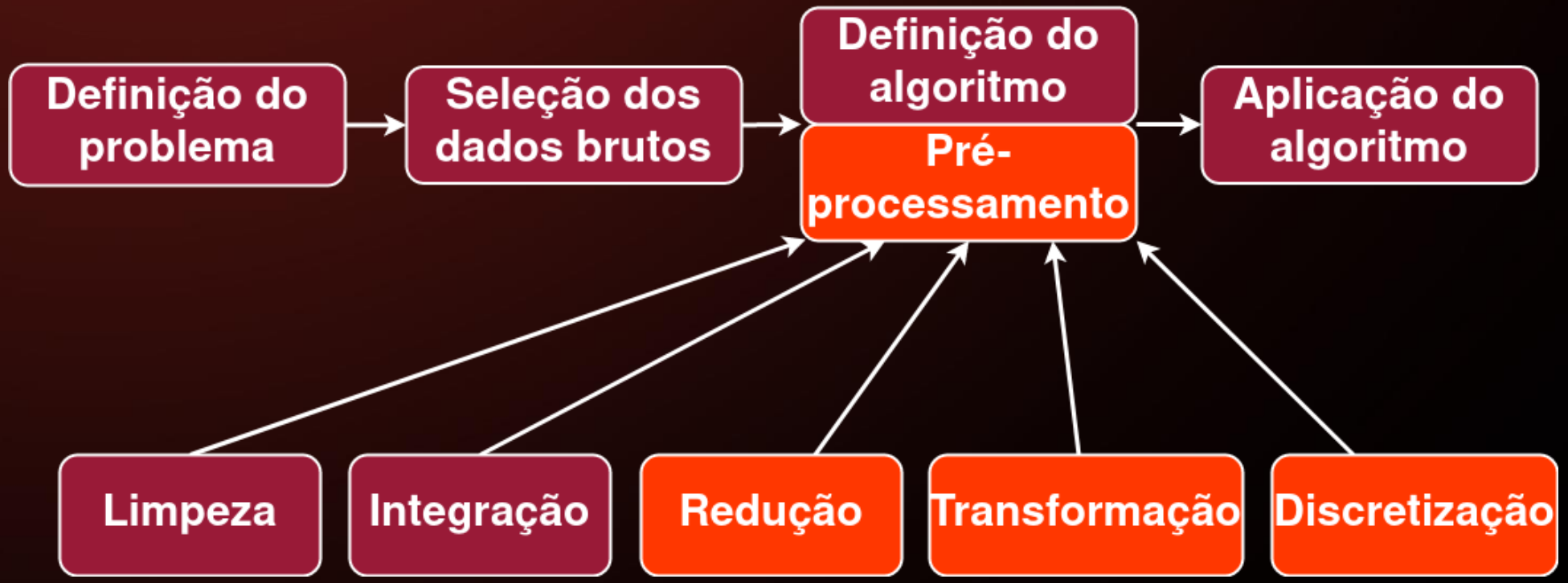
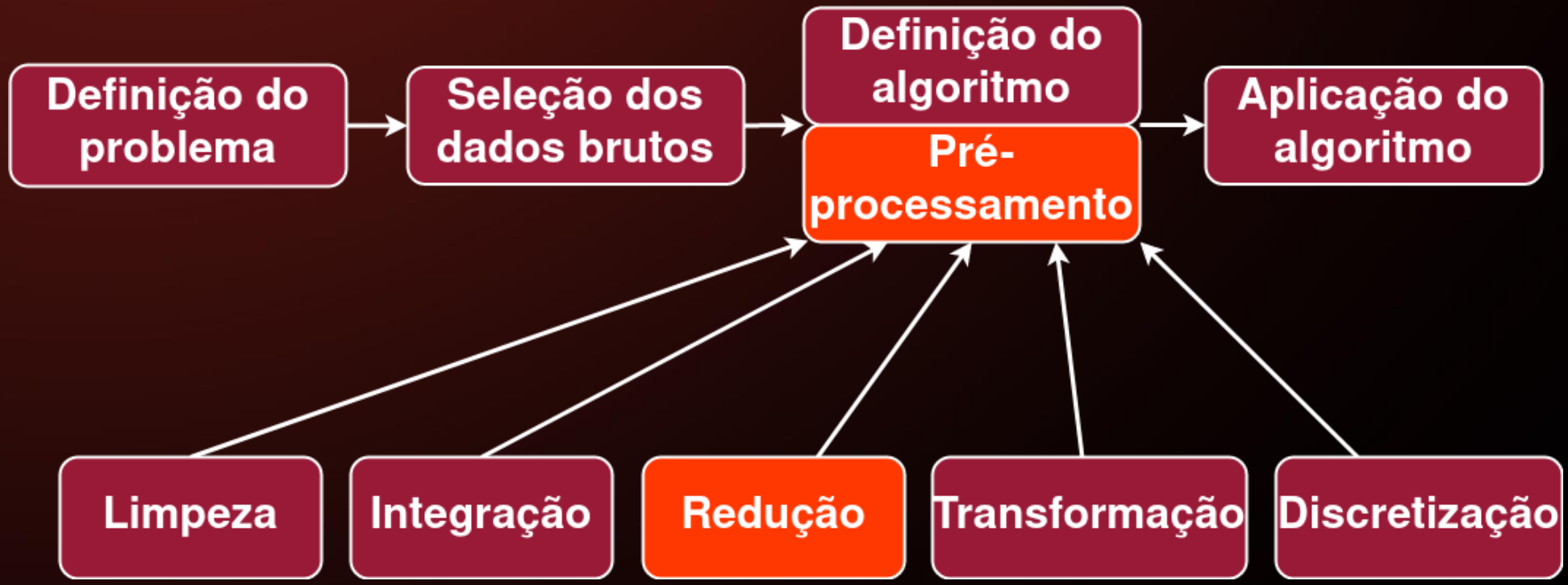
MINERAÇÃO DE DADOS

Redução e transformação dos dados

PROCESSO DE PREPARAÇÃO DOS DADOS




Definição do problema
Seleção dos dados brutos
Definição do algoritmo
Pré-processamento
Aplicação do algoritmo
Limpeza
Integração
Redução
Transformação
Discretização

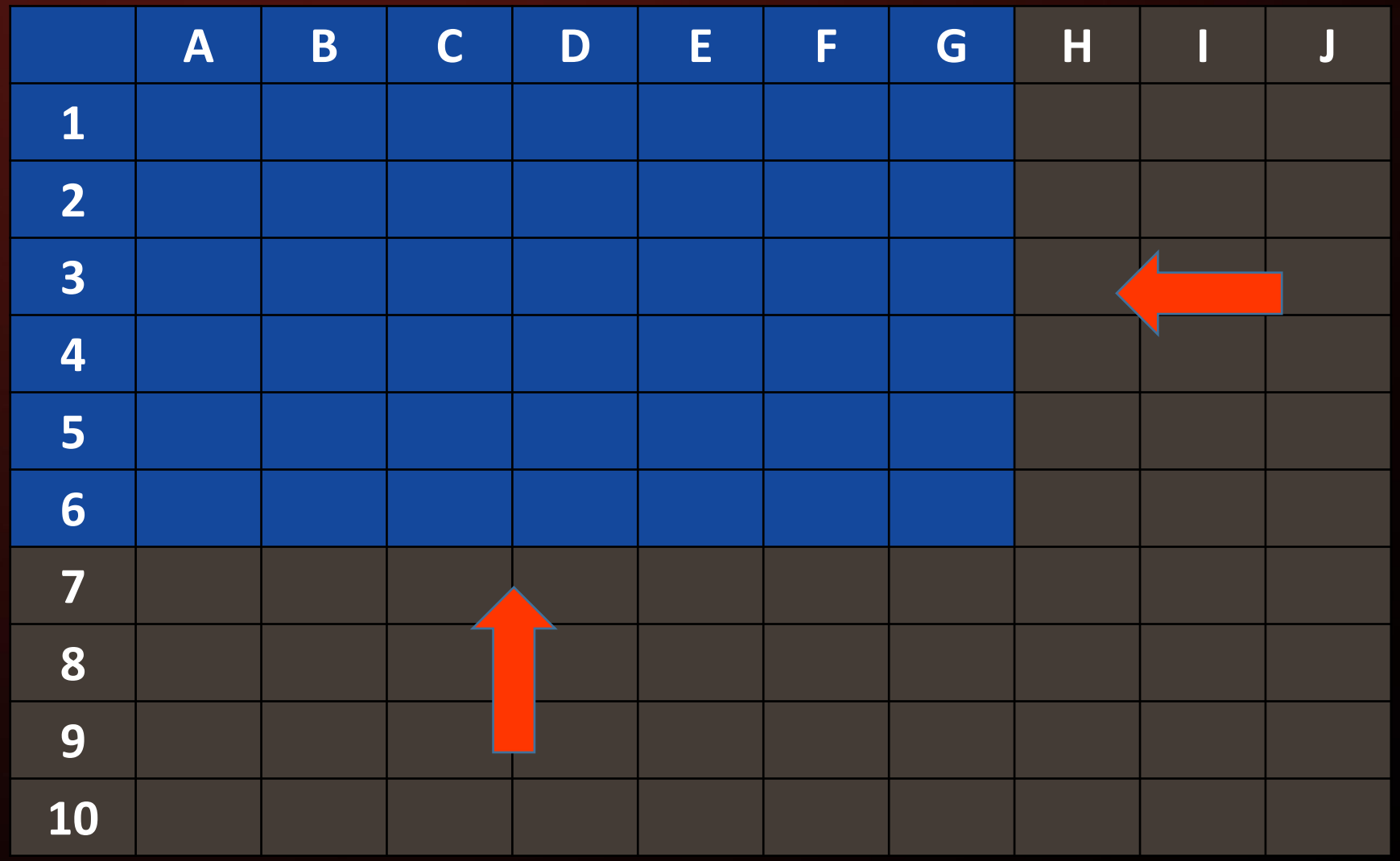
REDUÇÃO DOS DADOS

- Grandes quantidades de objetos e atributos:
  - Impactos negativos nos resultados de algoritmos
  - Alta complexidade de processamento
  - Geração de modelos complexos

REDUÇÃO DOS DADOS

- Exemplo: Algoritmo Apriori
- 2^n – 1 possíveis combinações de *n* atributos
  - n = 3:
    - 7 possíveis combinações
  - n = 20:
    - ~ 1 milhão de possíveis combinações

REDUÇÃO DOS DADOS



MÉTODOS DE REDUÇÃO DOS DADOS

- Seleção de atributos
- Compressão de atributos
- Redução do número de dados
- Discretização

SELEÇÃO DE ATRIBUTOS

- Remover atributos sem relevância ou redundantes
- Exemplos:
  - Identificador de registro (ID)
  - Nome x usuário
  - Data de nascimento x idade

COMPRESSÃO DE ATRIBUTOS

- Compactação dos dados
  - Ex: reduzir precisão das coordenadas de GPS
- Codificação / transformação de atributos
  - Análise de componentes principais (PCA)

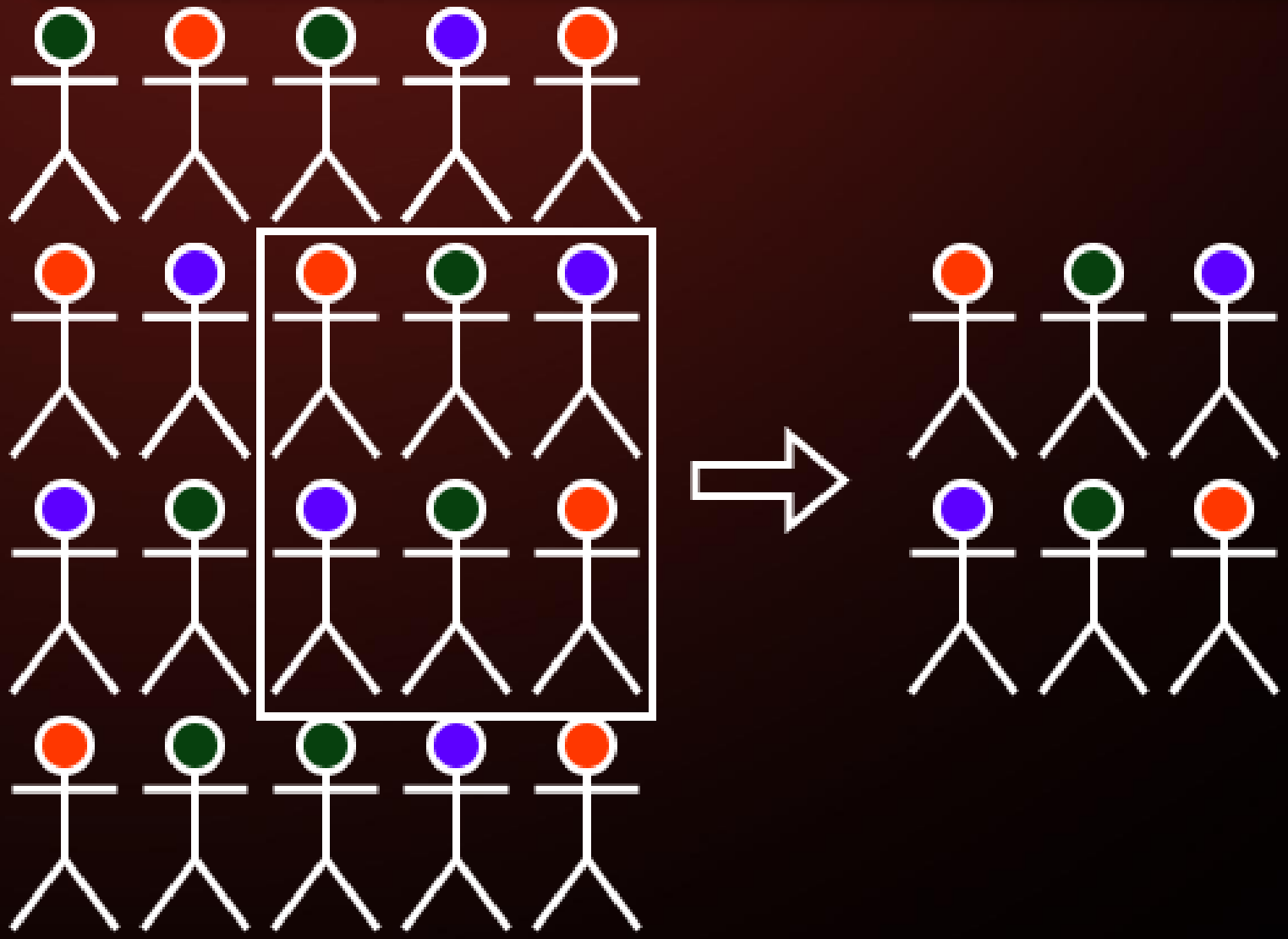
COMPRESSÃO DE ATRIBUTOS - MÉTODOS

- Análise de componentes principais (PCA)
  - Correlação entre atributos
    - Ex: Renda x Valor do imóvel ⟹ componente
  - Número de componentes <= número de atributos
  - Ordenados por variância:
    - Preservar as características dos dados que contribuem mais para a sua variância

REDUÇÃO DO NÚMERO DE DADOS

- Selecionar objetos da base
- Aproximar os objetos de modelos
- Métodos:
  - Amostragem
  - Modelos de aproximação

REDUÇÃO - AMOSTRAGEM



REDUÇÃO - AMOSTRAGEM

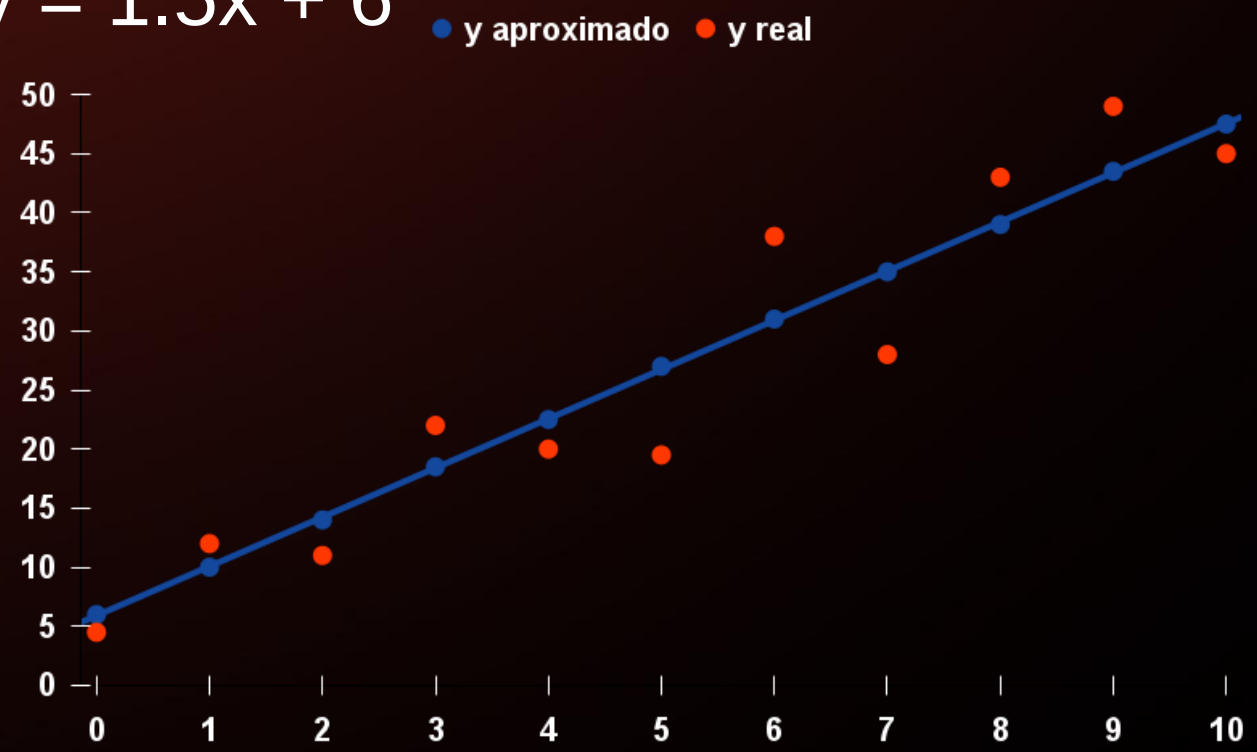
- Métodos de amostragem:
  - Aleatória sem substituição
  - Aleatória com substituição
  - Sistemática
  - Por grupo
  - Estratificada

REDUÇÃO - MODELOS DE APROXIMAÇÃO

- Representar ou ajustar os dados
- Substituem os valores dos atributos

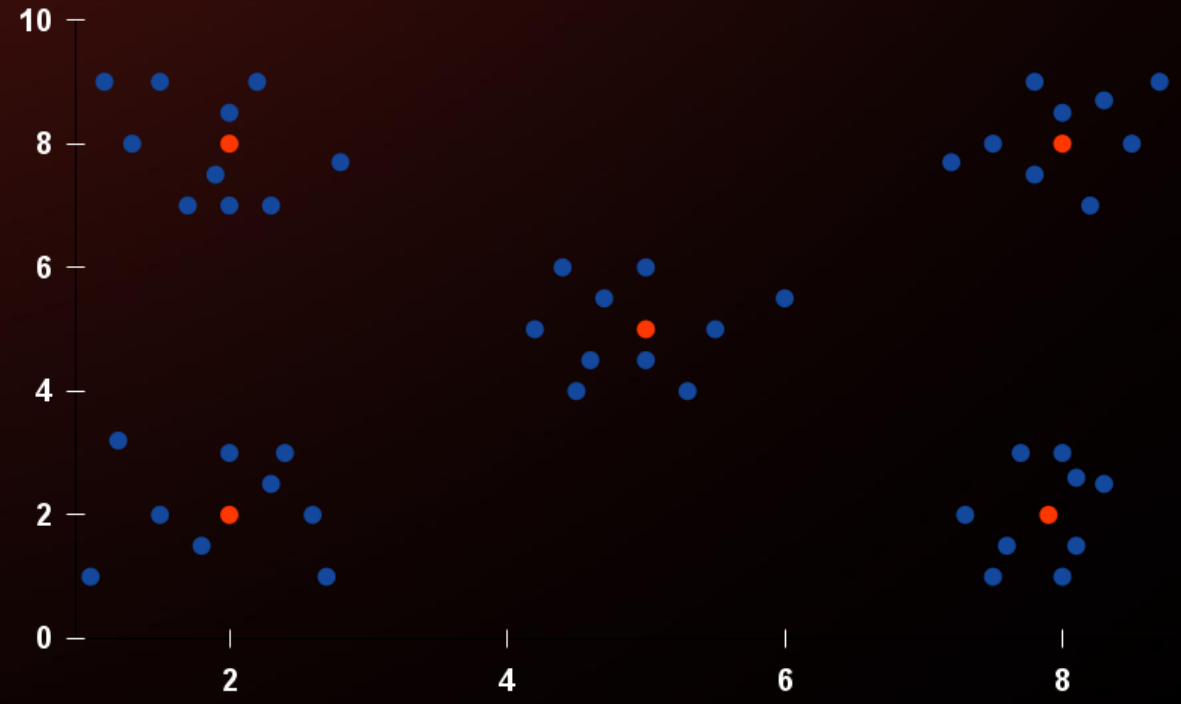
REDUÇÃO - MODELOS DE APROXIMAÇÃO

- Paramétrico:
  - Função de aproximação:
    - Ex: y = 1.5x + 6



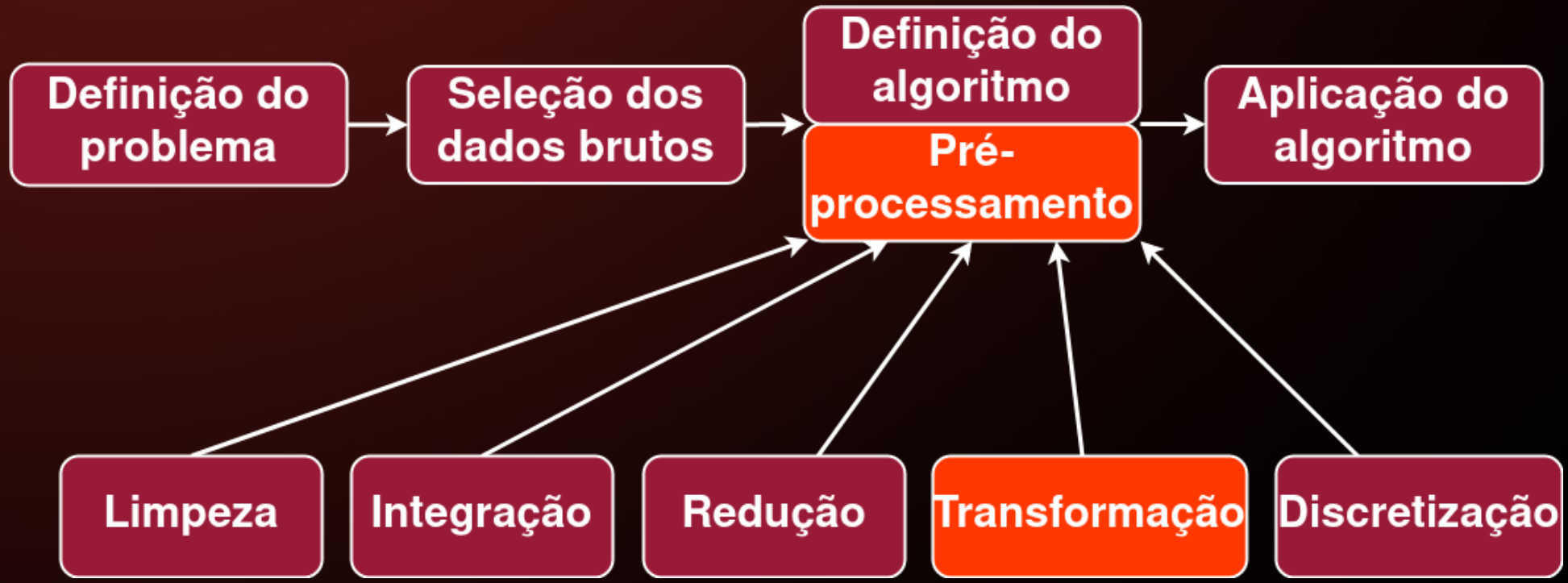
REDUÇÃO - MODELOS DE APROXIMAÇÃO

- Não paramétrico:
  - Sem estrutura ou forma definida
    - Ex: k-médias



DISCRETIZAÇÃO

- A discretização também é uma forma de redução de dados
- Veremos mais a seguir


Definição do problema
Seleção dos dados brutos
Definição do algoritmo
Pré-processamento
Aplicação do algoritmo
Limpeza
Integração
Redução
Transformação
Discretização

TRANSFORMAÇÃO DOS DADOS

- Bases brutas e integradas:
  - Inconsistências, ruídos
  - Ex:
    - Gênero: Feminino | feminino | F | 0 | Fem
    - Distância: 1 milha | 1,6 quilômetro | 1600 metros
- Atributos não uniformes:
  - Numéricos, categóricos, ...

TRANSFORMAÇÃO DOS DADOS

- Modificar ou consolidar os dados em formas apropriadas para uso nos processos de mineração
- Tipos de transformações:
  - Padronização
  - Normalização

PADRONIZAÇÃO

- Diferenças em unidades e escalas
- Capitalização
- Caracteres especiais
- Formatação
- Conversão de unidades

NORMALIZAÇÃO

- Adequar os dados para uso em algoritmos de mineração
- Tipos de normalização:
  - Max-Min
  - Escore-z
  - Escalonamento decimal
  - Range interquartil (RI)

NORMALIZAÇÃO MAX-MIN

- Mapeia o atributo dentro da faixa de valores definida
- $a' = \frac{a - min_a}{max_a - min_a} (novo\_max_a - novo\_min_a) + novo\_min_a$

- Exemplo:
  - [0,1]

NORMALIZAÇÃO ESCORE-Z

- Baseada na média e desvio padrão do atributo
  - Evitar influência de anomalias

- $a' = \frac{a - \bar{a}}{\sigma_a}$

NORMALIZAÇÃO PELO ESCALONAMENTO DECIMAL

- Mover a casa decimal
- Número de casas depende do valor máximo do atributo
- $a' = \frac{a}{10^j}$, onde j é o menor inteiro tal que max(|a'|) < 1
  - Ex: max(a) = 1000, j = 4, max(|a'|) = 0,1

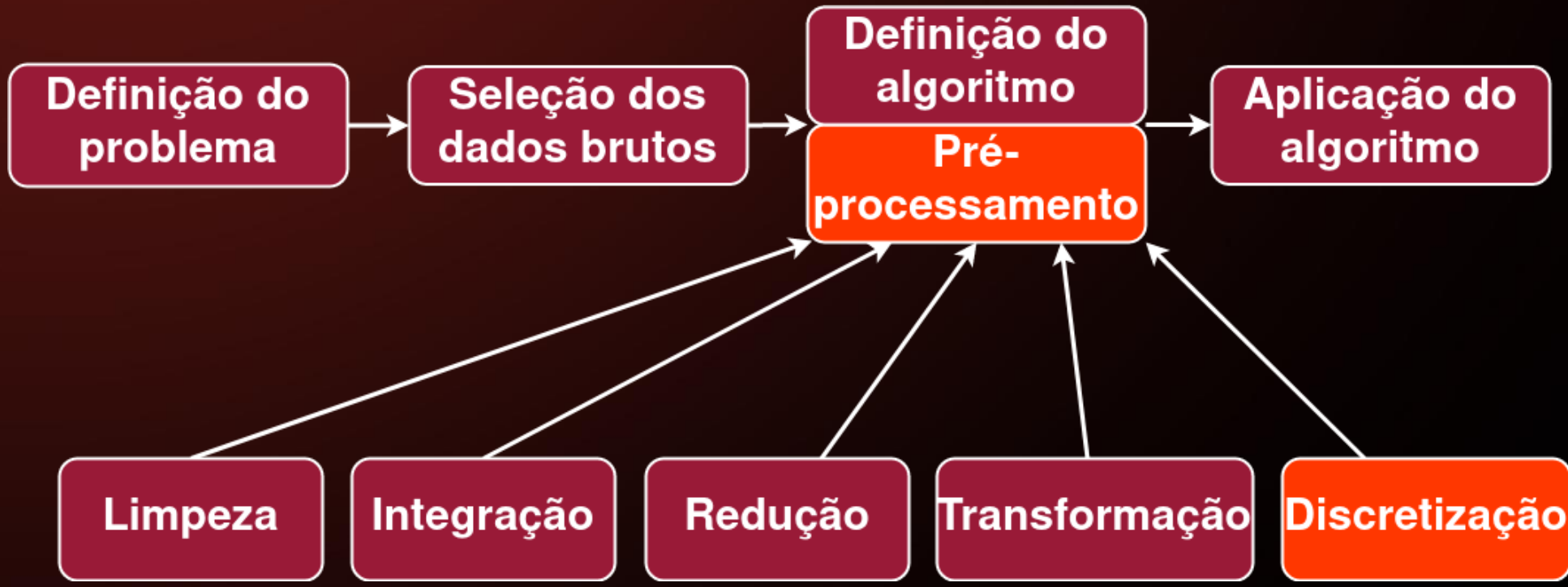
NORMALIZAÇÃO PELO RANGE INTERQUARTIL

- Divide os valores dos atributos em quartis

- $a' = \frac{a - Q_2}{Q_3 - Q_1}$, $RI = Q_3 - Q_1$

NORMALIZAÇÃO

Valor original	Max-Min	Escore-z	Escalonamento decimal	Range interquartil
67	0,85	0,73	0,67	0,40
43	0,33	-0,92	0,43	-0,80
58	0,65	0,11	0,58	-0,05
28	0,00	-1,96	0,28	-1,55
74	1,00	1,21	0,74	0,75
65	0,80	0,59	0,65	0,30


Definição do problema
Seleção dos dados brutos
Definição do algoritmo
Pré-processamento
Aplicação do algoritmo
Limpeza
Integração
Redução
Transformação
Discretização

DISCRETIZAÇÃO

- Transformar os valores dos atributos para dados categóricos
  - Valores numéricos → intervalo

1,33	7,68	5,21	3,23	6,78	9,75	4,58	8,65	7,24	6,35	1,25	8,36	6,25	2,21	4,56

↓

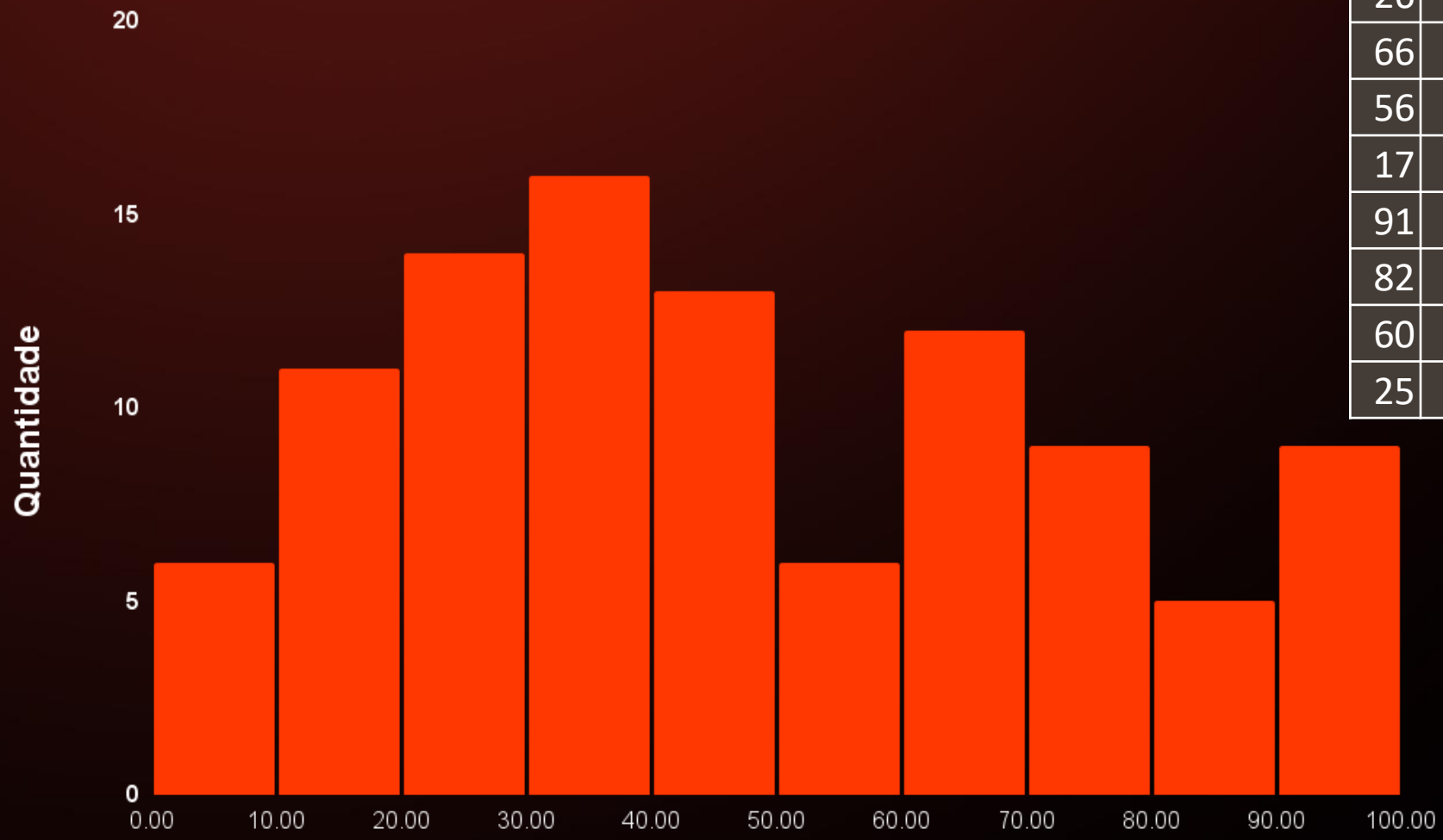
1-5	6-10

DISCRETIZAÇÃO

- **Métodos:**
  - **Intervalos predeterminados**
  - **Encaixotamento / Análise de histograma**
  - **Agrupamento**
  - **Entropia**

DISCRETIZAÇÃO - HISTOGRAMA

4	13	77	71	28	81	9	26	55	6
88	55	98	93	24	78	21	15	5	25
26	35	15	6	13	59	38	66	29	57
66	41	12	88	71	17	12	13	13	45
56	26	4	0	9	56	91	82	48	99
17	16	6	75	32	41	69	38	96	13
91	87	33	6	81	70	74	39	66	11
82	3	55	29	61	29	53	63	18	91
60	63	61	94	37	62	51	74	94	5
25	0	53	34	69	31	47	47	69	68



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.