Data Mining
Análise de Grupos

Prof. Dr. Joaquim Assunção

DEPARTAMENTO DE COMPUTAÇÃO APLICADA
CENTRO DE TECNOLOGIA
UFSM
2019

www.inf.ufsm.br/~joaquim

Fair user agreement

Este material foi criado para a disciplina de Mineração de Dados - Centro de Tecnologia da UFSM.

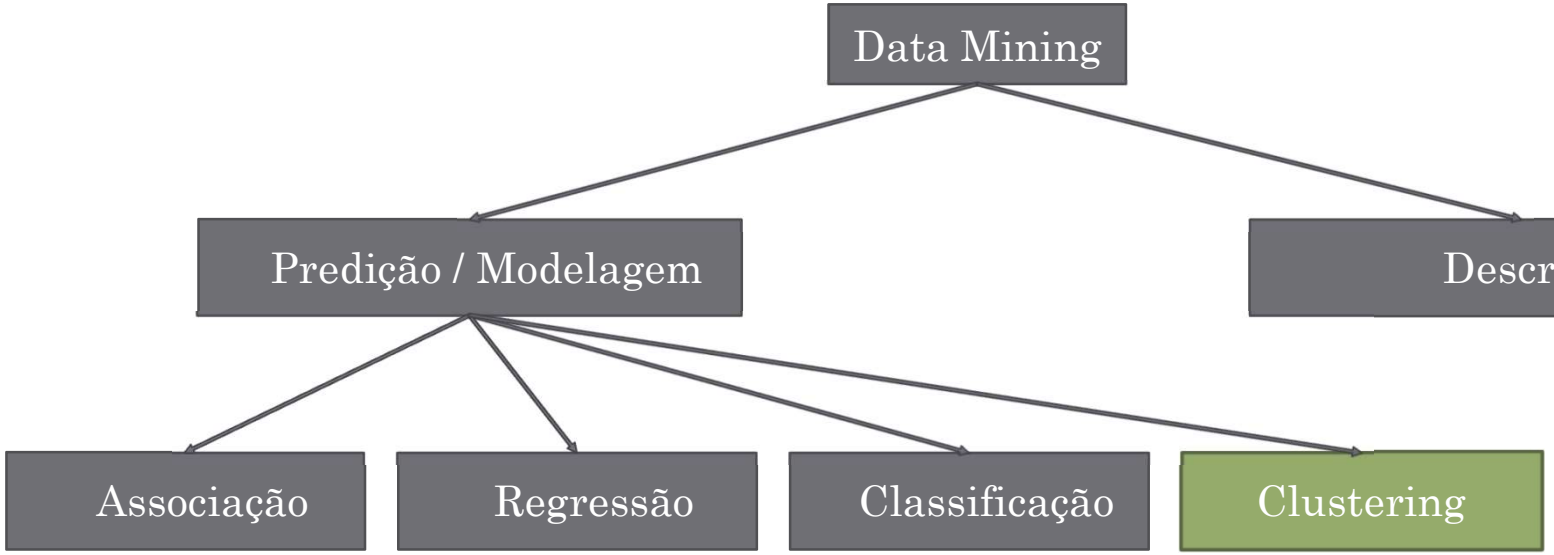
Você pode usar este material livremente*; porém, caso seja usado em outra instituição, **me envie um e-mail** avisando o nome da instituição e a disciplina.

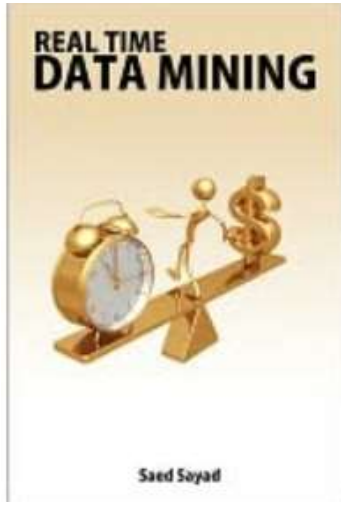
*Caso você queira usar algo desse material em alguma publicação, envie-me um e-mail com antecedência.

Prof. Dr. Joaquim Assunção.

joaquim@inf.ufsm.br

Mapa para Mineração de Dados*

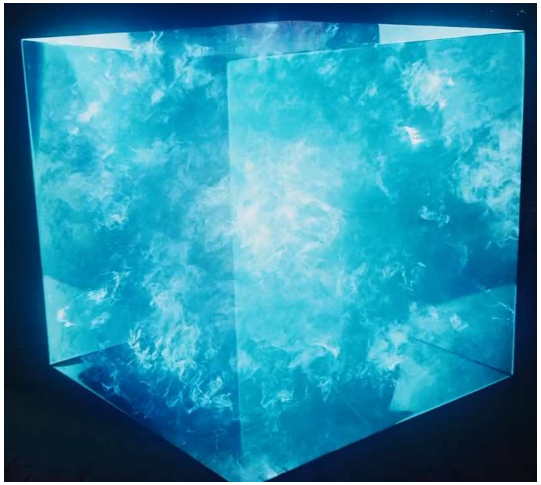




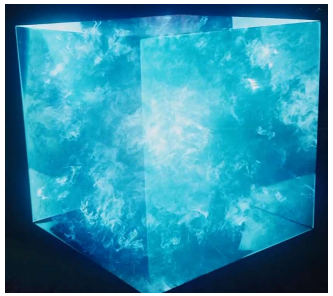
***http://www.saedsayad.com/data_mining_map.htm**

Um conto com final alternativo...


Este indivíduo,
roubou aquele cubo

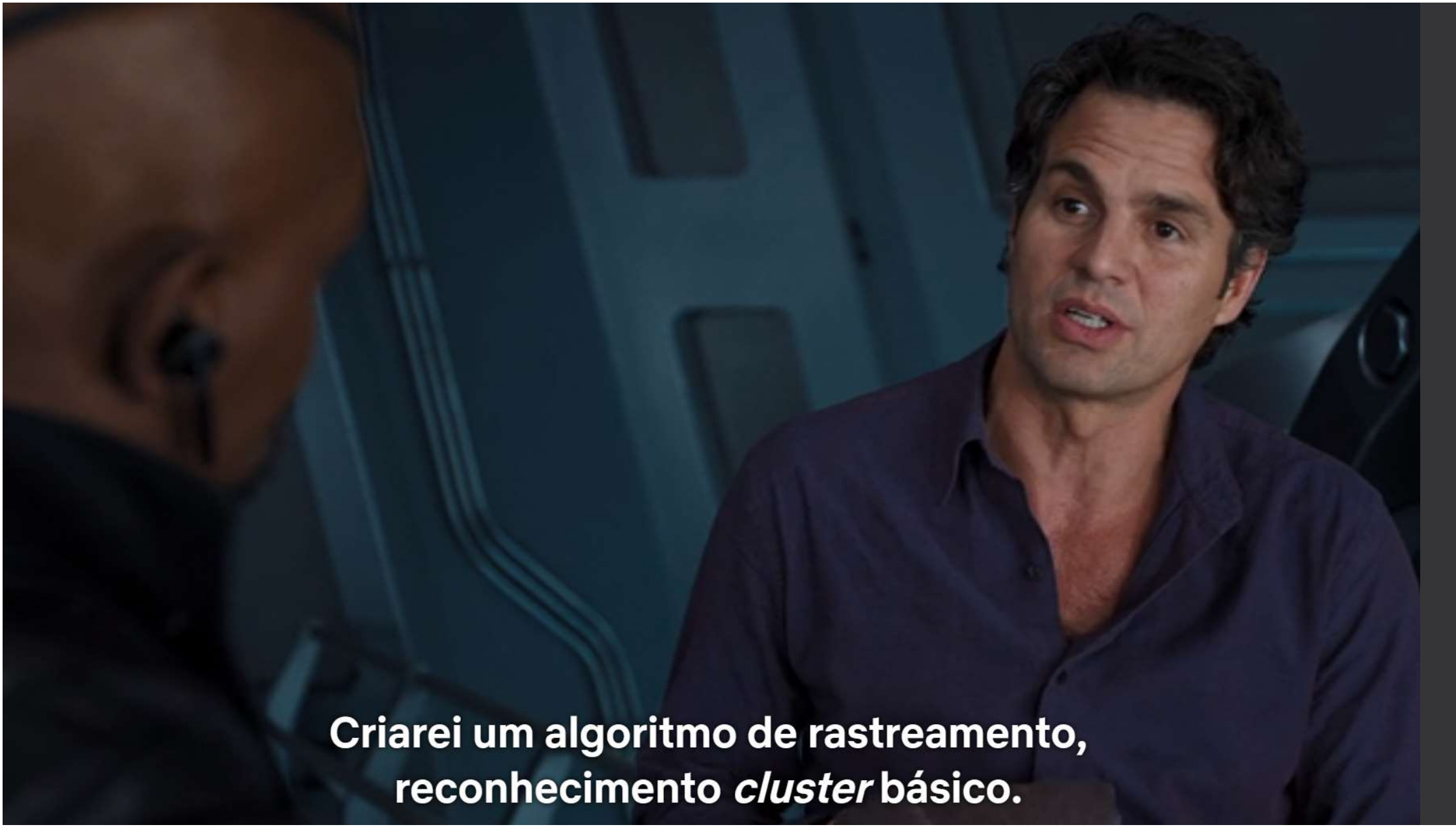


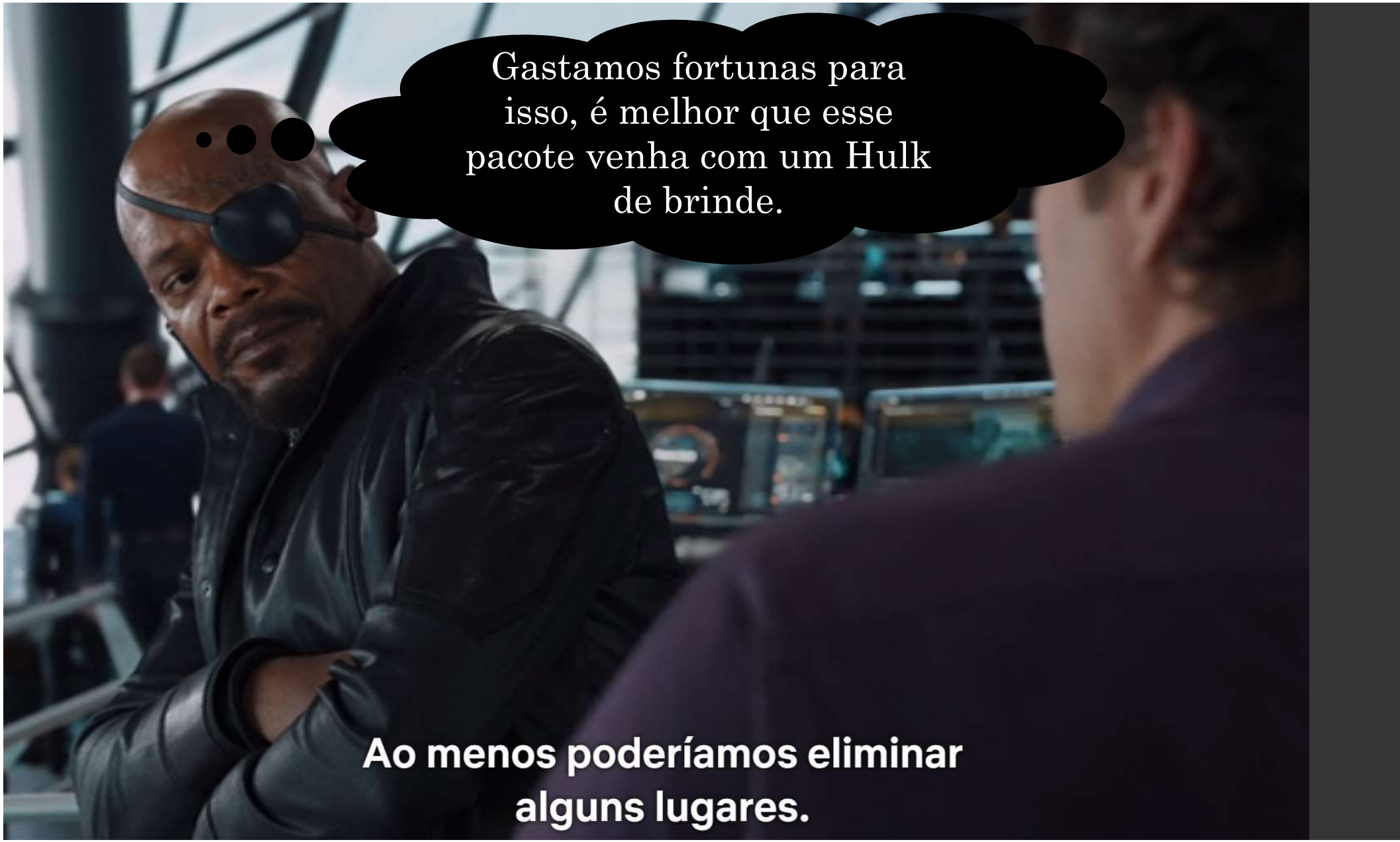
Aparentemente, isso deixa rastros de raios gama...



...E o diretor tem acesso a todos os espectrômetros do ~~mundo~~ planeta terra.

...Então, o doutor Banner tem acesso a todos os dados de radiação gama.


Criarei um algoritmo de rastreamento,
reconhecimento *cluster* básico.

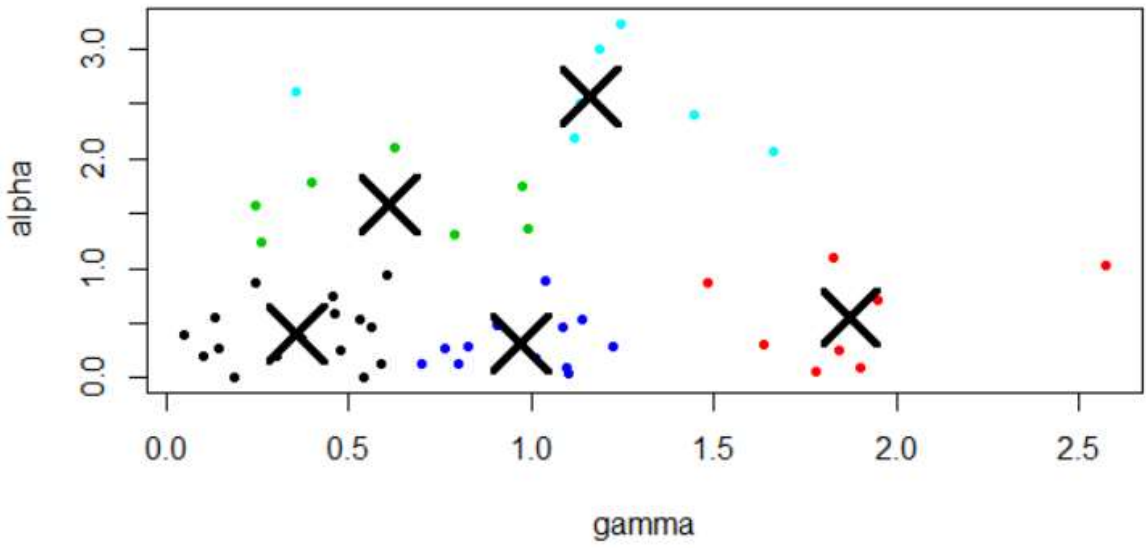

Gastamos fortunas para isso, é melhor que esse pacote venha com um Hulk de brinde.
Ao menos poderíamos eliminar alguns lugares.

Doutor Banner então coleta dados dos espectrômetros e coleta os seguintes dados.

`Baixe: "fakeBannerData"`

	gamma	alpha
1	1.13105465	2.502688827
2	0.52723426	0.536600346
3	1.66599090	2.071380937
4	1.13920064	0.538220622
5	0.14362323	0.275795692
6	1.09955094	0.047498389
7	0.90351643	0.474524377
8	1.48377949	0.864707838
9	1.95072101	1.713167149
10	0.79760066	0.124133868
11	1.84326625	0.255161459
12	1.24642391	3.234397240

Aqueles em vermelho estão fora do padrão, Vamos verificar onde estão...

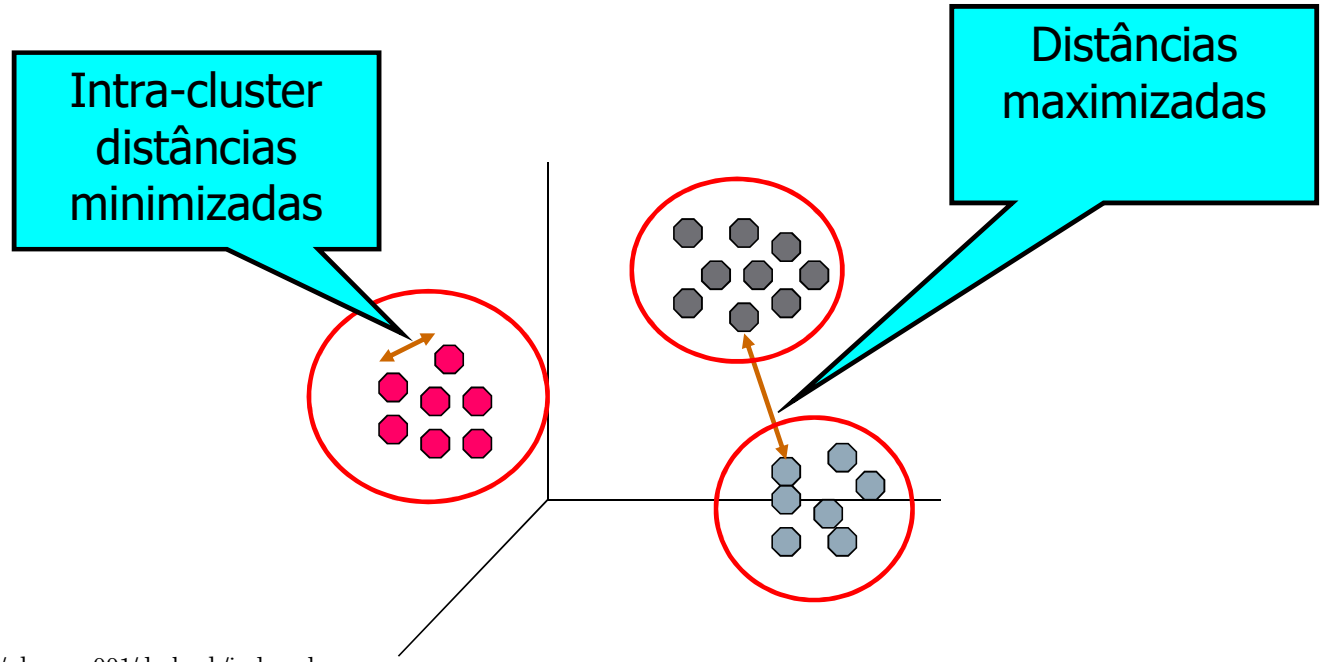


...Após algumas buscas, Dr. Banner encontra Dr.Selvig, o time captura o Tesseract, o portal nunca abre, Hulk smash ... Fim.

The end

O que é uma análise de Grupo?

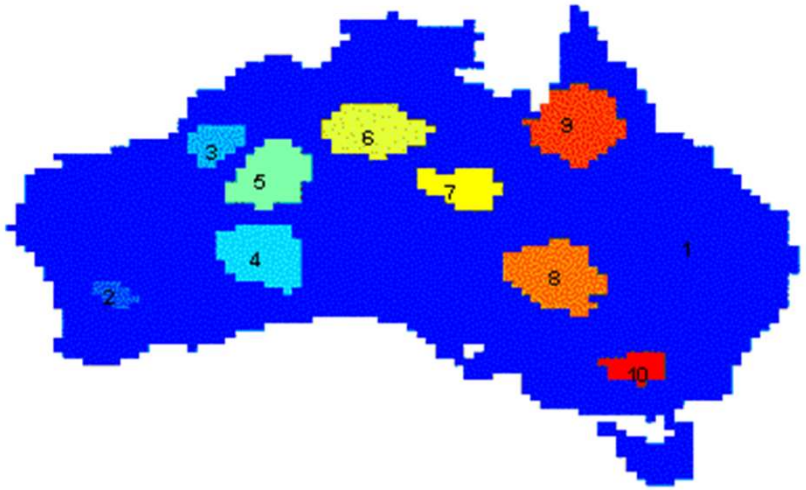
- Encontrar grupos de objetos de modo em que os objetos do grupo sejam similares (ou relacionados) um com o outro de acordo com alguma característica. Os demais grupos não devem ter essas características.



Fonte: https://www-users.cs.umn.edu/~kumar001/dmbook/index.php

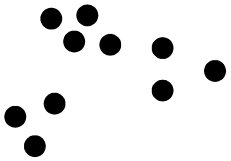
Exemplo

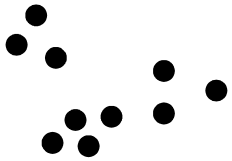
- *Clustering* precipitação na Austrália. Regiões agrupadas possuem um perfil chuvoso similar, ao passo que as demais são diferentes.



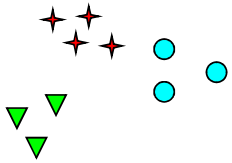
Fonte: https://www-users.cs.umn.edu/~kumar001/dmbook/index.php

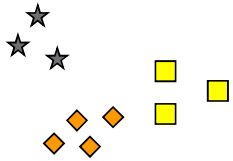
Noção de grupos pode ser ambíguo.



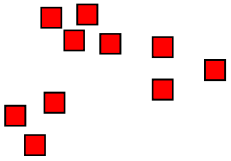


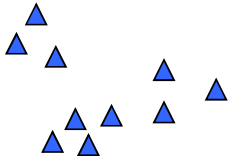
Quantos?



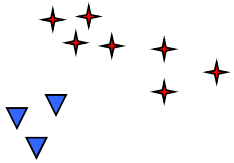


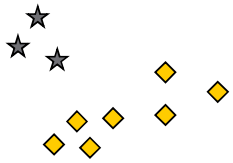
Seis grupos





Dois grupos





Quatro grupos

Fonte: https://www-users.cs.umn.edu/~kumar001/dmbook/index.php

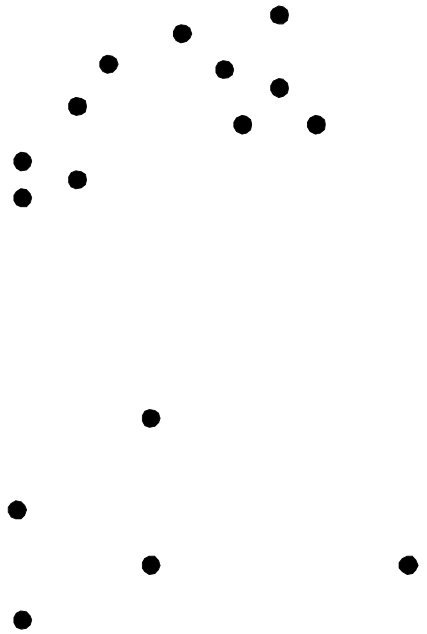
Tipos de agrupamentos

- Um agrupamento (clustering) é um conjunto de grupos (cluster)
- Há *clustering* hierarquico e particional

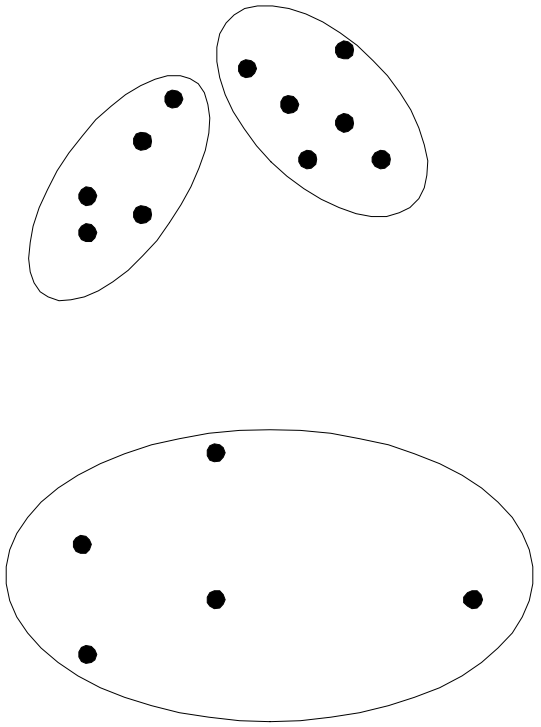
- *Partitional Clustering*
  - Uma divisão de objetos cujos grupos não se sobrepõem, de modo que cada objeto de dado está em somente um grupo.
- *Hierarchical clustering*
  - Um conjunto de grupos alinhados em uma árvore hierárquica.

Fonte: https://www-users.cs.umn.edu/~kumar001/dmbook/index.php

Agrupamento por partição



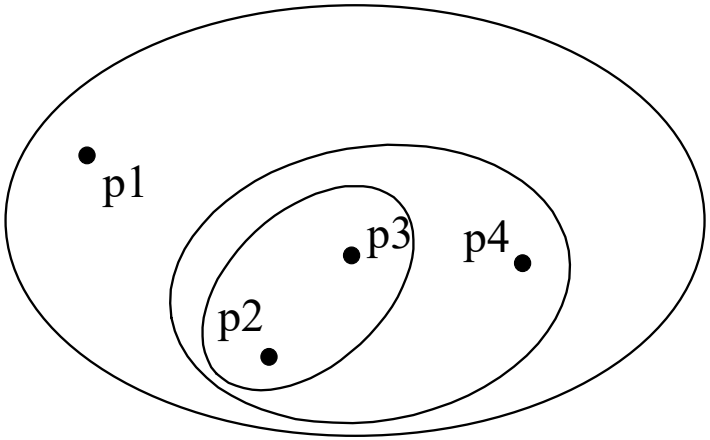
Pontos originais



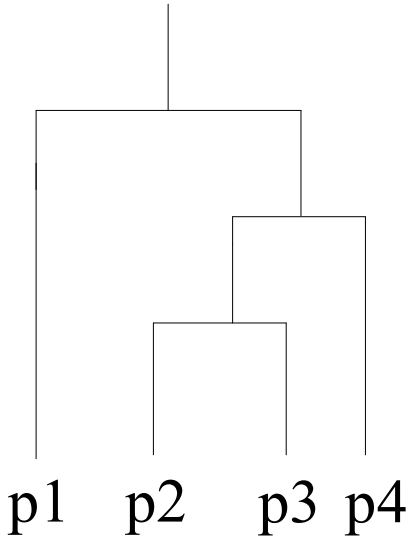
Agrupamento por partição

Fonte: https://www-users.cs.umn.edu/~kumar001/dmbook/index.php

Agrupamento hierarquico



Agrupamento hierarquico tradicional



Dendrograma tradicional

Fonte: https://www-users.cs.umn.edu/~kumar001/dmbook/index.php

Classificação geral dos agrupamentos

- **Exclusivo ou interseccionado**

Agrupamentos exclusivos são aqueles em que um ponto pertence somente à um grupo. Já os interseccionados estão ligados a dois ou mais grupos ao mesmo tempo. Por exemplo; elementos centrais em um diagrama de Venn são objetos pertencentes a múltiplos grupos. Um grupamento interseccionado.

Classificação geral dos agrupamentos

- **Parcial ou Completo**

Casos em que parte dos dados vão para um agrupamento são denominados parciais.

Classificação geral dos agrupamentos

- **Difuso ou Não-difuso**

Em um agrupamento difuso, um ponto pertence para grupos com alguma probabilidade.

Classificação geral dos agrupamentos

- **Heterogêneo ou Homogêneo**

Neste caso, heterogeneidade refere-se a diferentes formas, tamanhos e densidades.

K-means

- O k-means é um dos algoritmos mais conhecidos de *clustering*. Para particionar um conjunto X de *n* itens, o k-means precisa de um parâmetro k para indicar o número de clusters a serem formados.

K-means

- Centros são formados aleatoriamente e todas as instâncias são atribuídas ao centro do cluster mais próximo de acordo com a distância euclidiana*.

K-means

- Em seguida, o centroide, ou a média, das instâncias em cada cluster é calculado (“means”). Estes centroides são usados para serem os novos valores para cada respectivo cluster.
- Essa operação de atualizar o centroide continua, recursivamente, até que o centro do cluster esteja estabilizado.

K-means - exemplo

- No exemplo a seguir usamos o conjunto de dados Iris. Este conjunto de dados possui 150 amostras de 3 espécies de flores.
- O conjunto possui o comprimento e largura das sépalas e pétalas de cada uma das flores. Este é um típico caso em que um algoritmo de agrupamento pode ser útil, pois podemos definir a espécie da flor de acordo com estas informações.

Hands on!

- Carregue o conjunto de dados:

```
data("iris")
head(iris)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
```

Hands on!

- Para fazer o agrupamento, simplesmente carregamos `kmeans` passando como parâmetro o conjunto de dados o valor k:

```
cluster <- kmeans(iris[,c(3,4)],3)
```

Hands on!

```
cluster <- kmeans(iris[,c(3,4)],3)
```

- Neste caso usamos os dados das pétalas para gerar os grupos.
- Também usamos o parâmetro k = 3 porque sabemos que existem 3 espécies no conjunto. O plot a seguir é gerado usando a variável cluster gerada pelo k-means.

Hands on!

```
cluster <- kmeans(iris[,c(3,4)],3)
```

- Neste caso usamos os dados das pétalas para gerar os grupos.
- Também usamos o parâmetro k = 3 porque sabemos que existem 3 espécies no conjunto. O plot a seguir é gerado usando a variável cluster gerada pelo k-means.

Hands on!

```
cluster <- kmeans(iris[,c(3,4)],3)
```

```
plot(iris[,c(3,4)],
     col = cluster$cluster,
     pch = 20, cex = 1)
points(cluster$centers, pch = 4, cex = 4, lwd = 4)
```

K-means

- Pelo gráfico podemos ver os três clusters gerados.
- Dá para ver que um dos clusters foi facilmente identificado e outros dois ficaram bem próximos.
- Em nosso conjunto de dados temos 50 amostras de cada espécie. Podemos fazer uma verificação rápida contando quantos registros ficaram em cada cluster.

```
paste(sum(cluster$cluster==1),sum(cluster$clu
ster==2),sum(cluster$cluster==3))
```

K-means

- Por intuição podemos dizer que o cluster mais afastado é o que teve 50 registros. Mas como saber ao certo?
- Para isso podemos obter os centroides de cada cluster. Usamos a propriedade `centers`

Hands on!

- Use o k-means para criar grupos sobre os dados de "`fakeBannerData`". Crie plots com 5 e 6 clusters.

Exercícios

1) Qual a finalidade de um algoritmo de agrupamento? Cite uma possível característica para agrupar dados.

2) É correto afirmar que um algoritmo de cluster é um algoritmo de classificação não supervisionado? Justifique sua resposta.

3) Podemos dizer que um grupo bem separado também é um agrupamento particional?