

FCT/Unesp – Presidente Prudente
Departamento de Matemática e Computação

Fundamentos sobre Dados

Parte 3

Prof. Danilo Medeiros Eler
danilo.eler@unesp.br

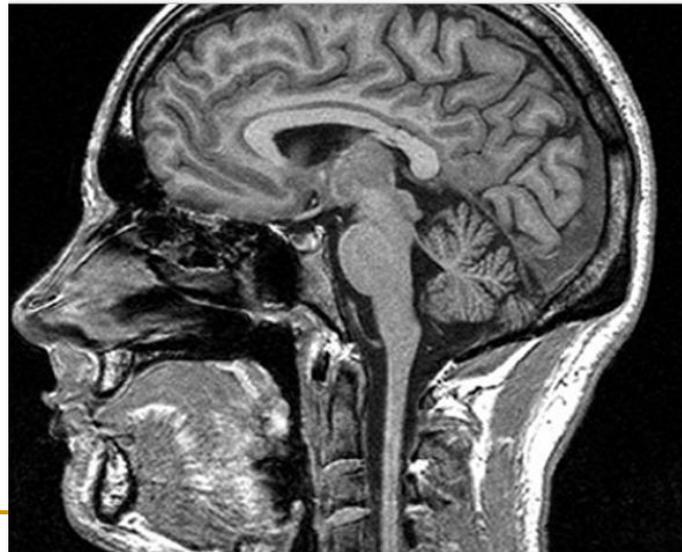
Sumário

- Tipos de Dados
- Estrutura dentro e entre instâncias
- Processamento dos dados

Processamento dos Dados

Processamento dos dados

- Em algumas circunstâncias, é preferível visualizar os dados brutos (*raw data*)
 - Por exemplo, na maioria das aplicações médicas os dados não sofrem modificações para serem visualizados, pois informações importante seriam perdidas e artefatos seriam adicionados



Processamento dos dados

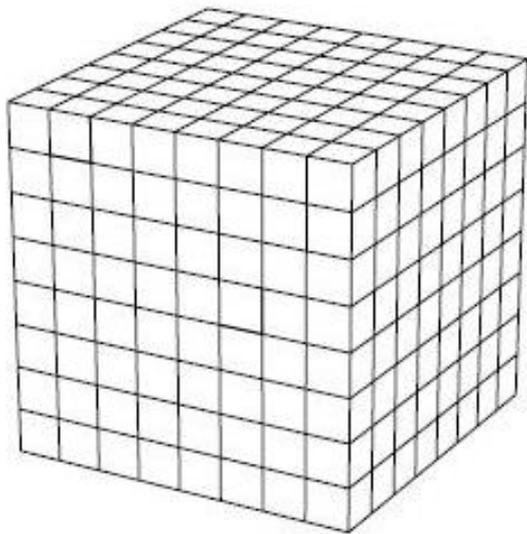
- Dependendo do tipo de dado ou da técnica de visualização a ser aplicada, os dados necessitam de pré-processamento
 - Dados faltantes, *outliers* ou erros

Processamento dos dados

- Alguns métodos para o pré-processamento dos dados são
 - Metadados
 - Estatística
 - Valores faltantes e limpeza dos dados
 - Normalização
 - Segmentação
 - Amostragem
 - Redução de Dimensionalidade
 - Agregação e Sumarização

Processamento dos dados

■ Metadados



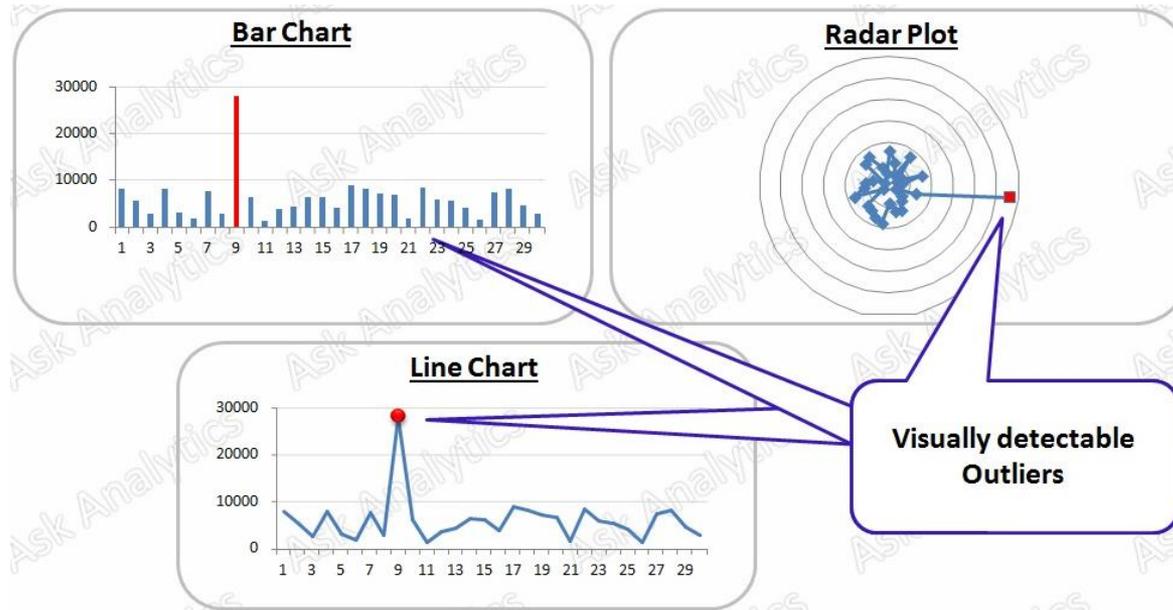
```
principal.log x footjpg.000.jpg.vvi x
1 <KWOpenFileProperties Version="1.5"
2   ClassName="vtkKWOpenFileProperties"
3   Spacing="0.95 0.95 1"
4   Origin="0 0 0"
5   ScalarType="3"
6   WholeExtent="0 101 0 246 0 199"
7   NumberOfScalarComponents="1"
8   IndependentComponents="1"
9   FileOrientation="4 2 0"
10  BigEndianFlag="0"
11  FilePattern="footjpg.%03d.jpg"
12  FileDimensionality="2"
13  Scope="2"/>
14
```

Processamento dos dados

- Metadados
 - Podem guiar o processamento dos dados
 - Fornecem informação para sua interpretação, tal como o formato dos campos de uma instância
 - Pode conter o ponto de referência base de alguma medida, unidade, símbolo ou número para indicar algum valor faltante
 - Essas informações são importantes para selecionar as operações apropriadas de processamento

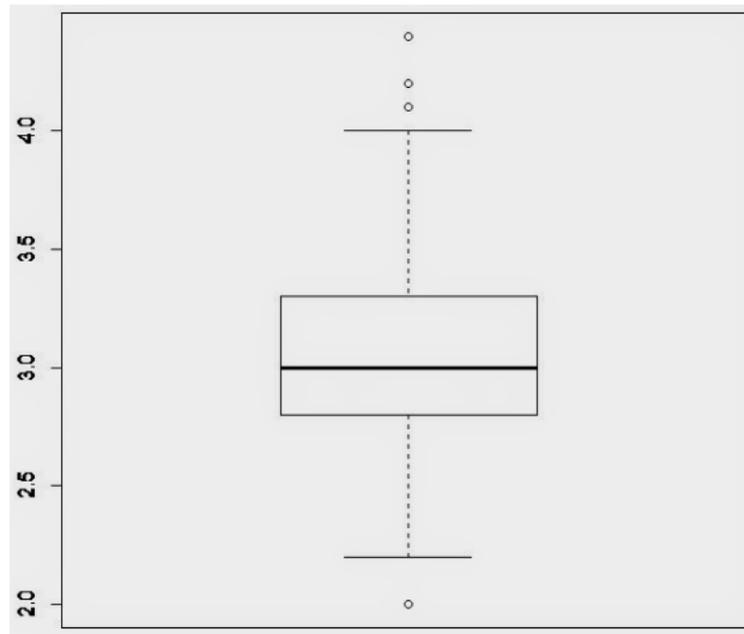
Processamento dos dados

- **Métodos de análise estatística** podem fornecer *insights* úteis
 - Detecção de *outlier*
 - Podem indicar instâncias com valores errados em determinados campos



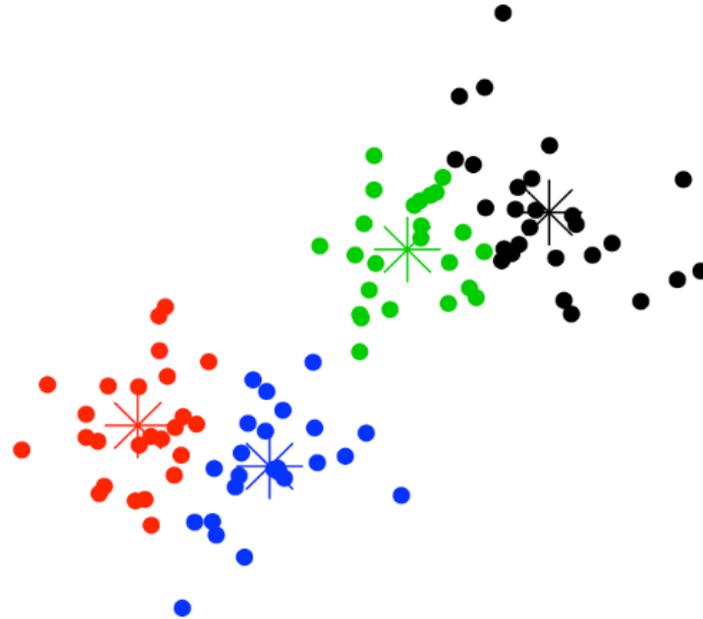
Processamento dos dados

- **Métodos de análise estatística** podem fornecer *insights* úteis
 - Detecção de *outlier*
 - Podem indicar instâncias com valores errados em determinados campos



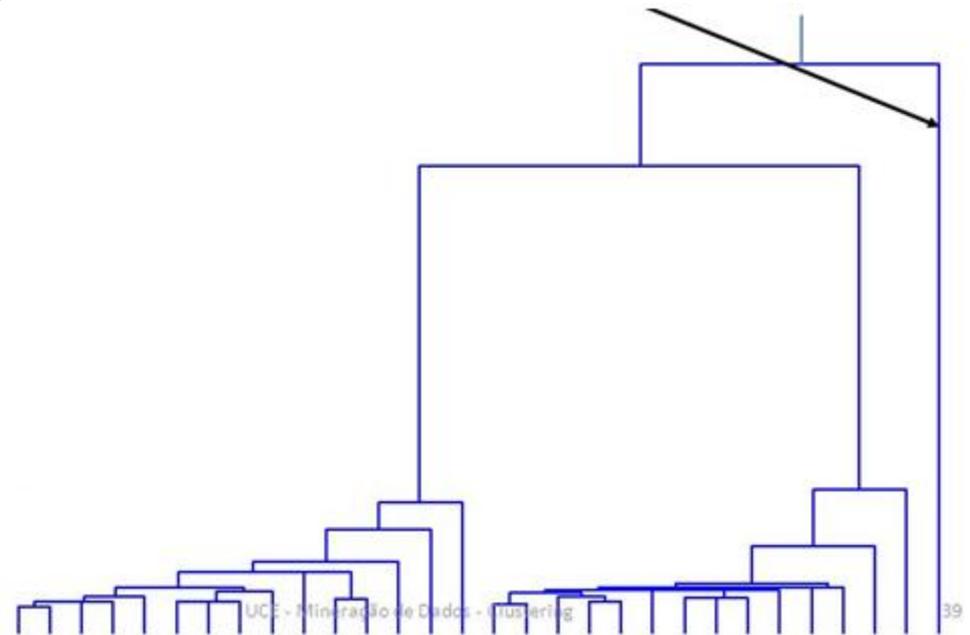
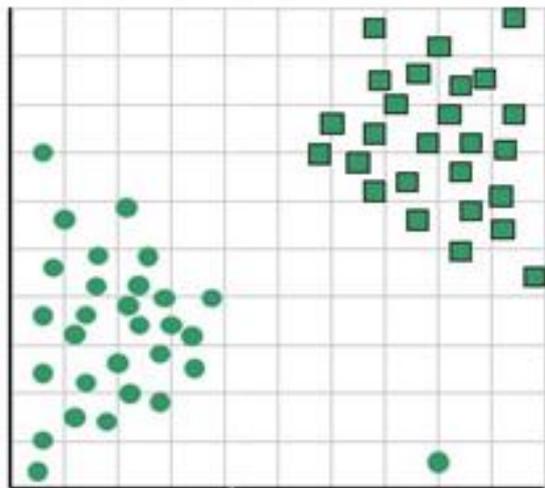
Processamento dos dados

- **Métodos de análise estatística** podem fornecer *insights* úteis
 - Análise de agrupamentos
 - Pode auxiliar na segmentação de um conjunto de dados em grupos muito similares



Processamento dos dados

- **Métodos de análise estatística** podem fornecer *insights* úteis
 - Análise de agrupamentos
 - Pode auxiliar na segmentação de um conjunto de dados em grupos muito similares



Processamento dos dados

- **Métodos de análise estatística** podem fornecer *insights* úteis
 - Análise de correlação
 - Pode auxiliar a eliminar campos redundantes ou destacar associação entre dimensões

5 X 5 Matrix

	Var1	Var2	Var3	Var4	Var5
Var1	1	-0.08071	0.098675	0.014625	0.061913
Var2	-0.08071	1	-0.10168	0.37678	0.103062
Var3	0.098675	-0.10168	1	0.049934	0.119171
Var4	0.014625	0.37678	0.049934	1	0.002249
Var5	0.061913	0.103062	0.119171	0.002249	1

Processamento dos dados

- Conjuntos de dados reais, geralmente, possuem **dados faltantes ou errôneos**
 - Por exemplo, o mal funcionamento de um sensor, uma entrada em branco em uma pesquisa ou omissão de algum dado
- Quando o valor de um atributo possui erro, frequentemente foi causado por uma falha humana e é difícil de detectar

Processamento dos dados

- Algumas estratégias para lidar com esses problemas são
 - Descartar a instância com erros
 - Associar um valor sentinela
 - Associar um valor médio
 - Associar um valor baseado nos vizinhos mais próximos

Identificador	Salário	Idade	Peso	Nível	Aprovado
P1	1500,89	30	87,6	1	0
P2	789,52	48	null	2	0
P3	1000,00	28	70,5	2	1
P4	589,36	39	90,5	3	1

Processamento dos dados

- Descartar a instância com erros
 - É uma medida drástica, mas é frequentemente praticada, desde que a qualidade das instâncias restantes seja significativa para análise

Identificador	Salário	Idade	Peso	Nível	Aprovado
P1	1500,89	30	87,6	1	0
P2	789,52	48	null	2	0
P3	1000,00	28	70,5	2	1
P4	589,36	39	90,5	3	1

Processamento dos dados

- Descartar a instância com erros
 - Pode levar a uma grande perda de informação, especialmente em conjuntos de dados com muitos dados faltante ou com erros
 - Além disso, as instâncias com dados faltantes podem ser as mais interessantes

Identificador	Salário	Idade	Peso	Nível	Aprovado
P1	1500,89	30	87,6	1	0
P2	789,52	48	null	2	0
P3	1000,00	28	70,5	2	1
P4	589,36	39	90,5	3	1

Processamento dos dados

- Associar um valor sentinela
 - Pode-se associar um valor fixo para designar o valor faltante ou com erro
 - Por exemplo, se os dados variam de 0 a 100, pode-se escolher um outro valor
 - Por exemplo: -5

Identificador	Salário	Idade	Peso	Nível	Aprovado
P1	1500,89	30	87,6	1	0
P2	789,52	48	-5	2	0
P3	1000,00	28	70,5	2	1
P4	589,36	39	90,5	3	1

Processamento dos dados

- Associar um valor sentinela
 - Assim, quando os dados forem visualizados, as instâncias com valores faltantes podem ser identificadas
 - Um cuidado deve ser tomado para não levar em conta esses valores sentinela em algum processamento, como alguma medição estatística

Identificador	Salário	Idade	Peso	Nível	Aprovado
P1	1500,89	30	87,6	1	0
P2	789,52	48	-5	2	0
P3	1000,00	28	70,5	2	1
P4	589,36	39	90,5	3	1

Processamento dos dados

- Associar um valor médio
 - Uma estratégia simples é substituir o valor faltante ou errôneo por um valor médio calculado da variável em questão
 - A vantagem é que pode afetar muito pouco medidas estatísticas dessa variável

Identificador	Salário	Idade	Peso	Nível	Aprovado
P1	1500,89	30	87,6	1	0
P2	789,52	48	82,86	2	0
P3	1000,00	28	70,5	2	1
P4	589,36	39	90,5	3	1

Processamento dos dados

- Associar um valor médio
 - Entretanto, essa abordagem pode mascarar a identificação de *outliers*, principalmente se esse é o foco da exploração

Identificador	Salário	Idade	Peso	Nível	Aprovado
P1	1500,89	30	87,6	1	0
P2	789,52	48	82,86	2	0
P3	1000,00	28	70,5	2	1
P4	589,36	39	90,5	3	1

Processamento dos dados

- Associar um valor baseado nos vizinhos mais próximos
 - Uma das melhores abordagens para substituição de valores é encontrar uma instância muito similar àquela em questão, com base nas outras variáveis
 - Uma vez que a instância mais similar é encontrada, os valores faltantes ou errôneos são substituídos pelos da instância mais similar

Processamento dos dados

- Associar um valor baseado nos vizinhos mais próximos

Identificador	Salário	Idade	Peso	Nível	Aprovado
P1	1500,89	30	87,6	1	0
P2	789,52	48	null	2	0
P3	1000,00	28	70,5	2	1
P4	589,36	39	90,5	3	1

$$D(P2, P1) = \text{sqrt}((789,52 - 1500,89)^2 + (48 - 30)^2 + (2 - 1)^2 + (0 - 0)^2)$$

$$D(P2, P1) = \text{sqrt}(506047,27 + 324 + 1 + 0)$$

$$D(P2, P1) = 711,59$$

Processamento dos dados

- Associar um valor baseado nos vizinhos mais próximos

Identificador	Salário	Idade	Peso	Nível	Aprovado
P1	1500,89	30	87,6	1	0
P2	789,52	48	null	2	0
P3	1000,00	28	70,5	2	1
P4	589,36	39	90,5	3	1

$$D(P2, P3) = \text{sqrt}((789,52 - 1000,00)^2 + (48 - 28)^2 + (2 - 3)^2 + (0 - 1)^2)$$

$$D(P2, P3) = \text{sqrt}(44301,83 + 400 + 0 + 1)$$

$$D(P2, P3) = 211,43$$

Processamento dos dados

- Associar um valor baseado nos vizinhos mais próximos

Identificador	Salário	Idade	Peso	Nível	Aprovado
P1	1500,89	30	87,6	1	0
P2	789,52	48	null	2	0
P3	1000,00	28	70,5	2	1
P4	589,36	39	90,5	3	1

$$D(P2, P4) = \text{sqrt}((789,52 - 589,36)^2 + (48 - 39)^2 + (2 - 3)^2 + (0 - 1)^2)$$

$$D(P2, P4) = \text{sqrt}(40064,02 + 81 + 1 + 1)$$

$$D(P2, P4) = 200,36$$

Processamento dos dados

- Associar um valor baseado nos vizinhos mais próximos

Identificador	Salário	Idade	Peso	Nível	Aprovado
P1	1500,89	30	87,6	1	0
P2	789,52	48	90,5	2	0
P3	1000,00	28	70,5	2	1
P4	589,36	39	90,5	3	1

$$D(P2, P1) = 711,59$$

$$D(P2, P3) = 211,43$$

$$D(P2, P4) = 200,36$$

Processamento dos dados

- Associar um valor baseado nos vizinhos mais próximos
 - Um problema é decidir se a utilização de todos os atributos para calcular a similaridade é uma boa estratégia ou se a melhor seria selecionar um subconjunto de atributos mais relevantes

Identificador	Salário	Idade	Peso	Nível	Aprovado
P1	1500,89	30	87,6	1	0
P2	789,52	48	90,5	2	0
P3	1000,00	28	70,5	2	1
P4	589,36	39	90,5	3	1



Processamento dos dados

- **Normalização** é o processo de transformar um conjunto de dados para que seja satisfeita uma propriedade estatística particular
 - As variáveis de um conjunto de dados podem estar em uma escala muito diferente

Identificador	Salário	Idade	Peso	Nível	Aprovado
P1	1500,89	30	87,6	1	0
P2	789,52	48	60,0	2	0
P3	1000,00	28	70,5	2	1
P4	589,36	39	90,5	3	1

Processamento dos dados

■ Normalização

- Um exemplo simples é a transformação da abrangência dos valores de dados para assumirem valores entre 0.0 e 1.0

Identificador	Salário	Idade	Peso	Nível	Aprovado
P1	1500,89	30	87,6	1	0
P2	789,52	48	60,0	2	0
P3	1000,00	28	70,5	2	1
P4	589,36	39	90,5	3	1

Processamento dos dados

■ Normalização

- Um exemplo simples é a transformação da abrangência dos valores de dados para assumirem valores entre 0.0 e 1.0
 - No exemplo, normalização pelo máximo

Identificador	Salário	Idade	Peso	Nível	Aprovado
P1	1,00	0,62	0,96	0,33	0,00
P2	0,52	1,00	0,66	0,66	0,00
P3	0,66	0,58	0,77	0,66	1,00
P4	0,39	0,81	1,00	1,00	1,00

Processamento dos dados

■ Não Normalizado

ID	Salário	Idade	Peso	Nível	Aprovado
P1	1500,89	30	87,6	1	0
P2	789,52	48	null	2	0
P3	1000,00	28	70,5	2	1
P4	589,36	39	90,5	3	1

$$D(P2, P1) = 711,59$$

$$D(P2, P3) = 211,43$$

$$D(P2, P4) = 200,36$$

■ Normalizado

ID	Salário	Idade	Peso	Nível	Aprovado
P1	1,00	0,62	0,96	0,33	0,00
P2	0,52	1,00	null	0,66	0,00
P3	0,66	0,58	0,77	0,66	1,00
P4	0,39	0,81	1,00	1,00	1,00

$$D(P2, P1) = 0,69$$

$$D(P2, P3) = 1,09$$

$$D(P2, P4) = 1,07$$

Processamento dos dados

■ Não Normalizado

ID	Salário	Idade	Peso	Nível	Aprovado
P1	1500,89	30	87,6	1	0
P2	789,52	48	null	2	0
P3	1000,00	28	70,5	2	1
P4	589,36	39	90,5	3	1

$$D(P2, P1) = 711,59$$

$$D(P2, P3) = 211,43$$

$$D(P2, P4) = 200,36$$

■ Normalizado

ID	Salário	Idade	Peso	Nível	Aprovado
P1	1,00	0,62	0,96	0,33	0,00
P2	0,52	1,00	null	0,66	0,00
P3	0,66	0,58	0,77	0,66	1,00
P4	0,39	0,81	1,00	1,00	1,00

$$D(P2, P1) = 0,69$$

$$D(P2, P3) = 0,43$$

$$D(P2, P4) = 0,40$$

Processamento dos dados

■ Normalizado pelo Máximo

ID	Salário	Idade	Peso	Nível	Aprovado
P1	1,00	0,62	0,96	1,00	0,00
P2	0,52	1,00	null	0,66	0,00
P3	0,66	0,58	0,77	0,33	1,00
P4	0,39	0,81	1,00	1,00	1,00

$$D(P2, P1) = 0,69$$

$$D(P2, P3) = 1,09$$

$$D(P2, P4) = 1,07$$

■ Normalizado pelo Mínimo e Máximo

ID	Salário	Idade	Peso	Nível	Aprovado
P1	1,00	0,10	0,87	0,00	0,00
P2	0,21	1,00	null	0,50	0,00
P3	0,45	0,00	0,13	0,50	1,00
P4	0,00	0,55	1,00	1,00	1,00

$$D(P2, P1) = 1,29$$

$$D(P2, P3) = 1,43$$

$$D(P2, P4) = 1,22$$

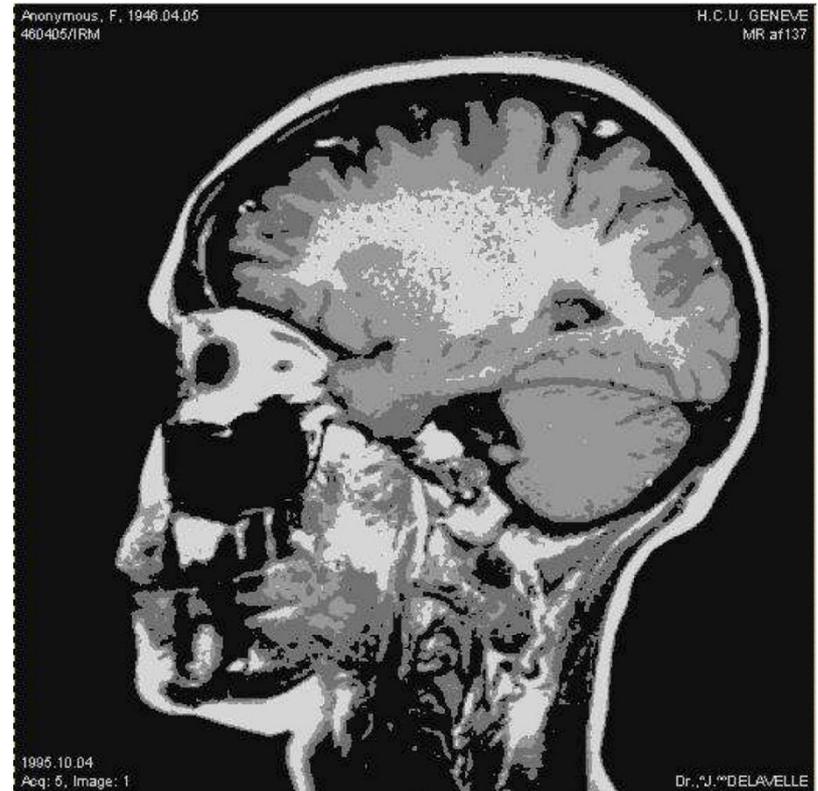
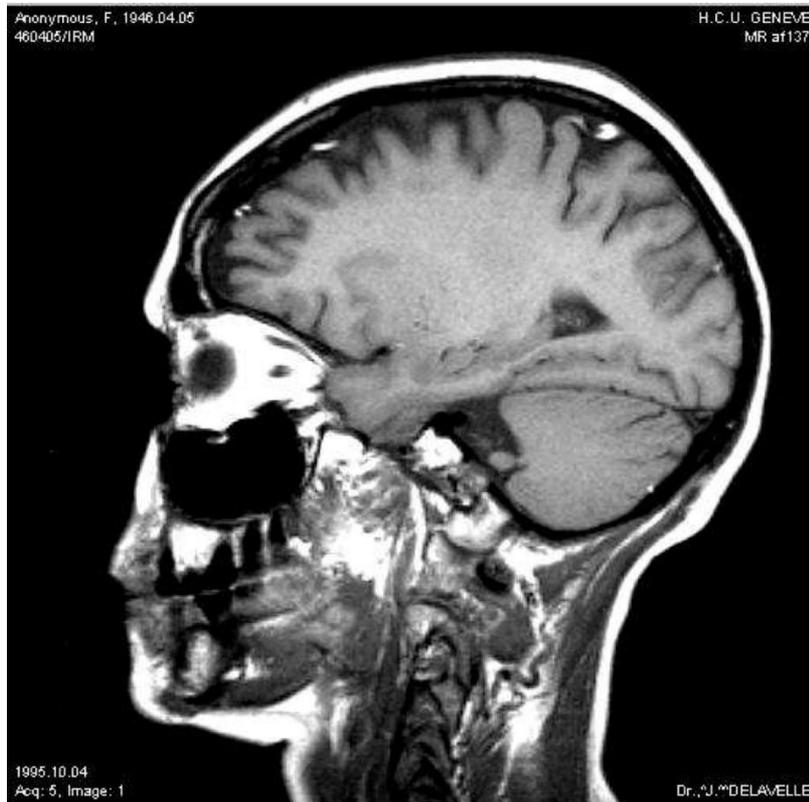
Processamento dos dados

■ Segmentação

- Em algumas situações, os dados podem estar separados em regiões contínuas, em que cada região corresponde a um particular classificação dos dados
- Por exemplo, em imagens de ressonâncias magnética, um conjunto de dados pode ter 256 possíveis valores para cada ponto, e pode ser segmentado em categorias específicas, tais como pele, músculo, ossos e gordura
 - Como pode haver ambiguidade, deve ser levado em consideração a vizinhança das regiões

Processamento dos Dados

■ Segmentação



Processamento dos Dados

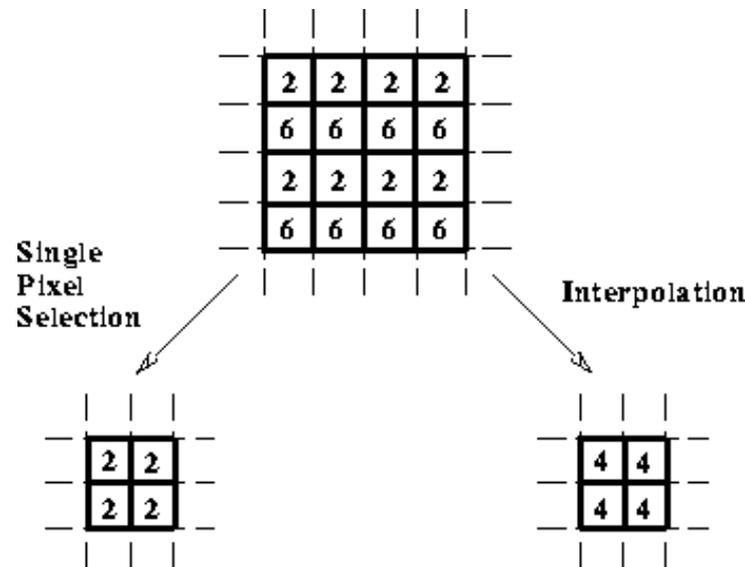
■ Amostragem

- Pode ser utilizada para reduzir o número de elementos que serão utilizados durante o processo de exploração
- Geralmente é aplicada algumas restrições e as instâncias que as satisfazem são selecionadas como amostras

Processamento dos Dados

■ Amostragem

- Em algumas situações é necessário transformar os dados de uma distribuição espacial para outra com uma resolução diferente
 - Para isso, é necessário fazer uma re-amostragem dos dados, com base nas instâncias de uma vizinhança
 - Geralmente, é aplicado algum processo de interpolação



Processamento dos Dados

■ Redução de Dimensionalidade

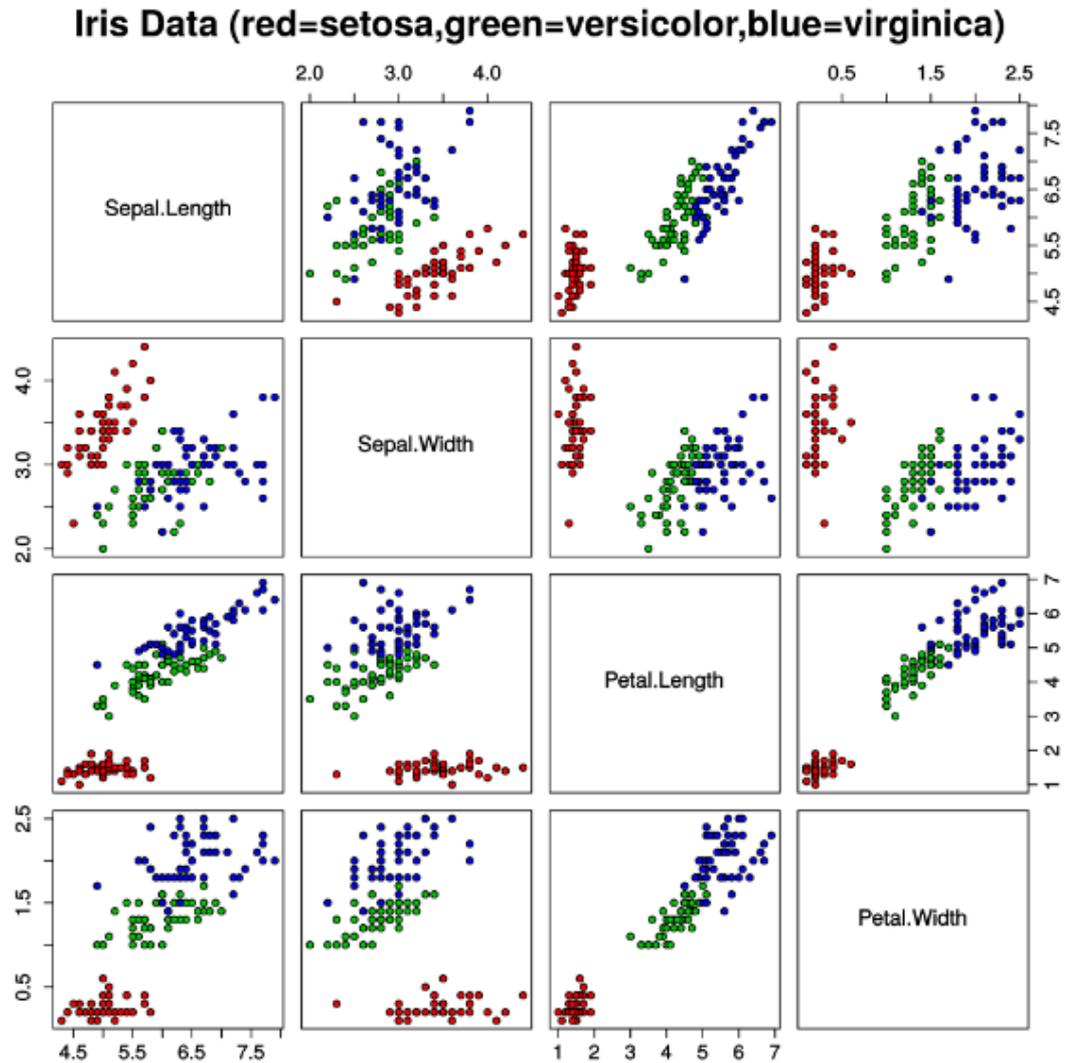
- É um processamento empregado em situações em que a dimensionalidade dos dados excede as capacidades das técnicas de análise de dados
- Assim, é necessário investigar meios de reduzir a dimensionalidade dos dados, tentando preservar o máximo possível a informação contida neles

Processamento dos Dados

- A **Redução de Dimensionalidade** pode ser realizada manualmente ou computacionalmente
- O analista pode selecionar atributos de interesse ou os mais relevantes
 - Técnicas automáticas também podem ser empregadas

Processamento dos Dados

- Exemplo de seleção de atributos dois a dois

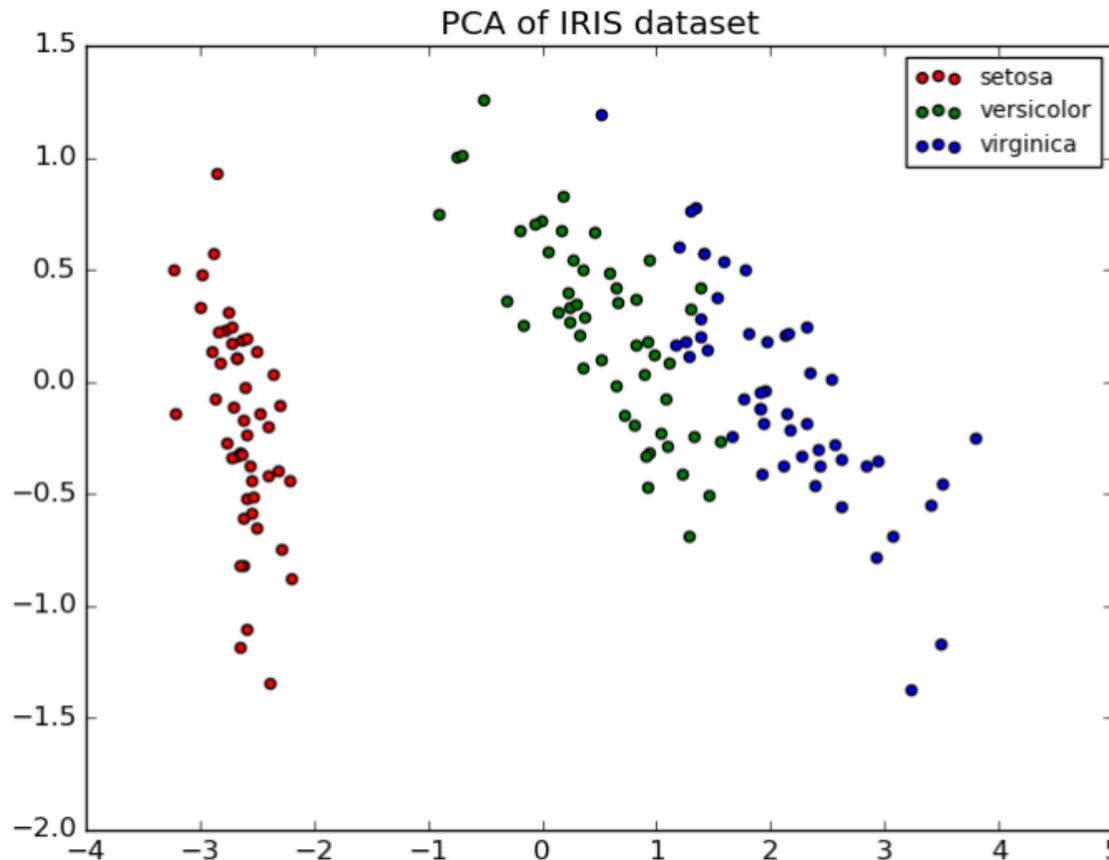


Processamento dos Dados

- Algumas técnicas podem reduzir a dimensionalidade preservando as relações e estruturas do espaço original
 - Ex.: Projeções Multidimensionais
- Para tanto, pode-se utilizar técnicas como
 - *Principal Component Analysis* (PCA)
 - *Multidimensional Scaling* (MDS)
 - *Self Organizing Maps* (SOM)
 - *Fastmap*

Processamento dos Dados

- Exemplo de redução de dimensionalidade



Processamento dos Dados

■ Mapear variáveis nominais para números

- Em alguns domínios os valores das dimensões são nominais
- Algumas alternativas podem ser empregadas para mapear esses valores para números
- Se o valor nominal for ranqueado, o mapeamento é direto, pois há uma relação de ordem
 - Por Exemplo, tamanho de camisas
 - $P = 0$
 - $M = 1$
 - $G = 2$
 - $GG = 3$

Processamento dos Dados

- **Mapear variáveis nominais para números**
 - Deve-se encontrar um mapeamento dos dados para elementos gráficos que não introduzam relacionamentos artificiais que não existam nos dados
 - Por exemplo, em um conjunto de dados de carros, o atributo marca é nominal
 - Se um valor inteiro for associado para cada marca, um falso relacionamento poderá prejudicar a análise
 - Honda = 0
 - VW = 1
 - Nissan = 2
 - Toyota = 3

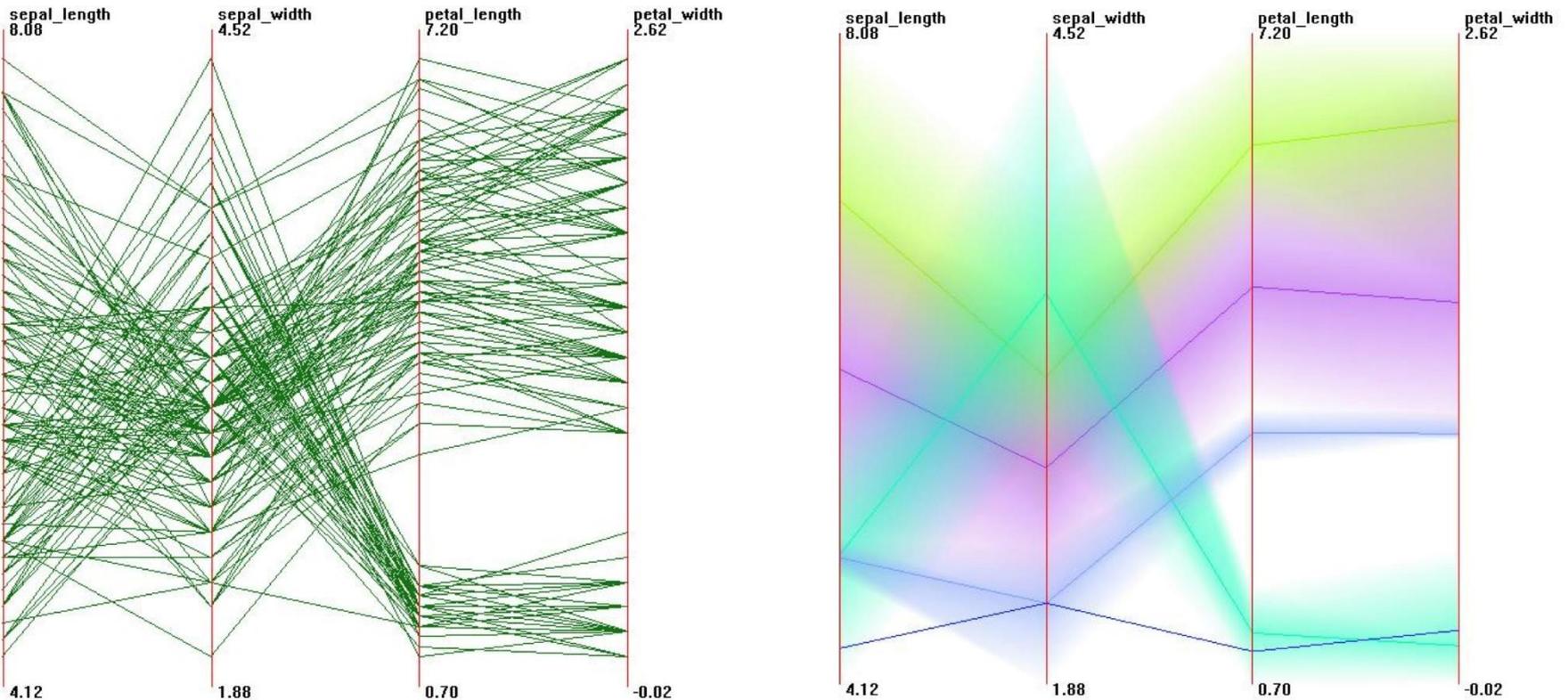
Processamento dos Dados

■ **Agregação e Sumarização**

- A visualização pode ficar sobrecarregada quando muitos dados são apresentados, havendo muita sobreposição
 - Uma alternativa é agrupar instâncias
- Primeiramente deve-se definir o método que executará a agregação e depois como o grupo será representado na visualização
- Deve-se exibir informação suficiente para o usuário decidir se ele deverá continuar a explorar um determinado grupo

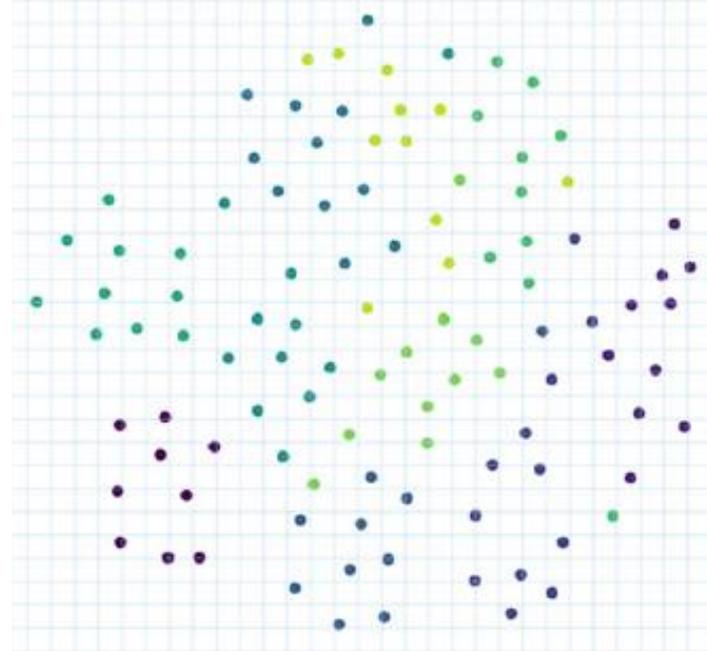
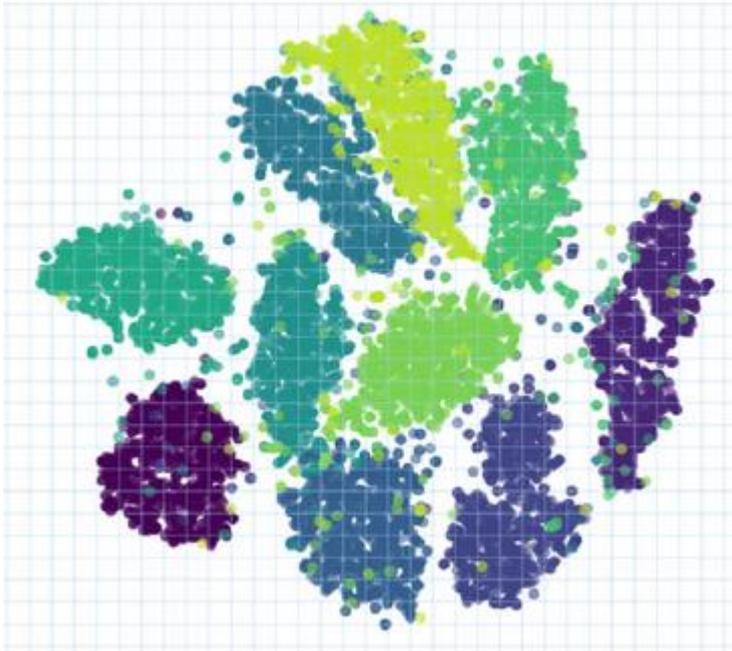
Processamento dos Dados

- Coordenadas paralelas original (esquerda) e coordenadas paralelas apresentando agregação dos dados (direita)



Processamento dos Dados

- Conjunto original (esquerda) e Conjunto Amostrado (direita)
 - Técnica SADIRE



Referências

- Ward, M., Grinstein, G. G., Keim, D. Interactive data visualization foundations, techniques, and applications. Natick, Mass., A K Peters, 2010.
 - Capítulo 2

Referências

- SADIRE: a context-preserving sampling technique for dimensionality reduction visualizations
 - Wilson Estécio Marcilio-Jr, Danilo Medeiros Eler
 - Journal of Visualization (2020)
- Aulas de visualização da wiki.icmc.usp.br
 - Prof. Dr. Fernando Paulovich (ICMC/USP)
 - Profa. Dra. Maria Cristina Ferreira de Oliveira (ICMC/USP)
 - Profa. Dra. Rosane Minghim (ICMC/USP)