

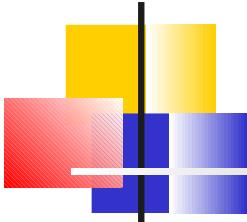
# SCC-275 - Ciência de Dados

## Exploração de dados

Revisado por  
Profa. Roseli Ap. Francelin  
Romero – SCC

Prof. Dr. André C. P. L. F. de Carvalho  
Dr. Isvani Frias-Blanco  
ICMC-USP

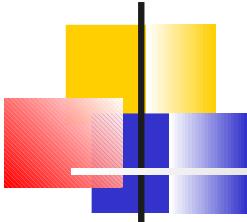




# Dados multivariados

---

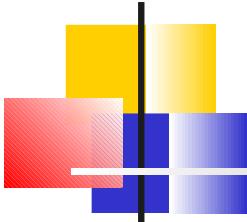
- Possuem mais de um atributo
  - Cada atributo é uma variável
- Medidas de localização (tendência central)
  - Podem ser obtidas calculando medida de localização de cada atributo separadamente
  - Ex.: média, mediana, ...
    - Média dos objetos de um conjunto de dados com  $m$  atributos é dada por:  $\bar{x} = (\bar{x}_1, \dots, \bar{x}_m)$



# Dados multivariados

---

- Medidas de espalhamento (dispersão)
  - Podem ser calculadas para cada atributo independentemente dos demais
    - Usando qualquer medida de espalhamento
      - Intervalo, variância, desvio padrão
  - Para dados multivariados numéricos é melhor usar uma matriz de covariância
    - Cada elemento da matriz é a covariância entre dois atributos



# Dados multivariados

---

- Cálculo de cada elemento  $s_{ij}$  de uma matriz de covariância  $S$  para um conjunto de  $n$  objetos

$$s_{ij} = \text{covariância } (x_i, x_j) = \frac{1}{n-1} \sum_{k=1}^n (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)$$

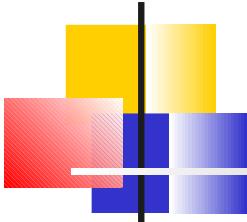
Onde:

$\bar{x}_i$  : Valor médio do i-ésimo atributo

$x_{ki}$ : Valor do i-ésimo atributo para o k-ésimo objeto

É de ordem  $n \times n$

- Obs: covariância  $(x_i, x_i) = \text{variância } (x_i)$ 
  - Matriz de covariância tem em sua diagonal as variâncias dos atributos

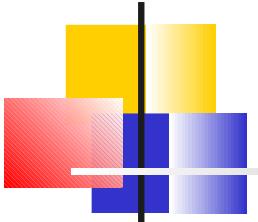


# Exercício

---

- Calcular a matriz de covariância para o conjunto de dados:

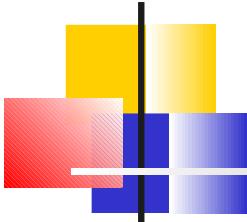
Peso	Altura	Temperatura
73	170	37
67	165	38
90	190	34
49	152	31



# Dados multivariados

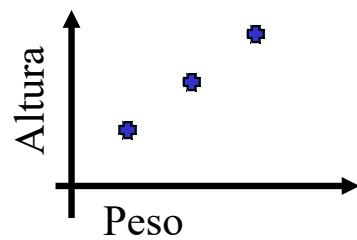
---

- Covariância de dois atributos
  - Mede o grau com que os atributos variam juntos (linearmente)
    - Valor próximo de 0:
      - Atributos não têm um relacionamento linear
    - Valor positivo:
      - Atributos diretamente relacionados
        - Quando o valor de um atributo aumenta, o do outro também aumenta
      - Valor negativo:
        - Atributos inversamente relacionados
    - Valor depende da magnitude dos atributos

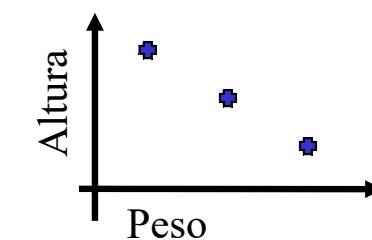


# Exemplo

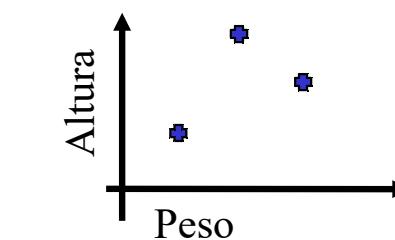
Peso	Altura
60	170
70	180
80	190

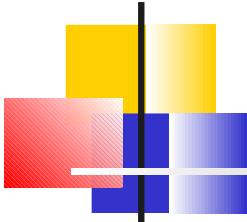


Peso	Altura
60	190
70	180
80	170



Peso	Altura
60	170
70	190
80	180

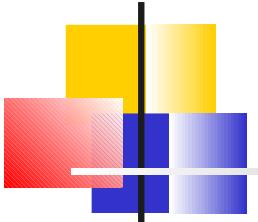




# Dados multivariados

---

- Covariância de dois atributos
  - É difícil avaliar o relacionamento entre dois atributos olhando apenas a covariância
    - Sofre influência da faixa de valores dos atributos
    - **Correlação linear** entre dois atributos ilustra mais claramente a força da relação linear entre eles
      - Mais popular que covariância
      - Elimina influência da faixa de valores



# Dados multivariados

---

- **Correlação linear**

- Indica força da relação linear entre dois atributos
- Matriz de correlação R

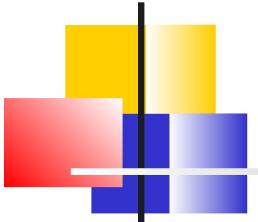
$$r_{ij} = \text{correlação}(x_i, x_j) = \frac{\text{covariância}(x_i, x_j)}{s_i s_j}$$

Onde:

$x_i$ : i-ésimo atributo

$s_i$ : Desvio padrão do atributo  $x_i$

- Obs: correlação  $(x_i, x_i) = 1$ 
  - Elementos da diagonal principal têm valor 1
  - Demais elementos têm valor entre  $-1$  e  $+1$



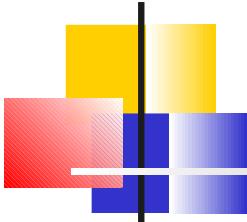
# Exercício

---

- Calcular a matriz de covariância e a matriz de correlação para o conjunto de dados:

Peso	Altura	Temperatura
73	170	37
67	165	38
90	190	34
49	152	31

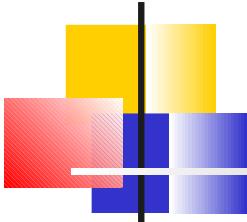
- Ilustrar graficamente pares de atributos direta e inversamente correlacionados



# Outras formas de summarizar dados

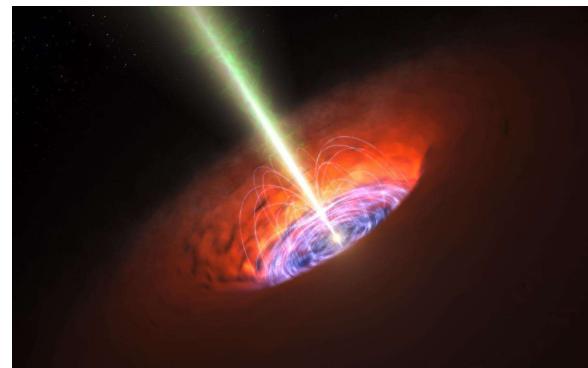
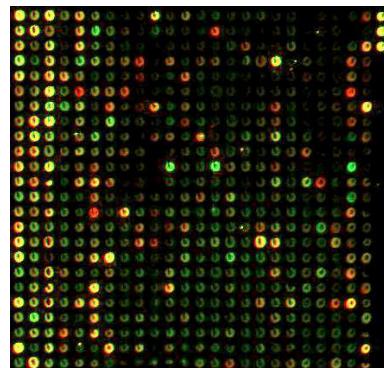
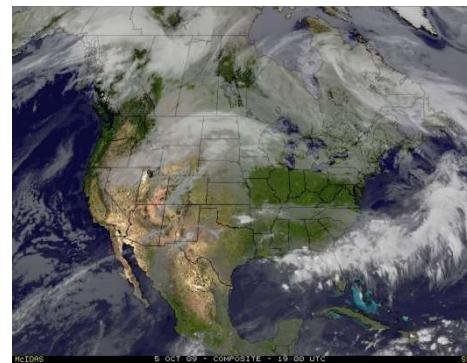
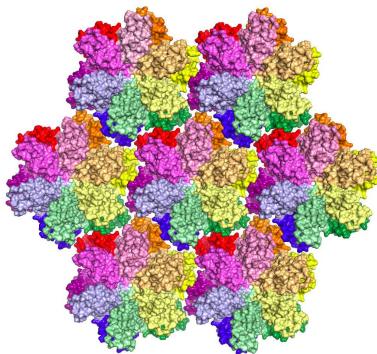
---

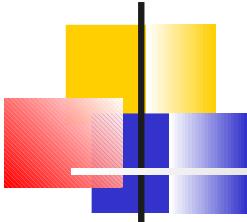
- Visualização gráfica
  - Em vários casos, facilita compreensão de padrões mais complexos nos dados
  - Exemplos simples
    - Histograma
    - Diagrama de torta
    - *Scatter plot*
    - Faces de Chernoff



# Exemplos

---





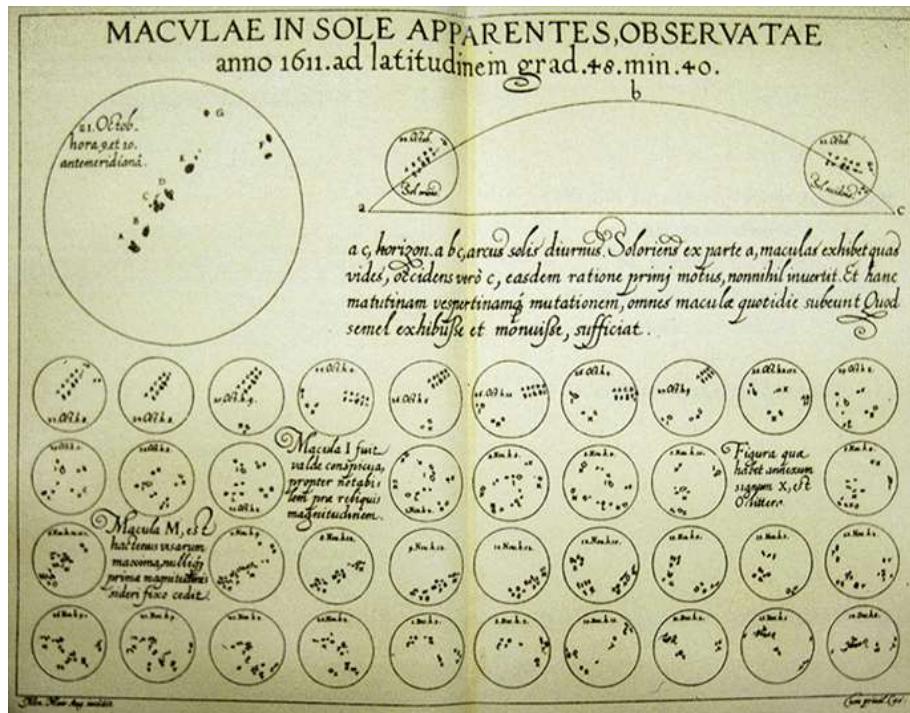
# Visualização

---

- Visualização tem um papel importante em análise de dados
  - Uma das técnicas mais poderosas para exploração dos dados
    - Facilita visualização de dados e resultados
    - *Visual data mining*
      - Usa técnicas de visualização em mineração de dados
      - Importante área de CD

# Um dos primeiros

## ■ Mapa solar de Galileu



# Outro dos primeiros usos



Napoleão

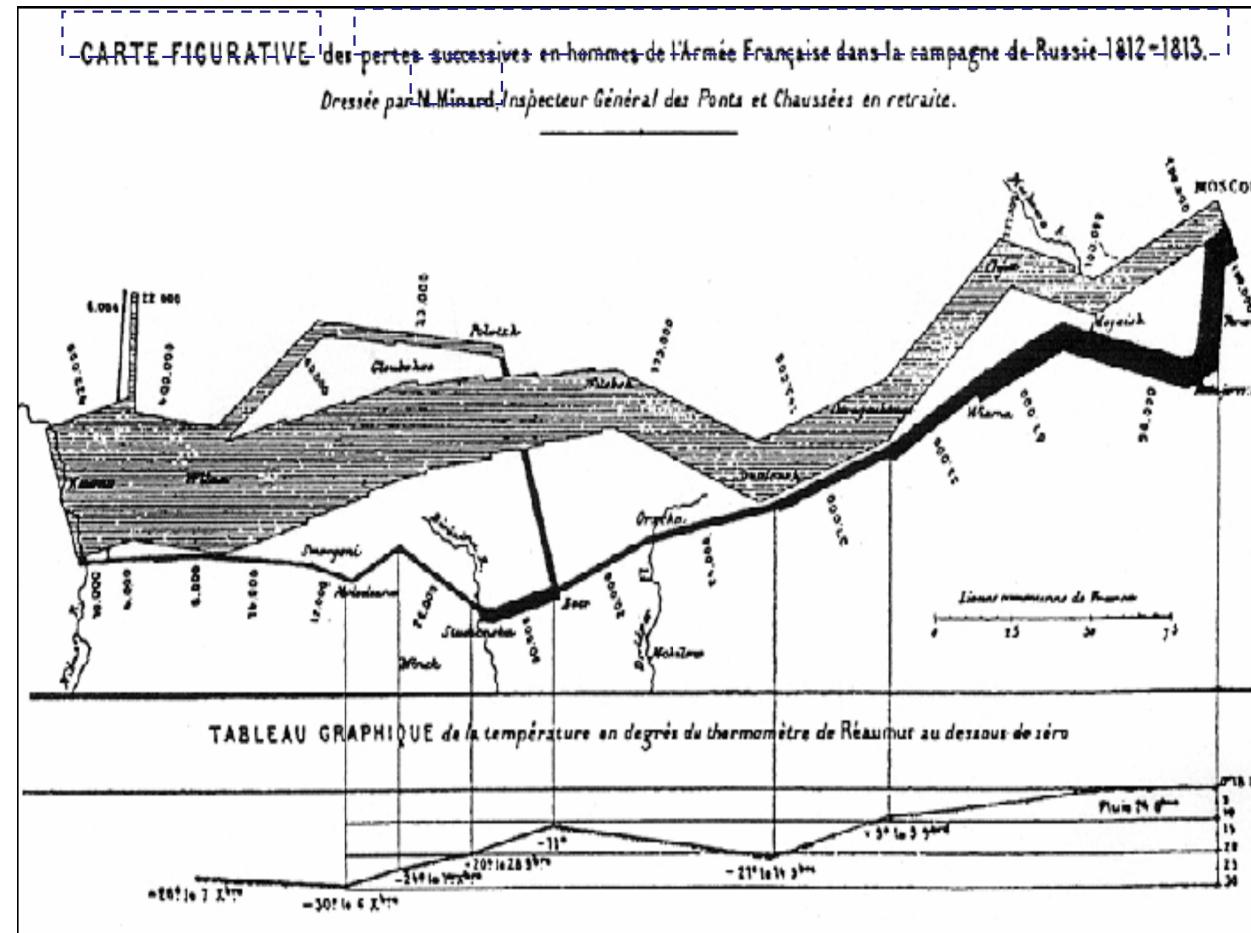
Exército francês, comandado por Napoleão  
Invade a Rússia, em 1812



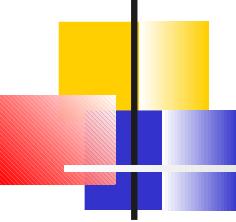
# Outro dos primeiros usos

Perdas humanas do exército francês, sob o comando de Napoleão, na invasão da Rússia em 1812

Charles Joseph Minard (1869)



# O que mostra



Cinza: entra na Rússia  
Preto: sai da Rússia

Cada mm de largura  
Equivale a 10.000 homens

Variação de temperatura

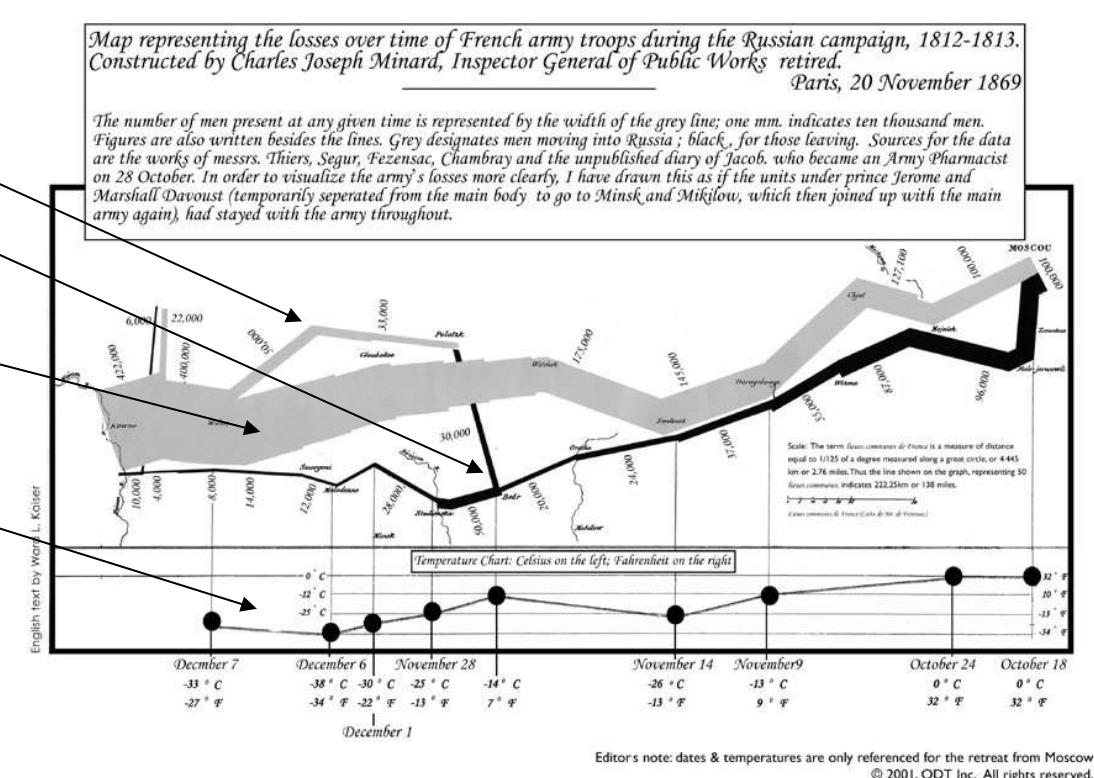


Figure 58. Minard's map of Napoleon's Russian campaign.  
This graphic has been translated from French to English and modified to most effectively display the temperature data.

© www.odt.org , from <http://www.odt.org/Pictures/minard.jpg>, used by permission

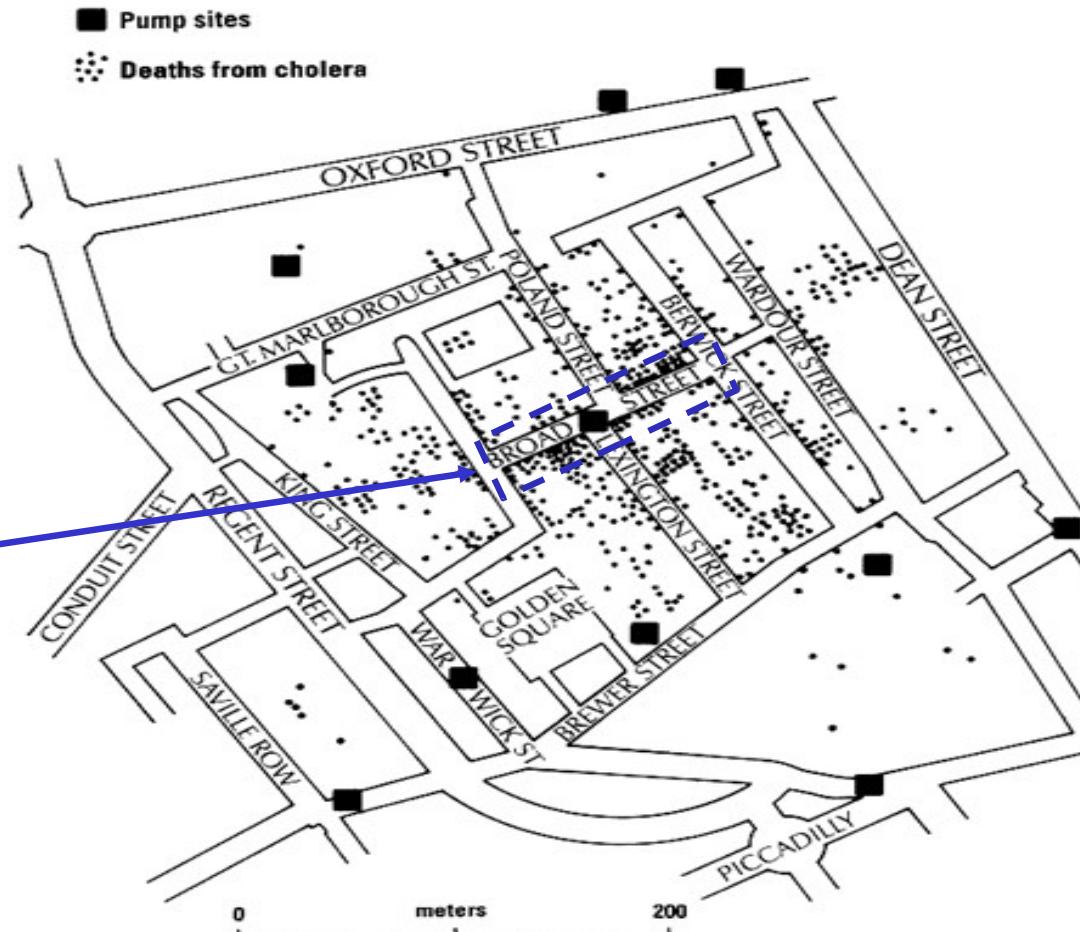
© André de Carvalho - ICMC/USP

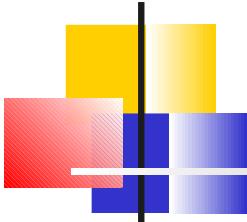
# Outro exemplo

Mapa da Cólica em Londres

(Snow) 1855

Distribuição da doença permitiu identificar que a fonte da cólera era uma bomba de água pública na Broad Street

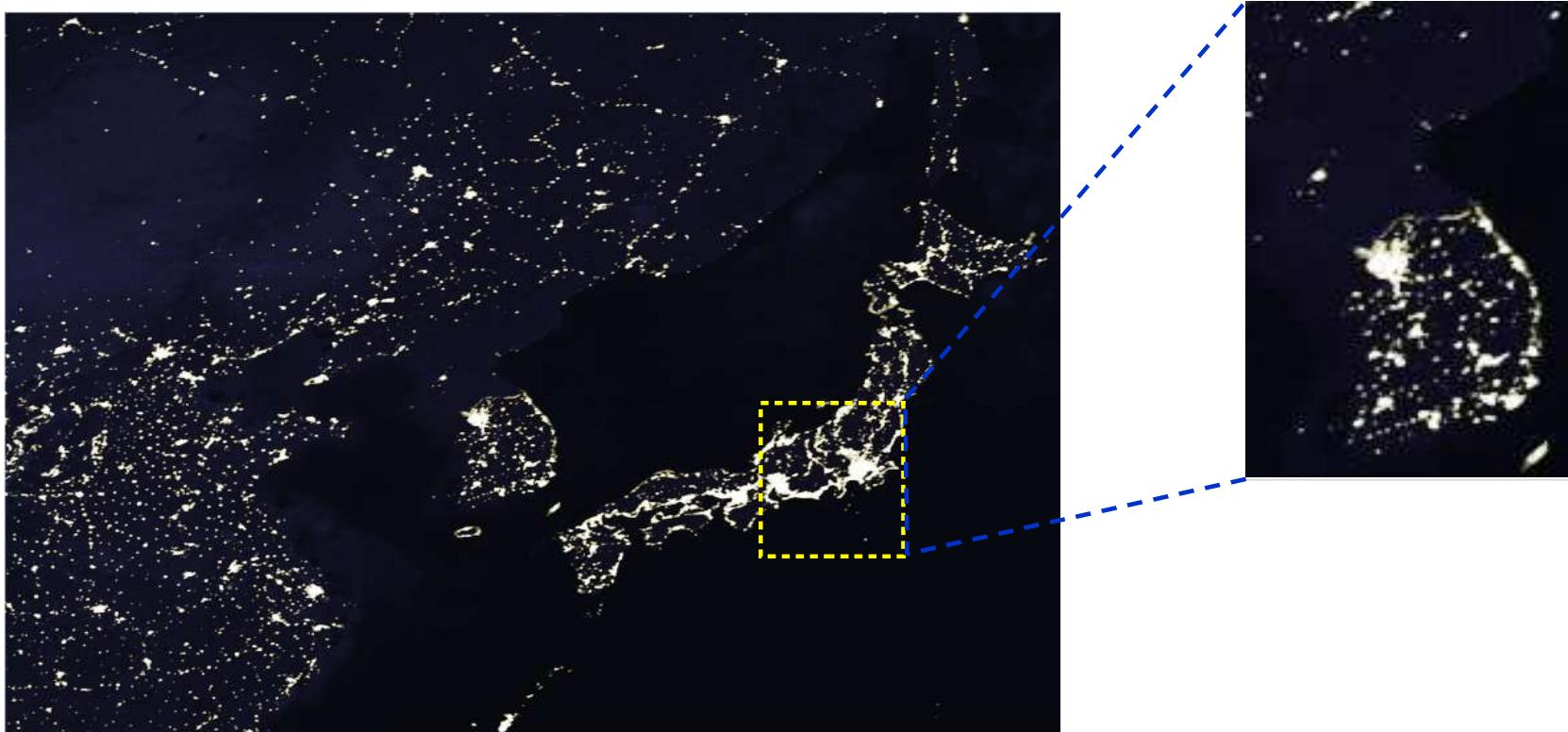


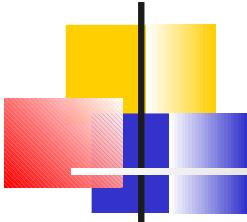


# Exemplo mais recentes

---

Ásia à noite

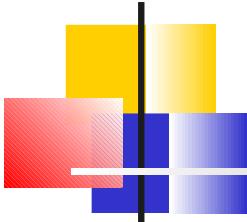




# Motivação

---

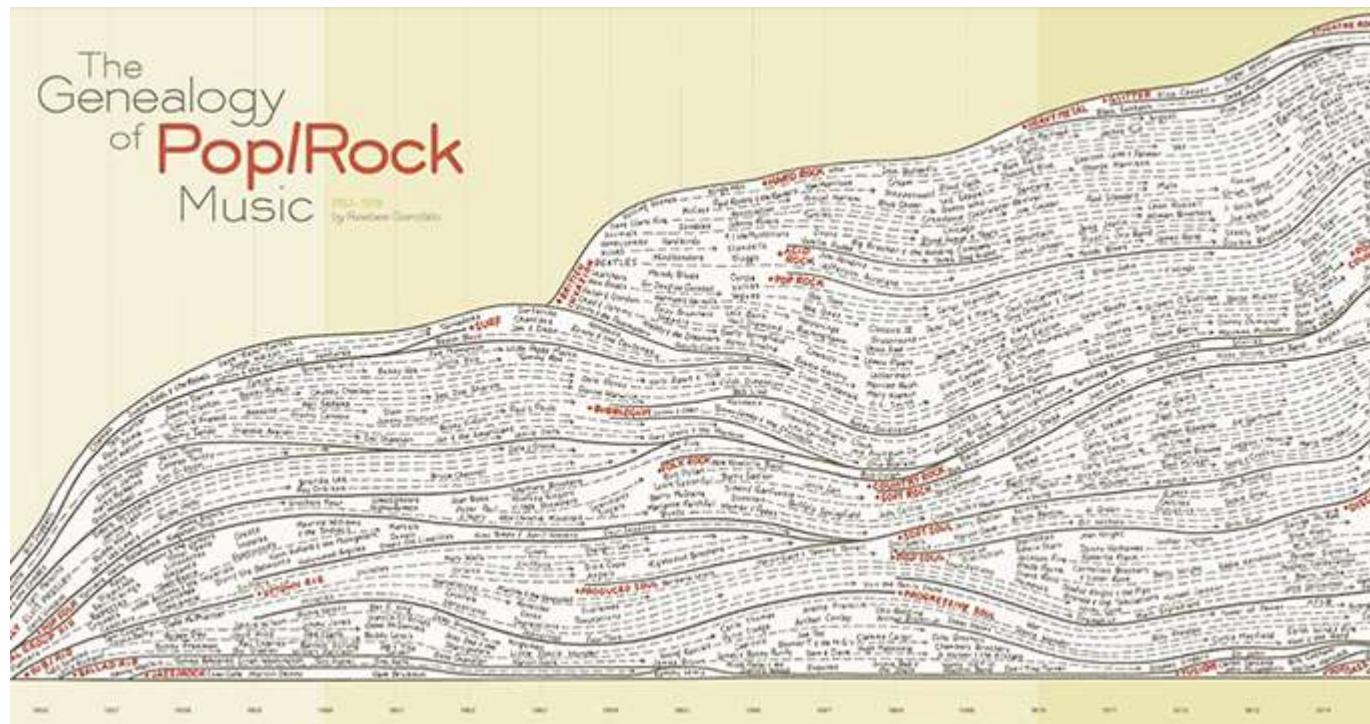
- Pessoas se saem bem na detecção de estruturas em imagens
  - Têm mais facilidade para entender informação representada visualmente
    - E encontrar padrões
  - Conseguem usar conhecimento do domínio que elas têm, mesmo sem estar cientes
    - Difícil fazer isso automaticamente com ferramentas estatísticas ou computacionais

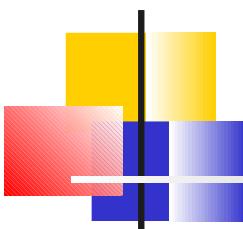


# Benefícios

---

- Resumo informação presente em dados e resultados experimentais
  - Torna mais claros
    - Padrões e tendências gerais
    - Anomalias e outliers
- Facilita análise de grandes conjuntos de dados

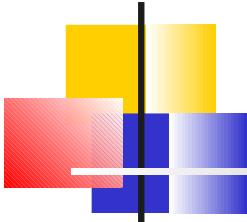




# Mais Claro?

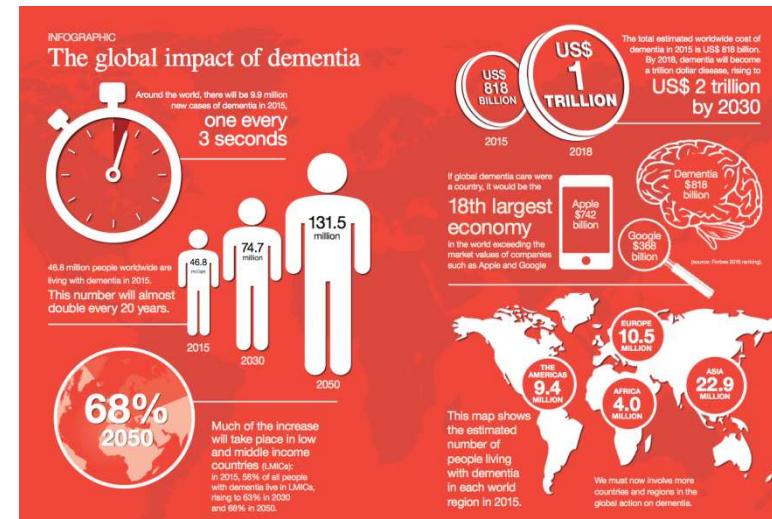
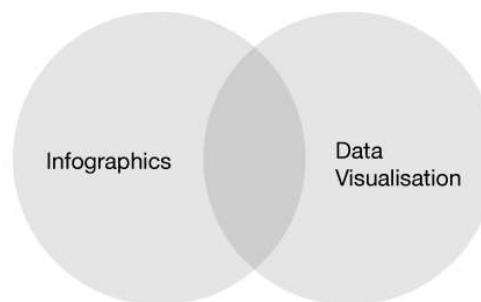


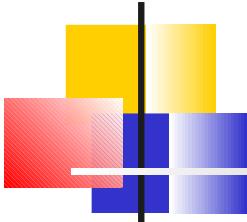
# Infográfico



# Visualização e infográficos

- Ambos transformam dados em uma imagem
- Importantes para ilustrar padrões

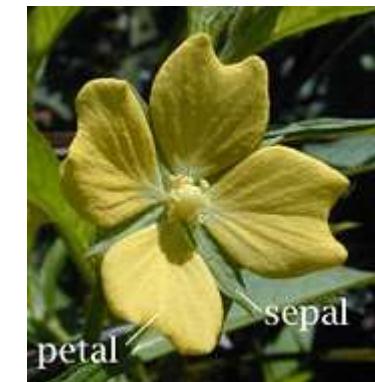


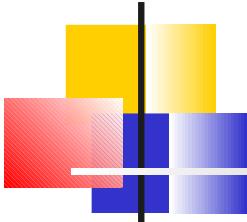


# Conjunto de dados iris

---

- Iris (lírio): planta com flor
  - Atributos de entrada numéricos
    - Tamanho sépala (cm)
    - Largura sépala (cm)
    - Tamanho pétala (cm)
    - Largura pétala (cm)
  - Classes
    - Iris Setosa
    - Iris Versicolour
    - Iris Virginica
  - 150 exemplos, com distribuição 50/50/50

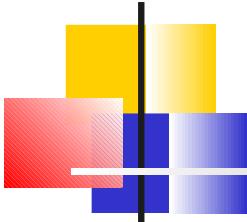




# Histogramas

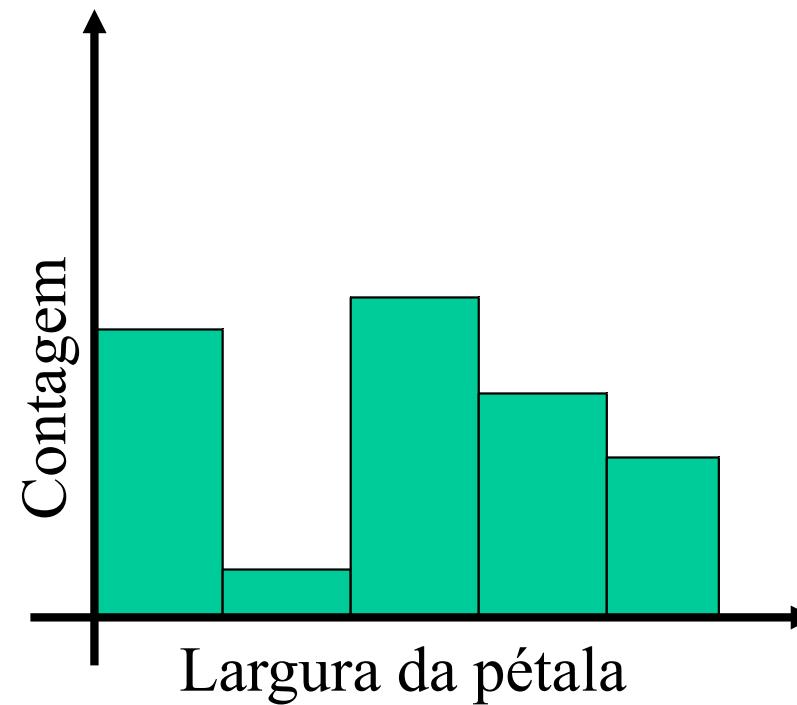
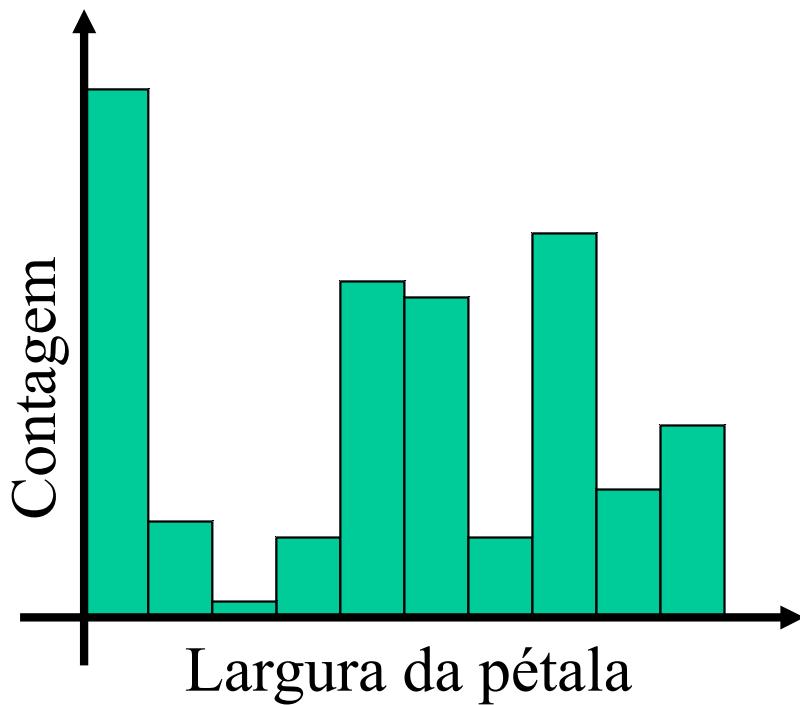
---

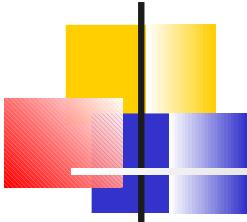
- Divide os valores em cestas e exibe uma barra para cada cesta
  - Comprimento da barra proporcional ao número de objetos na cesta
    - Valores ordinais ou numéricos:
      - São divididos nas cestas (geralmente, em intervalos do mesmo tamanho)
    - Valores nominais:
      - Cada valor (ou subconjunto de valores) é uma cesta
- Formato depende do número de cestas



# Histogramas

- Conjunto de dados Iris
  - Largura das pétalas usando 10 e 5 cestas





# Diagrama de torta

---

- Frequências relativas podem ser vistas no diagrama circular

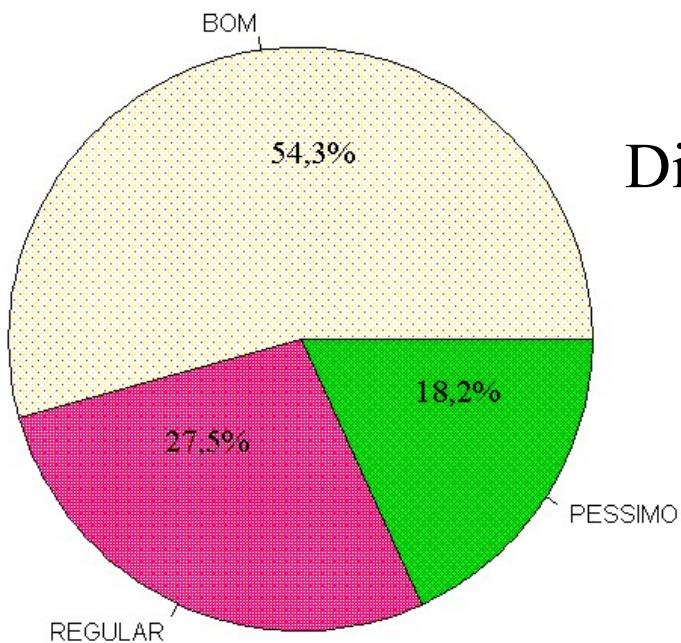
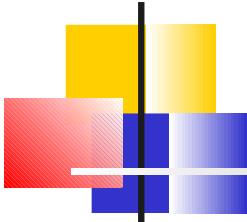


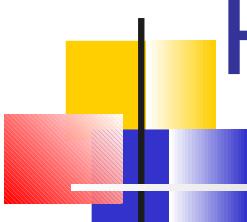
Diagrama de torta (pizza)



# Scatter Plot

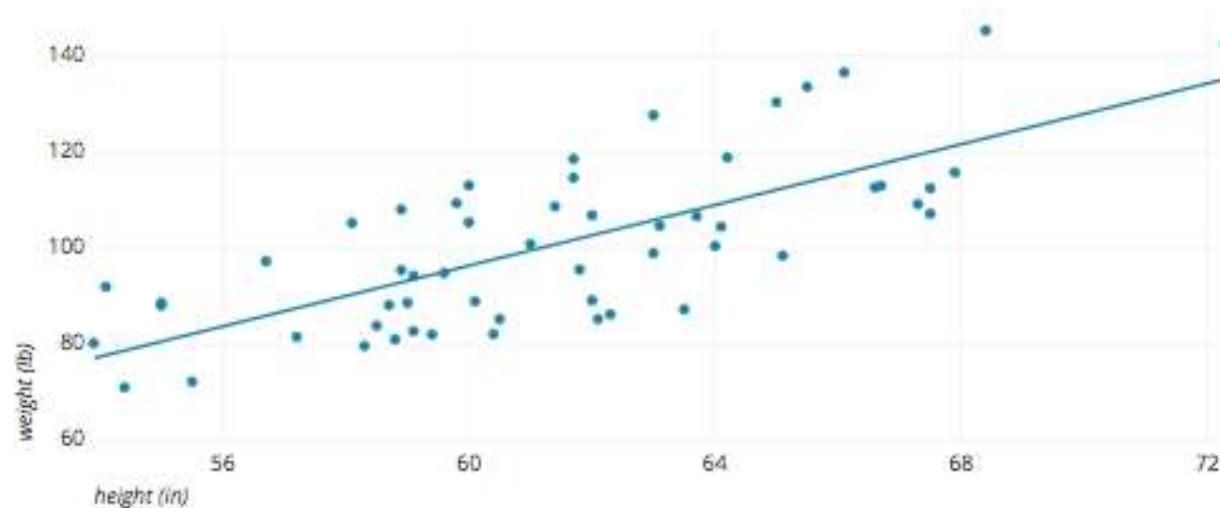
---

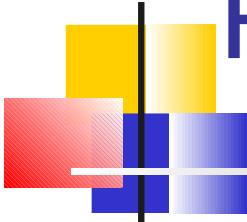
- Usado para ilustrar graficamente correlação linear entre dois atributos
- Cada objeto é associado a uma posição em um gráfico
  - Valores dos atributos definem sua posição
  - Valores podem ser inteiros ou reais
- Matrizes de scatter plot resumem relação para vários pares de atributos



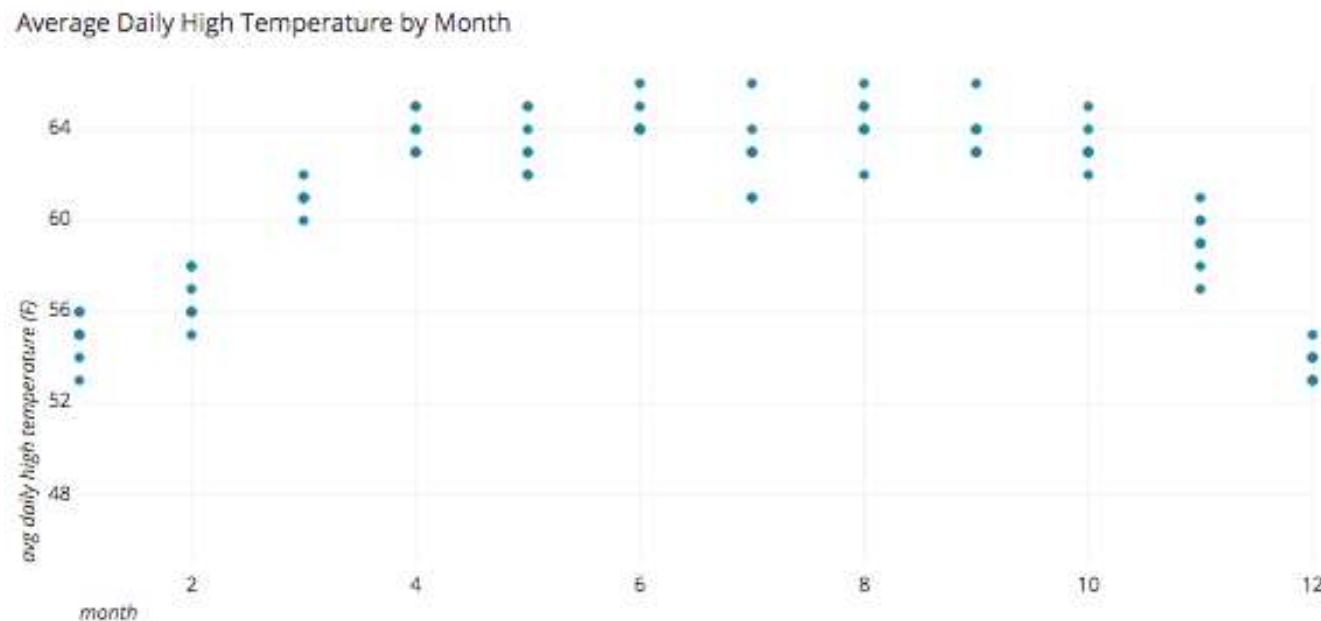
# Relação Linear

Weight and Height of Children



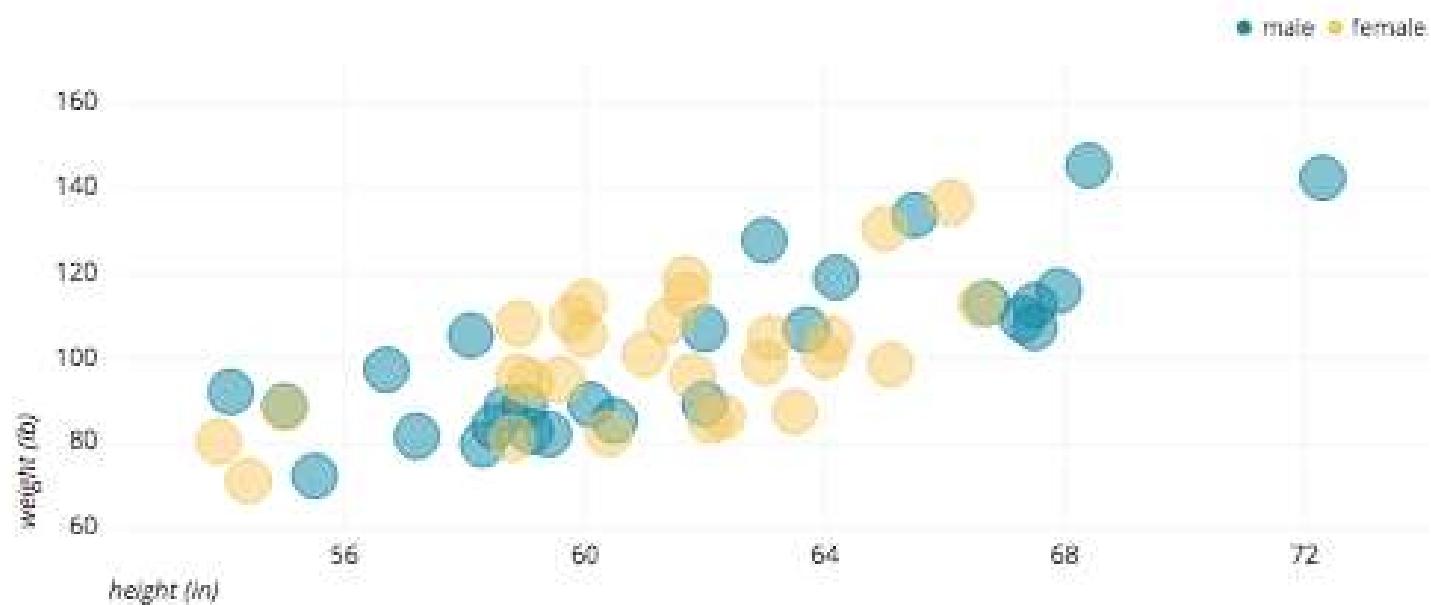


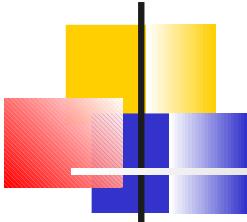
# Relação Não-Linear



# Pode adicionar mais info

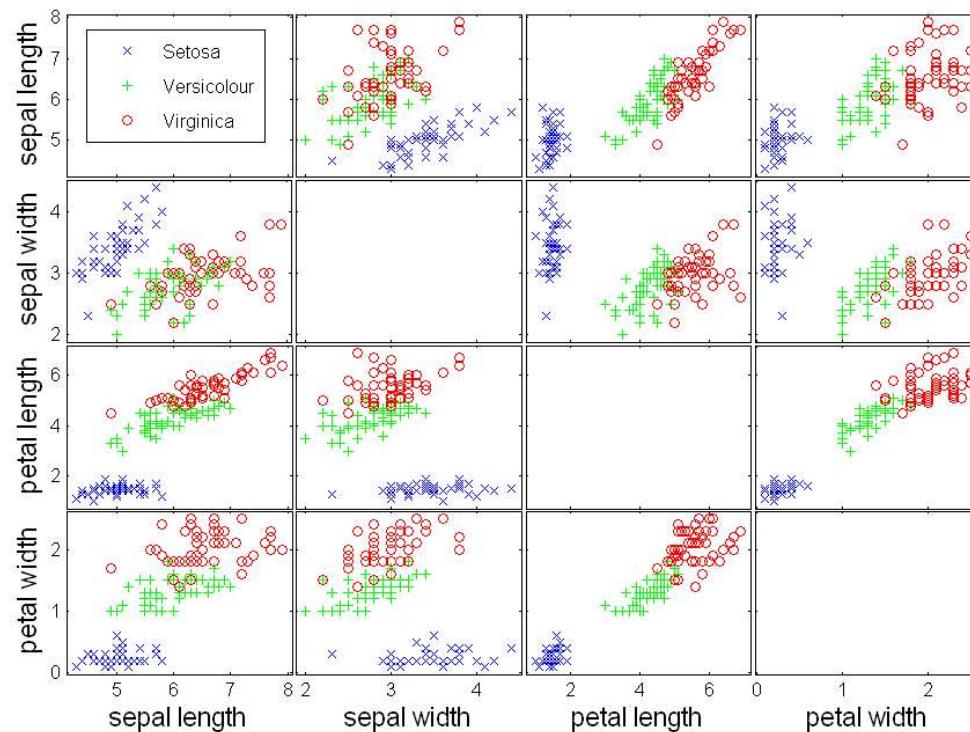
Weight and Height of Children by Gender



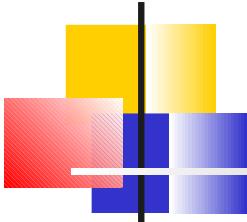


# Scatter Plot

- Matriz para atributos do conjunto iris



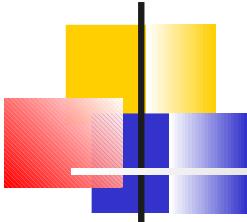
Diferentes classes  
são indicadas por  
cores diferentes



# Faces de Chernoff

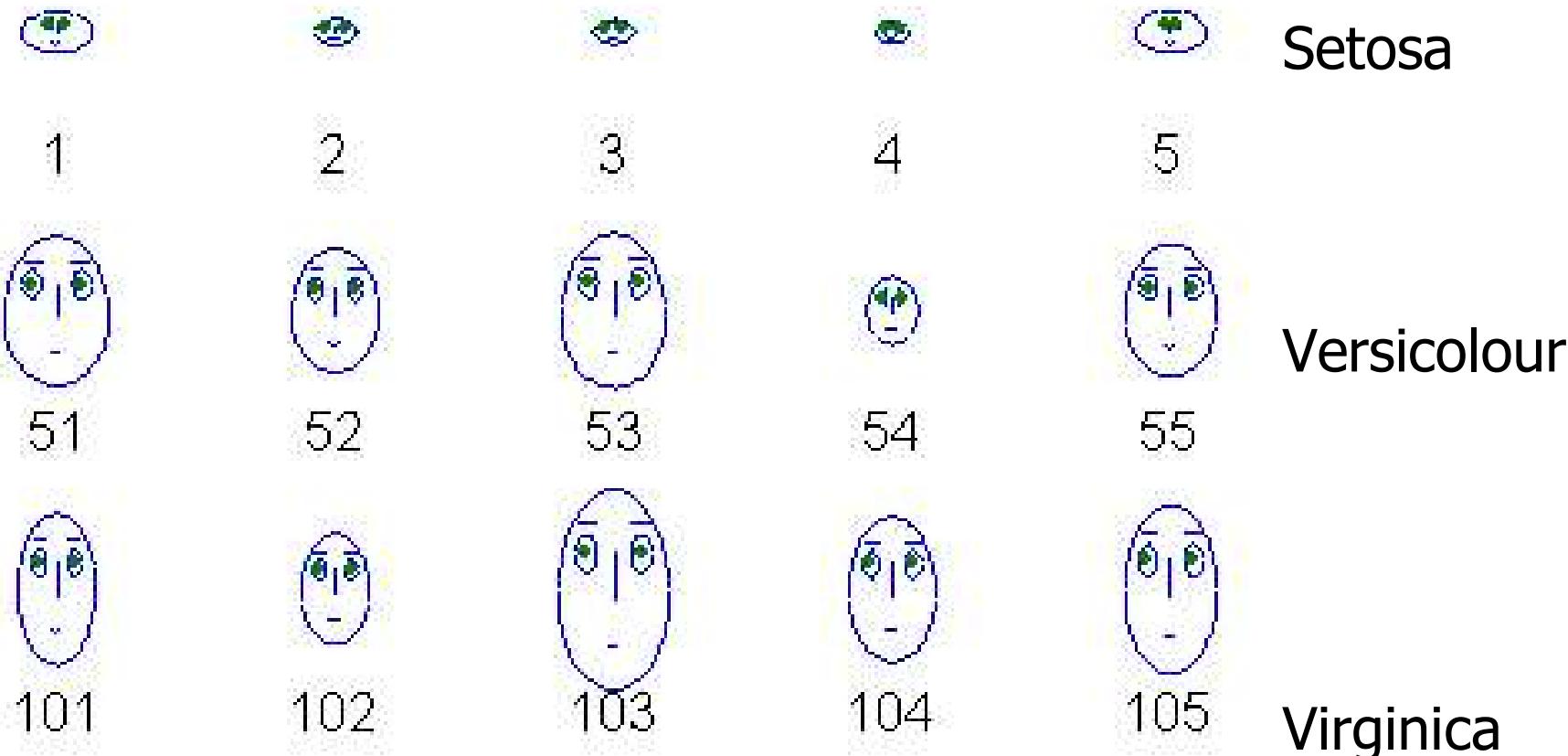
---

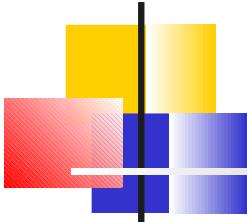
- Criado por Herman Chernoff
- Mapeia os valores dos atributos para imagens familiares para seres humanos: faces
  - Cada objeto é representado por uma face
  - Cada atributo é associado a uma característica específica da face
- Baseia-se na habilidade humana de distinguir faces



# Faces de Chernoff

---



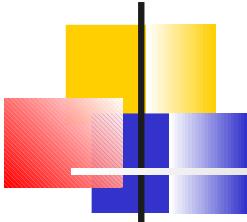


# Exercício

---

- Representar os dados a seguir usando faces de Chernoff

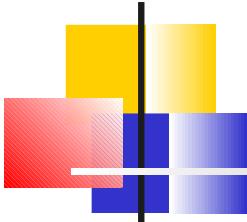
Febre	Idade	Batimento	Dor	Diagnóstico
sim	23	elevado	sim	doente
não	9	baixo	não	saudável
sim	61	elevado	não	saudável
sim	32	baixo	sim	doente
sim	21	elevado	sim	saudável
não	48	elevado	sim	doente



# Considerações Finais

---

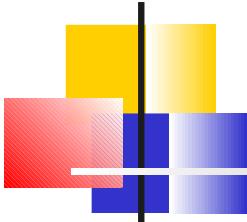
- Caracterização de dados
  - Objetos e atributos
  - Tipos de dados
- Exploração de dados
  - Dados univariados
  - Medidas de localidade, espalhamento e distribuição
  - Dados multivariados
  - Visualização de dados e de resultados



# Exercício

---

- Descrever e explorar os dados utilizados na aula de laboratório



# Perguntas

---

