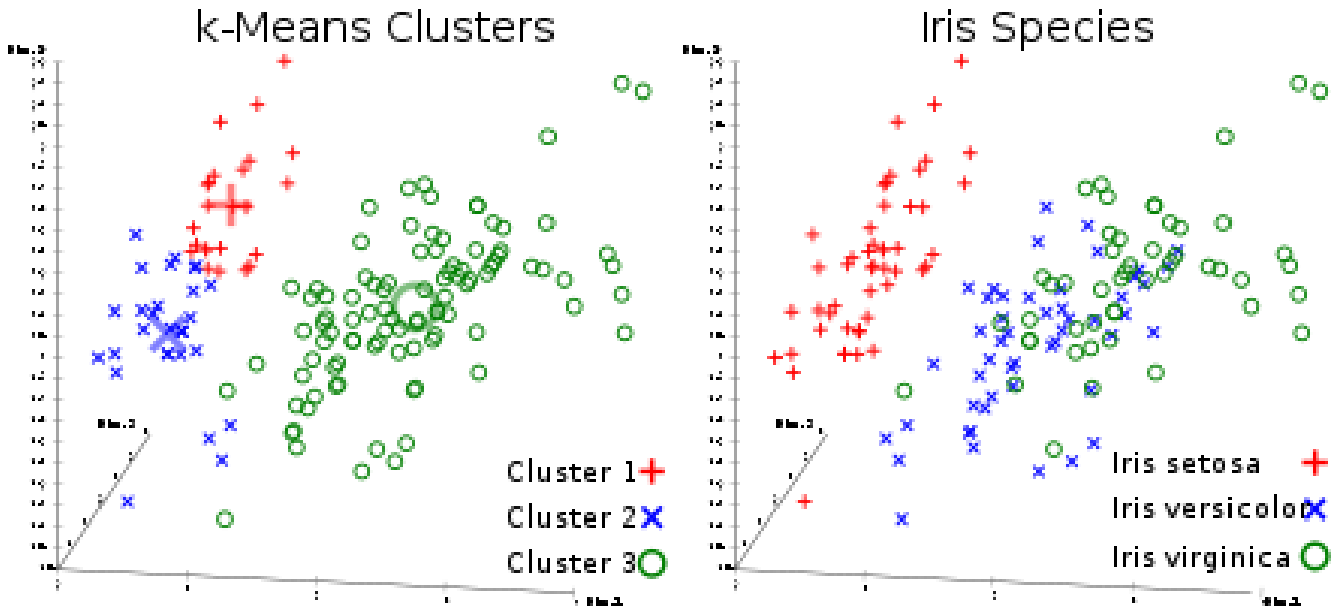


Business Intelligence

Knowledge Discovery in Databases

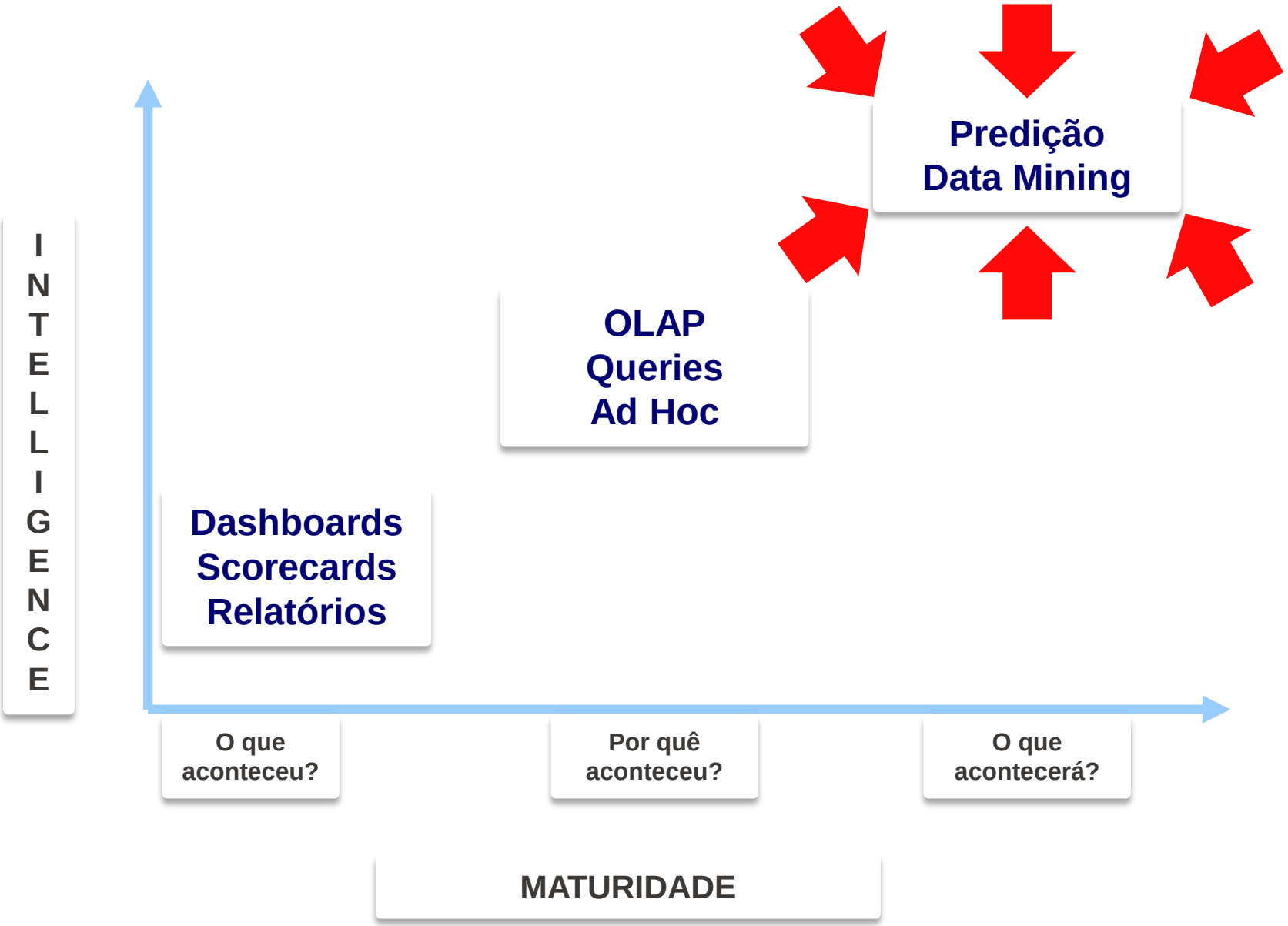
Prof. Leandro Guerra

E-mail: leandro.guerra@outspokenmarket.com.br **IG: @leandrowar**


“ Don’t ever let someone tell you
that you can’t do something.
Not even me. You got a dream,
you gotta protect it.
When people can’t do
something themselves,
they’re gonna tell you
that you can’t do it.
You want something,
go get it. Period.”
~ Will Smith
(The Pursuit of Happiness, film)

Business Intelligence
Nível de Maturidade

Relembrar é viver...

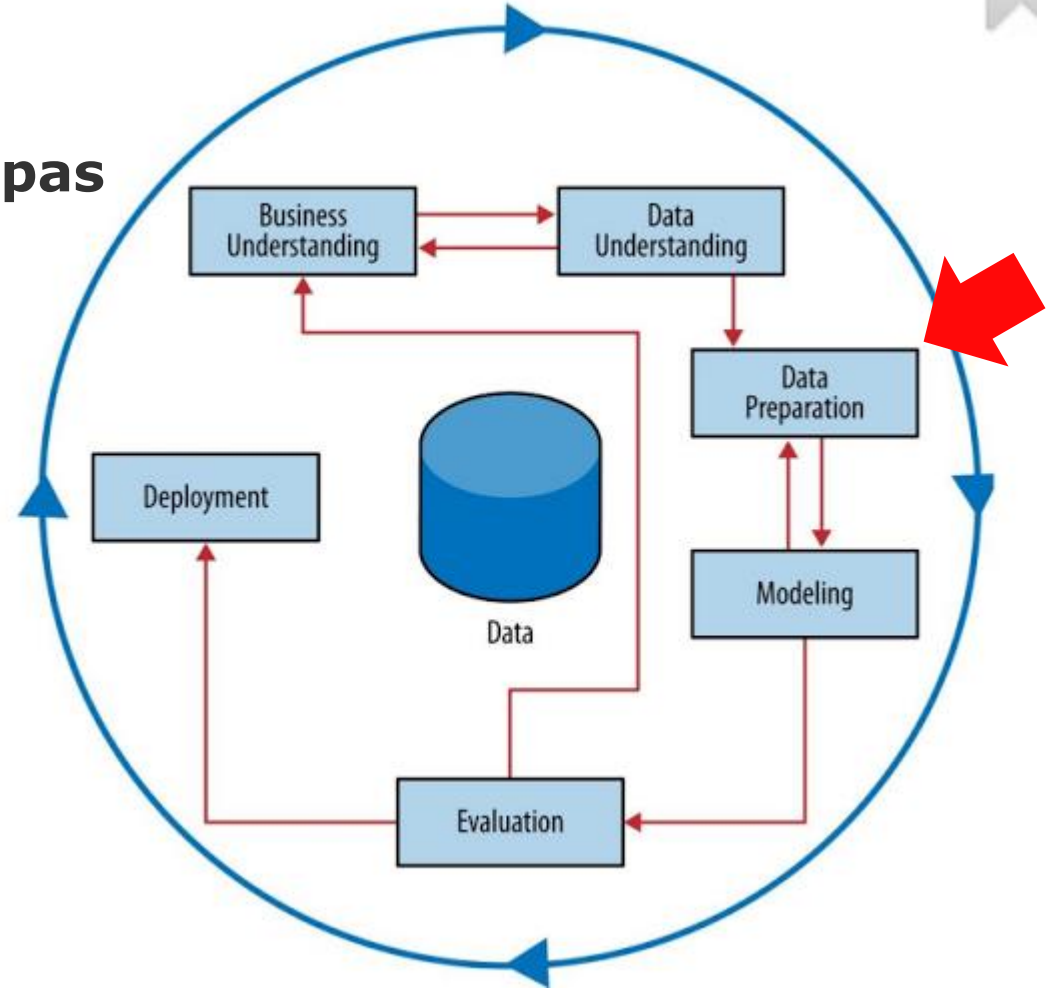


Business Intelligence
CRISP-DM

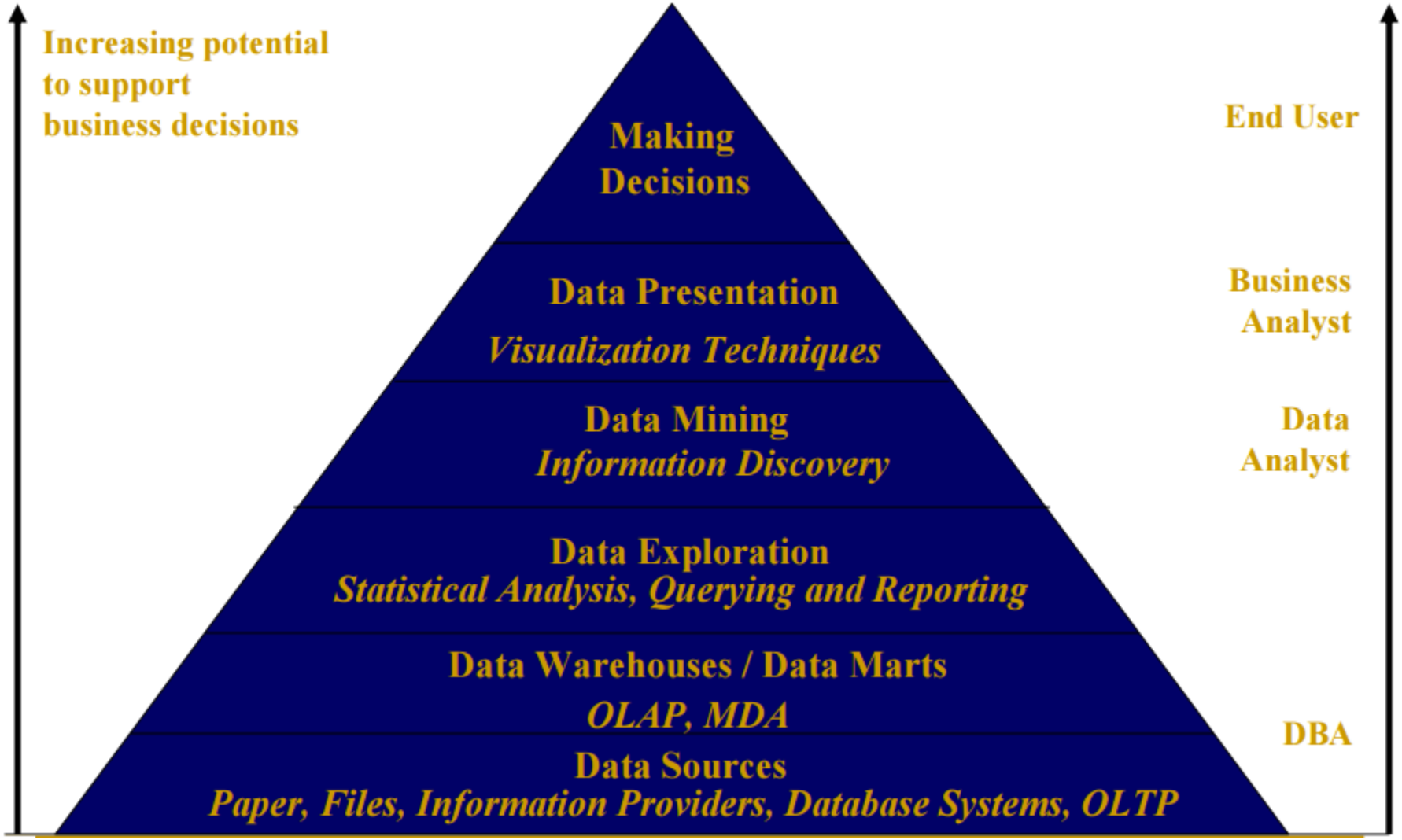
Relembrar é viver...

Ele é constituído de seis etapas

- Entendimento do Negócio
- Entendimento dos Dados
- **Preparação dos Dados**
- Modelagem
- Avaliação
- Entrega



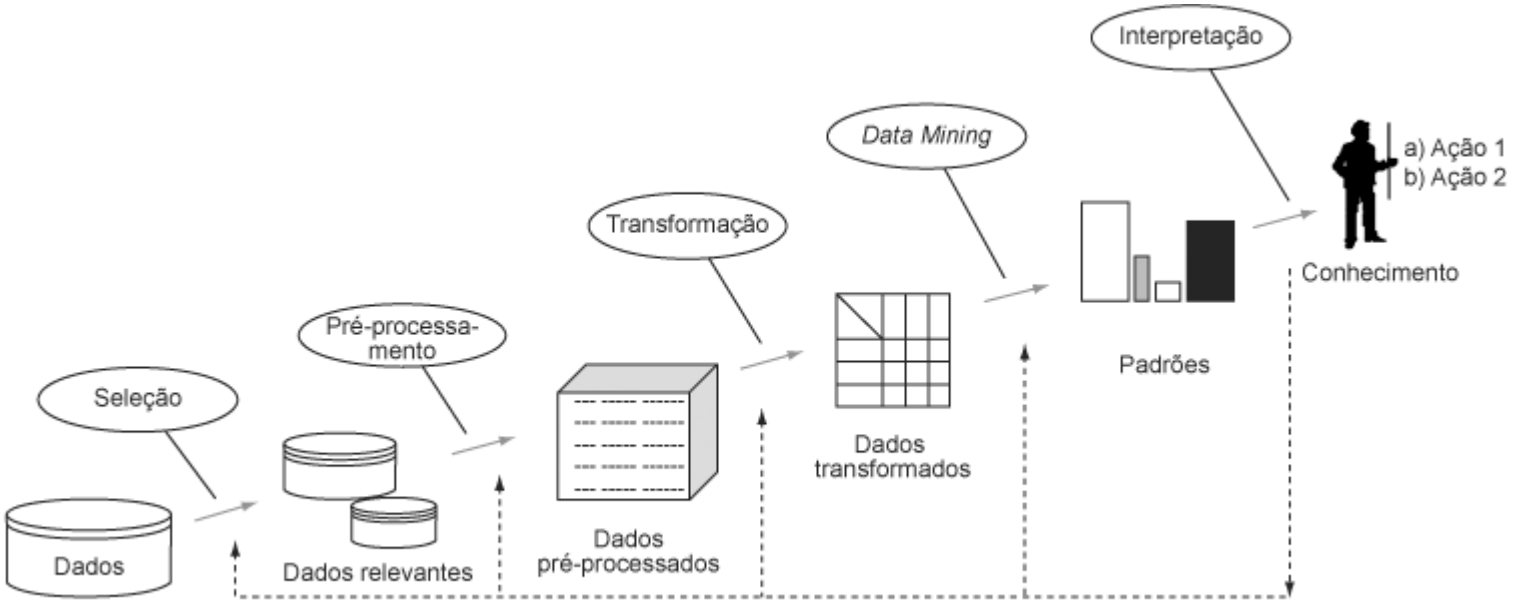
Business Intelligence
& Data Mining



*Introduction to KDD and data mining - http://www.mimuw.edu.pl/~son/datamining/DM/1-intro.pdf

KDD
Knowledge Discovery in Databases

"É o processo de descobrir conhecimento útil de uma ou mais bases de dados. É um processo amplamente utilizado, que inclui preparação dos dados, hieginação, seleção e técnicas de data mining para encontrar padrões que possam ser interpretados e transformado em conhecimento, auxiliando o processo de tomada de decisão"



*Etapas dos processo de KDD (Fayyad et. al. 1996)

KDD
Etapas

1 – Entendimento do Negócio

2 – Entendimento e escolha dos dados

3 – Data *cleaning* e pré-processamento
- Tratamento de *outliers*
- Tratamento de *missings*

4 – Featuring Engineering* e *Feature Selection

5 – Escolha da tarefa de *data mining*
- Classificação
- Regressão
- Agrupamento (*Clustering*)
- Associação

6 – Escolha do algoritmo

7 – Execução

8 – Interpretação dos resultados

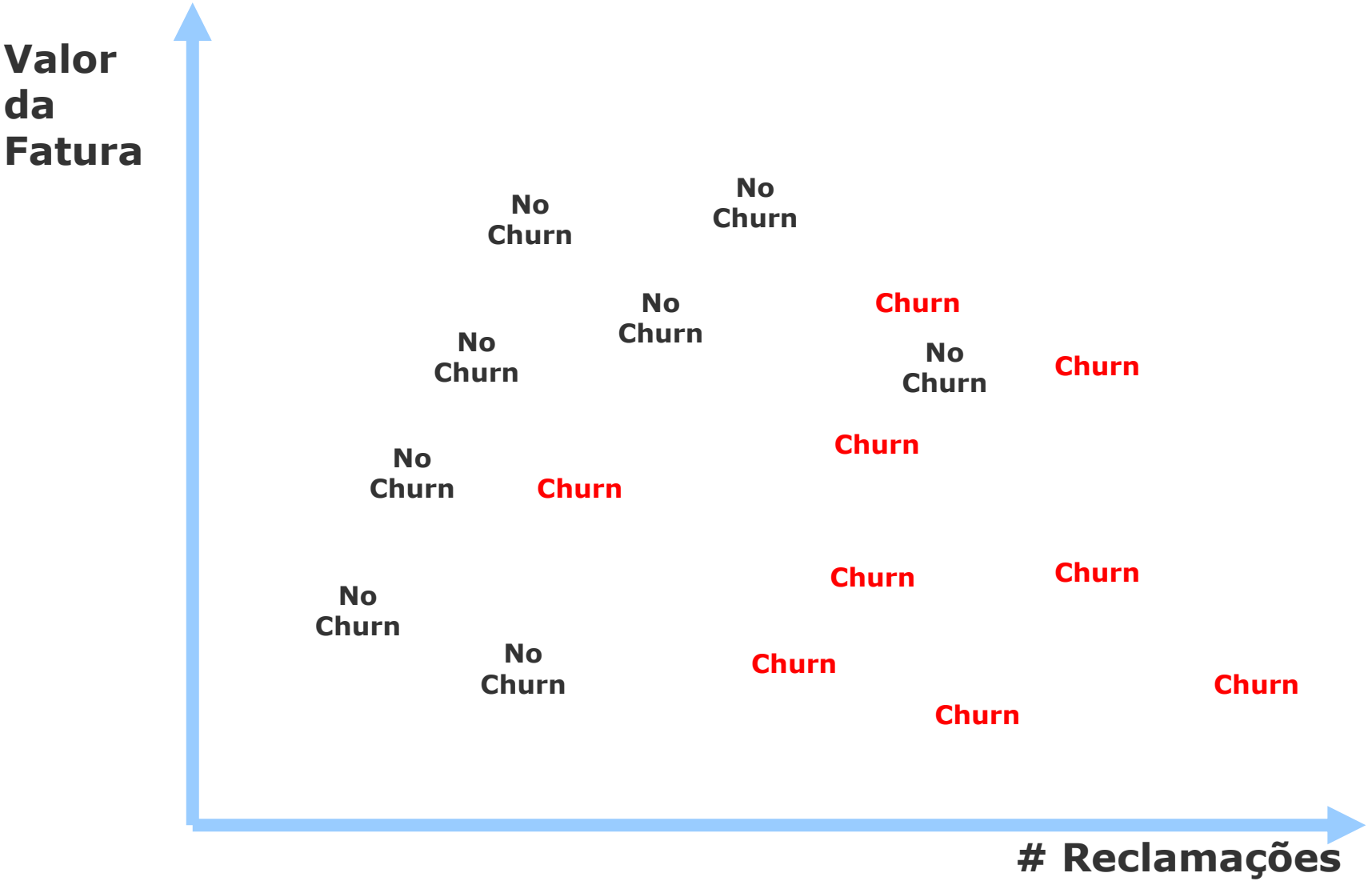
9 - Entrega

**Etapas dos processo de KDD - "From Data Mining to Knowledge Discovery: An Overview" (Fayyad et. al. 1996)*

KDD

Etapa 5 – Escolha da tarefa de data mining

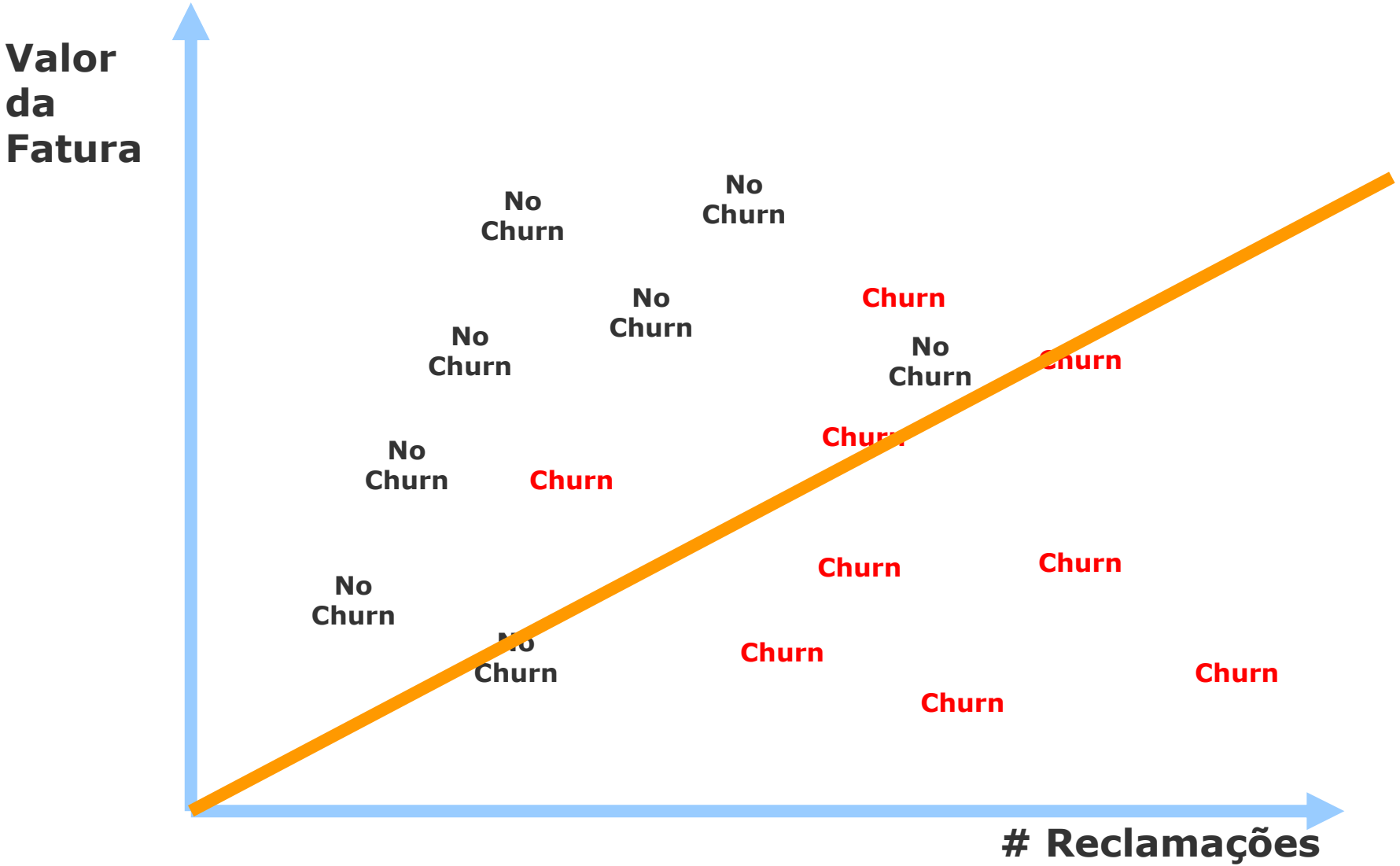
Exemplo: Customer Churn



KDD

Etapa 5 – Escolha da tarefa de data mining

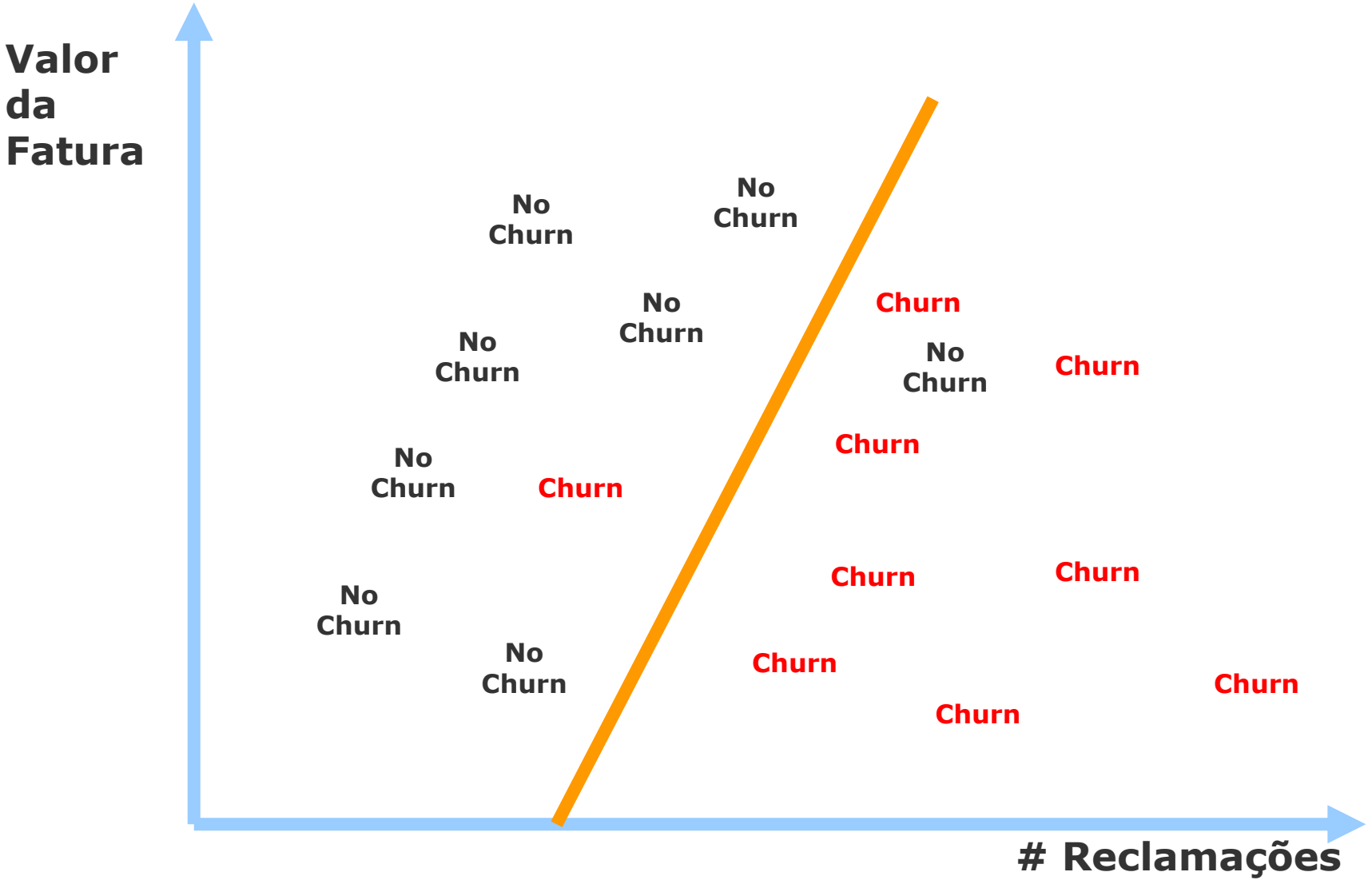
Exemplo: Regressão Linear



KDD

Etapa 5 – Escolha da tarefa de data mining

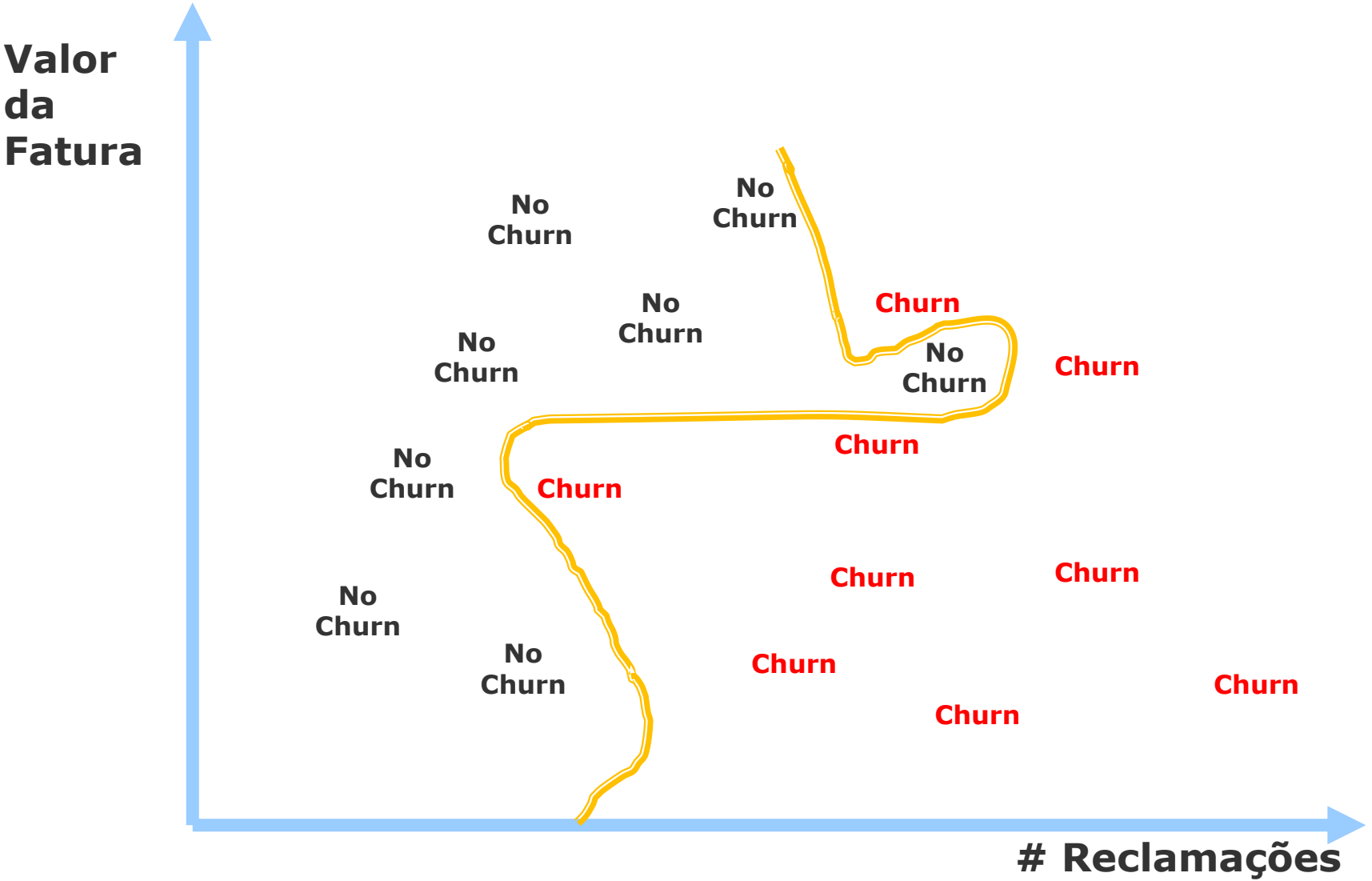
Exemplo: Classificação Linear



KDD

Etapa 5 – Escolha da tarefa de data mining

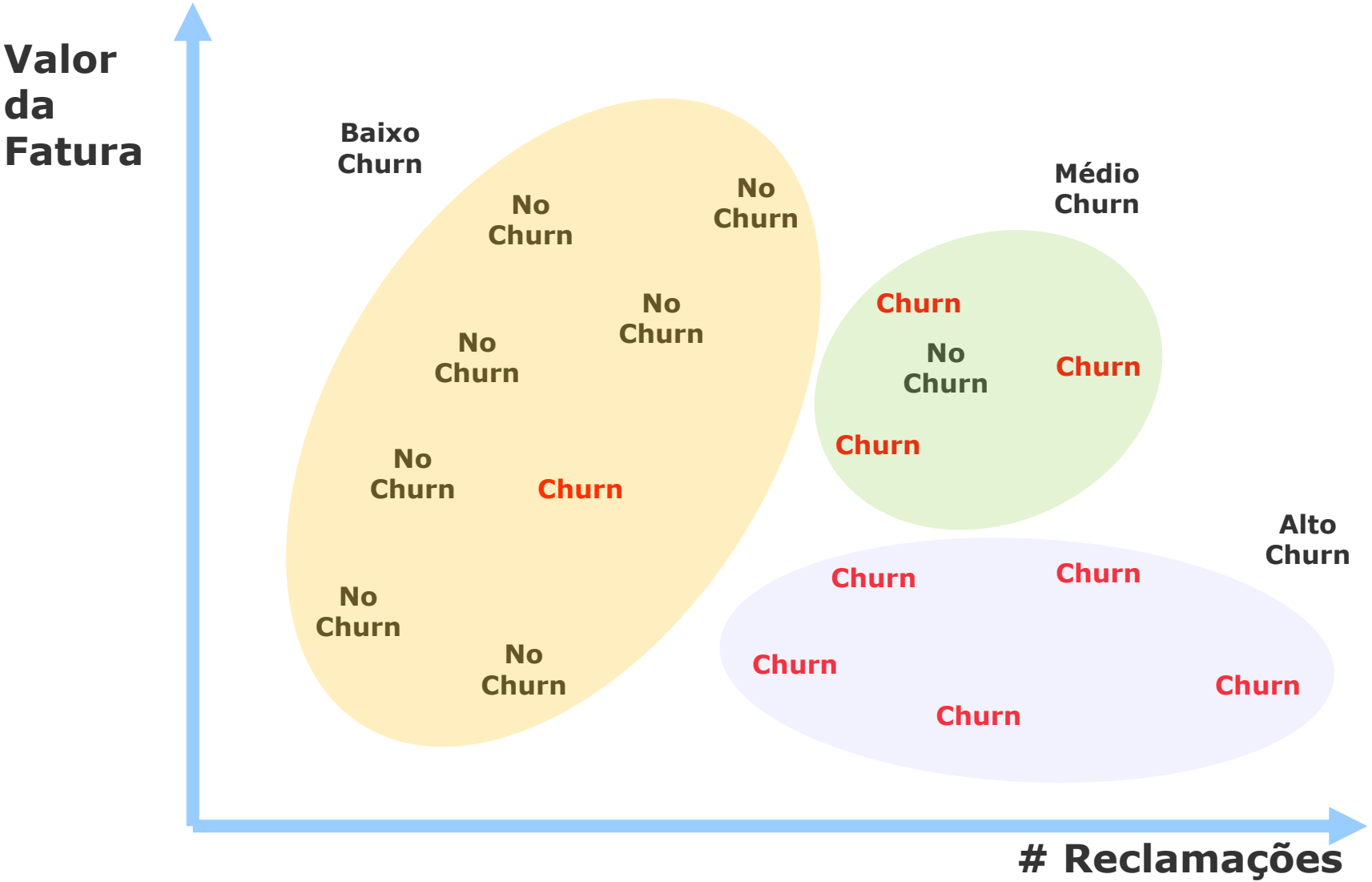
Exemplo: Classificação Não-Linear



KDD

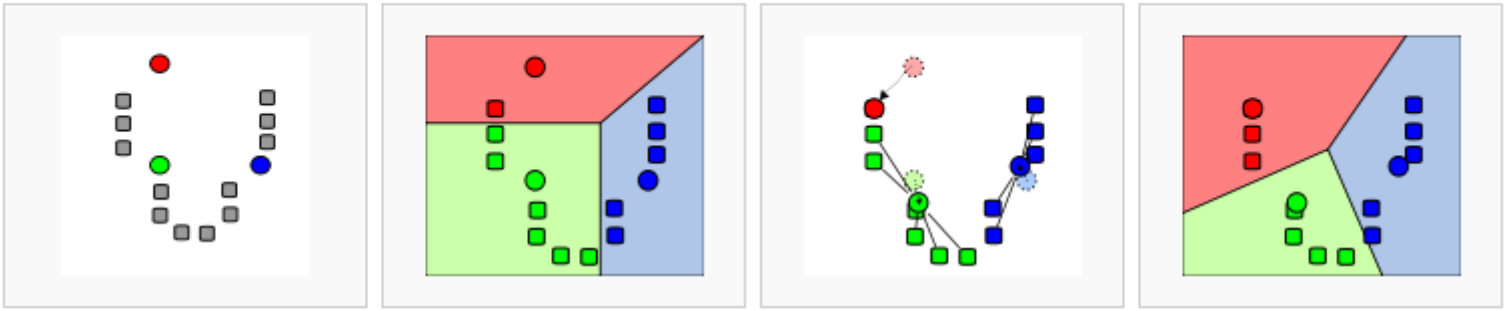
Etapa 5 – Escolha da tarefa de data mining

Exemplo: Clustering – K-Means



Clustering no R
K-Means – Demonstração do Algoritmo

Demonstration of the standard algorithm



1) *k* initial "means" (in this case *k*=3) are randomly generated within the data domain (shown in color).

2) *k* clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.

3) The centroid of each of the *k* clusters becomes the new mean.

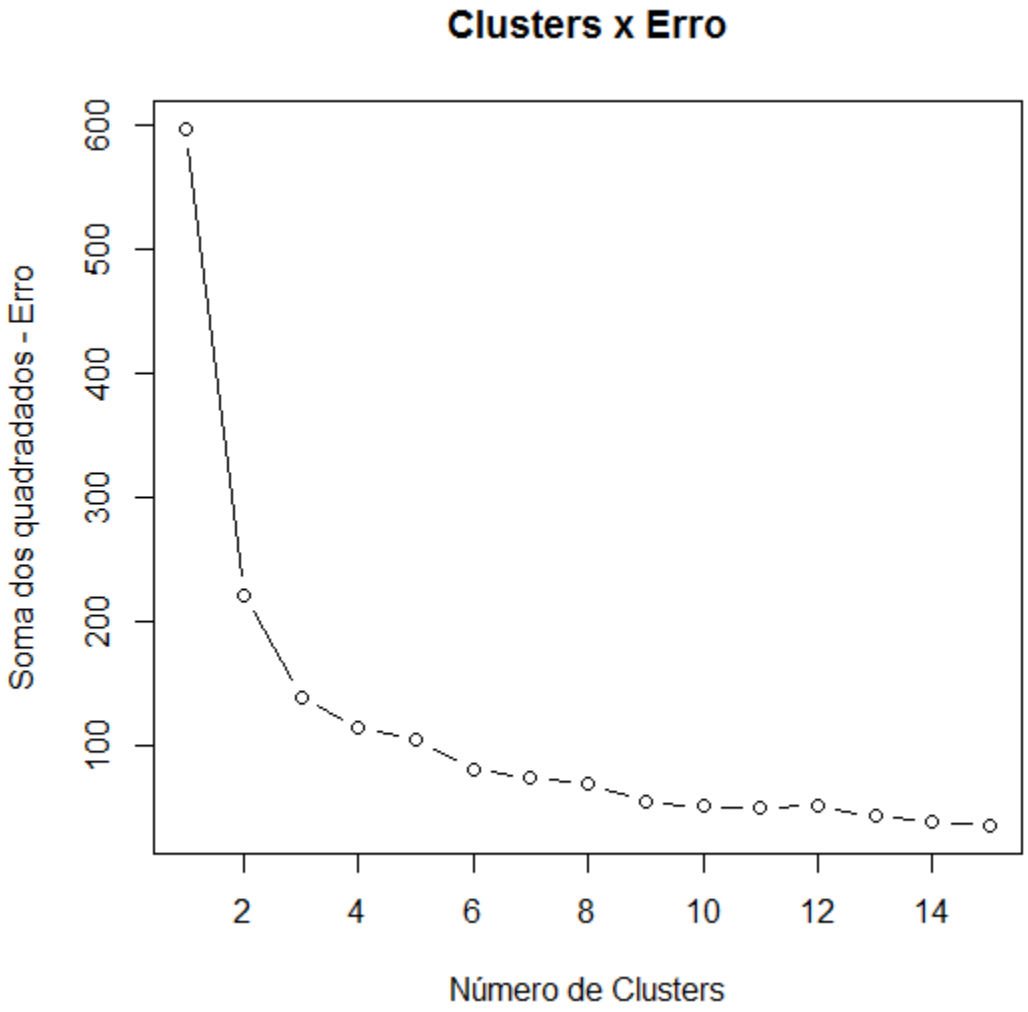
4) Steps 2 and 3 are repeated until convergence has been reached.

**http://en.wikipedia.org/wiki/K-means_clustering*

Clustering no R
K-Means

```
#Selecionando o diretorio de trabalho
source('C:/Users/Leandro/Google Drive/FMU/kmeans.R')

############################
### Preparação dos dados ###
############################
#Limpando o environment
rm(list=ls())

#Limpando os NAs
base <- na.omit(iris)

#Padronizando as variáveis
base2 <- data.frame(scale(base[,1:4]))
base2$Species <- base$Species
base <- base2

#Determinando o número de clusters
wss <- (nrow(base)-1)*sum(apply(base[,1:4],2,var))

#15 será o número de Cluster máximo para teste
for (i in 2:15){
      wss[i] <- sum(kmeans(base[,1:4],centers=i)$withinss)
      }

#Cria o gráfico mostrando o número de clusters x erro
plot(1:15, wss, type="b", xlab="Número de Clusters",
     ylab="Soma dos quadradados - Erro", main = "Clusters x Erro")
```

Clustering no R
K-Means



Clustering no R
K-Means

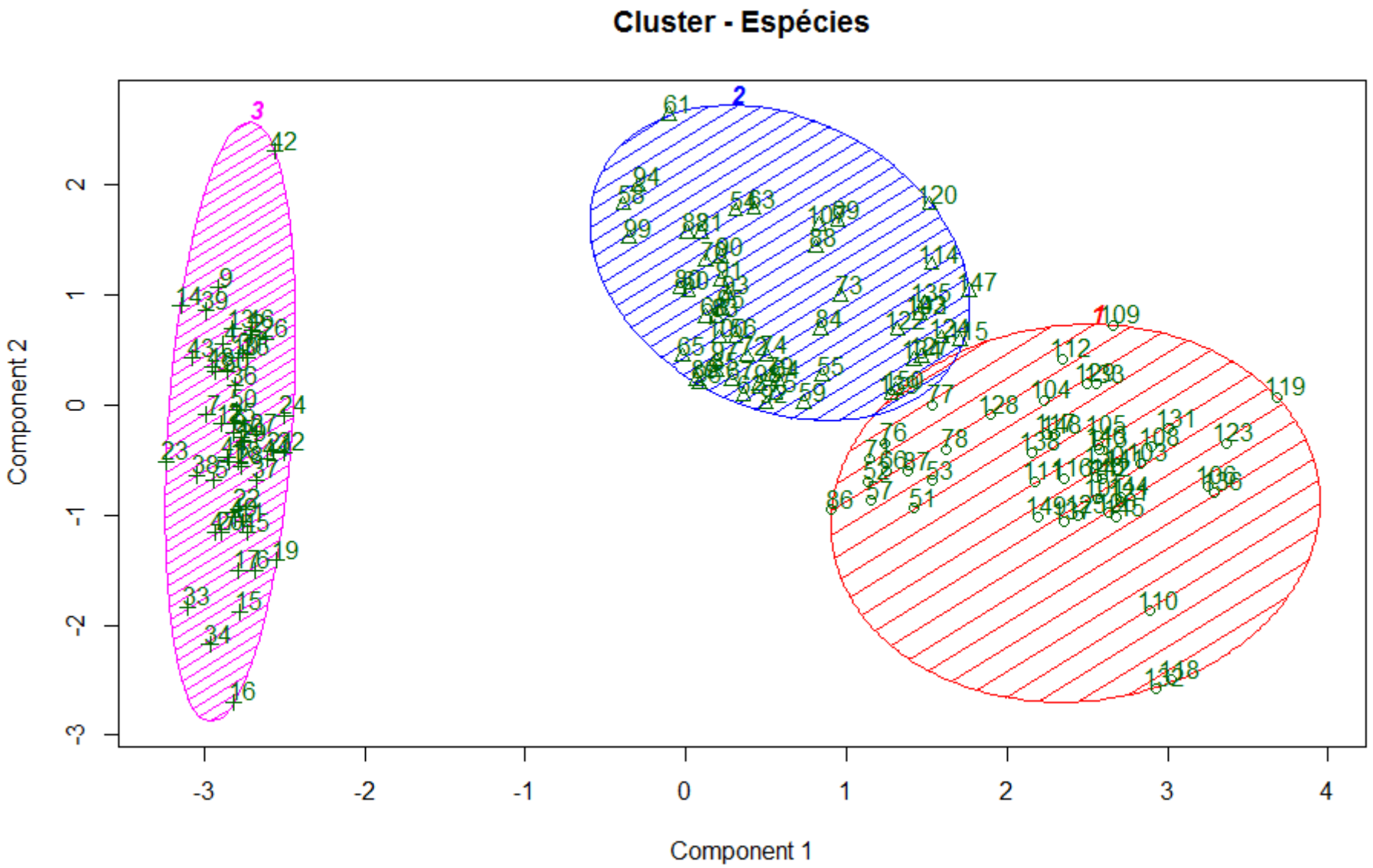
```
############################
### Executando o K-Means ###
############################

#Agrupamento com o K-Means
kmedias <- kmeans(base[,1:4], 3) #3 clusters

#Calcula a media dos clusters
aggregate(base[,1:4],by=list(kmedias$cluster),FUN=mean)

#Atribui o resultado dos cluster na base
base <- data.frame(base, kmedias$cluster)

#Verificação do Resultado
table(base$kmedias.cluster,base$Species)
```

```
> table(base$kmedias.cluster,base$Species)

   setosa versicolor virginica
1       0         11        36
2       0         39        14
3      50          0         0
```

Clustering no R

K-Means

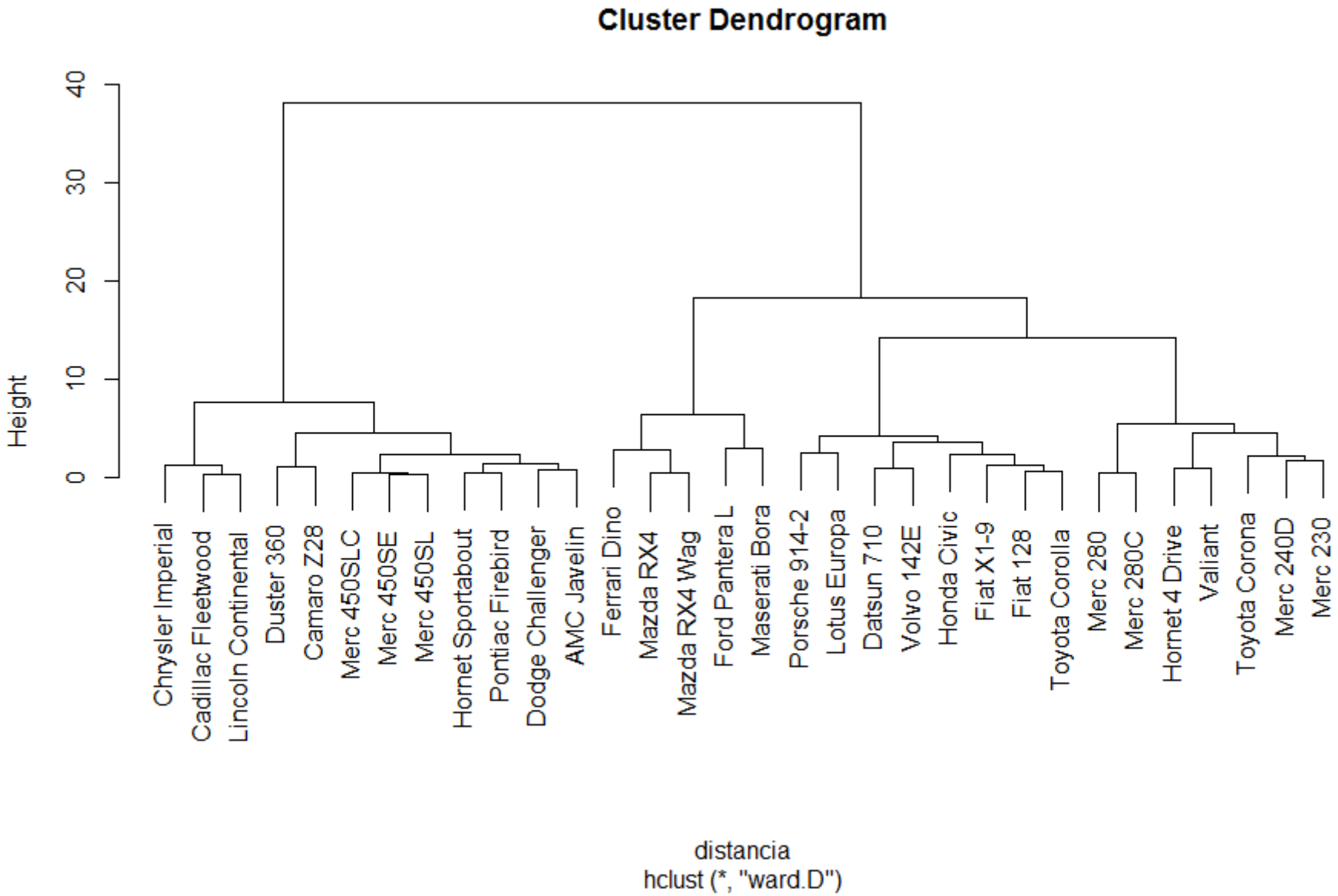
```
########################################
### Exibe os clusters, utilizando PCA ###
########################################
library(cluster)
clusplot(base, kmedias$cluster, color=TRUE, shade=TRUE,
         labels=2, lines=0,main = "Cluster - Espécies")
```



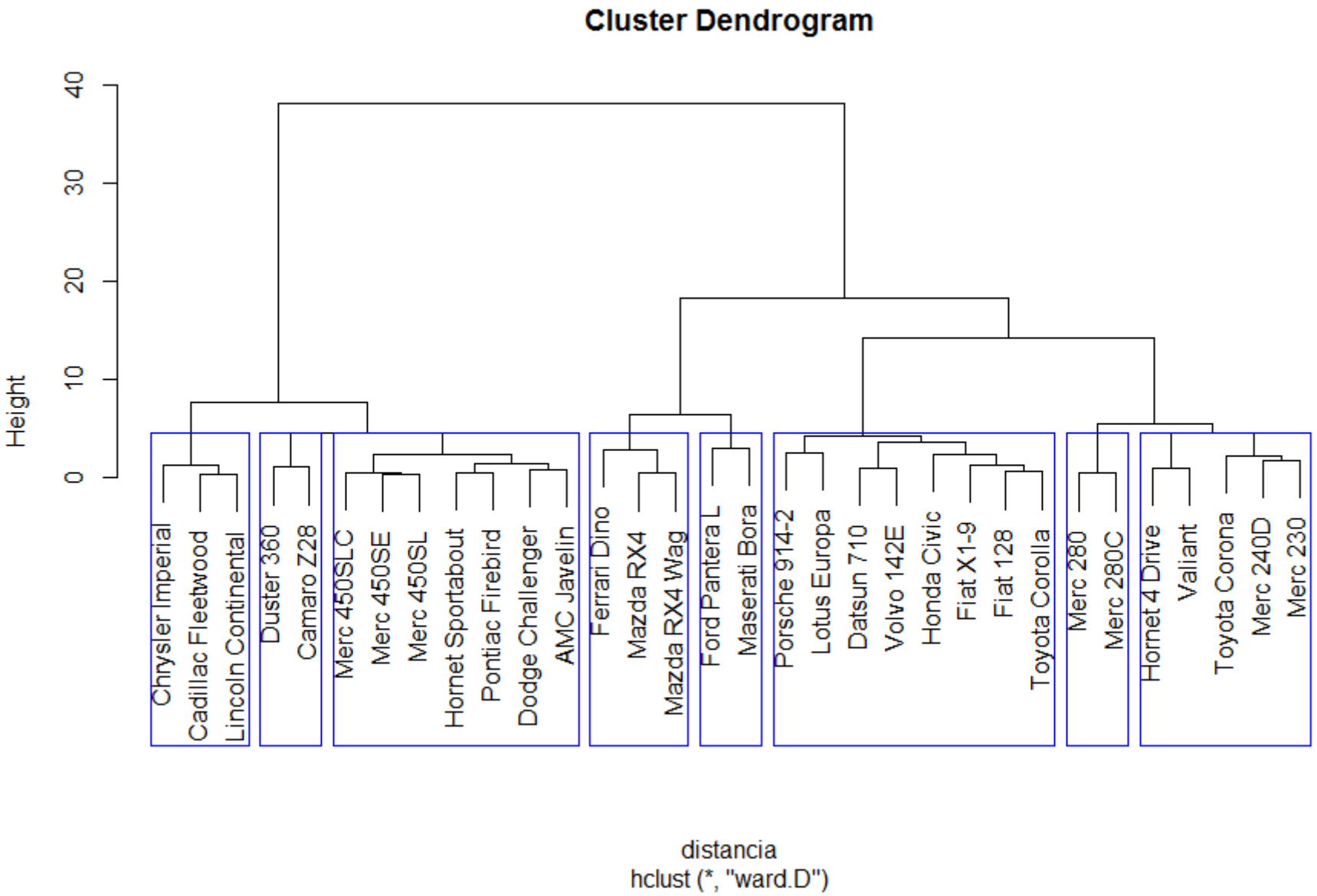
Clustering no R
Dendograma

```
############################
### Criando um Dendograma ###
############################

#Base mtcars
#Limpando os NAs
base_mtcars <- na.omit(mtcars)

#Padronizando as variáveis
base_mtcars <- scale(base_mtcars)

#Cria uma clusterização hierarquica
#Cria a matriz de distâncias
distancia <- dist(base_mtcars, method = "euclidean")
dendo <- hclust(distancia, method="ward.D")

#Plota o dendograma
plot(dendo)

#Divide o dendograma em 5 grupos
grupos <- cutree(dendo, k=5)
#Desenha uma borda azul
rect.hclust(dendo, k=8, border="blue")
```

Clustering no R
Dendograma

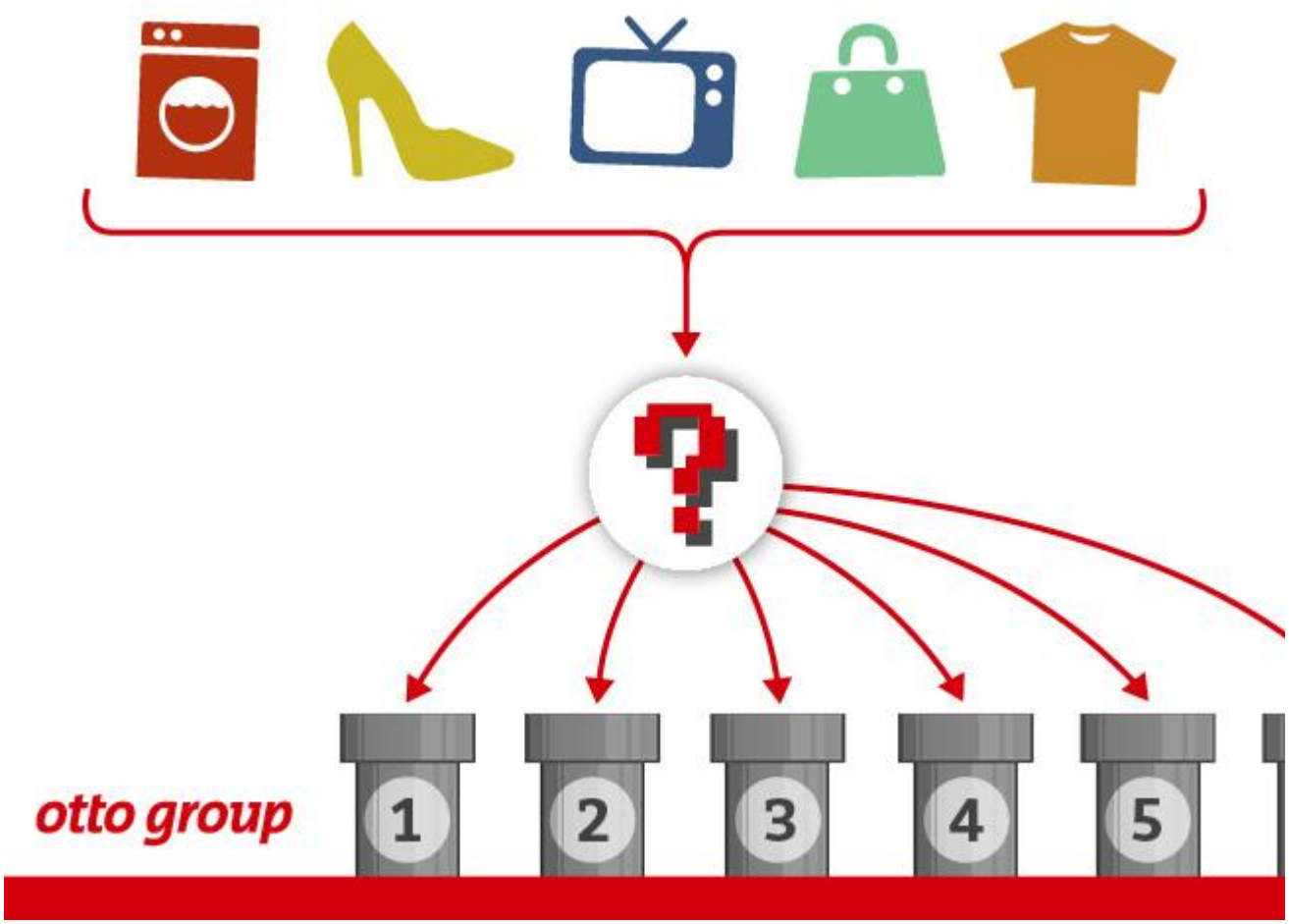


Clustering no R
Dendograma



Voltando ao Kaggle…

Otto Group Product Classification Challenge



Voltando ao Kaggle…
Otto Group Product Classification Challenge

896	new	**Leandro Guerra**	0.55719	4	Sat, 04 Apr 2015 12:46:45

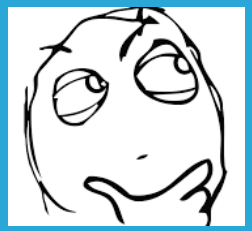
Your Best Entry ↑
You improved on your best score by 0.03766.

You just moved up 178 positions on the leaderboard. Tweet this!

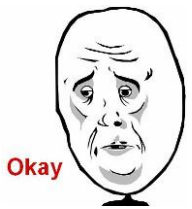
1054	new	**Leandro Guerra**	0.59485	3	Fri, 03 Apr 2015 19:22:33

Your Best Entry ↑
You improved on your best score by 14.72112.

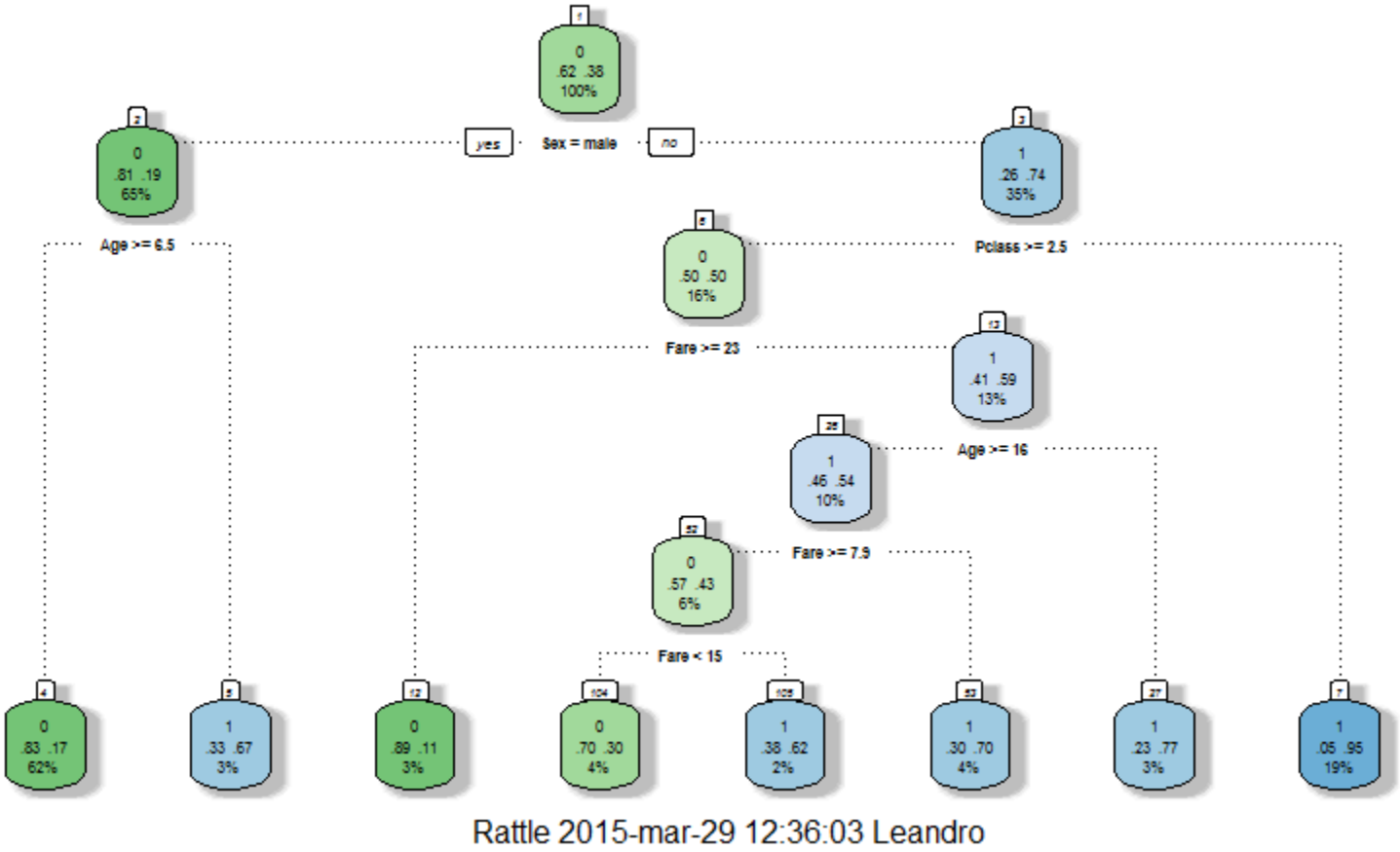
You just moved up 597 positions on the leaderboard. Tweet this!



Submitted an entry to Otto Group Product Classification Challenge, obtaining 15.31597



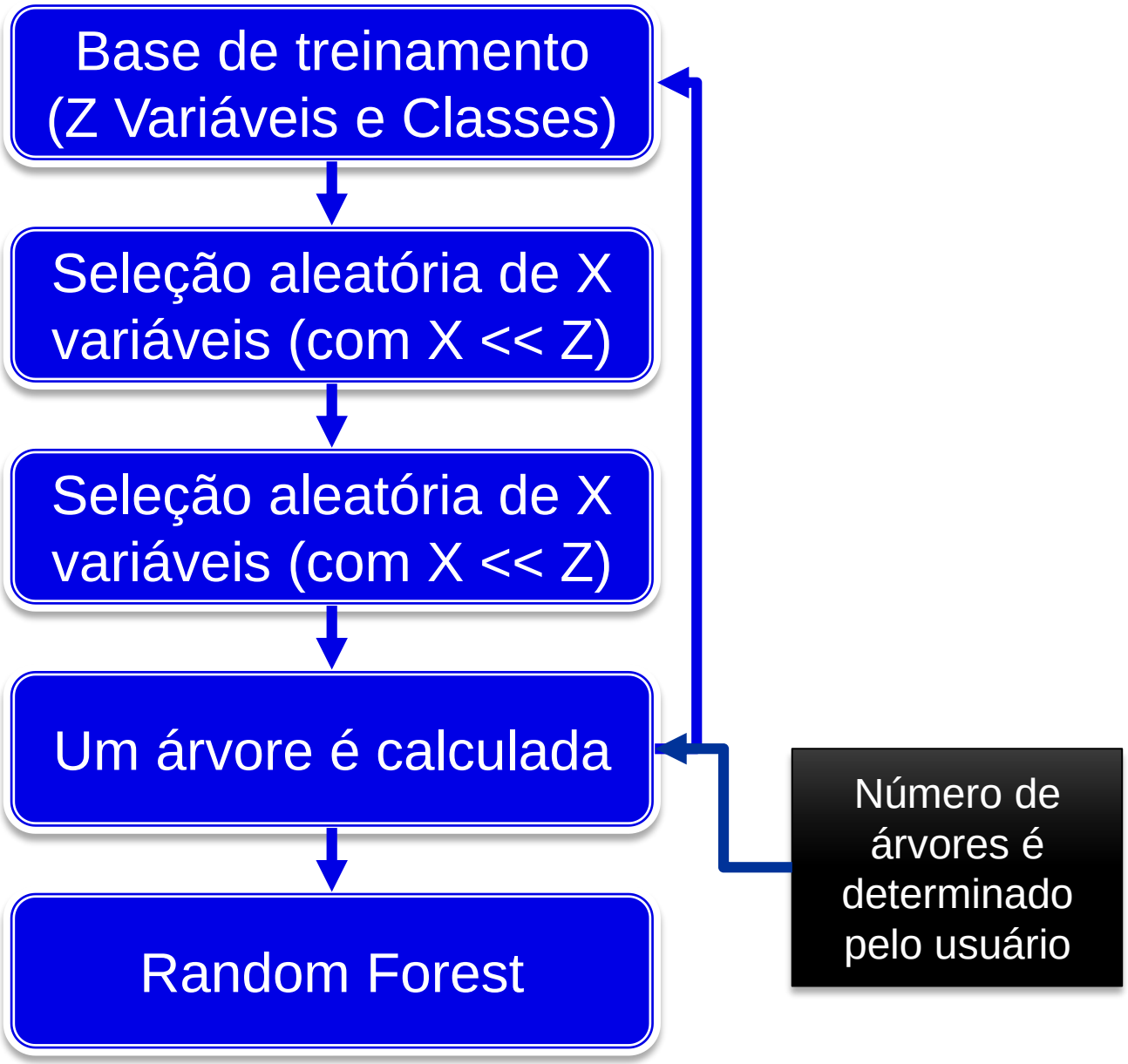
Por onde começar?
Árvore de Decisão?



Rattle 2015-mar-29 12:36:03 Leandro

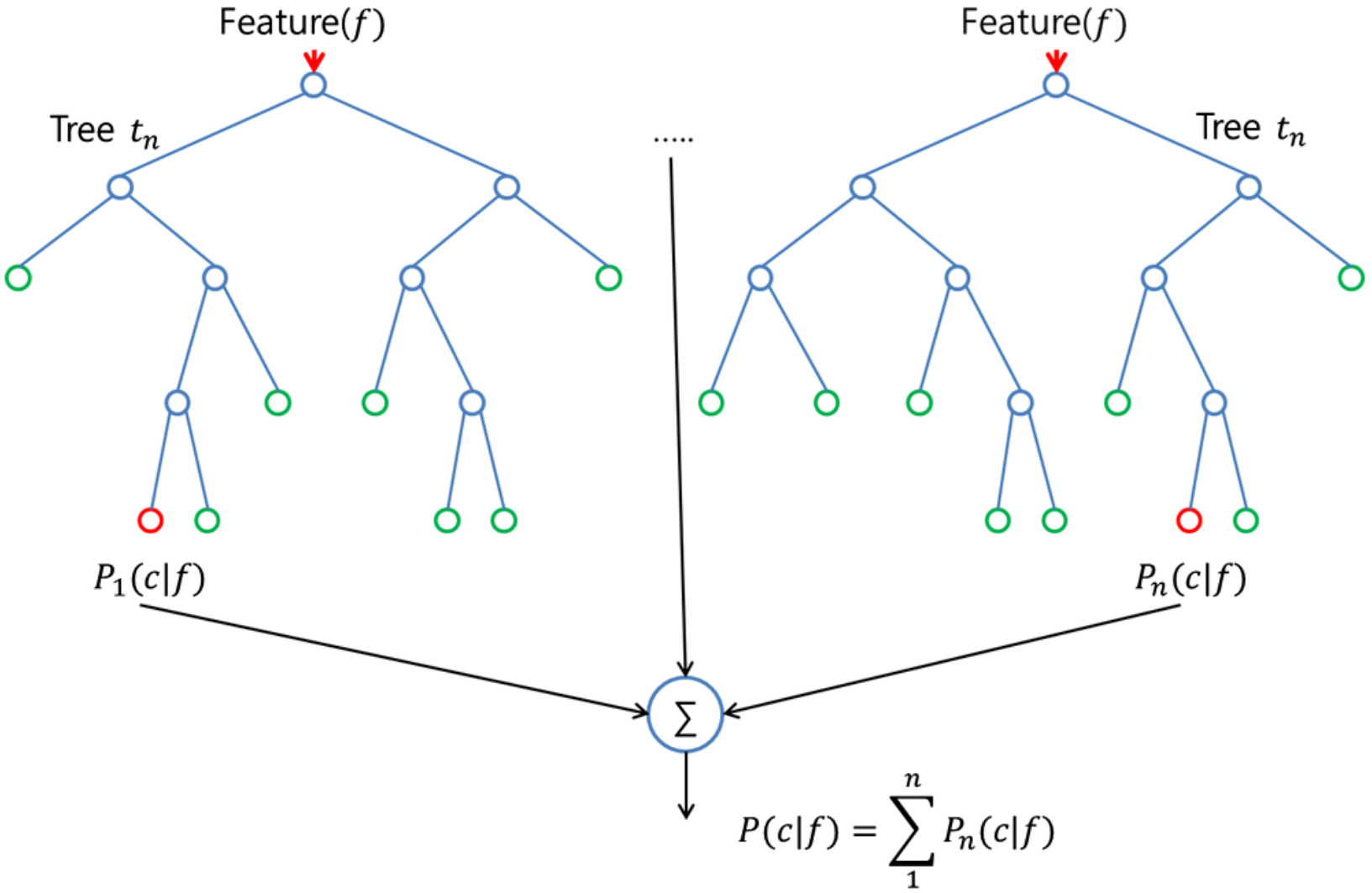
Random Forest

Uma floresta?



**David Roberts, University of Alberta - http://www.ualberta.ca/~drr3/random-forest.html*

Random forest
Decision Trees Ensemble



Random forest
Observações

Prós

Contras

Business Intelligence