

# CURSO SUPERIOR DE ADS

## Introdução à Ciência de Dados e Big Data



Prof. Fernando Marlon Soares Figueiredo  
Disciplina: Ciência de Dados e Big data



# Apresentação da Disciplina

---

## 1. Objetivo Geral:

- Desenvolver competências para a análise de grandes volumes de dados, utilizando técnicas e ferramentas de Big Data e Ciência de Dados para extrair informações relevantes e apoiar a tomada de decisão.

## 2. Específicos:

- Compreender os fundamentos de Ciência de Dados e Big Data.
- Aplicar técnicas de coleta, armazenamento e processamento de dados.
- Utilizar ferramentas de visualização de dados para análise exploratória.
- Implementar algoritmos de aprendizado de máquina em projetos de Big Data.
- Analisar e interpretar resultados obtidos a partir de grandes volumes de dados

# Lembrando nossa ementa

---

**Ementa:** Estudo prático dos fundamentos de Ciência de Dados e Big Data. Coleta, limpeza, transformação e análise de dados. Mineração de dados, aprendizado de máquina e visualização. Ferramentas práticas como Python (Pandas, Scikit-learn), Power BI, Tableau, Hadoop e Spark.

# Bibliografia

---

## Básica:

- BASSO, D. G. Big Data. 1ª Ed. Curitiba: Intersaberes, 2020. [acervo digital]
- ELMASRI, Ramez; NAVATHE, Shamkant B. Sistemas de banco de dados. 7ª Ed. São Paulo: Pearson, 2018. [acervo digital]
- VALDATI, A. B. Inteligência Artificial – IA. 1ª Ed. Curitiba: Intersaberes, 2020. [acervo digital]

## Complementar:

- ALVES, W. P. Banco de dados: Teoria e desenvolvimento. 2ª Ed. São Paulo: Érica, 2020. [acervo físico]
- ARAÚJO, R. C. A. Urban data analytics, urban big data e IOT. Curitiba: Intersaberes, 2020. [acervo digital]
- FORTA, Ben. SQL em 10 minutos por dia. 1ª Ed. São Paulo: Novatec, 2021. [acervo físico]
- NIELD, Thomas. Introdução à Linguagem SQL: Abordagem prática para iniciantes. 1ª Ed. São Paulo: Novatec, 2016. [acervo físico]
- REGO, B. L. Simplificando a governança de dados: Governe os dados de forma objetiva e inovadora. 1ª Ed. Rio de Janeiro: Brasport, 2020. [acervo digital]

# Avaliações

---

**1ª Unidade:** Prova - 100% **Data:** 11/03

**2ª Unidade:** Prova (60%) + Projetos (40%) **Data:** 15/04

**3ª Unidade:** Seminários (80%) + Prática Integradora (20%) **Data:** 27/05

**Media Final:**  $((1U + 2U + 3U) / 3) \Rightarrow 7 \rightarrow$  **Aprovado(a)**

**Prova Final (Data)** -> 10/06 ☹️



O QUE É  
DATA  
SCIENCE?



# O que é Ciência de Dados?

- Área interdisciplinar focada em extrair conhecimento, padrões e insights de dados para apoiar a tomada de decisões.
- A expressão Data Science tem origem na década de 60
- A Ciência de Dados estuda o dado desde a produção até o descarte!
- Ciência de dados que lidamos atualmente é nova, ainda em transformação e muitas vezes controversa
- **Não é apenas programação:** envolve método científico e pensamento crítico.



# Método Científico

---

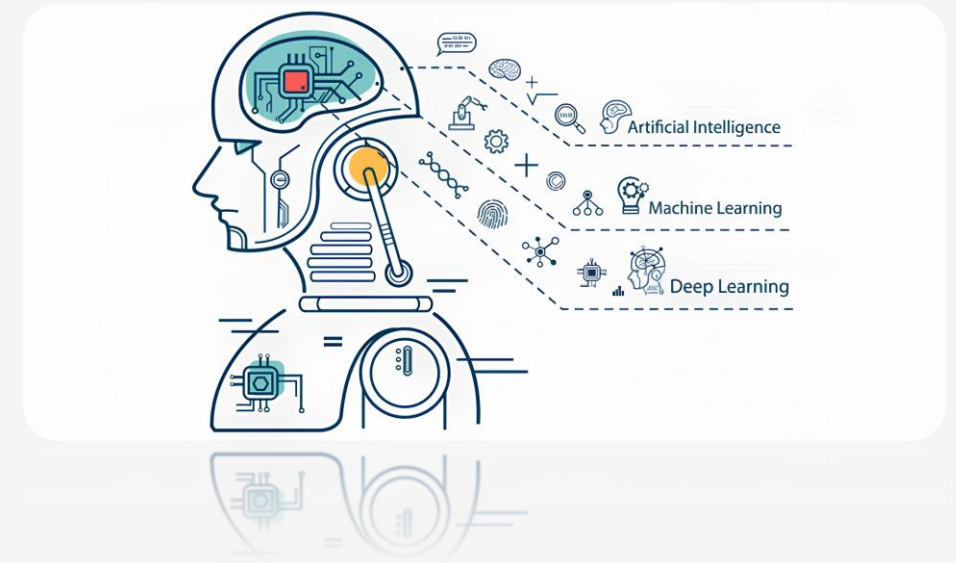
- O método científico é, basicamente, um conjunto de regras para se realizar uma experiência, com o objetivo de produzir um novo conhecimento, além de corrigir conhecimentos preexistentes.
- Essas regras são necessárias justamente para coibir a subjetividade, direcionando a pesquisa para a produção de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





# Por que estudar DS agora?

- Porque aumentou demais a geração e produção de dados (internet, dispositivos móveis, sensores....)
- **Big Data** é uma realidade;
- Custo baixo (mas ainda alto) para armazenar dados;
- Aumento da capacidade de processamento;
- Evolução das tecnologias;
- Pela certeza de que a análise de dados é capaz de gerar conhecimento e vantagens em tomadas de decisões nas mais variadas áreas.



# "DATA IS THE NEW OIL"

From the beginning of recorded time until 2003, we created **5 exabytes** of data.  
(5 billion gigabytes)

In 2011 the same amount was created **every two days**.

By 2013, it's expected that the time will shrink to **10 minutes**.

Every hour, we create enough Internet traffic to fill **7 billion DVDs**.

Side by side, that's **thirteen times** the height of Everest.

Collected in 2008 by Glaxo Humber, a British data communication entrepreneur, this new business was announced by the World Economic Forum in a 2011 report, which considered data to be an economic asset, like oil.

There are nearly as many bits of information in the digital universe as there are **atoms** in our entire universe.

As of August 2012, there were just over **4 million** articles in the English Wikipedia.

There are **133 million BLOGS** on the web.

**80%** of all humans own a mobile phone of some sort. Out of 5 billion mobiles, 1 billion are smartphones. (In Singapore, 84% of citizens are smartphone users.)

**247 billion EMAILS** are sent **every day**. (Up to 60% are spam.)

English is the dominant language of the web. But by 2014 it will be **Chinese**, if its current rate of increase continues.

Top languages used on the web (May 2011):



Just as a study of activity on Twitter gave residents, family members, and journalists advance warning of details about the devastating earthquake and tsunami in Japan, **high-frequency traders**, with the help of computer algorithms, use Big Data to follow trends and to act quickly on their findings.

These specialized algorithms make split-second decisions to buy or sell a commodity. New cable being laid under the Atlantic will shave **5 milliseconds** from the current 66 milliseconds it takes for trading instructions to travel between New York City and London.

With new fiber-optic cable, the round-trip time between New York and London will be 59.6 milliseconds.

This 6-millisecond saving is worth many millions of dollars to the trading firms who use the cable (and who will pay millions to do so).

**How they save 6 milliseconds**

The depth of the Atlantic Ocean varies.

The new cable will be an arc of the ocean floor that are up to 1,000 feet shallower than the current fastest cable. By taking a different route, the new cable is shorter, meaning that the time it takes for messages to travel along it is shortened.



The new cable takes a shallower, shorter route.

**60%** of all humans (5.4 billion people) are active traders. In 2010, 100,000 text messages were sent every second.

**10%** of all photos ever taken were taken in 2011.

**50%** of 8-year-old kids in the U.S. are given access to a smartphone.

# Vida Real (MUITOS EXEMPLOS)

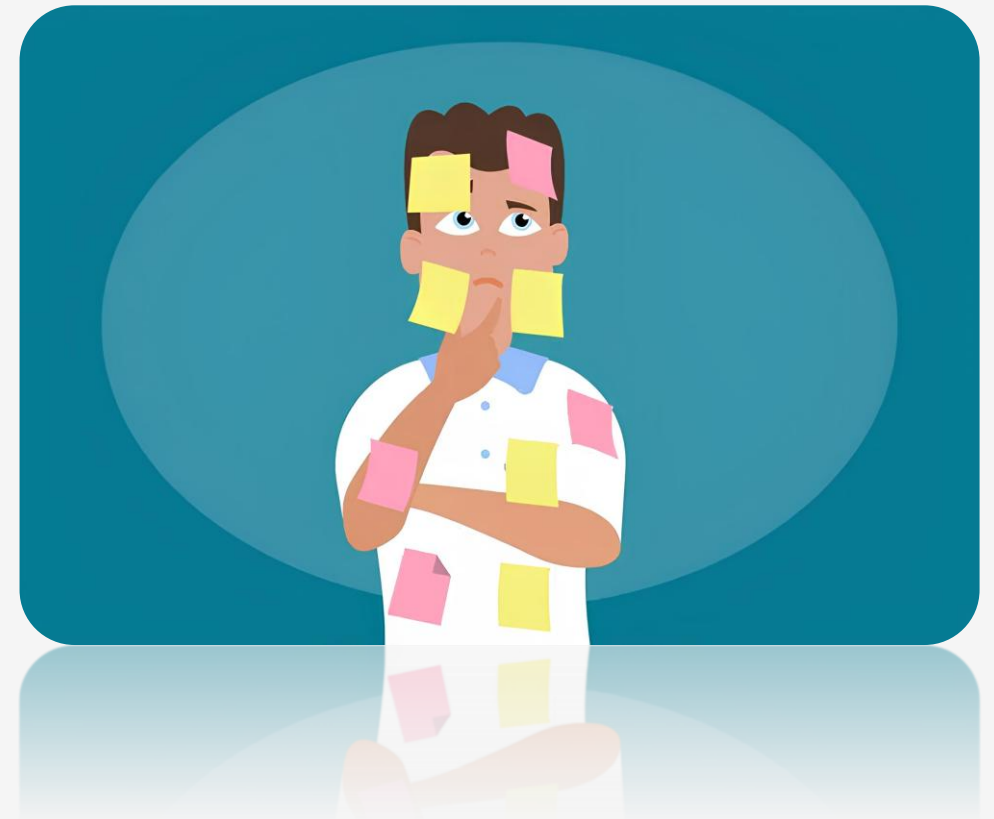
- Futebol: LIVERPOOL (Campeão de tudo em 2019)
  - **Ian Graham** (doutor em física teórica) que trabalha para o Liverpool elaborou o conceito de "domínio de campo", que agrega diversas estatísticas coletadas no monitoramento de partidas.
  - Link para o Artigo: <https://www.startse.com/artigos/ciencia-de-dados-liverpool/>
- Fleury cria Saúde iD, um 'one-stop shop' da saúde.
  - <https://braziljournal.com/fleury-cria-saude-id-um-one-stop-shop-da-saude/#:~:text=O%20Fleury%20anunciou%20a%20cria%C3%A7%C3%A3o,core%20business'%20de%20exames%20diagn%C3%B3stico>



## Portanto...

---

- DS é sobre você descobrir conhecimento a partir de dados.
- DS é uma ciência, lembre-se sempre disso em seus projetos!
- Dados são o novo petróleo.
- Nem todo excelente programador (inclusive em **Python** ou **R**) é um bom analista de dados.
- Essa teoria toda é tão importante? **Eu quero programar!!**
- Estudar Ciência de Dados tem muito a ver com sua capacidade de abstração sobre um conjunto de dados, é preciso ampliar o olhar, abrir a mente!



# Big Data e os 4Vs

---

- **Volume:** quantidade massiva de dados gerados.
- **Velocidade:** rapidez com que dados são criados e processados.
- **Variedade:** diferentes formatos (texto, imagem, vídeo, etc.).
- **Valor:** relevância e utilidade do dado para gerar insights.

**Desafios:** armazenamento, processamento, qualidade, governança





# Ciclo de vida dos dados



Coleta → Armazenamento → Pré-processamento → Análise/Modelagem → Visualização → Decisão

# PYTHON (python.org)

- Nesta disciplina usaremos massivamente a linguagem Python.
- Python é uma linguagem de programação de alto nível, interpretada de script, com tipagem dinâmica e forte. Além disso, é multiplataforma.
- Há muita conversa e às vezes competição em definir a melhor linguagem para DS (Python, linguagem R, ou outras menos utilizadas).
- Saiba:
  - Tanto o **Python** como o **R** são ótimos, mas é preciso se especializar em uma delas.
  - Penso que o grande diferencial seja a quantidade de oferta de **empregos** e também a amplitude de possibilidades que o **Python** oferece em relação ao **R**.



# LINGUAGEM R (r-project.org)

---

- Nesta disciplina mencionaremos a linguagem **R**, embora nosso foco seja **Python**.
- **R** é uma linguagem de programação e ambiente de software livre para análise estatística e gráficos.
- **R** é especialmente forte em estatística e visualização de dados, sendo muito utilizada em pesquisas acadêmicas e por estatísticos.





# Aplicações em Ciência de Dados (Python x R)

---



- Utilizado para análise de dados, machine learning, visualização de dados.
- Bibliotecas populares: **Pandas, NumPy, Matplotlib, SciPy, Scikit-Learn, TensorFlow.**
- Favorecido pela simplicidade e extensa comunidade.



- Principalmente usado em estatísticas e visualização de dados.
- Bibliotecas populares: **ggplot2, dplyr, tidyr, Shiny**
- Ferramenta favorita entre acadêmicos e pesquisadores para análises estatísticas avançadas.

# CONTEÚDO PRAGMÁTICO

| Conceito                                         | Descrição                                                                                                   |
|--------------------------------------------------|-------------------------------------------------------------------------------------------------------------|
| Ciência de Dados                                 | Estudo de técnicas e ferramentas para extrair conhecimento e insights a partir de grandes volumes de dados. |
| Descoberta de Conhecimento em Bases de Dados     | Processo de identificar padrões válidos, novos, potencialmente úteis e compreensíveis a partir de dados.    |
| Mineração de Dados                               | Técnica de análise que permite encontrar padrões ocultos em grandes volumes de dados.                       |
| Preparação de Dados e Pré-processamento de Dados | Processo de limpar, transformar e organizar dados brutos antes da análise.                                  |
| Modelagem de Dados                               | Processo de criar representações abstratas dos dados e suas relações.                                       |
| Estudo de Algoritmo Preditivo Simples (k-NN)     | Algoritmo de aprendizado supervisionado usado para classificação e regressão.                               |
| Conceitos de Business Intelligence               | Conjunto de estratégias e tecnologias usadas por empresas para análise de dados e gestão de informações.    |
| Arquitetura de Soluções de BI                    | Estrutura organizacional que define a configuração de sistemas, processos e tecnologias de BI.              |

# CONTEÚDO PRAGMÁTICO

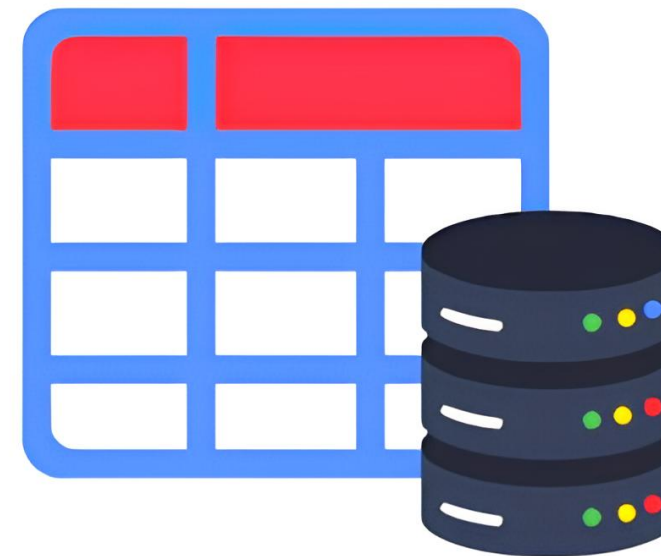
|                                                      |                                                                                                                                                             |
|------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Tecnologias de Business Intelligence                 | Softwares e aplicações usadas para coletar, processar e analisar dados de negócios.                                                                         |
| Metodologia para Desenvolvimento de Aplicações em BI | Conjunto de práticas e etapas para criar soluções de BI eficientes.                                                                                         |
| Ferramentas e Produtos                               | Softwares específicos usados no desenvolvimento e implementação de soluções de BI.                                                                          |
| Técnicas de ETL                                      | Processos de Extração, Transformação e Carregamento de dados.                                                                                               |
| Consultas OLAP                                       | Método de consulta e análise de dados multidimensionais em um data warehouse.                                                                               |
| Mineração de Dados e Conceitos de Big Data           | Processo de analisar grandes conjuntos de dados para descobrir padrões ocultos; conjunto de técnicas e tecnologias para grandes volumes de dados complexos. |

# Tipo de Dados

## Dados Estruturados

Possuem duas dimensões: linhas e colunas. De maneira técnica, são denominados como variáveis e observações e a dimensão de uma tabela é interpretada como linha versus coluna, ou variáveis por observações. Também conhecidos como tabelas ou matrizes, são utilizados em bancos de dados SQL e em data warehouses. Uma função bastante utilizada para a visualização e a quantidade de variáveis por observações é shape, que retorna uma tupla representando a dimensionalidade do DataFrame.

Exemplos: Tabelas de bancos de dados **SQL- Data warehouses**

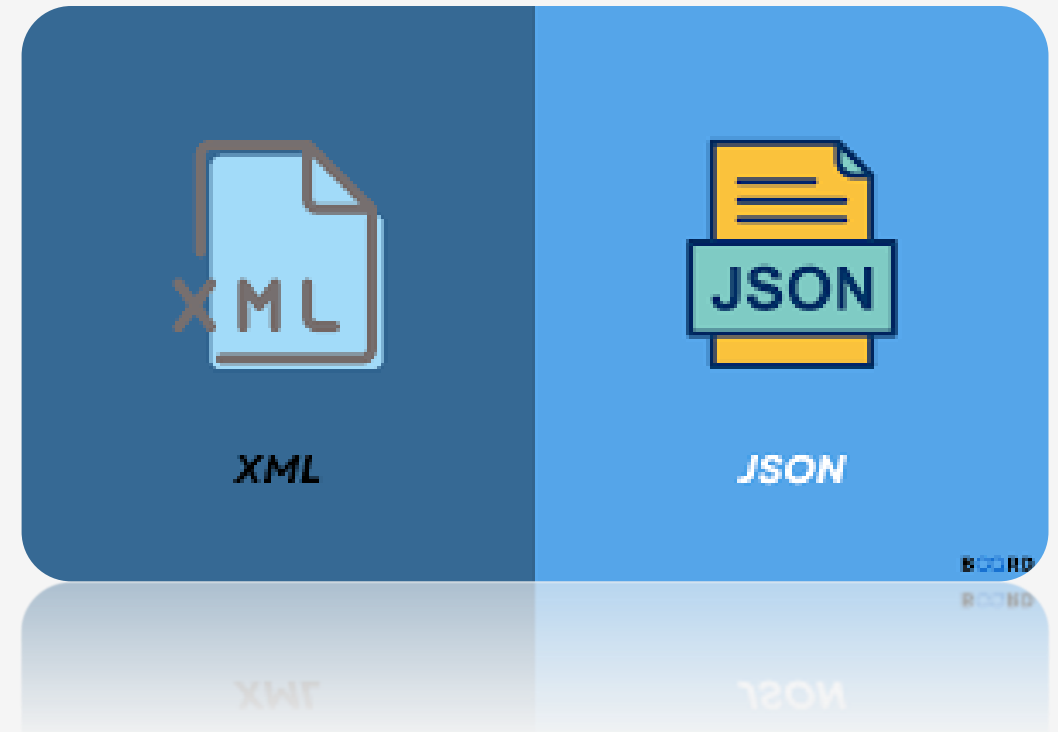


# Tipo de Dados

## Dados Semiestruturados

Apresentam uma estrutura não trivial, extrapolando a definição de dados estruturados. Possuem um esquema de apresentação autodescritivo, mas que exige uma visualização para que a estrutura seja identificada.

Exemplos: Formatos **XML** e **JSON**



# Tipo de Dados

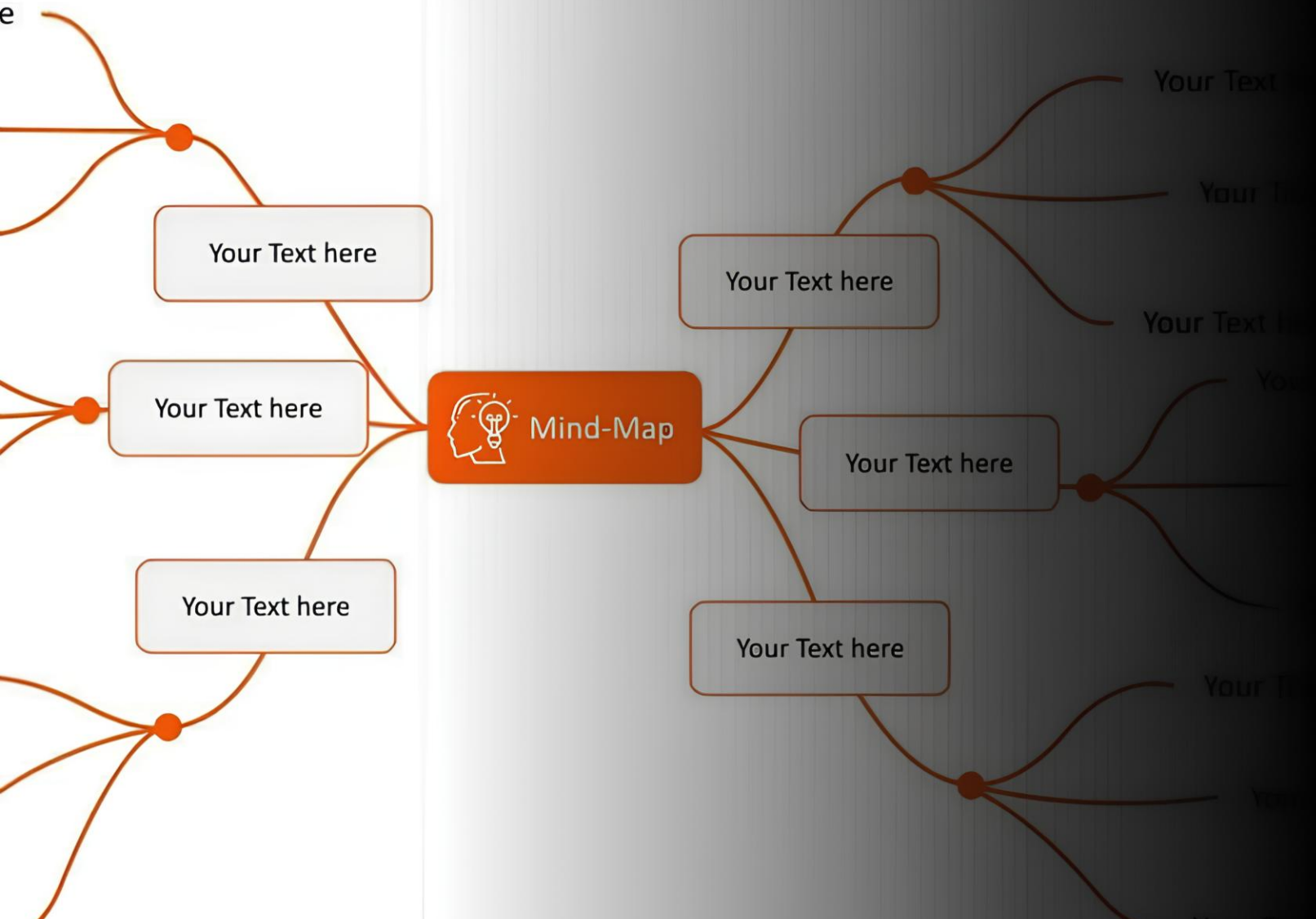
## Dados Não Estruturados

São encontrados de maneira abundante no dia a dia. Este tipo de informação possui estrutura não linear, dificultando a interpretação da informação.

Exemplos: **Imagens**, **Vídeos**, **Músicas** e **Texto**.



# Mind-Map Infographics



---

ATIVIDADE PRÁTICA:  
CRIANDO MIND MAP DA  
DISCIPLINA

---



# Atividade – Google Colab

```
import networkx as nx
import matplotlib.pyplot as plt
```

```
G = nx.Graph()
centro = "Ciência de Dados"
ramos = [
    "Fundamentos (4Vs, KDD)",
    "Coleta & Pré-processamento",
    "Análise & ML (Scikit-learn)",
    "Visualização (Power BI / Tableau)",
    "Big Data (Hadoop / Spark)",
    "Ética & Governança"
]
```

```
[3] G.add_node(centro)
    for r in ramos:
        G.add_edge(centro, r)
```

```
[4] pos = nx.spring_layout(G, seed=42)
    plt.figure(figsize=(8,6))
    nx.draw(G, pos, with_labels=True, node_size=2000, font_size=10)
    plt.title("Mapa mental – Ciência de Dados")
    plt.show()
```

