



EVROPSKÁ UNIE  
Evropské strukturální a investiční fondy  
Operační program Výzkum, vývoj a vzdělávání



MINISTERSTVO ŠKOLSTVÍ,  
MLÁDEŽE A TĚLOVÝCHOVY

Název projektu	Rozvoj vzdělávání na Slezské univerzitě v Opavě
Registrační číslo projektu	CZ.02.2.69/0.0./0.0/16_015/0002400

## Dolování dat

### Statistika v kontextu dolování dat

Jan Górecki



**SLEZSKÁ  
UNIVERZITA**  
OBCHODNĚ PODNIKATELSKÁ  
FAKULTA V KARVINĚ

# Obsah přednášky

---

- Typy statistických metod
  - Kontingenční tabulky
  - Regresní analýza
  - Diskriminační analýza
  - Shluková analýza
- 





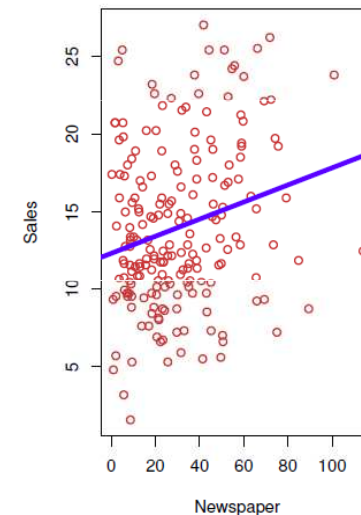
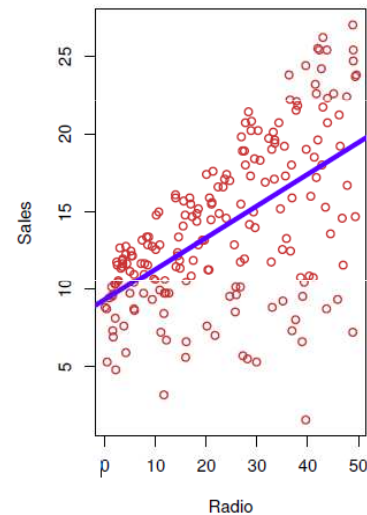
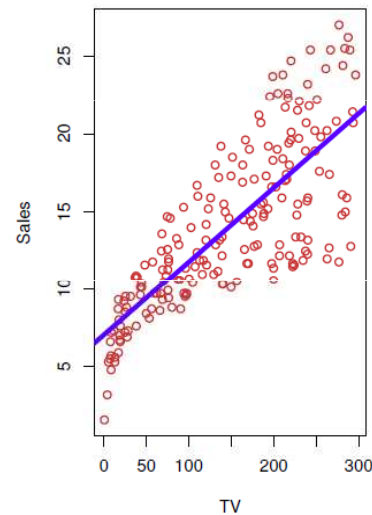
A formal science that deals with collection, analysis, interpretation, explanation and presentation of (usually numerical) data.

Metody:

- **Deskripční** – cílem je popsat základní charakteristiky daných dat
  - **Konfirmační** – cílem je potvrdit resp. vyvrátit zkoumanou hypotézu
  - **Explorační** – cílem je “objevit” možnou hypotézu, která je podporovaná daty
-

# Motivace

---



Sales – prodej produktu v tisících

TV, Radio, Newspaper - rozpočet na reklamu v daném médiu

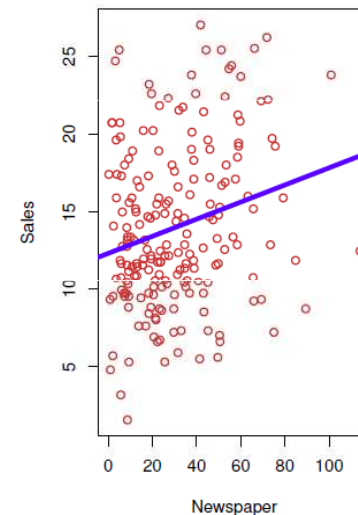
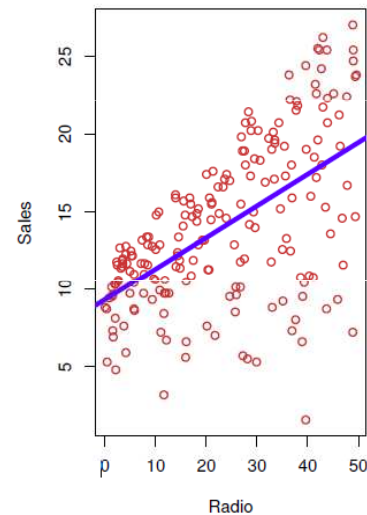
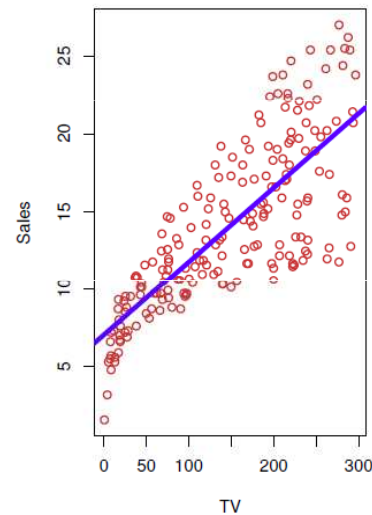
**Cíl:** Na základě těchto údajů navrhnete marketingový plán na příští rok, který povede k vysokým prodejům produktů.

**Otázka:** Jaké informace by byly užitečné pro poskytnutí takového doporučení?

---

# Motivace

---



1. Existuje vztah mezi rozpočtem na reklamu a prodejem?
  2. Jak silný je vztah mezi rozpočtem na reklamu a prodejem?
  3. Která média jsou spojena s prodejem?
  4. Jaký je vztah mezi každým médiem a prodejem?
  5. Jak přesně můžeme předpovídat budoucí prodeje?
  6. Je vztah lineární?
  7. Existuje synergie mezi reklamními médii?
-

# Lineární regrese



**SLEZSKÁ  
UNIVERZITA**  
OBCHODNĚ PODNIKATELSKÁ  
FAKULTA V KARVINĚ

## 1. Definice:

- Jednoduchá lineární regrese je statistická metoda, která zkoumá lineární vztah mezi dvěma kvantitativními proměnnými.

## 2. Matematický Model:

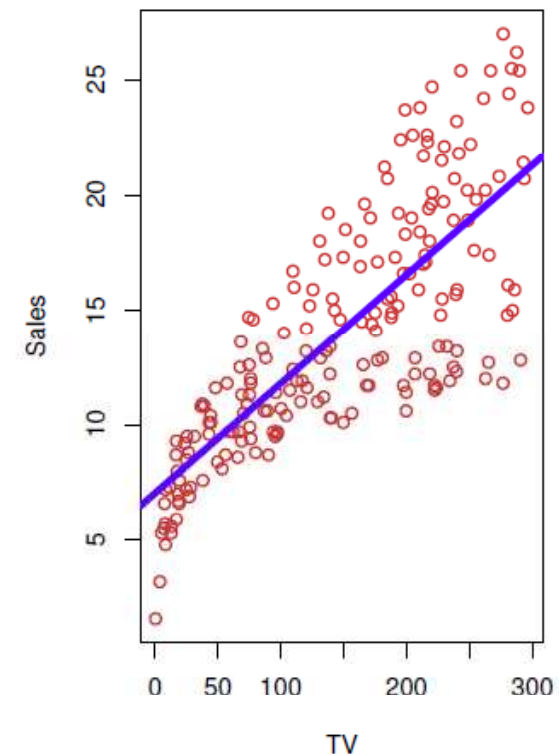
- $y = \beta_0 + \beta_1 x + \epsilon$
- kde:
  - $y$  je závislá proměnná
  - $x$  je nezávislá proměnná
  - $\beta_0$  je y-intercept (parameter modelu)
  - $\beta_1$  je sklon regresní čáry (parameter modelu)
  - $\epsilon$  je chyba

## 3. Příklad:

- Analýza vztahu mezi rozpočtem na reklamu ( $x$ ) a prodejem produktu ( $y$ ).

## 4. Význam:

- Pomáhá předpovídat hodnotu závislé proměnné na základě hodnoty nezávislé proměnné.
- Umožňuje odhadnout, jak se změna v hodnotě nezávislé proměnné ovlivní závislou proměnnou.



# Odhad parametrů modelu

## 1. Cíl:

- Odhad hodnot  $\beta_0$  a  $\beta_1$  tak, aby byla minimalizována chyba mezi skutečnými a předpovězenými hodnotami  $y$ .

## 2. Metoda Nejmenších Čtverců:

- Minimalizuje součet čtverců reziduí  $RSS = e_1^2 + e_2^2 + \dots + e_n^2$ , kde  $e_i = y_i - (\beta_0 + \beta_1 x_i)$  je rozdíl mezi skutečnou ( $y_i$ ) a předpovězenou hodnotou ( $\beta_0 + \beta_1 x_i$ ) pro  $i$ -tý bod. (RSS = Residual sum of squares)

- Výpočet:

- $$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

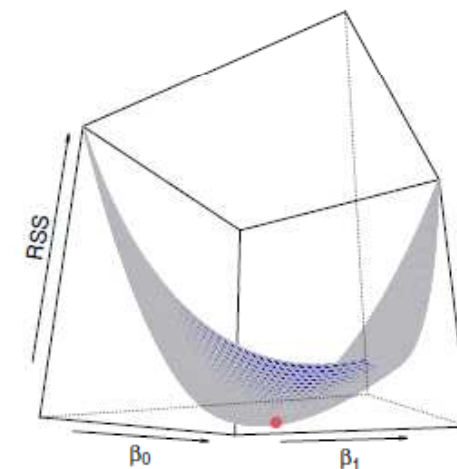
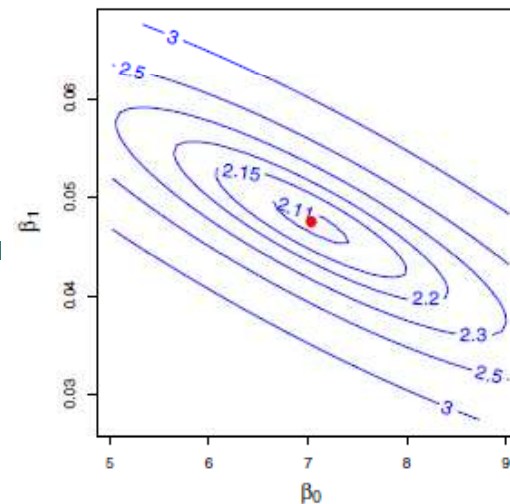
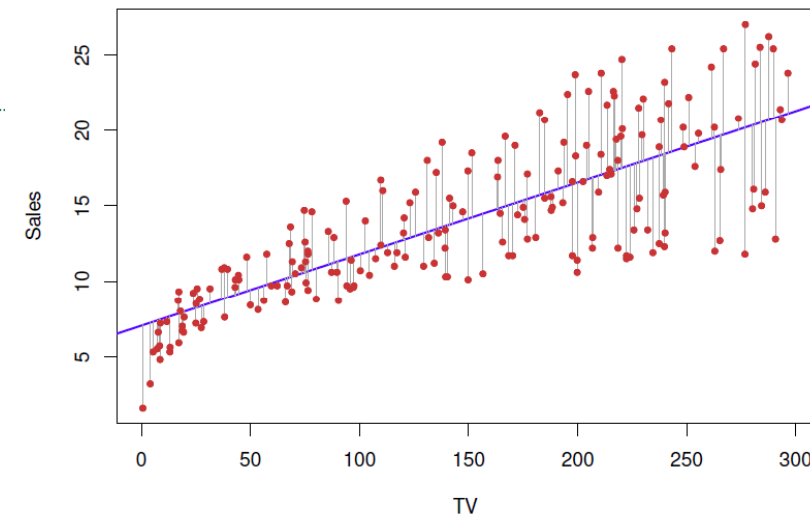
- $$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

## 3. Příklad s Daty:

- Použití dat o prodeji a reklamním rozpočtu k odhadu parametrů modelu lineární regrese.

## 4. Interpretace:

- Jak interpretovat odhadnuté parametry a jaký mají význam pro předpovězení prodeje na základě rozpočtu na reklamu.



# Odpovědi na 7 otázek



**SLEZSKÁ  
UNIVERZITA**  
OBCHODNĚ PODNIKATELSKÁ  
FAKULTA V KARVINĚ

## 1. Existuje vztah mezi rozpočtem na reklamu a prodejem?

- Analýza korelace a regresní analýza k identifikaci vztahu.

## 2. Jak silný je vztah mezi rozpočtem na reklamu a prodejem?

- Hodnocení koeficientu determinace ( $R^2$ ) k měření síly vztahu ( $R^2$  odpovídá kvadrátu korelace X a Y).

## 3. Která média jsou spojena s prodejem?

- Výsledky analýzy pro TV, Radio a Newspaper.

## 4. Jaký je vztah mezi každým médiem a prodejem?

- Hodnoty koeficientů regrese pro každé médium.

## 5. Jak přesně můžeme předpovídat budoucí prodeje?

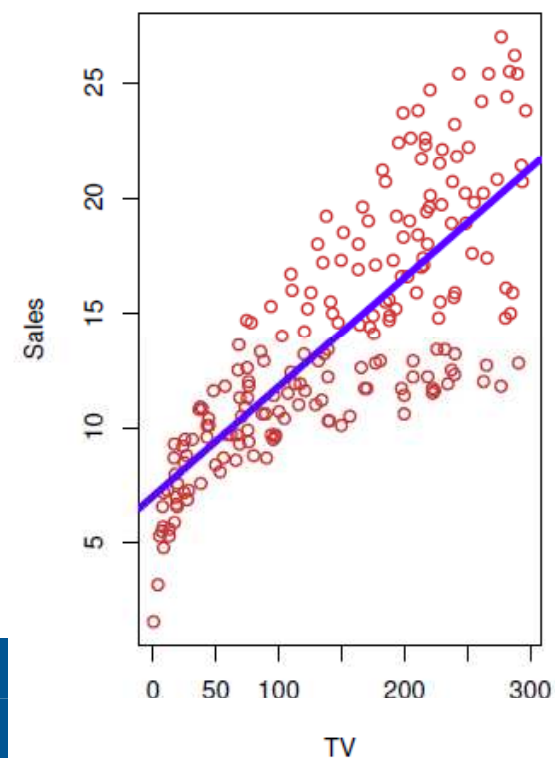
- Model predikce s odhadnutými parametry a intervaly spolehlivosti.

## 6. Je vztah lineární?

- Grafická a statistická analýza pro ověření linearitu vztahu.

## 7. Existuje synergie mezi reklamními médii?

- Vícerozměrná analýza interakcí mezi různými médii.





# Kontingenční tabulky



- zjišťování vztahu mezi dvěma **kategoriálními** veličinami

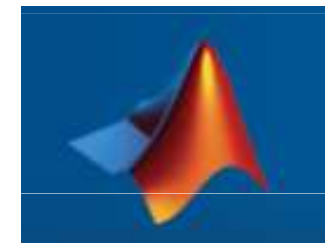
čtyřpolní tabulka

	Úvěr ano	Úvěr ne	$\Sigma$
Vysoký příjem	$a_{11}$	$a_{12}$	$r_1$
Nízký příjem	$a_{21}$	$a_{22}$	$r_2$
$\Sigma$	$s_1$	$s_2$	$n$

$$\chi^2 = \sum_{i=1}^R \sum_{j=1}^S \frac{(a_{ij} - o_{ij})^2}{o_{ij}} = n \times \sum_{i=1}^R \sum_{j=1}^S \frac{\left( a_{ij} - \frac{r_i \cdot s_j}{n} \right)^2}{r_i \cdot s_j}$$

pro  $\chi^2 \geq \chi^2_{(R-1)(S-1)}(\alpha)$  předpokládáme  
závislost mezi X a Y

příjem	úvěr
vysoký	ano
vysoký	ano
nízký	ne
nízký	ano
nízký	ano
nízký	ne
vysoký	ano
vysoký	ano
nízký	ne
vysoký	ano
nízký	ne
nízký	ano



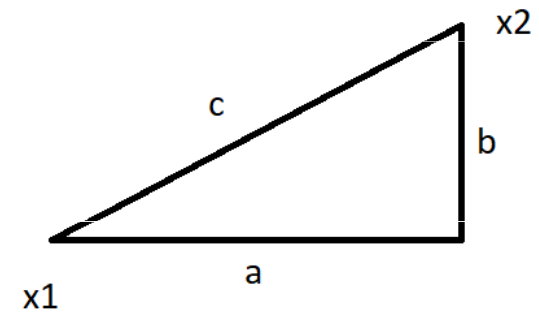
$o_{ij}$  ... očekávané množství při  
platnosti hypotézy o nezávislosti  
veličin

# Shluková analýza

- slouží pro nalezení skupin (shluků) navzájem si podobných příkladů
- Např. dva příklady  $\mathbf{x}_1 = [x_{11}, \dots, x_{1m}]$  a  $\mathbf{x}_2 = [x_{21}, \dots, x_{2m}]$

Eukleidovská vzdálenost

$$d_E(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^m (x_{1j} - x_{2j})^2}$$



$$d_E(\mathbf{x}_1, \mathbf{x}_2) = c$$

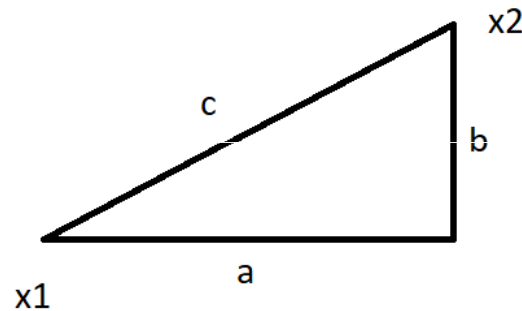
Jaká je tato vzdálenost pro body  $x_1 = [1, 2]$  a  $x_2 = [4, 6]$ ?

# Minkovského (Manhattan) vzdálenost



SLEZSKÁ  
UNIVERZITA  
OBCHODNĚ PODNIKATELSKÁ  
FAKULTA V KARVINĚ

$$d_H(\mathbf{x}_1, \mathbf{x}_2) = \sum_{j=1}^m |x_{1j} - x_{2j}|$$



$$d_E(\mathbf{x}_1, \mathbf{x}_2) = a + b$$

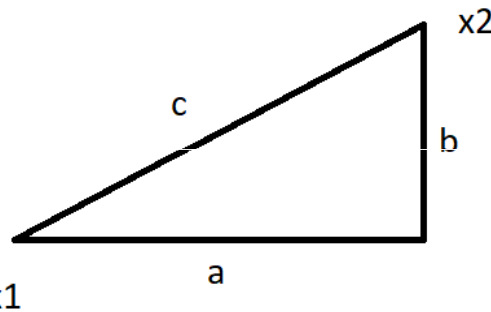
Jaká je tato vzdálenost pro body  $x_1 = [1, 2]$  a  $x_2 = [4, 6]$ ?

# Čebyševova vzdálenost



SLEZSKÁ  
UNIVERZITA  
OBCHODNĚ PODNIKATELSKÁ  
FAKULTA V KARVINĚ

$$d_C(\mathbf{x}_1, \mathbf{x}_2) = \max_j |x_{1j} - x_{2j}|$$



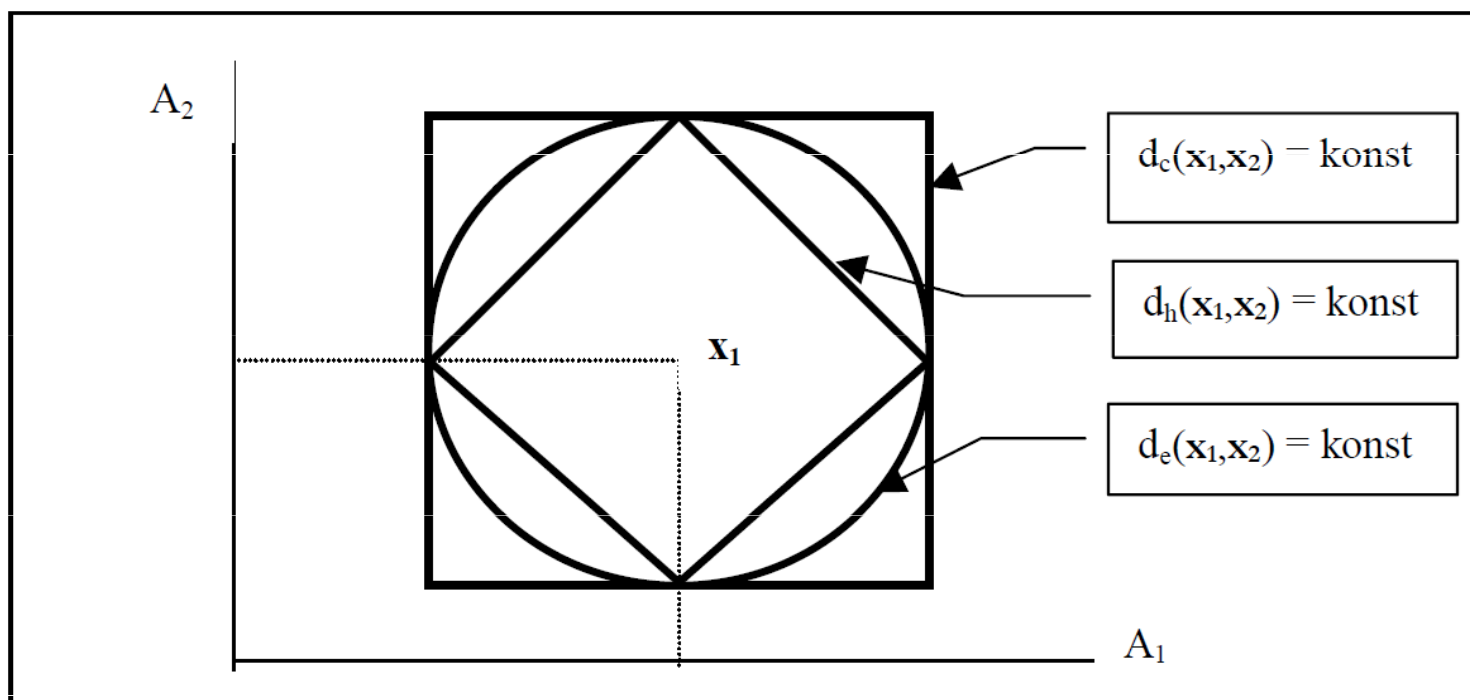
$$d_E(\mathbf{x}_1, \mathbf{x}_2) = \max(a, b) = a$$

Jaká je tato vzdálenost pro body  $\mathbf{x}_1 = [1, 2]$  a  $\mathbf{x}_2 = [4, 6]$ ?

## Rozdíl mezi $d_M(\mathbf{x}_1, \mathbf{x}_2)$ , $d_E(\mathbf{x}_1, \mathbf{x}_2)$ a $d_C(\mathbf{x}_1, \mathbf{x}_2)$ ve 2D



SLEZSKÁ  
UNIVERZITA  
OBCHODNĚ PODNIKATELSKÁ  
FAKULTA V KARVINĚ



Pozor: pro 1D všechny vzdálenosti splývají (dávají stejný výsledek)



- Výše uvedené míry vzdálenosti závisí na měřítku veličin. Proto je třeba veličiny normovat (normalizovat)
  - Konkrétní hodnota se obvykle dělí nějakou jinou hodnotou:
    - směrodatnou odchylkou
    - rozpětím (max-min).
-



- hierarchické shlukování,
  - metoda *K*-středů (*K*-means clustering).
-

# Hierarchické shlukování



Při hierarchickém shlukování se obvykle postupuje metodou „zdola nahoru“. Začíná se tedy v situaci, kdy každý příklad tvoří jeden samostatný shluk. Postupně se pak jednotlivé shluky spojují, až skončíme s jedním shlukem obsahujícím všechny příklady

## Algoritmus hierarchického shlukování

### Inicializace

1. urči vzájemné vzdálenosti mezi všemi příklady
2. zařaď každý příklad do samostatného shluku

### Hlavní cyklus

1. dokud je více než jeden shluk
  - 1.1. najdi dva navzájem nejbližší shluky a spoj je
  - 1.2. spočítej pro tento nový shluk jeho vzdálenost od ostatních shluků



# Vzdálenost mezi shluky

---



- *metoda nejbližšího souseda* - vzdálenost mezi shluky  $U$  a  $V$  je dána minimem ze vzdálenosti mezi jejich příklady

$$D(U, V) = \min_{k,l} d(\mathbf{x}_k, \mathbf{x}_l), \mathbf{x}_k \in U, \mathbf{x}_l \in V$$

- *metoda nejvzdálenějšího souseda* - vzdálenost mezi shluky  $U$  a  $V$  je dána maximem ze vzdálenosti mezi jejich příklady

$$D(U, V) = \max_{k,l} d(\mathbf{x}_k, \mathbf{x}_l), \mathbf{x}_k \in U, \mathbf{x}_l \in V$$

- *metoda průměrné vzdálenosti* - vzdálenost mezi shluky  $U$  a  $V$  je dána průměrem ze vzdálenosti mezi jejich příklady ( $n_U$  je počet příkladů ve shluku  $U$  a  $n_V$  je počet příkladů ve shluku  $V$ )

$$D(U, V) = \frac{1}{n_U n_V} \sum_{k=1}^{n_U} \sum_{l=1}^{n_V} d(\mathbf{x}_k, \mathbf{x}_l)$$

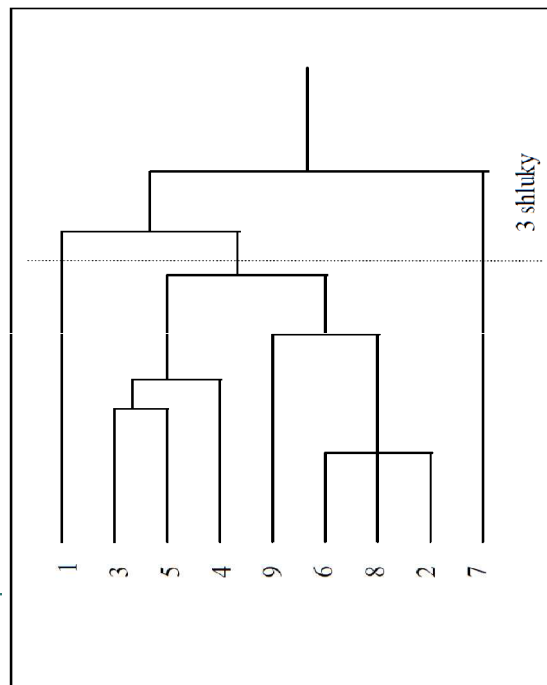
---

# Dendrogram



SLEZSKÁ  
UNIVERZITA  
OBCHODNĚ PODNIKATELSKÁ  
FAKULTA V KARVINĚ

- Proces hierarchického shlukování bývá zachycen v podobě tzv. dendrogramu. Ten ukazuje (odspoda nahoru) postupné spojování shluků počínaje očíslovanými příklady. Optimální počet shluků zde není předem znám, odvodíme ho až rozbořením výsledků – tak, že někde dendrogram „rozřízneme“



# Příklad

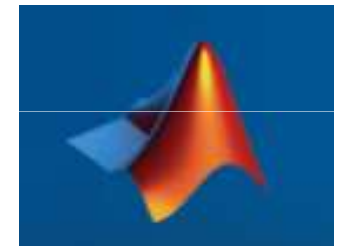
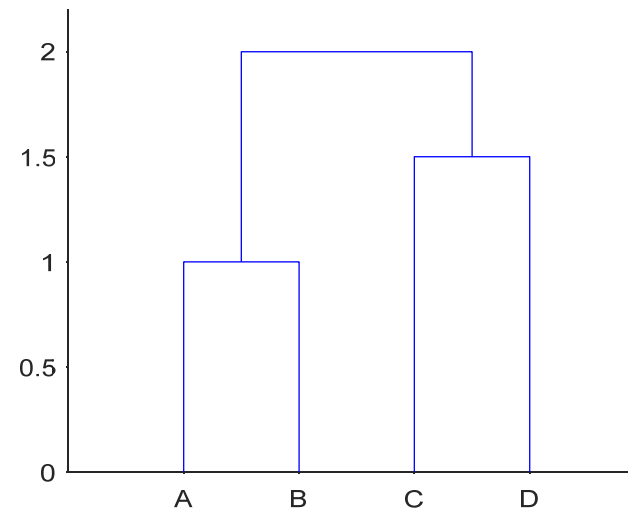


Jak proběhne hierarchické shlukování pro 4 jednorozměrné body

$A = [0]$ ,  $B = [1]$ ,  $C = [3]$  a  $D = [4,5]$

pro *eukleidovskou* vzdálenost a metodu *nejbližšího* souseda?

	A	B	C	D
A				
B				
C				
D				



# Metoda $K$ –středů - Algoritmus

---



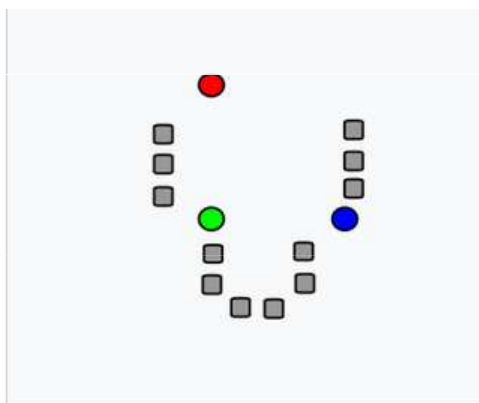
SLEZSKÁ  
UNIVERZITA  
OBCHODNĚ PODNIKATELSKÁ  
FAKULTA V KARVINĚ

1. urči centroidy pro všechny shluky v aktuálním rozkladu (v prvním opakování zcela náhodně)
  2. pro každý příklad  $\mathbf{x}$ 
    - 2.1. urči vzdálenosti  $d(\mathbf{x}, \mathbf{c}_k)$ ,  $k=1, \dots, K$  kde  $\mathbf{c}_k$  je centroid  $k$ -tého shluku
    - 2.2. urči centroid  $\mathbf{c}_l$  tak, že  $d(\mathbf{x}, \mathbf{c}_l) = \min_k d(\mathbf{x}, \mathbf{c}_k)$
    - 2.3. není-li  $\mathbf{x}$  součástí shluku  $l$  (k jehož centroidu  $\mathbf{c}_l$  má nejbližší) přesuň  $\mathbf{x}$  do shluku  $l$
  3. došlo-li k nějakému přesunu potom jdi na 1, jinak konec
-

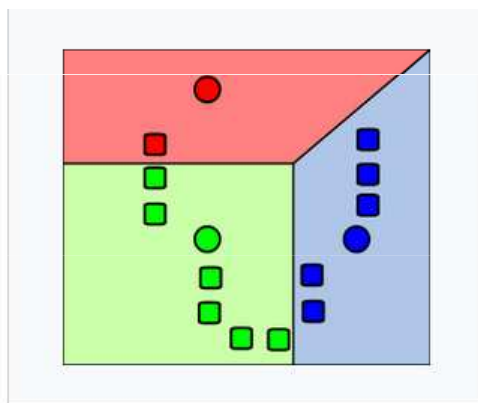
# Ukázka algoritmu K-středů



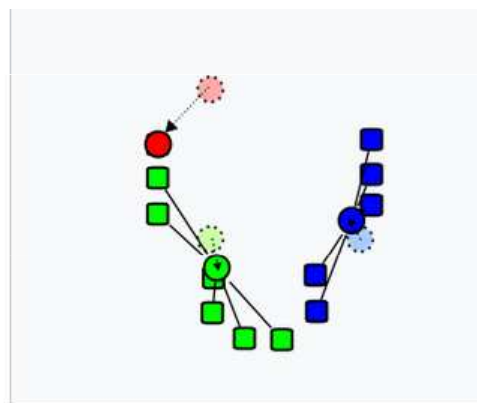
SLEZSKÁ  
UNIVERZITA  
OBCHODNĚ PODNIKATELSKÁ  
FAKULTA V KARVINĚ



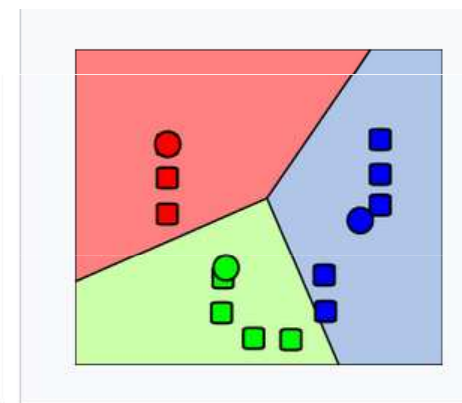
1.  $k$  výchozích centroidů (zde je  $k=3$ ) se náhodně umístí v prostoru dat (shlukované objekty šedé, centroidy barevné)



2. Objekty se přiřadí nejbližším centroidům, čímž vznikne  $k$  shluků. Centroidy tak definují [Voroného teselaci](#) prostoru.



3. Přepočtou se centroidy shluků tak, aby šlo o těžiště objektů, jež patří do těchto shluků.



4. Kroky 2 a 3 se opakují, dokud nedojde k ustálení ([konvergence](#)).

# Děkuji za pozornost

Některé snímky převzaty od:  
prof. Ing. Petr Berka, CSc. [berka@vse.cz](mailto:berka@vse.cz)