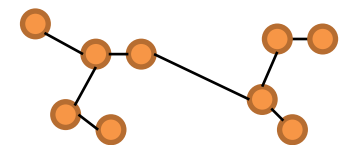
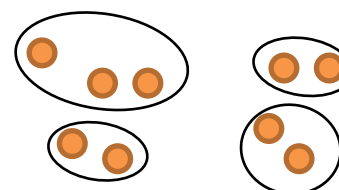
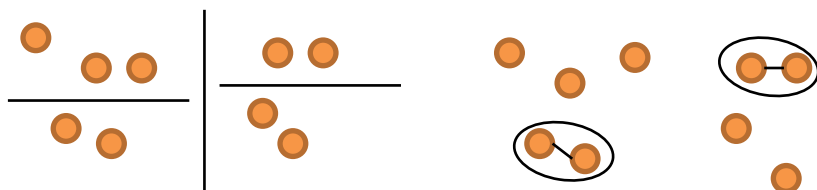
1. Krok



Kolik shluků chceme definovat? Například 4

Minimum spanning tree, Prim network

2. Krok



X. Krok

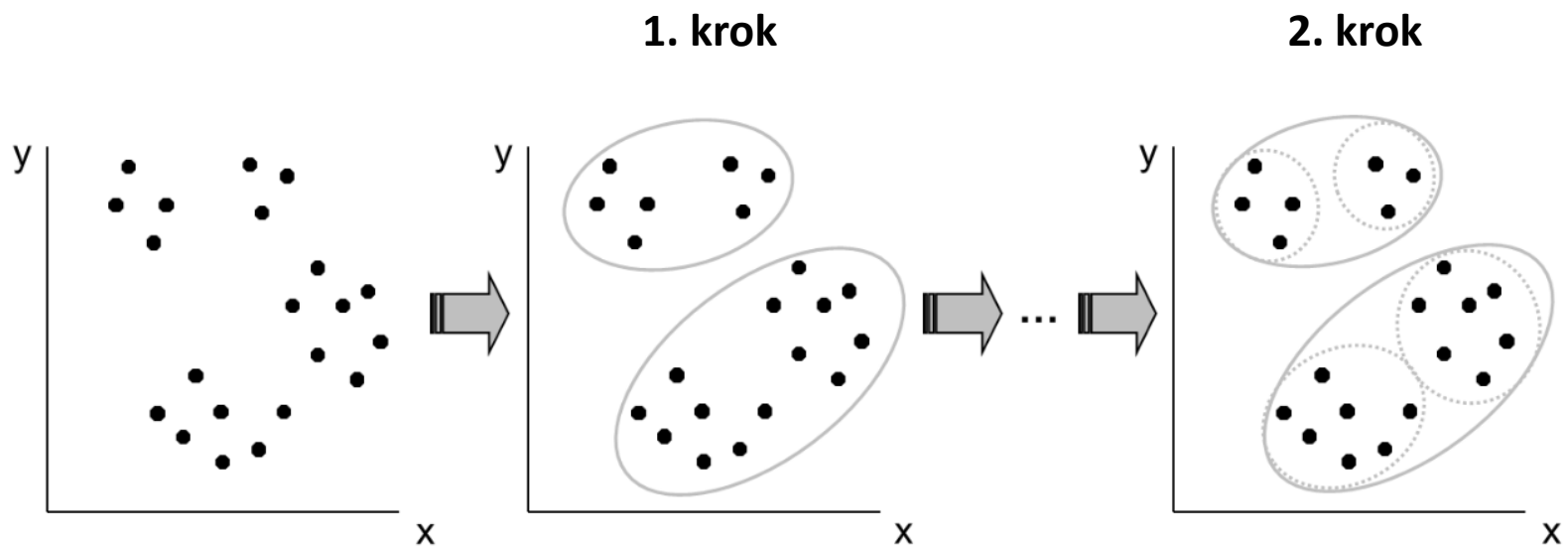
Atd.

Atd.

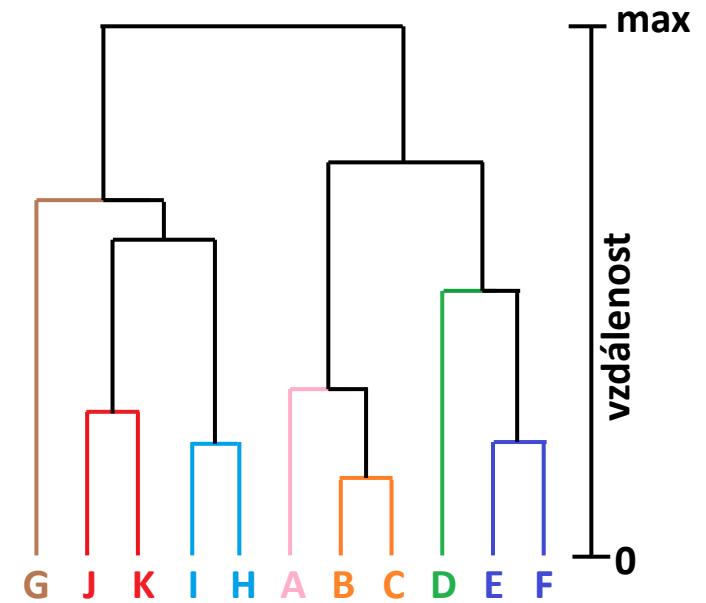
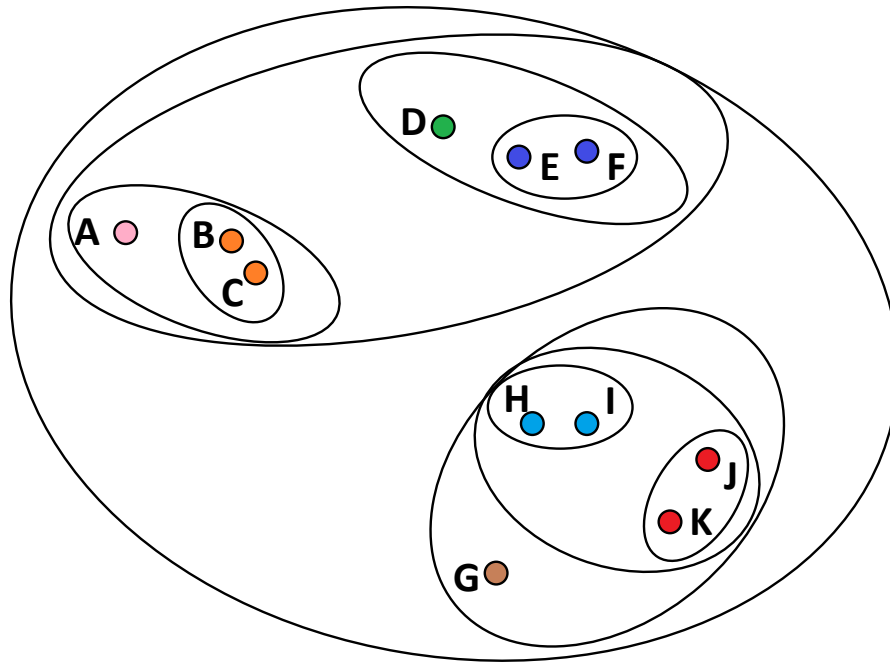
Výpočet ukončen

Výpočet ukončen

Pojmenujte shlukovací algoritmus I



Hierarchické aglomerativní algoritmy

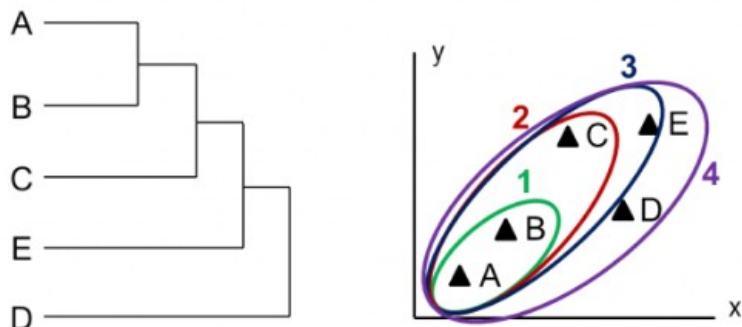


Shlukovací algoritmy

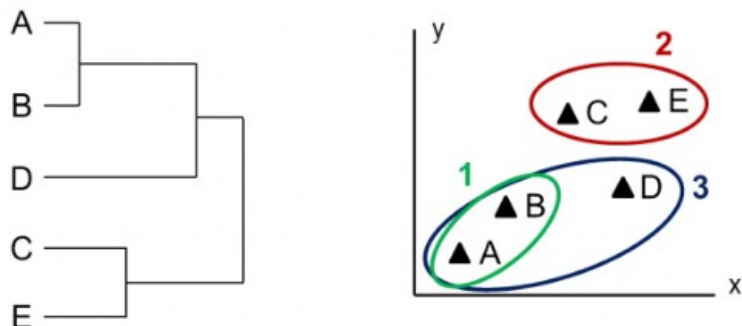
hierarchické aglomerativní

Metoda nejbližšího souseda („single linkage“)

- spojení na základě nejmenší minimální vzdálenosti dvou objektů

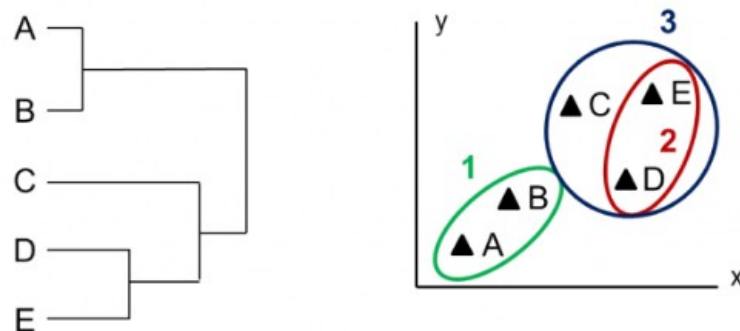


Metoda nejvzdálenějšího souseda („complete linkage“) - spojení na základě nejmenší maximální vzdálenosti dvou objektů



Centroidová metoda („centroids“)

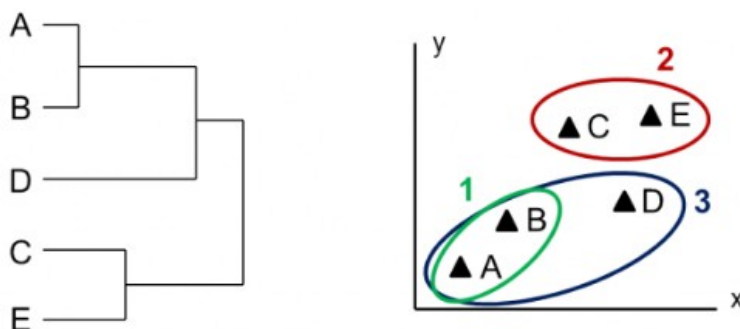
- spojení na základě minimální vzdálenosti centroidů (= průměrů) shluků



Wardova metoda („Ward's method“)

- shluky jsou vytvářeny tak, aby nově vzniklý shluk přispíval co nejméně k sumě čtverců vzdáleností objektů od centroidů jejich shluků

- vstupem je čtverec Euklidovy vzdálenosti



Úkol č. 1



- Na základě asociační matice sestrojte dendrogram pomocí algoritmu nejvzdálenějšího souseda.
 - Jaká je minimální vzdálenost dvou objektů?
 - Vykreslete spojení objektů v dendrogramu a přepočítejte asociační matici.

	A	B	C	D	E
A	0	Matice je symetrická podél diagonály			
B	9	0			
C	3	7	0		
D	6	5	9	0	
E	11	10	2	8	0

?

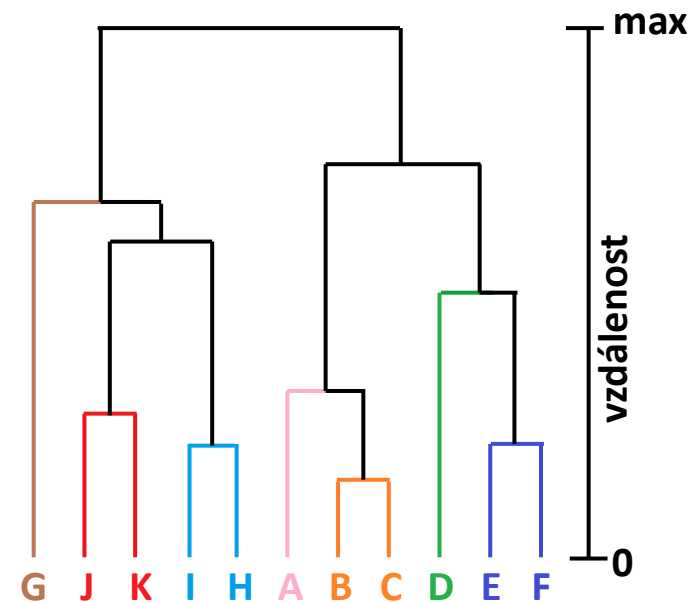
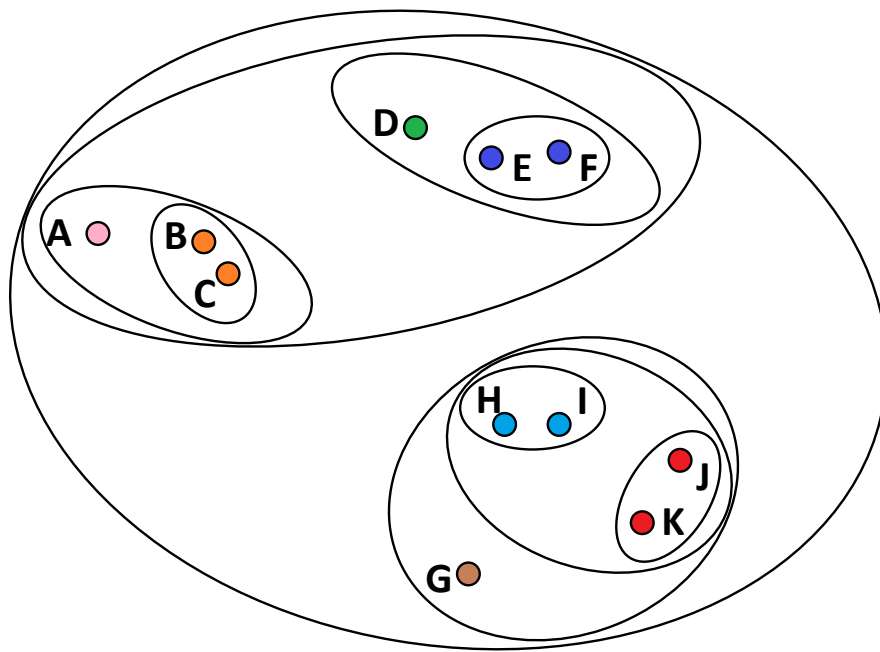
vzdálenost

0

Shluková analýza – rozhodovací proces



- 1) Výpočet **asociační matice** (pozor na správný výběr metriky vzdálenosti / podobnosti).
- 2) Výběr shlukovacího **algoritmu**. ←
- 3) Volba **počtu shluků**.



Výběr vhodného shlukovacího algoritmu



Kofenetická matice

- Matice dimenze $n \times n$ (n = počet objektů) popisující vzdálenost, kdy byly objekty poprvé spojeny do jednoho shluku.
- Hodnoty kofenetické matice závisí na typu algoritmu shlukování.

Dendrogram

Kofenetická matice



	A	B	C	D	E
A	0				
B		0			
C			0		
D				0	
E					0

Matice je symetrická
podél diagonály

Vzdálenost, kdy došlo k
prvnímu spojení D+C

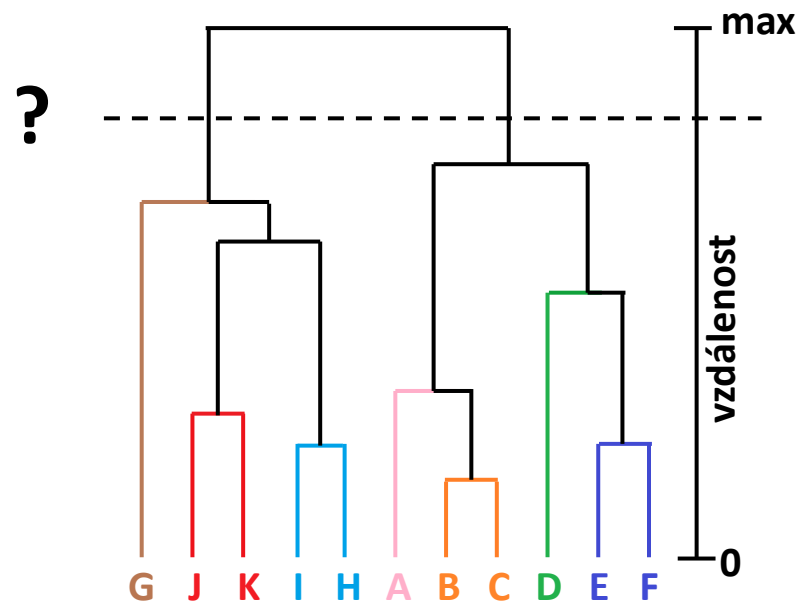
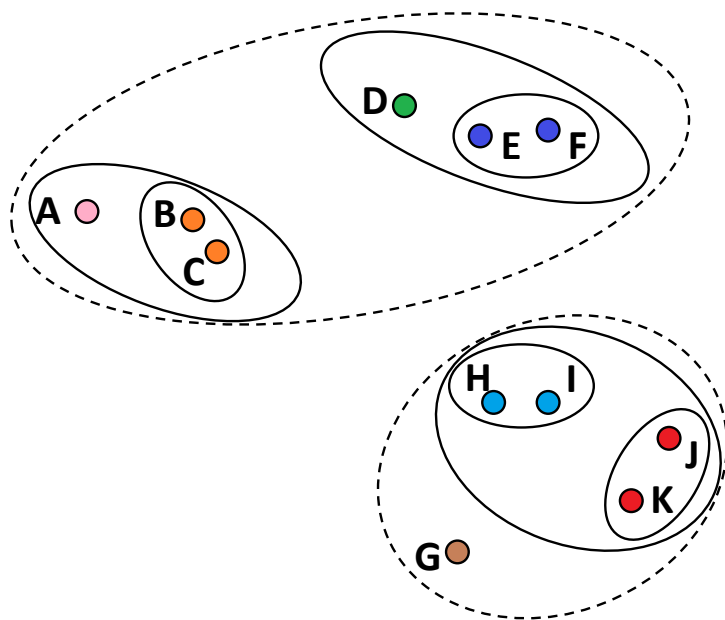
Kofenetický index

- Korelace kofenetické matice s původní maticí vzdáleností. Čím vyšší korelace, tím lepší algoritmus (algoritmus lépe popisuje realitu).

Shluková analýza – rozhodovací proces



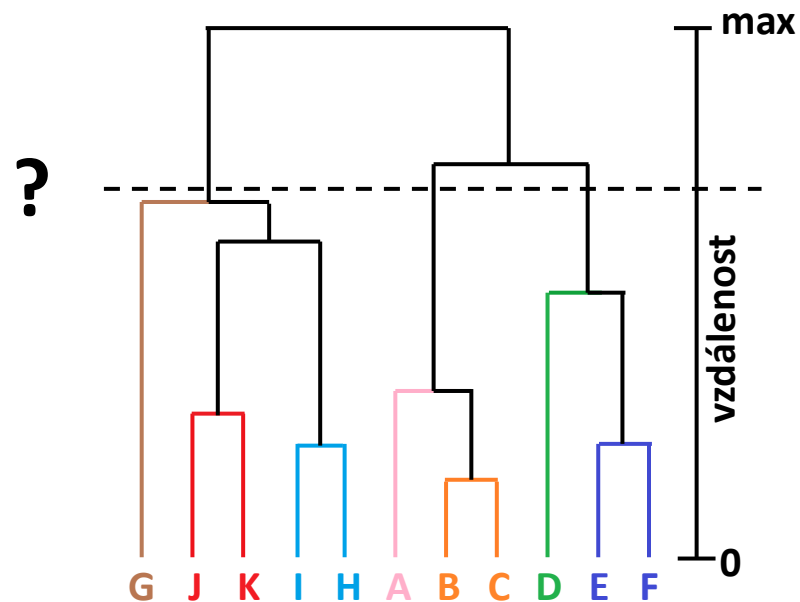
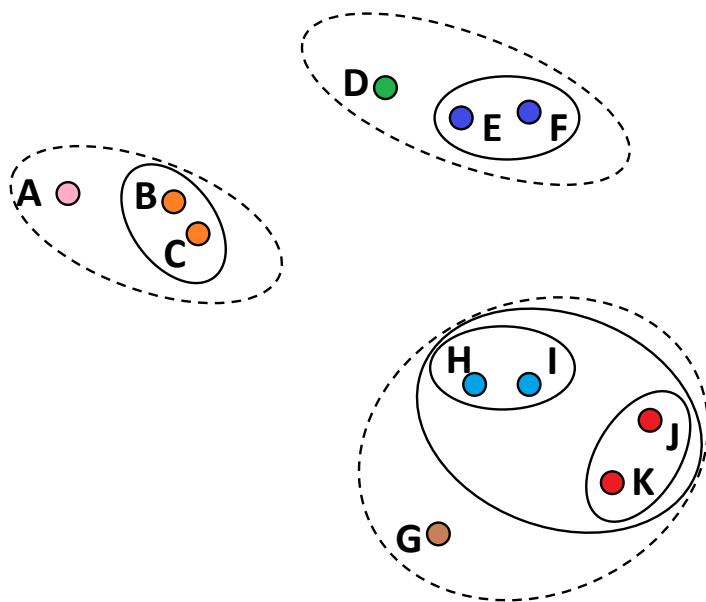
- 1) Výpočet **asociační matice** (pozor na správný výběr metriky vzdálenosti / podobnosti).
- 2) Výběr shlukovacího **algoritmu**.
- 3) Volba **počtu shluků**. ←



Shluková analýza – rozhodovací proces



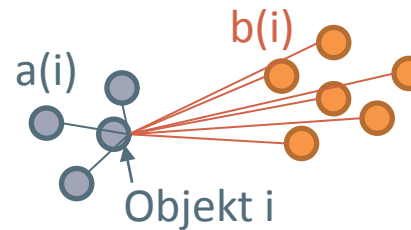
- 1) Výpočet **asociační matice** (pozor na správný výběr metriky vzdálenosti / podobnosti).
- 2) Výběr shlukovacího **algoritmu**.
- 3) Volba **počtu shluků**. 



Určení optimálního počtu shluků I



- **Subjektivní** rozhodování podle:
 - 1) počtu objektů ve shluku,
 - 2) vzdálenosti shluků,
 - 3) na základě charakteru dat.
- **Objektivní** např. pomocí **Silhouette indexu**, kde $a(i)$ je průměrná vzdálenost objektu ke všem ostatním objektům v daném shluku a $b(i)$ je nejmenší průměrná vzdálenost objektu i k objektům ostatních shluků (odkazuje tedy na vzdálenost k sousednímu shluku).



- Platí: $-1 \leq s(i) \leq 1$.
- $s(i)$ blízké -1 značí špatné zařazení do shluku, blízké 1 správné zařazení do shluku, hodnoty blízké 0 značí, že objekt leží na hranici dvou shluků.
- Počítá se průměr $s(i)$ v rámci shluků a do grafu vykreslujeme průměr $s(i)$ pro všechny shluky. Počet shluků s nejvyšší hodnotou celkového $s(i)$ odkazuje na nejlepší dělení souboru. →

Nakonec ale stejně může vyhrát naše subjektivní rozhodnutí 😊

Určení optimálního počtu shluků II



- Objektivní pomocí **Mantelova testu**.
- Hodnotíme korelaci původní asociační matice vzdáleností a asociační matice (vypočítanou pomocí Gowerova indexu), která obsahuje 1, pokud jsou spolu objekty ve shluku a 0 pokud nejsou. R si matici určující současný výskyt ve shluku převede na vzdálenosti – tedy 0 pokud jsou spolu objekty ve shluku a 1 pokud nejsou.

shluky A, B+C, D+E

	A	B	C	D	E
A	1	0	0	0	0
B	0	1	1	0	0
C	0	1	1	0	0
D	0	0	0	1	1
E	0	0	0	1	1



matice vzdáleností

	A	B	C	D	E
A	0	1	1	1	1
B	1	0	0	1	1
C	1	0	0	1	1
D	1	1	1	0	0
E	1	1	1	0	0

vs.

asociační matice

	A	B	C	D	E
A	0	5.0	6.2	11.8	11.7
B	5.0	0	3.5	11.0	9.3
C	6.2	3.5	0	4.0	4.8
D	11.8	11.0	4.0	0	2.4
E	11.7	9.3	4.8	2.4	0

- **Kladná korelace (nízká vzdálenost → objekty jsou spolu ve shluku) nám říká, že objekty sobě podobné leží spolu ve shluku.**
- Počet shluků s nejvyšší hodnotou korelace odkazuje na nejlepší dělení souboru.

Funkce v R – shluková analýza



`cluster<-hclust(dist(data), method='single')` = provede shlukovou analýzu

`plot(cluster)` = vykreslí dendrogram

`cutree(cluster, k=3)` = klasifikuje objekty do 3 skupin podle vzdáleností v dendrogramu

`cutree(cluster, h=3)` = klasifikuje objekty do skupin na vzdálenosti 3 v dendrogramu