

# Data Mining

## Transformação/Integração

Prof. Dr. Joaquim Assunção

DEPARTAMENTO DE COMPUTAÇÃO APLICADA  
CENTRO DE TECNOLOGIA  
UFSM  
2022

[www.inf.ufsm.br/~joaquim](http://www.inf.ufsm.br/~joaquim)



# *Fair user agreement*

Este material foi criado para a disciplina de Mineração de Dados - Centro de Tecnologia da UFSM.

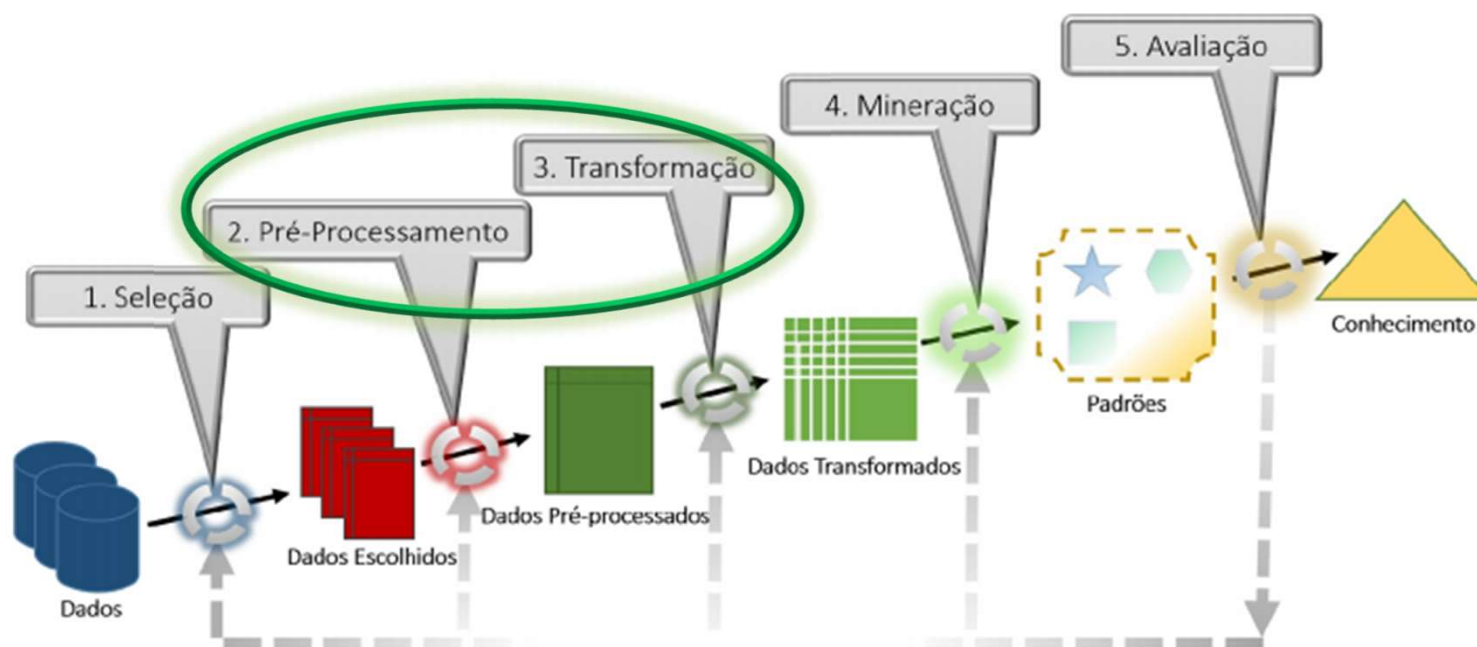
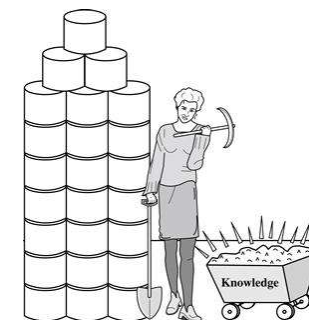
Você pode usar este material livremente\*; porém, caso seja usado em outra instituição, **me envie um e-mail** avisando o nome da instituição e a disciplina.

\*Caso você queira usar algo desse material em alguma publicação, envie-me um e-mail com antecedência.

Prof. Dr. Joaquim Assunção.

joaquim@inf.ufsm.br

# Transformação



# Problemas de Integração

- **Dados heterogêneos:** Em banco de dados, seriam dados sem uma chave comum.
- **Definição diferente:** Os mesmos dados com definições diferentes, como um esquema de banco de dados diferente.
- **Sincronização de tempo:** Dados com períodos de tempo diferentes. Isto é, coletados em períodos diferentes.
- **Dados legados:** Dados provenientes de um sistema antigo.

# Problemas de Integração

- **Redundância e análise de correlação:** algumas redundâncias podem ser detectadas por análise de correlação. Dados dois atributos, tal análise pode medir quão fortemente um atributo implica o outro, com base nos dados disponíveis.
- **Duplicação de Tuplas:** A duplicação deve ser detectada no nível da tupla para detectar redundâncias entre os atributos.
- **Detecção e resolução de conflitos de valores de dados:** os atributos podem diferir no nível de abstração, onde um atributo em um sistema é registrado em um nível de abstração diferente.

# Subconjuntos e ordenação

# Subconjuntos

- Conceito: Retirar parte de um conjunto de dados  $D$  de modo a formar um conjunto  $D'$  sem os dados indesejados.
- Exemplo 1, somente os heróis devem ficar.

Nome	Time	Super Força	Vilão/ Herói	QI
Tony Stark	Avengers	S	H	150
St. Rogers	Avengers	S	H	90
Dr.Strange	Vários	N	H	160
P. Quill	Guardiões	N	H	100
Drax	Guardiões	S	H	70
Thanos	Thanos	S	V	NA
C. glaive	Thanos	S	V	NA

Nome	Time	Super Força	Vilão/ Herói	QI
Tony Stark	Avengers	S	H	150
St. Rogers	Avengers	S	H	90
Dr.Strange	Vários	N	H	160
P. Quill	Guardiões	N	H	100
Drax	Guardiões	S	H	70

# Subconjuntos

- Exemplo 2, somente nome e time daqueles com super força.

Nome	Time	Super Força	Vilão/ Herói	QI
Tony Stark	Avengers	S	H	150
St. Rogers	Avengers	S	H	90
Dr.Strange	Vários	N	H	160
P. Quill	Guardiões	N	H	100
Drax	Guardiões	S	H	70
Thanos	Thanos	S	V	NA
C. glaive	Thanos	S	V	NA

Nome	Time	
Tony Stark	Avengers	
St. Rogers	Avengers	
Drax	Guardiões	
Thanos	Thanos	
C. glaive	Thanos	



# *Technical help*

Em R:


- Use: `set.seed(<numero>)` para definir uma semente.
- Posteriormente crie um DataFrame com:
- `X <- data.frame("var1"=sample(1:5),  
"var2"=sample(6:10),"var3"=sample(11:15))`

```
> X
  var1 var2 var3
1    2    6   12
2    3   NA   13
3    1   10   11
4    5    9   15
5    4   NA   14
```

## *Technical help*

- Podemos filtrar por colunas usando o índice da coluna ou o nome da variável.
- Observe que o primeiro index começa em 1.

```
> X
  var1 var2 var3
1     2     6  12
2     3    NA  13
3     1    10  11
4     5     9  15
5     4    NA  14
```



```
> X[,1]
[1] 2 3 1 5 4
> X[,"var1"]
[1] 2 3 1 5 4
```

## *Technical help*

- Também podemos filtrar ambos, usando linha e coluna no parâmetro.

```
> X
  var1 var2 var3
1    2    6   12
2    3   NA   13
3    1   10   11
4    5    9   15
5    4   NA   14
```

```
> X[1:2, "var2"]
[1] 6 NA
```

```
> X[c(1,3), "var2"]
[1] 6 10
```

```
> X[c(1,3), c(1,3)]
  var1 var3
1    2   12
3    1   11
```

## *Technical help*

- Podemos usar operações lógicas com `&` e `|`

```
> X
  var1 var2 var3
1     2     6  12
2     3    NA  13
3     1    10  11
4     5     9  15
5     4    NA  14
```

```
> X[(X$var1 <= 2 & X$var3 > 10),]
  var1 var2 var3
1     2     6  12
3     1    10  11
> X[(X$var1 <= 3 | X$var3 > 14),]
  var1 var2 var3
1     2     6  12
2     3    NA  13
3     1    10  11
4     5     9  15
```

# Technical help

- Para filtrar valores faltantes, podemos usar `which` e `is.na`

```
> X
  var1 var2 var3
1    2    6   12
2    3   NA   13
3    1   10   11
4    5    9   15
5    4   NA   14
```

```
> X[which(!is.na(X$var2)),]
  var1 var2 var3
1    2    6   12
3    1   10   11
4    5    9   15
> X[!is.na(X$var2),]
  var1 var2 var3
1    2    6   12
3    1   10   11
4    5    9   15
```

# *Hands on!*

Use `read.csv()` para ler o arquivo *'FakeMarvelData'*.

1. Extraia o subconjunto *'Avengers'* somente com as variáveis *nome*, *afiliação* e *QI*.
2. Extraia o subconjunto de todos os registros cujos personagem não possuem um valor para *QI*.
3. Extraia o subconjunto de todos os registros cujos personagens possuem mais de 120 de *QI*.
4. Use a saída da questão 3 para ordenar por *QI* o subconjunto\*.

\*dica, use `order` ou `arrange` da biblioteca `plyr`

# Abordagens

- **Internas:**

- A seleção de características (de dados, consequentemente dados), ocorre naturalmente como parte do algoritmo de mineração de dados. Um caso típico são os classificadores por árvore de decisão.

- **De Envoltório:**

- Uso de algoritmos de mineração como caixa preta para selecionar dados.

# Abordagens

- **De Filtro:**
- Características são selecionadas antes da mineração (dos algoritmos de mineração). Essa abordagem é independente da mineração de modo com que o algoritmo não está atrelado ao filtro.