

# Introdução à Ciência de Dados e Aprendizado de Máquina

## Exploração de Dados



# Robson Fernandes

## Acadêmico

Mestrando em Matemática, Estatística e Computação Aplicadas (Data Science & Machine Learning) - USP  
Especialização em Reconhecimento de Padrões e Análise de Imagens - UNICAMP  
Pós-Graduado em Arquitetura de Software Distribuído - PUC-MG  
MBA em Engenharia de Software Orientada a Serviços – SOA – METROCOMP  
Certificado – JavaScript e HTML5 Developer – W3C INTERNACIONAL  
Autor do Livro Gestão da Tecnologia da Informação: Teoria e Prática

## Profissional

Cientista de Dados Sênior – Finch Soluções  
Docente Pós-Graduação - MBA em Data Science & Machine Learning - UNIP  
Docente Pós-Graduação em Engenharia de Software - USC  
Docente Graduação em Ciência da Computação - UNIP

## Site

<http://robsonfernandes.net>

## e-mail

[robson.fernandes@usp.br](mailto:robson.fernandes@usp.br) / [robs.fernandes@outlook.com](mailto:robs.fernandes@outlook.com)

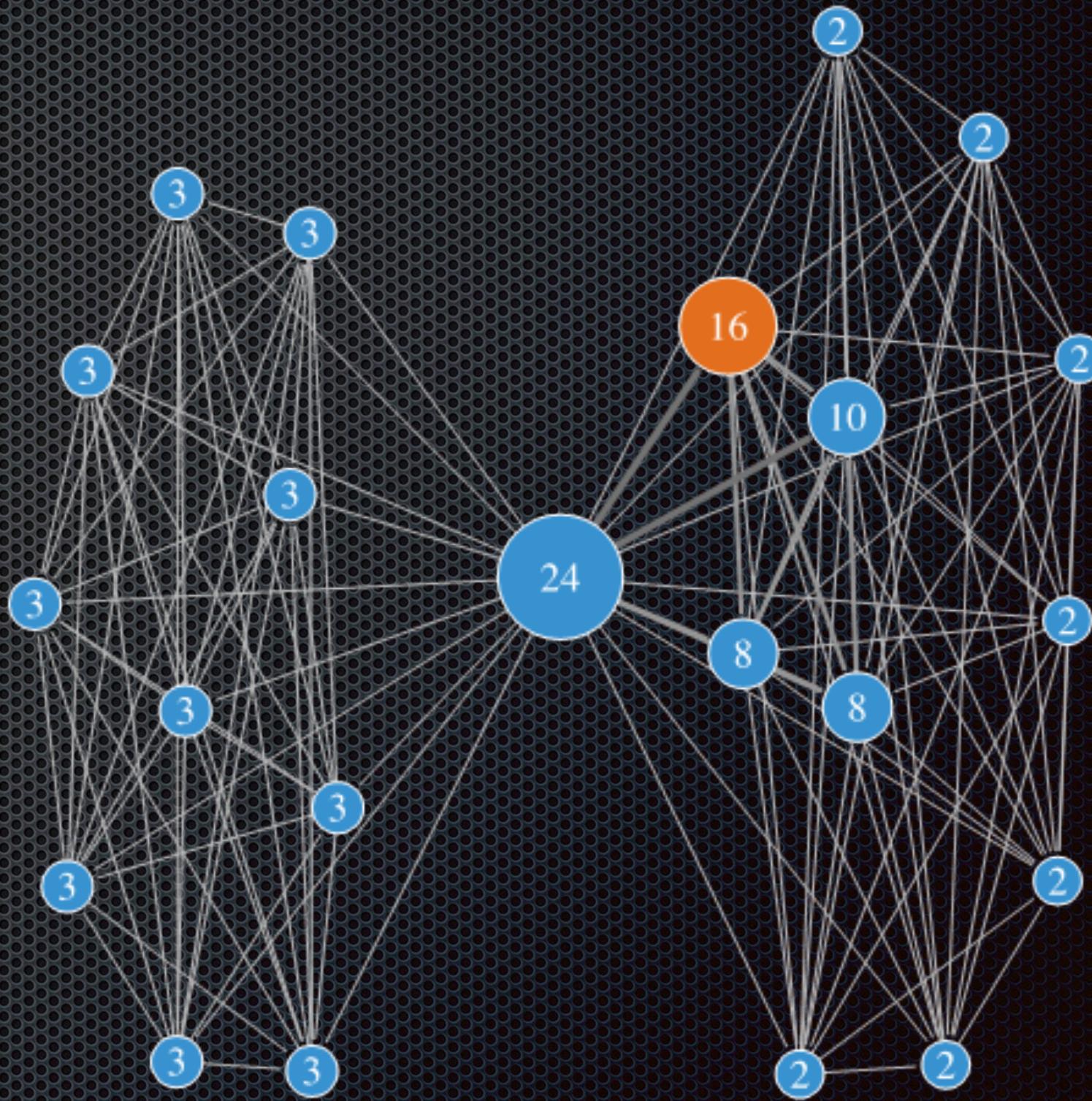
# Exploração de Dados

- Exploração preliminar dos dados facilita entendimento de suas características
- Principais motivações:
  - Ajudar a selecionar a melhor técnica para pré-processamento e/ou modelagem
- Ferramentas
  - Estatística descritiva
  - Visualização



# Conjunto de Dados

- Estruturados
    - Mais facilmente analisados por técnicas de Mineração de Dados (MD)
    - Ex.: Planilhas e tabelas atributo-valor
  - Não estruturados
    - Mais facilmente analisados por seres humanos.
    - Para MD, são geralmente convertidos em dados estruturados.
    - Ex.: Sequência de DNA, conteúdo de página na web, e-mails, vídeos, ...

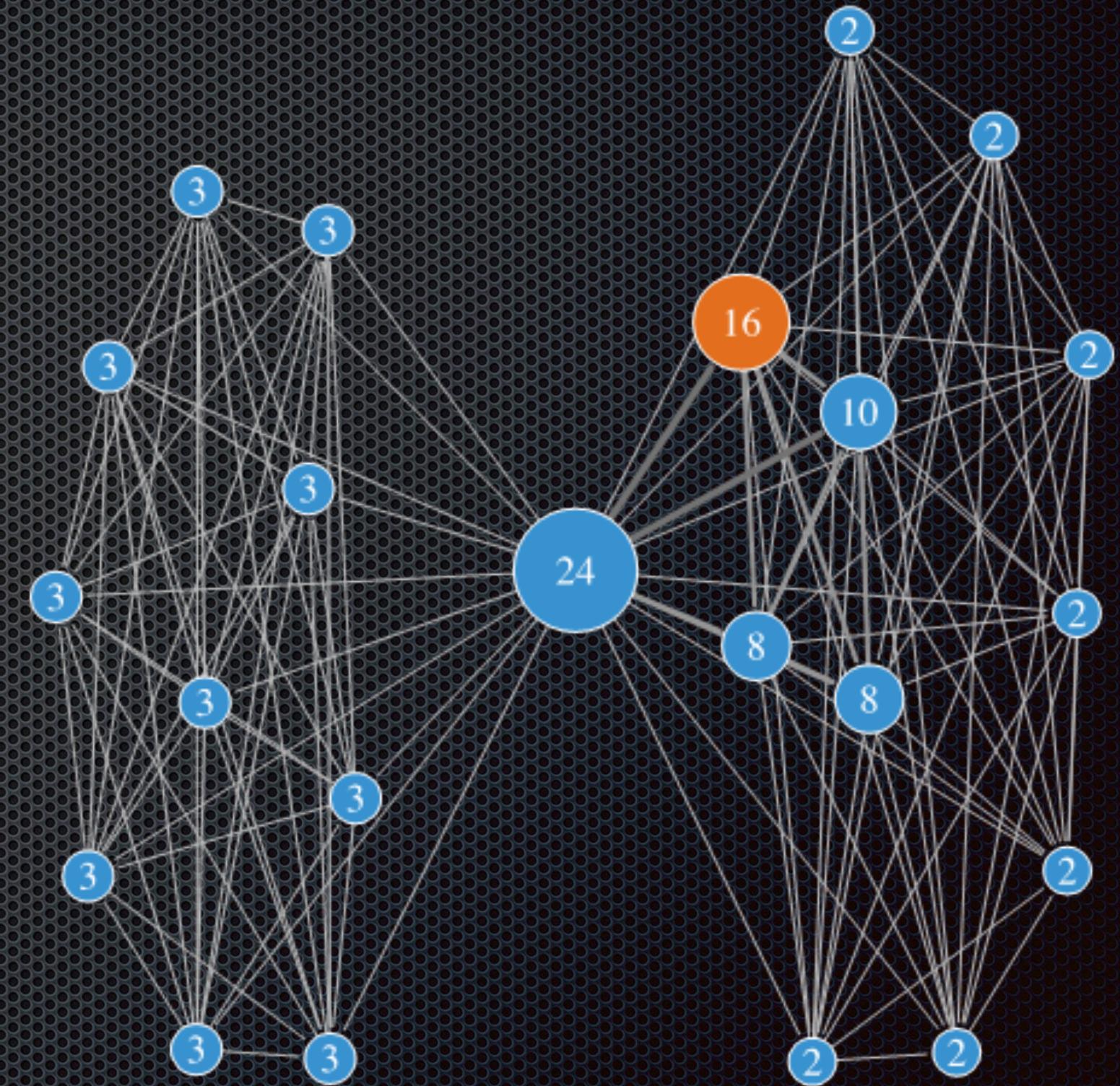


# Conjunto de Dados Estruturados

Exemplos de Objetos (Instâncias)	Atributos de entrada (preditivos)						Atributo Alvo
	Nome	Temperatura	Idade	Peso	Altura	Diagnóstico	
	João	37	70	94	190	Saudável	
	Maria	38	65	60	172	Doente	
	José	39	19	70	185	Doente	
	Sílvia	38	25	65	160	Saudável	
Pedro	37	70	90		168	Doente	

# Tipos de Atributos

- Simbólicos ou qualitativos
  - Nominal ou categórico
    - Ex.: *cor, código de identificação, profissão*
  - *Ordinal*
    - Ex.: *gosto (ruim, médio, bom), dias da semana*
- Numéricos, contínuos ou quantitativos
  - Intervalar
    - Ex.: *data, temperatura em Celsius*
  - Racional
    - Ex.: *peso, tamanho, idade*



# Tipos de Atributos : Exemplo

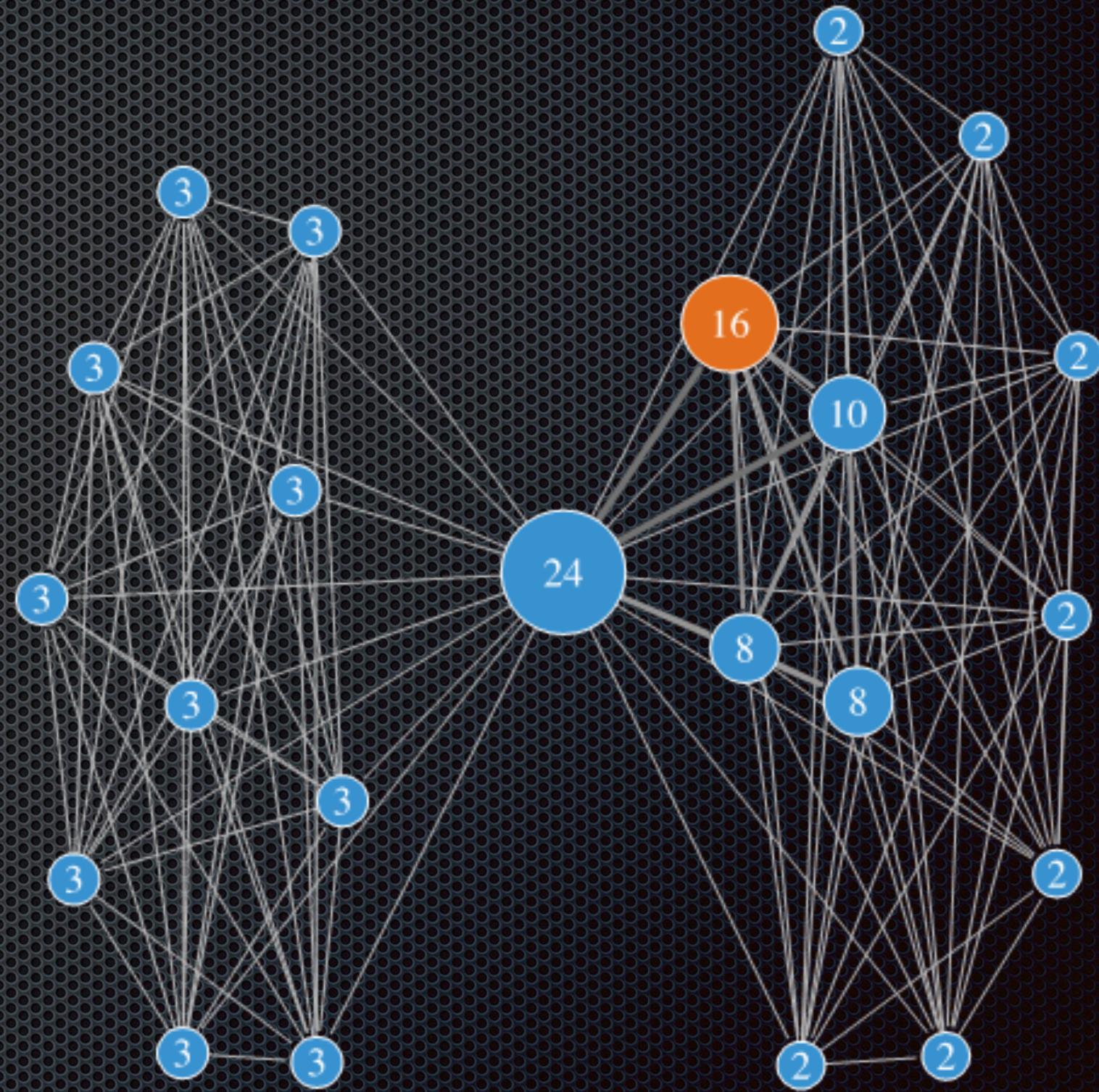
Nome	Temperatura	Batimento	Dor	Salário	Diagnóstico
João	37.7	baixo	sim	1000	Doente
Maria	37.0	normal	não	1100	Saudável
José	38.2	elevado	não	800	Saudável
Sílvia	39.0	baixo	sim	2000	Doente
Pedro	37.3	elevado	sim	1800	Saudável

|            |            |            |            |

Nominal      Intervalar      Ordinal      Racional

# Tipos de Atributos - Sentido

- Nominal ( = , <>)
    - Valores são apenas nomes diferentes
  - Ordinal (< , >)
    - Existe uma relação de ordem entre valores
  - Intervalar ( + , - )
    - Diferença entre valores faz sentido
  - Racional ( \* , +, -)
    - Razão e diferença entre valores fazem sentido



# Exercício 01

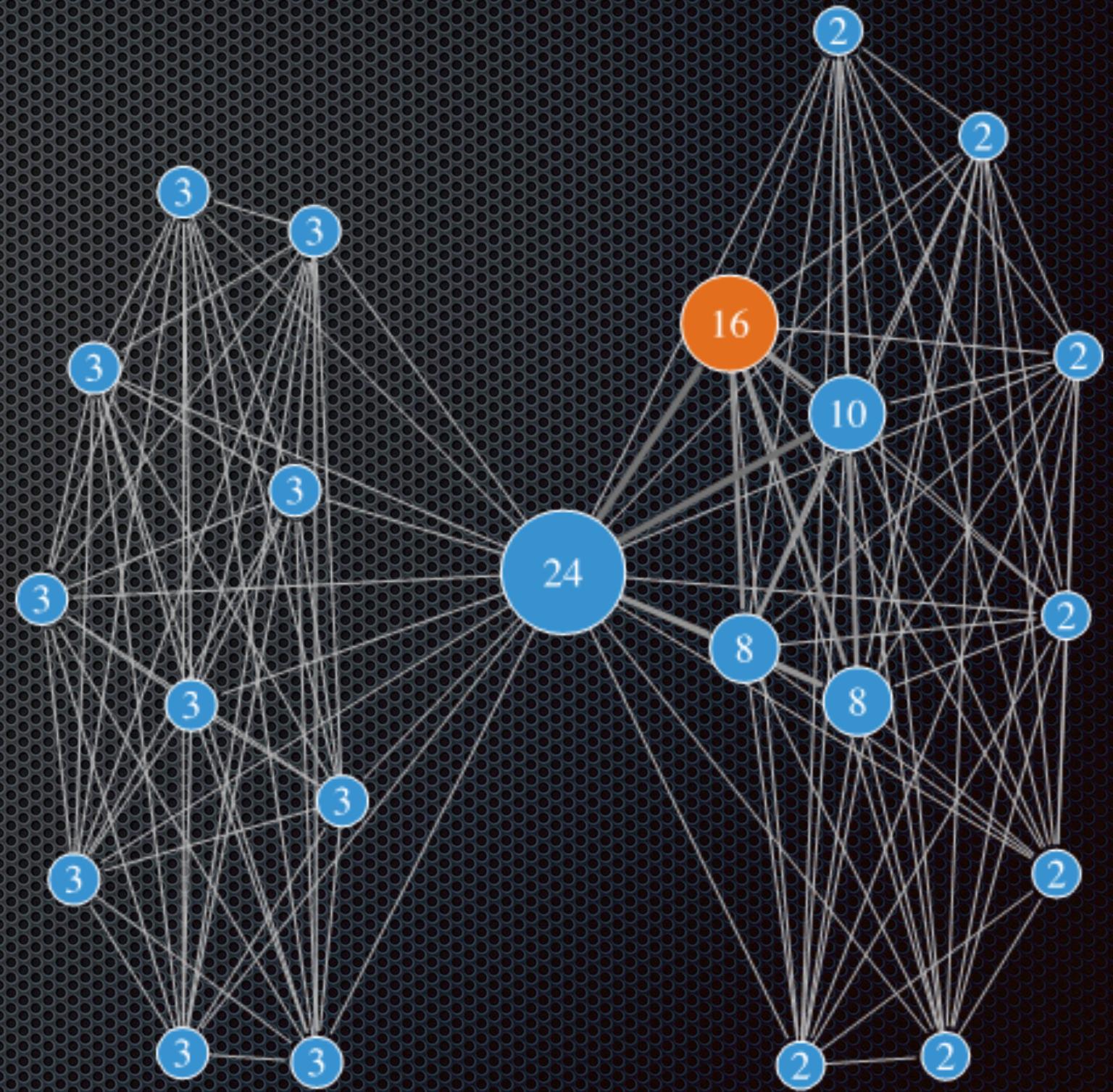
- Definir o tipo dos seguintes atributos:
  - Número de palavras de um texto
  - Fotografia
  - Número de RG
  - Data de nascimento
  - Código de disciplina
  - Posição em uma corrida
  - Expressão de um gene em um tecido
  - Sequência de aminoácidos



# Quantidade de Valores

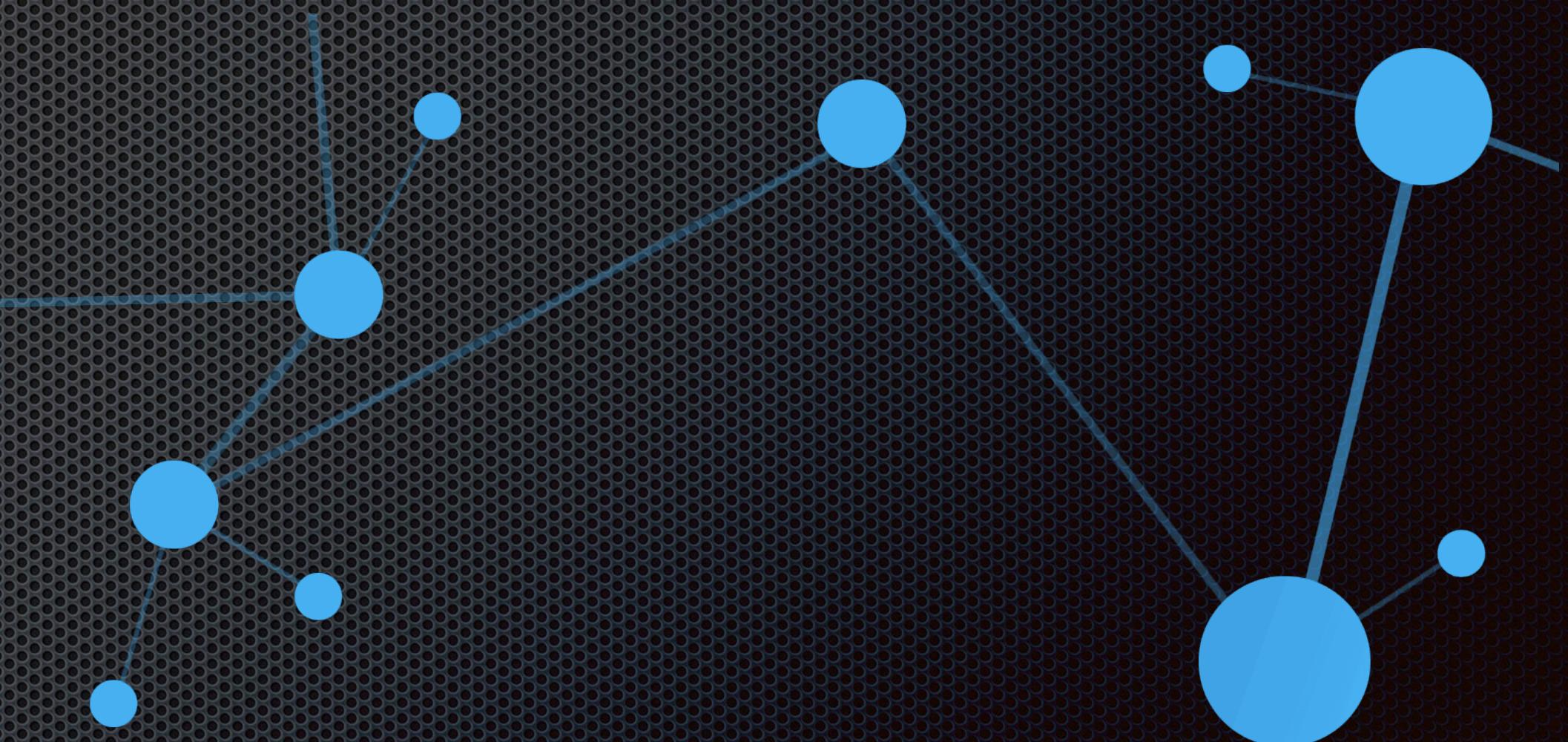
Atributos também se distinguem pela quantidade de valores

- Discretos
    - Número finito ou infinito e enumerável de valores, como números naturais.
      - Ex. código postal, contagem (quantidade de algum elemento)
  - Contínuos
    - Número infinito de valores, como números reais
      - Ex. temperatura, peso, distância



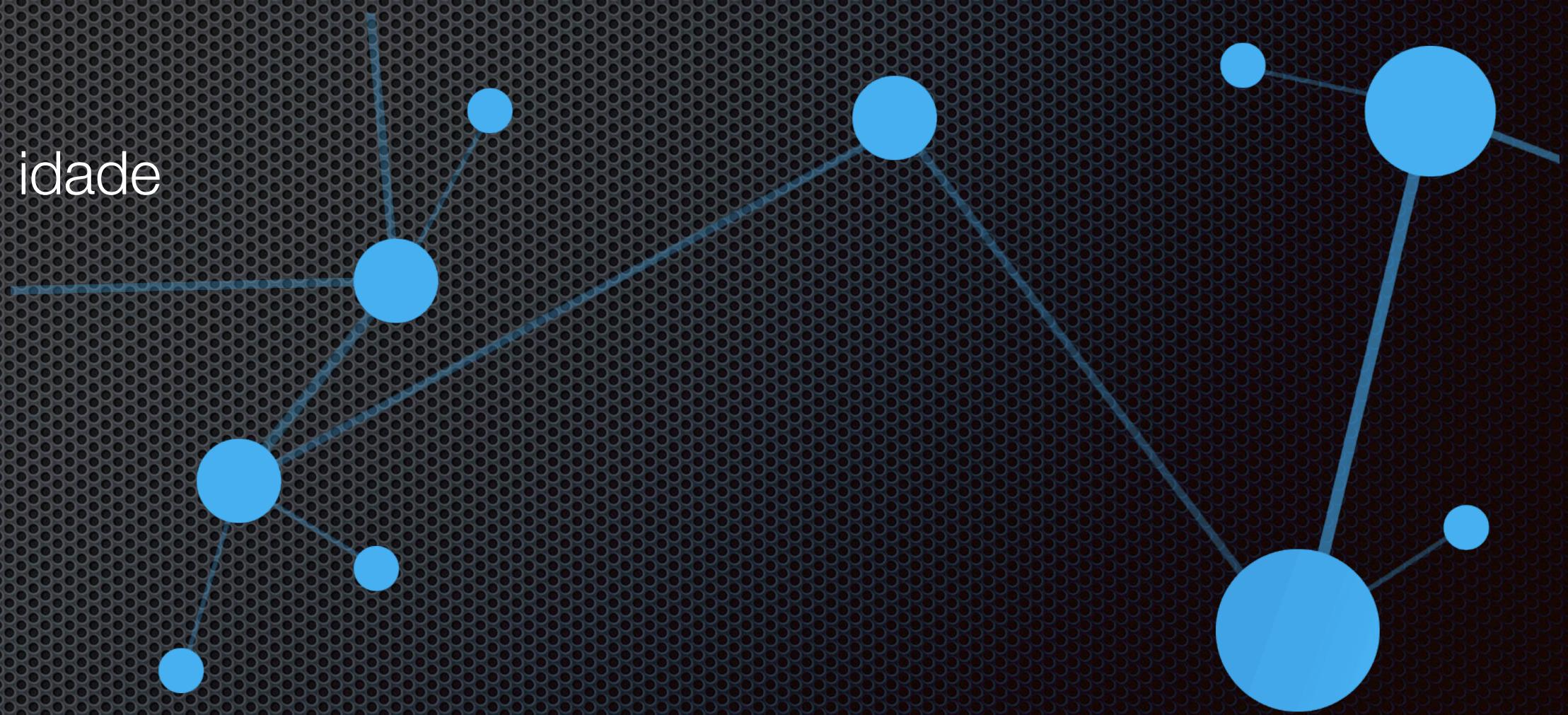
# Estatística Descritiva

- Descreve propriedades estatísticas de dados
- Produz valores que resumem características de um conjunto de dados
  - Na maioria das vezes por meio de cálculos muito simples
- Pode capturar medidas de:
  - Frequência
  - Localização ou tendência central
    - Ex.: Média
  - Dispersão ou espalhamento
    - Ex.: Desvio padrão
  - Distribuição ou formato
    - Ex. Momento



# Estatística Descritiva - Frequência

- Proporção de vezes que um atributo assume um dado valor
- Em um determinado conjunto de dados
- Muita usada para dados categóricos
  - Ex.: Em um BD de um hospital, 60% dos pacientes é maior de idade



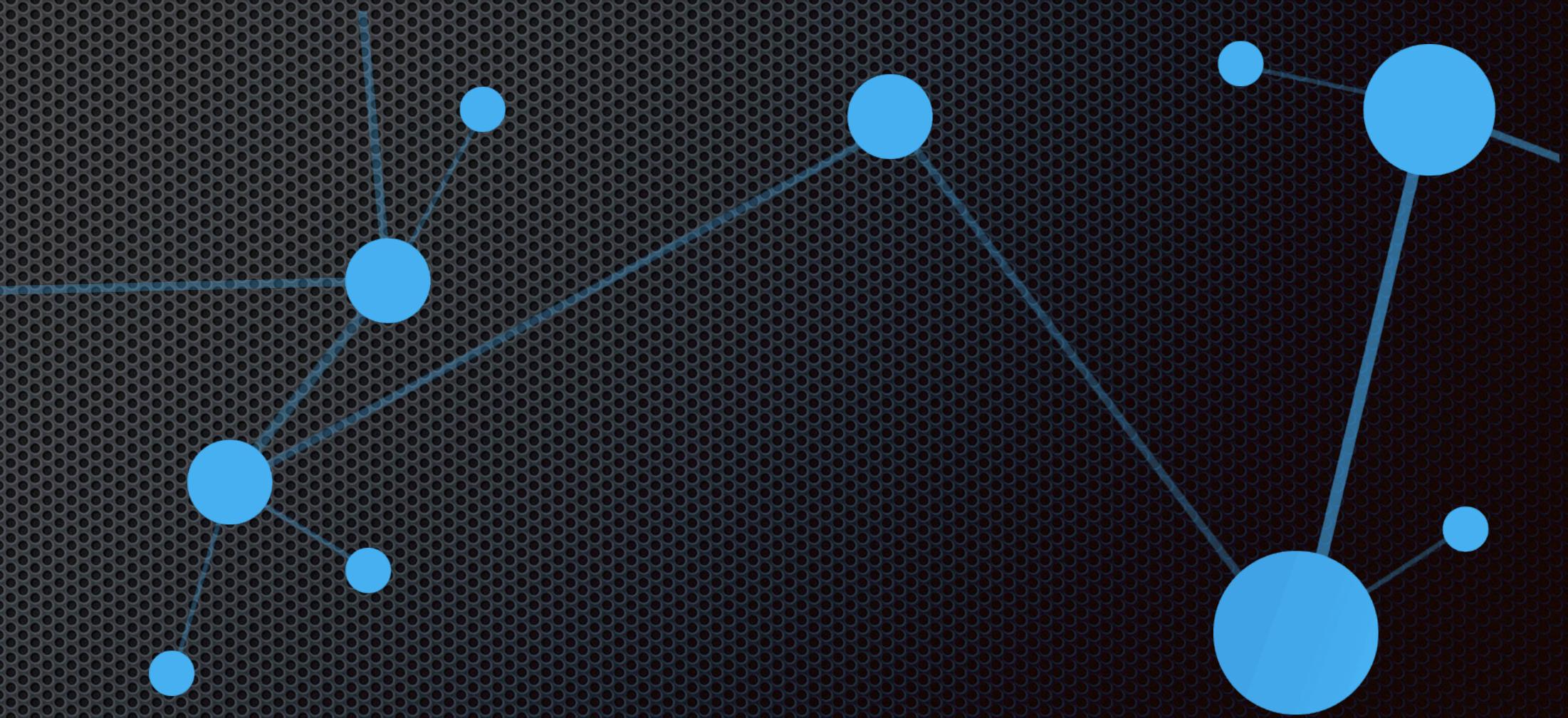
# Estatística Descritiva - Frequência - Exemplo

Nome	Temperatura	Batimento	Dor	Salário	Diagnóstico
João	37.7	baixo	sim	1000	Doente
Maria	37.0	normal	não	1100	Saudável
José	38.2	elevado	não	800	Saudável
Sílvia	39.0	baixo	sim	2000	Doente
Pedro	37.3	elevado	sim	1800	Saudável

40% das medidas de batimento cardíaco encontradas em pacientes são elevados.

# Estatística Descritiva - Medidas de Localidade

- Tendência central
- Valores quantitativos
  - Média
  - Mediana
  - Percentil
- Valores qualitativos ou quantitativos
  - Moda



## Estatística Descritiva - Medidas de Localidade - Média

- Pode ser calculada facilmente
- $\text{média}(x) = \frac{1}{n} \sum_{i=1}^n x_i$
- Problema: sensível a outliers

# Estatística Descritiva - Medidas de Localidade - Mediana

- Menos sensível a *outliers* que média
- Necessário ordenar valores

Se houver um número ímpar de elementos, o número do meio é o valor do meio

$$\frac{n+1}{2}$$

(na amostra de sete elementos  $\{1, 3, 3, 6, 7, 8, 9\}$ , a mediana é  $6$ )

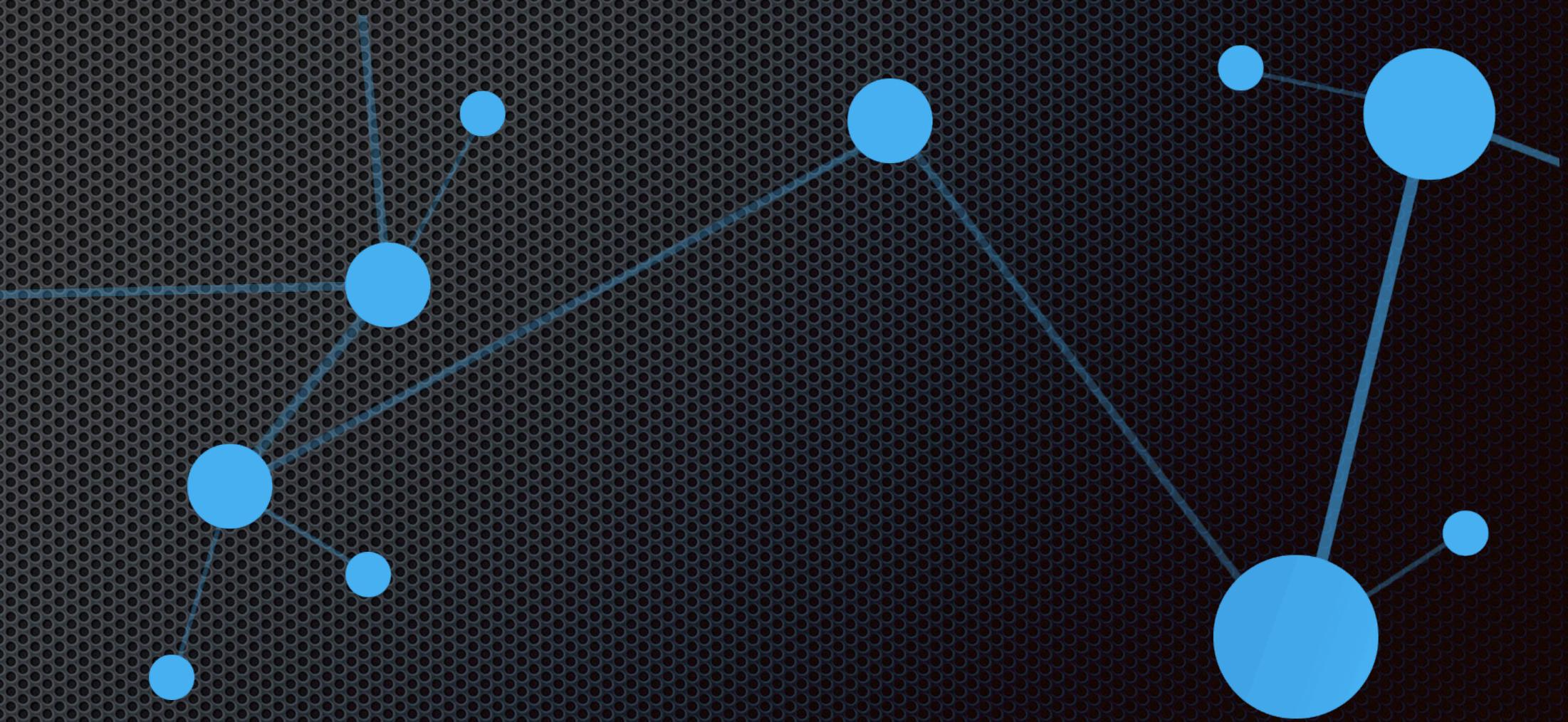
Se houver um número par de elementos, não há um único valor do meio. Então, a mediana é definida como a média dos dois valores do meio

$$\frac{n}{2} \text{ e } \frac{n+1}{2}$$

(na amostra de oito elementos  $\{1, 2, 3, 4, 5, 6, 8, 9\}$ , a mediana é a média  $(4+5)/2 = 4.5$ )

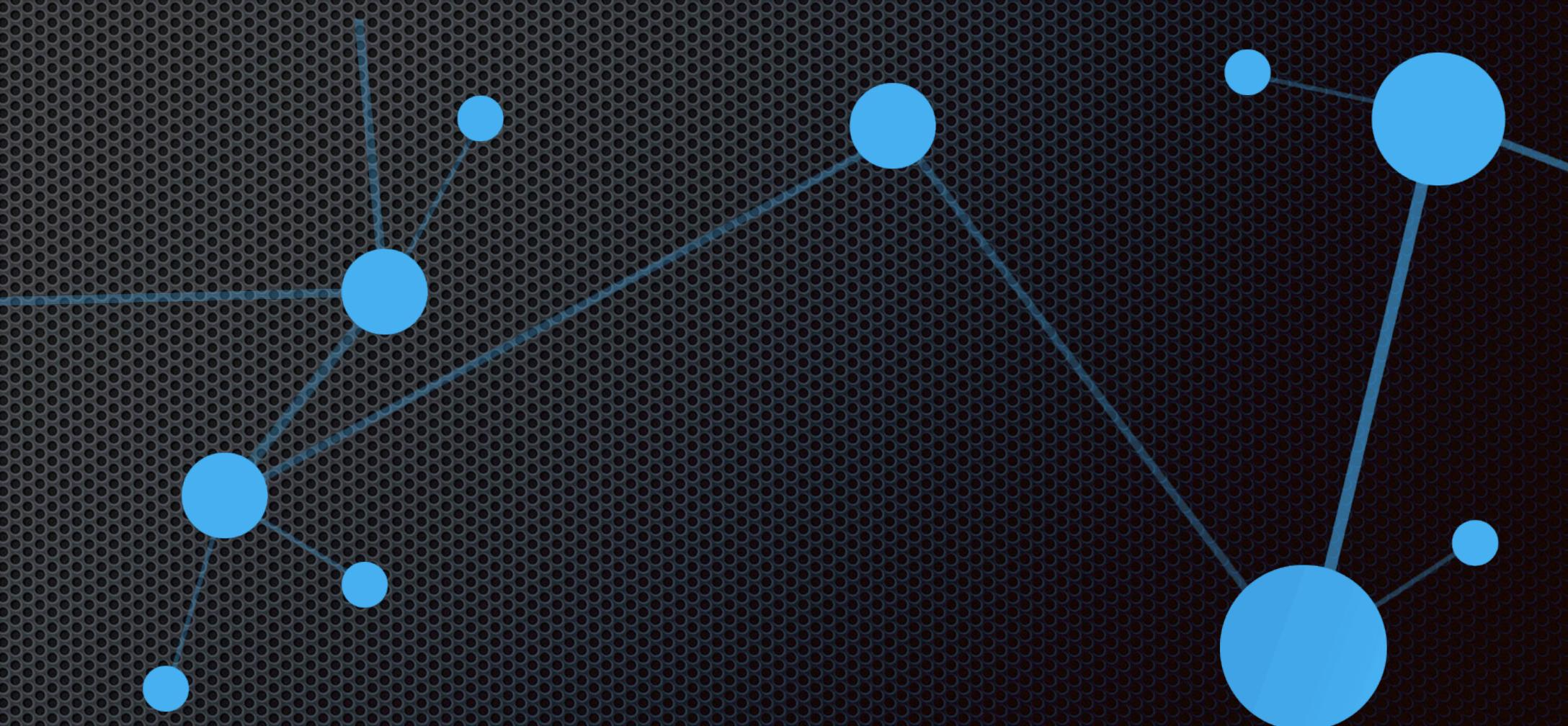
# Estatística Descritiva - Média vs Mediana

- Média é uma boa medida de localização quando os valores estão distribuídos simetricamente.
- Mediana indica melhor o centro
  - Se existem outliers



# Estatística Descritiva - Moda

- Valor mais frequente nos dados
  - Nenhuma moda: Todos os valores são iguais
  - Uma moda: Unimodal
  - Mais de uma moda: Multimodal (Bimodal, Trimodal, ...)
- Indicada quando existem poucos possíveis valores



# Estatística Descritiva - Moda - Exemplo

Nome	Temperatura	Batimento	Dor	Salário	Diagnóstico
João	37.7	baixo	sim	1000	Doente
Maria	37.0	normal	não	1100	Saudável
José	38.2	elevado	não	800	Saudável
Sílvia	39.0	baixo	sim	2000	Doente
Pedro	37.3	elevado	sim	1800	Saudável

Valor da **moda** para o atributo **diagnóstico** : Saudável

Valor da **moda** para o atributo **dor**: sim

# Exercício 02

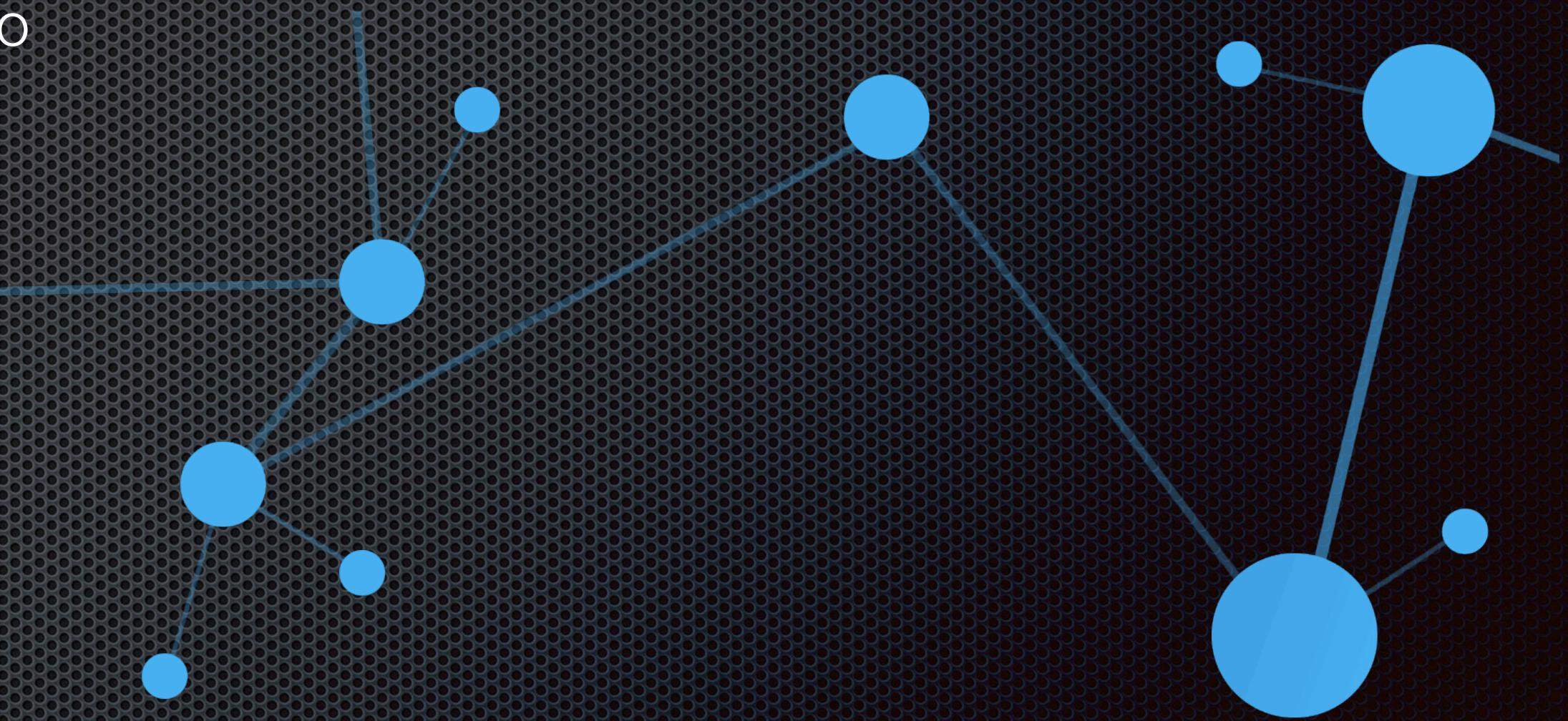
- Dado o conjunto de dados  $\{2, 2, 3, 3, 4, 5, 80\}$ , calcular:

- Média
- Mediana
- Moda



# Estatística Descritiva - Medidas de Espalhamento

- Medem variabilidade, dispersão ou espalhamento de um conjunto de valores
- Indicam se os dados estão:
  - Amplamente espalhados ou relativamente concentrados em torno de um ponto (ex. média)
- Medidas comuns
  - Intervalo ou amplitude
  - Variância
  - Desvio padrão



# Estatística Descritiva - Medidas de Espalhamento - Intervalo

- Medida mais simples
  - Mostra espalhamento máximo
  - Usada em controle de qualidade
- Sejam  $\{x_1, ..., x_n\}$  valores para um atributo  $x$

$$r(x) = \max(x) - \min(x)$$

- Pode não ser uma boa medida
  - Maioria dos valores próximos de um ponto e poucos valores próximos aos extremos

# Estatística Descritiva - Medidas de Espalhamento - Variância

- Medida mais utilizada para analisar espalhamento de valores
- $\text{variância}(x) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$
- Denominador  $n-1$  : correção de Bessel, usada para uma melhor estimativa da variância verdadeira
- Desvio padrão: [raiz quadrada da variância](#)
- Um dos momentos de uma distribuição de probabilidade

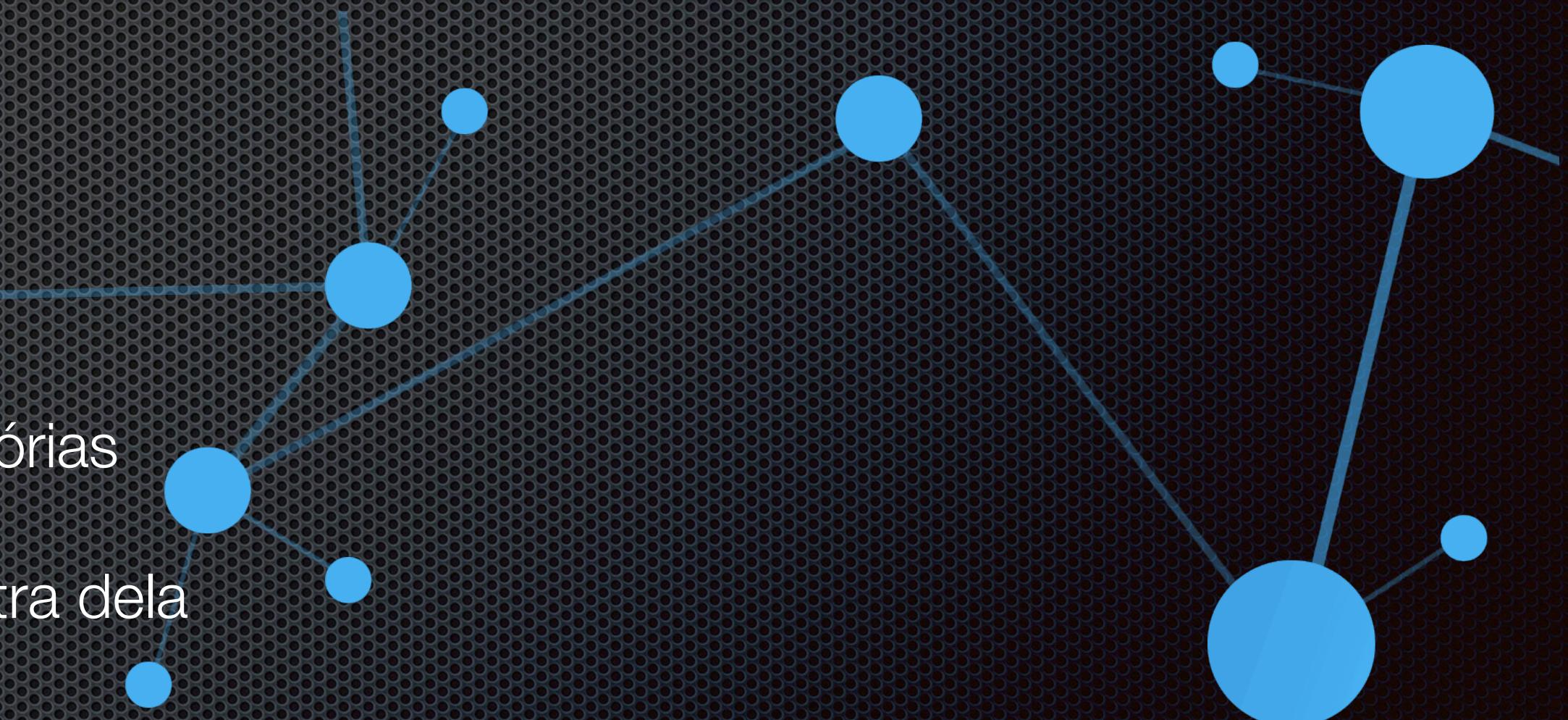
# Exercício 03

- Dado o conjunto de dados {16, 25, 4, 18, 11, 13, 20, 8, 11 e 9}, calcular:
  - Amplitude
  - Variância
  - Desvio Padrão



# Estatística Descritiva - Medidas de Espalhamento

- Definem como os valores de uma variável (atributo) estão distribuídos
- Calculada por meio de momentos
  - Medida quantitativa usada na estatística e na mecânica
  - Captura o formato da distribuição de um conjunto de valores
- Usados para caracterizar a distribuição de valores de variáveis aleatórias
- Estimam medidas de uma população de valores usando uma amostra dela
- Vários cálculos de momento
  - Cálculo de momento original
  - Cálculo de momento central
  - Cálculo de momento padronizado



# Bibliografia

- SILVA. L. A.; PERES. S. M; BOSCAROLI C. [Introdução à Mineração de Dados](#). Elsevier. 2016
- FACELI, Katti; Lorena, Ana Carolina; Gama, João ; de Carvalho, A. C. P. L. F. (2011). [Inteligência Artificial - Uma Abordagem de Aprendizado de Máquina](#). 1. ed. Rio de Janeiro: LTC.
- PROVOST, F.; Fawcett, T. [Data Science for Business: What you need to know about data mining and data-analytic thinking](#) by O'Reilly Media, 2013.
- FLACH, P. (2012). [Machine Learning: The Art and Science of Algorithms that Make Sense of Data](#). Cambridge University Press.
- ALPAYDIN, E. (2004). [Introduction to Machine Learning](#). MIT Press.