

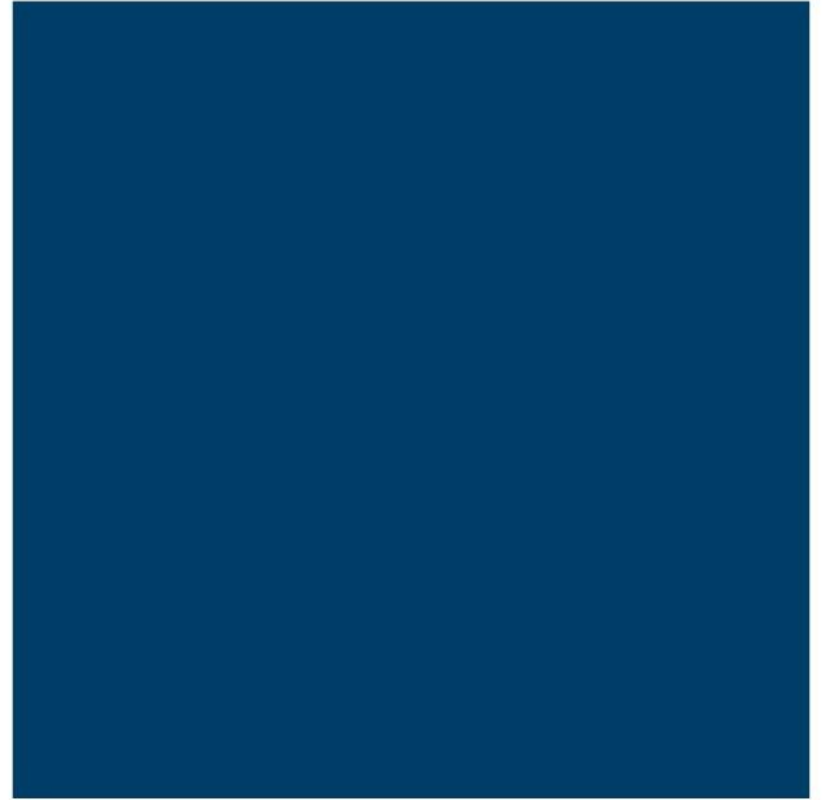
# **Análise, construção e visualização de dados**

Prof. Dr. Álvaro Campos  
Ferreira

[alvaro.ferreira@idp.edu.br](mailto:alvaro.ferreira@idp.edu.br)

# Fluxo de trabalho com dados

# Vamos construir um conjunto de dados



<https://forms.gle/rn1eZGAatVk6LNpk6>

# Dados

## Estruturados

- Tabelas com eixos definidos
- Bancos de dados
- Texto estruturados como:
  - CSV
  - JSON
  - XLS
  - DB
- Dados que possuem uma estrutura subjacente que permite o fácil acesso de uma ou múltiplas entradas.

## Não-estruturados ou Semi-estruturados

- Tweets, comentários, e-mails
- Logs e texto automatizado
- Texto como livros, artigos, wikis
- Páginas da web
- Dados que não possuem uma estrutura subjacente que permite o fácil acesso de cada entrada. Deve ser interpretado de acordo com suas características próprias.

# Dados

## Categorizados

- Escala Nominal
- Escala Ordinal
- Exemplos:
  - Lista de nomes
  - Lista de menções

## Quantitativos

- Escala Intervalar
- Exemplos:
  - Lista de salários
  - Preços diários
  - Número de usuários

# Dados

## Seção transversal

- Todas as observações são realizadas ao mesmo tempo

## Série temporal

- Observações realizadas ao longo do tempo

# Variáveis

# Tipos das variáveis



# Análise exploratória

**Quais perguntas  
responder?**

# Quais perguntas responder?

A partir de agora, a análise vai acontecer através de uma série de perguntas cujas respostas podem ser obtidas através da análise e interpretação dos dados.

Definir bem as perguntas é útil para guiar a análise exploratória.

# Quais perguntas responder?

Algumas perguntas que podemos responder de nosso conjunto de dados sobre nosso conjunto de dados:

- Quantos filmes estão no conjunto?
- Quantos de cada tipo?
- Quantos repetidos?

# Pandas e DataFrames

# Pandas e Dataframes

Vamos usar o Pandas para acessar os nossos dados.

```
import pandas as pd  
file = "filmes.csv"  
df = pd.read_csv(file)  
print(df)
```

# Indexação de DataFrames

O DataFrames é organizado em colunas que podem ser acessadas a partir de seu nome.

Para determinar os nomes das colunas, pode-se escrever na tela os primeiros valores com a função `head()`

```
df.head()
```



INSTITUTO BRASILEIRO DE ENSINO,  
DESENVOLVIMENTO E PESQUISA