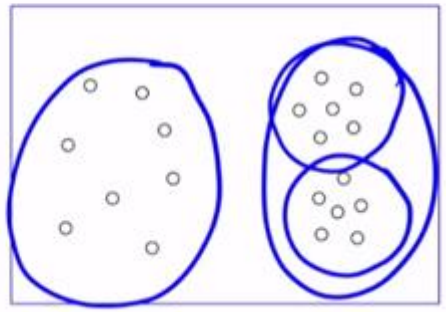
KLASTERYZACJA

Clustering in $\mathbb{R}^d$



Przypisywanie labels do zbioru danych bez labels

Vector quantization - wydzielenie, naturalnych grup danych (poszukiwanie skończonego zbioru reprezentantów w wielowymiarowej przestrzeni)

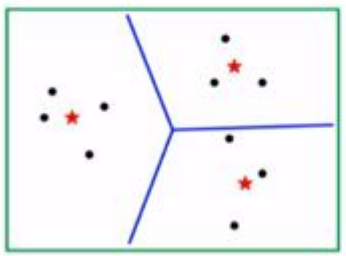
Poszukiwanie znaczących struktur w danych (Pogrupowanie danych w sensowne grupy (klastry))

K-Means algorytm

The k-means optimization problem

- Input: Points $x_1, \ldots, x_n \in \mathbb{R}^d$; integer k
- Output: "Centers", or representatives, $\mu_1, \ldots, \mu_k \in \mathbb{R}^d$
- Goal: Minimize average squared distance between points and their nearest

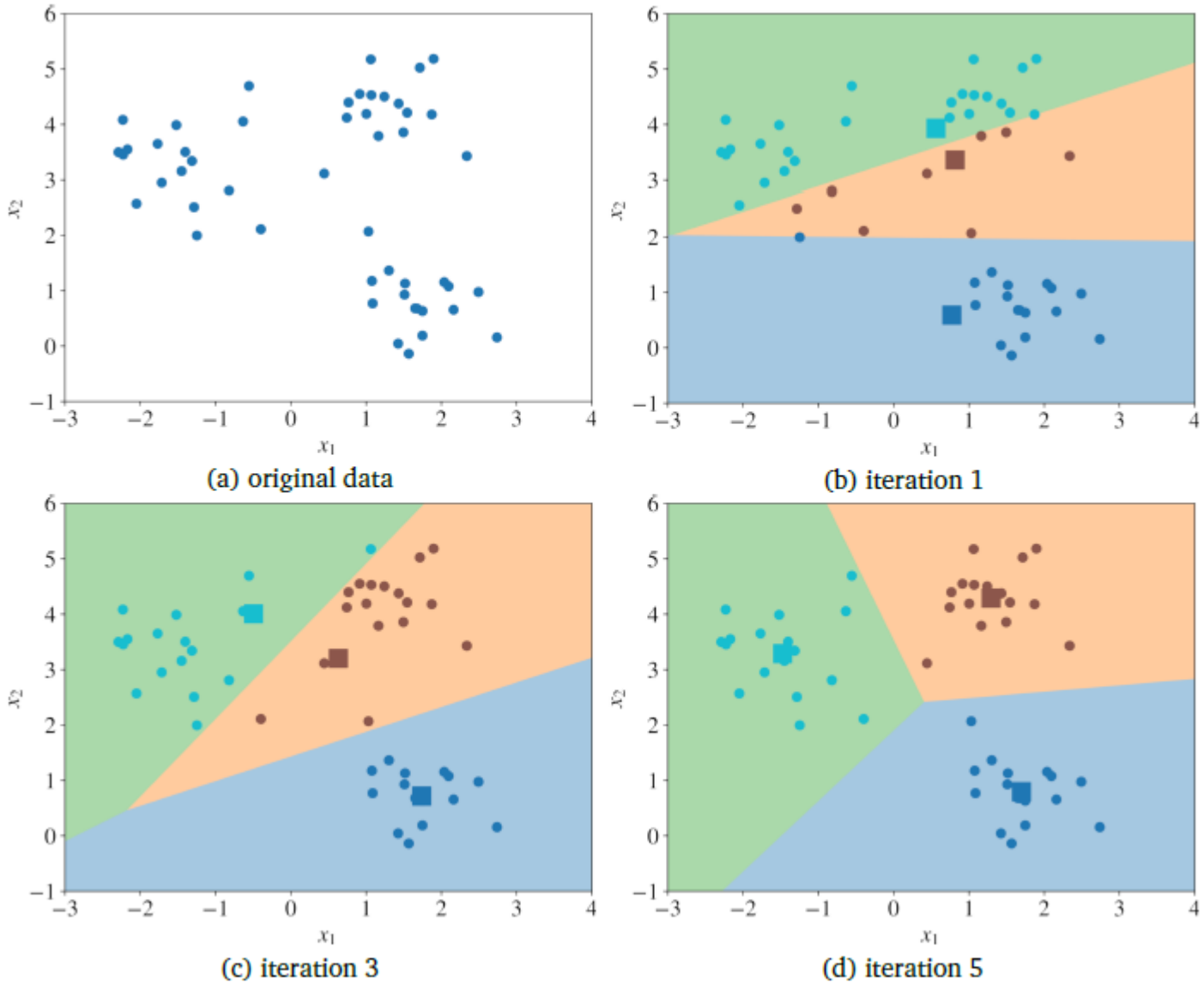
$$\text{cost}(\mu_1, \ldots, \mu_k) = \sum_{i=1}^{n} \min_j \|x_i - \mu_j\|^2$$



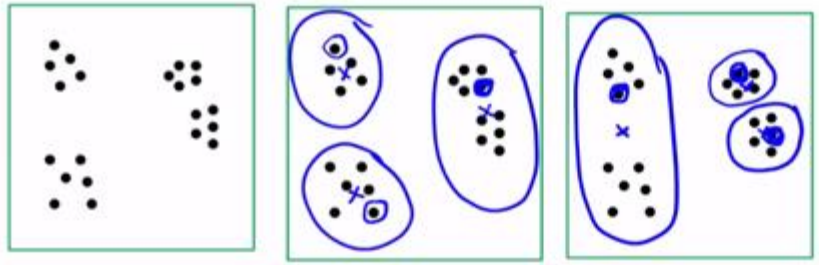
The centers partition $\mathbb{R}^d$ into k convex regions: μ_j's region consists of points for which it is the closest center.

Lloyd's k-means algorithm

- Initialize centers $\mu_1, \ldots, \mu_k$ in some manner.
- Repeat until convergence:
  - Assign each point to its closest center.
  - Update each μ_j to the mean of the points assigned to it.



Wartość k, określająca liczbę klastrów, określa się za pomocą różnych metod.



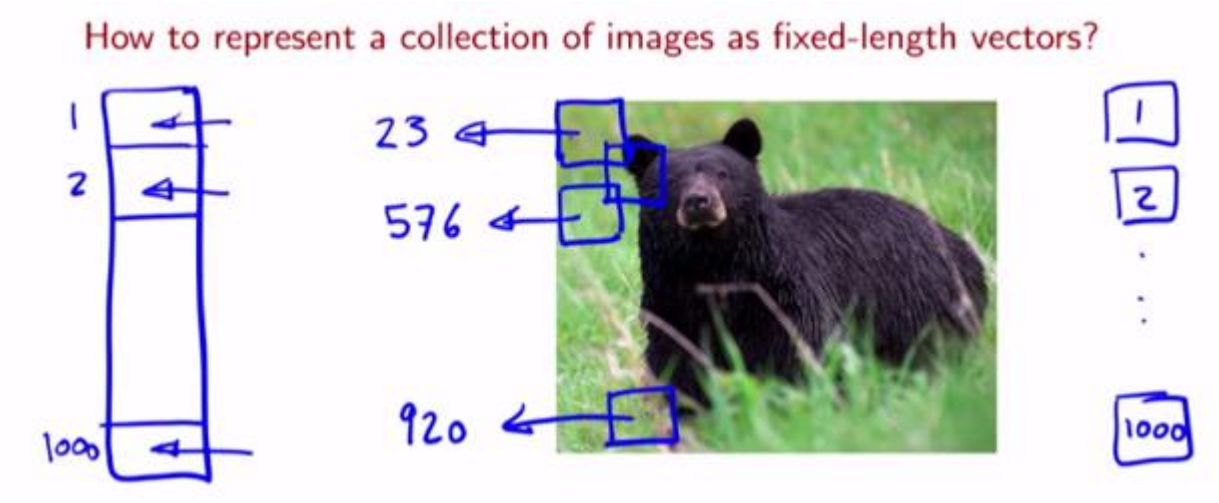
Initializing the k-means algorithm

Typical practice: choose k data points at random as the initial centers.

Another common trick: start with extra centers, then prune later.

Chcemy 10 robimy 20 następnie je usuwamy (zbyt blisko do innych, mają zbyt mało punktów)

Reprezentowanie obrazów za pomocą k-means codewords



1. Patche 10x10 pobierz ich bardzo dużo (zrób to dla każdego zdjęcia oddzielnie), Każdy patch to wektor o długości 100 (10x10). W ten sposób otrzymamy ogromną liczbę wektorów (reprezentujących każde zdjęcie).

2. Przeprowadzamy klasteryzację metodą k-means ze wszystkich wektorów dla np. 1000 centrów czyli otrzymujemy 1000 reprezentatywnych patches – ponumerowanych 1 – 1000.

3. Sprawdzamy dla każdego zdjęcia do którego reprezentatywnego patcha należy każdy z patch zdjęcia i zliczmy ich liczbę dla każdej klasy 1-1000, otrzymujemy 1000 elementowy wektor reprezentujący zdjęcie.