

# Bi8600: Vícerozměrné metody

## 2. cvičení



Vícerozměrné rozdělení dat  
Koeficienty podobnosti a vzdálenosti  
Asociační matice  
Shluková analýza

# Kendalovo tau - $\tau_k$



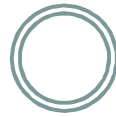
- Neparametrický přístup vyhodnocení asociace mezi dvěma spojitými/ordinálními parametry

extraverze (x)	3	4	5	5	6	7	10	12	13	14	
šťestí (y)	12	6	7	10	4	8	10	15	9	18	
konkordance	2	7	5	2	5	4	2	1	1	0	29
diskordance	7	1	1	3	0	0	1	1	0	0	14

$$t_k = (P-Q)/[n*(n-1)/2]$$

<https://is.muni.cz/el/1423/jaro2005/PSY117/um/t10/507763/>

# ANOVA



- **Počet faktorů:** jednoduché x dvojné x trojné, ... Třídění (podle počtu kategoriálních proměnných, jejichž vliv zkoumáme – one-way, two-way); možná interakce mezi faktory
- **Počet proměnných:** jednorozměrná x vícerozměrná analýza rozptylu (dle počtu spojitých parametrů, jejichž hodnoty v rámci skupin srovnáváme)

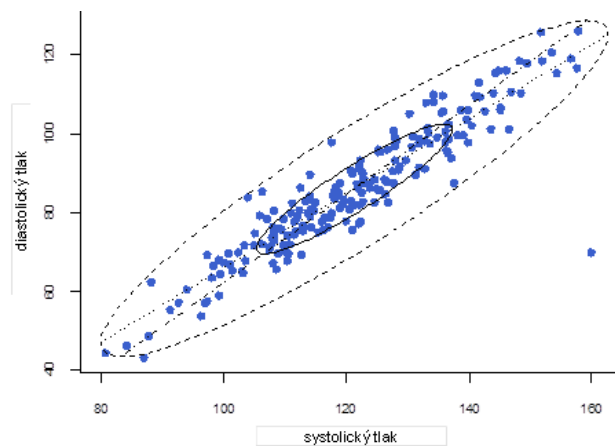
# Jak vizualizujeme vícerozměrný prostor?



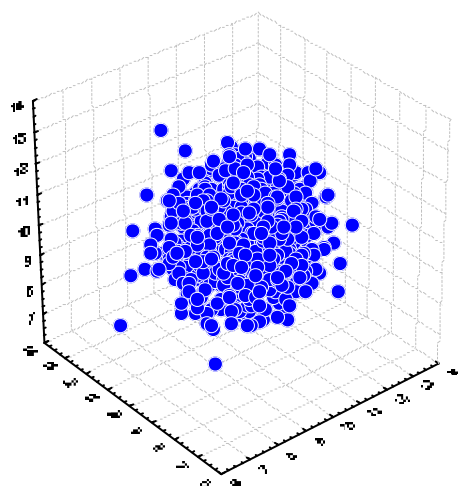
# Jak vizualizujeme vícerozměrný prostor I



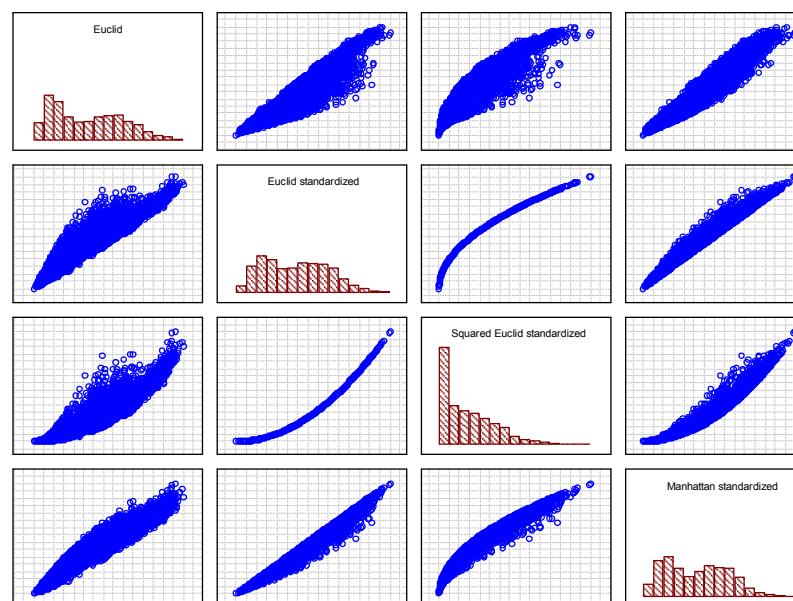
2D



3D



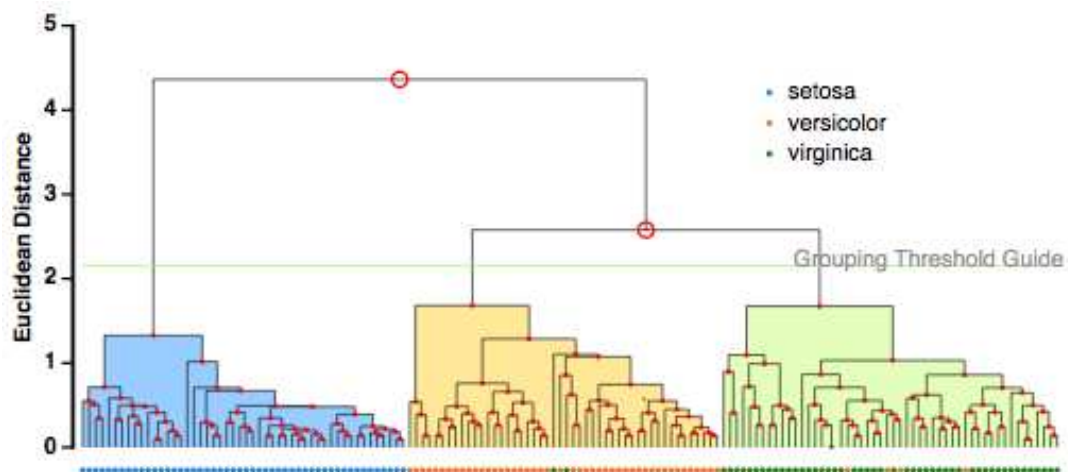
Maticové grafy



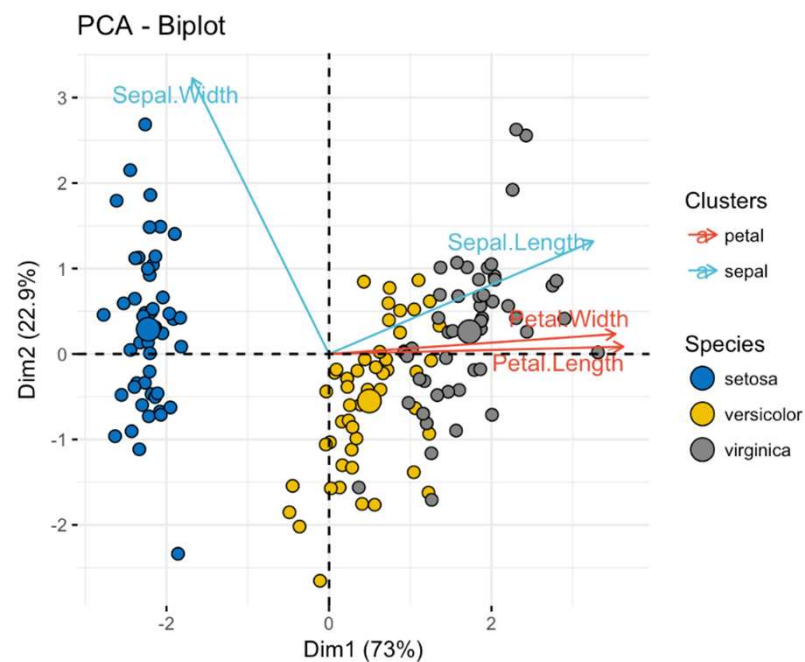
# Jak vizualizujeme vícerozměrný prostor II



## Dendrogram



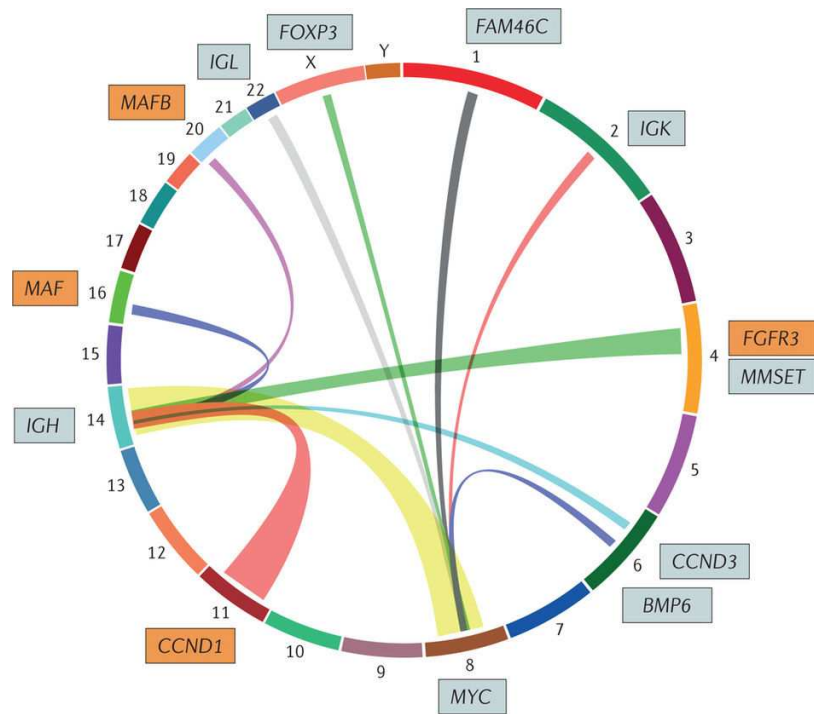
## Biplot korelací a vzdáleností



# Jak vizualizujeme vícerozměrný prostor - jiné

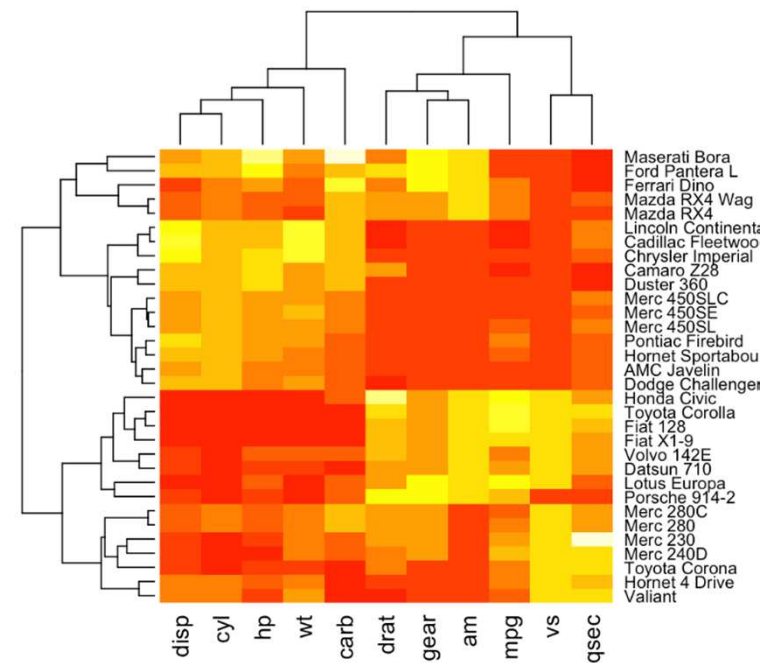


**Circos**



Nature Reviews | Clinical Oncology

**Heatmap**



# Jak popíšeme vícerozměrný prostor?





# Popisné statistiky vícerozměrných dat

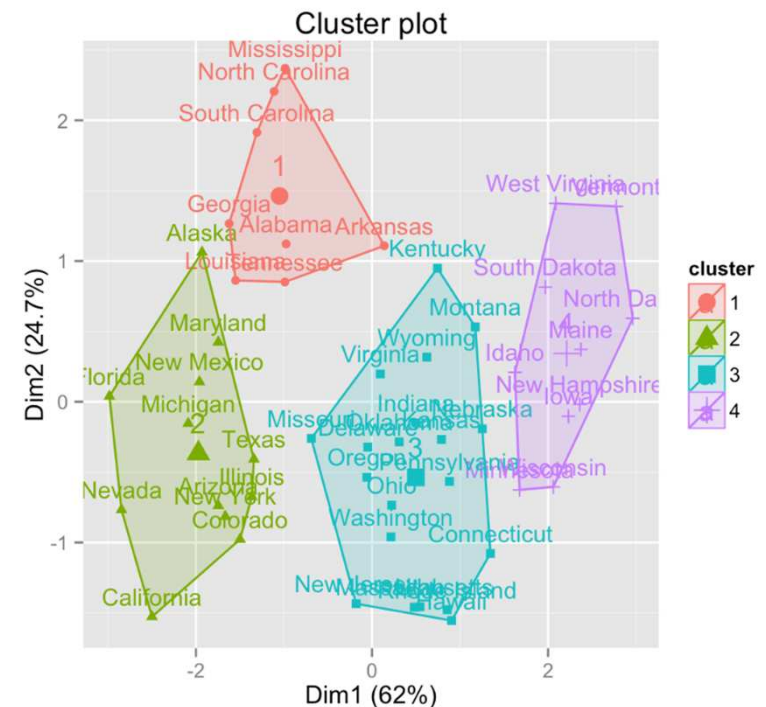


## Charakteristiky polohy středu

- Udávají, kolem jaké hodnoty se data centrují.
- Centroid = vektor průměrných hodnot, reprezentuje virtuální střed.
- Medoid = reprezentuje reálný objekt.

## Charakteristiky variability

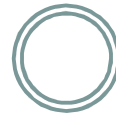
- Zachycují rozptýlení hodnot v souboru.
- Kovarianční matice.
- Korelační matice.



# Jaký je vztah mezi kovariancí a korelací?



# Jaký je vztah mezi kovariancí a korelací?



- **Kovariance** popisuje vztah dvou proměnných; její rozsah závisí na variabilitě dat.

$$C(x_1, x_2) = \frac{\sum_{i=1}^n (x_1 - \bar{x}_1)(x_2 - \bar{x}_2)}{n-1}; C \in (-\infty; \infty)$$

- **Korelace** = kovariance standardizovaná na rozptyl proměnných.

$$r(x_1, x_2) = \frac{C(x_1, x_2)}{\sqrt{D(x_1)}\sqrt{D(x_2)}}; r \in \langle -1; 1 \rangle$$

- Jaké hodnoty se nachází na diagonále korelační a kovarianční matice?
- Má smysl použít metody redukce dimenzionality dat v situaci, kdy jsou hodnoty kovariance/korelace blízké nule?
- Čemu odpovídá kovariance na standardizovaných datech?

# Chybějící data



Určete celkovou velikost souboru, která bude vstupovat do analýzy:

- A) Je průměrná hodnota systolického tlaku rovna 120 mmHg?
- B) Lze pacienty klasifikovat do skupin na základě systolického tlaku, tepové frekvence a saturace krve kyslíkem?

ID	Systolický tlak (mmHg)	Tepová frekvence (/min)	Saturace krve kyslíkem (%)
Xx_001	110	68	92
Xx_002	135	71	95
Xx_003	170	66	83
Xx_004	110	92	92
Xx_005	130		98
Xx_006	145	90	93
Xx_007	160	68	

# Chybějící data - řešení



- 1) **„Complete case analysis“** – do analýzy zahrnujeme pouze pacienty, kteří mají kompletně vyplněná data → můžeme přijít o velké množství dat a tedy i jejich reprezentativnost.
- 2) **Imputace chybějících hodnot** – pomocí statistických přístupů odhadneme chybějící data → vnášíme do dat chybu, ale zachováváme jejich reprezentativnost.

Např. balíček „mice“ (Multivariate Imputation by Chained Equations) 

## „Complete case analysis“

ID	Systolický tlak (mmHg)	Tepová frekvence (/min)	Saturace krve kyslíkem (%)
Xx_001	110	68	92
Xx_002	135	71	95
Xx_003	170	66	83
Xx_004	110	92	92
<del>Xx_005</del>	<del>130</del>		<del>98</del>
Xx_006	145	90	93
<del>Xx_007</del>	<del>160</del>	<del>68</del>	

## Imputace chybějících hodnot

ID	Systolický tlak (mmHg)	Tepová frekvence (/min)	Saturace krve kyslíkem (%)
Xx_001	110	68	92
Xx_002	135	71	95
Xx_003	170	66	83
Xx_004	110	92	92
Xx_005	130	80	98
Xx_006	145	90	93
Xx_007	160	68	95

# Co je to asociační matice?



- Jaké dimenze nabývá asociační matice?
- Co se nachází na diagonále asociační matice?
- Je matice symetrická kolem diagonály?

# 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  $n_1$ .

# Asociační matice – R mode analýza

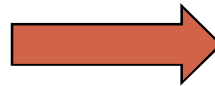


**NxP datová tabulka**

	$p_1$	$p_2$	$p_3$
$n_1$			
$n_2$			
$n_3$			
$n_4$			
$n_5$			

Hodnota subjektu  $n_5$  v parametru  $p_1$

Výpočet korelační/  
kovarianční matice



**PxP asociační matice**

	$p_1$	$p_2$	$p_3$
$p_1$	1		
$p_2$		1	
$p_3$			1

Vztah parametru  $p_3$  a parametru  $p_1$ .

## Obecně:

- Základní výběr koeficientu je často spjat s metodou/algoritmem.
- Dále je potřeba zohlednit typ vstupních dat: spojitá/kategoriální/mix.
- Výběrem metriky ovlivníme výsledky analýz.



# Koeficienty vzdálenosti



## Kvantitativní data

# Koeficienty vzdálenosti

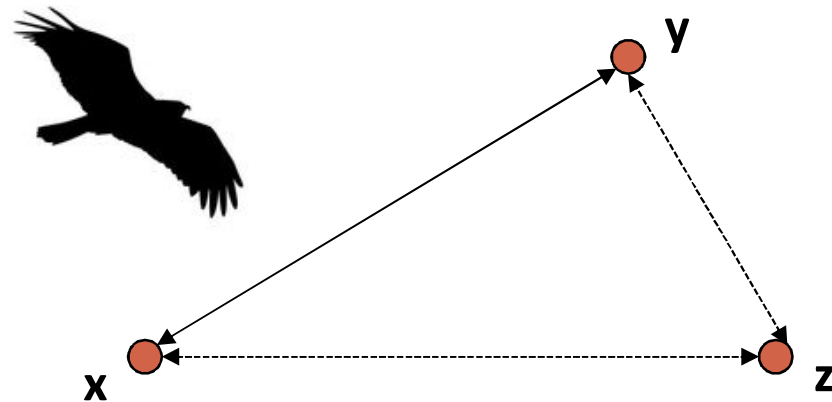


(i)  $d(x, x) = 0$ ;

(ii)  $d(x, y) > 0, x \neq y$ ;

(iii)  $d(x, y) = d(y, x)$ ;

(iv)  $d(x, z) \leq d(x, y) + d(y, z)$



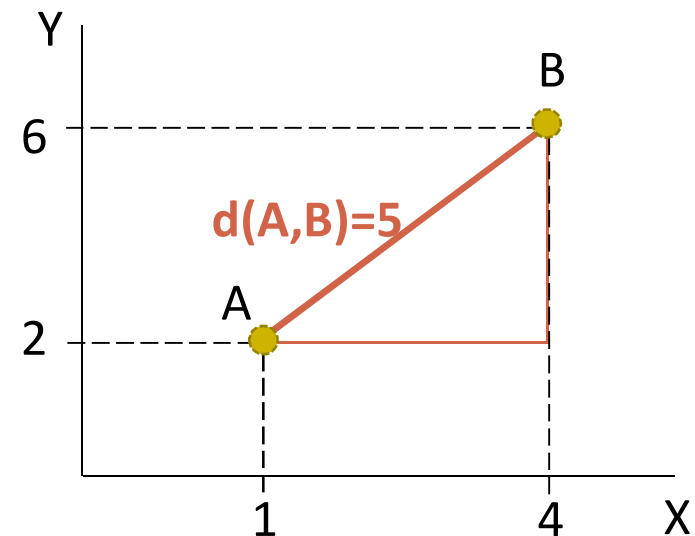
Podrobný přehled koeficientů vzdáleností a podobností najdete v knize **LEGENDRE, P. & LEGENDRE, L. (1998). *Numerical ecology*. Elsevier Science BV, Amsterdam.**

# Euklidova vzdálenost I



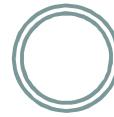
- Euklidova vzdálenost vychází z Pythagorovy věty:

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



- Jaká by byla vzdálenost bodů A a B dle Manhattanské metriky?

# 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?

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

# Euklidova vzdálenost - příklad



- Pomocí MS Excel spočítejte Euklidovu vzdálenost subjektu  $n_2$  a  $n_3$  pro následující dva datové zdroje.

	BMI	váha	výška	cholesterol
$n_2$	24.9	72	170	5,1
$n_3$	25.8	98	195	5,2

$$D(n_2, n_3) = ?$$

	BMI	váha	výška	cholesterol
$n_2$	24.9	72	170	5,1
$n_3$	25.8	98	195	2,9

$$D(n_2, n_3) = ?$$

# Euklidova vzdálenost - příklad



- Pomocí MS Excel spočítejte Euklidovu vzdálenost subjektu  $n_2$  a  $n_3$  pro následující dva datové zdroje.

	BMI	váha	výška	cholesterol
$n_2$	24.9	72	170	5,1
$n_3$	25.8	98	195	5,2

$$D(n_2, n_3) = 36,08$$

	BMI	váha	výška	cholesterol
$n_2$	24.9	72	170	5,1
$n_3$	25.8	98	195	2,9

$$D(n_2, n_3) = 36,15$$

Proč je vzdálenost téměř shodná, když se v druhém datovém souboru subjekty významně liší v hladině cholesterolu?

# Euklidova vzdálenost III



	BMI	váha	výška
$n_1$	35.6	80	150
$n_2$	24.9	72	170
$n_3$	25.8	98	195
$n_4$	22.2	54	156
$n_5$	19.3	55	169

## POZOR!

- U větších datových souborů, u kterých se často vyskytují korelované proměnné, dochází k nadhodnocení výsledků těmito korelovanými proměnnými = stejná informace je započtena více než jednou.
- Je potřeba zohlednit vztahy parametrů v datech → Mahalanobisova vzdálenost.

# Mahalanobisova vzdálenost



$$D^2 = (x - \bar{x})^T S^{-1} (x - \bar{x})$$

Matice vzdáleností hodnot  
od průměru

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} - \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_n \end{bmatrix}$$

Inverze kovarianční matice

$$\begin{bmatrix} s_1^2 & \dots & \text{Cov}(s_n, s_1) \\ \vdots & \ddots & \vdots \\ \text{Cov}(s_1, s_n) & \dots & s_n^2 \end{bmatrix}$$

- Odstraňuje vliv korelovaných parametrů.
- Dle volby  $(x - \bar{x})$  lze hodnotit:
  - 1) **vzdálenosti objektů od centroidů** (vstupem je matice rozdílů původních hodnot od průměru).
  - 2) **vzdálenosti skupin objektů** (vstupem je matice rozdílů průměrných hodnot).
  - 3) **párové vzdálenosti jednotlivých subjektů** (vstupem je matice rozdílů srovnávaných subjektů).



# Koeficienty podobnosti



## Binární data

# Koeficienty podobnosti



- Pokud proměnné popisují výskyt/nevýskyt = jsou tedy binárního typu, lze podobnost/odlišnost subjektů hodnotit dle tabulky níže:

	1	0	
1	a	b	a + b
0	c	d	c + d
	a + c	b + d	p = a + b + c + d

**Co je to problém „double zero“?**

# Koeficienty podobnosti příklad I

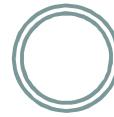


- Úkol: Na základě datové matice doplňte tabulku.

	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	$x_6$	$x_7$
$n_1$	1	1	0	0	0	0	1
$n_2$	1	0	1	1	1	0	0

		$n_1$		
		<b>1</b>	<b>0</b>	
$n_2$	<b>1</b>	a =	b =	a + b =
	<b>0</b>	c =	d =	c + d =
		a + c =	b + d =	p = a + b + c + d =

# Koeficienty podobnosti příklad I



- Úkol: Na základě datové matice doplňte tabulku.

	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	$x_6$	$x_7$
$n_1$	1	1	0	0	0	0	1
$n_2$	1	0	1	1	1	0	0

		$n_1$		
		1	0	
$n_2$	1	$a = 1$	$b = 3$	$a + b = 4$
	0	$c = 2$	$d = 1$	$c + d = 3$
		$a + c = 3$	$b + d = 4$	$p = a + b + c + d = 7$

**Já znáte koeficienty podobnosti?**

# Koeficienty podobnosti příklad II



- Úkol:

- 1) Přiřadte uvedené vzorce ke koeficientům podobnosti: „simple matching“, Jaccardův a Sørensenův koeficient podobnosti.
- 2) Na základě získané tabulky spočítejte uvedené koeficienty.
- 3) Podobnosti převedte na vzdálenosti.

		$n_1$		
		<b>1</b>	<b>0</b>	
$n_2$	<b>1</b>	a = 1	b = 3	a + b = 4
	<b>0</b>	c = 2	d = 1	c + d = 3
		a + c = 3	b + d = 4	p = a + b + c + d = 7

- $S_{\text{simple matching}} = \rightarrow D =$
- $S_{\text{Jaccard}} = \rightarrow D =$
- $S_{\text{Sørensen}} = \rightarrow D =$

1. 
$$\frac{2a}{2a+b+c}$$

2. 
$$\frac{a+d}{a+b+c+d}$$

3. 
$$\frac{a}{a+b+c}$$

# Koeficienty podobnosti příklad II



- Úkol:

- 1) Přiřadte uvedené vzorce ke koeficientům podobnosti: „simple matching“, Jaccardův a Sørensenův koeficient podobnosti.
- 2) Na základě získané tabulky spočítejte uvedené koeficienty.
- 3) Podobnosti převedte na vzdálenosti.

		$n_1$		
		<b>1</b>	<b>0</b>	
$n_2$	<b>1</b>	a = 1	b = 3	a + b = 4
	<b>0</b>	c = 2	d = 1	c + d = 3
		a + c = 3	b + d = 4	p = a + b + c + d = 7

- $S_{\text{simple matching}} = (1+1)/(1+2+3+1) = 0.3 \rightarrow D = 0.7$
- $S_{\text{Jaccard}} = 1/(1+2+3) = 0.2 \rightarrow D = 0.8$
- $S_{\text{Sørensen}} = 2*1/(2*1+2+3) = 0.3 \rightarrow D = 0.8$

1. 
$$\frac{2a}{2a+b+c}$$

2. 
$$\frac{a+d}{a+b+c+d}$$

3. 
$$\frac{a}{a+b+c}$$

# Sørensenův asymetrický koeficient podobnosti pro kvantitavní data



- Tabulka popisuje abundance živočichů na dvou lokalitách.
- **Úkol:** Pomocí Sørensenova koeficientu vyhodnoťte, zda jsou uvedené lokality podobné.

Lokalita	Výskyt/nevýskyt živočicha							aN	bN	jN
	žralok	velryba	had	ještěrka	velbloud	varan	tučňák			
Vysočina	0	0	2	3	0	0	0	5		
Sahara	0	0	4	1	5	6	0		16	
Minimum	0	0	2	1	0	0	0			3

$$C_N = \frac{2jN}{(aN + bN)} = \frac{2 \cdot 3}{(5 + 16)} = 0.3$$

# Gowerův obecný koeficient podobnosti



Mix kategoriálních a kvantitativních dat



# Gowerův obecný koeficient podobnosti



- Kombinuje různé typy deskriptorů.
- Podobnost mezi dvěma objekty je vypočítána jako průměr podobností, vypočítaných pro všechny deskriptory:

$$S_{15}(x_1, x_2) = \frac{1}{p} \sum_{j=1}^p s_{12j}$$

- ✓ Pro kategoriální deskriptory  $s = 1$  (shoda) nebo  $0$  (neshoda).
- ✓ Kvantitativní deskriptory (reálná čísla): rozdíl mezi stavy obou objektů je vydělen největším rozdílem ( $R_j$ ), nalezeným pro daný deskriptor mezi všemi objekty ve studii.

# Asociační matice



# Asociační matice vzdáleností



STATISTICA - [Data: Irisdat\* (5v by 150c)]

File Edit View Insert Format Statistics Data Mining Graphs Tools Data Window

Arial 10 B I U

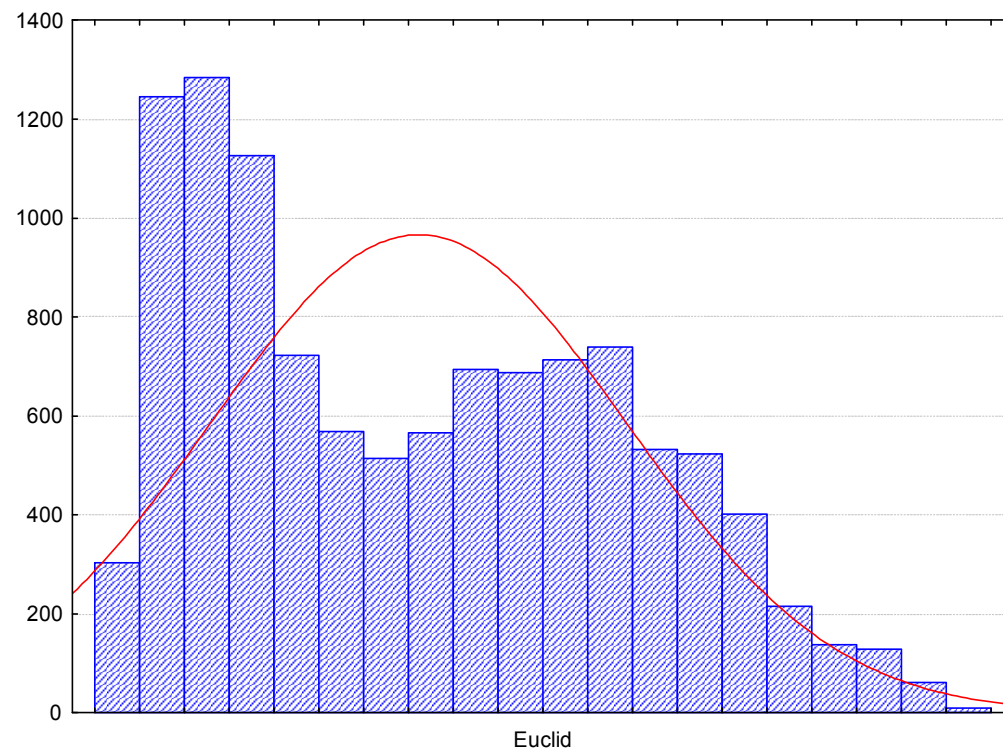
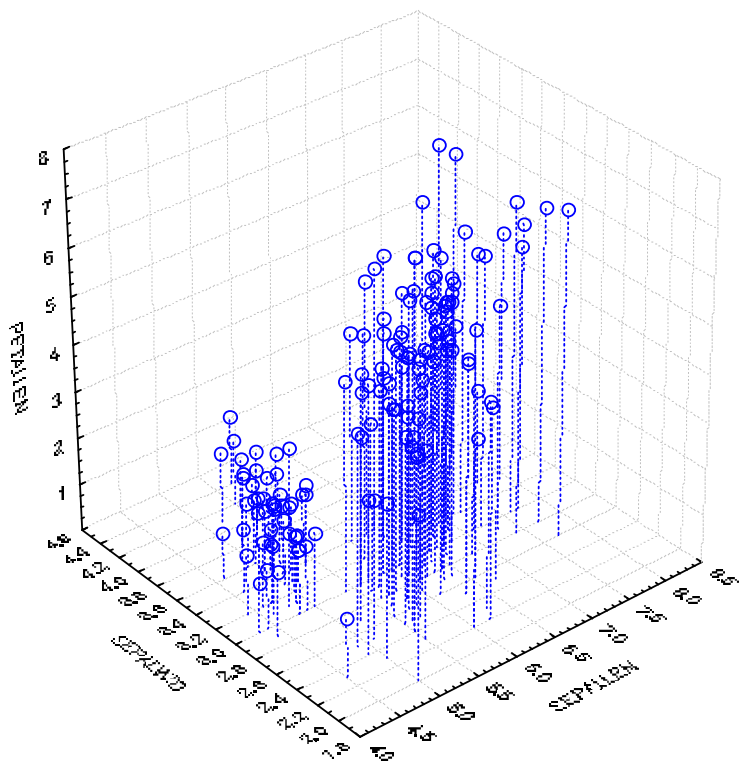
Fisher (1936) iris data: length & width of sepals and petals, 3 types of I

	1	2	3	4	5
	SEPALLEN	SEPALWID	PETALLEN	PETALWID	IRISTYPE
1	5.0	3.3	1.4	0.2	SETOSA
2	6.4	2.8	5.6		
3	6.5	2.8	4.6		
4	6.7	3.1	5.6		
5	6.3	2.8	5.1		1.5 VIRGINIC
6	4.6	3.4	1.4	0.3	SETOSA
7	6.9	3.1	5.1	2.3	VIRGINIC
8	6.2	2.2	4.5	1.5	VERSICO
9	5.9	3.2	4.8	1.8	VERSICO
10	4.6	3.6	1.0	0.2	SETOSA
11	6.1	3.0	4.6	1.4	VERSICO
12	6.0	2.7	5.1	1.6	VERSICO
13	6.5	3.0	5.2	2.0	VIRGINIC
14	5.6	2.5	3.9	1.1	VERSICO
15	6.5	3.0	5.5	1.8	VIRGINIC
16	5.8	2.7	5.1	1.9	VIRGINIC
17	6.8	3.2	5.9	2.3	VIRGINIC
18	5.1	3.3	1.7	0.5	SETOSA
19	5.7	2.8	4.5	1.3	VERSICO
20	6.2	3.4	5.4	2.3	VIRGINIC
21	7.7	3.8	6.7	2.2	VIRGINIC
22	6.3	3.3	4.7	1.6	VERSICO
23	6.7	3.3	5.7	2.5	VIRGINIC
24	7.6	3.0	6.6	2.1	VIRGINIC
25	4.9	2.5	4.5	1.7	VIRGINIC
26	5.5	3.5	1.3	0.2	SETOSA
27	6.7	3.0	5.2	2.3	VIRGINIC
28	7.0	3.2	4.7	1.4	VERSICO
29	6.4	3.2	4.5	1.5	VERSICO
30	6.1	2.8	4.0	1.3	VERSICO
31	4.8	3.1	1.6	0.2	SETOSA
32	5.9	3.0	5.1	1.8	VIRGINIC
33	5.5	2.4	3.8	1.1	VERSICO
34	6.3	2.5	5.0	1.9	VIRGINIC

Case No.	Euclidean distances (Irisdat)																
	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_10	C_11	C_12	C_13	C_14	C_15	C_16	C_17
C_1	0.00	4.88	3.80	5.04	4.16	0.42	4.66	3.73	3.87	0.64	3.60	4.12	4.47	2.84	4.66	4.19	5.28
C_2	4.88	0.00	1.22	0.47	0.87	4.98	0.77	1.45	1.10	5.39	1.33	0.88	0.50	2.20	0.47	0.84	0.65
C_3	3.80	1.22	0.00	1.39	0.54	3.96	1.07	0.68	0.81	4.35	0.46	0.72	0.81	1.24	0.97	0.95	1.61
C_4	5.04	0.47	1.39	0.00	1.14	5.15	0.55	1.75	1.28	5.54	1.54	1.24	0.61	2.48	0.65	1.21	0.35
C_5	4.16	0.87	0.54	1.14	0.00	4.29	1.04	0.85	0.71	4.69	0.58	0.33	0.58	1.48	0.57	0.65	1.30
C_6	0.42	4.98	3.96	5.15	4.29	0.00	4.80	3.88	3.94	0.46	3.72	4.22	4.59	2.95	4.78	4.26	5.40
C_7	3.73	1.45	1.07	0.55	1.04	4.80	0.00	1.52	1.16	5.17	1.31	1.21	0.52	2.22	0.76	1.24	0.81
C_8	3.87	1.10	0.81	1.28	0.71	3.94	1.16	1.13	0.00	4.30	0.82	0.81	1.21	0.98	1.35	0.96	1.99
C_9	3.87	1.10	0.81	1.28	0.71	3.94	1.16	1.13	0.00	4.34	0.53	0.62	0.77	1.37	0.94	0.60	1.51
C_10	0.64	5.39	4.35	5.54	4.69	0.46	5.17	4.30	4.34	0.00	4.12	4.64	4.98	3.38	5.17	4.69	5.78
C_11	3.60	1.33	0.46	1.54	0.58	3.72	1.31	0.82	0.53	4.12	0.00	0.62	0.94	1.04	1.06	0.82	1.74
C_12	4.12	0.88	0.72	1.24	0.33	4.22	1.21	0.81	0.62	4.64	0.62	0.00	0.71	1.37	0.73	0.36	1.42
C_13	4.47	0.50	0.81	0.61	0.58	4.59	0.52	1.21	0.77	4.98	0.94	0.71	0.00	1.89	0.36	0.77	0.84
C_14	2.84	2.20	1.24	2.48	1.48	2.95	2.22	0.98	1.37	3.38	1.04	1.37	1.89	0.00	2.03	1.47	2.71
C_15	4.66	0.47	0.97	0.65	0.57	4.78	0.76	1.35	0.94	5.17	1.06	0.73	0.36	2.03	0.00	0.87	0.73
C_16	4.19	0.84	0.95	1.21	0.65	4.26	1.24	0.96	0.60	4.69	0.82	0.36	0.77	1.47	0.87	0.00	1.43
C_17	5.28	0.65	1.61	0.35	1.30	5.40	0.81	1.99	1.51	5.78	1.74	1.42	0.84	2.71	0.73	1.43	0.00
C_18	0.44	4.48	3.41	4.63	3.77	0.62	4.25	3.36	3.46	0.96	3.21	3.73	4.07	2.47	4.26	3.79	4.88
C_19	3.40	1.58	0.83	1.87	0.87	3.49	1.70	0.81	0.73	3.91	0.47	0.74	1.29	0.71	1.39	0.86	2.08
C_20	4.68	0.67	1.32	0.62	1.05	4.75	0.82	1.70	0.86	5.14	1.27	1.05	0.62	2.20	0.71	0.95	0.81

Asociační matice euklidovských vzdáleností mezi rostlinami

# Histogram jako popis asociční matice



# Základní funkce v R pro výpočet asociační matice



# Funkce v R pro výpočet asociační matice



## Koeficienty vzdálenosti

`dist(data, method='euclidean')` = Euklidova vzdálenost

`mahalanobis(X, X.mean, X.cov)` = Mahalanobisova vzdálenost

`pairwise.mahalanobis(X, grouping)` = Mahalanobisova vzdálenost mezi skupinami {HDMD}

## Koeficienty podobnosti

`dist.binary(data, method = 2)` = „simple matching“ koeficient {ade4}

`vegdist(data, "jac", binary=T)` = Jaccardův koeficient {vegan}

`vegdist(data, "jac", binary=F)` = Sørensenův asymetrický koeficient

**Podobnosti jsou pomocí funkce 1-podobnost automaticky převáděny na vzdálenosti!!!**

Výpočet **Mantelova testu** (testuje korelaci dvou asociačních matic – např. průběh onemocnění vs. genetická výbava pacientů, charakteristiky lokalit vs. abundance druhů na lokalitách): `mantel{vegan}`

# Shluková analýza



# Shluková analýza – jaký je cíl?

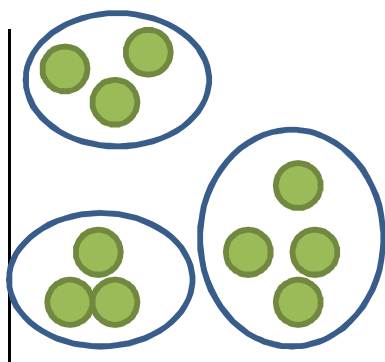




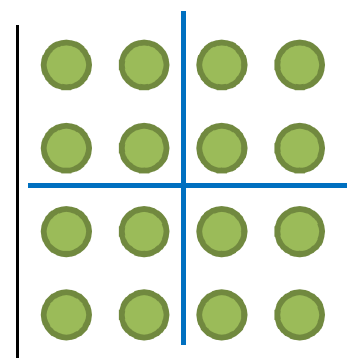
# Shluková analýza – jaký je cíl?



- 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.

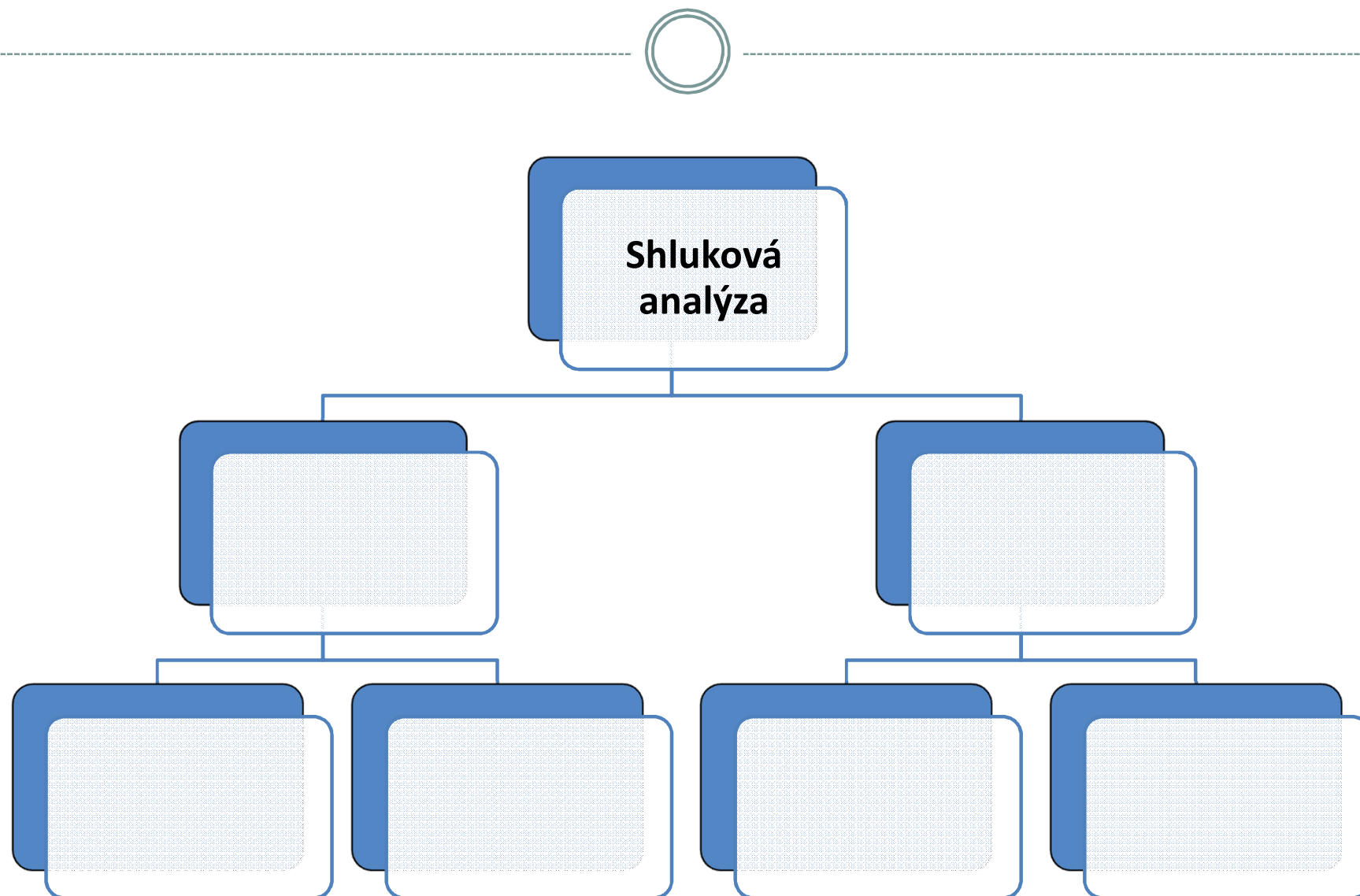


Jednoznačné odlišení existujících shluků v datech

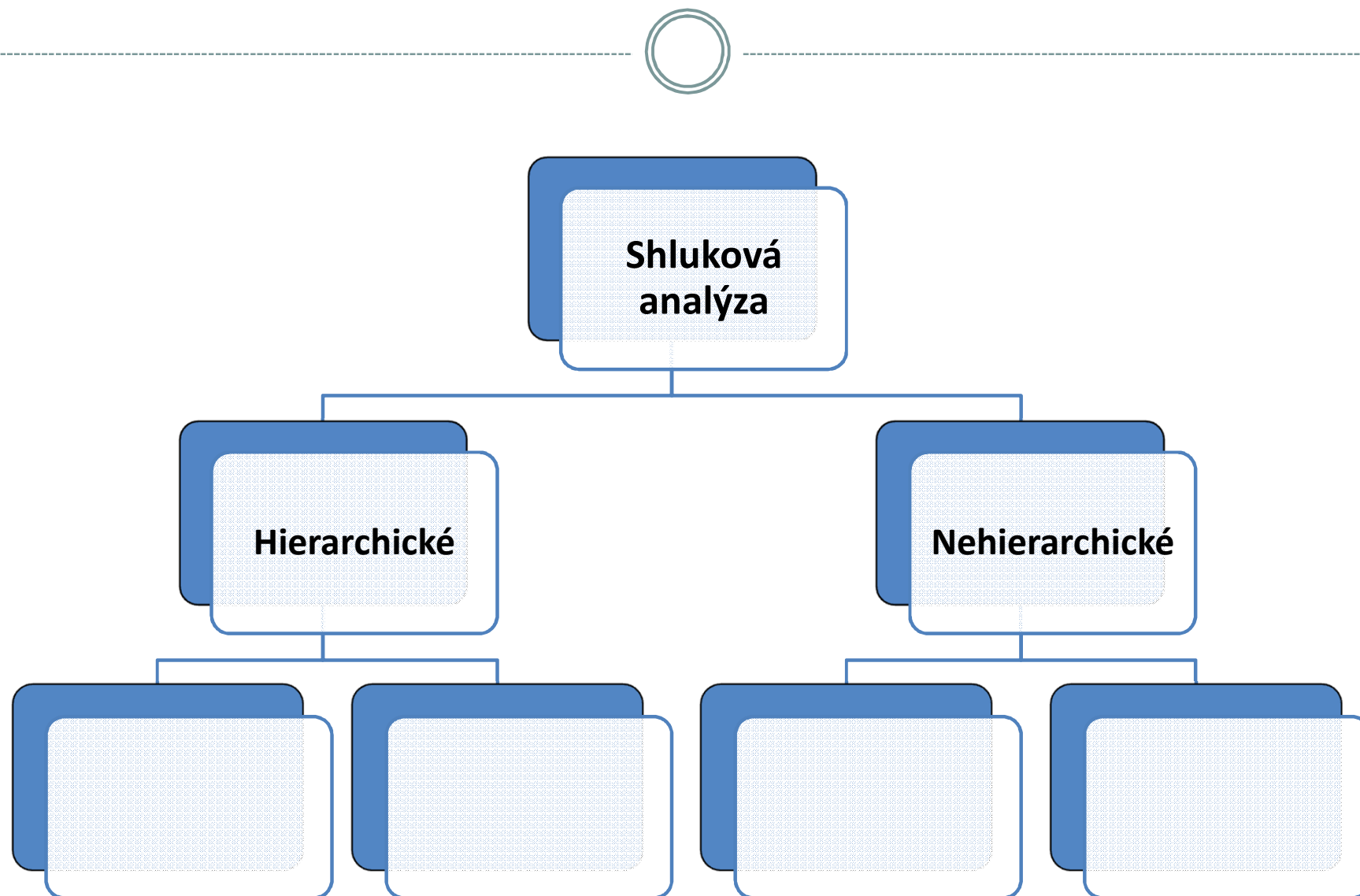


Shlukovou analýzu lze provést i na datech bez objektivní existence shluků

# Shluková analýza: typy metod



# Shluková analýza: typy metod



# Shluková analýza: typy metod

**Hierarchické**  
shluky jsou definovány postupným skládáním objektů

**Nehierarchické**  
Shluky jsou definovány v jednom kroku

**Divizivní**

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

**Aglomerativní**

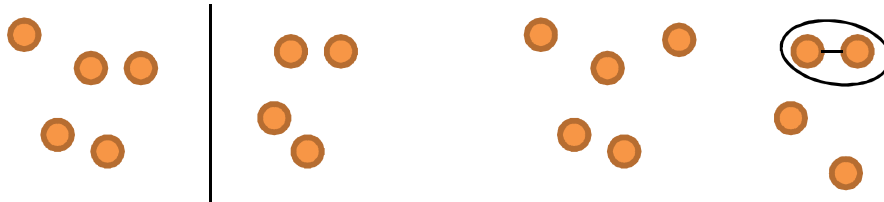
Po spojení první dvojice objektů dochází k postupnému napojování dalších objektů.

**Divizivní**

Objekty jsou rozděleny do předem nastaveného počtu shluků.

**Aglomerativní**  
sítí spojených bodů

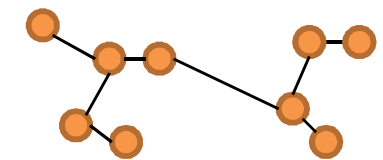
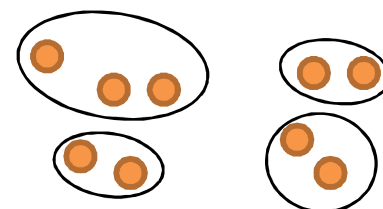
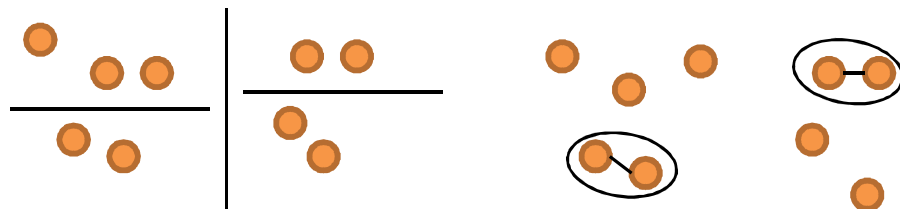
1. Krok



Kolik shluků chceme definovat? Například 4

Minimum spanning tree, Prim network

2. Krok



X. Krok

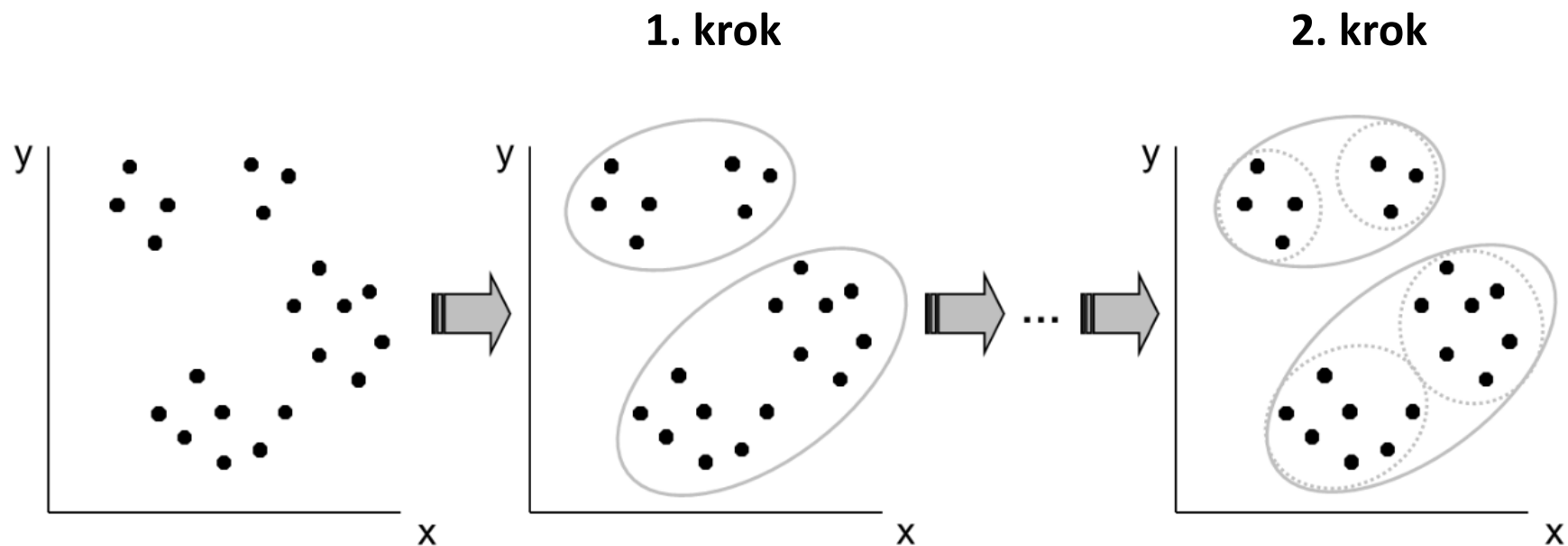
Atd.

Atd.

Výpočet ukončen

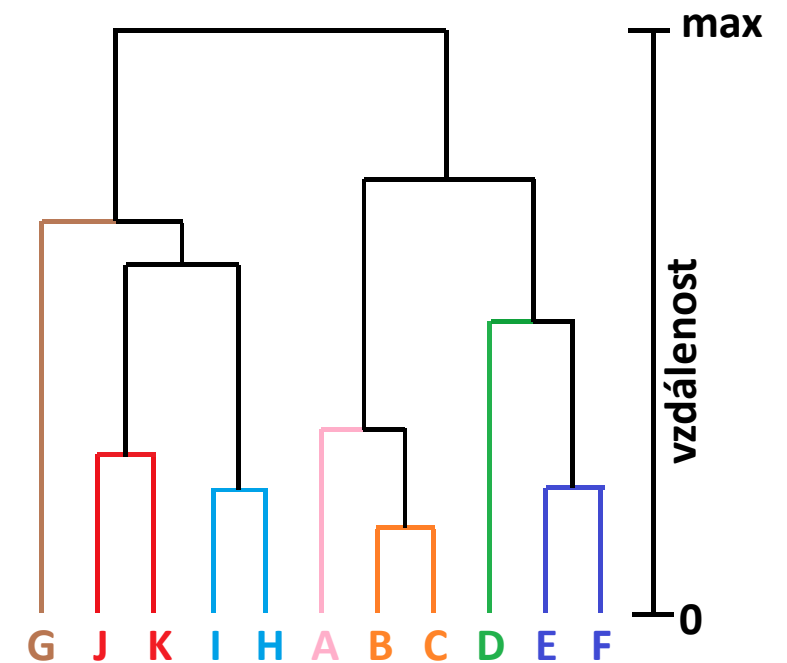
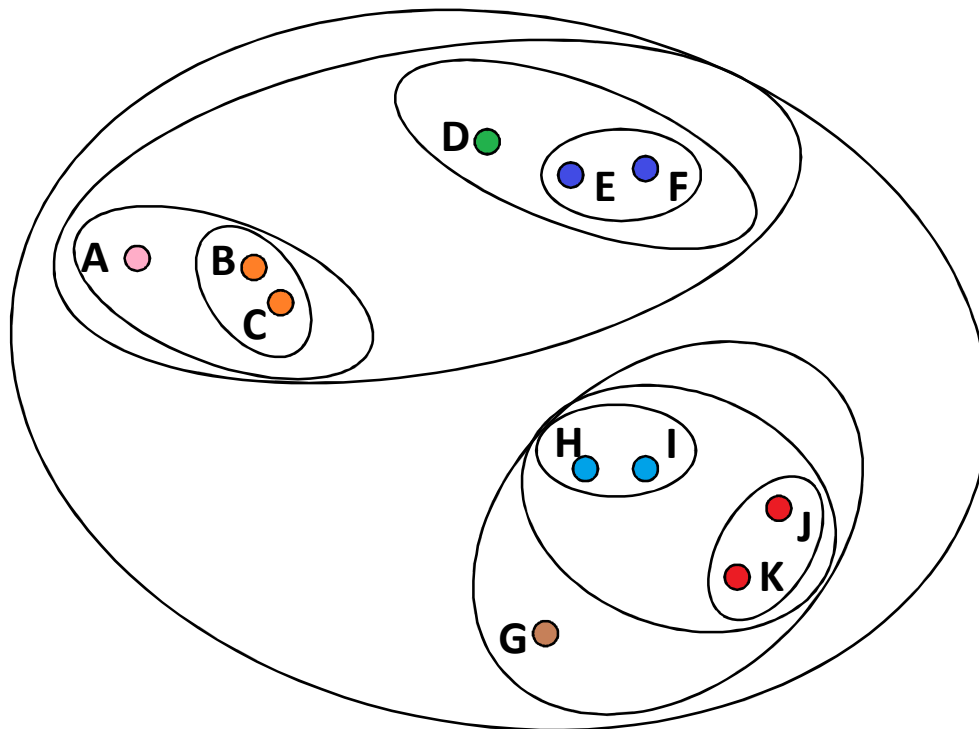
Výpočet ukončen

# Pojmenujte shlukovací algoritmus I



<http://portal.matematickabiologie.cz/>

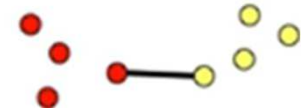
# Hierarchické aglomerativní algoritmy



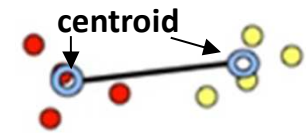
# Pojmenujte shlukovací algoritmus II



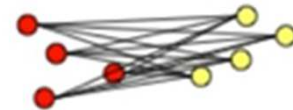
1)



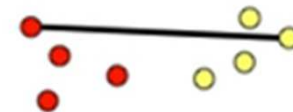
2)



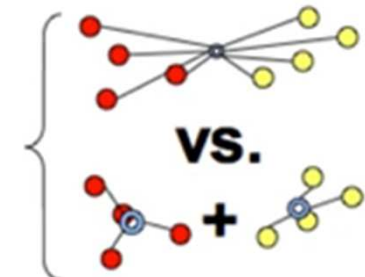
3)



4)



5)

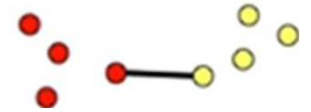


# Pojmenujte shlukovací algoritmus II



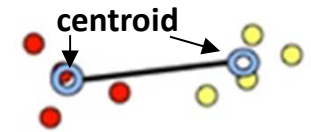
## 1) Metoda nejbližšího souseda („single linkage“)

- spojení na základě nejmenší minimální vzdálenosti dvou objektů



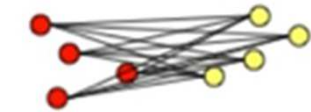
## 2) Metoda středospojná/centroidní („centroids“)

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



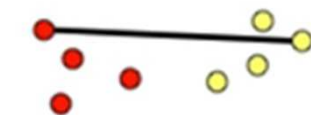
## 3) Metoda průměrné vzdálenosti („average linkage“)

- spojení na základě minimální průměrné vzdálenosti všech párů objektů dvou shluků



## 4) Metoda nejvzdálenějšího souseda („complete linkage“)

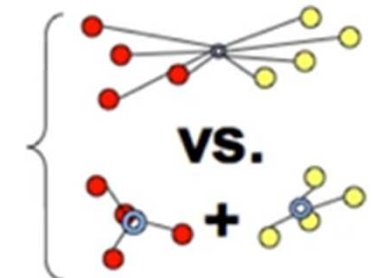
- spojení na základě nejmenší maximální vzdálenosti dvou objektů



## 5) 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

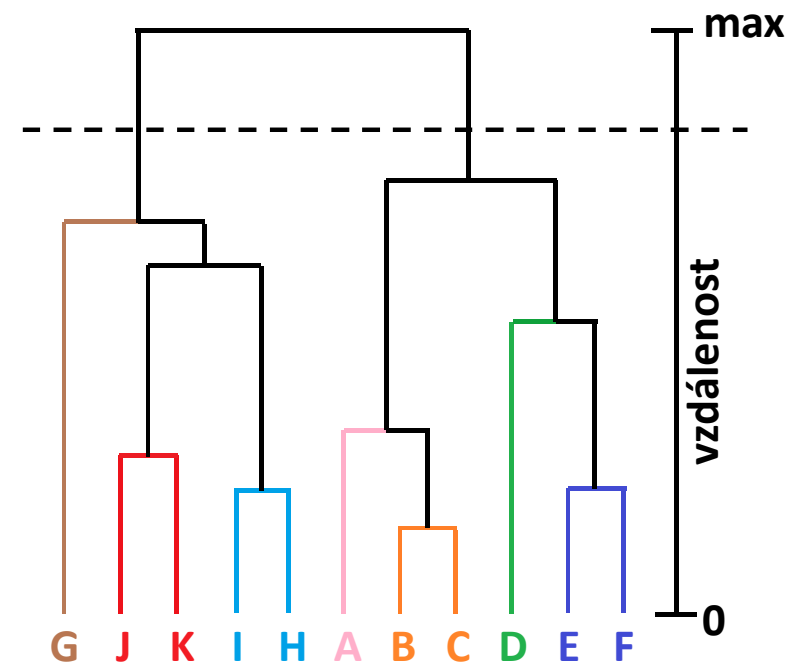
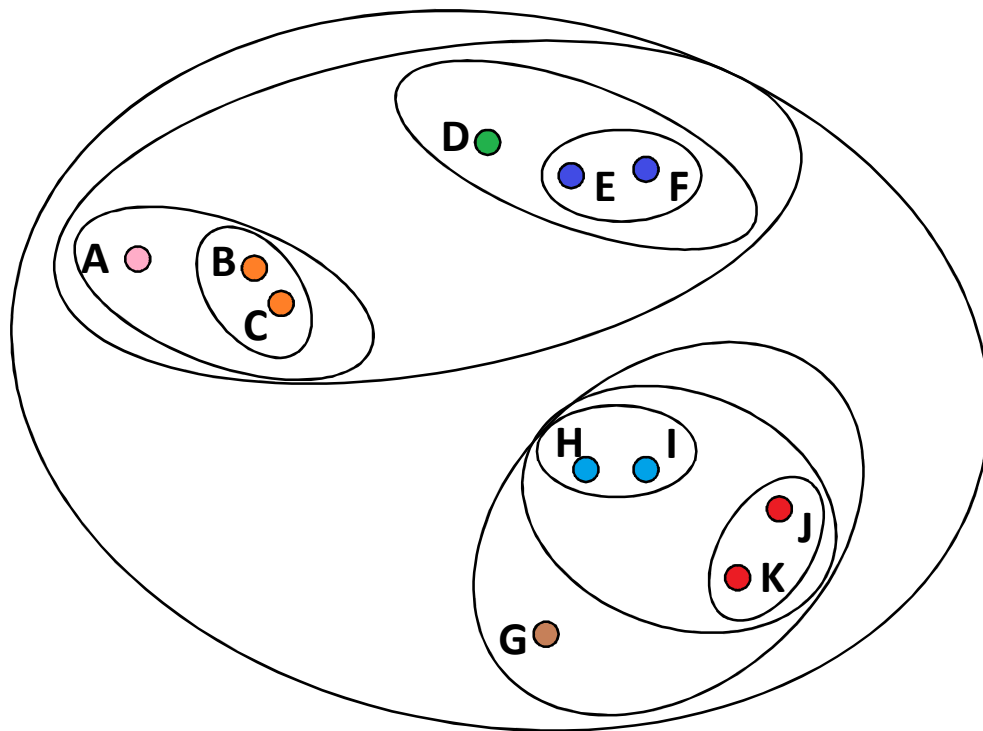
- vstupem je čtverec Euklidovy vzdálenosti





# Shluková analýza – rozhodovací proces

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



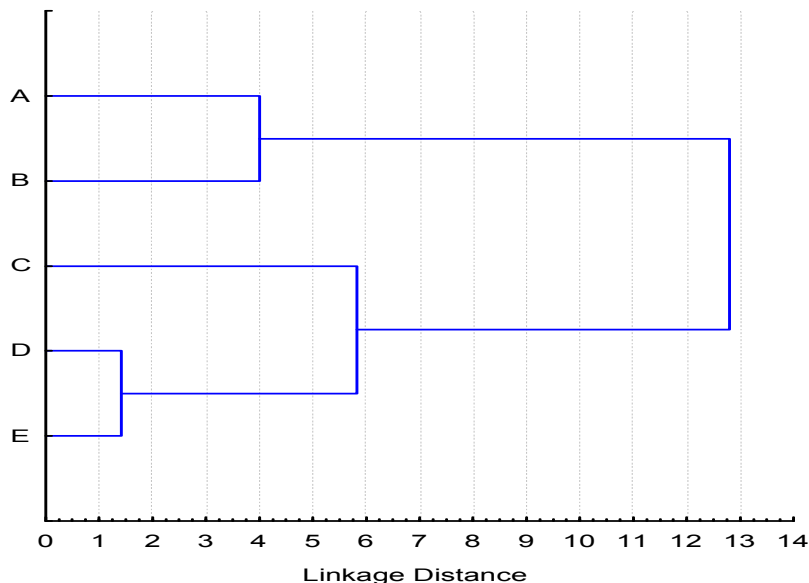
# Výběr vhodného algoritmu



## 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**

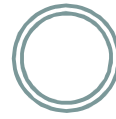
	A	B	C	D	E
A	0	4.0	12.7	12.7	12.7
B		0	12.7	12.7	12.7
C			0	5.7	5.7
D	Matice je symetrická podél diagonály			0	1.4
E					0

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).

# Určení optimálního počtu shluků I

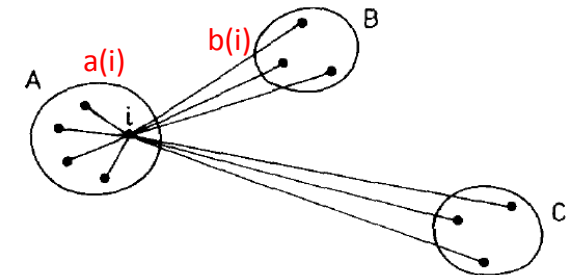



- **Subjektivní** rozhodování podle:

- 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).

$$s(i) = \frac{b(i) - a(i)}{\max\{b(i), a(i)\}}$$



- Platí:  $-1 \leq s(i) \leq 1$ .
- $s(i)$  blízke -1 značí špatné zařazení do shluku, blízke 1 správné zařazení do shluku, hodnoty blízke 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.

# 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	<i>Matice je symetrická podél diagonály</i>			
B	9	0			
C	3	7	0		
D	6	5	9	0	
E	11	10	2	8	0

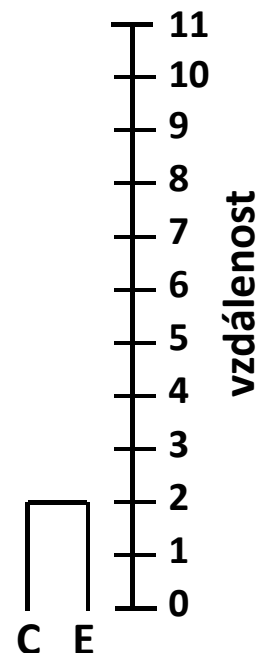
vzdálenost  
0 ?

# 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.

## 1. krok



## 2. krok

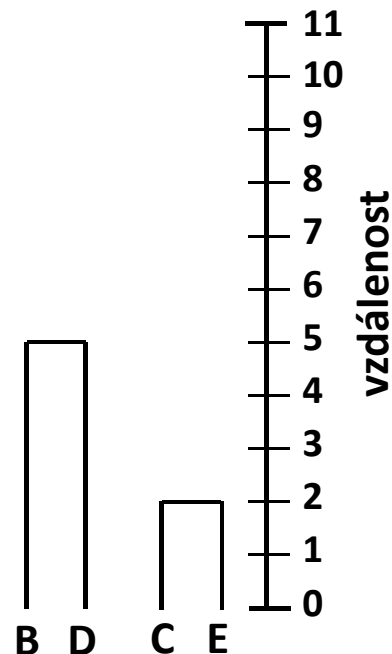
	A	B	D	C+E
A	0			
B	9	0		
D	6	5	0	
C+E	11	10	9	0

# 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.

## 2. krok



## 3. krok

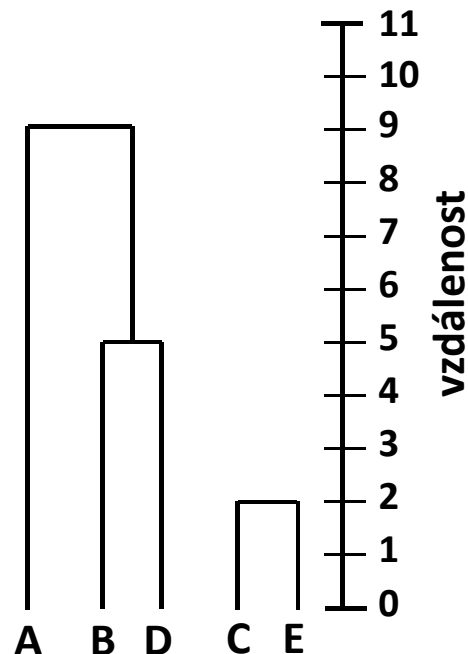
	A	B+D	C+E
A	0		
B+D	9	0	
C+E	11	10	0

# 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.

## 3. krok



## 4. krok

	A+B+D	C+E
A+B+D	0	
C+E	11	0

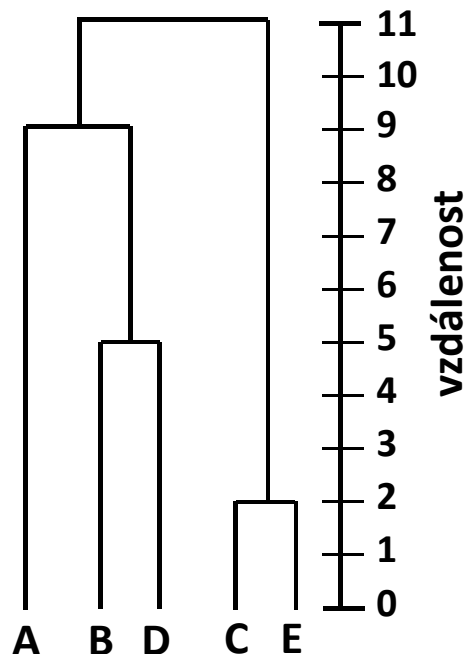


# Úkol č. 1



- 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.

## 4. krok



Všechny objekty jsou spojeny do jednoho shluku → již není co spojovat.

# 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