INSTITUTO FEDERAL
Santa Catarina

Introdução a Ciência de Dados

PROF. TIAGO G. MORAES

Roteiro

- Introdução
- Tarefas típicas
- Definições
- Tipos de aprendizado de máquina
- ML - aplicações

Introdução

- O que é ciência de dados?
  - Aplicação de métodos científicos combinando computação e estatística para extrair conhecimento e *insights* sobre dados estruturados ou não
  - Ciclo (*pipeline* dos dados) que transforma dados brutos em conhecimento útil
    - Coleta e limpeza dos dados, análise, criação de um modelo e análise dos resultados
  - Disciplina que usa diversas ferramentas (matemática, estatística e aprendizado de máquina) para extrair conhecimento e insights dos dados
- Conceito se mistura e confunde com outros:

Introdução

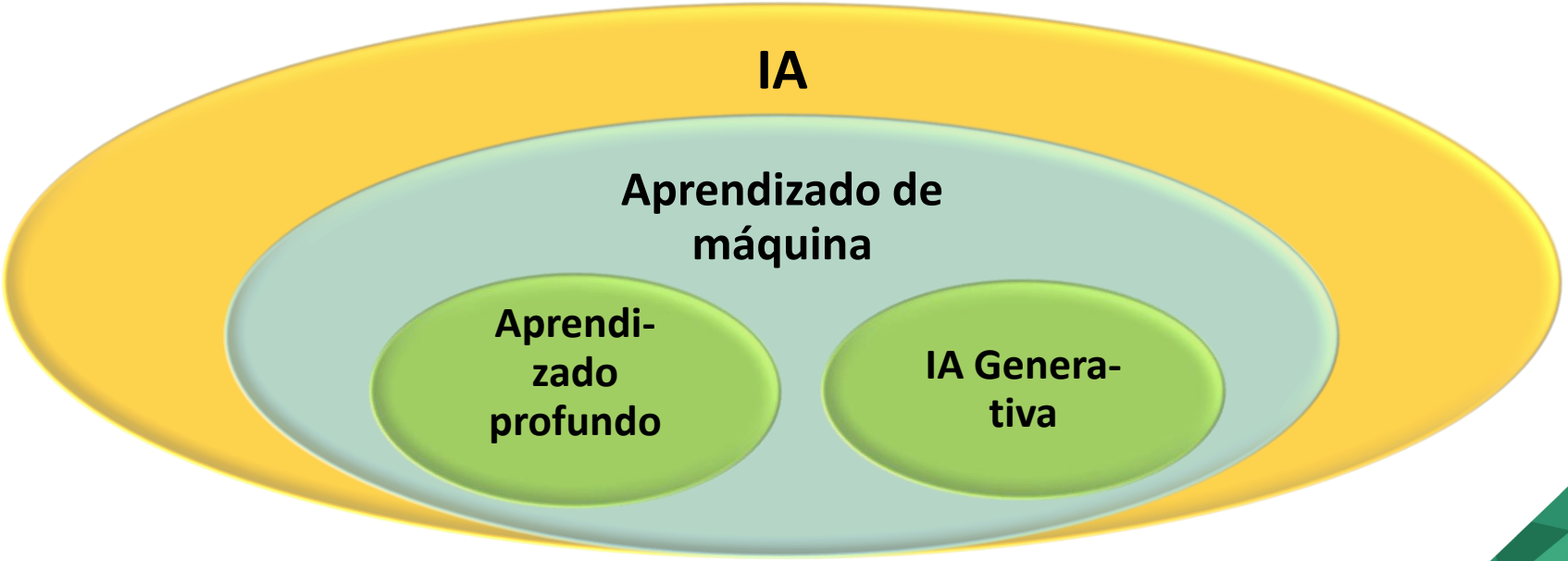
Obtenção de **conhecimento** e **insights** a partir de dados

É um conceito fortemente atrelado a **Aprendizado de Máquina**, que é uma parte da área de **Inteligência Artificial**

- Conceito se mistura e confunde com outros:
  - Inteligência Artificial
  - Aprendizado de Máquina
  - Engenharia de Dados

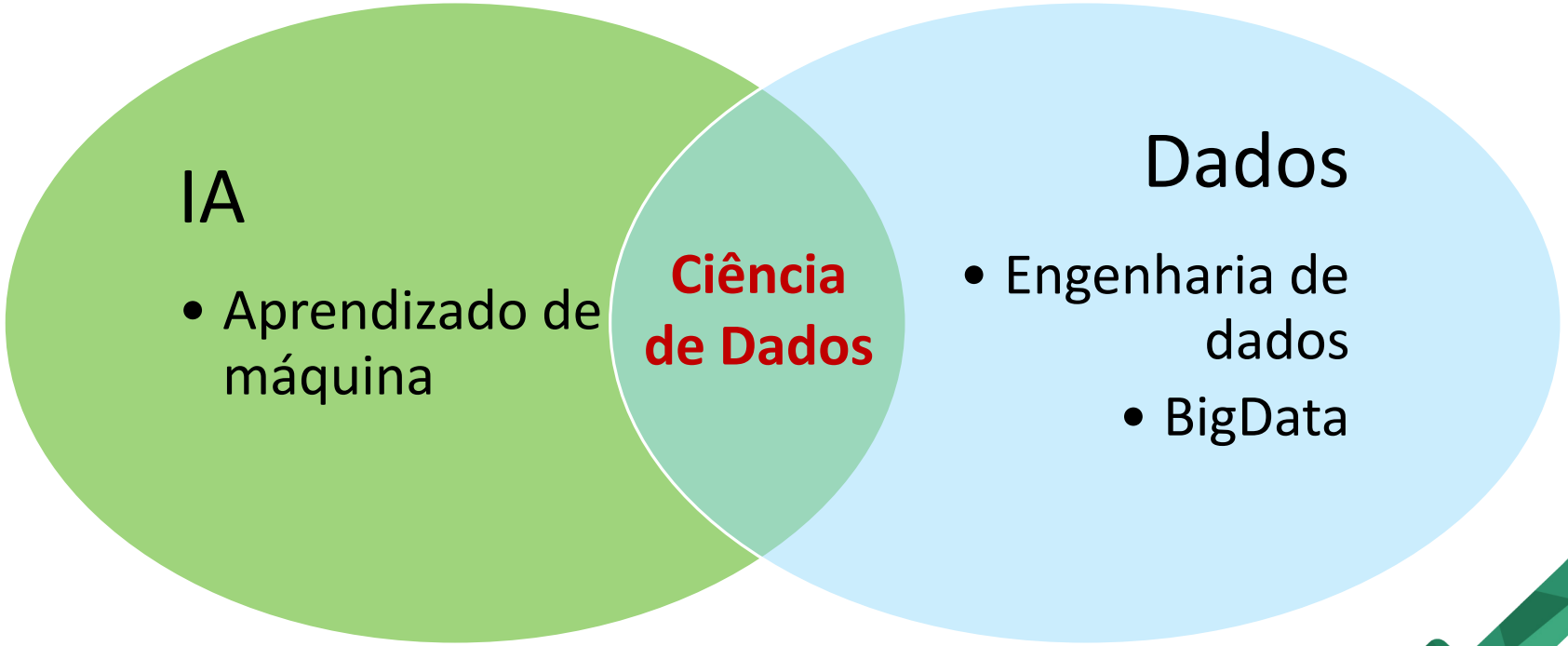
Introdução

- Diferentes conceitos



Introdução

- Diferentes conteitos



Introdução

- Diferentes carreiras:
  - Engenheiro de IA Generativa: desenvolve sistemas com IAs generativas
    - LLM (Large Language Model): ChatGPT, Gemini...
  - Engenheiro de Machine Learning (aprendizado de máquina)
    - Cria modelos para previsões
  - Engenheiro de dados
    - Captura e limpa dados
  - Ciêntista de dados
    - Análise e insigths

Introdução

- Por que estudar essa área
  - 1 EB equivale a um quintilhão de bytes

Source: IDC's Digital Universe Study, sponsored by EMC, December 2012

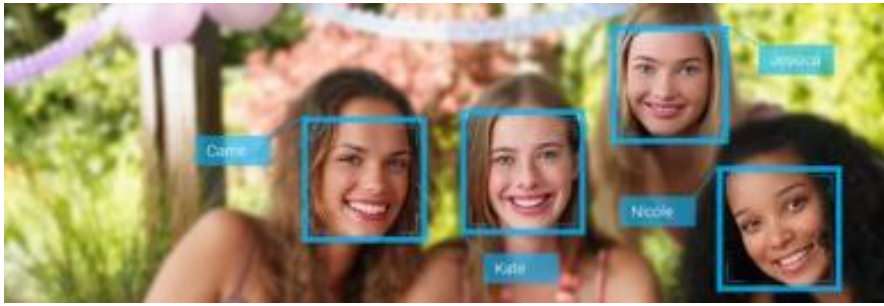
Introdução

- Por que estudar essa área
  - 1 EB: um quintilhão de bytes
  - 1 ZB: um sextilhão de bytes

- Fonte: pesquisas de mercado da IDC (International Data Corporation)

Introdução

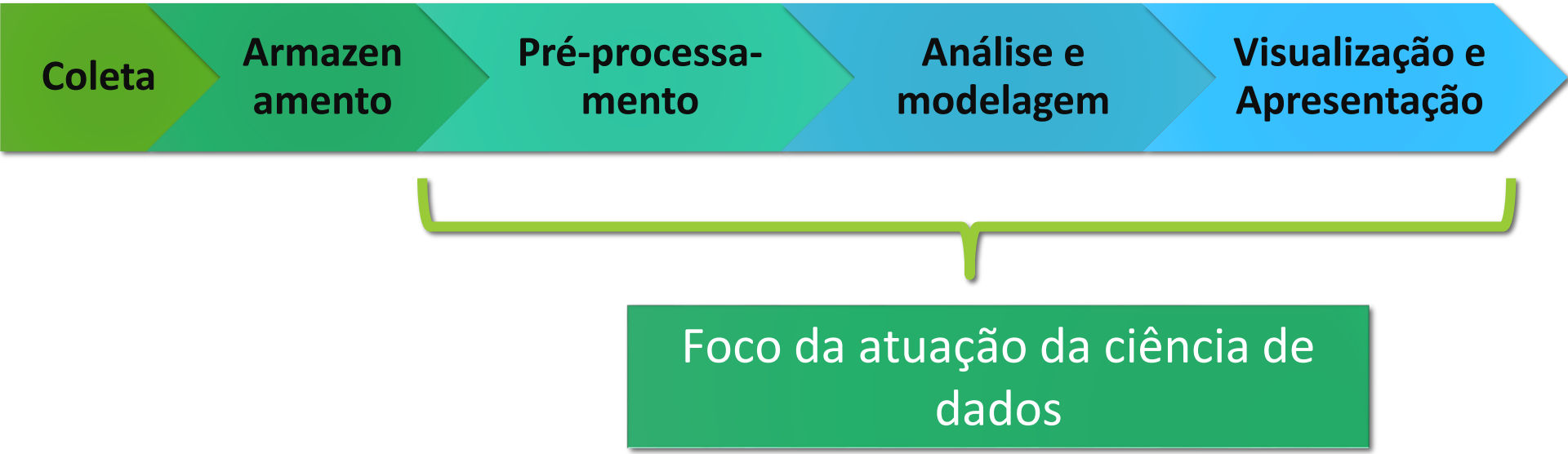
- Por que estudar esses conceitos?



- Aplicação nas mais diversas áreas
  - Reconhecimento facial
  - Reconhecimento de fraudes
  - Segmentação de mercado
  - Recomendação de conteúdo (spotify, netflix etc)
  - etc
- Utilizados por: Google, Meta, Bancos, Fintechs, Startups...
- Curiosidades:
  - Inteligência artificial detecta mal de Alzheimer uma década antes de sintomas
  - Inteligência artificial está escrevendo o fim de Game of Thrones

Tarefas Típicas

- Pipelide dados:



Tarefas Típicas

- Coleta e Ingestão de Dados:
  - Fonte:
    - API's, bancos de dados, sensores IoT, arquivos (texto, imagem e vídeo)
  - Importante verificar a fonte e confiabilidade dos dados
  - **Para ciência de dados é importante também saber se os dados apresentam algum viés ou limitação**

Tarefas Típicas

- Armazenamento e Gerenciamento:
  - Após a coleta os dados são armazenados e local apropriado:
    - Data Lake (dados brutos grandes – sem estrutura)
    - Data Warehouse
    - BD's relacionais
  - **Para ciência de dados é importante a forma que os dados são gerenciados e armazenados pois impacta na velocidade de validar modelos**

Tarefas Típicas

- Pré-processamento:
  - Aqui os dados são preparados para o Aprendizado de Máquina
  - Limpeza: valores inconsistentes, outliers ou faltantes
  - Normalização de dados
  - Redução de dimensionalidade
  - Obteção de novos atributos (exemplo o dia da semana de uma data)
  - **Para ciência de dados é importante o correto processamento inicial pois impacta fortemente na qualidade do modelo criado**

Tarefas Típicas

- Análise e Modelagem:
  - Utilização de algoritmo e modelo de aprendizado de máquina
  - Deve-se escolhe o(s) algoritmo(s) certo para a tarefa
  - Comparar modelos
  - **Etapa central para a ciência de dados**

Tarefas Típicas

- Verificação e Apresentação:
  - Comunicação e apresentação dos resultados para público não técnico
  - Os *insights* devem gerar valor de maneira clara para a organização
  - Nesse momendo se vende a ideia: “story telling”
  - Utiliza gráficos e *dashboards*
  - **Vital para o ciência de dados pois é a etapa que sua importância é demonstrada**

Definições

- Característica ou Atributo
  - Dado extraído de uma amostra por meio de medida e/ou processamento. Em geral são organizadas na forma de um vetor de características.
- Classe
  - Conjunto de padrões que possuem características em comum.
  - Padrão: é uma entidade, objeto, processo ou evento, vagamente definido, que pode assumir um nome.
  - A característica que se quer prever. Classe de interesse

Definições

- Classificação
  - atribuir classes para as amostras, baseado em suas características.
- Ruído
  - distorção, falha ou imprecisão que ocorre na aquisição dos dados.
- Classificadores
  - Usados para classificar ou descrever padrões ou objetos a partir de um conjunto de propriedades ou características.

Definições

- Exemplo:

Ruído

ID	Salário anual	Idade	Valor solicitado	Nota Bom Pagador	Aprovação
1	-	25	R$20000,00	A	Não
2	R$82002,22	160	R$100000,00	B	Não
3	R$115502,25	45	R$50000,00	A	Sim
4	R$42302,25	32	R$20000,00	A+	Sim

Atributo

Classe de interesse

Tipos de Aprendizado de Máquina

- Aprendizado de máquina (ML – Machine learning)
  - Aprender baseado nos dados
  - Ao invés de se programar regras explícitas (algoritmos tradicionais) se busca padrões a partir de um conjunto de dados para tomar decisões
- Existem três tipos de ML:
  - Aprendizado Supervisionado
  - Aprendizado não supervisionado
  - Aprendizado por Reforço

Tipos de Aprendizado de Máquina

- Aprendizado Supervisionado
  - Selecionam-se amostras representativas para cada uma das classes que se deseja classificar.
  - Conhecemos o padrão e classes que estamos procurando.
  - Também conhecido como Classificação supervisionado.
- Exemplos: detecção de SPAM a partir da avaliação prévia de conjunto de emails em SPAM e NÃO SPAM

Tipos de Aprendizado de Máquina

- Aprendizado não supervisionado
  - Não conhecemos o padrão, nem o número total de classes a serem encontradas 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
- Exemplo: segmentação de mercado – agrupar clientes por comportamento semelhante

Tipos de Aprendizado de Máquina

- Aprendizado por Reforço
  - Aprender com iterações: causa e efeito
  - Aprender com a própria experiência
- Exemplo:
  - Robo aspirador de pó mapeando a casa e calculando a rota de limpeza
  - Achar a saída mais rápida de um cenário (mapa) sem conhecê-lo antecipadamente

ML - aplicações

Aplicação	Padrão de Entrada	Classes (saída)
Reconhecimento óptico de caracteres	Imagem de um documento	Caracteres/palavra
Busca na internet	Documento texto/imagem	Categoria semântica
Filtro de e-mails	Email	Spam/normal
Identificação de pessoas	Face, íris, impressão digital	Acesso de usuários credenciados
Diagnóstico auxiliado por computador	Imagem microscópica	Células saudáveis/doentes
Reconhecimento de alvos militares	Imagem óptica ou infravermelho	Tipo do alvo
Seleção automática de qualidade	Imagem em esteira de produção	Níveis de qualidade
Análise de sequências de DNA	Sequência de DNA	Gene conhecido/desconhecido