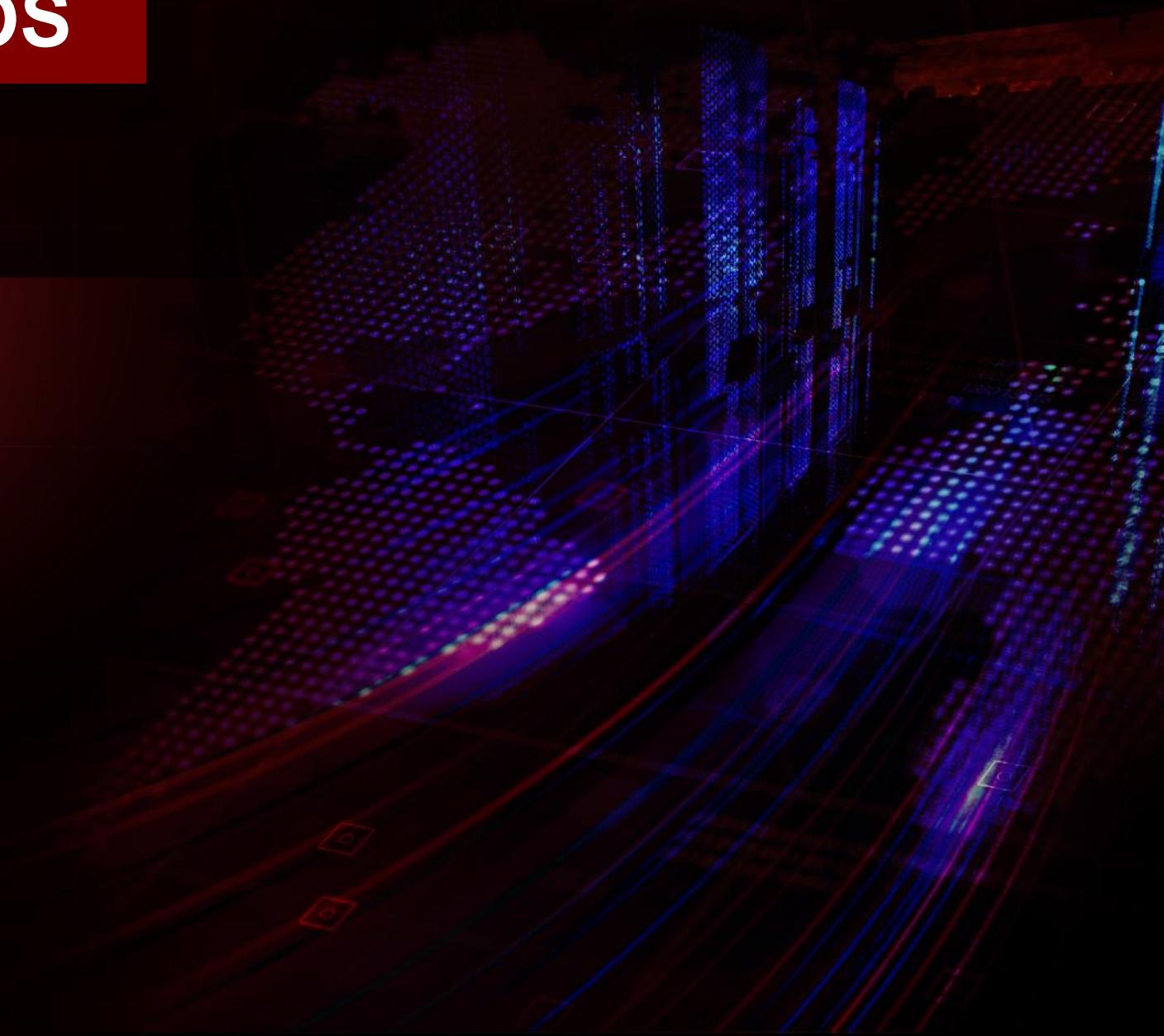


# MINERAÇÃO DE DADOS

## Agrupamento de dados



# APRENDIZADO NÃO SUPERVISIONADO

Há casos em que não existe uma rotulação prévia dos objetos

Objetos têm características próprias

Como identificar objetos semelhantes?



# APRENDIZADO NÃO SUPERVISIONADO



Fonte: pixabay

# APRENDIZADO NÃO SUPERVISIONADO



Fonte: pixabay

# AGRUPAMENTO DE DADOS

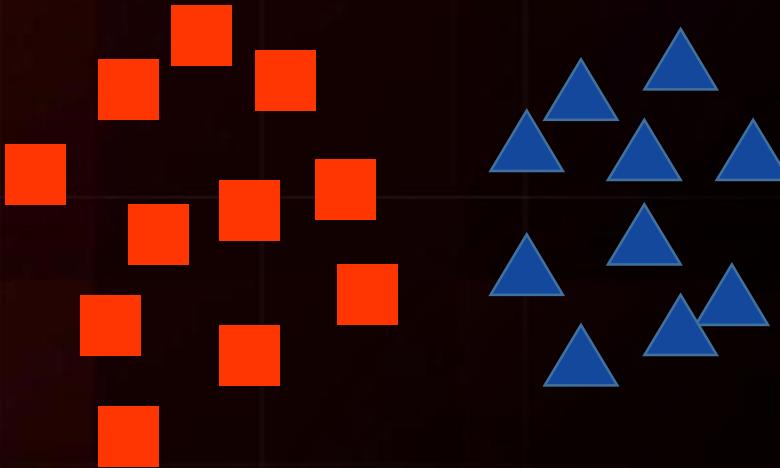
Métodos de análise de dados com o objetivo de descobrir grupos homogêneos

Baseiam-se na similaridade das características dos objetos

# AGRUPAMENTO DE DADOS

**Grupos (clusters): subconjuntos de objetos similares**

**Coesão interna X isolamento externo**



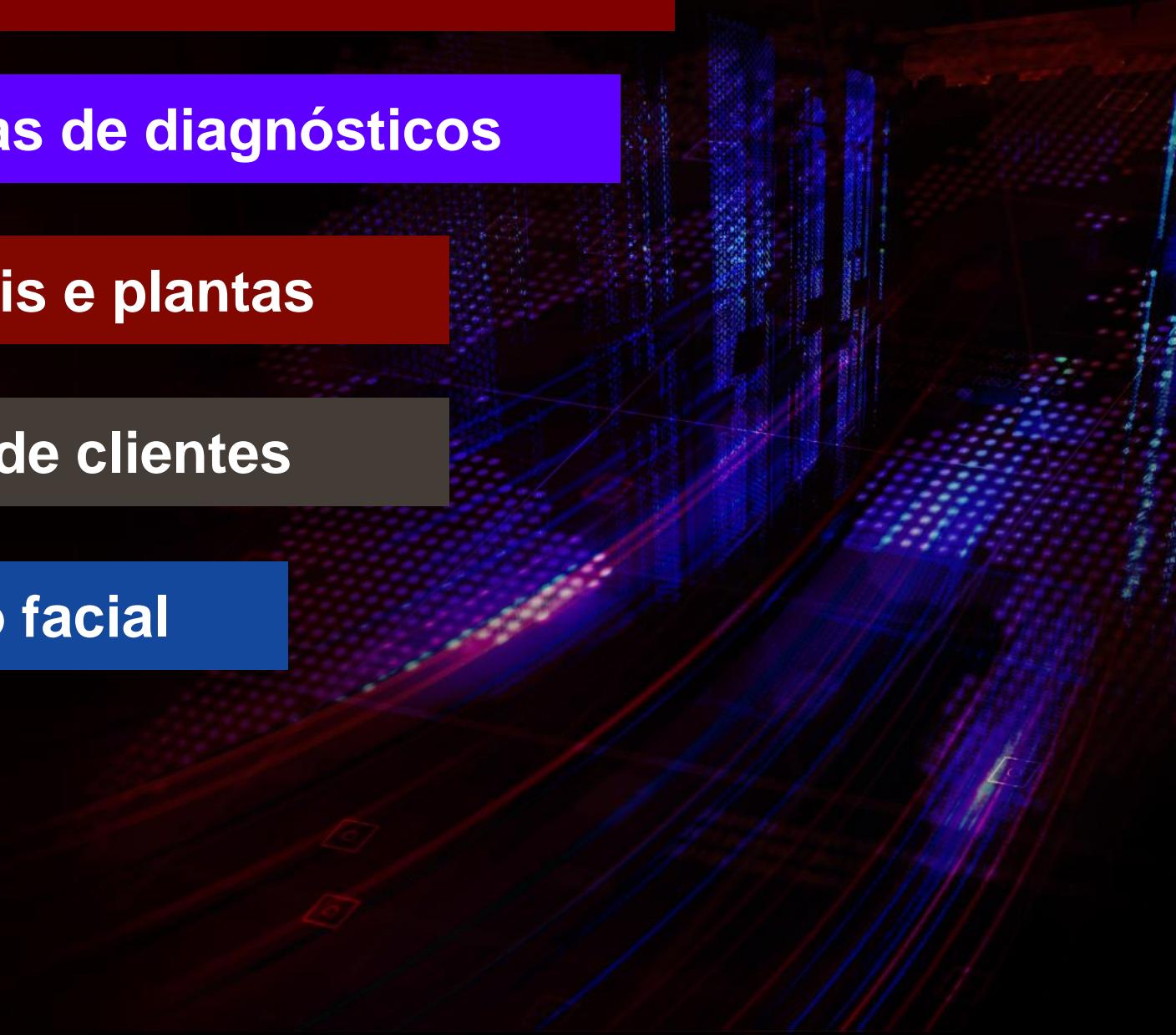
# USOS DE AGRUPAMENTO DE DADOS

Medicina: identificar categorias de diagnósticos

Biologia: taxonomia de animais e plantas

Marketing: identificar grupos de clientes

Análise de imagens: detecção facial



# AGRUPAMENTO X CLASSIFICAÇÃO

Agrupamento



Rótulos

# subconjuntos

Treinamento

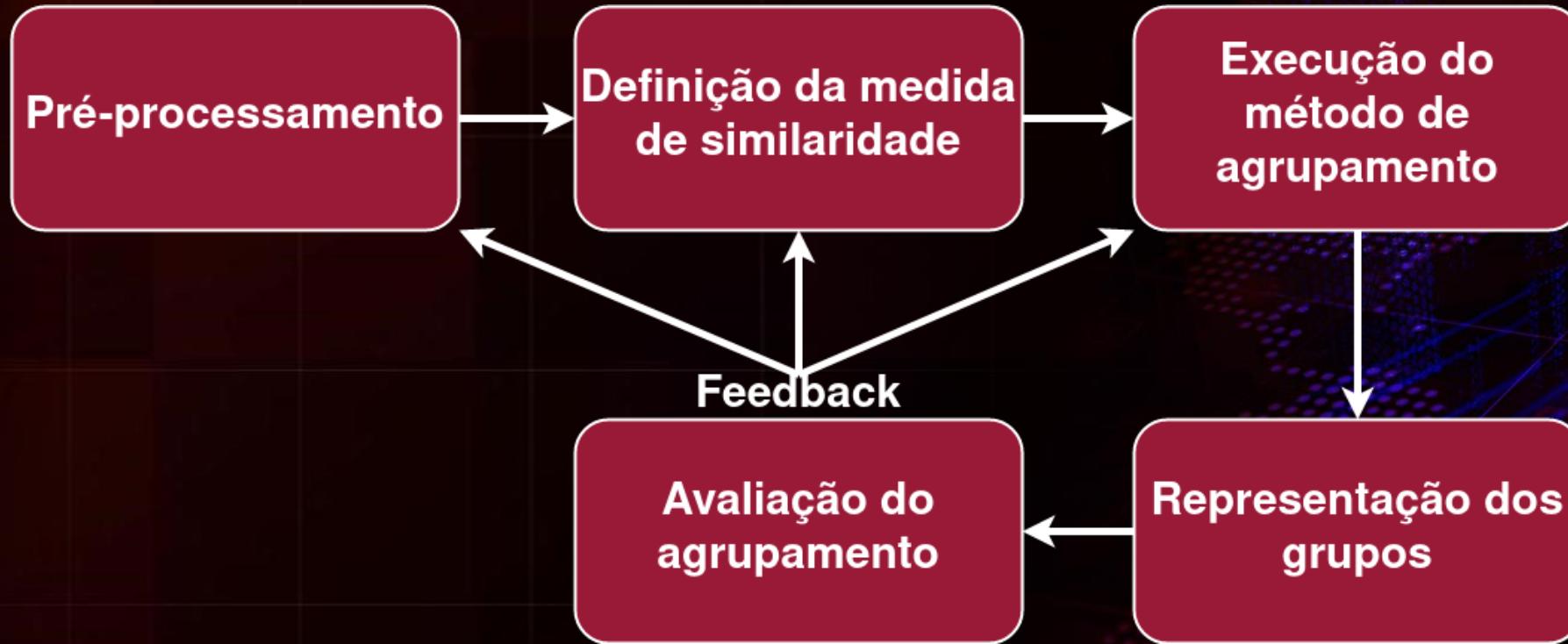
Classificação



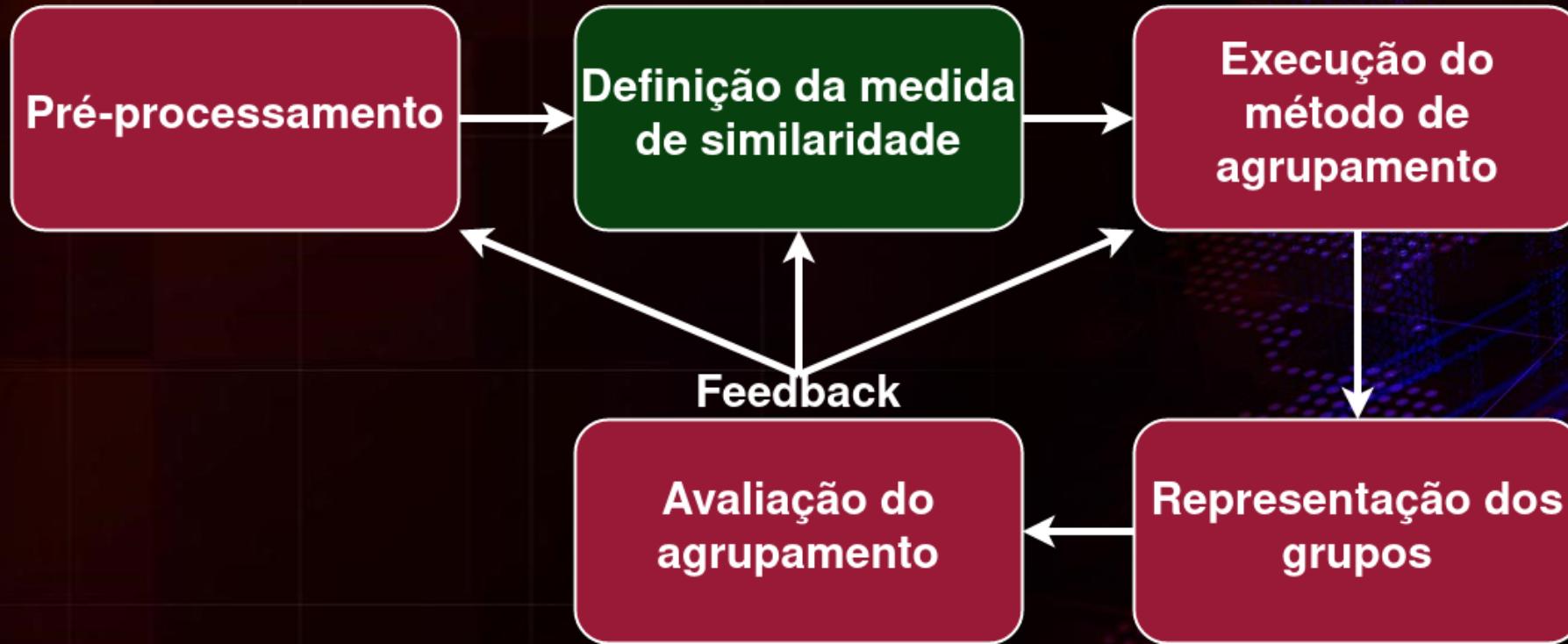
# ESCOLHA DO NÚMERO DE GRUPOS

- **Tentativa e erro**
  - Teste com diferentes valores
- **Baseado no conhecimento do domínio**
  - Interpretação representativa dos dados
  - Expectativa do especialista do domínio

# PROCESSO DE AGRUPAMENTO DE DADOS

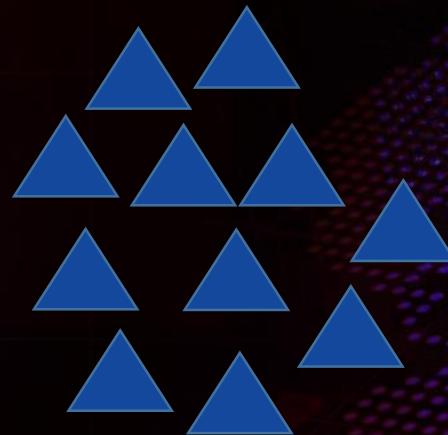


# PROCESSO DE AGRUPAMENTO DE DADOS



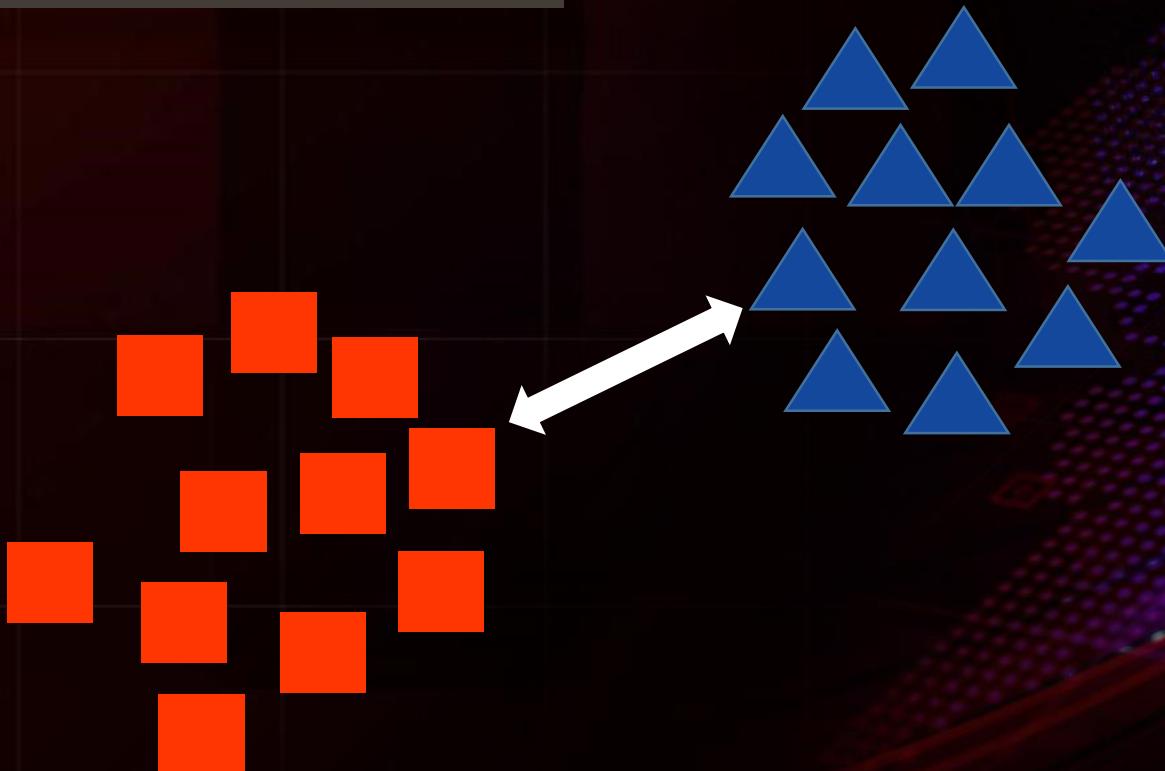
# MEDIDAS DE SIMILARIDADE

Similaridade: proximidade



# MEDIDAS DE SIMILARIDADE

Dissimilaridade: distância



# MEDIDAS DE SIMILARIDADE

Matriz de dados (X)

$$X = \begin{pmatrix} x_{11} & x_{1m} \\ x_{n1} & x_{nm} \end{pmatrix}$$

n objetos, m atributos

# MEDIDAS DE SIMILARIDADE

## Matriz de distância (D)

$$D = \begin{pmatrix} 0 & & & \\ d(2,1) & 0 & & \\ d(3,1) & d(3,2) & 0 & \\ & & & 0 \\ d(n,1) & d(n,2) & d(n,n-1) & 0 \end{pmatrix}$$

n objetos  
(linhas x colunas)

# MEDIDAS DE SIMILARIDADE

## ➤ Dados categóricos

- Valores normalizados
  - Intervalo  $[0,1]$  ou  $[0\%, 100\%]$

## ➤ Dados numéricos

- Valores dos próprios atributos

# DADOS CATEGÓRICOS BINÁRIOS

- **Distância de Hamming (H)**
  - Soma 1 a cada atributo diferente dos objetos comparados

Animal	Pelo	Penas	Mamífero	Aquático	Predador	H
Cobra do mar	0	0	0	1	1	-
Pato	0	1	0	1	0	2
Urso	1	0	1	0	1	3
Robalo	0	0	0	1	1	0

Fonte: Castro e Ferrari (2016)

Base: zoo

# DADOS CATEGÓRICOS BINÁRIOS

## ➤ Coeficientes de similaridade

- Função das diferenças entre  $l$  de  $m$  atributos
  - 1: presença, 0: ausência

		Atributo $l$ do objeto $i$		
		1	0	Total
Atributo $l$ do objeto $j$	1	a	b	$a + b$
	0	c	d	$c + d$
Total		$a + c$	$b + d$	$a + b + c + d$

# DADOS CATEGÓRICOS BINÁRIOS

## ➤ Coeficientes de similaridade

- Matching:  $s_{ij} = (a + d) / (a + b + c + d)$
- Jaccard:  $s_{ij} = (a) / (a + b + c)$
- Rogers & Tanimoto:  $s_{ij} = (a + d) / [(a + 2(b + c) + d)]$

Matching e Rogers & Tanimoto são simétricos, ou seja, consideram a ausência e presença dos atributos

# DADOS CATEGÓRICOS BINÁRIOS

## ➤ Coeficientes de similaridade

Animal	Pelo	Penas	Mamífero	Aquático	Predador	M	J	R&T
Cobra do mar	0	0	0	1	1	-	-	-
Pato	0	1	0	1	0	0,6	0,33	0,43
Urso	1	0	1	0	1	0,4	0,25	0,25
Robalo	0	0	0	1	1	1	1	1

# DADOS CATEGÓRICOS NOMINAIS

## ➤ Cálculo de dissimilaridade

- $d_{ij} = (m - M) / m$
- m: número total de atributos
- M: número de atributos iguais entre os objetos

*i e j, que estão sendo comparados*

# DADOS CATEGÓRICOS ORDINAIS

## ➤ Cálculo de dissimilaridade

- Os  $p$  valores distintos de um atributo são ordenados
- Cada valor recebe um ranking  $r$  entre 1 e  $p$
- Normalizar os valores de  $r$  entre 0 e 1 ( $z_{ij}$ )

$$z_{ij} = \frac{r_{ij} - 1}{p_{ij} - 1}$$

# DADOS CATEGÓRICOS ORDINAIS

Pernas ( $p_{ij}$ )	0	2	4	5	6	8
Ranking ( $r_{ij}$ )	1	2	3	4	5	6
Normalizado ( $z_{ij}$ )	0,0	0,2	0,4	0,6	0,8	1,0

Usar medidas de dados contínuos para a distância

# DADOS NUMÉRICOS CONTÍNUOS

## ➤ Medidas de proximidade ou distância

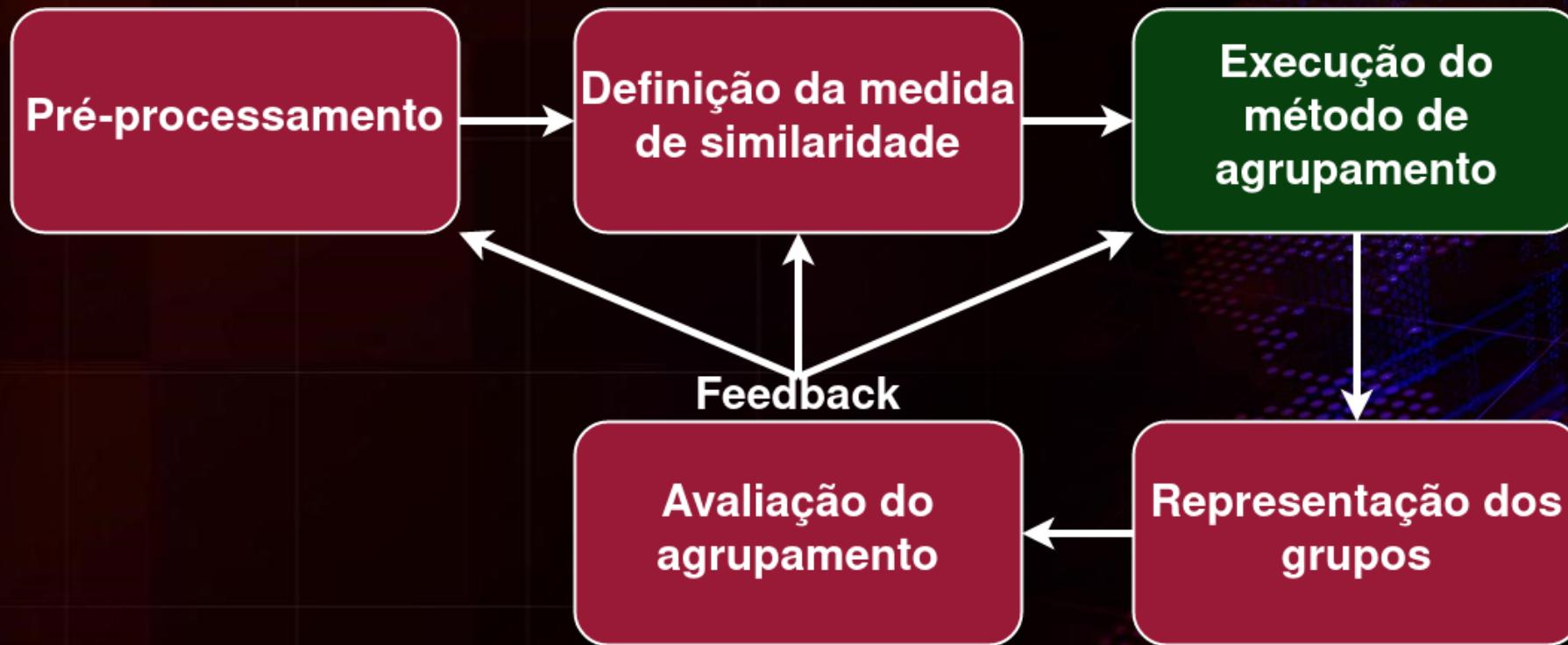
- Distância Euclidiana:

$$d_{ij} = \sqrt{\sum_{k=1}^m (x_{ik} - x_{jk})^2}$$

- Distância de Manhattan:

$$d_{ij} = \sum_{k=1}^m |x_{ik} - x_{jk}|$$

# PROCESSO DE AGRUPAMENTO DE DADOS



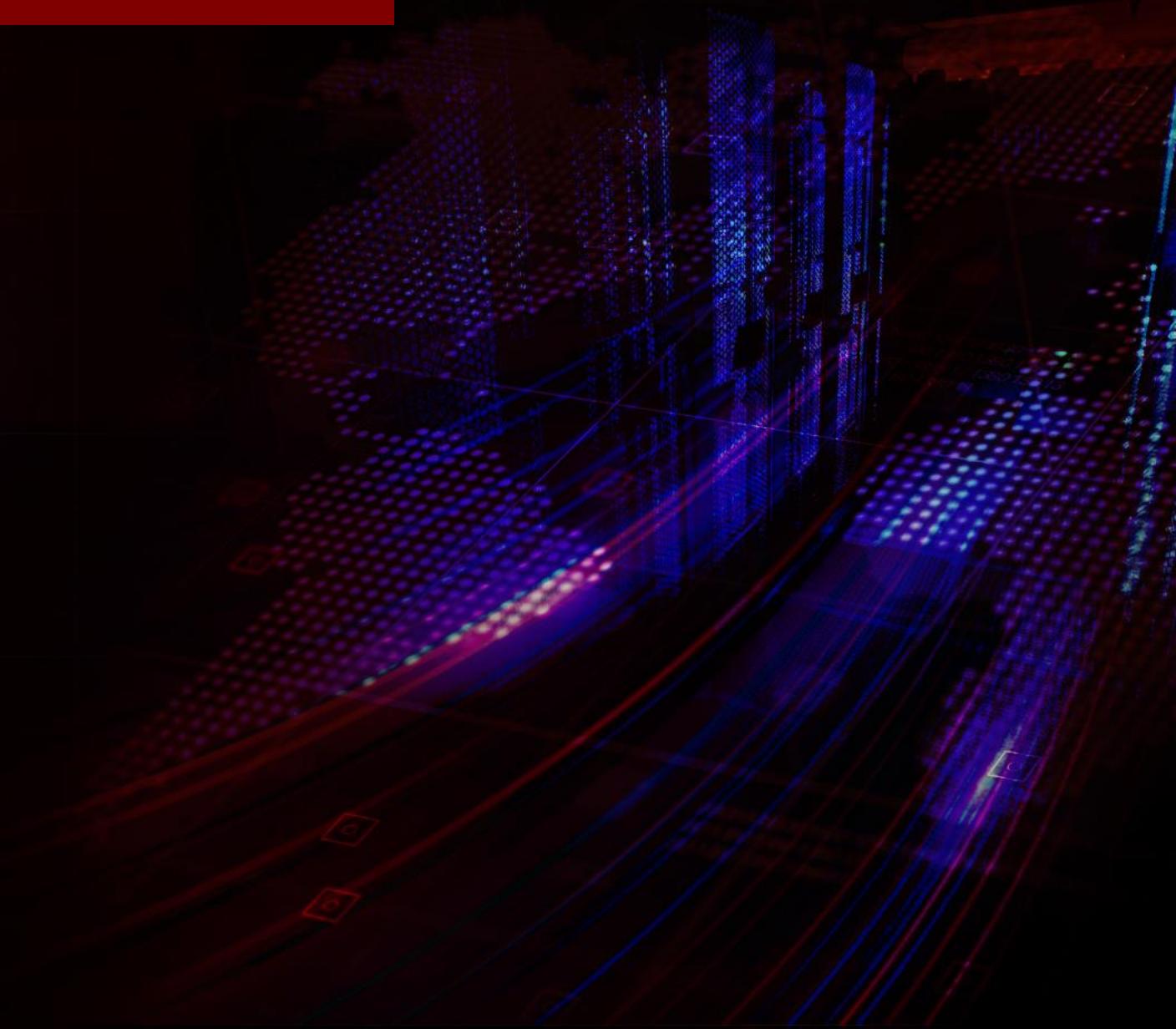
# MÉTODOS DE AGRUPAMENTO

## ➤ Hierárquicos

- Aglomerativos
- Divisivos

## ➤ Particionais

- Exclusivos
- Não exclusivos



# MÉTODOS DE AGRUPAMENTO

## ➤ Hierárquicos aglomerativos

- Cada objeto pertencendo a um grupo
- Objetos são agrupados até um critério de parada

## ➤ Hierárquicos divisivos

- Todos os objetos pertencem ao mesmo grupo
- Objetos são divididos até um critério de parada

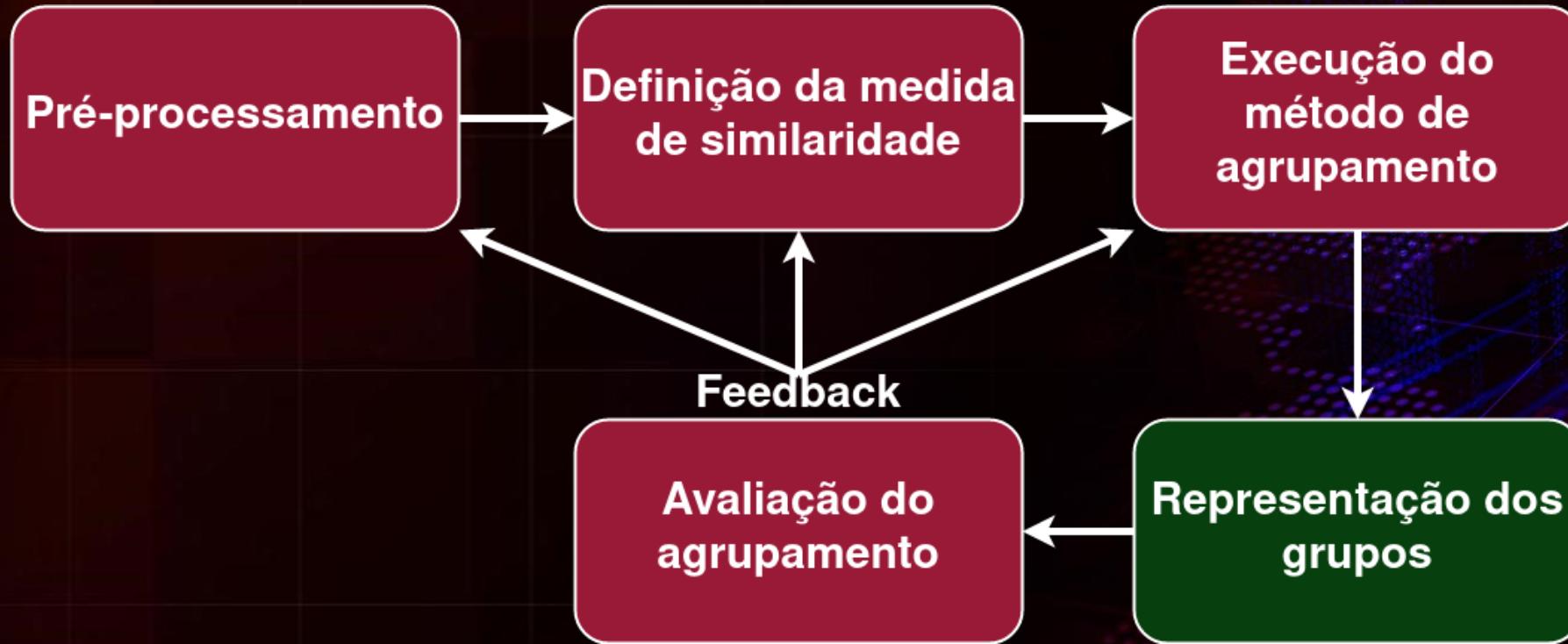
# MÉTODOS DE AGRUPAMENTO

- **Particionais exclusivos**
  - Cada objeto pertence a um único grupo
  
- **Particionais não exclusivos**
  - Cada objeto pode pertencer a mais de um grupo
  - Agrupamento *fuzzy*

# MÉTODOS DE AGRUPAMENTO

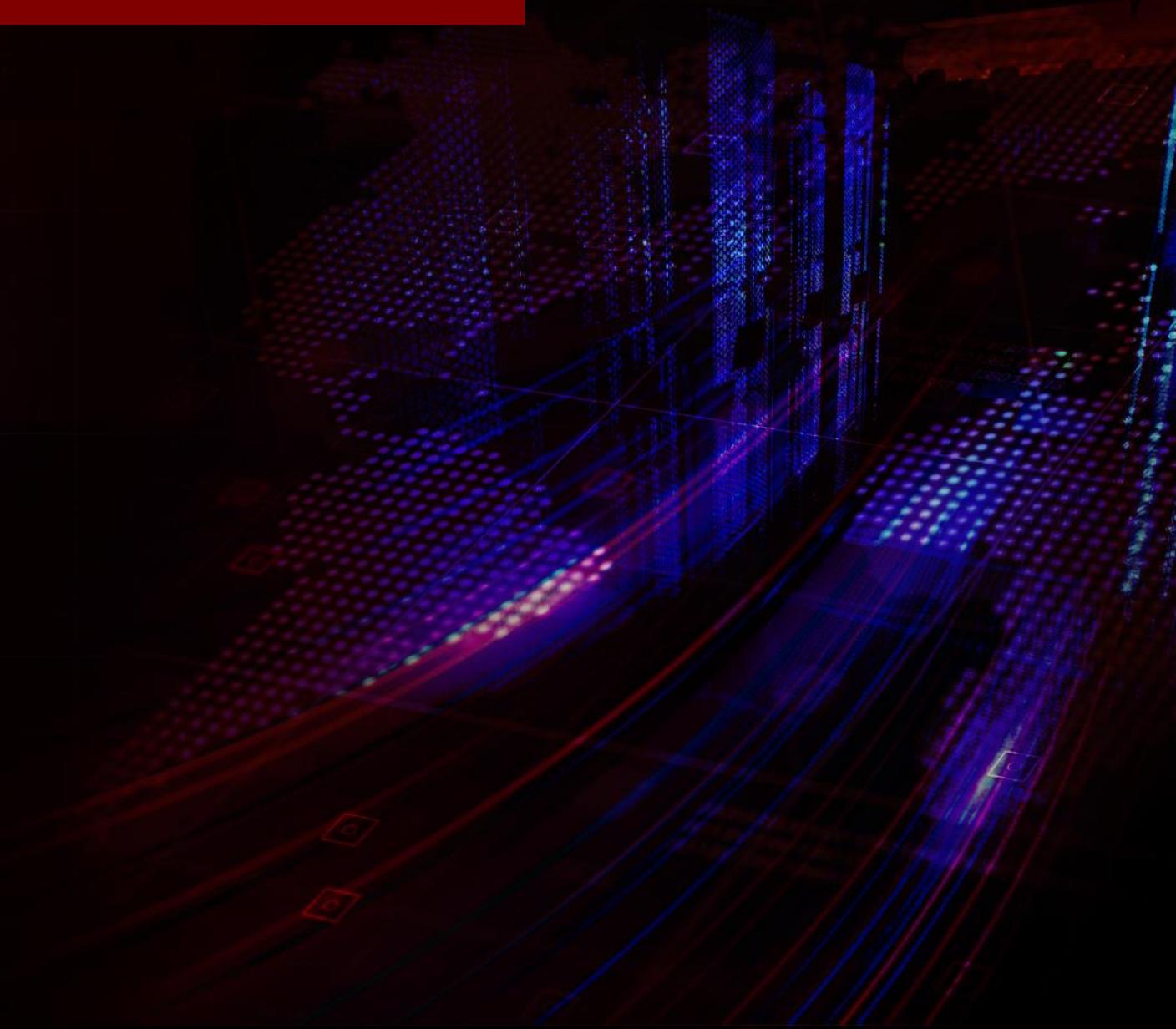
- **Monotéticos ou políticos**
  - Número de atributos usados para calcular a distância
- **Hard ou fuzzy**
  - Pertencimento integral ou parcial de um objeto aos grupos
- **Determinístico ou estocástico**
  - Resultado do agrupamento

# PROCESSO DE AGRUPAMENTO DE DADOS



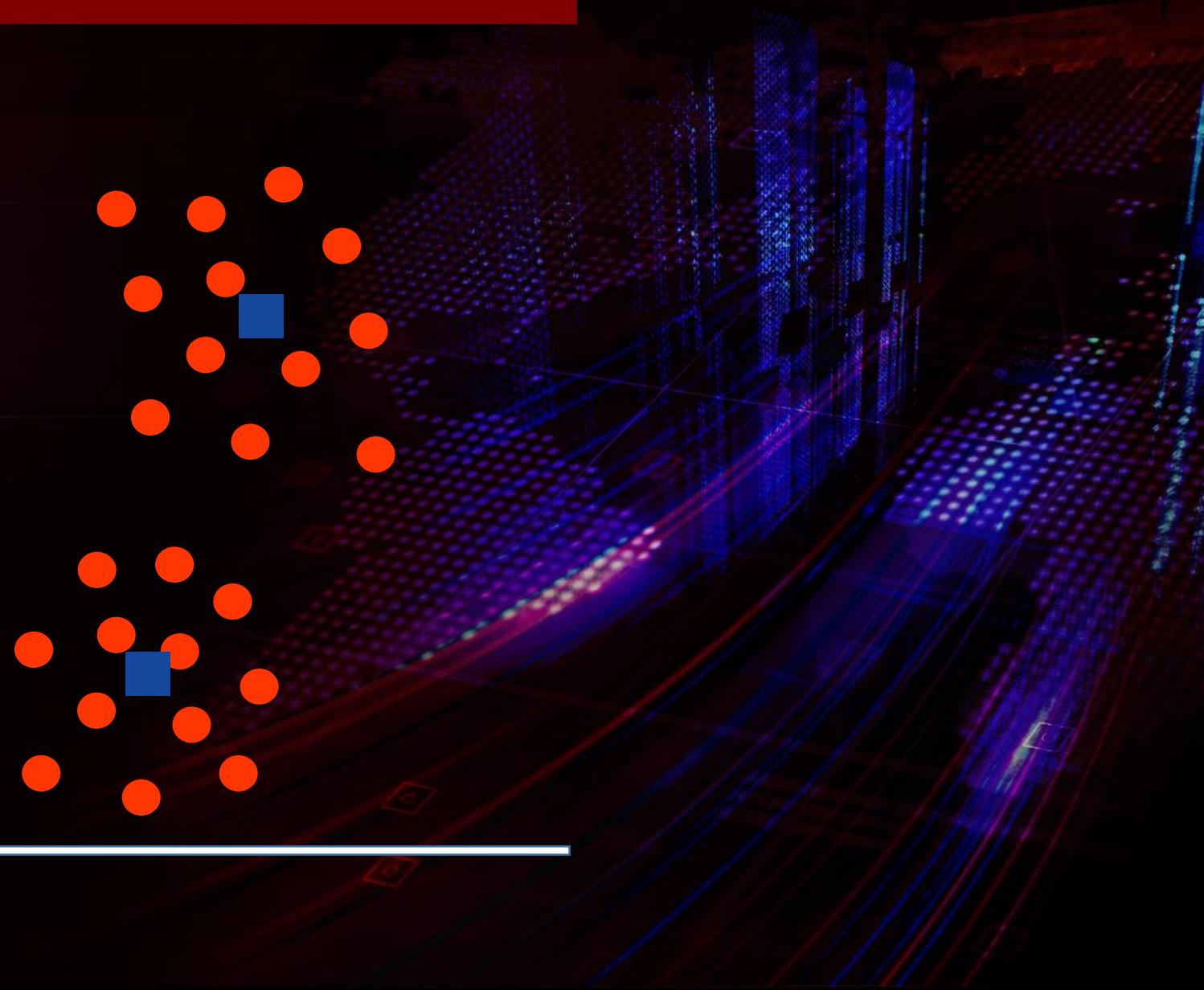
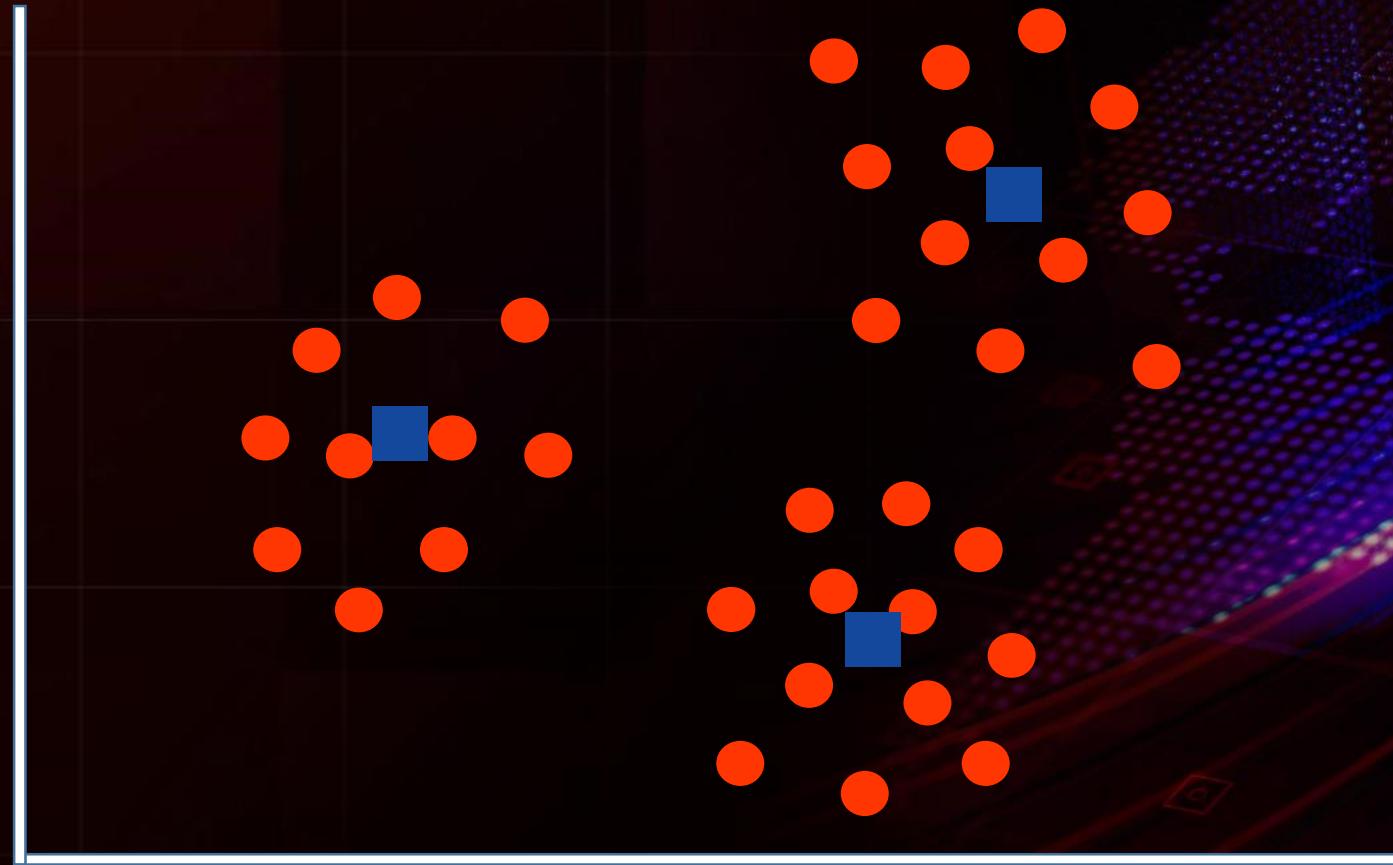
# REPRESENTAÇÃO DOS GRUPOS

- Protótipos
- Estruturas em grafo
- Estruturas em árvore
- Rotulação



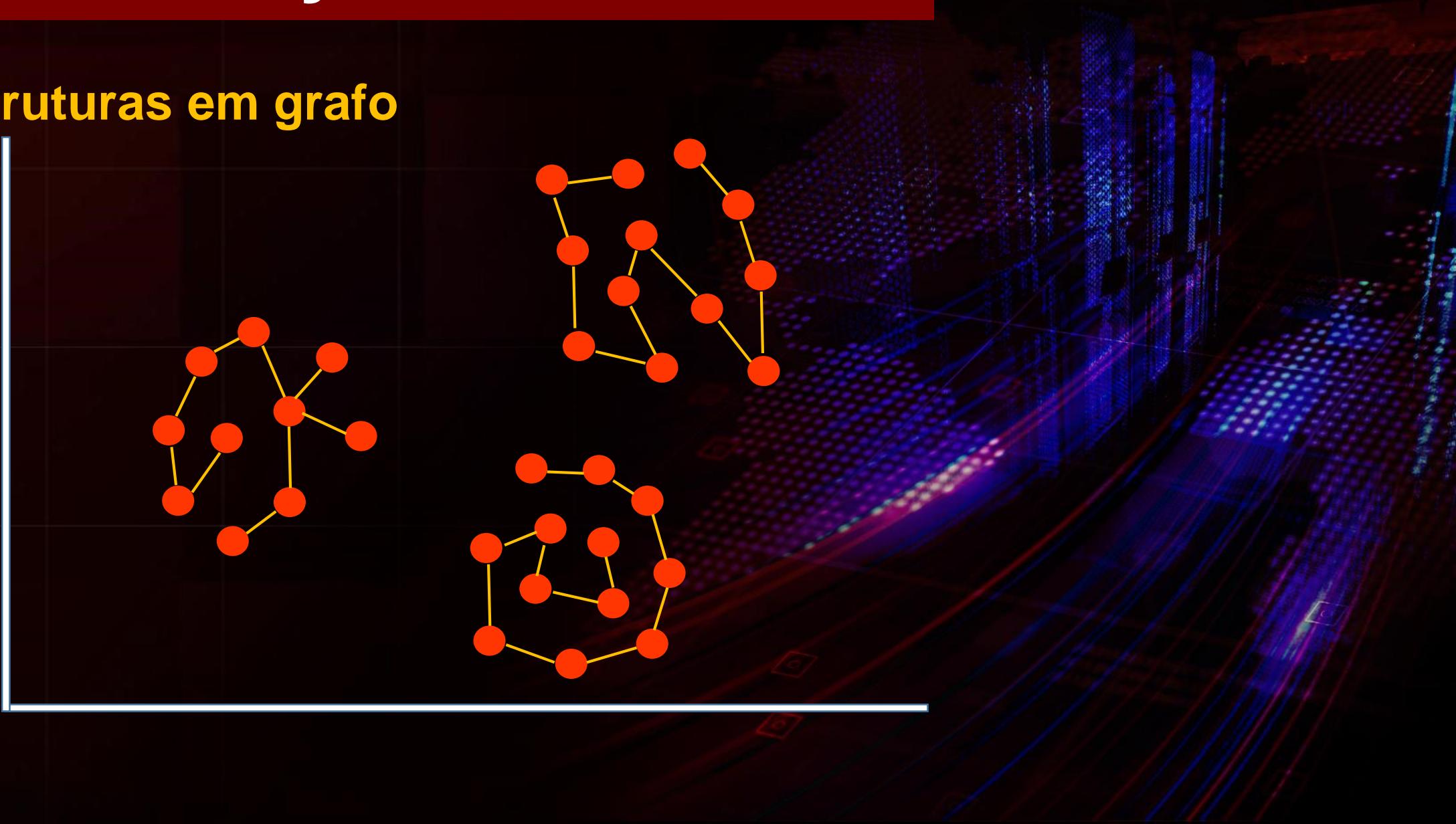
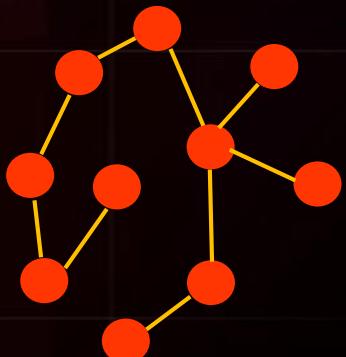
# REPRESENTAÇÃO DOS GRUPOS

## ➤ Protótipos



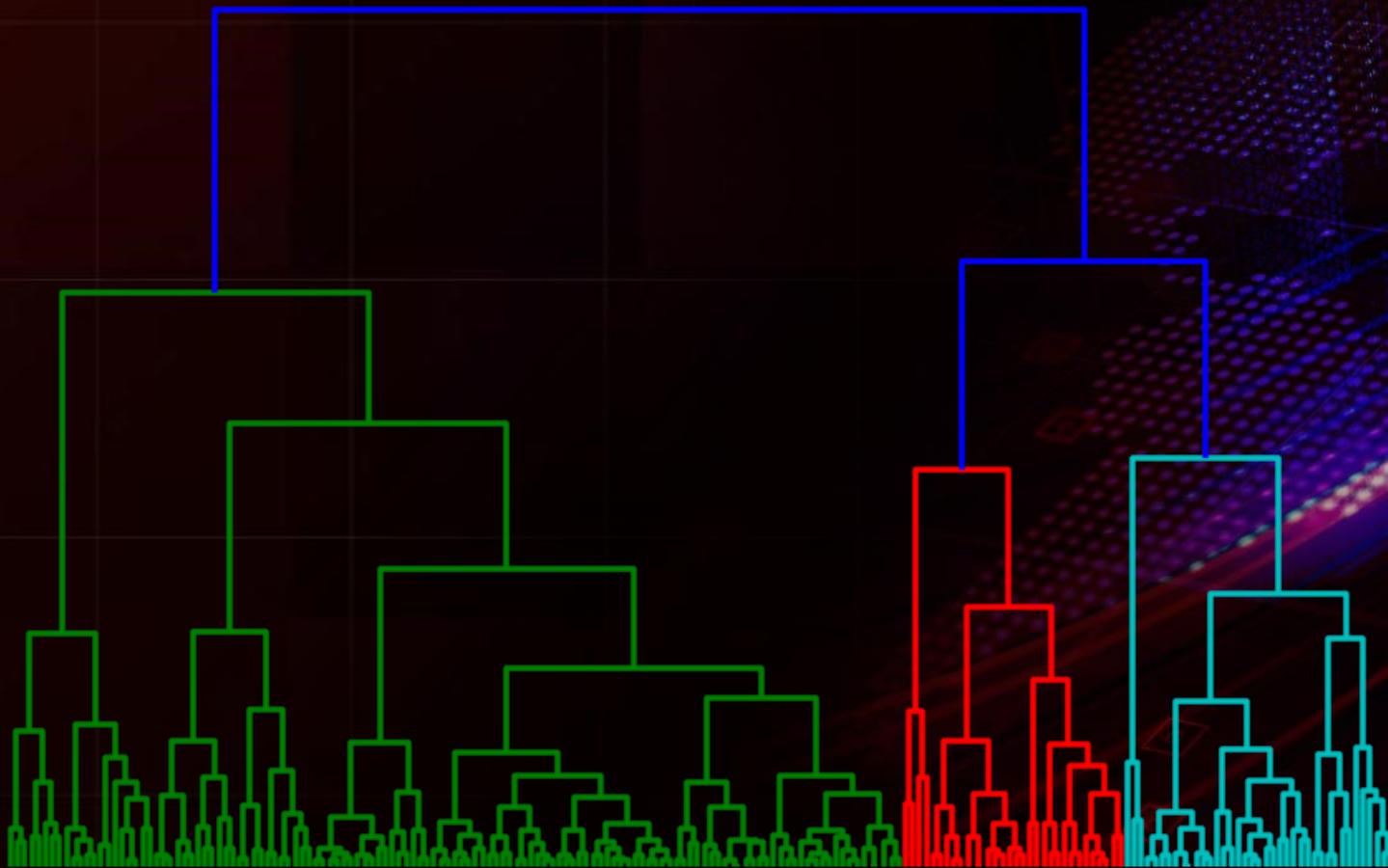
# REPRESENTAÇÃO DOS GRUPOS

## ➤ Estruturas em grafo



# REPRESENTAÇÃO DOS GRUPOS

## ➤ Estruturas em árvore

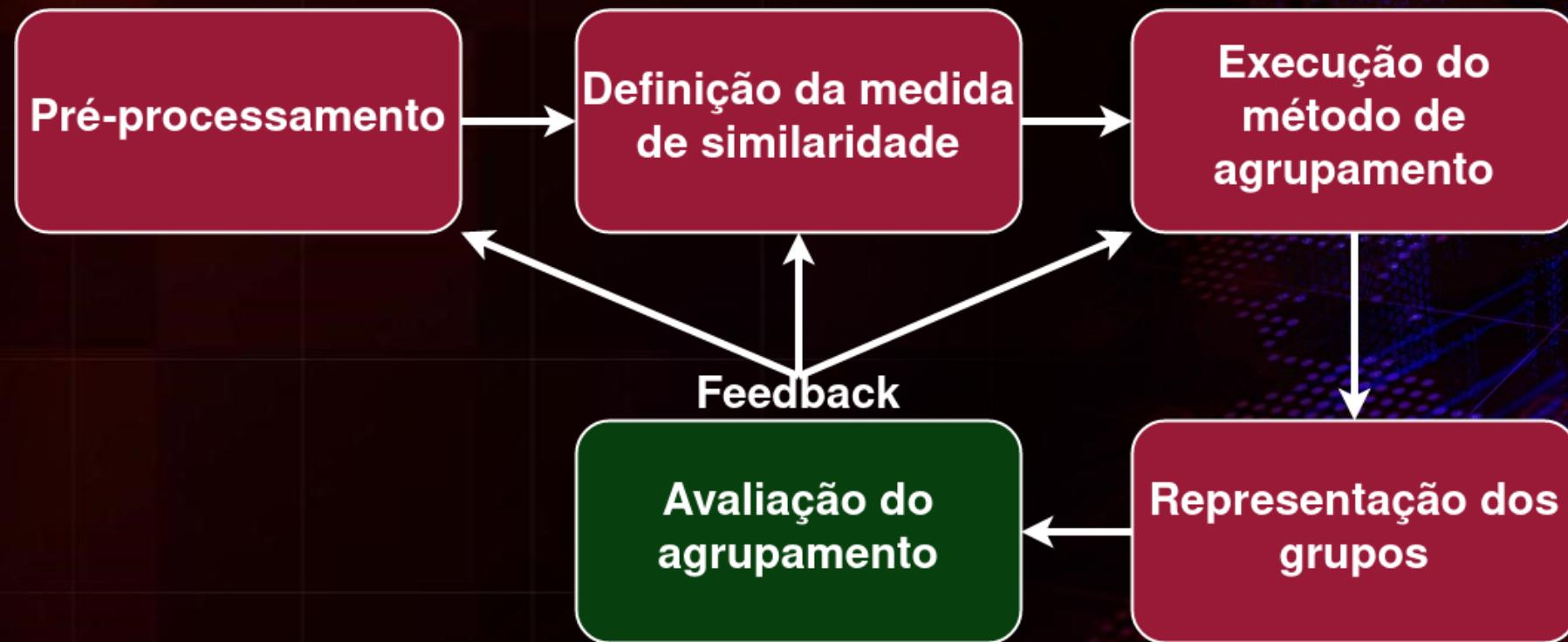


# REPRESENTAÇÃO DOS GRUPOS

## ➤ Rotulação

- Adicionar rótulos aos dados
- Atributo alvo

# PROCESSO DE AGRUPAMENTO DE DADOS



# AVALIAÇÃO DE AGRUPAMENTO

## ➤ Critérios para avaliação

- Compactação (intragrupo)
  - Objetos de grupo devem estar próximos
- Separação (intergrupo)
  - Grupos devem estar distantes

# AVALIAÇÃO DE AGRUPAMENTO

## ➤ Tipos de medidas de avaliação

- Internas
  - Avaliam as informações dos objetos
  - Distâncias intragruo e intergru
- Externas
  - Comparam com um agrupamento ideal
  - Dependem de conhecimento prévio

# MEDIDAS DE AVALIAÇÃO INTERNA

## ➤ Índice de Dunn

- Intra / intergrupos
- Intervalo:  $[0; \infty]$
- Quanto maior, melhor o agrupamento
- Equilíbrio entre compactação interna e separação externa

# MEDIDAS DE AVALIAÇÃO INTERNA

## ➤ Índice de Bezdek-Pal

- Separação entre grupos é mais importante
- Valor médio da medida intergrupos
- Intervalo:  $[0; \infty]$
- Quanto maior, melhor o agrupamento

# MEDIDAS DE AVALIAÇÃO EXTERNA

## ➤ Entropia

- Indica a homogeneidade das classes dos objetos nos grupos
- Baixa entropia: grupos mais homogêneos

## ➤ Pureza

- Proporção da classe dominante em relação ao tamanho do grupo
- Quanto maior o valor, maior a pureza

# REFERÊNCIAS

**Introdução à Mineração de Dados: Conceitos Básicos, Algoritmos e Aplicações: Cap. 4: Análise de grupos.**  
**Leandro Nunes de Castro e Daniel Gomes Ferrari. Editora Saraiva, 2016.**