

MINERAÇÃO DE DADOS

Preparação e limpeza de dados

CONJUNTO DE DADOS

- **Dados brutos:**
 - **Conjunto de dados que não foram processados para uso**
 - **Pode conter valores errados**
 - **Verificação antes do uso para mineração**



Fonte: James St. John, Wikimedia Commons



Fonte: James St. John, Flickr

CONJUNTO DE DADOS

- **Objeto:**
 - **Uma instância do conjunto de dados**
 - **Registro ou linha**
- **Atributo:**
 - **Característica de objeto**
 - **Coluna ou campo**

PROBLEMAS COM DADOS - MOTIVOS

- **Falta de preenchimento**
- **Preenchimento incorreto**
- **Erros de medição**
- **Uso de unidades de medidas diferentes**

DADOS BRUTOS

Nome	Idade	Nível educacional	Estado civil	Gênero	Cartão de crédito	Renda mensal (\$)
Roberto Felix	42	Especialização	Divorciado	M	Sim	5.000
Joana Pereira	10	Doutorado	Viúva	F	Sim	6.500
?	?	?	?	?	?	?
Isabela Assis	33	Graduação	Casada	F	?	3.900
Marco Araújo	29	Graduação	89 Kg	M	Não	3.100

DADOS BRUTOS

Nome	Idade	Nível educacional	Estado civil	Gênero	Cartão de crédito	Renda mensal (\$)
Roberto Felix	42	Especialização	Divorciado	M	Sim	5.000
Joana Pereira	10	Doutorado	Viúva	F	Sim	6.500
?	?	?	?	?	?	?
Isabela Assis	33	Graduação	Casada	F	?	3.900
Marco Araújo	29	Graduação	89 Kg	M	Não	3.100

Estado civil: 89 Kg?

DADOS BRUTOS

Nome	Idade	Nível educacional	Estado civil	Gênero	Cartão de crédito	Renda mensal (\$)
Roberto Felix	42	Especialização	Divorciado	M	Sim	5.000
Joana Pereira	10	Doutorado	Viúva	F	Sim	6.500
?	?	?	?	?	?	?
Isabela Assis	33	Graduação	Casada	F	?	3.900
Marco Araújo	29	Graduação	89 Kg	M	Não	3.100

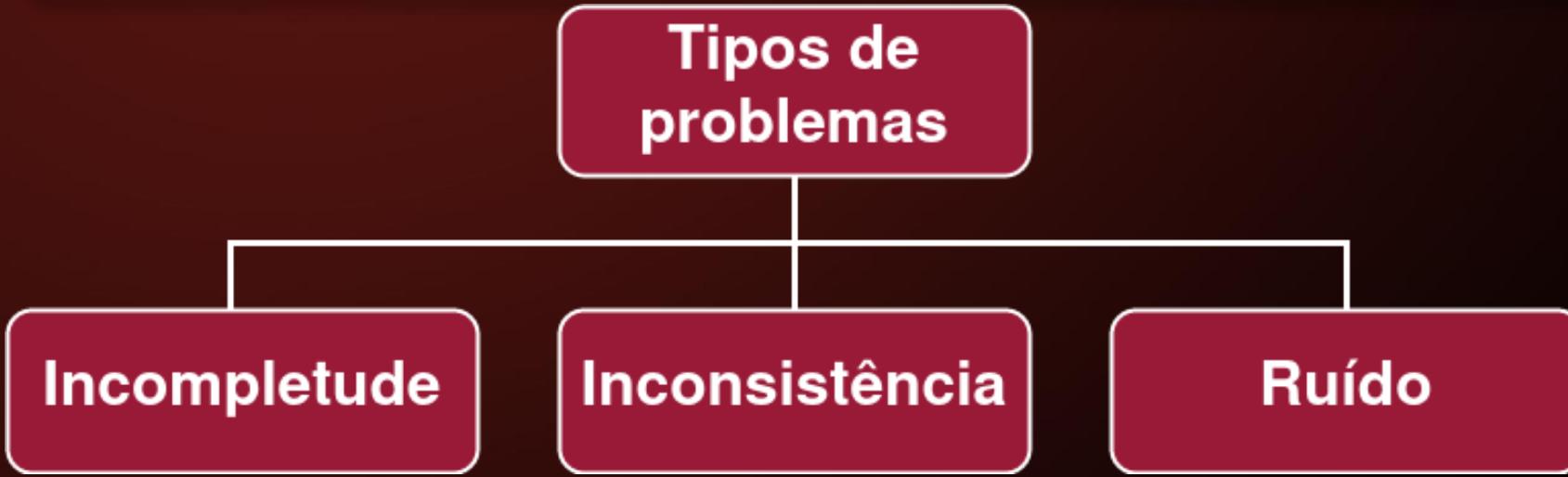
Doutorado com 10 anos?

DADOS BRUTOS

Nome	Idade	Nível educacional	Estado civil	Gênero	Cartão de crédito	Renda mensal (\$)
Roberto Felix	42	Especialização	Divorciado	M	Sim	5.000
Joana Pereira	10	Doutorado	Viúva	F	Sim	6.500
?	?	?	?	?	?	?
Isabela Assis	33	Graduação	Casada	F	?	3.900
Marco Araújo	29	Graduação	89 Kg	M	Não	3.100

Linha 3: ?

PROBLEMAS COM DADOS



PROBLEMAS COM DADOS



Valor, atributo ou objeto ausentes

INCOMPLETITUDE

Nome	Idade	Nível educacional	Estado civil	Gênero	Cartão de crédito	Renda mensal (\$)
Roberto Felix	42	Especialização	Divorciado	M	Sim	5.000
Joana Pereira	10	Doutorado	Viúva	F	Sim	6.500
?	?	?	?	?	?	?
Isabela Assis	33	Graduação	Casada	F	?	3.900
Marco Araújo	29	Graduação	89 Kg	M	Não	3.100

Linha 3: ?

PROBLEMAS COM DADOS



Violação do domínio

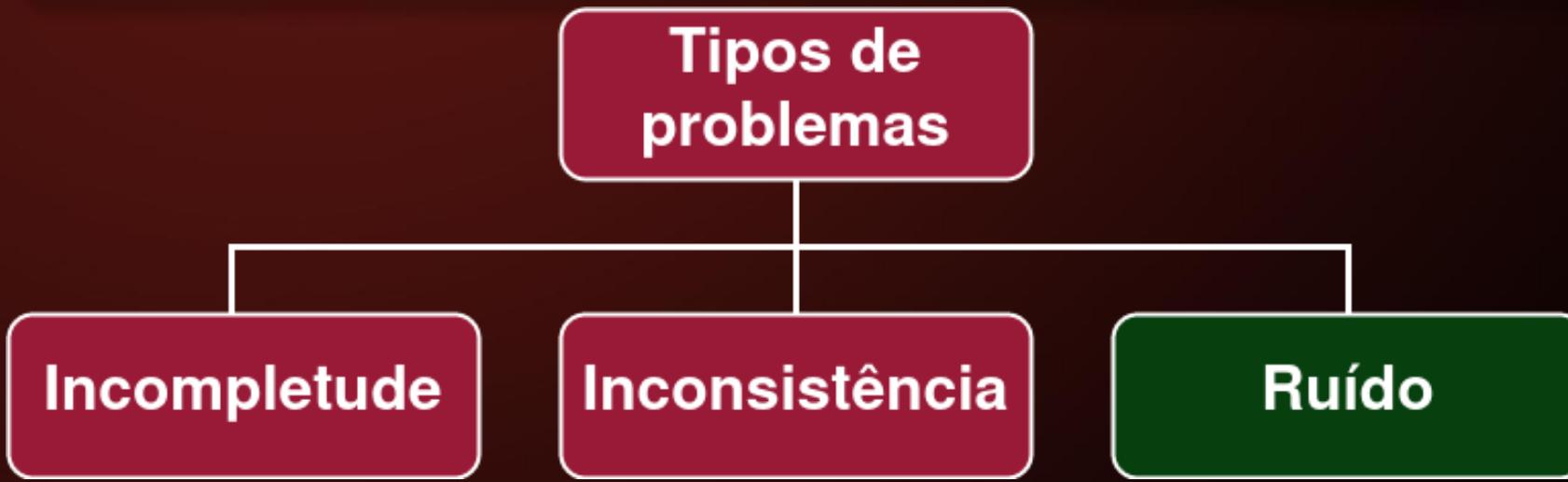
Valores diferentes de um atributo para
o mesmo objeto

INCONSISTÊNCIA

Nome	Idade	Nível educacional	Estado civil	Gênero	Cartão de crédito	Renda mensal (\$)
Roberto Felix	42	Especialização	Divorciado	M	Sim	5.000
Joana Pereira	10	Doutorado	Viúva	F	Sim	6.500
?	?	?	?	?	?	?
Isabela Assis	33	Graduação	Casada	F	?	3.900
Marco Araújo	29	Graduação	89 Kg	M	Não	3.100

Estado civil: 89 Kg?

PROBLEMAS COM DADOS



Valores inexplicáveis ou indesejados

Pode gerar inconsistência

RUÍDO

Nome	Idade	Nível educacional	Estado civil	Gênero	Cartão de crédito	Renda mensal (\$)
Roberto Felix	42	Especialização	Divorciado	M	Sim	5.000
Joana Pereira	10	Doutorado	Viúva	F	Sim	6.500
?	?	?	?	?	?	?
Isabela Assis	33	Graduação	Casada	F	?	3.900
Marco Araújo	29	Graduação	89 Kg	M	Não	3.100

Doutorado com 10 anos?

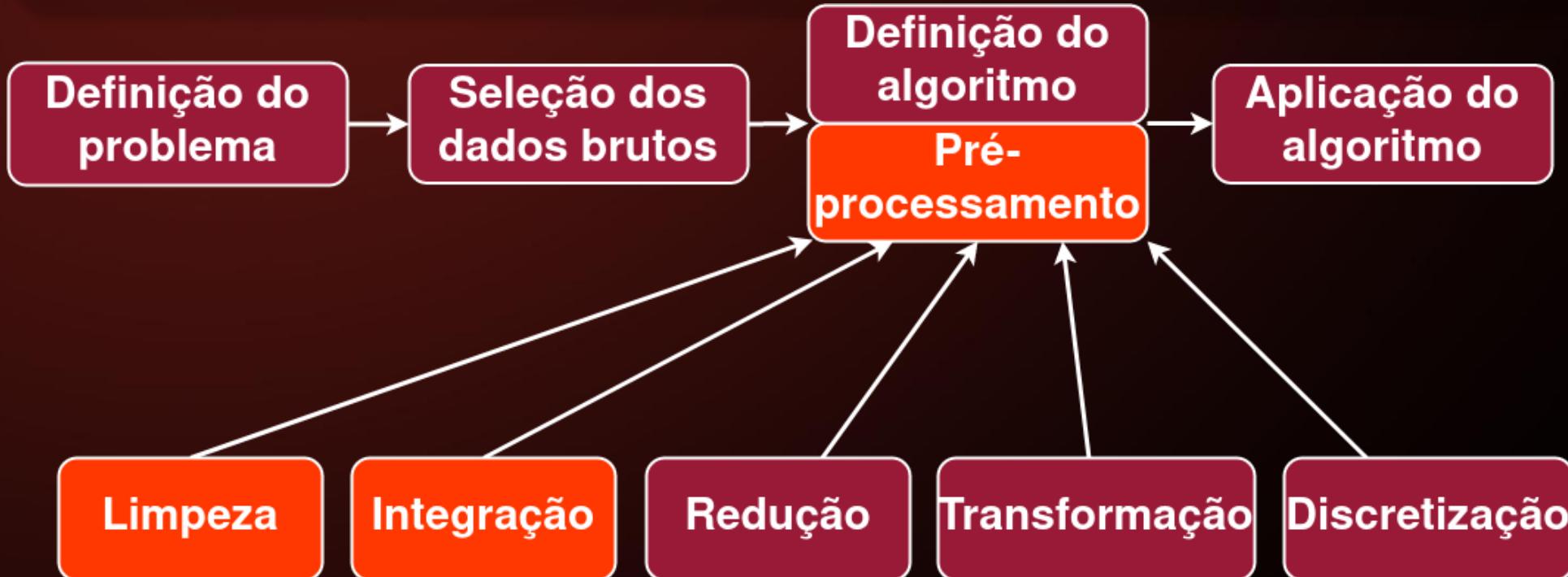
PROCESSO DE PREPARAÇÃO DOS DADOS



PROCESSO DE PREPARAÇÃO DOS DADOS



PROCESSO DE PREPARAÇÃO DOS DADOS



LIMPEZA DOS DADOS

- Valores ausentes:
 - Imputação
- Ruídos:
 - Agrupamento
 - Aproximação

LIMPEZA DOS DADOS - IMPUTAÇÃO

Métodos de imputação:

- Ignorar o objeto
- Manual
- Constante global
- *Hot-deck*: valor de objeto similar aleatório
- Valor da última observação:
 - Ordenação de um ou mais atributos

HOT - DECK

Nome	Idade	Nível educacional	Estado civil	Gênero	Cartão de crédito	Renda mensal (\$)
Roberto Felix	42	Especialização	Divorciado	M	Sim	5.000
Joana Pereira	10	Doutorado	Viúva	F	Sim	6.500
?	?	?	?	?	?	?
Isabela Assis	33	Graduação	Casada	F	?	3.900
Marco Araújo	29	Graduação	89 Kg	M	Não	?

HOT - DECK

Nome	Idade	Nível educacional	Estado civil	Gênero	Cartão de crédito	Renda mensal (\$)
Roberto Felix	42	Especialização	Divorciado	M	Sim	5.000
Joana Pereira	10	Doutorado	Viúva	F	Sim	6.500
?	?	?	?	?	?	?
Isabela Assis	33	Graduação	Casada	F	?	3.900
Marco Araújo	29	Graduação	89 Kg	M	Não	?

HOT - DECK

Nome	Idade	Nível educacional	Estado civil	Gênero	Cartão de crédito	Renda mensal (\$)
Roberto Felix	42	Especialização	Divorciado	M	Sim	5.000
Joana Pereira	10	Doutorado	Viúva	F	Sim	6.500
?	?	?	?	?	?	?
Isabela Assis	33	Graduação	Casada	F	?	3.900
Marco Araújo	29	Graduação	89 Kg	M	Não	3.900

VALOR DA ÚLTIMA OBSERVAÇÃO

Nome	Idade	Nível educacional	Estado civil	Gênero	Cartão de crédito	Renda mensal (\$)
Roberto Felix	42	Especialização	Divorciado	M	Sim	5.000
Joana Pereira	40	Doutorado	Viúva	F	Sim	6.500
João Carlos	25	Graduação	Solteiro	M	Sim	2.900
Isabela Assis	33	Graduação	Casada	F	Sim	3.900
Luísa Silva	27	Especialização	Solteira	F	Não	?
Marco Araújo	29	Graduação	Casado	M	Não	3.100

VALOR DA ÚLTIMA OBSERVAÇÃO

Nome	Idade	Nível educacional	Estado civil	Gênero	Cartão de crédito	Renda mensal (\$)
Roberto Felix	42	Especialização	Divorciado	M	Sim	5.000
Joana Pereira	40	Doutorado	Viúva	F	Sim	6.500
Isabela Assis	33	Graduação	Casada	F	Sim	3.900
Marco Araújo	29	Graduação	Casado	M	Não	3.100
Luísa Silva	27	Especialização	Solteira	F	Não	?
João Carlos	25	Graduação	Solteiro	M	Sim	2.900

VALOR DA ÚLTIMA OBSERVAÇÃO

Nome	Idade	Nível educacional	Estado civil	Gênero	Cartão de crédito	Renda mensal (\$)
Roberto Felix	42	Especialização	Divorciado	M	Sim	5.000
Joana Pereira	40	Doutorado	Viúva	F	Sim	6.500
Isabela Assis	33	Graduação	Casada	F	Sim	3.900
Marco Araújo	29	Graduação	Casado	M	Não	3.100
Luísa Silva	27	Especialização	Solteira	F	Não	3.100
João Carlos	25	Graduação	Solteiro	M	Sim	2.900

LIMPEZA DOS DADOS - IMPUTAÇÃO

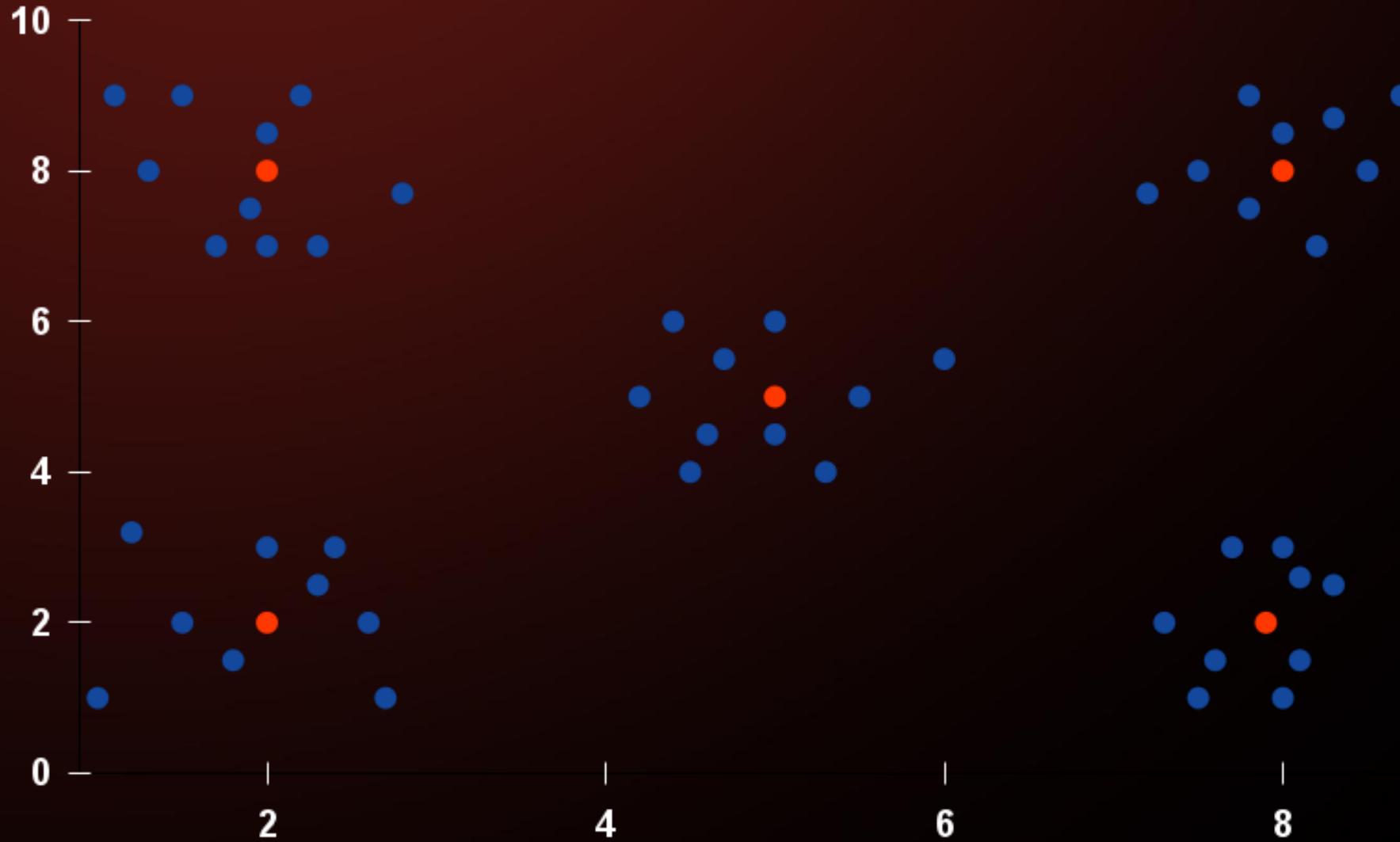
Métodos de imputação:

- Média ou moda:
 - Um atributo
 - Todos objetos da mesma classe
- Modelos preditivos:
 - Classificação / agrupamento

LIMPEZA DOS DADOS - AGRUPAMENTO

- Usar os demais atributos do objeto
- Atribuir o valor médio dos objetos do mesmo grupo

LIMPEZA DOS DADOS - AGRUPAMENTO

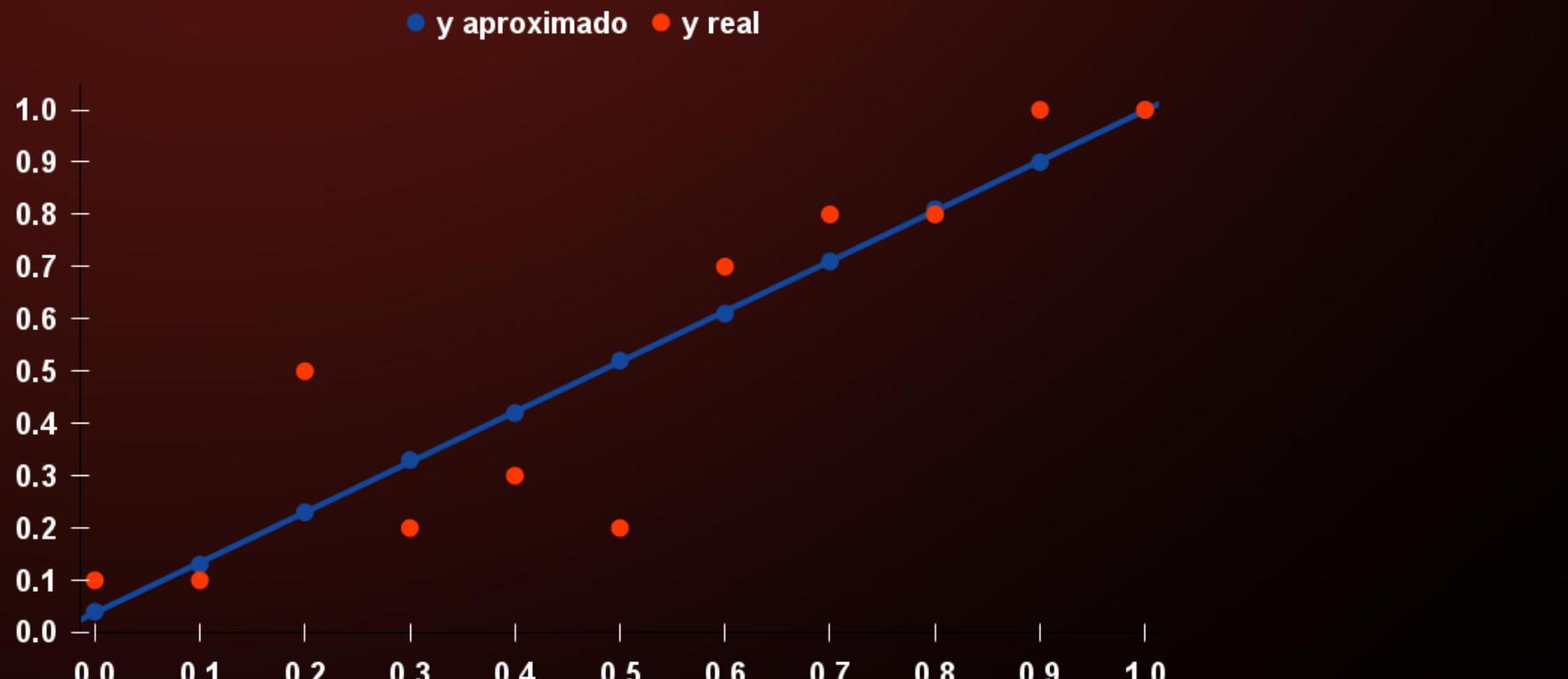


LIMPEZA DOS DADOS - APROXIMAÇÃO

- Função ou modelo de aproximação
- Substituir o valor real pelo valor obtido na função

LIMPEZA DOS DADOS - APROXIMAÇÃO

- Ex: Polinômio de grau 1 ($y = 0,9636x + 0,0364$)



x	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00
y real	0.10	0.10	0.50	0.20	0.30	0.20	0.70	0.80	0.80	1.00	1.00
y aproximado	0.04	0.13	0.23	0.33	0.42	0.52	0.61	0.71	0.81	0.90	1.00

INTEGRAÇÃO DE DADOS

- **Dados distribuídos em bases distintas:**
 - Departamentos, programas
 - Unir os dados em uma base única
 - Aspectos:
 - Redundância
 - Duplicidade
 - Conflitos

INTEGRAÇÃO DE DADOS - ASPECTOS

- Redundância:
 - Atributo obtido a partir de outros
 - Ex: data de nascimento e idade
- Duplicidade:
 - Mesmo dado em bases distintas
 - Tipo de redundância

INTEGRAÇÃO DE DADOS - ASPECTOS

- Conflitos:
 - Valores diferentes para um mesmo objeto em bases distintas
 - Representações diferentes:
 - Ex: distância em quilômetros e milhas
 - Ex: Feminino e F

REFERÊNCIAS

**Introdução à Mineração de Dados: Conceitos Básicos, Algoritmos e Aplicações: Cap. 2:
Pré-processamento de dados.**
Leandro Nunes de Castro e Daniel Gomes Ferrari. Editora Saraiva, 2016.