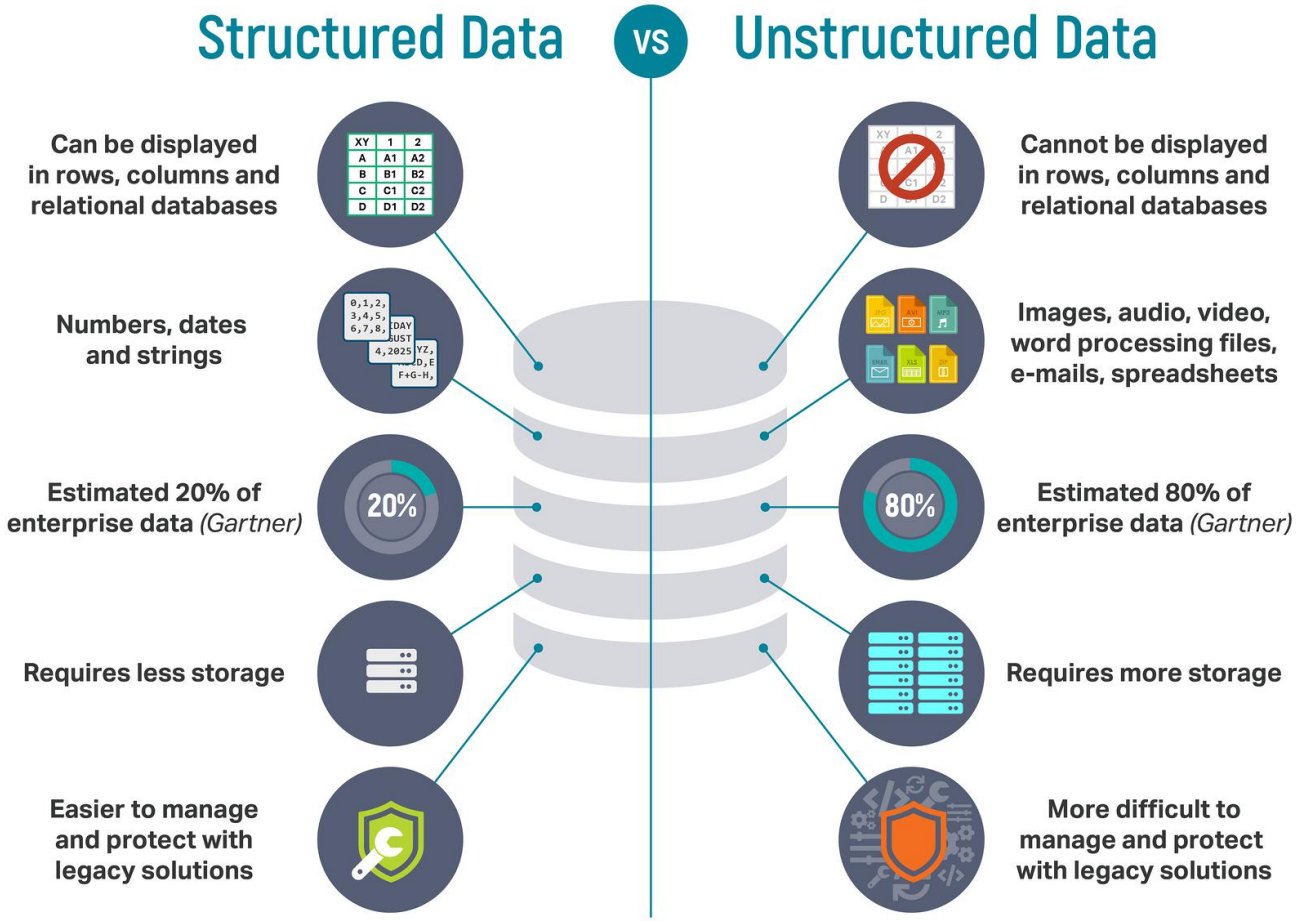
Pandas Essencial

Conceitos Básicos sobre Manipulação de Dados

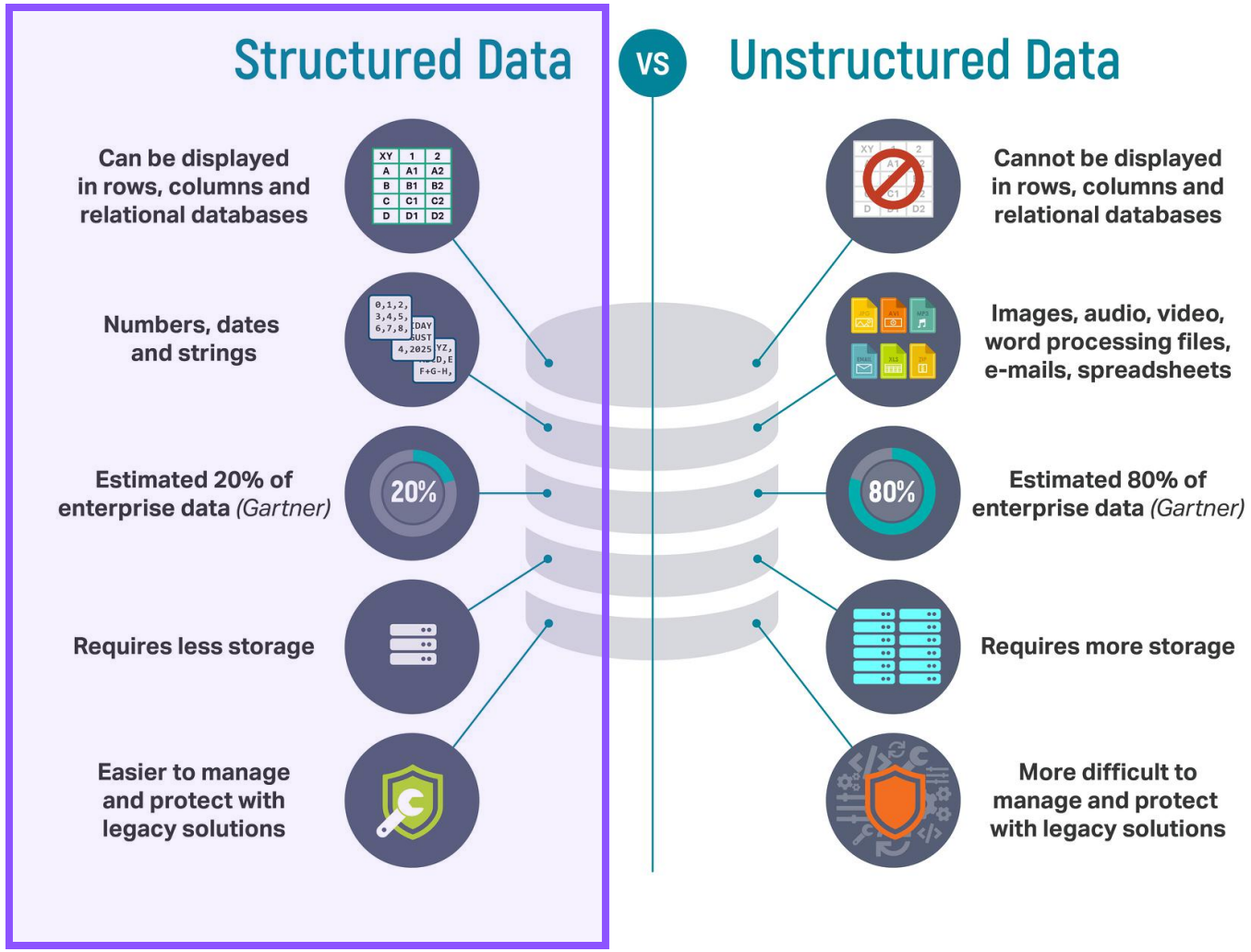
prof. dr. Samuel Martins (Samuka)
@xavecoding @hisamuka

Tipos de dados



https://lawtomated.com/structured-data-vs-unstructured-data-what-are-they-and-why-care/

Tipos de dados



https://lawtomated.com/structured-data-vs-unstructured-data-what-are-they-and-why-care/

Dados estruturados
Elementos Chave

	DATA INICIAL	DATA FINAL	MÊS	ANO	ESTADO	NÚMERO DE POSTOS PESQUISADOS	UNIDADE DE MEDIDA	PREÇO MÉDIO REVENDA
0	2004-05-09	2004-05-15	5	2004	DISTRITO FEDERAL	127	R$/l	1.288
1	2004-05-09	2004-05-15	5	2004	GOIAS	387	R$/l	1.162
2	2004-05-09	2004-05-15	5	2004	MATO GROSSO	192	R$/l	1.389
3	2004-05-09	2004-05-15	5	2004	MATO GROSSO DO SUL	162	R$/l	1.262
4	2004-05-09	2004-05-15	5	2004	ALAGOAS	103	R$/l	1.181
5	2004-05-09	2004-05-15	5	2004	BAHIA	408	R$/l	1.383

Dataset: Gas Prices in Brazil: https://www.kaggle.com/matheusfreitag/gas-prices-in-brazil

Dados estruturados
Elementos Chave

	DATA INICIAL	DATA FINAL	MÊS	ANO	ESTADO	NÚMERO DE POSTOS PESQUISADOS	UNIDADE DE MEDIDA	PREÇO MÉDIO REVENDA
0	2004-05-09	2004-05-15	5	2004	DISTRITO FEDERAL	127	R$/l	1.288
1	2004-05-09	2004-05-15	5	2004	GOIAS	387	R$/l	1.162
2	2004-05-09	2004-05-15	5	2004	MATO GROSSO	192	R$/l	1.389
3	2004-05-09	2004-05-15	5	2004	MATO GROSSO DO SUL	162	R$/l	1.262
4	2004-05-09	2004-05-15	5	2004	ALAGOAS	103	R$/l	1.181
5	2004-05-09	2004-05-15	5	2004	BAHIA	408	R$/l	1.383

Data Frame,
Table, Rectangular Data

Dataset: **Gas Prices in Brazil**: https://www.kaggle.com/matheusfreitag/gas-prices-in-brazil

Dados estruturados
Elementos Chave

Registro, Exemplo,
Observação,
Amostra (*Sample)**

	DATA INICIAL	DATA FINAL	MÊS	ANO	ESTADO	NÚMERO DE POSTOS PESQUISADOS	UNIDADE DE MEDIDA	PREÇO MÉDIO REVENDA
0	2004-05-09	2004-05-15	5	2004	DISTRITO FEDERAL	127	R$/l	1.288
1	2004-05-09	2004-05-15	5	2004	GOIAS	387	R$/l	1.162
2	2004-05-09	2004-05-15	5	2004	MATO GROSSO	192	R$/l	1.389
3	2004-05-09	2004-05-15	5	2004	MATO GROSSO DO SUL	162	R$/l	1.262
4	2004-05-09	2004-05-15	5	2004	ALAGOAS	103	R$/l	1.181
5	2004-05-09	2004-05-15	5	2004	BAHIA	408	R$/l	1.383

Data Frame,
Table, Rectangular Data

* O termo **Amostra/*Sample*** tem significados diferentes em **Estatística** e **Ciência de Dados** (veremos já)

Dataset: Gas Prices in Brazil: https://www.kaggle.com/matheusfreitag/gas-prices-in-brazil

Dados estruturados
Elementos Chave

Registro, Exemplo,
Observação,
Amostra (*Sample)**

Característica (*Feature*), Atributo, Variável, Entrada

	DATA INICIAL	DATA FINAL	MÊS	ANO	ESTADO	NÚMERO DE POSTOS PESQUISADOS	UNIDADE DE MEDIDA	PREÇO MÉDIO REVENDA
0	2004-05-09	2004-05-15	5	2004	DISTRITO FEDERAL	127	R$/l	1.288
1	2004-05-09	2004-05-15	5	2004	GOIAS	387	R$/l	1.162
2	2004-05-09	2004-05-15	5	2004	MATO GROSSO	192	R$/l	1.389
3	2004-05-09	2004-05-15	5	2004	MATO GROSSO DO SUL	162	R$/l	1.262
4	2004-05-09	2004-05-15	5	2004	ALAGOAS	103	R$/l	1.181
5	2004-05-09	2004-05-15	5	2004	BAHIA	408	R$/l	1.383

Data Frame,
Table, Rectangular Data

* O termo **Amostra/*Sample*** tem significados diferentes em **Estatística** e **Ciência de Dados** (veremos já)

Dataset: Gas Prices in Brazil: https://www.kaggle.com/matheusfreitag/gas-prices-in-brazil

Dados estruturados
Elementos Chave

Registro, Exemplo,
Observação,
Amostra (*Sample)**

Característica (*Feature*), Atributo, Variável, Entrada

	DATA INICIAL	DATA FINAL	MÊS	ANO	ESTADO	NÚMERO DE POSTOS PESQUISADOS	UNIDADE DE MEDIDA	PREÇO MÉDIO REVENDA
0	2004-05-09	2004-05-15	5	2004	DISTRITO FEDERAL	127	R$/l	1.288
1	2004-05-09	2004-05-15	5	2004	GOIAS	387	R$/l	1.162
2	2004-05-09	2004-05-15	5	2004	MATO GROSSO	192	R$/l	1.389
3	2004-05-09	2004-05-15	5	2004	MATO GROSSO DO SUL	162	R$/l	1.262
4	2004-05-09	2004-05-15	5	2004	ALAGOAS	103	R$/l	1.181
5	2004-05-09	2004-05-15	5	2004	BAHIA	408	R$/l	1.383

Índice
(comumente numérico, mas poderia ser textual)

Data Frame,
Table, Rectangular Data

* O termo **Amostra/*Sample*** tem significados diferentes em **Estatística** e **Ciência de Dados** (veremos já)

Dataset: Gas Prices in Brazil: https://www.kaggle.com/matheusfreitag/gas-prices-in-brazil

Dados estruturados

Elementos Chave

Series

("array/vetor", "feature vector")

DATA INICIAL	2004-05-09
DATA FINAL	2004-05-15
MÊS	5
ANO	2004
ESTADO	DISTRITO FEDERAL
NÚMERO DE POSTOS PESQUISADOS	127
UNIDADE DE MEDIDA	R$/l
PREÇO MÉDIO REVENDA	1.288

	DATA INICIAL	DATA FINAL	MÊS	ANO	ESTADO	NÚMERO DE POSTOS PESQUISADOS	UNIDADE DE MEDIDA	PREÇO MÉDIO REVENDA
0	2004-05-09	2004-05-15	5	2004	DISTRITO FEDERAL	127	R$/l	1.288
1	2004-05-09	2004-05-15	5	2004	GOIAS	387	R$/l	1.162
2	2004-05-09	2004-05-15	5	2004	MATO GROSSO	192	R$/l	1.389
3	2004-05-09	2004-05-15	5	2004	MATO GROSSO DO SUL	162	R$/l	1.262
4	2004-05-09	2004-05-15	5	2004	ALAGOAS	103	R$/l	1.181
5	2004-05-09	2004-05-15	5	2004	BAHIA	408	R$/l	1.383

Dataset: Gas Prices in Brazil: https://www.kaggle.com/matheusfreitag/gas-prices-in-brazil

Dados estruturados

Elementos Chave

Series
("array/vetor", "feature vector")

```
DATA INICIAL                          2004-05-09
DATA FINAL                            2004-05-15
MÊS                                            5
ANO                                         2004
ESTADO                          DISTRITO FEDERAL
NÚMERO DE POSTOS PESQUISADOS                 127
UNIDADE DE MEDIDA                            R$/l
PREÇO MÉDIO REVENDA                         1.288
```

Series
("array/vetor")

```
0    1.288
1    1.162
2    1.389
3    1.262
4    1.181
5    1.383
Name: PREÇO MÉDIO REVENDA, dtype: float64
```

	DATA INICIAL	DATA FINAL	MÊS	ANO	ESTADO	NÚMERO DE POSTOS PESQUISADOS	UNIDADE DE MEDIDA	PREÇO MÉDIO REVENDA
0	2004-05-09	2004-05-15	5	2004	DISTRITO FEDERAL	127	R$/l	1.288
1	2004-05-09	2004-05-15	5	2004	GOIAS	387	R$/l	1.162
2	2004-05-09	2004-05-15	5	2004	MATO GROSSO	192	R$/l	1.389
3	2004-05-09	2004-05-15	5	2004	MATO GROSSO DO SUL	162	R$/l	1.262
4	2004-05-09	2004-05-15	5	2004	ALAGOAS	103	R$/l	1.181
5	2004-05-09	2004-05-15	5	2004	BAHIA	408	R$/l	1.383

Dataset: Gas Prices in Brazil: https://www.kaggle.com/matheusfreitag/gas-prices-in-brazil

Dados estruturados

Elementos Chave

Variáveis Independentes

Variável Dependente
(Saída/Output, Target, Resposta)

	DATA INICIAL	DATA FINAL	MÊS	ANO	ESTADO	NÚMERO DE POSTOS PESQUISADOS	UNIDADE DE MEDIDA	PREÇO MÉDIO REVENDA
0	2004-05-09	2004-05-15	5	2004	DISTRITO FEDERAL	127	R$/l	1.288
1	2004-05-09	2004-05-15	5	2004	GOIAS	387	R$/l	1.162
2	2004-05-09	2004-05-15	5	2004	MATO GROSSO	192	R$/l	1.389
3	2004-05-09	2004-05-15	5	2004	MATO GROSSO DO SUL	162	R$/l	1.262
4	2004-05-09	2004-05-15	5	2004	ALAGOAS	103	R$/l	1.181
5	2004-05-09	2004-05-15	5	2004	BAHIA	408	R$/l	1.383

Dataset: **Gas Prices in Brazil**: https://www.kaggle.com/matheusfreitag/gas-prices-in-brazil

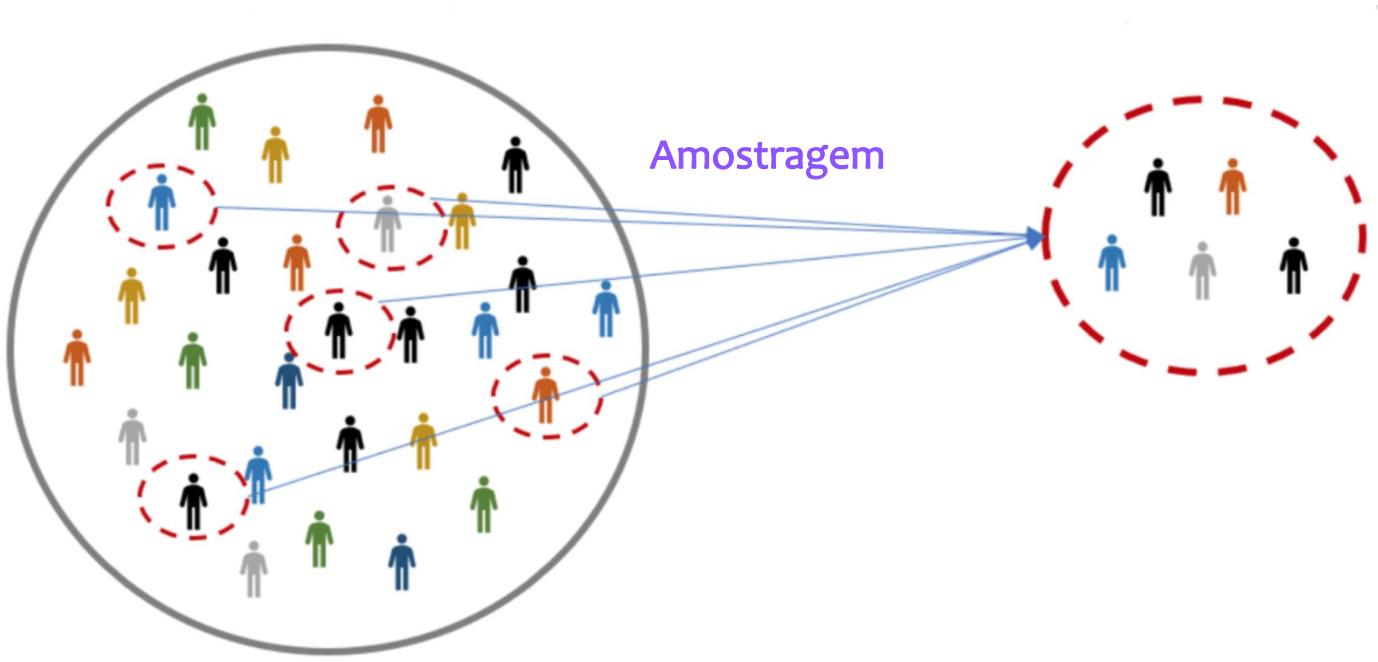
Dados estruturados
Elementos Chave

	variáveis independentes			variável dependente
observações	**Idade**	**Titulação**	**Experiência (anos)**	**Salário Anual ($)**
	21	Graduação	1	35,000.00
	25	Especialização	5	80,000.00
	35	Doutorado	10	120,000.00
	...	...	...	...

Outro Exemplo: Dataset Fictício sobre Salários nos USA

Diferenças de Terminologia

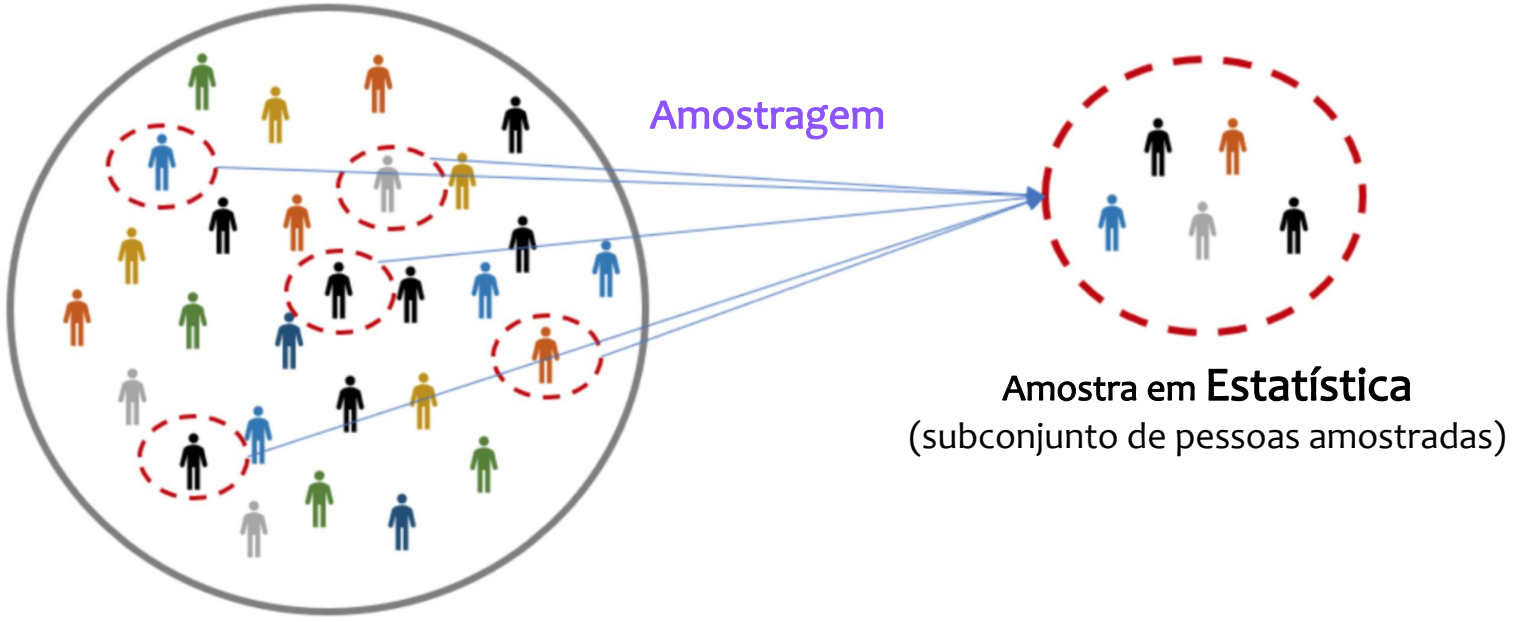
Exemplo: Pesquisa Eleitoral



População
(todos os eleitores do país)

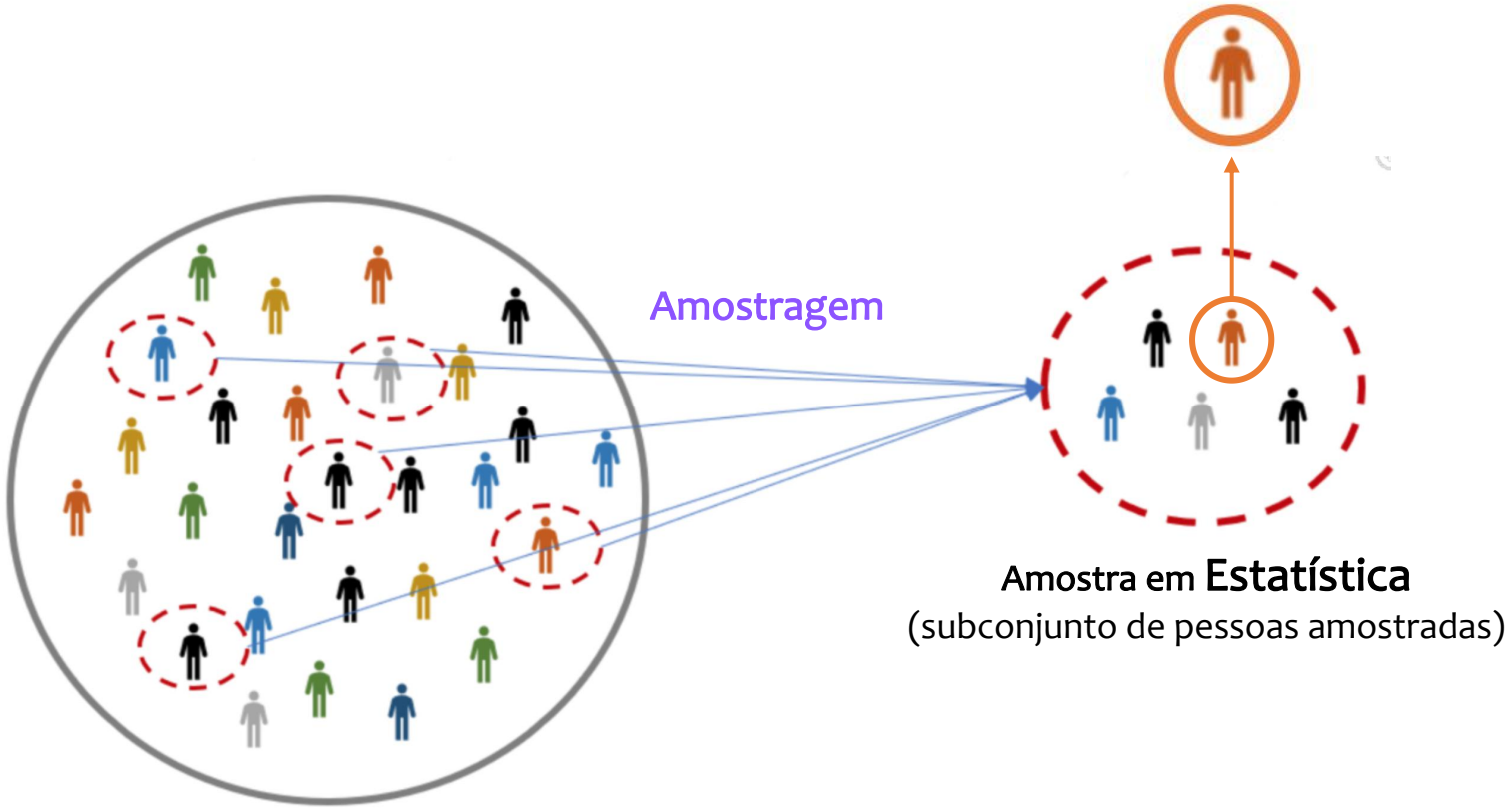
Diferenças de Terminologia

Exemplo: Pesquisa Eleitoral



Diferenças de Terminologia

Exemplo: Pesquisa Eleitoral

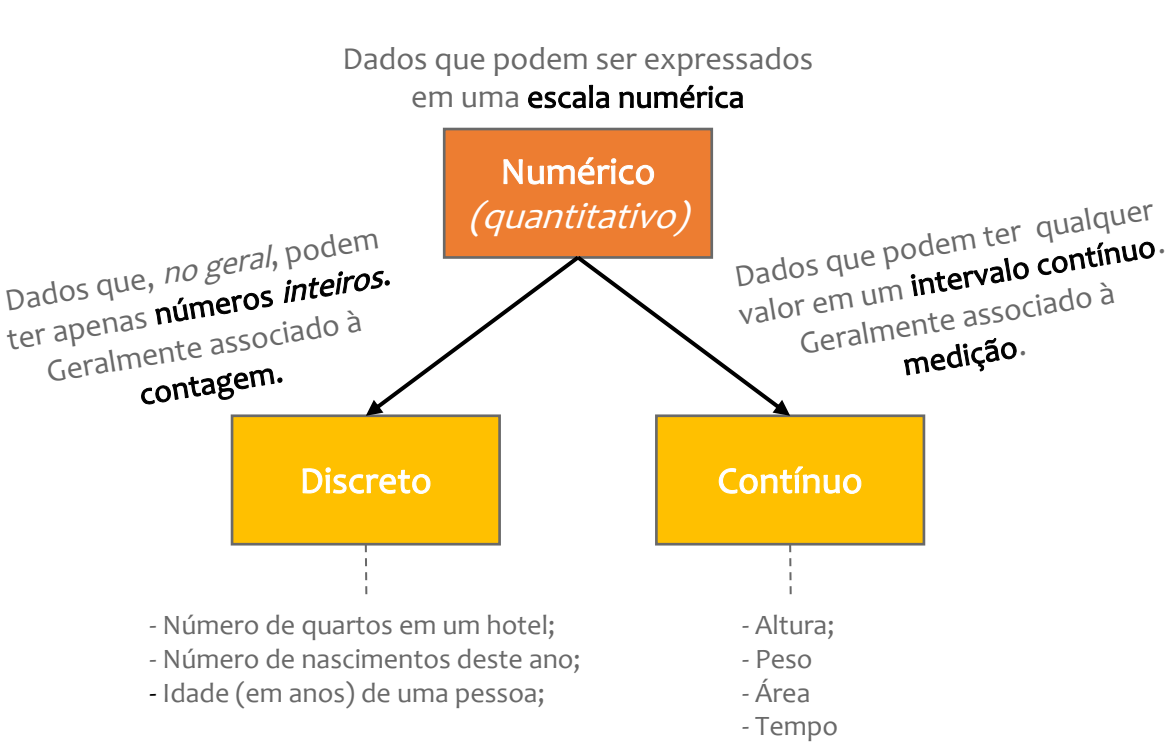


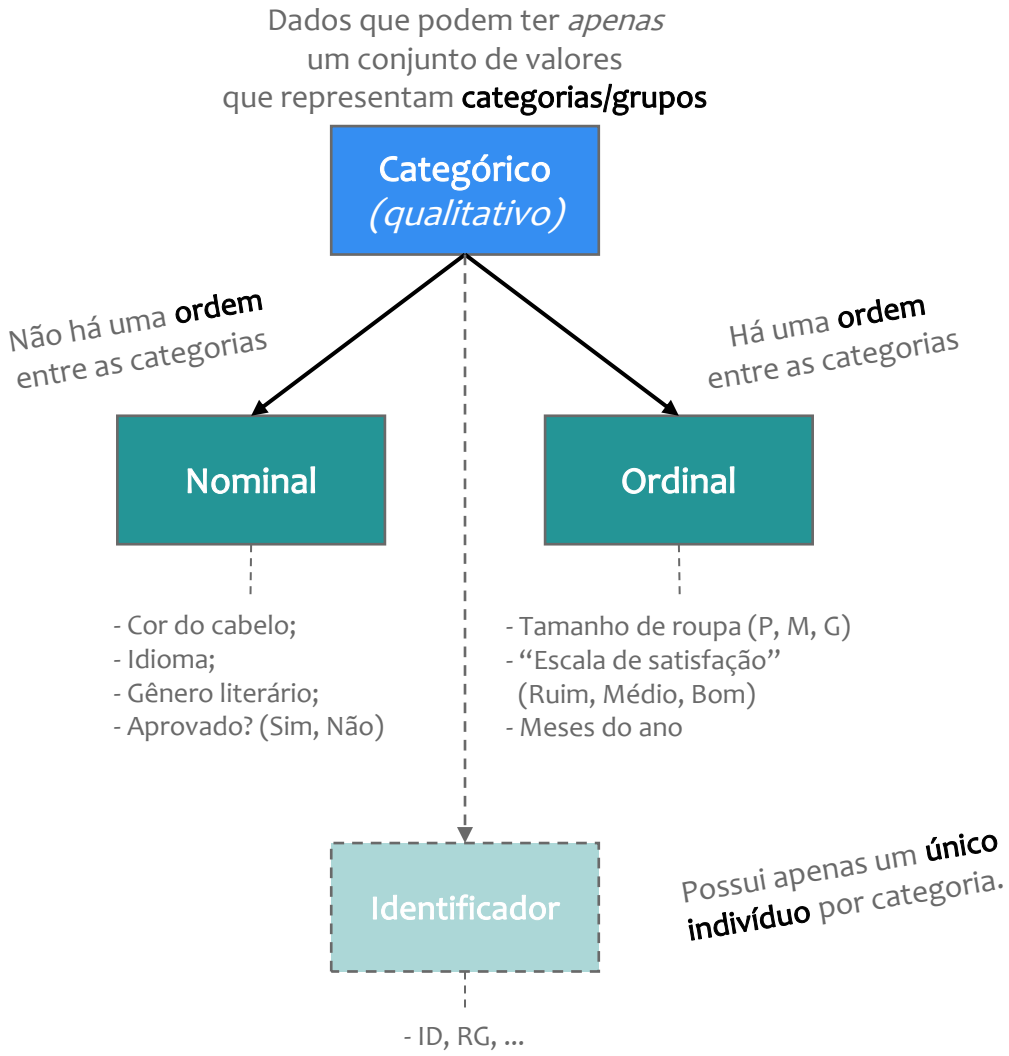
População
(todos os eleitores do país)

Amostra em Estatística
(subconjunto de pessoas amostradas)

Amostra em Ciência de Dados
(um exemplo do conjunto de dados)

Tipos de Dados





Outros Tipos:

Texto: String
- A sinopse de um filme
- Descrição de ativo na bolsa

Datas: String que representa datas.
Pode ser convertido em novas variáveis, como meses, anos, ...

Quais é o tipo de dados para cada feature abaixo?

	DATA INICIAL	DATA FINAL	MÊS	ANO	ESTADO	NÚMERO DE POSTOS PESQUISADOS	UNIDADE DE MEDIDA	PREÇO MÉDIO REVENDA
0	2004-05-09	2004-05-15	5	2004	DISTRITO FEDERAL	127	R$/l	1.288
1	2004-05-09	2004-05-15	5	2004	GOIAS	387	R$/l	1.162
2	2004-05-09	2004-05-15	5	2004	MATO GROSSO	192	R$/l	1.389
3	2004-05-09	2004-05-15	5	2004	MATO GROSSO DO SUL	162	R$/l	1.262
4	2004-05-09	2004-05-15	5	2004	ALAGOAS	103	R$/l	1.181
5	2004-05-09	2004-05-15	5	2004	BAHIA	408	R$/l	1.383

Dataset: Gas Prices in Brazil: https://www.kaggle.com/matheusfreitag/gas-prices-in-brazil

Quais é o tipo de dados para cada feature abaixo?

Proveniente dos atributos de datas.
Também poderiam ser **Categórico Ordinal.**

	Data	Data	Numérico Discreto	Numérico Discreto	Categórico Nomimal	Numérico Discreto	Categórico Nomimal	Numérico Contínuo
	DATA INICIAL	DATA FINAL	MÊS	ANO	ESTADO	NÚMERO DE POSTOS PESQUISADOS	UNIDADE DE MEDIDA	PREÇO MÉDIO REVENDA
0	2004-05-09	2004-05-15	5	2004	DISTRITO FEDERAL	127	R$/l	1.288
1	2004-05-09	2004-05-15	5	2004	GOIAS	387	R$/l	1.162
2	2004-05-09	2004-05-15	5	2004	MATO GROSSO	192	R$/l	1.389
3	2004-05-09	2004-05-15	5	2004	MATO GROSSO DO SUL	162	R$/l	1.262
4	2004-05-09	2004-05-15	5	2004	ALAGOAS	103	R$/l	1.181
5	2004-05-09	2004-05-15	5	2004	BAHIA	408	R$/l	1.383

Dataset: Gas Prices in Brazil: https://www.kaggle.com/matheusfreitag/gas-prices-in-brazil

Casos Especiais

Variáveis Discretas tradas como Variáveis Contínuas

O **dinheiro** ou o **preço** de algo varia em passos de **1 centavo**, então é uma variável discreta.

Porém, se você está lidando com **centenas de Reais**, os **passos são tão pequenos** que ele pode ser tratado como uma variável contínua.

Veja mais: link

Casos Especiais

Variáveis Categóricas representadas com Números

Às vezes, variáveis categóricas pode ser representadas com **números.**

Tais **números** não ter **sentido "númerico"**, no sentido que, não faz sentido você utilizar tais números em operações como, soma, subtração, etc.

Ex 1) Escala de Satisfação de um atendimento: 0: Péssimo; 1: Ruim; 2: Bom; 3: Excelente

Os números da escala não tem um **sentido aritmético:** Um atendimento Ruim (1) + uma atendimento Bom (2) não dá um atendimento Excelente (3).

Os números são utilizados, neste caso, como uma orientação ou apenas como um identificador.

Ex 2) Poderíamos representar a escolaridade com números: **1° grau (1), 2° grau (2) e 3° grau (3).**

Porém, não dá (e nem faz sentido) realizarmos **operações aritméticas** com tais valores.

P. ex, uma pessoa que tem o 1° e 2° grau não tem, por consequência, o 3° grau (1 + 2).

Ex 3) Avaliação de filmes.

Um filme nota 4.0 pode não ser, necessariamente, 2 vezes melhor do que um filme nota 2.0.

Tal escala de valor, pode não expressar, exatamente, a magnitude da diferença entre a qualidade de dois filmes.

Pandas Essencial

Conceitos Básicos sobre Manipulação de Dados

prof. dr. Samuel Martins (Samuka)
@xavecoding @hisamuka

Pandas Essencial

Conceitos Básicos sobre Manipulação de Dados

prof. dr. Samuel Martins (Samuka)
@xavecoding @hisamuka

xavecoding