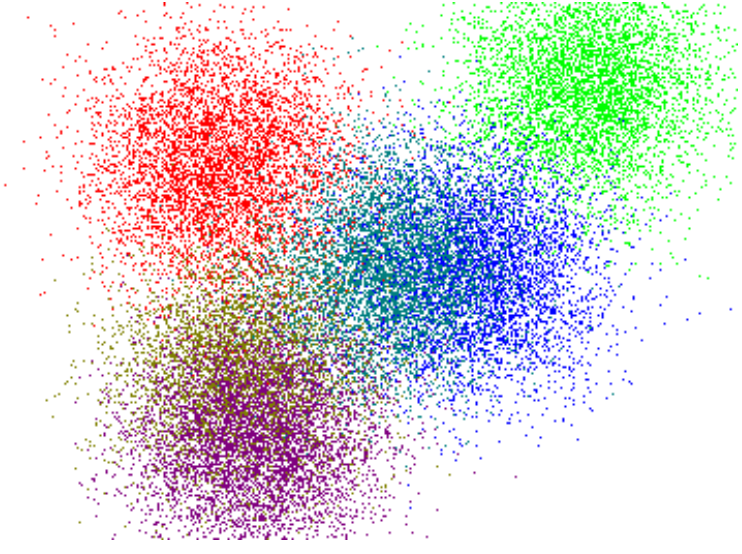
Agrupamento com k-means

Jones Granatyr

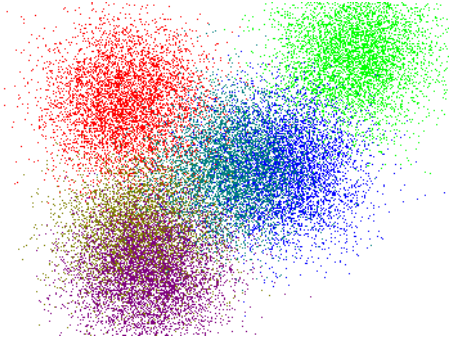
Agrupamento (cluster)

- Segmentação de mercado
- Encontrar grupos de clientes que irão comprar um produto (mala direta)
- Agrupamento de documentos/notícias
- Agrupamento de produtos similares
- Perfis de clientes (Netflix)
- Análise de redes sociais



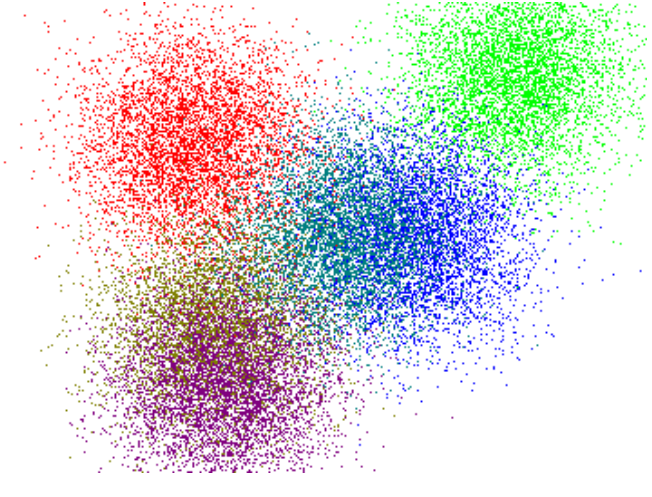
Agrupamento (cluster)

- Aprendizagem não supervisionada
- Classificação/regressão
  - Modelo que relaciona características com uma variável a ser prevista
- Agrupamento
  - Cria novos dados
  - Não tem um rótulo e o algoritmo aprende as relações entre os dados
- Identificar quando um grupo começa e outro termina
- Elementos dentro de um grupo devem ser similares e diferentes dos que estão fora do grupo (nearest mean)

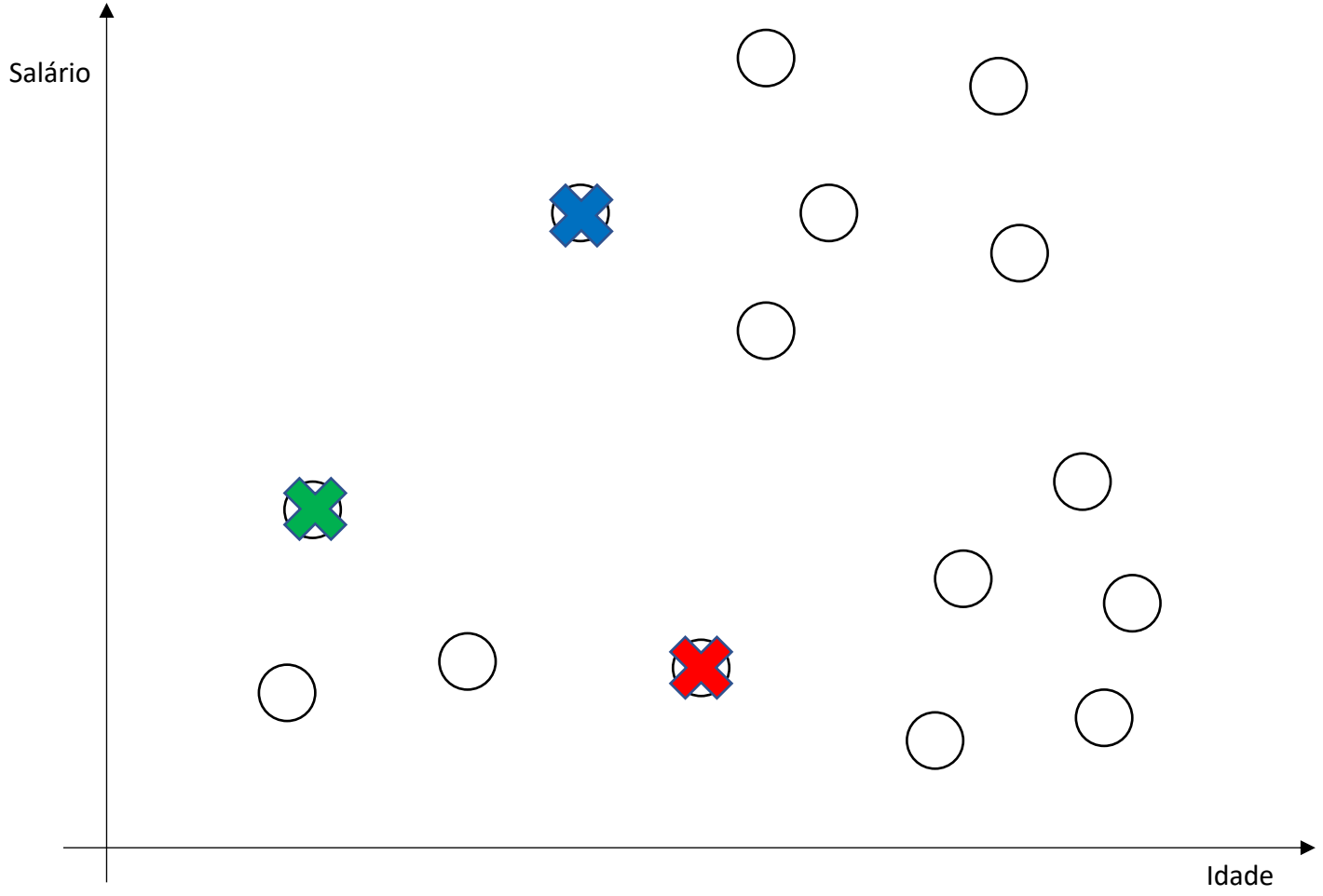


Algoritmo de Lloyd (k-means)

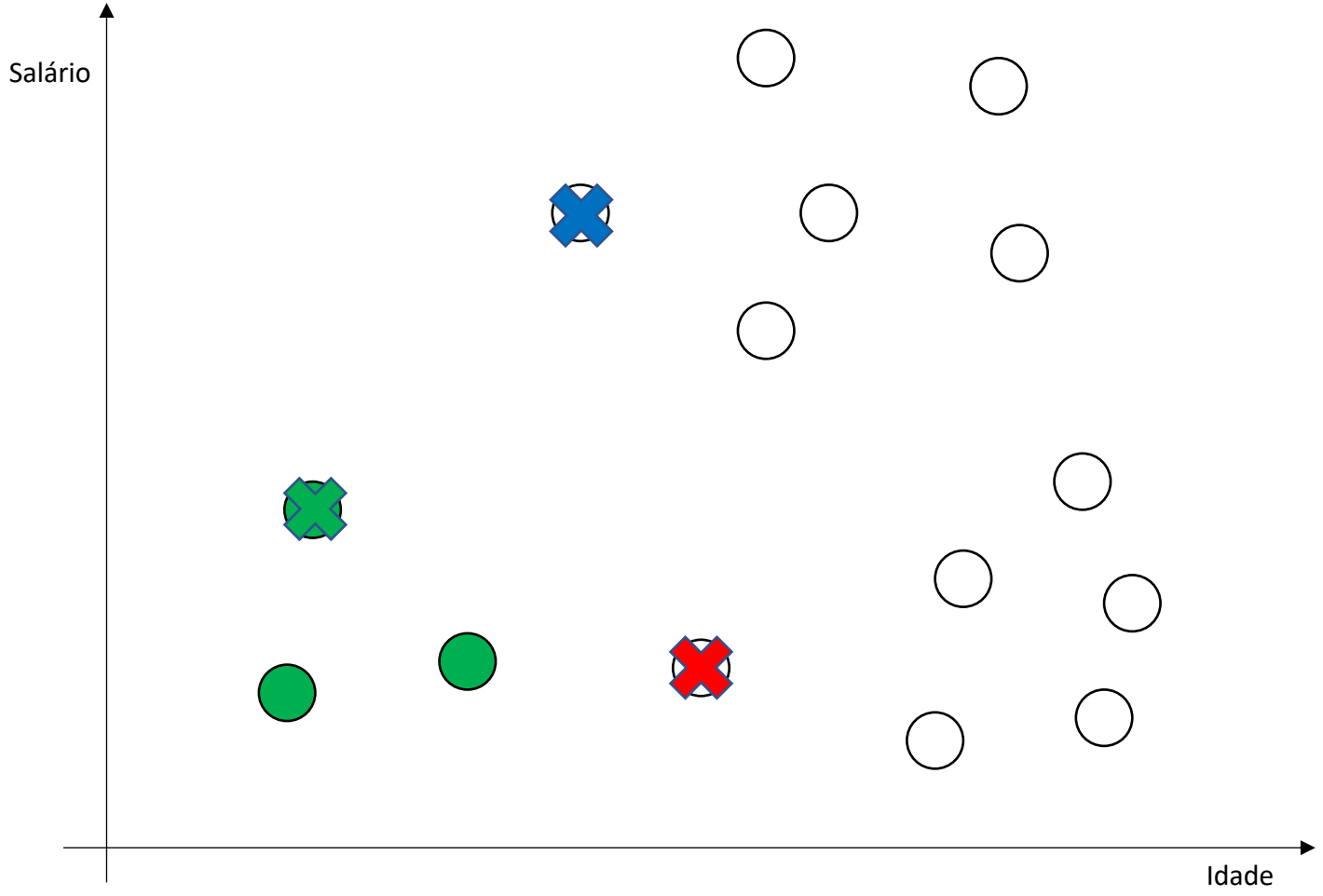
1. Inicializar os centroides aleatoriamente (centros de um cluster)
2. Para cada ponto na base de dados, calcular a distância para cada centroide e associar ao que estiver mais perto
3. Calcular a média de todos os pontos ligados a cada centroide e definir um novo centroide (repetir as etapas 2 e 3)



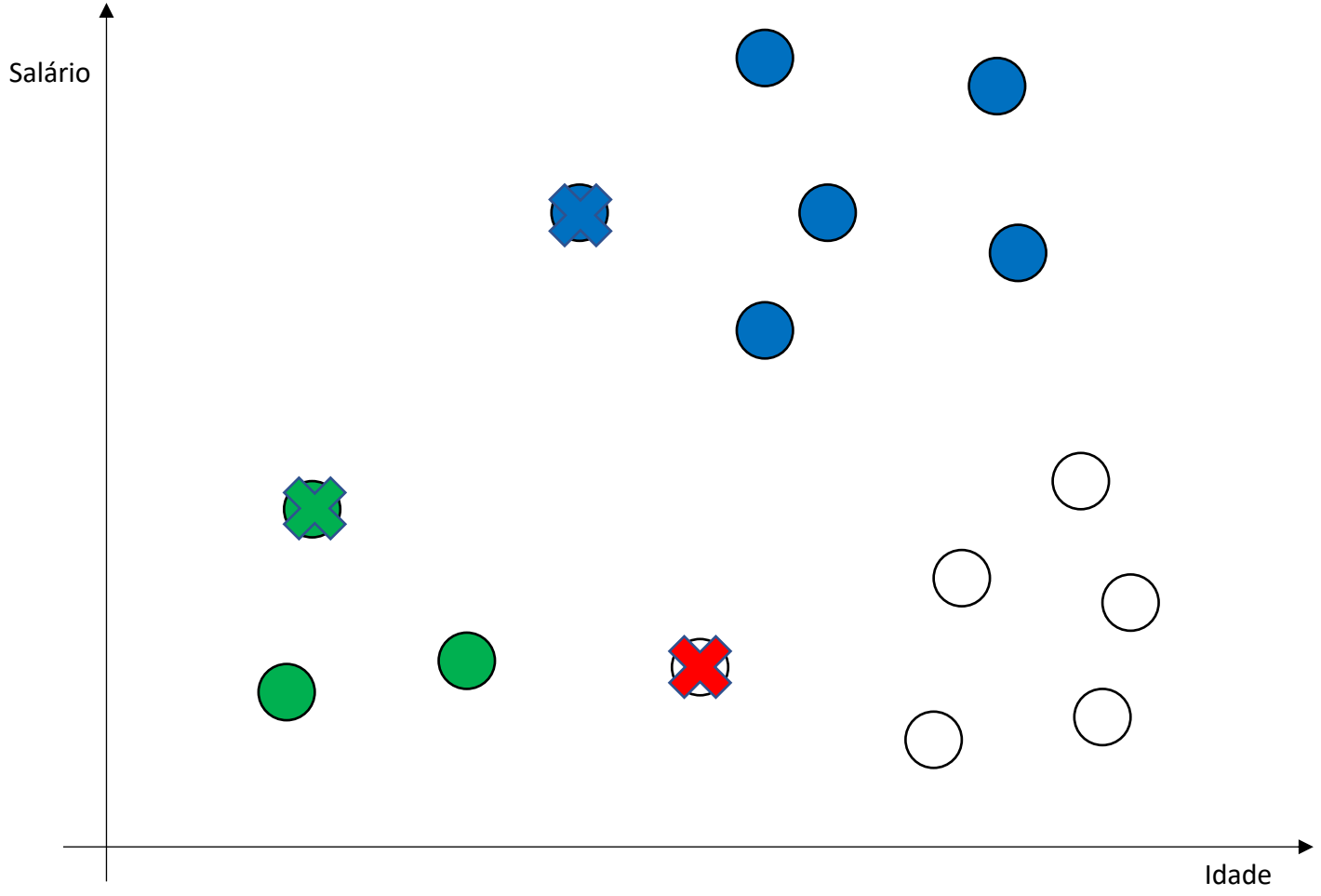
Algoritmo k-means



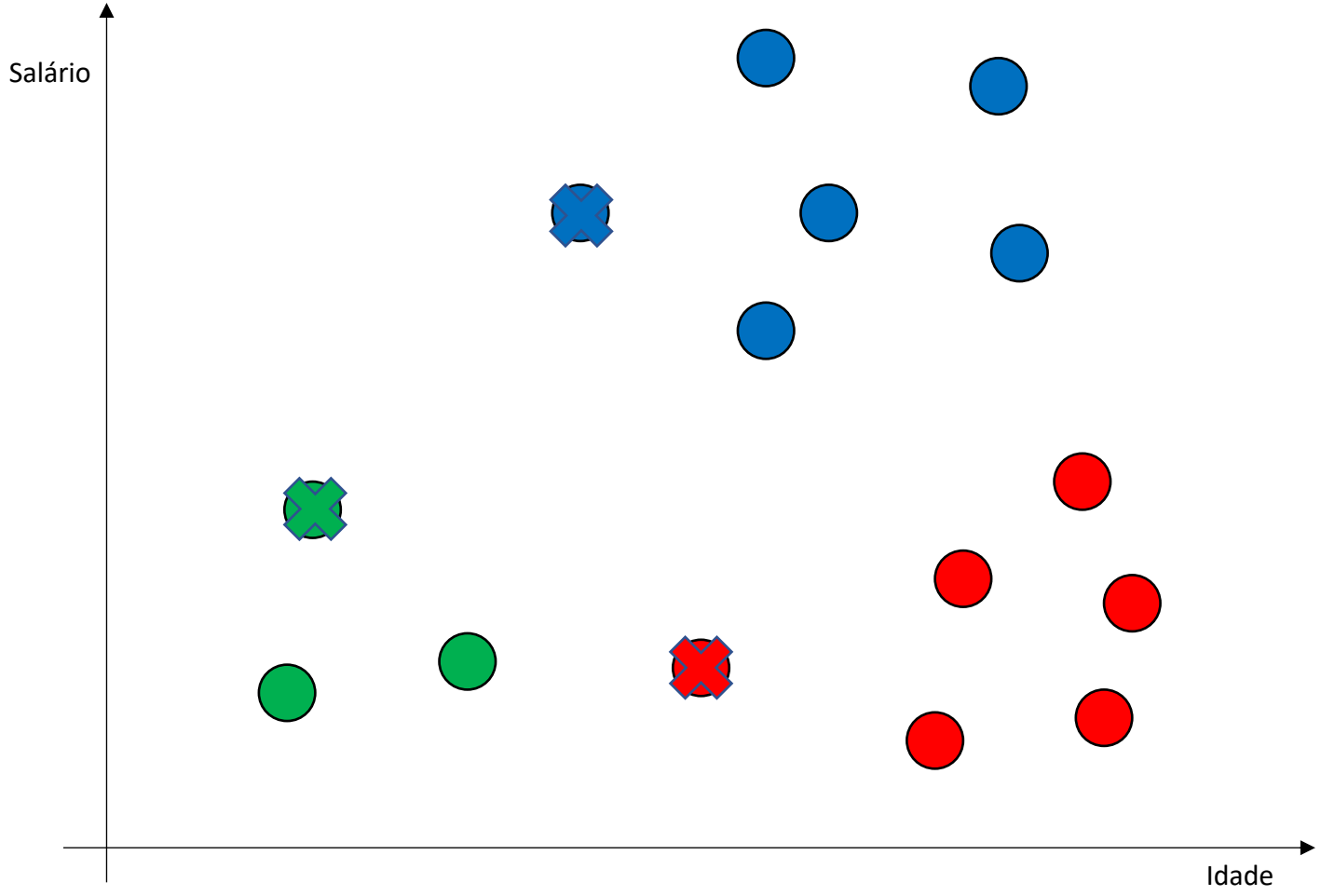
Algoritmo k-means



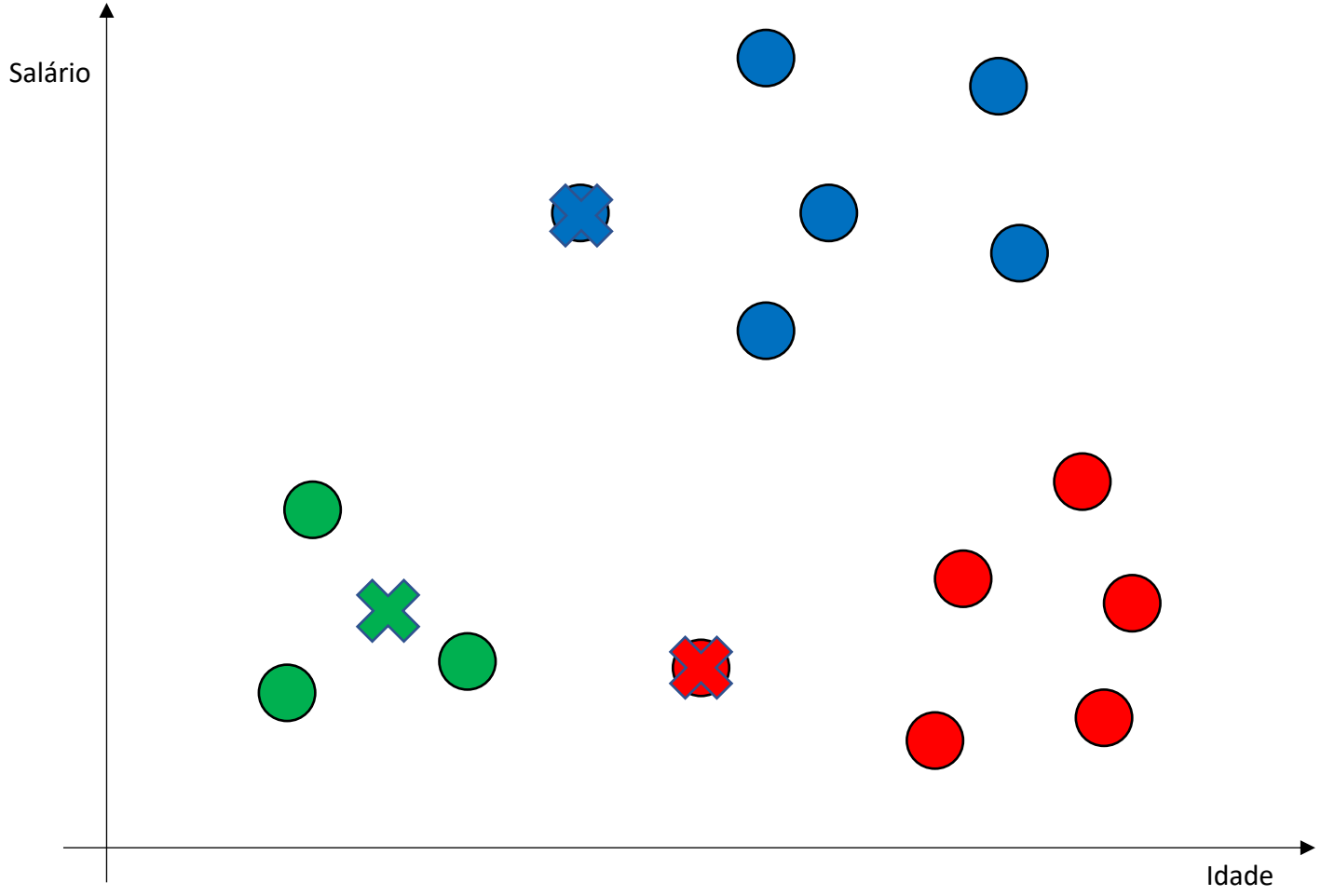
Algoritmo k-means



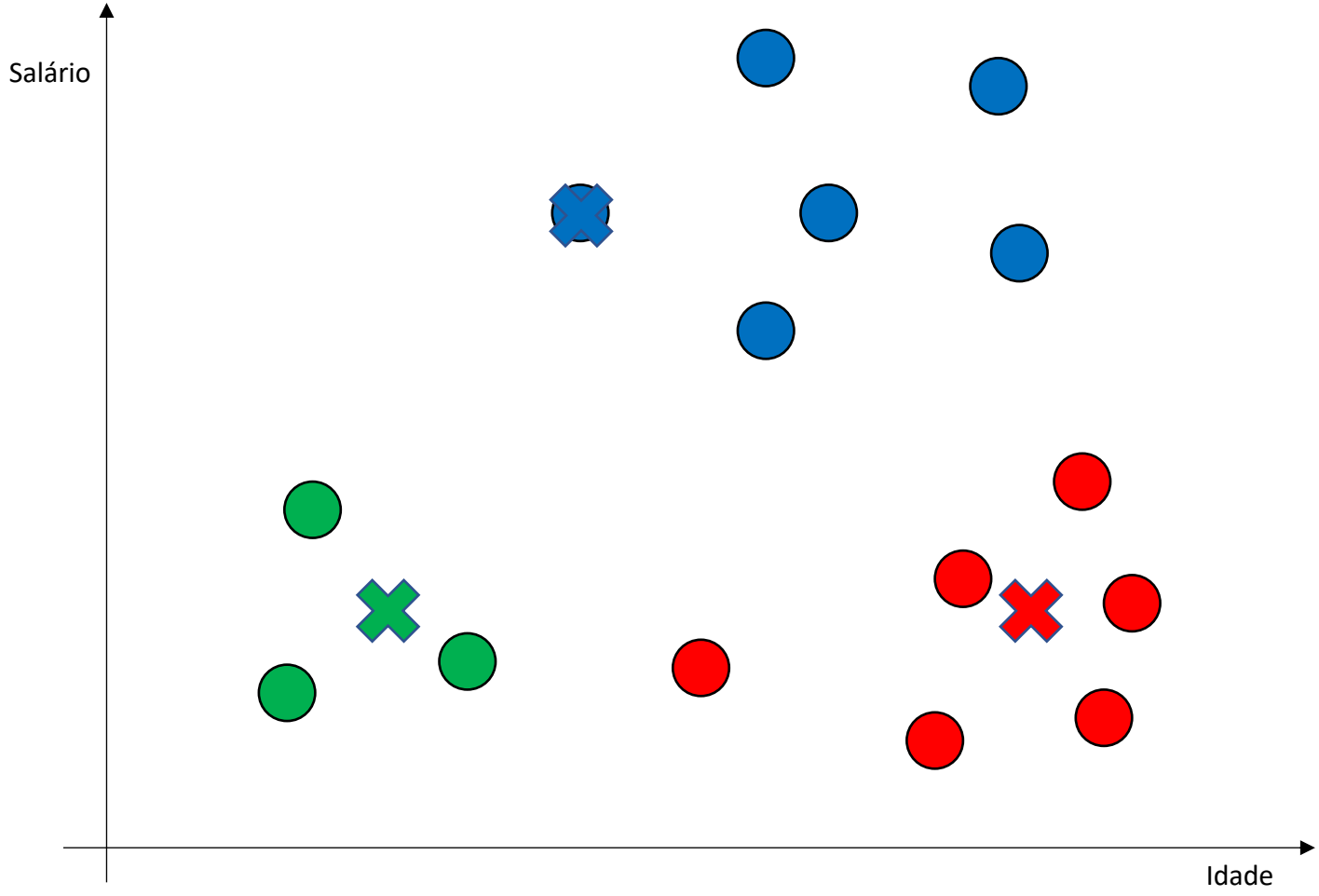
Algoritmo k-means



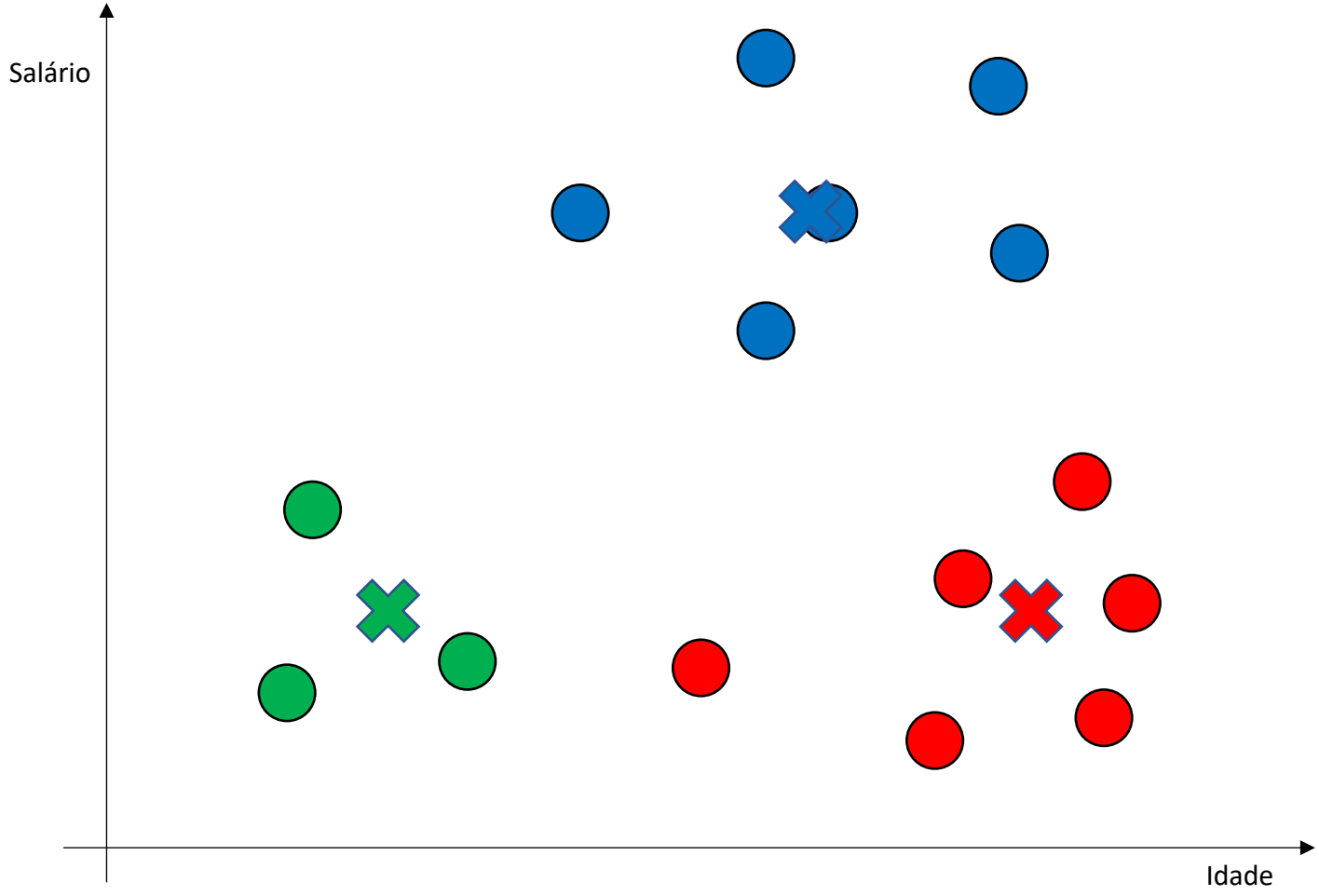
Algoritmo k-means



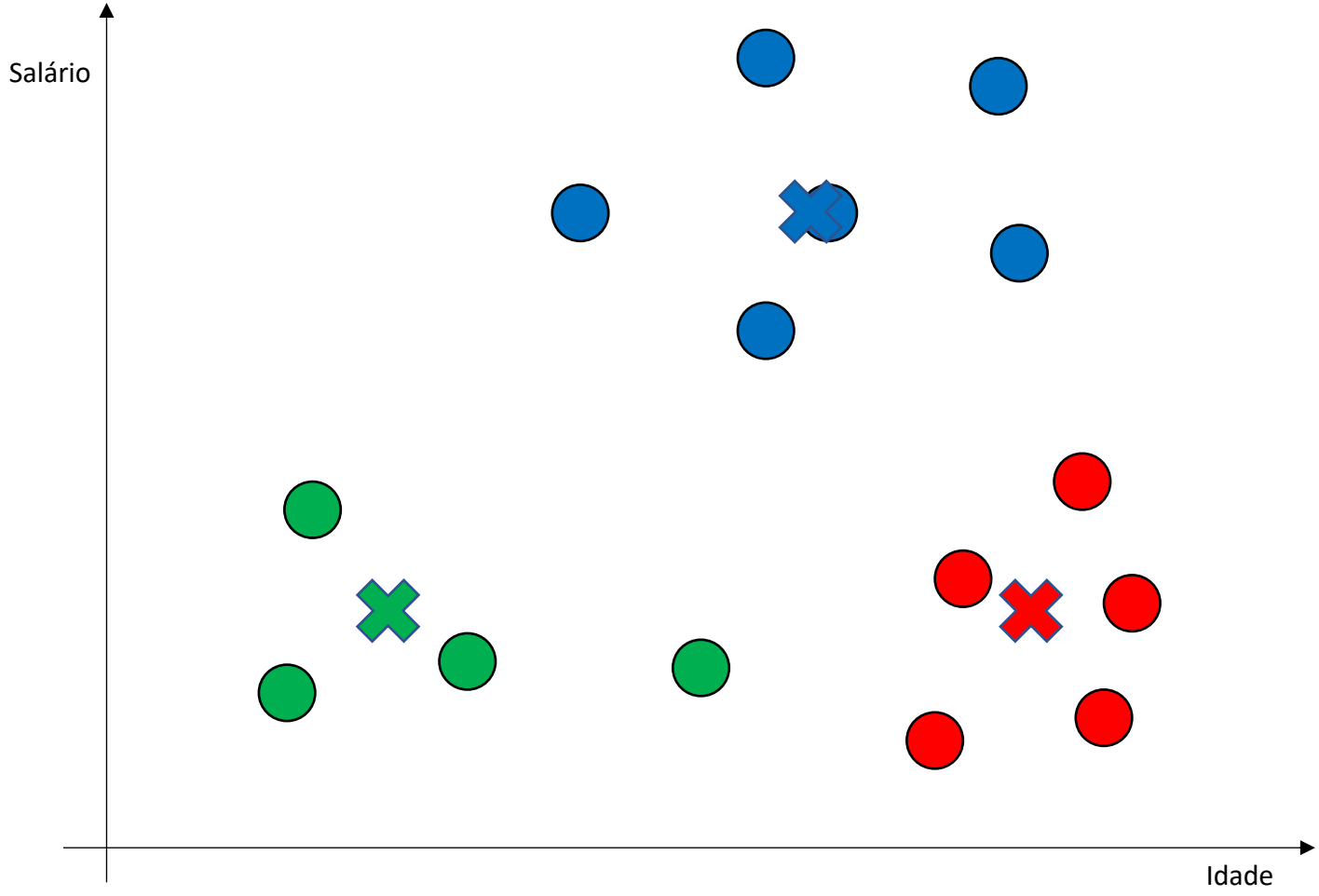
Algoritmo k-means



Algoritmo k-means



Algoritmo k-means

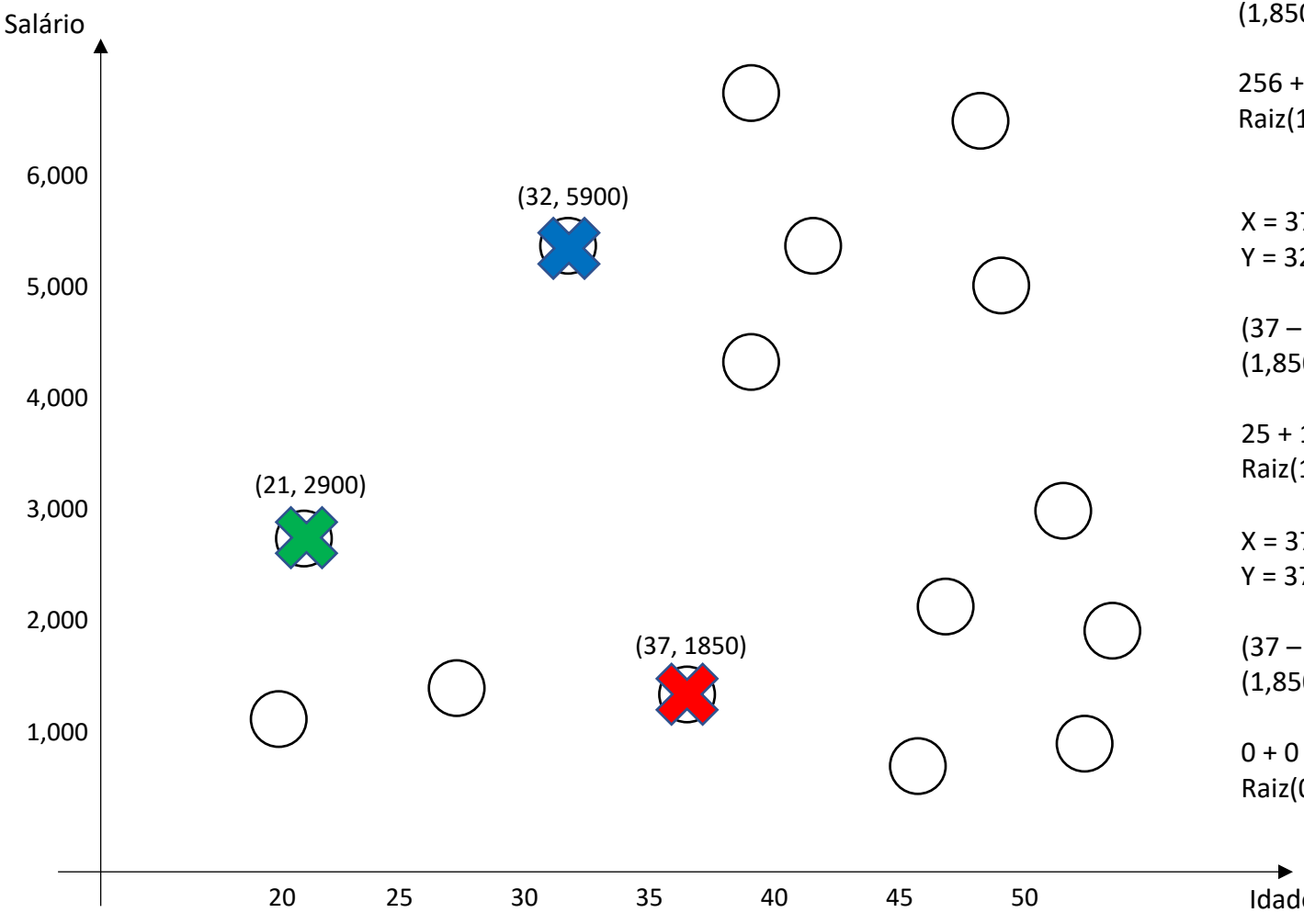


$$DE(x, y) = \sqrt{\sum_{i}^{p}(x_i - y_i)^2}$$

- x = 5, 7, 9
- y = 5, 5, 5
- Subtração de cada posição do vetor
  - 5 – 5 = 0
  - 7 – 5 = 2
  - 9 – 5 = 4
- Elevação ao quadrado
  - $0^2 = 0$
  - $2^2 = 4$
  - $4^2 = 16$
- Somatório
  - 0 + 4 + 16 = 20
- Raiz quadrada
  - Raiz(20) = 4,47

- **Distância Euclidiana = 4,47**

Algoritmo k-means



X = 37 1,850
Y = 21 2,900

$(37 - 21)^2 = 256$
$(1,850 - 2,900)^2 = 110250$

256 + 110250 = 1102756
Raiz(1102756) = **1050,12**

X = 37 1,850
Y = 32 5,900

$(37 - 32)^2 = 25$
$(1,850 - 5,900)^2 = 16402525$

25 + 16402500 = 1102756
Raiz(1102756) = **4050,00**

X = 37 1,850
Y = 37 1,850

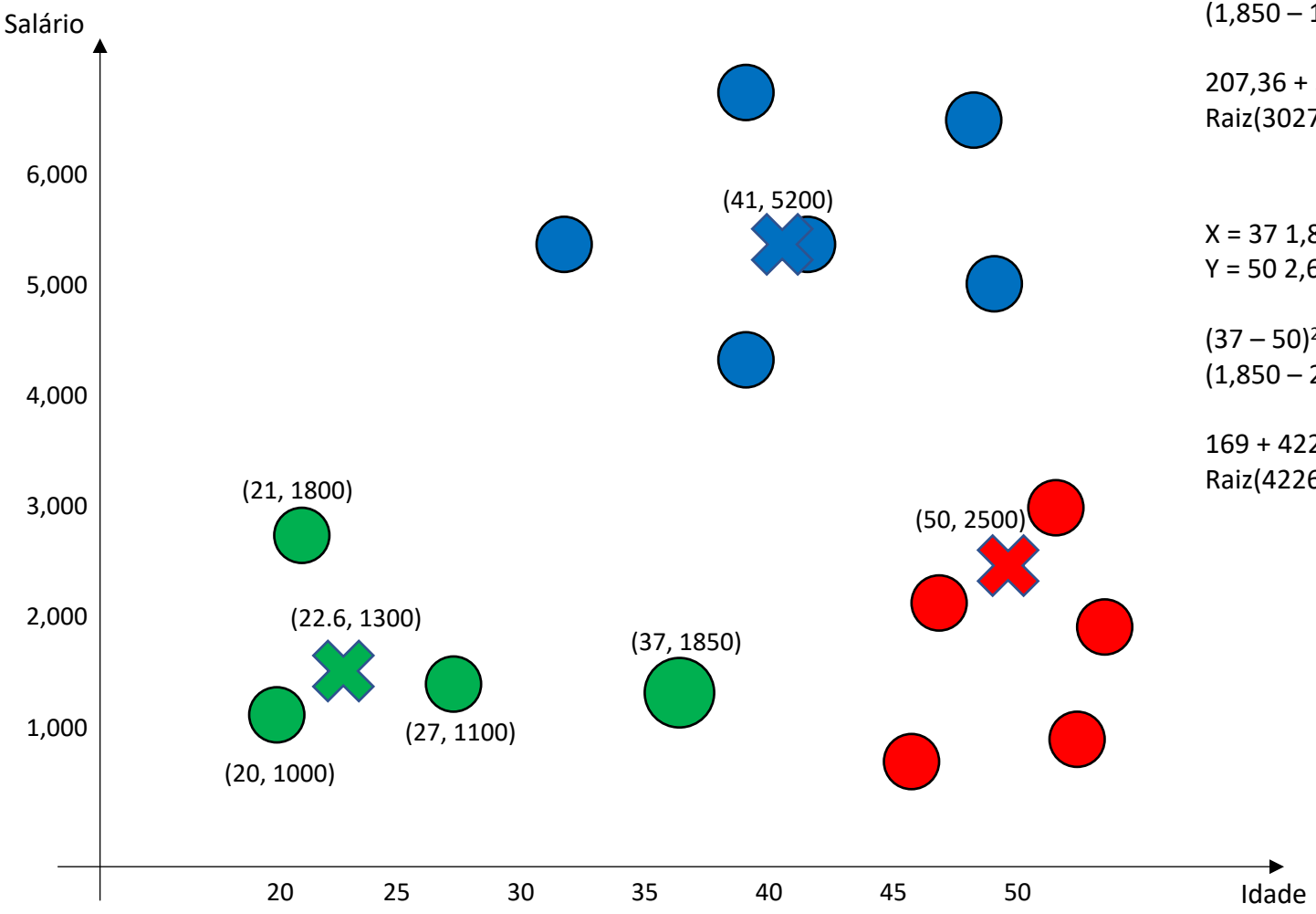
$(37 - 37)^2 = 0$
$(1,850 - 1,850)^2 = 0$

0 + 0 = 0
Raiz(0) = **0**

Algoritmo k-means

Idade = (21 + 20 + 27) / 3 = 22,6

Salário = (1800 + 1000 + 1100) / 3 = 1300



X = 37 1,850
Y = 22.6 1,300

$(37 – 22.6)^2 = 207,36$
$(1,850 – 1,300)^2 = 302500$

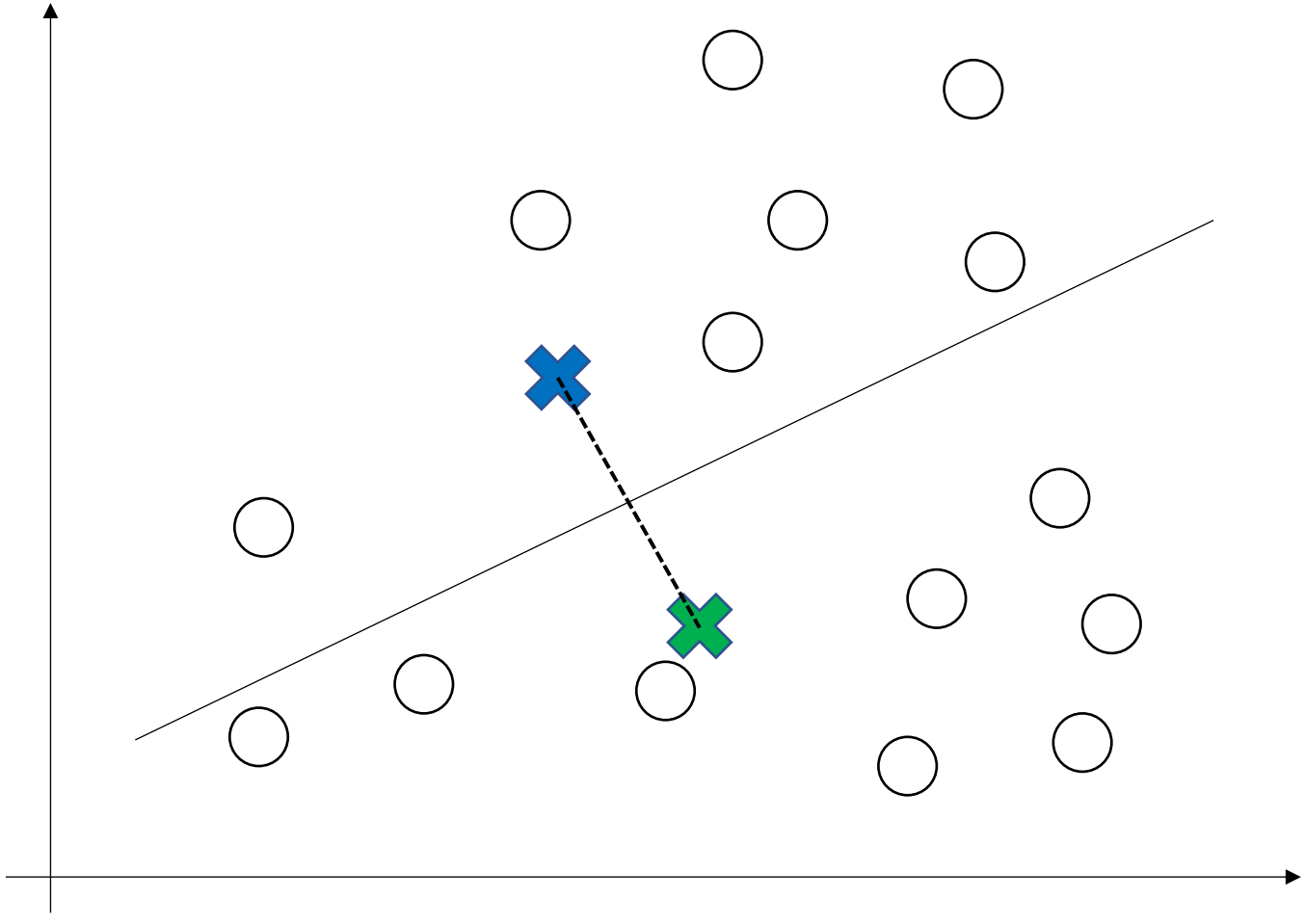
207,36 + 302500 = 302707,36
Raiz(302707,36) = **550,18**

X = 37 1,850
Y = 50 2,600

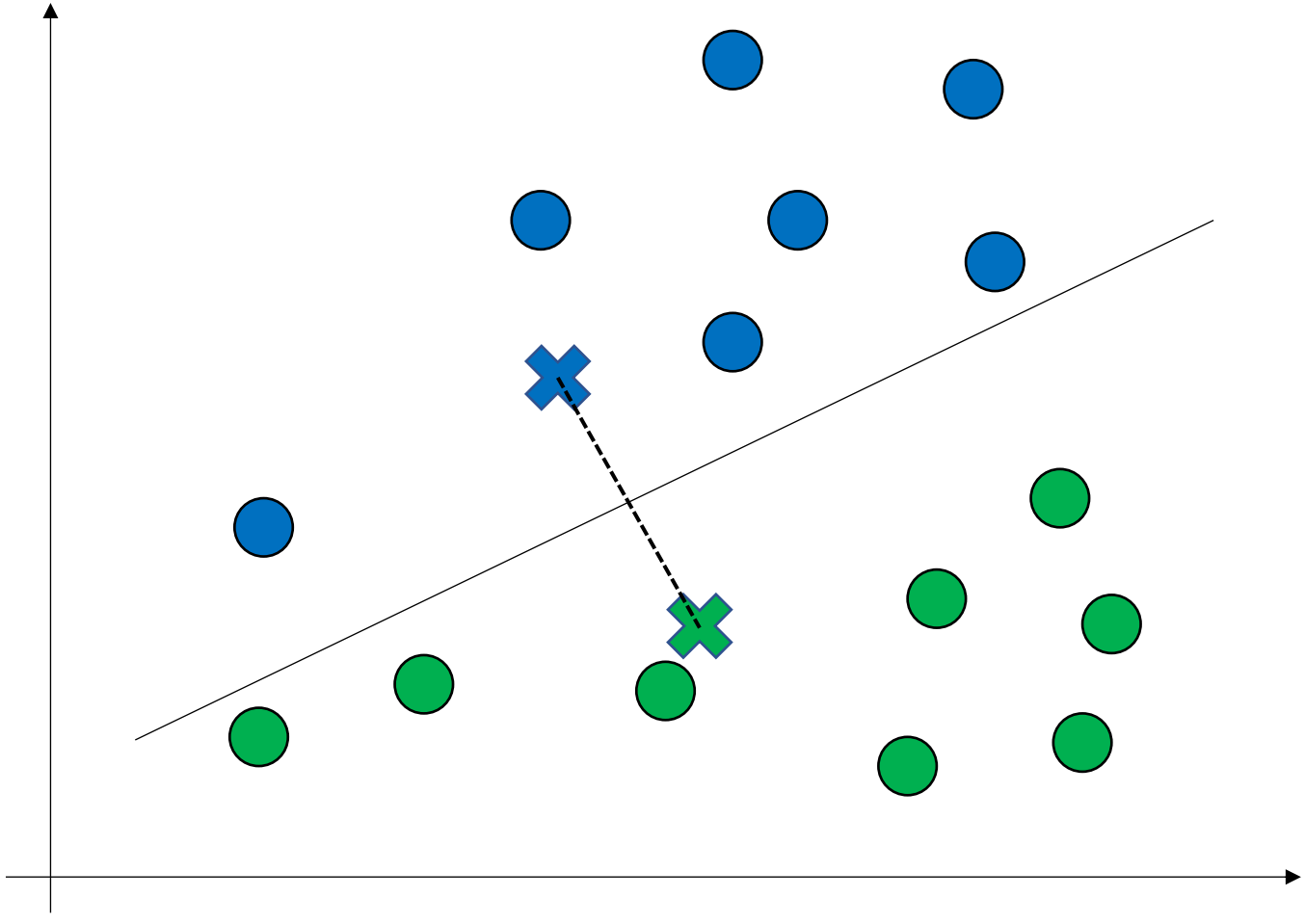
$(37 – 50)^2 = 169$
$(1,850 – 2,500)^2 = 422500$

169 + 422500 = 422669
Raiz(422669) = **650,12**

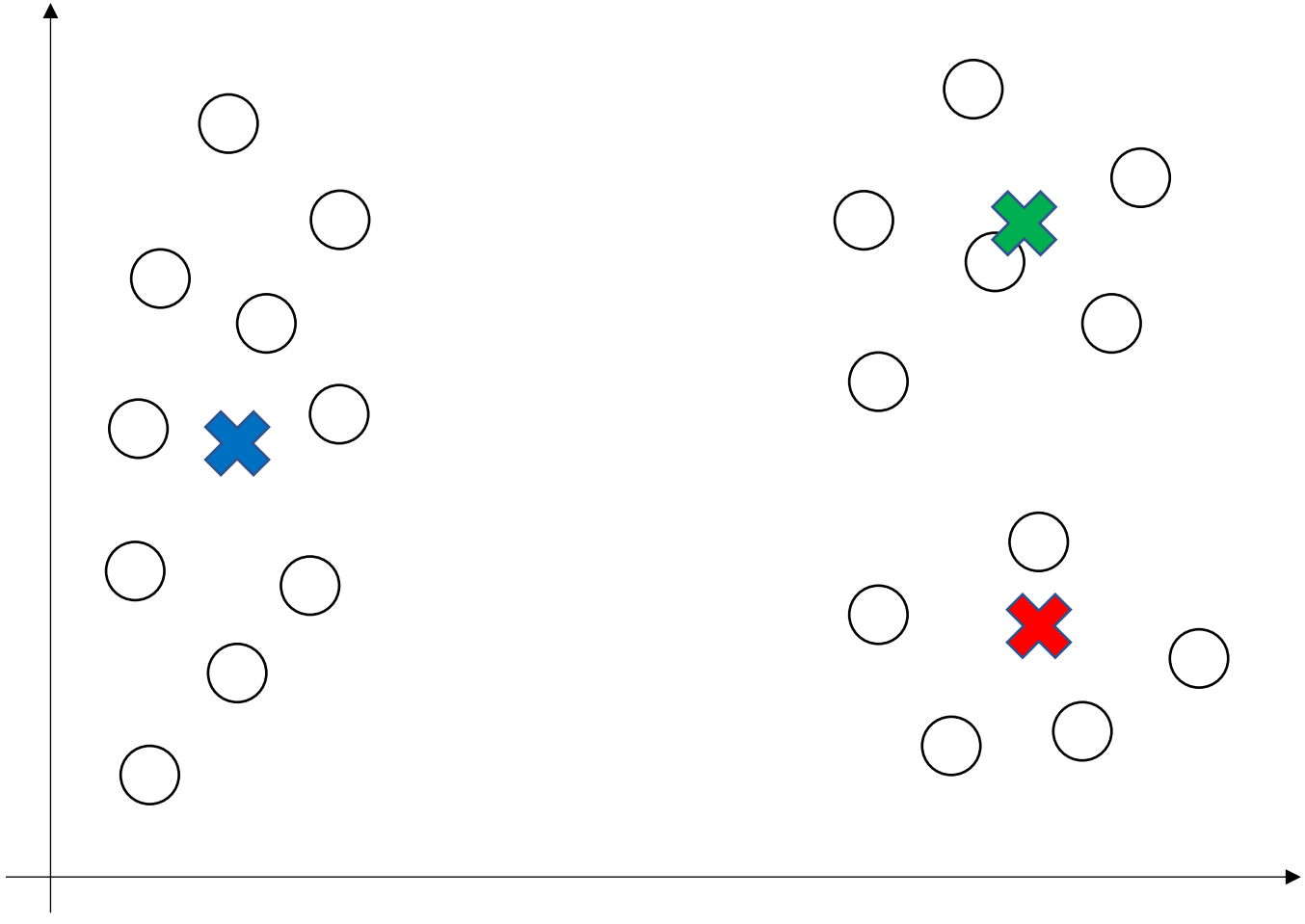
Algoritmo k-means



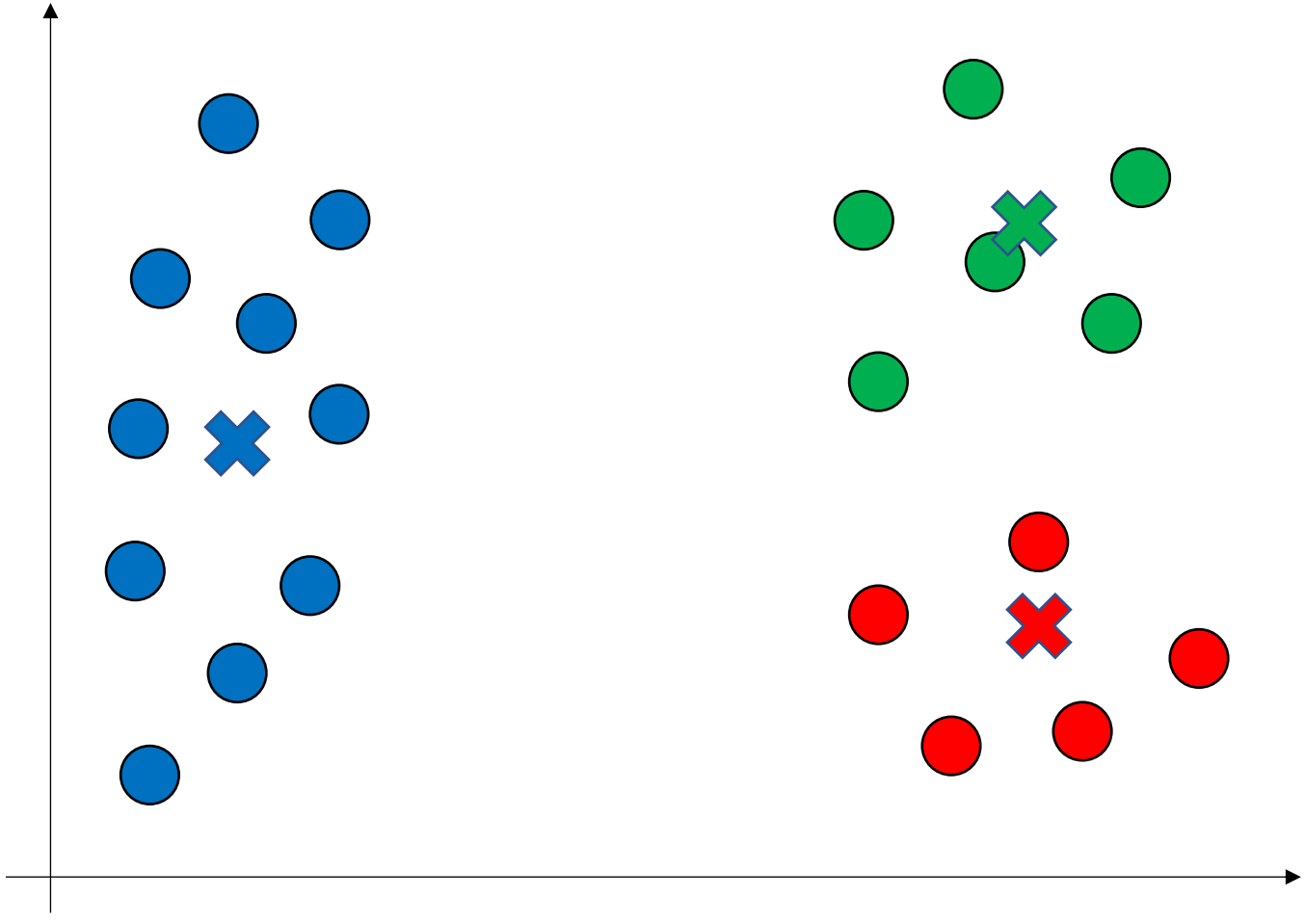
Algoritmo k-means



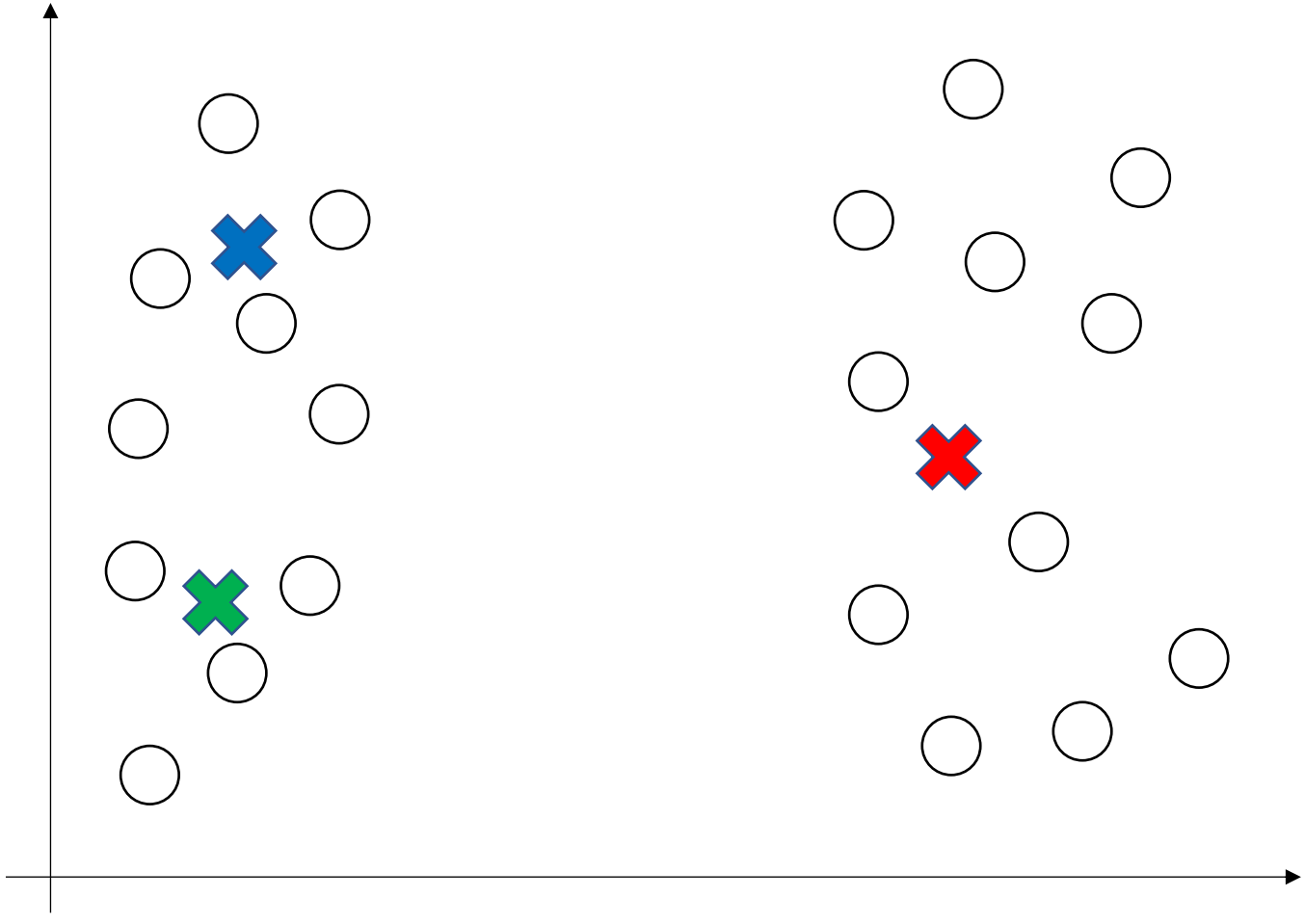
Algoritmo k-means



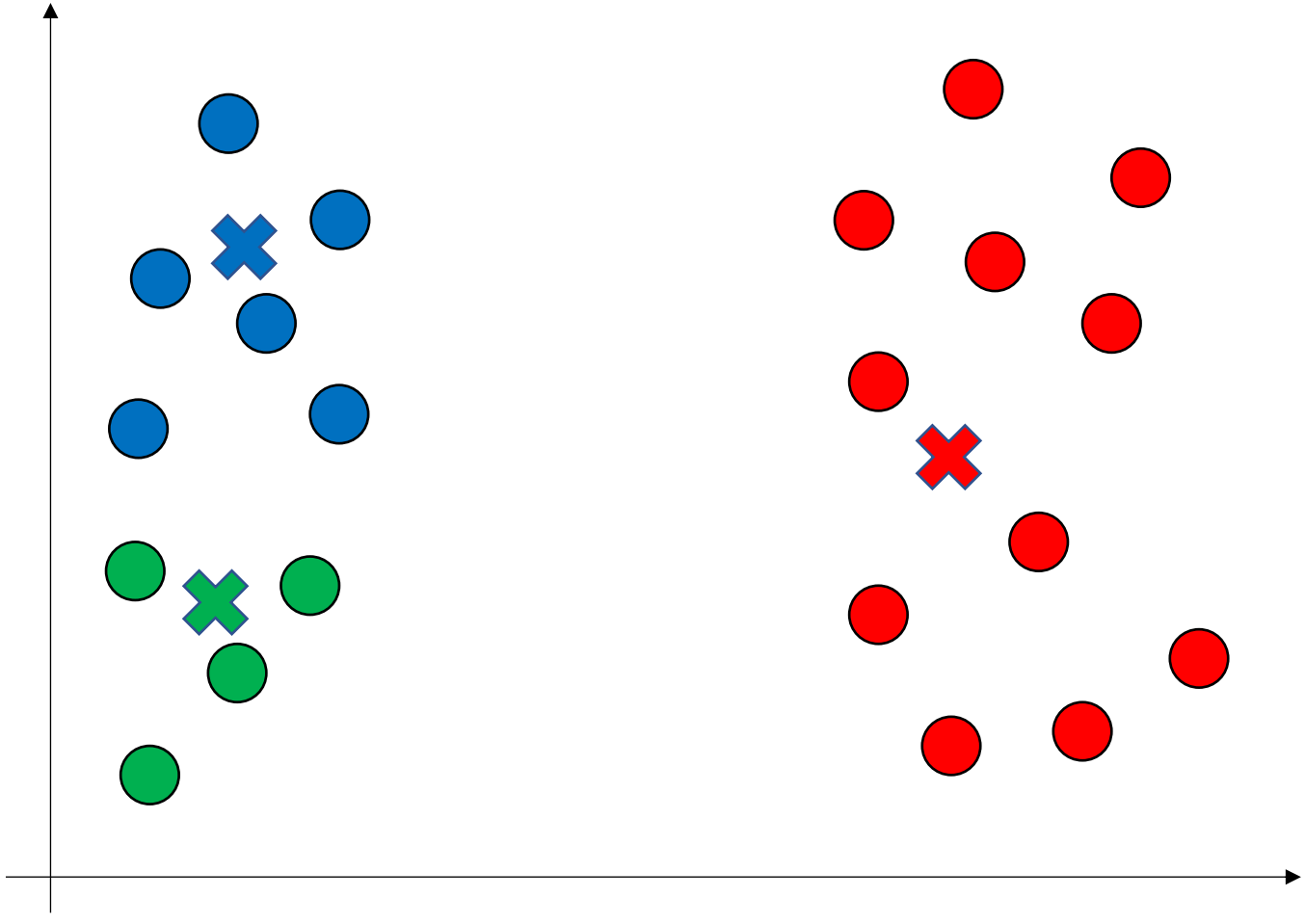
Algoritmo k-means



Algoritmo k-means

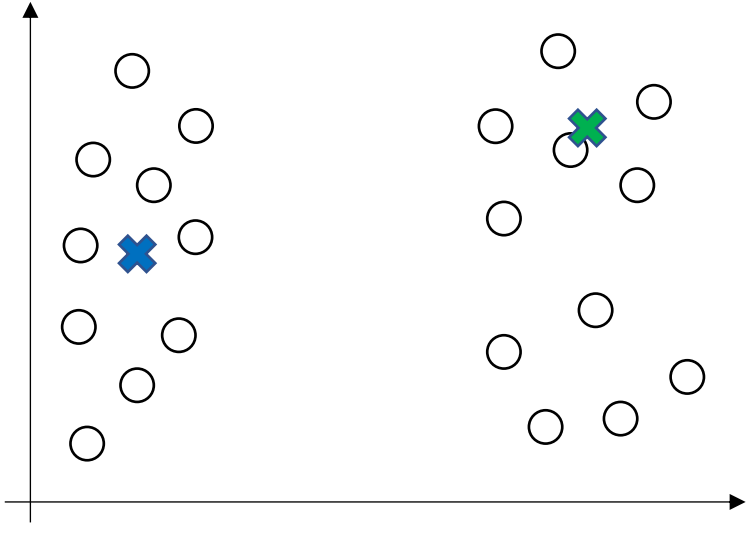


Algoritmo k-means



K-means++

- Reduz a probabilidade de inicializações ruins
- Seleciona os centroides iniciais que estão longes uns dos outros
- O primeiro centroide é selecionado randomicamente, porém, os outros são selecionados baseado na distância para o primeiro ponto



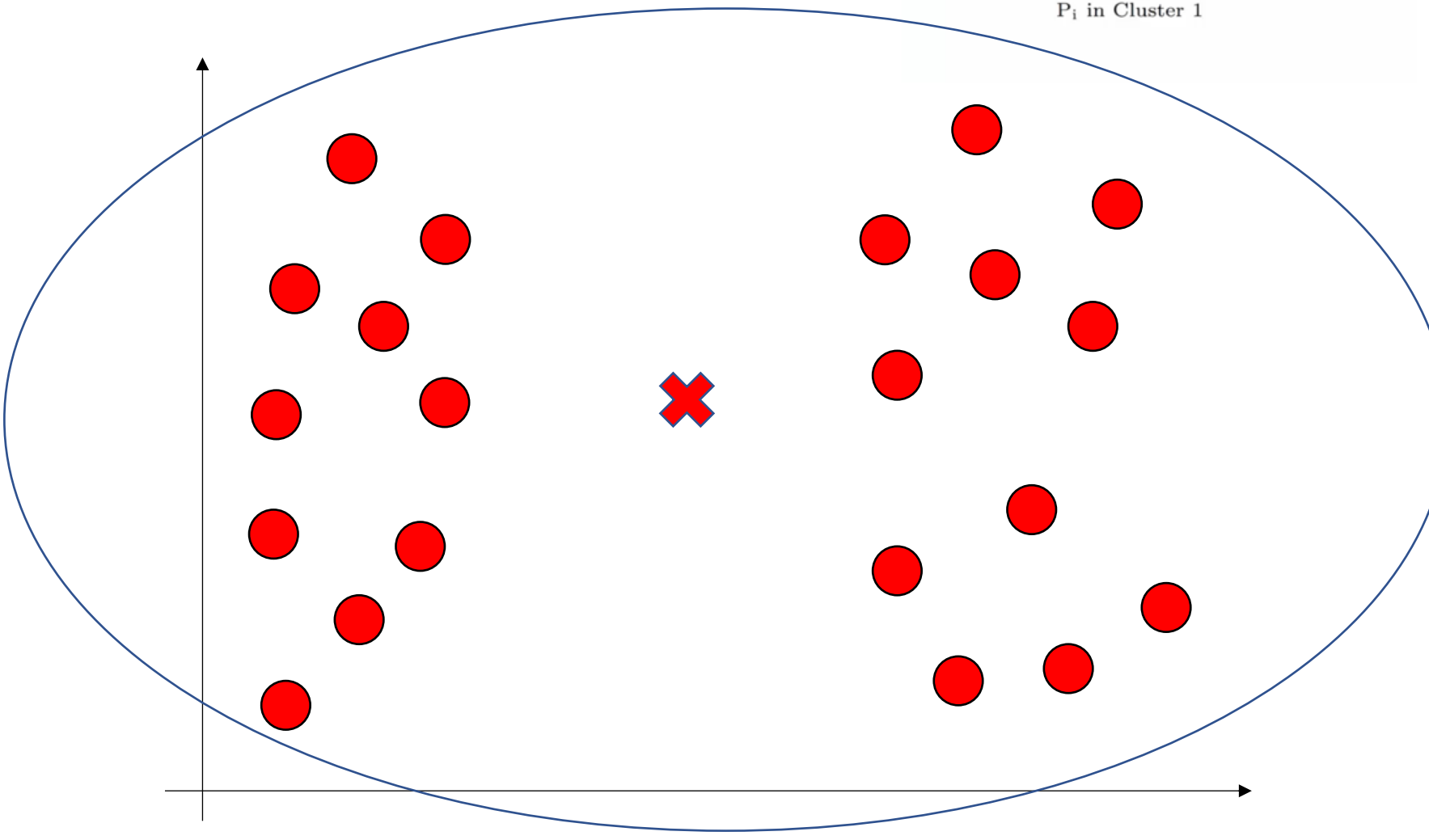
Definição do número de clusters

- Ter um conhecimento prévio de quantos grupos são necessários
- Se não tiver conhecimento prévio
  - $clusters = \sqrt{\frac{N}{2}}$
- Elbow method
  - Tenta vários valores de k
- Não existe garantia para encontrar o melhor conjunto de clusters

Elbow method

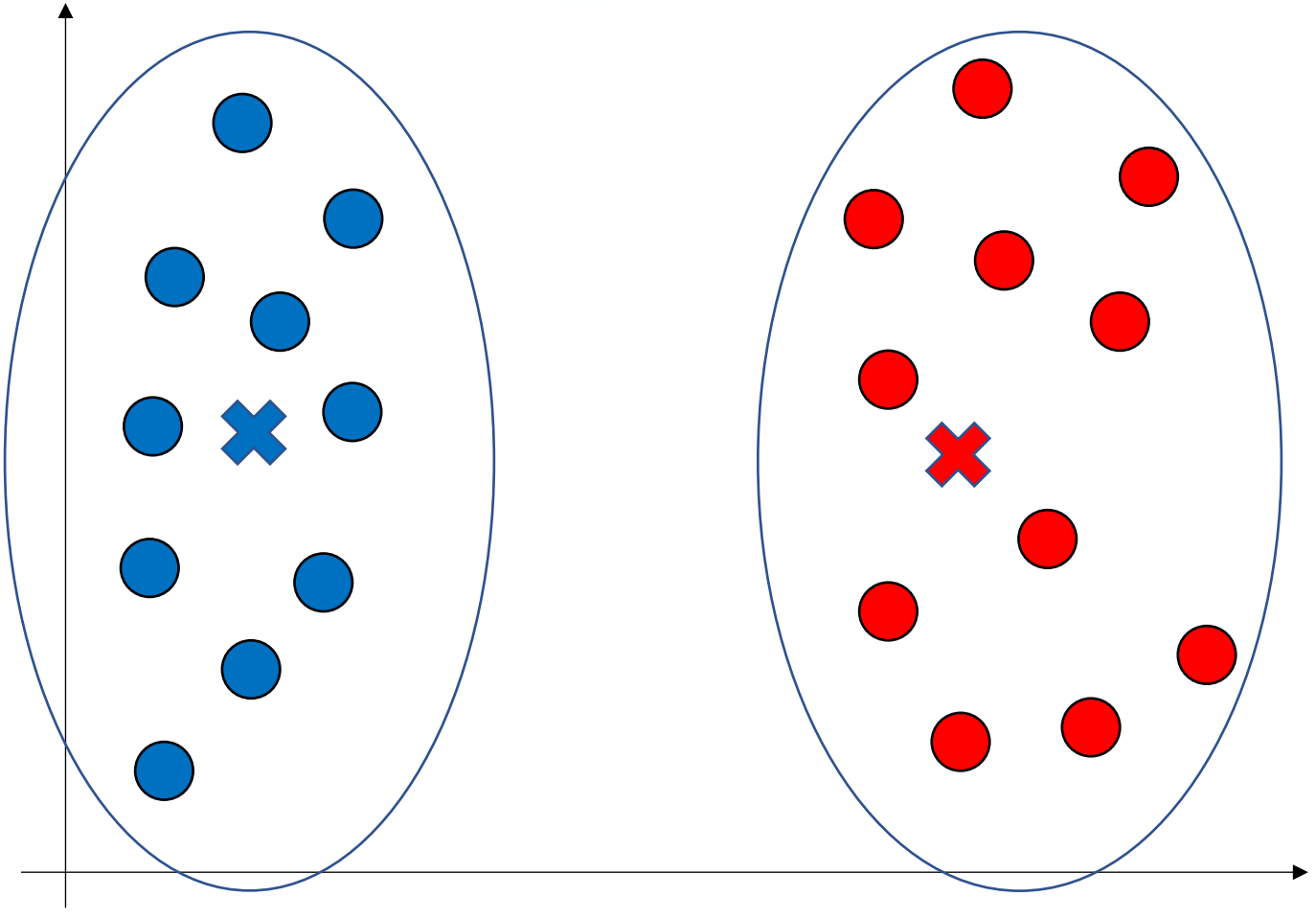
Within-cluster sum of squares

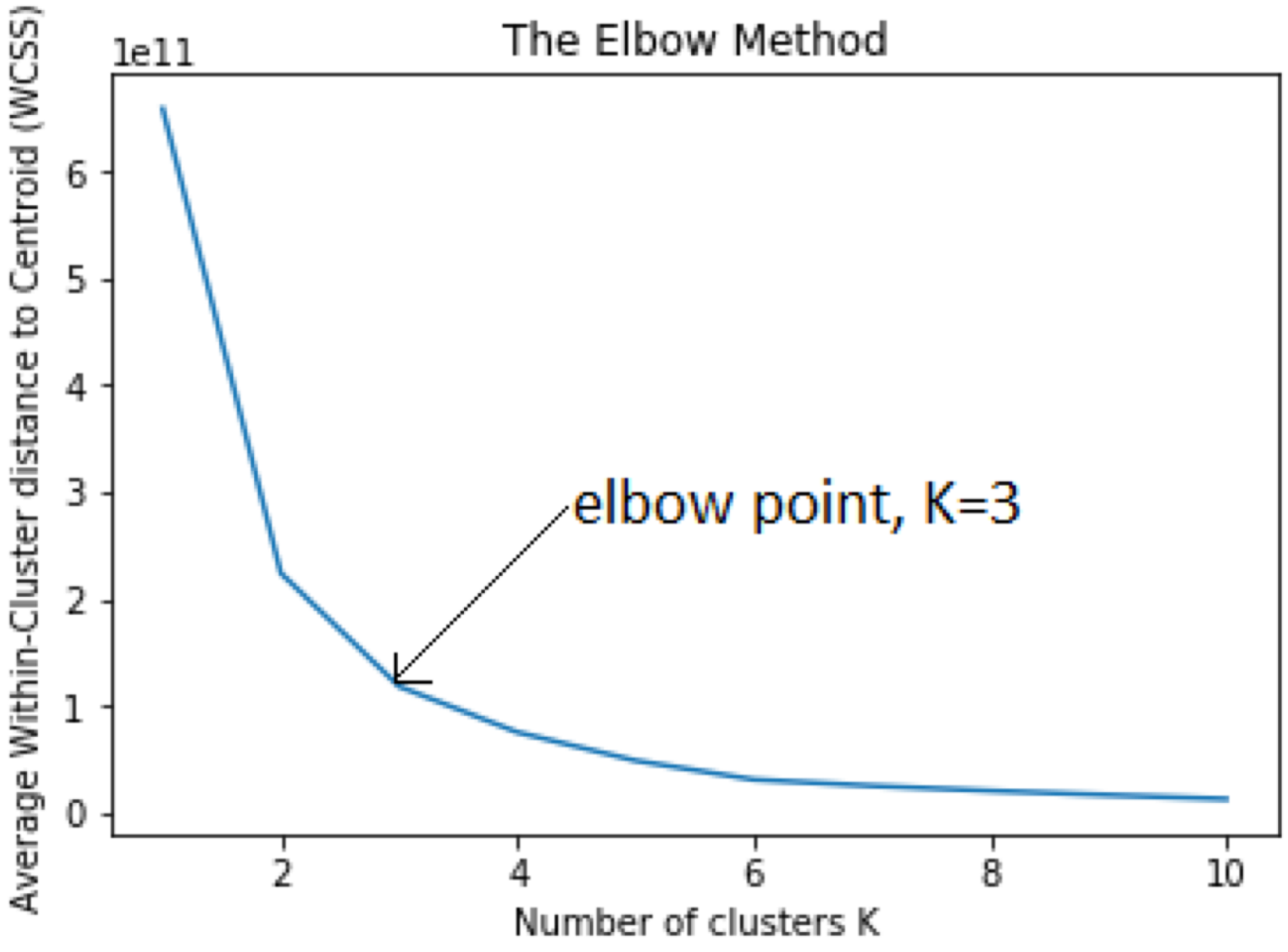
$$\mathrm{WCSS} = \sum_{P_i \text{ in Cluster } 1} \mathrm{distance}(P_i, C_1)^2$$



Elbow method

$$\text{WCSS} = \sum_{P_i \text{ in Cluster 1}} \text{distance}(P_i, C_1)^2 + \sum_{P_i \text{ in Cluster 2}} \text{distance}(P_i, C_2)^2$$





Fonte: http://www.patterns7tech.com/customer-segmentation-using-machine-learning-k-means-clustering/

Conclusão