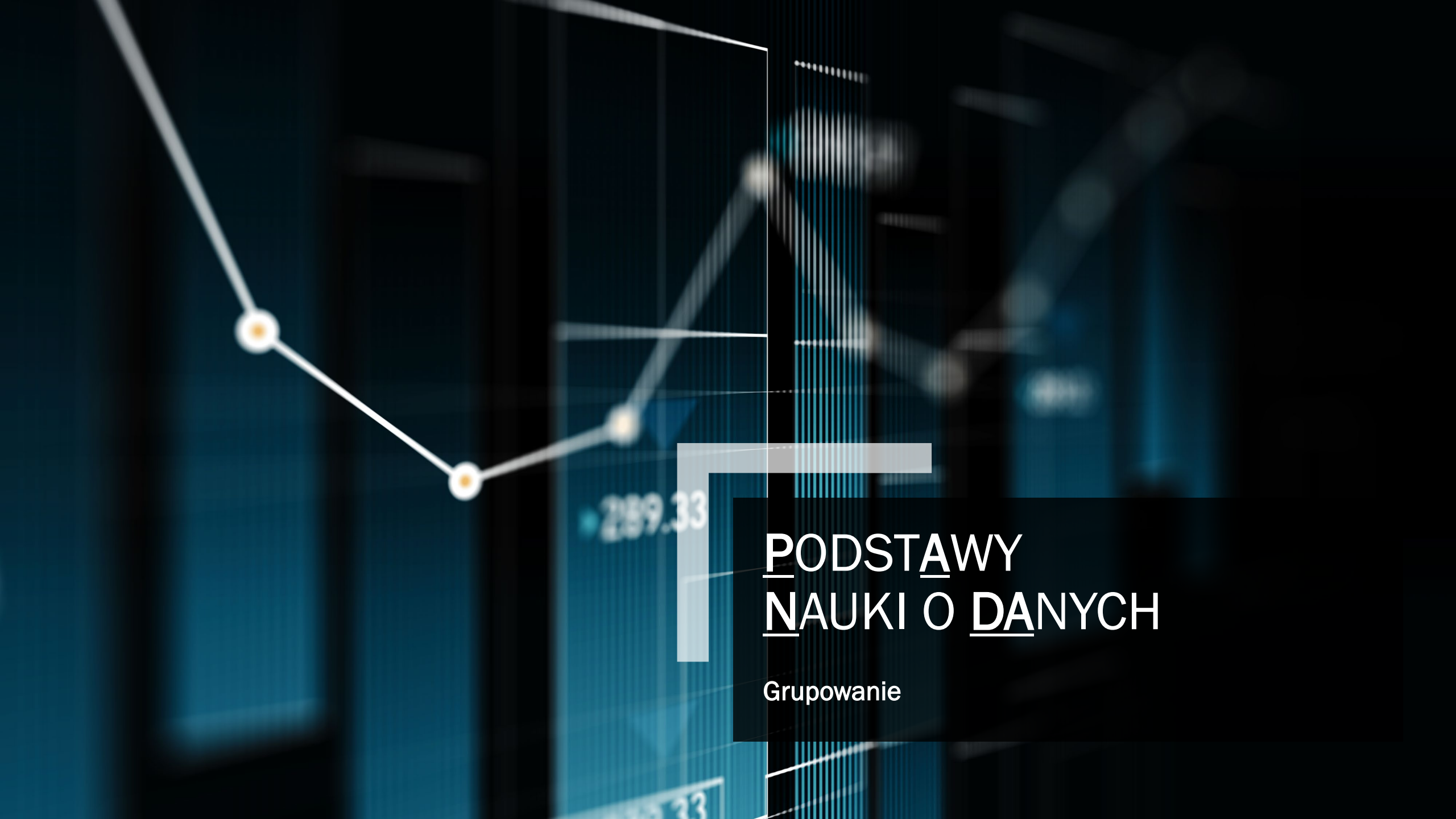


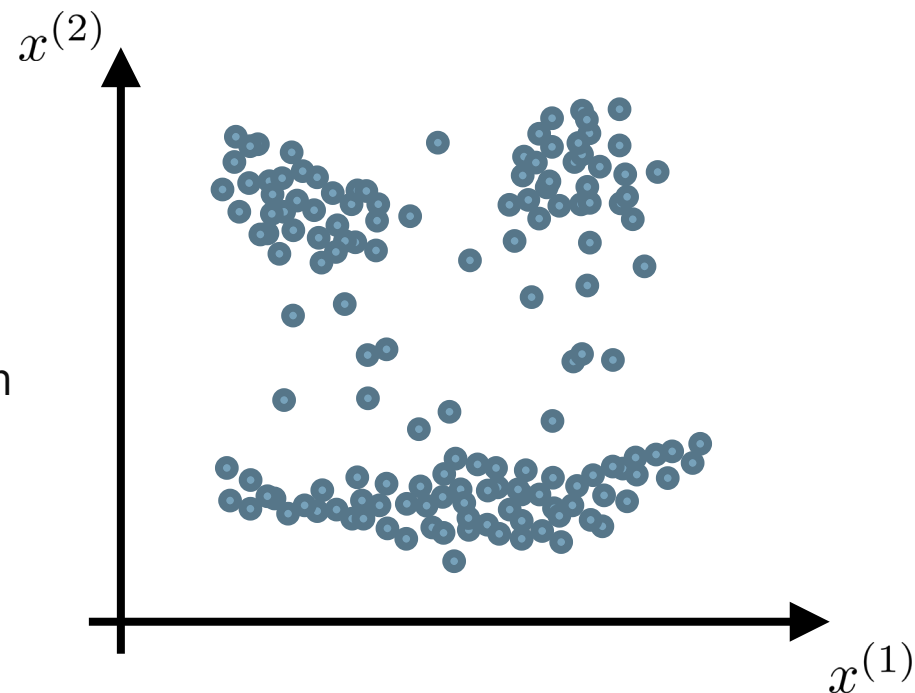
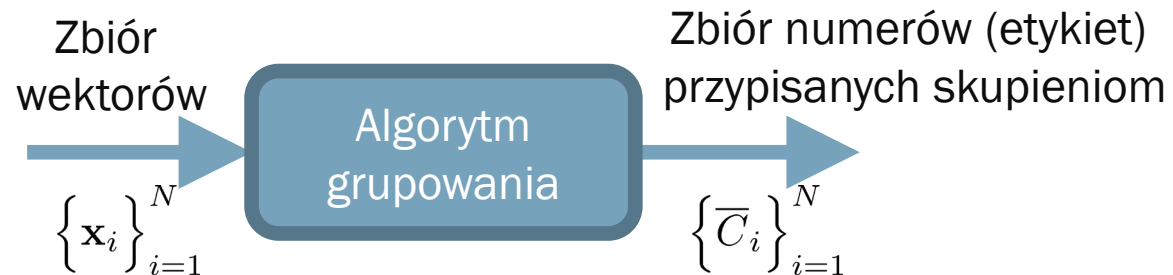


PODSTAWY NAUKI O DANYCH

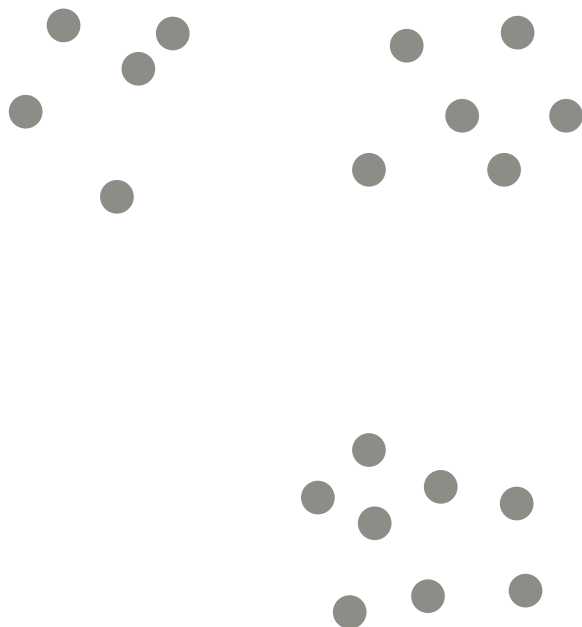
Grupowanie

Grupowanie

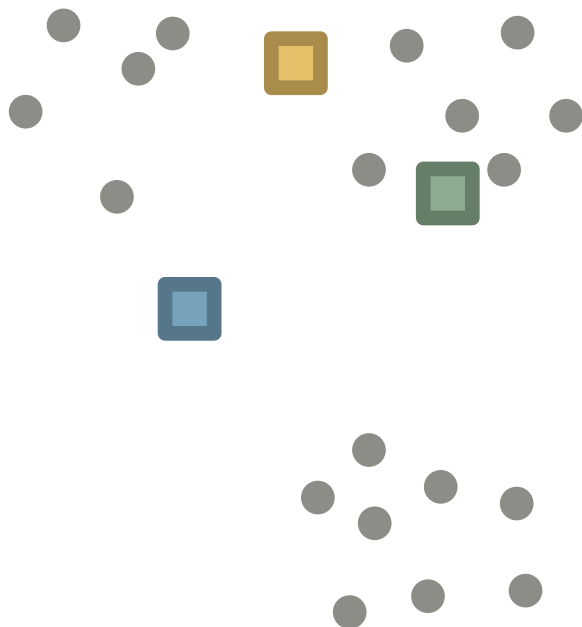
Dane
pomiarowe  $\{\mathbf{x}_i\}_{i=1}^N$



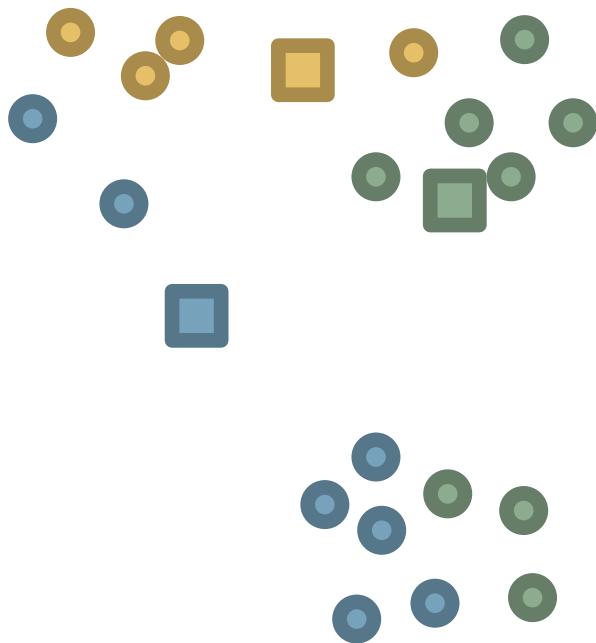
Algorytm k-średnich (ang. *K*-Means)



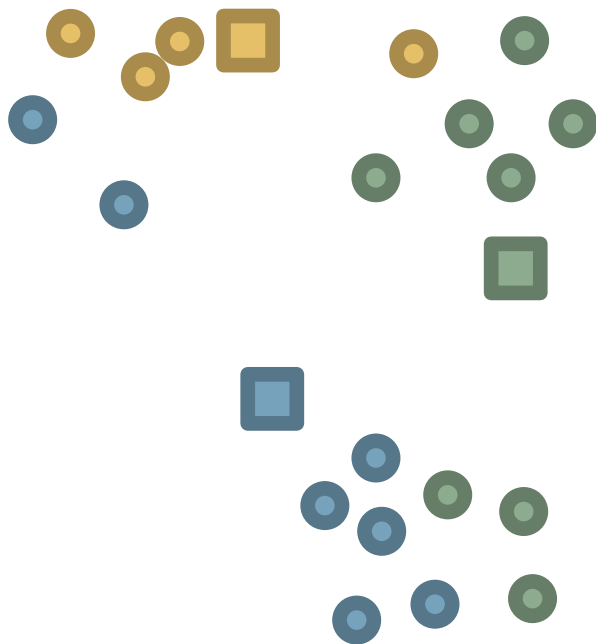
Algorytm k-średnich (ang. *K*-Means)



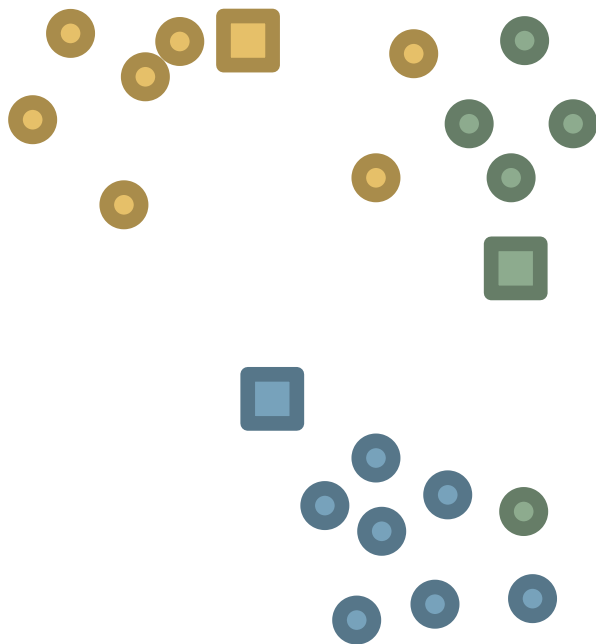
Algorytm k-średnich (ang. *K*-Means)



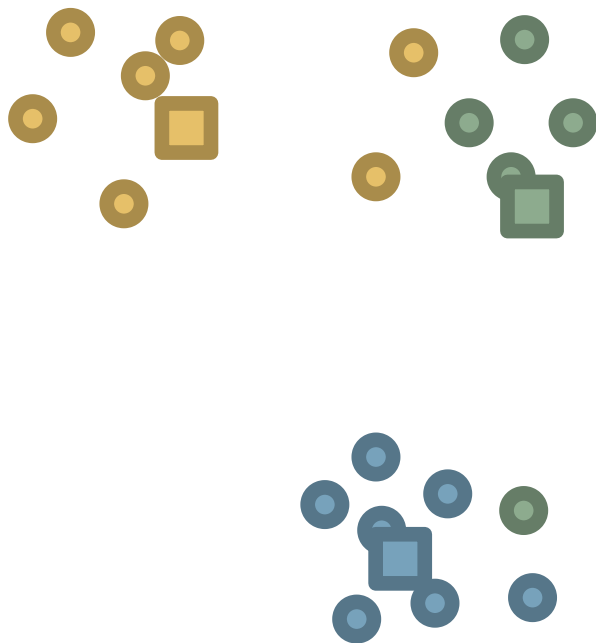
Algorytm k-średnich (ang. *K*-Means)



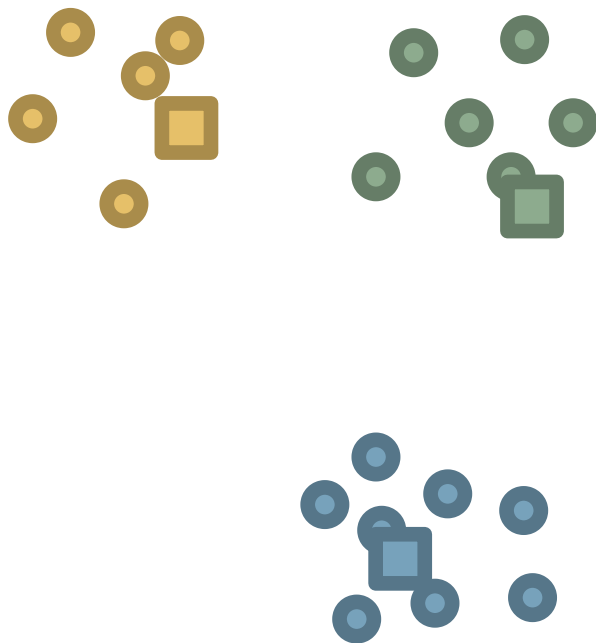
Algorytm k-średnich (ang. *K*-Means)



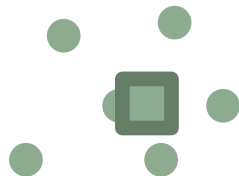
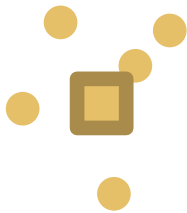
Algorytm k-średnich (ang. *K*-Means)



Algorytm k-średnich (ang. *K*-Means)



Algorytm k-średnich (ang. *K-Means*)



Ogólna procedura:

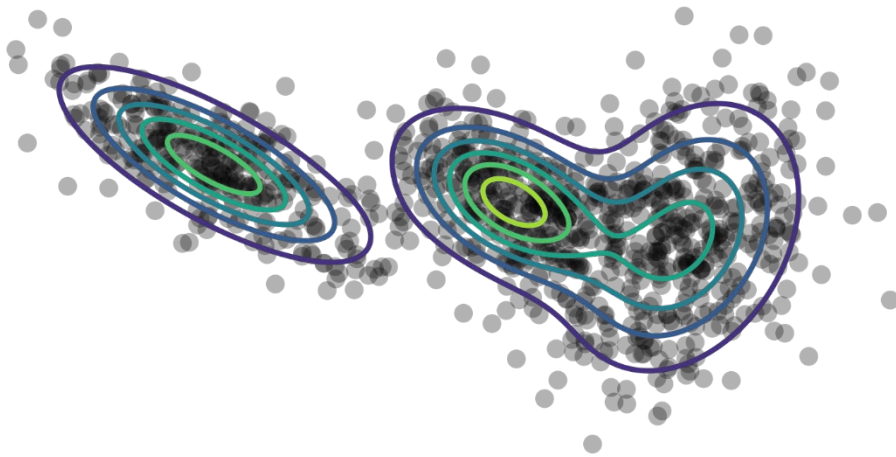
Metoda maksymalizacji oczekiwania

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

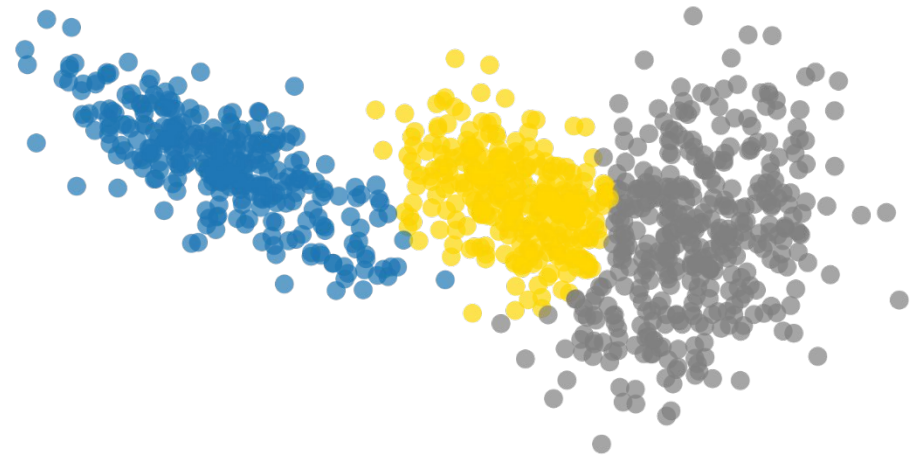
Mieszanina rozkładów Gaussa

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

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

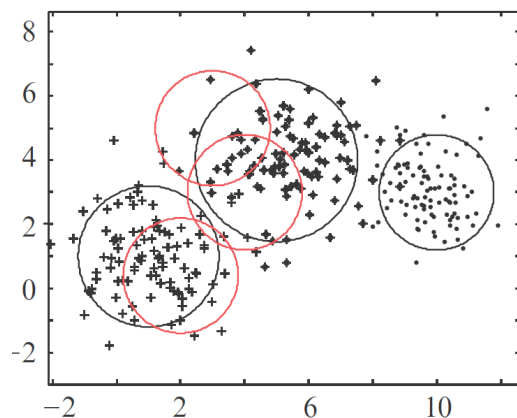


jako estymator rozkładu

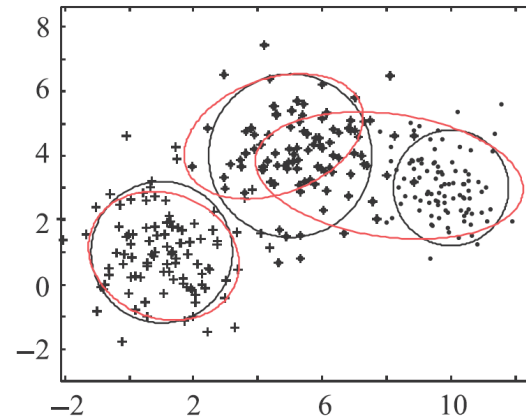


jako algorytm grupowania

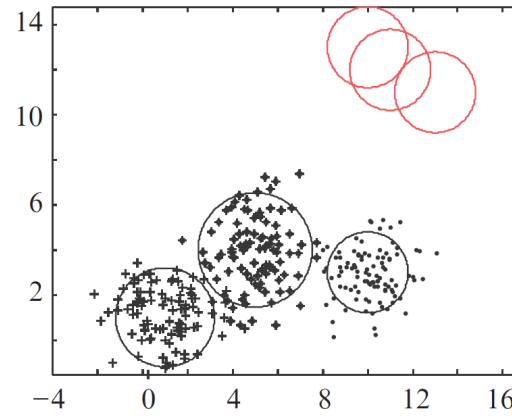
Mieszana rozkładów Gaussa



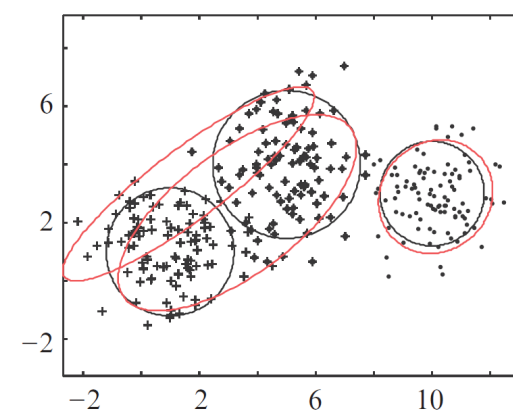
(a)



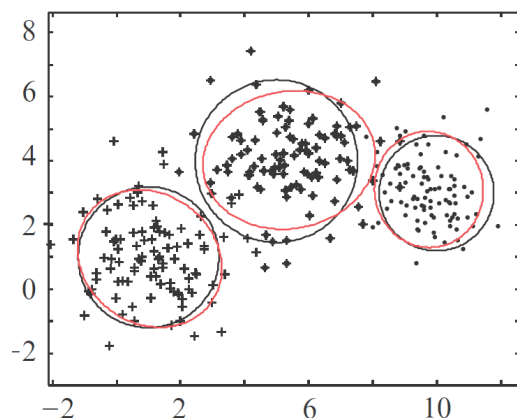
(b)



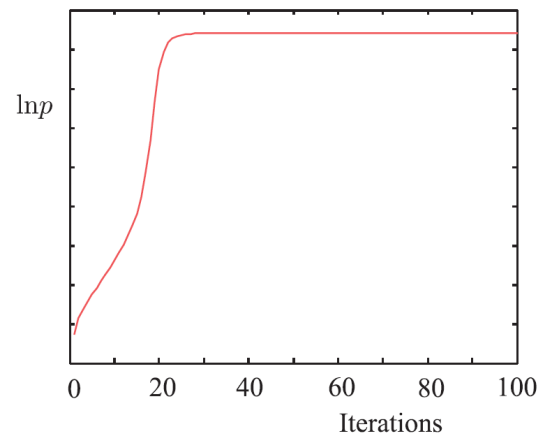
(a)



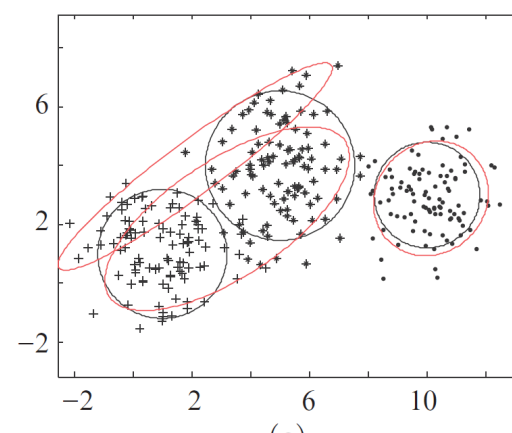
(b)



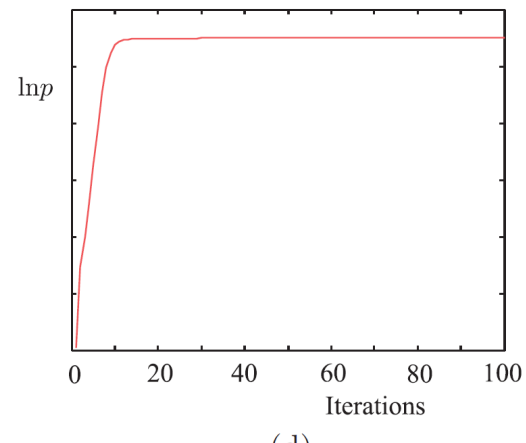
(c)



(d)



(c)

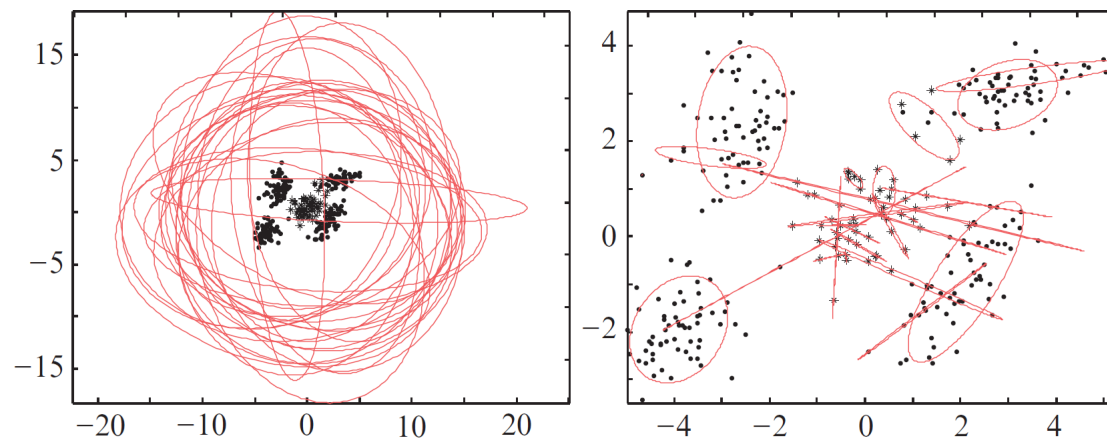


(d)

Estymacja wariacyjna Bayesa

ang. (*Variational Bayes*)

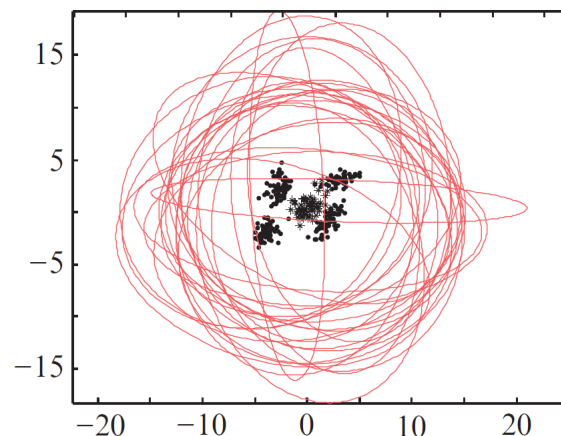
„Zwykła” mieszanina Gaussa



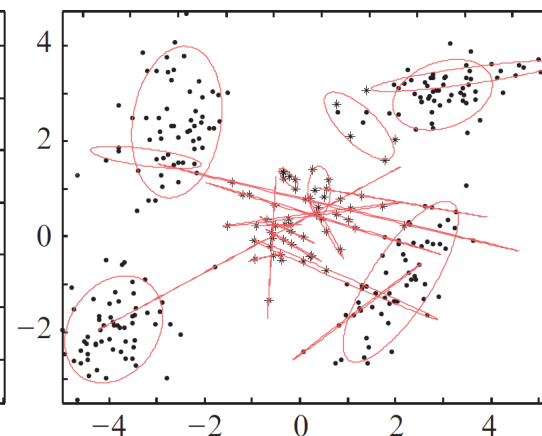
Estymacja wariacyjna Bayesa

ang. (*Variational Bayes*)

„Zwykła” mieszanina Gaussa

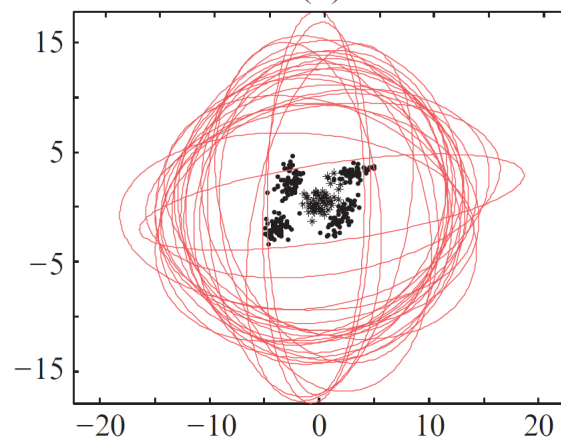


(a)

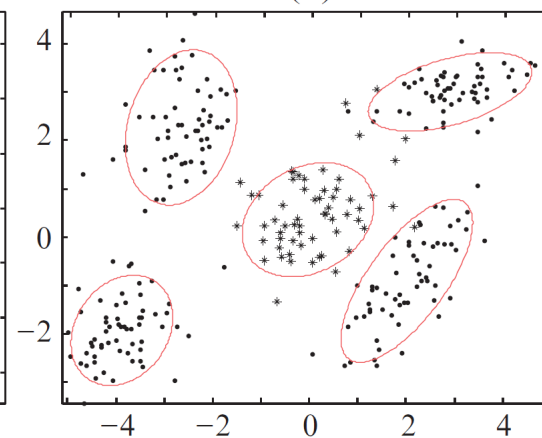


(b)

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
```

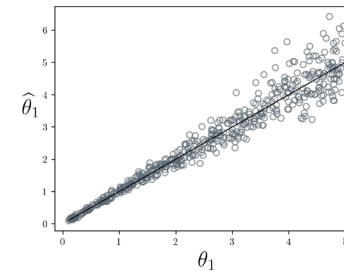
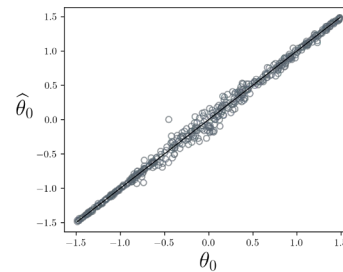
Uczenie maszynowe

jako

jako

Estymacja

Optymalizacja



**Dopasowanie do danych
z uwzględnieniem wiedzy**

Ewaluacja modelu

Ocena jakości grupowania

Odległość między dwoma wektorami

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

wewnątrzgrupowa (ang. *within-cluster*)
suma kwadratów błędów

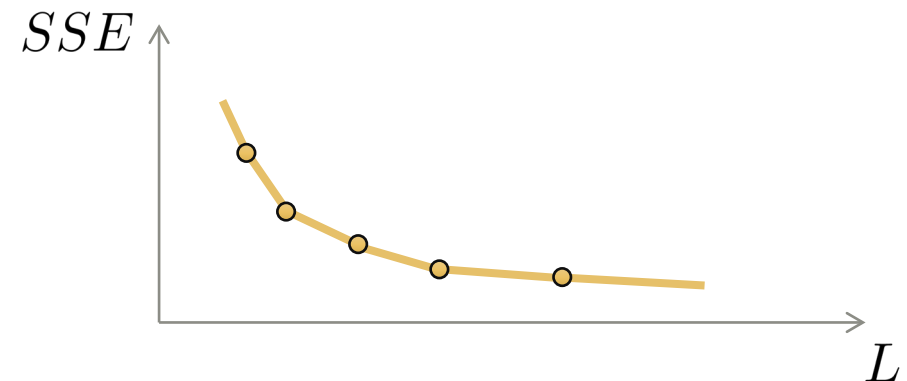
$$SSE = \sum_{l=1}^L \sum_{i=1}^N g_{il} \cdot d(\mathbf{x}_i, \boldsymbol{\mu}_l)$$

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

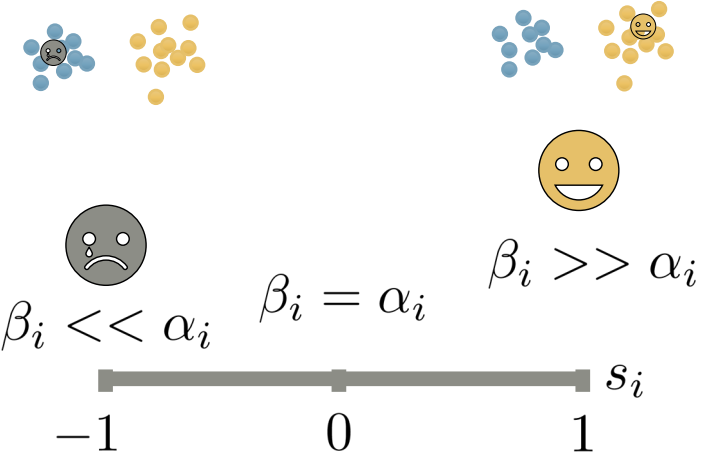
$$g_{il} = \begin{cases} 1 & \text{jeżeli } \mathbf{x}_i \in \mathcal{G}_l \\ 0 & \text{jeżeli } \mathbf{x}_i \notin \mathcal{G}_l \end{cases}$$

$\boldsymbol{\mu}_l$ – 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”

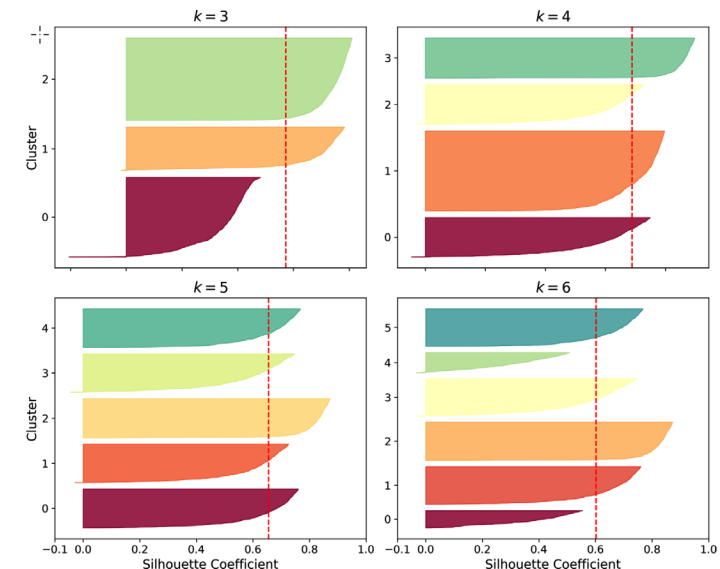
spójność grupy, czyli jak bardzo $\mathbf{x}_i$ przypomina elementy ze swojej grupy $\mathcal{G}_l$

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

średnia odległość od najbliższych „innych”

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

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



Przykład z rzeczywistymi danymi

	Hypertension	Diabetes mellitus	Hyper/Hypo-thyroidism	Insufficiencia renum chronica	Cerebral stroke in past	Myocardial infarction (MI) in past	MI STEMI	MI NSTEMI	Ejection Fraction (EF)
0	0	0	0	0	0	0	0	0	3
1	1	1	0	0	1	0	0	0	1
2	1	0	0	0	0	0	0	0	3
3	1	1	0	0	0	0	0	0	3
4	1	1	0	0	0	0	0	0	3
...	...	...	...	...	...	...	...	...	...
1063	0	0	0	1	0	0	0	0	3
1064	1	0	0	0	1	0	0	0	1
1065	1	1	2	1	0	1	0	1	3
1066	1	0	0	0	0	0	0	0	1
1067	0	0	2	1	0	1	0	1	3

Przykład z rzeczywistymi danymi

Algorytm k-średnich

Silhouette score: 0.2462636246354429

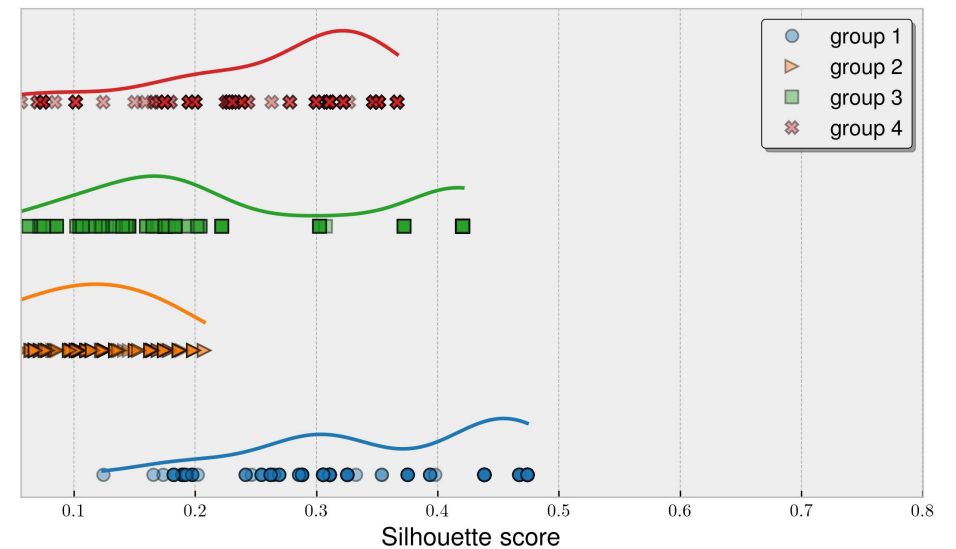
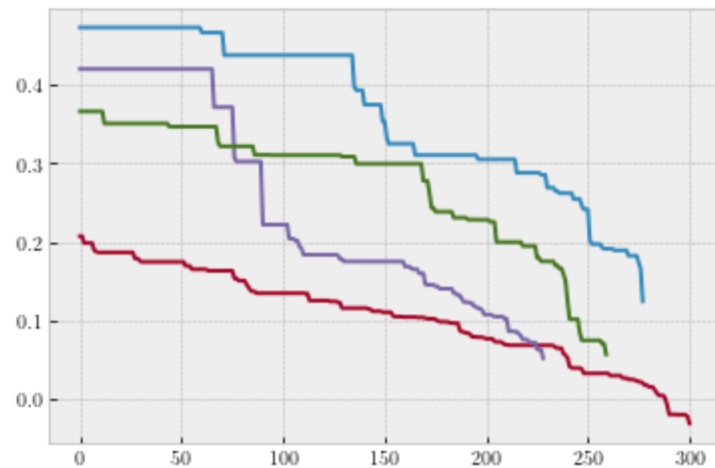
Silhouette samples:

min: 0.12419862001238247 - mean: 0.3686017970104325 - max: 0.4737554552800329

min: -0.03160587376035497 - mean: 0.10496007521610339 - max: 0.20741573709814043

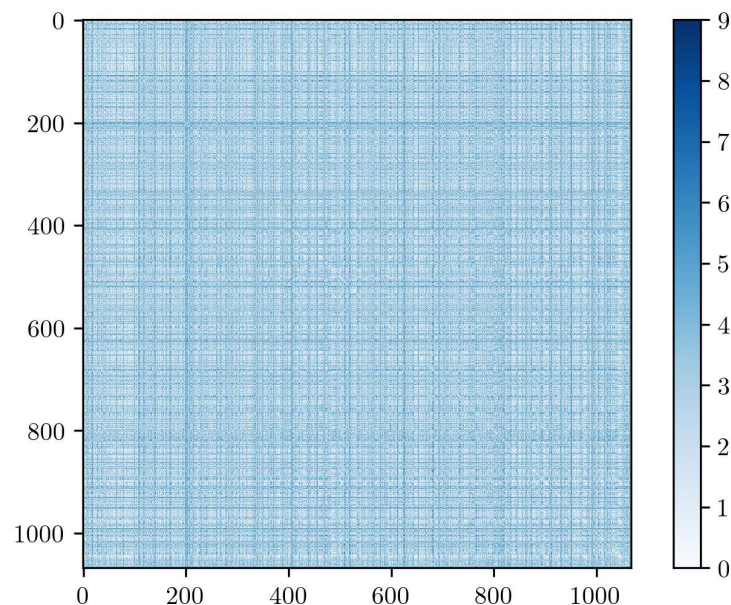
min: 0.05162826170957068 - mean: 0.24905195141256617 - max: 0.4207916336470454

min: 0.05635131085902938 - mean: 0.27658604626241556 - max: 0.36663559137835056



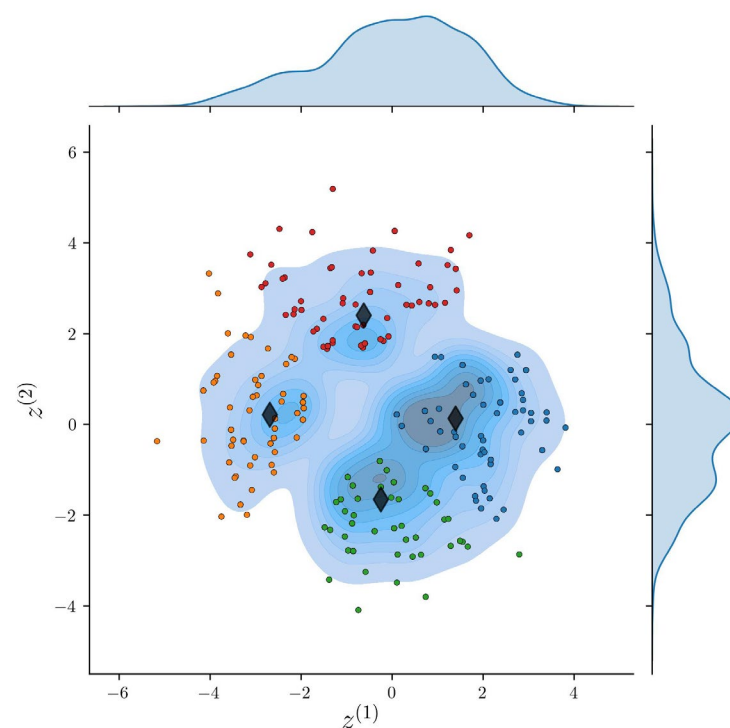
Przykład z rzeczywistymi danymi

Dwustopniowa procedura: mapowanie „dyskretne $\rightarrow$ ciągłe” + algorytm k -średnich



wzajemne odległości

$$d(\mathbf{x}_i, \mathbf{x}_j)$$



Skalowanie wielowymiarowe + k -średnich

Przykład z rzeczywistymi danymi

Dwustopniowa procedura: mapowanie „dyskretne $\rightarrow$ ciągłe” + algorytm k-średnich

Silhouette score: 0.5075580157657357

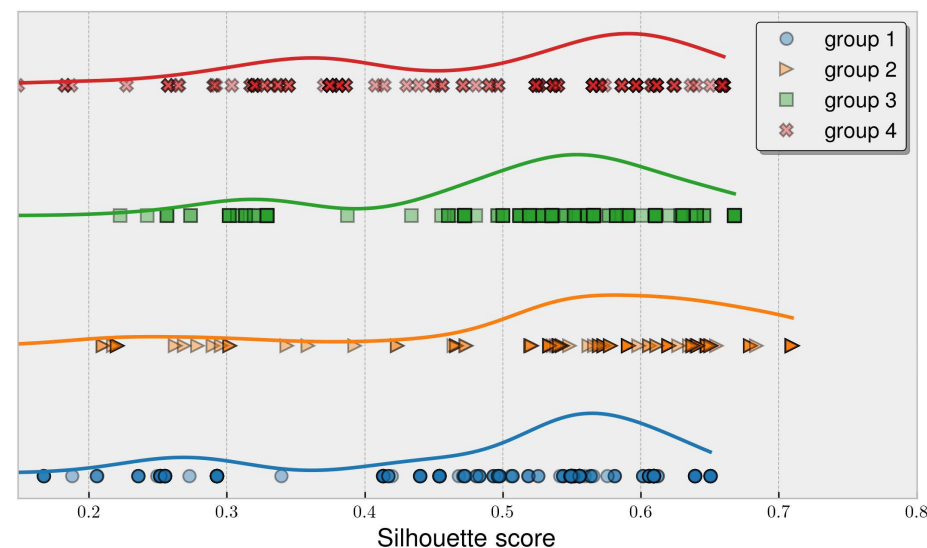
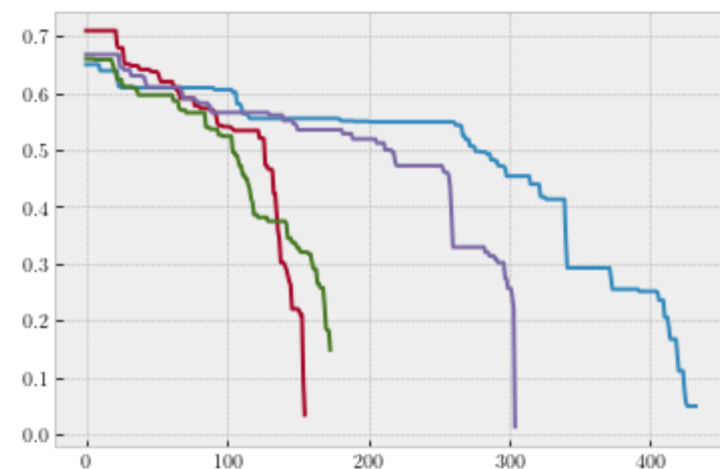
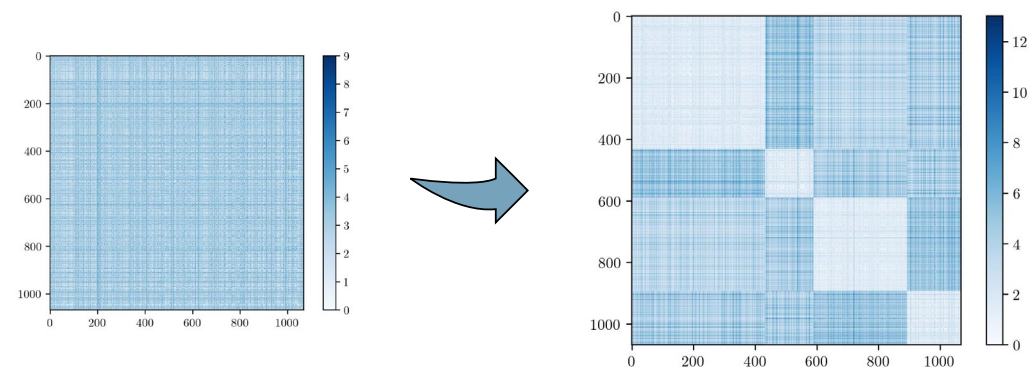
Silhouette samples:

min: 0.04952585823864839 - mean: 0.484265784367152 - max: 0.650284282753344

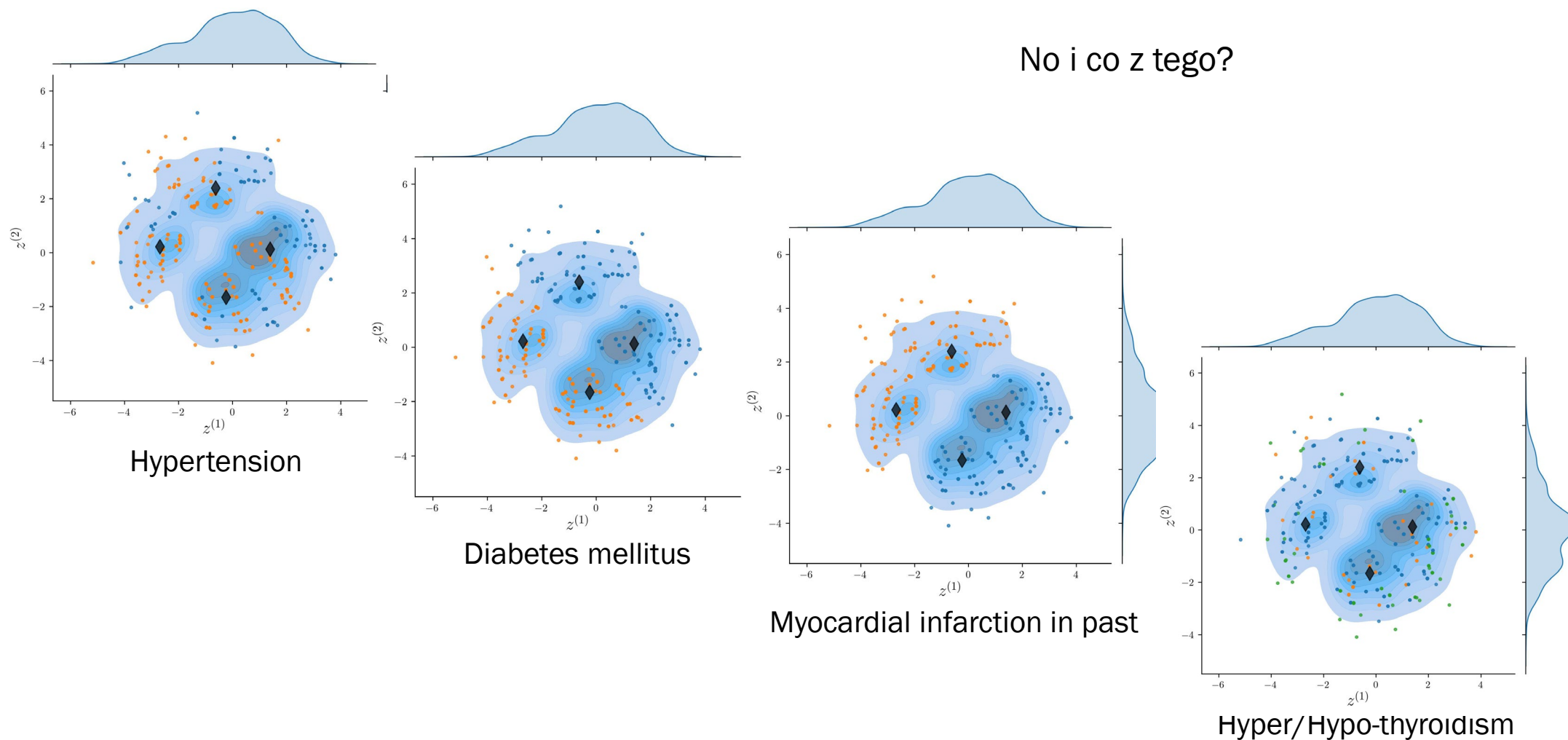
min: 0.034406633629570575 - mean: 0.5547471272405489 - max: 0.7096248564520437

min: 0.013668067157811431 - mean: 0.5216354333625124 - max: 0.6677271046247377

min: 0.14856462374568624 - mean: 0.498537455067454 - max: 0.6601389928677166



Przykład z rzeczywistymi danymi





THE END