Olá, aluno(a)!
Seja bem-vindo(a) à aula interativa!

Você entrará na reunião com a câmera e o microfone desligados.

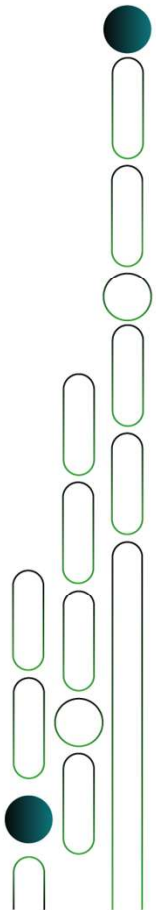
Sua presença será computada através da enquete.
Fique atento(a) e não deixe de respondê-la!

Aprenda com quem faz

> Coleta e Armazenamento de Dados de Renda Fixa

Segunda Aula Interativa

Prof. Joao Paulo Faria



Nesta aula

- Revisão do Conteúdo da 2ª Parte do Módulo
- Correção do Desafio
- Ciência de Dados Aplicada

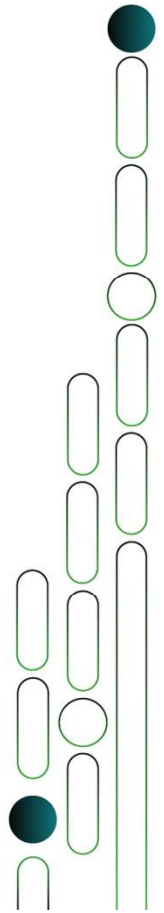
Dúvidas sobre o conteúdo do módulo

5. Fundamentos de Engenharia de Dados
    a) Tipos de dados e modelos de dados
    b) Modelos de dados: caso de uso

6. Pipeline de Ciência de Dados
    a) Pipeline de Ciência de Dados

Dúvidas sobre o conteúdo do módulo

7. Processamento de Linguagem Natural
    a) Introdução ao Processamento de Linguagem Natural
    b) Motivação do Desafio
    c) Demonstração: aplicando NLP às comunicações do Banco Central
    d) Revisão e Apresentação do Desafio

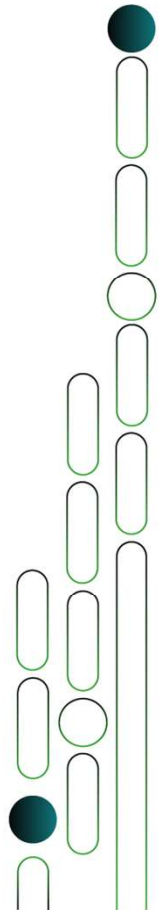


Correção do Desafio

Os alunos deverão desempenhar as seguintes atividades:

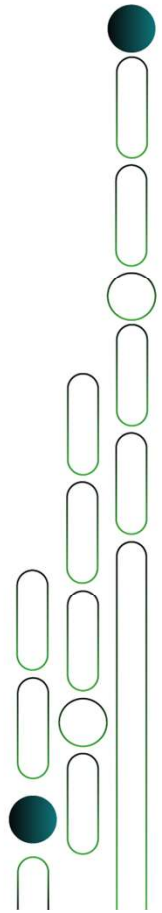
- Exercitar os seguintes conceitos trabalhados no Módulo:
  - 1. Tipos de dados e modelos de dados
  - 2. Pipeline de Ciência de Dados
  - 3. Processamento de Linguagem Natural

Fundamentos de Aprendizado de Máquina e Datacentric AI

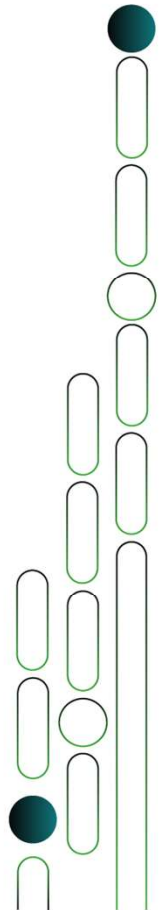


Fundamentos de Aprendizado de Máquina e Data Centric AI

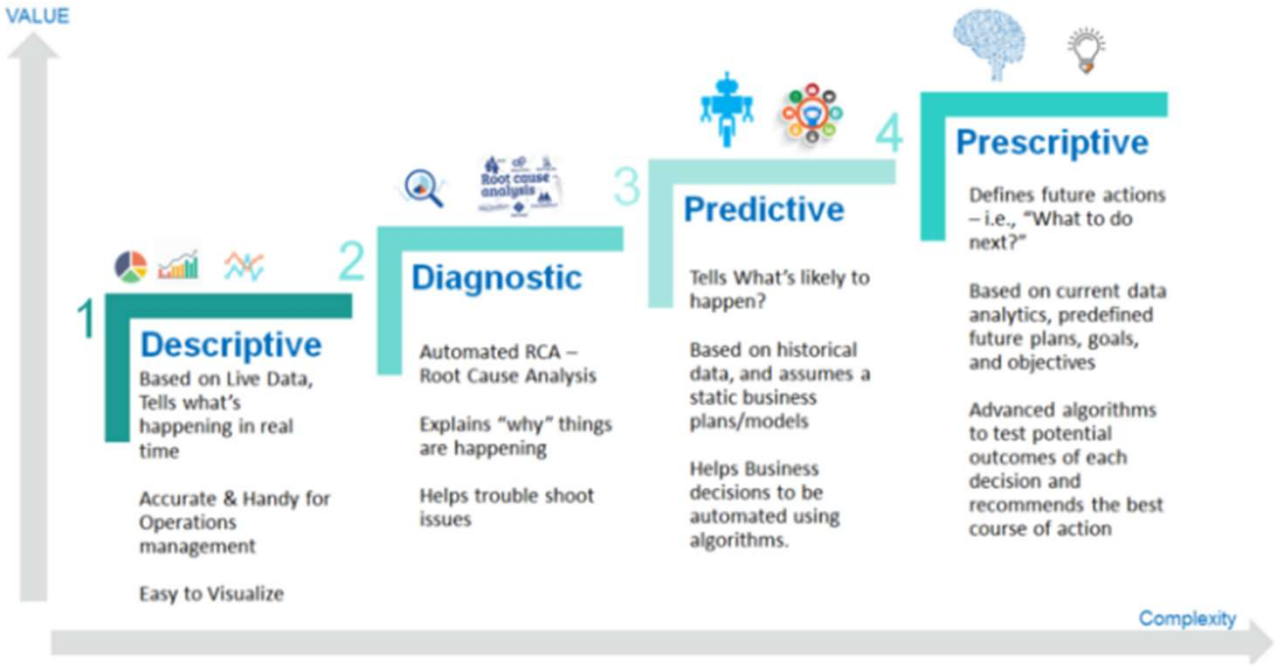
- Fundamentos de Aprendizado de Máquina
  - O que é Aprendizado de Máquina
  - Aprendizado Supervisionado
  - Underfitting e Overfitting
- Data Centric AI
  - Típico sistema de Machine Learning
  - Data Centric AI: exemplo real
  - Model vs Data Centric AI



Parte I: Fundamentos de Aprendizado de Máquina

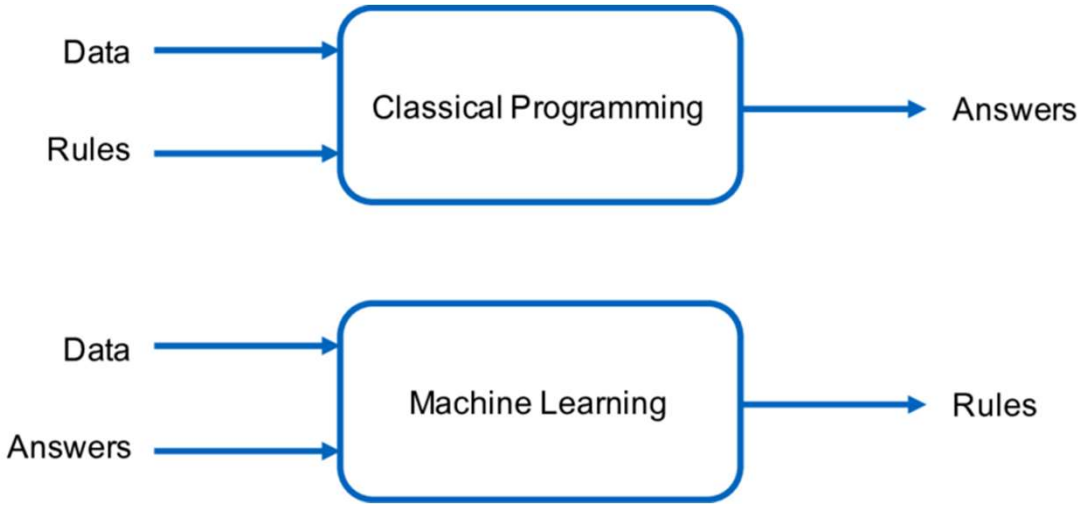


Tipos de Data Analytics



O que é Aprendizado de Máquina?

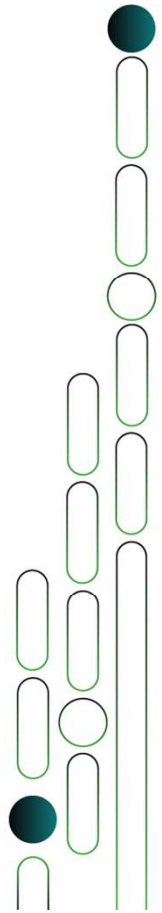
- Algoritmo capaz de aprender a partir dos dados.



O que é Aprendizado de Máquina?

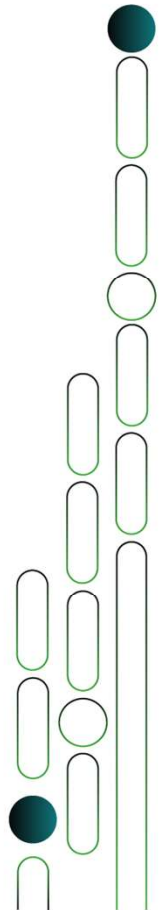
- Algoritmo capaz de aprender a partir dos dados.
  - Mas o que é aprender?

    “Se diz que um programa de computador aprende da experiência E com relação a determinada classe de tarefa T e métrica de desempenho P se seu desempenho com relação à tarefa T, mensurado por P, melhora através da experiência E”.

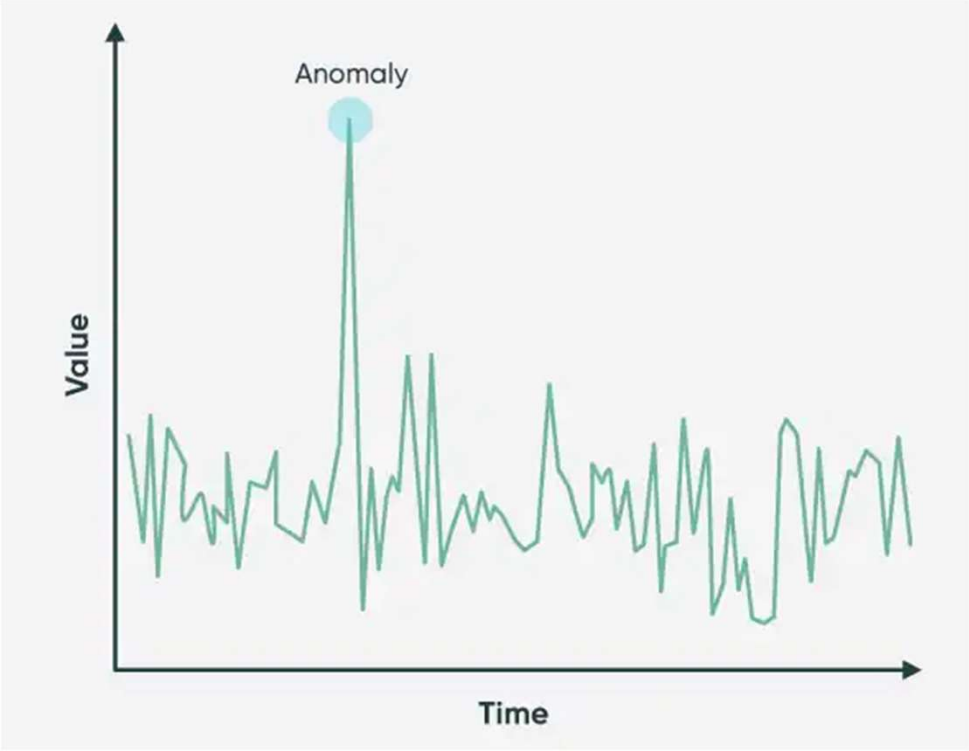


O que é Aprendizado de Máquina?

- Tarefa T
  - Classificação
  - Regressão
  - Detecção de Anomalia
- Métrica P
  - Acurácia
  - Erro Quadrático Médio (MSE)
- Experiência E
  - Dataset: conjunto de exemplos



O que é Aprendizado de Máquina?



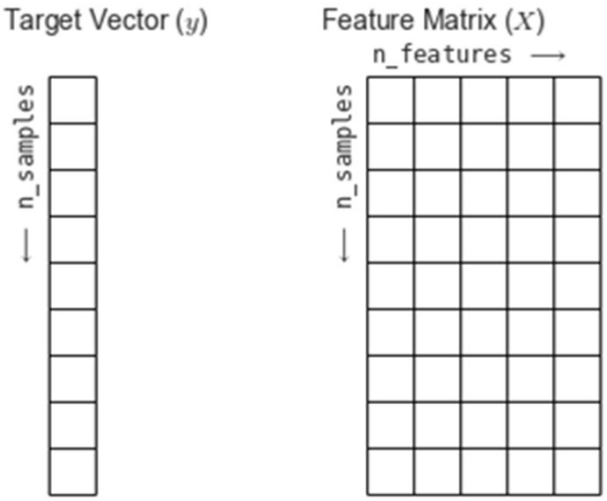
Aprendizado Supervisionado

$$y = f(x)$$

Target / Label
Output
Variável resposta

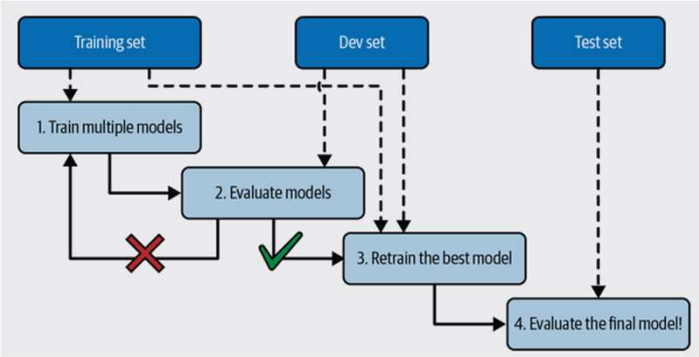
Input
Features
Variáveis explicativas

$$D = \{(x, y)\}$$



Objetivo

- Bom desempenho em *dados novos, nunca antes vistos.*
  - Erro de Treinamento
  - Erro de Generalização (Erro no conjunto de teste)

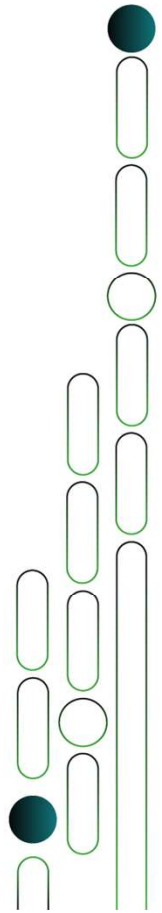


Objetivo

- Como moderar o desempenho de um modelo em um conjunto de dados não observável?

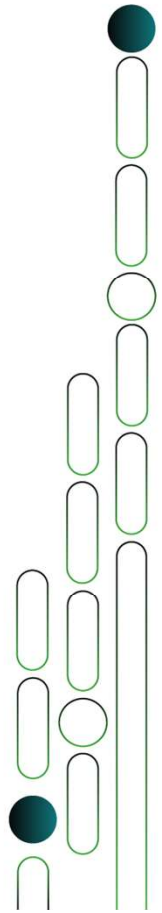
Disciplina conhecida como “Aprendizado Estatístico”

- Fatores determinarão o desempenho do modelo?
  - Minimizar erro de treinamento
  - Minimizar diferença entre erro de treinamento e erro de generalização

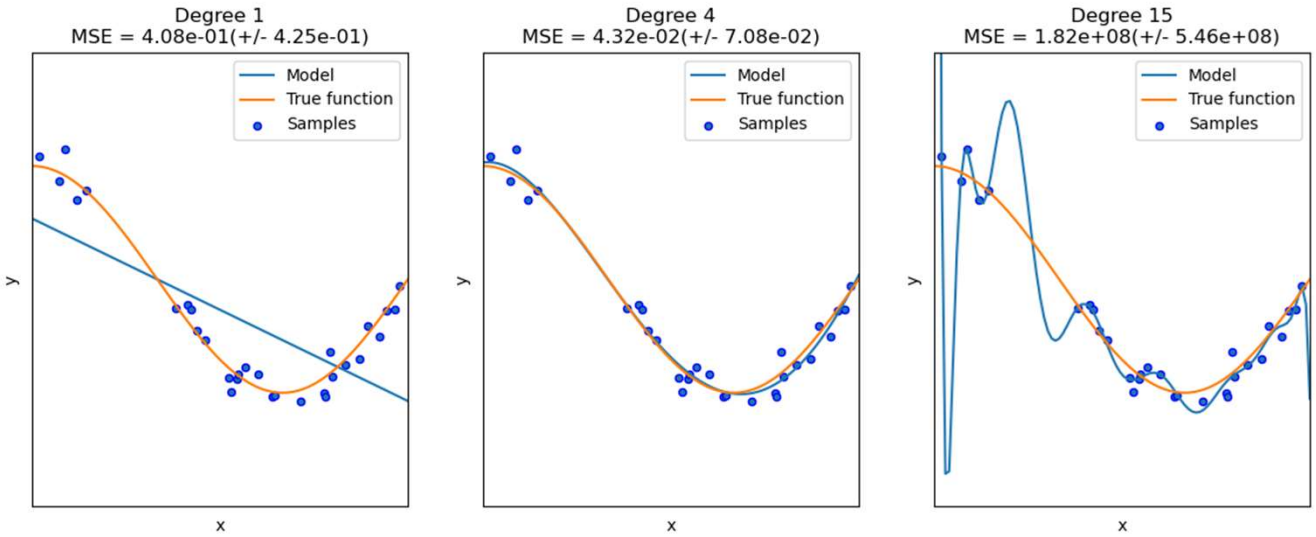


Underfitting e Overfitting

- *Underfitting*: modelo não é capaz de atingir valor suficiente, baixo de erro de treinamento.
- *Overfitting*: diferença entre erro de treinamento e de generalização é muito grande (i.e. modelo não é capaz de generalizar).
- Capacidade: métrica para classificarmos o produto do algoritmo.
    - Baixa Capacidade: dificuldade de ajustar-se ao dataset de treino.
    - Alta capacidade: pode sobreajustar, memorizando propriedades do dataset de treino que não serão úteis para outros dados.



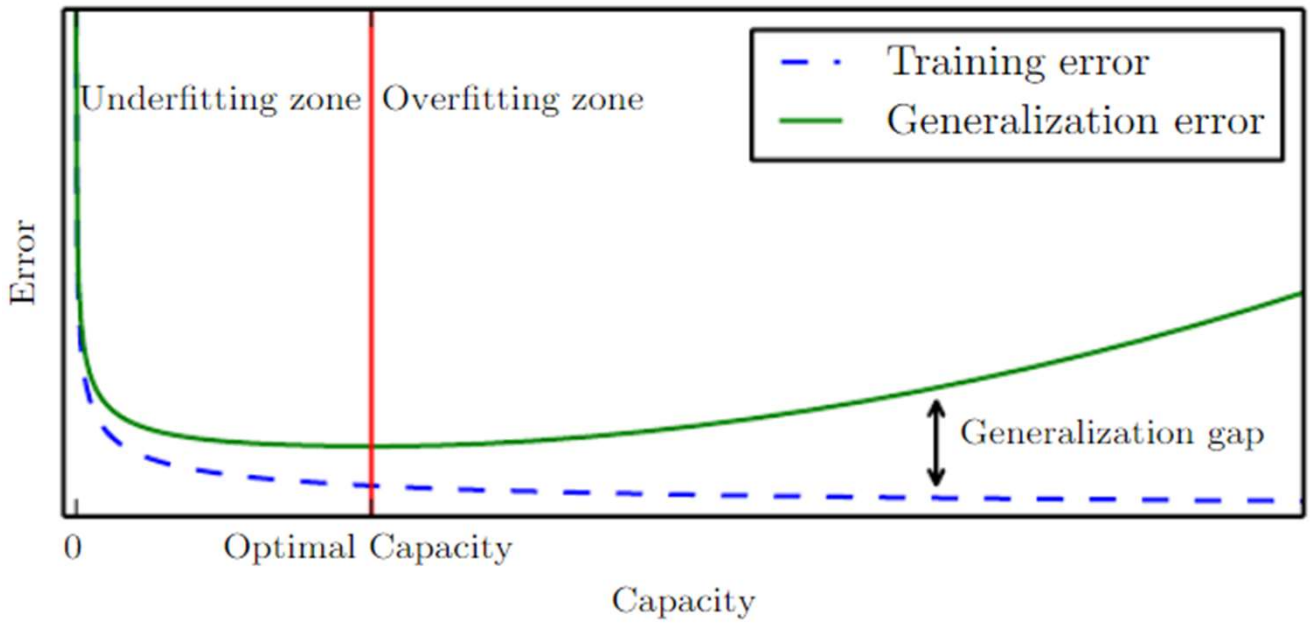
Underfitting, Overfitting e Capacity



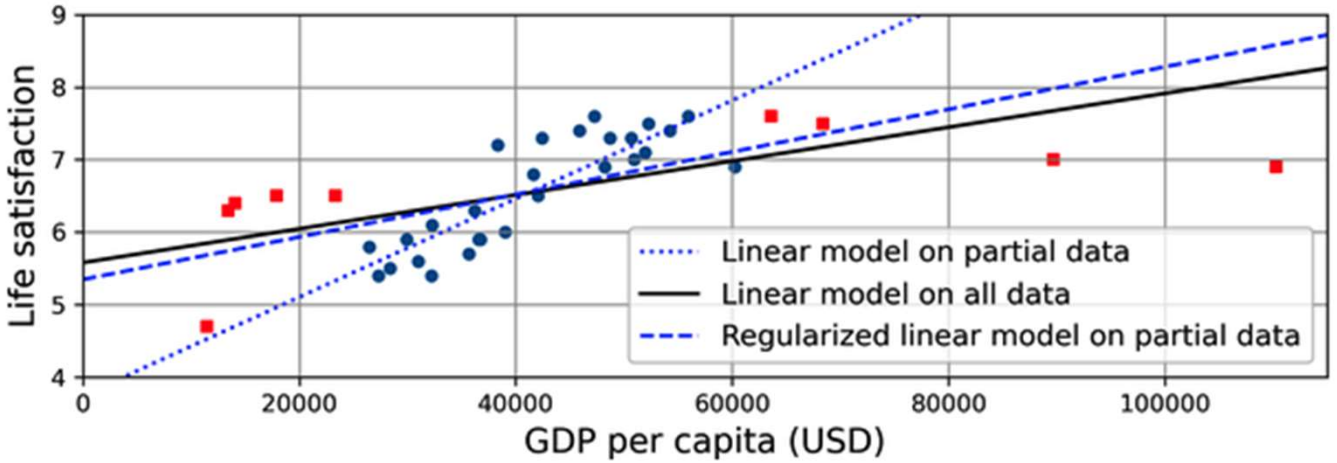
– Capacidade

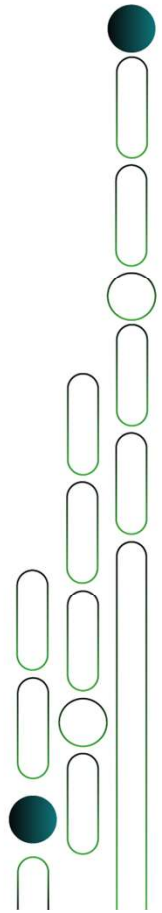
+ Capacidade

Underfitting, Overfitting e Capacity



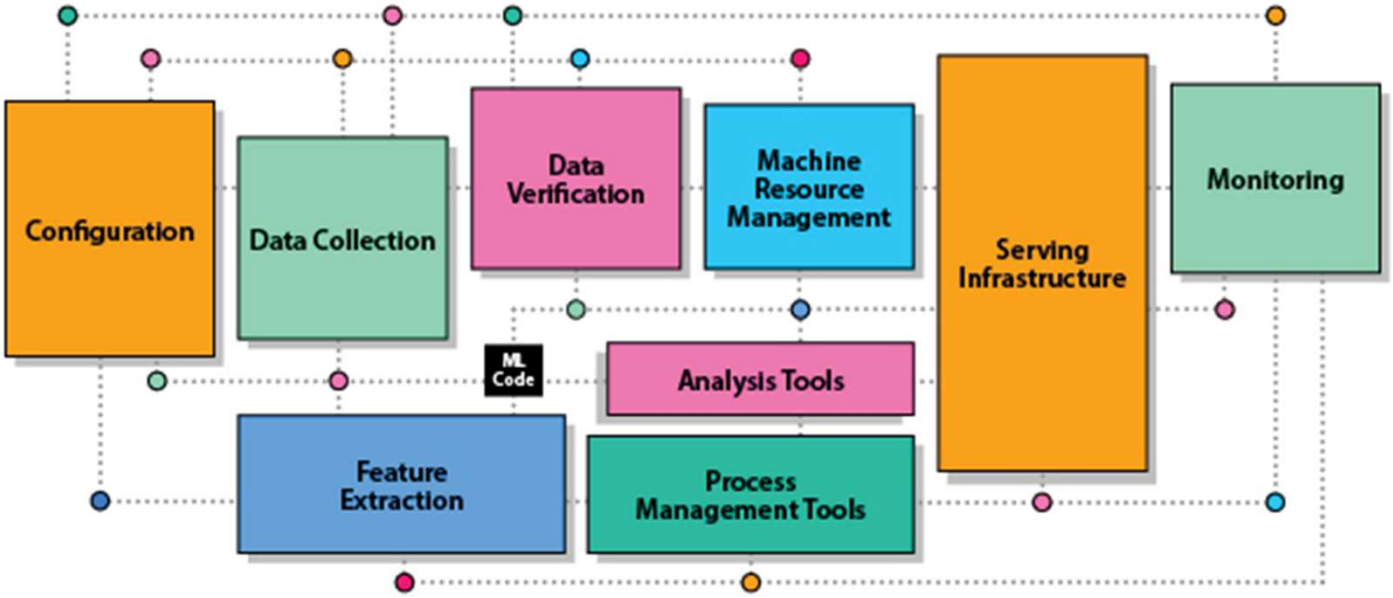
Regularização



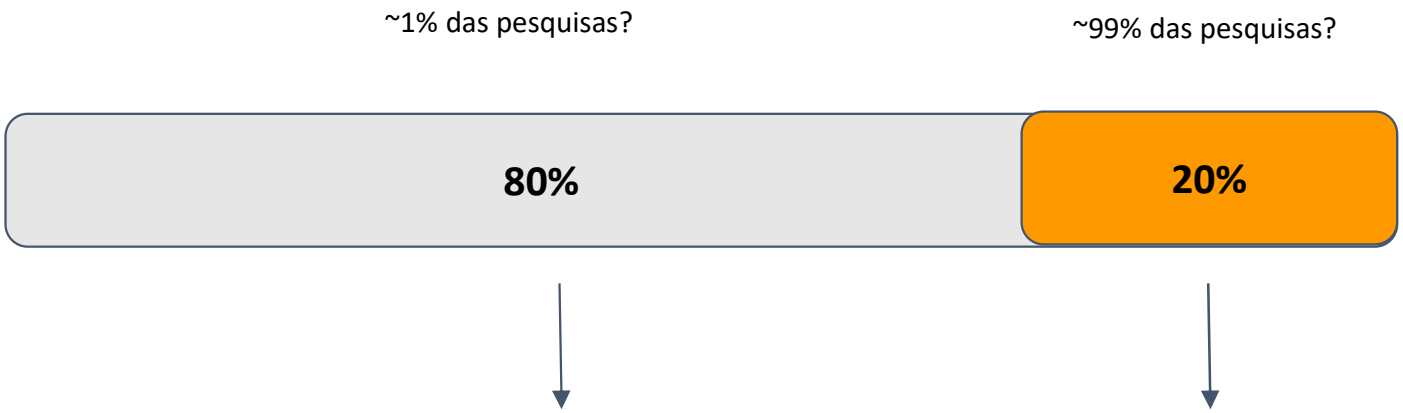


Parte II: Data Centric AI

Pipeline de Machine Learning



“Data is Food for AI”



Preparação

Obter e preparar ingredientes de alta qualidade

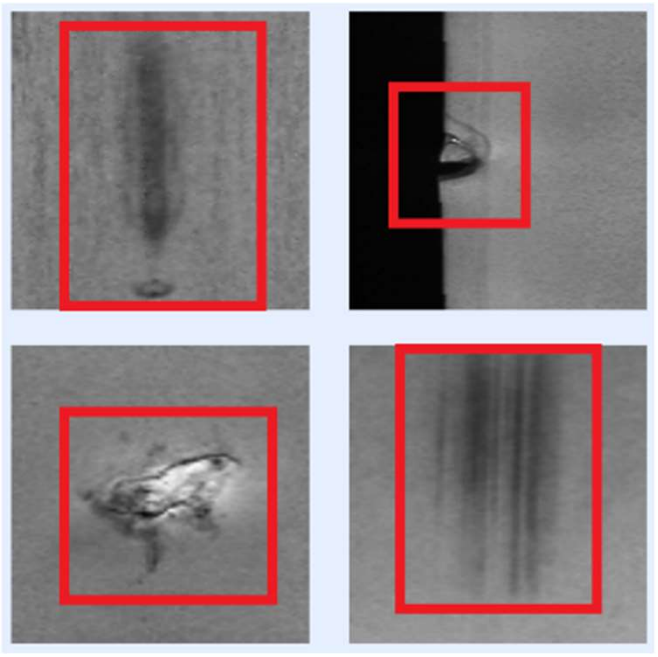
Obter e preparar dados de alta qualidade

Ação

Cozinhar

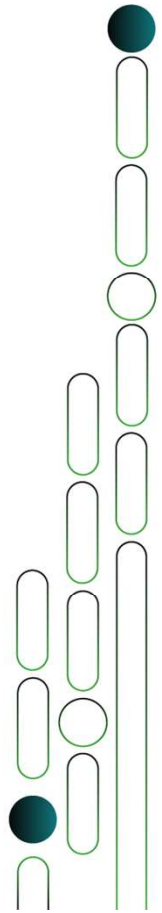
Treinar um modelo

Data Centric AI: Exemplo Real



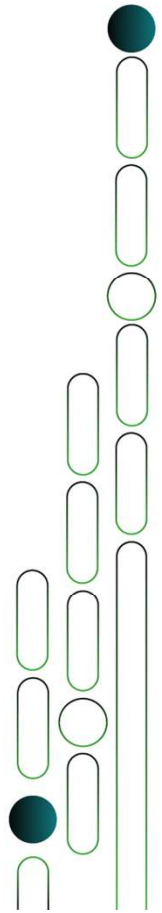
Data Centric AI: Exemplo Real

Modelo de detecção de defeito	**Acurácia**
Baseline	76,2%
Model-centric	+0%
Data-centric	+16,9% (93,1%)



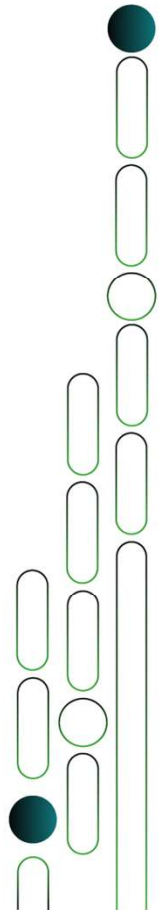
Model vs. Data Centric AI

AI = Model + Data

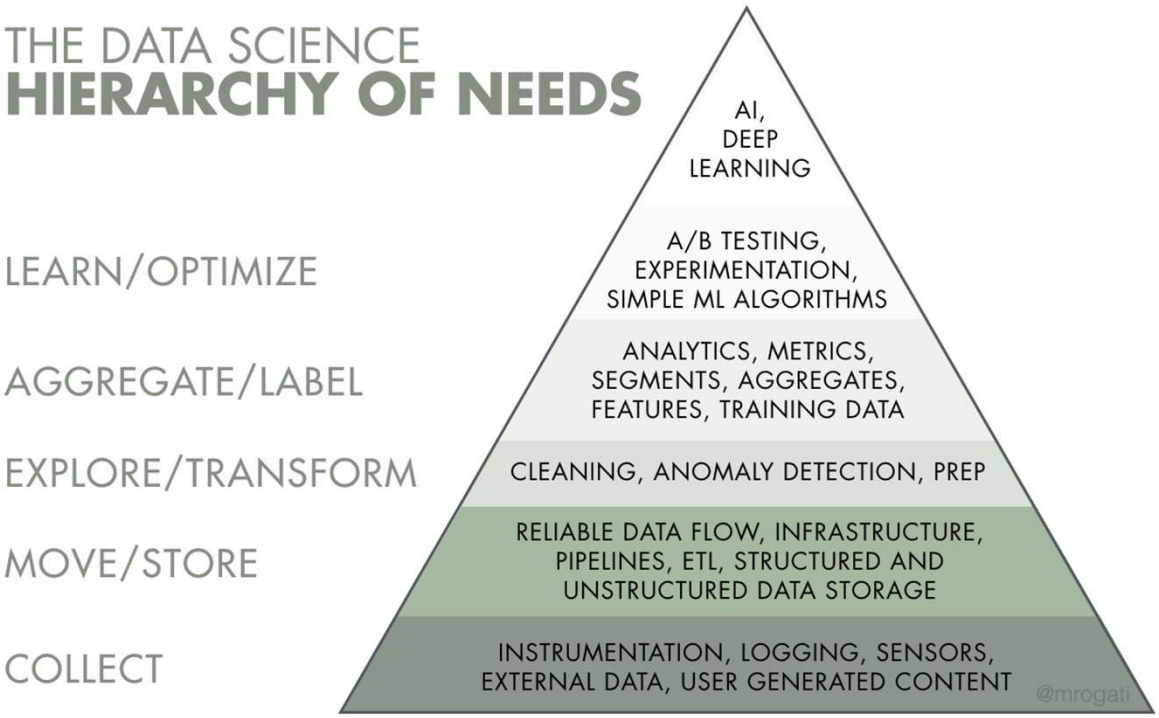


Model vs. Data Centric AI

Model-centric	Data-centric
Coletar dados e desenvolver um modelo bom o suficiente para lidar com o ruído no dado.	Foco na consistência do dado. Utilizar ferramentas para melhorar a qualidade do dado, permitindo que diversos modelos desempenhem bem.
Manter o dado fixo e iterativamente melhorar o modelo.	Manter o modelo fixo e iterativamente melhorar o dado.

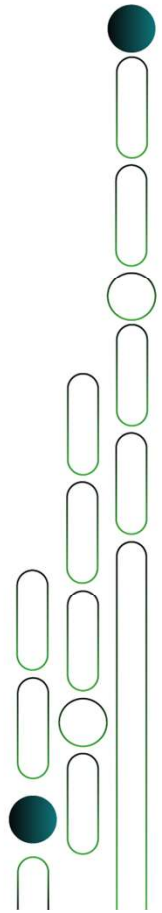


Model vs. Data Centric AI



Referências para seguir aprendendo

- Google's Rules of Machine Learning



Contatos

- E-mail: joao.faria.ext@igti.edu.br
- LinkedIn: https://www.linkedin.com/in/actsoft