AGRUPAMENTO

MANOELA KOHLER

Prof.manoela@ica.ele.puc-rio.br

TÓPICOS

R

Análise exploratória

Pré-processamento

- Balanceamento
- *Outliers*
- *Missing values*
- Normalização
- Seleção de atributos (Filtros, *Wrappers*, PCA)

Associação:

- *Apriori*
- *FP-Growth*
- *Eclat*

Classificação:

- Regressão logística
- *Support Vector Machine* (SVM)
- Árvores de Decisão
- *Random Forest*
- ~~Redes Neurais~~
- *K nearest neighbors*

Regressão

- Regressão linear simples
- Regressão linear múltipla
- Regressão não linear simples
- Regressão não linear múltipla

Agrupamento

- Particionamento (K-means, K-medoids)
- Hierárquico (DIANA, AGNES)
- Densidade (DBSCAN)

Séries Temporais

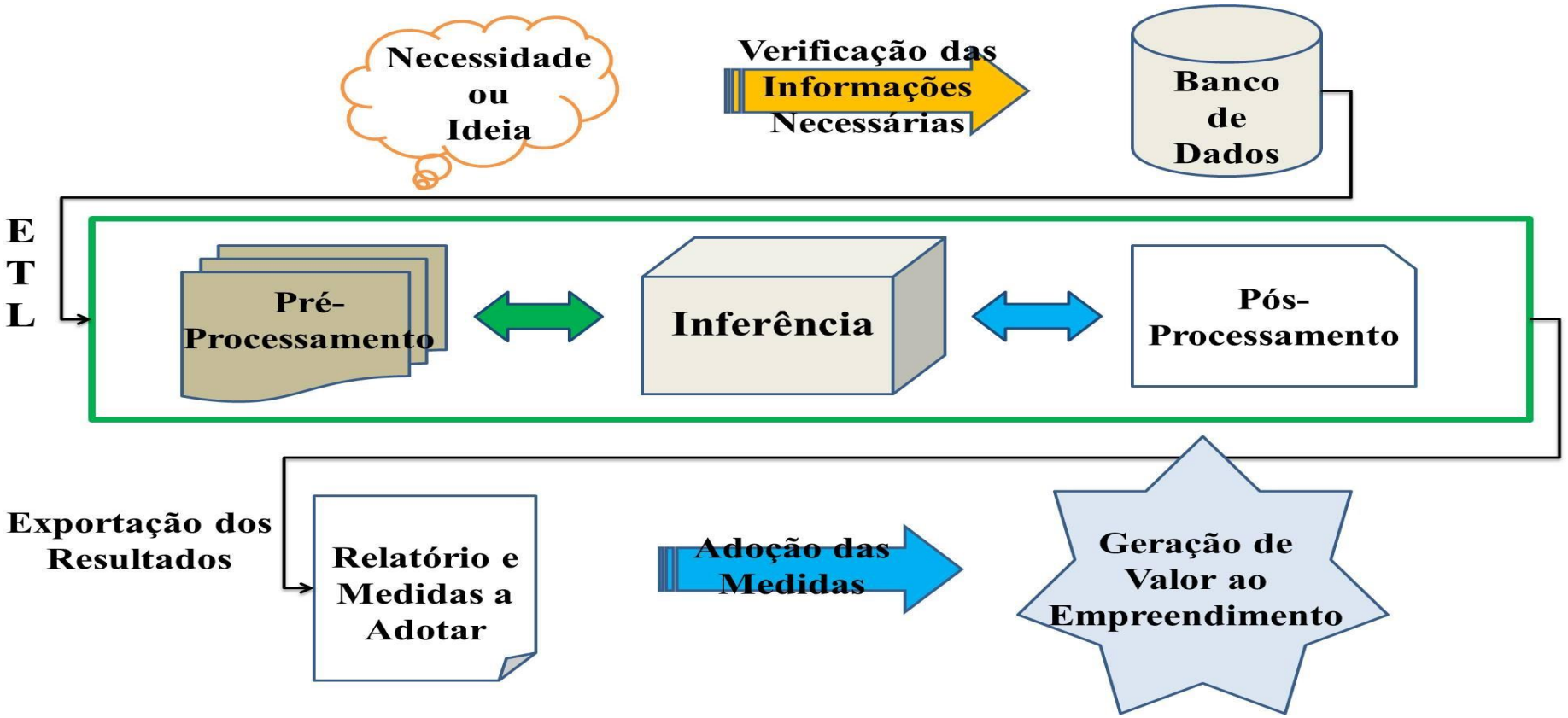
- Naive
- Média Móvel
- Amortecimento exponencial
- Auto-regressivo integrados de média móvel
- Auto regressivo não linear

BI Master

Recapitulação

ETAPAS DE UM PROJETO DE DATA MINING

ESQUEMA BÁSICO DE UM PROJETO DE DM



ASSOCIAÇÃO


Heineken

Apriori

FP-Growth

ECLAT

ESTUDO DE CASO

Transações de um Supermercado

Lista de transações (compras) em um mercado francês:

- Cada linha da base é uma transação;
- Cada transação tem de 1 a N itens;
- Existem 119 produtos diferentes no mercado;
- Base tem 7501 transações feitas no decorrer de 1 mês.

Aprendizado não supervisionado: Agrupamento

- K-means (Clusterização baseada em Particionamento)
- Clusterização Hierárquica
- Clusterização baseada em Densidade

AGRUPAMENTO (CLUSTERIZAÇÃO)

CLUSTERIZAÇÃO

Cluster: uma coleção de objetos

- Similares aos objetos do mesmo cluster
- Dissimilares aos objetos de outros clusters

Clusterização

- Agrupamento de conjuntos de dados em clusters.

Clusterização é uma classificação não supervisionada: sem classes predefinidas.

A NOÇÃO DE UM CLUSTER PODE SER AMBÍGUA

Quantos clusters?

Seis Clusters

Dois Clusters

Quatro Clusters

APLICAÇÕES GERAIS DE CLUSTERIZAÇÃO

Marketing: identifica grupos distintos de clientes (útil para desenvolver programas de marketing) (CHIANG, 2003).

Uso da terra: identifica a possibilidade de alocação de uso da terra para fins agrários e/ou urbanos em uma base de dados de observação via satélite de todo o planeta Terra (LEVIA JR, 2000).

Seguro: Identifica grupos de clientes que fazem comunicação de sinistro com alta frequência (YEOH, 2001).

Planejamento (cidade): Identifica grupos de casas de acordo com o tipo, valor e localização geográfica.

O QUE É UMA BOA CLUSTERIZAÇÃO?

Uma boa clusterização sempre produz clusters com:

- Alta similaridade nas classes;
- Baixa similaridade entre as classes.

A qualidade dos resultados depende do(a):

- Medida de similaridade usada;
- Método e sua implementação.

ANTES DE SEGUIR EM FRENTE VAMOS LEMBRAR

CLASSIFICAÇÃO NÃO SUPERVISIONADA

Não conhecemos o padrão, nem o número total de grupos a serem encontrados durante a classificação.

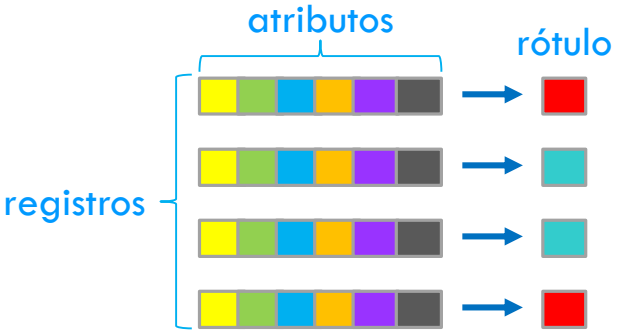
Também conhecido como aprendizado não supervisionado ou análise de agrupamentos (clusters).

O conjunto de dados é particionado em grupos, baseados em características específicas, tais que os pontos dentro de um grupo (cluster) sejam mais similares do que os pontos de outros grupos.

Machine Learning

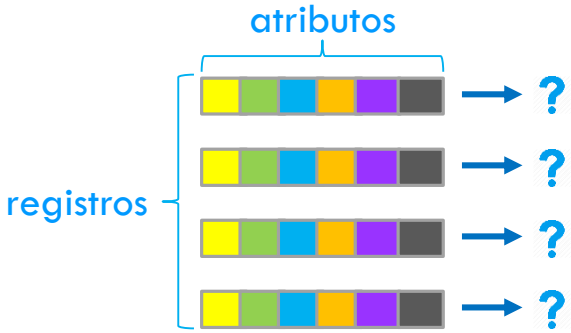
Supervisionado

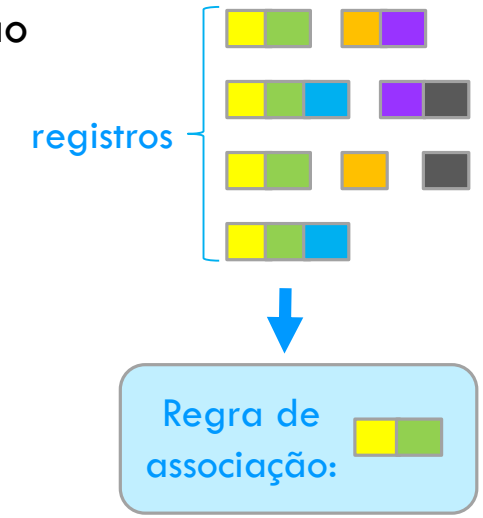
Classificação
Regressão
Previsão de séries temporais



Não Supervisionado

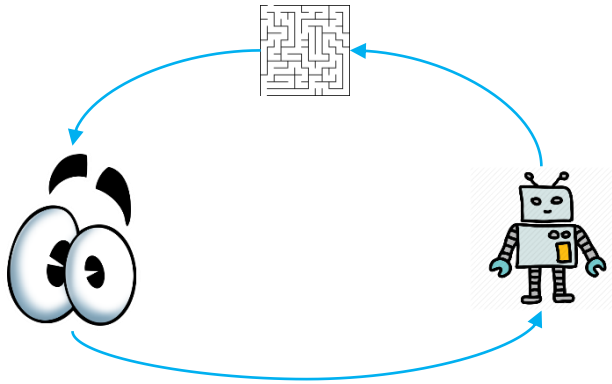
Agrupamento
Associação





Reforço

Aprendizado através da interação de agentes com um ambiente

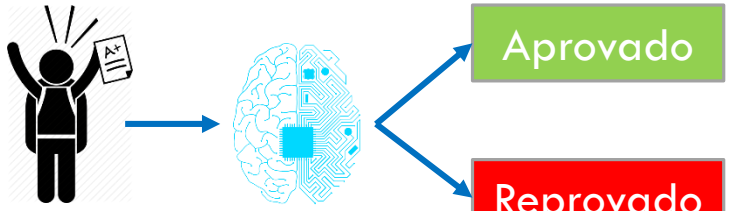


SUPERVISIONADO

- Aproximador: função mapeia entradas e saída.

Classificação

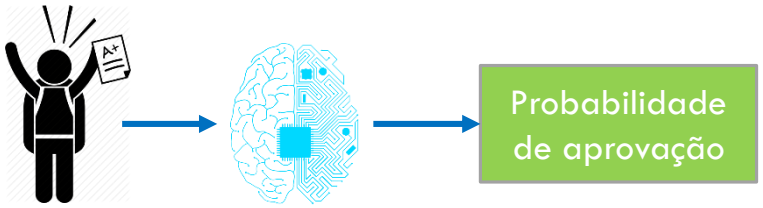
Rótulo é categórico.



Dados do estudante

Regressão

Rótulo é contínuo.



Dados do estudante

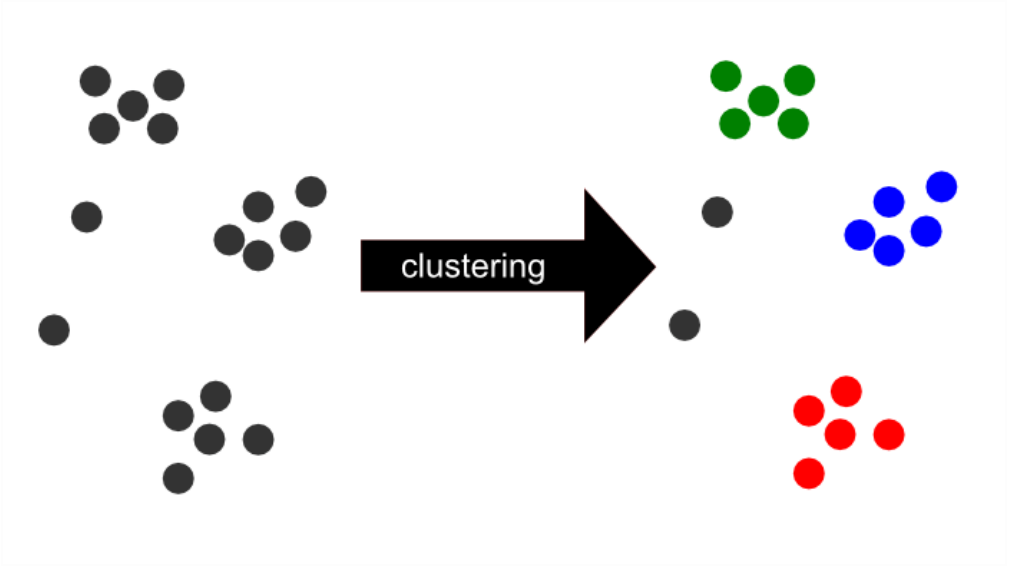
Previsão de Séries

Rótulo é contínuo e dependente do tempo.

NÃO SUPERVISIONADO

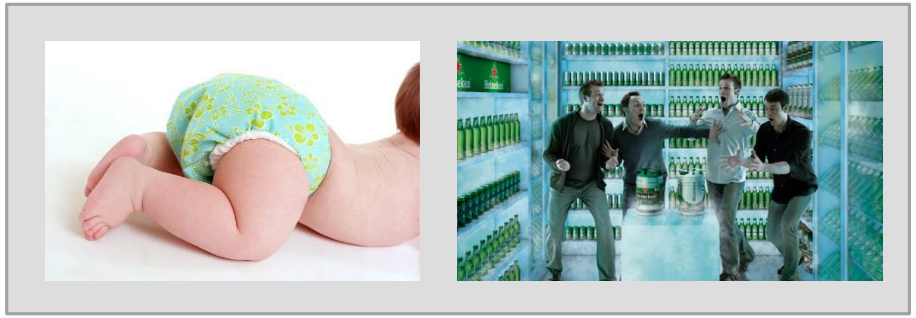
Agrupamento

Descoberta de semelhanças e grupos entre registros.



Associação

Descoberta de relações entre variáveis.



MÉTODOS DE CLUSTERIZAÇÃO

- **Particionamento:** Constrói várias partições e as avalia usando algum critério.
- **Hierárquico:** Cria uma decomposição hierárquica dos objetos usando algum critério.
- **Baseado em densidade:** Fundamenta-se em funções de conectividade e de densidade.

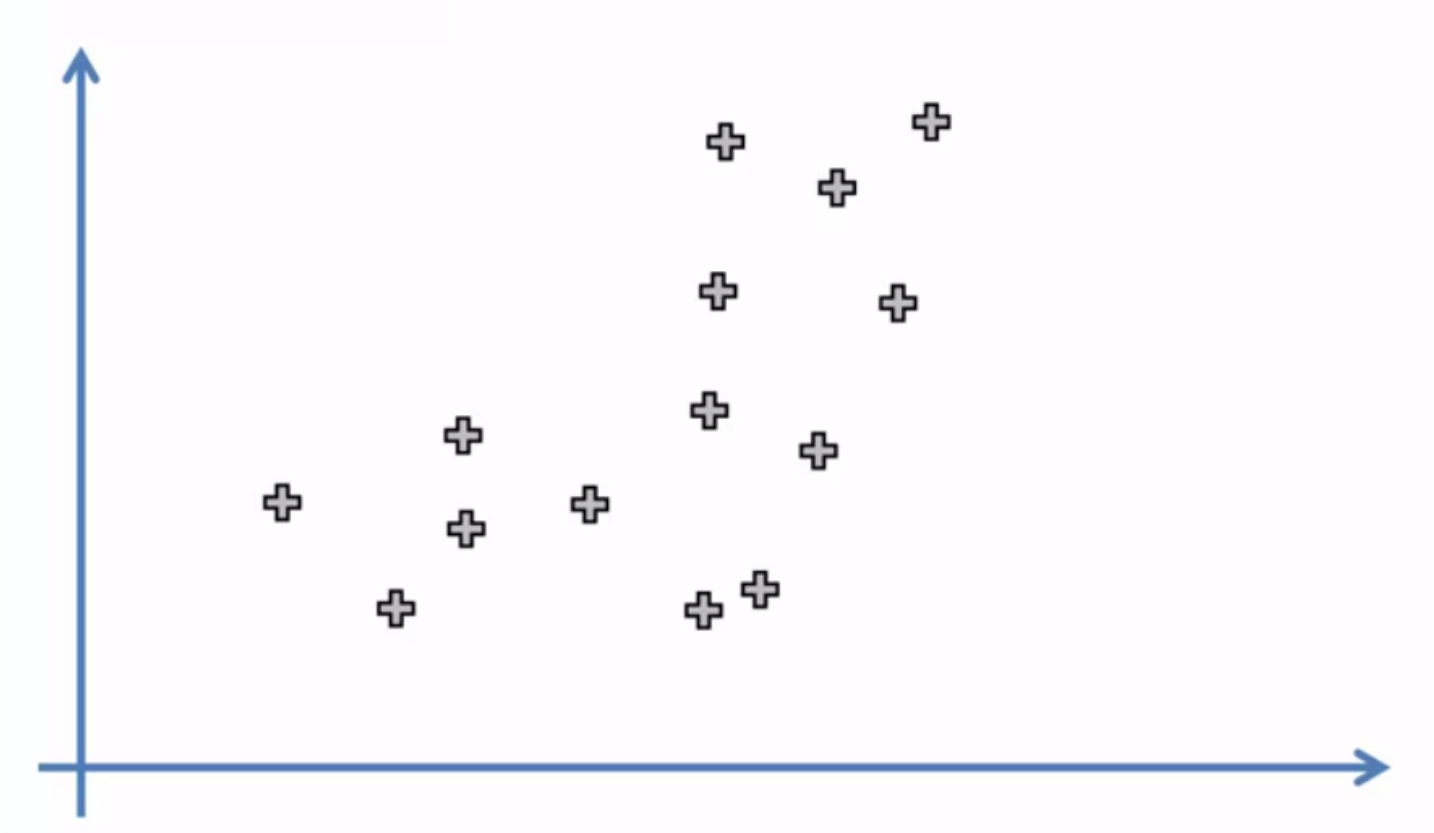
MÉTODOS BASEADOS EM PARTICIONAMENTO

Dado um valor de k, encontrar k clusters que otimizem um critério de particionamento escolhido:

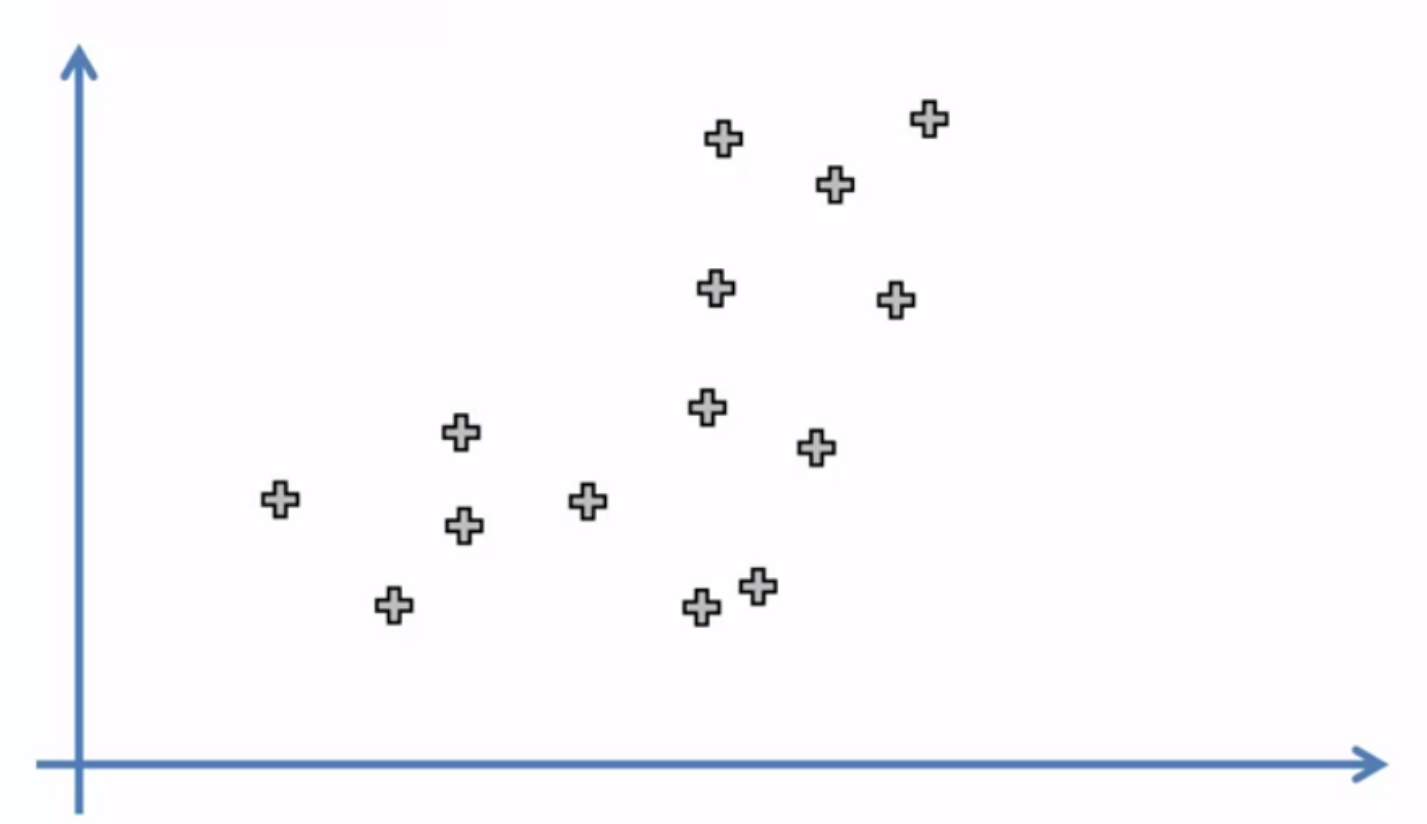
- Principais: algoritmos k-means e k-Medoids;
- K-means (MacQueen'67): Cada cluster é representado pelo centro (centroide) do cluster;
- K-medoids ou PAM (Partition Around Medoids) (Kaufman & Rousseeuw'87): Cada cluster é representado por um dos objetos no cluster.

K-MEANS

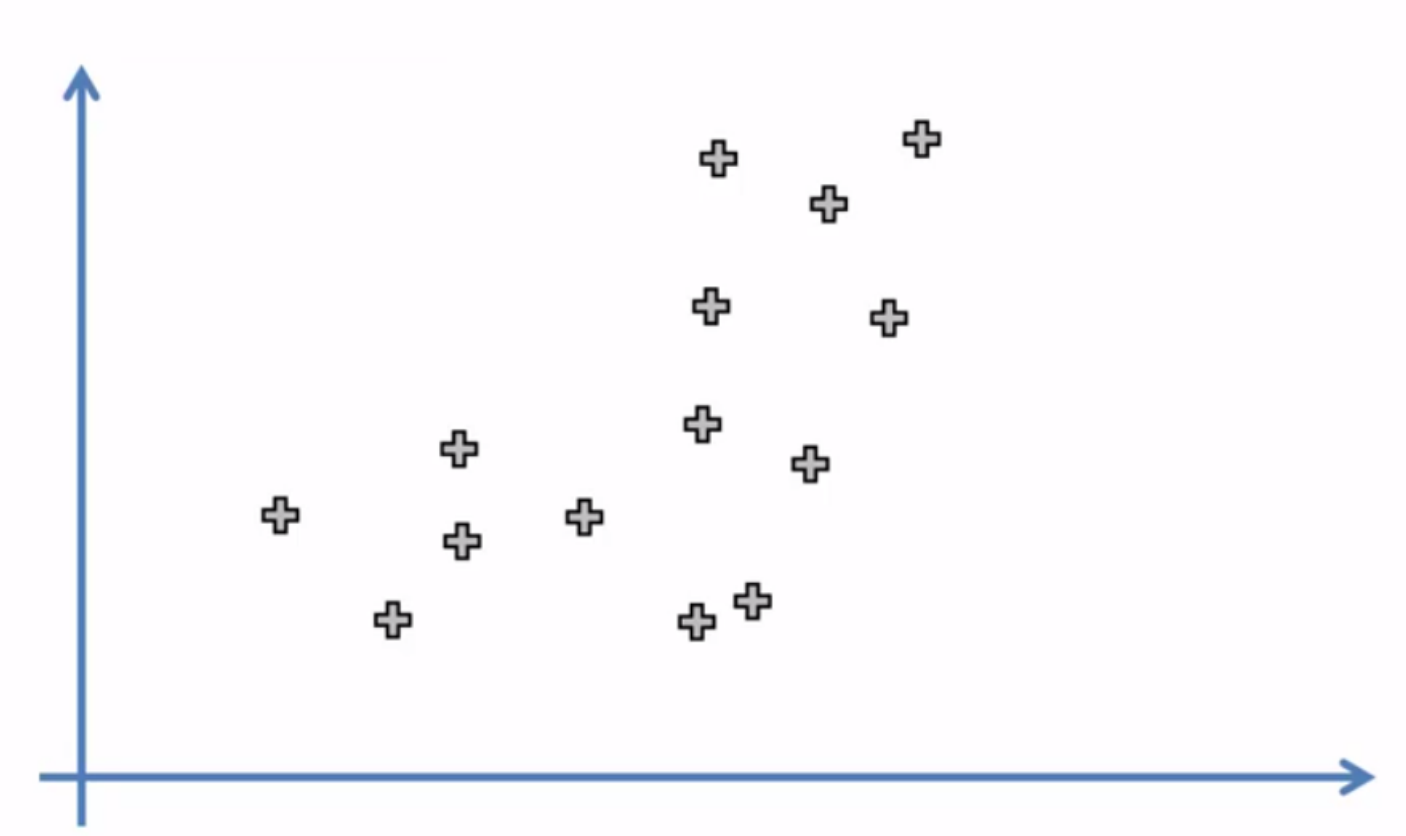
K-MEANS
PASSO 1: ESCOLHER O NÚMERO DE CLUSTERS



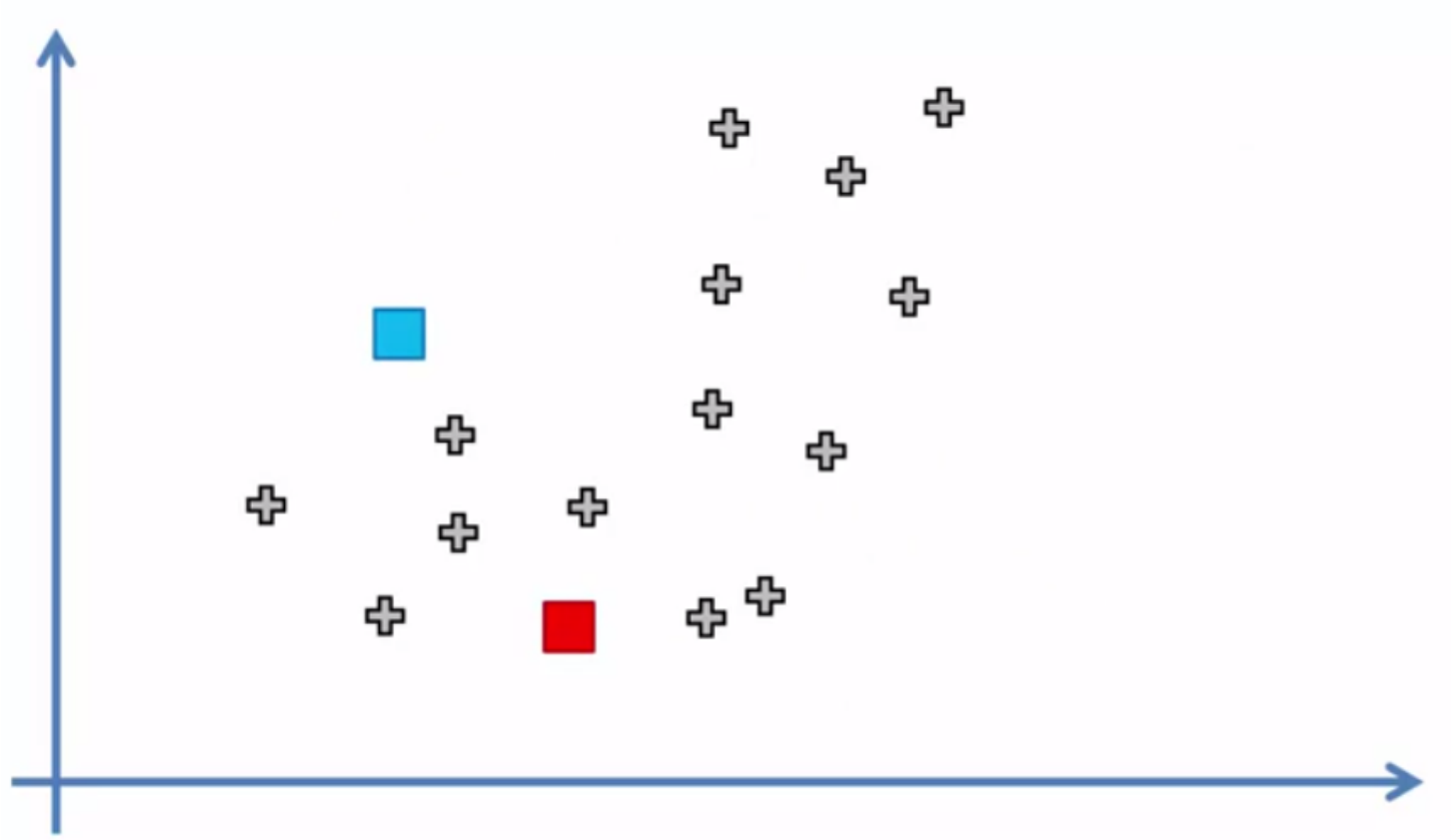
PASSO 1: K = 2



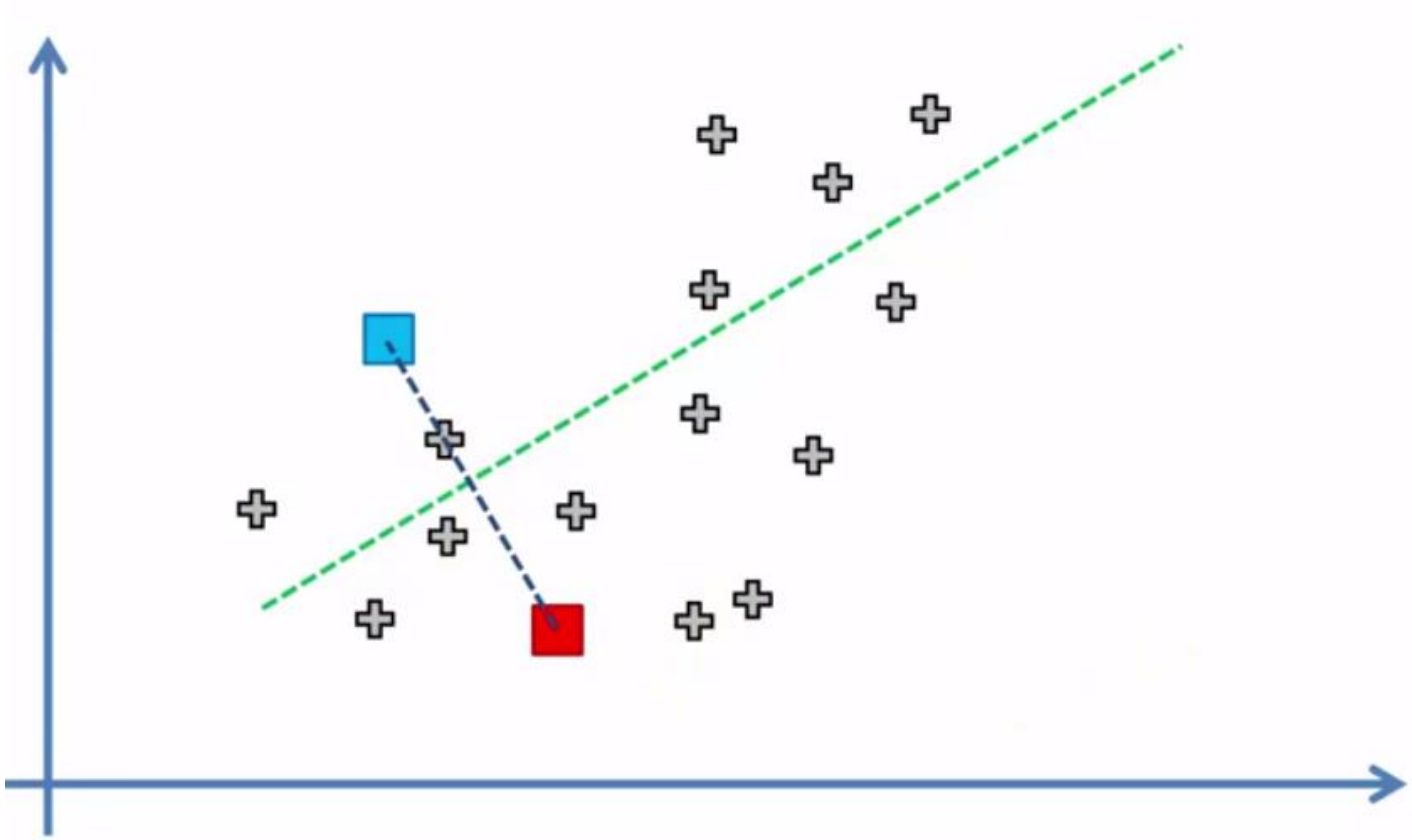
PASSO 2: SELECIONAR ARBITRARIAMENTE K PONTOS COMO OS CENTROIDES INICIAIS (NÃO NECESSARIAMENTE DA BASE DE DADOS)



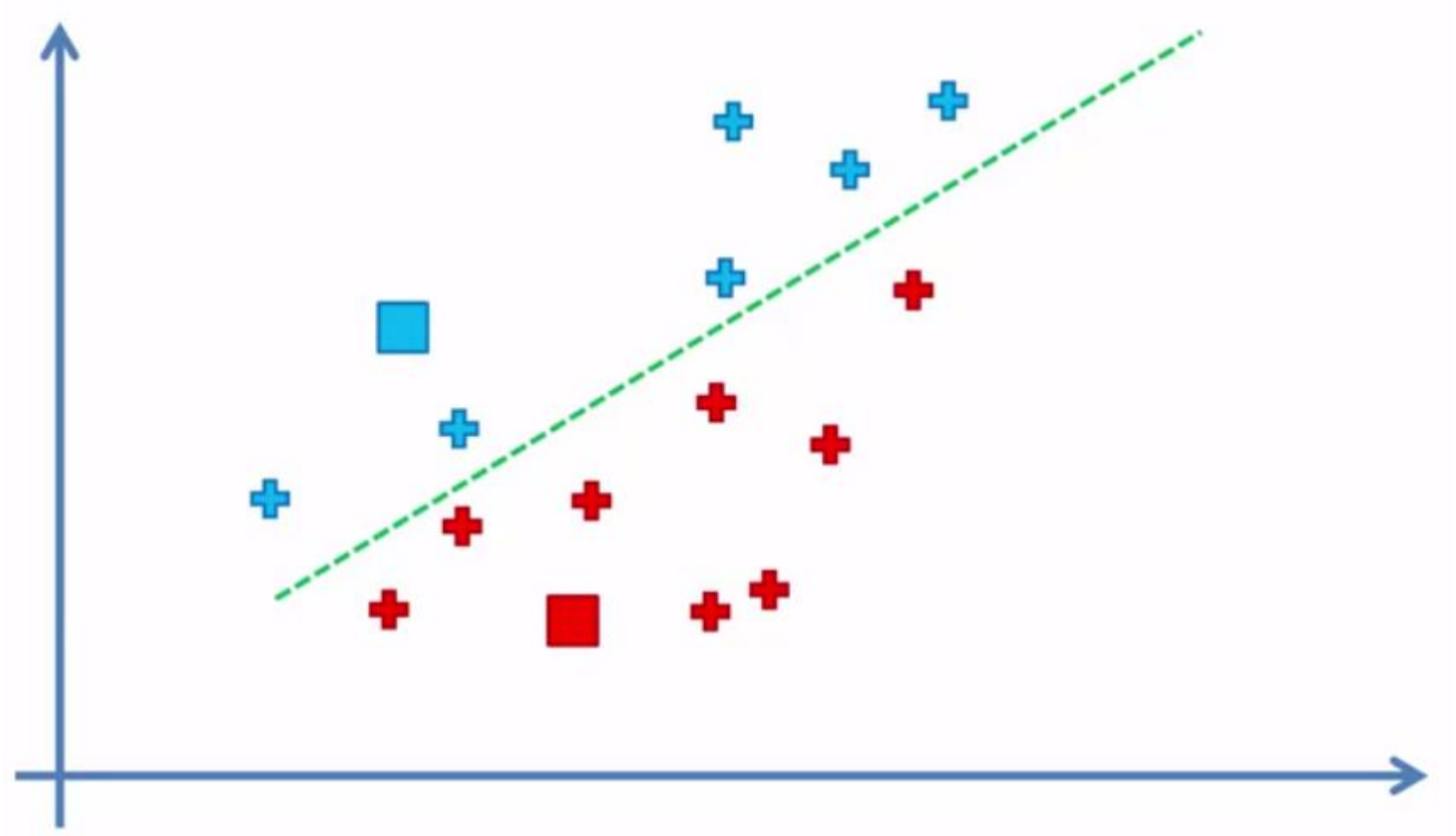
PASSO 2: SELECIONAR ARBITRARIAMENTE K PONTOS COMO OS CENTROIDES INICIAIS (NÃO NECESSARIAMENTE DA BASE DE DADOS)



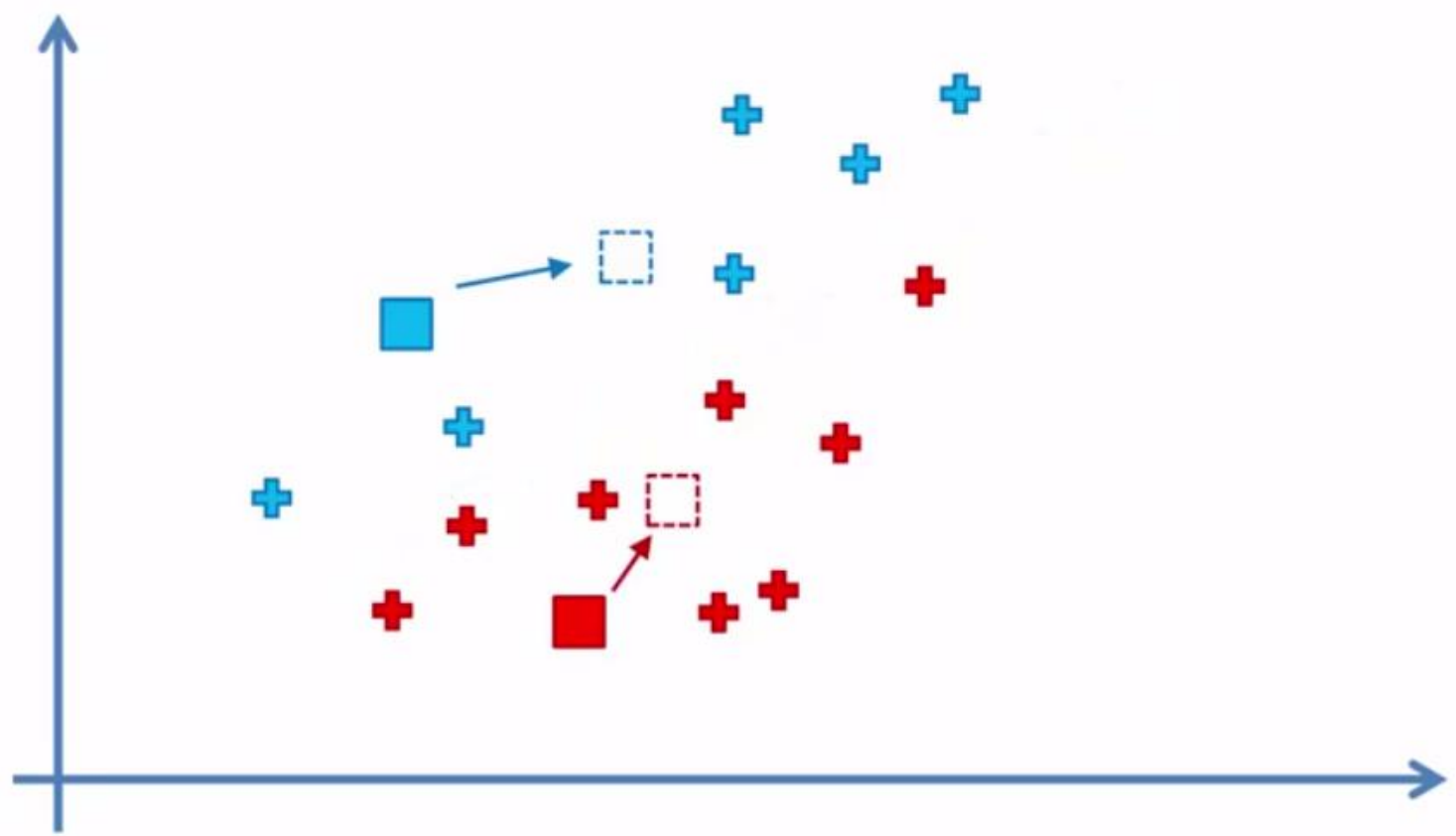
PASSO 3: ASSOCIAR CADA OBJETO AO CLUSTER (CENTROIDE) **MAIS PRÓXIMO** (MAIOR SIMILARIDADE), FORMANDO K CLUSTERS



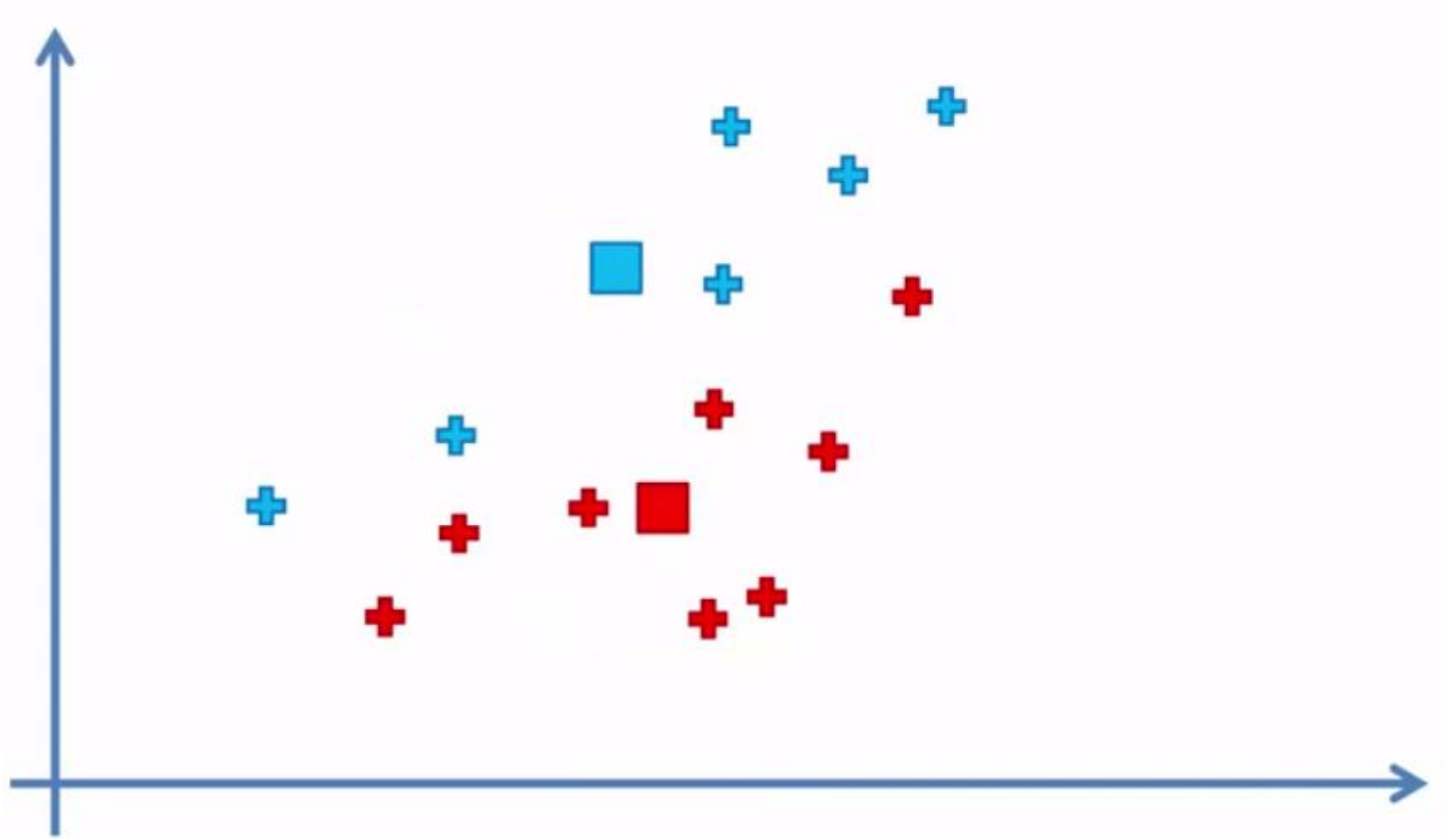
PASSO 3: ASSOCIAR CADA OBJETO AO CLUSTER (CENTROIDE) MAIS PRÓXIMO (MAIOR SIMILARIDADE), FORMANDO K CLUSTERS



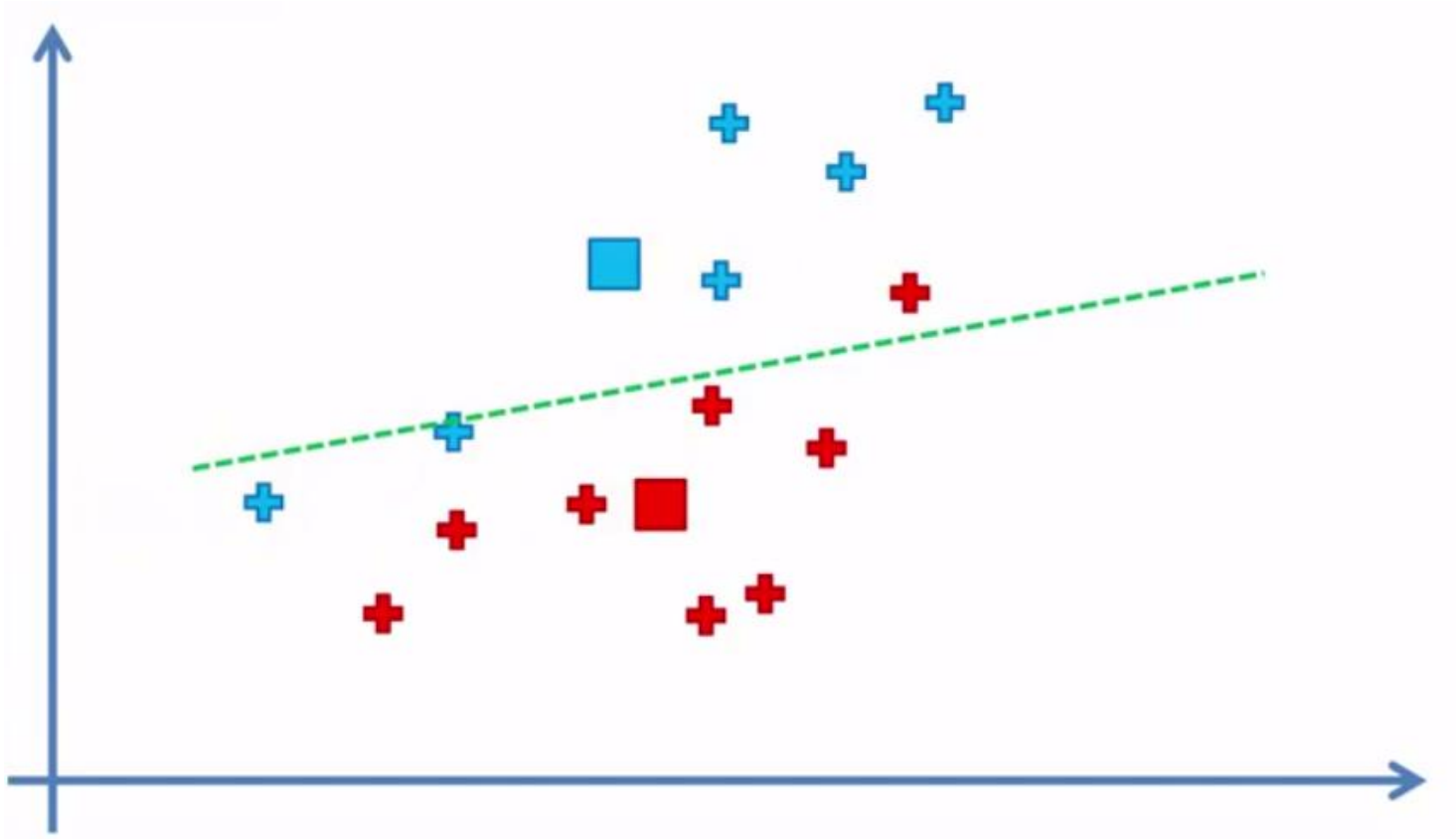
PASSO 4: CALCULAR E REALOCAR O NOVO CENTROIDE DE CADA CLUSTER (MÉDIA PARA CADA ATRIBUTO, POR EXEMPLO)



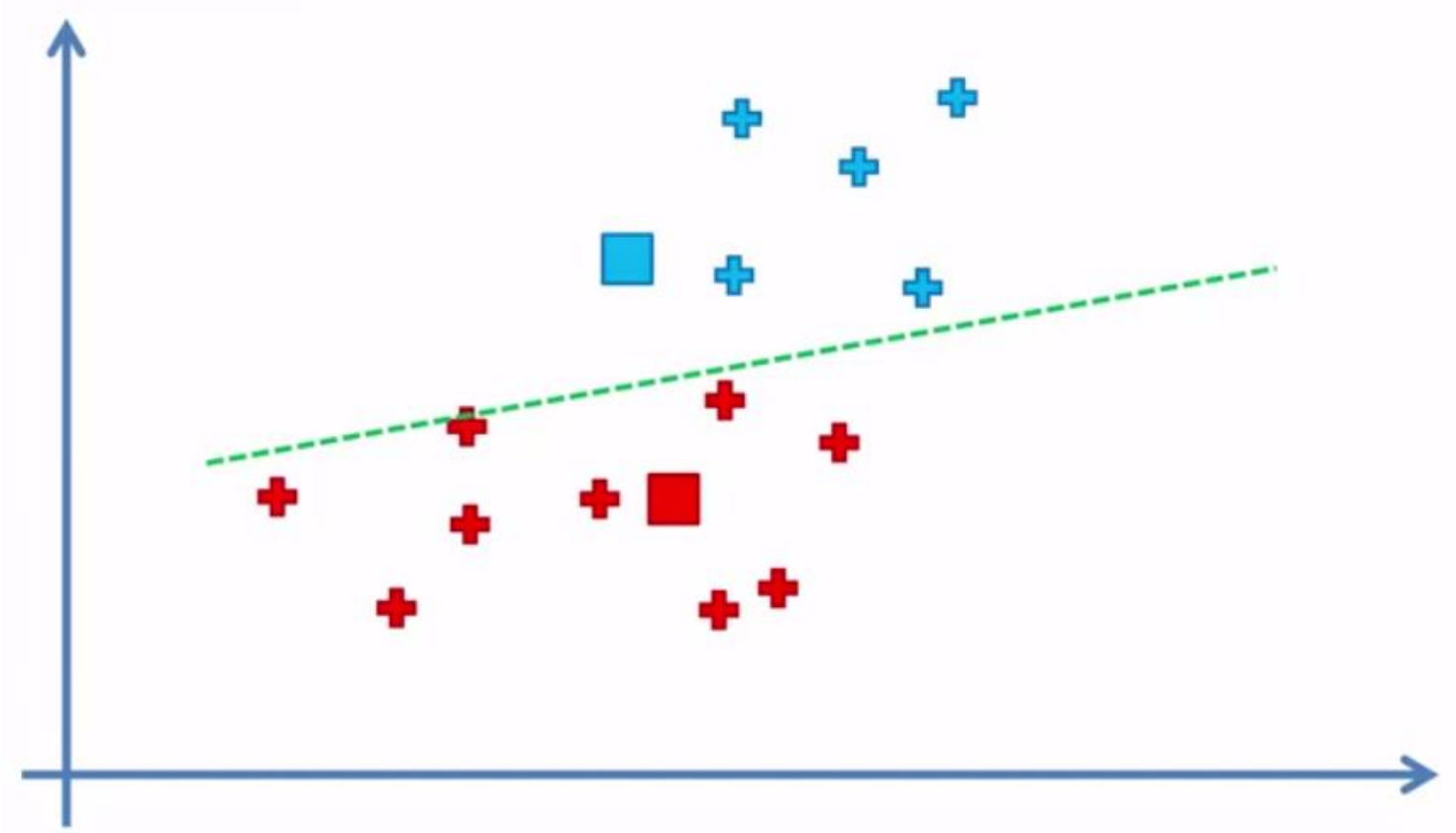
PASSO 4: CALCULAR E REALOCAR O NOVO CENTROIDE DE CADA CLUSTER



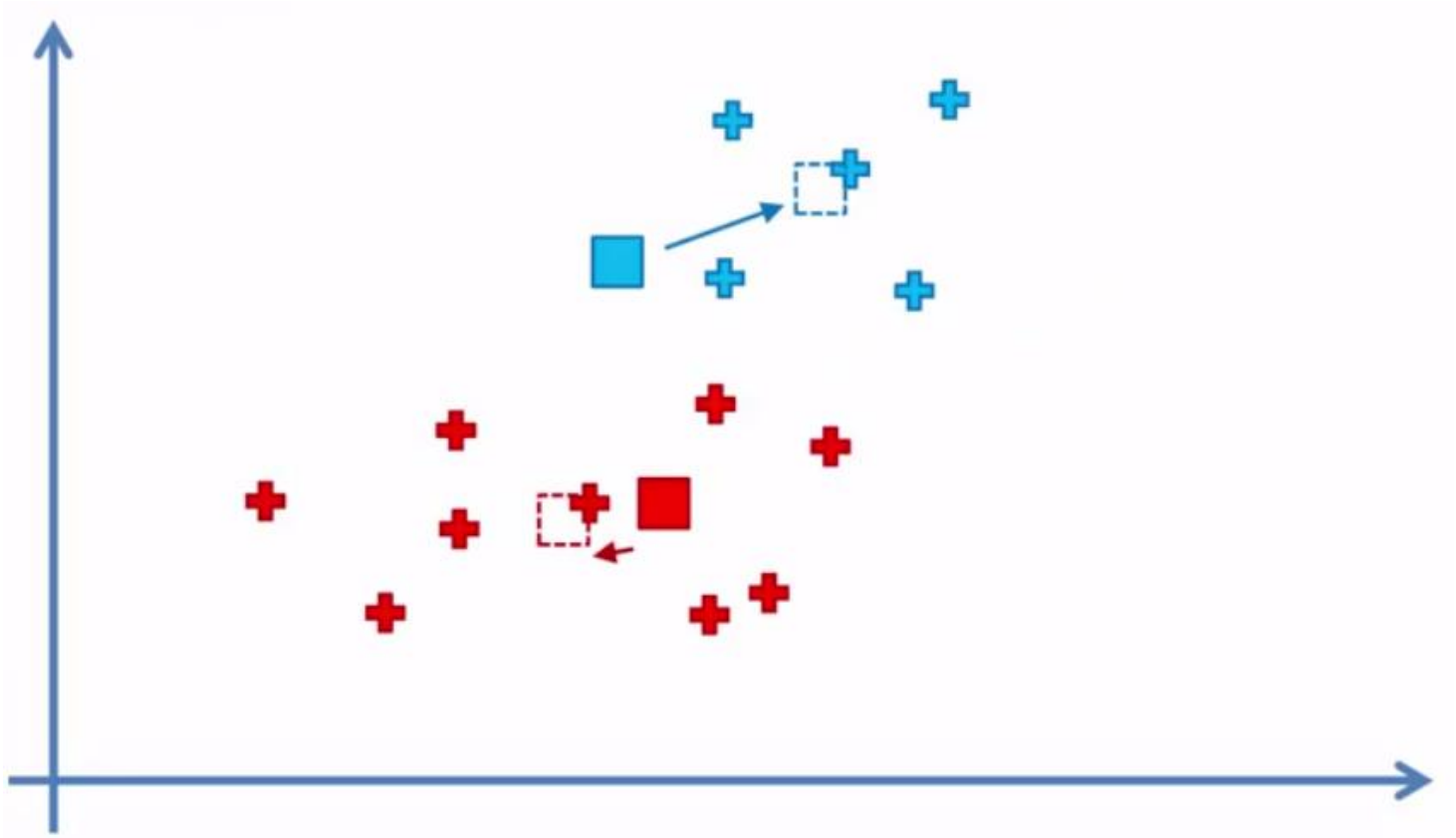
PASSO 5: ASSOCIAR CADA OBJETO AO CLUSTER MAIS PRÓXIMO. VOLTAR AO PASSO 4 SE ALGUM OBJETO FOI MOVIDO DE CLUSTER. TERMINAR CASO CONTRÁRIO.



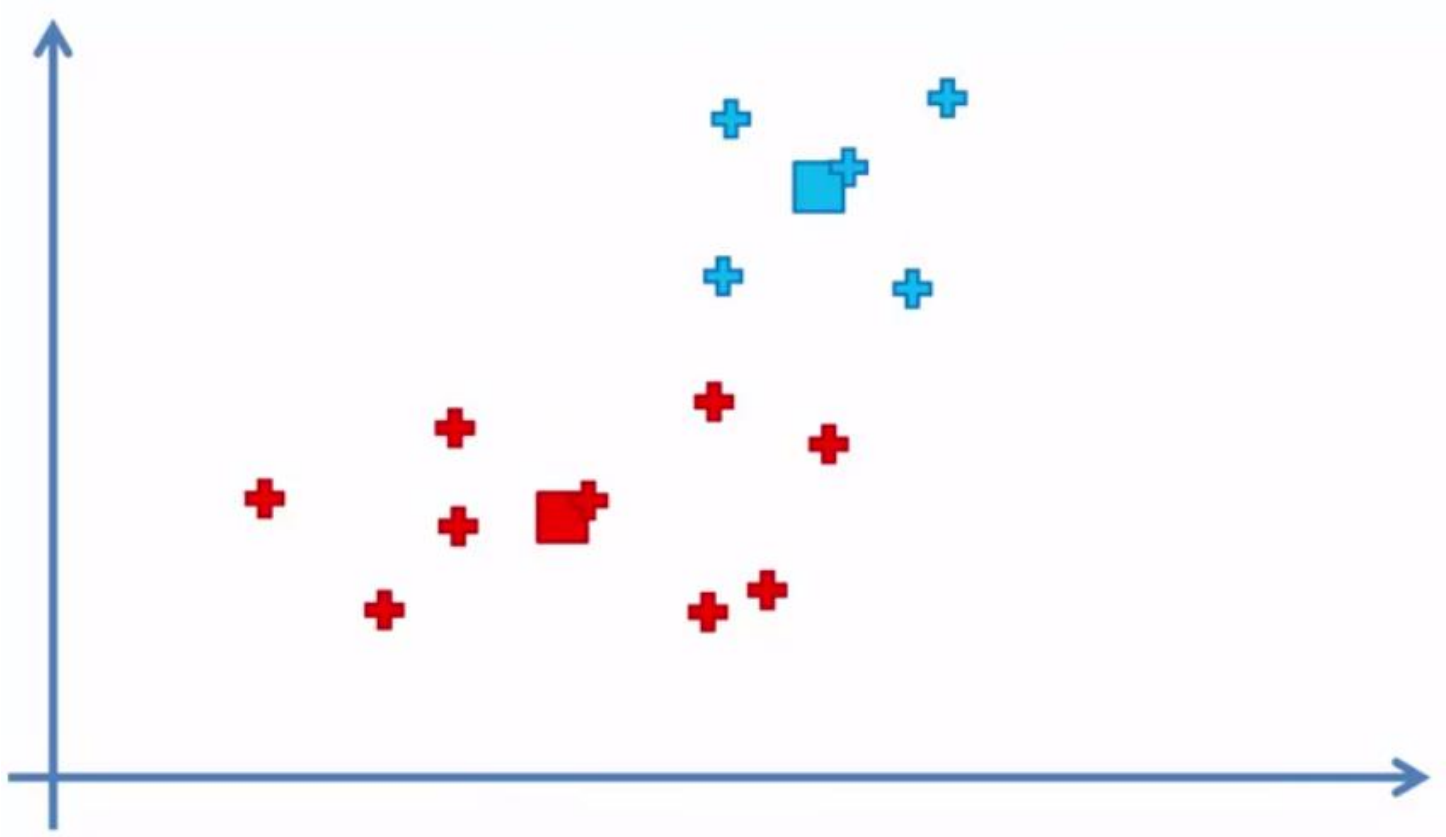
PASSO 5: ASSOCIAR CADA OBJETO AO CLUSTER MAIS PRÓXIMO. VOLTAR AO PASSO 4 SE ALGUM OBJETO FOI MOVIDO DE CLUSTER. TERMINAR CASO CONTRÁRIO.



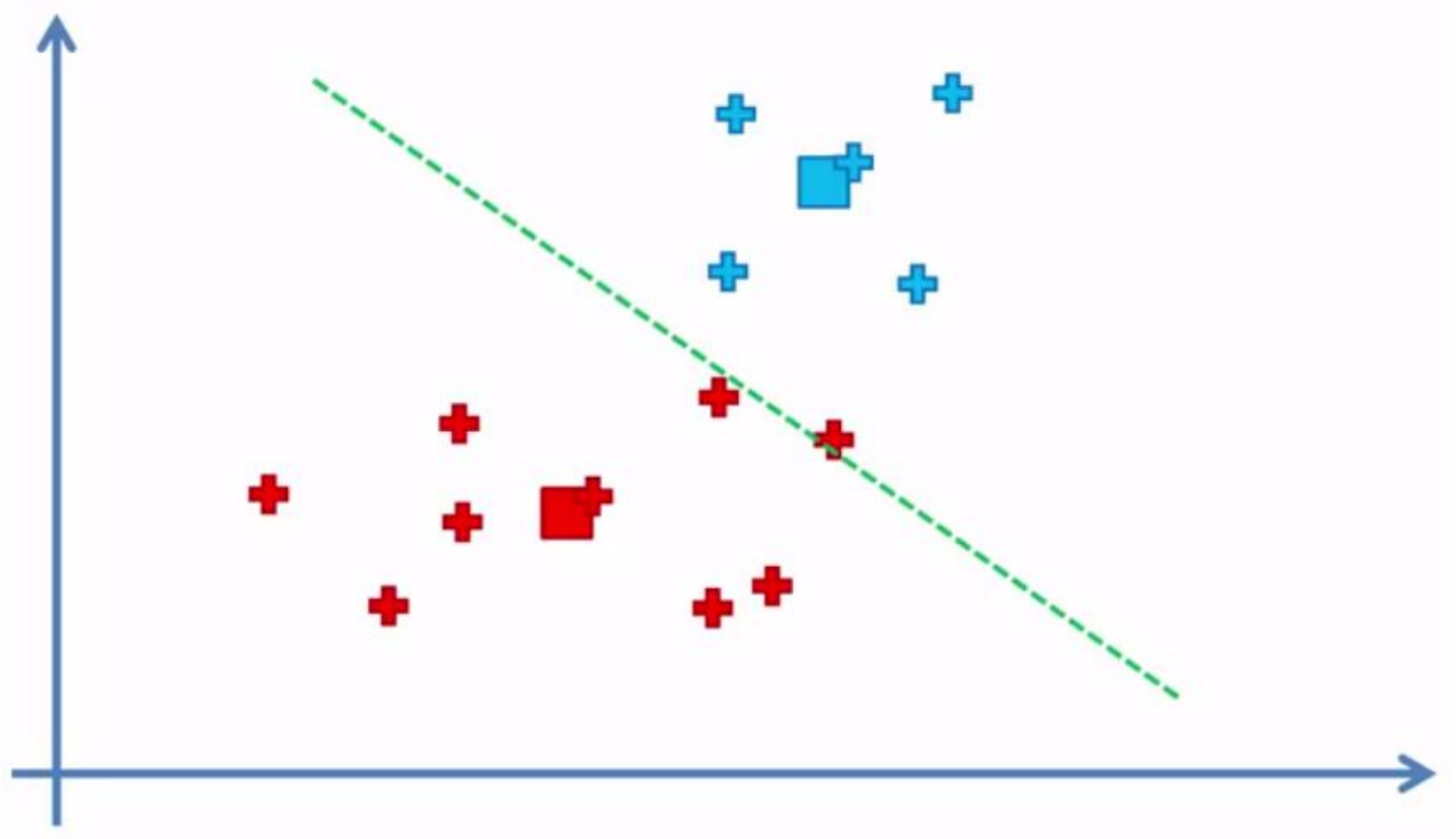
PASSO 4: CALCULAR E REALOCAR O NOVO CENTROIDE DE CADA CLUSTER



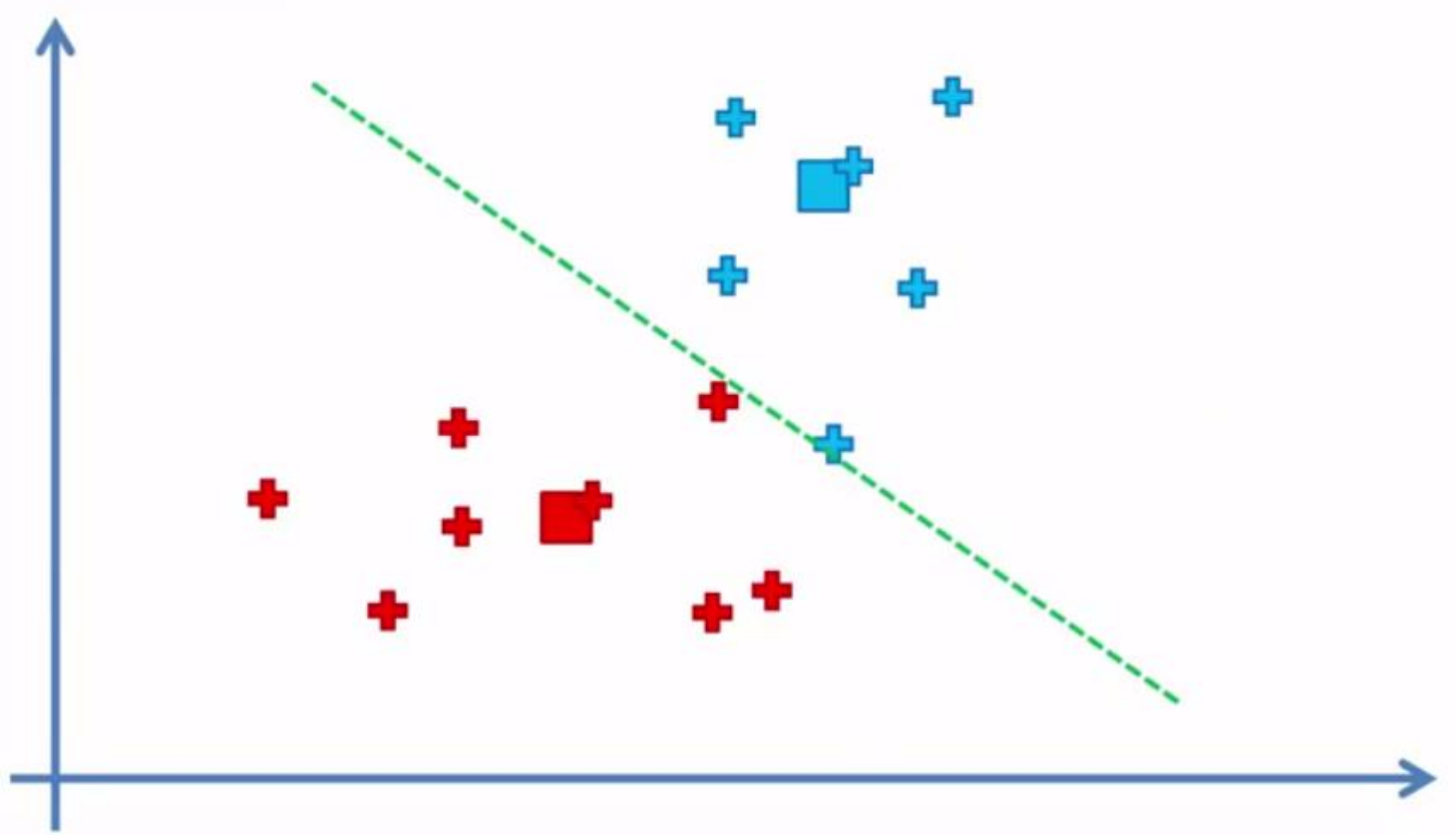
PASSO 4: CALCULAR E REALOCAR O NOVO CENTROIDE DE CADA CLUSTER



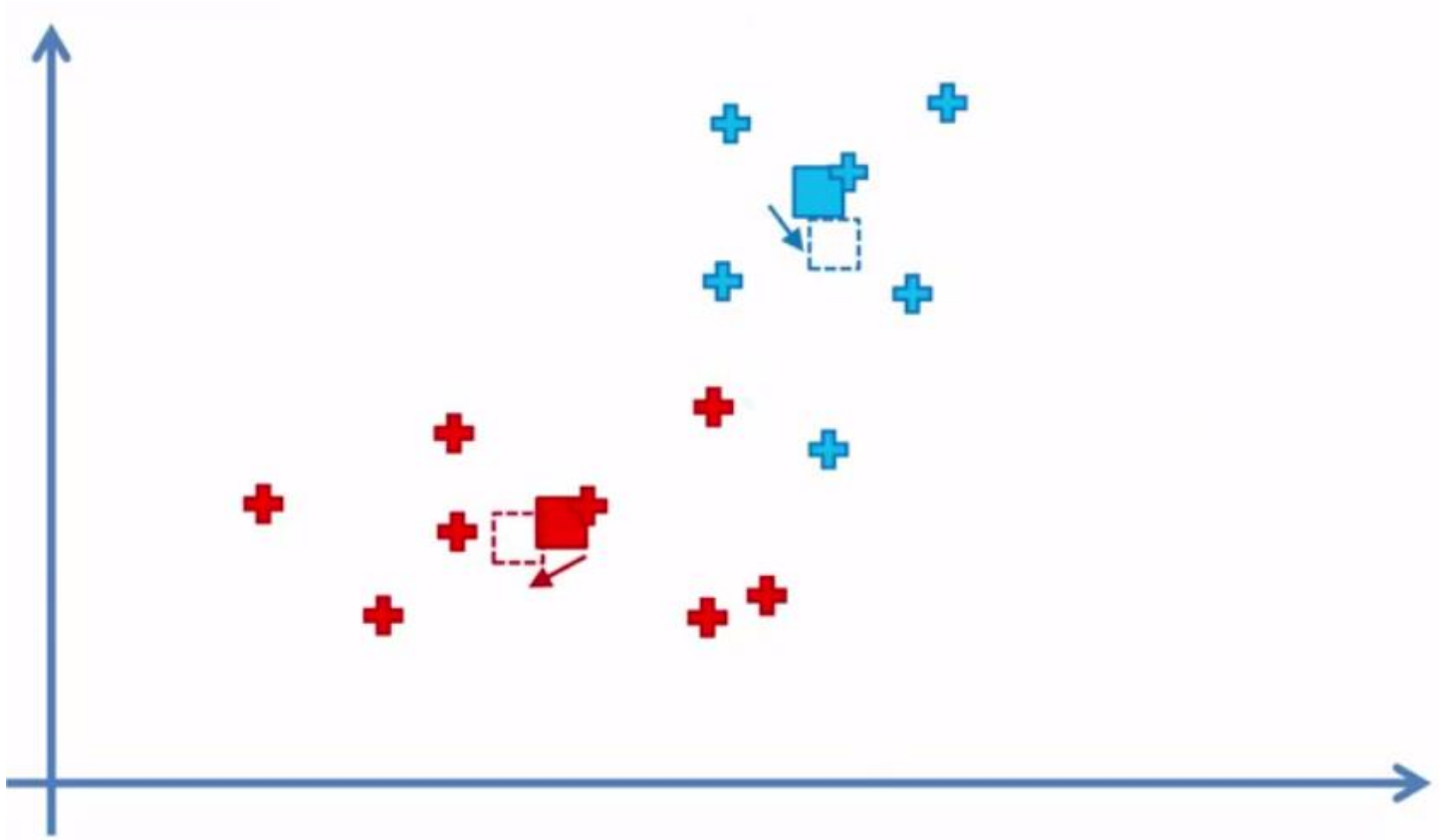
PASSO 5: ASSOCIAR CADA OBJETO AO CLUSTER MAIS PRÓXIMO. VOLTAR AO PASSO 4 SE ALGUM OBJETO FOI MOVIDO DE CLUSTER. TERMINAR CASO CONTRÁRIO.



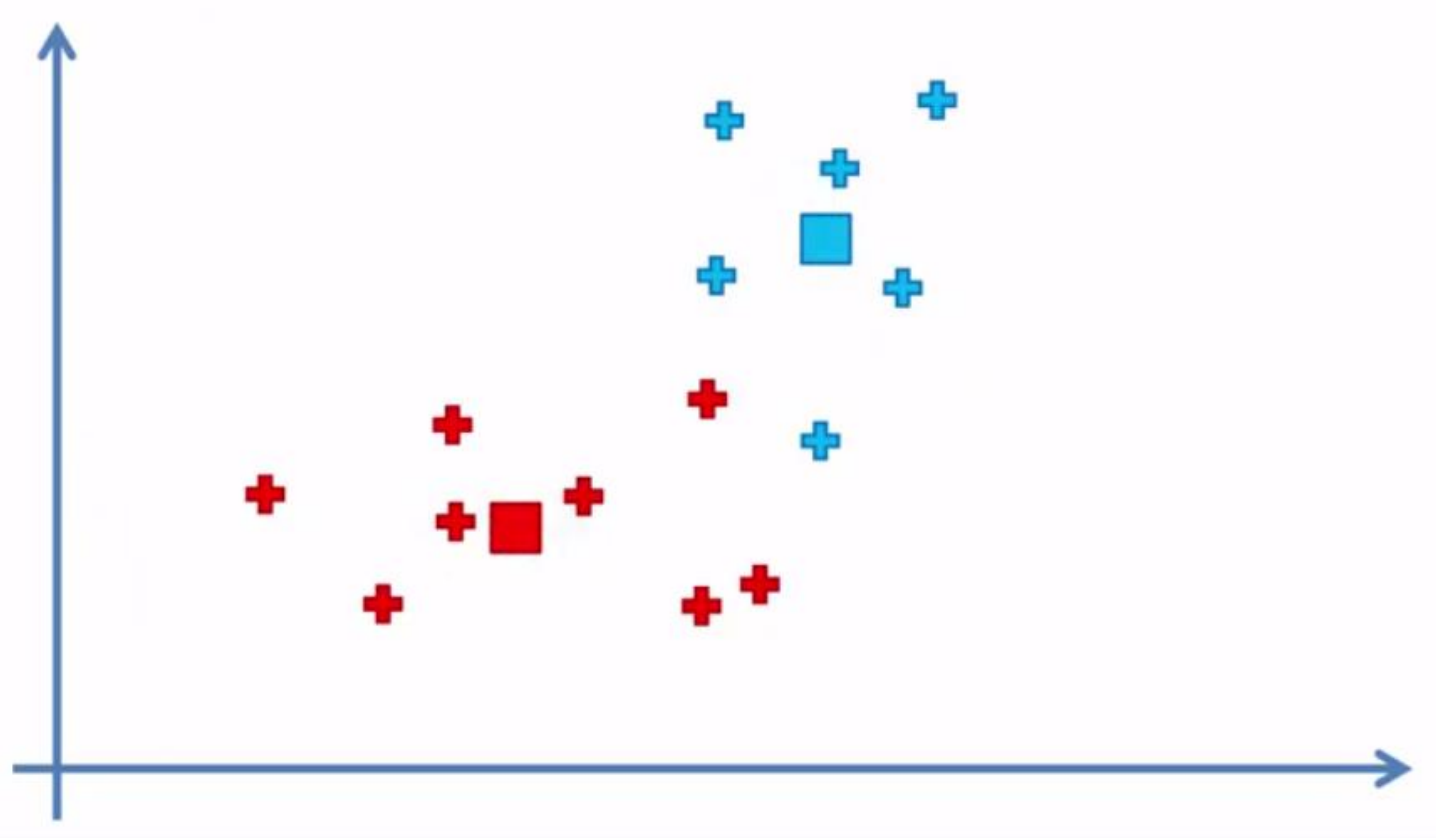
PASSO 5: ASSOCIAR CADA OBJETO AO CLUSTER MAIS PRÓXIMO. VOLTAR AO PASSO 4 SE ALGUM OBJETO FOI MOVIDO DE CLUSTER. TERMINAR CASO CONTRÁRIO.



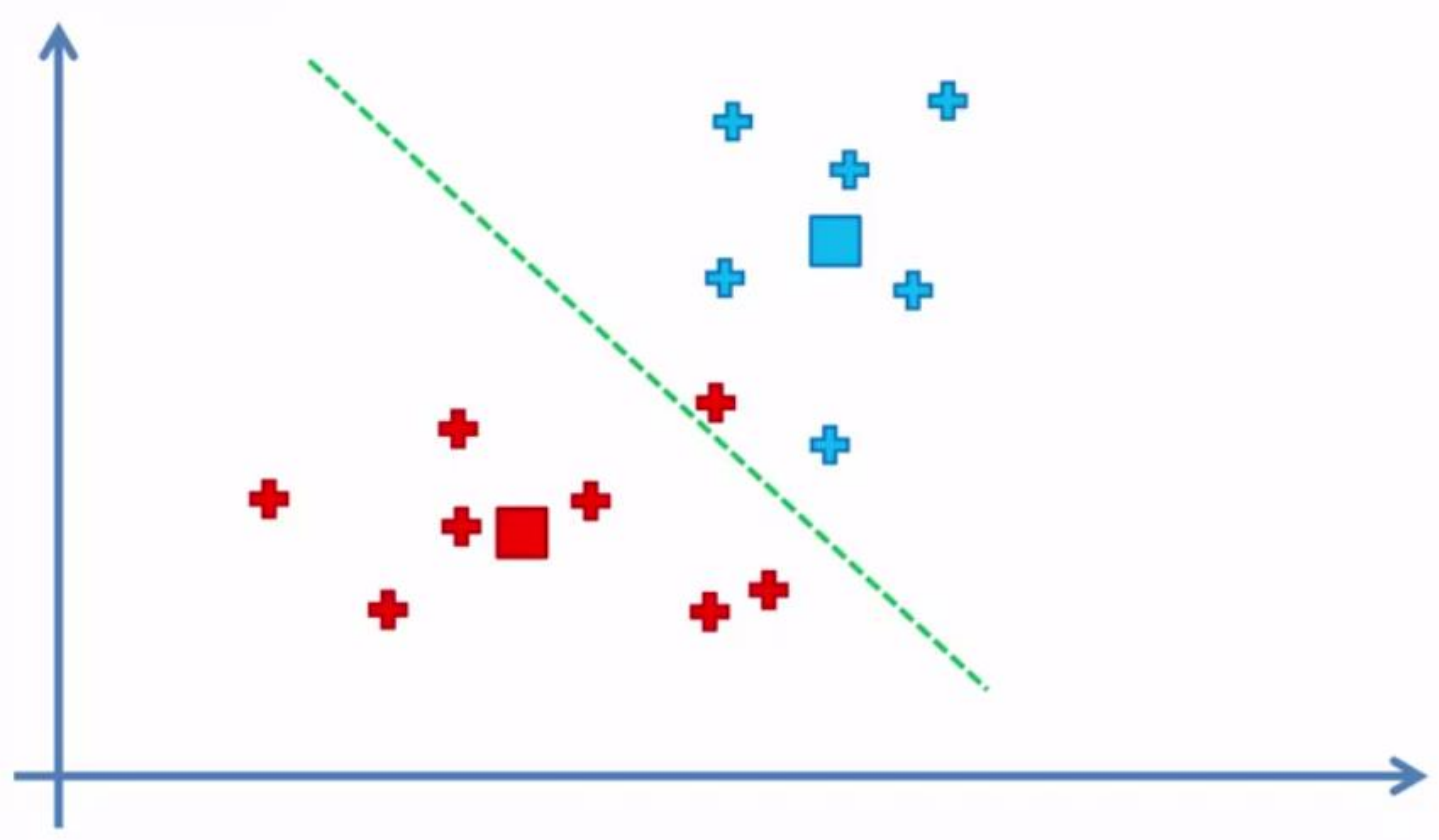
PASSO 4: CALCULAR E REALOCAR O NOVO CENTROIDE DE CADA CLUSTER



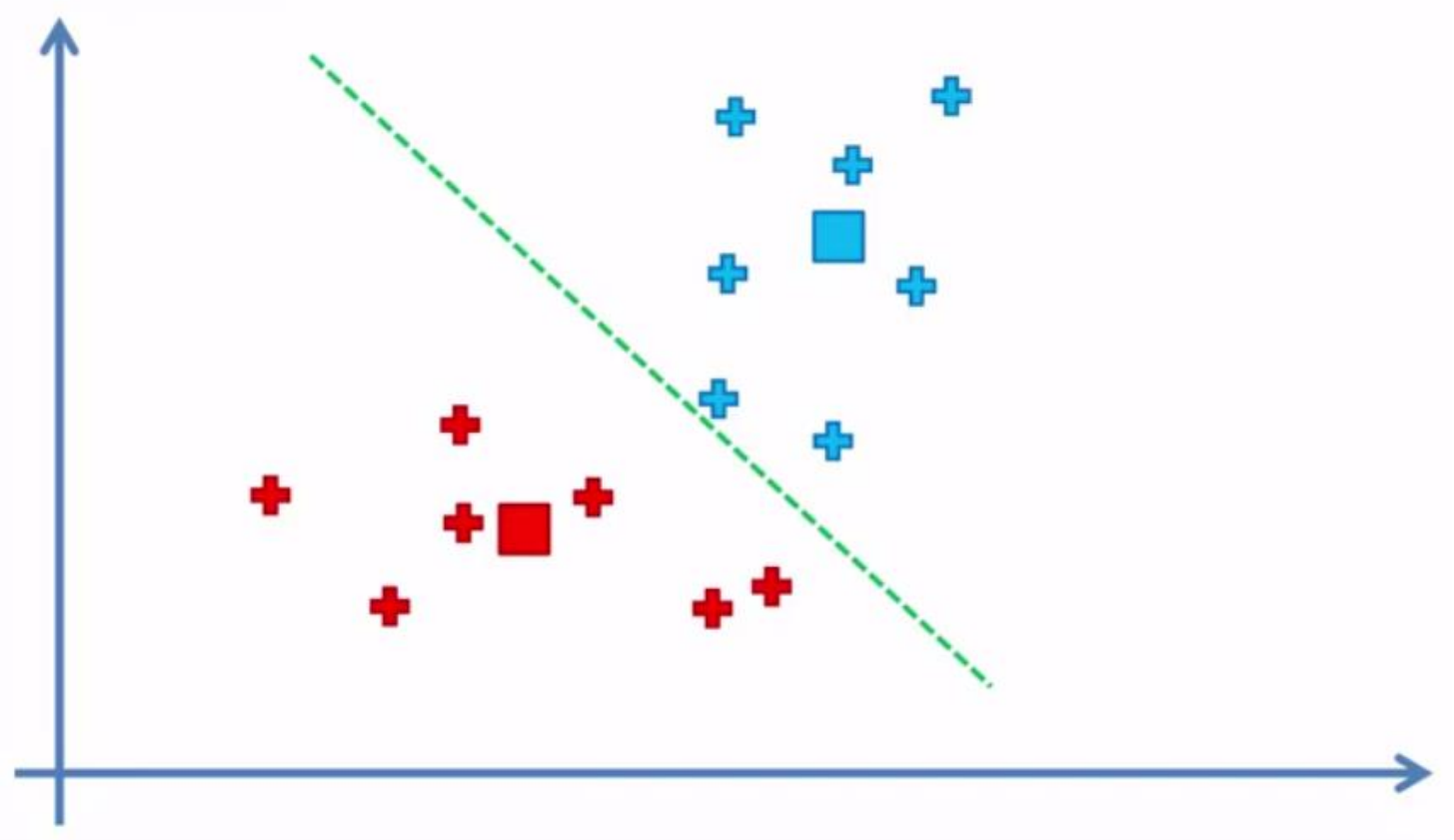
PASSO 4: CALCULAR E REALOCAR O NOVO CENTRÓIDE DE CADA CLUSTER



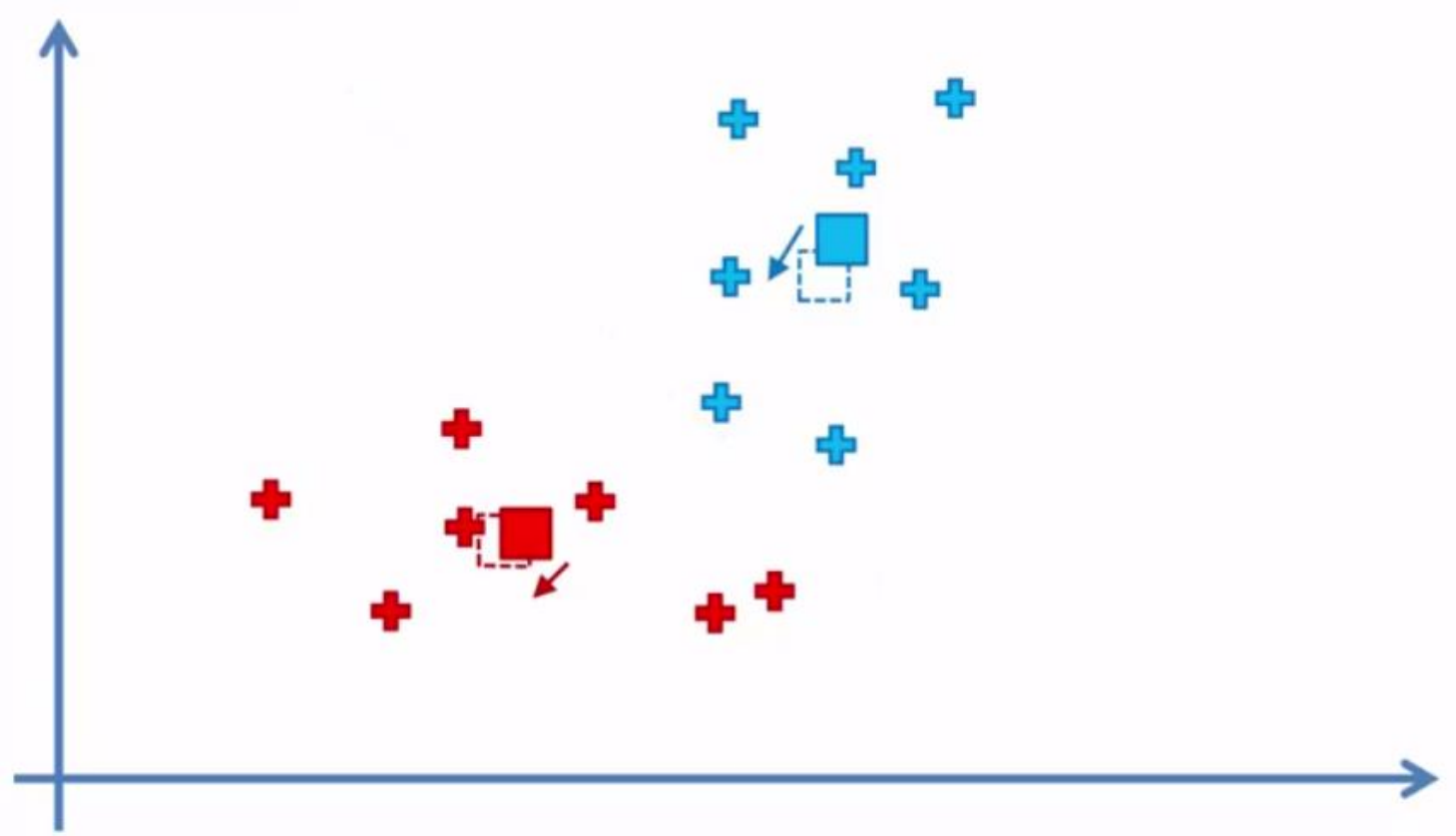
PASSO 5: ASSOCIAR CADA OBJETO AO CLUSTER MAIS PRÓXIMO. VOLTAR AO PASSO 4 SE ALGUM OBJETO FOI MOVIDO DE CLUSTER. TERMINAR CASO CONTRÁRIO.



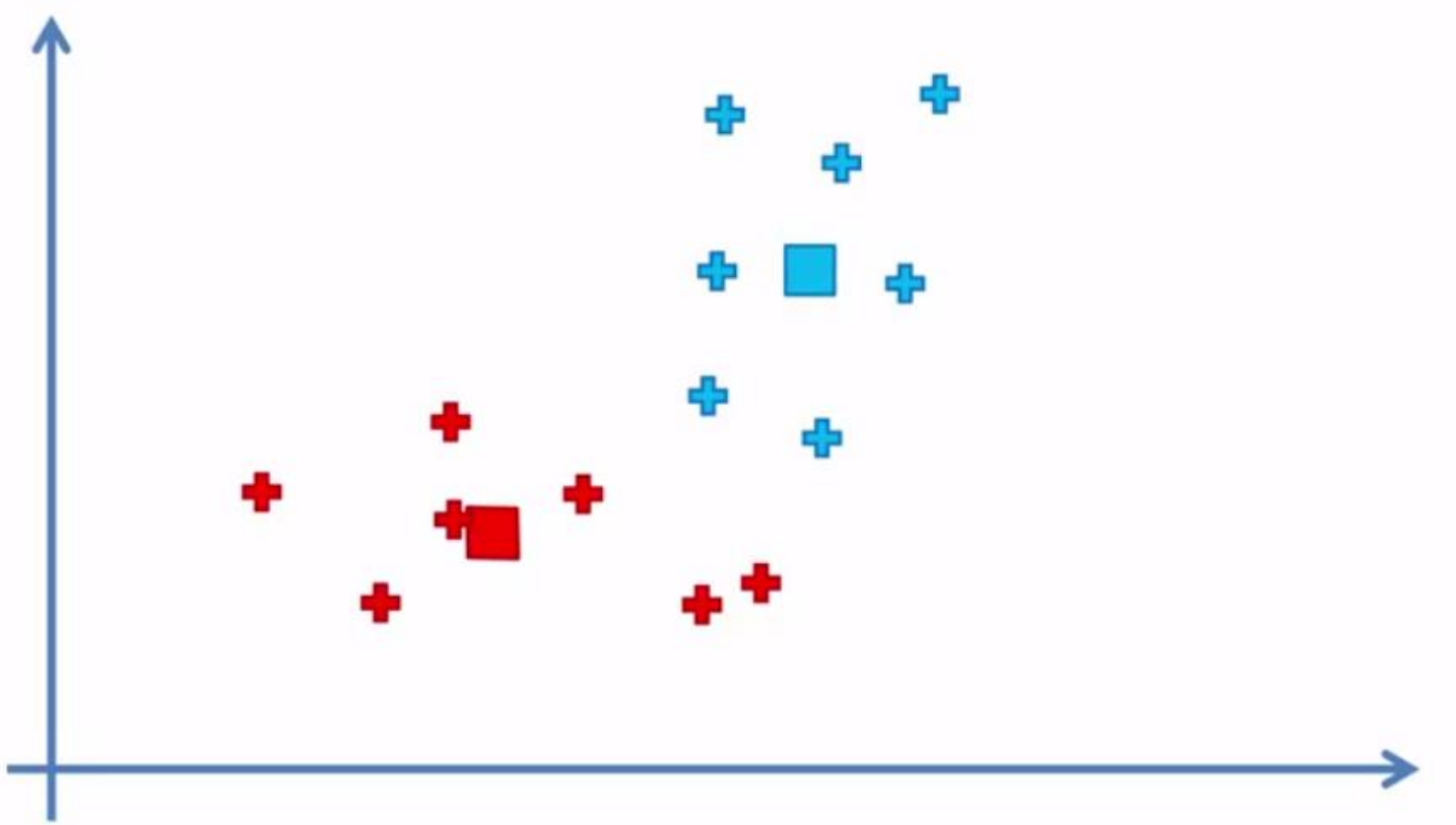
PASSO 5: ASSOCIAR CADA OBJETO AO CLUSTER MAIS PRÓXIMO. VOLTAR AO PASSO 4 SE ALGUM OBJETO FOI MOVIDO DE CLUSTER. TERMINAR CASO CONTRÁRIO.



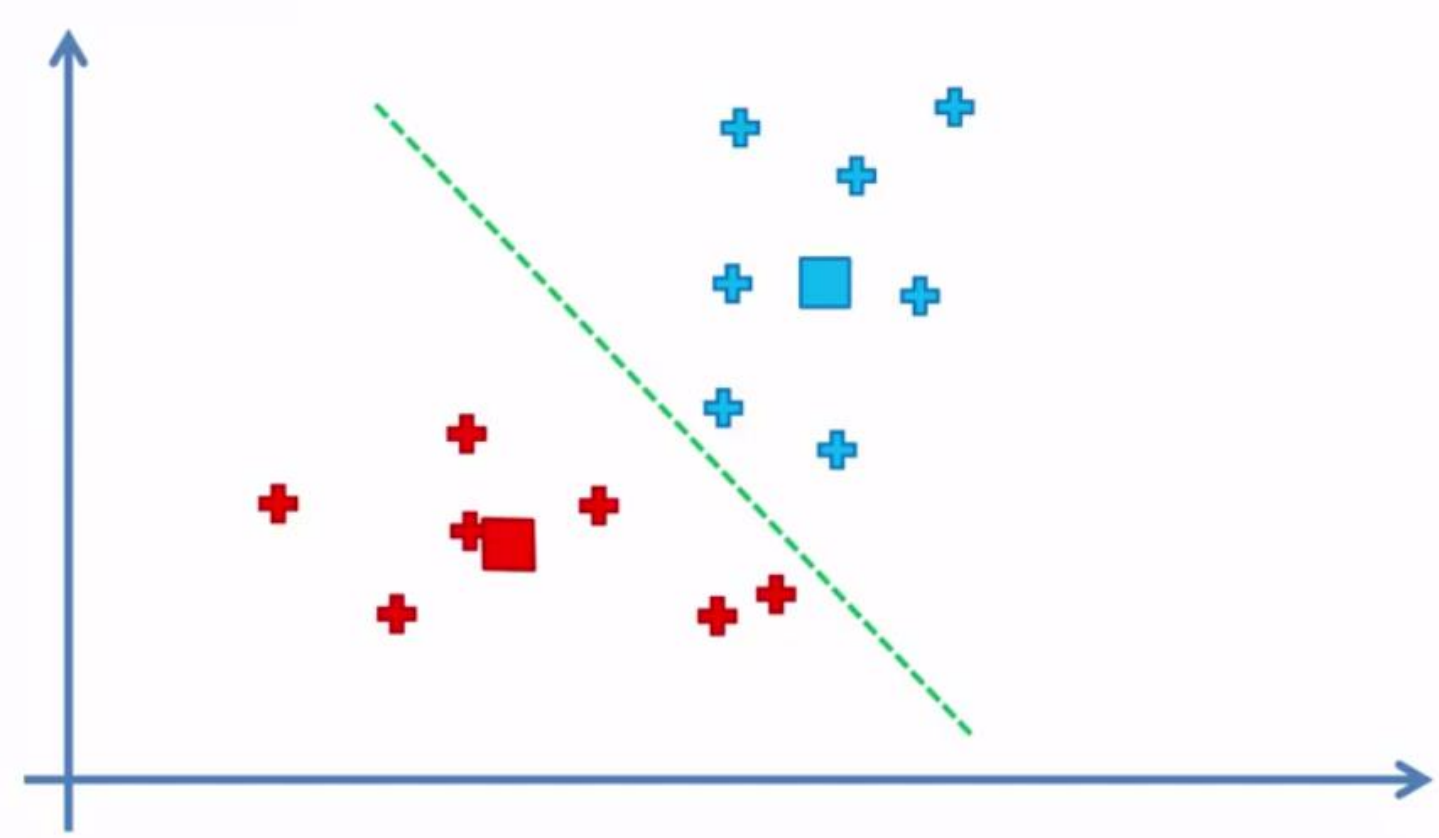
PASSO 4: CALCULAR E REALOCAR O NOVO CENTROIDE DE CADA CLUSTER



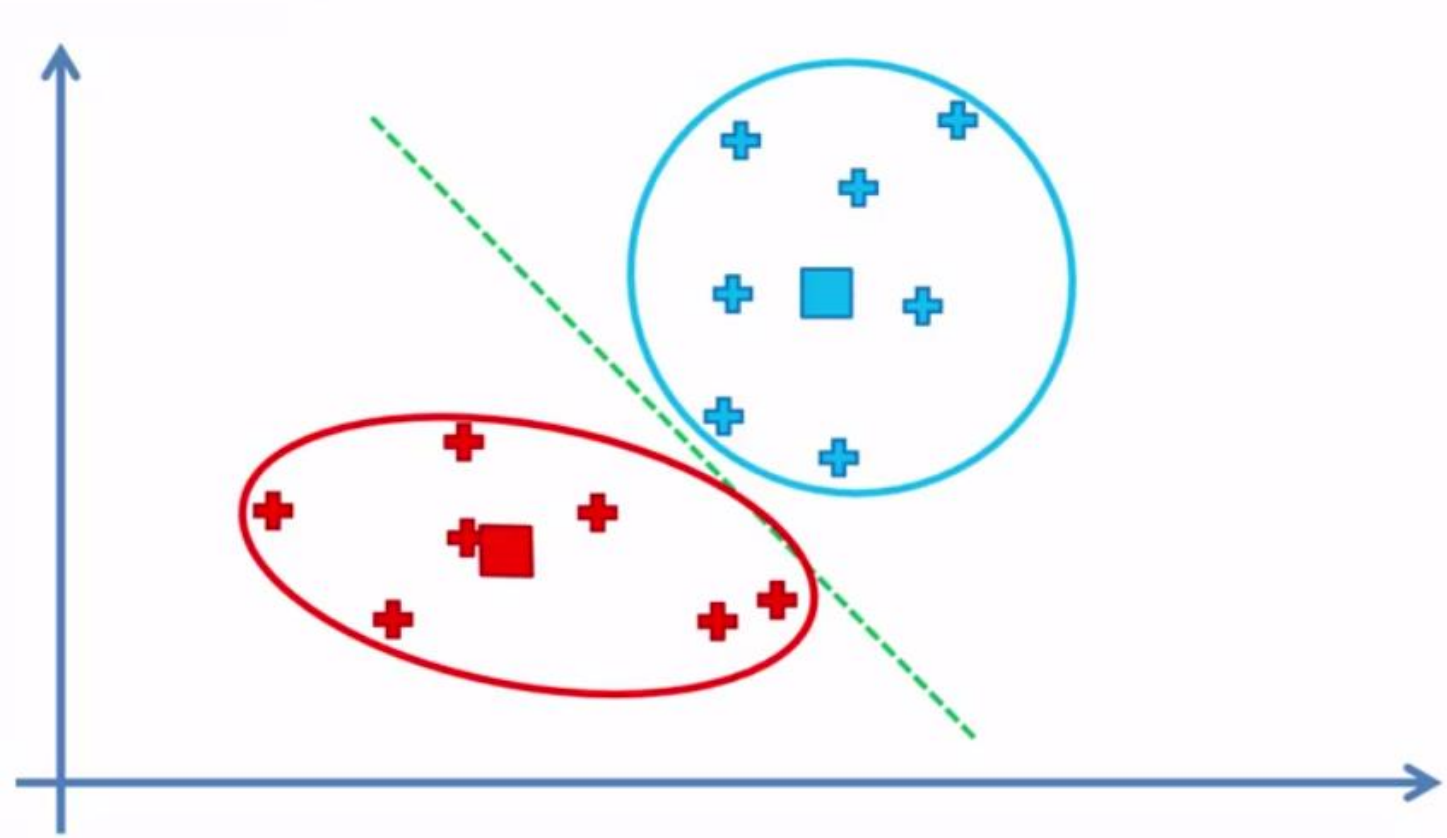
PASSO 4: CALCULAR E REALOCAR O NOVO CENTROIDE DE CADA CLUSTER



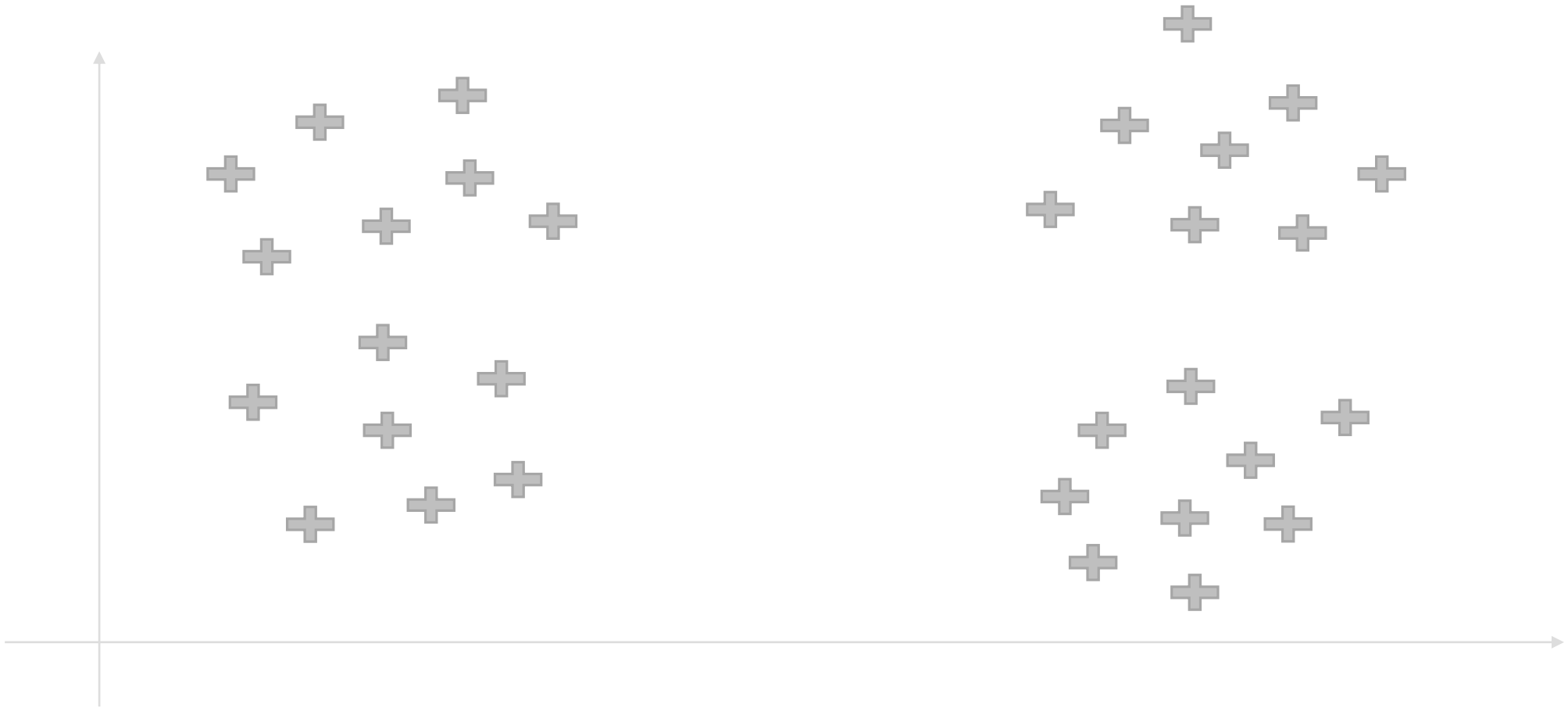
PASSO 5: ASSOCIAR CADA OBJETO AO CLUSTER MAIS PRÓXIMO. VOLTAR AO PASSO 4 SE ALGUM OBJETO FOI MOVIDO DE CLUSTER. TERMINAR CASO CONTRÁRIO.



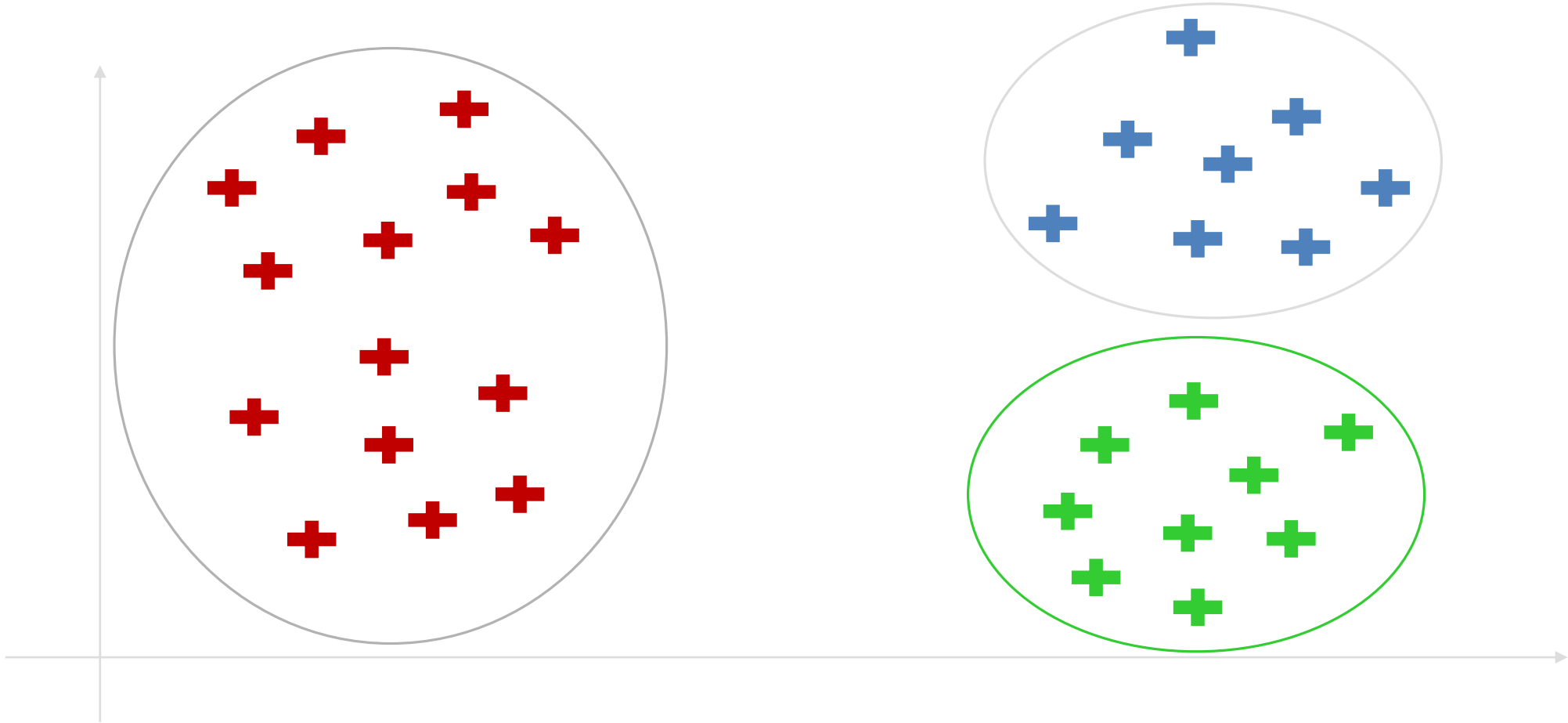
MODELO FINAL



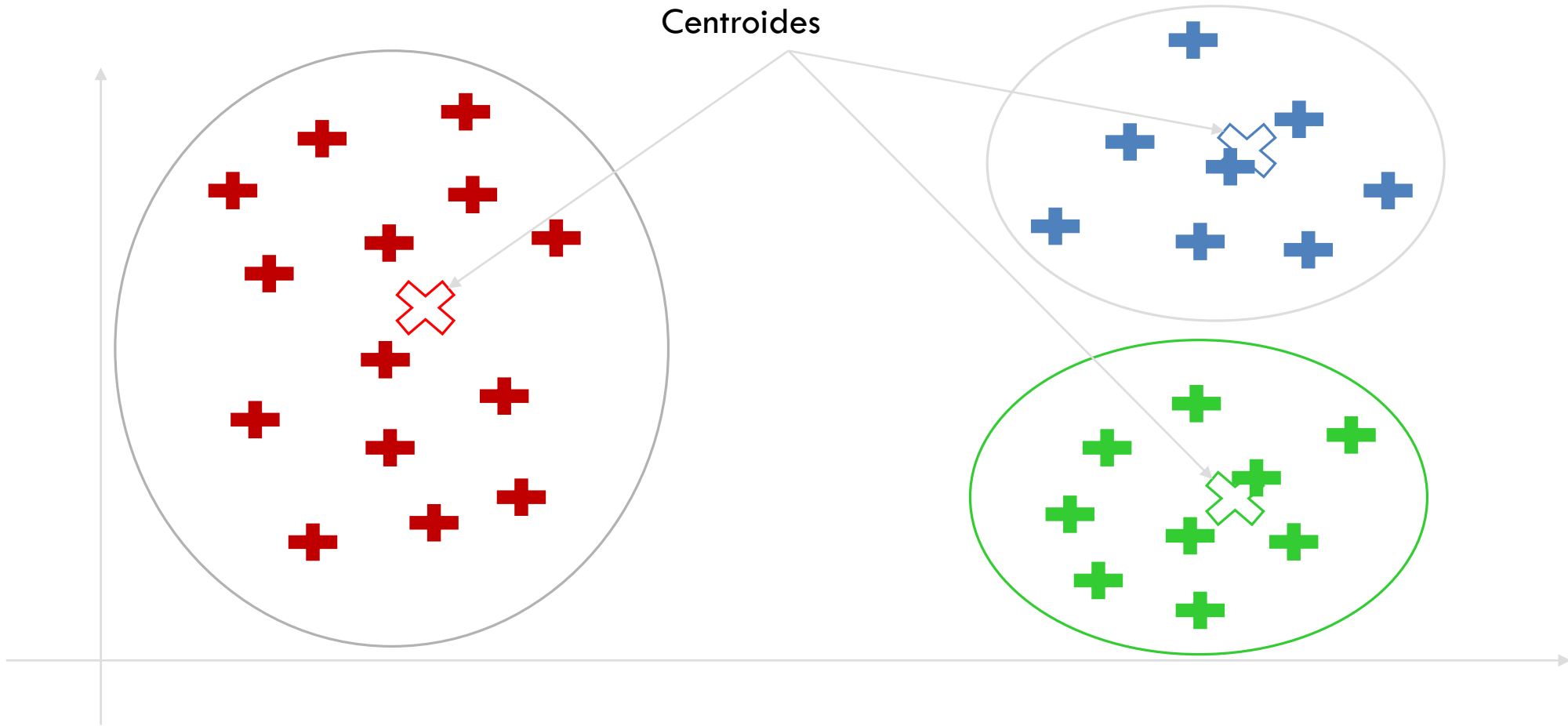
COMO DETERMINAR O NÚMERO DE CLUSTERS?



COMO DETERMINAR O NÚMERO DE CLUSTERS?



COMO DETERMINAR O NÚMERO DE CLUSTERS?

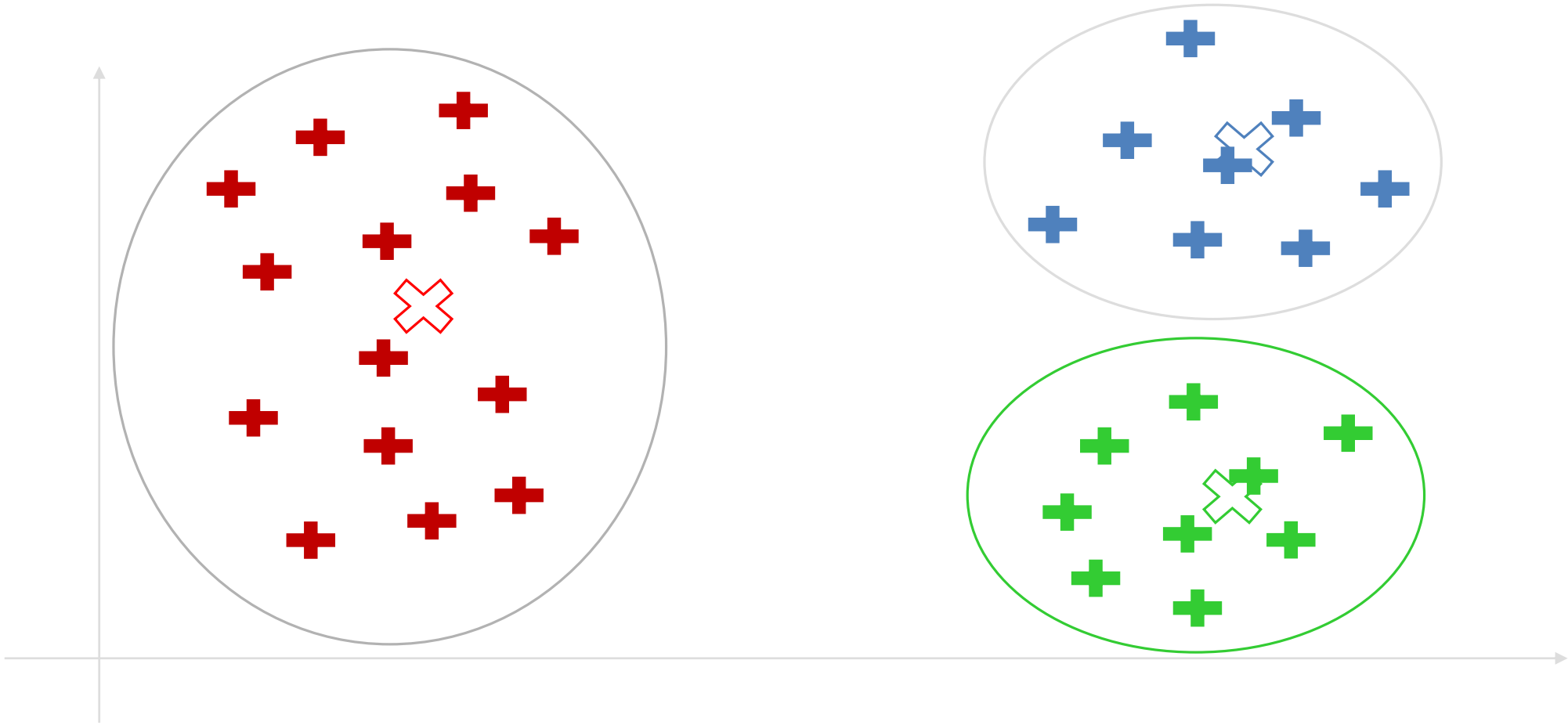


MÉTRICA PARA AVALIAR O NÚMERO DE CLUSTERS

WCSS (**W**ithin **C**luster **S**um of **S**quares)

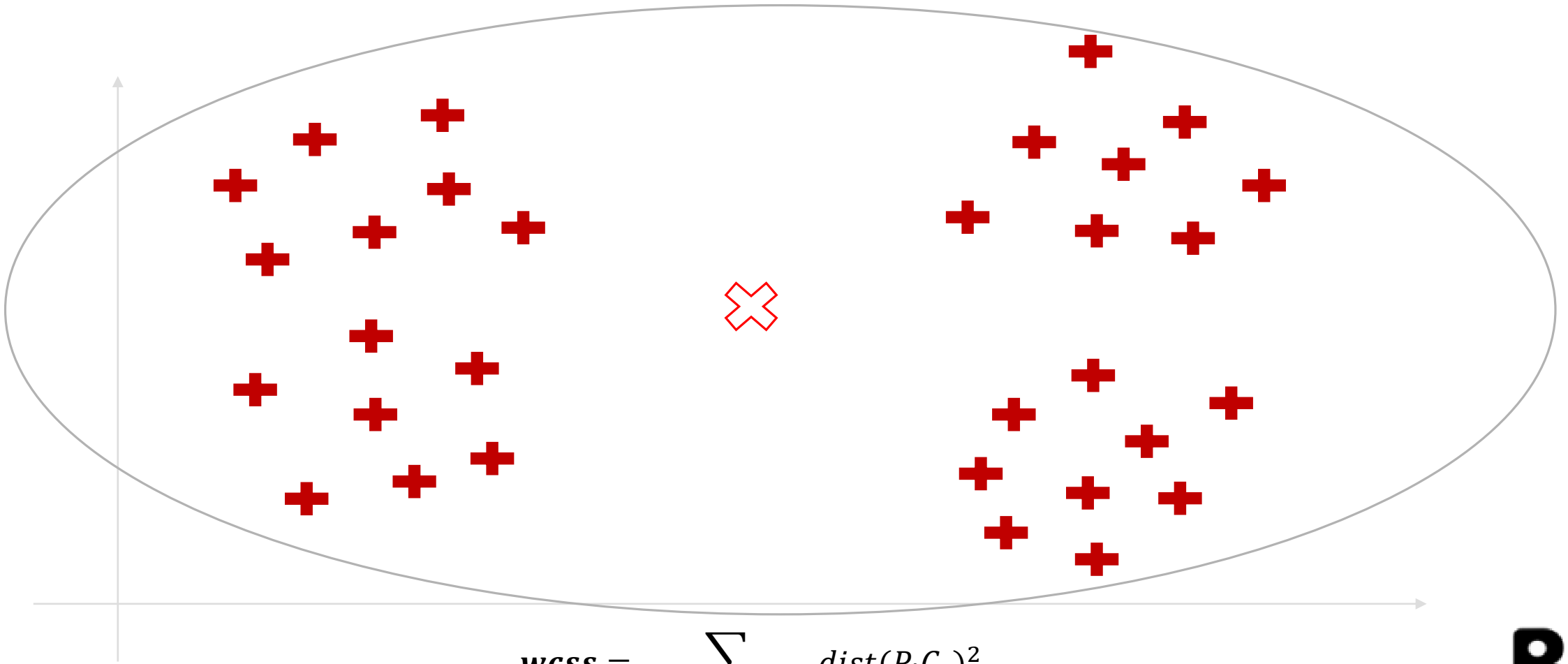
$$\boldsymbol{wcss} = \sum_{Cluster\ j} \sum_{P_i\ no\ cluster\ j} dist(P_i C_j)^2$$

Pi: pontos da base de índice *i*
C: centroide



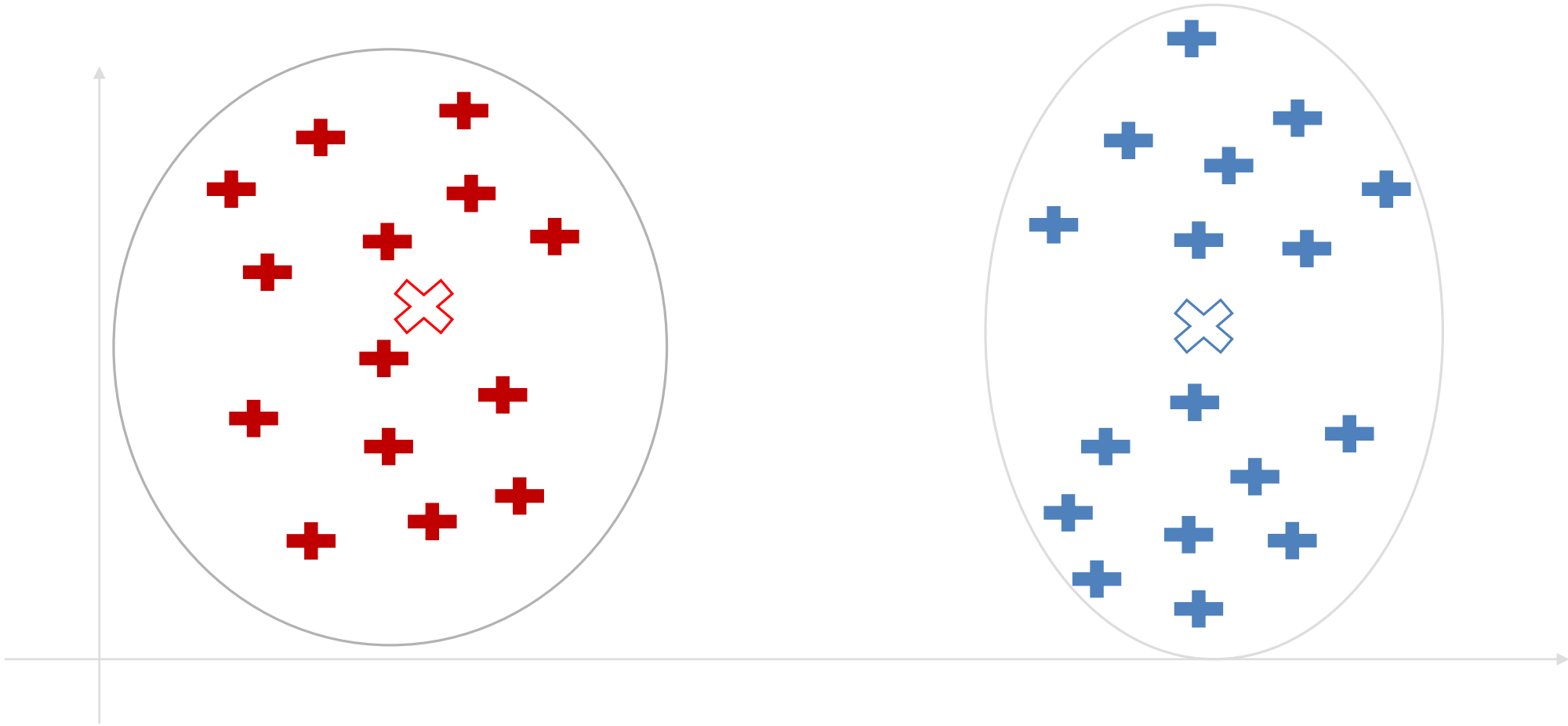
$$\boldsymbol{wcss} = \sum_{P_i \text{ no cluster } 1} dist(P_i C_1)^2 + \sum_{P_i \text{ no cluster } 2} dist(P_i C_2)^2 + \sum_{P_i \text{ no cluster } 3} dist(P_i C_3)^2$$

Intuitivamente, o WCSS nesse caso vai ser muito grande, pois a distância de cada ponto até o centroide é muito grande.



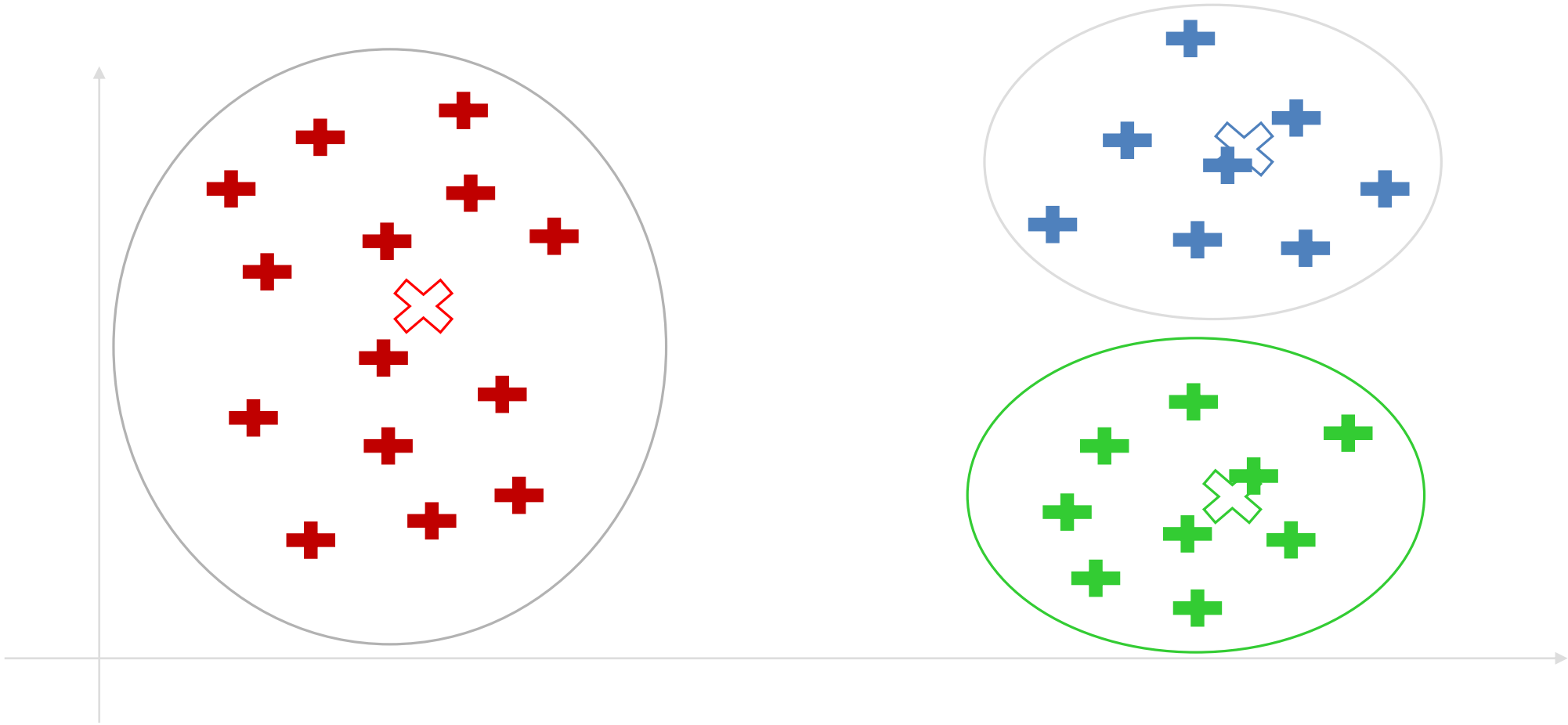
$$\boldsymbol{wcss} = \sum_{P_i \, no \, cluster \, 1} dist(P_i C_1)^2$$

Aqui, o WCSS vai ser menor, já que temos dois clusters e as distâncias entre eles e cada ponto será menor.



$$\boldsymbol{wcss} = \sum_{P_i \text{ no cluster } 1} dist(P_i C_1)^2 + \sum_{P_i \text{ no cluster } 2} dist(P_i C_2)^2$$

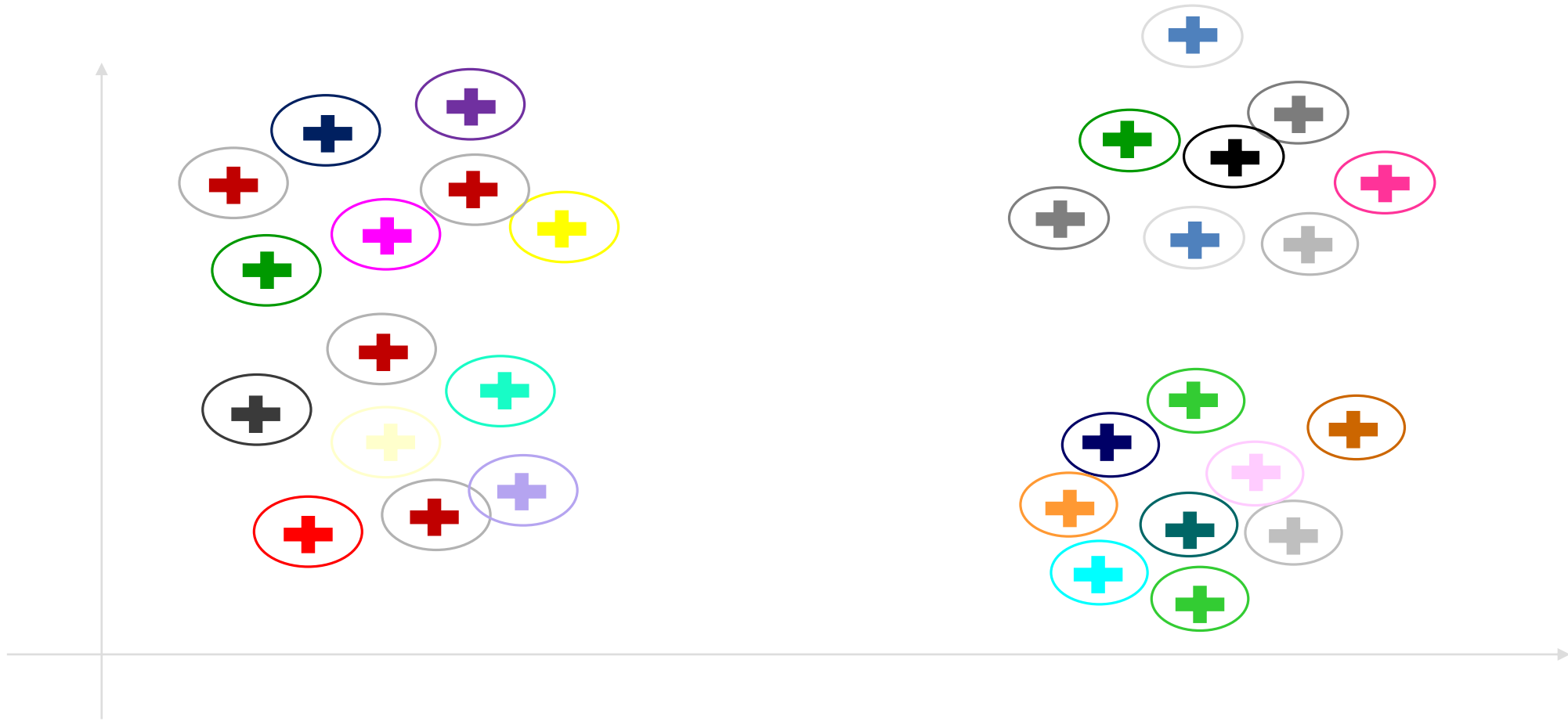
Aqui, o WCSS será menor ainda



$$\boldsymbol{wcss} = \sum_{P_i \text{ no cluster } 1} dist(P_i C_1)^2 + \sum_{P_i \text{ no cluster } 2} dist(P_i C_2)^2 + \sum_{P_i \text{ no cluster } 3} dist(P_i C_3)^2$$

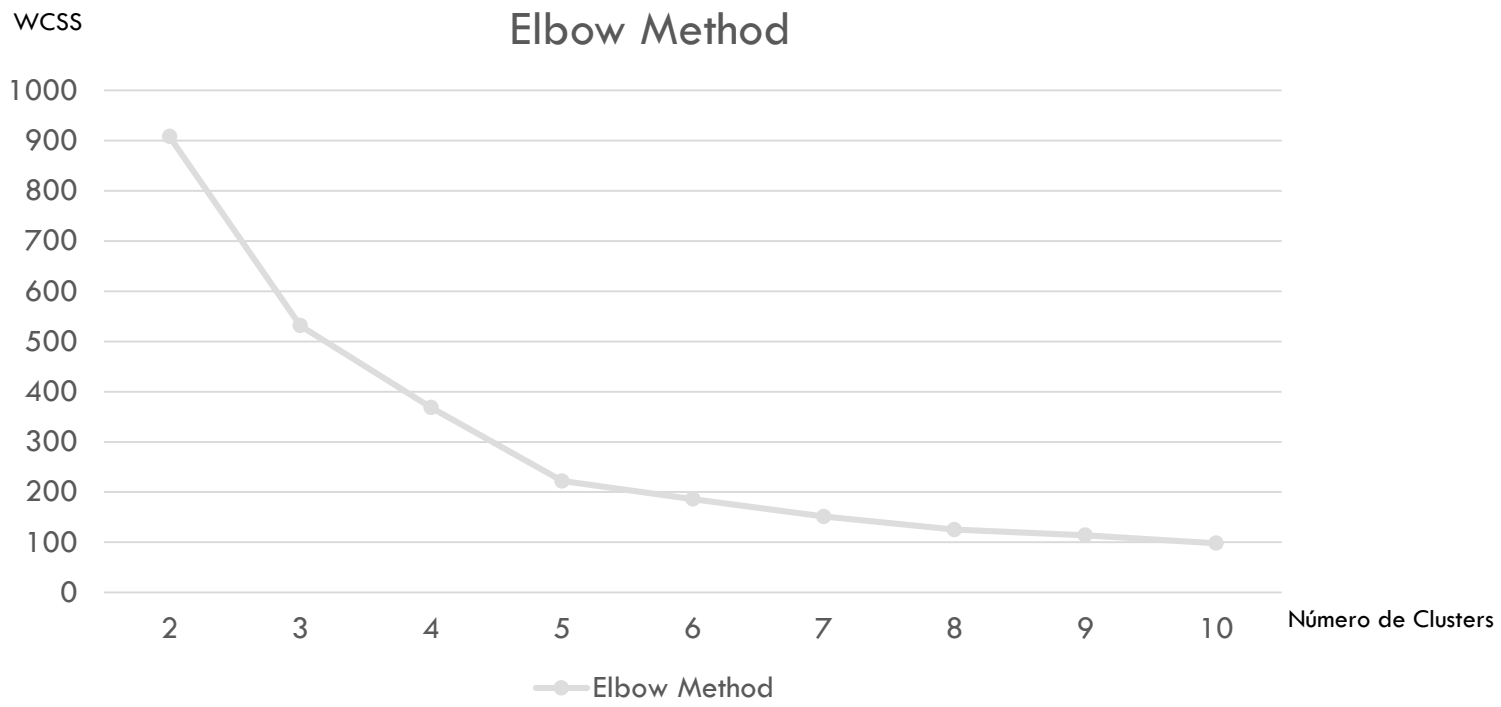
O número de clusters pode chegar até ao valor do número de dados da sua base, mas obviamente, essa não é uma abordagem boa, já que cada ponto será de um cluster, e seu centroide será igual ao ponto.

Nesse caso, WCSS = 0.



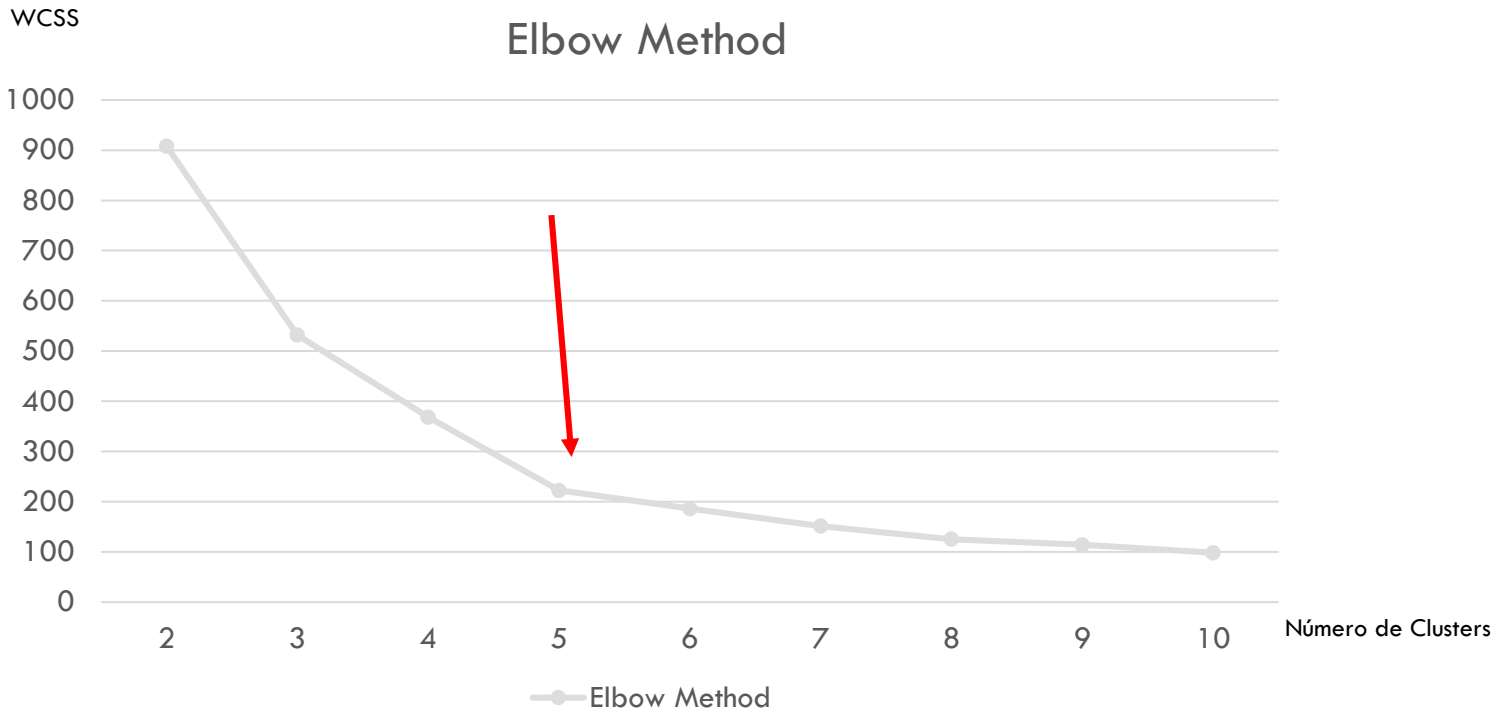
ELBOW METHOD

K	WCSS
2	908.329
3	531.742
4	368.399
5	222.242
6	186.169
7	151.367
8	125.059
9	113.953
10	98.218



ELBOW METHOD

K	WCSS
2	908.329
3	531.742
4	368.399
5	222.242
6	186.169
7	151.367
8	125.059
9	113.953
10	98.218



Usualmente, o número de clusters é definido pela inclinação da reta. Quando a melhora no aumento de cluster já não é tão significativa quando comparada com a melhora imediatamente anterior.

VARIAÇÕES DO MÉTODO K-MEANS

Algumas versões do K-means diferem em:

- Seleção dos pontos iniciais.
- Cálculo da similaridade entre os pontos.
- Estratégias para calcular os centróides dos clusters.

Para atributos nominais: K-modes (Huang'98)

- Substitui as médias dos clusters por modas.
- Usa medidas de similaridade para atributos nominais.
- Usa um método baseado em frequências para atualizar as modas dos clusters.

CLUSTERIZAÇÃO HIERÁRQUICA

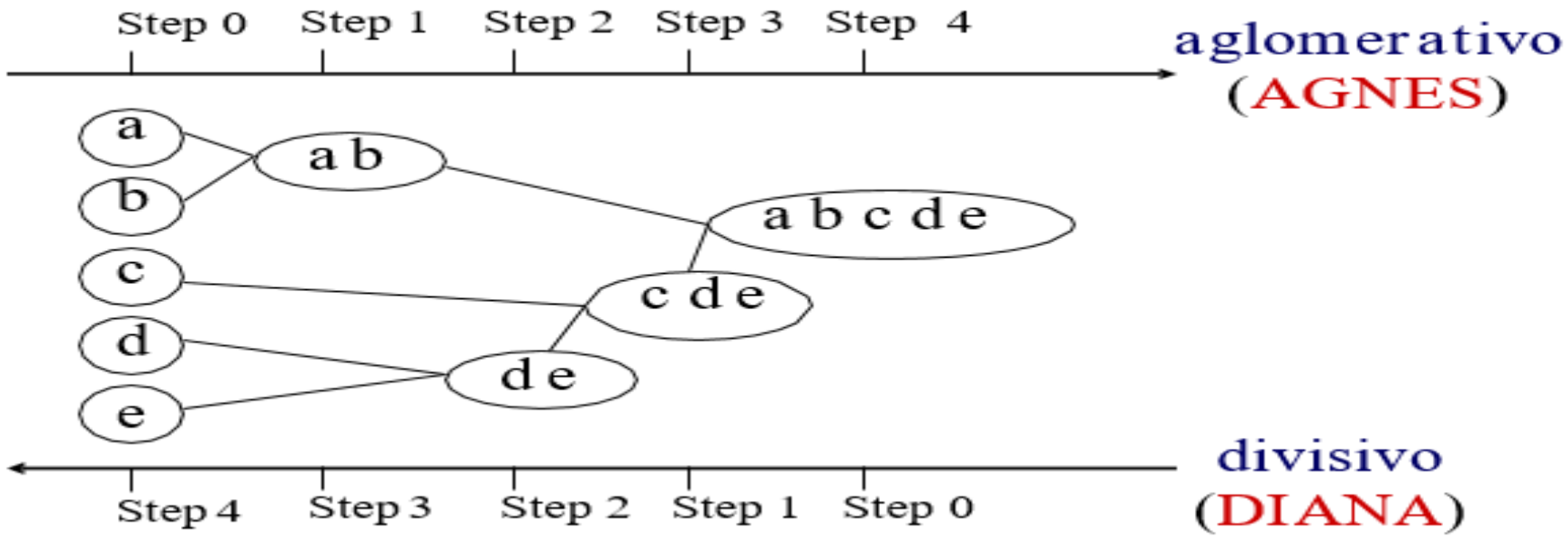
MÉTODOS HIERÁRQUICOS

- **MÉTODOS DIVISIVOS** → Todos Registros → Um "Grande Cluster".
  - Este "Grande Cluster" é dividido em dois ou mais "Clusters" menores até que cada Cluster tenha somente registros semelhantes.

- **MÉTODOS AGLOMERATIVOS** → Cada registro é um "Cluster"
  - A cada passo, combina-se Clusters com alguma característica comum até que se chegue a um "Grande Cluster".

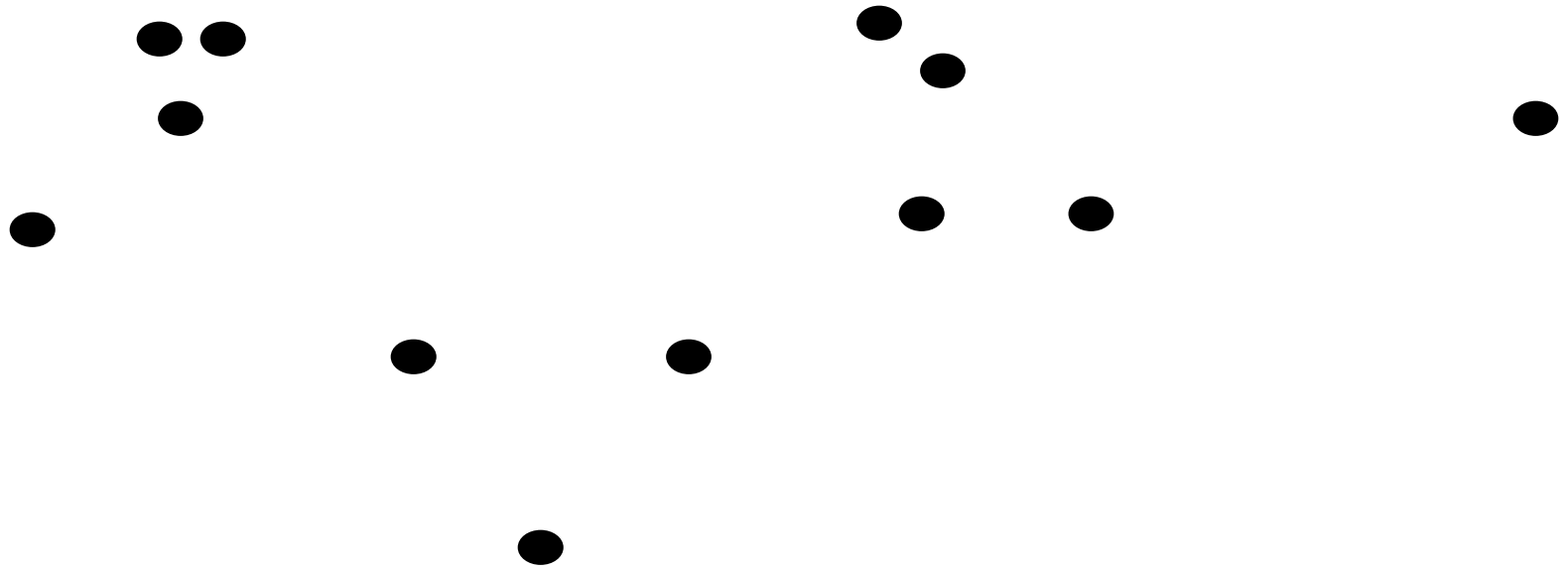
Master

MÉTODOS HIERÁRQUICOS



HIERARCHICAL CLUSTERING

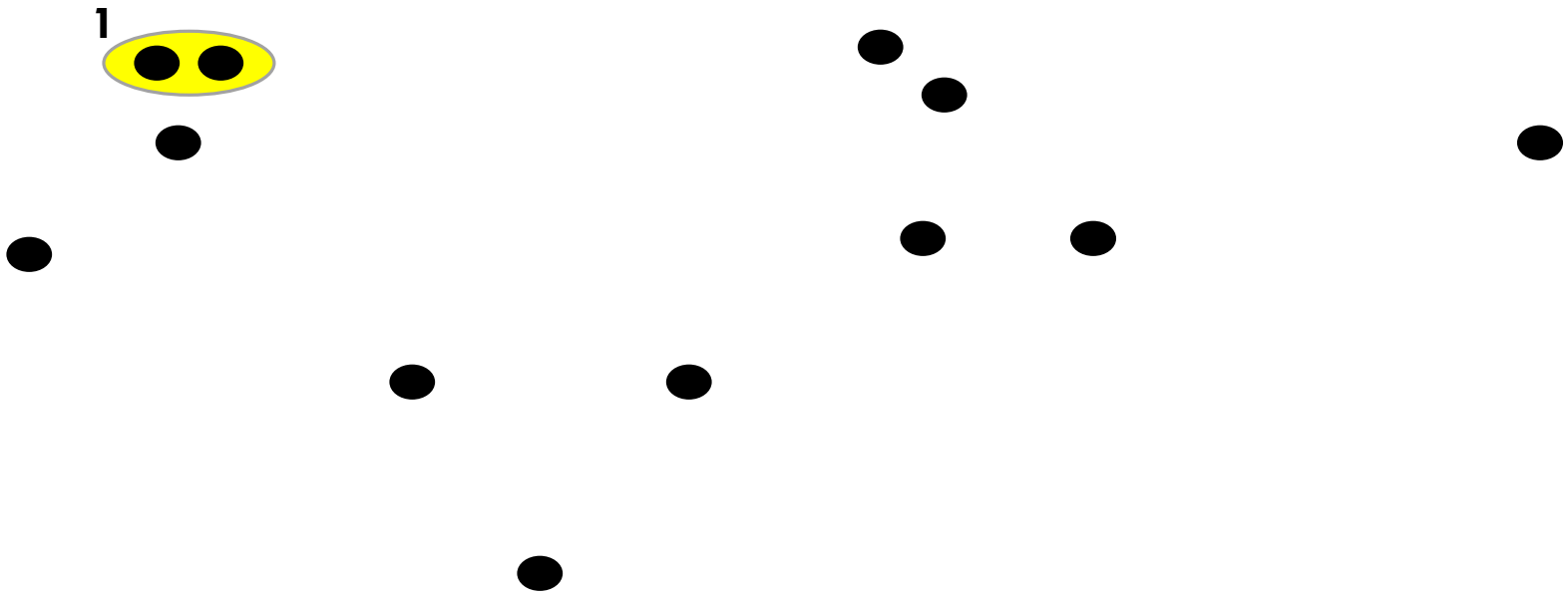
Aglomerativo (Bottom up)



HIERARCHICAL CLUSTERING

Aglomerativo (Bottom up)

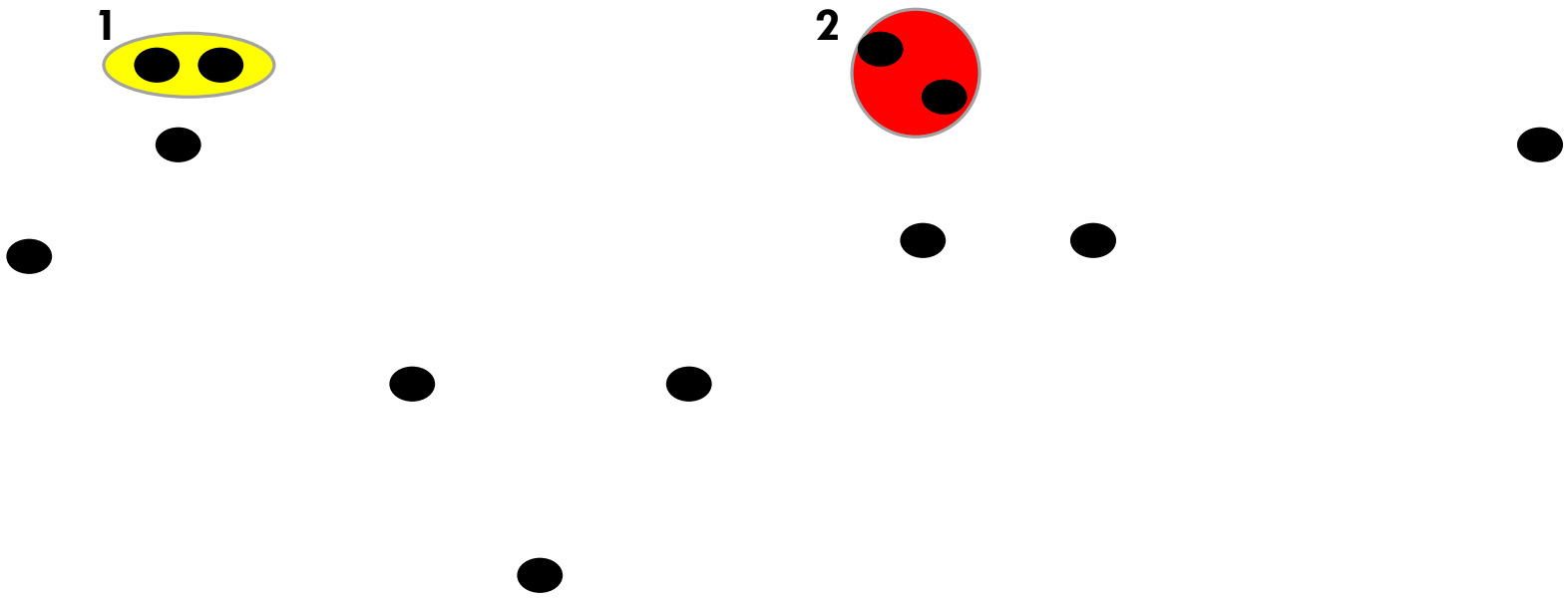
1ª iteração



HIERARCHICAL CLUSTERING

Aglomerativo (Bottom up)

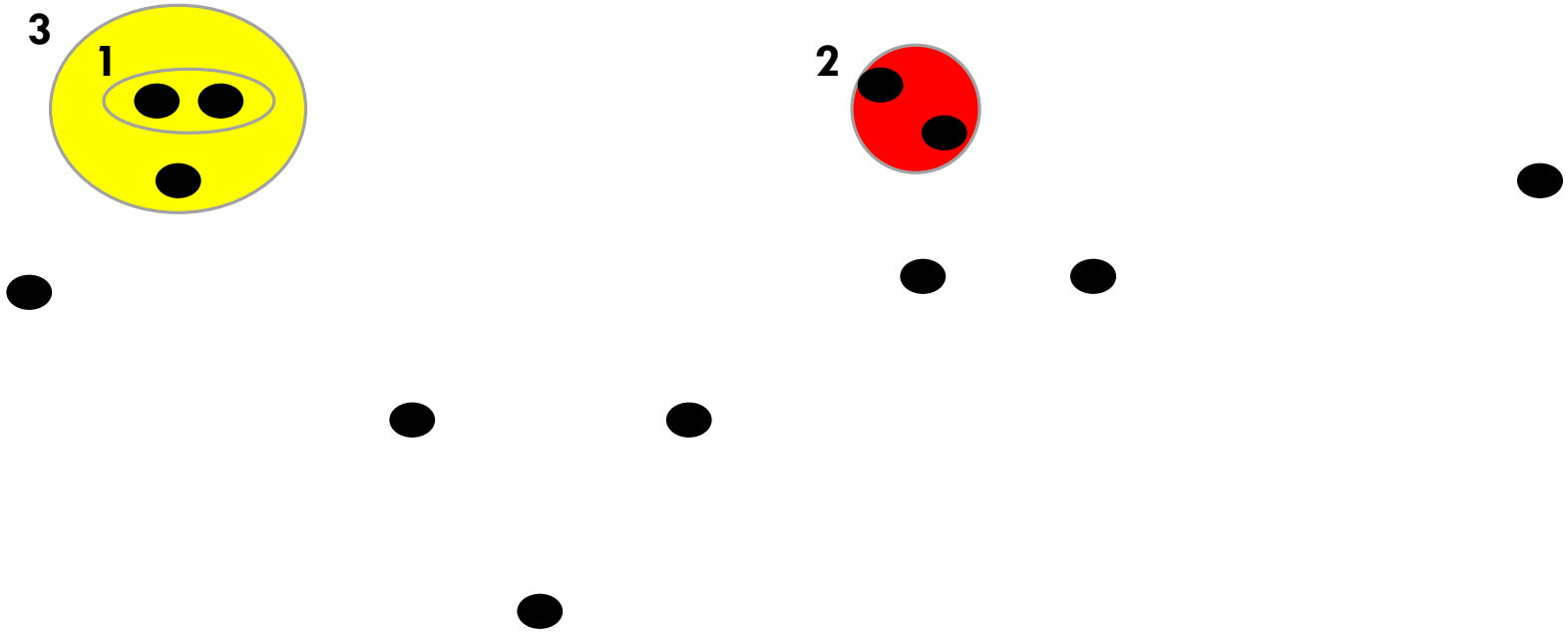
2ª iteração



HIERARCHICAL CLUSTERING

Aglomerativo (Bottom up)

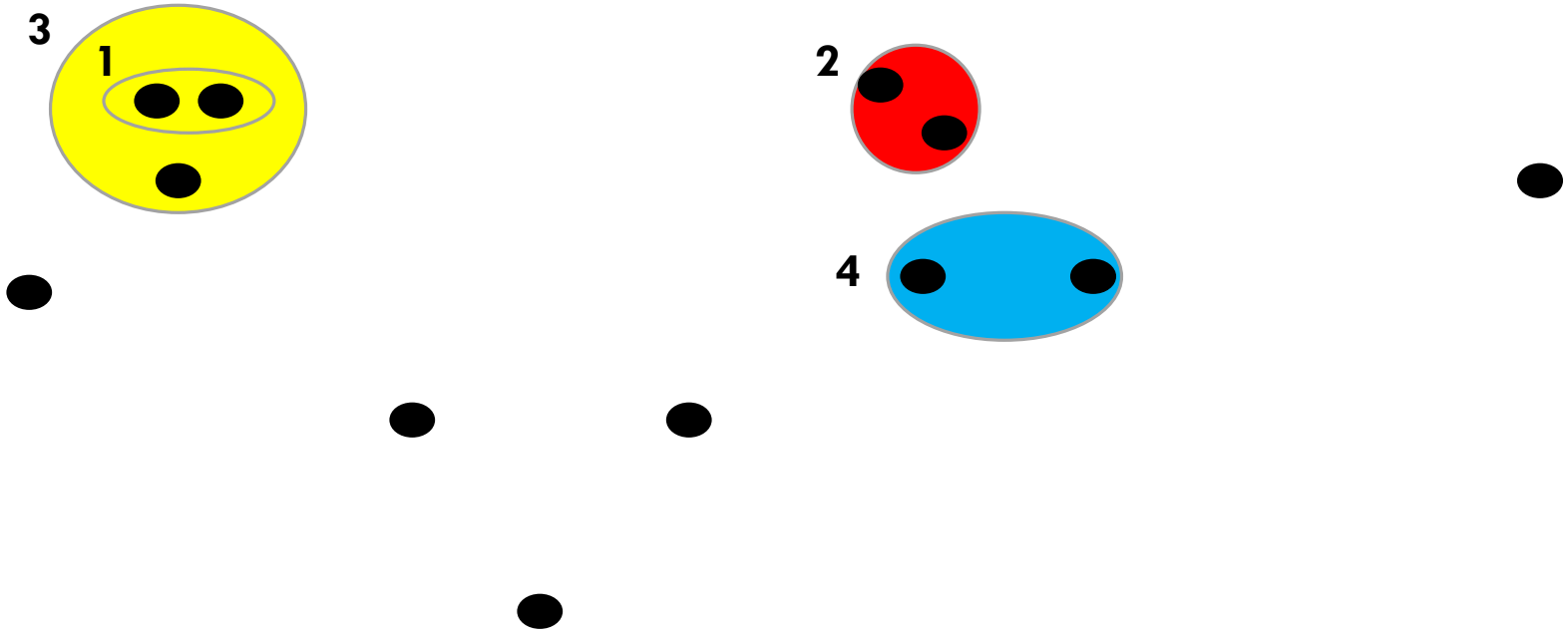
3ª iteração



HIERARCHICAL CLUSTERING

Aglomerativo (Bottom up)

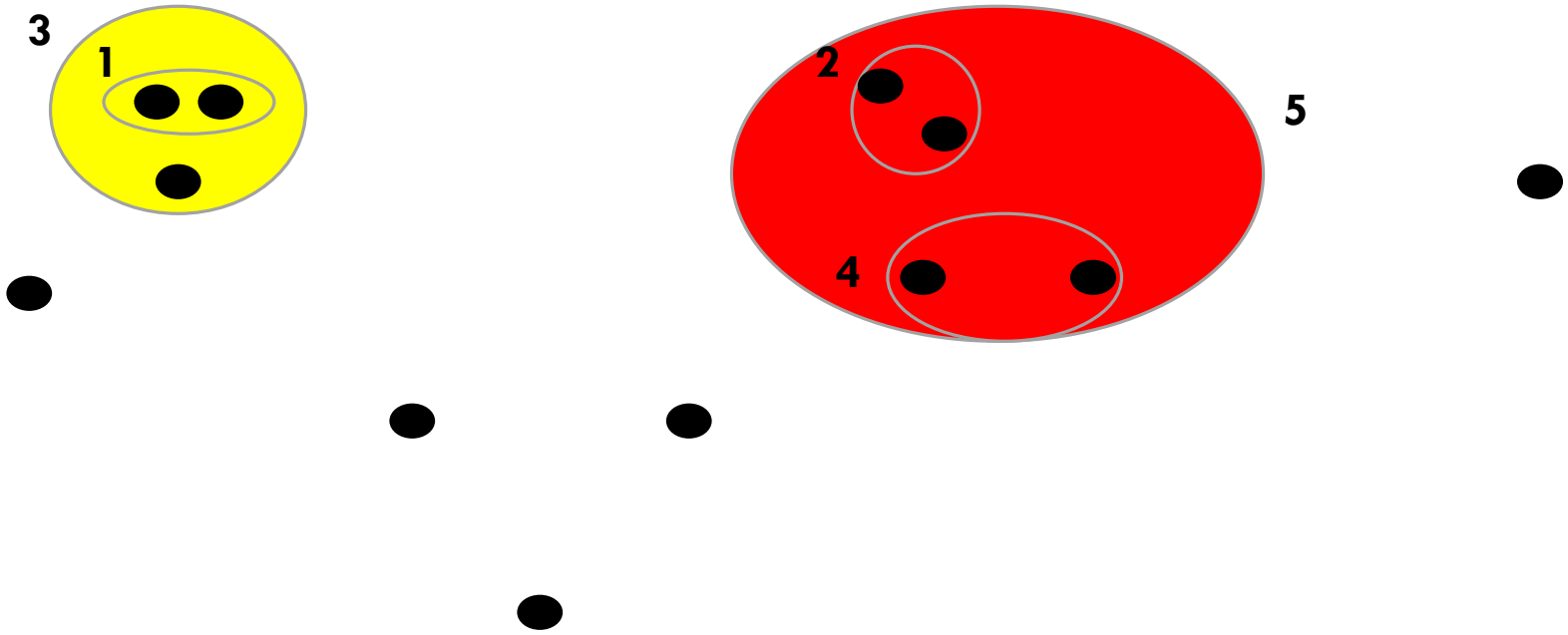
4ª iteração



HIERARCHICAL CLUSTERING

Aglomerativo (Bottom up)

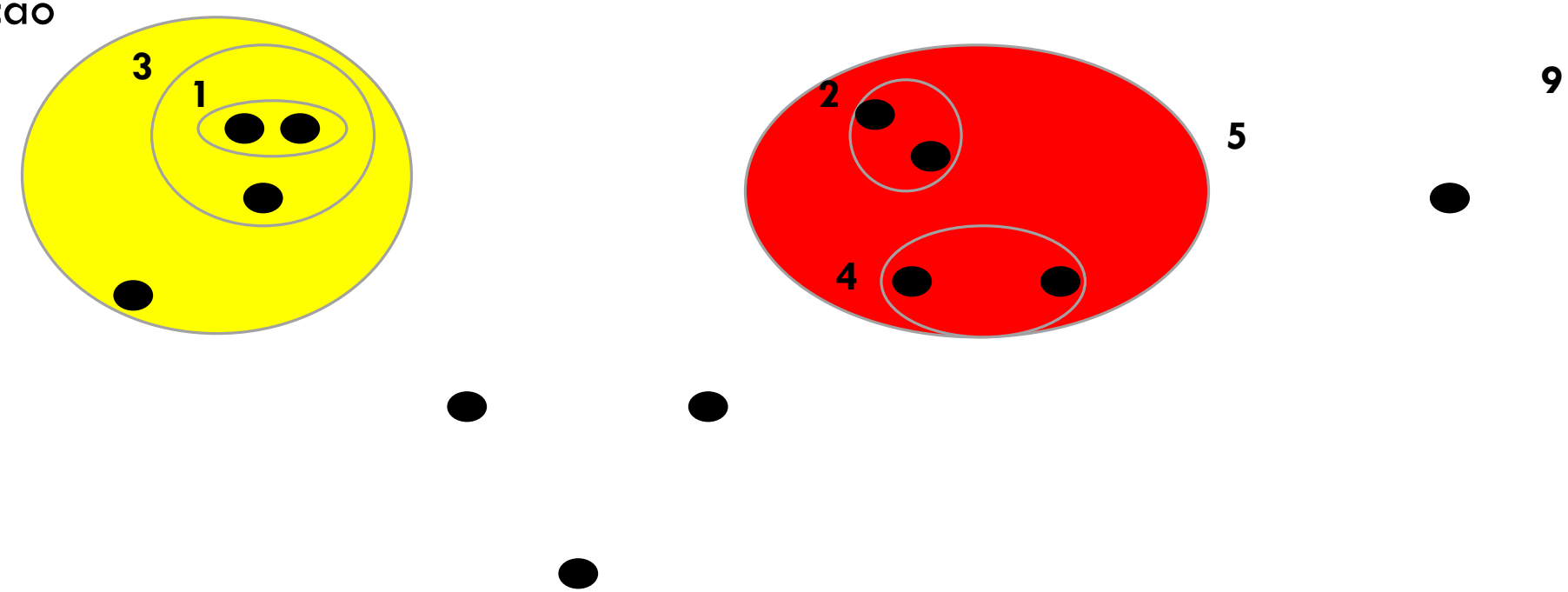
5ª iteração



HIERARCHICAL CLUSTERING

Aglomerativo (Bottom up)

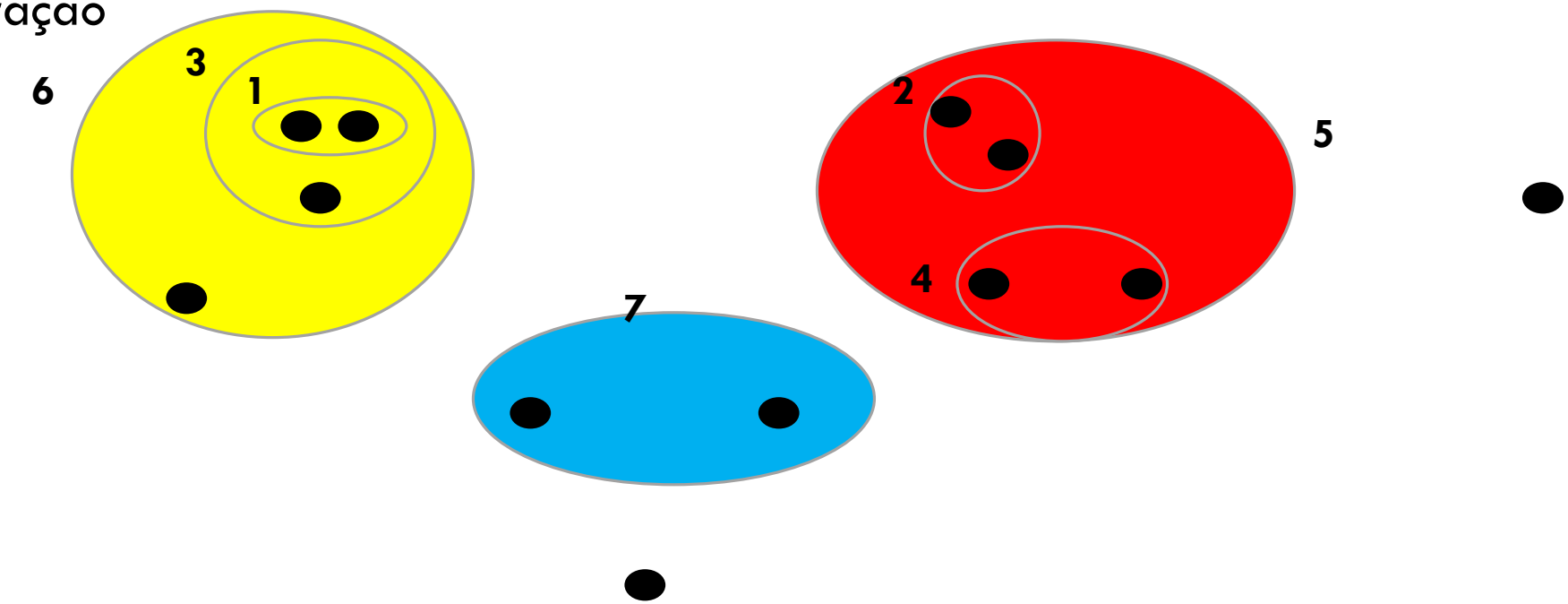
6ª iteração



HIERARCHICAL CLUSTERING

Aglomerativo (Bottom up)

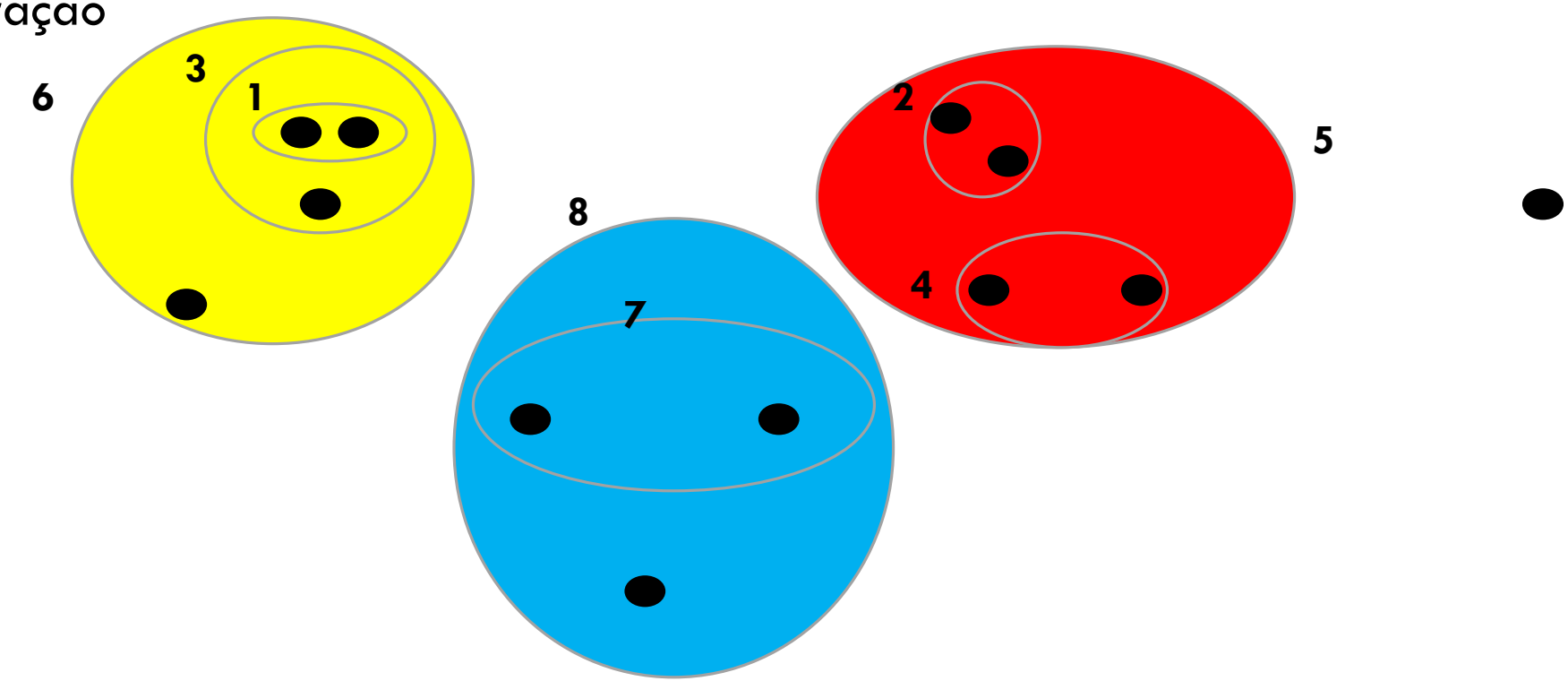
7ª iteração



HIERARCHICAL CLUSTERING

Aglomerativo (Bottom up)

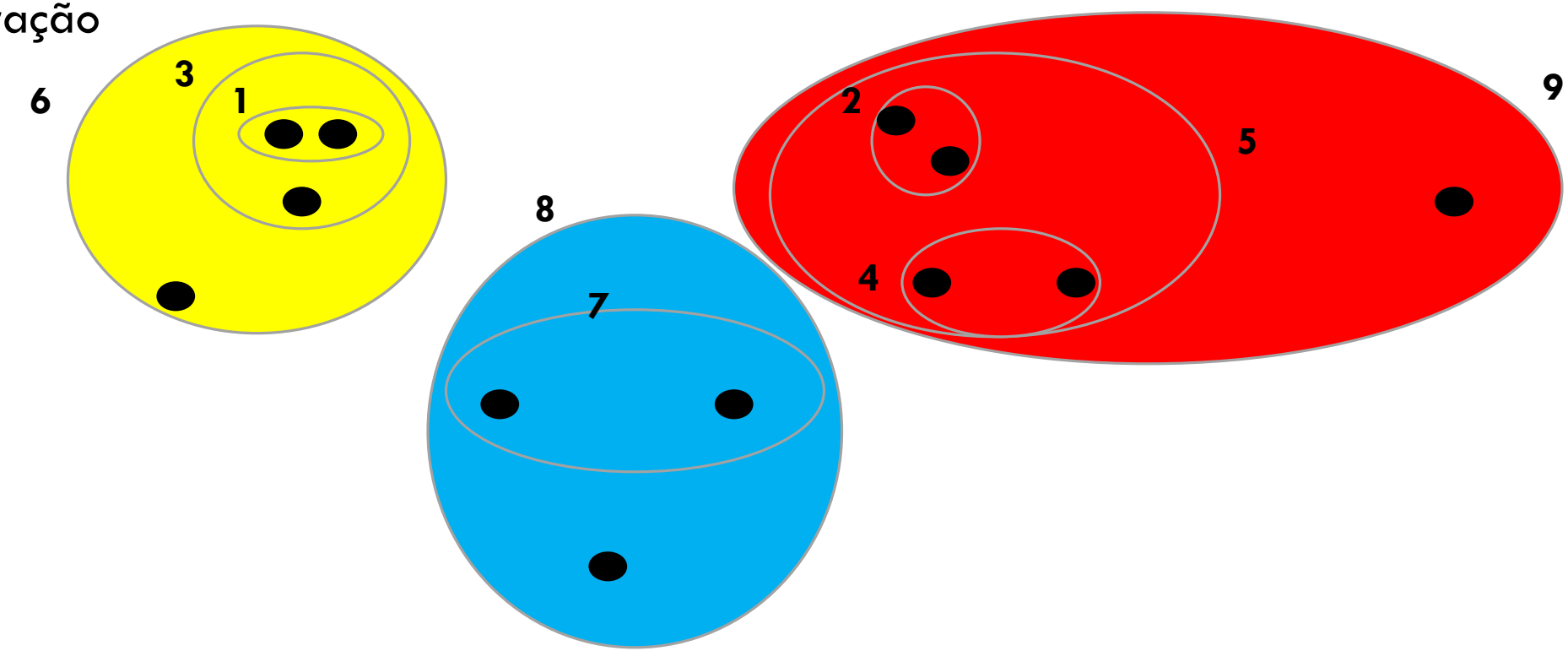
8ª iteração



HIERARCHICAL CLUSTERING

Aglomerativo (Bottom up)

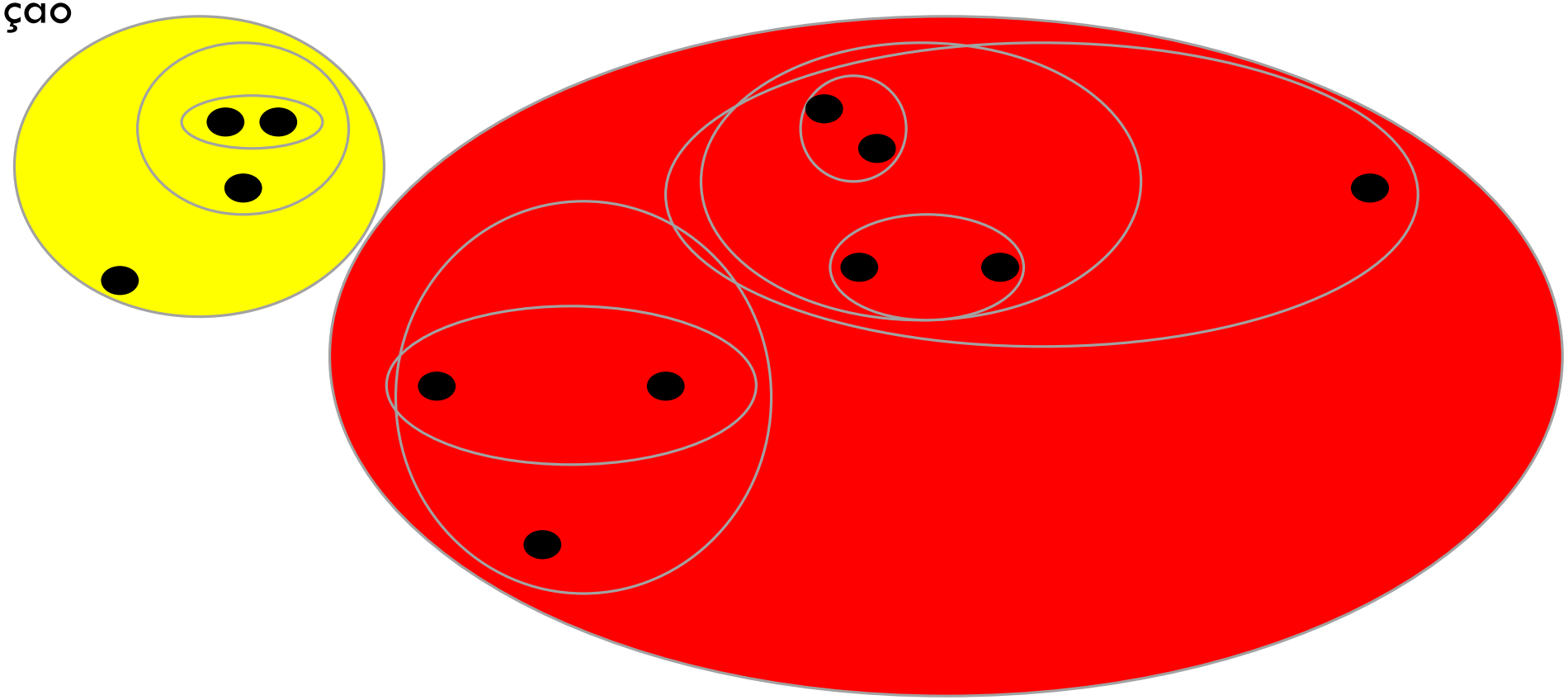
9ª iteração



HIERARCHICAL CLUSTERING

Aglomerativo (Bottom up)

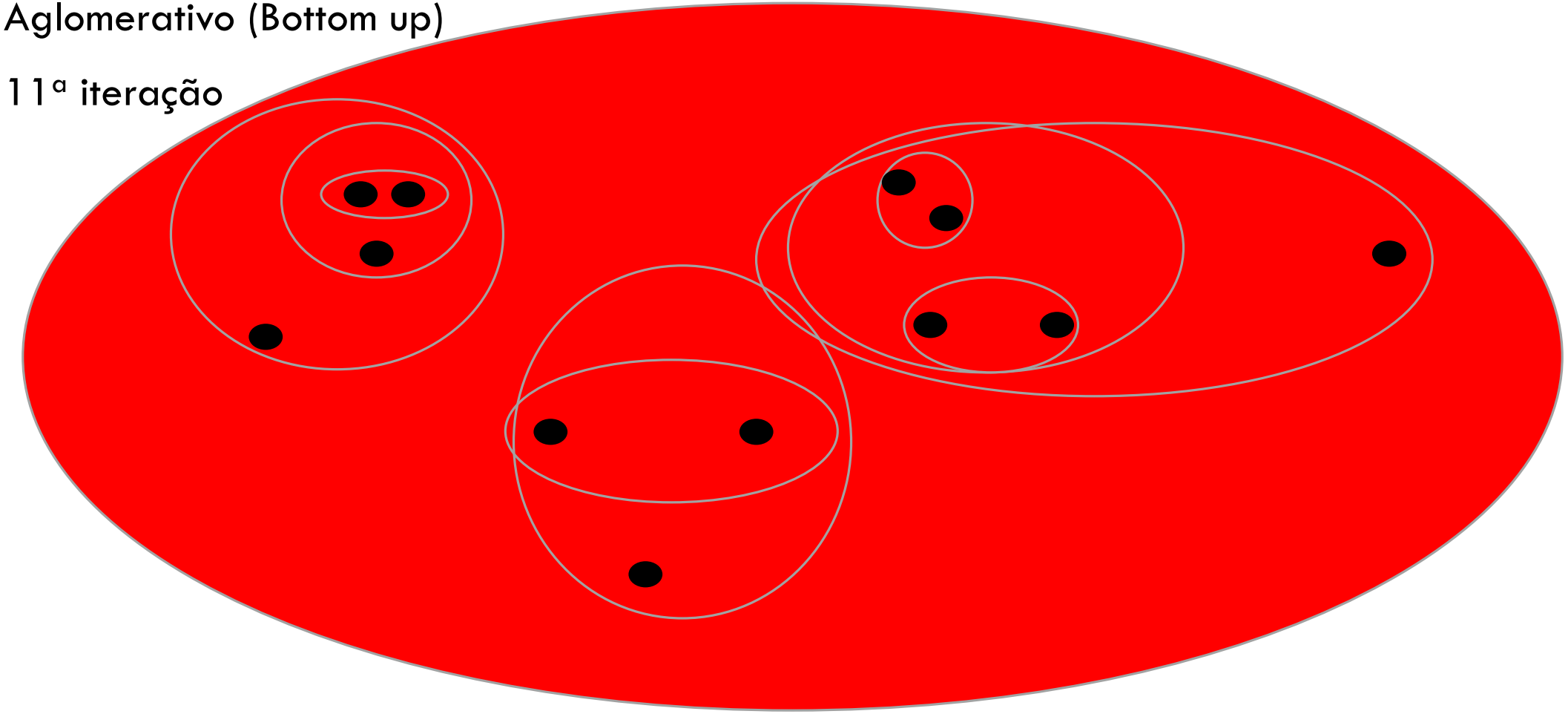
10ª iteração



HIERARCHICAL CLUSTERING

Aglomerativo (Bottom up)

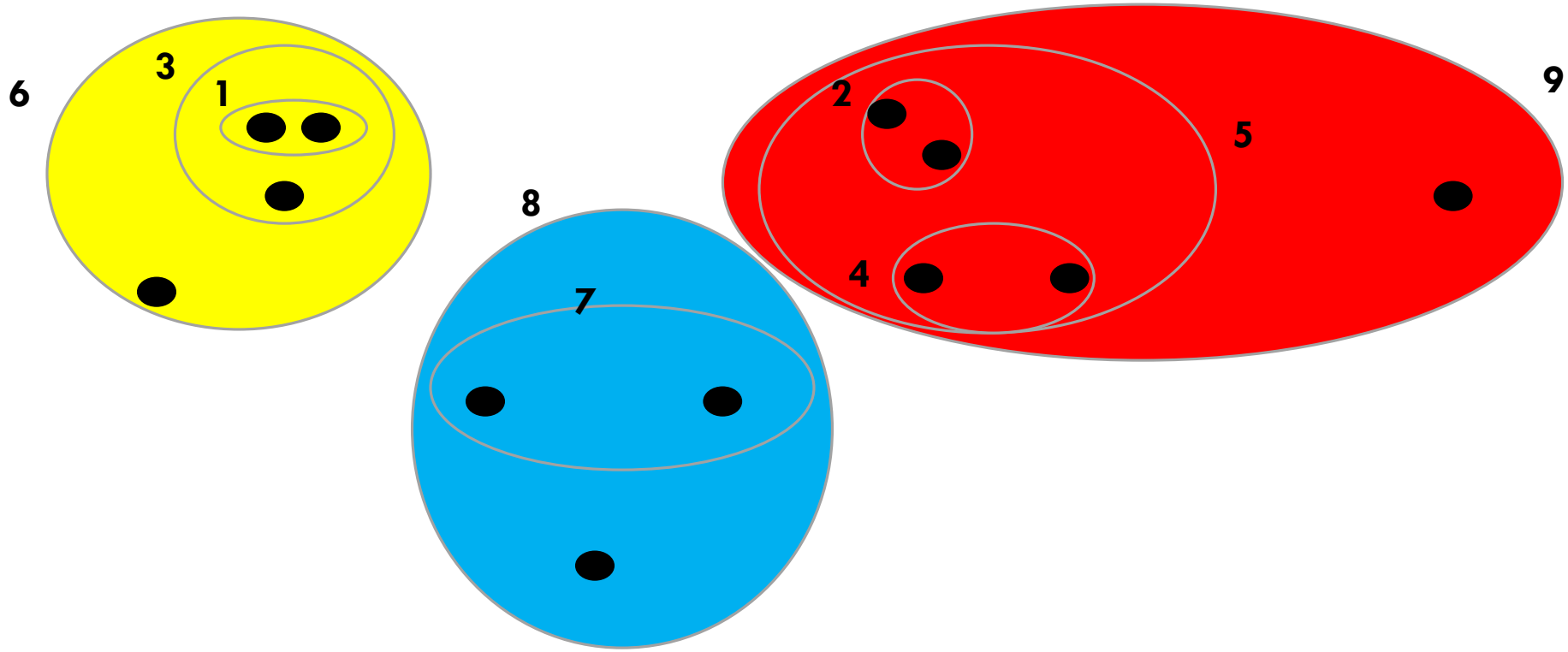
11ª iteração



HIERARCHICAL CLUSTERING

Aglomerativo (Bottom up)

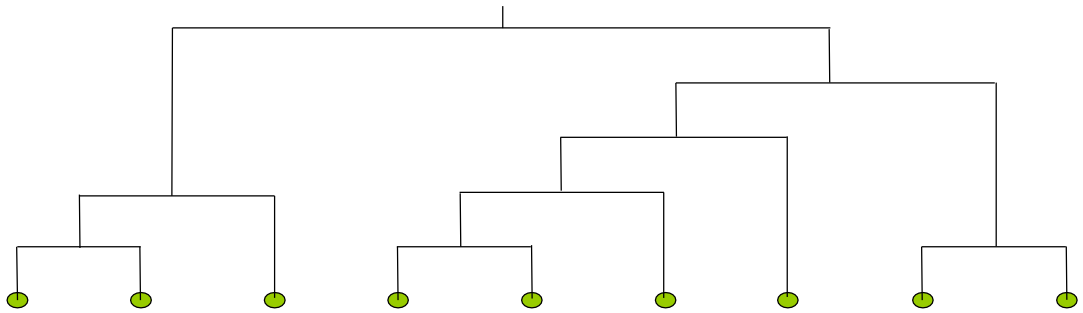
Ponto de corte (com k clusters)



EXEMPLO DE DENDROGRAMA: AGNES

- Decompõe objetos em vários níveis de particionamento aninhados (árvore de clusters), conhecida como dendrograma.

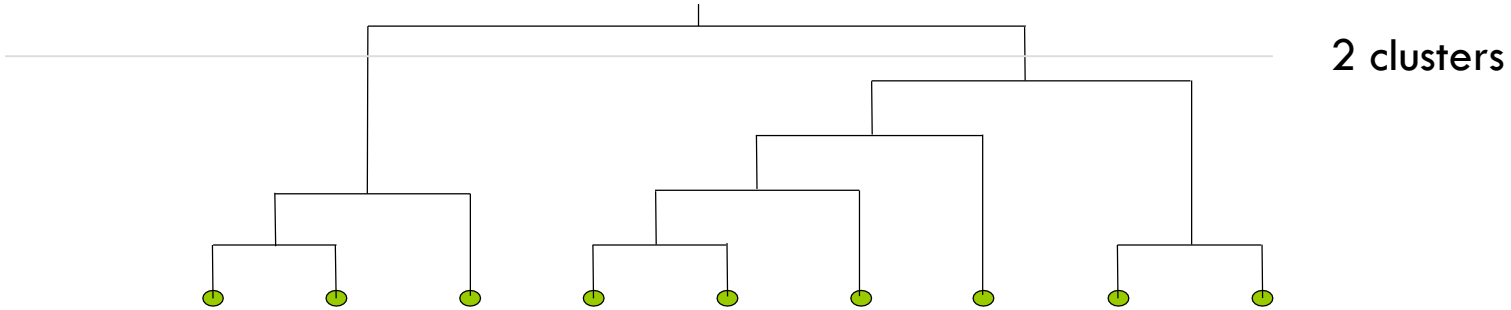
- Uma clusterização dos objetos é obtida particionando-se o dendrograma em um nível desejado. Cada componente conectado forma um cluster.



EXEMPLO DE DENDROGRAMA: AGNES

- Decompõe objetos em vários níveis de particionamento aninhados (árvore de clusters), conhecida como dendrograma.

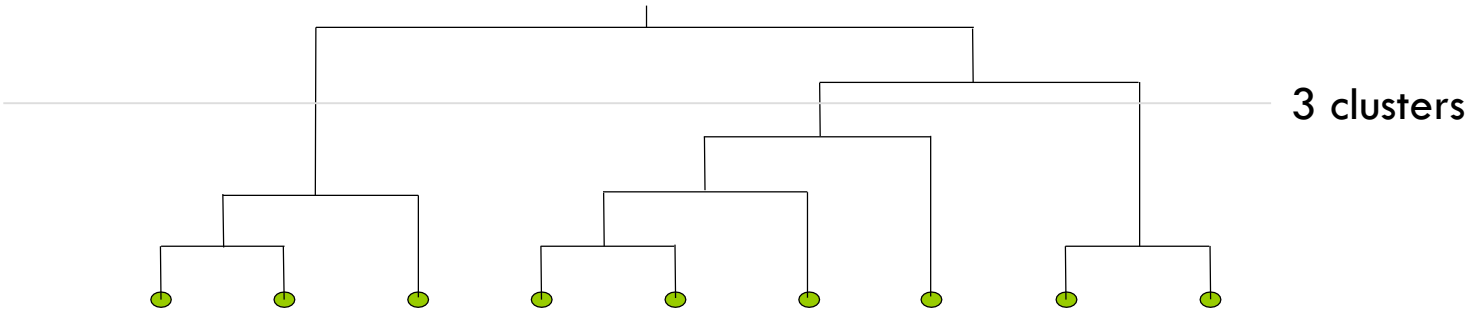
- Uma clusterização dos objetos é obtida particionando-se o dendrograma em um nível desejado. Cada componente conectado forma um cluster.



EXEMPLO DE DENDROGRAMA: AGNES

- Decompõe objetos em vários níveis de particionamento aninhados (árvore de clusters), conhecida como dendrograma.

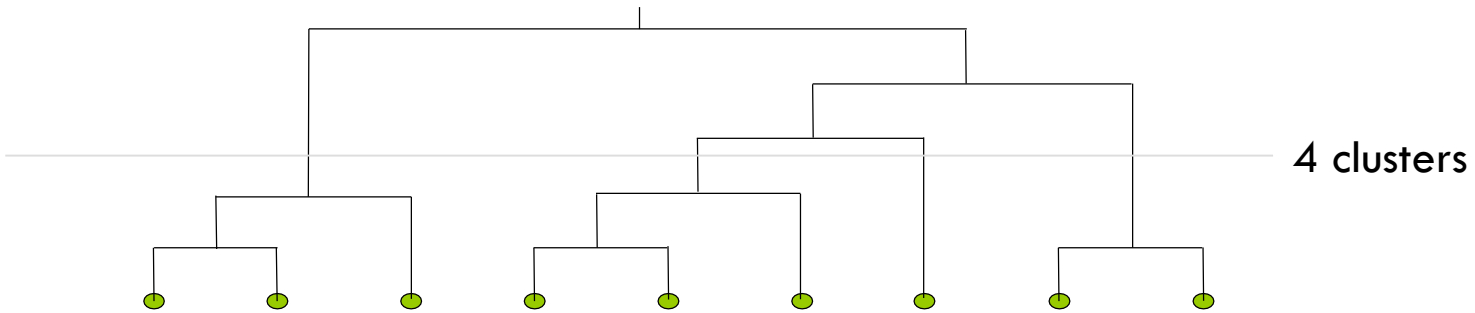
- Uma clusterização dos objetos é obtida particionando-se o dendrograma em um nível desejado. Cada componente conectado forma um cluster.



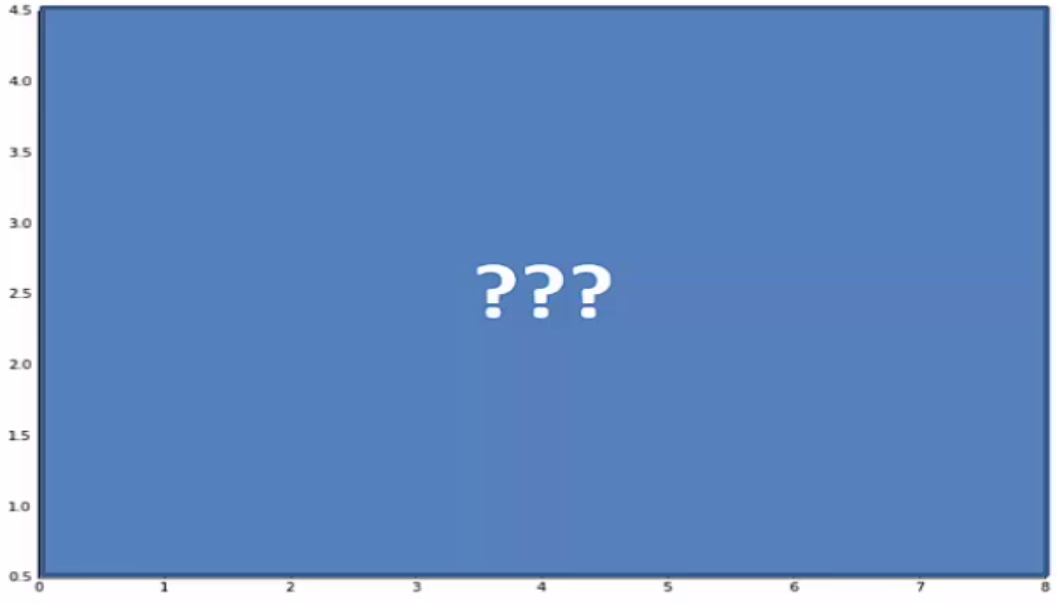
EXEMPLO DE DENDROGRAMA: AGNES

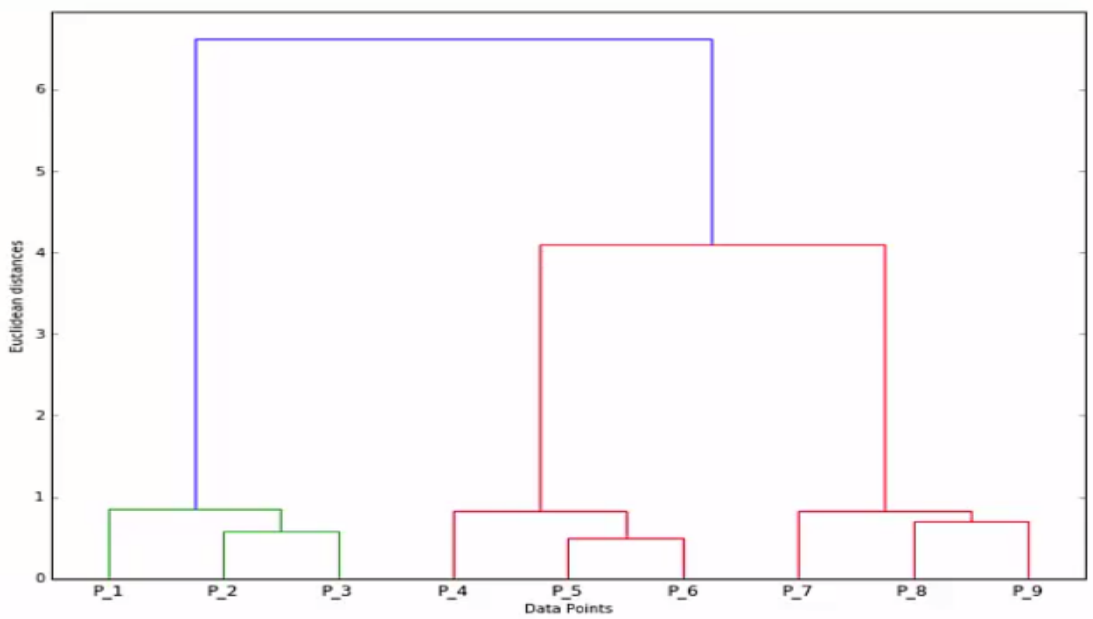
- Decompõe objetos em vários níveis de particionamento aninhados (árvore de clusters), conhecida como dendrograma.

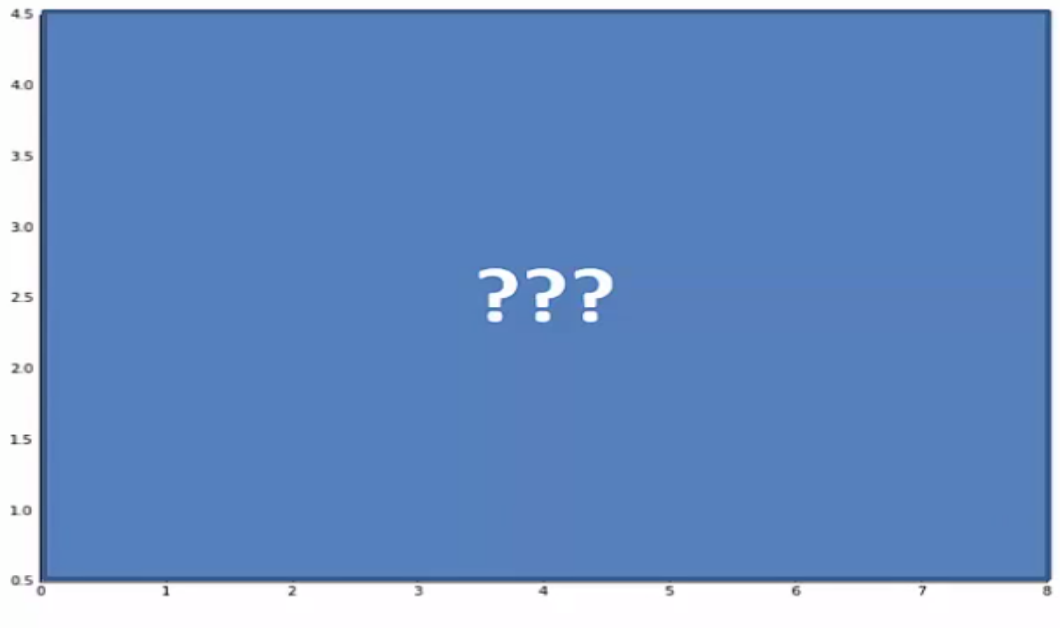
- Uma clusterização dos objetos é obtida particionando-se o dendrograma em um nível desejado. Cada componente conectado forma um cluster.

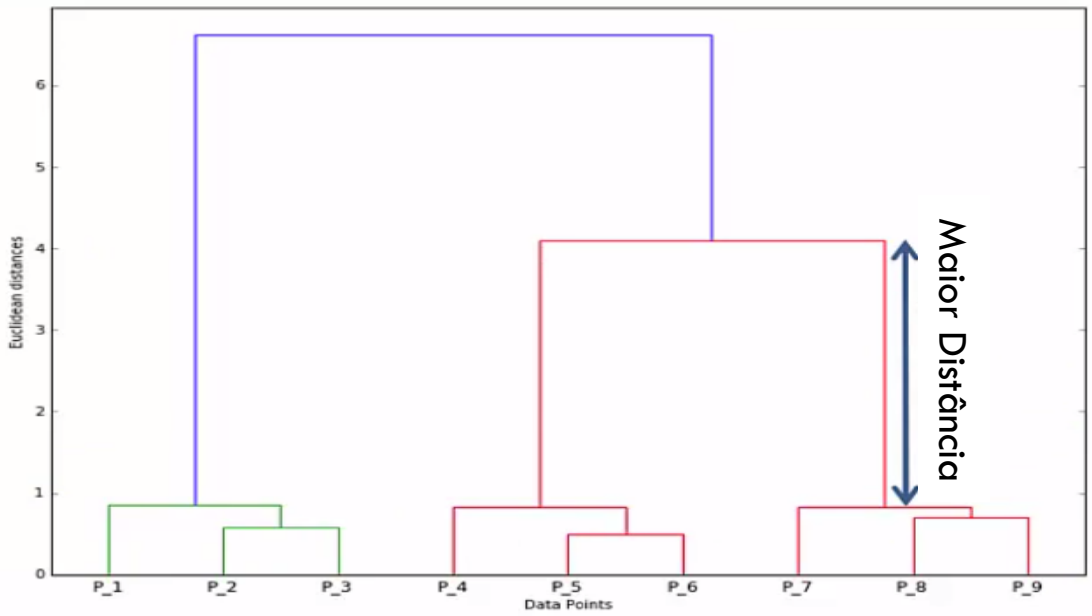


TESTE DE CONHECIMENTO

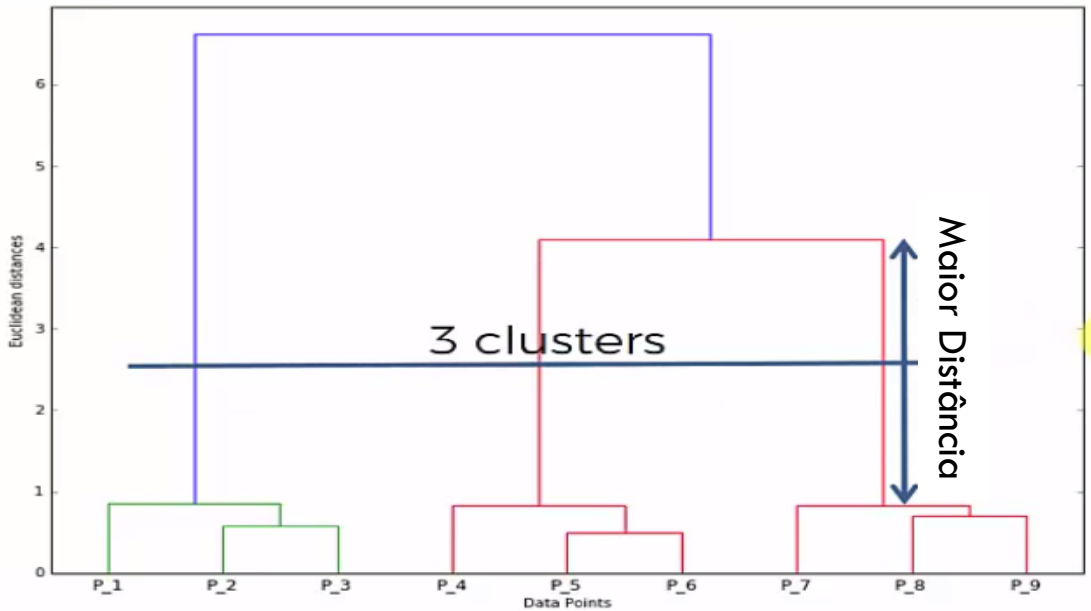


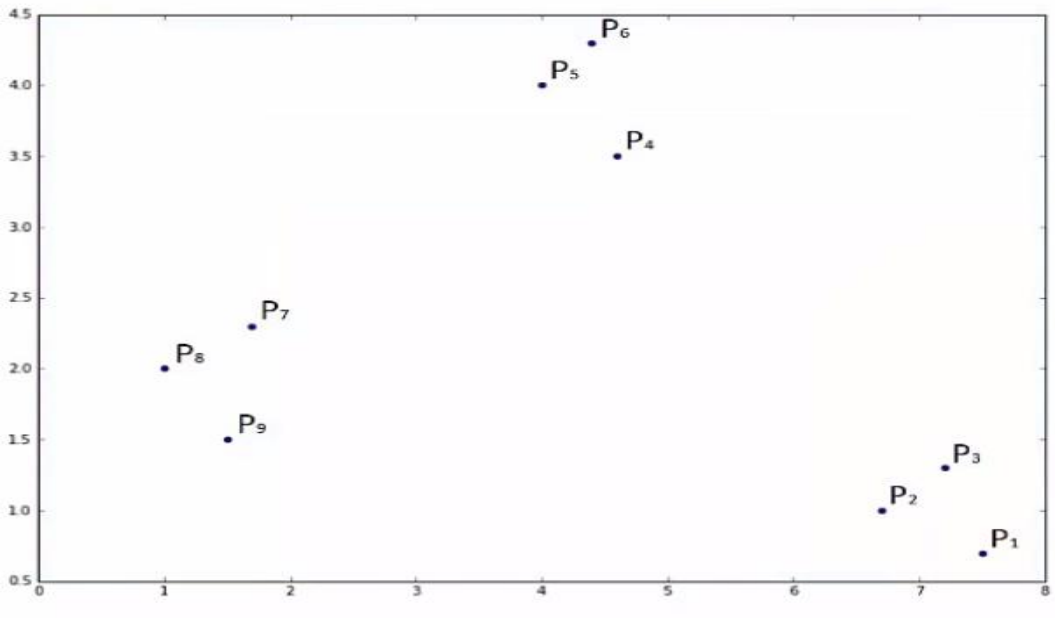


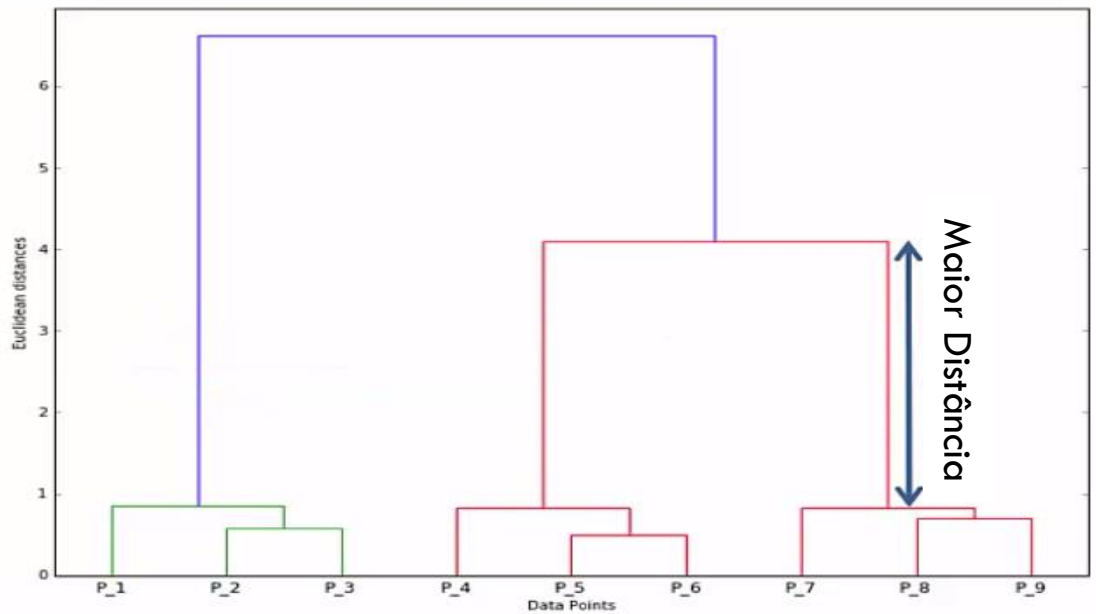

???

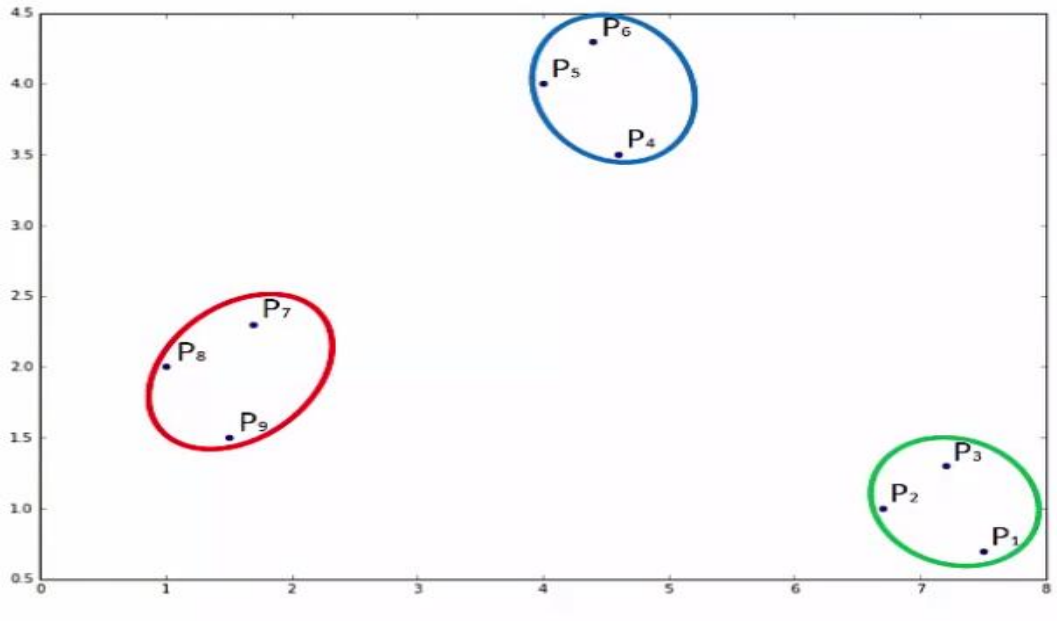

Euclidean distances
Data Points
P_1
P_2
P_3
P_4
P_5
P_6
P_7
P_8
P_9
Maior Distância

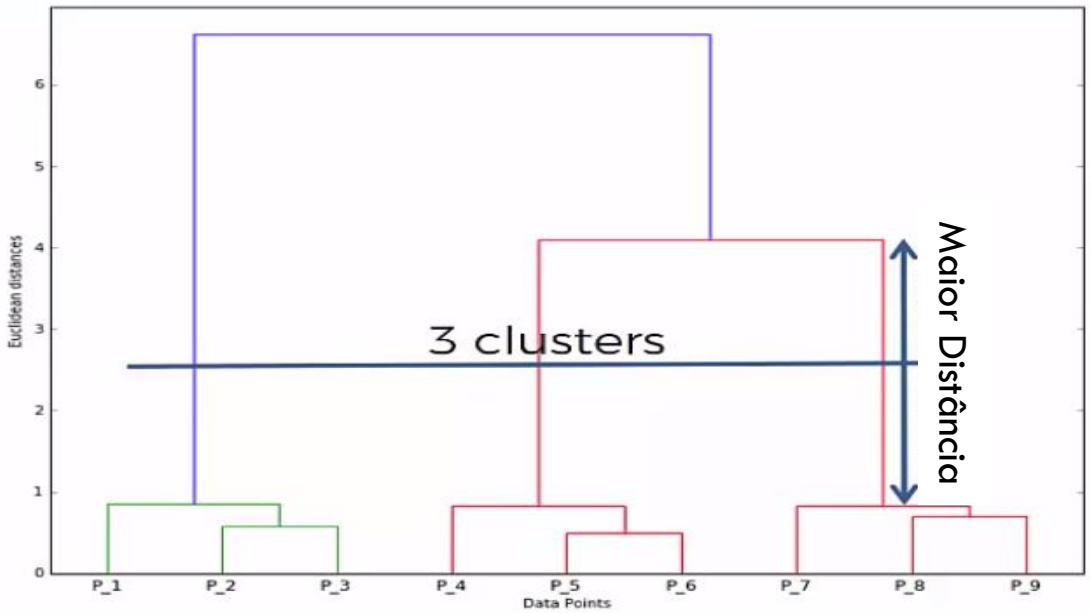

???


Euclidean distances
Data Points
P_1
P_2
P_3
P_4
P_5
P_6
P_7
P_8
P_9
3 clusters
Maior Distância


P6
P5
P4
P7
P8
P9
P3
P2
P1


Euclidean distances
Data Points
P_1
P_2
P_3
P_4
P_5
P_6
P_7
P_8
P_9
Maior Distância

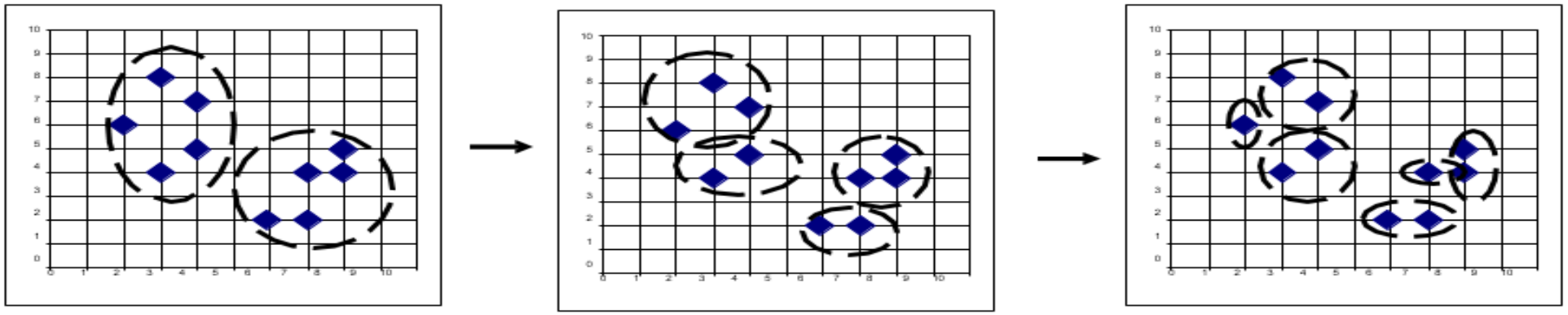

P6
P5
P4
P7
P8
P9
P3
P2
P1

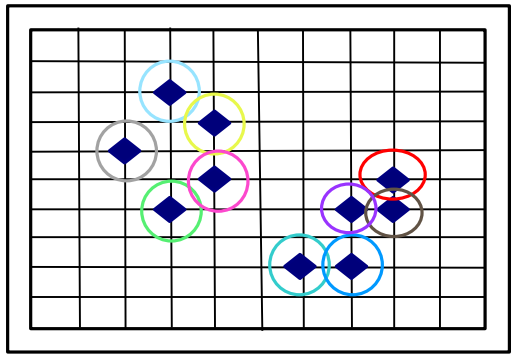

Euclidean distances
3 clusters
Maior Distância
P_1
P_2
P_3
P_4
P_5
P_6
P_7
P_8
P_9
Data Points

DIANA (DIVISIVE ANALYSIS)

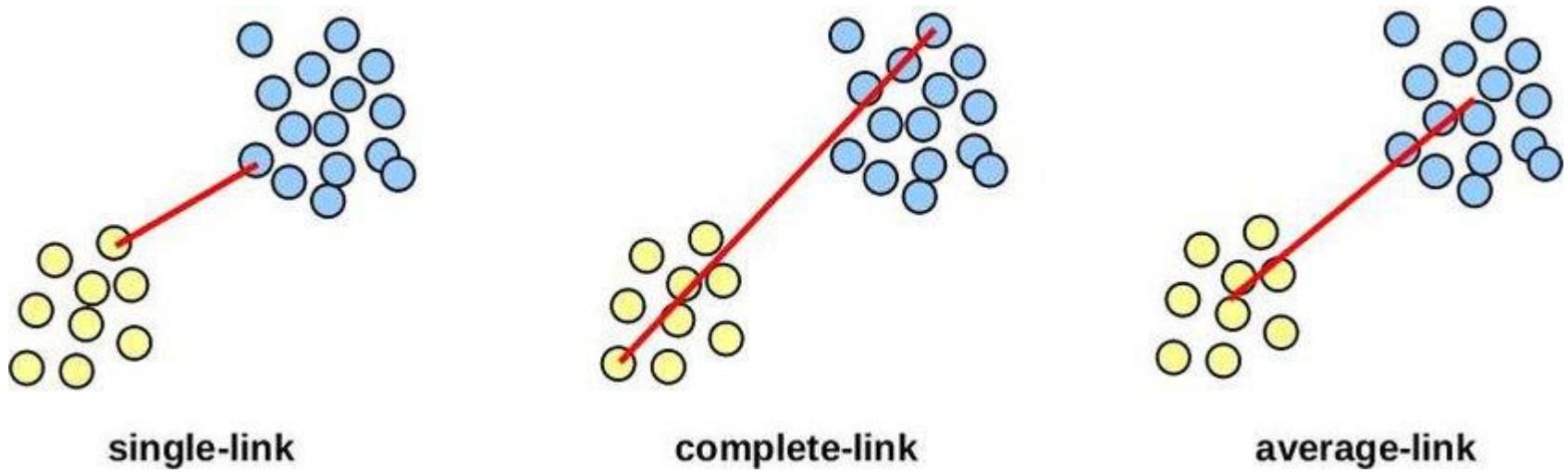
Procedimento: o inverso de AGNES.

Eventualmente cada nó forma um cluster.





OBS



K-MEANS VS. HIERÁRQUICO

Modelo de	Prós	Contras
K-means	Simples de entender; Trabalha bem em bases pequenas e grandes; Rápido e eficiente.	É preciso passar como parâmetro o número de clusters.
Hierárquico	O número ótimo de clusters pode ser obtido pelo próprio modelo; Visualização prática através de um dendrograma.	Não é apropriado para bases muito grandes.

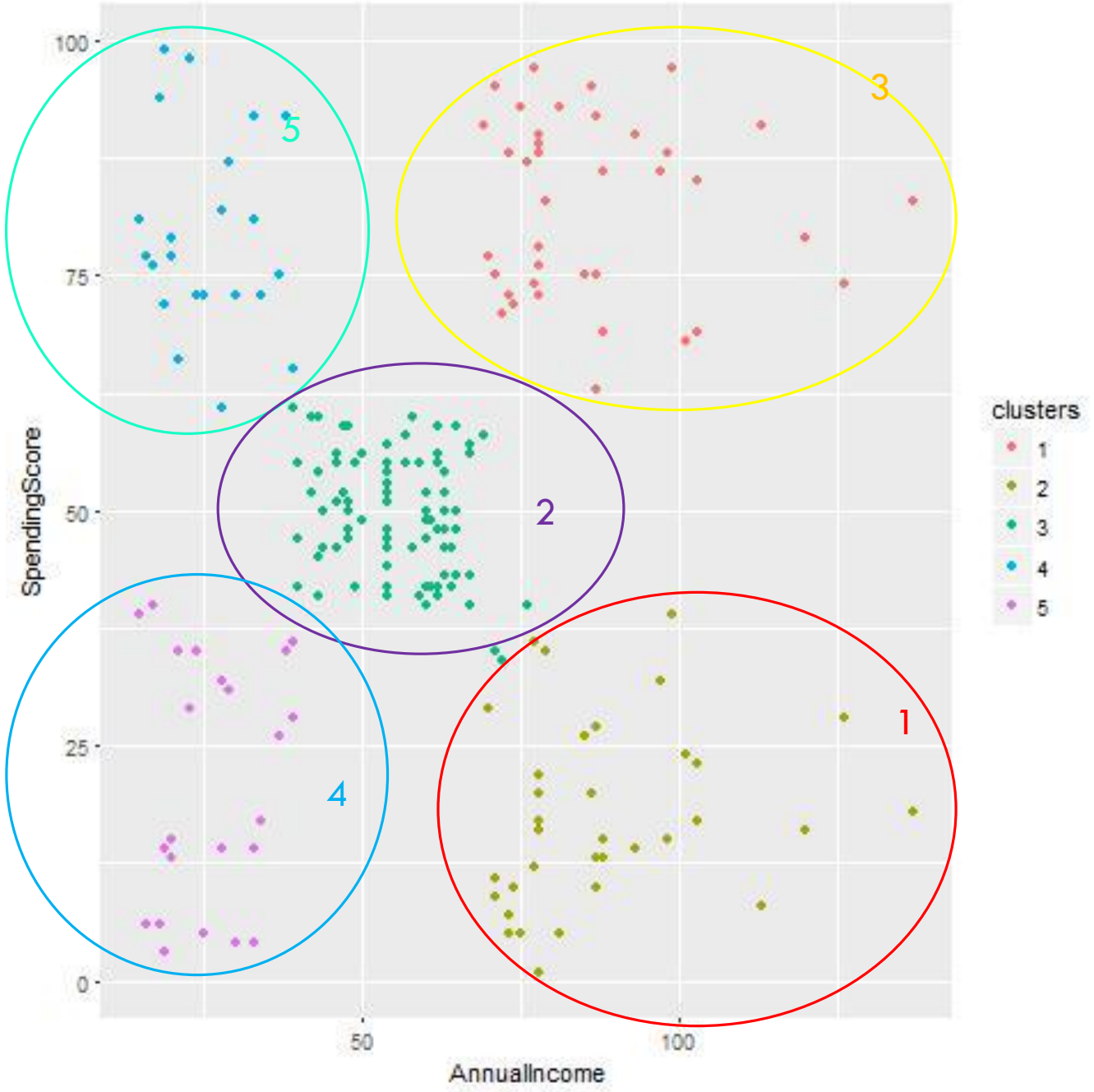
ESTUDO DE CASO

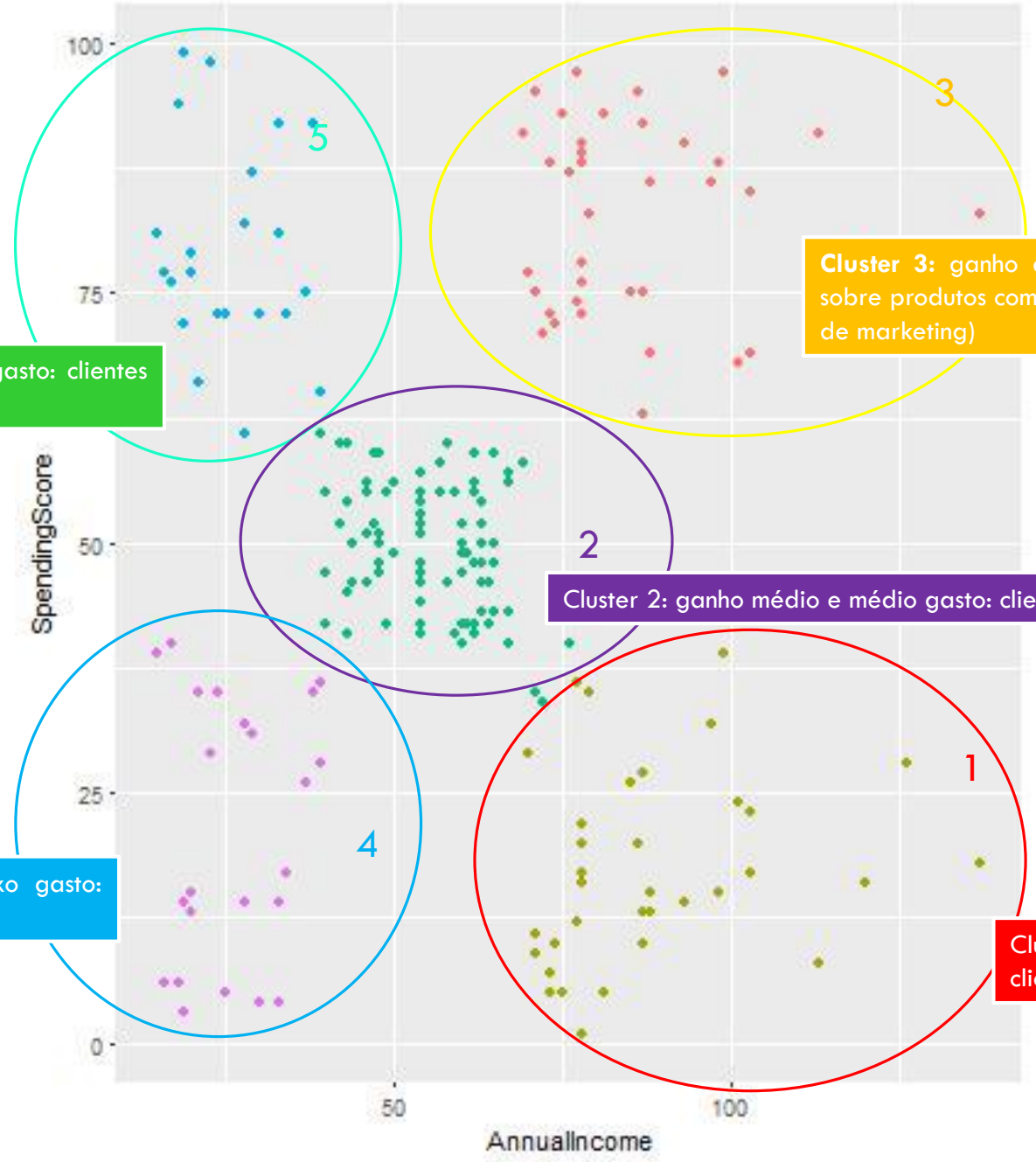
ESTUDO DE CASO

Clientes de um Shopping

- Ganho Anual
- Traço de Gastos

ANÁLISE DA CLUSTERIZAÇÃO


100
75
50
25
0
SpendingScore
50
100
AnnualIncome
5
3
2
4
1
clusters
1
2
3
4
5



Cluster 5: ganho baixo e alto gasto: clientes pouco cuidadosos

Cluster 3: ganho alto e alto gasto: clientes alvo (deve-se entender melhor sobre produtos comprados por esses clientes e direcionar melhor as campanhas de marketing)

Cluster 2: ganho médio e médio gasto: clientes padrão

Cluster 4: ganho baixo e baixo gasto: clientes sensíveis

Cluster 1: ganho alto e baixo gasto: clientes cuidadosos

CLUSTERIZAÇÃO BASEADA EM DENSIDADE

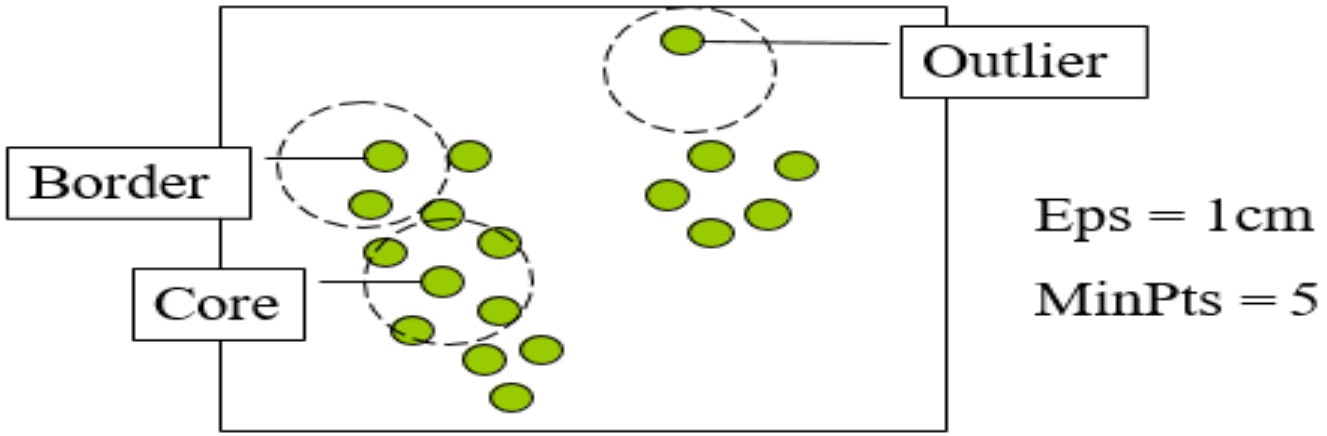
DBSCAN - MÉTODO BASEADO EM DENSIDADE

- **DBSCAN é um algoritmo baseado em densidade.**
  - Densidade = número de pontos dentro de um raio específico (*Eps*).
  - Um “border point” fica localizado na vizinhança de um “core point”.
  - Um “core point” tem um número mínimo de pontos especificados pelo usuário (*MinPts*) dentro do raio (*Eps).*
  - Um “noise point” é qualquer ponto que não se classifica como “core point” nem como “border point”.

BI Master

DBSCAN – IDEIA GERAL

Ideia: Um cluster é definido como um conjunto máximo de pontos densamente conectados.



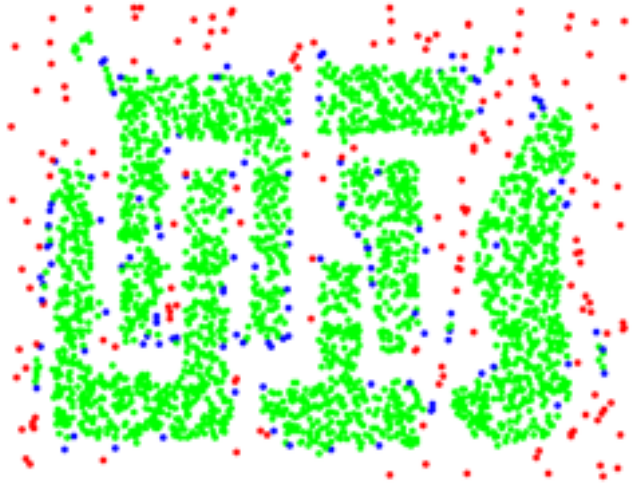
O ALGORITMO DBSCAN

- Arbitrariamente, seleciona um ponto *p.*
- Identifica todos os pontos densamente conectados a *p* com relação aos parâmetros *Eps* e *MinPts*.
- Se *p* é um “core point”, um cluster é formado.
- Se *p* é um “border point” e não há pontos densamente conectados a *p*, DBSCAN visita o próximo ponto do conjunto de dados.
- Continua o processo até que todos os pontos do conjunto de dados tenham sido analisados.

BI Master

DBSCAN: CORE, BORDER E NOISE POINTS

Pontos Originais

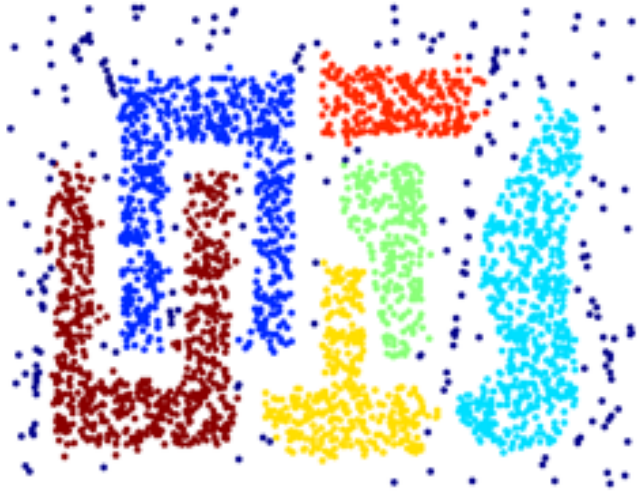


Tipos de pontos: core, border e noise

Eps = 10, MinPts = 4

QUANDO DBSCAN FUNCIONA BEM?

Pontos Originais



Clusters

- Na presença de ruídos (Noise)
- Na geração de clusters com diferentes formatos e tamanhos.

Master

MUNTAZ & DURAISWAMY (2010)