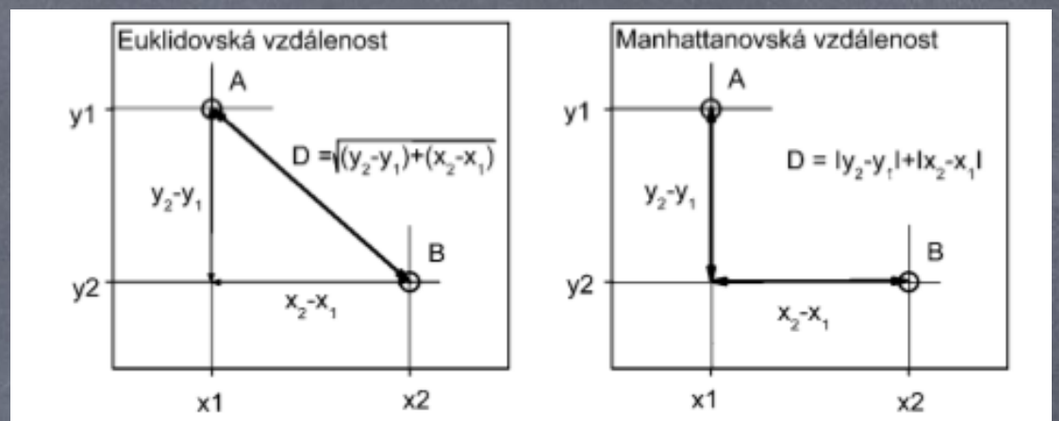
Shluková analýza

Klasifikace vícerozměrných dat

- vektory dat
- různé formy dat
- míra podobnosti
- mechanismus shlukování

Intervalová data



Obr. 1 Nejpoužívanější míry vzdálenosti: (a) Eukleidovská *D*, (b) Manhattanovská vzdálenost *D*

- normalizace
  - z-score (odečteme průměr a podělíme směrodatnou odchylkou)
  - transformace na <0..1>
- Minkovskeho vzdálenost (eukleidovská, hammingova)

$$d_M(x_k, x_l) = \sqrt[z]{\sum_{j=1}^{m} |x_{kj} - x_{lj}|^z}$$

Nominální a Ordinální data

- Nominální metrika

$$d(i,j) = \frac{p - m}{p}$$

p – počet proměnných
m – počet shod

- Ordinální data převedema na <0..1> a zpracováváme jako intervalová

Metody

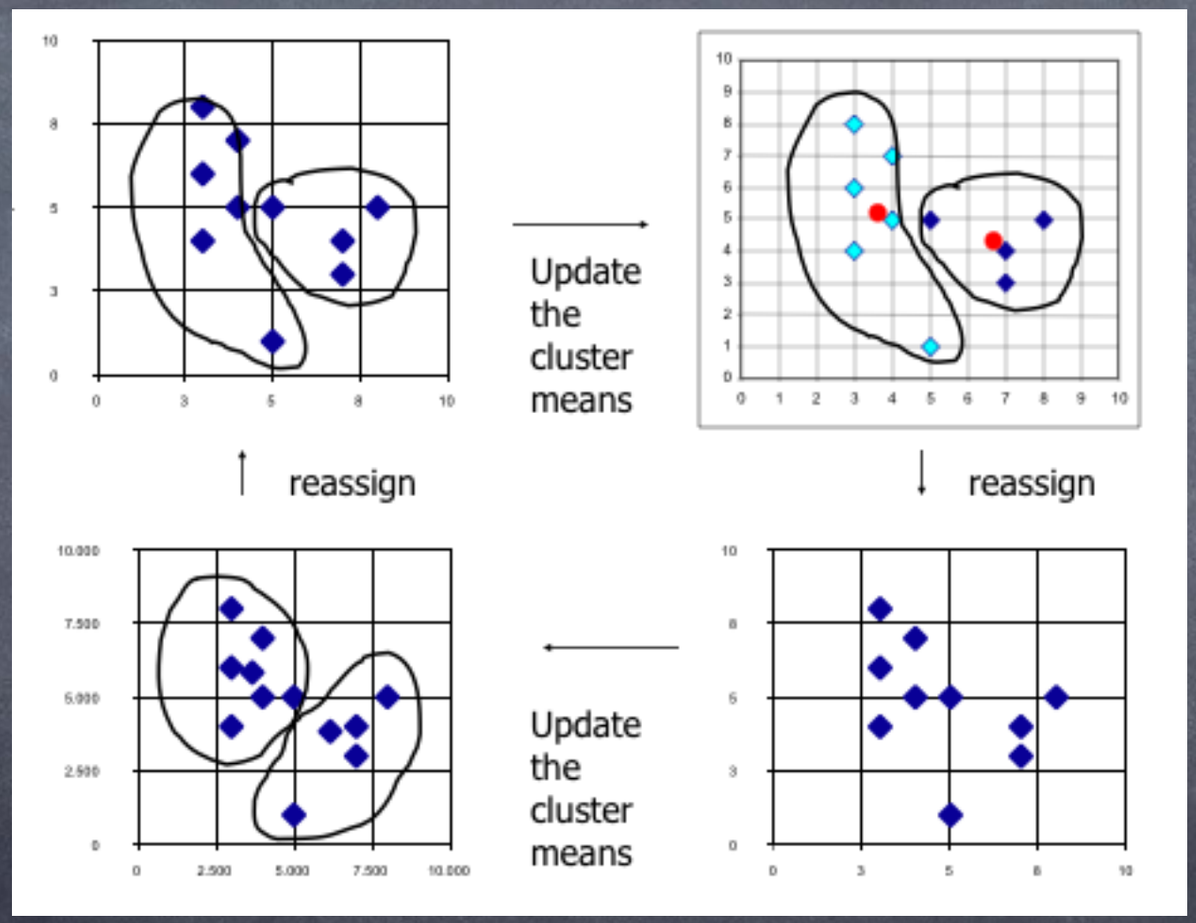
- Distanční matice
- Dělící
- Hierarchické
- Hustotní

Distanční matice

	A	B	C	D	E
A	0	1	1	1	2
B	1	0	1	2	2
C	1	1	0	1	1
D	1	2	1	0	2
E	2	2	1	2	0

- zkonstruujeme matici
- přehazujeme řádky (a sloupce)
  - male honoty k diagonale

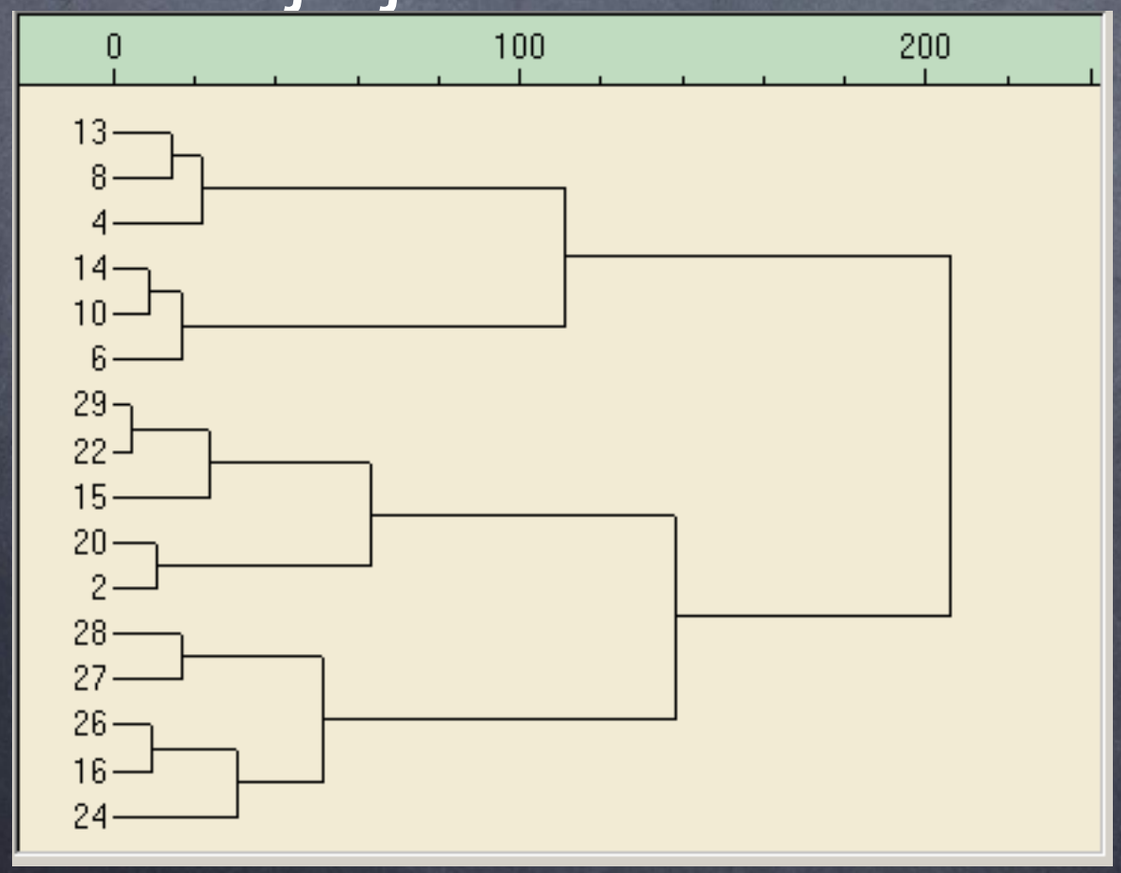
K-means

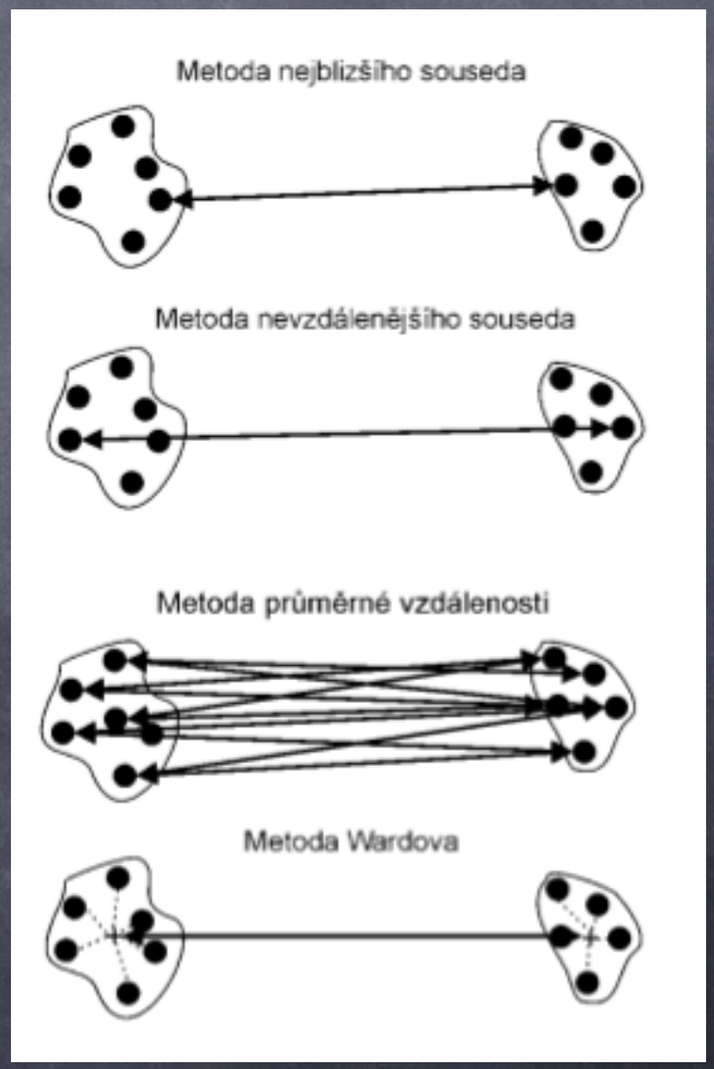


[Clustering - K-means demo]

Aglomerace

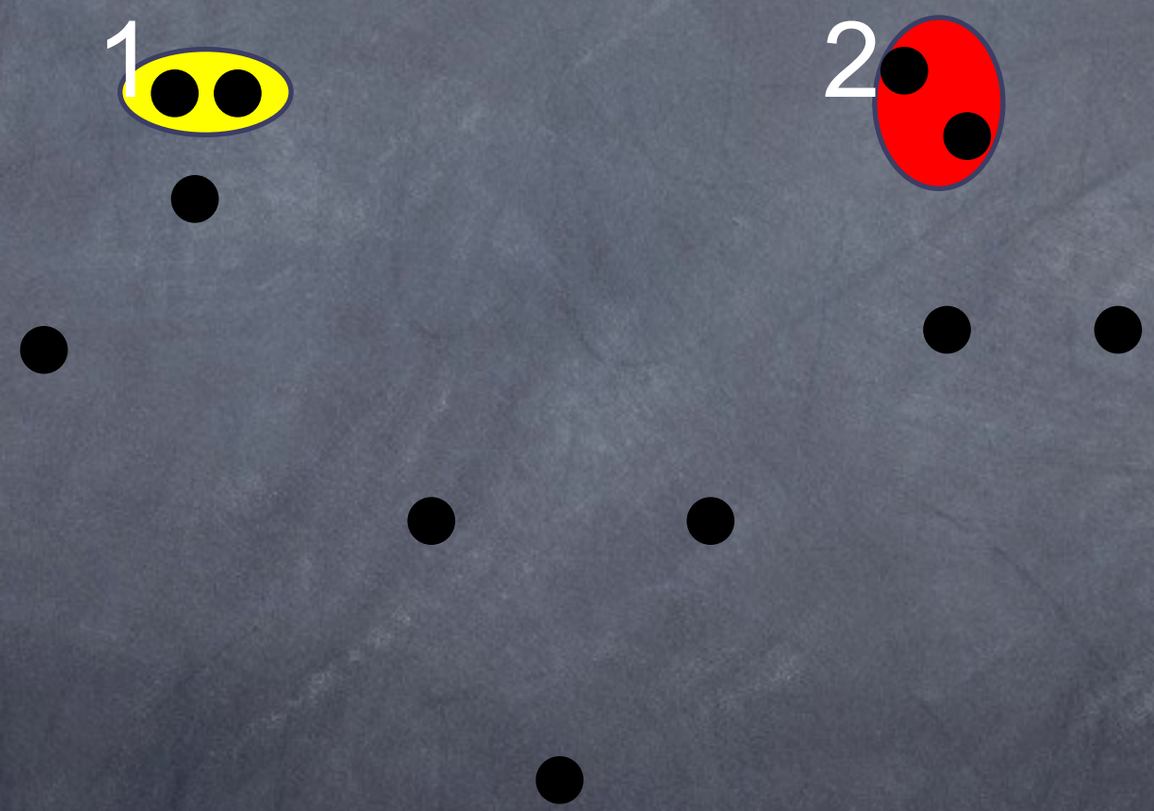
- postupně shlukujeme objekty na základě vzdálenosti
- na konci je jeden shluk



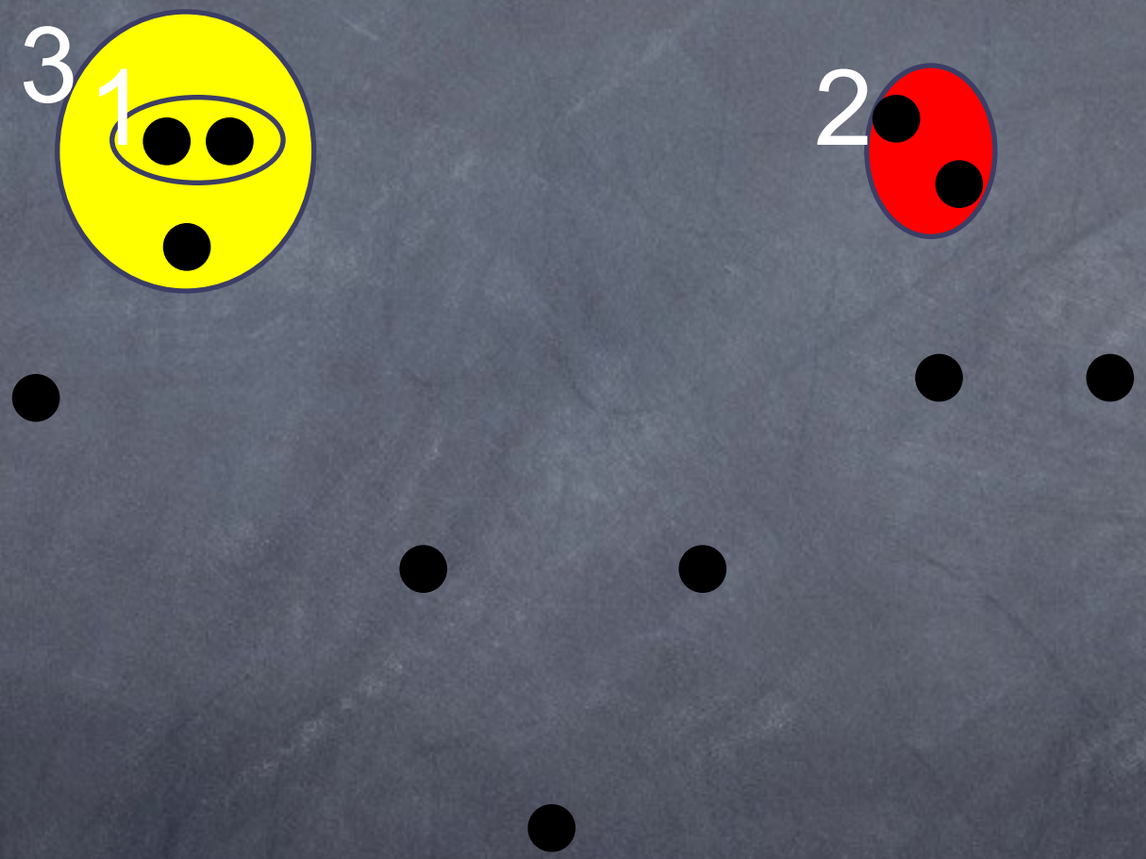


Hierarchical clustering

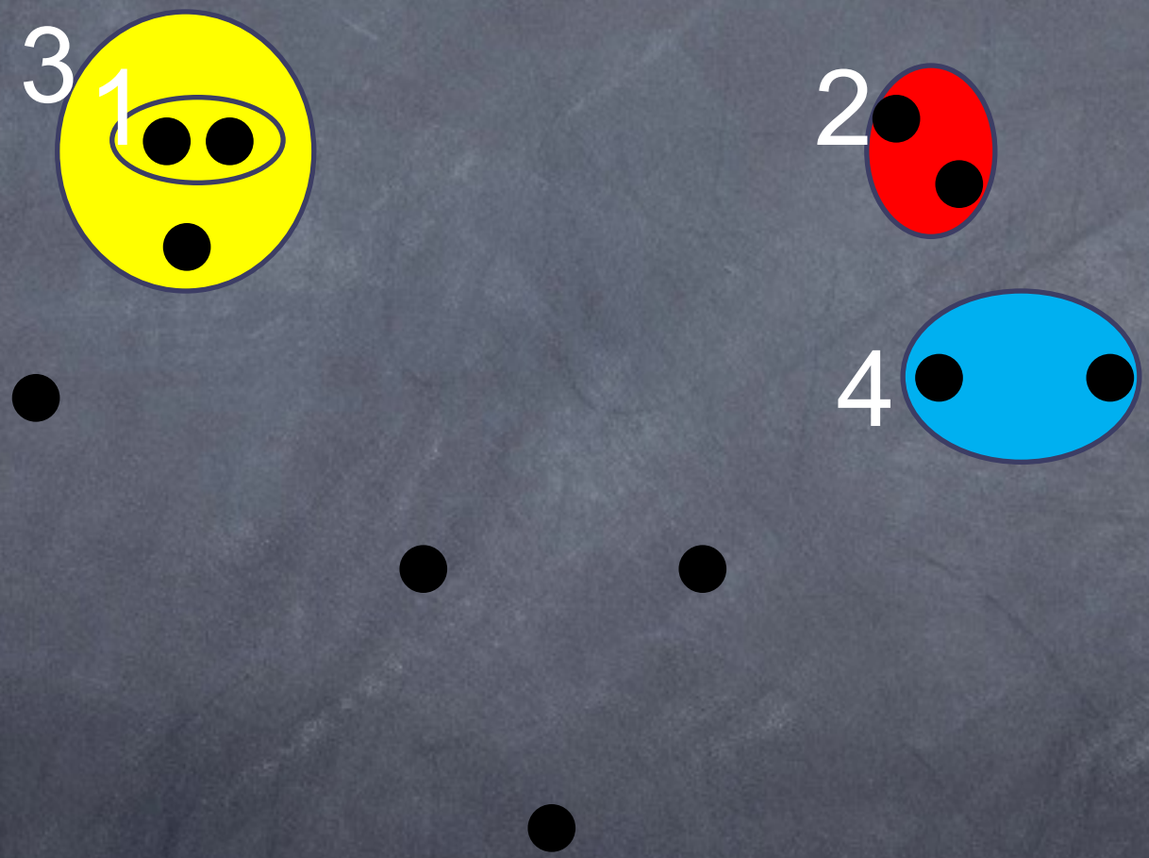
Hierarchical clustering



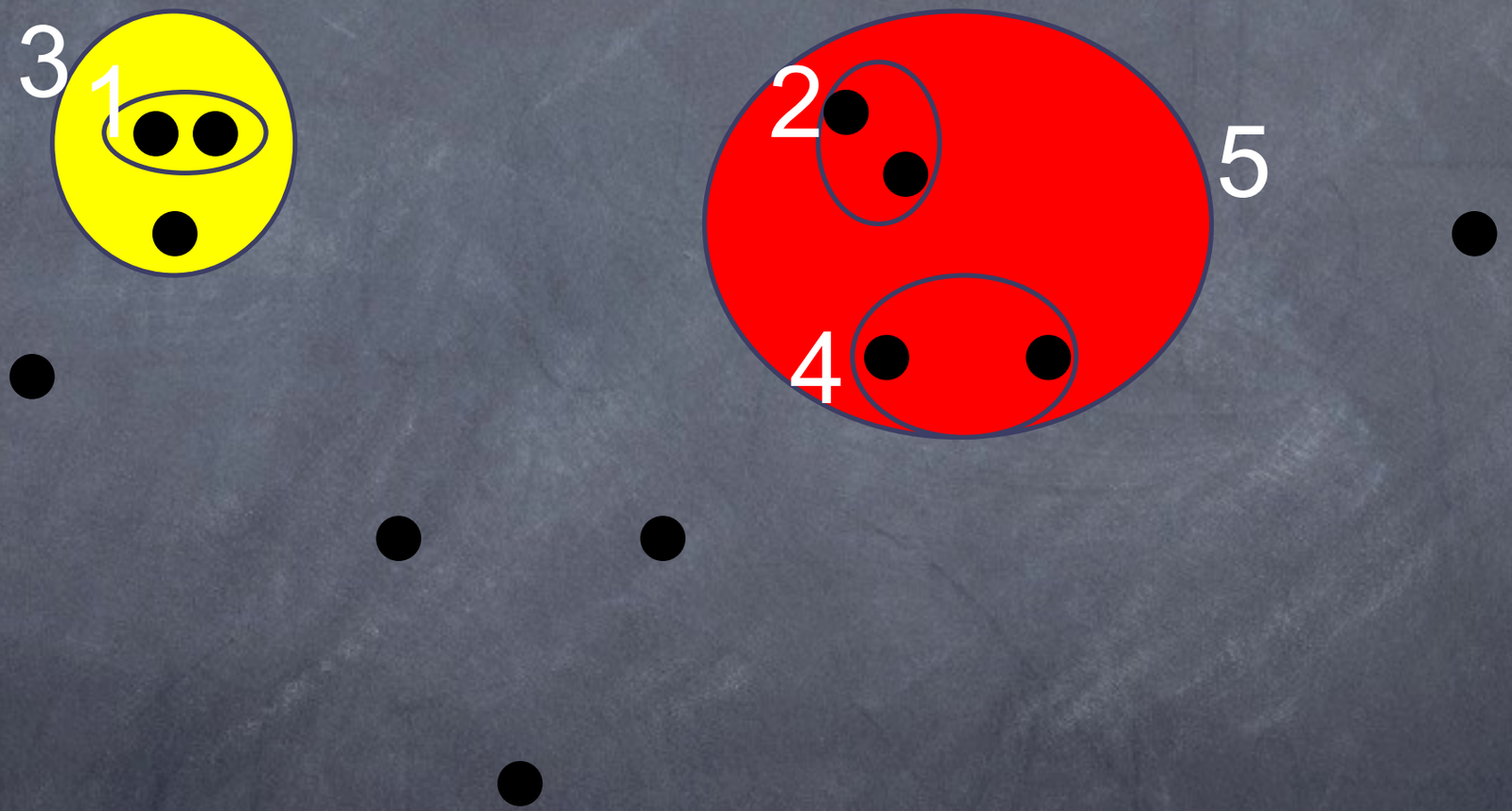
Hierarchical clustering



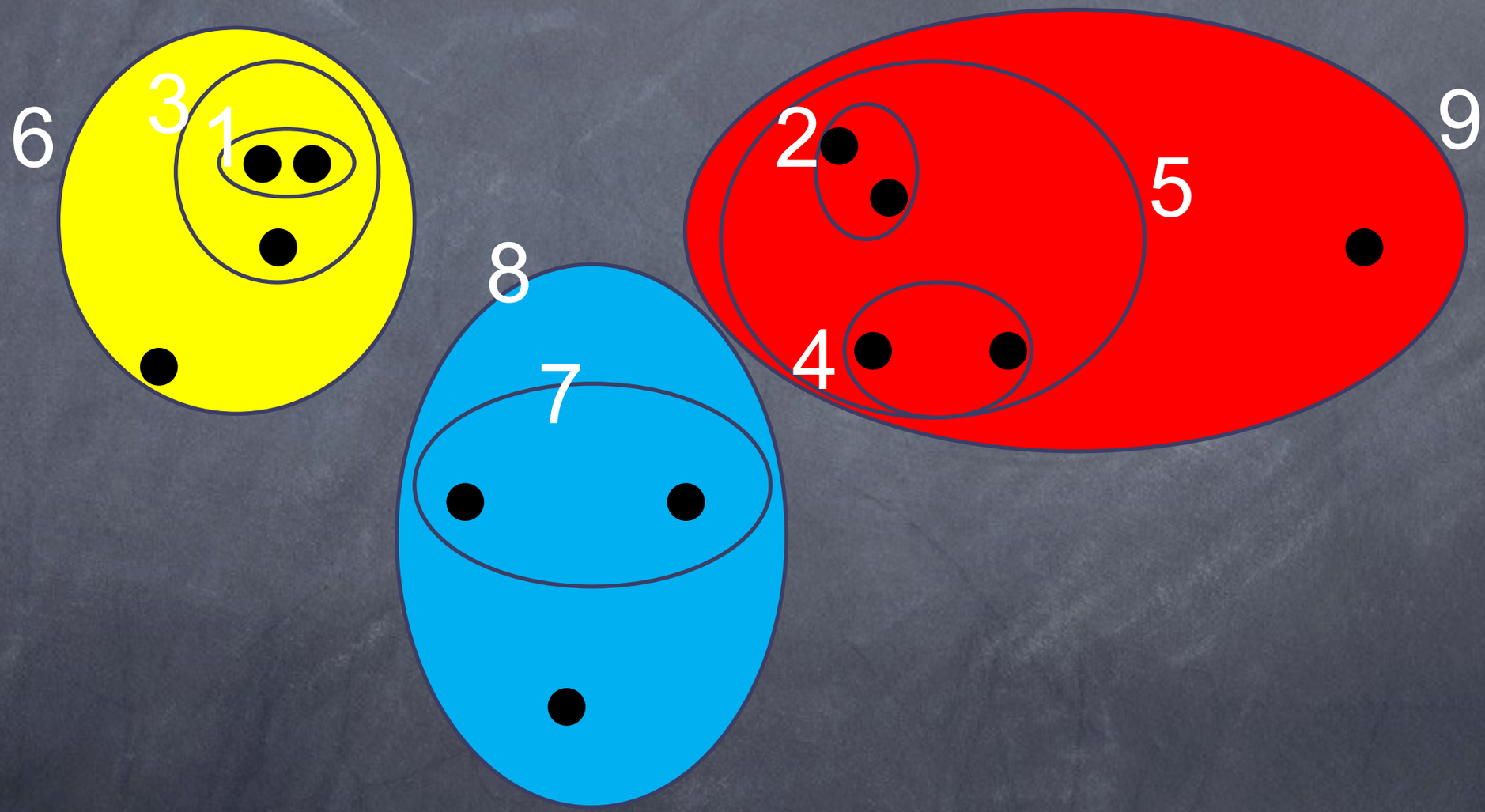
Hierarchical clustering



Hierarchical clustering



Hierarchical clustering



Hustotní

- radius
- minimální počet objektů
- jádro
- dosažitelnost
- šum