



DESENVOLVIMENTO DE SOFTWARE MULTIPLATAFORMA

Disciplina: ISW-039 - Mineração de Dados

Aula 01: Introdução à Ciência de Dados.

Data 10/08/2024

Prof. Me. Anderson Silva Vanin

Quem sou eu

- Técnico em Eletrônica
- Bacharel em Ciência da Computação
- Pós Graduado em Banco de Dados
- Mestre em Gestão do Conhecimento e Informática (Aplicado a Visão Computacional)
- Atuação no CPS desde 2006
- Aulas nas disciplinas em diversas de Programação
- Cursos extracurriculares em Inteligência Artificial e IoT

Ementa da Disciplina

- Conceitos Básicos;
- Descoberta de Conhecimento em Banco de Dados (KDD);
- Pré-processamento de dados: Extract, Transform and Load (ETL), limpeza, transformação, redução de dimensionalidade; Raspagem de dados;
- Técnicas de amostragem;
- Balanceamento de classes (undersampling e oversampling);
- Técnicas de visualização de dados;
- Análise descritiva de dados;
- Análises de redes sociais;
- Business Intelligence.

Avaliações e Trabalhos

- **P1:** 28/09/2024 (35%)

Avaliação Teórica/Prática em Laboratório de Informática

- **P2:** 23/11/2024 (35%)

Avaliação Teórica/Prática em Laboratório de Informática

- **P3:** 07/12/2024

Avaliação Prática em Laboratório de Informática

- **T:** 23/11/2024 - Trabalhos e atividades (30%)

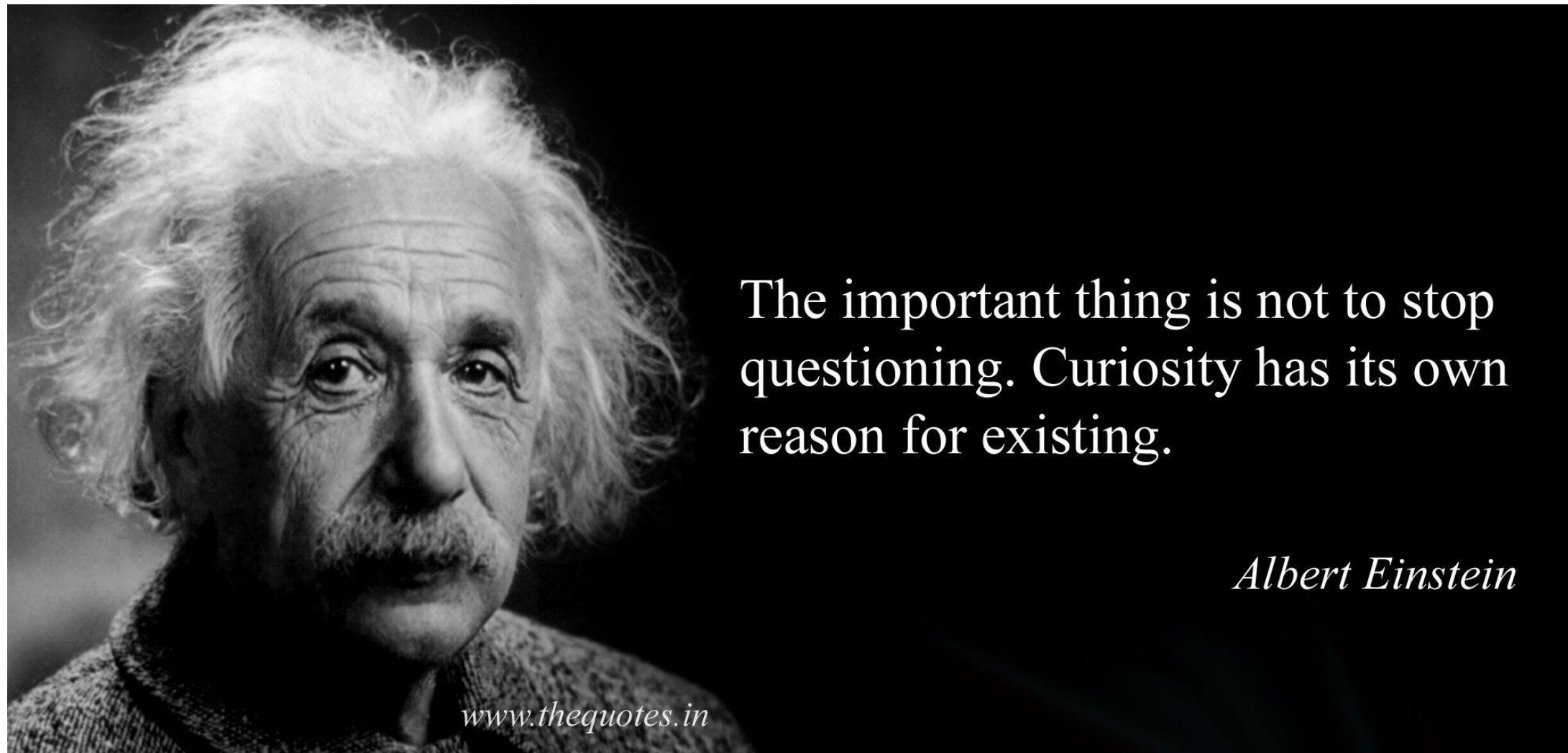
Conjunto de Atividades solicitadas durante o semestre letivo somadas

Material e recursos para as aulas

- **GITHUB:**
https://github.com/ProfAndersonVanin/FATEC_MINERACAO_DADOS_2SEM2024
- **EMAIL:** anderson.vanin@fatec.sp.gov.br

Introdução à Ciência de Dados

Ciência de Dados



Ciência de Dados

A expressão “Data Science” tem origem nos anos 1960, mas a “Ciência de Dados” que lidamos atualmente é algo novo, ainda em transformação e muitas vezes controversa.

Ciência

- A palavra “**ciência**” vem do latim “***scientia***”, que significa “conhecimento”.
- Ciência refere-se a **qualquer conhecimento ou prática sistemáticos**.
- Em sentido estrito, ciência refere-se ao **sistema de adquirir conhecimento baseado no método científico** bem como ao corpo organizado de conhecimento conseguido através de tais pesquisas.

Ciência

- A ciência é aquele tipo de conhecimento que **busca compreender verdades ou leis naturais** para explicar o funcionamento das coisas e do universo em geral.
- É por isso que cientistas fazem observações, verificações, medições, análises e classificações, procurando **entender os fatos e traduzi-los para uma linguagem estatística**.
- E é aí que entra o método científico.

Método Científico

- O método científico é, basicamente, um conjunto de regras para se realizar uma experiência, com o objetivo de produzir um novo conhecimento, além de corrigir conhecimentos preexistentes.
- Essas regras são necessárias justamente para coibir a subjetividade, direcionando a pesquisa para a produção de conhecimentos válidos – em suma, científicos.



Sobre Ciência de Dados

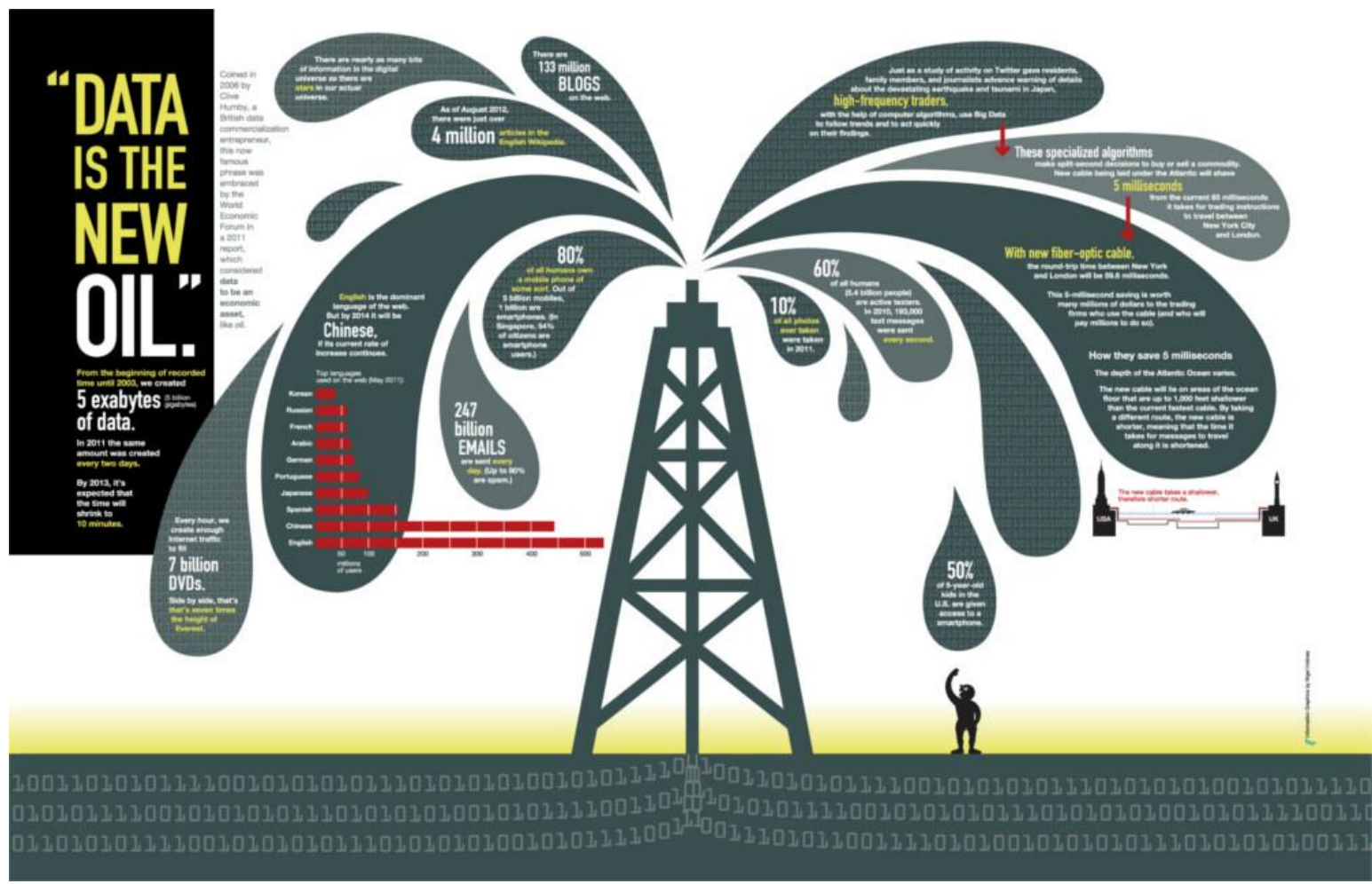
- **Habilidades de programação** e de uso de ferramentas tecnológicas.
- Muito **conhecimento de estatística e matemática** (análise de modelos, compreensão de técnicas de visualização, etc...).
- Competência significativa em uma **área do conhecimento**, com capacidade de analisar informações e resolver problemas.

Por que estudar Data Science (DS)

- Aumento da **geração e produção de dados**.
- **Big Data**.
- Custo mais baixo para **armazenamento de dados**.
- Aumento da **capacidade de processamento**.
- Evolução das tecnologias.
- Capacidade de gerar conhecimento e **vantagens em tomadas de decisão**.

Por que estudar Data Science (DS)

- Clive Humby – Data is the new Oil



Vida Real (Muitos exemplos)

Futebol: LIVERPOOL (**Campeão de tudo em 2019!**).

- Ian Graham (doutor em física teórica) que trabalha para o Liverpool, elaborou o conceito de “domínio de campo”, que agrega diversas estatísticas coletadas no monitoramento de partidas.
- (<https://www.startse.com/artigos/ciencia-de-dados-liverpool/>)

Dado – Informação - Conhecimento

Dado é o menor e mais simples elemento de um sistema.

É uma unidade indivisível, extremamente objetiva, geralmente abundante e que tem o papel de registrar um fato (evento).

Considerado um elemento de fácil manipulação e transporte.

Basicamente **informação** é um conjunto de dados dentro de um contexto.

Conhecimento: Uma informação que, devidamente tratada, muda o comportamento do sistema.

Dados Estruturados

- Aqueles que possuem formato e comprimento definido, como por exemplo, números, datas e grupos de palavras.
- Consistem de um conjunto de dados definidos a partir de um esquema formalmente definido.
- Exemplo: dados armazenados em banco de dados relacionais; dados num estrutura XML regida por um documento XSD; dados de planilhas com clareza estrutural; dados oriundos de sensores e equipamentos, desde que com uma estrutura de metadados bem definida.


Dados Não Estruturados

- Consistem em conjuntos de dados que não têm uma estrutura definida.
- Em razão de não haver uma estrutura formal, a extração de informações nesses conjuntos de dados torna-se complexa do ponto de vista computacional.
- Exemplos de dados são: áudios, vídeos, documentos em formato texto, imagens, dados de mídias sociais, entre outros.

Dados Semi-Estruturados

- Consistem em dados com uma estrutura implícita e flexível, geralmente um meio termo entre a estruturação e a falta total de estruturação.
- Mesmo a estrutura não sendo rígida, a existência de uma mínima estrutura implícita facilita a gestão dos dados.
- Exemplos são arquivos tabulares em planilhas (formatos TSV e CSV), arquivos XML, conteúdos web acompanhados de tags;

BI X DS




Diferenças entre Business Intelligence e Data Science

Business Intelligence

O que aconteceu?


Análise descritiva
Preparação de relatórios



Data Science

Por quê? O que acontecerá?
O que deveria fazer?

Análise preditiva
Análise prescritiva



CARATERÍSTICAS	Foco	Relatórios, KPIs (Indicadores-chave de desempenho), tendências	Padrões, correlações, modelos
	Processo	Estático, comparativo	Exploratório, experimental, visual
	Fontes de dados	Planejadas, adicionadas progressivamente	Em andamento, de acordo com as necessidades
	Transformação	Com antecedência, cuidadosamente planejada	No banco de dados, sob demanda, enriquecimento
	Qualidade dos dados	Única versão da verdade (Single version of the truth - SVOT)	Suficientemente boa (Good Enough)
	Análise	Retrospectiva, descritiva	Preditiva, prescritiva, preventiva

Fonte: Data Science Central (TechTarget).

Jupyter Notebook e o Google Colab

Python (python.org)

- Nesta disciplina usaremos muito a linguagem Python.
- Python é uma linguagem de programação de alto nível, interpretada de script, com tipagem dinâmica e forte. Além disso é multiplataforma.

Jupyter Notebook (jupyter.org)

- Aplicativo open source para desenvolvimento.
- Ambiente de trabalho Browser
- Excelente alternativa para desenvolvimento e prática de programação em python em praticamente todas as fases de atividade do profissional de DS.

Demonstração Jupyter Notebook – MÁQUINA LOCAL

Google Colab

- A Google disponibiliza uma plataforma para desenvolvimento Python, muito similar ao Jupyter Notebook que o desenvolvedor instala em sua máquina local.
- Mas lembre-se, para usar o Colab é necessário ter internet disponível.
- **Algumas vantagens:**
 - Não é necessário ficar gerenciando bibliotecas.
 - Acesso a GPUs e TPUs gratuitas.
 - Acesso de qualquer lugar.
 - Compartilhamento fácil.

Demonstração Google Colab