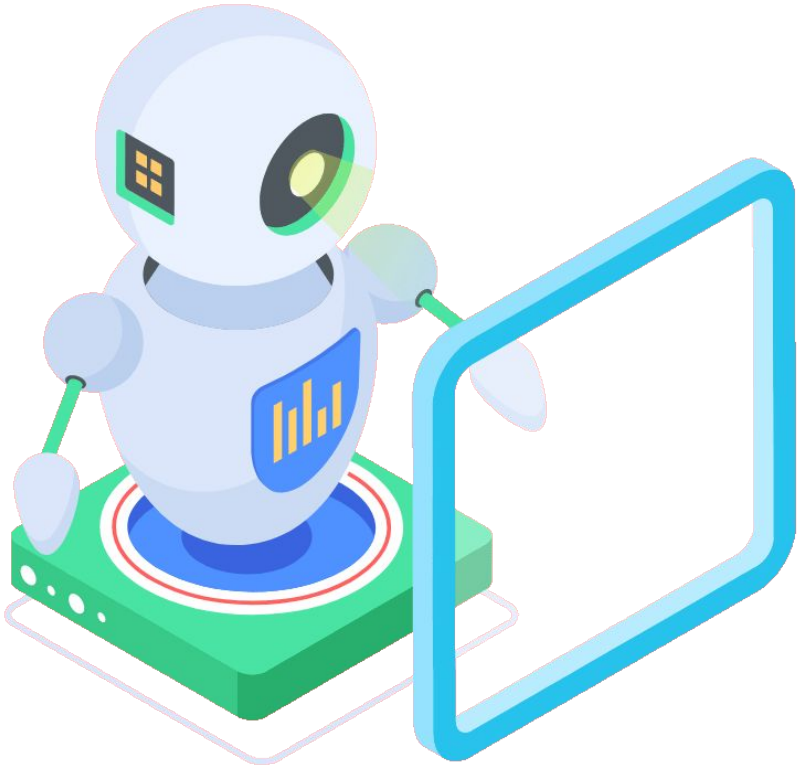
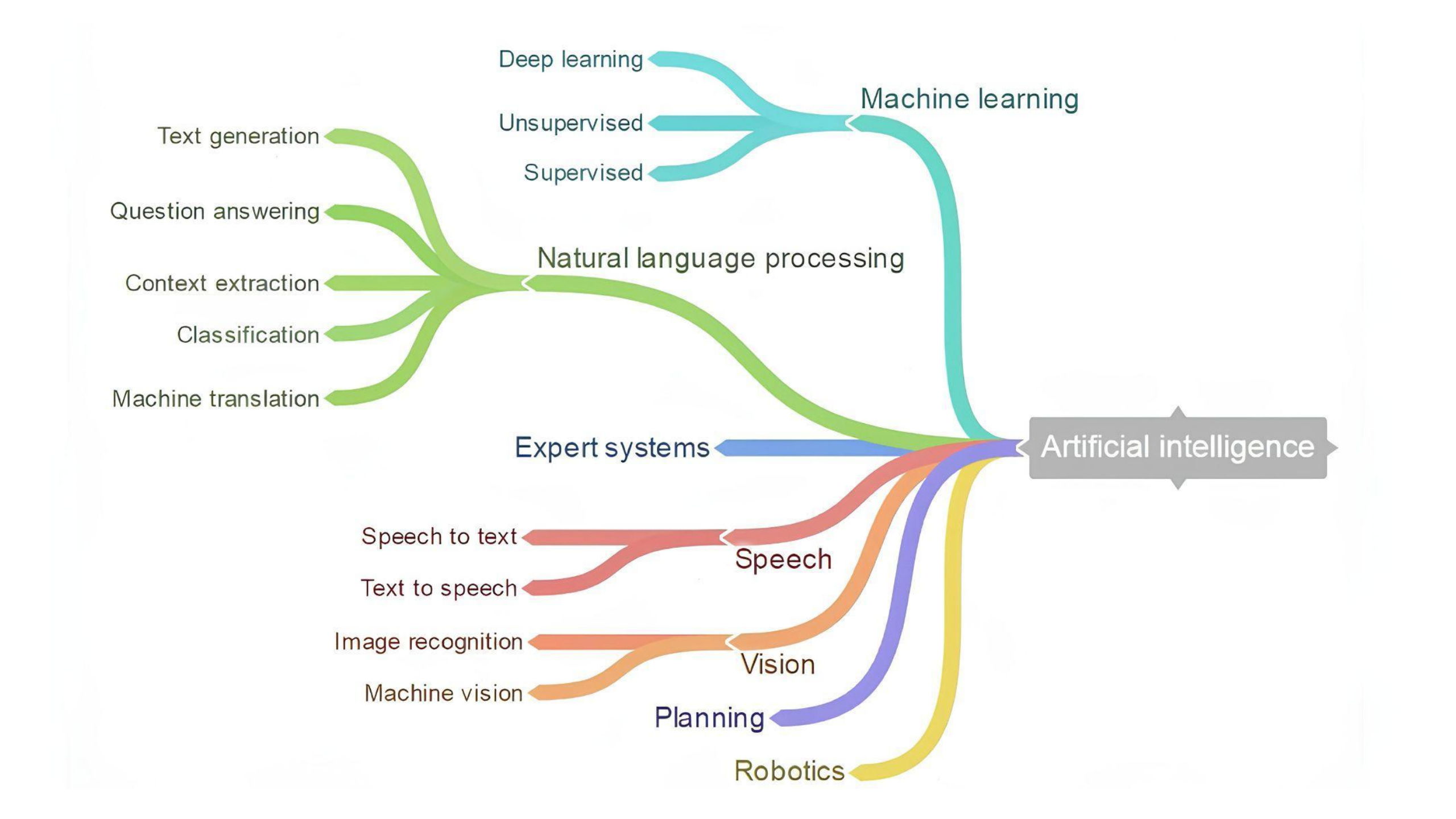
Machine Learning: uma introdução simples ao **aprendizado Não Supervisionado com *clustering(K-means*)**

O que é *Machine Learning*?



Machine learning é uma área de pesquisa no campo da Inteligência Artificial, voltada a ensinar máquinas a aprender a partir de dados.

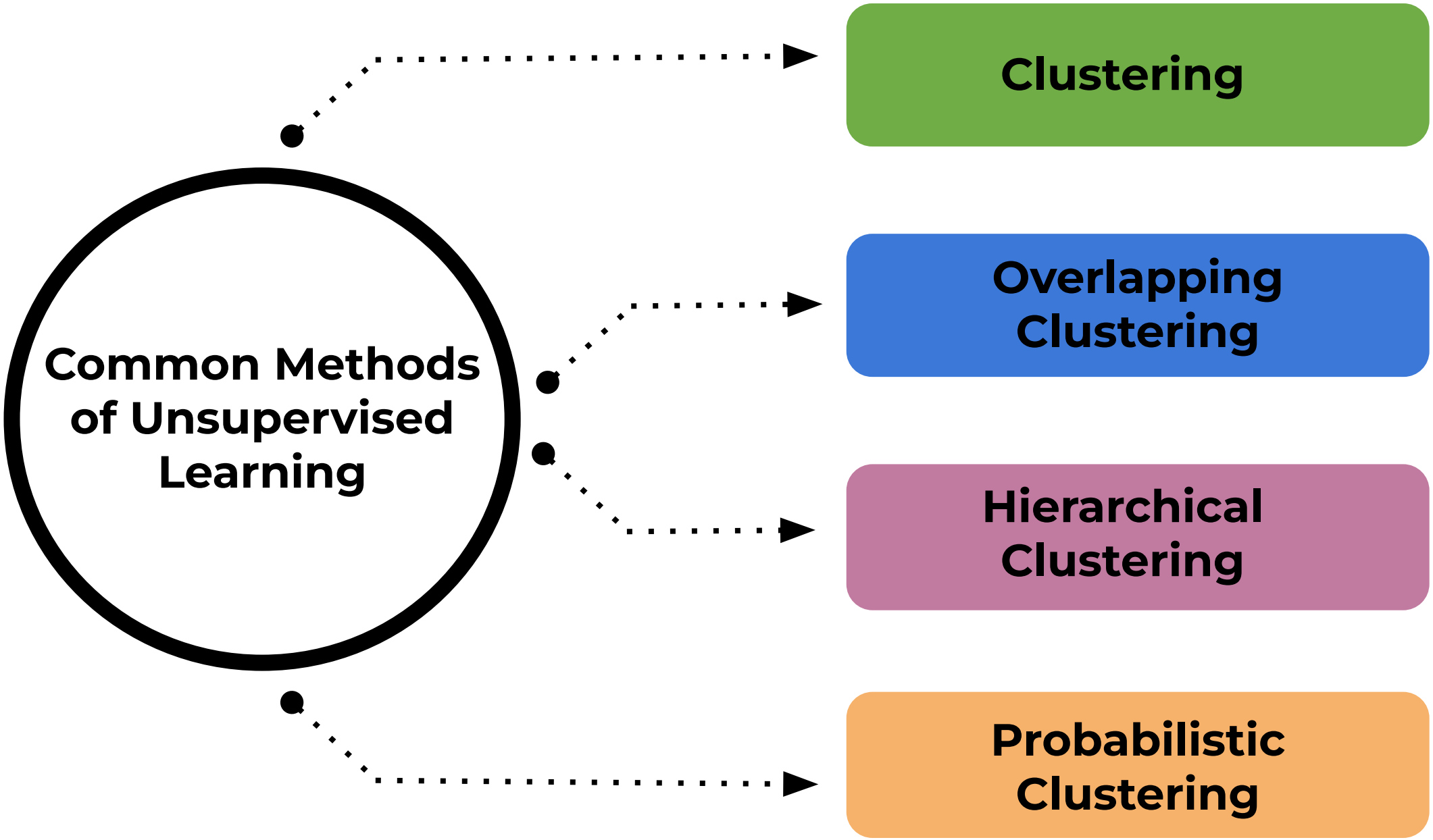
- a criação de algoritmos que podem analisar padrões em dados;
- gerar modelos para tarefas específicas;
- realizar previsões precisas;
- [...]


Deep learning
Unsupervised
Supervised
Machine learning
Text generation
Question answering
Context extraction
Classification
Machine translation
Natural language processing
Expert systems
Artificial intelligence
Speech to text
Text to speech
Speech
Image recognition
Machine vision
Vision
Planning
Robotics

O que é *Unsupervised Learning*?

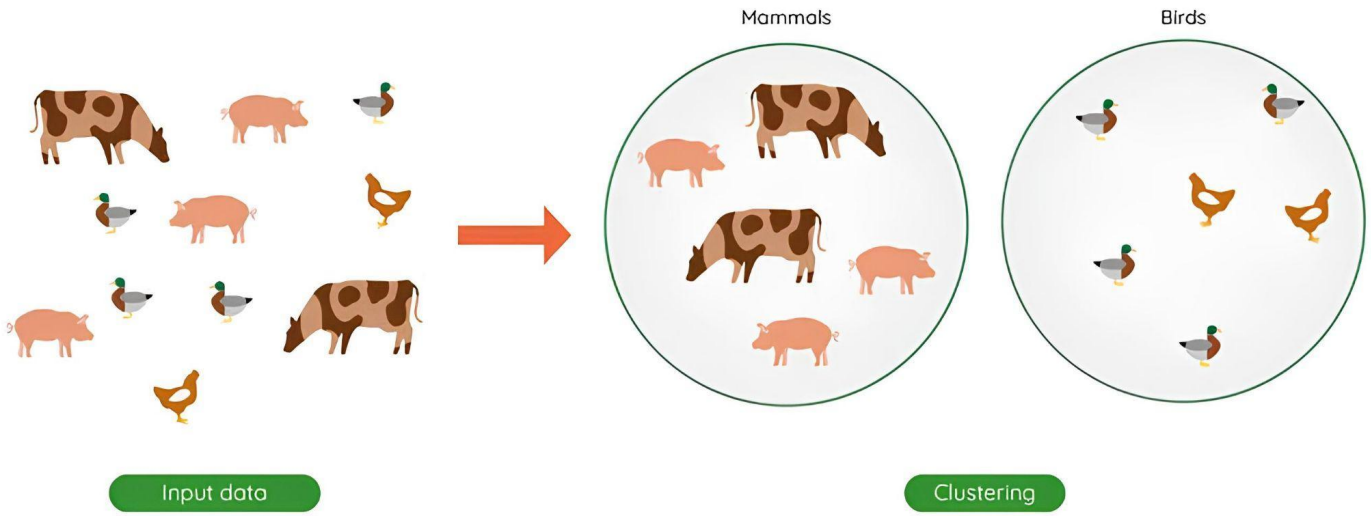
Unsupervised Learning é um método de aprendizado de máquina onde a máquina aprende a entender os dados por conta própria, sem qualquer orientação ou supervisão explícita. É tipo de aprendizado muito comum:

- em análise exploratória;
- quando se deseja identificar padrões e estruturas nos dados;
- situações onde os dados apresentam uma organização natural
- [...]


Common Methods of Unsupervised Learning
Clustering
Overlapping Clustering
Hierarchical Clustering
Probabilistic Clustering

O que é Clustering?

Clustering, ou ‘agrupamento’, é um método de aprendizado de máquina não supervisionado, amplamente usado para agrupar dados semelhantes em conjuntos chamados de "clusters" (ou grupos). O objetivo é que os dados dentro de um mesmo *cluster*/grupo sejam mais semelhantes entre si do que com dados de outros clusters.



Clustering

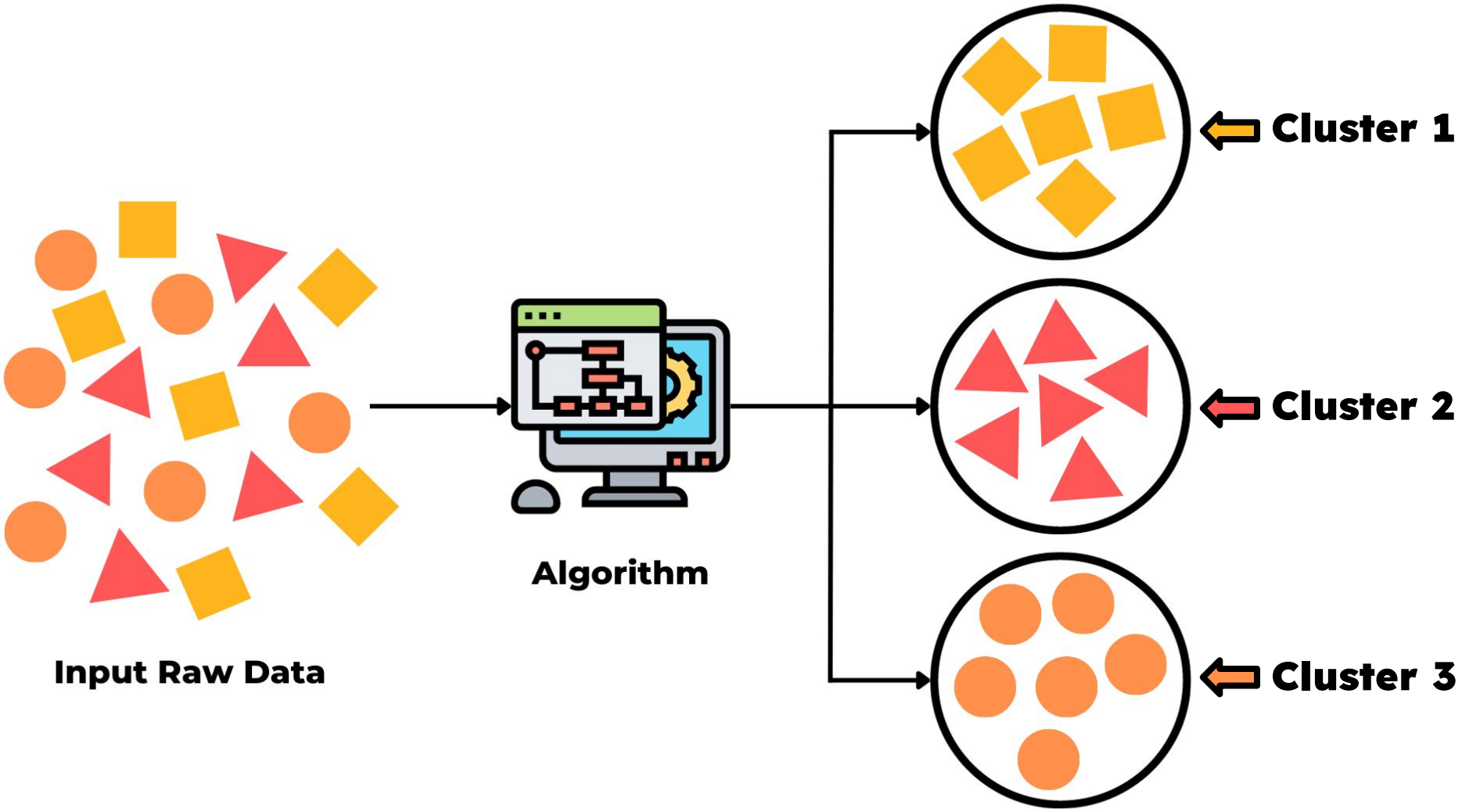
Existem muitos métodos para realizar o *clustering*, cada um com suas próprias abordagens/características.

- *K-means;*
- Hierárquico;
- DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*);
- *Mean Shift*;
- *Gaussian Mixture Models* (GMM);
- [...]

Clustering (*K-means*)

K-Means: um dos métodos de clustering/agrupamento mais amplamente utilizados para dividir um conjunto de dados em K *clusters*/grupos com base nas características dos dados.

O objetivo é que os dados dentro de um mesmo cluster sejam o mais semelhantes possível entre si e o mais diferente possível dos dados em outros *clusters*.


Cluster 1
Cluster 2
Cluster 3
Algorithm
Input Raw Data

Clustering (*K-means*)

Etapa 1 - Tenha os dados!

Etapa 2 - Escolha do Número de *Clusters* (K):

O valor de **K** (grupos) é uma escolha crítica e pode ser baseado em conhecimento prévio do usuário ou obtido a partir de técnicas específica de avaliação dos dados.

Ex.1: Em um estudo sobre imagens de satélite para classificar diferentes tipos de uso da terra, você pode saber que há cinco tipos principais de cobertura e uso da terra na região de interesse (floresta, cerradão, agricultura, água, urbano). Definir **K** como 5 pode ajudar a identificar essas categorias na imagem.

Ex.2: Com base no Índice de Desenvolvimento Humano (IDH) de vários municípios, você pode saber que há três grandes categorias de desenvolvimento humano (como "Alto", "Médio", e "Baixo"). Definir **K** como 3 pode ajudar a identificar e comparar os municípios dentro dessas categorias.

Clustering (*K-means*)

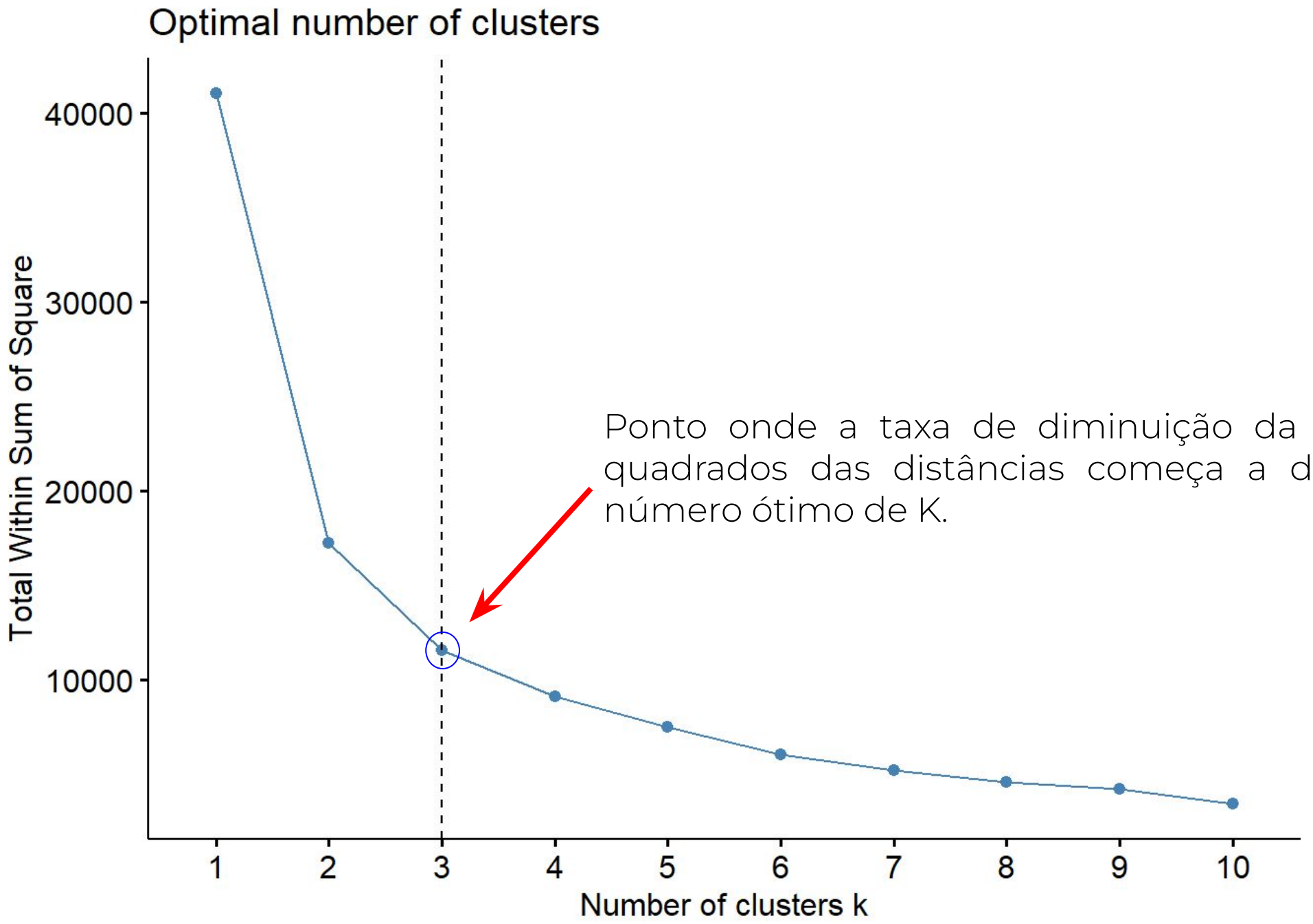
Etapa 2 - Escolha do Número de Clusters (K):

O valor de **K** também pode ser obtido a partir de técnicas de avaliação, a exemplo do Método do Cotovelo.

O Método do Cotovelo baseia-se na análise da soma dos quadrados das distâncias dentro dos clusters para diferentes valores de K. O objetivo é identificar um ponto de inflexão, ou "cotovelo", no gráfico da soma dos quadrados das distâncias em função de K.

Esse ponto de inflexão representa um valor de **K** onde a adição de mais *clusters* não melhora significativamente a qualidade dos *clusters*.

$$\sum_{i=0}^{n} (X_i - \overline{X})^2$$



Ponto onde a taxa de diminuição da soma dos quadrados das distâncias começa a desacelerar: número ótimo de K.

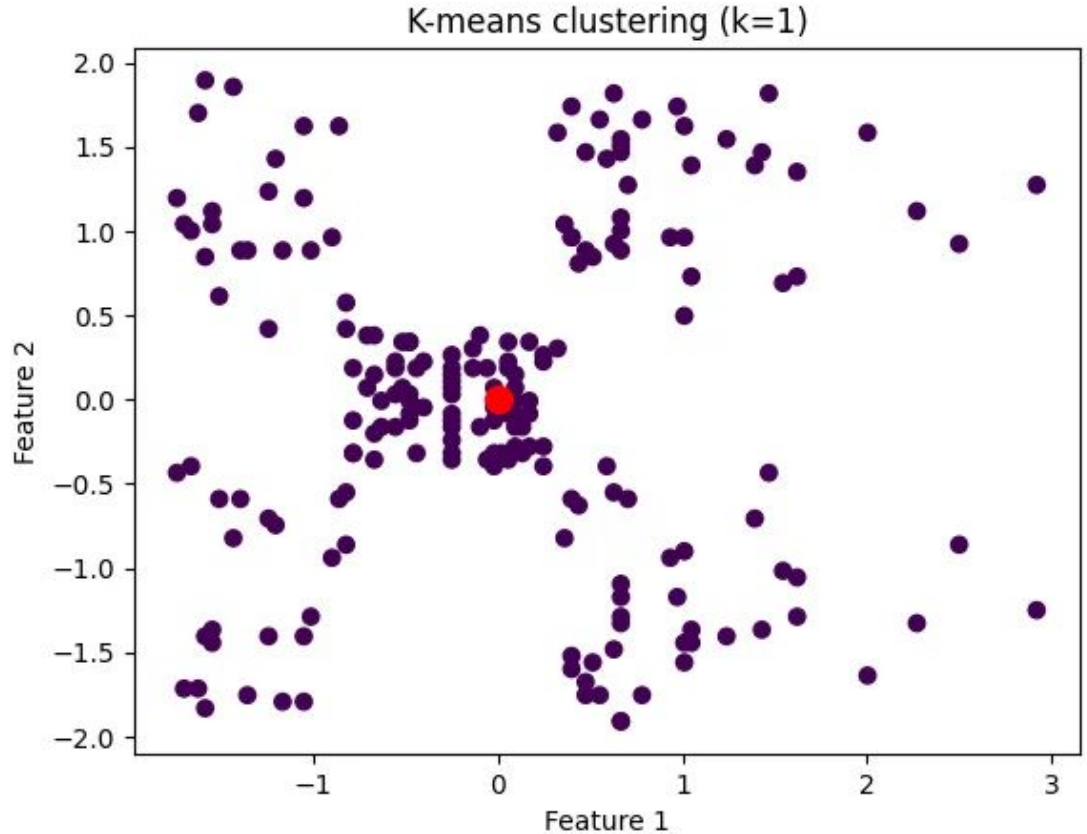
Clustering (*K-means*)

Etapa 3 - Inicialização dos Centróides:

Os centróides são os centros dos *clusters*. O *K-means* começa a tarefa de clusterização, de fato, com a escolha inicial desses centróides. Existem muitas maneiras de fazer isso.

Aleatoriamente: Selecionar aleatoriamente K pontos do conjunto de dados como os centróides iniciais.

Método *K-means++*: essa técnica ajuda a escolher centróides iniciais que estão mais distantes uns dos outros para 'melhorar a convergência do algoritmo'.


K-means clustering (k=1)
Feature 2
Feature 1
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−1
0
1
2
3

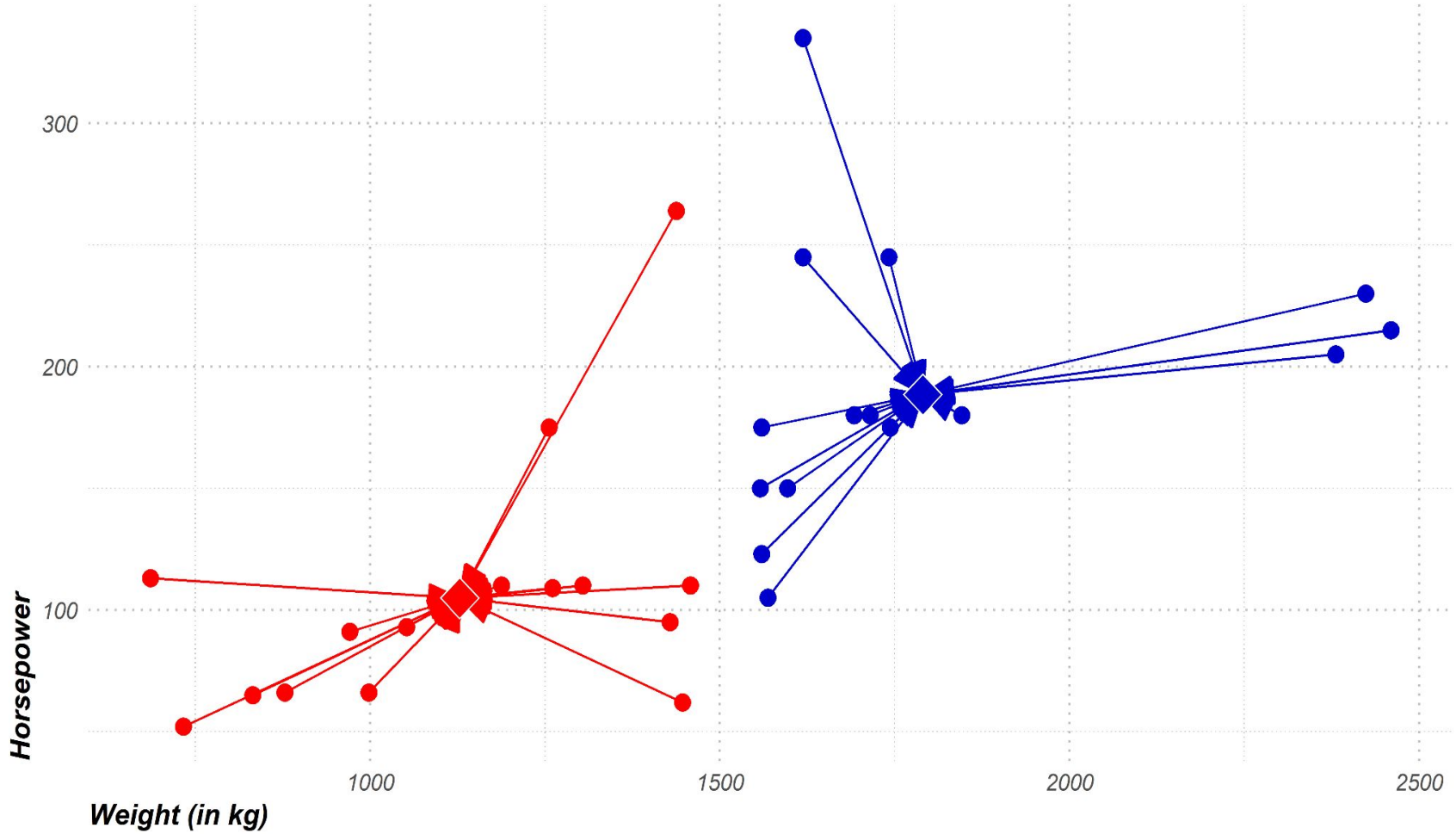
Clustering (*K-means*)

Etapa 4 - Atribuir cada ponto ao *cluster* mais próximo:

Considerando o valor de **K**, para cada ponto no conjunto de dados, o algoritmo calcula a distância entre o ponto e cada um dos centróides. O ponto é então atribuído ao *cluster* cujo centróide está mais próximo.

Por padrão, a distância é calculada usando a distância euclidiana, que é a raiz quadrada da soma dos quadrados das diferenças entre as coordenadas dos pontos. Existem várias outras métricas de distância ou similaridade que podem ser usadas, como distância de Manhattan, distância de Mahalanobis, entre outras.

$$D = \sqrt{\left[\sum_{i=1}^{n}(x_i - y_i)^2\right]}$$


300
200
100
Horsepower
1000
1500
2000
2500
Weight (in kg)

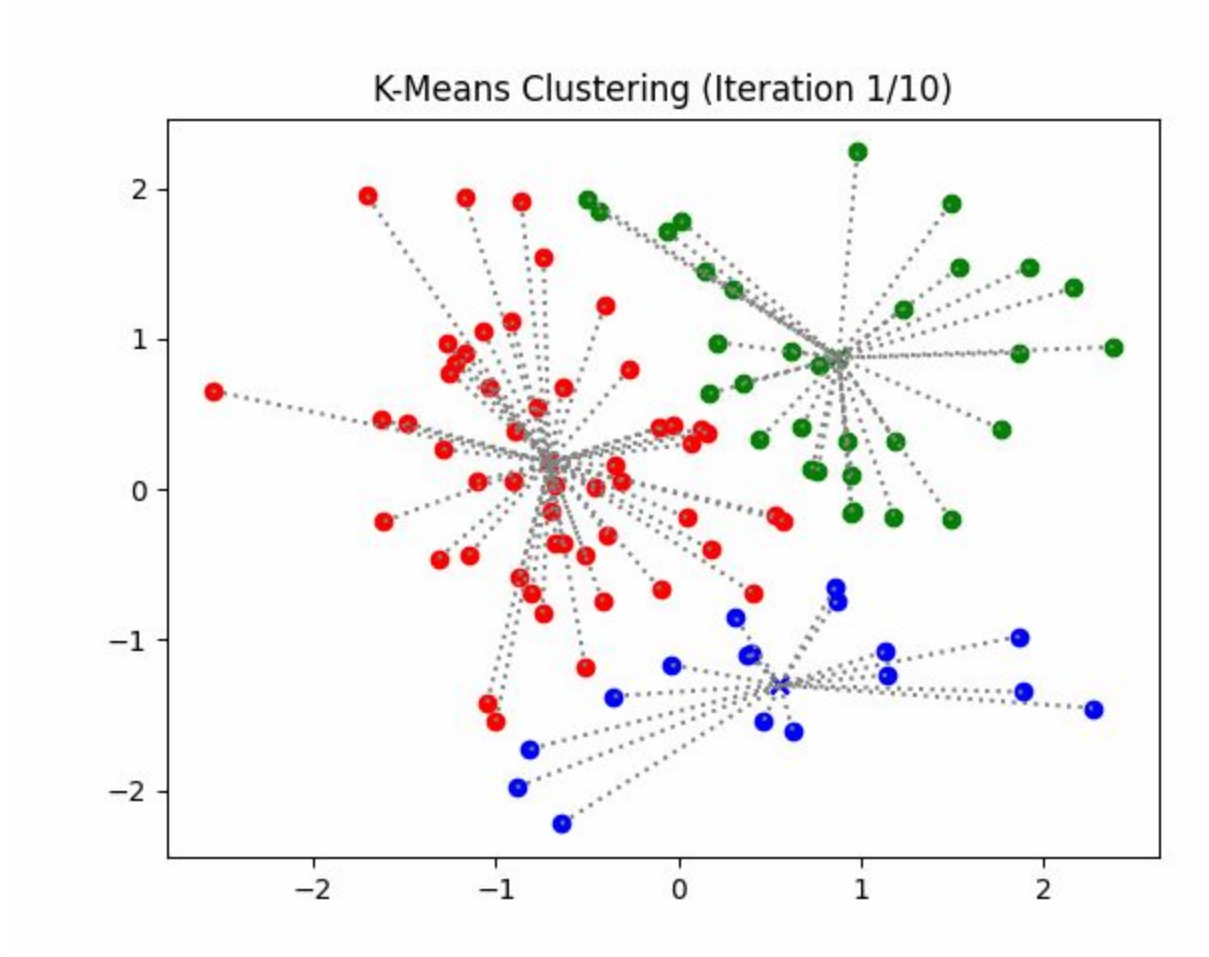
Clustering (*K-means*)

Etapa 4 - Recalcular os Centróides:

Depois de ter todos os pontos atribuídos aos clusters (etapa anterior), a próxima etapa é recalcular os centróides. O novo centróide de um cluster é a média (ou o centro geométrico) de todos os pontos que pertencem a esse cluster.

A "ideia" por trás de usar a média' é encontrar o ponto que está no "centro" dos pontos do cluster.

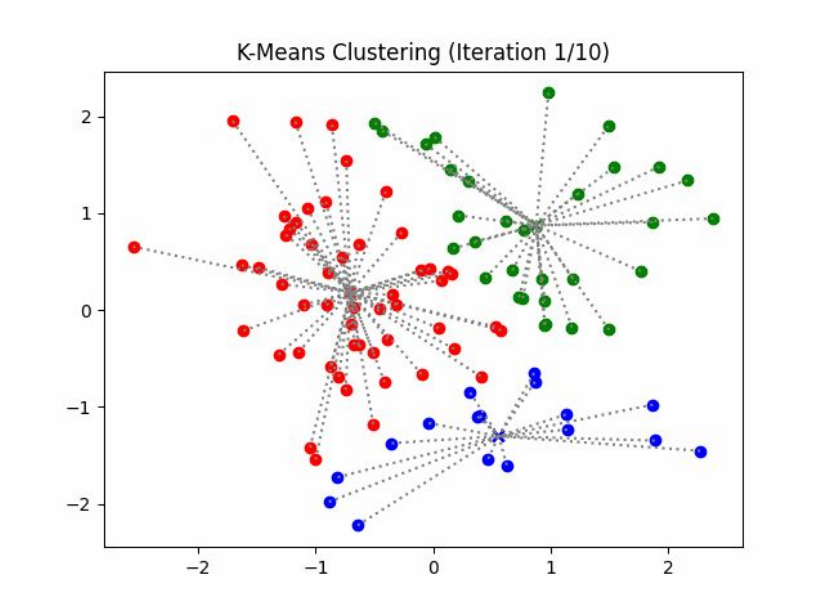
O objetivo é manter os clusters compactos e bem definidos, ajustando a posição dos centróides para refletir melhor a distribuição dos pontos em cada cluster.


K-Means Clustering (Iteration 1/10)
2
1
0
−1
−2
−2
−1
0
1
2

Clustering (*K-means*)

Etapa 4 - Repetir as etapas 3 e 4:

As etapas de atribuição dos pontos aos clusters (Etapa 3) e recalcular os centróides (Etapa 4) são repetidas até que os centróides não mudem mais significativamente, isto é, até que a mudança entre as iterações seja muito pequena, ou até que um número máximo de iterações seja atingido.

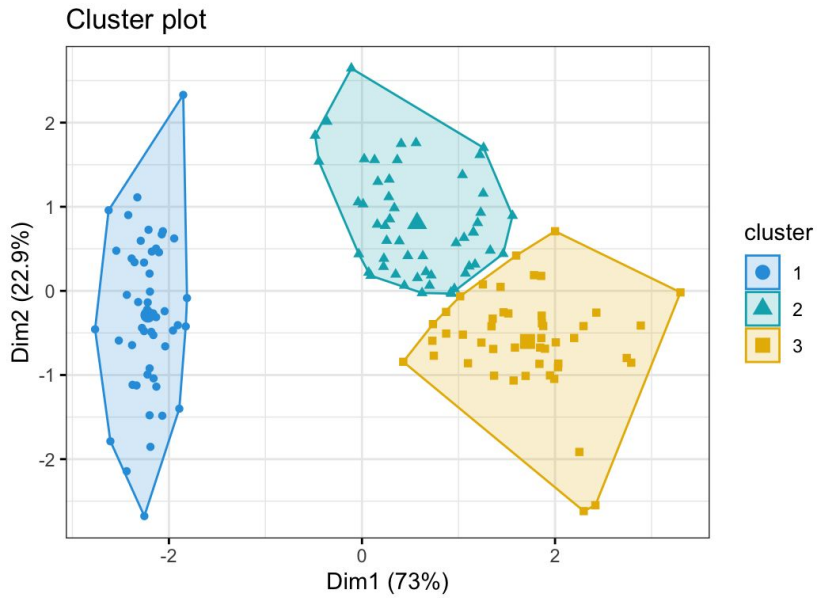


Clustering (*K-means*)

Etapa 5 - Analisar o resultado:

A saída do algoritmo é **K** *clusters*, em que os pontos são agrupados de maneira que pontos dentro de um *cluster* são mais similares entre si do que com pontos de outros *clusters*.

O resultado é uma divisão dos dados em **K** grupos, onde cada grupo é representado pelo centróide do *cluster*.



Tenha os dados prontos para análise!

Escolha do Número de *Clusters* (K).

- Conhecimento prévio;
- Obtido a partir de técnicas de avaliação;

Inicialização dos Centróides.

- Inicialização Aleatória;
- K-means++;
- Inicialização por Amostragem de Dados
- [...]

Recalcular os Centróides.

- manter os clusters compactos e bem definidos;

Analisar o resultado:

O resultado é uma divisão dos dados em **K** grupos.

O'REILLY®

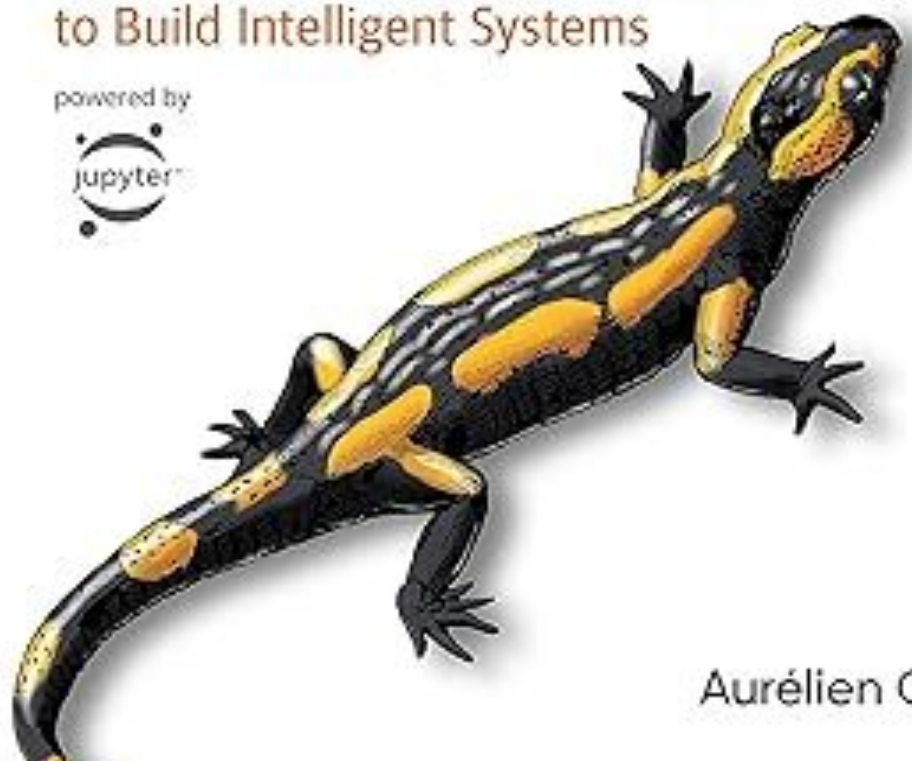
2nd Edition
Updated for
TensorFlow 2

Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow

Concepts, Tools, and Techniques to Build Intelligent Systems

powered by

jupyter



Aurélien Géron