

Machine Learning

INTRODUÇÃO

Profa Silvia Maria Wanderley Moraes



Conceito

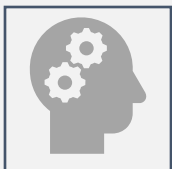
Machine Learning

(ou Aprendizado de máquina)
é uma subárea da inteligência artificial que trabalha com **algoritmos de computação que podem identificar, extrair e aprender padrões a partir de dados.**

Mas o
que é
**Machine
Learning**
?



Conceito



Capacidade de melhorar o desempenho na realização de alguma tarefa por meio da experiência. (Mitchell, 1997)



Aprendizado é qualquer mudança em um sistema que melhore o seu desempenho na segunda vez que ele repetir a mesma tarefa ou outra tarefa tirada da mesma população (Simon, 1983)



Processo de indução de uma hipótese a partir da experiência. (Facelli e outros, 2011).

Por que estudar Machine Learning?

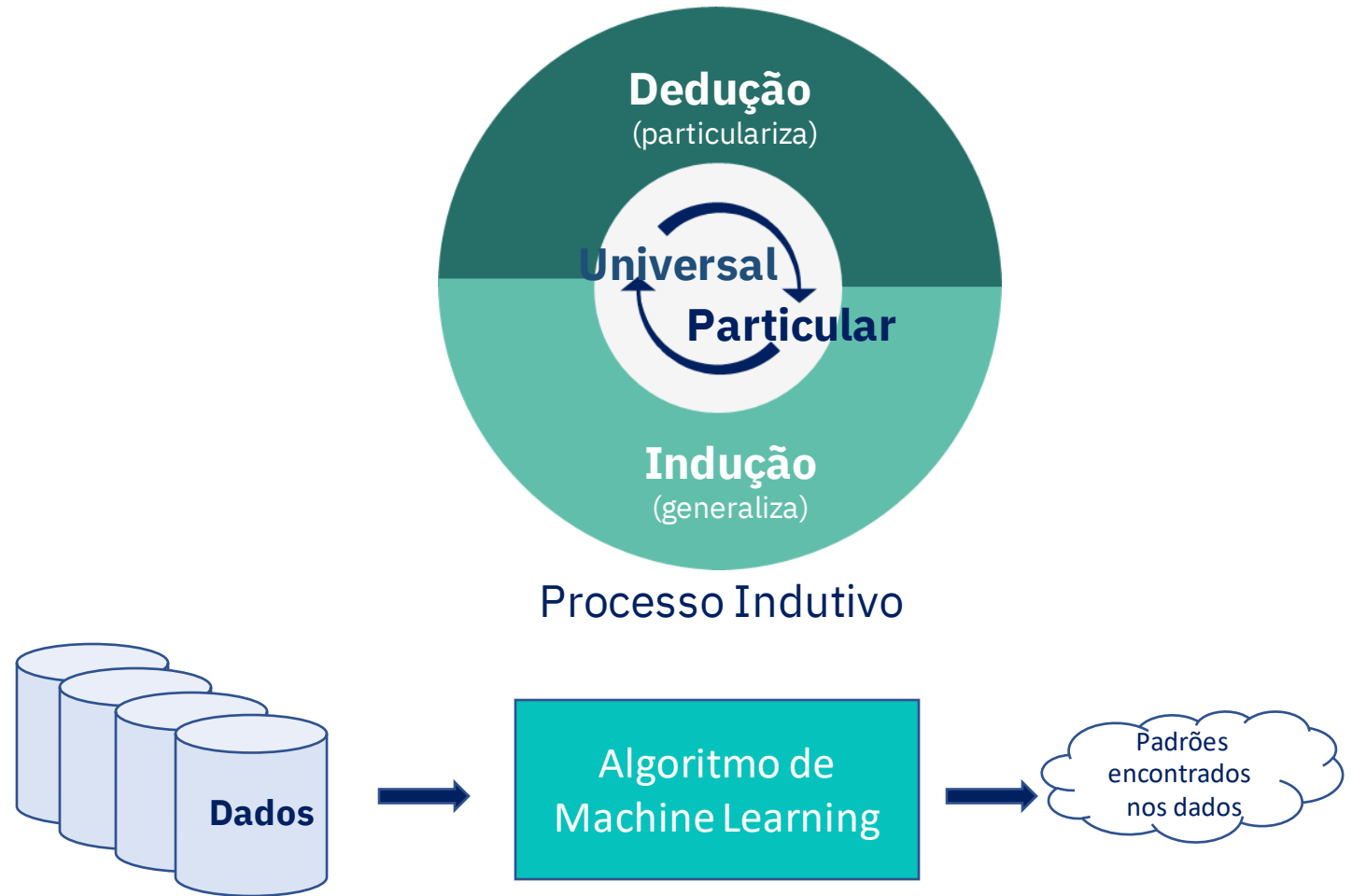
Razões Práticas:

- Reduz o tempo de programação
- Permite customizar produtos
- Resolve problemas que não sabemos resolver

Razões Filosóficas:

- Muda a forma como pensamos os problemas
- A abordagem deixa de ser matemática e passar a ser mais natural: aprender a partir da observação em um mundo incerto.

Machine
Learning
usa
aprendizado
indutivo.



Machine Learning usa aprendizado indutivo.

- Pedro, 18 anos, apresenta sintomas como febre, dor de cabeça e perda de olfato e seu teste para COVID resultou em **POSITIVO**.
- Maria, 30 anos, apresenta sintomas como dor de garganta, febre baixa e seu teste para COVID resultou em **NEGATIVO**.
- João, 55 anos, apresenta sintomas como dor no corpo, febre alta e seu teste para COVID resultou em **POSITIVO**.
- Ana, 78 anos, apresenta sintomas como dor de cabeça e febre baixa e seu teste para COVID resultou em **NEGATIVO**.
- ...



Base de dados de pacientes



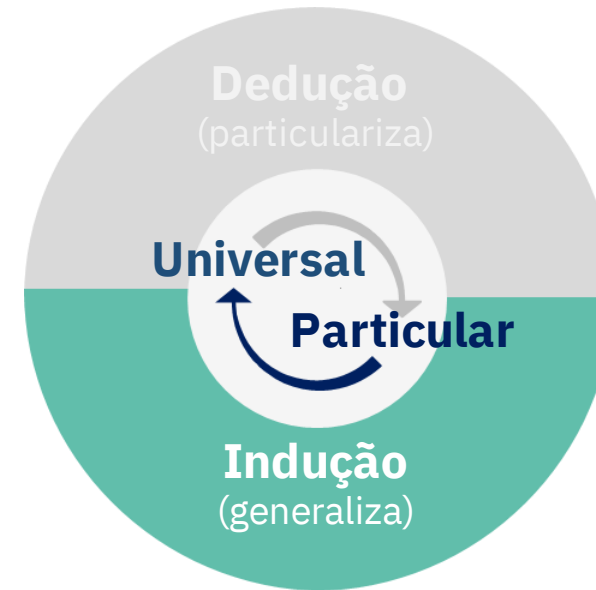
Algoritmo de
Machine Learning



Modelo COVID

Padrões indicativos de COVID, sintomas combinados:

- Febre
- Perda de olfato
- ...



Processo Indutivo

Modelo
Conjunto de padrões identificados a partir de “regularidades”, “comportamentos frequentes” existentes nos dados.

Machine Learning usa aprendizado indutivo.

- Pedro, 18 anos, apresenta sintomas como febre, dor de cabeça e perda de olfato e seu teste para COVID resultou em **POSITIVO**.
- Maria, 30 anos, apresenta sintomas como dor de garganta, febre baixa e seu teste para COVID resultou em **NEGATIVO**.
- João, 55 anos, apresenta sintomas como dor no corpo, febre alta e seu teste para COVID resultou em **POSITIVO**.
- Ana, 78 anos, apresenta sintomas como dor de cabeça e febre baixa e seu teste para COVID resultou em **NEGATIVO**.
- ...

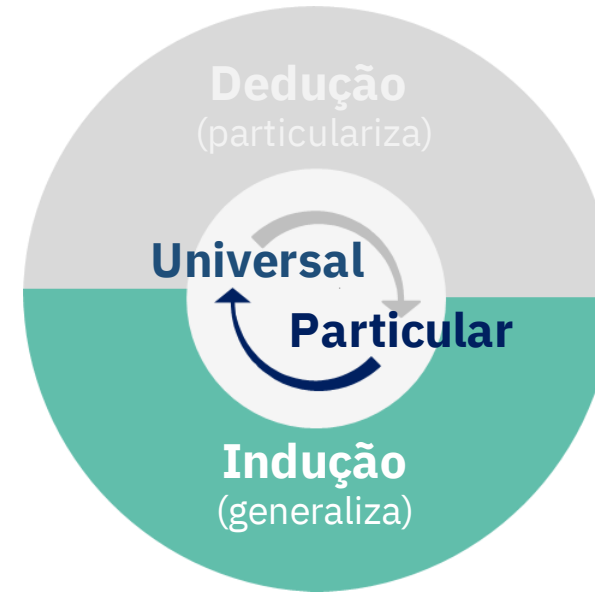


Base de dados de pacientes

Pré-processamento

Algoritmo de Machine Learning

Processo Indutivo



Modelo
Conjunto de padrões identificados a partir de “regularidades”, “comportamentos frequentes” existentes nos dados.

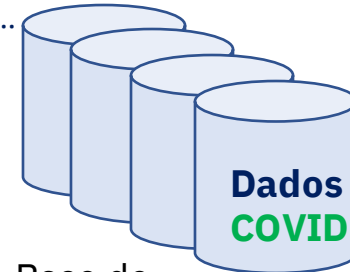


Padrões indicativos de COVID, sintomas combinados:

- Febre
- Perda de olfato
- ...

Machine Learning usa aprendizado indutivo.

- Pedro, 18 anos, apresenta sintomas como febre, dor de cabeça e perda de olfato e seu teste para COVID resultou em **POSITIVO**.
- Maria, 30 anos, apresenta sintomas como dor de garganta, febre baixa e seu teste para COVID resultou em **NEGATIVO**.
- João, 55 anos, apresenta sintomas como dor no corpo, febre alta e seu teste para COVID resultou em **POSITIVO**.
- Ana, 78 anos, apresenta sintomas como dor de cabeça e febre baixa e seu teste para COVID resultou em **NEGATIVO**.
- ...



Base de dados de pacientes

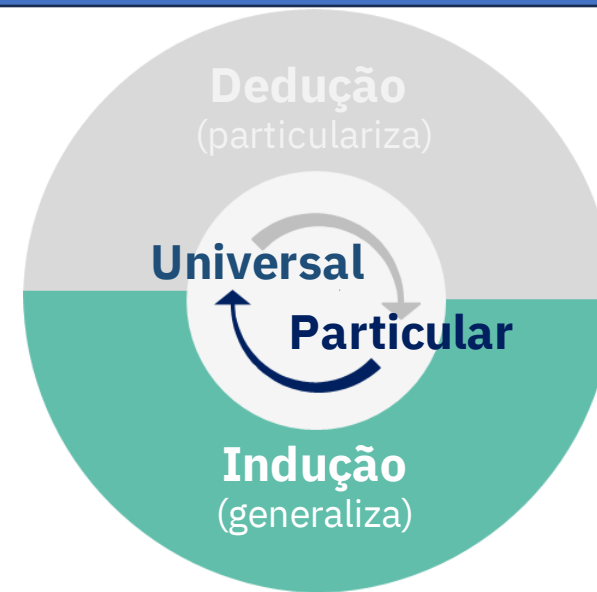
Pré-processamento

Algoritmo de Machine Learning

Modelo COVID

Padrões indicativos de COVID, sintomas combinados:

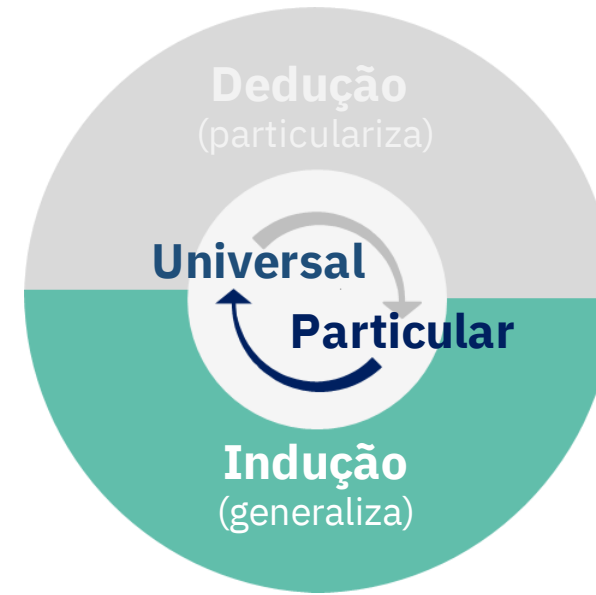
- Febre
- Perda de olfato
- ...



Etapa de Treinamento

Modelo
Conjunto de padrões identificados a partir de “regularidades”, “comportamentos frequentes” existentes nos dados.

Machine Learning usa aprendizado indutivo.



Etapa de Generalização

Matheus, 36
anos, apresenta sintomas
como febre, dor de cabeça
...

Entrada: Dados novos



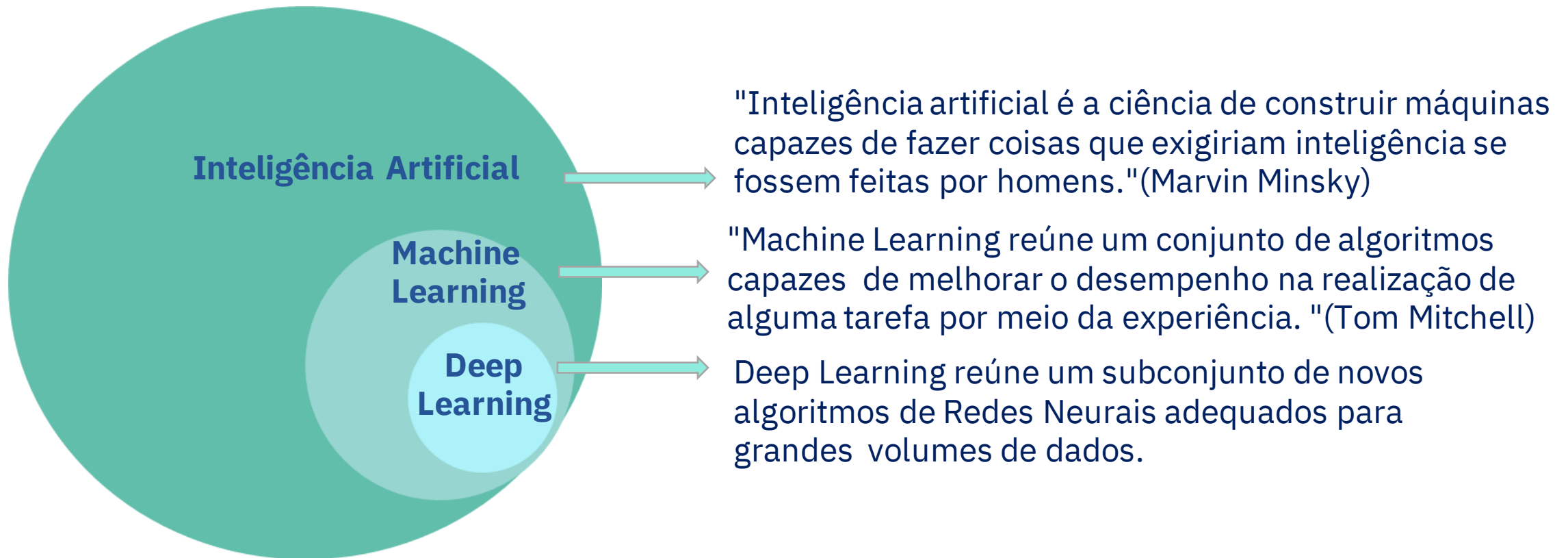
86% de chances de
estar com COVID

Saída: Predição



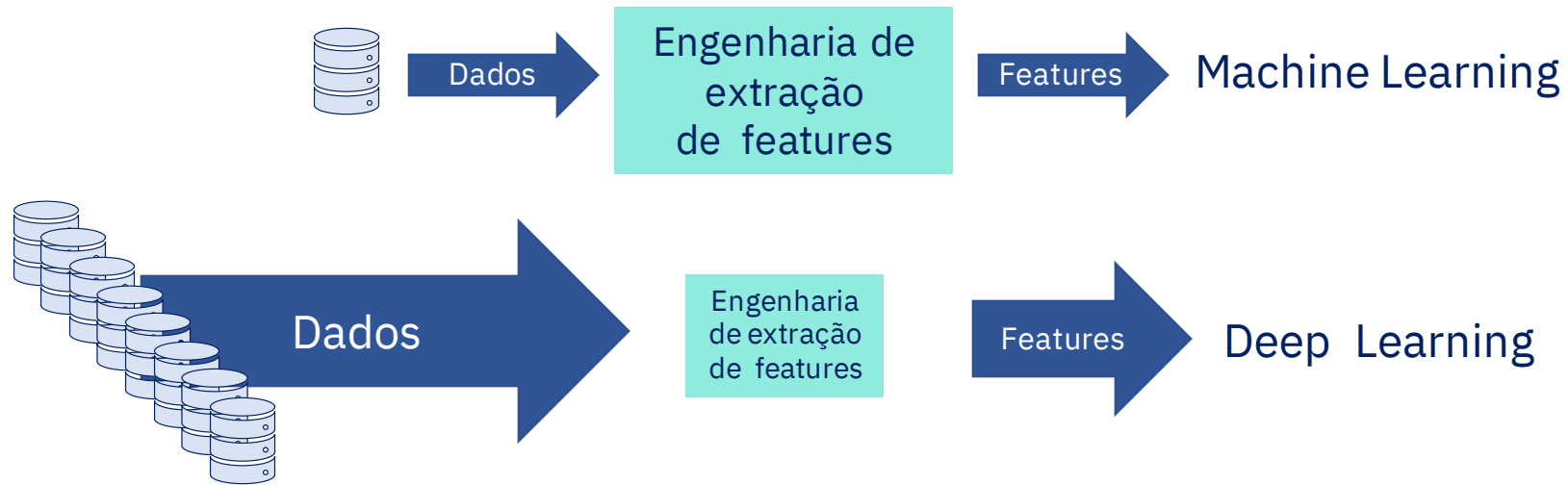
Machine Learning e
Deep Learning:
como essas áreas se
relacionam?

Machine Learning e Deep Learning



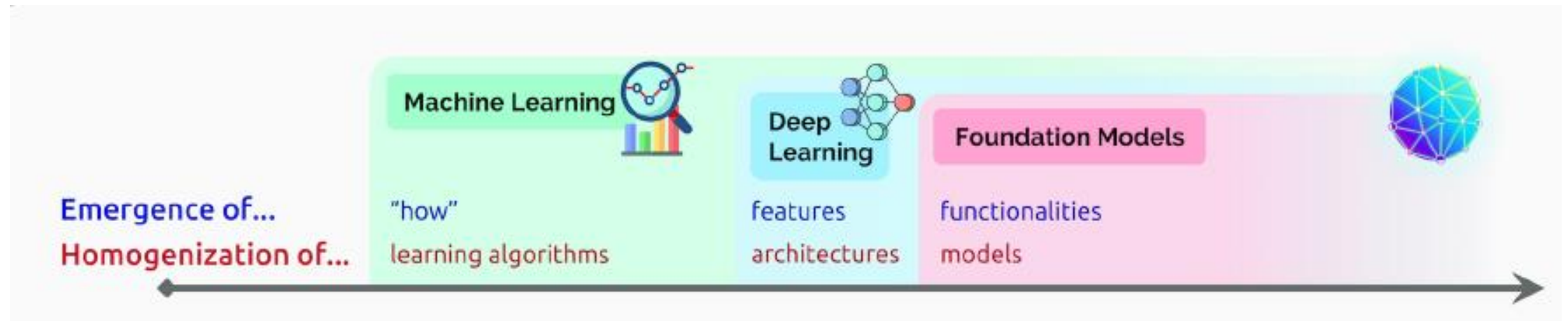
Deep Learning

- Uso de redes neurais profundas.
- Capacidade de trabalhar com grandes volumes de dados.
- Aprendizagem de representações automatizada, ou pelo menos simplificada.



Deep Learning

Mais recentemente, temos os chamados Foundations Models: BERT, DALL-E, Gato e ChatGPT.



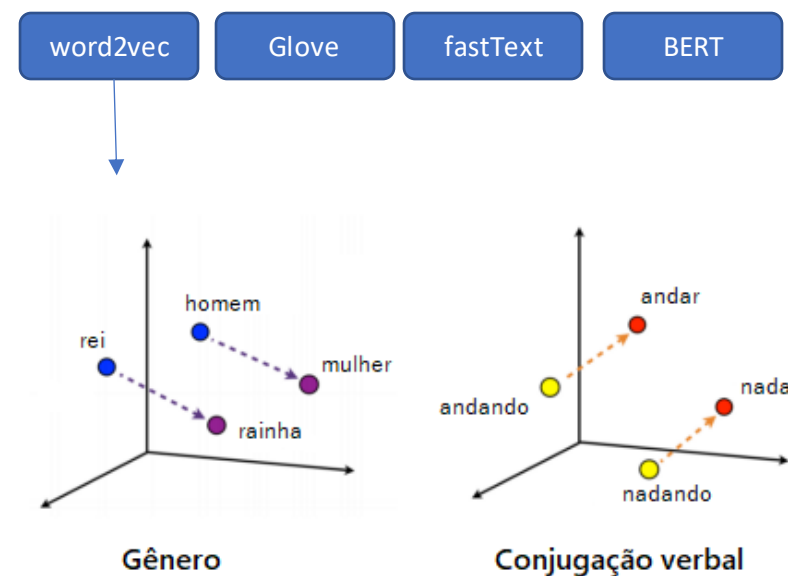
BERT

BERT é um modelo de linguagem.

Modelos de linguagem determinam a probabilidade de uma determinada sequência de palavras ocorrer em uma frase com base nas palavras anteriores.

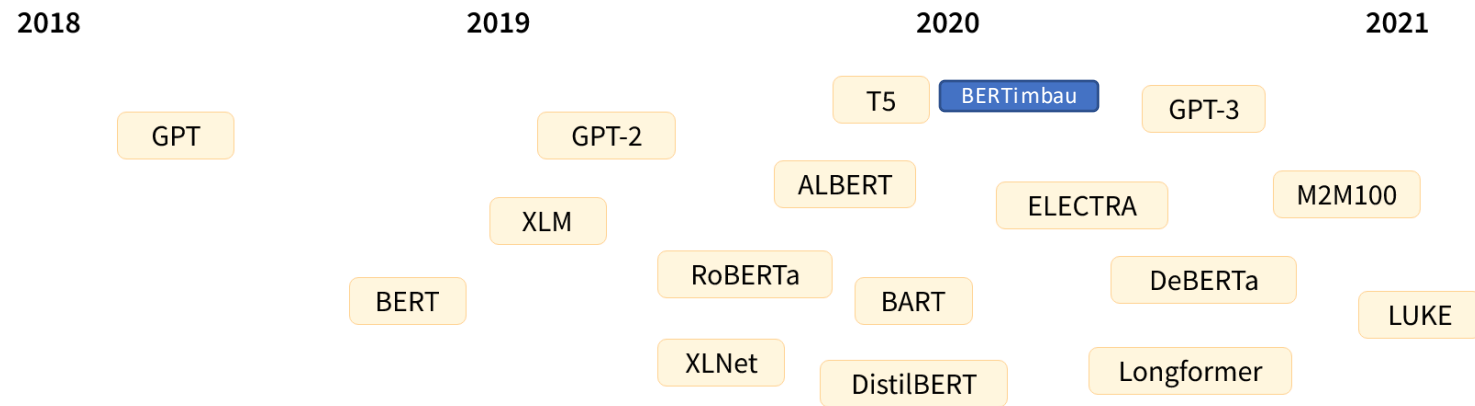
Esses modelos permitem que as palavras sejam representadas por vetor semânticos densos conhecidos como Word Embeddings.

Palavras que têm o mesmo significado têm uma representação semelhante – por exemplo, o famoso exemplo “Rei – Homem + Mulher = Rainha”.



BERT

- Bidirecional Encoder Representations from Transformers (BERT) : modelo pré-treinado para processamento de linguagem natural (NLP) desenvolvido pelo Google.



Fonte: <https://huggingface.co/course/chapter1/4>

BERT

Pode ser usado na Geração de Texto como em:

- Sistemas de Perguntas e Respostas
- Sumarizadores
- Geração de resposta para agente conversacional

Pode ser usado também em tarefas de compreensão de linguagem natural, como:

- Resolução de polissemia
- Resolução de correferência
- Desambiguação do sentido da palavra
- Inferência de linguagem natural
- Análise de sentimentos



DALL-E

DALL-E é um novo sistema de IA que pode criar imagens e arte realistas a partir de uma descrição em linguagem natural.

Entrada:

"a painting of a fox sitting in a field at sunrise in the style of Claude Monet"

uma pintura de uma raposa sentada em um campo ao nascer do sol ao estilo de Claude Monet



- **ChatGPT**

ChatGPT é um chatbot online de inteligência artificial desenvolvido pela OpenAI, lançado em novembro de 2022.

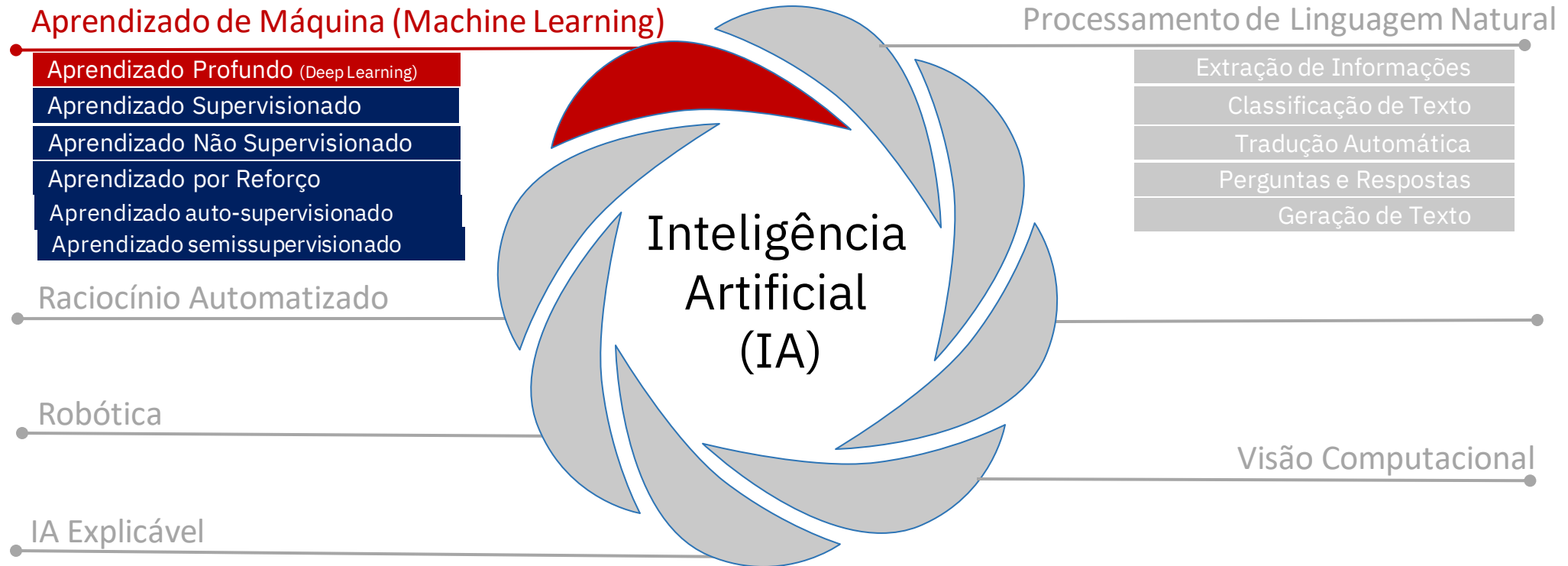
- O nome "ChatGPT" combina "Chat", referindo-se à sua funcionalidade de chatbot, e "GPT", que significa Generative Pre-trained Transformer, um tipo de modelo de linguagem grande (Large Language Models – LLMs)



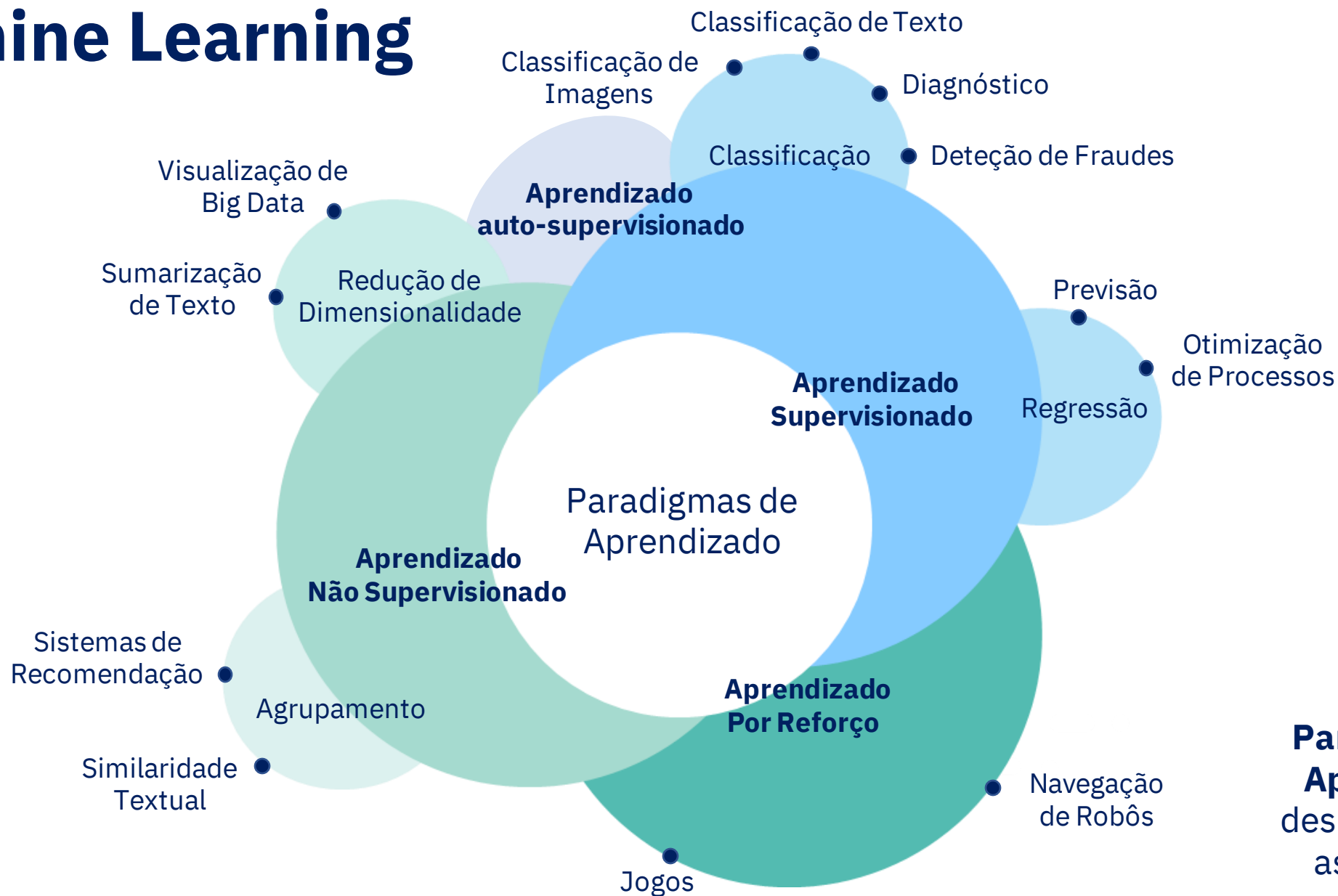
**Paradigmas de
Aprendizado:
quais os mais usados?**



Subáreas da Inteligência Artificial



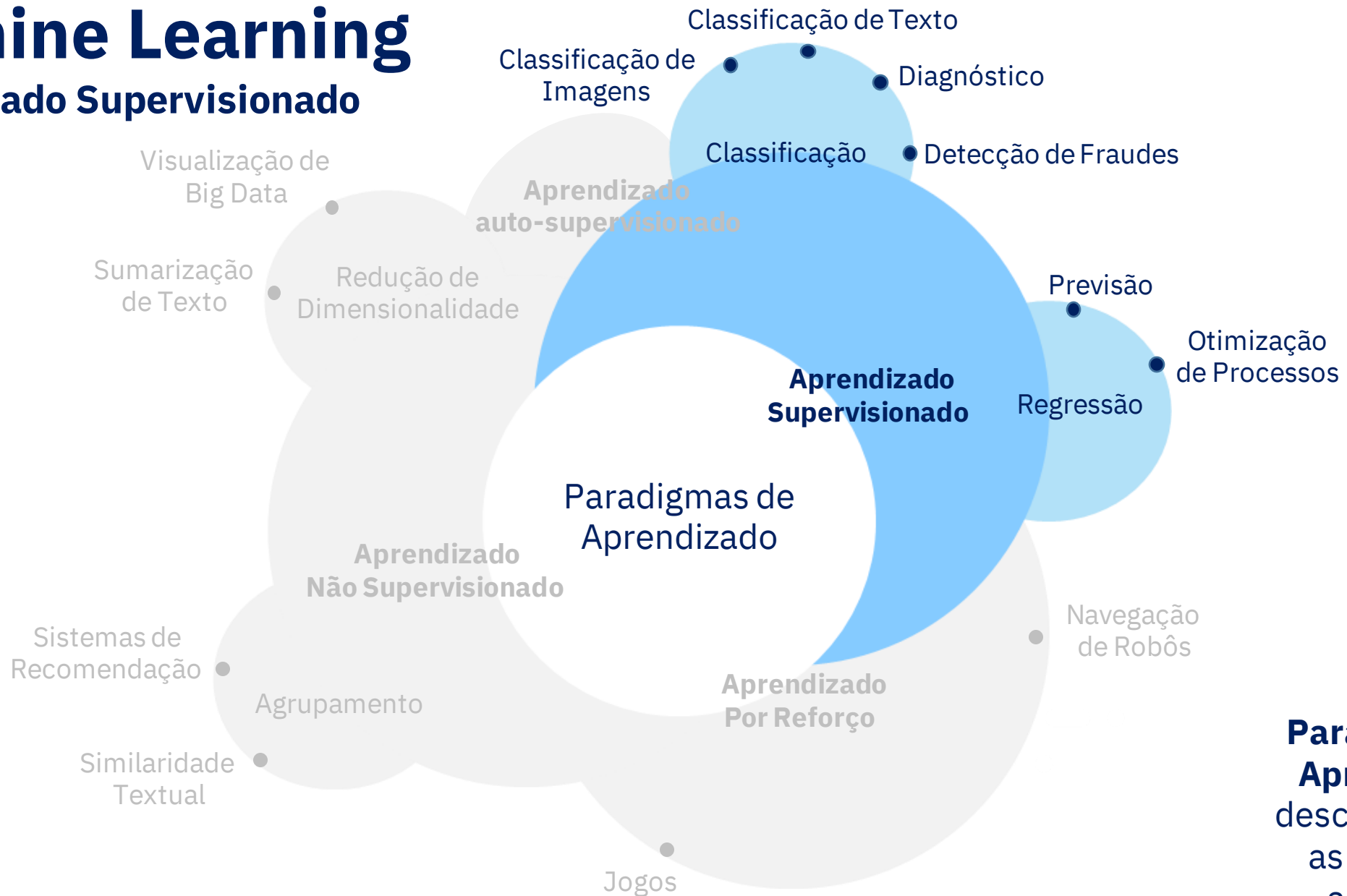
Machine Learning



Paradigmas de Aprendizado:
descrevem como
as máquinas
aprendem

Machine Learning

Aprendizado Supervisionado



Paradigmas de Aprendizado:
descrevem como
as máquinas
aprendem

Aprendizado Supervisionado

- Exige que os **dados** estejam **rotulados** (anotados com suas respectivas classes/valores de saída)
- Os algoritmos que seguem esse tipo de aprendizado recebem pares de valores:
 - os dados de entrada (x) e
 - os valores de saída (rótulos) correspondentes (y).



Aprendizado Supervisionado

Em um conjunto de dados (exemplos) rotulado :

- Cada dado corresponde a um indivíduo do domínio e é formado por uma tupla contendo características (features).

Atributo de entrada
(atributo previsor)

sepal length	sepal width	petal length	petal width	class
5,1	3,5	1,4	0,2	Iris-setosa
4,9	3,0	1,4	0,2	Iris-setosa
7,0	3,2	4,7	7,1	Iris-versicolor
6,4	3,2	4,5	1,5	Iris-versicolor
6,3	3,3	6,0	2,5	Iris-virginica
5,8	2,7	5,1	1,9	Iris-virginica

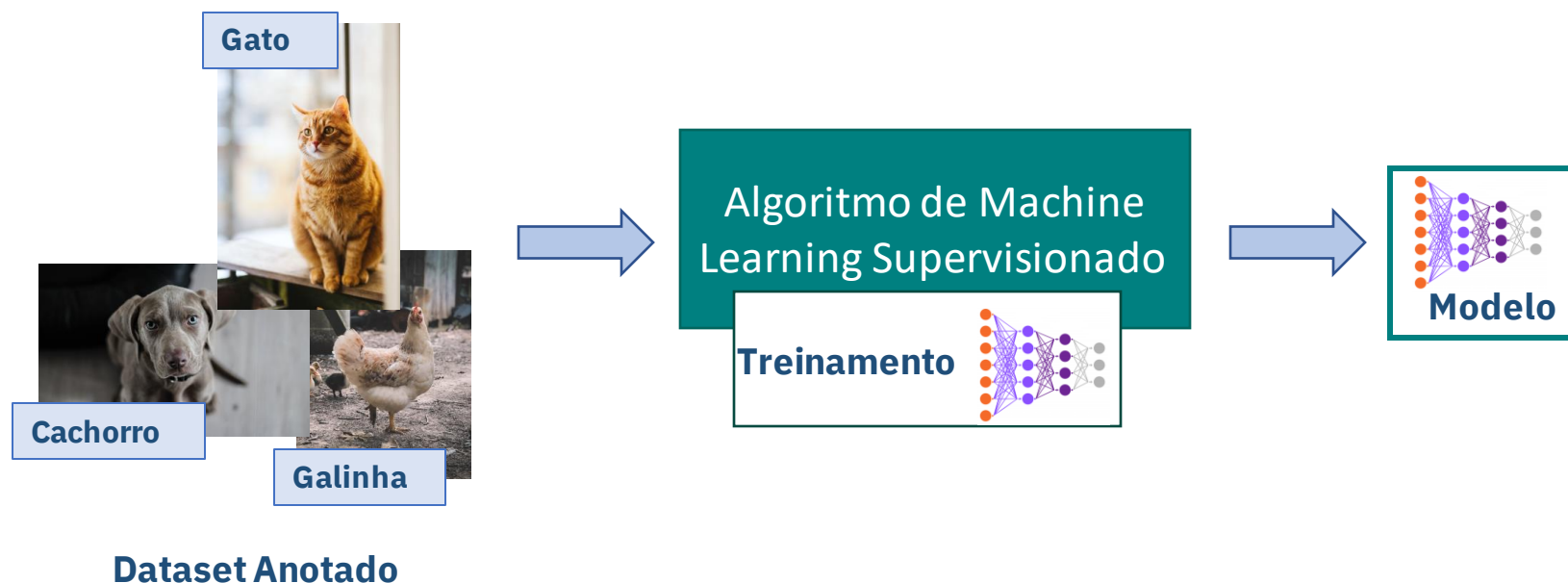
Atributo de saída
(atributo alvo ou meta)


Rótulo
(Classes)



Aprendizado Supervisionado

- O objetivo é encontrar um modelo capaz de mapear os valores de entrada (x) nos valores de saída y .
- Em outras palavras, que aproxime f , tal que $f(x) = y$.
- Supervisão: ajuste usando o erro em relação à saída esperada.





Modelo Probabilístico

Aprendizado Supervisionado

- **Tarefa preditiva:** encontra uma função (modelo) a partir dos dados de treino que possa ser usada para prever um rótulo (classe) ou valor de um novo exemplo.
- Pode ser:
 - **classificação** (rótulos discretos)
 - **regressão** (rótulos contínuos)

Classificação

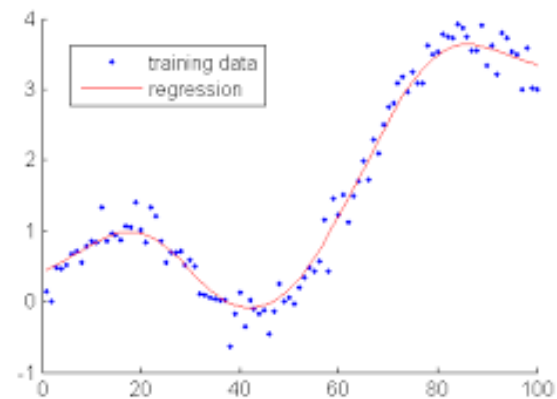
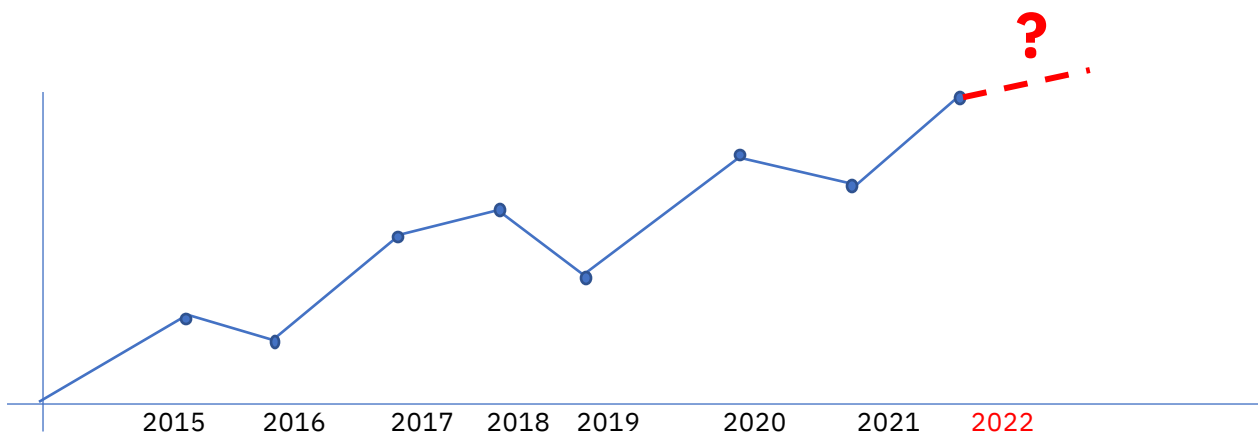


- É o processo de automaticamente atribuir rótulos a dados.
- Pode ser do tipo
 - Binária: possui apenas duas classes
 - Multiclasse: possui mais de duas classes
- Pode atribuir
 - Um único rótulo (single label)
 - Vários rótulos (multi-label)

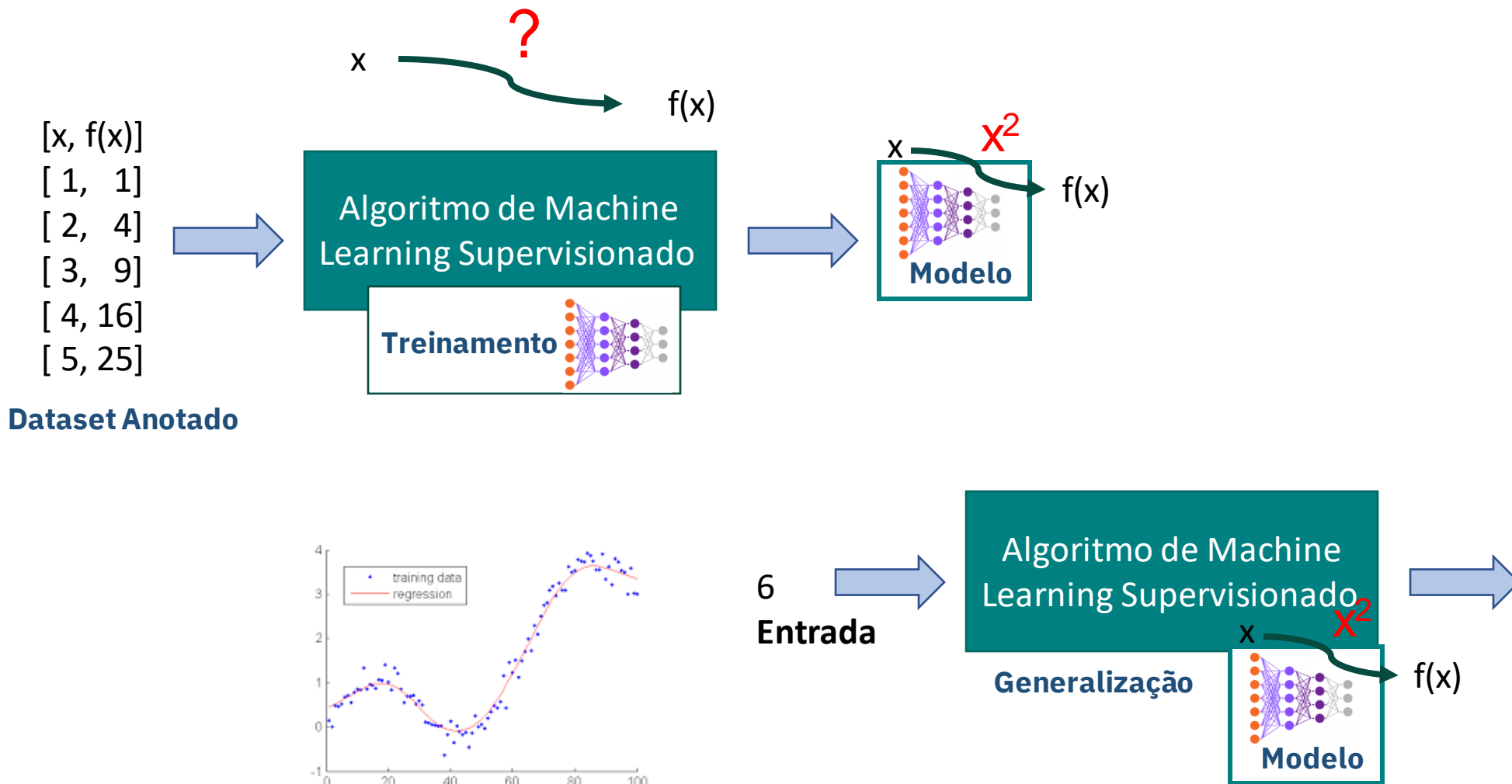


Regressão

- É o processo de automaticamente predizer novos valores y .
- Neste caso, os dados são anotados com valores.

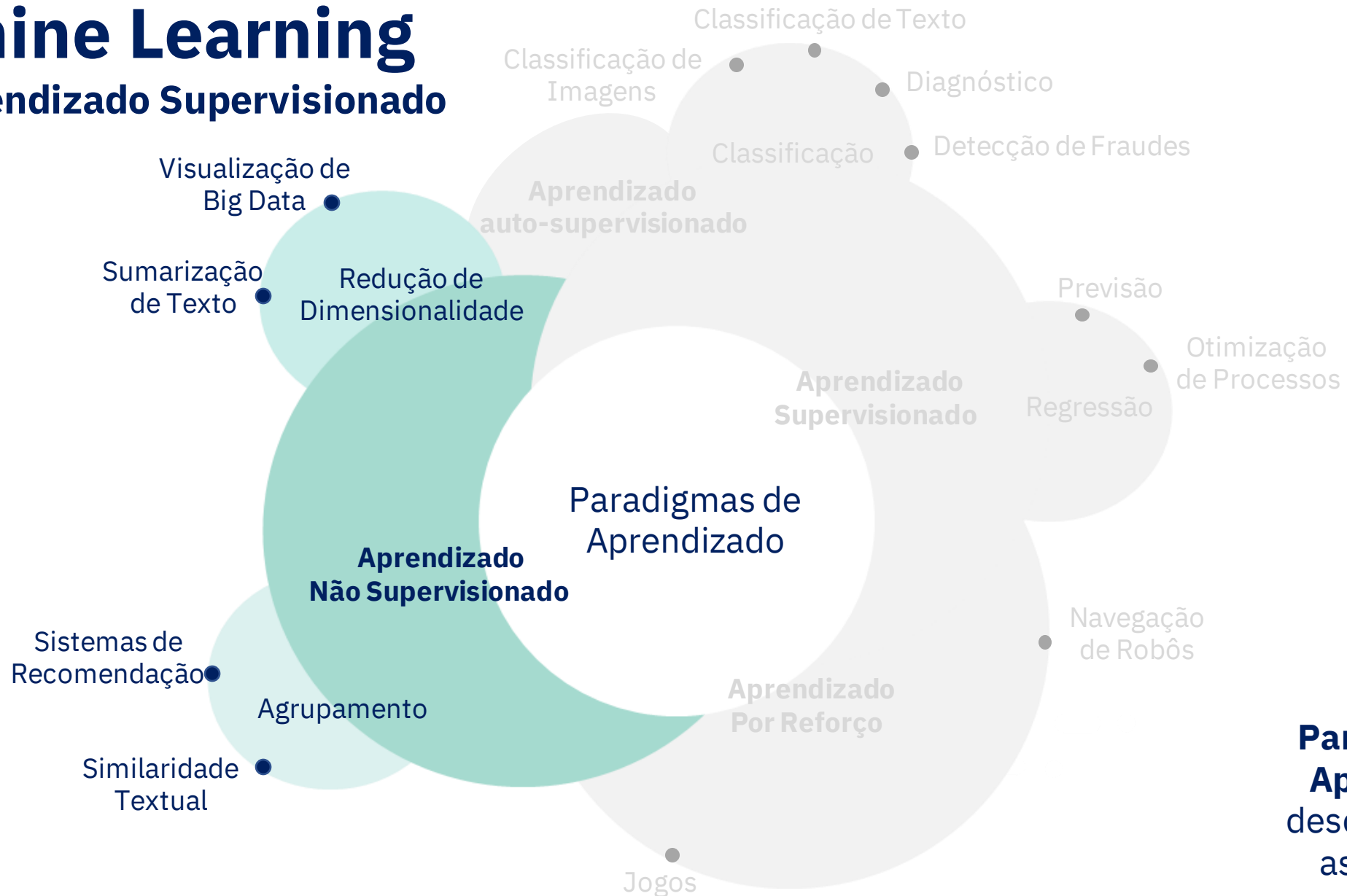


Regressão



Machine Learning

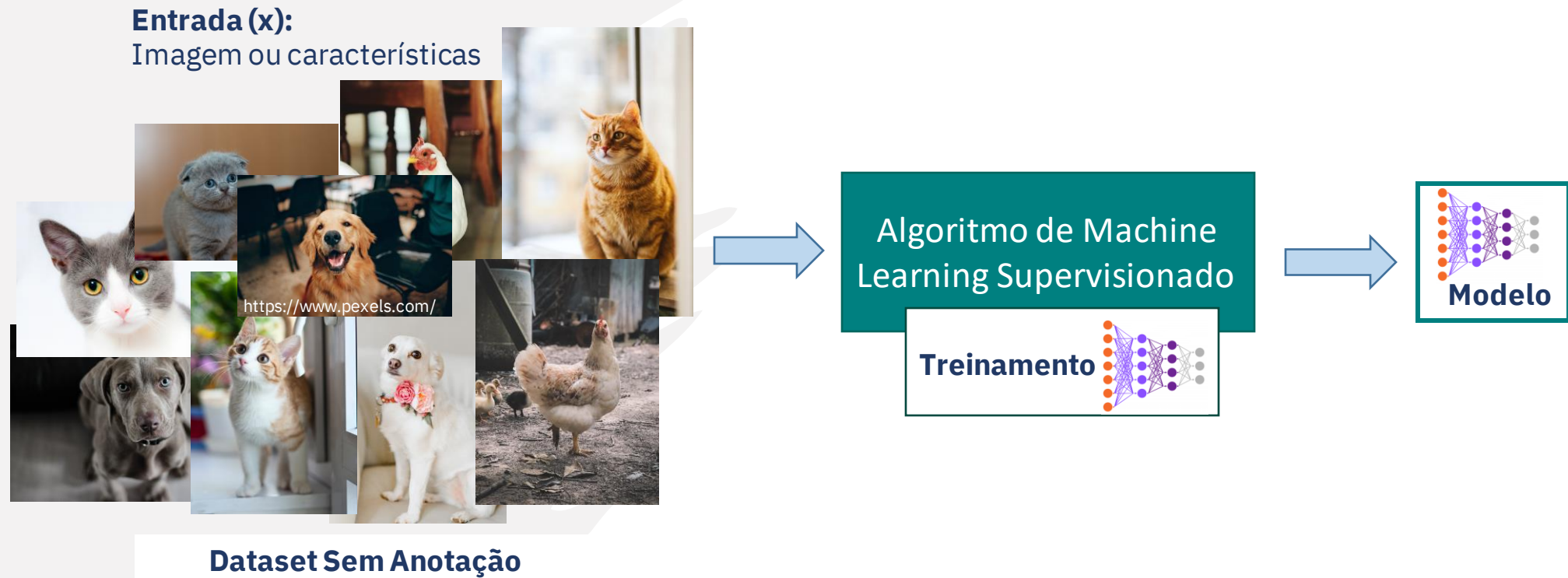
Não Aprendizado Supervisionado



Paradigmas de Aprendizado:
descrevem como
as máquinas
aprendem

Aprendizado Não Supervisionado

- **Não exige** que os **dados** estejam **rotulados**
- Sem crítica, **usa regularidades e propriedades estatísticas** dos dados.

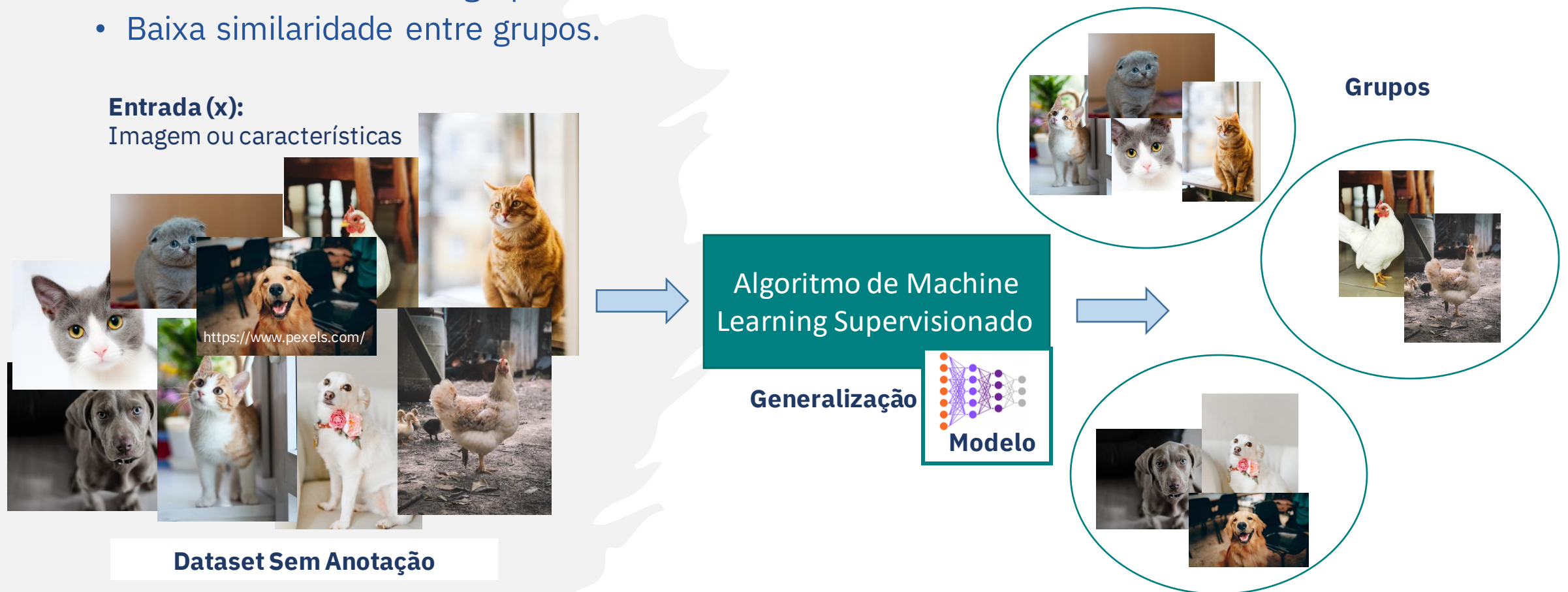


Aprendizado Não Supervisionado

- Executa **tarefas descritivas**: explora ou descreve um conjunto de dados.
- **Agrupamento**: divisão em grupos baseada em alguma regularidade ou similaridade
- **Sumarização**: descrição simples e compacta
- **Associação**: relações frequentes entre dados

Agrupamento

- Organiza dados (não classificados, sem rótulos) em grupos de acordo com alguma medida de similaridade, tal que exista:
 - Alta similaridade intra-grupo.
 - Baixa similaridade entre grupos.



Machine Learning:

Aprendizado Não Supervisionado



Dados não rotulados



Algoritmo de
Machine Learning



Resultados

sepal length	sepal width	petal length	petal width
5,1	3,5	1,4	0,2
4,9	3,0	1,4	0,2
7,0	3,2	4,7	7,1
6,4	3,2	4,5	1,5
6,3	3,3	6,0	2,5
5,8	2,7	5,1	1,9



Algoritmo
k-Means



Tarefa: agrupamento
(clustering)

Grupo 1

5,1	3,5	1,4	0,2
4,9	3,0	1,4	0,2

Grupo 2

7,0	3,2	4,7	7,1
6,4	3,2	4,5	1,5

Grupo 3

6,3	3,3	6,0	2,5
5,8	2,7	5,1	1,9



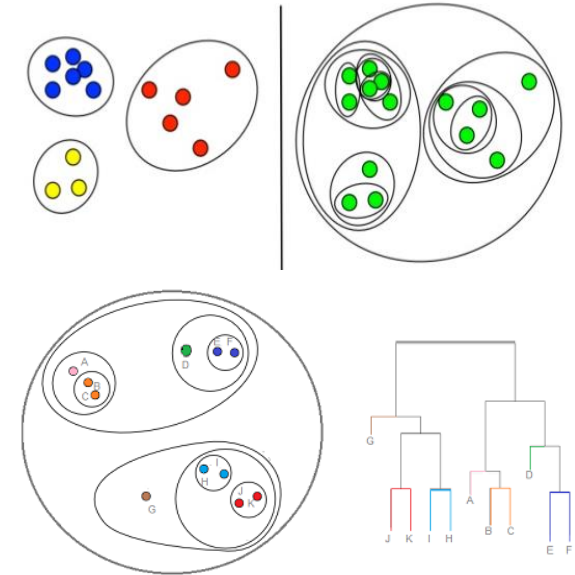
Machine Learning:

Aprendizado Não Supervisionado

**Agrupamento
produtos** de acordo
com as suas
características.



**Agrupamento de
clientes**
Identificação de perfil
para
recomendação de
produtos



Sumarização

68 Preliminarmente, requer-se a concessão dos benefícios
69 da justiça gratuita, nos termos do artigo 98 e seguintes do Código de Processo Civil,
70 uma vez que a Requerente está desempregada e não dispõe de recursos suficientes
71 arcar com as custas processuais ou qualquer despesas provenientes dos autos sem
72 comprometer sua subsistência e a de sua família, conforme declaração de
73 hipossuficiência e cópia da CTPS em anexo.
74 A lei não determina que tal benefício seja concedido
75 apenas àqueles que vivem miseravelmente, mas a todos os que se encontrem em
76 situação de insuficiência de recursos, ao passo que o dispêndio com o processo
77 representaria prejuízo ao sustento próprio e de sua família.
78 Deste modo, requer-se a concessão dos benefícios da
79 justiça gratuita, consoante ditames do artigo 98 e seguintes do Código de Processo
80 Civil.



Texto Original

Sumário Gerado



Resumo

Feito em <https://smodin.io/pt/resumidor-de-texto>

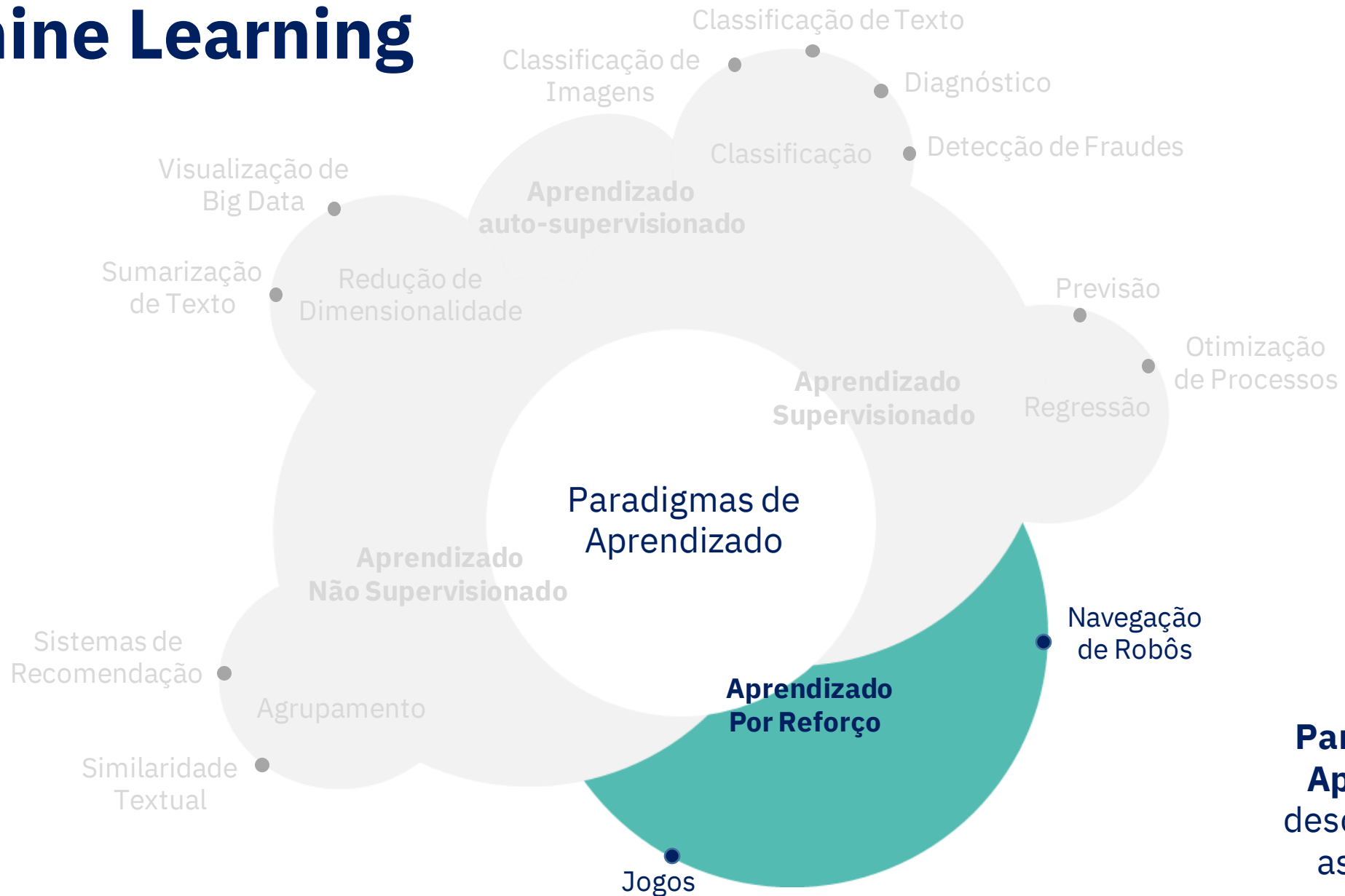
Preliminarmente, a concessão de benefícios é exigida nos termos dos artigos 98.º e seguintes. do Código de Processo Civil. A lei não determina que tal benefício seja concedido apenas a quem vive miseravelmente, mas a todos os que estão em situação de insuficiência de recursos. O gasto com o processo representaria prejuízo ao sustento dele e de sua família.

How do you like the result?



CÓPIA DE

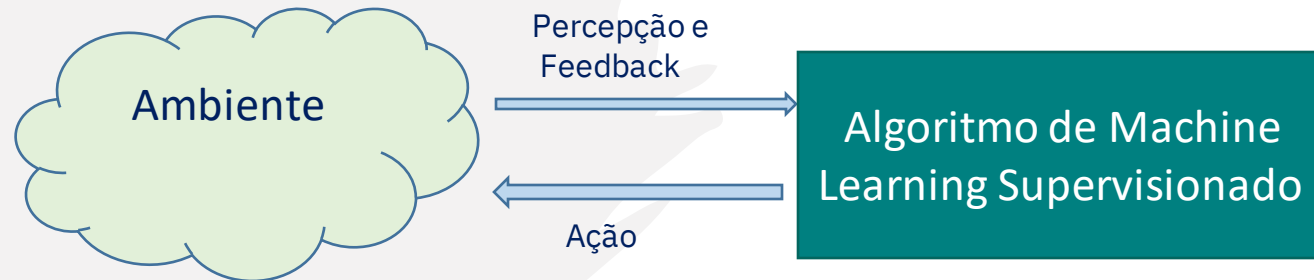
Machine Learning



Paradigmas de Aprendizado:
descrevem como
as máquinas
aprendem

Aprendizado por Reforço

- Processo de aprendizado baseado em punição e recompensa.
- Reforça uma ação positiva e penaliza uma negativa.
- Crítica apenas de desempenho.
- Aprendizado não depende de dados históricos
- Aprende a partir da experimentação por tentativa e erro

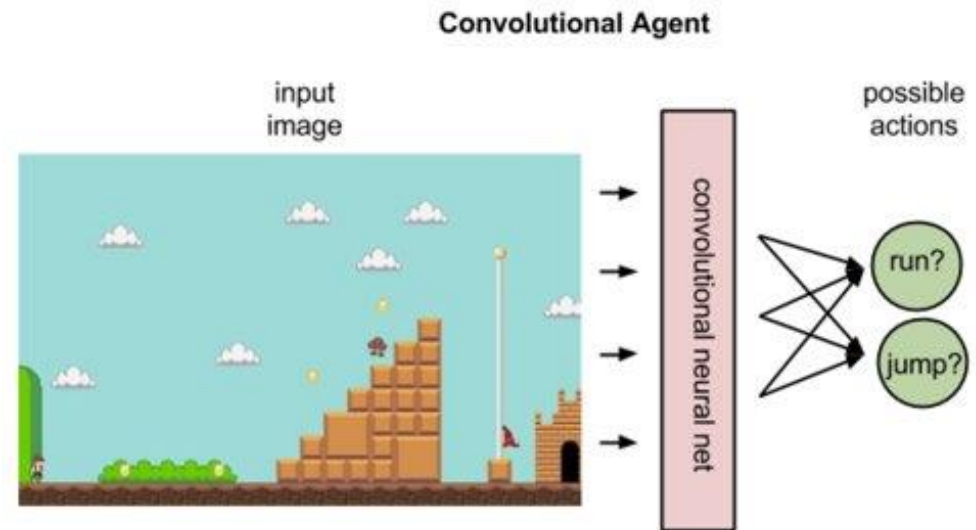


Feedback: indica apenas se a ação foi certa ou errada.

Aprendizado por Reforço

- **Aplicações:**

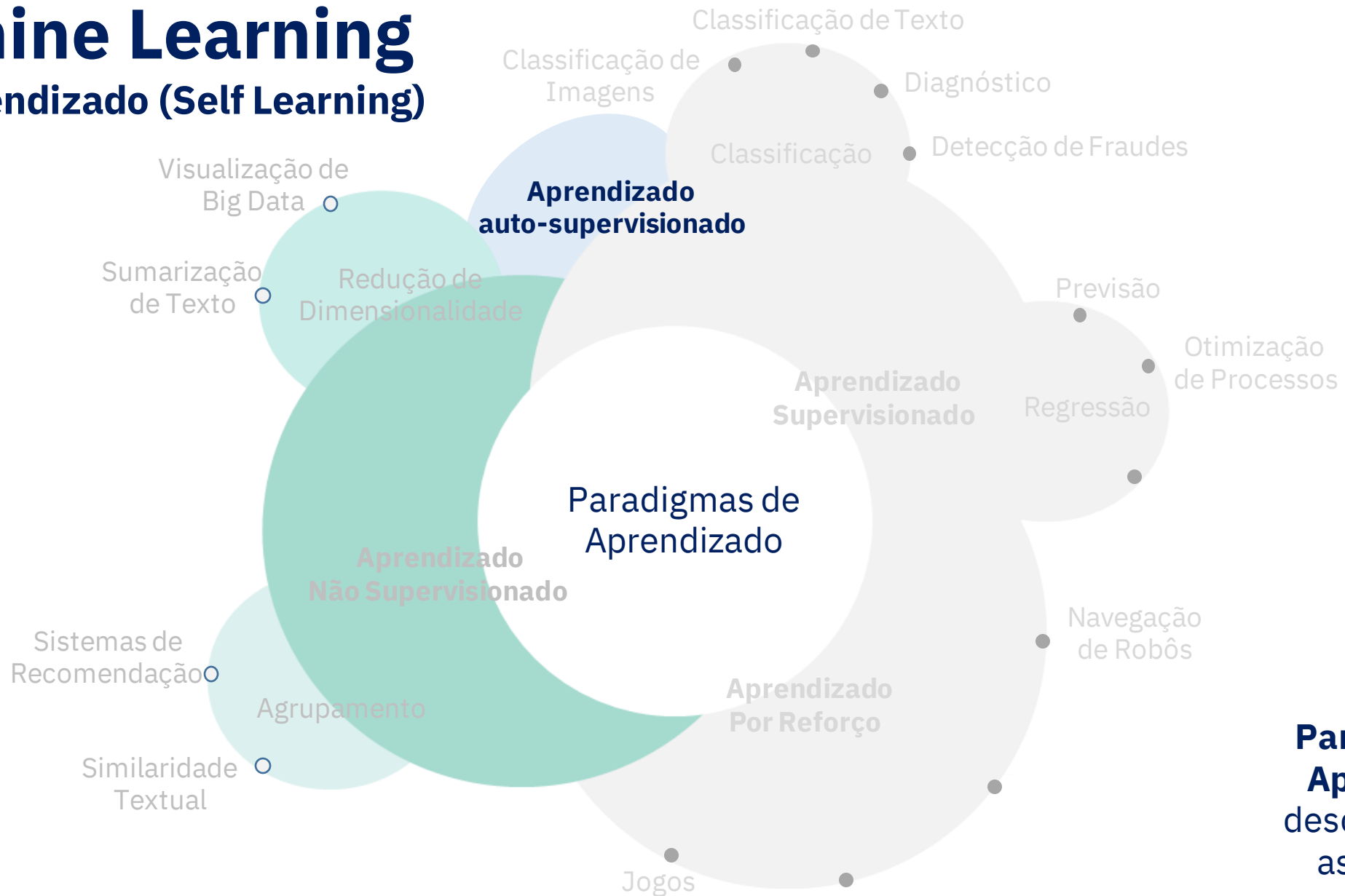
- Customização, personalização (aprende as preferências dos clientes ao longo das interações)
- Jogos
 - ALBERT: https://youtu.be/L_4BPjLBF4E



Fonte imagem: <https://wiki.pathmind.com/deep-reinforcement-learning>

Machine Learning

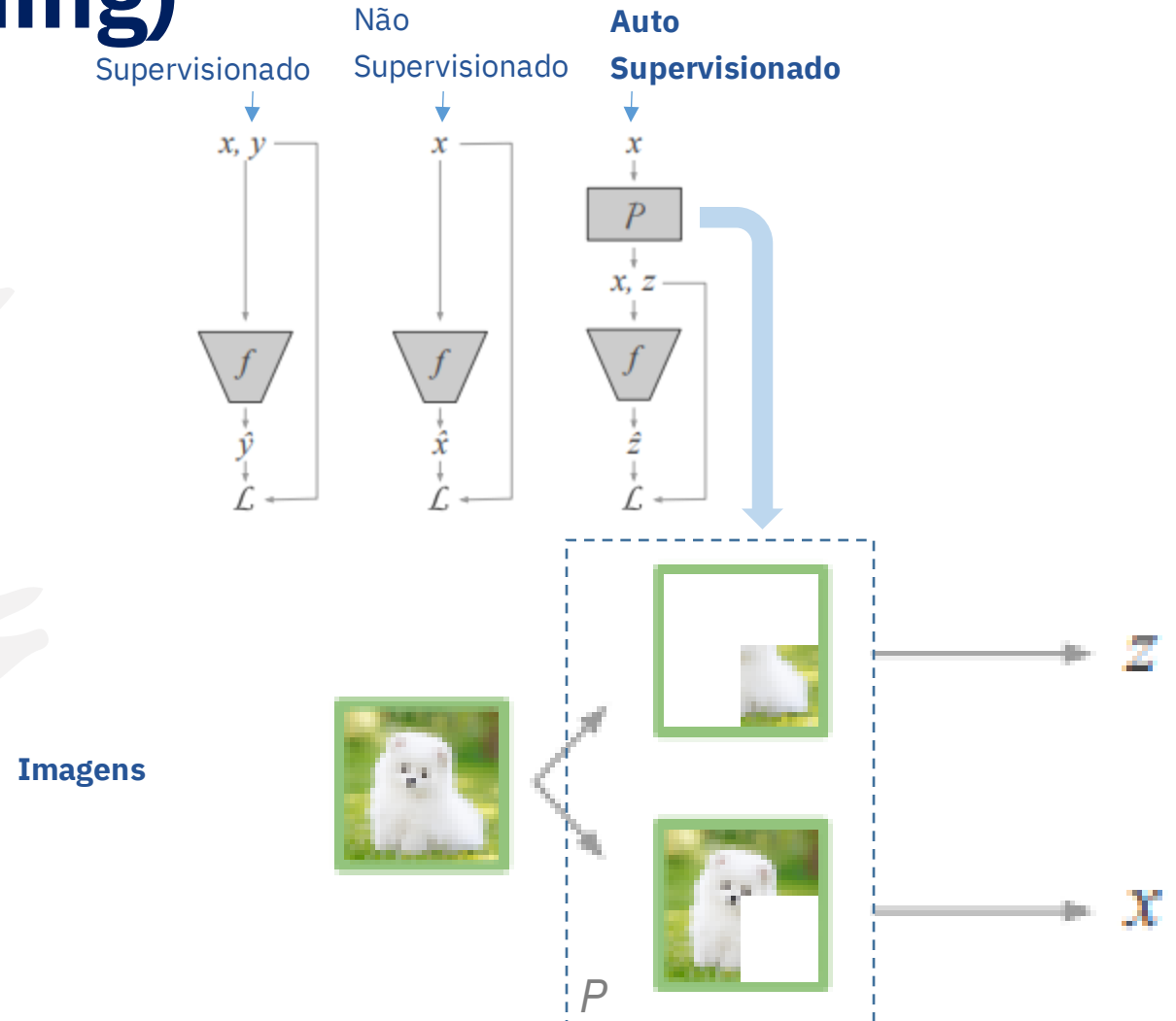
Autoaprendizado (Self Learning)



Paradigmas de Aprendizado:
descrevem como
as máquinas
aprendem

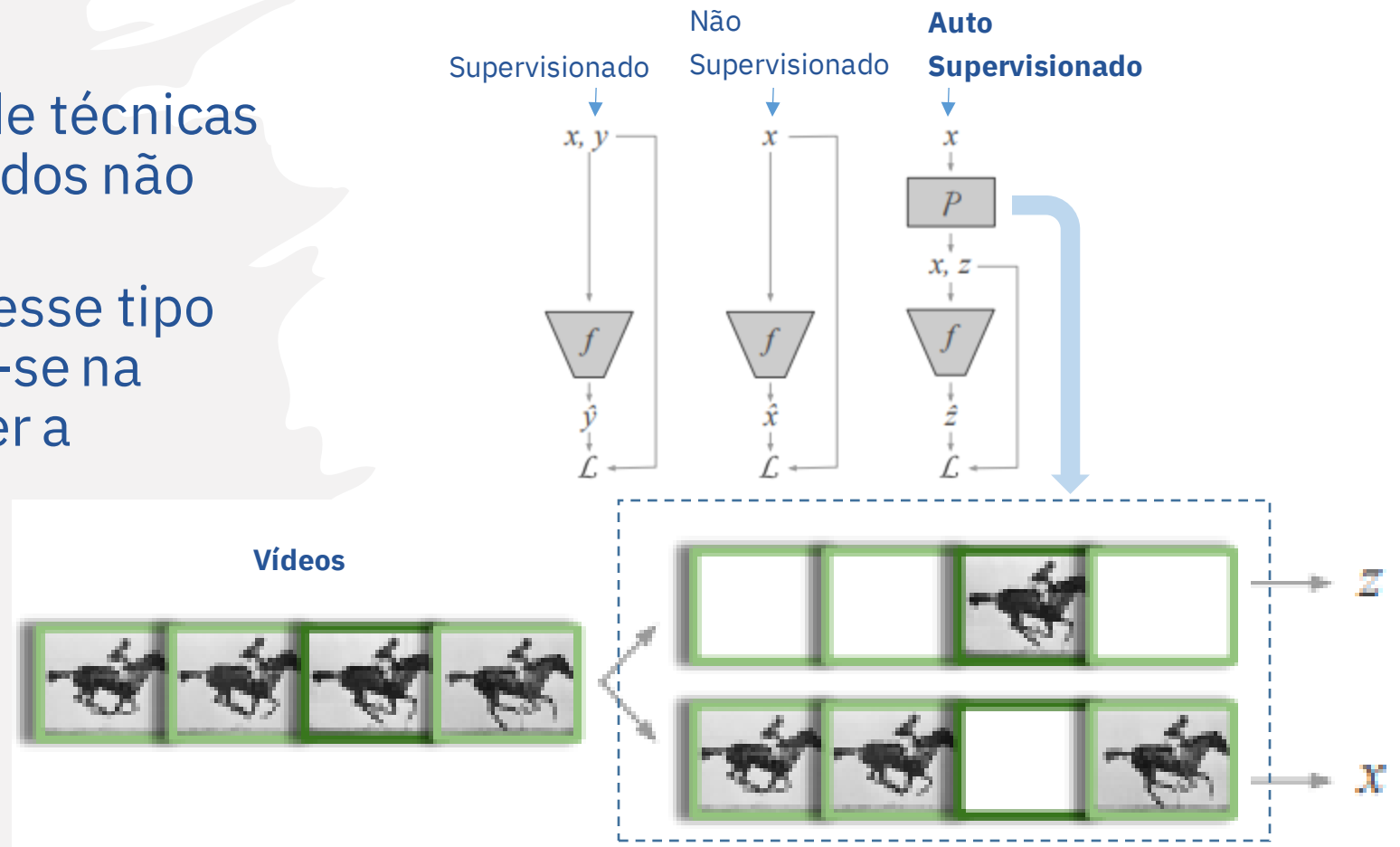
Aprendizado auto-supervisionado (Self Supervised Learning)

- Permite a aplicação de técnicas supervisionadas em dados não rotulados
- O conceito por trás desse tipo de aprendizado baseia-se na capacidade de aprender a preencher as lacunas.



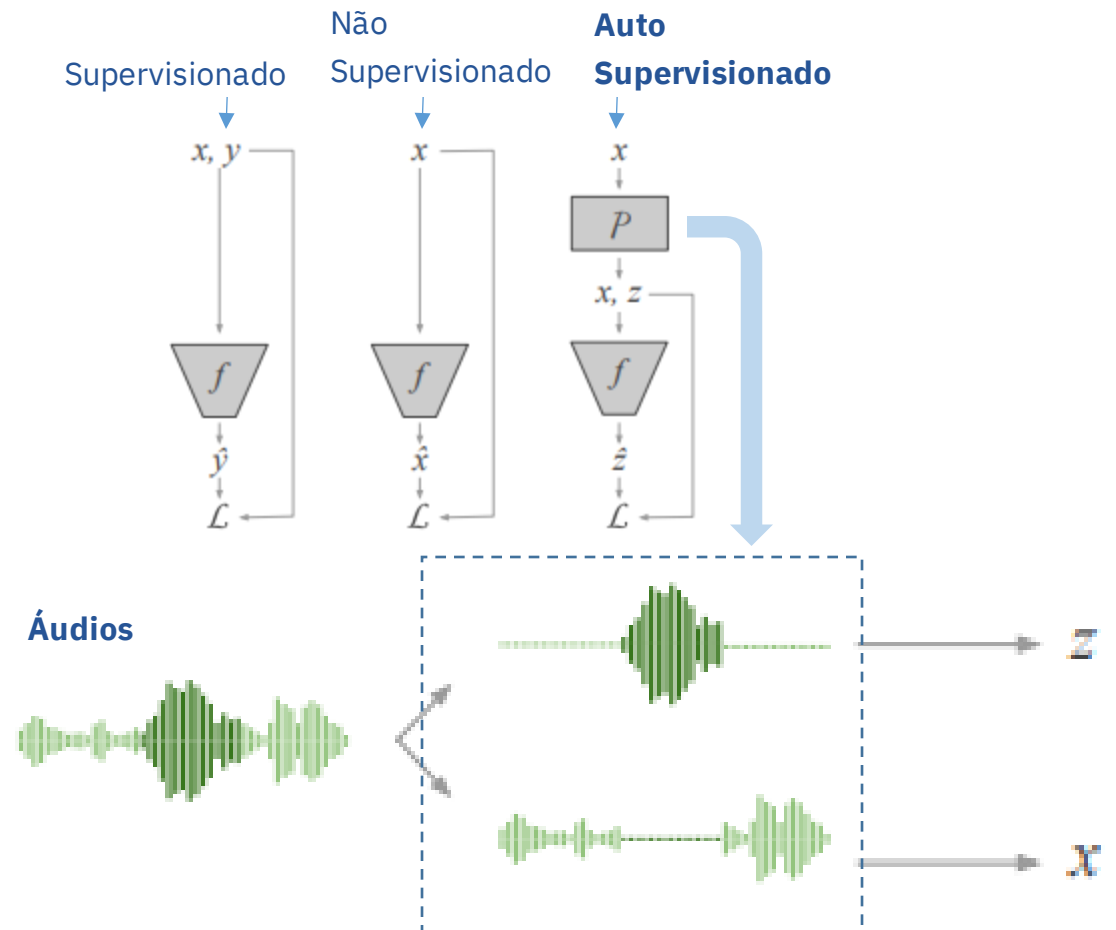
Aprendizado auto-supervisionado (Self Supervised Learning)

- Permite a aplicação de técnicas supervisionadas em dados não rotulados
- O conceito por trás desse tipo de aprendizado baseia-se na capacidade de aprender a preencher as lacunas.



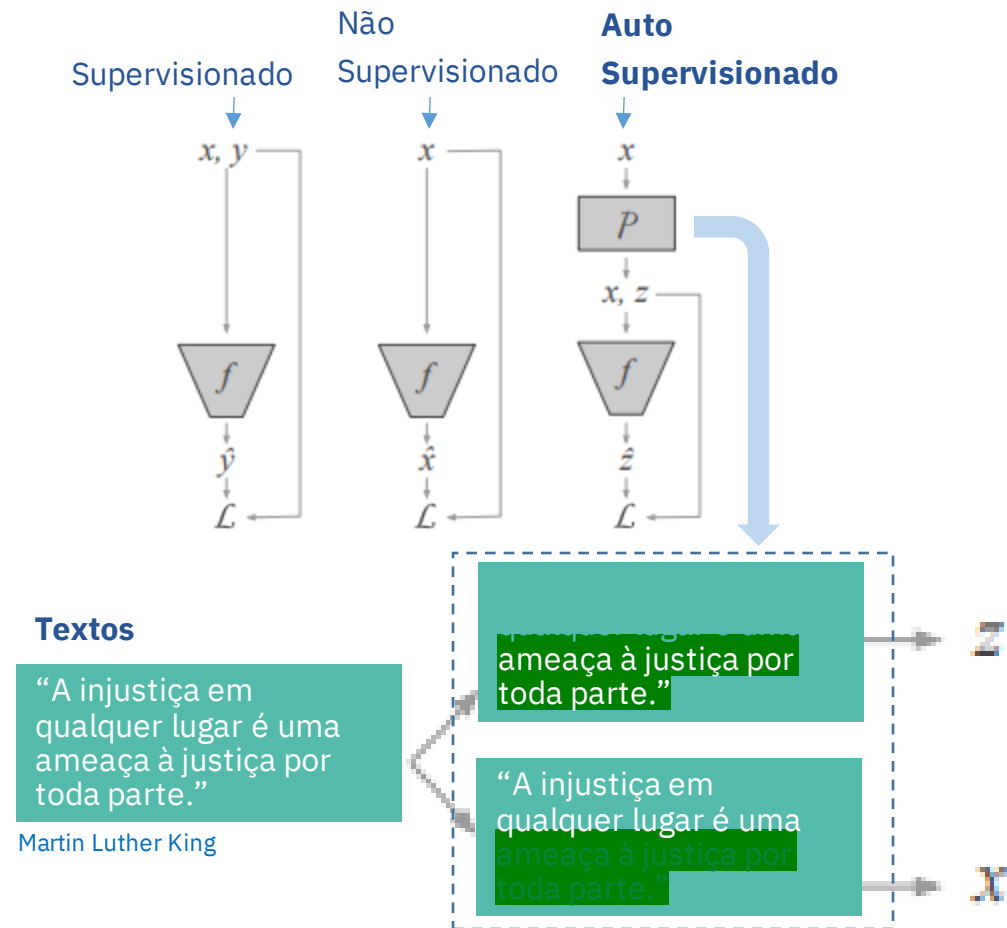
Aprendizado auto-supervisionado (Self Supervised Learning)

- Permite a aplicação de técnicas supervisionadas em dados não rotulados
- O conceito por trás desse tipo de aprendizado baseia-se na capacidade de aprender a preencher as lacunas.



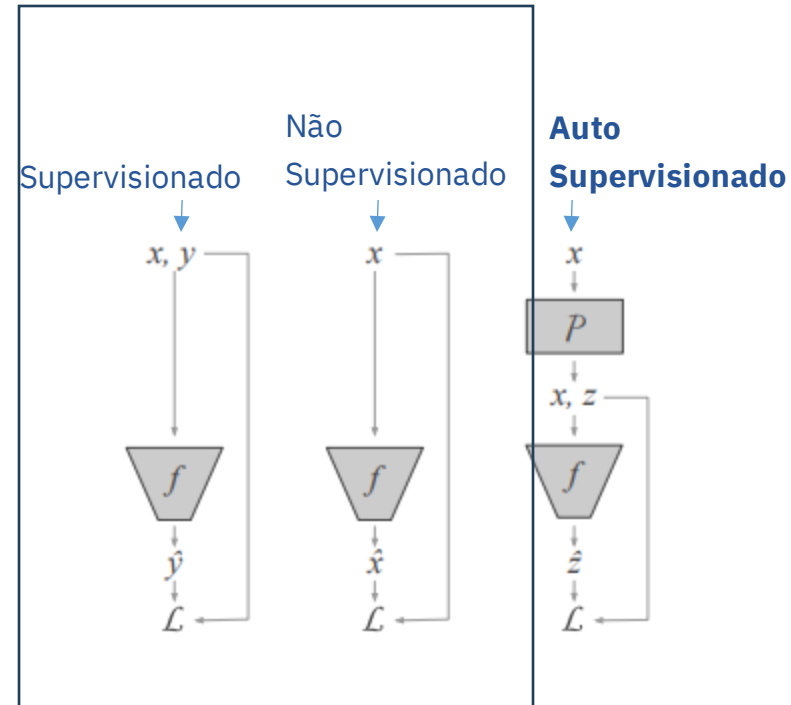
Aprendizado auto-supervisionado (Self Supervised Learning)

- Permite a aplicação de técnicas supervisionadas em dados não rotulados
- O conceito por trás desse tipo de aprendizado baseia-se na capacidade de aprender a preencher as lacunas.



Aprendizado semissupervisionado

- Híbrido, consegue trabalhar com dados rotulados e não rotulados



Machine Learning: Abordagens

- Abordagem (método) usado para aprender podem ser:
 - **Simbólica:** a capacidade de descrever os padrões extraídos em uma linguagem compreensível para os usuários. Usa descrições simbólicas. Ex: árvores de decisão
 - **Conexionista:** inspiradas no modelo biológico do sistema nervoso. Ex: redes neurais:
 - **Estatística:** usa modelos estatísticos para encontrar uma boa aproximação da hipótese. Ex: redes bayesianas
 - **Evolutiva:** baseada na teoria da evolução de Darwin. Ex: algoritmos genéticos

Etapas de Desenvolvimento de um Projeto de IA



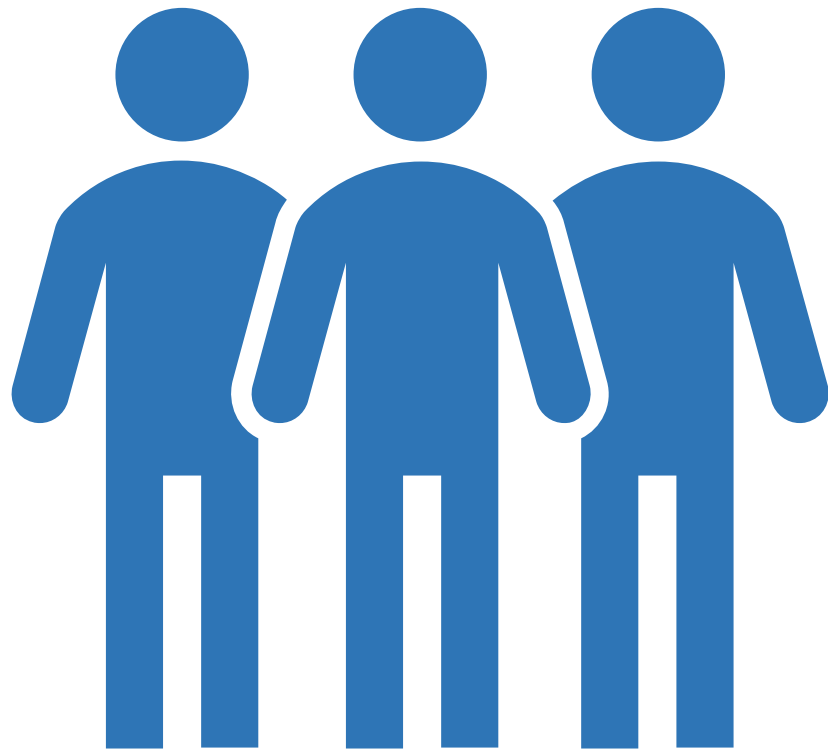
“A automação reduzirá o tempo de execução de tarefas que hoje são impossíveis para humanos... a quantidade de dados que as empresas possuem está além da capacidade do profissional ou cientista de dados mais sofisticado. Mas isso não é verdade para o aprendizado de máquina...”

Mark Hurd
CEO da Oracle.

“Existem muitas oportunidades para projetos valiosos de IA e uma das maneiras mais eficientes de executar projetos valiosos de IA é construir novas equipes e empresas.”

Andrew Ng

Professor de Ciência da Computação da Universidade de Stanford



1) Forme uma equipe

Para executar o projeto de IA, você precisará de uma **equipe multidisciplinar**:

- **Desenvolvedores** (TI, Analistas, Cientistas de Dados, ...)
- **Profissionais da área do Negócio** (que conheçam o domínio da aplicação)

2) Identifique o problema



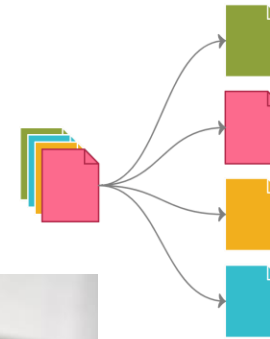
1. Escolha um **problema** cuja solução exige, no momento, **um esforço manual repetitivo considerável**.
2. Se for o seu primeiro projeto de IA, escolha algo mais simples mas que a **automatização e solução de IA agreguem valor ao trabalho dos usuários**, apoiando e acelerando a execução das tarefas.
3. Verifique se você **dispõe dos dados necessários** para a construção da solução em IA.

2) Identifique o problema



4. Entenda as **características dos problema** para determinar se a solução exigirá a construção de um modelo de IA:

- supervisionado: classificação, regressão, ...
- não supervisionado: agrupamento, ...
- por reforço



3) Construa um dataset

- Crie um **Dataset de Referência** (*benchmark*) para construir a sua solução de IA. Um **Dataset** é um conjunto de dados.
- Para isso, defina as **fontes de dados** que podem conter os dados necessários para a sua solução.
- **Estabeleça critérios e selecione dados** relevantes para sua solução, considerando volume, qualidade e adequação.

3) Construa um dataset



- **Os dados podem ser estruturados ou desestruturados (não estruturados).**

Dados estruturados são informações claramente definidas que incluem padrões e parâmetros facilmente pesquisáveis. São metadados, como por exemplo, nomes, endereços, datas de nascimento, números de telefone e de documentos.

Dados não estruturados não têm padrões, consistência ou uniformidade. Inclui textos, áudios, imagens, infográficos e e-mails.

Dados semiestruturados possuem algum formato. Ex: html



3) Construa um dataset

- **Execute ETL (*Extract, Transform and Load*):** processo de extração, transformação e carga de dados para construir o dataset.



Dataset é um conjunto de dados que foi identificado com útil para solução, extraído das suas fontes originais, preparado para uso e armazenado em um local de fácil acesso e manipulação.

3) Construa um dataset

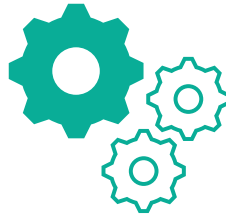


Execute ETL (*Extract, Transform and Load*): processo de extração, transformação e carga de dados para construir o dataset.



Extract

(Extração) de diferentes fontes de dados



Transform

(Transformação) dos dados consiste na preparação para uso.
Ex: conversão de formato pdf para txt.



Load

(Carga) dos dados em um repositório.

Anotação de dados manual exige tempo e esforço da equipe da área do negócio.

3) Construa um dataset



- No caso de **abordagens supervisionadas**:
 - **defina as classes** previamente;
 - verifique se as classes podem ser associadas aos dados partir de metadados ou se será necessária a anotação manual para isso.

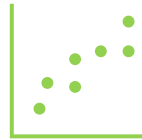
O processo de anotação também pode ser realizado de forma automática ou semiautomática (anotação realizada meio de um modelo de IA e revisada, posteriormente, manualmente).

Anotação de dados manual exige tempo e esforço da equipe da área do negócio.

4) Analise o dataset



- Realize **EDA** (*Exploratory Data Analysis*), ou seja, **Análise Exploratória de Dados**;
 - Etapa importante que visa investigar o conjunto de dados para descobrir padrões e anomalias (outliers) e formar hipótese;
 - Envolve a **geração de estatísticas e representações gráficas** para entender melhor os dados.



4) Analise o dataset

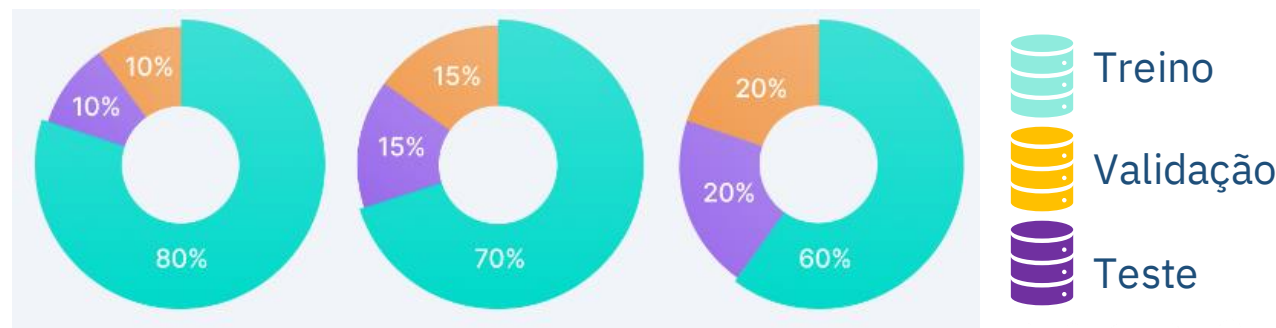


- Verifique se a **diversidade e quantidade de dados é suficiente** para construção da solução.
- No caso de abordagem supervisionada, verifique se o **dataset está balanceado**. Para estar balanceado, a quantidade de dados de cada classe deve ser a mesma ou muito próxima. Isso ajuda a evitar o viés.

Soluções de IA que usam algoritmos de Machine Learning são dependentes dos dados usados no seu treinamento. Portanto, para construir bons modelos de IA, seja criterioso na definição do seu dataset.

5) Divida o dataset

- **Divida o dataset** (harmonicamente, no caso de classes), em conjuntos disjuntos de:
 - **treino**: usado para o treinamento do modelo;
 - **validação**: usado para validar o modelo durante o treinamento;
 - **teste**: usado para verificar a solução, ou seja, a versão final do modelo, após o treinamento .

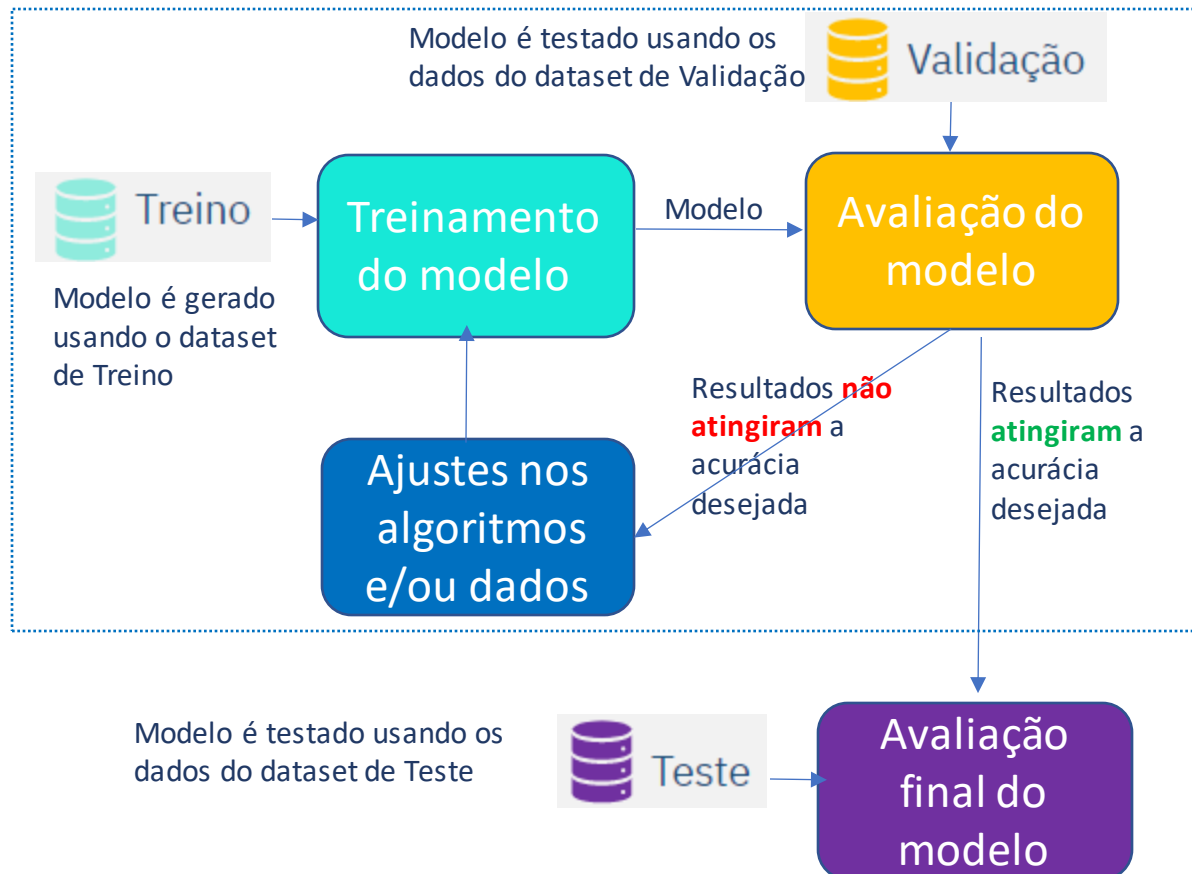


Fonte da Figura: <https://www.v7labs.com/blog/train-validation-test-set#:~:text=Training%20data%20is%20the%20set,after%20completing%20the%20training%20phase>.

5) Divida o dataset



- Entenda como os dados são usados para construção do modelo:



5) Divida o dataset



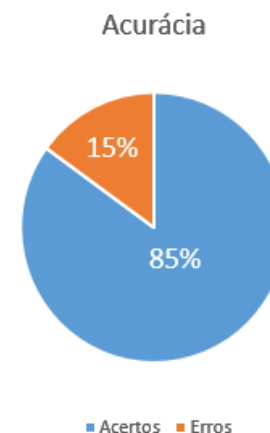
É importante viabilizar a avaliação do modelo de IA.

- Escolha a métrica adequada.

Ex: **Acurácia**: é uma das métricas usadas para medir o desempenho dos modelos classificadores.

$$\text{Acurácia} = \frac{\text{Número de predições corretas}}{\text{Total de predições}}$$

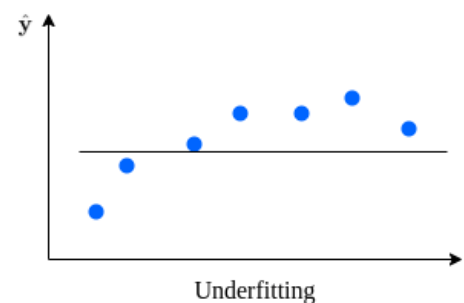
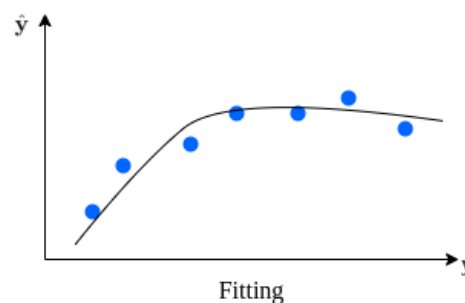
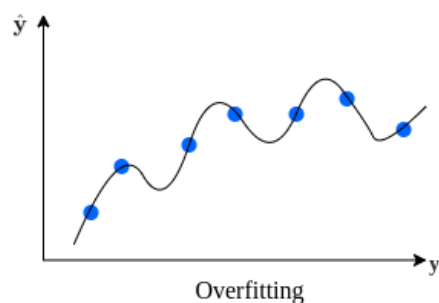
Se para os 100 dados do conjunto de validação (ou teste), o modelo acertar 85, a acurácia será de 0,85 ou de 85%.



5) Divida o dataset



- Os datasets de **treino/validação/teste** ajudam ainda a identificar o *overfitting* e *underfitting*.
 - **Overfitting**: é quando o modelo memoriza o dataset de treino e não consegue ter bom desempenho com dados não vistos antes.
 - **Underfitting**: é quando o modelo não aprende nem os padrões existentes nos dados de treino.



6) Construa a solução



- Delimite o escopo da sua solução, estabelecendo objetivos claros quanto à sua finalidade;
- Estude possibilidades e defina a arquitetura da sua solução;
- Escolha algoritmos, recursos, materiais e ferramentas adequados para apoiar o desenvolvimento;
- Sempre que possível, incorpore à sua arquitetura modelos já existentes e de qualidade comprovada. Isso reduzirá o tempo de desenvolvimento.
 - Valide sua solução junto aos usuários e aproveite os feedbacks;
 - Pense também em como essa solução se integrará aos sistemas existentes;
 - Considere aspectos como desempenho e escalabilidade.

Projetos de IA exigem experimentação. Portanto, precisam de tempo para serem desenvolvidos e exigem constante aprimoramento.