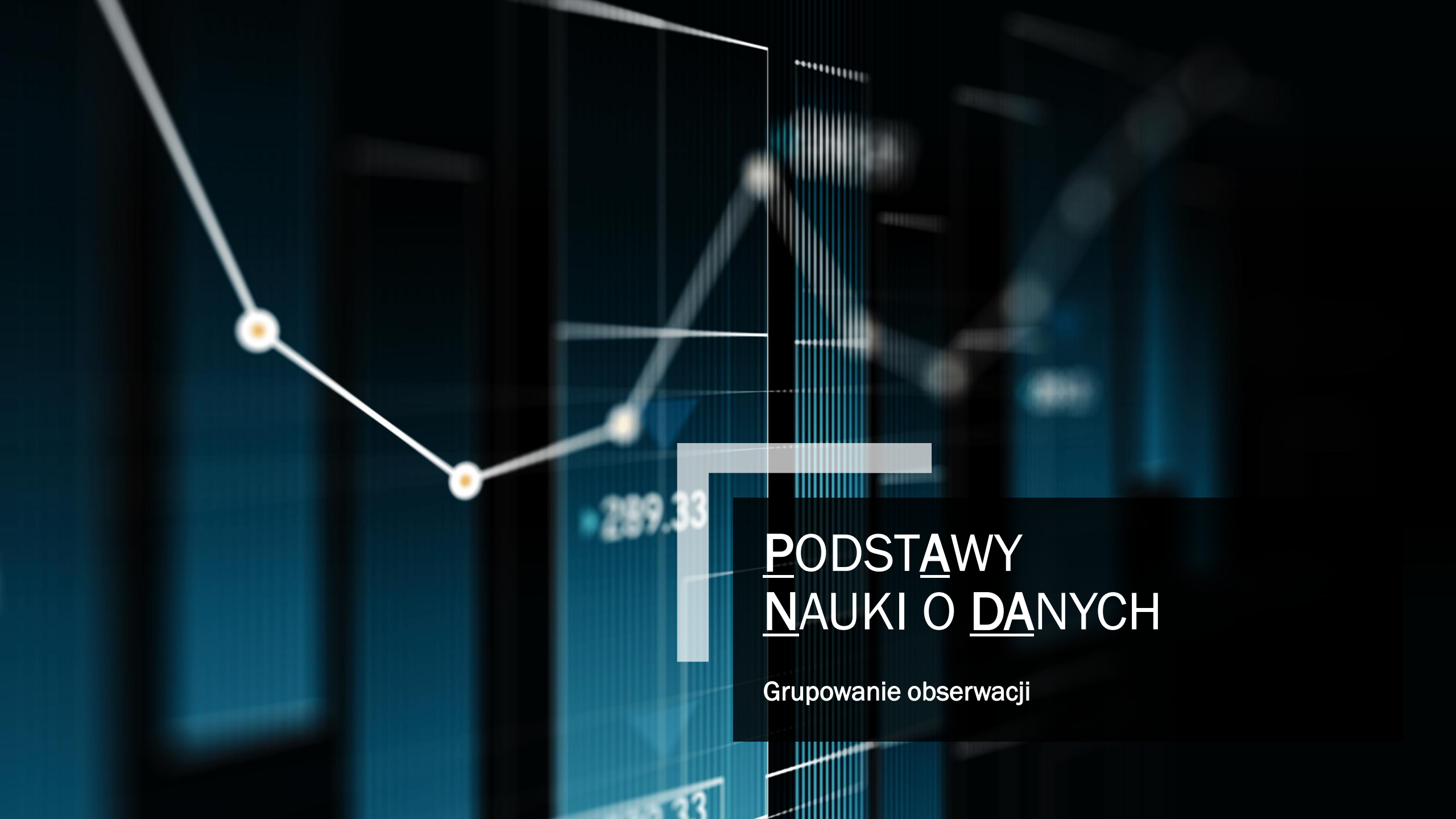




PODSTAWY NAUKI O DANYCH

Grupowanie obserwacji

Grupowanie

Dane
pomiarowe



$$\{\mathbf{x}_i\}_{i=1}^N$$

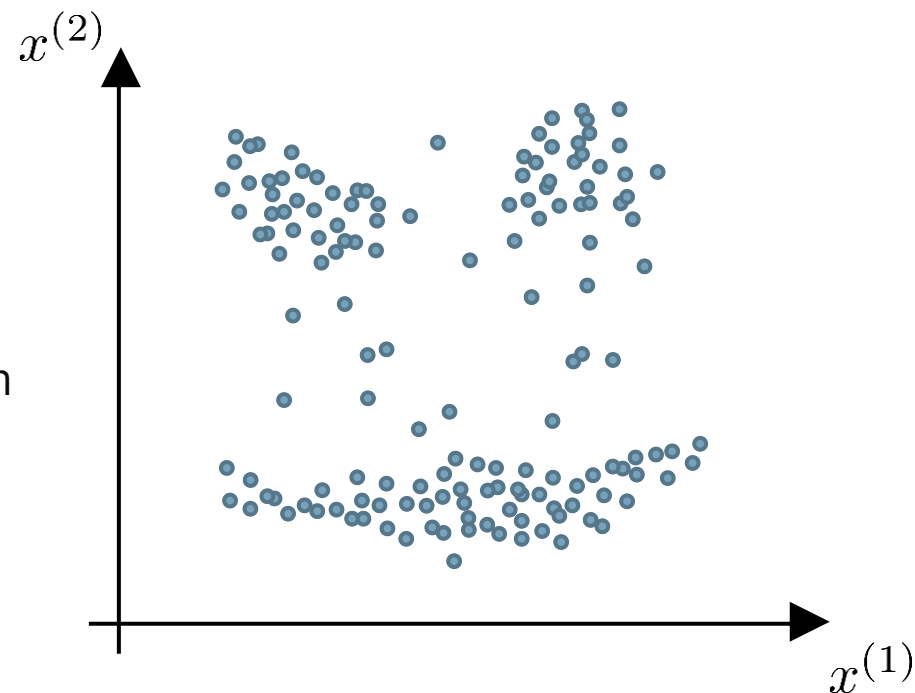
Zbiór
wektorów

$$\{\mathbf{x}_i\}_{i=1}^N$$

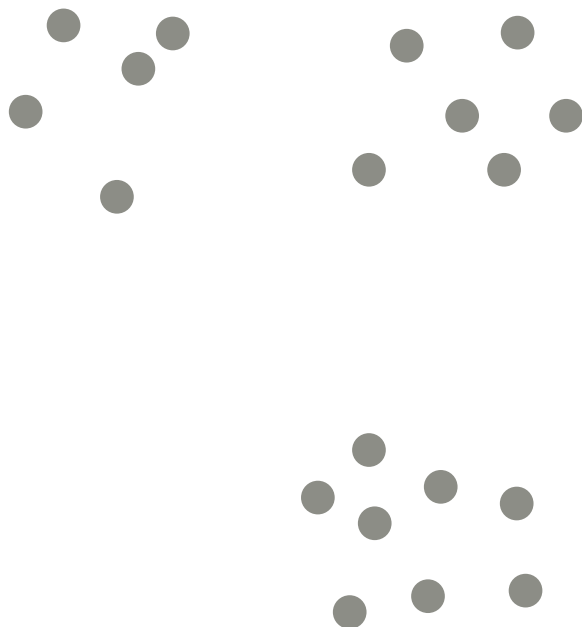
Algorytm
grupowania

Zbiór numerów (etykiet)
przypisanych skupieniom

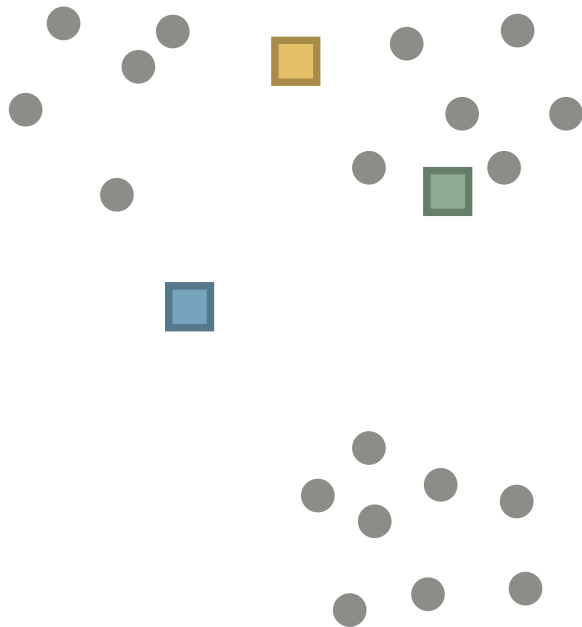
$$\{\hat{C}_i\}_{i=1}^N$$



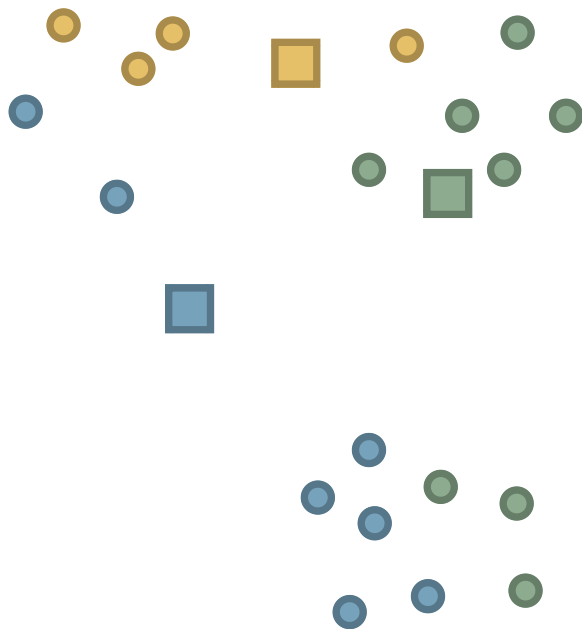
Algorytm k-średnich (ang. *K*-Means)



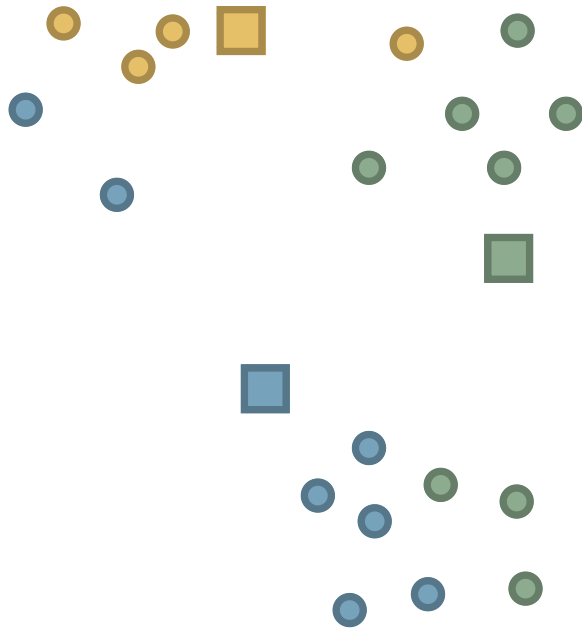
Algorytm k-średnich (ang. *K*-Means)



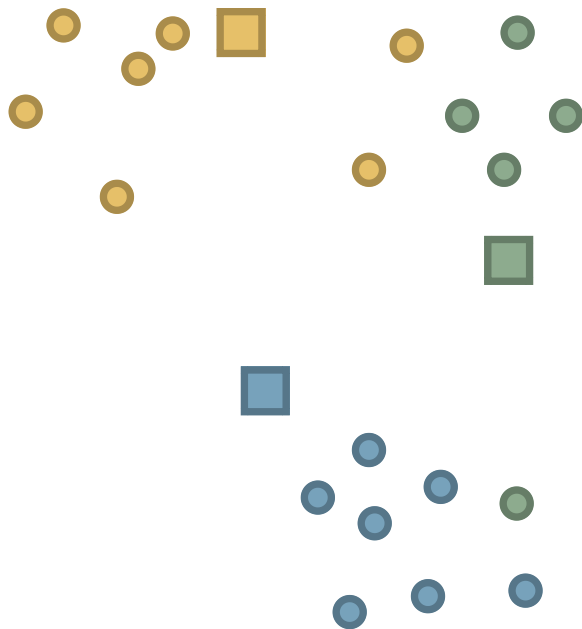
Algorytm k-średnich (ang. *K*-Means)



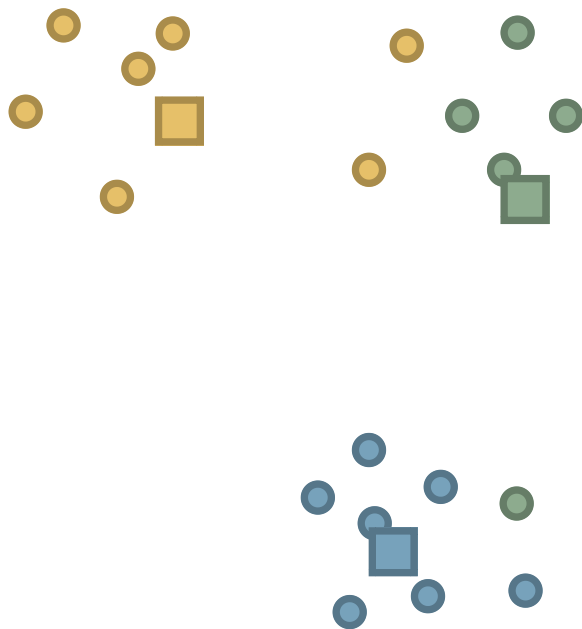
Algorytm k-średnich (ang. *K*-Means)



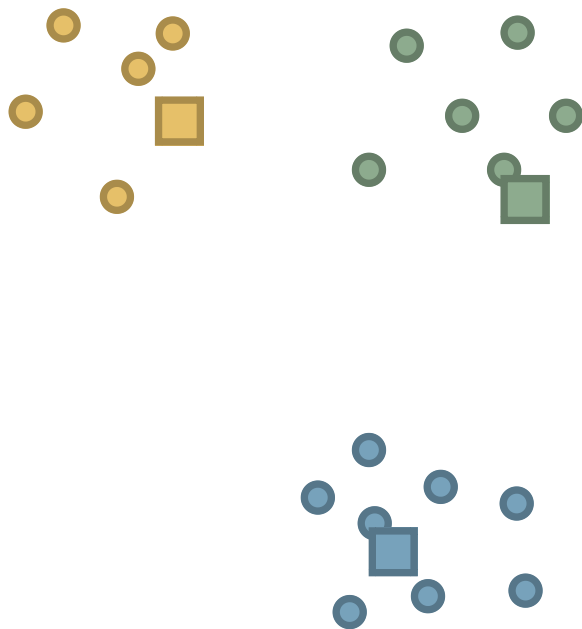
Algorytm k-średnich (ang. *K*-Means)



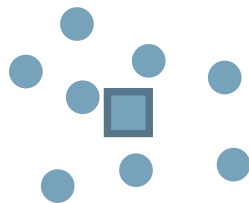
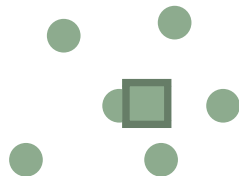
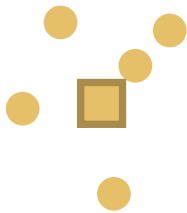
Algorytm k-średnich (ang. *K*-Means)



Algorytm k-średnich (ang. *K*-Means)



Algorytm k-średnich (ang. *K-Means*)



Ogólna procedura:

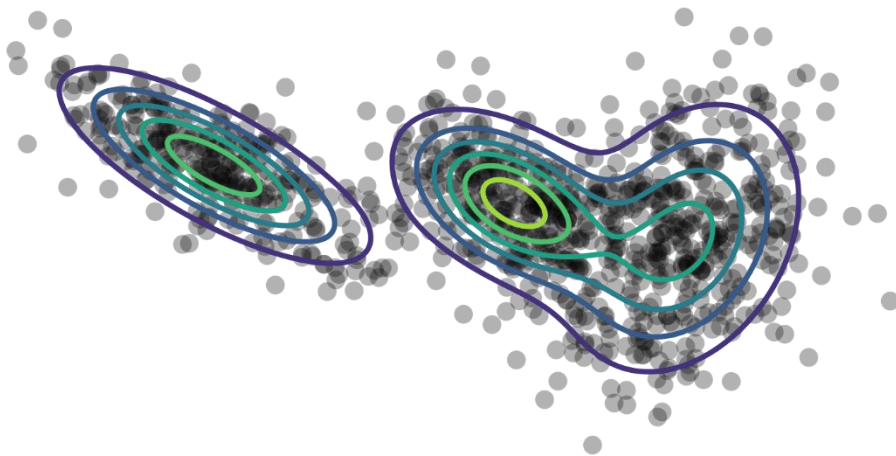
Metoda maksymalizacji oczekiwania

(ang. *Expectation – Maximization E-M*)

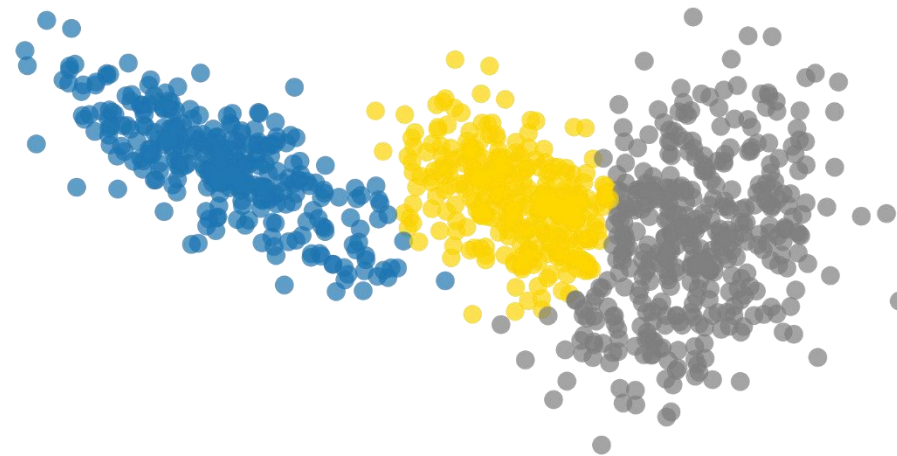
Mieszana rozkładów Gaussa

$$p(\mathbf{x}|\theta) = \sum_{m=1}^M \pi_m \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$$

$$0 \leq \pi_m \leq 1, \quad \sum_{m=1}^M \pi_m = 1$$

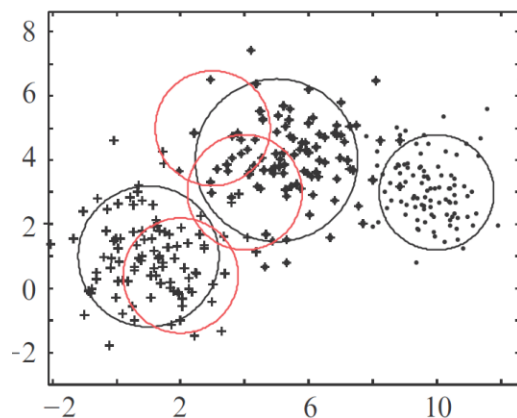


jako estymator rozkładu

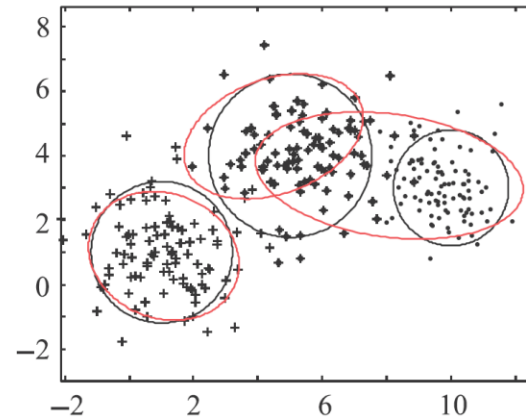


jako algorytm grupowania

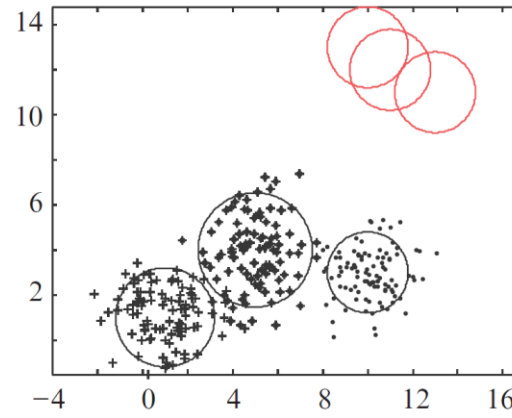
Mieszanina rozkładów Gaussa



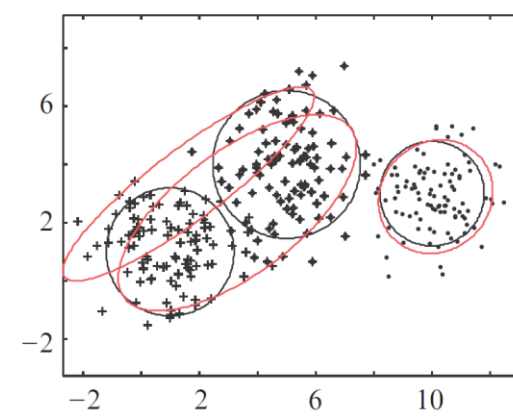
(a)



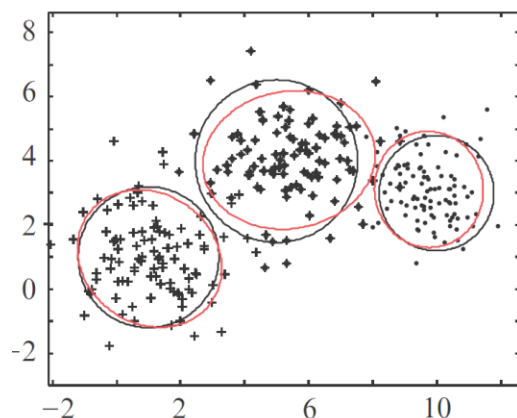
(b)



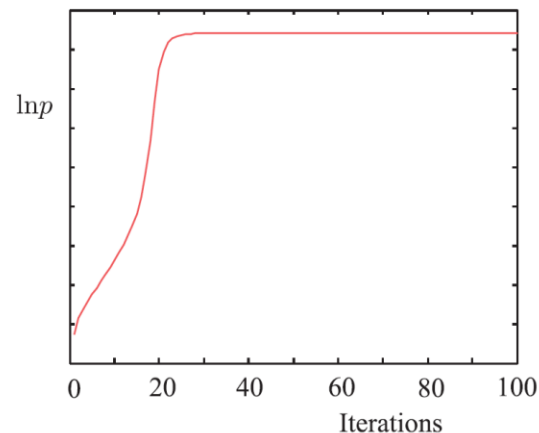
(a)



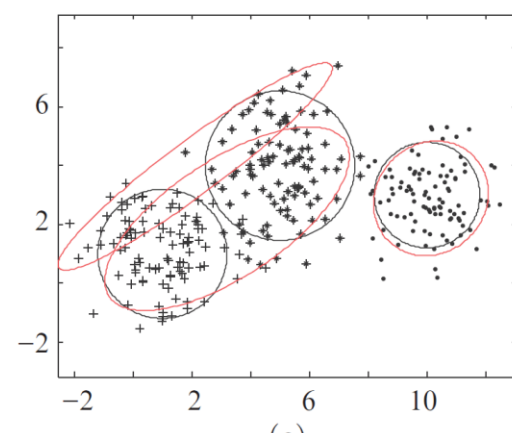
(b)



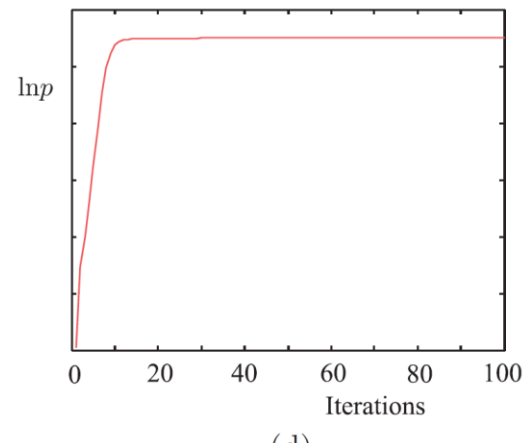
(c)



(d)



(c)

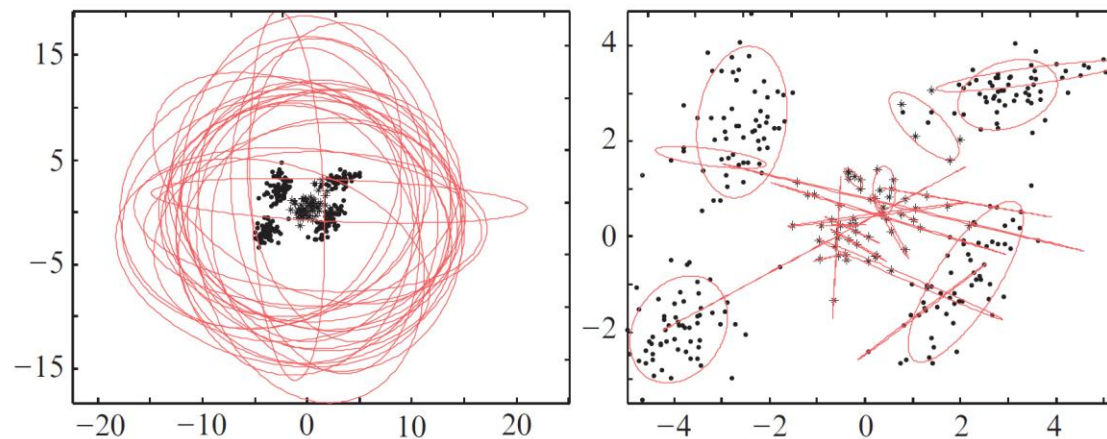


(d)

Estymacja wariacyjna Bayesa

ang. (*Variational Bayes*)

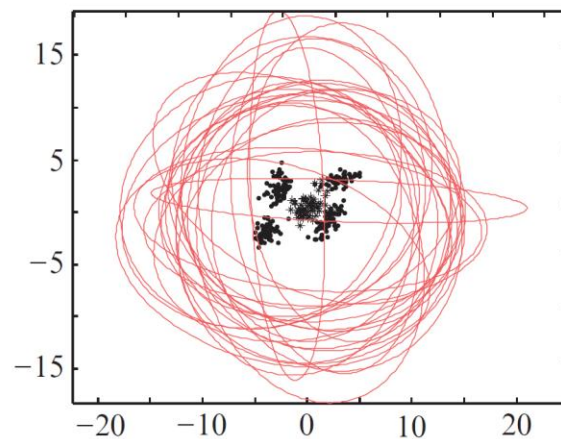
„Zwykła” mieszanina Gaussa



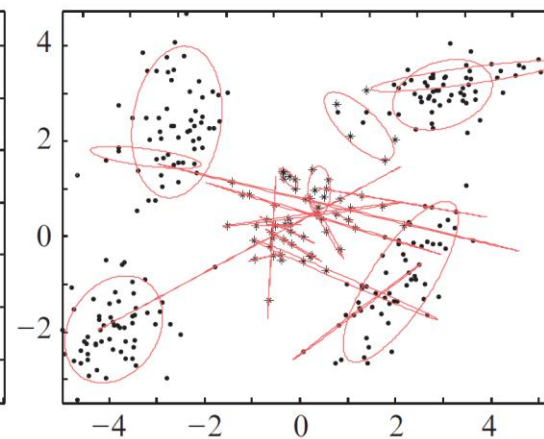
Estymacja wariacyjna Bayesa

ang. (*Variational Bayes*)

„Zwykła” mieszanina Gaussa

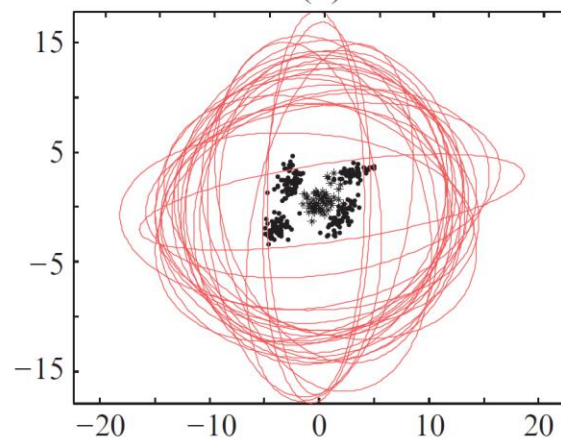


(a)

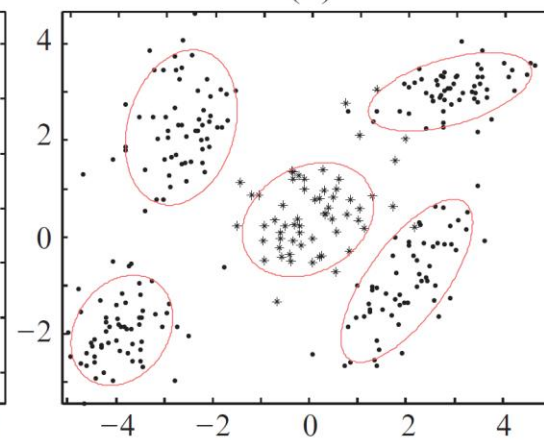


(b)

Wariacyjna metoda E-M



(c)



(d)

Przykłady algorytmów i podejść

- K-średnich (ang. *k-means*) / centroidów
- Mieszanina rozkładów Gaussa
 - *maksymalizacja oczekiwania (ang. **Expectation–Maximization**)*
- DBSCAN
- Widmowa analiza skupień
- Grupowanie twarde i miękkie
 - *rozmyte c-średnich (ang. *fuzzy c-means*)*
- Metody aglomeracyjne

```
sklearn.mixture.GaussianMixture  
                .BayesianGaussianMixture
```

```
sklearn.cluster.Kmeans  
                .DBSCAN  
                .AgglomerativeClustering  
                .SpectralClustering  
scipy.cluster.hierarchy.linkage  
                        .dendrogram
```

Ocena jakości grupowania

Odległość między dwoma wektorami

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{k=1}^K \left(\mathbf{x}_i^{(k)} - \mathbf{x}_j^{(k)} \right)^2}$$

Rozproszenie wewnątrzgrupowe
(ang. *Within-Cluster Sum of Squares*)

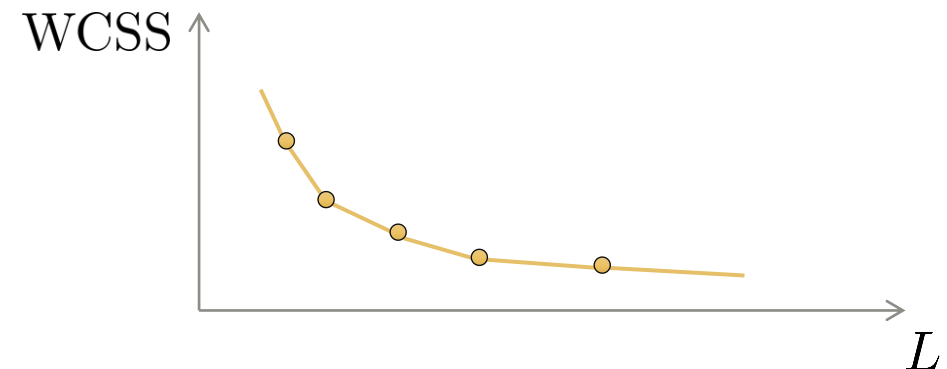
$$\text{WCSS} = \sum_{l=1}^L \sum_{i=1}^N g_{il} \cdot d(\mathbf{x}_i, \boldsymbol{\mu}_l)$$

$\mathcal{G}_l$ – grupa numer l ($l = 1, 2, \dots, L$)

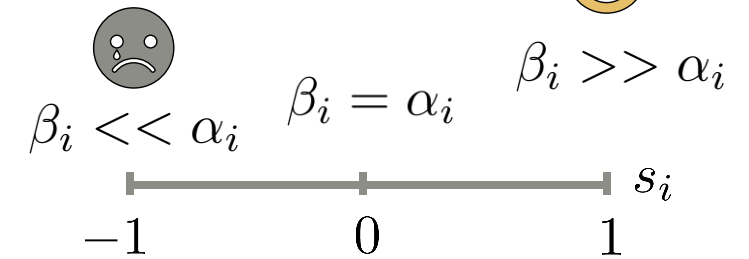
$$g_{il} = \begin{cases} 1 & \text{jeżeli } \mathbf{x}_i \in \mathcal{G}_l \\ 0 & \text{jeżeli } \mathbf{x}_i \notin \mathcal{G}_l \end{cases}$$

$\boldsymbol{\mu}_l$ – centroid grupy $\mathcal{G}_l$

■ metoda łokcia (ang. *elbow method*):



Ocena jakości grupowania



- Współczynnik profilu (ang. *silhouette coefficient*)

$$\alpha_i = \frac{1}{|\mathcal{G}_l| - 1} \sum_{\mathbf{x}_j \in \mathcal{G}_l, j \neq i} d(\mathbf{x}_i, \mathbf{x}_j) \quad \text{dla } \mathbf{x}_i \in \mathcal{G}_l$$

średnia odległość od „swoich”

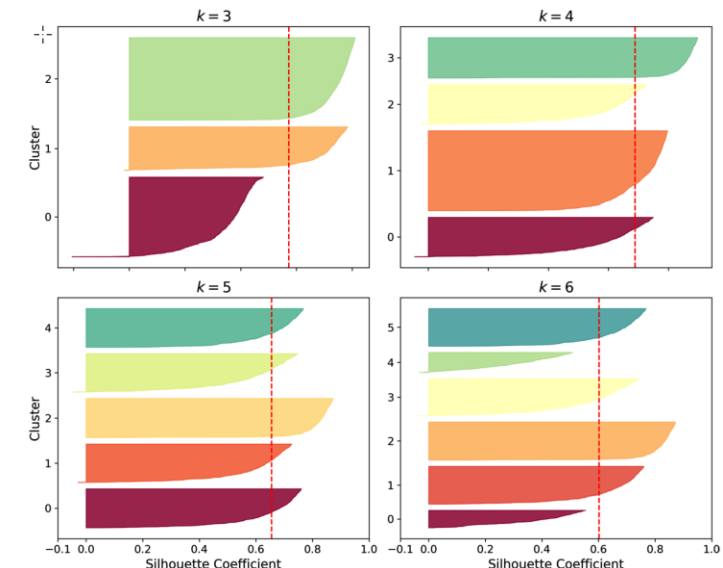
spójność grupy, czyli jak bardzo $\mathbf{x}_i$ przypomina elementy ze swojej grupy $\mathcal{G}_l$

$$\beta_i = \min_{r \neq l} \frac{1}{|\mathcal{G}_r|} \sum_{\mathbf{x}_j \in \mathcal{G}_r} d(\mathbf{x}_i, \mathbf{x}_j) \quad \text{dla } \mathbf{x}_i \in \mathcal{G}_l$$

średnia odległość od najbliższych „innych”

odstęp grupy, czyli jak bardzo $\mathbf{x}_i$ nie pasuje do innych grup

$$s_i = \frac{\beta_i - \alpha_i}{\max \{ \beta_i, \alpha_i \}}$$





THE END